All possible regressions and "best subset" regression

Two opposed criteria of selecting a model:

- **Including as many covariates as possible so that the fitted values are reliable.**
- **Including as few covariates so that the cost of obtaining information and monitoring is not a lot.**

Note:

There is not unique statistical procedure for selecting the best regression model.

Note:

Common sense, basic knowledge of the data being analyzed, and considerations related to invariance principle (shift and scale invariance) can not ever be set side.

Motivating example:

The "Hald" regression data

Y	X_1	X_2	X_3	X_4
78.5	7	26	6	60
74.3	1	29	15	52
104.3	11	56	8	20
87.6	11	31	8	47
95.9	7	52	6	33
109.2	11	55	9	22
102.7	3	71	17	6
72.5	1	31	22	44
93.1	2	54	18	22
115.9	21	47	4	26
83.8	1	40	23	34
113.3	11	66	9	12
109.4	10	68	8	12

$\Rightarrow$ Total 13 observations.

3 methods which can be used are:
(a) using the value of R^2 .
(b) using the value of s^2, the mean residual sum of square.
(c) using Mallow C_p statistic.

(a) R^2:

Example (continue):

In "Hald" data, there are 4 covariates, $X_1, X_2, X_3,$ and X_4. All possible models are divided into 5 sets:

Set A: $Y = \beta_0 + \varepsilon \Rightarrow \binom{4}{0} = 1$ possible model.

Set B: $Y = \beta_0 + \beta_i X_i + \varepsilon,\ i = 1,2,3,4 \Rightarrow \binom{4}{1} = 4$ possible models.

Set C: $Y = \beta_0 + \beta_i X_i + \beta_j X_j + \varepsilon,\ i \neq j,\ i,j = 1,2,3,4 \Rightarrow \binom{4}{2} = 6$ possible models.

Set D: $Y = \beta_0 + \beta_i X_i + \beta_j X_j + \beta_k X_k + \varepsilon,\ i \neq j \neq k,\ i,j,k = 1,2,3,4 \Rightarrow$

$\binom{4}{3} = 4$ possible models

Set E: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon, \Rightarrow \binom{4}{4} = 1$ possible model.

$\Rightarrow$ Total $2^4 = 16$ models.

For every set, **one or two models with large R^2 are picked**. They are the following:

Sets	Models	R^2
Set B	$Y = \beta_0 + \beta_2 X_2 + \varepsilon$	**0.666**
	$Y = \beta_0 + \beta_4 X_4 + \varepsilon$	**0.675**
Set C	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$	**0.979**
	$Y = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \varepsilon$	**0.972**
Set D	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \varepsilon$	**0.982**
	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$	**0.982**
Set E	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$	**0.982**

Principle based on R^2:

A model with **large** R^2 and **small number of covariates** should be a good choice since large R^2 implies the reliability of fitted values and a small number of covariates reduce the costs of obtaining information and monitoring.

Example (continue):

Based on the above principle, two models,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

and

$$Y = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \varepsilon$$

are sensible choices!!

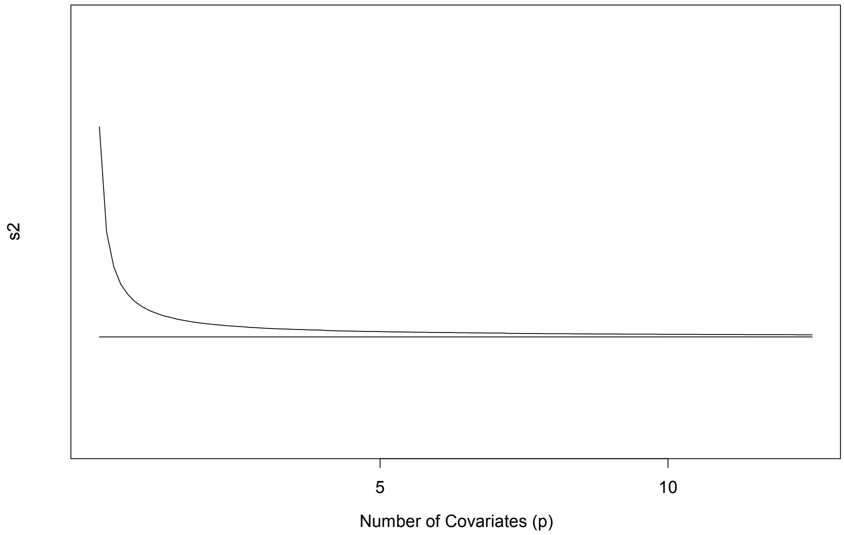
Note:

X_2 and X_4 are highly correlated. The correlation coefficient is -0.973. Therefore, it is not surprising that the two models have very close R^2.

