REGRESSION TECHNIQUES

BY TANUJIT CHAKRABORTY
(RESEARCH SCHOLAR, ISI KOLKATA)
Mail : tanujitisi@gmail.com

CONTENT **Page No.**

Simple Linear Regression	**3**
Multiple Linear Regression	**12**
Multicollinearity	**18**
Model Adequacy Checking	**26**
Transformations	**35**
Dummy Variable	**37**
Polynomial Regression Models	**42**
Generalized Linear Model	**47**
Non-linear Model	**50**
Autocorrelation Issue	**53**
Measurement Error	**56**
Problems & Solutions	**58**
Panel Data Models	**71**
Binary Outcome Model	**75**
Hazard Model	**77**
Time Series Models	**79**
Classification & Regression Tree	**82**
Regression Splines	**86**

REGRESSION ANALYSIS

Books:- Introduction to Linear Regression Analysis (3rd Ed) By Montgomery

Applied Regression Analysis (3rd Ed) by Draper & Smith.

- **Regression analysis** is a statistical tool for investigating the relationship between a dependent variable and one or more independent variable. This technique is widely used for prediction and forecasting.

Scatter plot is an essential tool for checking correlation.

- **Simple Linear Regression model** is a model with single regressor x that has a linear relationship with a response y.

Model is: $y = \beta_0 + \beta_1 x + \epsilon$

y: Response variable; β_0: intercept; β_1: slope; x: regressor variable; ϵ: random error component

We now make some basic assumption on the model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad ; \; i = 1(1)n$$

Given data (X_i, Y_i) ; checking scatter plot whether linear model is appropriate or not.

Assumptions:-

1. ϵ_i is an RV with mean '0' and s.d. σ (unknown). i.e. $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma^2$.
2. $Cov(\epsilon_i, \epsilon_j) = 0 \Rightarrow \epsilon_i$ & ϵ_j are uncorrelated. i.e. no autocorrelation.
3. $\epsilon_i \overset{IID}{\sim} N(0, \sigma^2)$.

$$E(Y_i) = E(\beta_0 + \beta_1 X_i + \epsilon_i) = \beta_0 + \beta_1 X_i$$

$$V(Y_i) = V(\beta_0 + \beta_1 X_i + \epsilon_i) = V(\epsilon_i) = \sigma^2$$

$$\epsilon_i \overset{IND}{\sim} N(0, \sigma^2)$$

$$Y_i \overset{ind.}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

The line fitted by least square is the one that makes the sum of squares of all vertical discrepancies as small as possible.

BY TANUJIT CHAKRABORTY,
RS, ISI KOLKATA, M: 9051152281
MAIL :- tanujitisi@gmail.com

We estimate β_0 & β_1 so that the sum of squares of all the diff. between the observation y_i and the fitted line is minimum.

$$S = SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$ is minimum.

$\hat{\beta}_0$ is the estimate of β_0, and $\hat{\beta}_1$ is the estimate of β_1.

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

Normal equations

Solving these two equations :-

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \Rightarrow \sum y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

$$\to n\hat{\beta}_0 = \sum y_i - \hat{\beta}_1 \sum x_i$$

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

$$\sum x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \Rightarrow \sum x_i (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) = 0$$

$$\to \sum x_i (y_i - \bar{y}) = \hat{\beta}_1 \sum (x_i - \bar{x}) x_i$$

$$\to \hat{\beta}_1 = \frac{\sum (y_i - \bar{y}) x_i}{\sum (x_i - \bar{x}) x_i}$$

$$= \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})(x_i - \bar{x})} = \frac{S_{xy}}{S_{xx}}$$

$$= \sum_{i=1}^{n} (x_i - \bar{x}) y_i \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$= \sum_{i=1}^{n} c_i y_i, \quad c_i = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}$$

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

$$= \frac{1}{n}\sum y_i - \hat{\beta}_1 \bar{x}$$

So, $\hat{\beta}_1, \hat{\beta}_0$ are linear combinations, so its called Linear estima

$E(\hat{\beta}_1) = \beta_1$ $\quad \bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{\epsilon}$.

$\longrightarrow$ $\hat{\beta}_1$ is a UE of β_1, $\hat{\beta}_0$ is a UE of β_0.

$$y_i - \bar{y} = \beta_1(x_i - \bar{x}) + \epsilon_i - \bar{\epsilon}$$

$$E(y_i - \bar{y}) = \beta_1(x_i - \bar{x}) + E(\epsilon_i - \bar{\epsilon}) = \beta_1(x_i - \bar{x}); \quad \epsilon_i \sim N(0, \sigma^2)$$

$$E(\hat{\beta}_1) = E\left[\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}\right] = \beta_1 \quad (\text{Proved})$$

$$E(\hat{\beta}_0) = E(\bar{y} - \hat{\beta}_1\bar{x}) = E(\beta_0 + \beta_1\bar{x} - \hat{\beta}_1\bar{x})$$

$$= \beta_0 \quad (\text{Proved})$$

$$V(\hat{\beta}_1) = V\left(\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}\right) = V\left(\frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2}\right) = V\left(\sum_{i=1}^{n} c_i y_i\right)$$

where, $c_i = \dfrac{(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$

$$= \sum c_i^2 V(y_i) = \sum c_i^2 \sigma^2 = \sigma^2 \cdot \frac{\sum(x_i - \bar{x})^2}{\left[\sum(x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{S_{xx}}$$

$$V(\hat{\beta}_0) = V(\bar{y} - \hat{\beta}_1\bar{x}) = V(\bar{y}) + V(\hat{\beta}_1\bar{x}) - 2\bar{x}\,\mathrm{Cov}(\bar{y}, \hat{\beta}_1)$$

$$= \frac{\sigma^2}{n} + \bar{x}^2 V(\hat{\beta}_1) - 2\bar{x} \times 0$$

$$= \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{S_{xx}} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right).$$

$$\left[\mathrm{Cov}(\bar{y}, \hat{\beta}_1) = \mathrm{Cov}\left(\frac{\sum y_i}{n}, \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2}\right) = \frac{\sum(x_i - \bar{x})V(y_i)}{n\sum(x_i - \bar{x})^2}\right.$$

$$= \frac{\sum(x_i - \bar{x})V(y_i)}{n\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})\sigma^2}{n\sum(x_i - \bar{x})^2} = \frac{\sigma^2\sum(x_i - \bar{x})}{n\sum(x_i - \bar{x})^2}$$

$$\left. = 0. \right]$$

<u>Estimation of σ^2:-</u>

$$SS_{Res} = \sum e_i^2 = \sum(y_i - \hat{y}_i)^2$$

$$= \sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$= \sum(y_i - \bar{y} - \hat{\beta}_1 x_i + \hat{\beta}_1\bar{x})^2$$

$$= \sum\left[y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x})\right]^2$$

$$= S_{yy} + \hat{\beta}_1^2 S_{xx} - 2\hat{\beta}_1 S_{xy}; \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$\therefore \boxed{E\left(\frac{SS_{Res}}{n-2}\right) = \sigma^2} \qquad \boxed{SS_{Res} = S_{yy} - \hat{\beta}_1^2 S_{xx}}$$

(7)

$$E(S_{yy}) = E\left(\sum(y_i-\bar{y})^2\right) = E\left[\sum y_i^2\right] - nE\left[\bar{y}^2\right]$$
$$= \sum E\left[y_i^2\right] - nE\left[\bar{y}^2\right]$$
$$= n\sigma^2 + \sum(\beta_0+\beta_1 x_i)^2 - \sigma^2 - n(\beta_0+\beta_1\bar{x})^2$$
$$= (n-1)\sigma^2 + \beta_1^2 S_{xx}$$

$$\left[\quad y_i = \beta_0 + \beta_1 x_i + \epsilon_i \;; \quad E(y_i) = \beta_0 + \beta_1 x_i\right.$$
$$V(y_i) = \sigma^2$$
$$E(y_i^2) = V(y_i) + [E(y_i)]^2$$
$$= \sigma^2 + (\beta_0 + \beta_1 x_i)^2$$
$$E(\bar{y}^2) = V(\bar{y}) + [E(\bar{y})]^2$$
$$\left. = \frac{\sigma^2}{n} + (\beta_0 + \beta_1\bar{x})^2\right]$$

$$E(\hat{\beta}_1^2 S_{xx}) = S_{xx} E(\hat{\beta}_1^2)$$
$$= \sigma^2 + \beta_1^2 S_{xx}$$

$$\left[\quad E(\hat{\beta}_1) = \beta_1\right.$$
$$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$
$$E(\hat{\beta}_1^2) = V(\hat{\beta}_1) + [E(\hat{\beta}_1)]^2$$
$$= \frac{\sigma^2}{S_{xx}} + \beta_1^2$$

$$E(SS_{Res}) = E(S_{yy}) - E(\hat{\beta}_1^2 S_{xx})$$
$$= (n-1)\sigma^2 + \beta_1^2 S_{xx} - \sigma^2 - \beta_1^2 S_{xx}$$
$$= (n-2)\sigma^2$$

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

$= MS_{Res}$ = Residual Mean Square

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

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

$$e_i \sim N(0, \sigma^2)$$

$$\frac{e_i}{\sigma} \sim N(0,1)$$

$$\frac{e_i^2}{\sigma^2} \sim \chi_1^2 .$$

$\hat{\beta}_0$ & $\hat{\beta}_1$ are LSE of β_0 & β_1 respectively.

$$e_i = y_i - \hat{y}_i$$

$$e_1 + e_2 + \cdots + e_n = 0$$

$$e_1 x_1 + e_2 x_2 + \cdots + e_n x_n = 0$$

There are $(n-2)$ degree of freedom of residuals.

$$\frac{SS_{Res}}{\sigma^2} = \frac{\sum e_i^2}{\sigma^2} \sim \chi^2_{n-2}$$

$$\therefore \boxed{\frac{(n-2) MS_{Res}}{\sigma^2} \sim \chi^2_{n-2}} \qquad MS_{Res} = \frac{SS_{Res}}{n-2}$$

Evaluate Model: Test of Slope Coefficient

Shows if there is a linear relationship between X & Y.

$H_0 : \beta_1 = 0$ (No linear relationship)

$H_1 : \beta_1 \neq 0$ (Linear relationship).

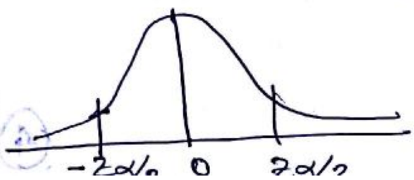
$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \sum c_i y_i \; ; \; y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right).$$

Thus $$Z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$$

If σ^2 is known, we can use $Z = \frac{\hat{\beta}_1}{\sqrt{\sigma^2/S_{xx}}}$, under $H_0 : \beta_1 = 0$ to test H_0.

Reject H_0 if $|Z| > Z_{\alpha/2}$



9

Usually σ^2 is not known:- $\sigma^2 = E(MS_{Res}) = E\left(\frac{SS_{Res}}{n-2}\right)$

Test statistic:- $t = \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}}$

$\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1)$

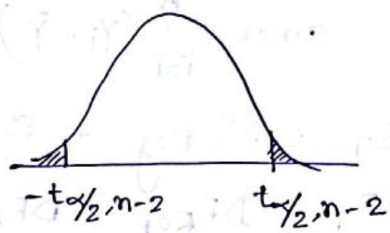
$\dfrac{(n-2)MS_{Res}}{\sigma^2} \sim \chi^2_{n-2}$

independently

Let $X \sim N(0,1)$, $Y \sim \chi^2_n$ ind.

Then $\dfrac{X}{\sqrt{\frac{Y}{n}}} \sim t_n$

So, $\dfrac{\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}}}{\sqrt{\frac{(n-2)MS_{Res}}{(n-2)\sigma^2}}} \sim t_{n-2}$



So, $\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \sim t_{n-2}$.

Test statistic:- $t = \dfrac{\hat{\beta}_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}}$ under $H_0 : \beta_1 = 0$

We reject $H_0 : \beta_1 = 0$ if $|t| > t_{\alpha/2; n-2}$

ANOVA

Total variation in data $= \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

How much of the variation is explained by the model?

Identity $\quad Y_i - \hat{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$

$$\sum (Y_i - \hat{Y}_i)^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 + 2\underbrace{\sum (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)}_{=0}$$

$$CPT = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)$$

$$= \sum \hat{\beta}_1 (x_i - \bar{x})\left[(Y_i - \bar{Y}) - \hat{\beta}_1 (x_i - \bar{x})\right]$$

$$= \hat{\beta}_1 S_{xy} - \hat{\beta}_1^2 S_{xx}$$

$$= \hat{\beta}_1 (S_{xy} - \hat{\beta}_1 S_{xx}) = 0$$

$Y_i - \bar{Y} = \hat{\beta}_1 (x_i - \bar{x})$

$Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

$= Y_i - \bar{Y} - \hat{\beta}_1 (x_i - \bar{x})$

Since $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

$$SS_T = SS_{Reg} + SS_{Res}$$

$$SS_{Regression} = \sum_1^n (\hat{y}_i - \bar{y})^2 = \sum \hat{\beta}_1^2 (x_i - \bar{x})^2 = \hat{\beta}_1^2 S_{xx}.$$

$$SS_{Residual} = \sum_1^n (y_i - \hat{y}_i)^2 = \sum e_i^2 \sim \chi^2_{n-2}$$

since $\sum e_i = 0$, $\sum x_i e_i = 0$ so, e_i's are not indep.

$$SS_T = \sum_1^n (y_i - \bar{y})^2 \text{ has DF } (n-1)$$

since $\sum_{i=1}^n (y_i - \bar{y}) = 0$.

$$SS_T = SS_{Reg} + SS_{Res}$$

$$DF_T = DF_{Reg} + DF_{Res} \Rightarrow (n-1) = 1 + (n-2)$$

ANOVA Table:-

Source of Variation	DF	SS	MS	F
Regression	1	SS_{Reg}	$MS_{Reg} = \frac{SS_{Reg}}{1}$	$F = \frac{MS_{Reg}}{MS_{Res}}$
Residual	$n-2$	SS_{Res}	$MS_{Res} = \frac{SS_{Res}}{n-2}$	
Total	$n-1$	SS_T		

$$E(MS_{Res}) = \sigma^2$$

$$E(MS_{Reg}) = \sigma^2 + \beta_1^2 S_{xx}$$

$$\left.\begin{array}{l} \frac{(n-2) MS_{Res}}{\sigma^2} \sim \chi^2_{n-2} \\ \frac{MS_{Reg}}{\sigma^2} \sim \chi^2_1 \end{array}\right\} \text{indep}$$

under $H_0: \beta_1 = 0$

$$F = \frac{MS_{Reg}}{MS_{Res}} \sim F_{1, n-2}$$

To test $H_0: \beta_1 = 0$

We compute F and reject H_0 if $F > F_{\alpha; 1, n-2}$.

Let $X \sim \chi^2_m$, $Y \sim \chi^2_n$ ind.

Then $\dfrac{X/m}{Y/n} \sim F_{m,n}$

'R^2': A measure of "Goodness of Fit":-

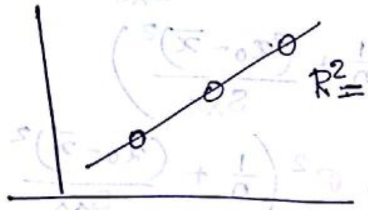
R^2 measures the proportion of variability in response variable that is explained by the regression model.

Coefficient of Determination:-

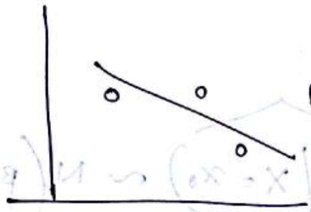
$$R^2 = \frac{SS_{Reg}}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$

$$0 \le R^2 \le 1$$

$R^2 = 1$, $SS_{Reg} = SS_T$, i.e. $SS_{Res} = 0$, if the fitted model explains all the variability in y



$R^2 = 1$



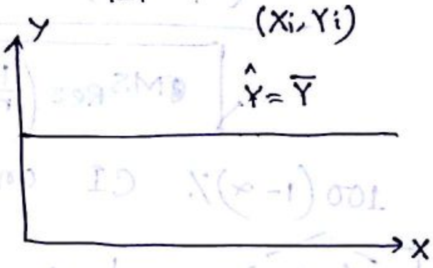
$R^2 = 0.8$

$R^2 = 0$, $SS_{Res} = SS_T \Rightarrow \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2$

$\Rightarrow \hat{Y}_i = \bar{Y}$

Fitted model: $\hat{Y} = \bar{Y}$.

i.e., when there is no relationship between Y and X.



Confidence Interval for β_1:-

$$y = \beta_0 + \beta_1 x + \epsilon$$

LSE of β_1 is $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$ $\quad E(\hat{\beta}_1) = \beta_1$ and $V(\hat{\beta}_1) = \dfrac{\sigma^2}{S_{xx}}$

$$= \sum c_i Y_i$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

So, $\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\sigma^2}{S_{xx}}}} \sim N(0,1)$; For σ^2 unknown $t = \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{MS_{Res}}{S_{xx}}}} \sim t_{n-2}$.

$$\therefore P\left\{ -t_{\alpha/2, n-2} \le \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{MS_{Res}}{S_{xx}}}} \le t_{\alpha/2, n-2} \right\} = 1 - \alpha; \quad \alpha = 0.05$$

$100(1-\alpha)\%$ CI for β_1 is

$$\hat{\beta}_1 - t_{\alpha/2, n-2}\sqrt{\frac{MS_{Res}}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2, n-2}\sqrt{\frac{MS_{Res}}{S_{xx}}}$$

Interval Estimation of Mean Response $E(Y)$ for given $X = X_0$:-

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$E(Y \mid X = X_0) = \beta_0 + \beta_1 X_0$$

An unbiased estimator of $E(Y \mid X = x_0)$ is

$$\widehat{E(Y \mid X = X_0)} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$V(\hat{\beta}_0 + \hat{\beta}_1 x_0) = V(\bar{y} + \hat{\beta}_1(x_0 - \bar{x})) \quad : \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

$$= V(\bar{y}) + V(\hat{\beta}_1(x_0 - \bar{x})) + \underbrace{2\,Cov(\bar{y}, \hat{\beta}_1(x_0 - \bar{x}))}_{=0}$$

$$= \frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \frac{\sigma^2}{S_{xx}}$$

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

$$\widehat{E(Y \mid X = x_0)} \sim N\left(\beta_0 + \beta_1 x_0,\ \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right)$$

$$\frac{\widehat{E(Y \mid X = x_0)} - E(Y \mid x = x_0)}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}$$

$100(1-\alpha)\%$ CI on $E(Y \mid X = x_0)$ is

$$\left[\widehat{E(Y \mid x_0)} \pm t_{\alpha/2, n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right]$$

CI is minimum at $x = x_0$. This widens as $|x_0 - \bar{x}|$ increases.

Prediction of new observation:-

y_0 corresponds to a specific value of regressor $x = x_0$

$$y_0 = \beta_0 + \beta_1 x_0 + \epsilon \qquad E(Y \mid X_0) = \beta_0 + \beta_1 x_0$$

If $x = x_0$, then $\hat{\beta}_0 + \hat{\beta}_1 x_0$ is point estimator of the response y_0

$$\psi = y_0 - \hat{y}_0, \quad E(\psi) = 0$$

$$V(\psi) = V(y_0 - \hat{y}_0) = V(y_0) + V(\hat{y}_0) = \sigma^2 + V(\hat{y}_0)$$

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

$$\frac{\psi - 0}{V(\psi)} \sim t_{n-2}$$

13

Multiple Linear Regression:-

More than one regressor variables, say K−1

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{K-1} x_{i,K-1} + \epsilon_i \; ; \; i = 1(1)n$$

This is linear of unknown parameters $\beta_0, \beta_1, \ldots, \beta_{K-1}$.

Assumption:- $\epsilon_i \overset{ind}{\sim} N(0, \sigma^2)$

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{K-1} \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

vector of obsn. Vector of parameters vector of errors

$n \times K$ matrix

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1K-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2K-1} \\ & \vdots & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nK-1} \end{pmatrix}$$

Model:- $Y = X\beta + \epsilon$

Estimation of Model parameters:

LSM determines the parameters by minimizing

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_{K-1} x_{K-1}$$

$$= \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_{K-1} x_{iK-1})^2$$

$$e = (Y - \hat{Y})$$

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = e'e = (Y - \hat{Y})'(Y - \hat{Y})$$

$$= Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

Normal equations:- $\dfrac{\partial SS_{Res}}{\partial \hat{\beta}_0} = 0 \Rightarrow \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_{K-1} x_{iK-1}) = 0$

$$\left.\begin{aligned} &\sum_{i=1}^{n} e_i = 0 \\ &\sum e_i x_{i1} = 0 \\ &\vdots \\ &\sum e_i x_{iK-1} = 0 \end{aligned}\right\} K \text{ constraints}$$

$$\frac{\partial SS_{Res}}{\partial \hat{\beta}} = 0$$

$$\Rightarrow -2X'Y + 2X'X\hat{\beta} = 0$$

$$\Rightarrow \hat{\beta} = (X'X)^{-1}(X'Y)$$

Statistical properties of LSE:-

$$\hat{\beta} = (X'X)^{-1}(X'Y)$$

$$E(\hat{\beta}) = E\left[(X'X)^{-1}X'Y\right]$$
$$= E\left[(X'X)^{-1}X'(X\beta+\epsilon)\right]$$
$$= E\left[(X'X)^{-1}(X'X)\beta\right] + E\left[(X'X)^{-1}X'\epsilon\right]$$
$$= \beta + 0 = \beta; \quad E(\epsilon)=0$$

$$V(\hat{\beta}) = V\left((X'X)^{-1}X'Y\right) = (X'X)^{-1}X' I\sigma^2 X(X'X)^{-1}$$
$$= \sigma^2 (X'X)^{-1}$$

$$SS_{Res} = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} \quad ; \quad \hat{\beta} = (X'X)^{-1}X'Y$$
$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'Y$$
$$= Y'Y - \hat{\beta}'X'Y$$
$$= \sum_{i=1}^{n} e_i \quad ; \quad e_i \sim N(0,\sigma^2)$$

You can choose only $(n-K)$ e_i's independently.

SS_{Res} has $(n-K)$ df. $\quad \dfrac{e_i^2}{\sigma^2} \sim \chi_1^2$

$MS_{Res} = \dfrac{SS_{Res}}{n-K}$ $\quad \dfrac{SS_{Res}}{\sigma^2} = \dfrac{\sum e_i^2}{\sigma^2} \sim \chi^2_{n-K}$

$E(MS_{Res}) = \sigma^2$

$$SS_T = \sum_{1}^{n}(Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2$$

SS_T has df $n-1$ since $\sum(Y_i - \bar{Y}) = 0$

$$SS_{Reg} = SS_T - SS_{Res} = \hat{\beta}'X'Y - n\bar{Y}^2$$

SS_{Reg} has $(K-1)$ DF.

