Unit 4: Pearson Product-Moment Correlations (*r*)

Unit 4 Post Hole:

Interpret a correlation matrix.

Unit 4 Technical Memo and School Board Memo:

Produce an appropriate table, and discuss a correlation matrix for five variables (from Memo 3, plus an additional continuous or dichotomous predictor of your choice).

*Note: To set us up for Unit 6, let's makes sure that there is at least one statistically significant relationship (whatever that means!) between your outcome and a predictor and that there is at least one statistically **in**significant relationship between your outcome and a predictor. Check with me before you finalize your choice of new predictor.*

Unit 4 Reading:

http://onlinestatbook.com/
Chapter 4, Describing Bivariate Data

Unit 4: Technical Memo and School Board Memo

Work Products (Part I of II):

I. Technical Memo: Have one section per biviariate analysis. For each section, follow this outline. (2 Sections)

- 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 the two 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 Short 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.
- C. Correlations. Provide an overview of the relationships between your variables using descriptive statistics.
  - i. Interpret all the correlations with your outcome variable. Compare and contrast the correlations in order to ground your analysis in substance. (1 Paragraph)
  - ii. Interpret the correlations among your predictors. Discuss the implications for your theory. As much as possible, tell a coherent story. (1 Paragraph)
  - iii. As you narrate, note any concerns regarding assumptions (e.g., outliers or non-linearity), and, if a correlation is uninterpretable because of an assumption violation, then do not interpret it.

Unit 4: 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. (1 Paragraph)

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 the 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. Describe the direction and magnitude of the relationship in your sample, preferably with illustrative examples. Draw out the substance of your findings through your narrative.

v. Use confidence intervals to describe the precision of your magnitude estimates so that you can discuss the magnitude in the population.

vi. If simple linear regression is inappropriate, then say so, briefly explain why, and forego any misleading analysis.

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)**

ii. **For the relationship between your outcome and predictor, use a coherent narrative to convey the results of your exploratory bivariate analysis of the data. (1 Paragraph)**

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 4: 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):

FREELUNCH, a dichotomous variable, 1 = Eligible for Free/Reduced Lunch and 0 = Not
RACE, a polychotomous variable, 1 = Asian, 2 = Latino, 3 = Black and 4 = White

- Unit 1: In our sample, is there a relationship between reading achievement and free lunch?
- Unit 2: In our sample, what does reading achievement look like (from an outlier resistant perspective)?
- Unit 3: In our sample, what does reading achievement look like (from an outlier sensitive perspective)?
- Unit 4: In our sample, how strong is the relationship between reading achievement and free lunch?
- Unit 5: In our sample, free lunch predicts what proportion of variation in reading achievement?
- Unit 6: In the population, is there a relationship between reading achievement and free lunch?
- Unit 7: In the population, what is the magnitude of the relationship between reading and free lunch?
- Unit 8: What assumptions underlie our inference from the sample to the population?
- Unit 9: In the population, is there a relationship between reading and race?
- Unit 10: In the population, is there a relationship between reading and race controlling for free lunch?
- Appendix A: In the population, is there a relationship between race and free lunch?

Unit 4: Roadmap (R Output)

```
> load("E:/User/Folder/RoadmapData.rda")
> library(abind, pos=4)
> numSummary(RoadmapData[,c("FREELUNCH", "READING")],
+   statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))
                mean       sd    0%   25%   50%   75%  100%    n
FREELUNCH  0.3353846 0.472155  0.00  0.00  0.00  1.00  1.00 7800
READING   47.4940397 8.569440 23.96 41.24 47.43 53.93 63.49 7800
```

Unit 3

Unit 2

```
> RegModel.1 <- lm(READING~FREELUNCH, data=RoadmapData)
> summary(RegModel.1, cor=FALSE)
Call:
lm(formula = READING ~ FREELUNCH, data = RoadmapData)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.1176     0.1147  428.17   <2e-16 ***
FREELUNCH    -4.8409     0.1981  -24.44   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.26 on 7798 degrees of freedom
Multiple R-squared: 0.07114, Adjusted R-squared: 0.07102
F-statistic: 597.3 on 1 and 7798 DF,  p-value: < 2.2e-16
```

Unit 1 Unit 8 Unit 6

Unit 5

Unit 9

```
> library(MASS, pos=4)
> Confint(RegModel.1, level=.95)
             Estimate     2.5 %     97.5 %
(Intercept) 49.117616 48.892742 49.342489
FREELUNCH   -4.840938 -5.229237 -4.452638
```

Unit 7

```
> cor(RoadmapData[,c("FREELUNCH","READING")])
            FREELUNCH    READING
FREELUNCH   1.0000000 -0.2667237
READING    -0.2667237  1.0000000
```

Unit 4

Unit 4: Roadmap (SPSS Output)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.267[a]	.071	.071	8.25952

a. Predictors: (Constant), FREELUNCH

Unit 5

Statistics

		READING	FREELUNCH
N	Valid	7800	7800
	Missing	0	0
Mean		47.4940	.3354
Std. Deviation		8.56944	.47216
Minimum		23.96	.00
Maximum		63.49	1.00
Percentiles	25	41.2400	.0000
	50	47.4300	.0000
	75	53.9300	1.0000

Unit 3

Unit 2

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	40744.322	1	40744.322	597.251	.000[a]
	Residual	531977.541	7798	68.220		
	Total	572721.864	7799			

a. Predictors: (Constant), FREELUNCH

b. Dependent Variable: READING

Unit 9

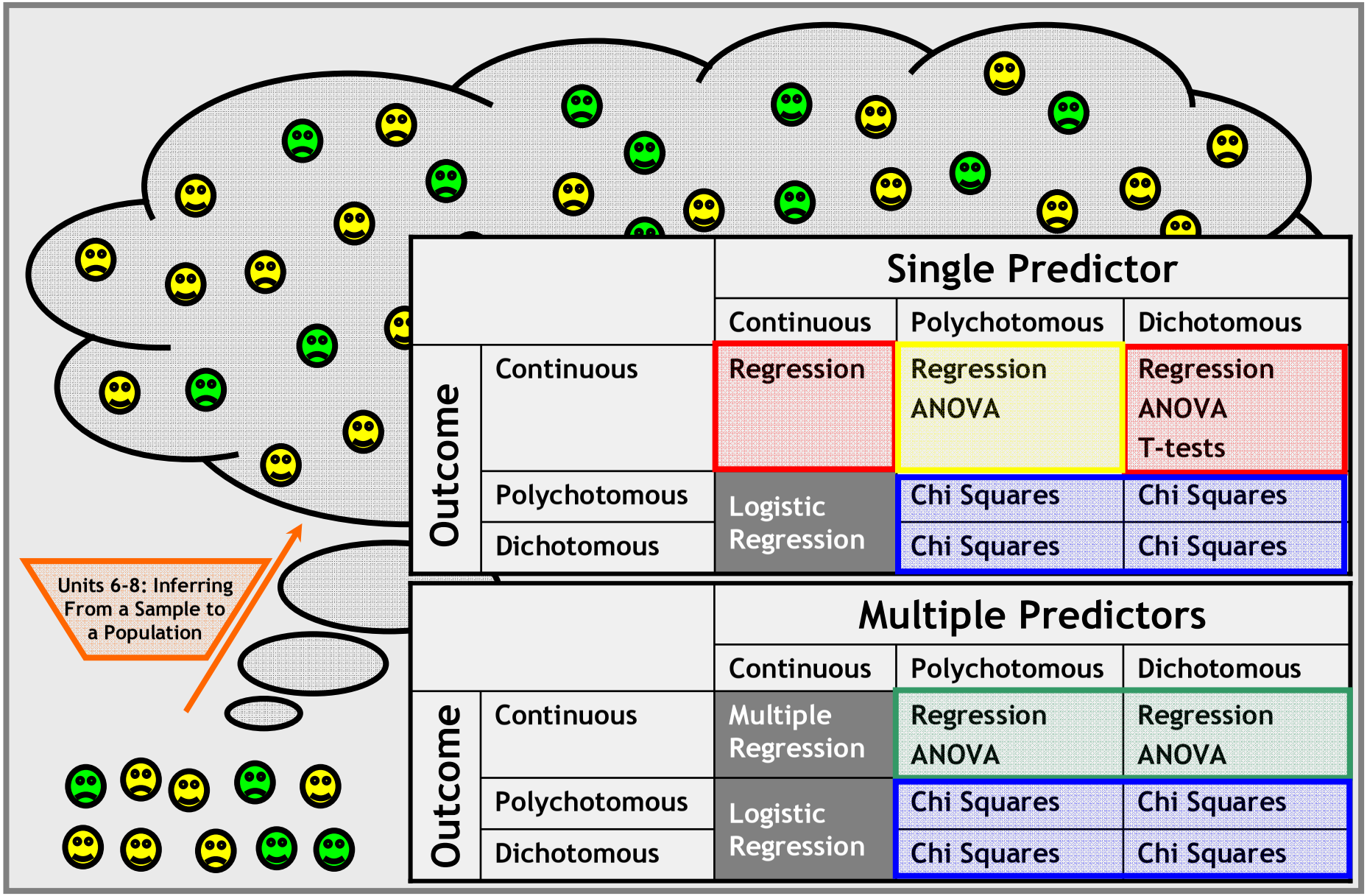
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	49.118	.115		428.169	.000	48.893	49.342
	FREELUNCH	-4.841	.198	-.267	-24.439	.000	-5.229	-4.453

a. Dependent Variable: READING

Unit 1 Unit 8 Unit 4 Unit 6 Unit 7

Unit 4: Road Map (Schematic)



Single Predictor

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

Multiple Predictors

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

Epistemological Minute

David Hume (1711-1776) is a great skeptical philosopher. He forces us to appreciate a concept that we often take for granted—causality. As data analysts, we are often inclined to conclude that our predictor causes our outcome. Hume argues that such a conclusion is only warranted if three conditions hold: Contiguity, Succession and Necessary Connexion. I am going to loosely (but usefully) translate "Contiguity" as correlation. If our predictor causes our outcome, then our predictor must be correlated with our outcome. "Succession" is about temporal order. If our predictor causes our outcome, then our predictor must precede (in time) our outcome. Finally, there is "Necessary Connexion." It is not enough that our predictor be correlated with our outcome and precede it temporally. After all, the rooster's crowing is correlated with the sunrise and precedes it temporarily. I hope that this simple but powerful example convinces you that there is something more to causality than correlation and succession. The precise nature of that something extra, which Hume calls "Necessary Connexion," is problematic.

For Hume, there are two sources of rational human understanding:

Relations Of Ideas, which derive from math and logic

Matters Of Facts, which derive from observation and experience

Correlation: We can observe that the predictor and outcome tend to go together, a *Matter Of Fact*.

Succession: We can observe that the predictor precedes the outcome, a *Matter Of Fact*.

Necessary Connexion: How do we observe that "something extra" beyond correlation and succession? In order to understand Necessary Connexion, we must observe it, because the connection between our predictor and outcome is not a matter of math or logic, it is not a *Relation of Ideas*, it is a *Matter of Fact*. Yet, all we observe is that our predictor happens and then our outcome happens; where do we observe the Necessary Connexion? Hume is convinced that we cannot observe Necessary Connexion and, therefore, causality is not rational. I don't ask you to be as skeptical about causality as Hume, but I do ask you to give the concept some thought.

Epistemological Minute

3 Causal Rules of 3

I. Causal conclusions require 3 conditions:
   A. Correlation
   B. Succession
   C. Necessary Connexion

II. In addition to your Predictor and Outcome, always consider the possible influence of a 3rd Hidden Confounding Variable.

III. When presenting your pet causal conclusion, present 2 other plausible causal conclusions for the sake of balance.



Predictor – – – – – – – Outcome

E.g., homework is positively correlated with test scores.

Predictor → – – – – – – Outcome

E.g., homework is positively correlated with test scores.

Doing homework gives students the necessary skill- and knowledge-building practice to increase their performance on examinations.

Predictor ← – – – – – – Outcome

E.g., homework is positively correlated with test scores.

Success on exams encourages students to fully participate in school and its academic requirements, including homework.

Predictor ↔ – – – – – – Outcome

E.g., homework is positively correlated with test scores.

Doing homework gives students the necessary skill- and knowledge-building practice to increase their performance on examinations, which in turn encourages students to fully participate in school and its academic requirements, including homework, which in turn gives students the necessary skill- and knowledge-building practice to increase their performance on examinations…

Third/Hidden/Confounding Variable
E.g., SES.

Predictor – – – – – – – Outcome

E.g., homework is positively correlated with test scores.




Causality can be very complicated. It often involves far more than two (or three) variables. A causal relationship between two variables can be mediated by hidden variables and moderated by other hidden variables. Meanwhile, the causal arrows are pointing every which way.

Unit 4: Research Question

Theory: Students who go to smaller schools will have better math achievement scores, because smaller schools form tighter communities, and consequently struggling and gifted students are less likely to fall through the cracks.

Research Question: Are students' math achievement scores negatively correlated with their school population size?

Data Set: (NELS88Math.sav)

Variables:

Outcome—Math Achievement Score (*MATHACH*)

Predictor—Number of Students in Student's School (*SchoolPop*)

