# COMPLETE EXAMPLE

MODERATION

**Data set:** data set 1.csv

**IVs:**

* Books – number of books a person read
* Attend – attendance in the class
* Interaction of books by attendance

**DV:**

* Grade – final course grade

**Research Question:** Is there an interaction between books and attendance in predicting final course grade?

**SPECIAL INSTRUCTIONS:**

* Centering: When working with two continuous variables in regression, you have to center them first. Centering means that you start with mean centered or z scores instead of the regular variables.
* Why? When you create an interaction, you are creating multicollinearity. Mean centering or Z-scoring helps.
  + Also, z score centering the variables creates SDs of 1 and a mean of zero.
  + When variables are NOT centered – the b/Beta values may be negative, but the slopes will look like they are positive (your figure will NOT match the values you are getting for each high and low group).
  + When variables are centered – the means are zero, so the b/Beta values will match your pictures.
* QuantPsyc will center the variables for you.

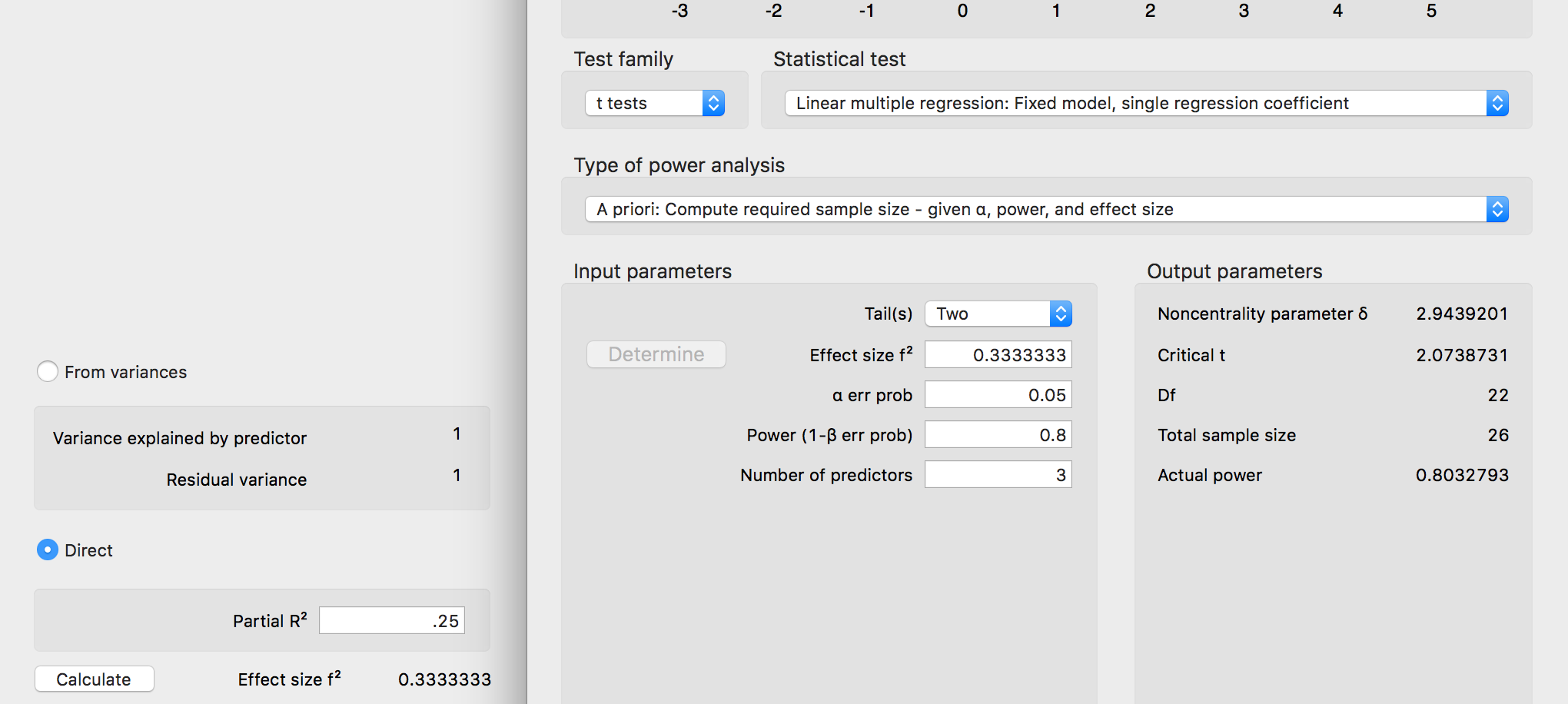
**Interaction:**

* The interaction is created by multiplying the scores of each variable together (like participant one book times participant one attendance) = participant one interaction score.
* The interaction is created after the variables are centered, so the interaction is centered as well.

