Applied Regression

Multiple Linear Regression Model (MLR Model)

Model Building - Part-1

Module 6 Lecture - 6-1

In the preceding materials, our focus was on techniques to ensure that the functional form of the model was correct and that the underlying assumptions were not violated.

In most practical problems, especially those involving historical data, the analyst has a rather large pool of possible candidate regressors, of which only a few are likely to be important. Finding an appropriate subset of regressors for the model is often called the **variable selection problem**.

Good variable selection methods are very important in the presence of multicollinearity. Variable selection does not guarantee elimination of multicollinearity. Multicollinearity is not the only reason to pursue variable selection techniques.

Building a regression model that includes only a subset of the available regressors involves two conflicting objectives.

(1) We would like the model to include as many regressors as possible so R^2 is high.

(2) We want the model to include as few regressors as possible because $\sqrt{MSE}$ as well as the variance of the prediction increases as the number of regressors increases.

The process of finding a model that is a compromise between these two objectives is called selecting the "best" regression equation. Unfortunately, as we will see in this chapter, there is no unique definition of "best." Furthermore, there are several algorithms that can be used for variable selection, and these procedures frequently specify different subsets of the candidate regressors as best.

Assume that there are k candidate regressors $x_1, x_2, \ldots, x_k$ and $n(\geq (k+1))$ observations on these regressors and the response variable is y. The full model, containing all k regressors, is

$$y_i = \beta_0 + \sum_{i=1}^{k} \beta_j x_{ij} + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

$$\text{or equivalently} \quad \mathbf{y} = \mathbf{X}\beta + \epsilon$$

Now suppose that r of the regressors are deleted and then the number of variables that are retained in a subset model is p = k + 1 - r.

$$\text{Hence the subset model is} \quad \mathbf{y} = \mathbf{X_p}\beta_{\mathbf{p}} + \epsilon$$

Two key aspects of the variable selection problem are generating the subset models and deciding if one subset is better than another. Following is list of key criteria for evaluating and comparing subset regression models.

(1) Coefficient of Multiple Determination - R^2.

(2) Adjusted R^2.

(3) Mean Square Error

(4) Mallow's C_p Statistic

(5) Akaike Information Criterion (AIC)

(6) Bayesian Information Criterion (BIC)

(7) PRESS statistic

(1) Coefficient of Multiple Determination (R^2)

A measure of the adequacy of a regression model that has been widely used by practitioners is the coefficient of multiple determination, R^2. Let R_p^2 denote the coefficient of multiple determination for a subset regression model with p terms (including β_0),

$$R_p^2 = \frac{SSR(p)}{SST} = 1 - \frac{SSE(p)}{SST}$$

where SSR(p) and SSE(p) denote the regression sum of squares and the residual sum of squares, respectively, for a p-term subset model.

By definition full model will have the highest R^2 and any the subset model will have less R^2, it is sometime difficult to compare R^2 due to its inherent nature.

Hence any subset model having $R_p^2 > R_0^2$, it can be considered satisfactory where R_0^2 computed from full model (i.e p= k+1) as

$$R_0^2 = 1 - (1 - R_{k+1}^2)(1 + d_{\alpha,n,k}) \quad \text{where } d_{\alpha,n,k} = \frac{k.F_{\alpha,k,n-k-1}}{n-k-1}$$

.

(2) Adjusted R^2:

To avoid the inherent properties, some analysts prefer to use the adjusted R_p^2 statistic, defined for a p-term equation as

$$adj - R_p^2 = 1 - \frac{SSE(p)/(n-p)}{SST/(n-1)} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R_p^2)$$

The $adj - R_p^2$ statistic does not necessarily increase as additional regressors are introduced into the model. In fact, it can be shown that $adj - R_p^2$ goes down with the addition of a bad variable. Consequently, one criterion for selection of an optimum subset model is to choose the model that has a maximum $adj - R_p^2$.

(3) Mean Square Error (MSE):

The MSE for a subset regression model, $MSE(p) = \dfrac{SSE(p)}{n-p}$

may be used as a model evaluation criterion.