15

Test for significance of Regression model:-

If there is linear relationship between the response and any one of the regressor variable $x_1, x_2, \ldots, x_{K-1}$.

$H_0: \beta_1 = \beta_2 = \cdots = \beta_{K-1} = 0$

Vs. $H_1: \beta_j \neq 0$ for at least one j.

$SS_T = SS_{Reg} + SS_{Res}$

$(n-1) = (n-K) + (K-1)$

$$MS_{Res} = \frac{SS_{Res}}{n-K}$$

$$E(MS_{Res}) = \sigma^2$$

$$\frac{SS_{Reg}}{\sigma^2} \sim \chi^2_{K-1}$$

$$E(MS_{Reg}) = \sigma^2 + \frac{\beta^{*\prime} X_c' X_c \beta^*}{(K-1)\sigma^2}$$

$$\frac{SS_{Res}}{\sigma^2} \sim \chi^2_{n-K}$$

$$\beta^* = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{K-1} \end{pmatrix}$$

$$F = \frac{SS_{Reg}/K-1}{SS_{Res}/n-K} \sim F_{K-1,n-K} \;;$$

$$F = \frac{MS_{Reg}}{MS_{Res}} \quad \text{at least one } \beta_j \neq 0 \;; \quad X_c = \begin{pmatrix} x_{11}-\bar{x} & \cdots & x_{1K-1}-\bar{x} \\ \vdots & & \\ x_{n1}-\bar{x} & \cdots & x_{nK-1}-\bar{x} \end{pmatrix}$$

We reject $H_0: \beta_1 = \beta_2 = \cdots = \beta_{K-1} = 0$ if $F > F_{\alpha, K-1, n-K}$.

ANOVA table easily one can do.

Test on individual regression coefficient (Partial/Marginal test):-

Test the significance of x_j in the presence of other regressors in the model.

$H_0: \beta_j = 0$ Vs $H_1: \beta_j \neq 0$

$$\hat{\beta} = (X'X)^{-1} X'Y$$

$$\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1})$$

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 (X'X)^{-1}_{jj}}} \sim N(0,1)$$

Test statistic:-

$$t = \frac{\hat{\beta}_j}{\sqrt{MS_{Res}(X'X)^{-1}_{jj}}} \sim t_{n-K} \text{ under } H_0.$$

$H_0: \beta_j = 0$ is rejected if $|t| > t_{\alpha/2, n-K}$.

Confidence Intervals on regression coefficients:-

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad V(\hat{\beta}) = \sigma^2 (X'X)^{-1} \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{K-1} \end{pmatrix}$$

$$\hat{\beta}_i \sim N\left(\beta_i, \sigma^2 (X'X)^{-1}_{ii}\right)$$

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{MS_{Res}(X'X)^{-1}_{ii}}} \sim t_{n-k}$$

$100(1-\alpha)\%$ CI for the parameters β_i is

$$\hat{\beta}_i - t_{\alpha/2, n-k}\sqrt{MS_{Res}(X'X)^{-1}_{ii}} \leq \beta_i \leq \hat{\beta}_i + t_{\alpha/2, n-k}\sqrt{MS_{Res}(X'X)^{-1}_{ii}}$$

Commonly used Linear Transformation :-

Equation	Transformation Y'	Transformation X'	Changed equation
$Y = \beta_0 x^{\beta_1}$	$Y' = \log y$	$x' = \log x$	$Y' = \log\beta_0 + \beta_1 x'$
$Y = \beta_0 e^{\beta_1 x}$	$Y' = \ln y$	x	$Y' = \ln\beta_0 + \beta_1 x$
$Y = \beta_0 + \beta_1 \log x$	~~$y' = \log x$~~ y	$x' = \log x$	$Y' = \beta_0 + \beta_1 x'$
$Y = \dfrac{x}{\beta_0 x - \beta_1}$	$Y' = 1/y$	$X' = 1/x$	$Y' = \beta_0 - \beta_1 x'$

NOTE:- Standardization of the data is needed when units are different and large scale variable value difference in the data.

All possible Regression

We need to consider all reg. equations involving

$$\begin{array}{lll} 0 \text{ regressors} & \binom{K-1}{0} & Y = \beta_0 + \epsilon \\ 1 \quad '' & \binom{K-1}{1} & \\ 2 \quad '' & \binom{K-1}{2} & \\ \vdots & & \\ K-1 \text{ regressors} & \binom{K-1}{K-1} & \\ \hline & 2^{K-1} & \end{array}$$

These equations are evaluated according to some suitable criteria:

- R^2 ~~Coefficient of Multiple Determination R_p^2~~
- Adjusted R^2
- MSRes
- Mallow's Statistic (C_p)

Sequential Selection: — Forward Selection
— Backward Selection
— Stepwise Selection
(Page:- 75)

If $k-1 = 4$, $k = 5$

There are $2^4 = 16$ possible regression equations.

Criteria for evaluating subset regression models:-

- **R_p^2 (Coefficient of multiple determination):-**

Let R_p^2 denote the coefficient of multiple determination for a subset reg. model with $(p-1)$ regressors and intercept β_0.

$$R_p^2 = \frac{SS_{Reg}(p)}{SS_T} = 1 - \frac{SS_{Res}(p)}{SS_T}$$

$SS_{Reg}(p)$, $SS_{Res}(p)$ denote Reg. SS and Res. SS for subset model with $(p-1)$ regressors.

$R_p^2 \uparrow$ as $p\uparrow$ and $SS_{Res}(p) \downarrow$ as $p\uparrow$

& is maximum when $p = K$

$$R_1^2 = 0 \; ; \; p = 1, \; p-1 = 0 \; ; \; y = \beta_0 + \epsilon$$

Suppose $R_p^2 = 53.4\%$.

explains 53% of the total variability in the response variable

So, calculate R_p^2 for all possible 2^{K-1} combinations, and higher R_p^2 means best regression model.

All possible models with (p-1) reg. are evaluated and the one giving the greatest R_p^2 are tabulated. Higher the value of R_p^2 indicates the best fit.

- Residual Mean Squares (MS_{Res}):-

$SS_{Res} \downarrow$ as $p \uparrow$ $\qquad MS_{Res}(p) = \dfrac{SS_{Res}(p)}{n-p}$.

lower the value of $MS_{Res}(p)$ indicates better fit.

- Adjusted Coefficient of Multiple Determination ($\bar{R}^2$):-

$$R_p^2 = \frac{SS_{Reg}(p)}{SS_T} = 1 - \frac{SS_{Res}(p)}{SS_T}$$

$SS_{Res}(p) \downarrow$ as $p \uparrow$

$\Rightarrow R_p^2 \uparrow$ as $p \uparrow$

R^2 is not a good measure of the quality of fit; it don't consider new included variable.

$$\bar{R}_p^2 = 1 - \frac{MS_{Res}(p)}{MS_T} = 1 - \frac{SS_{Res}(p)}{n-p} \times \frac{n-1}{SS_T}$$

$$= 1 - \frac{n-1}{n-p} \times \frac{SS_{Res}(p)}{SS_T}$$

$$= 1 - \frac{n-1}{n-p}\left(1 - R_p^2\right)$$

$\bar{R}_p^2$ value will not necessarily increase with the addition of any regressors.

- Mallow's Statistic (C_p):-

Measures the overall bias or Mean Square Error in the fitted model.

$\hat{y}_i$ is i^{th} fitted value.

$E(y_i)$ is the expected response for the reg. model.

$$MSE = \frac{\sum_1^n E\left(\hat{y}_i - E(y_i)\right)^2}{\sigma^2}$$

$$C_p = \frac{SS_{Res}(p)}{MS_{Res}^{Full}} - n + 2p$$

when $p = k$, full model, $SS_{Res}(p) = SS_{Res}^{Full}$

$\boxed{C_k = k}$ since $C_k = \dfrac{SS_{Res}(k)}{MS_{Res}} - n + 2k = n - k + (2k - n) = k$

low C_p value indicate better fit.

In general, look for the model where Mallow's C_p is small & close to p. (Software use)

Multicollinearity

The problem of multicollinearity exists when two or more regressor variables are strongly correlated or linearly dependent.

Suppose we have to fit the model $Y = X\beta + \epsilon$

LSE : $\hat{\beta} = (X'X)^{-1}(X'Y)$

If $(X'X)$ is singular then we can't perform the inverse.

This happens when at least one column of X is LD on the other.

Effect of Multicollinearity / problems due to Multicollinearity :-

MLR model with two regressors

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon \quad , i = 1(1)n$$

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ & \vdots & \\ 1 & X_{n1} & X_{n2} \end{bmatrix}$$

1. Strong multicollinearity between regressors result in large variance and covariance of regression coefficients.

Centering & Scaling Regression Data :-

$$x_{i1} = \frac{X_{i1} - \bar{X}_1}{\sqrt{S_{11}}}, \quad x_{i2} = \frac{X_{i2} - \bar{X}_2}{\sqrt{S_{22}}}, \quad y_i = \frac{Y_i - \bar{Y}}{\sqrt{S_{yy}}}$$

$$\bar{X}_1 = \frac{1}{n}\sum_{i=1}^{n} X_{i1}, \quad S_{11} = \sum_{i=1}^{n}(X_{i1} - \bar{X}_1)^2, \quad \bar{Y} = \frac{1}{n}\sum_{1}^{n} Y_i$$

$$\bar{X}_2 = \frac{1}{n}\sum_{i=1}^{n} X_{i2}, \quad S_{22} = \sum_{i=1}^{n}(X_{i2} - \bar{X}_2)^2, \quad S_{YY} = \sum_{1}^{n}(Y_i - \bar{Y})^2$$

$X_1 \quad X_2 \quad Y$

$\bar{x}_1$

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2$

$x_1 \quad x_2 \quad y$

$\bar{x}_1 = 0 \quad \bar{x}_2 = 0 \quad \bar{y} = 0$

$\sum_{i=1}^{n} x_{i1}^2 = 1 \quad \sum_{i=1}^{n} x_{i2}^2 = 1 \quad \sum_{i=1}^{n} y_i^2 = 1$

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2$

The model, assuming that X_1, X_2 and Y are centered and scaled, is

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

$$X = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \\ x_{n1} & x_{n2} \end{pmatrix} = \begin{pmatrix} \frac{X_{11}-\bar{X}_1}{\sqrt{S_{11}}} & \frac{X_{12}-\bar{X}_2}{\sqrt{S_{22}}} \\ \frac{X_{21}-\bar{X}_1}{\sqrt{S_{11}}} & \frac{X_{22}-\bar{X}_2}{\sqrt{S_{22}}} \\ \vdots & \\ \frac{X_{n1}-\bar{X}_1}{\sqrt{S_{11}}} & \frac{X_{n2}-\bar{X}_2}{\sqrt{S_{22}}} \end{pmatrix}, \quad Y = \begin{pmatrix} \frac{Y_1-\bar{Y}}{\sqrt{S_{YY}}} \\ \frac{Y_2-\bar{Y}}{\sqrt{S_{YY}}} \\ \vdots \\ \frac{Y_n-\bar{Y}}{\sqrt{S_{YY}}} \end{pmatrix}$$

<u>Normal equation:-</u>

correlation mtx.

$$(X'X)\hat{\beta} = X'Y \Rightarrow \begin{pmatrix} 1 & r_{12} \\ r_{21} & 1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} r_{1y} \\ r_{2y} \end{pmatrix}$$

where r_{12} is sample correlation between x_1 and x_2
r_{1y} " " " " x_1 & y
r_{2y} " " " " x_2 & y

$$r_{1y} = \frac{\sum_1^n (x_{i1}-\bar{x}_1)(y_i-\bar{y})}{\sqrt{S_{11} S_{yy}}} \quad \&$$

$$r_{12} = \frac{\sum (x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sqrt{S_{11} S_{22}}}$$

Inverse of $(X'X)$ is $(X'X)^{-1} = \begin{pmatrix} \frac{1}{1-r_{12}^2} & -\frac{r_{12}}{1-r_{12}^2} \\ \frac{-r_{21}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{pmatrix}$

$$\hat{\beta} = (X'X)^{-1} X'Y$$

$$= \begin{pmatrix} \frac{1}{1-r_{12}^2} & -\frac{r_{12}}{1-r_{12}^2} \\ \frac{-r_{21}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{pmatrix} \begin{pmatrix} r_{1y} \\ r_{2y} \end{pmatrix}$$

The estimates are

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12} r_{2y}}{1-r_{12}^2} \quad \& \quad \hat{\beta}_2 = \frac{r_{2y} - r_{21} r_{1y}}{1-r_{12}^2}$$

$$V(\hat{\beta}_j) = \sigma^2 (X'X)^{-1}_{jj}$$

$$V(\hat{\beta}_1) = \sigma^2 \cdot (X'X)^{-1}_{11} = \frac{\sigma^2}{1-r_{12}^2} \rightarrow \infty \text{ as } |r_{12}| \rightarrow 1$$

$$V(\hat{\beta}_2) = \sigma^2 \cdot (X'X)^{-1}_{22} = \frac{\sigma^2}{1-r_{12}^2} \rightarrow \infty \text{ as } |r_{12}| \rightarrow 1$$

If there is strong multicollinearity between x_1 and x_2, then the correlation coefficient will be large.

$$Cov(\hat{\beta}_1, \hat{\beta}_2) = \sigma^2 (X'X)^{-1}_{12} = \frac{-\sigma^2 r_{12}}{1-r_{12}^2} \rightarrow \pm\infty$$

depending on whether $r_{12} \rightarrow +1$ or $r_{12} \rightarrow -1$

More than two regressors (MLR):-

It can be shown that the diagonal elements of $(X'X)^{-1}$ matrix are $\frac{1}{1-R_j^2}$ $\forall\ j = 1(1)K-1$;

where R_j^2 is the coefficient of multiple determination ~~for~~ for the regression of x_j on the remaining $(K-2)$ regressors.

$$R_j^2 = \frac{SS\ reg}{SST}$$

$$V(\hat{\beta}_j) = \sigma^2 (X'X)^{-1}_{jj} = \frac{\sigma^2}{1-R_j^2} \rightarrow \infty \text{ as } R_j^2 \rightarrow 1$$

If there is strong multicollinearity between x_j and any subset of other $(K-2)$ reg., then R_j^2 will be close to unity.

(2) Multicollinearity tends to produce LSE $\hat{\beta}$ that are too far from the true parameter β:-

$$L^2 = \sum_{i=1}^{K-1} (\hat{\beta}_i - \beta_i)^2$$

Expected square distance $= E(L^2) = \sum_i E(\hat{\beta}_i - \beta_i)^2$; $E(\hat{\beta}_i) = \beta_i$.

$$= \sum_i E(\hat{\beta}_i - E(\hat{\beta}_i))^2$$

$$= \sum_i V(\hat{\beta}_i) = \sum_i \sigma^2 (X'X)^{-1}_{ii}$$

$$= \sum_i \frac{\sigma^2}{1-R_i^2} = \sigma^2 \sum \frac{1}{1-R_j^2}$$

When there is multicollinearity $\frac{1}{1-R_i^2}$ will be large for at least one i.

i.e. $\frac{1}{1-R_i^2} \rightarrow \infty$ as $R_i^2 \rightarrow 1$.

3. Model coefficient with '−'ve sign when '+'ve sign is expected.
4. High significance in a global F-test but in which none of the regressors are significant in partial F-test.

Example:- Consider the data in the following table:-

X_1	X_2	Y
1	8	6
4	2	8
9	−8	1
11	−10	0
3	6	5
8	−6	3
5	0	2
10	−12	−4
2	4	10
7	−2	−3
6	−4	5

Fitted model:-

$$\hat{Y} = 14 - 2X_1 - \frac{X_2}{2}$$

ANOVA table:-

Source of Variation	DF	SS	MS	F
Regression	2	122	61	7.17
Residual	8	68	8.5	
Total	10	190		

$$F = 7.17 > F_{0.05,2,8} = 4.46$$

We reject $H_0: \beta_1 = \beta_2 = 0$.
We accept $H_1: \beta_i \neq 0$ for at least one i

Global test says we rejecting null hypothesis, i.e., there is linear relationship between y and X_i's.

What does X_2 contributes, given that X_1 is already in the regression?

$H_0: \beta_2 = 0$

Vs $H_1: \beta_2 \neq 0$

$$t = \frac{\hat{\beta}_2}{\sqrt{MS_{Res}(X'X)^{-1}_{22}}} = -0.8348$$

$|t| \not> t_{0.025, 8} = 2.306$

We accept $H_0: \beta_2 = 0$.

What does X_1 contributes, given that X_2 is already in the regression?

$H_0: \beta_1 = 0$

Vs. $H_1: \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1}{\sqrt{MS_{Res}(X'X)^{-1}_{11}}} = 1.668$$

$|t| \not> t_{0.025, 8} = 2.308$

We accept $H_0: \beta_1 = 0$

None of the partial test are significant. So, there is the presence of multicollinearity in the data.

(4) Different model selection procedures yield different models.

Techniques for detecting Multicollinearity:-

- Examination of Correlation Matrix $(X'X)$

A simple measure of multicollinearity is inspection of off-diagonal elements r_{ij} in $X'X$. $|r_{ij}| > .9$ indicates multicollinearity problem. Examine the correlation matrix $(X'X)$ is helpful in detecting linear dependence between pairs of regressors.

Examining the correlation matrix $(X'X)$ is not helpful in detecting multicollinearity problem arising from linear dependence between more than two regressors.

- Eigen System Analysis on $X'X$:-

Multicollinearity can also be detected from the eigenvalues of the correlation matrix $X'X$.

For a $(k-1)$ regression model, there will be $(k-1)$ eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{k-1}$.

If there are one or more linear dependences in the data, then one or more eigen values will be small.

Define the condition number of $(X'X)$ as

$$K = \frac{\lambda_{max}}{\lambda_{min}}$$

As a general rule,

$K < 100$ indicates no serious problem with multicollinearity.

$100 \leq K \leq 1000$ " moderate to strong "

$K > 1000$ " severe problem " "

The condition indices of the $(X'X)$ matrix are

$$K_j = \frac{\lambda_{max}}{\lambda_j}, \quad j = 1(1)k-1$$

Clearly the Largest condition index is the condition number.

The number of $K_j > 1000$ is a useful measure of the number of near linear dependence in $X'X$.

The correlation matrix $(X'X)$ may be decomposed as

$$X'X = TDT'$$

where, $D = diag(\lambda_1, \lambda_2, \ldots, \lambda_{k-1})$ and

$$T_{\overline{k-1} \times \overline{k-1}} = (t_1, t_2, \ldots, t_{k-1})' \text{ where}$$

$t_i = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{k-1} \end{pmatrix}$ is the eigen vector associated with eigen value λ_i.

If the eigenvalue λ_i is close to zero, the elements of the associated eigenvector t_i describe the nature of linear dependence.

$$\sum_{i=1}^{k-1} a_i x_i = 0$$

- **Variance Inflation Factor:- (VIF):-**

Variance of the i^{th} regressor coefficient

$$V(\hat{\beta}_i) = \sigma^2 (X'X)^{-1}_{ii} = \frac{\sigma^2}{1-R_i^2}$$

R_i^2 is the coefficient of multiple determination when x_i is regressed on the remaining regressors.

If x_i is nearly orthogonal to the remaining regressors, R_i^2 is small and $\frac{1}{1-R_i^2} \to 1$.

If x_i is nearly linearly dependent on some subset of the remaining regressors, $R_i^2 \to 1$ and $\frac{1}{1-R_i^2} \to \infty$.

$V(\hat{\beta}_i)$ can be viewed as the factor by which the $V(\hat{\beta}_j)$ is increased due to linear dependence among the regressors.

The VIF associated with regressor x_i is defined by

$$VIF_i = \frac{1}{1-R_i^2}$$

large value of VIF_i indicates possible multicollinearity associated with x_i.

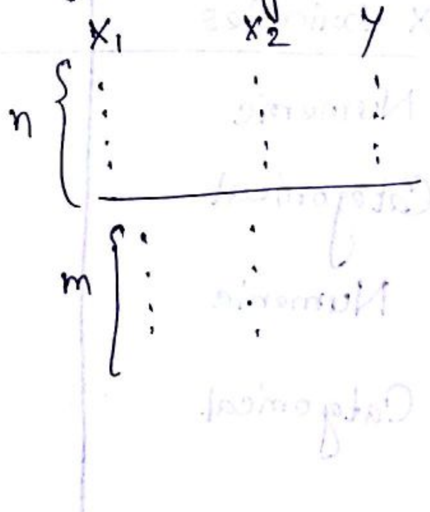
In general, $VIF_i \geq 5$ indicates possible multicollinearity problem.

$VIF_i \geq 10$ " at most certainly multicollinearity problem.

Dealing with Multicollinearity:-

- **Collect Additional data:-** Collecting additional data has been suggested as the best method of dealing with multicollinearity.

Additional data should be collected in manner to break up the multicollinearity in the existing data



- Remove Regressors from the model:-

If two regressors are linearly dependent, it means they contain redundant information. Thus we can pick one regressor to keep in the model and discard the other one. If x_1, x_2 and x_3 are linearly dependent, then eliminating one reg. may helpful to reduce the effect of multicollinearity.
Eliminating regressor to reduce multicollinearity may damage the predictive power of the model.

- Collapse Variables:- Combine two or more variables which are linearly dependent into Single composite variables.

SLR: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \forall\, i = 1(1)n$

MLR:- $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{K-1} x_{iK-1} + \epsilon_i$

Assumption:- $E(\epsilon_i) = 0$

$V(\epsilon_i) = \sigma^2$

errors are uncorrelated & normally distd.

$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

X Variables	Y Variables	Regression type
Numeric	Numeric	Ordinary L.S
Categorical	Numeric	Dummy Variable
Numeric	Categorical	Logistic Regression
Categorical	Categorical	Logistic using dummy Variable

(27)

MODEL ADEQUACY CHECKING

Residual $e_i = y_i - \hat{y}_i$ — y_i is the observation, $\hat{y}_i$ is the corresponding fitted value.

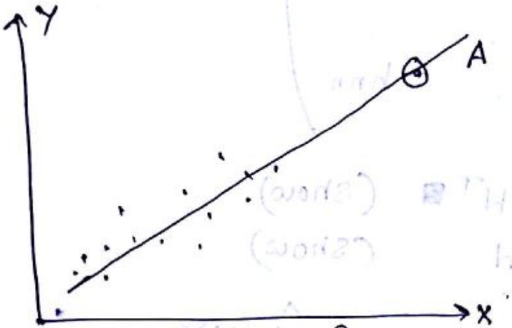
Check the assumption: $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

ϵ_i's are independent but residuals e_i's are not independent, as the n residuals have only $(n-k)$ DF.

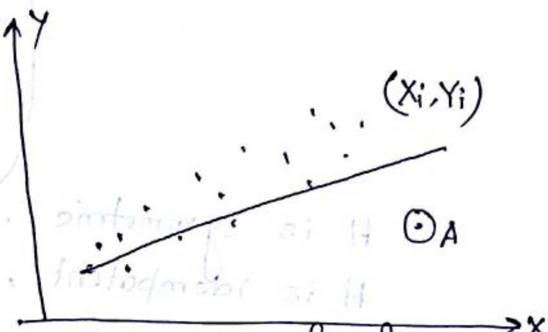
It is convenient to think of the residuals as the observed value of the errors.

