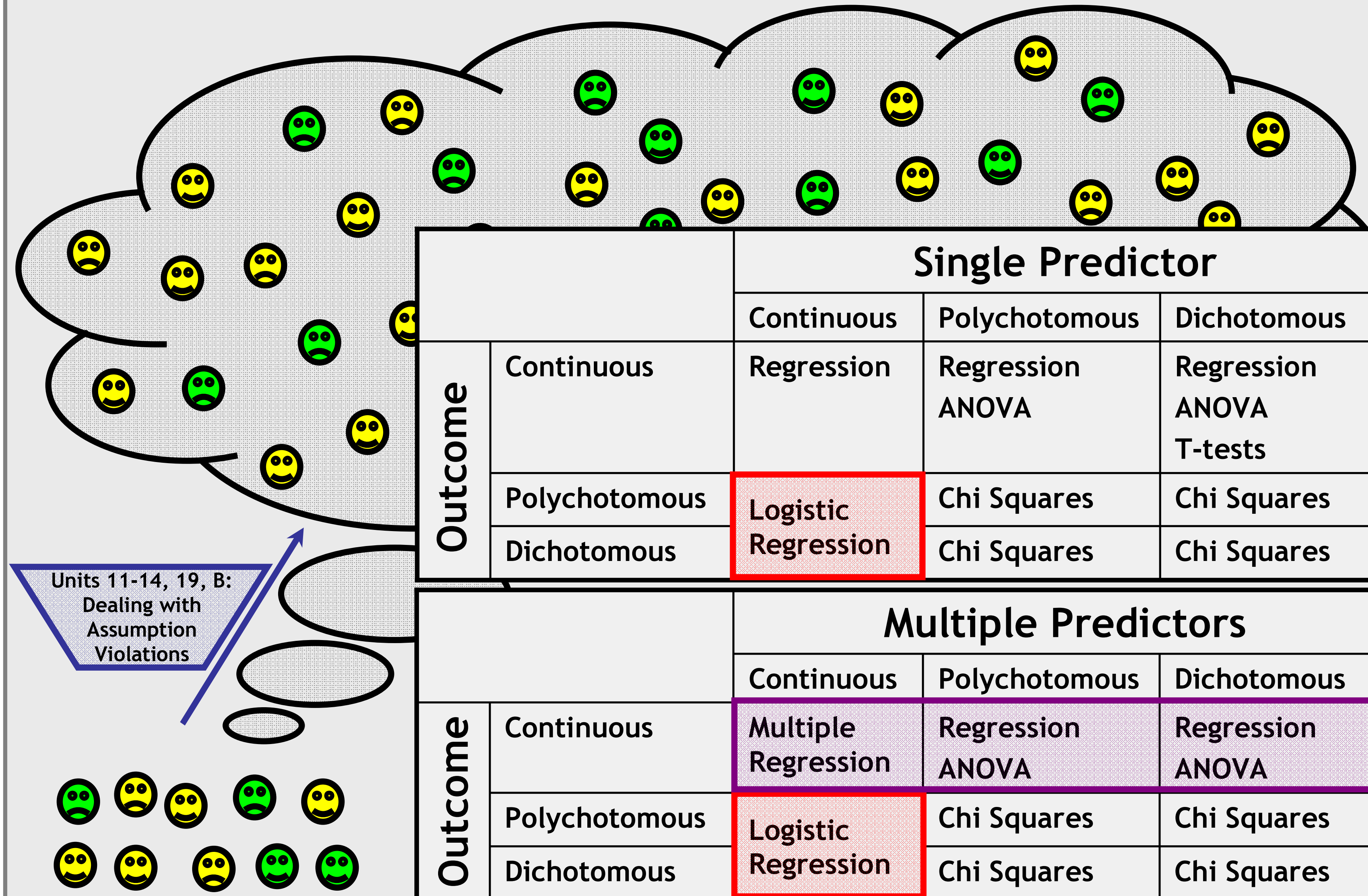
Unit 14: Road Map (VERBAL)

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

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

Unit 14: Road Map (Schematic)



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

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

Unit 14: Roadmap (SPSS Output)

Coefficients[a]

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)	48.338	.110		438.242	.000	48.122	48.554
	ASIAN	1.034	.383	.030	2.697	.007	.283	1.786
	BLACK	-4.889	.339	-.161	-14.423	.000	-5.554	-4.225
	LATINO	-4.418	.306	-.161	-14.447	.000	-5.017	-3.818
2	(Constant)	43.878	.280		156.558	.000	43.328	44.427
	ASIAN	.727	.377	.021	1.929	.054	-.012	1.465
	BLACK	-4.796	.333	-.158	-14.412	.000	-5.448	-4.144
	LATINO	-4.123	.301	-.151	-13.715	.000	-4.712	-3.534
	L2HOMEWORKP1	1.766	.102	.188	17.254	.000	1.565	1.967
3	(Constant)	45.381	.284		159.528	.000	44.823	45.938
	ASIAN	.461	.441	.013	1.045	.296	-.404	1.325
	BLACK	-3.622	.331	-.119	-10.956	.000	-4.270	-2.974
	LATINO	-3.311	.366	-.121	-9.035	.000	-4.029	-2.592
	L2HOMEWORKP1	1.603	.100	.170	15.974	.000	1.406	1.799
	ESL	.218	.363	.009	.600	.548	-.494	.930
	FREELUNCH	-3.867	.199	-.213	-19.452	.000	-4.256	-3.477
4	(Constant)	45.358	.288		157.560	.000	44.794	45.923
	ASIAN	-.377	.668	-.011	-.564	.573	-1.687	.933
	BLACK	-3.447	.498	-.113	-6.922	.000	-4.423	-2.471
	LATINO	-2.779	.517	-.102	-5.371	.000	-3.793	-1.765
	L2HOMEWORKP1	1.591	.100	.169	15.866	.000	1.394	1.788
	ESL	-.876	.638	-.035	-1.373	.170	-2.126	.374
	FREELUNCH	-3.574	.235	-.197	-15.208	.000	-4.035	-3.113
	ESLxASIAN	3.245	.999	.080	3.249	.001	1.287	5.202
	ESLxBLACK	5.872	1.885	.036	3.115	.002	2.177	9.568
	ESLxLATINO	.446	.858	.013	.520	.603	-1.235	2.127
	FREELUNCHxASIAN	-2.769	.853	-.041	-3.245	.001	-4.442	-1.096
	FREELUNCHxBLACK	-.751	.666	-.019	-1.127	.260	-2.058	.555
	FREELUNCHxLATINO	-.437	.604	-.012	-.724	.469	-1.622	.747

a. Dependent Variable: READING

Unit 18

Unit 9

Unit 16

Unit 17

Unit 8

Unit 11

Unit 12

Unit 13

Unit 14

Unit 15

Unit 14: Robust Standard Error to Meet the Homoskedasticity Assumption

Unit 14 Post Hole:

Judge whether robust standard errors are necessary for estimation.

Unit 14 Technical Memo and School Board Memo:

Use Stata and robust standard errors to fit your regression model (from Memos 11 and 12).

Unit 14: Technical Memo and School Board Memo

Work Products (Part I of II):

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

   A. Introduction

      i. State a theory (or perhaps hunch) for the relationship—think causally, be creative. (1 Sentence)

      ii. State a research question for each theory (or hunch)—think correlationally, be formal. Now that you know the statistical machinery that justifies an inference from a sample to a population, begin each research question, “In the population,…” (1 Sentence)

      iii. List your variables, and label them “outcome” and “predictor,” respectively.

      iv. Include your theoretical model.

   B. Univariate Statistics. Describe your variables, using descriptive statistics. What do they represent or measure?

      i. Describe the data set. (1 Sentence)

      ii. Describe your variables. (1 Paragraph Each)

         a. Define the variable (parenthetically noting the mean and s.d. as descriptive statistics).

         b. Interpret the mean and standard deviation in such a way that your audience begins to form a picture of the way the world is. Never lose sight of the substantive meaning of the numbers.

         c. Polish off the interpretation by discussing whether the mean and standard deviation can be misleading, referencing the median, outliers and/or skew as appropriate.

         d. Note validity threats due to measurement error.

   C. Correlations. Provide an overview of the relationships between your variables using descriptive statistics. Focus first on the relationship between your outcome and question predictor, second-tied on the relationships between your outcome and control predictors, second-tied on the relationships between your question predictor and control predictors, and fourth on the relationship(s) between your control variables.

      a. Include your own simple/partial correlation matrix with a well-written caption.

      b. Interpret your simple correlation matrix. Note what the simple correlation matrix foreshadows for your partial correlation matrix; “cheat” here by peeking at your partial correlation and thinking backwards. Sometimes, your simple correlation matrix reveals possibilities in your partial correlation matrix. Other times, your simple correlation matrix provides foregone conclusions. You can stare at a correlation matrix all day, so limit yourself to two insights.

      c. Interpret your partial correlation matrix controlling for one variable. Note what the partial correlation matrix foreshadows for a partial correlation matrix that controls for two variables. Limit yourself to two insights.

Unit 14: Technical Memo and School Board Memo

Work Products (Part II of II):

I. Technical Memo (continued)

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

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

III. Memo Metacognitive

Unit 14: Research Question

Theory: Because both the AHS curriculum and the MCAS math test are aligned with the Massachusetts Curriculum Frameworks, they are highly correlated. Some students, however, perform differently from what the correlation would lead one to expect. It would be helpful to identify those students in order to address their learning needs.

Research Question: Controlling for GPA, which students perform best on the Math MCAS and which students perform worst?

Data Set: MCAS and GPA (MCAS and GPA.sav): Math MCAS scaled scores and GPAs for the 223 sophomores at an anonymous Massachusetts high school (AHS), date withheld.

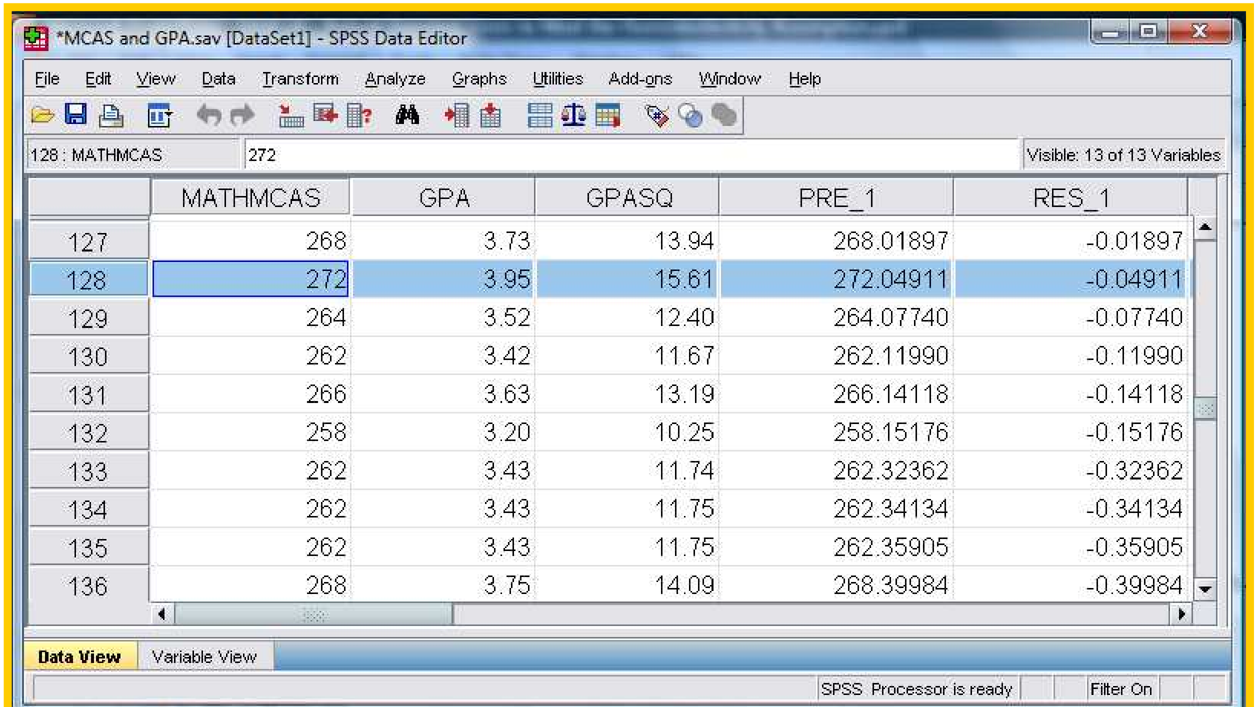
Variables:

Outcome—Math MCAS Scaled Score (*MATHMCAS*)

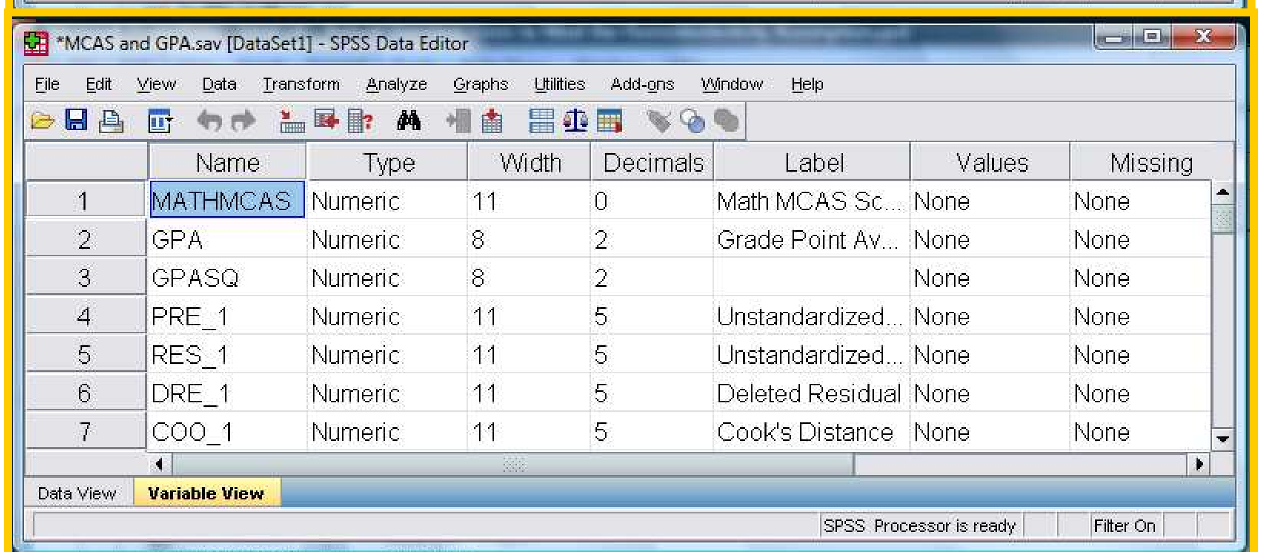
Predictor—Sophomore Grade Point Average (*GPA*)

Model: $MATHMCAS = \beta_0 + \beta_1 GPA + \varepsilon$

The MCAS and GPA Data Set

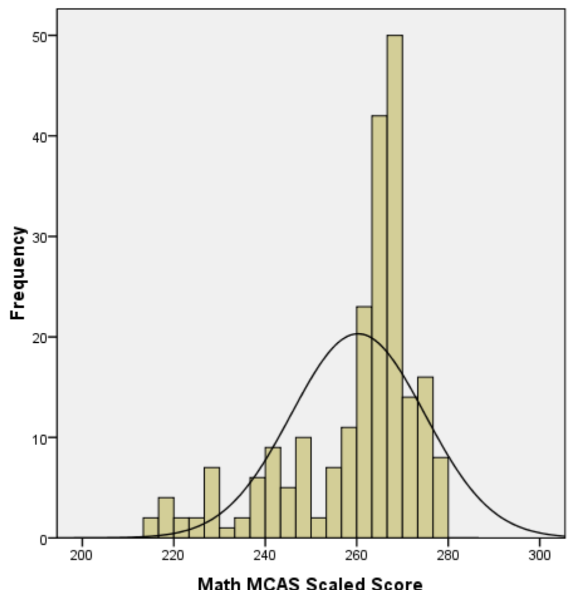


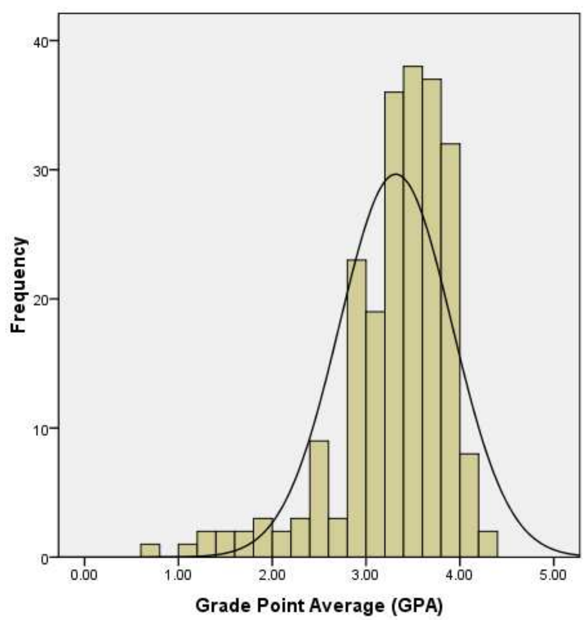
	MATHMCAS	GPA	GPASQ	PRE_1	RES_1
127	268	3.73	13.94	268.01897	-0.01897
128	272	3.95	15.61	272.04911	-0.04911
129	264	3.52	12.40	264.07740	-0.07740
130	262	3.42	11.67	262.11990	-0.11990
131	266	3.63	13.19	266.14118	-0.14118
132	258	3.20	10.25	258.15176	-0.15176
133	262	3.43	11.74	262.32362	-0.32362
134	262	3.43	11.75	262.34134	-0.34134
135	262	3.43	11.75	262.35905	-0.35905
136	268	3.75	14.09	268.39984	-0.39984

