Regression

**Description:**

Regression is a type of analysis where you want to be able to *predict* some scores using other information. Regression is more flexible than ANOVA because you do not have to have categorical or group variables for your independent variables (but you can).

**Definitions/Abbreviations:**

Predictors – these are your independent variables that you want to use to predict the dependent variable. You can think of them as IVs or Xs. You can use categorical, Likert, or continuous variables as predictors.

Criterion – this variable is your dependent variable or the information you are trying to predict/understand. You can think of this variable as DV or Y.

a = Constant, this value is added to the regression equation to help with prediction. Not often reported in research.

b (little b) = Coefficient, this value is the unstandardized slope for your regression equation. For every one point increase in X, you will get b points increase in Y. This score will be based on the scale of the variable you are using to predict.

β (beta) = Coefficient, this value is the standardized slope for your regression equation. With one X/predictor beta is equal to Pearson’s r. Since beta is standardized you can use it to compare across predictors at which IV best explained your DV.

Y-hat = The predicted value based on the equation. If your regression equation is good, it will be close to the actual value for that participant Y (DV).

Equation = Y-hat = a + b1X1 + b2X2 + …

R = Pearson’s product moment correlation or the relationship between your predicted values and people’s actual DV/Y values.

R2 = the amount of variance in the DV scores that your IV/predictors account for. This number is effect size for regression equations.

sr = semi-partial correlation, the variance from only *that IV* over the total variance. Tells you how much variance overall that variable accounts for. The unique contribution of that variable to R2 – increase in proportion explained Y when that variable is added to the equation.

pr = partial correlation, the variance from only *that IV* over the variance *not accounted for (error)*. Tells you how much variance your variance accounts for when you only look at variance that you can explain. Proportion of variance in Y not explain by other predictors.

PR > SR

D

C

B

A

R = B + C + D / A + B + C + D

r IV1 = B + C / A + B + C + D

sr IV1 = B / A + B + C + D

pr IV1 = B / A

Dummy coding - a way to code categorical predictors to better understand which category predicts the IV (see below).

**Types of Regression:**

1. Simple Linear Regression – regression with only one predictor variable (IV).
   1. It’s called simple because there’s only ONE thing predicting.
   2. In this case, beta = r.
2. Multiple Linear Regression – regression with more than one predictor variable (IVs).
   1. You can use a mix of variables – continuous, categorical, Likert, etc.
   2. You can use MLR to figure out which IVs are the most important.
3. Types MLR
   1. Standard/Simultaneous Linear Regression – all variables are entered into the equation at the same time.
      1. Each variable assessed as if it were the last variable entered
      2. This “controls” for the other IVs, as we talked about the interpretation of b.
      3. Evaluates sr > 0?
      4. If you have two highly correlated IVs the one with the biggest sr gets all the variance.
      5. Therefore the other IV will get very little variance associated with it and look unimportant (suppression).
   2. Hierarchical / Sequential Linear Regression – predictor variables are entered in as sets or steps. Variance gets assigned at each step to the first variables entered.
      1. IVs enter the regression equation in an order specified by the researcher.
      2. What order?
         1. Assigned by theoretical importance
         2. Or you can control for nuisance variables in the first step
      3. First IV is basically tested against r (since there’s nothing else in the equation it gets all the variance)
      4. Next IVs are tested against pr (they only get the left over variance).
      5. Using SETS of IVs instead of individuals
         1. So, say you have a group of IVs that are super highly correlated but you don’t know how to combine them or want to eliminate them.
         2. Instead you will process each step as a SET and you don’t care about each individual predictor
   3. Statistical / Stepwise Linear Regression – predictor variables are entered in steps, based on some statistical cutoff.
      1. Entry into the equation is solely based on statistical relationship and nothing to do with theory or your experiment
      2. Types
         1. Forward – biggest IV is added first, then each IV is added as long as it accounts for enough variance
         2. Backward – all are entered in the equation at first, and then each one is removed if it doesn’t account for enough variance
         3. Stepwise – mix between the two (adds them but then may later delete them if they are no longer important).

**Research Questions:**

1. How good is the equation? Can we predict people’s scores better than chance?
   1. Use the overall model ANOVA statistics, R2 for effect size.
2. Which IVs are the most important? Which contribute the most to prediction?
   1. Use the coefficient statistics (t values) with pr2 as the effect size.
3. Are groups of IVs important together? (sets)
   1. Use a hierarchical regression with variables as steps.
4. Controlling for X, can Y be predictive? (control)

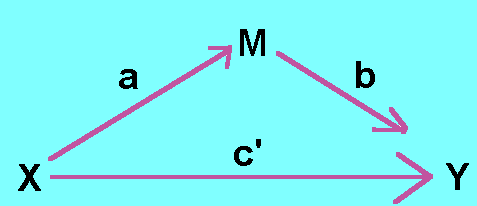
Y=Z, controlling for X, can Z (another predictor) be predictive? (control) Like this?

* 1. Use a hierarchical regression with the control variable in the first step.

**Special Research Questions:**

Mediation – analyzing that the relationship between X and Y is changed by the inclusion of another variable M.

1. c path: Simple regression analysis with X predicting Y
2. a path: Simple regression analysis with X predicting M
3. b path: Simple regression analysis with M predicting Y
4. c’ path: Multiple regression analysis with X and M predicting Y



Moderation – analyzing an interaction between two X variables in the prediction of Y. Similar to an interaction effect in ANOVA.

X

Y

M

**Plug in needed for mediation/moderation**:

<http://www.afhayes.com/introduction-to-mediation-moderation-and-conditional-process-analysis.html>

(google PROCESS Andrew Hayes)

**Power:**

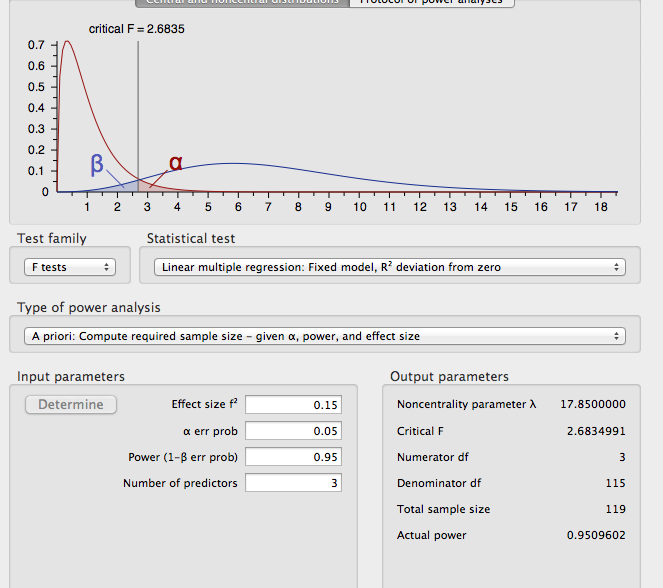
Power for regression depends on the type of analysis you decide to pick. Here’s an example of multiple linear regression.

Other things to think about:

* Ratio of cases to IVs
  + If you have less cases than IVs you will get a perfect solution (aka account for all the variance in the DV)
* Rules of thumb for participants
  + More than 50 + 8(K) (number of IVs)
  + Or 104 + K (for testing importance of predictors)
* How many people?
  + However…you can have too many people.
  + Any correlation or predictor will be significant with very large N
  + Practical versus statistical significance

G\*Power options include:

* F-test
* Linear multiple regression: fixed model R2 deviation from zero
  + (R2 increase would be for hierarchical regression)
* Effect size f2 – you would need to estimate based on research or you can hit determine
  + Rho = estimate based on r2
* Alpha = .05
* Beta = .80
* Number of predictors = number of IVs.



**Assumptions:**

* Accuracy – still the same rules to check and make sure.
* Missing data – missing data in a regression has the same rules of replacement as other analyses. For continuous data, linear trend at point, mean replacement, etc would help replace the data. For categorical data, it’s best to leave it out.
* Outliers – outliers in regression can either cause lots of problems or almost no problems. You’ll want to check a couple different things when deciding on outliers.
  + Mahalanobis distance = gives you the combination of predictor’s mean and sees how far away from the mean of means a person is.
  + Leverage – how much influence over the slope a person has (how much their score will change b values)
    - Cut off rule of thumb = (2K+2)/N
    - K is the number of IV(s) predictors
    - N is the number of people
  + Discrepancy – how far away from other data points a point is (no influence on the slope)
  + Cooks – influence – combination of both leverage and discrepancy
    - Cut off rule of thumb = 4/(N-K-1)
    - K is the number of IV/predictors
    - N is the number of people
* Multicollinearity – using two predictor variables that are the same will lower your power overall and one of the variables will appear not to be predictive. You want to make sure they are not too highly correlated to avoid this problem.
* Normality – you want both your predictors and your DV to be normally distributed.
* Linearity – since you are trying to predict values, you will want a linear relationship between your predictors and your DV.
* Homogeneity – the variance of your predictors and DV should be about the same. This assumption can be a little weird if you have categorical predictors and a continuous DV, so you really just want to make sure you have enough people in the study to control for this assumption.
* Homoscedasticity – you want the errors to be distributed evenly across all the values of the dependent variable. You look for this in a residual plot to make sure there are no megaphone shapes.
* Independence of Errors – you design your experiment to meet this assumption. You want to make sure that scores of one person are not based on other participants in the experiment. Each person’s score(s) should only be influenced by experimental manipulations (things you did to them) and their individualness. (Mud on Scale Example).

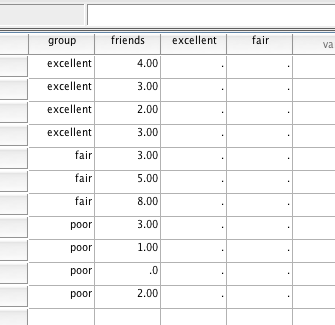
# Dummy Coding Example

Dataset: Dummy Coding

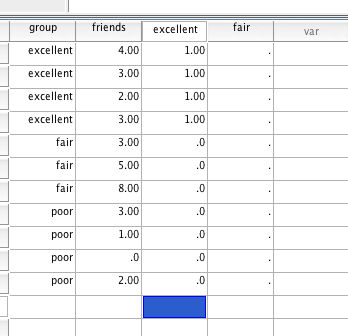
IV: Group health – excellent, fair, poor

DV: Number of Friends

1. First open your data set and figure out how many levels you have for your categorical variable (it’s the number of value labels).
2. In our example here, we have three levels; excellent, fair, and poor.
3. Create LEVELS – 1 new variables (you do NOT need three new variables for 3 levels, only 2).



1. Now, label the first group as 1 and everyone else as zero.
2. Here everyone labeled as one in the excellent variable is the excellent group.



1. Now do the same thing for the fair group. Label everyone in that group as 1, and not as 0.



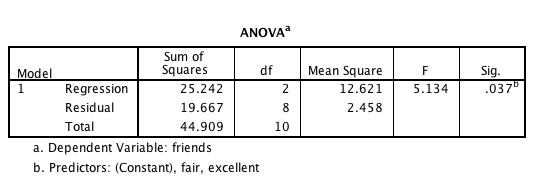
1. Now everyone with a one in the fair variable is fair.
2. But what about poor?! Well everyone who has a zero on both excellent AND fair are poor. You do not need another variable.
3. You will enter these as a set into a regression.

(Here we are going to test a simple linear regression without data screening to show how dummy coding in regression = post hoc tests in ANOVA, so please see other sections for how to screen/run regression).

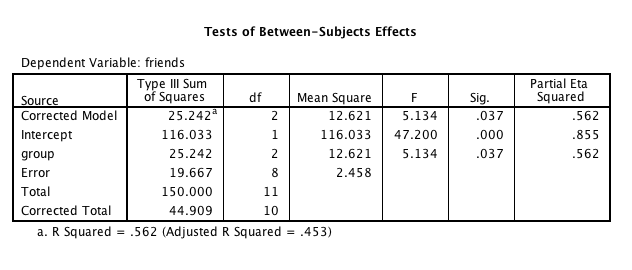
How to run a simple SLR without the pictures provided below:

1. Analyze > regression > linear.
2. DV into the dependent box.
3. IVs into the independent box.
4. Just for this example: Statistics: R2 change, part and partials.

