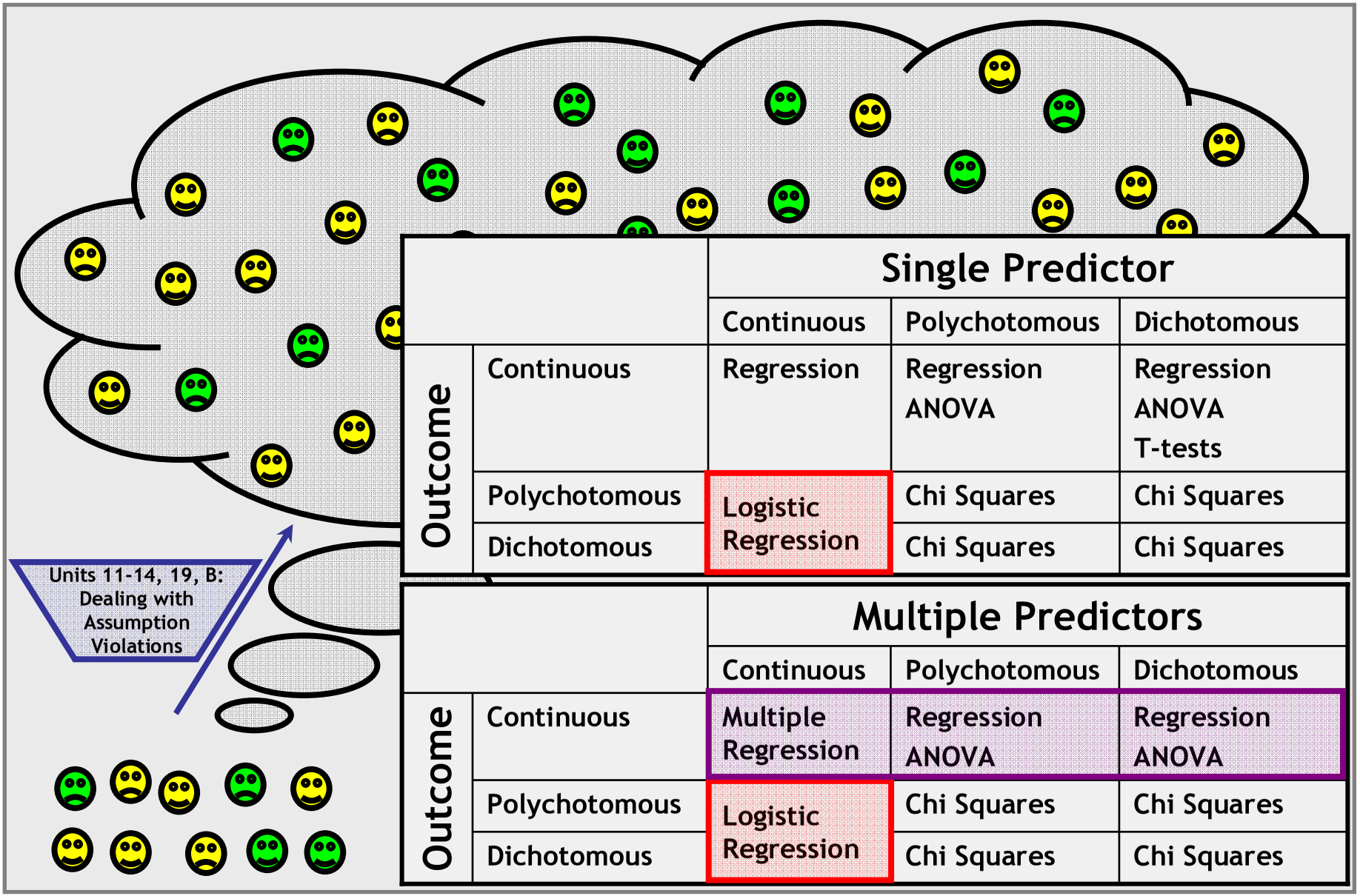
Unit 12: Road Map (VERBAL)

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 12: Road Map (Schematic)



Single Predictor

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

Multiple Predictors

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

Unit 12: Roadmap (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)	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 9

Unit 8

Unit 11

Unit 18

Unit 16

Unit 12

Unit 13

Unit 17

Unit 14

Unit 19

Unit 12: Checking GLM Assumptions with Regression Diagnostics

Unit 12 Post Hole:

Check your GLM assumptions by interpreting a residual-versus-fitted (RVF) plot, a histogram of residuals, a normal probability plot, residual statistics, leverage statistics, and influence statistics.

Unit 12 Technical Memo and School Board Memo:

Use regression diagnostics to evaluate the assumptions of your simple linear regression (from Memo 11).

Unit 12 Review:

Review Units 6, 7 and 8.

Unit 12 Reading:

Meyers et al., Chapters 3a and 3b.

Unit 12: 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 12: 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. 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)
  - 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

NELS88.sav Codebook

National Education Longitudinal Study

Source: U.S. Department of Education

Summary: Here are select variables from the NELS88 data set.

Notes: I created the FREELUNCH variable based on annual family income of less than $25,000. converted the HOMEWORK variable from an ordinal/categorical variable to a continuous variable, which is why it is so "binny." I removed from the data set students who self-identified as other than Asian, Black, Latino, or White. I then created a set of indicator variables from RACE: ASIAN, BLACK AND LATION with WHITE as an (implicit) reference category.

Sample: A nationally representative sample of 7,800 8th graders.

Variables:

READING, a continuous variable, test score, mean = 47 and standard deviation = 9

FREELUNCH, a dichotomous variable, 1 = Eligible for Free/Reduced Lunch and 0 = Not

