Inference about regression coefficients

Want to infer something about the parameters \( \beta_1, \beta_2, \ldots, \beta_m \). We’ll learn a method of hypothesis testing called the \textit{reduction method}. This method allows one to test the hypothesis that a specified subset of independent variables is not needed in the model.

Let \( B_1 \) denote the subset of regression coefficients not involved in inference, and \( B_2 \) the coefficients about which inference is desired.

For example, suppose we want to test a hypothesis about just \( \beta_1 \). Then \( B_2 = \beta_1 \) and \( B_1 = \{\beta_2, \beta_3, \ldots, \beta_m\} \).
The reduction method allows us to test the hypotheses

\[ H_0 : \text{all coefficients in } B_2 \text{ are 0} \]

\[ H_1 : \text{at least one coefficient in } B_2 \text{ is not 0}. \]

Now, let \( SSE \) be the error sum of squares for the model that has all \( m \) independent variables in it.

Let \( SSE(B_1) \) be the error sum of squares when only the independent variables associated with \( B_1 \) are in the model.

Define

\[ SSR(B_2|B_1) = SSE(B_1) - SSE, \]

which is called the reduction in error sum of squares due to adding the \( B_2 \) variables to a model that has the \( B_1 \) variables in it.
Remarks

• The reduction sum of squares also satisfies

\[ SSR(B_2 | B_1) = SSR - SSR(B_1). \]

• It must be true that \( SSR(B_2 | B_1) \geq 0 \). Why?

Let \( m_1 \) and \( m_2 \) be the numbers of parameters in \( B_1 \) and \( B_2 \), respectively. \( (m = m_1 + m_2) \)

Now define the \( F \)-statistic

\[ F = \frac{SSR(B_2 | B_1) / m_2}{MSE}. \]

At level of significance \( \alpha \), the null hypothesis is rejected if \( F \geq F_{m_2, n-m-1, \alpha} \).
### ANOVA Table for Reduction Method

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>SS</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_1$</td>
<td>$SSR(B_1)$</td>
<td>$m_1$</td>
</tr>
<tr>
<td>$B_2$</td>
<td>$SSR(B_2</td>
<td>B_1)$</td>
</tr>
<tr>
<td>Error</td>
<td>$SSE$</td>
<td>$n - m - 1$</td>
</tr>
<tr>
<td>Total</td>
<td>$SST$</td>
<td>$n - 1$</td>
</tr>
</tbody>
</table>

*(table continued)*

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>MS</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_1$</td>
<td>$SSR(B_1)/m_1$</td>
<td></td>
</tr>
<tr>
<td>$B_2$</td>
<td>$SSR(B_2</td>
<td>B_1)/m_2$</td>
</tr>
<tr>
<td>Error</td>
<td>$MSE$</td>
<td></td>
</tr>
</tbody>
</table>
Examples of using the reduction method

Testing whether or not any variable is useful

\[ H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0 \]

In this case \( B_1 \) is empty,

\[ SSR(B_2|B_1) = SSR, \]

\( m_1 = 0 \) and \( m_2 = m \). The \( F \)-statistic is the one found in the ANOVA table of the SPSS output.

For the L.A. Heart Study data, the \( F \)-statistic is 11.94. If we do the test at level of significance .05, the null hypothesis

\[ H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \]

is rejected if \( F \geq F_{3,22,.05} = 3.049 \).

Since 11.94 > 3.049, \( H_0 \) is rejected and we may conclude that at least one of the variables age, weight and cholesterol level have an impact on systolic blood pressure. Note: From SPSS output, \( P \)-value < .0001.
Testing whether a specified variable is useful

\[ H_0 : \beta_j = 0 \]

Here \( B_2 \) contains only \( \beta_j \) and \( B_1 \) contains the other \( m-1 \) regression coefficients, and so \( m_1 = m - 1 \) and \( m_2 = 1 \).

The necessary sums of squares may be obtained by fitting two models: (i) the “full” model containing all \( m \) variables, and (ii) a model containing all the variables except the one corresponding to \( \beta_j \).

For the L.A. Heart Study data, suppose we want to test

\[ H_0 : \beta_2 = 0, \]

which is equivalent to saying that weight is not needed in the model with age and cholesterol.
Running the model with only age and cholesterol level yields

$$SSR(B_1) = 2419.76042,$$

and so

$$SSR(B_2|B_1) = 3011.99455 - 2419.76042 = 592.23413.$$ 

We have

$$F = \frac{592.23413}{84.08591} = 7.04.$$ 

Since this is larger than $F_{1,22,.05} = 4.30$, we may reject $H_0$ at the .05 level of significance. Apparently, weight should be in the model.
It’s always true that the $F$-statistic for testing $H_0 : \beta_j = 0$ is such that

$$F = t^2,$$

where $t$ is the $t$-statistic from the SPSS output. (Check this out for the example on the previous page.)

So, we needn’t use the reduction method to test $H_0 : \beta_j = 0$. We can just use the $t$-statistic from the standard SPSS output.

**Testing a specific subset of variables**

In our heart study example, suppose we want to test

$$H_0 : \beta_2 = \beta_3 = 0.$$ 

We may use the reduction method with $B_2$ containing $\beta_2$ and $\beta_3$, and $B_1$ containing just $\beta_1$, in which case $m_1 = 1$ and $m_2 = 2$. 
To get $SSR(B_1)$, fit the model with only the age variable in it. This yields

$$SSR(B_1) = 2413.13112,$$

and

$$SSR(B_2|B_1) = 3011.99455 - 2413.13112 = 598.8634.$$ 

The $F$-statistic is

$$F = \frac{598.8634/2}{84.08591} = 3.56.$$

At level .05, we reject $H_0$ if $F \geq F_{2,22,.05} = 3.44$. The $F$-statistic is just larger than 3.44, so we may conclude that at least one of weight and cholesterol is needed in the model.