Unit 5: The R^2 Statistic

Unit 5 Post Hole:

Interpret an R^2 statistic verbally and, using Boolean circles, graphically.

Unit 5 Technical Memo and School Board Memo:

Fit and discuss four regression models (with your variables from Memo 4).

Unit 5 Reading:

http://onlinestatbook.com/
Chapter 12, Prediction (except Part F)

Unit 5: 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 <u>descriptive statistics</u>. 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 <u>descriptive statistics</u>.

    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 5: Technical Memo and School Board Memo

Work Products (Part II of II):

I. Technical Memo (continued)

  D. Regression Analysis. Answer your research question using <u>inferential statistics</u>. (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 5: 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 5: 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 5: 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

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

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

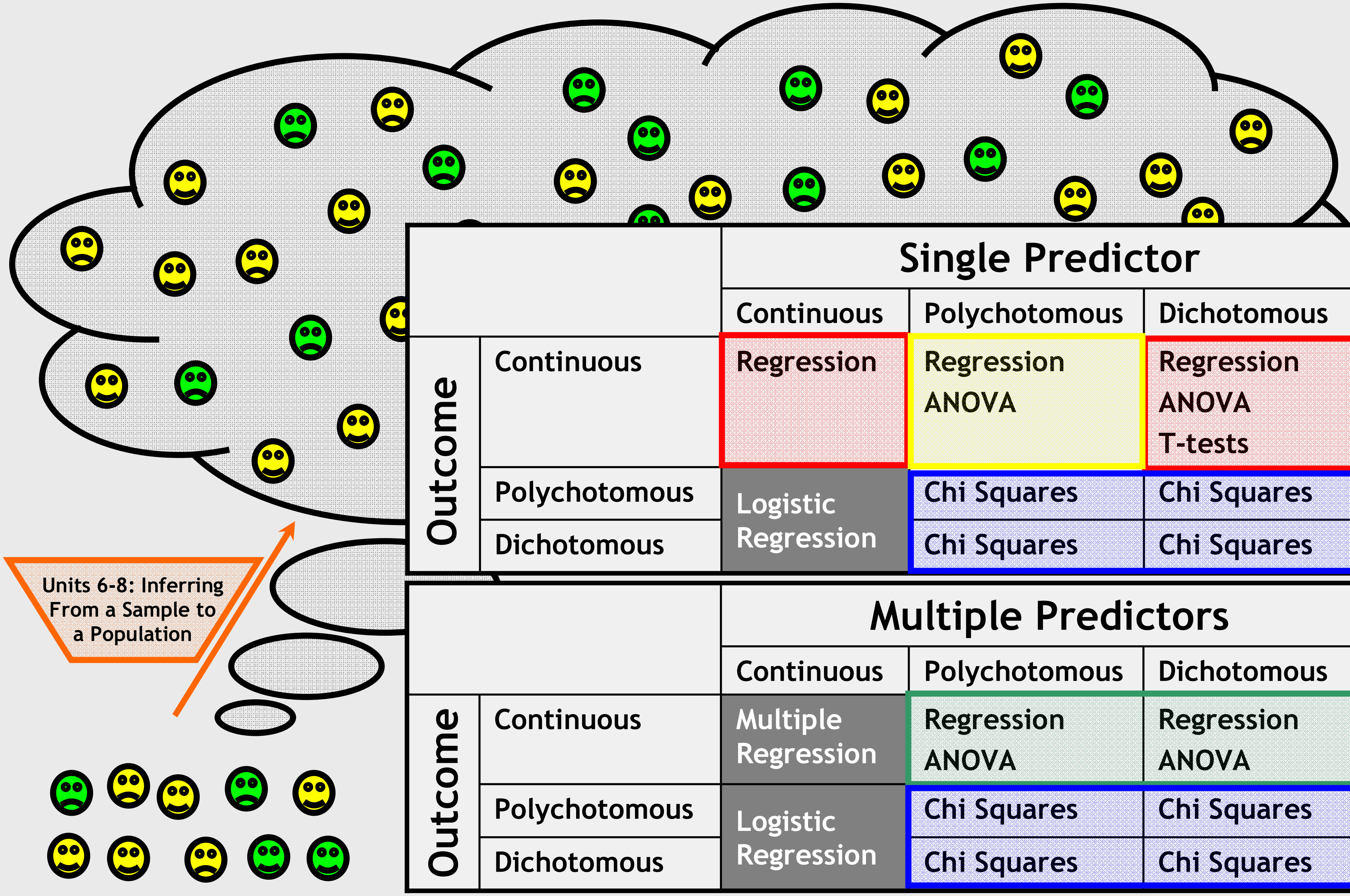
Coefficients[a]

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 5: Road Map (Schematic)

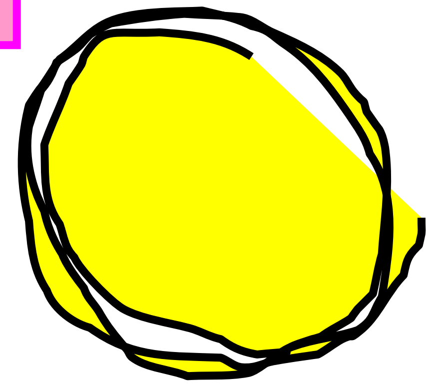


		Single Predictor		
		Continuous	Polychotomous	Dichotomous
Outcome	Continuous	Regression	Regression ANOVA	Regression ANOVA T-tests
	Polychotomous	Logistic Regression	Chi Squares	Chi Squares
	Dichotomous		Chi Squares	Chi Squares

		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

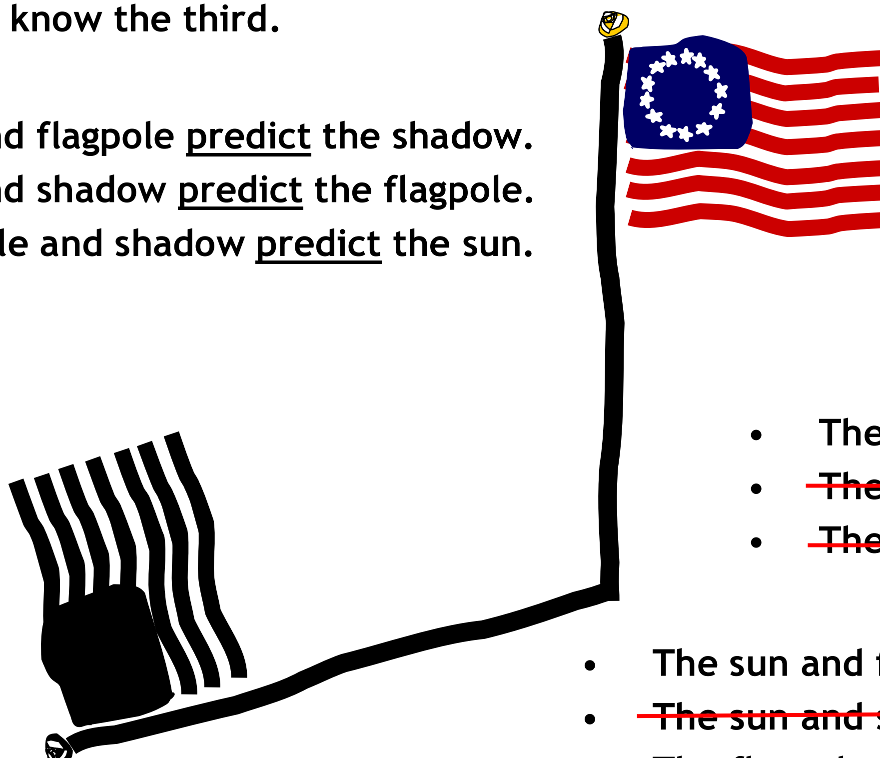
Epistemological Minute

"Explains" and "accounts for" imply causality. Correlations do not imply causality. Therefore, "explains" and "accounts for" are misleading interpretations of correlations.



According to Mr. Kochnowicz, my high school Earth Science teacher: the sun, the school's flagpole, and the flagpole's shadow each have a position, but we only need to know two of the positions in order to know the third.

- The sun and flagpole predict the shadow.
- The sun and shadow predict the flagpole.
- The flagpole and shadow predict the sun.



"Explains" and "accounts for" are common interpretations among statisticians, but data analysts should be held to a higher standard.

- The sun and flagpole explain the shadow.
- ~~The sun and shadow explain the flagpole.~~
- ~~The flagpole and shadow explain the sun.~~

- The sun and flagpole account for the shadow.
- ~~The sun and shadow account for the flagpole.~~
- ~~The flagpole and shadow account for the sun.~~

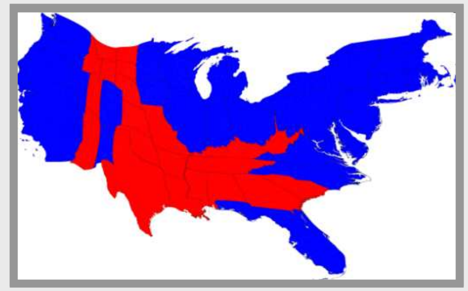
Why the higher standard for data analysts? The conclusions of data analysts impact the lives of children. That's why.

See "Explanatory Asymmetries" in the *Stanford Encyclopedia of Philosophy* entry on Scientific Explanation.

Unit 5: Research Question

Theory: Since all cognitively demanding tests are correlated, the verbal SAT and the math SAT will be correlated.

Research Question: In 1994, were state average verbal SAT scores positively correlated with state average math SAT scores?



2008 Presidential Election Results
http://www-personal.umich.edu/~mejn/election/2008/

Data Set: SAT Scores By State (SAT.sav)

Variables:

Outcome—State Average Math SAT Score (*MATH*)
Predictor—State Average Verbal SAT (*VERBAl*)

Model: $Math = \beta_0 + \beta_1 Verbal + \varepsilon$

SAT.sav Codebook

SAT Scores by State

Source: http://www.stat.ucla.edu/datasets/view_data.php?data=30

Dataset entered on: 2005-09-07

Summary
Is School Performance Related to Spending? This data set provides an example of the types of data that public policy makers consider when making decisions and crafting arguments.

Sample: The 50 United States, 1994-95.

Documentation
This data set includes eight variables:

- STATE: name of state
- COST: current expenditure per pupil (measured in thousands of dollars per average daily attendance in public elementary and secondary schools)
- RATIO: average pupil/teacher ratio in public elementary and secondary schools during Fall 1994
- SALARY: estimated average annual salary of teachers in public elementary and secondary schools during 1994-95 (in thousands of dollars)
- PERCENT: percentage of all eligible students taking the
- SAT in 1994-95
- VERBAL: average verbal SAT score in 1994-95
- MATH: average math SAT score in 1994-95
- TOTAL: average total score on the SAT in 1994-95

The SAT Data Set

*SAT.sav [DataSet1] - SPSS Data Editor

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

1 : STATE | Alabama | Visible: 8 of 8 Variables

	STATE	COST	RATIO	SALARY	PERCENT	VERBAL	MATH	SAT	var	var
7	Connecticut	9	14.400	50.045	81.000	431	477	908		
8	Delaware	7	16.600	39.076	68.000	429	468	897		
9	Florida	6	19.100	32.588	48.000	420	469	889		
10	Georgia	5	16.300	32.291	65.000	406	448	854		
11	Hawaii	6	17.900	38.518	57.000	407	482	889		
12	Idaho	4	19.100	29.783	15.000	468	511	979		
13	Illinois	6	17.300	39.431	13.000	488	560	1048		

Data View | Variable View

SPSS Processor is ready

*SAT.sav [DataSet1] - SPSS Data Editor

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

	Name	Type			Label					Measure
1	STATE	String		0	State			8		Nominal
2	COST	Numeric	9	0	Per Pupil Expenditure (in thousands of dollars)			3		Scale
3	RATIO	Numeric	5	3	Average Student/Teacher Ratio			8		Scale
4	SALARY	Numeric	6	3	Average Teacher Salary (in thousands of dollars)			6		Scale
5	PERCENT	Numeric	6	3	Percent of Eligible Students Who Take the SAT			6		Scale
6	VERBAL	Numeric	3	0	Average SAT Verbal Score			6		Scale
7	MATH	Numeric	3	0	Average SAT Math Score			5		Scale
8	SAT	Numeric	4	0	Average SAT Total Score			4		Scale
9										

Data View | **Variable View**

SPSS Processor is ready

SPSS Regression Output

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.970[a]	.941	.940	9.834

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

R^2 is the Pearson correlation squared. (Note that .97 *.97 = .941.) More interestingly, however, R^2 is the proportion of variation in the outcome predicted by the predictor. (Note that 74562/79204 = .941 and that 4641/79204 = .059 and that .941+.059 = 1.00 or that 74562+4641 = 79204.)

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

"Variation" is a non-technical term. We have many technical measures of variation, such as variance, sum of squares, standard deviation, midspread and range.

