Multiple Linear Regression

Look up Fundamental Assumptions.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots \beta_k x_{ki} + \epsilon_i \quad \text{—} \ (1)$$

$\beta_0 =$ Expected value of y when $x_i = 0$

$\beta_i =$ Change in y for unit change in x_i

$\epsilon_i =$ Random var with zero mean, std. dev $= \sigma_i$

Rewriting in matrix form.

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots \beta_k x_{1k} + \epsilon_1$$

$$\vdots$$

$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots \beta_k x_{nk} + \epsilon_n.$$

or

$$\underset{[n\times 1]}{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \qquad \underset{[(k+1)\times 1]}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

$$\underset{n\times(k+1)}{[X]} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ \vdots & & & & \\ 1 & x_{n1} & \cdots & & x_{nk} \end{bmatrix}$$

Thus we have

$$\underset{[n\times 1]}{Y} = X B$$

$$X \to n\times(k+1), \quad B \to (k+1)\times 1$$

Least Square estimator

We have. $\hat{y} = X\beta$

errors as: $(y-\hat{y})^2 = \hat{e}^2$

$$= [y - X\beta]'[y - X\beta]$$

thus $\hat{e}^2 = [y'y + (X\beta)'X\beta - (X\beta)'y - y'X\beta]$

$$\hat{e}^2 = [y'y + \beta'X'X\beta \underbrace{- \beta'X'y - y'X\beta}_{\text{same scalar Product}}]$$

Now for minimising squared error

$$\frac{\partial \hat{e}^2}{\partial \beta} = 0 \Rightarrow \quad 2X'X\beta - (X'y + y'X) = 0$$

Now using $X'y = y'X$ $\therefore$ $X'X\beta = X'y$

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

Soln for $\left[\hat{\beta} = (X'X)^{-1}X'y\right]$

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

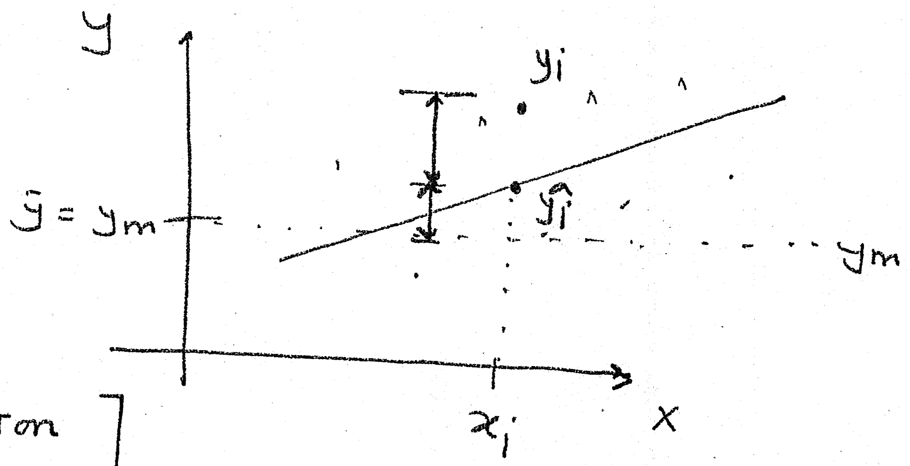
$\hat{y} = Hy$ $\quad$ H = hat matrix

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

$y_i = \hat{y} + e$ $\quad$ $e = y_i - \hat{y} = y_i - Hy_i$

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

R^2 and adjusted R^2



Explained Variation
Unexplained Variation
Total Variation.

Total Error = $y_i - \bar{y}$
error / unexplained = $y_i - \hat{y_i}$
Explained = $\hat{y}_i - \bar{y}$

Total error = $y_i - \bar{y}$ = Unexplained $(y_i - \hat{y_i})$ + Explained $(\hat{y_i} - \bar{y})$

Square bothsides

$$(y_i - \bar{y})^2 = \left[(y_i - \hat{y}_i) + (\hat{y_i} - \bar{y}) \right]^2$$

