北京师范大学
香港浸会大学
联合国际学院
BEIJING NORMAL UNIVERSITY · HONG KONG BAPTIST UNIVERSITY
UNITED INTERNATIONAL COLLEGE

STAT2013 Regression Analysis

Semester 1, 2018-2019

Room: T3-401-R1

Instructor: Ye, Huajun Terry

TA：Jenna Otto

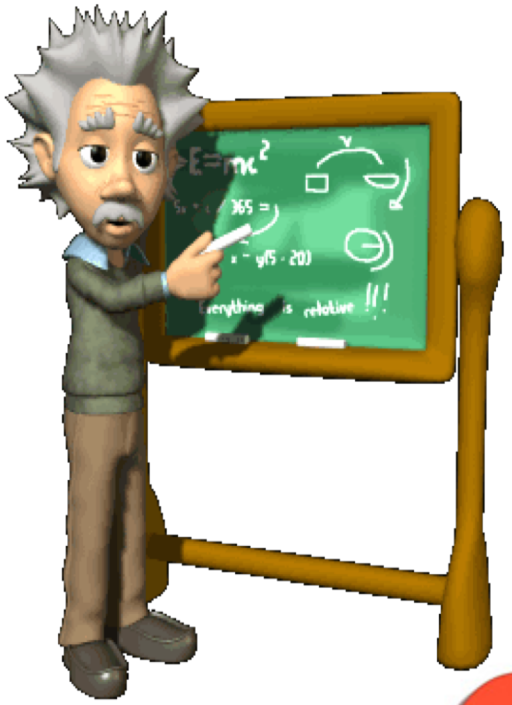
Email:hjye@uic.edu.hk

Jotto@uic.edu.hk (TA)

Chapter 4
Criteria For Choice of Best Model

In this Chapter, we will know some important concepts.

- Standard criteria for comparing models
- Cross validation for model selection and determination of model performance
- Conceptual predictive criteria
- Sequential variable selection procedures



4.0 Introduction

1. Why do we need variable selection?

   a) Reduce the expenses of collecting data;

   b) Singular Problem:
   If there are high correlations among independent variables:

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}$$

   $(\boldsymbol{X}'\boldsymbol{X})$ will be approximately singular.

   c) The effect of regression model is never good if some important predictors (independent variables) are missed.

2. What is the contribution of a predictor x_i to the current regression model?

The current model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$$

with the residual sum of squares $\boldsymbol{RSS} = \boldsymbol{SS}_{\text{Res}}$.

Delete x_i from the model above , we have

$$y = \beta_0^* + \beta_1^* x_1 + \ldots + \beta_{i-1}^* x_{i-1} + \beta_{i+1}^* x_{i+1} + \ldots + \beta_k^* x_k + \varepsilon^*$$

with the corresponding residual sum of squares $\boldsymbol{RSS}_i = \boldsymbol{SS}_{\text{Res},-i}$

Let $\boldsymbol{C} = (c_{ij}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}$, and $\hat{\boldsymbol{\beta}}$ be the LSE of $\boldsymbol{\beta}$ in the current model. It can be verified that

$$0 \le \mathbf{RRS}_i - \mathbf{RRS} = \hat{\beta}_j^2 \big/ c_{j+1,j+1} \equiv p_i$$

where p_i is called the **partial sum of squares of x_i** in the current model. The p_i can be regarded as the contribution of x_i to the current model.

The contribution of x_i strongly depends on the current model.

Example 1 (example from chapter 2)

Model (old): $y = \beta_0 + \beta_1 x1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model (old)	3	8779676741	2926558914	46.66	<.0001
Error	16	1003491259	62718204		
Total	19	9783168000			

$H_0: \beta_2 = 0$ vs $H_1: \beta_2 \neq 0$

When H0 is true, Model (new): $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model (new)	2	7833077048	3916538524	34.14	<.0001
Error	17	1950090952	114711232		
Total	19	9783168000			

Partial sum of squares of x_2 in the old model

$$= SS_{Res}(\text{new}) - SS_{Res}(\text{old})$$

$$= 1950090952 - 1003491259$$

$$= 946599693$$

It's easy to write down a code for this example.

4.1 Standard Criteria for Comparing Models

A. Coefficient of Determination R^2

$$R^2 = \frac{SS_{\text{Reg}}}{SS_{\text{Total}}}$$

If two models have different number of variables, it is unfair to compare the two models based on R^2..