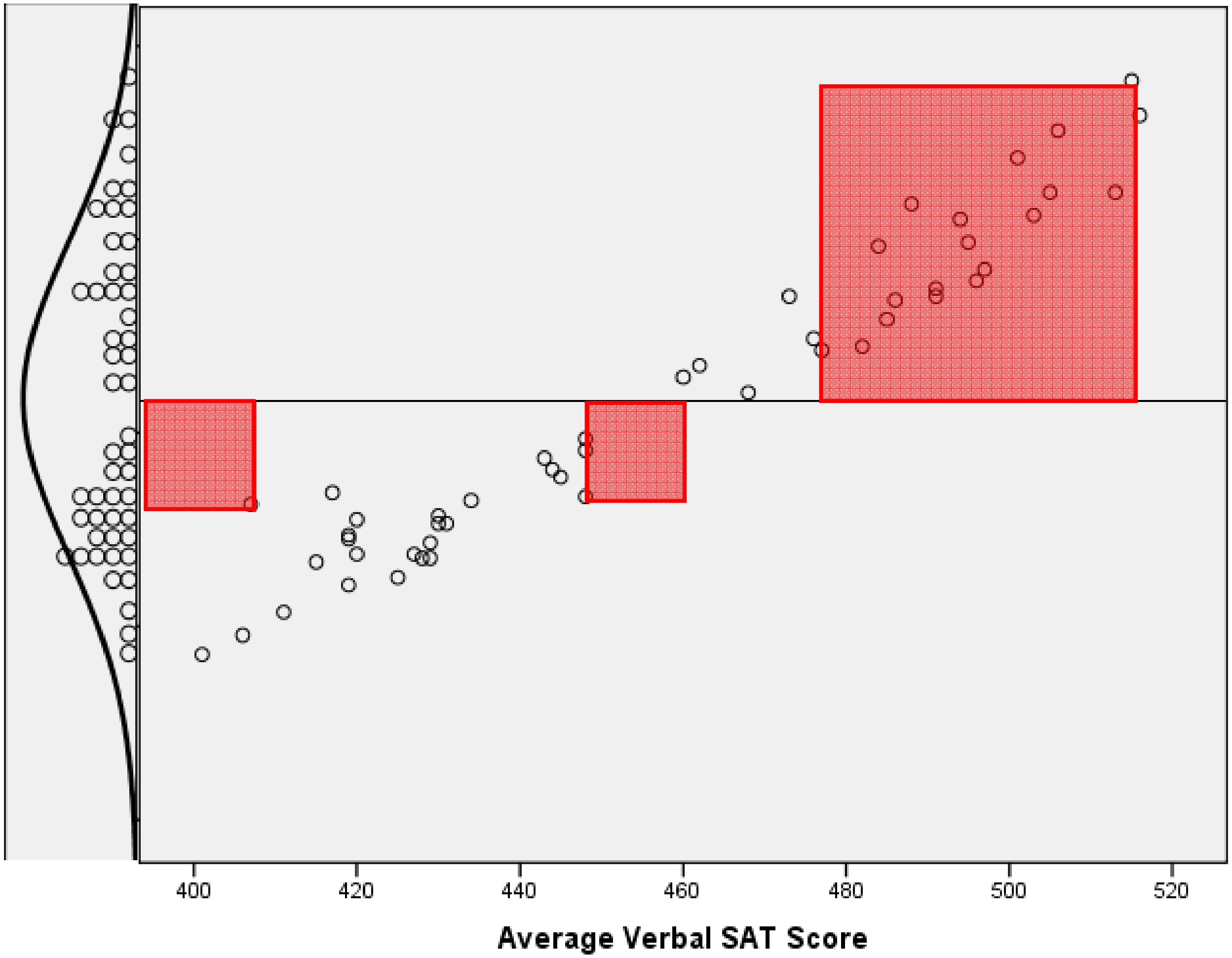
Coefficients[a]

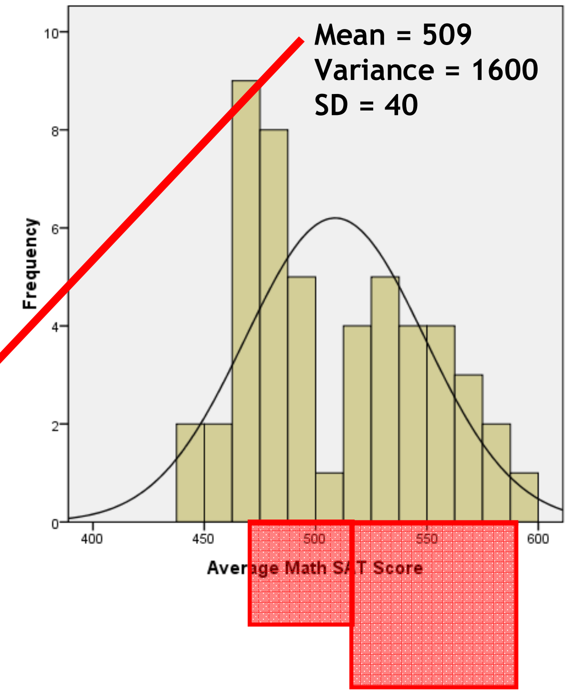
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.828	18.310		.100	.921
	Average Verbal SAT Score	1.109	.040	.970	27.768	.000

a. Dependent Variable: Average Math SAT Score

Note that, throughout the lecture, I will toggle back and forth between proportions and percentages. SPSS reports R Square as a proportion, .941, but we can just as easily think of it as a percentage, 94.1%. To convert a proportion to a percentage, multiply by 100 (i.e., move the decimal point two digits to the left) and add a "%" symbol.

Total Sum of Squares





The mean square total is the variance of the outcome:

Total sum of squares is just the sum of squared mean deviations in the outcome. We do not even need a predictor to calculate the total sum of squares. (In Post Hole 3, we calculate the total sum of squares *en route* to the variance and then the standard deviation.) It is a measure of the variation we hope to predict.

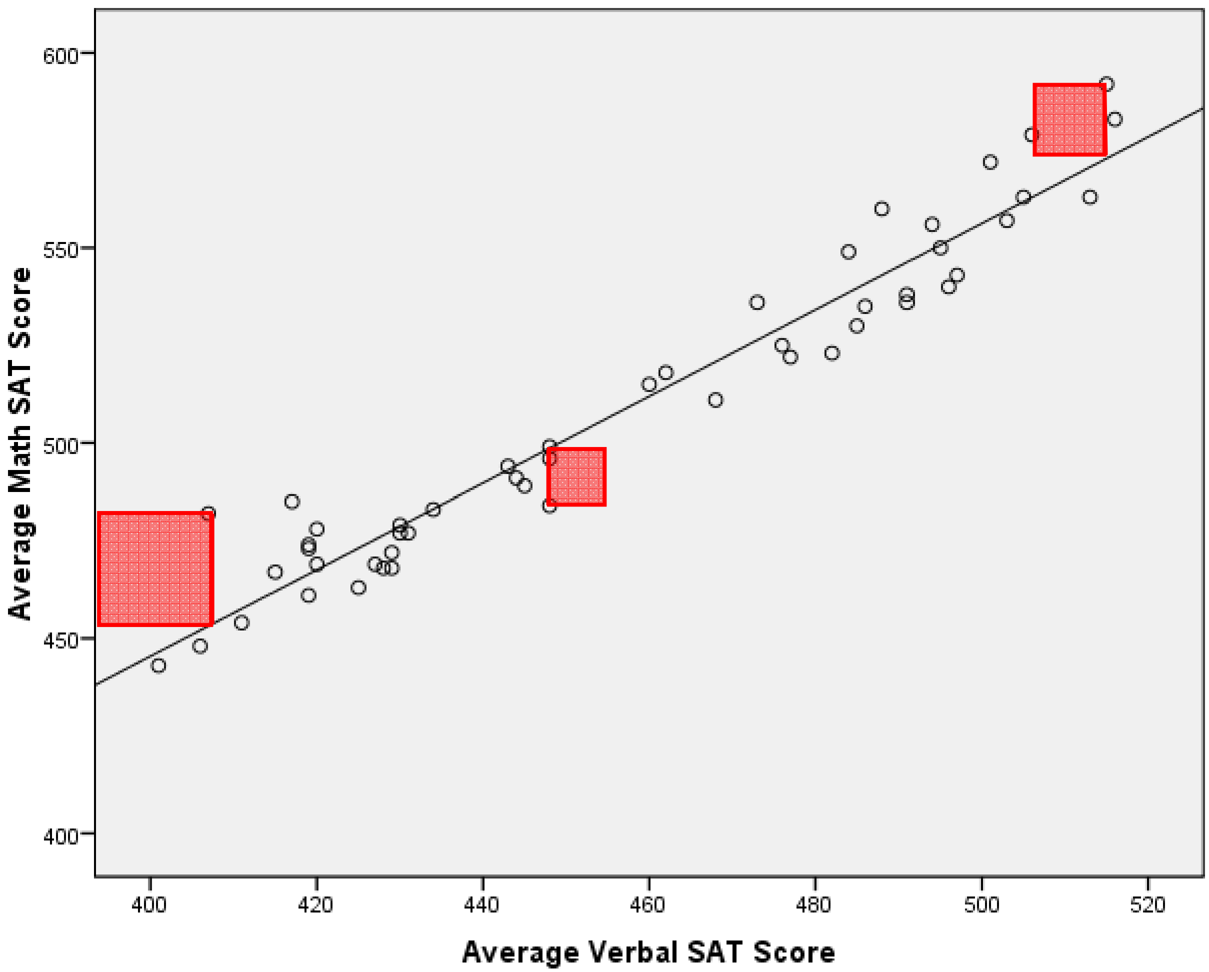
ANOVA[b]

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

Residual Sum Of Squares



Why Residuals?
- Unaccounted Variables
- Measurement Error
- Individual Variation

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

Question: If the mean square error is the variance of the residuals, what is the root mean square error (aka standard error of the estimate)?

*Not quite…notice the degrees of freedom.

Residual sum of squares is also known as "error sum of squares." It is the sum of squared residuals. It is a measure of how much variation remains to be predicted after regression, i.e., after we fit our model. The residual sum of squares is a "badness of fit" statistic. We want our errors to be small.

ANOVA[b]

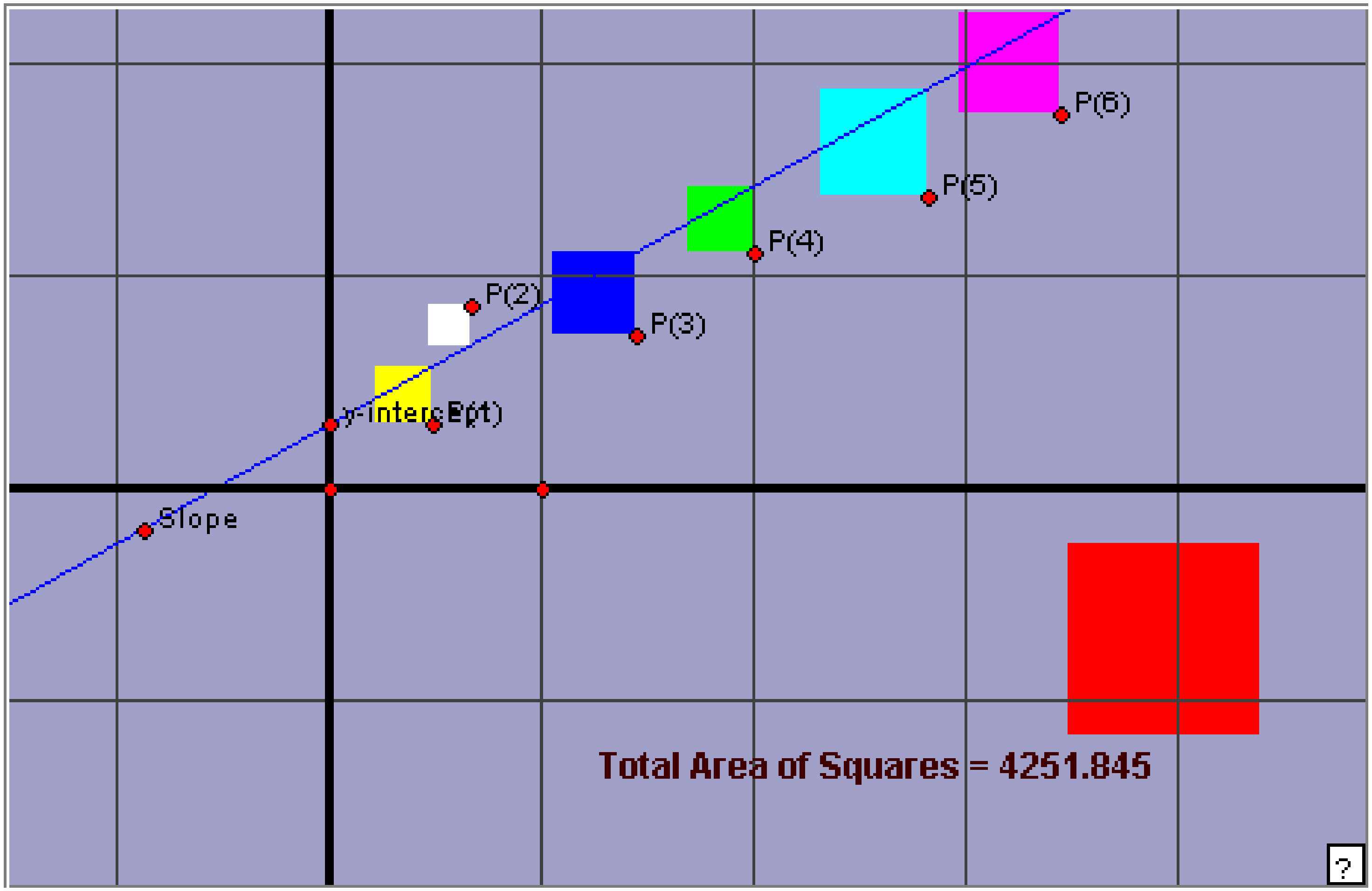
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	74562.936	1	74562.936	771.068	.000[a]
	Residual	4641.644	48	96.701		
	Total	79204.580	49			

