Chapter 9. Variable Selection and Model Building

We want: 1) bias errors to be small for prediction purposes

— include as many $X's$ as possible.

2) variance of the prediction to be small

— i.e. $\sum_{i=1}^{n} Var\ \hat{y}_i/n = \frac{p}{n}\sigma^2$, small — p to be small.

Seek for a compromise between 1) and 2) – no unique procedure.

Usually iterative approach is employed including

(1) a particular variable selection criterion

(2) diagnostic check for the resulting model.

Motivation for Selecting Variables

Assume there are K candidate regressors, $x_1, \ldots, x_K$.

$$\text{Full model :} \quad y_i = \beta_0 + \sum_{j=1}^{K} \beta_j x_{ij} + \epsilon_i, \quad i = 1, \ldots, n,$$

or

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Let r be the number of regressors that are deleted from the full model. Rewrite the full model as

$$\mathbf{y} = X_p\boldsymbol{\beta}_p + X_r\boldsymbol{\beta}_r + \boldsymbol{\epsilon}, \quad X = (X_p, X_r), \boldsymbol{\beta} = (\boldsymbol{\beta}_p', \boldsymbol{\beta}_r')'$$

such that the subset model is

$$\mathbf{y} = X_p\boldsymbol{\beta}_p + \boldsymbol{\epsilon}.$$

Then the LSE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}^{\star} = (X'X)^{-1}X'\mathbf{y} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_p^{\star} \\ \hat{\boldsymbol{\beta}}_r^{\star} \end{pmatrix}$ and we use

$\hat{\sigma^2_{\star}}$ to denote the estimate of σ^2;

$\hat{y}_i^{\star}$ to denote the fitted values in full model.

Also, let $\hat{\boldsymbol{\beta}}_p = (X_p'X_p)^{-1}X_p'\mathbf{y}$ be the LSE of $\boldsymbol{\beta}_p$ in the subset model;

$\hat{\sigma^2}$ and $\hat{y}_i$ for estimate of σ^2 and fitted values.

Results: 1. $\hat{\boldsymbol{\beta}}_p$ is a *biased* estimate of $\boldsymbol{\beta}_p$ (unless $\boldsymbol{\beta}_r = 0$ or $X_p' X_r = 0$).

2. $Var(\hat{\boldsymbol{\beta}}_p^\star) - Var(\hat{\boldsymbol{\beta}}_p)$ is *p.s.d.* ("≥ 0").

   i.e. deleting variables *never* increases the variances of the estimates of the (remaining) parameters.

3. $Var(\hat{\boldsymbol{\beta}}_p^\star) - MSE(\hat{\boldsymbol{\beta}}_p)$ "≥ 0", if $Var(\hat{\boldsymbol{\beta}}_r^\star) - \boldsymbol{\beta}_r \boldsymbol{\beta}_r'$ "≥ 0". ($|\beta_r| < s.e.(\hat{\beta}_r^\star)$ in one-dimension)

   $\therefore \hat{\boldsymbol{\beta}}_p$ has smaller MSE if $Var(\hat{\boldsymbol{\beta}}_r^\star) - \boldsymbol{\beta}_r \boldsymbol{\beta}_r'$ "≥ 0".

Results: 4. $E(\hat{\sigma^2_\star}) = \sigma^2$ but $E(\hat{\sigma^2}) \geq \sigma^2$.

5. $Var(\hat{y}^\star) \geq MSE(\hat{y})$ if $Var(\hat{\boldsymbol{\beta}}^\star_r) - \boldsymbol{\beta}_r\boldsymbol{\beta}'_r$ " ≥ 0".

Conclusions: 1. Deleting variables may improve the precision, but potentially introduces bias.

2. Retaining negligible variables may increase the variances of the estimates of the parameters and the predicted response.

