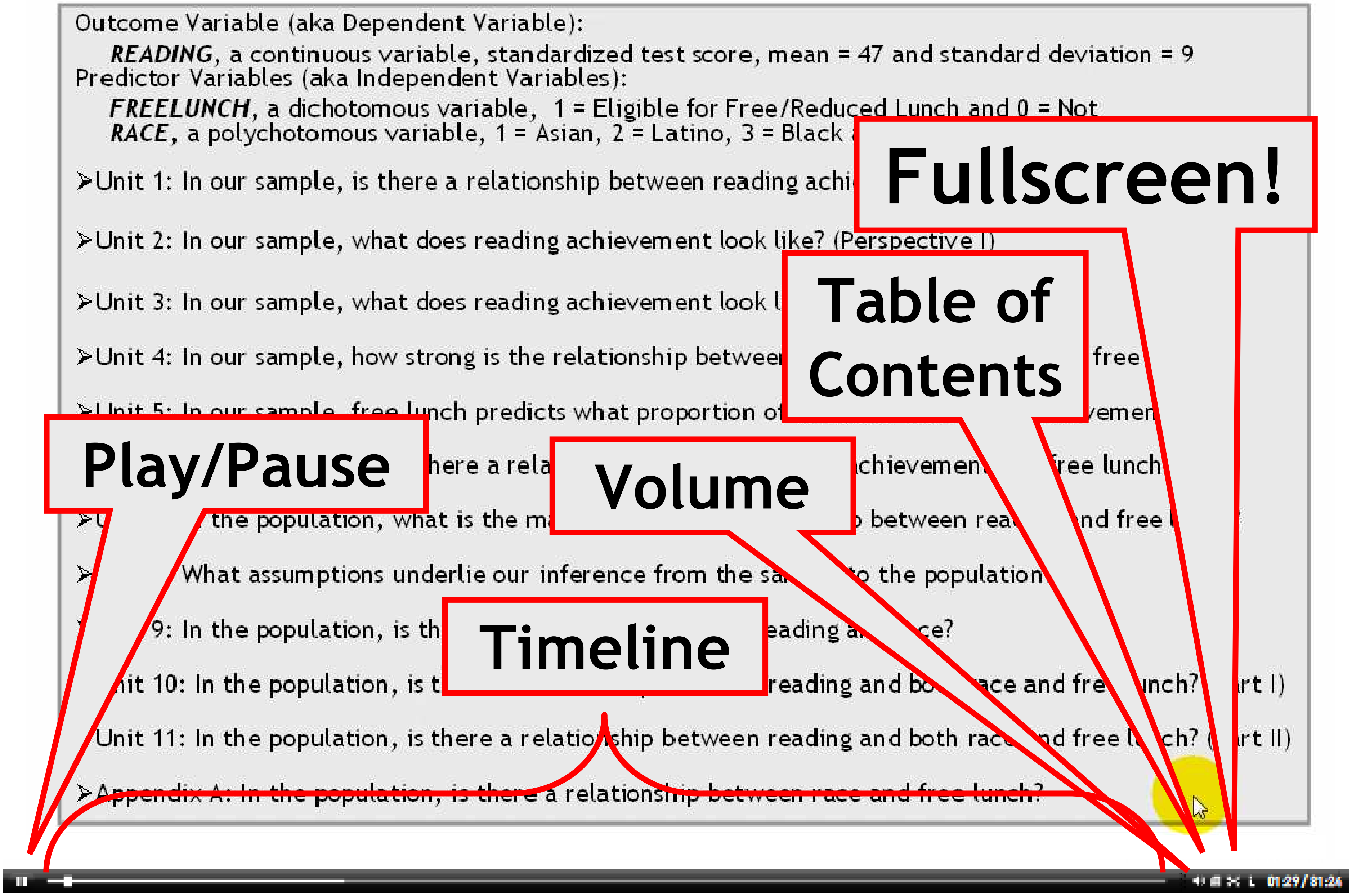
# Resizing Your Viewer



# Unit 1: Introduction to Simple Linear Regression

## Unit 1 Post Hole:

Use exploratory data analytic techniques to investigate the relationship between two variables.

## Unit 1 Technical Memo and School Board Memo:

Conduct two bivariate exploratory data analyses (with one continuous outcome, one continuous predictor and one dichotomous predictor of your choice).

# Unit 1: 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 1: 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**

# Memo Metacognitive Template

| |
|---|
| **Memo #1: Introduction to Simple Linear Regression** |
| Include Your Individual Technical Draft: |
| (Draft your individual technical memo here, or cut and paste it here from the word processor of your choice.) |
| Include Your Individual School Board Draft: |
| (Draft your individual school board memo here, or cut and paste it here from the word processor of your choice.) |
| Time Spent Outside Of Class On The Individual Memos: |
| •Programming: 0? Hours<br>•Technical Draft: 2.25? Hours<br>•School Board Draft: 0.75? Hours<br>•Time Sinks: 0? Hours  If so, what were they? |
| Comments, Questions, Concerns, Complaints, Compliments: |
| |
| Include Your Syntax: |
| |

"Metacognition" is thinking about thinking. I ask you to complete the memo *metacognitive* because not only do I want you to think about the memos but also I want you to think about your thinking. I want you to consider time sinks: What was valuable work, and what was busy work? I also want you take the opportunity to make any comments, ask any questions, voice any concerns, log any complaints and/or serve any compliments.

# Memo Metacognitive Exemplar (Part I of II)

**Memo #1: Introduction to Simple Linear Regression (Exemplar)**

Include Your Individual Technical Draft:

**Date: January 20, 2011**
**To: High School Data Team**
**From: Sean Parker**
**Subject: ELA MCAS Performance, A Preliminary Look**

**Extracurricular Activities and ELA MCAS Performance**

Introduction:
We theorize that participation in extracurricular activities builds school engagement and, in turn, school engagement builds school success, as measurable by the MCAS. Consequently, we hypothesize that participation in extracurricular activities will be positively correlated with ELA MCAS scores.
Outcome: ELAMCAS
Predictor: EXTRAC
Model: $ELAMCAS = \beta_0 + \beta_1 EXTRAC + \varepsilon$
Fitted Model: $\widehat{ELAMCAS} = 249 + 6.5EXTRAC$

Exploratory Data Analysis:
A linear model is appropriate for the relationship between ELAMCAS and EXTRAC because ELAMCAS is a continuous outcome and EXTRAC is a dichotomous predictor. In our sample, the relationship is positive such that students who participate in extracurricular activities on average outperform their non-participating counterparts by about 6.5 points on the ELA MCAS test. On average, students who participate extracurricular activities score 255.5 on the ELA MCAS, and students who do not participate score 249 on the ELA MCAS. This finding is consistent with our hypothesis. We can't, however, conclude that participation *causes* improved MCAS scores, because, for example, we don't know if students who participate in extracurricular activities are more proficient in ELA to begin with. Nevertheless, we observe a noteworthy trend in our sample, because 6.5 points on the ELA MCAS is substantial. The trend is not so strong that we would ever feel comfortable making predictions about individuals. In both groups (participants and non-participants) there is a wide range of scores such that many of the extracurricular participants score at the bottom and many of the non-participants score at the top. Because the relationship is so weak, there are no outliers that stand out from the pack (since, as per the weakness, the pack is already spread out wide). So, we see our predicted pattern in *group* performance, not *individual* performance.

**Absences and ELA MCAS Performance**

Introduction:
Students need to attend school in order to master the material. We ask: What is the magnitude of the presumably negative relationship between ELA MCAS scores and school absences?
Outcome: ELAMCAS
Predictor: ABSENCES
Model: $ELAMCAS = \beta_0 + \beta_1 ABSENCES + \varepsilon$
Fitted Model: $\widehat{ELAMCAS} = 255 + 0.2ABSENCES$

Exploratory Data Analysis:
The relationship between ELAMCAS and ABSENCES is negative and linear. As we hypothesized, students with more absences tend to do worse on the ELA MCAS. On average, students who differ in absences by 10 days tend to differ in ELA MCAS scores by 2 points. For example, we predict that a student with no absence will score a 255, but we predict that a student with 10 absences will score a 253. Frankly, we find this difference surprisingly small. There is a lot of variation in ELA MCAS performance even looking at students with the same number of absences, so the relationship is not particularly strong. No observation jumps out as an extreme outlier, but one observation may shed light on why the relationship seems so trivial. The only student with a perfect ELA MCAS score of 280 had 10 absences. Based on her absences we predicted that the student would score 253 but she in fact scored a 280. This prompts the question: why was she absent? Did she skip school? Did she go on safari with her family? Did she have a medical operation? There are many reasons why a student might be absent, and some reasons might be educationally beneficial (e.g., safari).

# Memo Metacognitive Exemplar (Part II of II)

## Include Your Individual School Board Draft:

**Date: January 20, 2011**
**To: School Board**
**From: Sean Parker**
**Subject: ELA MCAS Performance and Attendance**

In order to inform the reassessment of the high school attendance policy, we examined the relationship between ELA MCAS performance and attendance. We found that students with more absences tended to score lower on the ELA MCAS, but not by much. The average student with no absences scored 255, and the average student with 10 absences scored 253. This was a consistent pattern. In general, students who differed by 10 absences tended to differ by 2 points on the ELA MCAS. The numbers tell us *that* there is relationship, but they do not tell us *why* there is a relationship. Our next step will be to look at students who break the pattern to get insight into why there is a pattern. We will start by interviewing high-absence students who scored high on the ELA MCAS.

## Time Spent Outside Of Class On The Individual Memos:

- Programming: 0 Hours
- Technical Draft: 2.0 Hours
- School Board Draft: 0.5 Hours
- Time Sinks: .25 Hours If so, what were they? I had trouble with the equation editor. After 15 minutes, I quit fidgeting, because it seemed like a trick I could pick up later, as opposed to a real conceptual issue worth the struggle. Happily, my teammate, Jen, showed me how later.

## Comments, Questions, Concerns, Complaints, Compliments:

This stuff was really hard to put into words. Even when I had the post hole sketched out, there was still a lot of work to do!

## Include Your Syntax:

```
plot(ELAMCAS~EXTRAC, data=HS)
lm(ELAMCAS~EXTRAC , data=HS))
plot(HS$ELAMCAS~HS$ABSENCES)
lm(HS$ELAMCAS~HS$ABSENCES)
```

# Unit 1: Road Map (VERBAL)

Nationally Representative Sample of 7,800 8$^{th}$ 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 1: 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 1: 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

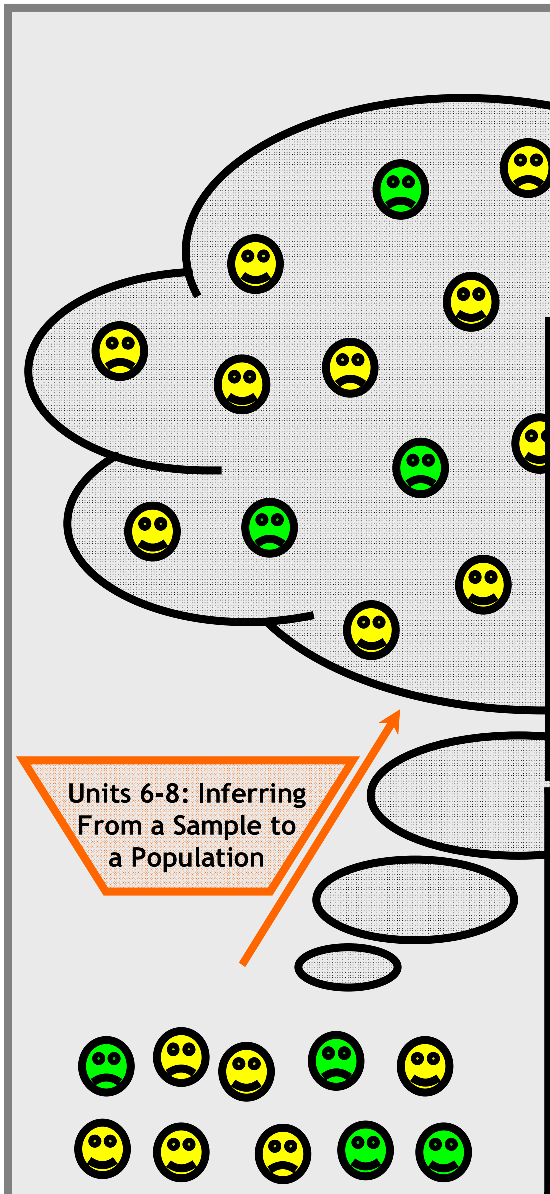
**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | | 95% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. | 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 1: Road Map (Schematic)




| | | Single Predictor | | |
|---|---|---|---|---|
| | | Continuous | Polychotomous | Dichotomous |
| Outcome | Continuous | Regression | Regression<br>ANOVA | Regression<br>ANOVA<br>T-tests |
| | Polychotomous | Logistic Regression | Chi Squares | Chi Squares |
| | Dichotomous | | Chi Squares | Chi Squares |

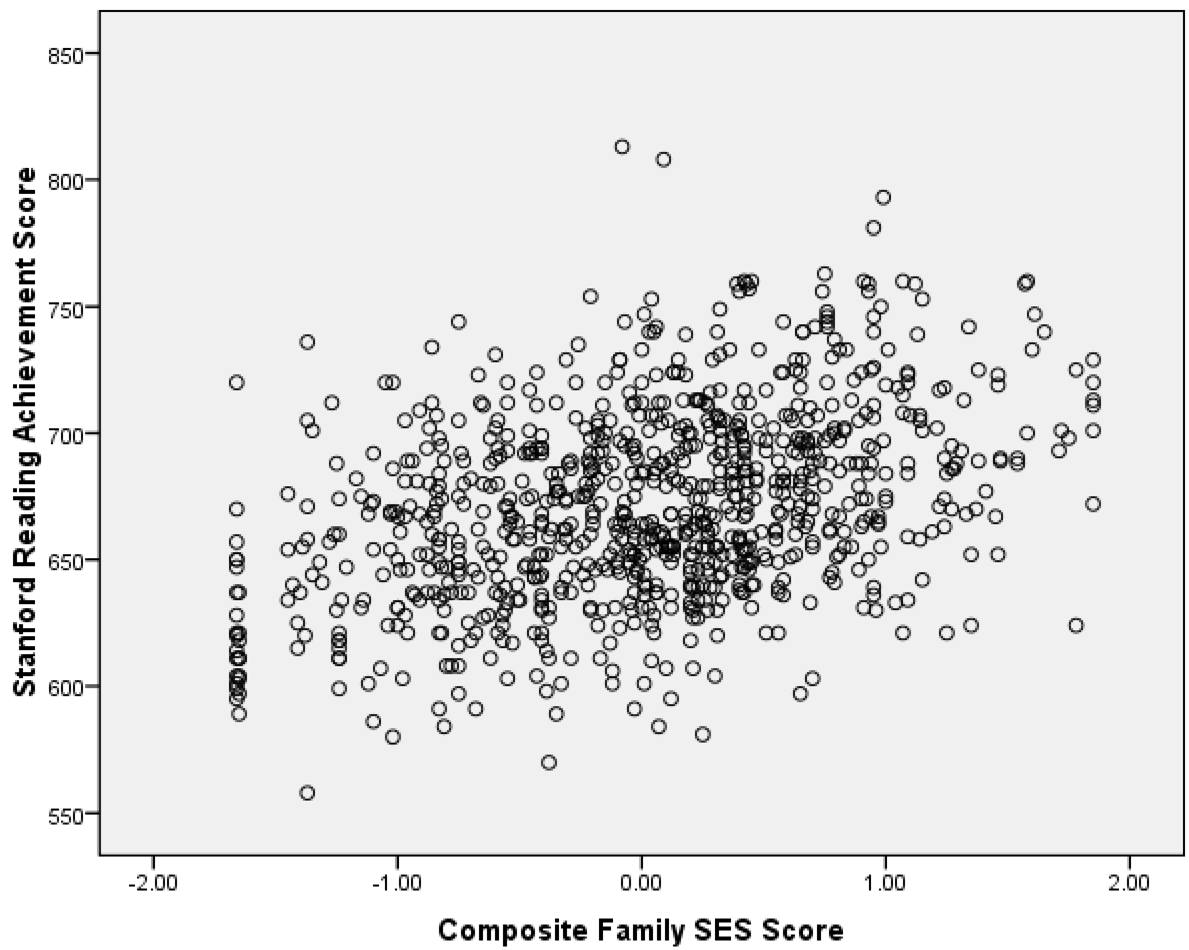
| | | Multiple Predictors | | |
|---|---|---|---|---|
| | | Continuous | Polychotomous | Dichotomous |
| Outcome | Continuous | Multiple Regression | Regression<br>ANOVA | Regression<br>ANOVA |
| | Polychotomous | Logistic Regression | Chi Squares | Chi Squares |
| | Dichotomous | | Chi Squares | Chi Squares |

# Epistemological Minute

In epistemology, the <u>curve-fitting problem</u> challenges us to consider the role of simplicity (or “parsimony”) in our theorizing. In this course, we are going to spend much of our time fitting linear models. Lines happen to be the simplest of curves; in fact, lines are so simple that most of us probably hesitate to consider them a kind of curve.