The general behavior of MSE(p) as p increases is illustrated in

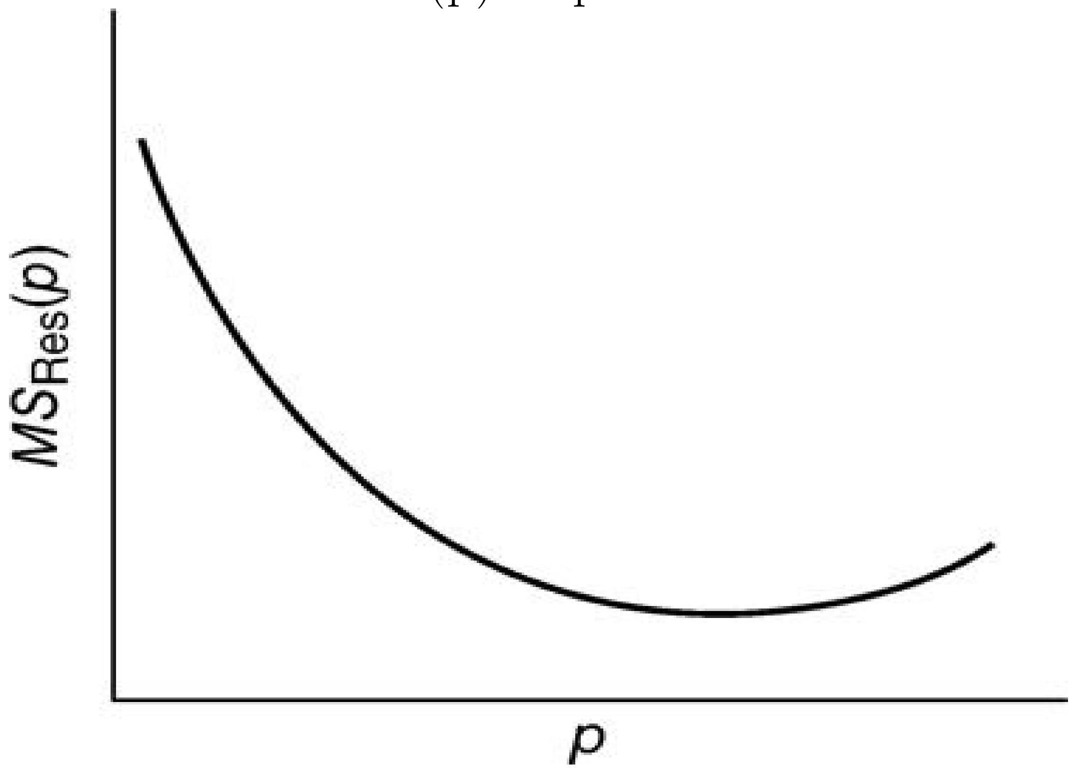

Figure 1: MSE vs p - Graph

Because SSE(p) always decreases as p increases, MSE(p) initially decreases, then stabilizes, and eventually may increase. The ideal choice is the model with least MSE(p).

But it can easily seen that it is the same criterion as choosing the model with highest $adj - R^2$, because

$$adj - R_p^2 = 1 - \frac{SSE(p)/(n-p)}{SST/(n-1)} = 1 - \frac{MSE(p)}{SST/(n-1)}$$

.

Note that SST is same for all the models.

(4) Mallow's C_p Statistic:

Mallows has proposed a criterion that is related to the mean square error of a fitted value (not MSE), that is,

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

Note that $E(y_i)$ is the true expected value of y_i (not model dependent) and $E(\hat{y}_i)$ is the expected value of the response from the p-term subset model. Thus, $(E(y_i) - E(\hat{y}_i))$ is the bias at the ith data point. Let the total squared bias for a p-term equation be

$$SS_B(p) = \sum_{i=1}^{n} \left[E(y_i) - E(\hat{y}_i)\right]^2$$

.

After going through the basic calculations, Mallows defined the estimate of the total means square error from the p-parameter model as

$$C_p = \frac{SSE(p)}{\hat{\sigma}^2} - n + 2p$$ where $\hat{\sigma}^2$is a good estimator like using pure error.

