NOTES for Wednesday 24 FEB 2010

We need to talk about inferences that we make with regression. First we will need to be very clear about the model assumptions. We’ll do this for simple regression ($K = 1$) and later for multiple regression ($K \geq 2$).

$$E(Y_i) = \beta_0 + \beta_1 x_i$$

This is the linearity assumption. The expected value of $Y$ is really a straight-line function of $x$.

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

This is the additive error (noise) assumption. In particular, the right side is not $(\beta_0 + \beta_1 x_i) \varepsilon_i$. It’s also not $(\beta_0 + \beta_1 x_i) e^{\varepsilon_i}$.

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent random variables, each with the same distribution

The independence means that $\varepsilon_i$ and $\varepsilon_j$ are independent random variables when $i \neq j$. It also means that $\varepsilon_i$ is independent of $x_i$ and of all the other independent variables. The “same distribution” assumption really means that the $\varepsilon$’s are a random sample.

$$E(\varepsilon_i) = 0 \text{ and } SD(\varepsilon_i) = \sigma$$

The $\varepsilon$’s are assumed to be taken from a population with mean zero. This is a costless assumption. If $E(\varepsilon_i) = \psi \neq 0$, then we would rewrite the model as $Y_i = (\beta_0 + \psi) + \beta_1 x_i + (\varepsilon_i - \psi)$. In this rewritten form $(\beta_0 + \psi)$ is an unknown intercept and the $(\varepsilon_i - \psi)$ are unobserved noise terms with mean 0.

The $\varepsilon$’s are assumed to all have the same standard deviation. We need to be wary about possible violations of this assumption, as it is fairly common for $SD(\varepsilon_i)$ to be larger for those data points that have large $x_i$.  

B90.2301 / C22.0017
This says also that $SD(Y_i) = SD(\beta_0 + \beta_1 x_i + \epsilon_i) = SD(\epsilon_i) = \sigma$. The “signal” $\beta_0 + \beta_1 x_i$ is not random, and it does not contributed to the standard deviation.

$\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are normally distributed.

Believe it or not, we don’t need to make this assumption in all cases. If our only concerns are about means, standard deviations, and correlations we do not need to invoke normality. The normal distribution assumption is critical for the inferential acts of confidence intervals, hypothesis tests, and predictions. Any activity that gets to using the $t$ distribution or the $F$ distribution must be based on the assumption of normality.

As a consequence of all these assumptions, including normality, we can show

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

The $\sim$ means “is distributed as.” The notation $N(\text{mean}, \text{variance})$ is also used here. That is, $b_1$ is normally distributed with mean $\beta_1$ and with variance $\frac{\sigma^2}{S_{xx}}$. The standard deviation is of course $\frac{\sigma}{\sqrt{S_{xx}}}$.

There is an interesting derivation for this.

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{S_{xx}}$$

It’s interesting that we can drop $\bar{Y}$ from this arithmetic.

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + \epsilon_i)}{S_{xy}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_1 x_i + \epsilon_i)}{S_{xx}}$$

Note that $\beta_0$ disappears.
\[
\beta_1 \sum_{i=1}^{n}(x_i - \bar{x})x_i + \sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i
\]
\[
= \beta_1 \frac{\sum_{i=1}^{n}(x_i - \bar{x})x_i + \sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{S_{xx}}
\]

Note that \(\bar{x}\) can be reintroduced.

\[
= \beta_1 S_{xx} + \sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i
\]
\[
= \frac{\beta_1 S_{xx} + \sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{S_{xx}}
\]

Recall that \(S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2\).

\[
= \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{S_{xx}}
\]

Since this is a linear combination of the \(\varepsilon_i\)'s, it must be normally distributed. (The \(x_i\)'s are treated as non-random.) Since \(E(\varepsilon_i) = 0\), this also shows that \(E(b_1) = \beta_1\).