Theorem 4.1

Let r be the correlation coefficient between $\{y_i\}$ and $\{\hat{y}_i\}$, that is,

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}}$$

Then $R^2 = r^2$.

$$r^2 = \frac{(\mathbf{y}'\mathbf{D}_n\hat{\mathbf{y}})^2}{(\mathbf{y}'\mathbf{D}_n\mathbf{y})(\hat{\mathbf{y}}'\mathbf{D}_n\hat{\mathbf{y}})} \quad \text{and} \quad R^2 = \frac{\hat{\mathbf{y}}'\mathbf{D}_n\hat{\mathbf{y}}}{\mathbf{y}'\mathbf{D}_n\mathbf{y}}$$

As $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, $\hat{\mathbf{y}}'\hat{\mathbf{y}} = \mathbf{y}'\mathbf{H}^2\mathbf{y} = \mathbf{y}'\mathbf{H}\mathbf{y} = \mathbf{y}'\hat{\mathbf{y}}$, we have

$$\begin{aligned}
\mathbf{y}'\mathbf{D}_n\hat{\mathbf{y}} &= (\mathbf{y}-\hat{\mathbf{y}})'\mathbf{D}_n\hat{\mathbf{y}} + \hat{\mathbf{y}}'\mathbf{D}_n\hat{\mathbf{y}} \\
&= (\mathbf{y}-\hat{\mathbf{y}})'\hat{\mathbf{y}} + \hat{\mathbf{y}}'\mathbf{D}_n\hat{\mathbf{y}} \qquad (\text{as } \mathbf{D}_n(\mathbf{y}-\hat{\mathbf{y}}) = \mathbf{y}-\hat{\mathbf{y}}) \\
&= \hat{\mathbf{y}}'\mathbf{D}_n\hat{\mathbf{y}}
\end{aligned}$$

and $$r^2 = \frac{(\hat{\mathbf{y}}'\mathbf{D}_n\hat{\mathbf{y}})^2}{(\mathbf{y}'\mathbf{D}_n\mathbf{y})(\hat{\mathbf{y}}'\mathbf{D}_n\hat{\mathbf{y}})} = \frac{\hat{\mathbf{y}}'\mathbf{D}_n\hat{\mathbf{y}}}{\mathbf{y}'\mathbf{D}_n\mathbf{y}} = R^2$$

R^2 is surely a measure of the model's capability to fit the present data.

B. Adjusted R^2

$$\overline{R}^2 = R^2_{\text{adj}} = 1 - \frac{\boldsymbol{SS}_{\text{Res}} / (n-p)}{\boldsymbol{SS}_{\text{Total}} / (n-1)}$$

$$= 1 - \frac{s^2 (n-1)}{\boldsymbol{SS}_{\text{Total}}}$$

C. Error Variance, s^2

It is also called residual mean square.

