Lecture 15: Introduction to Multiple Linear Regression

In multiple linear regression, a linear combination of two or more predictor variables is used to explain the variation in a response. In essence, the additional predictors are used to explain the variation in the response not explained by a simple linear regression fit.

As an illustration, I will consider the following problem. Anthropologists conducted a study to determine the long-term effects of an environmental change on systolic blood pressure. They measured the blood pressure and several other characteristics (weight, age, years since migration, pulse rate, skin fold measures) of 39 Indians who migrated from a very primitive environment high in the Andes into the mainstream of Peruvian society at a lower altitude. All of the Indians were males at least 21 years of age, and were born at a high altitude.

A question we consider concerns the long term effects of an environmental change on the systolic blood pressure. In particular, is there a relationship between the systolic blood pressure and how long the Indians lived in their new environment as measured by the fraction of their life spent in the new environment? (fraction = years since migration/age)

A plot of systolic blood pressure against fraction (see the scatterplot matrix in the JMP-IN output at the end) suggests a weak linear relationship. Nonetheless, consider fitting the regression model

\[ \text{sys bp} = \beta_0 + \beta_1 \text{ fraction} + \epsilon. \]

The least squares line is given by

\[ \hat{\text{sys bp}} = 133.49 - 15.75 \text{ fraction}, \]

and suggests that average systolic blood pressure decreases as the fraction of life spent in modern society increases. However, the t-test of \( H_0 : \beta_1 = 0 \) is not significant at the 5% level (p-value=.089). That is, the weak linear relationship observed in the data is not atypical of a population where there is no linear relationship between systolic blood pressure and the fraction of life spent in a modern society.

Even if this test were significant, the small value of \( R^2 = .076 \) suggests that fraction does not explain a substantial amount of the variation in the systolic blood pressures. If we omit
the individual with the highest blood pressure (see the plot) then the relationship would be weaker.

**Taking Weight into Consideration**

At best, there is a weak relationship between systolic blood pressure and fraction. However, it is usually accepted that systolic blood pressure and weight are related; see the scatterplot matrix for confirmation. A natural way to take weight into consideration is to include weight and fraction as predictors of systolic blood pressure in the multiple regression model:

\[
sys\ bp = \beta_0 + \beta_1 \ text{fraction} + \beta_2 \text{weight} + \epsilon.
\]

As in simple linear regression, the model is written in the form:

\[
\text{Response} = \text{Mean of Response} + \text{Residual},
\]

so the model implies that that average systolic blood pressure is a linear combination of fraction and weight. As in simple linear regression, the standard multiple regression analysis assumes that the responses are normally distributed with a constant variance \( \sigma^2_{Y|X} \). The parameters of the regression model \( \beta_0, \beta_1, \beta_2 \) and \( \sigma^2_{Y|X} \) are estimated by LS.

**JMP-IN** output for fitting the multiple regression model is given at the end of the handout. I omitted some of the default plots and output. Kyle will show you in this week’s lab how to fit multiple regression models using the **FIT MODEL** platform in **JMP-IN**.

**Important Points to Notice About the Regression Output**

1. (Parameter Estimates Tables) The LS estimates of the intercept and the regression coefficient for fraction, and their standard errors, change from the simple linear model to the multiple regression model. For the simple linear regression

\[
\widehat{sysbp} = 133.50 - 15.75 \text{fraction}.
\]

For the multiple regression model

\[
\widehat{sys\ bp} = 60.89 - 26.76 \text{fraction} + 1.21 \text{weight}.
\]
2. (ANOVA Tables) Comparing the simple linear regression and the multiple regression models we see that the Regression (model) \( df \) has increased to 2 from 1 (2=number of predictor variables) and the Residual (error) \( df \) has decreased from 37 to 36 (\( =n−1 \)- number of predictors). Adding a predictor increases the Regression \( df \) by 1 and decreases the Residual \( df \) by 1.

3. (ANOVA Tables) The Residual SS decreases by \( 6033.37 - 3441.36 = 2592.01 \) upon adding the weight term. The Regression SS increased by 2592.01 upon adding the weight term term to the model. The Total SS does not depend on the number of predictors so it stays the same. The Residual SS, or the part of the variation in the response unexplained by the regression model never increases when new predictors are added.

4. (Summary of Fit Tables) The proportion of variation in the response explained by the regression model:

\[
R^2 = \frac{\text{Regression SS}}{\text{Total SS}}
\]

never decreases when new predictors are added to a model. The \( R^2 \) for the simple linear regression was .076, whereas \( R^2 = .473 \) for the multiple regression model. Adding the weight variable to the model increases \( R^2 \) by 40%. That is, weight explains 40% of the variation in systolic blood pressure not already explained by fraction.

5. (ANOVA tables) The estimated variability about the regression line

\[
\text{Residual MS} = s^2_{Y|X}
\]

decreased dramatically after adding the weight effect. For the simple linear regression model \( s^2_{Y|X} = 163.06 \), whereas \( s^2_{Y|X} = 95.59 \) for the multiple regression model. This suggests that an important predictor has been added to model.

