INTRODUCTORY NOTES ON LINEAR REGRESSION

ANALYSIS OF VARIANCE TABLE

We have data of the form \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\). These will most often be presented to us as two columns of a spreadsheet.

As the topic develops we will see both upper case and lower case letters. Where reasonable, we use upper case \(X_i\) and \(Y_i\) to denote random quantities lower case \(x_i\) and \(y_i\) to denote non-random quantities which are possibly the observed values of random \(X_i\) and \(Y_i\).

We are not able to enforce this distinction in a consistent way, so for now do not be overly concerned.

Most regressions are summarized in the analysis of variance table. The development below is for simple regression, meaning \(K = 1\). The analysis of variance table for multiple regression, \(K \geq 2\), is presented also.

The quantity \(S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2\) measures variation in \(Y\). Indeed we get \(s_y\), the standard deviation of the \(Y\)'s, as \(s_y = \sqrt{\frac{S_{yy}}{n-1}}\).

We use the symbol \(\hat{Y}_i\) to denote the fitted value for point \(i\). This is computed from the estimated intercept and slope as \(\hat{Y}_i = b_0 + b_1 x_i\). Compare this expression to the model equation.

One can show that \(\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{Y}_i)^2\). These sums have the names \(SS_{total}\), \(SS_{regression}\), and \(SS_{error}\). They have other names or abbreviations. For instance

\[
SS_{total} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \text{ may be written as } SS_{tot}.
\]

\[
SS_{regression} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{y})^2 \text{ may be written as } SS_{reg}, SS_{fit} \text{, or } SS_{model}.
\]

\[
SS_{error} = \sum_{i=1}^{n} (y_i - \hat{Y}_i)^2 \text{ may be written as } SS_{err}, SS_{residual}, SS_{resid}, \text{ or } SS_{res}.
\]
This derivation starts from \( SS_{\text{total}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \).

\[
\begin{align*}
\sum_{i=1}^{n} (y_i - \bar{y})^2 &= \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} \left( (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \right)^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + 2 \sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\
&= SS_{\text{error}} + 2 \sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + SS_{\text{regression}}
\end{align*}
\]

Thus, all we have to deal with is the middle term; we need to show that it is equal to zero.

\[
\begin{align*}
\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))( (b_0 + b_1 x_i) - \bar{y}) \\
&= \sum_{i=1}^{n} (y_i - (\bar{y} - b_0 \bar{x} + b_1 x_i))( (\bar{y} - b_0 \bar{x}) + b_1 x_i) - \bar{y}) \\
&= \sum_{i=1}^{n} ( (y_i - \bar{y}) - b_1 (x_i - \bar{x}) )( b_1 (x_i - \bar{x})) = b_1 ( S_{xy} - b_1 S_{xx} )
\end{align*}
\]

However, \( b_1 = \frac{S_{xy}}{S_{xx}} \), so that this is indeed zero!

The more general proof (for two or more predictors) can be done through matrix algebra!

The degrees of freedom accounting is this:

- \( SS_{\text{total}} \) has \( n - 1 \) degrees of freedom
- \( SS_{\text{regression}} \) has \( K \) degrees of freedom (\( K \) is the number of independent variables)
- \( SS_{\text{error}} \) has \( n - 1 - K \) degrees of freedom
Here is how the quantities would be laid out in an analysis of variance table. For the case so far, simple regression, we just have \( K = 1 \). The table is in fact computable by hand, and computing formulas have been inserted.

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Degrees of freedom</th>
<th>Sum of Squares</th>
<th>Mean Squares</th>
<th>( F )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>( \frac{(S_{xy})^2}{S_{xx}} = \hat{\beta}<em>1 S</em>{xx} ) [a]</td>
<td>( \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 )</td>
<td>( \frac{MS_{\text{Regression}}}{MS_{\text{Error}}} ) [d]</td>
</tr>
<tr>
<td>Error</td>
<td>( n - 2 )</td>
<td>( S_{yy} - \frac{(S_{xy})^2}{S_{xx}} )</td>
<td>( \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 ) [b]</td>
<td>( n - 2 )</td>
</tr>
</tbody>
</table>
| Total               | \( n - 1 \)       | \( S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \) \[a\] | \[c\] |}

[a] The ratio \( \frac{SS_{\text{regression}}}{SS_{\text{total}}} \) is known as \( R^2 \), the fraction of variation in \( y \) that is explained by the regression.

[b] The square root of the mean square error is \( s \) (or \( s_e \)), the estimate of the noise standard deviation. It can be written explicitly in simple regression as \( s = \sqrt{\frac{1}{n-2} \left( \frac{S_{xy} - (S_{xy})^2}{S_{xx}} \right)} \).

