Unit 8: Statistical Inference and Assumption Checking

Unit 8 Post Hole:

Evaluate the assumptions underlying a simple linear regression.

Unit 7 and 8 Technical Memo and School Board Memo:

Continued from last week, fit and discuss two regression models being sure to check your regression assumptions.

Unit 8 (and Units 6 and 7) Reading:

http://onlinestatbook.com/

Chapter 5, Probability
Chapter 7, Sampling Distributions
Chapter 9, Logic Of Hypothesis Testing
Chapter 11, Power
Chapter 6, Normal Distributions
Chapter 8, Estimation
Chapter 10, Testing Means
Chapter 12, Prediction

Unit 8: 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. (4 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 8: Technical Memo and School Board Memo

Work Products (Part II of II):

I. Technical Memo (continued)

  D. Regression Analysis. Answer your research question using inferential statistics. (1 Paragraph)

    i. Include your fitted model.

    ii. Use the R^2 statistic to convey the goodness of fit for the model (i.e., strength).

    iii. To determine statistical significance, test the null hypothesis that the magnitude in the population is zero, reject (or not) the null hypothesis, and draw a conclusion (or not) from the sample to the population.

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

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

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

  X. Exploratory Data Analysis. Explore your data using outlier resistant statistics.

    i. For each variable, use a coherent narrative to convey the results of your exploratory univariate analysis of the data. Don't lose sight of the substantive meaning of the numbers. (1 Paragraph Each)

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

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

III. Memo Metacognitive

Unit 8: 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 8: 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 (mean, sd) — Unit 2 (quantiles, n)

```
> 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 (Estimate) — Unit 8 (Std. Error) — Unit 6 (t value, Pr(>|t|)) — Unit 5 (Adjusted R-squared) — Unit 9 (F-statistic)

```
> 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 (confidence intervals)

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

Unit 4 (correlation)

Unit 8: Roadmap (SPSS Output)

Model Summary

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

a. Predictors: (Constant), FREELUNCH

Unit 5

ANOVAb

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	40744.322	1	40744.322	597.251	.000a
	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

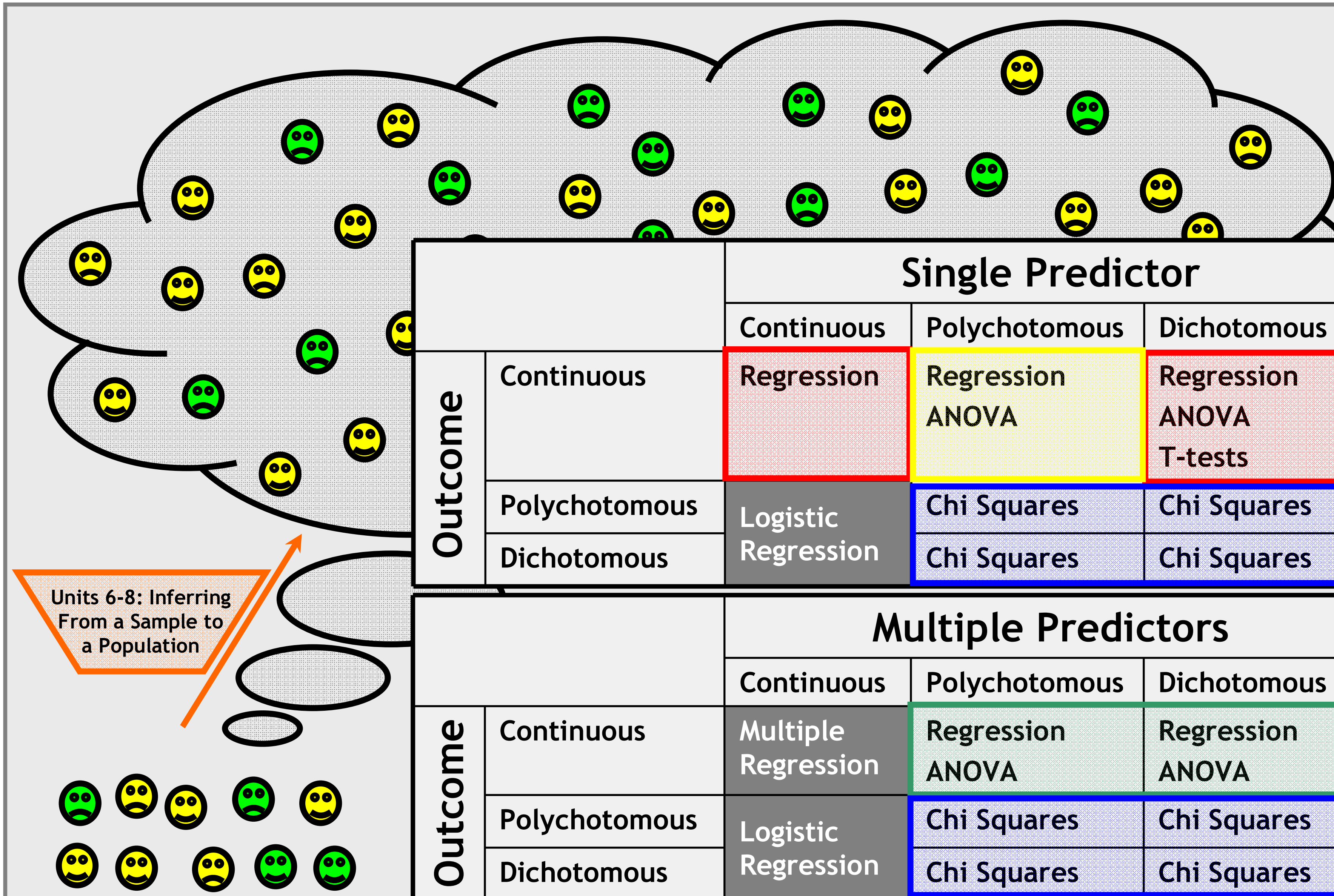
Coefficientsa

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 8: Road Map (Schematic)



Single Predictor

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

Multiple Predictors

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

Epistemological Minute

In his *Probability and the Logic of Rational Belief* (1961), Henry Kyburg asks you to consider three statements:

1. If you are 99.9999% certain of something, then it is rational for you to believe it.
2. If it is rational for you to believe *one thing*, *and* it is rational for you to believe *another thing*, then it is rational for you to believe *both things*.
3. It is *not* rational for you to believe a self-contradiction.

Do you find any of these statements objectionable?

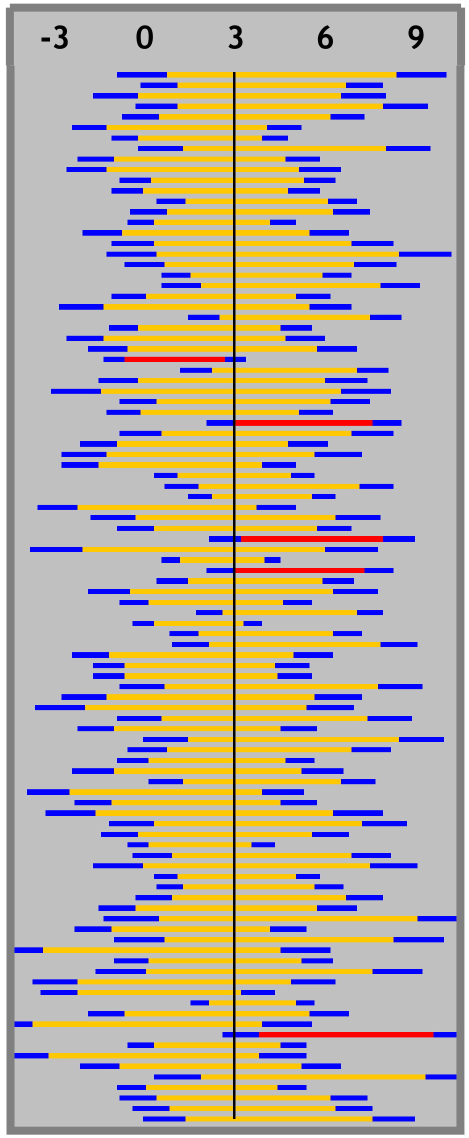
Kyburg asks you then to consider a fair lottery with 1 million tickets and 1 million contestants (each with a ticket) and 1 winner, and you are a contestant with a ticket:

1. You are 99.9999% certain that your ticket is a loser, therefore it is rational for you to believe that your ticket is a loser. Likewise, you are 99.9999% certain that Jo's ticket is a loser, and you are 99.9999% certain that Joe's ticket is a loser, and so on down the line.
2. If it is rational to believe each ticket is a loser, then it is rational for you to believe that all tickets are losers.
3. It is not rational for you to believe that one ticket will win and that no tickets will win.

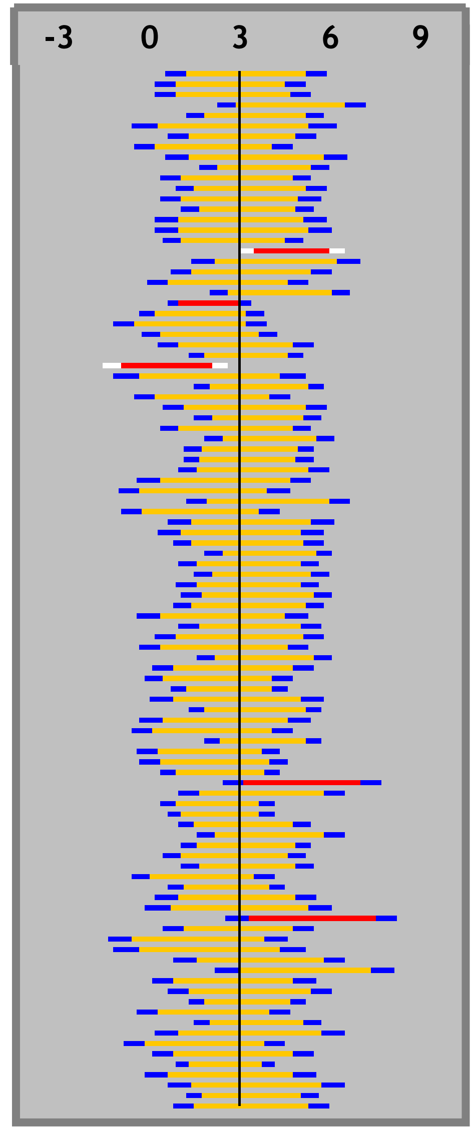
In the face of the "Lottery Paradox," which of the above three statements do you find objectionable?

I don't know the solution to the Lottery Paradox, but I do know that, when we build confidence intervals based on standard errors, we set the terms of the lottery. When we choose "95%" for our confidence interval, we make exactly 95% of our infinite tickets winners. Granted, the "exactly 95%" assumes our standard errors are unbiased, and it only takes into consideration error due to random sampling.

N = 10



N = 20

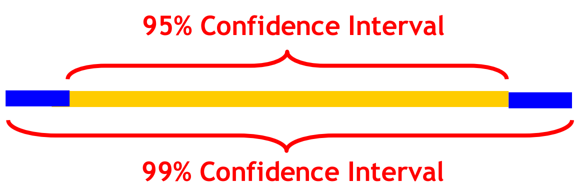


Confidence Interval Review (Part I of II)

Let us adapt and expand the applet from Unit 7:

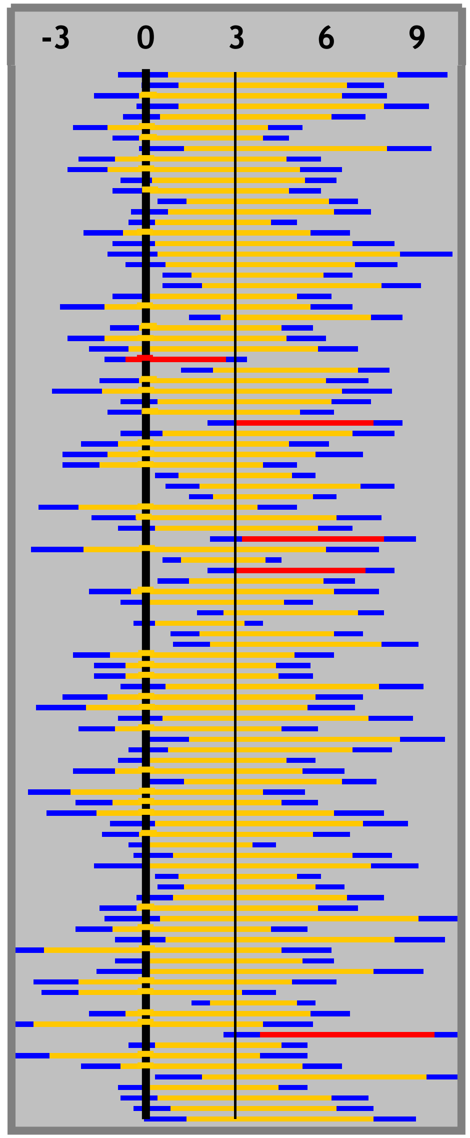
http://www.ruf.rice.edu/~lane/stat_sim/conf_interval/

Suppose the population slope is 2.98967438765. In life, we get one sample with which to estimate the population slope. We recognize that our estimate is, in all probability, wrong. In response, we estimate another population parameter, the standard error. The standard error gives us a measure of precision for our slope estimate. We can use the standard error to conduct a test of the null hypothesis (to determine statistical significance). Or, we can use the standard error to build a confidence interval (to determine a range of plausible values for the population slope). Let's focus on confidence intervals, but let's keep in mind that "alpha = .05" and "95% confidence intervals" are two sides of the same coin. In other words, if our 95% confidence interval contains zero, then we cannot reject the null hypothesis at alpha = .05.

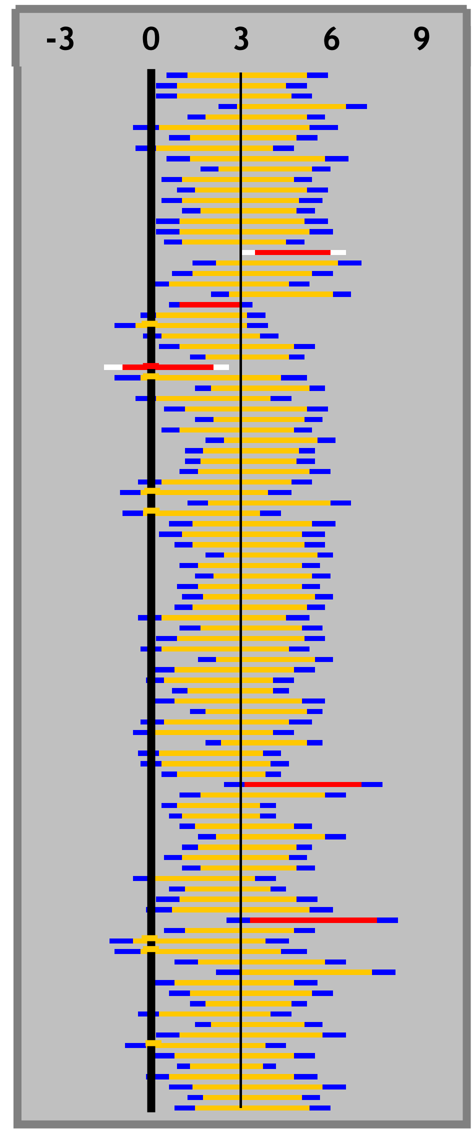


We get one sample, but if we got 100 samples, we would expect about 5 of our 95% confidence intervals to fail in their attempt to contain the true slope. That result is by design! If we don't like it, we can use 99% confidence intervals, for which we would expect only about 1 miss in 100 tries. The cost is that we are less likely to reject the null hypothesis. This is the eternal, perhaps infernal, trade-off between false positives and false negatives.

N = 10



N = 20



Confidence Interval Review (Part II of II)

Still assuming that the population slope is 2.98967438765:

Score Card , Alpha =.05, 95% Confidence Intervals:

When N = 10, we reject the null 58% of the time. Beta = .42, Power = .58

When N = 20, we reject the null 92% of the time. Beta = .08, Power = .92

Score Card , Alpha =.01, 99% Confidence Intervals:

When N = 10, we reject the null 21% of the time. Beta = .79, Power = .21

When N = 20, we reject the null 75% of the time. Beta = .25, Power = .75

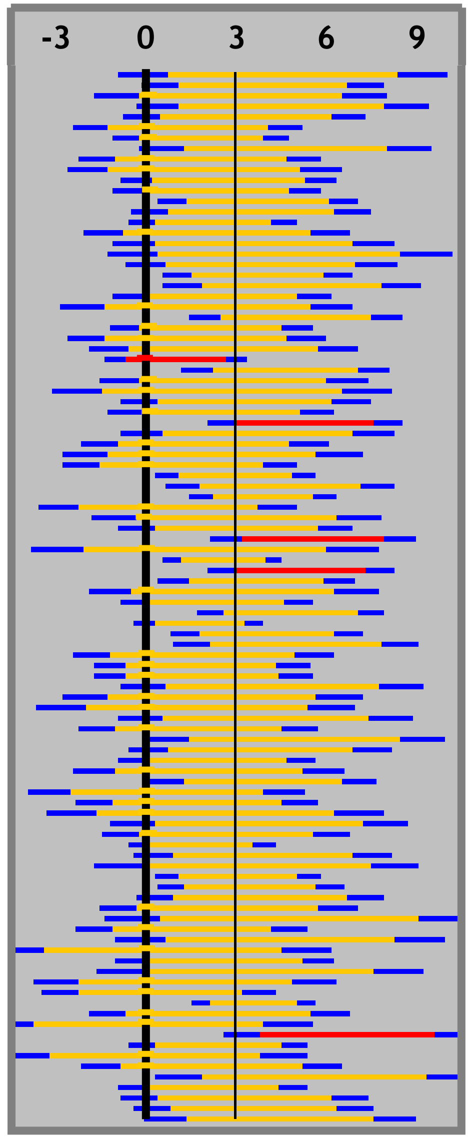
Given the choice, we always choose the larger sample size for the sake of greater precision. However, choosing our alpha level is less obvious. Which do you choose? Why?

Alpha level is the exact probability of Type I Error *when* the null is true.

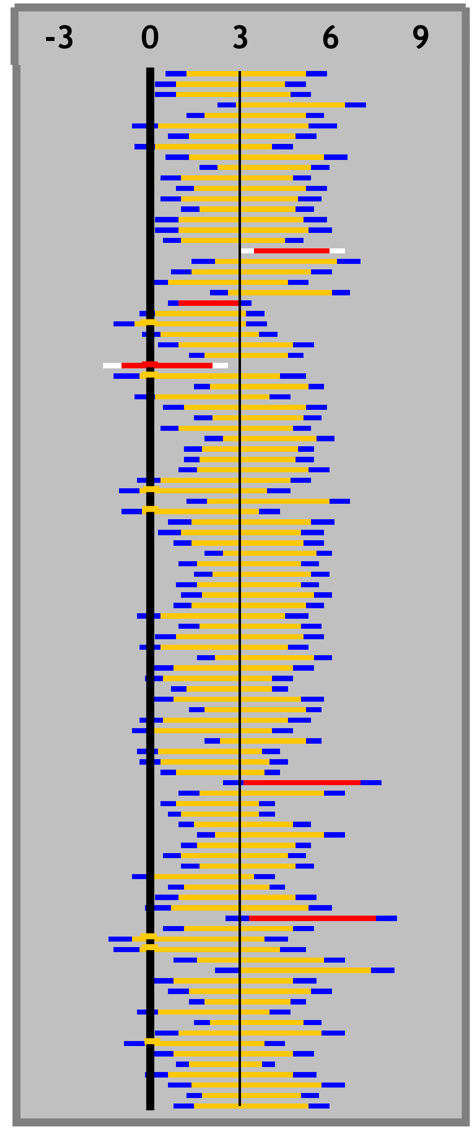
Beta level is the exact probability of Type II Error *if* a specified relationship is true.

