Psyc492 Multiple Regression Hypothses Testing Tour

These data were obtained from group of graduating High School seniors from a large eastern school district. All of the students in the data base went to college the fall after they graduated high school. The data for the predictor variables were all collected the week before high school graduation, or were taken from their high school transcript. Each student was contacted at the end of their first semester of college to collect the criterion variables.

Criterion variables → **College Performance (1st semester gpa)**

Predictor variables → **High School Performance variables (standardized tests in reading, writing, math, science and civics)**

→ **Demographic variables (Neighborhood they grew in urban/riral, Socioeconomic Status, Locus of Control (higher is more external)**

Lets take a look at four of the major kinds of “hypothesis testing” used in multiple regression

- #1 Building a single model
- #2 Comparing nested models
- #3 Comparing non-nested models
- #4 Comparing a model across populations

Before we start testing models!!

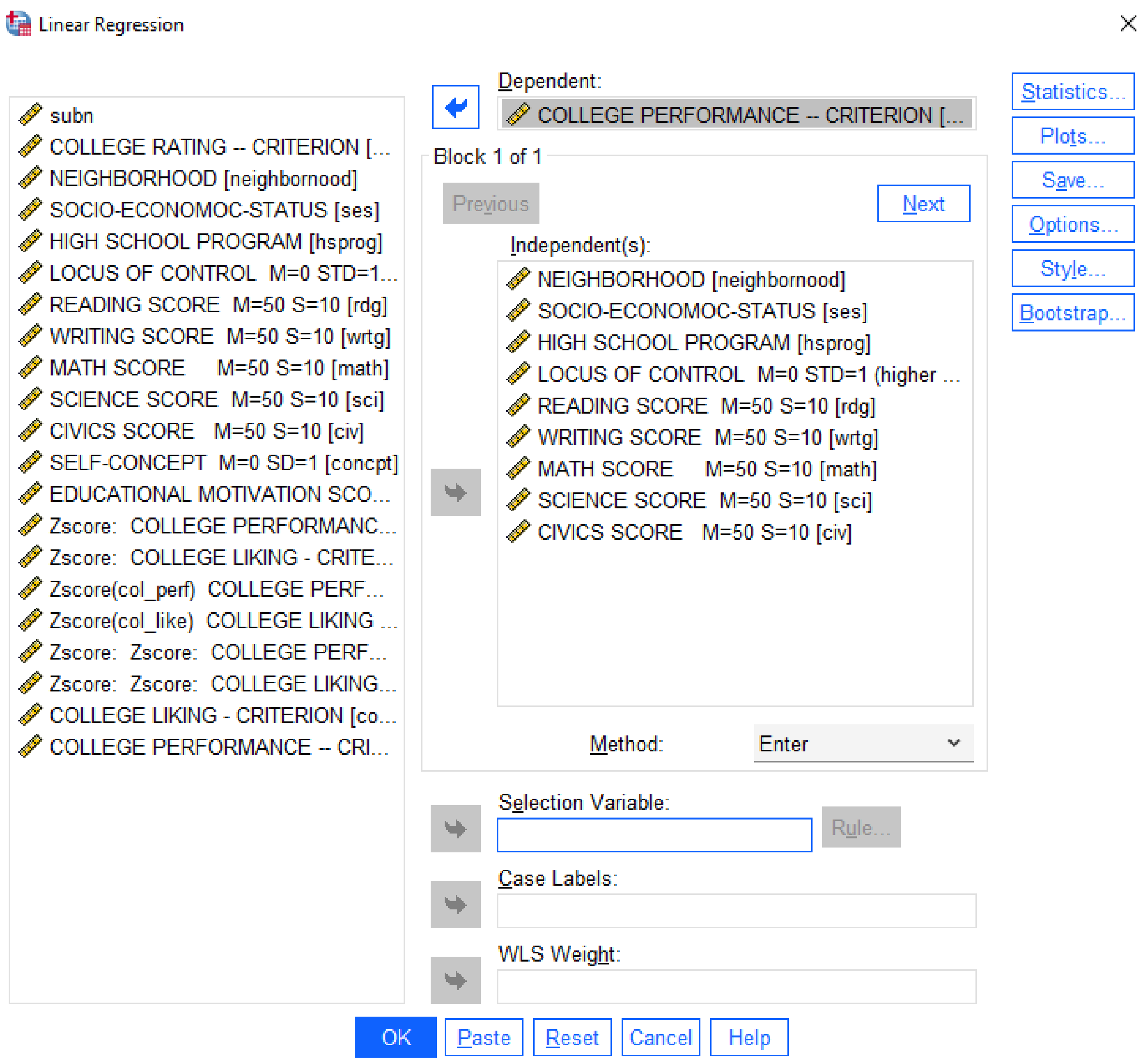
- Always look at the frequencies, mean, std, skewness, minimum value and maximum value of each variable to see if there is anything “squirrely” (negative values, a multiple-category variable, substantial skewing) and how it may influence how you analyze the data and your results
- Also, get the correlations -- all of them
  - Correlations between the criterion
  - Correlations of each criterion with all the predictors
  - Correlations among the predictors (called collinearities)

#1 – Getting a single model, using "all the predictors" and one criterion

Let's start with the performance variable (gpa) as the criterion variable.

When we say "all the predictors" we don't mean "every variable" or even "all the variables in the data set", we mean "all the variables we have decided to include".

Analyze → Regression → Linear



Load in the criterion and the predictors and click "OK"

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.795[a]	.633	.602	.82769

a. Predictors: (Constant), CIVICS SCORE M=50 S=10, NEIGHBORHOOD, HIGH SCHOOL PROGRAM, MATH SCORE M=50 S=10, SOCIO-ECONOMOC-STATUS, LOCUS OF CONTROL M=0 STD=1 (higher scores - more external), READING SCORE M=50 S=10, SCIENCE SCORE M=50 S=10, WRITING SCORE M=50 S=10

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	127.508	9	14.168	20.681	.000[b]
	Residual	73.987	108	.685		
	Total	201.496	117			

a. Dependent Variable: COLLEGE PERFORMANCE -- CRITERION

b. Predictors: (Constant), CIVICS SCORE M=50 S=10, NEIGHBORHOOD, HIGH SCHOOL PROGRAM, MATH SCORE M=50 S=10, SOCIO-ECONOMOC-STATUS, LOCUS OF CONTROL M=0 STD=1 (higher scores - more external), READING SCORE M=50 S=10, SCIENCE SCORE M=50 S=10, WRITING SCORE M=50 S=10

The Model Summary table tells us the R-square of the model (what proportion of the variance in the criterion is accounted for by the predictor model)

- This model accounts for about 63% of the variance in the criterion – a very strong model.

ANOVA table reports the test of whether the H0: that R-square = 0

- The p-value tells us to reject that null and conclude that the "model works better than is expected by chance, taking the sample size into account"

		Coefficients[a]				
		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	-3.014	.625		-4.823	.000
	NEIGHBORHOOD	.147	.158	.056	.930	.354
	SOCIO-ECONOMOC-STATUS	.027	.118	.015	.232	.817
	HIGH SCHOOL PROGRAM	.055	.145	.023	.381	.704
	LOCUS OF CONTROL M=0 STD=1 (higher scores - more external)	.030	.119	.016	.252	.802
	READING SCORE M=50 S=10	.033	.007	.403	5.101	.000
	WRITING SCORE M=50 S=10	.033	.012	.264	2.801	.006
	MATH SCORE M=50 S=10	.022	.006	.246	3.587	.001
	SCIENCE SCORE M=50 S=10	.011	.011	.086	1.029	.306
	CIVICS SCORE M=50 S=10	.014	.012	.101	1.164	.247

a. Dependent Variable: COLLEGE PERFORMANCE -- CRITERION

These are the “regression weights” or “coefficients” – we will use these to interpret the model

Note which predictors have significant loadings (H0: B = 0)

- Reading, writing and Math have (positive) significant regression weights – which means that these have “unique contributions to the model”.

The “Beta weights” can be used to consider the “relative importance” of the contributing predictors.

- Reading seems to be somewhat more important to the model than the other two significant predictors.

We would conclude that the model “works” and that Reading, Writing and Math have significant independent contributions to the model!

If we had specific hypotheses about which variables do and don’t contribute to the model, we would test those using the regression weights and significance test shown here.

