Notes On Econometric Analysis
Chapter 1
Introduction to Econometrics

Econometrics deals with the measurement of economic relationships. It is an integration of economics, mathematical economics and statistics with an objective to provide numerical values to the parameters of economic relationships. The relationships of economic theories are usually expressed in mathematical forms and combined with empirical economics. The econometrics methods are used to obtain the values of parameters which are essentially the coefficients of mathematical form of the economic relationships. The statistical methods which help in explaining the economic phenomenon are adapted as econometric methods. The econometric relationships depict the random behaviour of economic relationships which are generally not considered in economics and mathematical formulations.

It may be pointed out that the econometric methods can be used in other areas like engineering sciences, biological sciences, medical sciences, geosciences, agricultural sciences etc. In simple words, whenever there is a need of finding the stochastic relationship in mathematical format, the econometric methods and tools help. The econometric tools are helpful in explaining the relationships among variables.

Econometric Models:
A model is a simplified representation of a real world process. It should be representative in the sense that it should contain the salient features of the phenomena under study. In general, one of the objectives in modeling is to have a simple model to explain a complex phenomenon. Such an objective may sometimes lead to oversimplified model and sometimes the assumptions made are unrealistic. In practice, generally all the variables which the experimenter thinks are relevant to explain the phenomenon are included in the model. Rest of the variables are dumped in a basket called “disturbances” where the disturbances are random variables. This is the main difference between the economic modeling and econometric modeling. This is also the main difference between the mathematical modeling and statistical modeling. The mathematical modeling is exact in nature whereas the statistical modeling contains a stochastic term also.

An economic model is a set of assumptions that describes the behaviour of an economy, or more general, a phenomenon.
An econometric model consists of
- a set of equations describing the behaviour. These equations are derived from the economic model and have two parts – observed variables and disturbances.
- a statement about the errors in the observed values of variables.
- a specification of the probability distribution of disturbances.

Aims of econometrics:
The three main aims econometrics are as follows:

1. Formulation and specification of econometric models:
The economic models are formulated in an empirically testable form. Several econometric models can be derived from an economic model. Such models differ due to different choice of functional form, specification of stochastic structure of the variables etc.

2. Estimation and testing of models:
The models are estimated on the basis of observed set of data and are tested for their suitability. This is the part of statistical inference of the modeling. Various estimation procedures are used to know the numerical values of the unknown parameters of the model. Based on various formulations of statistical models, a suitable and appropriate model is selected.

3. Use of models:
The obtained models are used for forecasting and policy formulation which is an essential part in any policy decision. Such forecasts help the policy makers to judge the goodness of fitted model and take necessary measures in order to re-adjust the relevant economic variables.

Econometrics and statistics:
Econometrics differs both from mathematical statistics and economic statistics. In economic statistics, the empirical data is collected recorded, tabulated and used in describing the pattern in their development over time. The economic statistics is a descriptive aspect of economics. It does not provide either the explanations of the development of various variables or measurement of the parameters of the relationships.
Statistical methods describe the methods of measurement which are developed on the basis of controlled experiments. Such methods may not be suitable for economic phenomenon as they don’t fit in the framework of controlled experiments. For example, in real world experiments, the variables usually change continuously and simultaneously and so the set up of controlled experiments are not suitable.

Econometrics uses statistical methods after adapting them to the problems of economic life. These adopted statistical methods are usually termed as econometric methods. Such methods are adjusted so that they become appropriate for the measurement of stochastic relationships. These adjustments basically attempt to specify attempts to the stochastic element which operate in real world data and enters into the determination of observed data. This enables the data to be called as random sample which is needed for the application of statistical tools.

The theoretical econometrics includes the development of appropriate methods for the measurement of economic relationships which are not meant for controlled experiments conducted inside the laboratories. The econometric methods are generally developed for the analysis of non-experimental data.

The applied econometrics includes the application of econometric methods to specific branches of econometric theory and problems like demand, supply, production, investment, consumption etc. The applied econometrics involves the application of the tools of econometric theory for the analysis of economic phenomenon and forecasting the economic behaviour.

**Types of data**

Various types of data is used in the estimation of the model.

1. **Time series data**
   Time series data give information about the numerical values of variables from period to period and are collected over time. For example, the data during the years 1990-2010 for monthly income constitutes a time series data.

2. **Cross section data**
   The cross section data give information on the variables concerning individual agents (e.g., consumers or producers) at a given point of time. For example, a cross section of sample of consumers is a sample of family budgets showing expenditures on various commodities by each family, as well as information on family income, family composition and other demographic, social or financial characteristics.
3. Panel data:
The panel data are the data from repeated survey of a single (cross-section) sample in different periods of time.

4. Dummy variable data
When the variables are qualitative in nature, then the data is recorded in the form of indicator function. The values of the variables do not reflect the magnitude of data. They reflect only the presence/absence of a characteristic. For example, the variables like religion, sex, taste, etc. are qualitative variables. The variable ‘sex’ takes two values – male or female, the variable ‘taste’ takes values-like or dislike etc. Such values are denoted by dummy variable. For example, these values can be represented as ‘1’ represents male and ‘0’ represents female. Similarly, ‘1’ represents the liking of taste and ‘0’ represents the disliking of taste.

Aggregation problem:
The aggregation problems arise when aggregative variables are used in functions. Such aggregative variables may involve.

1. Aggregation over individuals:
For example the total income may comprise the sum of individual incomes.

2. Aggregation over commodities:
The quantity of various commodities may be aggregated over e.g., price or group of commodities. This is done by using suitable index.

3. Aggregation over time periods
Sometimes the data is available for shorter or longer time periods than required to be used in the functional form of economic relationship. In such cases, the data needs to be aggregated over time period. For example, the production of most of the manufacturing commodities is completed in a period shorter than a year. If annual figures are to be used in the model then there may be some error in the production function.

4. Spatial aggregation:
Sometimes the aggregation is related to spatial issues. For example, the population of towns, countries, or the production in a city or region etc..

Such sources of aggregation introduces “aggregation bias” in the estimates of the coefficients. It is important to examine the possibility of such errors before estimating the model.
Econometrics and regression analysis:

One of the very important role of econometrics is to provide the tools for modeling on the basis of given data. The regression modeling technique helps a lot in this task. The regression models can be either linear or non-linear based on which we have linear regression analysis and non-linear regression analysis. We will consider only the tools of linear regression analysis and our main interest will be the fitting of linear regression model to a given set of data.

Linear regression model

Suppose the outcome of any process is denoted by a random variable \( y \), called as dependent (or study) variable, depends on \( k \) independent (or explanatory) variables denoted by \( X_1, X_2, \ldots, X_k \). Suppose the behaviour of \( y \) can be explained by a relationship given by

\[
y = f(X_1, X_2, \ldots, X_k, \beta_1, \beta_2, \ldots, \beta_k) + \varepsilon
\]

where \( f \) is some well defined function and \( \beta_1, \beta_2, \ldots, \beta_k \) are the parameters which characterize the role and contribution of \( X_1, X_2, \ldots, X_k \), respectively. The term \( \varepsilon \) reflects the stochastic nature of the relationship between \( y \) and \( X_1, X_2, \ldots, X_k \) and indicates that such a relationship is not exact in nature. When \( \varepsilon = 0 \), then the relationship is called the mathematical model otherwise the statistical model. The term “model” is broadly used to represent any phenomenon in a mathematical framework.

A model or relationship is termed as linear if it is linear in parameters and nonlinear, if it is not linear in parameters. In other words, if all the partial derivatives of \( y \) with respect to each of the parameters \( \beta_1, \beta_2, \ldots, \beta_k \) are independent of the parameters, then the model is called as a linear model. If any of the partial derivatives of \( y \) with respect to any of the \( \beta_1, \beta_2, \ldots, \beta_k \) is not independent of the parameters, the model is called as nonlinear. Note that the linearity or non-linearity of the model is not described by the linearity or nonlinearity of explanatory variables in the model.

For example

\[
y = \beta_1 X_1^2 + \beta_2 \sqrt{X_2} + \beta_3 \log X + \varepsilon
\]

is a linear model because \( \frac{\partial y}{\partial \beta_i}, \ (i = 1, 2, 3) \) are independent of the parameters \( \beta_i, (i = 1, 2, 3) \). On the other hand,

\[
y = \beta_1^2 X_1 + \beta_2 X_2 + \beta_3 \log X + \varepsilon
\]
is a nonlinear model because \( \partial y / \partial \beta = 2 \beta X \) depends on \( \beta \) although \( \partial y / \partial \beta_2 \) and \( \partial y / \partial \beta_3 \) are independent of any of the \( \beta_1, \beta_2 \) or \( \beta_3 \).

When the function \( f \) is linear in parameters, then \( y = f(X_1, X_2, ..., X_k, \beta_1, \beta_2, ..., \beta_k) + \epsilon \) is called a linear model and when the function \( f \) is nonlinear in parameters, then it is called a nonlinear model. In general, the function \( f \) is chosen as

\[
f(X_1, X_2, ..., X_k, \beta_1, \beta_2, ..., \beta_k) = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k
\]
to describe a linear model. Since \( X_1, X_2, ..., X_k \) are pre-determined variables and \( y \) is the outcome, so both are known. Thus the knowledge of the model depends on the knowledge of the parameters \( \beta_1, \beta_2, ..., \beta_k \).

The statistical linear modeling essentially consists of developing approaches and tools to determine \( \beta_1, \beta_2, ..., \beta_k \) in the linear model

\[
y = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon
\]
given the observations on \( y \) and \( X_1, X_2, ..., X_k \).

Different statistical estimation procedures, e.g., method of maximum likelihood, principle of least squares, method of moments etc. can be employed to estimate the parameters of the model. The method of maximum likelihood needs further knowledge of the distribution of \( y \) whereas the method of moments and the principle of least squares do not need any knowledge about the distribution of \( y \).

The regression analysis is a tool to determine the values of the parameters given the data on \( y \) and \( X_1, X_2, ..., X_k \). The literal meaning of regression is “to move in the backward direction”. Before discussing and understanding the meaning of “backward direction”, let us find which of the following statements is correct:

- S1: model generates data or
- S2: data generates model.

Obviously, S1 is correct. It can be broadly thought that the model exists in nature but is unknown to the experimenter. When some values to the explanatory variables are provided, then the values for the output or study variable are generated accordingly, depending on the form of the function \( f \) and the nature of phenomenon. So ideally, the pre-existing model gives rise to the data. Our objective is to determine the
functional form of this model. Now we move in the backward direction. We propose to first collect the data on study and explanatory variables. Then we employ some statistical techniques and use this data to know the form of function $f$. Equivalently, the data from the model is recorded first and then used to determine the parameters of the model. The regression analysis is a technique which helps in determining the statistical model by using the data on study and explanatory variables. The classification of linear and nonlinear regression analysis is based on the determination of linear and nonlinear models, respectively.

Consider a simple example to understand the meaning of “regression”. Suppose the yield of crop ($y$) depends linearly on two explanatory variables, viz., the quantity of a fertilizer ($X_1$) and level of irrigation ($X_2$) as

$$y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$  

There exist the true values of $\beta_1$ and $\beta_2$ in nature but are unknown to the experimenter. Some values on $y$ are recorded by providing different values to $X_1$ and $X_2$. There exists some relationship between $y$ and $X_1, X_2$ which gives rise to a systematically behaved data on $y$, $X_1$ and $X_2$. Such relationship is unknown to the experimenter. To determine the model, we move in the backward direction in the sense that the collected data is used to determine the unknown parameters $\beta_1$ and $\beta_2$ of the model. In this sense such an approach is termed as regression analysis.

The theory and fundamentals of linear models lay the foundation for developing the tools for regression analysis that are based on valid statistical theory and concepts.

**Steps in regression analysis**

Regression analysis includes the following steps:

- Statement of the problem under consideration
- Choice of relevant variables
- Collection of data on relevant variables
- Specification of model
- Choice of method for fitting the data
- Fitting of model
- Model validation and criticism
- Using the chosen model(s) for the solution of the posed problem.
These steps are examined below.

1. **Statement of the problem under consideration:**

The first important step in conducting any regression analysis is to specify the problem and the objectives to be addressed by the regression analysis. The wrong formulation or the wrong understanding of the problem will give the wrong statistical inferences. The choice of variables depends upon the objectives of study and understanding of the problem. For example, height and weight of children are related. Now there can be two issues to be addressed.

   (i) Determination of height for given weight, or
   (ii) determination of weight for given height.

In the case 1, the height is response variable whereas weight is response variable in case 2. The role of explanatory variables are also interchanged in the cases 1 and 2.

2. **Choice of relevant variables:**

Once the problem is carefully formulated and objectives have been decided, the next question is to choose the relevant variables. It has to kept in mind that the correct choice of variables will determine the statistical inferences correctly. For example, in any agricultural experiment, the yield depends on explanatory variables like quantity of fertilizer, rainfall, irrigation, temperature etc. These variables are denoted by $X_1, X_2, ..., X_k$ as a set of $k$ explanatory variables.

3. **Collection of data on relevant variables:**

Once the objective of study is clearly stated and the variables are chosen, the next question arises is to collect data on such relevant variables. The data is essentially the measurement on these variables. For example, suppose we want to collect the data on age. For this, it is important to know how to record the data on age. Then either the date of birth can be recorded which will provide the exact age on any specific date or the age in terms of completed years as on specific date can be recorded. Moreover, it is also important to decide that whether the data has to be collected on variables as quantitative variables or qualitative variables. For example, if the ages (in years) are 15,17,19,21,23, then these are quantitative values. If the ages are defined by a variable that takes value 1 if ages are less than 18 years and 0 if the ages are more than 18 years, then the earlier recorded data is converted to 1,1,0,0,0. Note that there is a loss of information in converting the quantitative data into qualitative data. The methods and approaches for qualitative and quantitative data are also different. If the study variable is binary, then logistic and probit...
regressions etc. are used. If all explanatory variables are qualitative, then analysis of variance technique is used. If some explanatory variables are qualitative and others are quantitative, then analysis of covariance technique is used. The techniques of analysis of variance and analysis of covariance are the special cases of regression analysis.

Generally, the data is collected on $n$ subjects, then $y$ on data, then $y$ denotes the response or study variable and $y_1, y_2, \ldots, y_n$ are the $n$ values. If there are $k$ explanatory variables $X_1, X_2, \ldots, X_k$ then $x_{ij}$ denotes the $i^{th}$ value of $j^{th}$ variable $i = 1, 2, \ldots, n; j = 1, 2, \ldots, k$. The observation can be presented in the following table:

Notation for the data used in regression analysis

<table>
<thead>
<tr>
<th>Observation number</th>
<th>Response $y$</th>
<th>Explanatory variables $X_1$</th>
<th>$X_2$</th>
<th>$\ldots$</th>
<th>$X_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$y_1$</td>
<td>$x_{11}$</td>
<td>$x_{12}$</td>
<td>$\ldots$</td>
<td>$x_{1k}$</td>
</tr>
<tr>
<td>2</td>
<td>$y_2$</td>
<td>$x_{21}$</td>
<td>$x_{22}$</td>
<td>$\ldots$</td>
<td>$x_{2k}$</td>
</tr>
<tr>
<td>3</td>
<td>$y_3$</td>
<td>$x_{31}$</td>
<td>$x_{32}$</td>
<td>$\ldots$</td>
<td>$x_{3k}$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\ldots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$n$</td>
<td>$y_n$</td>
<td>$x_{n1}$</td>
<td>$x_{n2}$</td>
<td>$\ldots$</td>
<td>$x_{nk}$</td>
</tr>
</tbody>
</table>

4. Specification of model:

The experimenter or the person working in the subject usually help in determining the form of the model. Only the form of the tentative model can be ascertained and it will depend on some unknown parameters. For example, a general form will be like

$$y = f(X_1, X_2, \ldots, X_k; \beta_1, \beta_2, \ldots, \beta_k) + \varepsilon$$

where $\varepsilon$ is the random error reflecting mainly the difference in the observed value of $y$ and the value of $y$ obtained through the model. The form of $f(X_1, X_2, \ldots, X_k; \beta_1, \beta_2, \ldots, \beta_k)$ can be linear as well as nonlinear depending on the form of parameters $\beta_1, \beta_2, \ldots, \beta_k$. A model is said to be linear if it is linear in parameters. For example,
\[ y = \beta_1X_1 + \beta_2X_1^2 + \beta_3X_2 + \varepsilon \]
\[ y = \beta_1 + \beta_2 \ln X_2 + \varepsilon \]

are linear models whereas
\[ y = \beta_1X_1 + \beta_2^2X_2 + \beta_3X_2 + \varepsilon \]
\[ y = (\ln \beta_1)X_1 + \beta_2X_2 + \varepsilon \]

are non-linear models. Many times, the nonlinear models can be converted into linear models through some transformations. So the class of linear models is wider than what it appears initially.

If a model contains only one explanatory variable, then it is called as simple regression model. When there are more than one independent variables, then it is called as multiple regression model. When there is only one study variable, the regression is termed as univariate regression. When there are more than one study variables, the regression is termed as multivariate regression. Note that the simple and multiple regressions are not same as univariate and multivariate regressions. The simple and multiple regression are determined by the number of explanatory variables whereas univariate and multivariate regressions are determined by the number of study variables.

5. Choice of method for fitting the data:

After the model has been defined and the data have been collected, the next task is to estimate the parameters of the model based on the collected data. This is also referred to as parameter estimation or model fitting. The most commonly used method of estimation is the least squares method. Under certain assumptions, the least squares method produces estimators with desirable properties. The other estimation methods are the maximum likelihood method, ridge method, principal components method etc.

6. Fitting of model:

The estimation of unknown parameters using appropriate method provides the values of the parameter. Substituting these values in the equation gives us a usable model. This is termed as model fitting. The estimates of parameters \( \beta_1, \beta_2, \ldots, \beta_k \) in the model
\[ y = f(X_1, X_2, \ldots, X_k, \beta_1, \beta_2, \ldots, \beta_k) + \varepsilon \]
are denoted by \( \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k \) which gives the fitted model as
\[ y = f(X_1, X_2, \ldots, X_k, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k). \]
When the value of \( y \) is obtained for the given values of \( X_1, X_2, \ldots, X_k \), it is denoted as \( \hat{y} \) and called as fitted value.

The fitted equation is used for prediction. In this case, \( \hat{y} \) is termed as **predicted value**. Note that the fitted value is where the values used for explanatory variables correspond to one of the \( n \) observations in the data whereas predicted value is the one obtained for any set of values of explanatory variables. It is not generally recommended to predict the \( y \)-values for the set of those values of explanatory variables which lie outside the range of data. When the values of explanatory variables are the future values of explanatory variables, the predicted values are called forecasted values.

### 7. Model criticism and selection

The validity of statistical method to be used for regression analysis depends on various assumptions. These assumptions become essentially the assumptions for the model and the data. The quality of statistical inferences heavily depends on whether these assumptions are satisfied or not. For making these assumptions to be valid and to be satisfied, care is needed from beginning of the experiment. One has to be careful in choosing the required assumptions and to decide as well to determine if the assumptions are valid for the given experimental conditions or not? It is also important to decide that the situations is which the assumptions may not meet.

The validation of the assumptions must be made before drawing any statistical conclusion. Any departure from validity of assumptions will be reflected in the statistical inferences. In fact, the regression analysis is an iterative process where the outputs are used to diagnose, validate, criticize and modify the inputs. The iterative process is illustrated in the following figure.
8. Objectives of regression analysis

The determination of explicit form of regression equation is the ultimate objective of regression analysis. It is finally a good and valid relationship between study variable and explanatory variables. The regression equation helps in understanding the interrelationships of variables among them. Such regression equation can be used for several purposes. For example, to determine the role of any explanatory variable in the joint relationship in any policy formulation, to forecast the values of response variable for given set of values of explanatory variables.
Chapter 2
Simple Linear Regression Analysis

The simple linear regression model
We consider the modeling between the dependent and one independent variable. When there is only one independent variable in the linear regression model, the model is generally termed as simple linear regression model. When there are more than one independent variables in the model, then the linear model is termed as the multiple linear regression model.

The linear model
Consider a simple linear regression model
\[ y = \beta_0 + \beta_1 X + \varepsilon \]
where \( y \) is termed as the dependent or study variable and \( X \) is termed as independent or explanatory variable. The terms \( \beta_0 \) and \( \beta_1 \) are the parameters of the model. The parameter \( \beta_0 \) is termed as intercept term and the parameter \( \beta_1 \) is termed as slope parameter. These parameters are usually called as regression coefficients. The unobservable error component \( \varepsilon \) accounts for the failure of data to lie on the straight line and represents the difference between the true and observed realization of \( y \). There can be several reasons for such difference, e.g., the effect of all deleted variables in the model, variables may be qualitative, inherit randomness in the observations etc. We assume that \( \varepsilon \) is observed as independent and identically distributed random variable with mean zero and constant variance \( \sigma^2 \). Later, we will additionally assume that \( \varepsilon \) is normally distributed.

The independent variable is viewed as controlled by the experimenter, so it is considered as non-stochastic whereas \( y \) is viewed as a random variable with
\[ E(y) = \beta_0 + \beta_1 X \]
and
\[ Var(y) = \sigma^2. \]
Sometimes \( X \) can also be a random variable. In such a case, instead of simple mean and simple variance of \( y \), we consider the conditional mean of \( y \) given \( X = x \) as
\[ E(y \mid x) = \beta_0 + \beta_1 x \]
and the conditional variance of \( y \) given \( X = x \) as

\[
Var(y \mid x) = \sigma^2.
\]

When the values of \( \beta_0, \beta_1 \) and \( \sigma^2 \) are known, the model is completely described. The parameters \( \beta_0, \beta_1 \) and \( \sigma^2 \) are generally unknown in practice and \( \epsilon \) is unobserved. The determination of the statistical model \( y = \beta_0 + \beta_1 x + \epsilon \) depends on the determination (i.e., estimation) of \( \beta_0, \beta_1 \) and \( \sigma^2 \). In order to know the values of these parameters, \( n \) pairs of observations \((x_i, y_i)(i = 1, \ldots, n)\) on \((X, y)\) are observed/collection and are used to determine these unknown parameters.

Various methods of estimation can be used to determine the estimates of the parameters. Among them, the methods of least squares and maximum likelihood are the popular methods of estimation.

**Least squares estimation**

Suppose a sample of \( n \) sets of paired observations \((x_i, y_i)(i = 1, 2, \ldots, n)\) are available. These observations are assumed to satisfy the simple linear regression model and so we can write

\[
y_i = \beta_0 + \beta_1 x_i + \epsilon_i (i = 1, 2, \ldots, n).
\]

The method of least squares estimates the parameters \( \beta_0 \) and \( \beta_1 \) by minimizing the sum of squares of difference between the observations and the line in the scatter diagram. Such an idea is viewed from different perspectives. When the **vertical difference** between the observations and the line in the scatter diagram is considered and its sum of squares is minimized to obtain the estimates of \( \beta_0 \) and \( \beta_1 \), the method is known as **direct regression**.
Alternatively, the sum of squares of difference between the observations and the line in horizontal direction in the scatter diagram can be minimized to obtain the estimates of $\beta_0$ and $\beta_1$. This is known as reverse (or inverse) regression method.

Instead of horizontal or vertical errors, if the sum of squares of perpendicular distances between the observations and the line in the scatter diagram is minimized to obtain the estimates of $\beta_0$ and $\beta_1$, the method is known as orthogonal regression or major axis regression method.
Instead of minimizing the distance, the area can also be minimized. The **reduced major axis regression method** minimizes the sum of the areas of rectangles defined between the observed data points and the nearest point on the line in the scatter diagram to obtain the estimates of regression coefficients. This is shown in the following figure:

The method of **least absolute deviation regression** considers the sum of the absolute deviation of the observations from the line in the vertical direction in the scatter diagram as in the case of direct regression to obtain the estimates of $\beta_0$ and $\beta_1$.

No assumption is required about the form of probability distribution of $\epsilon_i$ in deriving the least squares estimates. For the purpose of deriving the statistical inferences only, we assume that $\epsilon_i$'s are random variable with $E(\epsilon_i) = 0, Var(\epsilon_i) = \sigma^2$ and $Cov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j (i, j = 1, 2, ..., n)$. This assumption is needed to find the mean, variance and other properties of the least squares estimates. The assumption that $\epsilon_i$'s are normally distributed is utilized while constructing the tests of hypotheses and confidence intervals of the parameters.

Based on these approaches, different estimates of $\beta_0$ and $\beta_1$ are obtained which have different statistical properties. Among them the direct regression approach is more popular. Generally, the direct regression estimates are referred as the **least squares estimates** or **ordinary least squares estimates**.
Direct regression method

This method is also known as the \textit{ordinary least squares estimation}. Assuming that a set of \( n \) paired observations on \((x_i, y_i), i = 1, 2, \ldots, n\) are available which satisfy the linear regression model \( y = \beta_0 + \beta_1 x + \varepsilon \).

So we can write the model for each observation as \( y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \), \((i = 1, 2, \ldots, n)\).

The direct regression approach minimizes the sum of squares

\[
S(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2
\]

with respect to \( \beta_0 \) and \( \beta_1 \).

The partial derivatives of \( S(\beta_0, \beta_1) \) with respect to \( \beta_0 \) is

\[
\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)
\]

and the partial derivative of \( S(\beta_0, \beta_1) \) with respect to \( \beta_1 \) is

\[
\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i.
\]

The solutions of \( \beta_0 \) and \( \beta_1 \) are obtained by setting

\[
\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0
\]

and

\[
\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0.
\]

The solutions of these two equations are called the \textit{direct regression estimators}, or usually called as the \textit{ordinary least squares (OLS)} estimators of \( \beta_0 \) and \( \beta_1 \).

This gives the ordinary least squares estimates \( b_0 \) of \( \beta_0 \) and \( b_1 \) of \( \beta_1 \) as

\[
b_0 = \bar{y} - b_1 \bar{x}
\]

\[
b_1 = \frac{s_{xy}}{s_{xx}}
\]

where

\[
s_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \quad s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.
\]
Further, we have
\[
\frac{\partial^2 S(\beta_0, \beta_1)}{\partial \beta_0^2} = -2 \sum_{i=1}^{n} (-1) = 2n,
\]
\[
\frac{\partial^2 S(\beta_0, \beta_1)}{\partial \beta_1^2} = 2 \sum_{i=1}^{n} x_i^2,
\]
\[
\frac{\partial^2 S(\beta_0, \beta_1)}{\partial \beta_0 \partial \beta_1} = 2 \sum_{i=1}^{n} x_i = 2n \bar{x}.
\]

The Hessian matrix which is the matrix of second order partial derivatives in this case is given as
\[
H^* = \begin{pmatrix}
\frac{\partial^2 S(\beta_0, \beta_1)}{\partial \beta_0^2} & \frac{\partial^2 S(\beta_0, \beta_1)}{\partial \beta_0 \partial \beta_1} \\
\frac{\partial^2 S(\beta_0, \beta_1)}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 S(\beta_0, \beta_1)}{\partial \beta_1^2}
\end{pmatrix}
\]
\[
= 2 \begin{pmatrix}
\sum_{i=1}^{n} x_i^2 & n \bar{x} \\
n \bar{x} & \sum_{i=1}^{n} x_i
\end{pmatrix}
\]
\[
= 2 \left( \ell' \ell, \ell' x \right)
\]

where \( \ell = (1, 1, ..., 1)' \) is a n-vector of elements unity and \( x = (x_1, ..., x_n)' \) is a n-vector of observations on \( X \).

The matrix \( H^* \) is positive definite if its determinant and the element in the first row and column of \( H^* \) are positive. The determinant of \( H \) is given by
\[
|H^*| = 2 \left( n \sum_{i=1}^{n} x_i^2 - n^2 \bar{x}^2 \right)
\]
\[
= 2n \left( \sum_{i=1}^{n} (x_i - \bar{x})^2 \right) \geq 0.
\]

The case when \( \sum_{i=1}^{n} (x_i - \bar{x})^2 = 0 \) is not interesting because all the observations in this case are identical, i.e. \( x_i = c \) (some constant). In such a case there is no relationship between \( x \) and \( y \) in the context of regression analysis. Since \( \sum_{i=1}^{n} (x_i - \bar{x})^2 > 0 \), therefore \( |H^*| > 0 \). So \( H^* \) is positive definite for any \( (\beta_0, \beta_1) \); therefore \( S(\beta_0, \beta_1) \) has a global minimum at \( (b_0, b_1) \).
The **fitted line** or the **fitted linear regression model** is

\[ y = b_0 + b_1 x. \]

The predicted values are

\[ \hat{y}_i = b_0 + b_1 x_i \quad (i = 1, 2, \ldots, n). \]

The difference between the observed value \( y_i \) and the fitted (or predicted) value \( \hat{y}_i \) is called as a **residual**. The \( i^{th} \) residual is defined as

\[ \hat{e}_i = y_i - \hat{y}_i \quad (i = 1, 2, \ldots, n). \]

We consider it as

\[ \hat{e}_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i). \]

**Properties of the direct regression estimators:**

**Unbiased property:**

Note that \( b_1 = \frac{\sum y_i}{\sum x_i^2} \) and \( b_0 = \bar{y} - b_1 \bar{x} \) are the linear combinations of \( y_i (i = 1, \ldots, n) \).

Therefore

\[ b_1 = \sum_{i=1}^{n} k_i y_i \]

where \( k_i = (x_i - \bar{x}) / s_{xx} \). Note that \( \sum k_i = 0 \) and \( \sum k_i x_i = 1 \), so

\[ E(b_1) = \sum_{i=1}^{n} k_i E(y_i) \]

\[ = \sum_{i=1}^{n} k_i (\beta_0 + \beta_1 x_i) \]

\[ = \beta_1. \]

This \( b_1 \) is an unbiased estimator of \( \beta_1 \). Next

\[ E(b_0) = E[\bar{y} - b_1 \bar{x}] \]

\[ = E[\beta_0 + \beta_1 \bar{x} + \bar{\varepsilon} - b_1 \bar{x}] \]

\[ = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} \]

\[ = \beta_0. \]

Thus \( b_0 \) is an unbiased estimator of \( \beta_0 \).
Variances:

Using the assumption that \( y_i \)'s are independently distributed, the variance of \( b_1 \) is

\[
Var(b_1) = \sum_{i=1}^{n} k_i^2 Var(y_i) + \sum_{i\neq j} k_i k_j Cov(y_i, y_j)
\]

\[
= \sigma^2 \frac{\sum (x_i - \bar{x})^2}{s_{xx}^2} \quad (Cov(y_i, y_j) = 0 \text{ as } y_1, ..., y_n \text{ are independent})
\]

\[
= \frac{\sigma^2 s_{xx}}{s_{xx}^2} = \frac{\sigma^2}{s_{xx}}.
\]

The variance of \( b_0 \) is

\[
Var(b_0) = Var(\bar{y}) + \bar{x}^2 Var(b_1) - 2\bar{x}Cov(\bar{y}, b_1).
\]

First we find that

\[
Cov(\bar{y}, b_1) = E\left[ (\bar{y} - E(\bar{y})) (b_1 - E(b_1)) \right]
\]

\[
= E\left[ E(k_i y_i - \beta_i) \right]
\]

\[
= \frac{1}{n} E \left[ \sum_{i} (\beta_0 k_i + \beta_1 \sum_{i} k_i x_i) - \beta_i \sum_{i} \epsilon_i \right]
\]

\[
= \frac{1}{n} [0 + 0 + 0 + 0] = 0
\]

So

\[
Var(b_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} \right).
\]

Covariance:

The covariance between \( b_0 \) and \( b_1 \) is

\[
Cov(b_0, b_1) = Cov(\bar{y}, b_1) - \bar{x}Var(b_1)
\]

\[
= -\frac{\bar{x}}{s_{xx}} \sigma^2.
\]
It can further be shown that the ordinary least squares estimators \( b_0 \) and \( b_1 \) possess the minimum variance in the class of linear and unbiased estimators. So they are termed as the Best Linear Unbiased Estimators (BLUE). Such a property is known as the Gauss-Markov theorem which is discussed later in multiple linear regression model.

**Residual sum of squares:**

The residual sum of squares is given as

\[
SS_{\text{res}} = \sum_{i=1}^{n} e_i^2
\]

\[
= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]

\[
= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2
\]

\[
= \sum_{i=1}^{n} [y_i - \bar{y} + b_1 \bar{x} - b_1 x_i]^2
\]

\[
= \sum_{i=1}^{n} [(y_i - \bar{y}) - b_1 (x_i - \bar{x})]^2
\]

\[
= \sum_{i=1}^{n} (y_i - \bar{y})^2 + b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 - 2b_1 \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})
\]

\[
= s_{yy} + b_1^2 s_{xx} - 2b_1 s_{xy}
\]

\[
= s_{yy} - b_1^2 s_{xx}
\]

\[
= s_{yy} - \left( \frac{s_{yx}}{s_{xx}} \right)^2 s_{xx}
\]

\[
= s_{yy} - \frac{s_{yy}^2}{s_{xx}}
\]

\[
= s_{yy} - b_1 s_{xy}
\]

where \( s_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \), \( \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \).
Estimation of $\sigma^2$

The estimator of $\sigma^2$ is obtained from the residual sum of squares as follows. Assuming that $y_i$ is normally distributed, it follows that $SS_{res}$ has a $\chi^2$ distribution with $(n-2)$ degrees of freedom, so

$$\frac{SS_{res}}{\sigma^2} \sim \chi^2(n-2).$$

Thus using the result about the expectation of a chi-square random variable, we have

$$E(SS_{res}) = (n-2)\sigma^2.$$

Thus an unbiased estimator of $\sigma^2$ is

$$s^2 = \frac{SS_{res}}{n-2}.$$

Note that $SS_{res}$ has only $(n-2)$ degrees of freedom. The two degrees of freedom are lost due to estimation of $b_0$ and $b_1$. Since $s^2$ depends on the estimates $b_0$ and $b_1$, so it is a **model dependent estimate** of $\sigma^2$.

Estimate of variances of $b_0$ and $b_1$:

The estimators of variances of $b_0$ and $b_1$ are obtained by replacing $\sigma^2$ by its estimate $\hat{\sigma}^2 = s^2$ as follows:

$$\hat{\text{Var}}(b_0) = s^2 \left( \frac{1}{n} + \frac{ \bar{x}^2}{s_{xx}} \right)$$

and

$$\hat{\text{Var}}(b_1) = \frac{s^2}{s_{xx}}.$$

It is observed that since $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$, so $\sum_{i=1}^{n} e_i = 0$. In the light of this property, $\hat{\epsilon}_i$ can be regarded as an estimate of unknown $\epsilon_i$ ($i = 1, \ldots, n$). This helps in verifying the different model assumptions on the basis of the given sample $(x_i, y_i), i = 1, 2, \ldots, n$.

Further, note that

(i) $\sum_{i=1}^{n} x_i e_i = 0$,

(ii) $\sum_{i=1}^{n} \hat{y}_i e_i = 0$. 

---

Econometrics | Chapter 2 | Simple Linear Regression Analysis | Shalabh, IIT Kanpur
(iii) $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$ and

(iv) the fitted line always passes through $(\bar{x}, \bar{y})$.

**Centered Model:**

Sometimes it is useful to measure the independent variable around its mean. In such a case, model

$$y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

has a centered version as follows:

$$y_i = \beta_0 + \beta_1 (x_i - \bar{x}) + \beta_2 \bar{x} + \epsilon_i$$  

$(i = 1, 2, ..., n)$

where $\beta_2^* = \beta_0 + \beta_1 \bar{x}$. The sum of squares due to error is given by

$$S(\beta_0^*, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left[ y_i - \beta_0^* - \beta_1 (x_i - \bar{x}) \right]^2.$$

Now solving

$$\frac{\partial S(\beta_0^*, \beta_1)}{\partial \beta_0^*} = 0$$

and

$$\frac{\partial S(\beta_0^*, \beta_1)}{\partial \beta_1} = 0,$$

we get the direct regression least squares estimates of $\beta_0^*$ and $\beta_1$ as

$$\hat{b}_0^* = \bar{y}$$

and

$$\hat{b}_1 = \frac{s_{xy}}{s_{xx}}$$

respectively.

Thus the form of the estimate of slope parameter $\beta_1$ remains same in usual and centered model whereas the form of the estimate of intercept term changes in the usual and centered models.

Further, the Hessian matrix of the second order partial derivatives of $S(\beta_0^*, \beta_1)$ with respect to $\beta_0^*$ and $\beta_1$ is positive definite at $\beta_0^* = \hat{b}_0^*$ and $\beta_1 = \hat{b}_1$ which ensures that $S(\beta_0^*, \beta_1)$ is minimized at $\beta_0^* = \hat{b}_0^*$ and $\beta_1 = \hat{b}_1$.
Under the assumption that \( E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2 \) and \( Cov(\varepsilon_i, \varepsilon_j) = 0 \) for all \( i \neq j = 1, 2, ..., n \), it follows that

\[
E(b^*_0) = \beta_0, \quad E(b_j) = \beta_j, \\
Var(b^*_0) = \frac{\sigma^2}{n}, \quad Var(b_j) = \frac{\sigma^2}{s_{xx}}.
\]

In this case, the fitted model of \( y_i = \beta^*_0 + \beta_j(x_i - \bar{x}) + \varepsilon_i \) is

\[
y = \bar{y} + b_j(x - \bar{x}),
\]

and the predicted values are

\[
\hat{y}_i = \bar{y} + b_j(x_i - \bar{x}) \quad (i = 1, ..., n).
\]

Note that in centered model

\[
Cov(b^*_0, b_j) = 0.
\]

**No intercept term model:**

Sometimes in practice, a model without an intercept term is used in those situations when \( x_i = 0 \Rightarrow y_i = 0 \) for all \( i = 1, 2, ..., n \). A no-intercept model is

\[
y_i = \beta_j x_i + \varepsilon_i \quad (i = 1, 2, ..., n).
\]

For example, in analyzing the relationship between the illumination of bulb \((y)\) and electric current \((X)\), the illumination is zero when current is zero.

Using the data \((x_i, y_i), i = 1, 2, ..., n\), the direct regression least squares estimate of \( \beta_j \) is obtained by minimizing \( S(\beta_j) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_j x_i)^2 \) and solving

\[
\frac{\partial S(\beta_j)}{\partial \beta_j} = 0
\]

gives the estimator of \( \beta_j \) as

\[
\hat{b}_j = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}.
\]

The second order partial derivative of \( S(\beta_j) \) with respect to \( \beta_j \) at \( \beta_j = b_j \) is positive which insures that \( b_j \) minimizes \( S(\beta_j) \).
Using the assumption that \( E(\varepsilon_i) = 0, \text{Var}(\varepsilon_i) = \sigma^2 \) and \( \text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \) for all \( i \neq j = 1, 2, \ldots, n \), the properties of \( b_i^* \) can be derived as follows:

\[
E(b_i^*) = \frac{\sum_{i=1}^{n} x_i E(y_i)}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i^2 \beta_i}{\sum_{i=1}^{n} x_i^2} = \beta_i
\]

This \( b_i^* \) is an unbiased estimator of \( \beta_i \). The variance of \( b_i^* \) is obtained as follows:

\[
\text{Var}(b_i^*) = \frac{\sum_{i=1}^{n} x_i^2 \text{Var}(y_i)}{\left( \sum_{i=1}^{n} x_i^2 \right)^2} = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{\left( \sum_{i=1}^{n} x_i^2 \right)^2} = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}
\]

and an unbiased estimator of \( \sigma^2 \) is obtained as

\[
\frac{\sum_{i=1}^{n} y_i^2 - b_i \sum_{i=1}^{n} y_i x_i}{n-1}.
\]

**Maximum likelihood estimation**
We assume that $e_i$'s ($i = 1, 2, ..., n$) are independent and identically distributed following a normal distribution $N(0, \sigma^2)$. Now we use the method of maximum likelihood to estimate the parameters of the linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad (i = 1, 2, ..., n),$$

the observations $y_i$ ($i = 1, 2, ..., n$) are independently distributed with $N(\beta_0 + \beta_1 x_i, \sigma^2)$ for all $i = 1, 2, ..., n$.

The likelihood function of the given observations $(x_i, y_i)$ and unknown parameters $\beta_0, \beta_1$ and $\sigma^2$ is

$$L(x_i, y_i; \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \left( \frac{1}{2\pi \sigma^2} \right)^{1/2} \exp \left[ -\frac{1}{2\sigma^2} (y_i - \beta_0 - \beta_1 x_i)^2 \right].$$

The maximum likelihood estimates of $\beta_0, \beta_1$ and $\sigma^2$ can be obtained by maximizing $L(x_i, y_i; \beta_0, \beta_1, \sigma^2)$ or equivalently $\ln L(x_i, y_i; \beta_0, \beta_1, \sigma^2)$ where

$$\ln L(x_i, y_i; \beta_0, \beta_1, \sigma^2) = -\left( \frac{n}{2} \right) \ln 2\pi - \left( \frac{n}{2} \right) \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

The normal equations are obtained by partial differentiation of log-likelihood with respect to $\beta_0, \beta_1$ and $\sigma^2$ and equating them to zero as follows:

$$\frac{\partial \ln L(x_i, y_i; \beta_0, \beta_1, \sigma^2)}{\partial \beta_0} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial \ln L(x_i, y_i; \beta_0, \beta_1, \sigma^2)}{\partial \beta_1} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

and

$$\frac{\partial \ln L(x_i, y_i; \beta_0, \beta_1, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^4} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = 0.$$

The solution of these normal equations give the maximum likelihood estimates of $\beta_0, \beta_1$ and $\sigma^2$ as

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x}$$

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{s_{xy}}{s_{xx}}$$

and

$$\tilde{s}^2 = \frac{\sum_{i=1}^{n} (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2}{n}$$

respectively.
It can be verified that the Hessian matrix of second order partial derivation of  \( \ln L \) with respect to \( \beta_0, \beta_1, \) and \( \sigma^2 \) is negative definite at \( \beta_0 = \tilde{\beta}_0, \beta_1 = \tilde{\beta}_1, \) and \( \sigma^2 = \tilde{s}^2 \) which ensures that the likelihood function is maximized at these values.

Note that the least squares and maximum likelihood estimates of \( \beta_0 \) and \( \beta_1 \) are identical. The least squares and maximum likelihood estimates of \( \sigma^2 \) are different. In fact, the least squares estimate of \( \sigma^2 \) is

\[
s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \bar{y})^2
\]

so that it is related to maximum likelihood estimate as

\[
\tilde{s}^2 = \frac{n-2}{n} s^2.
\]

Thus \( \tilde{\beta}_0 \) and \( \tilde{\beta}_1 \) are unbiased estimators of \( \beta_0 \) and \( \beta_1 \) whereas \( \tilde{s}^2 \) is a biased estimate of \( \sigma^2 \), but it is asymptotically unbiased. The variances of \( \tilde{\beta}_0 \) and \( \tilde{\beta}_1 \) are same as of \( \beta_0 \) and \( \beta_1 \) respectively but the mean squared error \( \text{MSE}(\tilde{s}^2) < \text{Var}(s^2) \).

**Testing of hypotheses and confidence interval estimation for slope parameter:**

Now we consider the tests of hypothesis and confidence interval estimation for the slope parameter of the model under two cases, viz., when \( \sigma^2 \) is known and when \( \sigma^2 \) is unknown.

**Case 1: When \( \sigma^2 \) is known:**

Consider the simple linear regression model \( y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (i = 1, 2, \ldots, n) \). It is assumed that \( \epsilon_i 's \) are independent and identically distributed and follow \( N(0, \sigma^2) \).

First we develop a test for the null hypothesis related to the slope parameter

\[
H_0 : \beta_1 = \beta_{10}
\]

where \( \beta_{10} \) is some given constant.
Assuming $\sigma^2$ to be known, we know that $E(b_i) = \beta_1$, $Var(b_i) = \frac{\sigma^2}{s_{xx}}$ and $b_i$ is a linear combination of normally distributed $y_i$’s. So

$$b_i \sim N\left(\beta_1, \frac{\sigma^2}{s_{xx}}\right)$$

and so the following statistic can be constructed

$$Z_i = \frac{b_i - \beta_{10}}{\sqrt{\frac{\sigma^2}{s_{xx}}}}$$

which is distributed as $N(0,1)$ when $H_0$ is true.

A decision rule to test $H_1: \beta_1 \neq \beta_{10}$ can be framed as follows:

Reject $H_0$ if $|Z_i| > z_{\alpha/2}$

where $Z_{\alpha/2}$ is the $\alpha/2$ percent points on normal distribution.

Similarly, the decision rule for one sided alternative hypothesis can also be framed.

The 100 $(1 - \alpha)\%$ confidence interval for $\beta_1$ can be obtained using the $Z_i$ statistic as follows:

$$P\left[-z_{\alpha/2} \leq Z_i \leq z_{\alpha/2}\right] = 1 - \alpha$$

$$P\left[-z_{\alpha/2} \leq \frac{b_i - \beta_1}{\sqrt{\frac{\sigma^2}{s_{xx}}}} \leq z_{\alpha/2}\right] = 1 - \alpha$$

$$P\left[b_i - z_{\alpha/2}\sqrt{\frac{\sigma^2}{s_{xx}}} \leq \beta_1 \leq b_i + z_{\alpha/2}\sqrt{\frac{\sigma^2}{s_{xx}}}\right] = 1 - \alpha.$$
Case 2: When $\sigma^2$ is unknown:

When $\sigma^2$ is unknown then we proceed as follows. We know that

$$\frac{SS_{res}}{\sigma^2} \sim \chi^2(n-2)$$

and

$$E\left( \frac{SS_{res}}{n-2} \right) = \sigma^2.$$

Further, $\frac{SS_{res}}{\sigma^2}$ and $b_i$ are independently distributed. This result will be proved formally later in next module on multiple linear regression. This result also follows from the result that under normal distribution, the maximum likelihood estimates, viz., the sample mean (estimator of population mean) and the sample variance (estimator of population variance) are independently distributed, so $b_i$ and $s^2$ are also independently distributed.

Thus the following statistic can be constructed:

$$t_0 = \frac{b_i - \beta_i}{\frac{s^2}{\sqrt{SS_{res}}} \sqrt{s_{xx}}},$$

which follows a $t$-distribution with $(n-2)$ degrees of freedom, denoted as $t_{n-2}$, when $H_0$ is true.

A decision rule to test $H_i: \beta_i \neq \beta_{i0}$ is to reject $H_0$ if $|t_0| > t_{n-2,\alpha/2}$

where $t_{n-2,\alpha/2}$ is the $\alpha / 2$ percent point of the $t$-distribution with $(n-2)$ degrees of freedom. Similarly, the decision rule for one sided alternative hypothesis can also be framed.

The $100(1-\alpha)$% confidence interval of $\beta_i$ can be obtained using the $t_0$ statistic as follows:
Consider
\[
P\left[-t_{a/2} \leq t_0 \leq t_{a/2}\right] = 1 - \alpha
\]
\[
P\left[-t_{a/2} \leq \frac{b_1 - \beta_1}{\sqrt{\sigma^2/s_{xx}}} \leq t_{a/2}\right] = 1 - \alpha
\]
\[
P\left[b_1 - t_{a/2}\sqrt{\frac{\sigma^2}{s_{xx}}} \leq \beta_1 \leq b_1 + t\alpha / 2\sqrt{\frac{\sigma^2}{s_{xx}}}\right] = 1 - \alpha.
\]
So the 100 (1 − α)% confidence interval β is
\[
\left[b_1 - t_{n-2,a/2}\sqrt{\frac{SS_{res}}{(n-2)s_{xx}}}, b_1 + t_{n-2,a/2}\sqrt{\frac{SS_{res}}{(n-2)s_{xx}}}\right].
\]

**Testing of hypotheses and confidence interval estimation for intercept term:**

Now, we consider the tests of hypothesis and confidence interval estimation for intercept term under two cases, viz., when \(\sigma^2\) is known and when \(\sigma^2\) is unknown.

**Case 1: When \(\sigma^2\) is known:**

Suppose the null hypothesis under consideration is
\[
H_0 : \beta_0 = \beta_{00},
\]
where \(\sigma^2\) is known, then using the result that \(E(b_0) = \beta_0, \text{Var}(b_0) = \sigma^2 \left(\frac{1}{n + \bar{x}^2/s_x}\right)\) and \(b_0\) is a linear combination of normally distributed random variables, the following statistic
\[
Z_0 = \frac{b_0 - \beta_{00}}{\sqrt{\sigma^2 \left(\frac{1}{n + \bar{x}^2/s_x}\right)}}
\]
has a \(N(0,1)\) distribution when \(H_0\) is true.

A decision rule to test \(H_1 : \beta_0 \neq \beta_{00}\) can be framed as follows:

Reject \(H_0\) if \(|Z_0| > Z_{a/2}\)

where \(Z_{a/2}\) is the \(\alpha/2\) percentage points on normal distribution. Similarly, the decision rule for one sided alternative hypothesis can also be framed.
The 100(1−\(\alpha\))% confidence intervals for \(\beta_0\) when \(\sigma^2\) is known can be derived using the \(Z_0\) statistic as follows:

\[
P\left[-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2}\right] = 1 - \alpha
\]

\[
P\left[-z_{\alpha/2} \leq \frac{b_0 - \beta_0}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)}} \leq z_{\alpha/2}\right] = 1 - \alpha
\]

\[
P\left[b_0 - z_{\alpha/2} \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)} \leq \beta_0 \leq b_0 + z_{\alpha/2} \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)}\right] = 1 - \alpha.
\]

So the 100(1−\(\alpha\))% of confidential interval of \(\beta_0\) is

\[
\left[b_0 - z_{\alpha/2} \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)}, b_0 + z_{\alpha/2} \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)}\right].
\]

**Case 2: When \(\sigma^2\) is unknown:**

When \(\sigma^2\) is unknown, then the following statistic is constructed

\[
t_0 = \frac{b_0 - \beta_{00}}{SS_{res} \left(\frac{1}{n} + \frac{\bar{x}^2}{n-2, s_{xx}}\right)}
\]

which follows a \(t\)-distribution with \((n-2)\) degrees of freedom, i.e., \(t_{n-2}\) when \(H_0\) is true.

A decision rule to test \(H_I: \beta_0 \neq \beta_{00}\) is as follows:

Reject \(H_0\) whenever \(|t_0| > t_{n-2, \alpha/2}\)

where \(t_{n-2, \alpha/2}\) is the \(\alpha/2\) percentage point of the \(t\)-distribution with \((n-2)\) degrees of freedom. Similarly, the decision rule for one sided alternative hypothesis can also be framed.

The 100(1−\(\alpha\))% confidence interval of \(\beta_0\) can be obtained as follows:
Consider

\[
P \left[ t_{n-2,\alpha/2} \leq t_0 \leq t_{n-2,\alpha/2} \right] = 1 - \alpha
\]

\[
P \left[ t_{n-2,\alpha/2} \leq \frac{b_0 - \beta_0}{\sqrt{\frac{SS_{res}}{n-2} \left( \frac{1}{n} + \frac{\bar{x}^2}{s^2} \right)}} \leq t_{n-2,\alpha/2} \right] = 1 - \alpha
\]

\[
P \left[ b_0 - t_{n-2,\alpha/2} \sqrt{\frac{SS_{res}}{n-2} \left( \frac{1}{n} + \frac{\bar{x}^2}{s^2} \right)} \leq \beta_0 \leq b_0 + t_{n-2,\alpha/2} \sqrt{\frac{SS_{res}}{n-2} \left( \frac{1}{n} + \frac{\bar{x}^2}{s^2} \right)} \right] = 1 - \alpha.
\]

So 100(1 - \alpha)% confidence interval for \( \beta_0 \) is

\[
\left[ b_0 - t_{n-2,\alpha/2} \sqrt{\frac{SS_{res}}{n-2} \left( \frac{1}{n} + \frac{\bar{x}^2}{s^2} \right)}, b_0 + t_{n-2,\alpha/2} \sqrt{\frac{SS_{res}}{n-2} \left( \frac{1}{n} + \frac{\bar{x}^2}{s^2} \right)} \right].
\]

**Test of hypothesis for \( \sigma^2 \)**

We have considered two types of test statistics for testing the hypothesis about the intercept term and slope parameter—when \( \sigma^2 \) is known and when \( \sigma^2 \) is unknown. While dealing with the case of known \( \sigma^2 \), the value of \( \sigma^2 \) is known from some external sources like past experience, long association of the experimenter with the experiment, past studies etc. In such situations, the experimenter would like to test the hypothesis like \( H_0 : \sigma^2 = \sigma^2_0 \) against \( H_1 : \sigma^2 \neq \sigma^2_0 \) where \( \sigma^2_0 \) is specified. The test statistic is based on the result

\[
\frac{SS_{res}}{\sigma^2} \sim \chi^2_{n-2}.
\]

So the test statistic is

\[
C_0 = \frac{SS_{res}}{\sigma^2_0} \sim \chi^2_{n-2} \text{ under } H_0.
\]

The decision rule is to reject \( H_0 \) if \( C_0 < \chi^2_{n-2,\alpha/2} \) or \( C_0 > \chi^2_{n-2,1-\alpha/2} \).
Confidence interval for $\sigma^2$

A confidence interval for $\sigma^2$ can also be derived as follows. Since $SS_{\text{res}} / \sigma^2 \sim \chi^2_{n-2}$, thus consider

$$P \left[ \frac{SS_{\text{res}}}{\sigma^2} \leq \frac{SS_{\text{res}}}{\chi^2_{n-2,\alpha/2}} \leq \frac{SS_{\text{res}}}{\chi^2_{n-2,1-\alpha/2}} \right] = 1 - \alpha$$

$$P \left[ \frac{SS_{\text{res}}}{\chi^2_{n-2,1-\alpha/2}} \leq \sigma^2 \leq \frac{SS_{\text{res}}}{\chi^2_{n-2,\alpha/2}} \right] = 1 - \alpha .$$

The corresponding 100(1 − $\alpha$)% confidence interval for $\sigma^2$ is

$$\left[ \frac{SS_{\text{res}}}{\chi^2_{n-2,1-\alpha/2}}, \frac{SS_{\text{res}}}{\chi^2_{n-2,\alpha/2}} \right] .$$

Joint confidence region for $\beta_0$ and $\beta_1$:

A joint confidence region for $\beta_0$ and $\beta_1$ can also be found. Such region will provide a 100(1 − $\alpha$)% confidence that both the estimates of $\beta_0$ and $\beta_1$ are correct. Consider the centered version of the linear regression model

$$y_i = \beta_0^* + \beta_1^*(x_i - \bar{x}) + \epsilon_i$$

where $\beta_0^* = \beta_0 + \beta_1 \bar{x}$. The least squares estimators of $\beta_0^*$ and $\beta_1^*$ are

$$b_0^* = \bar{y} \quad \text{and} \quad b_1 = \frac{s_{xy}}{s_{xx}},$$

respectively.

Using the results that

$$E(b_0^*) = \beta_0^*,$$
$$E(b_1) = \beta_1,$$
$$\text{Var}(b_0^*) = \frac{\sigma^2}{n},$$
$$\text{Var}(b_1) = \frac{\sigma^2}{s_{xx}} .$$

When $\sigma^2$ is known, then the statistic

$$\frac{b_0^* - \beta_0^*}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1) \quad \text{and} \quad \frac{b_1 - \beta_1}{\sqrt{\frac{\sigma^2}{s_{xx}}}} \sim N(0,1).$$
Moreover, both the statistics are independently distributed. Thus
\[
\left( \frac{b_0^* - \beta_0}{\sigma^2 / n} \right)^2 \sim \chi_i^2 \quad \text{and} \quad \left( \frac{b_1 - \beta_1}{s_{xx}} \right)^2 \sim \chi_i^2
\]
are also independently distributed because \( b_0^* \) and \( b_1 \) are independently distributed. Consequently, the sum of these two
\[
\frac{n(b_0^* - \beta_0)^2}{\sigma^2} + \frac{s_{xx}(b_1 - \beta_1)^2}{\sigma^2} \sim \chi_2^2.
\]
Since
\[
\frac{SS_{\text{reg}}}{\sigma^2} \sim \chi_{n-2}^2
\]
and \( SS_{\text{reg}} \) is independently distributed of \( b_0^* \) and \( b_1 \), so the ratio
\[
\left( \frac{n(b_0^* - \beta_0)^2}{\sigma^2} + \frac{s_{xx}(b_1 - \beta_1)^2}{\sigma^2} \right) / 2 \sim F_{2,n-2}.
\]
Substituting \( b_0^* = b_0 + b_1 \bar{x} \) and \( \beta_0^* = \beta_0 + \beta_1 \bar{x} \), we get
\[
\left( \frac{n-2}{2} \right) \left[ \frac{Q_f}{SS_{\text{reg}}} \right]
\]
where
\[
Q_f = n(b_0 - \beta_0)^2 + 2 \sum_{i=1}^{n} x_i (b_0 - \beta_0)(b_1 - \beta_1) + 2 \sum_{i=1}^{n} x_i^2 (b_1 - \beta_1)^2.
\]
Since
\[
P \left( \frac{n-2}{2} \frac{Q_f}{SS_{\text{reg}}} \leq F_{2,n-2} \right) = 1 - \alpha
\]
holds true for all values of \( \beta_0 \) and \( \beta_1 \), so the 100(1 - \( \alpha \))% confidence region for \( \beta_0 \) and \( \beta_1 \) is
\[
\left( \frac{n-2}{2} \right) \frac{Q_f}{SS_{\text{reg}}} \leq F_{2,n-2,1-\alpha}.
\]
This confidence region is an ellipse which gives the 100(1 - \( \alpha \))% probability that \( \beta_0 \) and \( \beta_1 \) are contained simultaneously in this ellipse.

*Econometrics*  |  Chapter 2  |  Simple Linear Regression Analysis  |  *Shalabh, IIT Kanpur*
Analysis of variance:
The technique of analysis of variance is usually used for testing the hypothesis related to equality of more than one parameters, like population means or slope parameters. It is more meaningful in case of multiple regression model when there are more than one slope parameters. This technique is discussed and illustrated here to understand the related basic concepts and fundamentals which will be used in developing the analysis of variance in the next module in multiple linear regression model where the explanatory variables are more than two.

A test statistic for testing $H_0: \beta_i = 0$ can also be formulated using the analysis of variance technique as follows.

On the basis of the identity
\[ y_i - \hat{y}_i = (y_i - \bar{y}) - (\hat{y}_i - \bar{y}), \]
the sum of squared residuals is
\[ S(b) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]
\[ = \sum_{i=1}^{n} (y_i - \bar{y})^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 - 2\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{y}). \]

Further consider
\[ \sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n} (y_i - \bar{y})b_i(x_i - \bar{x}) \]
\[ = b_i^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \]
\[ = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2. \]

Thus we have
\[ \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2. \]

The term $\sum_{i=1}^{n} (y_i - \bar{y})^2$ is called the sum of squares about the mean, corrected sum of squares of $y$ (i.e., $SS_{\text{corrected}}$), total sum of squares, or $s_{yy}$.\[ \]
The term \( \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \) describes the deviation: observation minus predicted value, viz., the residual sum of squares, i.e., \( SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \)

whereas the term \( \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \) describes the proportion of variability explained by regression, \( SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \).

If all observations \( y_i \) are located on a straight line, then in this case \( \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 0 \) and thus \( SS_{corrected} = SS_{reg} \).

Note that \( SS_{reg} \) is completely determined by \( b_1 \) and so has only one degrees of freedom. The total sum of squares \( s_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \) has \( (n-1) \) degrees of freedom due to constraint \( \sum_{i=1}^{n} (y_i - \bar{y}) = 0 \) and \( SS_{res} \) has \( (n-2) \) degrees of freedom as it depends on the determination of \( b_0 \) and \( b_1 \).

All sums of squares are mutually independent and distributed as \( \chi^2 \) with \( df \) degrees of freedom if the errors are normally distributed.

The mean square due to regression is
\[
MS_{reg} = \frac{SS_{reg}}{1}
\]
and mean square due to residuals is
\[
MSE = \frac{SS_{res}}{n-2}.
\]

The test statistic for testing \( H_0 : \beta_1 = 0 \) is
\[
F_0 = \frac{MS_{reg}}{MSE}.
\]

If \( H_0 : \beta_1 = 0 \) is true, then \( MS_{reg} \) and \( MSE \) are independently distributed and thus \( F_0 \sim F_{1,n-2} \).
The decision rule for \( \text{H}_i : \beta_i \neq 0 \) is to reject \( \text{H}_0 \) if

\[
F_0 > F_{1, n-2; 1-\alpha}
\]

at \( \alpha \) level of significance. The test procedure can be described in an Analysis of variance table.

### Analysis of variance for testing \( \text{H}_0 : \beta_i = 0 \)

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean square</th>
<th>( F )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>( SS_{reg} )</td>
<td>1</td>
<td>( MS_{reg} )</td>
<td>( MS_{reg} / MSE )</td>
</tr>
<tr>
<td>Residual</td>
<td>( SS_{res} )</td>
<td>( n-2 )</td>
<td>( MSE )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( s_{yy} )</td>
<td>( n-1 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Some other forms of \( SS_{reg}, SS_{res} \) and \( s_{yy} \) can be derived as follows:

The sample correlation coefficient then may be written as

\[
r_{xy} = \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}}.\]

Moreover, we have

\[
b_i = \frac{s_{xy}}{s_{xx}} = r_{xy} \frac{s_{yy}}{s_{xx}}.\]

The estimator of \( \sigma^2 \) in this case may be expressed as

\[
s^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2
= \frac{1}{n-2} SS_{res}.
\]

Various alternative formulations for \( SS_{res} \) are in use as well:

\[
SS_{res} = \sum_{i=1}^{n} [y_i - (b_0 + b_i x_i)]^2
= \sum_{i=1}^{n} [(y_i - \bar{y}) - b_i (x_i - \bar{x})]^2
= s_{yy} + b_i^2 s_{xx} - 2b_i s_{xy}
= s_{yy} - b_i^2 s_{xx}
= s_{yy} \left(1 - \frac{s_{xy}^2}{s_{xx}}\right).
\]
Using this result, we find that
\[ SS_{corrected} = s_{yy} \]
and
\[ SS_{reg} = s_{yy} - SS_{res} = \left( \frac{s_{xy}}{s_{xx}} \right)^2 s_{xx} = b^2_s s_{xx} = b_s^2 s_{xy}. \]

**Goodness of fit of regression**

It can be noted that a fitted model can be said to be good when residuals are small. Since \( SS_{res} \) is based on residuals, so a measure of quality of fitted model can be based on \( SS_{res} \). When intercept term is present in the model, a measure of goodness of fit of the model is given by
\[ R^2 = 1 - \frac{SS_{res}}{s_{yy}} = \frac{SS_{reg}}{s_{yy}}. \]

This is known as the **coefficient of determination**. This measure is based on the concept that how much variation in \( y \)'s stated by \( s_{yy} \) is explainable by \( SS_{reg} \) and how much unexplainable part is contained in \( SS_{res} \). The ratio \( SS_{reg} / s_{yy} \) describes the proportion of variability that is explained by regression in relation to the total variability of \( y \). The ratio \( SS_{res} / s_{yy} \) describes the proportion of variability that is not covered by the regression.

It can be seen that
\[ R^2 = r_{xy}^2 \]
where \( r_{xy} \) is the simple correlation coefficient between \( x \) and \( y \). Clearly \( 0 \leq R^2 \leq 1 \), so a value of \( R^2 \) closer to one indicates the better fit and value of \( R^2 \) closer to zero indicates the poor fit.
Prediction of values of study variable

An important use of linear regression modeling is to predict the average and actual values of study variable. The term prediction of value of study variable corresponds to knowing the value of $E(y)$ (in case of average value) and value of $y$ (in case of actual value) for a given value of explanatory variable. We consider both the cases.

Case 1: Prediction of average value

Under the linear regression model $y = \beta_0 + \beta_1 x + \epsilon$, the fitted model is $y = b_0 + b_1 x$ where $b_0$ and $b_1$ are the OLS estimators of $\beta_0$ and $\beta_1$ respectively.

Suppose we want to predict the value of $E(y)$ for a given value of $x = x_0$. Then the predictor is given by

$$\hat{E}(y | x_0) = \hat{\mu}_{y|x_0} = b_0 + b_1 x_0.$$

Predictive bias

Then the prediction error is given as

$$\hat{\mu}_{y|x_0} - E(y) = b_0 + b_1 x_0 - E(\beta_0 + \beta_1 x_0 + \epsilon) = b_0 + b_1 x_0 - (\beta_0 + \beta_1 x_0) = (b_0 - \beta_0) + (b_1 - \beta_1) x_0.$$

Then

$$E(\hat{\mu}_{y|x_0} - E(y)) = E(b_0 - \beta_0) + E(b_1 - \beta_1) x_0 = 0 + 0 = 0$$

Thus the predictor $\mu_{y|x_0}$ is an unbiased predictor of $E(y)$.

Predictive variance:

The predictive variance of $\hat{\mu}_{y|x_0}$ is

$$PV(\hat{\mu}_{y|x_0}) = Var(b_0 + b_1 x_0)
= Var[\overline{y} + b_1 (x_0 - \overline{x})]
= Var(\overline{y}) + (x_0 - \overline{x})^2 Var(b_1) + 2(x_0 - \overline{x}) Cov(\overline{y}, b_1)
= \sigma^2/n + \sigma^2 (x_0 - \overline{x})^2/s_{xx} + 0
= \sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{s_{xx}} \right].$$
Estimate of predictive variance

The predictive variance can be estimated by substituting $\sigma^2$ by $\hat{\sigma}^2 = MSE$ as

$$
\hat{PV}(\hat{y}_{x_0}) = \hat{\sigma}^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right] = MSE \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right].
$$

**Prediction interval estimation:**

The 100(1-\(\alpha\))% prediction interval for $E(y/ x_0)$ is obtained as follows:

The predictor $\hat{y}_{x_0}$ is a linear combination of normally distributed random variables, so it is also normally distributed as

$$
\hat{y}_{x_0} \sim N(\beta_0 + \beta_1 x_0, PV(\hat{y}_{x_0})).
$$

So if $\sigma^2$ is known, then the distribution of

$$
\frac{\hat{y}_{x_0} - E(y \mid x_0)}{\sqrt{PV(\hat{y}_{x_0})}}
$$

is $N(0,1)$. So the 100(1-\(\alpha\))% prediction interval is obtained as

$$
P \left[ -z_{\alpha/2} \leq \frac{\hat{y}_{x_0} - E(y \mid x_0)}{\sqrt{PV(\hat{y}_{x_0})}} \leq z_{\alpha/2} \right] = 1 - \alpha
$$

which gives the prediction interval for $E(y/ x_0)$ as

$$
[\hat{y}_{x_0} - z_{\alpha/2} \sqrt{\sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right]}, \hat{y}_{x_0} + z_{\alpha/2} \sqrt{\sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right]}].
$$

When $\sigma^2$ is unknown, it is replaced by $\hat{\sigma}^2 = MSE$ and in this case the sampling distribution of

$$
\frac{\hat{y}_{x_0} - E(y \mid x_0)}{\sqrt{MSE \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right]}}
$$

is $t$-distribution with $(n-2)$ degrees of freedom, i.e., $t_{n-2}$.
The 100(1 - \( \alpha \))% prediction interval in this case is
\[
P \left[ -t_{a/2,n-2} \leq \frac{\hat{\mu}_{\hat{y}|x_0} - E(y|\ x_0)}{\sqrt{\text{MSE} \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right)}} \leq t_{a/2,n-2} \right] = 1 - \alpha.
\]
which gives the prediction interval as
\[
\left[ \hat{\mu}_{\hat{y}|x_0} - t_{a/2,n-2} \sqrt{\text{MSE} \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right)}, \hat{\mu}_{\hat{y}|x_0} + t_{a/2,n-2} \sqrt{\text{MSE} \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right)} \right].
\]
Note that the width of prediction interval \( E(y|\ x_0) \) is a function of \( x_0 \). The interval width is minimum for \( x_0 = \bar{x} \) and widens as \( |x_0 - \bar{x}| \) increases. This is expected also as the best estimates of \( y \) to be made at \( x \)-values lie near the center of the data and the precision of estimation to deteriorate as we move to the boundary of the \( x \)-space.

**Case 2: Prediction of actual value**

If \( x_0 \) is the value of the explanatory variable, then the actual value predictor for \( y \) is
\[
\hat{y}_0 = b_0 + b_1 x_0.
\]
The true value of \( y \) in the prediction period is given by \( y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0 \) where \( \varepsilon_0 \) indicates the value that would be drawn from the distribution of random error in the prediction period. Note that the form of predictor is the same as of average value predictor but its predictive error and other properties are different. This is the dual nature of predictor.

**Predictive bias:**

The predictive error of \( \hat{y}_0 \) is given by
\[
\hat{y}_0 - y_0 = b_0 + b_1 x_0 - (\beta_0 + \beta_1 x_0 + \varepsilon_0) = (b_0 - \beta_0) + (b_1 - \beta_1) x_0 - \varepsilon.
\]
Thus, we find that
\[
E(\hat{y}_0 - y_0) = E(b_0 - \beta_0) + E(b_1 - \beta_1) x_0 - E(\varepsilon_0) = 0 + 0 + 0 = 0
\]
which implies that \( \hat{y}_0 \) is an unbiased predictor of \( y_0 \).
Predictive variance

Because the future observation $y_0$ is independent of $\hat{y}_0$, the predictive variance of $\hat{y}_0$ is

$$PV(\hat{y}_0) = E(\hat{y}_0 - y_0)^2$$

$$= E[(b_0 - \beta_0) + (x_0 - \bar{x})(b_1 - \beta_1) + (b_1 - \beta_1)\bar{x} - \epsilon_0]^2$$

$$= Var(b_0) + (x_0 - \bar{x})^2Var(b_1) + \bar{x}^2Var(b_1) + Var(\epsilon_0) + 2(x_0 - \bar{x})Cov(b_0, b_1) + 2\bar{x}Cov(b_0, b_1) + 2(x_0 - \bar{x})Var(b_1)$$

[rest of the terms are 0 assuming the independence of $\epsilon_0$ with $\epsilon_1, \epsilon_2, ..., \epsilon_n$]

$$= Var(b_0) + [(x_0 - \bar{x})^2 + \bar{x}^2 + 2(x_0 - \bar{x})]Var(b_1) + Var(\epsilon_0) + 2((x_0 - \bar{x}) + 2\bar{x})Cov(b_0, b_1)$$

$$= Var(b_0) + x_0^2Var(b_1) + Var(\epsilon_0) + 2x_0Cov(b_0, b_1)$$

$$= \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} \right] + x_0^2 \frac{\sigma^2}{s_{xx}} + \sigma^2 - 2x_0 \frac{\bar{x}\sigma^2}{s_{xx}}$$

$$= \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right].$$

Estimate of predictive variance

The estimate of predictive variance can be obtained by replacing $\sigma^2$ by its estimate $\hat{\sigma}^2 = MSE$ as

$$\overline{PV(\hat{y}_0)} = \hat{\sigma}^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right]$$

$$= MSE \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right].$$

Prediction interval:

If $\sigma^2$ is known, then the distribution of

$$\frac{\hat{y}_0 - y_0}{\sqrt{PV(\hat{y}_0)}}$$

is $N(0,1)$. So the 100$(1-\alpha)$% prediction interval is obtained as

$$P\left[ -z_{\alpha/2} \leq \frac{\hat{y}_0 - y_0}{\sqrt{PV(\hat{y}_0)}} \leq z_{\alpha/2} \right] = 1 - \alpha$$

which gives the prediction interval for $y_0$ as

$$\left[ \hat{y}_0 - z_{\alpha/2} \sqrt{\sigma^2 \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right)}, \hat{y}_0 + z_{\alpha/2} \sqrt{\sigma^2 \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right)} \right].$$
When $\sigma^2$ is unknown, then

$$\frac{\hat{y}_0 - y_0}{\sqrt{PV(\hat{y}_0)}}$$

follows a $t$-distribution with $(n - 2)$ degrees of freedom. The 100$(1- \alpha)\%$ prediction interval for $\hat{y}_0$ in this case is obtained as

$$P \left[ -t_{a/2, n-2} \leq \frac{\hat{y}_0 - y_0}{\sqrt{PV(\hat{y}_0)}} \leq t_{a/2, n-2} \right] = 1 - \alpha$$

which gives the prediction interval

$$\left[ \hat{y}_0 - t_{a/2, n-2} \sqrt{MSE \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right)}, \hat{y}_0 + t_{a/2, n-2} \sqrt{MSE \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right)} \right].$$

The prediction interval is of minimum width at $x_0 = \bar{x}$ and widens as $|x_0 - \bar{x}|$ increases.

