**INTRODUCTION TO LINEAR OR MULTIPLE REGRESSION**

**by Simon Moss**

|  |
| --- |
| **Introduction** |

 Linear regression, sometimes called multiple regression or ordinary least squares, is one of the most common statistical tests. Linear regression can be used in most circumstances—although is not always the most accurate or suitable option. In essence, linear regression is used in two circumstances

* to examine whether one set of variables, such as age, gender, and IQ, predicts or are related to some numerical outcome, such as motivation in PhD candidates
* to explore whether two numerical variables are related to each other, such as IQ and motivation in PhD candidates, after controlling other variables

|  |
| --- |
| **A simple example** |

**Example**

 To introduce you to linear regression, consider this example. Suppose you want to predict which research candidates are likely to be especially motivated. To investigate this topic, a researcher administers a survey to 500 research candidates. This survey includes questions that assess

* motivation, such as “On a scale of 1 to 10, how motivated do you feel”
* self-esteem, such as “On a scale of 1 to 10, to what extent do you feel proud of who you are”
* and IQ, such as “On a scale of 1 to 10, how intelligent do you feel you are”

An extract of the data appears in the following screen. Like most data files, each row corresponds to one person. Each column corresponds to a separate characteristic, called a variable. In the column called gender, 0 represents females, and 1 represents males.



Linear regression can be utilised to examine whether

* self-esteem, IQ, age, and sex predicts, or is associated with, the motivation of research candidates
* self-esteem is related to motivation after controlling IQ, age, and sex
* these aims will become clearer as you read.

Many software packages can be utilized to conduct linear regression. This example utilises SPSS. If you use another package, such as R or Stata, perhaps follow these examples anyway. Later, this document clarifies how to conduct linear regression in R and Stata. In SPSS, to generate the following screen, select the “Analyse” menu, and choose “Regression” and then “Linear”.



* Designate “Motivation” as the “Dependent” variable. That is, select “Motivation” and then press the top arrow. In regression, the dependent variable is sometimes called the outcome or criterion
* Designate “Self-esteem”, “IQ”, “Age”, and “Sex” as the “Independent” variables. In regression, the independent variables are sometimes called the predictors.
* Press Save and then tick “Unstandardized” Predicted Values and “Unstandardized” Residuals—the two top boxes.
* Press Continue and then OK. Here is an extract of the data.

|  |
| --- |
| **Coefficientsa** |
| Model | Unstandardized Coefficients | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
| 1 | (Constant) | 8.580 | 6.146 |  | 1.396 | .183 |
| Self\_esteem | .548 | .169 | .600 | 3.233 | .006 |
| IQ | -.069 | .060 | -.212 | -1.160 | .264 |
| Age | .001 | .038 | .007 | .039 | .969 |
| Gender | .958 | .855 | .214 | 1.120 | .280 |
| a. Dependent Variable: Motivation |

**Interpret the output**

 The key table is called “Coefficients”. To utilize this table, first interpret the p values. Specifically

* proceed to the column called “Sig”—a column that represents the p values
* in this example, the p value associated with self-esteem is less than .05 and thus significant
* consequently, we conclude that self-esteem is related to motivation after controlling IQ, age, and gender
* in contrast, the p value associated with IQ exceeds .05 and is thus not significant
* consequently, we conclude that IQ is not significantly related to motivation after controlling self-esteem, age, and gender
* these principles will be clarified later.

However, significance or p values do not clarify whether the association between self-esteem and motivation is positive or negative. Does self-esteem enhance motivation or diminish motivation? To answer this question

* proceed to the column called “B”—a column that represents something called B coefficients
* in this example, the B coefficient is positive
* consequently, we conclude that self-esteem is positively related to motivation after controlling IQ, age, and sex

**Generate an equation**

 This example shows how linear regression can be utilized to explore whether some predictor is related to some outcome after controlling other variables. In addition, linear regression can be used to predict some outcome, such as motivation, from an equation or formula. In particular, to construct this equation

* multiply each value in the B column by the corresponding predictor—and then sum these answers
* in this example, the equation is “Motivation = 8.580 + .548 x self-esteem - .069 x IQ + 0.001 x Age + 0.958 x Gender”
* as this example shows, the word “Constant” can be omitted from the equation

To illustrate the benefits of this equation

* suppose a person arrived with a self-esteem of 7, and IQ of 110, an age of 25, and a gender of 1, representing males
* you would then substitute these values in the formula
* in particular, motivation would equal 8.580 + .548 x 7 - .069 x 110 + 0.001 x 25 + 0.958 x 1 or 5.809
* consequently, you predict the motivation of this person is 5.809

|  |
| --- |
| **Controlling variables** |

**Spurious variables**

 The previous section showed that self-esteem is positively associated with motivation after controlling IQ, age, and gender. So, linear regression can be utilised to explore associations after controlling other variables. But, what does controlling variables actually mean? And, why would you want to control variables. To illustrate, consider the following table, in which each row represents one person.