Plotting the residuals is an effective way to investigate how well the reg. model fit the data or to check the model assumption.

Leverage & influential obsn.s



An example of leverage point, it's on the trend of the data set. $L_i > \frac{2p}{n}$ suggest leverage point (Software Rule)



An example of influential observation that has noticable impact on the model coefficient.

p: No. of parameters in your model
n: total no. of observation

For influential, $D_i > 1$ (Cook's statistic) shows influential obsn.

Various Types of Residuals

Regular residuals e_i
Standardize residuals d_i
Studentized residuals r_i
PRESS residuals $e_{(i)}$

Residual plots

Normal Probability plot
Plot of residuals (e_i) against the fitted values $(\hat{y}_i)$
Partial regression & partial residual plot

The hat matrix & the various types of residuals:

MLR : $Y = X\beta + \epsilon$; $V(\epsilon) = \sigma^2 I_n$.

Solution: $\hat{\beta} = (X'X)^{-1}X'Y$ if $(X'X)$ is non-singular.

Fitted model $\hat{Y} = X\hat{\beta}$

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

$= HY$, say, where $H = X(X'X)^{-1}X'$ = Hat matrix

(It maps Y to $\hat{Y}$, so called Hat mtx)

$$H = ((h_{ij})) = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \cdots & \cdots & & \cdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{pmatrix}$$

H is symmetric, i.e. $H = H^T$ (show)

H is idempotent, i.e. $H^2 = H$ (show)

$$\begin{aligned} e = Y - \hat{Y} &= Y - HY = (I-H)Y \quad \because \hat{Y} = HY \\ &= (I-H)(X\beta + \epsilon) \\ &= X\beta - HX\beta + (I-H)\epsilon \\ &= X\beta - X(X'X)^{-1}X'X\beta + (I-H)\epsilon \\ &= (I-H)\epsilon \end{aligned}$$

Variance - covariance matrix of e : $\text{Var}(e) = (I-H)\sigma^2 I(I-H)$

$= \sigma^2(I-H)^2$

$= \sigma^2(I-H)$

$$e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

$V(e_i) = \sigma^2(1 - h_{ii})$ where h_{ii} is the ith diagonal element of the hat matrix H.

$H = X(X'X)^{-1}X'$ $\quad X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix}$

$h_{ii} = x_i'(X'X)^{-1}x_i$

where x_i' is the ith row of X matrix.

h_{ii} measures the distance of ith observation from the center of x- coordinate.

Scanned by CamScanner

Studentized residuals

We define, $r_i = \dfrac{e_i}{\sqrt{MS_{Res}(1-h_{ii})}}$ $\quad V(e_i) = \sigma^2(1-h_{ii})$, $\hat{\sigma}^2 = MS_{Res}$

Studentized residuals have constant variance $V(r_i) = 1$ regardless of the location in x-coordinate, when the form of the model is correct.

Standardized residuals

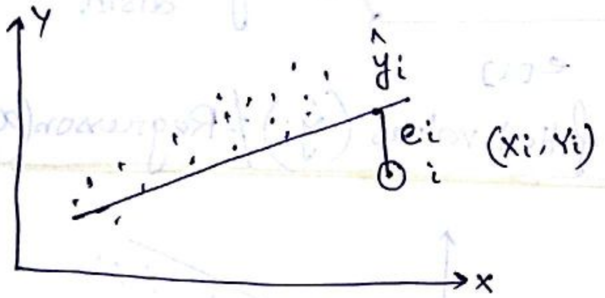
Define, $d_i = \dfrac{e_i}{\sqrt{MS_{Res}}}$; we approx. MS_{Res} as variance of i^{th} residual e_i.

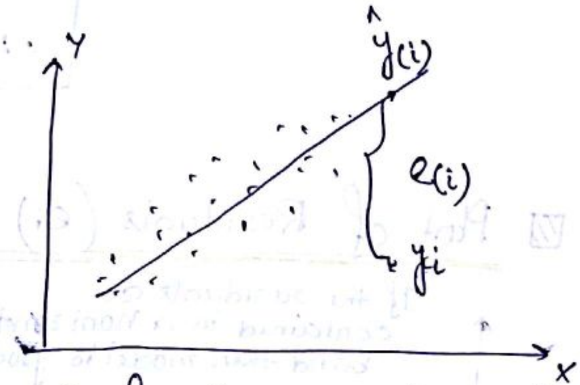
For influential observation h_{ii} is large ; $0 \le h_{ii} \le 1$, case of studentized residuals; for standardized residuals $h_{ii} = 0$.

PRESS Residual

i^{th} press residual $e_{(i)} = y_i - \hat{y}_{(i)}$

where, $\hat{y}_{(i)}$ is the fitted value of i^{th} response based on all all observations except i^{th} one.





We delete i^{th} observation (influential), fit the regression model to the remaining $(n-1)$ observations, and predict y_i.
It is possible to calculate PRESS residuals from the result of one single fit to all n obsn..

$$e_{(i)} = \frac{e_i}{1-h_{ii}}$$

Large PRESS residuals are useful in identifying obsn. where the model doesnot fit the data well.

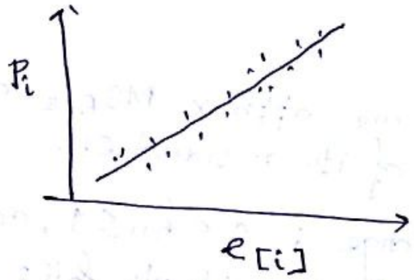
The PRESS statistic is $PRESS = \sum_1^n e_{(i)}^2 = \sum_1^n (y_i - \hat{y}_{(i)})^2$

$$= \sum \left(\frac{e_i}{1-h_{ii}}\right)^2$$

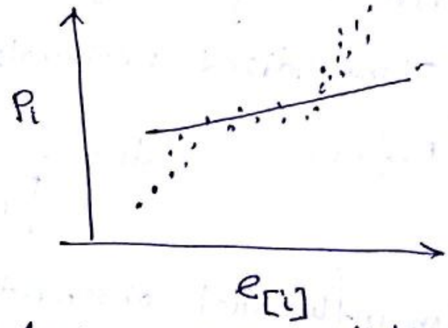
Residual Plots

Normal Probability Plot:- Let $e_1, e_2, \ldots, e_n$ be n residuals. Let $e_{[1]} < e_{[2]} < \cdots < e_{[n]}$ be the residuals ranked in increasing order.

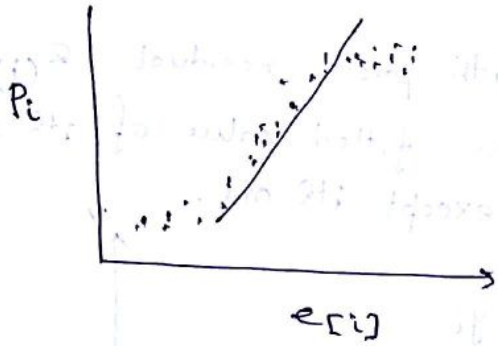
Plot $e_{[i]}$ Vs. cumulative probability $P_i = \frac{i - 1/2}{n} \;\forall\, i = 1(1)n$



(a) Normal distn.



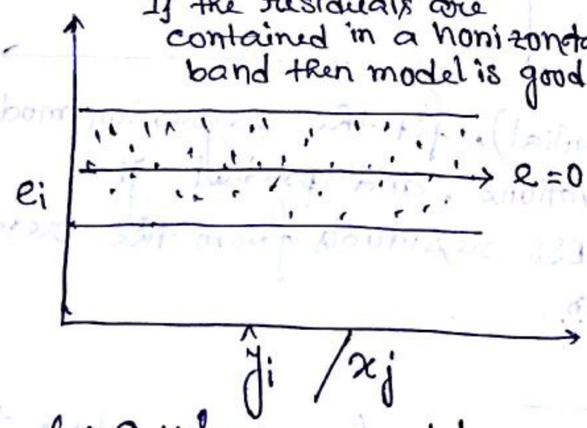
(b) heavy tailed distn.



(c) light-tailed distn.

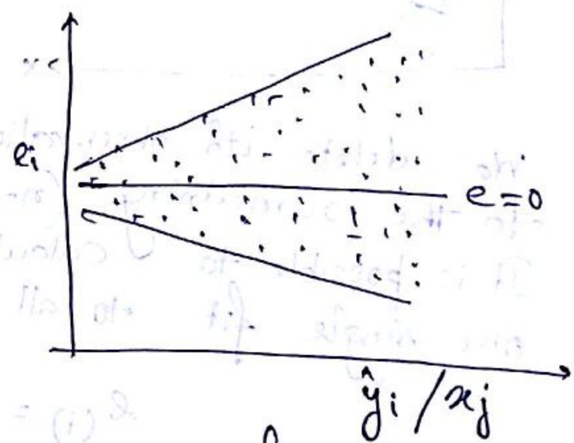
Plot of Residuals (e_i) Vs. fitted values ($\hat{y}_i$) / Regressor (x_j)

If the residuals are contained in a horizontal band then model is good.



(a) Satisfactory model

Good reg. model will produce a scatter in residuals that is roughly constant with $\hat{y}$ and centered about $e=0$



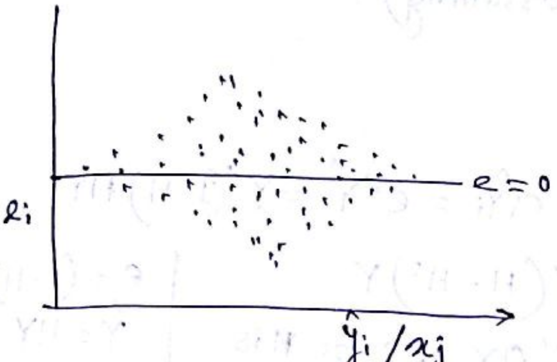
(b) Unsatisfactory model

Outward - opening Funnel

$V(\epsilon) \uparrow \quad y \uparrow$

For inward - open funnel

$V(\epsilon) \downarrow \quad y \uparrow$



(c) Double-bow
indicates non-constant variance, $V(\epsilon) = \sigma^2$
y is a proportion $0 \le y \le 1$

(d) Non-linear
Other regresson variables are needed in the model. Consider extra term (square term X^2) to the model or transform Y.

→ Why do we plot the residuals e_i against the $\hat{y}_i$ and not against y_i, for the usual linear model?

Ans. e_i and y_i are usually correlated
e_i and $\hat{y}_i$ " not "

SLR:
$$e_i = \beta_0 + \beta_1 y_i + \epsilon_i$$

LSE:
$$\hat{\beta}_1 = \frac{S_{ey}}{S_{yy}} = \frac{\sum(e_i - \bar{e})(y_i - \bar{y})}{\sum(y_i - \bar{y})^2} = \frac{\sum e_i(y_i - \bar{y})}{SS_T} = \frac{\sum e_i y_i}{SS_T}$$

$$= \frac{Y'e}{SS_T} = \frac{Y'(I-H)Y}{SS_T}$$

$$= \frac{Y'(I-H)(I-H)Y}{SS_T}\ ; \quad (I-H) \text{ is idempotent mtx.}$$

$$= \frac{e'e}{SS_T} = \frac{SS_{Res}}{SS_T}$$

$$= 1 - \frac{SS_{Reg}}{SS_T}$$

$$= 1 - R^2 \quad (R^2\text{: coefficient of multiple determination})$$

There is a linear relationship between y_i and e_i, since slope is $(1 - R^2)$.

Scanned by CamScanner

SLR: $e_i = \beta_0 + \beta_1 \hat{y}_i + \epsilon$ (assuming)

LSE $\hat{\beta} = \frac{S_{e\hat{Y}}}{S_{\hat{Y}\hat{Y}}}$

$$S_{e\hat{Y}} = \sum(e_i - \bar{e})(\hat{Y}_i - \hat{Y}) = \sum e_i \hat{Y}_i = e'\hat{Y} = Y'(I-H)HY$$

$$= Y'(H - H^2)Y$$

$$= Y'OY \text{ , since } H \text{ is idempotent}$$

$$= 0$$

$e = (I-H)Y$

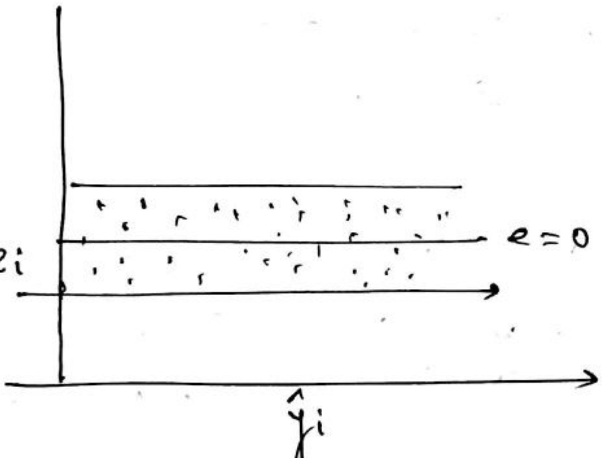
$\hat{Y} = HY$

So, e_i is not linearly related with $\hat{y}_i$, not with y_i.

In case of e_i and Y_i, $\hat{\beta}_1 = 1 - R^2$ (Show)

Unless $R^2 = 1$, there will be a slope of $(1 - R^2)$.

In e_i vs y_i plot, even if there is nothing wrong with the model,



In case of plotting the residuals vs regressors, it may not show the marginal effect of a regressor x_j, given the other regressors in the model.

So, next we will discuss Partial Residual Plot.

Show: $\hat{\beta}_1 = 1 - R^2$.

$\rightarrow$ $e = a + bY$

$Y = \beta_0 + \beta_1 x$

$\beta_1 = \frac{S_{xy}}{S_{xx}}$

$$b = \frac{e'Y}{Y'Y} = \frac{e'e}{Y'Y}$$

$$= \frac{SS_{Res}}{SS_T} = \frac{SS_T - SS_{Reg}}{SS_T} = 1 - R^2. \quad \left(\because R^2 = \frac{SS_{Reg}}{SS_T}\right)$$

Partial Residual Plot :-

Partial residual plot consider the marginal role of the reg. x_j given other reg. that are already in the model.

In this plot, the response variable and the reg. x_j (say) are both regressed ag. the other regressors in the model & residuals are obtained for each regression.

The plot of these residuals against each other show the marginal role of reg. x_j on response variable y in the presence of other regressors in the model.

Consider the MLR model with two reg. X_1 and X_2.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

We are interested in marginal role of X_1 on response variable Y in the response of other reg. in the model.

Regress Y on X_2 : $\hat{y}_{i(1)} = \hat{\theta}_0 + \hat{\theta}_1 x_{i2}$, $e^{y}_{i(1)} = y_i - \hat{y}_{i(1)}$

Regress X_1 on X_2 : $\hat{x}_{i1} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{i2}$, $e^{x}_{i1} = x_{i1} - \hat{x}_{i1}$

The partial residual of y for x_j is defined as

$$e^{y}_{i(j)} = y_i - \hat{y}_{i(j)}$$

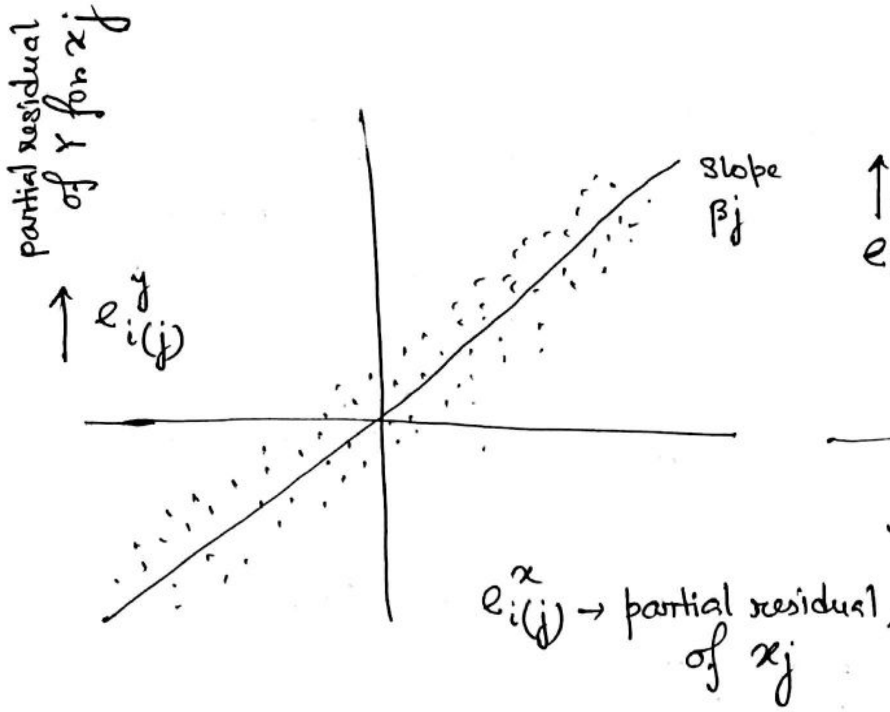
where $\hat{y}_{i(j)}$ is a prediction of y_i from a reg. model using all reg. except x_j.

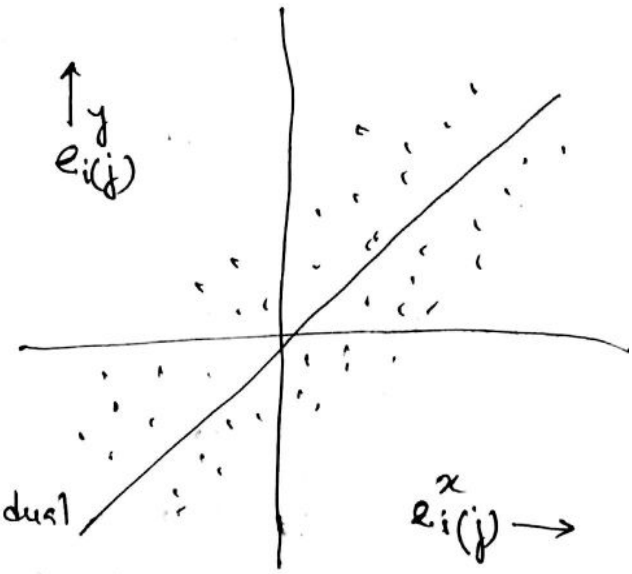
$e^{y}_{i(j)}$ represents the variability in y_i not explained by a model that excludes the regressor x_j.

The partial residual of x_j is defined as $e^{x}_{i(j)} = x_{ij} - \hat{x}_{i(j)}$

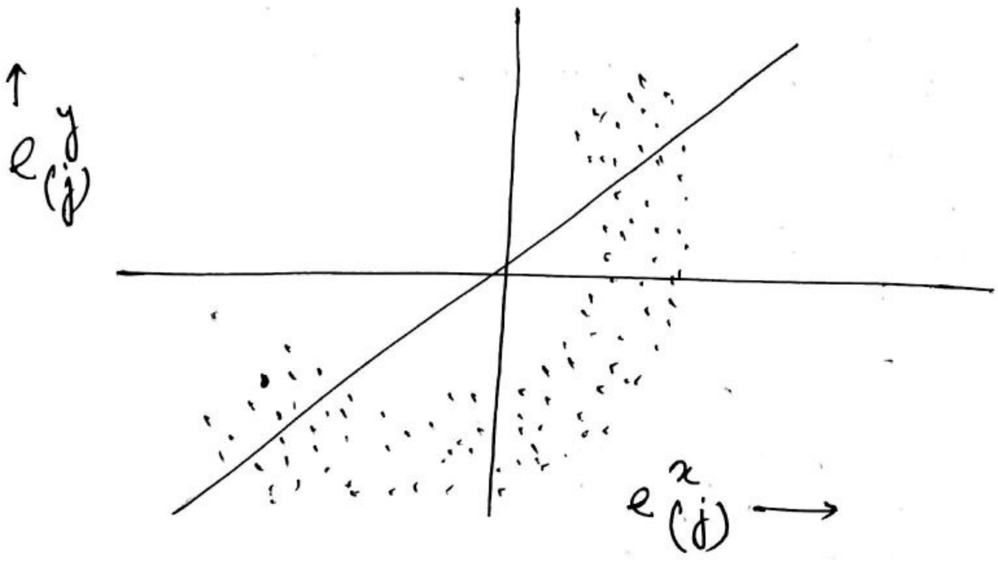
where, $\hat{x}_{i(j)}$ is a prediction of the reg. value x_{ij} from a regression of x_j on all other reg. variable.

$e^{x}_{i(j)}$ represents the variation in x_j that can't be explained by other regressors.





Partial residuals scatter around a line $y=\beta x$ is less scatter indicates strong relationship between x_j & Y.



<u>Curvilinear band</u>

higher order term in x_j or transformation such as $\left(\frac{1}{x_j}, \log x_j\right)$ may be helpful.

<u>Should influential obsn be discarded?</u>

If there is an error in recording the obsn., then it can be discarded.

Scanned by CamScanner

Determination of Influential Observation:-

Cook's statistic (D) for i^{th} observation is based on the diff. between predicted response $(\hat{Y})$ obtained using all the obs. and predicted response $\hat{Y}_{(i)}$ obtained without the i^{th} obsn.

$$D_i = \frac{(\hat{Y}_{(i)} - \hat{Y})'(\hat{Y}_{(i)} - \hat{Y})}{K\,MS_{Res}} \;;\quad \hat{Y} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix},\ \hat{Y}_{(i)} = \begin{pmatrix} \hat{y}_{(i)1} \\ \hat{y}_{(i)2} \\ \vdots \\ \hat{y}_{(i)n} \end{pmatrix}$$

$$= \frac{\sum (\hat{y}_{(i)j} - \hat{y}_j)^2}{K\,MS_{Res}}$$

Square Euclidean ~~[illegible]~~ distance between the vector of fitted values and vector of fitted values when i^{th} obsn is deleted.

$$D_1, D_2, \ldots, D_n$$

The value of D_i much larger than others indicates that i^{th} obsn. may be highly influential, preferably $D_i > 1$, is highly influential.

DFFITS (Difference between fit statistics) investigates deletion influence of the i^{th} observation on the fitted values.

For the i^{th} obsn. this statistic is defined as

$$DFFITS = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{MS_{Res}(i)\, h_{ii}}}$$

$\hat{y}_{(i)}$ is the fitted value of y_i obtained without the use of i^{th} obsn. $MS_{Res}(i)$ is the predicted value of MS_{Res} obtained without the use of i^{th} obsn.

A possible high influential observation is indicated by

$$|DFFITS_i| > 2\left(\frac{k}{n}\right)^{1/2}$$

DFBETAS : How much reg. coeff. $\hat{\beta}_j$ changes, if the i^{th} obsn. is deleted.

$$DFBETAS_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{MS_{Res}(i)\,(X'X)^{-1}_{jj}}}$$

$\hat{\beta}_{j(i)}$ is the j^{th} reg. coefficient computed without using i^{th} obsn.. As a general rule.

