How Do We Test Multiple Regression Coefficients?

Suppose you have constructed a multiple linear regression model and you have a specific hypothesis to test which involves more than one regression coefficient. How do we perform a hypothesis test that involves more than one regression coefficient?

First, in a multiple linear regression setting, you can perform either the likelihood ratio test (discussed in topic 2 lecture notes) or the analysis of deviance test.

Recall that you wish to determine if a set of “s” explanatory variables improve the fit of the model. Specifically, you have two models, called the null and extended of the form:

Null model:

\[ E(Y_i) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \]

Extended model:

\[ E(Y_i) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \beta_{p+1} X_{p+1} + \ldots + \beta_{p+s} X_{p+s} \]

You wish to test the following hypothesis:

\[ H_0: \beta_{p+1} = \beta_{p+2} = \ldots = \beta_{p+s} = 0 \]

This test can be performed using the deviance from the regression model. You need to obtain the SS(Error) from the null and extended model to perform the test.

\[
F = \frac{(SS(Error)_N - SS(Error)_E)/s}{SS(Error)_E/(n-p-s-1)}
\]

Under the null hypothesis, this F-statistic will follow an F distribution with s and n-p-s-1 degrees of freedom.

Now lets look at an example: You would like to determine the association between total medical expenditures and smoking status (never/current/former) after adjusting for age and gender.

Your variables are:

\[ \text{Logexp} = \log(\text{TOTALEXP}+100) \]

\[ \text{Smoke} = 0 \text{ if never, 1 if current, 2 if former} \]
Age = 40 – 94 (most plausible range of age for the disease)

Male = 1 if male, 0 if female

Our regression model is:

$$E[\log \text{exp}] = \beta_0 + \beta_1 S_1 + \beta_2 S_2 + \beta_3 \text{Age} + \beta_4 \text{Male} + \epsilon$$

Where

$$S_1 = \begin{cases} 1 & \text{if current} \\ 0 & \text{if never} \end{cases}$$

$$S_2 = \begin{cases} 1 & \text{if former} \\ 0 & \text{if never} \end{cases}$$

Therefore, you can write out a regression model for the never, current and former smokers.

Never smokers:

$$E[\log \text{exp}] = \beta_0 + \beta_3 \text{Age} + \beta_4 \text{Male} + \epsilon$$

Current smokers:

$$E[\log \text{exp}] = \beta_0 + \beta_1 + \beta_3 \text{Age} + \beta_4 \text{Male} + \epsilon$$

Former smokers:

$$E[\log \text{exp}] = \beta_0 + \beta_2 + \beta_3 \text{Age} + \beta_4 \text{Male} + \epsilon$$

So,

$$\beta_1 = \text{difference in the mean log total expenditures comparing current smokers to never smokers of the same age and gender.}$$

$$\beta_2 = \text{difference in the mean log total expenditures comparing former smokers to never smokers of the same age and gender.}$$

The test of interest is to determine if smoking is associated with total medical expenditures. To do this, we will compare the null model (includes age and gender) to the extended model (including dummy variables for smoking status and age and gender).

$$H_0: \beta_1 = \beta_2 = 0$$

Fit the null and extended model and perform the analysis of deviance.

(results below are based on a sample of the 1987 National Medical Expenditure Survey)
Null Model:

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs =     1365</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>167.057212</td>
<td>2</td>
<td>83.528606</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>2894.6933</td>
<td>1362</td>
<td>2.12532547</td>
<td>Adj R-squared = 0.0532</td>
</tr>
<tr>
<td>Total</td>
<td>3061.75051</td>
<td>1364</td>
<td>2.24468512</td>
<td>Root MSE = 1.4578</td>
</tr>
</tbody>
</table>

Extended Model:

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs =     1365</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>190.609655</td>
<td>4</td>
<td>47.6524137</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>2871.14085</td>
<td>1360</td>
<td>2.11113298</td>
<td>Adj R-squared = 0.0595</td>
</tr>
<tr>
<td>Total</td>
<td>3061.75051</td>
<td>1364</td>
<td>2.24468512</td>
<td>Root MSE = 1.453</td>
</tr>
</tbody>
</table>

F = (2894.69 – 2871.14)/2 / 2.11 = 5.58

Pr(F>5.58) with F distribution with 2 and 1360 degrees of freedom = 0.00385.