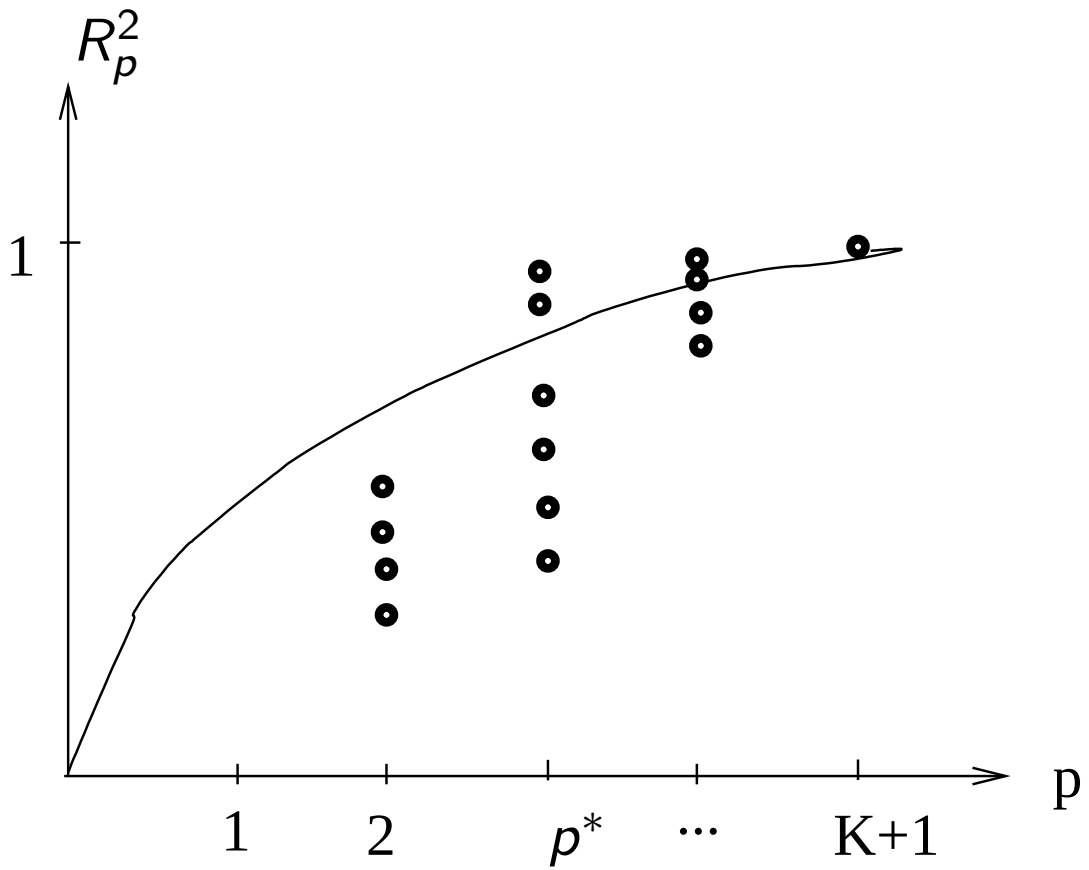
Criteria of Evaluating Subset Regression Models

