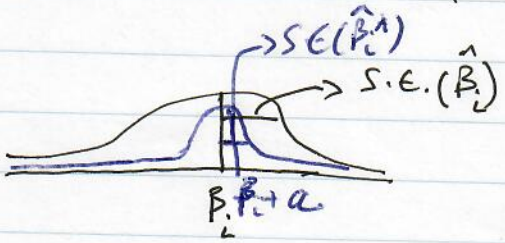
Notes

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

$$y = \beta_0^* + \beta_1^* X_1 + \beta_2^* X_2 + \cdots + \beta_K^* X_K + \varepsilon$$

$$S.E\,(\hat{\beta}_i^*) \leq S.E\,(\hat{\beta}_i)$$

$K < n$.

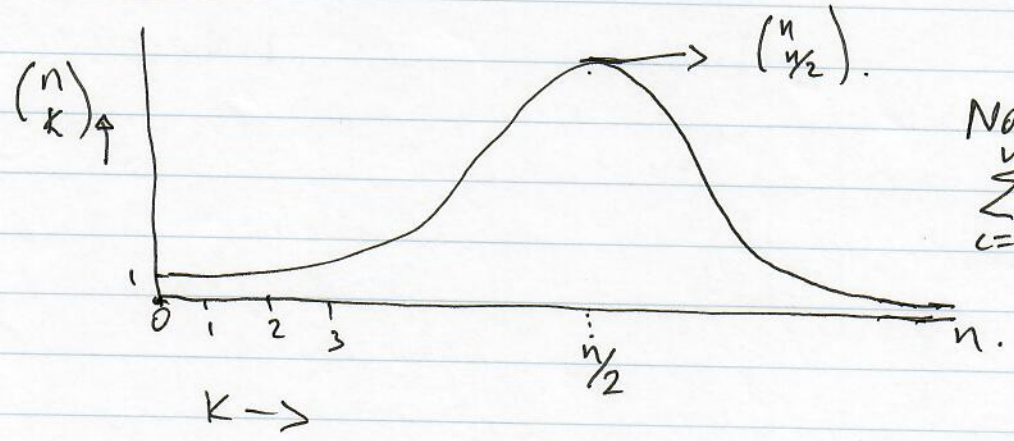


Finding all subset models with K variables it requires to assess $\binom{n}{K} = \frac{n!}{K!\,(n-K)!}$ subsets.

E.g.

$K=1$ $\binom{n}{K} = n$ and $K=n-1$ $\binom{n}{n-1} = n$.

$K=n$ $\binom{n}{n} = 1$. , $K=0$ $\binom{n}{0} = 1$. (note $0! = 1$).



Note:

$$\sum_{i=0}^{n} \binom{n}{i} = 2^n.$$

There are more combinations for $K = n/2$.

Comparing subsets with the same number of variables:

we can use R^2 or equivalently the RSS (Residual Sum of Squares).

C_p is used to compare subsets of different size.

FORWARD.

Consider that there are n variables $X_1, X_2, \ldots, X_n$.

First fit all single variable models, i.e.

$y = \beta_0 + \beta_1 X_1 + \varepsilon$ $\quad p^{(1)}$
$y = \beta_0 + \beta_1 X_2 + \varepsilon$ $\quad p^{(2)}$
$\vdots$ $\quad \vdots$
$y = \beta_0 + \beta_1 X_n + \varepsilon.$ $\quad p^{(n)}$

Find the p-value of β_1 in each model, say $p^{(i)}$ for subset $y = \beta_0 + \beta_1 X_i + \varepsilon$

Find the smallest p-value, say $p^{(2)}$, i.e $p^{(2)} = \min(p^{(1)}, p^{(2)}, \ldots p^{(n)})$.

Assume that we have $\alpha\%$ (α percentage point, eg 5%, 10%, 1%, etc).

If $p^{(2)} < \alpha\%$ then X_2 is significant and we select the variable X_2 to enter the model.
If $p^{(2)} \geq \alpha\%$ then X_2 is insignificant, we stop and select the previous model which is $y = \alpha_0 + \varepsilon$. (the empty model)

Assume that X_2 is significant. Then X_2 is retain for ever and we fit all remaining variables (one each time) with X_2. That is $n-1$ regressions (subsets) are evaluated:

$y = \beta_0 + \beta_1 X_2 + \beta_2 X_1 + \varepsilon$ $\quad p^{(1)}$
$y = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \varepsilon$ $\quad p^{(3)}$
$\vdots$ $\quad \vdots$
$y = \beta_0 + \beta_1 X_2 + \beta_2 X_n + \varepsilon$ $\quad p^{(n)}$

} P-values of β_2 in each of the $n-1$ regressions.

Let $p^{(1)} = \min(p^{(1)}, p^{(3)}, \ldots p^{(n)})$, i.e $p^{(1)}$ is the smallest p-value.

If $p^{(1)} \geq \alpha\%$ then X_1 is insignificant, we stop and select the previous model which is $y = \beta_0 + \beta_1 X_2 + \varepsilon$.