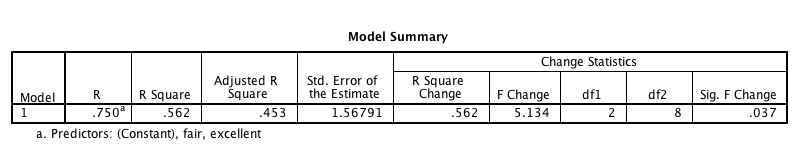
Check out the regression ANOVA:



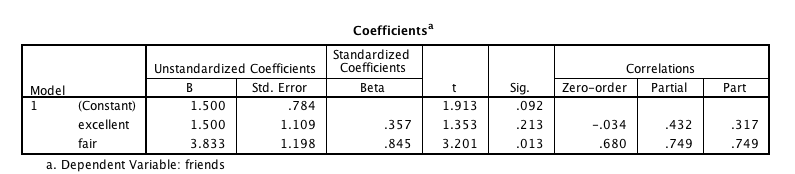
Check out the ANOVA (ANOVA):



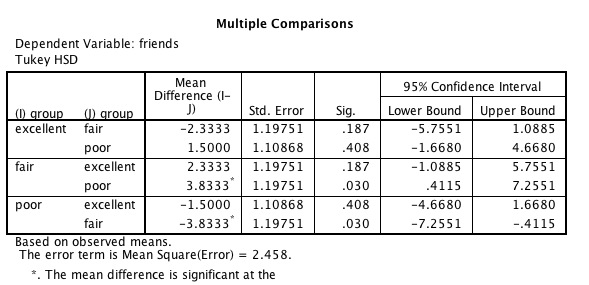
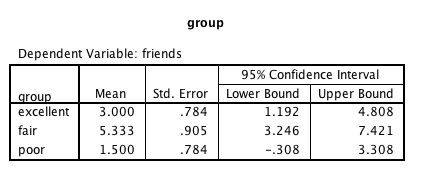
Check out the R2:



Check out the coefficients:



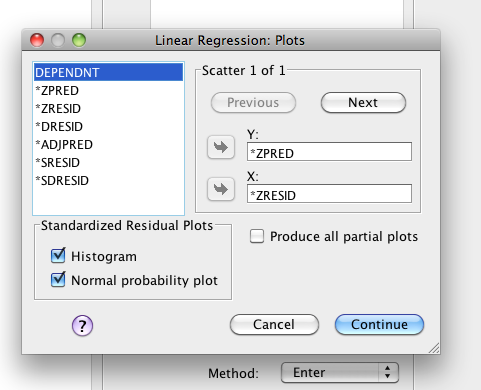
Check out the Tukey Post Hoc:



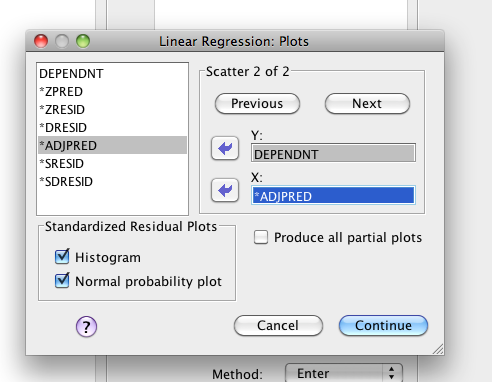
# Charts for Results Sections

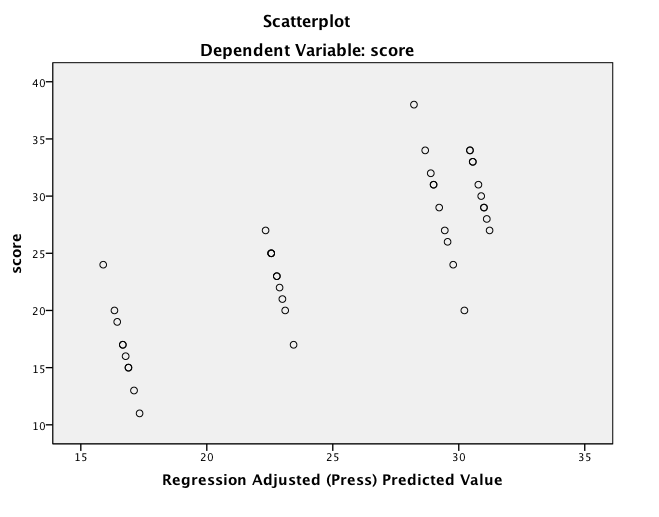
Note: There’s not a good place to put this section because there are a lot of examples below, and it’s the same rules for each one. So, I stuck it here at the beginning rather than once in each section.

1. If you want to add a chart of the regression, you can do this:
   1. Analyze > Regression > Linear
   2. Hit plots and then see the “next” button like when you enter steps into a regression.

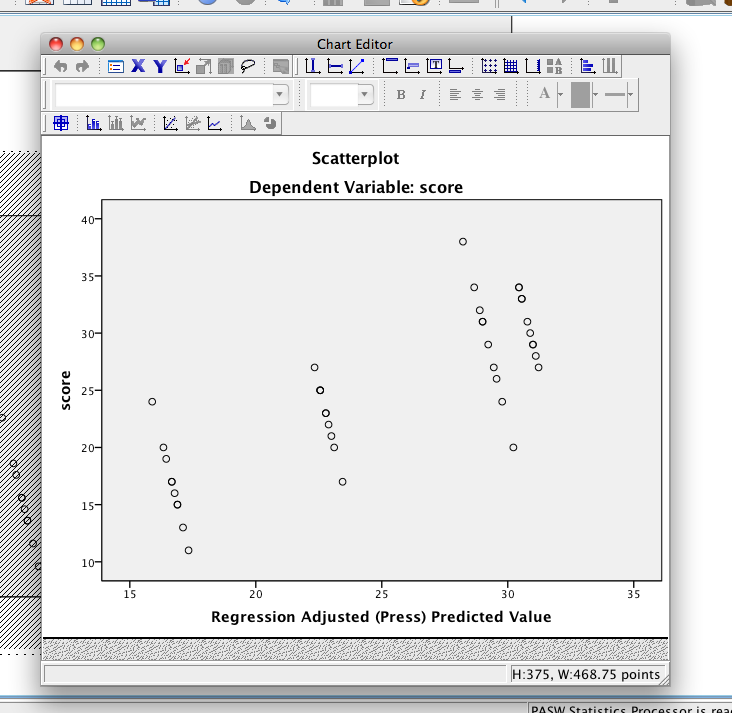


* 1. Hit Next.
  2. Enter Dependent in Y.
  3. Enter ADJPRED in X.
     1. This variable is ALL of the IVs together, not just one like a scatter plot.
  4. Hit continue. Now you’ll get the residual plots for outliers AND a regression graph.

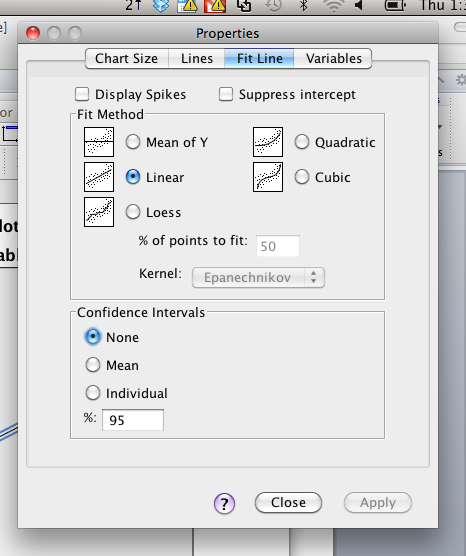




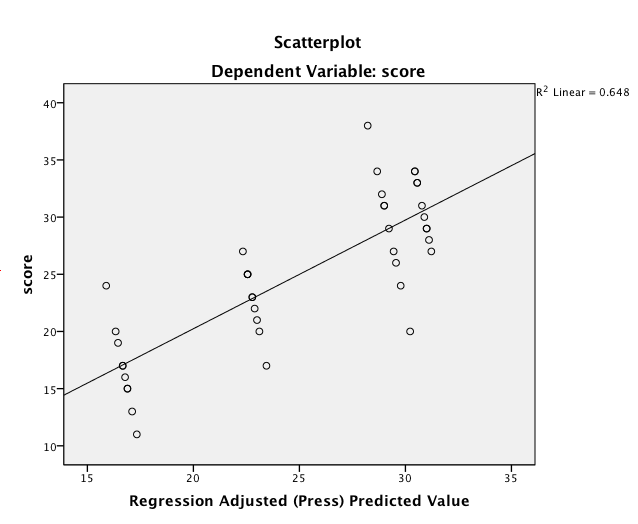
1. How to get the regression line on the graph.
2. Double click on the graph, which will bring up the chart editor.



1. Click the “add fit line at total button”. (note: the DV has to be scale for you to get this option on the variable view > measure column).
2. When the design box comes up, hit close. Then you can also close the chart editor menu.



1. Final Graph.



Clearly you’d want to clean up the X and Y axis for this graph (I’d say Predicted Scores on the X probably, with whatever Y actually is on the Y axis … so the label of your DV).

This graph put our predicted values (Y-hat) versus our real values (Y) and shows you how well you are capturing what people are scoring.

# Complete Example Simple Linear Regression

**Data Set 1 on blackboard.**

**IV(s):**

* Books – the number of books a person reads.
* Attend – the attendance of a person in a course.

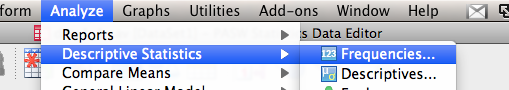
**DV:**

* Grade – final grade in the course.

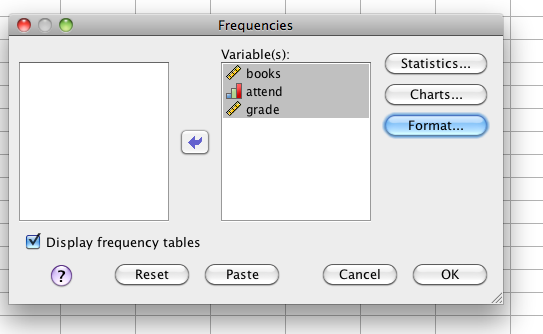
**Research Question:** Does the number of books predict a person’s final grade in a course?

Assumption Checks:

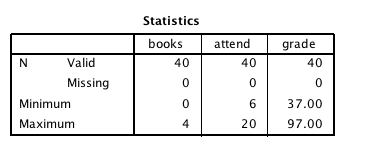
1. Missing data - any person with missing data on **any** variable will be excluded. You can fill in continuous data, but be careful with categorical data.
   1. Analyze > descriptive statistics > frequencies.



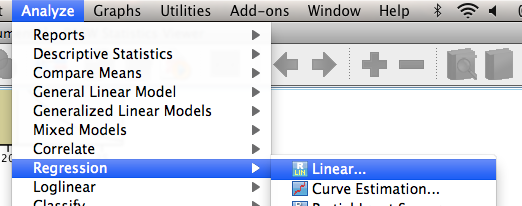
* 1. Move over all the variables to the right.



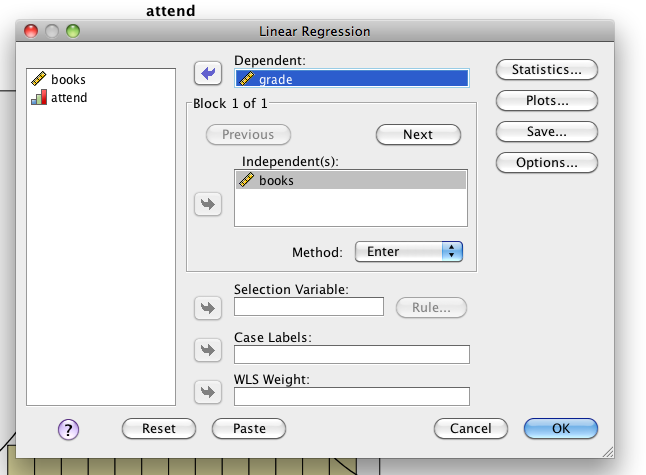
* 1. Uncheck frequency tables (with continuous variables you will get very large tables).
  2. Under statistics, check at least one box, so you get output.
  3. Look at the missing line under the Statistics box for missing data.



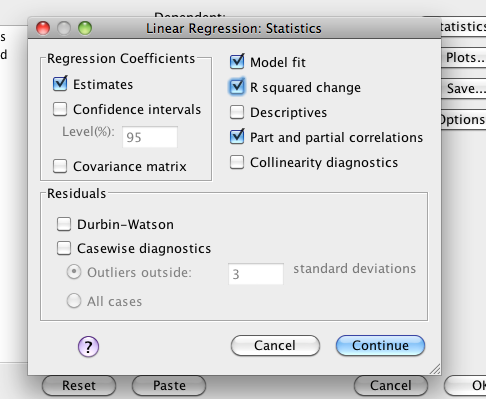
1. Outliers – outliers do **not** have to be run with a **fake** regression in this case. You can run your regression as you would for the analysis to get outlier (and below) analyses.
   1. Running - analyze > regression > linear.



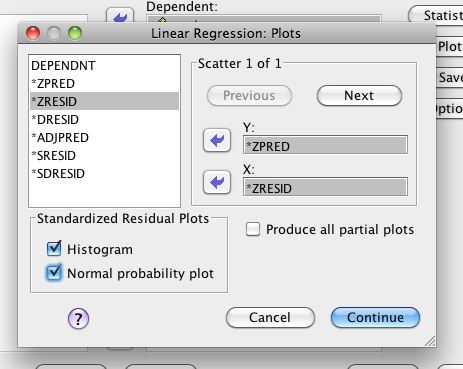
* + 1. Move over your **real** predictor IV into the independents box. Move over your **real** DV into the dependent box.



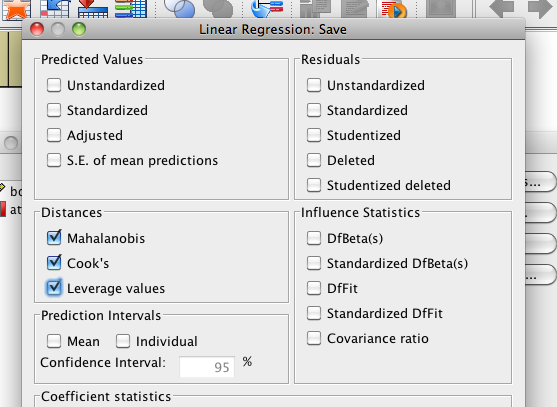
* + 1. **Note:** some of these steps are actually for running the real regression, but if you set it up the first time, you can hit analyze > linear > regression and just hit ok every time after that.
    2. Under statistics select R square changed and Part and Partial Statistics.



