**INTRODUCTION TO FRACTIONAL POLYNOMIAL REGRESSION**

**by Simon Moss**

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

Linear or multiple regression is often utilised to ascertain whether various predictors, such as IQ and emotional intelligence, are related to some numerical outcome, such as the motivation of research candidates. This technique is suitable whenever the relationship between the predictors and the outcome conform to a straight line. To illustrate

* the following graph depicts the association between the emotional intelligence of research candidates and their motivation
* each dot represents the emotional intelligence and motivation of one person
* the line is designed to represent the association between these two variables
* if the dots tend to be close to this line, linear or multiple regression is suitable.



However, in some circumstances, researchers might feel that a curve might correspond to the data more accurately. In these instances, either polynomial regression or fractional polynomial regression is preferable. The following display illustrates a curve. Admittedly, in this instance, the dots do not seem to lie close to the curve—so perhaps fractional polynomial regression would not be helpful.



This document will introduce polynomial regression and then fractional polynomial regression. This document assumes you have developed reasonable knowledge about linear regression, sometimes called multiple regression. If you are not familiar with linear regression, read the relevant document about this topic on CDU webpage about “Choosing your methodology and methods”.

|  |
| --- |
| **Example of a polynomial regression** |

**Linear regression**

To introduce you to fractional polynomial 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”
* IQ
* emotional intelligence or EQ—a measure of the extent to which individuals can decipher the emotions of other people and can readily control their own emotions

The following display presents a subset of the output. This output indicates that

* IQ and EQ are positively associated with motivation after controlling the other predictors.
* Motivation = .52 + .42 x IQ + .14 x EQ
* This equation can be derived from the unstandardised B values

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictor | Unstandardised B | SE | Standardised B or beta | t |
| Constant | .52 | .14 |  |  |
| IQ | .42 | .04 | .31 | 4.43\*\* |
| EQ | .14 | .02 | .14 | 2.82\* |

\* p < .05, \*\* p < .01

**Quadratic formulas**

Although these results seem convincing, the researcher questions whether linear regression is a suitable technique in these circumstances. That is, linear regression assumes the relationship between the predictors and outcome—such as the association between IQ or EQ and motivation—corresponds to a straight line. But the researcher had not evaluated this assumption. To circumvent this limitation, the researcher construct a scatterplot, depicting the relationship between motivation and one of the predictors: EQ. The following display presents this scatterplot.



Upon closer inspection, the researcher decides that perhaps the relationship between motivation and EQ is not linear. That is, rather than a straight line, perhaps the scatterplot conforms more closely to a U shape curve, as the following display suggests.



Memories of high school then saturate the awareness of this researcher. Specifically,

* the researcher vaguely remembers that such patterns correspond to the quadratic equation: y = a + b x2.
* that is, in this instance, motivation = a + b x EQ2.
* a and b are simply numbers, like 2 or 3.2

But, this excitement soon dwindles. The researcher is dubious about whether this scatterplot really conforms to this U shape. Instead

* perhaps the plot conforms to an equation that combines the straight line and the U shape
* for example, in the following display, the curve both ascends towards the top right, like a straight line, but also plateaus towards the end, like the beginning of a U shape
* if so, the formula would be motivation = a + b x EQ + c x EQ2.
* that is, the formula would include both the linear term, EQ, and the quadratic term, EQ2.



**Estimation of quadratic formulas**

So, how can the researcher determine whether this formula is indeed accurate? And how can the researcher determine the numbers a, b, and c in this formula? Actually, the researcher merely needs to complete a technique that resembles linear regression, besides a simple modification. Specifically

* the researcher first needs to convert the original data to z scores
* the researcher then computes the square of emotional intelligence
* the researcher then subjects these data to a linear or multiple regression