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

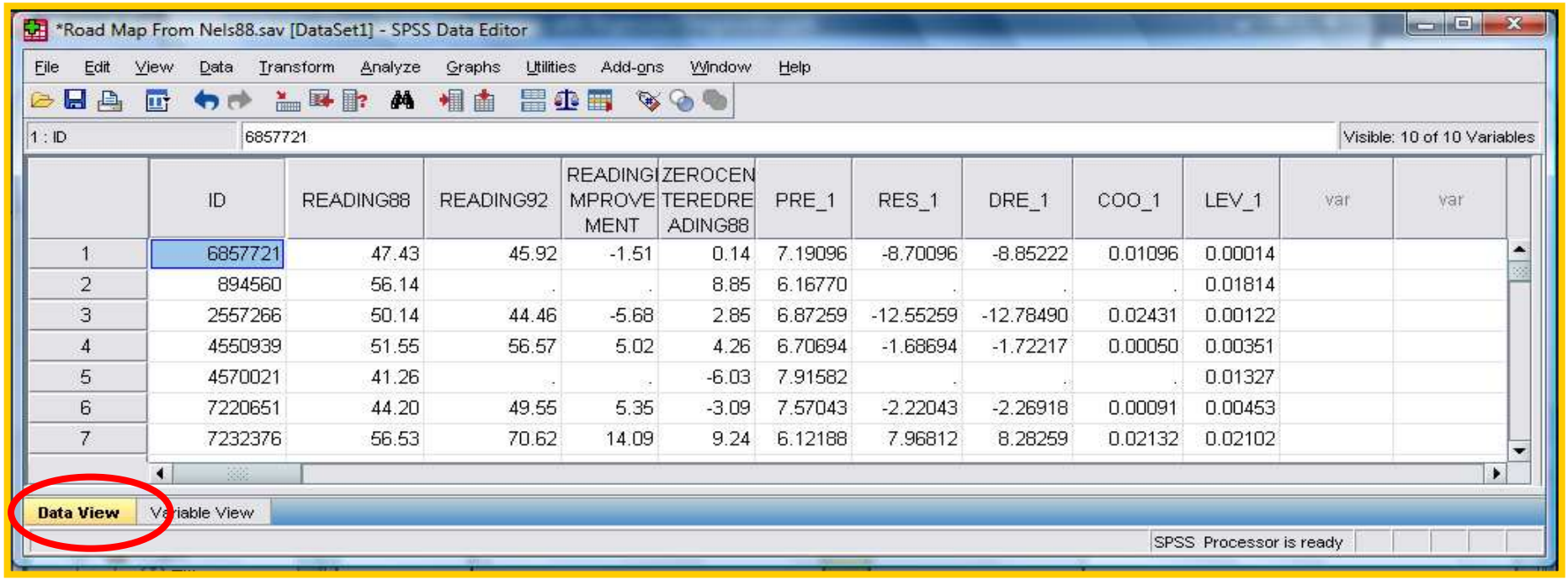
RACE, a polychotomous variable, 1 = Asian, 2 = Latino, 3 = Black and 4 = White

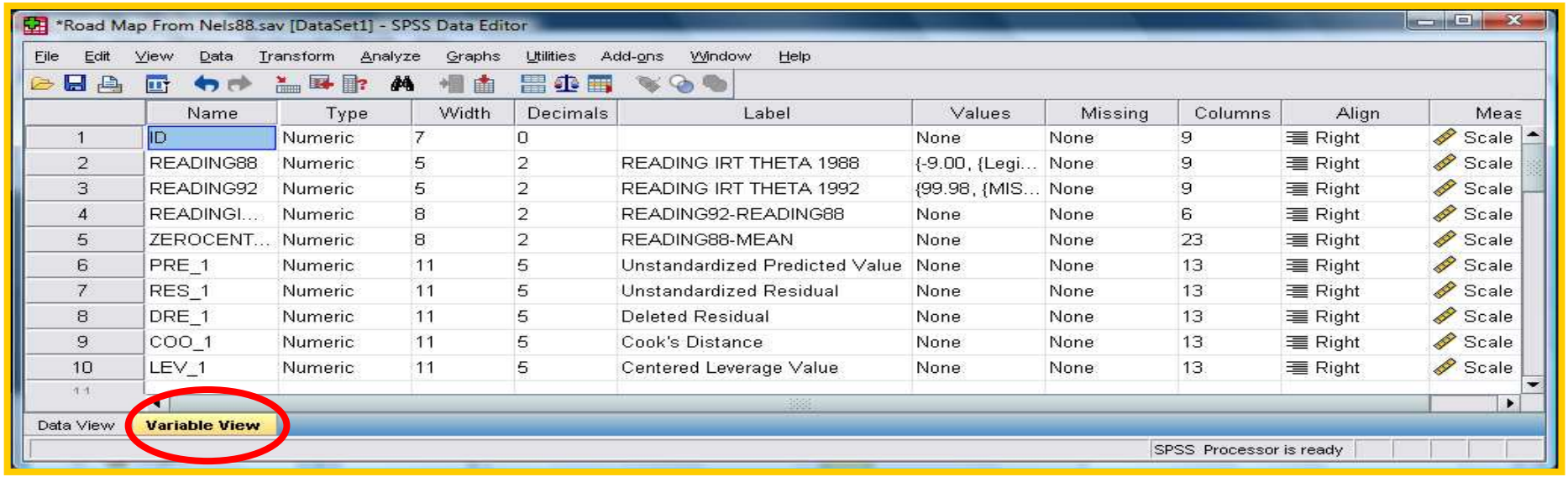
ASIAN, a dichotomous variable, 1 = Self-Identifies as Asian and 0 = Not