1) R^2. The R^2 in a p-term subset model including β_0 is

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

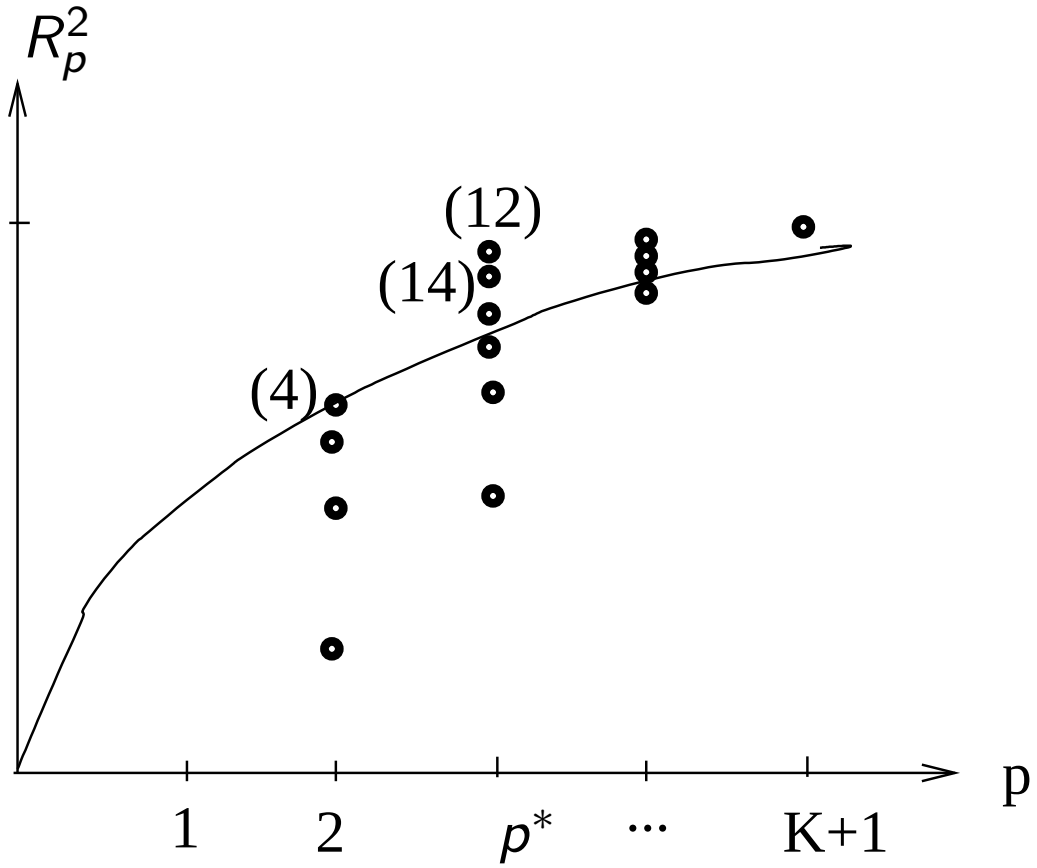
Note:

1. $R_p^2 \uparrow$ in p and attains maximum when $p = K + 1$.
2. There are $\begin{pmatrix} K \\ p-1 \end{pmatrix}$ subset models of size p.
3. Look at the point where an additional variable is not useful. i.e. it provides only a small increase in R_p^2.


R_p^2
1
p
1
2
p^*
...
K+1

Ex. The Hald Cement Data. $K = 4$.

p	I	II	III	IV
	.675(4)	.979 (12)*	.98234(124)*	.98237(1234)
	.666(2)	.972(14)*	.98228(123)*	
R_p^2	.534(1)	.935(34)	.98128(134)*	
	.286(3)	.847(23)	.97282(234)*	
		.680(24)		
		.548(13)		



(14) or (12) is the best, but (4) is the best when $p = 2$.

Read Page 271 in text.