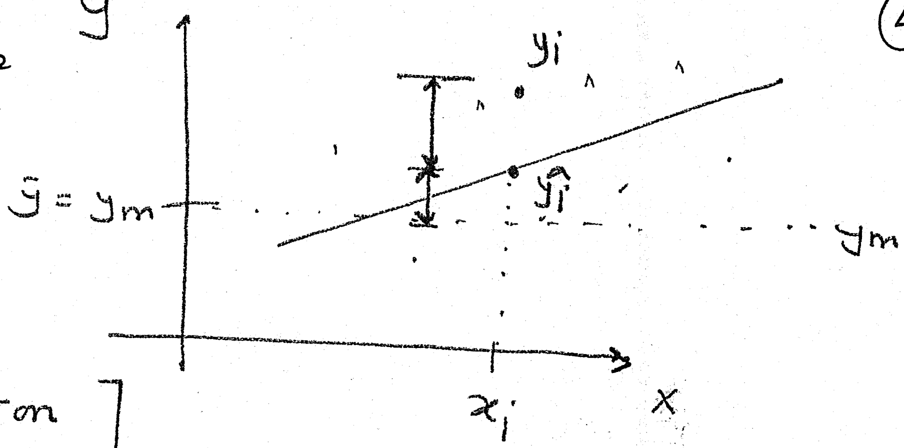
on Simplifying [SST] $= \underbrace{(y_i - \hat{y_i})^2}_{SSE} + \underbrace{(\hat{y_i} - \bar{y})^2}_{SSR}$

Tot S.S = SSE + SSR

$$R^2 = \frac{\text{Explained Variation}}{\text{Total variation}} = \frac{SSR}{SSE + SSR}$$

~~adjusted R~~

R^2 and adjusted R^2



Explained Variation
Unexplained Variation
Total Variation.

Total Error = $y_i - \bar{y}$

error / unexplained = $y_i - \hat{y}_i$

Explained = $\hat{y}_i - \bar{y}$

Total error = $y_i - \bar{y}$ = Unexplained $(y_i - \hat{y}_i)$ + Explained $(\hat{y}_i - \bar{y})$

Square bothsides

$$(y_i - \bar{y})^2 = \left[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \right]^2$$

on Simplifying [SST] = $\underbrace{(y_i - \hat{y}_i)^2}_{SSE} + \underbrace{(\hat{y}_i - \bar{y})^2}_{SSR}$

Tot S.S = SSE + SSR

$$R^2 = \frac{\text{Explained Variation}}{\text{Total variation}} = \frac{SSR}{SSE + SSR}$$

Adjusted $R^2 = \frac{SS\ Reg}{SST}$

Total = F + R

$\downarrow \quad \downarrow \quad \downarrow$

$(n-1) \qquad (k-1)$

$R^2_{adjusted} = 1 - \frac{SSE/dfe\ (n-p-1)}{SST/df_{total}\ (n-1)}$

$k = p+1$

F Test of Linear Regression.

Null

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

$H_1 \quad \beta_j \neq 0$ For at least one value of j

$$F_{stat} = \frac{MS_E \text{ of Reg}}{MS_E \text{ of error}} = \frac{SSR/df_r}{SSE/dfe}$$

Example:

$n = 35$ observations.

$k = 9$ independent variables $\qquad p = k+1 = 10$ parameters.

$SSE = 134$

$SSR = 289$

Test H_0 $\quad$ 0.05

$$F_{statistic} = \frac{MSR}{MSE} = \frac{SSE \times dfe}{} \qquad \frac{SSR/df_R\ (p-1)}{SSE/dfe\ (n-p)}$$