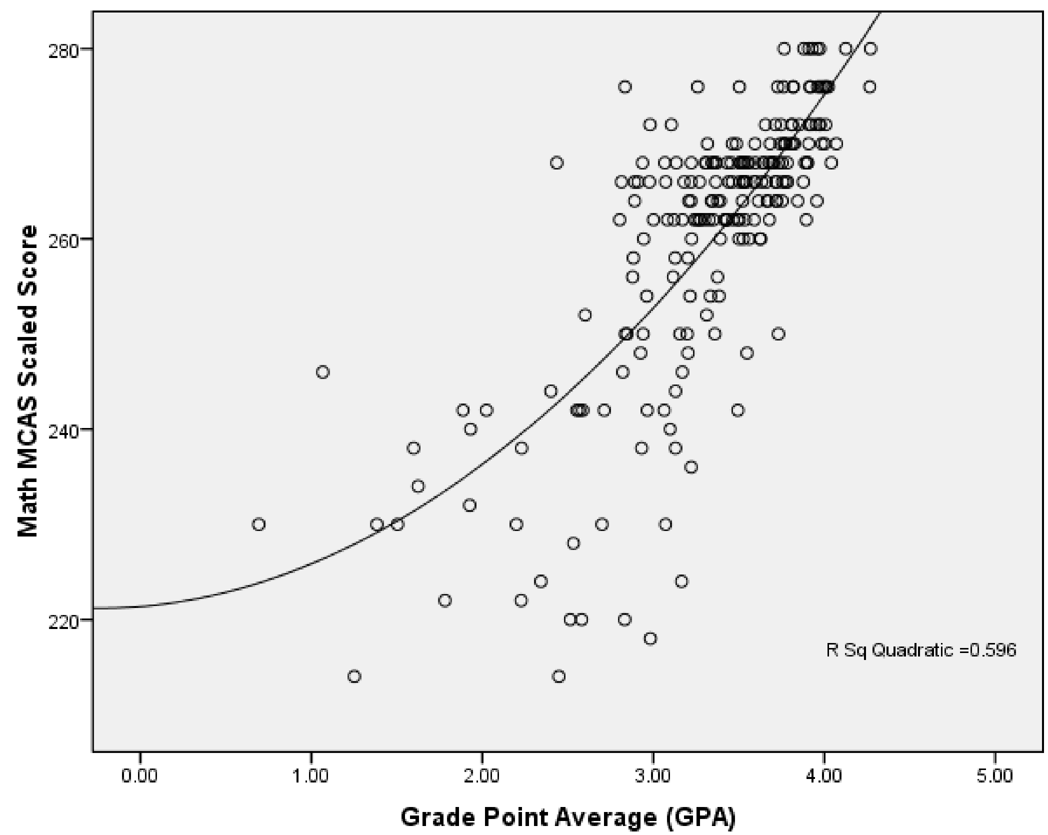


	Name	Type	Width	Decimals	Label	Values	Missing
1	MATHMCAS	Numeric	11	0	Math MCAS Sc...	None	None
2	GPA	Numeric	8	2	Grade Point Av...	None	None
3	GPASQ	Numeric	8	2		None	None
4	PRE_1	Numeric	11	5	Unstandardized...	None	None
5	RES_1	Numeric	11	5	Unstandardized...	None	None
6	DRE_1	Numeric	11	5	Deleted Residual	None	None
7	COO_1	Numeric	11	5	Cook's Distance	None	None

Exploratory Graphs







Using Tukey's Rule of the Bulge, should we try going up in X, down in X, up in Y or down in Y? Pick two. Of the two, which is the better as suggested by the histograms?

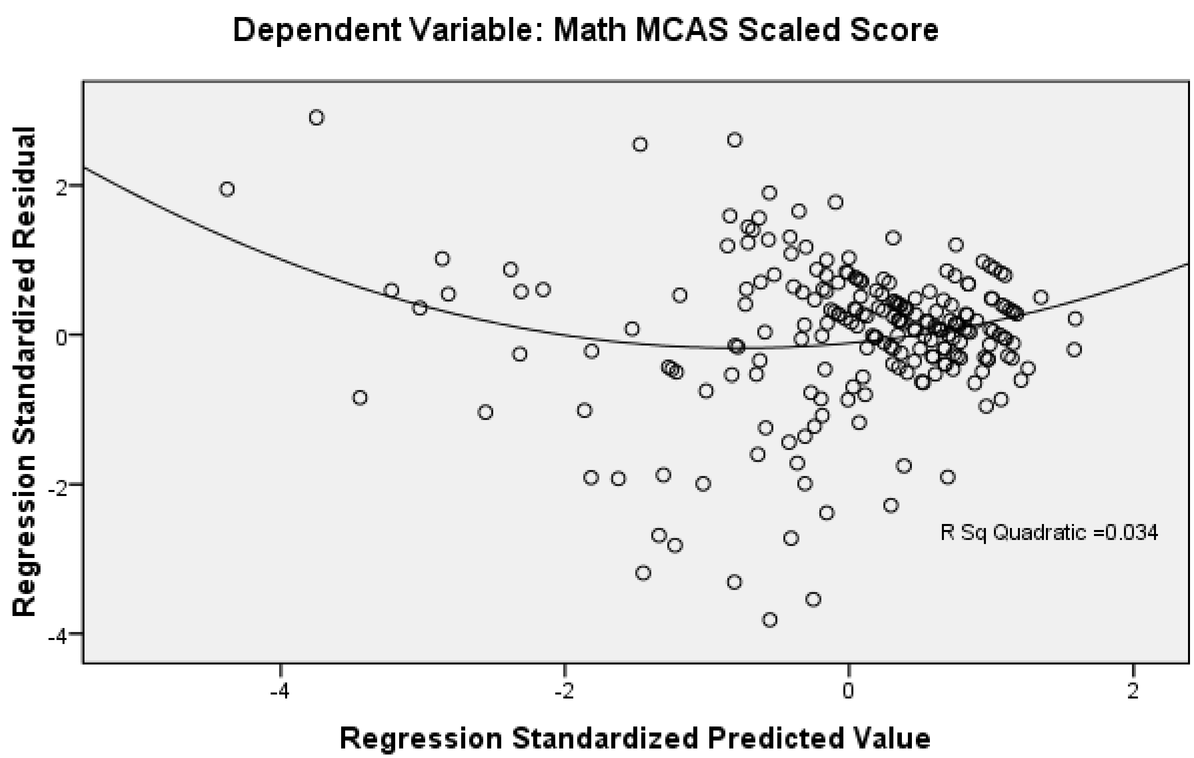
Fitting the Wrong Model

Coefficients[a]

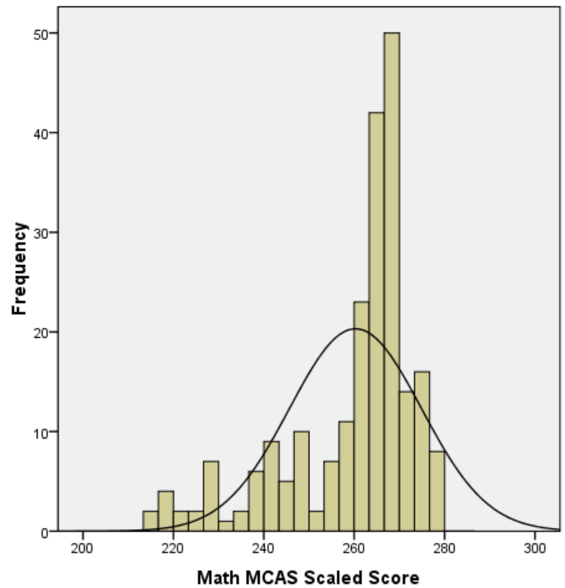
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	198.709	3.565		55.740	.000	191.684	205.735
	Grade Point Average (GPA)	18.565	1.058	.763	17.550	.000	16.481	20.650

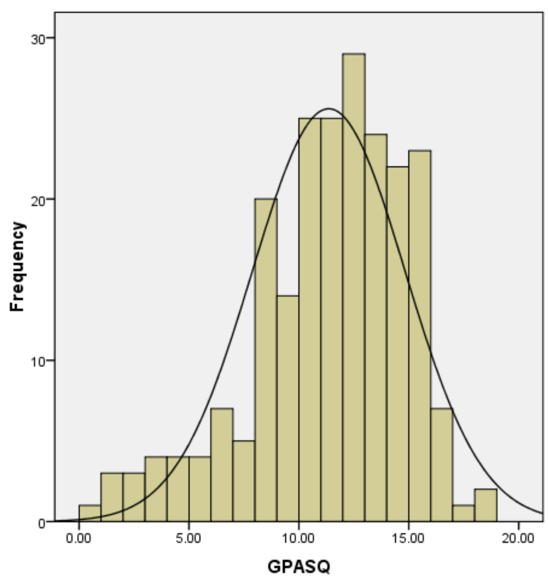
a. Dependent Variable: Math MCAS Scaled Score

I had SPSS drop a quadratic trend line instead of a mean line in order to guide my eye. There is a slight horseshoe. Note that in an RVF plot, there should be no pattern. The trend line and mean line should be identical.

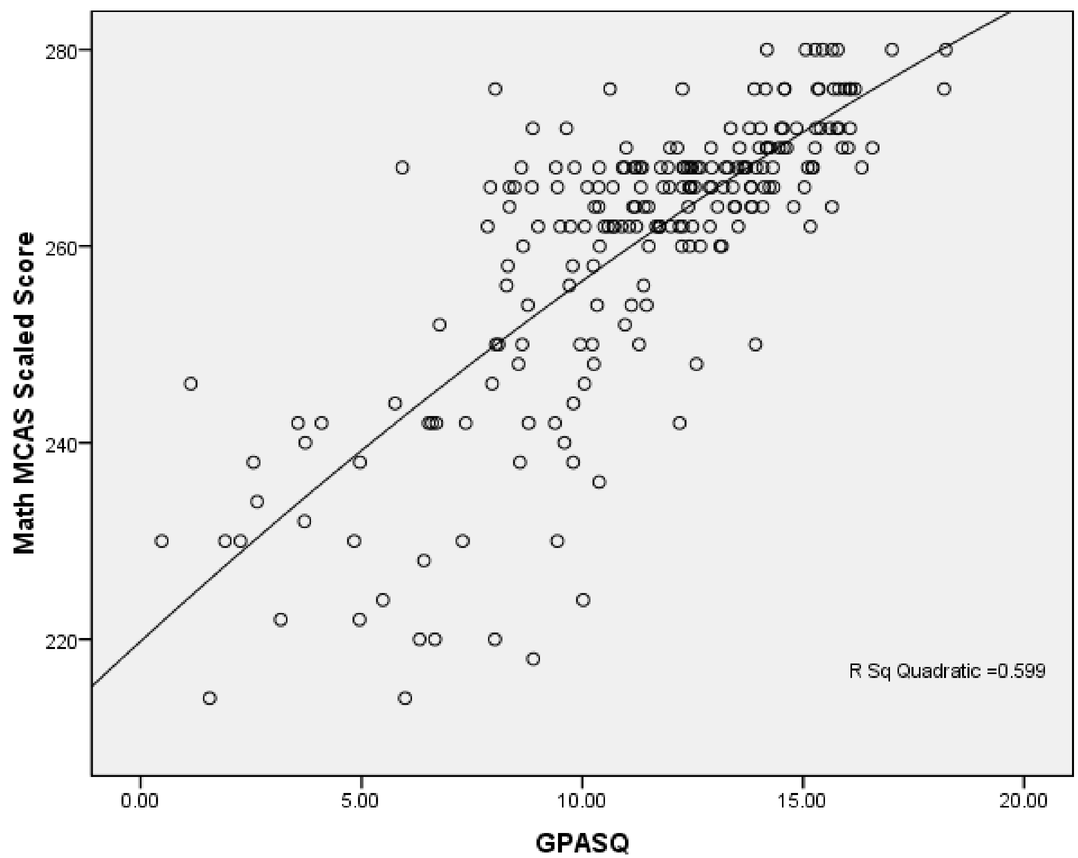


Exploratory Graphs (Second Iteration)






```
COMPUTE GPASQ=GPA**2.
EXECUTE.
```



Again, I had SPSS drop a quadratic curve to guide my eye. You can see a slight curvature but it's linear enough for government work.

Fitting the Right Model

$$MATHMCAS = \beta_0 + \beta_1 GPASQ + \varepsilon$$

Tolstoy begins his *Anna Karenina*, "Happy families are all alike; every unhappy family is unhappy in its own way."

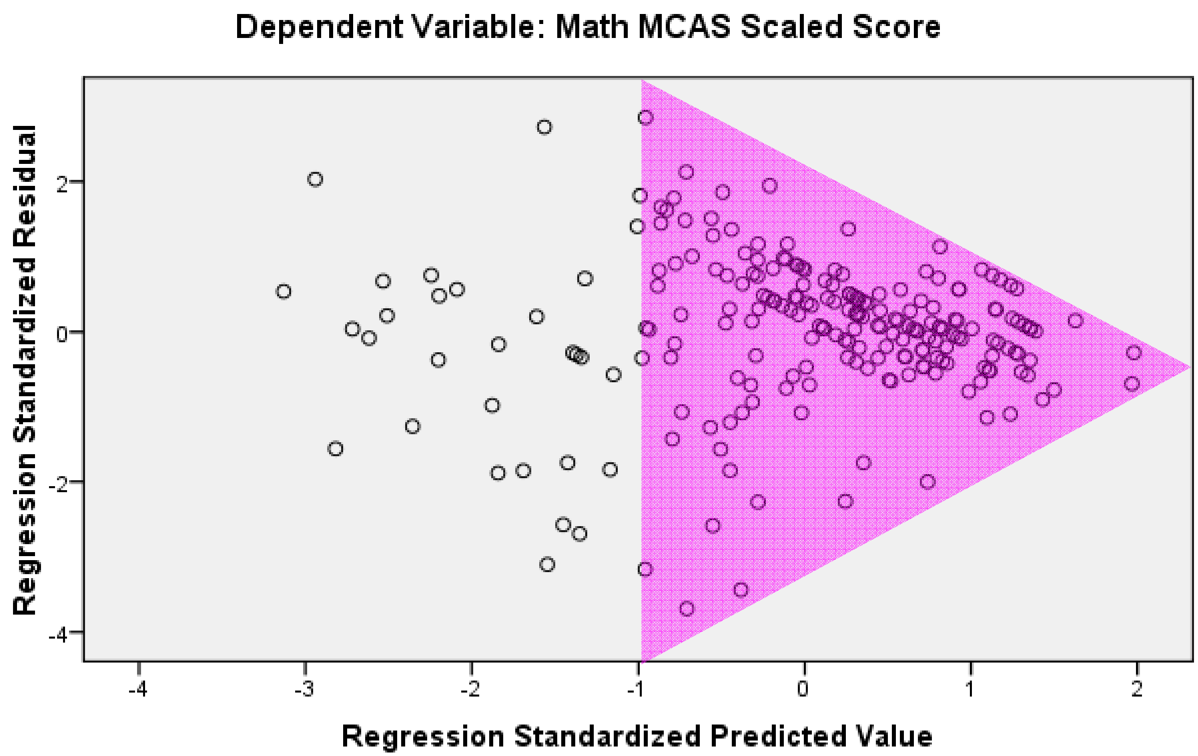
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	223.437	2.132		104.819	.000	219.236	227.638
	GPASQ	3.244	.180	.772	18.070	.000	2.890	3.598

a. Dependent Variable: Math MCAS Scaled Score

We've done a pretty good job of getting rid of the horseshoe, but what about the funnel?