If the p-term model has negligible bias, then $SS_B(p) \approx 0$.

$$\text{Consequently,} \quad E[SSE(p)] = (n–p)\sigma^2 \quad \Rightarrow \quad E(C_p) \ = \ p$$

,

Hence when using the Cp criterion, it can be helpful to visualize the plot of Cp as a function of p for each regression equation, such as shown in Figure

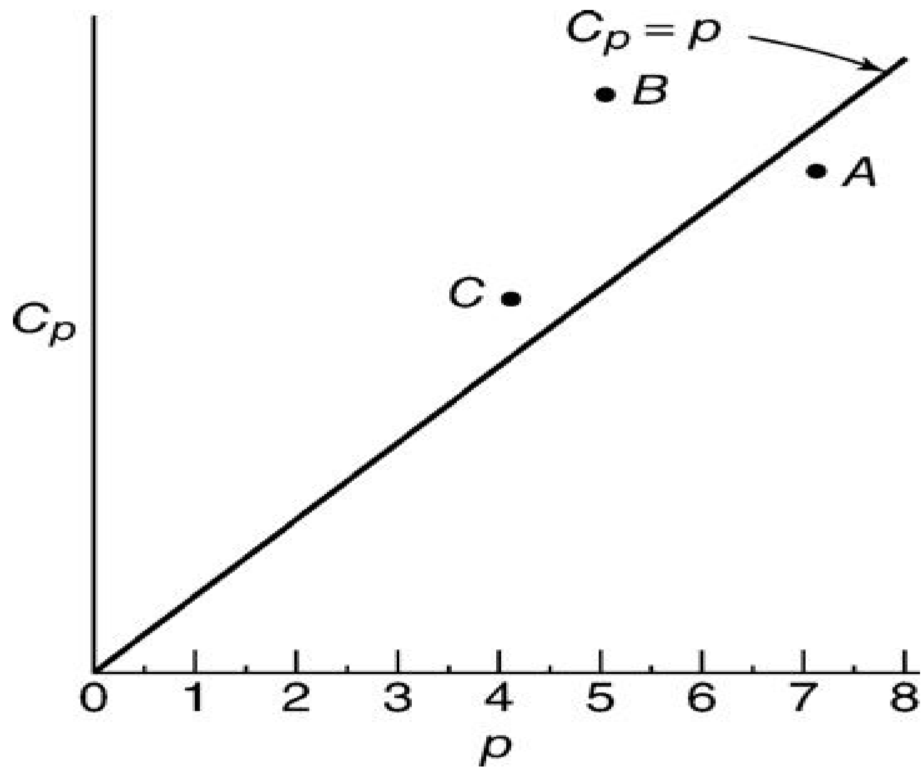

Figure 2: C_p vs p - Graph

Regression equations with little bias will have values of C_p that fall near the line $C_p = p$ (point A in the Figure). Generally, small values of C_p are desirable while those are closer to the line. For example, point C is preferred over point A. To calculate C_p, we need an unbiased estimate of σ^2. Frequently when we don't have repeat measures or other ways to get a "good" estimator for σ^2, we use the residual mean square for the full model equation for this purpose knowing that it may be biased.

(5) Akaike Information Criterion (AIC)

Akaike proposed an information criterion, AIC, based on maximizing the expected entropy of the model. Entropy is simply a measure of the expected information. Let L be the likelihood function for a specific model. The AIC is

$$AIC = -2ln(L) + 2p$$

where p is the number of parameters in the model.

In ordinary least square equation, it is

$$AIC = n\, ln\Big(\frac{SSE}{n}\Big) \;+\; 2p$$

The key insight to the AIC is similar to $adj - R^2$ and Mallows C_P. In practice it is considered that a lower AIC means a model is closer to the truth. So the model with lower AIC value is preferred.

(6) Bayesian Information Criterion (BIC)

There are several (but mainly two) Bayesian extensions of the AIC. Both are called BIC for "Bayesian information criterion".

One of them is (by Schwartz)

$$BIC = -2ln(L) + pln(n)$$

In Ordinary Least Square regression it is

$$BIC = n\,ln\Big(\frac{SSE}{n}\Big) \,+\, pln(n)$$

Remark: R uses the above formula while SAS uses a different one by Sawa.

Remark: To verify the numbers from computer print out take the numbers from R not SAS.

(7) PRESS statistic

As we have seen, there are several criteria that can be used to evaluate subset regression models. The criterion to be used for model selection should certainly be related to the intended use of the model. There are several possible uses of regression, including (1) data description, (2) prediction and estimation, (3) parameter estimation, and (4) control.

Frequently, regression equations are used for prediction of future observations or estimation of the mean response. In general, we would like to select the regressors such that the mean square error of prediction is minimized. That is where PRESS statistic comes in. For a p-term regression model PRESS statistic is described as

$$PRESS_p = \sum_{i=1}^{n} \left[y_i - \hat{y}_{(i)}\right]^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}}\right)^2$$

One then selects the subset regression model based on a small value of $PRESS_p$.