$$= (134) \times \frac{289/9}{134/25} \qquad F_{stat} = 5.99$$

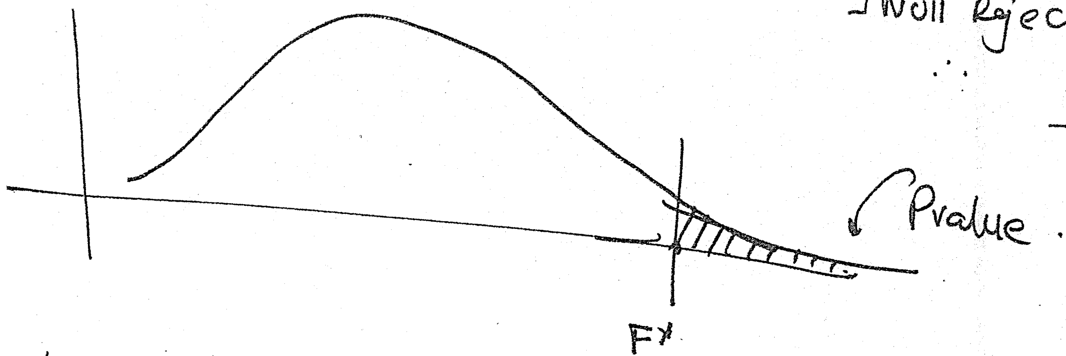
Look up Ftable $(9, 25)$

$\alpha = 0.05$

$F_{Critical}$ $(9, 25)$ $\alpha = 0.05$ $= 2.28$

[$F_{statistic}$ = $5.99 > F_{critical}$] Null Rejected

∴ Model is good



Now check using p values:

$P_{crit} = 0.05$

$F_{stat} = 5.99$ $(9, 25)$

We see from table For $F_{stat} = 5.99$, $(9,25)$
Pvalue will be less than 0.01

as $P_{less} < P_{crit}$ Null is Rejected

Significance of Individual Coefficients:

Methods → ① t test of individual coeff
② Partial F test.

S.d = $\sqrt{var}$ = $\sqrt{\sigma^2}$

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{(n-1)} \qquad \sigma = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$$

$$\sqrt{\frac{\sum(x_i - e_i)^2}{n-1}} = \sqrt{\hat{\sigma}^2} \qquad \hat{\sigma}^2 = \text{Error Mean square}$$

Regression Estimate

$$\hat{y} = x\hat{\beta} \qquad \hat{\beta} = (X^T X)^{-1} X^T y$$

Va. Covar.

$$Var = \frac{\sum(y_i - \bar{y})^2}{(n-1)} \qquad Covar = \sum \frac{(y_i - \bar{y})(x_i - \bar{x})}{n-1}$$

$$C = \hat{\sigma}^2 (X^T X)^{-1} \rightarrow \text{Var / Covar Matrix.}$$

$\hat{\sigma}^2$ - error Mean square.

$$SE(\beta_j) = \sqrt{C_{jj}}$$

Test statistic. [Significance of individual coefficients]

$\alpha = 0.05$

$$T_0 = \frac{\hat{\beta}_j}{SE(\beta_j)}$$

$H_0 \; \beta_j = 0$

$H_1 \; \beta_j \neq 0$

if $T_0 > t_{\alpha/2, n-2}$

or $T_0 < -t_{\alpha/2, n-2}$

also called Partial test.

Partial F test

$$y = \underbrace{\underbrace{\beta_0 + \beta_1 x_1 + \cdots}_{(k+1)-r} \quad \underbrace{\beta_k x_k}_{r}}_{(k+1)}.$$

$\theta_1 \rightarrow$ Contains First $(k+1)-r$ coeff.

$\theta_2 \rightarrow$ Contains last r coeff.

thus $\theta_1 = [\beta_0, \beta_1 \cdots \beta_{k-r}]$

$\theta_2 = [\beta_{k-r+1} \cdots \beta_k]$

Null $H_0: \theta_2 = 0$

$H_1: \theta_2 \neq 0.$

Example: Consider $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$

Partial F test for β_1

$$F_0 = \frac{SSR(\beta_1/\beta_2)/r}{MSE} \quad \text{as only 1 param. so } r = 1$$

$$= SSR(\beta_0, \beta_1, \beta_2) - SSR(\beta_0, \beta_2)$$

How to quickly compute $SSR(\beta_0, \beta_2)$

$$SSR(\beta_0, \beta_2) = y' \left[H_{\beta_0, \beta_2} - \left(\frac{1}{n}\right) J \right] y$$

Hat Matrix

We have for OLS $\beta = (X'X)^{-1} X' Y$

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

$$\hat{Y} = \underbrace{H}_{}Y \quad H = \text{hat matrix}$$

Var, Cov. Matrix

$$C = \hat{\sigma}^2 (X'X)^{-1}$$

$\hat{\sigma}^2 \to$ error Mean square.

$$S.E(\hat{\beta}_j) = \sqrt{C_{jj}}$$

t test $\dfrac{\hat{\beta}_j}{|se(\hat{\beta}_j)|}$

Properties of the hat Matrix.