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

b. Dependent Variable: Average Math SAT Score

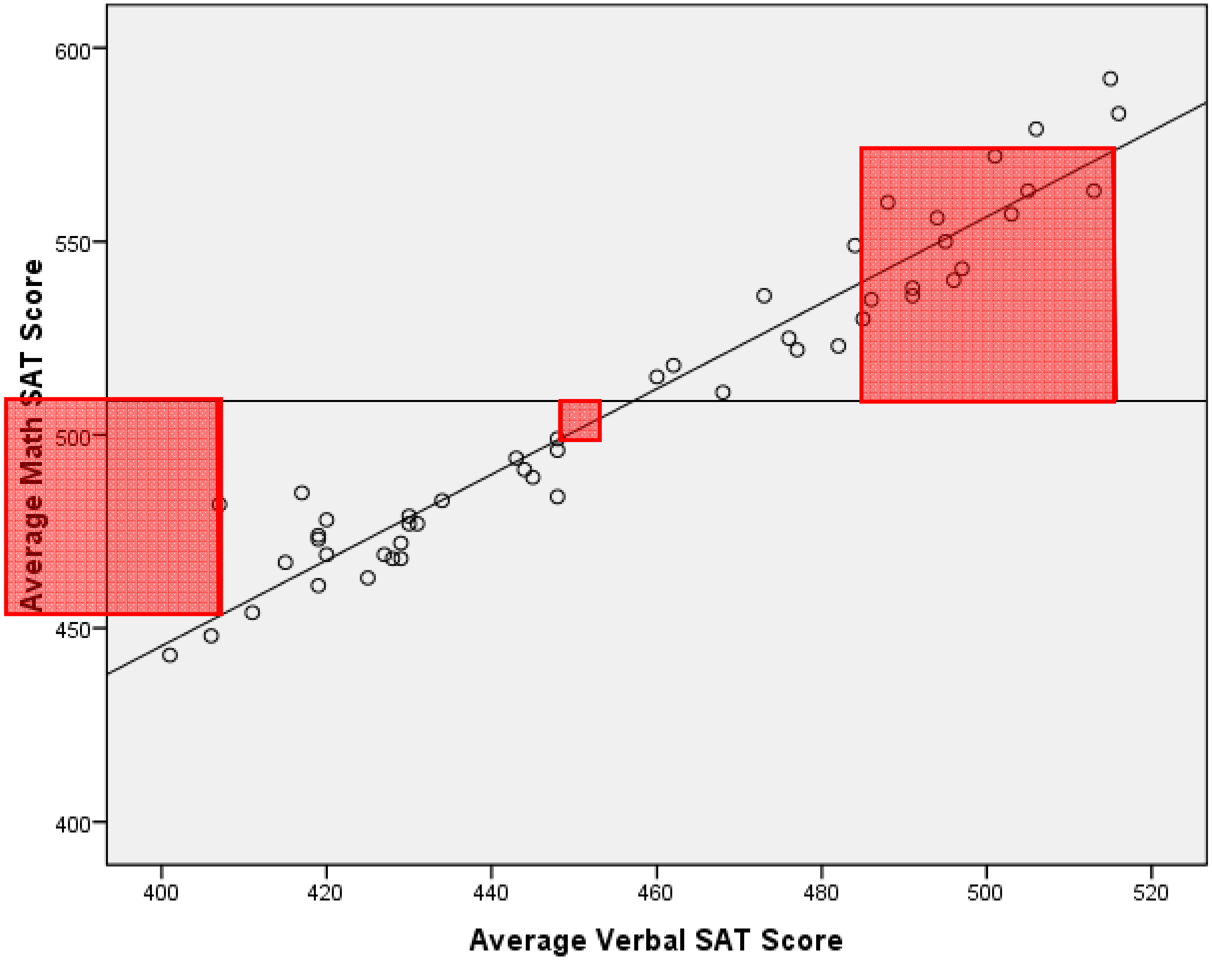
Ordinary Least Squares (OLS) Regression

How does SPSS fit the line? The Method of Ordinary Least Squares



http://www.dynamicgeometry.com/JavaSketchpad/Gallery/Other_Explorations_and_Amusements/Least_Squares.html

Regression Sum of Squares



If you understand total and residual sums of squares, you understand regression sum of squares, because the regression sum of squares is simply the total sum of squares minus the residual sum of squares.

For each observation, we can note the difference between the mean as a prediction and the regression line as a prediction. We can square these differences to get a sort of "value added" for the regression model. The R^2 statistic is the proportion of this regression sum of squares to the total sum of squares. It's easier for me, however, to conceptualize $1-R^2$: the variance* of the residuals as a proportion of the variance of the outcome. Then, since I understand $1-R^2$, I understand R^2. What makes regression sum of squares tricky is that it is 100% abstract, whereas deviations from the mean line or regression line at least involve concrete observations.

Regression sum of squares is also known as "model sum of squares." It is the sum of squared distance between the mean and our predictions. The residual sum of squares is a "goodness of fit" statistic. For the sake of prediction, we want our regression line to tell us more than the mean line.

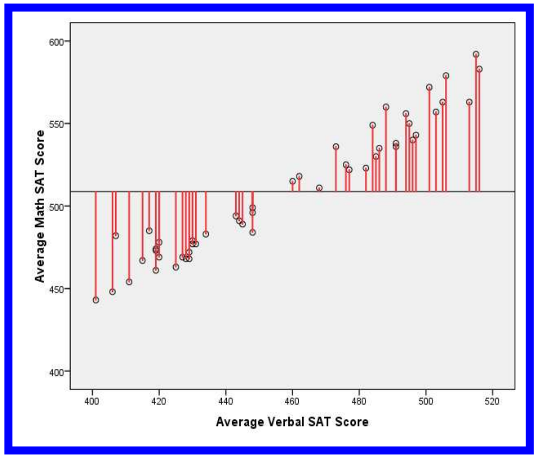
ANOVA[b]

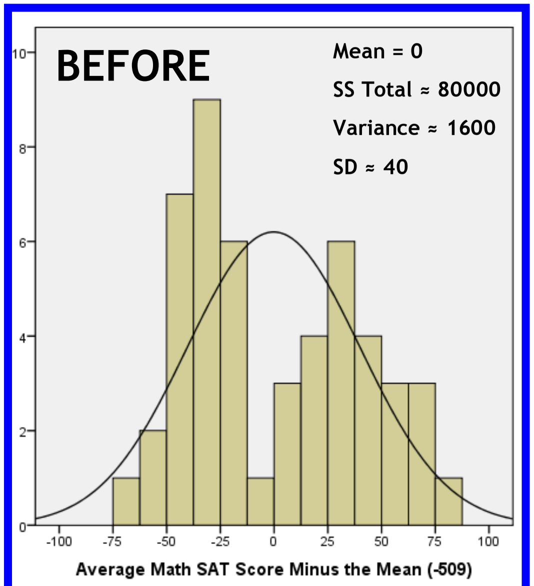
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	74562.936	1	74562.936	771.068	.000[a]
	Residual	4641.644	48	96.701		
	Total	79204.580	49			

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

b. Dependent Variable: Average Math SAT Score

$1-R^2$ and R^2





I am sorry for the flip, but there is a flip built into the nature of the R^2 statistic, and I think this is the easiest spot for it. Here is some quick flipping practice:

If we are ***not*** predicting 40% of the variation, how much of the variation ***are*** we predicting?

If we are ***not*** predicting 88% of the variation, how much of the variation ***are*** we predicting?

$$1-R^2 = \frac{\text{Sum of Squares } \textit{AFTER} \text{ Regression}}{\text{Sum of Squares } \textit{BEFORE} \text{ Regression}}$$

$$1-R^2 = \frac{\text{Sum of Squares Residual}}{\text{Sum of Squares Total}}$$

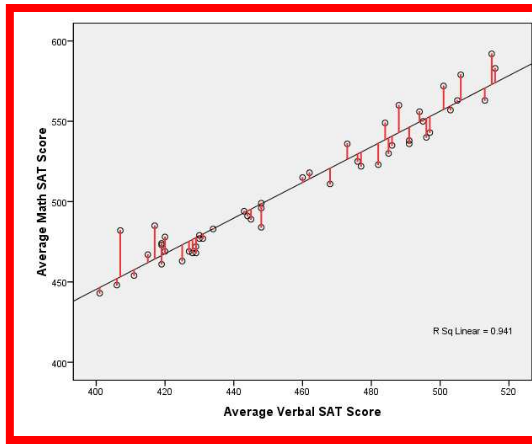
Note: All numbers are rounded for illustration.

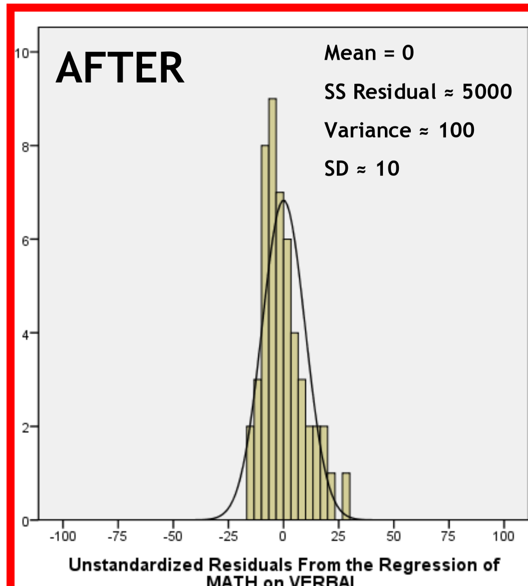
$$1-R^2 = \frac{5000}{80000} = .06 = 6\%$$

6% of the variation in Math SAT scores is ***not*** predicted by Verbal SAT scores, but let's think positive…

$$R^2 = .94 = 94\%$$

94% of the variation in Math SAT scores ***is*** predicted by Verbal SAT scores.





Consider the variance of the outcome as a baseline for what needs predicting *before* we fit our model.

Consider the variance of the residuals as what is still in need of predicting *after* we fit our model. Not bad, right?

$1-R^2$ is the proportion of variance still in need of predicting, and R^2 is the proportion of variance that we are predicting.

Interpreting the R^2 Statistic

Interpretations of the R^2 statistic:

Good: 94.1% of the variance for state average Math SAT scores is **associated** with state average Verbal SAT scores.

Good: We are **predicting** 94.1% of the variation in state average Math SAT scores with state average Verbal SAT scores.

Shady: Our model **explains** 94.1% of state average Math SAT scores.

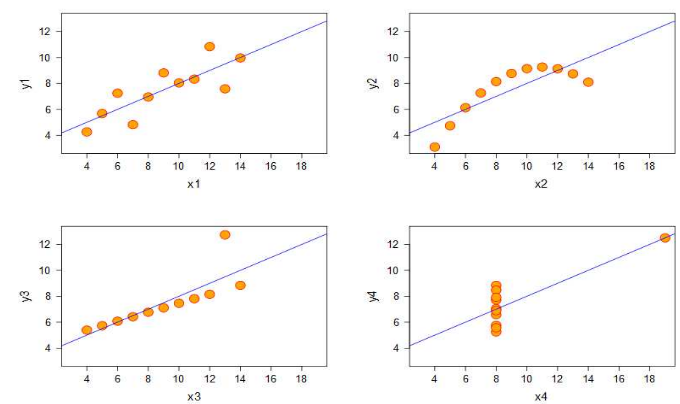
Shady: State average Verbal SAT scores **account** for 94.1% of state average Math SAT scores.

Evil: The model is 94.1% **accurate**.

Evil Evil Evil: <<<Any Causal Interpretation!>>>

The R^2 statistic is a goodness of fit statistic. Somewhat paradoxically, however, the R^2 statistic is a good goodness of fit statistic only when the model fits good, I mean, well. All the caveats that apply to Pearson's *r* (Unit 4) apply to the R^2 statistic.

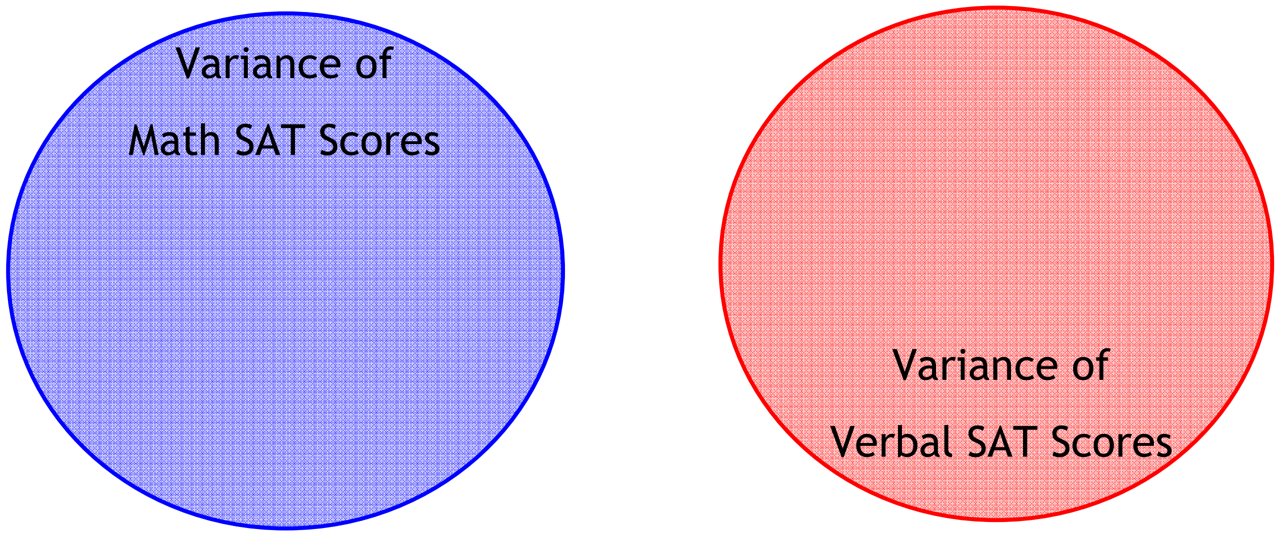
Cautionary Tale: $R^2 = 0.74$



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

Using Boolean Circles to Represent the R^2 Statistic

The R^2 statistic is symmetric exactly like Pearson correlations. If Verbal SAT scores predict 94.1% of the variance of Math SAT scores, then Math SAT scores predict 94.1% of the variance of Verbal SAT scores.

