33. INFERENCE FOR MULTIPLE REGRESSION

We can interpret most of the Minitab Multiple Regression output as we did in the simple regression case.

We should keep in mind, though, that we are now estimating several parameters.

This requires us to modify the degrees of freedom, and to think about the consequences of performing too many hypothesis tests at once.

Since we are estimating $k+1$ regression parameters ($\beta_0, \beta_1, ..., \beta_k$), we now have $n-(k+1) = n-k-1$ degrees of freedom.

In the housing example, we have $n=15$ and $k=3$, so $df=15-3-1=11$. 
Testing For Significance of an Individual Parameter

The Minitab $t$-statistics ($T$) can be used for testing that a given $\beta$ parameter is zero, that is, $H_0 : \beta_i = 0$. (Or to test $H_0 : \beta_0 = 0$).

If the model holds and $\beta_i = 0$, then the $t$-statistic corresponding to $\hat{\beta}_i$,

$$T = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}} = \frac{\text{"Coef"}}{\text{"SE Coef"}}$$

will have a $t$-distribution with $n-k-1$ degrees of freedom.

The $p$-value provided by Minitab corresponds to a two-tailed test of $H_0 : \beta_i = 0$ versus $H_A : \beta_i \neq 0$.

For the housing example, the coefficients of house size and lot size are statistically significant at the 5% level ($p < 0.05$), while the intercept and the coefficient for age are not significant ($p > 0.05$). Perhaps age should be deleted from the model, but we will leave it in for now.
Regression Analysis: Price versus Size, Age, Lot Size

The regression equation is
Price = - 161 + 41.5 Size - 2.36 Age + 48.3 Lot Size

Predictor    Coef  SE Coef      T      P
Constant   -160.6    190.7  -0.84  0.418
Size       41.462    7.512   5.52  0.000
Age        -2.361    8.812  -0.27  0.794
Lot Size   48.309    9.011   5.36  0.000

S = 68.9399   R-Sq = 91.6%   R-Sq(adj) = 89.3%

Analysis of Variance
Source          DF      SS      MS      F      P
Regression       3  570744  190248  40.03  0.000
Residual Error  11   52280    4753
Total           14  623024

Fit  SE Fit       95% CI            95% PI
1128.1    35.8  (1049.4, 1206.9)  (957.2, 1299.1)
Confidence intervals and tests for a general null hypothesis on a parameter

For each regression parameter, Minitab computes $t$-statistics and $p$-values for a null hypothesis of zero. For a general null hypothesis, you must construct your own $t$-statistic by hand. To decide whether it's statistically significant, you need to get the critical value from Table 6.

Even if $\beta_i$ is not zero, the estimator $\hat{\beta}_i$ is normally distributed, with $E[\hat{\beta}_i] = \beta_i$. Furthermore, the quantity $(\hat{\beta}_i - \beta_i) / s_{\hat{\beta}_i}$ has a $t$-distribution with $n-k-1$ degrees of freedom.

This allows us to get confidence intervals and perform general hypothesis tests for $\beta_i$. 
Eg: Suppose in the housing example that, before seeing the data, we had a hypothesis that every hundred square feet of house size adds an average of $10,000 to the selling price ($\beta_1=10$). To test $H_0 : \beta_1=10$ versus $H_A : \beta_1 \neq 10$ at level 0.05, we form the $t$-statistic,

\[
t = (41.462 - 10) / 7.512 = 4.19.
\]

If $H_0$ were true, such a $t$-statistic would have a $t$ distribution with 11 degrees of freedom. Thus, we need to compare our observed $t$-statistic to the critical value $t_{0.025} = 2.201$ from Table 6 (with df=11).

Conclusion: Reject the null hypothesis.
Next, let's construct a 95% confidence interval for $\beta_1$. From the Minitab output, we have $s_{\hat{\beta}_i} = 7.512$.