All that remains is finding \(\text{Var}(b_1)\).

\[
\text{Var}(b_1) = \text{Var}\left(\beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{S_{xx}}\right)
\]
\[
= \text{Var}\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{S_{xx}}\right)
\]

Since \(\beta_1\) is not random, it is not involved in the variance.

\[
= \frac{1}{(S_{xx})^2} \text{Var}\left(\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i\right)
\]
\[
\begin{align*}
&= \frac{1}{(S_{xx})^2} \sum_{i=1}^{n} \operatorname{Var}\{ (x_i - \bar{x}) e_i \} \\
&= \frac{1}{(S_{xx})^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \operatorname{Var}(e_i) = \frac{1}{(S_{xx})^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \sigma^2 \\
&= \frac{\sigma^2}{(S_{xx})^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{\sigma^2}{(S_{xx})^2} S_{xx} = \frac{\sigma^2}{S_{xx}}
\end{align*}
\]

\[
b_0 \sim N \left( \beta_0, \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right] \right)
\]

\[
\operatorname{Cov}(b_0, b_1) = -\bar{x} \frac{\sigma^2}{S_{xx}} \quad \text{(This does not require the normal assumption.)}
\]

\[
\frac{(n-2) s_e^2}{\sigma^2} \sim \chi^2_{n-2}
\]

This is the chi-squared distribution. You will also see this expressed as \( s_e^2 \sim \sigma^2 \frac{\chi^2_{n-2}}{n-2} \). It can be shown that the expected value of a chi-squared random variable is its degrees of freedom. Thus \( E(\chi^2_{n-2}) = n-2 \), leading us to \( E(s_e^2) = \sigma^2 \). We will say “\( s_e^2 \) is an unbiased estimate of \( \sigma^2 \).” It does not follow that \( s_e \) is an unbiased estimate of \( \sigma \). Since \( MS_{\text{resid}} = \frac{SS_{\text{resid}}}{n-2} = s_e^2 \), this result can be expressed as \( SS_{\text{resid}} \sim \sigma^2 \chi^2_{n-2} \).

\( s_e \) is statistically independent of \((b_0, b_1)\)

A derived quantity, such as \( b_1 \), has a standard deviation that depends on unknown things (usually \( \sigma \)). Here \( \text{SD}(b_1) = \frac{\sigma}{\sqrt{S_{xx}}} \). We can estimate this standard deviation by estimating the unknown thing, and we call this
estimate the standard error. Since \( s_e \) estimates \( \sigma \), we write \( \text{SE}(b_1) = \frac{s_e}{\sqrt{S_{xx}}} \).

Alas, not everyone uses this strict definition for standard error.

In many problems with normal distributions and estimates of \( \sigma \) from a chi-squared distribution, we can create

\[
\frac{b_1 - E(b_1)}{\text{SE}(b_1)} = \frac{b_1 - \beta_1}{s_e} = \frac{t(n-2)}{S_{xx}^{1/2}}
\]

and it will have a \( t \) distribution. The exact statements are a bit tricky, and that will be omitted here. The degrees of freedom is the degrees of freedom number from the chi-squared distribution.

This leads us to

\[
\frac{b_1 - E(b_1)}{\text{SE}(b_1)} = \frac{b_1 - \beta_1}{s_e} \sim t(n-2)
\]

This can be used to create a \( 1 - \alpha \) confidence interval for \( \beta_1 \). Thus

\[
1 - \alpha \text{ confidence interval for } \beta_1 \text{ is } b_1 \pm t_{\alpha/2; n-2} \frac{s_e}{\sqrt{S_{xx}}}.
\]

Similar logic gets a \( 1 - \alpha \) confidence interval for \( \beta_0 \).

\[
1 - \alpha \text{ confidence interval for } \beta_0 \text{ is } b_0 \pm t_{\alpha/2; n-2} s_e \sqrt{\frac{1}{n} + \frac{x^2}{S_{xx}}}
\]

The intervals for \( \beta_0 \) and \( \beta_1 \) are not statistically independent.

*If* it happens that \( \beta_1 = 0 \), then \( SS_{\text{regression}} \) is distributed as \( \sigma^2 \chi^2_1 \), independent of \( SS_{\text{resid}} \). If \( \beta_1 \neq 0 \), this distribution is a non-central chi-squared, but that’s another story.

