CHAPTER-3: MODEL SELECTION

VARIABLE SELECTION PROBLEM:
- ALL possible Regression Approach
- Sequential Selection Approach

All possible Regression

'Selecting the BEST regression model' is a crucial problem. In MLR we have large number of regressor variable, some variables can be redundant in the regression equation. We need to find the BEST possible regressors that can explain the variability in the response variable well and this problem is known as variable selection problem. One approach is "All possible regression" approach.

We consider all regression equations involving

0 regressor(s)	$\binom{K-1}{0}$	$Y = \beta_0 + \epsilon$
1 regressor (s)	$\binom{K-1}{1}$	$Y = \beta_0 + \beta_1 X + \epsilon$
⋮	⋮	⋮
K-1 regressors (s)	$\binom{K-1}{K-1}$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{K-1} X_{K-1} + \epsilon$
TOTAL	2^{K-1}	

[i.e., When $K=5$, i.e., $K-1=4$ (4 regressor variables), then there are $2^4 = 16$ possible regression equations to evaluate.]

These equations are evaluated to some suitable criteria:
- R^2
- Adjusted R^2
- Mallow's Statistic (C_p).
- AIC (Akaike's Information Criteria)
- BIC (Bayesian Information Criteria)

Sequential Selection Approach

Three approaches are there for sequential selection of BEST possible linear regression model for multi variable case:
- Forward Selection
- Backward Elimination
- Stepwise Selection

Criteria for Evaluating subset regression models:-

Coefficient of Multiple Determination (R_p^2):

Consider the following MLR model with $(p-1)$ regressors.

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + \epsilon$$

We denote regression sum of square (SS) by $SS_{Reg}(p)$ and residual SS by $SS_{Res}(p)$. Then R_p^2 is defined as follows:

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

[R_1^2 is when $p=1$, $p-1=0$, i.e., $y = \beta_0 + \epsilon$, $\Rightarrow R_1^2 = 0$.]

- Note that $R_p^2 \uparrow$ as $p \uparrow$ since $SS_{Res}(p) \downarrow$ as $p \uparrow$ and R_p^2 is maximum when $p = k$ (i.e., we have the full model).
- R_p^2 explains the percentage of total variability which is explained by the regression model for the response variable.
- For a given MLR problem, one may calculate R_p^2 for all possible 2^{k-1} regression equations and the 'rule of thumb' is to choose the model with highest R_p^2 as the 'BEST' linear regression model.
- Stopping criteria: Start with one regressor and add regressors to the model up to the point where an additional variable provides very small increase in R_p^2.

Adjusted coefficient of Multiple Determination ($\bar{R}_p^2$):

Note that $R_p^2 \uparrow$ as $p \uparrow$ since $SS_{Res}(p) \downarrow$ as $p \uparrow$

$$SS_{Res}(p) \geq SS_{Res}(p+1)$$

$$MS_{Res}(p) = \frac{SS_{Res}(p)}{n-p} < MS_{Res}(p+1) = \frac{SS_{Res}(p+1)}{n-p-1}.$$

$$\left[MS_{Res}(p) = \frac{SS_{Res}(p)}{n-p} \right]$$

- R_p^2 is not a good measure of quality of fit since it doesnot consider whether the newly added regressor variable is relevant or not. ($SS_{Res} \downarrow$ as $p \uparrow$ and $R_p^2 \uparrow$ as $p \uparrow$).
- Thus, R_p^2 can't distinguish between whether the newly added regressor variable is relevant in the model or redundant feature.
- To overcome this, $\bar{R}_p^2$ is introduced.

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

- $\bar{R}_p^2$ will not necessarily increase with the addition of any new regressor variable. For very large n, $R_p^2 \simeq \bar{R}_p^2$

- Criteria: All possible models with $(p-1)$ regressors are evaluated and model giving max $\bar{R}_p^2$ is tabulated. Select a p where R_p^2 reaches maximum.

▨ Mallow's C_p: It measures the overall bias or mean square error in the fitted model.