[c] This box is conventionally left empty. However, if you computed a mean square here, its square root would estimate \( \text{SD}(Y) = \sqrt{\frac{1}{n-1} S_{yy}} \).

[d] The \( F \) statistic tests the null hypothesis \( H_0: \beta_1 = 0 \).
This is the general form, with $K$ predictors.

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Degrees of freedom</th>
<th>Sum of Squares</th>
<th>Mean Squares</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$K$</td>
<td>$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$</td>
<td>$\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{K}$</td>
<td>$\frac{MS_{\text{Regression}}}{MS_{\text{Error}}}$</td>
</tr>
<tr>
<td>Error</td>
<td>$n - 1 - K$</td>
<td>$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$</td>
<td>$\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - 1 - K}$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$n - 1$</td>
<td>$\sum_{i=1}^{n}(y_i - \bar{y})^2$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A measure of quality of the multiple regression is the $F$ statistic. Formally, this $F$ statistic tests

$$H_0: \beta_1 = 0, \beta_2 = 0, \beta_3 = 0, \ldots, \beta_K = 0$$

versus

$$H_1: \text{at least one of } \beta_1, \beta_2, \beta_3, \ldots, \beta_K \text{ is not zero}$$

Note that $\beta_0$ is not involved in this test.

For simple regression ($K = 1$), the test is $H_0: \beta_1 = 0$ versus $H_1: \beta_1 \neq 0$.

Also, note that $s_e = \sqrt{MS_{\text{Error}}}$ is the estimate of $\sigma_e$. This has many names:

- standard error of estimate
- standard error of regression
- estimated noise standard deviation
- root mean square error (RMS error)
- root mean square residual (RMS residual)

The measure called $R^2$ is computed as $\frac{SS_{\text{Regression}}}{SS_{\text{Total}}}$. This is often described as the “fraction of the variation in $Y$ explained by the regression.”
You can show, by the way, that

$$\frac{s_e}{s_y} = \sqrt{\frac{n-1}{n-1-K}} (1-R^2)$$

The quantity $R_{adj}^2 = 1 - \frac{n-1}{n-1-K} (1-R^2)$ is called the adjusted $R$-squared. This is supposed to adjust the value of $R^2$ to account for both the sample size and the number of predictors. With a little simple arithmetic,

$$R_{adj}^2 = 1 - \left(\frac{s_e}{s_y}\right)^2$$

The analysis of variance table for simple regression might look like this:

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Degrees of Freedom</th>
<th>Sum Squares</th>
<th>Mean Squares</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>965.15</td>
<td>965.15</td>
<td>4.1304</td>
</tr>
<tr>
<td>Residual</td>
<td>18</td>
<td>4,206.06</td>
<td>233.67</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>4,871.21</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note that

The Degrees of Freedom column sums to its total.
The Sum Squares column sums to its total.
Each Mean Square is the corresponding Sum Square divided by its Degrees of Freedom.
The $F$ statistic is the ratio of Mean Square (Regression) / Mean Square (Residual).

There are other useful facts that we can derive from this table:

$$s_e = s_{Y|X} = \text{standard error of regression} = \sqrt{\text{MS}_{\text{Resid}}}.$$  
Here this is $s_e = \sqrt{233.67} \approx 15.29$.

$$\text{SD}(Y) = s_Y = \text{standard deviation of the dependent variable} = \sqrt{\frac{\text{SS}_{\text{Total}}}{n-1}}.$$  
Here this is $s_Y = \sqrt{\frac{4,871.21}{19}} \approx \sqrt{256.3795} \approx 16.01$.  

\[ R^2 = \text{percent variation in } Y \text{ explained by regression} = \frac{SS_{\text{Reg}}}{SS_{\text{Total}}} . \]

Here this is \( R^2 = \frac{965.15}{4,871.21} \approx 0.1981 = 19.81\% . \)

\[ R_{\text{adj}}^2 = \text{adjusted } R^2 = 1 - \left( \frac{s_e}{s_Y} \right)^2 . \] This takes some effort, and neither \( s_e \) nor \( s_Y \) is printed in the analysis of variance table.

Here this is \( R_{\text{adj}}^2 = 1 - \left( \frac{15.29}{16.01} \right)^2 \approx 0.0879 = 8.79\% . \)

Since it can happen that \( s_e > s_Y \) (in a really pathetic regression), you will sometimes see \( R_{\text{adj}}^2 < 0 . \)