COMPUTATIONAL TECHNIQUES FOR VARIABLE SELECTION

It is desirable to consider regression models with all possible combinations of the independent variables. It is easy to see that if the number of variables is few. To find the subset of variables to use in the final equation, it is natural to consider fitting models with various combinations of the candidate regressors and look at their criteria statistic. But if there are k many regressor variables are to be considered then there are 2^k models are possible and needs to be looked at which is overwhelming if $k > 4$. So here we will look at an example with 4 regressors and later we will look at other methods.

Example: The data in appendix B.21 contains the data concerning the heat evolved in calories per gram of cement (y) as a function of the amount of each of four ingredients in the mix: tricalcium aluminate (x_1), tricalcium silicate (x_2), tetracalcium alumino ferrite (x_3), and dicalcium silicate (x_4). Though the data other issues, we will use it just to summarize the process.

Since there are k = 4 candidate regressors, there are $2^4 = 16$ possible regression equations. The results of fitting these 16 equations are displayed in the next Table.

Summary Statistic for each Model

Number of Regressors in Model	p	Regressors in Model	$SS_{\text{Res}}(p)$	R_p^2	$R_{\text{Adj},p}^2$	$MS_{\text{Res}}(p)$	C_p
None	1	None	2715.7635	0	0	226.3136	442.92
1	2	x_1	1265.6867	0.53395	0.49158	115.0624	202.55
1	2	x_2	906.3363	0.66627	0.63593	82.3942	142.49
1	2	x_3	1939.4005	0.28587	0.22095	176.3092	315.16
1	2	x_4	883.8669	0.67459	0.64495	80.3515	138.73
2	3	x_1x_2	57.9045	0.97868	0.97441	5.7904	2.68
2	3	x_1x_3	1227.0721	0.54817	0.45780	122.7073	198.10
2	3	x_1x_4	74.7621	0.97247	0.96697	7.4762	5.50
2	3	x_2x_3	415.4427	0.84703	0.81644	41.5443	62.44
2	3	x_2x_4	868.8801	0.68006	0.61607	86.8880	138.23
2	3	x_3x_4	175.7380	0.93529	0.92235	17.5738	22.37
3	4	$x_1x_2x_3$	48.1106	0.98228	0.97638	5.3456	3.04
3	4	$x_1x_2x_4$	47.9727	0.98234	0.97645	5.3303	3.02
3	4	$x_1x_3x_4$	50.8361	0.98128	0.97504	5.6485	3.50
3	4	$x_2x_3x_4$	73.8145	0.97282	0.96376	8.2017	7.34
4	5	$x_1x_2x_3x_4$	47.8636	0.98238	0.97356	5.9829	5.00

Figure 3: Summary Statistic

Consider evaluating the subset models by the R_p^2 criterion. A plot of R_p^2 versus p is shown below.