Model: $MathAch = \beta_0 + \beta_1 SchoolPop + \varepsilon$

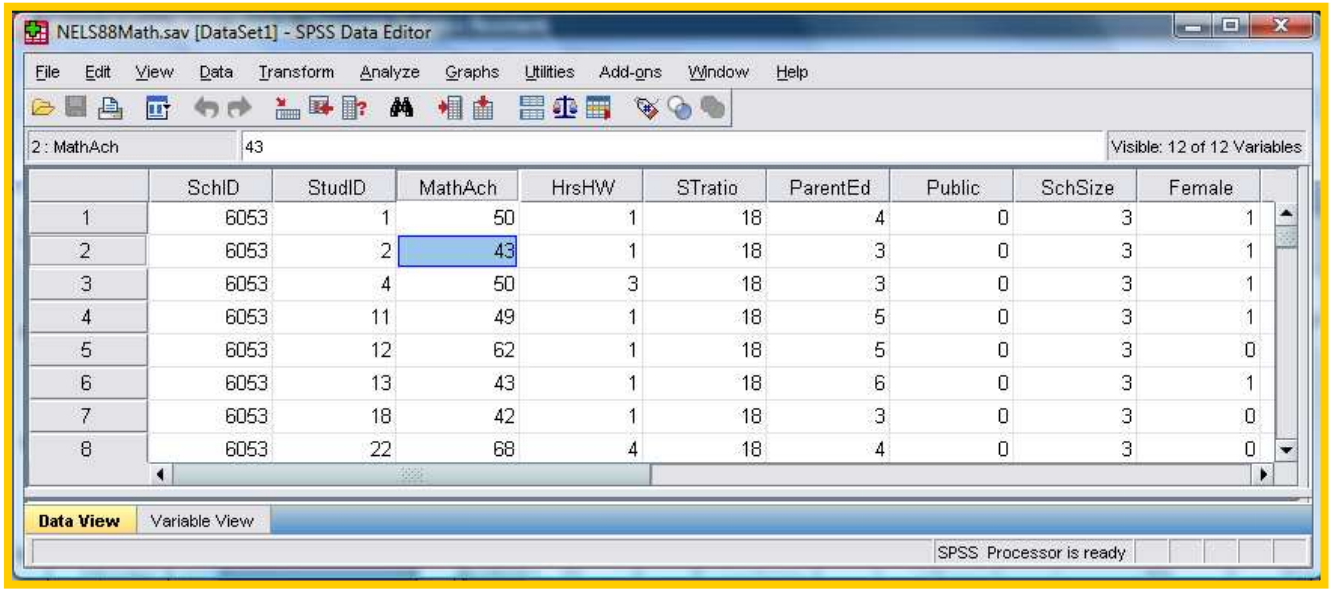
NELS88Math.Sav Codebook

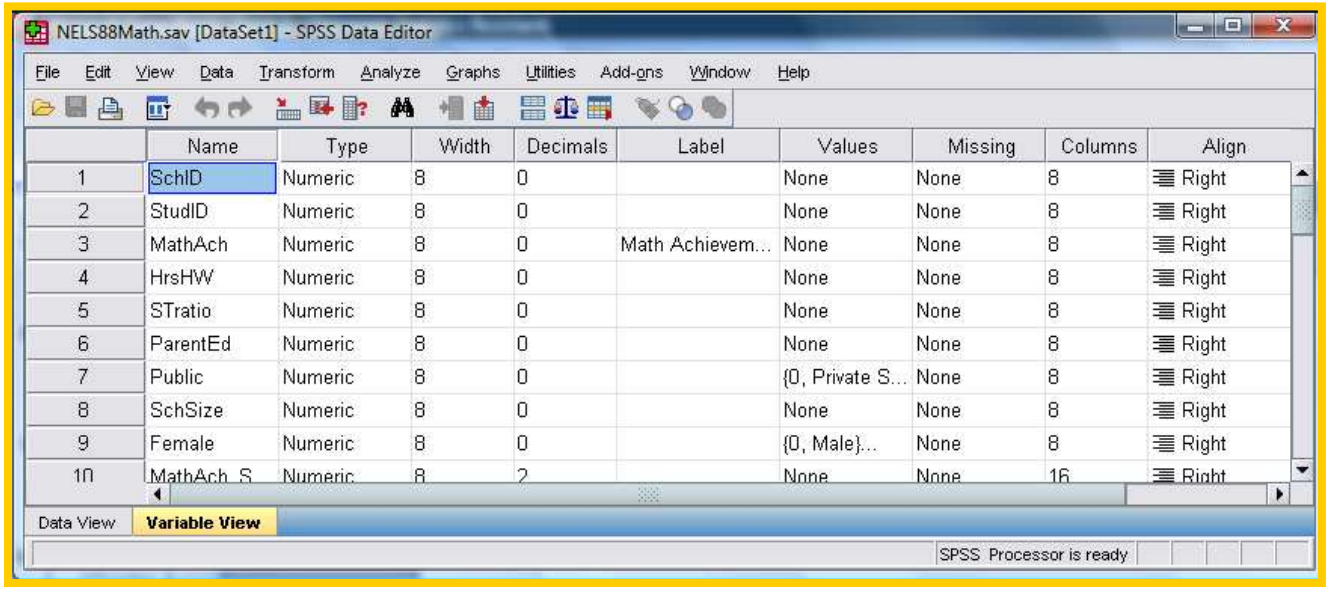
Dataset	NELS88Math.txt
Overview	Multilevel dataset on the mathematics achievement of 519 students in 23 schools, as a function of the number of hours of mathematics homework they complete each week and the student teacher ratio in their school, by selected controls.
Source	Kreft, I.G., & de Leeuw, J. *Introducing Multilevel Modeling*. Thousand Oaks, CA: Sage Publications, 1998, pp. 23-24. Data are a sub-sample from NELS-88 , which contains information on educational processes and outcomes for a nationally representative sample of eighth-graders first surveyed in 1988, and then again in 1990, 1992, 1994, and 2000. Students reported data on school, work, neighborhood, and home experiences; educational resources available to them; educational and occupational aspirations; substance abuse; and the education levels of parents and peers;. The reading, social studies, mathematics and science achievement of students were measured while they were in school. Background information was provided by teachers, parents, and school administrators. The public use dataset is available on CD-ROM and is free from NCES.
Sample size	23 schools, 519 students
Last updated	October 8, 2003

NELS88Math.Sav Codebook

Structure of Dataset			
Col. #	**Variable Name**	**Variable Description**	**Variable Metric/Labels**
1	SCHID	School identification code	Integer
2	STUID	Student identification code	Integer
3	MATHACH	Mathematics number-right achievement score	Continuous variable ranging from 30 to 71.
4	HOURSHW	Number of hours of mathematics homework completed each week	Ordinal variable: 0 = none 1 = less than 1 hour 2 = 1 hour 3 = 2 hours 4 = 3 hours 5 = 4 to 6 hours 6 = 7 to 9 hours 7 = 10 hours or more
5	STRATIO	Student/teacher ratio in the school	Continuous variable ranging from 10 to 28: 10 = 10 or less 11 = 11. etc.
6	PARENTED	Highest educational level attained by either parent.	Ordinal variable: 1 = Did not finish HS 2 = HS Grad/GED 3 = >HS & <4yr degree 4 = College grad. 5 = MA, or equiv. 6 = Ph.D., M.D., or equiv.
7	PUBLIC	Is the school in the public sector?	Dichotomous variable: 0 = no 1 = yes
8	SCHSIZE	Total school enrollment	Ordinal variable: 1 = 1–199 students 2 = 200-399 students 3 = 400-599 students 4 = 600-799 students 5 = 800-999 students 6 = 1000-1199 students 7 = 1200+ students
9	FEMALE	Is the student female?	Dichotomous variable: 0 = no 1 = yes

NELS88Math.sav





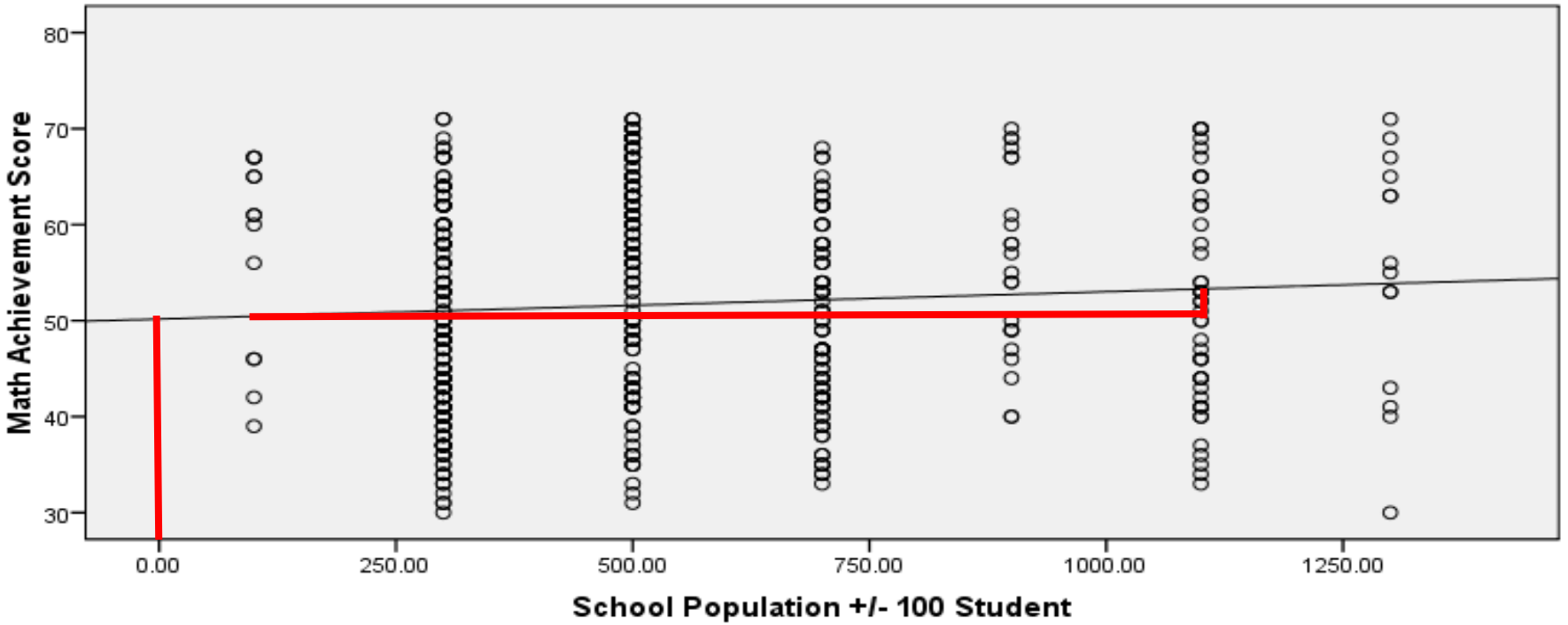
Exploring Math Achievement and School Size

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	50.167	1.029		48.767	.000
	School Population +/- 100 Student	.003	.002	.075	1.700	.090

a. Dependent Variable: Math Achievement Score

$$\hat{MathAch} = 50.2 + 0.003 SchoolPop$$



Exploring Math Achievement and School Size

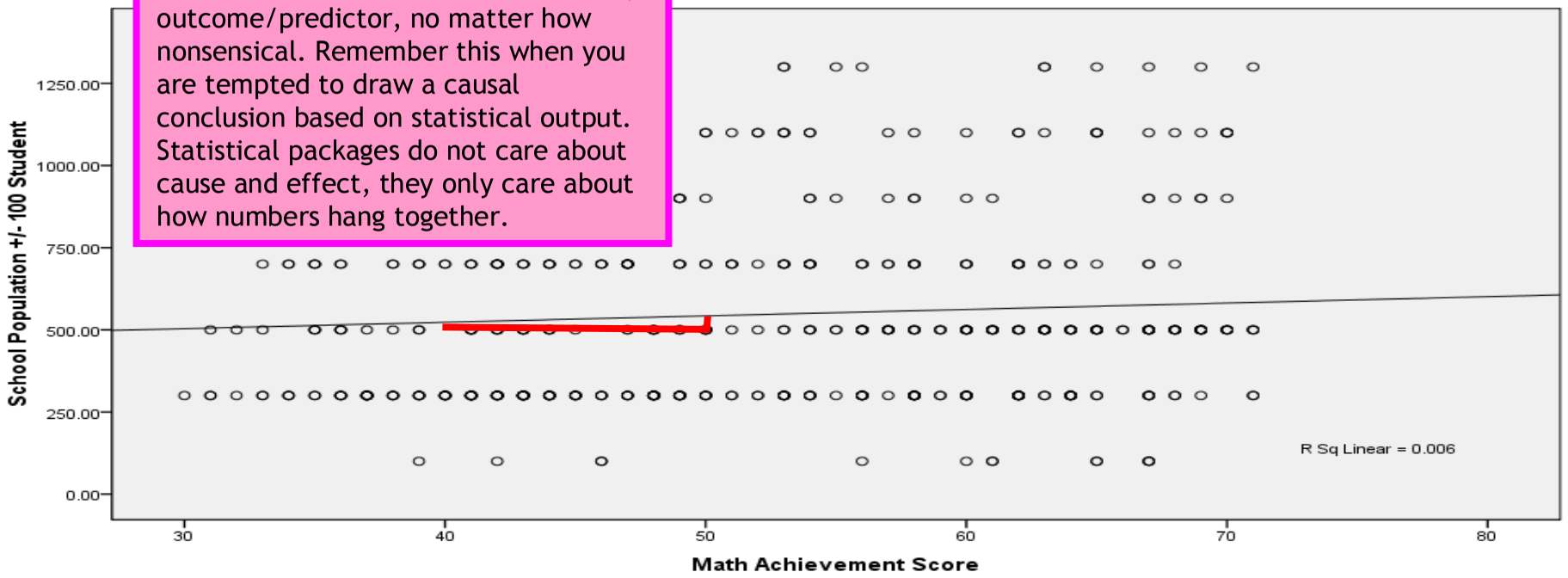
Before we leave this slide, let's note that the magnitude in this model (slope = 1.95) is bigger than the magnitude in the "reversed" model (slope = 0.003). Regression is not symmetric! When we reverse the outcome and predictor, we get different slopes. This make sense once we think about it. In the former model, the slope was telling us the predicted difference in math scores associated with a one-student difference in school population. In the current model, the slope is telling us the predicted difference in school population associated with a one-point difference in math scores.

Model		Unstandardized Coefficients		Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	445.035	60.578		7.346	.000
	Math Achievement Score	1.949	1.147	.075	1.700	.090

a. Dependent Variable: School Population +/- 100 Student

$$\hat{SchoolPop} = 445 + 1.95 MathAch$$

Notice that SPSS will fit a model to any outcome/predictor, no matter how nonsensical. Remember this when you are tempted to draw a causal conclusion based on statistical output. Statistical packages do not care about cause and effect, they only care about how numbers hang together.



Exploring Math Achievement and School Size

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-4.943E-17	.044		.000	1.000
	Zscore: School Population +/- 100 Student	.075	.044	.075	1.700	.090

a. Dependent Variable: Zscore: Math Achievement

$$\hat{ZMathAch} = 0.0 + 0.075 ZSchoolPop$$

Exploring Math Achievement and School Size

Notice that the slope is the same! The .075 here is the same .075 as the last slide.
Regression is not (necessarily) symmetric, but correlation is symmetric.

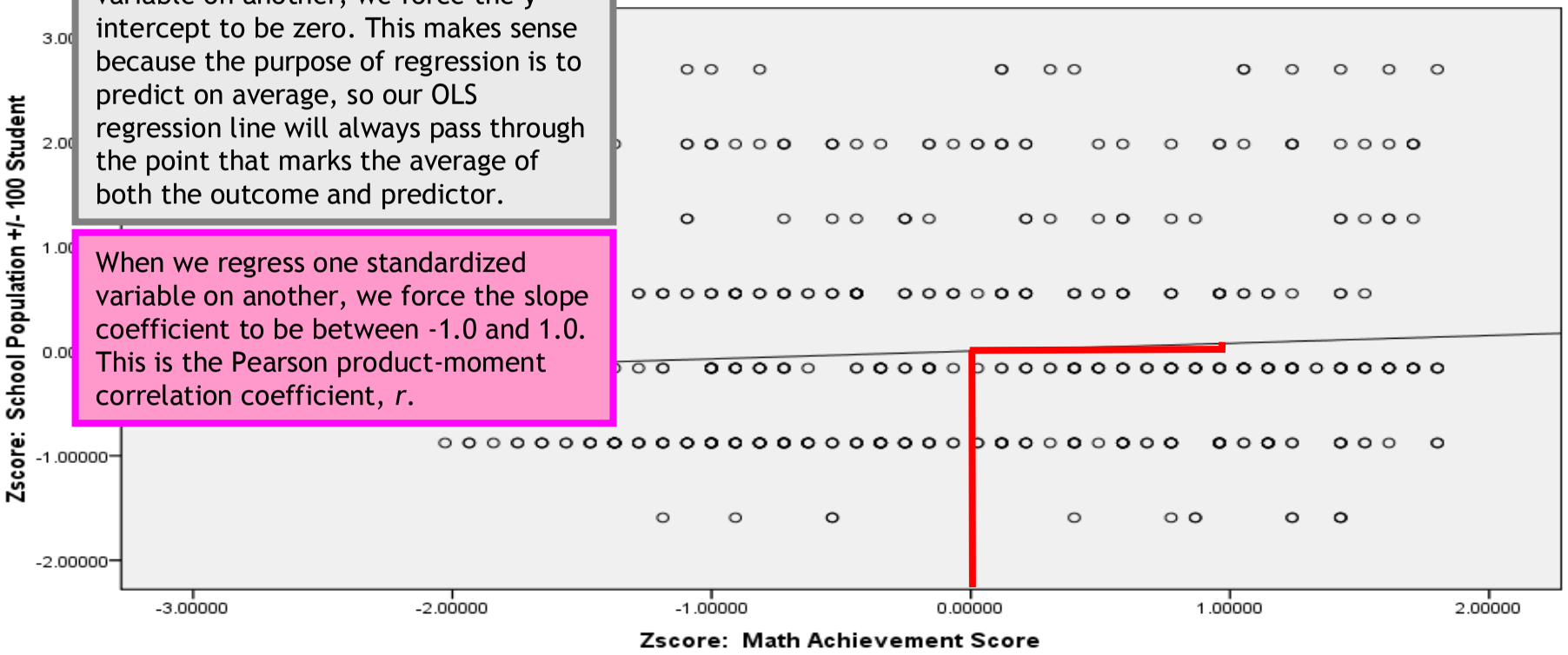
Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-1.052E-16	.044		.000	1.000
	Zscore: Math Achievement Score	.075	.044	.075	1.700	.090

a. Dependent Variable: Zscore: School Population +/- 100 Student

$$\hat{ZSchoolPop} = 0.0 + 0.075 ZMathAch$$

When we regress one standardized variable on another, we force the y-intercept to be zero. This makes sense because the purpose of regression is to predict on average, so our OLS regression line will always pass through the point that marks the average of both the outcome and predictor.