Our RVF plot should reveal a patternless cloud. A funnel pattern indicates heteroskedasticity (also spelled "heterosc̲edasticity") which mean inequality of variances conditional on X.

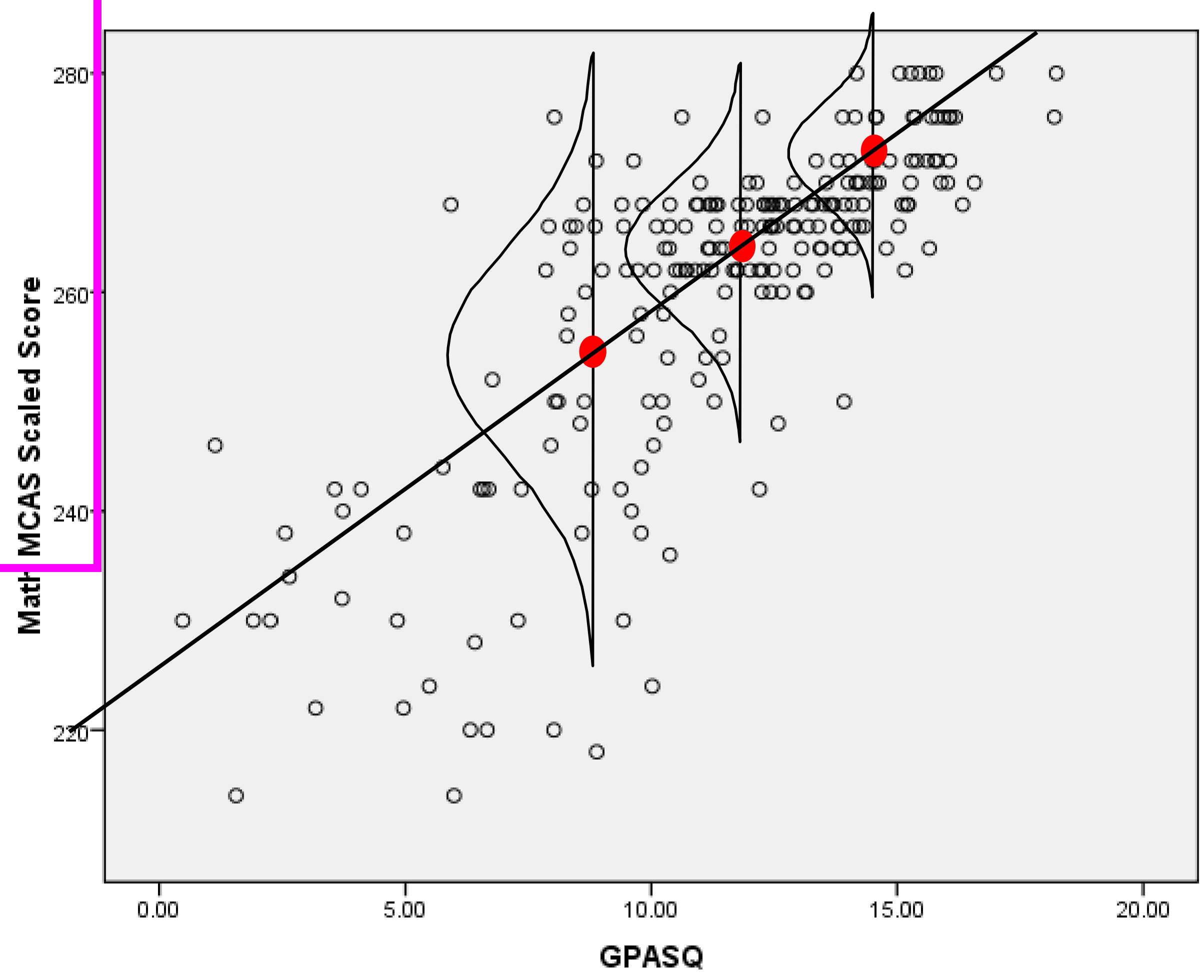
Just as with non-linearity, ask yourself, "what does heteroskedasticity mean substantively?"



Heteroskedasticity: The Problem and A Solution

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	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 heteroskedasticity 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.



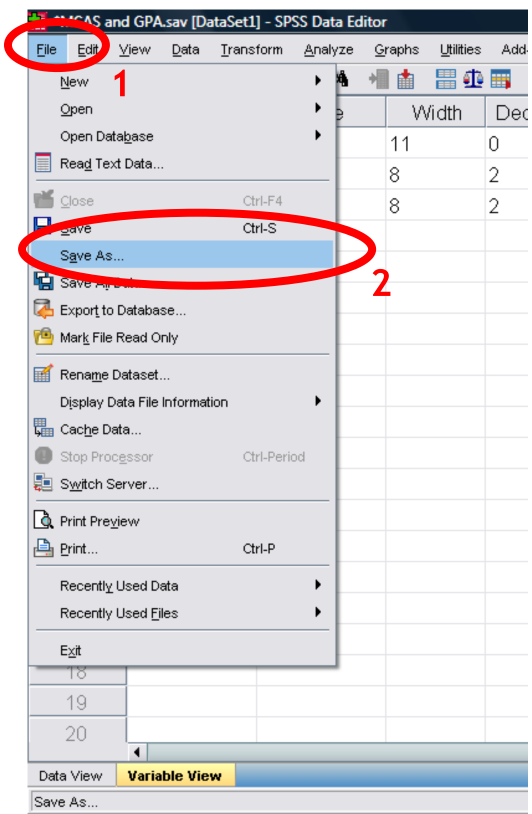
The problem is in our standard errors. Recall that a standard error is just a special kind of standard deviation—the standard deviation of THE sampling distribution. But, when there is a different sampling distribution at each level of X, which do we choose?

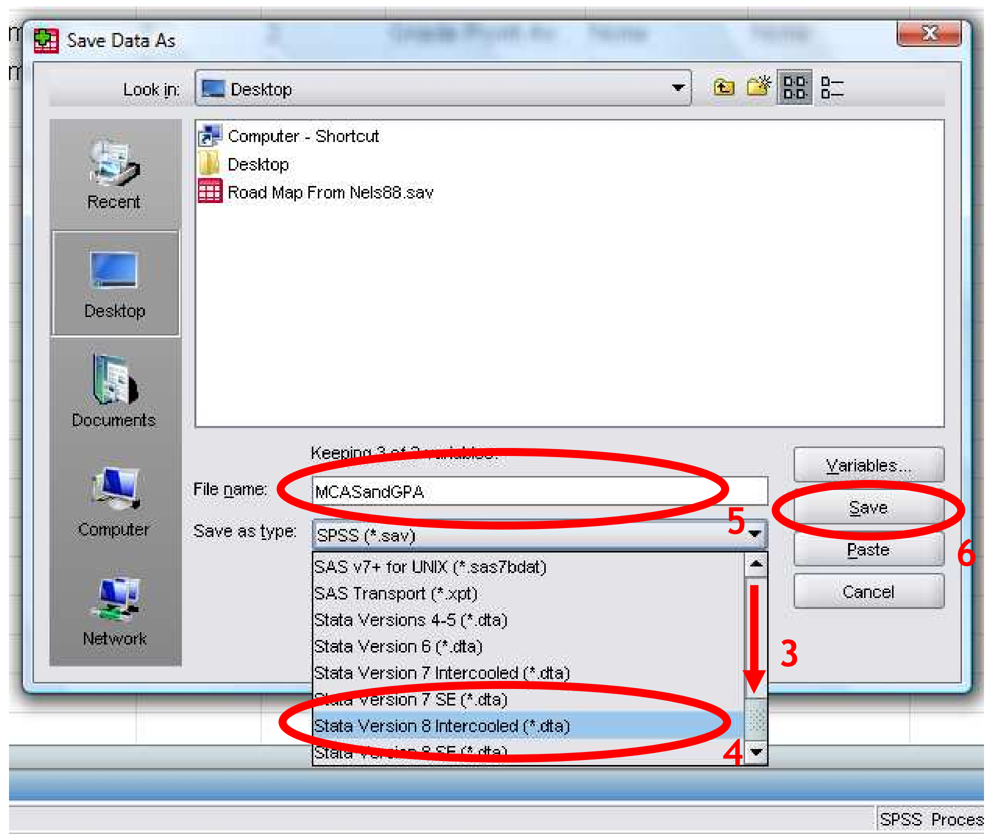
A solution is to calculate robust standard errors by using heteroskedasticity-consistent (HC) estimators.

Switching from SPSS to Stata

Sadly, SPSS offers no easy way to implement heteroskedasticity-consistent standard error (HCSE) estimation. In addition to a *relatively* comprehensible discussion of HCSE, Hayes and Cai (2007) provide two pages of SPSS syntax for HCSE: http://www.comm.ohio-state.edu/ahayes/BRM2007.pdf. Other statistical packages, however, have built-in HCSE functionality, so we will switch to one-Stata.

Roughly speaking, psychologists use SPSS, economists use Stata, and statisticians use SAS or R. Each package has its strengths and weaknesses, although you will find zealots for each who will disagree.





1. File
2. Save As…
3. Scroll down files types until you find your version of Stata (or something close).
4. Choose your version of Stata (or something close).
5. Name your file, but <u>do not use spaces</u>.
6. Save

Firing Up Stata

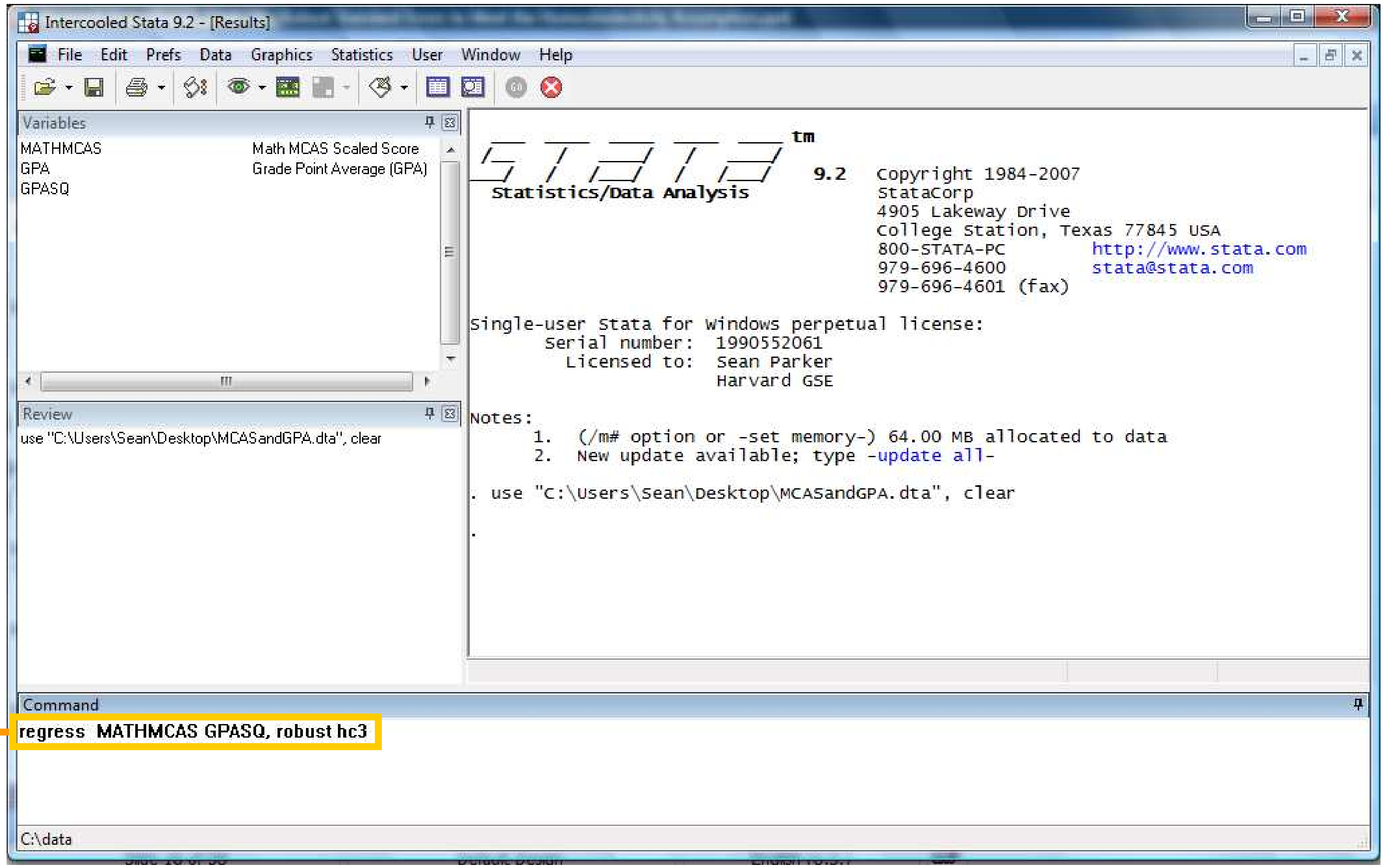
*Be careful. Stata is case sensitive.

regress MATHMCAS GPASQ, robust hc3

*If you have more than one predictor, list them:

regress OUTCOME PREDICTOR1 PREDICTOR2, robust hc3

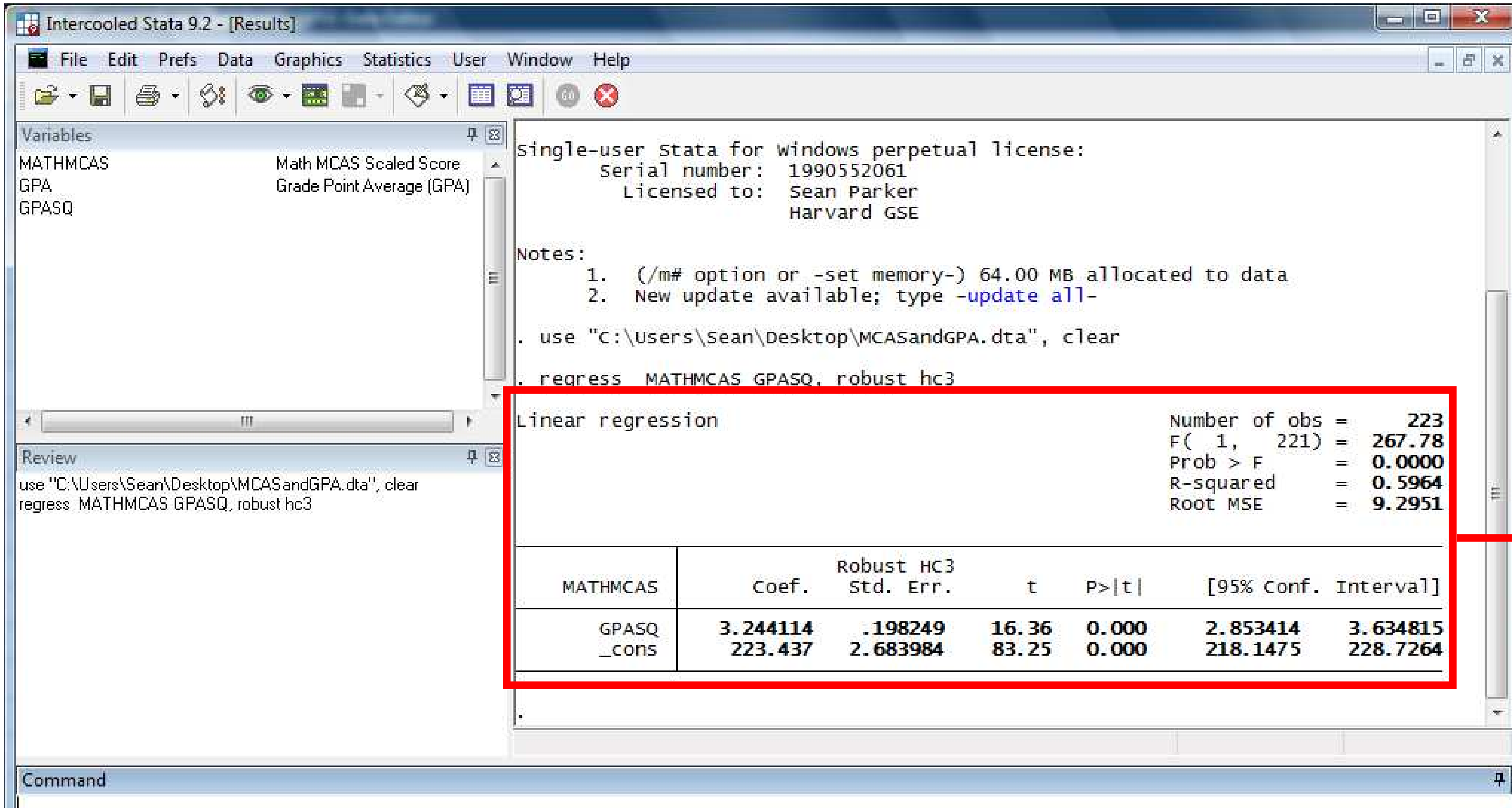
Through your MS Windows "Programs" menu, open your version of Stata. From Stata, open (File > Open…) your Stata formatted data set (*.dta). In the command window, enter your syntax, then hit enter.