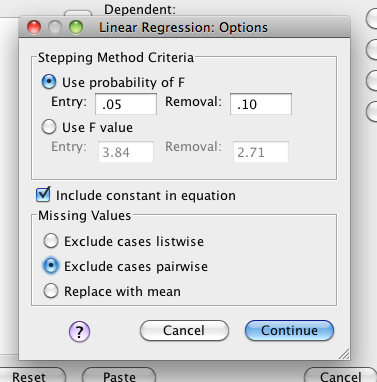
* + 1. Under plots put ZPRED into the Y box, ZRESID into the X box, check histogram, normal probability plot.



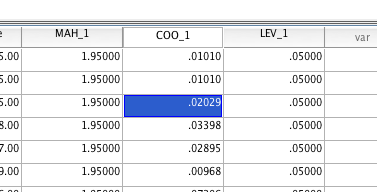
* + 1. Under Save, check Mahalanobis, Leverage, and Cooks.



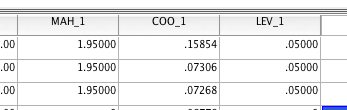
* + 1. Under options > exclude cases pairwise.



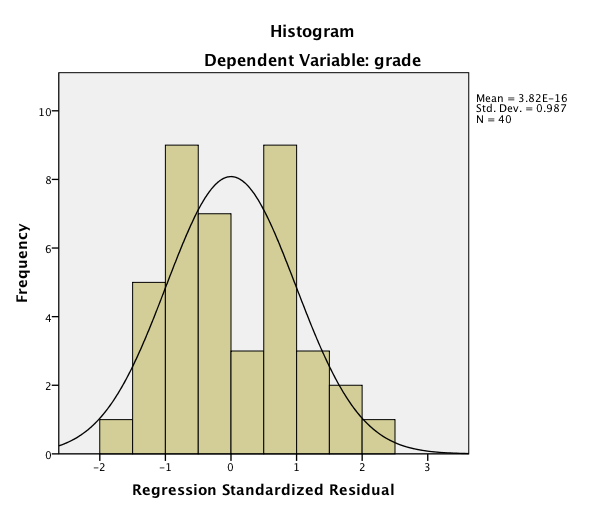
* + 1. When you hit ok, you will get a TON of output. First, you want to look at the outlier statistics. Go back to the dataset, where it created three new variables for you.
  1. Mahalanobis
     1. Cut off chi-square with 1 degrees of freedom (one X variable) p<.001, 10.83.
        1. Note some people suggest screening for both X and Y (because normally you would screen all the continuous, and X/Y are continuous). The addition of checking out cook’s and leverage helps us see more than just strange IV/DV combinations, so just screen for the IVs.
     2. Sort the data descending and check for values higher than that number.
     3. Nope!



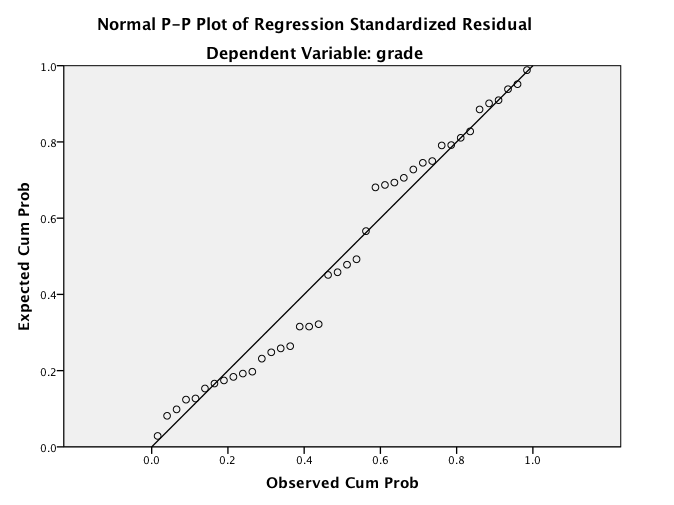
* 1. Leverage – (2k+ 2)/N
     1. K = 1 predictor
     2. N = 40 people
     3. Cut off = 4/40 = .10
     4. Nope!
  2. Cooks - 4/(N-K-1)
     1. 4 / (40-1-1) = 4/38 = .11
     2. Yes! – but since their value is not an outlier on leverage or Mahalanobis, I’ll leave them in. Mostly, people eliminate based on having two out of the three (sometimes only on Mahalanobis though).
  3. Remember the create new column rule.



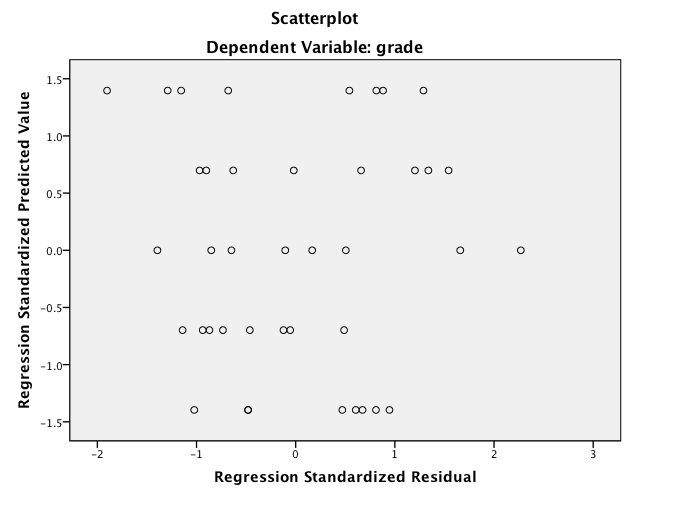
1. Multicollinearity – mostly for IVs that are too highly correlated. In this example, we only have *one* IV, so it’s not testable.
   1. To run > analyze > correlate > bivariate.
   2. Move the variables over to the right side.
   3. Hit ok.
   4. Look for Pearson’s r that is over .9.
   5. (will be shown below in other examples).
2. **NOTE:** 
   1. If you eliminated outliers – go back and rerun your regression.
      1. Analyze > Regression > Linear
      2. Hit Ok (you’ll get new outlier analyses but you can delete or ignore them). If you have a blank set up, see above on how to set up the analysis.
      3. This rerun will give you the new charts and graphs for the next assumptions **without** the outliers you eliminated.
   2. If you didn’t eliminate outliers – use the output you already have.
3. Normality – look at the histogram to see if it’s normal (notice that the DV now says the right DV name, rather than random).



1. Linearity – look at the PP Plot to make sure the dots line up on the line.



1. Homogeneity – look at the residual chart to make sure the spread is equal around zero.



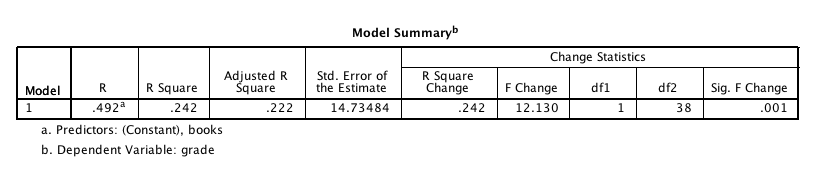
1. Homoscedasticity – look at the residual graph to make sure the dots are not bunched up on one part of the graph and spread out wide on the other part of the graph.
   1. Dashed lines above.
   2. No megaphone shapes.

**How to Run Analysis:**

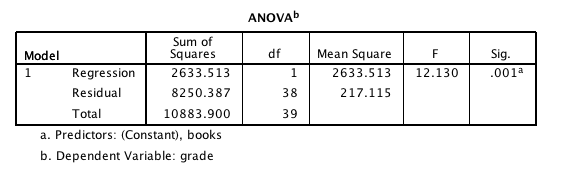
1. See above for pictures.
2. Analyze > regression > Linear.
3. Move the IV into the Independents box. Move the DV into the dependents box.
4. Under statistics, select R square changed and Part and Partial Statistics.
5. Under options, exclude cases pairwise.
6. Notes:
   1. Plots and Save options are more for assumptions checks. You can “turn them off” or just let them run again and ignore them.
   2. After you finish changing your variables due to missing data, outliers, multicollinearity, etc., you can rerun the same regression. If you don’t change anything, just use the output you already have.

**Reading the Output:**

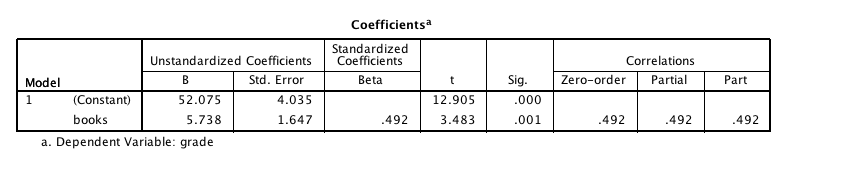
1. Model Summary – this will be slightly repetitive with only one variable. It always helps to ask for the output to get into the habit (because it’s very useful for later regression types).
2. The model summary contains the following:
   1. R – correlation between Y and Y hat (predicted values).
   2. R Square – the effect size you will want to report for your analysis. The total variance by all variables.
   3. Adjusted R Square – excludes some calculated error, sometimes journals asked for adjusted instead of R squared plain.
   4. Change statistics
      1. In SLR and simultaneous MLR, this box is a repeat of the ANOVA box (see information below).



1. ANOVA box.
   1. This box gives you the same information a regular ANOVA box gives you, only without all the corrected/intercept lines.
   2. Regression line – same as your “treatment” or group effect. This information is about your IV(s).
   3. Residual – error line in a regular ANOVA.
   4. You will use this information to report how well your regression line predicted.
      1. F(regression df, residual df) = F, *p*= Sig value, R2 = see above.
      2. *F*(1, 38) = 12.13, *p*=.001, *R2*=.24
      3. *Note*. This information is the same as the change statistics above.



1. Coefficients box – the last part of the information you will need to report for your write up.
   1. Constant line (going across in order)
      1. B box = a or the constant added to the regression equation.
      2. Std. Error = SE.
      3. Blank in the Std Coefficients box because the standardized regression equation does not have a constant.
      4. t = t-test to see if this value is different from zero.
      5. Sig = p value if this value is different from zero.
      6. Most of this information is completely ignored (depending on the research question).
   2. Books line (IV predictor information).
      1. B box = b for books predicting grades. This b value is the slope of the regression equation. It means that for every 1 book, grades go up by 5.7 points.
      2. Std Error = SE.
      3. Beta = standardized slope for your equation. This number is the usual one to report. For every 1 standard deviation in books, there is a .5 standard deviation increase in grades.
      4. t = the t-test to see if beta is different from zero. **DF is the error from your ANOVA.**
      5. Sig = p value to see if beta is different from zero.
      6. *t*(38) = 3.48, *p* = .001, *pr*2 = .24
   3. Correlations
      1. Zero-order = normal Pearson’s r for books to grades.
      2. Partial = partial, pr, for books to grades.
      3. Part = semi-partial, sr for books to grades.
      4. In SLR these are all the same.



Write up Example:

**Results**

The number of books a student read per semester was used to predict overall course grade. The data were screened for missing data, outliers and regression assumptions. Although one participant met the cut off for Cook’s values, they were included in the analysis due to not being an outlier using Mahalanobis distance. Linearity, normality, multicollinearity, homogeneity, and homoscedasticity were all met. Number of books significantly predicted overall course grade, *F*(1, 38) = 12.13, *p*=.001, R2=.24. Course grade improved approximately 6 points for every book a student read, β=.49, *t*(38) = 3.48, *p*=.001, *pr*2 = .24.

# Complete Example Simultaneous Multiple Linear Regression

**Data Set 1 from blackboard**

**IV(s):**

* Books – the number of books a person reads.
* Attend – the attendance of a person in a course.

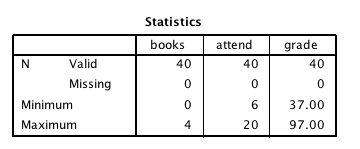
**DV:**

* Grade – final grade in the course.

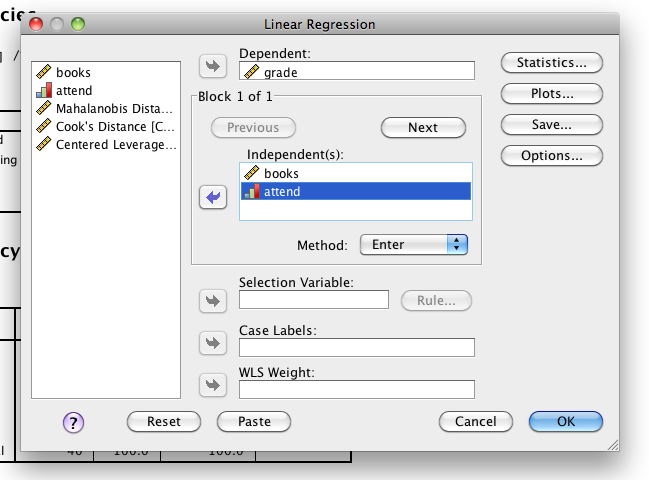
**Research Question:** Do attendance and books **both** predict overall course grade?

**Assumption Checks**: Since assumption checks are the same for all types of regression analyses, please see above for how to run this assumption check.

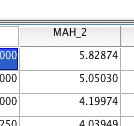
1. Missing data:



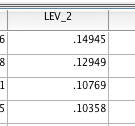
1. Outliers – note how BOTH IVs are in the independents box.



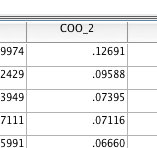
1. Mahalanobis cut off X2(2) *p* < .001 = 13.82.