The prediction interval for $\hat{y}_0$ is wider than the prediction interval for $\hat{\mu}_{y|x_0}$ because the prediction interval for $\hat{y}_0$ depends on both the error from the fitted model as well as the error associated with the future observations.

**Reverse regression method**

The reverse (or inverse) regression approach minimizes the sum of squares of horizontal distances between the observed data points and the line in the following scatter diagram to obtain the estimates of regression parameters.
The reverse regression has been advocated in the analysis of sex (or race) discrimination in salaries. For example, if \( y \) denotes salary and \( x \) denotes qualifications and we are interested in determining if there is a sex discrimination in salaries, we can ask:

“Whether men and women with the same qualifications (value of \( x \)) are getting the same salaries (value of \( y \)). This question is answered by the direct regression.”

Alternatively, we can ask:

“Whether men and women with the same salaries (value of \( y \)) have the same qualifications (value of \( x \)). This question is answered by the reverse regression, i.e., regression of \( x \) on \( y \).”

The regression equation in case of reverse regression can be written as

\[
x_i = \beta_0^* + \beta_1^* y_i + \delta_i \quad (i = 1, 2, \ldots, n)
\]

where \( \delta_i \)'s are the associated random error components and satisfy the assumptions as in the case of usual simple linear regression model. The reverse regression estimates \( \hat{\beta}_{or} \) of \( \beta_0^* \) and \( \hat{\beta}_{ir} \) of \( \beta_1^* \) for the model are obtained by interchanging the \( x \) and \( y \) in the direct regression estimators of \( \beta_0 \) and \( \beta_1 \). The estimates are obtained as

\[
\hat{\beta}_{or} = \bar{x} - \hat{\beta}_{ir} \bar{y}
\]

and

\[
\hat{\beta}_{ir} = \frac{s_{yx}}{s_{yy}}
\]

for \( \beta_0 \) and \( \beta_1 \) respectively. The residual sum of squares in this case is

\[
SS_{or}^* = s_{xx} - \frac{s_{xy}^2}{s_{yy}}.
\]

Note that

\[
\hat{\beta}_{ir} b_1 = \frac{s_{xy}^2}{s_{xx}s_{yy}} = r_{xy}^2
\]

where \( b_1 \) is the direct regression estimator of slope parameter and \( r_{xy} \) is the correlation coefficient between \( x \) and \( y \). Hence if \( r_{xy}^2 \) is close to 1, the two regression lines will be close to each other.

An important application of reverse regression method is in solving the calibration problem.
Orthogonal regression method (or major axis regression method)

The direct and reverse regression methods of estimation assume that the errors in the observations are either in $x$-direction or $y$-direction. In other words, the errors can be either in dependent variable or independent variable. There can be situations when uncertainties are involved in dependent and independent variables both. In such situations, the orthogonal regression is more appropriate. In order to take care of errors in both the directions, the least squares principle in orthogonal regression minimizes the squared perpendicular distance between the observed data points and the line in the following scatter diagram to obtain the estimates of regression coefficients. This is also known as major axis regression method. The estimates obtained are called as orthogonal regression estimates or major axis regression estimates of regression coefficients.

If we assume that the regression line to be fitted is $Y = \beta_0 + \beta_1 X$, then it is expected that all the observations $(x_i, y_i), i = 1, 2, \ldots, n$ lie on this line. But these points deviate from the line and in such a case, the squared perpendicular distance of observed data $(x_i, y_i)$ $(i = 1, 2, \ldots, n)$ from the line is given by

$$d_i^2 = (X_i - x_i)^2 + (Y_i - y_i)^2$$

where $(X_i, Y_i)$ denotes the $i^{th}$ pair of observation without any error which lie on the line.
The objective is to minimize the sum of squared perpendicular distances given by \( \sum_{i=1}^{n} d_i^2 \) to obtain the estimates of \( \beta_0 \) and \( \beta_1 \). The observations \((x_i, y_i) \ (i = 1, 2, \ldots, n)\) are expected to lie on the line 
\[ Y_i = \beta_0 + \beta_1 X_i, \]
so let
\[ E_i = Y_i - \beta_0 - \beta_1 X_i = 0. \]

The regression coefficients are obtained by minimizing \( \sum_{i=1}^{n} d_i^2 \) under the constraints \( E_i \)'s using the Lagrangian’s multiplier method. The Lagrangian function is
\[
L_0 = \sum_{i=1}^{n} d_i^2 - 2 \sum_{i=1}^{n} \lambda_i E_i
\]
where \( \lambda_1, \ldots, \lambda_n \) are the Lagrangian multipliers. The set of equations are obtained by setting
\[
\frac{\partial L_0}{\partial X_i} = 0, \quad \frac{\partial L_0}{\partial Y_i} = 0, \quad \frac{\partial L_0}{\partial \beta_0} = 0 \quad \text{and} \quad \frac{\partial L_0}{\partial \beta_1} = 0 \ (i = 1, 2, \ldots, n).
\]
Thus we find
\[
\frac{\partial L_0}{\partial X_i} = (X_i - x_i) + \lambda_i \beta_1 = 0
\]
\[
\frac{\partial L_0}{\partial Y_i} = (Y_i - y_i) - \lambda_i = 0
\]
\[
\frac{\partial L_0}{\partial \beta_0} = \sum_{i=1}^{n} \lambda_i = 0
\]
\[
\frac{\partial L_0}{\partial \beta_1} = \sum_{i=1}^{n} \lambda_i X_i = 0.
\]
Since
\[
X_i = x_i - \lambda_i \beta_1
\]
\[
Y_i = y_i + \lambda_i,
\]
so substituting these values in \( E_i \), we obtain
\[
E_i = (y_i + \lambda_i) - \beta_0 - \beta_1 (x_i - \lambda_i \beta_1) = 0
\]
\[
\Rightarrow \lambda_i = \frac{\beta_0 + \beta_1 x_i - y_i}{1 + \beta_1^2}.
\]
Also using this \( \lambda_i \) in the equation \( \sum_{i=1}^{n} \lambda_i = 0 \), we get

---

**Econometrics** | **Chapter 2** | **Simple Linear Regression Analysis** | **Shalabh, IIT Kanpur**
\[ \sum_{i=1}^{n} \frac{(\beta_0 + \beta_1 x_i - y_i)}{1 + \beta_1^2} = 0 \]

and using \( (X_i - x) + \lambda, \beta_i = 0 \) and \( \sum \lambda, X_i = 0 \), we get

\[ \sum_{i=1}^{n} \lambda_i (x_i - \lambda, \beta_i) = 0. \]

Substituting \( \lambda_i \) in this equation, we get

\[ \sum_{i=1}^{n} (\beta_0 x_i + \beta_1 x_i^2 - y_i) \]
\[ \frac{\beta_i \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i)^2}{(1 + \beta_i^2)^2} = 0. \] (1)

Using \( \lambda_i \) in the equation and using the equation \( \sum \lambda_i = 0 \), we solve

\[ \sum_{i=1}^{n} \frac{(\beta_0 + \beta_1 x_i - y_i)}{1 + \beta_i^2} = 0. \]

The solution provides an orthogonal regression estimate of \( \beta_0 \) as

\[ \hat{\beta}_{0\text{OR}} = \bar{y} - \hat{\beta}_{1\text{OR}} \bar{x} \]

where \( \hat{\beta}_{1\text{OR}} \) is an orthogonal regression estimate of \( \beta_1 \).

Now, substituting \( \beta_{0\text{OR}} \) in equation (1), we get

\[ \sum_{i=1}^{n} (1 + \beta_i^2) \left[ \bar{x} y_i - \beta_i \bar{x} y_i + \beta_i x_i^2 - x_i y_i \right] - \beta_i \sum_{i=1}^{n} \left( \bar{y} - \bar{x} \beta_i + x_i - y_i \right)^2 = 0 \]

or

\[ (1 + \beta_i^2) \sum_{i=1}^{n} x_i \left[ y_i - \bar{y} - \beta_i (x_i - \bar{x}) \right] + \beta_i \sum_{i=1}^{n} \left[ -(y_i - \bar{y}) + \beta_i (x_i - \bar{x}) \right]^2 = 0 \]

or

\[ (1 + \beta_i^2) \sum_{i=1}^{n} (u_i + \bar{x}) (v_i - \beta_i u_i) + \beta_i \sum_{i=1}^{n} (v_i - \beta_i u_i)^2 = 0 \]

where \( u_i = x_i - \bar{x} \),
\[ v_i = y_i - \bar{y}. \]
Since $\sum_{i=1}^{n} u_i = \sum_{i=1}^{n} v_i = 0$, so

$$\sum_{i=1}^{n} \left[ \beta_1^2 u_i v_i + \beta_1 (u_i^2 - v_i^2) - u_i v_i \right] = 0$$

or

$$\beta_1^2 s_{xy} + \beta_1 (s_{xx} - s_{yy}) - s_{xy} = 0.$$ Solving this quadratic equation provides the orthogonal regression estimate of $\beta_1$ as

$$\hat{\beta}_{1\text{OR}} = \frac{(s_{xy} - s_{xx}) + \text{sign}(s_{xy}) \sqrt{(s_{xx} - s_{yy})^2 + 4s_{xy}^2}}{2s_{xy}}$$

where $\text{sign}(s_{xy})$ denotes the sign of $s_{xy}$ which can be positive or negative. So

$$\text{sign}(s_{xy}) = \begin{cases} 1 & \text{if } s_{xy} > 0 \\ -1 & \text{if } s_{xy} < 0. \end{cases}$$

Notice that this gives two solutions for $\hat{\beta}_{1\text{OR}}$. We choose the solution which minimizes $\sum_{i=1}^{n} d_i^2$. The other solution maximizes $\sum_{i=1}^{n} d_i^2$ and is in the direction perpendicular to the optimal solution. The optimal solution can be chosen with the sign of $s_{xy}$. 

---

**Econometrics** | Chapter 2 | Simple Linear Regression Analysis | Shalabh, IIT Kanpur

36
Reduced major axis regression method:

The direct, reverse and orthogonal methods of estimation minimize the errors in a particular direction which is usually the distance between the observed data points and the line in the scatter diagram. Alternatively, one can consider the area extended by the data points in certain neighbourhood and instead of distances, the area of rectangles defined between corresponding observed data point and nearest point on the line in the following scatter diagram can also be minimized. Such an approach is more appropriate when the uncertainties are present in study as well as explanatory variables. This approach is termed as reduced major axis regression.

Suppose the regression line is \( Y_i = \beta_0 + \beta_1 X_i \) on which all the observed points are expected to lie. Suppose the points \((x_i, y_i), i = 1, 2, ..., n\) are observed which lie away from the line. The area of rectangle extended between the \(i^{th}\) observed data point and the line is

\[ A_i = (X_i - x_i)(Y_i - y_i) \quad (i = 1, 2, ..., n) \]

where \((X_i, Y_i)\) denotes the \(i^{th}\) pair of observation without any error which lie on the line.

The total area extended by \(n\) data points is

\[ \sum_{i=1}^{n} A_i = \sum_{i=1}^{n} (X_i - x_i)(Y_i - y_i). \]

All observed data points \((x_i, y_i), (i = 1, 2, ..., n)\) are expected to lie on the line

\[ Y_i = \beta_0 + \beta_1 X_i \]
and let

\[ E_i^* = Y_i - \beta_0 - \beta_1 X_i = 0. \]

So now the objective is to minimize the sum of areas under the constraints \( E_i^* \) to obtain the reduced major axis estimates of regression coefficients. Using the Lagrangian multiplies method, the Lagrangian function is

\[
L_R = \sum_{i=1}^{n} A_i - \sum_{i=1}^{n} \mu_i E_i^* = \sum_{i=1}^{n} (X_i - x_i)(Y_i - y_i) - \sum_{i=1}^{n} \mu_i E_i^*
\]

where \( \mu_1, \ldots, \mu_n \) are the Lagrangian multipliers. The set of equations are obtained by setting

\[
\frac{\partial L_R}{\partial X_i} = 0, \frac{\partial L_R}{\partial Y_i} = 0, \frac{\partial L_R}{\partial \beta_0} = 0, \frac{\partial L_R}{\partial \beta_i} = 0 \quad (i = 1, 2, \ldots, n).
\]

Thus

\[
\frac{\partial L_R}{\partial X_i} = (Y_i - y_i) + \beta_1 \mu_i = 0
\]

\[
\frac{\partial L_R}{\partial Y_i} = (X_i - x_i) - \mu_i = 0
\]

\[
\frac{\partial L_R}{\partial \beta_0} = \sum_{i=1}^{n} \mu_i = 0
\]

\[
\frac{\partial L_R}{\partial \beta_i} = \sum_{i=1}^{n} \mu_i X_i = 0.
\]

Now

\[
X_i = x_i + \mu_i
\]

\[
Y_i = y_i - \beta_1 \mu_i
\]

\[
\beta_0 + \beta_1 X_i = y_i - \beta_1 \mu_i
\]

\[
\beta_0 + \beta_1 (x_i + \mu_i) = y_i - \beta_1 \mu_i
\]

\[
\Rightarrow \mu_i = \frac{y_i - \beta_0 - \beta_1 x_i}{2\beta_1}.
\]

Substituting \( \mu_i \) in \( \sum_{i=1}^{n} \mu_i = 0 \), the reduced major axis regression estimate of \( \beta_0 \) is obtained as

\[
\hat{\beta}_{0RM} = \overline{y} - \hat{\beta}_{1RM} \overline{x}
\]

where \( \hat{\beta}_{1RM} \) is the reduced major axis regression estimate of \( \beta_1 \). Using \( X_i = x_i + \mu_i, \mu_i \) and \( \hat{\beta}_{0RM} \) in

\[
\sum_{i=1}^{n} \mu_i X_i = 0 \]

we get
\[
\sum_{i=1}^{n} \left( \frac{y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i}{2\beta_1} \right) \left( x_i - \frac{y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i}{2\beta_1} \right) = 0.
\]

Let \( u_i = x_i - \bar{x} \) and \( v_i = y_i - \bar{y} \), then this equation can be re-expressed as
\[
\sum_{i=1}^{n} (v_i - \beta_1 u_i)(v_i + \beta_1 u_i + 2\beta_1 \bar{x}) = 0.
\]

Using \( \sum_{i=1}^{n} u_i = \sum_{i=1}^{n} V_i = 0 \), we get
\[
\sum_{i=1}^{n} v_i^2 - \beta_1^2 \sum_{i=1}^{n} u_i^2 = 0.
\]

Solving this equation, the reduced major axis regression estimate of \( \beta_1 \) is obtained as
\[
\hat{\beta}_{RM} = \text{sign}(s_{xy}) \sqrt{\frac{s_{xy}}{s_{xx}}}
\]

where \( \text{sign}(s_{xy}) = \begin{cases} 1 & \text{if } s_{xy} > 0 \\ -1 & \text{if } s_{xy} < 0 \end{cases} \)

We choose the regression estimator which has same sign as of \( s_{xy} \).

**Least absolute deviation regression method**

The least squares principle advocates the minimization of sum of squared errors. The idea of squaring the errors is useful in place of simple errors because the random errors can be positive as well as negative. So consequently their sum can be close to zero indicating that there is no error in the model which can be misleading. Instead of the sum of random errors, the sum of absolute random errors can be considered which avoids the problem due to positive and negative random errors.

In the method of least squares, the estimates of the parameters \( \beta_0 \) and \( \beta_1 \) in the model
\[y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \ (i = 1, 2, ..., n)\]
are chosen such that the sum of squares of deviations \( \sum_{i=1}^{n} \varepsilon_i^2 \) is minimum. In the method of least absolute deviation (LAD) regression, the parameters \( \beta_0 \) and \( \beta_1 \) are estimated such that the sum of absolute deviations \( \sum_{i=1}^{n} |\varepsilon_i| \) is minimum. It minimizes the absolute vertical sum of errors as in the following scatter diagram:
The LAD estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of $\beta_0$ and $\beta_1$, respectively which minimize

$$LAD(\beta_0, \beta_1) = \sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i|$$

for the given observations $(x_i, y_i) (i = 1, 2, ..., n)$.

Conceptually, LAD procedure is simpler than OLS procedure because $|e|$ (absolute residuals) is a more straightforward measure of the size of the residual than $e^2$ (squared residuals). The LAD regression estimates of $\beta_0$ and $\beta_1$ are not available in closed form. Rather they can be obtained numerically based on algorithms. Moreover, this creates the problems of non-uniqueness and degeneracy in the estimates. The concept of non-uniqueness relates to that more than one best lines pass through a data point. The degeneracy concept describes that the best line through a data point also passes through more than one other data points. The non-uniqueness and degeneracy concepts are used in algorithms to judge the quality of the estimates. The algorithm for finding the estimators generally proceeds in steps. At each step, the best line is found that passes through a given data point. The best line always passes through another data point, and this data point is used in the next step. When there is non-uniqueness, then there are more than one best lines. When there is degeneracy, then the best line passes through more than one other data point. When either of the problem is present, then there is more than one choice for the data point to be used in the next step and the algorithm may go around in circles or make a wrong choice of the LAD regression line. The exact tests of hypothesis and confidence intervals for the LAD regression estimates can not be derived analytically. Instead they are derived analogous to the tests of hypothesis and confidence intervals related to ordinary least squares estimates.
Estimation of parameters when $X$ is stochastic

In a usual linear regression model, the study variable is supposed to be random and explanatory variables are assumed to be fixed. In practice, there may be situations in which the explanatory variable also becomes random.

Suppose both dependent and independent variables are stochastic in the simple linear regression model

$$y = \beta_0 + \beta_1 X + \varepsilon$$

where $\varepsilon$ is the associated random error component. The observations $(x_i, y_i), i = 1, 2, ..., n$ are assumed to be jointly distributed. Then the statistical inferences can be drawn in such cases which are conditional on $X$.

Assume the joint distribution of $X$ and $y$ to be bivariate normal $N(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ where $\mu_x$ and $\mu_y$ are the means of $X$ and $y$; $\sigma_x^2$ and $\sigma_y^2$ are the variances of $X$ and $y$; and $\rho$ is the correlation coefficient between $X$ and $y$. Then the conditional distribution of $y$ given $X = x$ is univariate normal conditional mean

$$E(y | X = x) = \mu_{y|x} = \beta_0 + \beta_1 x$$

and conditional variance of $y$ given $X = x$ is

$$Var(y | X = x) = \sigma_{y|x}^2 = \sigma_y^2 (1 - \rho^2)$$

where

$$\beta_0 = \mu_y - \mu_x \beta_1$$

and

$$\beta_1 = \frac{\sigma_y}{\sigma_x} \rho.$$

When both $X$ and $y$ are stochastic, then the problem of estimation of parameters can be reformulated as follows. Consider a conditional random variable $y | X = x$ having a normal distribution with mean as conditional mean $\mu_{y|x}$ and variance as conditional variance $Var(y | X = x) = \sigma_{y|x}^2$. Obtain $n$ independently distributed observation $y_i | x_i, i = 1, 2, ..., n$ from $N(\mu_{y|x}, \sigma_{y|x}^2)$ with nonstochastic $X$. Now the method of maximum likelihood can be used to estimate the parameters which yields the estimates of $\beta_0$ and $\beta_1$ as earlier in the case of nonstochastic $X$ as
\[ \hat{b} = \bar{y} - \hat{b}\bar{x} \]

and

\[ \hat{b}_i = \frac{s_{xy}}{s_{xx}} \]

respectively.

Moreover, the correlation coefficient

\[ \rho = \frac{E(y - \mu_y)(X - \mu_x)}{\sigma_y \sigma_x} \]

can be estimated by the sample correlation coefficient

\[
\hat{\rho} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}} = \hat{b}_i \frac{s_{xx}}{s_{yy}}
\]

Thus

\[
\hat{\rho}^2 = \hat{b}_i^2 \frac{s_{xx}}{s_{yy}} = \hat{b}_i^2 \frac{s_{xy}}{s_{yy}} = \frac{s_{xy} - \sum_{i=1}^{n} \hat{e}_i^2}{s_{yy}} = R^2
\]

which is same as the coefficient of determination. Thus \( R^2 \) has the same expression as in the case when \( X \) is fixed. Thus \( R^2 \) again measures the goodness of fitted model even when \( X \) is stochastic.
Chapter 3

Multiple Linear Regression Model

We consider the problem of regression when study variable depends on more than one explanatory or independent variables, called as multiple linear regression model. This model generalizes the simple linear regression in two ways. It allows the mean function \( E(y) \) to depend on more than one explanatory variables and to have shapes other than straight lines, although it does not allow for arbitrary shapes.

The linear model:

Let \( y \) denotes the dependent (or study) variable that is linearly related to \( k \) independent (or explanatory) variables \( X_1, X_2, \ldots, X_k \) through the parameters \( \beta_1, \beta_2, \ldots, \beta_k \) and we write

\[
y = X_1\beta_1 + X_2\beta_2 + \ldots + X_k\beta_k + \epsilon.
\]

This is called as the multiple linear regression model. The parameters \( \beta_1, \beta_2, \ldots, \beta_k \) are the regression coefficients associated with \( X_1, X_2, \ldots, X_k \) respectively and \( \epsilon \) is the random error component reflecting the difference between the observed and fitted linear relationship. There can be various reasons for such difference, e.g., joint effect of those variables not included in the model, random factors which can not be accounted in the model etc.

Note that the \( j^{th} \) regression coefficient \( \beta_j \) represents the expected change in \( y \) per unit change in \( j^{th} \) independent variable \( X_j \). Assuming \( E(\epsilon) = 0 \),

\[
\beta_j = \frac{\partial E(y)}{\partial X_j}.
\]

Linear model:

A model is said to be linear when it is linear in parameters. In such a case \( \frac{\partial y}{\partial \beta_j} \) (or equivalently \( \frac{\partial E(y)}{\partial X_j} \)) should not depend on any \( \beta^{'s} \). For example

i) \( y = \beta_0 + \beta_1 X \) is a linear model as it is linear in parameter.

ii) \( y = \beta_0 X^\beta \) can be written as

\[
\log y = \log \beta_0 + \beta_\log X
\]

\[
y^* = \beta_0^* + \beta_\log X^*
\]
which is linear is parameter $\beta_0^*$ and $\beta_1$, but nonlinear is variables $y^* = \log y, x^* = \log x$. So it is a linear model.

iii) $y = \beta_0 + \beta_1 X + \beta_2 X^2$

is linear in parameters $\beta_0, \beta_1$ and $\beta_2$ but it is nonlinear is variables $X$. So it is a linear model.

iv) $y = \beta_0 + \frac{\beta_1}{X - \beta_2}$

is nonlinear in parameters and variables both. So it is a nonlinear model.

v) $y = \beta_0 + \beta_1 X^{\beta_2}$

is nonlinear in parameters and variables both. So it is a nonlinear model.

vi) $y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$

is a cubic polynomial model which can be written as

$$y = \beta_0 + \beta_1 X + \beta_2 X_1 + \beta_3 X_2$$

which is linear in parameters $\beta_0, \beta_1, \beta_2, \beta_3$ and linear in variables $X_1 = X, X_2 = X^2, X_3 = X^3$. So it is a linear model.

Example:

The income and education of a person are related. It is expected that, on an average, higher level of education provides higher income. So a simple linear regression model can be expressed as

$$\text{income} = \beta_0 + \beta_1 \text{education} + \varepsilon.$$ 

Not that $\beta_1$ reflects the change in income with respect to per unit change in education and $\beta_0$ reflects the income when education is zero as it is expected that even an illiterate person can also have some income.

Further this model neglects that most people have higher income when they are older than when they are young, regardless of education. So $\beta_1$ will over-state the marginal impact of education. If age and education are positively correlated, then the regression model will associate all the observed increase in income with an increase in education. So better model is

$$\text{income} = \beta_0 + \beta_1 \text{education} + \beta_2 \text{age} + \varepsilon.$$
Often it is observed that the income tends to rise less rapidly in the later earning years than is early years. To accommodate such possibility, we might extend the model to

\[
\text{income} = \beta_0 + \beta_1 \text{education} + \beta_2 \text{age} + \beta_3 \text{age}^2 + \varepsilon
\]

This is how we proceed for regression modeling in real life situation. One needs to consider the experimental condition and the phenomenon before taking the decision on how many, why and how to choose the dependent and independent variables.

**Model set up:**

Let an experiment be conducted \( n \) times and the data is obtained as follows:

<table>
<thead>
<tr>
<th>Observation number</th>
<th>Response ( y )</th>
<th>Explanatory variables ( X_1 ) ( X_2 ) ( \cdots ) ( X_k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( y_1 )</td>
<td>( x_{11} ) ( x_{12} ) ( \cdots ) ( x_{1k} )</td>
</tr>
<tr>
<td>2</td>
<td>( y_2 )</td>
<td>( x_{21} ) ( x_{22} ) ( \cdots ) ( x_{2k} )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots ) ( \vdots ) ( \ddots ) ( \vdots )</td>
</tr>
<tr>
<td>( n )</td>
<td>( y_n )</td>
<td>( x_{n1} ) ( x_{n2} ) ( \cdots ) ( x_{nk} )</td>
</tr>
</tbody>
</table>

Assuming that the model is

\[
y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon,
\]

the \( n \)-tuples of observations are also assumed to follow the same model. Thus they satisfy

\[
\begin{align*}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \varepsilon_n.
\end{align*}
\]

These \( n \) equations can be written as

\[
\begin{pmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{pmatrix} =
\begin{pmatrix}
1 & x_{11} & x_{12} & \cdots & x_{1k} \\
1 & x_{21} & x_{22} & \cdots & x_{2k} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_{n1} & x_{n2} & \cdots & x_{nk}
\end{pmatrix}
\begin{pmatrix}
\beta_0 \\
\beta_1 \\
\vdots \\
\beta_k
\end{pmatrix}
+ 
\begin{pmatrix}
\varepsilon_1 \\
\varepsilon_2 \\
\vdots \\
\varepsilon_n
\end{pmatrix}
\]

or \( y = X\beta + \varepsilon. \)
In general, the model with \( k \) explanatory variables can be expressed as

\[
y = X\beta + \varepsilon
\]

where \( y = (y_1, y_2, \ldots, y_n)' \) is a \( n \times 1 \) vector of \( n \) observation on study variable, \( X = \begin{pmatrix}
x_{11} & x_{12} & \cdots & x_{1k} \\
x_{21} & x_{22} & \cdots & x_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{nk}
\end{pmatrix}
\)

is a \( n \times k \) matrix of \( n \) observations on each of the \( k \) explanatory variables, \( \beta = (\beta_1, \beta_2, \ldots, \beta_k)' \) is a \( k \times 1 \) vector of regression coefficients and \( \varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)' \) is a \( n \times 1 \) vector of random error components or disturbance term.

If intercept term is present, take first column of \( X \) to be \((1,1,\ldots,1)'\).

**Assumptions in multiple linear regression model**

Some assumptions are needed in the model \( y = X\beta + \varepsilon \) for drawing the statistical inferences. The following assumptions are made:

(i) \( E(\varepsilon) = 0 \)
(ii) \( E(\varepsilon\varepsilon') = \sigma^2 I_n \)
(iii) \( \text{Rank}(X) = k \)
(iv) \( X \) is a non-stochastic matrix
(v) \( \varepsilon \sim N(0, \sigma^2 I_n) \).

These assumptions are used to study the statistical properties of estimator of regression coefficients. The following assumption is required to study particularly the large sample properties of the estimators

(vi) \( \lim_{n \to \infty} \left( \frac{X'X}{n} \right) = \Delta \) exists and is a non-stochastic and nonsingular matrix (with finite elements).

The explanatory variables can also be stochastic in some cases. We assume that \( X \) is non-stochastic unless stated separately.

We consider the problems of estimation and testing of hypothesis on regression coefficient vector under the stated assumption.

*Econometrics*  |  Chapter 3  |  Multiple Linear Regression Model  |  Shalabh, IIT Kanpur  
--- | --- | --- | ---
**Estimation of parameters:**

A general procedure for the estimation of regression coefficient vector is to minimize

\[ \sum_{i=1}^{n} M(\varepsilon_i) = \sum_{i=1}^{n} M(y_i - x_{i1}\beta_1 - x_{i2}\beta_2 - \ldots - x_{ik}\beta_k) \]

for a suitably chosen function \( M \).

Some examples of choice of \( M \) are

- \( M(x) = |x| \)
- \( M(x) = x^2 \)
- \( M(x) = |x|^p \), in general.

We consider the principle of least square which is related to \( M(x) = x^2 \) and method of maximum likelihood estimation for the estimation of parameters.

**Principle of ordinary least squares (OLS)**

Let \( B \) be the set of all possible vectors \( \beta \). If there is no further information, the \( B \) is \( k \)-dimensional real Euclidean space. The object is to find a vector \( b' = (b_1, b_2, \ldots, b_k) \) from \( B \) that minimizes the sum of squared deviations of \( \varepsilon_i ' s \), i.e.,

\[ S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon' \varepsilon = (y - X\beta)'(y - X\beta) \]

for given \( y \) and \( X \). A minimum will always exist as \( S(\beta) \) is a real valued, convex and differentiable function. Write

\[ S(\beta) = y'y + \beta'X'X\beta - 2\beta'X'y \]

Differentiate \( S(\beta) \) with respect to \( \beta \)

\[ \frac{\partial S(\beta)}{\partial \beta} = 2X'X\beta - 2X'y \]

\[ \frac{\partial^2 S(\beta)}{\partial \beta^2} = 2X'X \] (atleast non-negative definite).

The normal equation is

\[ \frac{\partial S(\beta)}{\partial \beta} = 0 \]

\[ \Rightarrow X'Xb = X'y \]
where the following result is used:

**Result:** If \( f(z) = Z'AZ \) is a quadratic form, \( Z \) is a \( m \times 1 \) vector and \( A \) is any \( m \times m \) symmetric matrix then \( \frac{\partial}{\partial z} F(z) = 2Az \).

Since it is assumed that rank \( (X) = k \) (full rank), then \( X'X \) is positive definite and unique solution of normal equation is

\[
  b = (X'X)^{-1}X'y
\]

which is termed as **ordinary least squares estimator** (OLSE) of \( \beta \).

Since \( \frac{\partial^2 S(\beta)}{\partial \beta^2} \) is at least non-negative definite, so \( b \) minimize \( S(\beta) \).

In case, \( X \) is **not of full rank**, then

\[
  b = (X'X)^{-}X'y + [I - (X'X)^{-}X'X] \omega
\]

where \( (X'X)^{-} \) is the generalized inverse of \( X'X \) and \( \omega \) is an arbitrary vector. The generalized inverse \( (X'X)^{-} \) of \( X'X \) satisfies

\[
  X'X(X'X)^{-}X'X = X'X \\
  X(X'X)^{-}X'X = X \\
  X'X(X'X)^{-}X' = X'
\]

**Theorem:**

(i) Let \( \hat{y} = Xb \) be the empirical predictor of \( y \). Then \( \hat{y} \) has the same value for all solutions \( b \) of \( X'Xb = X'y \).

(ii) \( S(\beta) \) attains the minimum for any solution of \( X'Xb = X'y \).

**Proof:**

(i) Let \( b \) be any member in

\[
  b = (X'X)^{-}X'y + [I - (X'X)^{-}X'X] \omega
\]

Since \( X(X'X)^{-}X'X = X \), so then

\[
  Xb = X(X'X)^{-}X'y + X[I - (X'X)^{-}X'X] \omega
\]

\[
  = X(X'X)^{-}X'y
\]

which is independent of \( \omega \). This implies that \( \hat{y} \) has same value for all solution \( b \) of \( X'Xb = X'y \).
(ii) Note that for any $\beta$,

$$S(\beta) = \left[ y - Xb + X(b - \beta) \right] \left[ y - Xb + X(b - \beta) \right]$$

$$= (y - Xb)'(y - Xb) + (b - \beta)'X'X(b - \beta) + 2(b - \beta)'X'(y - Xb)$$

$$= (y - Xb)'(y - Xb) + (b - \beta)'X'X(b - \beta) \quad \text{(Using } X'Xb = X'y)$$

$$\geq (y - Xb)'(y - Xb) = S(b)$$

$$= y'y - 2y'Xb + b'X'Xb$$

$$= y'y - b'X'Xb$$

$$= y'y - \hat{y}'\hat{y}.$$ 

**Fitted values:**

If $\hat{\beta}$ is any estimator of $\beta$ for the model $y = X\beta + \varepsilon$, then the fitted values are defined as

$$\hat{y} = X\hat{\beta}$$

where $\hat{\beta}$ is any estimator of $\beta$.

In case of $\hat{\beta} = b$,

$$\hat{y} = Xb$$

$$= X(X'X)^{-1}X'y$$

$$= Hy$$

where $H = X(X'X)^{-1}X'$ is termed as **Hat matrix** which is

(i) symmetric

(ii) idempotent (i.e., $HH = H$) and

(iii) $tr H = tr X(X'X)^{-1}X' = tr X'X(X'X)^{-1} = tr I_k = k$.

**Residuals**

The difference between the observed and fitted values of study variable is called as residual. It is denoted as

$$e = y - \hat{y}$$

$$= y' - \hat{y}'$$

$$= y' - Xb$$

$$= y' - Hy$$

$$= (I - H)y$$

$$= \widetilde{H}y$$

where $\widetilde{H} = I - H$. 

---

*Econometrics  | Chapter 3  | Multiple Linear Regression Model  | Shalabh, IIT Kanpur*
Note that

(i)  \( \bar{H} \) is a symmetric matrix

(ii) \( \bar{H} \) is an idempotent matrix, i.e.,
\[
\bar{H}\bar{H} = (I - H)(I - H) = (I - H) = \bar{H}
\]
and

(iii) \( tr\bar{H} = trI_n - trH = (n - k) \).

Properties of OLSE

(i) Estimation error:
The estimation error of \( b \) is
\[
b - \beta = (X'X)^{-1}X'y - \beta \\
= (X'X)^{-1}X'(X\beta + \varepsilon) - \beta \\
= (X'X)^{-1}X'\varepsilon
\]

(ii) Bias
Since \( X \) is assumed to be nonstochastic and \( E(\varepsilon) = 0 \)
\[
E(b - \beta) = (X'X)^{-1}X'E(\varepsilon) \\
= 0.
\]
Thus OLSE is an unbiased estimator of \( \beta \).

(iii) Covariance matrix
The covariance matrix of \( b \) is
\[
V(b) = E(b - \beta)(b - \beta)' \\
= E\left[(X'X)^{-1}X'E\varepsilon X(X'X)^{-1}\right] \\
= (X'X)^{-1}X'E(\varepsilon\varepsilon')X(X'X)^{-1} \\
= \sigma^2(X'X)^{-1}X'I(X'X)^{-1} \\
= \sigma^2(X'X)^{-1}.
\]
(iv) Variance

The variance of $b$ can be obtained as the sum of variances of all $b_1, b_2, ..., b_k$ which is the trace of covariance matrix of $b$. Thus

$$Var(b) = tr[V(b)]$$

$$= \sum_{i=1}^{k} E(b_i - \beta_i)^2$$

$$= \sum_{i=1}^{k} Var(b_i).$$

Estimation of $\sigma^2$

The least squares criterion cannot be used to estimate $\sigma^2$ because $\sigma^2$ does not appear in $S(\beta)$. Since $E(\varepsilon_i^2) = \sigma^2$, so we attempt with residuals $e_i$ to estimate $\sigma^2$ as follows:

$$e = y - \hat{y}$$

$$= y - X(X'X)^{-1}X'y$$

$$= [I - X(X'X)^{-1}X']y$$

$$= \bar{Hy}.$$

Consider the residual sum of squares

$$SS_{res} = \sum_{i=1}^{n} e_i^2$$

$$= e'e$$

$$= (y - Xb)'(y - Xb)$$

$$= y'(I - H)(I - H)y$$

$$= y'(I - H)y$$

$$= y'\bar{Hy}.$$

Also

$$SS_{res} = (y - Xb)'(y - Xb)$$

$$= y'y - 2b'X'y + b'X'Xb$$

$$= y'y - b'X'y \quad (\text{Using } X'Xb = X'y)$$

$$SS_{res} = y'\bar{Hy}$$

$$= (X\beta + \varepsilon)'\bar{H}(X\beta + \varepsilon)$$

$$= \varepsilon'\bar{H}\varepsilon \quad (\text{Using } \bar{H}X = 0)$$
Since \( \varepsilon \sim N(0, \sigma^2 I) \).

So \( y \sim N(X\beta, \sigma^2 I) \).

Hence \( y'\bar{H}y \sim \chi^2(n-k) \).

Thus \( E[y'\bar{H}y] = (n-k)\sigma^2 \)

or \( E\left[\frac{y'\bar{H}y}{n-k}\right] = \sigma^2 \)

or \( E[MS_{res}] = \sigma^2 \)

where \( MS_{res} = \frac{SS_{res}}{n-k} \) is the mean sum of squares due to residual.

Thus an unbiased estimator of \( \sigma^2 \) is

\[ \hat{\sigma}^2 = MS_{res} = s^2 \text{ (say)} \]

which is a model dependent estimator.

**Variance of \( \hat{y} \)**

The variance of \( \hat{y} \) is

\[
V(\hat{y}) = V(Xb) \\
= XV(b)X' \\
= \sigma^2 X(X'X)^{-1}X' \\
= \sigma^2 H.
\]

**Gauss-Markov Theorem:**

The ordinary least squares estimator (OLSE) is the best linear unbiased estimator (BLUE) of \( \beta \).

**Proof:** The OLSE of \( \beta \) is

\[ b = (X'X)^{-1}X'y \]

which is a linear function of \( y \). Consider the arbitrary linear estimator

\[ b^* = a'y \]

of linear parametric function \( \ell'y\beta \) where the elements of \( a \) are arbitrary constants.

Then for \( b^* \),

\[ E(b^*) = E(a'y) = a'X\beta \]
and so $b^*$ is an unbiased estimator of $\ell' \beta$ when

$$E(b^*) = a'X\beta = \ell' \beta$$

$$\Rightarrow a'X = \ell'.$$

Since we wish to consider only those estimators that are linear and unbiased, so we restrict ourselves to those estimators for which $a'X = \ell'.$

Further

$$Var(a'y) = a'Var(y)a = \sigma^2 a'a$$

$$Var(\ell'b) = \ell'Var(b)\ell = \sigma^2 a'X(X'X)^{-1}X'a.$$ 

Consider

$$Var(a'y) - Var(\ell'b) = \sigma^2 \left[ a'a - a'X(X'X)^{-1}X'a \right]$$

$$= \sigma^2 a' \left[ I - X(X'X)^{-1}X' \right] a$$

$$= \sigma^2 a'(I - H)a.$$ 

Since $(I - H)$ is a positive semi-definite matrix, so

$$Var(a'y) - Var(\ell'b) \geq 0.$$ 

This reveals that if $b^*$ is any linear unbiased estimator then its variance must be no smaller than that of $b.$ Consequently $b$ is the best linear unbiased estimator, where ‘best’ refers to the fact that $b$ is efficient within the class of linear and unbiased estimators.

**Maximum likelihood estimation:**

In the model $y = X\beta + \varepsilon,$ it is assumed that the errors are normally and independently distributed with constant variance $\sigma^2$ or $\varepsilon \sim N(0, \sigma^2 I).$

The normal density function for the errors is

$$f(\varepsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[ -\frac{1}{2\sigma^2} \varepsilon_i^2 \right] \quad i = 1, 2, \ldots, n.$$ 

The likelihood function is the joint density of $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ given as
\[ L(\beta, \sigma^2) = \prod_{i=1}^{n} f(\varepsilon_i) \]
\[ = \frac{1}{(2\pi \sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \varepsilon_i^2 \right] \]
\[ = \frac{1}{(2\pi \sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \varepsilon' \varepsilon \right] \]
\[ = \frac{1}{(2\pi \sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} (y-X\beta)'(y-X\beta) \right]. \]

Since the log transformation is monotonic, so we maximize \(\ln L(\beta, \sigma^2)\) instead of \(L(\beta, \sigma^2)\).

\[ \ln L(\beta, \sigma^2) = -\frac{n}{2} \ln(2\pi \sigma^2) - \frac{1}{2\sigma^2} (y-X\beta)'(y-X\beta). \]

The maximum likelihood estimators (m.l.e.) of \(\beta\) and \(\sigma^2\) are obtained by equating the first order derivatives of \(\ln L(\beta, \sigma^2)\) with respect to \(\beta\) and \(\sigma^2\) to zero as follows:

\[ \frac{\partial \ln L(\beta, \sigma^2)}{\partial \beta} = \frac{1}{2\sigma^2} 2X'(y-X\beta) = 0 \]
\[ \frac{\partial \ln L(\beta, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} (y-X\beta)'(y-X\beta). \]

The likelihood equations are given by

\[ X'X\beta = X'y \]
\[ \sigma^2 = \frac{1}{n} (y-X\beta)'(y-X\beta). \]

Since \(\text{rank}(X) = k\), so that the unique mle of \(\beta\) and \(\sigma^2\) are obtained as

\[ \hat{\beta} = (X'X)^{-1} X'y \]
\[ \hat{\sigma}^2 = \frac{1}{n} (y-X\hat{\beta})'(y-X\hat{\beta}). \]

Further to verify that these values maximize the likelihood function, we find

\[ \frac{\partial^2 \ln L(\beta, \sigma^2)}{\partial \beta^2} = -\frac{1}{\sigma^2} X'X \]
\[ \frac{\partial^2 \ln L(\beta, \sigma^2)}{\partial^2 (\sigma^2)^2} = -\frac{n}{2\sigma^4} - \frac{1}{\sigma^6} (y-X\beta)'(y-X\beta) \]
\[ \frac{\partial^2 \ln L(\beta, \sigma^2)}{\partial \beta \partial \sigma^2} = -\frac{1}{\sigma^4} X'(y-X\beta). \]
Thus the Hessian matrix of second order partial derivatives of \( \ln L(\beta, \sigma^2) \) with respect to \( \beta \) and \( \sigma^2 \) is

\[
\begin{pmatrix}
\frac{\partial^2 \ln L(\beta, \sigma^2)}{\partial \beta^2} & \frac{\partial^2 \ln L(\beta, \sigma^2)}{\partial \beta \partial \sigma^2} \\
\frac{\partial^2 \ln L(\beta, \sigma^2)}{\partial \sigma^2 \partial \beta} & \frac{\partial^2 \ln L(\beta, \sigma^2)}{\partial (\sigma^2)^2}
\end{pmatrix}
\]

which is negative definite at \( \beta = \hat{\beta} \) and \( \sigma^2 = \hat{\sigma}^2 \). This ensures that the likelihood function is maximized at these values.

Comparing with OLSEs, we find that

(i) OLSE and m.l.e. of \( \beta \) are same. So m.l.e. of \( \beta \) is also an unbiased estimator of \( \beta \).

(ii) OLSE of \( \sigma^2 \) is \( s^2 \) which is related to m.l.e. of \( \sigma^2 \) as \( \hat{\sigma}^2 = \frac{n-k}{n} s^2 \). So m.l.e. of \( \sigma^2 \) is a biased estimator of \( \sigma^2 \).

**Consistency of estimators**

(i) **Consistency of \( \beta \):**

Under the assumption that \( \lim_{n \to \infty} \left( \frac{X'X}{n} \right) = \Delta \) exists as a nonstochastic and nonsingular matrix (with finite elements), we have

\[
\lim V(\hat{\beta}) = \sigma^2 \lim_{n \to \infty} \frac{1}{n} \left( \frac{X'X}{n} \right)^{-1} = \sigma^2 \lim_{n \to \infty} \frac{1}{n} \Delta^{-1} = 0.
\]

This implies that OLSE converges to \( \beta \) in quadratic mean. Thus OLSE is a consistent estimator of \( \beta \). This holds true for maximum likelihood estimators also.

Same conclusion can also be proved using the concept of convergence in probability.

An estimator \( \hat{\theta}_n \) converges to \( \theta \) in probability if

\[
\lim_{n \to \infty} P \left( |\hat{\theta}_n - \theta| \geq \delta \right) = 0 \quad \text{for any} \quad \delta > 0
\]

and is denoted as \( \text{plim}(\hat{\theta}_n) = \theta \).
The consistency of OLSE can be obtained under the weaker assumption that
\[ \text{plim} \left( \frac{X'X}{n} \right) = \Delta. \]
exists and is a nonsingular and nonstochastic matrix such that
\[ \text{plim} \left( \frac{X'e}{n} \right) = 0. \]

Since
\[ b - \beta = (X'X)^{-1}X'e \]
\[ = \left( \frac{X'X}{n} \right)^{-1} \frac{X'e}{n}. \]

So
\[ \text{plim}(b - \beta) = \text{plim} \left( \frac{X'X}{n} \right)^{-1} \text{plim} \left( \frac{X'e}{n} \right) \]
\[ = \Delta^{-1} \cdot 0 \]
\[ = 0. \]

Thus \( b \) is a consistent estimator of \( \beta \). Same is true for m.l.e. also.

(ii) Consistency of \( s^2 \)

Now we look at the consistency of \( s^2 \) as an estimate of \( \sigma^2 \) as
\[ s^2 = \frac{1}{n-k} e'e \]
\[ = \frac{1}{n-k} e'\bar{H}e \]
\[ = \frac{1}{n} \left( 1 - \frac{k}{n} \right)^{-1} \left[ e'e - e'X(X'X)^{-1}X'e \right] \]
\[ = \left( 1 - \frac{k}{n} \right)^{-1} \left[ \frac{e'e}{n} - \frac{e'X(X'X)^{-1}X'e}{n} \right]. \]

Note that \( \frac{e'e}{n} \) consists of \( \frac{1}{n} \sum_{i=1}^{n} e_i^2 \) and \( \{e_i^2, i = 1, 2, ..., n\} \) is a sequence of independently and identically distributed random variables with mean \( \sigma^2 \). Using the law of large numbers.
\[
plim \left( \frac{\varepsilon' \varepsilon}{n} \right) = \sigma^2
\]
\[
plim \left[ \frac{\varepsilon' X (X'X)^{-1} X' \varepsilon}{n} \right] = \left( \frac{\varepsilon' X}{n} \right) \left( \frac{\varepsilon' X}{n} \right)^{-1} \left( \frac{\varepsilon' X}{n} \right)
\]
\[
= 0. \Delta^{-1}. 0
\]
\[
\Rightarrow \text{plim}(s^2) = (1 - 0)^{-1} \left[ \sigma^2 - 0 \right]
\]
\[
= \sigma^2.
\]
Thus \( s^2 \) is a consistent estimator of \( \sigma^2 \). Same hold true for m.l.e. also.

**Cramer-Rao lower bound**

Let \( \theta = (\beta, \sigma^2)' \). Assume that both \( \beta \) and \( \sigma^2 \) are unknown. If \( E(\hat{\theta}) = \theta \), then the Cramer-Rao lower bound for \( \hat{\theta} \) is greater than or equal to the matrix inverse of

\[
I(\theta) = -E \left[ \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} \right]
\]

\[
= -E \left[ \frac{\partial \ln L(\theta, \sigma^2)}{\partial \beta^2} \right] - E \left[ \frac{\partial \ln L(\theta, \sigma^2)}{\partial \sigma^2 \partial \beta} \right] - E \left[ \frac{\partial \ln L(\theta, \sigma^2)}{\partial ^2 (\sigma^2)^2} \right]
\]

\[
= -E \left[ \frac{X'X}{\sigma^2} \right] - E \left[ \frac{X'(y-X\beta)}{\sigma^4} \right] - E \left[ \frac{n}{2\sigma^4} - \frac{(y-X\beta)'(y-X\beta)}{\sigma^6} \right]
\]

\[
= \begin{bmatrix}
\frac{X'X}{\sigma^2} & 0 \\
0 & \frac{n}{2\sigma^4}
\end{bmatrix}
\]

Then

\[
[I(\theta)]^{-1} = \begin{bmatrix}
\sigma^2(X'X)^{-1} & 0 \\
0 & \frac{2\sigma^4}{n}
\end{bmatrix}
\]

is the Cramer-Rao lower bound matrix of \( \beta \) and \( \sigma^2 \).
The covariance matrix of OLSEs of $\beta$ and $\sigma^2$ is

$$\Sigma_{OLS} = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n-k} \end{bmatrix}$$

which means that the Cramer-Rao bound is attained for the covariance of $b$ but not for $s^2$.

**Standardized regression coefficients:**

Usually it is difficult to compare the regression coefficients because the magnitude of $\hat{\beta}_j$ reflects the units of measurement of $j^{th}$ explanatory variable $X_j$. For example, in the following fitted regression model

$$\hat{y} = 5 + X_1 + 1000X_2,$$

$y$ is measured in liters, $X_1$ is liters and $X_2$ in milliliters. Although $\hat{\beta}_2 \gg \hat{\beta}_1$ but effect of both explanatory variables is identical. One liter change in either $X_1$ and $X_2$ when other variable is held fixed produces the same change is $\hat{y}$.

Sometimes it is helpful to work with scaled explanatory variables and study variable that produces dimensionless regression coefficients. These dimensionless regression coefficients are called as **standardized regression coefficients**.

There are two popular approaches for scaling which gives standardized regression coefficients. We discuss them as follows:

1. **Unit normal scaling:**

Employ unit normal scaling to each explanatory variable and study variable.

So define

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad i = 1,2,...,n, \quad j = 1,2,...,k$$

$$y_i^* = \frac{y_i - \bar{y}}{s_y}$$

where $s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$.
and \[ s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \]

are the sample variances of \( j^{th} \) explanatory variable and study variable respectively.

All scaled explanatory variable and scaled study variable have mean zero and sample variance unity, i.e., using these new variables, the regression model becomes
\[ y^*_i = \gamma_1 z_{i1} + \gamma_2 z_{i2} + \ldots + \gamma_k z_{ik} + \epsilon_i, \quad i = 1, 2, \ldots, n. \]
Such centering removes the intercept term from the model. The least squares estimate of \( \gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)' \) is
\[ \hat{\gamma} = (Z'Z)^{-1} Z' y^*. \]

This scaling has a similarity to standardizing a normal random variable, i.e., observation minus its mean and divided by its standard deviation. So it is called as a unit normal scaling.

2. Unit length scaling:

In unit length scaling, define
\[ \omega_j = \frac{x_{ij} - \bar{x}_j}{S_{ij}^{1/2}}, \quad i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, k \]
\[ y_{i0} = \frac{y_i - \bar{y}}{S_{S_0}^{1/2}} \]
where \( S_{ij} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \) is the corrected sum of squares for \( j^{th} \) explanatory variables \( X_j \) and \( S_T = S_{S_0} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \) is the total sum of squares. In this scaling, each new explanatory variable \( W_j \) has mean \( \bar{\omega}_j = \frac{1}{n} \sum_{i=1}^{n} \omega_j = 0 \) and length \( \sqrt{\sum_{i=1}^{n} (\omega_j - \bar{\omega}_j)^2} = 1. \)

In terms of these variables, the regression model is
\[ y_{i0} = \delta_1 \omega_{i1} + \delta_2 \omega_{i2} + \ldots + \delta_k \omega_{ik} + \epsilon_i, \quad i = 1, 2, \ldots, n. \]
The least squares estimate of regression coefficient \( \delta = (\delta_1, \delta_2, \ldots, \delta_k)' \) is
\[ \hat{\delta} = (W'W)^{-1} W' y_{i0}. \]
In such a case, the matrix $W'W$ is in the form of correlation matrix, i.e.,

$$
W'W = \begin{pmatrix}
1 & r_{12} & r_{13} & \cdots & r_{1k} \\
r_{12} & 1 & r_{23} & \cdots & r_{2k} \\
r_{13} & r_{23} & 1 & \cdots & r_{3k} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
r_{1k} & r_{2k} & r_{3k} & \cdots & 1
\end{pmatrix}
$$

where

$$
r_{ij} = \frac{\sum_{u=1}^{n} (x_{iu} - \bar{x}_i)(x_{uj} - \bar{x}_j)}{(S_i S_j)^{1/2}} = \frac{S_{ij}}{(S_i S_j)^{1/2}}
$$

is the simple correlation coefficient between the explanatory variables $X_i$ and $X_j$. Similarly

$$
W'y^o = (r_{1y}, r_{2y}, \ldots, r_{ky})'
$$

where

$$
r_{jy} = \frac{\sum_{u=1}^{n} (x_{uj} - \bar{x}_j)(y_u - \bar{y})}{(S_j S_T)^{1/2}} = \frac{S_{jy}}{(S_j S_T)^{1/2}}
$$

is the simple correlation coefficient between $j^{th}$ explanatory variable $X_j$ and study variable $y$.

Note that it is customary to refer $r_{ij}$ and $r_{jy}$ as correlation coefficient though $X_i$'s are not random variable.

If unit normal scaling is used, then

$$
Z'Z = (n-1)W'W.
$$

So the estimates of regression coefficient in unit normal scaling (i.e., $\hat{\gamma}$) and unit length scaling (i.e., $\hat{\delta}$) are identical. So it does not matter which scaling is used, so $\hat{\gamma} = \hat{\delta}$.

The regression coefficients obtained after such scaling, viz., $\hat{\gamma}$ or $\hat{\delta}$ usually called standardized regression coefficients.
The relationship between the original and standardized regression coefficients is

\[ b_j = \hat{\delta}_j \left( \frac{SS_T}{S_{jj}} \right)^{1/2}, \quad j = 1, 2, ..., k \]

and

\[ b_0 = \bar{y} - \sum_{j=1}^{k} b_j \bar{x}_j \]

where \( b_0 \) is the OLSE of intercept term and \( b_j \) are the OLSE of slope parameters.

**The model in deviation form**

The multiple linear regression model can also be expressed in the deviation form.

First all the data is expressed in terms of deviations from sample mean.

The estimation of regression parameters is performed in two steps:

- **First step**: Estimate the slope parameters.
- **Second step**: Estimate the intercept term.

The multiple linear regression model in deviation form is expressed as follows:

Let

\[ A = I - \frac{1}{n} \ell \ell' \]

where \( \ell = (1,1,...,1)' \) is a \( n \times 1 \) vector of each element unity. So

\[ A = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} - \frac{1}{n} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \]

Then

\[ \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} (1,1,...,1) \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \frac{1}{n} \ell' y \]

\[ Ay = y - \ell y = (y_1 - \bar{y}, y_2 - \bar{y}, ..., y_n - \bar{y})' \]
Thus pre-multiplication of any column vector by $A$ produces a vector showing those observations in deviation form:

Note that

$$A\ell = \ell - \frac{1}{n} \ell \ell' \ell$$

$$= \ell - \frac{1}{n} \ell n$$

$$= \ell - \ell$$

$$= 0$$

and $A$ is symmetric and idempotent matrix.

In the model

$$y = X\beta + \varepsilon,$$

the OLSE of $\beta$ is

$$b = \left(X'X\right)^{-1}X'y$$

and residual vector is

$$e = y - Xb.$$

Note that $Ae = e$.

If the $n \times k$ matrix is partitioned as

$$X = \begin{bmatrix} X_1 & X_2^* \end{bmatrix}$$

where $X_1 = (1,1,...,1)'$ is $n \times 1$ vector with all elements unity, $X_2^*$ is $n \times (k-1)$ matrix of observations of $(k-1)$ explanatory variables $X_2, X_3,..., X_k$ and OLSE $b = \begin{pmatrix} b_1 & b_2^* \end{pmatrix}$ is suitably partitioned with OLSE of intercept term $\beta_i$ as $b_1$ and $b_2$ as a $(k-1) \times 1$ vector of OLSEs associated with $\beta_2, \beta_3,..., \beta_k$.

Then

$$y = X_1b_1 + X_2^*b_2^* + e.$$
Premultiply by $A$,

$$Ay = AXb_1 + AX_2^*b_2^* + Ae$$

$$= AX_2^*b_2^* + e.$$  

Premultiply by $X_2^*$ gives

$$X_2^*Ay = X_2^*AX_2^*b_2^* + X_2^*e$$

$$= X_2^*AX_2^*b_2^*.$$  

Since $A$ is symmetric and idempotent,

$$\left(AX_2^*\right)'(Ay) = \left(AX_2^*\right)'(AX_2^*)b_2^*.$$  

This equation can be compared with the normal equations $X'y = X'Xb$ in the model $y = X\beta + \varepsilon$. Such a comparison yields the following conclusions:

- $b_2^*$ is the sub-vector of OLSE.
- $Ay$ is the study variables vector in deviation form.
- $AX_2^*$ is the explanatory variable matrix in deviation form.
- This is a normal equation in terms of deviations. Its solution gives OLS of slope coefficients as

$$b_2^* = \left[\left(AX_2^*\right)'(AX_2^*)\right]^{-1}\left(AX_2^*\right)'(Ay).$$

The estimate of intercept term is obtained in the second step as follows:

Premultiplying $y = Xb + e$ by $\frac{1}{n}\ell'$ gives

$$\frac{1}{n}\ell'y = \frac{1}{n}\ell'Xb + \frac{1}{n}\ell'e$$

$$\bar{y} = \begin{bmatrix} 1 & \bar{X}_2 & \bar{X}_3 & \ldots & \bar{X}_k \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} + 0$$

$$\Rightarrow b_1 = \bar{y} - b_2\bar{X}_2 - b_3\bar{X}_3 - \ldots - b_k\bar{X}_k.$$  

Now we explain various sums of squares in terms of this model.

The expression of total sum of squares (TSS) remains same as earlier and is given by

$$TSS = y'y.$$
Since
\[ Ay = AX^*b^*_2 + e \]
\[ y' Ay = y' AX^*b^*_2 + y'e \]
\[ = (Xb + e)' AX^*b^*_2 + y'e \]
\[ = (X_1^*b_1 + X_2^*b^*_2 + e)' AX^*b^*_2 + (X_1^*b_1 + X_2^*b^*_2 + e)'e \]
\[ = b^*_2' X^*_2' AX^*b^*_2 + e'e \]
\[ TSS = SS_{reg} + SS_{res} \]

where sum of squares due to regression is
\[ SS_{reg} = b^*_2' X^*_2' AX^*b^*_2 \]
and sum of squares due to residual is
\[ SS_{res} = e'e. \]

**Testing of hypothesis:**

There are several important questions which can be answered through the test of hypothesis concerning the regression coefficients. For example

1. What is the overall adequacy of the model?
2. Which specific explanatory variables seems to be important?

etc.

In order the answer such questions, we first develop the test of hypothesis for a general framework, viz., general linear hypothesis. Then several tests of hypothesis can be derived as its special cases. So first we discuss the test of a general linear hypothesis.

**Test of hypothesis for** \( H_0 : R\beta = r \)

We consider a general linear hypothesis that the parameters in \( \beta \) are contained in a subspace of parameter space for which \( R\beta = r \), where \( R \) is \((J \times k)\) matrix of known elements and \( r \) is a \((J \times 1)\) vector of known elements.

In general, the null hypothesis
\[ H_0 : R\beta = r \]

is termed as general linear hypothesis and
\[ H_1 : R\beta \neq r \]

is the alternative hypothesis.

We assume that \( \text{rank}(R) = J \), i.e., full rank so that there is no linear dependence in the hypothesis.

Some special cases and interesting example of \( H_0 : R\beta = r \) are as follows:

(i) \( H_0 : \beta_i = 0 \)

Choose \( J = 1, r = 0, R = [0, 0, \ldots, 0, 1, 0, \ldots, 0] \) where 1 occurs at the \( i^{th} \) position is \( R \).

This particular hypothesis explains whether \( X_i \) has any effect on the linear model or not.

(ii) \( H_0 : \beta_3 = \beta_4 \) or \( H_0 : \beta_3 - \beta_4 = 0 \)

Choose \( J = 1, r = 0, R = [0, 0, 1, 0, \ldots, 0] \)

(iii) \( H_0 : \beta_3 = \beta_4 = \beta_5 \)

or \( H_0 : \beta_3 - \beta_4 = 0, \beta_3 - \beta_5 = 0 \)

Choose \( J = 2, r = (0, 0)', R = \begin{bmatrix} 0 & 0 & 1 & -1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 & \ldots & 0 \end{bmatrix} \).

(iv) \( H_0 : \beta_3 + 5\beta_4 = 2 \)

Choose \( J = 1, r = 2, R = [0, 0, 1, 0, \ldots, 0] \)

(v) \( H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0 \)

\( J = k-1 \)

\( r = (0, 0, \ldots, 0)' \)

\[ R = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}_{(k-1) \times k} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{k-1} \times \begin{bmatrix} I_{k-1} \\ 0 \end{bmatrix} \]

This particular hypothesis explains the goodness of fit. It tells whether \( \beta_i \) has linear effect or not and are they of any importance. It also tests that \( X_2, X_3, \ldots, X_k \) have no influence in the determination of \( y \). Here \( \beta_i = 0 \) is excluded because this involves additional implication that the mean level of \( y \) is zero. Our main concern is to know whether the explanatory variables helps to explain the variation in \( y \) around its mean value or not.
We develop the likelihood ratio test for $H_0 : R\beta = r$.

**Likelihood ratio test:**

The likelihood ratio test statistic is

$$
\hat{\lambda} = \frac{\max L(\beta, \sigma^2 | y, X)}{\max L(\beta, \sigma^2 | y, X, R\beta = r)} = \frac{\hat{L}(\Omega)}{\hat{L}(\omega)}
$$

where $\Omega$ is the whole parametric space and $\omega$ is the sample space.

If both the likelihood are maximized, one constrained and the other unconstrained, then the value of the unconstrained will not be smaller than the value of the constrained. Hence $\lambda \geq 1$.

First we discuss the likelihood ratio test for a simpler case when $R = I_k$ and $r = \beta_0$, i.e., $\beta = \beta_0$. This will give us a better and detailed understanding for the minor details and then we generalize it for $R\beta = r$, in general.

**Likelihood ratio test for $H_0 : \beta = \beta_0$**

Let the null hypothesis related to $k \times 1$ vector $\beta$ is

$$H_0 : \beta = \beta_0$$

where $\beta_0$ is specified by the investigator. The elements of $\beta_0$ can take on any value, including zero. The concerned alternative hypothesis is

$$H_1 : \beta \neq \beta_0.$$  

Since $\varepsilon \sim N(0, \sigma^2 I)$ in $y = X\beta + \varepsilon$, so $y \sim N(X\beta, \sigma^2 I)$. Thus the whole parametric space and sample space are $\Omega$ and $\omega$ respectively given by

$$\Omega : \{(\beta, \sigma^2) : -\infty < \beta_i < \infty, \sigma^2 > 0, i = 1, 2, ..., k\}$$

$$\omega : \{(\beta, \sigma^2) : \beta = \beta_0, \sigma^2 > 0\}.$$

The unconstrained likelihood under $\Omega$.

$$L(\beta, \sigma^2 | y, X) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left[ -\frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta) \right].$$

---

*Econometrics* | Chapter 3 | Multiple Linear Regression Model | *Shalabh, IIT Kanpur*
This is maximized over \( \Omega \) when
\[
\hat{\beta} = (X'X)^{-1}X'y
\]
\[
\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})'(y - X\hat{\beta}).
\]
where \( \hat{\beta} \) and \( \hat{\sigma}^2 \) are the maximum likelihood estimates of \( \beta \) and \( \sigma^2 \) which are the values maximizing the likelihood function.

\[
\hat{L}(\Omega) = \max L(\beta, \sigma^2 \mid y, X)\]
\[
= \frac{1}{\left[\frac{2\pi}{n}(y - X\hat{\beta})'(y - X\hat{\beta})\right]^{n/2}} \exp\left[-\frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{2\left[(y - X\hat{\beta})'(y - X\hat{\beta})\right]/n}\right]
\]
\[
= \frac{n^{n/2} \exp\left(-\frac{n}{2}\right)}{(2\pi)^{n/2} \left[(y - X\hat{\beta})'(y - X\hat{\beta})\right]^{n/2}}.
\]
The constrained likelihood under \( \omega \) is
\[
\hat{L}(\omega) = \max L(\beta, \sigma^2 \mid y, X, \beta = \beta_0) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} (y - X\beta_0)'(y - X\beta_0)\right].
\]
Since \( \beta_0 \) is known, so the constrained likelihood function has an optimum variance estimator
\[
\hat{\sigma}^2_{\omega} = \frac{1}{n}(y - X\beta_0)'(y - X\beta_0)
\]
\[
\hat{L}(\omega) = \frac{n^{n/2} \exp\left(-\frac{n}{2}\right)}{(2\pi)^{n/2} \left[(y - X\beta_0)'(y - X\beta_0)\right]^{n/2}}.
\]
The likelihood ratio is
\[
\hat{L}(\Omega) = \frac{n^{n/2} \exp(-n / 2)}{(2\pi)^{n/2} \left[(y - X\hat{\beta})'(y - X\hat{\beta})\right]^{n/2}}\]
\[
= \frac{n^{n/2} \exp(-n / 2)}{(2\pi)^{n/2} \left[(y - X\hat{\beta}_0)'(y - X\hat{\beta}_0)\right]^{n/2}}\]
\[
= \left[\frac{(y - X\beta_0)'(y - X\beta_0)}{(y - X\hat{\beta})'(y - X\hat{\beta})}\right]^{n/2}
\]
\[
= \left[\frac{\hat{\sigma}^2_{\omega}}{\hat{\sigma}^2}\right]^{n/2} = (\lambda)^{n/2}
\]
where
\[ \lambda = \frac{(y - X\beta_0)'(y - X\beta_0)}{(y - X\beta)'(y - X\tilde{\beta})} \]

is the ratio of the quadratic forms. Now we simplify the numerator in \( \lambda \) as follows:
\[
(y - X\beta_0)'(y - X\beta_0) = [(y - X\tilde{\beta}) + X(\tilde{\beta} - \beta_0)] [(y - X\tilde{\beta}) + X(\tilde{\beta} - \beta_0)]
\]
\[
= (y - X\tilde{\beta})'(y - X\tilde{\beta}) + 2y'[I - X(X'X)^{-1}X']X(\tilde{\beta} - \beta_0) + (\tilde{\beta} - \beta_0)'X'X(\tilde{\beta} - \beta_0)
\]
\[
= (y - X\tilde{\beta})'(y - X\tilde{\beta}) + (\tilde{\beta} - \beta_0)'X'X(\tilde{\beta} - \beta_0).
\]

Thus
\[
\lambda = \frac{(y - X\tilde{\beta})'(y - X\tilde{\beta}) + (\tilde{\beta} - \beta_0)'X'X(\tilde{\beta} - \beta_0)}{(y - X\tilde{\beta})'(y - X\tilde{\beta})}
\]
\[
= 1 + \frac{(\tilde{\beta} - \beta_0)'X'X(\tilde{\beta} - \beta_0)}{(y - X\tilde{\beta})'(y - X\tilde{\beta})}
\]
or \[
\lambda - 1 = \frac{(\tilde{\beta} - \beta_0)'X'X(\tilde{\beta} - \beta_0)}{(y - X\tilde{\beta})'(y - X\tilde{\beta})}
\]

where \( 0 \leq \lambda_0 < \infty \).

**Distribution of ratio of quadratic forms**

Now we find the distribution of the quadratic forms involved is \( \lambda_0 \) to find the distribution of \( \lambda_0 \) as follows:
\[
(y - X\tilde{\beta})'(y - X\tilde{\beta}) = \tilde{\epsilon}'\tilde{\epsilon}
\]
\[
= y'[I - X(X'X)^{-1}X']Y
\]
\[
= y'\bar{H}Y
\]
\[
= (X\beta + \epsilon)'\bar{H}(X\beta + \epsilon)
\]
\[
= \epsilon'\bar{H}\epsilon \quad \text{(using } \bar{H}X = 0) \]
\[
= (n-k)\hat{\sigma}^2
\]

**Result:** If \( Z \) is a \( n \times 1 \) random vector that is distributed as \( N(0,\sigma^2I_n) \) and \( A \) is any symmetric idempotent \( n \times n \) matrix of rank \( p \) then \( \frac{Z'AZ}{\sigma^2} \sim \chi^2(p) \). If \( B \) is another \( n \times n \) symmetric idempotent matrix of rank \( q \), then \( \frac{Z'BZ}{\sigma^2} \sim \chi^2(q) \). If \( AB = 0 \) then \( Z'AZ \) is distributed independently of \( Z'BZ \).
So using this result, we have
\[
\frac{y'y_Hy}{\sigma^2} = \frac{(n-k)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-k).
\]

Further, if \( H_0 \) is true, then \( \beta = \beta_0 \) and we have the numerator in \( \lambda_0 \). Rewriting the numerator in \( \lambda_0 \), in general, we have
\[
(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) = \varepsilon'X(X'X)^{-1}X'X(X'X)^{-1}X'\varepsilon = \varepsilon'X(X'X)^{-1}X'\varepsilon = \varepsilon'H\varepsilon
\]
where \( H \) is an idempotent matrix with rank \( k \).

Thus using this result, we have
\[
\frac{\varepsilon'H\varepsilon}{\sigma^2} = \frac{\varepsilon'X(X'X)^{-1}X'\varepsilon}{\sigma^2} \sim \chi^2(k).
\]

Furthermore, the product of the quadratic form matrices in the numerator \( (\varepsilon'\hat{H}\varepsilon) \) and denominator \( (\varepsilon'H\varepsilon) \) of \( \lambda_0 \) is
\[
\left[I - X(X'X)^{-1}X'\right]X(X'X)^{-1}X' = X(X'X)^{-1}X' - X(X'X)^{-1}X'X(X'X)^{-1}X' = 0
\]
and hence the \( \chi^2 \) random variables in numerator and denominator of \( \lambda_0 \) are independent. Dividing each of the \( \chi^2 \) random variable by their respective degrees of freedom
\[
\lambda_0 = \left(\frac{(\hat{\beta} - \beta_0)'X'X(\hat{\beta} - \beta_0)}{\frac{(n-k)\hat{\sigma}^2}{\sigma^2} \frac{k}{n-k}}\right)
\]
\[
= \frac{(\hat{\beta} - \beta_0)'X'X(\hat{\beta} - \beta_0)}{k\hat{\sigma}^2}
\]
\[
= \frac{(y - X\beta_0)(y - X\beta_0) - (y - X\hat{\beta})(y - X\hat{\beta})}{k\hat{\sigma}^2}
\]
\[
\sim F(k, n-k) \text{ under } H_0.
\]

Note that
\[
(y - X\beta_0)'(y - X\beta_0) : \text{Restricted error sum of squares}
\]
\[
(y - X\hat{\beta})'(y - X\hat{\beta}) : \text{Unrestricted error sum of squares}
\]

**Econometrics** | Chapter 3 | Multiple Linear Regression Model | Shalabh, IIT Kanpur
Numerator in $\lambda_i$: Difference between the restricted and unrestricted error sum of squares.

The decision rule is to reject $H_0: \beta = \beta_0$ at $\alpha$ level of significance whenever

$$\lambda_i \geq F_{a}(k,n-k)$$

where $F_{a}(k,n-k)$ is the upper critical points on the central $F$-distribution with $k$ and $n-k$ degrees of freedom.

**Likelihood ratio test for $H_0: R\beta = r$**

The same logic and reasons used in the development of likelihood ratio test for $H_0: \beta = \beta_0$ can be extended to develop the likelihood ratio test for $H_0: R\beta = r$ as follows.

$$\Omega = \{(\beta,\sigma^2): -\infty < \beta_i < \infty, \sigma^2 > 0, i = 1,2,...,k\}$$

$$\omega = \{(\beta,\sigma^2): -\infty < \beta_i < \infty, R\beta = r, \sigma^2 > 0\}.$$

Let $\tilde{\beta} = (X'X)^{-1}X'y$.

