Multiple regression

Multiple regression is the obvious generalization of simple regression to the situation where we have more than one predictor. The model is

\[ y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i. \]

The assumptions previously given for simple regression still are required; indeed, simple regression is just a special case of multiple regression, with \( p = 1 \) (this is apparent in some of the formulas given below). The ways of checking the assumptions also remain the same: residuals versus fitted values plot, normal plot of the residuals, time series plot of the residuals (if appropriate), and diagnostics (standardized residuals, leverage values and Cook’s distances, which we haven’t talked about yet). In addition, a plot of the residuals versus each of the predicting variables is a good idea (once again, what is desired is the lack of any apparent structure).

There are a few things that are different for multiple regression, compared to simple regression:

*Interpretation of regression coefficients*

We must be very clear about the interpretation of a multiple regression coefficient. As usual, the constant term \( \hat{\beta}_0 \) is an estimate of the expected value of the target variable when the predictors equal zero (only now there are several predictors). \( \hat{\beta}_j, j = 1, \ldots, p \), represents the estimated expected change in \( y \) associated with a one unit change in \( x_j \) holding all else in the model fixed. Consider the following example. Say we take a sample of college students and determine their College grade point average (\( \text{COLGPA} \)), High school GPA (\( \text{HSGPA} \)), and SAT score (\( \text{SAT} \)). We then build a model of \( \text{COLGPA} \) as a function of \( \text{HSGPA} \) and \( \text{SAT} \):

\[ \text{COLGPA} = 1.3 + .7 \times \text{HSGPA} - .0003 \times \text{SAT}. \]

It is tempting to say (and many people do) that the coefficient for \( \text{SAT} \) has the “wrong sign,” because it says that higher values of SAT are associated with lower values of College GPA. **This is absolutely incorrect!** What it says is that higher values of SAT are associated with lower values of College GPA, holding High school GPA fixed. High school GPA and SAT are no doubt correlated with each other, so changing SAT by one unit holding High school GPA fixed may not ever happen! **The coefficients of a multiple regression must not be interpreted marginally!** If you really are interested in the
relationship between College GPA and just SAT, you should simply do a regression of College GPA on only SAT.

We can see what’s going on here with some simple algebra. Consider the two-predictor regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i.$$ 

The least squares coefficients solve \((X'X)\beta = X'y\). In this case those equations are as follows:

\[
\begin{align*}
n\beta_0 + \left(\sum x_{1i}\right) \beta_1 + \left(\sum x_{2i}\right) \beta_2 &= \sum y_i \\
\left(\sum x_{1i}\right) \beta_0 + \left(\sum x^2_{1i}\right) \beta_1 + \left(\sum x_{1i}x_{2i}\right) \beta_2 &= \sum x_{1i}y_i \\
\left(\sum x_{2i}\right) \beta_0 + \left(\sum x_{1i}x_{2i}\right) \beta_1 + \left(\sum x^2_{2i}\right) \beta_2 &= \sum x_{2i}y_i
\end{align*}
\]

It is apparent that calculation of \(\hat{\beta}_1\) involves the variable \(x_2\); similarly, the calculation of \(\hat{\beta}_2\) involves the variable \(x_1\). That is, the form (and sign) of the regression coefficients depend on the presence or absence of whatever other variables are in the model. In some circumstances, this conditional statement is exactly what we want, and the coefficients can be interpreted directly, but in many situations, the “natural” coefficient refers to a marginal relationship, which the multiple regression coefficients do not address.

One of the most useful aspects of multiple regression is its ability to statistically represent a conditioning action that would otherwise be impossible. In experimental situations, it is common practice to change the setting of one experimental condition while holding others fixed, thereby isolating its effect, but this is not possible with observational data. Multiple regression provides a statistical version of this practice. This is the reasoning behind the use of “control variables” in multiple regression — variables that are not necessarily of direct interest, but ones that the researcher wants to “correct for” in the analysis.

Having said this, we need to recognize that in many situations it is impossible (from a practical point of view) to change one predictor while holding all others fixed. Thus, while we would like to interpret a coefficient as accounting for the presence of other predictors in a physical sense, we have to recognize that technically all we can really guarantee is that the coefficient \(\beta_j\) is the regression of the response, linearly adjusted for the other predictors, on \(x_j\) also linearly adjusted for the other predictors. This is of course much less appealing than “holding all else fixed,” and in practical situations the latter is what we really need, but it is important when dealing with observational data to remember that linear regression is at best only an approximation to what is really going on and what we
really want.

**Hypothesis tests**

There are two types of hypothesis tests of immediate interest:

(a) A test of the overall significance of the regression:

\[ H_0 : \beta_1 = \cdots = \beta_p = 0 \]

versus

\[ H_a : \text{at least one } \beta_j \neq 0, \quad j = 1, \ldots, p \]

The test of these hypotheses is the **F-test**:

\[ F = \frac{\text{Regression MS}}{\text{Residual MS}} = \frac{\text{Regression SS}/p}{\text{Residual SS}/(n - p - 1)}. \]

This is compared to a critical value for an F-distribution on \((p, n - p - 1)\) degrees of freedom.

