# COMPLETE EXAMPLE

# MEDIATION

# Please note that these steps are the traditional Baron and Kenny (1986) steps designed to instruct the basic idea of mediation. Many modifications and suggestions for mediation are now available through several plug ins in *R*. Kenny has also provided syntax and further information on his website for those who wish to use meditational analyses (<http://davidakenny.net/dtt/mediate.htm>; <http://davidakenny.net/cm/mediate.htm>).

The stages:

1. Show that the X variable predicts Y (the *c* path).
2. Show that the X variable predicts the mediator M (the *a* path).
3. Show that M predicts Y controlling for X (the *b* path).
4. Show that X is lessened by including M in predicting Y (the *c’* path).
   1. Use a Sobel test for significance.

**Data set:** data 6.csv

**IV:**

* Baseball group – AL or NL

**Mediator:**

* Budget money for extra pitchers

**DV:**

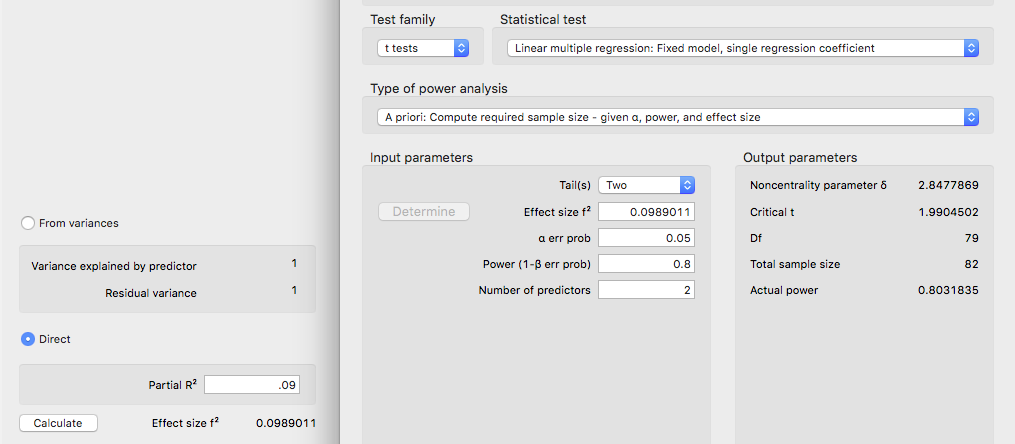
* Extra innings of pitching relief

**Research question:** Does the budget for pitchers mediate the relationship between baseball league and the innings of relief?

**Power:**

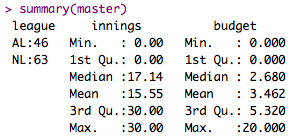
1. A tricky subject for mediation – you want X to originally predict Y, X to predict M, and X to stop predicting Y when M is included…
   1. The easiest guess for this type of test, is to estimate the effect size you’d expect for X to Y (originally), X to M, and M to Y.
   2. Others have written on this subject, there are R plug ins, but if you are truly using it for sample size estimation, then a best guess on those values will be adequate.
2. 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:
      1. For X 🡪 Y, you’d use 1
      2. For X 🡪 M, you’d use 1
      3. For M 🡪 Y, you’d use 2
      4. Pick the one you’d think has the lowest effect size since that will require the most people.
      5. For our examples, we will use M 🡪 Y.
3. Let’s estimate the following:
   1. Medium effect size (*R2* = .09)
   2. Number of predictors = 2 (X and M)

We would need 82 people to find a medium effect size of M to Y.



**Assumptions:**

1. Accuracy:
   1. Use the summary(*dataset name*) function to get the basic information for the data.
   2. Let’s check out minimum and maximum:
      1. League should be factored. (listen I know there aren’t that many teams … just pretend we measured across multiple years).
      2. Innings and budget should not be negative.

