Analysis of Variance (ANOVA) for Simple Linear Regression

The variability in $Y$ values can be partitioned into two pieces.

\[
\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2
\]

Total Sum of Squares $= \text{Regression Sum of Squares} + \text{Error (or Residual) Sum of Squares}$

\[
\text{SSTO} = \text{SSREG} + \text{SSE}
\]

We can organize the results of a simple linear regression analysis in an ANOVA table.

<table>
<thead>
<tr>
<th>Source</th>
<th>D.F.</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$df_{REG}$</td>
<td>SSREG</td>
<td>MSREG</td>
<td>$\frac{MSREG}{MSE}$</td>
<td>$P(T^2 \geq \frac{MSREG}{MSE})$</td>
</tr>
<tr>
<td>Error</td>
<td>$df_E$</td>
<td>SSE</td>
<td>MSE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$df_{TO}$</td>
<td>SSTO</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\text{Regression} & : 1 & \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 & \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 & \frac{MSREG}{MSE} & P(T^2 \geq \frac{MSREG}{MSE}) & T^2 \sim F(1, n-2) \\
\text{Error} & : n-2 & \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 & \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 & & & \\
\text{Total} & : n-1 & \sum_{i=1}^{n} (Y_i - \bar{Y})^2 & & & \\
\end{align*}
\]

The $F$-statistic $\frac{MSREG}{MSE}$ is used to test

\[ H_0 : \mu\{Y|X\} = \beta_0 \text{ versus } H_A : \mu\{Y|X\} = \beta_0 + \beta_1 X \text{ for some } \beta_1 \neq 0. \]

or $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$ for short. The test is equivalent to the $t$-test that we learned about previously because

\[
(1) \quad F = \frac{MSREG}{MSE} = \frac{\beta_1^2}{[SE(\hat{\beta}_1)]^2} = t^2 \quad \text{and} \quad (2) \quad T^2 \sim F \text{ with } 1 \text{ and } n-2 \text{ d.f.} \quad \iff T \sim t \text{ with } n-2 \text{ d.f.}
\]
The *F*-statistic \( \frac{MSREG}{MSE} \) is a special case of the \( F \)-statistic used to compare full and reduced models.

\[
F = \frac{[RSS(\text{red.}) - RSS(\text{full})]/[df_{RSS(\text{red.})} - df_{RSS(\text{full})}]}{RSS(\text{full})/df_{RSS(\text{full})}}
\]

Recall that our null and alternative hypotheses are

\[
H_0 : \mu \{Y|X\} = \beta_0 \quad \text{versus} \quad H_A : \mu \{Y|X\} = \beta_0 + \beta_1 X \quad \text{for some } \beta_1 \neq 0.
\]

The full model corresponds to the situation where \( \beta_1 \) can be any value. The reduced model forces \( \beta_1 \) to be 0, just like \( H_0 \). Write down formulas for \( RSS(\text{red.}) \), \( RSS(\text{full}) \), \( df_{RSS(\text{red.})} \), and \( df_{RSS(\text{full})} \) for the special case of simple linear regression; and show that the resulting reduced vs. full model \( F \)-statistic is the same as \( F = \frac{MSREG}{MSE} \).

Because \( SSTO = SSREG + SSE \), we may write

\[
1 = \frac{SSREG}{SSTO} + \frac{SSE}{SSTO}.
\]

\( \frac{SSE}{SSTO} \) is the proportion of total variation in the \( Y \) values that was not explained by the regression of \( Y \) on \( X \).

The remaining proportion of variation in the \( Y \) values is

\[
1 - \frac{SSE}{SSTO} = \frac{SSREG}{SSTO}.
\]

This quantity – known as the coefficient of determination – is the proportion of the variation in the \( Y \) values that was explained by the regression of \( Y \) on \( X \).

It can be shown that the coefficient of determination is equal to the square of the sample linear correlation coefficient between \( X \) and \( Y \).

\[
1 - \frac{SSE}{SSTO} = \frac{SSREG}{SSTO} = r^2
\]