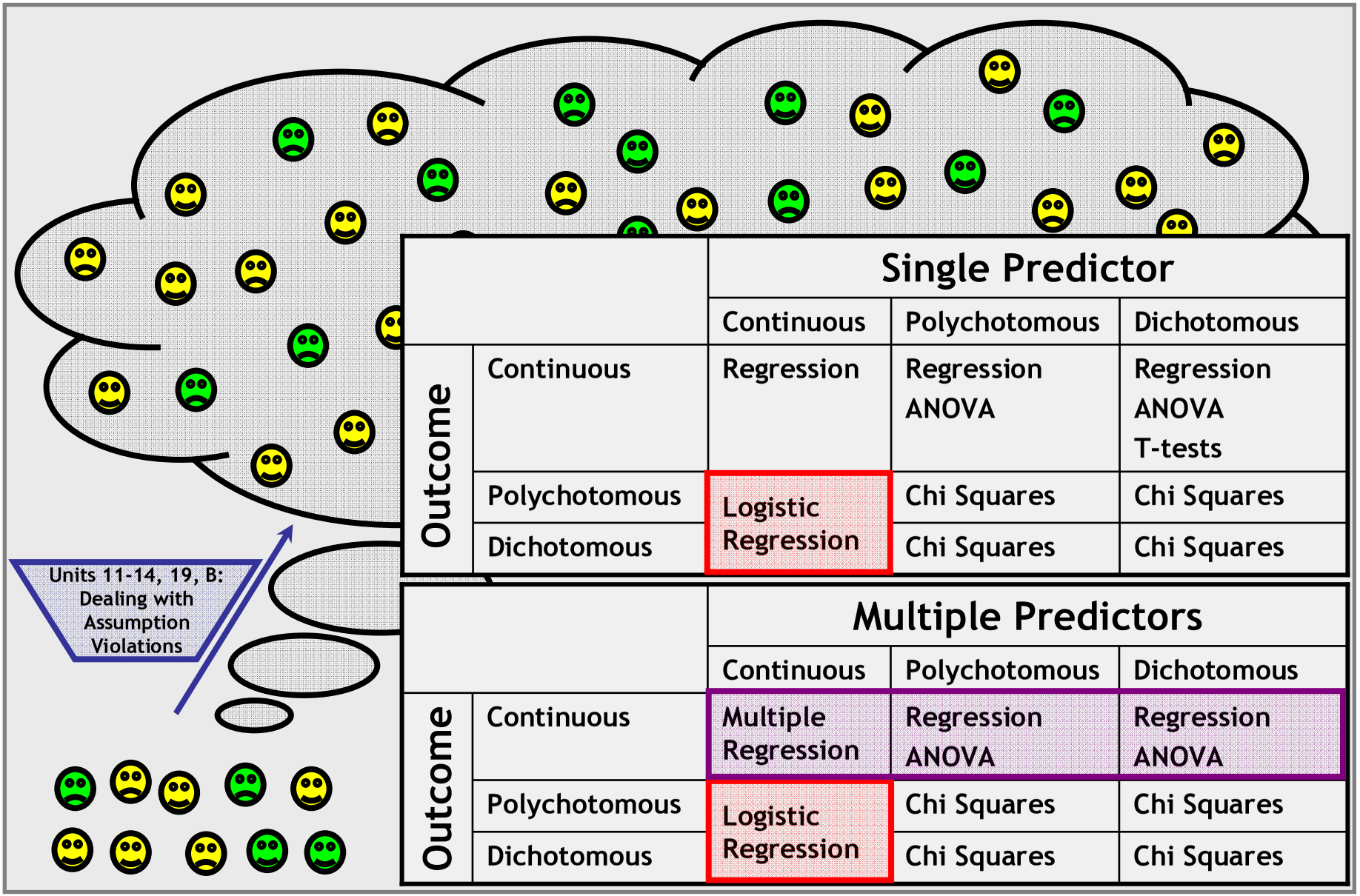
Unit 15: Road Map (VERBAL)

Nationally Representative Sample of 7,800 8th Graders Surveyed in 1988 (NELS 88).
Outcome Variable (aka Dependent Variable):
READING, a continuous variable, test score, mean = 47 and standard deviation = 9
Predictor Variables (aka Independent Variables):
Question Predictor-
RACE, a polychotomous variable, 1 = Asian, 2 = Latino, 3 = Black and 4 = White
Control Predictors-
HOMEWORK, hours per week, a continuous variable, mean = 6.0 and standard deviation = 4.7
FREELUNCH, a proxy for SES, a dichotomous variable, 1 = Eligible for Free/Reduced Lunch and 0 = Not
ESL, English as a second language, a dichotomous variable, 1 = ESL, 0 = native speaker of English

- Unit 11: What is measurement error, and how does it affect our analyses?
- Unit 12: What tools can we use to detect assumption violations (e.g., ***outliers***)?
- Unit 13: How do we deal with violations of the ***linearity*** and ***normality*** assumptions?
- Unit 14: How do we deal with violations of the ***homoskedasticity*** assumption?
- Unit 15: What are the correlations among reading, race, ESL, and homework, controlling for SES?
- Unit 16: Is there a relationship between reading and race, controlling for SES, ESL and homework?
- Unit 17: Does the relationship between reading and race vary by levels of SES, ESL or homework?
- Unit 18: What are sensible strategies for building complex statistical models from scratch?
- Unit 19: How do we deal with violations of the ***independence*** assumption (using ANOVA)?

Unit 15: Road Map (Schematic)



		Single Predictor		
		Continuous	Polychotomous	Dichotomous
Outcome	Continuous	Regression	Regression ANOVA	Regression ANOVA T-tests
	Polychotomous	Logistic Regression	Chi Squares	Chi Squares
	Dichotomous		Chi Squares	Chi Squares

		Multiple Predictors		
		Continuous	Polychotomous	Dichotomous
Outcome	Continuous	Multiple Regression	Regression ANOVA	Regression ANOVA
	Polychotomous	Logistic Regression	Chi Squares	Chi Squares
	Dichotomous		Chi Squares	Chi Squares

Unit 15: Roadmap (SPSS Output)

Correlations

		READING	NUMBER OF HRS SPENT ON HOMEWORK PER WEEK	ESL	FREELUNCH
READING	Pearson Correlation	1.000	.183**	-.053**	-.267**
	Sig. (2-tailed)		.000	.000	.000
	N	7800.000	7800	7800	7800
NUMBER OF HRS SPENT ON HOMEWORK PER WEEK	Pearson Correlation	.183**	1.000	.005	-.092**
	Sig. (2-tailed)	.000		.648	.000
	N	7800	7800.000	7800	7800
ESL	Pearson Correlation	-.053**	.005	1.000	.093**
	Sig. (2-tailed)	.000	.648		.000
	N	7800	7800	7800.000	7800
FREELUNCH	Pearson Correlation	-.267**	-.092**	.093**	1.000
	Sig. (2-tailed)	.000	.000	.000	
	N	7800	7800	7800	7800.000

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

Unit 4

Correlations

Control Variables			READING	NUMBER OF HRS SPENT ON HOMEWORK PER WEEK	ESL
FREELUNCH	READING	Correlation	1.000	.165	-.029
		Significance (2-tailed)	.	.000	.009
		df	0	7797	7797
	NUMBER OF HRS SPENT ON HOMEWORK PER WEEK	Correlation	.165	1.000	.014
		Significance (2-tailed)	.000	.	.222
		df	7797	0	7797
	ESL	Correlation	-.029	.014	1.000
		Significance (2-tailed)	.009	.222	.
		df	7797	7797	0

Unit 15

Unit 15: Partial Correlation Matrices

Unit 15 Post Hole:

Interpret a correlation matrix and/or partial correlation matrix and note what they may foreshadow about multiple regression.

Unit 15 Technical Memo and School Board Memo:

Use a correlation matrix and a partial correlation matrix to get a handle on four variables of your choice (one continuous outcome variable, one predictor variable, and two control variables) in preparation for multiple regression.

Unit 15 Review:

Review Units 4 and 5.

Unit 15: Technical Memo and School Board Memo

Work Products (Part I of II):

I. Technical Memo: Have one section per analysis. For each section, follow this outline.

- A. Introduction
  - i. State a theory (or perhaps hunch) for the relationship—think causally, be creative. (1 Sentence)
  - ii. State a research question for each theory (or hunch)—think correlationally, be formal. Now that you know the statistical machinery that justifies an inference from a sample to a population, begin each research question, "In the population,…" (1 Sentence)
  - iii. List your variables, and label them "outcome" and "predictor," respectively.
  - iv. Include your theoretical model.
- B. Univariate Statistics. Describe your variables, using descriptive statistics. What do they represent or measure?
  - i. Describe the data set. (1 Sentence)
  - ii. Describe your variables. (1 Paragraph Each)
    - a. Define the variable (parenthetically noting the mean and s.d. as descriptive statistics).
    - b. Interpret the mean and standard deviation in such a way that your audience begins to form a picture of the way the world is. Never lose sight of the substantive meaning of the numbers.
    - c. Polish off the interpretation by discussing whether the mean and standard deviation can be misleading, referencing the median, outliers and/or skew as appropriate.
    - d. Note validity threats due to measurement error.
- C. Correlations. Provide an overview of the relationships between your variables using descriptive statistics. Focus first on the relationship between your outcome and question predictor, second-tied on the relationships between your outcome and control predictors, second-tied on the relationships between your question predictor and control predictors, and fourth on the relationship(s) between your control variables.
  - a. Include your own simple/partial correlation matrix with a well-written caption.
  - b. Interpret your simple correlation matrix. Note what the simple correlation matrix foreshadows for your partial correlation matrix; "cheat" here by peeking at your partial correlation and thinking backwards. Sometimes, your simple correlation matrix reveals possibilities in your partial correlation matrix. Other times, your simple correlation matrix provides foregone conclusions. You can stare at a correlation matrix all day, so limit yourself to two insights.
  - c. Interpret your partial correlation matrix controlling for one variable. Note what the partial correlation matrix foreshadows for a partial correlation matrix that controls for two variables. Limit yourself to two insights.

Unit 15: Technical Memo and School Board Memo

Work Products (Part II of II):

I. Technical Memo (continued)

- D. Regression Analysis. Answer your research question using inferential statistics. Weave your strategy into a coherent story.
  - i. Include your fitted model.
  - ii. Use the R^2 statistic to convey the goodness of fit for the model (i.e., strength).
  - iii. To determine statistical significance, test each null hypothesis that the magnitude in the population is zero, reject (or not) the null hypothesis, and draw a conclusion (or not) from the sample to the population.
  - iv. Create, display and discuss a table with a taxonomy of fitted regression models.
  - v. Use spreadsheet software to graph the relationship(s), and include a well-written caption.
  - vi. Describe the direction and magnitude of the relationship(s) in your sample, preferably with illustrative examples. Draw out the substance of your findings through your narrative.
  - vii. Use confidence intervals to describe the precision of your magnitude estimates so that you can discuss the magnitude in the population.
  - viii. If regression diagnostics reveal a problem, describe the problem and the implications for your analysis and, if possible, correct the problem.
    - i. Primarily, check your residual-versus-fitted (RVF) plot. (Glance at the residual histogram and P-P plot.)
    - ii. Check your residual-versus-predictor plots.
    - iii. Check for influential outliers using leverage, residual and influence statistics.
    - iv. Check your main effects assumptions by checking for interactions before you finalize your model.
- X. Exploratory Data Analysis. Explore your data using outlier resistant statistics.
  - i. For each variable, use a coherent narrative to convey the results of your exploratory univariate analysis of the data. Don't lose sight of the substantive meaning of the numbers. (1 Paragraph Each)
    1. Note if the shape foreshadows a need to nonlinearly transform and, if so, which transformation might do the trick.
  - ii. For each relationship between your outcome and predictor, use a coherent narrative to convey the results of your exploratory bivariate analysis of the data. (1 Paragraph Each)
    1. If a relationship is non-linear, transform the outcome and/or predictor to make it linear.
    2. If a relationship is heteroskedastic, consider using robust standard errors.

II. School Board Memo: Concisely, precisely and plainly convey your key findings to a lay audience. Note that, whereas you are building on the technical memo for most of the semester, your school board memo is fresh each week. (Max 200 Words)

III. Memo Metacognitive

Unit 15: Research Question

Theory: Head Start programs provide educationally disadvantaged preschoolers the skills and knowledge to start kindergarten on a level playing field.

Research Question: Controlling for ***SES***, ***ESL*** and ***AGE***, is ***GENERALKNOWLEDGE*** positively correlated with ***HEADSTARTHOURS*** for Latina kindergarteners.

Data Set: ECLS (Early Childhood Longitudinal Study) subset of Latinas with no missing data for the variables below (n = 816)

Variables:

Outcome: (***GENERALKNOWLEDGE***) IRT Scaled Score on a Standardized Test of General Knowledge in Kindergarten

Question Predictor: (***HEADSTARTHOURS***) Hours Per Week of Head Start in the Year Before Kindergarten

Control Predictors:

(***SES***) A Composite Measure of the Family's Socioeconomic Status

(***ESL***) A Dichotomy for which 1 Denotes that English is a 2nd Language (0 = Not)

(***AGE***) Age in Months at Kindergarten Entry

Model: $$GENERALKNOWLEDGE = \beta_0 + \beta_1 HEADSTARTHOURS + \beta_2 SES + \beta_3 ESL + \beta_4 AGE + \varepsilon$$

SPSS DATA

*ECLSLATINASHSK.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1 : GENERALKNOWLEDGE 17.497 Visible: 13 of 13 Variables

	GENERALKNOWLEDGE	HEADSTARTHOURS	SES	ESL	AGE	var	var	var	var
1	17.50	0	-1.10	0	60				
2	16.19	0	-1.08	0	64				
3	20.63	17	-0.33	0	61				
4	17.76	0	-0.49	0	67				
5	18.42	3	0.67	0	68				

Data View Variable View

SPSS Processor is ready

ECLSLATINASHSK.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align
1	GENERALKNO...	Numeric	7	3	General Knowle...	None	None	10	Right
2	HEADSTARTH...	Numeric	2	0	Number of Hea...	None	None	9	Right
3	SES	Numeric	6	2	Socioeconomic...	None	None	8	Right
4	ESL	Numeric	8	2	English as a 2n...	{0.00, Englis...	None	10	Right
5	AGE	Numeric	8	2	Age in Months	None	None	10	Right

Data View Variable View

SPSS Processor is ready

Simple Correlation Matrix

Correlations

		General Knowledge IRT Scaled Score	Number of Head Start Hours Per Week	Age in Months	English as a 2nd Language	Socioeconomic Status Composite Score
General Knowledge IRT Scaled Score	Pearson Correlation	1.000	-.122**	.247**	-.332**	.433**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	816.000	816	816	816	816
Number of Head Start Hours Per Week	Pearson Correlation	$R^2 = .01$	1.000	.019	.152**	-.242**
	Sig. (2-tailed)			.581	.000	.000
	N		816.000	816	816	816
Age in Months	Pearson Correlation	$R^2 = .06$	$R^2 = .00$	1.000	-.038	.033
	Sig. (2-tailed)				.278	.354
	N			816.000	816	816
English as a 2nd Language	Pearson Correlation	$R^2 = .11$	$R^2 = .02$	$R^2 = .00$	1.000	-.201**
	Sig. (2-tailed)					.000
	N				816.000	816
Socioeconomic Status Composite Score	Pearson Correlation	$R^2 = .19$	$R^2 = .06$	$R^2 = .00$	$R^2 = .04$	1.000
	Sig. (2-tailed)					
	N					816.000

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

Let's call an R^2 statistic of .00 "no correlation" (even though, if we go out to enough decimal places, there will be some correlation)

Let's call an R^2 statistic from .01 to .05 a "weak correlation"

Let's call an R^2 statistic from .06 to .15 a "moderate correlation"

Let's call an R^2 statistic greater than .15 a "strong correlation"

Whether a correlation is strong or weak is relative. Never believe a chart that implies otherwise.

Simple Correlation Matrix

Correlations

		General Knowledge IRT Scaled Score	Number of Head Start Hours Per Week	Age in Months	English as a 2nd Language	Socioeconomic Status Composite Score
General Knowledge IRT Scaled Score	Pearson Correlation	1.000	-.122**	.247**	-.332**	.433**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	816.000	816	816	816	816
Number of Head Start Hours Per Week	Pearson Correlation		1.000	.019	.152**	-.242**
	Sig. (2-tailed)			.581	.000	.000
	N		816.000	816	816	816
Age in Months	Pearson Correlation			1.000	-.038	.033
	Sig. (2-tailed)				.278	.354
	N			816.000	816	816
English as a 2nd Language	Pearson Correlation				1.000	-.201**
	Sig. (2-tailed)					.000
	N				816.000	816
Socioeconomic Status Composite Score	Pearson Correlation					1.000
	Sig. (2-tailed)					
	N					816.000

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

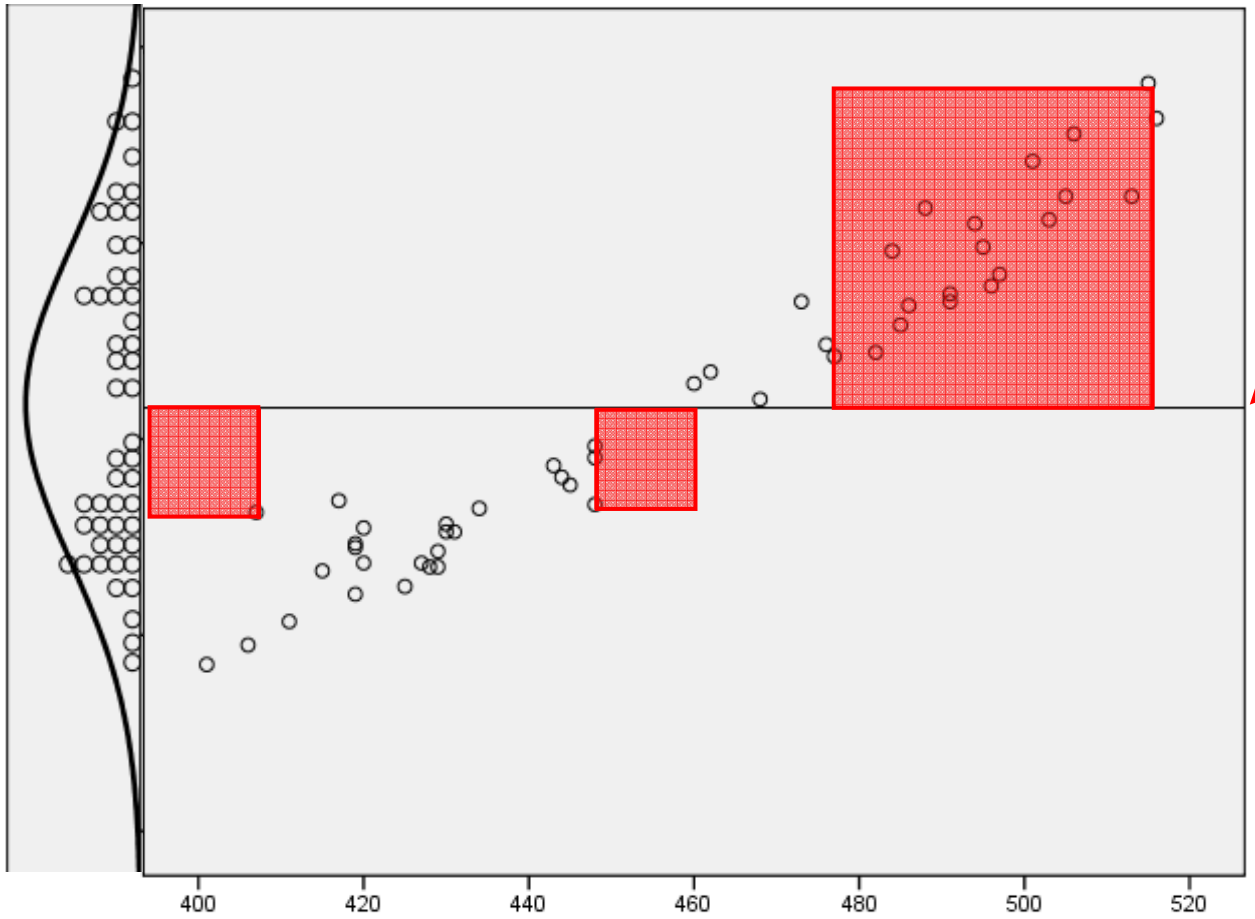
Notice that *GENERALKNOWLEDGE* and *SES* have a strong correlation.