*If* it happens that \( \beta_1 = 0 \), then \( F = \frac{MS_{\text{regression}}}{MS_{\text{resid}}} \) follows the \( F \) distribution with \( (1, n-2) \) degrees of freedom. If \( \beta_1 \neq 0 \), the distribution is non-central \( F \) and tends to be big. (That’s why we reject \( \beta_1 = 0 \) with large \( F \) statistics.)
Let’s suppose that we a target value $x_{\text{new}}$ in mind, and we wanted to predict the population average $Y$ that would result from this $x_{\text{new}}$. This is asking for a prediction of $E Y_{\text{new}} = \beta_0 + \beta_1 x_{\text{new}}$, which is something we generally do not need to do!

The point prediction is clearly $b_0 + b_1 x_{\text{new}}$. We can obtain the variance of this prediction:

$$\text{Var}(b_0 + b_1 x_{\text{new}}) = \text{Var}(b_0) + 2x_{\text{new}} \text{Cov}(b_0, b_1) + x_{\text{new}}^2 \text{Var}(b_1)$$

$$= \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right] + 2x_{\text{new}} \left( -\bar{x} \frac{\sigma^2}{S_{xx}} \right) + x_{\text{new}}^2 \frac{\sigma^2}{S_{xx}}$$

$$= \sigma^2 \left\{ \frac{1}{n} + \frac{1}{S_{xx}} \left( \bar{x}^2 - 2x_{\text{new}}\bar{x} + x_{\text{new}}^2 \right) \right\}$$

$$= \sigma^2 \left\{ \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}} \right\}$$

In this expression, $\bar{x}$ and $S_{xx}$ are based on the original $n$ data points, and $x_{\text{new}}$ is not part of the original data.

It follows that we can make a $t$ distribution out of this:

$$\frac{(b_0 + b_1 x_{\text{new}}) - (\beta_0 + \beta_1 x_{\text{new}})}{s_e \sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

Then, of course, we can give a $1 - \alpha$ confidence interval for $\beta_0 + \beta_1 x_{\text{new}}$. Here it is:

$$b_0 + b_1 x_{\text{new}} \pm t_{\alpha/2,n-2} s_e \sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}$$

This is all logically correct, but it begs the question. Why exactly might you want a confidence interval for $\beta_0 + \beta_1 x_{\text{new}}$?
Vastly more useful is a prediction for $Y_{\text{new}}$. The obvious prediction is $\hat{Y}_{\text{new}} = b_0 + b_1 + x_{\text{new}}$. In this expression, we regard $x_{\text{new}}$ as non-random, but we certainly must regard $b_0$ and $b_1$ (and thus $\hat{Y}_{\text{new}}$) as random.

Let’s use the symbol $Y_{\text{new}}$ as the new value, which we’ll see eventually. The quantity $\hat{Y}_{\text{new}} - Y_{\text{new}}$ is the prediction error. We can get to the statistical properties of this by using the model on $Y_{\text{new}}$. Specifically, we have $Y_{\text{new}} = \beta_0 + \beta_1 x_{\text{new}} + \epsilon_{\text{new}}$. The new noise term $\epsilon_{\text{new}}$ is statistically independent of everything that came before it.

Certainly the prediction error $\hat{Y}_{\text{new}} - Y_{\text{new}}$ will follow a normal distribution. We can show that the mean is zero.

$$E[\hat{Y}_{\text{new}} - Y_{\text{new}}] = E\left[ (b_0 + b_1 x_{\text{new}}) - (\beta_0 + \beta_1 x_{\text{new}} + \epsilon_{\text{new}}) \right]$$

$$= (E(b_0) + E(b_1) x_{\text{new}}) - (\beta_0 + \beta_1 x_{\text{new}} + E(\epsilon_{\text{new}})) = 0$$