**Power:**

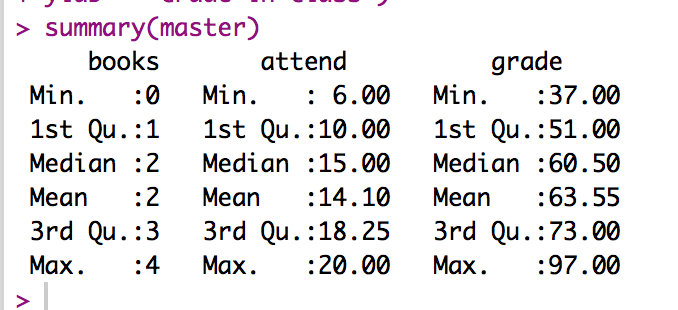
1. Open Gpower!
   1. Test family: t tests
   2. Statistical Test: Linear multiple regression: Fixed model, single regression coefficient
   3. Tails: two
   4. Effect size: click determine 🡪 direct 🡪 estimate partial R2 🡪 calculate and transfer to main window.
   5. Alpha = .05
   6. Power (1-beta of .20) = .80
   7. The number of predictors: 3
      1. One for X, M, and X\*M
      2. You can also include covariates.
2. Let’s estimate the following:
   1. Large effect size (*R2* = .25)
   2. Number of predictors = 3

We would need 26 people to find a large effect of the interaction.



**Assumptions:**

1. Accuracy:
   1. Use the summary(*dataset name*) function to get the basic information for the data.
      1. As with before, these values cannot be negative.

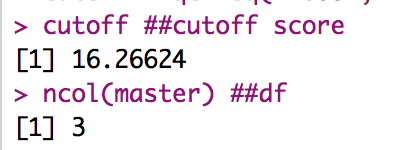


1. Missing:
   1. I can see from my summary function that I do not have missing data. Remember that you will need at least twenty variables to estimate missing data for participants – so mostly you won’t be estimating for regression.
2. Run the lm model for your data.
   1. You will run the FINAL model of your analysis for data screening.
   2. Therefore, if you are running a mediation/moderation/hierarchical be sure to run the model with ALL the variables here.
   3. Specifically, for moderation, there are three variables.
   4. output = lm(*DV* ~ scale(*X,* scale = F*)\**scale*(M,* scale = F*),* data = *dataset*)
      1. We use the scale function to make sure there is not multicollinearity created by multiplying the variables together.
      2. Remember:
         1. scale() has two components:
         2. center = T and scale = F creates mean centered variables
         3. center = T and scale = T creates z scores
         4. center = T is the default, so I left it out here to be able to read it.
         5. By using mean centers, we will match the QuantPsyc output.
3. Outliers:
   1. First: Mahalanobis scores:
      1. mahal = mahalanobis(*dataset*,

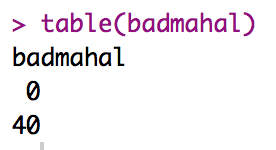
colMeans(*dataset*, na.rm = T),

cov(*dataset*, use = “pairwise.complete.obs”))

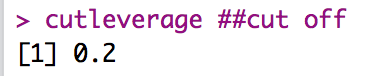
* + 1. Create the cut off score:
       1. cutoff = qchisq(1-.001, ncol(*dataset*))
    2. Remember you can use:
       1. cutoff to get the cutoff score
       2. ncol(*dataset*) to get the *df*



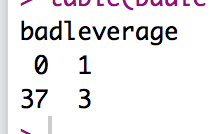
* 1. SAVE the scores:
     1. mahalout = as.numeric(mahal > cutoff) – notice that we have used > …
     2. We are checking if people are greater than the cutoff (that’s bad), and if so, giving them a 1 to mark they are an outlier. The as.numeric changes the TRUE for outlier to 1, while FALSE no outlier is a 0.
     3. This procedure is slightly different than ANOVA, because we are not simply going to keep people who are less than the cut off score – we need to keep a total of their bad scores.
     4. Check out the number of outliers (1 is bad!):
        1. table(badmahal)



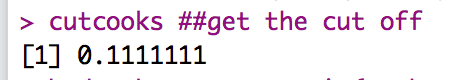
* + - 1. No outliers!
  1. Leverage scores:
     1. Remember that leverage is the influence of a single person over the slope.
     2. k = number of predictors, which is 3 for the final simple moderation model.
     3. To get leverage values:
        1. leverage = hatvalues(output)
     4. To get the cut off score:
        1. (2\*k+2)/N
        2. cutleverage = (2\*k+2) / nrow(*dataset*)
     5. Run cutleverage to see the cut off score:



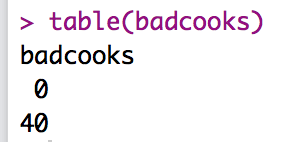
* + 1. Save the scores and see how many outliers:
       1. badleverage = as.numeric(leverage > cutleverage)
       2. table(badleverage)



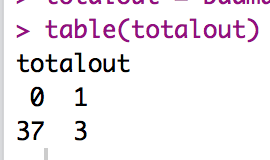
* + 1. We have three outliers.
  1. Cook’s scores:
     1. Remember that Cook’s is a measure of influence and discrepancy.
     2. To get Cook’s values:
        1. cooks = cooks.distance(output)
     3. Get the cutoff score:
        1. 4 / (N-k-1)
        2. cutcooks = 4 / (nrow(master) - k - 1)
        3. Run cutcooks to see the cut off score.



* + 1. Save the scores and see how many outliers:
       1. badcooks = as.numeric(cooks > cutcooks)
       2. table(badcooks)

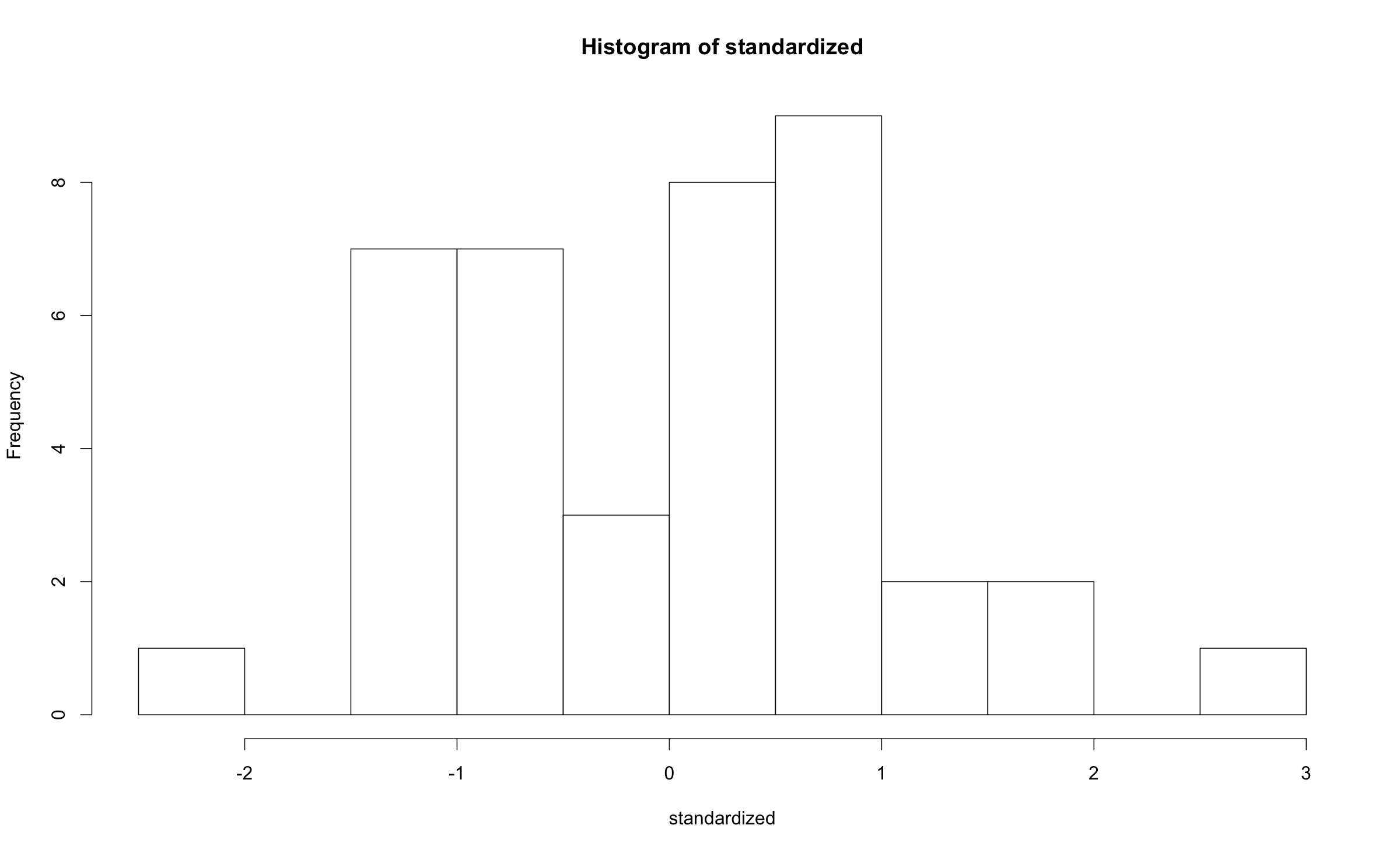


* + 1. We have no outliers!
  1. So, what does that mean overall?
     1. We want to create a total score for each participant of outliers.
     2. So, we add them up for total outlier-ness.
        1. totalout = badmahal + badleverage + badcooks
        2. table(totalout)
        3. Remember that top row = their score: 0, 1, 2, 3
        4. Bottom row is the number of people who have that score.

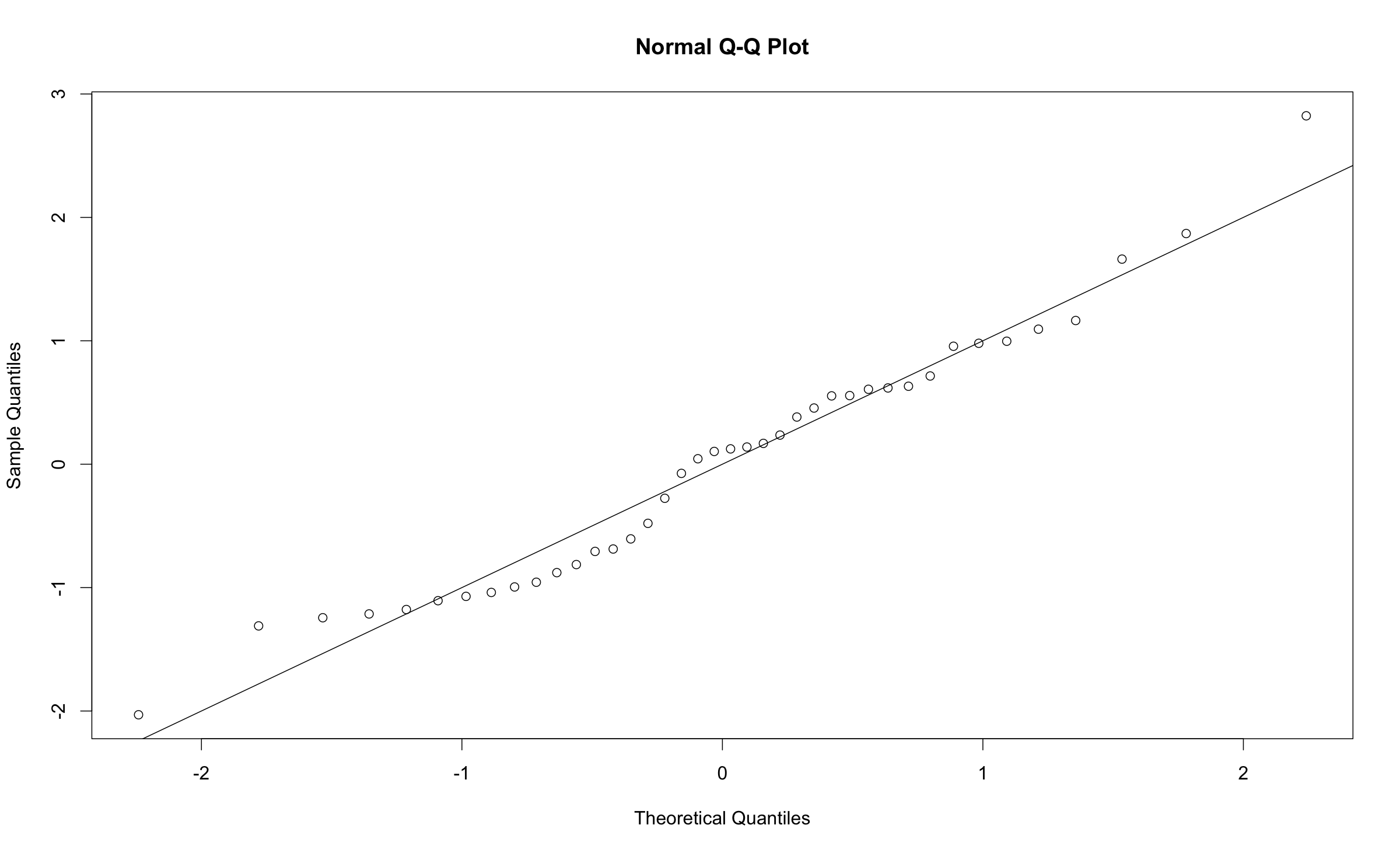


* + 1. Now, any people we have two or more problems need to get excluded:
       1. noout = subset(master, totalout < 2)
       2. We have no overall outliers.