* 1. Leverage = 6/40 = .15

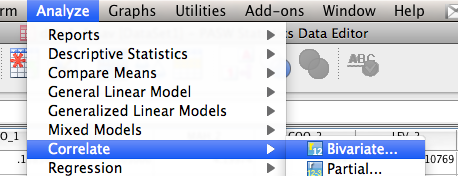


* 1. Cooks = 4/37 = .11

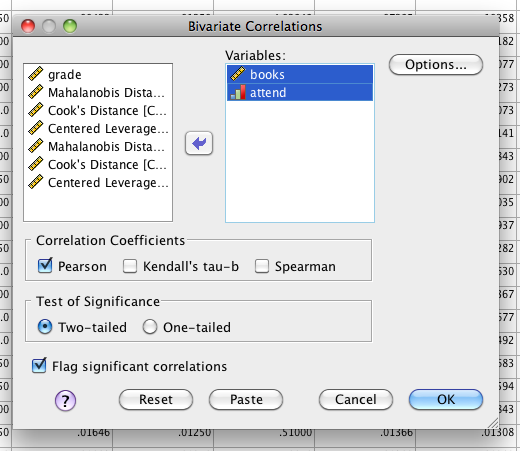


* + 1. Since this person was only outside Cook’s range, I’m going to leave them in.

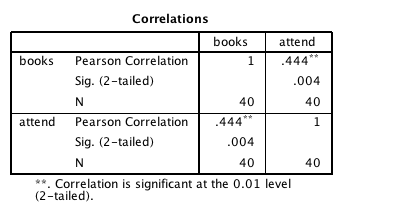
1. Multicollinearity
   1. Analyze > correlate > bivariate



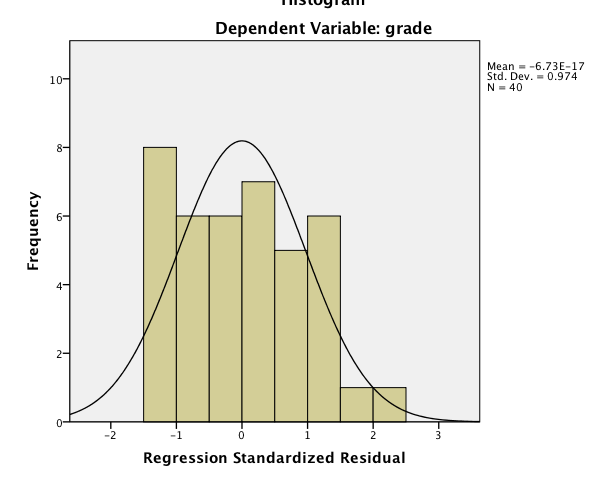
* 1. Move all the independent variables over (not the DV!)



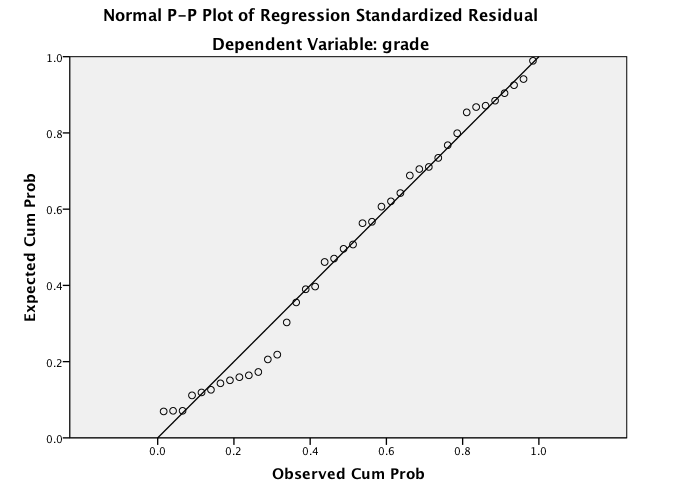
* 1. Make sure nothing is over .9



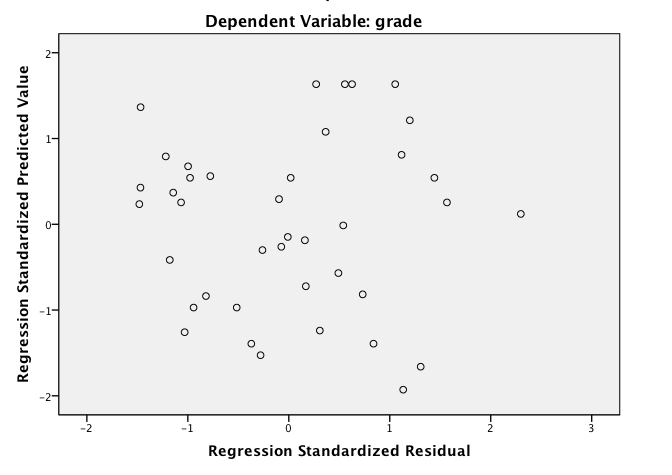
1. Normality



1. Linearity



1. Homogeneity/Homoscedasticity



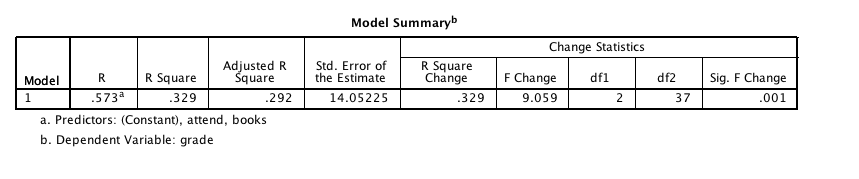
**How to Run Analysis:**

(see above for pictures).

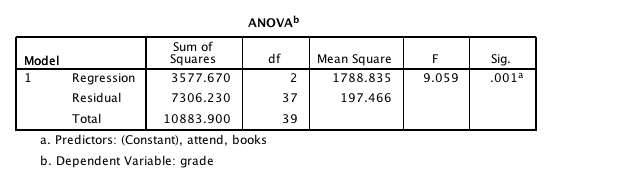
1. Analyze > regression > Linear.
2. Move the IVs into the Independents box. Move the DV into the dependents box.
3. Under statistics, select R square changed and Part and Partial Statistics.
4. Under options, exclude cases pairwise.

**Reading the Output:**

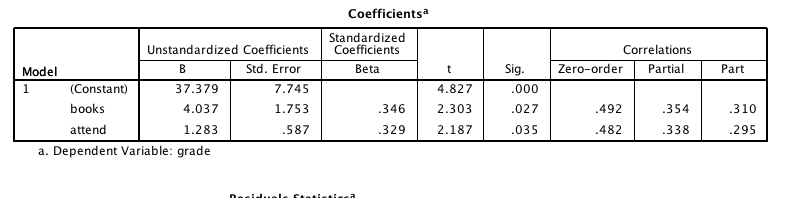
1. The model summary contains the following:
   1. R – correlation between Y and Y hat (predicted values).
   2. R Square – the effect size you will want to report for your analysis. The total variance by all variables.
   3. Adjusted R Square – excludes some calculated error, sometimes journals asked for adjusted instead of R squared plain.
   4. Change statistics
      1. In SLR and simultaneous MLR, this box is a repeat of the ANOVA box (see information below).



1. ANOVA box.
   1. Regression line – same as your “treatment” or group effect. This information is about your IV(s).
   2. Residual – error line in a regular ANOVA.
   3. You will use this information to report how well your regression line predicted.
      1. F(regression df, residual df) = F, *p*= Sig value, R2 = see above.
      2. *F*(2, 37) = 9.06, *p*=.001, *R2*=.29
      3. *Note*. This information is the same as the change statistics above.



1. Coefficients box
   1. Constant line (going across in order)
      1. B box = a or the constant added to the regression equation.
      2. Std. Error = SE.
      3. Blank in the Std Coefficients box because the standardized regression equation does not have a constant.
      4. t = t-test to see if this value is different from zero.
      5. Sig = p value if this value is different from zero.
      6. Most of this information is completely ignored (depending on the research question.
   2. Books and grades line (IV predictor information).
      1. B box = b for predictors predicting grades. This b value is the slope of the regression equation.
         1. Books = for every 1 book, there’s an increase in grades by 4 points
         2. Attend = for every 1 attend, there’s an increase in grades by 1.3 points.
      2. Std Error = SE.
      3. Beta = standardized slope for your equation. This number is the usual one to report.
         1. Books – for every 1 SD books, .346 SD grades.
         2. Attend – for every 1 SD attend, .329 SD grades.
         3. Since this is standardized, you can tell which one is strongest, by which value is largest.
      4. t = the t-test to see if beta is different from zero. DF is the error from your ANOVA.
      5. Sig = p value to see if beta is different from zero.
      6. Books *t*(37) = 2.30, *p* = .03, *pr2=*.13
      7. Attendance *t*(37)=2.19, *p*=.04, *pr2=*.11
      8. (so books is better with a higher pr as well as a higher beta)
   3. Correlations
      1. Zero-order = normal Pearson’s r for predictor to grades.
      2. Partial = partial, pr, for predictor to grades.
      3. Part = semi-partial, sr for predictor to grades.
      4. You can also tell which variable is the best by which one has the highest pr value.
      5. These are effect sizes for each predictor individually.



Write up MLR

**Results**

The number of books a student read per semester and their overall attendance in the semester was used to predict final course grade. The data were screened for missing data, outliers and regression assumptions. Although one participant met the cut off for Cook’s values, they were included in the analysis due to not being an outlier using Mahalanobis distance. Linearity, normality, multicollinearity, homogeneity, and homoscedasticity were all met.

The overall regression model was significant, indicating the books and attendance combined predicted final course grade, *F*(2, 37) = 9.06, *p*=.001, R2=.33. As students read more books throughout the semester, they were more likely to increase their course grade, β = .35, *t*(37)=2.30, *p*=.03, *pr2=*.13. Attendance was also a significant predictor of course grade, so that students who attended class more had higher grades, β = .33, *t*(37)=2.19, *p*=.04, *pr2=*.11.

# Complete Example Hierarchical Multiple Linear Regression

**Data Set 2 from blackboard**

**IV(s):**

* Sex – gender of the participant (0 = female, 1 = male)
* Age – age of the participant
* Extro – extroversion of the participant, low numbers are introverted, high numbers are extroverted.

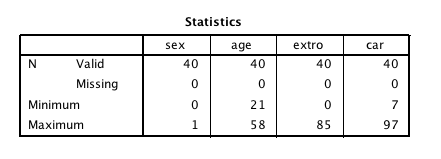
**DV:**

* Car – how well a person takes care of their car (regular washes, cleaned, oil changed, etc.).

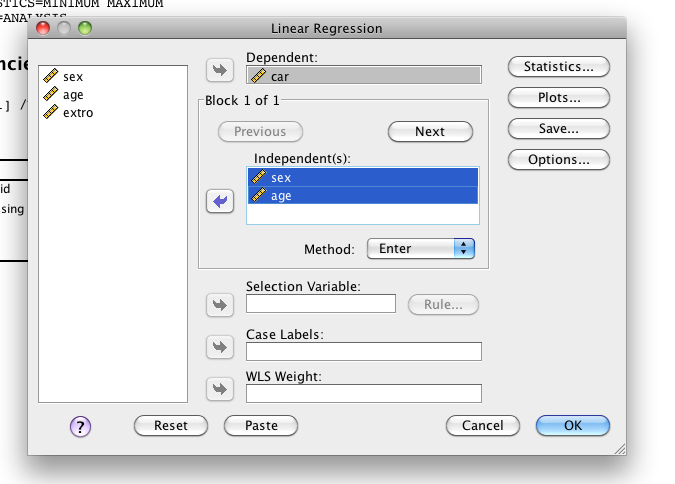
**Research Question:** After controlling for demographic variables, does the extroversion of the participant predict how well they take care of their car?

**Assumption Checks:** Since assumption checks are the same for all types of regression analyses, please see above for how to run this assumption check.

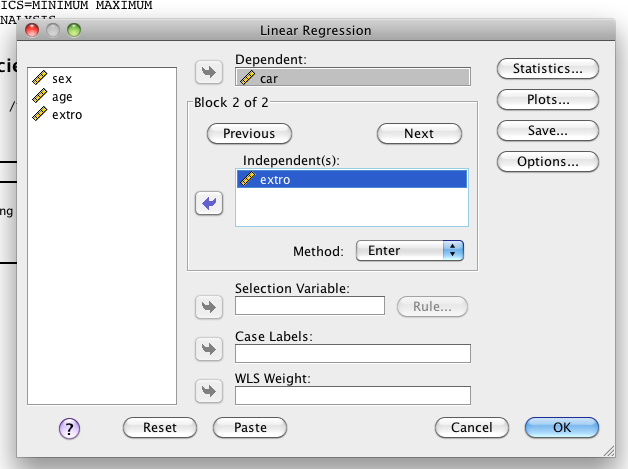
1. Missing data – none.



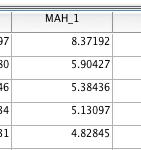
1. Outliers – this analysis is a little different because you will have “steps” of variables.
   1. First, move over your two IVs (control demographic variables) into the independents box. Move the DV over into the dependent variable box.



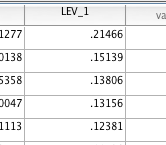
* 1. Then hit NEXT.
  2. Move over the second IV (or set of IVs) into the independents box. Notice how the DV stays at the top.



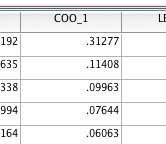
* 1. Use the same options from above analyses.
     1. Statistics – R square change, part and partial correlations.
     2. Plots – ZPRED in Y, ZRESID in X, Histogram, Normal PP Plot.
     3. Save – Mahalanobis, Cooks, Leverage
     4. Options – exclude cases pairwise.
  2. Mahalanobis – Chi square (3 df) p<.001 = 16.27\*\* note that sometimes people will pre-screen *without* the categorical variable.