Throughout this course, I invite you to consider the advantages and disadvantages of fitting lines to data. What do we gain? What do we lose?

Figure 1. A bivariate scatterplot depicting the relationship between Stanford Reading Achievement scores and socioeconomic status for a sample of 8th and 9th-grade children of immigrants (n = 880).

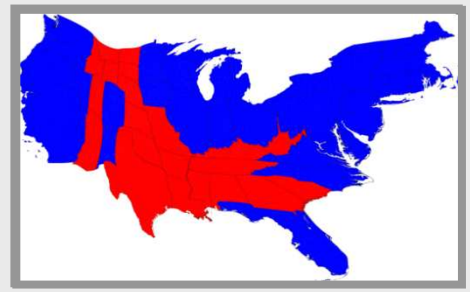


# Unit 1: Research Question

Theory: For interstate comparisons, SAT scores are deceptive because the relative number of test takers varies so widely from state to state. In particular, states with a low percentage of test takers will fare best since only the best of the best students comprise that low percentage.

Research Question: In 1994, were state average SAT scores negatively correlated with percentage of eligible students who take the SAT?

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



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

Variables:

Outcome—State Average SAT Score (*SAT*)

Predictor—%age of Eligible Students who Take the SAT (*PERCENT*)

Model: $SAT = \beta_0 + \beta_1 PERCENT + \varepsilon$

# 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
- SAT: average total score on the SAT in 1994-95

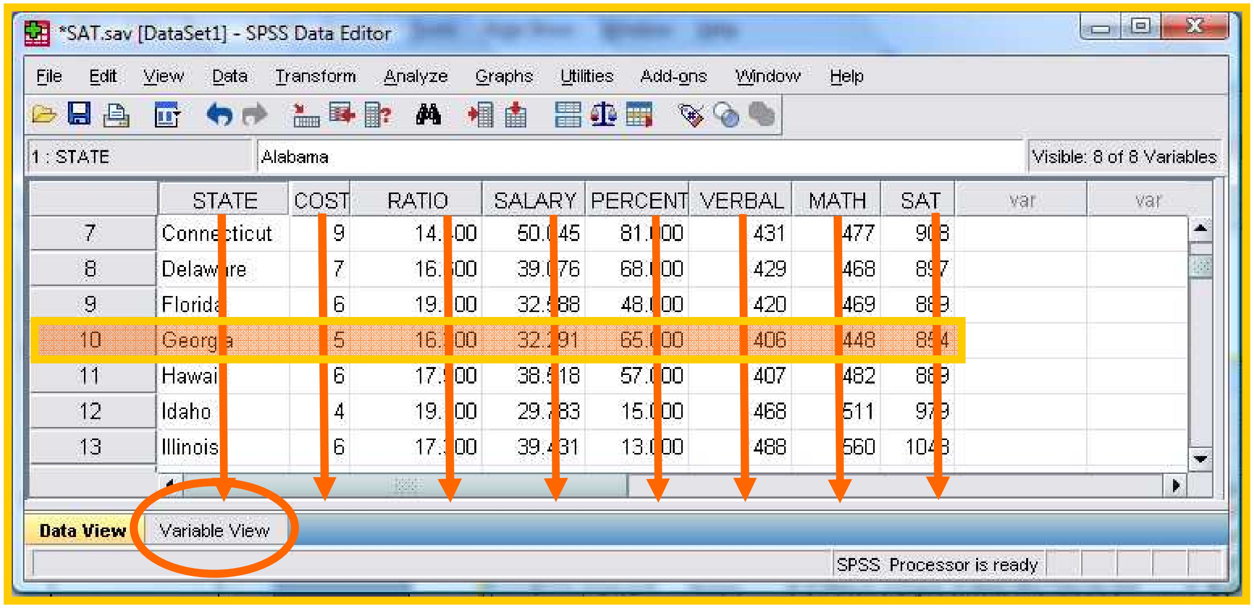
# The SAT Data Set (R)

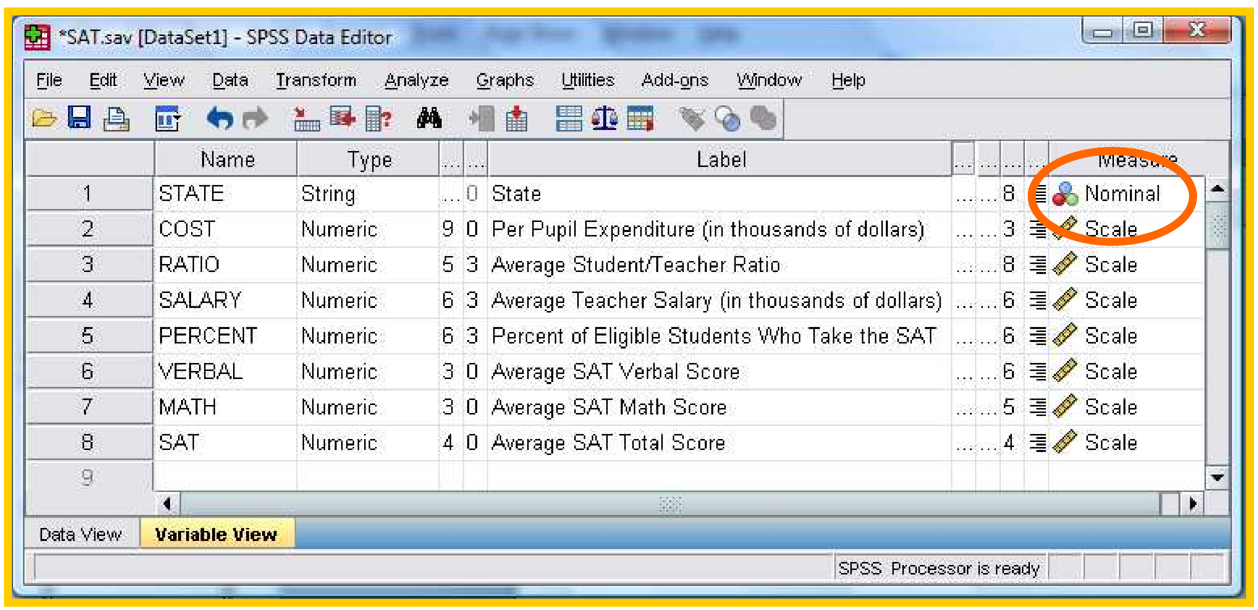
SAT

| | STATE | COST | RATIO | SALARY | PERCENT | VERBAL | MATH | TOTAL |
|---|---|---|---|---|---|---|---|---|
| 16 | Delaware | 7.030 | 16.6 | 39.076 | 68 | 429 | 468 | 897 |
| 17 | Florida | 5.718 | 19.1 | 32.588 | 48 | 420 | 469 | 889 |
| 18 | Georgia | 5.193 | 16.3 | 32.291 | 65 | 406 | 448 | 854 |
| 19 | Hawaii | 6.078 | 17.9 | 38.518 | 57 | 407 | 482 | 889 |
| 20 | Idaho | 4.210 | 19.1 | 29.783 | 15 | 468 | 511 | 979 |
| 21 | Illinois | 6.136 | 17.3 | 39.431 | 13 | 488 | 560 | 1048 |
| 22 | Indiana | 5.826 | 17.5 | 36.785 | 58 | 415 | 467 | 882 |
| 23 | Iowa | 5.483 | 15.8 | 31.511 | 5 | 516 | 583 | 1099 |

```
> str(SAT)
'data.frame': 50 obs. of 8 variables:
 $ STATE      : Factor w/ 50 levels "Alabama      ",..: 7 19 21 31 32 34 38 39 45 1 ...
 $ COST       : num  8.82 6.43 7.29 5.86 9.77 ...
 $ RATIO      : num  14.4 13.8 14.8 15.6 13.8 15.2 17.1 14.7 13.8 17.2 ...
 $ SALARY     : num  50 32 40.8 34.7 46.1 ...
 $ PERCENT    : num  81 68 80 70 70 74 70 70 68 8 ...
 $ VERBAL     : num  431 427 430 444 420 419 419 425 429 491 ...
 $ MATH       : num  477 469 477 491 478 473 461 463 472 538 ...
 $ TOTAL      : num  908 896 907 935 898 ...
```

# The SAT Data Set (SPSS)




Ray Charles says, “Keep **Georgia** on your mind, where **65%** of eligible students take the SAT, and the average combined SAT score is **854**.”

Data Take Different Forms (Scales):

**Nominal Scales** take values (either numeric or string) that stand for categories and that have no other meaning. Nominal variables include *FEMALE* and *RACE*.

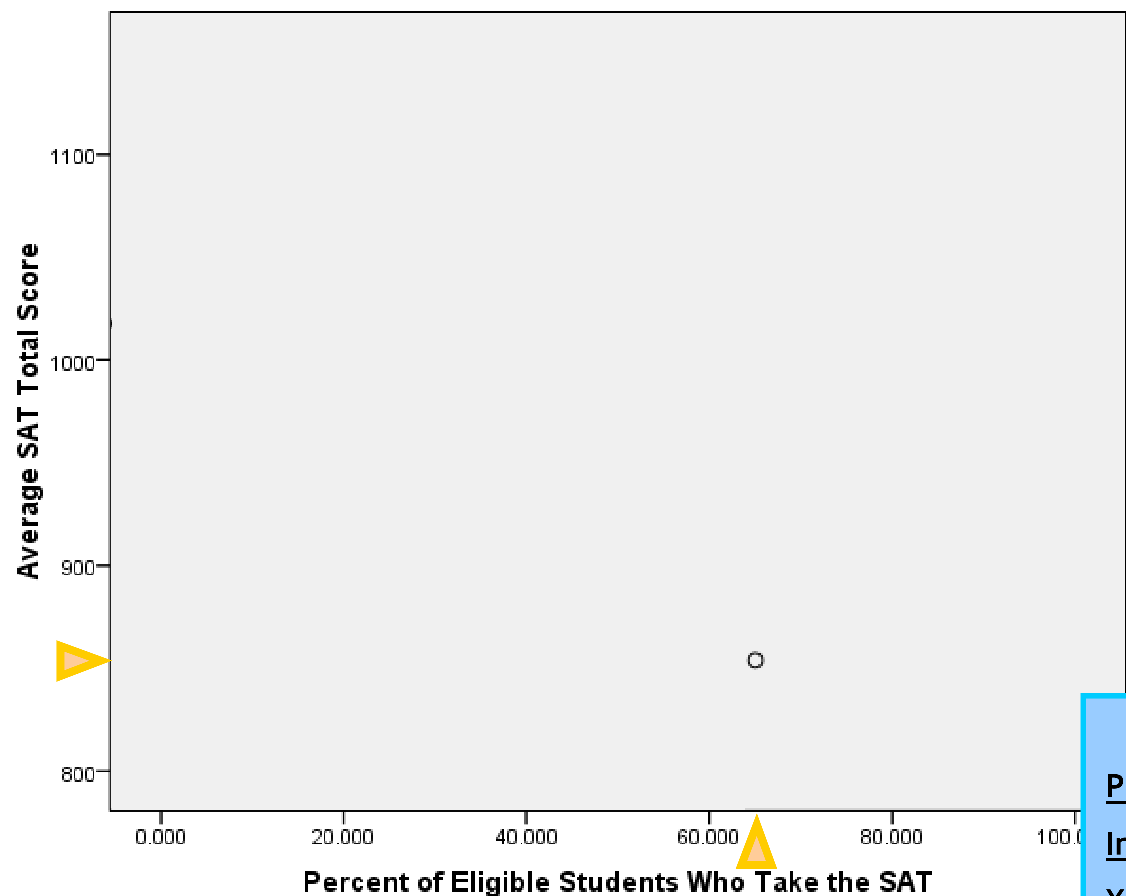
**Ordinal Scales** take numeric values that stand for ranked/ordered categories.

**Interval and Ratio Scales** take numeric values for which equal numeric intervals represent equal amounts of the construct. In ratio scales, zero means zero of the construct. Celsius and Fahrenheit are interval scales for temperature, whereas Kelvin is a ratio scale. (In order to confuse things, SPSS calls both types of scales “Scale.”) In this course, we will work exclusively (until Appendix A) with interval and ratio outcomes, but we will learn to work with predictors that are nominal, ordinal, interval or ratio.

# Bivariate Scatterplots (SPSS)

Figure A: Bivariate scatterplot of average SAT total score for a state versus percent of eligible students who take the SAT (n = 50 states).

Ray Charles says, “Keep **Georgia** on your mind, where **65%** of eligible students take the SAT, and the average combined SAT score is **854**.”