#2 – Comparing nested models

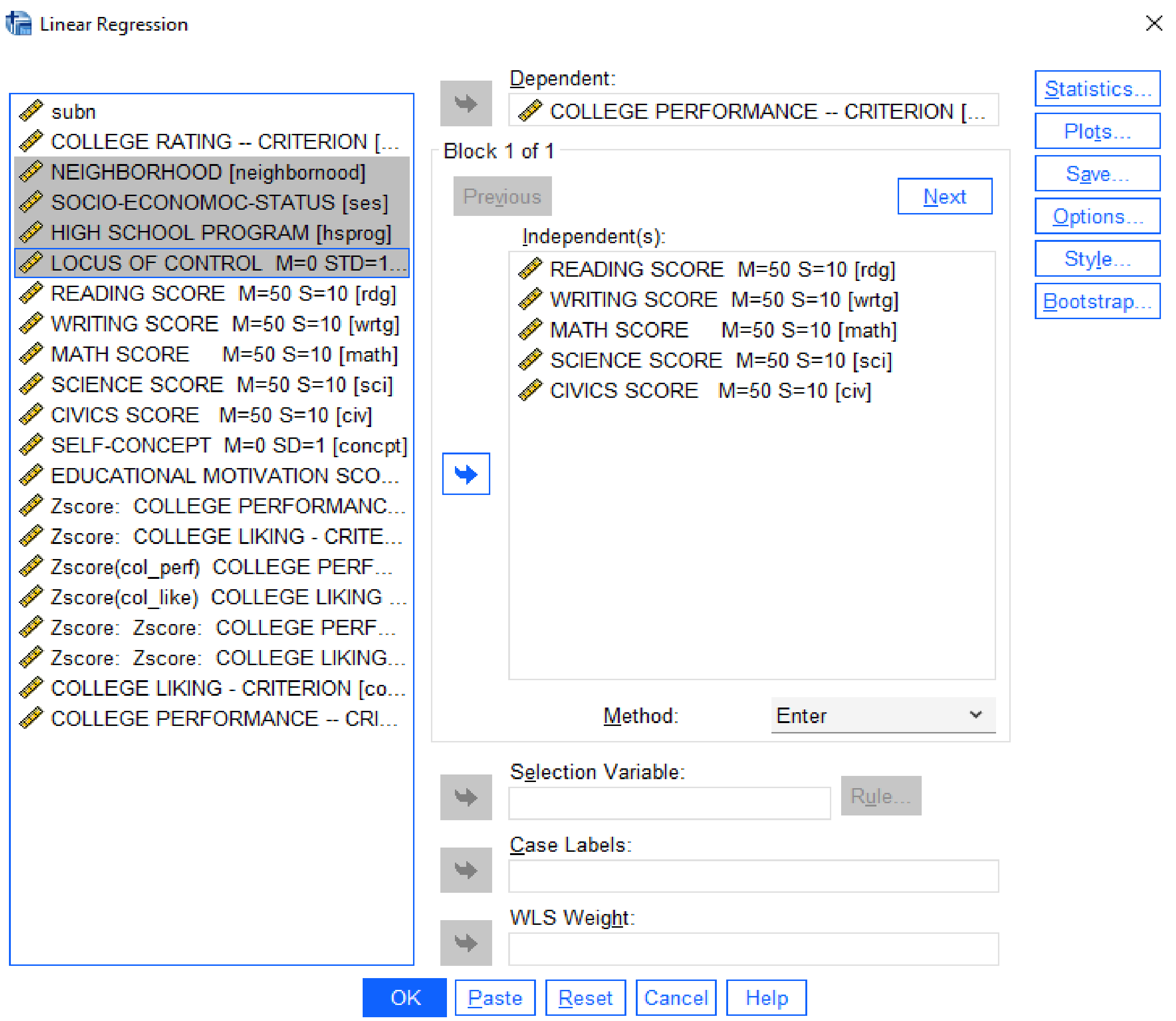
Since the criterion variable is college performance, it makes sense to predict college performance from high school performance! But, do the other variables in the model (demographic and "personality" variables) "add anything"? To do this, we will test nested models!

Full model → reading, writing, math, science, civics, neighborhood, ses, high school prog & Locus of control

Reduced model (the model we want to test if it is "sufficient") → reading, writing, math, science &civics

First we're going to build the **"reduced model",** with just the five high school performance variables, This is the "**reduced model**" because it only has a subset of the predictors in it.

Analyze → Regression → Linear



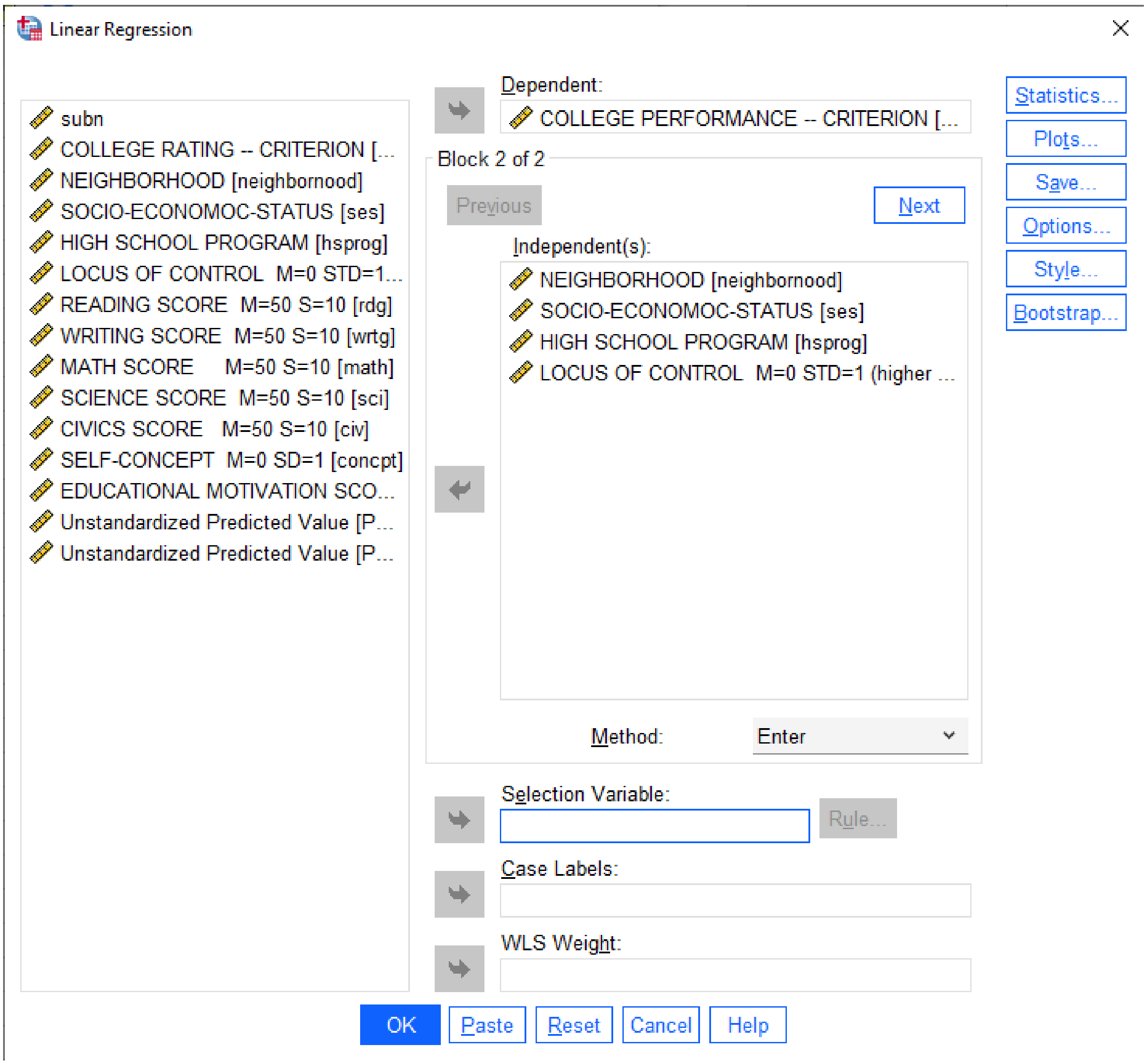
Put in College Performance as the criterion (Dependent).

Put in five high school performance as the predictors (Independents).

This is the "**reduced model**" because it only has a subset of the predictors in it. **Then click "Next"**

It should now say "Block 2 of 2"

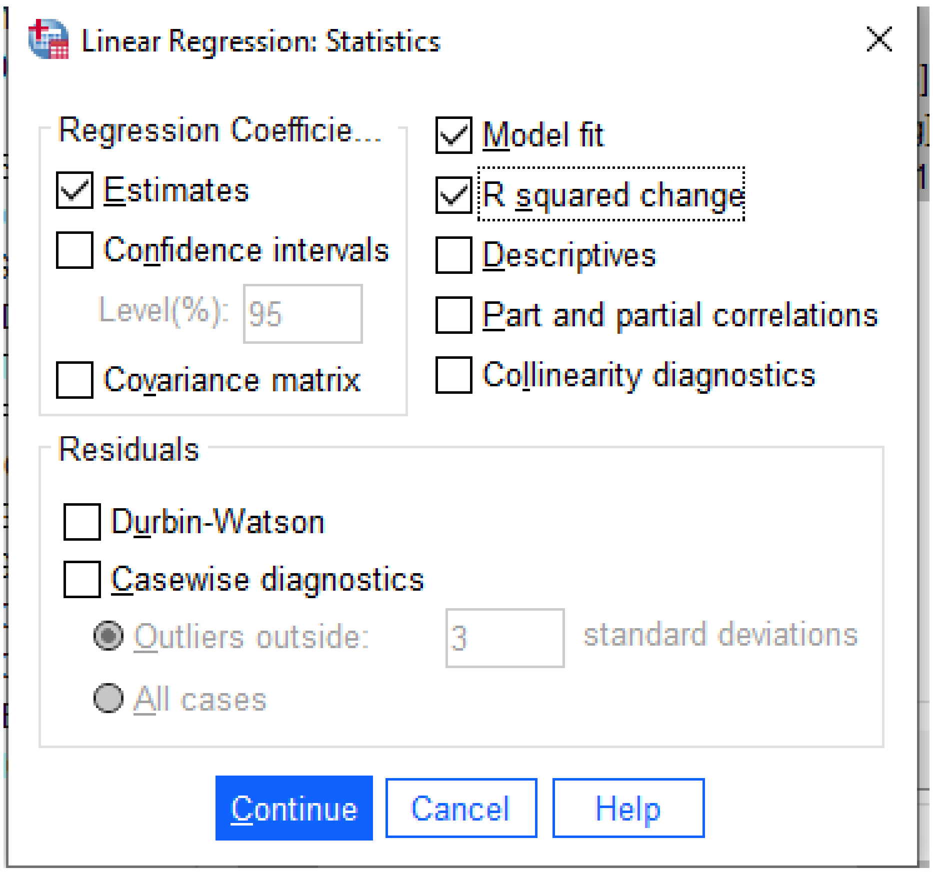
Then, we will build the "**full model**" by adding in the four demographic and personality variables. This will be the "**full model**" because it has all nine of the predictors in it.



Put in the four additional demographic and personality variables.

This is making the "**full model**" by adding in "the rest of the variables"

Now click on the “Statistics” button.



Be sure that all of these are checked

- Estimates
- Model fit
- R squared Change

Then click “**Continue**” on this window and “**OK**” on the main Linear Regression window.

Here's the first part of the output – you have to read these two tables together to understand each model and their comparison.