LATINO, a dichotomous variable, 1 = Self-Identifies as Latino and 0 = Not

BLACK, a dichotomous variable, 1 = Self-Identifies as Black and 0 = Not

Select Variables from the NELS Data Set



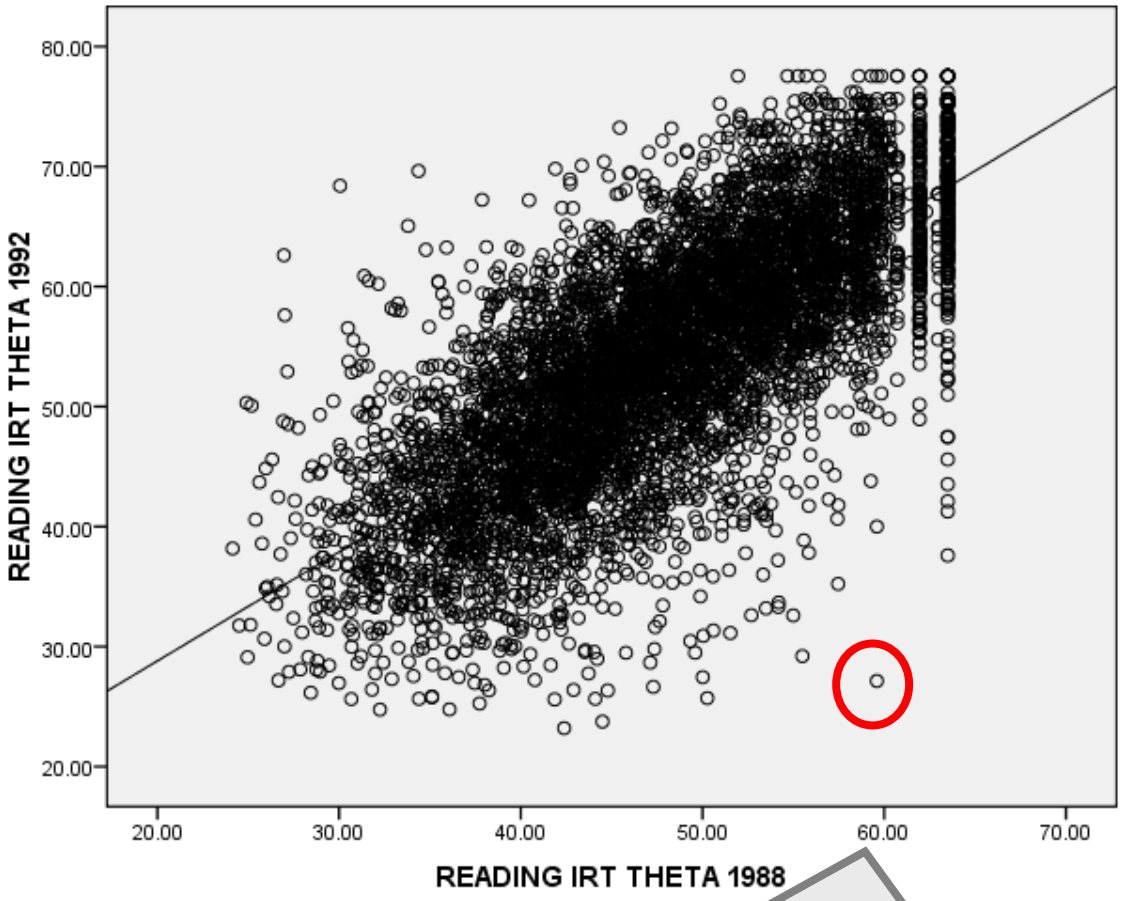


Introduction to Regression Diagnostics

Search HI-N-LO for Assumption Violations

- Heteroskedasticity
- Independence
- Normality
- Linearity
- Outliers

At least in simple linear regression, diagnostics provide information that we could conceivably glean from a bivariate scatterplot of the outome versus predictor; nevertheless they can provide a helpfully detailed view. In multiple regression, however, diagnostics provide information that we could never gather by eye.



Consider this outlier. We do not need any fancy regression diagnostics to see that this is an outlier. Describe the outlier. What will happen to our regression line if we remove the outlier? (You are thinking about the influence of the outlier.)

Setting Up Our Question (Part 1 of 3)

$$\widehat{READING92} = 10.7 + 0.9READING88$$

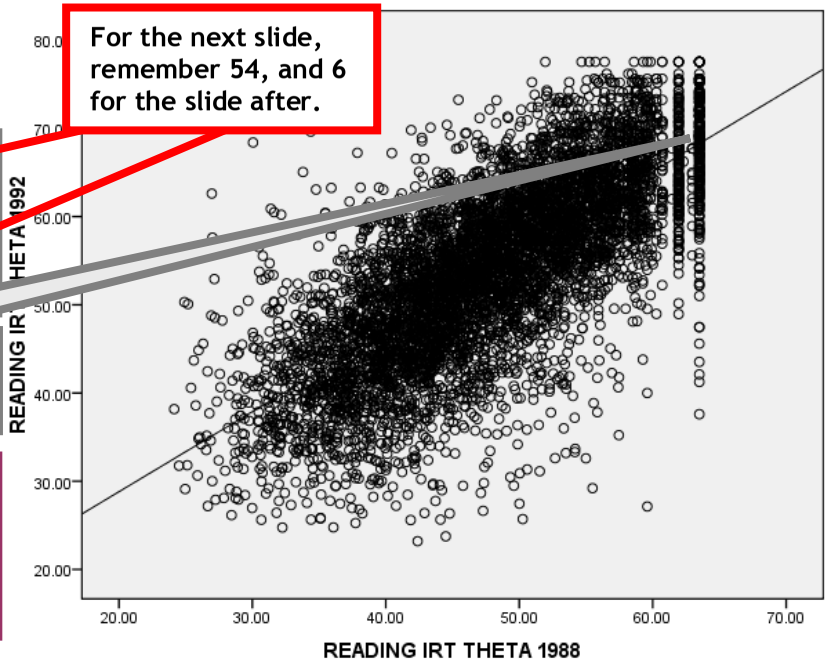
The average student scored 48 points as an 8th grader. How many points do you predict that she improved in the 12th grade?

$54 = 10.7 + 0.9*48$

Do we expect the same increase for a top 8th grader with a score of 69? (Regression To The Mean)

For the next slide, remember 54, and 6 for the slide after.

For the average student, we expect an increase of 6 reading points from the 8th grade to the 12th grade, from 48 points to 54 points.



Notice that I use "increase." The longitudinal data warrant the developmental conclusion!

Recall that the y-intercept (i.e., β_0 , e.g., 10.7) is our prediction when x is zero (e.g., *READING88* = 0). Since READING88 never equals zero within the range of our data, the y-intercept is merely a mathematical abstraction in our fitted model. But, we can make zero more interesting...

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	10.659	.515		20.715	.000	9.650	11.667
	READING IRT THETA 1988	.908	.011	.738	85.984	.000	.887	.928

a. Dependent Variable: READING IRT THETA 1992

Setting Up Our Question (Part 2 of 3)

*Example SPSS syntax for computing transformed variables.

*This linear transformation is not a z-transformation because I did not divide the difference by the standard deviation.

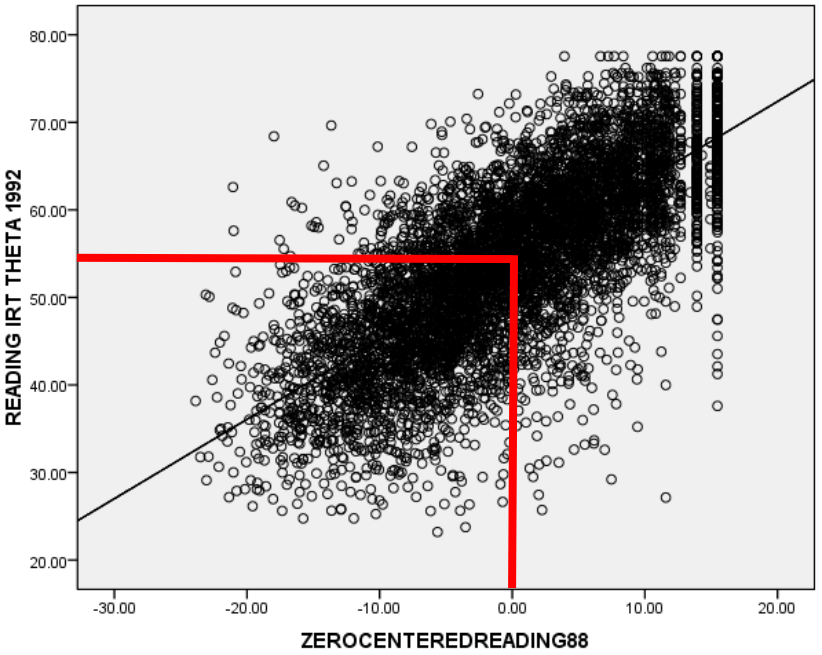
COMPUTE ZEROCENTEREDREADING88 = READING88 - 48.0155.

EXECUTE.

*This (goofy) transformation is non-linear because I do more than add/subtract and/or multiply/divide by a constant. I use powers and logs.

COMPUTE Sean_Is_A_Great_SPSS_Programmer = READING88 * 48.0155 - 1975/27 + FREELUNCH**(1/2) + LN(HOMEWORK+1).