|  |
| --- |
| Data from this study |
| Age | Self-esteem out of 10 | Motivation out of 10 |
| 21 | 3 | 5 |
| 23 | 4 | 3 |
| 21 | 3 | 4 |
| 24 | 5 | 2 |
| 20 | 3 | 5 |
| 24 | 2 | 3 |
| 49 | 7 | 8 |
| 52 | 8 | 9 |
| 47 | 9 | 7 |
| 51 | 8 | 8 |
| 46 | 7 | 8 |
| 52 | 9 | 9 |

 This table generates some interesting conclusions. If you scan the last two columns, you will conclude that self-esteem seems to be positively associated with motivation. That is, high scores on self-esteem—the final six rows—tend to coincide with high scores on motivation. Low scores on self-esteem tend to coincide with low scores on motivation. And yet, another explanation is possible:

* Perhaps age affects both self-esteem and motivation
* That is, as people age, their self-esteem and motivation might both tend to improve, as their life becomes more certain
* So, to assess whether a boost to self-esteem would really enhance motivation, the researcher needs to control age.
* For example, the researcher could survey only people who are aged in their twenties.

Indeed, as the following table shows, if you examine only people aged in their twenties, the association between self-esteem and motivation is not as apparent. That is, when you scan the second and third column now, the higher scores on self-esteem do not necessarily correspond to the higher scores on motivation. In short, we should control variables that could affect both the predictor and outcome, such as age—called spurious variables. Otherwise, the apparent relationship could be ascribed to this spurious variable.

|  |
| --- |
| Data from this study |
| Age | Self-esteem out of 10 | Motivation out of 10 |
| 21 | 3 | 5 |
| 23 | 4 | 3 |
| 21 | 3 | 4 |
| 24 | 5 | 2 |
| 20 | 3 | 5 |
| 24 | 2 | 3 |
| 49 | 7 | 8 |
| 52 | 8 | 9 |
| 47 | 9 | 7 |
| 51 | 8 | 8 |
| 46 | 7 | 8 |
| 52 | 9 | 9 |

**Confounds**

 Besides spurious variables, researchers might also want to control variables for other reasons. In particular, the measures are sometimes contaminated or confounded with other variables. To illustrate, perhaps the measure of IQ is confounded with self-esteem. For example

* if self-esteem is high, people often exaggerate their strengths
* therefore, people with a high self-esteem might inflate and thus bias their IQ
* if self-esteem was controlled, this bias would evaporate.

 In short, at times, you might want to control variables, such as age or IQ. You can apply two approaches to control variables:

* You can examine only a subset of participants, such as only people who are 18
* Or you can utilize statistical tests to predict what the results would be if you had controlled variables—such as if the participants were average in age. Linear regression is one of these tests. That is, linear regression can estimate what the association between motivation and self-esteem would have been had you controlled IQ and age.

 So, when should you control variables? You should control variables whenever you have collected information about a variable, such as age or IQ, that is likely to be strongly associated with the measures. IQ is likely to be associated motivation, so IQ, should be controlled if possible. Height is not as likely to be associated with motivation, so height might not need to be controlled.

|  |
| --- |
| **Testing the assumptions** |

 Unless particular assumptions—or patterns in the data—are fulfilled, linear regression may not be accurate. But, to understand these assumptions, you need to appreciate the concept of the predicted dependent variable and the residuals. To illustrate the predicted dependent variable, suppose you conducted a linear regression and generated the following equation:

Motivation = 2 + 0.1 x Age + .001 x IQ

 You can then utilize this equation to predict the motivation of participants from Age and IQ. Specifically, in the following table, the first two columns correspond to the Age and IQ of participants. The third column corresponds to the predicted Motivation of participants, using this formula. For example,

* The first person is 28 with an IQ of 104
* The motivation of this person is thus 2 + 0.1 x 28 + .001 x 104 = 4.90

|  |
| --- |
| Equation: Motivation = 2 + 0.1 x Age + .001 x IQ |
| Age | IQ | Predicted motivation  |  |  |
|  28 | 104  | 4.90 |   |  |
|  43 |  102 | 6.40 |   |  |
|  23 |  107 | 4.41 |   |  |
|  19 |  98 | 4.00 |   |  |
|  51 |  87 | 7.19 |   |  |