(b) Mean residual sum of square s^2 :

A useful result:

As more and more covariates are added to an already overfitted model, the mean residual sum of square will tend to stabilize and approach the true value of σ^2 provided that all important variables have been included. That is,

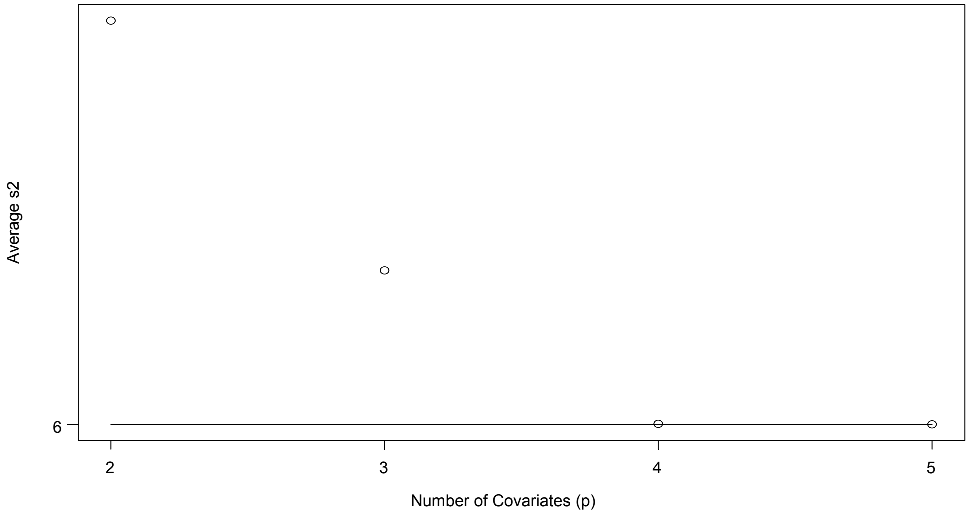


Example (continue):)

Again, we compute all s^2 for 16 possible models. We have the following table:

Sets	s^2	Average s^2
Set B	**115.06(X_1), 82.39(X_2), 176.31(X_3), 80.35(X_4)**	**113.53**
Set C	**5.79(X_1, X_2), 122.71(X_1, X_3), 7.48(X_1, X_4), 41.54(X_2, X_3) 86.89(X_2, X_4), 17.59(X_3, X_4)**	**47.00**
Set D	**5.35($X_1, X_2\ X_3$), 5.33(X_1, X_2, X_3),5.65(X_1, X_3, X_4) 8.20(X_2, X_3, X_4)**	**6.13**
Set E	**5.98($X_1, X_2\ X_3, X_4$)**	**5.98**

The plot of average s^2 against p (the number of covariates, including β_0) is



Principle based on s^2:

A model with mean sum of square s^2 **close to the estimate of σ^2 (the horizontal line) and with the fewest covariates might be a sensible model.**

Example:

The estimate of σ^2 could be 6. The model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

is sensible since its mean residual sum of square s^2 is 5.79 (close to 6) and the number of covariates are small compared with the other models with s^2 close to 6.

(c) Mallows C_p:

$$\text{Mallows } C_p = \frac{RSS(p)}{s^2} - (n - 2p),$$

where $\boldsymbol{n}$ is the sample size, $\boldsymbol{p}$ is the number of covariates including β_0, $RSS(p)$ is the residual sum of squares from a model containing $\boldsymbol{p}$ parameters, and s^2 **is the mean residual sum of squares from the model containing all possible covariates.**

Intuition of Mallows C_p:

Suppose s^2 is the mean residual sum of squares from the full model (containing all possible covariates)

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{r-1} X_{r-1} + \varepsilon,$$