Getting The Output

Once you hit enter, the output should appear.

If you want to copy and paste your output into a Word document, highlight and copy as normal. When you paste, you'll find that the format is screwy. Change the font to Courier (a monospace-type font), and things should line up.



Since there is no easy way to copy and paste the output into a Word document, copy the entire screen by hitting both your function key and your print screen key simultaneously. The screen shot is now on your clipboard ready for pasting.

C:\data

Comparing our Stata Output with our SPSS Output

SPSS Output

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	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

a. Dependent Variable: Math MCAS Scaled Score

Same **Different**

Stata Output

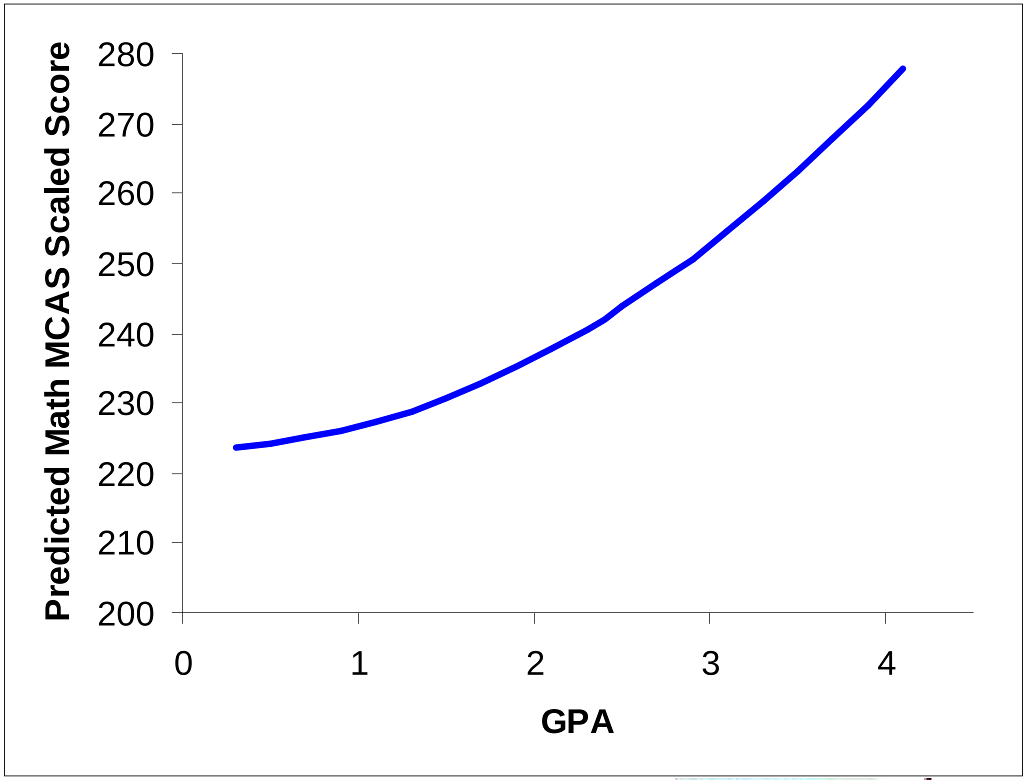
MATHMCAS	Coef.	Robust HC3 Std. Err.	t	P>|t	[95% Conf.	Interval]
GPASQ	3.244114	.198249	16.36	0.000	2.853414	3.634815
_cons	223.437	2.683984	83.25	0.000	218.1475	228.7264

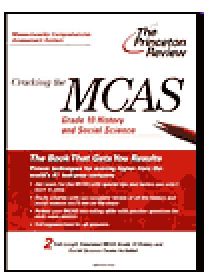
Do not memorize this, but be able to think in these terms based on your deepening understanding of standard errors:

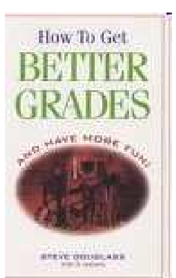
When we use HC standard errors, sometimes, as is the case here, they will be smaller than our run of the mill standard errors, but other times they will be larger. One way or the other, HC standard error are more trustworthy when we are analyzing heteroskedastic relationships. In terms of statistical significance, when our standard errors are too small (as is the case with our SPSS output), we are too prone to Type I Error (i.e., mistakenly rejecting the null hypothesis). When our standard errors are too large we are too prone to Type II Error (i.e., mistakenly failing to reject the null hypothesis). In terms of confidence intervals, when our standard errors are too small as is the case with or SPSS output, our 95% confidence intervals are really less than 95%. When our standard errors are too large, our 95% confidence intervals are really more than 95%.

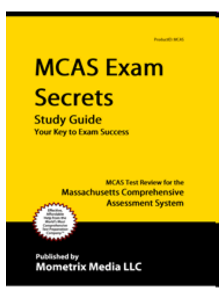
Interpreting our Results

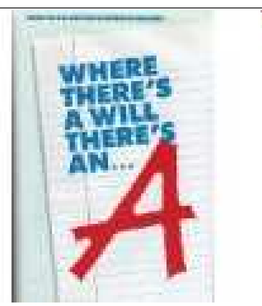
Figure 13.a. A fitted regression line showing the relationship between GPA and predicted math MCAS scaled scores for 223 AHS sophomores.











In our sample of 223 sophomores attending AHS, we found a statistically significant non-linear positive relationship between MATHMCAS and GPA^2, $t(221) = 16.36$, $p < .001$. We used robust standard errors (HC3) to derive our *t* statistic in order to correct for the heteroskedasticity in the data. Students with lower GPAs tended to exhibit greater variation in math MCAS scores than students with higher GPAs. We squared our predictor, GPA, in order to linearize the relationship for the purposes of OLS regression. Small differences among high GPAs lead us to predict fairly large differences in math MCAS scores, whereas small differences among low GPAs lead us to predict only small differences in math MCAS scores. Take for instance two students with A- and A overall GPAs, respectively. Based on averages, we predict that the higher GPA student will score more than 7 MCAS points higher. However, if we take two students with D- and D overall GPAs, respectively, we predict that the higher GPA student will score less than 2 MCAS points higher. Based on this trend, we can make controlled observations of students, identifying students who outperform or underperform our predictions, so that we can learn from their educational strategies in order to improve our educational strategies.

Bottom 10 and Top 10 MCAS Scores Controlling for GPA

The ten students who most underperformed on the math MCAS based on their GPA.

Case Summaries[a]

	Fictional Names	Math MCAS Scaled Score	Grade Point Average (GPA)	Unstandardized Residual
1	William	218	2.98	-34.28166
2	Susan	224	3.17	-31.94478
3	Serena	220	2.83	-29.45615
4	Patricia	214	2.45	-28.86243
5	Amanda	220	2.58	-25.04924
6	Arnold	230	3.07	-24.04306
7	Victor	220	2.51	-23.93737
8	Andrea	236	3.22	-21.11133
9	Emma	242	3.49	-21.02513
10	George	250	3.73	-18.60492

a. Limited to first 10 cases.

The ten students who most outperformed predictions based on their GPA.

Case Summaries[a]

	Fictional Names	Math MCAS Scaled Score	Grade Point Average (GPA)	Unstandardized Residual
1	Karen	276	2.83	26.51754
2	Jennifer	268	2.43	25.34155
3	Pavel	272	2.98	19.75525
4	Jonathan	246	1.07	18.87113
5	Kristy	276	3.26	18.09618
6	Krystal	272	3.10	17.28777
7	Andrew	266	2.81	16.88462
8	Patrick	268	2.94	16.59764
9	Robert	266	2.89	15.47228
10	Daniel	266	2.91	15.08627

a. Limited to first 10 cases.

Notice that William and Karen (in red) have nearly identical GPAs but widely different math MCAS scores.

```
SORT CASES BY RES_2(A).

SUMMARIZE
  /TABLES=FICTNAMES MATHMCAS GPA RES_2
  /FORMAT=VALIDLIST NOCASENUM TOTAL LIMIT=10
  /TITLE='Case Summaries'
  /MISSING=VARIABLE
  /CELLS=COUNT.
```

```
SORT CASES BY RES_2(D).

SUMMARIZE
  /TABLES=FICTNAMES MATHMCAS GPA RES_2
  /FORMAT=VALIDLIST NOCASENUM TOTAL LIMIT=10
  /TITLE='Case Summaries'
  /MISSING=VARIABLE
  /CELLS=COUNT.
```

Robust Standard Errors

Heteroskedasticity is inequality of variances in Y conditional on X.

Heteroskedasticity will not affect your parameter estimates, but it will distort your standard errors and, in turn, your t-tests and confidence intervals.

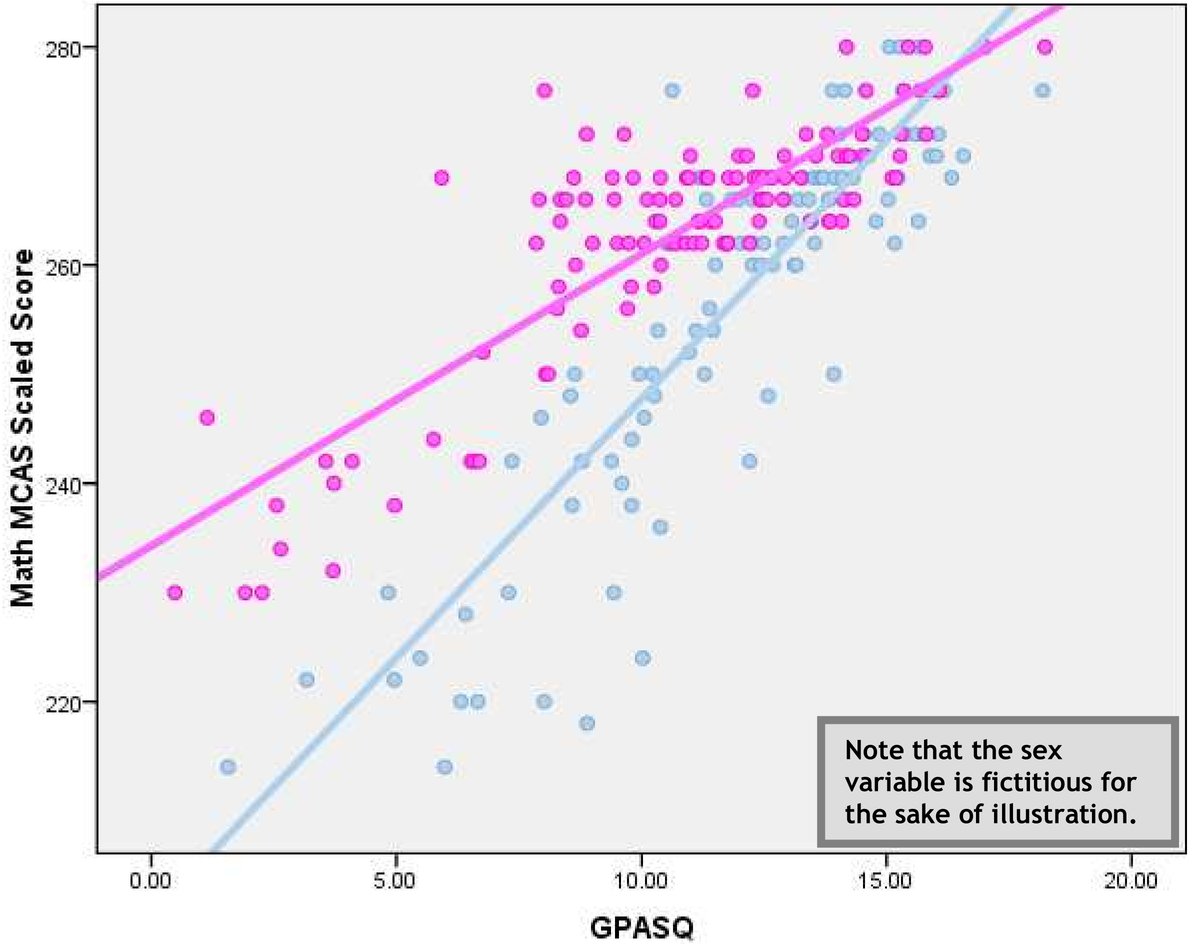
Identify heteroskedasticity problems by inspecting your residual versus fitted (RVF) plots for funnel shapes.

If you have a problem with heteroskedasticity, use robust standard errors, in particular, heteroskedasticity-consistent (HC3) standard errors.

Other Solutions: Statistical Interaction and Weighted Least Squares

Sometimes heteroskedasticity appears because we are specifying the wrong model. Sometimes the cure for heteroskedasticity is multiple regression with interactions, which we learn in Unit 17.

Imagine that we examine the relationship between MCAS and GPA for girls and boys separately. It is possible that two (or more) homoskedastic relationships can be heteroskedastic when taken as a whole.



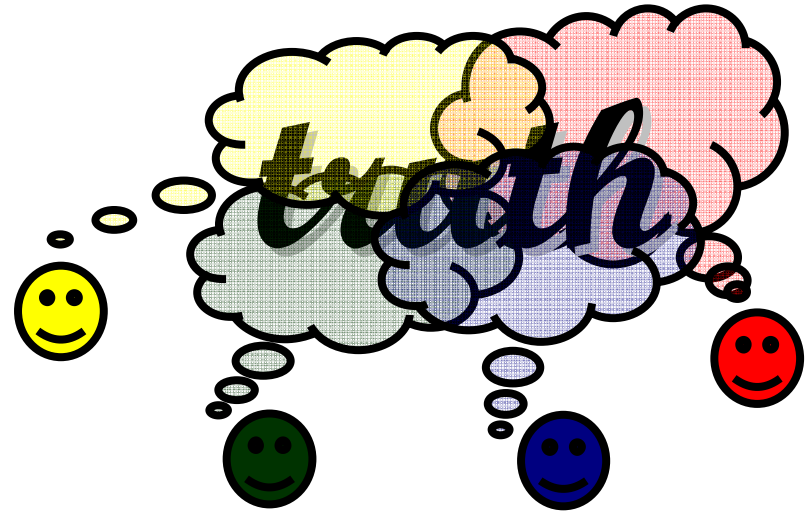
Note that the sex variable is fictitious for the sake of illustration.

$$MATHMCAS = \beta_0 + \beta_1 GPASQ + \beta_2 FEMALE + \beta_3 GPASQxFEMALE + \varepsilon$$

If we can model the pattern of heteroskedasticity in terms of the predictors, we can do weighted least squares (WLS) regression instead of ordinary least squares (OLS) regression. This WLS solution is very difficult, because it requires a theory of population variation so specific that we can mathematize it. However, if we have the right theory and the right formula for that theory, we can weight our observations appropriately and derive standard errors from a weighted average sampling distribution.

Cognitive Division of Labor

Underlying most statistical tests are mathematical proofs that you could learn to comprehend in a theoretical statistics course. At Harvard, the prerequisites for Statistics 101 are three courses in calculus (differentiation, integration, and multivariable), a course in linear algebra and a course in probability. Note that heteroskedasticity-consistent standard errors are too advanced for Statistics 101.