When we regress one standardized variable on another, we force the slope coefficient to be between -1.0 and 1.0. This is the Pearson product-moment correlation coefficient, *r*.



Pearson Correlations Are Standardized Slopes

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	50.167	1.029		48.767	.000
	School Population +/- 100 Student	.003	.002	.075	1.700	.090

a. Dependent Variable: Math Achievement Score

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	445.035	60.578		7.346	.000
	Math Achievement Score	1.949	1.147	.075	1.700	.090

a. Dependent Variable: School Population +/- 100 Student

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-4.943E-17	.044		.000	1.000
	Zscore: School Population +/- 100 Student	.075	.044	.075	1.700	.090

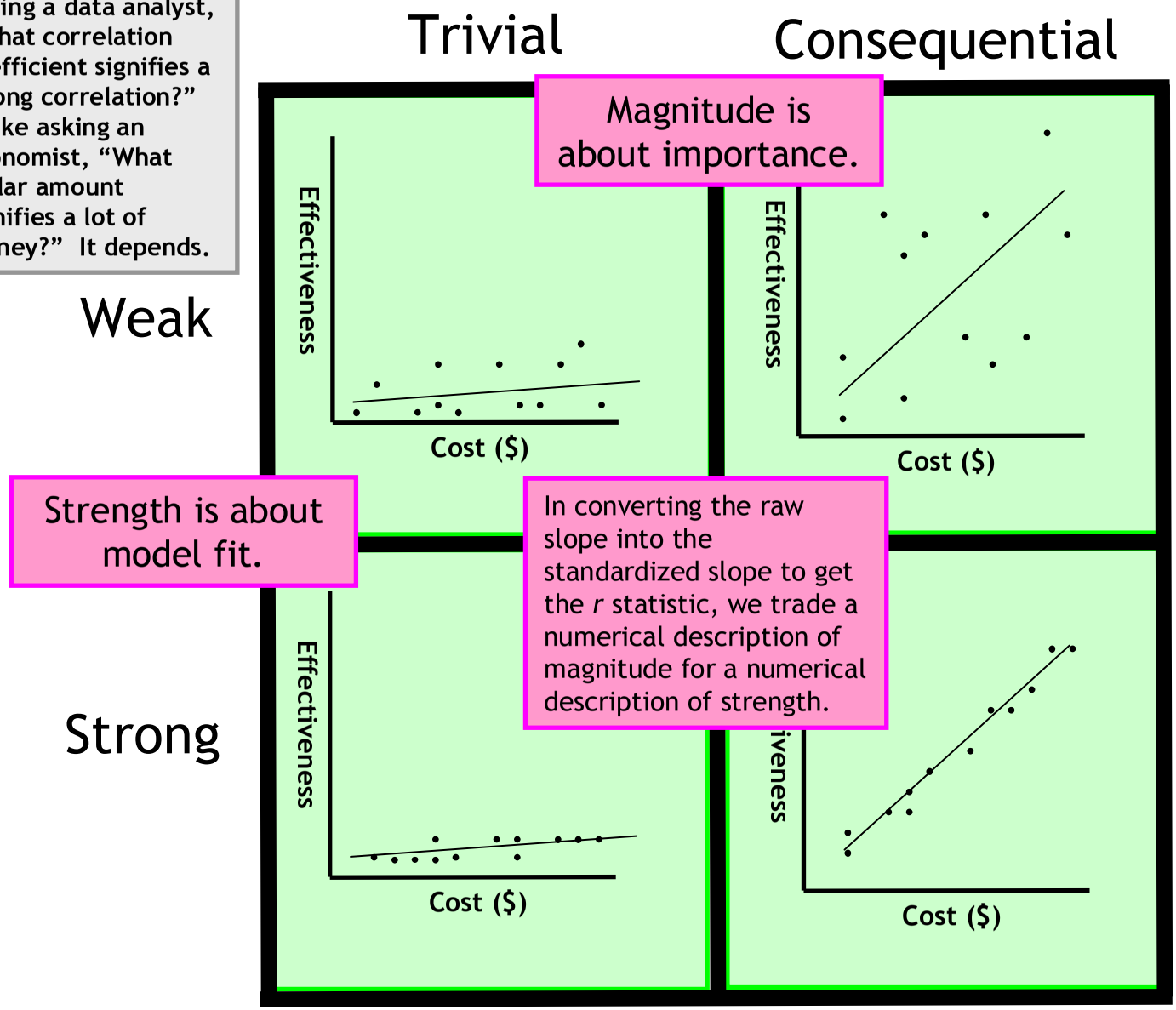
a. Dependent Variable: Zscore: Math Achievement Score

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.052E-16	.044		.000	1.000
	Zscore: Math Achievement Score	.075	.044	.075	1.700	.090

a. Dependent Variable: Zscore: School Population +/- 100 Student

Conceptually Distinct: Strength and Magnitude (Redux)

Asking a data analyst, "What correlation coefficient signifies a strong correlation?" is like asking an economist, "What dollar amount signifies a lot of money?" It depends.



What is the bang for your buck? This is magnitude.

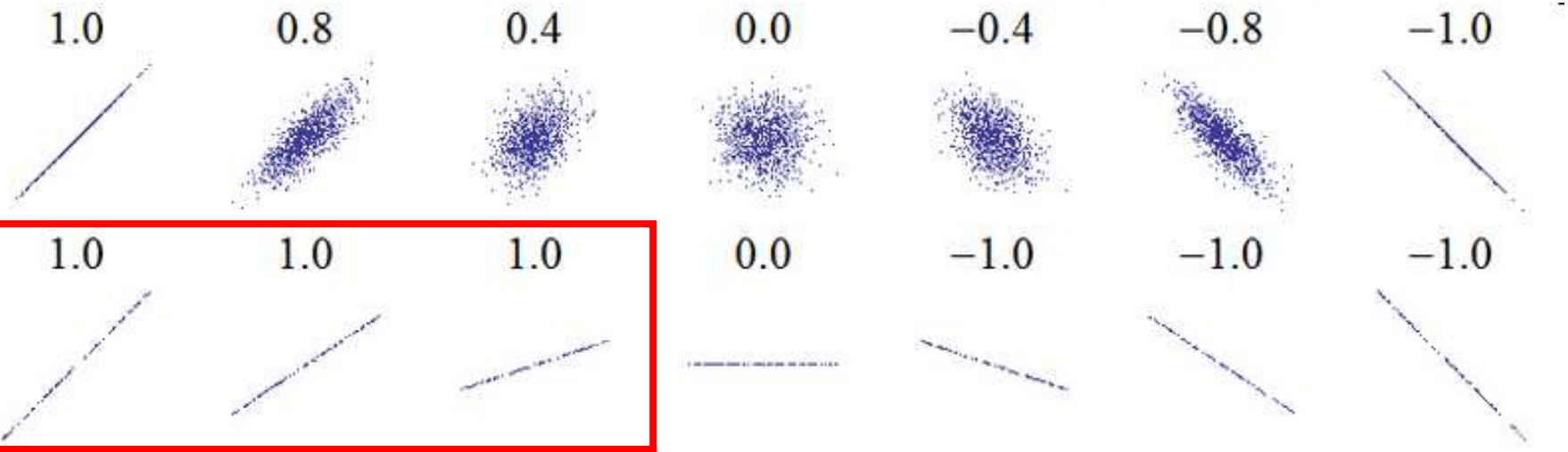
We cannot assess magnitude without a substantive understanding of the outcome and predictor. If we don't know the value of a ruble, we can't really answer, "what is the bang for your ruble?" Do not be charmed by the apparent slope, which fluctuates arbitrarily with the lengths of the X axis and Y axis.

How tightly do the data hug the trend line? Think vertically. This is strength.

Unlike magnitude, strength has nifty statistics such as the Pearson product-moment correlation coefficient (r).

Examples of Pearson Product-Moment Correlations

http://en.wikipedia.org/wiki/Correlation



Notice how different magnitudes can have the same correlation, and, in these cases, the same perfect correlation.

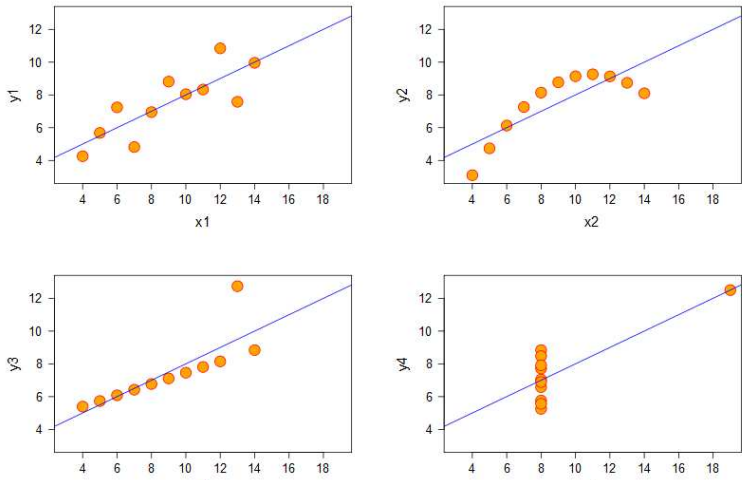
You want to know which values of the r statistic are big and which are small, but there are no easy answers. Beware easy answers. You'll find charts that give you easy answers, but they are junk. In fact, the indomitable Jacob Cohen, the originator of the most cited chart, says himself that they are junk.

The correlation between the SAT and first-year college GPA is 0.35 for students who attend college. http://professionals.collegeboard.com/profdownload/Validity_of_the_SAT_for_Predicting_First_Year_College_Grade_Point_Average.pdf

Is 0.35 high? Compare 0.35 to 0.4 above, and consider the life changing importance of being accepting to the right college. NO!

Is 0.35 high? Compare 0.35 to the correlation between high school GPA and first-year college GPA, 0.36. YES!

Cautionary Tale: r = 0.861



http://en.wikipedia.org/wiki/Anscombe%27s_quartet

Interpreting a Correlation Matrix

```
CORRELATIONS
  /VARIABLES=MathAch SCHPOP MathAch_SchMean STratio Public Female
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Note that every variable is perfectly correlated with itself, so ignore the diagonal.

Note $r_{xy} = r_{yx}$, thus the symmetry along the diagonal, so ignore a remaining half.

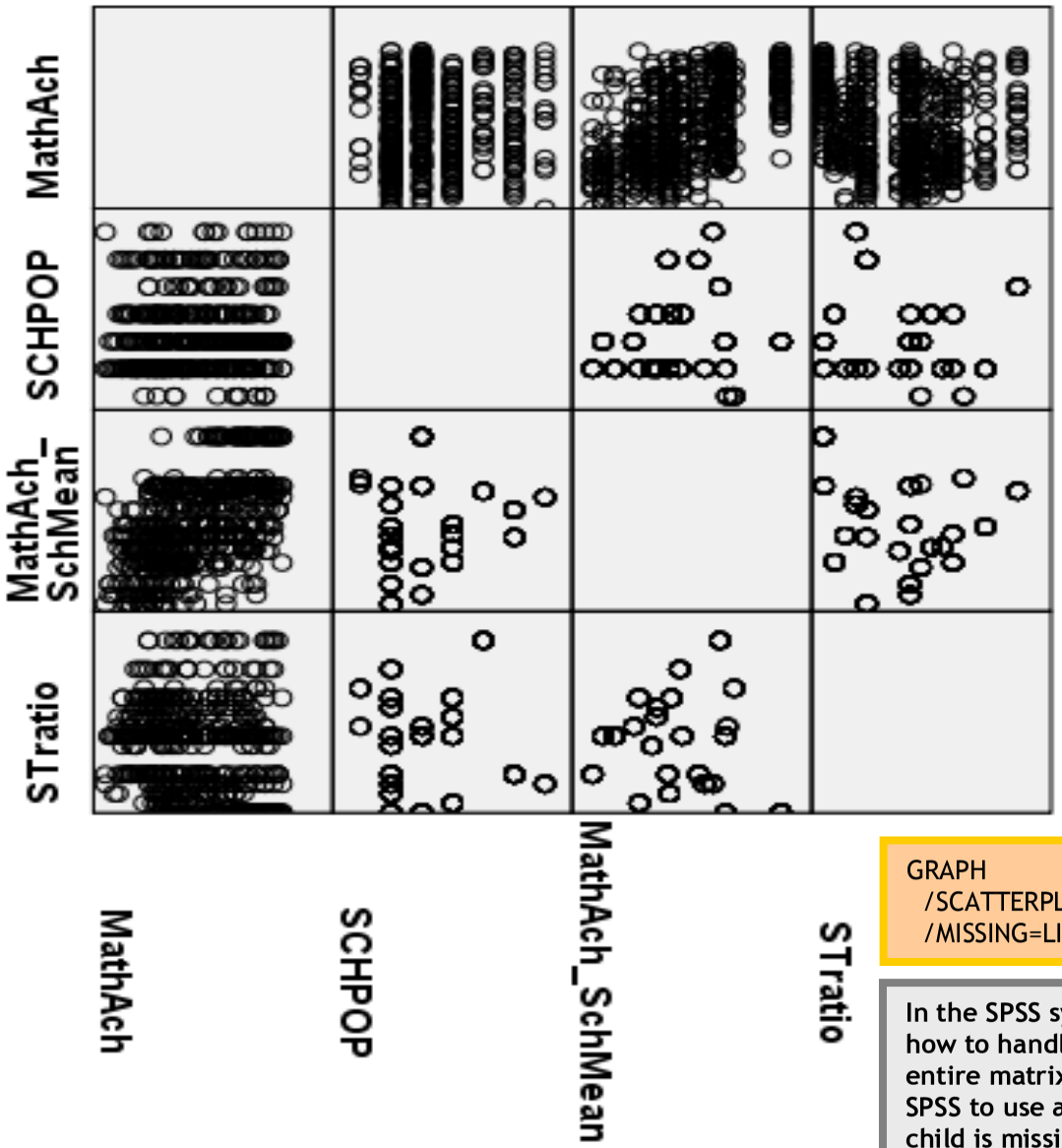
Correlations

		Math Achievement Score	School Population +/- 100 Student	School Average Math Achievement Score	Student/Teacher Ratio	Public School = 1, Private School = 0	Female = 1, Male = 0
Math Achievement Score	Pearson Correlation	1.000	.075	.567**	-.186**	-.356**	-.018
	Sig. (2-tailed)		.090	.000	.000	.000	.682
	N	519.000	519	519	519	519	519
School Population +/- 100 Student	Pearson Correlation	.075	1.000	.132**	-.047	.230**	.066
	Sig. (2-tailed)	.090		.003	.282	.000	.131
	N	519	519.000	519	519	519	519
School Average Math Achievement Score	Pearson Correlation	.567**	.132**	1.000	-.328**	-.628**	.002
	Sig. (2-tailed)	.000	.003		.000	.000	.959
	N	519	519	519.000	519	519	519
Student/Teacher Ratio	Pearson Correlation	-.186**	-.047	-.328**	1.000	.059	-.074
	Sig. (2-tailed)	.000	.282	.000		.177	.092
	N	519	519	519	519.000	519	519
Public School = 1, Private School = 0	Pearson Correlation	-.356**	.230**	-.628**	.059	1.000	.032
	Sig. (2-tailed)	.000	.000	.000	.177		.462
	N	519	519	519	519	519.000	519
Female = 1, Male = 0	Pearson Correlation	-.018	.066	.002	-.074	.032	1.000
	Sig. (2-tailed)	.682	.131	.959	.092	.462	
	N	519	519	519	519	519	519.000

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

Don't worry about statistical significance until Unit 6. And, the sample size for each cell could be interesting, but it's not here.

The Scatterplot Matrix



Before we go ga ga over any numbers, let's be sure to check the assumptions that underlie those numbers. Scatterplot matrices give us a quick and dirty picture of our correlation matrices.

Do you see any linearity problems?

Do you seen any outlier problems?

You may want to zoom in by creating a bivariate scatterplot.

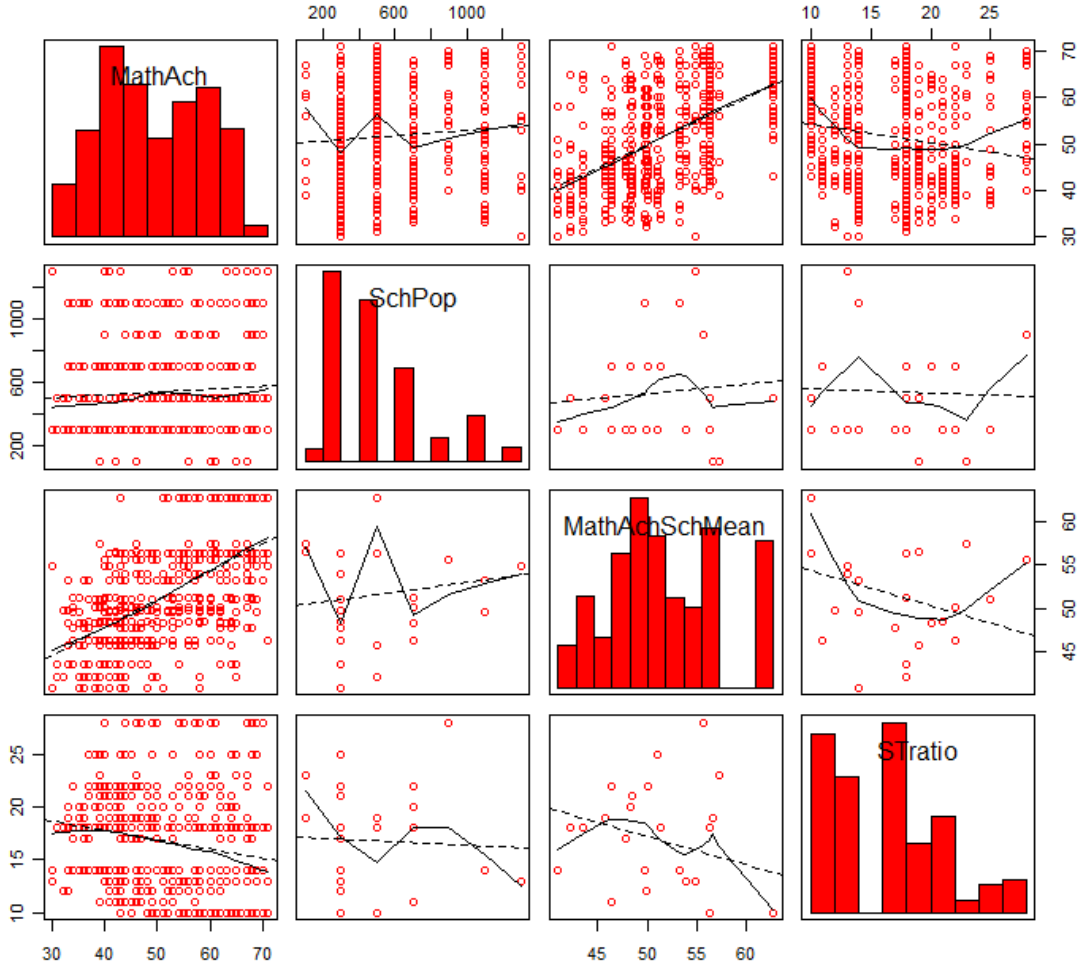
Note that SPSS does not want us to include dichotomous variables in our matrices. We can circumvent this by changing their scale classification from "nominal" to "scale." We will not have a linearity problem with dichotomies, because we are just connecting the means with our regression line, but we could have an outlier problem.