Synonyms:

**Outcome Variable**

**Dependent Variable**

**Y-Axis Variable**

**Y Variable**

Synonyms:

**Predictor Variable**

**Independent Variable**

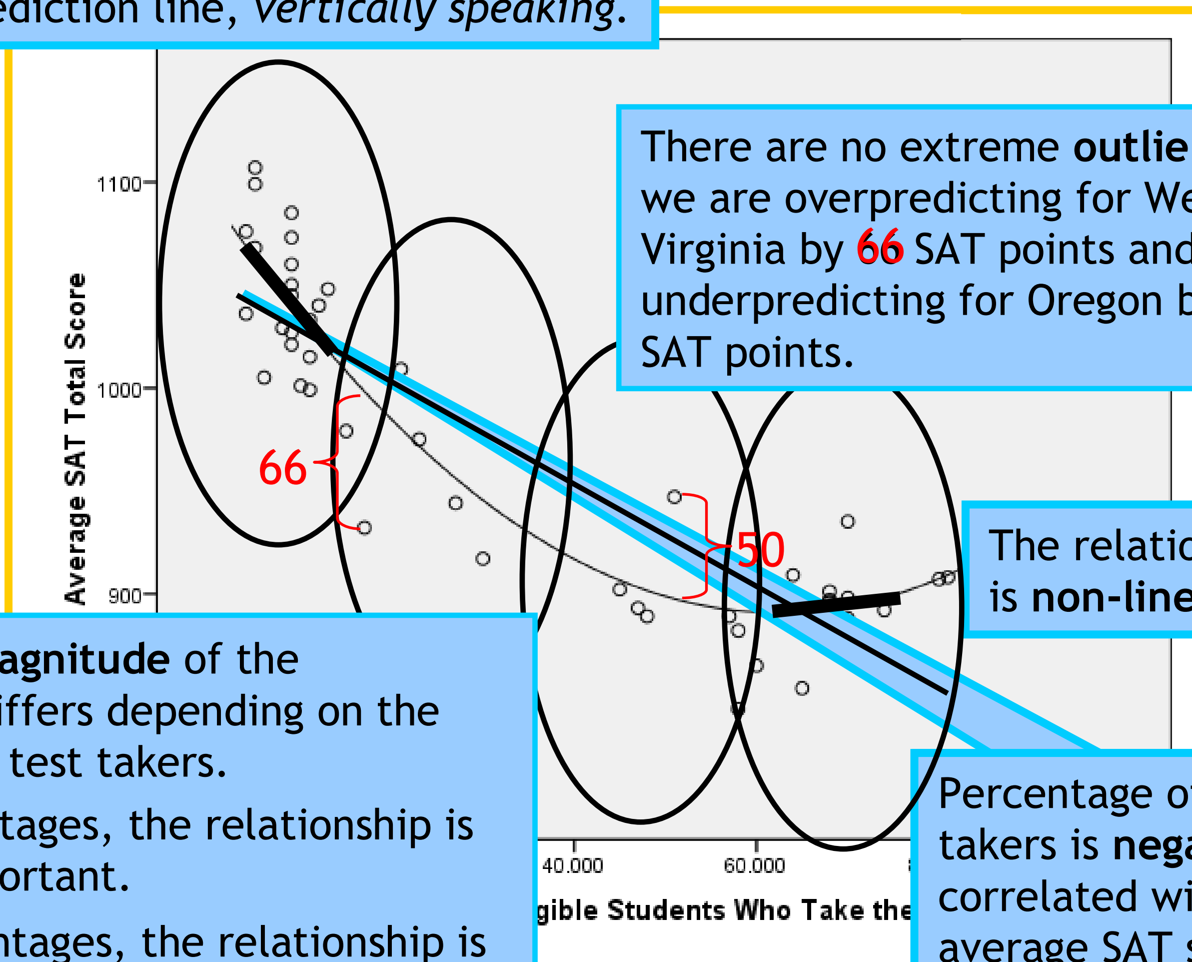
**X-Axis Variable**

**X Variable**

# Bivariate Exploratory Data Analysis

Figure A: Bivariate scatterplot of average SAT ... of eligible ... tes).

Our primary goal is to predict on average. The relationship is **strong**, because the data tend to hug our prediction line, *vertically speaking*.




There are no extreme **outliers**, but we are overpredicting for West Virginia by **66** SAT points and underpredicting for Oregon by **50** SAT points.

The relationship is **non-linear**.

Finally, the **magnitude** of the relationship differs depending on the percentage of test takers.

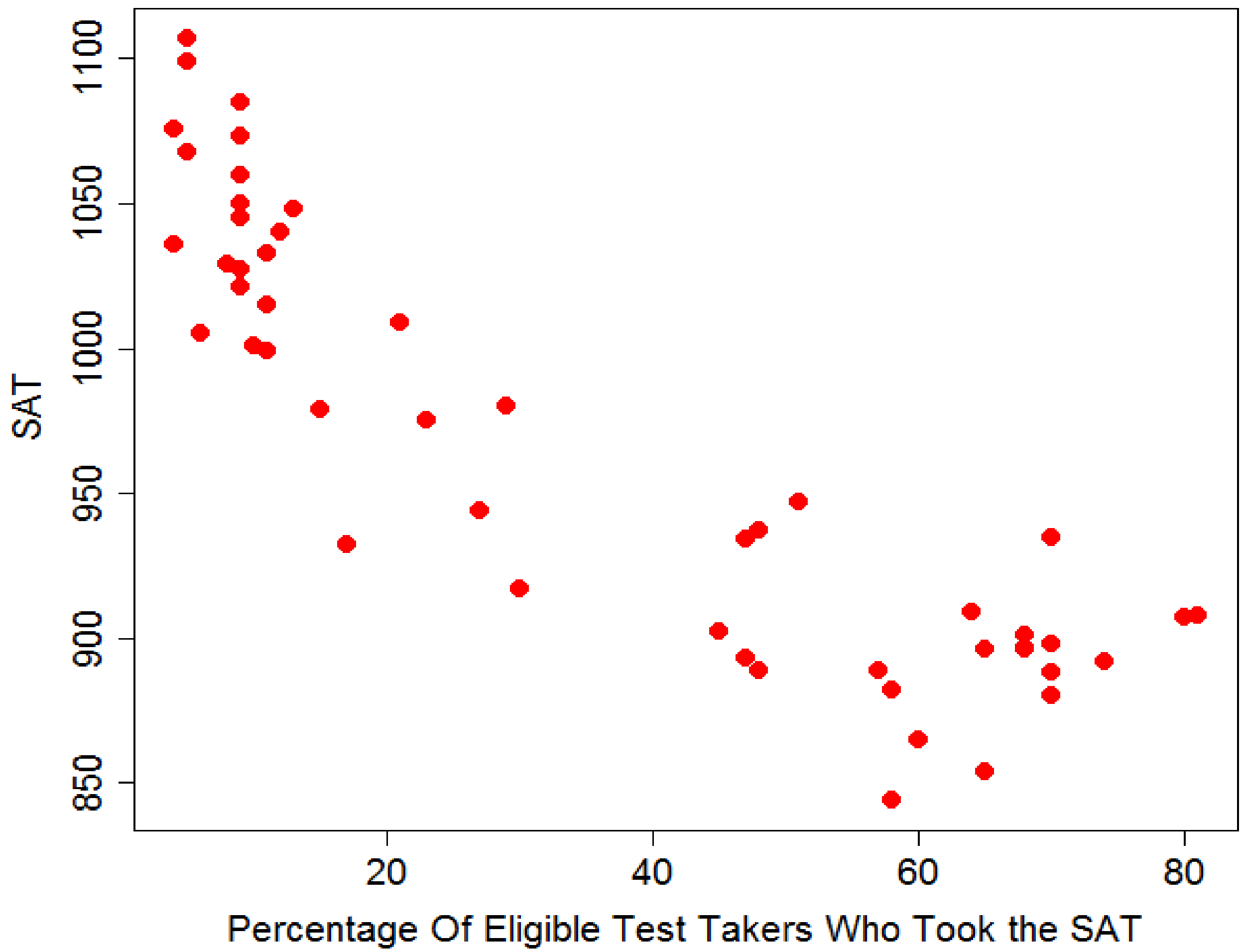
At low percentages, the relationship is relatively important.

At high percentages, the relationship is relatively trivial.

Percentage of test takers is **negatively** correlated with average SAT score.

# Bivariate Scatterplots (R)

Figure A: Bivariate scatterplot of average SAT total score for a state versus percent of eligible students who take the SAT (n = 50 states).




If we know a state's percentage of eligible test takers who take the SAT, then we can make a pretty good prediction of that state's average SAT score. If X helps us predict Y, then X and Y are ***correlated***. If X and Y are correlated, then X helps us predict Y (and Y helps us predict X). Prediction and correlation are two sides of the same coin.

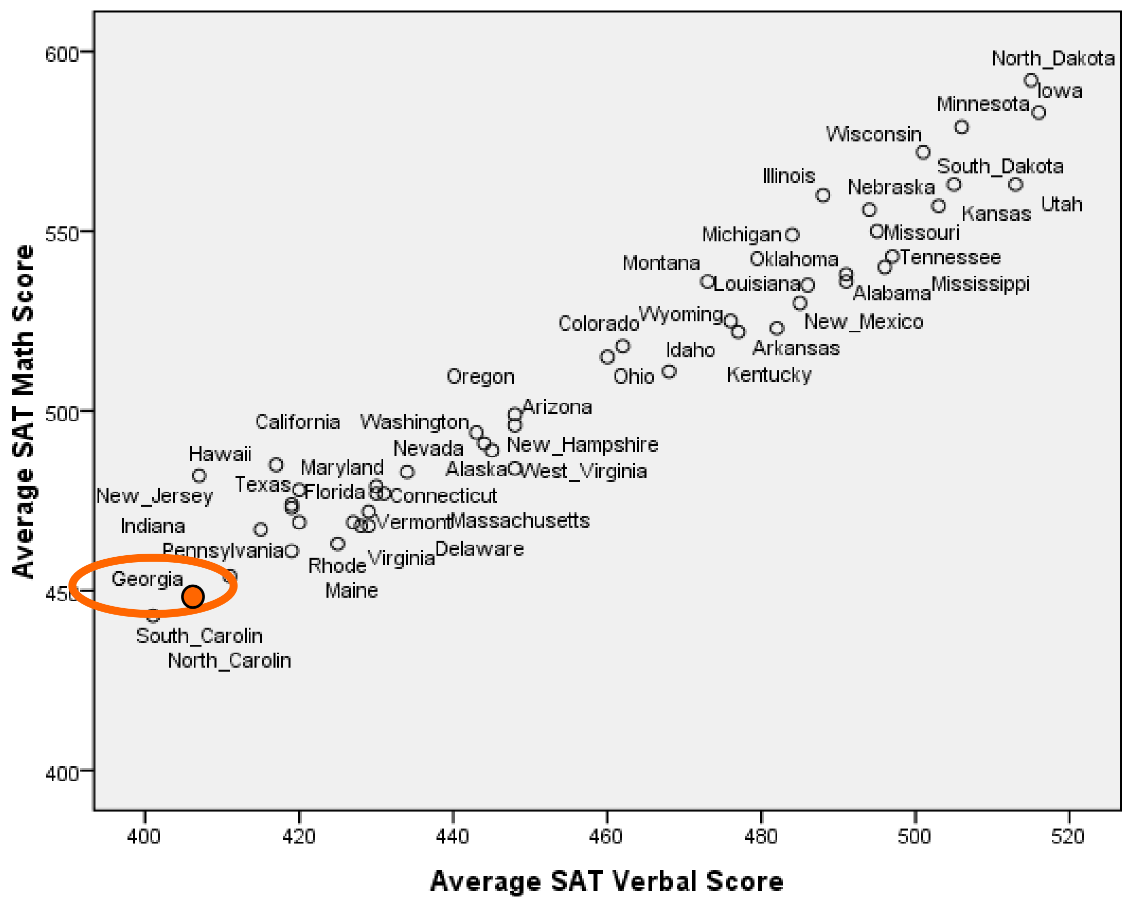
Throughout the course, I am going to use "predict" in a broad sense. When I use "predict," I generally don't mean predicting into the future. Forecasting is a narrow sense of "predict." The broad sense involves any predicting from the known to the unknown. Theories help us make predictions. We can have theories about past events or latent traits or unobserved patterns. When our theories make predictions that pan out, then such theories are supported by the data. This is why correlations are so interesting to researchers.

Note also, when I say "X predicts Y," I do not necessarily mean that X is a *good* predictor of Y. Rather, I just mean that, for the sake of predicting Y, X is better than nothing.

# Bivariate Exploratory Data Analysis

Figure B: Bivariate scatterplot of average SAT Math score for a state versus average SAT Verbal score (n = 50 states).

Let's examine another bivariate relationship.

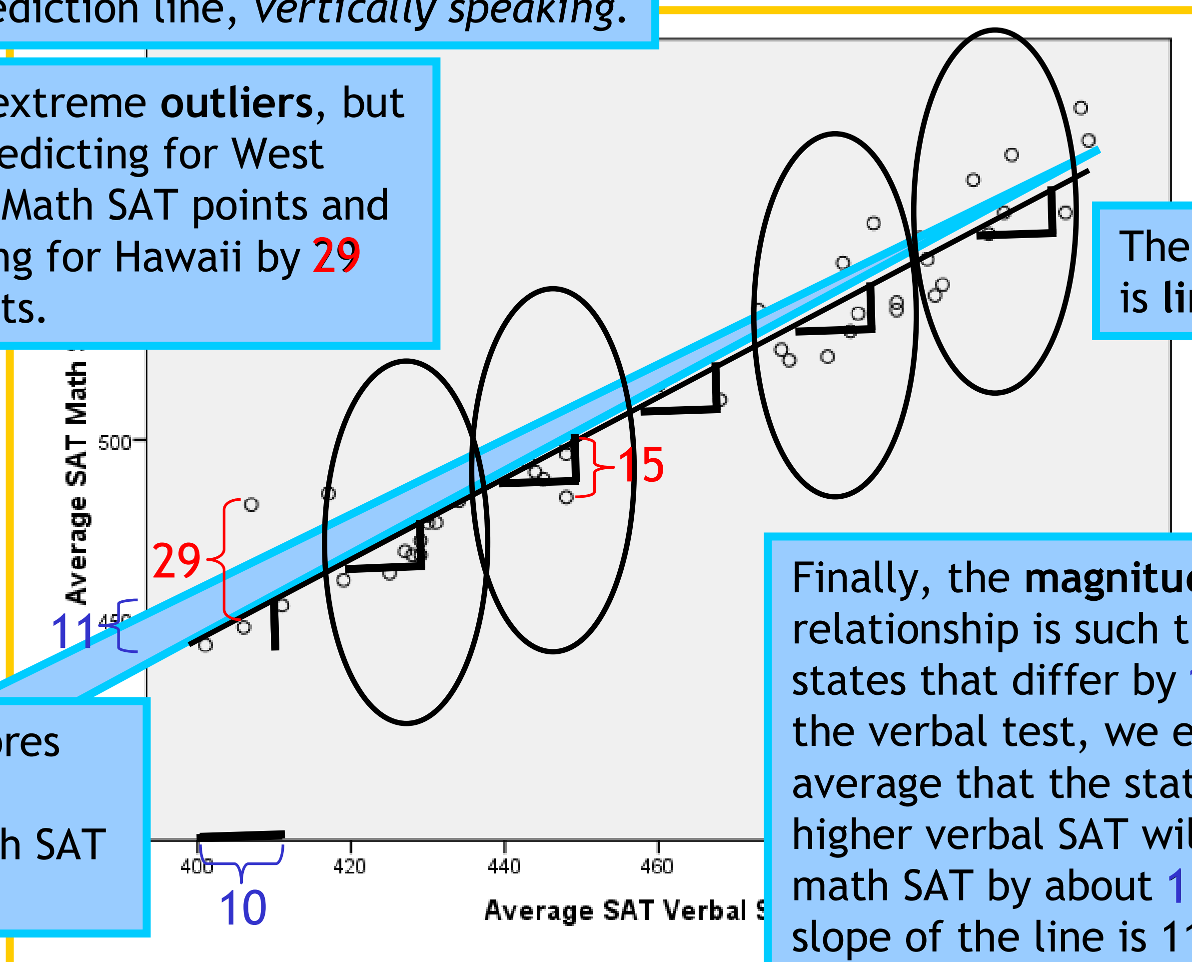


# Bivariate Exploratory Data Analysis

Figure B: Bivariate scatterplot of average SAT ... SAT Verbal

Our primary goal is to predict on average. The relationship is **strong**, because the data tend to hug our prediction line, *vertically speaking*.

There are no extreme **outliers**, but we are overpredicting for West Virginia by 15 Math SAT points and underpredicting for Hawaii by 29 Math SAT points.




The relationship is **linear**.

SAT verbal scores are **positively** correlated with SAT math scores.

Finally, the **magnitude** of the relationship is such that, given two states that differ by 10 points on the verbal test, we expect on average that the state with the higher verbal SAT will have a higher math SAT by about 11 points. The slope of the line is 11/10, or 1.1.