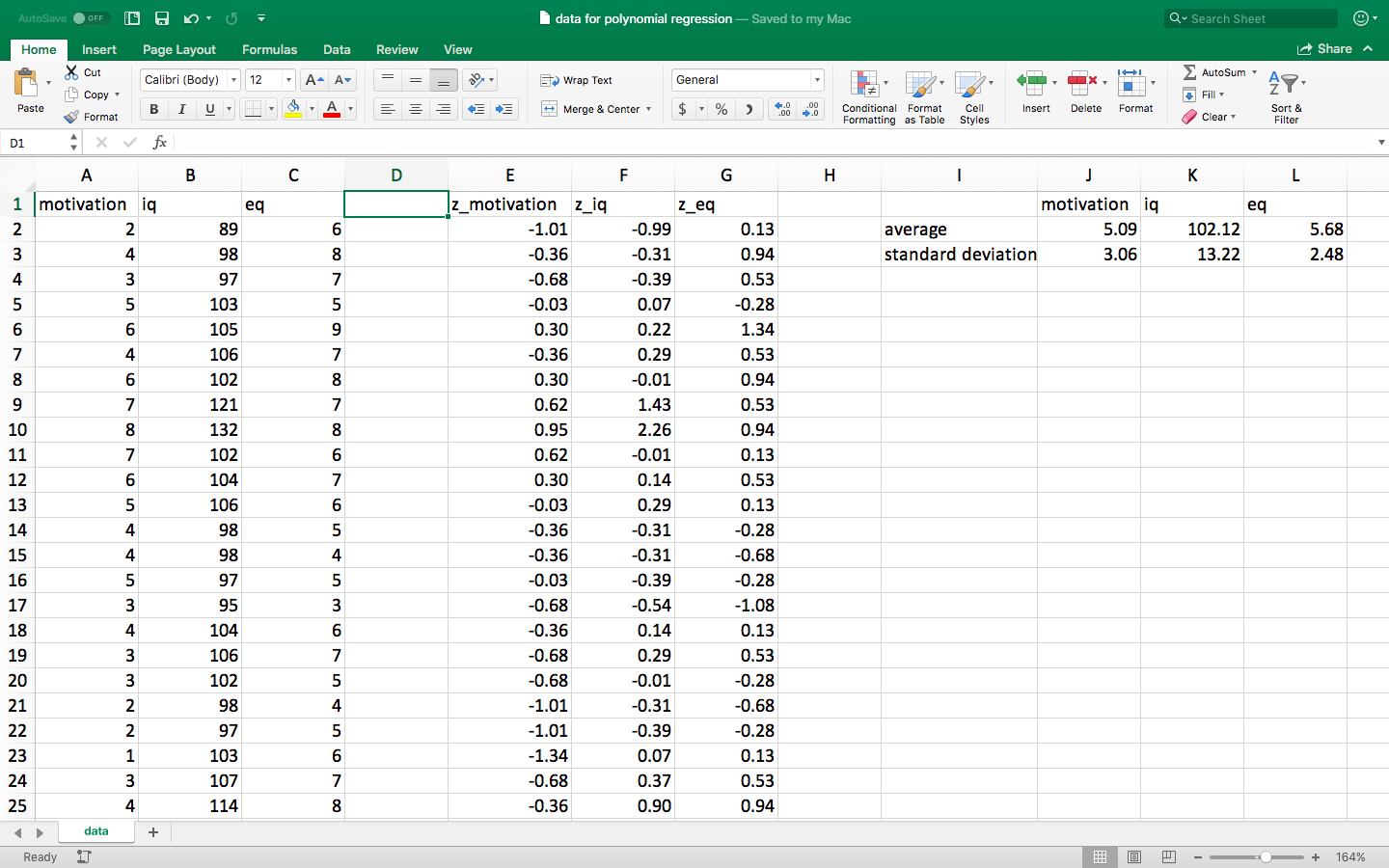
The rest of this example illustrates these calculations.

**Convert to z scores**

To examine formulas that include the square of some predictor, such as emotional intelligence, researchers often transform the original scores first. In particular, for various reasons, these equations tend to be more accurate whenever researchers utilise z scores or standard scores. To calculate z scores, the researcher

* calculates the average and standard deviation of each variable; these values appear towards the right of this spreadsheet
* subtracts this average from each score
* divides these answers by the standard deviation—to generate what is called z scores