A possible high influential obsn. is indicated by

$$|DFBETAS_{ij}| > \frac{2}{\sqrt{n}}.$$

TRANSFORMATIONS AND WEIGHTING TO CORRECT MODEL INADEQUACIES

- Variance Stabilizing Transformations
- Transformations to linearize the model
- Analytical models to select a Transformation.
- Generalized & weighted Least squares.

___ The usual approach to deal with inequality of variance is to apply suitable transformation to the response variable or regressor variable.

Variance Stabilizing Transformation:-

$V(\epsilon) = \sigma^2$ constant variance assumption.

If the constant variance assumption is violated, the cause is often that the response variable y doesnot follow a Normal distn.

Ex. 1. $Y \sim \text{Poisson}(\lambda)$

$E(Y) = V(Y) = \lambda$

$Y' = \sqrt{Y}$, then you regress $Y' = \sqrt{Y}$ on x

$V(\sqrt{Y})$ is independent of mean λ.

Ex. 2. Y is a proportion $0 < Y < 1$

$Y' = \sin^{-1}(\sqrt{Y})$

Constant variance assumption is violated.

- Y has mean μ and variance σ^2

Situation:- $g(\mu) = \sigma^2$

$$U = f(Y) = f(\mu) + \frac{f'(\mu)}{1!}(Y-\mu)$$

$$V(U) = V(f(Y)) = [f'(\mu)]^2 V(Y) = [f'(\mu)]^2 g(\mu)$$

If we chose the function f $\ni$

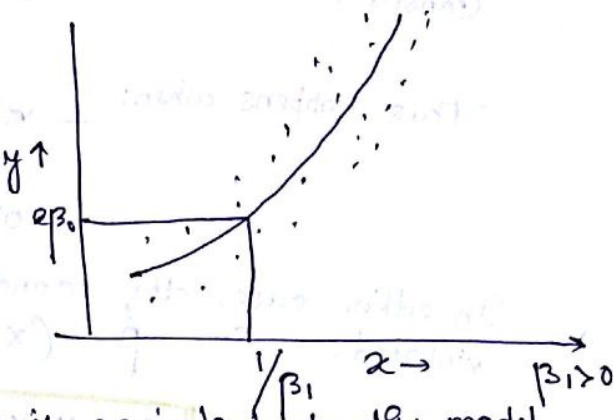
$$[f'(\mu)]^2 = \frac{1}{g(\mu)} \Rightarrow f'(\mu) = [g(\mu)]^{-2}$$

Then $V(U) = V(f(Y)) = 1$.

Scanned by CamScanner

Transformation to Linearize the model:-

~~Normality~~ Non-linearity may be detected via scatter plot/ residual plot.



Ex. If the scatter plot of y on x suggest an exponential relationship between x and y, then the appropriate model would be

$$y = \beta_0 e^{x\beta_1}$$

This model is linear, because it is equivalent to the model

$$\log y = \log\beta_0 + \beta_1 x.$$

$$\text{i.e. } y' = \beta_0' + \beta_1 x.$$

Weighted least Squares:- Linear regression model with non-constant variance can be fitted by the method of weighted least squares.

For SLR $\quad y = \beta_0 + \beta_1 x + \epsilon \qquad (x_i, y_i)$

The weighted least square function is

$$S = \sum w_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 = \sum w_i e_i^2 \; ; \quad \boxed{w_i \propto \frac{1}{\sigma_i^2}}$$

Normal equations:-

$$\frac{\partial S}{\partial \beta_0} = 0 \Rightarrow \sum w_i y_i = \hat{\beta}_0 \sum w_i + \hat{\beta}_1 \sum w_i x_i$$

$$\frac{\partial S}{\partial \beta_1} = 0 \Rightarrow \sum w_i y_i x_i = \hat{\beta}_0 \sum w_i x_i + \hat{\beta}_1 \sum w_i x_i^2$$

Gauss-Markov Theorem:- For reg. model (MLR) $Y = X\beta + \epsilon$ with $E(\epsilon) = 0$ and $V(\epsilon) = \sigma^2 I$, the LSEs are unbiased and minimum variance when compared with all other unbiased estimators that are linear combination of y_i's.

$$\text{LSE}: \quad \hat{\beta} = (X'X)^{-1} X'Y$$

LSEs are BLUE.

Generalised Least Squares :-

Weighted least square for multiple regression model.
Consider the same model: $Y = X\beta + \epsilon$ with $E(\epsilon) = 0$ &
$V(\epsilon) = V\sigma^2$ not, $I\sigma^2$; $V(\epsilon_i) = \sigma_i^2$
where V is a p.d. mtx.

This happens when: — observations Y have unequal variances
and/or
— observations are correlated.

In either case, the condition of Gauss-markov theorem are violated. So, $\hat{\beta} = (X'X)^{-1}X'Y$ is not BLUE.

~~Problems.~~

DUMMY VARIABLES

1. Dummy Variables to seperate Blocks of data
2. Interaction Terms involving Dummy variables

(1) → Suppose we wish to introduce into a model the idea that there are two types of machines (types A and B) that produces diff. levels of response, in addition to the variation that occurs due to the other regressors. $\boxed{Z = 0, 1}$

One way to do this is to add a dummy variable Z. Consider the simple model with one regressor variable X and one dummy variable Z.

Response: $y = \beta_0 + \beta_1 x + \alpha Z + \epsilon$

$Z = 0$ if the observation is from machine A

$Z = 1$ " " " " B

Let $\hat{\beta}_0, \hat{\beta}_1, \hat{\alpha}$ be LSE of β_0, β_1 & α respectively.

Fitted model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\alpha} Z$

Machine A data are estimated by setting $Z = 0$: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Machine B " " " " $Z = 1$: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\alpha}$

$\hat{\alpha}$ simply estimates the diff. in response level between machine A and B.

$$y = \beta_0 + \beta_1 x + \alpha z + \epsilon$$

$$Y = X\beta + \epsilon$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n_1+n_2} \end{pmatrix}, \quad X = \begin{array}{c} \begin{array}{ccc} x_0 & \text{obs } X's & z \end{array} \\ \left[\begin{array}{ccc} 1 & & 0 \\ 1 & & 0 \\ 1 & & 0 \\ \vdots & & \vdots \\ 1 & & 0 \\ \hline 1 & & 1 \\ 1 & & 1 \\ 1 & & 1 \\ \vdots & & \vdots \\ 1 & & 1 \end{array}\right] \end{array} \begin{array}{l} \text{Machine A} \\ n_1 \text{ obs}. \\ \text{Mac. B} \\ n_2 \text{ obs}. \end{array}$$

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \alpha \end{pmatrix}$$

Two Blocks require two dummy variables including x_0.

Three Blocks, Three dummy variables :-

$$(Z_1, Z_2) = (1, 0) \quad \text{for Machine A}$$
$$= (0, 1) \quad \text{for Machine B}$$
$$= (0, 0) \quad \text{for Machine C}$$

The model would be $Y = \beta_0 x_0 + \beta X + \alpha_1 Z_1 + \alpha_2 Z_2 + \epsilon$

$$\hat{\beta} = (X'X)^{-1}(X'Y)$$

Suppose the fitted equation is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta} X + \hat{\alpha}_1 Z_1 + \hat{\alpha}_2 Z_2$$

Machine A data are estimated by setting $(Z_1, Z_2) = (1, 0)$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta} X + \hat{\alpha}_1$$

" B " " " " " $(Z_1, Z_2) = (0, 1)$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta} X + \hat{\alpha}_2$$

" C " " " " " $(Z_1, Z_2) = (0, 0)$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta} X$$

$\hat{\alpha}_1$ estimates the diff. in response level between A & C

$\hat{\alpha}_2$ " " " " " " B & C

$\hat{\alpha}_1 - \hat{\alpha}_2$ " " " " " " A & B.

If desired, t test can be performed to test the diff. in response level between A & C.

$$H_0: \alpha_1 = 0 \quad \text{ag. } H_1: \alpha_1 \neq 0$$

$\hookrightarrow$ diff. in response model

Test statistic $t = \dfrac{\hat{\alpha}_1}{\sqrt{(X'X)^{-1}_{33} \, MS_{Res}}}$

Critical region: $|t| > t_{\alpha/2}$, Res d.f.

$H_0: \alpha_2 = 0$ ag. $H_1: \alpha_2 \neq 0$..
↳ diff. in response ~~model~~ level between B & C

Test statistic: $t = \dfrac{\hat{\alpha}_2}{\sqrt{(X'X)^{-1}_{44} MS_{Res}}}$

Critical region $|t| > t_{\alpha/2, \text{Res d.f.}}$

$H_0: \alpha_1 - \alpha_2 = 0$ Vs. $H_1: \alpha_1 - \alpha_2 \neq 0$
↳ diff. in response level between A & B

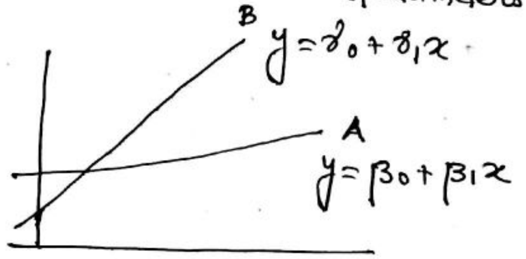
$t = \dfrac{\hat{\alpha}_1 - \hat{\alpha}_2}{\sqrt{V(\hat{\alpha}_1 - \hat{\alpha}_2)}}$; $V(\hat{\alpha}_1 - \hat{\alpha}_2) = V(\hat{\alpha}_1) + V(\hat{\alpha}_2) - 2\,Cov(\hat{\alpha}_1, \hat{\alpha}_2)$

Critical region ; $|t| > t_{\alpha/2, \text{Res df}}$

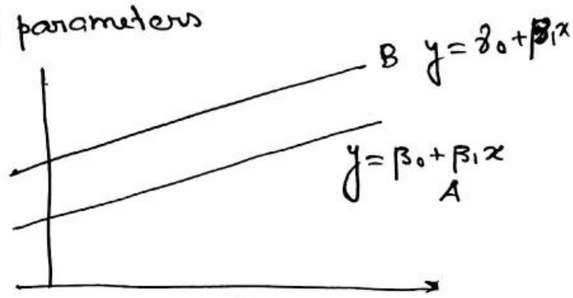
Interaction Terms Involving Dummy Variables

Two sets of data, straight line models.
Suppose A & B denote two sets of data and we are considerly fits involving straight lines. There are 4 possibilities;

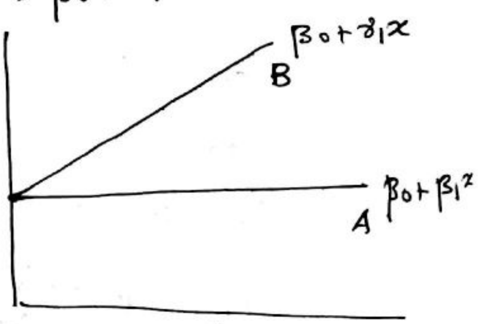
(a) Two distinct lines $\beta_0 + \beta_1 x$, $\gamma_0 + \gamma_1 x$ 4 parameters



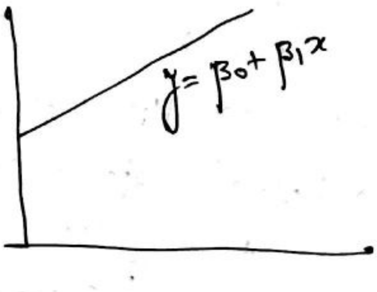
(b) Two parallel lines $\beta_0 + \beta_1 x$, $\gamma_0 + \beta_1 x$, 3 parameters



(c) Two lines with the same intercepts $\beta_0 + \beta_1 x$, $\beta_0 + \gamma_1 x$
3 parameters



(d) One line $\beta_0 + \beta_1 x$



We can take care of 4 possibilities at once by choosing two dummies, including X_0.

X_0	Z	
1	0	for A
1	1	for B

Then the model could be

$$Y = X_0(\beta_0 + \beta_1 x) + Z(\alpha_0 + \alpha_1 x) + \epsilon$$

$$= \beta_0 + \beta_1 x + \alpha_0 Z + \alpha_1 Z x + \epsilon \quad \text{——— } (*)$$

This model contains not only Z but an interaction term involving Z.

The seperate models for A & B are given by setting $Z=0$ & $Z=1$.

$$Y = \beta_0 + \beta_1 x \quad \text{for } A$$

$$= (\beta_0 + \alpha_0) + (\beta_1 + \alpha_1) x \quad \text{for } B$$

$$= \gamma_0 + \gamma_1 x$$

To test whether two parallel lines will do, i.e., to test the appropriateness of case (b) we could fit $(*)$ & then test.

$$H_0: \alpha_1 = 0 \text{ Vs. } H_1: \alpha_1 \neq 0$$

To test the appropriateness of the case (c) we could fit $(*)$ & then test

$$H_0: \alpha_0 = 0 \text{ Vs. } H_1: \alpha_0 \neq 0$$

To test the appropriateness of the case (d), we could test $H_0: \alpha_0 = \alpha_1 = 0$ Vs. H_1: H_0 is not true.

Three sets of data, straight line models:—

To allow the fitting of three seperate straight lines, we form the model: $Y = X_0(\beta_0 + \beta_1 X) + Z_1(\gamma_0 + \gamma_1 X) + Z_2(\delta_0 + \delta_1 X) + \epsilon$

$X_0 = 1$ & Z_1, Z_2 are two additional dummy variables.

	X_0	Z_1	Z_2
A →	1	1	0
B →	1	0	1
C →	1	0	0

$$Y = \beta_0 + \beta_1 X + \gamma_0 Z_1 + \gamma_1 X Z_1 + \delta_0 Z_2 + \delta_1 X Z_2 + \epsilon$$

Note that we have two interaction terms XZ_1 & XZ_2.

To test whether 3 lines are identical, we test

$$H_0: \gamma_0 = \gamma_1 = \delta_0 = \delta_1 = 0 \quad H_1: H_0 \text{ is not true.}$$

$$Y = (\beta_0 + \beta_1 X) + Z_1(\gamma_0 + \gamma_1 X) + Z_2(\delta_0 + \delta_1 X) + \epsilon.$$

42

$$F = \frac{\{SS_{Reg}(\text{Full model}) - SS_{Reg}(\text{Restricted Model})\}/4}{SS_{Res}/n-6} \sim F_{4,n-6}$$

$\rightarrow Y = \beta_0 + \beta_1 X$

Critical region: $F > F_{\alpha, 4, n-6}$

To test three lines are parallel,

$H_0: \gamma_1 = \delta_1 = 0$ Vs. H_1: H_0 is not true.

$\rightarrow Y = \beta_0 + \beta_1 X + \gamma_0 Z_1 + \gamma_1 Z_2 + \epsilon$

$$F = \frac{\{SS_{Reg}(\text{Full model}) - SS_{Reg}(\text{Restricted model})\}/2}{SS_{Res}/n-6} \sim F_{2,n-6}$$

$$F > F_{\alpha, 2, n-6}$$

Two sets of data, Quadratic Model:-

Suppose we have two sets of data on Y and X and we have in mind to model of the form

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \epsilon \qquad \boxed{Z = X^2}$$

We fit the model

$$Y = Z_0(\beta_0 + \beta_1 X + \beta_{11} X^2) + Z_1(\alpha_0 + \alpha_1 X + \alpha_{11} X^2) + \epsilon$$

Z_0	Z_1	
1	0	for A
1	1	for B

(1) Test: $H_0: \alpha_0 = \alpha_1 = \alpha_{11} = 0$ Vs H_1: H_0 is not true.
If H_0 is rejected then we conclude the models are not the same.

(2). If H_0 is (1) is rejected, test $H_0: \alpha_1 = \alpha_{11} = 0$ Vs. H_1: H_0 is not true.
If H_0 is accepted, we conclude that the two sets of data have the same slope & curvature.

(3.) If H_0 in (2) is rejected, then test $H_0: \alpha_{11} = 0$ Vs. $H_1: \alpha_{11} \neq 0$
Model differ only in zero & first order term.

Polynomial Regression Models

- Polynomial models in one variable — Orthogonal Polynomials — Piecewise Polynomial Fitting
- Polynomial models in two or more variables.

→ $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ is called second order model in one variance.

In general, k^{th} order polynomial in one variance is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon$$

Set $x_j = x^j$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

Then k^{th} order polynomial model in one variable becomes MLR model with k regressors $x_1, x_2, \ldots, x_k$.

- Order of the Polynomial :- $k \leq 2$
- Model building strategy :- Forward Selection: start with linear model

$$y = \beta_0 + \beta_1 x + \epsilon$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

Successively fit model of increasing orders unit the t-test for the highest order term is non-significant.

- ill-Conditioning :- As the order of the polynomial increases, the $(X'X)$ matrix becomes ill-contioned.

i.e., $(X'X)^{-1}$ calculation becomes inaccurate.

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

If the value of x are limited to a narrow range, there can be significant ill-conditioning problem in columns of X. Example:- $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon$

$$X = \begin{pmatrix} 1 & x & x^2 & \cdots \\ 1 & .11 & .0121 & \\ 1 & .12 & .0144 & \\ 1 & .13 & .0169 & \end{pmatrix} \qquad x^2 \approx 0.01 \times x$$

Centering the data may remove ill-conditioning.

We fit the model $Y = \beta_0 + \beta_1 (x - \bar{x}) + \beta_2 (x - \bar{x})^2 + \epsilon$

instead of $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$

Orthogonal Polynomials

Suppose we wish to fit the model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_K x^K + \epsilon$$

If we wish to add another term $\beta_{K+1} x^{K+1}$ we must recompute $(X'X)^{-1}$ and estimates of lower order parameters $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K$ will change.

$$X = \begin{pmatrix} 1 & x & x^2 & \cdots & x^K & x_{K+1} \\ & & & & & \end{pmatrix}$$

To avoid this problem we use orthogonal polynomials.

If we construct polynomials $P_0(x), P_1(x), \ldots, P_K(x)$ with the property that they are orthogonal polynomials

$$\sum_{i=1}^{n} P_r(x_i) P_s(x_i) = 0 \quad , r \neq s \; , r, s = 1(1)K.$$

We can rewrite the model as

$$y_i = \alpha_0 P_0(x_i) + \alpha_1 P_1(x_i) + \cdots + \alpha_K P_K(x_i) + \epsilon_i \,, \quad i = 1(1)n.$$

where, $P_r(x_i)$ is the r-th order orthogonal polynomial.

Example:- x values are equally spaced.

$$P_0(x_i) = 1$$

$$P_1(x_i) = \lambda_1\left(\frac{x_i - \bar{x}}{d}\right)$$

$$P_2(x_i) = \lambda_2\left[\left(\frac{x_i - \bar{x}}{d}\right)^2 - \left(\frac{n^2-1}{12}\right)\right]$$

where d is the spacing between the levels of x and λ_j are chosen so that the polynomial will have integer values.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_K x^K + \epsilon$$

$$\Downarrow$$

$$y = \alpha_0 P_0(x) + \alpha_1 P_1(x) + \alpha_2 P_2(x) + \cdots + \alpha_K P_K(x) + \epsilon$$

$$X = \begin{pmatrix} 1 & P_1(x_1) & P_2(x_1) & \cdots & P_K(x_1) \\ 1 & P_1(x_2) & P_2(x_2) & \cdots & P_K(x_2) \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & P_1(x_n) & P_2(x_n) & \cdots & P_K(x_n) \end{pmatrix}, \quad X'X = \begin{pmatrix} n & 0 & 0 & \cdots & 0 \\ 0 & \sum_1^n P_1^2(x_i) & 0 & \cdots & 0 \\ \vdots & & & & \\ 0 & 0 & \cdots & \cdots & \sum_1^n P_K^2(x_i) \end{pmatrix}$$

$$Y = X\alpha + \epsilon$$

LSE:- $\hat{\alpha} = (X'X)^{-1} X'Y$, $\quad \hat{\alpha}_0 = \dfrac{\sum y_i}{n} = \bar{y}$

$$\hat{\alpha}_j = \frac{\sum P_j(x_i)\, y_i}{\sum P_j^2(x_i)} \,, \quad j = 1(1)K.$$

Residual sum of square:-

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{1}^{n} (y_i - \hat{y}_i)^2 = (Y-\hat{Y})'(Y-\hat{Y})$$

$$= Y'Y - Y'X\hat{\alpha} \quad \text{MLR}$$

$$= \sum_{1}^{n} y_i^2 - \sum_{j=0}^{K} \hat{\alpha}_j \sum y_i P_j(x_i) \qquad \hat{\alpha}_0 = \bar{y}$$

$$= \sum_{1}^{n} y_i^2 - \hat{\alpha}_0 \sum y_i - \sum_{j=1}^{K} \hat{\alpha}_j \sum_{1}^{n} y_i P_j(x_i)$$

$$= \sum y_i^2 - n\bar{y}^2 - \sum_{j=1}^{K} \hat{\alpha}_j \sum_{1}^{n} y_i P_j(x_i)$$

$$= SS_T - \sum_{j=1}^{K} \hat{\alpha}_j \sum_{1}^{n} y_i P_j(x_i) = SS_T - SS_{Reg}$$

All sum of squares for $\alpha_1, \alpha_2, \ldots, \alpha_K$ are orthogonal & their values do not depend on the order of the polynomial.

ANOVA TABLE.

Source	df	SS	MS	F
α_1	1	$SS_{Reg}(\alpha_1)$	$MS_{Reg} = \frac{SS_{Reg}}{DF}$	
α_2	1	$SS_{Reg}(\alpha_2)$		$F = \frac{MS_{Reg}(\alpha_K)}{MS_{Res}}$
⋮	⋮	⋮	⋮	$\sim F_{1, n-K-1}$
α_K	1	$SS_{Reg}(\alpha_K)$		
Residual	$n-K-1$	SS_{Res}	$MS_{Res} = \frac{SS_{Res}}{n-K-1}$	
Total	$n-1$	SS_T		

$$SS_T = \sum_{1}^{n} (y_i - \bar{y})^2 \quad \because \sum (y_i - \bar{y}) = 0$$

Significance of highest order term α_K:-

$$H_0: \alpha_K = 0 \quad vs. \quad H_1: \alpha_K \neq 0$$

$$F = \frac{MS_{Reg}(\alpha_K)}{MS_{Res}} \sim F_{1, n-K-1}$$

Critical reg:- $F > F_{\alpha, 1, n-K-1}$

Piecewise Polynomial fitting (Splines).

This problem may occur when the function behavs diff. in different parts of the range of x.

Splines are piecewise polynomial of order k. The joint points of the pieces are called knots.

The cubic spline $(K=3)$ is usually adequate for most practical problems.