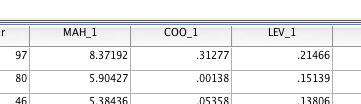
* 1. Leverage = (2\*3 predictors+ 2) / N (40 people) = 8/40 = .20



* 1. Cook’s = 4 / (40 – 3 – 1) = 4/36 = .11

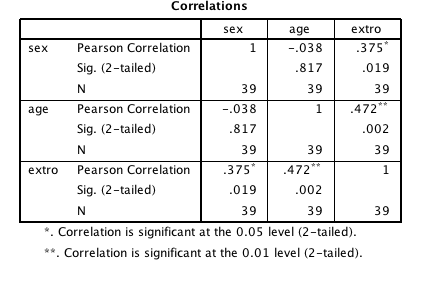


* 1. Outliers that have 2 out of the 3

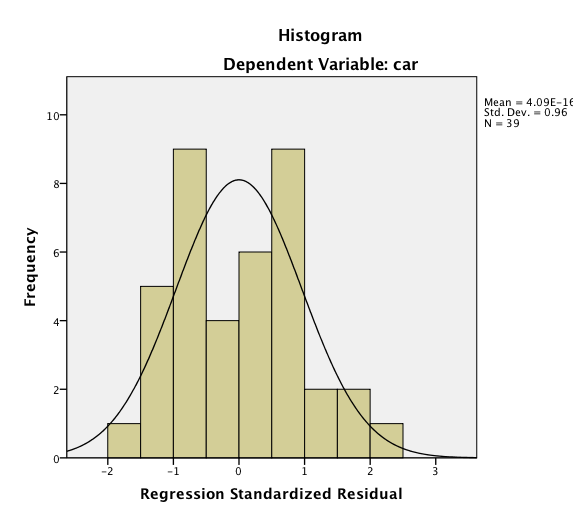


* 1. This case has problems with cooks and leverage, therefore we should get rid of them.

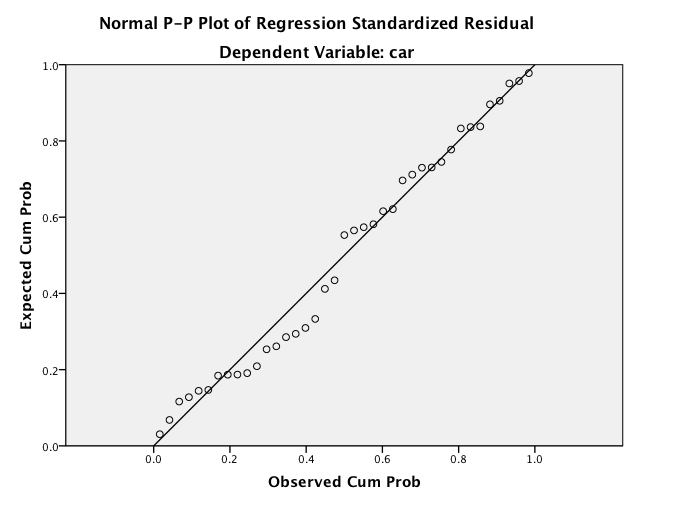
1. Multicollinearity – bivariate correlations check out ok.



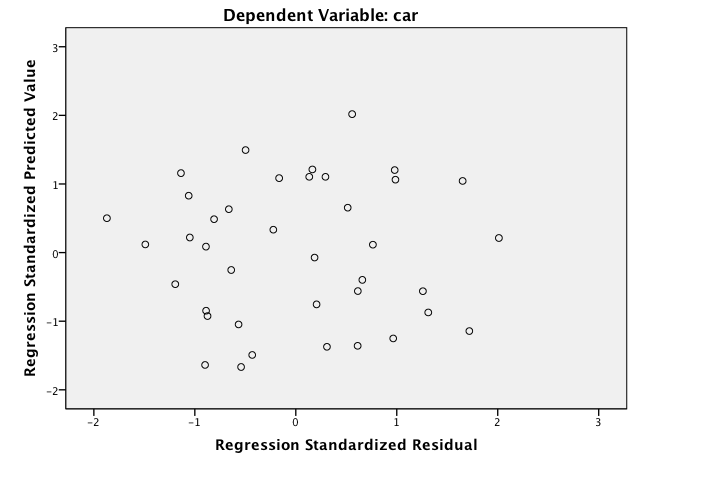
1. Note! I’m going to need to rerun my regression because I deleted an outlier.
2. Normality – checks out.



1. Linearity – checks out.



1. Homogeneity/ Homoscedasticity – checks out.

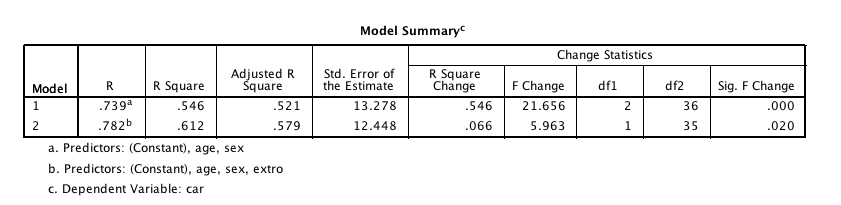


**How to Run Analysis:**

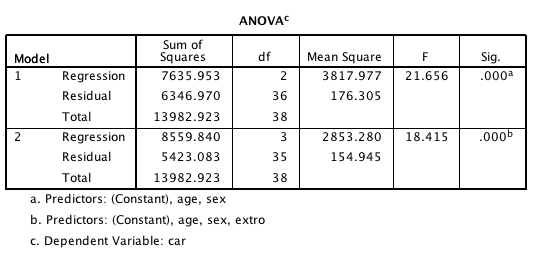
1. Repeat from above:
   1. First, move over your two IVs (control demographic variables) into the independents box. Move the DV over into the dependent variable box.
   2. Then hit NEXT.
   3. Move over the second IV (or set of IVs) into the independents box. Notice how the DV stays at the top.
2. Use the same options from above analyses.
   1. Statistics – R square change, part and partial correlations.
   2. Plots – ZPRED in Y, ZRESID in X, Histogram, Normal PP Plot.\*
   3. Save – Mahalanobis, Cooks, Leverage\*
   4. Options – exclude cases pairwise.
      1. \* you can leave off if you don’t want the numbers again.

**Reading the Output:**

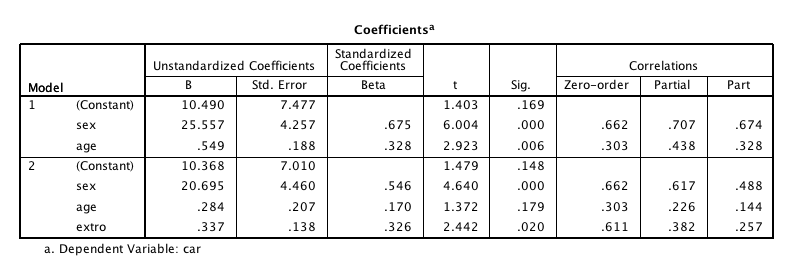
1. Model Summary
   1. R Square – the effect size you will want to report for your analysis. The total variance by all variables.
   2. **Change statistics – this section is important for a hierarchical MLR.**
      1. You get one line for each “model” or each step of the hierarchical MLR.
      2. When you report each step, you will use the R squared change to report how much each step “added” to the equation.
      3. Each model is a “set” of variables. So in the first step, the demographics set (age and sex) accounted for 55% of the variance.
         1. *F*(2, 36) = 21.66, *p* <.001, R2 = .55
      4. When we added extroversion, another 6.6% of the variance was included (which was significant, and then you get F change information).
         1. Δ*F*(1, 35) = 5.96, *p*=.02, ΔR2 = .07. (delta means change).



1. ANOVA box
   1. The tricky part about this section is that each ANOVA tells you if the *overall regression* is significant. Therefore, it is testing if regular R2 is different from zero.
   2. It does NOT test if the new variables in each step are useful, just if overall you are greater than zero. So you can get into the case where the change statistics in the model summary are NOT significant (meaning that the variable or set did not add anything to help prediction), but the ANOVA statistics are significant (meaning the model including step 1 is still useful). Most people interpret that using the change statistics.



1. Coefficients box
   1. You want to talk about the variables in in the step you first used them in.
   2. So, here, we would talk about the demographic set using R2 from above and then each individual predictor.
      1. Sex was a significant predictor (*t*(36) = 6.00, *p*<.001, *pr2* = .50).
      2. Age was a significant predictor (*t*(36) = 2.92, *p*<.01, *pr2* = .19).
         1. Notice *df* = 36 because we are talking about step 1 – match the df to the step you are talking about them.
   3. Now we want to talk about the addition of extroversion in step 2.
      1. We will NOT mention the new values for sex and age. This reporting style varies by journal. Generally, you want to talk about the addition of extro, so you ignore sex and age because you’ve already talked about them.
      2. Extro was a significant predictor (*t*(35) = 2.44, *p*=.02, *pr2* = .15).
         1. Notice how the addition of R2 was 7% but the variable accounts for 15% of unaccounted for variance (that’s why people like *pr*).



Write Up Example:

**Results**

Age, gender, and extroversion were used to predict a person’s overall care for their car. The data were screened for assumptions, and one participant was eliminated due to high Cook’s and Leverage values. Linearity, normality, multicollinearity, homogeneity, and homoscedasticity were all met.

Age and gender were entered first into a hierarchical regression to control for demographic differences in care maintenance. Overall, this model was significant, indicating that demographics predict how much a person takes care of their car, *F*(2, 36) = 21.66, *p*<.001, R2= .55. Gender was a stronger predictor of car maintenance, β=.68, *t*(36) = 6.00, *p*<.001, *pr2* = .50, which showed that males are more likely to take care of their cars by regular maintenance. Age was also positively related to car maintenance, β=.33, *t*(36) = 2.92, *p*<.01, *pr2* = .19; therefore, older participants indicated better car care. Next, extroversion was added in a second step to examine it’s predictive value after controlling for demographic variables. The addition of this variable was significant, *F*(1, 35) = 5.96, *p*=.02, ΔR2 = .07. Participants who were more extroverted took better care of their cars, β=.33, *t*(35) = 2.44, *p*=.02, *pr2* = .15.

Where did you get all those numbers?

* The overall model significance from step 1 can be found in the change statistics model summary OR the ANOVA box for model 1.
* pr2 = is the *partial* column, squared.
* Degrees of Freedom for t-tests are the second DF for your F test (either the change statistics or the ANOVA box residual DF).
* The second step significance is found in the change statistics model summary section.

# Install the PROCESS plug-in

<http://www.afhayes.com/introduction-to-mediation-moderation-and-conditional-process-analysis.html>

Go about half way down the page to this part:

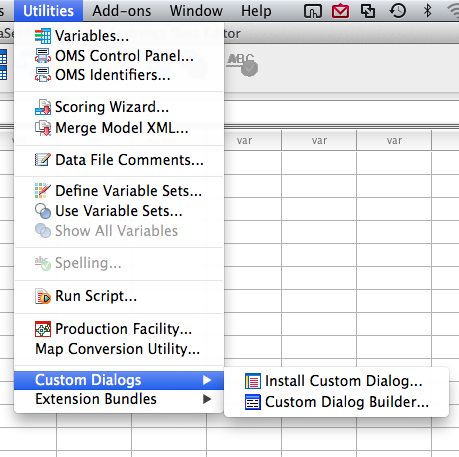


Click to download the file.

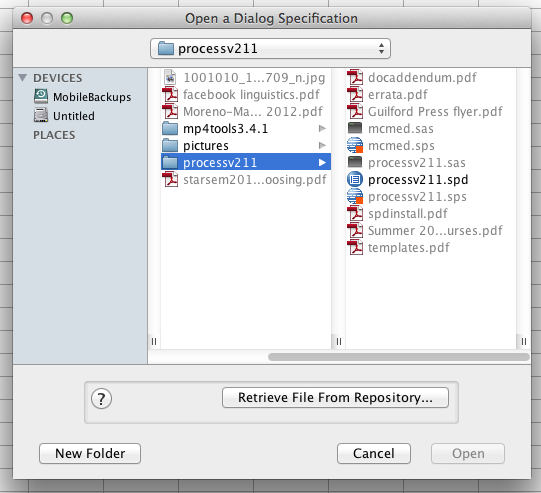
Unzip the file to somewhere on your computer that you are going to remember where it is (like the desktop or documents folder).

Go to SPSS.

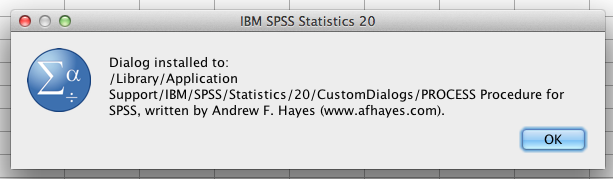
Utilities > custom dialogs > install custom dialog



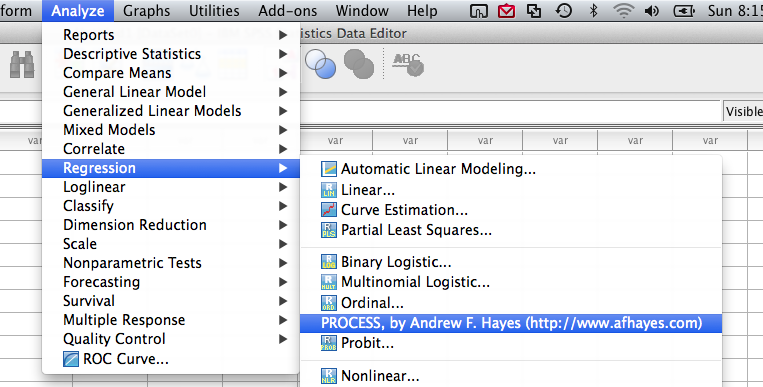
Find the .spd file in the window.