Then

$$E(R\tilde{\beta}) = R\beta$$

$$V(R\tilde{\beta}) = E\left[R(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'R'\right]$$

$$= RV(\tilde{\beta})R'$$

$$= \sigma^2 R(X'X)^{-1}R'.$$

Since $\tilde{\beta} \sim N\left[\beta, \sigma^2 (X'X)^{-1}\right]$ so

$$R\tilde{\beta} \sim N\left[R\beta, \sigma^2 R(X'X)^{-1}R'\right]$$

$$R\tilde{\beta} - r = R\tilde{\beta} - R\beta = R(\tilde{\beta} - \beta) \sim N\left[0, \sigma^2 R(X'X)^{-1}R'\right].$$

There exists a matrix $Q$ such that $R(X'X)^{-1}R' = QQ'$ and then

$$\xi = QR(b - \beta) \sim N(0, \sigma^2 I_n).$$

Therefore under $H_0: R\beta - r = 0$, so
\[
\frac{\xi' \xi}{\sigma^2} = \frac{(R\hat{\beta} - r)'QQ'(R\hat{\beta} - r)}{\sigma^2} \\
= \frac{(R\hat{\beta} - r)[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)}{\sigma^2} \\
= \frac{(\hat{\beta} - \beta)'R'[R(X'X)^{-1}R']^{-1}R(\hat{\beta} - \beta)}{\sigma^2} \\
= \frac{\varepsilon'X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'\varepsilon}{\sigma^2} \\
\sim \chi^2(J).
\]

which is obtained as \( X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X' \) is an idempotent matrix and its trace is \( J \) which is the associated degrees of freedom.

Also, irrespective of whether \( H_0 \) is true or not,
\[
\frac{\tilde{\varepsilon}'\tilde{\varepsilon}}{\sigma^2} = \frac{(y - X\hat{\beta})(y - X\hat{\beta})}{\sigma^2} = \frac{y'y_Hy}{\sigma^2} = \frac{(n-k)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-k).
\]

Moreover, the product of quadratic form matrices of \( \tilde{\varepsilon}'\tilde{\varepsilon} \) and
\( (\hat{\beta} - \beta)'R'[R(X'X)^{-1}R']^{-1}R(\hat{\beta} - \beta) \) is zero implying that both the quadratic forms are independent. So in terms of likelihood ratio test statistic
\[
\lambda = \left[ \left( \frac{(R\hat{\beta} - r)[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)}{\sigma^2} \right) \right]
\]
\[
= \left[ \left( \frac{(n-k)\hat{\sigma}^2}{\sigma^2} \right) \right]
\]
\[
= \left( R\hat{\beta} - r \right)[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)
\]
\[
\sim F(J, n-k) \text{ under } H_0.
\]

So the decision rule is to reject \( H_0 \) whenever
\[
\lambda \geq F_{\alpha}(J, n-k)
\]
where \( F_{\alpha}(J, n-k) \) is the upper critical points on the central \( F \) distribution with \( J \) and \( (n-k) \) degrees of freedom.
Test of significance of regression (Analysis of variance)

If we set \( R = [0 \ I_{k-1}] \), \( r = 0 \), then the hypothesis \( H_0 : R \beta = r \) reduces to the following null hypothesis:

\[
H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0
\]

against the alternative hypothesis

\[
H_1 : \beta_j \neq 0 \text{ for at least one } j = 2, 3, \ldots, k
\]

This hypothesis determines if there is a linear relationship between \( y \) and any set of the explanatory variables \( X_2, X_3, \ldots, X_k \). Notice that \( X_1 \) corresponds to the intercept term in the model and hence \( x_{i1} = 1 \) for all \( i = 1, 2, \ldots, n \).

This is an overall or global test of model adequacy. Rejection of the null hypothesis indicates that at least one of the explanatory variables among \( X_2, X_3, \ldots, X_k \) contributes significantly to the model. This is called as analysis of variance.

Since \( \varepsilon \sim N(0, \sigma^2 I) \),

so \( y \sim N(X \beta, \sigma^2 I) \)

\[
b = (X'X)^{-1}X'y \sim N\left[ \beta, \sigma^2 (X'X)^{-1} \right].
\]

Also \( \hat{\sigma}^2 = \frac{SS_{\text{res}}}{n-k} \)

\[
= \frac{(y - \hat{y})(y - \hat{y})}{n-k}
= \frac{y'[(I - X(X'X)^{-1}X')y}{n-k} = \frac{y'X'y - b'X'y}{n-k}
\]

Since \( (X'X)^{-1}X'H = 0 \), so \( b \) and \( \hat{\sigma}^2 \) are independently distributed.

Since \( y'Hy = \varepsilon'H\varepsilon \) and \( H \) is an idempotent matrix, so

\[
SS_{\text{res}} \sim \chi^2_{(n-k)},
\]

i.e., central \( \chi^2 \) distribution with \( n-k \) degrees of freedom.

Partition \( X = [X_1, X_2'] \) where the submatrix \( X_2' \) contains the explanatory variables \( X_2, X_3, \ldots, X_k \) and partition \( \beta = [\beta_1, \beta_2'] \) where the subvector \( \beta_2' \) contains the regression coefficients \( \beta_2, \beta_3, \ldots, \beta_k \).
Now partition the total sum of squares due to \( y ' s \) as

\[
SS_T = y ' Ay
\]
\[
= SS_{\text{reg}} + SS_{\text{res}}
\]

where \( SS_{\text{reg}} = b_2^* X_2^* A X_2^* b_2^* \) is the sum of squares due to regression and the sum of squares due to residuals is given by

\[
SS_{\text{res}} = (y - Xb)'(y - Xb)
\]
\[
= y ' \bar{H} y
\]
\[
= SS_T - SS_{\text{reg}}.
\]

Further

\[
\frac{SS_{\text{reg}}}{\sigma^2} \sim \chi^2_{k-1} \left( \frac{\beta_2^* X_2^* A X_2^* \beta_2^*}{2\sigma^2} \right), \text{ i.e., non-central } \chi^2 \text{ distribution with non-centrality parameter } \frac{\beta_2^* X_2^* A X_2^* \beta_2^*}{2\sigma^2},
\]

\[
\frac{SS_T}{\sigma^2} \sim \chi^2_{n-1} \left( \frac{\beta_2^* X_2^* A X_2^* \beta_2^*}{2\sigma^2} \right), \text{ i.e., non-central } \chi^2 \text{ distribution with non-centrality parameter } \frac{\beta_2^* X_2^* A X_2^* \beta_2^*}{2\sigma^2}.
\]

Since \( X_2 \bar{H} = 0 \), so \( SS_{\text{reg}} \) and \( SS_{\text{res}} \) are independently distributed. The mean squares due to regression is

\[
MS_{\text{reg}} = \frac{SS_{\text{reg}}}{k-1}
\]

and the mean square due to error is

\[
MS_{\text{res}} = \frac{SS_{\text{res}}}{n-k}.
\]

Then

\[
\frac{MS_{\text{reg}}}{MS_{\text{res}}} \sim F_{k-1,n-k} \left( \frac{\beta_2^* X_2^* A X_2^* \beta_2^*}{2\sigma^2} \right)
\]

which is a non-central \( F \)-distribution with \((k-1, n-k)\) degrees of freedom and noncentrality parameter

\[
\frac{\beta_2^* X_2^* A X_2^* \beta_2^*}{2\sigma^2}.
\]

Under \( H_0 : \beta_2 = \beta_3 = \ldots = \beta_k \),

\[
F = \frac{MS_{\text{reg}}}{MS_{\text{res}}} \sim F_{k-1,n-k}.
\]

The decision rule is to reject at \( \alpha \) level of significance whenever

\[
F \geq F_{\alpha} (k-1, n-k).
\]
The calculation of \( F \)-statistic can be summarized in the form of an analysis of variance (ANOVA) table given as follows:

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean squares</th>
<th>( F )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>( SS_{reg} )</td>
<td>( k - 1 )</td>
<td>( MS_{reg} = SS_{reg} / k - 1 )</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>( SS_{res} )</td>
<td>( n - k )</td>
<td>( MS_{res} = SS_{res} / (n - k) )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( SS_{T} )</td>
<td>( n - 1 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Rejection of \( H_0 \) indicates that it is likely that at least one \( \beta_i \neq 0 \) \( (i = 1, 2, ..., k) \).

**Test of hypothesis on individual regression coefficients**

In case, if the test in analysis of variance is rejected, then another question arises is that which of the regression coefficients is/are responsible for the rejection of null hypothesis. The explanatory variables corresponding to such regression coefficients are important for the model.

Adding such explanatory variables also increases the variance of fitted values \( \hat{y} \), so one need to be careful that only those regressors are added that are of real value in explaining the response. Adding unimportant explanatory variables may increase the residual mean square which may decrease the usefulness of the model.

To test the null hypothesis

\[ H_0 : \beta_j = 0 \]

versus the alternative hypothesis

\[ H_1 : \beta_j \neq 0 \]

has already been discussed is the case of simple linear regression model. In present case, if \( H_0 \) is accepted, it implies that the explanatory variable \( X_j \) can be deleted from the model. The corresponding test statistic is

\[ t = \frac{b_j}{se(b_j)} \sim t(n - k - 1) \quad \text{under} \quad H_0 \]

where the standard error of OLS \( b_j \) of \( \beta_j \) is

\[ se(b_j) = \sqrt{\hat{\sigma}^2 C_{jj}} \quad \text{where} \quad C_{jj} \text{ denotes the } j^{th} \text{ diagonal element of } (X'X)^{-1} \text{ corresponding to } b_j. \]
The decision rule is to reject $H_0$ at $\alpha$ level of significance if

$$|t| > t_{\alpha, n-k-1}.$$

Note that this is only a partial or marginal test because $\hat{\beta}_j$ depends on all the other explanatory variables $X_i (i \neq j)$ that are in the model. This is a test of the contribution of $X_j$ given the other explanatory variables in the model.

**Confidence interval estimation**

The confidence intervals in multiple regression model can be constructed for individual regression coefficients as well as jointly. We consider both of them as follows:

**Confidence interval on the individual regression coefficient:**

Assuming $\varepsilon_i$'s are identically and independently distributed following $N(0, \sigma^2)$ in $y = X\beta + \varepsilon$, we have

$$y \sim N(X\beta, \sigma^2 I)$$

$$b \sim N(\beta, \sigma^2 (X'X)^{-1}).$$

Thus the marginal distribution of any regression coefficient estimate

$$b_j \sim N(\beta_j, \sigma^2 C_{jj})$$

where $C_{jj}$ is the $j^{th}$ diagonal element of $(X'X)^{-1}$.

Thus

$$t_j = \frac{b_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t(n-k) \text{ under } H_0, \ j = 1, 2, \ldots$$

where $\hat{\sigma}^2 = \frac{SS_{\text{res}}}{n-k} = \frac{y'y - b'X'y}{n-k}$.

So the $100(1-\alpha)$% confidence interval for $\beta_j (j = 1, 2, \ldots, k)$ is obtained as follows:

$$P \left[ -t_{\alpha, n-k} \leq \frac{b_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \leq t_{\alpha, n-k} \right] = 1 - \alpha$$

$$P \left[ b_j - t_{\alpha, n-k} \sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq b_j + t_{\alpha, n-k} \sqrt{\hat{\sigma}^2 C_{jj}} \right] = 1 - \alpha.$$

So the confidence interval is

$$\left( b_j - t_{\alpha, n-k} \sqrt{\hat{\sigma}^2 C_{jj}}, b_j + t_{\alpha, n-k} \sqrt{\hat{\sigma}^2 C_{jj}} \right).$$
Simultaneous confidence intervals on regression coefficients:

A set of confidence intervals that are true simultaneously with probability \((1 - \alpha)\) are called simultaneous or joint confidence intervals.

It is relatively easy to define a joint confidence region for \(\beta\) in multiple regression model.

Since

\[
\frac{(b - \beta)'X'X(b - \beta)}{k MS_{res}} \sim F_{k,n-k}
\]

\[
\Rightarrow P \left[ \frac{(b - \beta)'X'X(b - \beta)}{k MS_{res}} \leq F_{a}(k,n-k) \right] = 1 - \alpha.
\]

So a 100\((1 - \alpha)\)% joint confidence region for all of the parameters in \(\beta\) is

\[
\frac{(b - \beta)'X'X(b - \beta)}{k MS_{res}} \leq F_{a}(k,n-k)
\]

which describes an elliptically shaped region.

Coefficient of determination \((R^2)\) and adjusted \(R^2\)

Let \(R\) be the multiple correlation coefficient between \(y\) and \(X_1, X_2, \ldots, X_k\). Then square of multiple correlation coefficient \((R^2)\) is called as coefficient of determination. The value of \(R^2\) commonly describes that how well the sample regression line fits to the observed data. This is also treated as a measure of goodness of fit of the model.

Assuming that the intercept term is present in the model as

\[
y_i = \beta_1 + \beta_2X_{i2} + \beta_3X_{i3} + \ldots + \beta_kX_{ik} + u_i, i = 1, 2, \ldots, n
\]

then

\[
R^2 = 1 - \frac{e'e}{\sum_{i=1}^{n} (y_i - \bar{y})^2}
\]

\[
= 1 - \frac{SS_{res}}{SS_{T}}
\]

\[
= \frac{SS_{reg}}{SS_{T}}
\]

where
$SS_{res}$ : sum of squares due to residuals,

$SS_T$ : total sum of squares

$SS_{reg}$ : sum of squares due to regression.

$R^2$ measure the explanatory power of the model which in turn reflects the goodness of fit of the model. It reflects the model adequacy in the sense that how much is the explanatory power of explanatory variable.

Since

$$e'e = y'(I - X(X'X)^{-1}X')y = y'y_Hy,$$

$$
\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2,
$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \ell'y$ with $\ell = (1,1,...,1)'$, $y = (y_1, y_2, ..., y_n)'$

Thus

$$
\sum_{i=1}^{n} (y_i - \bar{y})^2 = y'y - n \left( \frac{1}{n} \ell'y \ell'y \right)
\quad = y'y - y'y \frac{1}{n} \ell'y
\quad = y'y - y'y(\ell'\ell)^{-1} \ell'y
\quad = y'y[I - \ell(\ell'\ell)^{-1}\ell']y
\quad = y'yAy
$$

where $A = I - \ell(\ell'\ell)^{-1}\ell'$.

So $R^2 = 1 - \frac{y'y_Hy}{y'yAy}$.

The limits of $R^2$ are 0 and 1, i.e.,

$$0 \leq R^2 \leq 1.$$

$R^2 = 0$ indicates the poorest fit of the model.

$R^2 = 1$ indicates the best fit of the model.

$R^2 = 0.95$ indicates that 95% of the variation in $y$ is explained by $R^2$. In simple words, the model is 95% good.

Similarly any other value of $R^2$ between 0 and 1 indicates the adequacy of fitted model.
Adjusted $R^2$

If more explanatory variables are added to the model, then $R^2$ increases. In case the variables are irrelevant, then $R^2$ will still increase and gives an overly optimistic picture.

With a purpose of correction in overly optimistic picture, adjusted $R^2$, denoted as $\bar{R}^2$ or adj $R^2$ is used which is defined as

$$\bar{R}^2 = 1 - \frac{SS_{res} / (n-k)}{SS_r / (n-1)}$$

$$= 1 - \left( \frac{n-1}{n-k} \right)(1 - R^2).$$

We will see later that $(n-k)$ and $(n-1)$ are the degrees of freedom associated with the distributions of $SS_{res}$ and $SS_r$. Moreover, the quantities $\frac{SS_{res}}{n-k}$ and $\frac{SS_r}{n-1}$ are based on the unbiased estimators of respective variances of $e$ and $y$ is the context of analysis of variance.

The adjusted $R^2$ will decline if the addition if an extra variable produces too small a reduction in $(1 - R^2)$ to compensate for the increase is $\left( \frac{n-1}{n-k} \right)$.

Another limitation of adjusted $R^2$ is that it can be negative also. For example if $k = 3, n = 10, R^2 = 0.16$, then

$$\bar{R}^2 = 1 - \frac{9}{7} \times 0.97 = -0.25 < 0$$

which has no interpretation.
Limitations

1. If constant term is absent in the model, then $R^2$ can not be defined. In such cases, $R^2$ can be negative. Some ad-hoc measures based on $R^2$ for regression line through origin have been proposed in the literature.

**Reason that why $R^2$ is valid only in linear models with intercept term:**

In the model $y = X\beta + \varepsilon$, the ordinary least squares estimator of $\beta$ is $b = (X'X)^{-1}X'y$. Consider the fitted model as

$$y = Xb + (y - Xb)$$

$$= Xb + e$$

where $e$ is the residual. Note that

$$y - \bar{y} = Xb + e - \bar{y}$$

$$= \hat{y} + e - \bar{y}$$

where $\hat{y} = Xb$ is the fitted value and $l = (1,1,\ldots,1)'$ is a $n \times 1$ vector of elements unity. The total sum of squares $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$ is then obtained as

$$TSS = (y - \bar{y})(y - \bar{y}) = [(\hat{y} - \bar{y}) + e] [(\hat{y} - \bar{y}) + e]$$

$$= (\hat{y} - \bar{y})(\hat{y} - \bar{y}) + e'e + 2(\hat{y} - \bar{y})'e$$

$$\downarrow \quad \downarrow \quad \downarrow$$

$$= SS_{reg} + SS_{res} + 2(Xb - \bar{y})'e \quad \text{(because $\hat{y} = Xb$)}$$

$$= SS_{reg} + SS_{res} - 2\bar{y}l'e \quad \text{(because $X'e = 0$).}$$

The Fisher Cochran theorem requires $TSS = SS_{reg} + SS_{res}$ to hold true in the context of analysis of variance and further to define the $R^2$. In order that $TSS = SS_{reg} + SS_{res}$ holds true, we need that $l'e$ should be zero, i.e. $l'e = l'(y - \hat{y}) = 0$ which is possible only when there is an intercept term in the model. We show this claim as follows:

First we consider a no intercept simple linear regression model $y_i = \beta_1 x_i + \varepsilon_i$, $(i = 1, 2, \ldots, n)$ where the parameter $\beta_1$ is estimated as $b_1^* = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$. Then $l'e = \sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - b^*_1 x_i) \neq 0$, in general.
Similarly, in a no intercept multiple linear regression model \( y = X \beta + \varepsilon \), we find that
\[
l'(y - \hat{y}) = l'(X \beta + \varepsilon - Xb) = -l' (b - \beta) + l' \varepsilon = 0, \text{ in general.}
\]

Next we consider a simple linear regression model with intercept term \( y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \), \( (i = 1, 2, ..., n) \) where the parameters \( \beta_0 \) and \( \beta_1 \) are estimated as \( b_0 = \bar{y} - b_1 \bar{x} \) and \( b_1 = \frac{s_{xy}}{s_{xx}} \) respectively, where
\[
s_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \quad s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i. \quad \text{We find that}
\]
\[
l'(y - \hat{y}) = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)
\]
\[
= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)
\]
\[
= \sum_{i=1}^{n} (y_i - \bar{y} + b_1 \bar{x} - b_1 x_i)
\]
\[
= \sum_{i=1}^{n} [(y_i - \bar{y}) - b_1 (x_i - \bar{x})]
\]
\[
= \sum_{i=1}^{n} (y_i - \bar{y}) - b_1 \sum_{i=1}^{n} (x_i - \bar{x})
\]
\[
= 0.
\]

In a multiple linear regression model with an intercept term \( y = \beta_0 + X \beta + \varepsilon \) where the parameters \( \beta_0 \) and \( \beta \) are estimated as \( \hat{\beta}_0 = \bar{y} - \bar{x} \bar{x} \) and \( b = (X'X)^{-1} X'y \), respectively. We find that
\[
l'(y - \hat{y}) = l'(y - \hat{y})
\]
\[
= l'(y - \hat{\beta}_0 - Xb)
\]
\[
= l'(y - \bar{y} + \bar{x}b - Xb)
\]
\[
= l'(y - \bar{y}) + l'(X - \bar{x})b
\]
\[
= 0.
\]

Thus we conclude that for the Fisher Cochran to hold true in the sense that the total sum of squares can be divided into two orthogonal components, viz., sum of squares due to regression and sum of squares due to errors, it is necessary that \( l'e = l'(y - \hat{y}) = 0 \) holds and which is possible only when the intercept term is present in the model.
2. $R^2$ is sensitive to extreme values, so $R^2$ lacks robustness.

3. $R^2$ always increases with an increase in the number of explanatory variables in the model. The main drawback of this property is that even when the irrelevant explanatory variables are added in the model, $R^2$ still increases. This indicates that the model is getting better which is not really correct.

4. Consider a situation where we have following two models:

$$y_i = \beta_1 + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + u_i, \quad i = 1, 2, \ldots, n$$

$$\log y_i = \gamma_1 + \gamma_2 X_{i2} + \ldots + \gamma_k X_{ik} + v_i$$

The question is now which model is better?

For the first model,

$$R^2_1 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

and for the second model, an option is to define $R^2$ as

$$R^2_2 = 1 - \frac{\sum_{i=1}^{n} (\log y_i - \log \hat{y}_i)^2}{\sum_{i=1}^{n} (\log y_i - \log \bar{y})^2}$$

As such $R^2_1$ and $R^2_2$ are not comparable. If still, the two models are needed to be compared, a better proposition to define $R^2$ can be as follows:

$$R^2_3 = 1 - \frac{\sum_{i=1}^{n} (y_i - \text{anti log } \hat{y}_i^*)}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where $y_i^* = \log y_i$. Now $R^2_1$ and $R^2_3$ on comparison may give an idea about the adequacy of the two models.
Relationship of analysis of variance test and coefficient of determination

Assuming the $\beta_i$ to be an intercept term, then for $H_0: \beta_2 = \beta_3 = \ldots = \beta_k = 0$, the $F$-statistic in analysis of variance test is

$$F = \frac{MS_{reg}}{MS_{res}}$$

$$= \frac{(n-k)SS_{reg}}{(k-1)SS_{res}}$$

$$= \left(\frac{n-k}{k-1}\right)\frac{SS_{reg}}{SS_T} - SS_{reg}$$

$$= \left(\frac{n-k}{k-1}\right)\frac{SS_{reg}}{SS_T}$$

$$= \left(\frac{n-k}{k-1}\right)R^2$$

where $R^2$ is the coefficient of determination. So $F$ and $R^2$ are closely related. When $R^2 = 0$, then $F = 0$.

In limit, when $R^2 = 1, F = \infty$. So both $F$ and $R^2$ vary directly. Larger $R^2$ implies greater $F$ value. That is why the $F$ test under analysis of variance is termed as the measure of overall significance of estimated regression. It is also a test of significance of $R^2$. If $F$ is highly significant, it implies that we can reject $H_0$, i.e. $y$ is linearly related to $X'$s.

Prediction of values of study variable

The prediction in multiple regression model has two aspects

1. Prediction of average value of study variable or mean response.
2. Prediction of actual value of study variable.

1. Prediction of average value of $y$

We need to predict $E(y)$ at a given $x_0 = (x_{01}, x_{02}, \ldots, x_{0k})'$.

The predictor as a point estimate is

$$p = x_0'b = x_0'(X'X)^{-1}X'y$$

$$E(p) = x_0\beta.$$

So $p$ is an unbiased predictor for $E(y)$. 
Its variance is

\[ \text{Var}(p) = E[(p - E(y))(p - E(y))] = \sigma^2 x_0'(X'X)^{-1}x_0. \]

Then

\[ E(\hat{y}_0) = x_0\beta = E(y/x_0) \]

\[ \text{Var}(\hat{y}_0) = \sigma^2 x_0'(X'X)^{-1}x_0 \]

The confidence interval on the mean response at a particular point, such as \( x_{01}, x_{02}, \ldots, x_{0k} \) can be found as follows:

Define \( x_0 = (x_{01}, x_{02}, \ldots, x_{0k})' \). The fitted value at \( x_0 \) is \( \hat{y}_0 = x_0\beta \).

Then

\[ P \left[ t_{\frac{\alpha}{2},n-k} \leq \frac{\hat{y}_0 - E(y/x_0)}{\sqrt{\text{Var}(\hat{y}_0)}} \leq t_{\frac{\alpha}{2},n-k} \right] = 1 - \alpha \]

\[ P \left[ \hat{y}_0 - t_{\frac{\alpha}{2},n-k} \sqrt{\text{Var}(\hat{y}_0)} \leq E(y/x_0) \leq \hat{y}_0 + t_{\frac{\alpha}{2},n-k} \sqrt{\text{Var}(\hat{y}_0)} \right] = 1 - \alpha. \]

The 100(1 - \( \alpha \))% confidence interval on the mean response at the point \( x_{01}, x_{02}, \ldots, x_{0k} \), i.e., \( E(y/x_0) \) is

\[ \left[ \hat{y}_0 - t_{\frac{\alpha}{2},n-k} \sqrt{\text{Var}(\hat{y}_0)} \, \hat{y}_0 + t_{\frac{\alpha}{2},n-k} \sqrt{\text{Var}(\hat{y}_0)} \right]. \]

2. Prediction of actual value of \( y \)

We need to predict \( y \) at a given \( x_0 = (x_{01}, x_{02}, \ldots, x_{0k})' \).

The predictor as a point estimate is

\[ p_f = x_0\beta \]

\[ E(p_f) = x_0\beta \]

So \( p_f \) is an unbiased for \( y \). It's variance is

\[ \text{Var}(p_f) = E((p_f - y)(p_f - y)') = \sigma^2 [1 + x_0'(X'X)^{-1}x_0]. \]

The 100(1 - \( \alpha \))% confidence interval for this future observation is

\[ \left[ p_f - t_{\frac{\alpha}{2},n-k} \sqrt{\text{Var}(p_f)}, \, p_f + t_{\frac{\alpha}{2},n-k} \sqrt{\text{Var}(p_f)} \right]. \]
**Chapter 4**

**Predictions In Linear Regression Model**

**Prediction of values of study variable**

An important use of linear regression modeling is to predict the average and actual values of study variable. The term prediction of value of study variable corresponds to knowing the value of $E(y)$ (in case of average value) and value of $y$ (in case of actual value) for a given value of explanatory variable.

We consider both the cases. The prediction of values consists of two steps. In the first step, the regression coefficients are estimated on the basis of given observations. In the second step, these estimators are then used to construct the predictor which provides the prediction of actual or average values of study variables. Based on this approach of construction of predictors, there are two situations in which the actual and average values of study variable can be predicted: within sample prediction and outside sample prediction. We describe the prediction in both the situations.

**Within sample prediction in simple linear regression model**

Consider the linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$. Based on a sample of $n$ sets of paired observations $(x_i, y_i)$ $(i = 1, 2, \ldots, n)$ following $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $\varepsilon_i$’s are identically and independently distributed following $N(0, \sigma^2)$. The parameters $\beta_0$ and $\beta_1$ are estimated using the ordinary least squares estimation as $b_0$ of $\beta_0$ and $b_1$ of $\beta_1$ as

$$
\begin{align*}
\hat{b}_0 &= \bar{y} - \hat{b}_1 \bar{x} \\
\hat{b}_1 &= \frac{s_{xy}}{s_{xx}}
\end{align*}
$$

where

$$
egin{align*}
s_{xy} &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \\
s_{xx} &= \sum_{i=1}^{n} (x_i - \bar{x})^2, \\
\bar{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i, \\
\bar{y} &= \frac{1}{n} \sum_{i=1}^{n} y_i.
\end{align*}
$$

The fitted model is $y = \hat{b}_0 + \hat{b}_1 x$.

**Case 1: Prediction of average value of $y$**

Suppose we want to predict the value of $E(y)$ for a given value of $x = x_0$. Then the predictor is given by

$$
\hat{p}_m = b_0 + b_1 x_0.
$$

Here $m$ stands for mean value.
Predictive bias

The prediction error is given as
\[ p_m - E(y) = b_0 + b_1x_0 - E(\beta_0 + \beta_1x_0 + \varepsilon)\]
\[ = b_0 + b_1x_0 - (\beta_0 + \beta_1x_0)\]
\[ = (b_0 - \beta_0) + (b_1 - \beta_1)x_0.\]

Then the prediction bias is given as
\[ E[p_m - E(y)] = E(b_0 - \beta_0) + E(b_1 - \beta_1)x_0\]
\[ = 0 + 0 = 0.\]

Thus the predictor \( p_m \) is an unbiased predictor of \( E(y) \).

Predictive variance:

The predictive variance of \( p_m \) is
\[ PV(p_m) = Var(b_0 + b_1x_0)\]
\[ = Var(\bar{y} + b_1(x_0 - \bar{x}))\]
\[ = Var(\bar{y}) + (x_0 - \bar{x})^2 Var(b_1) + 2(x_0 - \bar{x}) Cov(\bar{y}, b_1)\]
\[ = \sigma^2 + \frac{\sigma^2}{n} s_{xx} + 0\]
\[ = \sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right].\]

Estimate of predictive variance

The predictive variance can be estimated by substituting \( \sigma^2 \) by \( \hat{\sigma}^2 = MSE \) as
\[ \hat{PV}(p_m) = \hat{\sigma}^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right]\]
\[ = MSE \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right].\]

Prediction interval:

The 100(1-\( \alpha \))% prediction interval for \( E(y) \) is obtained as follows:

The predictor \( p_m \) is a linear combination of normally distributed random variables, so it is also normally distributed as
\[ p_m \sim N\left(\beta_0 + \beta_1x_0, PV\left(p_m\right)\right).\]
So if $\sigma^2$ is known, then the distribution of

$$\frac{p_m - E(y)}{\sqrt{PV(p_m)}}$$

is $N(0,1)$. So the 100$(1-\alpha)$% prediction interval is obtained as

$$P\left[ -z_{\alpha/2} \leq \frac{p_m - E(y)}{\sqrt{PV(p_m)}} \leq z_{\alpha/2} \right] = 1 - \alpha$$

which gives the prediction interval for $E(y)$ as

$$\left[ p_m - z_{\alpha/2} \sqrt{\sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right]}, \ p_m + z_{\alpha/2} \sqrt{\sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right]} \right].$$

When $\sigma^2$ is unknown, it is replaced by $\hat{\sigma}^2 = MSE$ and in this case, the sampling distribution of

$$\frac{p_m - E(y)}{MSE \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right]}$$

is $t$-distribution with $(n-2)$ degrees of freedom, i.e., $t_{n-2}$.

The 100$(1-\alpha)$% prediction interval in this case is

$$P\left[ -t_{\alpha/2,n-2} \leq \frac{p_m - E(y)}{MSE \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right]} \leq t_{\alpha/2,n-2} \right] = 1 - \alpha.$$ 

which gives the prediction interval as

$$\left[ p_m - t_{\alpha/2,n-2} \sqrt{MSE \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right]}, \ p_m + t_{\alpha/2,n-2} \sqrt{MSE \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right]} \right].$$

Note that the width of prediction interval $E(y)$ is a function of $x_0$. The interval width is minimum for $x_0 = \bar{x}$ and widens as $|x_0 - \bar{x}|$ increases. This is expected also as the best estimates of $y$ to be made at $x$-values lie near the center of the data and the precision of estimation to deteriorate as we move to the boundary of the $x$-space.
Case 2: Prediction of actual value

If $x_0$ is the value of the explanatory variable, then the actual value predictor for $y$ is

$$p_a = b_0 + b_1 x_0.$$  

Here $a$ means “actual”. The true value of $y$ in the prediction period is given by $y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$ where $\varepsilon_0$ indicates the value that would be drawn from the distribution of random error in the prediction period. Note that the form of predictor is the same as of average value predictor but its predictive error and other properties are different. This is the dual nature of predictor.

Predictive bias:

The predictive error of $p_a$ is given by

$$p_a - y_0 = b_0 + b_1 x_0 - (\beta_0 + \beta_1 x_0 + \varepsilon_0) = (b_0 - \beta_0) + (b_1 - \beta_1)x_0 - \varepsilon_0.$$  

Thus, we find that

$$E(p_a - y_0) = E(b_0 - \beta_0) + E(b_1 - \beta_1)x_0 - E(\varepsilon_0) = 0 + 0 + 0 = 0$$

which implies that $p_a$ is an unbiased predictor of $y$.

Predictive variance

Because the future observation $y_0$ is independent of $p_a$, the predictive variance of $p_a$ is

$$PV(p_a) = E((p_a - y_0)^2) = E[(b_0 - \beta_0)^2 + (x_0 - \bar{x})(b_1 - \beta_1) + (b_1 - \beta_1)\bar{x} - \varepsilon_0]^2$$

$$= Var(b_0) + (x_0 - \bar{x})^2Var(b_1) + \bar{x}^2Var(b_1) + Var(\varepsilon_0) + 2(x_0 - \bar{x})Cov(b_0, b_1) + 2\bar{x}Cov(b_0, b_1) + 2(x_0 - \bar{x})Var(b_1)$$

[rest of the terms are 0 assuming the independence of $\varepsilon_0$ with $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$]

$$= Var(b_0) + [(x_0 - \bar{x})^2 + \bar{x}^2 + 2(x_0 - \bar{x})]Var(b_1) + Var(\varepsilon_0) + 2[(x_0 - \bar{x}) + 2\bar{x}]Cov(b_0, b_1)$$

$$= Var(b_0) + x_0^2Var(b_1) + Var(\varepsilon_0) + 2x_0Cov(b_0, b_1)$$

$$= \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} \right] + x_0^2 \frac{\sigma^2}{s_{xx}} + \sigma^2 - 2x_0 \frac{\bar{x}\sigma^2}{s_{xx}}$$

$$= \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right].$$
**Estimate of predictive variance**

The estimate of predictive variance can be obtained by replacing $\sigma^2$ by its estimate $\hat{\sigma}^2 = MSE$ as

$$\widehat{PV}(p_a) = \hat{\sigma}^2 \left[ 1 + \frac{1}{n} \cdot \frac{(x_0 - \bar{x})^2}{s_{xx}} \right] = MSE \left[ 1 + \frac{1}{n} \cdot \frac{(x_0 - \bar{x})^2}{s_{xx}} \right].$$

**Prediction interval:**

If $\sigma^2$ is known, then the distribution of

$$\frac{p_a - y_0}{\sqrt{PV(p_a)}}$$

is $N(0,1)$. So the 100(1-$\alpha$)% prediction interval for $y_0$ is obtained as

$$P \left[ -z_{a/2} \leq \frac{p_a - y_0}{\sqrt{PV(p_a)}} \leq z_{a/2} \right] = 1 - \alpha$$

which gives the prediction interval for $y_0$ as

$$\left[ p_a - z_{a/2} \sqrt{\sigma^2 \left( 1 + \frac{1}{n} \cdot \frac{(x_0 - \bar{x})^2}{s_{xx}} \right)}, p_a + z_{a/2} \sqrt{\sigma^2 \left( 1 + \frac{1}{n} \cdot \frac{(x_0 - \bar{x})^2}{s_{xx}} \right)} \right].$$

When $\sigma^2$ is unknown, then

$$\frac{p_a - y_0}{\sqrt{PV(p_a)}}$$

follows a $t$-distribution with $(n-2)$ degrees of freedom. The 100(1-$\alpha$)% prediction interval for $y_0$ in this case is obtained as

$$P \left[ -t_{a/2,n-2} \leq \frac{p_a - y_0}{\sqrt{PV(p_a)}} \leq t_{a/2,n-2} \right] = 1 - \alpha$$

which gives the prediction interval for $y_0$ as

$$\left[ p_a - t_{a/2,n-2} \sqrt{MSE \left( 1 + \frac{1}{n} \cdot \frac{(x_0 - \bar{x})^2}{s_{xx}} \right)}, p_a + t_{a/2,n-2} \sqrt{MSE \left( 1 + \frac{1}{n} \cdot \frac{(x_0 - \bar{x})^2}{s_{xx}} \right)} \right].$$

The prediction interval is of minimum width at $x_0 = \bar{x}$ and widens as $|x_0 - \bar{x}|$ increases.

The prediction interval for $p_a$ is wider than the prediction interval for $p_m$ because the prediction interval for $p_a$ depends on both the error from the fitted model as well as the error associated with the future observations.
Within sample prediction in multiple linear regression model

Consider the multiple regression model with $k$ explanatory variables as

$$y = X\beta + \varepsilon,$$

where $y = (y_1, y_2, ..., y_n)'$ is a $n \times 1$ vector of $n$ observation on study variable,

$$X = \begin{pmatrix}
  x_{11} & x_{12} & \cdots & x_{1k} \\
  x_{21} & x_{22} & \cdots & x_{2k} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{n1} & x_{n2} & \cdots & x_{nk}
\end{pmatrix}$$

is a $n \times k$ matrix of $n$ observations on each of the $k$ explanatory variables, $\beta = (\beta_1, \beta_2, ..., \beta_k)'$ is a $k \times 1$ vector of regression coefficients and $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)'$ is a $n \times 1$ vector of random error components or disturbance term following $N(0, \sigma^2I_n)$. If intercept term is present, take first column of $X$ to be $(1, 1, ..., 1)'$.

Let the parameter $\beta$ be estimated by its ordinary least squares estimator $b = (X'X)^{-1}X'y$. Then the predictor is $p = Xb$ which can be used for predicting the actual and average values of study variable. This is the dual nature of predictor.

**Case 1: Prediction of average value of $y$**

When the objective is to predict the average value of $y$, i.e., $E(y)$, then the estimation error is given by

$$p - E(y) = Xb - X\beta$$

$$= X(b - \beta)$$

$$= X(X'X)^{-1}X'\varepsilon$$

$$= H\varepsilon$$

where $H = X(X'X)^{-1}X'$.

Then

$$E[p - E(y)] = X\beta - X\beta = 0$$

which proves that the predictor $p = Xb$ provides unbiased prediction for average value.

The predictive variance of $p$ is

$$PV_m(p) = E\left[\left\{p - E(y)\right\}'\left\{p - E(y)\right\}\right]$$

$$= E[\varepsilon'HH\varepsilon]$$

$$= E(\varepsilon'HH\varepsilon)$$

$$= \sigma^2trH = \sigma^2k.$$
When $\sigma^2$ is known, then the distribution of

$$\frac{p - E(y)}{\sqrt{P\tilde{V}_m(p)}}$$

is $N(0,1)$. So the 100(1-$\alpha$)% prediction interval for $E(y)$ is obtained as

$$P\left[-z_{\alpha/2} \leq \frac{p - E(y)}{\sqrt{P\tilde{V}_m(p)}} \leq z_{\alpha/2}\right] = 1 - \alpha$$

which gives the prediction interval for $E(y)$ as

$$\left[ p - z_{\alpha/2}\sqrt{P\tilde{V}_m(p)}, \ p + z_{\alpha/2}\sqrt{P\tilde{V}_m(p)} \right].$$

When $\sigma^2$ is unknown, it is replaced by $\hat{\sigma}^2 = MSE$ and in this case, the sampling distribution of

$$\frac{p - E(y)}{\sqrt{P\tilde{V}_m(p)}}$$

is $t$-distribution with $(n-k)$ degrees of freedom, i.e., $t_{n-k}$.

The 100(1-$\alpha$)% prediction interval for $E(y)$ in this case is

$$P\left[-t_{\alpha/2,n-k} \leq \frac{p - E(y)}{\sqrt{P\tilde{V}_m(p)}} \leq t_{\alpha/2,n-k}\right] = 1 - \alpha.$$ 

which gives the prediction interval for $E(y)$ as

$$\left[ p - t_{\alpha/2,n-k}\sqrt{P\tilde{V}_m(p)}, \ p + t_{\alpha/2,n-k}\sqrt{P\tilde{V}_m(p)} \right].$$

**Case 2: Prediction of actual value of $y$**

When the predictor $p = Xb$ is used for predicting the actual value of study variable $y$, then its prediction error is given by

$$p - y = Xb - X\beta - \varepsilon$$

$$= X(b - \beta) - \varepsilon$$

$$= X(X'X)^{-1}X'^{\prime}\varepsilon - \varepsilon$$

$$= -\left[I - X(X'X)^{-1}X'^{\prime}\right]\varepsilon$$

$$= -\tilde{H}\varepsilon.$$ 

Thus

$$E(p - y) = 0$$
which shows that \( p \) provides unbiased predictions for the actual values of study variable.

The predictive variance in this case is

\[
PV_a(p) = E\left[ (p - y)(p - y) \right] \\
= E(\varepsilon' H\varepsilon) \\
= E(\varepsilon' H\varepsilon) \\
= \sigma^2 \text{tr} H \\
= \sigma^2(n-k).
\]

The predictive variance can be estimated by

\[
\hat{PV}_m(p) = \hat{\sigma}^2(n-k)
\]

where \( \hat{\sigma}^2 = MSE \) is obtained from analysis of variance based on OLSE.

Comparing the performances of \( p \) to predict actual and average values, we find that \( p \) in better predictor for predicting the average value in comparison to actual value when

\[
PV_m(p) < PV_a(p) \\
\text{or} \quad k < (n-k) \\
\text{or} \quad 2k < n.
\]

i.e. when the total number of observations are more than twice the number of explanatory variables.

Now we obtain the confidence interval for \( y \).

When \( \sigma^2 \) is known, then the distribution of

\[
\frac{p - y}{\sqrt{PV_a(p)}}
\]

is \( N(0,1) \). So the 100(1-\( \alpha \))\% prediction interval for \( y \) is obtained as

\[
P\left[ -z_{\alpha/2} \leq \frac{p - y}{\sqrt{PV_a(p)}} \leq z_{\alpha/2} \right] = 1-\alpha
\]

which gives the prediction interval for \( y \) as

\[
\left[ p - z_{\alpha/2}\sqrt{PV_a(p)}, p + z_{\alpha/2}\sqrt{PV_a(p)} \right].
\]

When \( \sigma^2 \) is unknown, it is replaced by \( \hat{\sigma}^2 = MSE \) and in this case, the sampling distribution of

\[
\frac{p - y}{\sqrt{PV_a(p)}}
\]

is \( t \)-distribution with \( (n-k) \) degrees of freedom, i.e., \( t_{n-k} \).
The 100(1-α)% prediction interval of $y$ in this case is obtained as

$$P \left[ -t_{a/2, n-k} \leq \frac{p - y}{\sqrt{PV_a(p)}} \leq t_{a/2, n-k} \right] = 1 - \alpha.$$ 

which gives the prediction interval for $y$ as

$$\left[ p - t_{a/2, n-k} \sqrt{PV_a(p)}, \ p + t_{a/2, n-k} \sqrt{PV_a(p)} \right],$$

### Outside sample prediction in multiple linear regression model

Consider the model

$$y = X\beta + \varepsilon \tag{1}$$

where $y$ is a $n \times 1$ vector of $n$ observations on study variable, $X$ is a $n \times k$ matrix of explanatory variables and $\varepsilon$ is a $n \times 1$ vector of disturbances following $N(0, \sigma^2 I_n)$.

Further, suppose a set of $n_f$ observations on the same set of $k$ explanatory variables are also available but the corresponding $n_f$ observations on the study variable are not available. Assuming that this set of observation also follows the same model, we can write

$$y_f = X_f\beta + \varepsilon_f \tag{2}$$

where $y_f$ is a $n_f \times 1$ vector of future values, $X_f$ is a $n_f \times k$ matrix of known values of explanatory variables and $\varepsilon_f$ is a $n_f \times 1$ vector of disturbances following $N(0, \sigma^2 I_{n_f})$. It is also assumed that the elements of $\varepsilon$ and $\varepsilon_f$ are independently distributed.

We now consider the prediction of $y_f$ values for given $X_f$ from model (2). This can be done by estimating the regression coefficients from model (1) based on $n$ observations and use it is formulating the predictor in model (2). If ordinary least squares estimation is used to estimate $\beta$ in model (1) as

$$b = (X'^X)^{-1}X'y$$

then the corresponding predictor is

$$p_f = X_fb = X_f(X'^X)^{-1}X'y.$$
Case 1: Prediction of average value of study variable

When the aim is to predict the average value $E(y_f)$, then the prediction error is

$$p_f - E(y_f) = X_f b - X_f \beta = X_f (b - \beta) = X_f (X'X)^{-1}X'_f \varepsilon.$$  

Then

$$E[p_f - E(y_f)] = X_f (X'X)^{-1}X'_f E(\varepsilon) = 0.$$  

Thus $p_f$ provides unbiased prediction for average value.

The predictive covariance matrix of $p_f$ is

$$Cov_m(p_f) = E\left[\{p_f - E(y_f)\}'\{p_f - E(y_f)\}\right]$$

$$= E\left[ X_f (X'X)^{-1}X'_f \varepsilon \varepsilon' X(X'X)^{-1}X'_f \right]$$

$$= X_f (X'X)^{-1}X'_f E(\varepsilon\varepsilon')X(X'X)^{-1}X'_f$$

$$= \sigma^2 X_f (X'X)^{-1}X'_f X(X'X)^{-1}X'_f$$

$$= \sigma^2 X_f (X'X)^{-1}X'_f.$$  

The predictive variance of $p_f$ is

$$PV_m(p_f) = E\left[\{p_f - E(y_f)\}'\{p_f - E(y_f)\}\right]$$

$$= tr\left[Cov_m(p_f)\right]$$

$$= \sigma^2 tr\left[(X'X)^{-1}X'_f X'_f\right].$$  

If $\sigma^2$ is unknown, then replace $\sigma^2$ by $\hat{\sigma}^2 = MSE$ in the expressions of predictive covariance matrix and predictive variance and there estimates are

$$\hat{Cov}_m(p_f) = \hat{\sigma}^2 X_f (X'X)^{-1}X'_f$$

$$\hat{PV}_m(p_f) = \hat{\sigma}^2 tr\left[(X'X)^{-1}(X'_f X'_f)\right].$$  

Now we obtain the confidence interval for $E(y_f)$.

When $\sigma^2$ is known, then the distribution of

$$\frac{p_f - E(y_f)}{\sqrt{PV_m(p_f)}}$$

is $N(0,1)$. So the 100(1-$\alpha$)% prediction interval of $E(y_f)$ is obtained as
\[
P \left[ -z_{\alpha/2} \leq \frac{p_f - E(y_f)}{\sqrt{P\hat{V}_m(p_f)}} \leq z_{\alpha/2} \right] = 1 - \alpha
\]

which gives the prediction interval for \( E(y_f) \) as

\[
\left[ p_f - z_{\alpha/2} \sqrt{P\hat{V}_m(p_f)}, \ p_f + z_{\alpha/2} \sqrt{P\hat{V}_m(p_f)} \right].
\]

When \( \sigma^2 \) is unknown, it is replaced by \( \hat{\sigma}^2 = MSE \) and in this case, the sampling distribution of

\[
\frac{p_f - E(y_f)}{\sqrt{P\hat{V}_m(p_f)}},
\]

is \( t \)-distribution with \((n - k)\) degrees of freedom, i.e., \( t_{n-k} \).

The 100(1-\( \alpha \))% prediction interval for \( E(y_f) \) in this case is

\[
P \left[ -t_{\alpha/2,n-k} \leq \frac{p_f - E(y_f)}{\sqrt{P\hat{V}_m(p_f)}} \leq t_{\alpha/2,n-k} \right] = 1 - \alpha.
\]

which gives the prediction interval for \( E(y_f) \) as

\[
\left[ p_m - t_{\alpha/2,n-k} \sqrt{P\hat{V}_m(p)}, \ p_m + t_{\alpha/2,n-k} \sqrt{P\hat{V}_m(p)} \right].
\]

**Case 2: Prediction of actual value of study variable**

When \( p_f \) is used to predict the actual value \( y_f \), then the prediction error is

\[
p_f - y_f = X_f b - X_f \beta - \varepsilon_f = X_f (b - \beta) - \varepsilon_f.
\]

Then

\[
E\left( p_f - y_f \right) = X_f E(b - \beta) - E(\varepsilon_f) = 0.
\]

Thus \( p_f \) provides unbiased prediction for actual values.

The predictive covariance matrix of \( p_f \) in this case is

\[
Cov_a \left( p_f \right) = E\left[ \left( p_f - y_f \right) \left( p_f - y_f \right)^\prime \right] = X_f V(b) X_f + E(\varepsilon_f \varepsilon_f^\prime) \quad (Using \ (b - \beta) = (X^\prime X)^{-1} X^\prime \varepsilon)
\]

\[
= \sigma^2 \left[ X_f (X^\prime X)^{-1} X_f + I_n \right].
\]
The predictive variance of $p_f$ is

$$PV_a(p_f) = E[(p_f - y_f)'(p_f - y_f)] = tr[Cov_a(p_f)] = \sigma^2[tr(X'X)^{-1}X_fX_f + n_f].$$

The estimates of covariance matrix and predictive variance can be obtained by replacing $\sigma^2$ by $\hat{\sigma}^2 = MSE$ as

$$\hat{Cov}_a(p_f) = \hat{\sigma}^2\left[X_f(X'X)^{-1}X_f + I_{n_f}\right]$$
$$\hat{PV}_a(p_f) = \hat{\sigma}^2\left[tr(X'X)^{-1}X_fX_f + n_f\right].$$

Now we obtain the confidence interval for $y_f$.

When $\sigma^2$ is known, then the distribution of

$$\frac{p_f - y_f}{\sqrt{PV_a(p_f)}}$$

is $N(0,1)$. So the 100(1-$\alpha$)% prediction interval is obtained as

$$P\left[-z_{\alpha/2} \leq \frac{p_f - y_f}{\sqrt{PV_a(p_f)}} \leq z_{\alpha/2}\right] = 1 - \alpha$$

which gives the prediction interval for $y_f$ as

$$\left[p_f - z_{\alpha/2}\sqrt{PV_a(p_f)}, p_f + z_{\alpha/2}\sqrt{PV_a(p_f)}\right].$$

When $\sigma^2$ is unknown, it is replaced by $\hat{\sigma}^2 = MSE$ and in this case, the sampling distribution of

$$\frac{p_f - y_f}{\sqrt{PV_a(p_f)}}$$

is $t$-distribution with $(n-k)$ degrees of freedom, i.e., $t_{n-k}$.

The 100(1-$\alpha$)% prediction interval for $y_f$ in this case is

$$P\left[-t_{\alpha/2,n-k} \leq \frac{p_f - y_f}{\sqrt{PV_a(p_f)}} \leq t_{\alpha/2,n-k}\right] = 1 - \alpha.$$ 

which gives the prediction interval for $y_f$ as

$$\left[p_f - t_{\alpha/2,n-k}\sqrt{PV_a(p_f)}, p_f + t_{\alpha/2,n-k}\sqrt{PV_a(p_f)}\right].$$
Simultaneous prediction of average and actual values of study variable

The predictions are generally obtained either for the average values of study variable or actual values of study variable. In many applications, it may not be appropriate to confine our attention to only to either of the two. It may be more appropriate in some situations to predict both the values simultaneously, i.e., consider the prediction of actual and average values of study variable simultaneously. For example, suppose a firm deals with the sale of fertilizer to the user. The interest of company would be in predicting the average value of yield which the company would like to use in showing that the average yield of the crop increases by using their fertilizer. On the other side, the user would not be interested in the average value. The user would like to know the actual increase in the yield by using the fertilizer. Suppose both seller and user, both go for prediction through regression modeling. Now using the classical tools, the statistician can predict either the actual value or the average value. This can safeguard the interest of either the user or the seller. Instead of this, it is required to safeguard the interest of both by striking a balance between the objectives of the seller and the user. This can be achieved by combining both the predictions of actual and average values. This can be done by formulating an objective function or target function. Such target function has to be flexible and should allow to assign different weights to the choice of two kinds of predictions depending upon their importance in any given application and also reducible to individual predictions leading to actual and average value prediction.

Now we consider the simultaneous prediction in within and outside sample cases.

Simultaneous prediction in within sample prediction

Define a target function

\[ \tau = \lambda y + (1 - \lambda)E(y); \quad 0 \leq \lambda \leq 1 \]

which is a convex combination of actual value \( y \) and average value \( E(y) \). The weight \( \lambda \) is a constant lying between zero and one whose value reflects the importance being assigned to actual value prediction. Moreover \( \lambda = 0 \) gives the average value prediction and \( \lambda = 1 \) gives the actual value prediction. For example, the value of \( \lambda \) in the fertilizer example depends on the rules and regulation of market, norms of society and other considerations etc. The value of \( \lambda \) is the choice of practitioner.

Consider the multiple regression model

\[ y = X\beta + \varepsilon, \quad E(\varepsilon) = 0, \quad E(\varepsilon \varepsilon') = \sigma^2 I_n. \]

Estimate \( \beta \) by ordinary least squares estimation and construct the predictor

\[ p = Xb. \]

Now employ this predictor for predicting the actual and average values simultaneously through the target function.
The prediction error is

\[ p - \tau = Xb - \lambda y - (1 - \lambda)E(y) \]
\[ = Xb - \lambda (X\beta + \varepsilon) - (1 - \lambda)X\beta \]
\[ = X(b - \beta) - \lambda \varepsilon. \]

Thus

\[ E(p - \tau) = XE(b - \beta) - \lambda E(\varepsilon) = 0. \]

So \( p \) provides unbiased prediction for \( \tau \).

The variance is

\[ \text{Var}(p) = E((p - \tau)'(p - \tau)) \]
\[ = E\left[\left((b - \beta)'X' - \lambda \varepsilon\right)'\left\{X(b - \beta) - \lambda \varepsilon\right\}\right] \]
\[ = E\left[\varepsilon'X(X'X)^{-1}X'X\varepsilon + \lambda^2 \varepsilon'\varepsilon - \lambda(b - \beta)'X'\varepsilon - \lambda \varepsilon'X(b - \beta)\right] \]
\[ = E\left[(1 - 2\lambda)\varepsilon'X(X'X)^{-1}X'\varepsilon + \lambda^2 \varepsilon'\varepsilon\right] \]
\[ = \sigma^2(1 - 2\lambda)\text{tr}\left[(X'X)^{-1}X'X\right] + \lambda^2 \sigma^2 \text{tr}I_n \]
\[ = \sigma^2\left[(1 - 2\lambda)k + \lambda^2 n\right]. \]

The estimates of predictive variance can be obtained by replacing \( \sigma^2 \) by \( \hat{\sigma}^2 = \text{MSE} \) as

\[ \text{Var}(p) = \hat{\sigma}^2\left[(1 - 2\lambda)k + \lambda^2 n\right]. \]

**Simultaneous prediction is outside sample prediction:**

Consider the model described earlier under outside sample prediction as

\[ y_{n \times 1} = X_{n \times k} \beta + \varepsilon_{n \times 1}, \quad E(\varepsilon) = 0, \quad V(\varepsilon) = \sigma^2 I_n \]
\[ y_{f \times 1} = X_{f \times k} \beta + \varepsilon_{f \times 1}, \quad E(\varepsilon_f) = 0, \quad V(\varepsilon_f) = \sigma^2 I_{n_f}. \]

The target function in this case is defined as

\[ \tau_f = \lambda y_f + (1 - \lambda)E(y_f); \quad 0 \leq \lambda \leq 1. \]

The predictor based on OLSE of \( \beta \) is

\[ p_f = X_f b; \quad b = \left(X'X\right)^{-1}X'y. \]

The predictive error of \( p \) is

\[ p_f - \tau_f = X_f b - \lambda y_f - (1 - \lambda)E(y_f) \]
\[ = X_f b - \lambda (X_f\beta + \varepsilon_f) - (1 - \lambda)X_f\beta \]
\[ = X_f (b - \beta) - \lambda \varepsilon_f. \]
So
\[ E(p_f - \tau_f) = X_f E(b - \beta) - \lambda E(\varepsilon_f) = 0. \]

Thus, \( p_f \) provides unbiased prediction for \( \tau_f \).

The variance of \( p_f \) is
\[
\text{Var}(p_f) = E((p_f - \tau)^\prime (p_f - \tau))
= E[(b - \beta)' \varepsilon_f - \lambda \varepsilon_f] [X_f (b - \beta) - \lambda \varepsilon_f]
= E[\varepsilon' X(X'X)^{-1} X_f (X'X)^{-1} X' \varepsilon + \lambda \varepsilon_f X_f (X'X)^{-1} X' \varepsilon - 2 \lambda \varepsilon_f X_f (X'X)^{-1} X' \varepsilon]
= \sigma^2 [tr \{ X(X'X)^{-1} X_f (X'X)^{-1} X' \} + \lambda^2 n_f]
\]
assuming that the elements are \( \varepsilon \) and \( \varepsilon_f \) are mutually independent.

The estimates of predictive variance can be obtained by replacing \( \sigma^2 \) by \( \hat{\sigma}^2 = \text{MSE} \) as
\[
\text{Var}(p_f) = \hat{\sigma}^2 [tr \{ X(X'X)^{-1} X_f (X'X)^{-1} X' \} + \lambda^2 n_f].
\]
Chapter 5
Generalized and Weighted Least Squares Estimation

The usual linear regression model assumes that all the random error components are identically and independently distributed with constant variance. When this assumption is violated, then ordinary least squares estimator of regression coefficient looses its property of minimum variance in the class of linear and unbiased estimators. The violation of such assumption can arise in anyone of the following situations:

1. The variance of random error components is not constant.
2. The random error components are not independent.
3. The random error components do not have constant variance as well as they are not independent.

In such cases, the covariance matrix of random error components does not remain in the form of an identity matrix but can be considered as any positive definite matrix. Under such assumption, the OLSE does not remain efficient as in the case of identity covariance matrix. The generalized or weighted least squares method is used in such situations to estimate the parameters of the model.

In this method, the deviation between the observed and expected values of $y_i$ is multiplied by a weight $\omega_i$ where $\omega_i$ is chosen to be inversely proportional to the variance of $y_i$.

For simple linear regression model, the weighted least squares function is

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \omega_i (y_i - \beta_0 - \beta_1 x_i)^2.$$ 

The least squares normal equations are obtained by differentiating $S(\beta_0, \beta_1)$ with respect to $\beta_0$ and $\beta_1$ and equating them to zero as

$$\hat{\beta}_0 \sum_{i=1}^{n} \omega_i + \hat{\beta}_1 \sum_{i=1}^{n} \omega_i x_i = \sum_{i=1}^{n} \omega_i y_i$$
$$\hat{\beta}_0 \sum_{i=1}^{n} \omega_i x_i + \hat{\beta}_1 \sum_{i=1}^{n} \omega_i x_i^2 = \sum_{i=1}^{n} \omega_i x_i y_i.$$ 

Solution of these two normal equations give the weighted least squares estimate of $\beta_0$ and $\beta_1$. 

**Generalized least squares estimation**

Suppose in usual multiple regression model

\[ y = X\beta + \varepsilon \quad \text{with} \quad E(\varepsilon) = 0, \quad V(\varepsilon) = \sigma^2 I, \]

the assumption \( V(\varepsilon) = \sigma^2 I \) is violated and become

\[ V(\varepsilon) = \sigma^2 \Omega \]

where \( \Omega \) is a known \( n \times n \) nonsingular, positive definite and symmetric matrix.

This structure of \( \Omega \) incorporates both the cases.

- when \( \Omega \) is diagonal but with unequal variances and
- when \( \Omega \) is not necessarily diagonal depending on the presence of correlated errors, then the off-diagonal elements are nonzero.

The OLSE of \( \beta \) is

\[ b = (X'X)^{-1}X'y \]

In such cases OLSE gives unbiased estimate but has more variability as

\[ E(b) = (X'X)^{-1}X'E(y) = (X'X)^{-1}X'X\beta = \beta \]
\[ V(b) = (X'X)^{-1}X'V(y)X(X'X)^{-1} = \sigma^2 (X'X)^{-1}X'\Omega X(X'X)^{-1}. \]

Now we attempt to find better estimator as follows:

Since \( \Omega \) is positive definite, symmetric, so there exists a nonsingular matrix \( K \) such that \( KK' = \Omega \).

Then in the model

\[ y = X\beta + \varepsilon, \]

premultiply by \( K^{-1} \) gives

\[ K^{-1}y = K^{-1}X\beta + K^{-1}\varepsilon \]

or

\[ z = B\beta + g \]

where \( z = K^{-1}y, B = K^{-1}X, g = K^{-1}\varepsilon \). Now observe that

\[ E(g) = K^{-1}E(\varepsilon) = 0 \]
\[ V(g) = E\left[ \{g - E(g)\} \{g - E(g)\}' \right] \\
= E(gg') \\
= E\left[ K^{-1} \varepsilon \varepsilon' K^{-1} \right] \\
= K^{-1} E(\varepsilon \varepsilon') K^{-1} \\
= \sigma^2 K^{-1} \Omega K^{-1} \\
= \sigma^2 K^{-1} K' K^{-1} \\
= \sigma^2 I. \]

Thus the elements of \( g \) have 0 mean and unit variance they are uncorrelated.

So either minimize \( S(\beta) = g' g \)

\[ = \varepsilon' \Omega^{-1} \varepsilon \\
= (y - X\beta)' \Omega^{-1} (y - X\beta) \]

and get normal equations as

\[ (X' \Omega^{-1} X) \hat{\beta} = X' \Omega^{-1} y \]
or \( \hat{\beta} = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y \).

Alternatively, we can apply OLS to transformed model and obtain OLSE of \( \beta \) as

\[ \hat{\beta} = (B' B)^{-1} B' z \]

\[ = (X' K^{-1} K^{-1} X)^{-1} X' K^{-1} K^{-1} y \]

\[ = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y \]

This is termed as **generalized least squares estimator (GLSE)** of \( \beta \).

The estimation error of GLSE is

\[ \hat{\beta} = (B' B)^{-1} B' (B \beta + g) \]

\[ = \beta + (B' B)^{-1} B' g \]
or \( \hat{\beta} - \beta = (B' B)^{-1} B' g. \)

Then

\[ E(\hat{\beta} - \beta) = (B' B)^{-1} B' E(g) = 0 \]

which shows that GLSE is an unbiased estimator of \( \beta \). The covariance matrix of GLSE is given by

\[ V(\hat{\beta}) = E\left[ \{\hat{\beta} - E(\hat{\beta})\} \{\hat{\beta} - E(\hat{\beta})\}' \right] \\
= E\left[ (B' B)^{-1} B' gg' B'(B' B)^{-1} \right] \\
= (B' B)^{-1} B' E(gg') B'(B' B)^{-1}. \]
\[
E(gg') = K^{-1}E(\varepsilon \varepsilon')K^{-1} \\
= \sigma^2 K^{-1} \Omega K^{-1} \\
= \sigma^2 K^{-1} K K' K^{-1} \\
= \sigma^2 I,
\]
so
\[
V(\hat{\beta}) = \sigma^2 (B' B)^{-1} B' B (B' B)^{-1} \\
= \sigma^2 (B' B)^{-1} \\
= \sigma^2 (X' K^{-1} K^{-1} X)^{-1} \\
= \sigma^2 (X' \Omega^{-1} X)^{-1}.
\]

Now we prove that GLSE is the best linear unbiased estimator of \( \beta \).

**The Gauss-Markov theorem for the case \( \text{Var}(\varepsilon) = \Omega \)**

The Gauss-Markov theorem establishes that the generalized least-squares (GLS) estimator of \( \beta \) given by \( \hat{\beta} = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y \), is BLUE (best linear unbiased estimator). By best \( \beta \), we mean that \( \hat{\beta} \) minimizes the variance for any linear combination of the estimated coefficients, \( \ell' \hat{\beta} \). We note that
\[
E(\hat{\beta}) = E \left[ (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y \right] \\
= (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} E(y) \\
= (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} X \beta \\
= \beta.
\]
Thus \( \hat{\beta} \) is an unbiased estimator of \( \beta \).

The covariance matrix of \( \hat{\beta} \) is given by
\[
V(\hat{\beta}) = \left[ (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} \right] V(y) \left[ (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} \right]' \\
= \left[ (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} \right] \Omega \left[ (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} \right]' \\
= \left[ (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} \right] \Omega \left[ \Omega^{-1} X (X' \Omega^{-1} X)^{-1} \right] \\
= (X' \Omega^{-1} X)^{-1}.
\]
Thus,
\[
\text{Var}(\ell' \hat{\beta}) = \ell' \text{Var}(\hat{\beta}) \ell \\
= \ell' \left[ (X' \Omega^{-1} X)^{-1} \right] \ell.
\]
Let $\tilde{\beta}$ be another unbiased estimator of $\beta$ that is a linear combination of the data. Our goal, then, is to show that $\text{Var}(\ell^T \tilde{\beta}) \geq \ell^T (X' \Omega^{-1} X)^{-1} \ell$ with at least one $\ell$ such that $\text{Var}(\ell^T \tilde{\beta}) \geq \ell^T (X' \Omega^{-1} X)^{-1} \ell$.

We first note that we can write any other estimator of $\beta$ that is a linear combination of the data as

$$\tilde{\beta} = \left( X' \Omega^{-1} X \right)^{-1} X' \Omega^{-1} + B \right) y + b_0^*$$

where $B$ is an $p \times n$ matrix and $b_0^*$ is a $p \times 1$ vector of constants that appropriately adjusts the GLS estimator to form the alternative estimate. Then

$$E(\tilde{\beta}) = E \left( \left( X' \Omega^{-1} X \right)^{-1} X' \Omega^{-1} + B \right) y + b_0^*$$

$$= \left( X' \Omega^{-1} X \right)^{-1} X' \Omega^{-1} + B \right) E(y) + b_0^*$$

$$= \left( X' \Omega^{-1} X \right)^{-1} X' \Omega^{-1} + B \right) X \beta + b_0^*$$

$$= \left( X' \Omega^{-1} X \right)^{-1} X' \Omega^{-1} X \beta + BX \beta + b_0^*$$

$$= \beta + BX \beta + b_0^*.$$

Consequently, $\tilde{\beta}$ is unbiased if and only if both $b_0^* = 0$ and $BX = 0$. The covariance matrix of $\tilde{\beta}$ is

$$\text{Var}(\tilde{\beta}) = \text{Var} \left( \left( X' \Omega^{-1} X \right)^{-1} X' \Omega^{-1} + B \right) y$$

$$= \left( X' \Omega^{-1} X \right)^{-1} X' \Omega^{-1} + B \right) \text{Var}(y) \left( X' \Omega^{-1} X \right)^{-1} X' \Omega^{-1} + B \right)$$

$$= \left( X' \Omega^{-1} X \right)^{-1} X' \Omega^{-1} + B \right) \Omega \left( X' \Omega^{-1} X \right)^{-1} X' \Omega^{-1} + B \right)$$

$$= \left( X' \Omega^{-1} X \right)^{-1} X' \Omega^{-1} + B \right) \Omega \left( X' \Omega^{-1} X \right)^{-1} X' \Omega^{-1} + B \right)$$

$$= \left( X' \Omega^{-1} X \right)^{-1} + B \Omega \Omega B'$$

because $BX = 0$, which implies that $(BX)' = X' B' = 0$. Then

$$\text{Var}(\ell^T \tilde{\beta}) = \ell^T \text{Var}(\tilde{\beta}) \ell$$

$$= \ell^T \left( \left( X' \Omega^{-1} X \right)^{-1} + B \Omega \Omega B' \right) \ell$$

$$= \ell^T (X' \Omega^{-1} X)^{-1} \ell + \ell^T B \Omega \Omega B' \ell$$

$$= \text{Var}(\ell^T \tilde{\beta}) + \ell^T B \Omega \Omega B' \ell.$$
which must be strictly greater than 0 for some $\ell \neq 0$ unless $B = 0$. For $\ell = (0,0,...,0,1,0,...,0)$ where $1$ occurs at $i^{th}$ place, $\ell' \hat{\beta} = \hat{\beta}_i$ is the best linear unbiased estimator of $\ell' \beta = \beta_i$ for all $i = 1,2,...,k$. Thus, the GLS estimate of $\beta$ is the best linear unbiased estimator.

**Weighted least squares estimation**

When $\varepsilon$'s are uncorrelated and have unequal variances, then

$$V(\varepsilon) = \sigma^2 \Omega = \sigma^2 \begin{bmatrix}
\frac{1}{\omega_1} & 0 & 0 & \cdots & 0 \\
0 & \frac{1}{\omega_2} & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & \frac{1}{\omega_n}
\end{bmatrix}.$$  

The estimation procedure is usually called as weighted least squares.

Let $W = \Omega^{-1}$ then the weighted least squares estimator of $\beta$ is obtained by solving normal equation

$$(X'WX)\hat{\beta} = X'Wy$$

which gives

$$\hat{\beta} = (X'WX)^{-1}X'Wy$$

where $\omega_1, \omega_2,...,\omega_n$ are called the **weights**.

The observations with large variances usual have smaller weights than observations with small variance.
One of the basic objectives in any statistical modeling is to find good estimators of the parameters. In the context of multiple linear regression model \( y = X\beta + \varepsilon \), the ordinary least squares estimator \( b = (X'X)^{-1}X'y \) is the best linear unbiased estimator of \( \beta \). Several approaches have been attempted in the literature to improve further the OLSE. One approach to improve the estimators is the use of extraneous information or prior information. In applied work, such prior information may be available about the regression coefficients. For example, in economics, the constant returns to scale imply that the exponents in a Cobb-Douglas production function should sum to unity. In another example, absence of money illusion on the part of consumers implies that the sum of money income and price elasticities in a demand function should be zero. These types of constraints or the prior information may be available from

(i) some theoretical considerations.
(ii) past experience of the experimenter.
(iii) empirical investigations.
(iv) some extraneous sources etc.

To utilize such information in improving the estimation of regression coefficients, it can be expressed in the form of

(i) exact linear restrictions
(ii) stochastic linear restrictions
(iii) inequality restrictions.

We consider the use of prior information in the form of exact and stochastic linear restrictions in the model \( y = X\beta + \varepsilon \) where \( y \) is a \((n\times1)\) vector of observations on study variable, \( X \) is a \((n\times k)\) matrix of observations on explanatory variables \( X_1, X_2, \ldots, X_k \), \( \beta \) is a \((k\times1)\) vector of regression coefficients and \( \varepsilon \) is a \((n\times1)\) vector of disturbance terms.
**Exact linear restrictions:**

Suppose the prior information binding the regression coefficients is available from some extraneous sources which can be expressed in the form of exact linear restrictions as

\[ r = R \beta \]

where \( r \) is a \((q \times 1)\) vector and \( R \) is a \((q \times k)\) matrix with \( \text{rank}(R) = q \) \((q < k)\). The elements in \( r \) and \( R \) are known.

Some examples of exact linear restriction \( r = R \beta \) are as follows:

(i) If there are two restrictions with \( k = 6 \) like

\[
\begin{align*}
\beta_2 &= \beta_4 \\
\beta_3 + 2\beta_4 + \beta_5 &= 1
\end{align*}
\]

then

\[
\begin{bmatrix}
0 \\
1
\end{bmatrix}, \quad
\begin{bmatrix}
0 & 1 & 0 & -1 & 0 & 0 \\
0 & 0 & 1 & 2 & 1 & 0
\end{bmatrix}
\]

(ii) If \( k = 3 \) and suppose \( \beta_2 = 3 \), then

\[
\begin{bmatrix}
3
\end{bmatrix}, \quad
\begin{bmatrix}
0 & 1 & 0
\end{bmatrix}
\]

(iii) If \( k = 3 \) and suppose \( \beta_1 : \beta_2 : \beta_3 :: ab : b :: 1 \)

then

\[
\begin{bmatrix}
0 \\
0 \\
0
\end{bmatrix}, \quad
\begin{bmatrix}
1 & -a & 0 \\
0 & 1 & -b \\
1 & 0 & -ab
\end{bmatrix}
\]

The ordinary least squares estimator \( \beta = (X'X)^{-1}X'y \) does not uses the prior information. It does not obey the restrictions in the sense that \( r \neq R \beta \). So the issue is how to use the sample information and prior information together in finding an improved estimator of \( \beta \).
**Restricted least squares estimation**

The restricted least squares estimation method enables the use of sample information and prior information simultaneously. In this method, choose $\beta$ such that the error sum of squares is minimized subject to linear restrictions $r = R\beta$. This can be achieved using the Lagrangian multiplier technique. Define the Lagrangian function

$$S(\beta, \lambda) = (y - X\beta)'(y - X\beta) - 2\lambda'(R\beta - r)$$

where $\lambda$ is a $(k \times 1)$ vector of Lagrangian multiplier.

Using the result that if $a$ and $b$ are vectors and $A$ is a suitably defined matrix, then

$$\frac{\partial}{\partial a} a' Aa = (A + A')a$$
$$\frac{\partial}{\partial a} a' b = b,$$

we have

$$\frac{\partial S(\beta, \lambda)}{\partial \beta} = 2X'X\beta - 2X'y - 2R'\lambda' = 0 \quad (*)$$
$$\frac{\partial S(\beta, \lambda)}{\partial \lambda} = R\beta - r = 0.$$

Pre-multiplying equation $(*)$ by $R(X'X)^{-1}$, we have

$$2R\beta - 2R(X'X)^{-1}X'y - 2R(X'X)^{-1}R'\lambda' = 0$$

or

$$R\beta - Rb - R(X'X)^{-1}R'\lambda' = 0$$

$$\Rightarrow \lambda' = -\left[R(X'X)^{-1}R\right]^{-1}(Rb - r)$$

using $R(X'X)^{-1}R' > 0$.