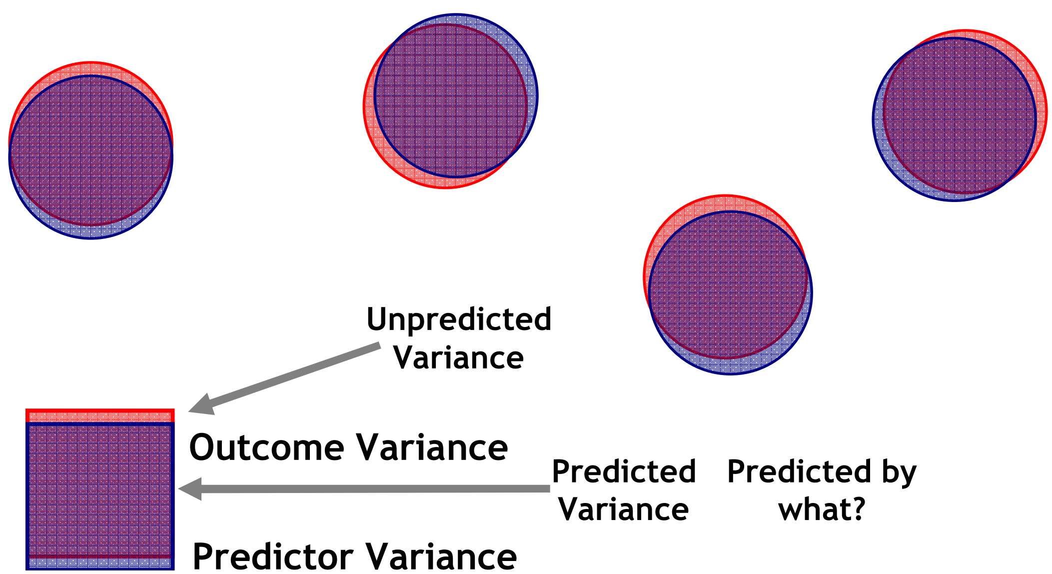


Never forget that variance is just a hardworking number trying to summarize variation. It is neither variation itself nor the subjects who vary. We have been using average squares to represent variance, so you can consider these Boolean circles as rounded versions of those squares. These circles do **not** contain people.

What Do Those Circles Really Represent? Variance

That small square is the variance in the outcome still in need of predicting AFTER the 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…

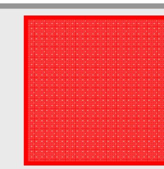
Now, what do those circles really represent? They just represent variance. Instead of squares, we use circles. Although I'm tempted to use "Boolean squares" in the future for the sake of clarity.



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!

Reporting the Results of Simple Linear Regression

When reporting the results of simple linear regression:

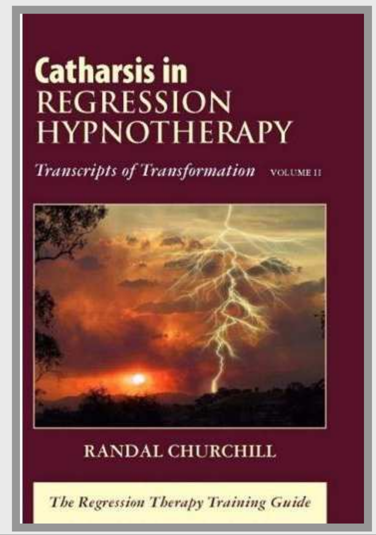
1. Include your fitted model.
2. Use the R^2 statistic to convey the goodness of fit for the model (i.e., strength).
3. To determine statistical significance, test the null hypothesis that the magnitude in the population is zero, reject (or not) the null hypothesis, and draw a conclusion (or not) from the sample to the population.
4. Describe the direction and magnitude of the relationship in your sample, preferably with illustrative examples. Make clear the substance of your findings.
5. Use confidence intervals to describe the magnitude of the relationship in the population
6. If simple linear regression is inappropriate for the data, then say so, briefly explain why, and forego any misleading analysis.

Interpret R^2 statistics using "predicts," not "explains" or "accounts."

A Causal Rule of Three Revisited:
Whenever you offer a causal interpretation of your findings, offer at least two alternatives to your pet interpretation. If you suggest that A may cause B, also suggest how B may cause A and how C may cause both A and B.

Never lose sight of the substantive meaning of the numbers.

You have what you need for the Unit 5 post hole. There is practice in back.



Dig the Post Hole

Unit 5 Post Hole:
Interpret an R^2 statistic verbally and, using Boolean circles, graphically.

Evidentiary materials: regression output (SPSS).

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

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			

Coefficients[a]

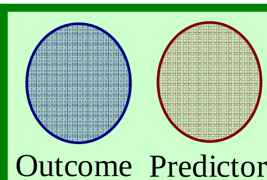
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	49.118	.115		428.169	.000
	FREELUNCH	-4.841	.198	-.267	-24.439	.000

a. Dependent Variable: READING

Evidentiary materials: regression output (R).

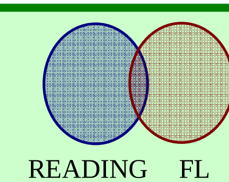
```
Call:
lm(formula = READING ~ FREELUNCH, data = RoadmapData)

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 '

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
```

Here is the answer blank:



Here is my answer:

***FREELUNCH* predicts 7% of the variation in *READING*.**



Answering our Roadmap Question

Unit 5: In our sample, free lunch predicts what proportion of variation in reading achievement?

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

FREELUNCH predicts 7% of the variation in *READING*.

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

Notice that:

531977.541 is 93% of 572721.864

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

$$\hat{Reading} = 49.1 - 4.8FreeLunch$$

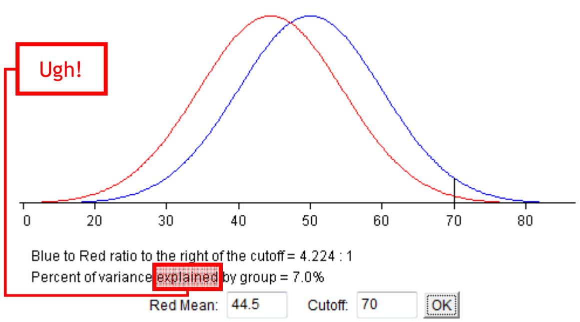
Does this mean that free lunch causes low test scores?

I only wish that were case, for then we could solve all our educational problems by charging $40 for school lunch. Now, that's silly.

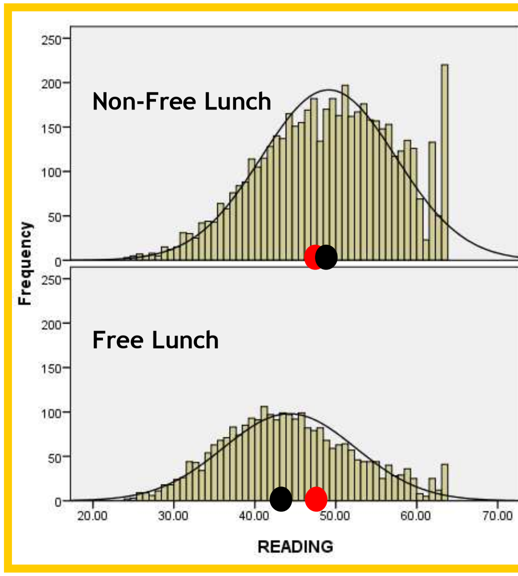
What's not silly is when people see the correlation between race and low test scores and conclude that being Black, for example, is the problem, and there is nothing the schools nor anybody can do about it. Furthermore, when data analysts use unwarranted causal language, they spur the madness. Do you think race *explains* reading scores?

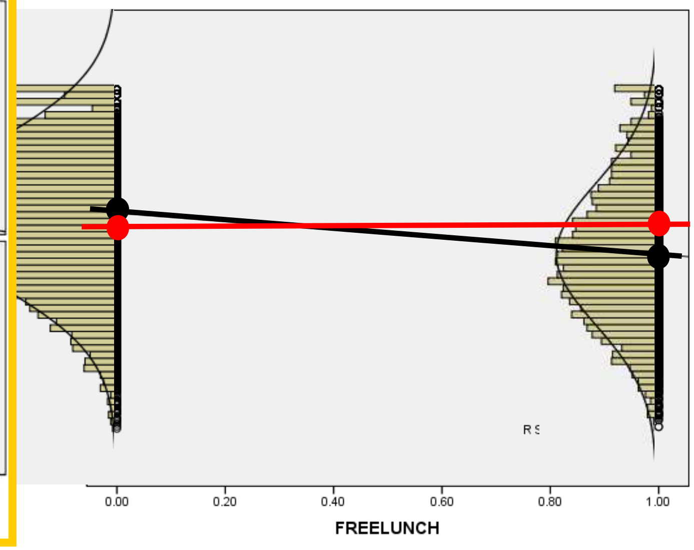
Answering our Road Map Question (Thinking Deeper)

For the sake of illustration, let's suppose that the Non-Free Lunch group and the Free Lunch group were exactly equal in size. What would the 7% mean for Tufts?



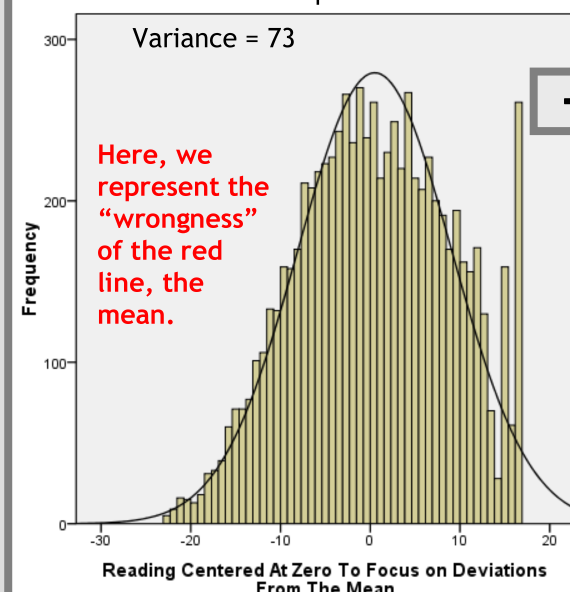
http://onlinestatbook.com/stat_sim/group_diff.html





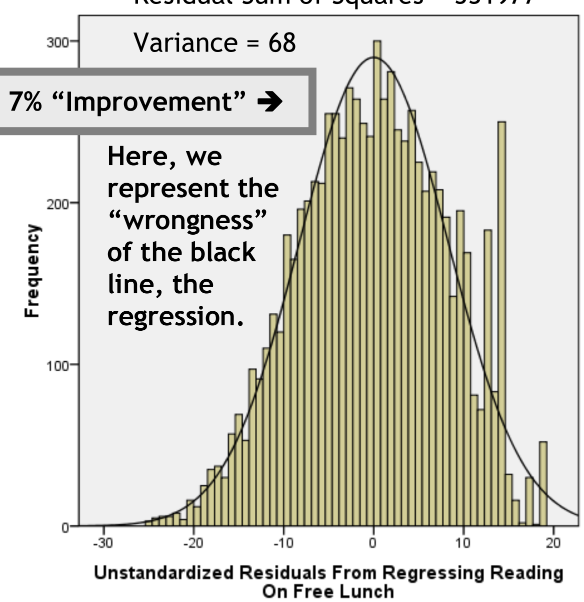
- See the histograms in the scatterplot.
- See the regression line as predicting conditional means.
- See the unconditional mean as the basis of comparison for regression.
- See that the spread of deviations from the regression line is smaller than the spread of deviations from the mean.
- See that we can compare the spreads by summing the squares, or we could compare the mean squares or, in other words, the variances.
- See that the residual sum of squares is a percentage of the total sum of squares, or the residual variance is a percentage of total variance; we are thus analyzing/predicting variance. "ANOVA" stands for analysis of variance.
- See that, if group membership predicts even a small amount of variance, the consequences can be huge.

Total Sum of Squares = 572721



➔ 7% "Improvement" ➔

Residual Sum of Squares = 531977



Answering our Roadmap Question (Revisited)

Unit 5: In our sample, free lunch predicts what proportion of variation in reading achievement?

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

FREELUNCH predicts 7% of the variation in *READING*.

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

Notice that:

531977.541 is 93% of 572721.864

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

$$\widehat{Reading} = 49.1 - 4.8 FreeLunch$$

Does this mean that free lunch causes low test scores?

I only wish that were case, for then we could solve all our educational problems by charging $40 for school lunch. Now, that's silly.

What's not silly is when people see the correlation between race and low test scores and conclude that being Black, for example, is the problem, and there is nothing the schools nor anybody can do about it. Furthermore, when data analysts use unwarranted causal language, they spur the madness. Do you think race *explains* reading scores?

Unit 5 Appendix: Key Concepts

R^2 is the Pearson correlation squared. (Note that .97*.97=.941.) More interestingly, however, R^2 is the proportion of variation in the outcome predicted by the predictor. (Note that 74562/79204 = .941 and that 4641/79204 = .059 and that .941+.059 = 1.00 or that 74562+4641 = 79204.)

Why Residuals?

- Unaccounted Variables
- Measurement Error
- Individual Variation

The R^2 statistic is a goodness of fit statistic. Somewhat paradoxically, however, the R^2 statistic is a good goodness of fit statistic only when the model fits good, I mean, well. All the caveats that apply to Pearson's r (Unit 4) apply to the R^2 statistic.

The R^2 statistic is symmetric exactly like Pearson correlations. If Verbal SAT scores predict 94.1% of the variance of Math SAT scores, then Math SAT scores predict 94.1% of the variance of Verbal SAT scores.

Never forget that variance is just a hardworking number trying to summarize variation. It is neither variation itself nor the subjects who vary. We have been using average squares to represent variance, so you can consider the Boolean circles as rounded versions of those squares. The circles do **not** contain people.

"Explains" and "accounts for" imply causality. Correlations do not imply causality. Therefore, "explains" and "accounts for" are misleading interpretations of correlations.

Unit 5 Appendix: Key Interpretations

Good: 94.1% of the variance for state average Math SAT scores is **associated** with state average Verbal SAT scores.

Good: We are **predicting** 94.1% of the variation in state average Math SAT scores with state average Verbal SAT scores.

Shady: Our model **explains** 94.1% of state average Math SAT scores.

Shady: State average Verbal SAT scores **account** for 94.1% of state average Math SAT scores.

Evil: The model is 94.1% **accurate**.

Evil Evil Evil: <<<Any Causal Interpretation!>>>

Unit 5 Appendix: Key Terminology

The R^2 statistic measures the goodness of model fit (as long as the model fits!). It is a close cousin to Pearson's r. Like Person's r, the R^2 statistic quantifies strength. Unlike Pearson's r, the R^2 statistic does not signify the direction of the interpretation, but it does have a nifty (albeit oft misunderstood) verbal interpretation as the proportion of variance predicted.