The variance is trickier:

$$\text{Var}[\hat{Y}_{\text{new}} - Y_{\text{new}}] = \text{Var}\left[ (b_0 + b_1 x_{\text{new}}) - (\beta_0 + \beta_1 x_{\text{new}} + \epsilon_{\text{new}}) \right]$$

$$= \text{Var}\left[ (b_0 + b_1 x_{\text{new}}) - \epsilon_{\text{new}} \right]$$

$$= \text{Var}(b_0) + 2x_{\text{new}} \text{Cov}(b_0, b_1) + x_{\text{new}}^2 \text{Var}(b_1) + \text{Var}(\epsilon_{\text{new}})$$

$$= \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right] + 2x_{\text{new}} \left( -\bar{x} \frac{\sigma^2}{S_{xx}} \right) + x_{\text{new}}^2 \frac{\sigma^2}{S_{xx}} + \sigma^2$$

$$= \sigma^2 \left\{ 1 + \frac{1}{n} \left( \frac{x_{\text{new}} - \bar{x}}{S_{xx}} \right) \right\}$$

This leads to a $t$ distribution:
You will frequently see the very useful prediction interval for $Y_{new}$:

$$\hat{Y}_{new} - Y_{new} \sim t_{n-2}$$

$$s_e \sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}$$

It should be noted that the intervals for two new values, say $x_{new_1}$ and $x_{new_2}$ are not statistically independent of each other.

You will also see plot of prediction bands. These are not a confidence region for the line.

It must be noted that these distributional results depend on the correctness of the model. It is curious to note that the results hold exactly only if we make no effort to check the integrity of the model!

Let’s think about that. If we check the model, perhaps by examining the data or looking at residual plots, we might take actions. We might replace $Y$ by its logarithm. We might delete a point as being an outlier. The actions are, of course, possible sources of errors.

As a simple illustration of that, let’s recall the simple (pre-regression) problem based on a sample of values $w_1, w_2, \ldots, w_n$. We make the model that these are a sample from a normal population with mean $\mu_w$ and standard deviation $\sigma_w$.

Then a 95% confidence interval for $\mu_w$ is $\bar{w} \pm t_{0.025; n-1} \frac{s}{\sqrt{n}}$. This is true if the model is true and we don’t examine our data! Suppose that we have a rule for finding outliers and throwing them away. Whether we make correct or incorrect decisions in any particular case, we will still be messing up the 95% confidence.

Of course, every practicing statistician would look at the data to check for assumptions. We would look very foolish if we did not. But let’s at least be honest and admit that looking at the data may alter our confidence values!
Residuals also help identify model failure, especially through using the residual-versus-fitted plot. (There’s a section in the Simple Linear Regression pamphlet on this; the commonly encountered is spreading residuals. The cure is taking logarithms of $Y$.

Let’s examine this in the context of the Salary data set. This is M\$SALARY.MTP

Here’s the residual versus fitted plot:

![Residuals Versus the Fitted Values](image)

This is a clear case of expanding residuals.

The cure for this one is to replace $Y$ by log $Y$.

* We recommend base-$e$ logarithms.
* If the $Y$’s have some zeroes or negative values, use $\log(Y + c)$, choosing $c$ to elevate all the values to just above zero.

We can point out also the “gross size” phenomenon. Variables that run over several orders of magnitude are almost always treated with logarithms.

The residual versus fitted plot will also let us detect curvature. It can find weird values (outliers?), but we have to be careful with such funny points.

If any point is weird, it must be checked for correct coding.

If the point is not recorded correctly, fix it and start over. If it cannot be fixed, then discard it, but record this action in the report.
If the point is correct, try to decide if it has undue impact on the regression. If removing the point gives a big change in $b_1$, then you might consider removing it. In any case, if you remove a point, that action must be included in the report.

For multiple regression the model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_K x_{iK} + \epsilon_i$$

We say that this is regression with $K$ predictors. For $K = 1$, we call it simple regression and for $K \geq 2$, it’s multiple regression.

The notation is a problem. The subscript-free version of the model is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_K x_K + \varepsilon$$

In this form, the independent variables would be described as $X_1, X_2, \ldots, X_K$. (We can use upper case letters when we think of them as variables.)