Cubic Spline :- A cubic spline with h knots, $t_1 < t_2 < \dots < t_h$ with the continuous first and second derivatives, can be written as

$$E(y) = S(x) = \sum_{j=0}^{3} \beta_{0j} x^j + \sum_{i=1}^{h} \beta_i (x - t_i)_+^3 \quad ;$$

$$(x-t_i)_+ = \begin{cases} (x-t_i), & x > t_i \\ 0, & x \le t_i \end{cases}$$

Deciding on the number and positions of the knots the order of the poly in each segment is not simple.

$\boxed{\text{Polynomial models in Two or more Variables}}$

Second order polynomial model in two variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$$

Linear effect parameters: β_1, β_2

Quadratic effect parameters: β_{11}, β_{22}

Interaction effect parameter: β_{12}

We usually call the reg. function

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2$$

Problem:- Fit a cubic equation using orthogonal polynomials to the Y-values 13, 4, 3, 4, 10, 22, which are equally spaced in the respective X-values given by X = −2.5, −1.5, −0.5, 0.5, 1.5, 2.5. Is the cubic term needed? If not what is the best quadratic fit.

If the model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$ has been fitted directly, how could the extra sum of squares $SS_{Reg}(\beta_1 | \beta_0) = 58.51$, $SS_{Reg}(\beta_2 | \beta_0, \beta_1) = 210.58$, $SS_{Reg}(\beta_3 | \beta_0, \beta_1, \beta_2) = 0.006$ relate to the sum of squares for the first, second and 3rd-order orthogonal polynomials.

Solution:-

X	Y
-2.5	13
-1.5	4
-0.5	3
0.5	4
1.5	10
2.5	12

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

$$Y = \alpha_0 + \alpha_1 P_1(x) + \alpha_2 P_2(x) + \alpha_3 P_3(x) + \epsilon$$

$$P_1(x) = \lambda_1\left(\frac{x_i - \bar{x}}{d}\right) \qquad \bar{x} = 3.5$$
$$= 2\left(\frac{1-3.5}{1}\right)$$

$P_0(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$
1	-5	5	-5
1	-3	-1	7
1	-1	-4	4
1	1	-4	-4
1	3	-1	-7
1	3	5	6

$$\hat{\alpha}_0 = \bar{y}, \quad \hat{\alpha}_j = \frac{\sum P_j(x_i) y_i}{\sum P_j^2(x_i)}$$

Test:- $H_0: \alpha_3 = 0$ Vs. $H_1: \alpha_3 \neq 0$

$$SS_{Reg}(\alpha_1) = \hat{\alpha}_1 \sum y_i P_1(x_i) = 58.51 \quad df\ 1$$

$$SS_{Reg}(\alpha_2) = \hat{\alpha}_2 \sum y_i P_2(x_i) = 210.58 \quad df\ 1$$

$$SS_{Reg}(\alpha_3) = \hat{\alpha}_3 \sum y_i P_3(x_i) = 0.006 \quad df\ 1$$

$$SS_{Reg} = \sum_1^3 SS_{Reg}(\alpha_i)$$

$SS_{Res} = 207.70 \quad df\ 2$ (because 6 x_i's, 4 α_i's)

$$F = \frac{SS_{Reg}(\alpha_3)/1}{SS_{Res}/2} = \frac{0.006/1}{207.7/2} = .000057 < F_{0.05,1,2} = 18.51$$

$\therefore$ α_3 is not significant.

So, $\hat{Y} = \hat{\alpha}_0 + \hat{\alpha}_1 P_1(x) + \hat{\alpha}_2 P_2(x)$ $\qquad P_1(x) = \lambda_1\left(\frac{x-\bar{x}}{d}\right)$

$$\hat{Y} = 4.273 + 1.8286x + 2.3750x^2$$

Here, $SS_{Reg}(\alpha_2) = SS_{Reg}(\beta_2 \mid \beta_0, \beta_1)$

$SS_{Reg}(\alpha_3) = SS_{Reg}(\beta_3 \mid \beta_0, \beta_1, \beta_2)$

Generalized Linear Model (GLM)

GLM analysis comes into account when errors distribution is not normal
and/or
when a vector of non-linear functions of the responses $\eta(Y) = (\eta(Y_1), \eta(Y_2), \ldots, \eta(Y_n))'$ are not Y itself, has expectation $X\beta$.

MLR: $Y = X\beta + \epsilon$ $\qquad E(\eta(Y)) = X\beta$

$E(Y) = X\beta$, $E(\epsilon) = 0$

In GLM, the response variable distribution must be a member of the exponential family.

The exponential family of distribution:-

A random variable u belongs to exponential family with single parameter θ has a pdf

$$f(u,\theta) = s(u)\, t(\theta)\, e^{a(u)\, b(\theta)}$$

where s, t, a, b are all known functions.

Rewrite: $f(u,\theta) = \exp\{a(u)\, b(\theta) + d(u) + c(\theta)\}$

where $d(u) = \ln(s(u))$, $c(\theta) = \ln(t(\theta))$

when $a(u) = u$, the distn is said to be in canonical form. $b(\theta)$ is called natural parameter, parameters other than the parameter of interest θ are called nuisance parameter.

Some member of exponential family:-

1. **Normal distribution:-**

$$f(\mu, x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}; -\infty \le x \le \infty$$

$$= \exp\left\{ x \cdot \frac{\mu}{\sigma^2} + \left[-\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln 2\pi\sigma^2\right] - \frac{\mu^2}{2\sigma^2} \right\}$$

where, $a(x) = x$, $b(\theta) = \frac{\mu}{\sigma^2}$, $c(\theta) = -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln 2\pi\sigma^2$, $d(\mu) = \frac{-\mu^2}{2\sigma^2}$

2. **Binomial distribution:-**

$$f(x, p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n$$

$$= \binom{n}{x}\left(\frac{p}{1-p}\right)^x (1-p)^n$$

$$= \exp\left\{ x\ln\left(\frac{p}{1-p}\right) + n\ln(1-p) + \ln\binom{n}{x} \right\}$$

where, $a(x) = x$, $b(\theta) = \ln\left(\frac{p}{1-p}\right)$, $c(\theta) = n\ln(1-p)$, $d(u) = \ln\binom{n}{x}$

Expected value and variance of $a(u)$:-

$$E(a(u)) = -\frac{c'(\theta)}{b'(\theta)}, \quad V(a(u)) = \frac{b''(\theta)\, c'(\theta) - c''(\theta)\, b'(\theta)}{[b'(\theta)]^3}$$

- Fitting GLM :- Suppose we have a set of independent obsn.s. $(Y_i, \tilde{x}_i')$, $i=1(1)n$, $\tilde{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ from some exponential type distribution of canonical form [i.e. $a(Y) = Y$]. The joint pdf is

$$f(Y_1, Y_2, \ldots, Y_n, \theta, \phi) = \exp\left\{ \sum_{i=1}^{n} Y_i b(\theta) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(Y_i) \right\}$$

where ϕ is a vector of a nuisance parameters that occur with in $b(\cdot), c(), \& d(\cdot)$.

$\theta = (\theta_1, \theta_2, \ldots, \theta_n)'$ vector of parameters of interest.

The variation in response variable (Y_i) can be explained in terms of $\tilde{x}_i$ values, $\tilde{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$

Consider the set of parameters $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$

We find some suitable link function $g(\cdot) \ni$

$$g(\mu_i) = \tilde{x}_i'\beta \qquad \Big| \qquad y_i = \tilde{x}_i'\beta + \epsilon$$

$$E(y_i) = \tilde{x}_i'\beta. \qquad \Big| \qquad E(y_i) = \tilde{x}_i'\beta$$

A link function that is often regarded as sensible one is natural parameter.

Example :- Binomial Distribution

Suppose we have data $(Y_i, \tilde{x}_i')$ from a binomial distn. $Bin(n_i, p_i)$. The single observation Y_i is of the form $\frac{r_i}{n_i}$, where r_i is the no. of success in n_i trials, each having prob. p_i of success & $\tilde{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is a set of observations of p regressors associated with Y_i. Binomial distn. is a member of exp. family.

Joint pdf $= f(Y_1, Y_2, \ldots, Y_n)$

$$= \prod_{i=1}^{n} \binom{n_i}{Y_i} p_i^{Y_i} (1-p_i)^{n_i - Y_i}$$

$$= \prod_{i=1}^{n} \exp\left\{ Y_i \ln\left(\frac{p_i}{1-p_i}\right) + n_i \ln(1-p_i) + \ln\binom{n_i}{Y_i} \right\}$$

$$= \exp\left\{ \sum_{i=1}^{n} Y_i \ln\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} n_i \ln(1-p_i) + \sum_{1}^{n} \ln\binom{n_i}{Y_i} \right\}$$

Natural parameter $= \ln\left(\frac{p_i}{1-p_i}\right)$

We would hope that the variation in the $Y_i \mid E(Y_i) = p_i$ could be explained in terms of $\tilde{x}_i'$ values, i.e., we could hope that we could find a suitable link function $g(\cdot) \ni$

$$g(p_i) = \tilde{x}_i'\beta$$

We fit the model $\ln\left(\frac{p_i}{1-p_i}\right) = \tilde{x}_i'\beta = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$

$$E(y_i) = p_i = \frac{\exp(\tilde{x}_i'\beta)}{1+\exp(\tilde{x}_i'\beta)}$$

Instead of fitting $y_i = \tilde{x}_i'\beta + \epsilon$ we fit here

$$y_i = p_i + \epsilon.$$

where $\tilde{x}_i'\beta = \beta_1 + \beta_2 x_{i2}$ —— (*) is called the logistic function.

Estimation via ML function:- To estimate β,

$$L = \exp\left\{ \sum_{1}^{n} y_i \ln\left(\frac{p_i}{1-p_i}\right) + \sum_{1}^{n} n_i \ln(1-p_i) + \sum_{1}^{n} \ln\binom{n_i}{y_i} \right\}$$

$$\ln L = \sum_{1}^{n} y_i \ln\left(\frac{p_i}{1-p_i}\right) + \sum_{1}^{n} n_i \ln(1-p_i) + \sum_{1}^{n} \ln\binom{n_i}{y_i}$$

$$= \sum y_i \tilde{x}_i'\beta - \sum_{1}^{n} n_i \ln\left(1+\exp(\tilde{x}_i'\beta)\right) + \sum_{1}^{n} \ln\binom{n_i}{y_i}$$

Maximize $\ln L$ w.r.t. β, use numerical search/ iteratively reweighted least square (IRLS) could be used to compute MLE of β.

Choice of Link function:-

Normal : $g(\mu) = \mu$ (Identity link)

Binomial :- $g(p) = \ln\left(\frac{p}{1-p}\right)$ (logistic link)

Poisson :- $g(\mu) = \ln \mu$ (log link)

Gamma/Exponential : $g(\mu) = \frac{1}{\mu}$ (reciprocal link)

Non-linear Estimation

- Linear models and non-linear models.
- Least squares in non-linear model.

Linear Models:- Models that are linear in parameters are called linear model.

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_{k-1} Z_{k-1} + \epsilon$$

where Z_i is any function of the basic regressors $X_1, X_2, \ldots, X_{p-1}$

$$Y = \beta_0 + \beta_1 (X_1 - X_2) + \beta_2 (X_1 - X_2)^2 + \epsilon$$ is a linear model.

Non-linear Models:- Models that are non-linear in parameters.

$$Y = e^{\theta_1 + \theta_2 t} + \epsilon \quad \text{——— (1)}$$

$$Y = \frac{\theta_1}{\theta_1 - \theta_2}\left[e^{-\theta_2 t} - e^{-\theta_1 t}\right] + \epsilon \quad \text{——— (2)}$$

t: regressor variable, θ: parameter.

(1) can be transformed to $\ln Y = \theta_1 + \theta_2 t + \epsilon$ linear model.

$$Y = f(t_1, t_2, \ldots, t_k ; \theta_1, \theta_2, \ldots, \theta_p) + \epsilon$$

write $\tilde{t} = (t_1, t_2, \ldots, t_k)'$, $\tilde{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$

$$Y = f(\tilde{t}, \tilde{\theta}) + \epsilon$$

or $E(Y) = f(\tilde{t}, \tilde{\theta})$ if we assume $E(\epsilon) = 0$, $V(\epsilon) = \sigma^2$;

$$\epsilon \overset{ind}{\sim} N(0, \sigma^2)$$

Suppose we have n observations $(Y_u, \tilde{t}_u)$, $u = 1(1)n$.

$$Y_u = f(\tilde{t}_u, \theta) + \epsilon_u$$

Error/Residual sum of squares $S(\theta) = \sum (Y_u - f(\tilde{t}_u, \tilde{\theta}))^2$

To find LSE's $\hat{\theta}$, we need to differentiate $S(\theta)$ w.r.t. $\tilde{\theta}$

p normal equations are

$$\sum \left(Y_u - f(\tilde{t}_u, \tilde{\theta})\right) \frac{\partial f(\tilde{t}_u, \tilde{\theta})}{\partial \theta_i}\Bigg|_{\theta = \hat{\theta}} = 0$$

where $f(\tilde{t}_u, \tilde{\theta})$ is linear $\frac{\partial f(\tilde{t}_u, \tilde{\theta})}{\partial \theta_i}$ is a functn of $\tilde{t}_u$ only and indep. of $\tilde{\theta}$.

When the model is non-linear in θ's, so will be the normal equations.

<u>Example:-</u> $Y = f(\theta, t) + \epsilon$ where $f(\theta, t) = e^{-\theta t}$

$Y = e^{-\theta t} + \epsilon$, $S(\theta) = \sum_u (Y_u - e^{-\theta t_u})^2$

Single normal equation:

$$\frac{\partial S(\theta)}{\partial \theta} = 0 \Rightarrow \sum_u (y_u - e^{-\theta t_u}) t_u e^{-\theta t_u} = 0.$$

$\rightarrow$ Finding $\hat{\theta}$ is not easy.

<u>Estimation of parameters of a non linear systems:-</u>

$$y_u = f(\tilde{t}_u, \tilde{\theta}) + \epsilon_u$$

Let $\theta_{10}, \theta_{20}, \dots, \theta_{p0}$ be initial values for the parameters $\theta_1, \theta_2, \dots, \theta_p$. Carrying out Taylor expansion of $f(\tilde{t}_u, \tilde{\theta})$ about the point $\tilde{\theta}_0 = (\theta_{01}, \theta_{02}, \dots, \theta_{0p})'$

Taylor series of a real or complex function $f(x)$ that is infinitely diff. in a neighbourhood of a real/complex no. 'a' is

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \dots$$

$$f(\tilde{t}_u, \tilde{\theta}) = f(\tilde{t}_u, \tilde{\theta}_0) + \sum_{i=1}^{p} \left. \frac{\partial f(\tilde{t}_u, \tilde{\theta})}{\partial \theta_i} \right|_{\theta_i = \theta_{i0}} (\theta_i - \theta_{i0})$$

Set $f_u^0 = f(\tilde{t}_u, \tilde{\theta}_0)$

$\beta_i^0 = (\theta_i - \theta_{i0})$

$$Z_{iu}^0 = \left. \frac{\partial f(\tilde{t}_u, \tilde{\theta})}{\partial \theta_i} \right|_{\theta_i = \theta_{i0}}$$

$$f(\tilde{t}_u, \tilde{\theta}) = f_u^0 + \sum_{i=1}^{p} Z_{iu}^0 \beta_i^0$$

$$Y_u = f(\tilde{t}_u, \tilde{\theta}) + \epsilon_u$$

$$Y_u - f_u^0 = \sum Z_{iu}^0 \beta_i^0 + \epsilon_u$$ is a linear model in β_i^0.

If we write

$$Z_0 = \begin{bmatrix} Z_{11}^0 & \cdots & Z_{p1}^0 \\ \cdots & \cdots & \cdots \\ Z_{1n}^0 & \cdots & Z_{pn}^0 \end{bmatrix}, \quad \tilde{\beta}_0 = \begin{pmatrix} \beta_1^0 \\ \beta_2^0 \\ \vdots \\ \beta_p^0 \end{pmatrix}, \quad \tilde{y}_0 = \begin{pmatrix} y_1 - f_1^0 \\ \vdots \\ y_n - f_n^0 \end{pmatrix}$$

$\tilde{y}_0 = Z_0 \tilde{\beta}_0 + \epsilon$ then the estimate of $\tilde{\beta}_0$ is given by

$$\tilde{b}_0 = \hat{\tilde{\beta}}_0 = (Z_0' Z_0)^{-1} Z_0' y_0$$

The vector b_0 minimize $\sum \left(y_u - f_u^0 - \sum \beta_i^0 Z_{iu}^0 \right)^2$ w.r.t. β_i^0, $i = 1(1)p$ where $\beta_i^0 = \theta_i - \theta_{i0}$, $i = 1(1)p$.

Let us write $b_i^0 = \theta_{i1} - \theta_{i0}$

$\theta_{i1} = \theta_{i0} + b_i^0$ is the revised best estimate of θ.

We can now place θ_{i1} in the same role as θ_{i0} and go through the same procedure. This will lead to another revised estimate θ_{i2}, and so on.

$$\theta_{j+1} = \theta_j + b_j = \theta_j + (Z_j' Z_j)^{-1} Z_j' (Y - f^j)$$

where $Z_j' = ((Z_{iu}^j))$ $\qquad Z_{iu} = \left. \dfrac{\partial f(\tilde{t}_u, \tilde{\theta}_i)}{\partial \theta_i} \right|_{\theta = \theta_j}$

$$f^j = (f_1^j, f_2^j, \ldots, f_n^j)'$$

$$\theta_j = (\theta_{1j}, \theta_{2j}, \ldots, \theta_{pj})'$$

This iterative process continue until

$$\left| \theta_{i(j+1)} - \theta_{i(j)} \right| < \delta = 0.00001 \text{ (let)}$$

Regression Models with Autocorrelated Errors

- Source & Effect of Autocorrelation
- Detecting the presence of Autocorrelation
- Parameter Estimation Models (Not discussed)

— Errors are autocorrelated / serially correlated means correlation between errors s steps apart are always the same. i.e., $\text{Corr}(\epsilon_t, \epsilon_{t+s}) = \rho_s$, $s = 1, 2, 3, \ldots$
Correlation between residuals one (or two or three) steps apart is called lag−1 (or 2 or 3) serial correlation.

— **Source of Autocorrelation:-** primary cause of autocorrelation in regression problem involving time series data is failure to include one or more important regressor(s) in the model.

— **Effect of Autocorrelation:-**

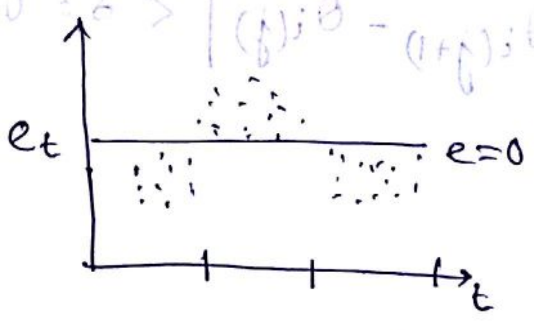
$$Y = X\beta + \epsilon \quad ; \quad \text{Cor}(\epsilon_i, \epsilon_j) \neq 0$$

$$\text{LSE}: \quad \hat{\beta} = (X'X)^{-1} X'Y \qquad \boxed{\text{Errors are correlated}}$$

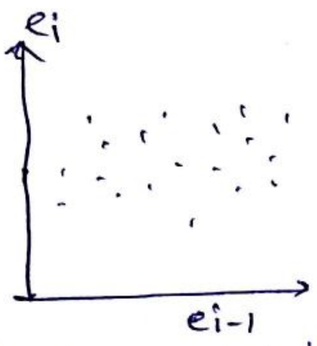
$\hat{\beta}$ is unbiased but not has minimum variance.

$\therefore$ $\hat{\beta}$ is not BLUE. $\qquad V(\epsilon) = \sigma^2 V \neq \sigma^2 I$

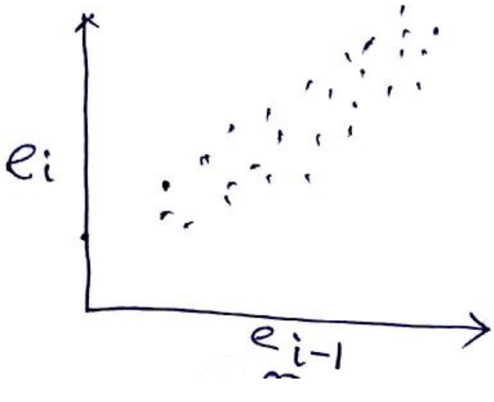
— **Detecting Autocorrelation:-** Residual plot is useful.



Residual of identical sign occur in a cluster then it implies positive autocorrelation.



ϵ_i's uncorrelated



lower left to upper right pattern indicates positive lag−1 autocorrelation.

$e_i = \hat{\rho} e_{i-1}$

The Durbin-Watson test:-

Suppose we wish to fit the model $y_u = \beta_0 + \sum \beta_i X_{iu} + \epsilon_u$ $u = 1(1)n$

by LS to obsn $(y_u, X_{1u}, X_{2u}, \ldots, X_{ku})$

We usually assume $\epsilon_u \overset{iid}{\sim} N(0, \sigma^2)$; $\rho_s = 0$.

We want to see if this assumption is justified.

$H_0: \rho_s = 0$ Vs. $H_1: \underset{\downarrow}{\rho_s = \rho^s}$ $(\rho \neq 0, |\rho| < 1)$

comes from the assumption that

$\epsilon_u = \rho \epsilon_{u-1} + z_u$ first order autoregressive errors.

where, $z_u \sim N(0, \sigma^2)$ & is independent of $\epsilon_{u-1}, \epsilon_{u-2}, \ldots$
& of $z_{u-1}, z_{u-2}, \ldots$

$$\begin{aligned}
\epsilon_u &= \rho \epsilon_{u-1} + z_u \\
&= \rho(\rho \epsilon_{u-2} + z_{u-1}) + z_u \\
&= \rho^2 \epsilon_{u-2} + \rho z_{u-1} + z_u \\
&= \rho^2(\rho \epsilon_{u-3} + z_{u-2}) + \rho z_{u-1} + z_u \\
&= \rho^3 \epsilon_{u-3} + \rho^2 z_{u-2} + \rho z_{u-1} + z_u \\
&= \sum_{k=0}^{u} \rho^k z_{u-k}
\end{aligned}$$

$E(\epsilon_u) = 0$; $V(\epsilon_u) = (1 + \rho^2 + \rho^4 + \cdots)\sigma^2$

$$= \frac{\sigma^2}{1-\rho^2}$$

$$Cov(\epsilon_u, \epsilon_{s+u}) = \rho^{|s|} \cdot \sigma^2 \cdot \frac{1}{1-\rho^2}$$

$$Corr(\epsilon_u, \epsilon_{s+u}) = \rho^{|s|}$$

$$\epsilon_u \sim N\left(0, \frac{\sigma^2}{1-\rho^2}\right)$$

Under $H_0: \rho = 0$, $\epsilon_u \overset{ind.}{\sim} N(0, \sigma^2)$.