Substituting $\lambda$ in equation $(*)$, we get

$$2X'X\beta - 2X'y + 2R\left[R(X'X)^{-1}R\right]^{-1}(Rb - r) = 0$$

or

$$X'X\beta = X'y - R\left[R(X'X)^{-1}R\right]^{-1}(Rb - r).$$

Pre-multiplying by $(X'X)^{-1}$ yields

$$\hat{\beta}_r = (X'X)^{-1}X'y + (X'X)^{-1}R\left[R(X'X)^{-1}R\right]^{-1}(r - Rb)$$

$$= b - (X'X)^{-1}R\left[R(X'X)^{-1}R\right]^{-1}(Rb - r).$$

This estimation is termed as **restricted regression estimator** of $\beta$. 

*Econometrics | Chapter 6 | Linear Restrictions and Preliminary Test Estimation | Shalabh, IIT Kanpur*
Properties of restricted regression estimator

1. The restricted regression estimator \( \hat{\beta}_R \) obeys the exact restrictions, i.e., \( r = R \hat{\beta}_R \). To verify this, consider

\[
R \hat{\beta}_R = R \left[ b + \left( X'X \right)^{-1} R' \left\{ R(X'X)^{-1} \right\}^{-1} \left( r - Rb \right) \right] \\
= Rb + r - Rb \\
= r.
\]

2. Unbiasedness

The estimation error of \( \hat{\beta}_R \) is

\[
\hat{\beta}_R - \beta = (b - \beta) + \left( X'X \right)^{-1} R' \left\{ R(X'X)^{-1} \right\}^{-1} R \beta - Rb
\]

\[
= \left[ I - \left( X'X \right)^{-1} R' \left\{ R(X'X)^{-1} \right\}^{-1} R \right] \left( b - \beta \right)
\]

\[
= D (b - \beta)
\]

where

\[
D = I - \left( X'X \right)^{-1} R \left\{ R(X'X)^{-1} \right\}^{-1} R.
\]

Thus

\[
E(\hat{\beta}_R - \beta) = DE (b - \beta) \\
= 0
\]

implying that \( \hat{\beta}_R \) is an unbiased estimator of \( \beta \).

3. Covariance matrix

The covariance matrix of \( \hat{\beta}_R \) is

\[
V(\hat{\beta}_R) = E(\hat{\beta}_R - \beta)(\hat{\beta}_R - \beta)^t
\]

\[
= DE (b - \beta)(b - \beta)^t D
\]

\[
= DV(b)D
\]

\[
= \sigma^2 D (X'X)^{-1} D
\]

\[
= \sigma^2 (X'X)^{-1} - \sigma^2 (X'X)^{-1} R \left\{ R(X'X)^{-1} \right\}^{-1} R' (X'X)^{-1}
\]

which can be obtained as follows:
Consider
\[
D(X'X)^{-1} = (X'X)^{-1} - (X'X)^{-1} R' \left[ R(X'X)^{-1} R \right]^{-1} R(X'X)^{-1}
\]
\[
D(X'X)^{-1} D' = \left[ (X'X)^{-1} - (X'X)^{-1} R' \left[ R(X'X)^{-1} R \right]^{-1} R(X'X)^{-1} \right] \left[ I - (X'X)^{-1} R' \left[ R(X'X)^{-1} R \right]^{-1} R \right]
\]
\[
= (X'X)^{-1} - (X'X)^{-1} R' \left[ R(X'X)^{-1} R \right]^{-1} R(X'X)^{-1} - (X'X)^{-1} R' \left[ R(X'X)^{-1} R \right]^{-1} R(X'X)^{-1}
\]
\[
+ (X'X)^{-1} R' \left[ R(X'X)^{-1} R \right]^{-1} R(X'X)^{-1} R' \left[ R(X'X)^{-1} R \right]^{-1} R(X'X)^{-1}
\]
\[
= (X'X)^{-1} - (X'X)^{-1} R' \left[ R(X'X)^{-1} R \right]^{-1} R(X'X)^{-1}.
\]

**Maximum likelihood estimation under exact restrictions:**

Assuming \( \varepsilon \sim N(0, \sigma^2 I) \), the maximum likelihood estimator of \( \beta \) and \( \sigma^2 \) can also be derived such that it follows \( r = R\beta \). The Lagrangian function as per the maximum likelihood procedure can be written as

\[
L(\beta, \sigma^2, \lambda) = \left( \frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} \exp \left[ -\frac{1}{2} \left( \frac{(y - X\beta)(y - X\beta)}{\sigma^2} - \lambda' (R\beta - r) \right) \right]
\]

where \( \lambda \) is a \( (q \times 1) \) vector of Lagrangian multipliers. The normal equations are obtained by partially differentiating the log-likelihood function with respect to \( \beta, \sigma^2 \) and \( \lambda \) and equated to zero as

\[
\frac{\partial \ln L(\beta, \sigma^2, \lambda)}{\partial \beta} = -\frac{1}{\sigma^2} (X'X\beta - X'y) + 2R'\lambda = 0 \quad (1)
\]

\[
\frac{\partial \ln L(\beta, \sigma^2, \lambda)}{\partial \lambda} = 2(R\beta - r) = 0 \quad (2)
\]

\[
\frac{\partial \ln L(\beta, \sigma^2, \lambda)}{\partial \sigma^2} = -\frac{2n}{\sigma^2} + \frac{2(y - X\beta)'(y - X\beta)}{\sigma^4} = 0. \quad (3)
\]

Let \( \tilde{\beta}_r, \tilde{\sigma}_r^2 \) and \( \tilde{\lambda} \) denote the maximum likelihood estimators of \( \beta, \sigma^2 \) and \( \lambda \) respectively which are obtained by solving equations (1), (2) and (3) as follows:

From equation (1), we get optimal \( \lambda \) as

\[
\tilde{\lambda} = \left[ R(X'X)^{-1} R' \right]^{-1} (r - R\tilde{\beta}).
\]

Substituting \( \tilde{\lambda} \) in equation (1) gives

\[
\tilde{\beta}_r = \tilde{\beta} + (X'X)^{-1} R' \left[ R(X'X)^{-1} R \right]^{-1} (r - R\tilde{\beta}).
\]
where \( \hat{\beta} = (X'X)^{-1}X'y \) is the maximum likelihood estimator of \( \beta \) without restrictions. From equation (3), we get

\[
\hat{\sigma}_R^2 = \frac{(y-X\hat{\beta})'(y-X\hat{\beta})}{n}.
\]

The Hessian matrix of second order partial derivatives of \( \beta \) and \( \sigma^2 \) is positive definite at \( \beta = \hat{\beta}_R \) and \( \sigma^2 = \hat{\sigma}_R^2 \).

The restricted least squares and restricted maximum likelihood estimators of \( \beta \) are same whereas they are different for \( \sigma^2 \).

**Test of hypothesis**

It is important to test the hypothesis

\[
H_0 : r = R\beta \quad H_1 : r \neq R\beta
\]

before using it in the estimation procedure.

The construction of the test statistic for this hypothesis is detailed in the module on multiple linear regression model. The resulting test statistic is

\[
F = \frac{\left( (r-Rb)\left[ R(X'X)^{-1}R' \right]^{-1} (r-Rb) \right)}{q}
\]

\[
\frac{\left( (y-Xb)'(y-Xb) \right)}{n-k}
\]

which follows a \( F \)-distribution with \( q \) and \( (n-k) \) degrees of freedom under \( H_0 \). The decision rule is to reject \( H_0 \) at \( \alpha \) level of significance whenever

\[
F \geq F_{1-\alpha}(q, n-k).
\]
Stochastic linear restrictions:
The exact linear restrictions assume that there is no randomness involved in the auxiliary or prior information. This assumption may not hold true in many practical situations and some randomness may be present. The prior information in such cases can be formulated as

\[ r = R\beta + V \]

where \( r \) is a \((q \times 1)\) vector, \( R \) is a \((q \times k)\) matrix and \( V \) is a \((q \times 1)\) vector of random errors. The elements in \( r \) and \( R \) are known. The term \( V \) reflects the randomness involved in the prior information \( r = R\beta \).

Assume

\[
\begin{align*}
E(V) &= 0 \\
E(VV') &= \psi \\
E(\varepsilon V') &= 0.
\end{align*}
\]

where \( \psi \) is a known \((q \times q)\) positive definite matrix and \( \varepsilon \) is the disturbance term is multiple regression model \( y = X\beta + \varepsilon \).

Note that \( E(r) = R\beta \).

The possible reasons for such stochastic linear restriction are as follows:

(i) Stochastic linear restrictions exhibit the unstability of estimates. An unbiased estimate with the standard error may exhibit stability. For example, in repetitive studies, the surveys are conducted every year. Suppose the regression coefficient \( \beta_i \) remains stable for several years. Suppose its estimate is provided along with its standard error. Suppose its value remains stable around the value 0.5 with standard error 2. This information can be expressed as

\[ r = \beta_i + V_i, \]

where \( r = 0.5, E(V_i) = 0, E(V_i^2) = 2^2. \)

Now \( \psi \) can be formulated with this data. It is not necessary that we should have information for all regression coefficients but we can have information on some of the regression coefficients only.

(ii) Sometimes the restrictions are in the form of inequality. Such restrictions may arise from theoretical considerations. For example, the value of a regression coefficient may lie between 3 and 5, i.e., \( 3 \leq \beta_i \leq 5 \), say. In another example, consider a simple linear regression model

\[ y = \beta_0 + \beta_1 x + \varepsilon \]
where \( y \) denotes the consumption expenditure on food and \( x \) denotes the income. Then the marginal propensity (tendency) to consume is
\[
\frac{dy}{dx} = \beta_i,
\]
i.e., if salary increase by rupee one, then one is expected to spend \( \beta_i \), amount of rupee one on food or save \( (1-\beta_i) \) amount. We may put a bound on \( \beta \) that either one can not spend all of rupee one or nothing out of rupee one. So \( 0 < \beta_i < 1 \). This is a natural restriction arising from theoretical considerations.

These bounds can be treated as \( p - \sigma \) limits, say 2-sigma limits or confidence limits. Thus
\[
\mu - 2\sigma = 0
\]
\[
\mu + 2\sigma = 1
\]
\[
\Rightarrow \mu = \frac{1}{2}, \sigma = \frac{1}{4}.
\]
These values can be interpreted as
\[
\beta_i + V_1 = \frac{1}{2}
\]
\[
E(V_1^2) = \frac{1}{16}.
\]

(iii) Sometimes the truthfulness of exact linear restriction \( r = R\beta \) can be suspected and accordingly an element of uncertainty can be introduced. For example, one may say that 95% of the restrictions hold. So some element of uncertainty prevails.

**Pure and mixed regression estimation:**

Consider the multiple regression model
\[
y = X\beta + \epsilon
\]
with \( n \) observations and \( k \) explanatory variables \( X_1, X_2, \ldots, X_k \). The ordinary least squares estimator of \( \beta \) is
\[
b = (X'X)^{-1} X'y
\]
which is termed as **pure estimator**. The pure estimator \( b \) does not satisfy the restriction \( r = R\beta + V \). So the objective is to obtain an estimate of \( \beta \) by utilizing the stochastic restrictions such that the resulting estimator
satisfies the stochastic restrictions also. In order to avoid the conflict between prior information and sample information, we can combine them as follows:

Write

\[ y = X\beta + \varepsilon \quad E(\varepsilon) = 0, E(\varepsilon') = \sigma^2 I_n \]
\[ r = R\beta + V \quad E(V) = 0, E(VV') = \psi, E(\varepsilon V') = 0 \]

jointly as

\[
\begin{bmatrix} y \\ r \end{bmatrix} = \begin{bmatrix} X \\ R \end{bmatrix}\beta + \begin{bmatrix} \varepsilon \\ V \end{bmatrix}
\]

or \( a = A\beta + w \)

where \( a = (y \quad r)', \quad A = (X \quad R)', \quad w = (\varepsilon \quad V)' \).

Note that

\[
E(w) = \begin{bmatrix} E(\varepsilon) \\ E(V) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\]
\[
\Omega = E(ww') = E\begin{bmatrix} \varepsilon\varepsilon' & \varepsilon V' \\ V\varepsilon' & VV' \end{bmatrix} = \begin{bmatrix} \sigma^2 I_n & 0 \\ 0 & \psi \end{bmatrix}.
\]

This shows that the disturbances \( w \) are non spherical or heteroskedastic. So the application of generalized least squares estimation will yield more efficient estimator than ordinary least squares estimation. So applying generalized least squares to the model

\[ a = AB + w \quad E(w) = 0, V(w) = \Omega, \]

the generalized least square estimator of \( \beta \) is given by

\[
\hat{\beta}_M = \left( A'\Omega^{-1}A \right)^{-1} A'\Omega^{-1}a.
\]

The explicit form of this estimator is obtained as follows:
\[
A’\Omega^{-1}a = [X’ \ R’] \begin{bmatrix}
\frac{1}{\sigma^2} I_n & 0 \\
0 & \Psi^{-1}
\end{bmatrix} \begin{bmatrix}
y’ \\
r’
\end{bmatrix}
\]

\[
= \frac{1}{\sigma^2} X’y + k’\Psi^{-1}r
\]

\[
A’\Omega^{-1}A = [X’ \ R’] \begin{bmatrix}
\frac{1}{\sigma^2} I_n & 0 \\
0 & \Psi^{-1}
\end{bmatrix} \begin{bmatrix}
X’ \\
R’
\end{bmatrix}
\]

\[
= \frac{1}{\sigma^2} X’X + R’\Psi^{-1}R.
\]

Thus

\[
\hat{\beta}_M = \left( \frac{1}{\sigma^2} X’X + R’\Psi^{-1}R \right)^{-1} \left( \frac{1}{\sigma^2} X’y + R’\Psi^{-1}r \right)
\]

assuming \( \sigma^2 \) to be unknown. This is termed as mixed regression estimator.

If \( \sigma^2 \) is unknown, then \( \sigma^2 \) can be replaced by its estimator \( \hat{\sigma}^2 = s^2 = \frac{1}{n-k} (y-Xb)’(y-Xb) \) and feasible mixed regression estimator of \( \beta \) is obtained as

\[
\hat{\beta}_f = \left( \frac{1}{s^2} X’X + R’\Psi^{-1}R \right)^{-1} \left( \frac{1}{s^2} X’y + R’\Psi^{-1}r \right)
\]

This is also termed as estimated or operationalized generalized least squares estimator.

Properties of mixed regression estimator:

(i) Unbiasedness:

The estimation error of \( \hat{\beta}_m \) is

\[
\hat{\beta}_m - \beta = (A’\Omega A)^{-1} A’\Omega^{-1}a - \beta
\]

\[
= (A’\Omega^{-1} A)^{-1} A’\Omega^{-1} (AB + w) - \beta
\]

\[
= (A’\Omega^{-1} A)^{-1} A’\Omega^{-1}w.
\]

\[
E \left( \hat{\beta}_m - \beta \right) = (A’\Omega^{-1} A)^{-1} A’\Omega^{-1}E(w)
\]

\[
= 0.
\]

So mixed regression estimator provides an unbiased estimator of \( \beta \). Note that the pure regression \( b = (X’X)^{-1} X’y \) estimator is also an unbiased estimator of \( \beta \).
(ii) Covariance matrix

The covariance matrix of $\hat{\beta}_M$ is

\[
V(\hat{\beta}_M) = E(\hat{\beta}_M - \beta)(\hat{\beta}_M - \beta)'
\]
\[
= (A'\Omega^{-1}A)^{-1} A'\Omega^{-1} E(VV')\Omega^{-1} A (A'\Omega^{-1} A)^{-1}
\]
\[
= (A'\Omega^{-1} A)^{-1}
\]
\[
= \left( \frac{1}{\sigma^2} X'X + R'\Omega^{-1}R \right)^{-1}.
\]

(iii) The estimator $\hat{\beta}_M$ satisfies the stochastic linear restrictions in the sense that

\[
r = R\hat{\beta}_M + V
\]
\[
E(r) = RE(\hat{\beta}_M) + E(V)
\]
\[
= R\beta + 0
\]
\[
= R\beta.
\]

(iv) Comparison with OLSE

We first state a result that is used further to establish the dominance of $\hat{\beta}_M$ over $b$.

Result: The difference of matrices $(A_1^{-1} - A_2^{-1})$ is positive definite if $(A_2 - A_1)$ is positive definite.

Let

\[
A_1 \equiv V(b) = \sigma^2 (X'X)^{-1}
\]
\[
A_2 \equiv V(\hat{\beta}_M) = \left( \frac{1}{\sigma^2} X'X + R'\Psi^{-1}R \right)^{-1}
\]

then $A_1^{-1} - A_2^{-1} = \frac{1}{\sigma^2} X'X + R'\Psi^{-1}R - \frac{1}{\sigma^2} X'X
\]
\[
= R'\Psi^{-1}R
\]

which is a positive definite matrix. This implies that

\[
A_1 - A_2 = V(b) - V(\hat{\beta}_M)
\]

is a positive definite matrix. Thus $\hat{\beta}_M$ is more efficient than $b$ under the criterion of covariance matrices or Loewner ordering provided $\sigma^2$ is known.
Testing of hypothesis:
In the prior information specified by stochastic restriction \( r = R\beta + V \), we want to test whether there is close relation between the sample information and the prior information. The test for the compatibility of sample and prior information is tested by \( \chi^2 \) - test statistic given by

\[
\chi^2 = \frac{1}{\sigma^2}(r - Rb)\left[R\left(X'X\right)^{-1}R' + \Psi\right]^{-1}(r - Rb)
\]

assuming \( \sigma^2 \) is known and \( b = \left(X'X\right)^{-1}X'y \). This follows a \( \chi^2 \)-distribution with \( q \) degrees of freedom.

If \( \Psi = 0 \), then the distribution is degenerated and hence \( r \) becomes a fixed quantity. For the feasible version of mixed regression estimator

\[
\hat{\beta}_f = \left(\frac{1}{s^2}X'X + R^i\Psi^{-1}R\right)^{-1}\left(\frac{1}{s^2}X'y + R^i\Psi^{-1}r\right),
\]

the optimal properties of mixed regression estimator like linearity unbiasedness and/or minimum variance do not remain valid. So there can be situations when the incorporation of prior information may lead to loss in efficiency. This is not a favourable situation. Under such situations, the pure regression estimator is better to use. In order to know whether the use of prior information will lead to better estimator or not, the null hypothesis \( H_0 : E(r) = R\beta \) can be tested.

For testing the null hypothesis

\( H_0 : E(r) = R\beta \)

when \( \sigma^2 \) is unknown, we use the \( F \) -statistic given by

\[
F = \frac{\left[(r - Rb)^\prime\left[R\left(X'X\right)^{-1}R' + \Psi\right]^{-1}(r - Rb)\right]/q}{s^2}
\]

where \( s^2 = \frac{1}{n-k}(y - Xb)'(y - Xb) \) and \( F \) follows a \( F \)-distribution with \( q \) and \((n-k)\) degrees of freedom under \( H_0 \).
Inequality Restrictions

Sometimes the restriction on the regression parameters or equivalently the prior information about the regression parameters is available in the form of inequalities. For Example, \( 1 < \beta_1 < 2, 5 < \beta_3 < 6, 2 < \beta_1 + 2\beta_2 < 5 \), etc. Suppose such information is expressible in the form of \( R\beta \geq r \). We want to estimate the regression coefficient \( \beta \) in the model \( y = X\beta + \epsilon \) subject to constraints \( R\beta \geq r \).

One can minimize \((y - X\beta)'(y - X\beta)\) subject to \( R\beta \geq r \) to obtain an estimator of \( \beta \). This can be formulated as a quadratic programming problem and can be solved using an appropriate algorithm, e.g., Simplex algorithm and a numerical solution is obtained. The advantage of this procedure is that a solution \( \hat{\beta} \) is found that fulfills the condition. The disadvantage is that the statistical properties of the estimates are not easily determined and no general conclusions about superiority can be made.

Another option to obtain an estimator of \( \beta \) is subject to inequality constraints is to convert the inequality constraints in the form of stochastic linear restrictions e.g., \( p - \sigma \text{a} \) limits, and use the framework of mixed regression estimation.

The minimax estimation can also be used to obtain the estimator of \( \beta \) under inequality constraints. The minimax estimation is based on the idea that the quadratic risk function for the estimate \( \hat{\beta} \) is not minimized over the entire parameter space but only over an area that is restricted by the prior knowledge or restrictions in relation to the estimate.

If all the restriction define a convex area, this area can be enclosed in an ellipsoid of the following form

\[
B(\beta) = \{ \beta : \beta'T\beta \leq k \}
\]

with the origin as center point or in

\[
B(\beta, \beta_0) = \{ \beta : (\beta - \beta_0)'(\beta - \beta_0) \leq k \}
\]

with the center point vector \( \beta_0 \) where \( k > 0 \) is a given constant and \( T \) is a known \( p \times p \) matrix which is assumed to be positive definite. Here \( B \) defines a concentration ellipsoid.

First we consider an example to understand how the inequality constraints are framed. Suppose it is known a priori that

\[
a_i \leq \beta_i \leq b_i (i = 1, 2, \ldots, n)
\]
when \( a_i \) and \( b_i \) \( (i = 1, 2, \ldots, n) \) are known and may include \( a_i = -\infty \) and \( b_i = \infty \). These restriction can be written as

\[
\left| \frac{\beta_i - a_i + b_i}{2} \right| \leq 1, \quad i = 1, 2, \ldots, n.
\]

Now we want to construct a concentration ellipsoid \((\beta - \beta_0)^T (\beta - \beta_0) = 1\) which encloses the cuboid and fulfills the following conditions:

(i) The ellipsoid and the cuboid have the same center point, \( \beta_0 = \frac{1}{2} (a_1 + b_1, \ldots, a_p + b_p) \).

(ii) The axes of the ellipsoid are parallel to the coordinate axes, that is, \( T = \text{diag}(t_1, \ldots, t_p) \).

(iii) The corner points of the cuboid are on the surface of the ellipsoid, which means we have

\[
\sum_{i=1}^{p} \left( \frac{a_i - b_i}{2} \right)^2 t_i = 1.
\]

(iv) The ellipsoid has minimal volume:

\[
V = c_p \prod_{i=1}^{p} t_i^{-\frac{1}{2}},
\]

with \( c_p \) being a constant dependent on the dimension \( p \).

We now include the linear restriction (iii) for the \( t_i \) by means of Lagrangian multipliers \( \lambda \) and solve (with \( c_p^{-2} V_p^2 = \Pi t_i^{-1} \))

\[
\min_{\{ t_i \}} \min_{\lambda} \left\{ \prod_{i=1}^{p} t_i^{-1} - \lambda \left[ \sum_{i=1}^{p} \left( \frac{a_i - b_i}{2} \right)^2 t_i - 1 \right] \right\}.
\]

The normal equations are then obtained as

\[
\frac{\partial \tilde{V}}{\partial t_i} = -t_i^{-2} \prod_{i<j} t_i^{-1} - \lambda \left( \frac{a_j - b_j}{2} \right)^2 = 0
\]

and

\[
\frac{\partial \tilde{V}}{\partial \lambda} = \sum \left( \frac{a_j - b_j}{2} \right)^2 t_i - 1 = 0.
\]
From $\frac{\partial \hat{V}}{\partial t_i} = 0$, we get

\[ \lambda = -t_i^2 \prod_{i=1}^{p} t_i \left( \frac{2}{a_j-b_j} \right)^2 \quad \text{(for all } j = 1, 2, \ldots, p) \]

\[ = -t_i^2 \prod_{i=1}^{p} t_i \left( \frac{2}{a_j-b_j} \right)^2, \]

and for any two $i, j$ we obtain

\[ t_i \left( \frac{a_j-b_j}{2} \right)^2 = t_j \left( \frac{a_j-b_j}{2} \right)^2, \]

and hence—after summation—according to $\frac{\partial \hat{V}}{\partial \lambda} = 0$ gives

\[ \sum_{i=1}^{p} \left( \frac{a_j-b_j}{2} \right)^2 t_j = pt_j \left( \frac{a_j-b_j}{2} \right)^2 = 1. \]

This leads to the required diagonal elements of $T$:

\[ t_j = \frac{4}{p} \left( \frac{a_j-b_j}{2} \right)^2 \quad \text{(j = 1, 2, \ldots, p)}. \]

Hence, the optimal ellipsoid $(\beta - \beta_0)'T(\beta - \beta_0) = 1$, which contains the cuboid, has the center point vector

\[ \beta_0' = \frac{1}{2} (a_1+b_1, \ldots, a_p+b_p) \]

and the following matrix, which is positive definite for finite limits $a_i, b_i (a_i \neq b_i)$,

\[ T = \text{diag} \left( \frac{4}{p} \left( b_1-a_1 \right)^{-2}, \ldots, \left( b_p-a_p \right)^{-2} \right). \]

**Interpretation:** The ellipsoid has a larger volume than the cuboid. Hence, the transition to an ellipsoid as a priori information represents a weakening, but comes with an easier mathematical handling.
Example: (Two real regressors) The center-point equation of the ellipsoid is (see Figure)

\[
\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,
\]

or

\[
(x, y) \begin{pmatrix}
\frac{1}{a^2} & 0 \\
0 & \frac{1}{b^2}
\end{pmatrix}
\begin{pmatrix}
x \\
y
\end{pmatrix} = 1
\]

with \(T = \text{diag} \left( \frac{1}{a^2}, \frac{1}{b^2} \right) = \text{diag} \left( t_1, t_2 \right)\)

and the area \(F = \pi ab = \pi t_1 t_2\).

The Minimax Principle:

Consider the quadratic risk \(R(\hat{\beta}, \beta, A) = \text{tr} \left[ AE \left( \hat{\beta} - \beta \right) \left( \hat{\beta} - \beta \right)' \right]\) and a class \(\{\hat{\beta}\}\) of estimators. Let \(B(\beta) \subseteq \mathbb{R}^p\) be a convex region of a priori restrictions for \(\beta\). The criterion of the minimax estimator leads to the following.

**Definition**: An estimator \(b^* \in \{\hat{\beta}\}\) is called a minimax estimator of \(\beta\) if

\[
\min_{\{\hat{\beta}\}} \sup_{\beta \in B} R(\hat{\beta}, \beta, A) = \sup_{\beta \in B} R(\hat{\beta}, \beta, A).
\]

An explicit solution can be achieved if the weight matrix is of the form \(A = aa'\) of rank 1.

Using the abbreviation \(D_s = \left( S + k^{-1} \sigma^2 T \right)\), we have following result:

**Result**: In the model \(y = X\beta + \varepsilon, \varepsilon \sim (0, \sigma^2 I)\), with the restriction \(\beta' T \beta \leq k\) with \(T > 0\), and the risk function \(R(\hat{\beta}, \beta, a)\), the linear minimax estimator is of the following form:

\[
b_s = \left( X' X + k^{-1} \sigma^2 T \right)^{-1} X'y
\]

\[
= D_s^{-1} X'y
\]

with the bias vector and covariance matrix as
\[ \text{Bias} (b_*, \beta) = -k^{-1} \sigma^2 D_r^{-1} T \beta, \]
\[ V (b_*) = \sigma^2 D_r^{-1} S D_r^{-1} \]

and the minimax risk is
\[ \sup_{\beta^T \beta \leq k} R (b_*, \beta, a) = \sigma^2 a' D_r^{-1} a. \]

**Result:** If the restrictions are \((\beta - \beta_0)^T (\beta - \beta_0) \leq k\) with center point \(\beta_0 \neq 0\), the linear minimax estimator is of the following form:
\[ b_*(\beta_0) = \beta_0 + D_r^{-1} X' (y - X \beta_0) \]

with bias vector and covariance matrix as
\[ \text{Bias} (b_*(\beta_0), \beta) = -k^{-1} \sigma^2 D_r^{-1} T (\beta - \beta_0), \]
\[ V (b_*(\beta_0)) = V (b_*), \]

and the minimax risk is
\[ \sup_{(\beta - \beta_0)^T (\beta - \beta_0) \leq k} R (b_*(\beta_0), \beta, a) = \sigma^2 a' D_r^{-1} a. \]

**Interpretation:** A change of the center point of the a priori ellipsoid has an influence only on the estimator itself and its bias. The minimax estimator is not operational because \(\sigma^2\) is unknown. The smaller the value of \(k\), the stricter is the a priori restriction for fixed \(T\). Analogously, the larger the value of \(k\), the smaller is the influence of \(\beta^T \beta \leq k\) on the minimax estimator. For the borderline case we have
\[ B (\beta) = \{ \beta : \beta^T \beta \leq k \} \rightarrow \mathbb{R}^k \quad \text{as} \quad k \rightarrow \infty \]
and
\[ \lim_{k \rightarrow \infty} b_* \rightarrow b = (X'X)^{-1} X'y. \]

**Comparison of \(b_*\) and \(b\):**

**Minimax Risk:** Since the OLS estimator is unbiased, its minimax risk is
\[ \sup_{\beta^T \beta \leq k} R (b_*, a) = \sigma^2 a' S^{-1} a. \]

The linear minimax estimator \(b_*\) has a smaller minimax risk than the OLS estimator, and
\[ R (b_*, a) - \sup_{\beta^T \beta \leq k} R (b_*, \beta, a) \]
\[ = \sigma^2 a' (S^{-1} - \left( k^{-1} \sigma^2 T + S \right)^{-1}) a \geq 0, \]
since \( S^{-1} - \left( k^{-1} \sigma^2 T + S \right)^{-1} \geq 0 \)

Considering the superiority with MSE matrices, we get

\[
M(b_*, \beta) = V(b_*) + Bias(b_*, \beta) Bias(b_*, \beta)' = \sigma^2 D_s^{-1} \left( S + k^{-2} \sigma^2 T \beta \beta' T' \right) D_s^{-1}
\]

Hence, \( b_s \) is superior to \( b \) under the criterion of Loewner ordering when

\[
\Delta(b, b_s) = V(b_s) - M(b, \beta) = \sigma^2 D_s^{-1} [D_s S^{-1} D_s - S - k^{-2} \sigma^2 T \beta \beta' T' ] D_s^{-1} \geq 0,
\]

which is possible if and only if

\[
B = D_s S^{-1} D_s - S - k^{-2} \sigma^2 T \beta \beta' T' = k^{-2} \sigma^2 T \left[ \left\{ S^{-1} + 2k \sigma^{-2} T^{-1} \right\} - \sigma^{-2} \beta \beta' T' \right] T \geq 0
\]

\[
= k^{-2} \sigma^2 T C_k \left[ I - \sigma^{-2} C^{-\frac{1}{2}} \beta \beta' C^{-\frac{1}{2}} \right] C^{-\frac{1}{2}} T \geq 0
\]

with \( C = S^{-1} + 2k \sigma^{-2} T^{-1} \). This is equivalent to

\[
\sigma^{-2} \beta' (S^{-1} + 2k \sigma^{-2} T^{-1}) \geq 0.
\]

Since \( \left( 2k \sigma^{-2} T^{-1} \right)^{-1} - \left( S^{-1} + 2k \sigma^{-2} T^{-1} \right) \geq 0 \),

\[
k^{-1} \leq \frac{2}{\beta' \beta}.
\]

**Preliminary Test Estimation:**

The statistical modeling of the data is usually done assuming that the model is correctly specified and the correct estimators are used for the purpose of estimation and drawing statistical inferences form a sample of data. Sometimes the prior information or constraints are available from outside the sample as non-sample information. The incorporation and use of such prior information along with the sample information leads to more efficient estimators provided it is correct. So the suitability of the estimator lies on the correctness of prior information. One possible statistical approach to check the correctness of prior information is through the framework of test of hypothesis. For example, if prior information is available in the form of exact linear restrictions \( r = R \theta \), there are two possibilities- either it is correct or incorrect. If the information is correct, then \( r = R \theta \) holds true in the model \( y = X \theta + \epsilon \) and then the restricted regression estimator (RRE) \( \beta_R = b + (X' X)^{-1} R [R (X' X)^{-1} R]^{-1} (R b - r) \) is used which is more efficient than OLSE \( b = (X' X)^{-1} X' y \) if \( \theta \). Moreover, RRE satisfies the restrictions, i.e. \( R \beta_R = r \). On the other hand, when
the information is incorrect, i.e., \( r \neq R\beta \), then OLSE is better than RRE. The truthfulness of prior information in terms of \( r = R\beta \) or \( r \neq R\beta \) is tested by the null hypothesis \( H_0: R\beta = r \) using the \( F \)-statistics.

- If \( H_0 \) is accepted at \( \alpha \) level of significance, then we conclude that \( R\beta = r \) and in such a situation, RRE is better than OLSE.
- On the other hand, if \( H_0 \) is rejected at \( \alpha \) level of significance, then we conclude that \( R\beta \neq r \) and OLSE is better than RRE under such situations.

So when the exact content of the true sampling model is unknown, then the statistical model to be used is determined by a preliminary test of hypothesis using the available sample data. Such procedures are completed in two stages and are based on a test of hypothesis which provides a rule for choosing between the estimator based on the sample data and the estimator is consistent with the hypothesis. This requires to make a test of the compatibility of OLSE (or maximum likelihood estimator) based on sample information only and RRE based on the linear hypothesis. The one can make a choice of estimator depending upon the outcome. Consequently, one can choose OLSE or RRE. Note that under the normality of random errors, the equivalent choice is made between the maximum likelihood estimator of \( \beta \) and the restricted maximum likelihood estimator of \( \beta \), which has the same form as OLSE and RRE, respectively. So essentially a pre-test of hypothesis is done for \( H_0: R\beta = r \) and based on that, a suitable estimator is chosen. This is called the pre-test procedure which generates the pre-test estimator that in turn, provides a rule to choose between restricted or unrestricted estimators.

One can also understand the philosophy behind the preliminary test estimation as follows. Consider the problem of an investigator who has a single data set and wants to estimate the parameters of a linear model that are known to lie in a high dimensional parametric space \( \Theta_2 \). However, the prior information about the parameter is available and it suggests that the relationship may be characterized by a lower dimensional parametric space \( \Theta_2 \subset \Theta_2 \). Under such uncertainty, if the parametric space \( \Theta_1 \) is estimated by OLSE, the result from the over specified model will be unbiased but will have larger variance. Alternatively, the parametric space \( \Theta_2 \) may incorrectly specify the statistical model and if estimated by OLSE will be biased. The bias may or may not overweigh the reduction in variance. If such uncertainty is represented in the form of general linear hypothesis, this leads to pre-test estimators.
Let us consider the conventional pre-test estimator under the model \( y = X\beta + \epsilon \) with usual assumptions and the general linear hypothesis \( H_0: R\beta = r \) which can be tested by using \( F \) statistics. The null hypothesis \( H_0 \) is rejected at \( \alpha \) level of significance when
\[
\lambda = F_{\text{calculated}} \geq F_{\alpha, p, n-p} = c
\]
where the critical value \( c \) is determined for given level of the test \( \alpha \) by
\[
\int_c^\infty F_{p, n-p} = P\left[ F_{p, n-p} \geq c \right] = \alpha .
\]

- If \( H_0 \) is true, meaning thereby that the prior information is correct, then use RRE \( \hat{\beta}_R = b + (X'X)^{-1}R[R(X'X)^{-1}R]^{-1}(Rb - r) \) to estimate \( \beta \).
- If \( H_0 \) is false, meaning thereby that the prior information is incorrect, then use OLSE \( b = (X'X)^{-1}X'y \) to estimate \( \beta \).

Thus the estimator to be used depends on the preliminary test of significance and is of the form
\[
\hat{\beta}_{PT} = \begin{cases} \hat{\beta}_R & \text{if } u < c \\ b & \text{if } u \geq c. \end{cases}
\]

This estimator \( \hat{\beta}_{PT} \) is called as preliminary test or pre-test estimator of \( \beta \). Alternatively,
\[
\hat{\beta}_{PT} = \hat{\beta}_R I_{(0, c)}(u) + b I_{(c, \infty)}(u) \\
= \hat{\beta}_R I_{(0, c)}(u) + \left[1 - I_{(0, c)}(u)\right] b \\
= b - (b - \hat{\beta}_R) I_{(0, c)}(u) \\
= b - (b - r) I_{(0, c)}(u)
\]

where the indicator functions are defined as
\[
I_{(0, c)}(u) = \begin{cases} 1 & \text{when } 0 < u < c \\ 0 & \text{otherwise} \end{cases}
\]
\[
I_{(c, \infty)}(u) = \begin{cases} 1 & \text{when } u \geq c \\ 0 & \text{otherwise}. \end{cases}
\]

- If \( \alpha = 0 \), then \( \hat{\beta}_{PT} = \hat{\beta}_R I_{(0, c)}(u) + b I_{(c, \infty)}(u) = \hat{\beta}_R \).
- If \( \alpha = 1 \), then \( \hat{\beta}_{PT} = \hat{\beta}_R I_{(0)}(u) + b I_{(0, \infty)}(u) = b \).

Note that \( \alpha = 0 \) and \( \alpha = 1 \) indicate that the probability of type 1 error (i.e., rejecting \( H_0 \) when it is true) is 0 and 1 respectively. So the entire area under the sampling distribution is the area of acceptance or the area of rejection of null hypothesis. Thus the choice of \( \alpha \) has a crucial role to play in determining the sampling
performance of the pre-test estimators. Therefore in a repeated sampling context, the data, the linear hypothesis, and the level of significance all determine the combination of the two estimators that are chosen on the average. The level of significance has an impact on the outcome of pretest estimator in the sense of determining the proportion of the time each estimator is used and in determining the sampling performance of pretest estimator.

We use the following result to derive the bias and risk of pretest estimator:

**Result 1**: If the \((K \times 1)\) random vector, \(Z/\sigma\), is distributed as a multivariate normal random vector with mean \(\delta/\sigma\) and covariance matrix \(I_k\) and is independent of \(\chi^2_{(n-K)}\) then

\[
E \left[ I_{(0,c)} \left( \frac{Z'Z}{\sigma^2 \chi^2_{(n-K)}} \right) \right] = \left( \frac{\delta}{\sigma} \right) P \left[ \frac{\chi^2_{(K+2,\lambda)}}{\chi^2_{(n-K)}} \leq c^* \right],
\]

where \(\lambda = \delta^2/2\sigma^2\) and \(c^* = cK/(n-K)\).

**Result 2**: If the \(K \times 1\) random vector, \(Z/\sigma\), is distributed as a multivariate normal random vector with mean \(\delta/\sigma\) and covariance \(I_k\) and is independent of \(\chi^2_{(n-K)}\) then

\[
E \left[ I_{(0,c^*)} \left( \frac{Z'Z}{\sigma^2 \chi^2_{(n-K)}} \right) \right] = KE \left[ I_{(0,c^*)} \left( \frac{\chi^2_{(K+2,\delta^2/2\sigma^2)}}{\chi^2_{(n-K)}} \right) \right] + \frac{\delta^2}{\sigma^2} E \left[ I_{(0,c^*)} \frac{\chi^2_{(K+4,\delta^2/2\sigma^2)}}{\chi^2_{(n-K)}} \right]
\]

\[= KP \left[ \frac{\chi^2_{(K+2,\delta^2/2\sigma^2)}}{\chi^2_{(n-K)}} \leq \frac{cK}{T-K} \right] + \frac{\delta^2}{\sigma^2} P \left[ \frac{\chi^2_{(K+4,\delta^2/2\sigma^2)}}{\chi^2_{(n-K)}} \leq \frac{cK}{n-K} \right],
\]

where \(c^* = cK/(n-K)\).

Using these results, we can find the bias and risk of \(\hat{\beta}_{PT}\) as follows:

**Bias:**

\[
E \left( \hat{\beta}_{PT} \right) = E (b) - E \left[ I_{(0,c)} \left( u \right) \left( b - r \right) \right] = \beta - \delta P \left[ \frac{\chi^2_{(p+2,\delta^2/2\sigma^2)}}{\chi^2_{(n-p,\delta^2/2\sigma^2)}} \leq \frac{c}{n-p} \right]
\]

\[= \beta - \delta \left[ F \left( p+2, \frac{n-p,\delta^2/2\sigma^2}{c} \right) \right].
\]
where \( \delta = r - R\beta, \chi^2_{(p+2, \frac{\delta'\delta}{2\sigma^2})} \) denotes the non-central \( \chi^2 \) distribution with noncentrality parameter \( \delta'\delta/2\sigma^2 \).

\( \delta'\delta/2\sigma^2, F_{(p+2,n-p, \delta'\delta/2\sigma^2)} \) denotes the non-central \( F \) distribution with noncentrality parameter \( \delta'\delta/2\sigma^2 \).

Thus, if \( \delta = 0 \), the pretest estimator is unbiased. Note that the size of bias is affected by the probability of a random variable with a non-central \( F \)-distribution being less than a constant that is determined by the level of the test, the number of hypothesis and the degree of hypothesis error \( \delta \). Since the probability is always less than or equal to one, so \( bias(\hat{\beta}_{PT}) \leq bias(b) \).

**Risk:**

The risk of pretest estimator is obtained as

\[
\rho(\beta, \hat{\beta}_{PT}) = E\left[ (\hat{\beta}_{PT} - \beta)'(\hat{\beta}_{PT} - \beta) \right] \\
= E\left[ (b - \beta - I_{(0,c)}(u)(b-r))'(b - \beta - I_{(0,c)}(u)(b-r)) \right] \\
= E\left[ (b - \beta)'(b - \beta) \right] - E\left[ I_{(0,c)}(u)(b - \beta)'(b - \beta) \right] + E\left[ I_{(0,c)}(u) \right] \delta'\delta \\
= \sigma^2 p + (2\delta'\delta - \sigma^2 p)P \left[ \chi^2_{(p+2, \frac{\delta'\delta}{2\sigma^2})} \leq \frac{cp}{n-p} \right] - \delta'\delta P \left[ \chi^2_{(n-p)} \leq \frac{cp}{n-p} \right]
\]

or compactly,

\[
\rho(\beta, \hat{\beta}_{PT}) = \sigma^2 p + (2\delta'\delta - \sigma^2 p)l(2) - \delta'\delta l(4)
\]

where

\[
l(2) = \frac{\chi^2_{(p+2, \frac{\delta'\delta}{2\sigma^2})}}{\chi^2_{(n-p)}}, \quad l(4) = \frac{\chi^2_{(p+4, \frac{\delta'\delta}{2\sigma^2})}}{\chi^2_{(n-p)}} \quad \frac{1}{2}, \quad 0 < l(4) < l(2) < 1.
\]

The risk function implies the following results:

1. If the restrictions are correct and \( (\delta = 0) \), the risk of the pretest estimator \( \sigma^2 K[1 - l(2)] \), where \( 1 > [1 - l(2)] > 0 \) for \( 0 < c < \alpha \). Therefore, the pretest estimator has risk less than that of the least squares estimator at the origin \( (\delta = 0) \) and the decrease in risk depends on the level of significance \( \alpha \) and, correspondingly, on the critical value of the test \( c \).

2. As the hypothesis error \( \beta - r = \delta \), and thus \( \delta'\delta/2\sigma^2 \), increases and approaches infinity, \( l(\cdot) \) and \( \delta'\delta l(\cdot) \) approach zero. The risk of the pretest estimator therefore approaches \( \sigma^2 p \), the risk of the unrestricted least squares estimator.
3. As the hypothesis error grows, the risk of the pretest estimator increases obtains a maximum after crossing the risk of the least squares estimator, and then monotonically decreases to approach $\sigma^2 p$, the risk of the OLSE.

4. The pretest estimator risk function defined on the $\delta'\delta/2\sigma^2$ parameter spaces crosses the risk function of the least squares estimator within the bounds $p/4 \leq \delta'\delta/2\sigma^2 \leq p/2$.

The sampling characteristic of the preliminary test estimator are summarized in Figure 1.

From these results we see that the pretest estimator does well relative to OLSE if the hypothesis is correctly specified. However, in the $\delta'\delta/2\sigma^2$ space representing the range of hypothesis are correctly specified. However, in the $\delta'\delta/2\sigma^2$ space representing the range of hypothesis errors, the pretest estimator is inferior to the least squares estimator over an infinite range of the parameter space. In figures 1 and 2, there is a range of the parameter space which the pretest estimator has risk that is inferior to (greater than) that of both the unrestricted and restricted least squares estimators. No one estimator depicted in Figure 1 dominates the other competitors. In addition, in applied problems the hypothesis errors, and thus the correct in the specification error parameter space, are seldom known. Consequently, the choice of the estimator is unresolved.

**The Optimal Level of Significance**

The form of the pretest estimator involves, for evaluation purposes, the probabilities of ratios of random variables $l(2)$ and $l(4)$ being less than a constant that depends on the critical value of the test $c$ or on the level of statistical significance $\alpha$. Thus as $\alpha \to 0$, the probabilities $l(\cdot) \to 1$, and the risk of the pretest estimator approaches that of the restricted regression estimator $\hat{\beta}_R$. In contrast, as $\alpha \to 1$, $l(\cdot)$ approaches zero and the
risk of the pretest estimator approaches that of the least squares estimator $b$. The choice of $c_0 \alpha$, which has a crucial impact on the performance of the pretest estimator, is portrayed in Figure 3.

Since the investigator is usually unsure of the degree of hypothesis specification error, and thus is unsure of the appropriate point in the $\delta$ space for evaluating the risk, the best of worlds would be to have a rule that mixes the unrestricted and restricted estimators so as to minimize risk regardless of the relevant specification error $\delta' \delta / 2 \sigma^2$. Thus the risk function traced out by the cross-hatched area in Figure 2 is relevant. Unfortunately, the risk of the pretest estimator, regardless of the choice of $\alpha$, is always equal to or greater than the minimum risk function for some range of the parameter space. Given this result, one criterion that has been proposed for choosing the $\alpha$ level might be to choose the critical value $c$ that would minimize the maximum regret of not being on the minimum risk function, reflected by the boundary of the shaded area. Another criterion that has been proposed for choosing $\alpha$ is to minimize the average regret over the whole $\delta' \delta / 2 \sigma^2$ space. Each of these criteria lead to different conclusions or rules for choice, and the question concerning the optimal level of the test is still open. One thing that is apparent is that conventional choices of 0.05 and 0.01 may have rather severe statistical consequences.
Chapter 7
Multicollinearity

A basic assumption in multiple linear regression model is that the rank of the matrix of observations on explanatory variables is same as the number of explanatory variables. In other words, such matrix is of full column rank. This in turn implies that all the explanatory variables are independent, i.e., there is no linear relationship among the explanatory variables. It is termed that the explanatory variables are orthogonal.

In many situations in practice, the explanatory variables may not remain independent due to various reasons. The situation where the explanatory variables are highly intercorrelated is referred to as multicollinearity.

Consider the multiple regression model

\[ y = X \beta + \varepsilon, \; \varepsilon \sim N(0, \sigma^2 I) \]

with \( k \) explanatory variables \( X_1, X_2, \ldots, X_k \) with usual assumptions including \( \text{Rank}(X) = k \).

Assume the observations on all \( X_i \)'s and \( y_i \)'s are centered and scaled to unit length. So

- \( X'X \) becomes a \( k \times k \) matrix of correlation coefficients between the explanatory variables and
- \( X'y \) becomes a \( k \times 1 \) vector of correlation coefficients between explanatory and study variables.

Let \( X = [X_1, X_2, \ldots, X_k] \) where \( X_j \) is the \( j^{th} \) column of \( X \) denoting the \( n \) observations on \( X_j \). The column vectors \( X_1, X_2, \ldots, X_k \) are linearly dependent if there exists a set of constants \( \ell_1, \ell_2, \ldots, \ell_k \), not all zero, such that

\[ \sum_{j=1}^{k} \ell_j X_j = 0. \]

If this holds exactly for a subset of the \( X_1, X_2, \ldots, X_k \), then \( \text{rank}(X'X) < k \). Consequently \( (X'X)^{-1} \) does not exist. If the condition \( \sum_{j=1}^{k} \ell_j X_j = 0 \) is approximately true for some subset of \( X_1, X_2, \ldots, X_k \), then there will be a near-linear dependency in \( X'X \). In such a case, the multicollinearity problem exists. It is also said that \( X'X \) becomes ill-conditioned.
Source of multicollinearity:

1. Method of data collection:
   It is expected that the data is collected over the whole cross-section of variables. It may happen that the data is collected over a subspace of the explanatory variables where the variables are linearly dependent. For example, sampling is done only over a limited range of explanatory variables in the population.

2. Model and population constraints
   There may exist some constraints on the model or on the population from where the sample is drawn. The sample may be generated from that part of population having linear combinations.

3. Existence of identities or definitional relationships:
   There may exist some relationships among the variables which may be due to the definition of variables or any identity relation among them. For example, if data is collected on the variables like income, saving and expenditure, then income = saving + expenditure. Such relationship will not change even when the sample size increases.

4. Imprecise formulation of model
   The formulation of the model may unnecessarily be complicated. For example, the quadratic (or polynomial) terms or cross product terms may appear as explanatory variables. For example, let there be 3 variables $X_1, X_2$ and $X_3$, so $k = 3$. Suppose their cross-product terms $X_1X_2, X_2X_3$ and $X_1X_3$ are also added. Then $k$ rises to 6.

5. An over-determined model
   Sometimes, due to over enthusiasm, large number of variables are included in the model to make it more realistic and consequently the number of observations ($n$) becomes smaller than the number of explanatory variables ($k$). Such situation can arise in medical research where the number of patients may be small but information is collected on a large number of variables. In another example, if there is time series data for 50 years on consumption pattern, then it is expected that the consumption pattern does not remain same for 50 years. So better option is to choose smaller number of variables and hence it results into $n < k$. But this is not always advisable. For example in-microarray experiments it is not advisable to choose smaller number of variables.
Consequences of multicollinearity

To illustrate the consequences of presence of multicollinearity, consider a model

\[ y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \quad E(\varepsilon) = 0, \quad V(\varepsilon) = \sigma^2 I \]

where \( x_1, x_2 \) and \( y \) are scaled to length unity.

The normal equation \((X'X)b = X'y\) in this model becomes

\[
\begin{pmatrix}
1 & r \\
r & 1
\end{pmatrix}
\begin{pmatrix}
b_1 \\
b_2
\end{pmatrix}
= 
\begin{pmatrix}
r_{1y} \\
r_{2y}
\end{pmatrix}
\]

where \( r \) is the correlation coefficient between \( x_1 \) and \( x_2 \); \( r_{jy} \) is the correlation coefficient between \( x_j \) and \( y(j=1,2) \) and \( b = (b_1, b_2)' \) is the OLS estimate of \( \beta \).

\[
(X'X)^{-1} = \begin{pmatrix}
1 & -r \\
-r & 1
\end{pmatrix}
\]

\[
\Rightarrow b_1 = \frac{r_{1y} - rr_{2y}}{1-r^2},
\]

\[
b_2 = \frac{r_{2y} - rr_{1y}}{1-r^2}.
\]

So the covariance matrix is \( V(b) = \sigma^2 (X'X)^{-1} \)

\[
\Rightarrow Var(b_1) = Var(b_2) = \frac{\sigma^2}{1-r^2}
\]

\[
Cov(b_1, b_2) = -\frac{r\sigma^2}{1-r^2}.
\]

If \( x_1 \) and \( x_2 \) are uncorrelated, then \( r = 0 \) and

\[
X'X = \begin{pmatrix}
1 & 0 \\
0 & 1
\end{pmatrix}
\]

\[ \text{rank}(X'X) = 2. \]

If \( x_1 \) and \( x_2 \) are perfectly correlated, then \( r = \pm 1 \) and \( \text{rank}(X'X) = 1. \)

If \( r \to \pm 1 \), then \( Var(b_1) = Var(b_2) \to \infty \).

So if variables are perfectly collinear, the variance of OLS estimates becomes large. This indicates highly unreliable estimates and this is an inadmissible situation.
Consider the following result

<table>
<thead>
<tr>
<th></th>
<th>0.99</th>
<th>0.9</th>
<th>0.1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$</td>
<td>50$\sigma^2$</td>
<td>5$\sigma^2$</td>
<td>1.01$\sigma^2$</td>
<td>$\sigma^2$</td>
</tr>
</tbody>
</table>

$Var(b_1) = Var(b_2)$

The standard errors of $b_1$ and $b_2$ rise sharply as $r \to \pm 1$ and they break down at $r = \pm 1$ because $X'X$ becomes non-singular.

- If $r$ is close to 0, then multicollinearity does not harm and it is termed as **non-harmful multicollinearity**.
- If $r$ is close to +1 or -1 then multicollinearity inflates the variance and it rises terribly. This is termed as **harmful multicollinearity**.

There is no clear cut boundary to distinguish between the harmful and non-harmful multicollinearity. Generally, if $r$ is low, the multicollinearity is considered as non-harmful and if $r$ is high, the multicollinearity is considered as harmful.

In case of near or high multicollinearity, following possible consequences are encountered.

1. The OLSE remains an unbiased estimator of $\beta$ but its sampling variance becomes very large. So OLSE becomes imprecise and property of BLUE does not hold anymore.
2. Due to large standard errors, the regression coefficients may not appear significant. Consequently, important variables may be dropped.
   For example, to test $H_0 : \beta_i = 0$, we use $t$-ratio as
   $$t_0 = \frac{b_i}{\sqrt{Var(b_i)}}.$$  
   Since $\widehat{Var}(b_i)$ is large, so $t_0$ is small and consequently $H_0$ is more oftenly accepted.
   Thus harmful multicollinearity intends to delete important variables.
3. Due to large standard errors, the large confidence region may arise. For example, the confidence interval is given by $\left( b_i \pm t_{\frac{\alpha}{2},n-1}\sqrt{\widehat{Var}(b_i)} \right)$. When $\widehat{Var}(b_i)$ becomes large, then confidence interval becomes wider.
4. The OLSE may be sensitive to small changes in the values of explanatory variables. If some observations are added or dropped, OLSE may change considerably in magnitude as well as in sign. Ideally, OLSE should not change with inclusion or deletion of few observations. Thus OLSE looses stability and robustness.

When the number of explanatory variables are more than two, say \( k = 1, 2, \ldots, k \) then the \( j^{th} \) diagonal element of \( C = (X'X)^{-1} \) is

\[
C_{jj} = \frac{1}{1 - R_j^2}
\]

where \( R_j^2 \) is the multiple correlation coefficient or coefficient of determination from the regression of \( X_j \) on the remaining \( (k - 1) \) explanatory variables.

If \( X_j \) is highly correlated with any subset of other \( (k - 1) \) explanatory variables then \( R_j^2 \) is high and close to 1. Consequently variance of \( j^{th} \) OLSE \( \text{Var}(b_j) = C_{jj}\sigma^2 = \frac{\sigma^2}{1 - R_j^2} \) becomes very high. The covariance between \( b_i \) and \( b_j \) will also be large if \( X_i \) and \( X_j \) are involved in the linear relationship leading to multicollinearity.

The least squares estimates \( b_j \) become too large in absolute value in the presence of multicollinearity. For example, consider the squared distance between \( b \) and \( \beta \) as

\[
L^2 = (b - \beta)'(b - \beta)
\]

\[
E(L^2) = \sum_{j=1}^{k} E[(b_j - \beta_j)^2]
\]

\[
= \sum_{j=1}^{k} \text{Var}(b_j)
\]

\[
= \sigma^2 \text{tr}(X'X)^{-1}.
\]

The trace of a matrix is same as the sum of its eigenvalues. If \( \lambda_1, \lambda_2, \ldots, \lambda_k \) are the eigenvalues of \( (X'X) \), then \( \frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \ldots, \frac{1}{\lambda_k} \) are the eigenvalues of \( (X'X)^{-1} \) and hence

\[
E(L^2) = \sigma^2 \sum_{j=1}^{k} \frac{1}{\lambda_j}, \quad \lambda_j > 0.
\]
If \((X'X)\) is ill-conditioned due to the presence of multicollinearity then at least one of the eigenvalue will be small. So the distance between \(b\) and \(\beta\) may also be large. Thus

\[
E(L^2) = E(b - \beta)'(b - \beta) \\
\sigma^2 \text{tr}(X'X)^{-1} = E(b'b - 2b'\beta + \beta'\beta) \\
\Rightarrow E(b'b) = \sigma^2 \text{tr}(X'X)^{-1} + \beta'\beta \\
\Rightarrow b \text{ is generally larger in magnitude than } \beta \\
\Rightarrow \text{OLSE are too large in absolute value.}
\]

The least squares produces bad estimates of parameters in the presence of multicollinearity. This does not imply that the fitted model produces bad predictions also. If the predictions are confined to \(x\)-space with non-harmful multicollinearity, then predictions are satisfactory.

**Multicollinearity diagnostics**

An important question arises that how to diagnose the presence of multicollinearity in the data on the basis of given sample information. Several diagnostic measures are available and each of them is based on a particular approach. It is difficult to say that which of the diagnostic is best or ultimate. Some of the popular and important diagnostics are described further. The detection of multicollinearity involves 3 aspects:

(i) Determining its presence.

(ii) Determining its severity.

(iii) Determining its form or location.

1. **Determinant of** \(X'X\) \(|X'X|\)

This measure is based on the fact that the matrix \(X'X\) becomes ill conditioned in the presence of multicollinearity. The value of determinant of \(X'X\), i.e., \(|X'X|\) declines as degree of multicollinearity increases.

If \(\text{Rank}(X'X) < k\) then \(|X'X|\) will be singular and so \(|X'X| = 0\). So as \(|X'X| \to 0\), the degree of multicollinearity increases and it becomes exact or perfect at \(|X'X| = 0\). Thus \(|X'X|\) serves as a measure of multicollinearity and \(|X'X| = 0\) indicates that perfect multicollinearity exists.
Limitations:

This measure has following limitations

(i) It is not bounded as \(0 < |X'X| < \infty\).

(ii) It is affected by dispersion of explanatory variables. For example, if \(k = 2\), then

\[
|X'X| = \frac{\sum_{i=1}^{n} x_{1i}^2 \sum_{i=1}^{n} x_{1i} x_{2i}}{\sum_{i=1}^{n} x_{2i} x_{1i} \sum_{i=1}^{n} x_{2i}^2}
\]

\[
= \left(\frac{\sum_{i=1}^{n} x_{1i}^2}{\sum_{i=1}^{n} x_{2i}^2}\right) \left(1 - r_{12}^2\right)
\]

where \(r_{12}\) is the correlation coefficient between \(x_1\) and \(x_2\). So \(|X'X|\) depends on correlation coefficient and variability of explanatory variable. If explanatory variables have very low variability, then \(|X'X|\) may tend to zero which will indicate the presence of multicollinearity and which is not the case so.

(iii) It gives no idea about the relative effects on individual coefficients. If multicollinearity is present, then it will not indicate that which variable in \(|X'X|\) is causing multicollinearity and is hard to determine.

2. Inspection of correlation matrix

The inspection of off-diagonal elements \(r_{ij}\) in \(X'X\) gives an idea about the presence of multicollinearity. If \(X_i\) and \(X_j\) are nearly linearly dependent then \(|r_{ij}|\) will be close to 1. Note that the observations in \(X\) are standardized in the sense that each observation is subtracted from mean of that variable and divided by the square root of corrected sum of squares of that variable.

When more than two explanatory variables are considered and if they are involved in near-linear dependency, then it is not necessary that any of the \(r_{ij}\) will be large. Generally, pairwise inspection of correlation coefficients is not sufficient for detecting multicollinearity in the data.
3. Determinant of correlation matrix

Let $D$ be the determinant of correlation matrix then $0 \leq D \leq 1$.

If $D = 0$ then it indicates the existence of exact linear dependence among explanatory variables.

If $D = 1$ then the columns of $X$ matrix are orthonormal.

Thus a value close to 0 is an indication of high degree of multicollinearity. Any value of $D$ between 0 and 1 gives an idea of the degree of multicollinearity.

**Limitation**

It gives no information about the number of linear dependencies among explanatory variables.

**Advantages over $X'X$**

(i) It is a bounded measure, $0 \leq D \leq 1$.

(ii) It is not affected by the dispersion of explanatory variables. For example, when $k = 2$,

$$
|X'X| = \begin{vmatrix}
\sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} \\
\sum_{i=1}^{n} x_{i1}x_{i2} & \sum_{i=1}^{n} x_{i2}^2 \\
\end{vmatrix} = (1 - r_{12}^2).
$$

4. Measure based on partial regression:

A measure of multicollinearity can be obtained on the basis of coefficients of determination based on partial regression. Let $R^2$ be the coefficient of determination in the full model, i.e., based on all explanatory variables and $R_i^2$ be the coefficient of determination in the model when $i^{th}$ explanatory variable is dropped, $i = 1, 2, ..., k$, and $R_k^2 = \text{Max}(R_1^2, R_2^2, ..., R_k^2)$.

**Procedure:**

(i) Drop one of the explanatory variable among $k$ variables, say $X_1$.

(ii) Run regression of $y$ over rest of the $(k-1)$ variables $X_2, X_3, ..., X_k$.

(iii) Calculate $R_i^2$.

(iv) Similarly calculate $R_2^2, R_3^2, ..., R_k^2$.

(v) Find $R_k^2 = \text{Max}(R_1^2, R_2^2, ..., R_k^2)$.

(vi) Determine $R^2 - R_k^2$. 

*Econometrics | Chapter 7 | Multicollinearity | Shalabh, IIT Kanpur*
The quantity \( R^2 - R_L^2 \) provides a measure of multicollinearity. If multicollinearity is present, \( R_L^2 \) will be high. Higher the degree of multicollinearity, higher the value of \( R_L^2 \). So in the presence of multicollinearity, \( R^2 - R_L^2 \) will be low.

Thus if \( R^2 - R_L^2 \) is close to 0, it indicates the high degree of multicollinearity.

**Limitations:**

(i) It gives no information about the underlying relations about explanatory variables, i.e., how many relationships are present or how many explanatory variables are responsible for the multicollinearity.

(ii) Small value of \( R^2 - R_L^2 \) may occur because of poor specification of the model also and it may be inferred in such situation that multicollinearity is present.

**5. Variance inflation factors (VIF):**

The matrix \( X'X \) becomes ill-conditioned in the presence of multicollinearity in the data. So the diagonal elements of \( C = (X'X)^{-1} \) helps in the detection of multicollinearity. If \( R_j^2 \) denotes the coefficient of determination obtained when \( X_j \) is regressed on the remaining \( (k-1) \) variables excluding \( X_j \), then the \( j^{th} \) diagonal element of \( C \) is

\[
C_{jj} = \frac{1}{1 - R_j^2}.
\]

If \( X_j \) is nearly orthogonal to remaining explanatory variables, then \( R_j^2 \) is small and consequently \( C_{jj} \) is close to 1.

If \( X_j \) is nearly linearly dependent on a subset of remaining explanatory variables, then \( R_j^2 \) is close to 1 and consequently \( C_{jj} \) is large.

Since the variance of \( j^{th} \) OLSE of \( \beta_j \) is

\[
Var(b_j) = \sigma^2 C_{jj}
\]

So \( C_{jj} \) is the factor by which the variance of \( b_j \) increases when the explanatory variables are near linear dependent. Based on this concept, the variance inflation factor for the \( j^{th} \) explanatory variable is defined as
This is the factor which is responsible for inflating the sampling variance. The combined effect of dependencies among the explanatory variables on the variance of a term is measured by the \( VIF \) of that term in the model.

One or more large \( VIF \)s indicate the presence of multicollinearity in the data.

In practice, usually a \( VIF > 5 \) or 10 indicates that the associated regression coefficients are poorly estimated because of multicollinearity. If regression coefficients are estimated by OLSE and its variance is \( \sigma^2(X'X)^{-1} \). So \( VIF \) indicates that a part of this variance is given by \( VIF_j \).

**Limitations:**

(i) It sheds no light on the number of dependencies among the explanatory variables.

(ii) The rule of \( VIF > 5 \) or 10 is a rule of thumb which may differ from one situation to another situation.

**Another interpretation of \( VIF_j \)**

The \( VIF \)s can also be viewed as follows.

The confidence interval of the \( j^{th} \) OLSE of \( \beta_j \) is given by

\[
\left( b \pm \hat{\sigma}^2 C_j t_{\alpha/2, n-k-1} \right).
\]

The length of the confidence interval is

\[
L_j = 2\hat{\sigma}^2 C_j t_{\alpha/2, n-k-1}.
\]

Now consider a situation where \( X \) is an orthogonal matrix, i.e., \( X'X = I \) so that \( C_j = 1 \), sample size is same as earlier and same root mean squares \( \left( \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \right) \), then the length of confidence interval becomes

\[
L^* = 2\hat{\sigma} t_{\alpha/2, n-k-1}.
\]
Consider the ratio

\[ \frac{L_j}{L^*} = \sqrt{C_{jj}}. \]

Thus \( \sqrt{VIF_j} \) indicates the increase in the length of confidence interval of \( j^{th} \) regression coefficient due to the presence of multicollinearity.

### 6. Condition number and condition index:

Let \( \lambda_1, \lambda_2, ..., \lambda_k \) be the eigenvalues (or characteristic roots) of \( X'X \). Let

\[ \lambda_{\text{max}} = \text{Max}(\lambda_1, \lambda_2, ..., \lambda_k) \]
\[ \lambda_{\text{min}} = \text{Min}(\lambda_1, \lambda_2, ..., \lambda_k). \]

The condition number \( (CN) \) is defined as

\[ CN = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}}, \quad 0 < CN < \infty. \]

The small values of characteristic roots indicates the presence of near-linear dependency in the data. The \( CN \) provides a measure of spread in the spectrum of characteristic roots of \( X'X \).

The condition number provides a measure of multicollinearity.

- If \( CN < 100 \), then it is considered as **non-harmful multicollinearity**.
- If \( 100 < CN < 1000 \), then it indicates that the multicollinearity is moderate to severe (or strong).
  
  This range is referred to as **danger level**.
- If \( CN > 1000 \), then it indicates a **severe** (or **strong**) **multicollinearity**.

The condition number is based only on two eigenvalues: \( \lambda_{\text{min}} \) and \( \lambda_{\text{max}} \). Another measures are condition indices which use information on other eigenvalues.

The **condition indices** of \( X'X \) are defined as

\[ C_j = \frac{\lambda_{\text{max}}}{\lambda_j}, \quad j = 1, 2, ..., k. \]

In fact, largest \( C_j = CN \).

The number of condition indices that are large, say more than 1000, indicate the number of near-linear dependencies in \( X'X \).

A limitation of \( CN \) and \( C_j \) is that they are unbounded measures as \( 0 < CN < \infty \), \( 0 < C_j < \infty \).
7. Measure based on characteristic roots and proportion of variances:
Let \( \lambda_1, \lambda_2, ..., \lambda_k \) be the eigenvalues of \( X'X \), \( \Lambda = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_k) \) is \( k \times k \) matrix and \( V \) is a \( k \times k \) matrix constructed by the eigenvectors of \( X'X \). Obviously, \( V \) is an orthogonal matrix. Then \( X'X \) can be decomposed as \( X'X = V \Lambda V' \). Let \( V_1, V_2, ..., V_k \) be the column of \( V \). If there is near-linear dependency in the data, then \( \lambda_j \) is close to zero and the nature of linear dependency in described by the elements of associated eigenvector \( V_j \).

The covariance matrix of OLSE is
\[
V(b) = \sigma^2(X'X)^{-1} = \sigma^2(V \Lambda V')^{-1} = \sigma^2V \Lambda^{-1}V' \\
\Rightarrow Var(b_j) = \sigma^2 \left( \frac{v_{1j}}{\lambda_1} + \frac{v_{2j}}{\lambda_2} + ... + \frac{v_{kj}}{\lambda_k} \right)
\]
where \( v_{ij}, v_{i2}, ..., v_{ik} \) are the elements in \( V \).

The condition indices are
\[
C_j = \frac{\lambda_{\text{max}}}{\lambda_j}, \quad j = 1, 2, ..., k.
\]

Procedure:
(i) Find condition index \( C_1, C_2, ..., C_k \).
(ii) (a) Identify those \( \lambda_i \)'s for which \( C_j \) is greater than the danger level 1000.
     (b) This gives the number of linear dependencies.
     (c) Don’t consider those \( C_j \)'s which are below the danger level.
(iii) For such \( \lambda_i \)'s with condition index above the danger level, choose one such eigenvalue, say \( \lambda_j \).
(iv) Find the value of proportion of variance corresponding to \( \lambda_j \) in \( Var(b_1), Var(b_2), ..., Var(b_k) \) as
\[
p_j = \frac{(v_{ij}^2 / \lambda_j)}{VIF_j} = \frac{v_{ij}^2 / \lambda_j}{\sum_{j=1}^{k}(v_{ij}^2 / \lambda_j)}.
\]
Note that \[ \left( \frac{v^2_j}{\lambda_j} \right) \] can be found from the expression

\[ Var(b_j) = \sigma^2 \left( \frac{v^2_1}{\lambda_1} + \frac{v^2_2}{\lambda_2} + \ldots + \frac{v^2_k}{\lambda_k} \right) \]

i.e., corresponding to \( j^{th} \) factor.

The proportion of variance \( p_{ij} \) provides a measure of multicollinearity.

If \( p_{ij} > 0.5 \), it indicates that \( b_i \) is adversely affected by the multicollinearity, i.e., estimate of \( \beta_i \) is influenced by the presence of multicollinearity.

It is a good diagnostic tool in the sense that it tells about the presence of harmful multicollinearity as well as also indicates the number of linear dependencies responsible for multicollinearity. This diagnostic is better than other diagnostics.

The condition indices are also defined by the singular value decomposition of \( X \) matrix as follows:

\[ X = UDV' \]

where \( U \) is \( n \times k \) matrix, \( V \) is \( k \times k \) matrix, \( U'U = I \), \( V'V = I \), \( D \) is \( k \times k \) matrix, \( D = \text{diag}(\mu_1, \mu_2, \ldots, \mu_k) \) and \( \mu_1, \mu_2, \ldots, \mu_k \) are the singular values of \( X \), \( V \) is a matrix whose columns are eigenvectors corresponding to eigenvalues of \( X'X \) and \( U \) is a matrix whose columns are the eigenvectors associated with the \( k \) nonzero eigenvalues of \( X'X \).

The condition indices of \( X \) matrix are defined as

\[ \eta_j = \frac{\mu_{\text{max}}}{\mu_j}, \; j = 1, 2, \ldots, k \]

where \( \mu_{\text{max}} = \text{Max}(\mu_1, \mu_2, \ldots, \mu_k) \).

If \( \lambda_1, \lambda_2, \ldots, \lambda_k \) are the eigenvalues of \( X'X \) then

\[ X'X = (UDV')'UDV' = VD^2V' = VAV', \]

so \( \mu_j^2 = \lambda_j, \; j = 1, 2, \ldots, k \).

---

*Econometrics | Chapter 7 | Multicollinearity | Shalabh, IIT Kanpur*
Note that with $\mu_j^2 = \lambda_j$,

$$\text{Var}(b_j) = \sigma^2 \sum_{i=1}^{k} \frac{v_{ji}^2}{\mu_i^2}$$

$$VIF_j = \sum_{i=1}^{k} \frac{v_{ji}^2}{\mu_i^2}$$

$$p_{ij} = \frac{v_{ij}^2 / \mu_i^2}{VIF_j}.$$

The ill-conditioning in $X$ is reflected in the size of singular values. There will be one small singular value for each non-linear dependency. The extent of ill conditioning is described by how small is $\mu_j$ relative to $\mu_{\text{max}}$.

It is suggested that the explanatory variables should be scaled to unit length but should not be centered when computing $p_{ij}$. This will help in diagnosing the role of intercept term in near-linear dependence. No unique guidance is available in literature on the issue of centering the explanatory variables. The centering makes the intercept orthogonal to explanatory variables. So this may remove the ill conditioning due to intercept term in the model.

**Remedies for multicollinearity:**

Various techniques have been proposed to deal with the problems resulting from the presence of multicollinearity in the data.

1. **Obtain more data**

The harmful multicollinearity arises essentially because rank of $X'X$ falls below $k$ and $|X'X|$ is close to zero. Additional data may help in reducing the sampling variance of the estimates. The data need to be collected such that it helps in breaking up the multicollinearity in the data.