"Variation" is a non-technical term. We have many technical measures of variation, such as variance, sum of squares, standard deviation, midspread and range.

Total sum of squares is just the sum of squared mean deviations in the outcome. We do not even need a predictor to calculate the total sum of squares. (In Post Hole 3, we calculate the total sum of squares *en route* to the variance and then the standard deviation.) It is a measure of the variation we hope to predict.

Residual sum of squares is also known as "error sum of squares." It is the sum of squared residuals. It is a measure of how much variation remains to be predicted after regression, i.e., after we fit our model. The residual sum of squares is a "badness of fit" statistic. We want our errors to be small.

Regression sum of squares is also known as "model sum of squares." It is the sum of squared distance between the mean and our predictions. The residual sum of squares is a "goodness of fit" statistic. For the sake of prediction, we want our regression line to tell us more than the mean line.

Unit 5 Appendix: Math

Every individual observation gets three squares:

The blue square represents the squared difference between the observed outcome for the individual and the mean of the outcome.

$$(Y_i - \bar{Y})^2$$

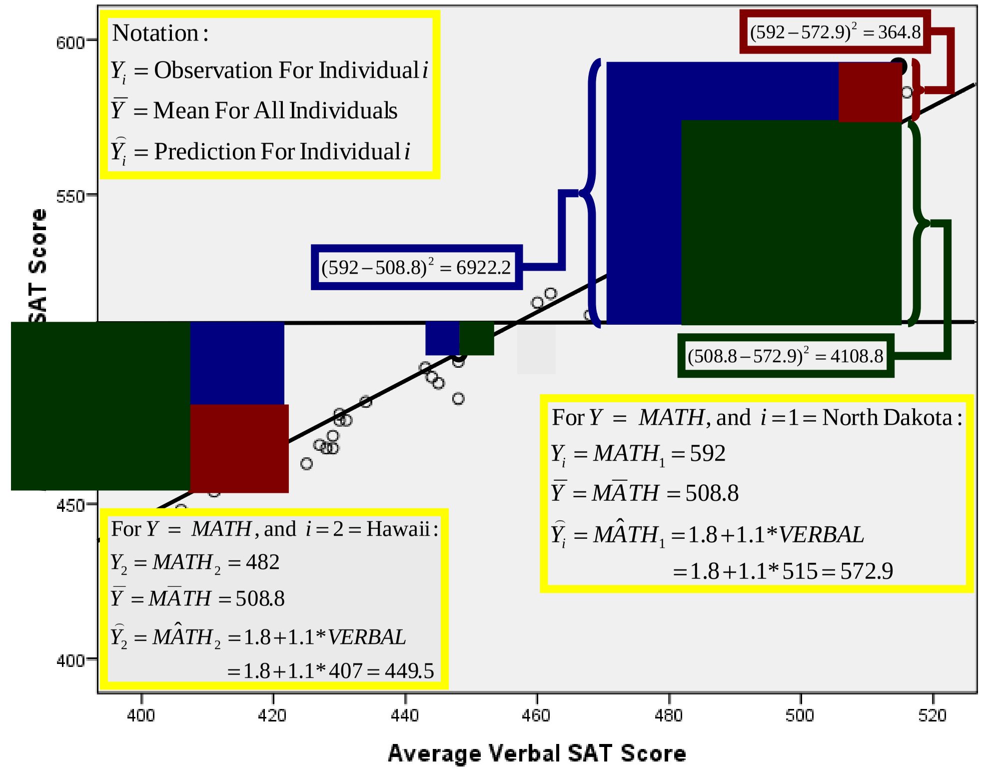
The red square represents the squared difference between the observed outcome for the individual and the predicted outcome for the individual.

$$(Y_i - \hat{Y}_i)^2$$

The green square represents the squared difference between the mean of the outcome and the predicted outcome for the individual.

$$(\bar{Y} - \hat{Y}_i)^2$$

Cool Algebraic Fact: Because of the way we fit our regression line, all the blue squares combined equal all the red squares combined plus all the green squares combined.
http://en.wikipedia.org/wiki/Sum_of_squares



$$\text{Sum of Squares Total} = SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$\text{Sum of Squares Residual/Error} = SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$\text{Sum of Squares Regression/Model} = SSM = \sum_{i=1}^{n}(\bar{Y} - \hat{Y}_i)^2$$

Cool Algebraic Fact: SST = SSE + SSM

Don't be afraid of capital sigma (Σ). It is the capital Greek letter S, and it stands for "sum." It just means "add 'em up"! You calculate SST for Post Hole 3.

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSM}{SST}$$

The R^2 statistic is a goodness of fit statistic:

$$R^2 = 1 - \frac{BAD}{BASELINE} = \frac{GOOD}{BASELINE}$$

Unit 5 Appendix: SPSS Syntax

```
*You can use my code by switching out my variables (circled) with your variables.
*You can make a comment by starting with an asterisk and ending with a period.
*SPSS will ignore anything between the asterisk and period.
*SPSS loves/needs to end chunks of command with a period, so if something is
acting funky, make sure that your periods are in order.
***************************************************************************.
*I'm going to create a scatterplot with PERCENT on the x-axis and
SAT on y-axis; the only thing that you can't decipher is the
"/MISSING=LISTWISE" line, but all this does is tell SPSS to ignore
anybody with missing data for the variables at play in this
chunk of code.
***************************************************************************.
GRAPH
  /SCATTERPLOT(BIVAR)=PERCENT WITH SAT
  /MISSING=LISTWISE.
***************************************************************************.
*I'm going to linearly regress SAT on PERCENT.
*NOTE THAT IT IS STUPID TO TAKE SERIOUSLY THE RESULTS SINCE THE RELATIONSHIP IS NONLINEAR.
*Notice our now familiar friend "LISTWISE".
*Notice that, against proper English, I put the last period outside the quotation marks!.
*I didn't want SPSS to "see" a dangling quotation mark and wonder what to do.
*Notice the last two lines; you should be able to decipher a little.
*Ignore the rest for now.
***************************************************************************.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT SAT
  /METHOD=ENTER PERCENT.
```

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)	Self Disclosure to Boyfriend (B_Seldis)
Trusts Mother (M_Trust)	Trusts Boyfriend (B_Trust)
Mutual Caring with Mother (M_Care)	Mutual Caring with Boyfriend (B_Care)
Risk Vulnerability with Mother (M_Vuln)	Risk Vulnerability with Boyfriend (B_Vuln)
Physical Affection with Mother (M_Phys)	Physical Affection with Boyfriend (B_Phys)
Resolves Conflicts with Mother (M_Cres)	Resolves Conflicts with Boyfriend (B_Cres)

Perceived Intimacy of Adolescent Girls (Intimacy.sav)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.731[a]	.534	.526	.80682

a. Predictors: (Constant), Self-disclose to boyfriend

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	43.280	1	43.280	66.487	.000[a]
	Residual	37.756	58	.651		
	Total	81.037	59			

a. Predictors: (Constant), Self-disclose to boyfriend

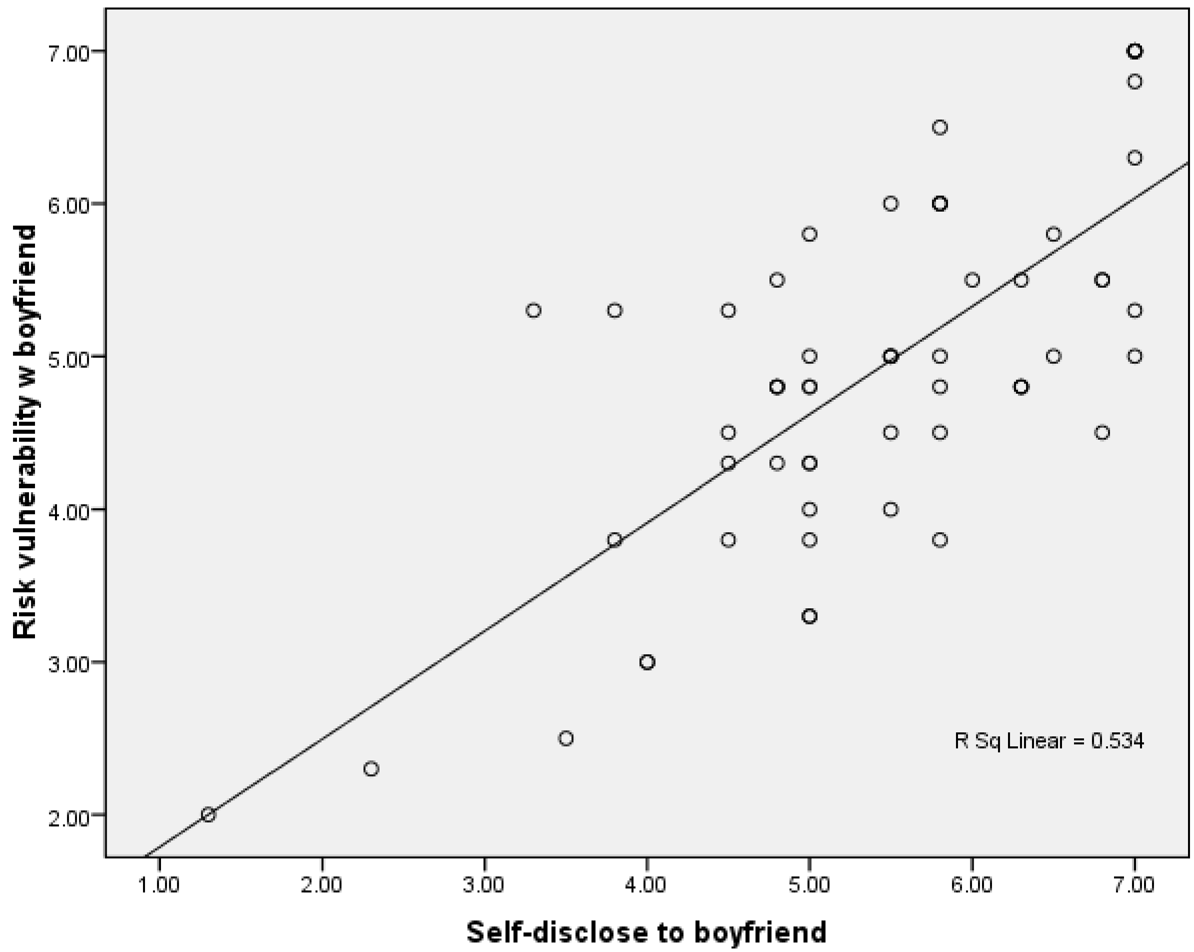
b. Dependent Variable: Risk vulnerability w boyfriend

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.081	.482		2.244	.029
	Self-disclose to boyfriend	.708	.087	.731	8.154	.000

a. Dependent Variable: Risk vulnerability w boyfriend

Perceived Intimacy of Adolescent Girls (Intimacy.sav)



Perceived Intimacy of Adolescent Girls (Intimacy.sav)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.002[a]	.000	-.017	1.19785

a. Predictors: (Constant), Self-disclose to mother

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.000	1	.000	.000	.985[a]
	Residual	83.221	58	1.435		
	Total	83.222	59			

a. Predictors: (Constant), Self-disclose to mother

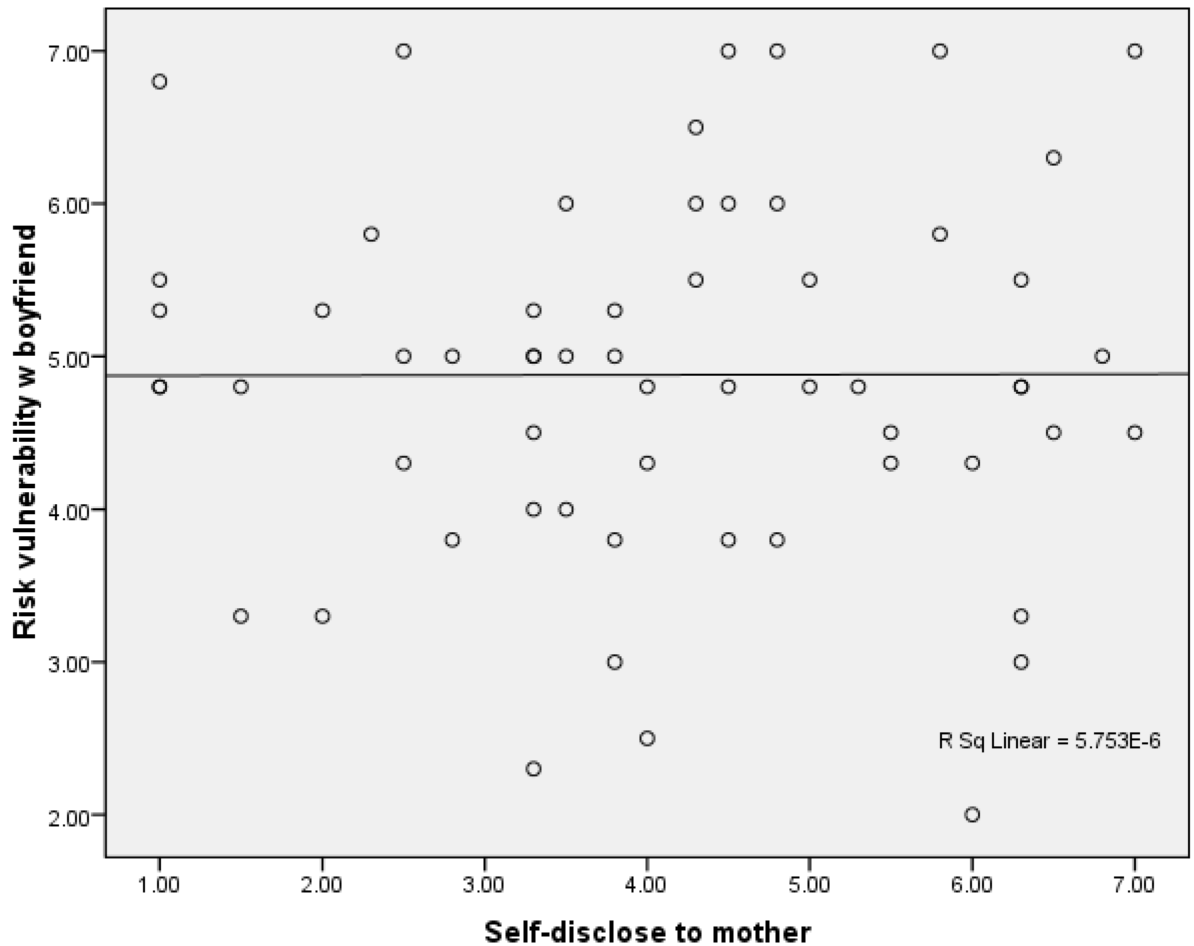
b. Dependent Variable: Risk vulnerability w boyfriend