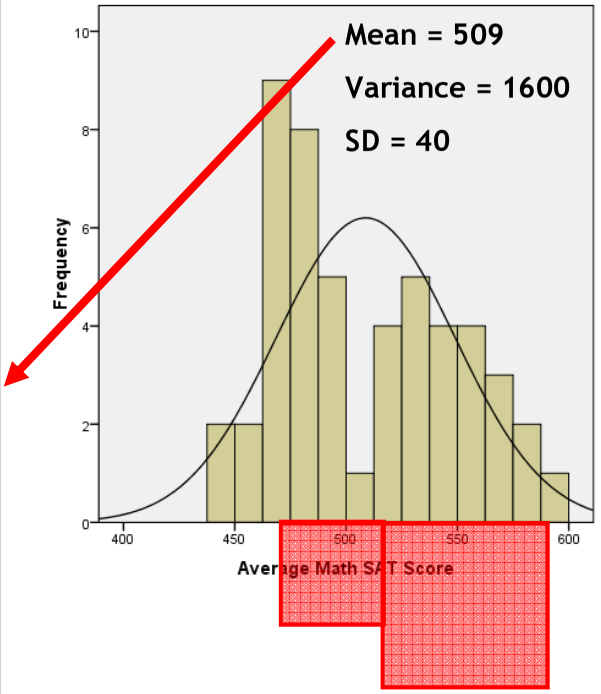
Notice that *GENERALKNOWLEDGE* and *HEADSTARTHOURS* have a weak correlation.

Notice that *HEADSTARTHOURS* and *SES* have a moderate correlation.

Also notice that AGE has a moderate correlation with *GENERALKNOWLEDGE* but no correlation with *HEADSTARTHOURS*, *ESL* or *SES*.

What Do Those Circles Really Represent? Variance (Unit 5 Redux I)





This square represents the average squared mean deviation, in a word, THE variance.

Variance is just a hard working number trying, trying, trying to summarize the variation of a univariate distribution. It is one of many statistical summaries of variation, including range, midspread and standard deviation. Variance is the average squared deviation from the mean.

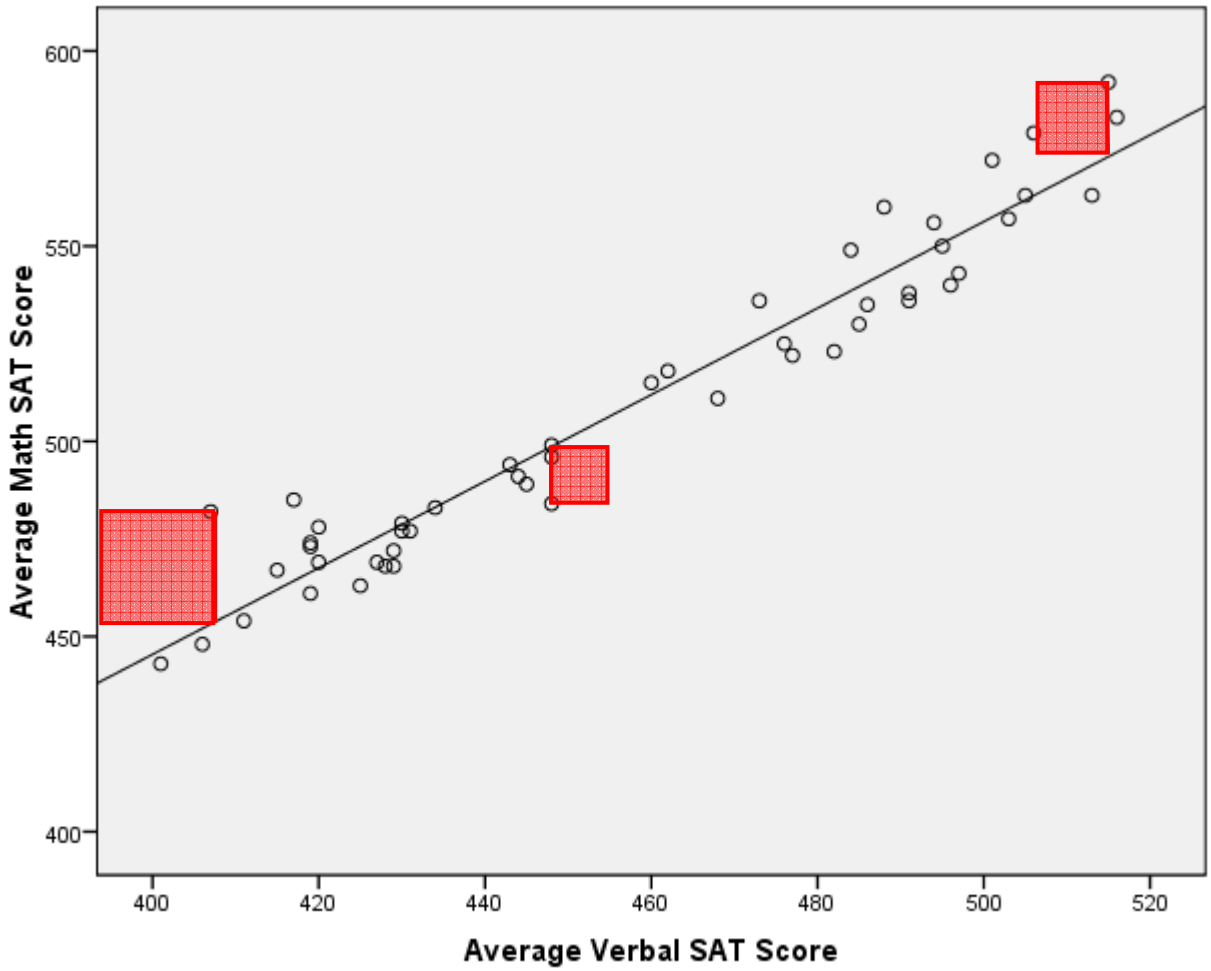
ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	74562.936	1	74562.936	771.068	.000[a]
	Residual	4641.644	48	96.701		
	Total	79204.580	49			

a. Predictors: (Constant), Average Verbal SAT Score

b. Dependent Variable: Average Math SAT Score

What Do Those Circles Really Represent? Variance (Unit 5 Redux II)



Why Residuals?
Unaccounted Variables
Measurement Error
Individual Variation

The mean square residual (or mean square error) is the variance* of the residuals:

*Not quite…notice the degrees of freedom.

The mean square residual (or error) represents the variance in the outcome that is left over after we fit our model. It is an average. Every observation has a residual. We can square that residual. The mean square residual is just the average squared residual.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	74562.936	1	74562.936	771.068	.000[a]
	Residual	4641.644	48	96.701		
	Total	79204.580	49			

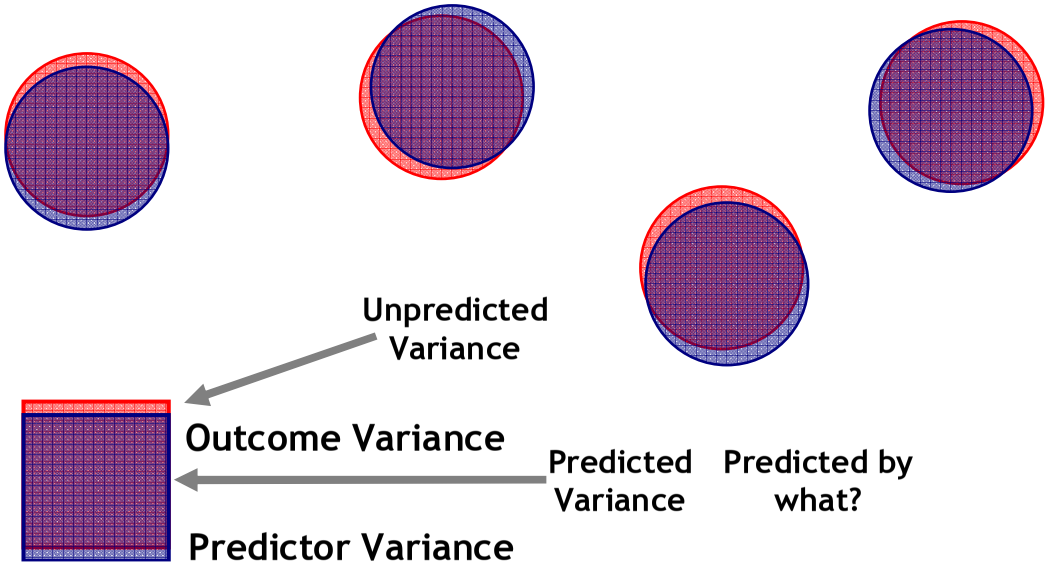
a. Predictors: (Constant), Average Verbal SAT Score

b. Dependent Variable: Average Math SAT Score

What Do Those Circles Really Represent? Variance (Unit 5 Redux III)

That small square is the variance in the outcome still in need of predicting AFTER the (one) predictor has done all its predictive work. To see how small is small, we can compare the variance-still-in-need-of-predicting with the original variance…

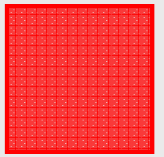
Now, what do those circles really represent? They just represent variance. Instead of squares, we use circles. Although I'm tempted to use "Boolean squares" in the future for the sake of clarity.



The mean square residual (or mean square error) is the variance* of the residuals:

*Not quite…notice the degrees of freedom.

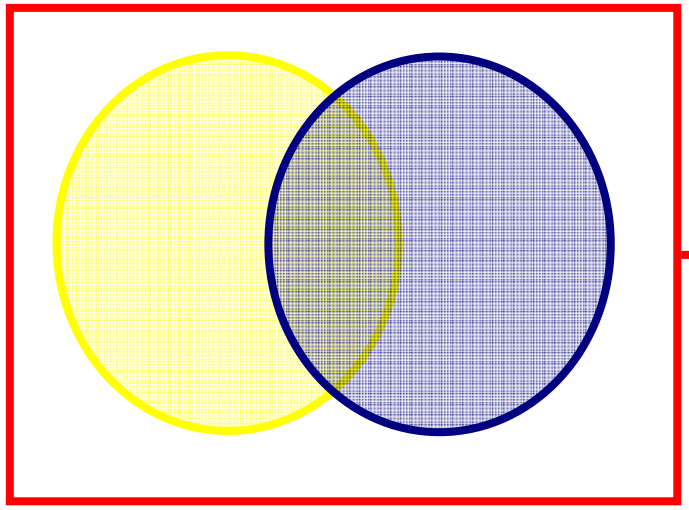
This square represents the average squared mean deviation, in a word, THE variance.



Notice that the outcome variance and the predictor variance are identical in size. That's because (for conceptual purposes) we standardized both the outcome and predictor so that each mean is zero and each standard deviation is one. If the standard deviation is one, then the variance is also one. I.e., if a side of the square is one, then the area of the square is also one. By standardizing, we compare apples to apples.

Also, notice that if the predictor overlaps 95% of the outcome, then the outcome overlaps 95% of the predictor. I.e., the outcome predicts the predictor just as well as the predictor predicts the outcome. Correlations are symmetrical!

Three's Company

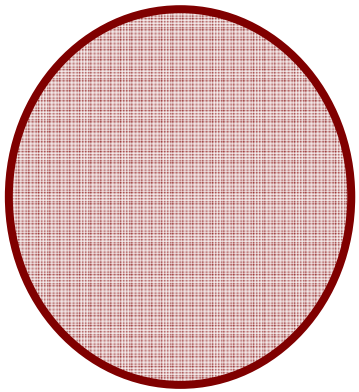


Correlations

		General Knowledge IRT Scaled Score	Number of Head Start Hours Per Week	Age in Months	English as a 2nd Language	Socioeconomic Status Composite Score
General Knowledge IRT Scaled Score	Pearson Correlation	1.000	-.122**	.247**	-.332**	.433**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	816.000	816	816	816	816
Number of Head Start Hours Per Week	Pearson Correlation		1.000	.019	.152**	-.242**
	Sig. (2-tailed)			.581	.000	.000
	N		816.000	816	816	816
Age in Months	Pearson Correlation			1.000	-.038	.033
	Sig. (2-tailed)				.278	.354
	N			816.000	816	816
English as a 2nd Language	Pearson Correlation				1.000	-.201**
	Sig. (2-tailed)					.000
	N				816.000	816
Socioeconomic Status Composite Score	Pearson Correlation					1.000
	Sig. (2-tailed)					
	N					816.000

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

If we look at all three circles overlapping simultaneously, the yellow and blue can stay where they are, and the red must take a small bite out of the yellow and a medium bite out of the blue.



Graphically, to get the right sized bites (more or less), we'll squish the red circle, but this is purely graphical, not conceptual.

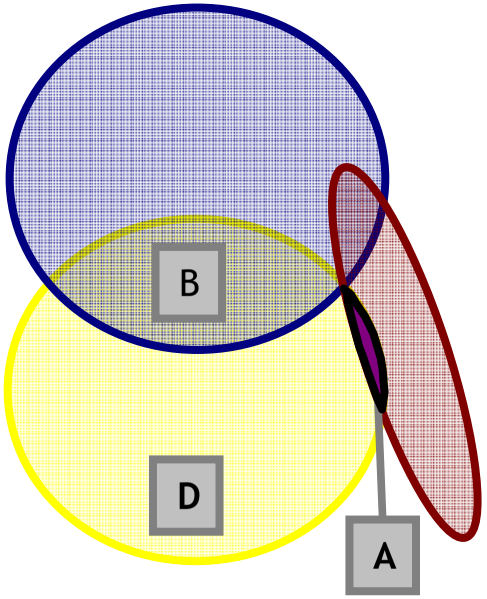
Conceptually, there are different ways to get the right size bites.

Notice that *GENERALKNOWLEDGE* and *SES* have a strong correlation.

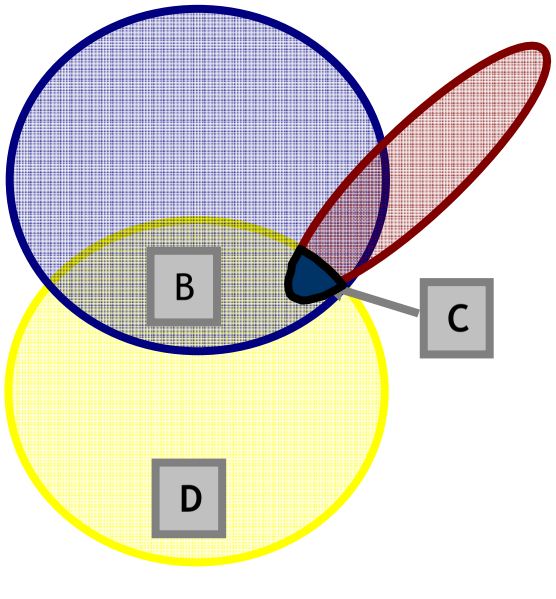
Notice that *GENERALKNOWLEDGE* and *HEADSTARTHOURS* have a weak correlation.

Notice that *HEADSTARTHOURS* and *SES* have a moderate correlation.

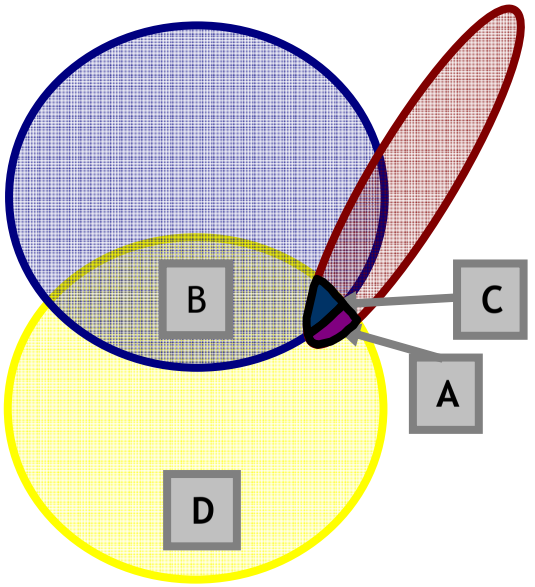
Three Possibilities



HEADSTARTHOURS and *SES* each uniquely predict variation in *GENERALKNOWLEDGE*, but they do not jointly predict variation in *GENERALKNOWLEDGE*.



HEADSTARTHOURS and *SES* jointly predict variation in *GENERALKNOWLEDGE*, but only *SES* uniquely predicts variation in *GENERALKNOWLEDGE*.



HEADSTARTHOURS and *SES* each uniquely predict variation in *GENERALKNOWLEDGE*, but they also jointly predict variation in *GENERALKNOWLEDGE*.

A: Variation in *GENERALKNOWLEDGE* uniquely predicted by *HEADSTARTHOURS*.

B: Variation in *GENERALKNOWLEDGE* uniquely predicted by *SES*.

