Multiple Linear Regression

Look up Fundamental Assumptions.

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

β_0 = Expected value of y when $x_i = 0$

β_i = Change in y For unit change in x_i

ϵ_i = Random var with zero Mean, Std. dev = σ_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

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

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

Thus we have

$$\underset{[n\times 1]}{Y} = \underset{n\times(k+1)}{X}\ \underset{(k+1)\times 1}{\beta}$$

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 $\hat{e}^2 = [y'y + (X\beta)'X\beta - (X\beta)'y - y'X\beta]$ ④

$$\hat{e}^2 = [y'y + \beta'X'X\beta - \beta'X'y - y'X\beta]$$ ④

Now For minimising squared error

$\beta'X'y$ and $y'X\beta$ → Same scalar Product

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

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

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

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

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

$$\hat{y} = Hy$$

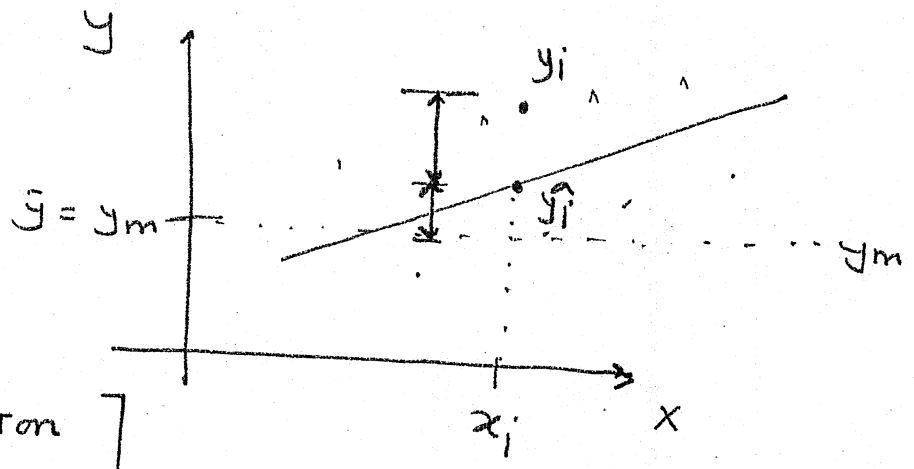
H = hat matrix

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

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

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

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 both sides

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

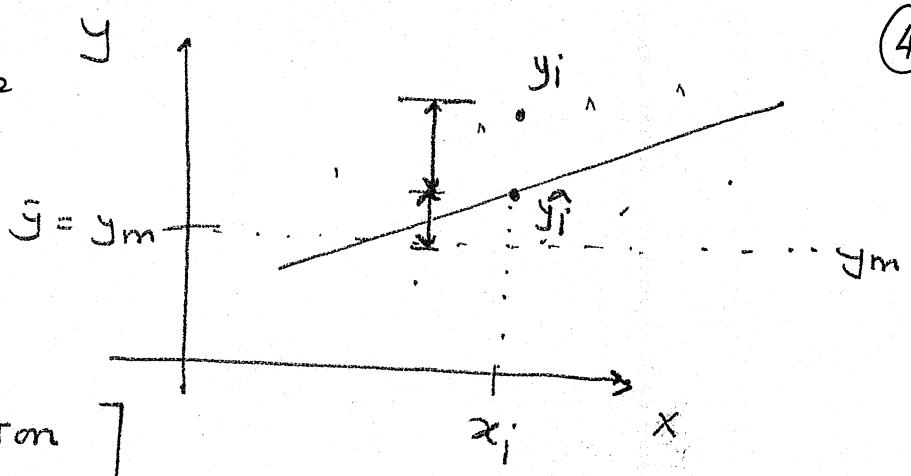
on Simplifying [SST] $= \underbrace{(y_i - \hat{y}_i)^2}_{SSE} + \underbrace{(\hat{y}_i - \bar{y})^2}_{SSR}$

Tot S.S = SSE + SSR

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

Regression

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 both sides

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

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

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

Total = E + R

$(n-1)$ $(k-1)$

$k = p+1$

F Test of Linear Regression.