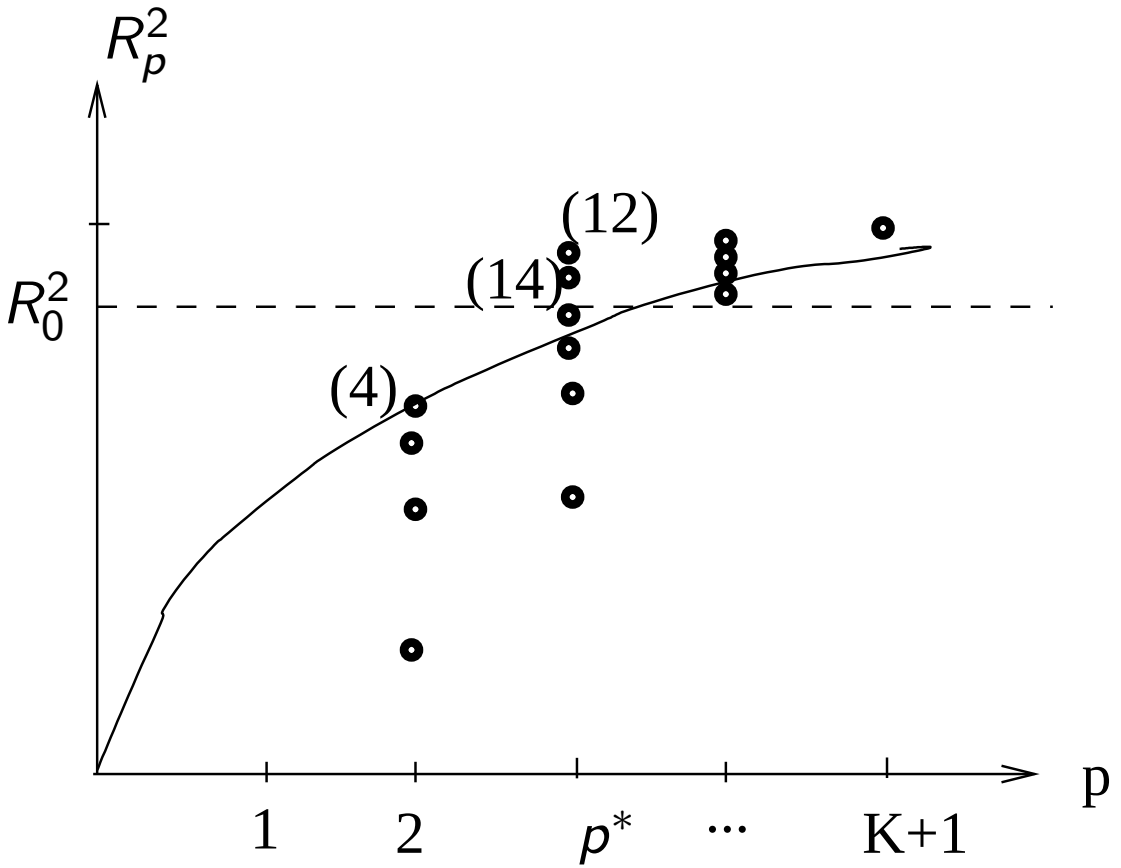
Note: 4. Large R_p^2 is preferred. Choose subsets of regressors producing an R^2 greater than

$R_0^2 = 1 - (1 - R_{K+1}^2)(1 + d_{\alpha,n,K})$, where

$d_{\alpha,n,K} = K\ F_{K,n-K-1;\alpha}/(n - K - 1)$, called $R^2_{adequate}(\alpha)$ subsets.

Ex. (cont'd)

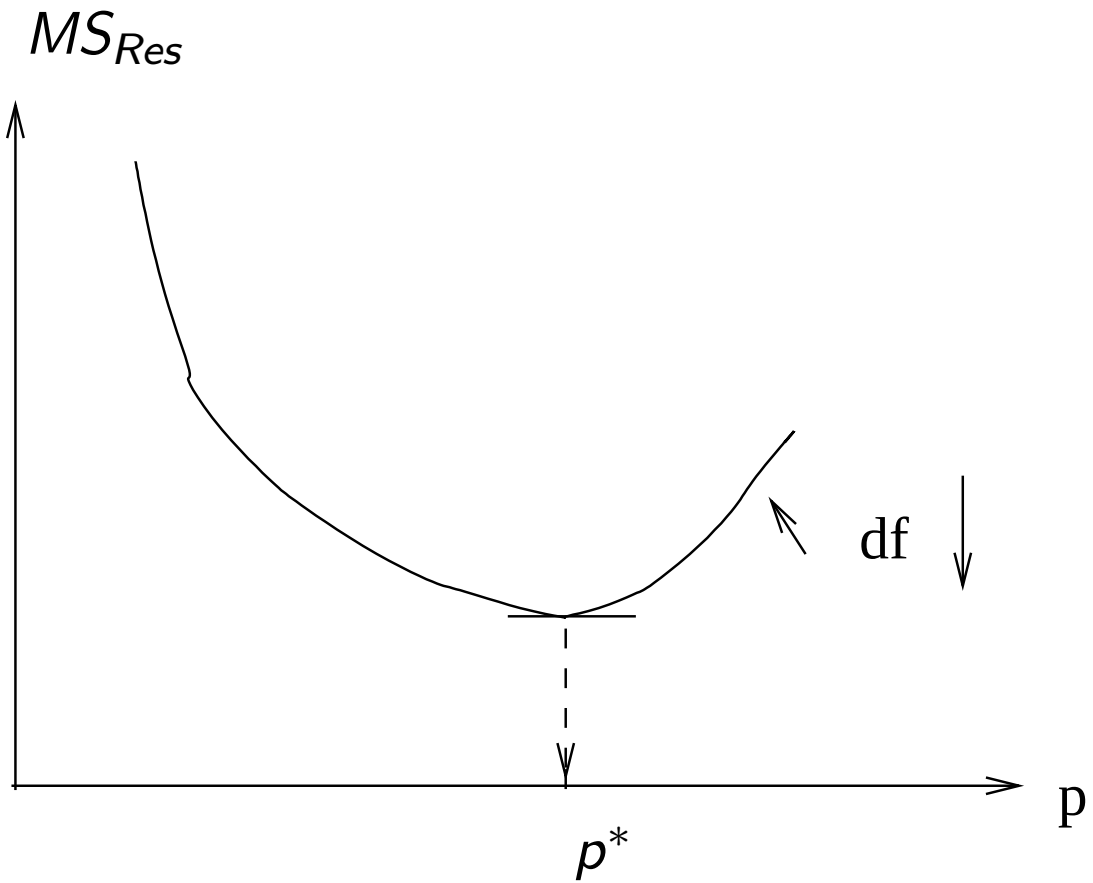
$$R_0^2 = 1 - (1 - R_5^2)(1 + \frac{5F_{4,8;.05}}{8}) = .94885.$$