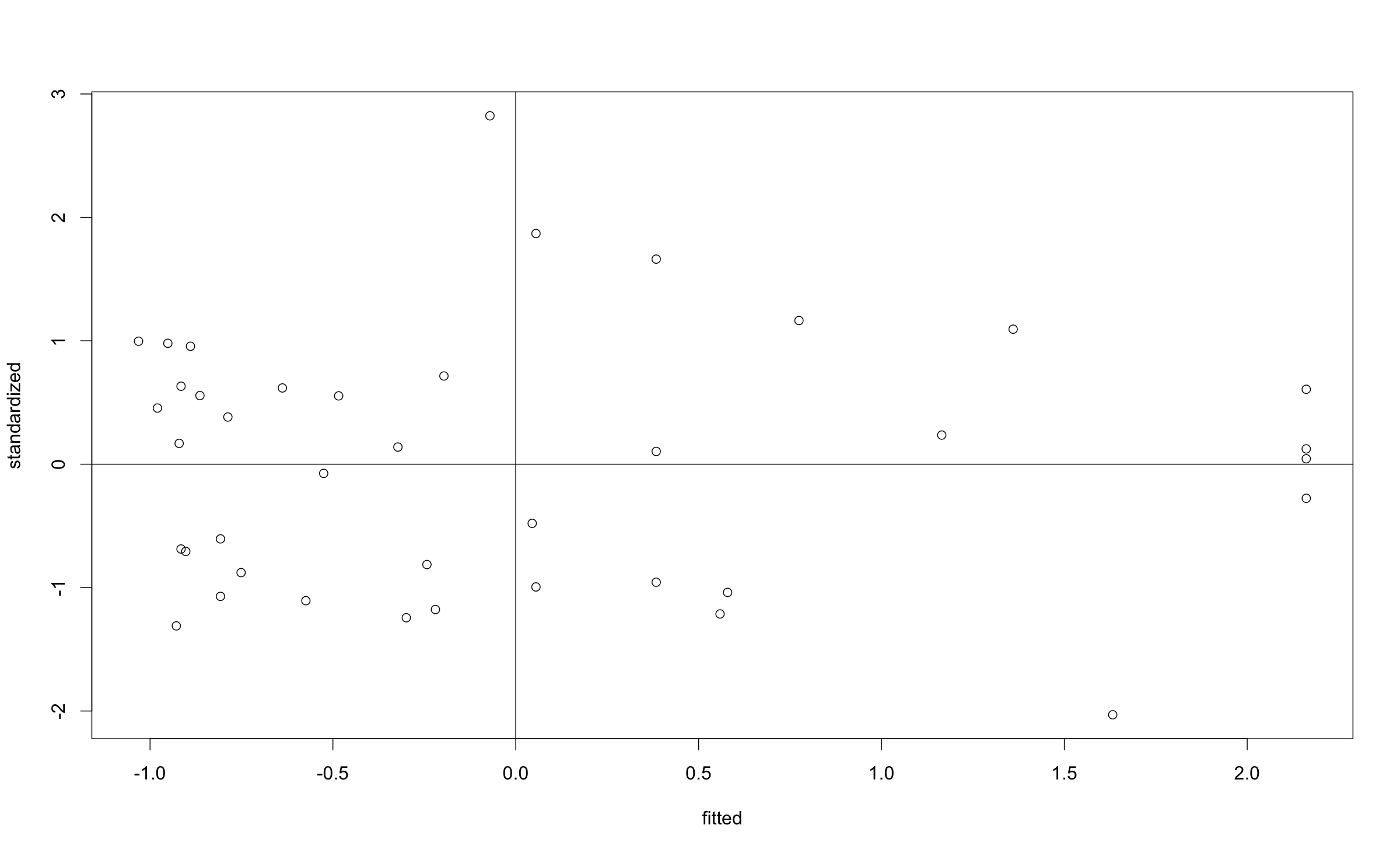
1. Additivity
   1. We expect additivity when doing simple moderation, which is why we control for it by centering the variables. If you have other control variables, you want to make sure they aren’t problematic though.
   2. Get the correlations:
      1. correl = cor(*dataset*, use = “pairwise.complete.obs”)
   3. Get the symbols chart:
      1. symnum(correl)
2. Set up the rest of the assumptions:
   1. Run the real analysis again with the no outliers dataset.
   2. No fake or randomness! It’s real regression!
   3. Create the standardized residuals:
      1. standardized = rstudent(output)
   4. Create the fitted values:
      1. fitted = scale(output$fitted.values)
3. Normality:
   1. hist(standardized)
   2. We see some bimodal problems here in this graph, but we have thirty people so we should be ok.