Model Summary

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.793[a]	.629	.613	.81679	.629	38.005	5	112	.000
2	.795[b]	.633	.602	.82769	.004	.268	4	108	.898

a. Predictors: (Constant), CIVICS SCORE M=50 S=10, MATH SCORE M=50 S=10, READING SCORE M=50 S=10, SCIENCE SCORE M=50 S=10, WRITING SCORE M=50 S=10

b. Predictors: (Constant), CIVICS SCORE M=50 S=10, MATH SCORE M=50 S=10, READING SCORE M=50 S=10, SCIENCE SCORE M=50 S=10, WRITING SCORE M=50 S=10, NEIGHBORHOOD, HIGH SCHOOL PROGRAM, SOCIO-ECONOMOC-STATUS, LOCUS OF CONTROL M=0 STD=1 (higher scores - more external)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	126.775	5	25.355	38.005	.000[b]
	Residual	74.721	112	.667		
	Total	201.496	117			
2	Regression	127.508	9	14.168	20.681	.000[c]
	Residual	73.987	108	.685		
	Total	201.496	117			

a. Dependent Variable: COLLEGE PERFORMANCE -- CRITERION

b. Predictors: (Constant), CIVICS SCORE M=50 S=10, MATH SCORE M=50 S=10, READING SCORE M=50 S=10, SCIENCE SCORE M=50 S=10, WRITING SCORE M=50 S=10

c. Predictors: (Constant), CIVICS SCORE M=50 S=10, MATH SCORE M=50 S=10, READING SCORE M=50 S=10, SCIENCE SCORE M=50 S=10, WRITING SCORE M=50 S=10, NEIGHBORHOOD, HIGH SCHOOL PROGRAM, SOCIO-ECONOMOC-STATUS, LOCUS OF CONTROL M=0 STD=1 (higher scores - more external)

What about the first model (the reduced model with only the high school performance variables)?

- Look at the Model Summary table – first row. It tells us (on the left) that the reduced model with the five high school performance variables accounts for 62.9% of the variance in college grades
- Look at the top part of the ANOVA – first 3 rows. It tells us that this model "works" (p= .000).

What about the second model (the full model with the demographic and personality variables added in)?

- Look at the Model Summary table – second ros. It tells us (on the left) that the full model with all nine predictors in it accounts for 63.3% of the variance in college grades.
- Look at the bottom part of the ANOVA table – bottom 3 rows. It tells us that the full model "works" (p=.000).

What about comparing the models (do the demographic and personality variables add anything to the reduced model using only the high school performance variables)?

- Look at the Model Summary Table – right side, bottom row
- This tells us that the full model (with all 9 predictors) accounts for .4% (R-square change = .004) more variance than the reduced model (with only the high school performance predictors).
- The F- test tell us that there is no difference between the R-square of the two models (p = .858)

So… We would conclude that the reduced model including the 5 high school performance predictors "works as well" to predict College performance as does the full model that also includes the demographic and personality predictors.

Here is the rest of the output – it gives the regression weights for each model and their significance tests.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-2.743	.441		-6.224	.000
	READING SCORE M=50 S=10	.034	.006	.410	5.378	.000
	WRITING SCORE M=50 S=10	.032	.012	.256	2.778	.006
	MATH SCORE M=50 S=10	.022	.006	.244	3.639	.000
	SCIENCE SCORE M=50 S=10	.012	.010	.095	1.168	.245
	CIVICS SCORE M=50 S=10	.015	.011	.112	1.347	.181
2	(Constant)	-3.014	.625		-4.823	.000
	READING SCORE M=50 S=10	.033	.007	.403	5.101	.000
	WRITING SCORE M=50 S=10	.033	.012	.264	2.801	.006
	MATH SCORE M=50 S=10	.022	.006	.246	3.587	.001
	SCIENCE SCORE M=50 S=10	.011	.011	.086	1.029	.306
	CIVICS SCORE M=50 S=10	.014	.012	.101	1.164	.247
	NEIGHBORHOOD	.147	.158	.056	.930	.354
	SOCIO-ECONOMOC-STATUS	.027	.118	.015	.232	.817
	HIGH SCHOOL PROGRAM	.055	.145	.023	.381	.704
	LOCUS OF CONTROL M=0 STD=1 (higher scores - more external)	.030	.119	.016	.252	.802

a. Dependent Variable: COLLEGE PERFORMANCE -- CRITERION

Looking at Model 1 (the reduced model with the 5 high school performance predictors), we see that Reading, Writing, and Math contribute significantly to the model (have significant p-values) and Science & Civics do not.

Looking at Model 2 (the full model with the 4 demographic and personality predictors added in with the 5 high school performance predictors), we see that again, only Reading, Writing and Civics contribute to the model.

Since none of the variables added in on the second step contribute to the model, it is easy to see why the model fit doesn't improve when the demographic and personality variables are added.

#3 – Comparing non-nested models

When comparing nested models, we usually end up specifying two different models: 1) the model defined by the reduced model (e.g., the 5 high school performance variables) and 2) the model defined by the variables added to the reduced model to form the full model (e.g., the 4 demographic and personality variables).

In addition to asking if one set "adds to" the other, we will probably want to compare them – which model "does better"?

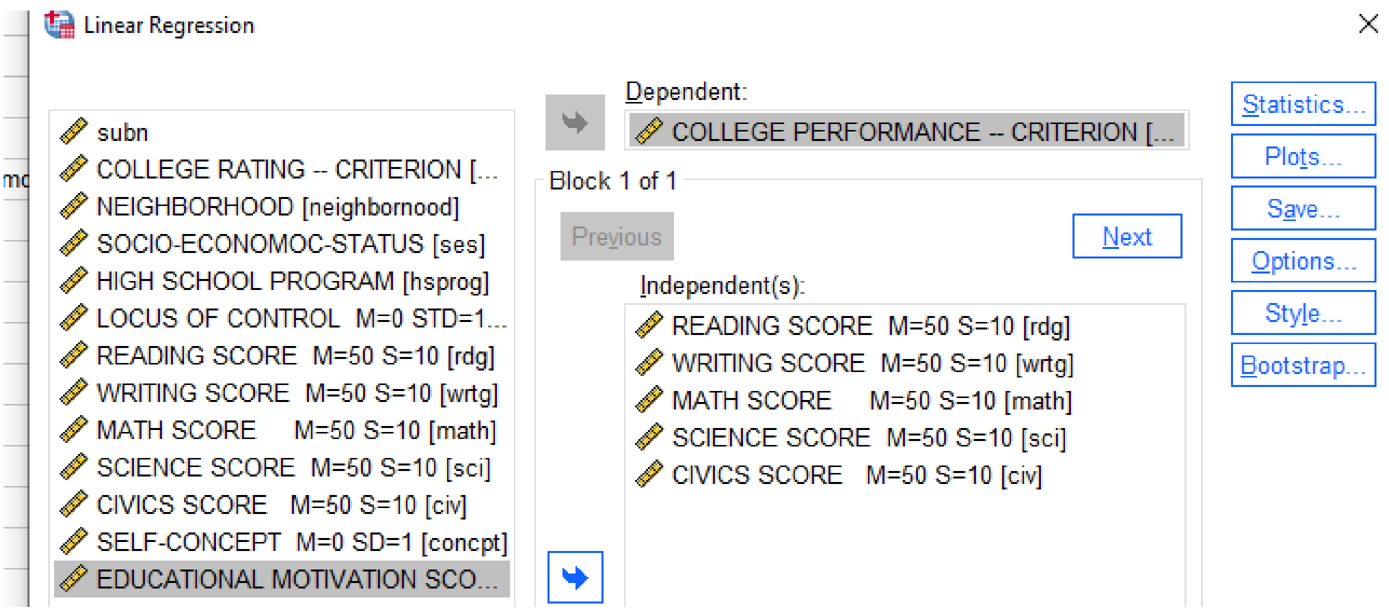
To do that we will first run each model separately – to see how well each model works. Then we will need to get the correlation between the two models (SPSS makes this really easy). Finally, we will use the Computator to perform a Steiger's Z-test to compare the two models, to see if one "accounts for the criterion" better?

Analyze → Regression → Linear

First → Getting the High School Performance model (and the predicted criterion values from it)

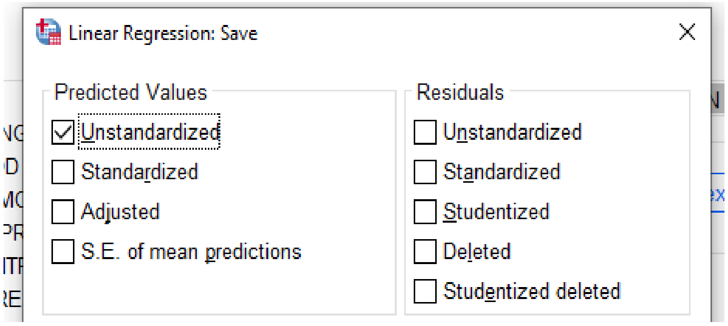
Analyze → Regression → Linear

This will look like the first step of the last analysis but I always do it over to I have things together in the output window).



Select the Dependent variable and then select the high school performance variables.

THEN → click on the SAVE button



Be sure to check "Unstandardized" box.

SPSS will build the model, and then will use that model to calculate a "predicted college GPA" score for each person.

We'll need that later!