Coefficients[a]

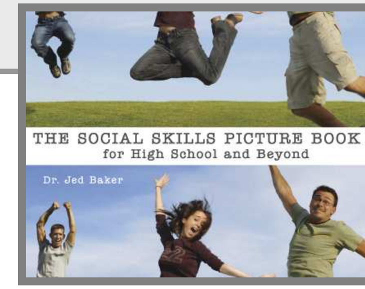
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.872	.404		12.050	.000
	Self-disclose to mother	.002	.091	.002	.018	.985

a. Dependent Variable: Risk vulnerability w boyfriend

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:**

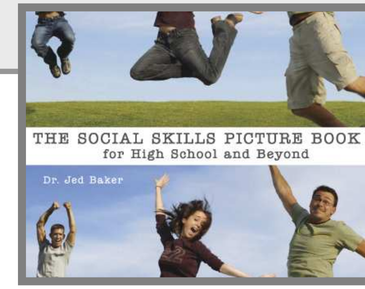
Variables about the student—

(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)



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.440[a]	.193	.192	7.71738

a. Predictors: (Constant), Base Year SES

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	14858.061	1	14858.061	249.473	.000[a]
	Residual	62059.321	1042	59.558		
	Total	76917.382	1043			

a. Predictors: (Constant), Base Year SES

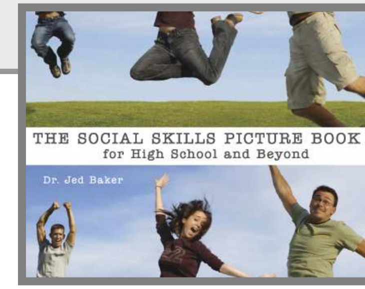
b. Dependent Variable: Base Year Composite Test

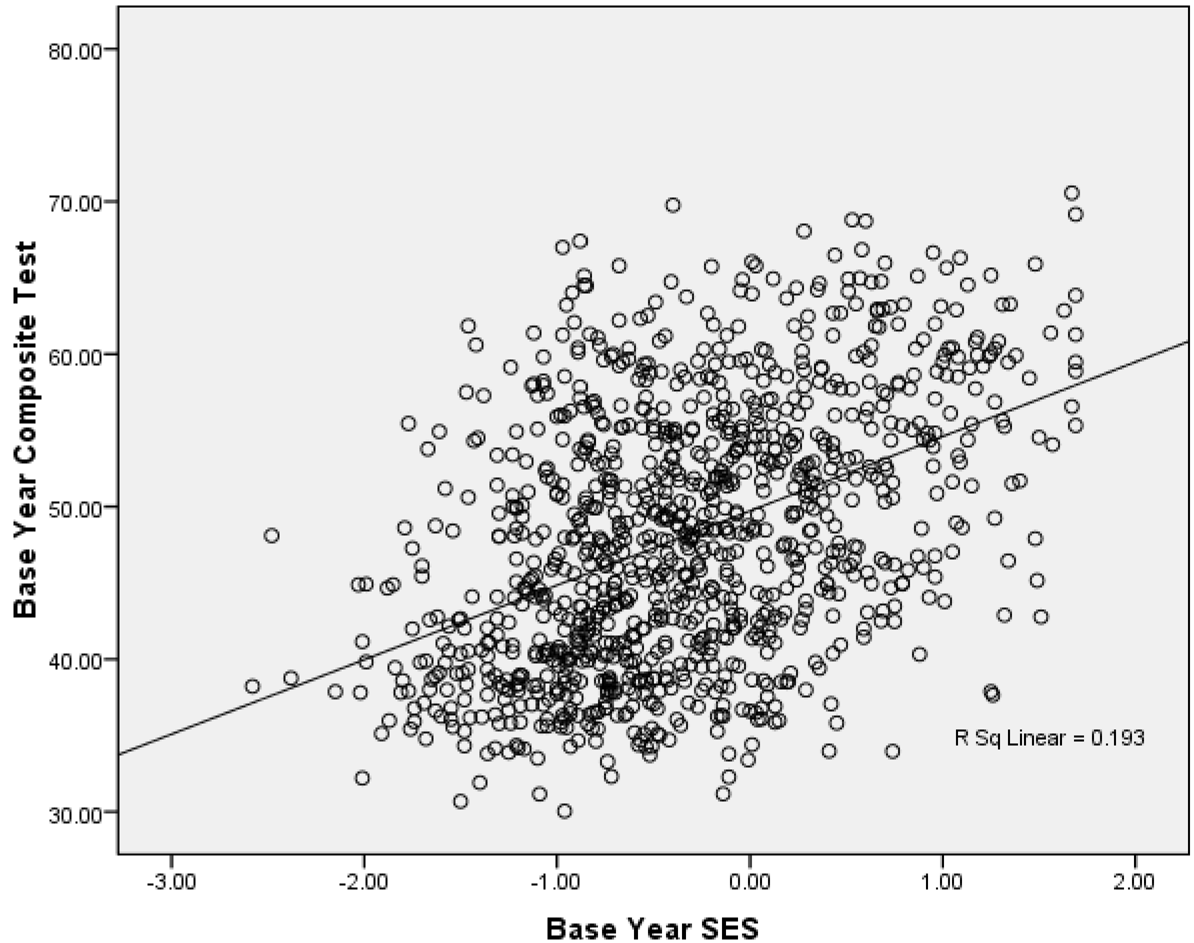
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	49.726	.260		191.448	.000	49.216	50.235
	Base Year SES	4.879	.309	.440	15.795	.000	4.273	5.485

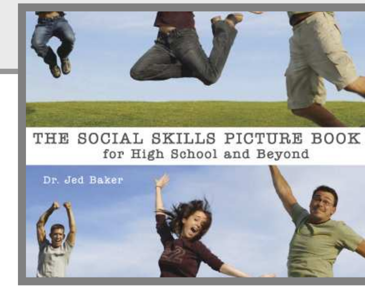
a. Dependent Variable: Base Year Composite Test

High School and Beyond (HSB.sav)





High School and Beyond (HSB.sav)



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.429[a]	.184	.184	7.75965

a. Predictors: (Constant), BY SES, School Avg

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	14176.284	1	14176.284	235.439	.000[a]
	Residual	62741.098	1042	60.212		
	Total	76917.382	1043			

a. Predictors: (Constant), BY SES, School Avg

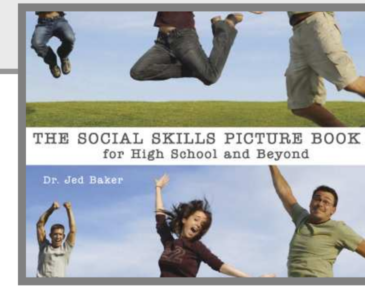
b. Dependent Variable: Base Year Composite Test

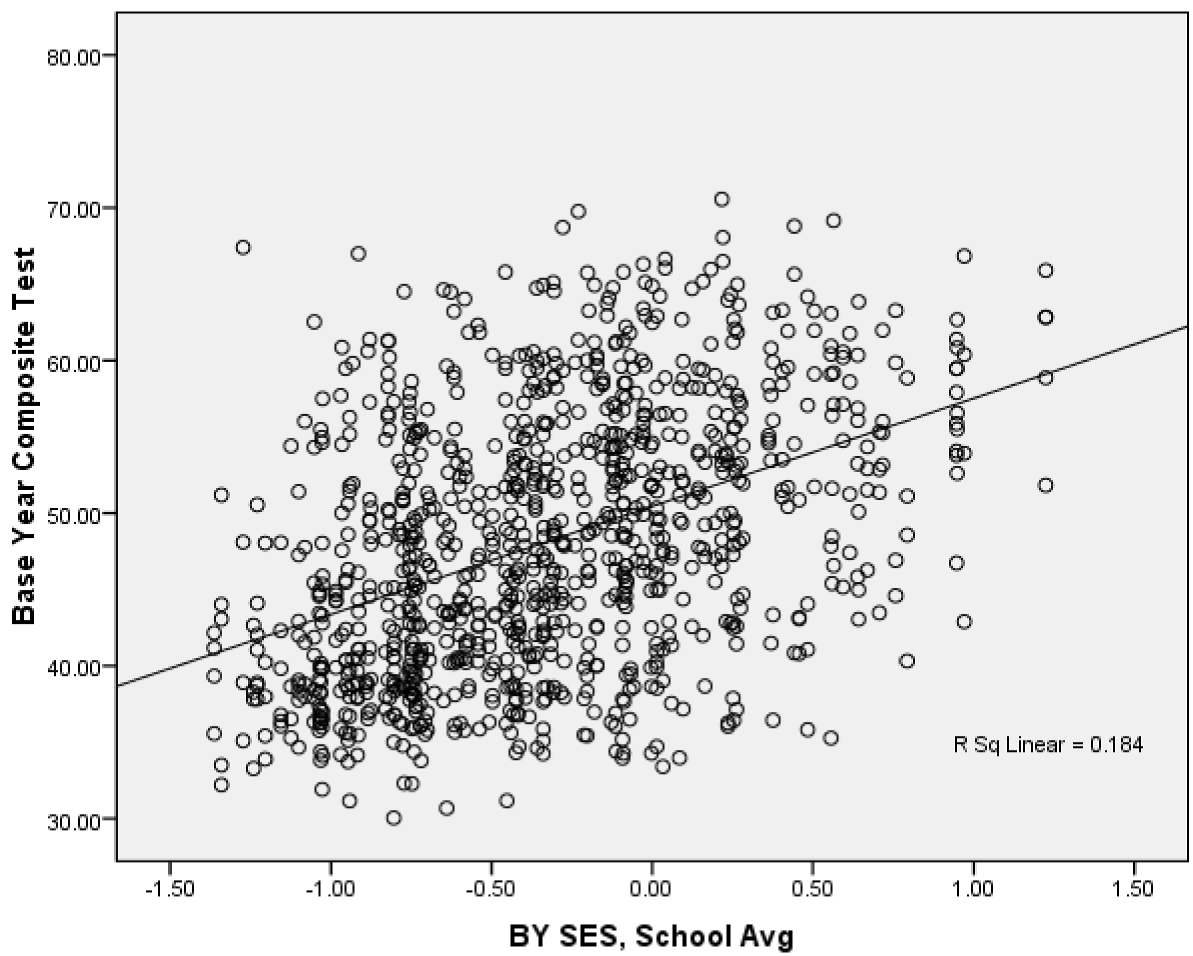
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	50.451	.284		177.397	.000	49.893	51.009
	BY SES, School Avg	7.075	.461	.429	15.344	.000	6.171	7.980

a. Dependent Variable: Base Year Composite Test

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)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.824[a]	.679	.678	.58181

a. Predictors: (Constant), General Reasoning

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	136.226	1	136.226	402.433	.000[a]
	Residual	64.316	190	.339		
	Total	200.542	191			

a. Predictors: (Constant), General Reasoning

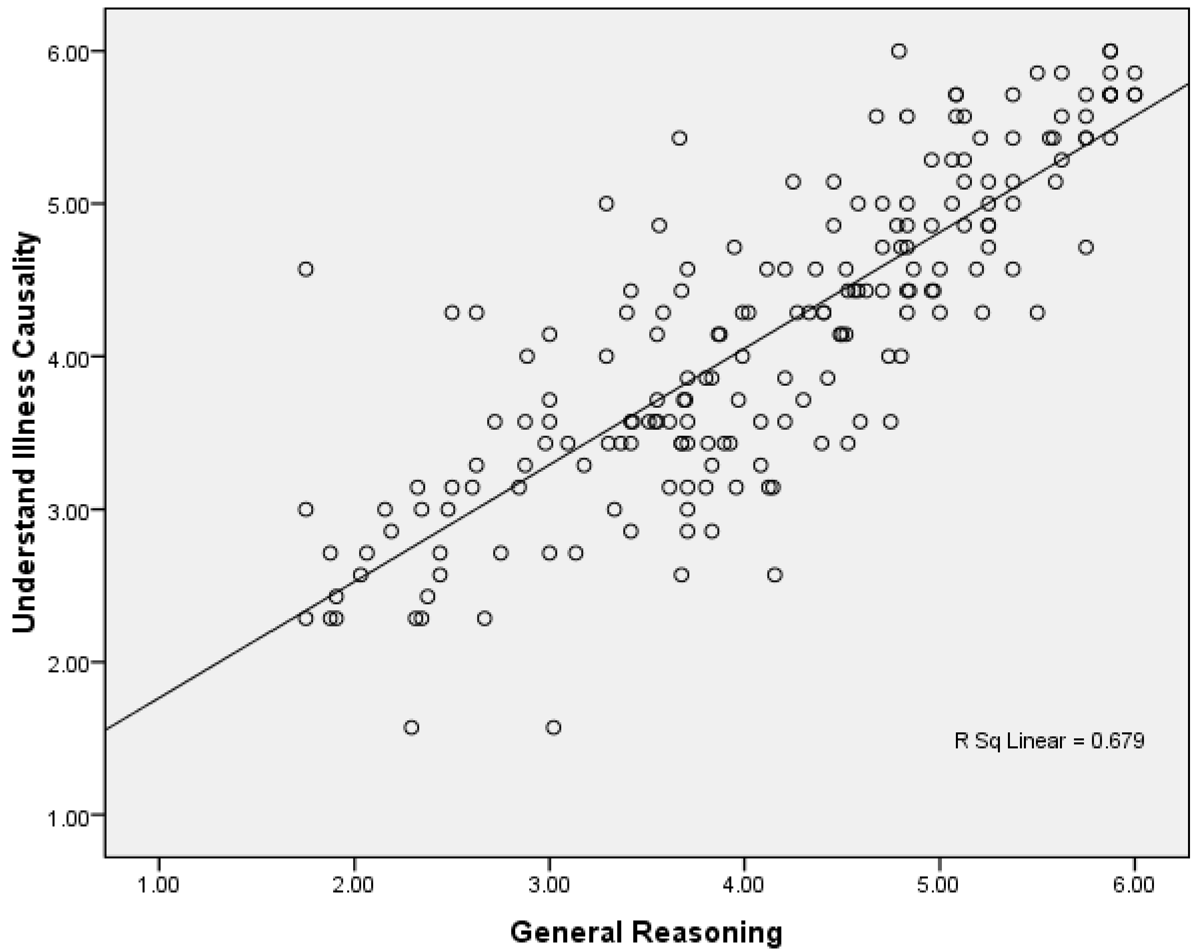
b. Dependent Variable: Understand Illness Causality