The confidence interval is

$$\hat{\beta}_1 \pm t_{\frac{a}{2}} s_{\hat{\beta}_i} = 41.462 \pm 2.201(7.512) = (24.9, 58.0).$$

Based on this interval, we can perform any two-tailed hypothesis test on $\beta_1$, without actually calculating the $t$-statistic.

For example, since the interval does not contain 10, we know that we can reject the hypothesis

$$H_0 : \beta_1 = 10 \text{ in favor of } H_A : \beta_1 \neq 10, \text{ at level } 0.05.$$
We can estimate the response surface $E(Y \mid x)$ and predict a future value using $\hat{y}$. We can also obtain confidence intervals for $E(Y \mid x)$ and prediction intervals for a future $y$ (for given values of the explanatory variables) from the Minitab output. For example, if a house has a ground area of 2000 feet ($x_1 = 20$), the house is 10 years old ($x_2 = 10$) and the lot size is 10,000 square feet ($x_3 = 10$), then $\hat{y} = 1128.1$, so the predicted selling price is $1,128,100$.

The 95 percent confidence interval for the mean selling price is ($1,049,400, 1,206,900$) and the 95 percent prediction interval for the price of the house is ($957,200, 1,299,100$).
The error variance $\sigma^2$ measures the dispersion of the data points from the true response surface.

We can estimate $\sigma^2$ without bias by $s^2 = \text{SSE}/(n-k-1)$.

The value of $s$ (an estimate of $\sigma$) is given by Minitab.

For the housing example, we get $s = 68.94$. 

**Estimating the Error Variance, $\sigma^2$**
Multiple Coefficient of Determination, $R^2$

Just as in simple regression, we have

$$SST = SSR + SSE,$$

where $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the total sum of squares,

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$ is the regression sum of squares, and

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$ is the residual sum of squares.

The coefficient of multiple determination is $R^2 = SSR/SST$.

The interpretations of these quantities are essentially the same as in simple linear regression. Thus, $R^2$ is the proportion of the variation in $y$ that is "explained" by the multiple regression model.
Regression Analysis: Price versus Size, Age, Lot Size

The regression equation is
Price = -161 + 41.5 Size - 2.36 Age + 48.3 Lot Size

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-160.6</td>
<td>190.7</td>
<td>-0.84</td>
<td>0.418</td>
</tr>
<tr>
<td>Size</td>
<td>41.462</td>
<td>7.512</td>
<td>5.52</td>
<td>0.000</td>
</tr>
<tr>
<td>Age</td>
<td>-2.361</td>
<td>8.812</td>
<td>-0.27</td>
<td>0.794</td>
</tr>
<tr>
<td>Lot Size</td>
<td>48.309</td>
<td>9.011</td>
<td>5.36</td>
<td>0.000</td>
</tr>
</tbody>
</table>

S = 68.9399  R-Sq = 91.6%  R-Sq(adj) = 89.3%

Analysis of Variance
<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</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>Residual 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>

Fit  SE Fit  95% CI  95% PI
1128.1  35.8  (1049.4, 1206.9)  (957.2, 1299.1)
In Minitab, $R^2$ is denoted by R-Sq.

For the housing data, we have $R^2 = 91.6\%$

A value of $R^2$ close to 1 is generally considered to imply that the model is good. We have to be somewhat careful, however, since $R^2$ is guaranteed to go up whenever we include a new variable.

The "Adjusted $R^2$" avoids this problem, but in my opinion it is not much more useful than $R^2$ itself for deciding which variables to use in the model.
In Minitab, we can read the values of SSR, SSE and SST from the "SS" column of the "Analysis of Variance" table.

For the housing data, we have SSR = 570744 ("Regression"), SSE = 52280 ("Residual Error"), and SST = 623024 ("Total").

We can check that 623024 = 570744 + 52280 (SST=SSR+SSE), and that 0.916 = 570744 / 623024 (R² = SSR/SST).

We can also compute s² = SSE/(n–k–1) = 52280/11 = 4753.

The square root of this is s = 68.94 as given earlier.