# Mathematizing the Magnitude (Slope)

Equation for a line (from 8th grade):

$$y = mx + b$$

$$m = slope = \frac{rise}{run} = \frac{11}{10} = 1.1$$

$b$ = y-intercept, which is the value of y when x equals zero

A little algebra, a little substitution:

$$y = b + mx$$

Let: $b = \beta_0$ Let: $m = \beta_1$

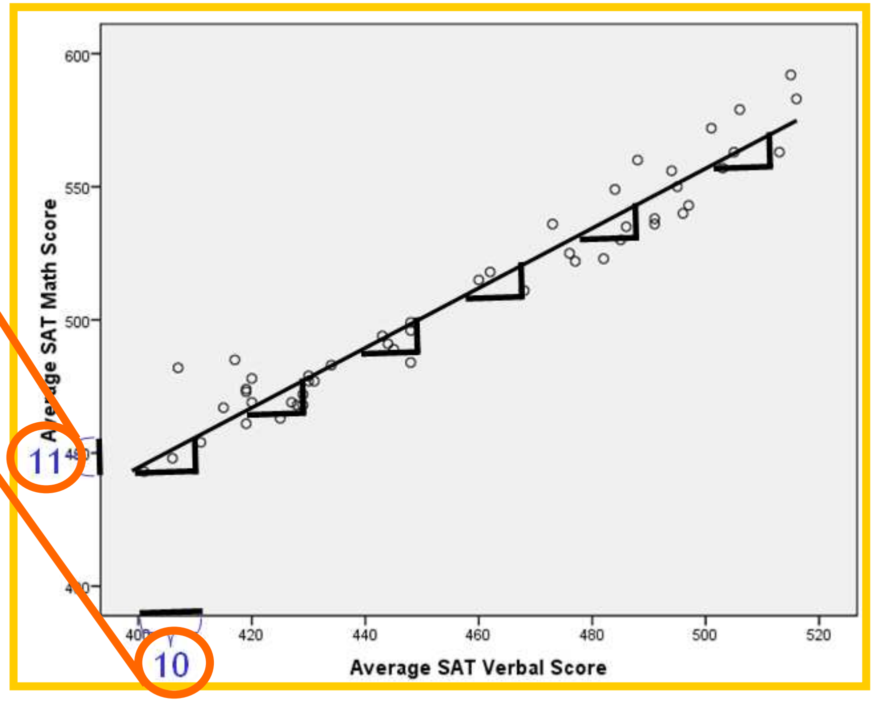
$$y = \beta_0 + \beta_1 x$$

The acknowledgement of error separates statistics from mathematics.

Let: $y = MathSAT$

Let: $x = VerbalSAT$

$$MathSAT = \beta_0 + \beta_1 VerbalSAT$$




Our (theoretical) regression model:

$$MathSAT = \beta_0 + \beta_1 VerbalSAT + \varepsilon$$

Our fitted regression model:

$$\widehat{MathSAT} = 1.8 + 1.1 VerbalSAT$$

## Interpreting the Slope (Magnitude)

$$\widehat{MathSAT} = 1.8 + 1.1 VerbalSAT$$

In our sample, there is a positive correlation between verbal SAT scores and math SAT scores such that for every 1 point difference in verbal SAT, we expect on average a 1.1 difference in math SAT.

In our sample, given two states that differ by 1 point in their verbal SAT scores, we predict that the state with the higher verbal SAT score will have a math SAT score that is 1.1 points higher.

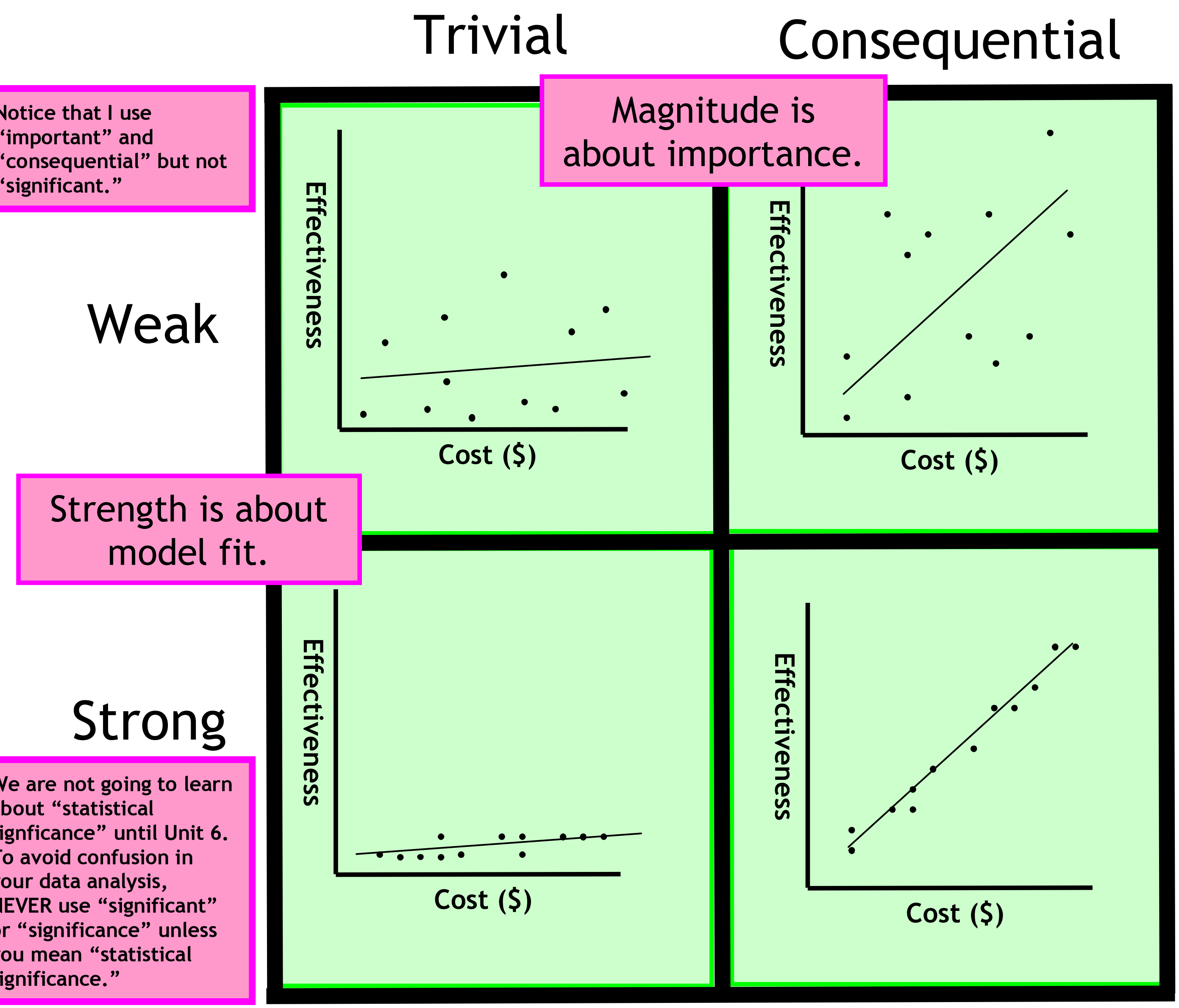
A positive correlation means that higher goes with higher.
A negative correlation means that higher goes with lower.

Correlation implies neither causation nor development.
Avoid unwarranted causal and developmental conclusions.

Do verbal SAT scores *cause* math SAT scores? $VerbalSAT \longrightarrow MathSAT$
Do math SAT scores *cause* verbal SAT scores? $MathSAT \longrightarrow VerbalSAT$
Are we observing development over time? $Maryland_{1996}$ $Verbal = 460$ $Math = 512$

# Conceptually Distinct: Strength and Magnitude

Notice that I use "important" and "consequential" but not "significant."




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

We <u>cannot</u> 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), which we'll learn about in Unit 4.

We are not going to learn about "statistical signficance" until Unit 6. To avoid confusion in your data analysis, NEVER use "significant" or "significance" unless you mean "statistical significance."

# SPSS Regression Output

The slope coefficient is the most important statistic in all of statistics.
The slope coefficient tells us the magnitude of the relationship.

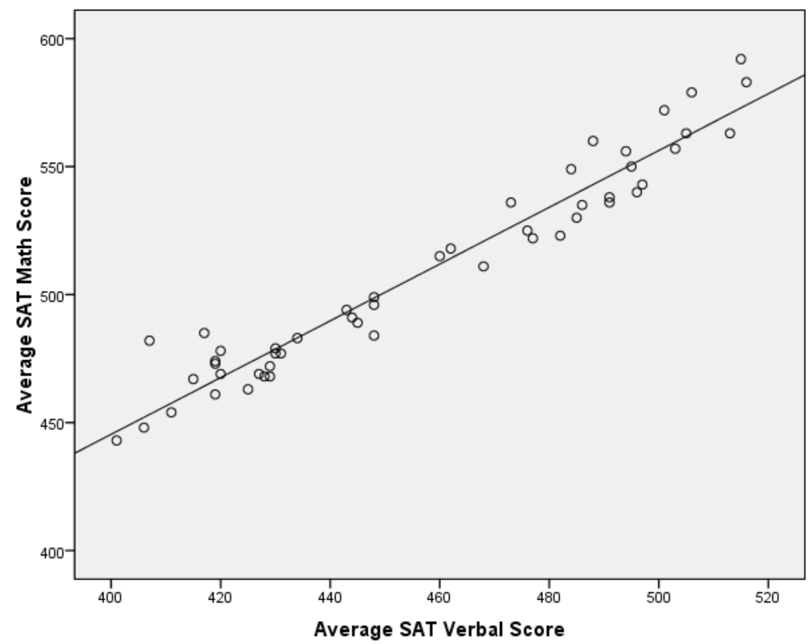
The magnitude of the relationship is the difference in the outcome (Y) associated with a one unit difference in the predictor (X).

**Theoretical Model:**

$$MathSAT = \beta_0 + \beta_1 VerbalSAT + \varepsilon$$

**Fitted Model:**

$$\widehat{MathSAT} = 1.8 + 1.1 VerbalSAT$$




**Coefficients**[a]

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

a. Dependent Variable: Average SAT Math Score

# R Regression Output

The slope coefficient has many names: *Slope*, *Magnitude*, *Parameter Estimate* (where $\beta_1$ is the parameter and 1.1 is the estimate), and *Regression Coefficient* (unstandardized).

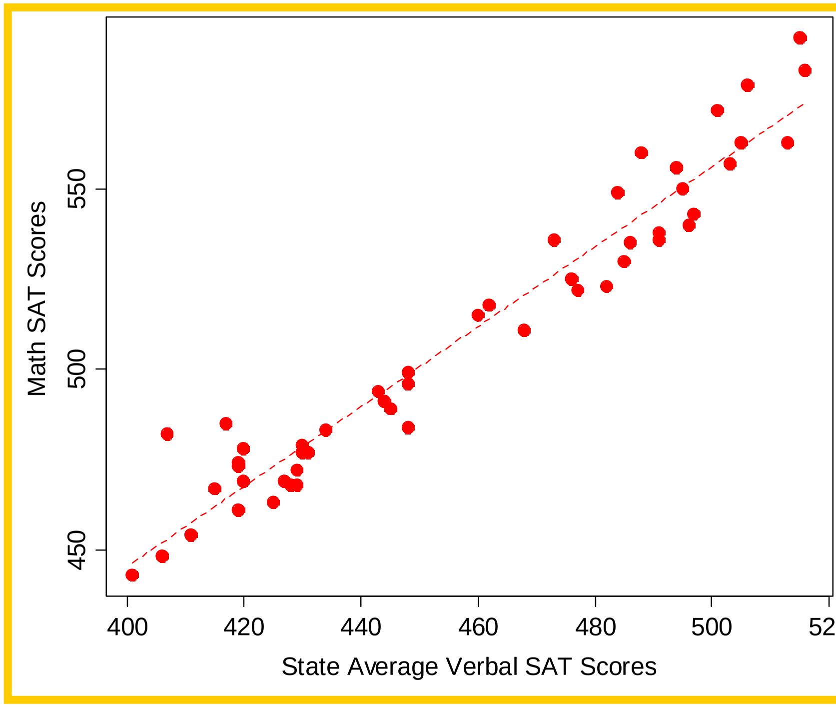
*Constant* and *Y-Intercept* are synonymous.

**Theoretical Model:**

$$MathSAT = \beta_0 + \beta_1 VerbalSAT + \varepsilon$$

**Fitted Model:**

$$\hat{MathSAT} = 1.8 + 1.1VerbalSAT$$




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.82797   18.30952    0.10     0.92
VERBAL       1.10896    0.03994   27.77   <2e-16 ***
```

The constant, or y-intercept, represents our predicted outcome (Y) when the predictor (X) equals zero. It's generally not substantively interesting because zero is often outside the range of our data (as it is here). Nevertheless, we need the constant, or y-intercept, to "anchor" our model. In addition to the slope, we need to know the vertical placement of the regression line. I hope to demonstrate this need in a few slides: Ordinary Least Squares (OLS) Regression.

# Bivariate Exploratory Data Analysis

**D**irection

**O**utliers

**L**inearity

**M**agnitude

**A**nd

**S**trength

When conducting exploratory data analysis on the relationship between two variables, look for DOLMAS: direction, outliers, linearity, magnitude and strength.

You probably want to assess linearity and direction first. For starters, draw a line (straight or curvilinear, but think vertically) by hand. Is this relationship what you expected?

Do your best to assess strength based on your (perhaps very limited) experience. In Unit 4, we'll learn to quantify strength using Pearson correlations, the *r* statistic.

Note the outliers, which will have jumped out at you during your assessment of strength. Wonder what's happening.

If the relationship is linear, calculate the slope to quantify the magnitude. If the relationship is nonlinear, consider where the magnitude is greatest and where it is least.

You can now do the Unit 1 Post Hole: Use exploratory data analytic techniques to investigate the relationship between two variables.

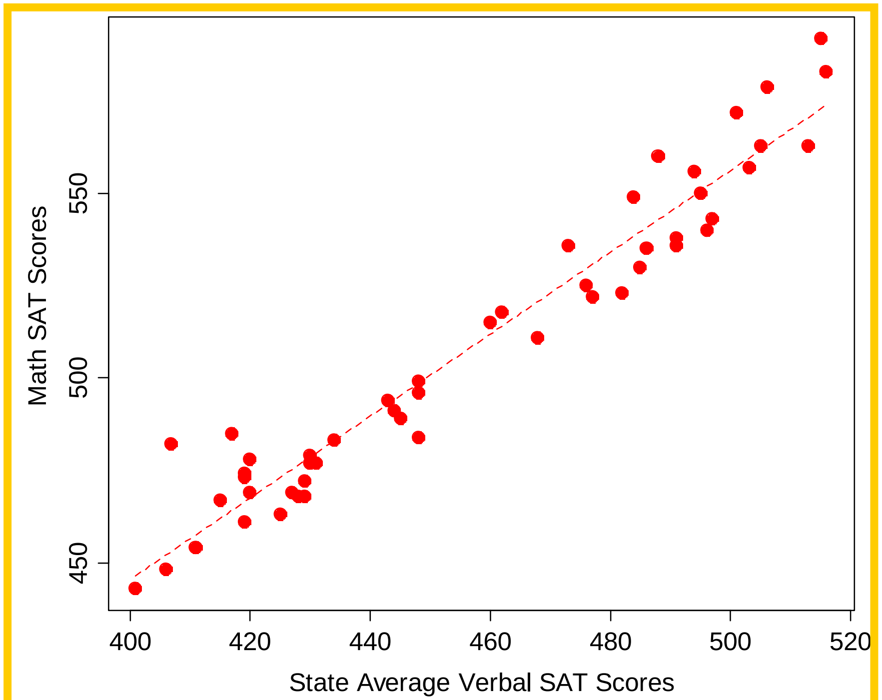
Practice problems are at the end of the presentation.

# Dig the Post Hole

## Unit 1 Post Hole:

## Use exploratory data analytic techniques to investigate the relationship between two variables.

**Evidentiary materials: a scatterplot and regression output.**




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.82797   18.30952    0.10     0.92
VERBAL       1.10896    0.03994   27.77   <2e-16 ***
```

**Post holes are sketches. In general be brief, but some post holes require a careful sentence. Post Hole 1 requires a careful sentence for the magnitude. For the magnitude, carefully interpret the slope coefficient, the most important number in all of statistics. Be sure to avoid unwarranted causal and developmental language, the most important caveat in all of statistics.**