It is always not possible to collect additional data to various reasons as follows.

- The experiment and process have finished and no longer available.
- The economic constrains may also not allow to collect the additional data.
- The additional data may not match with the earlier collected data and may be unusual.
- If the data is in time series, then longer time series may force to take data that is too far in the past.
• If multicollinearity is due to any identity or exact relationship, then increasing the sample size will not help.

• Sometimes, it is not advisable to use the data even if it is available. For example, if the data on consumption pattern is available for the years 1950-2010, then one may not like to use it as the consumption pattern usually does not remain the same for such a long period.

2. **Drop some variables that are collinear:**

If possible, identify the variables which seem to cause multicollinearity. These collinear variables can be dropped so as to match the condition of full rank of $X$–matrix. The process of omitting the variables may be carried out on the basis of some kind of ordering of explanatory variables, e.g., those variables can be deleted first which have smaller value of $t$-ratio. In another example, suppose the experimenter is not interested in all the parameters. In such cases, one can get the estimators of the parameters of interest which have smaller mean squared errors than the variance of OLSE by dropping some variables.

If some variables are eliminated, then this may reduce the predictive power of the model. Sometimes there is no assurance that how the model will exhibit less multicollinearity.

3. **Use some relevant prior information:**

One may search for some relevant prior information about the regression coefficients. This may lead to specification of estimates of some coefficients. More general situation includes the specification of some exact linear restrictions and stochastic linear restrictions. The procedures like restricted regression and mixed regression can be used for this purpose. The relevance and correctness of information plays an important role in such analysis but it is difficult to ensure it in practice. For example, the estimates derived in U.K. may not be valid in India.

4. **Employ generalized inverse**

If rank $(X'X)<k$, then the generalized inverse can be used to find the inverse of $X'X$. Then $\beta$ can be estimated by $\hat{\beta} = (X'X)^{-1} X'y$.

In such case, the estimates will not be unique except in the case of use of Moore-Penrose inverse of $(X'X)$. Different methods of finding generalized inverse may give different results. So applied workers will get different results. Moreover, it is also not known that which method of finding generalized inverse is optimum.
5. Use of principal component regression

The principal component regression is based on the technique of principal component analysis. The \( k \) explanatory variables are transformed into a new set of orthogonal variables called as principal components. Usually this technique is used for reducing the dimensionality of data by retaining some levels of variability of explanatory variables which is expressed by the variability in study variable. The principal components involves the determination of a set of linear combinations of explanatory variables such that they retain the total variability of the system and these linear combinations are mutually independent of each other. Such obtained principal components are ranked in the order of their importance. The importance being judged in terms of variability explained by a principal component relative to the total variability in the system. The procedure then involves eliminating some of the principal components which contribute in explaining relatively less variation. After elimination of the least important principal components, the set up of multiple regression is used by replacing the explanatory variables with principal components. Then study variable is regressed against the set of selected principal components using ordinary least squares method. Since all the principal components are orthogonal, they are mutually independent and so OLS is used without any problem. Once the estimates of regression coefficients for the reduced set of orthogonal variables (principal components) have been obtained, they are mathematically transformed into a new set of estimated regression coefficients that correspond to the original correlated set of variables. These new estimated coefficients are the principal components estimators of regression coefficients.

Suppose there are \( k \) explanatory variables \( X_1, X_2, \ldots, X_k \). Consider the linear function of \( X_1, X_2, \ldots, X_k \) like

\[
Z_1 = \sum_{i=1}^{k} a_i X_i
\]

\[
Z_2 = \sum_{i=1}^{k} b_i X_i \quad \text{etc.}
\]

The constants \( a_1, a_2, \ldots, a_k \) are determined such that the variance of \( Z_1 \) is maximized subject to the normalizing condition that \( \sum_{i=1}^{k} a_i^2 = 1 \). The constant \( b_1, b_2, \ldots, b_k \) are determined such that the variance of \( Z_2 \) is maximized subject to the normality condition that \( \sum_{i=1}^{k} b_i^2 = 1 \) and is independent of the first principal component.
We continue with such process and obtain $k$ such linear combinations such that they are orthogonal to their preceding linear combinations and satisfy the normality condition. Then we obtain their variances. Suppose such linear combinations are $Z_1, Z_2, \ldots, Z_k$ and for them, $\text{Var}(Z_1) > \text{Var}(Z_2) > \ldots > \text{Var}(Z_k)$. The linear combination having the largest variance is the first principal component. The linear combination having the second largest variance is the second largest principal component and so on. These principal components have the property that $\sum_{i=1}^{k} \text{Var}(Z_i) = \sum_{i=1}^{k} \text{Var}(X_i)$. Also, the $X_1, X_2, \ldots, X_k$ are correlated but $Z_1, Z_2, \ldots, Z_k$ are orthogonal or uncorrelated. So there will be zero multicollinearity among $Z_1, Z_2, \ldots, Z_k$.

The problem of multicollinearity arises because $X_1, X_2, \ldots, X_k$ are not independent. Since the principal components based on $X_1, X_2, \ldots, X_k$ are mutually independent, so they can be used as explanatory variables and such regression will combat the multicollinearity.

Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the eigenvalues of $X'X$, $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)$ is $k \times k$ diagonal matrix, $V$ is a $k \times k$ orthogonal matrix whose columns are the eigenvectors associated with $\lambda_1, \lambda_2, \ldots, \lambda_k$. Consider the canonical form of the linear model

\[ y = X\beta + \varepsilon = XV^T \beta + \varepsilon = Z\alpha + \varepsilon \]

where $Z = XV$, $\alpha = V^T \beta$, $V^T X'XV = Z'Z = \Lambda$.

Columns of $Z = (Z_1, Z_2, \ldots, Z_k)$ define a new set of explanatory variables which are called as principal components.

The OLS of $\alpha$ is

\[ \hat{\alpha} = (Z'Z)^{-1}Z'y = \Lambda^{-1}Z'y \]

and its covariance matrix is
\[ V(\hat{\alpha}) = \sigma^2 (Z'Z)^{-1} \]
\[ = \sigma^2 \Lambda^{-1} \]
\[ = \sigma^2 \text{diag}\left( \frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \ldots, \frac{1}{\lambda_k} \right) \]

Note that \( \lambda_j \) is the variance of \( j^{th} \) principal component and \( Z'Z = \sum_{i=1}^{k} \sum_{j=1}^{k} Z_i Z_j = \Lambda \). A small eigenvalue of \( X'X \) means that the linear relationship between the original explanatory variable exist and the variance of corresponding orthogonal regression coefficient is large which indicates that the multicollinearity exists. If one or more \( \lambda_j \) are small, then it indicates that multicollinearity is present.

**Retainment of principal components:**

The new set of variables, i.e., principal components are orthogonal, and they retain the same magnitude of variance as of original set. If multicollinearity is severe, then there will be at least one small value of eigenvalue. The elimination of one or more principal components associated with smallest eigenvalues will reduce the total variance in the model. Moreover, the principal components responsible for creating multicollinearity will be removed and the resulting model will be appreciably improved.

The principal component matrix \( Z = [Z_1, Z_2, \ldots, Z_k] \) with \( Z_1, Z_2, \ldots, Z_k \) contains exactly the same information as the original data in \( X \) in the sense that the total variability in \( X \) and \( Z \) is same. The difference between them is that the original data are arranged into a set of new variables which are uncorrelated with each other and can be ranked with respect to the magnitude of their eigenvalues. The \( j^{th} \) column vector \( Z_j \) corresponding to the largest \( \lambda_j \) accounts for the largest proportion of the variation in the original data. Thus the \( Z_j \)'s are indexed so that \( \lambda_1 > \lambda_2 > \ldots > \lambda_k > 0 \) and \( \lambda_j \) is the variance of \( Z_j \).

A strategy of elimination of principal components is to begin by discarding the component associated with the smallest eigenvalue. The idea behind to do so is that the principal component with smallest eigenvalue is contributing least variance and so is least informative.

Using this procedure, principal components are eliminated until the remaining components explain some preselected variance is terms of percentage of total variance. For example, if 90% of total variance is
needed, and suppose \( r \) principal components are eliminated which means that \((k-r)\) principal components contributes 90% variation, then \( r \) is selected to satisfy
\[
\frac{\sum_{i=1}^{k-r} \lambda_i}{\sum_{i=1}^{k} \lambda_i} > 0.90.
\]

Various strategies to choose required number of principal components are also available in the literature.

Suppose after using such a rule, the \( r \) principal components are eliminated. Now only \((k-r)\) components will be used for regression. So \( Z \) matrix is partitioned as
\[
Z = (Z_r, Z_{k-r}) = X(V'_r, V'_{k-r})
\]
where \( Z_r \) submatrix is of order \( n \times r \) and contains the principal components to be eliminated. The submatrix \( Z_{k-r} \) is of order \( n \times (k-r) \) and contains the principal components to be retained.

The reduced model obtained after the elimination of \( r \) principal components can be expressed as
\[
y = Z_{k-r} \alpha_{k-r} + \epsilon^*.
\]
The random error component is represented as \( \epsilon^* \) just to distinguish with \( \epsilon \). The reduced coefficients contain the coefficients associated with retained \( Z_j 's \). So
\[
Z_{k-r} = (Z_1, Z_2, ..., Z_{k-r}) \\
\alpha_{k-r} = (\alpha_1, \alpha_2, ..., \alpha_{k-r}) \\
V_{k-r} = (V_1, V_2, ..., V_{k-r}).
\]

Using OLS on the model with retained principal components, the OLSE of \( \alpha_{k-r} \) is
\[
\hat{\alpha}_{k-r} = (Z'_{k-r} Z_{k-r})^{-1} Z'_{k-r} y.
\]
Now it is transformed back to original explanatory variables as follows:
\[
\alpha = V' \beta \\
\alpha_{k-r} = V'_{k-r} \beta \\
\Rightarrow \hat{\beta}_{pc} = V'_{k-r} \hat{\alpha}_{k-r}
\]
which is the **principal component regression estimator** of \( \beta \).

This method improves the efficiency as well as multicollinearity.
6. Ridge regression

The OLSE is the best linear unbiased estimator of regression coefficient in the sense that it has minimum variance in the class of linear and unbiased estimators. However if the condition of unbiased can be relaxed then it is possible to find a biased estimator of regression coefficient say \( \hat{\beta} \) that has smaller variance than the unbiased OLSE \( \beta \). The mean squared error (MSE) of \( \hat{\beta} \) is

\[
MSE(\hat{\beta}) = E(\hat{\beta} - \beta)^2
= E\left[\left(\hat{\beta} - E(\hat{\beta})\right) + \left(E(\hat{\beta}) - \beta\right)\right]^2
= Var(\hat{\beta}) + \left[E(\hat{\beta}) - \beta\right]^2
= Var(\hat{\beta}) + \left[\text{Bias}(\hat{\beta})\right]^2.
\]

Thus \( MSE(\hat{\beta}) \) can be made smaller than \( Var(\hat{\beta}) \) by introducing small bias in \( \hat{\beta} \). One of the approach to do so is the ridge regression. The ridge regression estimator is obtained by solving the normal equations of least squares estimation. The normal equations are modified as

\[
(X'X + \delta I)\hat{\beta}_{\text{ridge}} = X'y
\Rightarrow \hat{\beta}_{\text{ridge}} = (X'X + \delta I)^{-1}X'y
\]

is the ridge regression estimator of \( \beta \) and \( \delta \geq 0 \) is any characterizing scalar termed as biasing parameter.

As \( \delta \rightarrow 0, \hat{\beta}_{\text{ridge}} \rightarrow b(\text{OLSE}) \) and as \( \delta \rightarrow \infty, \hat{\beta}_{\text{ridge}} \rightarrow 0.\)

So larger the value of \( \delta \), larger shrinkage towards zero. Note that the OLSE in inappropriate to use in the sense that it has very high variance when multicollinearity is present in the data. On the other hand, a very small value of \( \hat{\beta} \) may tend to accept the null hypothesis \( H_0 : \beta = 0 \) indicating that the corresponding variables are not relevant. The value of biasing parameter controls the amount of shrinkage in the estimates.
Bias of ridge regression estimator:

The bias of $\hat{\beta}_{ridge}$ is

\[
\text{Bias}(\hat{\beta}_{ridge}) = E(\hat{\beta}_{ridge}) - \beta \\
= (X'X + \delta I)^{-1}X'E(y) - \beta \\
= \left[(X'X + \delta I)^{-1}X'X - I\right]\beta \\
= (X'X + \delta I)^{-1}[X'X - X'X - \delta I] \beta \\
= -\delta(X'X + \delta I)^{-1} \beta.
\]

Thus the ridge regression estimator is a biased estimator of $\beta$.

Covariance matrix:

The covariance matrix of $\hat{\beta}_{ridge}$ is defined as

\[
V(\hat{\beta}_{ridge}) = E\left[\left\{\hat{\beta}_{ridge} - E(\hat{\beta}_{ridge})\right\}\left\{\hat{\beta}_{ridge} - E(\hat{\beta}_{ridge})\right\}'\right].
\]

Since

\[
\hat{\beta}_{ridge} - E(\hat{\beta}_{ridge}) = (X'X + \delta I)^{-1}X'y - (X'X + \delta I)^{-1}X'X\beta \\
= (X'X + \delta I)^{-1}X'(y - X\beta) \\
= (X'X + \delta I)^{-1}X'\varepsilon,
\]

so

\[
V(\hat{\beta}_{ridge}) = (X'X + \delta I)^{-1}X'V(\varepsilon)X(X'X + \delta I)^{-1} \\
= \sigma^2(X'X + \delta I)^{-1}X'X(X'X + \delta I)^{-1}.
\]

Mean squared error:

The mean squared error of $\hat{\beta}_{ridge}$ is

\[
\text{MSE}(\hat{\beta}_{ridge}) = \text{Var}(\hat{\beta}_{ridge}) + \left[\text{bias}(\hat{\beta}_{ridge})\right]^2 \\
= tr\left[V(\hat{\beta}_{ridge})\right] + \left[\text{bias}(\hat{\beta}_{ridge})\right]^2 \\
= \sigma^2 tr\left[(X'X + \delta I)^{-1}X'X(X'X + \delta I)^{-1}\right] + \delta^2 \beta'(X'X + \delta I)^{-2} \beta \\
= \sigma^2 \sum_{j=1}^{k} \frac{\lambda_j}{(\lambda_j + \delta)^2} + \delta^2 \beta'(X'X + \delta I)^{-2} \beta
\]

where $\lambda_1, \lambda_2, ..., \lambda_k$ are the eigenvalues of $X'X$. 
Thus as $\delta$ increases, the bias in $\hat{\beta}_{\text{ridge}}$ increases but its variance decreases. The trade off between bias and variance hinges upon the value of $\delta$. It can be shown that there exists a value of $\delta$ such that

$$\text{MSE}(\hat{\beta}_{\text{ridge}}) < \text{Var}(b)$$

provided $\beta'\beta$ is bounded.

**Choice of $\delta$:**

The estimation of ridge regression estimator depends upon the value of $\delta$. Various approaches have been suggested in the literature to determine the value of $\delta$. The value of $\delta$ can be chosen on the basis of criteria like

- stability of estimators with respect to $\delta$.
- reasonable signs.
- magnitude of residual sum of squares etc.

We consider here the determination of $\delta$ by the inspection of ridge trace.

**Ridge trace:**

Ridge trace is the graphical display of ridge regression estimator versus $\delta$.

If multicollinearity is present and is severe, then the instability of regression coefficients is reflected in the ridge trace. As $\delta$ increases, some of the ridge estimates vary dramatically and they stabilizes at some value of $\delta$. The objective in ridge trace is to inspect the trace (curve) and find the reasonable small value of $\delta$ at which the ridge regression estimators are stable. The ridge regression estimator with such a choice of $\delta$ will have smaller MSE than the variance of OLSE.

An example of ridge trace for a model with 6 parameters is as follows. In this ridge trace, the $\hat{\beta}_{\text{ridge}}$ is evaluated for various choices of $\delta$ and the corresponding values of all regression coefficients $\hat{\beta}_{j(\text{ridge})}$'s, $j=1,2,\ldots,6$ are plotted versus $\delta$. These values are denoted by different symbols and are joined by a smooth curve. This produces a ridge trace for respective parameter. Now choose the value of $\delta$ where all the curves stabilize and become nearly parallel. For example, the curves in following figure become nearly parallel starting from $\delta = \delta_4$ or so. Thus one possible choice of $\delta$ is $\delta = \delta_4$ and parameters can be estimated as $\hat{\beta}_{\text{ridge}} = (X'X + \delta_4 I)^{-1} X'y$. 

---

*Econometrics | Chapter 7 | Multicollinearity | Shalabh, IIT Kanpur*
The figure drastically exposes the presence of multicollinearity in the data. The behaviour of $\hat{\beta}_{(\text{ridge})}$ at $\delta_0 \approx 0$ is very different than at other values of $\delta$. For small values of $\delta$, the estimates change rapidly. The estimates stabilize gradually as $\delta$ increases. The value of $\delta$ at which all the estimates stabilize gives the desired value of $\delta$ because moving away from such $\delta$ will not bring any appreciable reduction in the residual sum of squares. If multicollinearity is present, then the variation in ridge regression estimators is rapid around $\delta = 0$. The optimal $\delta$ is chosen such that after that value of $\delta$, almost all traces stabilize.

**Limitations:**

1. The choice of $\delta$ is data dependent and therefore is a random variable. Using it as a random variable violates the assumption that $\delta$ is a constant. This will disturb the optimal properties derived under the assumption of constancy of $\delta$.

2. The value of $\delta$ lies in the interval $(0, \infty)$. So large number of values are required for exploration. This results is wasting of time. This is not a big issue when working with software.

3. The choice of $\delta$ from graphical display may not be unique. Different people may choose different $\delta$ and consequently the values of ridge regression estimators will be changing. However, $\delta$ is chosen so that all the estimators of all coefficients stabilize. Hence small variation in choosing the value of $\delta$ may not produce much change in the ridge estimators of the coefficients. Another choice of $\delta$ is

$$\delta = \frac{k\hat{\sigma}^2}{\hat{b}\hat{b}}$$

where $b$ and $\hat{\sigma}^2$ are obtained from the least squares estimation.
4. The stability of numerical estimates of $\hat{\beta}_i$'s is a rough way to determine $\delta$. Different estimates may exhibit stability for different $\delta$ and it may often be hard to strike a compromise. In such situation, generalized ridge regression estimators are used.

5. There is no guidance available regarding the testing of hypothesis and for confidence interval estimation.

**Idea behind ridge regression estimator:**

The problem of multicollinearity arises because some of the eigenvalues roots of $X'X$ are close to zero or are zero. So if $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the characteristic roots, and if

$$X'X = \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)$$

then

$$\hat{\beta}_{\text{ridge}} = (\Lambda + \delta I)^{-1} X' y = (I + \delta \Lambda^{-1})^{-1} b$$

where $b$ is the OLSE of $\beta$ given by

$$b = (X'X)^{-1} X' y = \Lambda^{-1} X' y.$$ 

Thus a particular element will be of the forms

$$\frac{1}{\lambda_i + \delta} x_i y = \frac{\lambda_i}{\lambda_i + \delta} b_i.$$ 

So a small quantity $\delta$ is added to $\lambda_i$ so that if $\lambda_i = 0$, even then $\frac{1}{\lambda_i + \delta}$ remains meaningful.

**Another interpretation of ridge regression estimator:**

In the model $y = X\beta + \epsilon$, obtain the least squares estimator of $\beta$ when $\sum_{i=1}^{k} \beta_i^2 = C$, where $C$ is some constant. So minimize

$$\delta(\beta) = (y - X\beta)'(y - X\beta) + \delta(\beta'\beta - C)$$

where $\delta$ is the Lagrangian multiplier. Differentiating $S(\beta)$ with respect to $\beta$, the normal equations are obtained as

$$\frac{\partial S(\beta)}{\partial \beta} = 0 \Rightarrow -2X'y + 2X'X\beta + 2\delta\beta = 0$$

$$\Rightarrow \hat{\beta}_{\text{ridge}} = (X'X + \delta I)^{-1} X'y.$$ 

Note that if $\delta$ is very small, it may indicate that most of the regression coefficients are close to zero and if $\delta$ is large, then it may indicate that the regression coefficients are away from zero. So $\delta$ puts a sort of penalty on the regression coefficients to enable its estimation.
Chapter 8  
Heteroskedasticity

In the multiple regression model

\[ y = X \beta + \varepsilon, \]

it is assumed that

\[ V(\varepsilon) = \sigma^2 I, \]

i.e.,

\[ \text{Var}(\varepsilon_i^2) = \sigma^2, \]

\[ \text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \ i \neq j = 1, 2, \ldots, n. \]

In this case, the diagonal elements of covariance matrix of \( \varepsilon \) are same indicating that the variance of each \( \varepsilon_i \) is same and off-diagonal elements of covariance matrix of \( \varepsilon \) are zero indicating that all disturbances are pairwise uncorrelated. This property of constancy of variance is termed as \textbf{homoskedasticity} and disturbances are called as \textbf{homoskedastic disturbances}.

In many situations, this assumption may not be plausible and the variances may not remain same. The disturbances whose variances are not constant across the observations are called \textbf{heteroskedastic disturbance} and this property is termed as \textbf{heteroskedasticity}. In this case

\[ \text{Var}(\varepsilon_i) = \sigma_i^2, \ i = 1, 2, \ldots, n \]

and disturbances are pairwise uncorrelated.

The covariance matrix of disturbances is

\[ V(\varepsilon) = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}. \]
Graphically, the following pictures depict homoskedasticity and heteroskedasticity.

**Heteroskedasticity**

Examples: Suppose in a simple linear regression model, $x$ denote the income and $y$ denotes the expenditure on food. It is observed that as the income increases, the variation in expenditure on food increases because the choice and varieties in food increase, in general, up to a certain extent. So the variance of observations on $y$ will not remain constant as income changes. The assumption of homoscedasticity implies that the consumption pattern of food will remain same irrespective of the income of the person. This may not generally be a correct assumption in real situations. Rather the consumption pattern changes and hence the variance of $y$ and so the variances of disturbances will not remain constant. In general, it will be increasing as income increases.
In another example, suppose in a simple linear regression model, \( x \) denotes the number of hours of practice for typing and \( y \) denotes the number of typing errors per page. It is expected that the number of typing mistakes per page decreases as the person practices more. The homoskedastic disturbances assumption implies that the number of errors per page will remain same irrespective of the number of hours of typing practice which may not be true is practice.

**Possible reasons for heteroskedasticity:**

There are various reasons due to which the heteroskedasticity is introduced in the data. Some of them are as follows:

1. The nature of phenomenon under study may have an increasing or decreasing trend. For example, the variation in consumption pattern on food increases as income increases, similarly the number of typing mistakes decreases as the number of hours of typing practice increases.

2. Sometimes the observations are in the form of averages and this introduces the heteroskedasticity in the model. For example, it is easier to collect data on the expenditure on clothes for the whole family rather than on a particular family member. Suppose in a simple linear regression model

   \[
   y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}, i = 1, 2, \ldots, n, j = 1, 2, \ldots, m_i
   \]

   \( y_{ij} \) denotes the expenditure on cloth for the \( j^{th} \) family having \( m_j \) members and \( x_{ij} \) denotes the age of the \( i^{th} \) person in \( j^{th} \) family. It is difficult to record data for individual family member but it is easier to get data for the whole family. So \( y_{ij}'s \) are known collectively.

Then instead of per member expenditure, we find the data on average expenditure for each family member as

\[
\overline{y}_i = \frac{1}{m_j} \sum_{j=1}^{m_j} y_{ij}
\]

and the model becomes

\[
\overline{y}_i = \beta_0 + \beta_1 \overline{x}_i + \overline{\varepsilon}_i.
\]

It we assume \( E(\varepsilon_{ij}) = 0, Var(\varepsilon_{ij}) = \sigma^2 \), then

\[
E(\overline{\varepsilon}_i) = 0
\]

\[
Var(\overline{\varepsilon}_i) = \frac{\sigma^2}{m_j}
\]
which indicates that the resultant variance of disturbances does not remain constant but depends on the number of members in a family \( m_j \). So heteroskedasticity enters in the data. The variance will remain constant only when all \( m_j \)'s are same.

3. Sometimes the theoretical considerations introduces the heteroskedasticity in the data. For example, suppose in the simple linear model

\[
y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n,
\]

\( y_i \) denotes the yield of rice and \( x_i \) denotes the quantity of fertilizer in an agricultural experiment. It is observed that when the quantity of fertilizer increases, then yield increases. In fact, initially the yield increases when quantity of fertilizer increases. Gradually, the rate of increase slows down and if fertilizer is increased further, the crop burns. So notice that \( \beta_i \) changes with different levels of fertilizer. In such cases, when \( \beta_i \) changes, a possible way is to express it as a random variable with constant mean \( \bar{\beta}_i \) and constant variance \( \theta^2 \) like

\[
\beta_{ii} = \bar{\beta}_i + v_i, \quad i = 1, 2, \ldots, n
\]

with

\[
E(v_i) = 0, \text{Var}(v_i) = \theta^2, E(\epsilon_i v_i) = 0.
\]

So the complete model becomes

\[
y_i = \beta_0 + \beta_{ii} x_i + \epsilon_i
\]

\[
\beta_{ii} = \bar{\beta}_i + v_i
\]

\[
\Rightarrow y_i = \beta_0 + \bar{\beta} x_i + (\epsilon_i + x_i v_i)
\]

\[
= \beta_0 + \bar{\beta} x_i + w_i
\]

where \( w_i = \epsilon_i + x_i v_i \) is like a new random error component. So

\[
E(w_i) = 0
\]

\[
\text{Var}(w_i) = E(w_i^2)
\]

\[
= E(\epsilon_i^2) + x_i^2 E(v_i^2) + 2x_i E(\epsilon_i v_i)
\]

\[
= \sigma^2 + x_i^2 \theta^2 + 0
\]

\[
= \sigma^2 + x_i^2 \theta^2.
\]

So variance depends on \( i \) and thus heteroskedasticity is introduced in the model. Note that we assume homoskedastic disturbances for the model

\[
y_i = \beta_0 + \beta_i x_i + \epsilon_i, \quad \beta_{ii} = \bar{\beta}_i + v_i
\]

but finally end up with heteroskedastic disturbances. This is due to theoretical considerations.
4. The skewness in the distribution of one or more explanatory variables in the model also causes heteroskedasticity in the model.

5. The incorrect data transformations and incorrect functional form of the model can also give rise to the heteroskedasticity problem.

**Tests for heteroskedasticity**

The presence of heteroskedasticity affects the estimation and test of hypothesis. The heteroskedasticity can enter into the data due to various reasons. The tests for heteroskedasticity assume a specific nature of heteroskedasticity. Various tests are available in literature for testing the presence of heteroskedasticity, e.g.,

1. Bartlett test
2. Breusch Pagan test
3. Goldfeld Quandt test
4. Glesjer test
5. Test based on Spearman’s rank correlation coefficient
6. White test
7. Ramsey test
8. Harvey Phillips test
9. Szroeter test
10. Peak test (nonparametric) test

We discuss the first five tests.

**1. Bartlett’s test**

It is a test for testing the null hypothesis

\[ H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_i^2 = \ldots = \sigma_n^2 \]

This hypothesis is termed as the hypothesis of homoskedasticity. This test can be used only when replicated data is available.

Since in the model

\[ y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i, \ E(\epsilon_i) = 0, \ Var(\epsilon_i) = \sigma_i^2, \ i = 1, 2, \ldots, n, \]

only one observation \( y_i \) is available to find \( \sigma_i^2 \), so the usual tests cannot be applied. This problem can be overcome if replicated data is available. So consider the model of the form

\[ y^*_i = X_i \beta + \epsilon^*_i \]
where $y_i^*$ is a $m_i \times 1$ vector, $X_i$ is $m_i \times k$ matrix, $\beta$ is $k \times 1$ vector and $\epsilon_i^*$ is $m_i \times 1$ vector. So replicated data is now available for every $y_i^*$ in the following way:

\[
y_1^* = X_1 \beta + \epsilon_1^* \text{ consists of } m_1 \text{ observations}
\]

\[
y_2^* = X_2 \beta + \epsilon_2^* \text{ consists of } m_2 \text{ observations}
\]

\[
\vdots
\]

\[
y_n^* = X_n \beta + \epsilon_n^* \text{ consists of } m_n \text{ observations}
\]

All the individual model can be written as

\[
\begin{pmatrix}
y_1^* \\
y_2^* \\
\vdots \\
y_n^*
\end{pmatrix} =
\begin{pmatrix}
X_1 \\
X_2 \\
\vdots \\
X_n
\end{pmatrix} \beta +
\begin{pmatrix}
\epsilon_1^* \\
\epsilon_2^* \\
\vdots \\
\epsilon_n^*
\end{pmatrix}
\]

or $y^* = X\beta + \epsilon^*$

where $y^*$ is a vector of order $\left( \sum_{i=1}^{n} m_i \right) \times 1$, $X$ is $\left( \sum_{i=1}^{n} m_i \right) \times k$ matrix, $\beta$ is $k \times 1$ vector and $\epsilon^*$ is $\left( \sum_{i=1}^{n} m_i \right) \times 1$ vector. Application of OLS to this model yields

\[
\hat{\beta} = (X'^X)^{-1}X'^y^*
\]

and obtain the residual vector

\[
e_i^* = y_i^* - X_i \hat{\beta}.
\]

Based on this, obtain

\[
s_i^2 = \frac{1}{m_i - k} \epsilon_i^* \epsilon_i^*
\]

\[
s^2 = \frac{\sum_{i=1}^{n} (m_i - k)s_i^2}{\sum_{i=1}^{n} (m_i - k)}.
\]

Now apply the Bartlett’s test as

\[
\chi^2 = \frac{1}{C} \sum_{i=1}^{n} (m_i - k) \log \frac{s^2}{s_i^2}
\]

which has asymptotic $\chi^2$ distribution with $(n-1)$ degrees of freedom where

\[
C = 1 + \frac{1}{3(n-1)} \left[ \sum_{i=1}^{n} \left( \frac{1}{m_i - k} \right) - \frac{1}{\sum_{i=1}^{n} (m_i - k)} \right].
\]
Another variant of Bartlett’s test

Another variant of Bartlett’s test is based on likelihood ratio test statistic. If there are \( m \) independent normal random samples where there are \( n_i \) observations in the \( i^{th} \) sample. Then the likelihood ratio test statistic for testing \( H_0 = \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 \) is

\[
u = \sum_{i=1}^{m} \left( \frac{s_i^2}{s^2} \right)^{n_i/2}
\]

where

\[
s_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2, \quad i = 1, 2, \ldots, m; \quad j = 1, 2, \ldots, n_i
\]

\[
s^2 = \frac{1}{n} \sum_{i=1}^{m} n_i s_i^2
\]

\[
n = \sum_{i=1}^{m} n_i.
\]

To obtain an unbiased test and a modification of \(-2 \ln u\) which is a closer approximation to \(\chi^2_{m-1}\) under \(H_0\), Bartlett test replaces \(n_i\) by \((n_i - 1)\) and divide by a scalar constant. This leads to the statistic

\[
M = \frac{(n-m) \log \hat{\sigma}^2 - \sum_{i=1}^{m} (n_i - 1) \log \hat{\sigma}_i^2}{1 + \frac{1}{3(m-1)} \left[ \sum_{i=1}^{m} \left( \frac{1}{n_i - 1} \right) - \frac{1}{n-m} \right]}
\]

which has a \(\chi^2\) distribution with \((m-1)\) degrees of freedom under \(H_0\) and

\[
\hat{\sigma}_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2
\]

\[
\hat{\sigma}^2 = \frac{1}{n-m} \sum_{i=1}^{m} (n_i - 1) \hat{\sigma}_i^2.
\]

In experimental sciences, it is easier to get replicated data and this test can be easily applied. In real life applications, it is difficult to get replicated data and this test may not be applied. This difficulty is overcome in Breusch Pagan test.
2. Breusch Pagan test

This test can be applied when the replicated data is not available but only single observations are available. When it is suspected that the variance is some function (but not necessarily multiplicative) of more than one explanatory variable, then Breusch Pagan test can be used.

Assuming that under the alternative hypothesis $\sigma_i^2$ is expressible as

$$\sigma_i^2 = h(Z_i'\gamma) = h(\gamma_1 + Z_i'\gamma^*)$$

where $h$ is some unspecified function and is independent of $i$,

$$Z_i = (1,Z_{i1}'(1,Z_{i2},Z_{i3},...,Z_{ip})$$

is the vector of observable explanatory variables with first element unity and $\gamma' = (\gamma_1,\gamma^*) = (\gamma_1,\gamma_2,...,\gamma_p)$ is a vector of unknown coefficients related to $\beta$ with first element being the intercept term. The heterogeneity is defined by these $p$ variables. These $Z_i$'s may also include some $X$'s also.

Specifically, assume that

$$\sigma_i^2 = \gamma_1 + \gamma_2 Z_{i2} + ... + \gamma_p Z_{ip}.$$  

The null hypothesis

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2$$

can be expressed as

$$H_0 : \gamma_1 = \gamma_2 = ... = \gamma_p = 0.$$  

If $H_0$ is accepted, it implies that $Z_{i2}, Z_{i3},...,Z_{ip}$ do not have any effect on $\sigma_i^2$ and we get $\sigma_i^2 = \gamma_1$.

The test procedure is as follows:

1. Ignoring heteroskedasticity, apply OLS to

$$y_i = \beta_1 + \beta_2 X_{i1} + ... + \beta_k X_{ik} + \varepsilon_i$$

and obtain residual

$$e = y - Xb$$

$$b = (X'X)^{-1} X'Y.$$  

2. Construct the variables

$$g_i = \frac{e_i^2}{SS_{res}} = \frac{ne_i^2}{SS_{res}}$$
where $SS_{res}$ is the residual sum of squares based on $e_i$'s.

3. Run regression of $g$ on $Z_1, Z_2, \ldots, Z_p$ and get residual sum of squares $SS_{res}^*$.

4. For testing, calculate the test statistic

$$Q = \frac{1}{2} \left( \sum_{i=1}^{n} g_i^2 - SS_{res}^* \right)$$

which is asymptotically distributed as $\chi^2$ distribution with $(p-1)$ degrees of freedom.

5. The decision rule is to reject $H_0$ if $Q > \chi_{1-\alpha}^2 (m-1)$.

- This test is very simple to perform.
- A fairly general form is assumed for heteroskedasticity, so it is a very general test.
- This is an asymptotic test.
- This test is quite powerful in the presence of heteroskedasticity.

3. Goldfeld Quandt test

This test is based on the assumption that $\sigma_i^2$ is positively related to $X_{ij}$, i.e., one of the explanatory variable explains the heteroskedasticity in the model. Let $j^{th}$ explanatory variable explains the heteroskedasticity, so

$$\sigma_i^2 \propto X_{ij}$$

or $\sigma_i^2 = \sigma^2 X_{ij}$.

The test procedure is as follows:

1. Rank the observations according to the decreasing order of $X_j$.

2. Split the observations into two equal parts leaving $c$ observations in the middle.

So each part contains $\frac{n-c}{2}$ observations provided $\frac{n-c}{2} > k$.

3. Run two separate regression in the two parts using OLS and obtain the residual sum of squares $SS_{res1}$ and $SS_{res2}$.

4. The test statistic is $F_0 = \frac{SS_{res2}}{SS_{res1}}$

which follows a $F$-distribution, i.e., $F\left(\frac{n-c}{2} - k, \frac{n-c}{2} - k\right)$ when $H_0$ true.
5. The decision rule is to reject $H_0$ whenever $F_0 > F_{1-\alpha}\left(\frac{n-c-k}{2}, \frac{n-c-k}{2}\right)$.

- This test is simple test but it is based on the assumption that one of the explanatory variable helps is determining the heteroskedasticity.
- This test is an exact finite sample test.
- One difficulty in this test is that the choice of $c$ is not obvious. If large value of $c$ is chosen, then it reduces the degrees of freedom $\frac{n-c-k}{2}$ and the condition $\frac{n-c}{2} > k$ may be violated.

On the other hand, if smaller value of $c$ is chosen, then the test may fail to reveal the heteroskedasticity. The basic objective of ordering of observations and deletion of observations in the middle part may not reveal the heteroskedasticity effect. Since the first and last values of $\sigma^2_i$ give the maximum discretion, so deletion of smaller value may not give the proper idea of heteroskedasticity. Keeping those two points in view, the working choice of $c$ is suggested as $c = \frac{n}{3}$.

Moreover, the choice of $X_{ij}$ is also difficult. Since $\sigma^2_i \propto X_{ij}$, so if all important variables are included in the model, then it may be difficult to decide that which of the variable is influencing the heteroskedasticity.

4. Glesjer test:

This test is based on the assumption that $\sigma^2_i$ is influenced by one variable $Z$, i.e., there is only one variable which is influencing the heteroskedasticity. This variable could be either one of the explanatory variable or it can be chosen from some extraneous sources also.

The test procedure is as follows:
1. Use OLS and obtain the residual vector $e$ on the basis of available study and explanatory variables.
2. Choose $Z$ and apply OLS to
   $$|e_i| = \delta_0 + \delta_1 Z_i^h + v_i$$
   where $v_i$ is the associated disturbance term.
3. Test $H_0: \delta_1 = 0$ using $t$-ratio test statistic.
4. Conduct the test for $h = \pm 1, \pm \frac{1}{2}$. So the test procedure is repeated four times.

In practice, one can choose any value of $h$. For simplicity, we choose $h = 1$.

- The test has only asymptotic justification and the four choices of $h$ give generally satisfactory results.
- This test sheds light on the nature of heteroskedasticity.
5. Spearman’s rank correlation test

It $d_i$ denotes the difference in the ranks assigned to two different characteristics of the $i^{th}$ object or phenomenon and $n$ is the number of objects or phenomenon ranked, then the Spearman’s rank correlation coefficient is defined as

$$r = 1 - 6 \left( \frac{\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \right); -1 \leq r \leq 1.$$ 

This can be used for testing the hypothesis about the heteroskedasticity.

Consider the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$ 

1. Run the regression of $y$ on $X$ and obtain the residuals $e$.
2. Consider $|e|$.
3. Rank both $|e|$ and $X_i$ (or $\hat{y}_i$) in an ascending (or descending) order.
4. Compute rank correlation coefficient $r$ based on $|e|$ and $X_i$ (or $\hat{y}_i$).
5. Assuming that the population rank correlation coefficient is zero and $n > 8$, use the test statistic

$$t_0 = r \sqrt{\frac{n-2}{1-r^2}}$$

which follows a $t$-distribution with $(n-2)$ degrees of freedom.

6. The decision rule is to reject the null hypothesis of heteroskedasticity whenever $t_0 \geq t_{1-\alpha}(n-2)$.

If there are more than one explanatory variables, then rank correlation coefficient can be computed between $|e|$ and each of the explanatory variables separately and can be tested using $t_0$. 

---

Econometrics  | Chapter 8  | Heteroskedasticity  | Shalabh, IIT Kanpur
**Estimation under heteroskedasticity**

Consider the model

\[ y = X\beta + \varepsilon \]

with \( k \) explanatory variables and assume that

\[ E(\varepsilon) = 0 \]
\[ V(\varepsilon) = \Omega = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}. \]

The OLSE is

\[ b = (X'X)^{-1}X'y. \]

Its estimation error is

\[ b - \beta = (X'X)^{-1}X'\varepsilon \]

and

\[ E(b - \beta) = (X'X)^{-1}X'E(\varepsilon) = 0. \]

Thus OLSE remains unbiased even under heteroskedasticity.

The covariance matrix of \( b \) is

\[ V(b) = E((b - \beta)(b - \beta)') = (X'X)^{-1}X'E(\varepsilon\varepsilon')X(X'X)^{-1} = (X'X)^{-1}X'\Omega X(X'X)^{-1} \]

which is not the same as conventional expression. So OLSE is not efficient under heteroskedasticity as compared under homokedasticity.

Now we check if \( E(e_i^2) = \sigma_i^2 \) or not where \( e_i \) is the \( i^{th} \) residual.

The residual vector is

\[ e = y - Xb = \bar{H}\varepsilon \]

\[ e_i = [0, 0, \ldots, 0, 1, 0, \ldots 0]^\top \]

\[ = \ell_i'\bar{H}\varepsilon\varepsilon'\bar{H}\ell_i \]
where $\ell_i$ is a $n \times 1$ vector with all elements zero except the $i^{th}$ element which is unity and 
\[ H = I - X(X'X)^{-1}X'. \]
Then
\[ e_i^2 = \ell_i'e_i'e_i = \ell_i'HH\ell_i \]
\[ E(e_i^2) = \ell_i'HE(e'e')H\ell_i = \ell_i'HH\Omega H\ell_i \]
\[ H\ell_i = \begin{bmatrix} \overline{h}_{i1} & \cdots & \overline{h}_{in} \\ \vdots & \ddots & \vdots \\ \overline{h}_{n1} & \cdots & \overline{h}_{nn} \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} \overline{h}_{1i} \\ \vdots \\ \overline{h}_{ni} \end{bmatrix} \]
\[ E(e_i^2) = \begin{bmatrix} \overline{h}_{n1}i & \overline{h}_{n2}i & \cdots & \overline{h}_{ni}i \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} \begin{bmatrix} \overline{h}_{1i} \\ \vdots \\ \overline{h}_{ni} \end{bmatrix} \]

Thus $E(e_i^2) \neq \sigma_i^2$ and so $e_i^2$ becomes a biased estimator of $\sigma_i^2$ in the presence of heteroskedasticity.

In the presence of heteroskedasticity, use the generalized least squares estimation. The generalized least squares estimator (GLSE) of $\beta$ is
\[ \hat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y. \]
Its estimation error is obtained as
\[ \hat{\beta} - \beta = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}(X\beta + \varepsilon) \]
\[ = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}E(\varepsilon). \]
Thus
\[ E(\hat{\beta} - \beta) = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}E(\varepsilon) = 0 \]
\[ V(\hat{\beta}) = E(\hat{\beta} - \beta)(\hat{\beta} - \beta) \]
\[ = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}E(\varepsilon)\Omega^{-1}X(X'\Omega^{-1}X)^{-1} \]
\[ = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\Omega^{-1}X(X'\Omega^{-1}X)^{-1} \]
\[ = (X'\Omega^{-1}X)^{-1}. \]
Example: Consider a simple linear regression model

\[ y_i = \beta_0 + \beta_i x_i + \varepsilon_i, \quad i = 1, 2, ..., n. \]

The variances of OLSE and GLSE of \( \beta \) are

\[
Var(b) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sigma_i^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

\[
Var(\hat{\beta}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sqrt{\left[ \sum_{i=1}^{n} (x_i - \bar{x})^2 \sigma_i^2 \right] \left[ \sum_{i=1}^{n} (x_i - \bar{x})^2 \frac{1}{\sigma_i^2} \right]}}
\]

\[= \text{Square of the correlation coefficient between } \sigma_i (x_i - \bar{x}) \text{ and } \left( \frac{x_i - \bar{x}}{\sigma_i} \right) \leq 1 \]

\[\Rightarrow \text{Var}(\hat{\beta}) \leq \text{Var}(b). \]

So efficient of OLSE and GLSE depends upon the correlation coefficient between \((x_i - \bar{x})\sigma_i\) and \(\frac{(x_i - \bar{x})}{\sigma_i}\).

The generalized least squares estimation assumes that \( \Omega \) is known, i.e., the nature of heteroscedasticity is completely specified. Based on this assumption, the possibilities of following two cases arise:

- \( \Omega \) is completely specified or
- \( \Omega \) is not completely specified.

We consider both the cases as follows:

**Case 1: \( \sigma_i^2 \)'s are pre-specified:**

Suppose \( \sigma_1^2, \sigma_2^2, ..., \sigma_n^2 \) are completely known in the model

\[ y_i = \beta_1 + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \varepsilon_i. \]

Now deflate the model by \( \sigma_i \), i.e.,

\[ \frac{y_i}{\sigma_i} = \beta_1 \frac{1}{\sigma_i} + \beta_2 \frac{X_{i2}}{\sigma_i} + ... + \beta_k \frac{X_{ik}}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i}. \]

Let \( \varepsilon_i^* = \frac{\varepsilon_i}{\sigma_i} \), then \( E(\varepsilon_i^*) = 0, \text{Var}(\varepsilon_i^*) = \frac{\sigma_i^2}{\sigma_i^2} = 1 \). Now OLS can be applied to this model and usual tools for drawing statistical inferences can be used.

Note that when the model is deflated, the intercept term is lost as \( \beta_i / \sigma_i \) is itself a variable. This point has to be taken care in a software output.
Case 2: $\Omega$ may not be completely specified

Let $\sigma_i^2, \sigma_2^2, ..., \sigma_n^2$ are partially known and suppose

$$\sigma_i^2 \propto X_{ij}^{2.5}$$

or

$$\sigma_i^2 = \sigma^2 X_{ij}^{2.5}$$

but $\sigma^2$ is not available. Consider the model

$$y_i = \beta_1 + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \varepsilon_i$$

and deflate it by $X_{ij}^{2.5}$ as

$$\frac{y_i}{X_{ij}^{2.5}} = \frac{\beta_1}{X_{ij}^{2.5}} + \beta_2 \frac{X_{i2}}{X_{ij}^{2.5}} + ... + \beta_k \frac{X_{ik}}{X_{ij}^{2.5}} + \frac{\varepsilon_i}{X_{ij}^{2.5}}.$$

Now apply OLS to this transformed model and use the usual statistical tools for drawing inferences.

A caution is to be kept in mind while doing so. This is illustrated in the following example with one explanatory variable model.

Consider the model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$ 

Deflate it by $x_i$, so we get

$$\frac{y_i}{x_i} = \frac{\beta_0}{x_i} + \beta_1 + \frac{\varepsilon_i}{x_i}.$$

Note that the roles of $\beta_0$ and $\beta_1$ in original and deflated models are interchanged. In original model, $\beta_0$ is intercept term and $\beta_1$ is slope parameter whereas in deflated model, $\beta_1$ becomes the intercept term and $\beta_0$ becomes the slope parameter. So essentially, one can use OLS but need to be careful in identifying the intercept term and slope parameter, particularly in the software output.
Chapter 9  
Autocorrelation

One of the basic assumptions in linear regression model is that the random error components or disturbances are identically and independently distributed. So in the model \( y = X\beta + u \), it is assumed that

\[
E(u_t, u_{t-s}) = \begin{cases} 
\sigma_u^2 & \text{if } s = 0 \\
0 & \text{if } s \neq 0 
\end{cases}
\]
i.e., the correlation between the successive disturbances is zero.

In this assumption, when \( E(u_t, u_{t-s}) = \sigma_u^2, s = 0 \) is violated, i.e., the variance of disturbance term does not remains constant, then problem of heteroskedasticity arises. When \( E(u_t, u_{t-s}) = 0, s \neq 0 \) is violated, i.e., the variance of disturbance term remains constant though the successive disturbance terms are correlated, then such problem is termed as problem of autocorrelation.

When autocorrelation is present, some or all off diagonal elements in \( E(\text{uu}') \) are nonzero.

Sometimes the study and explanatory variables have a natural sequence order over time, i.e., the data is collected with respect to time. Such data is termed as **time series data**. The disturbance terms in time series data are serially correlated.

The **autocovariance** at lag \( s \) is defined as

\[
\gamma_s = E(u_t, u_{t-s}), \ s = 0, \pm 1, \pm 2, \ldots
\]

At zero lag, we have constant variance, i.e.,

\[
\gamma_0 = E(u_t^2) = \sigma^2.
\]

The **autocorrelation coefficient** at lag \( s \) is defined as

\[
\rho_s = \frac{E(u_t, u_{t-s})}{\sqrt{Var(u_t)Var(u_{t-s})}} = \frac{\gamma_s}{\gamma_0}; \ s = 0, \pm 1, \pm 2, \ldots
\]

Assume \( \rho_s \) and \( \gamma_s \) are symmetrical in \( s \), i.e., these coefficients are constant over time and depend only on length of lag \( s \). The autocorrelation between the successive terms \((u_2 \text{ and } u_1), (u_3 \text{ and } u_2), \ldots, (u_n \text{ and } u_{n-1})\) gives the autocorrelation of order one, i.e., \( \rho_1 \). Similarly, the autocorrelation

---

*Econometrics | Chapter 9 | Autocorrelation | Shalabh, IIT Kanpur*
between the successive terms \((u_1\text{ and } u_0), (u_4\text{ and } u_2), \ldots, (u_n\text{ and } u_{n-2})\) gives the autocorrelation of order two, i.e., \(\rho_2\).

**Source of autocorrelation**

Some of the possible reasons for the introduction of autocorrelation in the data are as follows:

1. Carryover of effect, at least in part, is an important source of autocorrelation. For example, the monthly data on expenditure on household is influenced by the expenditure of preceding month. The autocorrelation is present in cross-section data as well as time series data. In the cross-section data, the neighboring units tend to be similar with respect to the characteristic under study. In time series data, the time is the factor that produces autocorrelation. Whenever some ordering of sampling units is present, the autocorrelation may arise.

2. Another source of autocorrelation is the effect of deletion of some variables. In regression modeling, it is not possible to include all the variables in the model. There can be various reasons for this, e.g., some variable may be qualitative, sometimes direct observations may not be available on the variable etc. The joint effect of such deleted variables gives rise to autocorrelation in the data.

3. The misspecification of the form of relationship can also introduce autocorrelation in the data. It is assumed that the form of relationship between study and explanatory variables is linear. If there are log or exponential terms present in the model so that the linearity of the model is questionable then this also gives rise to autocorrelation in the data.

4. The difference between the observed and true values of variable is called measurement error or errors-in-variable. The presence of measurement errors on the dependent variable may also introduce the autocorrelation in the data.
Structure of disturbance term:

Consider the situation where the disturbances are autocorrelated,

\[ E(\varepsilon \varepsilon') = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n-1} & \gamma_{n-2} & \cdots & \gamma_0 \end{bmatrix} = \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{bmatrix} \]

Observe that now there are \((n+k)\) parameters- \(\beta_1, \beta_2, \ldots, \beta_k, \sigma_u, \rho_1, \rho_2, \ldots, \rho_{n-1}\). These \((n+k)\) parameters are to be estimated on the basis of available \(n\) observations. Since the number of parameters are more than the number of observations, so the situation is not good from the statistical point of view. In order to handle the situation, some special form and the structure of the disturbance term is needed to be assumed so that the number of parameters in the covariance matrix of disturbance term can be reduced.

The following structures are popular in autocorrelation:

1. **Autoregressive (AR) process.**
2. **Moving average (MA) process.**
3. **Joint autoregression moving average (ARMA) process.**

1. **Autoregressive (AR) process**

The structure of disturbance term in autoregressive process (AR) is assumed as

\[ u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \ldots + \phi_q u_{t-q} + \varepsilon_t, \]

i.e., the current disturbance term depends on the \(q\) lagged disturbances and \(\phi_1, \phi_2, \ldots, \phi_k\) are the parameters (coefficients) associated with \(u_{t-1}, u_{t-2}, \ldots, u_{t-q}\) respectively. An additional disturbance term is introduced in \(u_t\) which is assumed to satisfy the following conditions:
\[ E(\varepsilon_t) = 0 \]
\[ E(\varepsilon_t\varepsilon_{t-s}) = \begin{cases} \sigma^2_{\varepsilon} & \text{if } s = 0 \\ 0 & \text{if } s \neq 0. \end{cases} \]

This process is termed as $AR(q)$ process. In practice, the $AR(1)$ process is more popular.

2. Moving average (MA) process:

The structure of disturbance term in the moving average (MA) process is
\[ u_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_p \varepsilon_{t-p}, \]
i.e., the present disturbance term $u_t$ depends on the $p$ lagged values. The coefficients $\theta_1, \theta_2, \ldots, \theta_p$ are the parameters and are associated with $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-p}$ respectively. This process is termed as $MA(p)$ process.

3. Joint autoregressive moving average (ARMA) process:

The structure of disturbance term in the joint autoregressive moving average (ARMA) process is
\[ u_t = \phi_1 u_{t-1} + \ldots + \phi_q u_{t-q} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_p \varepsilon_{t-p}. \]
This is termed as $ARMA(q, p)$ process.

The method of correlogram is used to check that the data is following which of the processes. The correlogram is a two dimensional graph between the lag $s$ and autocorrelation coefficient $\rho_s$ which is plotted as lag $s$ on $X$-axis and $\rho_s$ on $y$-axis.

In $MA(1)$ process
\[ u_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} \]
\[ \rho_s = \begin{cases} \frac{\theta_1}{1 + \theta_1^2} & \text{for } s = 1 \\ 0 & \text{for } s \geq 2 \end{cases} \]
\[ \rho_0 = 1 \]
\[ \rho_s = 0 \]
\[ \rho_i = 0 \quad i = 2, 3, \ldots \]
So there is no autocorrelation between the disturbances that are more than one period apart.
In $ARMA(1,1)$ process

$$u_t = \phi u_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$$

$$\rho_s = \begin{cases} 
(1 + \phi \theta)(\phi + \theta) & \text{for } s = 1 \\
(1 + \theta^2 + 2\phi \theta) & \phi \rho_{s-1} & \text{for } s \geq 2 
\end{cases}$$

$$\sigma_u^2 = \left(\frac{1 + \theta^2 + 2\phi \theta}{1 - \phi^2}\right) \sigma^2.$$

The autocorrelation function begins at some point determined by both the $AR$ and $MA$ components but thereafter, declines geometrically at a rate determined by the $AR$ component.

In general, the autocorrelation function

- is nonzero but is geometrically damped for $AR$ process.
- becomes zero after a finite number of periods for $MA$ process.

The $ARMA$ process combines both these features.

The results of any lower order of process are not applicable in higher order schemes. As the order of the process increases, the difficulty in handling them mathematically also increases.

**Estimation under the first order autoregressive process:**

Consider a simple linear regression model

$$y_t = \beta_0 + \beta X_t + u_t, \ t = 1, 2, \ldots, n.$$ 

Assume $u_t$'s follow a first order autoregressive scheme defined as

$$u_t = \rho u_{t-1} + \varepsilon_t$$

where $|\rho| < 1$, $E(\varepsilon_t) = 0$,

$$E(\varepsilon_t, \varepsilon_{t+s}) = \begin{cases} 
\sigma^2 & \text{if } s = 0 \\
0 & \text{if } s \neq 0
\end{cases}$$

for all $t = 1, 2, \ldots, n$ where $\rho$ is the first order autocorrelation between $u_t$ and $u_{t-1}$, $t = 1, 2, \ldots, n$. Now

$$u_t = \rho u_{t-1} + \varepsilon_t$$

$$= \rho(u_{t-2} + \varepsilon_{t-1}) + \varepsilon_t$$

$$= \vdots$$

$$= \varepsilon_t + \rho \varepsilon_{t-1} + \rho^2 \varepsilon_{t-2} + \ldots$$

$$= \sum_{r=0}^{\infty} \rho^r \varepsilon_{t-r}.$$
\( E(u_t) = 0 \)
\[ E(u_t^2) = E(\varepsilon_t^2) + \rho^2 E(\varepsilon_{t-1}^2) + \rho^4 E(\varepsilon_{t-2}^2) + \ldots \]
\[ = (1 + \rho^2 + \rho^4 + \ldots)\sigma_\varepsilon^2 \] (\( \varepsilon_t \)'s are serially independent)
\[ E(u_t^2) = \sigma_u^2 = \frac{\sigma_\varepsilon^2}{1 - \rho^2} \text{ for all } t. \]

\[
E(u_{t},u_{t-1}) = E\left[ \left( \varepsilon_t + \rho \varepsilon_{t-1} + \rho^2 \varepsilon_{t-2} + \ldots \right) \times \left( \varepsilon_{t-1} + \rho \varepsilon_{t-2} + \rho^2 \varepsilon_{t-3} + \ldots \right) \right]
= E\left[ \varepsilon_t^2 + \rho E(\varepsilon_{t-1}^2) + \rho^2 E(\varepsilon_{t-2}^2) + \ldots \right]
= \rho \sigma_u^2.
\]

Similarly,
\[ E(u_{t},u_{t-2}) = \rho^2 \sigma_u^2. \]

In general,
\[
E(u_{t},u_{t-s}) = \rho^s \sigma_u^2 \left( \begin{array}{c c c c}
1 & \rho & \rho^2 & \ldots & \rho^{n-1} \\
\rho & 1 & \rho & \ldots & \rho^{n-2} \\
\rho^2 & \rho & 1 & \ldots & \rho^{n-3} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \ldots & 1
\end{array} \right).
\]

Note that the disturbance terms are no more independent and \( E(uu') \neq \sigma^2 I \). The disturbance are nonspherical.

**Consequences of autocorrelated disturbances:**

Consider the model with first order autoregressive disturbances
\[
y_n = X_{nx1} \beta + u_{nkx1} + \varepsilon_{n} \\
u_t = \rho u_{t-1} + \varepsilon_t, \quad t = 1, 2, \ldots, n
\]
with assumptions
\[ E(u) = 0, \quad E(uu') = \Omega \]
\[ E(\varepsilon) = 0, \quad E(\varepsilon_i\varepsilon_{t+s}) = \begin{cases} \sigma_\varepsilon^2 & \text{if } s = 0 \\ 0 & \text{if } s \neq 0 \end{cases} \]
where \( \Omega \) is a positive definite matrix.
The ordinary least squares estimator of $\beta$ is

$$b = (X'X)^{-1} X'y$$

$$= (X'X)^{-1} X'(X\beta+u)$$

$$b - \beta = (X'X)^{-1} X'u$$

$$E(b-\beta) = 0.$$ 

So OLSE remains unbiased under autocorrelated disturbances.

The covariance matrix of $b$ is

$$V(b) = E(b-\beta)(b-\beta)'$$

$$= (X'X)^{-1} X' E(uu')X(X'X)^{-1}$$

$$= (X'X)^{-1} X' \Omega X(X'X)^{-1}$$

$$\neq \sigma_u^2 (X'X)^{-1}.$$ 

The residual vector is

$$e = y - Xb = \bar{Hy} = \bar{Hu}$$

$$e'e = y'\bar{Hy} = u'\bar{Hu}$$

$$E(e'e) = E(u'u) - E\left[u'X(X'X)^{-1}X'u\right]$$

$$= n\sigma_u^2 - tr(X'X)^{-1}X'\Omega X.$$ 

Since $s^2 = \frac{e'e}{n-1}$, so

$$E(s^2) = \frac{\sigma_u^2}{n-1} - \frac{1}{n-1} tr(X'X)^{-1}X'\Omega X,$$

so $s^2$ is a biased estimator of $\sigma^2$. In fact, $s^2$ has downward bias.

Application of OLS fails in case of autocorrelation in the data and leads to serious consequences as

— overly optimistic view from $R^2$.
— narrow confidence interval.
— usual $t$-ratio and $F$-ratio tests provide misleading results.
— prediction may have large variances.

Since disturbances are nonspherical, so generalized least squares estimate of $\beta$ yields more efficient estimates than OLSE.
The GLSE of $\beta$ is

$$\hat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$$

$$E(\hat{\beta}) = \beta$$

$$V(\hat{\beta}) = \sigma^2_u (X'\Omega^{-1}X)^{-1}.$$

The GLSE is best linear unbiased estimator of $\beta$.

**Tests for autocorrelation:**

**Durbin Watson test:**

The Durbin-Watson ($D$-$W$) test is used for testing the hypothesis of lack of first order autocorrelation in the disturbance term. The null hypothesis is

$$H_0 : \rho = 0$$

Use OLS to estimate $\beta$ in $y = X\beta + u$ and obtain residual vector

$$e = y - Xb = \bar{H}y$$

where $b = (X'X)^{-1}X'y$, $\bar{H} = I - X(X'X)^{-1}X'$.

The $D$-$W$ test statistic is

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

$$= \frac{\sum_{t=2}^{n} e_t^2 - \sum_{t=2}^{n} e_{t-1}^2 - 2 \sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2 - \sum_{t=1}^{n} e_t^2}.$$

For large $n$,

$$d \approx 1 + 1 - 2r$$

$$d = 2(1 - r)$$

where $r$ is the sample autocorrelation coefficient from residuals based on OLSE and can be regarded as the regression coefficient of $e_t$ on $e_{t-1}$. Here
positive autocorrelation of \( e_i \)'s \( \Rightarrow d < 2 \)

negative autocorrelation of \( e_i \)'s \( \Rightarrow d > 2 \)

zero autocorrelation of \( e_i \)'s \( \Rightarrow d \approx 2 \)

As \(-1 < r < 1\), so

if \(-1 < r < 0\), then \(2 < d < 4\) and

if \(0 < r < 1\), then \(0 < d < 2\).

So \(d\) lies between 0 and 4.

Since \(e\) depends on \(X\), so for different data sets, different values of \(d\) are obtained. So the sampling distribution of \(d\) depends on \(X\). Consequently exact critical values of \(d\) cannot be tabulated owing to their dependence on \(X\). Durbin and Watson therefore obtained two statistics \(d\) and \(\overline{d}\) such that

\[d < \overline{d} \]

and their sampling distributions do not depend upon \(X\).

Considering the distribution of \(d\) and \(\overline{d}\), they tabulated the critical values as \(d_L\) and \(d_U\) respectively.

They prepared the tables of critical values for \(15 < n < 100\) and \(k \leq 5\). Now tables are available for \(6 < n < 200\) and \(k \leq 10\).

The test procedure is as follows:

<table>
<thead>
<tr>
<th>(H_0: \rho = 0)</th>
<th>Reject (H_0) when</th>
<th>Retain (H_0) when</th>
<th>The test is inconclusive when</th>
</tr>
</thead>
<tbody>
<tr>
<td>(H_1: \rho &gt; 0)</td>
<td>(d &lt; d_L)</td>
<td>(d &gt; d_U)</td>
<td>(d_L &lt; d &lt; d_U)</td>
</tr>
<tr>
<td>(H_1: \rho &lt; 0)</td>
<td>(d &gt; (4 - d_L))</td>
<td>(d &lt; (4 - d_U))</td>
<td>((4 - d_U) &lt; d &lt; (4 - d_L))</td>
</tr>
<tr>
<td>(H_1: \rho \neq 0)</td>
<td>(d &lt; d_L) or (d &gt; (4 - d_L))</td>
<td>(d_U &lt; d &lt; (4 - d_U))</td>
<td>(d_L &lt; d &lt; d_U) or ((4 - d_U) &lt; d &lt; (4 - d_L))</td>
</tr>
</tbody>
</table>

Values of \(d_L\) and \(d_U\) are obtained from tables.
Limitations of D-W test

1. If $d$ falls in the inconclusive zone, then no conclusive inference can be drawn. This zone becomes fairly larger for low degrees of freedom. One solution is to reject $H_0$ if the test is inconclusive. A better solution is to modify the test as
   - Reject $H_0$ when $d < d_U$.
   - Accept $H_0$ when $d \geq d_U$.
   This test gives satisfactory solution when values of $x_i$’s change slowly, e.g., price, expenditure etc.

2. The D-W test is not applicable when intercept term is absent in the model. In such a case, one can use another critical values, say $d_M$ in place of $d_L$. The tables for critical values $d_M$ are available.

3. The test is not valid when lagged dependent variables appear as explanatory variables. For example,

   $y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \beta_r y_{t-r} + \beta_{r+1} x_{t-1} + \ldots + \beta_k x_{t-k-r} + u_t,$

   $u_t = \rho u_{t-1} + \varepsilon_t$.

   In such case, Durbin’s $h$ test is used which is given as follows.

Durbin’s $h$-test

Apply OLS to

$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \beta_r y_{t-r} + \beta_{r+1} x_{t-1} + \ldots + \beta_k x_{t-k-r} + u_t,$

$u_t = \rho u_{t-1} + \varepsilon_t$,

and find OLSE $b_1$ of $\beta_1$. Let its variance be $\hat{Var}(b_1)$ and its estimator is $\hat{Var}(b_1)$. Then the Durbin’s $h$-statistic is

$$h = r \sqrt{\frac{n}{1 - n \hat{Var}(b_1)}}$$

which is asymptotically distributed as $N(0,1)$ and

$$r = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_t^2}.$$
This test is applicable when \( n \) is large. When \( \left[ 1 - n\hat{\text{Var}}(b_1) \right] < 0 \), then test breaks down. In such cases, the following test procedure can be adopted.

Introduce a new variable \( \varepsilon_{t-1} \) to \( u_t = \rho u_{t-1} + \varepsilon_t \). Then

\[
e_t = \delta \rho_{t-1} + y_t.
\]

Now apply OLS to this model and test \( H_{0,A}: \delta = 0 \) versus \( H_{1,A}: \delta \neq 0 \) using \( t \)-test. If \( H_{0,A} \) is accepted then accept \( H_0: \rho = 0 \).

If \( H_{0,A}: \delta = 0 \) is rejected, then reject \( H_0: \rho = 0 \).

4. If \( H_0: \rho = 0 \) is rejected by \( D-W \) test, it does not necessarily mean the presence of first order autocorrelation in the disturbances. It could happen because of other reasons also, e.g.,

- distribution may follow higher order \( AR \) process.
- some important variables are omitted.
- dynamics of model is misspecified.
- functional term of model is incorrect.

**Estimation procedures with autocorrelated errors when autocorrelation coefficient is known**

Consider the estimation of regression coefficient under first order autoregressive disturbances and autocorrelation coefficient is known. The model is

\[
y = X \beta + u,
\]

\[
u_t = \rho u_{t-1} + \varepsilon_t,
\]

and assume that \( E(u) = 0, E(u'u') = \psi \neq \sigma^2 I, E(\varepsilon) = 0, E(\varepsilon\varepsilon') = \sigma^2 \varepsilon I \).

The OLSE of \( \beta \) is unbiased but not, in general, efficient and estimate of \( \sigma^2 \) is biased. So we use generalized least squares estimation procedure and GLSE of \( \beta \) is

\[
\hat{\beta} = (X'\psi^{-1}X)^{-1}X'\psi^{-1}y
\]

where
To employ this, we proceed as follows:

1. Find a matrix $P$ such that $P'P = \psi^{-1}$. In this case

$$
P = \begin{bmatrix}
\sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 & 0 \\
-\rho & 1 & 0 & \cdots & 0 & 0 \\
0 & -\rho & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & 0 \\
0 & 0 & 0 & \cdots & -\rho & 1
\end{bmatrix}.
$$

2. Transform the variables as

$$
y^* = Py, \quad X^* = PX, \quad e^* = Pe.
$$

Such transformation yields

$$
y^* = \begin{bmatrix}
\sqrt{1-\rho^2} y_1 \\
y_2 - \rho y_1 \\
y_3 - \rho y_2 \\
\vdots \\
y_n - \rho y_{n-1}
\end{bmatrix}, \quad X^* = \begin{bmatrix}
\sqrt{1-\rho^2} x_{i1} & \sqrt{1-\rho^2} x_{i2} & \cdots & \sqrt{1-\rho^2} x_{i,k} \\
1-\rho & x_{21} - \rho x_{i1} & \cdots & x_{2,k} - \rho x_{i,k} \\
1-\rho & x_{31} - \rho x_{21} & \cdots & x_{3,k} - \rho x_{2,k} \\
\vdots & \vdots & \ddots & \vdots \\
1-\rho & x_{n1} - \rho x_{n-1,1} & \cdots & x_n - \rho x_{n-1,n}
\end{bmatrix}.
$$

Note that the first observation is treated differently than other observations. For the first observation,

$$
\left(\sqrt{1-\rho^2}\right) y_1 = \left(\sqrt{1-\rho^2}\right) x_i \beta + \left(\sqrt{1-\rho^2}\right) u_i
$$

whereas for other observations

$$
y_t = \rho y_{t-1} = (x_t - \rho x_{t-1})' \beta + (u_t - \rho u_{t-1}); \quad t = 2, 3, \ldots, n
$$

where $x_t'$ is a row vector of $X$. Also, $\sqrt{1-\rho^2} u_i$ and $(u_t - \rho u_{t-1})$ have same properties. So we expect these two errors to be uncorrelated and homoscedastic.
If first column of $X$ is a vector of ones, then first column of $X^*$ is not constant. Its first element is $\sqrt{1-\rho^2}$.

Now employ OLSE with observations $y^*$ and $X^*$, then the OLSE of $\beta$ is

$$\beta^* = (X^* X^*)^{-1} X^* y^*,$$

its covariance matrix is

$$V(\hat{\beta}) = \sigma^2 (X^* X^*)^{-1}$$

$$= \sigma^2 (X'\psi^{-1} X)^{-1}$$

and its estimator is

$$\hat{V}(\hat{\beta}) = \hat{\sigma}^2 (X'\psi^{-1} X)^{-1}$$

where

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta})'\psi^{-1}(y - X\hat{\beta})}{n - k}.$$

### Estimation procedures with autocorrelated errors when autocorrelation coefficient is unknown

Several procedure have been suggested to estimate the regression coefficients when autocorrelation coefficient is unknown. The feasible GLSE of $\beta$ is

$$\hat{\beta}_F = (X'\hat{\Omega}^{-1} X)^{-1} X'\hat{\Omega}^{-1} y$$

where $\hat{\Omega}^{-1}$ is the $\Psi^{-1}$ matrix with $\rho$ replaced by its estimator $\hat{\rho}$.

#### 1. Use of sample correlation coefficient

Most common method is to use the sample correlation coefficient $r$ between successive residuals as the natural estimator of $\rho$. The sample correlation can be estimated using the residuals in place of disturbances as

$$r = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_t^2}$$

where $e_t = y_t - \hat{x}_t\hat{b}$, $t = 1, 2, ..., n$ and $\hat{b}$ is OLSE of $\beta$.

Two modifications are suggested for $r$ which can be used in place of $r$. 

---

*Econometrics | Chapter 9 | Autocorrelation | Shalabh, IIT Kanpur*
1. \( r^* = \left( \frac{n-k}{n-1} \right) r \) is the Theil’s estimator.

2. \( r^{**} = 1 - \frac{d}{2} \) for large \( n \) where \( d \) is the Durbin Watson statistic for \( H_0 : \rho = 0 \).

2. Durbin procedure:

In Durbin procedure, the model
\[
y_t - \rho y_{t-1} = \beta_0(1 - \rho) + \beta(x_t - \rho x_{t-1}) + \varepsilon_t, \ t = 2,3,...,n
\]
is expressed as
\[
y_t = \beta_0(1 - \rho) + \rho y_{t-1} + \beta x_t - \rho \beta x_{t-1} + \varepsilon_t
\]
\[
= \beta_0^* + \rho y_{t-1} + \beta x_t + \beta^* x_{t-1} + \varepsilon_t, \ t = 2,3,...,n \quad (*)
\]
where \( \beta_0^* = \beta_0(1 - \rho) \), \( \beta^* = -\rho \beta \).

Now run regression using OLS to model (*) and estimate \( r^* \) as the estimated coefficient of \( y_{t-1} \).

Another possibility is that since \( \rho \in (-1,1) \), so search for a suitable \( \rho \) which has smaller error sum of squares.

3. Cochrane-Orcutt procedure:

This procedure utilizes \( P \) matrix defined while estimating \( \beta \) when \( \rho \) is known. It has following steps:

(i) Apply OLS to \( y_t = \beta_0 + \beta_1 x_t + u_t \) and obtain residual vector \( e \).

(ii) Estimate \( \rho \) by \( r = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e^2_{t-1}} \).

Note that \( r \) is a consistent estimator of \( \rho \).

(iii) Replace \( \rho \) by \( r \) is
\[
y_t - \rho y_{t-1} = \beta_0^*(1 - \rho) + \beta(x_t - \rho x_{t-1}) + \varepsilon_t
\]
and apply OLS to transformed model
\[
y_t - r y_{t-1} = \beta_0^* + \beta(x_t - r x_{t-1}) + \text{disturbance term}
\]
and obtain estimators of \( \beta_0^* \) and \( \beta \) as \( \hat{\beta}_0^* \) and \( \hat{\beta} \) respectively.

This is Cochrane-Orcutt procedure. Since two successive applications of OLS are involved, so it is also called as two-step procedure.
This application can be repeated in the procedure as follows:

(I) Put \( \hat{\beta}_0 \) and \( \hat{\beta} \) in original model.