Plot - R_p^2 vs p - Example

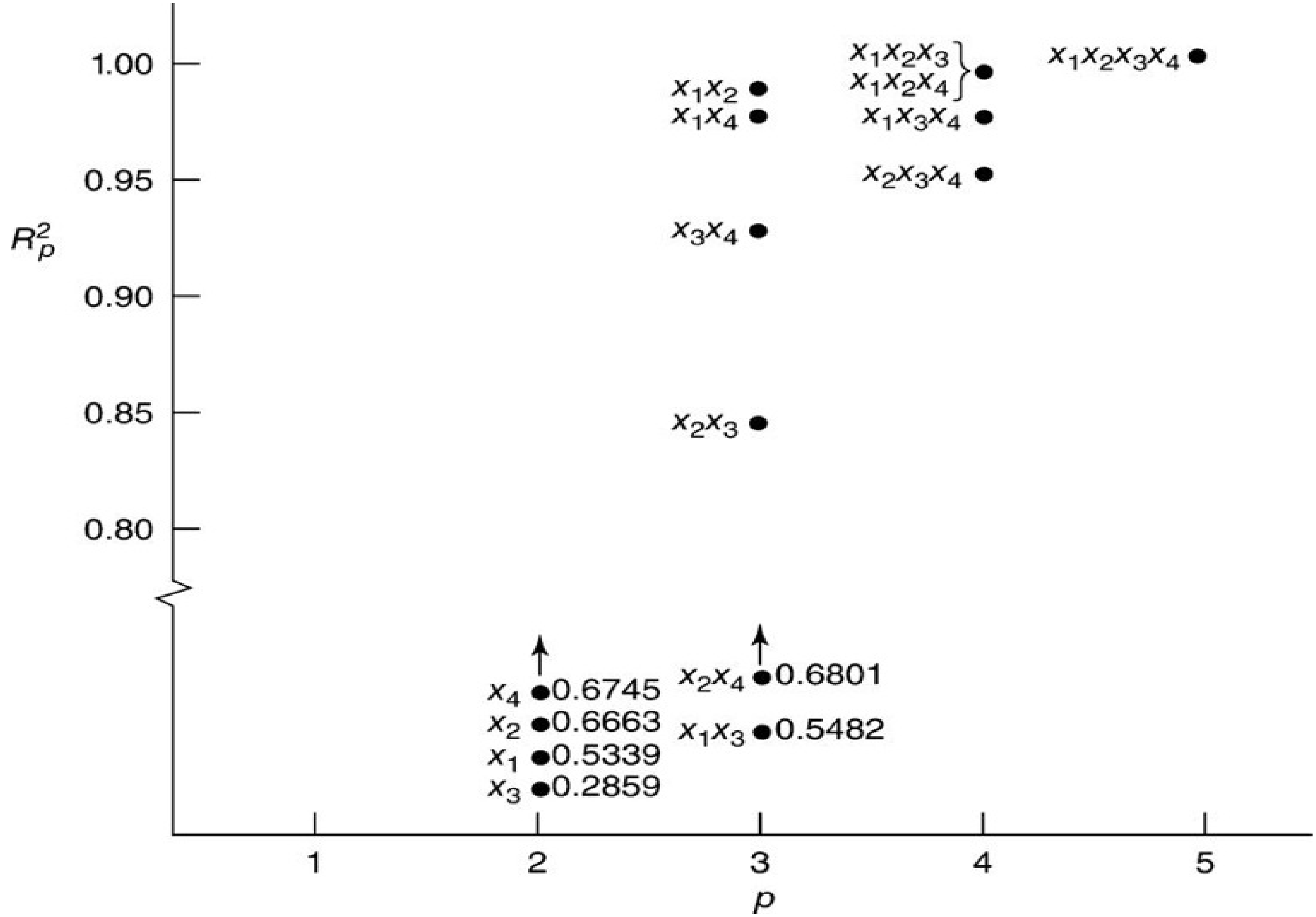

Figure 4: Summary Statistic

(1) Among the four SLR models, the model with only $X_2, (R^2 = 0.666)$ and only $X_4, (R^2 = 0.675)$ have the highest R^2, but when they are combined it did not increase that much($R^2 = 0.68$). Hence X_2 & X_4 are bringing the the similar type of information about Y and adding them does not increase very much.

(2) But examining the graph,it is clear that after two regressors are in the model, there is little to be gained in terms of R^2 by introducing additional variables. Both of the two-regressor models (x_1, x_2) and (x_1, x_4) have essentially the same R^2 values. To check we need to compute

$$\begin{aligned} R_0^2 &= 1 - (1 - R_{4+1}^2)(1 + \frac{4.F_{0.05,4,8}}{8}) \\ &= 1 - (1 - 0.98238)(1 + \frac{4.(3.84)}{8}) \\ &= 0.94855 \end{aligned}$$

As the models with (x_1, x_2) and (x_1, x_4) both have R^2 higher than R_0^2.

A plot of MSE(p) versus p reveals that the minimum residual mean square model is with (x_1, x_2, x_4), with MSE(4) = 5.3303. Note that, as expected, the model that minimizes MSE(p) also maximizes the adj_R^2.