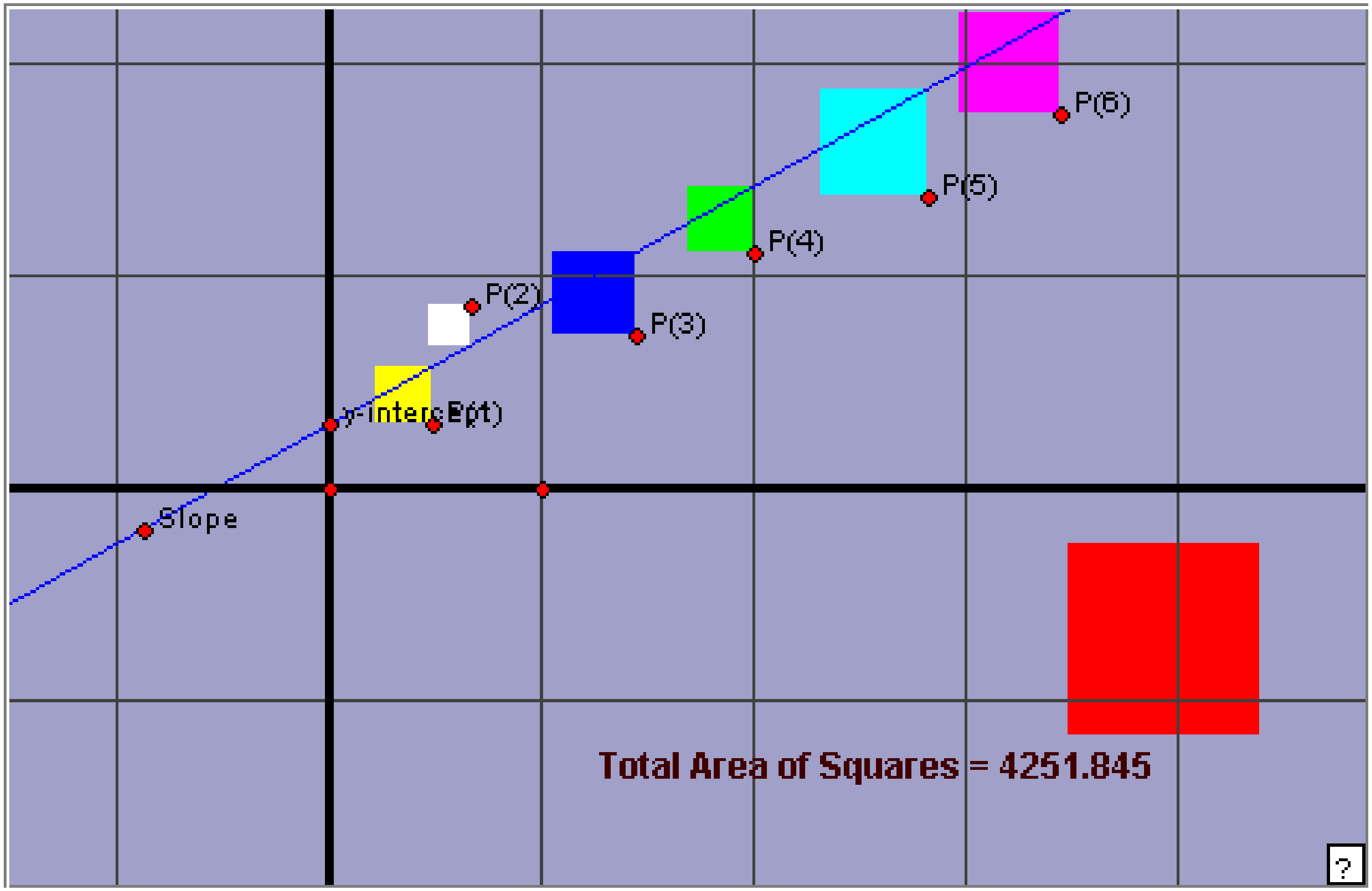
**You get a blank green box to fill in your exploratory analysis:**

**Here is my shot (the parenthetical comments are optional but nice):**

Direction: Positive (hi goes with hi, lo with lo)
Outliers: None (there is a low verbal with relatively high math but not extremely)
Linearity: Linear
Magnitude: A one point difference in Verbal SAT scores is associated with a 1.1 point difference in Math SAT scores. If we compare Math SAT scores from two states that differ in Verbal SAT scores by 100, we expect on average that the state with the higher Verbal score will have a higher Math score by about 111 points.
Strength: Strong (the data vertically hug the line tight)

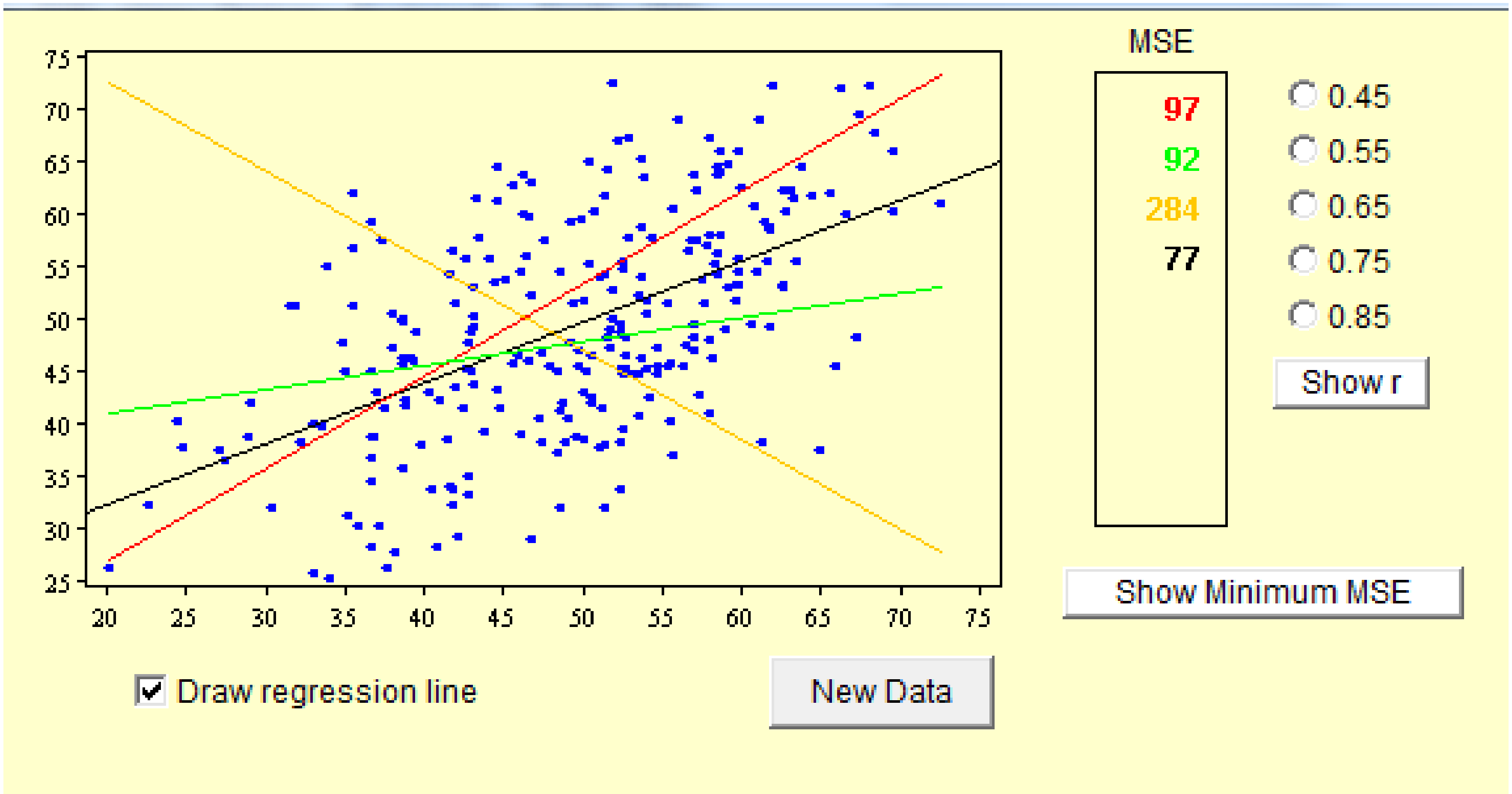
# Ordinary Least Squares (OLS) Regression

## How do SPSS and R fit the line? The Method of Ordinary Least Squares




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

# OLS Regression By Eye




http://www.ruf.rice.edu/~lane/stat_sim/reg_by_eye/

"MSE" is short for "mean square error." It is the average error square that we saw in the last slide. It is the same thing to minimize the mean square as it is to minimize the sum of squares; both are "least squares."

# Answering our Roadmap Question

Unit 1: In our sample, is there a relationship between reading achievement and free lunch?

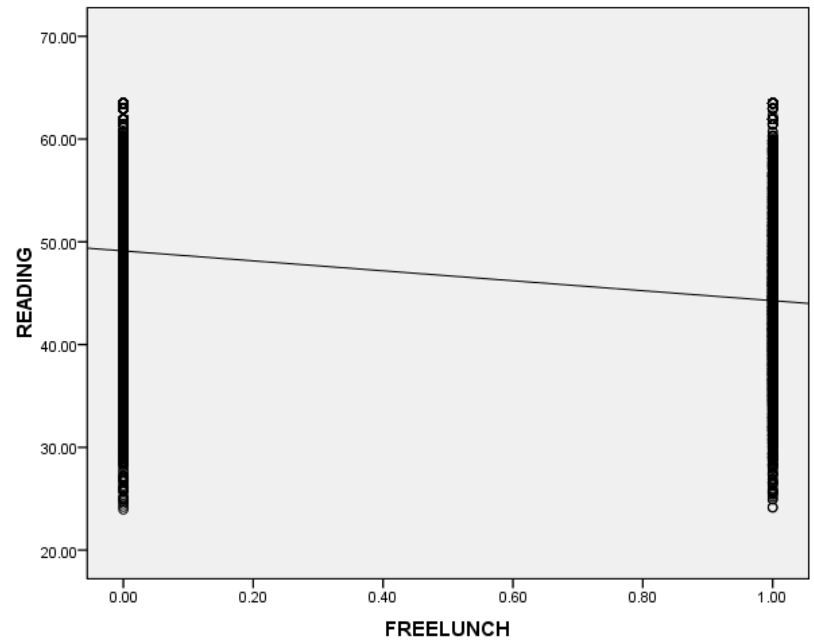
**Theoretical Model:**

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

**Fitted Model:**

$$\hat{Reading} = 49 - 5FreeLunch$$

*Freelunch* takes on only two values: 0 and 1. It is therefore a dichotomous variable. Our "prediction machine" (i.e., fitted model) gives us two predictions. One prediction for students who are eligible for free/reducted lunch (*FreeLunch* = 1), and another prediction for students who are NOT eligible (*FreeLunch* = 0).




Continuous variables take on a *continuum* of values. Dichotomous variables take on *two* values.

**Coefficients[a]**

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

a. Dependent Variable: READING

## Interpreting the Slope (Magnitude)

$$\widehat{Reading} = 49 - 5FreeLunch$$

In our sample, there is a negative correlation between free lunch eligibility and reading scores such that for every 1 unit difference in free lunch status, we expect on average a 5 point difference in reading scores.

But, since one unit is the whole shebang, we can simply say:

In our sample, students eligible for free lunch tend to score 5 points lower on the reading test than their ineligible counterparts.

In our sample, given two students who differ in free lunch eligibility, we predict that the student who is eligible for free lunch will, on average, have a reading score that is 5 points lower.

A negative correlation means that higher goes with lower.

Avoid unwarranted causal and developmental conclusions!

If free lunch *caused* low reading achievement, wouldn't that be great?
We could then solve all our educational problems by charging $75 for lunch.

# Unit 1 Appendix: Key Concepts

- **If we know a state's percentage of eligible test takers who take the SAT, then we can make a pretty good prediction of that state's average SAT score. If X helps us predict Y, then X and Y are *correlated*. If X and Y are correlated, then X helps us predict Y (and Y helps us predict X). Prediction and correlation are two sides of the same coin.**
- **Magnitude is represented by the slope of the fitted line. In order to understand magnitude, we need to have a fundamental understanding of the outcome and predictor scales.**
- **Strength refers to the model fit. If the data vertically "hug" the line (or is it the line that "hugs" the data?), then the model (as geometrically represented by the line) does a good job of predicting.**
- **Notice that I use "important" and "consequential" but not "signficant."**
- **We are not going to learn about "statistical signficance" until Unit 6. To avoid confusion in your data analysis, NEVER use "significant" or "significance" unless you mean "statistical significance."**
- **The acknowledgement of error separates statistics from mathematics. Residuals, the difference between our predictions and our observations, represent error. There are three sources of statistical error:**
  - Measurement Error
  - Unobserved Variables
  - Individual Variation
- **In our fitted models, we acknowledge error by talking about the predicted outcome and not the outcome itself. We are predicting on average. Symbolically, we represent this by putting a hat (or carrot) over the outcome.**
- **Correlation implies neither causation nor development. Avoid unwarranted causal and developmental conclusions. (This may be the most important concept of the entire course!)**

# Unit 1 Appendix: Key Interpretations

**The Unstandardized Slope Coefficient, Magnitude, Slope Parameter Estimate:**

“In our sample, there is a positive correlation between Verbal SAT scores and math SAT scores such that for every 1 point difference in Verbal SAT, we expect on average a 1.1 difference in Math SAT.”

“In our sample, given two states that differ by 1 point in their Verbal SAT scores, we predict that the state with the higher Verbal SAT score will have a Math SAT score that is 1.1 points higher.”

“In our sample, students eligible for free lunch tend to score 5 points lower on the reading test than their ineligible counterparts.”

“In our sample, given two students who differ in free lunch eligibility, we predict that the student who is eligible for free lunch will, on average, have a reading score that is 5 points lower.”

- Avoid causal language (unless warranted).
- Avoid developmental language (unless warranted).

# Unit 1 Appendix: Key Terminology

- Outcome Variable = Dependent Variable = Y-Axis Variable = Y Variable = Y
- Predictor Variable = Independent Variable = X-Axis Variable = X Variable = X

Note: I prefer "outcome/predictor" over "dependent/independent" because our Y variable *does not depend* on our X variable (unless perhaps we are doing a true experiment). Instead, we are making predictions in the Y variable based on available information, the X variable. Or, we are concluding that there are systematic differences in our subjects associated with the X variable because the X variable predicts our Y variable.

- Direction, Outliers, Linearity, Magnitude and Strength (DOLMAS)
  - Direction: Positive or Negative?
    - A positive correlation means that higher goes with higher (and lower goes with lower).
    - A negative correlation means that higher goes with lower (and lower goes with higher).
  - Outliers: Are there data points that wildly break the pattern?
  - Linearity: Is a straight line the reasonable curve to fit?
  - Magnitude: What's the bang for your buck? Does a little of X "buy" you a lot of Y? The difference in the outcome (Y) associated with a one unit difference in the predictor (X).
  - Strength: Do the data (vertically, vertically, vertically) hug the line closely?

- <u>Continuous variables</u> take on a *continuum* of values.
- <u>Dichotomous variables</u> take on *two* values.

