Partial F-test on Subsets of Coefficients

Suppose we wish to test the hypothesis

$H_0: \beta_2 = \beta_3 = 0 \qquad H_1: \beta_2$ and β_3 are ~~note~~ not both 0.

in a regression model with 3 explanatory variables

i.e. $y = \beta_0 + \beta_1 x_1 + \overset{\overset{0}{\shortparallel}}{\beta_2} x_2 + \overset{\overset{0}{\shortparallel}}{\beta_3} x_3 = 0$

We may then compute

$SS_R(\beta_1, \beta_2, \beta_3 | \beta_0) = SS_R$ for the model with 3 regression variables and $SS_R(\beta_1 | \beta_0) = R(\beta_1)$ for the model with only x_1 as regression variable

Let $R(\beta_2, \beta_3 | \beta_1) = SS_R - R(\beta_1)$

The F statistics is then $F = \dfrac{\frac{R(\beta_2, \beta_3 | \beta_1)}{2}}{\frac{SS_E}{n-4}} \sim F_{2, n-4}$

$SS_E \rightarrow$ full model

We reject H_0 when $F_{obs} \geq f_{\alpha, 2, n-4}$

In general: $H_0: \beta_{i+1} = \beta_{i+2} = \cdots = \beta_k = 0$, H_1: at least one of the coefficients $\neq 0$

$R(\beta_{i+1} \cdots \beta_k | \beta_1, \ldots, \beta_i) = SS_R(\beta_1, \ldots, \beta_k | \beta_0) - SS_R(\beta_1, \ldots, \beta_i | \beta_0)$

and we reject H_0 if

$$\frac{\frac{R(\beta_{i+1}, \ldots, \beta_k | \beta_1, \ldots, \beta_i)}{k-i}}{\frac{SS_E}{n-k-1}} \geq f_{\alpha, k-i, n-k-1}$$

This test can be used to test if one variable can be taken out and if the regression is significant.

In case we test on one variable for instance

$H_0: \beta_2 = 0 \qquad H_1: \beta_2 \neq 0$ in a model

with 3 regression variables we compute

$$R(\beta_2 \mid \beta_1, \beta_3) = SS_R - R(\beta_1, \beta_3)$$

The F statistics is

$$\frac{\frac{R(\beta_2 \mid \beta_1, \beta_3)}{1}}{\frac{SS_E}{n-4}} \sim F_{1, n-4}$$

this is equivalent to a t-test.

12.9 Sequential Methods for Model selection and Multicollinearity.

We have multicollinearity when two or more columns are strongly correlated (almost linear dependent). The degree of multicollinearity can be measured by the sample correlation coefficient

$$r_{ij} = \frac{\sum_{k=1}^{m} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\left(\sum_{k=1}^{m} (x_{ik} - \bar{x}_i)^2 \sum_{k=1}^{m} (x_{jk} - \bar{x}_j)^2 \right)^{\frac{1}{2}}}.$$

NB. Since the regression variables are not considered random, this expression does not estimate a true correlation.

Multicollinearity occurs often when the variation range of one or more explanatory variables are small.

Example If $x_{1i} \in (0.95, 1.05)$, x_1 will be strongly correlated by 1 and x_1^2.

The best measure of multicollinearity is given by the VIF-factor.

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the ~~multiple~~ multiple determination coefficient that is obtained by regressing x_j on all the other regression variables.

VIF should be below 10. If R_j^2 is greater than 0.99 MINITAB will remove variable x_j

What can be done if there exist strong multicollinearity

Remove column / columns

Collect more data

Respecify the model $(x_1, x_2, x_3) \longrightarrow \begin{cases} \dfrac{x_1 + x_2}{x_3} \\ x_1 \cdot x_2 \cdot x_3 \end{cases}$

Try to center variables: $(x_1 - \bar{x}_1)$. polynomial models

Center and scale variables. $\dfrac{(x_1 - \bar{x}_1)}{s_{x_1}}$

PCA regression

PLS regression

RIDGE regression

12.9 Sequential methods for variable selection

An alternative to best subset regression or to try all regression models with $0, 1, 2, \ldots, k$ variable

Forward selection

1. Start with β_0 in the model
2. Find $\max_j R(\beta_j) = \max_j SS_R(\beta_j | \beta_0) = \max_j \{ SS_R(\beta_0, \beta_j) - SS_R(\beta_0) \}$
3. If $\max_j \dfrac{R(\beta_j)}{\frac{SSE}{n-2}} < f_{\alpha, 1, n-2}$, stop no variables in the model.
4. If $\max_j \dfrac{R(\beta_j)}{\frac{SSE}{n-2}} \geq f_{\alpha, 1, n-2}$ occurs for $j = m$ add $\cancel{x_{jm}}$ x_m to the model

Find $\max_{i \neq m} R(\beta_i | \beta_m) = \max_{i \neq m} SS_R(\beta_i | \beta_0, \beta_m) = \max_{i \neq m} \{ SS_R(\beta_0, \beta_m, \beta_i) - SS_R(\beta_0, \beta_m) \}$

and continue the same way.

Backward ~~selection~~ elimination

Define $\underline{\beta} \setminus \beta_j = (\beta_0, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_k)$

1. Start with all the variables in the model

2. Find $\min_j R(\beta_j \mid \underline{\beta} \setminus \beta_j) = \min_j \{ SS_R(\beta_0, \dots, \beta_k) - SS_R(\beta_0, \dots, \beta_{j-1}, \beta_{j+1} \dots \beta_k) \}$

3. If $\min_j \dfrac{R(\beta_j \mid \underline{\beta} \setminus \beta_j)}{\frac{SS_E}{n-k-1}} \geq f_{\alpha, 1, n-k-1}$ stop, no variable is removed.

4. If $\min_j \dfrac{R(\beta_j \mid \underline{\beta} \setminus \beta_j)}{\frac{SSE}{n-k-1}} = \dfrac{R(\beta_m \mid \underline{\beta} \setminus \beta_m)}{\frac{SSE}{n-k-1}} < f_{\alpha, 1, n-k-1}$

remove X_m

Let $\underline{\beta} = \{\beta_1, \beta_2, \dots, \beta_k\} \setminus \beta_m$, $k = k-1$ and continue until

all the variables are significant.

Stepwise regression

1. Start like forward selection

Assume X_1 and X_2 are chosen to enter the model in step 1 and step 2. Let $\underline{\beta} = \{\beta_1, \beta_2\}$.

2. Find $\min_{j=1,2} R(\beta_j \mid \underline{\beta} \setminus \beta_j) = R(\beta_m \mid \underline{\beta} \setminus \beta_m)$

and investigate if X_m should be removed as for backward elimination.

3. Continue as for forward selection, but test in each step if any of the variable chosen to enter can be removed.