Plot of MSE vs p - Example

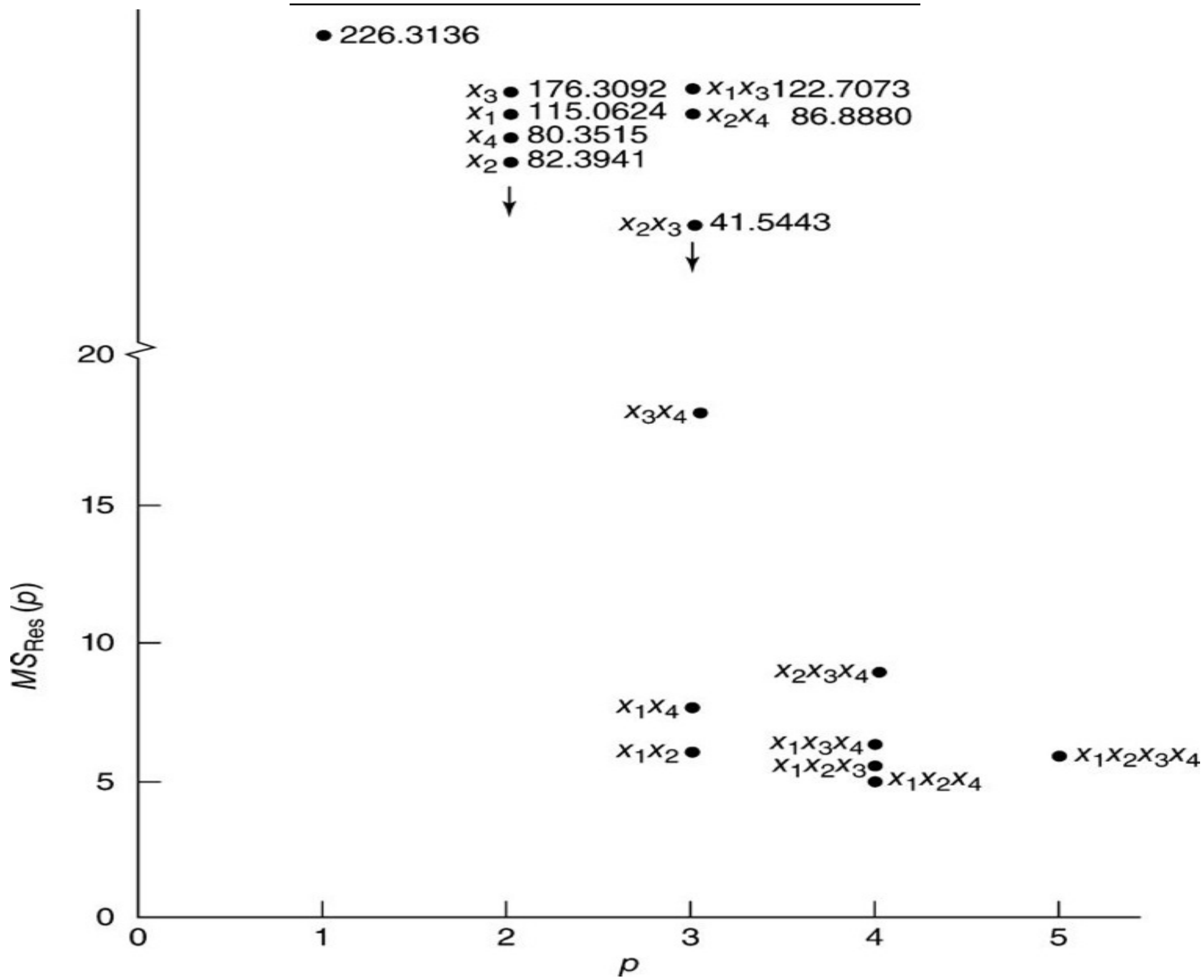

Figure 5: Summary Statistic

However, two of the other three-regressor models $[(x_1, x_2, x_3)$ & $(x_1, x_3, x_4)]$ and the two-regressor models $[(x_1, x_2)$ & $(x_1, x_4)]$ have comparable values of the residual mean square. Hence we need to look at the parameter estimates table of all of these models to justify addition of the third variable on top of the two.

To calculate C_p values, the MSE of the full model has been used to estimate the $\hat{\sigma}^2$. For example to calculate C_3 for the model with (x_1, x_4) we see that

$$C_3 = \frac{SSE(3)}{\hat{\sigma}^2} - n + 2p = \frac{74.7621}{5.9829} - 13 + 2 \times 3 = 5.50$$

Finally, from examination of the table and the plot we find that

(1) Between the 2 two-variable models (a) (x_1, x_4) and (b) (x_1, x_2), we that (x_1, x_2) is better than (x_1, x_4) almost in every respect. So (x_1, x_4) model is not a potential candidate.

(3) Among the three variable models, the model with (x_1, x_2, x_4) is best among the three in all three criterion like R^2, MSE, C_p (though we have not look at the other ones).

A C_p plot is shown in Figure below.