Here's the output for the High School Performance model

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.793[a]	.629	.613	.81679

a. Predictors: (Constant), CIVICS SCORE M=50 S=10, MATH SCORE M=50 S=10, READING SCORE M=50 S=10, SCIENCE SCORE M=50 S=10, WRITING SCORE M=50 S=10

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	126.775	5	25.355	38.005	.000[b]
	Residual	74.721	112	.667		
	Total	201.496	117			

a. Dependent Variable: COLLEGE PERFORMANCE -- CRITERION

b. Predictors: (Constant), CIVICS SCORE M=50 S=10, MATH SCORE M=50 S=10, READING SCORE M=50 S=10, SCIENCE SCORE M=50 S=10, WRITING SCORE M=50 S=10

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2.743	.441		-6.224	.000
	READING SCORE M=50 S=10	.034	.006	.410	5.378	.000
	WRITING SCORE M=50 S=10	.032	.012	.256	2.778	.006
	MATH SCORE M=50 S=10	.022	.006	.244	3.639	.000
	SCIENCE SCORE M=50 S=10	.012	.010	.095	1.168	.245
	CIVICS SCORE M=50 S=10	.015	.011	.112	1.347	.181

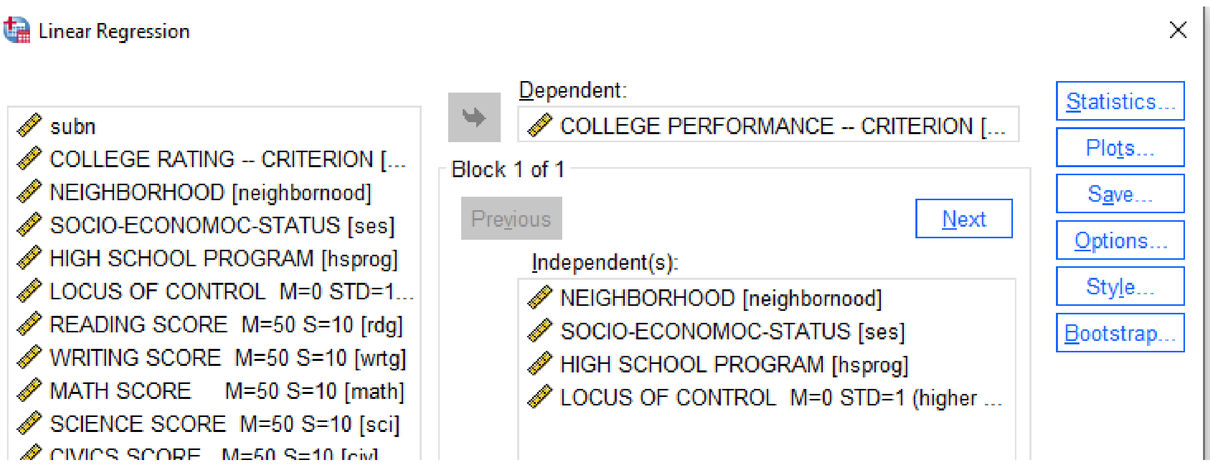
a. Dependent Variable: COLLEGE PERFORMANCE -- CRITERION

We see that the model has an R-square of .629, which is significant. We can also see that (like before) only Reading, Writing and Science have significant contributions to the model.

Later we will need the R from this analysis → The R = .793 for the High School Performance model

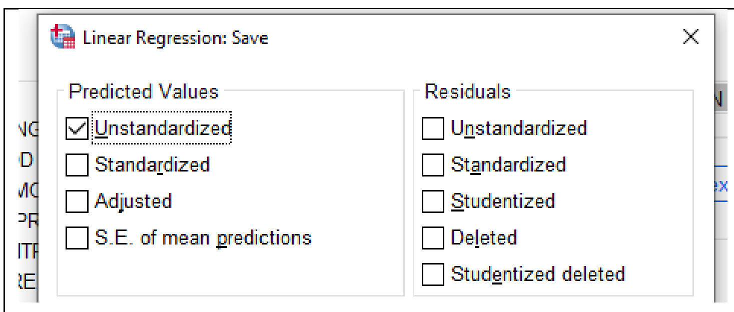
Second → Getting the Demographic and Personality model (and the predicted criterion values from it)

This will different from anything we've run yet (we've never made a model just with these variables).



Select the Dependent variable and then select the demographic and personality variables.

THEN → click on the SAVE button



Be sure to check "Unstandardized" box.

Like on the last step, SPSS will build the model, and then will use that model to calculate a "predicted college GPA" score for each person.

We'll need that later!

Here's the output for the Demographics & Personality model

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.448[a]	.201	.172	1.19385

a. Predictors: (Constant), LOCUS OF CONTROL M=0 STD=1 (higher scores - more external), HIGH SCHOOL PROGRAM, NEIGHBORHOOD, SOCIO-ECONOMOC-STATUS

b. Dependent Variable: COLLEGE PERFORMANCE -- CRITERION

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	40.439	4	10.110	7.093	.000[b]
	Residual	161.056	113	1.425		
	Total	201.496	117			

a. Dependent Variable: COLLEGE PERFORMANCE -- CRITERION

b. Predictors: (Constant), LOCUS OF CONTROL M=0 STD=1 (higher scores - more external), HIGH SCHOOL PROGRAM, NEIGHBORHOOD, SOCIO-ECONOMOC-STATUS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.806	.595		3.036	.003
	NEIGHBORHOOD	.288	.225	.109	1.280	.203
	SOCIO-ECONOMOC-STATUS	.491	.157	.268	3.138	.002
	HIGH SCHOOL PROGRAM	-.251	.202	-.105	-1.246	.215
	LOCUS OF CONTROL M=0 STD=1 (higher scores - more external)	.554	.155	.303	3.569	.001

a. Dependent Variable: COLLEGE PERFORMANCE -- CRITERION

We see that the model has an R-square of .201, which is significant. We can also see that only Socio economic level and Locus of Control have significant contributions to the model.

Later we will need the R from this analysis → The R = .448 for the Demographics and Personality model

Third → We want to compare the R from the two models, to see if one is significantly larger than the other!

We will do that using the Steiger's Z-test page in the Computator, Here's what that part of the Computator looks like.

A	B	C	D	E	F	G
Steiger's Z-test - Comparing Correlated Correlations						
		r(1,2) =>	0.793			
		r(1,3) =>	0.448			
		r(2,3) =>	0.538			
		N =>	118			
		Z =	5.574			
		p =	2.49E-08			
Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. Psy						

To get the necessary values for this, we need the correlations of each model -- r(1,2) & r(1,3). And we also need the correlation between the teo models – r()2,3) .

We found R = .793 for the 5 predictor High School Performance model

We found R = .448 for the 4 predictor Demographic & Personality model

All we need is the correlation between the two models! Remember that, when we got each model, we asked SPSS to compute and save the predicted score based on that model? Well, the correlation between the predicted scores for the two models IS the correlation between the models!!

If we look at the bottom of the data set, we will se two new variables – PRE_1 and PRE_2. These are the predicted scores for each model.

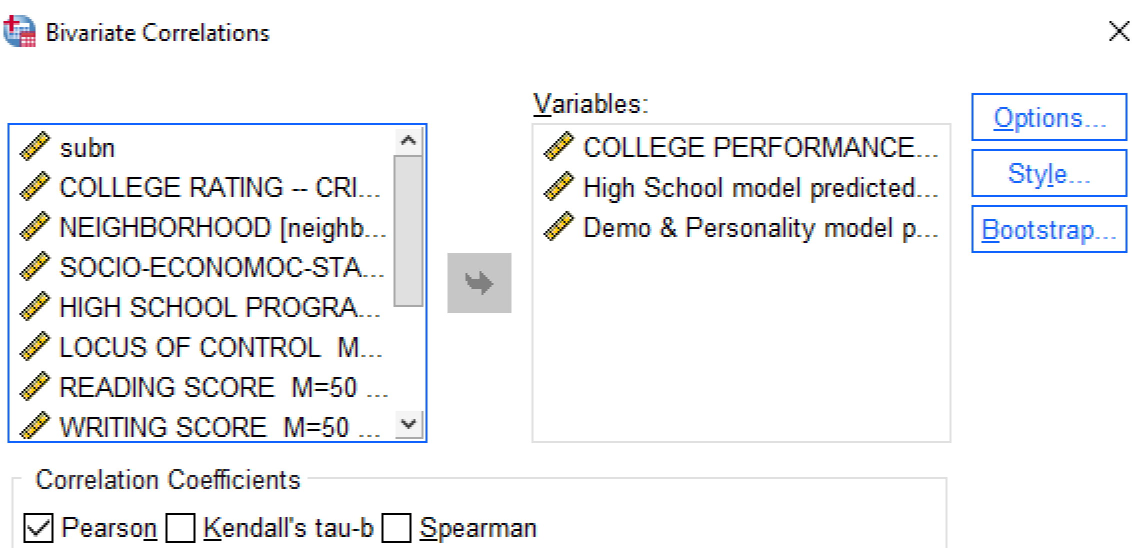
11	sci	Numeric	8	2	SCIENCE SCORE M=50 S=10
12	civ	Numeric	8	2	CIVICS SCORE M=50 S=10
13	concpt	Numeric	8	2	SELF-CONCEPT M=0 SD=1
14	motv	Numeric	8	2	EDUCATIONAL MOTIVATION SCORE
15	PRE_1	Numeric	11	5	Unstandardized Predicted Value
16	PRE_2	Numeric	11	5	Unstandardized Predicted Value
17					
18					

I suggest editing the variable name of these, so you'll know what they are in correlation output. Here's what I called them.

1	sci	Numeric	8	2	SCIENCE SCORE M=50 S=10	None
2	civ	Numeric	8	2	CIVICS SCORE M=50 S=10	None
3	concpt	Numeric	8	2	SELF-CONCEPT M=0 SD=1	None
4	motv	Numeric	8	2	EDUCATIONAL MOTIVATION SCORE	None
5	PRE_1	Numeric	11	5	High School model predicted values	None
6	PRE_2	Numeric	11	5	Demo & Personality model predicted valuesValue	None

So, to get the correlation between the models, we need only get the correlation between these predicted score variables

We do that running a regular correlation. I also added in the criterion variable, just to have all the values we need in one place…



The results are…

Correlations

		COLLEGE PERFORMANCE -- CRITERION	High School model predicted values	Demo & Personality model predicted valuesValue
COLLEGE PERFORMANCE -- CRITERION	Pearson Correlation	1	.793**	.448**
	Sig. (2-tailed)		.000	.000
	N	118	118	118
High School model predicted values	Pearson Correlation	.793**	1	.538**
	Sig. (2-tailed)	.000		.000
	N	118	118	118
Demo & Personality model predicted valuesValue	Pearson Correlation	.448**	.538**	1
	Sig. (2-tailed)	.000	.000	
	N	118	118	118

**. Correlation is significant at the 0.01 level (2-tailed).

The correlation of college performance and the High School Perf model of .793 matches the R from the High School Perf model from above.