(II) Calculate the residual sum of squares.

(III) Calculate \( \rho \) by

\[
\rho = \frac{\sum_{t=2}^{n} e_{t-1}^2}{\sum_{t=2}^{n} e_{t-1}^2}
\]

and substitute it in the model

\[
y_t - \rho y_{t-1} = \beta_0(1 - \rho) + \beta(x_t - \rho x_{t-1}) + \epsilon_t,
\]

and again obtain the transformed model.

(IV) Apply OLS to this model and calculate the regression coefficients.

This procedure is repeated until convergence is achieved, i.e., iterate the process till the two successive estimates are nearly same so that stability of estimator is achieved.

This is an iterative procedure and is numerically convergent procedure. Such estimates are asymptotically efficient and there is a loss of one observation.

4. Hildreth-Lu procedure or Grid-search procedure:

The Hildreth-Lu procedure has following steps:

(i) Apply OLS to

\[
(y_t - \rho y_{t-1}) = \beta_0(1 - \rho) + \beta(x_t - \rho x_{t-1}) + \epsilon_t, \quad t = 2, 3, ..., n
\]

using different values of \( \rho \) \((-1 \leq \rho \leq 1)\) such as \( \rho = \pm 0.1, \pm 0.2, ... \).

(ii) Calculate residual sum of squares in each case.

(iii) Select that value of \( \rho \) for which residual sum of squares is smallest.

Suppose we get \( \rho = 0.4 \). Now choose a finer grid. For example, choose \( \rho \) such that \( 0.3 < \rho < 0.5 \) and consider \( \rho = 0.31, 0.32, ..., 0.49 \) and pick up that \( \rho \) with smallest residual sum of squares. Such iteration can be repeated until a suitable value of \( \rho \) corresponding to minimum residual sum of squares is obtained.

The selected final value of \( \rho \) can be used and for transforming the model as in the case of Cochrane-Orcutt procedure. The estimators obtained with this procedure are as efficient as obtained by Cochrane-Orcutt procedure and there is a loss of one observation.
5. Prais-Winston procedure

This is also an iterative procedure based on two step transformation.

(i) Estimate $\rho$ by $\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=3}^{n} e_t^2}$ where $e_t$’s are residuals based on OLSE.

(ii) Replace $\rho$ by $\hat{\rho}$ is the model as in Cochrane-Orcutt procedure

$$\begin{align*}
\left(\sqrt{1-\hat{\rho}^2}\right)y_t & = \left(\sqrt{1-\hat{\rho}^2}\right)\beta_0 + \beta \left(\sqrt{1-\hat{\rho}^2}\right)x_t + \left(\sqrt{1-\hat{\rho}^2}\right)\epsilon_t \\
y_t - \hat{\rho}y_{t-1} & = (1-\hat{\rho})\beta_0 + \beta(x_t - \hat{\rho}x_{t-1}) + (\epsilon_t - \hat{\rho}\epsilon_{t-1}), \ t = 2,3,...,n.
\end{align*}$$

(iii) Use OLS for estimating the parameters.

The estimators obtained with this procedure are asymptotically as efficient as best linear unbiased estimators. There is no loss of any observation.

(6) Maximum likelihood procedure

Assuming that $y \sim N(X\beta, \sigma^2\psi)$, the likelihood function for $\beta, \rho$ and $\sigma^2$ is

$$L = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \left| \begin{array}{c} \psi \\ \psi \end{array} \right|^{\frac{n}{2}} \exp\left[ -\frac{1}{2\sigma^2} (y - X\beta)'\psi^{-1}(y - X\beta) \right].$$

Ignoring the constant and using $|\psi| = \frac{1}{1-\rho^2}$, the log-likelihood is

$$\ln L = \ln L(\beta, \sigma^2, \rho) = \frac{-n}{2} \ln \sigma^2 + \frac{1}{2} \ln(1-\rho^2) - \frac{1}{2\sigma^2} (y - X\beta)'\psi^{-1}(y - X\beta).$$

The maximum likelihood estimators of $\beta, \rho$ and $\sigma^2$ can be obtained by solving the normal equations

$$\frac{\partial \ln L}{\partial \beta} = 0, \quad \frac{\partial \ln L}{\partial \rho} = 0, \quad \frac{\partial \ln L}{\partial \sigma^2} = 0.$$

There normal equations turn out to be nonlinear in parameters and can not be easily solved.

One solution is to

- first derive the maximum likelihood estimator of $\sigma^2$. 

Econometrics | Chapter 9 | Autocorrelation | Shalab, IIT Kanpur
- Substitute it back into the likelihood function and obtain the likelihood function as the function of $\beta$ and $\rho$.

- Maximize this likelihood function with respect to $\beta$ and $\rho$.

Thus

$$\frac{\partial \ln L}{\partial \sigma^2_{\epsilon}} = 0 \Rightarrow -\frac{n}{2\sigma^2_{\epsilon}} + \frac{1}{2\sigma^2_{\epsilon}} (y - X \beta)'\psi^{-1}(y - X \beta) = 0$$

$$\Rightarrow \hat{\sigma}^2_{\epsilon} = \frac{1}{n} (y - X \beta)'\psi^{-1}(y - X \beta)$$

is the estimator of $\sigma^2_{\epsilon}$.

Substituting $\hat{\sigma}^2_{\epsilon}$ in place of $\sigma^2_{\epsilon}$ in the log-likelihood function yields

$$\ln L^* = \ln L^*(\beta, \rho) = -\frac{n}{2} \ln \left[ \frac{1}{n} (y - X \beta)'\psi^{-1}(y - X \beta) \right] + \frac{1}{2} \ln(1 - \rho^2) - \frac{n}{2}$$

$$= -\frac{n}{2} \left[ \ln \left( (y - X \beta)'\psi^{-1}(y - X \beta) \right) - \frac{1}{n} \ln(1 - \rho^2) \right] + k$$

$$= k - \frac{n}{2} \ln \left( \frac{(y - X \beta)'\psi^{-1}(y - X \beta)}{(1 - \rho^2)^{1/2}} \right)$$

where $k = \frac{n}{2} \ln n - \frac{n}{2}$.

Maximization of $\ln L^*$ is equivalent to minimizing the function

$$\frac{(y - X \beta)'\psi^{-1}(y - X \beta)}{(1 - \rho^2)^{1/2}}.$$

Using optimization techniques of non-linear regression, this function can be minimized and estimates of $\beta$ and $\rho$ can be obtained.

If $n$ is large and $|\rho|$ is not too close to one, then the term $(1 - \rho^2)^{-1/n}$ is negligible and the estimates of $\beta$ will be same as obtained by nonlinear least squares estimation.
In general, the explanatory variables in any regression analysis are assumed to be quantitative in nature. For example, the variables like temperature, distance, age etc. are quantitative in the sense that they are recorded on a well defined scale.

In many applications, the variables can not be defined on a well defined scale and they are qualitative in nature.

For example, the variables like sex (male or female), colour (black, white), nationality, employment status (employed, unemployed) are defined on a nominal scale. Such variables do not have any natural scale of measurement. Such variables usually indicate the presence or absence of a “quality” or an attribute like employed or unemployed, graduate or non-graduate, smokers or non-smokers, yes or no, acceptance or rejection, so they are defined on a nominal scale. Such variables can be quantified by artificially constructing the variables that take the values, e.g., 1 and 0 where “1” indicates usually the presence of attribute and “0” indicates usually the absence of attribute. For example, “1” indicates that the person is male and “0” indicates that the person is female. Similarly, “1” may indicate that the person is employed and then “0” indicates that the person is unemployed.

Such variables classify the data into mutually exclusive categories. These variables are called *indicator variables* or *dummy variables*.

Usually, the dummy variables take on the values 0 and 1 to identify the mutually exclusive classes of the explanatory variables. For example,

\[
D = \begin{cases} 
1 & \text{if person is male} \\
0 & \text{if person is female,}
\end{cases}
\]

\[
D = \begin{cases} 
1 & \text{if person is employed} \\
0 & \text{if person is unemployed.}
\end{cases}
\]

Here we use the notation $D$ in place of $X$ to denote the dummy variable. The choice of 1 and 0 to identify a category is arbitrary. For example, one can also define the dummy variable in above examples as

---

*Econometrics* | Chapter 10 | Dummy Variable Models | *Shalabh, IIT Kanpur*
$D = \begin{cases} 
1 & \text{if person is female} \\
0 & \text{if person is male,} 
\end{cases}$

$D = \begin{cases} 
1 & \text{if person is unemployed} \\
0 & \text{if person is employed.} 
\end{cases}$

It is also not necessary to choose only 1 and 0 to denote the category. In fact, any distinct value of $D$ will serve the purpose. The choices of 1 and 0 are preferred as they make the calculations simple, help in easy interpretation of the values and usually turn out to be a satisfactory choice.

In a given regression model, the qualitative and quantitative variables may also occur together, i.e., some variables may be qualitative and others are quantitative.

When all explanatory variables are
- **quantitative**, then the model is called a **regression model**,  
- **qualitative**, then the model is called an **analysis of variance model** and  
- **quantitative and qualitative both**, then the model is called a **analysis of covariance model**.

Such models can be dealt within the framework of regression analysis. The usual tools of regression analysis can be used in case of dummy variables.

**Example:**

Consider the following model with $x_i$ as quantitative and $D_2$ as dummy variable

$$ y = \beta_0 + \beta_1 x_i + \beta_2 D_2 + \epsilon, \quad E(\epsilon) = 0, Var(\epsilon) = \sigma^2 $$

$$ D_2 = \begin{cases} 
0 & \text{if an observation belongs to group } A \\
1 & \text{if an observation belongs to group } B. 
\end{cases} $$

The interpretation of result is important. We proceed as follows:

If $D_2 = 0$, then

$$ y = \beta_0 + \beta_1 x_i + \beta_2 0 + \epsilon $$

$$ = \beta_0 + \beta_1 x_i + \epsilon $$

$$ E(y | D_2 = 0) = \beta_0 + \beta_1 x_i $$

which is a straight line relationship with intercept $\beta_0$ and slope $\beta_1$. 

---

*Econometrics* | Chapter 10 | Dummy Variable Models | Shalabh, IIT Kanpur
If $D_2 = 1$, then

$$y = \beta_0 + \beta_1 x_1 + \beta_2 + \varepsilon$$

$$= (\beta_0 + \beta_2) + \beta_1 x_1 + \varepsilon$$

$$E(y|D_2 = 1) = (\beta_0 + \beta_2) + \beta_1 x_1$$

which is a straight line relationship with intercept $(\beta_0 + \beta_2)$ and slope $\beta_1$.

The quantities $E(y|D_2 = 0)$ and $E(y|D_2 = 1)$ are the average responses when an observation belongs to group $A$ and group $B$, respectively. Thus

$$\beta_2 = E(y|D_2 = 1) - E(y|D_2 = 0)$$

which has an interpretation as the difference between the average values of $y$ with $D_2 = 0$ and $D_2 = 1$.

Graphically, it looks like as in the following figure. It describes two parallel regression lines with same variances $\sigma^2$.

If there are three explanatory variables in the model with two dummy variables $D_2$ and $D_3$ then they will describe three levels, e.g., groups $A$, $B$ and $C$. The levels of dummy variables are as follows:

1. $D_2 = 0$, $D_3 = 0$ if the observation is from group $A$
2. $D_2 = 1$, $D_3 = 0$ if the observation is from group $B$
3. $D_2 = 0$, $D_3 = 1$ if the observation is from group $C$

The concerned regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 D_2 + \beta_3 D_3 + \varepsilon, \ E(\varepsilon) = 0, \ \text{var}(\varepsilon) = \sigma^2.$$
In general, if a qualitative variable has $m$ levels, then $(m-1)$ dummy variables are required and each of them takes value 0 and 1.

Consider the following examples to understand how to define such dummy variables and how they can be handled.

**Example:**
Suppose $y$ denotes the monthly salary of a person and $D$ denotes whether the person is graduate or non-graduate. The model is

$$y = \beta_0 + \beta_1 D + \varepsilon, \ E(\varepsilon) = 0, \ \text{var}(\varepsilon) = \sigma^2.$$ 

With $n$ observations, the model is

$$y_i = \beta_0 + \beta_1 D_i + \varepsilon_i, \ i = 1, 2, \ldots, n$$

$$E(y_i | D_i = 0) = \beta_0$$

$$E(y_i | D_i = 1) = \beta_0 + \beta_1$$

$$\beta_1 = E(y_i | D_i = 1) - E(y_i | D_i = 0)$$

Thus

- $\beta_0$ measures the mean salary of a non-graduate.
- $\beta_1$ measures the difference in the mean salaries of a graduate and non-graduate person.

Now consider the same model with two dummy variables defined in the following way:

$$D_{1i} = \begin{cases} 
1 & \text{if person is graduate} \\
0 & \text{if person is non-graduate},
\end{cases}$$

$$D_{2i} = \begin{cases} 
1 & \text{if person is non-graduate} \\
0 & \text{if person is graduate}.
\end{cases}$$

The model with $n$ observations is

$$y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \varepsilon_i, \ E(\varepsilon_i) = 0, \ Var(\varepsilon_i) = \sigma^2, i = 1, 2, \ldots, n.$$ 

Then we have

1. $E\left[y_i | D_{1i} = 0, D_{2i} = 1 \right] = \beta_0 + \beta_2$ : Average salary of non-graduate

2. $E\left[y_i | D_{1i} = 1, D_{2i} = 0 \right] = \beta_0 + \beta_1$ : Average salary of graduate

3. $E\left[y_i | D_{1i} = 0, D_{2i} = 0 \right] = \beta_0$ : cannot exist

4. $E\left[y_i | D_{1i} = 1, D_{2i} = 1 \right] = \beta_0 + \beta_1 + \beta_2$ : cannot exist.
Notice that in this case
\[ D_{i1} + D_{i2} = 1 \text{ for all } i \]
which is an exact constraint and indicates the contradiction as follows:
\[ D_{i1} + D_{i2} = 1 \Rightarrow \text{person is graduate} \]
\[ D_{i1} + D_{i2} = 1 \Rightarrow \text{person is non-graduate} \]

So multicollinearity is present in such cases. Hence the rank of matrix of explanatory variables falls short by 1. So \( \beta_0, \beta_1 \) and \( \beta_2 \) are indeterminate and least squares method breaks down. So the proposition of introducing two dummy variables is useful but they lead to serious consequences. This is known as **dummy variable trap**.

If the intercept term is ignored, then the model becomes
\[
y_i = \beta_1 D_{i1} + \beta_2 D_{i2} + \varepsilon_i, \quad E(\varepsilon_i) = 0, \ Var(\varepsilon_i) = \sigma^2, \ i = 1, 2, ..., n
\]
then
\[
E(y_i \mid D_{i1} = 1, D_{i2} = 0) = \beta_1 \Rightarrow \text{Average salary of a graduate.}
\]
\[
E(y_i \mid D_{i1} = 0, D_{i2} = 1) = \beta_2 \Rightarrow \text{Average salary of a non-graduate.}
\]

So when intercept term is dropped, then \( \beta_1 \) and \( \beta_2 \) have proper interpretations as the average salaries of a graduate and non-graduate persons, respectively.

Now the parameters can be estimated using ordinary least squares principle and standard procedures for drawing inferences can be used.

**Rule:** When the explanatory variable leads to \( m \) mutually exclusive categories classification, then use \((m - 1)\) dummy variables for its representation. Alternatively, use \( m \) dummy variables but drop the intercept term.
Interaction term:

Suppose a model has two explanatory variables – one quantitative variable and another a dummy variable. Suppose both interact and an explanatory variable as the interaction of them is added to the model.

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 D_{i2} + \beta_3 x_{i1} D_{i2} + \epsilon_i, \quad E(\epsilon_i) = 0, Var(\epsilon_i) = \sigma^2, i = 1, 2, \ldots, n. \]

To interpret the model parameters, we proceed as follows:

Suppose the dummy variables are given by

\[ D_{i2} = \begin{cases} 1 & \text{if } i^{th} \text{ person belongs to group } A \\ 0 & \text{if } i^{th} \text{ person belongs to group } B \end{cases} \]

Then

\[ y_i = \text{salary of } i^{th} \text{ person.} \]

Then

\[
E(y_i | D_{i2} = 0) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1} = \beta_0 + \beta_1 x_{i1}.
\]

This is a straight line with intercept \( \beta_0 \) and slope \( \beta_1 \). Next

\[
E(y_i | D_{i2} = 1) = \beta_0 + \beta_1 x_{i1} + \beta_2 + \beta_3 x_{i1} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{i1}.
\]

This is a straight line with intercept term \( \beta_0 + \beta_2 \) and slope \( \beta_1 + \beta_3 \).

The model

\[ E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 D_{i2} + \beta_3 x_{i1} D_{i2} \]

has different slopes and different intercept terms.

Thus

\( \beta_2 \) reflects the change in intercept term associated with the change in the group of person i.e., when group changes from \( A \) to \( B \).

\( \beta_3 \) reflects the change in slope associated with the change in the group of person, i.e., when group changes from \( A \) to \( B \).

Fitting of the model

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 D_{i2} + \beta_3 x_{i1} D_{i2} + \epsilon_i \]

is equivalent to fitting two separate regression models corresponding to \( D_{i2} = 1 \) and \( D_{i2} = 0 \), i.e.
\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i \]
\[ y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{i1} D_{i2} + \epsilon_i \]

and
\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2, 0 + \beta_3 x_{i3}, 0 + \epsilon_i \]
\[ y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i \]

respectively.

The test of hypothesis becomes convenient by using a dummy variable. For example, if we want to test whether the two regression models are identical, the test of hypothesis involves testing

\[ H_0 : \beta_2 = \beta_3 = 0 \]
\[ H_1 : \beta_2 \neq 0 \text{ and/or } \beta_3 \neq 0. \]

Acceptance of \( H_0 \) indicates that only a single model is necessary to explain the relationship.

In another example, if the objective is to test that the two models differ with respect to intercepts only and they have same slopes, then the test of hypothesis involves testing

\[ H_0 : \beta_3 = 0 \]
\[ H_1 : \beta_3 \neq 0. \]

**Dummy variables versus quantitative explanatory variable**

The quantitative explanatory variables can be converted into dummy variables. For example, if the ages of persons are grouped as follows:

Group 1: 1 day to 3 years
Group 2: 3 years to 8 years
Group 3: 8 years to 12 years
Group 4: 12 years to 17 years
Group 5: 17 years to 25 years

then the variable “age” can be represented by four different dummy variables.

Since it is difficult to collect the data on individual ages, so this will help in easy collection of data. A disadvantage is that some loss of information occurs. For example, if the ages in years are 2, 3, 4, 5, 6, 7 and suppose the dummy variable is defined as
Then these values become 0, 0, 0, 1, 1, 1. Now looking at the value 1, one can not determine if it corresponds to age 5, 6 or 7 years.

Moreover, if a quantitative explanatory variable is grouped into \( m \) categories, then \((m - 1)\) parameters are required whereas if the original variable is used as such, then only one parameter is required.

Treating a quantitative variable as qualitative variable increases the complexity of the model. The degrees of freedom for error are also reduced. This can effect the inferences if data set is small. In large data sets, such effect may be small.

The use of dummy variables does not require any assumption about the functional form of the relationship between study and explanatory variables.
Chapter 11
Specification Error Analysis

The specification of a linear regression model consists of a formulation of the regression relationships and of statements or assumptions concerning the explanatory variables and disturbances. If any of these is violated, e.g., incorrect functional form, incorrect introduction of disturbance term in the model etc., then specification error occurs. In narrower sense, the specification error refers to explanatory variables.

The complete regression analysis depends on the explanatory variables present in the model. It is understood in the regression analysis that only correct and important explanatory variables appears in the model. In practice, after ensuring the correct functional form of the model, the analyst usually has a pool of explanatory variables which possibly influence the process or experiment. Generally, all such candidate variables are not used in the regression modeling but a subset of explanatory variables is chosen from this pool.

While choosing a subset of explanatory variables, there are two possible options:

1. In order to make the model as realistic as possible, the analyst may include as many as possible explanatory variables.
2. In order to make the model as simple as possible, one may include only fewer number of explanatory variables.

In such selections, there can be two types of incorrect model specifications.

1. Omission/exclusion of relevant variables.
2. Inclusion of irrelevant variables.

Now we discuss the statistical consequences arising from the both situations.

1. Exclusion of relevant variables:

In order to keep the model simple, the analyst may delete some of the explanatory variables which may be of importance from the point of view of theoretical considerations. There can be several reasons behind such decisions, e.g., it may be hard to quantify the variables like taste, intelligence etc. Sometimes it may be difficult to take correct observations on the variables like income etc.
Let there be \( k \) candidate explanatory variables out of which suppose \( r \) variables are included and \( (k-r) \) variables are to be deleted from the model. So partition the \( X \) and \( \beta \) as

\[
X = \begin{pmatrix} X_1 & X_2 \\ \end{pmatrix}_{nk} \quad \text{and} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}_{r(k-r)}.
\]

The model \( y = X\beta + \epsilon \), \( E(\epsilon) = 0 \), \( V(\epsilon) = \sigma^2 I \) can be expressed as

\[
y = X_1\beta_1 + X_2\beta_2 + \epsilon
\]

which is called as **full model** or **true model**.

After dropping the \( r \) explanatory variable in the model, the new model is

\[
y = X_1\beta_1 + \delta
\]

which is called as **misspecified model** or **false model**.

Applying OLS to the false model, the OLSE of \( \beta_1 \) is

\[
b_{1F} = (X_1'X_1)^{-1}X_1'y.
\]

The estimation error is obtained as follows:

\[
b_{1F} = (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \epsilon)
\]

\[
= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\epsilon
\]

\[
b_{1F} - \beta_1 = \theta + (X_1'X_1)^{-1}X_1'\epsilon
\]

where \( \theta = (X_1'X_1)^{-1}X_1'X_2\beta_2 \).

Thus

\[
E(b_{1F} - \beta_1) = \theta + (X_1'X_1)^{-1}E(\epsilon)
\]

\[
= \theta
\]

which is a linear function of \( \beta_2 \), i.e., the coefficients of excluded variables. So \( b_{1F} \) is biased, in general. The bias vanishes if \( X_1'X_2 = 0 \), i.e., \( X_1 \) and \( X_2 \) are orthogonal or uncorrelated.

The mean squared error matrix of \( b_{1F} \) is

\[
MSE(b_{1F}) = E(b_{1F} - \beta_1)(b_{1F} - \beta_1)'
\]

\[
= E \left[ \theta\theta' + \theta\epsilon'X_1(X_1'X_1)^{-1} + (X_1'X_1)^{-1}X_1'\epsilon\theta' + (X_1'X_1)^{-1}X_1'\epsilon\epsilon'X_1(X_1'X_1)^{-1} \right]
\]

\[
= \theta\theta' + 0 + 0 + \sigma^2(X_1'X_1)^{-1}X_1'IX_1(X_1'X_1)^{-1}
\]

\[
= \theta\theta' + \sigma^2(X_1'X_1)^{-1}.
\]
So efficiency generally declines. Note that the second term is the conventional form of MSE.

The residual sum of squares is

\[ \hat{\sigma}^2 = \frac{SS_{res}}{n-r} = \frac{e'e}{n-r} \]

where \( e = y - X_bF = \overline{H}_1y \),
\( \overline{H}_1 = I - X_1(X'_1X_1)^{-1}X'_1 \).

Thus

\[ \overline{H}_1y = \overline{H}_1(X_1\beta_1 + X_2\beta_2 + \epsilon) \]
\[ = 0 + \overline{H}_1(X_2\beta_2 + \epsilon) \]
\[ = \overline{H}_1(X_2\beta_2 + \epsilon). \]

\[ y'\overline{H}_1y = (X_1\beta_1 + X_2\beta_2 + \epsilon)\overline{H}_1(X_2\beta_2 + \epsilon) \]
\[ = (\beta_1X'_1\overline{H}_1X_1\beta_1 + \beta_1X'_1\overline{H}_1\epsilon + \beta_1X'_1\overline{H}_1X_2\beta_2 + \beta_1X'_1\overline{H}_1\epsilon + \epsilon'\overline{H}_1X_2\beta_2 + \epsilon'\overline{H}_1\epsilon). \]

\[ E(s^2) = \frac{1}{n-r}\left[ E(\beta_1X'_1\overline{H}_1X_1\beta_1) + 0 + E(\epsilon'\overline{H}_1\epsilon) \right] \]
\[ = \frac{1}{n-r}\left[ \beta_1X'_1\overline{H}_1X_1\beta_1 + (n-r)s^2 \right] \]
\[ = \sigma^2 + \frac{1}{n-r}\beta_1X'_1\overline{H}_1X_1\beta_1. \]

Thus \( s^2 \) is a biased estimator of \( \sigma^2 \) and \( s^2 \) provides an over estimate of \( \sigma^2 \). Note that even if \( X'_1X_2 = 0 \), then also \( s^2 \) gives an overestimate of \( \sigma^2 \). So the statistical inferences based on this will be faulty. The \( t \)-test and confidence region will be invalid in this case.

If the response is to be predicted at \( x' = (x'_1, x'_2) \), then using the full model, the predicted value is
\[ \hat{y} = x'b = x'(X'X)^{-1}X'y \]

with
\[ E(\hat{y}) = x'\beta \]
\[ Var(\hat{y}) = \sigma^2\left[1 + x'(X'X)^{-1}x\right]. \]

When subset model is used then the predictor is
\[ \hat{y}_1 = x'_1b_{1F} \]

and then
\[ E(\hat{y}_1) = \hat{x}_i (X'_iX_i)^{-1} X'_i E(y) \]
\[ = \hat{x}_i (X'_iX_i)^{-1} X'_i (X_1\beta_1 + X_2\beta_2 + \varepsilon) \]
\[ = \hat{x}_i (X'_iX_i)^{-1} X'_i(\hat{x}_i\beta_1 + \hat{x}_i\beta_2) \]
\[ = \hat{x}_i\beta_1 + \hat{x}_i (X'_iX_i)^{-1} X'_iX_2\beta_2 \]
\[ = \hat{x}_i\beta_1 + \hat{x}_i\theta. \]

Thus \( \hat{y}_1 \) is a biased predictor of \( y \). It is unbiased when \( X'_iX_2 = 0 \). The MSE of predictor is
\[ MSE(\hat{y}_1) = \sigma^2 \left[ 1 + \hat{x}_i (X'_iX_i)^{-1} x_i \right] + \left( \hat{x}_i\theta - \hat{x}_i\beta_2 \right)^2. \]

Also
\[ Var(\hat{y}) \geq MSE(\hat{y}_1) \]
provided \( V(\hat{\beta}_2) - \beta_2\beta_2 \) is positive semidefinite.

2. Inclusion of irrelevant variables

Sometimes due to enthusiasm and to make the model more realistic, the analyst may include some explanatory variables that are not very relevant to the model. Such variables may contribute very little to the explanatory power of the model. This may tend to reduce the degrees of freedom \((n-k)\) and consequently the validity of inference drawn may be questionable. For example, the value of coefficient of determination will increase indicating that the model is getting better which may not really be true.

Let the true model be
\[ y = X\beta + \varepsilon, \quad E(\varepsilon) = 0, V(\varepsilon) = \sigma^2 I \]
which comprise \( k \) explanatory variable. Suppose now \( r \) additional explanatory variables are added to the model and resulting model becomes
\[ y = X\beta + Z\gamma + \delta \]
where \( Z \) is a \( n \times r \) matrix of \( n \) observations on each of the \( r \) explanatory variables and \( \gamma \) is \( r \times 1 \) vector of regression coefficient associated with \( Z \) and \( \delta \) is disturbance term. This model is termed as false model.

Applying OLS to false model, we get
\[
\begin{pmatrix}
    b_F \\
    c_F
\end{pmatrix} = \begin{pmatrix}
    X'X & X'Z \\
    Z'X & Z'Z
\end{pmatrix}^{-1}
\begin{pmatrix}
    X'y \\
    Z'y
\end{pmatrix}
\]

\[
\begin{pmatrix}
    X'X & X'Z \\
    Z'X & Z'Z
\end{pmatrix} \begin{pmatrix}
    b_F \\
    c_F
\end{pmatrix} = \begin{pmatrix}
    X'y \\
    Z'y
\end{pmatrix}
\]

\[\Rightarrow X'Xb_F + X'Zc_F = X'y \quad (1)\]
\[Z'Xb_F + Z'Zc_F = Z'y \quad (2)\]

where \( b_F \) and \( c_F \) are the OLSEs of \( \beta \) and \( \gamma \) respectively.

Premultiply equation (2) by \( X'Z(Z'Z)^{-1} \), we get
\[X'Z(Z'Z)^{-1}Z'Xb_F + X'Z(Z'Z)^{-1}Z'Zc_F = X'Z(Z'Z)^{-1}Z'y. \quad (3)\]

Subtracting equation (1) from (3), we get
\[
\left[ X'X - X'Z(Z'Z)^{-1}Z'X \right] b_F = X'y - X'Z(Z'Z)^{-1}Z'y
\]
\[X'[I - Z(Z'Z)^{-1}Z']Xb_F = X'[I - Z(Z'Z)^{-1}Z']y
\]
\[\Rightarrow b_F = (X'H_zX)^{-1}X'H_zy\]

where \( H_z = I - Z(Z'Z)^{-1}Z' \).

The estimation error of \( b_F \) is
\[
b_F - \beta = (X'H_zX)^{-1}X'H_zy - \beta
\]
\[= (X'H_zX)^{-1}X'H_z(X\beta + \varepsilon) - \beta
\]
\[= (X'H_zX)^{-1}X'H_z\varepsilon.
\]

Thus
\[E(b_F - \beta) = (X'H_zX)^{-1}X'H_zE(\varepsilon) = 0\]

so \( b_F \) is unbiased even when some irrelevant variables are added to the model.

The covariance matrix is
\[
V(b_F) = E((b_F - \beta)(b_F - \beta)')
\]
\[= E\left[ (X'H_zX)^{-1}X'H_z\varepsilon\varepsilon'X'H_zX(X'H_zX)^{-1} \right]
\]
\[= \sigma^2 (X'H_zX)^{-1}X'H_zX(X'H_zX)^{-1}
\]
\[= \sigma^2 (X'H_zX)^{-1}.
\]
If OLS is applied to true model, then
\[ b_\tau = (X'X)^{-1}X'y \]
with \( E(b_\tau) = \beta \)
\[ V(b_\tau) = \sigma^2(X'X)^{-1}. \]

To compare \( b_F \) and \( b_\tau \), we use the following result.

Result: If \( A \) and \( B \) are two positive definite matrices then \( A - B \) is at least positive semi definite if \( B^{-1} - A^{-1} \) is also at least positive semi definite.

Let
\[
A = (X'\bar{H}_z X)^{-1} \\
B = (X'X)^{-1} \\
B^{-1} - A^{-1} = X'X - X'\bar{H}_z X \\
= X'X - X'X + X'Z(Z'Z)^{-1}Z'X \\
= X'Z(Z'Z)^{-1}Z'X
\]
which is at least positive semi definite matrix. This implies that the efficiency declines unless \( X'Z = 0 \). If \( X'Z = 0 \), i.e., \( X \) and \( Z \) are orthogonal, then both are equally efficient.

The residual sum of squares under false model is
\[ SS_{res} = e_F'e_F \]
where
\[
e_F = y - Xb_F - Zc_F \\
b_F = (X'\bar{H}_z X)^{-1}X'\bar{H}_z y \\
c_F = (Z'Z)^{-1}Z'y - (Z'Z)^{-1}Z'Xb_F \\
= (Z'Z)^{-1}Z'(y - Xb_F) \\
= (Z'Z)^{-1}Z'[I - X(X'\bar{H}_z X)^{-1}X'\bar{H}_z]y \\
= (Z'Z)^{-1}Z'\bar{H}_{xz}y \\
\bar{H}_z = I - Z(Z'Z)^{-1}Z' \\
\bar{H}_{xz} = I - X(X'\bar{H}_z X)^{-1}X'\bar{H}_z \\
\bar{H}_{xz}^2 = \bar{H}_{xz} : \text{idempotent.}
\]
So

\[ e_F = y - X(X' \bar{H}_Z X)^{-1} X' \bar{H}_Z y - Z(Z' Z)^{-1} Z' \bar{H}_{XZ} y \]
\[ = [I - X(X' \bar{H}_Z X)^{-1} X' \bar{H}_Z - Z(Z' Z)^{-1} Z' \bar{H}_{XZ}] y \]
\[ = \bar{H}_{XZ} - (I - \bar{H}_Z) \bar{H}_{XZ} \]
\[ = \bar{H}_Z \bar{H}_{XZ} y \]
\[ = \bar{H}^*_{XZ} y \text{ where } \bar{H}^*_{XZ} = \bar{H}_Z \bar{H}_{XZ}. \]

Thus

\[ SS_{res} = e_F' e_F \]
\[ = y' \bar{H}_Z \bar{H}_{XZ} \bar{H}_{XZ} \bar{H}_Z y \]
\[ = y' \bar{H}_Z \bar{H}_{XZ} y \]
\[ = y' \bar{H}^*_{XZ} y \]
\[ E(SS_{res}) = \sigma^2 tr(\bar{H}^*_{XZ}) \]
\[ = \sigma^2 (n - k - r) \]
\[ E \left( \frac{SS_{res}}{n-k-r} \right) = \sigma^2. \]

So \[ \frac{SS_{res}}{n-k-r} \] is an unbiased estimator of \( \sigma^2 \).

A comparison of exclusion and inclusion of variables is as follows:

<table>
<thead>
<tr>
<th></th>
<th>Exclusion type</th>
<th>Inclusion type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimation of coefficients</td>
<td>Biased</td>
<td>Unbiased</td>
</tr>
<tr>
<td>Efficiency</td>
<td>Generally declines</td>
<td>Declines</td>
</tr>
<tr>
<td>Estimation of disturbance term</td>
<td>Over-estimate</td>
<td>Unbiased</td>
</tr>
<tr>
<td>Conventional test of hypothesis and confidence region</td>
<td>Invalid and faulty inferences</td>
<td>Valid though erroneous</td>
</tr>
</tbody>
</table>
Chapter 12

Tests for Structural Change and Stability

A fundamental assumption in a regression modeling is that the pattern of data on dependent and independent variables remains the same throughout the period over which the data is collected. Under such an assumption, a single linear regression model is fitted over the entire data set. The regression model is estimated and used for prediction assuming that the parameters remain same over the entire time period of estimation and prediction. When it is suspected that there exist a change in the pattern of data, then the fitting of single linear regression model may not be appropriate and more than one regression models may be required to be fitted. Before taking such a decision to fit a single or more than one regression models, a question arises how to test and decide if there is a change in the structure or pattern of data. Such changes can be characterized by the change in the parameters of the model and are termed as structural change.

Now we consider some examples to understand the problem of structural change in the data. Suppose the data on the consumption pattern is available for several years and suppose there was a war in between the years over which the consumption data is available. Obviously, the consumption pattern before and after war does not remains the same as the economy of the country gets disturbed. So if a model

$$y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} + \varepsilon_i, \ i = 1, 2, \ldots, n$$

is fitted then the regression coefficients before and after the war period will change. Such a change is referred to as structural break or structural change in the data. A better option in this case would be to fit two different linear regression models- one for the data before war and another for the data after war.

In an another example, suppose the study variable is the salary of a person and explanatory variable is the number of years of schooling. Suppose the objective is to find if there is any discrimination is the salaries of males and females. To know this, two different regression models can be fitted-one for male employees and another for females employees. By calculating and comparing the regression coefficients of both the models, one can check the presence of sex discrimination in the salaries of male and female employees.

Consider another example of structural change. Suppose an experiment is conducted to study certain objectives and data is collected in USA and in India. Then a question arises whether the data sets from both the countries can be pooled together or not. The data sets can be pooled if they originate from the same model in the sense that there is no structural change present in the data. In such case, the presence of
structural change in the data can be tested and if there is no change, then both the data sets can be merged and single regression model can be fitted. If structural change is present, then two models are needed to be fitted.

The objective is now how to test for the presence of structural change in the data and stability of regression coefficients. In other words, we want to test the hypothesis that some of or all the regression coefficients differ in different subsets of data.

**Analysis**

We consider here a situation where only one structural change is present in the data. The data in this case be divided into two parts. Suppose we have a data set of \( n \) observations which is divided into two parts consisting of \( n_1 \) and \( n_2 \) observations such that

\[
n_1 + n_2 = n.
\]

Consider the model

\[
y = \alpha \ell + X \beta + \varepsilon
\]

where \( \ell \) is a \((n \times 1)\) vector with all elements unity, \( \alpha \) is a scalar denoting the intercept term, \( X \) is a \((n \times k)\) matrix of observations on \( k \) explanatory variables, \( \beta \) is a \((k \times 1)\) vector of regression coefficients and \( \varepsilon \) is a \((n \times 1)\) vector of disturbances.

Now partition \( \ell \), \( X \) and \( \varepsilon \) into two subgroups based on \( n_1 \) and \( n_2 \) observation as

\[
\ell = \begin{pmatrix} \ell_1 \\ \ell_2 \end{pmatrix}, \quad X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}
\]

where the orders of \( \ell_1 \) is \((n_1 \times 1)\), \( \ell_2 \) is \((n_2 \times 1)\), \( X_1 \) is \((n_1 \times k)\), \( X_2 \) is \((n_2 \times k)\), \( \varepsilon_1 \) is \((n_1 \times 1)\) and \( \varepsilon_2 \) is \((n_2 \times 1)\).

Based on this partitions, the two models corresponding to two subgroups are

\[
y_1 = \alpha \ell_1 + X_1 \beta + \varepsilon_1
\]

\[
y_2 = \alpha \ell_2 + X_2 \beta + \varepsilon_2.
\]
In matrix notations, we can write

\[
\begin{pmatrix}
  y_1 \\
  y_2
\end{pmatrix} = \begin{pmatrix}
  \ell_1 & X_1 \\
  \ell_2 & X_2
\end{pmatrix} \begin{pmatrix}
  \alpha \\
  \beta
\end{pmatrix} + \begin{pmatrix}
  \epsilon_1 \\
  \epsilon_2
\end{pmatrix}
\]

and term it as Model (1).

In this case, the intercept terms and regression coefficients remain same for both the submodels. So there is no structural change in this situation.

The problem of structural change can be characterized if intercept terms and/or regression coefficients in the submodels are different.

If the structural change is caused due to change in the intercept terms only then the situation is characterized by the following model:

\[
y_1 = \alpha_1 \ell_1 + X_1 \beta + \epsilon_1 \\
y_2 = \alpha_2 \ell_2 + X_2 \beta + \epsilon_2
\]

or

\[
\begin{pmatrix}
  y_1 \\
  y_2
\end{pmatrix} = \begin{pmatrix}
  \ell_1 & 0 & X_1 \\
  0 & \ell_2 & X_2
\end{pmatrix} \begin{pmatrix}
  \alpha_1 \\
  \alpha_2 \\
  \beta
\end{pmatrix} + \begin{pmatrix}
  \epsilon_1 \\
  \epsilon_2
\end{pmatrix}.
\]

If the structural change is due to different intercept terms as well as different regression coefficients, then the model is

\[
y_1 = \alpha_1 \ell_1 + X_1 \beta_1 + \epsilon_1 \\
y_2 = \alpha_2 \ell_2 + X_2 \beta_2 + \epsilon_2
\]

or

\[
\begin{pmatrix}
  y_1 \\
  y_2
\end{pmatrix} = \begin{pmatrix}
  \ell_1 & 0 & X_1 \\
  0 & \ell_2 & X_2
\end{pmatrix} \begin{pmatrix}
  \alpha_1 \\
  \alpha_2 \\
  \beta_1
\end{pmatrix} + \begin{pmatrix}
  \epsilon_1 \\
  \epsilon_2
\end{pmatrix}.
\]

The test of hypothesis related to the test of structural change is conducted by testing anyone of the null hypothesis depending upon the situation.
(I) \( H_0 : \alpha_1 = \alpha_2 \)

(II) \( H_0 : \beta_1 = \beta_2 \)

(III) \( H_0 : \alpha_1 = \alpha_2, \beta_1 = \beta_2. \)

To construct the test statistic, apply ordinary least squares estimation to models (1), (2) and (3) and obtain the residual sum of squares as \( RSS_1, RSS_2 \) and \( RSS_3 \) respectively.

Note that the degrees of freedom associated with

- \( RSS_1 \) from model (1) is \( n - (k + 1) \).
- \( RSS_2 \) from model (2) is \( n - (k + 1 + 1) = n - (k + 2) \).
- \( RSS_3 \) from model (3) is \( n - (k + 1 + k + 1) = n - 2(k + 1) \).

The null hypothesis \( H_0 : \alpha_1 = \alpha_2 \) i.e., different intercept terms is tested by the statistic

\[
F = \frac{(RSS_1 - RSS_2) / 1}{RSS_2 / (n - k - 2)}
\]

which follows \( F(1, n-k-2) \) under \( H_0 \). This statistic tests \( \alpha_1 = \alpha_2 \) for model (2) using model (1), i.e., model (1) contrasted with model (2).

The null hypothesis \( H_0 : \beta_1 = \beta_2 \), i.e., different regression coefficients is tested by

\[
F = \frac{(RSS_2 - RSS_3) / k}{RSS_3 / (n - 2k - 2)}
\]

which follows \( F(k, n - 2k - 2) \) under \( H_0 \). This statistic tests \( \beta_1 = \beta_2 \) from model (3) using model (2), i.e., model (2) contrasted with model (3).

The test of null hypothesis \( H_0 : \alpha_1 = \alpha_2, \beta_1 = \beta_2 \), i.e., different intercepts and different slope parameters can be jointly tested by the test statistic

\[
F = \frac{(RSS_1 - RSS_3) / (k + 1)}{RSS_3 / (n - 2k - 2)}
\]

which follows \( F(k + 1, n - 2k - 2) \) under \( H_0 \). This statistic tests jointly \( \alpha_1 = \alpha_2 \) and \( \beta_1 = \beta_2 \) for model (3) using model (1), i.e., model (1) contrasted with model (3). This test is known as **Chow test**. It requires
\( n_1 > k \) and \( n_2 > k \) for the stability of regression coefficients in the two models. The development of this test is as follows which is based on the set up of analysis of variance test.

**Development of Chow test:**

Consider the models

\[
\begin{align*}
y_1 &= X_1 \beta_1 + \varepsilon_1 \\
y_2 &= X_2 \beta_2 + \varepsilon_2 \\
y &= X \beta + \varepsilon
\end{align*}
\]

where \( p = k + 1 \) which includes \( k \) explanatory variables and an intercept term.

Define

\[
\begin{align*}
\bar{H}_1 &= I_1 - X_1 (X_1'X_1)^{-1} X_1' \\
\bar{H}_2 &= I_2 - X_2 (X_2'X_2)^{-1} X_2' \\
\bar{H} &= I - X (X'X)^{-1} X'
\end{align*}
\]

where \( I_1 \) and \( I_2 \) are the identity matrices of orders \( (n_1 \times n_1) \) and \( (n_2 \times n_2) \). The residual sums of squares based on models (i), (ii) and (iii) are obtained as

\[
\begin{align*}
\text{RSS}_1 &= \varepsilon_1' \bar{H}_1 \varepsilon_1 \\
\text{RSS}_2 &= \varepsilon_2' \bar{H}_2 \varepsilon_2 \\
\text{RSS} &= \varepsilon' \bar{H} \varepsilon.
\end{align*}
\]

Then define

\[
\begin{align*}
\bar{H}_1^* &= \begin{bmatrix} \bar{H}_1 & 0 \\ 0 & 0 \end{bmatrix}, \\
\bar{H}_2^* &= \begin{bmatrix} 0 & 0 \\ 0 & \bar{H}_2 \end{bmatrix}.
\end{align*}
\]

Both \( \bar{H}_1^* \) and \( \bar{H}_2^* \) are \((n \times n)\) matrices. Now the \( \text{RSS}_1 \) and \( \text{RSS}_2 \) can be re-expressed as

\[
\begin{align*}
\text{RSS}_1 &= \varepsilon_1' \bar{H}_1^* \varepsilon_1 \\
\text{RSS}_2 &= \varepsilon_2' \bar{H}_2^* \varepsilon_2
\end{align*}
\]

where \( \varepsilon \) is the disturbance term related to model (iii) based on \( (n_1 + n_2) \) observations.

Note that \( \bar{H}_1^* \bar{H}_2^* = 0 \) which implies that \( \text{RSS}_1 \) and \( \text{RSS}_2 \) are independently distributed.
We can write
\[
\bar{H} = I - X (X'X)^{-1} X' \\
= \begin{pmatrix}
I_1 & 0 \\
0 & I_2
\end{pmatrix} - \begin{pmatrix}
X_1 \\
X_2
\end{pmatrix} (X'X)^{-1} \begin{pmatrix}
X_1' \\
X_2'
\end{pmatrix} \\
= \begin{pmatrix}
\bar{H}_{11} & \bar{H}_{12} \\
\bar{H}_{21} & \bar{H}_{22}
\end{pmatrix}
\]
where
\[
\bar{H}_{11} = I_1 - X_1 (X'X)^{-1} X_1' \\
\bar{H}_{12} = -X_1 (X'X)^{-1} X_2' \\
\bar{H}_{21} = -X_2 (X'X)^{-1} X_1' \\
\bar{H}_{22} = I_2 - X_2 (X'X)^{-1} X_2'.
\]

Define
\[
\bar{H}^* = \bar{H}_1^* + \bar{H}_2^*
\]
so that
\[
\text{RSS}_3 = \varepsilon' \bar{H}^* \varepsilon \\
\text{RSS}_1 = \varepsilon' \bar{H} \varepsilon.
\]

Note that \((\bar{H} - \bar{H}^*)\) and \(\bar{H}^*\) are idempotent matrices. Also \((\bar{H} - \bar{H}^*) \bar{H}^* = 0\). First we see how this result holds.

Consider
\[
(\bar{H} - \bar{H}_1^*) \bar{H}_1^* = \begin{pmatrix}
\bar{H}_{11} - \bar{H}_1 \\
\bar{H}_{21}
\end{pmatrix} \begin{pmatrix}
\bar{H}_1 & 0 \\
0 & 0
\end{pmatrix} \\
= \begin{pmatrix}
(\bar{H}_{11} - \bar{H}_1) \bar{H}_1 & 0 \\
\bar{H}_{21} \bar{H}_1 & 0
\end{pmatrix}.
\]
Since \(X_1' \bar{H}_1 = 0\), so we have
\[
\bar{H}_{21} \bar{H}_1 = 0 \\
\bar{H}_{11} \bar{H}_1 = 0.
\]
Also, since $\bar{H}_1$ is idempotent, it follows that
\[
(\bar{H}_1 - \bar{H}_1) \bar{H}_1 = 0.
\]
Thus $(\bar{H} - \bar{H}_1^*) \bar{H}_1^* = 0$
or $\bar{H} \bar{H}_1^* = \bar{H}_1^*$.

Similarly, it can be shown that
\[
(\bar{H} - \bar{H}_2^*) \bar{H}_2^* = 0
\]
or $\bar{H} \bar{H}_2^* = \bar{H}_2^*$.

Thus this implies that
\[
(\bar{H} - \bar{H}_1^* - \bar{H}_2^*)(\bar{H}_1^* + \bar{H}_2^*) = 0
\]
or $(\bar{H} - \bar{H}^*) \bar{H}^* = 0$.

Also, we have
\[
\text{tr } \bar{H} = n - p
\]
\[
\text{tr } \bar{H}^* = \text{tr } \bar{H}_1^* + \text{tr } \bar{H}_2^*
\]
\[
= n_1 - p + n_2 - p
\]
\[
= n - 2p
\]
\[
= n - 2k - 2.
\]

Hence $RSS_1$ and $RSS_3$ are independently distributed. Further
\[
\frac{RSS_1 - RSS_3}{\sigma^2} \sim \chi^2_p,
\]
\[
\frac{RSS_3}{\sigma^2} \sim \chi^2_{(n-2p)}.
\]

Also, $\left( \frac{RSS_1 - RSS_3}{\sigma^2} \right)$ and $\frac{RSS_3}{\sigma^2}$ are independently distributed. Hence under the null hypothesis
\[
\frac{\left( \frac{RSS_1 - RSS_3}{\sigma^2} \right)}{p} \sim F(p, n-2p)
\]
\[
\frac{\left( \frac{RSS_3}{\sigma^2} \right)}{(n-2p)} \sim F(k+1, n-2k-2)
\]
or $\frac{RSS_3}{(n-2k-2)} \sim F(k+1, n-2k-2)$. 

Econometrics  |  Chapter 12  |  Tests for Structural Change and Stability  |  Shalabh, IIT Kanpur
Limitations of these tests

1. All tests are based under the assumption that $\sigma^2$ remains same. So first the stability of $\sigma^2$ should be checked and then these tests can be used.

2. It is assumed in these tests that the point of change is exactly known. In practice, it is difficult to find such a point at which the change occurs. It is more difficult to know such point when the change occurs slowly. These tests are not applicable when the point of change is unknown. An adhoc techniques when the point of change is unknown is to delete the data of transition period.

3. When there are more than one points of structural change, then the analysis becomes difficult.
Chapter 13
Asymptotic Theory and Stochastic Regressors

The nature of explanatory variable is assumed to be non-stochastic or fixed in repeated samples in any regression analysis. Such an assumption is appropriate for those experiments which are conducted inside the laboratories where the experimenter can control the values of explanatory variables. Then the repeated observations on study variable can be obtained for fixed values of explanatory variables. In practice, such an assumption may not always be satisfied. Sometimes, the explanatory variables in a given model are the study variable in another model. Thus the study variable depends on the explanatory variables that are stochastic in nature. Under such situations, the statistical inferences drawn from the linear regression model based on the assumption of fixed explanatory variables may not remain valid.

We assume now that the explanatory variables are stochastic but uncorrelated with the disturbance term. In case, they are correlated then the issue is addressed through instrumental variable estimation. Such a situation arises in the case of measurement error models.

Stochastic regressors model

Consider the linear regression model

\[ y = X \beta + \varepsilon \]

where \( X \) is a \((n \times k)\) matrix of \( n \) observations on \( k \) explanatory variables \( X_1, X_2, ..., X_k \) which are stochastic in nature, \( y \) is a \((n \times 1)\) vector of \( n \) observations on study variable, \( \beta \) is a \((k \times 1)\) vector of regression coefficients and \( \varepsilon \) is the \((n \times 1)\) vector of disturbances. Under the assumption \( E(\varepsilon) = 0, \sigma^2 I \), the distribution of \( \varepsilon_i \), conditional on \( x_i' \), satisfy these properties for all all values of \( X \) where \( x_i' \) denotes the \( i^{th} \) row of \( X \). This is demonstrated as follows:

Let \( p(\varepsilon_i | x_i') \) be the conditional probability density function of \( \varepsilon_i \) given \( x_i' \) and \( p(\varepsilon_i) \) is the unconditional probability density function of \( \varepsilon_i \). Then

\[
E(\varepsilon_i | x_i') = \int \varepsilon_i p(\varepsilon_i | x_i') d\varepsilon_i \\
= \int \varepsilon_i p(\varepsilon_i) d\varepsilon_i \\
= E(\varepsilon_i) \\
= 0
\]
\[ E(e_i^2 | x_i) = \int e_i^2 p(e_i | x_i) \, de_i \]
\[ = \int e_i^2 p(e_i) \, de_i \]
\[ = E(e_i^2) \]
\[ = \sigma^2. \]

In case, \( e_i \) and \( x_i \) are independent, then \( p(e_i | x_i) = p(e_i) \).

**Least squares estimation of parameters**

The additional assumption that the explanatory variables are stochastic poses no problem in the ordinary least squares estimation of \( \beta \) and \( \sigma^2 \). The OLSE of \( \beta \) is obtained by minimizing \((y - X\beta)'(y - X\beta)\) with respect \( \beta \) as

\[ b = (X'X)^{-1} X' y \]

and estimator of \( \sigma^2 \) is obtained as

\[ s^2 = \frac{1}{n-k}(y - Xb)'(y - Xb). \]

**Maximum likelihood estimation of parameters:**

Assuming \( e \sim N(0, \sigma^2 I) \) in the model \( y = X\beta + e \) along with \( X \) is stochastic and independent of \( e \), the joint probability density function \( e \) and \( X \) can be derived from the joint probability density function of \( y \) and \( X \) as follows:

\[ f(e, X) = f(e_1, e_2, \ldots, e_n, x_1, x_2, \ldots, x_n) \]
\[ = \left( \prod_{i=1}^n f(e_i) \right) \left( \prod_{i=1}^n f(x_i) \right) \]
\[ = \left( \prod_{i=1}^n f(y_i | x_i) \right) \left( \prod_{i=1}^n f(x_i) \right) \]
\[ = \prod_{i=1}^n \left( f(y_i | x_i) f(x_i) \right) \]
\[ = \prod_{i=1}^n \left( f(y_i, x_i) \right) \]
\[ = f(y_1, y_2, \ldots, y_n, x_1, x_2, \ldots, x_n) \]
\[ = f(y, X). \]
This implies that the maximum likelihood estimators of $\beta$ and $\sigma^2$ will be based on

$$
\prod_{i=1}^{n} f \left( y_i \mid x_i' \right) = \prod_{i=1}^{n} f \left( \varepsilon_i \right)
$$

so they will be same as based on the assumption that $\varepsilon_i'$s, $i=1,2,\ldots,n$ are distributed as $N(0,\sigma^2)$. So the maximum likelihood estimators of $\beta$ and $\sigma^2$ when the explanatory variables are stochastic are obtained as

$$
\hat{\beta} = (X'X)^{-1} X' y
$$

$$
\hat{\sigma}^2 = \frac{1}{n} (y - X\hat{\beta})' (y - X\hat{\beta}).
$$

**Alternative approach for deriving the maximum likelihood estimates**

Alternatively, the maximum likelihood estimators of $\beta$ and $\sigma^2$ can also be derived using the joint probability density function of $y$ and $X$.

**Note:** Note that the vector $x'$ is represented by an underscore in this section to denote that it’s order is $[1 \times (k-1)]$ which excludes the intercept term.

Let $x'_i$, $i=1,2,\ldots,n$ are from a multivariate normal distribution with mean vector $\mu_x$ and covariance matrix $\Sigma_{xx}$, i.e., $x'_i \sim N(\mu_x, \Sigma_{xx})$ and the joint distribution of $y$ and $x'$ is

$$
\begin{pmatrix} y \\ x' \end{pmatrix} \sim N \left( \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix} \right).
$$

Let the linear regression model is

$$
y_i = \beta_0 + x'_i \beta_i + \varepsilon_i, \quad i=1,2,\ldots,n
$$

where $x'_i$ is a $[1 \times (k-1)]$ vector of observation of random vector $x$, $\beta_0$ is the intercept term and $\beta_i$ is the $[(k-1) \times 1]$ vector of regression coefficients. Further $\varepsilon_i$ is disturbance term with $\varepsilon_i \sim N(0,\sigma^2)$ and is independent of $x'$. 

---

_Econometrics | Chapter 13 | Asymptotic Theory and Stochastic Regressors | Shalabh, IIT Kanpur_
Suppose
\[
\begin{pmatrix} y \\ x \end{pmatrix} \sim N \left( \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix} \right).
\]

The joint probability density function of \( (y, x') \) based on random sample of size \( n \) is
\[
f(y, x') = \frac{1}{(2\pi)^{\frac{k}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left[ -\frac{1}{2} \left( y - \mu_y \right)^\prime \Sigma^{-1} \left( y - \mu_y \right) \right].
\]

Now using the following result, we find \( \Sigma^{-1} \):

**Result:** Let \( A \) be a nonsingular matrix which in partitioned suitably as
\[
A = \begin{pmatrix} B & C \\ D & E \end{pmatrix},
\]
where \( E \) and \( F = B - CE^{-1}D \) are nonsingular matrices, then
\[
A^{-1} = \begin{pmatrix} F^{-1} & -F^{-1}CE^{-1} \\ -E^{-1}DF^{-1} & E^{-1} + E^{-1}DF^{-1}CE^{-1} \end{pmatrix}.
\]

Note that \( AA^{-1} = A^{-1}A = I \).

Thus
\[
\Sigma^{-1} = \frac{1}{\sigma^2} \begin{pmatrix} 1 & -\Sigma_{yx} \Sigma_{xx}^{-1} \\ -\Sigma_{sx} \Sigma_{sy}^{-1} & \sigma^2 \Sigma_{xx}^{-1} + \Sigma_{yx} \Sigma_{sy}^{-1} \Sigma_{xx} \Sigma_{sy}^{-1} \end{pmatrix},
\]
where
\[
\sigma^2 = \Sigma_{yy} - \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}.
\]

Then
\[
f(y, x') = \frac{1}{(2\pi)^{\frac{k-1}{2}} |\Sigma_{xx}|^{\frac{1}{2}}} \exp \left[ -\frac{1}{2} \left( x - \mu_x \right)^\prime \Sigma_{xx}^{-1} \left( x - \mu_x \right) \right].
\]

The marginal distribution of \( x' \) is obtained by integrating \( f(y, x') \) over \( y \) and the resulting distribution is \((k-1)\) variate multivariate normal distribution as
\[
g(x') = \frac{1}{(2\pi)^{\frac{k-1}{2}} |\Sigma_{xx}|^{\frac{1}{2}}} \exp \left[ -\frac{1}{2} \left( x - \mu_x \right)^\prime \Sigma_{xx}^{-1} \left( x - \mu_x \right) \right].
\]
The conditional probability density function of \( y \) given \( x' \) is

\[
f(y | x') = \frac{f(y, x')}{g(x')}
= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{1}{2\sigma^2} \left( (y - \mu_y) - (x - \mu_x)' \Sigma_{xy}^{-1} \Sigma_{xy}' \right)^2 \right]
\]

which is the probability density function of normal distribution with

- conditional mean
  \[ E(y | x') = \mu_y + (x - \mu_x)' \Sigma_{xy}^{-1} \Sigma_{xy} \]
- conditional variance
  \[ Var(y | x') = \sigma_y^2 \left( 1 - \rho^2 \right) \]

where

\[ \rho^2 = \frac{\Sigma_{y} \Sigma_{xy}^{-1} \Sigma_{xy}}{\sigma_y^2} \]

is the population multiple correlation coefficient.

In the model

\[ y = \beta_0 + x' \beta_i + \varepsilon, \]

the conditional mean is

\[
E(y_i | x_i) = \beta_0 + x'_i \beta_i + E(\varepsilon | x_i)
= \beta_0 + x'_i \beta_i.
\]

Comparing this conditional mean with the conditional mean of normal distribution, we obtain the relationship with \( \beta_0 \) and \( \beta_i \) as follows:

\[
\beta_i = \Sigma_{xx}^{-1} \Sigma_{xy},
\beta_0 = \mu_y - \mu_x' \beta_i.
\]

The likelihood function of \((y, x')\) based on a sample of size \( n \) is

\[
L = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^\frac{n}{2}} \exp \left[ -\frac{1}{2} \sum_{i=1}^{n} \left( y_i - \mu_y \right)' \Sigma_{xx}^{-1} \left( y_i - \mu_y \right) \right].
\]

Maximizing the log likelihood function with respect to \( \mu_y, \mu_x, \Sigma_{xx}, \) and \( \Sigma_{xy} \), the maximum likelihood estimates of respective parameters are obtained as
\[ \hat{\mu}_y = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \]
\[ \hat{\mu}_x = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = (\bar{x}_2, \bar{x}_3, \ldots, \bar{x}_k) \]
\[ \hat{\Sigma}_{xx} = S_{xx} = \frac{1}{n} \left( \sum_{i=1}^{n} x_i \bar{x}_i' - n \bar{x} \bar{x}' \right) \]
\[ \hat{\Sigma}_{xy} = S_{xy} = \frac{1}{n} \left( \sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y}' \right) \]

where \( \bar{x}_i = (x_{i2}, x_{i3}, \ldots, x_{ik}) \), \( S_{xx} \) is \([(k-1) \times (k-1)] \) matrix with elements \( \frac{1}{n} \sum_{i} (x_{ii} - \bar{x}_i)(x_{ij} - \bar{x}_j) \) and \( S_{xy} \) is \([(k-1) \times 1] \) vector with elements \( \frac{1}{n} \sum_{i} (x_{ii} - \bar{x}_i)(y_i - \bar{y}). \)

Based on these estimates, the maximum likelihood estimators of \( \beta_1 \) and \( \beta_0 \) are obtained as
\[ \hat{\beta}_1 = S_{x'y}S_{xy} \]
\[ \hat{\beta}_0 = \bar{y} - \bar{x}' \hat{\beta}_1 \]
\[ \hat{\beta} = \left( \begin{array}{c} \hat{\beta}_0 \\ \hat{\beta}_1 \end{array} \right) = (X'X)^{-1} X'y. \]

**Properties of the estimators of least squares estimator:**

The estimation error of OLSE \( b = (X'X)^{-1} X'y \) of \( \beta \) is
\[ b - \beta = (X'X)^{-1} X'y - \beta \]
\[ = (X'X)^{-1} X'(X\beta + \epsilon) - \beta \]
\[ = (X'X)^{-1} X'\epsilon. \]

Then assuming that \( E \left[ (X'X)^{-1} X' \right] \) exists, we have
\[ E(b - \beta) = E \left[ (X'X)^{-1} X'\epsilon \right] \]
\[ = E \left[ E \left[ (X'X)^{-1} X'\epsilon | X \right] \right] \]
\[ = E \left[ (X'X)^{-1} X' \right] E(\epsilon) \]
\[ = 0 \]

because \( (X'X)^{-1} X' \) and \( \epsilon \) are independent. So \( b \) is an unbiased estimator of \( \beta \).

---

*Econometrics  | Chapter 13  | Asymptotic Theory and Stochastic Regressors  | Shalabh, IIT Kanpur*
The covariance matrix of \( b \) is obtained as

\[
V(b) = E(b - \beta)(b - \beta)'
\]

\[
= E[(X'X)^{-1}X'e \varepsilon'e X(X'X)^{-1}]
\]

\[
= E[E[(X'X)^{-1}X'e \varepsilon'e X(X'X)^{-1}|X]]
\]

\[
= E[(X'X)^{-1}X'e E(\varepsilon' \varepsilon)X(X'X)^{-1}|X]
\]

\[
= E[(X'X)^{-1}X' \sigma^2 X(X'X)^{-1}]
\]

\[
= \sigma^2 E[(X'X)^{-1}].
\]

Thus the covariance matrix involves a mathematical expectation. The unknown \( \sigma^2 \) can be estimated by

\[
\hat{\sigma}^2 = \frac{e'e}{n-k}
\]

\[
= \frac{(y - Xb)'(y - Xb)}{n-k}
\]

where \( e = y - Xb \) is the residual and

\[
E(\hat{\sigma}^2) = E[E(\hat{\sigma}^2 | X)]
\]

\[
= E[E\left(\frac{e'e}{n-k}\right) | X]
\]

\[
= E(\sigma^2).
\]

Note that the OLSE \( b = (X'X)^{-1}X'y \) involves the stochastic matrix \( X \) and stochastic vector \( y \), so \( b \) is not a linear estimator. It is also no more the best linear unbiased estimator of \( \beta \) as in the case when \( X \) is nonstochastic. The estimator of \( \sigma^2 \) as being conditional on given \( X \) is an efficient estimator.

**Asymptotic theory:**

The asymptotic properties of an estimator concerns the properties of the estimator when sample size \( n \) grows large.

For the need and understanding of asymptotic theory, we consider an example. Consider the simple linear regression model with one explanatory variable and \( n \) observations as

\[
y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad E(\varepsilon_i) = 0, \quad Var(\varepsilon_i) = \sigma^2, \quad i = 1, 2, ..., n.
\]
The OLSE of $\beta_1$ is

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

and its variance is

$$Var(b_1) = \frac{\sigma^2}{n}.$$  

If the sample size grows large, then the variance of $b_1$ gets smaller. The shrinkage in variance implies that as sample size $n$ increases, the probability density of OLSE $b$ collapses around its mean because $Var(b)$ becomes zero.

Let there be three OLSEs $b_1, b_2$ and $b_3$ which are based on sample sizes $n_1, n_2$ and $n_3$ respectively such that $n_1 < n_2 < n_3$, say. If $c$ and $\delta$ are some arbitrarily chosen positive constants, then the probability that the value of $b$ lies within the interval $\beta \pm c$ can be made to be greater than $(1 - \delta)$ for a large value of $n$. This property is the consistency of $b$ which ensure that even if the sample is very large, then we can be confident with high probability that $b$ will yield an estimate that is close to $\beta$.

**Probability in limit**

Let $\hat{\beta}_n$ be an estimator of $\beta$ based on a sample of size $n$. Let $\gamma$ be any small positive constant. Then for large $n$, the requirement that $b_n$ takes values with probability almost one in an arbitrary small neighborhood of the true parameter value $\beta$ is

$$\lim_{n \to \infty} P\left[ |\hat{\beta}_n - \beta| < \gamma \right] = 1$$

which is denoted as

$$\text{plim} \ \hat{\beta}_n = \beta$$

and it is said that $\hat{\beta}_n$ converges to $\beta$ in probability. The estimator $\hat{\beta}_n$ is said to be a consistent estimator of $\beta$.

A sufficient but not necessary condition for $\hat{\beta}_n$ to be a consistent estimator of $\beta$ is that

$$\lim_{n \to \infty} E\left[ \hat{\beta}_n \right] = \beta$$

and

$$\lim_{n \to \infty} Var\left[ \hat{\beta}_n \right] = 0.$$
**Consistency of estimators**

Now we look at the consistency of the estimators of $\beta$ and $\sigma^2$.

**(i) Consistency of $b$**

Under the assumption that $\lim_{n \to \infty} \left( \frac{X'X}{n} \right) = \Delta$ exists as a nonstochastic and nonsingular matrix (with finite elements), we have

$$
\lim_{n \to \infty} V(b) = \sigma^2 \lim_{n \to \infty} \frac{1}{n} \left( \frac{X'X}{n} \right)^{-1} = \sigma^2 \lim_{n \to \infty} \frac{1}{n} \Delta^{-1} = 0.
$$

This implies that OLSE converges to $\beta$ in quadratic mean. Thus OLSE is a consistent estimator of $\beta$.

This also holds true for maximum likelihood estimators also.

Same conclusion can also be proved using the concept of convergence in probability.

The consistency of OLSE can be obtained under the weaker assumption that

$$
\text{plim} \left( \frac{X'X}{n} \right) = \Delta_.,
$$

exists and is a nonsingular and nonstochastic matrix and

$$
\text{plim} \left( \frac{X'e}{n} \right) = 0.
$$

Since

$$
b - \beta = (X'X)^{-1} X'e = \left( \frac{X'X}{n} \right)^{-1} \frac{X'e}{n}.
$$

So

$$
\text{plim}(b - \beta) = \text{plim} \left( \frac{X'X}{n} \right)^{-1} \text{plim} \left( \frac{X'e}{n} \right) = \Delta^{-1} \cdot 0 = 0.
$$

Thus $b$ is a consistent estimator of $\beta$. The same is true for maximum likelihood estimators also.
(ii) Consistency of $s^2$

Now we look at the consistency of $s^2$ as an estimate of $\sigma^2$. We have

\[
s^2 = \frac{1}{n-k}e'e = \frac{1}{n-k}e'\bar{H}e = \frac{1}{n}\left(1-\frac{k}{n}\right)^{-1}\left[e'e - e'X(X'X)^{-1}X'e\right] = \left(1-\frac{k}{n}\right)^{-1}\left[e'e - \frac{e'X}{n}(X'X)^{-1}X'e\right].
\]

Note that $\frac{e'\varepsilon}{n}$ consists of $\frac{1}{n}\sum_{i=1}^{n}e_i^2$ and $\{e_i^2, i=1,2,\ldots,n\}$ is a sequence of independently and identically distributed random variables with mean $\sigma^2$. Using the law of large numbers

\[
\text{plim}\left(\frac{e'\varepsilon}{n}\right) = \sigma^2.
\]

Then

\[
\text{plim}\left[\frac{e'X}{n}\left(\frac{X'X}{n}\right)^{-1}\frac{X'e}{n}\right] = \left(\text{plim} \frac{e'X}{n}\right)\left(\text{plim} \left(\frac{X'X}{n}\right)^{-1}\right)\left(\text{plim} \frac{X'e}{n}\right) = 0.\Delta^{-1}.0 = 0.
\]

\[
\Rightarrow \text{plim}(s^2) = (1-0)^{-1}\left[\sigma^2 - 0\right] = \sigma^2.
\]

Thus $s^2$ is a consistent estimator of $\sigma^2$. The same holds true for maximum likelihood estimates also.

**Asymptotic distributions:**

Suppose we have a sequence of random variables $\{\alpha_n\}$ with a corresponding sequence of cumulative density functions $\{F_n\}$ for a random variable $\alpha$ with cumulative density function $F$. Then $\alpha_n$ converges in distribution to $\alpha$ if $F_n$ converges to $F$ point wise. In this case, $F$ is called the asymptotic distribution of $\alpha_n$.

Note that since convergence in probability implies the convergence in distribution, so $\text{plim} \alpha_n = \alpha \Rightarrow \alpha_n \xrightarrow{D} \alpha$ ($\alpha_n$ tend to $\alpha$ in distribution), i.e., the asymptotic distribution of $\alpha_n$ is $F$ which is the distribution of $\alpha$.
Note that

\[ E(\alpha): \text{ Mean of asymptotic distribution} \]

\[ Var(\alpha): \text{ Variance of asymptotic distribution} \]

\[ \lim_{n \to \infty} E(\alpha_n): \text{ Asymptotic mean} \]

\[ \lim_{n \to \infty} E \left[ \alpha_n - \lim_{n \to \infty} E(\alpha_n) \right]^2: \text{ Asymptotic variance.} \]

**Asymptotic distribution of sample mean and least squares estimation**

Let \( \alpha_n = \bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i \) be the sample mean based on a sample of size \( n \). Since sample mean is a consistent estimator of population mean \( \bar{Y} \), so

\[ \text{plim} \bar{Y}_n = \bar{Y} \]

which is constant. Thus the asymptotic distribution of \( \bar{Y}_n \) is the distribution of a constant. This is not a regular distribution as all the probability mass is concentrated at one point. Thus as sample size increases, the distribution of \( \bar{Y}_n \) collapses.

Suppose consider only the one third observations in the sample and find sample mean as

\[ \bar{Y}_n^* = \frac{3}{n} \sum_{i=1}^{n} Y_i. \]

Then \( E(\bar{Y}_n^*) = \bar{Y} \)

and

\[ Var(\bar{Y}_n^*) = \frac{9}{n^2} \sum_{i=1}^{n} Var(Y_i) \]

\[ = \frac{9}{n^2} \frac{n}{3} \sigma^2 \]

\[ = \frac{3}{n} \sigma^2 \]

\[ \to 0 \text{ as } n \to \infty. \]
Thus \( \text{plim} \ Y_n^* = \bar{Y} \) and \( Y_n^* \) has the same degenerate distribution as \( \bar{Y}_n \). Since \( \text{Var}(Y_n^*) > \text{Var}(Y_n) \), so \( Y_n^* \) is preferred over \( Y_n \).

Now we observe the asymptotic behaviour of \( \bar{Y}_n \) and \( 
Y_n^* \). Consider a sequence of random variables \( \{\alpha_n\} \).

Thus for all \( n \), we have

\[
\begin{align*}
\alpha_n &= \sqrt{n} (\bar{Y}_n - \bar{Y}) \\
\alpha_n^* &= \sqrt{n} (\bar{Y}_n^* - \bar{Y}) \\
E(\alpha_n) &= \sqrt{n} E(\bar{Y}_n - \bar{Y}) = 0 \\
E(\alpha_n^*) &= \sqrt{n} E(\bar{Y}_n^* - \bar{Y}) = 0 \\
\text{Var}(\alpha_n) &= nE(\bar{Y}_n - \bar{Y})^2 = n \frac{\sigma^2}{n} = \sigma^2 \\
\text{Var}(\alpha_n^*) &= nE(\bar{Y}_n^* - \bar{Y})^2 = n \frac{3\sigma^2}{n} = 3\sigma^2.
\end{align*}
\]

Assuming the population to be normal, the asymptotic distribution of

- \( \bar{Y}_n \) is \( N(0, \sigma^2) \)
- \( \bar{Y}_n^* \) is \( N(0, 3\sigma^2) \).