R_p^2
R_0^2
(12)
(14)
(4)
p
1
2
p^*
···
K+1

□

2) MS_{Res}. $MS_{Res}(p) = \dfrac{SS_{Res}(p)}{n-p} \downarrow$ then $\uparrow$ as $p \uparrow$.

Ex.(Cont'd)



Note: 1. For each p, there are $\binom{K}{p-1}$ possible subset models.

1) Average over all possibilities for each p.

2) Find the smallest p that 'begins' to get close to

   $MS_{Res}(\text{full})$.

3) Choose the subset of size p that yields

   $MS_{Res} \approx MS_{Res}(\text{full})$.

Note: 2. $R^2_{Adj,p} = 1 - \frac{n-1}{n-p}(1 - R^2_p) = 1 - \frac{n-1}{n-p}(\frac{SS_{Res}(p)}{SS_T})$

$= 1 - \frac{n-1}{SS_T} MS_{Res}(p)$.

$\therefore$ min $MS_{Res}(p)$ and max $R^2_{Adj,p}$ are equivalent.

Ex.(Cont'd) Again (14) or (12). See P. 274. $\square$

3) Mallow's C_p statistics.

Recall that $E(SS_{Res}(p)) = (n-p)\sigma^2 + \sum_{i=1}^{n} B_i^2$.

Define

$$C_p = SS_{Res}(p)/\hat{\sigma^2} - (n-2p), \text{ where } \hat{\sigma^2} = MS_{Res}(\text{full}).$$

Note: 1. If the model is adequate (unbiased), $\sum_{i=1}^{n} B_i^2 = 0$

$$\Longrightarrow C_p \approx (n-p)\sigma^2/\sigma^2 - (n-2p) = p, \text{ ideally.}$$