4.2 Cross Validation for model selection and Determination of Model Performance

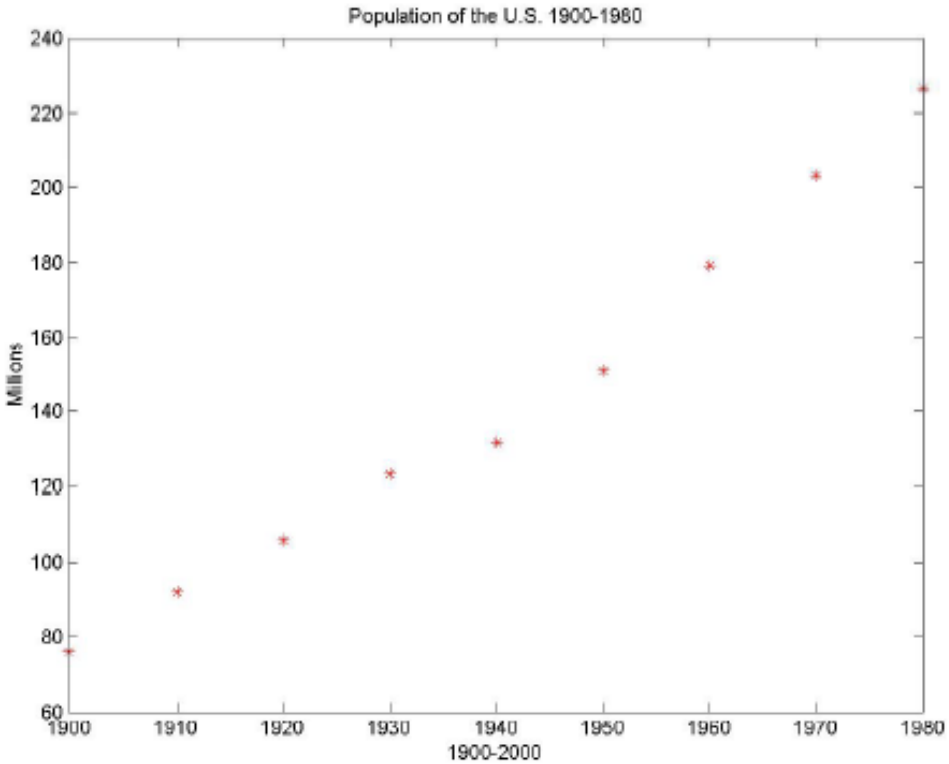
1. Fitting and Prediction

So far we consider only about fitting. The term "prediction" implies the process of estimating $E(y(\boldsymbol{X}_0))$ with $\hat{y}(\boldsymbol{X}_0)$ at $x = \boldsymbol{X}_0$. The ordinary residual $y_i - \hat{y}_i = e_i$ are not generally indicative of how the regression model will predict.

For example, the fitting is fine, but the prediction is poor.

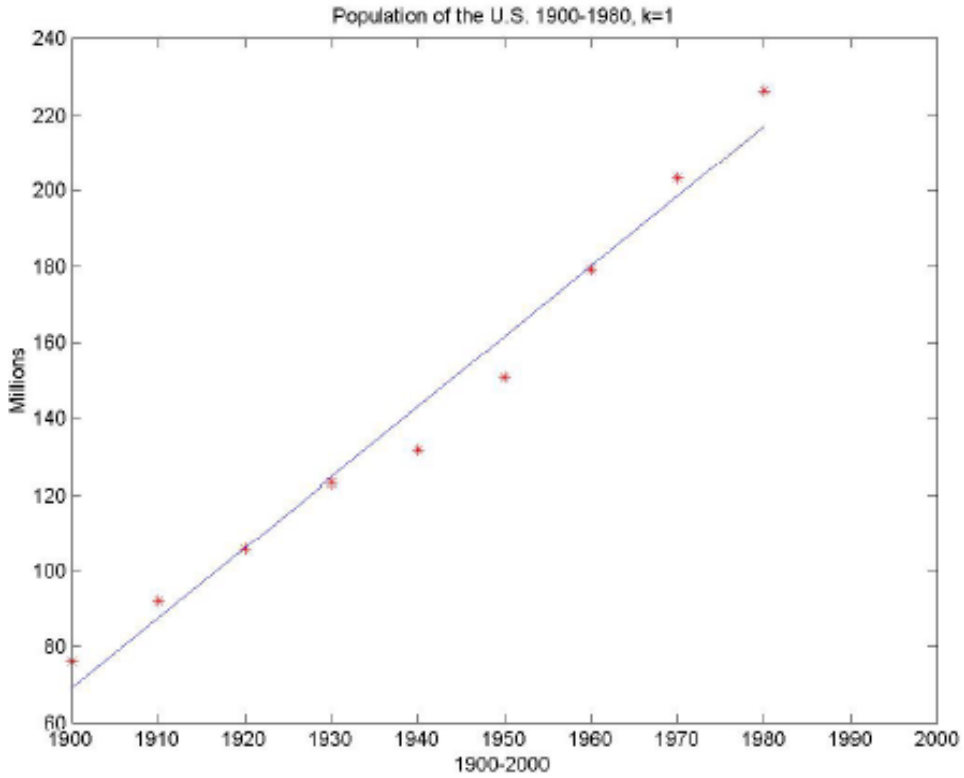
Data set

y	75.995	91.972	105.711	123.203	131.669	150.697	179.323	203.212	226.505
x	0	1	2	3	4	5	6	7	8