The correlation of college performance and the Demo & Personality model of .448 matches the R from the Demographics and Personalty model from above also.

The correlation of the two models is the correlation between the two sets of predicted score – r = .538.

We will also need the sample size – N = 118

A	B	C	D	E	F	G
Steiger's Z-test - Comparing Correlated Correlations						
		r(1,2) =>	**0.793**			
		r(1,3) =>	**0.448**			
		r(2,3) =>	**0.538**			
		N =>	**118**			
		Z =	**5.574**			
		p =	**2.49E-08**			
Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. Psy						

When we plug in the correlations and the N, get a Z-value of 5.574, which is significant with a p-value of .0000000.249.

We would conclude that models fit the data differentially well. Specifically we would conclude that the High School Performance model predicts College GPA significantly better than does the Demographic and Personality model.

#4 -- Comparing a model across populations

A lot of the time we are working with a single population and want do compare different models within that single population. But, sometime, we have more that one population/group and want to know if a particular model "works differently" for two different populations/groups.

There are two different questions we can ask about if a model "works differently" for two different populations:

- Does the mode "work better" for one population versus the other? Asking if the R-square of the model is higher for one population/group than the other?
- Are the regression weights for the predictors in the model different for the two populations?

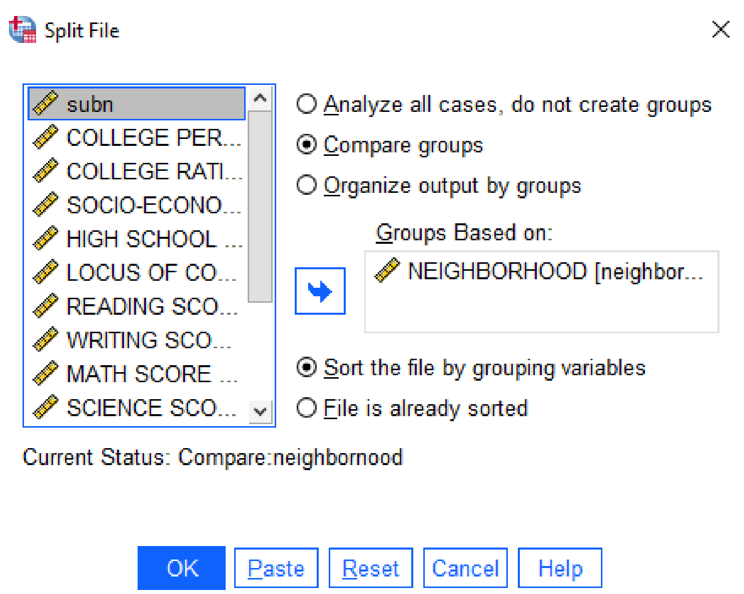
What we are going to do is to get the same model for the two different groups and then make these two comparisons.

The two populations/groups we are going to compare are those who were raised in an "Urban" versus raised in a "Rural" neighborhood'

The model we will use is the full model from above (without the Neighborhood variable – since it is the grouping variable).

First → we have to "split" the sample into the two groups/populations. To do that we will use the "Split File" function in SPSS.

Date → Split File

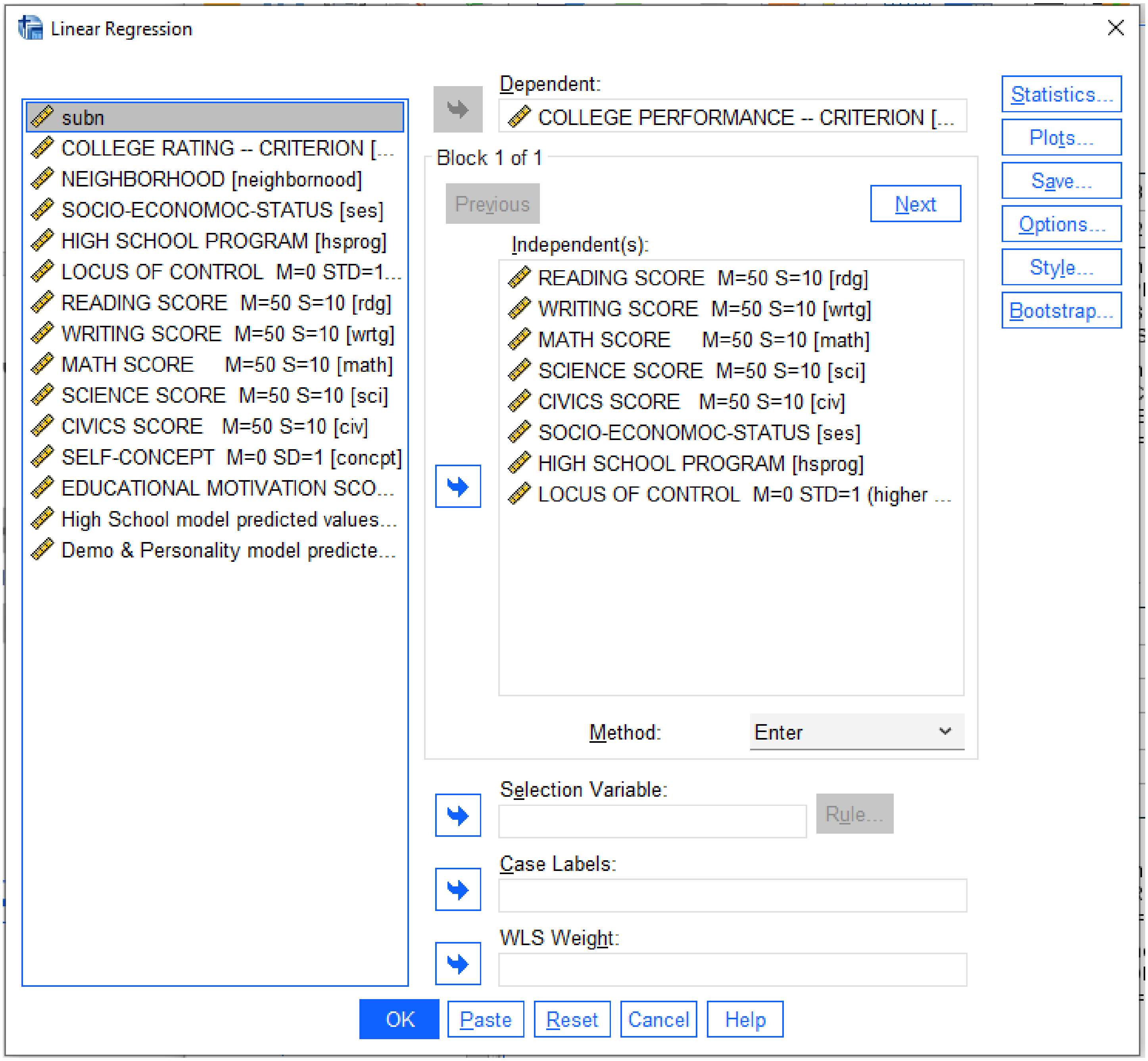


Click the "Compare groups" button

Move the grouping variable into the window and click "OK"

SPSS will sort the data into the two groups. **Every analysis you ask for will now be done twice – once on each group.**

We request the multipe regression model just like before, but now it will do that analysis twice, once for the “Urban” neighborhood group and once for the “Rural” neighborhood group.



Here's the first part of the output – for each group.

Model Summary

NEIGHBORHOOD	Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
URBAN	1	.883[a]	.780	.737	.74431
RURAL	1	.763[b]	.583	.526	.82455

a. Predictors: (Constant), LOCUS OF CONTROL M=0 STD=1 (higher scores - more external), SOCIO-ECONOMOC-STATUS, HIGH SCHOOL PROGRAM, MATH SCORE M=50 S=10, READING SCORE M=50 S=10, CIVICS SCORE M=50 S=10, SCIENCE SCORE M=50 S=10, WRITING SCORE M=50 S=10

b. Predictors: (Constant), LOCUS OF CONTROL M=0 STD=1 (higher scores - more external), HIGH SCHOOL PROGRAM, SOCIO-ECONOMOC-STATUS, MATH SCORE M=50 S=10, WRITING SCORE M=50 S=10, SCIENCE SCORE M=50 S=10, CIVICS SCORE M=50 S=10, READING SCORE M=50 S=10

ANOVA[a]

NEIGHBORHOOD	Model		Sum of Squares	df	Mean Square	F	Sig.
URBAN	1	Regression	80.514	8	10.064	18.167	.000[b]
		Residual	22.714	41	.554		
		Total	103.228	49			
RURAL	1	Regression	56.031	8	7.004	10.302	.000[c]
		Residual	40.113	59	.680		
		Total	96.143	67			

a. Dependent Variable: COLLEGE PERFORMANCE -- CRITERION

b. Predictors: (Constant), LOCUS OF CONTROL M=0 STD=1 (higher scores - more external), SOCIO-ECONOMOC-STATUS, HIGH SCHOOL PROGRAM, MATH SCORE M=50 S=10, READING SCORE M=50 S=10, CIVICS SCORE M=50 S=10, SCIENCE SCORE M=50 S=10, WRITING SCORE M=50 S=10

c. Predictors: (Constant), LOCUS OF CONTROL M=0 STD=1 (higher scores - more external), HIGH SCHOOL PROGRAM, SOCIO-ECONOMOC-STATUS, MATH SCORE M=50 S=10, WRITING SCORE M=50 S=10, SCIENCE SCORE M=50 S=10, CIVICS SCORE M=50 S=10, READING SCORE M=50 S=10

The model has an R-square of .780 for Urban, which is significant.

The model has an R-square of .583 for Rural, which is also significant

To test if the model fits one group/population better than the other, we will use the Computator to perform a Fisher's Z-test of the correlations associated with each model

Here's what that part of the Computator looks like.