1) Symmetric $H = H^T$

$(AB)^T = B^T A^T$

$$H = X(X^T X)^{-1} X^T.$$

$$H^T = \left[X (X^T X)^{-1} X^T\right]^T = X \left[X (X^T X)^{-1}\right]^T$$

$$= X \left[(X^T X)^{-1}\right]^T X^T$$

We know $X^T X$ is symmetric

$\therefore \left[(X^T X)^{-1}\right]^T = (X^T X)^{-1}$ $= X (X^T X)^{-1} X^T = H$ $X^T X$ is Symmetric.

$\therefore H^T = H.$

2) $(I-H)$ is idempotent

Means $(I-H) = (I-H)(I-H)$

$$(I-H)(I-H) = I(I-H) - H(I-H)$$

$$= I - H - H + H \cdot H$$

$$= I - H \cancel{-H} \cancel{+H} = I - H$$

$[H \cdot H = H]$ We use this

$$X(X^T X)^{-1} X^T \cdot X (X^T X)^{-1} X^T$$

$$= X (X^T X)^{-1} X^T = H$$

Var ($\hat{e}$)

$$\hat{e} = y - \hat{y} \qquad \hat{y}_i = H y_i$$

$$\hat{y} = Hy$$

$$\therefore \hat{e} = y - Hy = (I-H)y$$

$$\text{Var}(\hat{e}) = \text{Var}[(I-H)y]$$

I For MLR

$$\beta_j \pm t_{\alpha/2,\, n-(k+1)} \sqrt{c_{jj}}$$

— x — x — x — x —

R Studentised Residual

$$\text{DFFITS}_j = \hat{y}_j - \hat{y}_j(j)$$

Change in y pred. jth obs due to deletion of jth obs.

Std Form

$$= \frac{\hat{y}_j - \hat{y}_j(-j)}{\hat{\sigma}_e(-j)\sqrt{h_{jj}}}$$

If Large jth obs was influential.

Residual Analysis

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

Standardised Residuals $d_i = \dfrac{e_i}{\sqrt{MSE}}$

Studentised Residual $r_i = \dfrac{e_i}{\sqrt{MSE(1-h_{ii})}}$

→ Find outliers

dof = $n-(k+1)$

t test, 2 tailed

$t_{\alpha/2,\, dof}$, $-t_{\alpha/2,\, dof}$

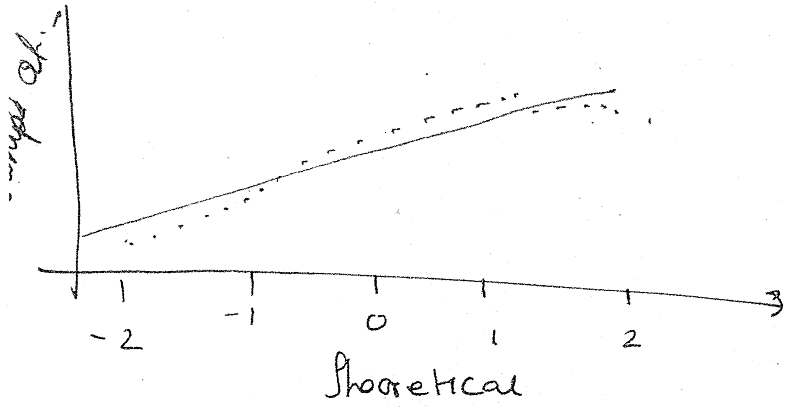
Cook's Distance 'D' [Outlier detection]

$$D_i = \frac{r_i^2}{(k+1)} \cdot \frac{h_{ii}}{(1-h_{ii})} \qquad r_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