Plot of C_p vs p - Example

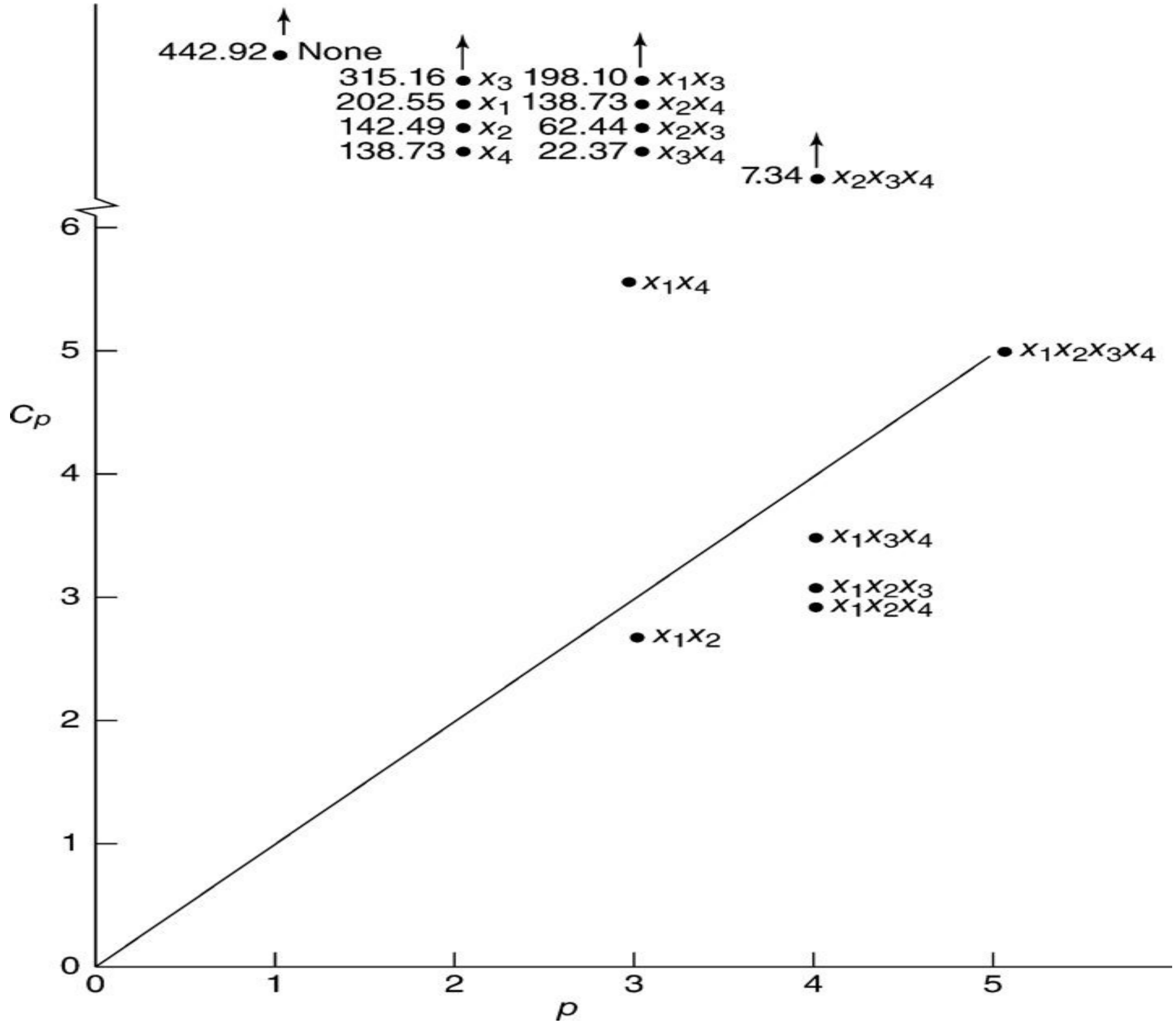

Figure 6: Summary Statistic

So to compare between these two potential winners, we look at two things (1) The PRESS statistic (2) Parameter estimates Table.