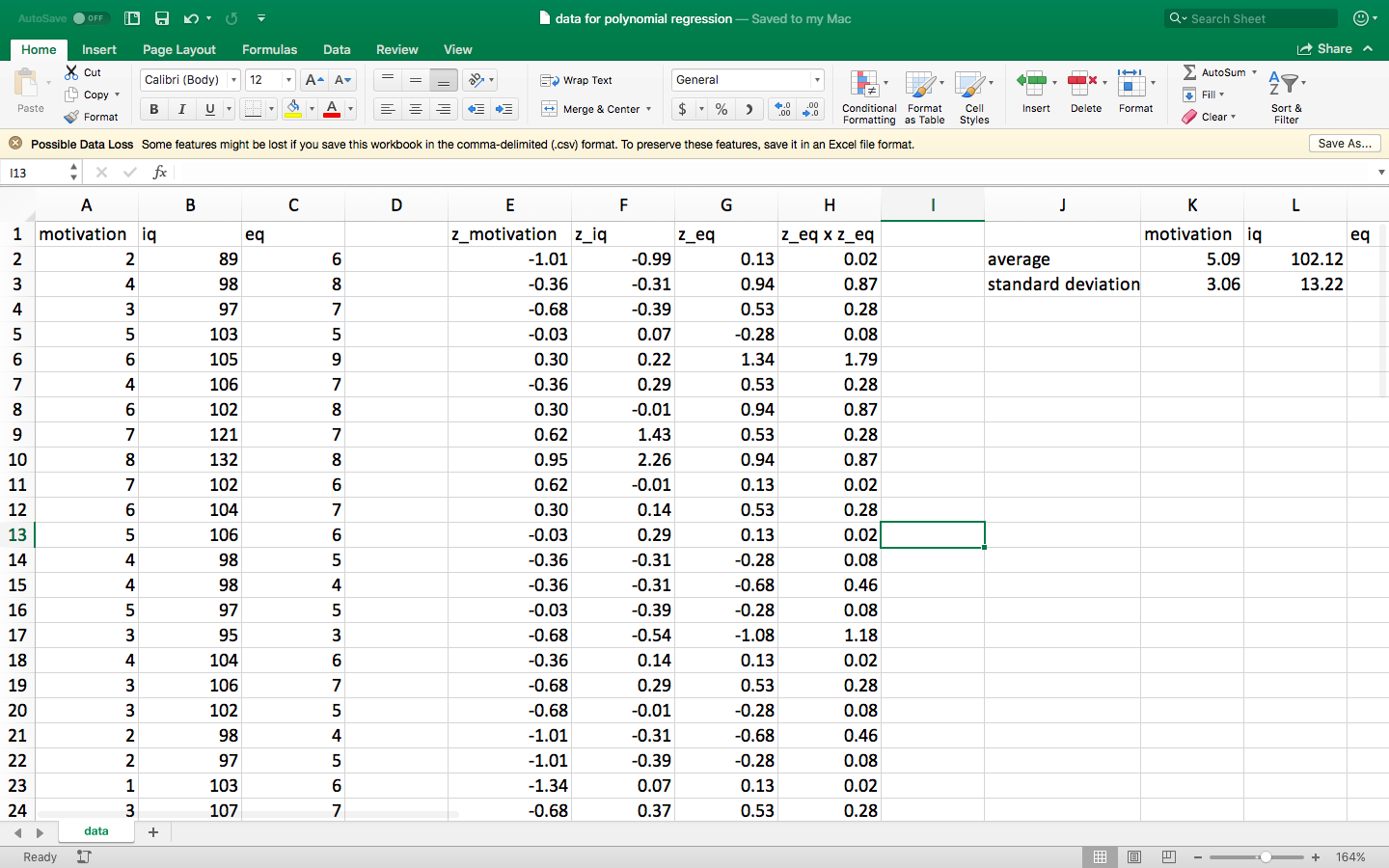
These calculations are simple, provided you have learned to compute formulas in Excel. These z scores appear in columns E, F, and G. For example, the first motivation score, 2, is converted to -1.01. The second motivation score, 5, is converted to -.36.