Fisher's Z-test - comparing a corrrelation across groups		
Group 1	**r(1,2)** =>	**0.883**
	n =>	**50**
Group 2	**r(1,2)** =>	**0.763**
	n =>	**68**
	Z =	2.015
	p =	0.0439

Snedcor, G. W. & Cochran, W. G. (1967). Statistical Methods (6th ed.). Iowa State University Press.

Notice! Even though we are intending to compare the R-square from the two models the Fisher's Z-test compares the R values.

Notice that the .883 and .763 are the R values (not the R-square values) from the Urban and Rural models, respectively.

The "N" for each model is derived from the Total degrees of freedom (df) given in the ANOVA table for each model. Total df = N-1, so… N = Total df + 1

The N for the Urban model is 49 + 1 = 50

The N for the Rural model is 67 + 1 = 68

With these values, we get a Z-value of 2.015 and a p of .0439

We would conclude that the model works better for the Rural group than for the Urban group.

Here are the regression weights for the model applied to the two groups.

NEIGHBORHOOD	Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
URBAN	1	(Constant)	-4.020	.792		-5.073	.000
		READING SCORE M=50 S=10	.035	.009	.385	3.901	.000
		WRITING SCORE M=50 S=10	.022	.016	.162	1.355	.183
		MATH SCORE M=50 S=10	.033	.011	.288	2.967	.005
		SCIENCE SCORE M=50 S=10	.025	.016	.184	1.615	.114
		CIVICS SCORE M=50 S=10	.030	.017	.204	1.791	.081
		SOCIO-ECONOMOC-STATUS	-.305	.160	-.158	-1.906	.064
		HIGH SCHOOL PROGRAM	.258	.183	.109	1.409	.166
		LOCUS OF CONTROL M=0 STD=1 (higher scores - more external)	.069	.174	.033	.398	.693
RURAL	1	(Constant)	-1.945	.768		-2.532	.014
		READING SCORE M=50 S=10	.030	.010	.392	3.072	.003
		WRITING SCORE M=50 S=10	.048	.016	.412	2.923	.005
		MATH SCORE M=50 S=10	.018	.008	.233	2.186	.033
		SCIENCE SCORE M=50 S=10	-.008	.014	-.068	-.571	.570
		CIVICS SCORE M=50 S=10	.001	.016	.009	.070	.944
		SOCIO-ECONOMOC-STATUS	.350	.163	.199	2.145	.036
		HIGH SCHOOL PROGRAM	-.044	.214	-.018	-.208	.836
		LOCUS OF CONTROL M=0 STD=1 (higher scores - more external)	-.029	.155	-.018	-.185	.854

a. Dependent Variable: COLLEGE PERFORMANCE -- CRITERION

We would conclude that the “structure” of the mode is different for the two groups.

For the Urban model only Reading & Math have significant individual contributions.

For the Rural model Reading, Writing, Math & Socio-Economic Status have significant individual contributions.

Interestingly, even though the Rural model has more significant contributors, it has a poorer fit to the model (lower R-square).

Supplement to #4 – Defining "groups" with a quantitative variable

Sometimes we will use a quantitative variable to define the "groups" we want to compare. To do this we will "recode" the quantitative variable into a new variable that has two "groups".

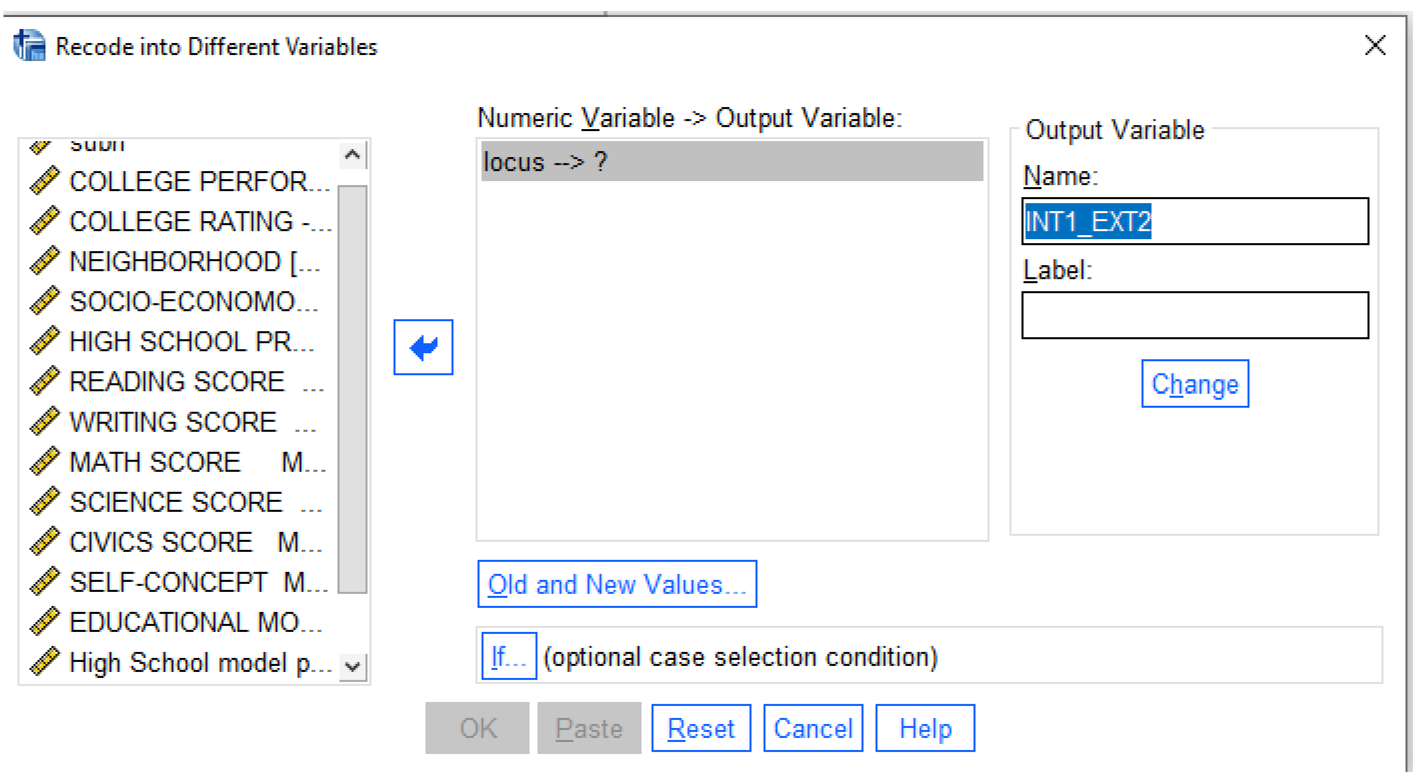
There are three common ways of doing this. Here's an example of each.

Using Previously Defined Values to Assign Groups

Sometimes previous use of a variable has established "cutoff values" for defining groups. Take for example, for the Internal/External Locus of control variable. Sometimes people use this as a quantitative variable – people who have higher scores tend to attribute their success to "external forces" such as luck or other persons or groups, while people who have lower scores tend to attribute that they are responsible for their own success. However, "middle scores" on this variable tend to be sort of "mushy", resulting from a combination of "internal" and "external" attributions. Some prefer to use the variable to identify "internalizers" and "externalizers". For the particular measure of I/E used in this study, previous research has led to the use of ".25" as a cutoff: 1) scores below -.25 define Internalizers, 2) scores above .25 define externalizers, and 3) people with scores between these cutoffs are not grouped (and are dropped from the analysis).

We would do this in SPSS using **Transform → Recode Into Different Variables**

Highlight the locus of control variable and use the arrow to move it into the middle box. Then type the name of the new variable you are making in to the "Name" box (I chose the name INT1_EXT2, meaning that those identified as internal are coded 1 and those identified as external are coded 2).