```
GRAPH
  /SCATTERPLOT(MATRIX)=MathAch SCHPOP MathAch_SchMean STratio
  /MISSING=LISTWISE.
```

In the SPSS syntax, you may notice the "MISSING" code. This tells SPSS how to handle missing data. "LISTWISE" tells SPSS to exclude from the entire matrix any student who is missing any data. "PAIRWISE" tells SPSS to use as much data as possible for each scatterplot, so that if a child is missing data on MATHACH, she can still be in a scatterplot of STratio vs. SCHPOP.

Interpreting Correlation and Scatterplot Matrices (R Output)

	MathAch	SchPop	SchMean	STratio	Public	Female
MathAch	1.00	0.07	0.57	-0.19	-0.36	-0.02
SchPop	0.07	1.00	0.13	-0.05	0.23	0.07
MathAchSchMean	0.57	0.13	1.00	-0.33	-0.63	0.00
STratio	-0.19	-0.05	-0.33	1.00	0.06	-0.07
Public	-0.36	0.23	-0.63	0.06	1.00	0.03
Female	-0.02	0.07	0.00	-0.07	0.03	1.00




```
attach(SMD.data)
{for.my.matrices <-
     data.frame(MathAch, SchPop,
     SchMean, StRatio, Public,
     Female)}
detach(SMD.data)
x <- cor(for.my.matrices)
round(x, digits=2)
plot(for.my.matrices)
```

Rcmdr produces pretty fancy scatterplot matrices as a default. It includes regression lines in addition to LOESS lines. (LOESS lines are just "moving averages," to help in exploratory work.) Also, Rcmdr fills in the blank boxes with univariate information. The default is a kernal density plot, a "smoothed out histogram" or "soft-eyed histogram," which like the LOESS line can help in exploratory work. I chose to switch the option from kernal density plots to good old histograms.

When the regression line and LOESS line overlap, that provides support for the linearity assumption. Don't read too much, however, into bouncy LOESS lines. LOESS lines have "soft eyes," but not always soft enough!

Interpreting a Correlation Matrix.

Correlations

		Math Achievement Score	School Population +/- 100 Student	School Average Math Achievement Score	Student/Teacher Ratio
Math Achievement Score	Pearson Correlation	1.000	.075	.567**	-.186**
	Sig. (2-tailed)		.090	.000	.000
	N	519.000	519	519	519
School Population +/- 100 Student	Pearson Correlation	.075	1.000	.132**	-.047
	Sig. (2-tailed)	.090		.003	.282
	N	519	519.000	519	519
School Average Math Achievement Score	Pearson Correlation	.567**	.132**	1.000	-.328**
	Sig. (2-tailed)	.000	.003		.000
	N	519	519	519.000	519
Student/Teacher Ratio	Pearson Correlation	-.186**	-.047	-.328**	1.000
	Sig. (2-tailed)	.000	.282	.000	
	N	519	519	519	519.000

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

In our sample, a school's average math score is a strong predictor of individual math scores (r = 0.57), whereas school population size (r = 0.08) and student/teacher ratio (r = -0.19) are relatively weak predictors of individual math scores. The correlation between school-average and individual math scores is artificially inflated, because the individual contributes her score to her school's average, and an individual's score is perfectly correlated with itself. We cannot tell the extent of "part-whole problem" unless we know whether the school averages were calculated from a small sample or the entire school. The more students who contributed to the averages, the less the bias.

Among the predictors, the strongest relationship is between student-teacher ratio and school average math achievement score (r = -0.33). This is not surprising: smaller class sizes are associated with higher scores. Although a causal conclusion is tempting here, it is worth a reminder that correlation does not imply causation. Small class sizes may be a proxy for SES, for example, where wealthier school districts can afford more teachers.

Discussing Pearson Product-Moment Correlations

When discussing Pearson product-moment correlations, make sure, first and foremost, that the relationship is linear and that it is not determined by outliers.

(i) Discuss the correlations between each predictor and your outcome. Note the implications for your theories (hunches). Tell a story. Group your correlations into "surprising" and "unsurprising," and note why. Perhaps group your correlations into "positive," "negative" and "virtually none."

(ii) Discuss the correlations among your predictors. Use the same techniques as above, but perhaps more briefly: don't steal the show from your research questions.

Never lose sight of the substantive meaning of the numbers.

You will find that you have all you need for the Unit 4 Post Hole. You can find examples at the end of the slides.

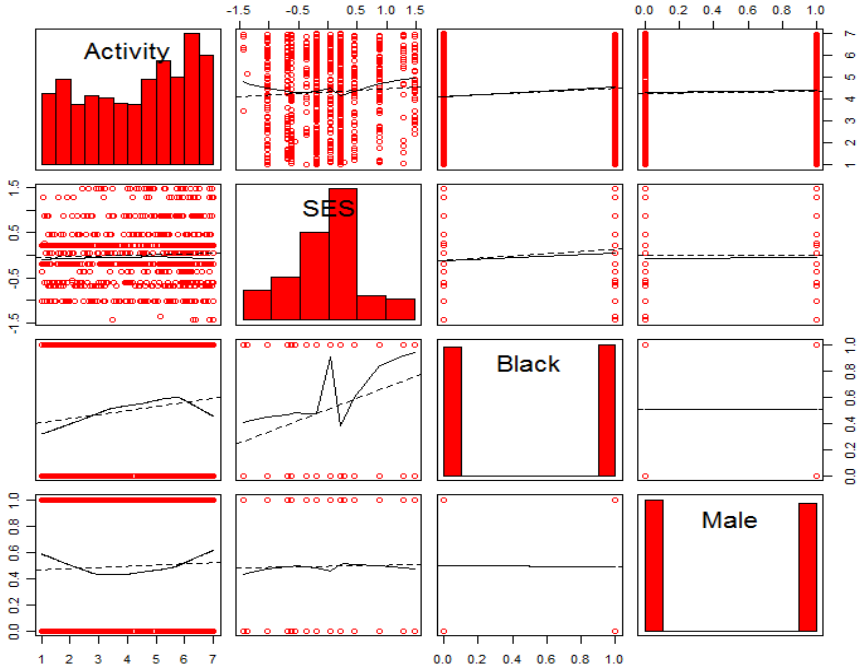
Dig the Post Hole

Unit 4 Post Hole:

Interpret a correlation matrix.

Evidentiary materials: correlation and scatterplot matrices.

	Activity	SES	Black	Male
Activity	1.00	0.05	0.11	0.03
SES	0.05	1.00	0.21	0.01
Black	0.11	0.21	1.00	0.00
Male	0.03	0.01	0.00	1.00



Brief description of the variables:
Activity, a measure of teenage sexual activity on a scale of 1-7.
SES, a measure of SES on a scale of -1.5 to 1.5.
Black, an indicator for race, where 1 = Black, 0 = Else.
Male, an indicator for sex, where 1 = Male, 0 = Female
Data from: Brown, J.D. et al. "Sexy Media Matter: Exposure to Sexual Content in Music, Movies, Television, and Magazines Predicts Black and White Adolescents' Sexual Behavior" PEDIATRICS Vol. 117 No. 4 April 2006, pp. 1018-1027.

Here is my try:

None of the three predictors has a strong relationship with the outcome, sexual activity. Of the three predictors, *Black* has the strongest, but nonetheless weak, relationship with *Activity* (r = 0.11).

As to be expected, *Male* has no relationship with *SES* or *Black*. Also, as to be expected, given the known inequalities in our society, there is a moderate relationship between *Black* and *SES* (r = 0.21).

An inspection of the scatterplot matrix reveals no linearity or outler problems.

Here is my minimally acceptable try:

None of the three predictors has a strong relationship with the outcome, sexual activity, with correlations ranging from 0.03 to 0.11. Perhaps the only meaningful relationship among the variables is between the predictors *Black* and *SES* (r=.0.21). There appears to be no linearity or outlier problems.

Dig the Post Hole

Unit 4 Post Hole:

Interpret a correlation matrix.

Evidentiary materials: correlation and scatterplot matrices.

Correlations

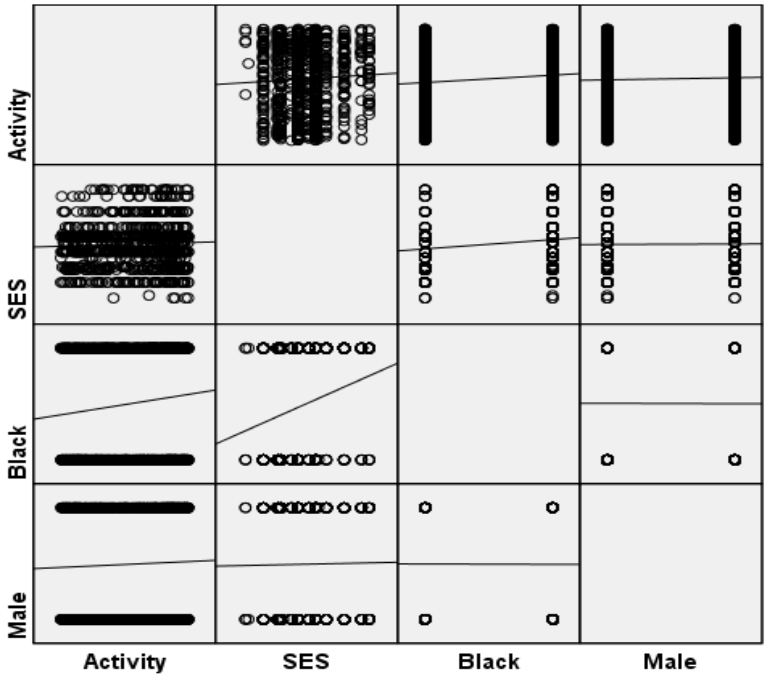
		Activity	SES	Black	Male
Activity	Pearson Correlation	1.000	.048	.110**	.031
	Sig. (2-tailed)		.152	.001	.351
	N	887.000	887	887	887
SES	Pearson Correlation	.048	1.000	.205**	.010
	Sig. (2-tailed)	.152		.000	.770
	N	887	887.000	887	887
Black	Pearson Correlation	.110**	.205**	1.000	-.003
	Sig. (2-tailed)	.001	.000		.922
	N	887	887	887.000	887
Male	Pearson Correlation	.031	.010	-.003	1.000
	Sig. (2-tailed)	.351	.770	.922	
	N	887	887	887	887.000

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

Here is my minimally acceptable try:

None of the three predictors has a strong relationship with the outcome, sexual activity, with correlations ranging from 0.03 to 0.11. Perhaps the only meaningful relationship among the variables is between the predictors *Black* and *SES* (r=.0.21). There appears to be no linearity or outlier problems.

Brief description of the variables:
Activity, a measure of teenage sexual activity on a scale of 1-7.
SES, a measure of SES on a scale of -1.5 to 1.5.
Black, an indicator for race, where 1 = Black, 0 = Else.
Male, an indicator for sex, where 1 = Male, 0 = Female
Data from: Brown, J.D. et al. "Sexy Media Matter: Exposure to Sexual Content in Music, Movies, Television, and Magazines Predicts Black and White Adolescents' Sexual Behavior" PEDIATRICS Vol. 117 No. 4 April 2006, pp. 1018-1027.



Correlation Tables for Publication

Table 1

Pearson Correlations Between Sexual Activity and Demographic Variables (n = 887)

	Sexual Activity	SES	Black
SES	.048		
Black	.110**	.205**	
Male	.031	.010	-.003

Note. **p < 0.01

Some people care deeply about the details of table style or format. I am not one of those people. I do value the goals of maximizing information and minimizing clutter. Everything else is just dandy. Here, I present an APA style correlation matrix, although I bet an APA guru could find fault.

http://psychology.about.com/od/apastyle/ig/APA-Format-Examples/apa-table.htm

APA style tables never include vertical lines. This is a legacy of old type-setting restrictions, where vertical lines were hard to produce. Go figure.

For your tech memos, don't worry about the details. For journal submissions, look through recently published articles within the journal; find a table that presents the same statistics as your table; COPY the table format! For your own self-publications, use your best judgment: maximize info, minimize clutter.

Table 1

Pearson Correlations Between Sexual Activity and Demographic Variables (n = 887)

	Sexual Activity	SES	Black
SES	.048		
Black	.110**	.205**	
Male	.031	.010	-.003

Note. **p < 0.01

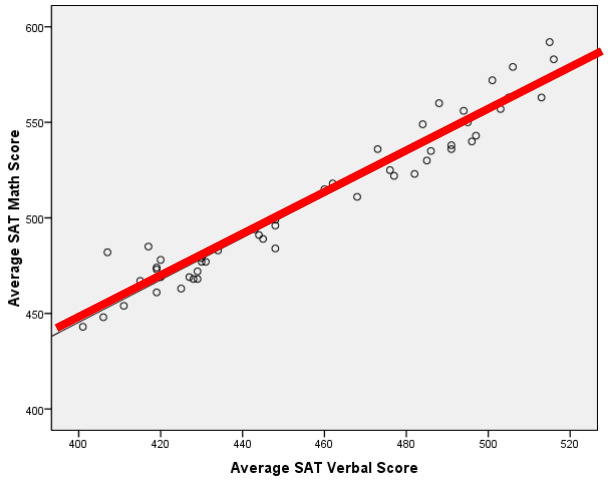
Include a fully descriptive title or caption. Include the sample size if the format permits.

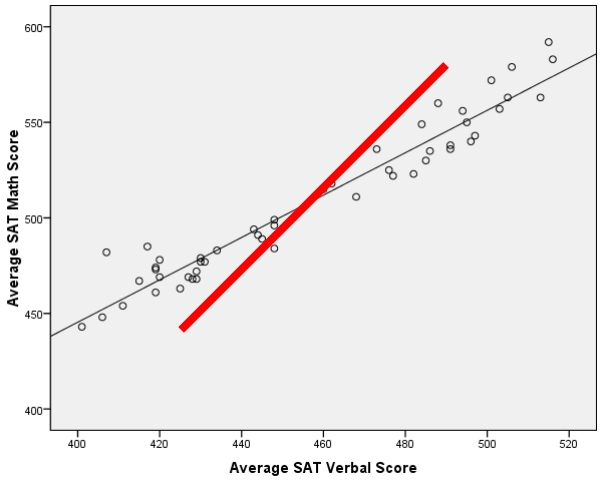
Exclude clutter.

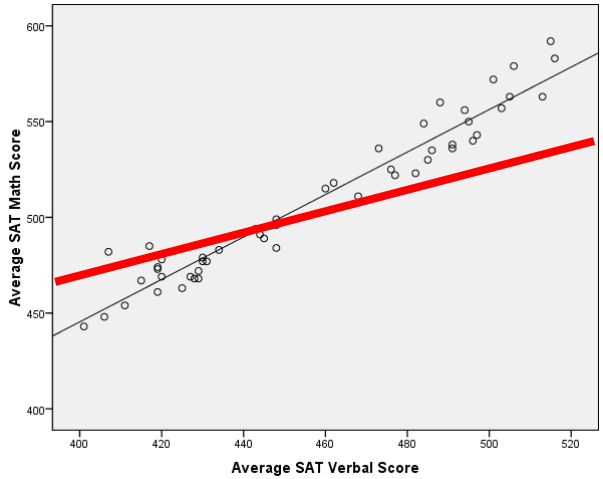
- Exclude redundant correlations.
- Exclude self-correlations (*r* =1.00).
- Exclude senseless decimal precision.
- Exclude distracting lines.