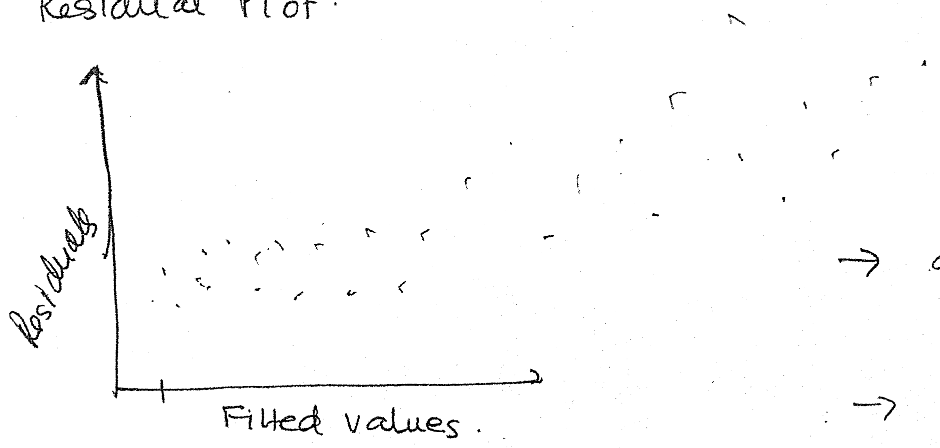
if D_i > 50th Percentile in F table

ith observation is influential

Residual plots . the QQ norm.



Residual Plot.



→ at Lower values of y e_i is Less

→ e_i ↑ at higher Values of y

→ Variance of e_i is higher at higher Values of y_i

Multi Collinearity: VIF] R_i = Regression → when ith var is ~~treated~~ as dep. regressed on

When Linear relationship involves more ~~than~~ two features. (K-1) other indep. vars.

$$VIF = \frac{1}{1-R_i^2}$$ Variance inflation factor.

Thumb Rule

VIF > 4 Suspect

VIF > 10 Strong Multicollinearity.

Feature ~~Column~~ Selections

1] Check Variance of the feature, remove features which have Low variation.

2] Forward Selection.

Forward selection   AIC OR Adjusted R^2

1] Start with Null Model. 1

2] Take one feature Run model

3] Repeat with each feature → Run 'k' times
1st Feature is set (one with Lowest AIC Value)

4] ⋮

$$AIC = 2k - 2\ln(L)$$

L = Max Value of Likelihood Function.

Continue adding features till adding another feature does not improve the Model Significantly.

$$1 + (k + (k-1) + \cdots + 1) = 1 + \sum_{i=1}^{k} i = 1 + \frac{k(k+1)}{2}$$

Note: order of Magnitude of computation is of the scale of k^2 which is Less than 2^k which would come from full grid search.

Usages of the hat Matrix.

We have. $\hat{y} = Hy$ $\therefore$ $y - \hat{y} = (I - H)y$

The Variance operator.

$$Var(\hat{e}) = Var[(I-H)y] = (I-H)\, Var(y)\, (I-H)'$$

Digression

$$Var(y) = \begin{bmatrix} Var\, y_1 & Cov(y_1, y_2) & & \\ & Var\, y_2 & & \\ & & \ddots & \\ & & & Var\, y_N \end{bmatrix}$$

Now $Var(ax+by) = a^2 Var(x) + b^2 Var(y) + 2ab\, Cov(x,y)$

Now $y_i = x_i \beta + e_i$

$$Var(y_i) = Var(x_i \beta) + Var(e_i) + 2\, Cov(x_i \beta, e_i)$$

Now $Var(\beta) = 0$

$Cov(x_i \beta, e_i) = 0$ $\therefore$ x_i is not corr. with e_i

$\therefore$ $Var(y_i) = Var(e_i) = \sigma^2_{e_i}$

Now Homoskedasticity assumption.

$$Var\, y = \begin{bmatrix} \sigma_e^2 & & \\ & \sigma_e^2 & \\ & & \ddots \\ & & & \sigma_e^2 \end{bmatrix} = \sigma_e^2 [I]$$

Thus $Var(\hat{e}) = \sigma_e^2 (I-H)(I-H)'$ as $Var(y) = \sigma_e^2 I$

$\therefore$ $Var(\hat{e}) = \sigma_e^2 (I - H)$

eg: $\left[Var(\hat{e}_{10}) = \sigma_e^2 (1 - h_{10,10}) \right]$

OLS - Estimators.