C: Variation in *GENERALKNOWLEDGE* jointly predicted by *HEADSTARTHOURS* and *SES*.

D: Variation in *GENERALKNOWLEDGE* unpredicted by *HEADSTARTHOURS* and *SES*.

Determining Uniquely Predicted Variation: R^2 Change (I of III)

"Unique" is relative to the other predictors in the model. In other words, uniquely predicted variation is predicted variation unique from the variation predicted by the "control" predictors in the model.

Model 1:

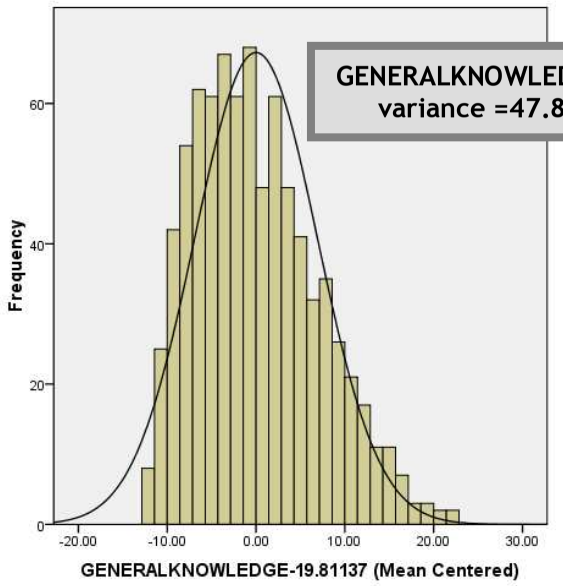
$$GENERALKNOWLEDGE = \beta_0 + \beta_1 SES + \varepsilon$$

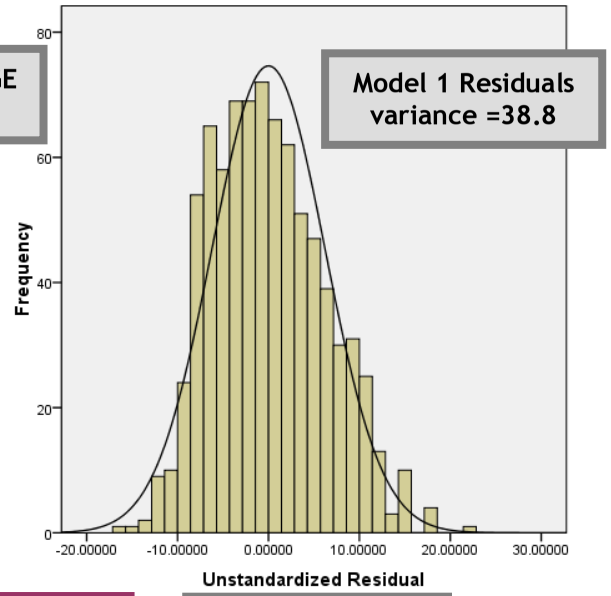
$$R^2 = 1 - \frac{\sigma^2_{\text{Residual}}}{\sigma^2_{\text{Outcome}}}$$

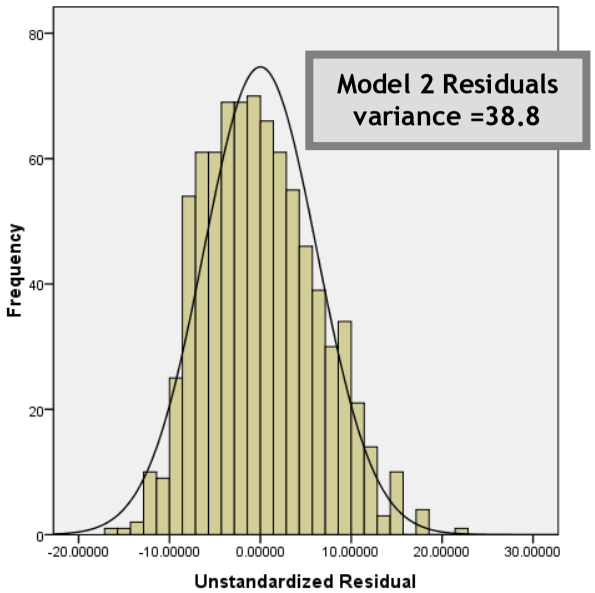
The addition of *HEADSTARTHOURS* to our model does not decrease the residual variance. I.e., it does not tell us anything we did not know with *SES* alone!

Model 2:

$$GENERALKNOWLEDGE = \beta_0 + \beta_1 SES + \beta_2 HEADSTARTHOURS + \varepsilon$$







HEADSTARTHOURS does not uniquely predict variation in *GENERALKNOWLEDGE* over and above the variation predicted by *SES*.

Model 1: R^2 = .19

Model 2: R^2 = .19

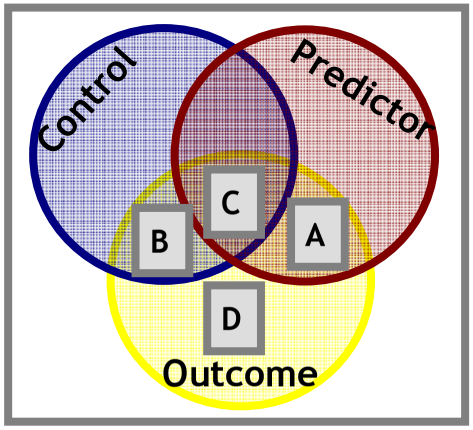
Determining Uniquely Predicted Variation: R^2 Change (II of III)

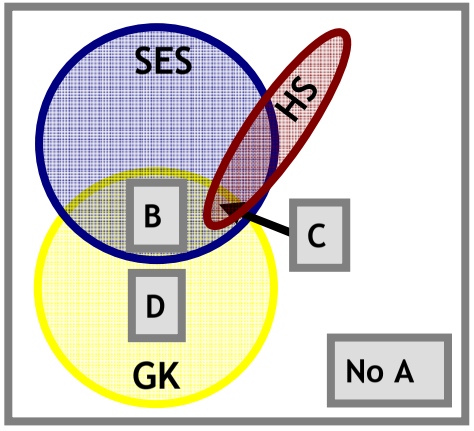
$$\text{Model 1: } R^2 = 1 - \frac{A+D}{A+B+C+D} = \frac{B+C}{A+B+C+D} = .19$$

$$\text{Model 2: } R^2 = 1 - \frac{D}{A+B+C+D} = \frac{A+B+C}{A+B+C+D} = .19$$

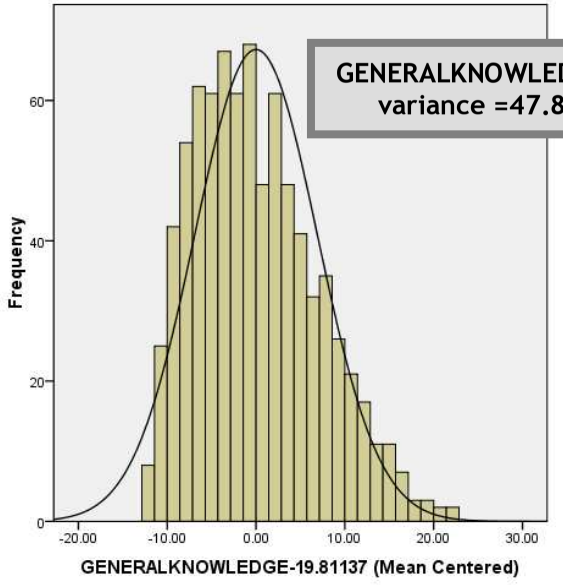
We have been training ourselves to think of variables in terms of "outcomes" and "predictors."

$$\therefore A \approx 0$$

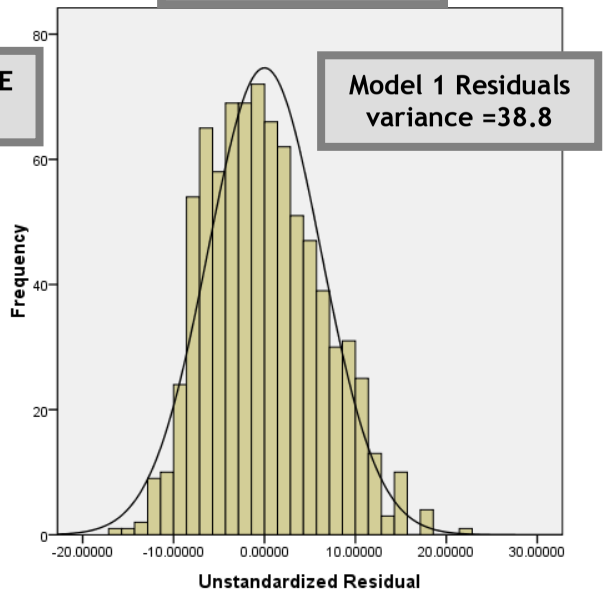




A + B + C + D

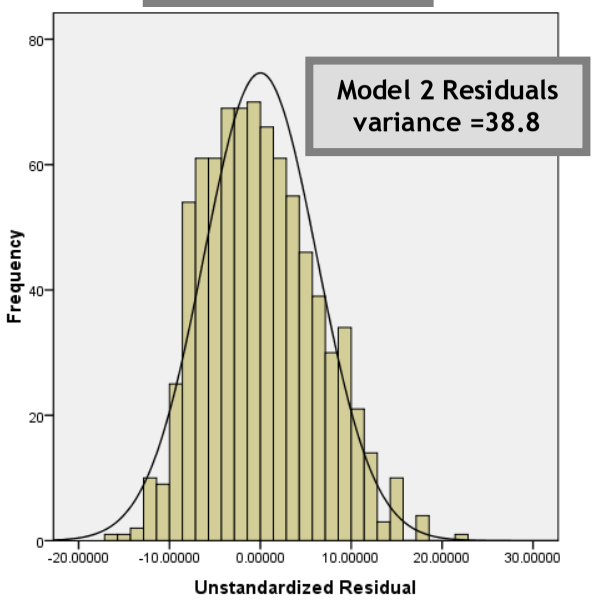


A + D



Model 1: R^2 = .19

D



Model 2: R^2 = .19

Now we are seeing that there are two types of predictors: question predictors ("predictors," for short) and control predictors ("controls," for short).

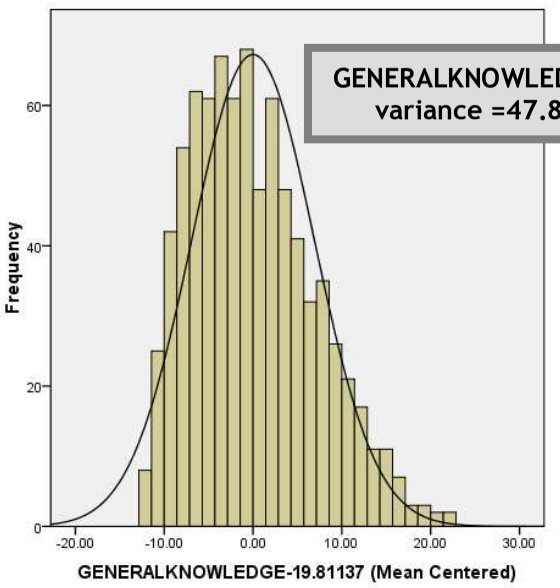
Determining Uniquely Predicted Variation: R^2 Change (III of III)

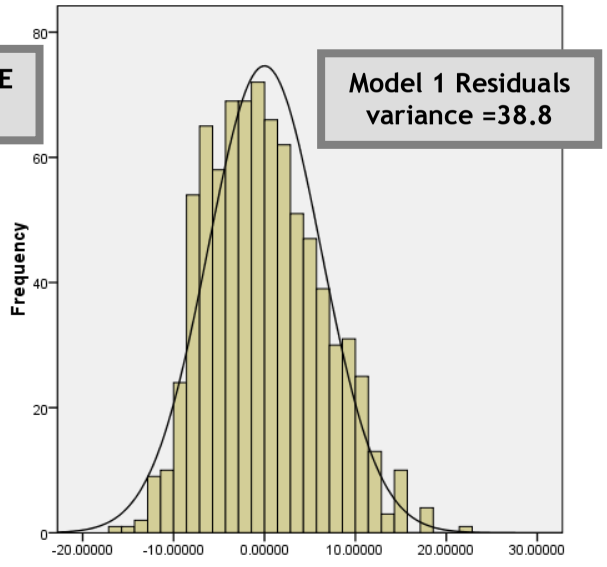
Change in the R^2 statistic is one way to determine uniquely predicted variation. Think of models nested within models. Model 1 is **tightly nested** within Model 2 if Model 2 has not only the same outcome and predictors as Model 1 but also one additional predictor. The additional predictor uniquely predicts variation in the outcome if and only if there is an increase in the R^2 statistic from the tightly nested model to the tightly nesting model; this increase is called the partial R^2 statistic.

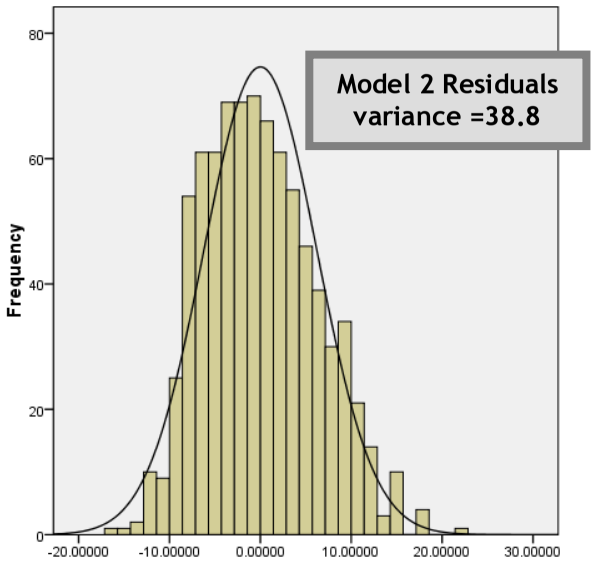
Model 1: $GENERALKNOWLEDGE = \beta_0 + \beta_1 SES + \varepsilon$

Models 1 and 2 form a (small) set of hierarchically nested models.

Model 2: $GENERALKNOWLEDGE = \beta_0 + \beta_1 SES + \beta_2 HEADSTARTHOURS + \varepsilon$







Model 1: R^2 = .19 — INCREASE? → Model 2: R^2 = .19

Determining Uniquely Predicted Variation: Partial Correlation (I of IV)

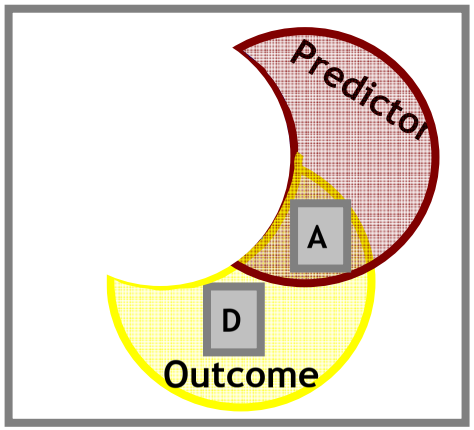
Partial correlation (i.e., the partial *r* statistic) is another way to determine uniquely predicted variation. The partial correlation measures the relationship after we partial out a control variable (or set of control variables). A partial correlation can be greater or less than the simple correlation.

Partial correlations can change signs from their simple correlations!

If we ignore positive/negative signs, we can get a good handle on partial correlations through the R^2 statistic.* Recall that when we square a Pearson correlation (*r*), we get an R^2 statistic. We lose the sign but we get a cool interpretation in terms of proportion of predicted variance. Because we lose the sign, we cannot get back to the Pearson correlation by square rooting the R^2 statistic, but we can get to the absolute value of the Pearson correlation: $|r|$.

The square of the Pearson product moment correlation:

$$r^2 = R^2 = 1 - \frac{\sigma^2_{\text{Simple-Model Residual}}}{\sigma^2_{\text{Outcome}}} = 1 - \frac{B+D}{A+B+C+D} = \frac{A+C}{A+B+C+D}$$



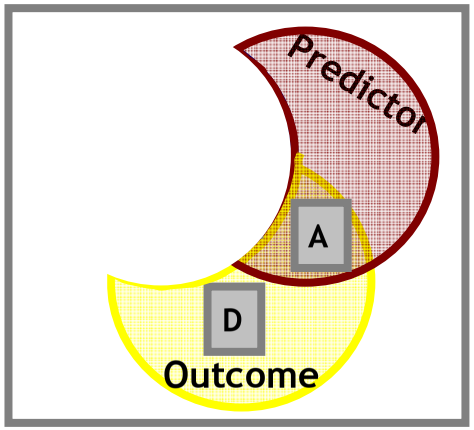
The square of a partial correlation between a predictor and an outcome controlling for one or more variables:

$$(\text{partial } r)^2 = 1 - \frac{\sigma^2_{\text{Control-Model-Plus-Predictor Residual}}}{\sigma^2_{\text{Control-Model Residual}}} = 1 - \frac{D}{A+D} = \frac{A}{A+D}$$

A control model is a model in which all the predictor variables are control predictors.

* The partial R^2 from the previous slides is NOT directly analogous to the partial *r*.

Determining Uniquely Predicted Variation: Partial Correlation (II of IV)

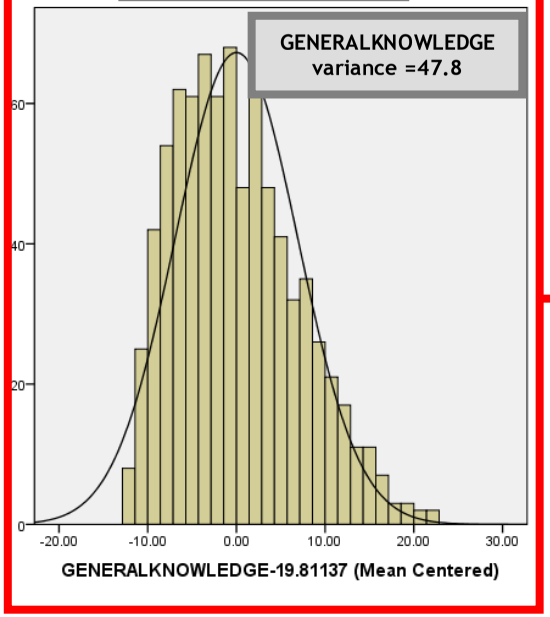


Whereas the R^2 statistic uses the outcome variance as its base, the (partial r)^2 statistic for *HEADSTARTHOURS* uses the residual variance from the control model (e.g., Model 1) as its base.