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	1.004	.162		6.204	.000	.685	1.323
	General Reasoning	.762	.038	.824	20.061	.000	.687	.837

a. Dependent Variable: Understand Illness Causality

Understanding Causes of Illness (ILLCAUSE.sav)



Understanding Causes of Illness (ILLCAUSE.sav)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.440[a]	.194	.189	.94848

a. Predictors: (Constant), 1 = Asthmatic, 0 = Healthy

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	34.383	1	34.383	38.219	.000[a]
	Residual	143.040	159	.900		
	Total	177.423	160			

a. Predictors: (Constant), 1 = Asthmatic, 0 = Healthy

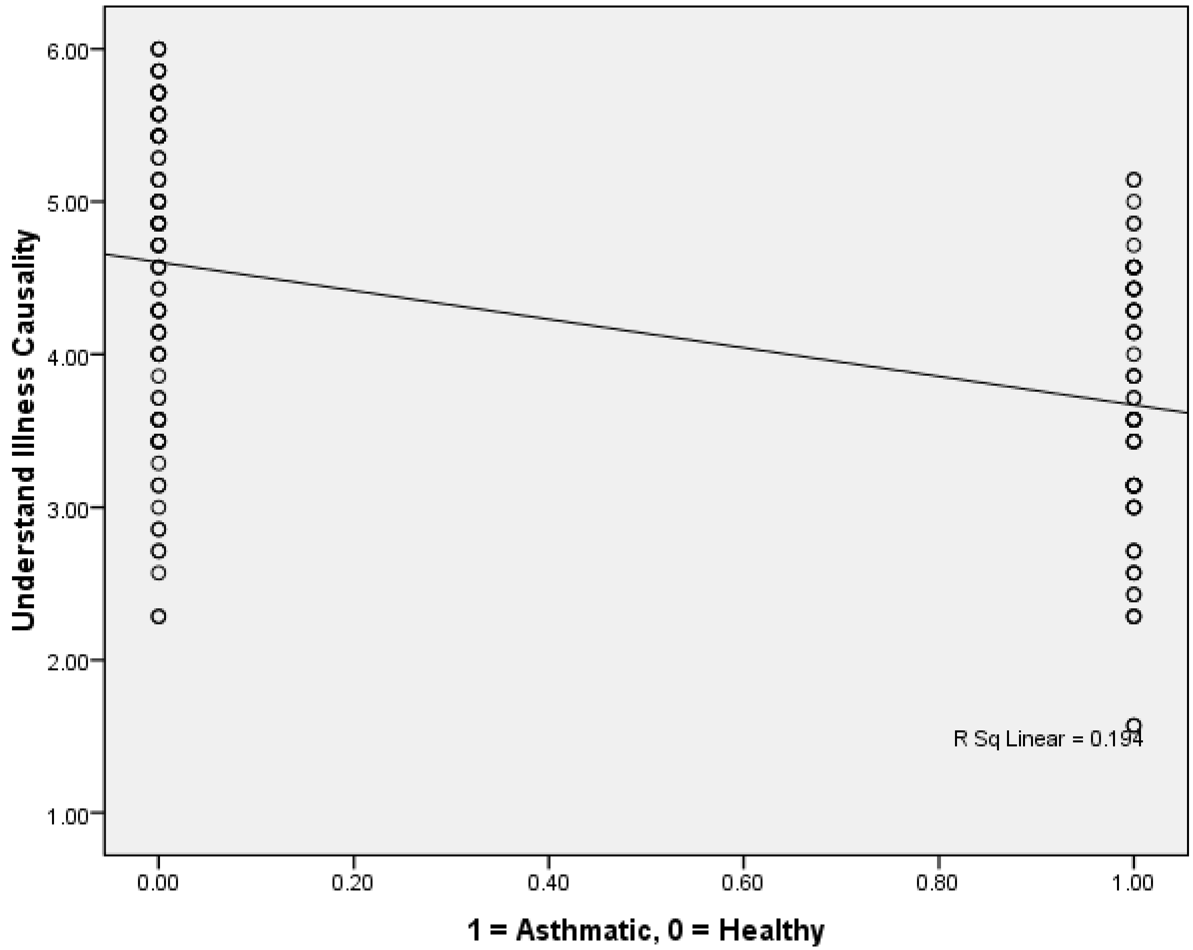
b. Dependent Variable: Understand Illness Causality

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	4.604	.098		46.807	.000	4.409	4.798
	1 = Asthmatic, 0 = Healthy	-.936	.151	-.440	-6.182	.000	-1.234	-.637

a. Dependent Variable: Understand Illness Causality

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)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.353[a]	.125	.124	35.624

a. Predictors: (Constant), % of Students in Child's School Eligible for Free Lunch

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	158680.746	1	158680.746	125.040	.000[a]
	Residual	1114213.431	878	1269.036		
	Total	1272894.177	879			

a. Predictors: (Constant), % of Students in Child's School Eligible for Free Lunch

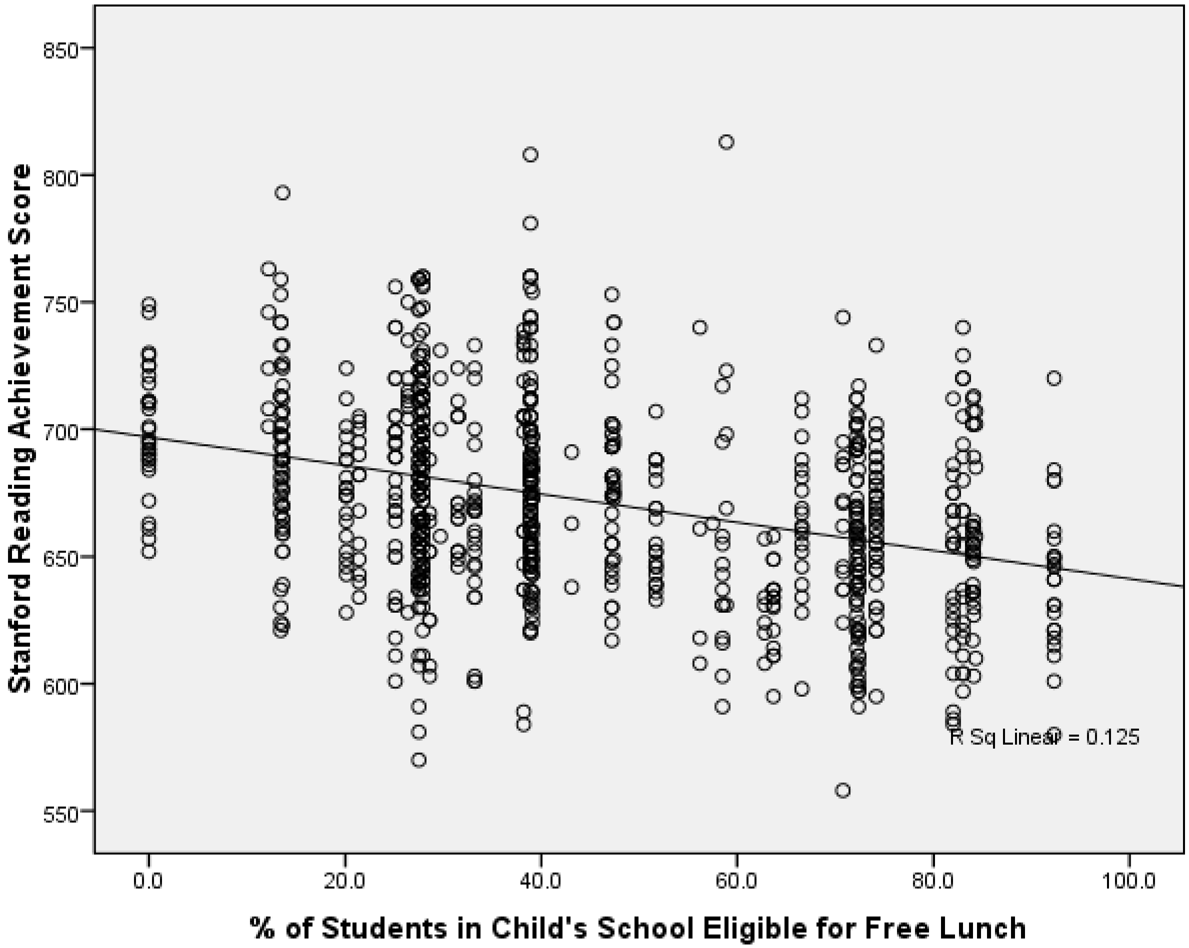
b. Dependent Variable: Stanford Reading Achievement Score

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	696.847	2.540		274.325	.000
	% of Students in Child's School Eligible for Free Lunch	-.555	.050	-.353	-11.182	.000

a. Dependent Variable: Stanford Reading Achievement Score

Children of Immigrants (ChildrenOfImmigrants.sav)



Children of Immigrants (ChildrenOfImmigrants.sav)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.404[a]	.163	.162	34.837

a. Predictors: (Constant), Composite Family SES Score

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	207358.576	1	207358.576	170.863	.000[a]
	Residual	1065535.601	878	1213.594		
	Total	1272894.177	879			

a. Predictors: (Constant), Composite Family SES Score

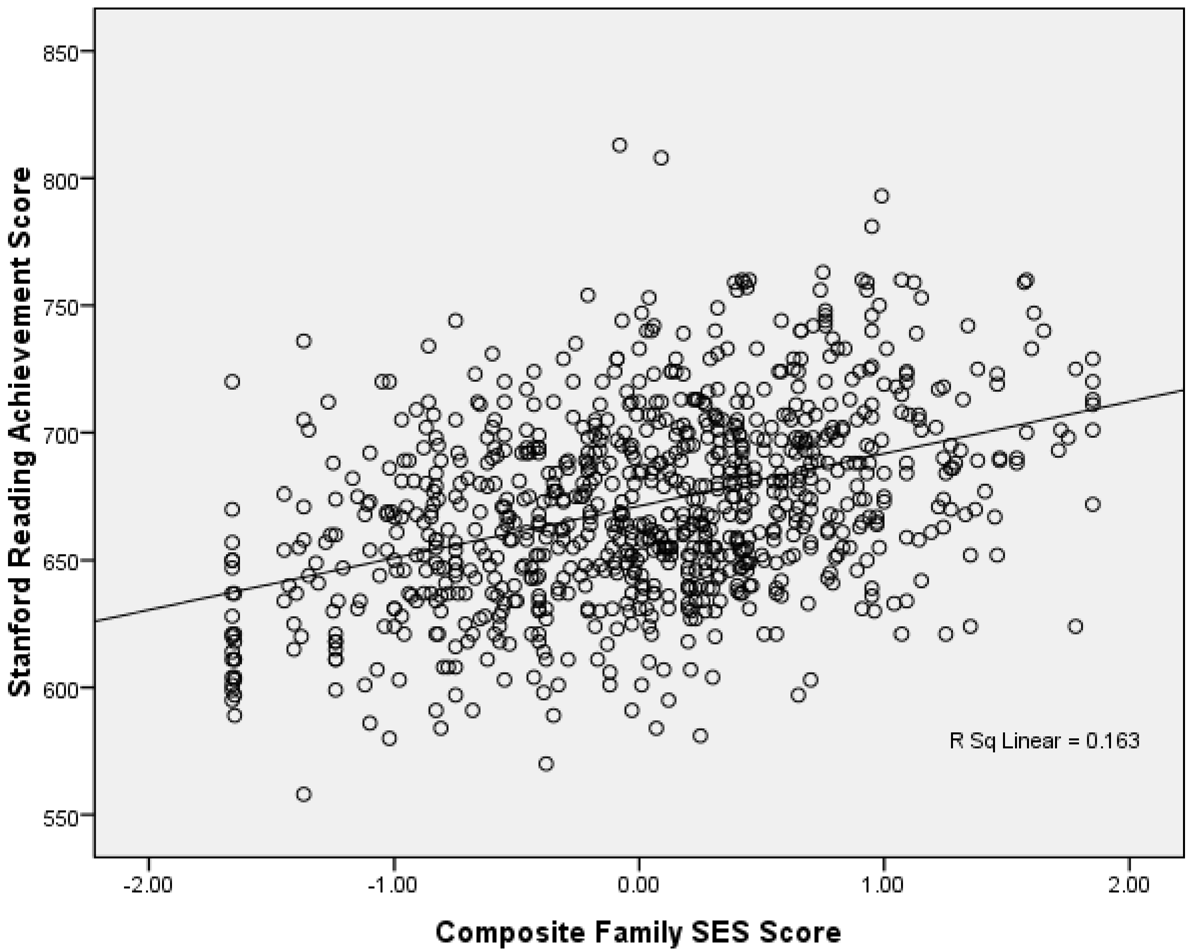
b. Dependent Variable: Stanford Reading Achievement Score