Brief Recap:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{k1} \\ \vdots & & \\ 1 & x_{1n} \; \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

$$n\times 1 \quad = \quad [n\times(k+1)]\,[(k+1)\times 1] \qquad n\times 1$$

$$y = X\beta + e \quad \therefore \quad e = y - X\hat{\beta}$$

$$\therefore SSE = (y - X\hat{\beta})'(y - X\hat{\beta})$$

Now $e'e = y'y - 2\hat{\beta}'X'y + \beta'X'X\beta$

$$\therefore \quad \frac{\partial e'e}{\partial \beta} = -2X'y + 2X'X\hat{\beta} = 0 \; \Big] \text{ Normal equation}$$

$$\Rightarrow \quad \underbrace{X'X\hat{\beta} = X'y}_{\text{Normal Eqn's.}} \quad \Rightarrow \quad \hat{\beta} = (X'X)^{-1}X'y$$

$$y = \underbrace{X\hat{\beta} + e} \qquad (X'X)\hat{\beta} = X'(X\hat{\beta} + e)$$

$$\underbrace{X'X\hat{\beta}} = \underbrace{X'X\hat{\beta}} + X'e$$

$$\Rightarrow X'e = 0$$

$\longrightarrow$ n obs

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & & & \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix}_{n\times(k+1)} \quad X' = \begin{bmatrix} 1 & 1 & \\ x_{11} & x_{12} & \cdots \\ x_{21} & x_{22} & \\ \vdots & \vdots & \\ x_{k1} & x_{k2} & \end{bmatrix}_{(k+1)\times n} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{n\times 1}$$

$\Rightarrow$

Multiplying 2nd Row out

$$\underbrace{x_{11}}\, e_1 + \underbrace{x_{12}}\, e_2 + \cdots \underbrace{x_{1n}}\, e_n = 0$$

$$x_{i1} e_1 + x_{i2} e_2 + \cdots x_{in} e_n = 0$$

$\Rightarrow$ Each Regressor has zero correlation with residuals.

First Row Multiplied out:

$$\sum e_i = 0$$

Predicted Value of 'y'

$$\hat{y} = X\hat{\beta} \qquad \underbrace{\hat{y}'e} = (X\hat{\beta})'e = \hat{\beta}'\underbrace{X'e}_{=0} = 0$$

$\hat{y}$ Predicted is uncorrelated with residuals.

Gauss Markov Assumptions:

1] $y = X\beta + e$ y, X are Linearly related.

2] X is a $n \times k$ Matrix of full rank

$\rightarrow$ Cols of X have No perfect multicollinearity.

$\rightarrow$ Cols of X are linearly independent

$E(y) = X\beta$

$\Omega = \sigma^2 I \rightarrow$ homoskedasticity.

X] May be fixed or random but must be generated by a mechanism that is unrelated to e

$$\epsilon [X \sim N[0, \sigma^2 I]$$

Gauss Markov Theorem:

OLS is the best linear unbiased and efficient Estimator.

Proof: $\hat{\beta}$ is the unbiased estimator of β

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y \qquad y = X\beta + e$$

Thus $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T (X\beta + e)$

$$= \beta + (X^T X^{-1}) X^T e$$

$$E[\hat{\beta}] = E[\beta] + (X^T X^{-1}) X^T E(e) \qquad \text{but } E(e) = 0$$

$$\therefore E[\hat{\beta}] = \beta$$

Proof $\hat{\beta}$ is a linear estimator of β.

$$\hat{\beta}_{OLS} = \beta + (X^T X)^{-1} X^T e$$

$$\text{or } \hat{\beta} = \beta + Ae$$

CROSS VALIDATION

- A Model validation Technique for assessing how the results of a statistical analysis will generalize to an independent data set.
- Involves partitioning a sample of data into Complimentary subsets, performing the analysis on one subset (training), validating the analysis on the other (Validation set)

'k' fold Cross validation

- Original sample is randomly partitioned in 'k' equal sized sub samples.
- One sub sample is retained as validation set, the other k-1 sub samples form the training set.
- The cross validation process is repeated 'k' times.
- Each 'k' sub samples are exactly used once.
- The 'k' results can then be averaged to produce a single estimation.

Adv] All observations are used for both training and validation.

Can be used for feature selection.

Q] How?

e.g.] Say 20 protein expression levels are being used to test whether a cancer patient will respond to a drug.

→ Which subset of features give the best in-sample error rates?

Limitations: Cross validation only yields meaningful results if the validation set and training set are drawn from the same population and only if human biases are controlled