# Instructions

1. Read through "collegedistance-data-description.rtf" and learn about the data file
2. Open "collegedistance.csv" in SPSS
3. Add variable labels.
4. Convert all string categorical variables into numeric variables with value labels (see automatic - recode)
5. Examine frequencies and graphs where appropriate to understand the data file
6. Test whether the effect of parental college education on child years of education is mediated by the child's high school test scores; (a) create a composite variable that indicates the number of parents that have a college education (i.e., zero for none; 1 for either mum or dad; 2 for both mum and dad (b) run the mediation and report a, b, c, and c-prime; (c) re-run the mediation test with standardised variables (d)) test the significance of the indirect effect.
7. We are going to use the categorical predictors in a regression model: (a) convert all the yes-no variables into a form where 0=no and 1=yes; (b) re-label gender, income, region, so that 0=something and 1=something else and the label corresponds to what 1 is; e.g., if 0 = female and 1=male, then label the variable "male"; (c) dummy code ethnicity so that you have african american and hispanic indicator variables.
8. Run a multiple regression predicting years of education from all other variables in the data file (use the newly created numeric variables for the binary and categorical variables(.
9. Using the previous overall model, perform a series of hierarchical regressions where you try adding in an interaction term in second step of a multiple hierarchical regression: try adding the following one at a time (a) two binary variables (e.g., father college and mother college); (b) a binary and a three categorical variable (e.g., gender and ethnicity); (c) a numeric and a binary variable (e.g., base line test score and gender); (d) numeric and three categorical variable (e.g., score and ethnicity); (e) two numeric variables (e.g., tuition and score).
10. For each of the preceding interaction terms, (a) what does the direction of the interaction effect imply; (b) was the interaction statistically significant; (c) consider creating a plot to show the relationship

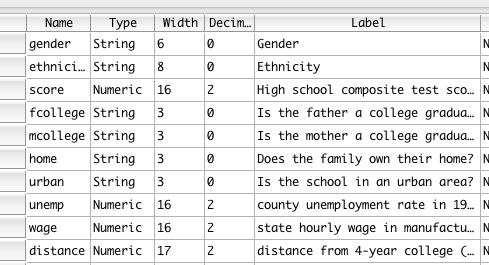
# Answers

## 2. Open "collegedistance.csv" in SPSS

This is a comma separated value file so use the text import wizard in SPSS.

## 3. Add variable labels.

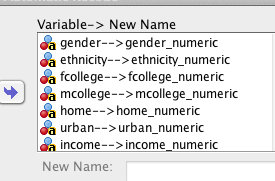
I.e., take the variable descriptions from the data description file and paste into the label field in the variable view in SPSS.



## 4. Convert all string categorical variables into numeric variables with value labels (see automatic - recode)

This is often useful in SPSS as string variables are often not permitted in various analysis methods.

Use "transform - automatic recode". You can do all the variables at once. You just need to give each variable a new name. A simple approach is to append some consistent word to the end of each of the original variable names.



## 5. Examine frequencies and graphs where appropriate to understand the data file

"Analyze - descriptives - frequencies" is a good tool for this optionally requesting histograms is often nice to get a feel for the distribution of the continuous variables.

This should allow you to understand each variable, what it represents, and what kind of variation and distribution there is in each variable.

## 6. Test whether the effect of parental college education on child years of education is mediated by the child's high school test scores; (a) create a composite variable that indicates the number of parents that have a college education (i.e., zero for none; 1 for either mum or dad; 2 for both mum and dad

There are a few ways you could do this. You could sum the numeric variables that you created in the previous step (which are most likely coded 1=no, 2=yes) and you could then subtract 2 from that total. Or you could do step seven where you recode the no-yes variables into no=0, yes=1 and sum that version of the variables using transform - compute.

## (b) run the mediation and report a, b, c, and c-prime;

There are various ways to do this

I grabbed the Sobel test from here.

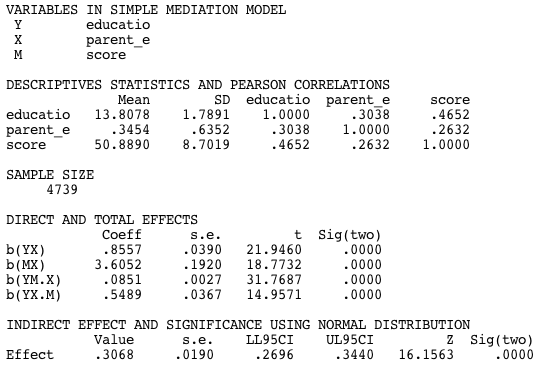
http://www.afhayes.com/public/sobel.sps

Save the file as "sobel.sps" and open in SPSS. Highlight all the text in the file and run all this syntax. This adds the macro to SPSS.