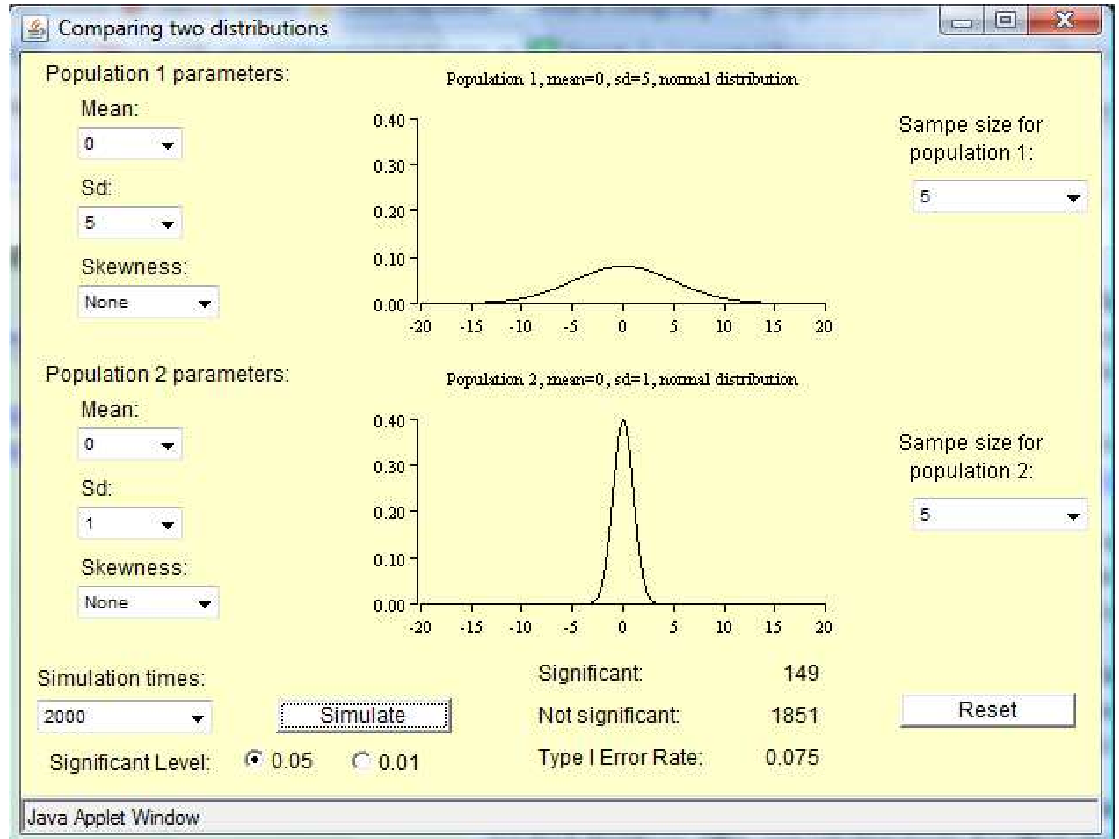
- The statisticians tell me that the HCSE math works. (I am working on the skills to check myelf.)
- The Monte Carlo methodologists can't come up with a heteroskedastic population in which HCSEs fail. (Again, I am working on the skills to try my hand at creating a problematic population.)
- The economists have adopted HCSEs nearly unanimously. (A million economists can't be wrong, right? Right?)

White's (1980) article on HCSEs is the most cited article in economics since being published. He invented (or is it "discovered"?) HC0 estimators which were improved upon by other statisticians who found HC0 estimators were not robust in small samples, and they invented/discovered HC3 estimators to solve the small sample problem.

White; Halbert (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity", *Econometrica* 48 (4): 817--838, doi:10.2307/1912934

The Concept Of Monte Carlo Methods

Monte Carlo methods provide statistical tests for statistical tests. A statistical test can justify our inference from a sample to a population, but what justifies our reliance on the statistical test? Underlying most statistical tests are mathematical proofs that you could learn to comprehend in a theoretical statistics course. However, in the meantime, there are more comprehensible methods for support. A statistical programmer can create her own population (with carefully specified parameters) and draw a random sample from it. She can then use a statistical test to draw conclusions from the sample to the population, but because she created the population, she already knows the right answer, so she can determine whether the test is working as advertised. These methods are called *Monte Carlo methods* after the famous casino city of Monte Carlo.



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

We can do our own Monte Carlo test of non-robust standard errors. We can specify a population in which there is no average difference between boys and girls, but in which there is a difference in variances (e.g., heteroskedasticity). Our Type I Error Rate should be 0.05 because our alpha level is 0.05. Is our Type I Error Rate in fact 0.05?

All GLM-based statistical tests are fairly robust to small violations of the homoskedasticity (and normality) assumptions.

We Can Calculate "Robust" Standard Errors for Two-Sample T-Tests

A two-sample T-test is just the T-test for the slope coefficient from regressing a continuous outcome on a dichotomous predictor. Because they are so easy to calculate by hand, two-sample T-tests are the focus of many statistics courses which use them to introduce the concept of statistical hypothesis testing. (I don't focus on two-sample T-tests, because I focus on the more general tool of regression which subsumes two-sample T-tests.)

A Translation Guide

Two-Sample T-Test Language	Regression Language
Two Populations	One sample from one population, with two groups indicated by a dichotomous predictor variable
Two Samples	
Estimated Mean Difference	Slope Parameter Estimate (β_1)

The key to any statistical null-hypothesis test is to see if your estimate is more than two standard errors from zero. Thus, the key is calculating standard errors, by which you can divide your estimate.

"Non-Robust" Standard Error (i.e., Pooled)	Robust Standard Error (i.e., Satterthwaite)
$se_p = \sqrt{\left(\boxed{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$	$se_s = \sqrt{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)}$ Notice that the variances are ***weighted*** by the sample sizes.

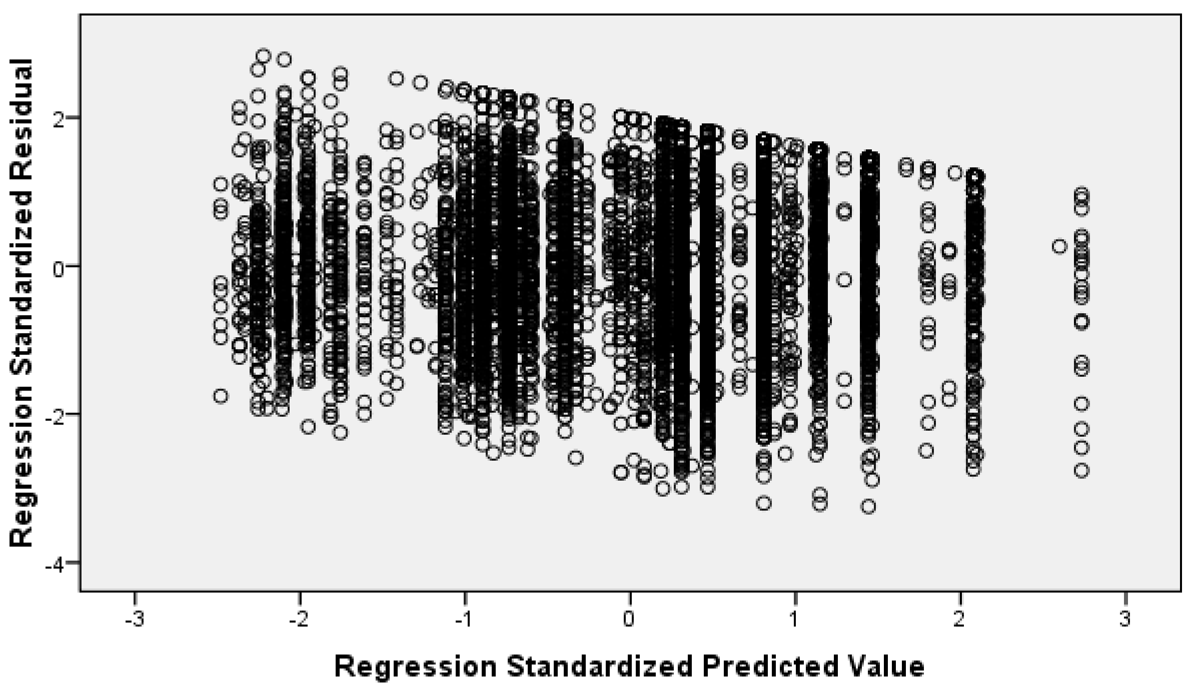
$$\text{where}: n_1 = \text{sample size for group 1},\ n_2 = \text{sample size for group 2}$$
$$s_1^2 = \text{variance for group 1},\ s_2^2 = \text{variance for group 2}$$

With pooled standard errors, we pool/combine the variances in one numerator, but we shouldn't combine them if they differ in the population.

Answering our Roadmap Question

Unit 14: How do we deal with violations of the ***heteroskedasticity*** assumptions?

$$READING = \beta_0 + \beta_1 ASIAN + \beta_2 BLACK + \beta_3 LATINO + \beta_4 L2HOMEWORKP1 + \beta_5 ESL + \beta_6 FREELUNCH + \beta_7 ESLxASIAN + \beta_8 ESLxBLACK + \beta_9 ESLxLATINO + \beta_{10} FREELUNCHxASIAN + \beta_{11} FREELUNCHxBLACK + \beta_{12} FREELUNCHxLATINO + \varepsilon$$



As we can see from this RVF plot, there is a ceiling effect, which is easy to mistake for heteroskedasticity. If we look closely at the slices, however, we see that their variances are roughly equal. We'll try robust standard errors anyway, but no data analytic decision is going to undo the ceiling effect. In the words of John Willett, "We cannot fix with data analysis what we bungled by design."

Answering our Roadmap Question

Unit 14: How do we deal with violations of the ***heteroskedasticity*** assumptions?

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	45.358	.288		157.560	.000	44.794	45.923
	ASIAN	-.377	.668	-.011	-.564	.573	-1.687	.933
	BLACK	-3.447	.498	-.113	-6.922	.000	-4.423	-2.471
	LATINO	-2.779	.517	-.102	-5.371	.000	-3.793	-1.765
	L2HOMEWORKP1	1.591	.100	.169	15.866	.000	1.394	1.788
	ESL	-.876	.638	-.035	-1.373	.170	-2.126	.374
	FREELUNCH	-3.574	.235	-.197	-15.208	.000	-4.035	-3.113
	ESLxASIAN	3.245	.999	.080	3.249	.001	1.287	5.202
	ESLxBLACK	5.872	1.885	.036	3.115	.002	2.177	9.568
	ESLxLATINO	.446	.858	.013	.520	.603	-1.235	2.127
	FREELUNCHxASIAN	-2.769	.853	-.041	-3.245	.001	-4.442	-1.096
	FREELUNCHxBLACK	-.751	.666	-.019	-1.127	.260	-2.058	.555
	FREELUNCHxLATINO	-.437	.604	-.012	-.724	.469	-1.622	.747

a. Dependent Variable: READING

READING	Coef.	Robust HC3 Std. Err.	t	P>|t|	[95% Conf.	Interval]
ASIAN	-.3769188	.6598317	-0.57	0.568	-1.670366	.9165286
BLACK	-3.447231	.5009758	-6.88	0.000	-4.429278	-2.465184
LATINO	-2.779094	.5320832	-5.22	0.000	-3.82212	-1.736068
L2HOMEWORKP1	1.591023	.1042259	15.27	0.000	1.386713	1.795334
ESL	-.8757125	.6685416	-1.31	0.190	-2.186234	.4348086
FREELUNCH	-3.574192	.2397379	-14.91	0.000	-4.044142	-3.104241
ESLXASIAN	3.244724	1.017383	3.19	0.001	1.250379	5.239069
ESLXBLACK	5.872272	1.970221	2.98	0.003	2.01011	9.734434
ESLXLATINO	.4460223	.8730504	0.51	0.609	-1.265391	2.157436
FREELUNCHx~N	-2.769305	.8643695	-3.20	0.001	-4.463701	-1.074908
FREELUNCHx~K	-.751243	.6584616	-1.14	0.254	-2.042005	.5395186
FREELUNCHx~O	-.4372372	.5866277	-0.75	0.456	-1.587185	.7127107
_cons	45.35835	.3019198	150.23	0.000	44.76651	45.9502

Std. Error		Robust HC3 Std. Err.
.288		
.668	>	.6598317
.498	<	.5009758
.517	<	.5320832
.100	<	.1042259
.638	<	.6685416
.235	<	.2397379
.999	<	1.017383
1.885	<	1.970221
.858	<	.8730504
.853	<	.8643695
.666	>	.6584616
.604	>	.5866277
		.3019198

In order to address heteroskedasticity concerns, we used robust standard errors (HC3).

We can see differences in our p-values, but none is big enough to change the results of our hypothesis tests.

Unit 14 Appendix: Key Concepts

Our RVF plot should reveal a patternless cloud. A funnel pattern indicates heteroskedasticity (also spelled "heterosc͟edasticity") which mean inequality of variances conditional on X.

Just as with non-linearity, ask yourself, "what does heteroskedasticity mean substantively?"

The problem with heteroskedasticity 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.

A solution is to calculate robust standard errors by using heteroskedasticity-consistent (HC) estimators.

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?

Do not memorize this, but be able to think in these terms based on your deepening understanding of standard errors: When we use HC standard errors, sometimes, as is the case here, they will be smaller than our run of the mill standard errors, but other times they will be larger. One way or the other, HC standard error are more trustworthy when we are analyzing heteroskedastic relationships. In terms of statistical significance, when our standard errors are too small (as is the case with our SPSS output), we are too prone to Type I Error (i.e., mistakenly rejecting the null hypothesis). When our standard errors are too large we are too prone to Type II Error (i.e., mistakenly failing to reject the null hypothesis). In terms of confidence intervals, when our standard errors are too small as is the case with or SPSS output, our 95% confidence intervals are really less than 95%. When our standard errors are too large, our 95% confidence intervals are really more than 95%.

Sometimes heteroskedasticity appears because we are specifying the wrong model. Sometimes the cure for heteroskedasticity is multiple regression with interactions, which we learn in Unit 17.

All GLM-based statistical tests are fairly robust to small violations of the homoskedasticity (and normality) assumptions.

Unit 14 Appendix: Key Interpretations

In our sample of 223 sophomores attending AHS, we found a statistically significant non-linear positive relationship between MATHMCAS and GPA2, $t(221) = 16.36$, $p < .001$. We used robust standard errors (HC3) to derive our t statistic in order to correct for the heteroskedasticity in the data. Students with lower GPAs tended to exhibit greater variation in math MCAS scores than students with higher GPAs. We squared our predictor, GPA, in order to linearize the relationship for the purposes of OLS regression. Small differences among high GPAs lead us to predict fairly large differences in math MCAS scores, whereas small differences among low GPAs lead us to predict only small differences in math MCAS scores. Take for instance two students with A- and A overall GPAs, respectively. Based on averages, we predict that the higher GPA student will score more than 7 MCAS points higher. However, if we take two students with D- and D overall GPAs, respectively, we predict that the higher GPA student will score less than 2 MCAS points higher. Based on this trend, we can make controlled observations of students, identifying students who outperform or underperform our predictions, so that we can learn from their educational strategies in order to improve our educational strategies.

In order to address heteroskedasticity concerns, we used robust standard errors (HC3).

Unit 14 Appendix: Key Terminology

Heteroskedasticity is inequality of variances conditional on our predictor(s).

Unit 14 Appendix: Formulas

packages—the estimator that assumes homoskedasticity (OLSE)—is

$$\text{OLSE}_j = \sqrt{\frac{MS_{\text{residual}}}{n\left(s_j^2\right)\left(1 - R_j^2\right)}}, \qquad (3)$$

where s_j is the sample standard deviation of X_j, R_j is the multiple correlation estimating X_j from the other $(p - 1)$ X variables in the model, and MS_{residual} is the mean squared residual. The extent of the problem produced by heteroskedasticity depends on both the form and the severity of heteroskedasticity. When the errors are heteroskedastic, OLSE is both *biased* and *inconsistent*. The net result is Type I error inflation or reduced statistical power for tests of hypotheses involving the regression coefficients, and inaccuracy in the estimation of the upper and lower bounds on confidence intervals. Increasing the sample size does not eliminate the bias; indeed, it can actually exacerbate the problems produced by heteroskedasticity when using OLSE (Hayes, 1996; Long & Ervin, 2000).

http://www.comm.ohio-state.edu/ahayes/BRM2007.pdf

Unit 14 Appendix: SPSS Syntax

Do not be afraid to dip in Stata or any other statistical package to meet your data analytic needs.