 Next, the researcher can compare this predicted motivation to the actual motivation of each participant. In the following table, the fourth column shows the actual motivation of participants. The fifth column shows the residual—defined as the difference between the actual motivation and predicted motivation. For the first person, the difference between the actual motivation, 6, and predicted motivation, 4.90, is 1.10.

|  |
| --- |
| Equation: Motivation = 2 + 0.1 x Age + .001 x IQ |
| Age | IQ | Predicted motivation  | Actual motivation | Residual |
|  28 | 104  | 4.90 | 6 | 1.10 |
|  43 |  102 | 6.40 | 7 | 0.60 |
|  23 |  107 | 4.41 | 4 | -0.41 |
|  19 |  98 | 4.00 | 3 | -1.00 |
|  51 |  87 | 7.19 | 8 | 0.81 |

 Fortunately, rather than compute these numbers yourself, the software will calculate these values. For example, in SPSS, if you tick Unstandardized predictors and Unstandardized residuals, the predicted outcome and the residual will appear in the datafile—labelled pred\_1 and res\_1 respectively.

**Normality of residuals**

 The first assumption of linear regression is these residuals are normally distributed. That is, if you constructed a frequency distribution of these residuals, the shape would resemble a bell curve. The following figure illustrates this shape. In this example

* 2 participants generated a residual of about -4
* 3 participants generated a residual of about -3 and so forth

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| -4 | -3 | -2 | -1 | 0 | 1 | 2 | 3 | 4 |

 You can apply a variety of tests to assess whether these residuals are normally distributed. Some researchers, for example, choose the “Analyze” menu and then “Descriptives” and “Explore”, generating the following screen.



* In the “Dependent List” box, specify the unstandardized residuals
* Click “Plots” and then tick “Normality plots with tests”—before clicking “Continue” and then “OK”
* In a table called “Tests of Normality”, you will then receive significance or p values associated with two tests: the Kolmogorov-Smirnov test and the Shapiro-Wilk test.

|  |
| --- |
| **Tests of Normality** |
|  | Kolmogorov-Smirnova | Shapiro-Wilk |
| Statistic | df | Sig. | Statistic | df | Sig. |
| Unstandardized Residual | .164 | 20 | .161 | .949 | 20 | .347 |
| a. Lilliefors Significance Correction |

* If the number of participants is more than 2000, use the Kolmogorov-Smirnov test
* If the number of participants is less than 2000, use the Shapiro-Wilk test
* In both instances, if the p value is significant—that is, less than .05—the assumption of normality is violated
* If the p value is not significant—that is, greater than .05—the assumption of normality is not violated. In this example, the assumption is not violated.

**Homoscedasticity and linearity**

To assess two other key assumptions—homoscedasticity and linearity—you need to construct a scatterplot between the residuals and predicted dependent variable. To achieve this goal, in the “Graphs” menu, specify “Legacy Dialogs” and then “Scatter/Dot”, “Simple Scatter”, and “Define” to generate this screen.



* In the box labelled “Y axis”, specify “Unstandardized residuals”
* In the box labelled “X axis”, specify “Unstandardized predicted values”
* Press OK

The shape of this scatterplot signifies whether the assumptions have been fulfilled. To illustrate, consider the following graph. According to the assumption called homoscedasticity, the spread or variability of residuals at one predicted value—represented by one arrow—should be similar to the spread or variability of residuals at other predicted values—represented by the other arrow. Therefore

* In the following figure, the assumption of homoscedasticity is fulfilled
* In the next figure, the assumption is violated; the residuals are more variable at high levels of the predicted values. The data fans outwards.





Furthermore, if the pattern is wavy, like a U or inverted U, the assumption of linearity might not have been fulfilled. That is, the relationship between one or more predictors and the outcome might not conform to a straight line.



**Responses to violated assumptions**

When the assumption of normality or homoscedasticity is violated, you should probably use a more conservative alpha. That is, you could use an alpha of 0.01 instead of 0.05. You would conclude that p values are significant if they are less than .01 rather than .05.

