WEEK-6 - PREDICTIVE MODELLING

CORRELATION:

n observations for x and y variables

$$\bar{x} = \frac{\sum x_i}{n}, \quad \bar{y} = \frac{\sum y_i}{n}$$

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

$$S_{xy} = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

Correlation the strength of association b/w 2 vars

Pearson's Correlation

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2}\sqrt{\sum y_i^2 - n\bar{y}^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

- r_{xy} takes a value b/w -1 and 1
- $r_{xy} = 0$ means no correlation

- used to measure degree of linear dependence
- can't be applied to ordinal variables
- sample size : (20-30)
- outliers can lead to misleading values.

Spearman Rank Correlation:

↳ Degree of association b/w 2 variables
↳ linear or non linear association.
↳ x increases, y increases / decrease monotonically.

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

d_i is the diff in the ranks given to the 2 variables values for each item of the data.

- r_s takes a value b/w -1 and 1
- $r_s = 0$ means no association
- Monotonicaly ↑sing $r_s = 1$
- Monotonicaly ↓sing $r_s = -1$
- can be used when association is non linear.
- can be applied for ordinal variables.

Kendall rank correlation coefficient:

- eff to measure association b/w 2 ordinal vars
- concordant pair:

a pair of observations (x_1, y_1) and (x_2, y_2) that follows the property $x_1 > x_2$ and $y_1 > y_2$

of $x_1 < x_2$ and $y_1 < y_2$

→ <u>Discordant pair:</u> A pair of observations (x_1, y_1) and (x_2, y_2) that follows the property $x_1 > x_2$ and $y_1 < y_2$ or $x_1 < x_2$ and $y_1 > y_2$

→ $$\boxed{\tau = \frac{(\text{No. of concordant pairs}) - (\text{No. of discordant pairs})}{n(n-1)/2}}$$

→ the pair for which $x_1 = x_2$ and $y_1 = y_2$ are not classified as concordant or discordant and are ignored.

LINEAR REGRESSION:

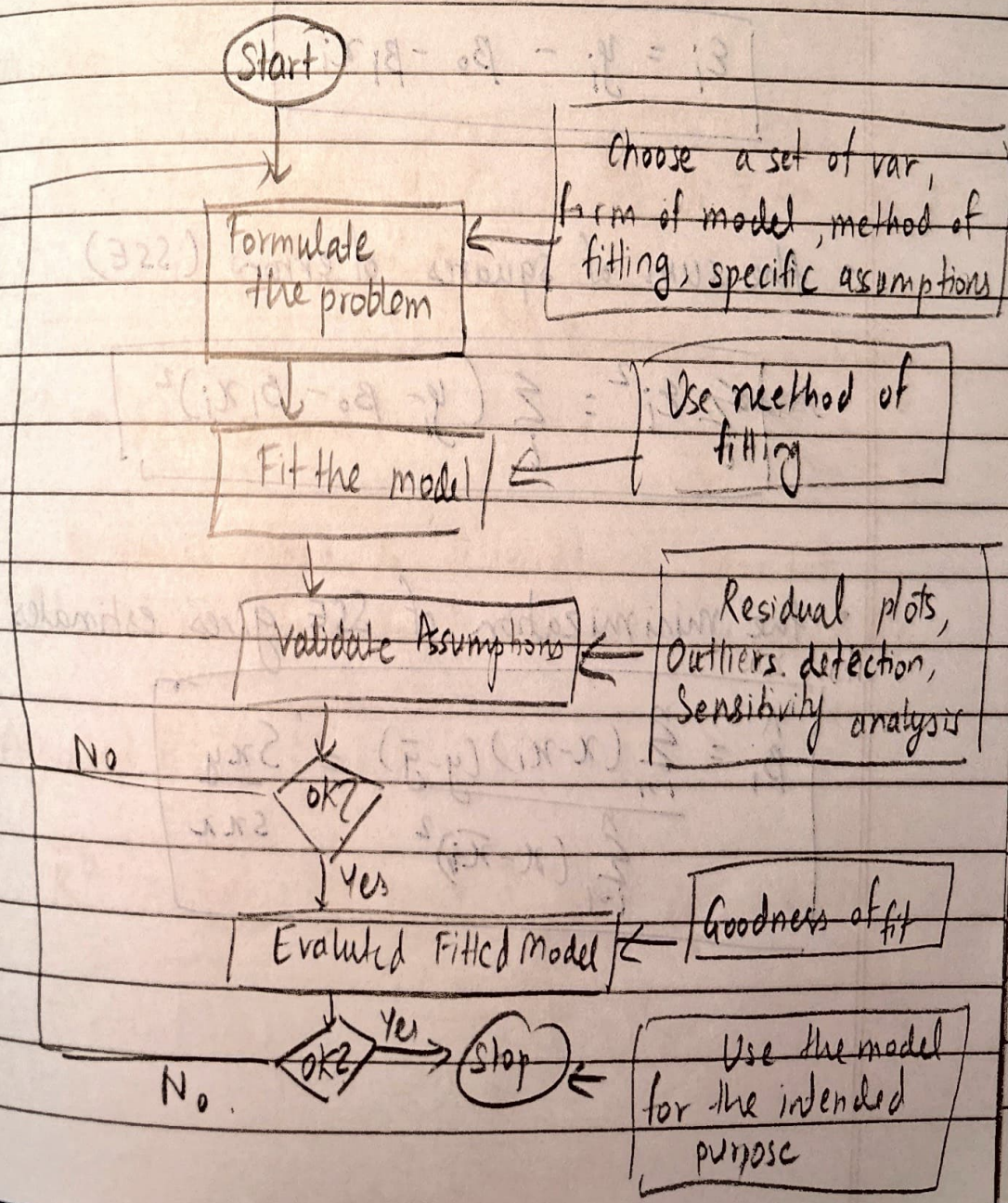
- <u>Dependent variable</u> – response variable, Regressand, Predicted variable, output var denoted by y