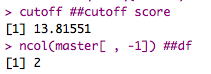


1. Missing:
   1. I can see from my summary function that I do not have missing data. Remember that you will need at least twenty variables to estimate missing data for participants – so mostly you won’t be estimating for regression.
2. Run the lm model for your data.
   1. You will run the FINAL model 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 mediation, there are two variables.
   4. output = lm(*DV* ~ *X + M,* data = *dataset*)
3. Outliers:
   1. First: Mahalanobis scores:
      1. mahal = mahalanobis(*dataset*,

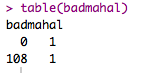
colMeans(*dataset*, na.rm = T),

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*



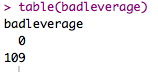
* 1. SAVE the scores:
     1. mahalout = as.numeric(mahal > cutoff) – notice that we have used > …
     2. We are checking if people are greater than the cutoff (that’s bad), and if so, giving them a 1 to mark they are an outlier. The as.numeric changes the TRUE for outlier to 1, while FALSE no outlier 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. One outlier.
  1. Leverage scores:
     1. Remember that leverage is the influence of a single person over the slope.
     2. k = number of predictors, which is 2 for the final mediation 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:



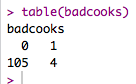
* + 1. Save the scores and see how many outliers:
       1. badleverage = as.numeric(leverage > cutleverage)
       2. table(badleverage)



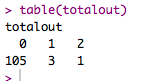
* + 1. We have no outliers.
  1. Cook’s scores:
     1. Remember that Cook’s is a measure of influence and discrepancy.
     2. To get Cook’s values:
        1. cooks = cooks.distance(output)
     3. Get the cutoff score:
        1. 4 / (N-k-1)
        2. cutcooks = 4 / (nrow(master) - k - 1)
        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)