You can then run the macro like this in syntax, notice how the spss variable name for the dv, iv, and mv has been entered. I choose 100 bootstrap samples just to make sure things run quickly. But 2,000 or more would be better if you have the computational time and it is a final analysis for reporting in a thesis or journal article (.

SOBEL y=education /x=parent\_education /m=score /boot=100.

Probably



Partly due to the large sample size all direct and indirect effects are statistically highly significant. That said, the output above shows the unstandardised coefficients.

* A one unit increase in the number of parents with college education predicts .86 years of additional education for the child.
* A one unit increase in the number of parents with college education predicts an 3.6 unit increase in child High School test scores (i.e., about .4 of standard deviation)
* A one unit increase in test scores controlling for parent college education predicts .08 years increase in child years of education
* A one unit increase increase in parents with college education controlling for child test scores leads to a .55 increase in child years of education
* The indirect effect suggests of parental education through test scores on child years of education is .30 (95% bootstrap CI = (.27 - .34).

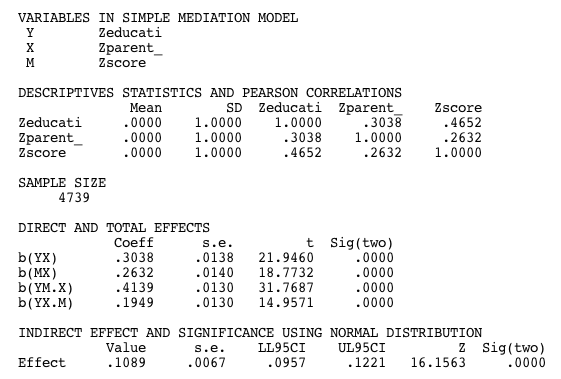
## (c) re-run the mediation test with standardised variables

Create z-scores for the three variables in analyze - descriptives - descriptives and click save as standardised variables.

That will create a set of z-score variables with z prefixed to the original.

You can then re-run the macro with the new z-score variables:

SOBEL y=zeducation /x=zparent\_education /m=zscore /boot=100.



Now the various coefficients and the indirect effect are in some sense easier to compare. Thus, the total effect of number of parents with college education is .30 of a standard deviation, but the direct effect is .19 and the indirect effect is .11. Note how the sum of the indirect and direct effects equals the total effect.

## (d) test the significance of the indirect effect.

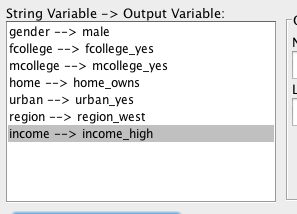
As is shown by the bootstrap, the 95% confidence interval of the indirect effect does not include zero, therefore it is statistically significant.

A note of caution: Just because we have a significant indirect effect does not mean that this is the way the causal process operates.

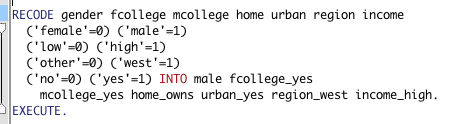
## 7. We are going to use the categorical predictors in a regression model: (a) convert all the yes-no variables into a form where 0=no and 1=yes; (b) re-label gender, income, region, so that 0=something and 1=something else and the label corresponds to what 1 is; e.g., if 0 = female and 1=male, then label the variable "male";

There are a few ways to do this; the main option is "transform - recode into different variable".

You can do all the variables at once, which is efficient when some of the "old and new values" are the same across variables. Note the naming convention I use below and how it makes it clear what the 1 value will represent.



After pasting the syntax, the code looked like this:



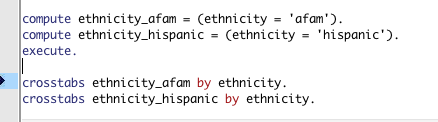
Note that to get this to work, you have to check the wording of the value labels.