# Unit 1 Appendix: Math

## Anatomy of A Simple Linear Regression Model

$$\boxed{OUTCOME = \beta_0 + \beta_1 PREDICTOR + \varepsilon}$$

$$\boxed{\beta_0 = \text{y} - \text{intercept} = \text{the predicted value of our outcome (Y) when our predictor (X) equals zero}}$$

$$\boxed{\beta_1 = \text{slope} = \frac{\text{rise}}{\text{run}} = \text{magnitude}}$$

$$\boxed{\beta_1 = \text{the difference in the outcome (Y) associated with a unit difference in the predictor (X)}}$$

$$\boxed{\varepsilon = \text{error (due to measurement error, hidden variables and individual variation)}}$$

# Unit 1 Appendix: Math (Very Optional)

If you want to fit by hand a linear model using ordinary least squares (OLS) regression, you'll need multivariable calculus (although we'll see a shortcut in Unit 4). Calculus is very good at finding minimums and maximums. When we do OLS regression, we want to find a y-intercept ($\beta_0$) and slope ($\beta_1$) that minimizes the sum of squared errors (i.e., sum of squared residuals). A statistical error (i.e., residual) is the difference between our observation and prediction. Say that we have three observations:

| *NAME* | *READING* | *FREELUNCH* |
|---|---|---|
| Sean | 90 | 0 |
| Betsy | 100 | 0 |
| Waverly | 80 | 1 |

We propose a model:

$$\boxed{READING = \beta_0 + \beta_1 FREELUNCH + \varepsilon}$$

Thus:

$$\boxed{READING - \beta_0 - \beta_1 FREELUNCH = \varepsilon}$$

Thus:

$$\boxed{(READING - \beta_0 - \beta_1 FREELUNCH)^2 = (\varepsilon)^2}$$

Each subject has a squared error:

$$\boxed{(90 - \beta_0 - \beta_1 0)^2 = (\varepsilon_{Sean})^2}$$

$$\boxed{(100 - \beta_0 - \beta_1 0)^2 = (\varepsilon_{Betsy})^2}$$

$$\boxed{(80 - \beta_0 - \beta_1 1)^2 = (\varepsilon_{Wavy})^2}$$

The sum of squared errors (SSE) is a function of two variables, $\beta_0$ and $\beta_1$:

$$\boxed{SSE(\beta_0, \beta_1) = (90 - \beta_0 - \beta_1 0)^2 + (100 - \beta_0 - \beta_1 0)^2 + (80 - \beta_0 - \beta_1 1)^2}$$

# Unit 1 Appendix: R Syntax

```
# Any line that begins with a pound sign (#) is commented out.
# The tilde sign (~) tells R that you are statistically modeling.
# Order matter. Your outcome goes first.
# To tilde key is probably in the upper left-hand corner of your keyboard.
# You’ll need to hit that key while holding down SHIFT to create a tilde.
# The plot function produces different results depending on the input.
# Since we are inputing a simple linear regression model,
#   the plot function produces a bivariate scatterplot.
# The lm function stands for “linear model.”
# Without further code, the lm function only produces the intercept and slope,
#   but that is exactly what we need now. (We’ll get more when we need it.)
# Here, the dataset name just happens to be the same as the outcome name.
# There are at least two ways to tell R the dataset of the variables:
#   After a comma, specify the dataset.
#   with a dollar sign ($), attach the dataset before the variable name.

plot(SAT~PERCENT, data=SAT)
lm(SAT~PERCENT, data=SAT)

plot(SAT$SAT~SAT$PERCENT)
lm(SAT$SAT~SAT$PERCENT)
```

# Unit 1 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!.
*Why? 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)
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)

**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

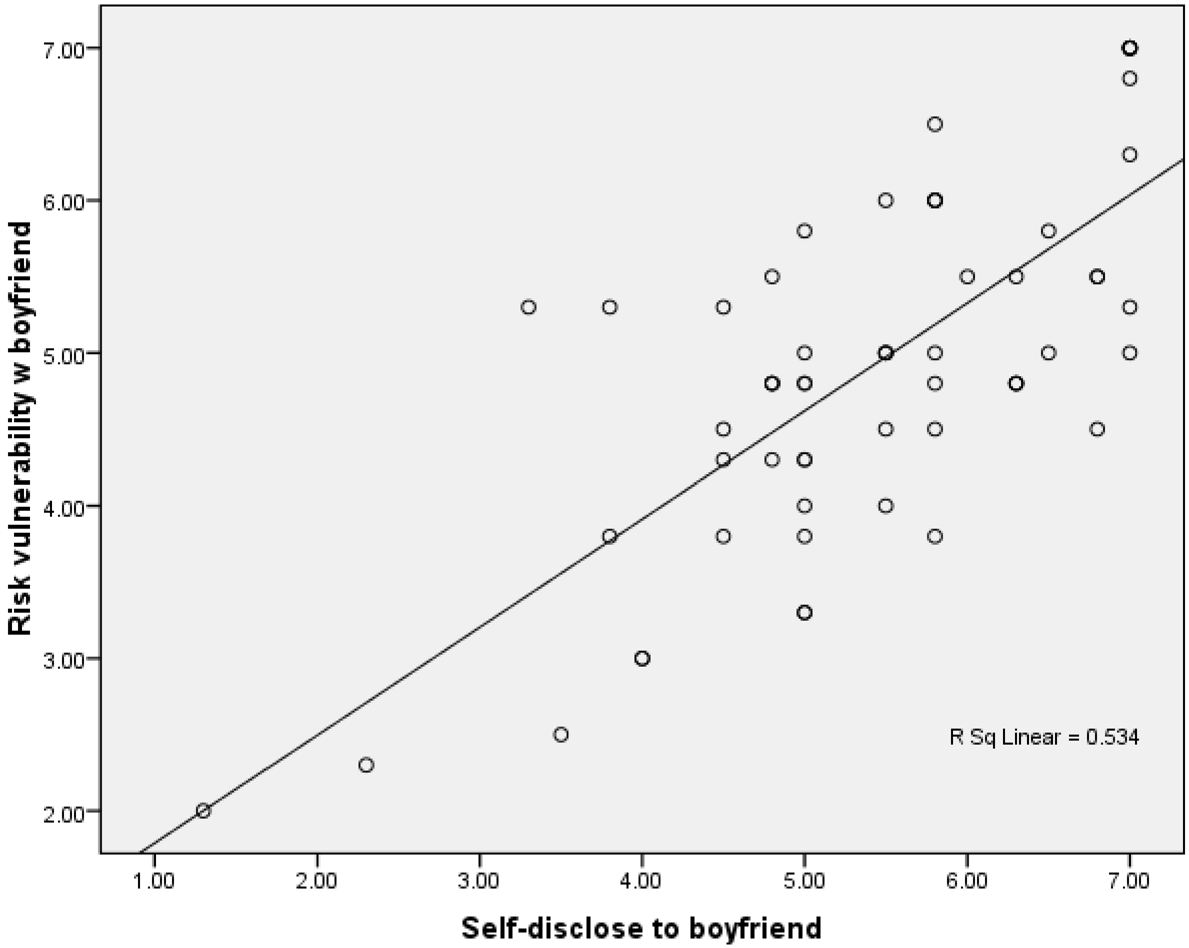
b. Dependent Variable: Risk vulnerability w boyfriend

**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

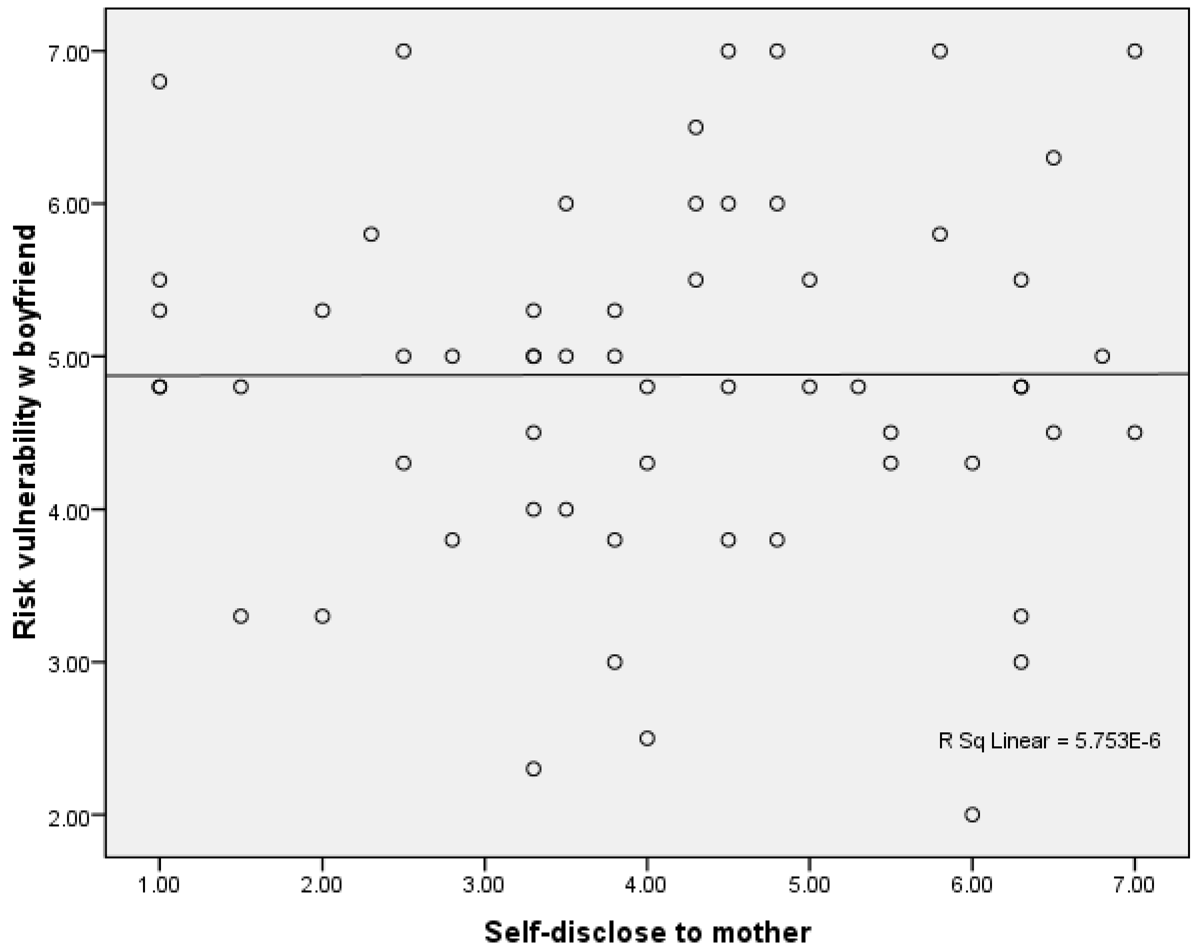
b. Dependent Variable: Risk vulnerability w boyfriend

**Coefficients[a]**

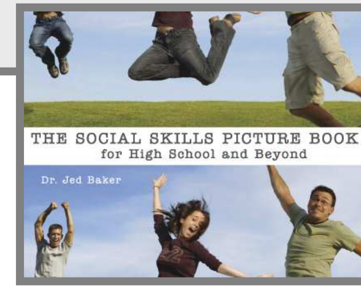
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 4.872 | .404 | | 12.050 | .000 |
| | Self-disclose to mother | .002 | .091 | .002 | .018 | .985 |

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:**

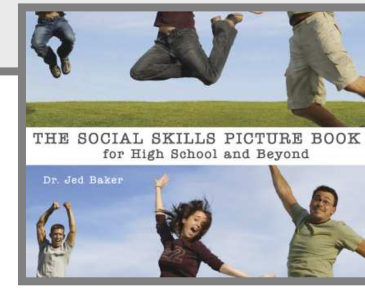
Variables about the student—

(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

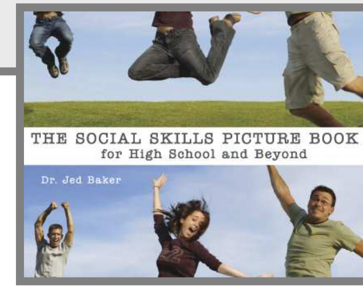
b. Dependent Variable: Base Year Composite Test

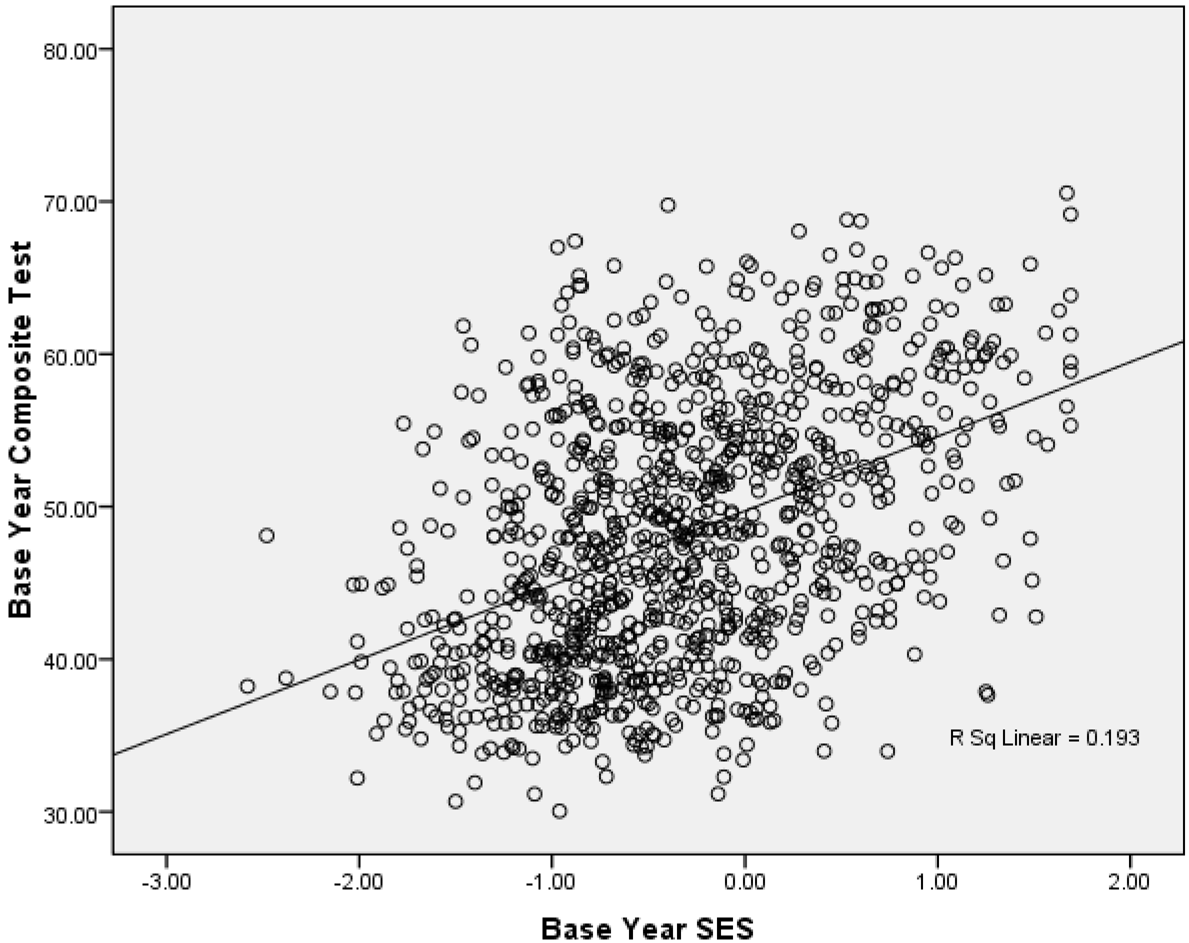
**Coefficients[a]**

| 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 |

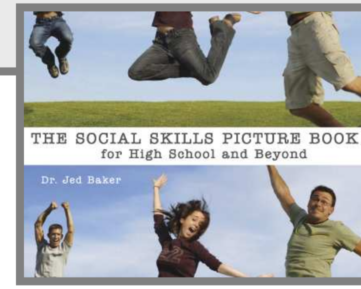
a. Dependent Variable: Base Year Composite Test