2. Special case: $p = K + 1$, full model (unbiased).

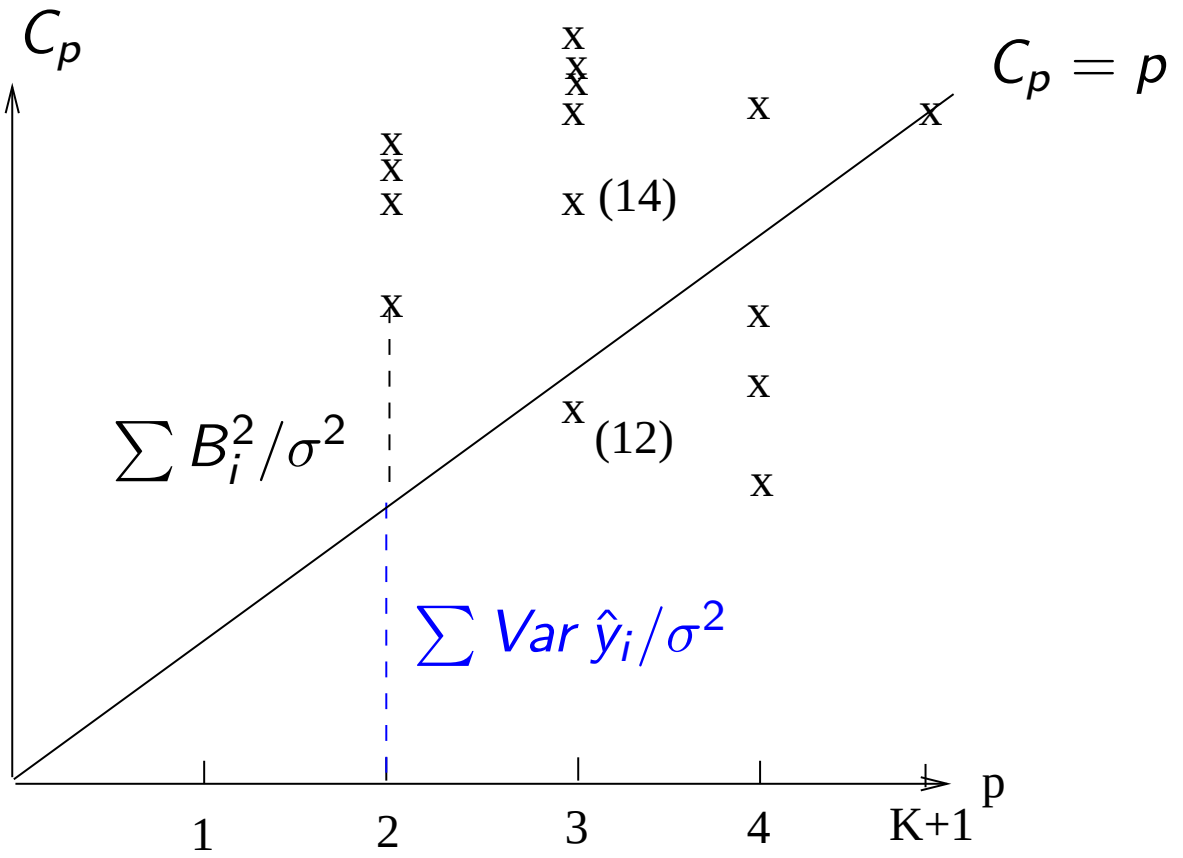
$$C_p \equiv (n - K - 1) - (n - 2(K + 1)) = K + 1 = p.$$

3. If the model is biased, $\sum_{i=1}^{n} B_i^2 > 0$.

$$
\begin{aligned}
C_p &\approx \frac{(n-p)\sigma^2 + \sum B_i^2}{\sigma^2} - (n-2p) = \frac{\text{bias}^2}{\sigma^2} + \frac{p\sigma^2}{\sigma^2} \\
&= (\text{bias}^2 + \sum_{i=1}^{n} Var\hat{y}_i)/\sigma^2 \\
&= MSE_p/\sigma^2.
\end{aligned}
$$

$\therefore$ we want to choose a model with $C_p \approx p$ (such that bias ≈ 0) and p as small as possible (such that predicted variance is small).

Ex.(Cont'd)



Theoretically $E(C_p) \geq p$, but due to random variation C_p may be $< p$.

□

4) $PRESS_p$.

Recall

$$PRESS_p = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2, \quad \text{in a } p\text{-term model.}$$

The model with smaller $PRESS_p$ values are preferrable.

Computational Techniques

- **Best Subset Regressions**.

  1) Specify M = number of best subsets considered.

  2) According to the above criteria, give a list of the selected best subsets for each p.

  Note: 1. Not all 2^K equations are considered, so may leave out sensible ones.

  2. Unless K is small, otherwise, it is impossible to make a detailed examination of *all* possible regressions.

Read pp. 270-277 in text.

II Stepwise-Type Procedures

(1) **Stepwise regression**:

i) Fit a simple linear regression for each of the K variables.

Test: slope=0. (i.e. $y = \beta_0 + \epsilon$ versus $y = \beta_0 + \beta_k x_k + \epsilon$.)