**Compute the square of emotional intelligence**

After computing the z scores, the researcher can then compute the square of emotional intelligence—that is, the square of all the values that appear in Column G. In particular

* these squared values now appear in Column H in the following spreadsheet.
* for example, in Column Hthe first z EQ score is .13.
* the square of this value is about .02



**Subject these data to a linear or multiple regression**

The researcher can then subject these data—in this instance, Columns E to H, to a traditional linear or multiple regression. In this instance

* the outcome or dependent variable is the z score of motivation
* the predictor or independent variables are the z scores of IQ, EQ, and EQ squared
* a subset of the output appears in the following table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictor | Unstandardised B | SE | Standardised B or beta | t |
| constant | .55 | .12 |  |  |
| IQ | .42 | .04 | .31 | 4.43\*\* |
| EQ | .14 | .03 | .14 | 2.82\* |
| EQ2 | .19 | .04 | .15 | 2.72\* |

\* p < .05, \*\* p < .01

As this output indicates

* IQ, EQ, and EQ squared are all positively associated with motivation
* the unstandardized B values are .42, .14, and .19 respectively
* accordingly, the data conform to the following formula

|  |
| --- |
| motivation = .55 + .42 x IQ + .14 x EQ + .19 x EQ2 |

In short, to calculate quadratic formula—formulas that include the square of a predictor—researchers merely need to construct another column in the data file. That is, researchers calculate the square of this predictor. Finally, researchers subject this column, as well as the original columns, to a linear regression.

**Polynomial regression**

In the previous example, the researcher includes both EQ and EQ2 in the regression analysis. However, researches might also

* subject EQ, EQ2 , and EQ3 into the regression analysis
* subject EQ, EQ2 , EQ3, and EQ4 into the regression analysis, and so forth

Before clarifying the purpose of this analysis, some definitions may be useful. In particular

* in this example, all the superscripts are positive *integers*—that is, numbers with no decimal places
* these superscripts are called *exponents*.
* whenever these exponents are integers, the analysis is called *polynomial regression;* polynomial merely refers to many numbers, such as 2, 3, and 4.

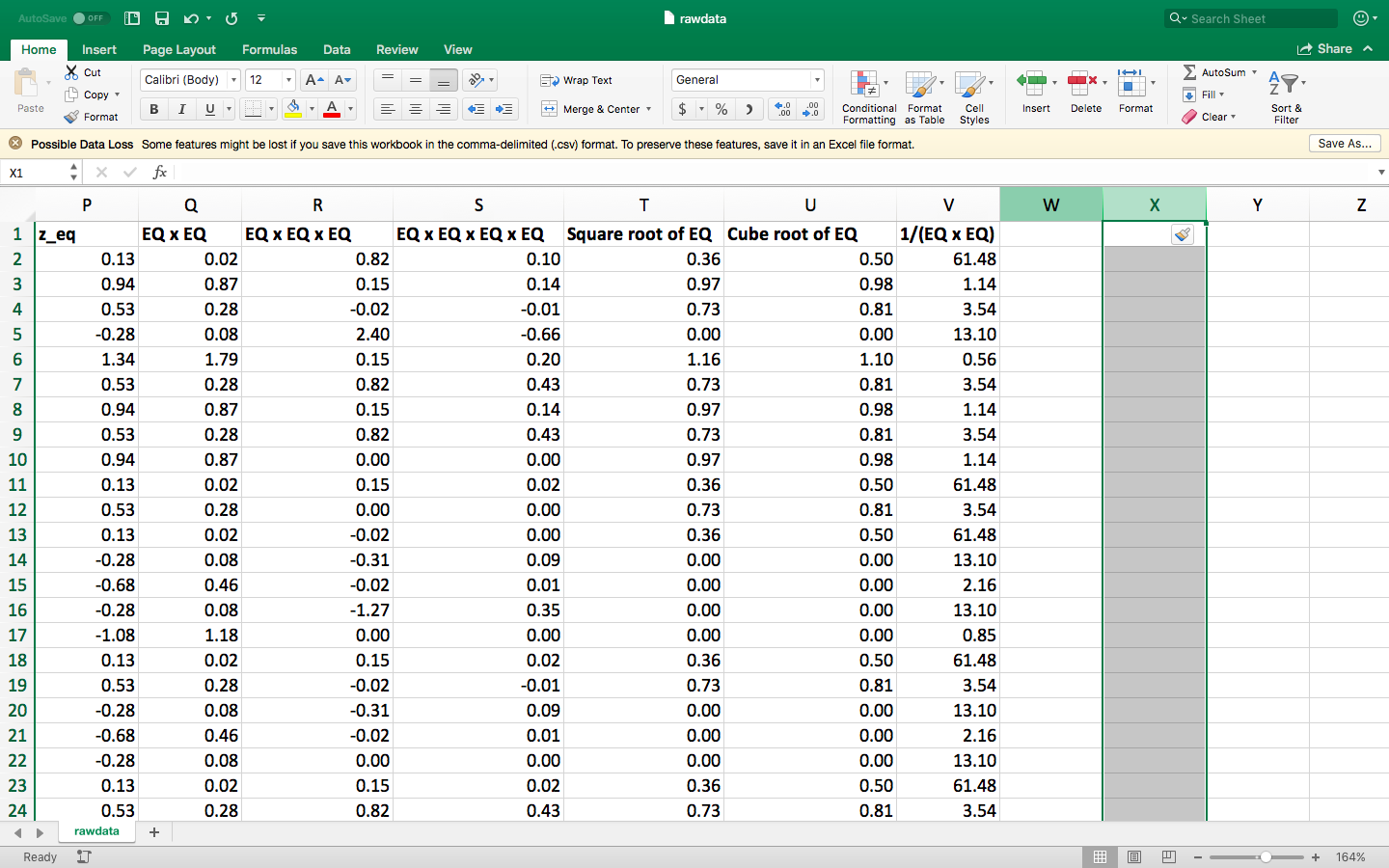
So, why do researchers sometimes use polynomial regression. The reason is simple. Polynomial regression can generate a slightly more squiggly curve—a curve that might resembles the data more closely. In other words, polynomial regression might be more accurate. In this example, if researchers utilise polynomial regression, they might be able to predict the motivation of individuals from their IQ and EQ more accurately.