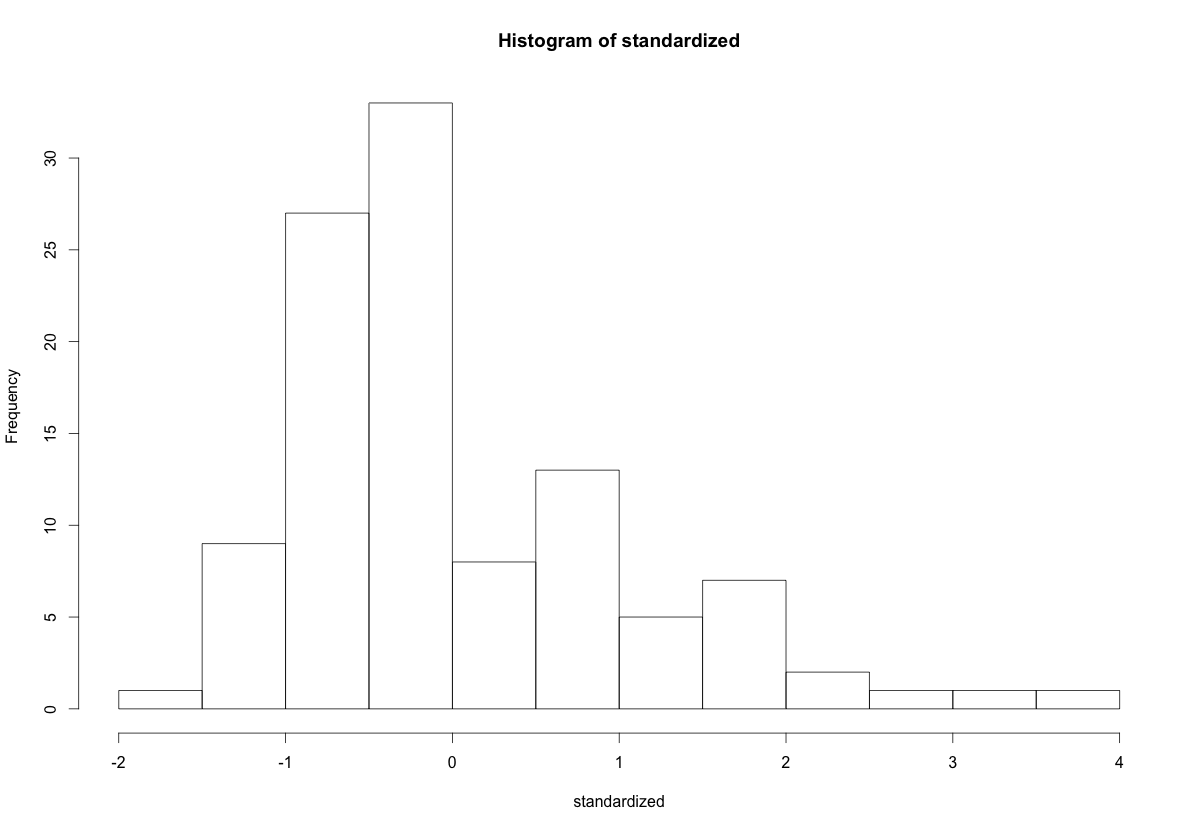


* + 1. We have four outliers!
  1. So, what does that mean overall?
     1. We want to create a total score for each participant of outliers.
     2. So, we add them up for total outlier-ness.
        1. totalout = badmahal + badleverage + badcooks
        2. table(totalout)
        3. Remember that top row = their score: 0, 1, 2, 3
        4. Bottom row is the number of people who have that score.