- <u>Independent variable</u> – predictor variable, regressor, exploratory variable, input variable denoted by x.

Regression Type

1. Univariate vs Multivariate
2. Linear vs Nonlinear
3. Simple vs Multiple.

Regression Process (Iterative):



Ordinary Least Squares (OLS):

- Linear model b/w y_i and x_i x_i $i=1 \ldots n$

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \epsilon_i}$$

- Error in only dependent var.

$$\boxed{\varepsilon_i = y_i - \beta_0 - \beta_1 x_i}$$

- the sum of squares of errors (SSE)

$$\boxed{\sum_i \varepsilon_i^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2}$$

- the minimization of SSE gives estimates β_0 and β_1

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x - x_i)(y - \bar{y})}{\sum_{i=1}^{n} (x - \bar{x}_i)^2} = \frac{S_{xy}}{S_{xx}}}$$

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

$$\boxed{\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i}$$

→ Coefficient of determination – R^2 is a measure of variability in o/p ~~var~~ var explained be input variable.

$$\boxed{R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}}$$

→ variability explained by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

→ total variability in y

R^2 → b/w 0 and 1

↳ values close to 0 indicates poor fit

↳ values close to 1 indicates a good fit.

↳ Adjusted $\bar{R}^2$

$$\bar{R}^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2 / (n-p-1)}{\sum (y_i - \bar{y})^2 / n-1}$$

→ R code:

lm(formula = Minutes ~ Units)

MODEL ASSESSMENT:

Assumptions about errors:

- Normality: Errors are normaly distributed
- Homoscedaetty sticity: Errors in diff samples have same variance.
  - $\epsilon_i \sim N(0, \sigma^2)$

→ OLS: Properties of Estimates:

- Both $\hat{\beta}_0$ and $\hat{\beta}_1$ estimates are unbiased.

$$E[\hat{\beta}_0] = \beta_0 \quad , \quad E[\hat{\beta}_1] = \beta_1$$

- Variance:

$$var[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}} \quad , \quad var[\hat{\beta}_0] = \frac{\sigma^2 \sum x_i^2}{n S_{xx}}$$

- Estimate of σ^2

$$\hat{\sigma}^2 = \frac{\Sigma(y_i - \hat{y}_i)^2}{n-2} = \boxed{\frac{SSE}{n-2}}$$

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

⇒ OLS : Confidence Intervals on regression coefficients:

① 95% 2 sided CI for $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\beta_1 \in \left[\hat{\beta}_1 - \underbrace{2.18}_{t_{0.025,12}} s_{\hat{\beta}_1}, \hat{\beta}_1 + 2.18 s_{\hat{\beta}_1}\right]$$

$$s_{\hat{\beta}_1} = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{(n-2) S_{xx}}}$$

$$\beta_0 \in \left[\hat{\beta}_0 - 2.18 s_{\hat{\beta}_0}, \hat{\beta}_0 + 2.18 s_{\hat{\beta}_0}\right]$$

$$s_{\hat{\beta}_0} = s_e \sqrt{\frac{\Sigma x_i^2}{n S_{xx}}}, \quad s_e = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{(n-2)}}$$

