**INTRODUCTION TO LOGISTIC REGRESSION**

**by Simon Moss**

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

Logistic regression—also called binary logistic regression—is commonly utilized in many fields, such as the health sciences. In essence, logistic regression is used

* to examine whether one set of variables, such as age, gender, and IQ, predict one of two outcomes, such as whether or not candidates will complete their PhD
* to compare two conditions or groups on a set of variables.

A similar technique, called multinomial logistic regression, is used if you want to predict more than two outcomes or compare more than two conditions. This document will primarily introduce logistic regression, but will also broach multinomial logistic regression as well. This document does not assume extensive knowledge in statistics, but may be easier to grasp if you are familiar with linear regression—a technique that is discussed in another document.

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

**Example**

 To introduce you to logistic regression, consider this example. Suppose you want to predict which research candidates are likely to complete their thesis on time. To investigate this topic, a researcher administers a survey to 500 individuals who had enrolled in a PhD or Masters by Research over 10 years ago. This survey includes questions that assess

* whether they had completed their thesis on time
* 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 completion, 0 represents did not complete on time, and 1 represents completed on time. In the column called gender, 0 represents females, and 1 represents males.



Logistic regression can be utilised to examine whether

* self-esteem, IQ, age, and sex predicts, or is associated with, whether research candidates completed on time
* self-esteem is related to whether candidates complete on time after controlling IQ, age, and sex
* these aims will become clearer as you read.

Many software packages can be utilized to conduct logistic 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 “Binary Logistic”.



* Designate “Completion” as the “Dependent” variable. That is, select “Completion” and then press the top arrow.
* Designate “Self-esteem”, “IQ”, “Age”, and “Gender” as the “Covariate” variables. These variables are sometimes called predictors instead of covariates.
* Press Continue and then OK.
* You will receive several tables of output. Here is the most important table, called “Variables in the equation”.

|  |
| --- |
| **Variables in the Equation** |
|  | B | S.E. | Wald | df | Sig. | Exp(B) |
| Step 1a | Self\_esteem | .441 | .177 | 6.229 | 1 | .013 | 1.555 |
| IQ | .007 | .032 | .053 | 1 | .818 | 1.007 |
| Age | -.002 | .027 | .007 | 1 | .932 | .998 |
| Gender | .409 | .545 | .563 | 1 | .453 | 1.505 |
| Constant | -2.668 | 3.751 | .506 | 1 | .477 | .069 |
| a. Variable(s) entered on step 1: Self\_esteem, IQ, Age, Gender. |

**Interpret the output**

 To utilize the output called “Variables in the equation”, 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 whether candidates complete on time 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 whether candidates complete on time after controlling self-esteem, age, and gender
* these principles will be clarified later.

However, significance or p values do not clarify whether self-esteem is positively or negatively related to completion on time. Does self-esteem improve or impede completion? To answer this question

* proceed to the column called “B”—a column that represents something called B coefficients
* in this example, the B coefficient associated with self-esteem is positive
* consequently, we conclude that self-esteem is positively related complete on time after controlling IQ, age, and sex. That is, self-esteem seems to facilitate completions.

**Interpret the magnitude of this effect: Conditional odds ratios**

The B coefficients also provide some insight into the extent to which the variables, such as self-esteem or IQ, differentiate the groups. More specifically, the column labelled Exp(B) is especially informative. In particular

* technically, Exp(B) represents eB. The e is a constant, sometimes called Euler’s number, that approximates 2.718
* therefore, this column equals 2.718B.
* for example, for self-esteem, B is .441; the value in the column labelled Exp(B) is thus 2.718.441 and thus 1.555.