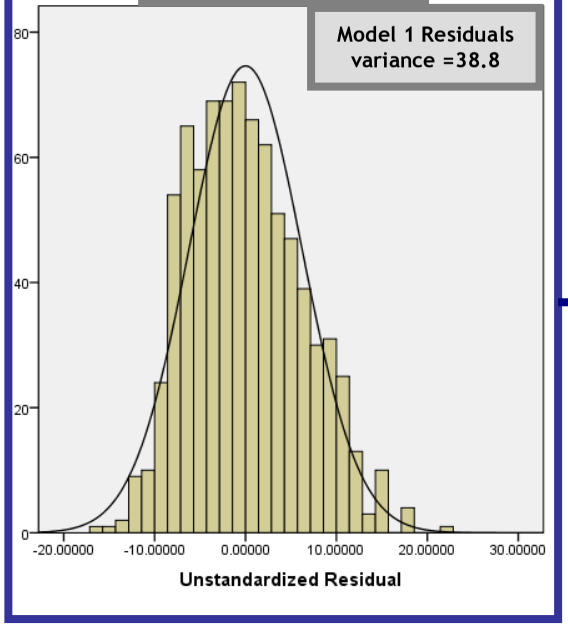
$$R^2 = 1 - \frac{\sigma^2_{\text{Model 2 Residual}}}{\sigma^2_{\text{Outcome}}} = 1 - \frac{D}{A+B+C+D} = \frac{A+B+C}{A+B+C+D} = .19$$

$$(\text{partial } r)^2 = 1 - \frac{\sigma^2_{\text{Model 2 Residual}}}{\sigma^2_{\text{Model 1 Residual}}} = 1 - \frac{D}{A+D} = \frac{A}{A+D} = .00$$

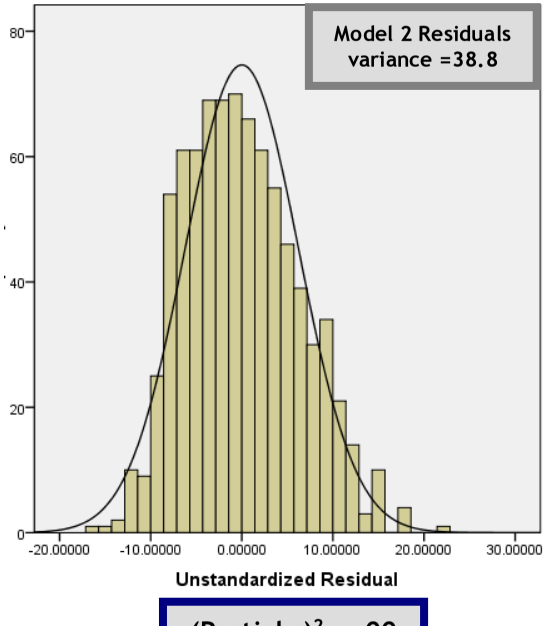
A + B + C + D



A + D



D



The insight here is that we can use residuals as the basis for our calculations.

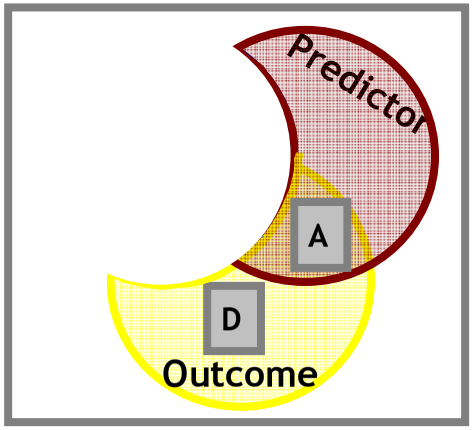
(Partial r)^2 = .00

Model 2: R^2 = .19

Determining Uniquely Predicted Variation: Partial Correlation (III of IV)

Partial correlation (partial *r*) is a correlation between two sets of residuals. Here, we are using residuals as controlled observations (which we have done in previous units to identify subjects who were performing better or worse than expected). One set of residuals comes from a regression of our *outcome variable* on our control variable(s). The other set of residuals comes from a regression of our *predictor variable* on our control variable(s). The correlation between the two sets of residuals (i.e., the partial correlation) tells us not whether the observations are correlated, but rather the partial correlation tells us whether the *controlled* observations are correlated.

Notice in the diagram from Part I of our exposition on partial correlation that, after we partial out the control variable, we have less variation in the outcome variable and the predictor variable (I.e., each full moon becomes a crescent moon). The crescents represent residuals, and where they overlap, the overlap represents their correlation.



Model GK ON SES: $GENERALKNOWLEDGE = \beta_0 + \beta_1 SES + \varepsilon$

Let ε from Model 1 be called *GKONSESERROR* and its z-transformation *ZGKONSESERROR*.

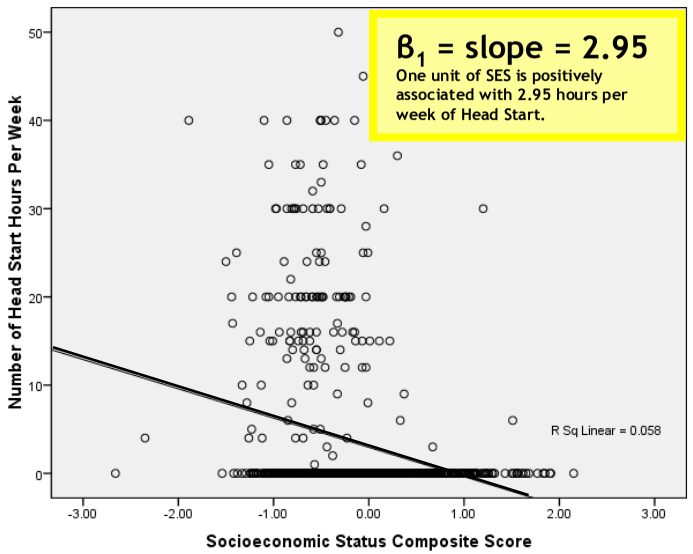
Model HS ON SES: $HEADSTARTHOURS = \beta_0 + \beta_1 SES + \varepsilon$

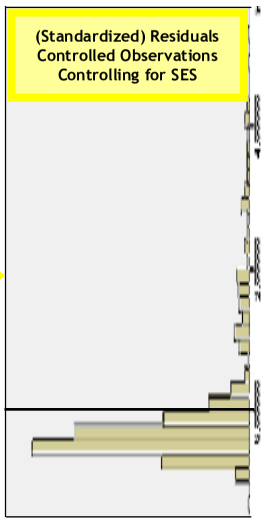
Let ε from Model 2 be called *HSONSESERROR* and its z-transformation *ZHSONSESERROR*.

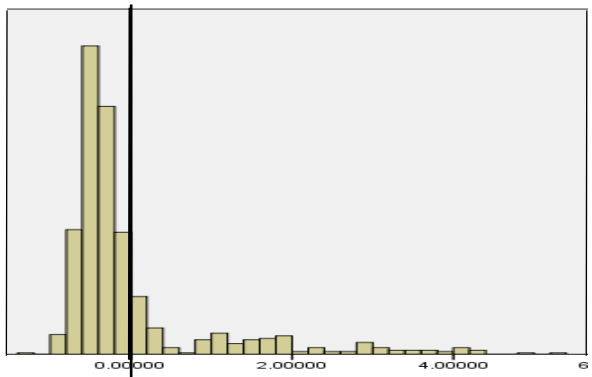
Model Voila: $ZGKONSESERROR = \beta_0 + \beta_1 ZHSONSESERROR + \varepsilon$

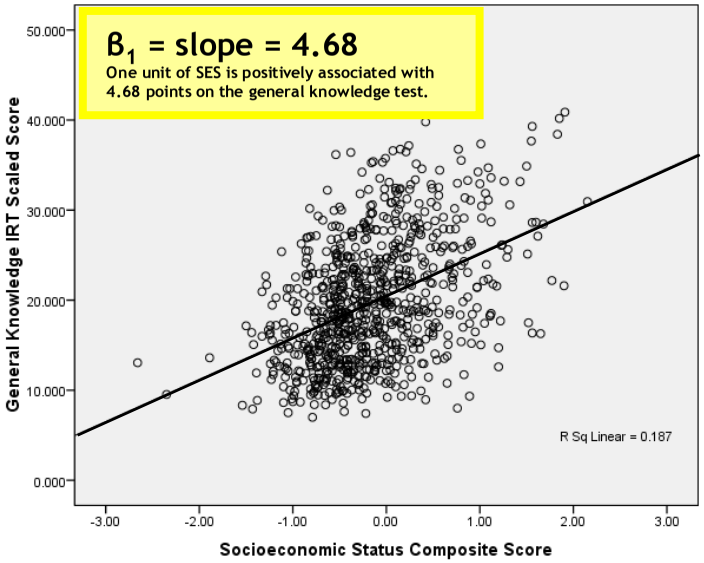
B_1 equals the partial correlation between GENERALKNOWLEDGE and HEADSTARTHOURS, controlling for SES. Recall that when we regress a standardized outcome on a standardized predictor the slope coefficient is the Pearson correlation (*r*).

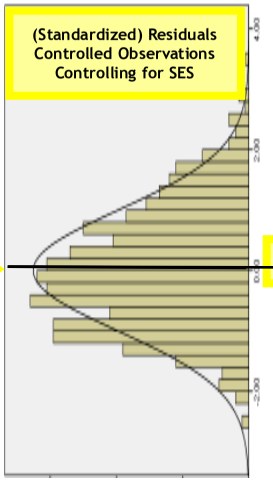
Determining Uniquely Predicted Variation: Partial Correlation (IV of IV)

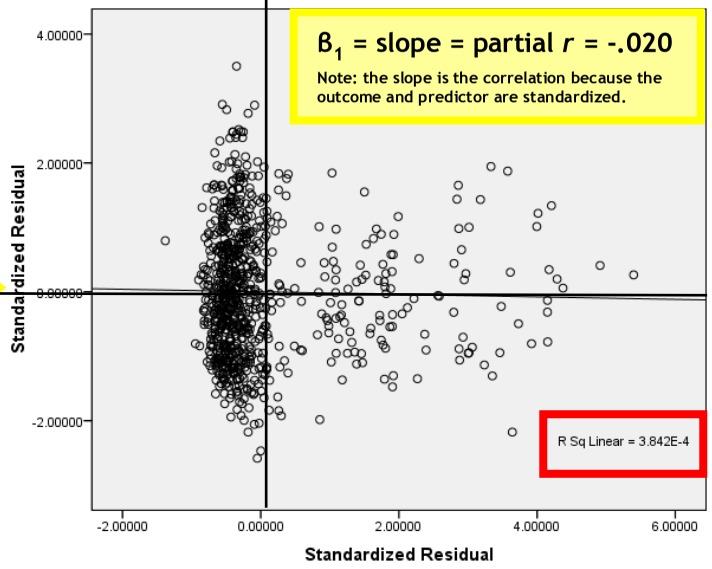












When controlling for SES, hours per week of Head Start has a partial correlation of -.020 with scores on the general knowledge test.

Comparing the Simple Correlation to the Partial Correlation

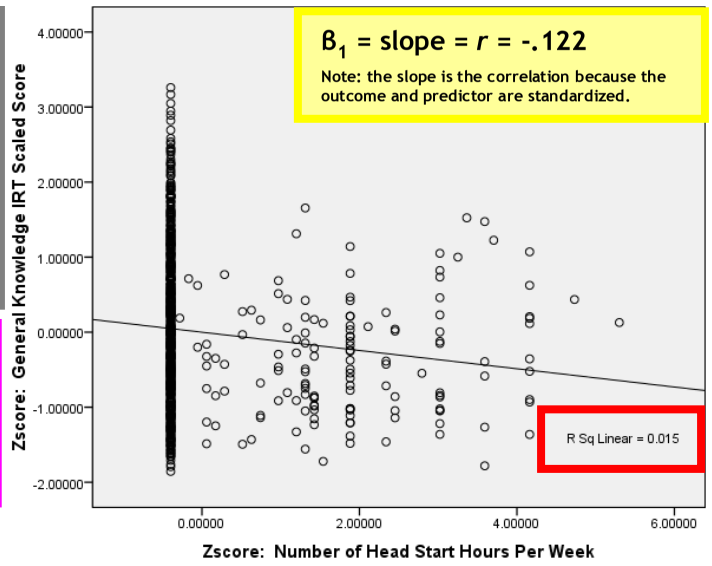
Surprising things can happen upon statistical control. The correlations between residuals (i.e., controlled observations) can be very different from the correlations between variables (i.e., uncontrolled observations).

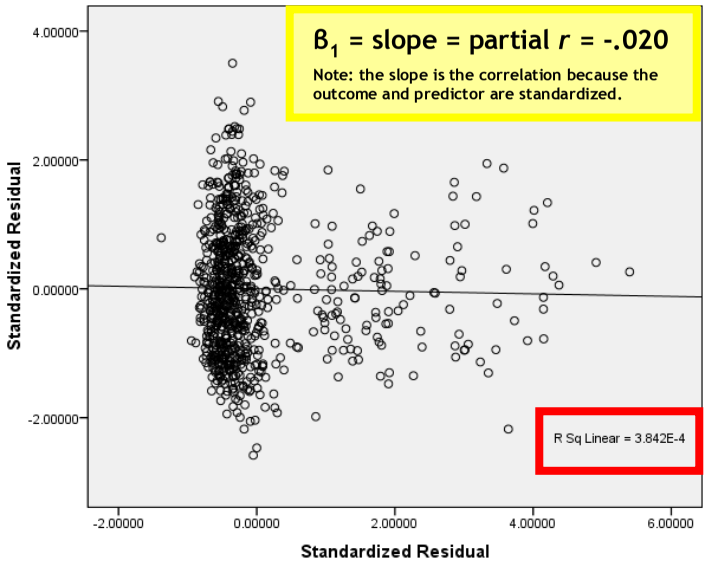
We can make a few observations about the differences between the controlled relationship and the uncontrolled relationship (controlling for *SES*) of *GENERALKNOWLEDGE* and *HEADSTARTHOURS*.

The controlled relationship is weaker with a partial *r* of -.020 from an *r* of -.122. Upon statistical control, the relationship becomes statistically insignificant (p = .576 from p < .001 for the uncontrolled relationship.) This makes substantive sense to me. Head Start is a program for educationally at risk children, with low SES being a primary risk factor. Head Start participants are likely to read worse, and that is precisely why they are Head Start participants. The question is not whether Head Start participants read better or worse than non-participants. Rather, the question is whether they read better than they would if they hadn't participated in Head Start. We need treatment and control groups that are equal (in expectation) to answer that question. In the absence of a control group, we can use statistical control, which is infinitely less valid but often the best we have. A randomized control group controls for all variables observed, unobserved and unobservable, whereas statistical control controls for a few observed variables.

It appears that the normality assumption (and perhaps the linearity assumption) is better met in the controlled relationship.

GLM assumption violations can appear or disappear upon statistical control.





A Partial Correlation Matrix (Partialling Out SES)

Correlations

Control Variables			General Knowledge IRT Scaled Score	Number of Head Start Hours Per Week	Age in Months	English as a 2nd Language
Socioeconomic Status Composite Score	General Knowledge IRT Scaled Score	Correlation	1.000	-.020	.258	-.277
		Significance (2-tailed)	.	.576	.000	.000
		df	0	813	813	813
	Number of Head Start Hours Per Week	Correlation	(partial r)^2 = .00	1.000	.028	.109
		Significance (2-tailed)		.	.424	.002
		df		0	813	813
	Age in Months	Correlation	(partial r)^2 = .07	(partial r)^2 = .00	1.000	-.032
		Significance (2-tailed)			.	.359
		df			0	813
	English as a 2nd Language	Correlation	(partial r)^2 = .08	(partial r)^2 = .01	(partial r)^2 = .00	1.000
		Significance (2-tailed)				.
		df				0

You can see that, as with simple correlation matrices, partial correlation matrices are symmetric about the diagonal, so which variables we consider the outcome or predictor in any given cell is arbitrary.

A Partial Correlation Matrix (Partialling Out SES)

Correlations

Control Variables			General Knowledge IRT Scaled Score	Number of Head Start Hours Per Week	Age in Months	English as a 2nd Language
Socioeconomic Status Composite Score	General Knowledge IRT Scaled Score	Correlation	1.000	-.020	.258	-.277
		Significance (2-tailed)	.	.576	.000	.000
		df	0	813	813	813
	Number of Head Start Hours Per Week	Correlation		1.000	.028	.109
		Significance (2-tailed)		.	.424	.002
		df		0	813	813
	Age in Months	Correlation			1.000	-.032
		Significance (2-tailed)			.	.359
		df			0	813
	English as a 2nd Language	Correlation				1.000
		Significance (2-tailed)				.
		df				0

A Simple/Partial Correlation Matrix

Figure 15.1. A simple/partial correlation matrix in which the top entry in each cell denotes the simple correlation and bottom entry of each cell denotes the partial correlation controlling for *SES* (n = 816).

	GENERAL KNOWLEDGE	*HEADSTART HOURS*	*AGE*	*ESL*
HEADSTARTHOURS	-.122***			
	-.020			
AGE	.247***	.019		
	.258***	.028		
ESL	-.332***	.152***	-.038	
	-.277***	.109**	-.032	
SES	.433***	-.242***	.033	-.201***
	--	--	--	--

Key: * p < .05, ** p < .01, *** p < .001

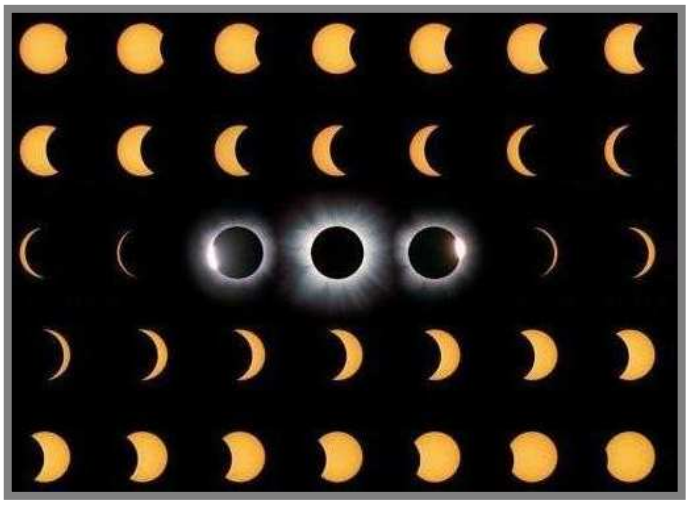
GENERALKNOWLEDGE and *HEADSTARTHOURS* have a weak negative correlation (*r*= -.122, p < .001)) that all but disappears when we control for *SES* (partial *r* = -.020, not statistically significant). The correlations between *GENERALKNOWLEDGE* and *AGE* and between *GENERALKNOWLEDGE* and *ESL* are moderate, and they remain moderate when we partial out *SES*. Of particular interest is *ESL* which not only remains moderately correlated with our outcome *GENERALKNOWLEDGE* upon statistical control of *SES* (as we just mentioned) but also which remains correlated with our question predictor, *HEADSTARTHOURS*. This suggests that if we control for *ESL* in addition to *SES*, the relationship between *GENERALKNOWLEDGE* and *HEADSTART* may differ from the simple and partial (controlling for SES) correlations. On the other hand, *AGE* is not correlated with both *GENERALKNOWLEDGE* and *HEADSTARTHOURS* but only *GENERALKNOWLEDGE*. This suggests that if we control for *AGE* in addition to *SES*, the correlation between *GENERALKNOWLEDGE* and *HEADSTARTHOURS* will increase. (Note: You should be able to nail the first two sentences. For the following sentences, I want you to try your hand at foreshadowing. Use the "Extreme Scenarios" slide as a guide.)