Do NOT worry about p-values and statistical significance, yet. Wait until Unit 6.

We Can Alter the Slope By Changing the Axes/Scales







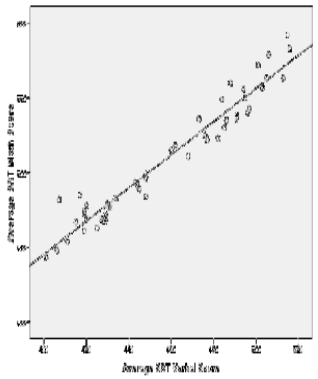
Pearson Correlations Provide Standardized Comparisons

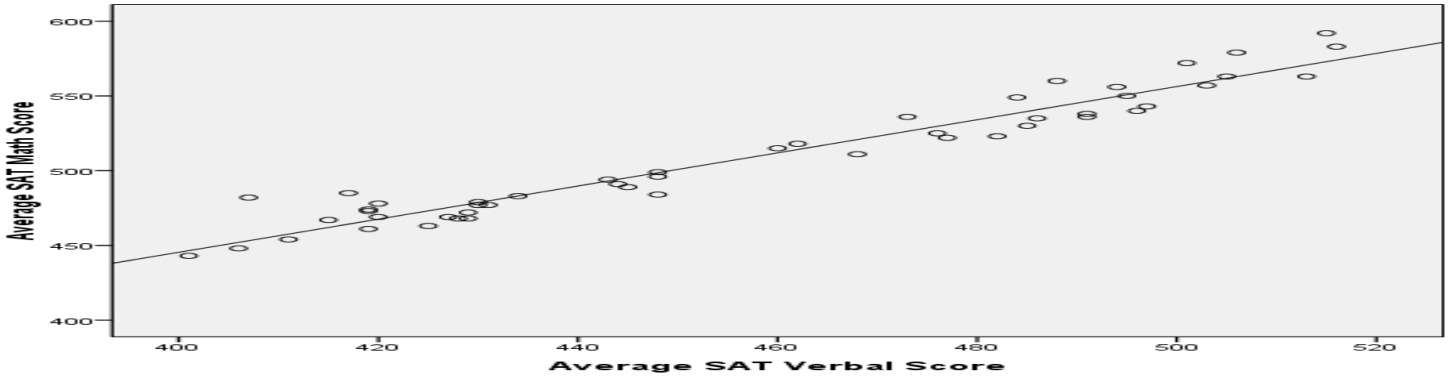
Graphically, we can change the apparent slope by expanding or contracting the width of the x-axis or the height of the y-axis.

Numerically, we can change the slope coefficient by rescaling our outcome or predictor. For example, if our predictor is *AGE* (in months), we can increase the slope coefficient by measuring *AGE* in years, or we can decrease the slope by measuring *AGE* in days.

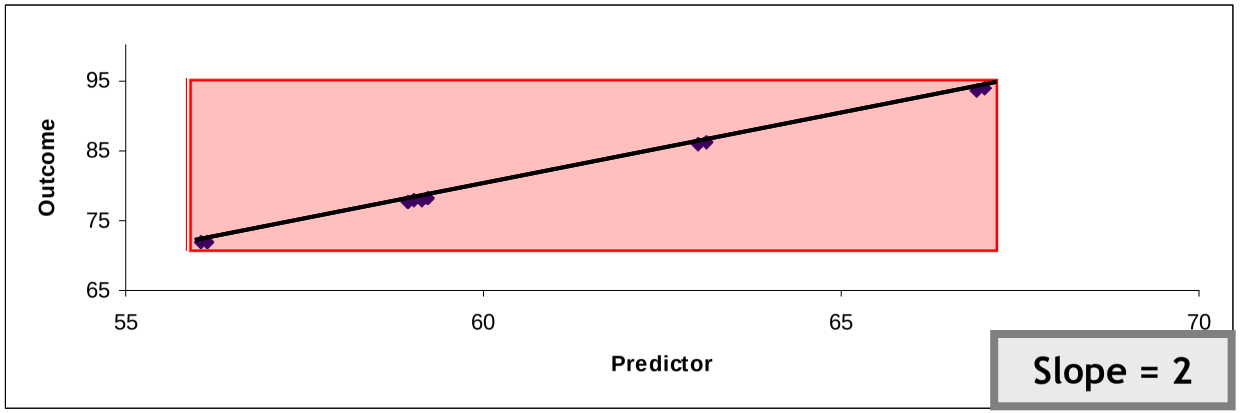
The great thing about magnitude is that we can compare apples to oranges. For example, we can compare *AGE* to *READING*. However, we must know whether *AGE* is measured in days, months or years, and we must likewise be familiar with the *READING* scale. IF we do understand our scales, then magnitude is excellent for making policy decisions.

By z-transforming, we (sort of) put the outcome and predictor on the same scale. Instead of comparing *AGE* to *READING*, or instead of comparing verbal SAT scores to math SAT scores, or instead of comparing number of students in a school to math achievement scores, we are comparing standard deviations to standard deviations. Graphically, this is analogous to making sure that the x-axis and y-axis are the same length; thus, we "square-frame" the scatterplot so that there is no monkey business as per the previous slide. (Rarely, if ever, do we insist on square-framed scatterplots, but often we insist that similar scatterplots be presented similarly within the same presentation.)



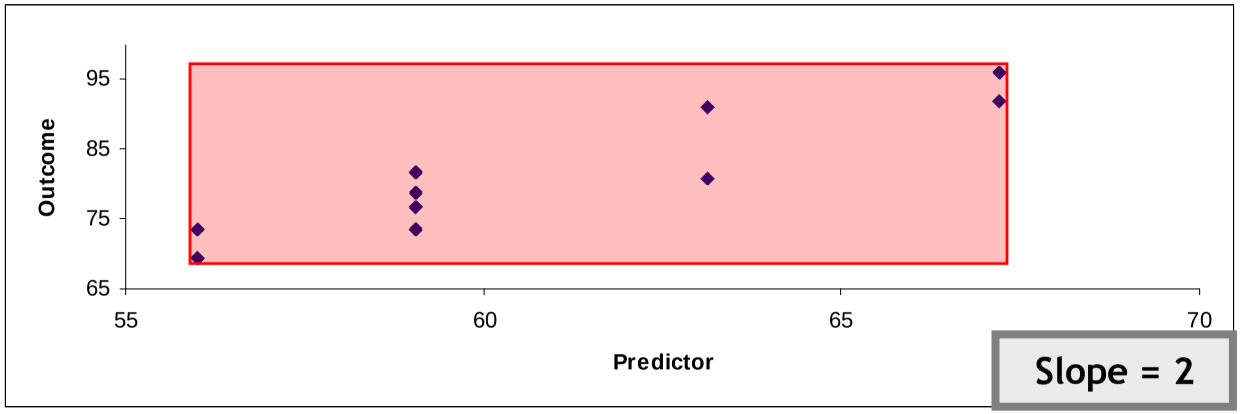


By Standardizing the Axes/Scales, We Can Quantify Strength



Slope = 2

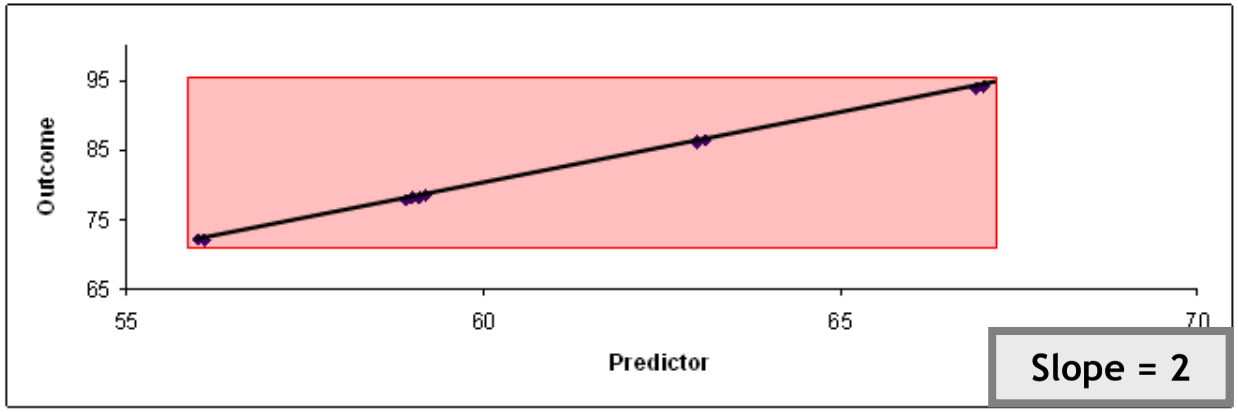
Notice that, when we "frame" the *perfect* relationship, the regression line goes from corner to corner.



Slope = 2

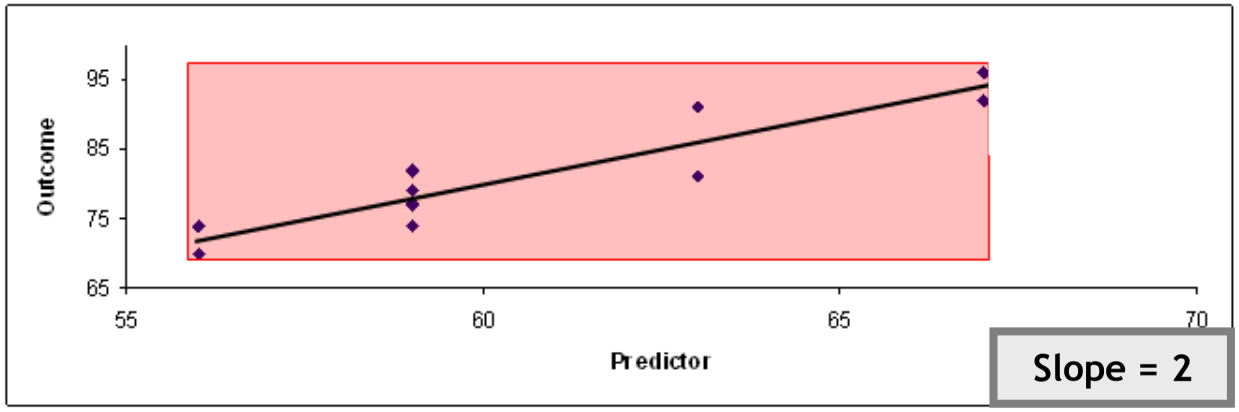
Notice that, when we "frame" the *imperfec*t relationship, the regression line falls short of the corners. The more imperfect the relationship, the more the regression line will fall short. The regression lines are the same, but the boxes are different (in order to include every data point). Our imperfect relationship needs a taller box because it has a greater vertical spread. The widths are identical here because the horizontal spreads happen to be identical.

By Standardizing the Axes/Scales, We Can Quantify Strength



Slope = 2

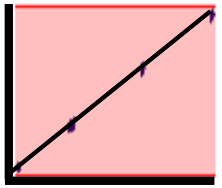
Notice that, when we "frame" the *perfect* relationship, the regression line goes from corner to corner.



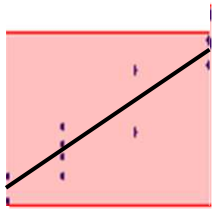
Slope = 2

Notice that, when we "frame" the *imperfect* relationship, the regression line falls short of the corners. The more imperfect the relationship, the more the regression line will fall short. The regression lines are the same, but the boxes are different (in order to include every data point). Our imperfect relationship needs a taller box because it has a greater vertical spread. The widths are identical here because the horizontal spreads happen to be identical.

$r = 1.00$



"Square-framing" makes the axes equal and the slopes comparable for the purposes of judging strength.



$r = 0.93$

A Pearson correlation, *r*, is a standardized slope. The trade-off is that, by standardizing, we lose the original meaning of the scales and, consequently, the magnitude.

Numerically, We "Square-Frame" By Standardizing

Perfect Correlation (r = 1.00)			
Raw Predictor	Raw Outcome	Z Predictor	Z Outcome
56	72		
56	72		
59	78		
59	78		
59	78		
59	78		
63	86		
63	86		
67	94		
67	94		
Mean = 60.8	Mean = 81.6		
S.D. = 4.0	S.D. = 8.0		

$$\hat{RawOutcome} = -40 + 2(RawPredictor)$$

$$\hat{ZOutcome} = 0.00 + 1.00(ZPredictor)$$

Imperfect (But Strong) Correlation (r = 0.93)			
Raw Predictor	Raw Outcome	Z Predictor	Z Outcome
56	71		
56	73		
59	77		
59	79		
59	74		
59	82		
63	81		
63	91		
67	95		
67	93		
Mean = 60.8	Mean = 81.6		
S.D. = 4.0	S.D. = 8.6		

$$\hat{RawOutcome} = -40 + 2(RawPredictor)$$

$$\hat{ZOutcome} = 0.00 + 0.93(ZPredictor)$$

Calculating Correlations by Hand is Easy! (And Fun!)

A Pearson product-moment correlation (i.e., *r* statistic) is the average product of the z-scores.

1. Take your first observation, multiply the z-transformed outcome by the z-transformed predictor.
2. Do this for every observation, and you get a product of z-scores for each observation.
3. Add up all the products of z-scores.
4. Divide the sum by the degrees of freedom (n-1) to get an average product of z-scores, *r*.

Imperfect (But Strong) Correlation (*r* = 0.93)

Raw Predictor	Raw Outcome	Z Predictor	Z Outcome	ZPredictor *x* ZOutcome
56	71	-1.1934	-1.2282	1.4657
56	73	-1.1934	-0.9964	1.1891
59	77	-0.4475	-0.5330	0.2385
59	79	-0.4475	-0.3013	0.1348
59	74	-0.4475	-0.8806	0.3941
59	82	-0.4475	0.0463	-0.0207
63	81	0.5470	-0.0695	-0.0380
63	91	0.5470	1.0891	0.5957
67	95	1.5415	1.5526	2.3933
67	93	1.5415	1.3209	2.0361
Mean = 60.8	Mean = 81.6	Mean = 0	Mean = 0	Sum = 8.3885
S.D. = 4.0	S.D. = 8.6	S.D. = 1	S.D. = 1	Sum/(n-1) = 0.9321

Positive Relationships:
Since a negative times a negative is a positive, almost all our products are positive. Here, negative ZPredictors tend to go with negative ZOutcomes, and positive ZPredictors tend to go with positive ZOutcomes. This tendency is strong so the sum is very large. Only twice do we have a negative product, and those products happen to be small. This is a very strong positive relationship.

Negative Relationships:
In a negative relationship, negative ZPredictors tend to go with positive ZOutcomes, and positive ZPredictors tend to go with negative ZOutcomes, so the products tend to be negative, and when this tendency is strong, the sum of products will be a very large negative number, signifying a negative correlation.

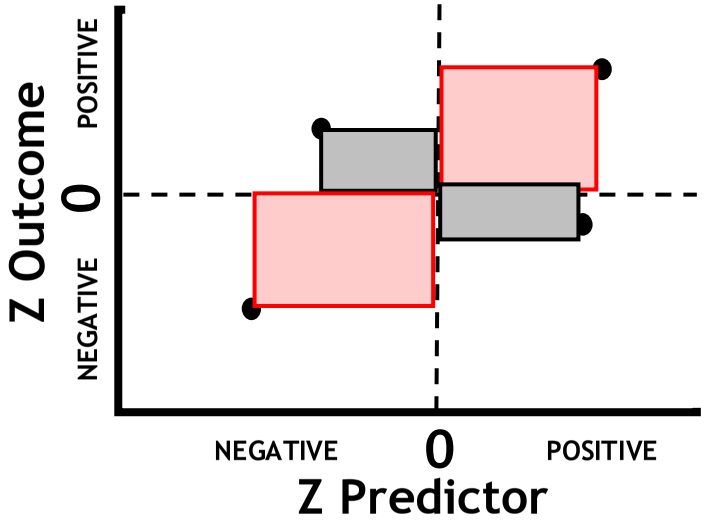
Zero Relationships:
When there is no relationship, negative ZPredictors are just as likely to go with negative ZOutcomes as positive ZOutcomes. Therefore, the products are a mix of positives and negatives that cancel each other out and sum to zero, signifying zero correlation.

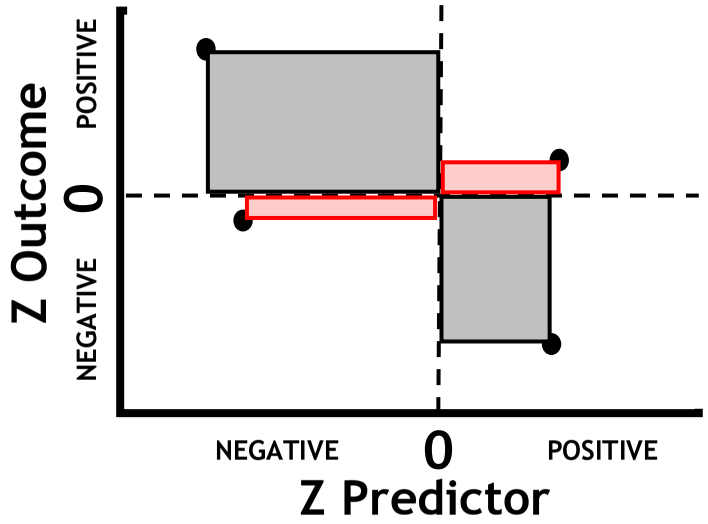
Notice that it doesn't matter which order we multiply. 2x3 = 3x2. This is why correlations are symmetric. From the math perspective, it does not matter which variable is our predictor and which is our outcome.

$$r_{x,y} = \frac{\sum_{i=1}^{n} zx_i \cdot zy_i}{n-1}$$

Isn't it comforting to know that, if the zombies were to attack, and we were engulfed in a post-apocalyptic dark age, you could still calculate correlations, without a computer? Okay, maybe not so comforting.

(Yet) Another Way to Think About Correlations

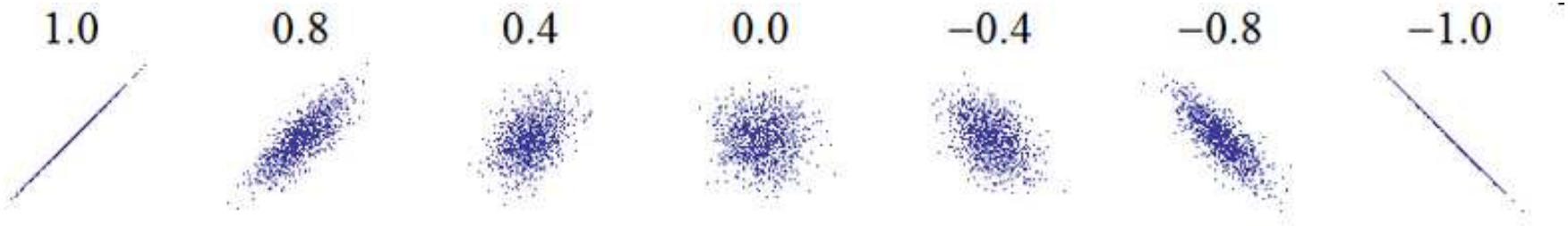


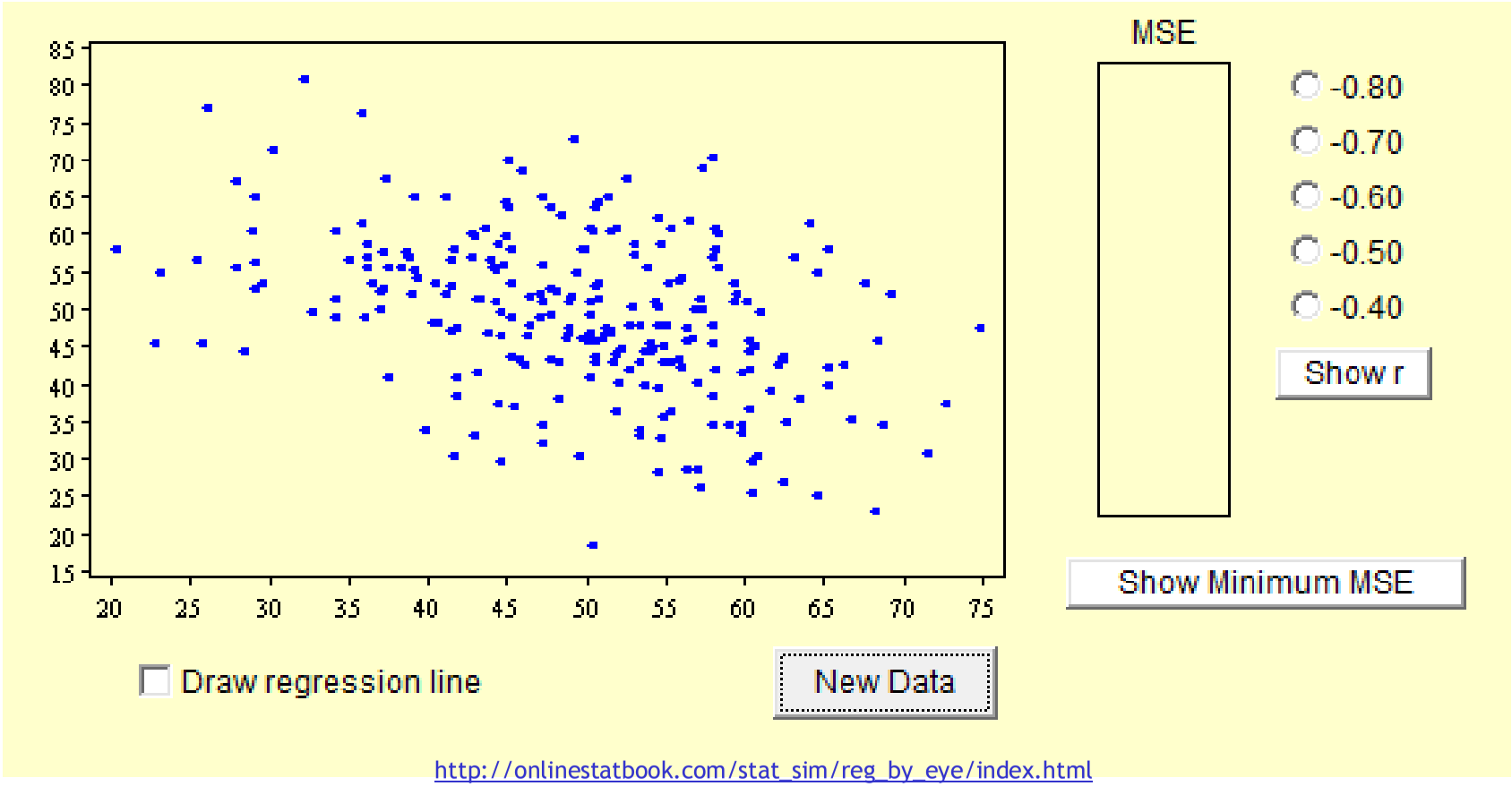


1. Divide the scatterplot into four quadrants.
   I. Quadrant I: Above Average Outcome, Above Average Predictor
   II. Quadrant II: Above Average Outcome, Below Average Predictor
   III. Quadrant III: Below Average Outcome, Below Average Predictor
   IV. Quadrant IV: Below Average Outcome, Above Average Predictor
2. Multiply the ZOutcome by the ZPredictor to get rectangles that can be positive or negative.
   I. Quadrant I: Positive Rectangles
   II. Quadrant II: Negative Rectangles
   III. Quadrant III: Positive Rectangles
   IV. Quadrant IV: Negative Rectangles
3. Sum the positive rectangles.
4. Sum the negative rectangles.
5. Take the difference of the sums.
   A. If the area of the positive rectangles is bigger than the area of the negative rectangles, then there is a positive correlation.
   B. If the area of the negative rectangles is bigger than the area of the positive rectangles, then there is a negative correlation.
   C. If the area of the positive rectangles is equal to the area of the negative rectangles, then there is zero correlation.
6. Divide the difference by the degrees of freedom (n-1) to get the correlation coefficient, *r*.

This is not how I think of correlations. There are many mathematically equivalent ways to think about correlations, but my way is the best. It's best to think of correlations as slope coefficients from the linear regression of a standardized outcome on a standardized predictor. Why is my way the best? My way is the best because the linearity and outlier assumptions are obvious. Just as I would never fit a linear regression model without conducting a bivariate exploratory analysis, I would not calculate a correlation coefficient without conducting a bivariate exploratory data analysis, because I know that a regression coefficient and a correlation coefficient are essentially the same thing.

SURVIVE-*r*





http://onlinestatbook.com/stat_sim/reg_by_eye/index.html

http://istics.net/stat/Correlations/ **Group ID: edstats**

Answering our Roadmap Question

Unit 4: In our sample, how strong is the relationship between reading achievement and free lunch?

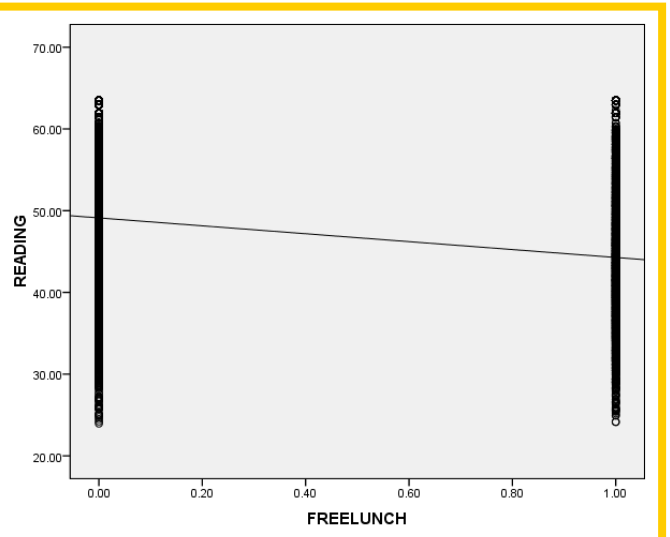
Coefficients[a]

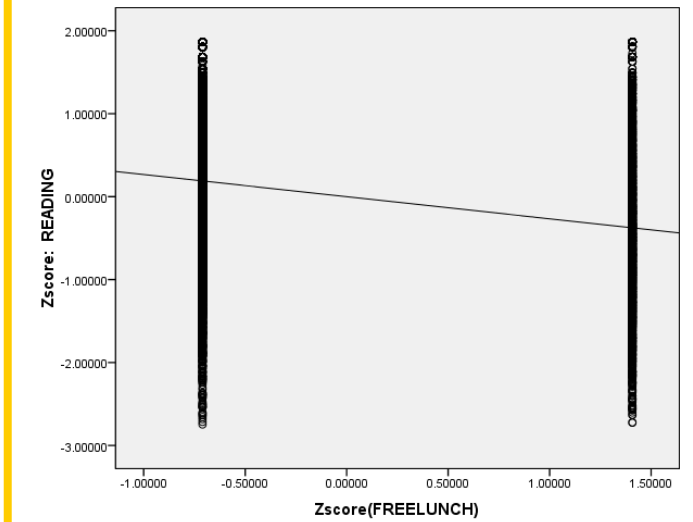
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	49.118	.115		428.169	.000
	FREELUNCH	-4.841	.198	-.267	-24.439	.000

a. Dependent Variable: READING