- Let $\hat{y}_i$ is the i^{th} fitted value and $E(y_i)$ is the expected response for the regression model. MSE is mean square error.
- C_p can be thought of: $C_p = MSE + (penalty)$.
  ↓ estimate of bias / cost on models for having extra parameter(s).
- $C_p = \frac{SS_{Res}(p)}{MS_{Res}^{Full}} - n + 2p$
- When $p = k$, $SS_{Res}(p) = SS_{Res}^{Full}$; i.e., $C_k = C_p = \frac{SS_{Res}}{MS_{Res}} - n + 2k$
  $= (n-k) - n + 2k$
  $= k$.
- Low C_p value indicates better fit.
- Criteria: Low C_p and $C_p \simeq p$ denotes the BEST model.

SEQUENTIAL SELECTION APPROACH

Recall PARTIAL t-test for MLR: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{K-1} X_{K-1} + \epsilon$

$H_0: \beta_i = 0$ vs. $H_1: \beta_i \neq 0$. Used for testing the significance of a given regressor X_i in the presence of other regressors in the model.

Test statistic: $t = \frac{\hat{\beta}_i}{\sqrt{MS_{Res}(X'X)^{-1}_{ii}}} \sim t_{n-k}$.

Reject H_0 if $|t| > t_{\alpha, n-k}$.

This can be also tested using F-statistic (extra sum of square approach).

This is known as PARTIAL F-test.

PARTIAL F-test: → We would like to determine if, other than X_i, all other regressors contribute significantly to the regression model.

$H_0: \beta_i = 0$
Vs. $H_1: \beta_i \neq 0$

large means F is large (Reject H_0)

$$F = \frac{SS_{Reg}(\text{Full model}) - SS_{Reg}(\text{Full model except } X_i)\,/\,1}{MS_{Res}^{Full}}$$

Extra sum of square due to regressor X_i (reg.)

[DF = n − (n−1)]

$\sim F_{1,n-k}$.

Reject H_0 if $F > F_{\alpha;1,n-k}$. (means X_i (ith regressor) has significant contribution to explain the variability in Y in the presence of other regressors in the model)

$Y = X\beta + \epsilon$ and $\bar{Y} = \frac{1}{n}\sum Y_i$

$SS_{Reg}^{Full} = \hat{\beta}'X'Y - n\bar{Y}^2$ has DF $(k-1)$.

$MS_{Res}^{Full} = \frac{Y'Y - \hat{\beta}'X'Y}{n-k}$;

$SS_{Reg}^{Full} - SS_{Reg}^{Rest}$ is the extra sum of square due to β_i given that other regressors in the model.

$$\frac{SS_{Reg}^{Full} - SS_{Reg}^{Rest}}{\sigma^2} \sim \chi^2_1$$

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

} independently.

Thus, $$F = \frac{\left(SS_{Reg}^{Full} - SS_{Reg}^{Rest}\right)/1}{SS_{Res}^{Full}/n-k} \sim F_{1,n-k}$$, under H_0.

If $F > F_{\alpha;1,n-k}$ we reject H_0 ⇒ X_i (ith regressor) is significant regressor variable.

– We use PARTIAL F-test in the sequential selection approach for choosing the BEST possible regression model in MLR.

Backward Elimination:

- Start with full model.
- Compute partial F statistic for each regressor in the presence of other regressors in the model.
- The regressor with smallest partial F value is removed from the model if $F < F_{OUT} = F_{0.05,1,\text{error DF}}$.
- Partial F statistics are computed for this new model and process is repeated.
- Backward elimination also terminates when the smallest partial $F > F_{OUT}$.

Example:

THE HALD CEMENT DATA

x_1	x_2	x_3	x_4	Y
7	26	6	60	78.5
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	52	6	33	95.9
11	55	9	22	109.2
3	71	17	6	102.7
1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4

So, 4 regressors and one response variable

$k-1=4$ i.e., $k=5$ coefficients.

We fit $\hat{Y} = f(x_1, x_2, x_3, x_4)$ [full model]

$\hat{Y} = 62.41 + 1.55x_1 + 0.51x_2 + 0.1x_3 - 0.144x_4$ [after OLS estimation and fitting]

→ We want to find out less significant regressor in the MLR model.