Compute $F_k = \frac{SS(\beta_k|\beta_0)}{SS_{Res}(\beta_k,\beta_0)/(n-2)}, \quad k = 1, 2, \ldots, K.$

The regressor with the largest F value is the *candidate* for first addition. Without loss of generality, assume $F_1 = \max_k F_k = F^\star$. If $F^\star > F_{1,n-2;\alpha} \equiv F_{IN}$, add into the model; otherwise, *stop* with no variable entering the model.

ii) Suppose now $y = \beta_0 + \beta_1 x_1 + \epsilon$.

Compute the partial-F statistics for each $k \neq 1$.

$$F_{k|1} = \frac{SS(\beta_k|\beta_1, \beta_1)}{MS_{Res}(\beta_k, \beta_1, \beta_0)} = \left(\frac{\hat{\beta}_k}{s.e.(\hat{\beta}_k)} \right)^2 = t_k^2, k \neq 1.$$

If $F_{2|1} = \max_k F_{k|\cdot}$, then test

$$H_0 : \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = 0 \quad (\text{ versus full: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon).$$

x_2 enters the model if the test is significant; otherwise, stop.

iii) Suppose now $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$.

Examine whether *any* of the other variables already in the model should be *dropped*.

Compute $F_{2|1}$ and $F_{1|2}$.

Let $F^\star = \min(F_{2|1}, F_{1|2})$. If $F^\star < F_{OUT}$, insignificant, then the corresponding variable is removed from the model.

iv) Examine the next candidate for addition, then examine *any* of the variables already in the model should be deleted.

Note: 1. The stepwise regression allows an x-variable brought into the model at an early stage to be dropped subsequently if it is no longer helpful in conjunction with variables added at the later stages.

2. The partial F test can be replaced by t-test ($\because F_{1,\nu} = t_\nu^2$).

3. The test values for the partial $F's$ are often called "F to enter" and "F to remove".

Ex. Hald Cement Data (Cont'd). $K = 4$, $n = 13$.

1) $F_{1|-} = 12.6, F_{2|-} = 21.96, F_{3|-} = 4.4, F^{\star}_{4|-} = 22.8$. First to enter? x_4? $y = \beta_0 + \beta_4 x_4 + \epsilon$?

   $F^{\star} = 22.8 > F_{1,,11;.05} = 4.84$, $\therefore$ x_4 enters.

2) Next to enter? $F_{\cdot|4}$=?

   $F^{\star}_{1|4} = 108.22, F_{2|4} = 0.17, F_{3|4} = 40.29$. x_1?

   (Is $y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \epsilon$ significant?)

   Test $H_0 = \begin{pmatrix} \beta_4 \\ \beta_1 \end{pmatrix} = 0$ in $y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \epsilon$.

   $F^{\star} = 176.63$, significant! $\therefore$ x_1 enters.

3) Next to exit? $y = \beta_0 + \beta_1 x_1 + \beta_4 x_4 + \epsilon$, x_1 or x_4?

$F_{4|1} = 159.30, F^{\star}_{1|4} = 108.22 > F_{1,10;.05} = 4.96$. Both are significant. $\therefore$ no exit.

4) Next to enter? $F_{\cdot|14}$=?

$F^{\star}_{2|14} = 5.03, F_{3|14} = 4.24$, x_2?

Test significance for $y = \beta_0 + \beta_1 x_1 + \beta_4 x_4 + \beta_2 x_2 + \epsilon$. Overall $F = 166.83$, significant! x_2 enters.

5) Next to exit? x_1, x_2 or x_4?

$F_{1|24} = 154.01, F_{2|14} = 5.03, F^{\star}_{4|12} = 1.86 < F_{1,9;.05} = 5.12,$

insignificant!

$\therefore$ x_4 should be removed from the model.

$\therefore \ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$

6) Next to enter? Only x_3? $F_{3|12} = 1.83 < 5.12$, insignificant!

No entrance!

7) Exit? x_1 or x_2?

$F^{\star}_{1|2} = 146.52, F_{2|1} = 208.58$, significant! No exit!