6. (ANOVA table) The F-statistic for the multiple regression model

\[
F_{obs} = \frac{\text{Regression MS}}{\text{Residual MS}} = 16.163
\]
(which is compared to a F-table with 2 and 36 df) tests $H_0 : \beta_1 = \beta_2 = 0$ against $H_A : \text{ not } H_0$. This is a test of no relationship between the average systolic blood pressure and fraction and weight, assuming the relationship is linear. If this test is significant than either fraction or weight, or both, are important for explaining the variation in systolic blood pressure.

7. (Parameter Estimates Table) Given the model

$$sys \, bp = \beta_0 + \beta_1 \, fraction + \beta_2 \, weight + \epsilon,$$

one interest is testing $H_0 : \beta_2 = 0$ against $H_A : \beta_2 \neq 0$. The $t$-statistic for this test

$$t_{obs} = \frac{b_2 - 0}{SE(b_2)} = \frac{1.217}{.233} = 5.207$$

is compared to a $t$-critical value with Residual $df = 36$. **JMP-IN** gives a p-value of .0001, which suggests $\beta_2 \neq 0$. The $t$-test of $H_0 : \beta_2 = 0$ in the multiple regression model tests whether adding weight to the simple linear regression model explains a significant part of the variation in systolic blood pressure not explained by fraction. In some sense, the $t$-test of $H_0 : \beta_2 = 0$ will be significant if the increase in $R^2$ (or decrease in Residual SS) obtained by adding weight to this simple linear regression model is substantial. We saw a big increase in $R^2$, which is deemed significant by the $t$-test. A similar interpretation is given to the $t$-test for $H_0 : \beta_1 = 0$.

8. The $t$-tests for $\beta_0 = 0$ and $\beta_1 = 0$ are conducted, assessed, and interpreted in the same manner. The p-value for testing $H_0 : \beta_0 = 0$ is .0001, whereas the p-value for testing $H_0 : \beta_1 = 0$ is .0007. This implies that fraction is important in explaining the variation in systolic blood pressure after weight is taken into consideration (by including weight in the model as a predictor).

9. (Effect Test Table) This table gives F-tests for $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$. The SS for each effect (weight and fraction) is the reduction in the Residual SS achieved when the given effect is added last to the model. The F-statistic is the MS effect divided by the Res MS from the multiple regression model. The p-value is obtained from an F-table.
with numerator df of 1 (because each effect involves estimating one parameter—more
on this later) and denominator df equal to the Res df from the ANOVA table. The
p-values for the F-tests agree with the p-values from the t-tests described above.

10. We compute CIs for the regression parameters in the usual way: \( b_i + t_{crit}SE(b_i) \), where
\( t_{crit} \) is the \( t \)−critical value for the corresponding CI level with \( df = \) Residual \( df \).

Understanding the Model

The \( t \)−test for \( H_0 : \beta_1 = 0 \) is highly significant (p-value=.0007), which implies that fraction
is important in explaining the variation in systolic blood pressure after weight is taken
into consideration  (by including weight in the model as a predictor). Weight is called a
suppressor variable. Ignoring weight suppresses the relationship between systolic blood
pressure and fraction.

The implications of this analysis are enormous! Essentially, the correlation between a
predictor and a response says very little about the importance of the predictor in a regression
model with one or more additional predictors. This conclusion also holds in situations where
the correlation is high, in the sense that a predictor that is highly correlated with the response
may be unimportant in a multiple regression model once other predictors are included in the
model.

Another issue that I wish to address concerns the interpretation of the regression coef-
ficients in a multiple regression model. For our problem, let us first focus on the fraction
coefficient in the fitted model

\[
\text{sys bp} = 60.89 - 26.76 \text{ fraction} + 1.21 \text{ weight}.
\]

The negative coefficient indicates that the predicted systolic blood pressure decreases as
fraction increases holding weight constant. In particular, the predicted systolic blood
pressure decreases by 26.76 for each unit increase in fraction, holding weight constant at any
value. Similarly, the predicted systolic blood pressure increases by 1.21 for each unit increase
in weight, holding fraction constant at any level.
Another Multiple Regression Example

The data below are selected from a larger collection of data referring to candidates for the General Certificate of Education (GCE) who were being considered for a special award. Here, \( Y \) denotes the candidate’s TOTAL mark, out of 1000, in the GCE exam, while \( X_1 \) is the candidate’s score in the compulsory part of the exam (COMP), which has a maximum score of 200 of the 1000 points on the exam. \( X_2 \) denotes the candidates’ score, out of 100, in a School Certificate English Language (SCEL) paper taken on a previous occasion.