Click open, and the process plug in will install.



The look under analyze > regression to make sure it properly installed.



# Moderation

*Note*. We are only going to practice interactions with two continuous variables, but the old notes attached online include two categorical, half-half, and two continuous variables.

The process plug-in will not run the same type of assumption checks we’ve been doing with a real regression. You will want to screen the data with the variables you have in a regression analysis before you start. That means you will run a real regression with IVs and the DV, but you wouldn’t interpret the regression output because you would instead run the process plug in.

**Data set 1 on blackboard.**

**IVs:**

* Books – number of books a person read
* Attend – attendance in the class
* Interaction (created for you with the plug-in)

**DVs:**

* Grade – final course grade

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

**SPECIAL INSTRUCTIONS:**

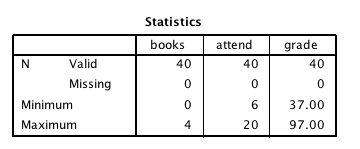
* Centering: When working with two continuous variables in regression, you have to center them first. Centering means that you start with ZScores instead of the regular variables.
* Why? When you create an interaction, you are creating multicollinearity. Z-scoring helps.
  + Also, centering the variables creates SDs of 1 and a mean of zero.
  + That means you can write out the equation and understand what the slopes mean. In English:
  + When variables are NOT centered – the 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 values will match your pictures.
* PROCESS will center the variables for you.

**Interaction:**

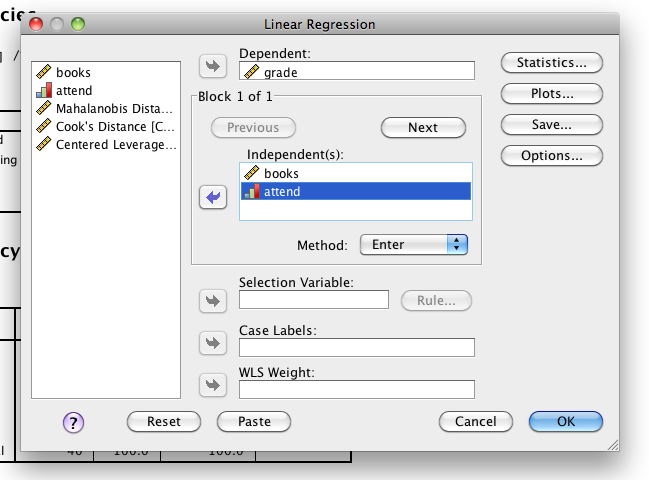
* PROCESS will also create the interaction for you.
* The interaction is created by multiplying the scores of each variable together (like participant one book times participant one grades) = participant one interaction score.
* The interaction is created after the variables are centered, so the interaction is centered as well.

**Assumption Checks:** Since assumption checks are the same for all types of regression analyses, please see above for how to run this assumption check.

1. Missing data – does not appear that we have any.

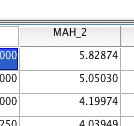


1. Outliers – note how BOTH IVs are in the independents box.

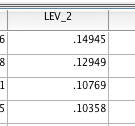


(so you won’t have an interaction yet here because we are just screening the variables before running the moderation analysis).

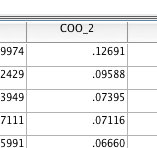
1. Mahalanobis cut off X2(2) *p* < .001 = 13.82.



* 1. Leverage = 6/40 = .15

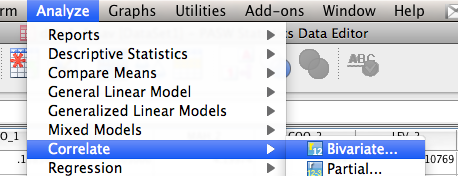


* 1. Cooks = 4/37 = .11

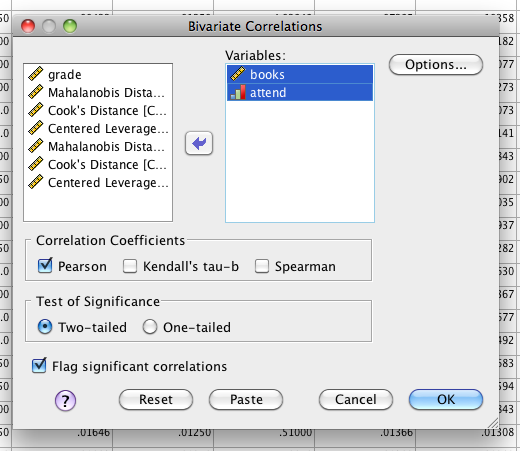


* + 1. Since this person was only outside Cook’s range, I’m going to leave them in.

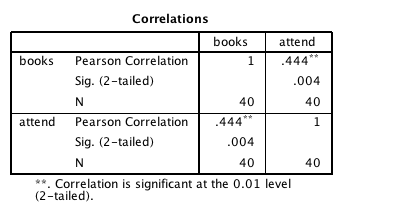
1. Multicollinearity
   1. Analyze > correlate > bivariate



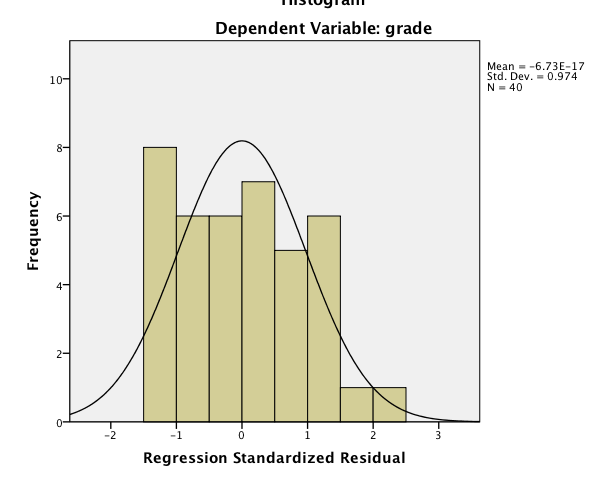
* 1. Move all the independent variables over (not the DV!)



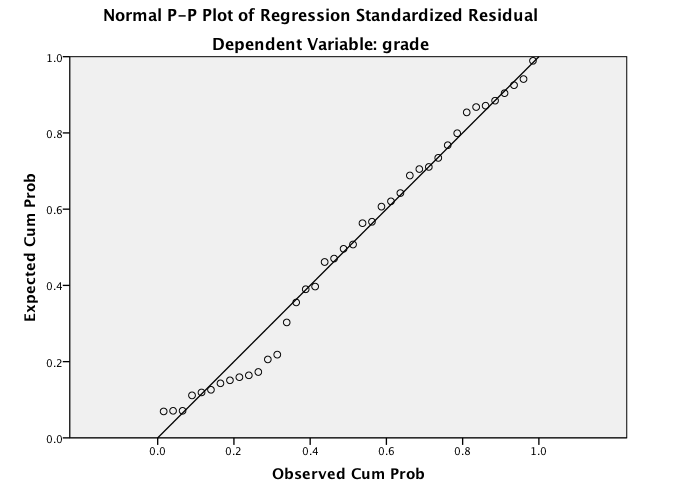
* 1. Make sure nothing is over .9



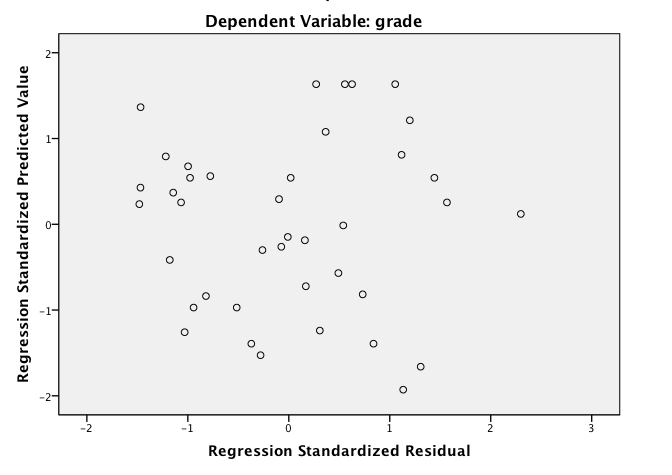
1. Normality



1. Linearity

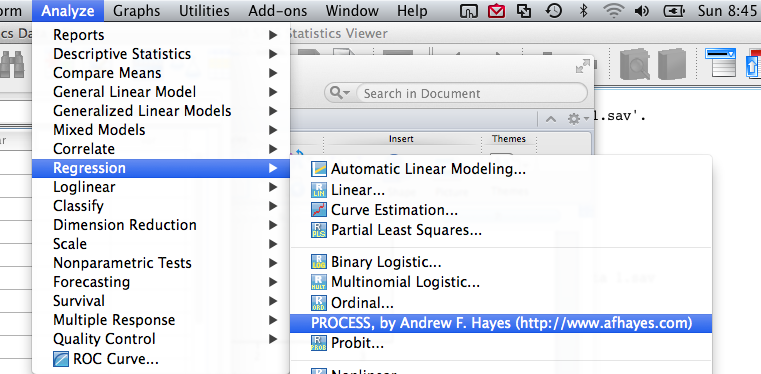


1. Homogeneity/Homoscedasticity

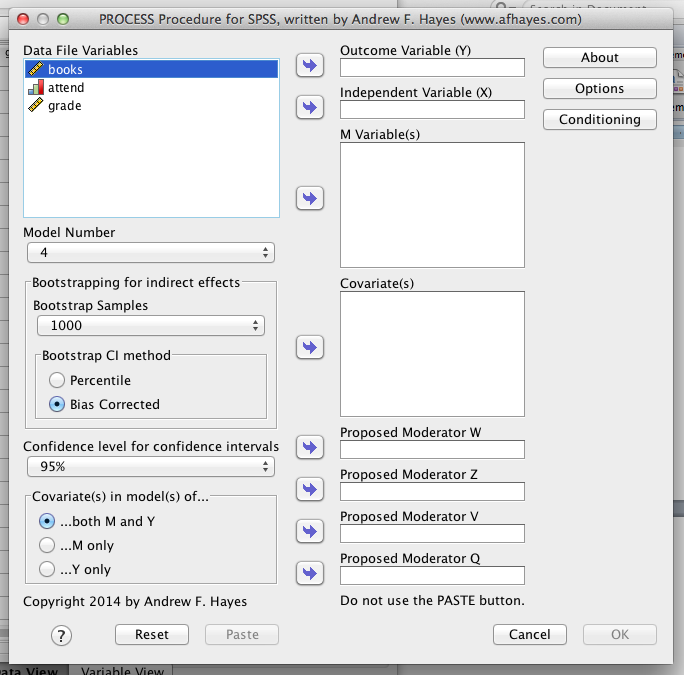


**How to run the analysis in PROCESS:**

Analyze > regression > PROCESS



Here is the PROCESS window.



Outcome variable = dependent variable

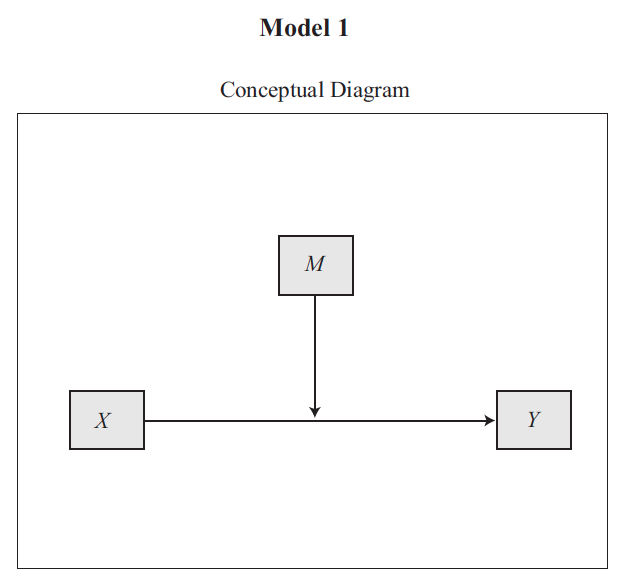
Independent variable = one of your independent variables.

M variables = moderator variables = your other independent variable.

Model number (IMPORTANT!).

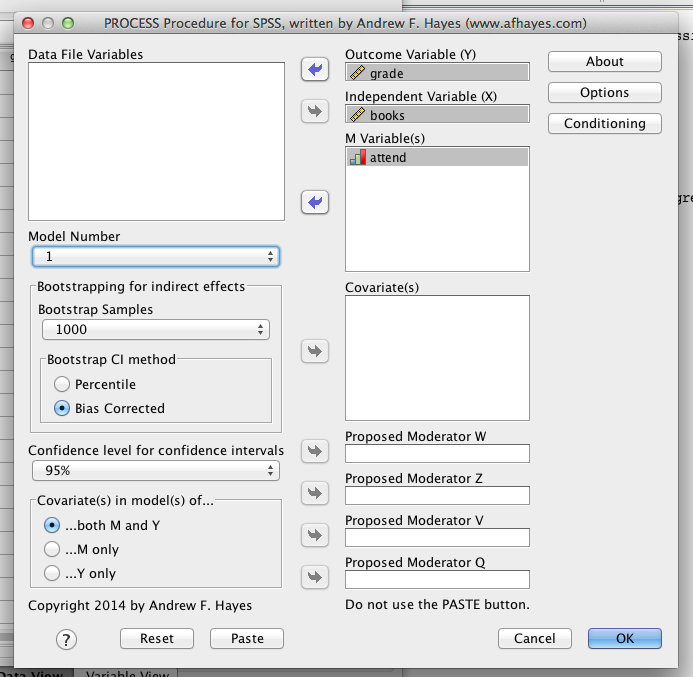
Go to the template.pdf you have in the process folder – it has all the model pictures in it.