→ We need to compute the partial F-statistic associated with the regressor x_1, x_2, x_3, x_4, respectively.

→ Using standard notation, we compute

$F_{1|234}$: Partial F-statistic associated with regressor x_1 in the presence of x_2, x_3, x_4. Similarly,

$F_{2|134}$, $F_{3|124}$, and $F_{4|123}$.

→ To compute, $F_{1|234} = \dfrac{SS_{Reg}(1,2,3,4) - SS_{Reg}(2,3,4)}{MS_{Res}(1,2,3,4)}$

$= \dfrac{SS_{Reg}(x_1,x_2,x_3,x_4) - SS_{Reg}(x_2,x_3,x_4)}{MS_{Res}(x_1,x_2,x_3,x_4)}$

6

→ We need the ANOVA Table for the model

$$\hat{Y} = 62.41 + 1.55X_1 + 0.51X_2 + 0.1X_3 - 0.144X_4$$

ANOVA Table

Source of Variation	DF	SS	MS	F
Regression	4	2667.90	666.97	111.48
Residual	8 (13-5)	47.86	5.98	
Total	12 (13-1)	2715.76		

and the model $\hat{Y} = 204 - 0.923X_2 - 1.45X_3 - 1.56X_4$

(neg. model with regressors X_2, X_3 and X_4)

ANOVA Table

Source of Variation	DF	SS	MS	F
Regression	3	2641.95	880.65	107.38
Residual	9	73.81	8.20	
Total	12	2715.76		

Thus, we have $F_{1|234} = \dfrac{2667.90 - 2641.95}{5.98} = \dfrac{\text{Extra SS due to } X_1}{MS_{Res}(Full)}$

$= 4.34$

Similarly, one may compute $F_{2|134} = 0.50$,

$F_{3|124} = 0.02$,

$F_{4|123} = 0.04$.

Smallest partial F-value is $F_{3|124} = 0.02 = F_{X_3|X_1X_2X_4}$

From statistical table, we have $F_{0.05,1,8} = 5.32$

So, $F_{3|124} = 0.02 < F_{0.05,1,8} = 5.32 = F_{OUT}$

Thus, we can remove X_3 from the model.

So, we are left with X_1, X_2, and X_4.

Then, we fit least square equation $\hat{Y} = f(X_1, X_2, X_4)$.

After calculation, we have $\hat{Y} = 71.65 + 1.452X_1 + 0.416X_2 - 0.237X_4$

Again, we need to compute three partial F-values:

$$F_{1|24} = \frac{SSReg(X_1, X_2, X_4) - SSReg(X_2, X_4)}{MSRes(X_1, X_2, X_4)} = \frac{2667.79 - 1846.88}{5.33}$$

$$= 154.01$$

$$F_{2|14} = 5.03$$

$$F_{4|12} = 1.86$$

Smallest partial F-value:

$$F_{4|12} = 1.86 \quad < \quad F_{0.05,1,9} = 5.12 = F_{OUT}$$

We may remove X_4 from the model.

Now, we are left with X_1 and X_2.

We fit least square equation: $\hat{Y} = f(X_1, X_2)$

$$\hat{Y} = 52.58 + 1.468X_1 + 0.662X_3$$

Compute: $$F_{1|2} = \frac{(SSReg(X_1, X_2) - SSReg(X_2))/1}{MSRes(X_1, X_2)}$$

$$= \frac{2657.9 - 1809.4}{5.8}$$

$$= 146.52$$

$$F_{2|1} = 208.58$$

Smallest partial F-value:

$$F_{1|2} = 146.52 \not< F_{0.05,1,10} = 4.96 = F_{OUT}$$

This means, the model with X_1 and X_2; X_1 is significant in the presence of X_2 and vice versa. Thus, Backward elimination algorithm terminates and yields the final regression equation: $\hat{Y} = 52.58 + 1.468X_1 + 0.662X_2$

(BEST subset regression model using Backward Elimination algorithm)

Forward Selection: Opposite approach of Backward elimination.