So now \( \bar{Y}_n \) is preferable over \( \bar{Y}_n^* \). The central limit theorem can be used to show that \( \alpha_n \) will have an asymptotically normal distribution even if the population is not normally distributed.

Also, since

\[
\sqrt{n} (\bar{Y}_n - \bar{Y}) \sim N(0, \sigma^2) \\
\Rightarrow Z = \frac{\sqrt{n} (\bar{Y}_n - \bar{Y})}{\sigma} \sim N(0,1)
\]

and this statement holds true in finite sample as well as asymptotic distributions.

Consider the ordinary least squares estimate \( b = (X'X)^{-1}X'y \) of \( \beta \) in linear regression model \( y = X\beta + \varepsilon \). If \( X \) is nonstochastic then the finite covariance matrix of \( b \) is

\[
V(b) = \sigma^2 (X'X)^{-1}.
\]
The asymptotic covariance matrix of $b$ under the assumption that $\lim_{n \to \infty} \frac{XX}{n} = \Sigma_{xx}$ exists and is nonsingular.

It is given by

$$\sigma^2 \lim_{n \to \infty} \left( X'X \right) = \sigma^2 \lim_{n \to \infty} \left( \frac{1}{n} \right) \lim_{n \to \infty} \left( \frac{X'X}{n} \right)^{-1}$$

$$= \sigma^2 \cdot 0 \cdot \Sigma_{xx}^{-1}$$

$$= 0$$

which is a null matrix.

Consider the asymptotic distribution of $\sqrt{n} (b - \beta)$. Then even if $\varepsilon$ is not necessarily normally distributed, then asymptotically

$$\sqrt{n} (b - \beta) \sim N \left( 0, \sigma^2 \Sigma_{xx}^{-1} \right)$$

$$\frac{n (b - \beta)' \Sigma_{xx} (b - \beta)}{\sigma^2} \sim \chi^2_k.$$

If $\frac{XX}{n}$ is considered as an estimator of $\Sigma_{xx}$, then

$$\frac{n (b - \beta)' \frac{XX}{n} (b - \beta)}{\sigma^2} = \frac{(b - \beta)' X' X (b - \beta)}{\sigma^2}$$

is the usual test statistic as is in the case of finite samples with $b \sim N \left( \beta, \sigma^2 \left( X'X \right)^{-1} \right)$. 

---

Econometrics  |  Chapter 13  |  Asymptotic Theory and Stochastic Regressors  |  Shalabh, IIT Kanpur
Chapter 14
Stein-Rule Estimation

The ordinary least squares estimation of regression coefficients in linear regression model provides the estimators having minimum variance in the class of linear and unbiased estimators. The criterion of linearity is desirable because such estimators involve less mathematical complexity, they are easy to compute and it is easier to investigate their statistical properties. The criterion of unbiasedness is attractive because it is intuitively desirable to have an estimator whose expected value, i.e., the mean of the estimator should be same as the parameter being estimated. Considerations of linearity and unbiased estimators sometimes may lead to an unacceptably high price to be paid in terms of the variability around the true parameter. It is possible to have a nonlinear estimator with better properties. It is to be noted that one of the main objectives of estimation is to find an estimator whose values have high concentration around the true parameter. Sometimes it is possible to have a nonlinear and biased estimator that has smaller variability than the variability of best linear unbiased estimator of the parameter under some mild restrictions.

In the multiple regression model
\[
y = X \beta + \varepsilon, \quad E(\varepsilon) = 0, \quad V(\varepsilon) = \sigma^2 I,
\]
the ordinary least squares estimator (OLSE) of \( \beta \) is
\[
b = (X'X)^{-1}X'y
\]
which is the best linear unbiased estimator of \( \beta \) in the sense that it is linear in \( y, E(b) = \beta \) and \( b \) has smallest variance among all linear and unbiased estimators of \( \beta \). Its covariance matrix is
\[
V(b) = E((b - \beta)(b - \beta)') = \sigma^2 (X'X)^{-1}.
\]
The weighted mean squared error of any estimator \( \hat{\beta} \) is defined as
\[
E((\hat{\beta} - \beta)' W (\hat{\beta} - \beta)) = \sum_i \sum_j w_{ij} E((\hat{\beta}_i - \beta_i)(\hat{\beta}_j - \beta_j))
\]
where \( W \) is \( k \times k \) fixed positive definite matrix of weights \( w_{ij} \). The two popular choice of weight matrix \( W \) are
(i) \( W \) is an identity matrix, i.e. \( W = I \), then \( E((\hat{\beta} - \beta)' (\hat{\beta} - \beta)) \) is called as the total mean squared error (MSE) of \( \hat{\beta} \).
(ii) \( W = X'X \), then

\[
E\left( \hat{\beta} - \beta \right)\left( X'X \right)\left( \hat{\beta} - \beta \right) = E\left( X\hat{\beta} - X\beta \right)\left( X\hat{\beta} - X\beta \right)
\]

is called as the predictive mean squared error of \( \hat{\beta} \). Note that \( \hat{E}(y) = X\hat{\beta} \) is the predictor of average value \( E(y) = X\beta \) and \( (X\hat{\beta} - X\beta) \) is the corresponding prediction error.

There can be other choices of \( W \) and it depends entirely on the analyst how to define the loss function so that the variability is minimum.

If a random vector with \( k \) elements \( (k > 2) \) is normally distributed as \( N(\mu, I) \), \( \mu \) being the mean vector, then Stein established that if the linearity and unbiasedness are dropped, then it is possible to improve upon the maximum likelihood estimator of \( \mu \) under the criterion of total MSE. Later, this result was generalized by James and Stein for linear regression model. They demonstrated that if the criteria of linearity and unbiasedness of the estimators are dropped, then a nonlinear estimator can be obtained which has better performance than the best linear unbiased estimator under the criterion of predictive MSE. In other words, James and Stein established that OLSE is inadmissible for \( k > 2 \) under predictive MSE criterion, i.e., for \( k > 2 \), there exists an estimator \( \hat{\beta} \) such that

\[
E\left( \hat{\beta} - \beta \right)\left( X'X \right)\left( \hat{\beta} - \beta \right) \leq E\left( b - \beta \right)\left( X'X \right)\left( b - \beta \right)
\]

for all values of \( \beta \) with strict inequality holding for some values of \( \beta \). For \( k \leq 2 \), no such estimator exists and we say that \( "b \) can be beaten" in this sense. Thus it is possible to find estimators which will beat \( b \) in this sense. So a nonlinear and biased estimator can be defined which has better performance than OLSE. Such an estimator is **Stein-rule estimator** given by

\[
\hat{\beta} = \left[ 1 - c \frac{\sigma^2}{b'X'Xb} \right] b \quad \text{when } \sigma^2 \text{ is known}
\]

and

\[
\hat{\beta} = \left[ 1 - c \frac{e'e}{b'X'Xb} \right] b \quad \text{when } \sigma^2 \text{ is unknown.}
\]

Here \( c \) is a fixed positive characterizing scalar, \( e'e \) is the residuum sum of squares based on OLSE and \( e = y - Xb \) is the residual. By assuming different values to \( c \), we can generate different estimators. So a class of estimators characterized by \( c \) can be defined. This is called as a family of Stein-rule estimators.
Let
\[
\left[1-c\frac{\sigma^2}{b'X'Xb}\right] = \delta
\]
be a scalar quantity. Then
\[
\hat{\beta} = \delta b.
\]

So essentially we say that instead of estimating \( \beta_1, \beta_2, \ldots, \beta_k \) by \( b_1, b_2, \ldots, b_k \) we estimate them by \( \delta\hat{\beta}_1, \delta\hat{\beta}_2, \ldots, \delta\hat{\beta}_k \), respectively. So in order to increase the efficiency, the OLSE is multiplied by a constant \( \delta \). Thus \( \delta \) is called the **shrinkage factor**. As Stein-rule estimators attempts to shrink the components of \( b \) towards zero, so these estimators are known as **shrinkage estimators**.

First we discuss a result which is used to prove the dominance of Stein-rule estimator over OLSE.

**Result:** Suppose a random vector \( Z \) of order \((k \times 1)\) is normally distributed as \( N(\mu, I) \) where \( \mu \) is the mean vector and \( I \) is the covariance matrix.

Then
\[
E \left[ \frac{Z'(Z-\mu)}{Z'Z} \right] = (k-2)E \left( \frac{1}{Z'Z} \right).
\]

An important point to be noted in this result is that the left hand side depends on \( \mu \) but right hand side is independent of \( \mu \).

Now we consider the Stein-rule estimator \( \hat{\beta} \) when \( \sigma^2 \) is known. Note that
\[
E(\hat{\beta}) = E(b) - c\sigma^2 E \left( \frac{b}{b'X'Xb} \right)
= \beta - (\text{In general, a non-zero quantity})
\neq 0,
\]
in general. Thus the Stein-rule estimator is biased while OLSE \( b \) is unbiased for \( \beta \).

The predictive risk of \( b \) and \( \hat{\beta} \) are
\[
PR(b) = E((b-\beta)'X'X(b-\beta))
PR(\hat{\beta}) = E((\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta)).
\]
The Stein-rule estimator \( \hat{\beta} \) is better than OLS \( \beta \) under the criterion of predictive risk if

\[
PR(\hat{\beta}) < PR(\beta).
\]

Solving the expressions, we get

\[
\begin{align*}
b - \beta &= (X'X)^{-1}X'\varepsilon \\
PR(\hat{\beta}) &= E\left[ \varepsilon'X(X'X)^{-1}X'X(X'X)^{-1}X'\varepsilon \right] \\
&= E\left[ \varepsilon'X(X'X)^{-1}X'\varepsilon \right] \\
&= E\left[ tr(X'X)^{-1}X'\varepsilon\varepsilon'X \right] \\
&= tr(X'X)^{-1}X'E(\varepsilon\varepsilon')X \\
&= \sigma^2 tr(X'X)^{-1}X'X \\
&= \sigma^2 trI_k \\
&= \sigma^2 k. 
\end{align*}
\]

\[
PR(\hat{\beta}) = E\left[ (b - \beta) - \frac{c\sigma^2}{b'X'Xb} b \right] X'X \left[ (b - \beta) - \frac{c\sigma^2}{b'X'Xb} b \right] \\
= E(b - \beta)'X'X(b - \beta) - E\left[ \frac{c\sigma^2}{b'X'Xb} \{b'X'X(b - \beta) + (b - \beta)'X'Xb\} \right] \\
+ E\left[ \frac{c^2\sigma^4}{(b'X'Xb)^2} b'X'Xb \right] \\
= \sigma^2 k - 2E\left[ \frac{c\sigma^2 (b - \beta)'X'Xb}{b'X'Xb} \right] + E\left[ \frac{c^2\sigma^4}{b'X'Xb} \right].
\]

Suppose

\[
Z = \frac{1}{\sigma}(X'X)^{1/2}b
\]

or

\[
b = \sigma(X'X)^{-1/2}Z
\]

\[
\mu = \frac{1}{\sigma}(X'X)^{1/2}\beta
\]

or

\[
\beta = \sigma(X'X)^{-1/2}\mu
\]

and \( Z \sim N(\mu, I) \), i.e., \( Z_1, Z_2, \ldots, Z_k \) are independent. Substituting these values in the expressions for

\[
PR(\hat{\beta}),
\]

we get
\[
PR(\hat{\beta}) = \sigma^2 k - 2E\left[c\sigma^2 \frac{(Z - \mu)'Z}{\sigma^2 Z'Z}\right] + E\left[\frac{c^2\sigma^2}{\sigma^2 Z'Z}\right] \\
= \sigma^2 k - 2E\left[c\sigma^2 \frac{Z'(Z - \mu)}{Z'Z}\right] + E\left[\frac{c^2\sigma^2}{Z'Z}\right] \\
= \sigma^2 k - 2c\sigma^2 E\left[\frac{Z'(Z - \mu)}{Z'Z}\right] + c^2\sigma^2 E\left[\frac{1}{Z'Z}\right] \\
= \sigma^2 k - c\sigma^2 \left[2(k-2) - c\right] E\left[\frac{1}{Z'Z}\right] \quad \text{(using the result)} \\
= PR(b) - c\sigma^2 \left[2(k-2) - c\right] E\left[\frac{1}{Z'Z}\right].
\]

Thus
\[
PR(\hat{\beta}) < PR(b)
\]
if and only if
\[
c\sigma^2 \left[2(k-2) - c\right] E\left[\frac{1}{Z'Z}\right] > 0.
\]

Since \( Z \sim N(\mu, I) \), \( \sigma^2 > 0 \). So \( Z'Z \) has a noncentral Chi-square distribution. Thus
\[
E\left[\frac{1}{Z'Z}\right] > 0 \\
\Rightarrow c\left[Z(k-2) - c\right] > 0.
\]

Since \( c > 0 \) is assumed, so this inequality holds true when
\[
2(k-2) - c > 0
\]
or
\[
0 < c < 2(k-2) \quad \text{provided} \quad k > 2.
\]

So as long as \( 0 < c < 2(k-2) \) is satisfied, the Stein-rule estimator will have smaller predictive risk than OLSE.

This inequality is not satisfied for \( k = 1 \) and \( k = 2 \). The largest gains efficiency arises when \( c = k - 2 \). So if the number of explanatory variables are more than two, then it is always possible to construct an estimator which is better than OLSE.
When $\sigma^2$ is unknown then it can be shown that the Stein-rule estimator

$$\hat{\beta} = \left[ 1 - c \frac{e'e}{b' X' X b} \right] b$$

is better than OLSE $b$ if and only if

$$0 < c < \frac{2(k-2)}{n-k+2}; \quad k > 2.$$  

The optimum choice of $c$ giving largest gain is efficiency is

$$c = \frac{k-2}{n-k+2}.$$
A basic assumption in analyzing the performance of estimators in a multiple regression is that the explanatory variables and disturbance terms are independently distributed. The violation of such assumption disturbs the optimal properties of the estimators. The instrumental variable estimation method helps in estimating the regression coefficients in multiple linear regression model when such violation occurs.

Consider the multiple linear regression model

\[ y = X\beta + \varepsilon \]

where \( y \) is \((n \times 1)\) vector of observation on study variable, \( X \) is \((n \times k)\) matrix of observations on \( X_1, X_2, \ldots, X_k \), \( \beta \) is a \((k \times 1)\) vector of regression coefficient and \( \varepsilon \) is a \((n \times 1)\) vector of disturbances.

Suppose one or more explanatory variables is correlated with the disturbances in the limit, then we can write

\[ \text{plim } \left( \frac{1}{n} X' \varepsilon \right) \neq 0. \]

The consequences of such an assumption on ordinary least squares estimator are as follows:

\[ b = (X'X)^{-1} X' y \]

\[ = (X'X)^{-1} X' (X\beta + \varepsilon) \]

\[ b - \beta = (X'X)^{-1} X' \varepsilon \]

\[ = \left( \frac{X'X}{n} \right)^{-1} \left( \frac{X'\varepsilon}{n} \right) \]

\[ \text{plim } (b - \beta) = \text{plim } \left( \frac{X'X}{n} \right)^{-1} \text{plim } \left( \frac{X'\varepsilon}{n} \right) \neq 0 \]

assuming \( \text{plim } \left( \frac{X'X}{n} \right) = \Sigma_{XX} \) exists and is nonsingular. Consequently \( \text{plim } b \neq \beta \) and thus the OLSE becomes an inconsistent estimator of \( \beta \).

To overcome this problem and to obtain a consistent estimator of \( \beta \), the instrumental variable estimation can be used.
Consider the model
\[ y = X\beta + \varepsilon \text{ with } \text{plim}\left(\frac{1}{n}X'\varepsilon\right) \neq 0. \]

Suppose that it is possible to find a data matrix \( Z \) of order \((n \times k)\) with the following properties:

(i) \[ \text{plim} \left( \frac{Z'X}{n} \right) = \Sigma_{zx} \text{ is a finite and nonsingular matrix of full rank.} \]
This interprets that the variables in \( Z \) are correlated with those in \( X \), in the limit.

(ii) \[ \text{plim} \left( \frac{Z'\varepsilon}{n} \right) = 0, \]
i.e., the variables in \( Z \) are uncorrelated with \( \varepsilon \), in the limit.

(iii) \[ \text{plim} \left( \frac{Z'Z}{n} \right) = \Sigma_{zz} \text{ exists.} \]

Thus \( Z \) – variables are postulated to be
- uncorrelated with \( \varepsilon \), in the limit and
- to have nonzero cross product with \( X \).

Such variables are called instrumental variables.

If some of \( X \) variables are likely to be uncorrelated with \( \varepsilon \), then these can be used to form some of the columns of \( Z \) and extraneous variables are found only for the remaining columns.

First we understand the role of the term \( X'\varepsilon \) in the OLS estimation. The OLS estimate \( \hat{\beta} \) of \( \beta \) is derived by solving the equation
\[
\frac{\partial (y - X\beta)'(y - X\beta)}{\partial \beta} = 0
\]
or
\[ X'y = X'X\hat{\beta} \]
or
\[ X'(y - X\hat{\beta}) = 0. \]

Now we look at this normal equation as if it is obtained by pre-multiplying \( y = X\beta + \varepsilon \) by \( X' \) as
\[ X'y = X'X\beta + X'\varepsilon \]
where the term $X'\varepsilon$ is dropped and $\beta$ is replaced by $b$. The disappearance of $X'\varepsilon$ can be explained when $X$ and $\varepsilon$ are uncorrelated as follows:

\[
\frac{X'y}{n} = \frac{X'X}{n}\beta + \frac{X'\varepsilon}{n}
\]

\[
\text{plim} \left( \frac{X'y}{n} \right) = \text{plim} \left( \frac{X'X}{n} \right) \beta + \text{plim} \left( \frac{X'\varepsilon}{n} \right)
\]

\[
\Rightarrow \beta = \text{plim} \left( \frac{X'X}{n} \right)^{-1} \left[ \text{plim} \left( \frac{X'y}{n} \right) - \text{plim} \left( \frac{X'\varepsilon}{n} \right) \right].
\]

Let

\[
\text{plim} \left( \frac{X'X}{n} \right) = \Sigma_{XX}
\]

\[
\text{plim} \left( \frac{X'y}{n} \right) = \Sigma_{XY}
\]

where population cross moments $\Sigma_{XX}$ and $\Sigma_{XY}$ are finite, $\Sigma_{XX}$ is finite and nonsingular.

If $X$ and $\varepsilon$ are uncorrelated so that

\[
\text{plim} \left( \frac{X'\varepsilon}{n} \right) = 0,
\]

then

\[
\beta = \Sigma_{XX}^{-1} \Sigma_{XY}.
\]

If $\Sigma_{XX}$ is estimated by sample cross moment $\frac{X'X}{n}$ and $\Sigma_{XY}$ is estimated by sample cross moment $\frac{X'y}{n}$, then the OLS estimator of $\beta$ is obtained as

\[
b = \left( \frac{X'X}{n} \right)^{-1} \left( \frac{X'y}{n} \right)
\]

\[
= \left( X'X \right)^{-1} X'y.
\]

Such an analysis suggests to use $Z$ to pre-multiply the multiple regression model as follows:
\[ Z'y = Z'X\beta + Z'\varepsilon \]
\[
\frac{Z'y}{n} = \left( \frac{Z'X}{n} \right)\beta + \left( \frac{Z'\varepsilon}{n} \right)
\]
\[
\text{plim} \left( \frac{Z'y}{n} \right) = \text{plim} \left( \frac{Z'X}{n} \right)\beta + \text{plim} \left( \frac{Z'\varepsilon}{n} \right)
\]
\[
\Rightarrow \beta = \text{plim} \left( \frac{Z'X}{n} \right)^{-1} \left[ \text{plim} \left( \frac{Z'y}{n} \right) - \text{plim} \left( \frac{Z'\varepsilon}{n} \right) \right]
\]
\[
= \Sigma_{zx}^{-1} \Sigma_{zy}.
\]

Substituting the sample cross moment
- \( \frac{Z'X}{n} \) of \( \Sigma_{zx} \) and
- \( \frac{Z'y}{n} \) of \( \Sigma_{zy} \),

Thus the following instrumental variable estimator of \( \beta \) is obtained:
\[
\hat{\beta}_{IV} = (Z'X)^{-1} Z'y
\]

which is termed as \textbf{instrumental variable estimator} of \( \beta \) and this method is called as \textbf{instrumental variable method}.

Since
\[
\hat{\beta}_{IV} - \beta = (Z'X)^{-1} Z'(X\beta + \varepsilon) - \beta
\]
\[
= (Z'X)^{-1} Z'\varepsilon
\]
\[
\text{plim} (\hat{\beta}_{IV} - \beta) = \text{plim} \left( (Z'X)^{-1} Z'\varepsilon \right)
\]
\[
= \text{plim} \left( \frac{Z'X}{n} \right)^{-1} \text{plim} \left( \frac{Z'\varepsilon}{n} \right)
\]
\[
= \Sigma_{zx}^{-1} \cdot 0
\]
\[
\Rightarrow \text{plim} \hat{\beta}_{IV} = \beta.
\]

Thus the instrumental variable estimator is consistent. Note that the variables \( Z_1, Z_2, \ldots, Z_k \) in \( Z \) are chosen such that they are uncorrelated with \( \varepsilon \) and correlated with \( X \), at least asymptotically, so that the second order moment matrix \( \Sigma_{zx} \) exists and is nonsingular.
**Asymptotic distribution:**

The asymptotic distribution of

\[
\sqrt{n}(\hat{\beta}_w - \beta) = \left(\frac{1}{n} Z'X\right)^{-1} \frac{1}{\sqrt{n}} Z'\epsilon
\]

is normal with mean vector 0 and asymptotic covariance matrix given by

\[
Asy Var(\hat{\beta}_w) = \text{plim} \left[ \left(\frac{Z'X}{n}\right)^{-1} \frac{1}{n} Z' E(\epsilon \epsilon') Z \left(\frac{X'Z}{n}\right)^{-1} \right]
\]

\[
= \sigma^2 \text{plim} \left[ \left(\frac{Z'X}{n}\right)^{-1} \left(\frac{Z'Z}{n}\right) \left(\frac{X'Z}{n}\right)^{-1} \right]
\]

\[
= \sigma^2 \text{plim} \left(\frac{Z'X}{n}\right)^{-1} \text{plim} \left(\frac{Z'Z}{n}\right) \text{plim} \left(\frac{X'Z}{n}\right)^{-1}
\]

\[
= \sigma^2 \Sigma_{xz}^{-1} \Sigma_{zz}^{-1} \Sigma_{xz}^{-1}.
\]

For large sample,

\[
V(\hat{\beta}_w) = \frac{\sigma^2}{n} \Sigma_{xz}^{-1} \Sigma_{zz}^{-1} \Sigma_{xz}^{-1}
\]

which can be estimated by

\[
\hat{V}(\hat{\beta}_w) = \frac{\sigma^2}{n} \hat{\Sigma}_{xz}^{-1} \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{xz}^{-1}.
\]

\[
= \frac{s^2}{n} \left(\frac{Z'X}{n}\right)^{-1} Z'Z \left(\frac{X'Z}{n}\right)^{-1}
\]

where

\[
s^2 = \frac{1}{n-k} (y - Xb)'(y - Xb),
\]

\[
b = \left(\frac{X'X}{n}\right)^{-1} X'y.
\]

The variance of $\hat{\beta}_w$ is not necessarily a minimum asymptotic variance because there can be more than one sets of instrumental variables that fulfill the requirement of being uncorrelated with $\epsilon$ and correlated with stochastic regressors.
Chapter 16
Measurement Error Models

A fundamental assumption in all the statistical analysis is that all the observations are correctly measured. In the context of multiple regression model, it is assumed that the observations on study and explanatory variables are observed without any error. In many situations, this basic assumption is violated. There can be several reasons for such violation.

- For example, the variables may not be measurable, e.g., taste, climatic conditions, intelligence, education, ability etc. In such cases, the dummy variables are used and the observations can be recorded in terms of values of dummy variables.
- Sometimes the variables are clearly defined but it is hard to take correct observations. For example, the age is generally reported in complete years or in multiple of five.
- Sometimes the variable is conceptually well defined but it is not possible to take correct observation on it. Instead, the observations are obtained on closely related proxy variables, e.g., the level of education is measured by the number of years of schooling.
- Sometimes the variable is well understood but it is qualitative in nature. For example, intelligence is measured by intelligence quotient (IQ) scores.

In all such cases, the true value of variable can not be recorded. Instead, it is observed with some error. The difference between the observed and true values of the variable is called as measurement error or errors-in-variables.

**Difference between disturbances and measurement errors:**

The disturbances in the linear regression model arise due to factors like unpredictable element of randomness, lack of deterministic relationship, measurement error in study variable etc. The disturbance term is generally thought of as representing the influence of various explanatory variables that have not actually been included in the relation. The measurement errors arise due to the use of an imperfect measure of true variables.
Large and small measurement errors

If the magnitude of measurement errors is small, then they can be assumed to be merged in the disturbance term and they will not affect the statistical inferences much. On the other hand, if they are large in magnitude, then they will lead to incorrect and invalid statistical inferences. For example, in the context of linear regression model, the ordinary least squares estimator (OLSE) is the best linear unbiased estimator of regression coefficient when measurement errors are absent. When the measurement errors are present in the data, the same OLSE becomes biased as well as inconsistent estimator of regression coefficients.

Consequences of measurement errors:

We first describe the measurement error model. Let the true relationship between correctly observed study and explanatory variables be

\[ \hat{y} = \hat{X} \beta \]

where \( \hat{y} \) is a \((n \times 1)\) vector of true observation on study variable, \( \hat{X} \) is a \((n \times k)\) matrix of true observations on explanatory variables and \( \beta \) is a \((k \times 1)\) vector of regression coefficients. The value \( \hat{y} \) and \( \hat{X} \) are not observable due to the presence of measurement errors. Instead, the values of \( \tilde{y} \) and \( \tilde{X} \) are observed with additive measurement errors as

\[ y = \tilde{y} + u \]
\[ X = \tilde{X} + V \]

where \( y \) is a \((n \times 1)\) vector of observed values of study variables which are observed with \((n \times 1)\) measurement error vector \( u \). Similarly, \( X \) is a \((n \times k)\) matrix of observed values of explanatory variables which are observed with \((n \times k)\) matrix \( V \) of measurement errors in \( X \). In such a case, the usual disturbance term can be assumed to be subsumed in \( u \) without loss of generality. Since our aim is to see the impact of measurement errors, so it is not considered separately in the present case.

Alternatively, the same setup can be expressed as

\[ y = \tilde{X} \beta + u \]
\[ X = \tilde{X} + V \]

where it can be assumed that only \( X \) is measured with measurement errors \( V \) and \( u \) can be considered as the usual disturbance term in the model.
In case, some of the explanatory variables are measured without any measurement error then the corresponding values in $V$ will be set to zero.

We assume that
\[ E(u) = 0, E(u'u') = \sigma^2 I \]
\[ E(V) = 0, E(V'V) = \Omega, E(V'u) = 0. \]

The following set of equations describes the measurement error model
\[
\tilde{y} = \tilde{X} \beta \\
y = \tilde{y} + u \\
X = \tilde{X} + V
\]
which can be re-expressed as
\[
y = \tilde{y} + u \\
= \tilde{X} \beta + u \\
= (X - V) \beta + u \\
= X \beta + (u - V \beta) \\
= X \beta + \omega
\]
where $\omega = u - V \beta$ is called as the composite disturbance term. This model resemble like a usual linear regression model. A basic assumption in linear regression model is that the explanatory variables and disturbances are uncorrelated. Let us verify this assumption in the model $y = X \beta + w$ as follows:
\[
E \left[ \{X - E(X)\}' \{\omega - E(\omega)\} \right] = E \left[ V'(u - V \beta) \right] \\
= E \left[ V'u \right] - E \left[ V'V \right] \beta \\
= 0 - \Omega \beta \\
= -\Omega \beta \\
\neq 0.
\]
Thus $X$ and $\omega$ are correlated. So OLS will not provide efficient result.

Suppose we ignore the measurement errors and obtain the OLSE. Note that ignoring the measurement errors in the data does not mean that they are not present. We now observe the properties of such an OLSE under the setup of measurement error model.
The OLSE is
\[
b = (X'X)^{-1} X' y
\]
\[
b - \beta = (X'X)^{-1} X'(X\beta + \omega) - \beta
\]
\[
= (X'X)^{-1} X' \omega
\]
\[
E(b - \beta) = E[(X'X)^{-1} X' \omega]
\]
\[
\neq (X'X)^{-1} X' E(\omega)
\]
\[
\neq 0
\]
as \( X \) is a random matrix which is correlated with \( \omega \). So \( b \) becomes a biased estimator of \( \beta \).

Now we check the consistency property of OLSE. Assume
\[
\text{plim} \left( \frac{1}{n} X' \tilde{X} \right) = \Sigma_{xx}
\]
\[
\text{plim} \left( \frac{1}{n} V'V \right) = \Sigma_{vv}
\]
\[
\text{plim} \left( \frac{1}{n} \tilde{X}'V \right) = 0
\]
\[
\text{plim} \left( \frac{1}{n} V'u \right) = 0.
\]
Then
\[
\text{plim}(b - \beta) = \text{plim} \left[ \left( \frac{X'X}{n} \right)^{-1} \left( \frac{X'\omega}{n} \right) \right]
\]
\[
\frac{1}{n} X'X = \frac{1}{n} (\tilde{X} + V)'(\tilde{X} + V)
\]
\[
= \frac{1}{n} \tilde{X}'\tilde{X} + \frac{1}{n} \tilde{X}'V + \frac{1}{n} V'\tilde{X} + \frac{1}{n} V'V
\]
\[
\text{plim} \left( \frac{1}{n} X'X \right) = \text{plim} \left( \frac{1}{n} \tilde{X}'\tilde{X} \right) + \text{plim} \left( \frac{1}{n} \tilde{X}'V \right) + \text{plim} \left( \frac{1}{n} V'\tilde{X} \right) + \text{plim} \left( \frac{1}{n} V'V \right)
\]
\[
= \Sigma_{xx} + 0 + 0 + \Sigma_{vv}
\]
\[
= \Sigma_{xx} + \Sigma_{vv}
\]
\[
\frac{1}{n} X'\omega = \frac{1}{n} \tilde{X}'\omega + \frac{1}{n} V'\omega
\]
\[
= \frac{1}{n} \tilde{X}'(u - V\beta) + \frac{1}{n} V'(u - V\beta)
\]
\[
\text{plim} \left( \frac{1}{n} X' \omega \right) = \text{plim} \left( \frac{1}{n} \tilde{X}' u \right) - \text{plim} \left( \frac{1}{n} \tilde{X}' V \right) \beta + \text{plim} \left( \frac{1}{n} V' u \right) - \text{plim} \left( \frac{1}{n} V' V \right) \beta
\]
\[
= 0 - 0 - \Sigma_{\omega} \beta
\]
\[
\text{plim} (b - \beta) = \text{plim} \left( \frac{X' X}{n} \right)^{-1} \text{plim} \left( \frac{X' \omega}{n} \right)
\]
\[
= - \left( \Sigma_{xx} + \Sigma_{vv} \right)^{-1} \Sigma_{\omega} \beta
\]
\[
\neq 0.
\]

Thus \( b \) is an inconsistent estimator of \( \beta \). Such inconsistency arises essentially due to correlation between \( X \) and \( \omega \).

**Note:** It should not be misunderstood that the OLSE \( b = (X'X)^{-1} X'y \) is obtained by minimizing \( S = \omega' \omega = (y - X \beta)'(y - X \beta) \) in the model \( y = X \beta + \omega \). In fact \( \omega' \omega \) cannot be minimized as in the case of usual linear regression, because the composite error \( \omega = u - V \beta \) is itself a function of \( \beta \).

To see the nature of consistency, consider the simple linear regression model with measurement error as
\[
\tilde{y}_i = \beta_0 + \beta_1 \tilde{x}_i, \quad i = 1, 2, ..., n
\]
\[
y_i = \tilde{y}_i + u_i
\]
\[
x_i = \tilde{x}_i + v_i.
\]

Now
\[
X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \tilde{X} = \begin{pmatrix} 1 & \tilde{x}_1 \\ 1 & \tilde{x}_2 \\ \vdots & \vdots \\ 1 & \tilde{x}_n \end{pmatrix}, \quad V = \begin{pmatrix} 0 & v_1 \\ 0 & v_2 \\ \vdots & \vdots \\ 0 & v_n \end{pmatrix}
\]
and assuming that
\[
\text{plim} \left( \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i \right) = \mu
\]
\[
\text{plim} \left( \frac{1}{n} \sum_{i=1}^{n} (\tilde{x}_i - \mu)^2 \right) = \sigma_x^2,
\]
we have
\[
\Sigma_{xx} = \text{plim} \left( \frac{1}{n} \mathbf{X}' \mathbf{X} \right) \\
= \text{plim} \left( \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \right) \\
= \left( \begin{array}{cc}
\mu & \sigma_x^2 + \mu^2 \\
\mu & \sigma_x^2 + \mu^2
\end{array} \right).
\]

Also,
\[
\Sigma_{vv} = \text{plim} \left( \frac{1}{n} \mathbf{V}' \mathbf{V} \right) \\
= \left( \begin{array}{cc}
0 & 0 \\
0 & \sigma_v^2
\end{array} \right).
\]

Now
\[
\text{plim} \left( b - \beta \right) = - \left( \Sigma_{xx} + \Sigma_{vv} \right)^{-1} \Sigma_{vn} \beta \\
\text{plim} \left( b_0 - \beta_0 \right) = - \left( \frac{1}{\mu} \cdot \frac{1}{\sigma_x^2 + \mu^2 + \sigma_v^2} \right)^{-1} \left( \begin{array}{c}
0 \\
0
\end{array} \right) \left( \begin{array}{c}
\beta_0 \\
\beta_1
\end{array} \right) \\
= - \left( \frac{1}{\sigma_x^2 + \sigma_v^2} \right) \left( \begin{array}{cc}
\sigma_x^2 + \mu^2 + \sigma_v^2 & -\mu \\
-\mu & 1
\end{array} \right) \left( \begin{array}{c}
0 \\
\beta \sigma_v^2
\end{array} \right) \\
= \left( \begin{array}{c}
\frac{\sigma_v^2}{\sigma_x^2 + \sigma_v^2} \mu \beta \\
-\frac{\sigma_v^2}{\sigma_x^2 + \sigma_v^2} \beta
\end{array} \right).
\]

Thus we find that the OLSEs of \( \beta_0 \) and \( \beta_1 \) are biased and inconsistent. So if a variable is subjected to measurement errors, it not only affects its own parameter estimate but also affect other estimator of parameter that are associated with those variable which are measured without any error. So the presence of measurement errors in even a single variable not only makes the OLSE of its own parameter inconsistent but also makes the estimates of other regression coefficients inconsistent which are measured without any error.
Forms of measurement error model:

Based on the assumption about the true values of explanatory variable, there are three forms of measurement error model.

Consider the model

\[ \hat{y}_i = \beta_0 + \beta_1 \tilde{x}_i, \quad i = 1, 2, ..., n \]

\[ y_i = \hat{y}_i + u_i \]

\[ x_i = \tilde{x}_i + v_i. \]

1. **Functional form:** When the \( \tilde{x}_i \)'s are unknown constants (fixed), then the measurement error model is said to be in its functional form.

2. **Structural form:** When the \( \tilde{x}_i \)'s are identically and independently distributed random variables, say, with mean \( \mu \) and variance \( \sigma^2 (\sigma^2 > 0) \), the measurement error model is said to be in the structural form.

Note that in case of functional form, \( \sigma^2 = 0. \)

3. **Ultrastructural form:** When the \( \tilde{x}_i \)'s are independently distributed random variables with different means, say \( \mu_i \) and variance \( \sigma^2 (\sigma^2 > 0) \), then the model is said to be in the ultrastructural form. This form is a synthesis of function and structural forms in the sense that both the forms are particular cases of ultrastructural form.

**Methods for consistent estimation of \( \beta \):**

The OLSE of \( \beta \) which is the best linear unbiased estimator becomes biased and inconsistent in the presence of measurement errors. An important objective in measurement error models is how to obtain the consistent estimators of regression coefficients. The instrumental variable estimation and method of maximum likelihood (or method of moments) are utilized to obtain the consistent estimates of the parameters.
**Instrumental variable estimation:**

The instrumental variable method provides the consistent estimate of regression coefficients in linear regression model when the explanatory variables and disturbance terms are correlated. Since in measurement error model, the explanatory variables and disturbance are correlated, so this method helps. The instrumental variable method consists of finding a set of variables which are correlated with the explanatory variables in the model but uncorrelated with the composite disturbances, at least asymptotically, to ensure consistency.

Let \( Z_1, Z_2, \ldots, Z_k \) be the \( k \) instrumental variables. In the context of the model

\[ y = X\beta + \omega, \quad \omega = u - V\beta, \]

let \( Z \) be the \( n \times k \) matrix of \( k \) instrumental variables \( Z_1, Z_2, \ldots, Z_k \), each having \( n \) observations such that

- \( Z \) and \( X \) are correlated, at least asymptotically and
- \( Z \) and \( \omega \) are uncorrelated, at least asymptotically.

So we have

\[
\lim_{n \to \infty} \frac{1}{n} Z' X = \Sigma_{ZX} \\
\lim_{n \to \infty} \frac{1}{n} Z' \omega = 0.
\]

The instrumental variable estimator of \( \beta \) is given by

\[
\hat{\beta}_{iv} = (Z'X)^{-1} Z' y \\
= (Z'X)^{-1} Z' (X\beta + \omega) \\
\hat{\beta}_{iv} - \beta = (Z'X)^{-1} Z' \omega \\
\lim_{n \to \infty} \left( \hat{\beta}_{iv} - \beta \right) = \lim_{n \to \infty} \left( \frac{1}{n} Z' X \right)^{-1} \lim_{n \to \infty} \left( \frac{1}{n} Z' \omega \right) \\
= \Sigma_{ZX}^{-1} \cdot 0 \\
= 0.
\]

So \( \hat{\beta}_{iv} \) is consistent estimator of \( \beta \).

Any instrument that fulfills the requirement of being uncorrelated with the composite disturbance term and correlated with explanatory variables will result in a consistent estimate of parameter. However, there can be various sets of variables which satisfy these conditions to become instrumental variables. Different

*Econometrics* | Chapter 16 | Measurement Error Models | Shalabh, IIT Kanpur
choices of instruments give different consistent estimators. It is difficult to assert that which choice of instruments will give an instrumental variable estimator having minimum asymptotic variance. Moreover, it is also difficult to decide that which choice of the instrumental variable is better and more appropriate in comparison to other. An additional difficulty is to check whether the chosen instruments are indeed uncorrelated with the disturbance term or not.

**Choice of instrument:**

We discuss some popular choices of instruments in a univariate measurement error model. Consider the model

\[ y_i = \beta_0 + \beta_1 x_i + \omega_i, \quad \omega_i = u_i - \beta_1 v_i, \quad i = 1, 2, \ldots, n. \]

A variable that is likely to satisfy the two requirements of an instrumental variable is the discrete grouping variable. The Wald’s, Bartlett’s and Durbin’s methods are based on different choices of discrete grouping variables.

**1. Wald’s method**

Find the median of the given observations \( x_1, x_2, \ldots, x_n \). Now classify the observations by defining an instrumental variable \( Z \) such that

\[
Z_i = \begin{cases} 
1 & \text{if } x_i > \text{median } (x_1, x_2, \ldots, x_n) \\
-1 & \text{if } x_i < \text{median } (x_1, x_2, \ldots, x_n).
\end{cases}
\]

In this case,

\[
Z = \begin{pmatrix} 
1 & Z_1 \\
1 & Z_2 \\
\vdots & \vdots \\
1 & Z_n
\end{pmatrix}, \quad X = \begin{pmatrix} 
1 & x_1 \\
1 & x_2 \\
\vdots & \vdots \\
1 & x_n
\end{pmatrix}.
\]

Now form two groups of observations as follows.

- One group with those \( x_i \)'s below the median of \( x_1, x_2, \ldots, x_n \). Find the means of \( y_i \)'s and \( x_i \)'s, say \( \bar{y}_1 \) and \( \bar{x}_1 \), respectively in this group.

- Another group with those \( x_i \)'s above the median of \( x_1, x_2, \ldots, x_n \). Find the means of \( y_i \)'s and \( x_i \)'s, say \( \bar{y}_2 \) and \( \bar{x}_2 \), respectively in this group.
Now we find the instrumental variable estimator under this set up as follows. Let \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \), \( \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \).

\[
\hat{\beta}_{IV} = (Z'X)^{-1} Z'y
\]

\[
Z'X = \left( \sum_{i=1}^{n} Z_i \right) \left( \sum_{i=1}^{n} x_i \right) = \begin{pmatrix} n & n \bar{x} \\ 0 & \frac{n}{2} (\bar{x}_2 - \bar{x}_1) \end{pmatrix}
\]

\[
Z'y = \sum_{i=1}^{n} y_i = \begin{pmatrix} n \bar{y} \\ \frac{n}{2} (\bar{y}_2 - \bar{y}_1) \end{pmatrix}
\]

\[
\begin{pmatrix} \hat{\beta}_{0IV} \\ \hat{\beta}_{1IV} \end{pmatrix} = \begin{pmatrix} n & n \bar{x} \\ 0 & \frac{n}{2} (\bar{x}_2 - \bar{x}_1) \end{pmatrix}^{-1} \begin{pmatrix} n \bar{y} \\ \frac{n}{2} (\bar{y}_2 - \bar{y}_1) \end{pmatrix}
\]

\[
= \frac{2}{(\bar{x}_2 - \bar{x}_1)} \begin{pmatrix} \bar{x}_2 - \bar{x}_1 \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ \bar{y}_2 - \bar{y}_1 \end{pmatrix} = \begin{pmatrix} \bar{y} - \frac{(\bar{y}_2 - \bar{y}_1)}{(\bar{x}_2 - \bar{x}_1)} \bar{x} \\ \frac{(\bar{y}_2 - \bar{y}_1)}{(\bar{x}_2 - \bar{x}_1)} \bar{x} \end{pmatrix}
\]

\[
\Rightarrow \hat{\beta}_{1IV} = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1}
\]

\[
\hat{\beta}_{0IV} = \bar{y} - \frac{(\bar{y}_2 - \bar{y}_1)}{(\bar{x}_2 - \bar{x}_1)} \bar{x} = \bar{y} - \hat{\beta}_{1IV} \bar{x}.
\]

If \( n \) is odd, then the middle observations can be deleted. Under fairly general conditions, the estimators are consistent but are likely to have large sampling variance. This is the limitation of this method.
2. Bartlett’s method:

Let $x_1, x_2, ..., x_n$ be the $n$ observations. Rank these observation and order them in an increasing or decreasing order. Now three groups can be formed, each containing $n/3$ observations. Define the instrumental variable as

$$Z_i = \begin{cases} 
1 & \text{if observation is in the top group} \\
0 & \text{if observation is in the middle group} \\
-1 & \text{if observation is in the bottom group.}
\end{cases}$$

Now discard the observations in the middle group and compute the means of $y_i$'s and $x_i$'s in

- bottom group, say $\bar{y}_1$ and $\bar{x}_1$ and
- top group, say $\bar{y}_3$ and $\bar{x}_3$.

Substituting the values of $X$ and $Z$ in $\hat{\beta}_{IV} = (Z'X)^{-1}Z'y$ and on solving, we get

$$\hat{\beta}_{IV} = \frac{\bar{y}_3 - \bar{y}_1}{\bar{x}_3 - \bar{x}_1},$$

$$\hat{\beta}_{0IV} = \bar{y} - \hat{\beta}_{IV} \bar{x}.$$ 

These estimators are consistent. No conclusive evidences are available to compare the Bartlett’s method and Wald’s method but three grouping method generally provides more efficient estimates than two grouping method is many cases.

3. Durbin’s method

Let $x_1, x_2, ..., x_n$ be the observations. Arrange these observations in an ascending order. Define the instrumental variable $Z_i$ as the rank of $x_i$. Then substituting the suitable values of $Z$ and $X$ in $\hat{\beta}_{IV} = (Z'X)^{-1}Z'y$, we get the instrumental variable estimators

$$\hat{\beta}_{IV} = \frac{\sum_{i=1}^{n} Z_i (y_i - \bar{y})}{\sum_{i=1}^{n} Z_i (x_i - \bar{x})}.$$
When there are more than one explanatory variables, one may choose the instrument as the rank of that particular variable.

Since the estimator uses more information, it is believed to be superior in efficiency to other grouping methods. However, nothing definite is known about the efficiency of this method.

In general, the instrumental variable estimators may have fairly large standard errors in comparison to ordinary least square estimators which is the price paid for inconsistency. However, inconsistent estimators have little appeal.

**Maximum likelihood estimation in structural form**

Consider the maximum likelihood estimation of parameters in the simple measurement error model given by

\[
\tilde{y}_i = \beta_0 + \beta_\tilde{x}_i, \ i = 1,2,...,n
\]

\[
y_i = \tilde{y}_i + u_i
\]

\[
x_i = \tilde{x}_i + v_i.
\]

Here \((\tilde{x}_i, \tilde{y}_i)\) are unobservable and \((x_i, y_i)\) are observable.

Assume

\[
E(u_i) = 0, \ E(u_iu_j) = \begin{cases} \sigma_u^2 & \text{if } i = j \\ 0 & \text{if } i \neq j, \end{cases}
\]

\[
E(v_i) = 0, \ E(v_i v_j) = \begin{cases} \sigma_v^2 & \text{if } i = j \\ 0 & \text{if } i \neq j, \end{cases}
\]

\[
E(u_i v_j) = 0 \ \text{for all } i = 1,2,...,n; \ j = 1,2,...,n.
\]

For the application of method of maximum likelihood, we assume the normal distribution for \(u_i\) and \(v_i\).

We consider the estimation of parameters in the structural form of the model in which \(\tilde{x}_i\)'s are stochastic.

So assume

\[
\tilde{x}_i \sim N(\mu, \sigma^2)
\]

and \(\tilde{x}_i\)'s are independent of \(u_i\) and \(v_i\).
Thus

\[ E(\tilde{x}_i) = \mu \]
\[ Var(\tilde{x}_i) = \sigma^2 \]
\[ E(x_i) = \mu \]
\[ Var(x_i) = E\left[(x_i - E(x_i))^2\right] \]
\[ = E\left[(\tilde{x}_i + v_i - \mu)^2\right] \]
\[ = E(\tilde{x}_i - \mu)^2 + E(v_i^2) - 2(\tilde{x}_i - \mu)v_i \]
\[ = \sigma^2 + \sigma_v^2 \]
\[ E(y_i) = \beta_0 + \beta_1 E(\tilde{x}_i) \]
\[ = \beta_0 + \beta_1 \mu. \]

\[ Var(y_i) = E\left[(y_i - E(y_i))^2\right] \]
\[ = E\left[\beta_0 + \beta_1 \tilde{x}_i + u_i - \beta_0 - \beta_1 \mu\right]^2 \]
\[ = \beta_1^2 E(\tilde{x}_i - \mu)^2 + E(u_i^2) - 2\beta_1 E(\tilde{x}_i - \mu)u_i \]
\[ = \beta_1^2 \sigma^2 + \sigma_u^2 \]

\[ Cov(x_i, y_i) = E\left[(x_i - E(x_i))\{y_i - E(y_i)\}\right] \]
\[ = E\left[\{\tilde{x}_i + v_i - \mu\}\{\beta_0 + \beta_1 \tilde{x}_i + u_i - \beta_0 - \beta_1 \mu\}\right] \]
\[ = \beta_1 E(\tilde{x}_i - \mu)^2 + E(\tilde{x}_i - \mu)u_i + \beta_1 E(\tilde{x}_i - \mu)v_i + E(u_i v_i) \]
\[ = \beta_1 \sigma^2 + 0 + 0 + 0 \]
\[ = \beta_1 \sigma^2. \]

So

\[
\begin{pmatrix}
  y_i \\
  x_i
\end{pmatrix}
\sim N\left[
\begin{pmatrix}
  \beta_0 + \beta_1 \mu \\
  \mu
\end{pmatrix},
\begin{pmatrix}
  \beta_1^2 \sigma^2 + \sigma_u^2 & \beta_1 \sigma^2 \\
  \beta_1 \sigma^2 & \sigma^2 + \sigma_v^2
\end{pmatrix}\right].
\]

The likelihood function is the joint probability density function of \( u_i \) and \( v_i, i = 1, 2, \ldots, n \) as
The log-likelihood is

\[ L^* = \ln L = \text{constant} - \frac{n}{2} \left( \ln \sigma_u^2 + \ln \sigma_v^2 \right) - \frac{\sum_{i=1}^{n} (x_i - \bar{x}_i)^2}{2\sigma_u^2} - \frac{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma_v^2}. \]

The normal equations are obtained by equating the partial differentiations equals to zero as

1. \( \frac{\partial L^*}{\partial \beta_0} = \frac{1}{\sigma_v^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 \bar{x}_i) = 0 \)
2. \( \frac{\partial L^*}{\partial \beta_1} = \frac{1}{\sigma_v^2} \sum_{i=1}^{n} \bar{x}_i (y_i - \beta_0 - \beta_1 \bar{x}_i) = 0 \)
3. \( \frac{\partial L^*}{\partial \bar{x}_i} = \frac{1}{\sigma_u^2} (x_i - \bar{x}_i) + \frac{\beta}{\sigma_v^2} (y_i - \beta_0 - \beta_1 \bar{x}_i) = 0, \ i = 1, 2, \ldots, n \)
4. \( \frac{\partial L^*}{\partial \sigma_u^2} = -\frac{n}{2\sigma_u^2} + \frac{1}{2\sigma_v^2} \sum_{i=1}^{n} (x_i - \bar{x}_i)^2 \)
5. \( \frac{\partial L^*}{\partial \sigma_v^2} = -\frac{n}{2\sigma_v^2} + \frac{1}{2\sigma_v^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 \bar{x}_i)^2. \)

These are \((n + 4)\) equations in \((n + 4)\) parameters but summing equation (3) over \(i = 1, 2, \ldots, n\) and using equation (4), we get

\[ \sigma_v^2 = \beta_1^2 \sigma_u^2 \]

which is undesirable.

These equations can be used to estimate the two means \((\mu \text{ and } \beta_0 + \beta_1 \mu)\), two variances and one covariance. The six parameters \(\mu, \beta_0, \beta_1, \sigma_u^2, \sigma_v^2\) and \(\sigma^2\) can be estimated from the following five structural relations derived from these normal equations.
(i) \( \bar{x} = \mu \)
(ii) \( \bar{y} = \beta_0 + \beta_1 \mu \)
(iii) \( m_{xx} = \sigma_x^2 + \sigma_y^2 \)
(iv) \( m_{yy} = \beta_1^2 \sigma_x^2 + \sigma_u^2 \)
(v) \( m_{xy} = \beta_1 \sigma_x \sigma_u \)

where \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \), \( \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \), \( m_{xx} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \), \( m_{yy} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \) and \( m_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \).

These equations can be derived directly using the sufficiency property of the parameters in bivariate normal distribution using the definition of structural relationship as

\[
E(x) = \mu \\
E(y) = \beta_0 + \beta_1 \mu \\
Var(x) = \sigma_x^2 + \sigma_y^2 \\
Var(y) = \beta_1^2 \sigma_x^2 + \sigma_u^2 \\
Cov(x,y) = \beta_1 \sigma_x \sigma_u
\]

We observe that there are six parameters \( \beta_0, \beta_1, \mu, \sigma_x^2, \sigma_y^2, \sigma_u^2 \) to be estimated based on five structural equations (i)-(v). So no unique solution exists. Only \( \mu \) can be uniquely determined while remaining parameters can not be uniquely determined. So only \( \mu \) is identifiable and remaining parameters are unidentifiable. This is called as the problem of identification. One relation is short to obtain a unique solution, so additional a priori restrictions relating any of the six parameters is required.

Note: The same equations (i)-(v) can also be derived using the method of moments. The structural equations are derived by equating the sample and population moments. The assumption of normal distribution for \( u_i, v_i \) and \( \bar{x}_i \) is not needed in case of method of moments.
Additional information for the consistent estimation of parameters:

The parameters in the model can be consistently estimated only when some additional information about the model is available.

From equations (i) and (ii), we have

\[ \hat{\mu} = \bar{x} \]

and so \( \mu \) is clearly estimated. Further

\[ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \]

is estimated if \( \hat{\beta}_1 \) is uniquely determined. So we consider the estimation of \( \beta_1, \sigma^2, \sigma_u^2 \) and \( \sigma_v^2 \) only. Some additional information is required for the unique determination of these parameters. We consider now various type of additional information which are used for estimating the parameters uniquely.

1. \( \sigma_v^2 \) is known:

Suppose \( \sigma_v^2 \) is known a priori. Now the remaining parameters can be estimated as follows:

\[ m_{xx} = \sigma^2 + \sigma_v^2 \Rightarrow \hat{\sigma}^2 = m_{xx} - \sigma_v^2 \]
\[ m_{xy} = \beta_1 \sigma^2 \Rightarrow \hat{\beta}_1 = \frac{m_{xy}}{m_{xx} - \sigma_v^2} \]
\[ m_{yy} = \beta_1 \sigma^2 + \sigma_v^2 \Rightarrow \hat{\sigma}_v^2 = m_{yy} - \hat{\beta}_1 \hat{\sigma}^2 \]

\[ = m_{yy} - \frac{m_{xy}^2}{m_{xx} - \sigma_v^2}. \]

Note that \( \hat{\sigma}^2 = m_{xx} - \sigma_v^2 \) can be negative because \( \sigma_v^2 \) is known and \( m_{xx} \) is based upon sample. So we assume that \( \hat{\sigma}^2 > 0 \) and redefine

\[ \hat{\beta}_1 = \frac{m_{xy}}{m_{xx} - \sigma_v^2}; m_{xx} > \sigma_v^2. \]

Similarly, \( \hat{\sigma}_u^2 \) is also assumed to be positive under suitable condition. All the estimators \( \hat{\beta}_1, \hat{\sigma}^2 \) and \( \hat{\sigma}_u^2 \) are the consistent estimators of \( \beta, \sigma^2 \) and \( \sigma_u^2 \) respectively. Note that \( \hat{\beta}_1 \) looks like as if the direct regression estimator of \( \beta_1 \) has been adjusted by \( \sigma_v^2 \) for its inconsistency. So it is termed as adjusted estimator also.
2. $\sigma_u^2$ is known

Suppose $\sigma_u^2$ is known a priori. Then using $m_{xy} = \beta_1 \sigma^2$, we can rewrite

$$m_{yy} = \beta_1^2 \sigma^2 + \sigma_u^2$$

$$= m_{xy} \beta_1 + \sigma_u^2$$

$$\Rightarrow \hat{\beta}_1 = \frac{m_{yy} - \sigma_u^2}{m_{xy}}; \quad m_{yy} > \sigma_u^2$$

$$\hat{\sigma}^2 = \frac{m_{xy}}{\hat{\beta}_1}$$

$$\hat{\sigma}_v^2 = m_{xx} - \hat{\sigma}^2.$$

The estimators $\hat{\beta}_1, \hat{\sigma}^2$ and $\hat{\sigma}_v^2$ are the consistent estimators of $\beta_1, \sigma^2$ and $\sigma_v^2$ respectively. Note that $\hat{\beta}_1$ looks like as if the reverse regression estimator of $\beta_1$ is adjusted by $\sigma_u^2$ for its inconsistency. So it is termed as adjusted estimator also.

3. $\lambda = \frac{\sigma_u^2}{\sigma_v^2}$ is known

Suppose the ratio of the measurement error variances is known, so let

$$\lambda = \frac{\sigma_u^2}{\sigma_v^2}$$

is known.

Consider

$$m_{yy} = \beta_1^2 \sigma^2 + \sigma_u^2$$

$$= \beta_1 m_{xy} + \lambda \sigma_v^2$$  \hspace{1cm} (using (iv))

$$= \beta_1 m_{xy} + \lambda (m_{xx} - \sigma^2)$$  \hspace{1cm} (using (iii))

$$= \beta_1 m_{xy} + \lambda \left(m_{xx} - \frac{m_{xy}}{\hat{\beta}_1}\right)$$  \hspace{1cm} (using iv)

$$\beta_1^2 m_{xy} + \lambda \left(\beta_1 m_{xx} - m_{xy}\right) - \beta_1 m_{yy} = 0 \quad (\beta_1 \neq 0)$$

$$\beta_1^2 m_{xy} + \beta \left(\lambda m_{xx} - m_{yy}\right) - \lambda m_{yy} = 0.$$
Solving this quadratic equation
\[
\hat{\beta}_1 = \frac{(m_{xy} - \lambda m_{xx}) \pm \sqrt{(m_{xy} - \lambda m_{xx})^2 + 4 \lambda m_{xy}^2}}{2m_{xy}}
\]
\[= \frac{U}{2m_{xy}}, \text{ say.}
\]
Since \( m_{xy} = \beta_1 \sigma^2 \) and 
\[
\hat{\sigma}^2 \geq 0
\]
\[\Rightarrow \frac{m_{xy}}{\hat{\beta}_1} \geq 0
\]
\[\Rightarrow \frac{2m_{xy}^2}{U} \geq 0
\]
\[\Rightarrow \text{since } m_{xy}^2 \geq 0, \text{ so } U \text{ must be nonnegative.}
\]

This implies that the positive sign in \( U \) has to be considered and so 
\[
\hat{\beta}_1 = \frac{(m_{xy} - \lambda m_{xx}) + \sqrt{(m_{xy} - \lambda m_{xx})^2 + 4 \lambda m_{xy}^2}}{2m_{xy}}
\]

Other estimates are 
\[
\hat{\sigma}_v^2 = \frac{m_{xy} - 2\hat{\beta}_1 m_{xy} + \hat{\beta}_1^2 s_{xx}}{\lambda + \hat{\beta}_1^2}
\]
\[
\hat{\sigma}^2 = \frac{m_{xy}}{\hat{\beta}_1}
\]

Note that the same estimator \( \hat{\beta}_1 \) of \( \beta_1 \) can be obtained by orthogonal regression. This amounts to transform 
\( x_i \) by \( x_i / \sigma_u \) and \( y_i \) by \( y_i / \sigma_v \) and use the orthogonal regression estimation with transformed variables.

4. Reliability ratio is known

The reliability ratio associated with explanatory variable is defined as the ratio of variances of true and observed values of explanatory variables, so
\[
K_x = \frac{\text{Var}(\hat{x})}{\text{Var}(x)} = \frac{\sigma^2}{\sigma^2 + \sigma_v^2}; \quad 0 \leq K_x \leq 1
\]
is the reliability ratio. Note that \( K_x = 1 \), when \( \sigma_v^2 = 0 \) which means that there is no measurement error in the explanatory variable and \( K_x = 0 \), means \( \sigma^2 = 0 \) which means explanatory variable is fixed. Higher value
of \( K_x \) is obtained when \( \sigma^2_v \) is small, i.e., the impact of measurement errors is small. The reliability ratio is a popular measure in psychometrics.

Let \( K_x \) be known a priori. Then

\[
m_{xx} = \sigma^2 + \sigma^2_v
\]

\[
m_{xy} = \beta_1 \sigma^2
\]

\[
\Rightarrow \frac{m_{xy}}{m_{xx}} = \frac{\beta_1 \sigma^2}{\sigma^2 + \sigma^2_v} = \beta_1 K_x
\]

\[
\Rightarrow \hat{\beta}_1 = \frac{m_{xy}}{K_x m_{xx}}
\]

\[
\sigma^2 = \frac{m_{xy}}{\beta_1}
\]

\[
\Rightarrow \hat{\sigma}^2 = K_x m_{xx}
\]

\[
m_{xx} = \sigma^2 + \sigma^2_v
\]

\[
\Rightarrow \hat{\sigma}^2_v = (1 - K_x) m_{xx}.
\]

Note that \( \hat{\beta}_1 = K_x^{-1} b \)

where \( b \) is the ordinary least squares estimator \( b = \frac{m_{xy}}{m_{xx}} \).

5. \( \beta_0 \) is known

Suppose \( \beta_0 \) is known a priori and \( E(x) \neq 0 \). Then

\[
\bar{y} = \beta_0 + \beta_1 \mu
\]

\[
\Rightarrow \hat{\beta}_1 = \frac{\bar{y} - \beta_0}{\mu} = \frac{\bar{y} - \beta_0}{\bar{x}}
\]

\[
\hat{\sigma}^2 = \frac{m_{xy}}{\hat{\beta}_1}
\]

\[
\hat{\sigma}^2_u = m_{xy} - \hat{\beta}_1 m_{xy}
\]

\[
\hat{\sigma}^2_v = m_{xx} - \frac{m_{xy}^2}{\hat{\beta}_1}.
\]
6. Both $\sigma_u^2$ and $\sigma_v^2$ are known

This case leads to over-identification in the sense that the number of parameters to be estimated are smaller than the number of structural relationships binding them. So no unique solutions are obtained in this case.

**Note:** In each of the cases 1-6, note that the form of the estimate depends on the type of available information which is needed for the consistent estimator of the parameters. Such information can be available from various sources, e.g., long association of the experimenter with the experiment, similar type of studies conducted in the part, some extraneous source etc.

**Estimation of parameters in function form:**

In the functional form of the measurement error model, $\tilde{x}_i$’s are assumed to be fixed. This assumption is unrealistic in the sense that when $\tilde{x}_i$’s are unobservable and unknown, it is difficult to know if they are fixed or not. This can not be ensured even in repeated sampling that the same value is repeated. All that can be said in this case is that the information in this case is conditional upon $\tilde{x}_i$’s. So assume that $\tilde{x}_i$’s are conditionally known. So the model is

$$
\tilde{y}_i = \beta_0 + \beta_i \tilde{x}_i \\
x_i = \tilde{x}_i + v_i \\
y_i = \tilde{y}_i + u_i
$$

then

$$
\begin{pmatrix}
  y_i \\
  x_i
\end{pmatrix} \sim N\left( \begin{pmatrix}
  \beta_0 + \beta_i \tilde{x}_i \\
  \tilde{x}_i
\end{pmatrix}, \begin{pmatrix}
  \sigma_u^2 & 0 \\
  0 & \sigma_v^2
\end{pmatrix} \right).
$$

The likelihood function is

$$
L = \left( \frac{1}{2\pi\sigma_u^2} \right)^{\frac{n}{2}} \exp \left[ -\frac{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_i x_i)^2}{2\sigma_u^2} \right] \left( \frac{1}{2\pi\sigma_v^2} \right)^{\frac{n}{2}} \exp \left[ -\frac{\sum (x_i - \tilde{x}_i)^2}{2\sigma_v^2} \right].
$$

The log-likelihood is

$$
L^* = \ln L = \text{constant} - \frac{n}{2} \ln \sigma_u^2 - \frac{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_i \tilde{x}_i)^2}{2\sigma_u^2} - \frac{n}{2} \ln \sigma_v^2 - \frac{1}{2\sigma_v^2} \sum_{i=1}^{n} (x_i - \tilde{x}_i)^2.
$$
The normal equations are obtained by partially differentiating $L^*$ and equating to zero as

\[(I) \quad \frac{\partial L^*}{\partial \beta_0} = 0 \Rightarrow \frac{1}{\sigma_u^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_i \tilde{x}_i) = 0 \]

\[(II) \quad \frac{\partial L^*}{\partial \beta_i} = 0 \Rightarrow \frac{1}{\sigma_v^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_i \tilde{x}_i) \tilde{x}_i = 0 \]

\[(III) \quad \frac{\partial L^*}{\partial \sigma_u^2} = 0 \Rightarrow -\frac{n}{2\sigma_u^2} + \frac{1}{2\sigma_v^4} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_i \tilde{x}_i)^2 = 0 \]

\[(IV) \quad \frac{\partial L^*}{\partial \sigma_v^2} = 0 \Rightarrow -\frac{n}{2\sigma_v^2} + \frac{1}{2\sigma_u^4} \sum_{i=1}^{n} (x_i - \tilde{x}_i)^2 = 0 \]

\[(V) \quad \frac{\partial L^*}{\partial \tilde{x}_i} = 0 \Rightarrow \frac{\beta}{\sigma_u^2} (y_i - \beta_0 - \beta_i \tilde{x}_i) + \frac{1}{\sigma_v^2} (x_i - \tilde{x}_i) = 0. \]

Squaring and summing equation (V), we get

\[\sum_{i} \left[ \frac{\beta}{\sigma_u^2} (y_i - \beta_0 - \beta_i \tilde{x}_i) \right]^2 = \sum_{i} \left[ -\frac{1}{\sigma_v^2} (x_i - \tilde{x}_i) \right]^2 \]

or \[\frac{\beta^2}{\sigma_u^2} \sum_{i} (y_i - \beta_0 - \beta_i \tilde{x}_i)^2 = \frac{1}{\sigma_v^4} \sum_{i} (x_i - \tilde{x}_i)^2. \]

Using the left hand side of equation (III) and right hand side of equation (IV), we get

\[\frac{n\beta^2}{\sigma_u^2} = \frac{n}{\sigma_v^2} \]

\[\Rightarrow \beta_1 = \frac{\sigma_u}{\sigma_v} \]

which is unacceptable because $\beta$ can be negative also. In the present case, as $\sigma_u > 0$ and $\sigma_v > 0$, so $\beta$ will always be positive. Thus the maximum likelihood breaks down because of insufficient information in the model. Increasing the sample size $n$ does not solve the purpose. If the restrictions like $\sigma_u^2$ known, $\sigma_v^2$ known or $\sigma_u^2 / \sigma_v^2$ known are incorporated, then the maximum likelihood estimation is similar to as in the case of structural form and the similar estimates may be obtained. For example, if $\lambda = \sigma_u^2 / \sigma_v^2$ is known, then substitute it in the likelihood function and maximize it. The same solution as in the case of structural form are obtained.
In any regression modeling, generally an equation is considered to represent a relationship describing a phenomenon. Many situations involve a set of relationships which explain the behaviour of certain variables. For example, in analyzing the market conditions for a particular commodity, there can be a demand equation and a supply equation which explain the price and quantity of commodity exchanged in the market at market equilibrium. So there are two equations to explain the whole phenomenon - one for demand and another for supply. In such cases, it is not necessary that all the variables should appear in all the equations. So estimation of parameters under this type of situation has those features that are not present when a model involves only a single relationship. In particular, when a relationship is a part of a system, then some explanatory variables are stochastic and are correlated with the disturbances. So the basic assumption of a linear regression model that the explanatory variable and disturbance are uncorrelated or explanatory variables are fixed is violated and consequently ordinary least squares estimator becomes inconsistent.

Similar to the classification of variables as explanatory variable and study variable in linear regression model, the variables in simultaneous equation models are classified as endogenous variables and exogenous variables.

**Endogenous variables (Jointly determined variables)**

The variables which are explained by the functioning of system and values of which are determined by the simultaneous interaction of the relations in the model are endogenous variables or jointly determined variables.

**Exogenous variables (Predetermined variables)**

The variables that contribute to provide explanations for the endogenous variables and values of which are determined from outside the model are exogenous variables or predetermined variables.

Exogenous variables help in explaining the variations in endogenous variables. It is customary to include past values of endogenous variables in the predetermined group. Since exogenous variables are predetermined, so they are independent of disturbance term in the model. They satisfy those assumptions.
which explanatory variables satisfy in the usual regression model. Exogenous variables influence the endogenous variables but are not themselves influenced by them. One variable which is endogenous for one model can be exogenous variable for the other model.

Note that in linear regression model, the explanatory variables influence study variable but not vice versa. So relationship is one sided.

The classification of variables as endogenous and exogenous is important because a necessary condition for uniquely estimating all the parameters is that the number of endogenous variables is equal to the number of independent equations in the system. Moreover, the main distinction of predetermined variable in estimation of parameters is that they are uncorrelated with disturbance term in the equations in which they appear.

**Simultaneous equation systems:**

A model constitutes a system of simultaneous equations if all the relationships involved are needed for determining the value of at least one of the endogenous variables included in the model. This implies that at least one of the relationships includes more than one endogenous variable.

**Example 1:**

Now we consider the following example in detail and introduce various concepts and terminologies used in describing the simultaneous equations models.

Consider a situation of an ideal market where transaction of only one commodity, say wheat, takes place. Assume that the number of buyers and sellers is large so that the market is perfect competitive market. It is also assumed that the amount of wheat that comes into market in a day is completely sold out on the same day. No seller takes it back. Now we develop a model for such mechanism.

Let

\[ d_t \] denotes the demand of the commodity, say wheat, at time \( t \),

\[ s_t \] denotes the supply of the commodity, say wheat, at time \( t \), and

\[ q_t \] denotes the quantity of the commodity, say wheat, transacted at time \( t \).
By economic theory about ideal market, we have the following condition:

\[ d_t = s_t, \quad t = 1, 2, \ldots, n. \]

Observe that

- demand of wheat depends on
  - price of wheat \((p_t)\) at time \(t\).
  - income of buyer \((i_t)\) at time \(t\).
- supply of wheat depends on
  - price of wheat \((p_t)\) at time \(t\).
  - rainfall \((r_t)\) at time \(t\).

From market conditions, we have

\[ q_t = d_t = s_t. \]

Demand, supply and price are determined from each other.

Note that

- income can influence demand and supply but demand and supply cannot influence the income.
- supply is influenced by rainfall but rainfall is not influenced by the supply of wheat.

Our aim is to study the behaviour of \(s_t, p_t\) and \(r_t\) which are determined by the simultaneous equation model.

Since endogenous variables are influenced by exogenous variables but not vice versa, so

- \(s_t, p_t\) and \(r_t\) are endogenous variables.
- \(i_t\) and \(r_t\) are exogenous variables.

Now consider an additional variable for the model as lagged value of price \(p_t\), denoted as \(p_{t-1}\). In a market, generally the price of the commodity depends on the price of the commodity on previous day. If the price of commodity today is less than the previous day, then buyer would like to buy more. For seller also, today’s price of commodity depends on previous day’s price and based on which he decides the quantity of commodity (wheat) to be brought in the market.
So the lagged price affects the demand and supply equations both. Updating both the models, we can now write that

- demand depends on \( p_t, i_t \) and \( p_{t-1} \).
- supply depends on \( p_t, r_t \) and \( p_{t-1} \).

Note that the lagged variables are considered as exogenous variable. The updated list of endogenous and exogenous variables is as follows:

- Endogenous variables: \( p_t, d_t, s_t \)
- Exogenous variables: \( p_{t-1}, i_t, r_t \).

The mechanism of market is now described by the following set of equations.

- demand \( d_t = \alpha_t + \beta_1 p_t + \epsilon_1 \)
- supply \( s_t = \alpha_2 + \beta_2 p_t + \epsilon_2 \)
- equilibrium condition \( d_t = s_t = q_t \)

where \( \alpha \)'s denote the intercept terms, \( \beta \)'s denote the regression coefficients and \( \epsilon \)'s denote the disturbance terms.

These equations are called structural equations. The error terms \( \epsilon_1 \) and \( \epsilon_2 \) are called structural disturbances. The coefficients \( \alpha_1, \alpha_2, \beta_1 \) and \( \beta_2 \) are called the structural coefficients.

The system of equations is called the structural form of the model.