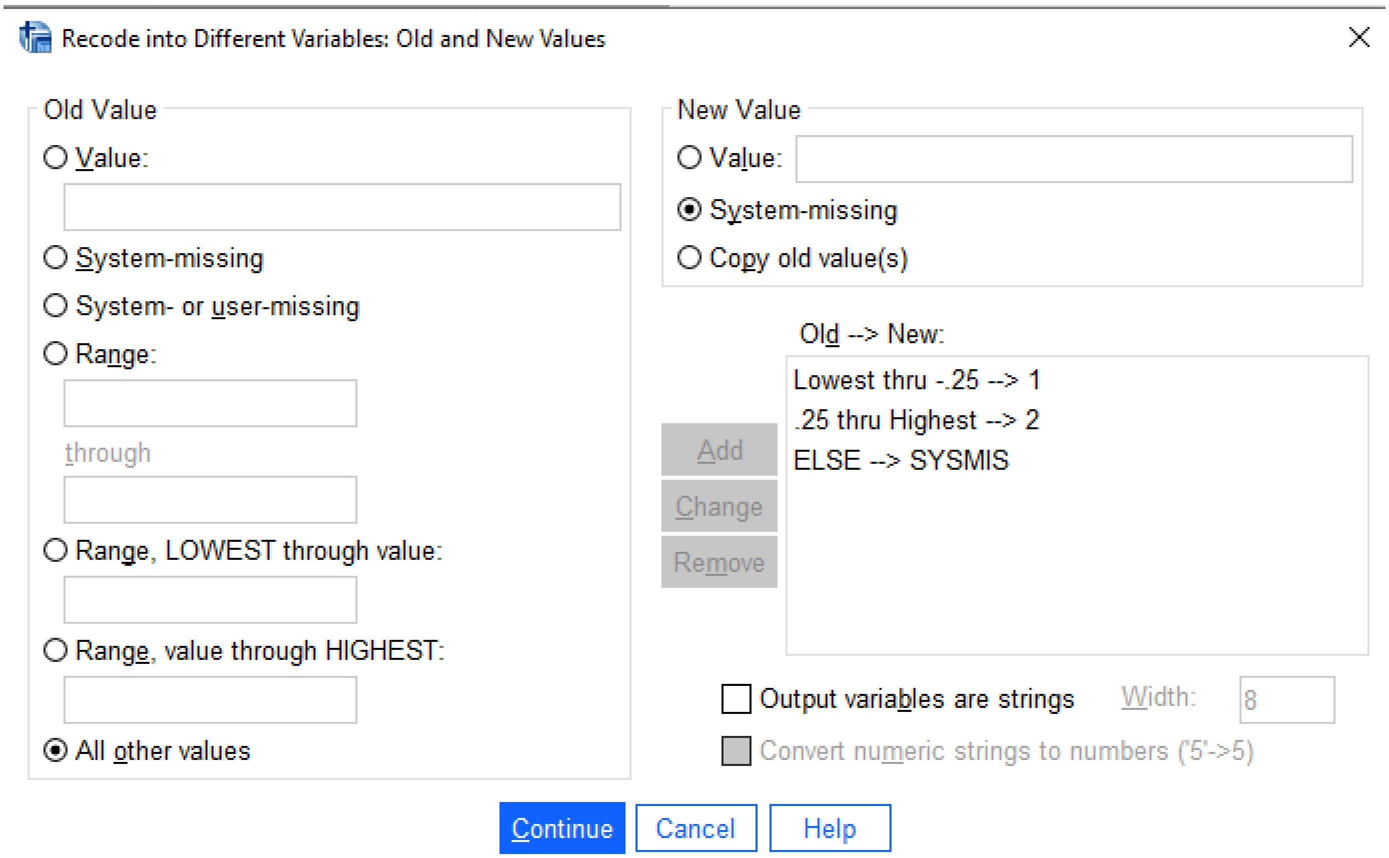
Then click the "Change" box

Then click on the "Old and New Values" box.

Use the "Old Value" and "New Value" choices to define the groups.

Lowest through -.25 → 1 .25 through highest → 2 "All other values" → System Missing

The click "Continue" and click "OK" on the Recode window.



Here's a frequency analysis for this new variable

INT1_EXT2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	34	28.8	38.2	38.2
	2.00	55	46.6	61.8	100.0
	Total	89	75.4	100.0	
Missing	System	29	24.6		
Total		118	100.0		

We identified 34 "internalizers" & 55 "externalizers" and 29 cases were given missing values.

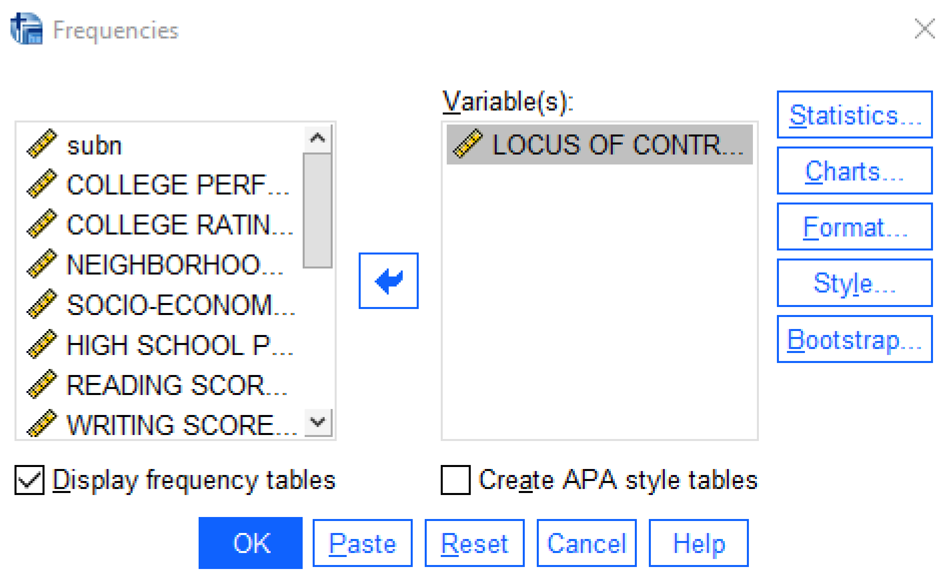
Using a "Median-Split" to Assign Groups

One of the "classic" ways of turning a quantitative variable in to a 2-group variable is to divide the distribution of scores "in half". The median of a set of scores defines the value which has ½ of the data values smaller than it and ½ of the data values larger than it.

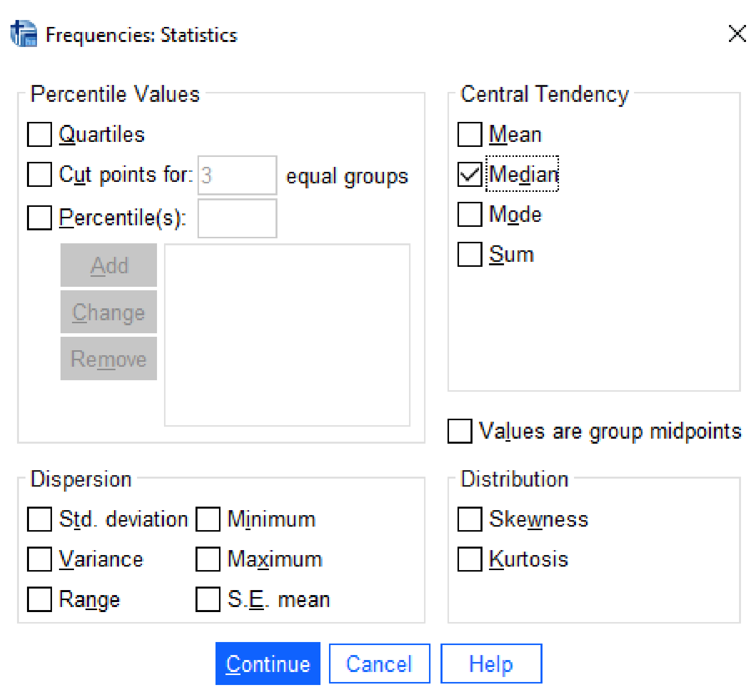
First we have to get the median, and then we use that value to perform the recode.

Analyze → Descriptive Statistics → Frequencies

Move the quantitative variable we are starting with into the box.



Then click on the "Statistics" box.



Be sure “Median” is checked. Click “Continue” and then click “OK” on the Frequencies window.

Frequencies

Statistics

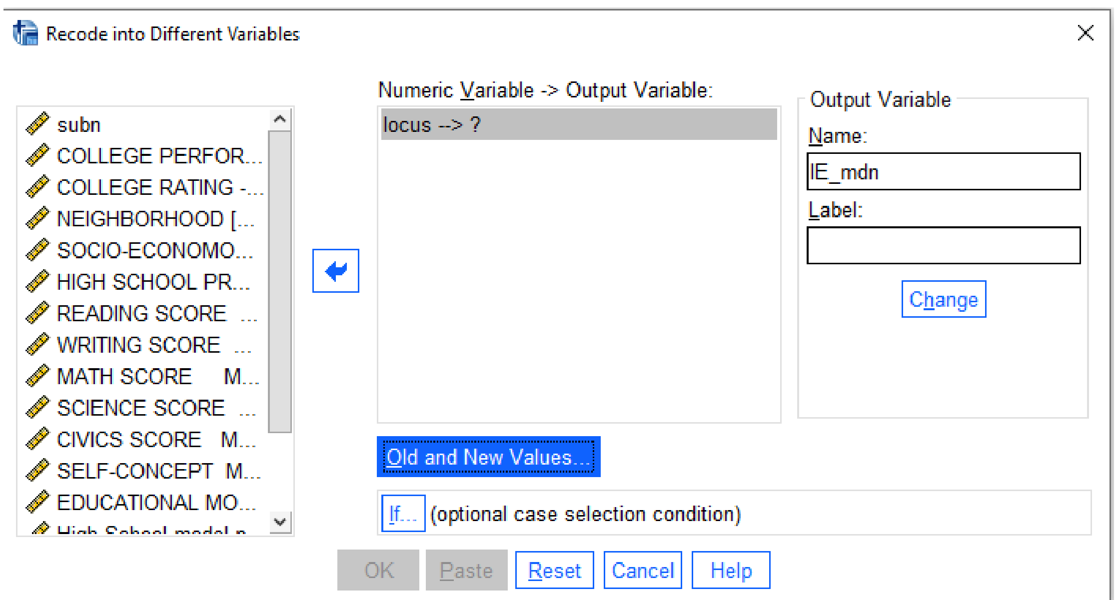
LOCUS OF CONTROL M=0 STD=1 (higher scores - more external)

N	Valid	118
	Missing	0
Median		.2150

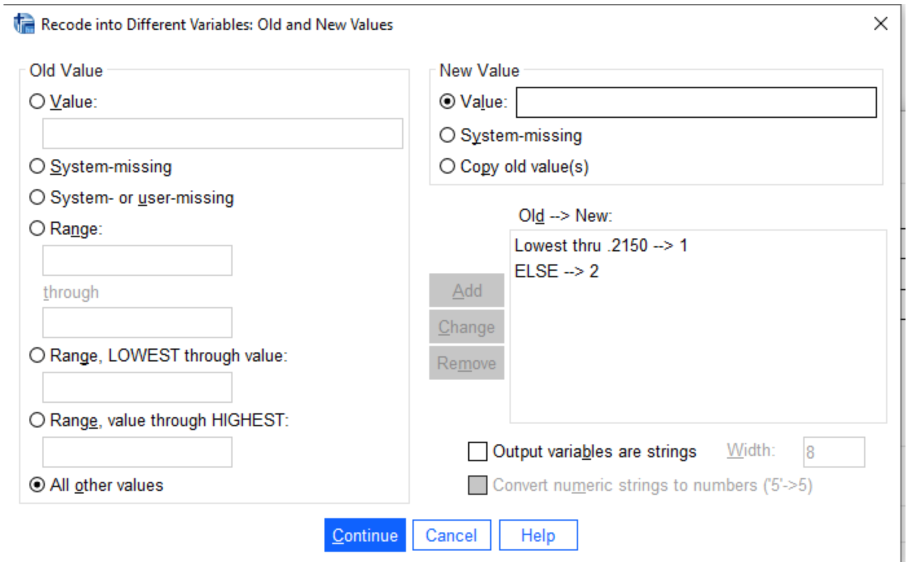
Based on this, we’d use the value of .2150 to split the sample into two groups.