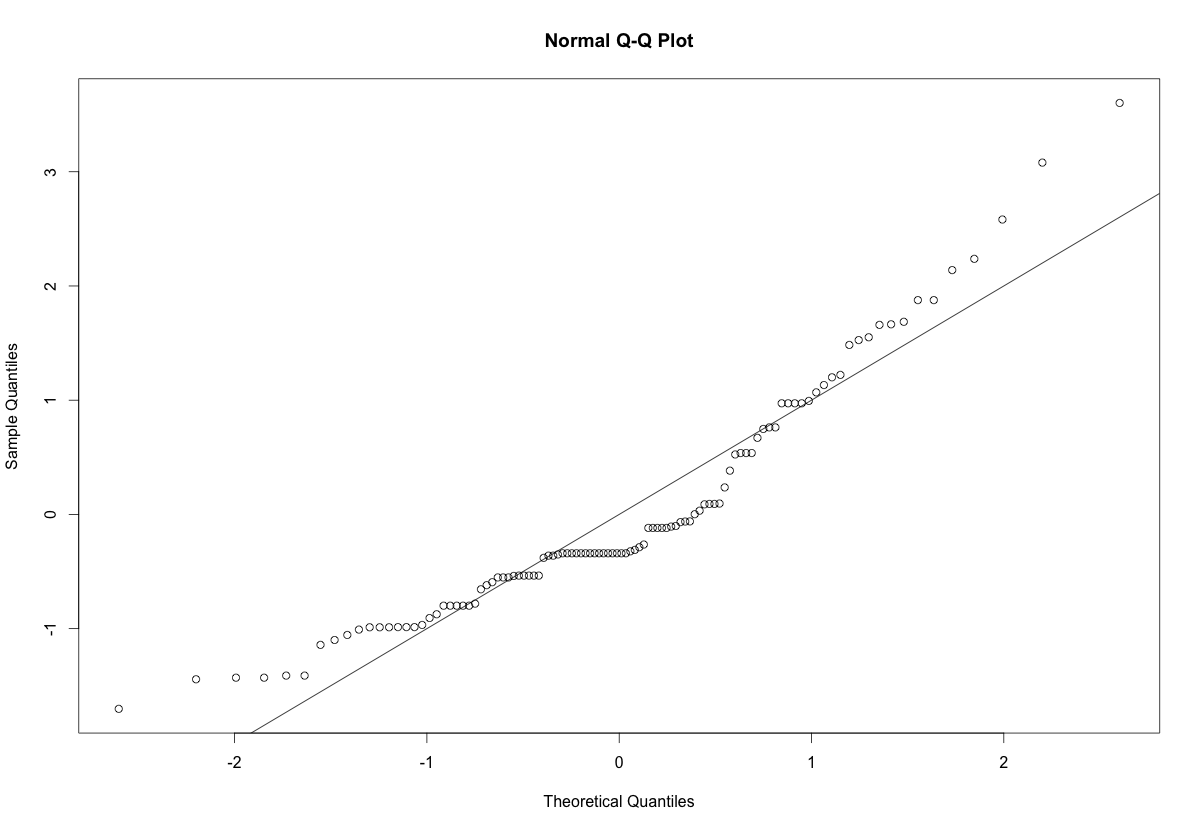


* + 1. Now, any people we have two or more problems need to get excluded:
       1. noout = subset(master, totalout < 2)
       2. We have one overall outlier.

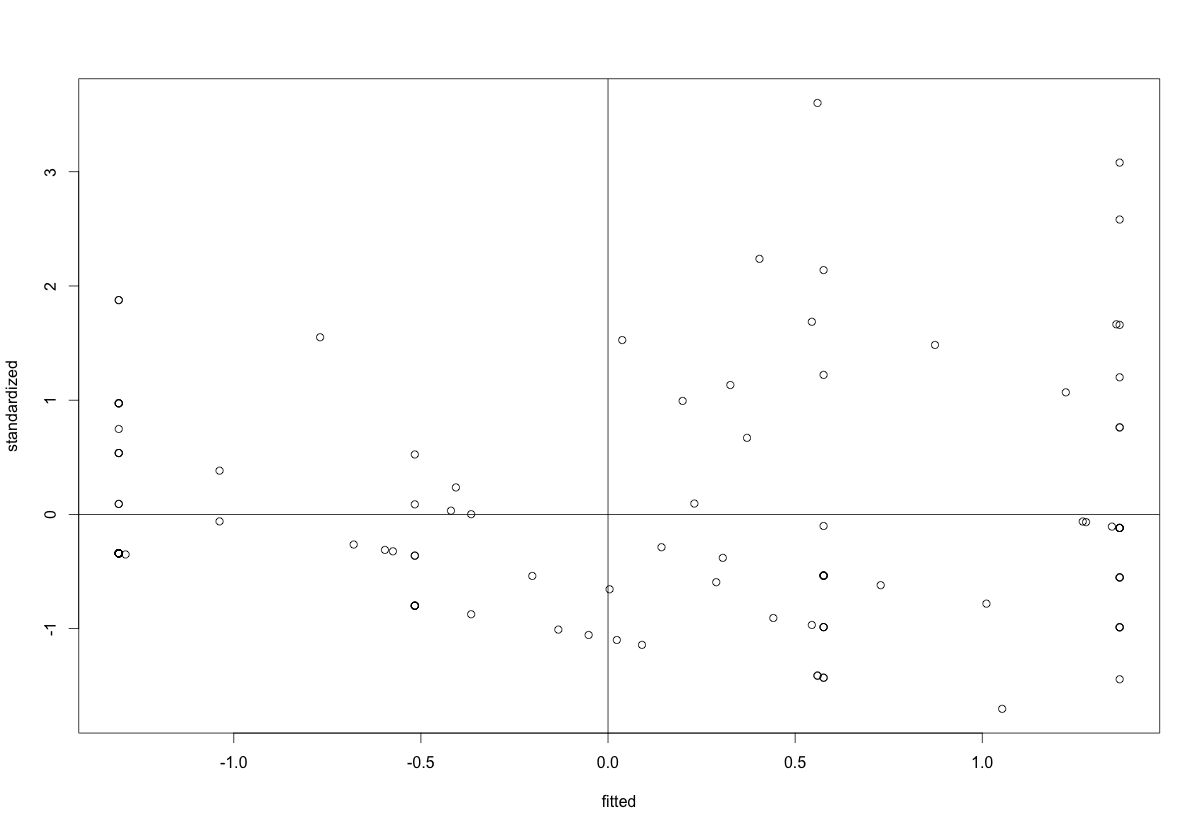
1. Additivity
   1. Here’s the thing about mediation – it’s the point that X and M and Y are correlated in specific ways. So, you do kind of want them to be highly correlated – therefore, you should run this analysis if you have other control variables for mediation. If you are not including other variables, I might skip this step.
   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 do see some skew here, not a lot of data on the right side of zero.