$$Reading = \beta_0 + \beta_1 FreeLunch + \varepsilon$$

$$ZReading = \beta_0 + \beta_1 ZFreeLunch + \varepsilon$$





Answering our Roadmap Question

Unit 4: In our sample, how strong is the relationship between reading achievement and free lunch?

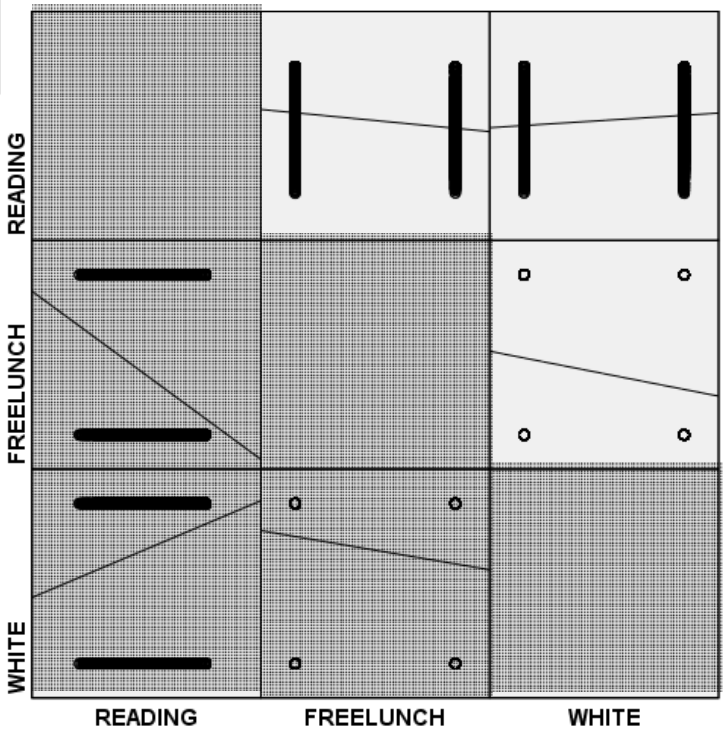
Correlations

		READING	FREELUNCH	WHITE
READING	Pearson Correlation	1.000	-.267**	.165**
	Sig. (2-tailed)		.000	.000
	N	7800.000	7800	7800
FREELUNCH	Pearson Correlation	-.267**	1.000	-.183**
	Sig. (2-tailed)	.000		.000
	N	7800	7800.000	7800
WHITE	Pearson Correlation	.165**	-.183**	1.000
	Sig. (2-tailed)	.000	.000	
	N	7800	7800	7800.000

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

There is a lot of redundant information in correlation matrices, so I gray out the redundancy. Usually, I gray out the diagonal and the UPPER half of the matrix, but here I grayed out the diagonal and the LOWER part of the matrix, only because I'm more comfortable with the UPPER half of the scatterplot matrix, even though the LOWER half shows the same thing.

In our sample, there is a negative correlation between *FREELUNCH* and *READING* such that students eligible for free lunch tend to perform worse on the reading test than their ineligible counterparts (r = -.27). There is a positive correlation between WHITE and READING such that White students tend to perform better on the reading test than their non-White counterparts (r = .17). *FREELUNCH* and *WHITE* are also correlated (r = -.18) which raises the question of confounding variables. It is unclear how much of the relationship between race and reading scores is really attributable to SES, since race and SES are correlated with each other and with reading scores. Likewise, it is unclear how much of the relationship between SES and reading scores is attributable to race. Of course, there may be a fourth variable (unobserved, perhaps unobservable) that is the root cause.



Unit 4 Appendix: Key Concepts

- Causal conclusions require 3 conditions:
  - Correlation
  - Succession
  - Necessary Connexion
- In addition to your Predictor and Outcome, consider the possible influence of a 3rd Hidden Confounding Variable.
- When presenting your pet causal conclusion, present 2 other plausible causal conclusions for the sake of balance.
- Causality can be very complicated. It often involves far more than two (or three) variables. A causal relationship between two variables can be mediated by hidden variables and moderated by other hidden variables. Meanwhile, the causal arrows are pointing every which way.
- In converting the raw slope into the standardized slope to get the *r* statistic, we trade a numerical description of magnitude for a numerical description of strength.
- You want to know which values of the *r* statistic are big and which are small, but there are no easy answers. Use thoughtful comparisons instead of rules of thumb. The only good rule of thumb is that all rules of thumb will eventually fail you.
- Correlation does not imply causation. Whenever you think about a correlation, consider which is the cause and which is effect. Are they reciprocally causal? What may be a third (hidden) variable at play?
- Check your assumptions before reporting any number. In the case of the *r* statistic, check for linearity and outliers.

Unit 4 Appendix: Key Interpretations

In our sample, a school's average math score is a strong predictor of the individual math scores (r = 0.57), whereas school population size (r = 0.08) and student/teacher ratio (r = -0.19) are weak predictors.

In our sample, students from private schools tend to perform better in math than students from public schools (r = 0.36).

Student teacher ratio is negatively correlated with math achievement (r = -0.19) such that, given two students who attend schools with student/teacher ratios that differ by 5, we expect on average the student who attends the school with the smaller student/teacher ratio to score 2 points higher than the student who attends the school with larger student/teacher ratio. Because student/teacher ratio is also correlated with SES, among other things, we can draw no causal relationship."

In our sample, there is a negative correlation between *FREELUNCH* and *READING* such that students eligible for free lunch tend to perform worse on the reading test than their ineligible counterparts (r = -.27). There is a positive correlation between WHITE and READING such that White students tend to perform better on the reading test than their non-White counterparts (r = .17). *FREELUNCH* and *WHITE* are also correlated (r = -.18) which raises the question of confounding variables. It is unclear how much of the relationship between race and reading scores is really attributable to SES, since race and SES are correlated with each other and with reading scores. Likewise, it is unclear how much of the relationship between SES and reading scores is attributable to race. Of course, there may be a fourth variable (unobserved, perhaps unobservable) that is the root cause.

Unit 4 Appendix: Key Terminology

The *r* statistic (or Pearson product-moment correlation) is a standard measure of the strength of a bivariate relationship. It can take values from -1.0 to 1.0, where negative values denote negative relationships and positive values denote positive relationships.

Unit 4 Appendix: Math

$$\text{variance}_x = \frac{\sum (x-\bar{x})(x-\bar{x})}{n-1} = \text{covariance}_{x,x}$$

$$\text{covariance}_{x,y} = \frac{\sum (x-\bar{x})(y-\bar{y})}{n-1}$$

$$\text{standard deviation}_x = \text{sd}_x = \sqrt{\text{variance}_x} = \sqrt{\frac{\sum (x-\bar{x})(x-\bar{x})}{n-1}}$$

$$r_{x,y} = \frac{\text{covariance}_{x,y}}{\text{sd}_x * \text{sd}_y} = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y}$$

IF: $ZOUTCOME = \beta_0 + \beta_1 ZPREDICTOR + \varepsilon$

THEN: $r_{ZPREDICTOR,ZOUTCOME} = r_{PREDICTOR,OUTCOME} = r_{x,y} = \beta_1$

Unit 4 Appendix: Math (draft)

Covariance is about as interesting as variance—not very. However, like variance, covariance is good for getting interesting things. Variance gets us standard deviations, for example, and covariance gets us Pearson correlations, for example.

Geometrically, when we plot X vs. X, and we take mean deviations, we get positive squares (assuming the axes are on the same scales). When we plot Y vs. X and we take mean deviations, we get positive and negative rectangles which can sum to zero indicating no relationship, but they can also sum to positive values, indicating positive relationships, or they can sum to negative values, indicating negative relationships .

Now that we can compute Pearson correlations by hand, we can also calculate by hand slope coefficients for regression equations. Recall that the Pearson correlation is just the slope (β_1) from a regression of a standardized outcome on a standardized predictor. Recall that a slope tells us the difference in the outcome associated with a one unit difference in the predictor. Due to standardization, our units are in standard deviations. Therefore, the Pearson correlation (or standardized slope) tells us the difference (measured in standard deviations) in the outcome associated with a one standard deviation difference in the predictor.

$$\beta_1 = \text{slope} = \frac{\text{rise}}{\text{run}} = r_{x,y}\left(\frac{sd_y}{sd_x}\right) \qquad r_{x,y} = \text{standardized slope} = \frac{\frac{\text{rise}}{sd_y}}{\frac{\text{run}}{sd_x}} = \beta_1\left(\frac{sd_x}{sd_y}\right)$$

Now that we can calculate correlations by hand, we can calculate β_1 by hand. And, now that we can calculate β_1 by hand, we can calculate the entire regression equation by hand. We only need to calculate the y-intercept, β_1. We will use the fact that, with linear models, mean of the predictor always predicts the mean of the outcome.

$$mean_y = \beta_0 + r_{x,y}\left(\frac{sd_y}{sd_x}\right)mean_x \qquad \beta_0 = mean_y - r_{x,y}\left(\frac{sd_y}{sd_x}\right)mean_x$$

Unit 4 Appendix: Math (draft)

$$MathAch = \beta_0 + \beta_1 SchoolPop + \varepsilon$$

One might think that, if we have a "prediction machine" for math achievement based on school population, we could algebraically reverse the machine to get a prediction for school population based on math achievement. NO.

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	50.167	1.029		48.767	.000
	School Population +/- 100 Student	.003	.002	.075	1.700	.090

a. Dependent Variable: Math Achievement Score

Based on our fitted equation, what math achievement score would we predict for a student in our sample who goes to a school with 1000 students?

$$\hat{MathAch} = 50.2 + 0.003 SchoolPop$$

$$\left[\hat{MathAch} \mid SchoolPop = 1000\right] = 50.2 + 0.003(1000) = 53.2$$

Does this mean that, for a student who scores a 53.2 in math achievement, we predict she goes to a school of 1000 students? NO.

$$SchoolPop = \beta_0 + \beta_1 MathAch + \varepsilon$$

Why? Regression to the mean.

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	445.035	60.578		7.346	.000
	Math Achievement Score	1.949	1.147	.075	1.700	.090

a. Dependent Variable: School Population +/- 100 Student

$$\hat{SchoolPop} = 445 + 1.95 MathAch$$

$$\left[\hat{SchoolPop} \mid MathAch = 53.2\right] = 445 + 1.95(53.2) = 548.74$$

Unit 4 Appendix: SPSS Syntax