Dig the Post Hole

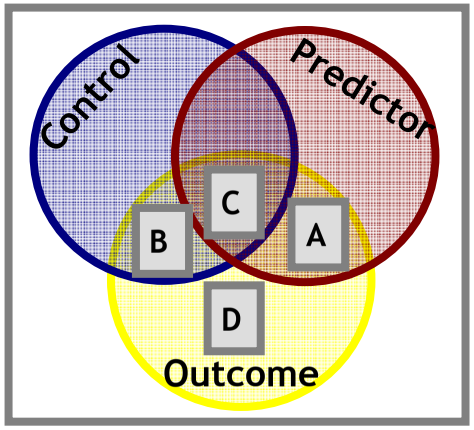
Unit 15 Post Hole:

Interpret a correlation matrix and/or partial correlation matrix and note what they may foreshadow about multiple regression.

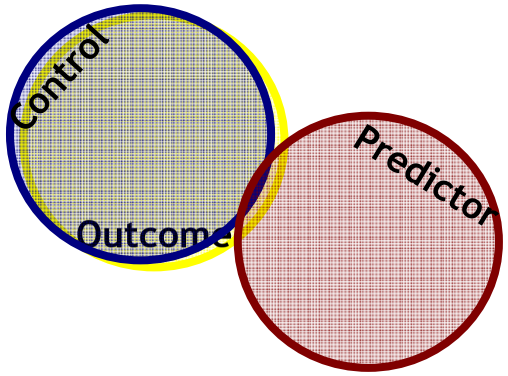
- Interpret the partial correlation matrix in the same way as you would a simple correlation matrix, but be sure to note, "Controlling for…."
- Try your best with the foreshadowing. After a few minutes, take a stab.
  - ➢ Use extreme correlations, high (near ±1) or low (near 0), in conjunction with the necessary consequences from the following "Extreme Scenarios" slide.
  - ➢ When the outcome, predictor and control are all moderately correlated among themselves, anything can happen!



Partial Correlations Can Be Greater/Less Than Their Simple Correlations

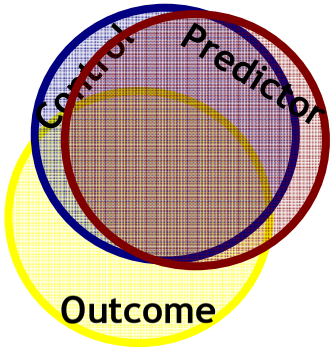


Small Simple Correlation
Large Partial Correlation



$$\frac{A+C}{A+B+C+D} < \frac{A}{A+D}$$

Large Simple Correlation
Small Partial Correlation



$$\frac{A+C}{A+B+C+D} > \frac{A}{A+D}$$

Extreme Scenarios For Conceptual/Foreshadowing Purposes

Outcome Variable: *READING* – a standardized reading score

Predictor Variable: *HOMEWORK* – self-reported hours spent per week on homework

Control Variable: *SES* – a socio-economic status composite score

We are interested in the relationship between READING and HOMEWORK, and we want to look past the universal confound of *SES*.

Extreme Scenario	Consequence for the correlation between READING and HOMEWORK controlling for SES
#1: SES is perfectly correlated with *READING* (r = 1.00).	*partial r* = 0.00 Why? There is no unique variation left in *READING* for *HOMEWORK* to predict.
#2: *SES* is perfectly uncorrelated with *READING* (r = 0.00).	*partial r* = *simple r* Why? Any variation that *HOMEWORK* predicts in *READING* will be unique from the variation that *SES* predicts (because SES does not predict any!).
#3: *HOMEWORK* is perfectly correlated with *READING* (r = 1.00).	*partial r* = 1.00 (Unless #1) Why? *HOMEWORK* predicts all the unique variation in *READING* after *SES* predicts its variation.
#4: *HOMEWORK* is perfectly uncorrelated with *READING* (r = 0.00).	*partial r* = 0.00 Why? *HOMEWORK* predicts no variation at all, so it cannot predict any unique variation.
#5: *HOMEWORK* is perfectly correlated with *SES* (r = 1.00).	*partial r* = 0.00 Why? *HOMEWORK* predicts the same variation as *SES*, so it cannot predict any unique variation.
#6: *HOMEWORK* is perfectly uncorrelated with *SES* (r = 0.00).	*partial r* ≥ *simple r* Why? Any variation that *HOMEWORK* predicts in *READING* will be unique from the variation that *SES* predicts in *READING*, but *SES* will decrease the variation in need of predicting insofar as it is correlated with *READING*.

Answering our Roadmap Question

Unit 15: What are the correlations among reading, ESL, and homework, controlling for SES?

Correlations

		READING	NUMBER OF HRS SPENT ON HOMEWORK PER WEEK	ESL	FREELUNCH
READING	Pearson Correlation	1.000	.183**	-.053**	-.267**
	Sig. (2-tailed)		.000	.000	.000
	N	7800.000	7800	7800	7800
NUMBER OF HRS SPENT ON HOMEWORK PER WEEK	Pearson Correlation	.183**	1.000	.005	-.092**
	Sig. (2-tailed)	.000		.648	.000
	N	7800	7800.000	7800	7800
ESL	Pearson Correlation	-.053**	.005	1.000	.093**
	Sig. (2-tailed)	.000	.648		.000
	N	7800	7800	7800.000	7800
FREELUNCH	Pearson Correlation	-.267**	-.092**	.093**	1.000
	Sig. (2-tailed)	.000	.000	.000	
	N	7800	7800	7800	7800.000

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

First, let's speculate based on this simple correlation matrix and our substantive knowledge or hunches (or prejudices?). We know that free lunch eligibility, our proxy for low SES, is negatively correlated with reading scores. We see that homework hours is correlated with reading scores, but we have to wonder:

Is SES a confounding third variable in the correlation between homework and reading? Perhaps the homework/reading correlation is just the SES/reading correlation in disguise? We must wonder this insofar as homework and SES are correlated. In fact, SES and reading are not highly correlated (r = -0.09). Nevertheless, we still have to wonder how much of the homework/reading correlation is uniquely predicted (aside from the SES/reading correlation). It is possible, perhaps likely, that at least some of the same variation in reading scores is jointly predicted by both homework and SES.

Answering our Roadmap Question

Unit 15: What are the correlations among reading, ESL, and homework, controlling for SES?

Correlations

Control Variables			READING	NUMBER OF HRS SPENT ON HOMEWORK PER WEEK	ESL
FREELUNCH	READING	Correlation	1.000	.165	-.029
		Significance (2-tailed)	.	.000	.009
		df	0	7797	7797
	NUMBER OF HRS SPENT ON HOMEWORK PER WEEK	Correlation	.165	1.000	.014
		Significance (2-tailed)	.000	.	.222
		df	7797	0	7797
	ESL	Correlation	-.029	.014	1.000
		Significance (2-tailed)	.009	.222	.
		df	7797	7797	0

Controlling for free lunch eligibility, there remains a a positive correlation between hours spent on homework per week and reading scores (*partial r* = .165, $p < .001$). Thus, homework predicts unique variation in reading scores over and above the variation predicted by free lunch eligibility. We may consider further controlling for ESL status, but its correlations with both homework and reading scores are so low that it will probably not inform the relationship between homework and reading scores.

Answering our Roadmap Question

Unit 15: What are the correlations among reading, ESL, and homework, controlling for SES?

Figure 15.2. A simple/partial correlation matrix in which the top entry in each cell denotes the simple correlation and bottom entry of each cell denotes the partial correlation controlling for free lunch eligibility (n = 816).

	READING	*HOMEWORK*	*ESL*
HOMEWORK	.183***		
	.165***		
ESL	-.053***	.005	
	-.029***	.014	
FREELUNCH	-.267***	-.092***	.093***
	--	--	--

Key: * p < .05, ** p < .01, *** p < .001

Notice that the partial correlations for READING/HOMEWORK and READING/ESL are less than their simple correlations, but the partial correlation for HOMEWORK/ESL is greater than its simple correlation. Substantively, the differences seem trivial, but, pedagogically, this is a good illustration of the possibilities. Sometimes, the direction of correlation can switch upon statistical control. We will see why in Unit 16.

Unit 15 Appendix: Key Concepts

Why Residuals? Unaccounted Variables, Measurement Error, Individual Variation

"Unique" is relative to the other predictors in the model. In other words, uniquely predicted variation is predicted variation unique from the variation predicted by the "control" predictors in the model.

Partial correlations can change signs from their simple correlations!

* The partial R2 from the previous slides is NOT directly analogous to the partial *r*.

Surprising things can happen upon statistical control. The correlations between residuals (i.e., controlled observations) can be very different from the correlations between variables (i.e., uncontrolled observations).

GLM assumption violations can appear or disappear upon statistical control.

Unit 15 Appendix: Key Interpretations

HEADSTARTHOURS and *SES* each uniquely predict variation in *GENERALKNOWLEDGE*, but they do not jointly predict variation in *GENERALKNOWLEDGE*.

HEADSTARTHOURS and *SES* jointly predict variation in *GENERALKNOWLEDGE*, but only *SES* uniquely predicts variation in *GENERALKNOWLEDGE*.

HEADSTARTHOURS and *SES* each uniquely predict variation in *GENERALKNOWLEDGE*, but they also jointly predict variation in *GENERALKNOWLEDGE*.

HEADSTARTHOURS does not uniquely predict variation in *GENERALKNOWLEDGE* over and above the variation predicted by *SES*.

When controlling for SES, hours per week of Head Start has a partial correlation of -.020 with scores on the general knowledge test.

GENERALKNOWLEDGE and *HEADSTARTHOURS* have a weak negative correlation (r= -.122, p < .001)) that all but disappears when we control for *SES* (partial r = -.020, not statistically significant). The correlations between *GENERALKNOWLEDGE* and *AGE* and between *GENERALKNOWLEDGE* and *ESL* are moderate, and they remain moderate when we partial out *SES*. Of particular interest is *ESL* which not only remains moderately correlated with our outcome *GENERALKNOWLEDGE* upon statistical control of *SES* (as we just mentioned) but also which remains correlated with our question predictor, *HEADSTARTHOURS*. This suggests that if we control for *ESL* in addition to *SES*, the relationship between *GENERALKNOWLEDGE* and *HEADSTART* may differ from the simple and partial (controlling for SES) correlations. On the other hand, *AGE* is not correlated with both *GENERALKNOWLEDGE* and *HEADSTARTHOURS* but only *GENERALKNOWLEDGE*. This suggests that if we control for *AGE* in addition to *SES*, the correlation between *GENERALKNOWLEDGE* and *HEADSTARTHOURS* will increase. (Note: You should be able to nail the first two sentences. For the following sentences, I want you to try your hand at foreshadowing. Use the "Extreme Scenarios" slide as a guide.)

Controlling for free lunch eligibility, there remains a a positive correlation between hours spent on homework per week and reading scores (*partial* r = .165, p < .001). Thus, homework predicts unique variation in reading scores over and above the variation predicted by free lunch eligibility. We may consider further controlling for ESL status, but its correlations with both homework and reading scores are so low that it will probably not inform the relationship between homework and reading scores.

Unit 15 Appendix: Key Terminology

Variance is just a hard working number trying, trying, trying to summarize the variation of a univariate distribution. It is one of many statistical summaries of variation, including range, midspread and standard deviation. Variance is the average squared deviation from the mean.

The mean square residual (or error) represents the variance in the outcome that is left over after we fit our model. It is an average. Every observation has a residual. We can square that residual. The mean square residual is just the average squared residual.

We have been training ourselves to think of variables in terms of "outcomes" and "predictors." Now we are seeing that there are two types of predictors: question predictors ("predictors," for short) and control predictors ("controls," for short).

Change in the R2 statistic is one way to determine uniquely predicted variation. Think of models nested within models. Model 1 is tightly nested within Model 2 if Model 2 has not only the same outcome and predictors as Model 1 but also one additional predictor. The additional predictor uniquely predicts variation in the outcome if and only if there is an increase in the R2 statistic from the tightly nested model to the tightly nesting model; this increase is called the partial R2 statistic.

Partial correlation (i.e., the partial *r* statistic) is another way to determine uniquely predicted variation. The partial correlation measures the relationship after we partial out a control variable (or set of control variables). A partial correlation can be greater or less than the simple correlation.

A control model is a model in which all the predictor variables are control predictors.

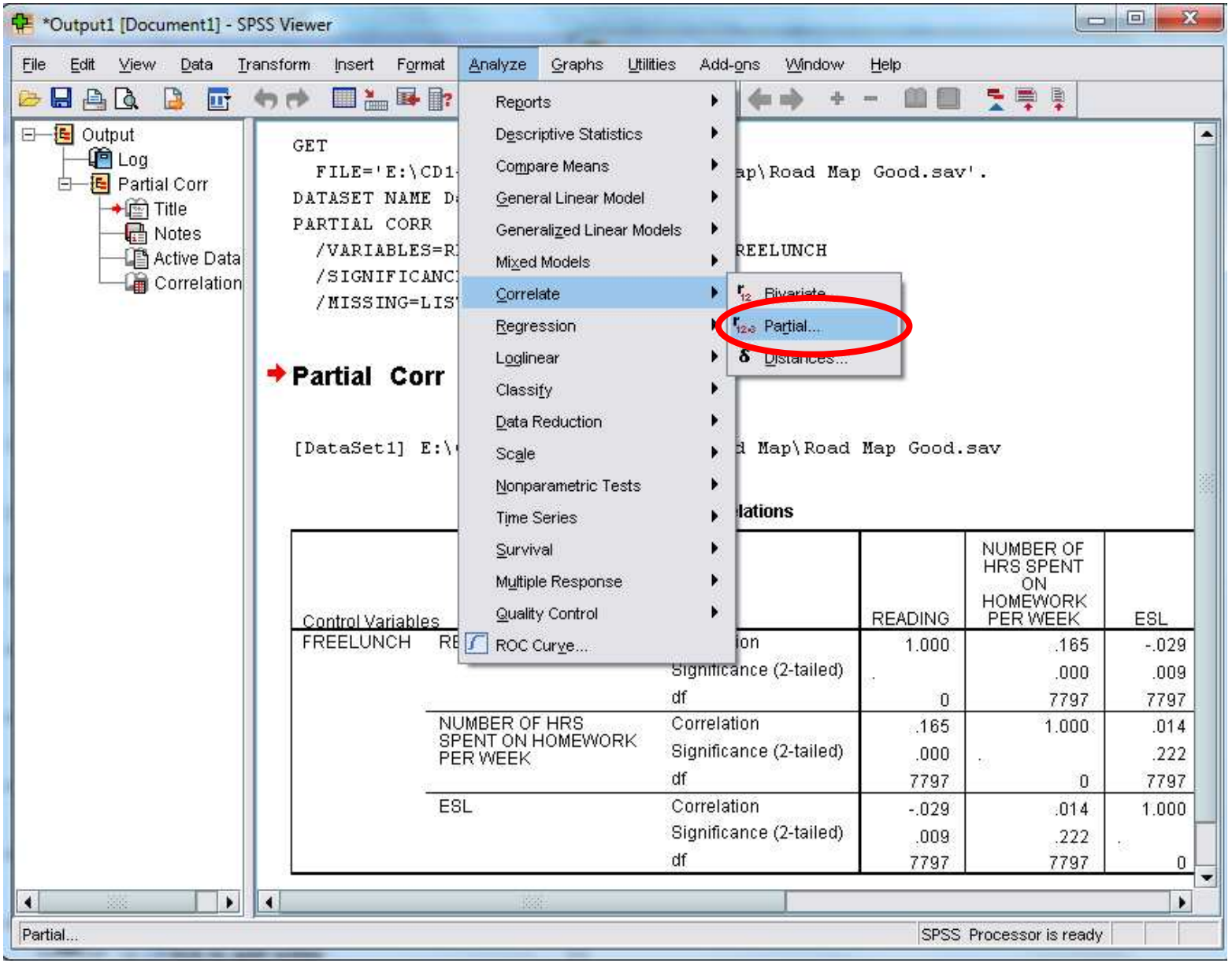
Partial correlation (partial *r*) is a correlation between two sets of residuals. Here, we are using residuals as controlled observations (which we have done in previous units to identify subjects who were performing better or worse than expected). One set of residuals comes from a regression of our *outcome variable* on our control variable(s). The other set of residuals comes from a regression of our *predictor variable* on our control variable(s). The correlation between the two sets of residuals (i.e., the partial correlation) tells us not whether the observations are correlated, but rather the partial correlation tells us whether the *controlled* observations are correlated.