⇒ OLS : Hypotheses test on regression coeff

↳ in order to check if linear model fit is good or not we can test whether $\hat{\beta}_1$ is significant (different from zero) or not

- Null Hypothesis: $H_0 : \beta_1 = 0$
  Alternative Hypothesis: $H_1 : \beta_1 \neq 0$

* Null Hypothesis implies $\boxed{\hat{y}_i = \hat{\beta}_0 + \epsilon_i}$ → Reduced Model

Alternative Hypothesis implies $\boxed{\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i}$ ↙ Full model

☆ Do not Reject null hypothesis if CI for β_1 includes 0

Similarly if CI for $\hat{\beta}_0$ includes 0, then intercept term is insignificant.

⇒ OLS : Sum Squared Quantities

used for goodness of linear model.

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

error

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

residuance

$$SST = \sum (y_i - \bar{y})^2$$

↳ total

* $SST > SSE$

⇒ OLS : F-Test for choosing b/w models

↳ F test for rejecting reduced model

↳ SST is goodness of fit for reduced model (null hypothesis)

↳ SSE is goodness of fit for full model (alternative hypothesis.

↳ F-statistic $\boxed{F_0 = \frac{SST - SSE}{SSE/(n-2)} = \frac{SSR}{SSE/(n-2)}}$

At 5% level of significance reject null hypothesis if $F_0 \geq F_{(1, n-2, 0.05)}$ (upper critical value of F distribution with 1 and n-2 dfs

*Numerator has 1 df

OLS: Residual plots:

A straightforward method for assessment of a model is by analysing residuals using RESIDUAL PLOTS

residual definition for OLS

$$\boxed{e_i = y_i - \hat{y}_i} \qquad i = 1, 2, \ldots n.$$

variance of e_i is not same for all data points and also correlated.

$$\boxed{Var(e_i) = \sigma^2(1 - p_{ii})}, \quad \boxed{p_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

$$\boxed{Cov(e_i e_j) = -\sigma^2 (p_{ij})}, \quad \boxed{p_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\Sigma (x_i - \bar{x})^2}}$$

→ Standardized residual

$$\boxed{z_i = \frac{e_i}{s_e \sqrt{1 - p_{ii}}}}$$

↳ if residual variance is estimated from data then standardized residual has a t distribution with (n-2) df.

→ Residual plots:

Plot of residuals vs predicted (fitted) value of dependent variable.

→ Residual plots are used for assessing:
① validity of the linear model
② Normality of the errors
③ Homoscedastic vs heteroscadastic error.

Normal Q-Q Plot:

↳ Plot of sample quantiles against quantiles from normal distribution.

↳ quantile (or percentile) is the data value below which a certain percentage of data lies

↳ quantiles for a standard normal distribution

10%	20%	30%	40%	50%	60%	70%	80%	90%
-1.28	-0.84	-0.52	-0.25	0.00	0.25	0.52	0.84	1.28

↳ Sample quantiles:
- Arrange standardized samples in ↑sing order.
- Choose no. of quantiles.
- Find corresponding quantiles of standard normal distribution

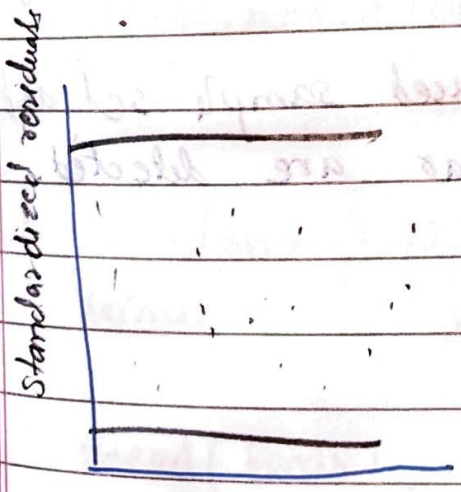
- If data points fall on 45° line, then data is from Normal distribution

↳ Function $\boxed{\text{qqnorm}(x)}$ in R

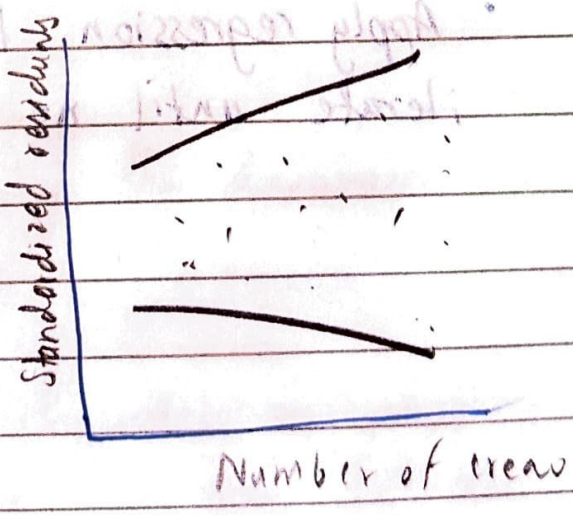
→ OLS: <u>Checking for non-uniform error variance:</u>

Check for heteroscedasticity of error:

Standardized residuals vs. estimated y. (or x for univariate case)



Uniform var



Heteroscedastic errors.