```
**********************************************************************.
*Correlation matrices make sense syntactically.
*On the other hand, scatterplot matrices demand dropdown menus.
**********************************************************************.

CORRELATIONS
  /VARIABLES=MathAch SCHPOP MathAch_SchMean STratio
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.

GRAPH
  /SCATTERPLOT(MATRIX)=MathAch SCHPOP MathAch_SchMean STratio
  /MISSING=LISTWISE.
```

Unit 4 Appendix: R Syntax

```
# Here is the simplest of syntax.

# To get one correlation at a time:
cor(schmoutcome,schmedictor)

# To get a correlation matrix for your whole data frame:
cor(schmataframe)

# To get a scatterplot matrix for your whole data frame:
plot(schmataframe)

# To create a smaller data frame from a larger data frame,
# first, attach your larger data frame:
attach(schmataframe)

# Second, assign the variables of your choosing from the larger data frame
# to a smaller data frame of your naming:
my.smaller.data.frame <- data.frame(schmariable1, schmariable2, schmariable3, schmariable4)

# Third, detach your larger data frame:
detach(schmataframe)

# Now, you can create exactly the correlation and scatterplot matrices you want:
cor(my.smaller.data.frame)
plot(my.smaller.data.frame)

# In general, the correlation matrix will provide annoyingly many decimal places, so round:
round(cor(my.smaller.data.frame), digits=2)
# In general, the round function goes: round(x, digits=#)
# Since we want to  round our whole correlation matrix, we substitute the code in for x.
```

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)

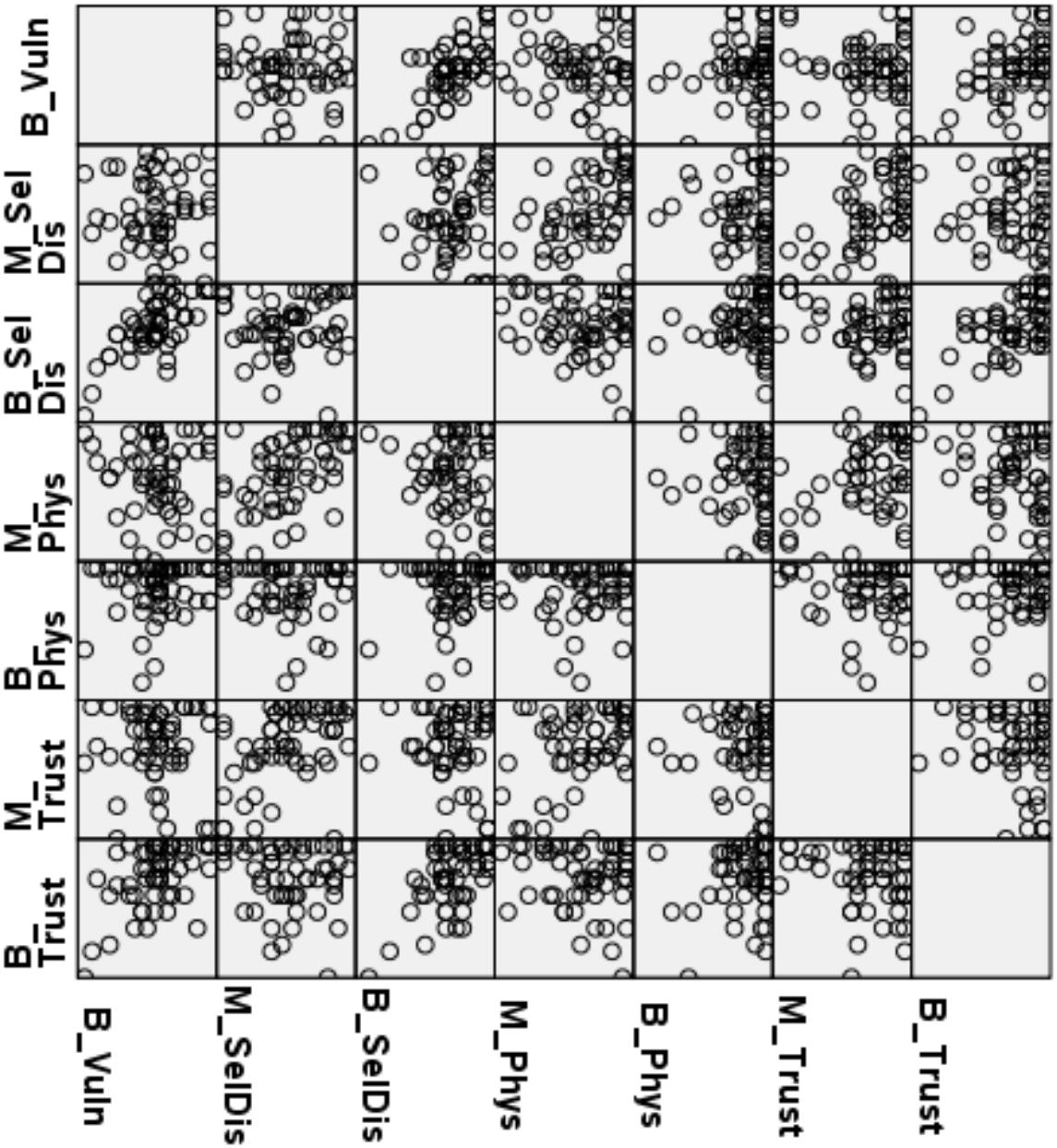
Perceived Intimacy of Adolescent Girls (Intimacy.sav)

Correlations

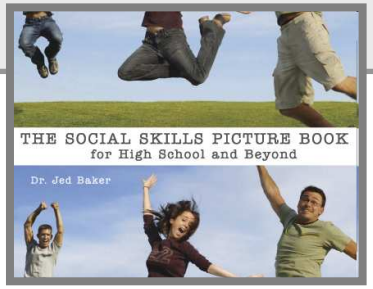
		Risk vulnerability w boyfriend	Self-disclose to mother	Self-disclose to boyfriend	Phys affection w mother	Phys affection w boyfriend	Trust mother	Trust boyfriend