In applications, these will have names. The model might look like

$$\text{SALES}_i = \beta_0 + \beta_{\text{SOCKS}} \text{SOCKS}_i + \beta_{\text{SHIRTS}} \text{SHIRTS}_i + \beta_{\text{PANTS}} \text{PANTS}_i + \epsilon_i$$

This would avoid the confusion of the double subscripts on the $x$ terms.

The estimating criterion is again least squares. Specifically, we try to minimize

$$\sum_{i=1}^{n} \left( Y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + \ldots + b_K x_{iK}) \right)^2$$

You’ll also see this as

$$\sum_{i=1}^{n} \left( Y_i - \sum_{j=0}^{K} b_j x_{ij} \right)^2$$

This requires that we identify $x_{i0} \equiv 1$ for every $i$.

We have a handout on the matrix notation.

A critical issue is that the solution vector $b = (X'X)^{-1}X'Y$ requires the inversion of a $p$-by-$p$ matrix. This can fail for a number of reasons:
* Rank(\(X\)) < \(p\). The \(p\) columns of \(X\) are not linearly independent vectors. The handout on matrix notation shows this.

* Not enough data, meaning \(n < p\). In that case, the matrix \(X\) has more columns than rows. This automatically makes Rank(\(X\)) \(\leq n < p\).

There is also the strange computational concern in which Rank(\(X\)) is just barely \(p\). This happens when one of the columns of \(X\) is very, very close to being an exact linear combination of other columns. We won’t discuss this in detail, as the problem shows up as collinearity.

The measure of this concern is the ratio \[
\frac{\text{largest eigenvalue of } (X'X)^{-1}}{\text{smallest eigenvalue of } (X'X)^{-1}}.
\]

There is this difficult interpretation issue with regard to the estimated coefficients. Examine the animal set. Consider the regression of Gestation on (Bweight, Tsleep).

Get matrix plot. Discover the Bweight needs logs. Then run the regression, finding that Gestation also needs logs.

For coherence, we’ll need to eliminate any animal that is not complete on these three variables.

Here’s the regression:

\[\text{Regression Analysis: log(G) versus log(W), Tsleep}\]

\[\text{The regression equation is}\]
\[\text{log}(G) = 5.01 + 0.178 \text{ log}(W) - 0.0773 \text{ Tsleep}\]

<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>5.0135</td>
<td>0.2959</td>
<td>16.95</td>
<td>0.000</td>
</tr>
<tr>
<td>log(W)</td>
<td>0.17754</td>
<td>0.03585</td>
<td>4.95</td>
<td>0.000</td>
</tr>
<tr>
<td>Tsleep</td>
<td>-0.07728</td>
<td>0.02424</td>
<td>-3.19</td>
<td>0.002</td>
</tr>
</tbody>
</table>

\[S = 0.652116 \quad R^2 = 62.2\% \quad R^2(\text{adj}) = 60.7\%\]

\[\text{Analysis of Variance}\]

<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>2</td>
<td>35.709</td>
<td>17.854</td>
<td>41.98</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>51</td>
<td>21.688</td>
<td>0.425</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>53</td>
<td>57.397</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This is a good regression (although it’s dependent in an absurd way on the style in which the animals were chosen). Let’s look at the estimated coefficient 0.178. The interpretation here is that
a one unit change in log(bw) leads to a change of 0.178 in log(G), holding Tsleep constant

There’s another way to look at this. Let log(bw)[Tsleep] be the residual from regressing log(bw) on Tsleep. That is, this is the part of log(bw) that cannot be explained by Tsleep.

Similarly, let log(G)[Tsleep] be the residual from regressing log(G) on Tsleep. This is the part of log(G) that cannot be explained by Tsleep.