1. Linearity:
   1. qqnorm(standardized)
   2. abline(0,1)
   3. This example is NOT very linear.

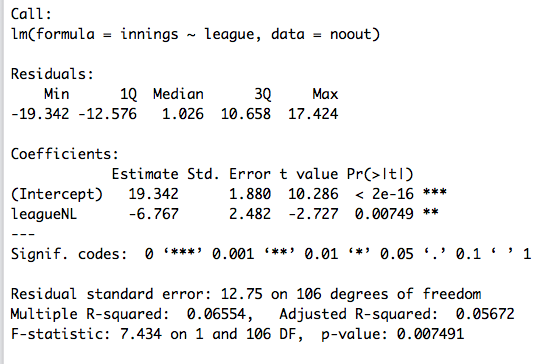


1. Homogeneity:
   1. plot(fitted,standardized)
   2. abline(0,0)
   3. abline(v = 0)
   4. Spread looks ok – the vertical access is not the best.
2. Homoscedasticity:
   1. The spread down the graph is ok – it does seem larger at the ends and smaller in the middle, which isn’t that good.

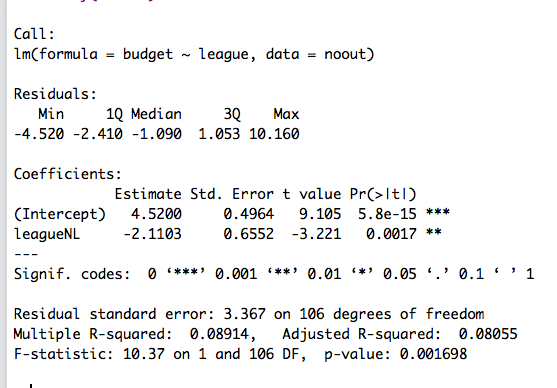


**Running the real analysis:**

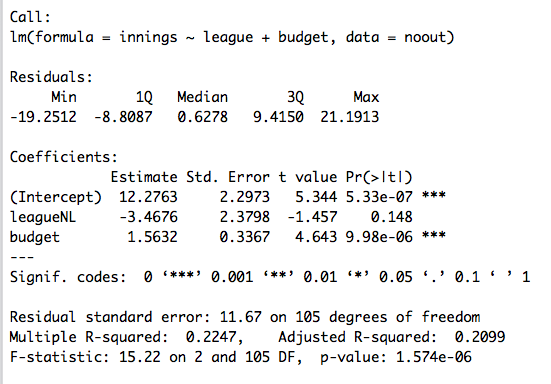
1. We are going to run the regressions one step at a time to make sure each requirement for mediation is met along the way.
   1. C path X 🡪 Y:
      1. modelc = lm(*Y ~ X,* data = *dataset*)
      2. summary(modelc)

****

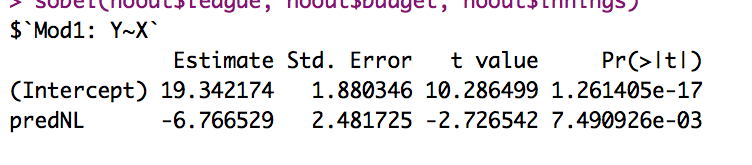
* + 1. Is the c path significant?
       1. Yes, *b* = -6.77, *t*(106) = -2.73, *p* = .007
       2. Put the overall model statistics in a table.
       3. *F*(1, 106) = 7.43, *p* = .007, *R2* = .07
  1. A path X 🡪 M:
     1. modela = lm(*M ~ X,* data = *dataset*)
     2. summary(modela)