## (c) dummy code ethnicity so that you have african american and hispanic indicator variables.

To dummy code a three categorical variables, check out this tutorial:

<http://www.ats.ucla.edu/stat/spss/code/dummy.htm>

So we can just use "transform - compute" or simply type in the syntax directly:



The first compute syntax creates the dummy variables and the crosstabs line checks that it worked (note crosstabs could also be obtained by going to "analyze - descriptives - crosstabs").

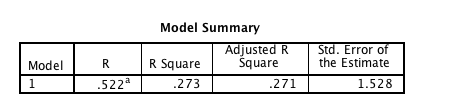
## 8. Run a multiple regression predicting years of education from all other variables in the data file (use the newly created numeric variables for the binary and categorical variables).

Go to Analyze - regression linear

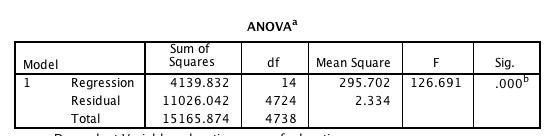
IVs: Enter all the continuous IVs and the previously coded binary variables and the dummy coded predictors.

DV: education

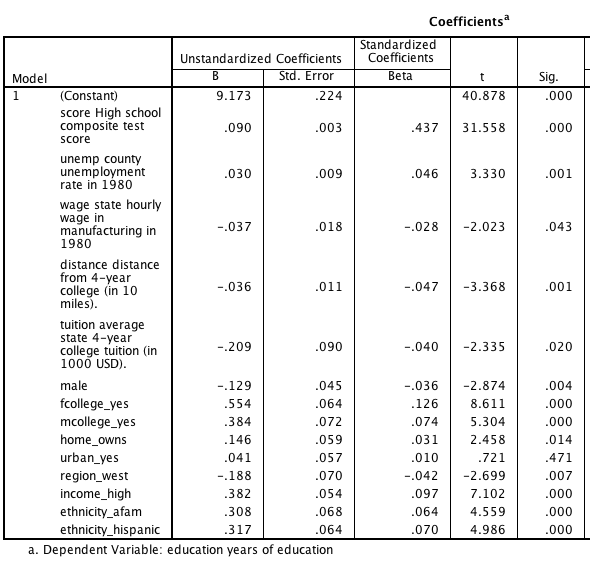
You might also want to select a few other things like statistics - descriptives, part and partial, collinearity, residuals etc. but I'll try to keep it simple here.



So overall, we are explaining 27.3% of variance in educational attainment.



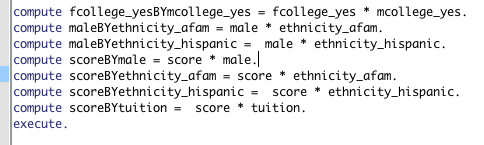
This is statistically significant.



Most of these predictors are statistically significant, but looking at the standardised betas only a few are large. The strongest predictor is high school test score. Then you have family's income and both parent's education being important factors related to more education.

## 9. Using the previous overall model, perform a series of hierarchical regressions where you try adding in an interaction term in second step of a multiple hierarchical regression: try adding the following one at a time

Here is the general syntax for creating all the interaction variables:

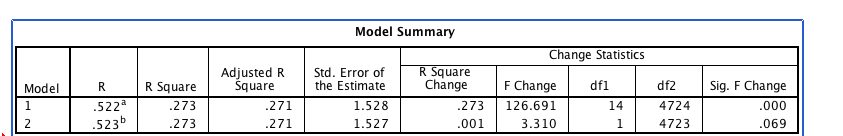


You could do this using transform - compute, but it's simpler just to use the syntax directly at this point. Note the "\*" symbol is used for multiplication. Note also that if you have a three category variable then you need to create separate interaction terms for each of the dummy coded variables.

To run the hierarchical regression with the interaction term, you could just include the interaction variable as another predictor, but it will be easier to see the incremental prediction if we perform a hierarchical regression which involves pressing the next button in "analyze - regression - linear" and putting the interaction variable or variables in the second step. Also make sure you request under "statistics" "r-square change".

## 10. For each of the preceding interaction terms, (a) what does the direction of the interaction effect imply; (b) was the interaction statistically significant; (c) consider creating a plot to show the relationship

## (a) two binary variables (e.g., father college and mother college);

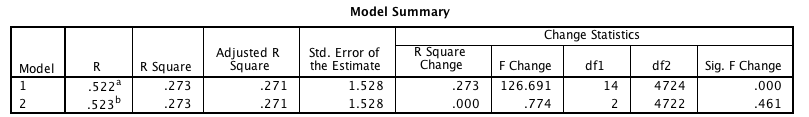


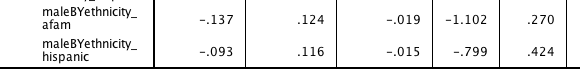
The interaction was not quite statistically significant.