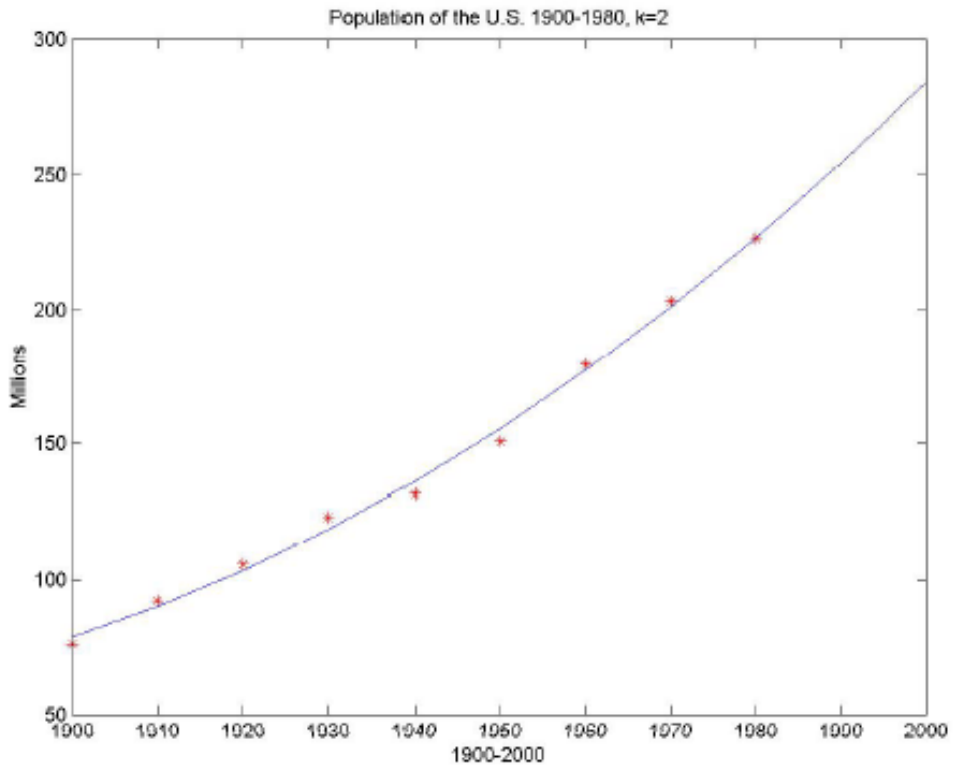
Model:

$$E(y) = \beta_0 + \beta_1 x$$



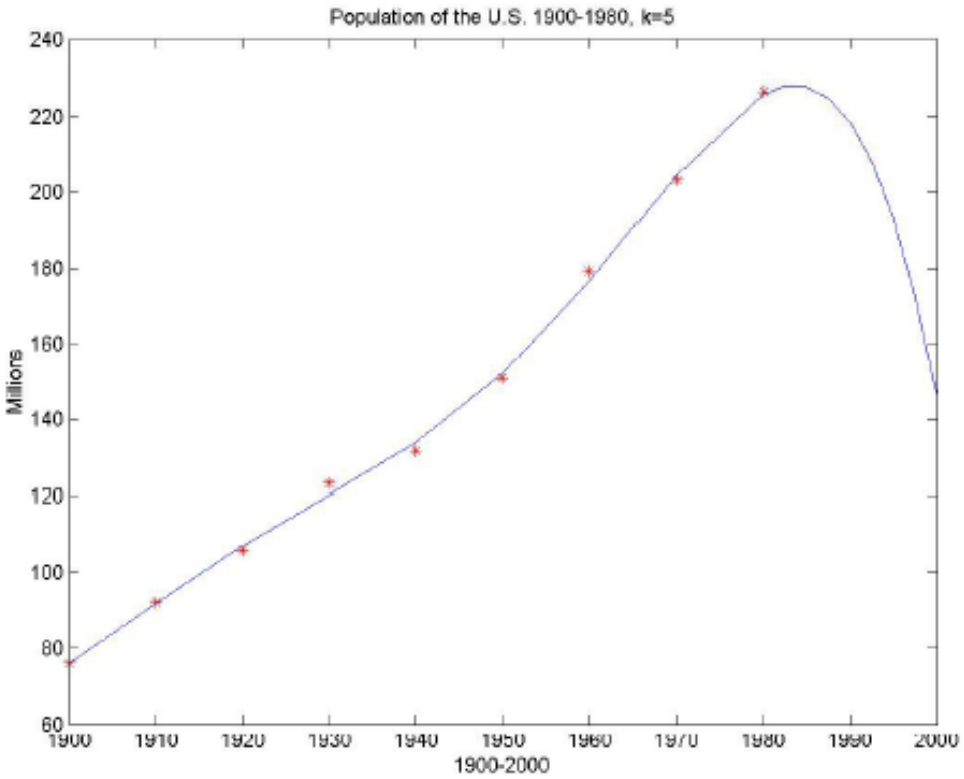
Model:

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$$



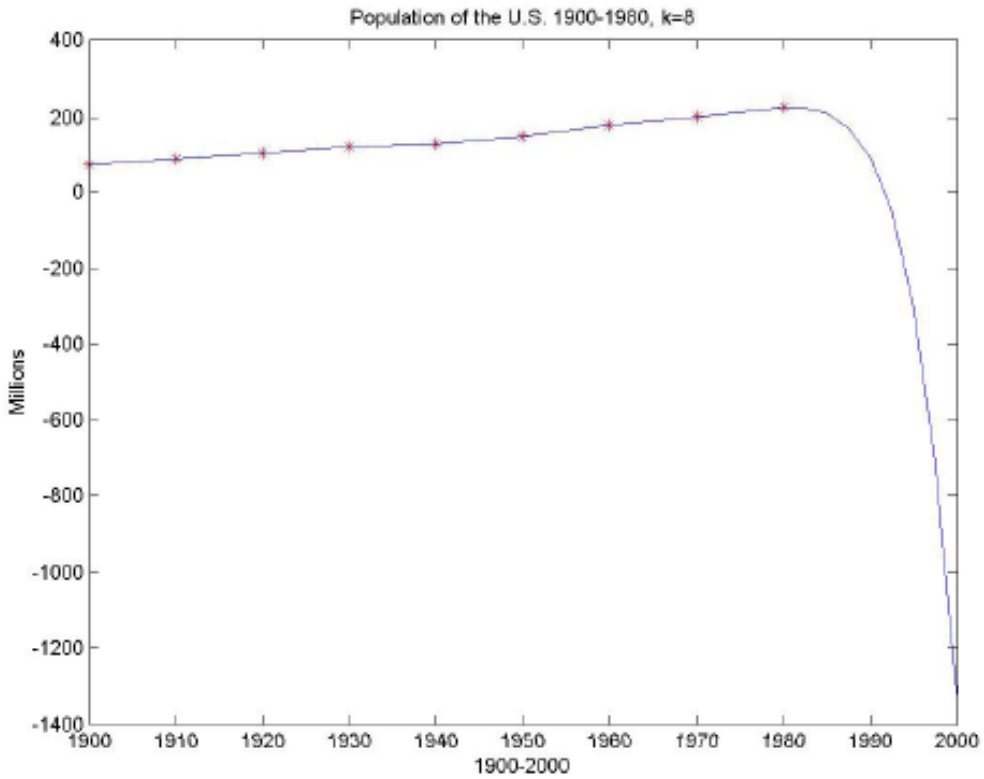
Model:

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5$$



Model:

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5 + \beta_6 x^6 + \beta_7 x^7 + \beta_8 x^8$$



2. Training sample & Validation sample

Split the data into two groups:

$$\left.\begin{array}{cccc} y_1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ y_m & x_{m1} & \cdots & x_{mk} \end{array}\right\} \text{Training sample (fitting sample)}$$

$$\left.\begin{array}{cccc} y_{m+1} & x_{m+1,1} & \cdots & x_{m+1,k} \\ \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & \cdots & x_{nk} \end{array}\right\} \text{Validation Sample}$$

Based on the training sample, we get this regression model:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

Calculate $\hat{y}_i$ for $i = m+1, \cdots, n$.

Note that $\hat{y}_i$ is independent of y_i, $i = m+1, \cdots, n$.

$$\sum_{j=m+1}^{n} (y_j - \hat{y}_j)^2 \quad \text{or} \quad \sum_{j=m+1}^{n} \left| y_j - \hat{y}_j \right|$$

are possible criteria to determine the best candidate model.

3. *PRESS* Statistics

We establish the model without the ith observation $\{y_i, x_{i1}, \ldots, x_{ik}\}$, and predict y_i by this model, denote it as:

$$\boxed{\hat{y}_{i,-i}}$$

Obviously, y_i and $\hat{y}_{i,-i}$ are independent.
The criteria:

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2 = \sum_{i=1}^{n} e_{i,-i}^2$$

can be used for model / variable selection.
A R^2-like statistics:

$$R_{\text{pred}}^2 = 1 - \frac{PRESS}{SS_{\text{Total}}}.$$

Reason: If y_i is really unusual, then the regression model based on all the observations may be overly influenced by this observations. This could produce a $\hat{y}_i$ that is very similar to y_i and consequently e_i will be small. So it will be difficult to detect any outlier.

If y_i is deleted, then $\hat{y}_{(i)}$ cannot be influenced by that observation, so the resulting residual should be likely to indicate the presence of the outlier.

Procedure

- Delete the ith observation,
- Fit the regression model to remaining $(n-1)$ observations,
- Calculate the predicted value of y_i corresponding to the deleted observation.
- The corresponding prediction error

$$e_{(i)} = y_i - y_{(i)}$$

- Calculate $e_{(i)}$ for each $i = 1, 2, \ldots, n$.

Computation of *PRESS*:

Denote the hat matrix $\boldsymbol{H} = (h_{ij}) = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$

$$e_{i,-i} = \frac{y_i - \hat{y}_i}{1 - x_i'(X'X)^{-1}x_i} = \frac{e_i}{1 - h_{ii}}$$

where

$\hat{y}_i$: based on the whole data set;

$x_i = (x_{i1}, \cdots, x_{ik})'$.