XXX

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)

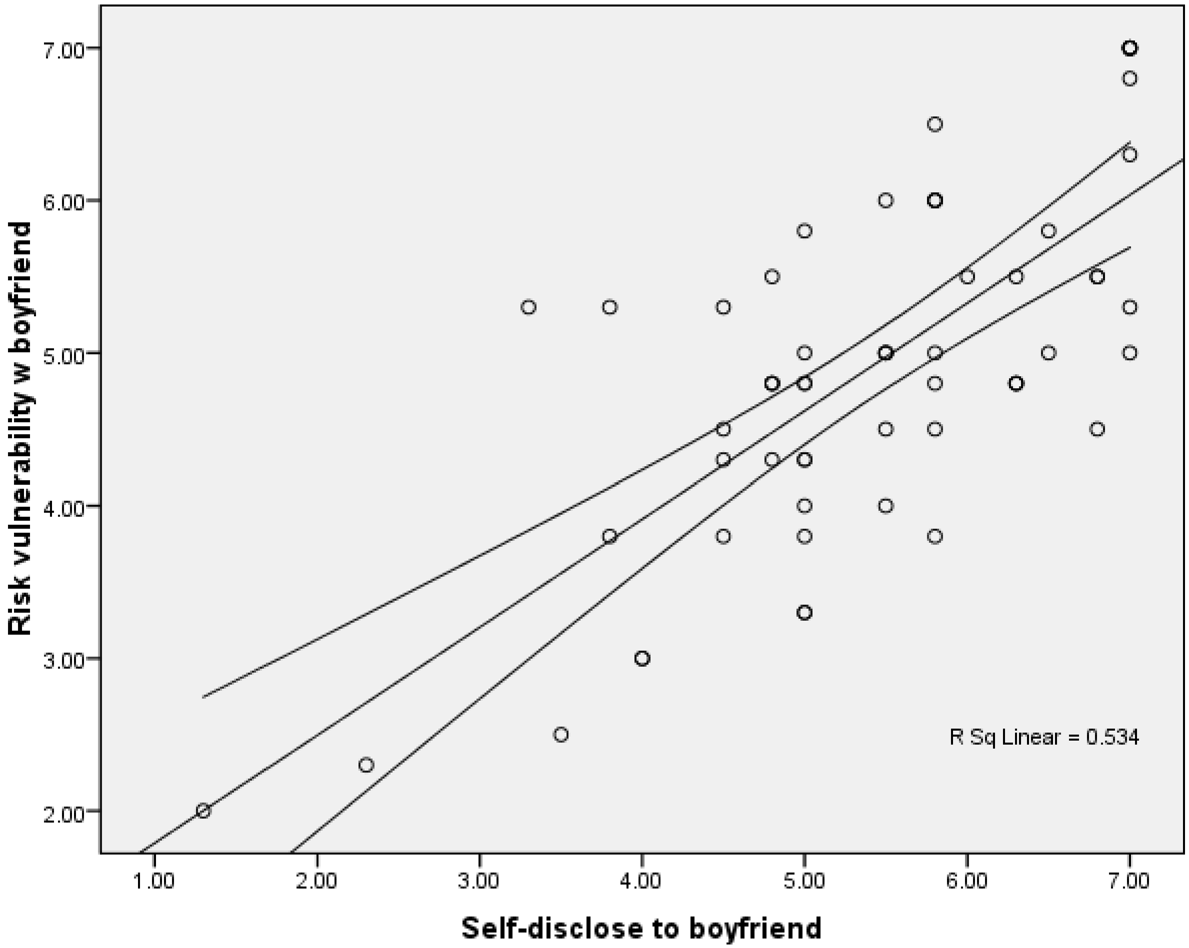
Coefficients[a]

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

B_Vuln	Coef.	Robust HC3 Std. Err.	t	P>|t|	[95% Conf.	Interval]
B_SelDis	.7078333	.0933239	7.58	0.000	.5210252	.8946414
_cons	1.080903	.5230132	2.07	0.043	.0339782	2.127828

Perceived Intimacy of Adolescent Girls (Intimacy.sav)



Perceived Intimacy of Adolescent Girls (Intimacy.sav)

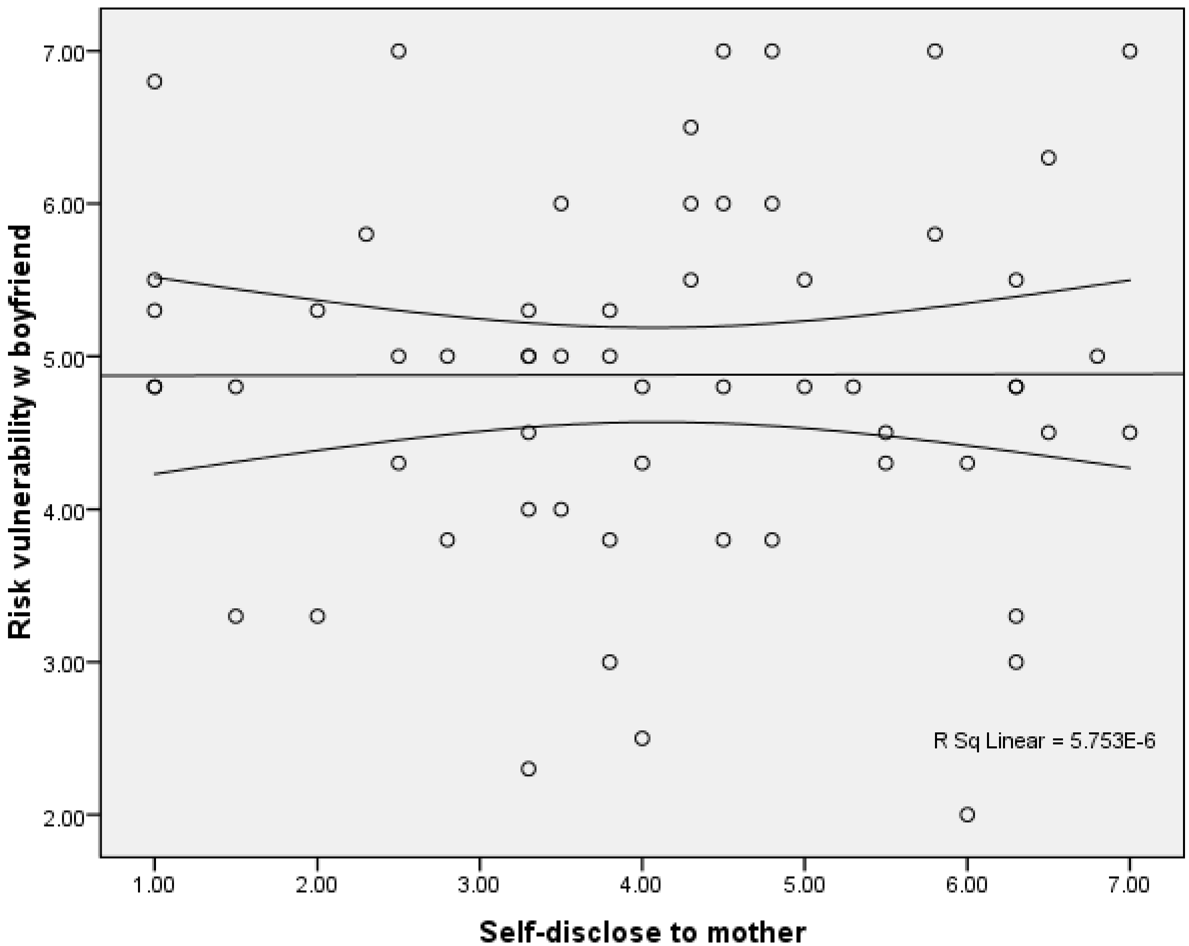
Coefficients[a]

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

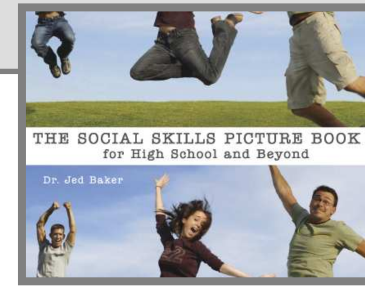
a. Dependent Variable: Risk vulnerability w boyfriend

B_Vuln	Coef.	Robust HC3 Std. Err.	t	P>|t|	[95% Conf.	Interval]
M_SelDis	.0016683	.0926282	0.02	0.986	-.1837472	.1870838
_cons	4.87151	.3856098	12.63	0.000	4.099628	5.643392

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:**

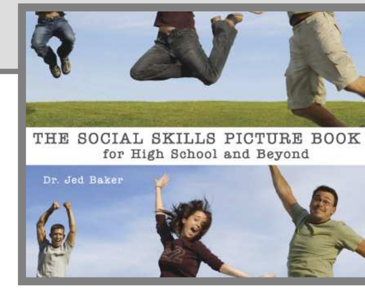
Variables about the student—

(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)



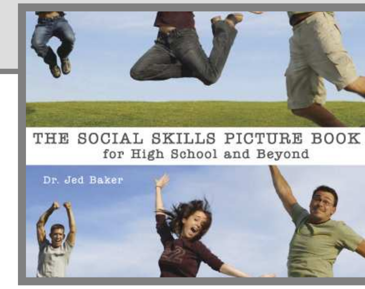
Coefficients[a]

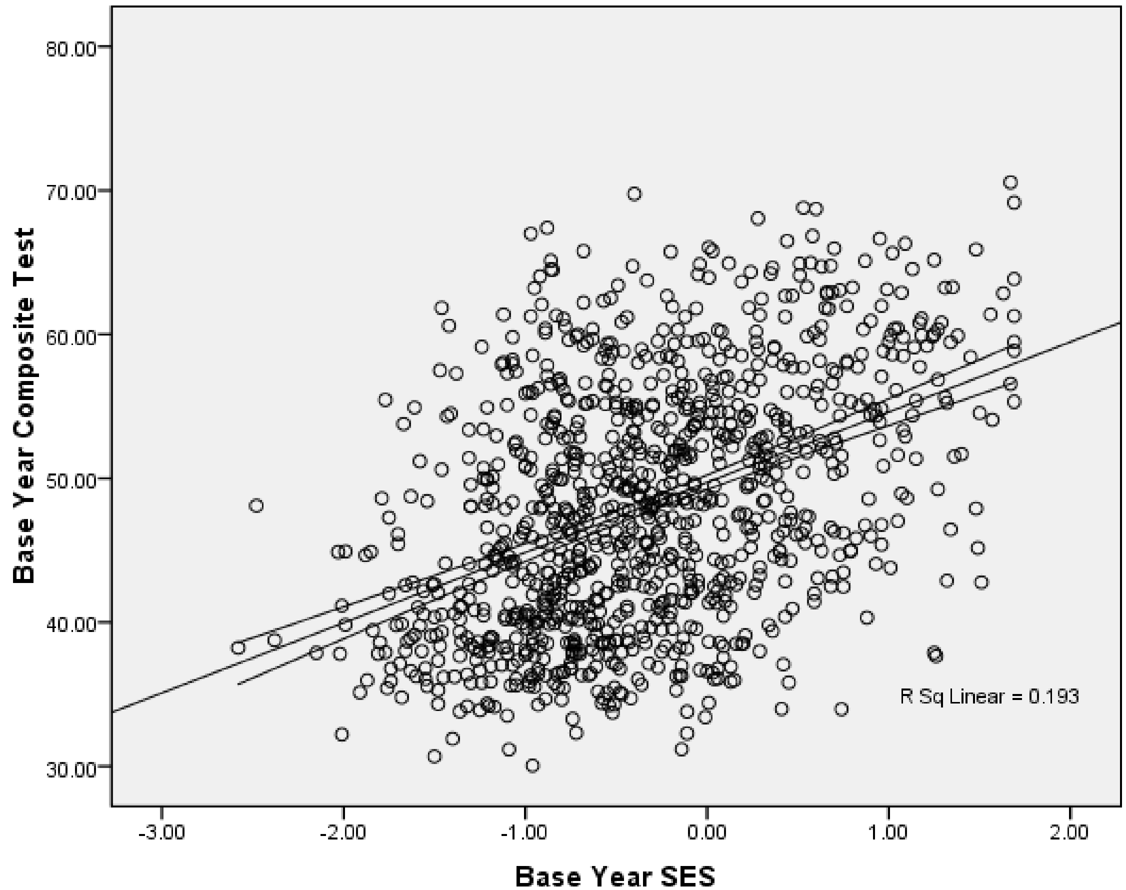
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	49.726	.260		191.448	.000	49.216	50.235
	Base Year SES	4.879	.309	.440	15.795	.000	4.273	5.485

a. Dependent Variable: Base Year Composite Test

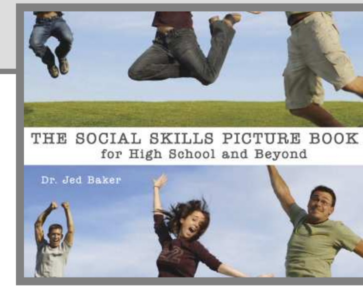
BYTest	Coef.	Robust HC3 Std. Err.	t	P>|t|	[95% Conf.	Interval]
BYSES	4.878948	.289184	16.87	0.000	4.311498	5.446397
_cons	49.72573	.2593947	191.70	0.000	49.21673	50.23472

High School and Beyond (HSB.sav)





High School and Beyond (HSB.sav)



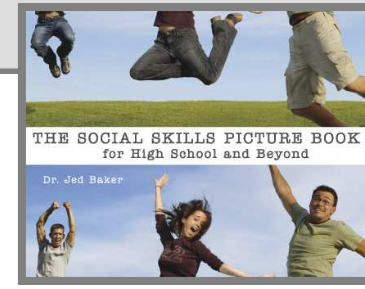
Coefficients[a]

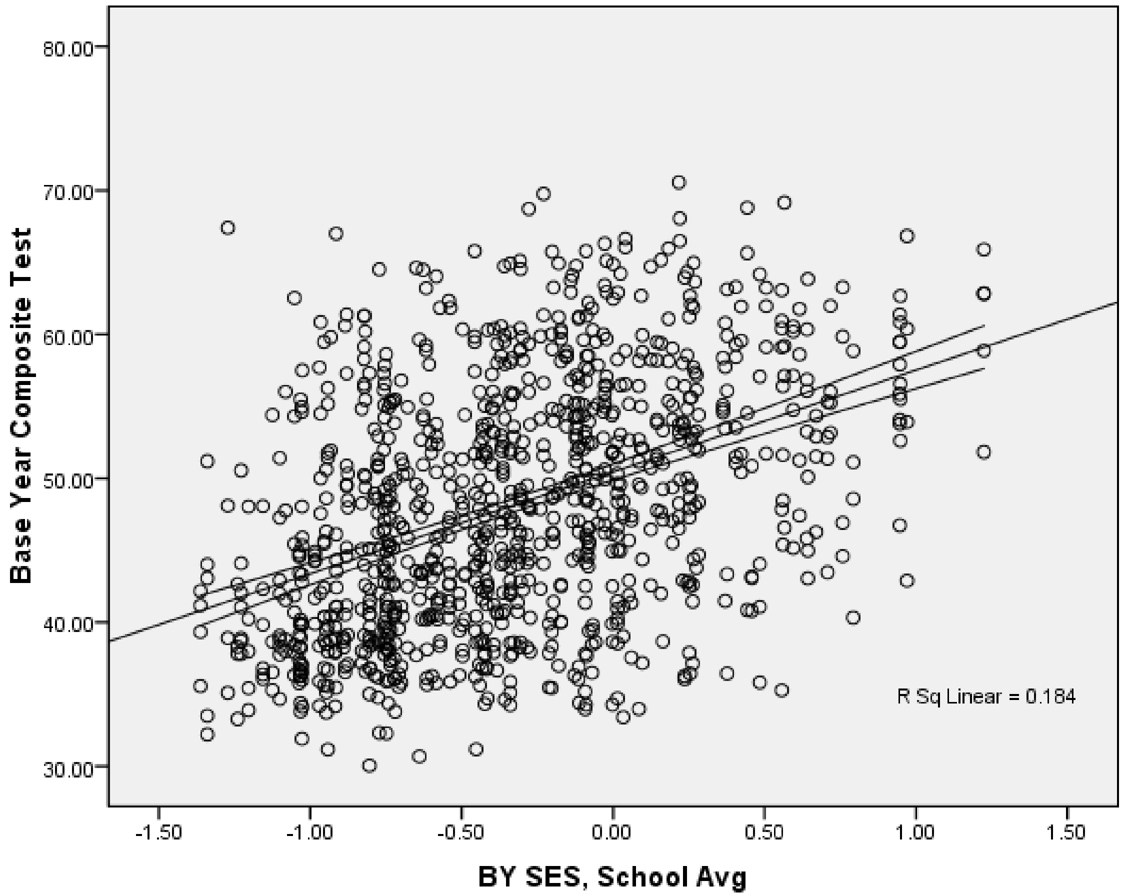
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	50.451	.284		177.397	.000	49.893	51.009
	BY SES, School Avg	7.075	.461	.429	15.344	.000	6.171	7.980

a. Dependent Variable: Base Year Composite Test

BYTest	Coef.	Robust HC3 Std. Err.	t	P>|t|	[95% Conf.	Interval]
BYSES_S	7.075463	.435426	16.25	0.000	6.221052	7.929875
_cons	50.4514	.2841164	177.57	0.000	49.89389	51.0089

