23. TESTING WHETHER THE MODEL IS USEFUL FOR PREDICTING $Y$

It is always possible that our “explanatory variables” are completely useless for predicting $y$. We can formulate this as the null hypothesis that all regression parameters (except the intercept) are zero, that is,

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 .$$

Under $H_0$, the true regression function does not depend at all on our explanatory variables, although our estimates $\hat{\beta}_1, \ldots, \hat{\beta}_k$ will almost certainly be different from zero, due to natural variability, i.e., “noise”, as opposed to “signal”.

To determine if the model is useful, we would like to perform a hypothesis test of $H_0$.

The alternative hypothesis $H_A$ is that $H_0$ is false.

So $H_A$ says that at least one of $\beta_1, \ldots, \beta_k$ is nonzero.

In other words, under $H_A$, the model is not completely useless.

(But notice that $H_A$ does not say that all coefficients are nonzero).

You might think that you could test $H_0$ versus $H_A$ by checking to see if any of the estimated coefficients is significantly different from zero. The problem with this approach is that you would be performing $k$ different hypothesis tests at the same time. Even if $H_0$ is true, the probability of finding at least one significant $t$-statistic at (let’s say) the 5% level of significance is actually much larger than 0.05.

The larger the value of $k$ is, the worse the problem becomes.

The level of significance of the $t$-test gives the probability of a Type I error for that test by itself. But in the scenario above, we have many chances to find a significant $t$-statistic. This affects the overall probability of a Type I error.

Here’s an analogy: If we draw a card at random from a shuffled deck, we only have a 1/52 chance of obtaining the Ace of Spades. But if we repeat this experiment 100 times, the probability that we will draw the Ace of Spades at least once is much larger. In fact, it’s almost 1.

A better way to test $H_0$ above is to use the "$F$-test".

This allows you to test $H_0$ versus $H_A$, with an actual Type I error rate of $\alpha$.

The test statistic is

$$F = \frac{SSR/k}{SSE/(n-k-1)} .$$

To understand the $F$-statistic better, we need to return to the Analysis of Variance part of the Minitab output.
Regression Analysis: Price versus Size, Age, Lot Size

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>3</td>
<td>570744</td>
<td>190248</td>
<td>40.03</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>11</td>
<td>52280</td>
<td>4753</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>623024</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Model Summary

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>R-sq</th>
<th>R-sq(adj)</th>
<th>R-sq(pred)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>68.9399</td>
<td>91.61%</td>
<td>89.32%</td>
<td>87.66%</td>
</tr>
</tbody>
</table>

Coefficients

<table>
<thead>
<tr>
<th>Term</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-161</td>
<td>191</td>
<td>-0.84</td>
<td>0.418</td>
</tr>
<tr>
<td>Size</td>
<td>41.46</td>
<td>7.51</td>
<td>5.52</td>
<td>0.000</td>
</tr>
<tr>
<td>Age</td>
<td>-2.36</td>
<td>8.81</td>
<td>-0.27</td>
<td>0.794</td>
</tr>
<tr>
<td>Lot Size</td>
<td>48.31</td>
<td>9.01</td>
<td>5.36</td>
<td>0.000</td>
</tr>
</tbody>
</table>

We can write the $F$-statistic as

$$F = \frac{MSR}{MSE},$$

where $MSR = \frac{SSR}{k}$ and $MSE = \frac{SSE}{n-k-1}$.

Here, MS stands for "Mean Square".

So MSE is the "Mean Square for Residuals", that is, the ratio of $SSE$ to the (Residual) degrees of freedom, $n-k-1$. Note that $MSE$ is the same as $s^2$.

In the housing example, from the Minitab output, we find that $MSE = 4753$, which is the ratio $SSE/DF(Residual) = 52280/11$.

Minitab gives three entries for degrees of freedom in the DF column: DF(Regression), DF(Residual), and DF(Total).

We have already mentioned $DF(Residual) = n-k-1$. This is the DF we have been using to calculate confidence intervals and tests on the individual regression parameters.

In fact, $DF(Residual)$ is the number of degrees of freedom available for estimating $\sigma^2$. The estimator, $s^2$, is based on the residuals, which have lost $k+1$ degrees of freedom from the original $n$ because of the need to estimate $k+1$ parameters in order to calculate the residuals.