Now . . . if we regress log(G)[Tsleep] on log(bw)[Tsleep], we get

Regression Analysis: log(G)[Tsleep] versus log(bw)[Tsleep]

The regression equation is

log(G)[Tsleep] = -0.0000 + 0.178 log(bw)[Tsleep]

<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>-0.0000</td>
<td>0.08788</td>
<td>-0.00</td>
<td>1.000</td>
</tr>
<tr>
<td>log(bw)[Tsleep]</td>
<td>0.17754</td>
<td>0.03551</td>
<td>5.00</td>
<td>0.000</td>
</tr>
</tbody>
</table>

S = 0.645815  R-Sq = 32.5%  R-Sq(adj) = 31.2%

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>1</td>
<td>10.429</td>
<td>10.429</td>
<td>25.00</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>52</td>
<td>21.688</td>
<td>0.417</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>53</td>
<td>32.117</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note that the slope coefficient is 0.178!

As a side thought question, why is the intercept zero?

Many things need to be checked in a multiple regression. The set CONDO209.mtp is useful for this.

* The residual versus fitted should be patternless. If the residuals expand to the right, you’ll need to use log(Yi), or perhaps log(Yi + c).

* With other strange patterns in the residual versus fitted plot, you may wish to plot the residuals against each independent variable.

This will suggest that we have curvature in DELEV. It’s hard to know if this was worth correcting, but we did.
Some people worry about interactions of predictors. The general method for doing this is to add a product term. In the animal regression of log(G) on \{log(W), Tsleep\}, consider also the regression of log(G) on \{log(W), Tsleep, log(W) \times Tsleep\}. That is, the regression is extended with an ordinary product.

For this example, the improvement in \(R^2\) is trifling, and the new variable (the product) is not statistically significant. With a \(p\)-value of 0.10 it does look interesting.

Warning:
If you do variable selection and you decide to keep the log(W) \times Tsleep interaction, you must also keep the predictors log(W) and Tsleep. In other words, never consider regressing log(G) on \{log(W), log(W) \times Tsleep\}.

In a multiple regression, the \(t\) statistics are used to test hypotheses about individual coefficients. If the model is \(Y_i = \beta_0 + \beta_H H_i + \beta_K K_i + \beta_Q Q_i + \beta_V V_i + \epsilon_i\), then the \(t\) statistic printed in the line for \(b_Q\) (the estimate of \(\beta_Q\)) tests the hypothesis \(H_0(Q): \beta_Q = 0\) against the alternative \(H_1(Q): \beta_Q \neq 0\). This must be interpreted in the context of this regression and with the predictors \{\(H, K, V\}\}. This is most definitely not independent of the tests on \(\beta_H, \beta_K\), or \(\beta_V\).

Suppose instead that you wished to test the hypothesis \(H_0: \beta_H = \beta_K\) against the alternative \(H_1: \beta_H \neq \beta_K\). This kind of inference is not automated in Minitab. We can get it, but we’ll have to work for it.

As a side note, you can only consider \(\beta_H = \beta_K\) if variables \(H\) and \(K\) are in the same units.

We are going to use the “partial \(F\) test.” This is explained neatly in Chatterjee and Hadi. Let’s suppose that we have a full model and a reduced model. In our example,

the full model is \(Y_i = \beta_0 + \beta_H H_i + \beta_K K_i + \beta_Q Q_i + \beta_V V_i + \epsilon_i\)

the reduced model is \(Y_i = \beta_0 + \beta_S S_i + \beta_Q Q_i + \beta_V V_i + \epsilon_i\)

where \(S_i = H_i + K_i\)
A requirement for using this partial $F$ test is that the reduced model must be a sub-model of the full model. This means that there is some choice for the parameters in the full model that produces the reduced model. Start from

$$Y_i = \beta_0 + \beta_H H_i + \beta_K K_i + \beta_Q Q_i + \beta_V V_i + \varepsilon_i$$

and then make the choice that $\beta_H = \beta_K$ to write

$$Y_i = \beta_0 + \beta_H H_i + \beta_H K_i + \beta_Q Q_i + \beta_V V_i + \varepsilon_i$$

This can be regrouped as

$$Y_i = \beta_0 + \beta_H (H_i + K_i) + \beta_Q Q_i + \beta_V V_i + \varepsilon_i$$

Now renaming $H_i + K_i$ as $S_i$ and also letting $\beta_S$ be another name for the assumed common value of $\beta_H$ and $\beta_K$ will produce the reduced model.