1. Linearity:
   1. qqnorm(standardized)
   2. abline(0,1)
   3. This example clearly meets the linearity assumption.

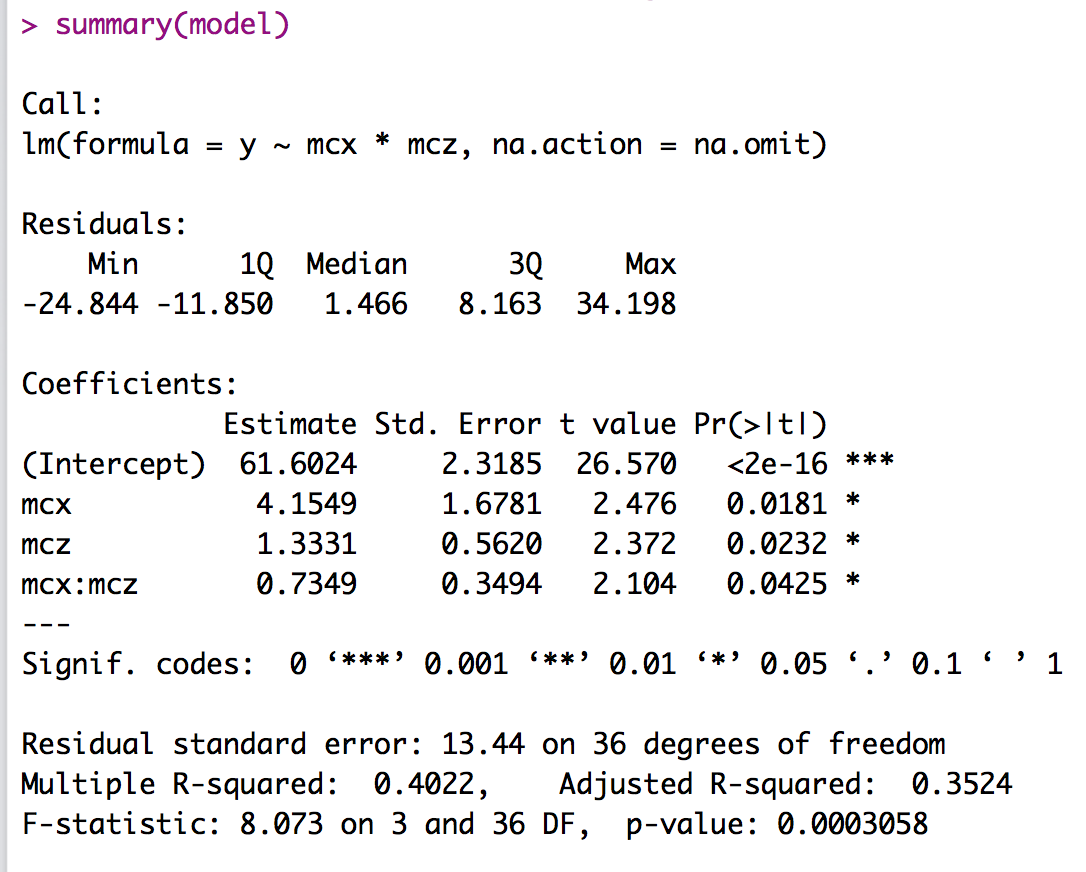


1. Homogeneity:
   1. plot(fitted,standardized)
   2. abline(0,0)
   3. abline(v = 0)
   4. Spread looks ok – the horizontal access is not the best.
2. Homoscedasticity:
   1. The spread down the graph is ok.



**Running the real analysis:**

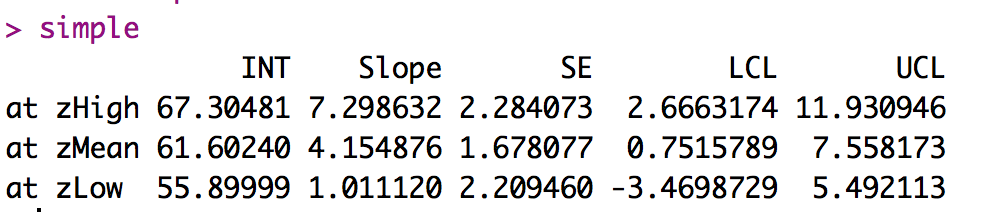
1. Load the QuantPsyc library.
2. Run the moderation with the interaction:
   1. model = moderate.lm(*X, M, Y, dataset*)
   2. summary(model)