EXECUTE.



$$\hat{READING92} = 54.2 + 0.9 ZEROCENTEREDREADING88$$

Look familiar? The y-intercept now has an interesting interpretation. It is our prediction for the average student now that the average student has an x-value of 0. (Also notice that the slope ***has not*** changed.)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	54.235	.089		609.885	.000	54.060	54.409
	ZEROCENTEREDREADING88	.908	.011	.738	85.984	.000	.887	.928

a. Dependent Variable: READING IRT THETA 1992

Setting Up Our Question (Part 3 of 3)

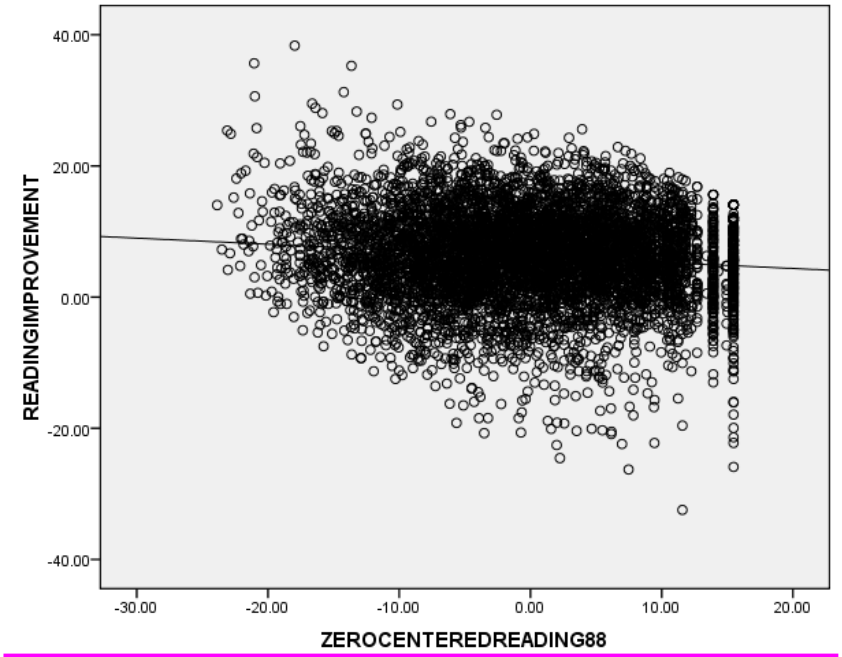
*If we are interested in changes, let's compute a change score and use it as our outcome. This is not a linear transformation because I add/subtract and/or multiply/divide by a variable, not a constant.

COMPUTE READINGIMPROVEMENT = READING92 - READING88.

EXECUTE.

READING92 has measurement error, and READING88 has measurement error, when I take their difference, their difference has more measurement error than either, since they are positively correlated! Ugh.

Whenever there is an element of randomness in the outcome, we expect regression to the mean. Measurement error is one possible source of randomness, but not the only possible source of randomness. If we predict adult height by mother's height, we will get regression to the mean, even though there is only trivial measurement error with height. Why? There is genetic and environmental luck involved in height. To be extremely tall (or short) requires luck, but there is no guarantee that the luck is hereditary. Most extremely tall moms have not-as-tall daughters.*



* But, if variance in the outcome is greater than variance in the predictor, there can be egression *from* the mean!

Look familiar? The y-intercept now has another interesting interpretation. It is our predicted change for the average student now that the average student has an x-value of 0 and our outcome variable is a change variable. The slope now tells us the difference in change associated with a I point difference in 1988 reading score. If we take two students who differed by 10 points in 1988, we expect the higher scoring student to have improved her score less, by about 1 point less.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	6.219	.089		69.936	.000	6.045	6.393
	ZEROCENTEREDREADING88	-.092	.011	-.111	-8.760	.000	-.113	-.072

a. Dependent Variable: READINGIMPROVEMENT

Unit 12: Research Question I

Theory: Some students acquire reading strategies that allow them to improve their reading skills at a faster than typical rate. If we can detect those students, we can use qualitative research methods to understand their reading strategies and perhaps bring them to scale by teaching them as part of the curriculum.

Research Question: Which students improve their reading through high school more than we would expect based on their 8th grade reading?

Data Set: Random Subsample (n = 96) From The National Education Longitudinal Study (NELS88.sav)

Variables:

- Outcome—8th-12th Grade Change in Reading Scores (*READINGIMPROVEMENT*)
- Predictor—8th Grade Reading Score (*ZEROCENTEREDREADING88*)

Model:

$$READINGIMPROVEMENT = \beta_0 + \beta_1 ZEROCENTEREDREADING88 + \varepsilon$$

Unit 12: Research Question II

Theory: Reading growth is exponential such that if there is a reading gap between two students (or two groups of students) in the 8th grade, the gap will be greater in the 12th grade. Furthermore, the strength of this relationship will be greater for top readers than bottom readers, because bottom readers have such variable access to reading support.

Research Question: Is the relationship between reading improvement and 8th grade reading non-linear, in particular exponential? Is it heteroskedastic, with more variance in improvement for low scorers?

Data Set: Random Subsample (n = 96) From The National Education Longitudinal Study (NELS88.sav)

Variables:

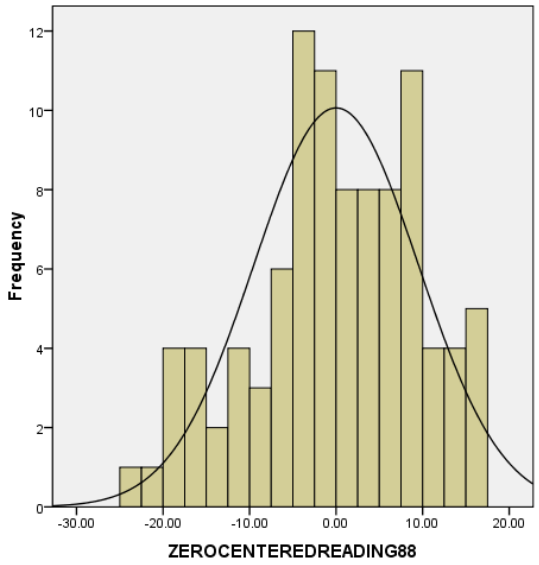
Outcome—8th-12th Grade Change in Reading Scores (*READINGIMPROVEMENT*)

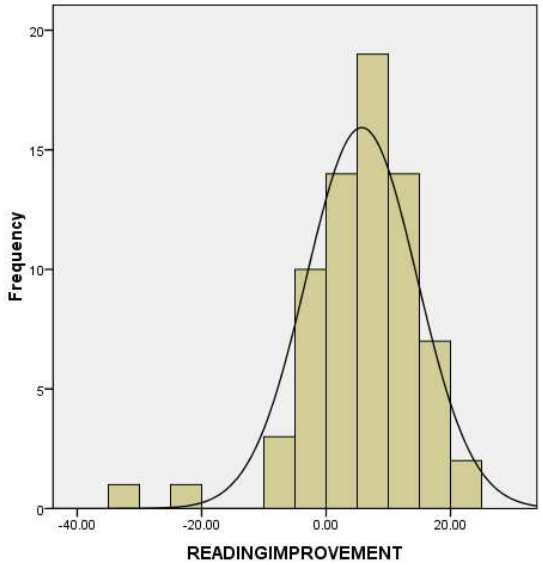
Predictor—8th Grade Reading Score (*ZEROCENTEREDREADING88*)