When the assumption of linearity is violated, you might need to consider another statistical technique. An example is generalized additive models.

|  |
| --- |
| **What do the other statistics mean?** |

 Linear regression also generates some other important statistics, such as R, R2, and Beta. Here is an extract of this output. The following table clarifies the meaning of these statistics.

|  |
| --- |
| **Model Summaryb** |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | .751a | .565 | .448 | 1.70847 |
| a. Predictors: (Constant), Gender, IQ, Age, Self\_esteem |
| b. Dependent Variable: Motivation |

|  |
| --- |
| **ANOVAa** |
| Model | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 56.767 | 4 | 14.192 | 4.862 | .010b |
| Residual | 43.783 | 15 | 2.919 |  |  |
| Total | 100.550 | 19 |  |  |  |
| a. Dependent Variable: Motivation |
| b. Predictors: (Constant), Gender, IQ, Age, Self\_esteem |

|  |  |
| --- | --- |
| Statistic | Interpretation |
| R | * The correlation between the predicted and actual values of the outcome or dependent variable
* The significance or p value in the ANOVA table indicates whether this correlation is significant
* If this p value is not significant, the predicted and actual values on the dependent variable are not correlated
* This pattern arises only when the predictors, such as Age and IQ, are unrelated to the dependent variable, such as Motivation
 |
| R square or R2 | * The square of R
* This value represents the proportion of variance in the dependent variable that is explained by the predictors
* For example, suppose that R2 =.40
* You would thus include than 0.40 or 40% of the variance in Motivation can be explained by self-esteem, IQ, age, and gender
* In other words, if you controlled self-esteem, IQ, age, and gender, the variance or variability in Motivation would diminish by 40%
 |
|  Adjusted R2 | * An estimate of what the R2 would have been had you included the entire population in your sample
 |
| The column called Beta in the previous output | * Indicates what the B coefficients would have been had you standardized all the measures first—that is, what the B coefficients would have been if you had converted each variable to a z score by subtracting the mean and dividing by the standard deviation.
* The higher Beta coefficients represent the most important predictors of the outcome
 |

|  |
| --- |
| **Dummy coding** |

 When researchers conduct linear regression, the predictors and outcome—sometimes called the independent variables and dependent variables—are usually numerical. For example, motivation, self-esteem, and IQ were assumed to be numerical variables. Yet, in some instances, the predictors can be categorical variables, such as gender or hair colour.

**Categorical variables comprising two categories**

 If the categorical variable comprises two categories—sometimes called a dichotomous or binary variable—the analysis is straightforward. In the data file, the researcher merely represents one category as 1 and the other category as 0. For example, if gender is one variable, males may be represented as 1 and females might be represented as 0. Consequently

* if the B value for gender is positive and significant, the researcher would conclude that motivation is positively associated with gender
* in other words, motivation is higher in the category labelled 1, males, than in the category labelled 0, females.
* conversely, if the B value for gender is negative and significant, the researcher would conclude that motivation is negatively associated with gender
* in other words, motivation is higher in the category labelled 0, females, than in the category labelled 1, males.

**Categorical variables comprising more than two categories**

 If the categorical variable comprises more than two categories, the analysis is not quite as straightforward. The researcher, instead, needs to apply an approach called dummy coding or dummy variables. To illustrate, suppose the sample comprised males, females, and intersex participants. In the following table, males, females, and intersex participants are coded as 1, 2, and 3 respectively.

|  |
| --- |
| Data from this study |
| Motivation  | Self-esteem out of 10 | Gender |
| 5 | 3 | 1 |
| 3 | 4 | 2 |
| 4 | 3 | 2 |
| 2 | 5 | 3 |
| 5 | 3 | 1 |
| 3 | 2 | 1 |
| 8 | 7 | 3 |
| 9 | 8 | 2 |
| 7 | 9 | 3 |
| 8 | 8 | 1 |
| 8 | 7 | 1 |
| 9 | 9 | 3 |

 The problem is the software might assume these three genders are numbers—and thus assume that males are more similar to females than to intersex participants, for example. Instead, researchers need to convert each category to a separate column of 1s and 0s. For instance, in the following table, 1s in the male column correspond to males and 0s in the male column correspond to non-males. Likewise, 1s in the female column correspond to females and 0s in the female column correspond to non-females.

|  |
| --- |
| Data from this study |
| Motivation  | Self-esteem out of 10 | Males | Females | Intersex |
| 5 | 3 | 1 | 0 | 0 |
| 3 | 4 | 0 | 1 | 0 |
| 4 | 3 | 0 | 1 | 0 |
| 2 | 5 | 0 | 0 | 1 |
| 5 | 3 | 1 | 0 | 0 |
| 3 | 2 | 1 | 0 | 0 |
| 8 | 7 | 0 | 0 | 1 |
| 9 | 8 | 0 | 1 | 0 |
| 7 | 9 | 0 | 0 | 1 |
| 8 | 8 | 1 | 0 | 0 |
| 8 | 7 | 1 | 0 | 0 |
| 9 | 9 | 0 | 0 | 1 |