Decision: Smoking is statistically significantly associated with medical expenditures after adjusting for age and gender.

Now, lets look at another example using logistic regression:

You would like to determine the association between COPD and smoking status (never/current/former) after adjusting for age and gender.

Your variables are:

\[
\text{COPD} = 1 \text{ if present, } 0 \text{ if absent}
\]

\[
\text{Smoke} = 0 \text{ if never, } 1 \text{ if current, } 2 \text{ if former}
\]

\[
\text{Age} = 40 – 94 \text{ (most plausible range of age for the disease)}
\]
Male = 1 if male, 0 if female

Our logistic regression model becomes:

\[ \log \text{it}[\text{COPD} = 1] = \beta_0 + \beta_1 S_1 + \beta_2 S_2 + \beta_3 \text{Age} + \beta_4 \text{Male} + \varepsilon \]

Where

\[ S_1 = \begin{cases} 1 & \text{if current} \\ 0 & \text{if never} \end{cases} \]

\[ S_2 = \begin{cases} 1 & \text{if former} \\ 0 & \text{if never} \end{cases} \]

Therefore, you can write out a regression model for the never, current and former smokers.

Never smokers:

\[ \log \text{it}[\text{COPD} = 1] = \beta_0 + \beta_3 \text{Age} + \beta_4 \text{Male} + \varepsilon \]

Current smokers:

\[ \log \text{it}[\text{COPD} = 1] = \beta_0 + \beta_1 + \beta_3 \text{Age} + \beta_4 \text{Male} + \varepsilon \]

Former smokers:

\[ \log \text{it}[\text{COPD} = 1] = \beta_0 + \beta_2 + \beta_3 \text{Age} + \beta_4 \text{Male} + \varepsilon \]

So,

\[ \beta_1 = \log \text{difference in the odds of COPD comparing current smokers to never smokers of the same age and gender, or the log OR comparing current smokers to never smokers, of the same age and gender.} \]

\[ \beta_2 = \log \text{difference in the odds of COPD comparing former smokers to never smokers of the same age and gender, or the log OR comparing former smokers to never smokers, of the same age and gender.} \]

The test of interest is to determine if smoking is associated with COPD. To do this, we will compare the null model (includes age and gender) to the extended model (including dummy variables for smoking status and age and gender).

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

Fit the null and extended model and obtain the log-likelihood and perform your test as in the notes for topic 2.

(results below are based on a sample of the 1987 National Medical Expenditure Survey)

Null Model:
Logit estimates
Number of obs = 1000
LR chi2(2) = 12.62
Prob > chi2 = 0.0018
Log likelihood = -71.57169  Pseudo R2 = 0.0810

lc5  Coef.   Std. Err. z P>z     [95% Conf. Interval]
lastage  .0627023  .0226038 2.77 0.006     .0183996 .107005
male  1.191797  .5559726 2.14 0.032     .1021104 2.281483
_cons  -8.89722  1.655434 -5.37 0.000    -12.14181 -5.652629

Extended Model:

Logit estimates
Number of obs = 1000
LR chi2(4) = 21.27
Prob > chi2 = 0.0003
Log likelihood = -67.245224 Pseudo R2 = 0.1366

lc5       Coef. Std. Err. z P>z     [95% Conf. Interval]
S1   -.3273701 1.186354 -0.28 0.783    -2.652581 1.997841
S2    1.550378 .685478  2.26 0.024     .206866 2.89389
age     .0578063 .0235927 2.45 0.014     .0115655 .1040471
male    .8386232 .5803814 1.44 0.148    -.2989034 1.97615
_cons   -9.140392 1.78537  -5.12 0.000    -12.63965 -5.641131

Perform your likelihood ratio test:

-2(-71.57 – (-67.24)) = 8.66

Compare this value to the 0.05 critical-value from the Chi-square distribution with 2 df, which is 5.99.

Hence, our decision is to reject the null hypothesis and we conclude that there is evidence in the data to suggest an association between COPD and smoking status.