		Our Conclusion	
		Positive	Negative
		We reject the null, so we find a relationship in the population.	We fail to reject the null, so our findings are inconclusive.
Our World	There is in fact a relationship in the population.	True Positive	False Negative "Type II Error"
	There is in fact *no* relationship in the population	False Positive "Type I Error"	True Negative

N = 10



N = 20



Dialectic of Statistical Inference (Part I of II)

In my random sample, I found that the intervention group scored 5 points higher on average than the control group ($r = .17$, $p < .05$). Intervention predicted 3% of the score variation.

I suspect that there is no intervention effect, that the relationship you observe in your sample is merely an artifact of sampling error and not reflective of the population.

Well, the $p < .05$ tells us that you might be right (it's not $p = 0$), but if you were right, and thus there were no relationship in the population, we would only observe a relationship so strong (or stronger) less than 5% of the time. That is pretty unlikely.

I see. My suspicion of 0.00000 relationship is not very plausible.

Okay, so we do not want to conclude that the relationship is exactly zero in the population, but that does not mean the relationship is exactly .17 as you observe in your sample.

That's right. We don't want to conclude that the Pearson correlation in the population is exactly .17. Likewise, we do not want to conclude an intervention effect of exactly 5 points. However, they are unbiased estimates of the population values.

How precise are those estimates?

Using the same standard errors that we used to calculate $p < .05$ and reject the null hypothesis, we can construct 95% confidence intervals. Thus, our estimate for the population correlation is $.17 \pm .14$ and, for the intervention effect, 5 ± 4.

Dialectic of Statistical Inference (Part II of II)

Your intervention is boosting children's scores on average. A 1-point boost is your lower-bound estimate for that average, and a 9-point boost is your upper-bound estimate for that average. Should your intervention be an educational funding priority?

That is a difficult question. I can tell you all about ***statistical*** significance, but your question is about ***practical*** significance. To determine whether the intervention is worth implementing, we need to conduct a benefits-costs analysis: Do the benefits of the intervention outweigh its costs? Among the costs, we must consider opportunity costs: How does our intervention stack up against similarly targeted interventions?

My wife is an economist. I'll have her people call your people.

Back to statistical significance—The whole process of inference from a sample to a population is heavily laden with assumptions. If those assumptions do not hold, then it's all lies.

Yes. If my assumptions weren't tenable, my standard errors and, consequently, p-values and confidence intervals would be biased, and I would not be reporting them. Give me a break.

Hey, we're all friends. Were there *any* worrisome assumptions?

Independence, normality, linearity and outliers were okay, but there was a little Heteroscedasticity; there was a little less variation in the intervention group than the control group. Heteroscedasticity won't bias our magnitude and strength estimates, but it will bias our precision estimate (i.e., standard error). Next semester, Sean will show me how to fix it.

Unit 8: Research Questions

Theory 1: Since depression leads to introversion, and reading is an introverted activity, depressed children will be stronger readers than non-depressed children.

Research Question 1: In children of immigrants, reading achievement is positively correlated with depression levels.

Theory 2: Since depression conflicts with cognitive functioning, and reading is a cognitively demanding activity, depressed children will be weaker readers than non-depressed children.

Research Question 2: In children of immigrants, reading achievement is negatively correlated with depression levels.

Data Set: ChildrenOfImmigrants.sav

Variables:
- Outcome—Reading Achievement Score (*READING*)
- Predictor—Depression Level (*DEPRESS*)

Model: $READING = \beta_0 + \beta_1 DEPRESS + \varepsilon$

SAT.sav Codebook

Portes, Alejandro and Ruben G. Rumbaut
Children of Immigrants Longitudinal Study
(1992, 1995)

"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).

Subset of data: Random sample of 880 participants obtained through the website.

Selected references:

Portes, Alejandro, & Ruben G. Rumbaut (2001). *Legacies: The Story of the Immigrant Second Generation*. Berkeley CA: University of California Press.

More information is available at: http://cmd.princeton.edu/ (Center for Migration and Development, Princeton University)

ChildrenOfImmigrants.sav Codebook

Variable Name	Variable Description	Characteristics
ID	Identification #	Integers
Reading	Stanford Reading Achievement Score	Range: 527-830 Mean: 669
FreeLunch	% students in school who are eligible for free lunch program	Range: 0-92.30 Mean: 45.27
Male	Sex dummy variable	1=Male 0=Female
Depress	Depression scale (Higher score means more depressed)	Range: -1.68 - 5.57 Mean: 0.00
SES	Composite family SES score	Range: -1.66 - 2.09 Mean: -0.04

The Children of Immigrants Data Set

ChildrenOfImmigrants.sav [DataSet1] - SPSS Data Editor

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

1 : ID 4414

	ID	Reading	FreeLunch	Male	Depress	SES
1	4414	558	70.8	0	0.95820	-1.37
2	282	570	27.5	0	0.50933	-0.38
3	3848	580	92.3	1	0.94126	-1.02
4	342	581	27.5	1	-0.72716	0.25
5	3805	584	38.2	1	1.96360	0.07
6	4301	584	82.0	1	0.33707	-0.81
7	3548	586	82.0	0	-0.47796	-1.10
8	2593	589	38.2	0	2.45614	-1.65
9	3545	589	82.0	0	3.14966	-0.35

ChildrenOfImmigrants.sav [DataSet1] - SPSS Data Editor

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

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	ID	Numeric	4	0		None	None	8	Right	Scale
2	Reading	Numeric	3	0	Stanford Readi...	None	None	8	Right	Scale
3	FreeLunch	Numeric	4	1	% of Students i...	None	None	8	Right	Scale
4	Male	Numeric	1	0	Male = 1, Fem...	{0, Female}...	None	8	Right	Nominal
5	Depress	Numeric	8	5	Depression Sc...	None	None	8	Right	Scale
6	SES	Numeric	5	2	Composite Fa...	None	None	8	Right	Scale
7										

Location Location Location: The Null Hypothesis Approach

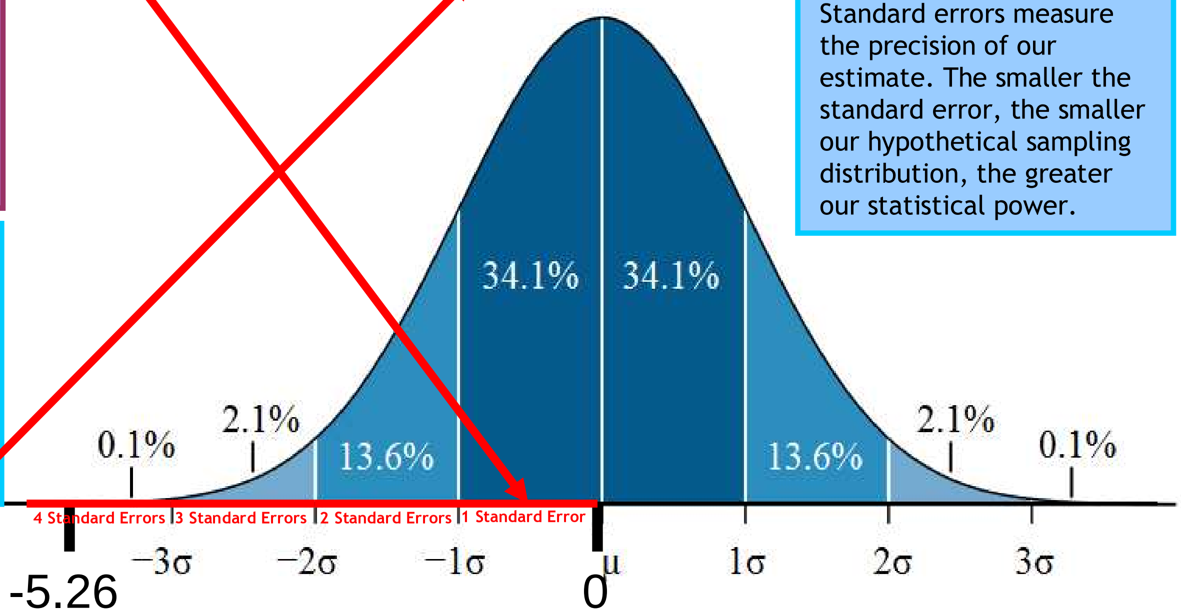
We know that our trend line belongs to a butterfly, and we can guesstimate the spread of the butterfly. We know that our slope belongs to a bell curve, and we can guesstimate the spread of the bell curve. We do not know the location of the butterfly or the bell curve, but we can use our knowledge of the shapes and spreads of the butterfly and bell curve.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	671.607	1.275		526.746	.000	669.105	674.110
	Depress	-5.260	1.429	-.123	-3.680	.000	-8.066	-2.455

There is a statistically significant relationship ($p < 0.05$) between depression levels and reading scores in our sample ($n = 880$) of children of immigrants.

Standard errors measure the precision of our estimate. The smaller the standard error, the smaller our hypothetical sampling distribution, the greater our statistical power.

A t-test is a test to see if our observation is a sufficient number of standard errors away from zero to scare us into rejecting the null hypothesis. A t statistic of ± 2, indicating that our observation is ± 2 standard errors from zero, will have a two-tailed significance level (or p value) of about 0.05.



Our observed slope is -3.68 standard errors from zero.

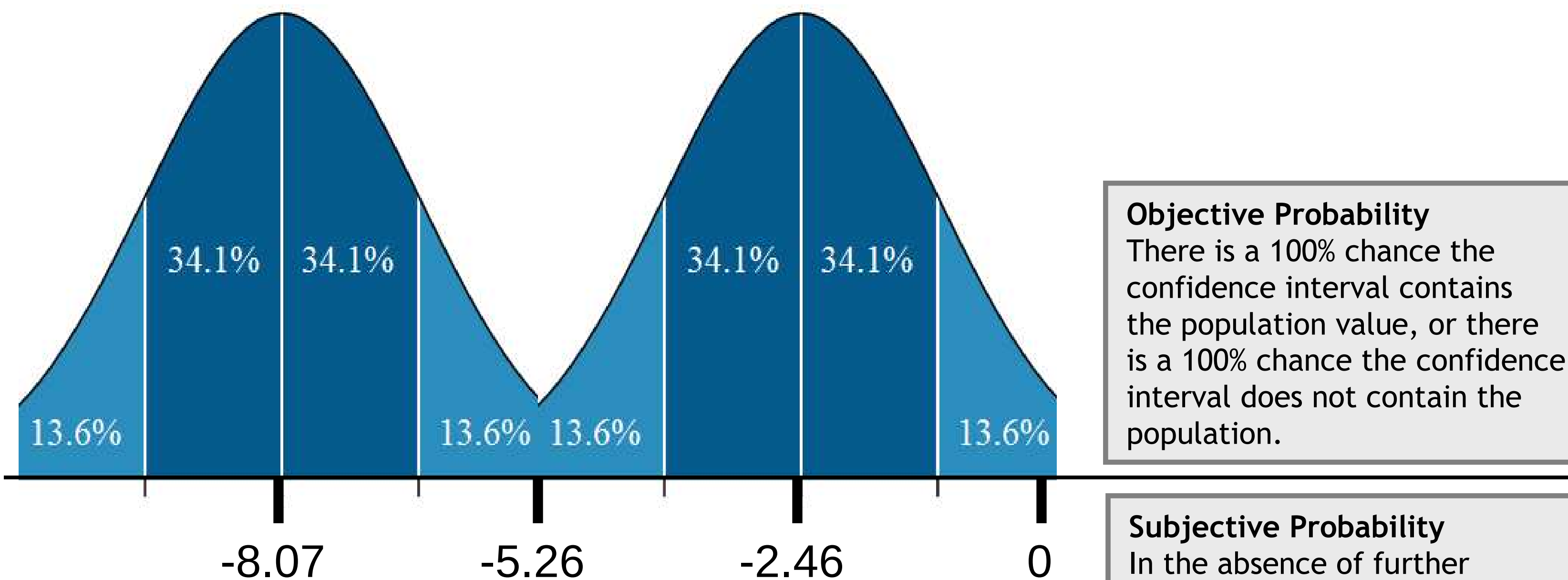
Quiz: What is -5.26 divided by 1.429?

We observe a regression coefficient of -5.26 in our sample. If there were no relationship in the population, we would observe a coefficient this large or larger in less than 0.01% of our samples.

Location Location Location: The Confidence Interval Approach

The purpose of confidence intervals is to contain the true population parameter. Over your lifetime, 95% of (95%) confidence intervals will succeed and 5% will fail. You will not know which are the unlucky 5%! Analogous reasoning holds for alpha level.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	671.607	1.275		526.746	.000	669.105	674.110
	Depress	-5.260	1.429	-.123	-3.680	.000	-8.066	-2.455



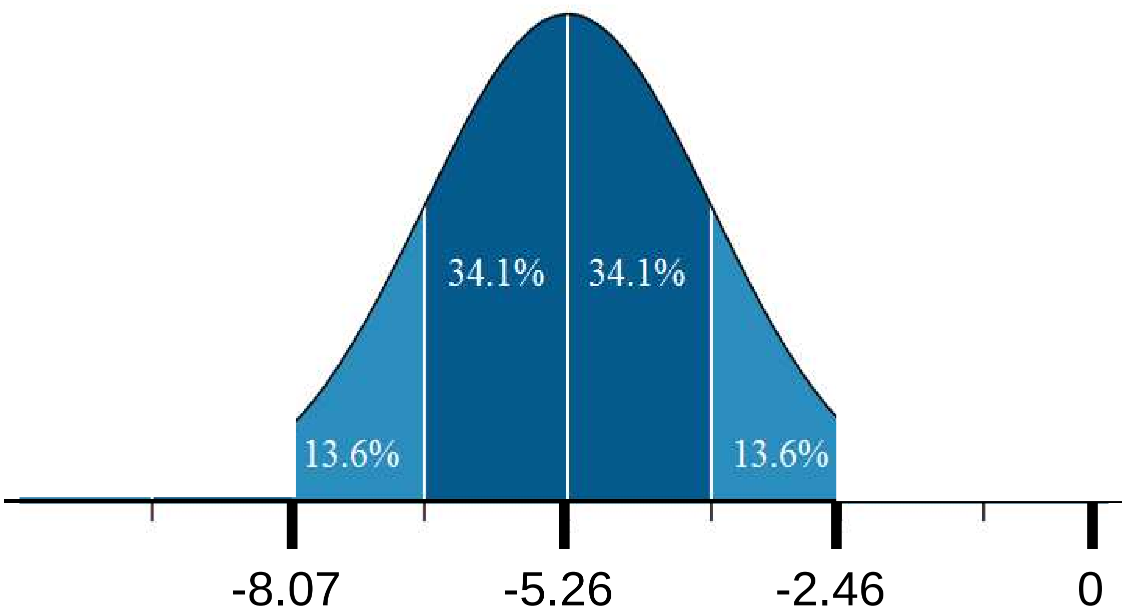
Objective Probability
There is a 100% chance the confidence interval contains the population value, or there is a 100% chance the confidence interval does not contain the population.

Subjective Probability
In the absence of further information, it is reasonable to conclude that there is a 95% chance the confidence interval contains the population value.

Location Location Location: The Confidence Interval Approach

The purpose of confidence intervals is to contain the true population parameter. Over your lifetime, 95% of (95%) confidence intervals will succeed and 5% will fail. You will not know which are the unlucky 5%! Analogous reasoning holds for alpha level.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	671.607	1.275		526.746	.000	669.105	674.110
	Depress	-5.260	1.429	-.123	-3.680	.000	-8.066	-2.455



Objective Probability
There is a 100% chance the confidence interval contains the population value, or there is a 100% chance the confidence interval does not contain the population.

Subjective Probability
In the absence of further information, it is reasonable to conclude that there is a 95% chance the confidence interval contains the population value.

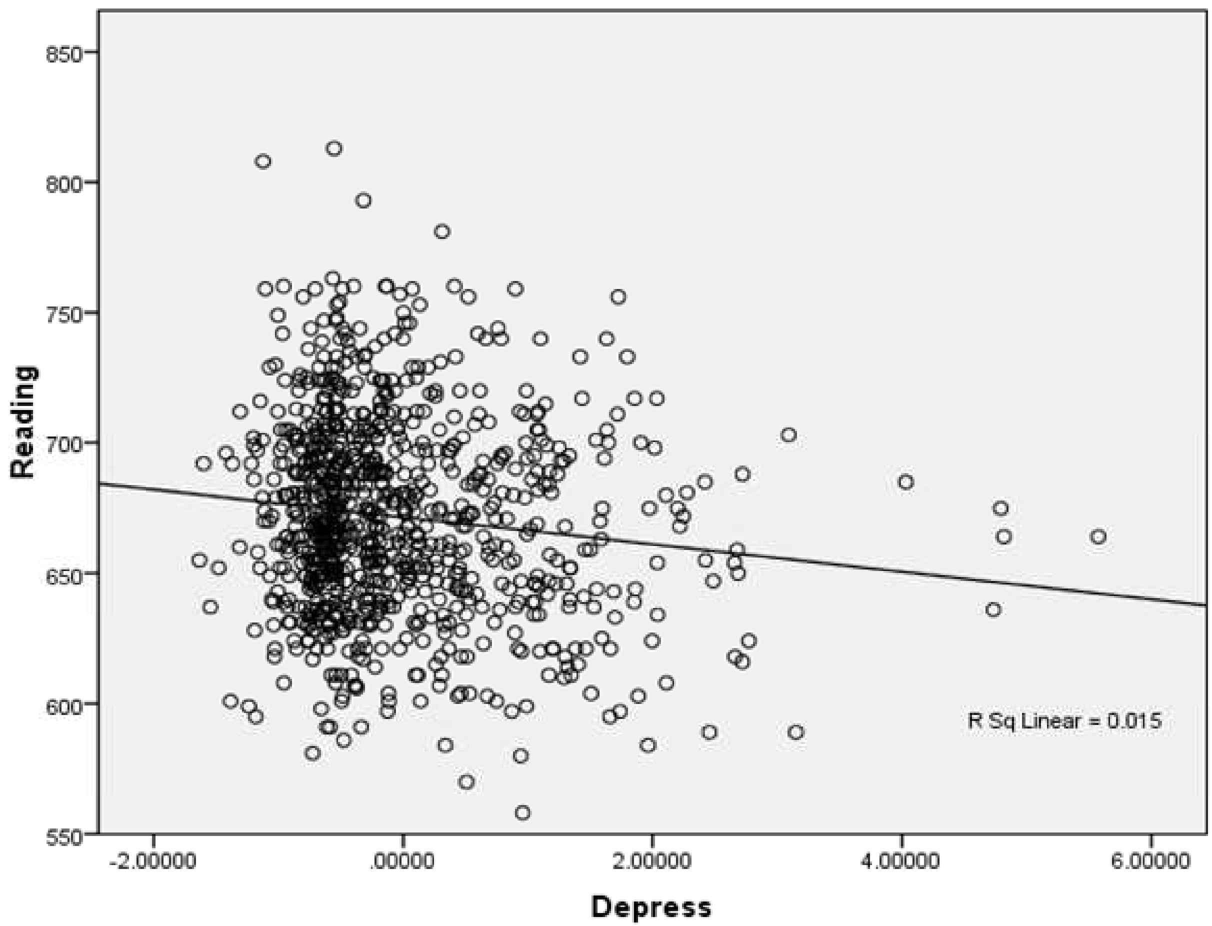
Linear Regression Assumptions

Search HI-N-LO for assumption violations that will threaten your statistical inference from the sample to the population.