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)

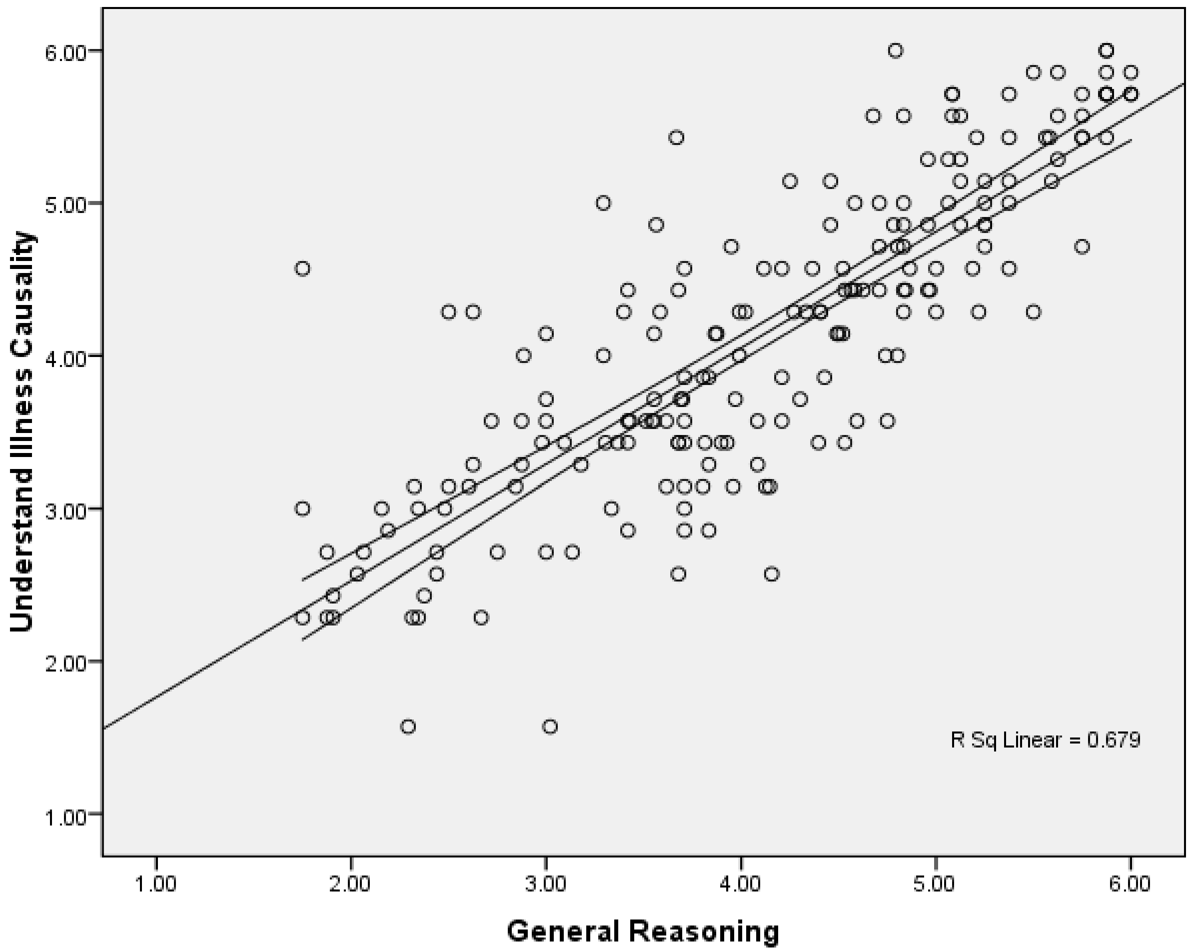
Coefficients[a]

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

IllCause	Coef.	Robust HC3 Std. Err.	t	P>|t|	[95% Conf.	Interval]
GenReas	.7618263	.038762	19.65	0.000	.6853672	.8382855
_cons	1.003958	.1780077	5.64	0.000	.652833	1.355083

Understanding Causes of Illness (ILLCAUSE.sav)



Understanding Causes of Illness (ILLCAUSE.sav)

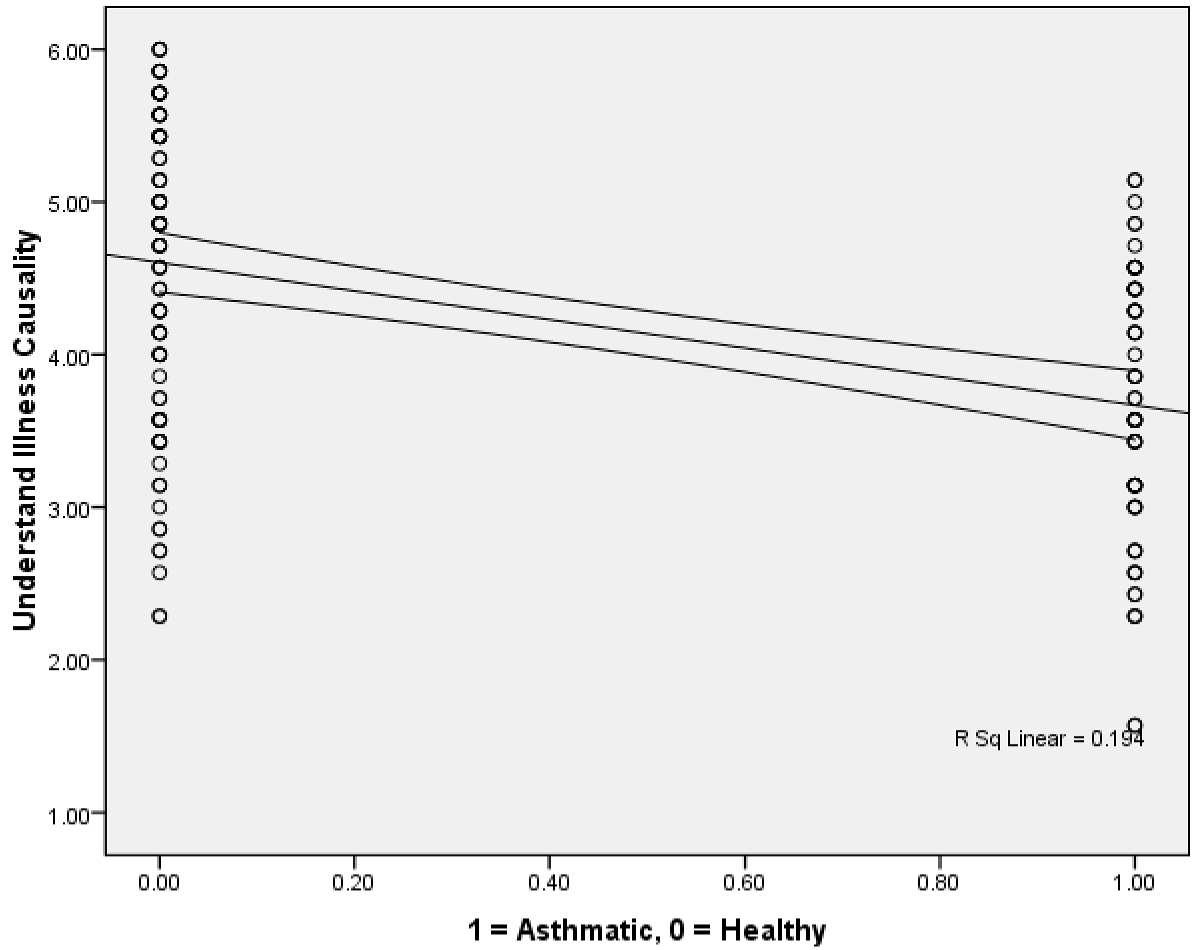
Coefficients[a]

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

IllCause	Coef.	Robust HC3 Std. Err.	t	P>|t|	[95% Conf.	Interval]
Asthmatic	-.9355971	.1489578	-6.28	0.000	-1.229788	-.641406
_cons	4.603656	.1045334	44.04	0.000	4.397203	4.810109

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)

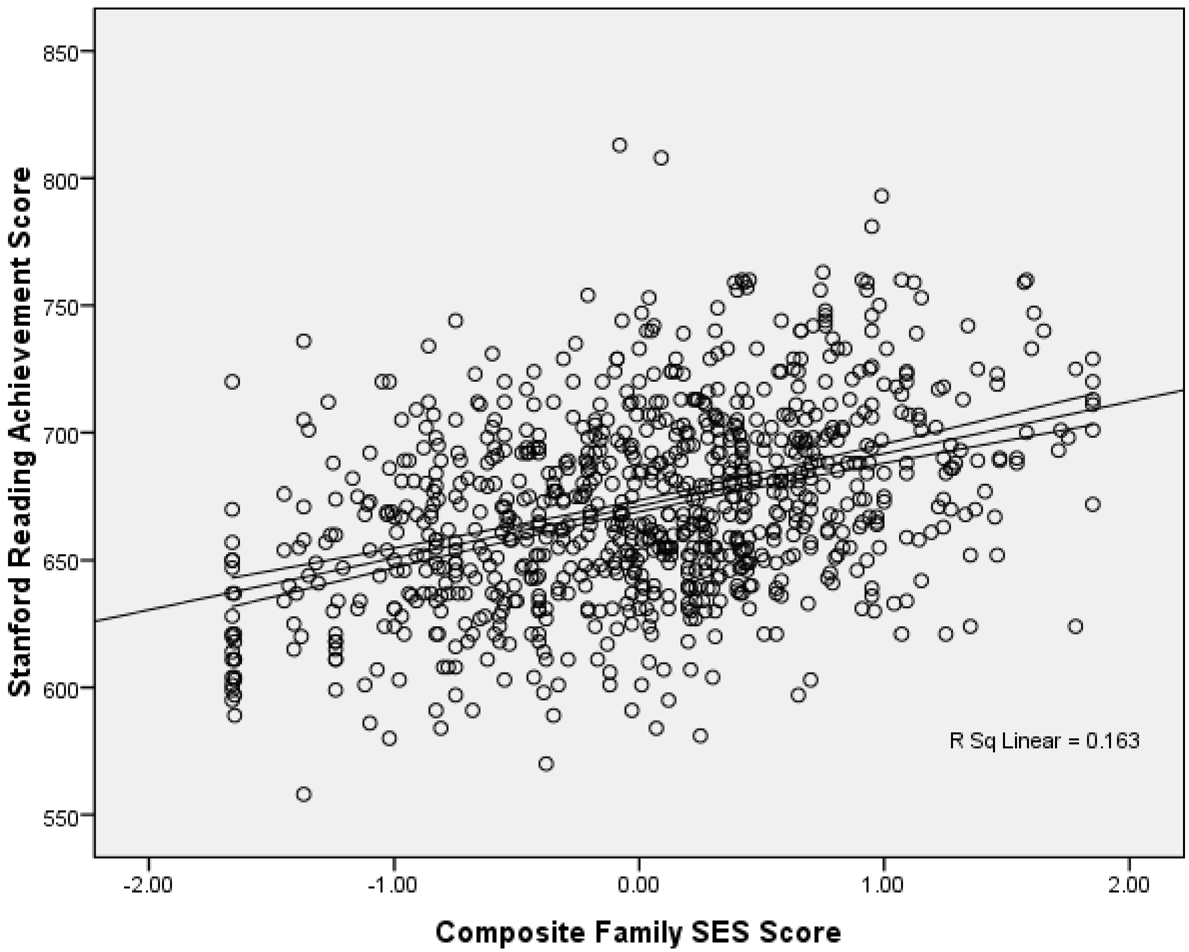
Coefficients[a]

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

Reading	Coef.	Robust HC3 Std. Err.	t	P>|t|	[95% Conf.	Interval]
SES	20.41768	1.531877	13.33	0.000	17.41111	23.42425
_cons	671.3502	1.175045	571.34	0.000	669.044	673.6565

Children of Immigrants (ChildrenOfImmigrants.sav)



Children of Immigrants (ChildrenOfImmigrants.sav)

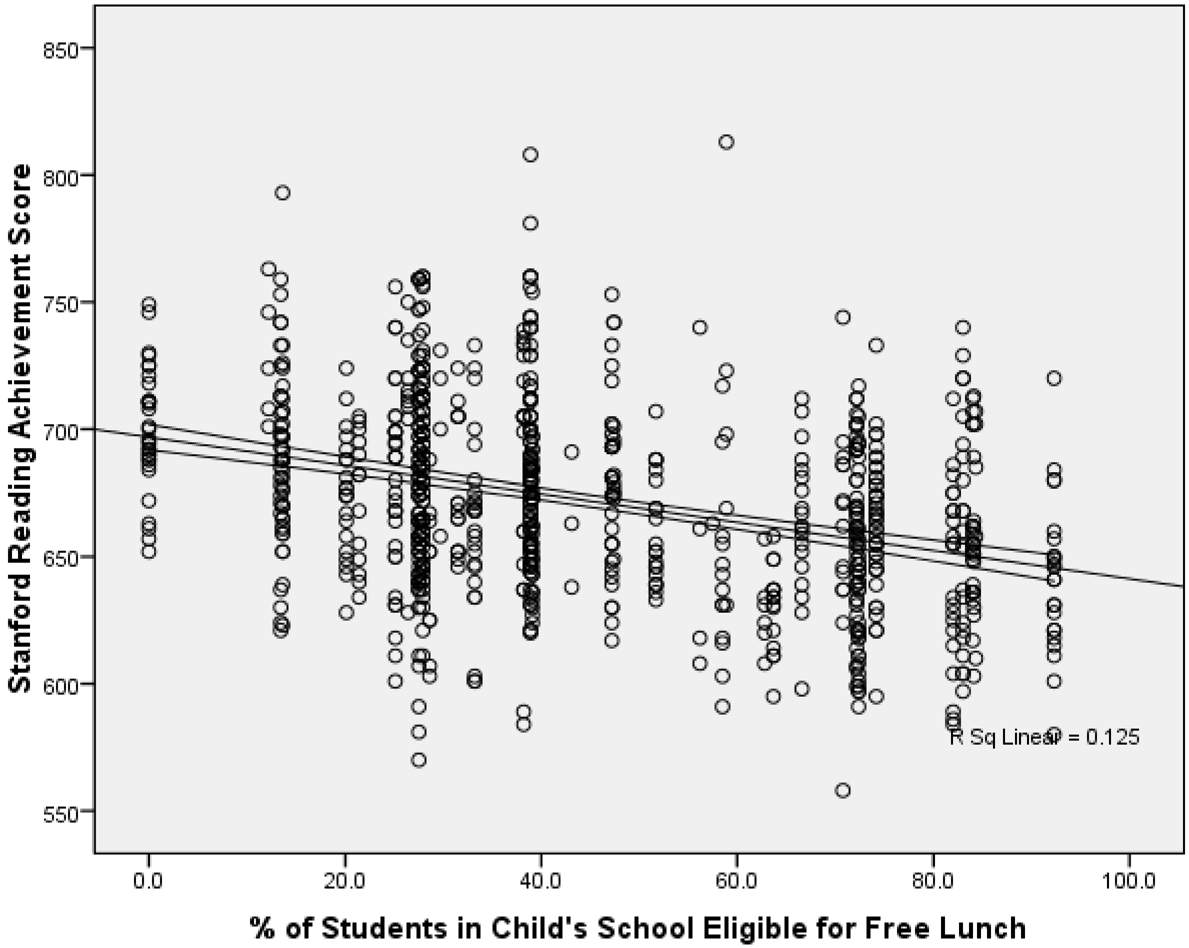
Coefficients[a]

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

Reading	Coef.	Robust HC3 Std. Err.	t	P>|t|	[95% Conf.	Interval]
FreeLunch	-.5553343	.0468145	-11.86	0.000	-.6472157	-.4634528
_cons	696.8465	2.449119	284.53	0.000	692.0397	701.6533

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)