Here is the PRESS table for two models:

PRESS Statistic Calculation - Example

Observation	$\hat{y} = 52.58 + 1.468x_1 + 0.662x_2$[a]			$\hat{y} = 71.65 + 1.452x_1 + 0.416x_2 - 0.237x_4$[b]		
i	e_i	h_{ii}	$[e_i/(1-h_{ii})]^2$	e_i	h_{ii}	$[e_i/(1-h_{ii})]^2$
1	−1.5740	0.25119	4.4184	0.0617	0.52058	0.0166
2	−1.0491	0.26189	2.0202	1.4327	0.27670	3.9235
3	−1.5147	0.11890	2.9553	−1.8910	0.13315	4.7588
4	−1.6585	0.24225	4.7905	−1.8016	0.24431	5.6837
5	−1.3925	0.08362	2.3091	0.2562	0.35733	0.1589
6	4.0475	0.11512	20.9221	3.8982	0.11737	19.5061
7	−1.3031	0.36180	4.1627	−1.4287	0.36341	5.0369
8	−2.0754	0.24119	7.4806	−3.0919	0.34522	22.2977
9	1.8245	0.17195	4.9404	1.2818	0.20881	2.6247
10	1.3625	0.55002	9.1683	0.3539	0.65244	1.0368
11	3.2643	0.18402	16.0037	2.0977	0.32105	9.5458
12	0.8628	0.19666	1.1535	1.0556	0.20040	1.7428
13	−2.8934	0.21420	13.5579	−2.2247	0.25923	9.0194
		PRESS x_1, x_2 =	93.8827		PRESS x_1, x_2, x_4 =	85.3516

Figure 7: PRESS Statistic

Both models have very similar values of PRESS though model with (x_1, x_2, x_4) is higher. However, x_2 and x_4 are highly multicollinear ($r_{24} = -0.973$). Now looking at the two parameter estimates table shows that x_4 should be dropped from the model with all 3 as p-value is 0.2054 which results in the model with x_1 and x_2. If we take a look at the AIC and BIC criterion (in the last slide) we notice that the model with x_1 and x_2 is the winner as they have the smallest value. Hence the model with x_1 and x_2 should be the final model. But the other aspects of the model still needs to carried out like outlier detection and model adequacy checking etc.

Parameter Estimates - Example

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	52.57735	2.28617	23.00	<.0001
x1	x1	1	1.46831	0.12130	12.10	<.0001
x2	x2	1	0.66225	0.04585	14.44	<.0001

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	71.64831	14.14239	5.07	0.0007
x1	x1	1	1.45194	0.11700	12.41	<.0001
x2	x2	1	0.41611	0.18561	2.24	0.0517
x4	x4	1	-0.23654	0.17329	-1.37	0.2054

Summary Stat - All Models

The SAS System

Obs	ModelIndex	VarsInModel	NumInModel	SSE	MSE	Adjrsq	RSquare	Cp	AIC	BIC
1	1	x1 x2 x4	3	47.97273	5.33030	0.9764	0.9823	3.0182	24.9739	31.1723
2	2	x1 x2 x3	3	48.11061	5.34562	0.9764	0.9823	3.0413	25.0112	31.1839
3	3	x1 x3 x4	3	50.83612	5.64846	0.9750	0.9813	3.4968	25.7276	31.4057
4	4	x1 x2	2	57.90448	5.79045	0.9744	0.9787	2.6782	25.4200	29.2437
5	5	x1 x2 x3 x4	4	47.86364	5.98295	0.9736	0.9824	5.0000	26.9443	34.4130
6	6	x1 x4	2	74.76211	7.47621	0.9670	0.9725	5.4959	28.7417	30.9805
7	7	x2 x3 x4	3	73.81455	8.20162	0.9638	0.9728	7.3375	30.5759	32.9997
8	8	x3 x4	2	175.73800	17.57380	0.9223	0.9353	22.3731	39.8526	37.8866
9	9	x2 x3	2	415.44273	41.54427	0.8164	0.8470	62.4377	51.0371	46.8392
10	10	x4	1	883.86692	80.35154	0.6450	0.6745	138.7308	58.8516	55.5401
11	11	x2	1	906.33634	82.39421	0.6359	0.6663	142.4864	59.1780	55.8498
12	12	x2 x4	2	868.88013	86.88801	0.6161	0.6801	138.2259	60.6293	55.5085
13	13	x1	1	1265.68675	115.06243	0.4916	0.5339	202.5488	63.5195	60.0035
14	14	x1 x3	2	1227.07206	122.70721	0.4578	0.5482	198.0947	65.1167	59.7425
15	15	x3	1	1939.40047	176.30913	0.2210	0.2859	315.1543	69.0674	65.3850