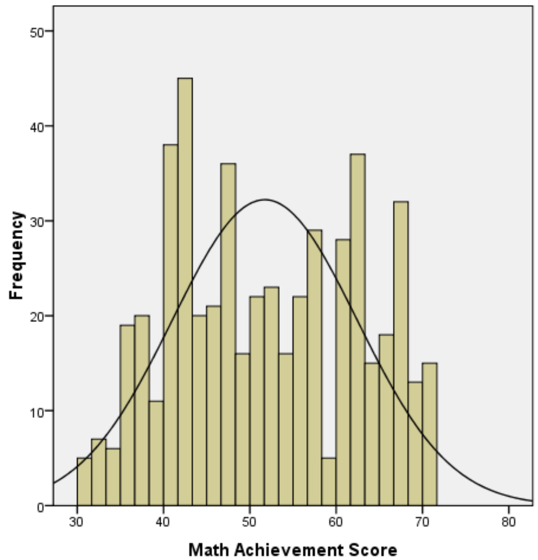
- **H**omoscedasticity
- **I**ndependence
- **N**ormality
- **L**inearity
- **O**utliers

Other assumptions (that we will not cover this semester, because we need measurement theory):

- Zero measurement error in our *predictors*. In our *outcome*, measurement error is okay (not ideal, but okay).
- All continuous variables are on an interval scale where units have the same meaning all along the continuum.

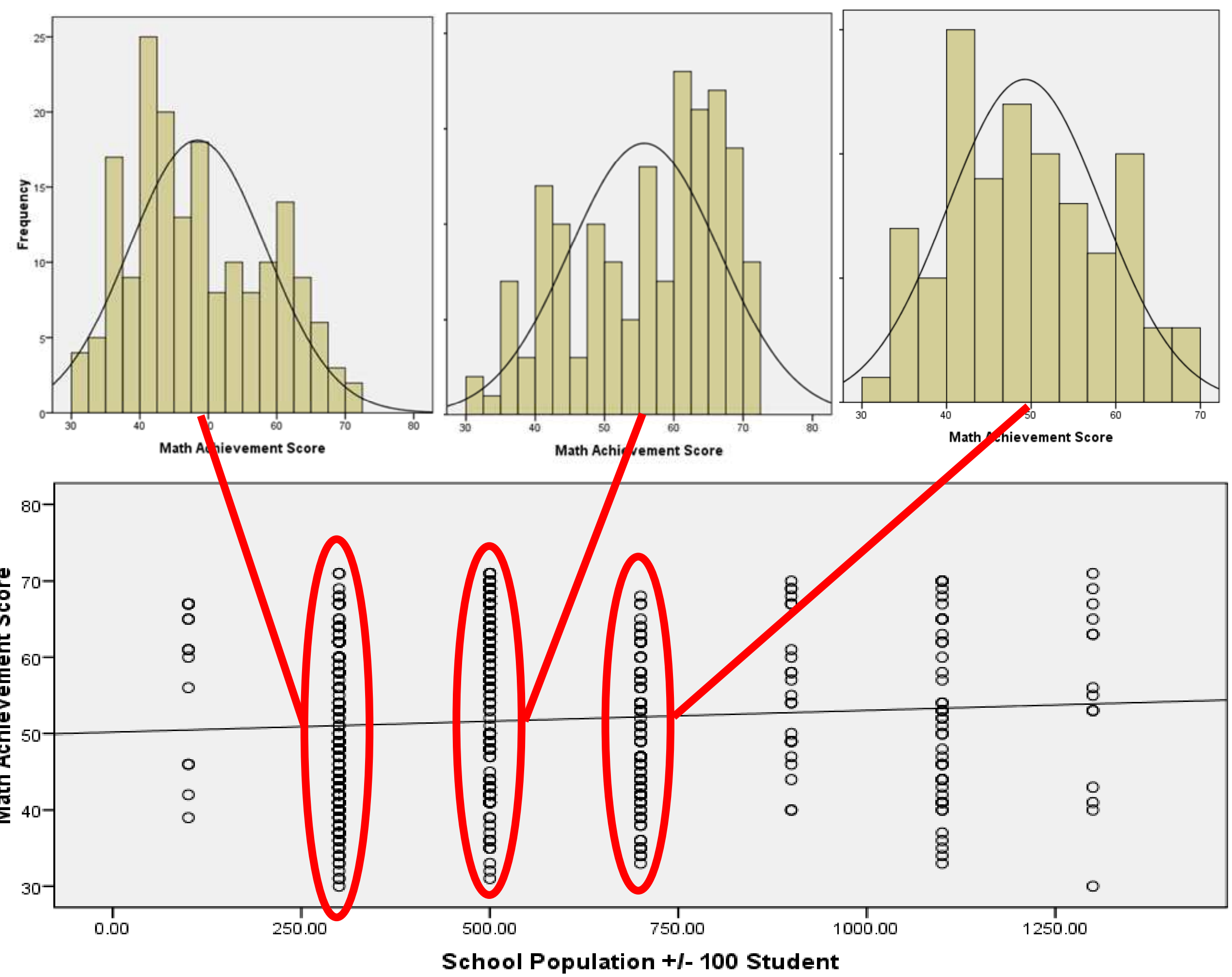


Exploring Math Achievement and School Size (Reprise)



Mean =51.72
Std. Dev. =10.709
N =519

Figure 6.?. Histograms of math achievement scores for students from schools with populations of about 300 (n=181), 500 (n=154), and 700 (n=89).



Think vertically!

The assumptions of homoscedasticity and normality are about the distributions of the outcome conditional on the predictor.

Directly above, we see the distribution of the outcome (Y), unconditional.

To the right, we see distributions of the outcome (Y) conditional on the predictor (X).

Homoscedasticity (Good) vs. Heteroscedasticity (Bad)

A bivariate relationship is <u>homoscedastic</u> when the distributions of Y conditional on X have equal variances (i.e., equal spreads).

A funnel shape gives a visual clue to ***violations of homo***scedasticity. I.e., funnels show ***hetero***scedasticity.

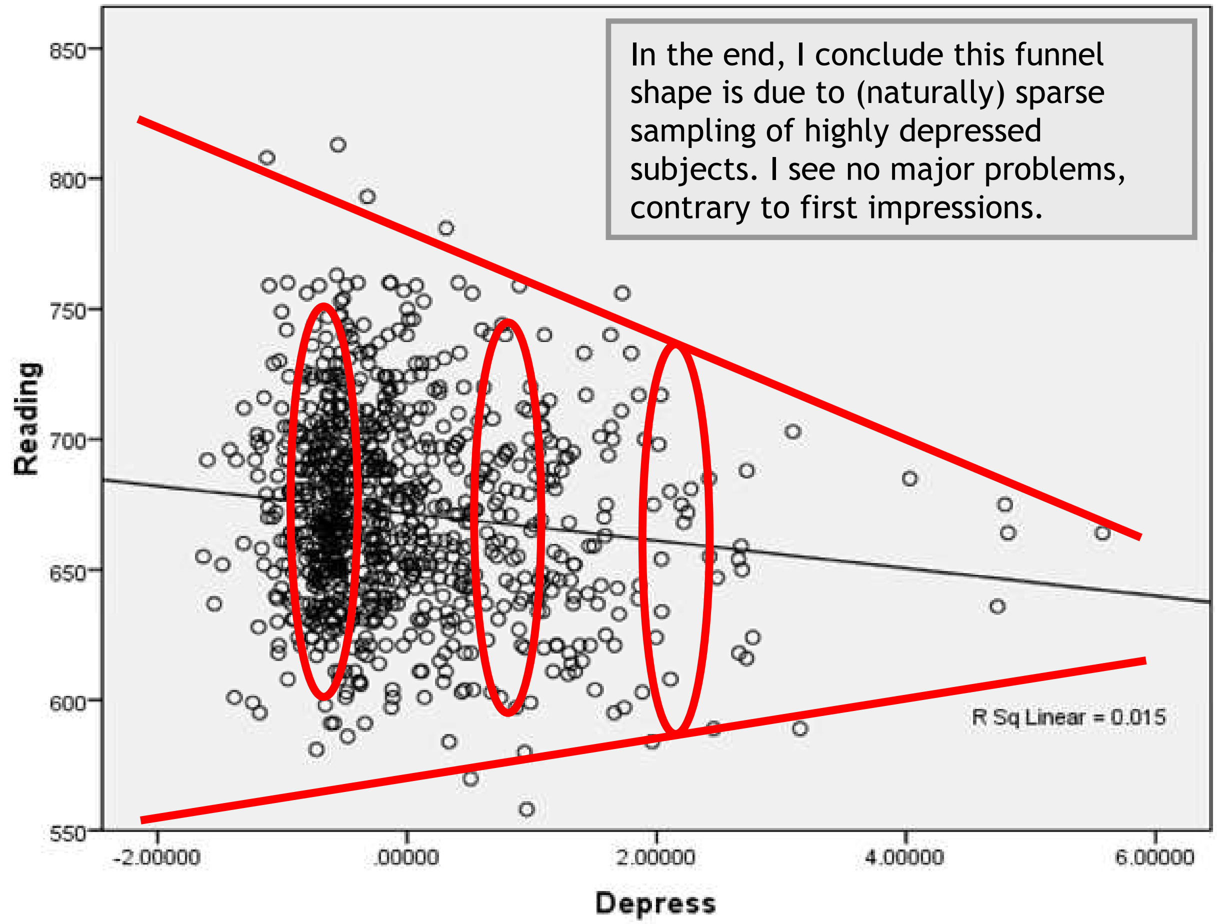
However, this assumption is about the population, so don't be fooled by small sample sizes of Y conditional on X. You must conjecture about how the distribution would spread out if you added more observations from the sparse values of X.

Fixes For Future Reference:

Two-sample t-tests can be calculated with weighted standard errors.

Regression t-tests can be calculated with robust standard errors.

Sometimes the inclusion of an interaction term can solve a heteroscedasticity problem.



In the end, I conclude this funnel shape is due to (naturally) sparse sampling of highly depressed subjects. I see no major problems, contrary to first impressions.

We guesstimate a standard error from the observed variance around the regression line, but, if the observed variance varies by level of the predictor, i.e., the relationship is heteroscedastic, then which variance should we use? Heteroscedasticicity biases (upwards) our estimate of the standard error, but it does not bias our estimate of the slope.

Normality (Good) vs. Non-Normality (Bad)

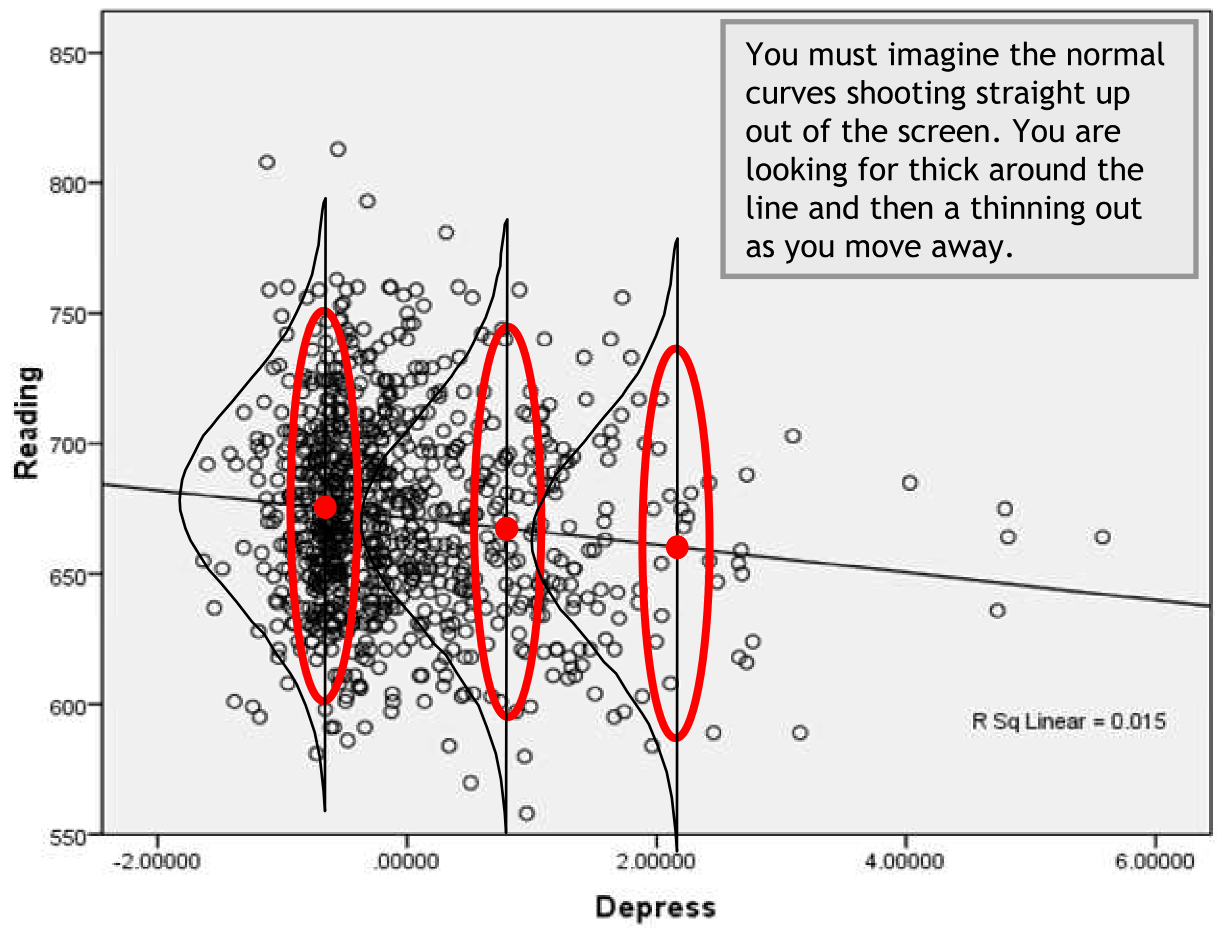
A bivariate relationship meets the <u>normality</u> assumption when Y is normally distributed conditional on X.

Lopsided outliers give a clue to non-normality.

However, this assumption is about the population, so don't be fooled by small sample sizes of Y conditional on X. You must conjecture about how the distribution would shape up if you added more observations from the sparse values of X.

Fixes For Future Reference:

A nonlinear transformation of Y will change the shapes of the distributions.



The Central Limit Theorem tells us that the sampling distribution of the slope estimate is approximately normal, no matter what. Great. We also need to know the standard deviation (i.e., standard error) of the sampling distribution to test null hypotheses and construct confidence intervals. We don't ***know*** the standard error, but we can estimate it. Whenever we estimate from a sample, we have to worry about sampling error, so we have to consider the sampling distribution of the standard error estimate (an additional sampling distribution). The Central Limit Theorem tells us that the sampling distribution of the standard error is NOT normal. Whereas sampling distributions for slopes are always approximately the same shape (normal), sampling distributions for the standard errors are skewed and skewed to different extents depending on the distributions of Y conditional on X. If we want to understand exactly the additional uncertainty from estimating standard errors, we need to know the distribution of Y conditional on X. If, for example, we know that the distribution of Y condition on X is normal, then we can account for the extra uncertainty perfectly in our calculations. Hence, the normality assumption. (If we have a great estimate of our standard error, then the normality assumption is needless.)

T-Tests Incorporate the Uncertainty From Our Standard Error Estimates

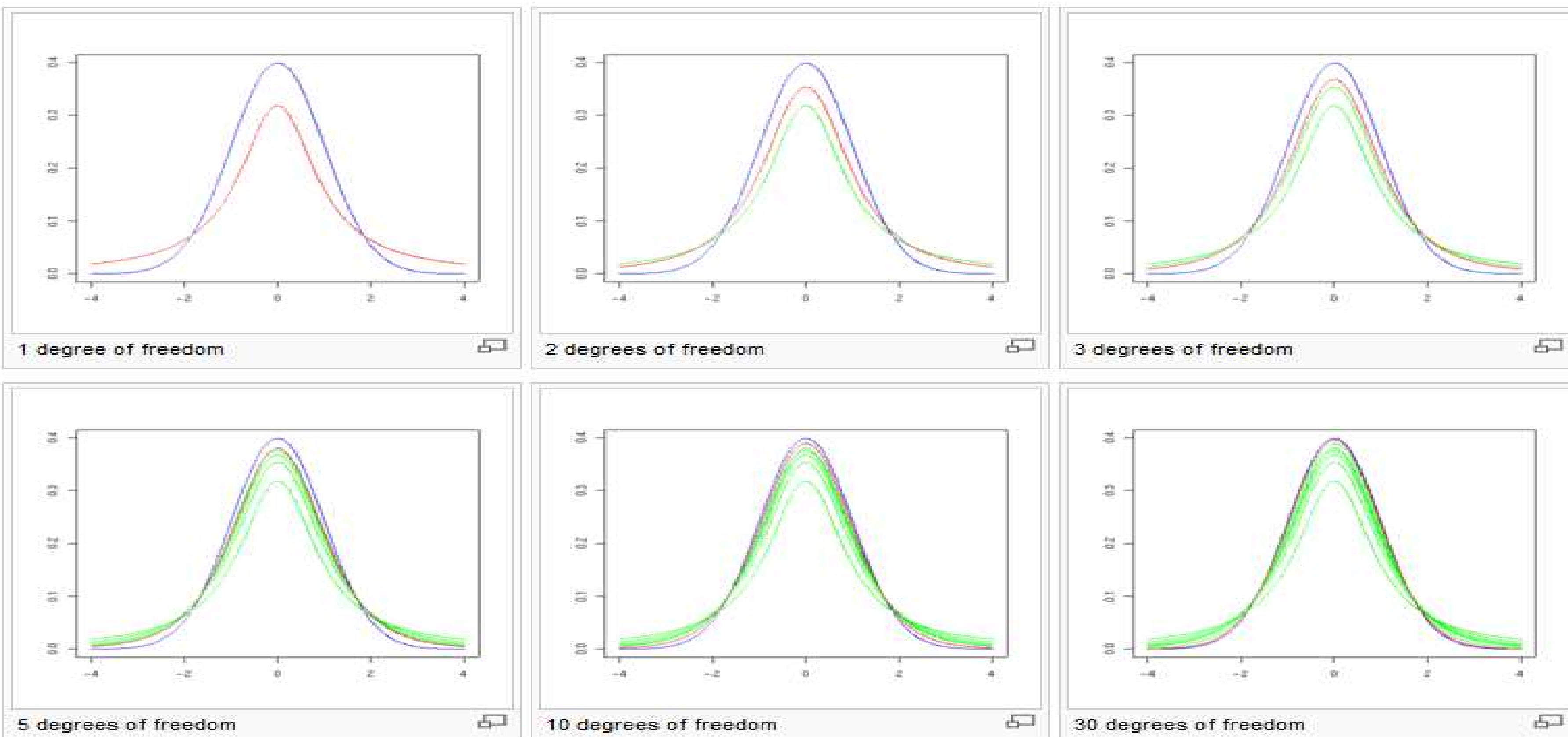
When we recognize the problem of sampling error for our slope (i.e., we recognize that our sample slope is only an estimate of the population slope), what do we do? We use our standard errors to test the statistical significance of the sample slope or to build confidence intervals around the sample slope. But, but, but… our standard errors are also subject to sampling error (i.e., our sample standard error is only an estimate of the population standard error). Ugh! We are using one estimate to check another estimate! Three responses:

First, welcome to the human condition. Each of us has a system of understanding, but that system is merely a network of more or less uncertain relations. Our best estimates from a web of belief. Best estimates are all we have. We feel better about an estimate when it fits snugly with our other best estimates.

Second, if the normality assumption holds, then we can puff out the sampling distribution of the slope to account for the added uncertainty.

Third, if our sample size is big enough, then we'll get a good estimate of the population standard error with minimal added uncertainty.

Density of the *t*-distribution (red) for 1, 2, 3, 5, 10, and 30 df compared to the standard normal distribution (blue). Previous plots shown in green. From Wikipedia.