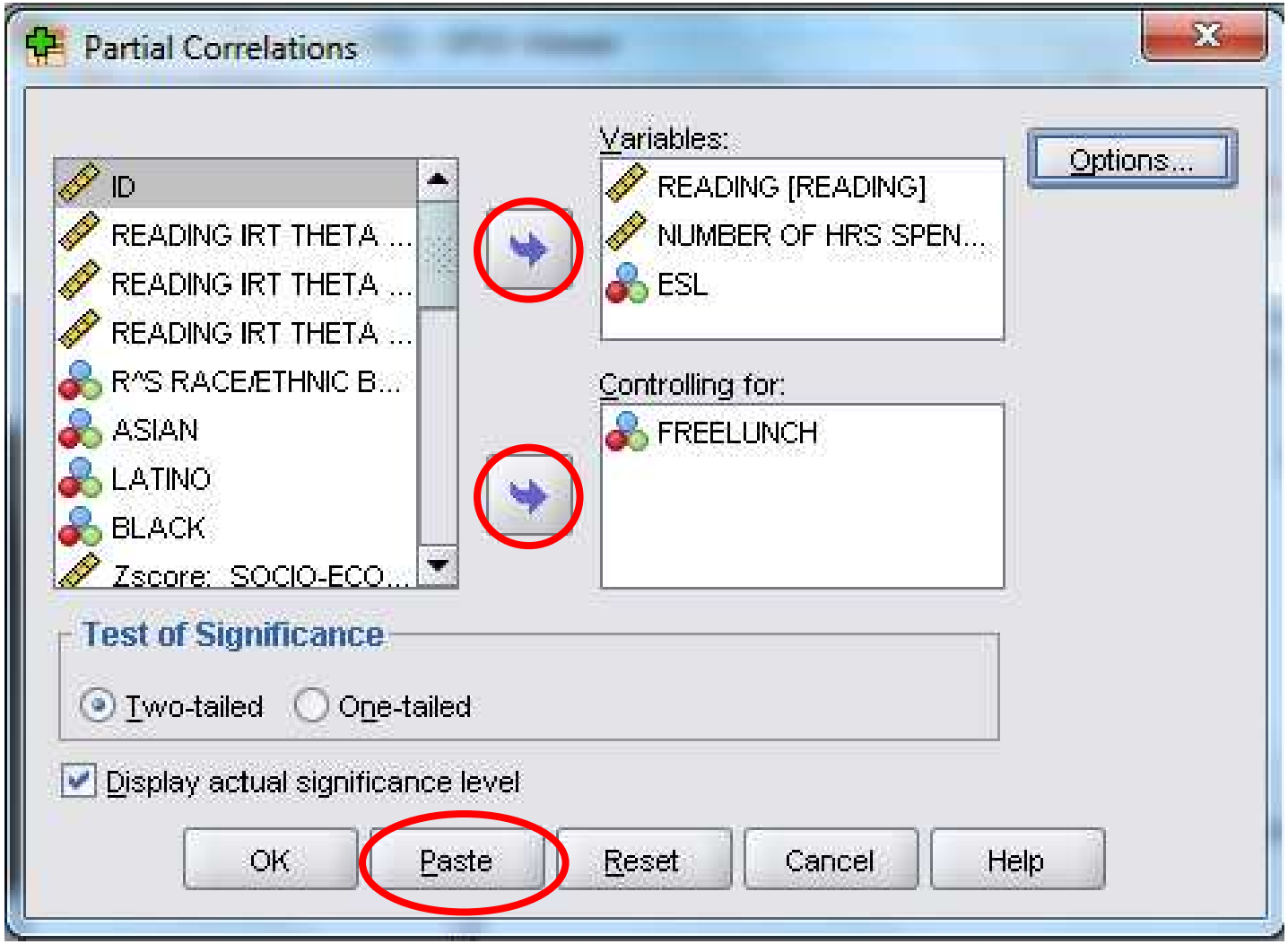
Unit 15 Appendix: Formulas

Unit 15 Appendix: SPSS Syntax

```
PARTIAL CORR
  /VARIABLES=READING HOMEWORK ESL BY FREELUNCH
  /SIGNIFICANCE=TWOTAIL
  /MISSING=LISTWISE.
```

Analyze > Correlate > Partial…




Partial Correlations
ID
READING IRT THETA ...
READING IRT THETA ...
READING IRT THETA ...
R^S RACE/ETHNIC B...
ASIAN
LATINO
BLACK
Zscore: SOCIO-ECO...
Variables:
READING [READING]
NUMBER OF HRS SPEN...
ESL
Options...
Controlling for:
FREELUNCH
Test of Significance
Two-tailed
One-tailed
Display actual significance level
OK
Paste
Reset
Cancel
Help

Perceived Intimacy of Adolescent Girls (Intimacy.sav)

- **Overview: Dataset contains self-ratings of the intimacy that adolescent girls perceive themselves as having with: (a) their mother and (b) their boyfriend.**

- **Source: HGSE thesis by Dr. Linda Kilner entitled Intimacy in Female Adolescent's Relationships with Parents and Friends (1991). Kilner collected the ratings using the Adolescent Intimacy Scale.**
- **Sample: 64 adolescent girls in the sophomore, junior and senior classes of a local suburban public school system.**
- **Variables:**

Self Disclosure to Mother (M_Seldis)
Trusts Mother (M_Trust)
Mutual Caring with Mother (M_Care)
Risk Vulnerability with Mother (M_Vuln)
Physical Affection with Mother (M_Phys)
Resolves Conflicts with Mother (M_Cres)

Self Disclosure to Boyfriend (B_Seldis)
Trusts Boyfriend (B_Trust)
Mutual Caring with Boyfriend (B_Care)
Risk Vulnerability with Boyfriend (B_Vuln)
Physical Affection with Boyfriend (B_Phys)
Resolves Conflicts with Boyfriend (B_Cres)

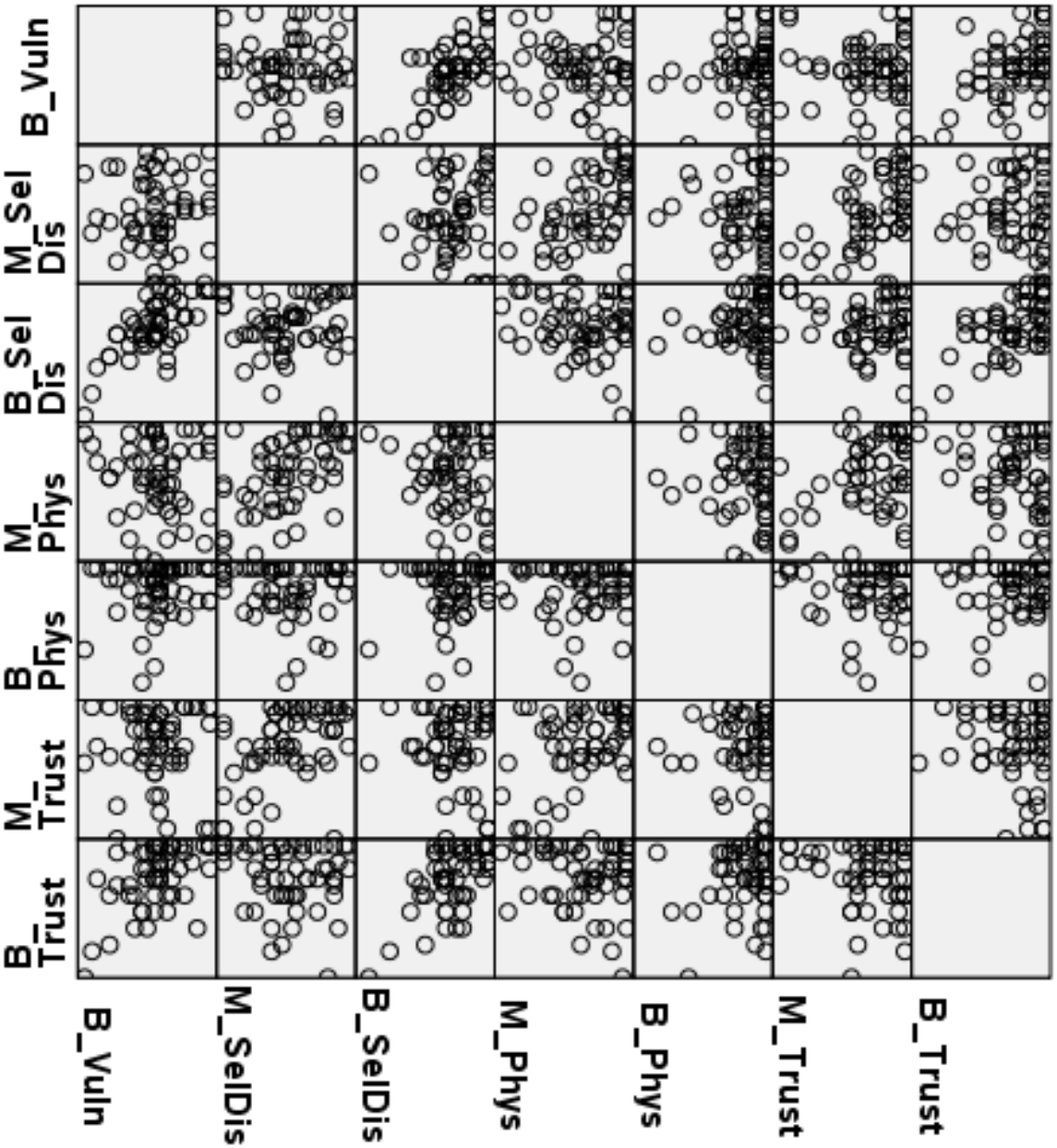
Perceived Intimacy of Adolescent Girls (Intimacy.sav)

Correlations

		Risk vulnerability w boyfriend	Self-disclose to mother	Self-disclose to boyfriend	Phys affection w mother	Phys affection w boyfriend	Trust mother	Trust boyfriend
Risk vulnerability w boyfriend	Pearson Correlation	1.000	.002	.731**	-.053	.094	.052	.553**
	Sig. (2-tailed)		.985	.000	.689	.476	.690	.000
	N	61.000	60	60	60	60	61	61
Self-disclose to mother	Pearson Correlation	.002	1.000	-.019	.539**	-.068	.483**	-.132
	Sig. (2-tailed)	.985		.888	.000	.606	.000	.309
	N	60	63.000	60	62	59	63	61
Self-disclose to boyfriend	Pearson Correlation	.731**	-.019	1.000	-.086	.162	-.076	.607**
	Sig. (2-tailed)	.000	.888		.512	.221	.562	.000
	N	60	60	61.000	60	59	61	61
Phys affection w mother	Pearson Correlation	-.053	.539**	-.086	1.000	.029	.422**	-.135
	Sig. (2-tailed)	.689	.000	.512		.827	.001	.299
	N	60	62	60	63.000	59	63	61
Phys affection w boyfriend	Pearson Correlation	.094	-.068	.162	.029	1.000	.027	.143
	Sig. (2-tailed)	.476	.606	.221	.827		.839	.277
	N	60	59	59	59	60.000	60	60
Trust mother	Pearson Correlation	.052	.483**	-.076	.422**	.027	1.000	-.126
	Sig. (2-tailed)	.690	.000	.562	.001	.839		.330
	N	61	63	61	63	60	64.000	62
Trust boyfriend	Pearson Correlation	.553**	-.132	.607**	-.135	.143	-.126	1.000
	Sig. (2-tailed)	.000	.309	.000	.299	.277	.330	
	N	61	61	61	61	60	62	62.000

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

Perceived Intimacy of Adolescent Girls (Intimacy.sav)

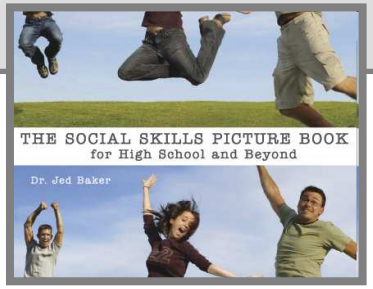


Perceived Intimacy of Adolescent Girls (Intimacy.sav)

Correlations

Control Variables			Risk vulnerability w boyfriend	Self-disclose to mother	Self-disclose to boyfriend	Phys affection w mother	Phys affection w boyfriend	Trust mother
Trust boyfriend	Risk vulnerability w boyfriend	Correlation	1.000	.073	.612	.050	.036	.149
		Significance (2-tailed)	.	.591	.000	.716	.794	.273
		df	0	54	54	54	54	54
	Self-disclose to mother	Correlation	.073	1.000	.111	.553	-.030	.432
		Significance (2-tailed)	.591	.	.414	.000	.827	.001
		df	54	0	54	54	54	54
	Self-disclose to boyfriend	Correlation	.612	.111	1.000	.022	.077	.038
		Significance (2-tailed)	.000	.414	.	.872	.574	.782
		df	54	54	0	54	54	54
	Phys affection w mother	Correlation	.050	.553	.022	1.000	.057	.422
		Significance (2-tailed)	.716	.000	.872	.	.676	.001
		df	54	54	54	0	54	54
	Phys affection w boyfriend	Correlation	.036	-.030	.077	.057	1.000	.061
		Significance (2-tailed)	.794	.827	.574	.676	.	.655
		df	54	54	54	54	0	54
	Trust mother	Correlation	.149	.432	.038	.422	.061	1.000
		Significance (2-tailed)	.273	.001	.782	.001	.655	.
		df	54	54	54	54	54	0

High School and Beyond (HSB.sav)



- **Overview: High School & Beyond – Subset of data focused on selected student and school characteristics as predictors of academic achievement.**
- **Source: Subset of data graciously provided by Valerie Lee, University of Michigan.**
- **Sample: This subsample has 1044 students in 205 schools. Missing data on the outcome test score and family SES were eliminated. In addition, schools with fewer than 3 students included in this subset of data were excluded.**
- **Variables:**

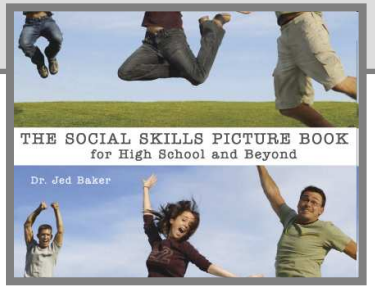
Variables about the student—

(Black) 1=Black, 0=Other
(Latin) 1=Latino/a, 0=Other
(Sex) 1=Female, 0=Male
(BYSES) Base year SES
(GPA80) HS GPA in 1980
(GPS82) HS GPA in 1982
(BYTest) Base year composite of reading and math tests
(BBConc) Base year self concept
(FEConc) First Follow-up self concept

Variables about the student's school—

(PctMin) % HS that is minority students Percentage
(HSSize) HS Size
(PctDrop) % dropouts in HS Percentage
(BYSES_S) Average SES in HS sample
(GPA80_S) Average GPA80 in HS sample
(GPA82_S) Average GPA82 in HS sample
(BYTest_S) Average test score in HS sample
(BBConc_S) Average base year self concept in HS sample
(FEConc_S) Average follow-up self concept in HS sample

High School and Beyond (HSB.sav)



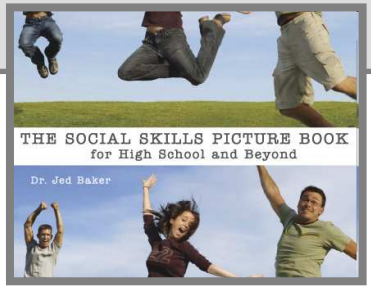
Correlations

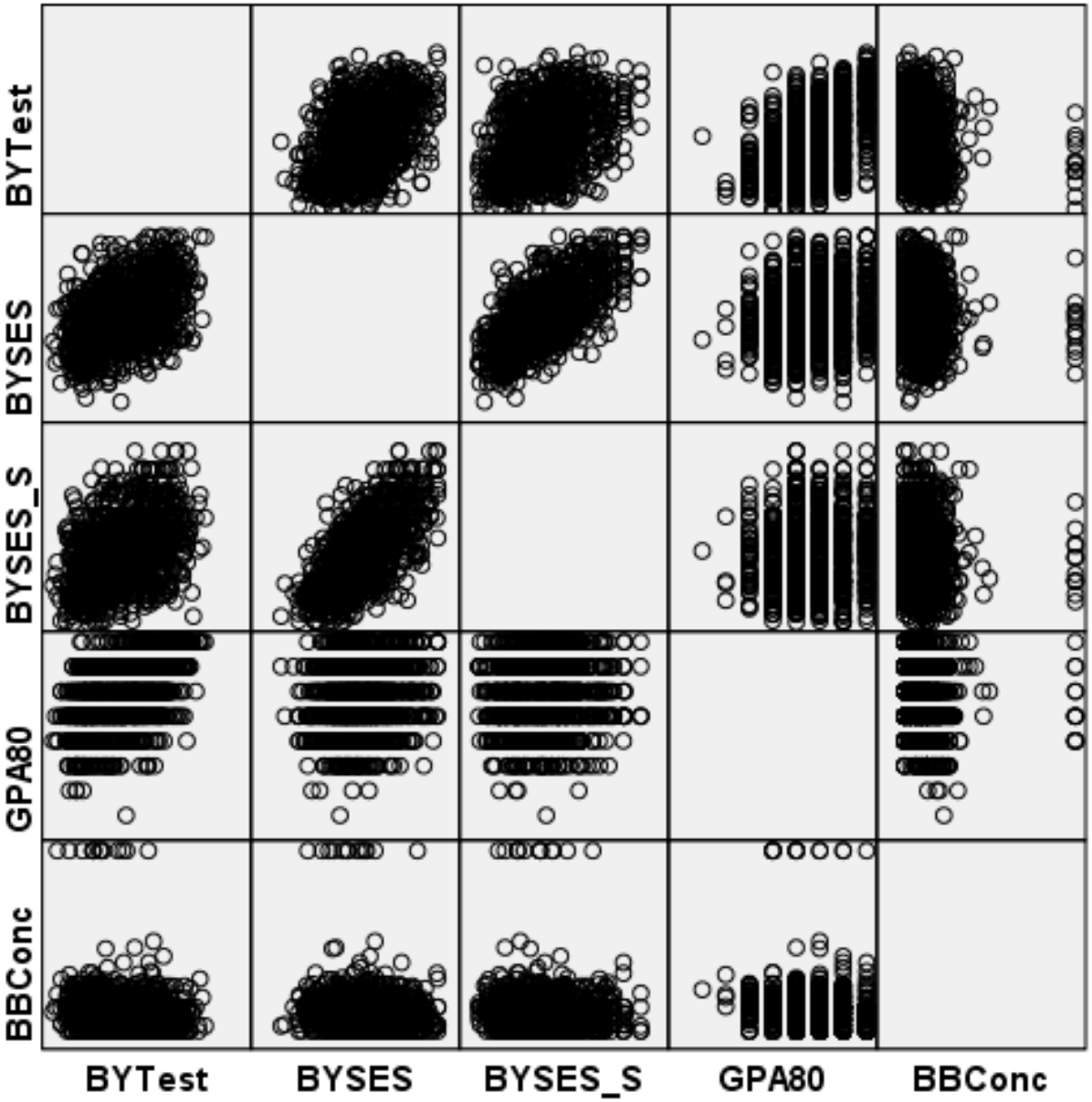
		Base Year Composite Test	Base Year SES	BY SES, School Avg	GPA 1980	Base Year Self Concept	1 = Black, 0 = Other	1 = Latino/a, 0 = Other	1 = Female, 0 = Other
Base Year Composite Test	Pearson Correlation	1.000	.440**	.429**	.508**	-.110**	-.303**	-.157**	-.158**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000
	N	1044.000	1044	1044	1039	1044	1044	1044	1044
Base Year SES	Pearson Correlation	.440**	1.000	.674**	.180**	-.053	-.227**	-.190**	-.085**
	Sig. (2-tailed)	.000		.000	.000	.086	.000	.000	.006
	N	1044	1044.000	1044	1039	1044	1044	1044	1044
BY SES, School Avg	Pearson Correlation	.429**	.674**	1.000	.099**	-.034	-.223**	-.190**	-.064*
	Sig. (2-tailed)	.000	.000		.001	.270	.000	.000	.038
	N	1044	1044	1044.000	1039	1044	1044	1044	1044
GPA 1980	Pearson Correlation	.508**	.180**	.099**	1.000	-.096**	-.179**	-.116**	.075*
	Sig. (2-tailed)	.000	.000	.001		.002	.000	.000	.015
	N	1039	1039	1039	1039.000	1039	1039	1039	1039
Base Year Self Concept	Pearson Correlation	-.110**	-.053	-.034	-.096**	1.000	.033	-.018	.010
	Sig. (2-tailed)	.000	.086	.270	.002		.291	.569	.742
	N	1044	1044	1044	1039	1044.000	1044	1044	1044
1 = Black, 0 = Other	Pearson Correlation	-.303**	-.227**	-.223**	-.179**	.033	1.000	-.413**	.086**
	Sig. (2-tailed)	.000	.000	.000	.000	.291		.000	.005
	N	1044	1044	1044	1039	1044	1044.000	1044	1044
1 = Latino/a, 0 = Other	Pearson Correlation	-.157**	-.190**	-.190**	-.116**	-.018	-.413**	1.000	-.048
	Sig. (2-tailed)	.000	.000	.000	.000	.569	.000		.118
	N	1044	1044	1044	1039	1044	1044	1044.000	1044
1 = Female, 0 = Other	Pearson Correlation	-.158**	-.085**	-.064*	.075*	.010	.086**	-.048	1.000
	Sig. (2-tailed)	.000	.006	.038	.015	.742	.005	.118	
	N	1044	1044	1044	1039	1044	1044	1044	1044.000

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

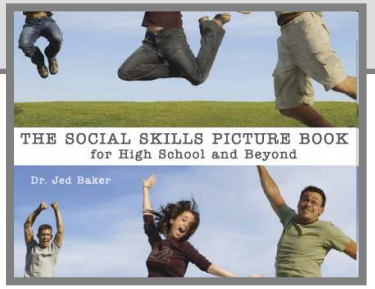
*. Correlation is significant at the 0.05 level (2-tailed).