Motivation: Instead of Backward elimination (we start with the full model and in each step we try to eliminate the less significant regressors from the model), here we start with a model having 'no' regressor variable and then in every step we try to find the most significant regressor variable and at every step we add one regressor variable to the model.

Algorithm:

Step-1. No regressors in the model.

Step-2. All possible models with one regressor are considered and F-statistic for each regressor is computed. The regressor having highest F-statistic value is added to the model if

$$F > F_{\alpha, 1, \text{residual DF}}.$$

Step-3. Partial F-statistics are computed for all of the remaining regressors in the presence of previously selected regressors and the one yielding the highest F is added to the model if

$$F > F_{\alpha, 1, \text{residual DF}}.$$

Step-4. Forward selection terminates when the highest partial F statistic at a particular stage does not exceed $F_{IN} = F_{\alpha, 1, \text{res DF}}$ or when the last candidate regressor is added to the model.

Example: THE HALD CEMENT DATA $\quad Y \mid X_1, X_2, X_3, X_4$

- No regressors in the model
- Partial F-value for regressor X_1 in the presence of no regressor in the model: $F_{1|-} = \underline{?}$; Model: $Y = \beta_0 + \beta_1 X_1 + \epsilon$.

Similarly: $F_{2|-} = \underline{?}$; Model: $Y = \beta_0 + \beta_2 X_2 + \epsilon$.

$F_{3|-} = \underline{?}$; Model: $Y = \beta_0 + \beta_3 X_3 + \epsilon$

$F_{4|-} = \underline{?}$; Model: $Y = \beta_0 + \beta_4 X_4 + \epsilon$.

[9]

Fitted equation for the Htald Cement Data with one regressors (X_1):

$$\hat{Y} = 81.5 + 1.87X_1$$

$$SS_{Res} = \sum_{1}^{13} e_i^2 = \sum_{1}^{13} (Y - \hat{Y})^2 = 1265$$

$$SS_{Total} = 2715.8 \;, \quad SS_{Reg} = 1450.1$$

ANOVA TABLE

Source of Variation	DF	SS	MS	F (partial $F_{1|-}$)
Reg	1	1450.1	1450.1	12.6
Res	11	1265	115.1	
Total	12	2715.8		

$$R^2 = \frac{1450.1}{2715.8} = 53.4\%$$

- $F_{1|-} = 12.6$; $F_{2|-} = 21.96$; $F_{3|-} = 4.40$; $F_{4|-} = 22.8$

$Y = \beta_0 + \beta_1 X_1 + \epsilon$; $Y = \beta_0 + \beta_2 X_2 + \epsilon$; $Y = \beta_0 + \beta_3 X_3 + \epsilon$; $Y = \beta_0 + \beta_4 X_4 + \epsilon$

Highest F-value is $F_{4|-} = 22.8 > F_{0.05,1,11} = 4.84$.

So, X_4 is added to the model.

- Partial F for all of the remaining regressors in the presence of X_4.

$F_{1|4}$; $F_{2|4}$; $F_{3|4}$.

- $Y = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \epsilon$

$$F_{1|4} = \frac{SS_{Reg}(1,4) - SS_{Reg}(4)}{MS_{Res}(1,4)} = \frac{2641 - 1831.9}{7.5} = 108.22$$

$F_{2|4} = 0.17$

$F_{3|4} = 40.29$

Highest partial F-value : $F_{1|4} = 108.22 > F_{0.05,1,10} = 4.96$

— X_1 is now added to the model.

— Now $Y = f(X_1, X_4)$ since X_1 and X_4 have been selected.

- Now, we will check two partial F-statistic value, i.e.,

$F_{2|14} = 5.03$ (significance of X_2 in the presence of X_1 and X_4)

$F_{3|14} = 4.29$ (significance of X_3 in the presence of X_1 and X_4)

$$= \frac{SS_{Reg}(X_1, X_2, X_4) - SS_{Reg}(X_1, X_4)}{MS_{Res}(X_1, X_2, X_4)}$$

$$= \frac{2667.79 - 2641}{5.33}$$

$$= 5.03$$

Highest F-value (partial): $F_{2|14} = 5.03 \not> F_{0.05,1,9} = 5.12$

- We note that X_2 can't be selected in the model.
- Forward selection algorithm terminates here and yields the model $\hat{Y} = 103 + 1.44X_1 - 0.614X_4$.