# 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

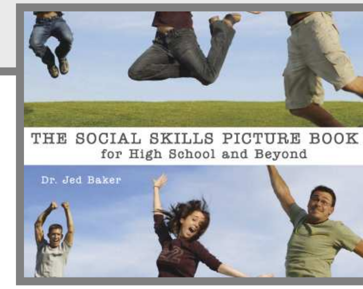
b. Dependent Variable: Base Year Composite Test

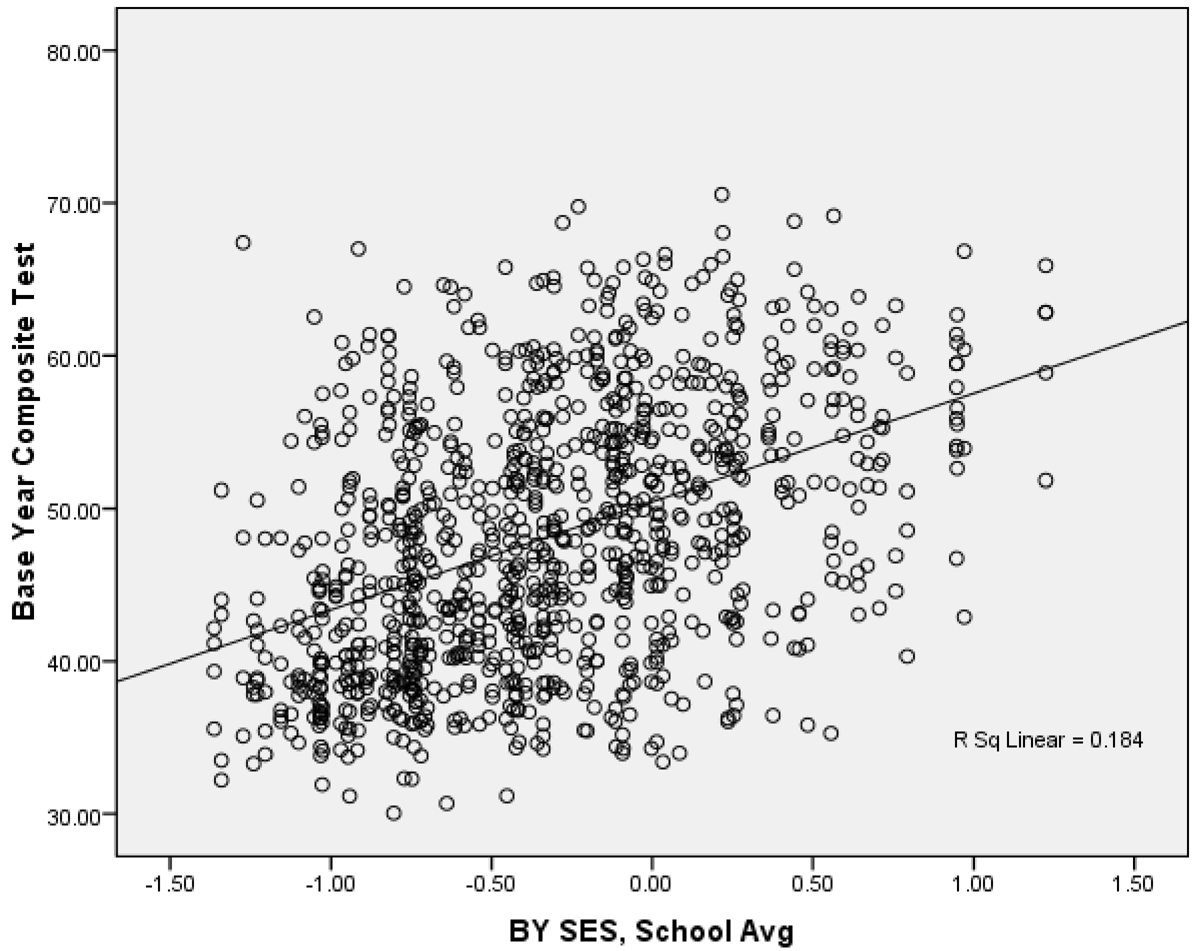
**Coefficients[a]**

| 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

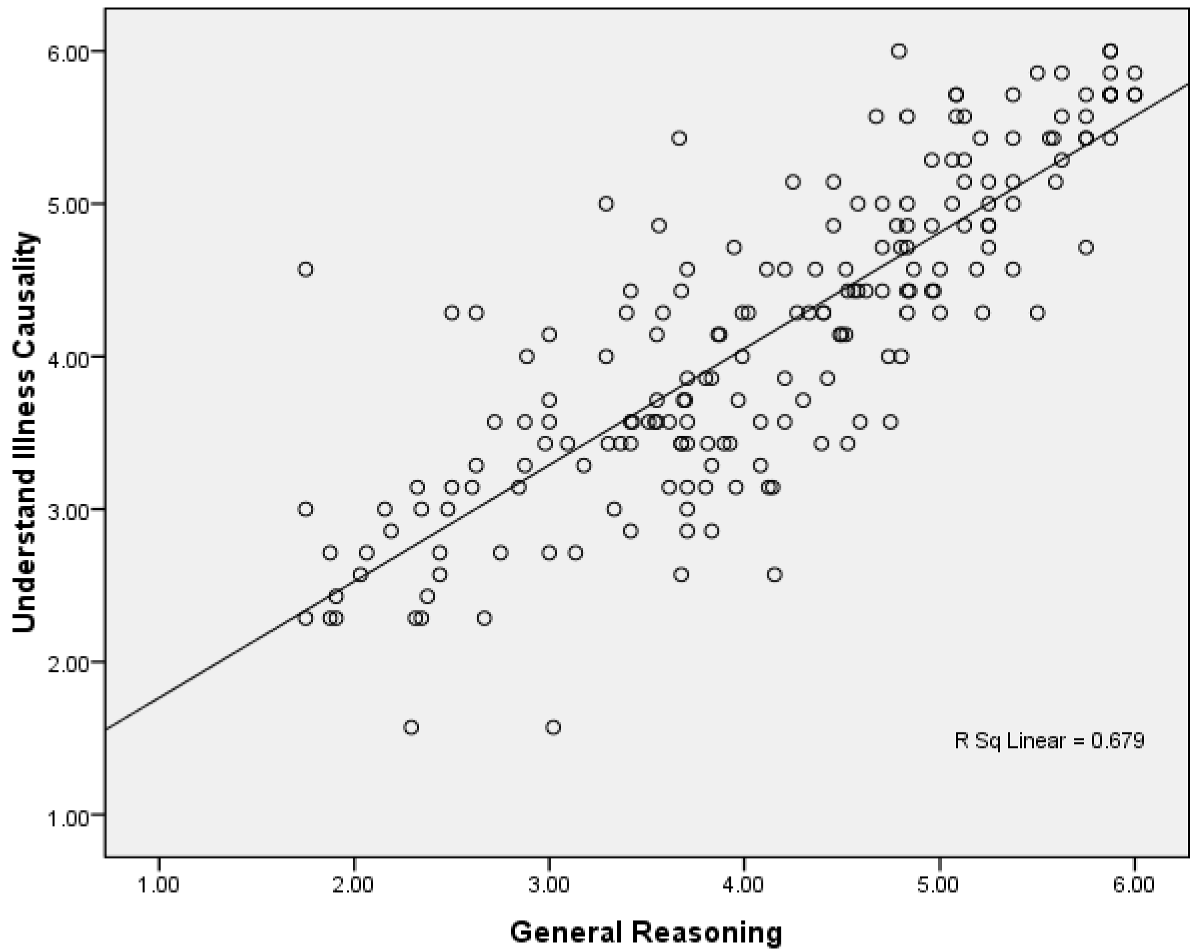
b. Dependent Variable: Understand Illness Causality

**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

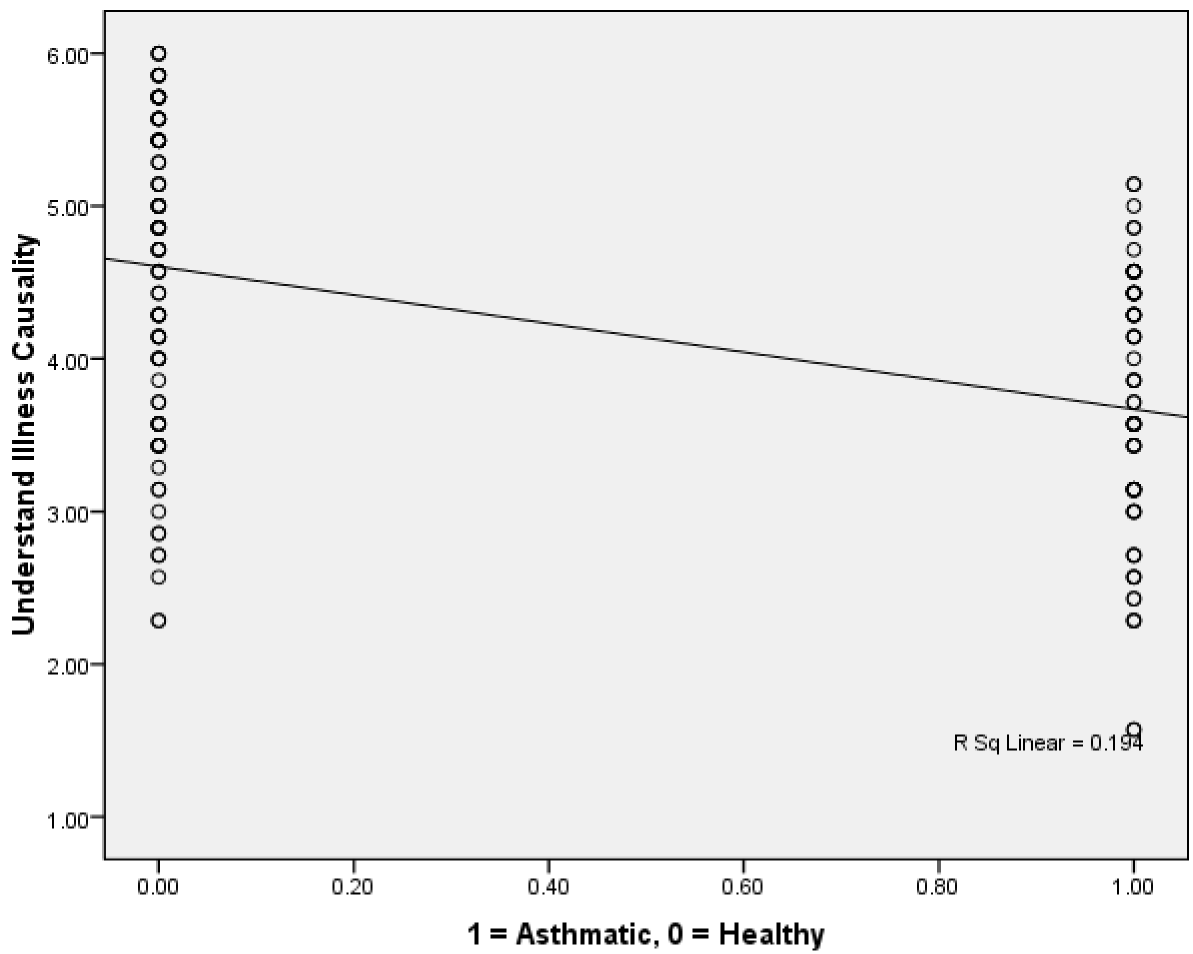
b. Dependent Variable: Understand Illness Causality

**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

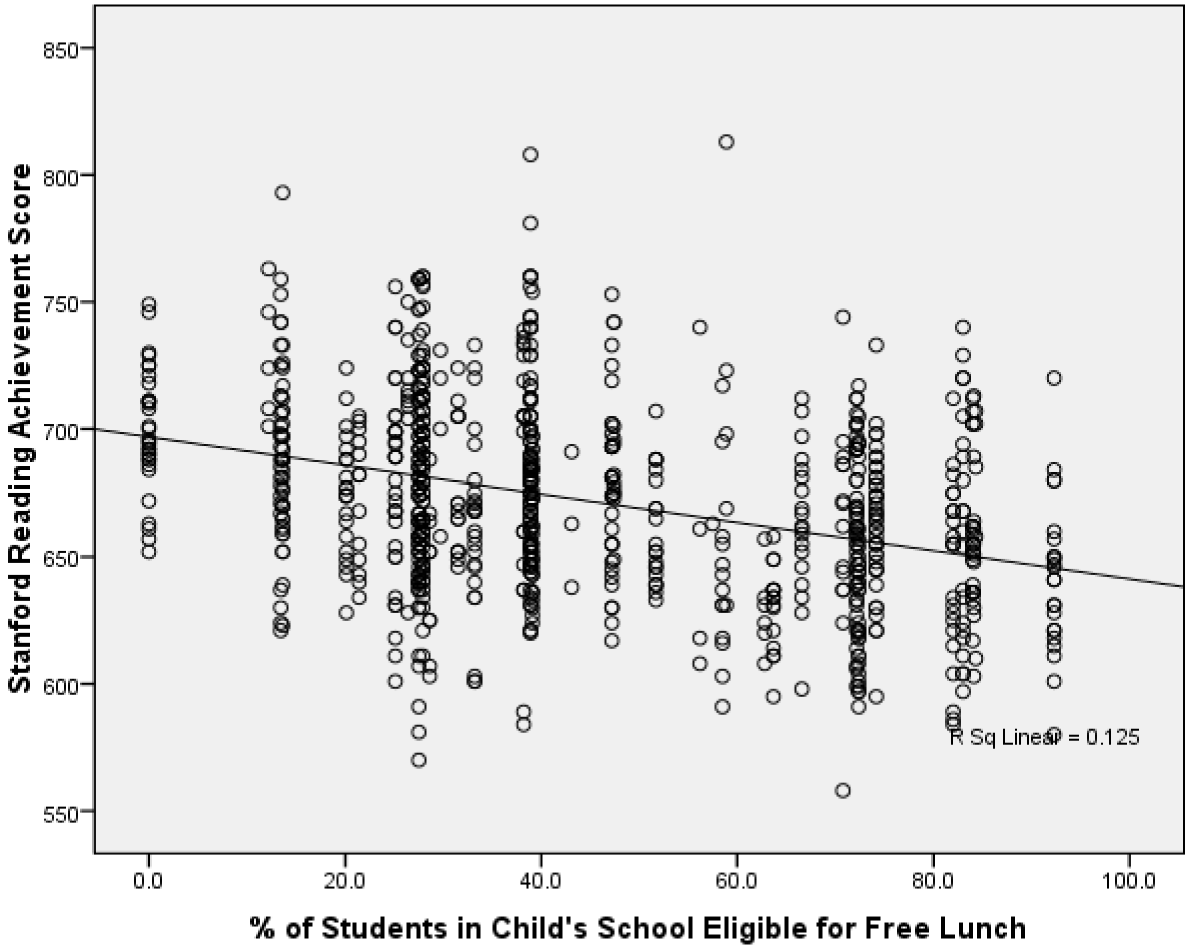
b. Dependent Variable: Stanford Reading Achievement Score