So, what does this number mean? How do you interpret this 1.555? To understand the answer, you first need to appreciate the concept of odds. To clarify this concept of odds,

* suppose that 80% or .80 of research candidates complete their PhD on time
* the odds equals the probability they complete their PhD on time over the probability they do not complete their PhD on time
* in this instance, the odds they complete their PhD on time is thus .80/.02 = 4.
* in other words, PhD candidates are 4 times as likely to complete on time than not complete on time

So, how is this concept of odds related to the column Exp(B)? Roughly, Exp(B) indicates the degree to which the covariate, such as self-esteem, affects the odds. Strictly speaking, an increase in one unit on the covariate affects the odds by a multiple of Exp(B). To illustrate

* in this example, Exp(B) for self-esteem is 1.555
* therefore, if you increased self-esteem by one unit—such as from 8 to 9 out of 10—you would multiply the odds by 1.555
* for example, suppose the odds of completing a PhD on time is 4 in people with a self-esteem of 8
* consequently, the odds of completing a PhD on time will be 4 x 1.555 or 6.22 in people with a self-esteem of 9.

|  |
| --- |
| **The underlying rationale** |

**The underlying equation**

 Logistic regression can be utilized to generate equations that predict the likelihood of some outcome, such as the probability of PhD completion, from a set of predictors or covariates, such as self-esteem and IQ. These equations are not only useful but could also help you understand the rationale that underpins logistic regression. In particular, logistic regression assumes that

|  |
| --- |
| Loge (odds that a person is in Group 1) = B1 x covariate 1 + B2 x covariate 2 + … constant |

 Initially, this formula might seem meaningless. But, to illustrate how you could utilize this equation

* to calculate the right side of this equation, multiply each value in the B column by the corresponding predictor—and then sum these answers
* in this example, the left side is .441 x self-esteem + .007 x IQ - 0.002 x Age + 0.409 x Gender – 2.668
* as this example shows, the word “Constant” can be omitted from the equation
* therefore, in this example, the equation is

|  |
| --- |
| Loge (odds that a person is in Group 1) = .441 x self-esteem + .007 x IQ - 0.002 x Age + 0.409 x  Gender – 2.668 |

To illustrate how you would utilize 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, Loge (odds the person will complete) = .441 x 7 + .007 x 110 -.002 x 25 + .409 x 1 - .2668 = 1.548

But, what does this value of 1.548 mean? What does Loge (odds the person will complete) imply? This expression does not seem intuitive at all. Fortunately, you can then utilize the following formula

|  |
| --- |
| Probability (person is in Group 1) = 1 / [1+ Loge (odds that a person is in Group 1)] |

In this instance, the probability a person is in Group 1 = 1/(1 + 1.548) = .0175. Hence, the probability this person will complete a thesis on time is .0175. This formula can thus be used to predict the probability of an outcome, such as the probability a person will complete a thesis, from a set of covariates, such as self-esteem, IQ, age, and gender.

**How to generate the B values**

But, how does SPSS, or any software, generate the B values? Which formulas or procedures does the computer need to complete? In essence, to estimate these B values the software utilizes the previous formula to predict the likelihood each person is in Group 1—that is, the likelihood that each person will complete the thesis on time. These values appear in the following spreadsheet, in the column called Probability. In practice, these probabilities would not appear in the datasheet, but are merely presented here to facilitate learning.



According to this formula

* the probability the first individual pertains to group 1 and thus will complete the thesis on time is 0.87.
* in reality, this individual did not complete the thesis on time
* hence, this estimated probability is not appropriate.
* the software will gradually adjust the B values to improve the equation
* Specifically, the software continues to adjust the B values until all of the individuals in group 0 yield low probabilities and all the individuals in group 1 yield high probabilities, if possible

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

**Spurious variables**

 The previous section showed that self-esteem is positively associated with the likelihood a person will complete the thesis on time after controlling IQ, age, and gender. So, logistic regression, like 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 | Did the person complete on time: 1 = Yes |
| 21 | 3 | 0 |
| 23 | 4 | 0 |
| 21 | 3 | 0 |
| 24 | 5 | 0 |
| 20 | 3 | 0 |
| 24 | 2 | 1 |
| 49 | 7 | 0 |
| 52 | 8 | 1 |
| 47 | 9 | 1 |
| 51 | 8 | 1 |
| 46 | 7 | 1 |
| 52 | 9 | 1 |

 This table generates some interesting conclusions. If you scan the last two columns, you will conclude that self-esteem seems to coincide with completion. That is, people with high scores on self-esteem—the final six rows—tend to complete thesis thesis. People with low self-esteem did not tend to complete their thesis. And yet, another explanation is possible:

* Perhaps age affects both self-esteem and the inclination of people to complete the thesis
* That is, as people age, their self-esteem and motivation to complete a thesis on time might both tend to improve, as their life becomes more certain
* So, to assess whether a boost to self-esteem would really affect whether people complete their thesis on time, 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 whether a person completed a thesis 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 people who completed the thesis on time. 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 | Did the person complete on time: 1 = Yes |
| 21 | 3 | 0 |
| 23 | 4 | 0 |
| 21 | 3 | 0 |
| 24 | 5 | 0 |
| 20 | 3 | 0 |
| 24 | 2 | 1 |
| 49 | 7 | 0 |
| 52 | 8 | 1 |
| 47 | 9 | 1 |
| 51 | 8 | 1 |
| 46 | 7 | 1 |
| 52 | 9 | 1 |

**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. Logistic regression is one of these tests. That is, logistic regression can estimate what the association between whether a person completed a thesis 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 outcome—in this instance, whether the person completed the thesis. IQ is likely to be associated completion, so IQ, should be controlled if possible. Height is not as likely to be associated with completion, so height might not need to be controlled.

|  |
| --- |
| **Benefits and limitations of logistic regression** |

 Other techniques, such as MANOVA and discriminant function analyses, can also be used to compare groups on multiple variables. Nevertheless, whenever you want to compare only two groups—such as people who completed their thesis on time and people who did not complete their thesis on time—logistic regression is preferable. In particular

* logistic regression is preferable when the sample size is reasonably large, such as more than 100 individuals or units
* the main reason is that, whenever the sample size is sufficiently large, the underlying assumptions of logistic regression will be fulfilled

**Multinomial regression**

 Logistic regression, or least binary logistic regression, can compare only two groups, such as people who completed their thesis on time and people who did not complete their thesis on time. However, if you want to compare more than two groups—such as candidates who completed on time, candidates who completed late, and candidates who never completed—you need to utilize a variant of logistic regression called multinomial regression. In practice, multinomial regression is very similar except

* if using SPSS, you select “Multinomial regression” instead of “Logistic regression”
* the output presents information that compares each group to a reference group

To illustrate, suppose that SPSS generates the following output. According to this output

* self-esteem associated with group 0 is not significant; p = .258
* thus, self-esteem does not differ between group 0 and group 2, the reference category.

|  |
| --- |
| **Parameter Estimates** |
| Completiona | B | Std. Error | Wald | df | Sig. | Exp(B) | 95% Confidence Interval for Exp(B) |
| Lower Bound | Upper Bound |
| .00 | Intercept | 7.167 | 7.825 | .839 | 1 | .360 |  |  |  |
| Self\_esteem | -.293 | .259 | 1.282 | 1 | .258 | .746 | .449 | 1.239 |
| IQ | -.043 | .068 | .396 | 1 | .529 | .958 | .839 | 1.095 |
| Age | .010 | .053 | .033 | 1 | .856 | 1.010 | .910 | 1.120 |
| 1.00 | Intercept | 5.744 | 7.657 | .563 | 1 | .453 |  |  |  |
| Self\_esteem | .083 | .229 | .131 | 1 | .717 | 1.087 | .693 | 1.703 |
| IQ | -.040 | .067 | .367 | 1 | .545 | .960 | .843 | 1.094 |
| Age | .000 | .053 | .000 | 1 | .993 | 1.000 | .902 | 1.108 |
| a. The reference category is: 2.00. |

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

**R**

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

* Model1 <- glm(completion ~ selfesteem + IQ + age + gender, data = mydata, family = "binomial")
* Summary(Model1)

To conduct multinomial regression, researchers tend to use a different package and function:

* Model1 <- multinom(completion ~ selfesteem + IQ + age + gender, data = mydata)
* Summary(Model1)

**Stata**

In Stata, to conduct logistic regression or multinomial logistic regression, you specify the categorical variable and then the covariates, such as

* logit completion selfesteem IQ Age Gender
* mlogit completion selfesteem IQ Age Gender base(2)

Note that base(2) is optional, but can be used to specify which group should be assigned as the reference category.