The tendency of the effect was suggesting that the effect of having with parents with college education was not entirely additive. Thus, having two parents with college education was less than the sum of the benefits of having either one parent with education. To simplify, father's college meant .6 of a year more education, mother's college meant .5 of a year more education, but to have both was not 1.1 extra years of education it was more like 1.1 - .26 equals around .8 years extra education. That said, this interaction was not statistically significant.

## (b) a binary and a three categorical variable (e.g., gender and ethnicity);

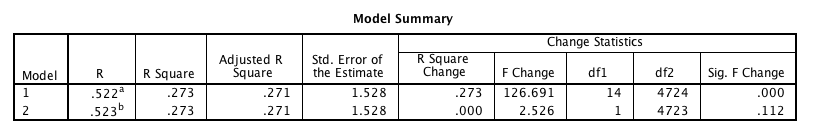




There was no significant gender by ethnicity interaction.

The direction of the effect suggested that the effect of being males relative to females who were aferican-american or hispanic received less education in relative to males versus females who were another ethnicity (presumably white).

## (c) a numeric and a binary variable (e.g., base line test score and gender);

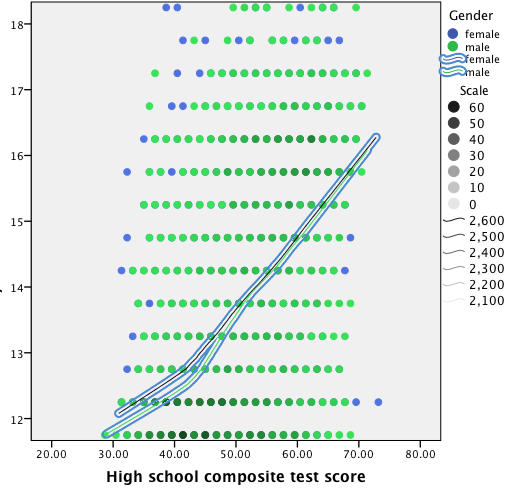




No significant interaction of test scores by gender. Thus, test scores are pretty much equally predictive of educational attainment for both males and females.

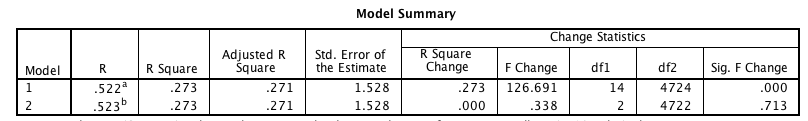
The coefficient suggests that there might a slightly stronger relationship between test scores and educational attainment for males.

You could create a scatter plot with education on the y-axis, score on the x-axis, and set markers by gender. It's a little tricky to read given the discrete nature of the data and the huge sample size. You can use binning and perhaps fit lines at sub-groups to make things a little better:

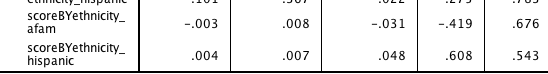


Note that this technically isn't quite the same as the model coefficients because the regression statistically controlled for all other variables. The graph shows that there is a strong relationship between scores and years of education, but that it doesn't vary much by gender.

## (d) numeric and three categorical variable (e.g., score and ethnicity);

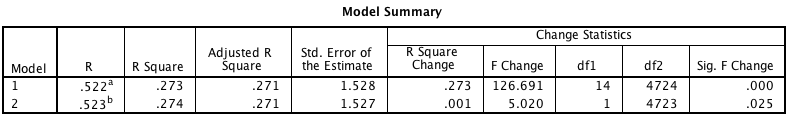


This interaction was not statistically significant. Thus, test scores are pretty much equally predictive of educational attainment for all ethnic groups.



The direction of the interaction effect suggests that test scores were better predictors of education for African Americans and poorer predictors for Hispanics. (but as mentioned this wasn't a significant interaction).

## (e) two numeric variables (e.g., tuition and score).



Interestingly, this was a significant interaction.



Test score appears to be a stronger predictor of educational attainment where tuition fees are greater. I don't have obvious explanation for that.