Model for moderation = model 1



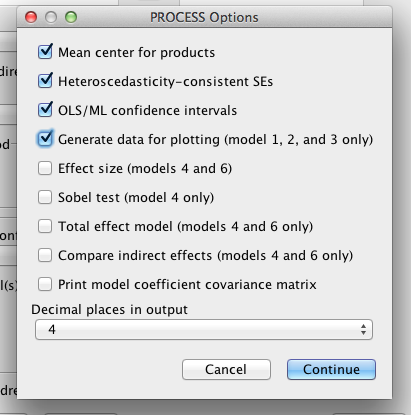
So put in grade in the output variable, books in the independent variable, attend in the m variables (you could switch the books/grades).

Change the model number 1.

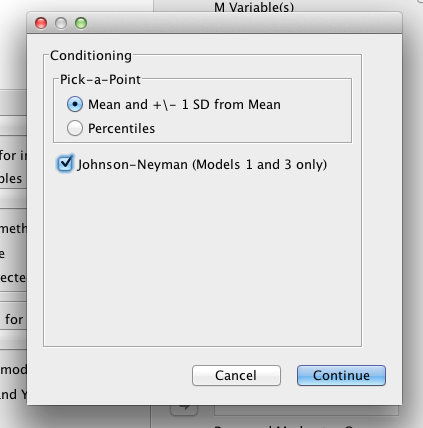


Click options, for moderation you will want to click the first four options.

* Mean center for products = centering of the variables.
* Heteroscedasticity consistent SEs = bootstraps the standard error, means they are better estimated because they have been estimated 1000 times (see above).
* OLS/ML confidence intervals = ordinary least squares/maximum likelihood confidence intervals, type of estimation for confidence intervals since it runs a lot of bootstrapped samples.
* Generate data for plotting = creates the numbers needed to create a graph for you.



Click conditioning.



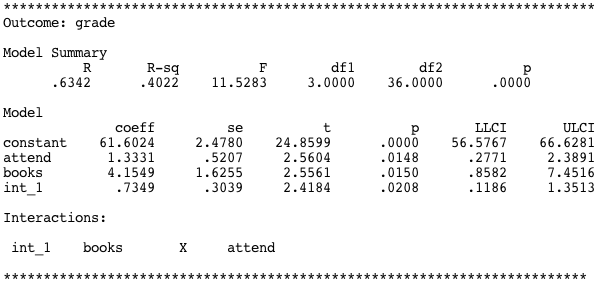
You want to leave mean and +/- 1SD mean – that’s the normal way to create groups for an interaction.

Click on the Johnson-Neyman zones of significance.

**Output:**



First part gives you the variables you had entered into the equation.



Model summary = the same box as the first model summary box in a regular regression.

The overall model including all three variables (attend, books, and interaction) is significant, *F*(3, 36) = 11.53, *p* < .001, *R2 =* .40.

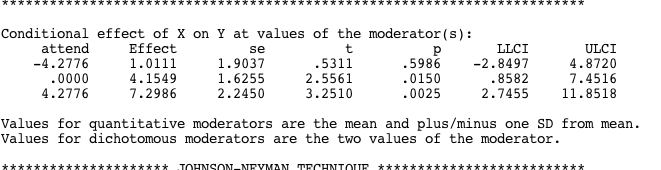
The next model part is the coefficients box.

* Coeff = unstandardized b (little b).
* SE = standard error.
  + You don’t get Beta in this output … if you want beta, you can do beta = coeff / SE.
* t = the t value for each coefficient, remember that the *df* is the same as the second *df* in the ANOVA.
* *p* = the p value.
* LLCI, ULCI = confidence interval for the unstandardized coefficient.
  + You do NOT want this value to cross zero. That implies that the coefficient is not different from zero, which would be non-significant/important.
* One bad thing about this output is that you do not get *pr*. So you will not really have an effect size for moderation (on each coefficient, you can get the overall).

So from this output:

* Attendance is significant, b = 1.33, *t*(36) = 2.56, *p* = .01
* Book is significant, b = 4.15, *t*(36) = 2.56, *p* = .02
* The interaction is significant, b = .73, *t*(36) = 2.42, *p* = .02

Same rules as ANOVA apply – mostly you are just going to analyze the interaction.



Ok, so what’s going on here? This box is the analysis of the interaction.

First column = the moderator groups. To be able to understand the interaction, we’ve broken it down the continuous variables into chunks. The chunks are:

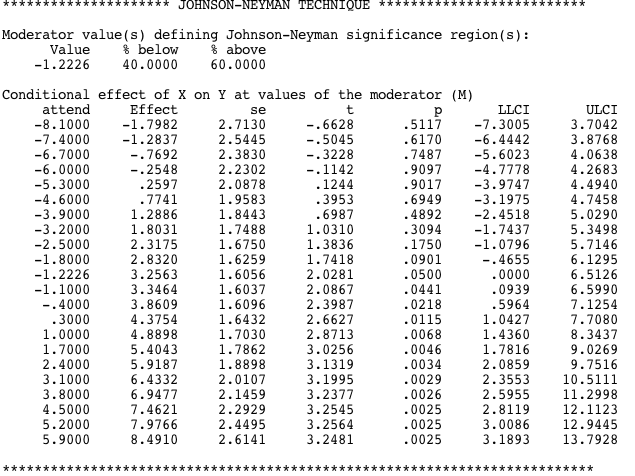
* Low = 1 standard deviation below the mean (the standard deviation depends on the variable … so here the SD = 4.28 points for attendance).
* Average = 0 … so we centered the variables, which means that zero is the average of the variable.
* High = 1 standard deviation above the mean.

The purpose of breaking it down is just to see what the interaction is – to help you visualize the interaction.

Second column is the coefficient for that individual slope, i.e. low attendance group slope for books predicting grade; average attendance group for books predicting grade; high attendance group for books predicting grade.

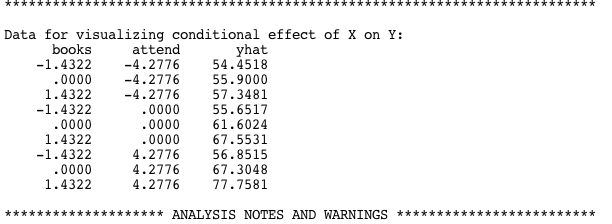
(the rest of the columns are the same as above).

In this example, the low attendance group is not significant, but the average and high attendance groups are significant.

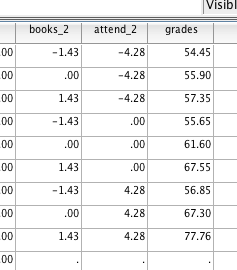


The Johnson-Neyman technique finds the spot were p = .05, and works in intervals to find the areas that are significant and not. For this example:

* Attendance is -1.22 standard deviations below the mean, below that there is no relationship between books and grade.
* Above this point, there is a significant relationship between books and grades, where the effect gets slowly larger (because the coefficients get larger).



Data for graphing! Put these values into SPSS.



Graphs:

Graphs > Chart Builder

Click line graphs.

Double click the second line graph option (multiple line).

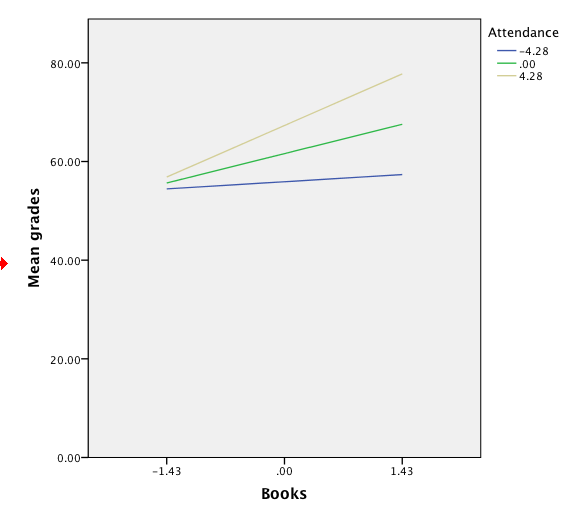
Note: the IVs will need to be labeled as nominal in SPSS.

Normally you will stick the moderator in the “set color” option.

Stick the other IV in the x-axis.

Stick the DV in the y-axis.

Note: you will not be able to do error bars because we have a summary of the data, rather than the actual data.



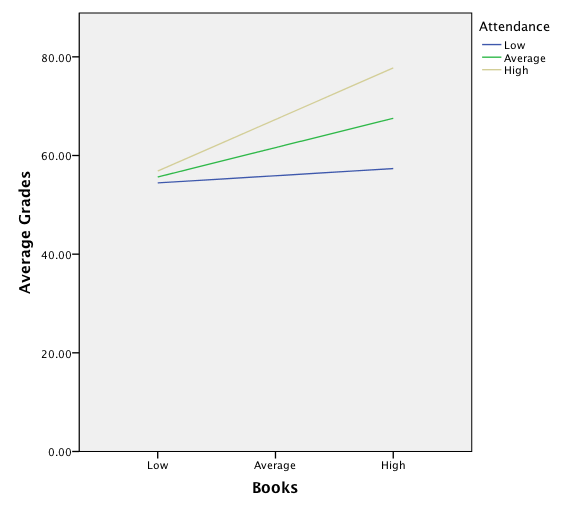
To clean up the graph, you want to change the numbers to low, average, high (for books and for attendance).

Write up Example

**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 PROCESS plug-in (Hayes, 2013) 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* = .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, *t*(36) = 2.56, *p*=.02. For low attendance, there was a non-significant difference in scores when reading more books, *b* = 1.01, *t*(36) = .53, *p*=.60. Finally, high attending participants showed the largest increase when reading more books, *b* = 7.30, *t*(36) = 3.25, *p* < .01.



*Figure* *1.*

# Mediation

Please note that these steps are the traditional Baron and Kenny (1986) steps designed to instruct the basic idea of mediation. which the PROCESS plug-in gives you automatically. Many modifications and suggestions for mediation are now available (a lot of these are implemented with process). 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 Baron and Kenny (1986) steps used to be c, a, b, c’ to test the paths, but the PROCESS output gives you the paths as a, b, c’, c in alphabetical order.

**Data set 6 on blackboard**

IV: Treatment condition – treated or control groups for housing

Mediator: Housing contacts

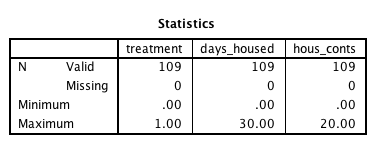
DV: Days housed

**Research question:** Does the number of housing contacts mediate the relationship between treatment condition and the number of days housed?

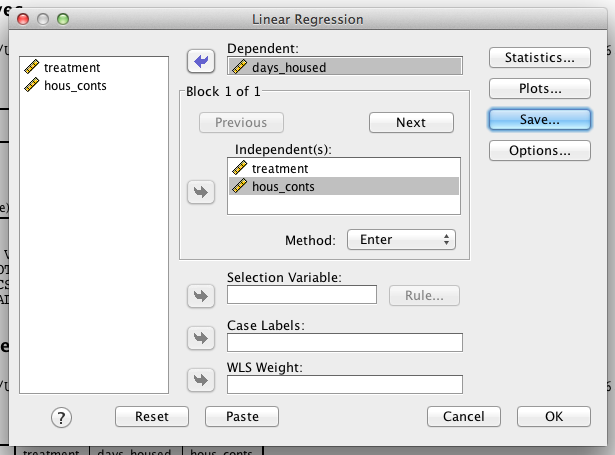
*Note*. As there are many examples on how to run everything above, this example does not have complete walk through steps. See examples above if you need help running the steps in SPSS.

**Assumptions:** Again, you will not get the assumptions running through PROCESS. You will screen everything as if both variables are in the equation (mediation includes steps with only one variable).

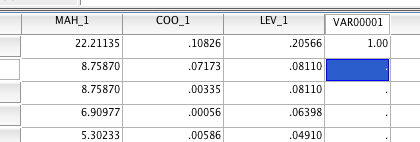
1. Missing data – no missing data is present in this dataset.



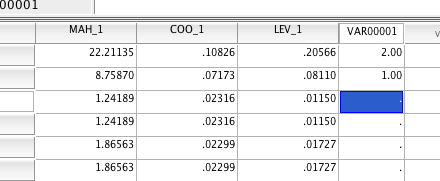
1. Outliers.
   1. Remember – you do not need to run a fake regression.
   2. Analyze > regression > linear.
   3. Put the DV in the dependent box, the IV AND MEDIATOR in the independent box.
      1. This step will not match the actual tests for mediation. However, you WANT the mediation, so you want to screen all the variables that are involved in the outcome you wish to happen.