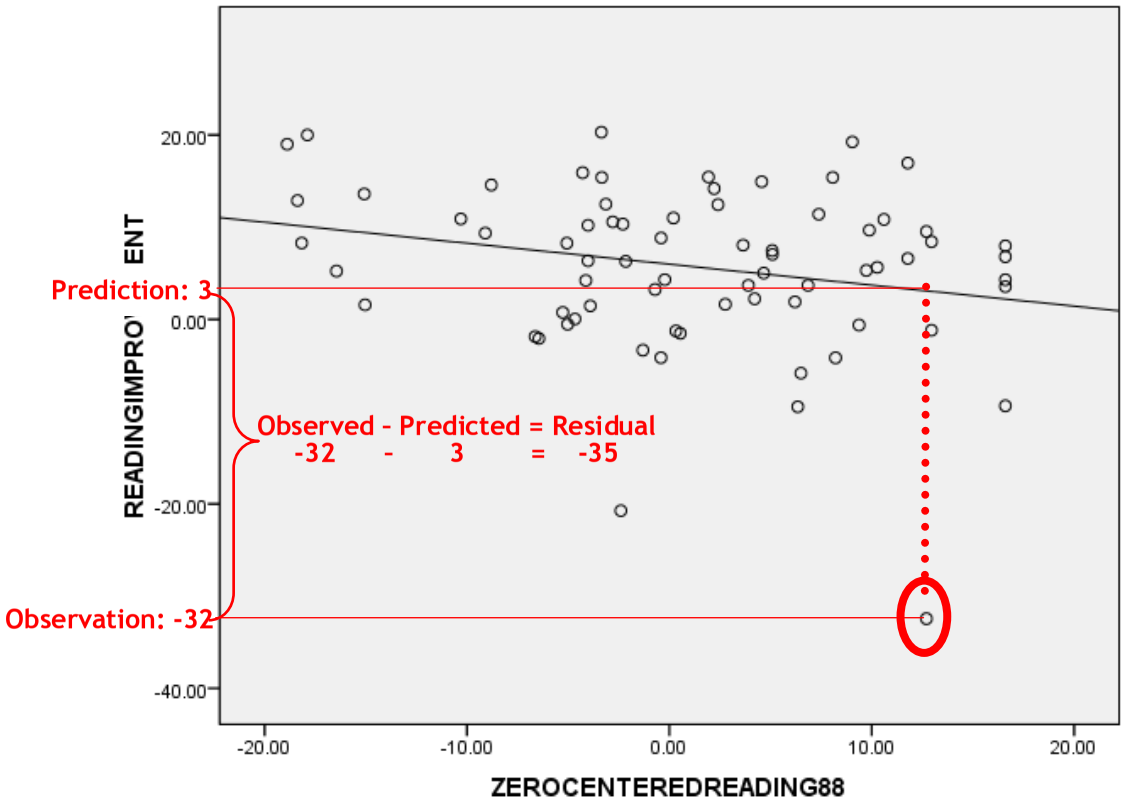
Model:

$$READINGIMPROVEMENT = \beta_0 + \beta_1 ZEROCENTEREDREADING88 + \varepsilon$$

Exploratory Graphs



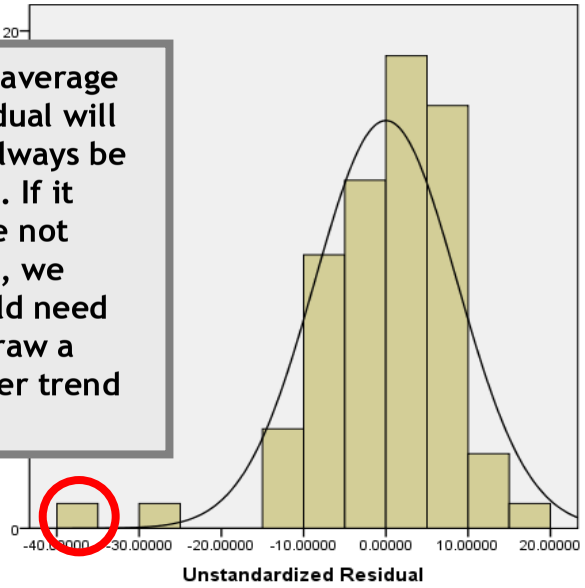




A residual (aka error) is the difference between our observed outcome and our predicted outcome. If the residual is negative that means we should have predicted lower (i.e., we overpredicted). If the residual is positive, we should have predicted higher (i.e., we underpredicted). Of course, we expect residuals because of individual variation, hidden variables, and measurement error.

Residuals

Every datum has an associated residual, and we can graph the residuals with a histogram:



The average residual will be always be zero. If it were not zero, we would need to draw a better trend line.

What would happen to our trend line if we removed the outlier with a residual of -35? You can think of every datum as pulling the line with a rubber band. What happens when our outlier lets go of its rubber band?

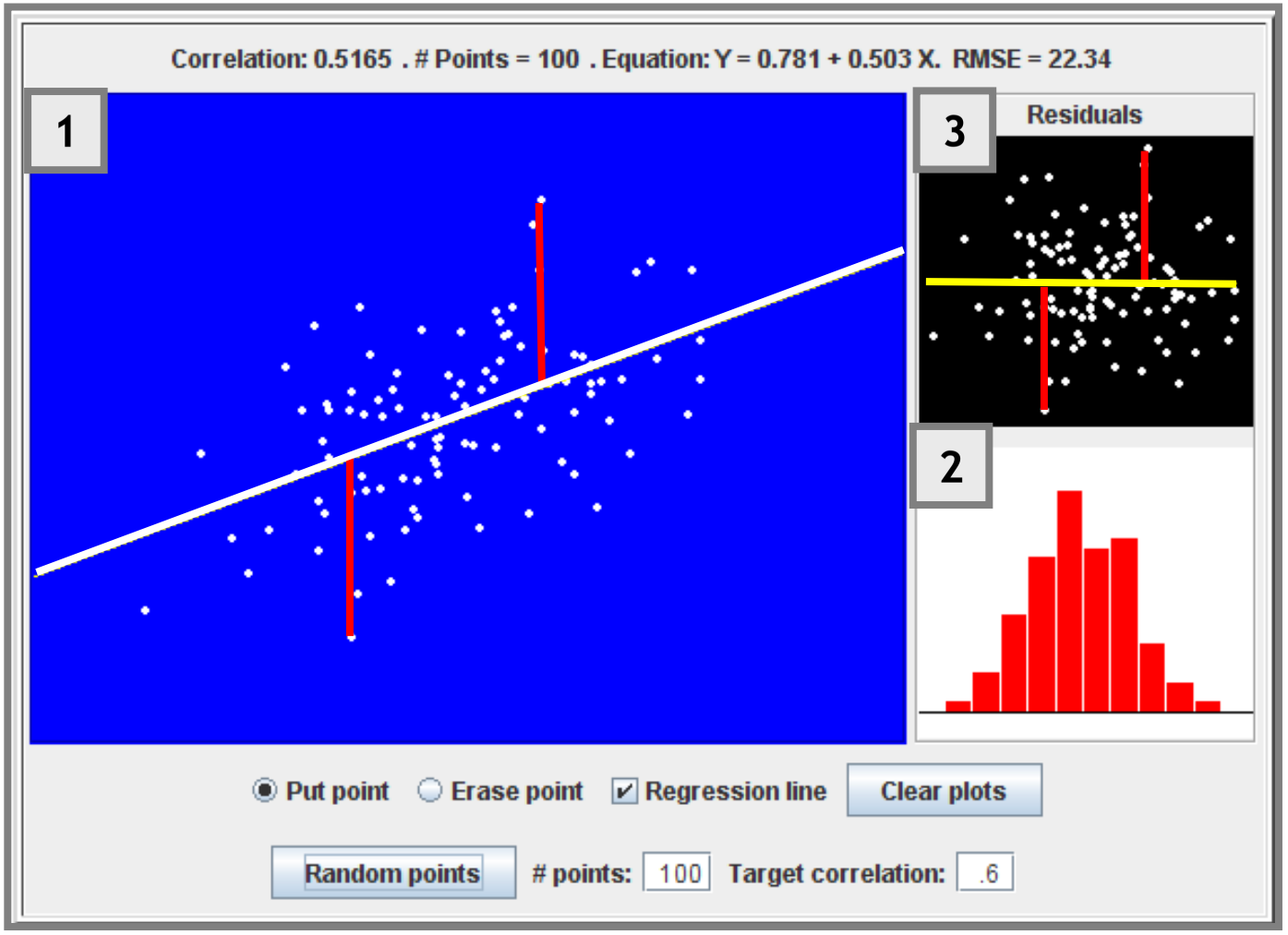
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	5.994	1.041		5.759	.000	3.918	8.071
	ZEROCENTEREDREADING88	-.227	.113	-.236	-2.021	.047	-.452	-.003