****

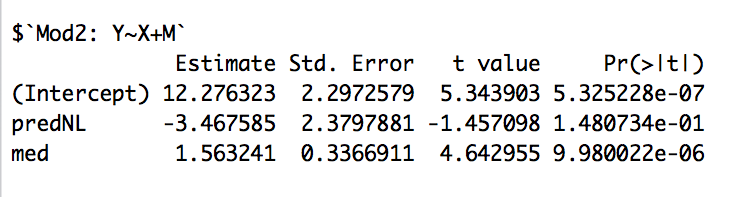
* + 1. Is the a path significant?
       1. Yes, *b* = -2.11, *t*(106) = -3.22, *p* = .002
       2. Put the overall model statistics in a table.
       3. *F*(1, 106) = 10.37, *p* = .002, *R2* = .09
  1. B path M 🡪 Y … controlling for X:
     1. modelb = lm(*Y ~ X + M,* data = *dataset*)
     2. summary(modelb)

****

* + 1. Is the b path significant?
       1. Yes, *b* = 1.56, *t*(105) = 4.64, *p* < .001
       2. Put the overall model statistics in a table.
       3. *F*(2,105) = 15.22, *p* < .001, *R2* = .22
  1. C’ path X 🡪 Y controlling for M:
     1. You’ve already run this model (it’s the same as model b), but instead we are looking at the X predictor to see if it has changed from the original c path.
     2. In a perfect world, X is now nonsignificant…but that doesn’t always happen. If it’s significant, that’s ok, you will test if c and c’ are *different* with the Sobel test.
     3. Our c’ is non-significant:
        1. *b* = -3.47, *t*(105) = -1.46, *p* = .15
        2. The model statistics are the same, so don’t repeat them.
  2. The Sobel test:
     1. This test tells you if c and c’ are significant different. You could find that X predicts M which predicts Y but that c and c’ aren’t really changed by including M in the equation (which implies a lot of correlated variables, not mediation).
     2. Additionally, you want c’ to be *closer to zero* than c, indicating a decrease in the relationship. If c’ is greater or stronger than c, that indicates you might actually have an interactive effect, rather than a mediator (i.e. moderation).
     3. Use the multilevel library to run the Sobel test (and make sure you got the right output above):
        1. library(multilevel)
        2. sobel(*X, M, Y*)