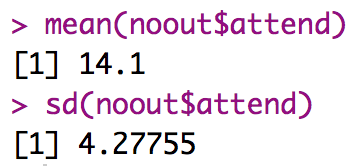
1. Interpret the output:
   1. MC stands for mean centered.
   2. Is the overall model significant? YES!
      1. *F*(3, 36) = 8.07, *p* < .001, *R2* = .40
   3. Are the main effects significant?
      1. You can think about these as main effects for ANOVA.
      2. Look at the individual predictors:
      3. Books (X): *b* = 4.15, *t*(36) = 2.48, *p* = .02
         1. As the number of books goes up, grade goes up – after controlling for attendance and the interaction.
      4. Attendance (M): *b* = 1.33, *t*(36) = 2.37, *p* = .02
         1. As attendance goes up, grade goes up – after controlling for books and the interaction.
   4. Is the interaction significant?
      1. Interaction: *b* = 0.73, *t*(36) = 2.10, *p* = .04
      2. There’s not a simple interpretation for this – just like ANOVA, you have to break down the interaction by doing a post hoc test.

**Simple slopes:**

1. Simple slopes are the post hoc test for a significant interaction.
   1. DO NOT think about simple slopes as creating low, average, and high groups – although the output will appear that way.
   2. What happens is that you create regressions that pretend that each person is one SD above their score (high), their current score (average), or one SD below their score (low).
      1. Remember we mean centered these scores – so the regression you ran to start is the average group.
      2. By mean centering, we are saying that the regression slopes in the main model (above) are the slopes for an average number of books and attendance.
   3. By pretending each person is one SD higher or lower than average allows us to calculate the slopes for those types of scores.
   4. Can I switch X and M?
      1. Yep. With mediation, the X to M direction is hypothesized, but just like interactions in regression, you can decide to switch them.
2. Run the simple slopes:
   1. simple = sim.slopes(model, meanCenter(*dataset$M*))
   2. simple



* 1. What are the slopes?
     1. The *at* on the side is the fact that we picked attendance to be the moderator.
     2. We have the slopes for BOOKs at each level of attendance (high, average, low).
  2. Terms:
     1. INT = intercept.
     2. Slope = b value for x at that level of moderation.
     3. SE = standard error.
     4. LCL = lower confidence limit
     5. UCL = upper confidence limit
  3. Interpret the slopes:
     1. At high levels of attendance: Books *b* = 7.30, 95% CI [2.66, 11.93]
     2. At average levels of attendance: Books *b* = 4.15, 95% CI [0.75, 7.55]
     3. At low levels of attendance: Books *b* = 1.01, 95% CI [-3.47, 5.49]
  4. Understand the slopes:
     1. How can I tell if they are significant?
        1. We can use the CI to determine if the slopes are better than 0.
        2. If the CI includes 0 (i.e. one positive side, one negative side), then that slope is not really different from zero.
     2. This example:
        1. When students are going to class a lot (high attendance), each book is adding 7.30 points to their grade. (a significant predictor)
        2. When students are going to class on average, each book is adding 4.15 points to their grade. (a significant predictor)
        3. When students aren’t really going to class (low attendance), books are adding 1.01 points to their grade. (not a significant predictor)
        4. Therefore, as attendance and books go up together, we get this additive effect to their grade.
     3. How do I know what you mean by low, average, high of the moderator?
        1. Get the mean and sd.
        2. mean(*dataset$column*)
        3. sd(*dataset$column*)