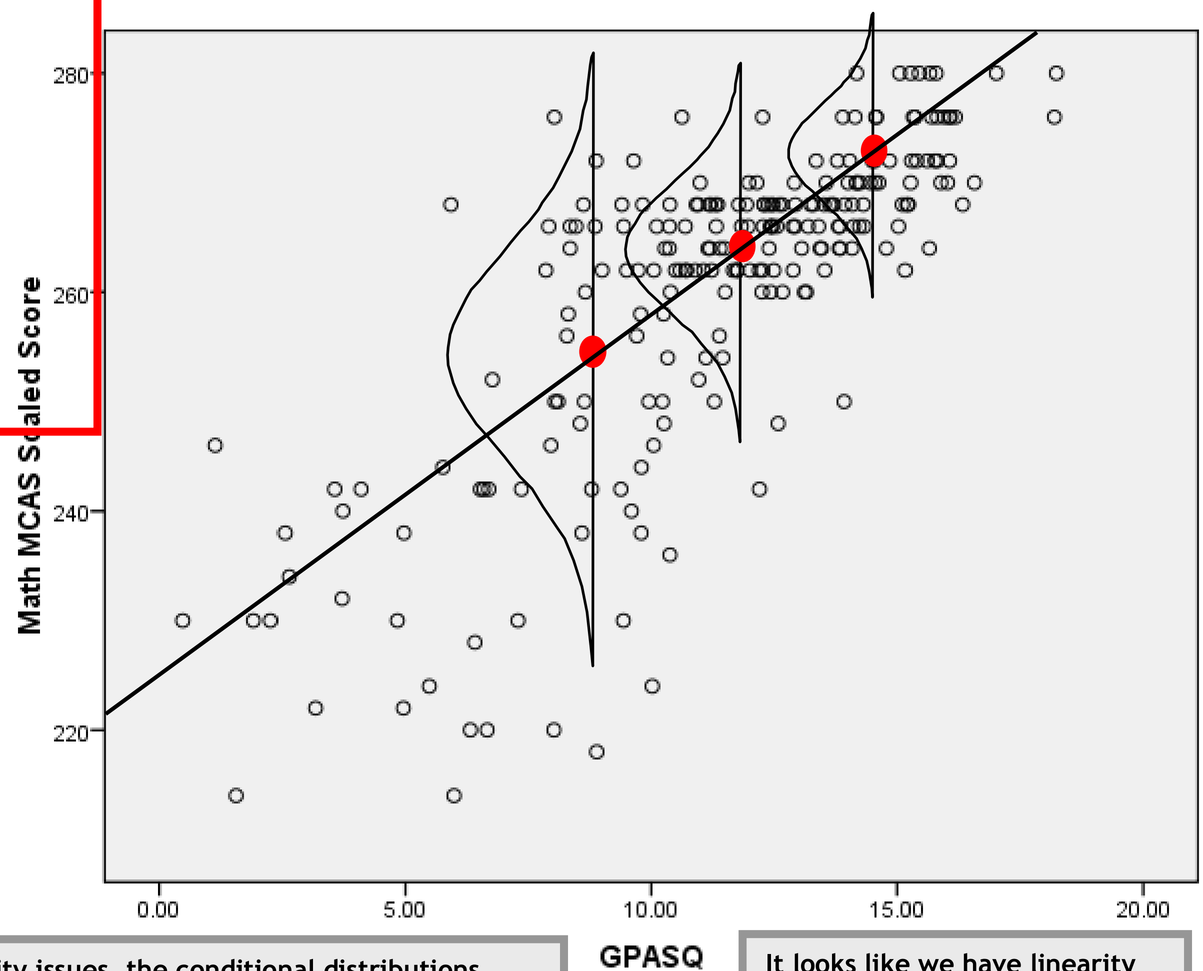
Homoscedasticity and Normality: Another Example

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	223.437	2.132		104.819	.000	219.236	227.638
	GPASQ	3.244	.180	.772	18.070	.000	2.890	3.598

The problem with heteroscedasticity is not in our parameter estimates: the intercept and slope coefficients will be unbiased, because they are conditional averages, and conditional averages do not care about conditional variances.

The problem is in our standard errors. Recall that a standard error is just a special kind of standard deviation—the standard deviation of THE sampling distribution. But, when there is a different sampling distribution at each level of X, which do we choose?

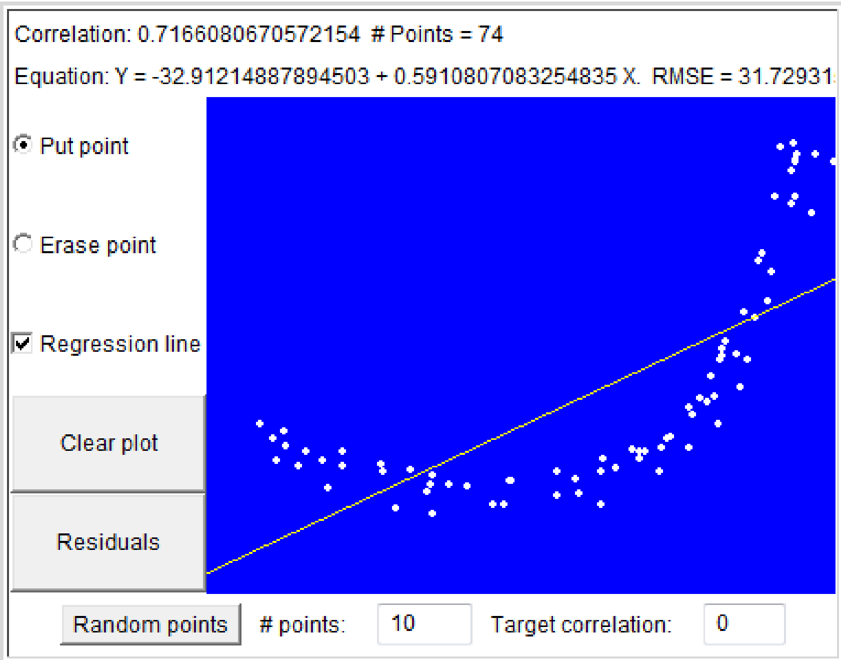
The problem with non-normality derives from our standard error being only an estimate. This adds a second layer of uncertainty (to our slope being only an estimate). However, if the normality assumption holds, we can perfectly account for that second layer of uncertainty in our t-test.



This relationship appears to have heteroscedasticity issues, the conditional distributions differ in variance. The conditional distributions appear normal, or at least symmetric.

It looks like we have linearity issues! This is scariest!

Linearity and Outliers (We Know This From Unit 1!)



Non-Linearity Fixes For Future Reference:

Non-linearily transform Y and/or X.
http://onlinestatbook.com/stat_sim/transformations/index.html

Use non-linear regression.

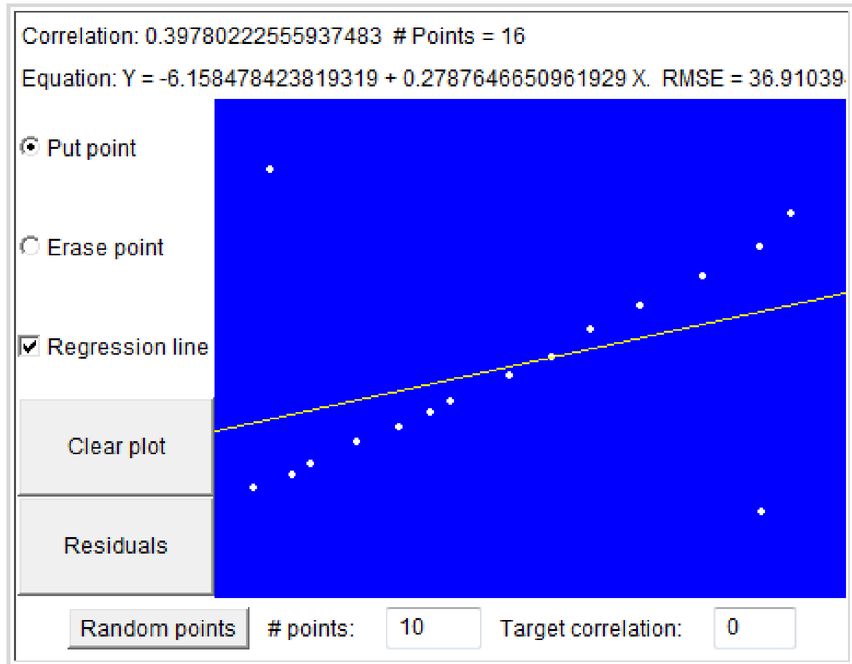
Whereas herterscedasticity, non-independence, and non-normality bias the standard errors but not the slope coefficient, non-linearity and outliers bias the slope coefficient and, consequently, the standard errors.

http://www.people.vcu.edu/~rjohnson/regression/

Outlier Fixes:

Sometime outliers are due to coding errors. (I've seen a study where the correlation was determined by a 9 month old fourth grader!) If it's just a coding error, fix the error or discard the observation.

Sometimes, you may want remove the outlier(s) and refit the model. If you do this, make sure your audience knows that you did it. Consider giving your audience members both sets of results and allowing them to choose.



Independence (Good) vs. Non-Independence (Bad)

A bivariate relationship meets the <u>independence</u> assumption when the errors are uncorrelated.

A scatterplot yields no visual clues. Generally, you must rely on substantive knowledge of the data-collection method.

Students are clustered within classrooms which are clustered within schools which are clustered within districts, which are clustered within states…

If students in your sample hang together in clusters, you may have less information than appears.

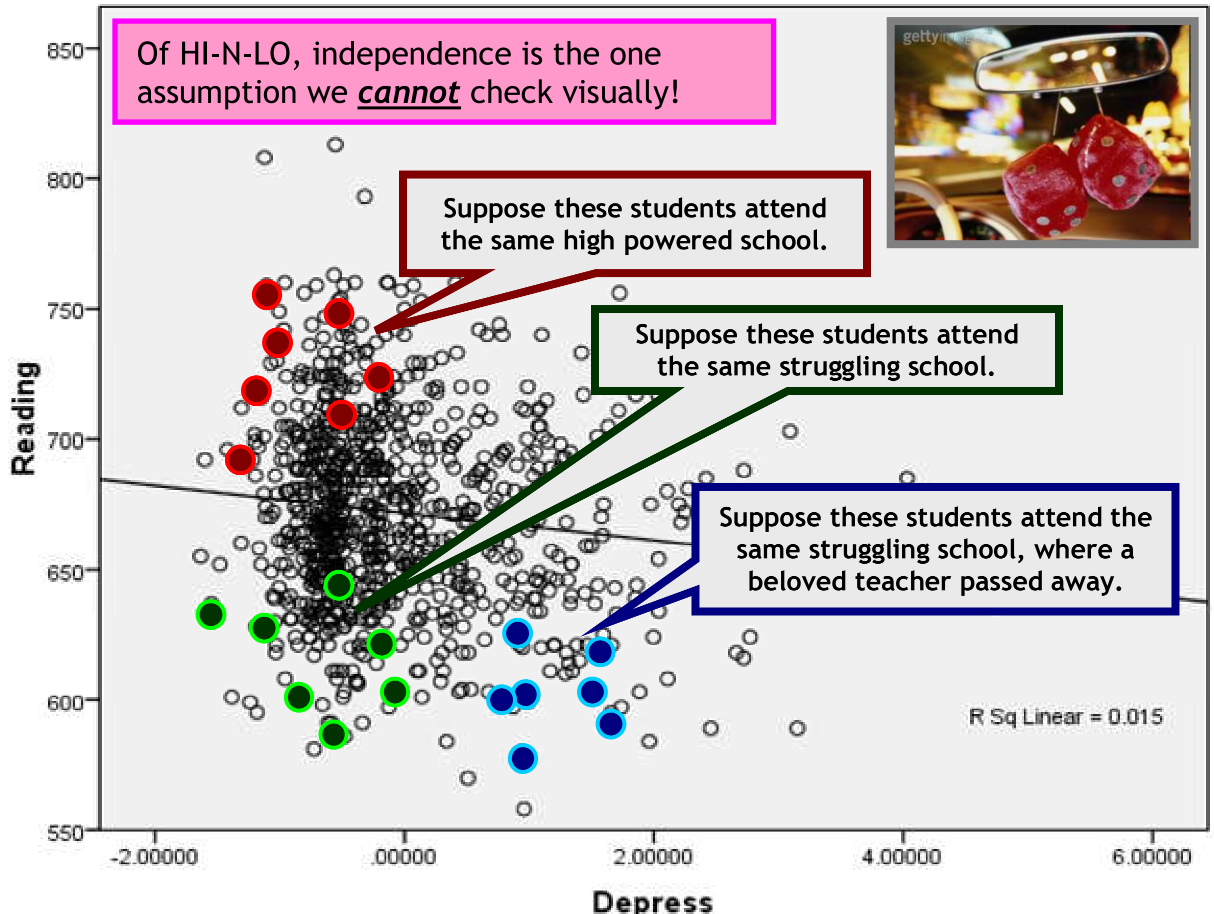
Fixes For Future Reference:

Use within subjects ANOVA.

Use multilevel models. Hierarchical linear modeling (HLM) is one type.

Fixes For Now:

If you have pretest and posttest results, you can use a within subjects t-test that accounts for the correlation due to the fact that, in your scatterplot, you have two data points for each student. (See Math Appendix.)



We use the Central Limit Theorem to guesstimate our standard errors, and the Central Limit Theorem relies on randomness. When observations are clustered, there is order in what should be essentially random. Thus, non-independent data bias our standard errors. Our effective sample size may be smaller than the number of observations if the observations hang together in clusters.

The Problem of Non-Independence: What is your Real Sample Size?

In studies of students nested within schools, what is the extent of the non-independence? The answer is going to depend on our outcome. Reading scores? Emotional disorders? Community service? Self esteem? Locus of control? For giggles, suppose that our outcome has to do with school clothing, and our data include students clustered within schools. Below are two school-clothing studies, each with its own data set. Which of the two data sets has the bigger problem of non-independence?

What is your sample size? Is your sample size equal to the number of students, or the number of schools, or something in between? Next semester, in Unit 19 we will answer this question with intraclass correlation, and we'll account for intraclass correlations with multilevel models.

Study 1

Study 2

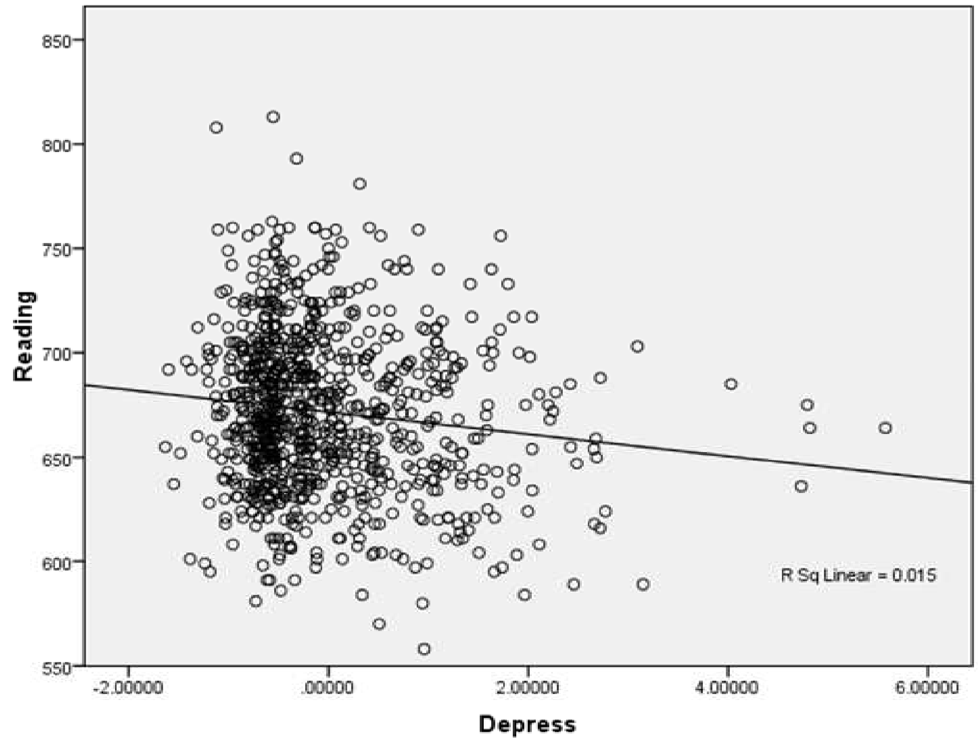
Linear Regression Assumptions (HI-N-LO)

Homoscedasticity:
Although the variation in reading scores appears to be less at higher levels of depression, we believe that this is merely an appearance due to sparse sampling of extremely depressed subjects. If we were to sample more extremely depressed subjects, we suspect that we would find their variation in reading equal to the variation of non-depressed subjects.

Independence:
Although we cannot test for independence visually, we suspect that students are clustered within schools, and this may be biasing our standard error. We may need to use multi-level regression models to account for the non-independence.

Normality:
At each value of depression, the distribution of reading scores appears roughly normal. For example, it appears that, for students with depression scores of 0.0, their depression scores are roughly normal, Likewise for students with depression scores of -1.0, 2.0 and 5.0.

Search **HI-N-LO** for assumption violations that will threaten your statistical inference from the sample to the population.



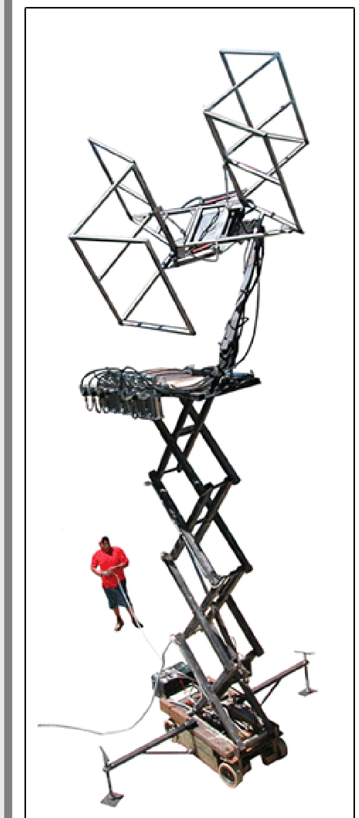
You already have a lot of practice looking for linearity and outliers. There's no need to rehash here, but don't forget the LO in HI-N-LO.

Linearity and (no) Outliers are by far the most important regression assumptions!

Checking Regression Assumptions

Search **HI-N-LO** for assumption violations that will threaten your statistical inference from the sample to the population.

Homoscedasticity
Independence
Normality
Linearity
Outliers



REACH FOR THE SKY: Rubén Ortiz-Torres shows off his customized scissors lift, 'High 'n' Low Rider,' which will be unveiled at the opening ceremonies at SJ01 and then displayed at MACLA.

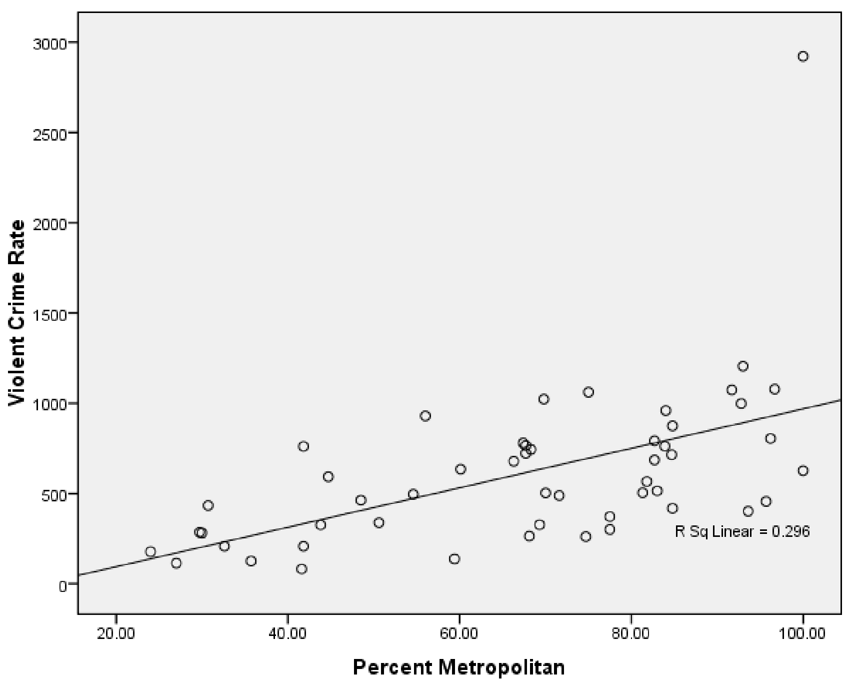
You have the concepts you need to complete the Unit 8 Post Hole. Practice is in back.