(b) Tests of the significance of an individual coefficient:

\[ H_0 : \beta_j = 0, \quad j = 0, \ldots, p \]

versus

\[ H_a : \beta_j \neq 0 \]

This is tested using a **t-test**:

\[ t_j = \frac{\hat{\beta}_j}{\text{s.e.}(\hat{\beta}_j)}, \]

which is compared to a critical value for a \(t\)-distribution on \(n - p - 1\) degrees of freedom. Of course, other values of \(\beta_j\) can be specified in the null hypothesis (say \(\beta_j^0\)), with the \(t\)-statistic becoming

\[ t_j = \frac{\hat{\beta}_j - \beta_j^0}{\text{s.e.}(\hat{\beta}_j)}. \]

**Proportion of variability accounted for by the regression**

As before, the \(R^2\) estimates the proportion of variability in the target variable accounted for by the regression. Also as before, the \(R^2\) equals

\[ R^2 = 1 - \frac{\text{Residual SS}}{\text{Total SS}}. \]

The adjusted \(R^2\) is different, however:

\[ R_a^2 = R^2 - \frac{p}{n - p - 1} (1 - R^2) \]
Estimation of \( \sigma^2 \)

As was the case in simple regression, the variance of the errors \( \sigma^2 \) is estimated using the residual mean square. The difference is that now the degrees of freedom for the residual sum of squares is \( n - p - 1 \), rather than \( n - 2 \), so the residual mean square has the form

\[
\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - p - 1}.
\]

Multicollinearity

A issue related to the interpretation of regression coefficients is that of multicollinearity. When predicting \((x)\) variables are highly correlated with each other, this can lead to instability in the regression coefficients, and the \(t\)-statistics for the variables can be deflated. From a practical point of view, this can lead to two problems:

1. If one value of one of the \(x\)-variables is changed only slightly, the fitted regression coefficients can change dramatically.
2. It can happen that the overall \(F\)-statistic is significant, yet each of the individual \(t\)-statistics is not significant. Another indication of this problem is that the \(p\)-value for the \(F\) test is considerably smaller than those of any of the individual coefficient \(t\)-tests.

One problem that multicollinearity does not cause to any serious degree is inflation or deflation of overall measures of fit (\(R^2\)), since adding unneeded variables cannot reduce \(R^2\) (it can only leave it roughly the same).

Another problem with multicollinearity comes from attempting to use the regression model for prediction. In general, simple models tend to forecast better than more complex ones, since they make fewer assumptions about what the future must look like. That is, if a model exhibiting collinearity is used for prediction in the future, the implicit assumption is that the relationships among the predicting variables, as well as their relationship with the target variable, remain the same in the future. This is less likely to be true if the predicting variables are collinear.

How can we diagnose multicollinearity? We can get some guidance by looking again at a two-predictor model:

\[
y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i.
\]

It can be shown that in this case

\[
var(\hat{\beta}_1) = \sigma^2 \left[ \sum x_{1i}^2 (1 - r_{12}^2) \right]^{-1}
\]
and

\[ \text{var}(\hat{\beta}_2) = \sigma^2 \left[ \sum x_{2i}^2 (1 - r_{12}^2) \right]^{-1}, \]

where \( r_{12} \) is the correlation between \( x_1 \) and \( x_2 \). Note that as collinearity increases (\( r_{12} \to \pm 1 \)), both variances tend to \( \infty \). This effect can be quantified as follows:

<table>
<thead>
<tr>
<th>( r_{12} )</th>
<th>Ratio of ( \text{var}(\hat{\beta}<em>1) ) to that if ( r</em>{12} = 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.50</td>
<td>1.33</td>
</tr>
<tr>
<td>0.70</td>
<td>1.96</td>
</tr>
<tr>
<td>0.80</td>
<td>2.78</td>
</tr>
<tr>
<td>0.90</td>
<td>5.26</td>
</tr>
<tr>
<td>0.95</td>
<td>10.26</td>
</tr>
<tr>
<td>0.97</td>
<td>16.92</td>
</tr>
<tr>
<td>0.99</td>
<td>50.25</td>
</tr>
<tr>
<td>0.995</td>
<td>100.00</td>
</tr>
<tr>
<td>0.999</td>
<td>500.00</td>
</tr>
</tbody>
</table>

This ratio describes by how much the variance of the estimated coefficient is inflated due to observed collinearity relative to when the predictors are uncorrelated.

