Multiple Linear Regression Model

Multiple Linear Regression refers to regression applications in which there are more than one independent variables, \( x_1, x_2, \ldots, x_k \). A multiple linear regression model with \( k \) independent variables has the equation

\[
y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon
\]  

(1)

The \( \varepsilon \) is a random variable with mean 0 and variance \( \sigma^2 \). A prediction equation for this model fitted to data is

\[
\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k
\]  

(2)

where \( \hat{y} \) denotes the “predicted” value computed from the equation, and \( \hat{\beta}_i \) denotes an estimate of \( \beta_i \). These estimates are usually obtained by the method of least squares. This means finding among the set of all possible values for the parameter estimates the ones which minimize the sum of squared residuals, \( \Sigma(y - \hat{y})^2 \). This yields the best fitting equation in terms of minimizing the sum of squared distances of the fitted plane to the data points.

An example of a multiple linear regression with two independent variables is given by the KWH data, but now with \( x_1 = \text{AC} \) and \( x_2 = \text{DRYER} \). Figure 1 shows a plot of KWH versus DRYER.

![Figure 1. Plot of KWH versus DRYER.](image-url)
The plot clearly shows KWH increases with increasing runs of the dryer. The model equation would be

\[ \text{KWH} = \beta_0 + \beta_1 \text{AC} + \beta_2 \text{DRYER} + \epsilon. \]

Least squares parameter estimates are

\[ b_0 = 8.11, \ b_1 = 5.47, \text{ and } b_2 = 13.22. \]

Computation of the estimates by hand is tedious, and infeasible for more than two variables. Estimates are ordinarily obtained using a regression computer program. Standard errors also are usually part of output from a regression program.

The prediction equation is

\[ \text{KWH} = 8.11 + 5.47(\text{AC}) + 13.22(\text{DRYER}). \]

This model ascribes 5.47 KWH to hourly use of the AC and 13.22 KWH to each use of the DRYER, and 8.11 to all other electrical devices.

Compare this prediction equation with the one including only AC in the model,

\[ \text{KWH} = 27.85 + 5.43(\text{AC}). \]

The intercept estimate has changed substantially from 27.85 to 8.11. This change occurs because KWH consumption due to DRYER usage is combined into the intercept estimate in the model that does not contain DRYER.

The estimate of the coefficient on AC has changed very little, from 5.34 to 5.47. This is related to the fact that AC and DRYER usage are relatively uncorrelated. In other words, use of one is not related to use of the other. (See Figure 2.) Generally speaking, if AC and DRYER were positively (negatively) correlated, then the regression coefficient on AC would be reduced (increased) when DRYER was added to the model.
Figure 2. Plot of AC versus DRYER

Compare the values of predicted KWH from the two models. Previously, AC=10 was inserted in the simple linear prediction equation to get

\[ KWH = 27.85 + 5.34(10) = 81.25. \]

A value of DRYER must also be inserted into the multiple regression equation to get a predicted KWH value. Trying DRYER = 0, 1, and 2 gives

\[ KWH = 8.11 + 5.47(10) + 13.22(0) = 62.81, \]
\[ KWH = 8.11 + 5.47(10) + 13.22(1) = 76.03, \]
\[ KWH = 8.11 + 5.47(10) + 13.22(2) = 89.25. \]

An analysis of variance for a multiple linear regression model with k independent variables fitted to a data set with n observations is

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>k</td>
<td>SSR</td>
<td>MSR</td>
</tr>
<tr>
<td>Error</td>
<td>n-k-1</td>
<td>SSE</td>
<td>MSE</td>
</tr>
<tr>
<td>Total</td>
<td>n-1</td>
<td>SSTot</td>
<td></td>
</tr>
</tbody>
</table>

The sums of squares SSR, SSE, and SST have the same definitions in relation to the model as in simple linear regression:
\[
SSR = \sum_{j=1}^{n}(\hat{y}_j - \overline{y})^2, \quad SSE = \sum_{j=1}^{n}(y_j - \hat{y})^2, \quad SSTot = \sum_{j=1}^{n}(y_j - \overline{y})^2 \tag{4}
\]