a. Dependent Variable: READINGIMPROVEMENT

Playing Around For A Few Minutes

Expanding our View of The Scatterplot

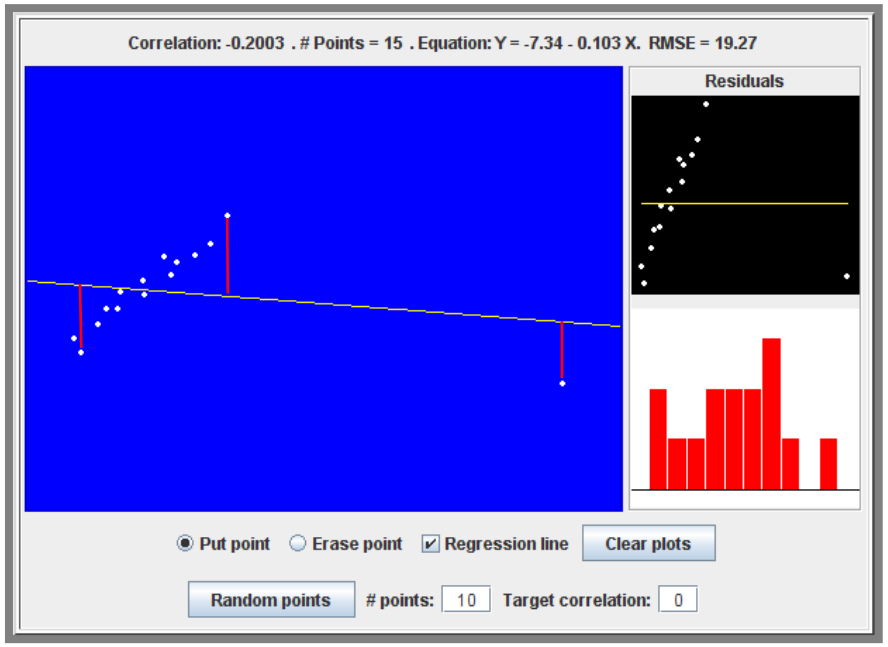
1. Scatterplot
2. Residual Histogram
3. Residual Vs. Fitted Plot (RVF Plot)



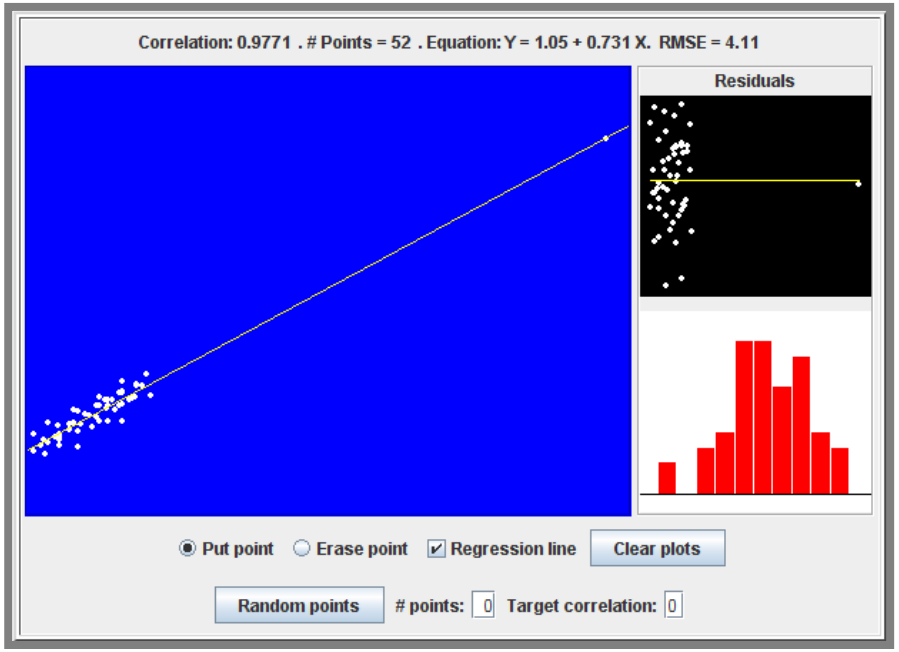
http://www.istics.net/stat/PutPoints/

Playing Around For A Few Minutes

Extreme Example 1:
The part/whole problem solved by deleted residuals.