Therefore,

$$PRESS = \sum_{i=1}^{n} \left[\frac{e_i}{1 - h_{ii}} \right]^2.$$

Let $b_{(i)}$ be the vector of regression coefficients estimated by with holding the *ith* observations. Then

$$b_{(i)} = \left(X'_{(i)}X_{(i)}\right)^{-1} X'_{(i)}y_{(i)}$$

where $X_{(i)}$ is the *X*-matrix without the vector of *ith* observation and $y_{(i)}$ is the *y*-vector without the *ith* observation. Then

$$\begin{aligned} e_{(i)} &= y_i - \hat{y}_{(i)} \\ &= y_i - x_i\hat{b}_{(i)} \\ &= y_i - x_i(X'_{(i)}X_{(i)})^{-1}X'_{(i)}y_{(i)} \end{aligned}$$

Result: If $X'X$ is a $k \times k$ matrix and x be its ith row vector then $(X'X - x'x)$ denotes the $X'X$ – matrix with the ith row withheld. Then

$$\left[X'X - x'x\right]^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x'x(X'X)^{-1}}{1 - x(X'X)^{-1}x'}.$$

Using this result, we can write

$$\left[X'_{(i)}X_{(i)}\right]^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x'_i x_i (X'X)^{-1}}{1 - h_{ii}}$$

where $h_{ii} = x_i (X'X)^{-1} x'_i$.

Then

$$
\begin{aligned}
e_{(i)} &= y_i - x_i\left(X'_{(i)}X_{(i)}\right)^{-1} X'_{(i)}y_{(i)} \\
&= y_i - x_i\left[(X'X)^{-1} + \frac{(X'X)^{-1}x_i'x_i(X'X)^{-1}}{1-h_{ii}}\right]X'_{(i)}y_{(i)} \\
&= y_i - x_i(X'X)^{-1}X'_{(i)}y_{(i)} - \frac{x_i(X'X)^{-1}x_i'x_i(X'X)^{-1}X'_{(i)}y_{(i)}}{1-h_{ii}} \\
&= y_i - x_i(X'X)^{-1}X'_{(i)}y_{(i)} - \frac{h_{ii}x_i(X'X)^{-1}X'_{(i)}y_{(i)}}{1-h_{ii}} \\
&= \frac{(1-h_{ii})y_i - (1-h_{ii})x_i(X'X)^{-1}X'_{(i)}y_{(i)} - h_{ii}x_i(X'X)^{-1}X'_{(i)}y_{(i)}}{1-h_{ii}} \\
&= \frac{(1-h_{ii})y_i - x_i(X'X)^{-1}X'_{(i)}y_{(i)}}{1-h_{ii}}.
\end{aligned}
$$

Using $X'y = X'_{(i)}y_{(i)} + x_i' y_i$ (as x_i is $1 \times k$ vector), we can write

$$e_{(i)} = \frac{(1-h_{ii})y_i - x_i(X'X)^{-1}(X'y - x_i' y_i)}{1-h_{ii}}$$

$$= \frac{(1-h_{ii})y_i - x_i(X'X)^{-1}X'y + x_i(X'X)^{-1}x_i' y_i}{1-h_{ii}}$$

$$= \frac{(1-h_{ii})y_i - x_i b + h_{ii} y_i}{1-h_{ii}}$$

$$= \frac{y_i - x_i b}{1-h_{ii}}$$

$$= \frac{e_i}{1-h_{ii}}.$$

4.3 Conceptual Predictive Criteria (The *Cp Statistics)*

The current (full) model

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

where β_1 and β_2 contain p and $m - p$ parameters, respectively.

If we use model:

$$y = X_1\beta_1 + \varepsilon$$

for prediction.

How do we measure the performance?

$$C_p = p + \frac{(S_p^2 - \sigma^2)(n-p)}{\sigma^2}$$

σ^2 known;

$S_p^2 : \hat{\sigma}^2$ based on p parameters model.

n : # of observations.

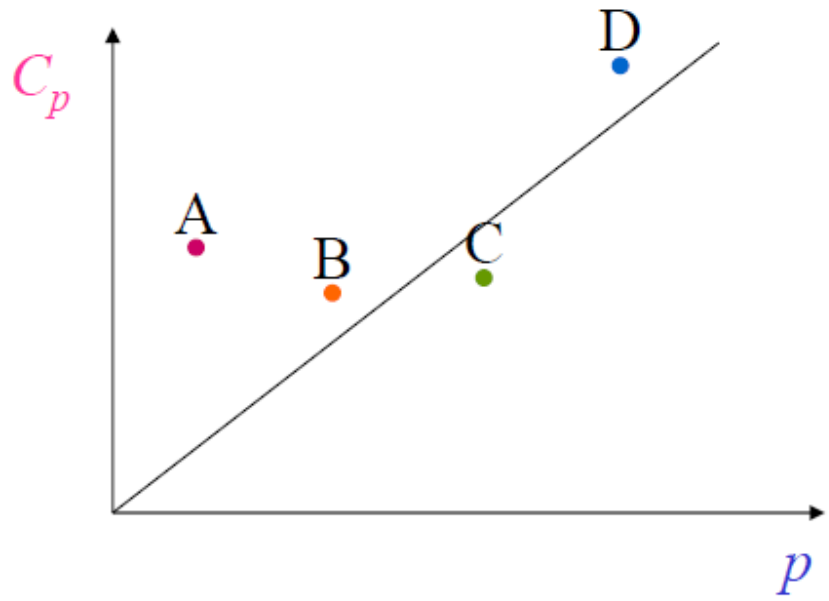
or $$C_p = p + \frac{(S_p^2 - s^2)(n-p)}{s^2}$$