Dig the Post Hole

Unit 8 Post Hole:
Evaluate the assumptions underlying a simple linear regression.

Evidentiary material: bivariate scatterplot (the same as Unit 1).

Scatterplot of violent crime rate vs. urbanicity of the States (n = 51).



Notice that there is less variation in the sample at the lowest levels of urbanicity. Also notice that the sample size is smaller. If we had more rural states, would we see homoscedasticity? That's the question. The homoscedasticity (and normality) assumptions are about the population. Likewise, we really don't have large enough conditional samples to be confident about the normality assumption, but it's plausible, so we'll stick with it.

Here is my answer:

Homoscedasticity: No problem. No fan shape.
Independence: Can't see. Maybe clustering by region.
Normality: No prob. Thick around line, then spreads vertically.
Linearity: No prob.
Outliers: Yikes! Major top-right outlier.

For your consideration:

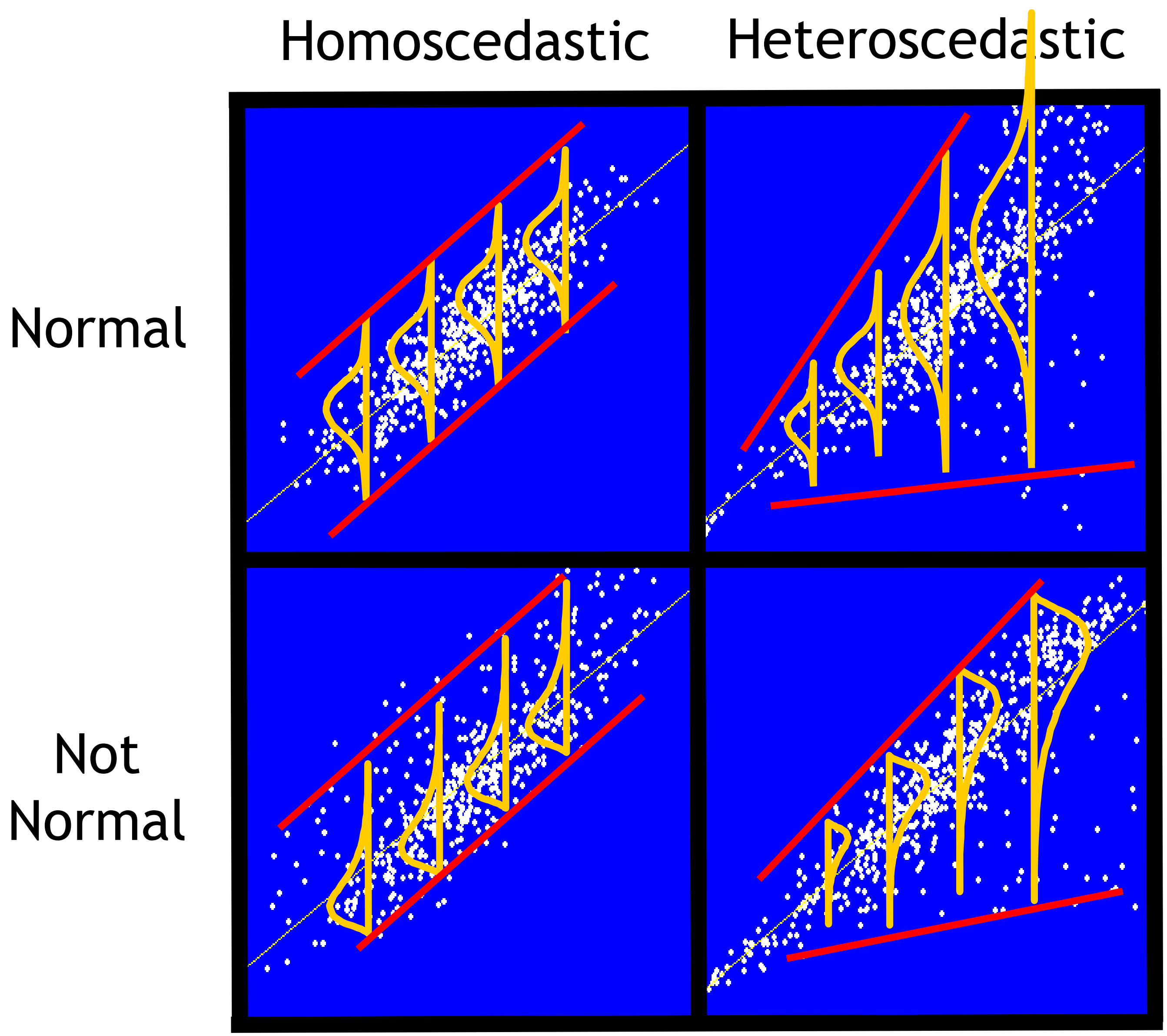
The outlier is Washington D.C. The 51st "state." I think any researcher could make a good case for exclusion here.

Q: What will happen to the slope if we remove D.C.?
A: It will decrease. Why?

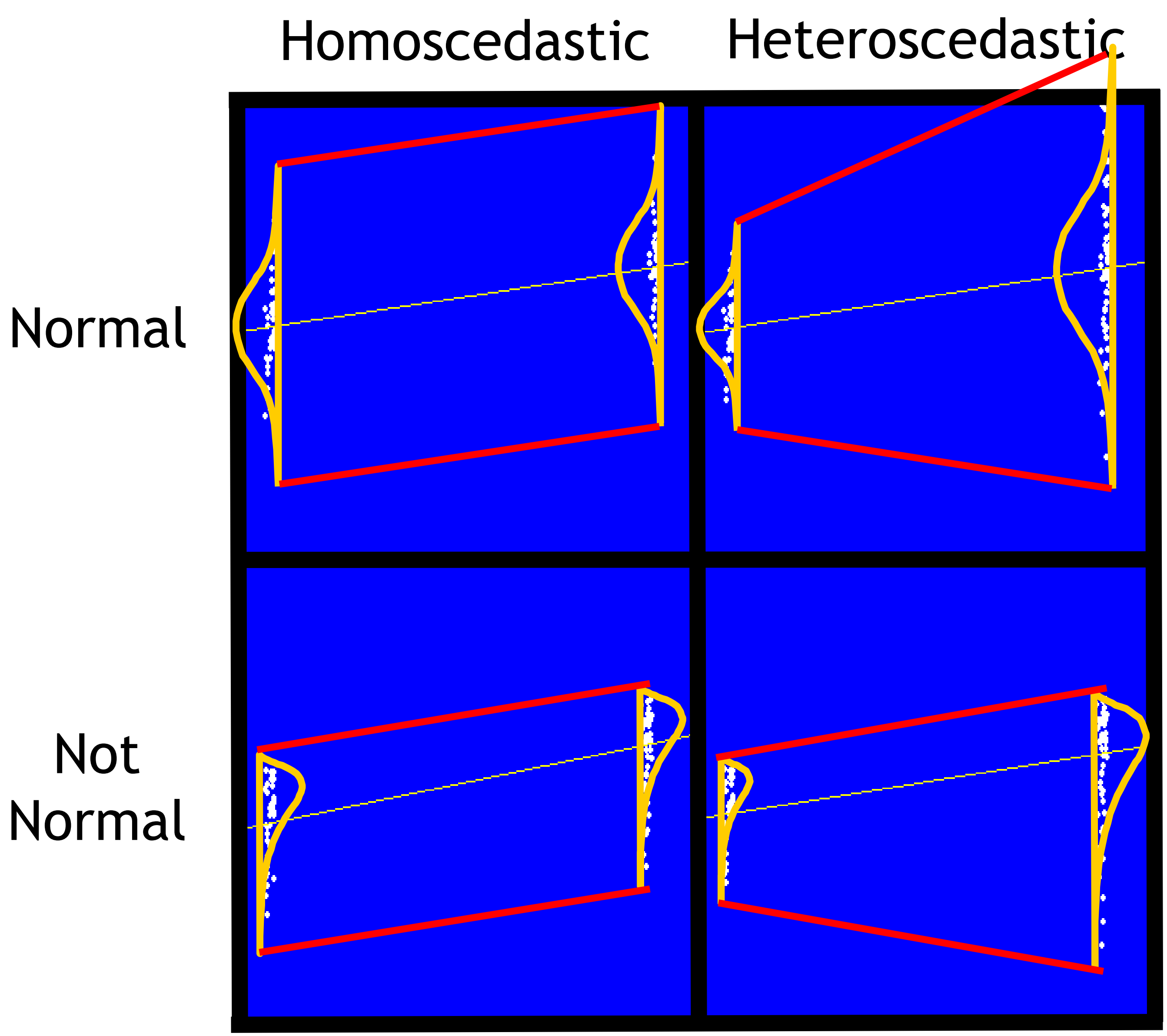
Q: What will happen to the R-square statistic if we remove D.C.?
A: It will increase. Why?

Q: If we have a "sample" of all 50 states, why do we have to make an inference from the "sample" to "population"? Don't we have the whole population?
A: If we are interested in demographics and census, then we *do* have the whole population. If, however, we are interested in *THEORIES OF WHY*, then we want to rule out randomness as the explanation for the relationship we see. To that end, we treat our sample as a random sample from a population (to which we'll never have further access). The null hypothesis is that the relationship in the sample is due purely to randomness (and if we had more states, the relationship would disappear).

Homoscedasticity and Normality (Abstractly) Exemplified (Part I of II)

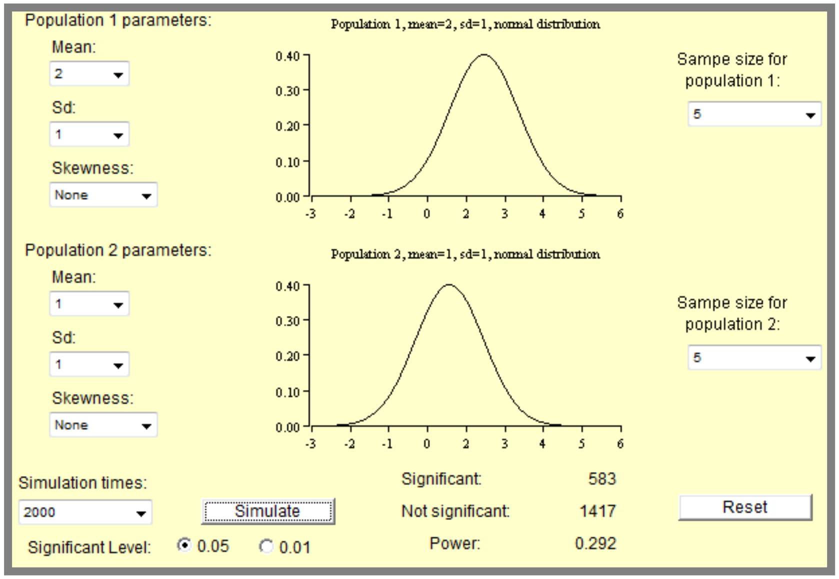


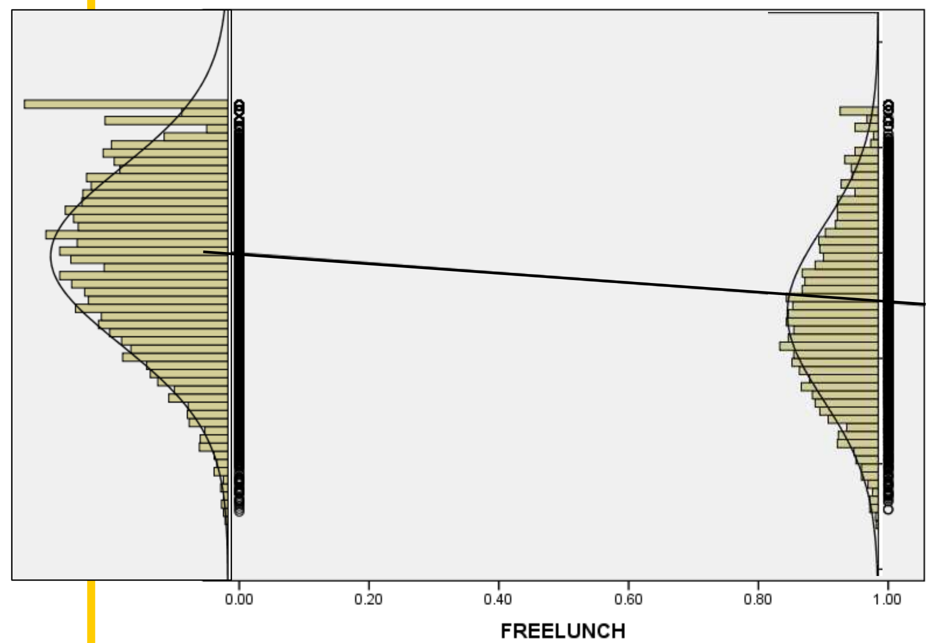
Homoscedasticity and Normality (Abstractly) Exemplified (Part II of II)

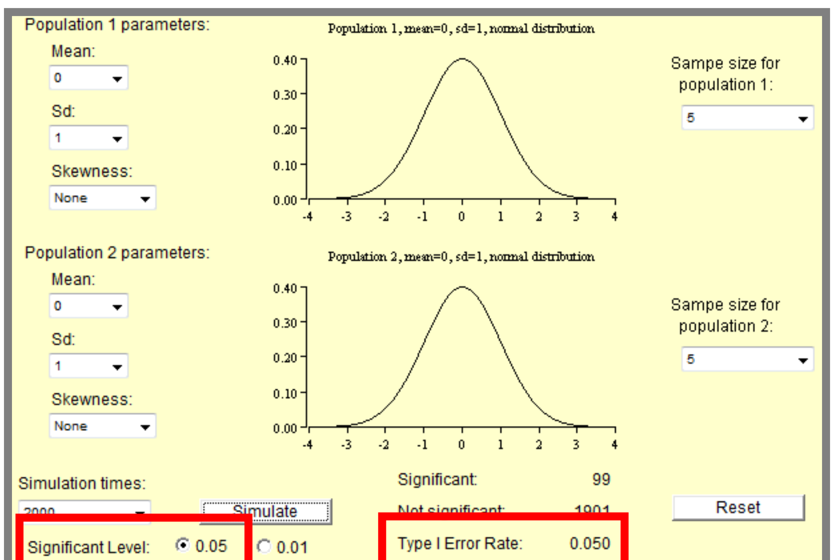


Unbiased Confidence Intervals Due To Homoscedasticity and Normality

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







When our CI is unbiased, we set the terms of the lottery.

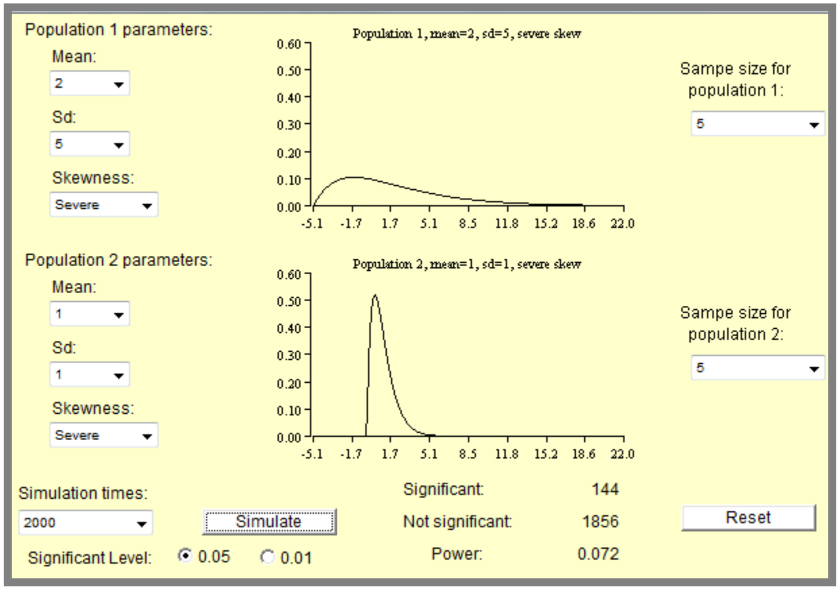
When our t-test is unbiased, our Alpha Level *exactly* equals our Type I Error Rate (when the null hypothesis is true). In other words, we set our Type I Error Rate by setting our Alpha Level:

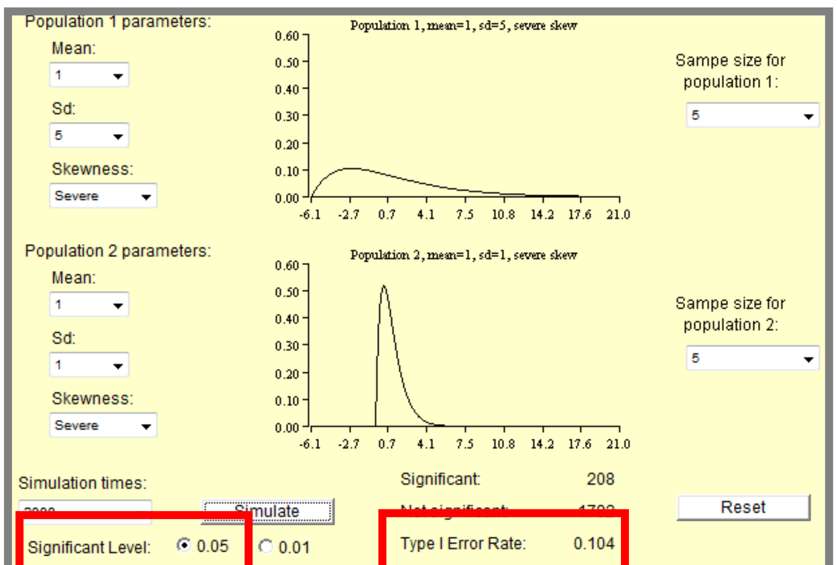
Alpha = .05

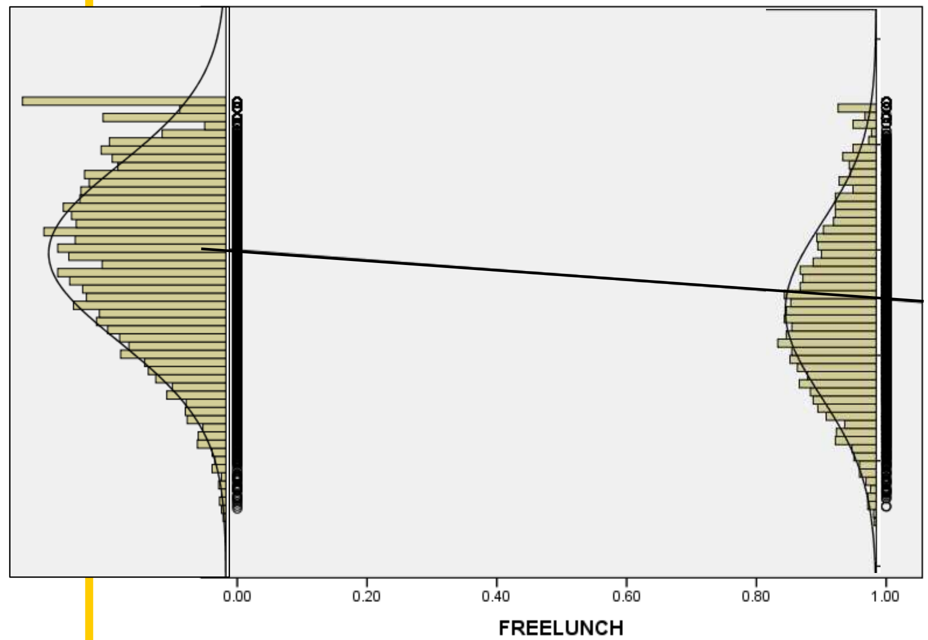
Probability of Type I Error (When the Null is True) = .05

Biased Confidence Intervals Due To Heteroscedasticity and Non-Normality

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







When our Ci is biased, the terms of the lottery are unclear.