$H_0: \rho_s = 0$ Vs. $H_1: \rho_s = \rho^{|s|}$

To test we fit the model $Y = X\beta + \epsilon$ & compute the residuals e_i, then Durbin-Watson statistic is

$$d = \frac{\sum_{u=2}^{n} (e_u - e_{u-1})^2}{\sum_{u=1}^{n} e_u^2}$$

The distn. of d lies between 0 & 4 and symmetric about 2.

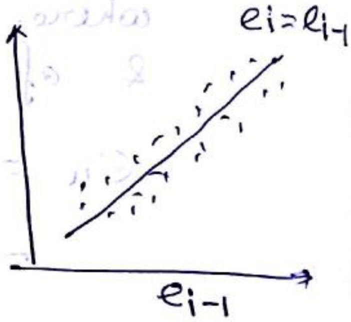
① <u>One sided test against the alternative:-</u>

$H_0: \rho = 0$ Vs $H_1: \rho > 0$

If $d < d_L$ reject H_0

If $d > d_U$ accept H_0

If $d_L < d < d_U$, test is inconclusive.



possitive autocorrelation indicates successive error term are of similar magnitude & the diff. in residuals $e_i - e_{i-1}$ will be small.

② $H_0: \rho = 0$ Vs. $H_1: \rho < 0$

If $4 - d < d_L$ reject H_0

If $4 - d > d_U$ accept H_0

If $d_L < 4 - d < d_U$, test is inconclusive.

③ $H_0: \rho = 0$ Vs. $H_1: \rho \neq 0$

If $d < d_L$ or $4 - d < d_L$ reject H_0

If $d > d_U$ or $4 - d > d_U$ accept H_0

Otherwise test is inconclusive.

For $n = 20$, $d_L = 1.20$, $d_U = 1.41$

$\alpha = 0.05$

Measurement errors & Calibration Problem

- Case where Response & Regressors are jointly distributed RVs.
- Measurement Error in Regressors
- The Calibration problem (inverse problem) [Not discussed]

(1) X & Y are jointly normally distd.

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{y-\mu_1}{\sigma_1}\right)\left(\frac{x-\mu_2}{\sigma_2}\right)\right]\right\}$$

$E(Y) = \mu_1$, $V(Y) = \sigma_1^2$, $E(X) = \mu_2$, $V(X) = \sigma_2^2$

$$\rho = \frac{E[(Y-\mu_1)(X-\mu_2)]}{\sigma_1\sigma_2} = \frac{\sigma_{12}}{\sigma_1\sigma_2}$$ is correlation coefficient between Y and X.

The conditional distn. of Y given X is

$$Y|x \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x-\mu_2),\ (1-\rho^2)\sigma_1^2\right)$$

$$E(Y|x) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x-\mu_2)$$

$$E(Y|x) = \beta_0 + \beta_1 x$$ where $\beta_0 = \mu_1 - \rho\mu_2\frac{\sigma_1}{\sigma_2}$, $\beta_1 = \rho\frac{\sigma_1}{\sigma_2}$

$$E(Y) = \beta_0 + \beta_1 x$$

$$Y = \beta_0 + \beta_1 x + \epsilon \ ; \ \epsilon \sim N(0, \sigma^2)$$

$$Y_i | x_i \overset{ind}{\sim} N\left(\beta_0 + \beta_1 x_i,\ (1-\rho^2)\sigma_1^2\right)$$

MLE:- $$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}} e^{-\frac{1}{2\sigma_1^2(1-\rho^2)}(y_i - \beta_0 - \beta_1 x_i)^2}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}}\right)^n e^{-\frac{1}{2\sigma_1^2(1-\rho^2)}\sum(y_i - \beta_0 - \beta_1 x_i)^2}$$

We find β_0 & β_1 $\ni$ $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ is minimum.

$$\hat\beta_0 = \bar y - \hat\beta_1 \bar x \quad \text{and} \quad \hat\beta_1 = \frac{\sum(y_i - \bar y)(x_i - \bar x)}{\sum(x_i - \bar x)^2} = \frac{S_{xy}}{S_{xx}}$$

identical to those given by LSE in case where x is a controlled variable.

$\rho = \text{Corr}(X, Y)$; estimate of ρ is $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$

$H_0: \rho = 0$ Vs $H_1: \rho \neq 0$

Test statistic, $t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$, under H_0.

Reject H_0 if $|t_0| > t_{\alpha/2, n-2}$.

Measurement errors in Regressors :-

We wish to fit the simple linear reg. model, but the regressor is measured with error.

$$X_i = x_i + a_i \quad ; \quad E(X_i) = x_i \quad ; \quad E(a_i \epsilon_i) = 0$$

X_i: Observed value; x_i: true value; a_i: measurement error

with $E(a_i) = 0$ and $V(a_i) = \sigma_a^2$

The response variable is subject to the usual error ϵ_i, $i = 1(1)n$.

The reg. model is
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$= \beta_0 + \beta_1 (X_i - a_i) + \epsilon_i$$
$$= \beta_0 + \beta_1 X_i + (\epsilon_i - \beta_1 a_i)$$
$$= \beta_0 + \beta_1 X_i + \gamma_i \quad ; \quad \gamma_i = \epsilon_i - \beta_1 a_i$$

$$Cov(X_i, \gamma_i) = E\left[(X_i - E(X_i))(\gamma_i) \right]$$
$$= E\left[(X_i - x_i)(\epsilon_i - \beta a_i) \right]$$
$$= E\left[a_i(\epsilon_i - \beta_1 a_i) \right]$$
$$= -\beta_1 E(a_i^2)$$
$$= -\beta_1 \sigma_a^2$$

If we apply standard LSM to the data, the estimates of the model parameters are no longer unbiased.

$$\hat{\beta}_1 = \frac{\sum (y_i - \bar{y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$$

$$E(\hat{\beta}_1) = \frac{\beta_1}{1+\theta}$$

where $\theta = \dfrac{\sigma_a^2}{\sigma_x^2}$,

$$\sigma_x^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

$\hat{\beta}_1$ is a biased estimator of β_1, unless $\sigma_a^2 = 0$, that is there is no measurement error in regressors.

If σ_a^2 is small relative to σ_x^2, the bias will be small.

If variability in the measurement errors is small, relative to the variability of the x's, then measurement error can be ignored & OLS method can be applied.

PROBLEMS & SOLUTIONS

1. A study was made on the effect of temperature on the yield of a chemical process, the following data were collected:

X	−5	−4	−3	−2	−1	0	1	2	3	4	5
Y	1	5	4	7	10	8	9	13	14	13	18

(a) Assuming a model, $Y = \beta_0 + \beta_1 X + \epsilon$, what are the least squares equation of β_0 and β_1? What is the fitted equation?

(b) Construct the ANOVA table and test the hypothesis $H_0: \beta_1 = 0$ with $\alpha = 0.05$

(c) What are the confidence limits $(\alpha = 0.05)$ for β_1?

(d) What are the confidence limits $(\alpha = 0.05)$ for the true mean value of Y when $X = 3$?

Solution:- (a) (x_i, y_i), $i = 1(1)11$

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

$$S = \sum_{i=1}^{11} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{158}{110} = 1.44$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{102}{11} = 9.27$$

$\boxed{\hat{y}_i = 9.27 + 1.44 x_i}$ is the fitted model.

(b) ANOVA TABLE

SV	DF	SS	MS	F
Reg	1	226.94	226.94	$96.17 > F_{0.05,1,9} = 5.12$
Residual	9	22.23	2.36	
Total	10	248.18		

$$SS_T = \sum_{i=1}^{11} (y_i - \bar{y})^2 = 248.18$$

$$SS_{Reg} = \hat{\beta}_1^2 S_{xx} = \left(\frac{158}{110}\right)^2 \times 110 = 226.94$$

$$= \sum_{i=1}^{11} e_i^2$$

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$. Here, Reject H_0.

$\therefore$ Y & X are linearly related.

Scanned by CamScanner

(c) $$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \sim t_{n-2}$$

$$P_r\left\{-t_{\alpha/2, n-2} \le \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \le t_{\alpha/2, n-2}\right\} = 1-\alpha \text{ ;}$$

$$\hat{\beta}_1 - t_{\alpha/2, n-2}\sqrt{\frac{MS_{Res}}{SS_x}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2, n-2}\sqrt{\frac{MS_{Res}}{S_{xx}}} \text{ ;}$$

$$1.44 - 2.263 \times 0.146 \le \beta_1 \le 1.44 + 2.263 \times 0.146$$

$$\Rightarrow 1.11 \le \beta_1 \le 1.77$$

(d) $$E(Y \mid X = 3)$$

95% CI for $E(Y \text{ at } X = x_0)$ is $\beta_0 + \beta_1 x_0$

An unbiased estimator $E(Y \text{ at } x = x_0)$ is $\hat{\beta}_0 + \hat{\beta}_1 x_0$

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N\left(\beta_0 + \beta_1 x_0 + \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$$

$$A = \frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{\sqrt{MS_{Res}\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}} \sim t_{n-2}$$

$$\therefore P_r\left\{-t_{\alpha/2, n-2} \le A \le t_{\alpha/2, n-2}\right\} = 1-\alpha$$

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2, n-2} * \sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$ is the

Confidence Interval for $\beta_0 + \beta_1 x_0$.

$\boxed{2}$. Fit the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ the data below, provide an ANOVA table, and perform the partial F-tests to test $H_0: \beta_i = 0$ Vs. $H_1: \beta_i \neq 0$ for $i = 1, 2$, given the other variable is already in the model. Comment on the relative contributions of the variables X_1 & X_2, depending on whether they enter the model first or second.

X_1	-5	-4	-1	2	2	3	3
X_2	5	4	1	-3	-2	-2	-3
Y	11	11	8	2	5	5	4

Solution:-

(a) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

Fit the model with both X_1, X_2

$$X = \begin{array}{c} \begin{matrix} 1 & X_1 & X_2 \end{matrix} \\ \begin{pmatrix} 1 & -5 & 5 \\ 1 & -4 & 4 \\ 1 & -1 & 1 \\ 1 & 2 & -3 \\ 1 & 2 & -2 \\ 1 & 3 & -2 \\ 1 & 3 & -3 \end{pmatrix} \end{array} \qquad \hat{\beta} = (X'X)^{-1} X'Y$$

$$= \begin{pmatrix} 7 & 0 & 0 \\ 0 & 68 & -67 \\ 0 & -67 & 68 \end{pmatrix}^{-1} \begin{pmatrix} 46 \\ -66 \\ 69 \end{pmatrix} = \begin{pmatrix} 46/7 \\ 1 \\ 2 \end{pmatrix}$$

$$\therefore \hat{Y} = \frac{46}{7} + X_1 + 2X_2$$

ANOVA Table

SV	DF	SS	MS	F
Regression	2	72.00	36.00	83.72 $F \sim F_{2,4}$
Residual	4	1.71	0.43	
Total	6	73.71		

$SS_T = \sum (y_i - \bar{y})^2 = 73.71$, $e_i = Y_i - \hat{Y}_i$, $SS_{Res} = \sum_{i=1}^{n} e_i^2$

$H_0: \beta_1 = \beta_2 = 0$ Vs. $H_1: H_0$ is not true

$F = 83.72 > F_{0.05, 2, 4} = 6.94$ | H_0 is rejected.

Scanned by CamScanner

62

Partial F-test:-

$H_0: \beta_2 = 0$ Vs $H_1: \beta_2 \neq 0$ $\quad Y = \beta_0 + \beta_1 X_1 + \epsilon$

Fit the model with X_1 alone:

$$\hat{Y} = \frac{46}{7} - \frac{66}{68} X_1$$

$$F = \frac{\{SS_{Reg}(\text{Full}) - SS_{Reg}(\text{Restricted model})\}/1}{MS_{Res}}$$

$$= \frac{72 - 64.06}{0.43} = 18.53 > F_{0.05,1,4} = 7.71$$

$\therefore$ H_0 is rejected.

Partial F-test:-

$H_0: \beta_1 = 0$ Vs. $H_1: \beta_1 \neq 0$ $\quad Y = \beta_0 + \beta_2 X_2 + \epsilon$

Fit the model with X_2 alone:

$$\hat{Y} = \frac{46}{7} + \frac{69}{68} X_2$$

$$F = \frac{\{SS_{Reg}(\text{Full}) - SS_{Reg}(\text{Restd. model})\}/1}{MS_{Res}}$$

$$= \frac{72 - 70.01}{0.43}$$

$$= 4.64 < F_{0.05,1,4} = 7.71$$

$\therefore$ H_0 is accepted.

Implication:- If X_2 is in model, we donot need X_1.
If X_1 is in model, X_2 helps out significantly.

Then X_2 is clearly has more useful variance & it explains

$$R^2 = \frac{70.01}{73.71} = 95\%$$ of the variability is Y about mean, where as X_1 explains $R^2 = \frac{64.06}{73.71} = 86\%$ of total variability in Y about mean.

And X_1 and X_2 together explain $\frac{72.00}{73.71} = 97\%$ of total variability

(63)

$\boxed{3}$ Can we use the data below to get a unique fit to the model $Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

X_1	−4	3	1	4	−3	−1
X_2	1	2	3	4	5	6
X_3	3	−5	−4	−8	−2	−5
Y	7.4	14.7	13.9	18.2	12.1	14.8

Solution:- $Y = X\beta + \epsilon$, $\hat{\beta} = (X'X)^{-1} X'Y$

$$X = \begin{array}{c} \begin{matrix} X_0 & X_1 & X_2 & X_3 \end{matrix} \\ \left[\begin{matrix} & & & \end{matrix} \right] \end{array}$$

$X_1 + X_2 + X_3 = 0$

$(X'X)$ is singular,

'NO'

$\boxed{4}$ Show that in GLR situation with a β_0 term in the model:

(a) the correlation between the vector e and Y is $(1-R^2)^{1/2}$. The implication of this result is that it is a mistake to attempt to find defective regressions by a plot of residuals e_i versus observations Y_i as this always shows a slope.

(b) Show that this slope is $1-R^2$.

(c) Show further that the correlation between e and $\hat{Y}$ is zero.

Solution:- (a) $\text{Corr}(e, Y) = \dfrac{\sum (e_i - \bar{e})(Y_i - \bar{Y})}{\sqrt{\sum (e_i - \bar{e})^2 \sum (Y_i - \bar{Y})^2}}$

$\sum_{i=1}^{n} (e_i - \bar{e})(Y_i - \bar{Y}) = \sum_{i=1}^{n} e_i (Y_i - \bar{Y})$ $\quad$ [$\because \bar{e} = 0$ if β_0 is in the model]

$= \sum_{i=1}^{n} e_i Y_i = e'Y = e'e = \sum e_i^2 = SS_{Res}$

$Y = X\beta + \epsilon$; $\hat{\beta} = (X'X)^{-1} X'Y$

$\hat{Y} = X(X'X)^{-1} X'Y = X\beta$

$\hat{Y} = HY$; $\hat{Y} = HY$; $H = X(X'X)^{-1} X'$.

$e = Y - \hat{Y} = (I - H)Y$

$e'e = Y'(I-H)'(I-H)Y = Y'(I-H)Y = Y'e$ $\quad$ [$\because H^2 = H$]

$\text{Corr}(e, Y) = \dfrac{e'e}{\sqrt{(e'e) SS_T}} = \sqrt{\dfrac{e'e}{SS_T}} = \sqrt{\dfrac{SS_{Res}}{SS_T}} = \sqrt{1 - \dfrac{SS_{Reg}}{SS_T}} = \sqrt{1 - R^2}$.

(b) done in Page: 32.

(c) $\text{Corr}(e, \hat{Y})$

$$\sum (e_i - \bar{e})(\hat{Y}_i - \bar{\hat{Y}}) = e'\hat{Y} \quad , \quad \hat{Y} = HY$$
$$e = (I-H)Y$$

$$e'\hat{Y} = Y'(I-H)\,HY$$
$$= Y'(H - H^2)Y \quad ; \quad H^2 = H$$
$$= 0.$$

$$\text{Corr}(e, \hat{Y}) = 0.$$

$\boxed{5}$. Prove that the multiple correlation coefficient R^2 is equal to the square of the correlation coefficient between Y and $\hat{Y}$.

<u>Solution:-</u>

$$[\text{Corr}(Y, \hat{Y})]^2 = R^2 = \frac{SS_{Reg}}{SS_T}$$

$$r_{Y\hat{Y}} = \frac{\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})}{\sqrt{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{Y})^2}}$$

$$= \frac{\sum (\hat{Y}_i - \bar{Y})(\hat{Y}_i - \bar{Y}) + \sum (\hat{Y}_i - \bar{Y}) e_i}{\sqrt{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{Y})^2}}$$

$\sum e_i = 0$,

$\sum (Y_i - \hat{Y}_i) = 0$

$\sum Y_i = \sum \hat{Y}_i$

$\therefore \bar{Y} = \bar{\hat{Y}}$.

$\because Y_i = \hat{Y}_i + e_i$

$\because \text{Corr}(e, \hat{Y}) = 0$

$$= \sqrt{\frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}} = \sqrt{\frac{SS_{Reg}}{SS_T}} = \sqrt{R^2}.$$

$\boxed{6}$. A new born baby was weighted weekly. Twenty such weights are shown below, recorded in ounces. Fit to the data, using orthogonal polynomials, a polynomial model of degree d justified by the accuracy of the figures, that is, test as you go along for the significance of the linear, quadratic and so fourth, terms.

No of weeks:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Weights:	141	144	148	150	158	161	166	170	175	181	189	194	196	206	218	229	239	242	257	

Solution:- We wish to fit the model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_K x^K + \epsilon$$

$$\Downarrow$$

$$y = \alpha_0 + \alpha_1 P_1(x) + \alpha_2 P_2(x) + \cdots + \alpha_K P_K(x) + \epsilon$$

$SS_{Reg}(\alpha_1) = \hat{\alpha}_1 \sum y_i P_1(x_i) = 25,438.75$

$SS_{Reg}(\alpha_2) = \hat{\alpha}_2 \sum y_i P_2(x_i) = 489$

$SS_{Reg}(\alpha_3) = \hat{\alpha}_3 \sum y_i P_3(x_i) = 1.15$

$SS_T = \sum (y_i - \bar{y})^2 = 26,018$

$\hat{\alpha}_0 = \bar{y}$

$$\hat{\alpha}_j = \frac{\sum P_j(x_i) y_i}{\sum P_j^2(x_i)}$$

SV	DF	SS	MS	F
Reg(α_1)	1	25438.75	25,438.75	4558.98
Reg(α_2)	1	489	489	87.63
Reg(α_3)	1	1.15	1.15	0.21 $< 4.49 = F_{0.05,1,16}$
Res	16	89.30	5.58	
Total	19	26,018		

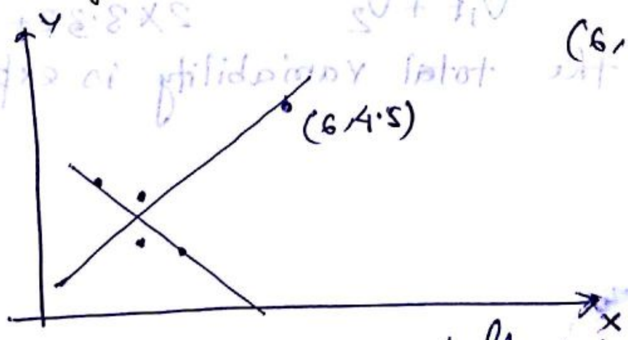
$$Y = 136.227 + 2.68x + 0.167x^2$$

If residual SS is large, then go for fourth degree polynomial.

$\boxed{7}$. If you are asked to fit a straight line to the data $(X, Y) = (1,3), (2,2.5), (2,1.2), (3,1)$ and $(6,4.5)$

What could you about it?

Solution:-



(6,4.5) is an influential observation.

Recommendation:- You can ignore influential observation if it's small in number.

Some observation between $x = 3$ & $x = 6$ would be useful here.

8. Your friend says he has fitted a plane to $n = 33$ obsns. on (X_1, X_2, Y) and that his overall regression (given b_0) is just significant at the $\alpha = 0.05$ level. You ask him for R^2 value but he doesnot know. You work it out for him on the basis of what he has told you.

Solution:- $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, $n = 33$

ANOVA

SV	DF	SS	MS	F
Reg	2	SS_{Reg}	MS_{Reg}	F
Res	30	SS_{Res}	MS_{Res}	
Total	32	SS_T		

$F = \frac{MS_{Reg}}{MS_{Res}}$

$F \sim F_{2,30}$

$F \geq F_{0.05,2,30} = 3.32$

$R^2 = \frac{SS_{Reg}}{SS_T}$

$$R^2 = \frac{SS_{Reg}}{SS_T} = \frac{SS_{Reg}}{SS_{Reg} + SS_{Res}}$$

$$= \frac{SS_{Reg}/MS_{Res}}{\frac{SS_{Reg}}{MS_{Res}} + \frac{SS_{Res}}{MS_{Res}}}$$

$\nu_1 \longrightarrow$ Reg DF $= 2$

$\nu_2 \rightarrow$ Res DF $= 30$

$F \approx 3.32$

$$= \frac{\nu_1 \frac{MS_{Reg}}{MS_{Res}}}{\frac{\nu_1 MS_{Reg}}{MS_{Res}} + \frac{\nu_2 MS_{Res}}{MS_{Res}}}$$

$$= \frac{\nu_1 F}{\nu_1 F + \nu_2} = \frac{2 \times 3.32}{2 \times 3.32 + 30} = 0.1812$$

18% of the total variability is explained by the model.

9. The following 24 residuals from a straight line fit are equally spaced and are given in time sequential order. Is there any evidence of lag-1 serial correlation? (67)

8, −5, 7, 1, −3, −6, 1, −2, 10, 1, −1, 8, −6, 1, −6, −8, 10, −6, 9, −3, 3, −5, 1, −9

Use a two-sided test at level $\alpha = 0.05$

Solution:- e_i, $i = 1(1)24$

$$\text{Corr}(\epsilon_u, \epsilon_{u+1}) = \rho.$$

$H_0: \rho = 0$ Vs. $H_1: \rho \neq 0$

Compute Durbin-Watson test statistic:

$$d = \frac{\sum_{u=2}^{24}(e_u - e_{u-1})^2}{\sum_{u=1}^{24} e_u^2} = \frac{2225}{834} \approx 2.67$$

$4 - d = 1.33$

Compare with d_L and d_U | for $\alpha = 0.025$ (two sided test)
$n = 24$, $K = 1$
$d_L = 1.16$, $d_U = 1.33$

$d = 2.67$, $4 - d = 1.33$

- If $d < d_L$ or $4 - d < d_L$ reject H_0.
  Accept H_0 as $d = 2.67 \nless 1.16$
- If $d > d_U$ and $4 - d > d_U$ accept $H_0: \rho = 0$
  $2.67 > 1.33$ & $1.33 \geq 1.33$

there is no lag-1 autocorrelation/serial correlation in the data.

