VARIABLE SELECTION:-

30 November 2017
18:38

Variable Selection For BestFit

- All Possible Regression
- Sequantial Selection

All possible Regression:

we need to consider all regression equations involving zero regressors i.e $y = \beta_0 + \epsilon$

→ Assume that, there are $(k-1)$ Regressors (Regressors means 'x' variables)

→ As part of the model development we need to estimate 'k' parameters. like $\beta_0, \beta_1, \beta_2, \ldots \beta_{k-1}$

∴ Total number of regressor model is 2^{k-1}

# of Regressors	# of Model
0	${}^{k-1}C_0$
1	${}^{k-1}C_1$
2	${}^{k-1}C_2$
⋮	⋮
k-1	${}^{k-1}C_{k-1}$
	2^{k-1}

Ex:- If we have 4 'x' variables, we have these following all possible Regressor models.

No	Regression Model	R^2	MS_{res}	$AdjR^2$	Cp
1	y=β0+ε	0	226.3	0	442.92
2	y=β0+β1X1+ε	53.4	115.06	49.2	202.55
3	y=β0+β2X2+ε	66.6	82.39	63.6	142.49
4	y=β0+β3X3+ε	28.6	176.3	22.1	315.16
5	y=β0+β4X4+ε	67.5	80.35	64.5	138.73
6	y=β0+β1X1+β2X2+ε	97.9	5.7	97.5	2.68
7	y=β0+β1X1+β3X3+ε	54.8	122.7	45.8	198.1
8	y=β0+β1X1+β4X4+ε	97.2	7.47	96.7	5.5
9	y=β0+β2X2+β3X3+ε	84.7	41.54	81.7	62.44
10	y=β0+β2X2+β4X4+ε	68	86.88	61.2	38.23
11	y=β0+β3X3+β4X4+ε	93.5	17.57	92.2	22.37
12	y=β0+β1X1+β2X2+β3X3+ε	98.2	5.34	97.6	3.04
13	y=β0+β1X1+β2X2+β4X4+ε	98.2	5.33	97.6	3.02
14	y=β0+β1X1+β3X3+β4X4+ε	98.1	5.64	97.5	3.5
15	y=β0+β2X2+β3X3+β4X4+ε	97.3	8.2	96.3	7.34
16	y=β0+β1X1+β2X2+β3X3+β4X4+ε	98.2	5.9	97.3	5

∴ $2^{(k-1)} = 2^4 = 16$

∵ $k-1$ = # of Regressors

∴ 16 models has developed

⇒ These (2^{k-1}) equations are evaluated by below criteria.

1. R^2
2. Adjusted R^2
3. MS Residual
4. Mallow Statistic (C p)

Mallow Statistic [Cp] :-

Mallow statistic measures the overall bias (or) Mean Square Error in the fitted model.

$\hat{y}$ = Predicted Response

$E(y)$ = True Response , 'k' is # of parameters (i.e, # of β's)

$$C_p = \frac{\sum E(\hat{y} - E(y))^2}{\sigma^2}$$

, Here we are dividing M.S.E with its S.D

⇒ Standardized Mean Square Error.

Here, $\hat{y} - E(y)$ is nothing but (predicted value - population value) i.e. nothing but Bias

So, In Realtime we cannot find population value i.e $E(y)$ so, we can estimated by

$$\boxed{\hat{C}_p = \frac{SS_{Res}(k)}{MS_{Res}(Full)} - n + 2k}$$

$$\boxed{C_p = \frac{\quad}{MS_{Res}(\text{Full})}}$$

where k = # of parameters including β_0

$MS_{Res}(\text{Full})$ = This is always for the full model

$SS_{Res}(k)$ = This is for 'k' parameters (i.e, (k-1) regressors)

ex:
$$\hat{C}_p = \frac{SS_{Res}(x_1)}{MS_{Res}(x_1, x_2, x_3, x_4)} - 13 + 2(2)$$

$$= 202.55$$

But the question is how do you find Best Cp from all calculated ...?

→ if any calculated Cp value is equivalent to the no of parameters i.e, going to be best Cp.