σ^2 unknown;

$s^2 : \hat{\sigma}^2$ based on the full model.

One favors the candidate model with the smallest C_p value.

A plot C_p VS. p is useful for model selection:



Models **A** and **D** appear to be undesirable.

Models **D** is clearly the poorest performer.

Models **B** and **C** appear to be reasonable candidates.

A reasonable norm by which to judge the C_p-*value* is $C_p = p$.

✓ Sales data for asphalt shingles

✓A maker of asphalt roofing shingles in interested in the relationship between sales for a particular year and factors that obviously influence sales: *promotional accounts*, *number of active accounts*, *number of competing brands*, and *district potential for the sales districts*.

✓ Fifteen districts were used, and it is of interest to predict sales from the resulting regression. *Table 4.1* supplies the data.

Table 4.1

Sales data for asphalt shingles

District	x_1 (Promotional Accounts)	x_2 (Active Accounts)	x_3 (Competing Brands)	x_4 (Potential)	y (Sales in Thousands of Dollars)
1	5.5	31	10	8	79.3
2	2.5	55	8	6	200.1
3	8.0	67	12	9	163.2
4	3.0	50	7	16	200.1
5	3.0	38	8	15	146.0
6	2.9	71	12	17	177.7
7	8.0	30	12	8	30.9
8	9.0	56	5	10	291.9
9	4.0	42	8	4	160.0
10	6.5	73	5	16	339.4
11	5.5	60	11	7	159.6
12	5.0	44	12	12	86.3
13	6.0	50	6	6	237.5
14	5.0	39	10	4	107.2
15	3.5	55	10	4	155.0

John Neter, William Wasserman, and Michael Kutner. *Applied Linear Regression Models* (Homewood, Illinois: Richard D. Irwin, Inc., 1983), p. 439.

Three models are under serious considerations, based on what is known about the regressor variables. We plan to study the combinations (x_1, x_2, x_3, x_4), (x_1, x_2, x_3), and (x_2, x_3).

The following represents a table of pertinent statistics that can be used for comparing the models.

Model	R^2	R^2_{Pred}	SS_{Res}	S^2	PRESS	C_p
X_2, X_3	0.9940	0.9913	534.6621	44.5552	782.1896	11.40
X_1, X_2, X_3	0.9970	0.9928	272.7516	24.7956	643.3578	3.41
X_1, X_2, X_3, X_4	0.9971	0.9917	262.0728	26.2073	741.7557	5

4.4 Sequential Variable Selection Procedures

Suppose that there are k independent variables $x_1, \ldots, x_k$ in the problem. We want to find the "best" model.

1. Forward Selection

Begin with the simple regression model with that single predictor that has the largest correlation in absolute value with the response y. Next, we add to the model that predictor such that the new model with two predictors has the largest correlation with y. Thus, in forward procedure, we start with a subset of size 1, and, at each step, adding one variable at a time, until a stopping rule is met.

We add to the model that predictor with three equivalent criteria:

a) Adding the variable will increase R^2 more than any other single variable;

b) Adding the variable will reduce $\boldsymbol{SS}_{\text{Res}}$ more than other single variable;

c) The variable added would have the largest F-statistic of any of the variables that are not already in the model.

The stopping rules:

1) Stop with a subset of predetermined size p^*, say;

2) Stop if the F-test for each of the variables not get entered would be less than some predetermined number F_{in}.

2. Backward Elimination

The backward procedure is similar to the forward procedure, except that we start with the full model, and, at each step, remove one variable, until a stopping rule is met. The variable to be removed is the one that has the smallest contribution in the current model. The stopping rule for the backward procedure is similar to that for the forward procedure.

The stopping rules:

1) Stop with a subset of predetermined size p^*, say;

2) Stop if the F-test for each of the variables now in the model is bigger that some predetermined number F_{out}.