When our t-test is biased, our Alpha Level *__is not quite__* our Type I Error Rate (when the null hypothesis is true). The difference is not necessarily huge; __t-tests are robust to assumption violations__.

Alpha = .05

Probability of Type I Error (When the Null is True) = .10

The case is terrible due to: Extreme Heteroscedasticity and Extreme Non-Normality and a Small Sample Size to boot. A "real" Alpha Level of .10, however, is not so horrific. Just know that our "95% confidence intervals" are really 90%!

Intermediate Data Analysis: Road Map (VERBAL)

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

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

"Statistically Significant": Abuses

- **The vastness of this country, the high mobility rate of many of its inhabitants and its statistically significant immigrant population all contribute to the need for an efficient postal service. - *The New York Times***
- **The numbers involved in this comparison were considered too small to be statistically significant. - *The New York Times***
- **It is difficult to test our results … against observations because no statistically significant global record of temperature back to 1600 has been constructed. - *Science***

Excerpted by Judith Singer for Unit 3 of *S-030: Applied Data Analysis*, Spring 2008, Harvard Graduate School of Education.

The p-value does not "give the probability that the null hypothesis is true." The null hypothesis states that the population slope is 0.000000000.... What are the chances of that? (Pop Quiz: What does the p-value give us?)

Never confuse statistical significance with practical significance.

If you do not find a statistically significant relationship, consider whether your analysis was underpowered. If you have a small sample size, you simply cannot detect small relationships, and small relationships can have huge consequences. (See http://onlinestatbook.com/stat_sim/index.html.)

Question 1: If all the regression assumptions are met perfectly, what does a p-value of 0.049 mean? How about a p-value of 0.00001? Or, a p-value of 0.89? Or, a p-value of 0.051?

Question 2: The relationship in the sample is one thing, and the relationship in the population is another thing (due to sampling error). If you say a relationship is statistically significant, are you talking about the relationship in the sample, or the relationship in the population?

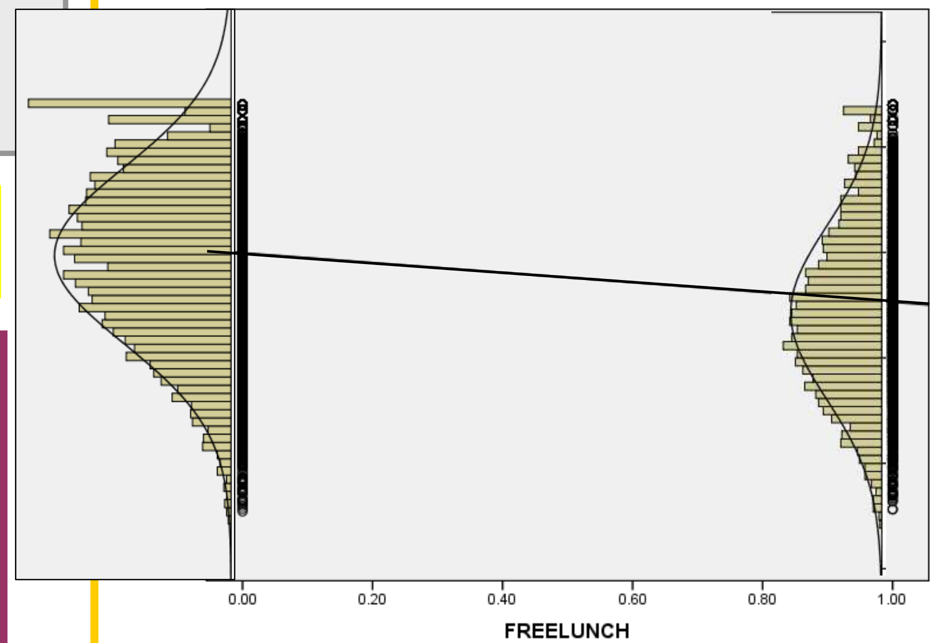
Question 3: A research institute (funded by Big Tobacco) finds no statistically significant relationship between smoking and cancer. All the regression assumptions are met perfectly. Their sample was a random sample of 77 Boston taxpayers, age 50 to 60. What's wrong with this finding, if anything?

Answering our Roadmap Question

Unit 8: What assumptions underlie our inference from the sample to the population?

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

We tentatively conclude that the assumptions of our linear model are met, but we would be more confident in our normality assumption were there no ceiling effect for *READING*, and we would be more confident in our independence assumption if we accounted for nesting of students within schools.



Homoscedasticity: The variances appear roughly equal.

Independence: There may be clustering of students within schools.

Normality: The conditional distributions are normal but for the ceiling effect.

Linearity: Linearity will never be a problem when our predictor is dichotomous.

Outliers: There are no egregious outliers.

Unit 8 Appendix: Key Concepts

•Standard errors are based on our guesstimate of the population standard deviation from the sample standard deviation and our sample size. The bigger the sample size, the smaller the standard error.

•A standard error is guesstimated from the observed variance, but, if the observed variance varies by level of the predictor, i.e., the relationship is heteroscedastic, then which variance should you use? Heteroscedasticicity biases (upwards) our estimate of the standard error, but it does not bias our estimate of the slope.

•The Central Limit Theorem tells us that the sampling distribution of the slope estimate is approximately normal, no matter what. Great. We also need to know the standard deviation (i.e., standard error) of the sampling distribution to test null hypotheses and construct confidence intervals. We don't ***know*** the standard error, but we can estimate it. Whenever we estimate from a sample, we have to worry about sampling error, so we have to consider the sampling distribution of the standard error estimate (an additional sampling distribution). The Central Limit Theorem tells us that the sampling distribution of the standard error is NOT normal. Whereas sampling distributions for slopes are always approximately the same shape (normal), sampling distributions for the standard errors are skewed and skewed to different extents depending on the distributions of Y conditional on X. If we want to understand exactly the additional uncertainty from estimating standard errors, we need to know the distribution of Y conditional on X. If, for example, we know that the distribution of Y condition on X is normal, then we can account for the extra uncertainty perfectly in our calculations. Hence, the normality assumption. (If we have a great estimate of our standard error, then the normality assumption is needless.)

•**Whereas heterscedasticity, non-independence, and non-normality bias the standard errors but not the slope coefficient, non-linearity and outliers bias the slope coefficient and, consequently, the standard errors.**

•We use the Central Limit Theorem to guesstimate our standard errors, and the Central Limit Theorem relies on randomness. When observations are clustered (i.e., the independence assumption is violated), there is order in what should be essentially random. Our effective sample size may be smaller than the number of observations if the observations hang together in clusters. Of HI-N-LO, independence is the one assumption we ***cannot*** check visually!

•Don't forget the LO in HI-N-LO. **L**inearity and (no) **O**utliers are by far the most important regression assumptions!

•The p-value does not "give the probability that the null hypothesis is true."

•Never confuse statistical significance with practical significance.

•If you do not find a statistically significant relationship, consider whether your analysis was underpowered. If you have a small sample size, you simply cannot detect small relationships, and small relationships can have huge consequences.

Unit 8 Appendix: Key Interpretations

Homoscedasticity:
Although the variation in reading scores appears to be less at higher levels of depression, we believe that this is merely an appearance due to sparse sampling of extremely depressed subjects. If we were to sample more extremely depressed subjects, we suspect that we would find their variation in reading equal to the variation of non-depressed subjects.

Independence:
Although we cannot test for independence visually, we suspect that students are clustered within schools, and this may be biasing our standard error. We may need to use multi-level regression models to account for the non-independence.

Normality:
At each value of depression, the distribution of reading scores appears roughly normal. For example, it appears that, for students with depression scores of 0.0, their depression scores are roughly normal, Likewise for students with depression scores of -1.0, 2.0 and 5.0.\

We tentatively conclude that the assumptions of our linear model are met, but we would be more confident in our normality assumption were there no ceiling effect for *READING*, and we would be more confident in our independence assumption if we accounted for nesting of students within schools.

Unit 8 Appendix: Key Terminology

A bivariate relationship is homoscedastic when the distributions of Y conditional on X have equal variances (i.e., equal spreads).

A bivariate relationship meets the normality assumption when Y is normally distributed conditional on X.

A bivariate relationship meets the independence assumption when the errors are uncorrelated.

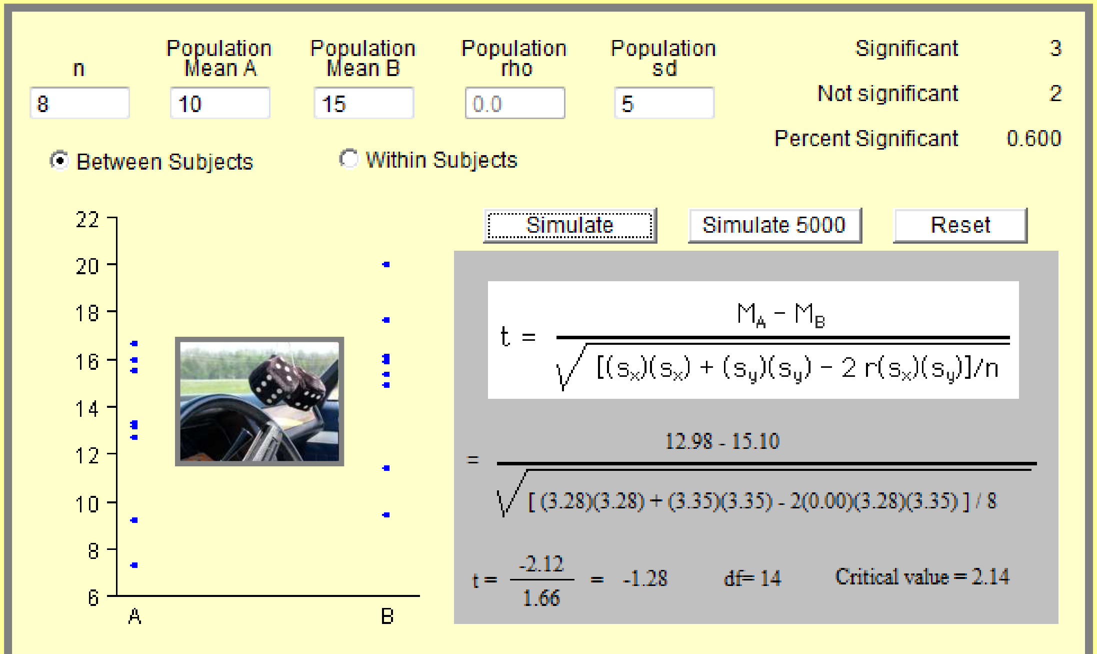
Unit 8 Math Appendix: Two-Sample T-Tests

Recall from the Central Limit Theorem that the standard deviation (i.e., standard error) of the sampling distribution of a mean is:

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}$$

All standard error have roughly the same form:

$$\text{StandardError} = \frac{\text{Guesstimation of the Population Variation}}{\text{SampleSize}}$$



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

(The notation is funky here. That's because there are all sorts of statistical notations, and we have become used to one, but this is another.)

A t-statistic is the slope divided by the standard error. When we regress a continuous outcome on a dichotomous predictor, the slope is simply the difference in the averages of the two groups defined by the dichotomous predictor. This simplicity makes the math very doable by hand, especially when the groups are equally sized.

In the following simulation, start with the "Between Subjects" radio button. This makes the simulated data independent. Thus, the population correlation between the two groups (rho) is 0.0, and the $-2r(s_x)(s_y)$ drops out of the standard error, leaving:

$$SE = \sqrt{\frac{(s_A)(s_A) + (s_B)(s_B)}{n}}$$

Where:

S_A is the standard deviation of Group A. (Which is strangely labeled S_x in the applet.)

n is the sample size of each group. (Which should be degrees of freedom n-1, but the applet ignores the nuance.)

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

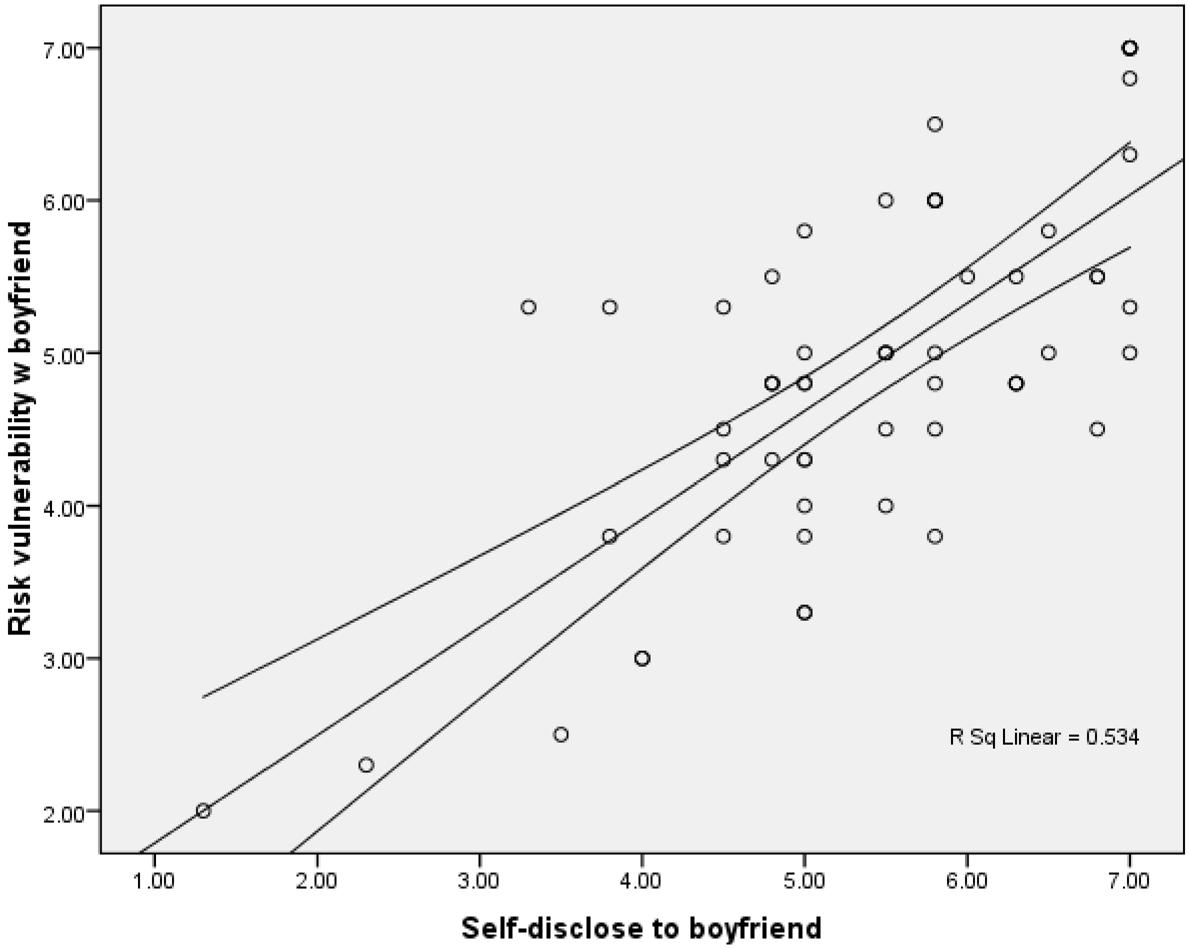
b. Dependent Variable: Risk vulnerability w boyfriend

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	1.081	.482		2.244	.029	.117	2.045
	Self-disclose to boyfriend	.708	.087	.731	8.154	.000	.534	.882

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

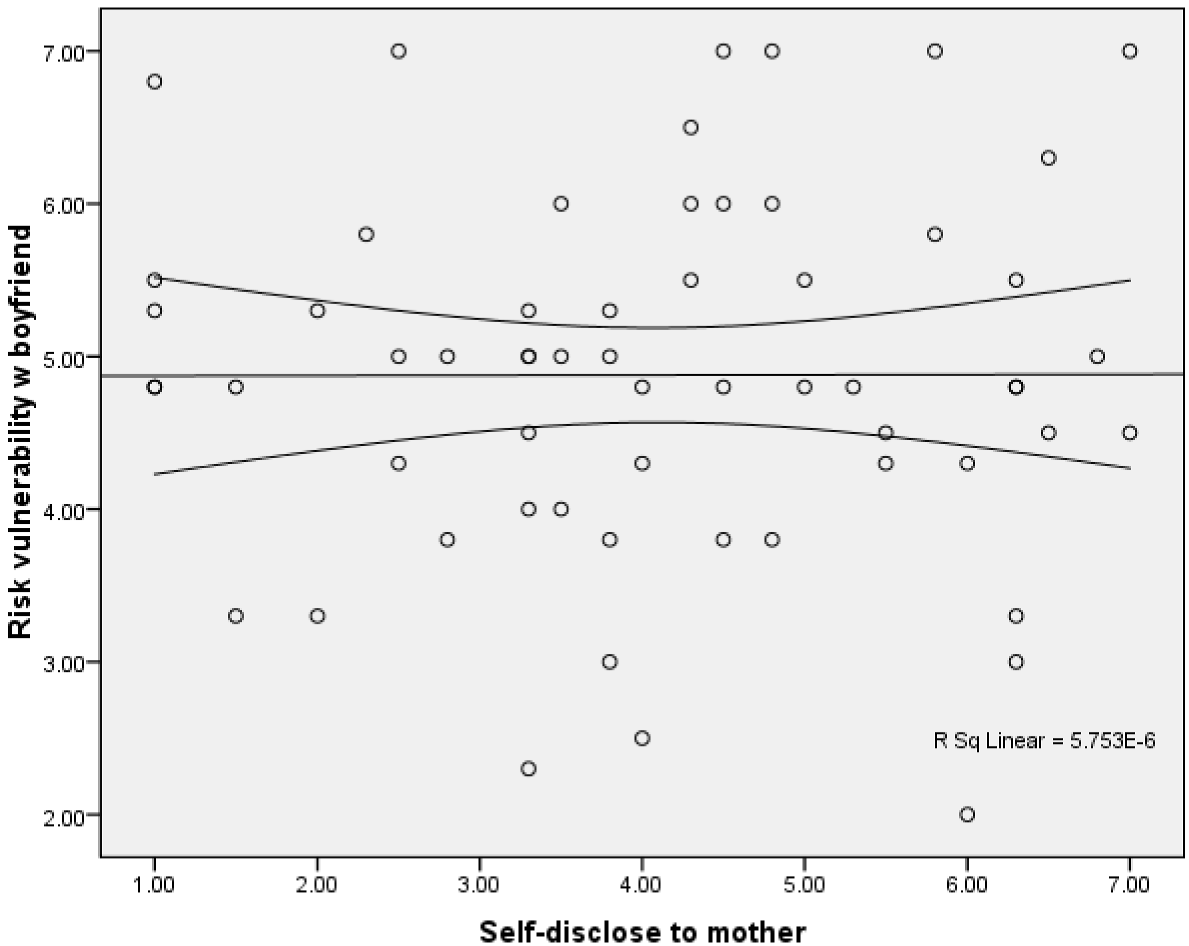
b. Dependent Variable: Risk vulnerability w boyfriend

Coefficients[a]