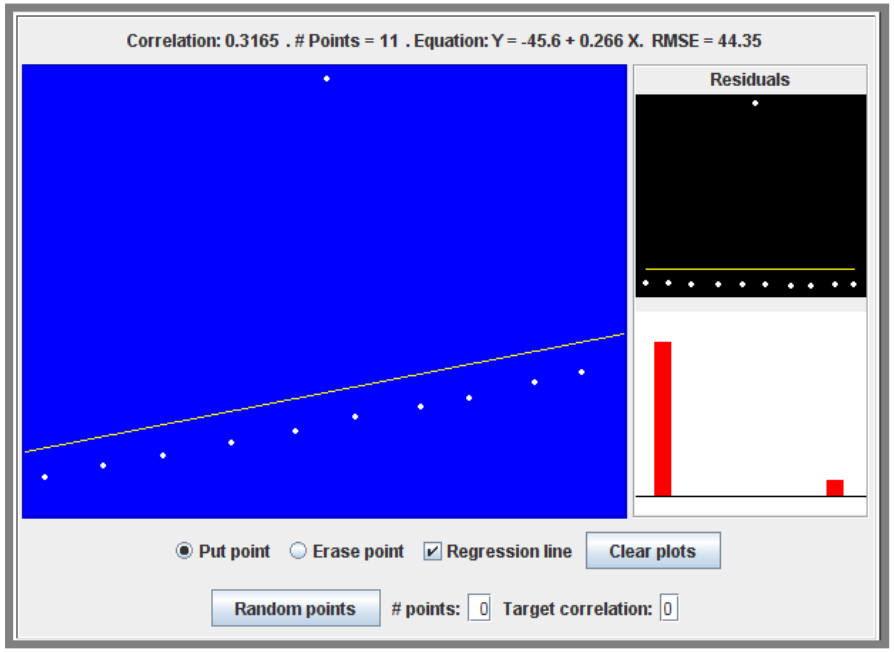


Extreme Example 2:
High leverage is not necessarily high influence.

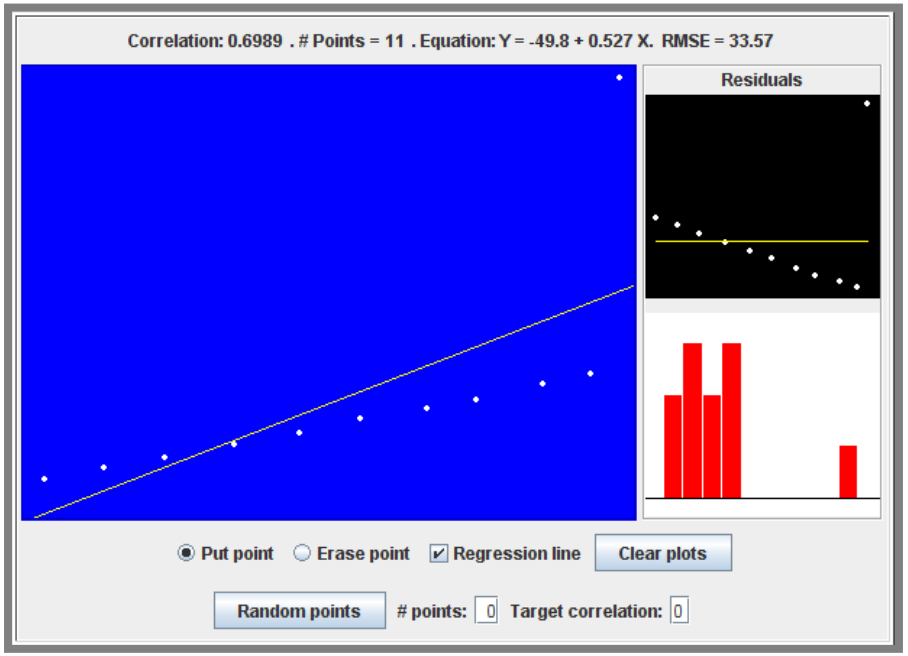


Playing Around For A Few Minutes

Extreme Example 3:
Low leverage high residuals influence the y-intercept.

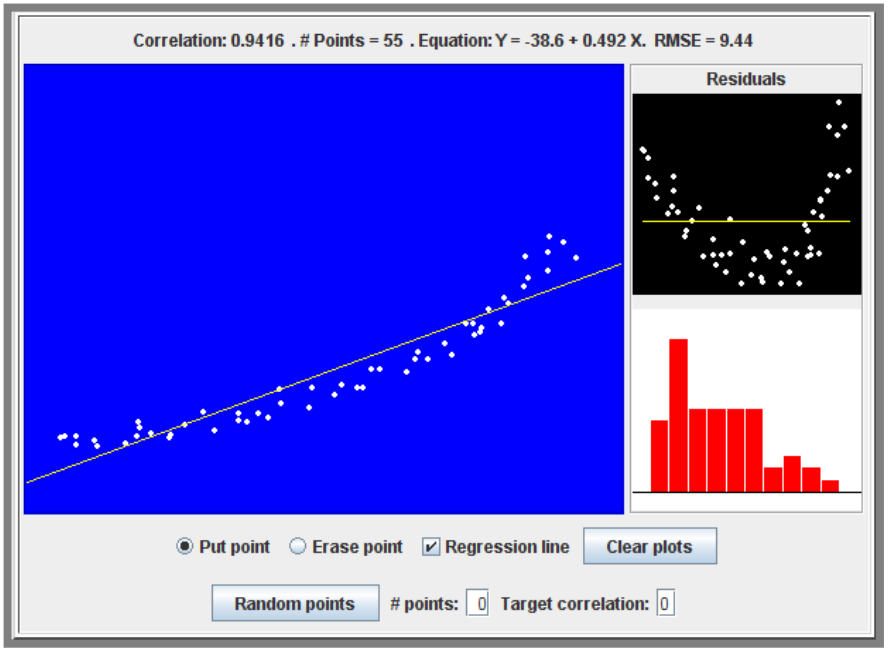


Extreme Example 4:
High leverage high residuals influence the slope.

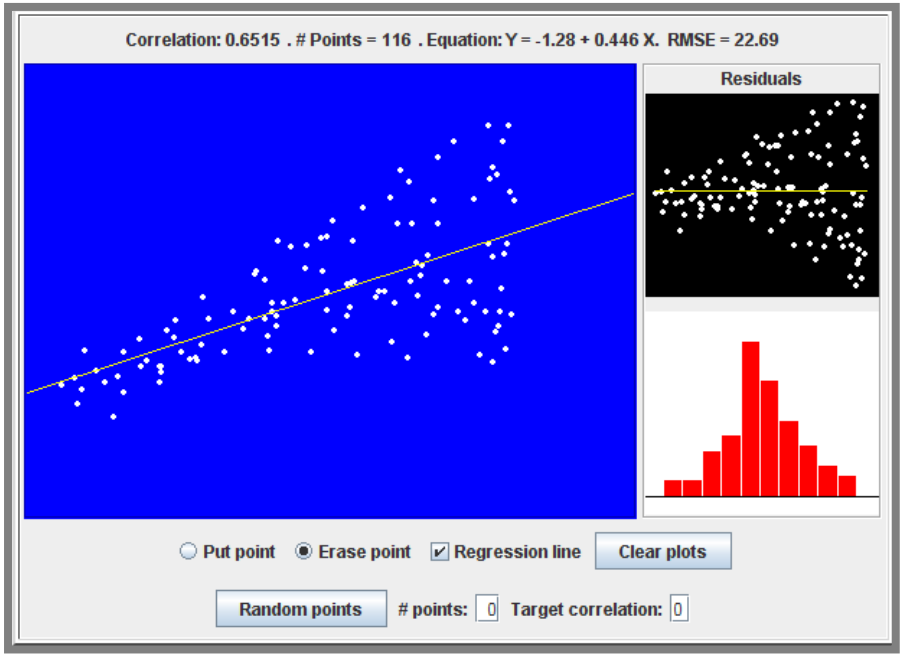


Playing Around For A Few Minutes

Extreme Example 5:
RVF Plots blow up non-linear horseshoes.