3. Stepwise Regression

The stepwise technique starts as in the forward procedure.

At each step after the first, we consider four alternatives:

Add a variable	Delete a variable	Exchange 2 variables	Stop

The rules are:

a) If there are at least two variables in the current model, and one or more has a value of F less than F_{out}, the variable with the smallest value of F is removed from the model.

b) If there are two or more variables in the model, the one with the smallest F-value is removed if its removal result in a value of R^2 that is larger than the R^2 obtained for the same number of variables previously. This can be happened in the process since variables are added and deleted at the various steps.

c) If two or more variables are in the model, one of them will be exchanged with a variable not in the model if the exchange increases R^2.

d) A variable is added to the model if it has the highest F-value as in the forward procedure, provided that F is greater than F_{in} and the tolerance criterion is satisfied.

4. MAXR Procedure

- MAXR begins in the same fashion as forward selection.
- No F-tests are conducted.
- An important feature of MAXR, distinguishing if from other methods, is that, at each stage, it allows for replacement of a model regressor, (exchanging regressor in the model and regressor out the model such that R^2 increases). This continues until no one-on-one replacement will increase R^2.
- The output of the MAXR procedure is, then k models, one for each model size.

The MAXR has several advantages over the other methods, there is still no assurance that:

The "best" model of a particular size is being found in the sense of R^2.

The best model of a particular size is being found in the sense of *PRESS*, C_p, or any other criteria.

5. Best Subset

- Consider & compare all 2^k subset models
- An algorithm to generate these 2^k models
- Criteria of comparing these 2^k models

Results for all possible regressions

Number of Variables in Model	$\sum_{i=1}^{n} \lvert e_{i,-i} \rvert$	R^2	C_p	PRESS	s^2	Variables in Model
1	61324.3	0.3348	785.26	455,789,688	21,940,359	x_5
1	22889.9	0.8843	125.87	70,075,727	3,816,879	x_4
1	28031.2	0.8934	114.97	158,243,072	3,517,337	x_2
1	13665.5	0.9715	21.20	22,243,776	939,990	x_1
1	13431.2	0.9722	20.38	21,841,500	917,487	x_3
2	25251.5	0.9104	96.50	71,164,912	3,165,704	x_4, x_5
2	25960.3	0.9239	80.30	128,703,914	2,688,542	x_2, x_5
2	28165.9	0.9306	72.29	140,743,693	2,452,812	x_2, x_4
2	13957.1	0.9725	21.99	22,564,832	971,399	x_1, x_3
2	17327.9	0.9741	20.04	35,786,830	914,181	x_1, x_4
2	16825.2	0.9754	18.57	32,454,824	870,761	x_3, x_4
2	11074.2	0.9840	8.16	12,977,100	564,291	x_1, x_5
2	10914.9	0.9848	7.29	12,628,470	538,720	x_3, x_5
2	12872.9	0.9861	5.66	18,036,834	490,585	x_1, x_2
2	12742.1	0.9867	4.94	17,853,441	469,456	x_2, x_3
3	26241.6	0.9523	48.28	107,102,405	1,816,626	x_2, x_4, x_5
3	18045.5	0.9785	16.80	34,400,018	818,290	x_1, x_3, x_4
3	12296.1	0.9846	9.41	17,828,243	583,845	x_1, x_4, x_5
3	11801.2	0.9850	9.00	16,275,120	570,794	x_3, x_4, x_5
3	11103.7	0.9850	8.97	13,036,635	569,830	x_1, x_3, x_5
3	16227.1	0.9861	7.66	32,814,300	528,257	x_1, x_2, x_4
3	15780.3	0.9868	6.90	30,139,304	504,218	x_2, x_3, x_4
3	14243.3	0.9873	6.21	22,794,229	482,286	x_1, x_2, x_3
3	12163.0	0.9894	3.71	18,051,730	403,215	x_1, x_2, x_5
3	12019.0	0.9901	2.92	17,846,717	377,954	x_2, x_3, x_5
4	12773.4	0.9851	10.92	18,621,398	615,741	x_1, x_3, x_4, x_5
4	17951.9	0.9879	7.54	37,719,883	499,469	x_1, x_2, x_3, x_4
4	13364.2	0.9905	4.34	22,464,723	389,793	x_1, x_2, x_3, x_5
4	14992.3	0.9906	4.26	30,255,902	387,001	x_1, x_2, x_4, x_5
4	14606.1	0.9908	4.03	28,629,419	378,826	x_2, x_3, x_4, x_5
5	16025.0	0.9908	6.00	32,195,222	412,277	x_1, x_2, x_3, x_4, x_5