**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

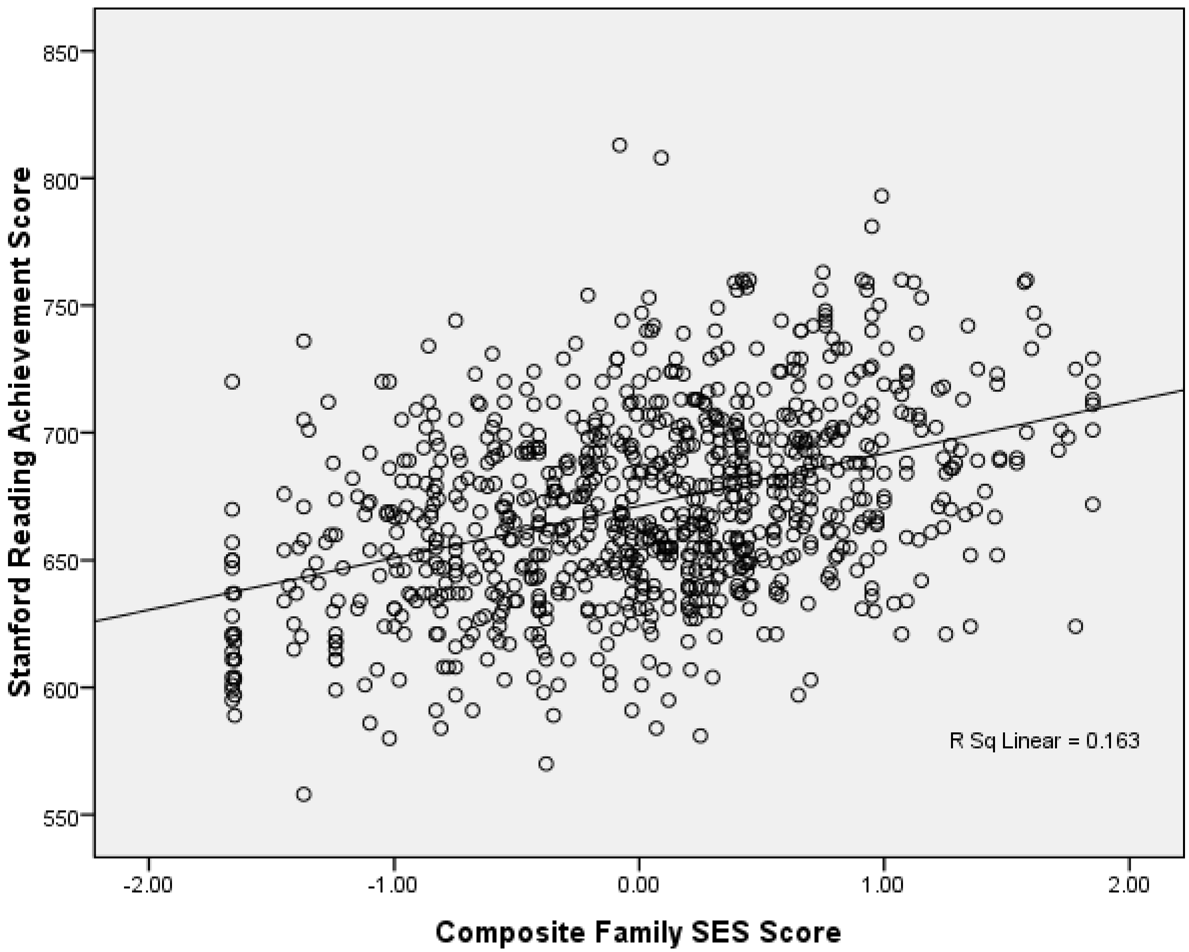
b. Dependent Variable: Stanford Reading Achievement Score

**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

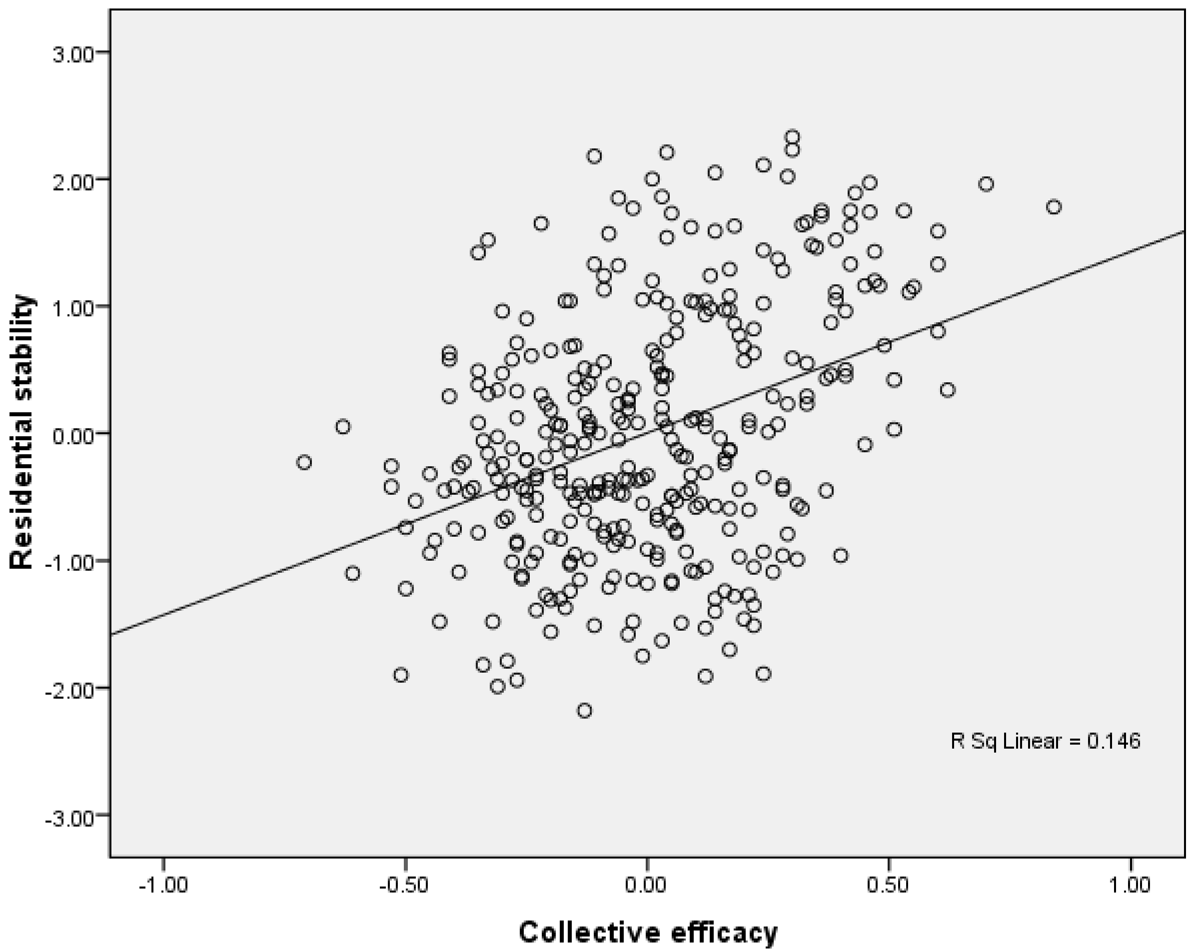
b. Dependent Variable: Residential stability

**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

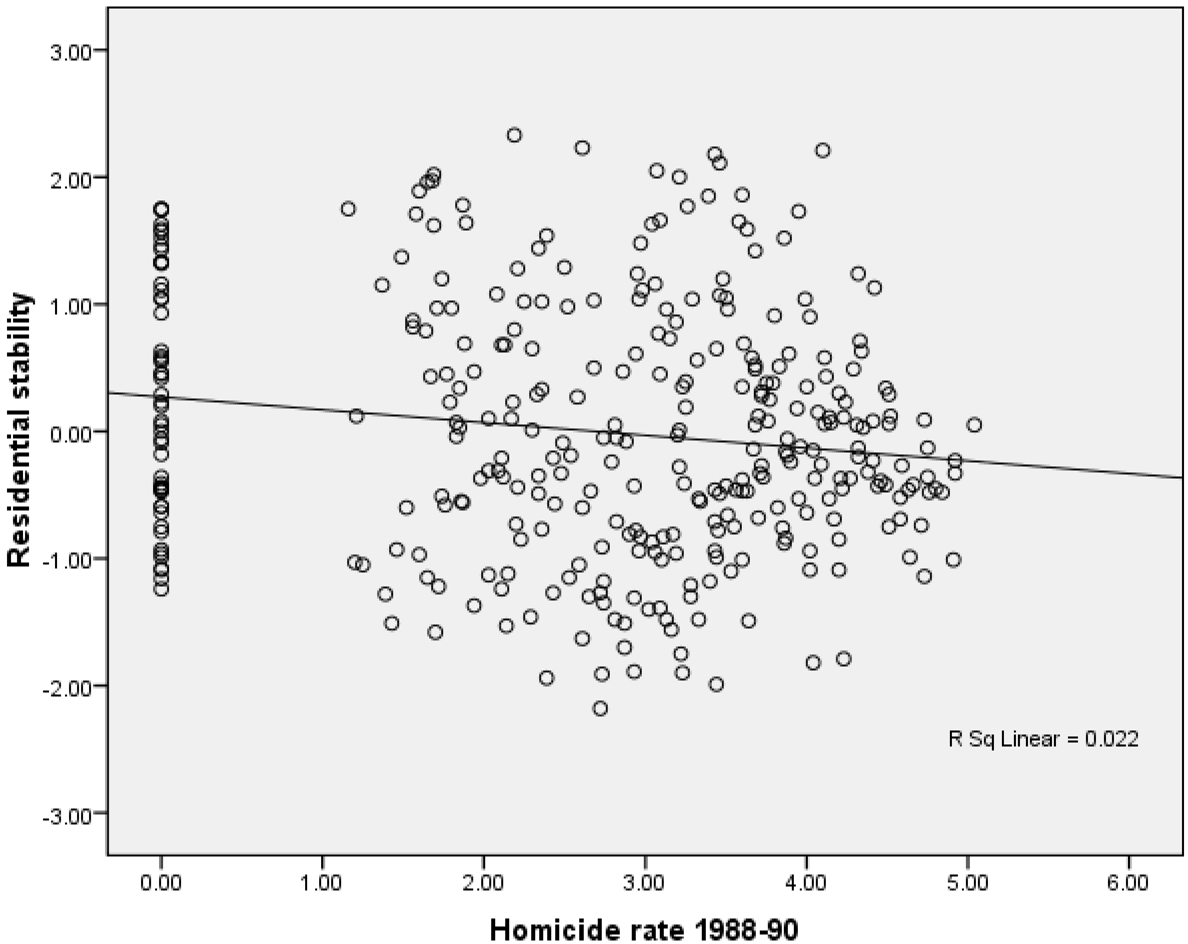
b. Dependent Variable: Residential stability

**Coefficients[a]**

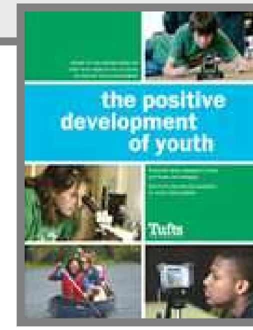
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound |
| 1 | (Constant) | .270 | .111 | | 2.432 | .016 | .052 | .489 |
| | Homicide rate 1988-90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

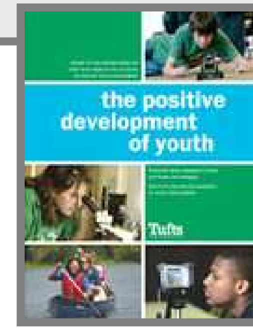



- **4-H Study of Positive Youth Development**
- **Source: Subset of data from IARYD, Tufts University**

- **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 | (AcadComp) | Self-Perceived Academic Competence |
| (MothEd) | Years of Mother's Education | (SocComp) | Self-Perceived Social Competence |
| (Grades) | Self-Reported Grades | (PhysComp) | Self-Perceived Physical Competence |
| (Depression) | Depression (Continuous) | (PhysApp) | Self-Perceived Physical Appearance |
| (FrInfl) | Friends' Positive Influences | (CondBeh) | Self-Perceived Conduct Behavior |
| (PeerSupp) | Peer Support | (SelfWorth) | Self-Worth |
| (Depressed) | 0 = (1-15 on Depression)<br>1 = Yes (16+ on Depression) | | |

# 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

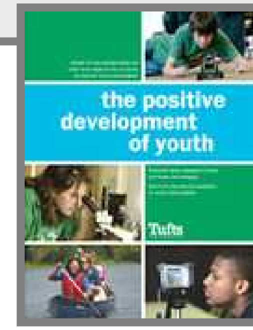
b. Dependent Variable: Self-Worth

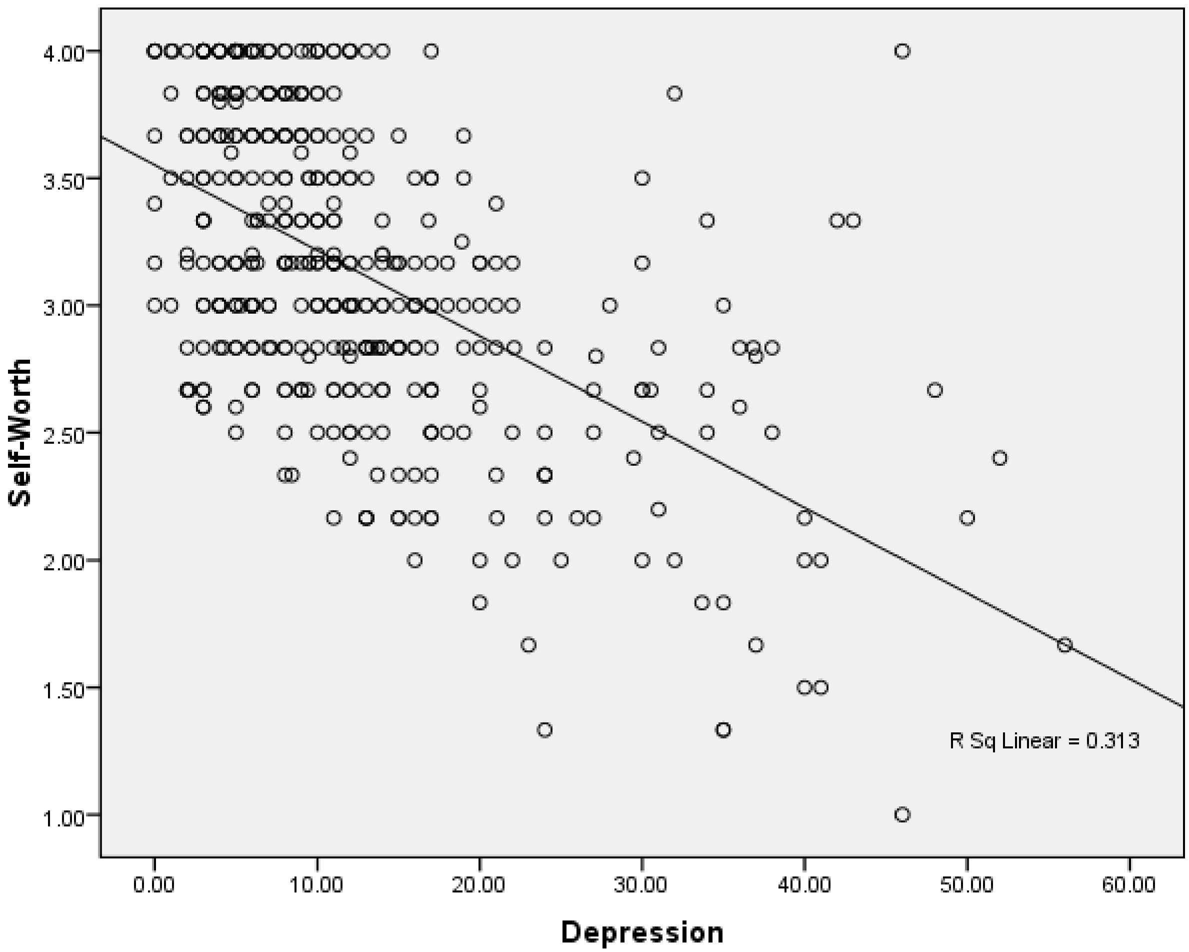
**Coefficients[a]**

| 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 |

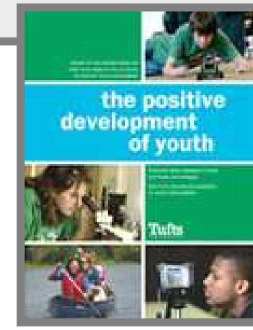
a. Dependent Variable: Self-Worth

# 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)