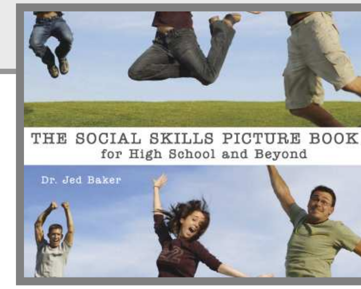
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	4.872	.404		12.050	.000	4.062	5.681
	Self-disclose to mother	.002	.091	.002	.018	.985	-.181	.184

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:**

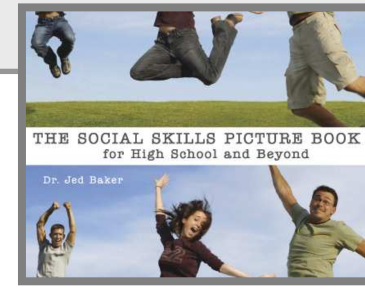
Variables about the student—

(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

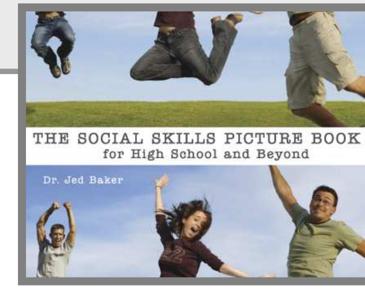
b. Dependent Variable: Base Year Composite Test

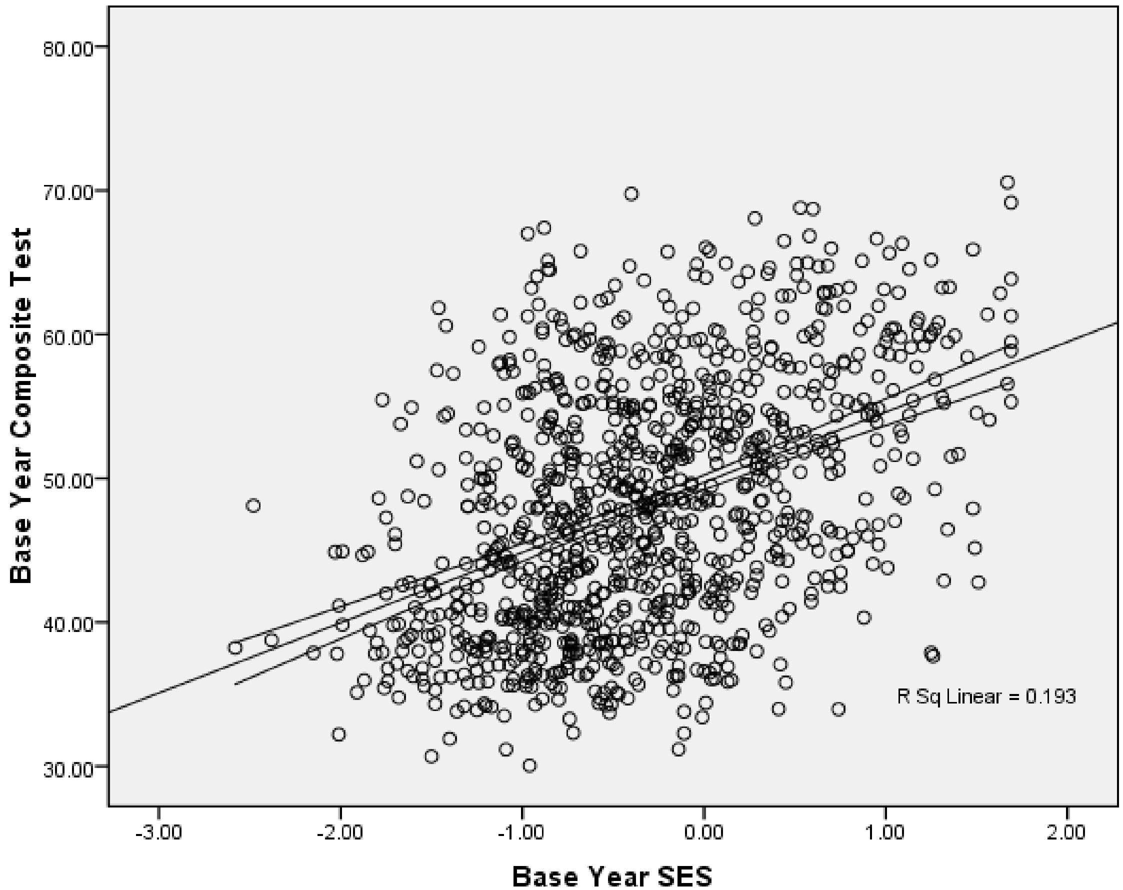
Coefficients[a]

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

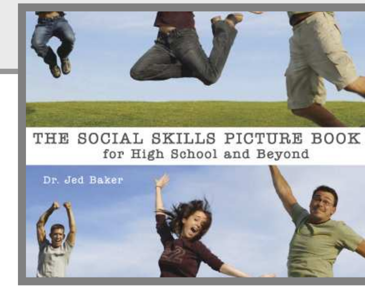
a. Dependent Variable: Base Year Composite Test

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

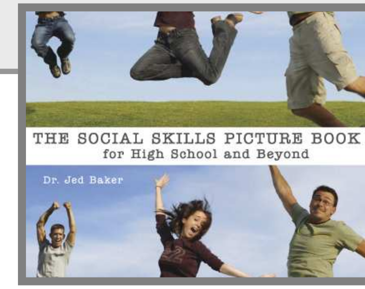
b. Dependent Variable: Base Year Composite Test

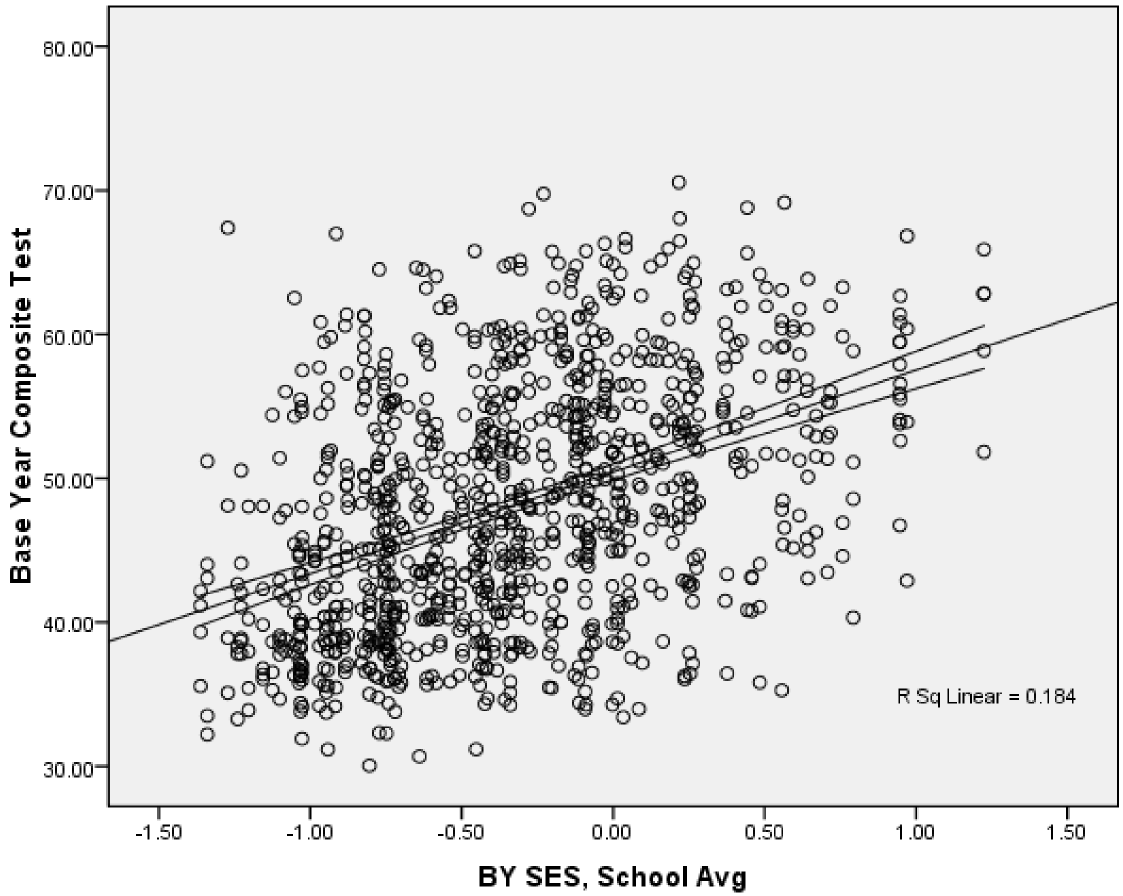
Coefficients[a]

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

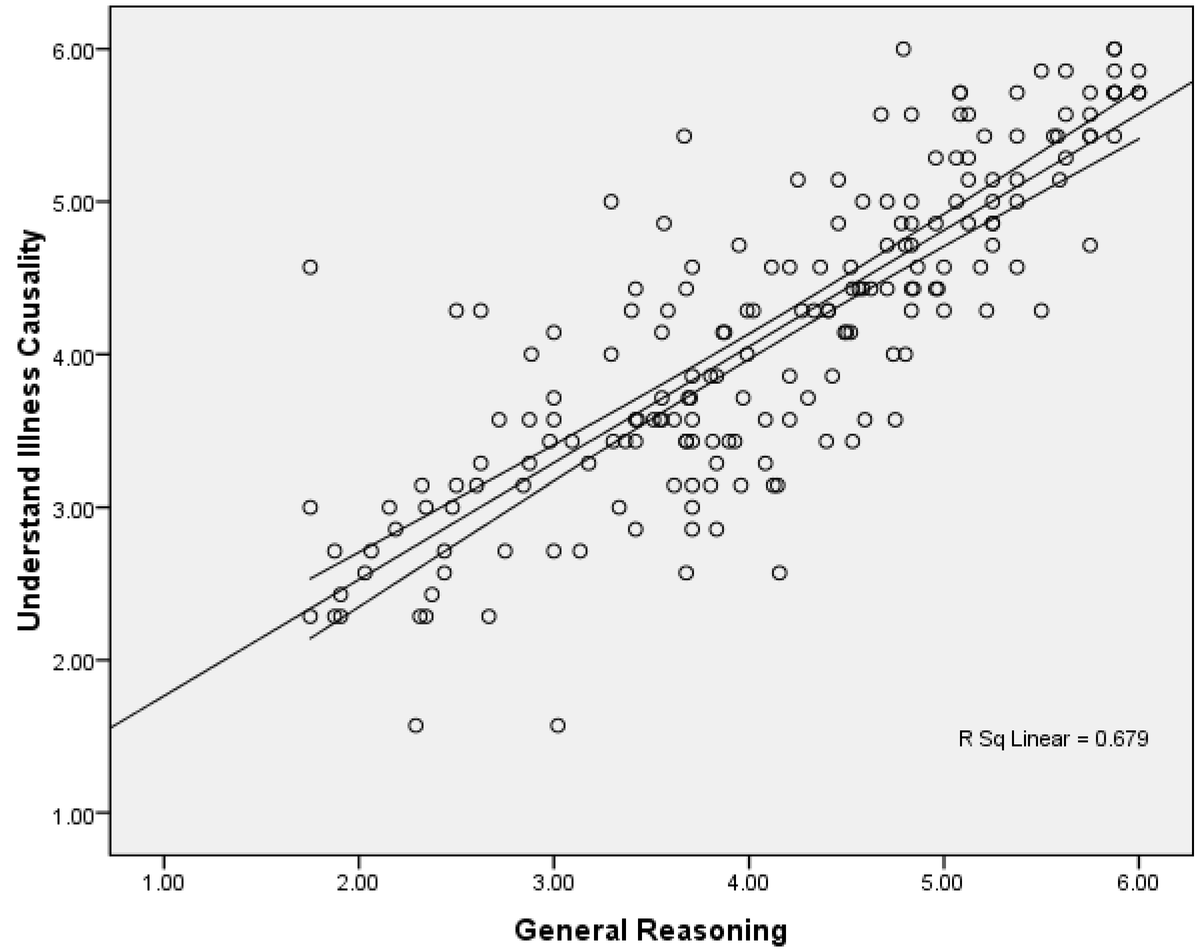
b. Dependent Variable: Understand Illness Causality

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

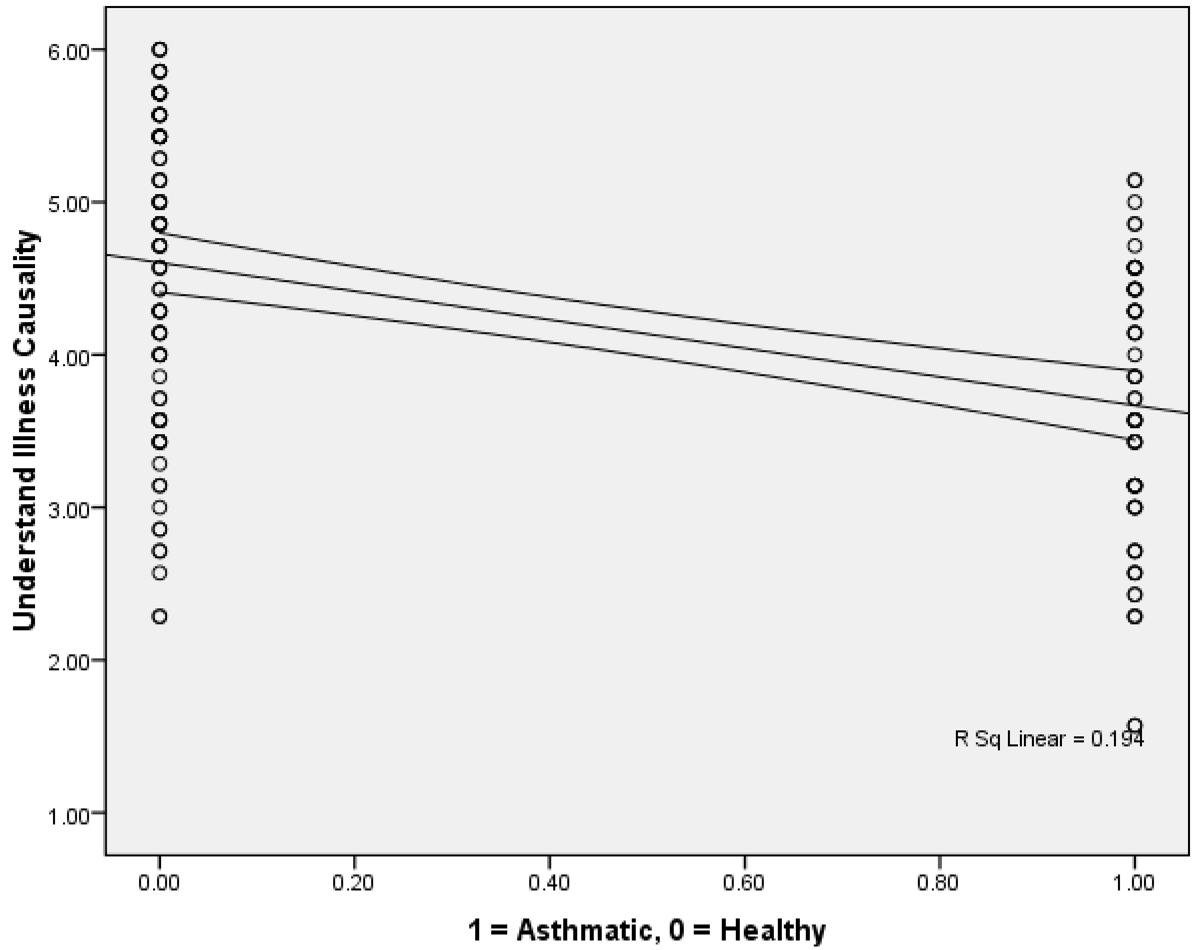
b. Dependent Variable: Understand Illness Causality

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	.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

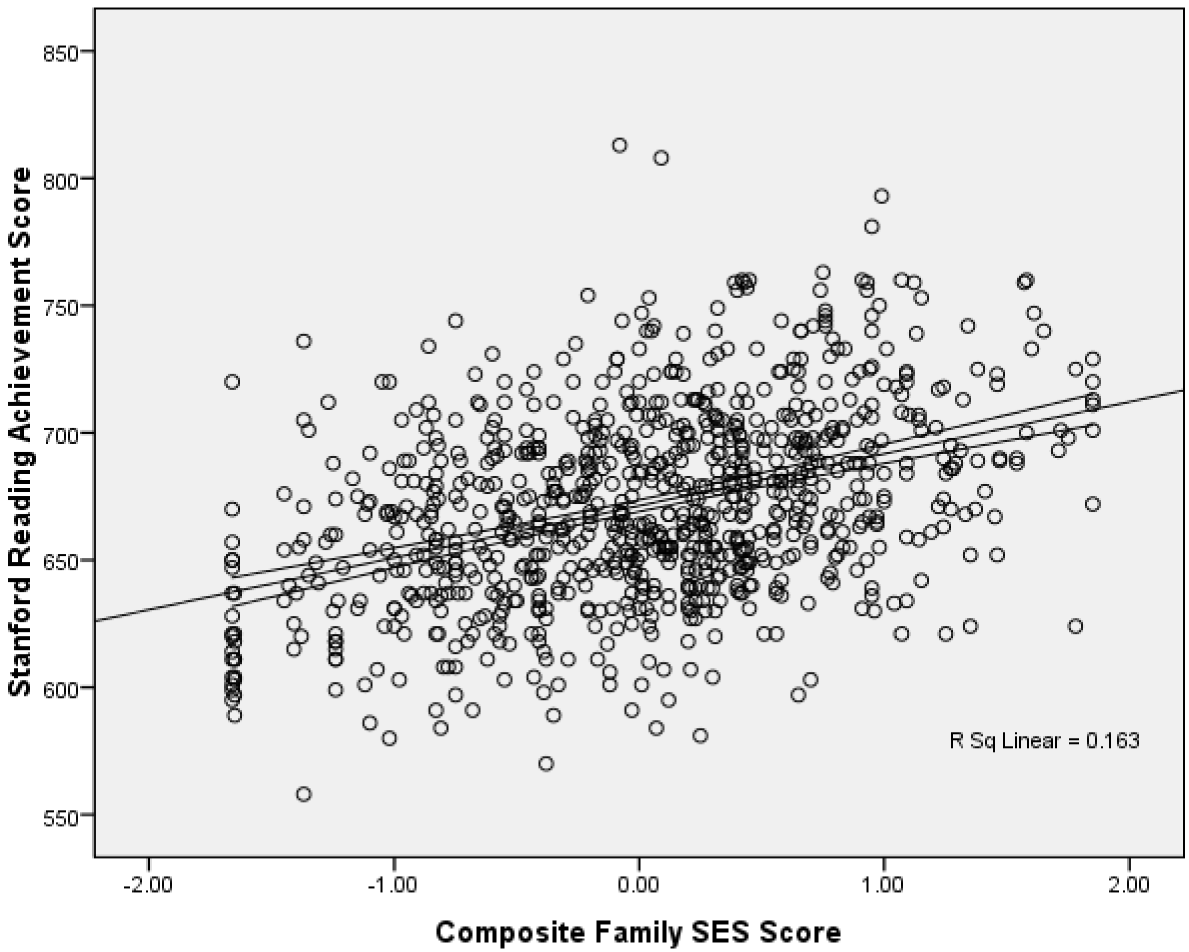
b. Dependent Variable: Stanford Reading Achievement Score

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	671.350	1.175		571.418	.000	669.044	673.656
	Composite Family SES Score	20.418	1.562	.404	13.071	.000	17.352	23.483

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	.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