→ OLS: Checking for outliers in data:

↳ Outliers: Points which do not conform to the pattern in the bulk of data.

↳ outliers can be identified using hypothesis test of ~~red~~ residual of each sample.

- for 5% level a sample is considered an outlier if the corresponding standardized residuals ~~et~~ lie outside [-2, 2]
- even if several residuals lie outside confidence region, identify only one outlier at every iteration.
- Apply regression to reduced sample set and iterate until no outliers are detected.

Simple Linear Regression Model building:

↳ dataset 'bonds.txt'

↳ to load to dataset

↓

$\boxed{\text{read.delim()}}$

↳ reads a file in table format and creates a data frame from it.

read.delim (file, row.names=1)

bonds ← read.delim('bonds.txt', row.name=1)

↳ View (bonds) will display the dataframe in a tabular format

↳ head(bonds) and tail (bonds) will display the first and ~~ta~~ last six rows from the dataframe.

→ Structure of the data

- Each variable and its data type
- str() → input is dataty frame.

→ see whether each of the variable datatypes are same as you expect them to be, if not coerce them (as. datatype)

→ Summary of the data:

- Gives mean and 5 number summary.

Min
1st Qu
Median
3rd Qu
Max

→ Plotting the data

plot (bonds$ CouponRate, bonds $ BidPrice, main: 'Bid Price vs Coupon Rate, xlab= "Coupon Rate", ylab='Bid Price")

→ Budding linear regression Model.

- Building linear model using the function lm()

lm(formula, data)
lm(dependent var ~ independent var)

bondsmod ← lm(bonds$BidPrice ~ bonds$CouponRate)
(or)

bondsmod ← lm(BidPrice ~ CouponRate, data = bonds)

$$\star \boxed{\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i}$$

intercept, slope, residuals.

→ Fitting the regression line over the plot

abline(bondsmod)

summary(bondsmod)
↳ gives summary

⇒ Model Assessment:

→ First level of model assessment:

- R^2 value
- Hypothesis testing
  - ↳ coeff
  - ↳ Full model and reduced model.

→ First level assessment - Hypothesis test on coeff

- inorder to check whether linear model is good we can check if the estimate $\hat{\beta}_1$ is significant.
- Hypothesis testing

  Null hypothesis $H_0 : \hat{\beta}_1 = 0 \Rightarrow \boxed{\hat{y}_i = \hat{\beta}_0 + \varepsilon_i}$ ↗ reduced model

  Alternate hypothesis $H_1 : \hat{\beta}_1 \neq 0 \Rightarrow \boxed{\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i}$

  ↓ Full model.

  ↙ 2 sided test

- at $\alpha = 0.05$ ie 95% confidence level.

alpha = 0.05

$n = 35 \rightarrow$ no. of observation.

$p = 1 \rightarrow$ no. of independent var.

qt(p = 1 - (alpha/2), df = n-p-1)

↳ confidence interval on $\hat{\beta}_1$

$$\hat{\beta}_1 - (qt(value) * s_{\hat{\beta}_1})$$

to

$$\hat{\beta}_1 + (qt(value) * s_{\hat{\beta}_1})$$

→ Hypothesis test on models:

- Computing F statistics

$$\boxed{F_0 = \frac{SST - SSE}{SSE/(n-2)} = \frac{SSR}{SSE/(n-2)}}$$

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

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

SSE ← sum((bonds$BidPrice - bondsmod$fitted values)^2)

SSR ← sum((bondsmod$fitted.values − mean(bonds$BidPrice))^2)

n = 35

(SSR/SSE) * (n-2)

* this F statistic is returned by the ~~summ~~ summary command.