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

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

$F\ stat = \frac{MS_R\ of\ Reg}{MS_E\ of\ error.} = \frac{SSR/df_R}{SSE/dfe.}$

Example:

$n = 35$ observations.

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

$SSE = 134$

$SSR = 289$

Test α_0 0.05

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

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

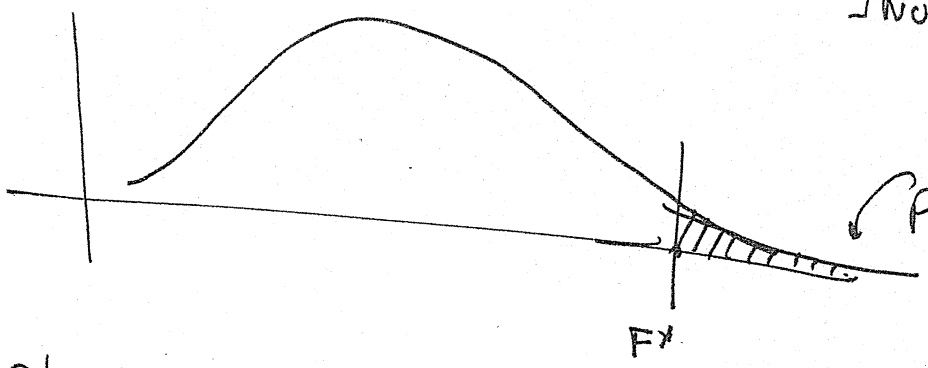
Look up Ftable (9, 25)

$\alpha = 0.05$

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

[$F_{Statistic} = 5.99 > F_{Critical}$] Null Rejected

∴ Model is good



Now check using p values:

$P_{crt} = 0.05$

Fstat = 5.99] (9, 25)

We see from table for Fstat = 5.99, (9,25)
Pvalue will be less than 0.01

$\Rightarrow P_{less} < P_{crt}$ Null is Rejected

Significance of Individual Coefficients:

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

S.d - Var = $V\sigma^2$

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

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

Regression Estimate

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

Var. Covar.

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

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

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

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

Test Statistic. [Significance of Individual Coefficients]

$\alpha = 0.05$ $\quad T_0 = \frac{\hat{\beta}_j}{SE(\beta_j)}$ $\quad H_0\ \beta_j = 0$ $\quad H_1\ \beta_j \neq 0$

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

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

also called Partial test.

Partial F test

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

$(k+1)$

$\Theta_1 \rightarrow$ Contains First $(k+1)-q$ Coeff.

$\Theta_2 \rightarrow$ Contains last q coeff.

thus $\Theta_1 = [\beta_0, \beta_1 \cdots \beta_{k-q}]$

$\Theta_2 = [\beta_{k-q+1} \cdots \beta_k]$

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

$$F_0 = \frac{SSR(\beta_1/\beta_2)}{MSE} \Big/ q$$

as only 1 Param.

$\therefore q = 1$

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

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

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

$$\hat{y} = HY$$

H = hat matrix

Var, Cov. Matrix

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

↳ error Mean square.

$$S.E(\beta_j) = \sqrt{C_{jj}}$$

t test $\frac{\beta_j}{|se(\beta_j)|}$

Properties of the hat Matrix.

1) Symmetric $H = H^T$

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

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

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

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

We know X^TX is symmetric

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

X^TX is Symmetric.

$\therefore H^T = H$

2) $(I-H)$ is idempotent

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

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

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

$$= I - H - H + H - H$$

$$= I - H - H + H = I - H$$

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

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

Var ($\hat{e}$)

$\hat{e} = y - \hat{y}$ $\quad y \& \hat{y} = Hy$

$\hat{y} = Hy$

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

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

I For MLR

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

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 = $n-(R+1)$

t test, 2 tailed

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

Cook's Distance 'D' [Outlier detection]

$D_i = \dfrac{r_i^2}{(R+1)} \cdot \dfrac{h_{ii}}{(1-h_{ii})}$ $\quad r_i = \dfrac{e_i}{\sqrt{MSE(1-h_{ii})}}$