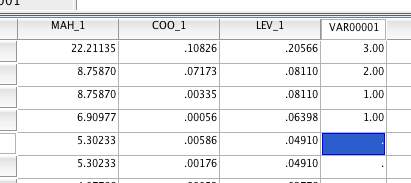
* 1. Plots > zpred Y, zresid X, histogram, normal probability plot
  2. Save > mahalanobis cooks leverage
  3. Check the values:
     1. Mahalanobis – 2 variables = 13.82
        1. Found one outlier



* + 1. Cooks = (4 / (109 – 2 – 1) ) = 0.038
       1. Found two outliers

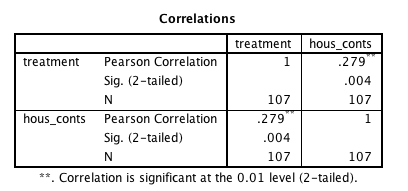


* + 1. Leverage = (2\*2 + 2) / 109 = 0.055
       1. Found four outliers

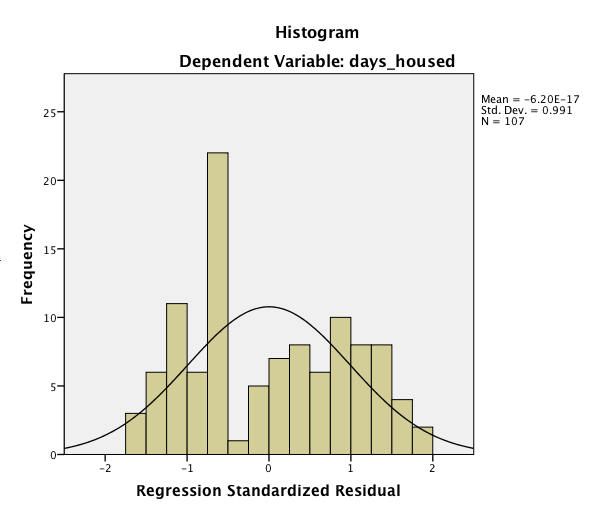


* + 1. I will delete the top two because they have two out of the three outlier markers.
  1. Because I deleted, I want to rerun the screening (the regression) to get the rest of these.

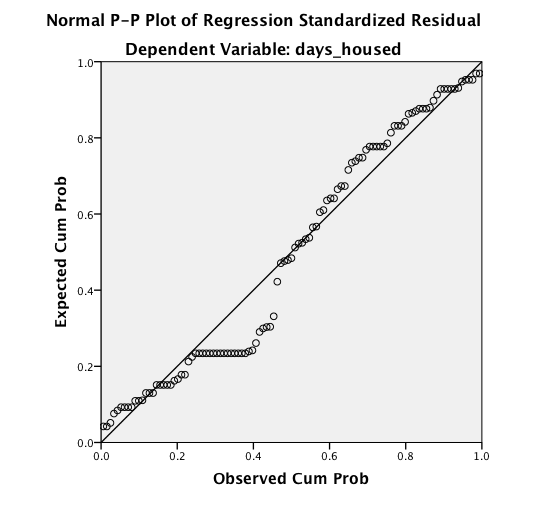
1. Multicollinearity
   1. Look at the correlation between the X variable and the mediator.
   2. Analyze correlate bivariate.
   3. Great, not too correlated (*r* = .28).



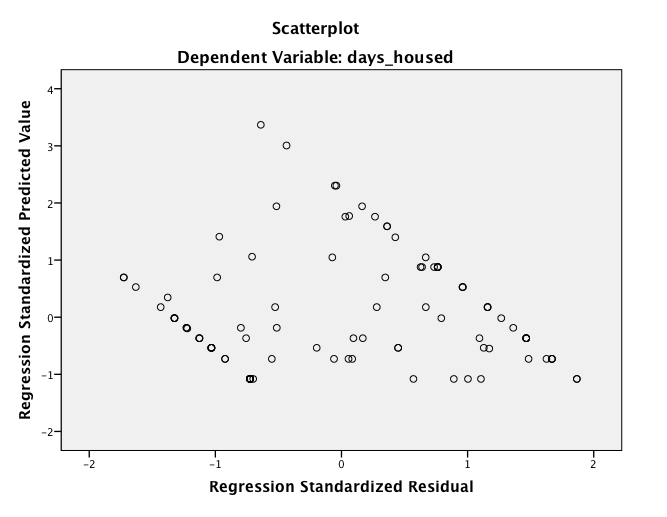
1. Normality – looks ok – one issue here is that we are using a categorical predictor (so probably why it looks like two humps, you can try screening without the predictor and see if it’s better, then you would know it’s that predictor).



1. Linear – mostly linear relationship – there’s one area with a small problem (again same issue, see above).



1. Homogeneity – most data points are between 2 and -2 both ways (runs up to -4 but that’s only 2 points outside that range). Looks ok.
2. Homoscedasticity – not the best, but should be ok (again may be that categorical variable).



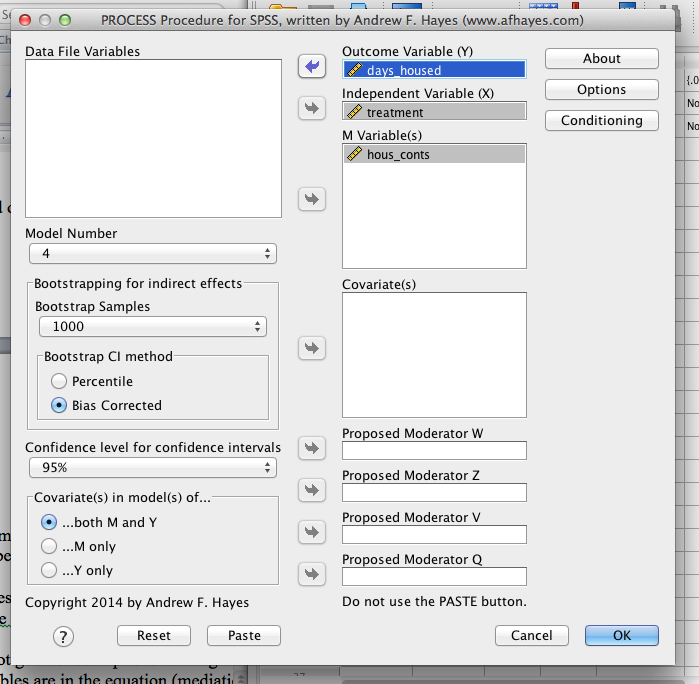
**Mediation Steps:**

Analyze > regression > PROCESS.

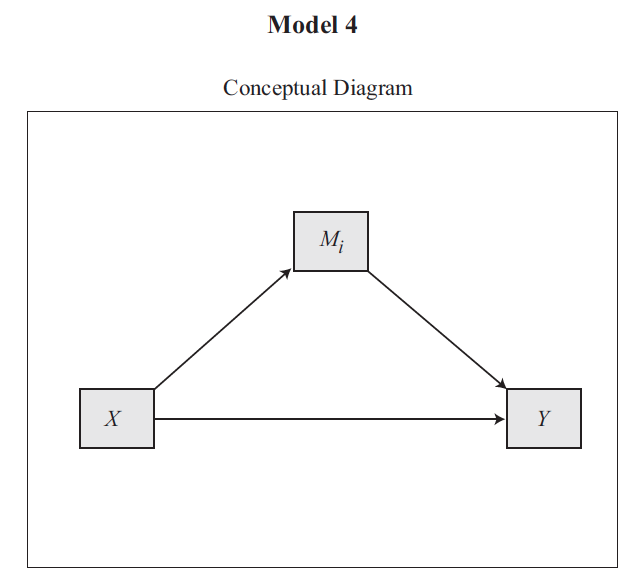
Independent = treatment (your IV).

Outcome variable = days housed (your DV).

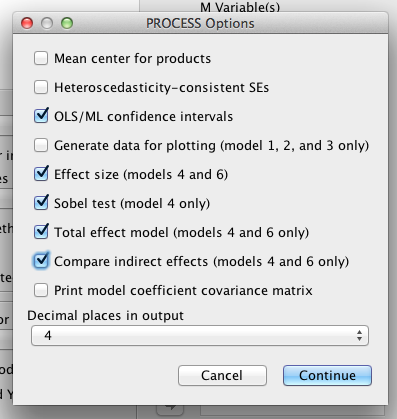
M variable = house contacts (your mediator – note you shouldn’t switch the IV and mediator, theory will dictate which one you think is the mediator).



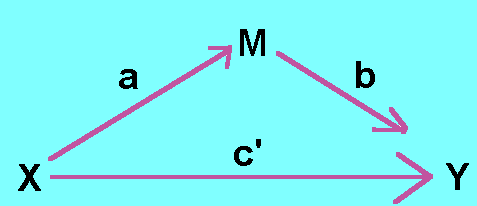
Make sure the model number is 4.

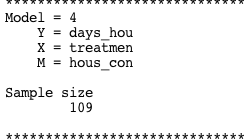


Click options.

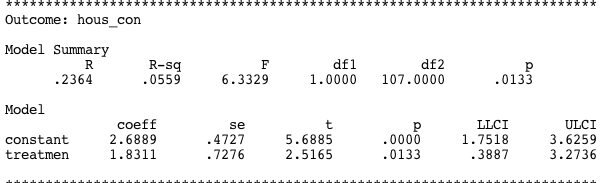


* OLS/ML confidence intervals = ordinary least squares/maximum likelihood confidence intervals, type of estimation for confidence intervals since it runs a lot of bootstrapped samples.
* Effect size – you’ll get the effect size for the mediation (plus the overall model but this effect size is more useful).
* Sobel test – tells you if the mediation is significant (yeah!).
* Total effect model – both of the predictors in the model.
* Compare indirect effects – gives you the comparison of c and c’ (see below).



****

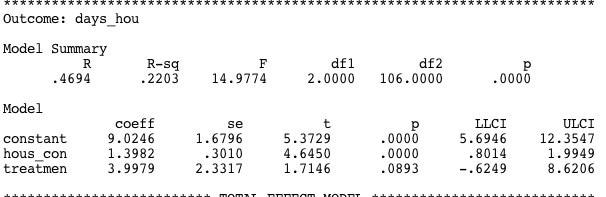
First, you have the variables you entered (helpful for reference).



This section tells you if the *a* path is significant.

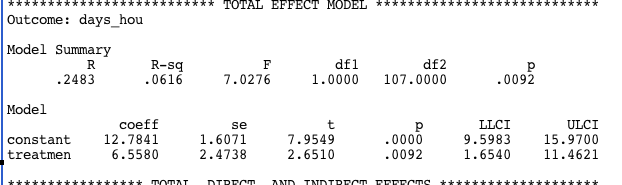
* Overall model: *F*(1, 107) = 6.33, *p* = .01, *R2* = .06
* *b* = 1.89, *t*(107) = 2.52, *p* = .01

So, treatment variable does predict housing contacts, which is important for mediation (you have to have the X variable predict the mediator or you can’t say the x > y relationship is changed by M).



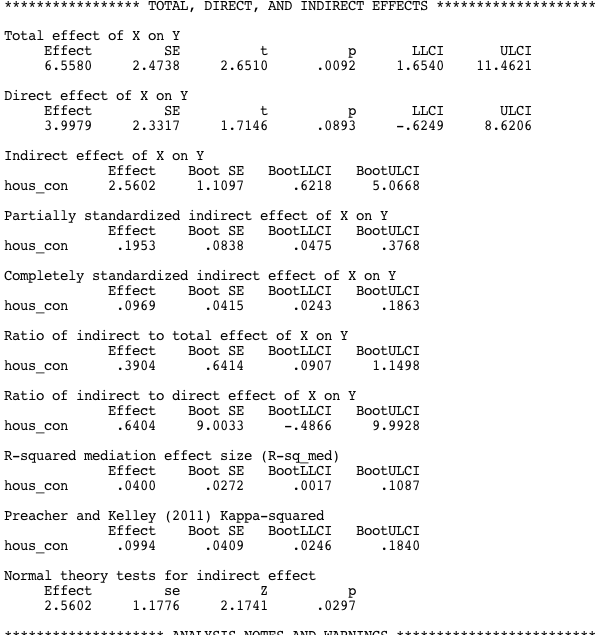
This model includes both X and M predicting Y, which includes both path *b* and path *c’*.

* The overall model with both predictors is significant: *F*(2, 106) = 14.98, *p* < .001, *R2 =* 22.
* The *b* path is significant (mediator to y variable, housing contacts to days housed): *b* = 1.40, *t*(106) =4.65, *p* < .001.
* The *c’* path is NOT significant (x to y, treatment condition to days housed): *b* = 4.00, *t*(106) = 1.71, *p* = .09.
  + You actually WANT this. That means that when the mediator is in the equation, the relationship between x and y is NOT significant.



This model is x predicting y (treatment condition to days housed) without the mediator in the equation. This model needs to be significant or you cannot have meditation.

* Overall model is significant, *F*(1, 107) = 7.03, *p* = .01, *R2* = .06
* The *c* path is significant, *b* = 6.56, *t*(107) = 2.65, *p* = .01.



Total effect of X on Y = the c path.

Direct effect of X on Y = the c’ path.

Indirect effect on X on Y = c – c’ path.

The rest of the mumbo jumbo is different types of effect sizes. You want to use the Preacher and Kelley Kappa-squared for the indirect effect. κ2 = .10.

How do I know if this value is significantly greater than zero? Use the Sobel test. It’s a Z-test for the indirect effect.

Yes, our mediation was significant, *Z* = 2.17, *p* = .03, κ2 = .10.

**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). Full 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* = .03, κ2 = .10.

|  |
| --- |
| *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 |