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

Overall $F = 229.5$, significant! Done!

$$\hat{y} = 52.58 + 1.468 x_1 + 0.662 x_2.$$

□

Note: x_4 entered the model in the earliest stage, but was out later!

(2) **Forward selection**. Only the most recent *entrance* is tested.

i) $F^{\star}_{1|-} = \max_k F_{k|-}$, $k = 1$, say. Significant! x_1 is added.

ii) Find $\max_{k \neq 1} F_{k|1}$, say $k = 2$. x_2 enters?

Test $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$. Overall $F(x_1, x_2)$ say.

Significant!

Test $H_0 : \beta_2 = 0$ in $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$.

$F_{2|1}$, significant. $\Longrightarrow$ x_2 enters. (Partial-F.)

iii) Check $\max_{k \neq 1,2} F_{k|12}$, say $k = 3$, x_3?

Overall significant, and $F_{3|12}$ significant, enters!

iv) $\vdots$

Until no entrance!

Ex.(Cont'd).

1) $F^{\star}_{4|-} = 22.8 = \max_k F_{k|-} > F_{1,11;.05} = 4.84$, significant!

   $y = \beta_0 + \beta_4 x_4 + \epsilon$.

2) $F^{\star}_{1|4} = 108.22$ is the largest. x_1?

   (i) Test $H_0 : (\beta_1, \beta_4) = (0, 0)$ vs. $H_1 : \beta_1 \neq 0$ or $\beta_4 \neq 0$,

   significant!

   (ii) Partial $F = F_{1|4}$, significant, $\therefore$ x_1 enters.

3) $F^{\star}_{2|14} = 5.03 > F_{3|14} = 4.24$, x_2 ?

   (i) Overall: significant! (ii) $F_{2|14} < F_{1,9;.05} = 5.12$,

   insignificant. No entrance!

   Stop! $y = \beta_0 + \beta_1 x_1 + \beta_4 x_4 + \epsilon$. □

Note: 1. Only the last possible entrance is tested.

2. Not recommended unless it is specially desired never to remove variables that were retained in the model.

(3) **Backward elimination**.

i) Include *all* variables in the model.

ii) Find $F^\star = \min_k F_{k|\text{all others}}$, say $k = K$. x_K removed?

If $F^\star < F_{OUT}$, insignificant, x_K is eliminated.

iii) Model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1} + \epsilon$, overall significant!

Go to ii) until no variable is deleted.

Ex.(Cont'd).

1) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$.

2) $F^{\star}_{1|234} = 4.34, F_{2|134} = 0.5, F^{\star}_{3|124} = .02, F_{4|123} = .04$. x_3?

   $.02 < F_{1,8;.05} = 5.32$, insignificant, $\therefore$ remove x_3.

3) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \epsilon$. Overall

   $F = 166.83 > F_{3,9;.05}$, significant!

4) $F_{1|24} = 154.01, F_{2|14} = 5.03, F^{\star}_{4|12} = 1.86 < F_{1,9;.05} = 5.12$,

   insignificant. x_4 removed.

5) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$. $F = 229.5$, significant!

6) $F_{1|2} = 146.52, F^{\star}_{2|1} = 208.58 > F_{1,10;.05}$, significant!

   No deletion!

   $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, done!

Note: Once a variable has been removed, it is gone *forever*. Thus, all alternative models using the eliminated variables are not considered for possible examination.

Significance levels.

1. If a large α is selected, more x-variables would be admitted (Large rejection region).
2. Most often, use the same α in testing for entrance and exit.
3. One may like to set the 'exit α' $>$ 'entry α' to 'protect' predictors already in the model.
4. One may always use $F_{IN} = F_{OUT} = 4$.

Note: 1. All the procedures discussed do not necessarily select the best model, but an acceptable one.

2. Run the stepwise regression procedure to determine the number of variables used, say q. Do *all* possible sets of size q, and choose the best one.

Ex. Two competitive models for $q = 2$. x_1, x_2 or x_1, x_4 ???

Need apriori consideration or the experimenter's judgement. □

Homework 9: (Page 300) 9.1, 9.2, 9.7, 9.8, 9.12, 9.13, 9.17, 9.18, 9.19, 9.20.

Due: Jan. 7, 2009.