Stepwise Selection: This is a combination of forward selection and Backward elimination.

Step-1: • No regressors in the model.

Step-2: All possible models with one regressor are comidered and F-statistic for each regressor is computed. The regressor having highest F statistic value is added to the model.

Step-3: Partial F-statistics are computed for all of the remaining regressors in the presence of previously selected regressors and the one yielding the highest F value is added to the model if

$F >$ a specified threshold value (5).

Step-4: (exit step) All variables in the model are evaluated with partial F-test to see if each one is still significant. At this stage, any regressor that is no longer significant is dropped from the model.

Step-5: The step-wise regression (selection) terminates when no other regressors yields a partial F greater than the threshold value and all regressors in the model remains significant.

Example: **HALD CEMENT DATA** (DRAPER & SMITH BOOK, Pg. 337)

PARTIAL F-values of variables — exit step

Variables already in Regression	x_1	x_2	x_3	x_4
—	12.6	21.96	4.40	22.80 ✓1
x_1	—	208.58	0.31	159.30 — (>5)
x_2	146.52	—	11.82	0.43
x_3	5.81	36.68	—	100.36 = $F_{4\|3}$
x_4 ✓2	108.22 ✓ — (>5)	0.17	40.29	—
x_1, x_2	—	—	1.83 ✓ <5	1.86 ✓ — <5
x_1, x_3	—	220.55	—	208.24
x_1, x_4 ✓3	—	✓5.03 = $F_{2\|14}$	4.24 = $F_{3\|14}$	—
x_2, x_3	68.72	—	—	41.65
x_2, x_4	154.01 = $F_{1\|2,4}$	—	96.94	—
x_3, x_4	22.11	12.43	—	—
x_1, x_2, x_3	—	—	—	0.09
x_1, x_2, x_4	—	—	0.02	—
x_1, x_3, x_4	—	0.50	—	—
x_2, x_3, x_4	4.34	—	—	—

- No regressors in the model at step-1.
- $F_{1|-} = 12.6$, $F_{2|-} = 21.96$, $F_{3|-} = 4.40$, $F_{4|-} = 22.8$

  ⇨ x_4 is added to the model.
- Seek the next best X:

  $F_{1|4} = 108.22$, $F_{2|4} = 0.17$, $F_{3|4} = 40.29$

  ⇨ x_1 is added to the model. So, $Y = f(x_1, x_4)$ at this step.

- Check for a possible exit: $Y = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \epsilon$

$F_{1|4} = 108.22$, $F_{4|1} = 159.30$ (both are > 5)

Both are significant (X_1 and X_4) in the presence of the other.

- Seek for the next best X:

$F_{2|14} = 5.03$ $\quad$ $F_{3|14} = 4.24$

> 5

Thus, X_2 is added in the model.

- Now, we have: $Y = f(X_1, X_2, X_4)$.

$F_{2|14} = 5.03$, $\quad$ $S_{1|24} = 159.01$, $\quad$ $F_{4|12} = 1.89$

> 5 $\quad$ > 5 $\quad$ < 5

We remove X_4 from the model. We have

$\hat{Y} = f(X_1, X_2)$.

- We seek for new candidate:

$F_{3|12} = 1.83$ $\quad$ and $F_{4|12} = 1.86$

< 5 $\quad$ < 5

No new variables are included.

- Check for possible exit:

$F_{1|2} = 146.52$, $\quad$ $F_{2|1} = 208.58$

> 5 $\quad$ > 5

We cannot remove any regressor from the model.

- Stepwise selection terminates and yields the model:

$Y = f(X_1, X_2)$.

Remark: Results of diffuent sequential selection algorithms can be diffeunt. For the HALD CEMENT DATA:

Backward elimination selects X_1, X_2.

Forward selection selects X_1, X_4.

Stepwise Selection selects X_1, X_2.