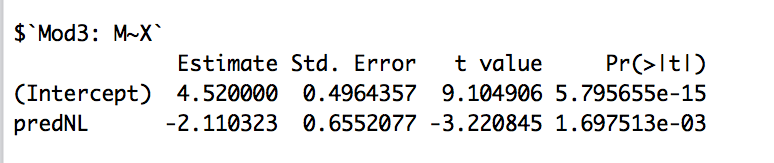
****

c path

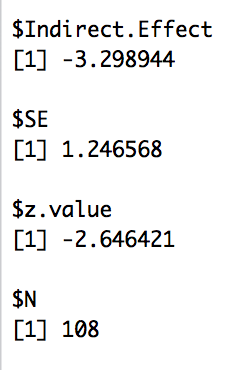


c’ path

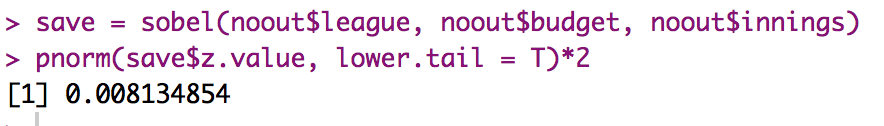
b path

****

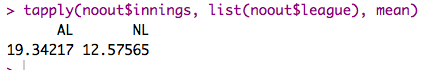
a path



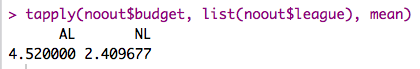
* + 1. The indirect effect is the c – c’ change, which is sometimes reported.
    2. SE = standard error of the change between c and c’.
    3. Z = z score (indirect effect / SE) of the change.
    4. What to report:
       1. *Z* = -2.65 … wait where’s p?
       2. Calculate p using pnorm:
          1. pnorm(abs(z score), lower.tail = F)\*2
          2. Why times two? Because we want a two tailed test.
          3. I pulled the z score by saving the sobel output and using $z.value, so I don’t have to manually type it every time.
       3. *Z* = -2.64, *p* = .008
       4. Therefore, significant mediation occurred by including budget between league and innings



1. Interpreting the whole thing:
   1. League predicted innings:
      1. Remember, this variable is dummy coded, so I ran the means to figure it out.



* + 1. The AL needed more pitching relief innings … not too surprising since NL has much better pitching.
  1. League predicted budget:
     1. However, it appears that the AL spends more money on pitching (maybe a Red Sox/Yankees bias here…).



* 1. Budget predicts innings: this effect is positive (b = 1.56) which indicates that as budget increases, you have more pitching relief innings (maybe because you can afford it?).
  2. When we put in budget, the effect between league and relief innings disappears. So really what is happening is that the leagues predict different budgets, which then predicts differences in relief innings.

1. Mostly, graphs are not presented because the DV changes between steps of the mediation process. However, you will see the triangle picture presented (you can copy mine below and edit it).
2. What about beta?
   1. There is some debate about presenting standardized effects in these models because the IVs and DV are changing as you work through the mediation steps. In this type of analysis, it is likely best to present values in b to stay true to the scale. However, I have seen beta presented to be able to compare effects when X and M are on different scales.

**An example write up using the original dataset example before it became an awesome baseball example:**

**Results**

Treatment condition for housing (either treated or control group) was used to predict days in housing, with housing contacts expected to mediate the relationship between treatment condition and days in housing. Data were screened for multivariate outliers, leverage and influence and two cases were removed as outliers and influential data points. All other assumptions of regression were checked and appeared satisfactory.

See Figure 1 for visual diagram of the mediated relationship. First, using steps described by Baron and Kenny (1986), treatment was a significant predictor of days in housing (the *c* pathway), as shown in Table 1. The treatment condition showed a higher number of days in housing than the control condition, *t*(105) = 2.72, *p* = .01. Second, treatment condition was used to predict the mediator variable of housing contacts (the *a* pathway), which showed that treatment condition was positively related to housing contacts, *t*(105) = 2.98, *p* = .01. Third, the relationship between the mediator housing contacts and days in housing was examined controlling for the treatment condition (the *b* pathway). Number of housing contacts was positively related to the number of days in housing, *t*(104) = 4.96, *p* <.001. Lastly, the mediated relationship between treatment condition and days in housing was examined for a drop in prediction when the mediator was added to the model (the *c’* pathway). Mediation was found, showing that the relationship between treatment condition and days in housing was no longer significant after controlling for housing contacts, *t*(104) = 1.50, *p* = .14. The Sobel test was used to determine that the *ab* effect was significantly greater than zero, *Z* = 2.17, *p* = .02.

|  |
| --- |
| *b* 1.74  *a* 1.89  Housing Contacts  *c'* 3.55  *c* 6.82  Days in Housing  Treatment Condition |

*Figure 1.* Mediated relationship between treatment condition and days in housing with housing contacts as the mediator.

Table 1

*Model Summaries for Mediation Analysis.*

|  |  |  |  |
| --- | --- | --- | --- |
| Model | *F* | *p* | *R2* |
| Treatment Condition predicting Days in Housing | (1, 105) = 7.38 | <.01 | .07 |
| Treatment Condition predicting Housing Contacts | (1, 105) = 8.87 | <.01 | .08 |
| Treatment Condition and Housing Contacts predicting Days in Housing | (1, 104) = 16.82 | <.001 | .24 |