If $r_i >$ 50th Percentile in F table

ith observation is influential

R Studentised Residual

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

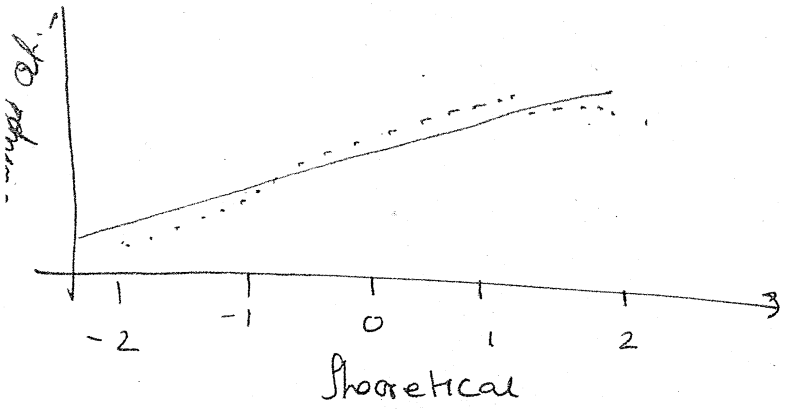
↳ Change in $y_{prod.}$ jth obs due to deletion of jth obs.

↳ Std form

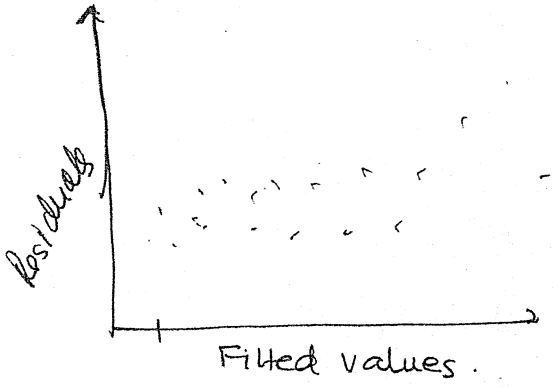
$= \dfrac{\hat{y}_j - \hat{y}_{j(-j)}}{\hat{\sigma}_{e(-j)} \sqrt{h_{jj}}}$

If Large jth obs was influential.

Residual plots . the QQ norm.



Residual Plot.



→ at Lower values of y e_i is Less

→ e_i ↑ at higher Values

→ 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 Multi Collinearity.

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.

2] Take one feature Run model
Repeat with each feature → Run 'k' times

3] 1st Feature is set (one with Lowest AIC Value)

4] ⋮

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

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

L = Max Value of Likelihood Function.

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

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. $\hat{y} = Hy \quad \therefore \quad y - \hat{y} = (1 - H)y$

The Variance operator

$$Var(\hat{e}) = Var[(I-H)y] = (I-H)\,Var(y)\,(1-H)'$$

Digression

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

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 $Var(\beta) = 0$

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

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

Now Homoskedasticity assumption.

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

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

$\therefore \quad Var(\hat{e}) = \sigma^2_e [I-H]$

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

OLS - Estimators.

Brief Recap:

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

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

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

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

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

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

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

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

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

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

$$\Rightarrow X'e = 0$$

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & & & \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix} \quad n \times (k+1)$$

$$X' = \begin{bmatrix} 1 & 1 & & \\ x_{11} & x_{12} & \cdots & \\ x_{21} & x_{22} & & \\ \vdots & \vdots & & \\ x_{k1} & x_{k2} & & \end{bmatrix} \quad (k+1) \times n \qquad \longrightarrow n \text{ obs}$$

$$e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \quad n \times 1$$

$\Rightarrow$

Multiplying 2nd Row out

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

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

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

First Row Multiplied out:

$$\Sigma e_i = 0$$

Predicted Value of 'y'

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

$\hat{y}$ 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

→ Cols of X have No perfect Multicollinearity.

→ Cols of X are linearly independent

$$E(y) = X\beta$$

$$\Omega = \sigma^2 I \rightarrow \text{homoskedasticity}$$

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

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

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.

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

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