A diagnostic to determine this in general is the variance inflation factor (VIF) for each predicting variable, which is defined as

\[ \text{VIF}_j = \frac{1}{1 - R^2_j}, \]

where \( R^2_j \) is the \( R^2 \) of the regression of the variable \( x_j \) on the other predicting variables. The VIF gives the proportional increase in the variance of \( \hat{\beta}_j \) compared to what it would have been if the predicting variables had been completely uncorrelated. Minitab supplies these values under Options for a multiple regression fit. How big a VIF indicates a problem? A good guideline is that values satisfying

\[ \text{VIF} < \max \left( 10, \frac{1}{1 - R^2_{\text{model}}} \right), \]

where \( R^2_{\text{model}} \) is the usual \( R^2 \) for the regression fit, mean that either the predictors are more related to the target variable than they are to each other, or they are not related to each other very much. In these circumstances coefficient estimates are not very likely to be very unstable, so collinearity is not a problem.
What can we do about multicollinearity? The simplest solution is to simply drop out any collinear variables; so, if High school GPA and SAT are highly correlated, you don’t need to have to both in the model, so use only one. Note, however, that this advice is only a general guideline — sometimes two (or more) collinear predictors are needed in order to adequately model the target variable.

Linear restrictions and hypothesis tests

It is sometimes the case that we believe that a simpler version of the full model (a subset model) might be adequate to fit the data. For example, say we take a sample of college students and determine their College grade point average (GPA), SAT reading score (Reading) and SAT math score (Math). The full regression model to fit to these data is

\[ \text{GPA}_i = \beta_0 + \beta_1 \text{Reading}_i + \beta_2 \text{Math}_i + \varepsilon_i. \]

However, we might very well wonder if all that really matters in prediction of GPA is the total SAT score — that is, Reading + Math. This subset model is

\[ \text{GPA}_i = \gamma_0 + \gamma_1 (\text{Reading} + \text{Math})_i + \varepsilon_i \]

with \( \beta_1 = \beta_2 \equiv \gamma_1 \). This equality condition is called a linear restriction, because it defines a linear condition on the parameters of the regression model (that is, it only involves additions, subtractions and equalities).

We can now state our question about whether the total SAT score is all that is needed as a hypothesis test about this linear restriction. As always, the null hypothesis is what we believe unless convinced otherwise; in this case, that is the simpler (subset) model that the sum of Reading and Math is adequate, since it says that only one predictor is needed, rather than two. The alternative hypothesis is simply the full model (with no conditions on \( \beta \)). That is,

\[ H_0 : \beta_1 = \beta_2 \]

versus

\[ H_a : \beta_1 \neq \beta_2. \]

These hypotheses are tested using a partial F-test. The F-statistic has the form

\[ F = \frac{(\text{Residual SS}_{\text{subset}} - \text{Residual SS}_{\text{full}})/d}{\text{Residual SS}_{\text{full}}/(n - p - 1)}, \]
where \( n \) is the sample size, \( p \) is the number of predictors in the full model, and \( d \) is the difference between the number of parameters in the full model and the number of parameters in the subset model. Some packages (such as SAS and Systat) allow the analyst to specify a linear restriction to test when fitting the full model, and will provide the appropriate \( F \)-statistic automatically. To calculate the statistic using other packages, the appropriate regressions have to be run manually. For the GPA/SAT example, a regression on Reading and Math would provide Residual \( SS_{\text{full}} \). Creating the variable TotalSAT = Reading + Math, and then doing a regression of GPA on TotalSAT, would provide Residual \( SS_{\text{subset}} \).

This statistic is compared to an \( F \) distribution on \((d, n - p - 1)\) degrees of freedom. So, for example, for the GPA/SAT example, \( p = 2 \) and \( d = 3 - 2 = 1 \), so the observed \( F \)-statistic would be compared to an \( F \) distribution on \((1, n - 3)\) degrees of freedom. The tail probability of the test can be determined, for example, using Minitab.

An alternative form for the \( F \)-test above might make a little clearer what’s going on:

\[
F = \frac{(R^2_{\text{full}} - R^2_{\text{subset}})/d}{(1 - R^2_{\text{full}})/(n - p - 1)}.
\]

That is, if the \( R^2 \) of the full model isn’t much larger than the \( R^2 \) of the subset model, the \( F \)-statistic is small, and we do not reject using the subset model; if, on the other hand, the difference in \( R^2 \) values is large, we do reject the subset model in favor of the full model.

Note, by the way, that the only reason to even look at linear restriction hypotheses is if they make sense from a physical point of view; you shouldn’t do them for hypotheses that are not of intrinsic interest. Note also that the \( F \)-statistic to test the overall significance of the regression is a special case of this construction (with restriction \( \beta_1 = \cdots = \beta_p = 0 \)), as are the individual \( t \)-statistics that test the significance of any variable (with restriction \( \beta_j = 0 \), and then \( F_j = t^2_j \)).