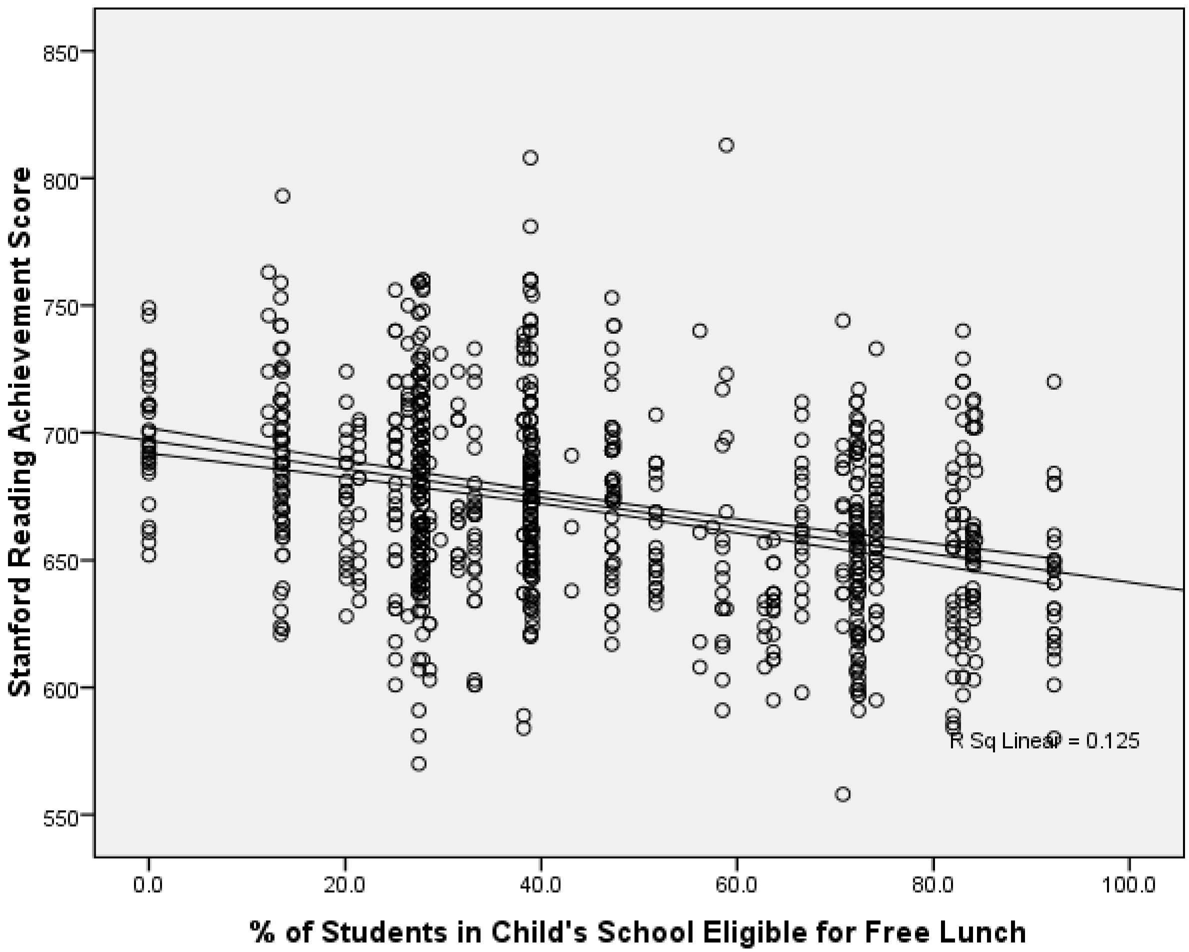
b. Dependent Variable: Stanford Reading Achievement Score

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	696.847	2.540		274.325	.000	691.861	701.832
	% of Students in Child's School Eligible for Free Lunch	-.555	.050	-.353	-11.182	.000	-.653	-.458

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

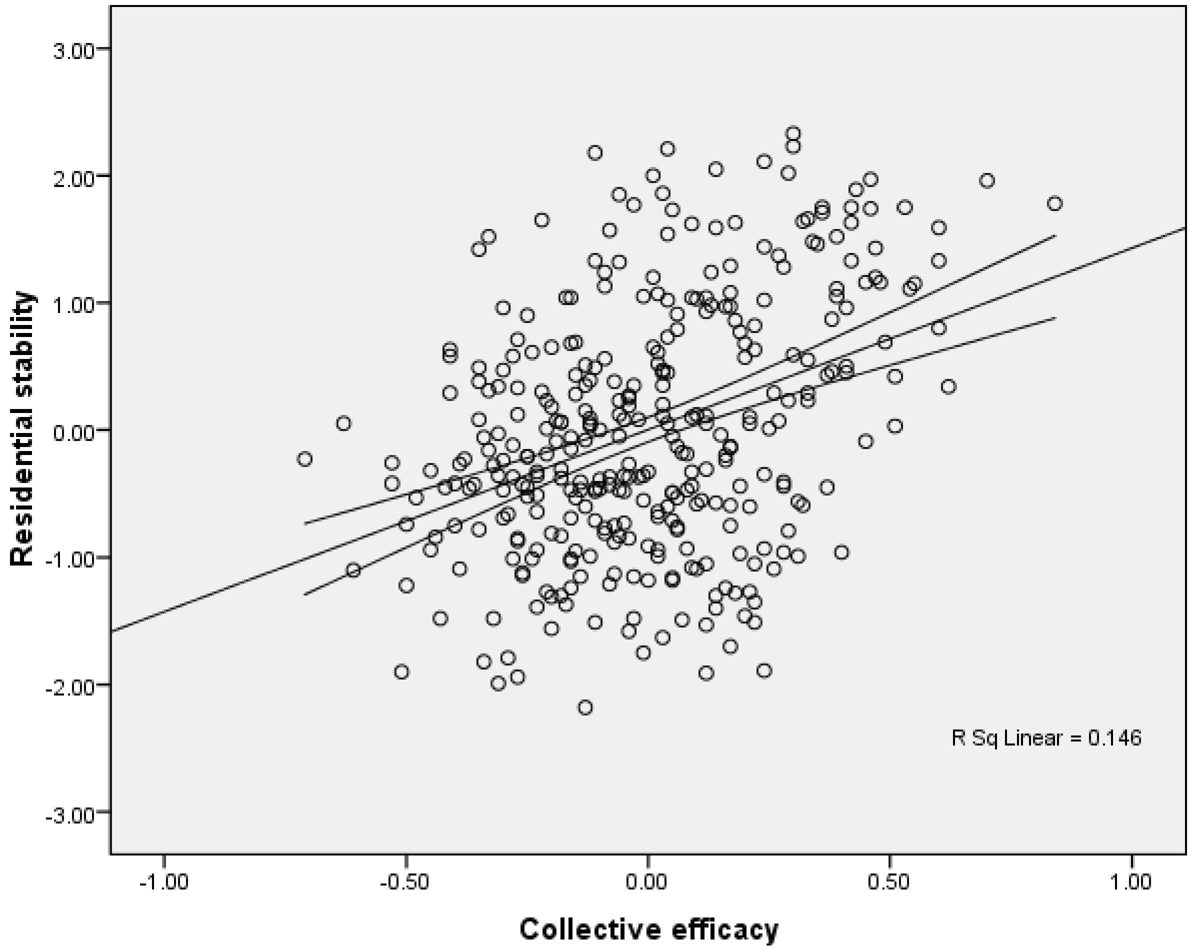
b. Dependent Variable: Residential stability

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

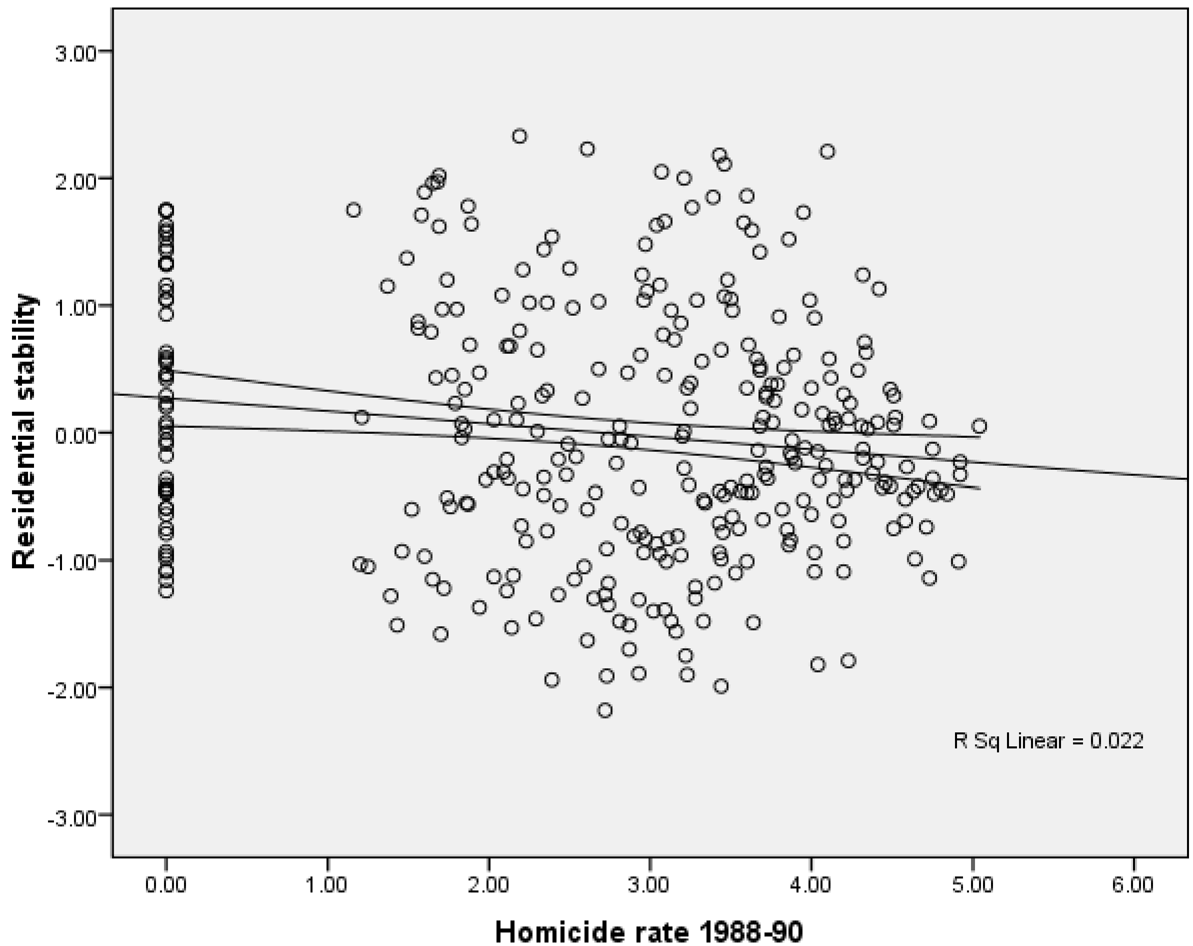
b. Dependent Variable: Residential stability

Coefficients[a]

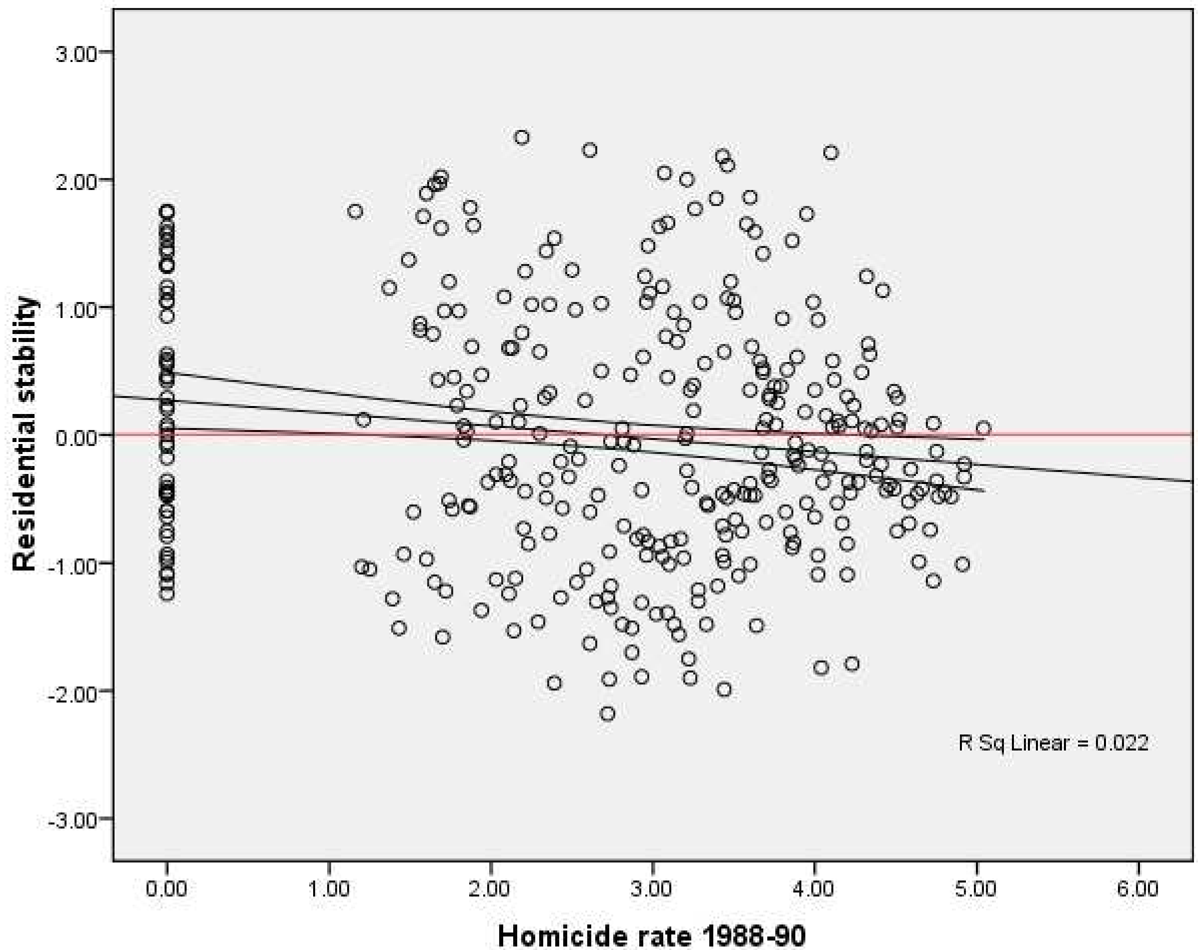
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.270	.111		2.432	.016	.052	.489
	Homicide rate 1988-90	-.100	.037	-.147	-2.735	.007	-.173	-.028

a. Dependent Variable: Residential stability

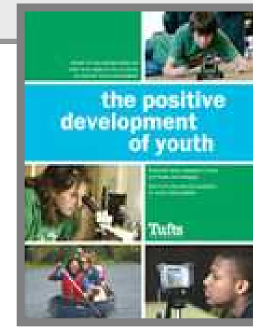
Human Development in Chicago Neighborhoods (Neighborhoods.sav)



Human Development in Chicago Neighborhoods (Neighborhoods.sav)



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



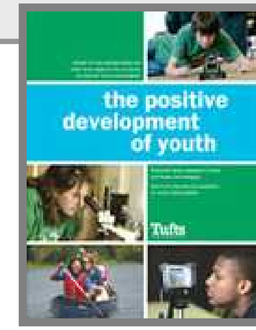
- **4-H Study of Positive Youth Development**
- **Source: Subset of data from IARYD, Tufts University**

- **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

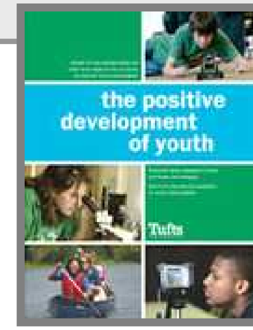
b. Dependent Variable: Self-Worth

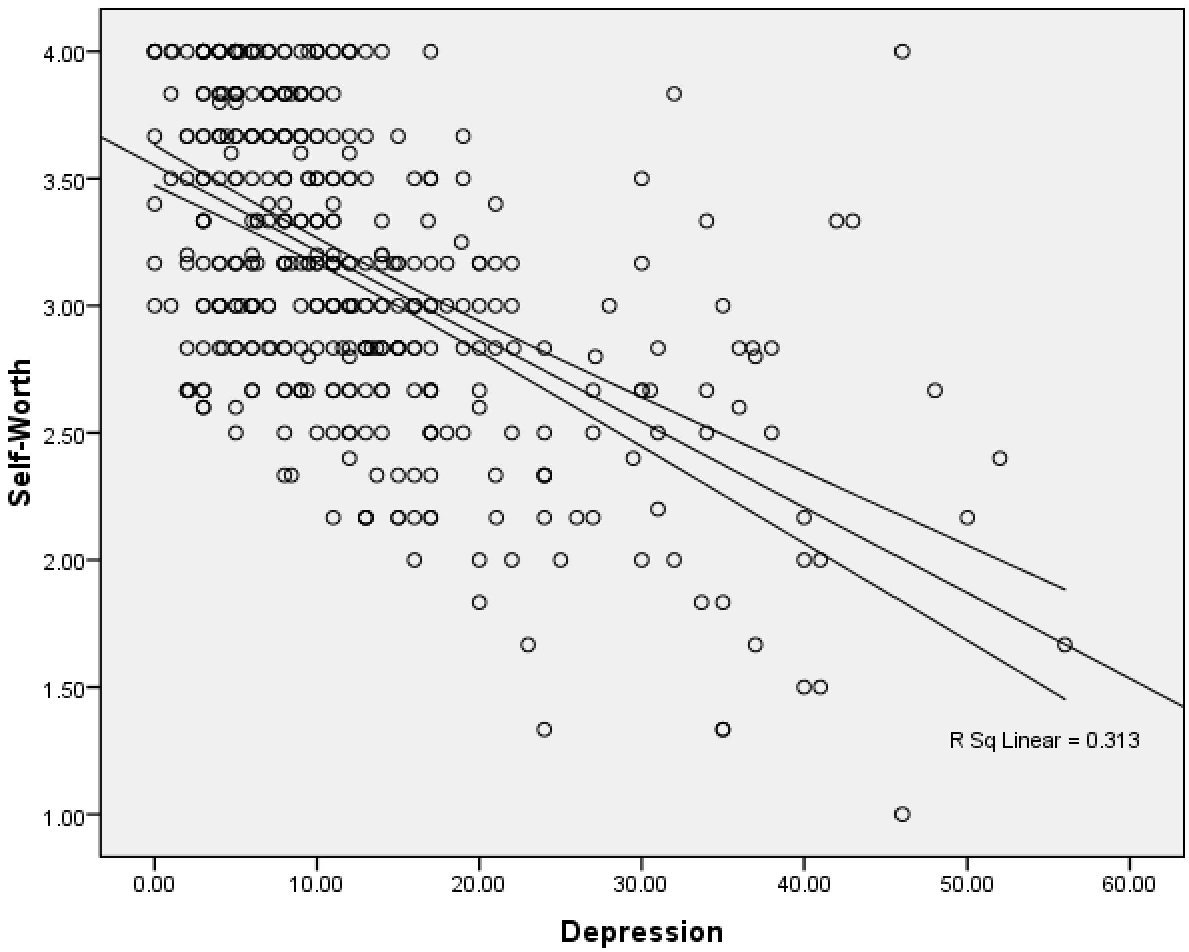
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	3.552	.040		88.146	.000	3.473	3.631
	Depression	-.034	.002	-.559	-13.606	.000	-.038	-.029

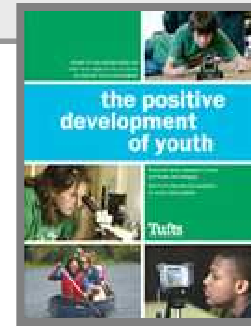
a. Dependent Variable: Self-Worth

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.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	3.307	.030		108.824	.000	3.247	3.367
	Depressed = 1, Not Depressed = 0	-.686	.058	-.504	-11.758	.000	-.801	-.571

a. Dependent Variable: Self-Worth

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