Test on significant impact of variables

In a multiple linear regression model we have

$$E[Y] = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

Thus β_i, $i = 1, 2, \dots, n$ is the change in $E[Y]$ if x_i is changed with one unit and all the rest of the variables are kept unchanged.

We want to test. Does a variable have a significant impact on the response given that the other variables are in the model.

Such a test for x_j, $j = 1, 2, \dots, k$ is

$H_0: \beta_j = 0$ against $H_1: \beta_j \neq 0$

We know $\hat{\underline{\beta}}$ is independent of $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-k-1}$ if

$\underline{\varepsilon} \sim N(\underline{0}, \sigma^2 \underline{I})$

Let c_{jj} be the $(j+1)$th diagonal element in $(\underline{X}^T \underline{X})^{-1}$

Then
$$T = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{c_{jj}}} = \frac{\dfrac{\hat{\beta}_j}{\sigma\sqrt{c_{jj}}}}{\dfrac{\hat{\sigma}}{\sigma}} = \frac{\dfrac{\hat{\beta}_j}{\sigma\sqrt{c_{jj}}}}{\sqrt{\dfrac{\hat{\sigma}^2(n-k-1)}{(n-k-1)\sigma^2}}} \sim \frac{N(0,1)}{\sqrt{\dfrac{\chi^2(n-k-1)}{n-k-1}}} \text{ if } H_0 \text{ is true}$$

and therefore t-distributed with $n-k-1$ degrees of freedom. We reject if $|T_{obs}| \geq t_{\frac{\alpha}{2}, n-k-1}$

For the test $H_0: \beta_j = \beta_{j0}$ against $H_1: \beta_j \gtreqless \beta_{j0}$, we use

$$T = \frac{\hat{\beta}_j - \beta_{j0}}{\hat{\sigma}\sqrt{c_{jj}}}$$ the same way except that

$\frac{\alpha}{2}$ is substituted by α if the test is one sided.

Choice of a fitted model

Explanatory variables with small or no influence on the response can give the model a bad prediction ability.

We have: $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

or $SS_T = SS_E + SS_R$

In a good model $y_i - \hat{y}_i$ should be small and $\hat{y}_i - \bar{y} \approx y_i - \bar{y}$, $i = 1, 2, \dots, n$. A measure of how much variation in the data that can be explained by the model is the cofficient of multiple determination.

$$R^2 = \frac{SS_R}{SS_T} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad 0 \le R^2 \le 1$$

$R^2 = 0{,}84$ tells us that 84% of the variation in the data can be explained by the model.

Note that $R^2 = \frac{SS_R}{SS_T} = \frac{SS_T - SS_E}{SS_T} = 1 - \frac{SS_E}{SS_T}$

A problem with R^2 is that it will always increase when we increase the number of explanatory variables.

To avoid that we have introduced $R^2_{adjusted}$, defined as

$$R^2_{adj} = 1 - \frac{\frac{SS_E}{n-k-1}}{\frac{SS_T}{n-1}} = 1 - \frac{(n-1)s^2}{SS_T}$$

Maximizing R^2_{adj} $\Leftrightarrow$ minimizing $\hat{\sigma}^2_{LS} = s^2$

<u>Other criteria</u>

Let $\hat{\sigma}^2 = \frac{SS_E}{n}$ i.e. the ML estimator

Smallest value of.

$$AIC \triangleq n \ln(\hat{\sigma}^2) + 2(k+2)$$

now σ^2 is also counted as parameter.

$$\hat{\sigma}^2 = \frac{SS_E}{n} = \frac{SS_E}{n-k-1}\frac{(n-k-1)}{n} = s^2\frac{(n-k-1)}{n}$$

Therefore $AIC = n\left(\ln s^2 - \ln\left(\frac{n}{n-k-1}\right) + \frac{2(k+2)}{n}\right)$

Smallest value of.

$$BIC \triangleq n\ln(\hat{\sigma}^2) + \ln(n)(k+2)$$

full model

$$\text{Mallows' } C_p \triangleq \frac{SS_E^k}{s^2} - (n - 2(k+1))$$

Note $E[C_p] \approx n-(k+1) - n + 2(k+1) = k+1$

if the model is correct.

Variable selection methods

These algorithms are normally performed as partial F-tests, adding and removing one factor at the time

Forward selection

1. Start with only β_0 in the model
2. Find $\max_j R(\beta_j) = \max_j \{SS_R(\beta_0, \beta_j) - SS_R(\beta_0)\}$
3. If $\max_j \dfrac{R(\beta_j)}{\frac{SS_E}{n-2}} = \dfrac{R(\beta_m)}{\frac{SS_E}{n-2}} < f_{\alpha,1,n-2}$ stop, no variable

   is entered into the model
4. If $\dfrac{R(\beta_m)}{\frac{SS_E}{n-2}} \geq f_{\alpha,1,n-2}$, add X_m to the model

Find $\max_{j \neq m} R(\beta_j|\beta_m) = \max_{j \neq m} \{SS_R(\beta_0, \beta_m, \beta_j) - SS_R(\beta_0, \beta_m)\}$ and

go to step 3. If $\max_{j \neq m} \dfrac{R(\beta_j|\beta_m)}{\frac{SS_E}{n-3}} < f_{\alpha,1,n-3}$ no variable

is entered. Otherwise proceed in the same way.

Note that the degrees of freedom in the partial F-test is reduced by one for each variable that is entered.

Backward elimination

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

1. Start with all variables in the model

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

3. If $\min_j \dfrac{R(\beta_j \mid \underline{\beta} \setminus \beta_j)}{\frac{SS_E}{n-k-1}} = \dfrac{R(\beta_m \mid \underline{\beta} \setminus \beta_m)}{\frac{SS_E}{n-k-1}} \geq f_{\alpha, 1, n-k-1}$ stop.

No variable is removed from the model

4 If $\dfrac{R(\beta_m \mid \underline{\beta} \setminus \beta_m)}{\frac{SS_E}{n-k-1}} < f_{\alpha, 1, n-k-1}$ remove X_m

Find $\min_{j \neq m} R(\beta_j \mid \underline{\beta} \setminus \{\beta_m, \beta_j\}) = \min_{j \neq m} \{ SS_R(\beta_0, \dots, \beta_{m-1}, \beta_{m+1}, \dots, \beta_k) - SS_R(\underline{\beta} \setminus \{\beta_m, \beta_j\}) \}$

If. $\min_{j \neq m} \dfrac{R(\beta_j \mid \underline{\beta} \setminus \{\beta_m, \beta_j\})}{\frac{SS_E}{n-k}} \geq f_{\alpha, 1, n-k}$ stop.

Otherwise proceed in the same way