Risk vulnerability w boyfriend	Pearson Correlation	1.000	.002	.731**	-.053	.094	.052	.553**
	Sig. (2-tailed)		.985	.000	.689	.476	.690	.000
	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)



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:**

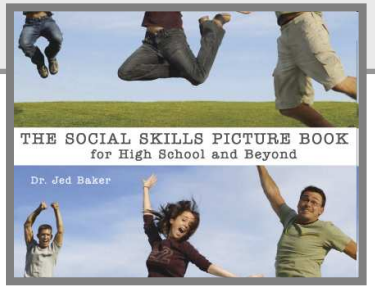
Variables about the student—

(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)



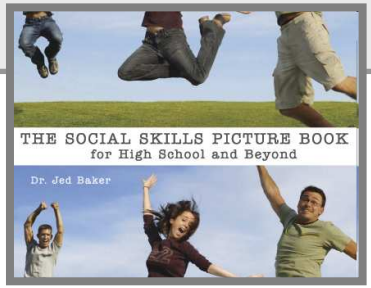
Correlations

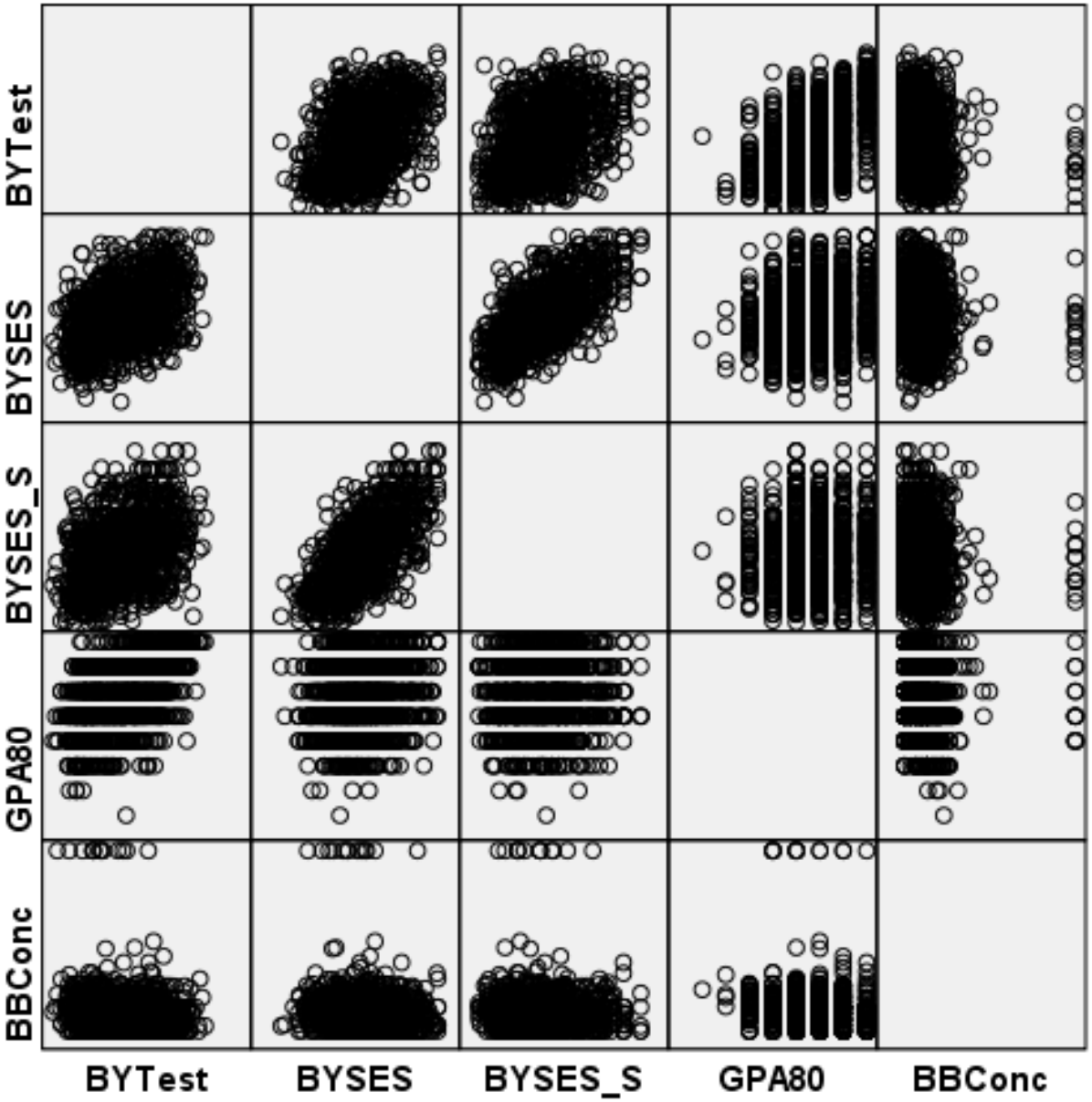
		Base Year Composite Test	Base Year SES	BY SES, School Avg	GPA 1980	Base Year Self Concept	1 = Black, 0 = Other	1 = Latino/a, 0 = Other	1 = Female, 0 = Other
Base Year Composite Test	Pearson Correlation	1.000	.440**	.429**	.508**	-.110**	-.303**	-.157**	-.158**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.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)





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)

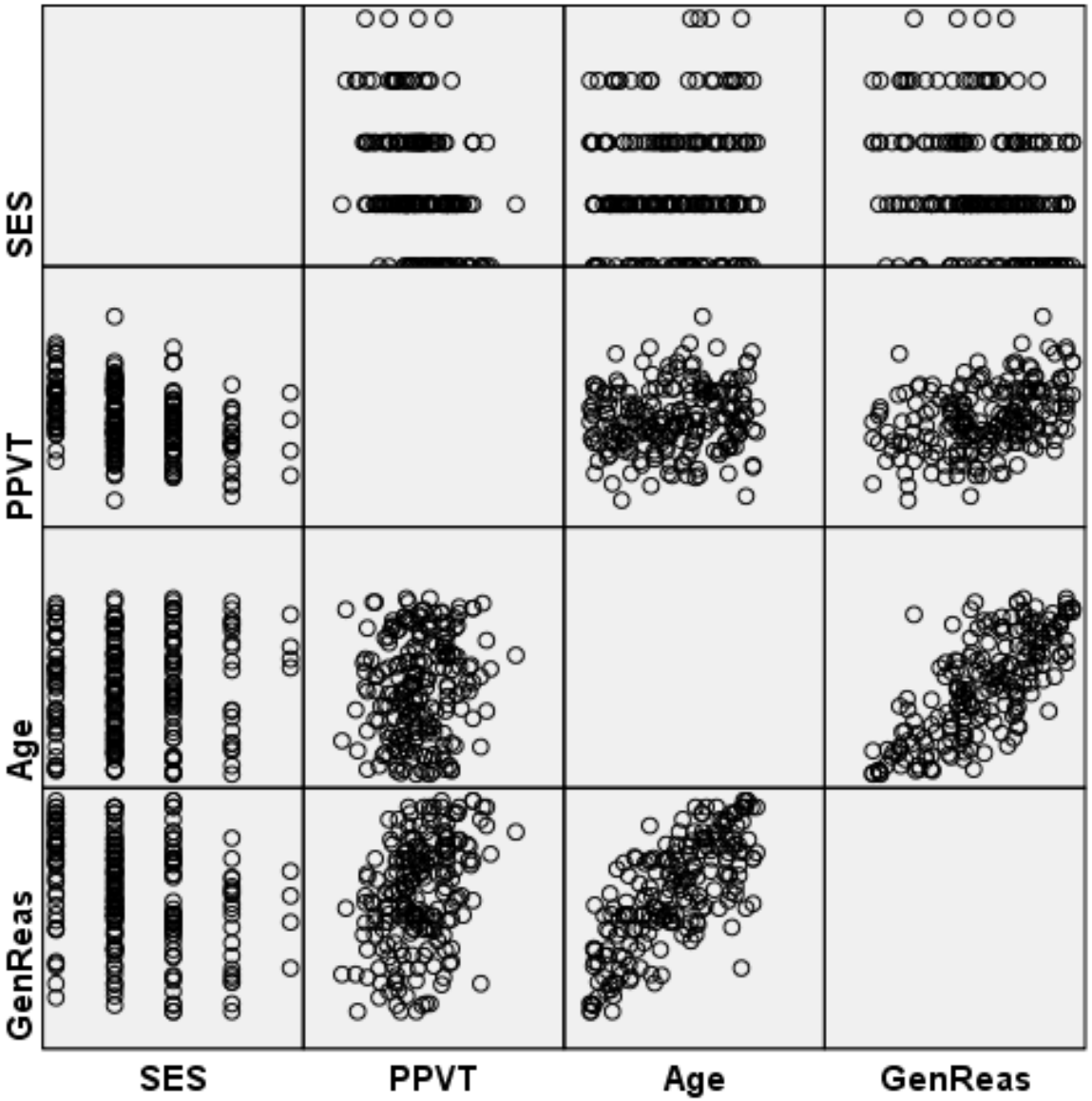
Correlations

		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)



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)

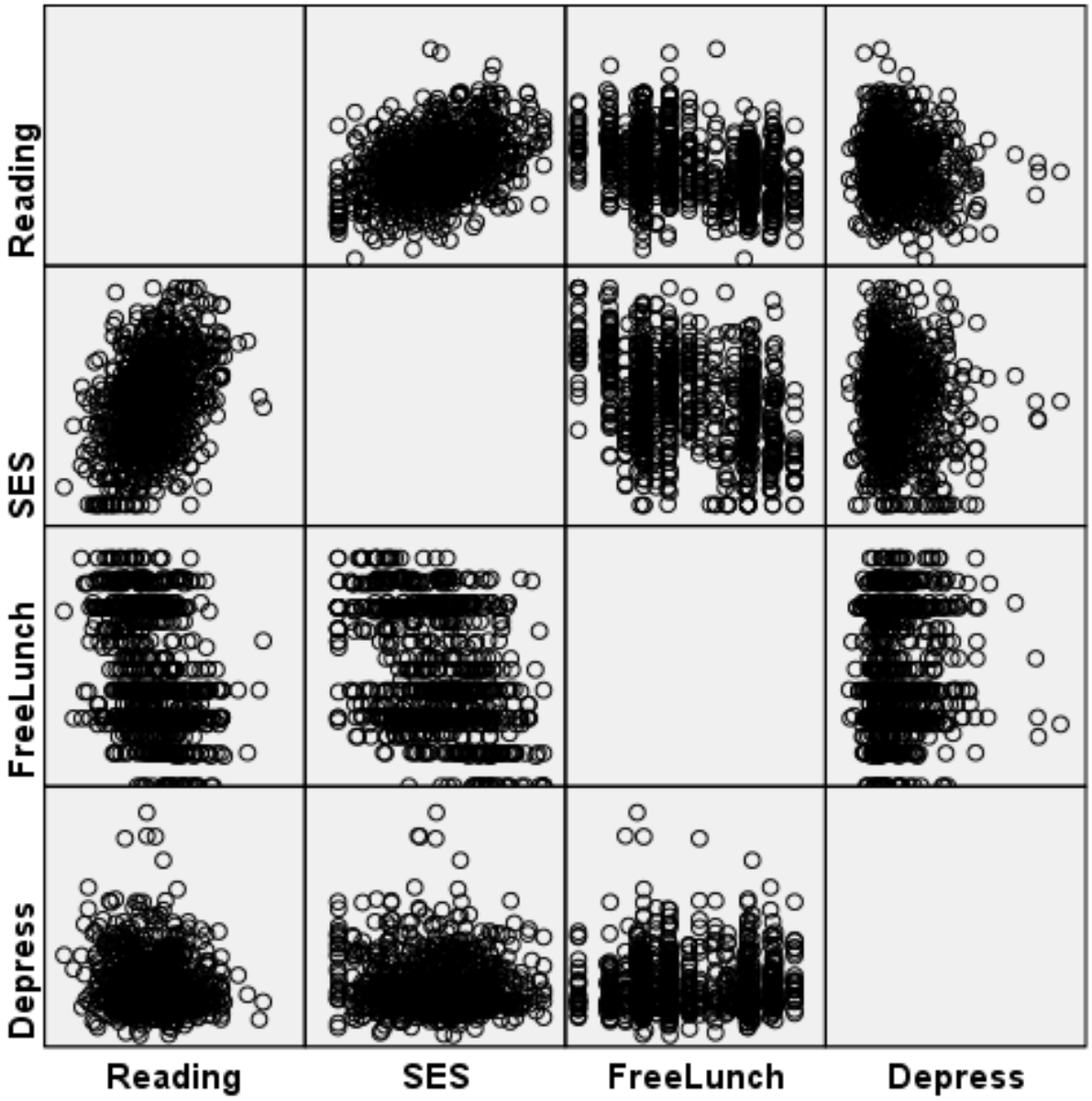
Correlations

		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)



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

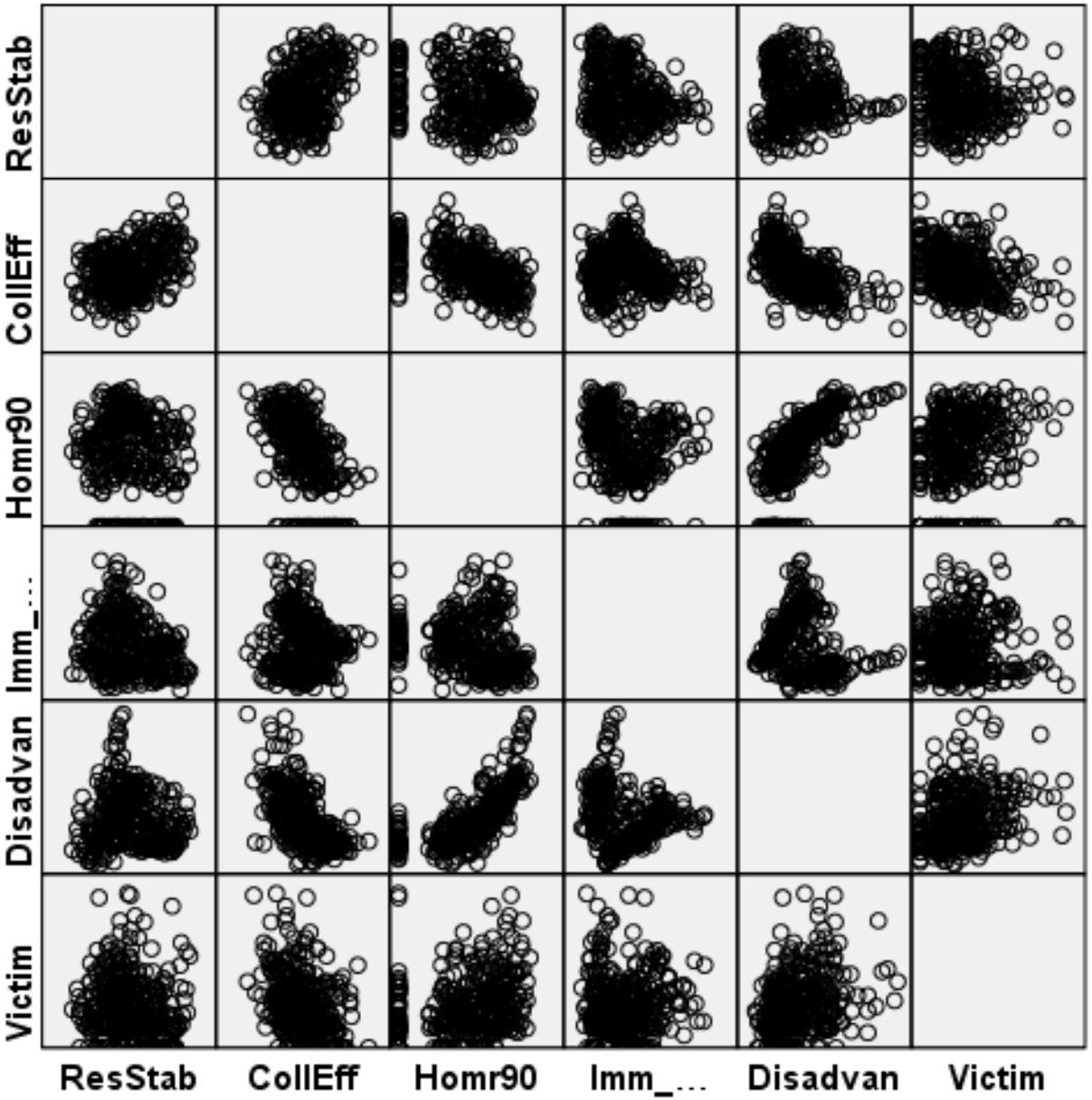
Human Development in Chicago Neighborhoods (Neighborhoods.sav)

Correlations

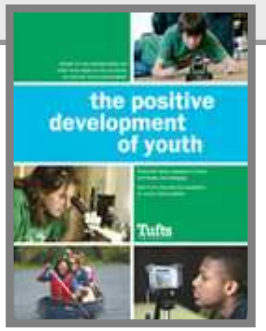
		Residential stability	Collective efficacy	Homicide rate 1988-90	Immigrant concentration	Concentrated disadvantage	% resp who were victims
Residential stability	Pearson Correlation	1.000	.382**	-.147**	-.216**	-.046	-.102
	Sig. (2-tailed)		.000	.007	.000	.400	.060
	N	342.000	342	342	342	342	342
Collective 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)



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



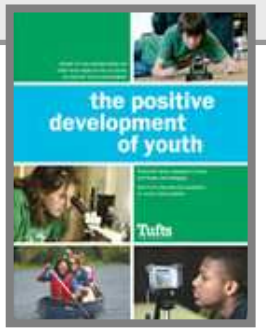
- 4-H Study of Positive Youth Development
- Source: Subset of data from IARYD, Tufts University

- 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)