→ Checking for outliers

→ even if several residuals lie outside confidence region, identify only one outliear at every iterat.
→ apply regression to reduced sample set.
→ iterate until no outliers are detected

→ Residual analysis.

plot(bondsmod$fitted.values, rstandard(bondsmod), main = "Residual Plot", xlab = "Predicted values for Bid Price", ylab = "Standardized Residuals")

abline(h=2, lty=2) > line type

abline(h=-2, lty=2)

* to Know the indices of the outliers we use the function

identify()

↳ reads the position of the graphics pointers when the mouse button is pressed.
↳ it then searches the coordinates given in x and y for the point closest to the pointer.
↳ if this point is close enough to the pointer its index will be returned as a part of the value of the call.

indentify (x, y)

identify (bondsmod$fitted.values, rstandard (bondsmod))

↳ first remove the farthest ~~po~~ ~~same~~ outlier.
eg:
bonds_new ← bond [-13,]
bondsmod1 ← lm (bonds_new$BidPrice ~ bonds_new$CouponRate.

↳ repeat the same process until there are no outliers.

```
plot (bonds$CouponRate[-c(4,13,34,35)],
      bonds$BidPrice[-c(4,13,34,35)],
      main = "Bid Price vs Coupon Rate without
      outliers", xlab = "Coupon rate", ylab = "Bid Price")

abline (bondsmod1)
```

Multiple Linear Regression:

↳ dependent variable (y) depends ~~on~~ on p independent variables x_j, $j = 1, 2, \ldots p$.

↳ general linear model.

$$\boxed{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p + \epsilon}$$

for ith observation

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots \beta_p x_{p,i} + \varepsilon_i$$

↳ <u>Objective</u>: Using n observations, estimate regression coeff.

↳ approach is similar to simple regression.

↓

Minimize the sum of squares of the errors.

$$y = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}, \quad X = \begin{bmatrix} x_{1,1} - \bar{x}_1 & x_{2,1} - \bar{x}_2 & \dots & x_{p,1} - \bar{x}_p \\ \vdots & & & \\ x_{1,n} - \bar{x}_1 & x_{2,n} - \bar{x}_2 & \dots & x_{p,n} - \bar{x}_p \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

↳ the linear model in matrix form

$$\boxed{y = X\beta + \epsilon} \quad , \quad \boxed{E(\epsilon) = 0, \; Var(\epsilon) = \sigma^2 I}$$

↳ SSE

$$\boxed{S(\beta) = \epsilon^T \epsilon = (y - X\beta)^T (y - X\beta)}$$

↳ Minimization of the SSE leads to the normal eq:

$$\boxed{(X^T X)\hat{\beta} = X^T y}$$

• Assumption : $(X^T X)$ is of full rank p (invertible)

• the coeff vector

$$\boxed{\hat{\beta} = (X^T X)^{-1} X^T y} \; ; \; \boxed{\beta_0 = \bar{y} - \bar{x}^T \hat{\beta}}$$

• the properties of the estimators

$$\boxed{\begin{aligned} E(\hat{\beta}) &= \beta \\ Var(\hat{\beta}) &= \sigma^2 (X^T X)^{-1} \end{aligned}}$$

- $\hat{\beta}$ is the best linear unbiased estimator

↳ Estimate of the error variance:

$$\boxed{\hat{\sigma}^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-p-1}}$$

$-df$

↳ $1-\alpha$ confidence intervals for β_j, $j = 0, 1, \ldots, p$

$$\boxed{\beta_j \in \left[\hat{\beta}_j - t_{(n-p-1,\, \alpha/2)}\, s.e.(\hat{\beta}_j),\ \hat{\beta}_j + t_{(n-p-1,\, \alpha/2)}\, s.e.(\hat{\beta}_j)\right]}$$

$(1-\alpha/2)$ percentile point of the t-distribution with $(n-p-1)$ df

$$\boxed{\begin{gathered} s.e.(\hat{\beta}_j) = \hat{\sigma}\sqrt{c_{jj}} \\ C = (X^T X)^{-1} \end{gathered}}$$

- Multiple correlation coeff

$$\boxed{Cor(y, \hat{y}) = \frac{\Sigma(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\Sigma(y_i - \bar{y})^2 \Sigma(\hat{y}_i - \bar{\hat{y}})^2}}}$$

- the coeff of determination R^2

$$\boxed{R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2}}$$

- Adjusted R_a^2

$$\boxed{R_a^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}}$$

Reduced model

→ testing 2 models : RM with k parameters

F statistic :

$$\boxed{F_0 = \frac{[SSE(RM) - SSE(FM)] / (p+1-k)}{SSE(FM) / (n-p-1)}}$$

$$SSE(RM) \geq SSE(FM)$$

- for α significance level : Reject H_0 if

$$F_0 \geq F_{(p+1-k,\ n-p-1,\ \alpha)}$$