|  |
| --- |
| **Introduction to fractional polynomial regression** |

**Fractional and negative exponents**

The previous section indicated that researcher could have calculated several polynomial terms, such as EQ2, EQ3, and EQ4, and then subjected all these terms to a multiple regression. In this example, all the exponents—the superscripts—were positive integers. In other instances, however, not all the exponents are positive integers. For example

* researchers might calculate EQ1/2.
* this exponent is a fraction and, therefore, seems frightening.
* but, actually, an exponent of 1/2 merely represents the square root
* in the following display, Column T is the square root of EQ—and thus equals EQ1/2
* likewise, researchers might calculate EQ1/3 or the cube root



In addition to these fractional exponents, researchers may also calculate negative exponents. To illustrate

* researchers might calculate EQ-2.
* this exponent is a negative value and, therefore, also seems frightening.
* but, actually, an exponent of -2 merely represents 1 / (the square)
* in the previous display, Column V equals 1/ (EQ x EQ)—and thus equals EQ - 2

In principle, although uncommon, the exponents could be any fraction or number, such as EQ2/3. To illustrate,

* to calculate EQ2/3 , the researcher would calculate both EQ2  and EQ1/3.
* after all, EQ2  is merely the square of EQ, and EQ1/3 is merely the cube root of the EQ
* the researcher would then multiply these two answers

**How to choose which exponents to include**

But, which exponents should researchers include in their regression analysis? How should they choose? To answer this question, Royston and Altman (1994) recommended that researchers should, roughly speaking

* first attempt some exponents that often improve the accuracy of equations—such as -2, -1, -0.5, 0.5, 1, 2, and 3
* then exclude the terms that do not significantly improve the equation.

This sequence of calculations is called a fractional polynomial regression. The remainder of this document shows how to compute a fractional polynomial regression.

|  |
| --- |
| **Step 1: Install and use R** |

You can use a variety of statistical packages to apply this technique. This document will show you how to conduct this technique in software called R. If you have not used R before, you can download and install this software at no cost. To achieve this goal

* visit <https://www.cdu1prdweb1.cdu.edu.au/files/2020-08/Introduction%20to%20R.docx> to download an introduction to R
* read the section called Download R and R studio
* although not essential, you could also skim a few of the other sections of this document to familiarize yourself with R.