Also, SSTot=SSR+SSE. The value of SSTot does not change with the model. It depends only on the values of the dependent variable \(y\). But SSE decreases as variables are added to a model, and SSR increases by the same amount. This amount of increase in SSR is the amount of variation due to variables in the larger model that was not accounted for by variables in the smaller model. This increase in regression sum of squares is sometimes denoted

\[
SSR(added \ variables \mid original \ variables), \tag{5}
\]

where original variables represents the list of independent variables that were in the model prior to adding new variables, and added variables represents the list of variables that were added to obtain the new model. The overall SSR for the new model can be partitioned into the variation attributable to the original variables plus the variation due to the added variables that is not due to the original variables,

\[
SSR(all \ variables) = SSR(original \ variables) + SSR(added \ variables \mid original \ variables). \tag{6}
\]

Generally speaking, larger values of the coefficient of determination \(R^2=\frac{SSR}{SST}\) indicate a better fitting model. The value of \(R^2\) must necessarily increase as variables are added to the model. However, this does not necessarily mean that the model has actually been improved. The amount of increase in \(R^2\) can be a mathematical artifact rather than a meaningful indication of an improved model. Sometimes an adjusted \(R^2\) is used to overcome this shortcoming of the usual \(R^2\). Most regression computer programs include both versions of \(R^2\).

The analysis of variance for the two-variable model fitted to the KWH data is

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>9299.8</td>
<td>4649.9</td>
</tr>
<tr>
<td>Error</td>
<td>18</td>
<td>278.8</td>
<td>15.5</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>9578.6</td>
<td></td>
</tr>
</tbody>
</table>

Adding DRYER to the model affected a dramatic change in the value of SSR, which increased from 5609.7 to 9299.8. The value of SSE dropped accordingly from 3968.9 to 278.8. The coefficient of determination is now \(R^2=\frac{9299.8}{9578.6}=0.97\). The two variables, AC and DRYER, account for 97% of the variability in KWH consumption in the house. This is up from \(R^2=\frac{5609.7}{9578.6}=0.58\) for the variable AC alone.

The regression sum of squares partitioned into the amount due to AC alone plus the amount due to DRYER that was not attributable to AC, is
SSR(AC and DRYER) = SSR(AC) + SSR(DRYER|AC),

\[ 9299.8 = 5609.7 + 3690.1. \]

Thus, 3690.1 is the amount of variation due to DRYER that was not accounted for by AC.

Statistical inference about the parameters requires standard errors of the estimates. A 95\% confidence interval for \( \hat{\beta}_i \) is

\[ \hat{\beta}_i \pm t_{df,.025} (\hat{\sigma}_{\hat{\beta}_i}) \]  

(7)

where \( t_{df,.025} \) is the critical value from a \( t \) distribution with \( df=n-k-1 \), the degrees of freedom for error, and \( \hat{\sigma}_{\hat{\beta}_i} \) is the standard error of \( \hat{\beta}_i \).

Standard errors for parameters in the two-variable model are

\[ \hat{\sigma}_{\hat{\beta}_0} = 2.48, \hat{\sigma}_{\hat{\beta}_1} = 0.28, \hat{\sigma}_{\hat{\beta}_2} = 0.86. \]  

(8)

The critical value from a \( t \) distribution with \( df=18 \) is \( t_{18,.025}=2.1 \). Thus, a 95\% confidence interval for \( \beta_1 \) is

\[ \hat{\beta}_1 \pm t_{18,.025} \hat{\sigma}_{\hat{\beta}_1} = 5.47 \pm 2.1(0.28) = 5.47 \pm 0.59. \]

We are 95\% confident that the “true” hourly KWH consumption of the AC is between 4.88 and 6.06. This is a considerably shorter interval than the interval 5.34±2.16 that was obtained from the simple linear regression model because the variance estimate (MSE) has been reduced from 208.9 to 15.5.