Transform → Recode Into Different Variables

Select the starting variable and name the new variable (IE_mdn means a median split of the Internal/External variable).



Use the median value of .2150 to identify the two groups.



Here's a frequency analysis of the new variable.

IE_mdn

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	59	50.0	50.0	50.0
	2.00	59	50.0	50.0	100.0
	Total	118	100.0	100.0	

We identified 59 "internalizers" & 59 "externalizers".

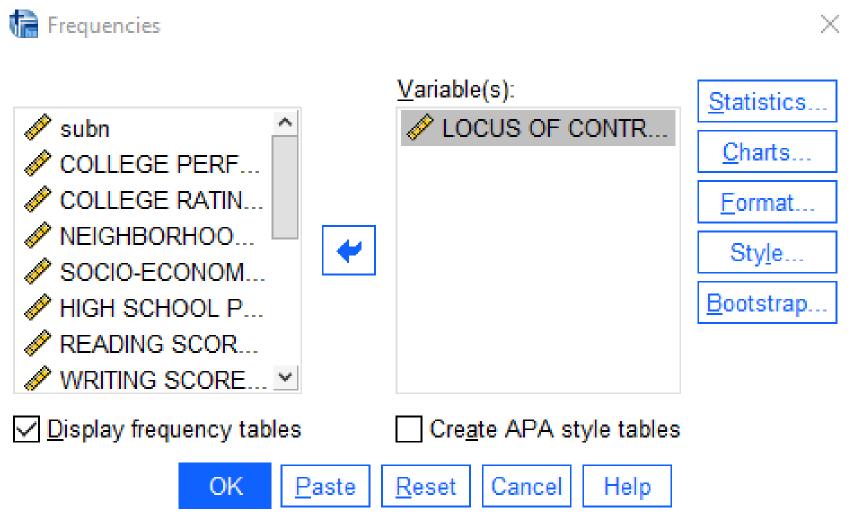
Using a "Outside-Splits" to Assign Groups

Some people don't like using a median split because people with very similar scores, but just above versus just below the median are "qualitatively different" (i.e., put into different groups). To avoid this problem people will use "outside groups" or "extreme groups" and drop the "cases in the middle". Common versions of this are to take the upper and lower one-fourths or to take the upper and lower one-thirds.

Here's an example using Frequencies to obtain the values to split the sample into "thirds". Then we'll use those values to identify the "top" and "bottom" of the distribution (and discard the middle third).

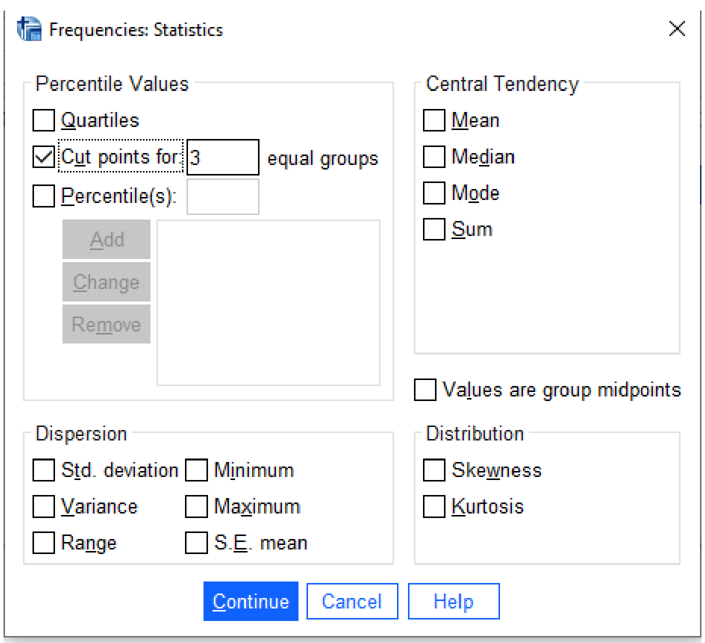
Analyze → Descriptive Statistics → Frequencies

Move the quantitative variable we are starting with into the box.



Then click on the "Statistics" box.

Be sure the "Cut points for" box is checked. Put "3" into the textbox, asking for the values to split the distribution into three equal groups



Click “Continue” and then click “OK” on the Frequencies window.

Here’s the output

Frequencies

Statistics

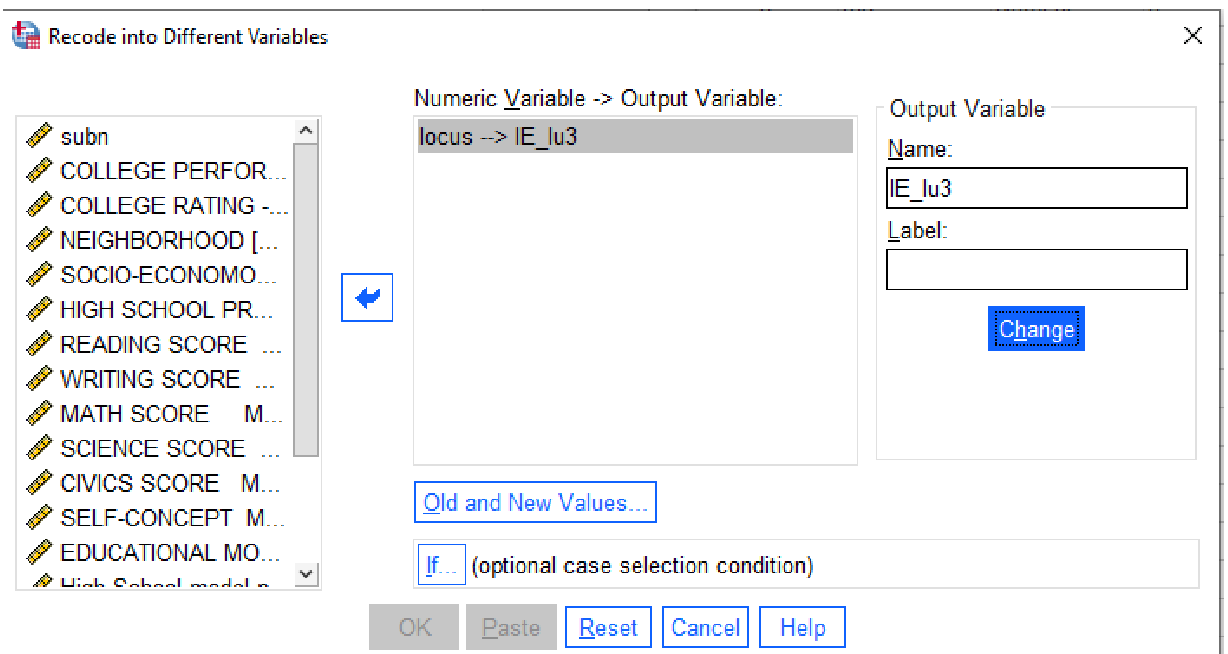
LOCUS OF CONTROL M=0 STD=1 (higher scores - more external)

N	Valid	118
	Missing	0
Percentiles	33.33333333	-.1667
	66.66666667	.4500

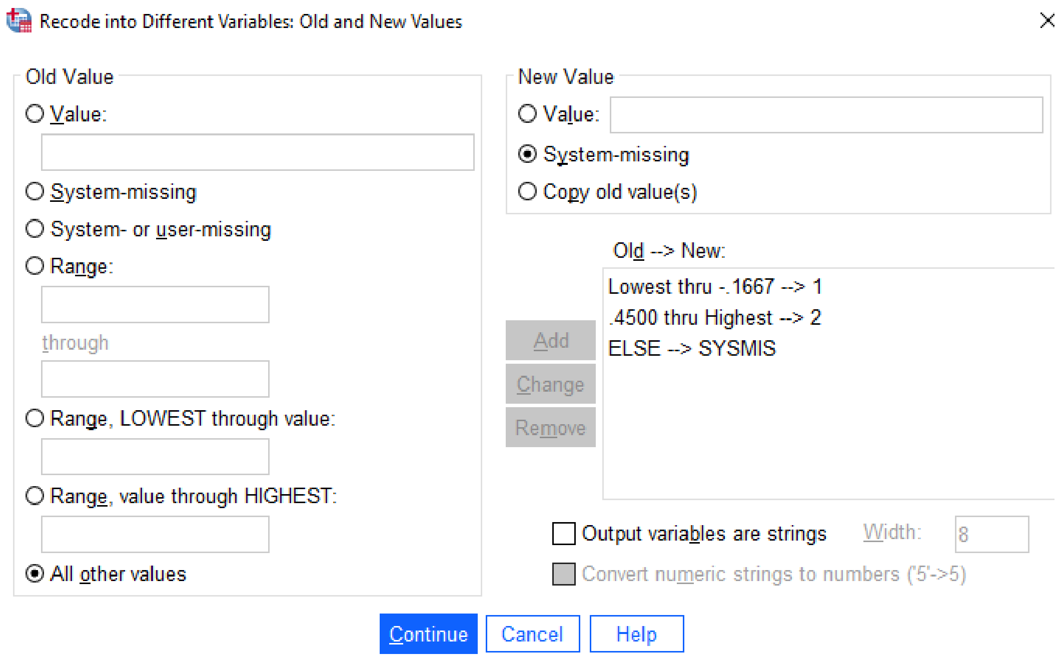
These are the values we’ll use to identify the “bottom third” as “internalizers” and the “top third” as “externalizers” (and we’ll set the “middle third” as missing values).

Transform → Recode Into Different Variables

Select the starting variable and name the new variable (IE_lu3 means form groups that are the lower and upper third of the IE distribution).



Now we use the values from the Frequencies to form the groups



Here's a Frequencies of the newly created variable

IE_Iu3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	39	33.1	48.8	48.8
	2.00	41	34.7	51.2	100.0
	Total	80	67.8	100.0	
Missing	System	38	32.2		
Total		118	100.0		

We identified 39 “internalizers” & 41 “externalizers” and 38 cases were given missing values.