Do regressions for both the full model and the reduced model. Observe that

$$SS_{\text{regression}(\text{full})} > SS_{\text{regression}(\text{reduced})}$$

$$SS_{\text{resid}(\text{full})} < SS_{\text{resid}(\text{reduced})}$$

$$SS_{\text{regression}(\text{full})} - SS_{\text{regression}(\text{reduced})} = SS_{\text{resid}(\text{reduced})} - SS_{\text{resid}(\text{full})}$$
The partial $F$ statistic can be written in several ways:

\[
F = \frac{\{SS_{\text{regression}}(\text{full}) - SS_{\text{regression}}(\text{reduced})\} \div \nu}{SS_{\text{resid}}(\text{full}) \div DF_{\text{resid}}(\text{full})}
\]

\[
= \frac{\{SS_{\text{resid}}(\text{reduced}) - SS_{\text{resid}}(\text{full})\} \div \nu}{SS_{\text{resid}}(\text{full}) \div DF_{\text{resid}}(\text{full})}
\]

\[
= \frac{\{SS_{\text{regression}}(\text{full}) - SS_{\text{regression}}(\text{reduced})\} \div \nu}{MS_{\text{resid}}(\text{full})}
\]

\[
= \frac{\{SS_{\text{resid}}(\text{reduced}) - SS_{\text{resid}}(\text{full})\} \div \nu}{MS_{\text{resid}}(\text{full})}
\]

In every form, the numerator is based on a sum squares difference. How much better is the fit with the full model? The value $\nu$ is the number of parameters that distinguish $H_0$ from $H_1$. For this example, the full model has $E(Y_i) = \beta_0 + \beta_H H_i + \beta_K K_i + \beta_O Q_i + \beta_V V_i$ and it takes five parameters to specify this ($\beta_0$, $\beta_H$, $\beta_K$, $\beta_O$, $\beta_V$)

the reduced model has $E(Y_i) = \beta_0 + \beta_S S_i + \beta_O Q_i + \beta_V V_i$

and it takes four parameters to specify this ($\beta_0$, $\beta_S$, $\beta_O$, $\beta_V$)

For this example, $\nu = 1$.

The denominator of the partial $F$ statistic is the mean square residual from the full model. This denominator estimates $\sigma^2$ whether $H_0$ is true or not.

The degrees of freedom for the partial $F$ are $(\nu, DF_{\text{resid}(\text{full})})$. The rule is to reject $H_0$ at level $\alpha$ if the partial $F$ equals or exceeds $F^{\alpha}_{\nu, DF_{\text{resid}(\text{full})}}$, the upper $\alpha$ point for the $F$ distribution with $(\nu, DF_{\text{resid}(\text{full})})$ degrees of freedom.

We are out of the habit of looking up cutoff points for the $F$ distribution because most software prints the $p$-value. We would just reject at level $\alpha$ if $p \leq \alpha$.

The partial $F$ is not arranged by Minitab. You have to do two runs and assemble this yourself. This also means that you won’t get a $p$-value. You could then use printed tables of the $F$ distribution, which give the upper 5% and 1% points. Since Minitab is
already active for the work you’re doing, you can get \( p \)-value yourself. Let’s suppose that your partial \( F \) works out to the value 7.22 with (2, 71) degrees of freedom. Use **Calc \( \Rightarrow \) Probability Distributions \( \Rightarrow \) F** and then fill in the resulting panel as follows:

![F Distribution](image)

This starts with the default radio button setting at **Cumulative probability**, which is exactly what you want. The output is this:

```
Cumulative Distribution Function

F distribution with 2 DF in numerator and 71 DF in denominator

\[ x \quad P( X \leq x ) \]

7.22 \quad 0.998601
```

The probability to the left of your 7.22 is found to be 0.9986. The probability to the right (corresponding to the rejection region for \( H_0 \) versus \( H_1 \)) is then 0.0014. You can report then \( p = 0.0014 \).