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>476</td>
<td>111</td>
<td>68</td>
</tr>
<tr>
<td>2</td>
<td>457</td>
<td>92</td>
<td>46</td>
</tr>
<tr>
<td>3</td>
<td>540</td>
<td>90</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>551</td>
<td>107</td>
<td>59</td>
</tr>
<tr>
<td>5</td>
<td>575</td>
<td>98</td>
<td>50</td>
</tr>
<tr>
<td>6</td>
<td>698</td>
<td>150</td>
<td>66</td>
</tr>
<tr>
<td>7</td>
<td>574</td>
<td>118</td>
<td>54</td>
</tr>
<tr>
<td>8</td>
<td>645</td>
<td>117</td>
<td>59</td>
</tr>
<tr>
<td>9</td>
<td>690</td>
<td>114</td>
<td>80</td>
</tr>
<tr>
<td>10</td>
<td>634</td>
<td>130</td>
<td>57</td>
</tr>
<tr>
<td>11</td>
<td>537</td>
<td>118</td>
<td>51</td>
</tr>
<tr>
<td>12</td>
<td>390</td>
<td>91</td>
<td>44</td>
</tr>
<tr>
<td>13</td>
<td>562</td>
<td>118</td>
<td>61</td>
</tr>
<tr>
<td>14</td>
<td>560</td>
<td>109</td>
<td>66</td>
</tr>
</tbody>
</table>

A goal here is to compute a multiple regression of the TOTAL score \( Y \) on COMP (\( X_1 \)) and SCEL (\( X_2 \)), and make the necessary tests to enable you to comment intelligently on the extent to which current performance in the compulsory test (COMP) may be used to predict aggregate TOTAL performance on the GCE exam, and on whether previous performance in the School Certificate English Language (SCEL) has any predictive value independently of what has already emerged from the current performance in the compulsory papers.

I will lead you through a number of steps to help you answer this question. Let us answer the following straightforward questions based on the JMP-IN output.

1. Plot TOTAL against COMP and SCEL individually, and comment on the form (i.e. linear, non-linear, logarithmic, etc.), strength, and direction of the relationships.

2. Plot COMP against SCEL and comment on the form, strength, and direction of the relationship.
3. Compute the correlation between all pairs of variables. Do the correlation values appear reasonable, given the plots?

In parts 4 through 9, ignore the possibility that TOTAL, COMP or SCEL might ideally need to be transformed.

4. Which of COMP and SCEL explains a larger proportion of the variation in TOTAL? Which would appear to be a better predictor of TOTAL? (Explain).

5. Consider 2 simple linear regression models for predicting TOTAL one with COMP as a predictor, and the other with SCEL as the predictor. Do COMP and SCEL individually appear to be important for explaining the variation in TOTAL (i.e. test that the slopes of the regression lines are zero). Which, if any, of the output, support, or contradicts, your answer to the previous question?

6. Fit the multiple regression model

\[ \text{TOTAL} = \beta_0 + \beta_1 \text{COMP} + \beta_2 \text{SCEL} + \epsilon. \]

Test \( H_0 : \beta_1 = \beta_2 = 0 \) at the 5% level. Describe in words what this test is doing, and what the results mean here.

7. In the multiple regression model, test \( H_0 : \beta_1 = 0 \) and \( H_0 : \beta_2 = 0 \) individually. Describe in words what these tests are doing, and what the results mean here.

8. How does the \( R^2 \) from the multiple regression model compare to the \( R^2 \) from the individual simple linear regressions? Is what you are seeing here appear reasonable, given the tests on the individual coefficients?

9. Do your best to answer the question posed above, in the paragraph on page 1 that begins “A goal .... ”. Provide an equation (LS) for predicting TOTAL.
Comments on the GCE Analysis

I will give you my thoughts on these data, and how I would attack this problem, keeping the ultimate goal in mind. As a first step, I plot the data and check whether transformations are needed. The plot of TOTAL against COMP is fairly linear, but the trend in the plot of TOTAL against SCEL is less clear. You might see a non-linear trend here, but the relationship is not very strong. When I assess plots I try to not allow a few observations affect my perception of trend, and with this in mind, I do not see any strong evidence at this point to transform any of the variables.

One difficulty that we must face when building a multiple regression model is that these two-dimensional (2D) plots of a response against individual predictors may have little information about the appropriate scales for a multiple regression analysis. In particular, the 2D plots only tell us whether we need to transform the data in a simple linear regression analysis. If a 2D plot shows a strong non-linear trend, I would do an analysis using the suggested transformations, including any other effects that are important. However, it might be that no variables need to be transformed in the multiple regression model.

Although SCEL appears to be useful as a predictor of TOTAL on its own, the multiple regression output indicates that SCEL does not explain a significant amount of the variation in TOTAL, once the effect of COMP has been taken into account. In particular, the SCEL ($X_2$) effect in the multiple regression model is far from significant (p-value=.30). Hence, previous performance in the SCEL exam ($X_2$) has little predictive value independently of what has already emerged from the current performance in the compulsory papers ($X_1$ or COMP).

What are my conclusions? Given that SCEL is not a useful predictor in the multiple regression model, I would propose a simple linear regression model to predict TOTAL from COMP:

$$\text{Predicted TOTAL} = 128.55 + 3.95\text{COMP}.$$ 

Output from the fitted model was given earlier. A residual analysis of the model showed no serious deficiencies.