and

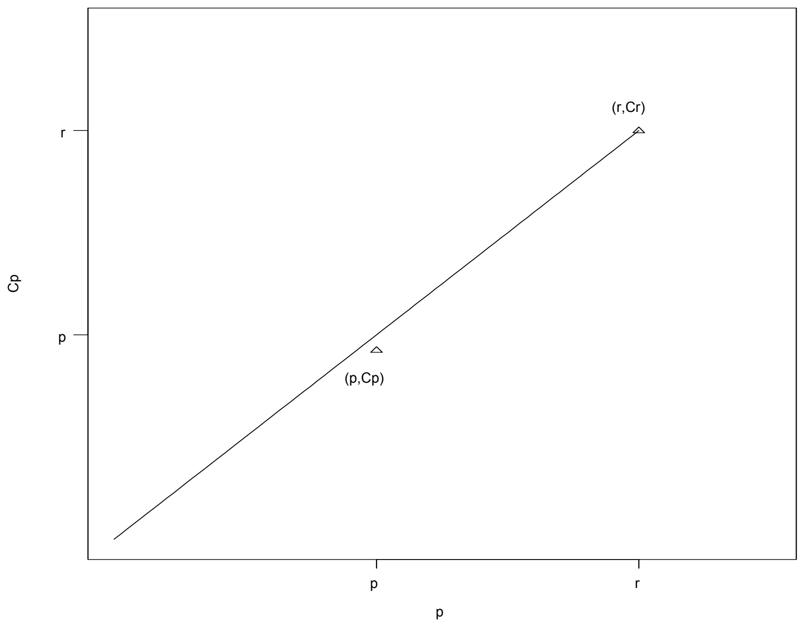
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon$$

is the true model, $p < r$. Then

$\frac{RSS(p)}{n-p}$, the mean residual sum of squares from model p, should be a sensible estimate σ^2 accurately. That is, $\frac{RSS(p)}{n-p} \approx \sigma^2$. Thus, $RSS(p) \approx (n-p)\sigma^2$. Also, the mean residual sum of squares s^2 for the overfitted model $s^2 \approx \sigma^2$.

$$\Rightarrow C_p = \frac{RSS(p)}{s^2} - (n-2p) \approx \frac{(n-p)\sigma^2}{\sigma^2} - (n-2p) = (n-p) - (n-2p) = p$$

Thus, (p, C_p) will falls close to the line of $Y = X$.



Note: $C_r = r$.

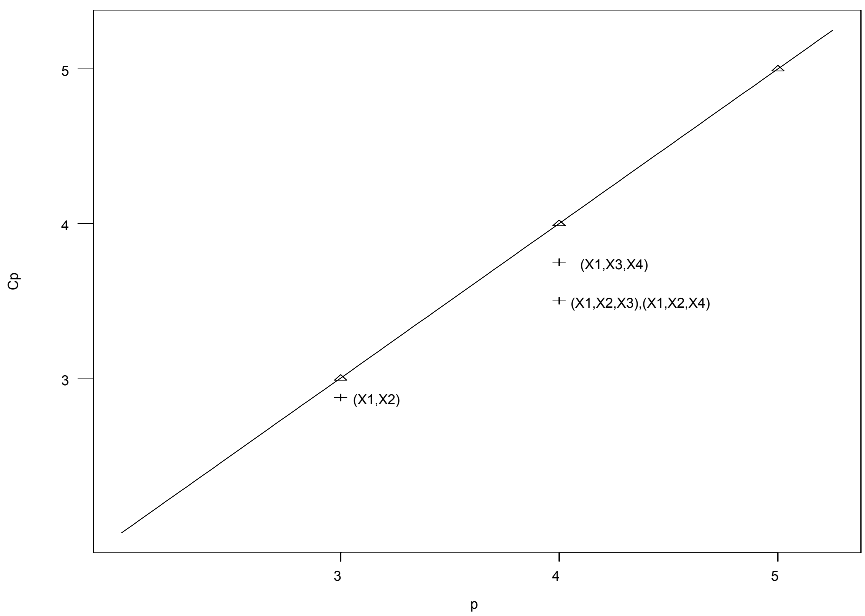
Principle based Mallows C_p:

The principle of selecting a best regression equation is to plot C_p versus p for every possible models. Then, choose some models with fewer covariates close to the line $Y = X$.

Example (continue):

For the motivating example, we calculate C_p for all 16 possible models. We then have the following table:

	C_p
Set A	443.2
Set B	202.5 (X_1), 142.5 (X_2), 315.2 (X_3), 138.7 (X_4)
Set C	2.7 (X_1, X_2), 198.1 (X_1, X_3), 5.5 (X_1, X_4), 62.4 (X_2, X_3), 138.2 (X_2, X_4), 22.4 (X_3, X_4)
Set D	3 (X_1, X_2, X_3), 3 (X_1, X_2, X_4), 3.5 (X_1, X_3, X_4), 7.3 (X_2, X_3, X_4)
Set E	5 (X_1, X_2, X_3, X_4)



The point (p, C_p) value for the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ is close to the line $Y = X$. and the model also has fewer parameters. Therefore, we recommend this model as a sensible choice.