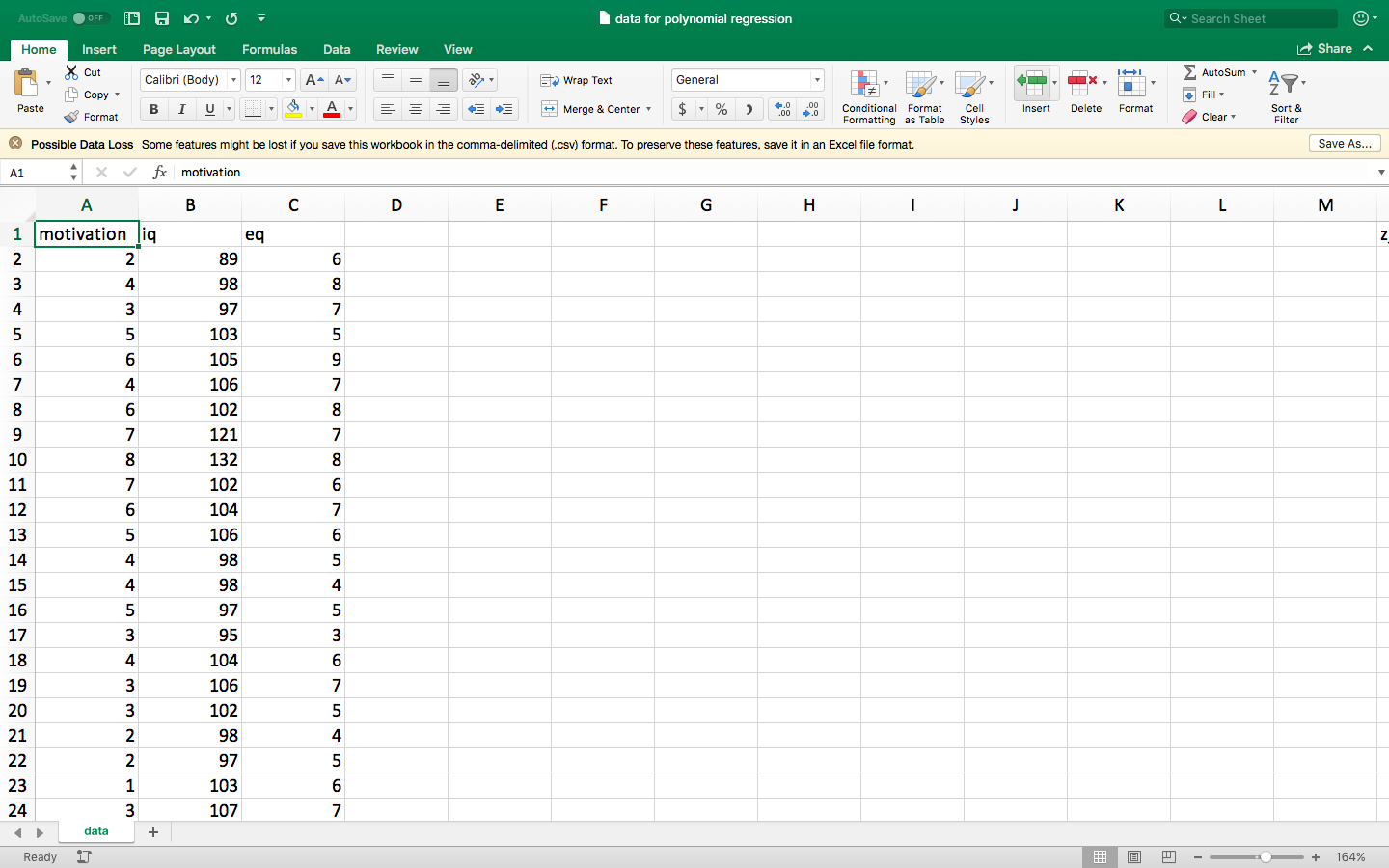
|  |
| --- |
| **Step 2: Upload the data file** |

Your next step is to upload the data into R. To achieve this goal

* open Microsoft Excel
* enter your data into Excel; you might need to copy your data from another format. Or your data might already have been entered into Excel

In particular, as the following example shows

* each column should correspond to one variable
* each row should correspond to one unit—such as one person, one animal, one specimen, and so forth
* the first row labels the variables
* to prevent complications, use labels that comprise only lowercase letters—although you could end the label with a number, such as age3



To convert this file into a csv file—such as a file called rawdata—and then to upload this file into R studio