68

10. Estimate the parameters α & β in the non-linear model

$$Y = \alpha + (0.49-\alpha)e^{-\beta(X-8)} + \epsilon$$

from the following observations —

X	8	10	12	14	16	18	20	22	24
Y	0.490	0.475	0.450	0.433	0.450	0.423	0.407	0.407	0.407

X	26	28	30	32	34	36	38	40	42
Y	0.405	0.393	0.405	0.400	0.395	0.400	0.390	0.407	0.390

Solution:- The problem is to estimate α & β of the non-linear model using the data, residual sum of square can be written as

$$S(\alpha,\beta) = \sum_u \left(Y_u - f(x_u,\alpha,\beta)\right)^2$$

$$= \sum_u \left(Y_u - \alpha - (0.49-\alpha)e^{-\beta(x_u-8)}\right)^2$$

$$f(x_u,\alpha,\beta) = \alpha + (0.49-\alpha)e^{-\beta(x_u-8)}$$

$$\frac{\partial f}{\partial \alpha} = 1 - e^{-\beta(x_u-8)}$$

$$\frac{\partial f}{\partial \beta} = -(0.49-\alpha)^{-\beta(x_u-8)}$$

Taylor series expansion of $f(x_u,\alpha,\beta)$ about the point (α_0,β_0) is

$$f(x_u,\alpha,\beta) = f(x_u,\alpha_0,\beta_0) + \left(1-e^{-\beta_0(x_u-8)}\right)(\alpha-\alpha_0) + \left[-(0.49-\alpha_0)e^{-\beta_0(x_u-8)}\right](\beta-\beta_0)$$

$$= f_u^0 + z_{1u}^0(\alpha-\alpha_0) + z_{2u}^0(\beta-\beta_0)$$ → Linear function

$$Y_u = f_u^0 + z_{1u}^0(\alpha-\alpha_0) + z_{2u}^0(\beta-\beta_0) + \epsilon_u$$

$$Y_u - f_u^0 = z_{1u}^0(\alpha-\alpha_0) + z_{2u}^0(\beta-\beta_0) + \epsilon_u$$

$$Y_0 = Z_0\theta_0 + \epsilon$$

LSE:- $\hat{\theta}_0 = (Z_0'Z_0)^{-1}Z_0'Y_0$

69

$$Y_0 = \begin{pmatrix} Y_1 - f_1^0 \\ \vdots \\ Y_n - f_n^0 \end{pmatrix} \quad Z_0 = \begin{pmatrix} Z_{11}^0 & Z_{21}^0 \\ \vdots & \\ Z_{1n}^0 & Z_{2n}^0 \end{pmatrix} \quad \theta_0 = \begin{pmatrix} \alpha_1 - \alpha_0 \\ \beta - \beta_0 \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

If we begin the iteration with initial guesses $\alpha_0 = 0.30$, $\beta_0 = 0.02$

iteration	α_0	β_0
0	0.30	0.02
1	0.84	0.10

The process goes until $|\alpha_{j+1} - \alpha_j| < \delta$ and

$|\beta_{j+1} - \beta_j| < \delta = 0.0001$

So,

iteration	α_0	β_0
2	0.3901	0.1009
3	0.3901	0.1016
4	0.3901	0.1016

$\longrightarrow$ stop here.

11. Look at these data. I don't know whether to fit two straight line, one straight line or what. How to solve this this dilemma?

Set A:

X	Y
8	5.3
0	0.9
12	7.1
2	2.4

X	Y
9	5.1
7	4.4
8	5.2
6	3.8

<u>Solution:</u>— If we attach a dummy variable Z to distinguish the two groups, we can look at all 4 possibilities.

$$Y = (\beta_0 + \beta_1 X) + Z(\alpha_0 + \alpha_1 X) + \epsilon \qquad Z = 0 \text{ for set A}$$
$$= 1 \text{ for set B}$$

$$= \beta_0 + \beta_1 X + Z\alpha_0 + \alpha_1 ZX + \epsilon$$

$$X = \begin{array}{cccc} 1 & X & Z & ZX \end{array}$$
$$\begin{pmatrix} 1 & 8 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 12 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ \hline 1 & 9 & 1 & 9 \\ 1 & 7 & 1 & 7 \\ 1 & 8 & 1 & 8 \\ 1 & 6 & 1 & 6 \end{pmatrix}$$

$$Y = X\beta + \epsilon$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \alpha_0 \\ \alpha_1 \end{pmatrix}$$

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$\hat{Y} = 1.142 + 0.506X - 0.0418Z - 0.036XZ$$

Case:- A single straight line is sufficient

$H_0: \alpha_0 = \alpha_1 = 0$

$$F = \frac{\{SS_{Reg}(\text{Full}) - SS_{Reg}(\text{Rest. model})\}/2}{MS_{Res}} = \frac{0.1818/2}{0.3272/4} = 1.11$$

$F \sim F_{2,4}$

$F < F_{0.05,2,4} = 6.94$

$\therefore$ H_0 is accepted.
We can go for single straight line fit.

ANOVA Table (Full model)

SV	DF
Reg	3
Res	7
Total	7

$\boxed{12}$. Suppose we have n observations of variables $X_1, X_2, \ldots, X_K, Y$ where X: predictors, Y: response. If Y_i's are poisson variables with mean μ_i; what type of analysis is feasible?

Solution:- $Y \sim \text{Poisson}(\mu_i)$

$f(y, \mu) = \exp\{y \ln\mu - \mu - \ln y!\}$; $b(\theta) = \ln y!$
= natural parameter

The variation in Y_i could be explained in terms of the $\tilde{x}_i$ values. We fit the model

$$g(\mu_i) = \tilde{x}_i'\beta = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}$$

$$\ln\mu_i = \tilde{x}_i'\beta$$

$$\mu_i = e^{\tilde{x}_i'\beta} \; ; \; E(y_i) = e^{\tilde{x}_i'\beta}$$

This is the model we fit for $Y_i \sim \text{Pois}(\mu_i)$.

If $Y_i \sim$ Normal then we fit $E(y_i) = \tilde{x}_i'\beta$.

TERMS AND DEFINITIONS

Time Series Data:- A time series is a sequence of data points, typically consisting of successive measurements made over a time interval.

Cross-sectional Data:- Cross sectional data is a type of data collected by observing many subjects (such as individuals, firms, countries, regions) at the same point of time, or without regard to differences in time.

Pooled Data:- Randomly sampled cross sections of individuals at different points in time. Ex. Current Population Survey 2014.

Panel Data:- Observe cross sections of the same individuals at different points in time.

Longitudinal Data:- A dataset is longitudinal if it tracks the same type of information on the same subjects at multiple points in time. Used in Biostat. same as Panel data.

Micropanel Data:- A micro-panel data set is a panel for which the time dimension T ~~is similar to the~~ is largely less important than the individual dimension N.

Data: y_{it}, $i = 1(1)N$, $t = 1(1)T$

Ratio Scale:- A scale of measurement of data which permits the comparison of differences of values; a scale having a fixed zero value.

Interval Scale:- A scale of measurement of data according to which the differences between values can be quantified in absolute but not relative terms and for which any zero is merely arbitrary.

Ordinal Scale:- A scale on which data is shown simply in order of magnitude since there is no standard of measurement of differences.

Nominal Scale:- A discrete classification of data, in which data are neither measured nor ordered but subjects are merely allocated to distinct categories.

Linear Regression

Examples:-
- The effect of hours studied on student grades
- The effect of education on income
- The effect of recession on stock returns

LR Variables:-
- The dependent variable is a continuous variable.
- The independent variables can be of any form,
- The multiple linear regression model has two or more independent variables.
- Regression Analysis does not establish a cause-and-effect relationship, just that there is a relationship.

Assumption of the OLS estimator:-
- Exogeneity of regressors
- Homoscedasticity
- Uncorrelated observations

Panel Data Models

Examples:-
- Labour Economics: effect of education on income, with data across time and individuals.
- Economics: effect of income on savings, with data across years and countries.

Characteristics:-
- Panel Data provides information on individual behavior, both across individuals and over-time, they have both cross-sectional & time-series dimensions.
- Panel data can be balanced when all individuals are observed in all time periods or unbalanced when individuals are not observed in all time periods.
- We assume correlation (clustering) over time for a given individual, with independence over individuals. Ex. the income for the same individuals is correlated over time but it is independent across individuals.

Panel data types :-

- short panel: many individuals and few time periods
- Long panel: many time periods and few individuals
- Both: many time periods and many individuals

Regressors:-

- Varying regressors x_{it},
  - annual income for a person, annual consumption of a product.
- Time-invariant regressors $x_{it} = x_i$ for all t.
  - gender, race, education
- Individual-invariant regressors $x_{it} = x_t$ for all i.
  - time trend, economy trends such as unemployment rate.

Formulae:- Individual mean, $\bar{x}_i = \frac{1}{T}\sum_t x_{it}$

Overall mean, $\bar{x} = \frac{1}{NT}\sum_i\sum_t x_{it}$

Overall variance, $S_0^2 = \frac{1}{NT-1}\sum_i\sum_t (x_{it}-\bar{x})^2$

Within variance, $S_W^2 = \frac{1}{NT-1}\sum_i\sum_t (x_{it}-\bar{x}_i)^2$

Between variance, $S_B^2 = \frac{1}{N-1}\sum_i (\bar{x}_i-\bar{x})^2$

$$S_0^2 = S_W^2 + S_B^2 .$$

- Time-invariant regressors have zero within variation.
- Individual-invariant regressors have zero between variation.

There are three types of models:- the pooled model, the fixed effects model, the random effect model.

Pooled model:- the pooled model specifies constant coefficients, the usual assumptions for cross-sectional analysis.

$$y_{it} = \alpha + x_{it}'\beta + \epsilon_{it}$$

Fixed-effects model:- The FE model allows the individual-specific effects α_i to be correlated with the regressors x.

$$y_{it} = \alpha_i + x_{it}'\beta + \epsilon_{it}$$

We can recover the individual specific effects after estimation as:

$$\hat{\alpha}_i = \bar{y}_i - \bar{x}_i'\hat{\beta}$$

Random effects model (RE):- The RE model assumes that the individual-specific effects α_i are distributed independently of the regressors. We include α_i in the error term. Each individual has the same slope parameters and a composite error term $\varepsilon_{it} = \alpha_i + e_{it}$.

$$y_{it} = x'_{it}\beta + (\alpha_i + e_{it})$$

Here $Var(\varepsilon_{it}) = \sigma_\alpha^2 + \sigma_e^2$ and $Cov(\varepsilon_{it}, \varepsilon_{is}) = \sigma_\alpha^2$

So, $\rho_\varepsilon = Cor(\varepsilon_{it}, \varepsilon_{is}) = \sigma_\alpha^2 / \sigma_\alpha^2 + \sigma_e^2$

Panel data estimators:-

- The panel data models can be estimated with several estimators.
- The estimators differ based on whether they consider the between or within variation in the data.
- Their properties (consistency) differ based on which model is appropriate.
- We prefer estimators that are consistent and efficient. We check for consistency first and then for efficiency.

Choosing between fixed and random effects:-

▨ Breusch-Pagan Lagrange Multiplier test:-

- This is a test for the random effect model based on the OLS (Ordinary least square)
- Test whether σ_α^2 or equivalently $Cor(e_{it}, e_{is})$ is significantly different from zero.
- If the test is significant, use random effect model instead of the OLS model.
- We still need to test for fixed Vs. Random effects.

▨ Hausman Test:-

- The Hausman test checks whethers there is significant difference between the fixed and random effects estimators.
- The Hausman test statistic can be calculated only for the time-varying regressors.
- Test statistic: $H = (\hat{\beta}_{RE} - \hat{\beta}_{FE})'(V(\hat{\beta}_{RE}) - V(\hat{\beta}_{FE}))^{-1}(\hat{\beta}_{RE} - \hat{\beta}_{FE})$

which is chi-square distributed with DF equal to the number of parameters for the time-varying regressors.

- If the test is significant use fixed effects, ow mixed effects.

Scanned by CamScanner

Sequential Selection in Regression:-

1) **Forward Regression:-** Forward selection of variables chooses the subset models by adding one variable at a time to the previously chosen subset. Forward selection starts by choosing as the one-variable subset the independent variable that accounts for the largest amount of variation in the dependent variable. This will be the variable having the highest simple correlation with Y. At each successive step, the variable in the subset of variables not already in the model that causes the largest decreases in the residual sum of squares is added to the subset. Without a termination rule, forward selection continues until all variables are in the model.

2) **Backward Regression:-** Backward elimination of variables chooses the subset models by starting with the full model and then eliminating at each step the one variable whose deletion will cause the residual sum of squares to increase the least. This will be the variable in the current subset model that has the smallest partial sum of squares. Without a termination rule, backward elimination continues until the subset model contains only one variable.

3) **Stepwise Regression:-** Neither forward selection nor backward elimination takes into account the effect that the addition or deletion of a variable can have on the contributions of other variables to the model. A variable added early to the model in forward selection can become unimportant after other variables are added, or variables previously dropped in backward elimination can become important after other variables are dropped from the model. Stepwise regression is a forward selection process that rechecks at each step the importance of all previously included variables. If the partial sum of squares for any previously included variables do not meet a minimum criterion to stay in the model, the selection procedure changes to backward ~~selection~~ elimination and variables are dropped one at a time until all remaining variables meet the minimum criterion. Then, forward selection resumes.

Probit and Logit Models (Binary Outcome Models) (76)

Binary outcome examples:-

- Consumer Economics: whether a consumer makes a purchase or not.
- Labor Economics: whether an individual participates in the labor market or not.
- Agricultural Economics: whether or not a farmer adopts or uses organic practices, marketing/production, etc.

Binary outcome dependent variable:-

- The decision/choice is whether or not to have, do, use or adopt.
- The dependent variable is a binary response.
- It takes on two values: 0 and 1.

$$y = \begin{cases} 0 & \text{if no} \\ 1 & \text{if yes} \end{cases}$$

Binary outcome models:-

- Binary outcome models are among most used in applied economies.
- A look at the OLS model: $y = \underset{\sim}{x}'\beta + \epsilon$
- Binary outcome models estimate the probability that $y=1$ as a function of the independent variables.

$$p = \Pr[y=1 \mid \underset{\sim}{x}] = F(x'\beta)$$

Three models depending on $F(x'\beta)$

(i) **Regression model** (linear Probability model):-

Here $F(x'\beta) = x'\beta = P[y=1 \mid \underset{\sim}{x}] = p$

- A problem with the regression model is that the predicted probabilities will not be limited between 0 and 1.
- We do not use the regression model with binary outcome data.

(ii) **Logit model**:- For the logit model,

$$F(x'\beta) = \Lambda(x'\beta) = \frac{e^{x'\beta}}{1+e^{x'\beta}} = \frac{\exp(x'\beta)}{1+\exp(x'\beta)}$$

The predicted probabilities are limited between 0 and 1.

(iii) **Probit model**:-

$$F(x'\beta) = \Phi(x'\beta) = \int_{-\infty}^{x'\beta} \phi(z)\,dz$$

The predicted probabilities are limited between 0 and 1.

77

Model coefficients:- Probit and logit models are estimated using the maximum likelihood method.

Interpretation of coefficients:-

- An increase in x increases / decreases the likelihood that $y=1$ (makes that outcome more or less likely).
- We interpret the sign of the coefficient but not the magnitude. The magnitude can't be interpreted using the coefficient because different models have different scales of coefficients.

Marginal effects:-

- When estimating probit and logit models, it is common to report the marginal effects after reporting the coefficients.
- The marginal effects reflect the change in the Probability of $y=1$ given a 1 unit change in an independent variable x.

Marginal effect for regression model:-

- For the OLS regression model, the marginal effects are the coefficients and they don't depend on x.

$$\frac{\partial p}{\partial x_j} = \beta_j$$

For logit model, marginal effect is ——

$$\frac{\partial p}{\partial x_j} = \Lambda(X'\beta)\left[1-\Lambda(X'\beta)\right]\beta_j = \frac{e^{X'\beta}}{(1+e^{X'\beta})^2}\beta_j$$

For probit model, marginal effect is ——

$$\frac{\partial p}{\partial x_j} = \phi(X'\beta)\beta_j .$$

Interpretation:-

- An increase in x increases (decreases) the prob. that $y=1$ by the marginal effect expressed as a percent.
  - For dummy independent variables, the ME is expressed in comparison to the base category.
  - For continuous independent variable, the ME is expressed for a one-unit change in x.
- We interpret both the sign & the magnitude of the ME.
- The probit & logit models produce almost identical ME.

Odds Ratio/Relative Risk for the logit model:-

Odds Ratio/Relative Risk $= \frac{p}{1-p}$ and measures the prob. that $y=1$ relative to the prob. that $y=0$.

$$p = \frac{\exp(x'\beta)}{1+\exp(x'\beta)}$$

$$\frac{p}{1-p} = \exp(x'\beta)$$

$$\therefore \; x'\beta = \ln\left(\frac{p}{1-p}\right).$$

- An odds ratio of 2 means that outcome $y=1$ is twice more likely as the outcome of $y=0$.
- Odds ratios are estimated with the logistic model.
- Reporting marginal effects instead of odds ratio is more popular in economics.

SURVIVAL ANALYSIS

Examples:- — Finance: Loan performance (borrowers obtain loans and then they either default or continue to reply their loans)

— Economics: Firm survival and exit.
Time to retirement, finding a new job, etc.
Adoption of new technology (firm either adopt the new technology or not).

Set up:-

— Subjects are tracked until an event happens (failure) or we lose them from the sample (censored observation).

— We are interested in how long they stay in the sample (survival)

— We are interested in their risk of failure (hazard rates).

Functions:- The dependent variable duration is assumed to have a continuous probability distribution $f(t)$.

The probability that the duration time will be less than t is:

$$F(t) = \text{Prob}(T \le t) = \int_0^t f(s)\,ds$$

Survival function is the probability that the duration will be at least

$$S(t) = 1 - F(t) = P(T \ge t)$$

Hazard rate is the prob. that the duration will end after time t, given that it has lasted until time t:

$$\lambda(t) = \frac{f(t)}{S(t)}.$$

- Nonparametric estimation is useful for descriptive purposes and to see the shape of the hazard or survival function before a parametric model with regressors is introduced.

Procedure:-

- Sort the observations based on duration from the smallest to largest $t_1 \le t_2 \le \dots \le t_n$.
- From each duration, determine the number of observations at risk n_j (those still in the sample), the number of events d_j and the number of censored observations m_j.

$$\lambda(t_j) = \text{hazard function} = \frac{d_j}{n_j}$$

- Nelson-Aalen estimator of the cumulative hazard function,

$$\Lambda(t_j) = \sum \frac{d_j}{n_j}$$

- The Kaplan-Meier estimator of the survival function

$$S(t_j) = \prod \frac{n_j - d_j}{n_j}$$

- Unlike the non-parametric estimation, the parametric models also allow the inclusion of independent variables.

Parametric model	Hazard function λ	Survival function S
Exponential	γ	$e^{-\gamma t}$
Weibull	$\gamma \alpha t^{\alpha-1}$	$e^{-\gamma t^{\alpha}}$
Gompertz	$\gamma e^{\alpha t}$	$\exp(-(\gamma/\alpha)(e^{\alpha t}-1))$
Log-logistic	$\dfrac{\alpha \gamma^{\alpha} t^{\alpha-1}}{(1+(\gamma t)^{\alpha})}$	$1/(1+(\gamma t)^{\alpha})$

- The exponential model has a constant hazard rate over time.

Cox-proportional hazard model:-

Hazard rate defined as $\lambda(t \mid x, \beta) = \lambda_0(t) e^{x'\beta}$

Rule:-

Coefficient	Hazard rate	Conclusion
Positive	>1	Lower duration, higher hazard rate (more likely for the event to happen)
Negative	$(0,1)$	Higher duration, lower hazard rate (less likely for the event to happen)

TIME SERIES ARIMA MODELS

Examples :-
- Modelling relationship using data collected over time — prices, quantity, GDP, etc.
- Forecasting — predicting economic growth.
- Time series involves decomposition into a trend, seasonal, cyclical, and irregular component.

White Noise :-
- White noise describes the assumption that each element in a series is a random draw from a population with mean zero and constant variance.
- Autoregressive (AR) and moving average (MA) models correct for violation of this white noise assumption.

Ques. How to model a time series data?

A major assumption in time series analysis is the stationarity of the series, this means that the average value and the variation of the series should be constant with respect to time. If the series is not stationary we make it stationary by using differencing method or other transformation.

Stationary Test → Unit Root test (ADF Test)
Stationary Test → KPSS Test

Box-Jenkins Modelling (ARIMA Modelling) :-

The methodology was introduced by Box and Jenkins assumes that the data is dependent on itself. And the very first thing is to decide on is the number of lags. Then a number of parameters are estimated, the residuals are checked and finally a forecast is made.

The general ARIMA (p, q, d) model looks like:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q};$$ where

c = constant, ϕ_i and θ_j $(i = 1(1)p,\ j = 1(1)q)$ are model parameters.

p: number of terms in AR model.
q: number of terms in MA model.

Forecasting Methods:-

Method	Data Pattern	Data Points	Forecast Horizon	Quantitative Skills
Moving Average	Stationary	At least the number of periods in MA	Very short	Little
Single Exponential smoothing	Stationary	5-10	Short	Little
ARIMA Methodology	Stationary	4-5 per season	Medium	High

Model Selection criteria:-

- Akaike Information Criteria (AIC)
- Bayesian Information Criteria (BIC)

The best model is that which minimizes AIC & BIC.

Residual Analysis:-
- Normality test } assumption of normality checking
- Whiteness Test.
- Ljung - Box test. } Autocorrelation test.

Measurement of Forecast Accuracy:-

- Mean Absolute Percentage Error (MAPE): $MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_t}{y_t}\right| \times 100$.

  Criteria: MAPE < 10% $\Rightarrow$ model is reasonably good.

  MAPE < 5% $\Rightarrow$ model is very good.
- Mean square Error; $MSE = \sum_{i=1}^{n}\frac{e_t^2}{n}$.
- Root Mean square Error: $RMSE = \sqrt{MSE}$
- Mean Absolute Error (MAE): $MAE = \frac{1}{n}\sum_{i=1}^{n}|e_t|$

MORE ABOUT REGRESSION

- Sometimes it may happen that the mean of the data is correlated with its variance.
- The distribution of such data is typically skewed.
- In this case a transformation may be required to make the distribution symmetrical (normal).
- Result of any transformation pertain only to the transformed response.
- However backtransforming the analysis will make inferences to the original response.

Check these before modelling:-

1. Check normality of each predictors
2. Errors must be normal with mean zero & constant variance
3. Errors are uncorrelated.
4. Errors and predictors must be uncorrelated.

Need of Transformation of Data:-

- For stabilizing response variance
- Making the distribution of the response variable closer to normal distribution.
- Improving the fit of the model to the data.

* On Page: 16, commonly used transformations are given.