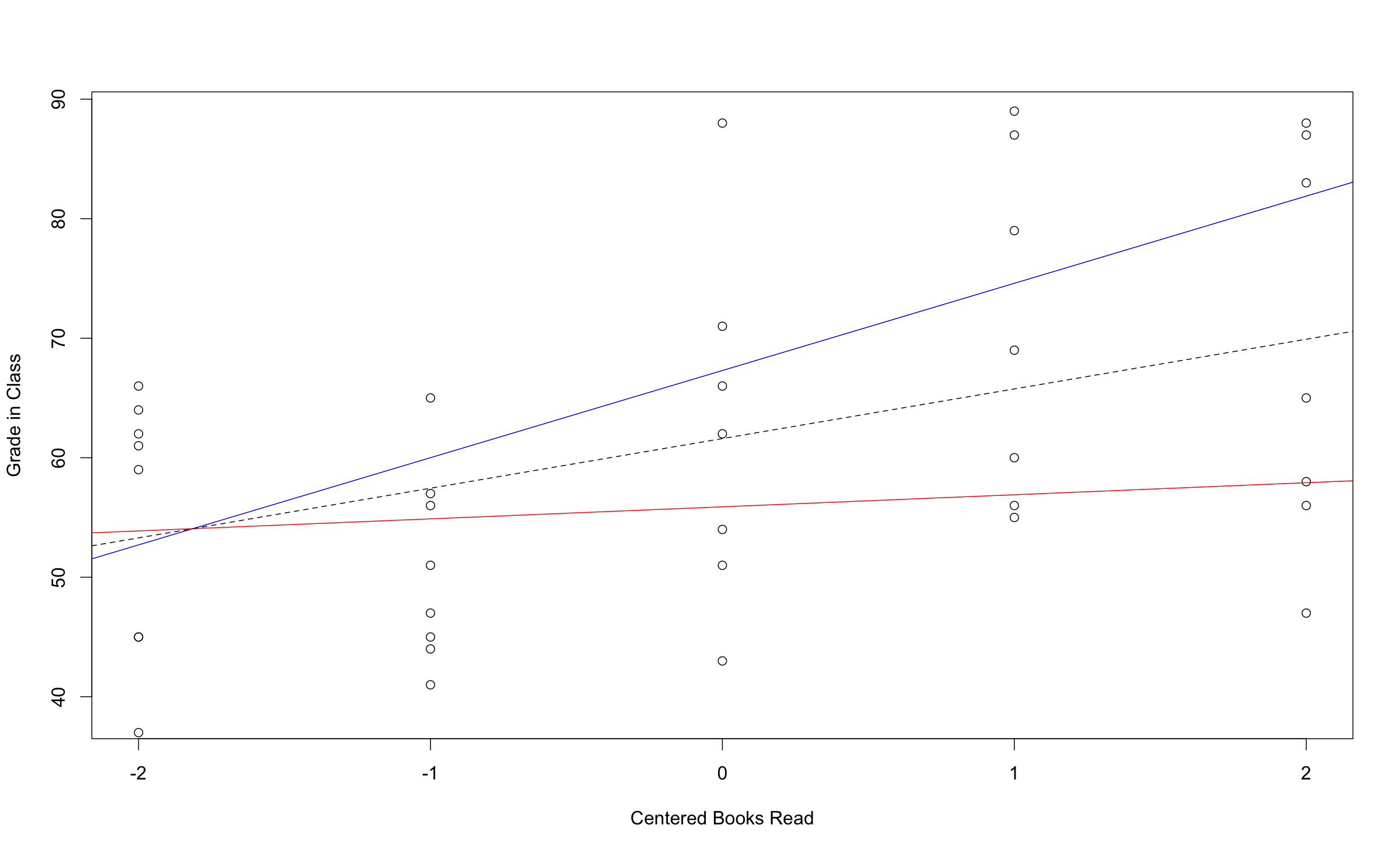
* + - 1. So, the high level is going to class 18 times, average 14 times, and low 10 times.

1. Create a graph (that is mostly decent):
   1. graph.mod(simple, *X*, *Y, dataset* ,

xlab = "Centered X LABEL",

ylab = "Y LABEL")

* 1. You will get a fan shaped pattern if your interaction was significant (not always like this, but a fan and NOT parallel lines).



*Figure 1*. Attendance moderating the relationship between books read and grades. Red line indicates low attendance, dashed line is average attendance, and blue line is high attendance.

**Results**

Attendance and number of books read during a semester were used to predict final class grade. Data were checked for outliers and assumptions of regression, and no violations were found. The *QuantPsyc* package was used to center variables, and analyze the interaction between books and attendance predicting final class grade.

The overall model of attendance and books were significant predictors of grades, *F*(3, 36) = 11.53, *p* < .001, *R2 =* .40. As a person attended more classes, their course grade increased significantly, *b* = 1.33, *t*(36) = 2.56, *p* = .01. Students could also increase their course grades by reading more books throughout the semester, *b* = 4.15, *t*(36) = 2.56, *p* = .02. Course grades were also predicted by the interaction between books read and attendance in the course, *b* = 0.73, *t*(36) = 2.42, *p* = .02. Figure 1 shows the interaction between our predictors. For average attendance, there was a significant increase in grades when reading more books, *b* = 4.15, 95% CI [0.75, 7.55]. For low attendance, there was a non-significant difference in scores when reading more books, *b* = 1.01, 95% CI [-3.47, 5.49]. Finally, high attending participants showed the largest increase when reading more books, *b* = 7.30, 95% CI [2.66, 11.93].