# 12 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.

It can be a lot more interesting than that sounds, however, since predictors can operate much differently together than alone. Fitting multiple predictors *adjusts* the estimated effects of a predictor for the other predictors. An apparently important predictor can have little effect if adjusted for other variables, or an apparently insignificant predictor can appear very important after adjusting for other variables. The upshot is that the simple linear regression models we worked with last week are inadequate in many circumstances (though they are the basis for what we will see in this section).

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 in the separate Minitab output) suggests at best 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

$$\widehat{\text{sys bp}} = 133 - 15.8 \text{ fraction},$$

and suggests that average systolic blood pressure decreases as the fraction of life spent in modern society increases (if half of life is spent in modern society then that should account for almost an 8 point drop in systolic blood pressure). 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 (and if we constructed a 95% CI for $\beta_1$ it would contain 0).

Even if this test were significant, the small value of $R^2 = 7.6\%$ 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 still. Minitab identifies two unusual observations, one of which is that individual. We should identify the other one too, and see if the fit changes much with those two observations removed. I will show you how to do that later in this section, and Erik will help you with it in lab.

### Taking Weight into Consideration

At best, there is a weak relationship between systolic blood pressure and fraction alone. 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 both as predictors of systolic blood pressure in the multiple regression model:

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

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.

Minitab output from fitting the multiple regression model is given at the end of the handout.

### Important Points to Notice About the Regression Output

1. (Parameter Estimates ) 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

   $$\widehat{\text{sys bp}} = 133 - 15.8 \text{ fraction}.$$

   For the multiple regression model

   $$\widehat{\text{sys bp}} = 60.9 - 26.8 \text{ fraction} + 1.22 \text{ weight}.$$

2. (ANOVA Table) 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 Table) The Residual SS decreases by 6033.4 - 3441.4 = 2592.0 upon adding the weight term. The Regression SS increased by 2592.0 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. The proportion of variation in the response explained by the regression model:

   $$R^2 = \text{Regression SS / Total SS}$$

never decreases when new predictors are added to a model. The $R^2$ for the simple linear regression with fraction as the only predictor was 7.6%, whereas $R^2 = 47.3\%$ for the multiple regression model. Adding the weight variable to the model increases $R^2$ by 40%. That is, weight explains an additional 40% of the variation in systolic blood pressure not already explained by fraction alone.

This is not as simple as it sounds. If you use weight as the only predictor, then $R^2 = 27.2\%$ (confirm this), so adding fraction to that model gives $R^2 = 47.3\%$ and fraction can be seen as explaining an additional 20% of variation in systolic blood pressure not already explained by weight alone. This is far more than explained by fraction alone! The two variables together act in a much more interesting manner than a simple sum of individual behavior.

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

$$\sqrt{\text{Residual MS}} = s_{Y|X}$$

decreased dramatically after adding the weight effect. For the simple linear regression model (fraction alone) $s_{Y|X} = 12.77$, whereas $s_{Y|X} = 9.78$ 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

$$F_{obs} = \text{Regression MS / Residual MS} = 16.16$$

(which is compared to a F-table with 2 and 36 *df*) tests $H_0 : \beta_1 = \beta_2 = 0$ against $H_A :$ 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. The p-value for this test is 0.

7. (Parameter Estimates Table) Given the model

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

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

$$t_{obs} = \frac{b_2 - 0}{SE(b_2)} = \frac{1.2169}{.2337} = 5.21$$

is compared to a $t-$critical value with Residual $df = 36$. Minitab gives a p-value of .000, 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 .000, whereas the p-value for testing $H_0 : \beta_1 = 0$ is .001. 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. (Seq SS Table) I wish Minitab gave you p-values here. The t-tests above do not depend upon what order terms are entered into the model. This table shows how much of the total SS is accounted for in a sequential manner. If fraction alone is entered into the model, the regression SS is 498.1 (also seen in the simple linear regression model). If weight then is added, an additional 2592.0 is accounted for (as calculated above). These also are known as SAS Type I SS. An F-test is constructed by dividing the sequential SS by residual MS (MSE = 95.96), and the critical value for each test is an $F$ with numerator degrees of freedom 1 and denominator degrees of freedom 36 (residual df). Minitab gives you a value of $F_{crit} = 4.11317$. Those test statistics are $F_s = \frac{498.1}{95.6} = 5.033$ and $F_s = \frac{2592.0}{95.6} = 27.11$ for fraction and weight, respectively. Both are significant. It seems contradictory that fraction is significant here but not in the simple linear regression. The difference is that MSE is much smaller in the 2-predictor model.

10. We could 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 $df$.

11. Residual plots show no striking problems.

12. Minitab does flag 3 unusual observations (outliers in some sense).

### 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 coefficients in a multiple regression model. For our problem, let us first focus on the fraction coefficient in the fitted model

$$\widehat{\text{sys bp}} = 60.89 - 26.76 \text{ fraction} + 1.21 \text{ weight}.$$

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.

### Considering the effect of unusual observations

Minitab has warned us about a large residual and two influential values, observations 1, 8, and 39. We should make certain that these few values are not driving our entire analysis. I created a new variable named flag, where I assigned a value of 0 to those three observations and a value of 1 to the others. Next I constructed a matrix plot with groups where flag is the group variable. This plots the three special points separately. You can see from the plots that these correspond to the most extreme values of weight and systol.

Under regression options you can enter a variable of weights. I constructed flag with this in mind. A weight of 0 throws the observation out of the analysis, while a weight of 1 keeps it in normally. There other possibilities we will not consider this semester. Using flag as the weight variable throws those three unusual observations out of the analysis, so we can see how much effect they really have.

The values of coefficients are a little different with these values out of the analysis, with an indication that both fraction and weight have smaller effects (though still significant) than before. Most striking is the drop in $R^2$ from 47.3% to 27.3%. The effects of both fraction and weight are still pronounced, but the three observations are making a big difference in total variability.

### 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, Total denotes the candidate's Total mark, out of 1000, in the GCE exam, while Comp is the candidate's score in the compulsory part of the exam, which has a maximum score of 200 of the 1000 points on the exam. SCEL denotes the candidate's score, out of 100, in a School Certificate English Language paper taken on a previous occasion.

```
Data Display
Row  Total  Comp  SCEL
  1    476   111    68
  2    457    92    46
  3    540    90    50
  4    551   107    59
  5    575    98    50
  6    698   150    66
  7    545   118    54
  8    574   110    51
  9    645   117    59
 10    690   114    80
 11    634   130    57
 12    637   118    51
 13    390    91    44
 14    562   118    61
 15    560   109    66
```

A goal here is to compute a multiple regression of the Total score on Comp and SCEL, 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. You also want to know 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 Minitab 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 117 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 effect in the multiple regression model is far from significant (p-value=.30). Hence, previous performance in the SCEL exam has little predictive value independently of what has already emerged from the current performance in the compulsory papers (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:

$$\text{Predicted Total} = 128.5 + 3.95\text{Comp}.$$

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