If $p^{(1)} < \alpha\%$, then X_1 is significant, thus, X_1 enters the model (for ever), and we consider all remaining variables (one at a time) in addition to (X_2, X_1). That is, we consider the $n-2$ regressions

$y = \beta_0 + \beta_1 X_2 + \beta_2 X_1 + \beta_3 X_3 + \varepsilon$ $\quad p^{(3)}$
$y = \beta_0 + \beta_1 X_2 + \beta_2 X_1 + \beta_3 X_4 + \varepsilon$ $\quad p^{(4)}$
$\vdots$ $\quad \vdots$
$y = \beta_0 + \beta_1 X_2 + \beta_2 X_1 + \beta_3 X_n + \varepsilon$ $\quad p^{(n)}$

} $n-2$ P-values

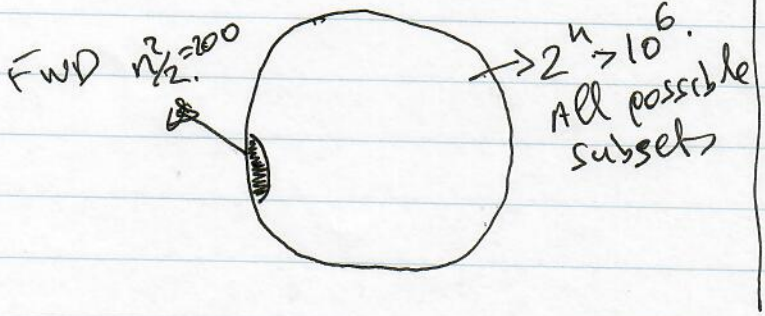
we repeat the same process until a p-value is found to be insignificant, or until all variables have been selected.

If all variables enter the model then the number of subsets (regressions) that have been evaluated are $n + (n-1) + (n-2) + \cdots + (1)$

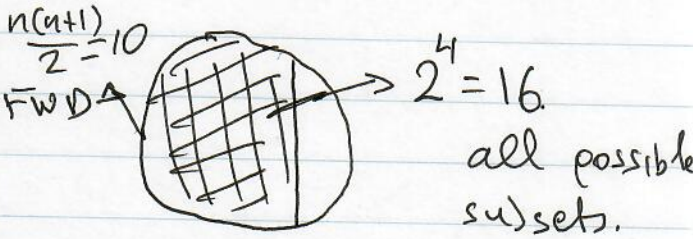
i.e $\sum_{i=1}^{n} i = \frac{n(n+1)}{2} \simeq n^2/2$

This implies that in the best case scenario the FORWARD method compares $n^2/2$ out of 2^n possible subsets.

E.g. if $n = 20$, $n^2/2 = 200$ and $2^n > 10^6$ (1 millia
Thus, the FORWARD checks 0.02% of all possible subsets



For $n = 4$



BACKWARD.

Fit all the variables, check the less significant variable and, (a) if the variable is significant then stop and select the model, or (b) if the variable is insignificant then delete it, and repeat the process. i.e.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon.$$

$p^{(1)}$ $p^{(2)}$ $p^{(n)}$ → p-values of each variable

Let $p^{(2)} = \max(p^{(1)}, p^{(2)}, \dots, p^{(n)})$.

If $p^{(2)} < \alpha\%$, then keep X_2 and stop with the full model above.

If $p^{(2)} \geq \alpha\%$, then delete the variable X_2 (from ever) and continue without it, i.e

fit the model

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \cdots + \beta_n X_n + \varepsilon.$$

$p^{(1)}$ $p^{(3)}$ ... $p^{(n)}$ $\rightarrow$ n-1 P values.

Repeat the process until you find an insignificant variable and STOP. Otherwise continue until all the variables are deleted.

If all variables are deleted then at most $n(n+1)/2 \approx n^2/2$ p-values have been assessed. The same as the FORWARD.

STEPWISE.

Combines both FORWARD AND BACKWARD so that when a variable enters the model can be deleted at a latter stage, or when a variable ~~enters~~ is deleted from the model, then it can enter later

First apply the Forward to enter a variable and then a backward to delete all insignificant variables. This is repeated until the Forward does not enter a variable.

Example. Stepwise

	variables in model	candidate variables
step 0:	$(\emptyset)$	$(X_1, X_2, \ldots, X_n)$

step 1: FORWARD

$y = \beta_0 + \beta_1 X_1 + \varepsilon$

$\boxed{y = \beta_0 + \beta_1 X_2 + \varepsilon}$

$\vdots$

$y = \beta_0 + \beta_1 X_n + \varepsilon$

Suppose X_2 is the most significant and enter the model

model list (X_2) candidate List $(X_1, X_3, \ldots X_n)$

The Backward to $y = \beta_0 + \beta_1 X_2 + \varepsilon$ would not result to any change.

Step 2:

FWD

$\boxed{y = \beta_0 + \beta_1 X_2 + \beta_2 X_1 + \varepsilon}$ X_1 has the smallest P-value and significant

$y = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \varepsilon$

$\vdots$

$y = \beta_0 + \beta_1 X_2 + \beta_2 X_n + \varepsilon$

BWD: $y = \beta_0 + \beta_1 X_2 + \beta_2 X_1 + \varepsilon$. (no this happens!)

Step 3:

FWD selects say X_n i.e $y = \beta_0 + \beta_1 X_2 + \beta_2 X_1 + \beta_3 X_n + \varepsilon$

BWD ~~can~~ may eliminate a variable (only X_1 or X_2 but actually only X_2 ....).

The stepwise in the best case considers ~~n^2~~ n^2 subsets (double Pass FWD and BWD).

This is still small when n is not small.