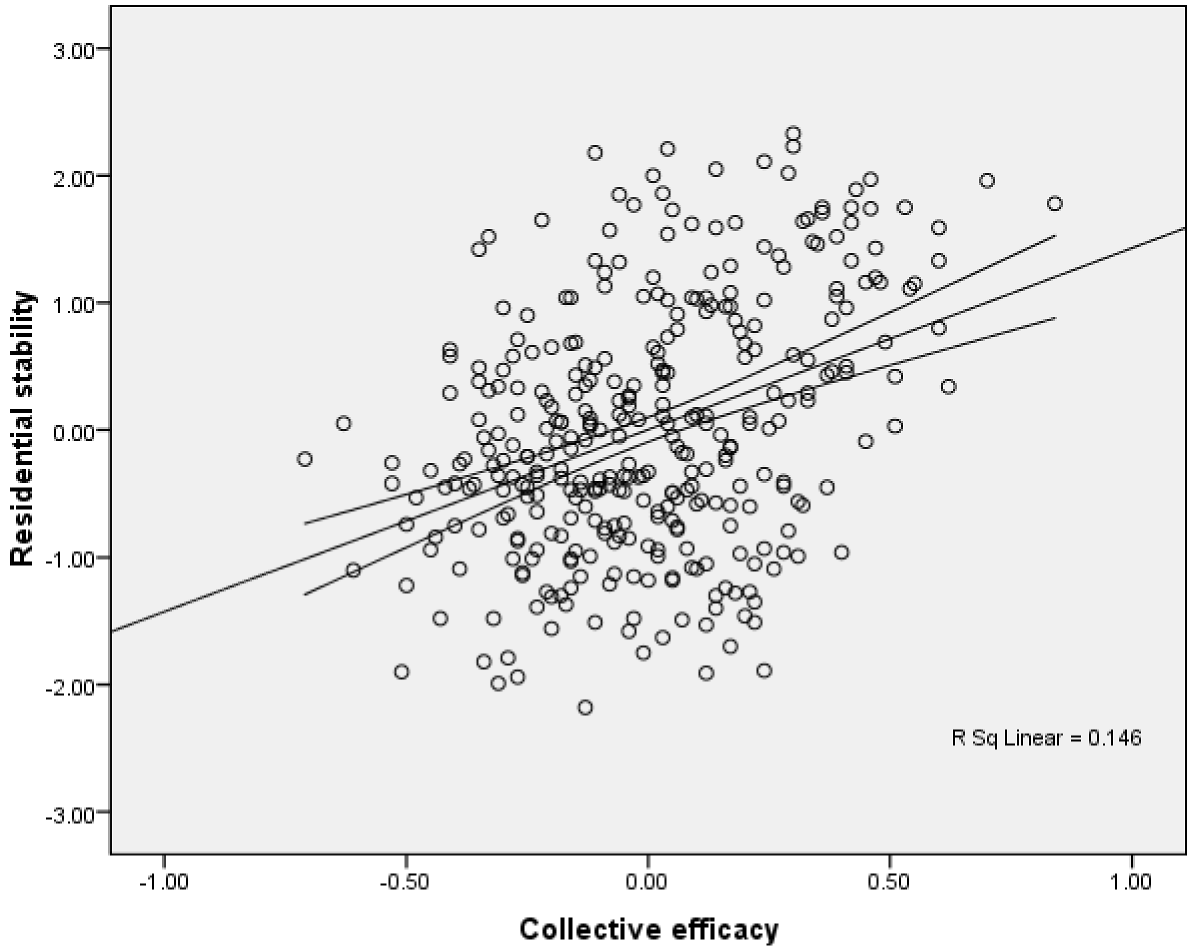
Coefficients[a]

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

ResStab	Coef.	Robust HC3 Std. Err.	t	P>|t|	[95% Conf.	Interval]
CollEff	1.428681	.1696641	8.42	0.000	1.094957	1.762404
_cons	.0024394	.049374	0.05	0.961	-.0946775	.0995563

Human Development in Chicago Neighborhoods (Neighborhoods.sav)



Human Development in Chicago Neighborhoods (Neighborhoods.sav)

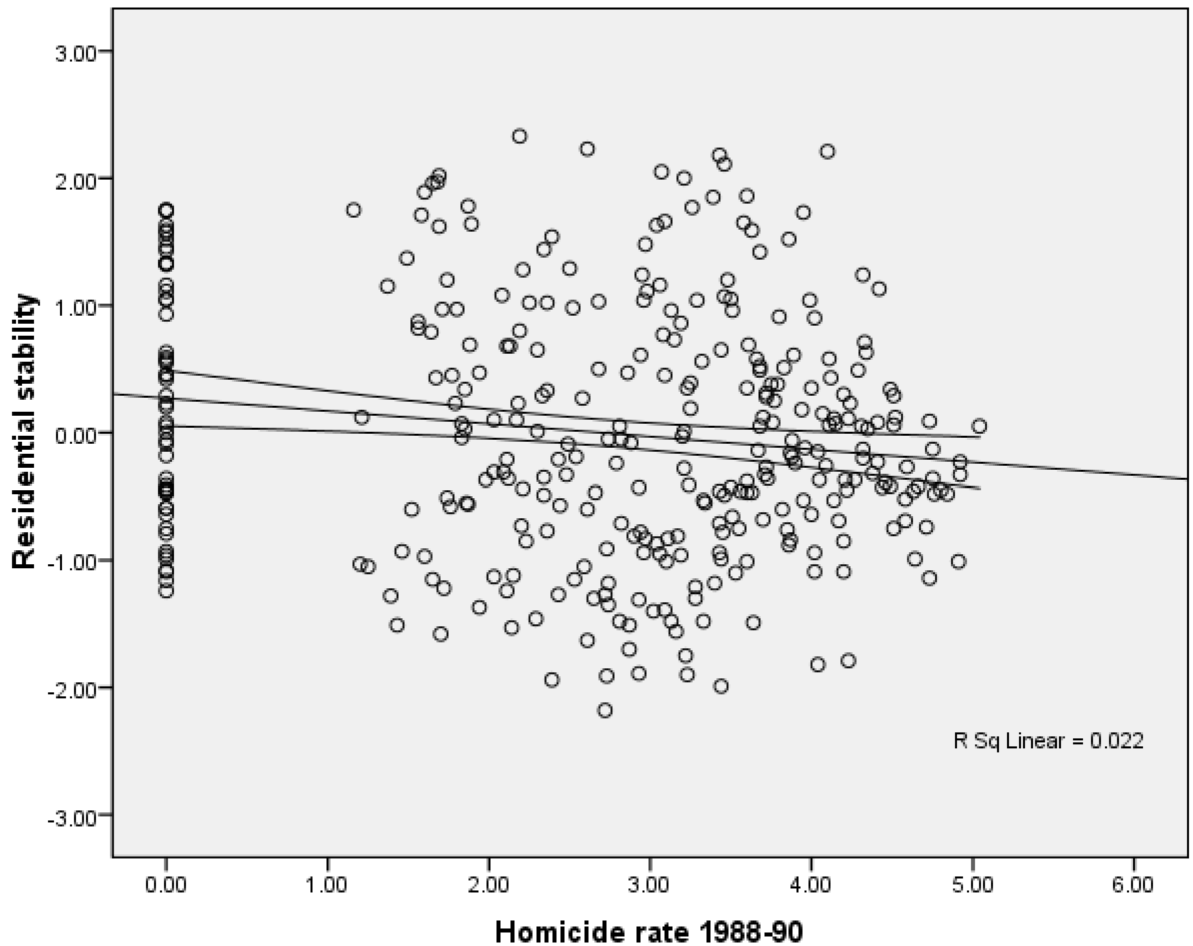
Coefficients[a]

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

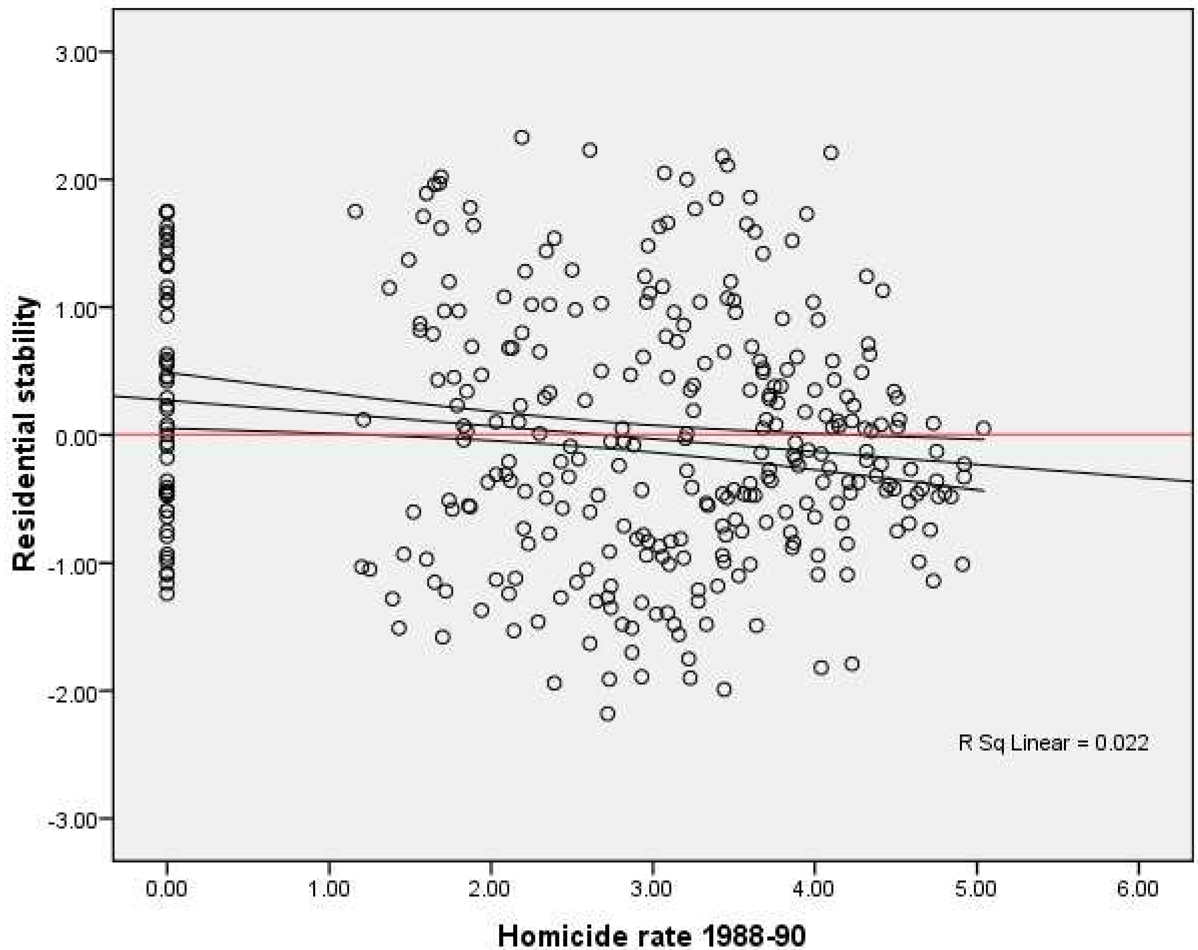
a. Dependent Variable: Residential stability

ResStab	Coef.	Robust HC3 Std. Err.	t	P>|t|	[95% Conf.	Interval]
Homr90	-.1004749	.0329669	-3.05	0.002	-.1653196	-.0356302
_cons	.2704146	.109388	2.47	0.014	.0552522	.4855769

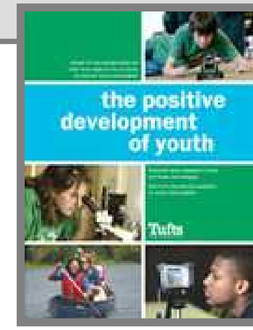
Human Development in Chicago Neighborhoods (Neighborhoods.sav)



Human Development in Chicago Neighborhoods (Neighborhoods.sav)



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



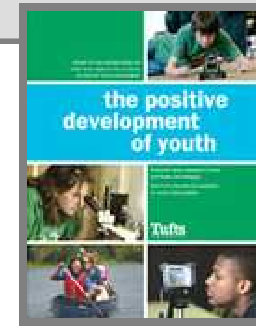
- **4-H Study of Positive Youth Development**
- **Source: Subset of data from IARYD, Tufts University**

- **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)



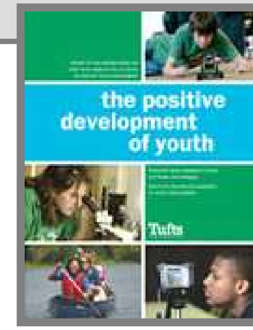
Coefficients[a]

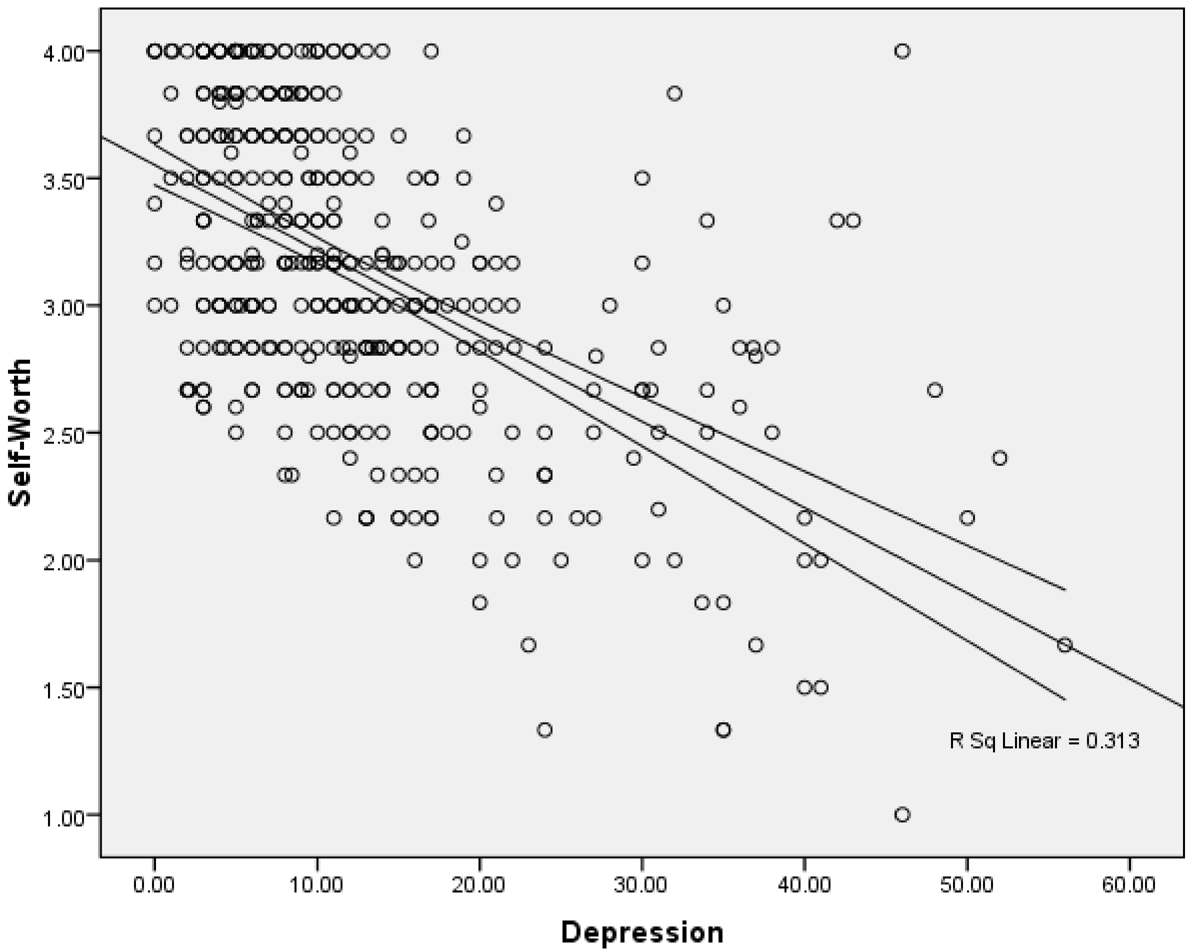
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	3.552	.040		88.146	.000	3.473	3.631
	Depression	-.034	.002	-.559	-13.606	.000	-.038	-.029

a. Dependent Variable: Self-Worth

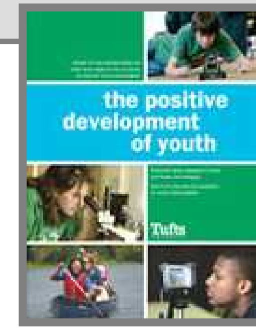
Selfworth	Coef.	Robust HC3 Std. Err.	t	P>|t|	[95% Conf.	Interval]
Depression	-.0336348	.0033104	-10.16	0.000	-.0401423	-.0271272
_cons	3.55208	.04353	81.60	0.000	3.466508	3.637652

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





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



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

SelfWorth	Coef.	Robust HC3 Std. Err.	t	P>|t|	[95% Conf.	Interval]
Depressed	-.6858758	.0634105	-10.82	0.000	-.8105287	-.5612229
_cons	3.307047	.0288516	114.62	0.000	3.25033	3.363764

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