* visit <https://www.cdu1prdweb1.cdu.edu.au/files/2020-08/Introduction%20to%20R.docx> to download the introduction to R—unless you have already downloaded this document
* read the section called “Upload some data”

|  |
| --- |
| **Step 3: Enter the code** |

To execute a fractional polynomial regression, you need to enter some code. The code might resemble the following display. At first glance, this code looks frightening. But actually this code is straightforward once explained.

|  |
| --- |
| install.packages("mfp")  library(mfp)  model1 <- mfp(motivation ~ fp(iq, df = 4, select = 0.1) + fp(eq, df = 4, select = 0.1), data = rawdata)  model1 |

To enter code, you could write one command at a time in the Console. But, if you want to enter code more efficiently,

* in R studio, choose the File menu and then “New File” as well as “R script”
* in the file that opens, paste the code that appears in the left column of the following table
* to execute this code, highlight all the instructions and press the “Run” button—a button that appears at the top of this file

|  |  |
| --- | --- |
| Code to enter | Explanation or clarification |
| install.packages("mfp")  library(mfp) | * This code installs, and then activates, a series of functions or procedures called “mfp”. * You do not really need to understand this code |
| model1  ~ mfp(formula = **motivation** ~z\_ **iq** +  fp(z\_**eq**), data = **rawdata**) | This code is designed to conduct the fractional polynomial regression. The bold symbols represent code you would modify to suit your needs. In particular   * mfp stands for multivariable fractional polynomial and merely instructs the computer to conduct fractional polynomial regression * the outcome variable, in this instance *motivation*, appears before the ~ sign * one of the predictors is iq; you could have included other predictors, using code like “z\_iq + age + gender” * **fp** indicates the researcher wants to examine various fractional polynomials of the predictor z\_eq—in which the default exponents are -2, -1, -0.5, 0.5, 1, 2, and 3 * **data = rawdata** merely specifies the file in which the data is stored. * the results are stored in a container called model1 |
| model1 | * this code simply displays the output of model1—that is, the output of this analysis   Deviance table:  Resid. Dev  Null model 524.5614  Linear model 219.5445  Final model 219.5445  Fractional polynomials:  df.initial select alpha df.final power1 power2  z\_eq 4 1 0.05 1 1 .  z\_iq 1 1 0.05 1 1 .  Transformations of covariates:  formula  z\_iq z\_iq  z\_eq I((eq/10+2.1)^1)  Rescaled coefficients:  Intercept z\_eq.1 z\_iq.1  -0.066064 0.943133 -0.002029 |

This output looks very complex. How can you interpret this output? Here are some clues:

* the residual deviations are numbers that represent the accuracy of these formulas
* in this instance, the linear model and final model were equally accurate
* accordingly, all the non-linear terms—such as EQ-2, EQ-1, EQ0.5, EQ2, and EQ3—did not improve the accuracy of this formula and were thus discarded
* the final equation therefore included one EQ term and one IQ term—as represented by the column called df.final
* the B coefficients that correspond to these terms were .943 and -.002 respectively
* if other EQ terms had been significant, more coefficients would have appeared

You might, therefore, conclude the regression uncovered the following formula, similar to a simple linear regression.

|  |
| --- |
| motivation = -0.066064 + 0.943133 x EQ + -.002 x IQ |

Actually, this conclusion would not be entirely correct. The reason is the computer cannot calculate the square root of negative numbers. Therefore the computer had to transform or change the EQ scores. That is, the computer needed to

* add 2.1 to each z\_eq score

|  |
| --- |
| **Use** |

In practice, fractional polynomial regression has often been applied in health research. Specifically, researchers often want to examine the association between some risk factor, such as inactivity, and some health outcome, such as obesity. Sometimes, they suspect the relationship might not be linear. That is, increases in the risk factor might not invariably increase the likelihood of some health outcome. In these circumstances, fractional polynomial regression might be useful.

|  |
| --- |
| **References** |

Faes, C., Aerts, M., H., G., and Molenberghs, G. (2007). Model averaging using fractional polynomials to estimate a safe level of exposure. Risk Analysis, 27(1):111–123.

Royston, P. and Altman, D. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. JRSSA, 43(3):429–467.

Sauerbrei, W. and Royston, P. (1999). Building multivariable prognostic and diagnostic models: Transformation of the predictors by using fractional polynomials. JRSSA, 162(1):71–94.