High School and Beyond (HSB.sav)





High School and Beyond (HSB.sav)



Correlations

Control Variables			Base Year Composite Test	Base Year SES	GPA 1980	Base Year Self Concept	1 = Black, 0 = Other	1 = Latino/a, 0 = Other	1 = Female, 0 = Other
BY SES, School Avg	Base Year Composite Test	Correlation	1.000	.221	.519	-.095	-.232	-.089	-.146
		Significance (2-tailed)	.	.000	.000	.002	.000	.004	.000
		df	0	1036	1036	1036	1036	1036	1036
	Base Year SES	Correlation	.221	1.000	.154	-.040	-.102	-.090	-.063
		Significance (2-tailed)	.000	.	.000	.202	.001	.004	.044
		df	1036	0	1036	1036	1036	1036	1036
	GPA 1980	Correlation	.519	.154	1.000	-.092	-.162	-.099	.082
		Significance (2-tailed)	.000	.000	.	.003	.000	.001	.008
		df	1036	1036	0	1036	1036	1036	1036
	Base Year Self Concept	Correlation	-.095	-.040	-.092	1.000	.015	-.023	.003
		Significance (2-tailed)	.002	.202	.003	.	.631	.463	.932
		df	1036	1036	1036	0	1036	1036	1036
	1 = Black, 0 = Other	Correlation	-.232	-.102	-.162	.015	1.000	-.475	.073
		Significance (2-tailed)	.000	.001	.000	.631	.	.000	.019
		df	1036	1036	1036	1036	0	1036	1036
	1 = Latino/a, 0 = Other	Correlation	-.089	-.090	-.099	-.023	-.475	1.000	-.062
		Significance (2-tailed)	.004	.004	.001	.463	.000	.	.046
		df	1036	1036	1036	1036	1036	0	1036
	1 = Female, 0 = Other	Correlation	-.146	-.063	.082	.003	.073	-.062	1.000
		Significance (2-tailed)	.000	.044	.008	.932	.019	.046	.
		df	1036	1036	1036	1036	1036	1036	0

Understanding Causes of Illness (ILLCAUSE.sav)

- **Overview: Data for investigating differences in children's understanding of the causes of illness, by their health status.**
- **Source: Perrin E.C., Sayer A.G., and Willett J.B. (1991). Sticks And Stones May Break My Bones: Reasoning About Illness Causality And Body Functioning In Children Who Have A Chronic Illness, *Pediatrics*, 88(3), 608-19.**
- **Sample: 301 children, including a sub-sample of 205 who were described as asthmatic, diabetic,or healthy. After further reductions due to the *list-wise deletion* of cases with missing data on one or more variables, the analytic sub-sample used in class ends up containing: 33 diabetic children, 68 asthmatic children and 93 healthy children.**
- **Variables:**

(ILLCAUSE)	Child's Understanding of Illness Causality
(SES)	Child's SES (Note that a high score means low SES.)
(PPVT)	Child's Score on the Peabody Picture Vocabulary Test
(AGE)	Child's Age, In Months
(GENREAS)	Child's Score on a General Reasoning Test
(ChronicallyIll)	1 = Asthmatic or Diabetic, 0 = Healthy
(Asthmatic)	1 = Asthmatic, 0 = Healthy
(Diabetic)	1 = Diabetic, 0 = Healthy

Understanding Causes of Illness (ILLCAUSE.sav)

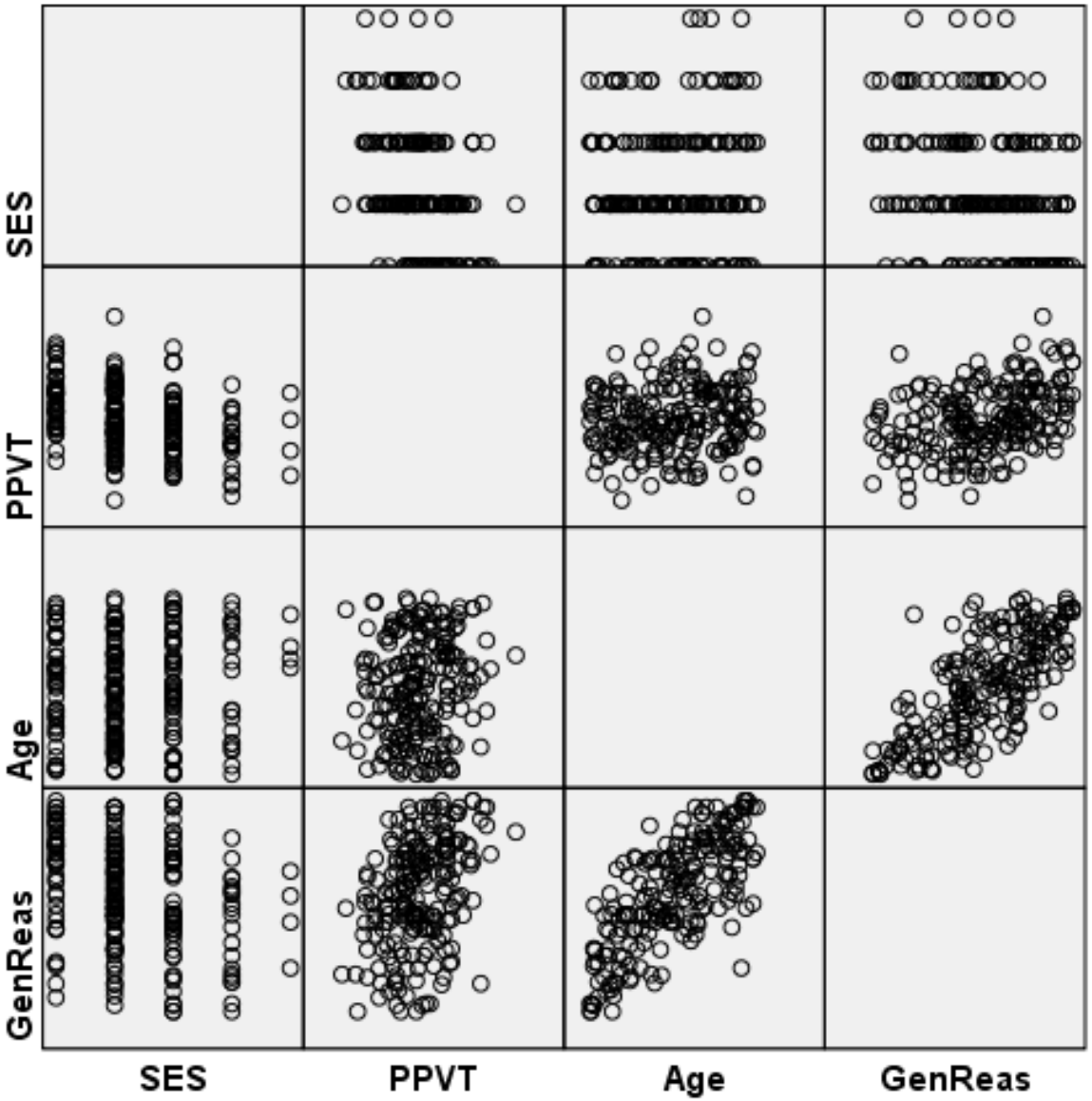
Correlations

		Understand Illness Causality	Social Class, Hollingshead	Normed PPVT	Age in Months	General Reasoning	1 = Asthmatic or Diabetic, 0 = Healthy	1 = Diabetic, 0 = Healthy	1 = Asthmatic, 0 = Healthy
Understand Illness Causality	Pearson Correlation	1.000	-.247**	.314**	.671**	.824**	-.443**	-.365**	-.440**
	Sig. (2-tailed)		.001	.000	.000	.000	.000	.000	.000
	N	194.000	194	194	194	192	194	126	161
Social Class, Hollingshead	Pearson Correlation	-.247**	1.000	-.378**	.060	-.298**	.484**	.464**	.498**
	Sig. (2-tailed)	.001		.000	.394	.000	.000	.000	.000
	N	194	205.000	205	205	203	205	132	169
Normed PPVT	Pearson Correlation	.314**	-.378**	1.000	.120	.389**	-.252**	-.274**	-.223**
	Sig. (2-tailed)	.000	.000		.087	.000	.000	.001	.004
	N	194	205	205.000	205	203	205	132	169
Age in Months	Pearson Correlation	.671**	.060	.120	1.000	.737**	-.005	.053	-.035
	Sig. (2-tailed)	.000	.394	.087		.000	.947	.548	.652
	N	194	205	205	205.000	203	205	132	169
General Reasoning	Pearson Correlation	.824**	-.298**	.389**	.737**	1.000	-.355**	-.276**	-.370**
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.001	.000
	N	192	203	203	203	203.000	203	131	168
1 = Asthmatic or Diabetic, 0 = Healthy	Pearson Correlation	-.443**	.484**	-.252**	-.005	-.355**	1.000	1.000**	1.000**
	Sig. (2-tailed)	.000	.000	.000	.947	.000		.000	.000
	N	194	205	205	205	203	205.000	132	169
1 = Diabetic, 0 = Healthy	Pearson Correlation	-.365**	.464**	-.274**	.053	-.276**	1.000**	1.000	.a
	Sig. (2-tailed)	.000	.000	.001	.548	.001	.000		.000
	N	126	132	132	132	131	132	132.000	96
1 = Asthmatic, 0 = Healthy	Pearson Correlation	-.440**	.498**	-.223**	-.035	-.370**	1.000**	.a	1.000
	Sig. (2-tailed)	.000	.000	.004	.652	.000	.000	.000	
	N	161	169	169	169	168	169	96	169.000

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

a. Cannot be computed because at least one of the variables is constant.

Understanding Causes of Illness (ILLCAUSE.sav)



Understanding Causes of Illness (ILLCAUSE.sav)

Correlations

Control Variables			Understand Illness Causality	Social Class, Hollingshead	Normed PPVT	General Reasoning	1 = Asthmatic or Diabetic, 0 = Healthy	1 = Diabetic, 0 = Healthy	1 = Asthmatic, 0 = Healthy
Age in Months	Understand Illness Causality	Correlation	1.000	-.106	.238	.496	.	.	.
		Significance (2-tailed)	.	.313	.022	.000	.	.	.
		df	0	90	90	90	90	90	90
	Social Class, Hollingshead	Correlation	-.106	1.000	-.289	-.217	.	.	.
		Significance (2-tailed)	.313	.	.005	.038	.	.	.
		df	90	0	90	90	90	90	90
	Normed PPVT	Correlation	.238	-.289	1.000	.329	.	.	.
		Significance (2-tailed)	.022	.005	.	.001	.	.	.
		df	90	90	0	90	90	90	90
	General Reasoning	Correlation	.496	-.217	.329	1.000	.	.	.
		Significance (2-tailed)	.000	.038	.001	.	.	.	.
		df	90	90	90	0	90	90	90
	1 = Asthmatic or Diabetic, 0 = Healthy	Correlation	.	.	.	.	1.000	.	.
		Significance (2-tailed)	.	.	.	.	.	.	.
		df	90	90	90	90	0	90	90
	1 = Diabetic, 0 = Healthy	Correlation	.	.	.	.	.	1.000	.
		Significance (2-tailed)	.	.	.	.	.	.	.
		df	90	90	90	90	90	0	90
	1 = Asthmatic, 0 = Healthy	Correlation	.	.	.	.	.	.	1.000
		Significance (2-tailed)	.	.	.	.	.	.	.
		df	90	90	90	90	90	90	0

Children of Immigrants (ChildrenOfImmigrants.sav)

- **Overview: "CILS is a longitudinal study designed to study the adaptation process of the immigrant second generation which is defined broadly as U.S.-born children with at least one foreign-born parent or children born abroad but brought at an early age to the United States. The original survey was conducted with large samples of second-generation children attending the 8th and 9th grades in public and private schools in the metropolitan areas of Miami/Ft. Lauderdale in Florida and San Diego, California" (from the website description of the data set).**
- **Source: Portes, Alejandro, & Ruben G. Rumbaut (2001). *Legacies: The Story of the Immigrant SecondGeneration*. Berkeley CA: University of California Press.**
- **Sample: Random sample of 880 participants obtained through the website.**
- **Variables:**

(Reading)	Stanford Reading Achievement Score
(Freelunch)	% students in school who are eligible for free lunch program
(Male)	1=Male 0=Female
(Depress)	Depression scale (Higher score means more depressed)
(SES)	Composite family SES score

Children of Immigrants (ChildrenOfImmigrants.sav)

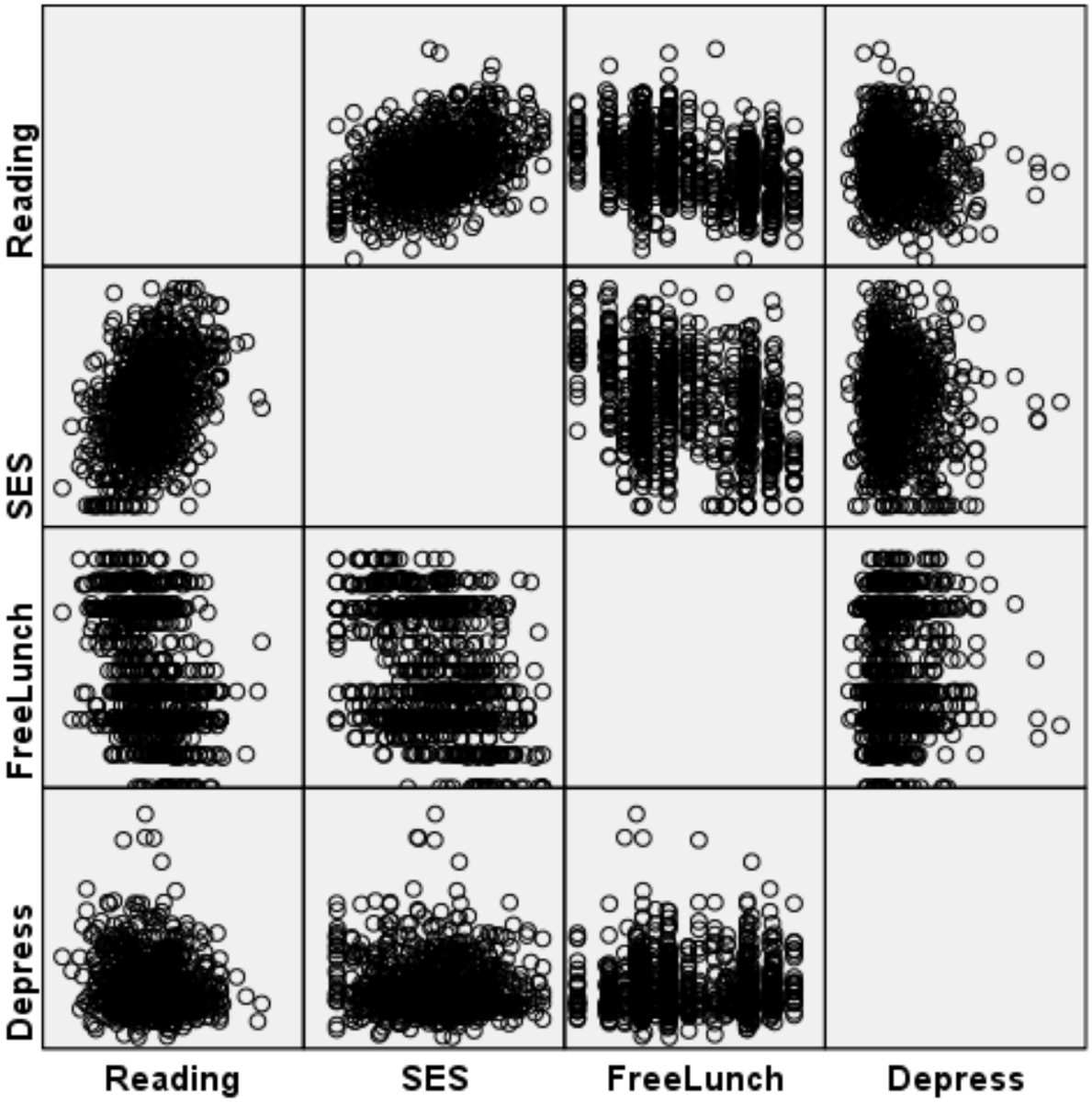
Correlations

		Stanford Reading Achievement Score	Composite Family SES Score	% of Students in Child's School Eligible for Free Lunch	Depression Scale (Higher = Greater Depression)	Male = 1, Female = 0
Stanford Reading Achievement Score	Pearson Correlation	1.000	.404**	-.353**	-.123**	-.045
	Sig. (2-tailed)		.000	.000	.000	.186
	N	880.000	880	880	880	880
Composite Family SES Score	Pearson Correlation	.404**	1.000	-.398**	-.065	.111**
	Sig. (2-tailed)	.000		.000	.054	.001
	N	880	880.000	880	880	880
% of Students in Child's School Eligible for Free Lunch	Pearson Correlation	-.353**	-.398**	1.000	.076*	-.073*
	Sig. (2-tailed)	.000	.000		.023	.031
	N	880	880	880.000	880	880
Depression Scale (Higher = Greater Depression)	Pearson Correlation	-.123**	-.065	.076*	1.000	.057
	Sig. (2-tailed)	.000	.054	.023		.088
	N	880	880	880	880.000	880
Male = 1, Female = 0	Pearson Correlation	-.045	.111**	-.073*	.057	1.000
	Sig. (2-tailed)	.186	.001	.031	.088	
	N	880	880	880	880	880.000

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

*. Correlation is significant at the 0.05 level (2-tailed).