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	671.350	1.175		571.418	.000
	Composite Family SES Score	20.418	1.562	.404	13.071	.000

a. Dependent Variable: Stanford Reading Achievement Score

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)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.382[a]	.146	.143	.91099

a. Predictors: (Constant), Collective efficacy

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	48.191	1	48.191	58.068	.000[a]
	Residual	282.170	340	.830		
	Total	330.361	341			

a. Predictors: (Constant), Collective efficacy

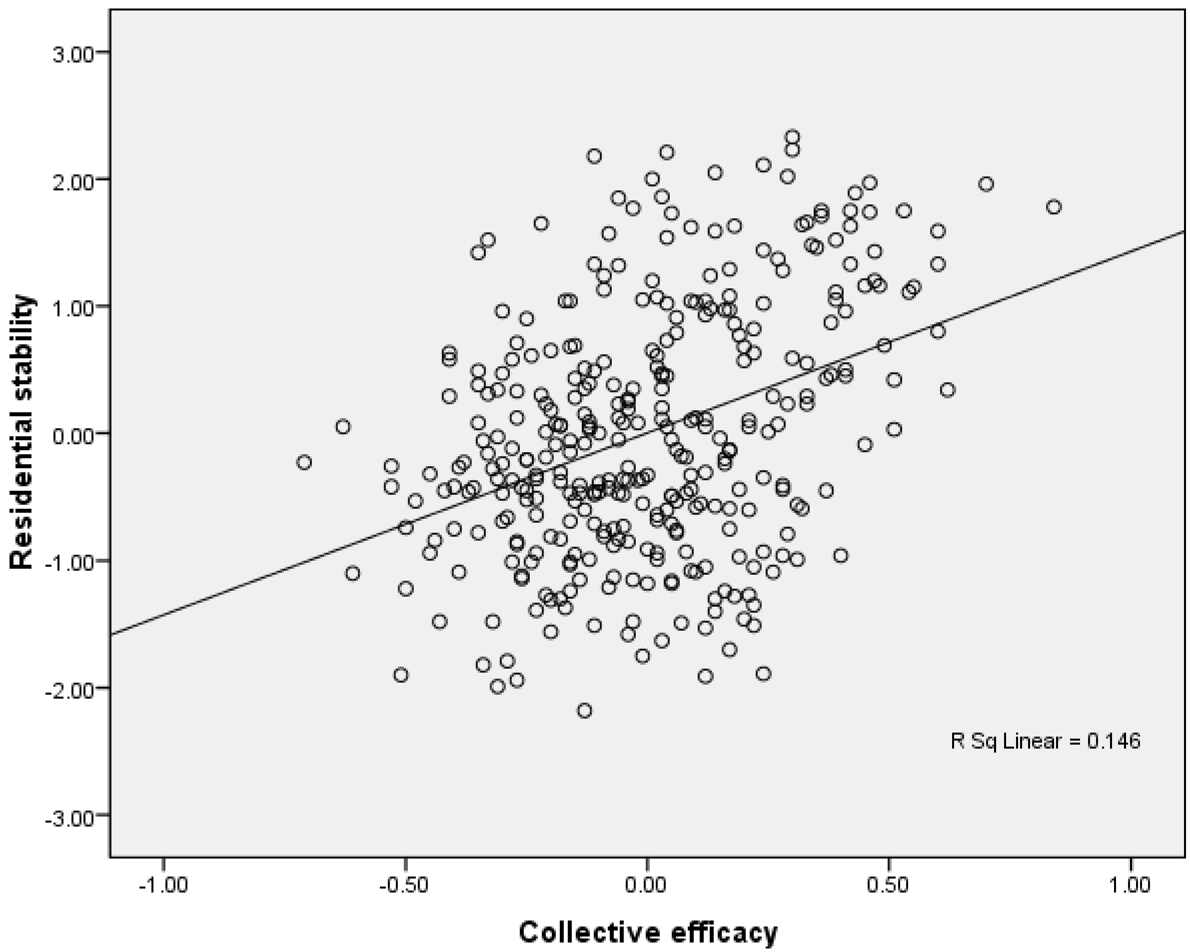
b. Dependent Variable: Residential stability

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.002	.049		.050	.961	-.094	.099
	Collective efficacy	1.429	.187	.382	7.620	.000	1.060	1.797

a. Dependent Variable: Residential stability

Human Development in Chicago Neighborhoods (Neighborhoods.sav)



Human Development in Chicago Neighborhoods (Neighborhoods.sav)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.147[a]	.022	.019	.97506

a. Predictors: (Constant), Homicide rate 1988-90

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7.112	1	7.112	7.480	.007[a]
	Residual	323.249	340	.951		
	Total	330.361	341			

a. Predictors: (Constant), Homicide rate 1988-90

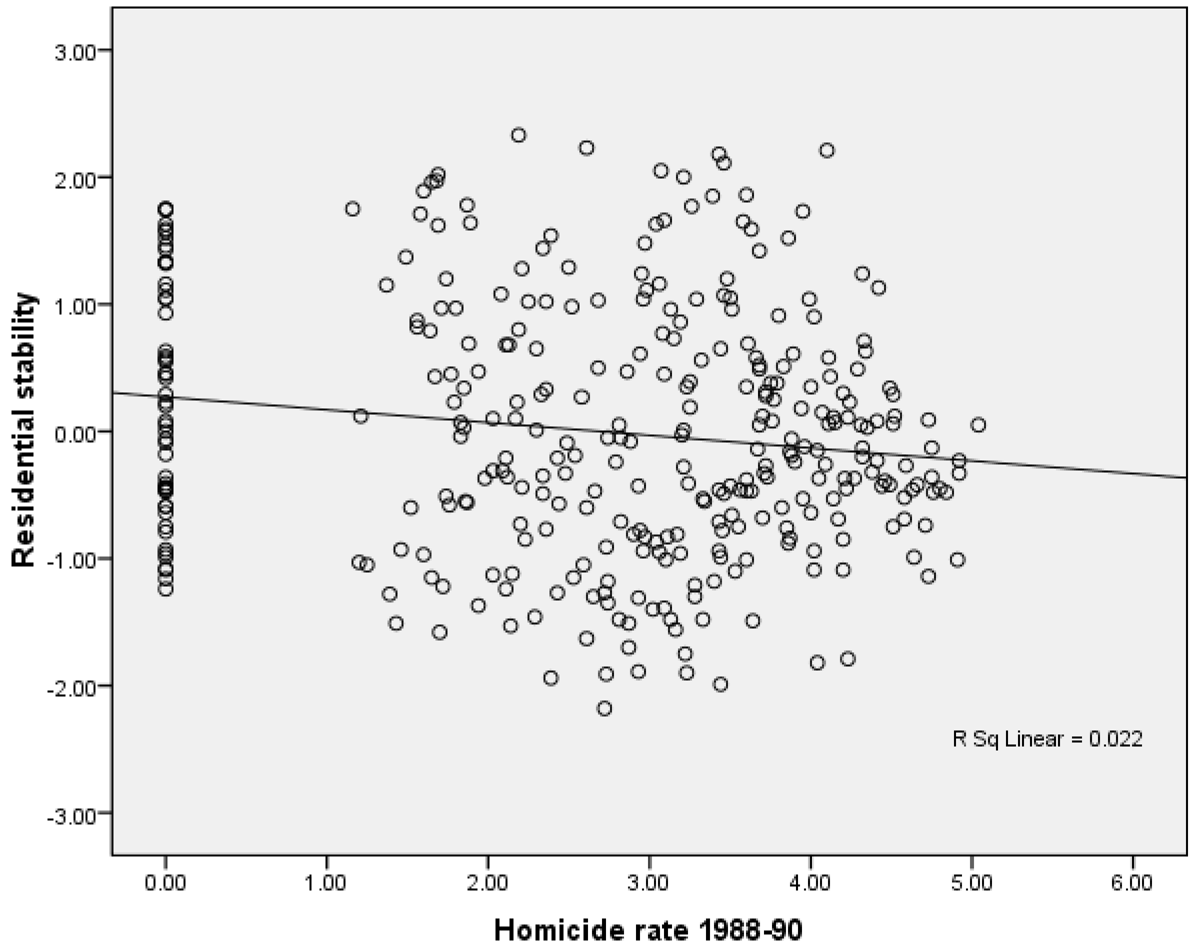
b. Dependent Variable: Residential stability

Coefficients[a]

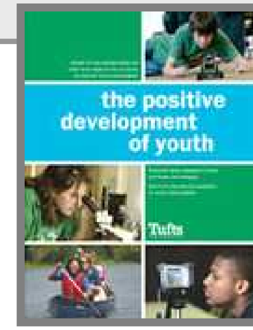
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.270	.111		2.432	.016	.052	.489
	Homicide rate 1988-90	-.100	.037	-.147	-2.735	.007	-.173	-.028

a. Dependent Variable: Residential stability

Human Development in Chicago Neighborhoods (Neighborhoods.sav)



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



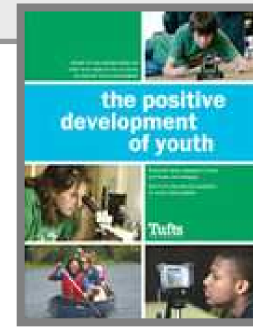
- **4-H Study of Positive Youth Development**
- **Source: Subset of data from IARYD, Tufts University**

- **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)



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.559[a]	.313	.311	.50341

a. Predictors: (Constant), Depression

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	46.912	1	46.912	185.115	.000[a]
	Residual	103.141	407	.253		
	Total	150.053	408			

a. Predictors: (Constant), Depression

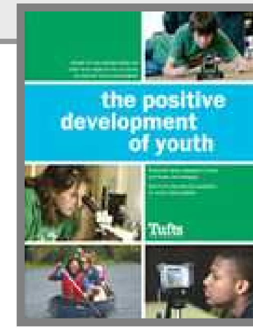
b. Dependent Variable: Self-Worth

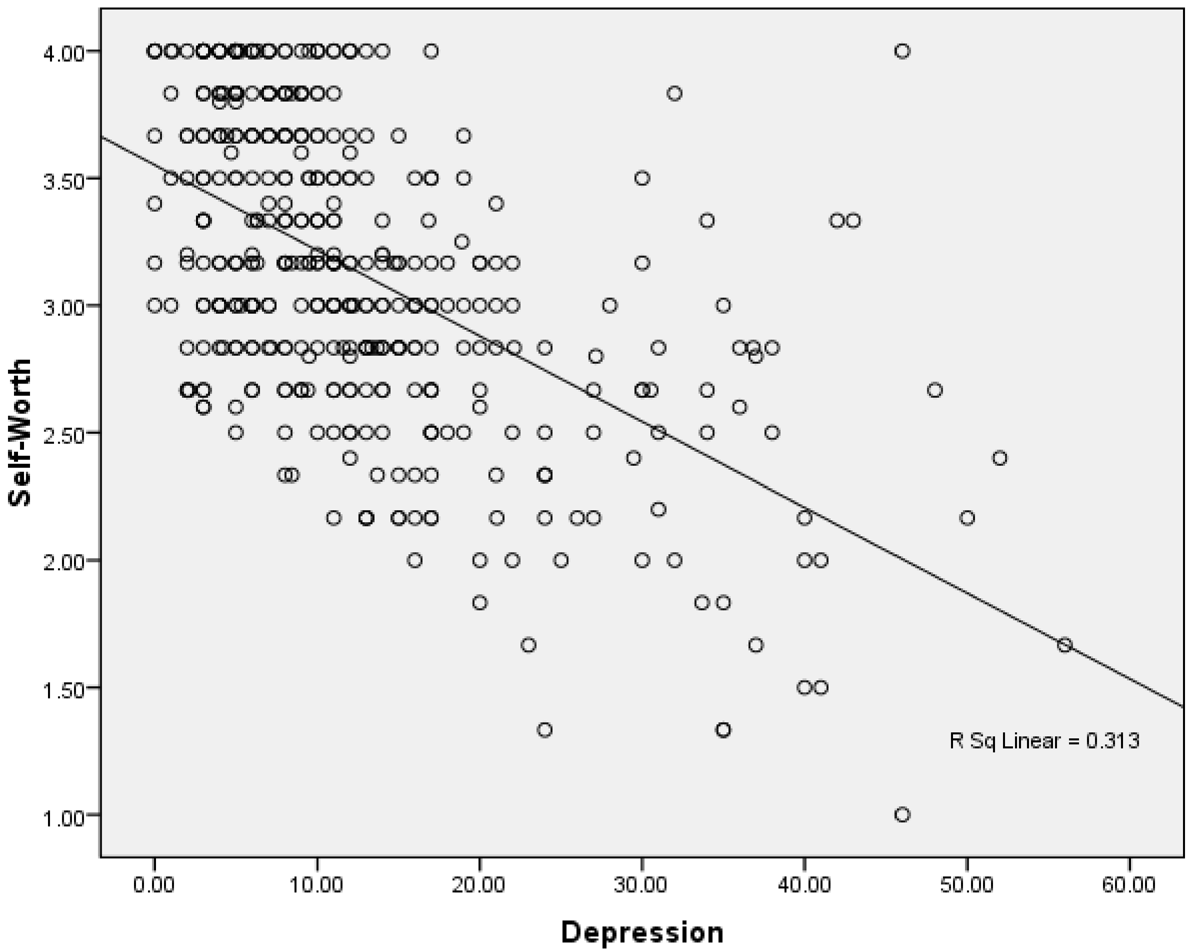
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.552	.040		88.146	.000
	Depression	-.034	.002	-.559	-13.606	.000

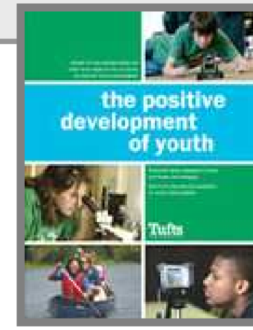
a. Dependent Variable: Self-Worth

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





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



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.504[a]	.254	.252	.52460

a. Predictors: (Constant), Depressed = 1, Not Depressed = 0

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	38.046	1	38.046	138.247	.000[a]
	Residual	112.007	407	.275		
	Total	150.053	408			

a. Predictors: (Constant), Depressed = 1, Not Depressed = 0

b. Dependent Variable: Self-Worth

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.307	.030		108.824	.000
	Depressed = 1, Not Depressed = 0	-.686	.058	-.504	-11.758	.000

a. Dependent Variable: Self-Worth

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