 Unfortunately, if all three genders were included in the analysis, a problem would unfold. In particular, one of these columns is redundant. That is, each gender column equals 1 – the other gender columns. For example, intersex = 1 – males – females. When the data include this redundancy, called singularity, linear regression does not work. So, the researcher needs to prevent this problem. In particular

* the researcher excludes one of these genders from the analysis, such as males
* this excluded gender is called the reference category
* hence, the predictors include females, intersex, self-esteem, IQ, and age, but not males

|  |
| --- |
| **Coefficientsa** |
| Model | Unstandardized Coefficients | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
| 1 | (Constant) | 5.840 | 6.294 |  | .928 | .369 |
| Self\_esteem | .650 | .219 | .713 | 2.968 | .010 |
| IQ | -.040 | .061 | -.121 | -.645 | .529 |
| Age | .013 | .039 | .062 | .341 | .738 |
| Male | -.877 | .943 | -.186 | -.929 | .368 |
| Female | -.841 | 1.280 | -.162 | -.657 | .522 |
| a. Dependent Variable: Motivation |

So, how do you interpret the B values? In this example, what does the positive B value for females indicate?

* The B value represents the extent to which this gender differs from the reference category
* For example, if female generates a positive B value, the researcher would conclude that motivation is higher in females relative to males
* If female generates a negative B value, the researcher would conclude that motivation is lower in females relative to males
* In this example, females do not differ significantly from males, because this predictor is not significant
* If you wanted to compare females and intersex participants, you would need to repeat the linear regression and exclude either females or intersex participants instead.

As an aside, you could represent the reference category with -1, as shown in the following table. If you utilize this approach, the B represents the extent to which this gender differs from the average.

* For example, if female generates a positive B value, the researcher would conclude that motivation is higher in females relative to the average participant
* If female generates a negative B value, the researcher would conclude that motivation is lower in females relative to the average participant

|  |
| --- |
| Data from this study |
| Motivation  | Self-esteem out of 10 | Males | Females | Intersex |
| 5 | 3 | 1 | -1 | -1 |
| 3 | 4 | 0 | 1 | 0 |
| 4 | 3 | 0 | 1 | 0 |
| 2 | 5 | 0 | 0 | 1 |
| 5 | 3 | 1 | -1 | -1 |
| 3 | 2 | 1 | -1 | -1 |
| 8 | 7 | 0 | 0 | 1 |
| 9 | 8 | 0 | 1 | 0 |
| 7 | 9 | 0 | 0 | 1 |
| 8 | 8 | 1 | -1 | -1 |
| 8 | 7 | 1 | -1 | -1 |
| 9 | 9 | 0 | 0 | 1 |

|  |
| --- |
| **Rationale** |

 Linear regression is not hard to conduct. But, how does linear regression generate these B coefficients? What is the rationale?

* In essence, the software utilizes a formula to generate these B coefficients
* To appreciate the formula, you need to know matrix algebra.
* But, in essence, this formula is designed to maximize the R squared value. That is, any other B values would generate a lower R squared—a lower association between the predictors and outcome

Furthermore, this formula is designed to minimize the sum of squared residuals. To illustrate, consider the following table.

* In the last column, to circumvent the negative values, the residuals were squared
* In the final box, these squared residuals were summed, generating 3.39
* If any other B coefficients had been utilized, this sum of squared residuals would have been higher.
* The R squared value equals 1 – sum of squared residuals

|  |  |
| --- | --- |
| Equation: Motivation = 2 + 0.1 x Age + .001 x IQ |  |
| Age | IQ | Predicted motivation  | Actual motivation | Residual | Residual squared |
|  28 | 104  | 4.90 | 6 | 1.10 | 1.21 |
|  43 |  102 | 6.40 | 7 | 0.60 | 0.36 |
|  23 |  107 | 4.41 | 4 | -0.41 | 0.1681 |
|  19 |  98 | 4.00 | 3 | -1.00 | 1 |
|  51 |  87 | 7.19 | 8 | 0.81 | 0.6561 |
|  |  |  |  |  | Sum = 3.39 |

 This discussion about sum of squared residuals may not seem especially interesting. Nevertheless, statisticians feel this discussion is important. They even tend to refer to linear regression as “ordinary least squares regression”—primarily to highlight that linear regression minimizes these squared residuals.

|  |
| --- |
| **Software** |

**R**

 If you use R, linear regression is simple. In essence, the code resembles

* summary(lm(Motivation~ SelfEsteem + IQ + Age +Gender))

Or you could write

* Model1 = lm(Motivation~ SelfEsteem + IQ + Age +Gender)
* summary(Model1)

**Stata**

In Stata, you specify the outcome and then the predictor, such as

* regress Motivation SelfEsteem IQ Age Gender