DF(Regression) is $k$, the number of regression parameters, not counting the intercept.

DF(Total) is $n-1$, the same as the df we have used in estimating the variance of the $y$'s.

For the housing example, DF(Regression) = 3, DF(Residual) = 11, and DF(Total) = 14.

Note: The value of $n$ does not appear anywhere in the output, but we can get it by adding one to DF(Total).
The Mean Square for Regression is defined as
\[ MSR = \frac{SSR}{k} = \frac{SSR}{DF(Regression)} \]

MSR represents the "average" squared fluctuation of the \( \hat{y} \) values about their mean, \( \bar{y} \).

In the housing example, \( MSR = 190248 = 570744 / 3 \).

The F statistic is the ratio \( F = \frac{MSR}{MSE} \).

In the housing example, \( F = 40.03 = 190248 / 4753 \).

It is clear from its definition that the \( F \) statistic is always positive.

If \( F = \frac{MSR}{MSE} \) is sufficiently large, we can reject \( H_0 \) in favor of \( H_A \).

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>( k )</td>
<td>SSR</td>
<td>MSR = ( \frac{SSR}{k} )</td>
<td>( F = \frac{MSR}{MSE} )</td>
<td></td>
</tr>
<tr>
<td>Residual (&quot;Error&quot;)</td>
<td>( n-k-1 )</td>
<td>SSE</td>
<td>MSE = ( \frac{SSE}{n-k-1} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( n-1 )</td>
<td>SST = SSR + SSE</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If \( H_0 \) is true, then the \( F \)-statistic has an \( F \) distribution with \( k \), and \( n-k-1 \) degrees of freedom.

Note that the \( F \) distribution has two different df: one for the numerator (\( \nu_1 \)), one for the denominator (\( \nu_2 \)).

We have \( \nu_1 = k \), \( \nu_2 = n-k-1 \).

The \( F \) distribution is skewed to the right, since the \( F \)-statistic can never be negative. If the \( F \)-statistic exceeds \( F_\alpha \) (Table 9), then we reject \( H_0 \) at level \( \alpha \). This test is inherently right-tailed, since we are looking for a large value of the \( F \) statistic as evidence that the regression is useful. Thus, we use the critical value \( F_\alpha \), and not \( F_\alpha/2 \).
Fortunately, we can avoid using tables here by just looking at the \( p \)-value reported by Minitab.

In the housing example, Minitab gives a \( p \)-value for the \( F \)-test of \( p = 0.000 \). So we can reject \( H_0 \) at level 0.05, and also at level 0.01. The model does seem to have predictive power.

For a manual calculation, suppose we wanted to do a formal test at level 0.05. From Table 9 with \( \nu_1 = 3 \), \( \nu_2 = 11 \), we find \( F_{0.05} = 3.59 \). Since the observed \( F \) statistic exceeds 3.59, we reject \( H_0 \) at level 0.05. (We already knew this had to happen.)

Going back to the individual \( t \)-statistics for each parameter and picking those which are significant does not solve the problem, due partly to the multiple testing issues we discussed earlier.

It is tempting to conclude that if \( H_0 \) is rejected, the model, with all \( k \) variables, must be "good". But this is not necessarily true.

In the housing example, the coefficient for age is not significant \( (p=0.794) \).

So even though we have a highly significant \( F \)-statistic, \( (p=0.000) \), it is not clear that the age variable has predictive power, given that the other two variables are in the model. Perhaps the age variable should be deleted. The \( F \)-statistic tells us nothing about this.

### Practical Interpretation of the F-Test

Most books recommend that in analyzing multiple regression computer output we look first at the \( F \)-test. If the \( F \)-statistic is not significant, we do not continue with the analysis, since the regression is not useful for prediction.

From a practical point of view, there are some problems with the \( F \)-test. If the \( F \)-statistic exceeds the critical value, then we have some indication that at least one of the \( \beta_j \) is nonzero. However, the test gives us no clue as to which of the \( \beta_j \) is (are) nonzero.

Unfortunately, this is precisely what we will typically want to know in practice.

So even if \( H_0 \) is rejected, we need to worry about the possibility that we have too many variables in our model. We also may be missing some variable that is of great importance.

To recap: A significant \( F \)-statistic simply suggests that the model is not completely useless. It does not indicate that the model is "correct", or even "adequate".