Box-Cox Transformation:- Transformation can be defined as

$$y_T = \begin{cases} \dfrac{y^{\lambda}-1}{\lambda Y^{\lambda-1}} & \text{for } \lambda \neq 0 \\ Y \ln Y & \text{for } \lambda = 0 \end{cases}$$

$$Y = (y_1 y_2 \cdots y_n)^{1/n} ;$$

Box-cox procedure evaluates the change in sum of squares for error for a model with a specific value of λ.

Generally, $-5 \leq \lambda \leq 5$.

To use Box-Cox transformation all data must be > 0.

- Box-cox is a procedure to identify an appropriate exponent (λ) to use transformation into normal shape.

Note:- For smoothing the data log transformation is suggested.

- First of all we should do transformation of X's (predictors).
- If improvement is there then we are done but if no improvement is there we should go for the transformation of response.
- When decision is taken based on scatter plot, that time don't drop any variable.
- Always start with linear regression model.

CLASSIFICATION & REGRESSION TREE

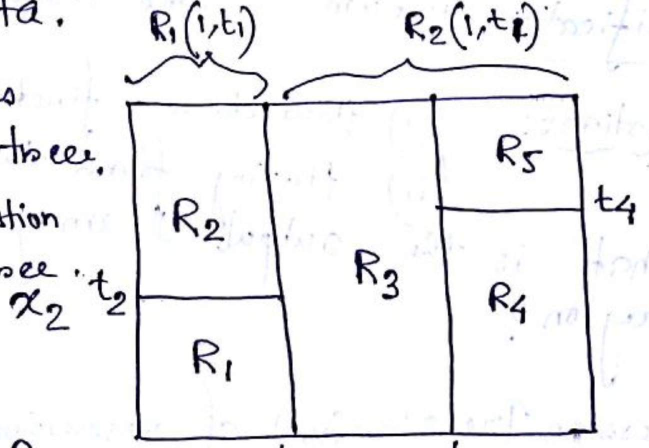
- Machine learning technique ; Supervised learning technique.
- Decision tree based approach for building regression model.

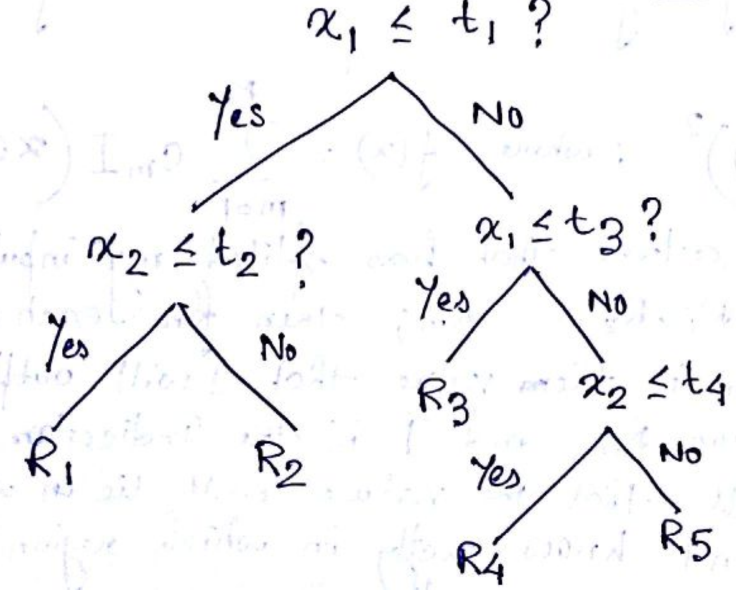
— Partition the input space into rectangles by drawing axis parallel lines. (Why)

because these lines can be specified very easily by just comparing against one of those dimensions ($x_1/x_2/x_3$) of the input data.

These lines will help us to construct decision tree.

Every point a binary question is asked, so it's a binary tree.





This algorithm is called Branch & Bound Algorithm.

Advantage:- (i) Decision trees are fantastic because they are the most interpretable of all of the classifiers that we are going to look at even more so than Linear Regression. Interpretability is high for CART.

(ii) They can work well with Mixed mode data. X : Continuous or discrete, Y : Discrete or Continuous When Y is discrete, we use classification tree, when Y is continuous we use Regression tree.

Regression Problem:- Same 'real valued output for each region.

Regardless where the data point is falling in R_p, we are going to predict the same output.

Classification Problem:- Same class level for the region.

Questions: (i) How do we find out the region?
(ii) Having found the region how do I decide what is the output I am going to produce for that region?

Regression Trees:- Goal of regression is to minimize sum of square of errors.

Minimize $\sum (y_i - f(x_i))^2$; where $f(x) = \sum_{m=1}^{M} c_m I(x \in R_m)$

Let suppose that I have a tree that has splited my input space into m regions ; $R_1, R_2, \ldots, R_m$; then for each of these regions and c_m is the value that I will output that lies in the region R_m and I is the Indicator function which will tell that the value will lie in which region R_m. We don't know exactly in which region it's lying, so using summation.

$f(x)$ will be non-zero for only one term, i.e. the data point is lying in. Suppose data point is lying in R_2, then $f(x) = c_2$.

Best $\hat{c}_m$ are $(y_i \mid x_i \in R_m)$.

So, this gives the solution of question (ii). Error measure in Regression tree is by MSE (Mean squared Error).

Finding best $R_m(s)$:- It's a tough combinatorial problem.

Split Variable:-

$$R_1(j, s) = \{x \mid x_j \le s\}$$ [Shown in fig (Pg: 80)]

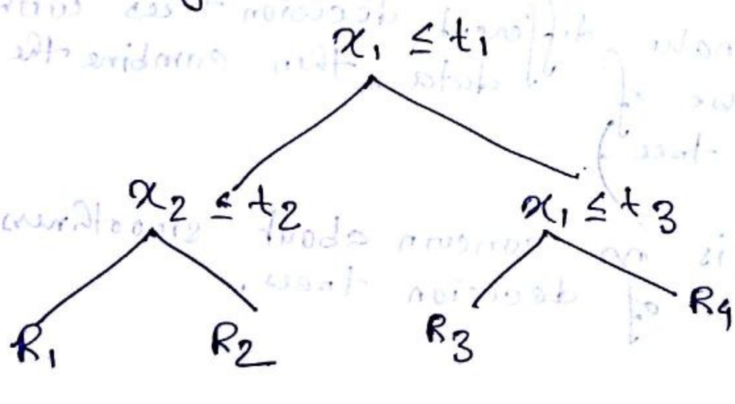
$$R_2(j, s) = \{x \mid x_j > s\}$$

We are to find j and s such that we can minimize this

$$\left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right]$$

Since number of data points is less & finite so we can find best (j, s) and proceed.

Stopping Rule:- Until each region has few data points we will continue. Grow a large tree, then Prune the tree (collapse the internal nodes of the tree such that you come up with a better tree).



Now, look first tree & prune tree for validation data, accordingly take decision which one to use for training data.

Classification Trees:-

prediction, $\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I\{y_i = k\}$

where, N_m = total number of points in region m.

where $\hat{p}_{mk}$ = probability that the data point in region m belongs to class k.

$$\text{Class}(m) = \arg\max_{k} \hat{p}_{mk}.$$

Error Measures:-

Misclassification Error:- $\frac{1}{N_m} \sum_{i \in R_m} I\left(y_i \neq \text{class}(m)\right) = 1 - \hat{p}_{m\,\text{class}(m)}$

the number of times the actual level doesn't match the prediction that we make by our classifiers.

Cross Entropy: $-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$.

If we have sufficient training data, $\hat{p}_{mk}$ is true actual distribution of output data.

Disadvantages:-

1. Trees are notoriously unstable. If there is a very small change in the training data, the tree would be very different. This is not the case in Logistics, SVM.
(So, we make different decision trees with slightly different use of data, then combine the trees into a single tree)

2. There is no concern about smoothness in the use of decision trees.

REGRESSION SPLINES

In a simple regression problem, given fixed $x_1, x_2, \ldots, x_n$, we obtain $y_1, y_2, \ldots, y_n$, where $y_i = f(x_i) + e_i$; where e_i's are iid with mean zero and variance σ^2 (unknown). The problem is to estimate the function 'f'.

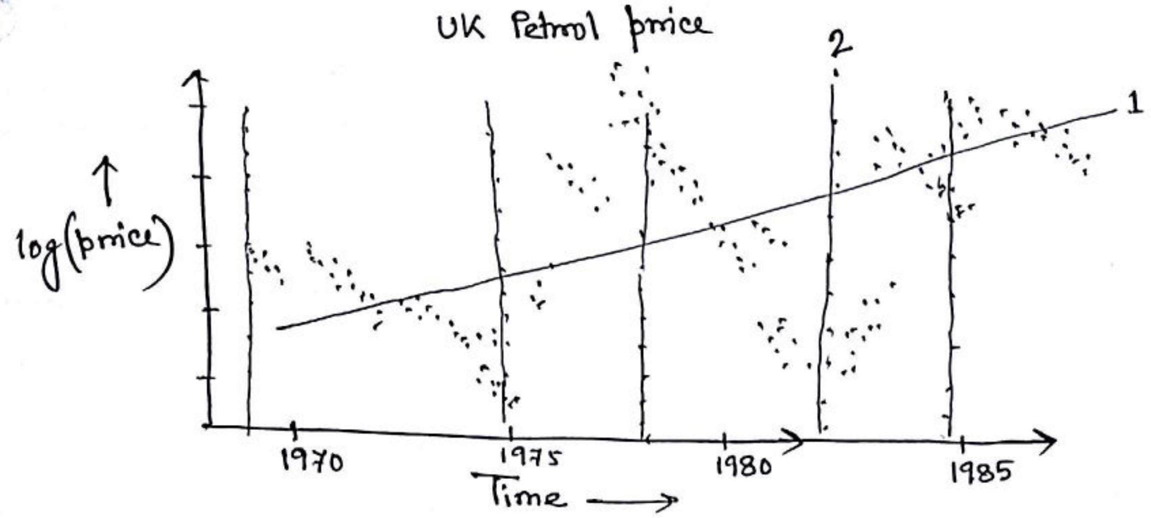
Parametric Regression :- The parametric approach is quite flexible in a sense that we are not constrained to just linear predictors but can incorporate polynomials and other functions of the variable in the model to attain higher degree of precision.

Non-parametric Regression :-

- Non-parametric approach is to choose 'f' from some smooth family of functions.
- The range of potential fits to the data is much larger than the parametric approach.
- Although some assumptions are made about 'f' (eg. degree of smoothness and continuity), these restrictions are far less than that in the parametric way.
- Non-parametric models do not have a formulaic way of describing the relationship between predictors and the response.
- Unlike parametric models/methods, which is prone to choose the wrong model and hence introduce bias in the model, non-parametric approach assumes less and hence is likely to generate less bias.

Spline regression is one of the efficient tools of non-parametric regression.

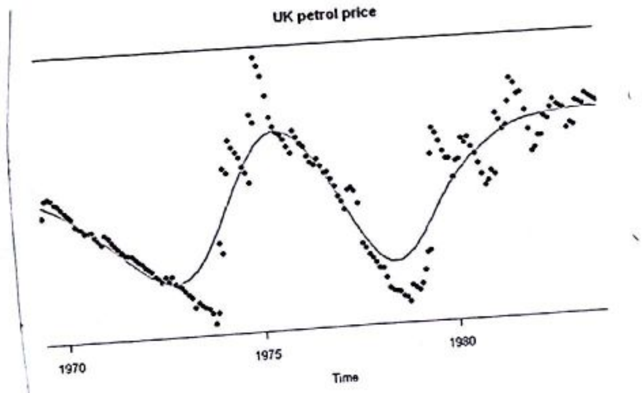
Ref. BOOK:- Nonparametric Regression and Spline Smoothing by Randall.



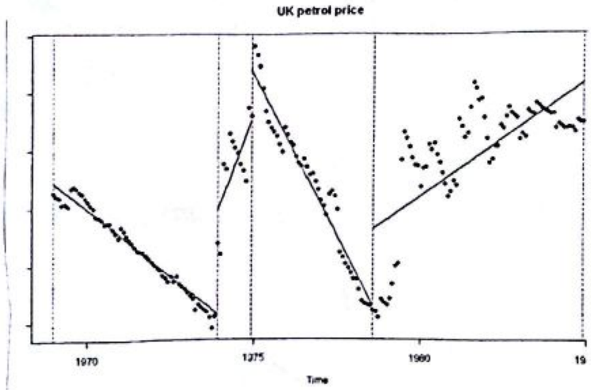
Interesting trends over time is visible.

Points to be noted:- 1. Linear regression is inadequate.

2. We can split time series in a number of parts, then perform regression on each part, then also regression pieces don't have to be linear, but they have to be connected. So, each regression line uses information in other parts.



When using higher order polynomial pieces:
Derivatives are `also connected'



Splitting either via evenly spaced `knots', or via known knot locations based on external information

Concept of Spline enters.

• What is Spline?

A spline is a smooth polynomial function that is piece-wise defined, and possesses a high degree of smoothness at the places where the polynomial pieces connect (which are known as knots).

• Mathematically, a spline is a piecewise polynomial real function $S:[a,b] \to \mathbb{R}$ on an interval $[a,b]$ composed of K ordered disjoint subintervals $[t_{i-1}, t_i]$ with $a = t_1 < t_2 < \cdots < t_{K-1} < t_K = b$.

The restriction of S to an interval 'i' is a polynomial

$$P = [t_{i-1}, t_i] \longrightarrow \mathbb{R}, \text{ so that}$$

$$S(t) = P_1(t), \quad t_0 < t < t_1$$
$$S(t) = P_2(t), \quad t_1 < t < t_2$$
$$\vdots$$
$$S(t) = P_K(t), \quad t_{K-1} < t < t_K.$$

• The highest order of the polynomials $P_i(t)$ is said to be the order of the spline 'S'.

• If all subintervals are of the same length, the spline is said to be uniform and non-uniform otherwise.

• The idea is to choose the polynomials in a way that guarantees sufficient smoothness of 'S'. Specifically, for a spline of order 'n', 'S' is required to be continuously differentiable to order $(n-1)$ at the interior point $t_i \ \forall \ i = 1, 2, \ldots, (K-1)$ and $\forall \ j \ni 0 \le j \le (n-1)$;

$$P_i^{(j)}(t_i) = P_{i+1}^{(j)}(t_i)$$

— Otherwise, the curve will not be smooth at the knot points.

- TYPES OF SPLINE :-
  - Smoothing Splines
  - Regression Splines
  - Interpolating splines (hybrid of smoothing and regression splines)

Smoothing Splines :-

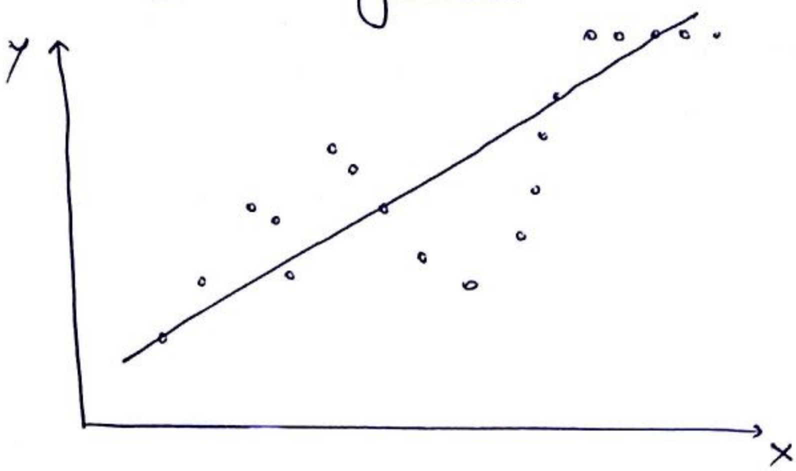
- The smoothing spline is a method of smoothing (fitting a smooth curve to a set of noisy observations) using a spline function.
- Let (X_i, Y_i) ; $x_1 < x_2 < \cdots < x_n$, $i \in \mathbb{Z}$ be a sequence of obsns, modeled by the relation $Y_i = f(x_i)$.

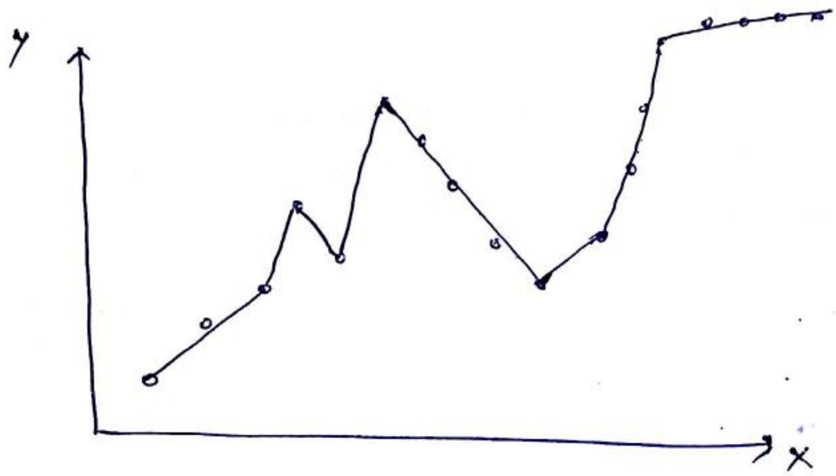
  The smoothing spline estimates $\hat{f}$ of the function 'f' is defined to be the minimizer (over the class of twice differentiable functions)

$$s(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{f}(x_i)\right)^2 + \lambda \int_{x_1}^{x_n} \left[\hat{f}''(x)\right]^2 dx$$

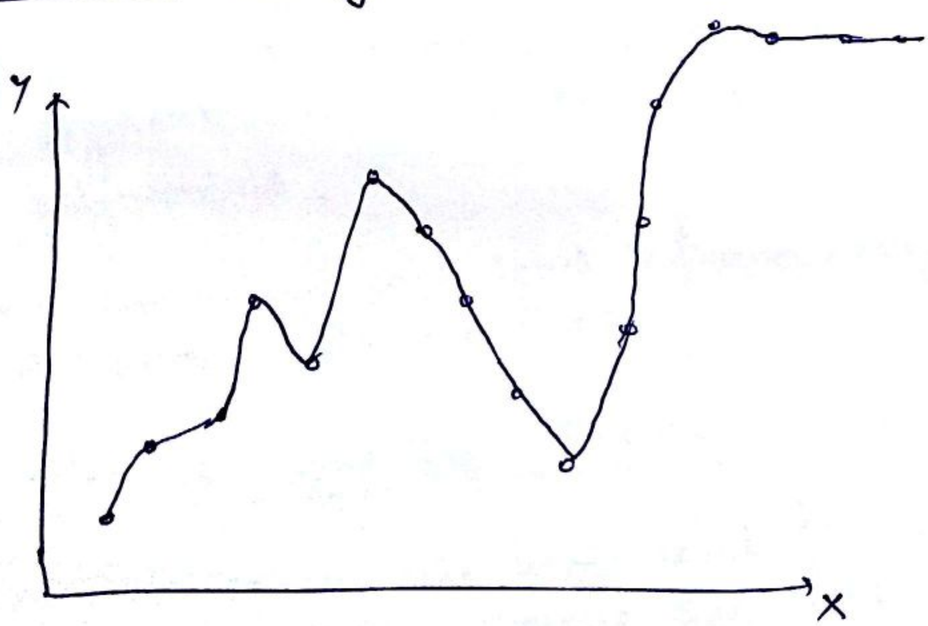
- Here $\lambda > 0$ is the smoothing parameter controlling the trade-off between fidelity to the data and roughness of the function estimate.
- $\int_{x_1}^{x_n} \left[f''(x)\right]^2 dx$ is a roughness penalty.
- As $\lambda > 0$, (no smoothing), the smoothing spline converges to the interpolating spline.
- As $\lambda \to \infty$ (infinite smoothing), the roughness penalty becomes paramount and the estimate converges to a linear least squares estimate.
- The roughness penalty based on the second derivative is the most common in modern statistics literature, although the method can easily be adapted to penalties based on other derivatives.

Linear Regression





After Spline Smoothing:



Regression Splines (B - Splines) :-

- The term "B - Spline" was coined by Isaac Jacob Schoenberg. B - splines (basis - splines) constitute an appealing method to the non - parametric regression of a range of statistical objects of interest.

- Every spline function of a given degree, smoothness, and domain partition, can be uniquely represented as a linear combination of B - splines of that same degree and smoothness and over that same partition.

- Spline regression estimates different linear slopes for different ranges of the independent variables. The endpoints of the ranges are called knots. It is the freedom to choose the number of knots that makes the method non - parametric.

Advantages:- • Polynomials (Parametric regression) have the advantage of smoothness, but the disadvantage that each data points affect the fit globally.

• Broken stick regression method (Non - parametric regression) localizes the influence of each data point to its particular segment but do not have the same smoothness as with the polynomials.

• Smoothness and local influence may be combined by using B - spline basis functions.

Regression Spline Vs. Smoothing Spline :-

- For regression splines, the knots used for the basis are much smaller than the sample size and the no. of knots chosen, than the smoothing splines.

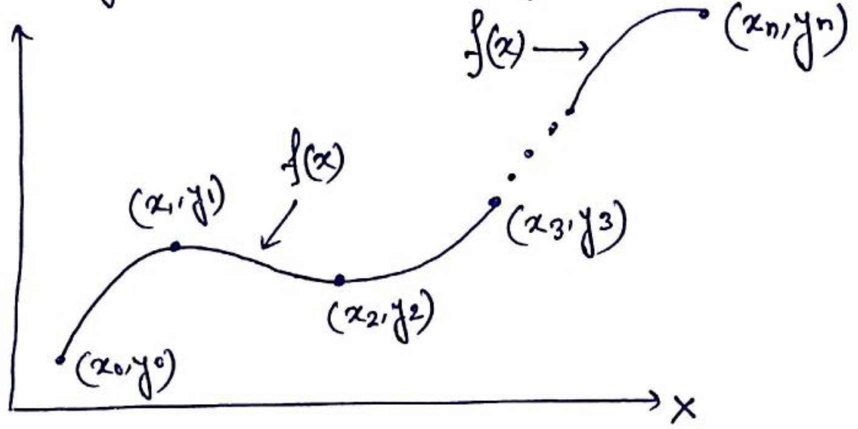
- Smoothing spline explicitly penalize roughness and use the data points themselves as potential knots where as regression splines place knots as equidistant / equiquantile points.

How to choose a Spline?

- Hermite curves are good for single segments where you know the parametric derivative or want easy control of it.
- B-splines are good for large continuous curves and surfaces,

What is Interpolation?

Given $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$, find the value of '$y$' at a value of '$x$' that is not given.



Interpolating spline:-

- Spline interpolation is a form of interpolation where the interpolant is a special type of piecewise polynomial called a spline.
- Spline interpolation is preferred over polynomial interpolation because the interpolation error can be made small even when using low degree polynomials for the spline.

— BY TANUJIT CHAKRABORTY
RS, ISI KOLKATA, M: 8420253573
MAIL: tanujitisi@gmail.com