Since \( q_t = d_t = s_t \), so the demand and supply equations can be expressed as

\[
q_t = \alpha_t + \beta_1 p_t + \epsilon_1 \quad \text{(I)}
\]

\[
q_t = \alpha_2 + \beta_2 p_t + \epsilon_2 \quad \text{(II)}
\]

So there are only two structural relationships. The price is determined by the mechanism of market and not by the buyer or supplier. Thus \( q_t \) and \( p_t \) are the endogenous variables. Without loss of generality, we can assume that the variables associated with \( \alpha_1 \) and \( \alpha_2 \) are \( X_1 \) and \( X_2 \) respectively such that \( X_1 = 1 \) and \( X_2 = 1 \). So \( X_1 = 1 \) and \( X_2 = 1 \) are predetermined and so they can be regarded as exogenous variables.
From statistical point of view, we would like to write the model in such a form so that the OLS can be directly applied. So writing equations (I) and (II) as

\[ \alpha_i + \beta_i p_i + \varepsilon_{1i} = \alpha_2 + \beta_2 p_i + \varepsilon_{2i} \]

or

\[ p_i = \frac{\alpha_1 - \alpha_2}{\beta_2 - \beta_1} + \frac{\varepsilon_{1i} - \varepsilon_{2i}}{\beta_2 - \beta_1} \]

or

\[ = \pi_{1i} + \nu_{1i} \quad \text{(III)} \]

\[ q_i = \frac{\beta_2 \alpha_1 - \beta_1 \alpha_2}{\beta_2 - \beta_1} + \frac{\beta_2 \varepsilon_{1i} - \beta_1 \varepsilon_{2i}}{\beta_2 - \beta_1} \]

\[ = \pi_{2i} + \nu_{2i} \quad \text{(IV)} \]

where

\[ \pi_{1i} = \frac{\alpha_1 - \alpha_2}{\beta_2 - \beta_1}, \quad \pi_{2i} = \frac{\beta_2 \alpha_1 - \beta_1 \alpha_2}{\beta_2 - \beta_1} \]

\[ \nu_{1i} = \frac{\varepsilon_{1i} - \varepsilon_{2i}}{\beta_2 - \beta_1}, \quad \nu_{2i} = \frac{\beta_2 \varepsilon_{1i} - \beta_1 \varepsilon_{2i}}{\beta_2 - \beta_1}. \]

Each endogenous variable is expressed as a function of exogenous variable. Note that the exogenous variable \( l \) (from \( X_1 = l \), or \( X_2 = l \)) is not clearly identifiable.

The equations (III) and (IV) are called the **reduced form relationships** and in general, called as the **reduced form of the model**.

The coefficients \( \pi_{1i} \) and \( \pi_{2i} \) are called **reduced form coefficients** and errors \( \nu_{1i} \) and \( \nu_{2i} \) are called the **reduced form disturbances**. The reduced from essentially express every endogenous variable as a function of exogenous variable. This presents a clear relationship between reduced form coefficients and structural coefficients as well as between structural disturbances and reduced form disturbances. The reduced form is ready for the application of OLS technique. The reduced form of the model satisfies all the assumptions needed for the application of OLS.

Suppose we apply OLS technique to equations (III) and (IV) and obtained the OLS estimates of \( \pi_{1i} \) and \( \pi_{12} \) as \( \hat{\pi}_{1i} \) and \( \hat{\pi}_{12} \) respectively which are given by

\[ \hat{\pi}_{1i} = \frac{\alpha_1 - \alpha_2}{\beta_2 - \beta_1} \]

\[ \hat{\pi}_{2i} = \frac{\beta_2 \alpha_1 - \beta_1 \alpha_2}{\beta_2 - \beta_1}. \]
Note that $\hat{\alpha}_{11}$ and $\hat{\alpha}_{21}$ are the numerical values of the estimates. So now there are two equations and four unknown parameters $\alpha_1, \alpha_2, \beta_1$ and $\beta_2$. So it is not possible to derive the unique estimates of parameters of the model by applying OLS technique to reduced form. This is known as problem of identifications.

By this example, the following have been described up to now:

- Structural form relationship.
- Reduced form relationship.
- Need for reducing the structural form into reduced form.
- Reason for the problem of identification.

Now we describe the problem of identification in more detail.

**The identification problem:**

Consider the model in earlier Example 1 which describes the behaviour of a perfectly competitive market in which only one commodity, say wheat, is transacted. The models describing the behaviour of consumer and supplier are prescribed by demand and supply conditions given as

- Demand: $d_t = \alpha_1 + \beta_1 p_t + \varepsilon_{1t}$, $t = 1, 2, ..., n$
- Supply: $s_t = \alpha_2 + \beta_2 p_t + \varepsilon_{2t}$
- Equilibrium condition: $d_t = s_t$.

If quantity $q_t$ is transacted at time $t$ then

$$d_t = s_t = q_t.$$  

So we have two structural equations model in two endogenous variables ($q_t$ and $p_t$) and one exogenous variable (value is 1 given by $X_1 = 1, X_2 = 1$). The set of three equations is reduced to a set of two equations as follows:

- Demand: $q_t = \alpha_1 + \beta_1 p_t + \varepsilon_{1t}$  \hspace{1cm} (1)
- Supply: $q_t = \alpha_2 + \beta_2 p_t + \varepsilon_{2t}$  \hspace{1cm} (2)

Before analysis, we would like to check whether it is possible to estimate the parameters $\alpha_1, \alpha_2, \beta_1$ and $\beta_2$ or not.
Multiplying equations (1) by $\lambda$ and (2) by $(1-\lambda)$ and then adding them together gives

$$
\lambda q_t + (1-\lambda)q_t = \left[ \lambda \alpha_1 + (1-\lambda)\alpha_2 \right] + \left[ \lambda \beta_1 + (1-\lambda)\beta_2 \right] p_t + \left[ \lambda \varepsilon_{1t} + (1-\lambda)\varepsilon_{2t} \right]
$$

or

$$
q_t = \alpha + \beta p_t + \varepsilon_t
$$

(3)

where $\alpha = \lambda \alpha_1 + (1-\lambda)\alpha_2$, $\beta = \lambda \beta_1 + (1-\lambda)\beta_2$, $\varepsilon_t = \lambda \varepsilon_{1t} + (1-\lambda)\varepsilon_{2t}$, and $\lambda$ is any scalar lying between 0 and 1.

Comparing equation (3) with equations (1) and (2), we notice that they have same form. So it is difficult to say that which is supply equation and which is demand equation. To find this, let equation (3) be demand equation. Then there is no way to identify the true demand equation (1) and pretended demand equation (3).

Similar exercise can be done for supply equation and we find that there is no way to identify the true supply equation (2) and pretended supply equation (3).

Suppose we apply OLS technique to these models. Applying OLS to equation (1) yields

$$
\hat{\beta}_1 = \frac{\sum_{t=1}^{n} (p_t - \bar{p})(q_t - \bar{q})}{\sum_{t=1}^{n} (p_t - \bar{p})^2} = 0.6, \text{ say}
$$

where $\bar{p} = \frac{1}{n} \sum_{t=1}^{n} p_t$, $\bar{q} = \frac{1}{n} \sum_{t=1}^{n} q_t$.

Applying OLS to equation (3) yields

$$
\hat{\beta} = \frac{\sum_{t=1}^{n} (p_t - \bar{p})(q_t - \bar{q})}{\sum_{t=1}^{n} (p_t - \bar{p})^2} = 0.6.
$$

Note that $\hat{\beta}_1$ and $\hat{\beta}$ have same analytical expressions, so they will also have same numerical values, say 0.6.

Looking at the value 0.6, it is difficult to say that the value 0.6 determines equation (1) or (3).

Applying OLS to equation (3) yields

$$
\hat{\beta}_2 = \frac{\sum_{t=1}^{n} (p_t - \bar{p})(q_t - \bar{q})}{\sum_{t=1}^{n} (p_t - \bar{p})^2} = 0.6
$$

Econometrics  |  Chapter 17  |  Simultaneous Equations Models  |  Shalabh, IIT Kanpur
because $\hat{\beta}_2$ has same analytical expression as of $\hat{\beta}_1$ and $\hat{\beta}$, so

$$\hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3.$$ 

Thus it is difficult to decide and identify whether $\hat{\beta}_1$ is determined by the value 0.6 or $\hat{\beta}_2$ is determined by the value 0.6. Increasing the number of observations also does not help in the identification of these equations. So we are not able to identify the parameters. So we take the help of economic theory to identify the parameters. 

The economic theory suggests that when price increases then supply increases but demand decreases. So the plot will look like

![Plot showing supply and demand](image)

and this implies $\beta_1 < 0$ and $\beta_2 > 0$. Thus since $0.6 > 0$, so we can say that the value 0.6 represents $\hat{\beta}_2 > 0$ and so $\hat{\beta}_2 = 0.6$. But one can always choose a value of $\lambda$ such that pretended equation does not violate the sign of coefficients, say $\beta > 0$. So it again becomes difficult to see whether equation (3) represents supply equation (2) or not. So none of the parameters is identifiable.

Now we obtain the reduced form of the model as

$$p_t = \frac{\alpha_1 - \alpha_2 + \epsilon_{1t} - \epsilon_{2t}}{\beta_2 - \beta_1}$$

or $p_t = \pi_{1t} + \nu_{1t}$

(4)

$$q_t = \frac{\alpha_1 \beta_2 - \alpha_2 \beta_1 + \beta_2 \epsilon_{1t} - \beta_1 \epsilon_{2t}}{\beta_2 - \beta_1}$$

or $q_t = \pi_{2t} + \nu_{2t}$.

(5)
Applying OLS to equations (4) and (5) and obtain OLSEs \( \hat{\alpha}_{11} \) and \( \hat{\pi}_{21} \) which are given by

\[
\hat{\alpha}_{11} = \frac{\alpha_1 - \alpha_2}{\beta_2 - \beta_1},
\]
\[
\hat{\pi}_{21} = \frac{\alpha_1 \beta_2 - \alpha_2 \beta_1}{\beta_2 - \beta_1}.
\]

There are two equations and four unknowns. So the unique estimates of the parameters \( \alpha_1, \alpha_2, \beta_1 \) and \( \beta_2 \) cannot be obtained. Thus the equations (1) and (2) cannot be identified. Thus the model is not identifiable and estimation of parameters is not possible.

Suppose a new exogenous variable income \( i_t \) is introduced in the model which represents the income of buyer at time \( t \). Note that the demand of commodity is influenced by income. On the other hand, the supply of commodity is not influenced by the income, so this variable is introduced only in the demand equation. The structural equations (1) and (2) now become

Demand: \( q_t = \alpha_1 + \beta_1 p_t + \gamma_1 i_t + \epsilon_{1t} \) \hspace{1cm} (6)
Supply: \( q_t = \alpha_2 + \beta_2 p_t + \epsilon_{2t} \) \hspace{1cm} (7)

where \( \gamma_1 \) is the structural coefficient associate with income. The pretended equation is obtained by multiplying the equations (6) and (7) by \( \lambda \) and \( (1-\lambda) \), respectively, and then adding them together. This is obtained as follows:

\[
\lambda q_t + (1-\lambda)q_t = \lfloor \lambda \alpha_1 + (1-\lambda)\alpha_2 \rfloor + \lfloor \lambda \beta_1 + (1-\lambda)\beta_2 \rfloor p_t + \lambda \gamma_1 i_t + \lfloor \lambda \epsilon_{1t} + (1-\lambda)\epsilon_{2t} \rfloor
\]

or \( q_t = \alpha + \beta p_t + \gamma i_t + \epsilon_t \) \hspace{1cm} (8)

where \( \alpha = \lambda \alpha_1 + (1-\lambda)\beta_2 \), \( \beta = \lambda \beta_1 + (1-\lambda)\beta_2 \), \( \gamma = \lambda \gamma_1 \), \( \epsilon_t = \lambda \epsilon_{1t} + (1-\lambda)\epsilon_{2t} \), \( 0 \leq \lambda \leq 1 \) is a scalar.

Suppose now if we claim that equation (8) is true demand equation because it contains \( p_t \) and \( i_t \) which influence the demand. But we note that it is difficult to decide that between the two equations (6) or (8), which one is the true demand equation.

Suppose now if we claim that equation (8) is the true supply equation. This claim is wrong because income does not affect the supply. So equation (6) is the supply equation.

Thus supply equation is now identifiable but demand equation is not identifiable. Such situation is termed as **partial identification**.
Now we find the reduced form of structural equations (6) and (7). This is achieved by first solving equation (6) for $p_t$ and substituting it in equation (7) to obtain an equation in $q_t$. Such an exercise yields the reduced form equations of the following form:

$$p_t = \pi_{11} + \pi_{12} i_t + \nu_t,$$

$$q_t = \pi_{21} + \pi_{22} i_t + \nu_{2t},$$

Applying OLS to equations (9) and (10), we get OLSEs $\hat{\pi}_{11}, \hat{\pi}_{22}, \hat{\pi}_{12}, \hat{\pi}_{21}$. Now we have four equations ((6),(7),(9),(10)) and there are five unknowns $(\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1)$. So the parameters are not determined uniquely. Thus the model is not identifiable.

However, here

$$\beta_2 = \frac{\pi_{22}}{\pi_{11}},$$

$$\alpha_2 = \pi_{21} - \pi_{22} \frac{\pi_{11}}{\pi_{22}}.$$

If $\hat{\pi}_{11}, \hat{\pi}_{22}, \hat{\pi}_{12}$ and $\hat{\pi}_{21}$ are available, then $\hat{\alpha}_2$ and $\hat{\beta}_2$ can be obtained by substituting $\hat{\pi}'s$ in place of $\pi'$s. So $\alpha_2$ and $\beta_2$ are determined uniquely which are the parameters of supply equation. So supply equation is identified but demand equation is still not identifiable.

Now, as done earlier, we introduce another exogenous variable—rainfall, denoted as $r_t$ which denotes the amount of rainfall at time $t$. The rainfall influences the supply because better rainfall produces better yield of wheat. On the other hand, the demand of wheat is not influenced by the rainfall. So the updated set of structural equations is

Demand: $q_t = \alpha_1 + \beta_1 p_t + \gamma_1 i_t + \epsilon_{1t}$

Supply: $q_t = \alpha_2 + \beta_2 p_t + \delta_2 r_t + \epsilon_{2t}.$

The pretended equation is obtained by adding together the equations obtained after multiplying equation (11) by $\lambda$ and equation (12) by $(1-\lambda)$ as follows:

$$\lambda q_t + (1-\lambda) q_t = [\lambda \alpha_1 + (1-\lambda) \alpha_2] + [\lambda \beta_1 + (1-\lambda) \beta_2] p_t + \lambda \gamma_1 i_t + (1-\lambda) \delta_2 r_t + [\lambda \epsilon_{1t} + (1-\lambda) \epsilon_{2t}]$$

or $q_t = \alpha + \beta p_t + \gamma i_t + \delta r_t + \epsilon_t$

where $\alpha = \lambda \alpha_1 + (1-\lambda) \alpha_2, \beta = \lambda \beta_1 + (1-\lambda) \beta_2, \gamma = \lambda \gamma_1, \delta = (1-\lambda) \delta_2, \epsilon_t = \lambda \epsilon_{1t} + (1-\lambda) \epsilon_{2t}$ and $0 \leq \lambda \leq 1$ is a scalar.
Now we claim that equation (13) is a demand equation. The demand does not depend on rainfall. So unless \( \lambda = 1 \) so that \( r_t \) is absent in the model, the equation (13) cannot be a demand equation. Thus equation (11) is a demand equation. So demand equation is identified.

Now we claim that equation (13) is supply equation. The supply is not influenced by the income of buyer, so (13) cannot be a supply equation. Thus equation (12) is the supply equation. So now the supply equation is also identified.

The reduced form model from structural equations (11) and (12) can be obtained which have the following forms:

\[ p_t = \pi_{11} + \pi_{12}i_t + \pi_{13}r_t + v_{1t}, \quad (14) \]
\[ q_t = \pi_{21} + \pi_{22}i_t + \pi_{23}r_t + v_{2t}, \quad (15) \]

Application of OLS technique to equations (14) and (15) yields the OLSEs \( \hat{\pi}_{11}, \hat{\pi}_{12}, \hat{\pi}_{13}, \hat{\pi}_{21}, \hat{\pi}_{22}, \) and \( \hat{\pi}_{23} \). So now there are six such equations and six unknowns \( \alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \) and \( \delta_2. \) So all the estimates are uniquely determined. Thus the equations (11) and (12) are exactly identifiable.

Finally, we introduce a lagged endogenous variable \( p_{t-1} \) which denotes the price of commodity on previous day. Since only the supply of wheat is affected by the price on the previous day, so it is introduced in the supply equation only as

\[
\begin{align*}
\text{Demand:} & \quad q_t = \alpha_1 + \beta_1 p_t + \gamma_1 i_t + \varepsilon_{1t} \quad (16) \\
\text{Supply:} & \quad q_t = \alpha_2 + \beta_2 p_t + \delta_2 r_t + \theta_2 p_{t-1} + \varepsilon_{2t} \quad (17)
\end{align*}
\]

where \( \theta_2 \) is the structural coefficient associated with \( p_{t-1} \).

The pretended equation is obtained by first multiplying equations (16) by \( \lambda \) and (17) by \( (1 - \lambda) \) and then adding them together as follows:

\[
\begin{align*}
\lambda q_t + (1 - \lambda) q_t = & \quad \alpha + \beta p_t + \gamma i_t + \delta r_t + (1 - \lambda) \theta_2 p_{t-1} + \varepsilon_t \\
\text{or} \quad q_t = & \quad \alpha + \beta p_t + \gamma i_t + \delta r_t + \theta_2 p_{t-1} + \varepsilon_t \quad (18)
\end{align*}
\]

Now we claim that equation (18) represents the demand equation. Since rainfall and lagged price do not affect the demand, so equation (18) cannot be demand equation. Thus equation (16) is a demand equation and the demand equation is identified.
Now finally, we claim that equation (18) is the supply equation. Since income does not affect supply, so equation (18) cannot be a supply equation. Thus equation (17) is supply equation and the supply equation is identified.

The reduced from equations from equations (16) and (17) can be obtained as of the following form:

\[
\begin{align*}
    p_t &= \pi_{11}t + \pi_{12}t + \pi_{13}r_t + \pi_{14}p_{t-1} + v_{1t} \quad (19) \\
    q_t &= \pi_{21}t + \pi_{22}t + \pi_{23}r_t + r_{24}p_{t-1} + v_{2t}. \quad (20)
\end{align*}
\]

Applying OLS technique to equations (19) and (20) gives the OLSEs as \( \hat{\pi}_{11}, \hat{\pi}_{12}, \hat{\pi}_{13}, \hat{\pi}_{14}, \hat{\pi}_{21}, \hat{\pi}_{22}, \hat{\pi}_{23} \) and \( \hat{\pi}_{24} \). So there are eight equations in seven parameters \( \alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \delta_2 \) and \( \theta_2 \). So unique estimates of all the parameters are not available. In fact, in this case, the supply equation (17) is identifiable and demand equation (16) is **overly identified** (in terms of multiple solutions).

The whole analysis in this example can be classified into three categories –

(i) **Under identifiable case**

The estimation of parameters is not at all possible in this case. No enough estimates are available for structural parameters.

(2) **Exactly identifiable case**:

The estimation of parameters is possible in this case. The OLSE of reduced form coefficients leads to unique estimates of structural coefficients.

(3) **Over identifiable case**:

The estimation of parameters in this case is possible. The OLSE of reduced form coefficients leads to multiple estimates of structural coefficients.
Analysis:

Suppose there are $G$ jointly dependent (endogenous) variables $y_1, y_2, ..., y_G$ and $K$ predetermined (exogenous) variables $x_1, x_2, ..., x_K$. Let there are $n$ observations available on each of the variable and there are $G$ structural equations connecting both the variables which describe the complete model as follows:

\[
\begin{align*}
\beta_{11}y_{1t} + \beta_{12}y_{2t} + \cdots + \beta_{1G}y_{Gt} + \gamma_{11}x_{1t} + \gamma_{12}x_{2t} + \cdots + \gamma_{1K}x_{Kt} &= \varepsilon_{1t} \\
\beta_{21}y_{1t} + \beta_{22}y_{2t} + \cdots + \beta_{2G}y_{Gt} + \gamma_{21}x_{1t} + \gamma_{22}x_{2t} + \cdots + \gamma_{2K}x_{Kt} &= \varepsilon_{2t} \\
\vdots \\
\beta_{G1}y_{1t} + \beta_{G2}y_{2t} + \cdots + \beta_{GG}y_{Gt} + \gamma_{G1}x_{1t} + \gamma_{G2}x_{2t} + \cdots + \gamma_{GK}x_{Kt} &= \varepsilon_{Gt}.
\end{align*}
\]

These equations can be expressed in matrix form as

\[
\begin{bmatrix}
\beta_{11} & \beta_{12} & \cdots & \beta_{1G} \\
\beta_{21} & \beta_{22} & \cdots & \beta_{2G} \\
\vdots & \vdots & \ddots & \vdots \\
\beta_{G1} & \beta_{G2} & \cdots & \beta_{GG}
\end{bmatrix}
\begin{bmatrix}
y_{1t} \\
y_{2t} \\
\vdots \\
y_{Gt}
\end{bmatrix}
+
\begin{bmatrix}
\gamma_{11} & \gamma_{12} & \cdots & \gamma_{1K} \\
\gamma_{21} & \gamma_{22} & \cdots & \gamma_{2K} \\
\vdots & \vdots & \ddots & \vdots \\
\gamma_{G1} & \gamma_{G2} & \cdots & \gamma_{KK}
\end{bmatrix}
\begin{bmatrix}
x_{1t} \\
x_{2t} \\
\vdots \\
x_{Kt}
\end{bmatrix}
= \begin{bmatrix}
\varepsilon_{1t} \\
\varepsilon_{2t} \\
\vdots \\
\varepsilon_{Gt}
\end{bmatrix}
\]

or

\[
S: By_t + \Gamma x_t = \varepsilon_t, \quad t = 1, 2, ..., n
\]

where $B$ is a $G \times G$ matrix of unknown coefficients of predetermined variables, $y_t$ is a $(n \times 1)$ vector of observations on $G$ jointly dependent variables, $\Gamma$ is $(G \times K)$ matrix of structural coefficients, $x_t$ is $(K \times 1)$ vector of observations on $K$ predetermined variables and $\varepsilon_t$ is $(G \times 1)$ vector of structural disturbances. The structural form $S$ describes the functioning of model at time $t$.

Assuming $B$ is nonsingular, premultiplying the structural form equations by $B^{-1}$, we get

\[
B^{-1}By_t + B^{-1}\Gamma x_t = B^{-1}\varepsilon_t
\]

or $y_t = x_t + v_t, \quad t = 1, 2, ..., n$.

This is the reduced form equation of the model where $\pi = B^{-1}\Gamma$ is the matrix of reduced form coefficients and $v_t = B^{-1}\varepsilon_t$ is the reduced form disturbance vectors.

If $B$ is singular, then one or more structural relations would be a linear combination of other structural relations. If $B$ is non-singular, such identities are eliminated.
The structural form relationship describes the interaction taking place inside the model. In reduced form relationship, the jointly dependent (endogenous) variables are expressed as linear combination of predetermined (exogenous) variables. This is the difference between structural and reduced form relationships.

Assume that \( \varepsilon_t \)'s are identically and independently distributed following \( N(0, \Sigma) \) and \( V_t \)'s are identically and independently distributed following \( N(0, \Omega) \) where \( \Omega = B^{-1} \Sigma B^{-1} \) with

\[
E(\varepsilon_t) = 0, \quad E(\varepsilon_t, \varepsilon_t') = \Sigma, \quad E(\varepsilon_t, \varepsilon_t') = 0 \text{ for all } t \neq t'.
\]
\[
E(\varepsilon_t) = 0, \quad E(V_t, V_t') = \Omega, \quad E(V_t, V_t') = 0 \text{ for all } t \neq t'.
\]

The joint probability density function of \( y_t \) given \( x_t \) is

\[
p(y_t \mid x_t) = p(v_t) \frac{\partial \varepsilon_t}{\partial v_t} = p(\varepsilon_t) \left| \text{det}(B) \right|
\]

where \( \frac{\partial \varepsilon_t}{\partial v_t} \) is the related Jacobian of transformation and \( \left| \text{det}(B) \right| \) is the absolute value of determinant of \( B \).

The likelihood function is

\[
L = p(y_1, y_2, ..., y_n \mid x_1, x_2, ..., x_n) = \prod_{t=1}^{n} p(y_t \mid x_t) = \left| \text{det}(B) \right|^n \prod_{t=1}^{n} p(\varepsilon_t).
\]

Applying a nonsingular linear transformation on structural equation \( S \) with a nonsingular matrix \( D \), we get

\[
DBy_t + D\Gamma x_t = D\varepsilon_t
\]

or \( S^*: B^* y_t + \Gamma^* x_t = \varepsilon_t^*, \quad t = 1, 2, ..., n \)

where \( B^* = DB, \Gamma^* = D\Gamma, \varepsilon_t^* = D\varepsilon_t \) and structural model \( S^* \) describes the functioning of this model at time \( t \). Now find \( p(y_t \mid x_t) \) with \( S^* \) as follows:

\[
p(y_t \mid x_t) = p(\varepsilon_t^*) \left| \frac{\partial \varepsilon_t^*}{\partial v_t} \right| = p(\varepsilon_t^*) \left| \text{det}(B^*) \right|.
\]
Also

\[ \varepsilon_t^* = D \varepsilon_t \]

\[ \Rightarrow \frac{\partial \varepsilon_t^*}{\partial \varepsilon_t} = D. \]

Thus

\[ p(y_t|x_t) = p(\varepsilon_t^*) \left| \frac{\partial \varepsilon_t^*}{\partial \varepsilon_t} \right| \left| \det(B^*) \right| \]

\[ = p(\varepsilon_t) \left| \det(D^{-1}) \right| \left| \det(DB) \right| \]

\[ = p(\varepsilon_t) \left| \det(D^{-1}) \right| \left| \det(D) \right| \left| \det(B) \right| \]

\[ = p(\varepsilon_t) \left| \det(B) \right|. \]

The likelihood function corresponding to \( S^* \) is

\[ L^* = \left| \det(B^*) \right|^n \prod_{t=1}^{n} p(\varepsilon_t^*) \]

\[ = \left| \det(D^*) \right|^n \left| \det(B) \right|^n \prod_{t=1}^{n} p(\varepsilon_t) \left| \frac{\partial \varepsilon_t^*}{\partial \varepsilon_t} \right| \]

\[ = \left| \det(D^*) \right|^n \left| \det(B) \right|^n \prod_{t=1}^{n} p(\varepsilon_t) \left| \det(D^{-1}) \right| \]

\[ = L. \]

Thus both the structural forms \( S \) and \( S^* \) have same likelihood functions. Since the likelihood functions form the basis of statistical analysis, so both \( S \) and \( S^* \) have same implications. Moreover, it is difficult to identify whether the likelihood function corresponds to \( S \) and \( S^* \). Any attempt to estimate the parameters will result into failure in the sense that we cannot know whether we are estimating \( S \) and \( S^* \). Thus \( S \) and \( S^* \) are observationally equivalent. So the model is not identifiable.

A parameter is said to be identifiable within the model if the parameter has the same value for all equivalent structures contained in the model.

If all the parameters in structural equation are identifiable, then the structural equation is said to be identifiable.
Given a structure, we can thus find many observationally equivalent structures by non-singular linear transformation.

The apriori restrictions on $B$ and $\Gamma$ may help in the identification of parameters. The derived structures may not satisfy these restrictions and may therefore not be admissible.

The presence and/or absence of certain variables helps in the identifiability. So we use and apply some apriori restrictions. These apriori restrictions may arise from various sources like economic theory, e.g. it is known that the price and income affect the demand of wheat but rainfall does not affect it. Similarly, supply of wheat depends on income, price and rainfall. There are many types of restrictions available which can solve the problem of identification. We consider zero-one type restrictions.

**Zero-one type restrictions:**
Suppose the apriori restrictions are of zero-one type, i.e., some of the coefficients are one and others are zero. Without loss of generality, consider $S$ as

$$By_t + \Gamma x_t = \varepsilon_t, \ t = 1,2,\ldots,n.$$ 

When the zero-one type restrictions are incorporated in the model, suppose there are $G_\Delta$ jointly dependent and $K_\ast$ predetermined variables in $S$ having nonzero coefficients. Rest $(G-G_\Delta)$ jointly dependent and $(K-K_\ast)$ predetermined variables are having coefficients zero.

Without loss of generality, let $\beta_\Delta$ and $\gamma_\ast$ be the row vectors formed by the nonzero elements in the first row of $B$ and $\Gamma$ respectively. Thus the first row of $B$ can be expressed as $(\beta_\Delta \ 0).$ So $B$ has $G_\Delta$ coefficients which are one and $(G-G_\Delta)$ coefficients which are zero.

Similarly the first row of $\Gamma$ can be written as $(\gamma_\ast \ 0).$ So in $\Gamma,$ there are $K_\ast$ elements present (those take value one) and $(K-K_\ast)$ elements absent (those take value zero).

The first equation of the model can be rewritten as

$$\beta_{11}y_{1t} + \beta_{12}y_{2t} + \ldots + \beta_{1G_\Delta}y_{G_\Delta t} + \gamma_{11}x_{1t} + \gamma_{12}x_{2t} + \ldots + \gamma_{1K_\ast}x_{K_\ast t} = \varepsilon_{1t}$$

or

$$(\beta_\Delta \ 0)y_t + (\gamma_\ast \ 0)x_t = \varepsilon_{1t}, \ t = 1,2,\ldots,n.$$
Assume every equation describes the behaviour of a particular variable, so that we can take $\beta_{11} = 1$.

If $\beta_{11} \neq 1$, then divide the whole equation by $\beta_{11}$ so that the coefficient of $y_{1t}$ is one.

So the first equation of the model becomes

$$y_{1t} + \beta_{12} y_{2t} + \ldots + \beta_{1G} y_{Gi} + \gamma_{11} x_{1t} + \gamma_{12} x_{2t} + \ldots + \gamma_{1K} x_{K*} = \varepsilon_{1t}.$$ 

Now the reduced form coefficient relationship is

$$\pi = -B^{-1}\Gamma$$

or $B\pi = -\Gamma$

or $(\beta_{\Delta} \ 0)\pi = -(\gamma_* \ 0)$.

### Partition

$$\pi = \begin{pmatrix}
\pi_{\Delta*} & \pi_{\Delta*} \\
\pi_{\Delta*} & \pi_{\Delta*}
\end{pmatrix}$$

where the orders of $\pi_{\Delta*}$ is $(G_{\Delta} \times K_*)$, $\pi_{\Delta*}$ is $(G_{\Delta} \times K_*)$, $\pi_{\Delta*}$ is $(G_{\Delta} \times K_*)$ and $\pi_{\Delta*}$ is $(G_{\Delta} \times K_*)$

where $G_{\Delta\Delta} = G - G_{\Delta}$ and $K_* = K - K_*$.  

We can re-express

$$(\beta_{\Delta} \ 0)\pi = -(\gamma_* \ 0)$$

$$(\beta_{\Delta} \ 0_{\Delta*})\pi = -(\gamma_* \ 0_{\Delta*})$$

or

$$(\beta_{\Delta} \ 0_{\Delta*})\begin{pmatrix}
\pi_{\Delta*} & \pi_{\Delta*} \\
\pi_{\Delta*} & \pi_{\Delta*}
\end{pmatrix} = -(\gamma_* \ 0_{\Delta*})$$

$$\Rightarrow \beta_{\Delta} \pi_{\Delta*} = -\gamma_* \quad (i)$$

$$\beta_{\Delta} \pi_{\Delta*} = 0_{\Delta*} \quad (ii)$$

Assume $\pi$ is known. Then (i) gives a unique solution for $\gamma_*$ if $\beta_{\Delta}$ is uniquely found from (ii). Thus identifiability of $S$ lies upon the unique determination of $\beta_{\Delta}$. Out of $G_{\Delta}$ elements in $\beta_{\Delta}$, one has coefficient 1, so there are $(G_{\Delta} - 1)$ unknown elements in $\beta_{\Delta}$ that are unknown.
Note that
\[ \beta_\Lambda = (1, \beta_{12}, \ldots, \beta_{1G_\Lambda}) . \]

As
\[ \beta_\Lambda \pi_{\Lambda^*} = 0^* \]
or \[ (1 \beta) \pi_{\Lambda^*} = 0^* . \]

So \((G_\Lambda - 1)\) elements of \(\beta_\Lambda\) are uniquely determined as non-trivial solution when
\[ \text{rank}(\pi_{\Lambda^*}) = G_\Lambda - 1 . \]

Thus \(\beta_0\) is identifiable when
\[ \text{rank}(\pi_{\Lambda^*}) = G_\Lambda - 1 . \]

This is known as **rank condition** for the identifiability of parameters in \(S\). This condition is **necessary and sufficient**.

Another condition, known as **order condition** is only necessary and not sufficient. The order condition is derived as follows:

We now use the result that for any matrix \(A\) of order \(m \times n\), the \(\text{rank}(A) = \text{Min}(m, n)\). For identifiability, if \(\text{rank}(A) = m\) then obviously \(n > m\).

Since \(\beta_\Lambda \pi_{\Lambda^*} = 0\) and \(\beta_\Lambda\) has only \((G_\Lambda - 1)\) elements which are identifiable when
\[ \text{rank}(\pi_{\Lambda^*}) = G_\Lambda - 1 \]
\[ \Rightarrow K - K_* \geq G_\Lambda - 1 . \]

This is known as the **order condition** for identifiability.

There are various ways in which these conditions can be represented to have meaningful interpretations. We discuss them as follows:

1. \((G - G_\Lambda) + (K - K_*) \geq (G - G_\Lambda) + (G_\Lambda - 1)\)
   or \((G - G_\Lambda) + (K - K_*) \geq G - 1\)
Here

\( G - G_\Delta : \) Number of jointly dependent variables left out in the equation

\[ y_{it} + \beta_{12} y_{2t} + \ldots + \beta_{1G_\Delta} y_{G_\Delta} + \gamma_{11} x_{1t} + \ldots + \lambda_{1K} x_{Kt} = \varepsilon_{it}, \]

\( K - K_* : \) Number of predetermined variables left out in the equation

\[ y_{it} + \beta_{12} y_{2t} + \ldots + \beta_{1G_\Delta} y_{G_\Delta} + \gamma_{11} x_{1t} + \ldots + \lambda_{1K} x_{Kt} = \varepsilon_{it}, \]

\( G - 1 : \) Number of total equations - 1.

So left hand side of the condition denotes the total number of variables excluded is this equation.

Thus if total number of variables excluded in this equation exceeds \((G - 1)\), then the model is identifiable.

2. \( K \geq K_* + G_\Delta - 1 \)

Here

\( K_* : \) Number of predetermined variables present is the equation.

\( G_\Delta : \) Number of jointly dependent variables present is the equation.

3. Define \( L = K - K_* - (G_\Delta - 1) \)

\( L \) measures the degree of overidentification.

If \( L = 0 \) then the equation is said to be exactly identified.

\( L > 0 \), then the equation is said to be over identified.

\( L < 0 \Rightarrow \) the parameters cannot be estimated.

\( L = 0 \Rightarrow \) the unique estimates of parameters are obtained.

\( L > 0 \Rightarrow \) the parameters are estimated with multiple estimators.

So by looking at the value of \( L \), we can check the identifiability of the equation.

Rank condition tells whether the equation is identified or not. If identified, then the order condition tells whether the equation is exactly identified or over identified.
Note
We have illustrated the various aspects of estimation and conditions for identification in the simultaneous equation system using the first equation of the system. It may be noted that this can be done for any equation of the system and it is not necessary to consider only the first equation. Now onwards, we do not restrict to first equation but we consider any, say $i^{th}$ equation ($i = 1, 2, \ldots, G$).

Working rule for rank condition
The checking of rank condition sometimes, in practice, can be a difficult task. An equivalent form of rank condition based on the partitioning of structural coefficients is as follows.

Suppose
\[
B = \begin{pmatrix}
\beta_{\Delta} & 0_{\Delta}\beta \\
B_{\Delta} & B_{\Delta}\Delta
\end{pmatrix}, \quad \Gamma = \begin{pmatrix}
\gamma_{\ast} & 0_{\ast}
\\
\Gamma_{\ast} & \Gamma_{\ast}\ast
\end{pmatrix},
\]
where $\beta_{\Delta}, \gamma_{\ast}, 0_{\Delta}\beta$ and $0_{\ast}$ are row vectors consisting of $G_{\Delta}, K_{\ast}, G_{\Delta}\Delta$ and $K_{\ast} \ast$ elements respectively in them.
Similarly the orders of $B_{\Delta}$ is $((G-1) \times G_{\Delta})$, $B_{\Delta}\Delta$ is $((G-1) \times G_{\Delta}\Delta)$, $\Gamma_{\ast}$ is $((G-1) \times K_{\ast})$ and $\Gamma_{\ast}\ast$ is $((G-1) \times K_{\ast}\ast)$. Note that $B_{\Delta}\Delta$ and $\Gamma_{\ast}\ast$ are the matrices of structural coefficients for the variable omitted from the $i^{th}$ equation ($i = 1, 2, \ldots, G$) but included in other structural equations.

Form a new matrix
\[
\Delta = \begin{pmatrix}
B & \Gamma
\end{pmatrix}
= \begin{pmatrix}
\beta_{\Delta} & 0_{\Delta}\beta & \gamma_{\ast} & 0_{\ast}
\\
B_{\Delta} & B_{\Delta}\Delta & \Gamma_{\ast} & \Gamma_{\ast}\ast
\end{pmatrix}
\]
\[
B^{-1}\Delta = \begin{pmatrix}
B^{-1}B & B^{-1}\Gamma
\end{pmatrix}
= (I - \pi)
= \begin{pmatrix}
I_{\Delta}\Delta & 0 & -\pi_{\Delta\ast} & -\pi_{\Delta\ast\ast}
\\
0 & I_{\Delta\Delta} & -\pi_{\Delta\ast\ast} & -\pi_{\Delta\ast\ast}
\end{pmatrix}.
\]
If
\[
\Delta_{\ast} = \begin{pmatrix}
0_{\Delta\Delta} & 0_{\ast}
\\
B_{\Delta}\Delta & \Gamma_{\ast}\ast
\end{pmatrix}
\]
then clearly the rank of $\Delta_{\ast}$ is same as the rank of $\Delta$ since the rank of a matrix is not affected by enlarging the matrix by a rows of zeros or switching any columns.
Now using the result that if a matrix $A$ is multiplied by a nonsingular matrix then the product has the same rank as of $A$, we can write

$$\text{rank}(\Delta) = \text{rank}(B^{-1}\Delta)$$

$$\text{rank}(B^{-1}\Delta) = \text{rank}\left(\begin{bmatrix} 0 & -\pi_{\Delta^{**}} \\ I_{\Delta^{\Delta},\Delta^{\Delta}} & -\pi_{\Delta^{\Delta^{**}}} \end{bmatrix}\right)$$

$$= \text{rank}(I_{\Delta^{\Delta},\Delta^{\Delta}}) + \text{rank}(\pi_{\Delta^{\Delta^{**}}})$$

$$= (G - G_{\Delta}) + (G_{\Delta} - 1)$$

$$= G - 1$$

$$\text{rank}(B^{-1}\Delta) = \text{rank}(\Delta) = \text{rank}(B_{\Delta\Delta} \Gamma_{**}).$$

So

$$\text{rank}(B_{\Delta\Delta} \Gamma_{**}) = G - 1$$

and then the equation is identifiable.

Note that $(\beta_{\Delta\Delta} \Gamma_{**})$ is a matrix constructed from the coefficients of variables excluded from that particular equation but included in other equations of the model. If $\text{rank}(\beta_{\Delta\Delta} \Gamma_{**}) = G - 1$, then the equation is identifiable and this is a necessary and sufficient condition. An advantage of this term is that it avoids the inversion of matrix. A working rule is proposed like following.

**Working rule:**

1. Write the model in tabular form by putting ‘X’ if the variable is present and ‘0’ if the variable is absent.
2. For the equation under study, mark the 0’s (zeros) and pick up the corresponding columns suppressing that row.
3. If we can choose $(G - 1)$ rows and $(G - 1)$ columns that are not all zero then it can be identified.

**Example 2:**

Now we discuss the identifiability of following simultaneous equations model with three structural equations

$$\begin{align*}
(1) & \quad y_1 = \beta_{11}y_3 + \gamma_{11}x_1 + \gamma_{12}x_3 + \varepsilon_1 \\
(2) & \quad y_2 = \gamma_{21}x_1 + \gamma_{23}x_3 + \varepsilon_2 \\
(3) & \quad y_2 = \beta_{13}y_3 + \gamma_{31}x_1 + \gamma_{32}x_2 + \varepsilon_3.
\end{align*}$$

First we represent the equation (1)-(3) in tabular form as follows
• $G = \text{Number of equations} = 3$.
• ‘X’ denotes the presence and '0' denotes the absence of variables in an equation.
• $G_\Delta = \text{Numbers of ‘X’ in } (y_1, y_2, y_3)$.
• $K^* = \text{Number of ‘X’ in } (x_1, x_2, x_3)$.

Consider equation (1):

- Write columns corresponding to '0' which are columns corresponding to $y_2$ and $x_2$ and write as follows

<table>
<thead>
<tr>
<th>Equation number</th>
<th>$y_1$</th>
<th>$y_2$</th>
<th>$y_3$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$G_\Delta -1$</th>
<th>$K^*$</th>
<th>$L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>0</td>
<td>X</td>
<td>X</td>
<td>0</td>
<td>X</td>
<td>2 - 1 = 1</td>
<td>2</td>
<td>3 - 3 = 0</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>0</td>
<td>0</td>
<td>X</td>
<td>0</td>
<td>X</td>
<td>1 - 1 = 0</td>
<td>2</td>
<td>3 - 2 = 1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>0</td>
<td>2 - 1 = 1</td>
<td>2</td>
<td>3 - 3 = 0</td>
</tr>
</tbody>
</table>

- Check if any row/column is present with all elements ‘0’. If we can pick up a block of order $(G - 1)$ is which all rows/columns are all ‘0’, then the equation is identifiable.

Here $G - 1 = 3 - 1 = 2$, $B_{\Delta} = \begin{pmatrix} 0 \\ X \end{pmatrix}$, $\Gamma^* = \begin{pmatrix} 0 \\ X \end{pmatrix}$. So we need $(2 \times 2)$ matrix in which no row and no column has all elements ‘0’. In case of equation (1), the first row has all ‘0’, so it is not identifiable.

Notice that from order condition, we have $L = 0$ which indicates that the equation is identified and this conclusion is misleading. This is happening because order condition is just necessary but not sufficient.
Consider equation (2):

- Identify ‘0’ and then

\[
\begin{array}{c|ccc}
 & y_2 & y_3 & x_2 \\
\hline
\text{Equation (1)} & 0 & X & 0 \\
\text{Equation (3)} & X & X & X \\
\end{array}
\]

- \( B_{\Delta\Delta} = \begin{pmatrix} 0 & X \\ X & X \end{pmatrix}, \quad \Gamma_{\Delta} = \begin{pmatrix} 0 \\ X \end{pmatrix}. \)

- \( G - 1 = 3 - 1 = 2. \)
- We see that there is at least one block in which no row is ‘0’ and no column is ‘0’. So the equation (2) is identified.
- Also, \( L = 1 > 0 \Rightarrow \text{Equation (2) is over identified.} \)

Consider equation (3)

- Identify ‘0’.

\[
\begin{array}{c|cc}
 & y_1 & x_3 \\
\hline
\text{Equation (1)} & X & X \\
\text{Equation (2)} & X & X \\
\end{array}
\]

- So \( B_{\Delta\Delta} = \begin{pmatrix} X \\ X \end{pmatrix}, \quad \Gamma_{\Delta} = \begin{pmatrix} X \\ X \end{pmatrix}. \)
- We see that there is no ‘0’ present in the block. So equation (3) is identified.
- Also \( L = 0 \Rightarrow \text{Equation (3) is exactly identified.} \)
Estimation of parameters
To estimate the parameters of the structural equation, we assume that the equations are identifiable.

Consider the first equation of the model
\[ \beta_{11} y_{1t} + \beta_{12} y_{2t} + \ldots + \beta_{1G} y_{Gt} + \gamma_{11} x_{1t} + \gamma_{12} x_{2t} + \ldots + \gamma_{1K} x_{Kt} = \varepsilon_{1t}, t = 1, 2, \ldots, n. \]
Assume \( \beta_{11} = 1 \) and incorporate the zero-one type restrictions. Then we have
\[ y_{1t} = \beta_{12} y_{2t} - \ldots - \beta_{1G} y_{Gt} - \gamma_{11} x_{1t} - \gamma_{12} x_{2t} - \ldots - \gamma_{1K} x_{Kt} + \varepsilon_{1t}, t = 1, 2, \ldots, n. \]
Writing this equation in vector and matrix notations by collecting all \( n \) observations, we can write
\[ y_1 = Y_1 \beta + X_1 \gamma + \varepsilon_1 \]
where
\[
\begin{align*}
y_1 &= \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n} \end{pmatrix}, & Y_1 &= \begin{pmatrix} y_{21} & y_{31} & \cdots & y_{G1} \\ y_{22} & y_{32} & \cdots & y_{G2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{2n} & y_{3n} & \cdots & y_{Gn} \end{pmatrix}, \\
X_1 &= \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{K1} \\ x_{12} & x_{22} & \cdots & x_{K2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{Kn} \end{pmatrix}, & \beta &= \begin{pmatrix} -\beta_{12} \\ -\beta_{13} \\ \vdots \\ -\beta_{1G} \end{pmatrix}, & \gamma &= \begin{pmatrix} -\gamma_{11} \\ -\gamma_{12} \\ \vdots \\ -\gamma_{1K} \end{pmatrix}.
\end{align*}
\]
The order of \( y_1 \) is \( (n \times 1) \), \( Y_1 \) is \( (n \times (G - 1)) \), \( X_1 \) is \( (n \times K_s) \), \( \beta \) is \( ((G - 1) \times 1) \) and \( \gamma \) is \( (K_s \times 1) \).
This describes one equation of the structural model.
Now we describe the general notations in this model.

Consider the model with incorporation of zero-one type restrictions as
\[ y_i = Y_i \beta + X_i \gamma + \varepsilon \]
where \( y_i \) is a \( (n \times 1) \) vector of jointly dependent variables, \( Y_i \) is \( (n \times (G - 1)) \) matrix of jointly dependent variables where \( G \) denote the number of jointly dependent variables present on the right hand side of the equation, \( \beta \) is a \( ((G - 1) \times 1) \) vector of associated structural coefficients, \( X_i \) is a \( (n \times K_s) \) matrix of \( n \) observations on each of the \( K_s \) predetermined variables and \( \varepsilon \) is a \( (n \times 1) \) vector of structural disturbances. This equation is one of the equations of complete simultaneous equation model.
Stacking all the observations according to the variance, rather than time, the complete model consisting of $G$ equations describing the structural equation model can be written as

$$YB + X\Gamma = \Phi$$

where $Y$ is a $(n \times G)$ matrix of observation on jointly dependent variables, $B$ is $(G \times G)$ matrix of associated structural coefficients, $X$ is $(n \times K)$ matrix of observations on predetermined variables, $\Gamma$ is $(K \times G)$ matrix of associated structural coefficients and $\Phi$ is a $(n \times G)$ matrix of structural disturbances.

Assume $E(\Phi) = 0, \frac{1}{n}E(\Phi') = \Sigma$ where $\Sigma$ is positive definite symmetric matrix.

The reduced form of the model is obtained from structural equation model by post multiplying by $B^{-1}$ as

$$YBB^{-1} + X\Gamma B^{-1} = \Phi B^{-1}$$
$$Y = X\pi + V$$

where $\pi = -\Gamma B^{-1}$ and $V = \Phi B^{-1}$ are the matrices of reduced form coefficients and reduced form disturbances respectively.

The structural equation is expressed as

$$y = Y_i\beta + X_i\gamma + \epsilon_i$$

$$= (Y_i \quad X_i) \begin{bmatrix} \beta \\ \gamma \end{bmatrix} + \epsilon_i$$

$$= A\delta + \epsilon$$

where $A = (Y_i, \quad X_i)$, $\delta = (\beta \quad \gamma)'$ and $\epsilon = \epsilon_i$. This model looks like a multiple linear regression model.

Since $X_i$ is a submatrix of $X$, so it can be expressed as

$$X_i = XJ_i$$

where $J_i$ is called as a **select matrix** and consists of two types of elements, viz., 0 and 1. If the corresponding variable is present, its value is 1 and if absent, its value is 0.
Method of consistent estimation of parameters

1. Indirect least squares (ILS) method

This method for the consistent estimation of parameters is available only for exactly identified equations.

If equations are exactly identifiable, then

\[ K = G_\lambda + K_s - 1. \]

**Step 1:** Apply ordinary least squares to each of the reduced form equations and estimate the reduced form coefficient matrix.

**Step 2:** Find algebraic relations between structural and reduced form coefficients. Then find the structural coefficients.

The structural model at time \( t \) is

\[ \beta y_i + \Gamma x_i = \varepsilon_i; \quad t = 1, 2, \ldots, n \]

where \( y_i = (y_{1t}, y_{2t}, \ldots, y_{Gt})', \quad x_i = (x_{1t}, x_{2t}, \ldots, x_{Kt})' \).

Stacking all \( n \) such equations, the structural model is obtained as

\[ BY + \Gamma X = \Phi \]

where \( Y \) is \((n \times G)\) matrix, \( X \) is \((n \times K)\) matrix and \( \Phi \) is \((n \times K)\) matrix.

The reduced form equation is obtained by premultiplication of \( B^{-1} \) as

\[ B^{-1}BY + B^{-1}\Gamma X = B^{-1}\Phi \]

\[ Y = X\pi + V \]

where \( \pi = -B^{-1}\Gamma \) and \( B^{-1}\Phi \).

Applying OLS to reduced form equation yields the OLSE of \( \pi \) as

\[ \hat{\pi} = (X'X)^{-1}X'Y. \]

This is the first step of ILS procedure and yields the set of estimated reduced form coefficients.

Suppose we are interested in the estimation of following structural equation

\[ y = Y_i\beta + X_i\gamma + \varepsilon \]
where \( y \) is \((n \times 1)\) vector of \( n \) observation on dependent (endogenous) variable, \( Y_i \) is \((n \times (G_A - 1))\) matrix of observations on \( G_i \) current endogenous variables, \( X_i \) is \((n \times K_i)\) matrix of observations on \( K_i \) predetermined (exogenous) variables in the equation and \( \varepsilon \) is \((n \times 1)\) vector of structural disturbances.

Write this model as

\[
\begin{pmatrix} y_1 & Y_i & X_i \end{pmatrix} \begin{pmatrix} 1 \\ -\beta \\ -\gamma \end{pmatrix} = \varepsilon
\]

or more general

\[
\begin{pmatrix} y_1 & Y_1 & Y_2 & X_1 & X_2 \end{pmatrix} \begin{pmatrix} 1 \\ -\beta \\ 0 \\ 0 \\ -\gamma \end{pmatrix} = \varepsilon
\]

where \( Y_2 \) and \( X_2 \) are the matrices of observations on \((G - G_A + 1)\) endogenous and \((K - K_A)\) predetermined variables which are excluded from the model due to zero-one type restrictions.

Write

\[
\pi B = -\Gamma
\]

or \[
\begin{pmatrix} 1 \\ -\beta \end{pmatrix} = \begin{pmatrix} \gamma \\ 0 \end{pmatrix}.
\]

Substitute \( \pi \) as \( \hat{\pi} = (X'X)^{-1} X'Y \) and solve for \( \beta \) and \( \gamma \). This gives indirect least squares estimators \( b \) and \( c \) of \( \beta \) and \( \gamma \) respectively by solving

\[
\hat{\pi} \begin{pmatrix} 1 \\ -b \\ 0 \end{pmatrix} = \begin{pmatrix} c \\ 0 \end{pmatrix}
\]

or

\[
(X'X)^{-1} X'Y \begin{pmatrix} 1 \\ -b \\ 0 \end{pmatrix} = \begin{pmatrix} c \\ 0 \end{pmatrix}
\]

or

\[
(X'X)^{-1} X'(y_1 Y_i Y_2) \begin{pmatrix} 1 \\ -b \\ 0 \end{pmatrix} = \begin{pmatrix} c \\ 0 \end{pmatrix}
\]

\[
\Rightarrow X'y_1 - X'Yb = X'X \begin{pmatrix} c \\ 0 \end{pmatrix}
\]
Since \( X = (X_1 \quad X_2) \)

\[ \Rightarrow (X_1^t Y_1)b + (X_1^t X_1)c = X_1^t y_1 \]  

(i)

\[ (X_2^t Y_1)b + (X_2^t X_1)c = X_2^t y_1. \]  

(ii)

These equations (i) and (ii) are \( K \) equations is \( (G_\Delta + K_s - 1) \) unknowns. Solving the equations (i) and (ii) gives ILS estimators of \( \beta \) and \( \gamma \).

2. **Two stage least squares (2SLS) or generalized classical linear (GCL) method:**

This is more widely used estimation procedure as it is applicable to exactly as well as overidentified equations. The least squares method is applied in two stages in this method.

Consider equation \( y_1 = Y_1 \beta + X_1 \gamma + \varepsilon \).

**Stage 1:** Apply least squares method to estimate the reduced form parameters in the reduced form model

\[ Y_i = X \pi_i + V_i \]

\[ \Rightarrow \hat{\pi}_i = (X'X)^{-1}X'Y_i \]

\[ \Rightarrow Y_i = X \hat{\pi}_i. \]

**Stage 2:** Replace \( Y_i \) is structural equation \( y_1 = Y_1 \beta + X_1 \gamma + \varepsilon \) by \( \hat{Y}_i \) and apply OLS to thus obtained structural equation as follows:

\[ y_1 = \hat{Y}_i \beta + X_1 \gamma + \varepsilon 
\]

\[ = X \left( \left( X'X \right)^{-1} X' Y_i \right) \beta + X_1 \gamma + \varepsilon 
\]

\[ = \left[ X \left( \left( X'X \right)^{-1} X' \hat{Y}_i \right) X_1 \right] \left[ \begin{array}{c} \beta \\ \gamma \end{array} \right] + \varepsilon 
\]

\[ = \hat{A} \delta + \varepsilon. \]

where \( \hat{A} = \left[ X \left( \left( X'X \right)^{-1} X' \hat{Y}_i \right) X_1 \right], A = \left[ X \left( \left( X'X \right)^{-1} X' Y_i \right) X_1 \right] = \bar{H} A, \quad \bar{H} = X \left( \left( X'X \right)^{-1} X' \right) \) is idempotent and \( \delta = (\beta \quad \gamma)' \).

Applying OLS to \( Y_i = \hat{A} \delta + \varepsilon \) gives OLSE of \( \delta \) as
\[ \hat{\delta} = (\hat{A}' \hat{A})^{-1} \hat{A} \hat{y}_1 \]
\[ = (A' \hat{H} A)^{-1} A' \hat{H} \hat{y}_1 \]

or
\[ \left( \begin{array}{c} \hat{\beta} \\ \hat{\gamma} \end{array} \right) = \left( \begin{array}{cc} Y_i' \hat{H} Y_i & Y_i' X_1 \\ X_1 Y_i & X_1' X_1 \end{array} \right)^{-1} \left( \begin{array}{c} Y_i' \hat{Y}_1 \\ X_1' X_1 \end{array} \right) \left( \begin{array}{c} Y_i' \hat{V}_1 \\ X_1' \hat{V}_1 \end{array} \right) \]

where \( V_i = Y_i - X \pi_i \) is estimated by \( \hat{V}_i = Y_i - X \hat{\pi}_i = (I - \hat{H}) Y_i = HY_i, H = I - \hat{H} \). Solving these two equations, we get \( \hat{\beta} \) and \( \hat{\gamma} \) which are the two stage least squares estimators of \( \beta \) and \( \gamma \) respectively.

Now we see the consistency of \( \hat{\delta} \).

\[ \hat{\delta} = (\hat{A}' \hat{A})^{-1} \hat{A} \hat{y}_1 \]
\[ = (\hat{A}' \hat{A})^{-1} \hat{A}' (\hat{A} \delta + \varepsilon) \]
\[ \hat{\delta} - \delta = (\hat{A}' \hat{A})^{-1} \hat{A}' \varepsilon \]
\[ \text{plim} \left( \hat{\delta} - \delta \right) = \text{plim} \left( \frac{1}{n} \hat{A}' A \right)^{-1} \text{plim} \left( \frac{1}{n} \hat{A}' \varepsilon \right) \]

The 2SLS estimator \( \hat{\delta} \) is consistent if
\[ \text{plim} \left( \frac{1}{n} \hat{A}' \varepsilon \right) = \begin{bmatrix} \text{plim} \frac{1}{n} \hat{Y}_i' \varepsilon \\ \text{plim} \frac{1}{n} X_i' \varepsilon \end{bmatrix} = 0. \]

Since by assumption, \( \varepsilon \) is uncorrelated with \( \varepsilon \) in limit, so
\[ \text{plim} \left( \frac{1}{n} X_i' \varepsilon \right) = 0. \]

For \( \text{plim} \left( \frac{1}{n} \hat{Y}_i' \varepsilon \right) \), we observe that
\[ \text{plim} \left( \frac{1}{n} \hat{Y}_i' \varepsilon \right) = \text{plim} \left( \frac{1}{n} \hat{\pi}_i X_i' \varepsilon \right) \]
\[ = (\text{plim} \hat{\pi}_i) \left( \text{plim} \frac{1}{n} X_i' \varepsilon \right) \]
\[ = \pi_i 0 \]
\[ = 0. \]
Thus \( \lim (\hat{\delta} - \delta) = 0 \) and so the 2SLS estimators are consistent.

The asymptotic covariance matrix of \( \hat{\delta} \) is

\[
\text{Asy Var}(\hat{\delta}) = n^{-1} \lim \left[ n (\hat{\delta} - \delta) (\hat{\delta} - \delta)' \right]
\]

\[
= n^{-1} \lim \left[ n (\hat{A}' \hat{A})^{-1} \hat{A}' \varepsilon \varepsilon' \hat{A} (\hat{A}' \hat{A})^{-1} \right]
\]

\[
= n^{-1} \lim \left( \frac{1}{n} \hat{A}' \hat{A} \right)^{-1} \lim \left( \frac{1}{n} \hat{A}' \varepsilon \varepsilon' \hat{A} \right) \lim \left( \frac{1}{n} \hat{A}' \hat{A} \right)^{-1}
\]

\[
= n^{-1} \sigma^2 \varepsilon \lim \left( \frac{1}{n} \hat{A}' \hat{A} \right)^{-1}
\]

where \( \text{Var}(\varepsilon) = \sigma^2 \varepsilon \).

The asymptotic covariance matrix is estimated by

\[
s^2 \left( \hat{A}' \hat{A} \right)^{-1} = s^2 \left[ Y_i' X (X' X)^{-1} X' Y_i \quad Y_i' X_i \right]
\]

\[
\left. \begin{array}{c}
X_i' Y_i \\
X_i' X_i 
\end{array} \right]
\]

where

\[
s^2 = \frac{(y - Y_i \hat{\beta} - X_i \hat{\gamma})' (y - Y_i \hat{\beta} - X_i \hat{\gamma})}{n - G - K}
\]
Chapter 18
Seemingly Unrelated Regression Equations Models

A basic nature of multiple regression model is that it describes the behaviour of a particular study variable based on a set of explanatory variables. When the objective is to explain the whole system, there may be more than one multiple regression equations. For example, in a set of individual linear multiple regression equations, each equation may explain some economic phenomenon. One approach to handle such a set of equations is to consider the set up of simultaneous equations model is which one or more of the explanatory variables in one or more equations are itself the dependent (endogenous) variable associated with another equation in the full system. On the other hand, suppose that none of the variables is the system are simultaneously both explanatory and dependent in nature. There may still be interactions between the individual equations if the random error components associated with at least some of the different equations are correlated with each other. This means that the equations may be linked statistically, even though not structurally – through the jointness of the distribution of the error terms and through the non-diagonal covariance matrix. Such a behaviour is reflected in the seemingly unrelated regression equations (SURE) model in which the individual equations are in fact related to one another, even though superficially they may not seem to be.

The basic philosophy of the SURE model is as follows. The jointness of the equations is explained by the structure of the SURE model and the covariance matrix of the associated disturbances. Such jointness introduces additional information which is over and above the information available when the individual equations are considered separately. So it is desired to consider all the separate relationships collectively to draw the statistical inferences about the model parameters.

Example:
Suppose a country has 20 states and the objective is to study the consumption pattern of the country. There is one consumption equation for each state. So all together there are 20 equations which describe 20 consumption functions. It may also not necessary that the same variables are present in all the models. Different equations may contain different variables. It may be noted that the consumption pattern of the neighbouring states may have the characteristics in common. Apparently, the equations may look distinct individually but there may be some kind of relationship that may be existing among the equations. Such equations can be used to examine the jointness of the distribution of disturbances. It seems reasonable to
assume that the error terms associated with the equations may be contemporaneously correlated. The equations are apparently or “seemingly” unrelated regressions rather than independent relationships.

**Model:**

We consider here a model comprising of $M$ multiple regression equations of the form

$$y_{it} = \sum_{j=1}^{k_i} x_{itj} \beta_{ij} + \epsilon_{it}, \quad t = 1, 2, ..., T; \quad i = 1, 2, ..., M; \quad j = 1, 2, ..., k_i$$

where $y_{it}$ is the $t^{\text{th}}$ observation on the $i^{\text{th}}$ dependent variable which is to be explained by the $i^{\text{th}}$ regression equation, $x_{itj}$ is the $t^{\text{th}}$ observation on $j^{\text{th}}$ explanatory variable appearing in the $i^{\text{th}}$ equation, $\beta_{ij}$ is the coefficient associated with $x_{itj}$ at each observation and $\epsilon_{it}$ is the $t^{\text{th}}$ value of the random error component associated with $i^{\text{th}}$ equation of the model.

These $M$ equations can be compactly expressed as

$$y_i = X_i \beta_i + \epsilon_i, \quad i = 1, 2, ..., M$$

where $y_i$ is $(T \times 1)$ vector with elements $y_{it}$; $X_i$ is $(T \times K_i)$ matrix whose columns represent the $T$ observations on an explanatory variable in the $i^{\text{th}}$ equation; $\beta_i$ is a $(k_i \times 1)$ vector with elements $\beta_{ij}$; and $\epsilon_i$ is a $(T \times 1)$ vector of disturbances. These $M$ equations can be further expressed as

$$\begin{pmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_M
\end{pmatrix} =
\begin{pmatrix}
  X_1 & 0 & \cdots & 0 \\
  0 & X_2 & \cdots & 0 \\
  \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & \cdots & X_M
\end{pmatrix} \begin{pmatrix}
  \beta_1 \\
  \beta_2 \\
  \vdots \\
  \beta_M
\end{pmatrix} +
\begin{pmatrix}
  \epsilon_1 \\
  \epsilon_2 \\
  \vdots \\
  \epsilon_M
\end{pmatrix}$$

or

$$y = X \beta + \epsilon$$

where the orders of $y$ is $(TM \times 1)$, $X$ is $(TM \times k^*)$, $\beta$ is $(k^* \times 1)$, $\epsilon$ is $(TM \times 1)$ and $k^* = \sum_i k_i$. 

---

*Econometrics  |  Chapter 18  |  SURE Models  |  Shalabh, IIT Kanpur*
Treat each of the $M$ equations as the classical regression model and make conventional assumptions for $i = 1, 2, \ldots, M$ as

- $X_i$ is fixed.
- $\text{rank}(X_i) = k_i$.
- $\lim_{T \to \infty} \left( \frac{1}{T} X_i'X_i \right) = Q_{ii}$ where $Q_{ii}$ is nonsingular with fixed and finite elements.
- $E(u_i) = 0$.
- $E(u_i'u_i) = \sigma_{ii} I_T$ where $\sigma_{ii}$ is the variance of disturbances in $i^{th}$ equation for each observation in the sample.

Considering the interactions between the $M$ equations of the model, we assume

- $\lim_{T \to \infty} \frac{1}{T} X_i'X_j = Q_{ij}$
- $E(u_i'u_j) = \sigma_{ij} I_T$; $i, j = 1, 2, \ldots, M$

where $Q_{ij}$ is non-singular matrix with fixed and finite elements and $\sigma_{ij}$ is the covariance between the disturbances of $i^{th}$ and $j^{th}$ equations for each observation in the sample.

Compactly, we can write

$$E(\varepsilon) = 0$$

$$E(\varepsilon\varepsilon') = \begin{pmatrix} \sigma_{11} I_T & \sigma_{12} I_T & \cdots & \sigma_{1M} I_T \\ \sigma_{21} I_T & \sigma_{22} I_T & \cdots & \sigma_{2M} I_T \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{M1} I_T & \sigma_{M2} I_T & \cdots & \sigma_{MM} I_T \end{pmatrix} = \Sigma \otimes I_T = \psi$$

where $\otimes$ denotes the Kronecker product operator, $\psi$ is $(MT \times MT)$ matrix and $\Sigma = \left(\sigma_{ij}\right)$ is $(M \times M)$ positive definite symmetric matrix. The definiteness of $\Sigma$ avoids the possibility of linear dependencies among the contemporaneous disturbances in the $M$ equations of the model.

The structure $E(uu') = \Sigma \otimes I_T$ implies that

- variance of $\varepsilon_{it}$ is constant for all $t$.
- contemporaneous covariance between $\varepsilon_{it}$ and $\varepsilon_{jt}$ is constant for all $t$.
- intertemporal covariance between $\varepsilon_{it}$ and $\varepsilon_{it^*}$ ($t \neq t^*$) are zero for all $i$ and $j$. 

_Econometrics_ | Chapter 18 | SURE Models | _Shalabh, IIT Kanpur_
By using the terminologies “contemporaneous” and “intertemporal” covariance, we are implicitly assuming that the data are available in time series form but this is not restrictive. The results can be used for cross-section data also. The constancy of the contemporaneous covariances across sample points is a natural generalization of homoskedastic disturbances in a single equation model.

It is clear that the \( M \) equations may appear to be not related in the sense that there is no simultaneity between the variables in the system and each equation has its own explanatory variables to explain the study variable. The equations are related stochastically through the disturbances which are serially correlated across the equations of the model. That is why this system is referred to as SURE model.

The SURE model is a particular case of simultaneous equations model involving \( M \) structural equations with \( M \) jointly dependent variable and \( k (\geq k_i \text{ for all } i) \) distinct exogenous variables and in which neither current nor logged endogenous variables appear as explanatory variables in any of the structural equations.

The SURE model differs from the multivariate regression model only in the sense that it takes account of prior information concerning the absence of certain explanatory variables from certain equations of the model. Such exclusions are highly realistic in many economic situations.

**OLS and GLS estimation:**

The SURE model is

\[
y = X \beta + \varepsilon, \quad E(\varepsilon) = 0, \quad V(\varepsilon) = \Sigma \otimes I_T = \psi.
\]

Assume that \( \psi \) is known.

The OLS estimator of \( \beta \) is

\[
b_0 = (X'X)^{-1}X'y
\]

Further

\[
E(b_0) = \beta
\]

\[
V(b_0) = E(b_0 - \beta)(b_0 - \beta)'
\]

\[
= (X'X)^{-1}X'\psi X(X'X)^{-1}.
\]

The generalized least squares (GLS) estimator of \( \beta \)
\[ \hat{\beta} = (X'\psi^{-1}X)^{-1} X'\psi^{-1}y \]
\[ \quad = \left[ X'\left(\Sigma^{-1} \otimes I_T\right)X \right]^{-1} X'\left(\Sigma^{-1} \otimes I_T\right)y \]
\[ E\left(\hat{\beta}\right) = \beta \]
\[ V\left(\hat{\beta}\right) = E\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)' \]
\[ \quad = (X'\psi^{-1}X)^{-1} \]
\[ \quad = \left[ X'\left(\Sigma^{-1} \otimes I_T\right)X \right]^{-1}. \]

Define
\[ G = (X'X)^{-1} X' - \left(X'\psi^{-1}X\right)^{-1} X'\psi^{-1} \]
then \( GX = 0 \) and we find that
\[ V\left(b_0\right) - V\left(\hat{\beta}\right) = G\psi G'. \]
Since \( \psi \) is positive definite, so \( G\psi G' \) is atleast positive semidefinite and so GLSE is, in general, more efficient than OLSE for estimating \( \beta \). In fact, using the result that GLSE best linear unbiased estimator of \( \beta \), so we can conclude that \( \hat{\beta} \) is the best linear unbiased estimator in this case also.

**Feasible generalized least squares estimation:**
When \( \Sigma \) is unknown, then GLSE of \( \beta \) cannot be used. Then \( \Sigma \) can be estimated and replaced by \((M \times M)\) matrix \( S \). With such replacement, we obtain a feasible generalized least squares (FGLS) estimator of \( \beta \) as
\[ \hat{\beta}_f = \left[ X'\left(S^{-1} \otimes I_T\right)X \right]^{-1} X'\left(S^{-1} \otimes I_T\right)y. \]
Assume that \( S = \left(\bar{s}_y\right) \) is nonsingular matrix and \( s_y \) is some estimator of \( \sigma_y \).

**Estimation of \( \Sigma \)**
There are two possible ways to estimate \( \sigma_y \)'s.
1. Use of unrestricted residuals

Let \( K \) be the total number of distinct explanatory variables out of \( k_1, k_2, \ldots, k_m \) variables in the full model

\[
y = X \beta + \varepsilon, \quad E(\varepsilon) = 0, \quad V(\varepsilon) = \Sigma \otimes I_T
\]

and let \( Z \) be a \( T \times K \) observation matrix of these variables.

Regress each of the \( M \) study variables on the column of \( Z \) and obtain \((T \times 1)\) residual vectors

\[
\hat{e}_i = y_i - Z(Z'Z)^{-1}Z'y_i \quad i = 1, 2, \ldots, M
\]

where \( \hat{H}_Z = I_T - Z(Z'Z)^{-1}Z' \).

Then obtain

\[
s_{ij} = \frac{1}{T} \hat{e}_i \hat{e}_j
\]

\[
= \frac{1}{T} y_i \hat{H}_Z y_j
\]

and construct the matrix \( S = \left( \left( s_{ij} \right) \right) \) accordingly.

Since \( X_i \) is a submatrix of \( Z \), so we can write

\[
X_i = ZJ_i
\]

where \( J_i \) is a \((K \times k_i)\) selection matrix. Then

\[
\hat{H}_Z X_i = X_i - Z(Z'Z)^{-1}Z'X_i
\]

\[
= X_i - ZJ_i
\]

\[
= 0
\]

and thus

\[
y_i \hat{H}_Z y_j = (\beta_i X_i + \varepsilon_i) \hat{H}_Z (X_j \beta_j + \varepsilon_j)
\]

\[
= \varepsilon_i \hat{H}_Z \varepsilon_j.
\]

Hence
\[ E\left(s_{ij}\right) = \frac{1}{T}E\left(\varepsilon_i^tH_z\varepsilon_j\right) \]
\[ = \frac{1}{T}\sigma_{ij}tr \left(H_z\right) \]
\[ = \left(1-\frac{K}{T}\right)\sigma_{ij} \]
\[ E\left(\frac{T}{T-K}s_{ij}\right) = \sigma_{ij}.\]

Thus an unbiased estimator of \( \sigma_{ij} \) is given by \( \frac{T}{T-K}s_{ij} \).

### 2. Use of restricted residuals

In this approach to find an estimator of \( \sigma_{ij} \), the residuals obtained by taking into account the restrictions on the coefficients which distinguish the SURE model from the multivariate regression model are used as follows.

Regress \( y_i \) on \( X_i \), i.e., regress each equation, \( i = 1, 2, \ldots, M \) by OLS and obtain the residual vector

\[ \tilde{u}_i = \left[I - X_i\left(X_i^tX_i\right)^{-1}X_i^t\right]y_i \]
\[ = H_{X_i}y_i.\]

A consistent estimator of \( \sigma_{ij} \) is obtained as

\[ s_{ij} = \frac{1}{T}\tilde{u}_i\tilde{u}_j \]
\[ = \frac{1}{T}y_i^tH_{X_i}\tilde{H}_{X_i}y_j \]

where

\[ H_{X_i} = I - X_i\left(X_i^tX_i\right)^{-1}X_i \]
\[ \tilde{H}_{X_i} = I - X_j\left(X_j^tX_j\right)^{-1}X_j.\]

Using \( s_{ij}^* \), a consistent estimator of \( S \) can be constructed.

If \( T \) in \( s_{ij}^* \) is replaced by

\[ tr \left(H_{X_i}\tilde{H}_{X_j}\right) = T - k_i - k_j + tr \left(X_i^tX_i\right)^{-1}X_i^tX_j\left(X_j^tX_j\right)^{-1}X_j^tX_i \]

then \( s_{ij}^* \) is an unbiased estimator of \( \sigma_{ij} \).