Children of Immigrants (ChildrenOfImmigrants.sav)



Children of Immigrants (ChildrenOfImmigrants.sav)

Correlations

Control Variables			Stanford Reading Achievement Score	% of Students in Child's School Eligible for Free Lunch	Depression Scale (Higher = Greater Depression)	Male = 1, Female = 0
Composite Family SES Score	Stanford Reading Achievement Score	Correlation	1.000	-.229	-.106	-.098
		Significance (2-tailed)	.	.000	.002	.004
		df	0	877	877	877
	% of Students in Child's School Eligible for Free Lunch	Correlation	-.229	1.000	.055	-.031
		Significance (2-tailed)	.000	.	.101	.354
		df	877	0	877	877
	Depression Scale (Higher = Greater Depression)	Correlation	-.106	.055	1.000	.065
		Significance (2-tailed)	.002	.101	.	.053
		df	877	877	0	877
	Male = 1, Female = 0	Correlation	-.098	-.031	.065	1.000
		Significance (2-tailed)	.004	.354	.053	.
		df	877	877	877	0

Human Development in Chicago Neighborhoods (Neighborhoods.sav)

- **These data were collected as part of the Project on Human Development in Chicago Neighborhoods in 1995.**
- **Source: Sampson, R.J., Raudenbush, S.W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, 277, 918-924.**
- **Sample: The data described here consist of information from 343 Neighborhood Clusters in Chicago Illinois. Some of the variables were obtained by project staff from the 1990 Census and city records. Other variables were obtained through questionnaire interviews with 8782 Chicago residents who were interviewed in their homes.**
- **Variables:**

(Homr90)	Homicide Rate c. 1990
(Murder95)	Homicide Rate 1995
(Disadvan)	Concentrated Disadvantage
(Imm_Conc)	Immigrant
(ResStab)	Residential Stability
(Popul)	Population in 1000s
(CollEff)	Collective Efficacy
(Victim)	% Respondents Who Were Victims of Violence
(PercViol)	% Respondents Who Perceived Violence

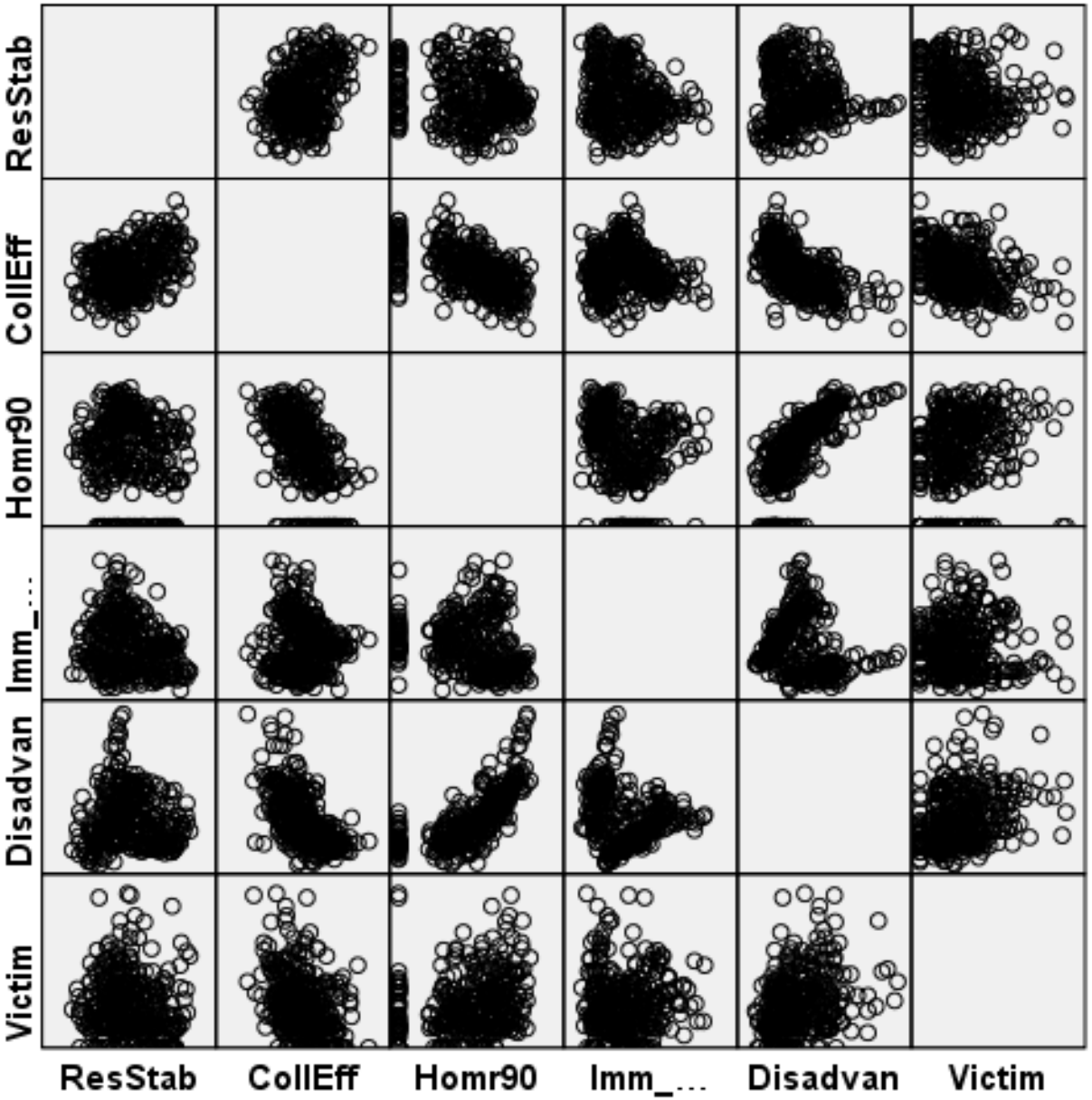
Human Development in Chicago Neighborhoods (Neighborhoods.sav)

Correlations

		Residential stability	Collective efficacy	Homicide rate 1988-90	Immigrant concentration	Concentrated disadvantage	% resp who were victims
Residential stability	Pearson Correlation	1.000	.382**	-.147**	-.216**	-.046	-.102
	Sig. (2-tailed)		.000	.007	.000	.400	.060
	N	342.000	342	342	342	342	342
Collective efficacy	Pearson Correlation	.382**	1.000	-.579**	-.047	-.624**	-.366**
	Sig. (2-tailed)	.000		.000	.385	.000	.000
	N	342	342.000	342	342	342	342
Homicide rate 1988-90	Pearson Correlation	-.147**	-.579**	1.000	-.201**	.731**	.242**
	Sig. (2-tailed)	.007	.000		.000	.000	.000
	N	342	342	342.000	342	342	342
Immigrant concentration	Pearson Correlation	-.216**	-.047	-.201**	1.000	-.217**	.033
	Sig. (2-tailed)	.000	.385	.000		.000	.543
	N	342	342	342	342.000	342	342
Concentrated disadvantage	Pearson Correlation	-.046	-.624**	.731**	-.217**	1.000	.318**
	Sig. (2-tailed)	.400	.000	.000	.000		.000
	N	342	342	342	342	342.000	342
% resp who were victims	Pearson Correlation	-.102	-.366**	.242**	.033	.318**	1.000
	Sig. (2-tailed)	.060	.000	.000	.543	.000	
	N	342	342	342	342	342	342.000

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

Human Development in Chicago Neighborhoods (Neighborhoods.sav)

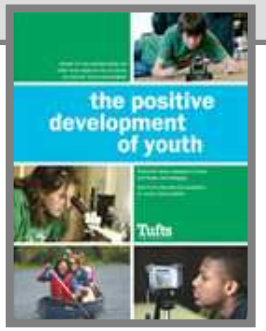


Human Development in Chicago Neighborhoods (Neighborhoods.sav)

Correlations

Control Variables			Residential stability	Collective efficacy	Homicide rate 1988-90	Immigrant concentration	% resp who were victims
Concentrated disadvantage	Residential stability	Correlation	1.000	.453	-.166	-.232	-.092
		Significance (2-tailed)	.	.000	.002	.000	.090
		df	0	339	339	339	339
	Collective efficacy	Correlation	.453	1.000	-.231	-.240	-.226
		Significance (2-tailed)	.000	.	.000	.000	.000
		df	339	0	339	339	339
	Homicide rate 1988-90	Correlation	-.166	-.231	1.000	-.064	.015
		Significance (2-tailed)	.002	.000	.	.240	.783
		df	339	339	0	339	339
	Immigrant concentration	Correlation	-.232	-.240	-.064	1.000	.110
		Significance (2-tailed)	.000	.000	.240	.	.042
		df	339	339	339	0	339
	% resp who were victims	Correlation	-.092	-.226	.015	.110	1.000
		Significance (2-tailed)	.090	.000	.783	.042	.
		df	339	339	339	339	0

4-H Study of Positive Youth Development (4H.sav)

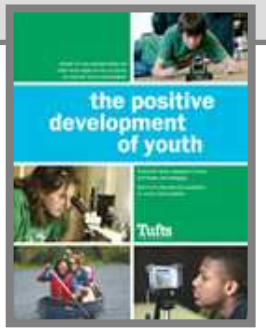


- 4-H Study of Positive Youth Development
- Source: Subset of data from IARYD, Tufts University
- Sample: These data consist of seventh graders who participated in Wave 3 of the 4-H Study of Positive Youth Development at Tufts University. This subfile is a substantially sampled-down version of the original file, as all the cases with any missing data on these selected variables were eliminated.
- Variables:

(SexFem)	1=Female, 0=Male
(MothEd)	Years of Mother's Education
(Grades)	Self-Reported Grades
(Depression)	Depression (Continuous)
(FrInfl)	Friends' Positive Influences
(PeerSupp)	Peer Support
(Depressed)	0 = (1-15 on Depression) 1 = Yes (16+ on Depression)

(AcadComp)	Self-Perceived Academic Competence
(SocComp)	Self-Perceived Social Competence
(PhysComp)	Self-Perceived Physical Competence
(PhysApp)	Self-Perceived Physical Appearance
(CondBeh)	Self-Perceived Conduct Behavior
(SelfWorth)	Self-Worth

4-H Study of Positive Youth Development (4H.sav)

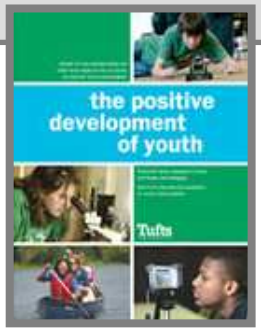


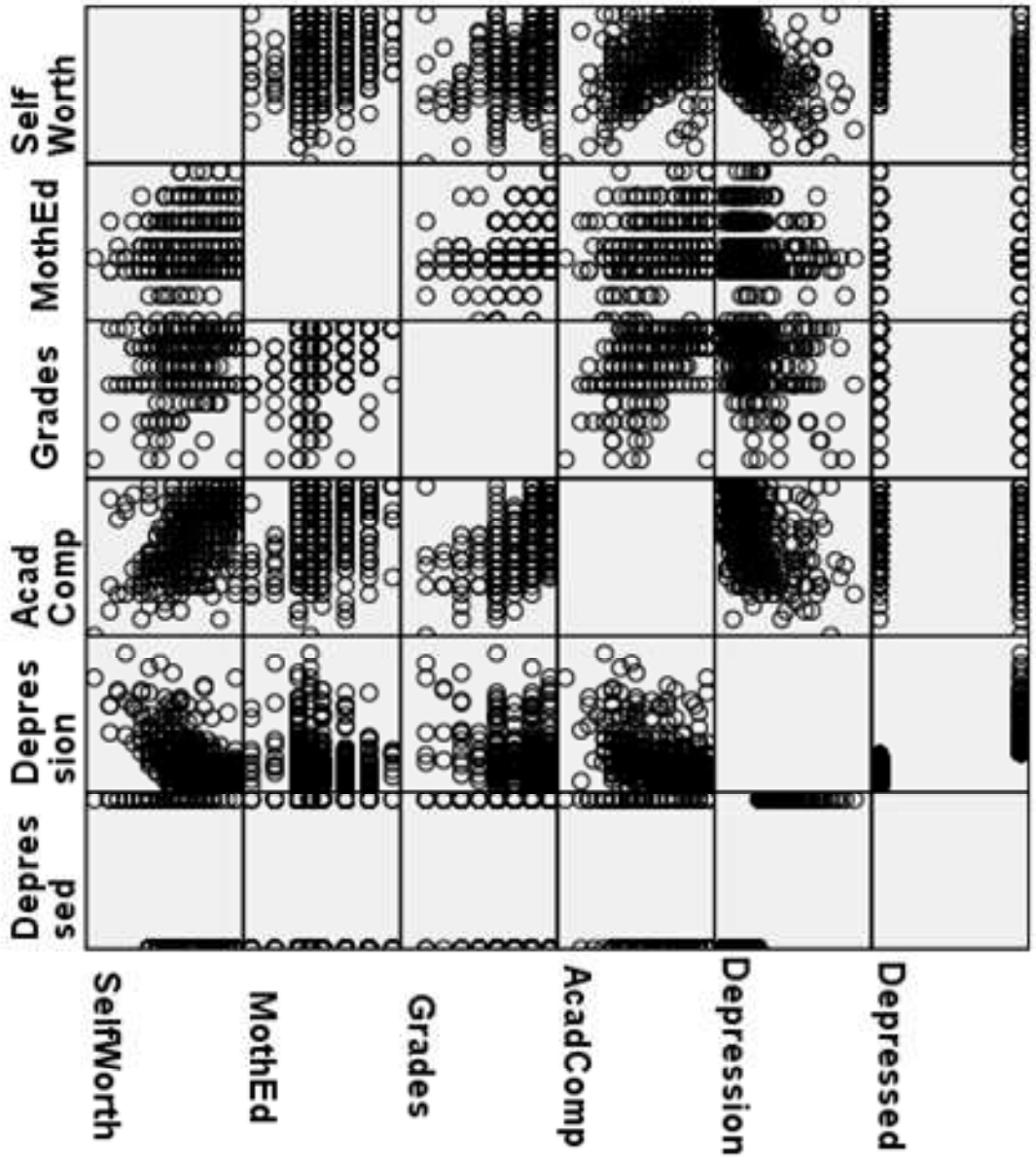
Correlations

		Self-Worth	Birth Mother Education	Grades in School	Self-Percieved Academic Competence	Depression	Depressed = 1, Not Depressed = 0
Self-Worth	Pearson Correlation	1.000	.172**	.345**	.531**	-.559**	-.504**
	Sig. (2-tailed)		.000	.000	.000	.000	.000
	N	409.000	409	409	409	409	409
Birth Mother Education	Pearson Correlation	.172**	1.000	.267**	.322**	-.165**	-.129**
	Sig. (2-tailed)	.000		.000	.000	.001	.009
	N	409	409.000	409	409	409	409
Grades in School	Pearson Correlation	.345**	.267**	1.000	.560**	-.375**	-.291**
	Sig. (2-tailed)	.000	.000		.000	.000	.000
	N	409	409	409.000	409	409	409
Self-Percieved Academic Competence	Pearson Correlation	.531**	.322**	.560**	1.000	-.414**	-.350**
	Sig. (2-tailed)	.000	.000	.000		.000	.000
	N	409	409	409	409.000	409	409
Depression	Pearson Correlation	-.559**	-.165**	-.375**	-.414**	1.000	.803**
	Sig. (2-tailed)	.000	.001	.000	.000		.000
	N	409	409	409	409	409.000	409
Depressed = 1, Not Depressed = 0	Pearson Correlation	-.504**	-.129**	-.291**	-.350**	.803**	1.000
	Sig. (2-tailed)	.000	.009	.000	.000	.000	
	N	409	409	409	409	409	409.000

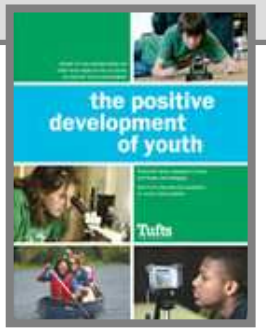
**. Correlation is significant at the 0.01 level (2-tailed).

4-H Study of Positive Youth Development (4H.sav)





4-H Study of Positive Youth Development (4H.sav)



Correlations

Control Variables			Self-Worth	Birth Mother Education	Self-Percieved Academic Competence	Depression	Depressed = 1, Not Depressed = 0
Grades in School	Self-Worth	Correlation	1.000	.088	.435	-.494	-.449
		Significance (2-tailed)	.	.074	.000	.000	.000
		df	0	406	406	406	406
	Birth Mother Education	Correlation	.088	1.000	.216	-.073	-.056
		Significance (2-tailed)	.074	.	.000	.143	.259
		df	406	0	406	406	406
	Self-Percieved Academic Competence	Correlation	.435	.216	1.000	-.266	-.236
		Significance (2-tailed)	.000	.000	.	.000	.000
		df	406	406	0	406	406
	Depression	Correlation	-.494	-.073	-.266	1.000	.782
		Significance (2-tailed)	.000	.143	.000	.	.000
		df	406	406	406	0	406
	Depressed = 1, Not Depressed = 0	Correlation	-.449	-.056	-.236	.782	1.000
		Significance (2-tailed)	.000	.259	.000	.000	.
		df	406	406	406	406	0