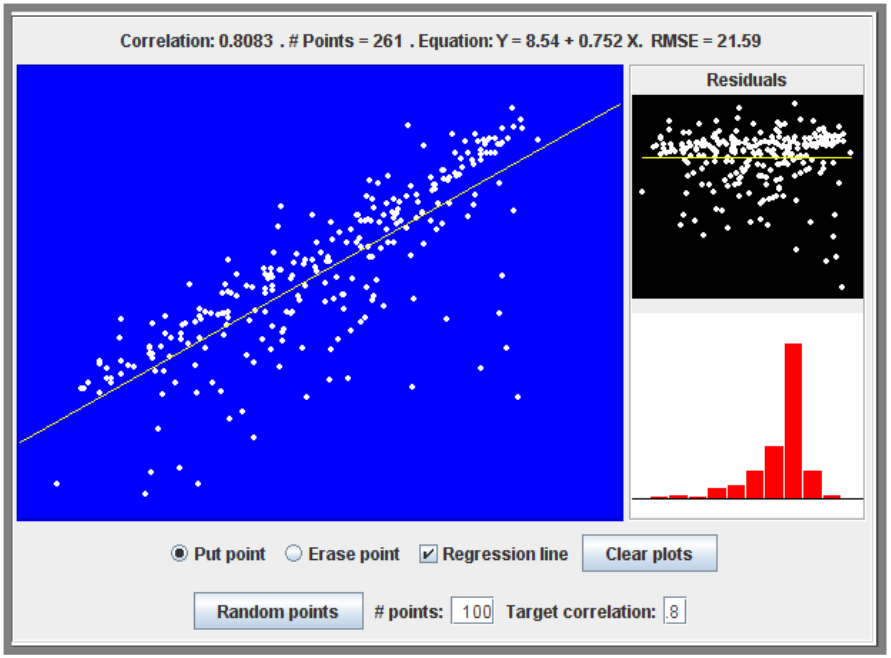


Extreme Example 6:
RVF Plots blow up heteroskedastic funnels.

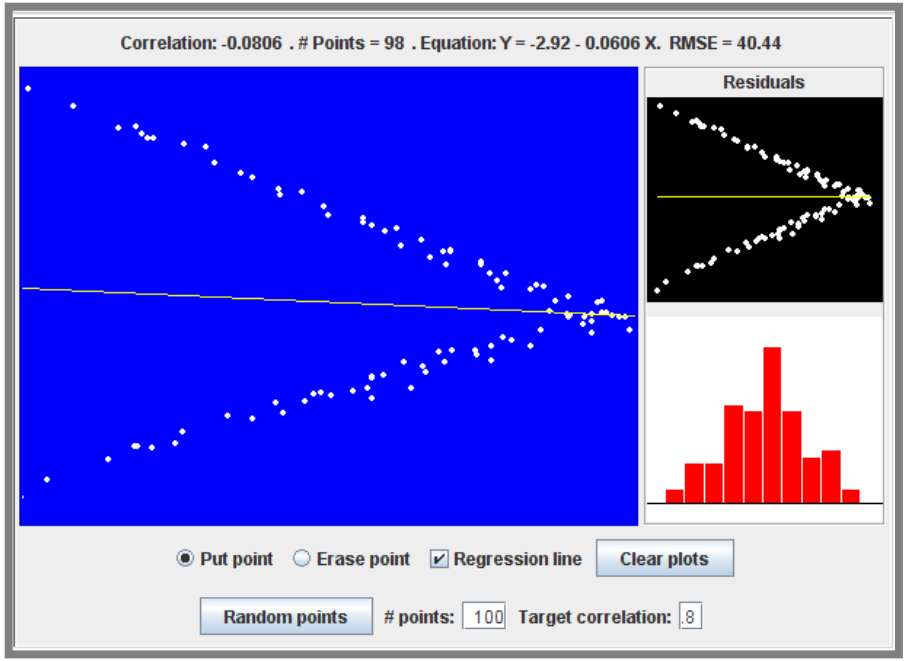


Playing Around For A Few Minutes

Extreme Example 7:
Residual histograms provide insight into normality.

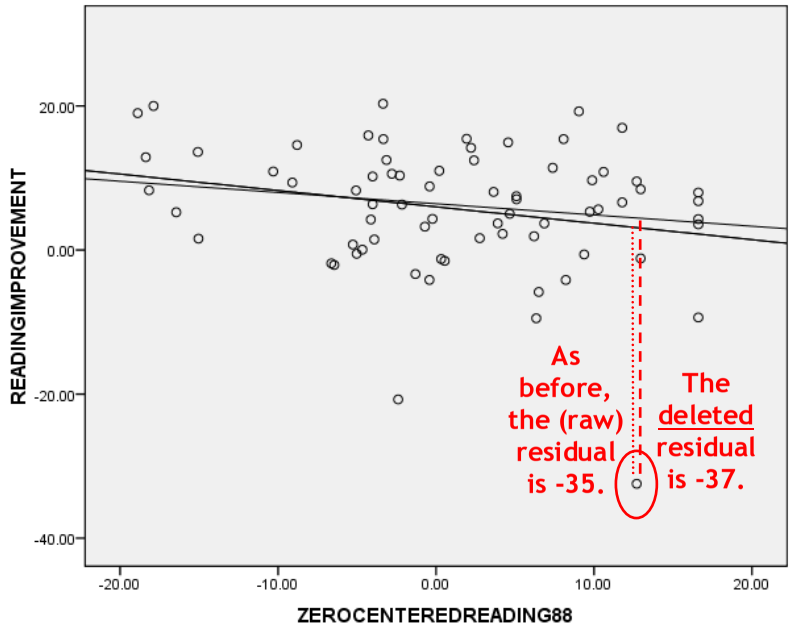


Extreme Example 8:
Residual histograms don't show <u>conditional</u> normality.



Outlier Detection: Deleted Residuals (Part 1 of 3)

```
* Identify the residual, temporarily remove it, and refit the line.
TEMPORARY.
SELECT IF NOT (ID = 2999973).
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT READINGIMPROVEMENT
  /METHOD=ENTER ZEROCENTEREDREADING88.
```



Original Regression (n = 71):

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	5.994	1.041		5.759	.000	3.918	8.071
	ZEROCENTEREDREADING88	-.227	.113	-.236	-2.021	.047	-.452	-.003

a. Dependent Variable: READINGIMPROVEMENT

Regression With Outlier Removed (n = 70):

Notice that the slope is no longer stat sig!

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	6.430	.912		7.051	.000	4.610	8.250
	ZEROCENTEREDREADING88	-.156	.099	-.188	-1.575	.120	-.354	.042

a. Dependent Variable: READINGIMPROVEMENT

Outlier Detection: Deleted Residuals (Part 2 or 3)

Create a residual vs. fitted plot (i.e., a residual vs. predicted plot).

Create a histogram of residuals and a normal probability plot.

Create five new variables:

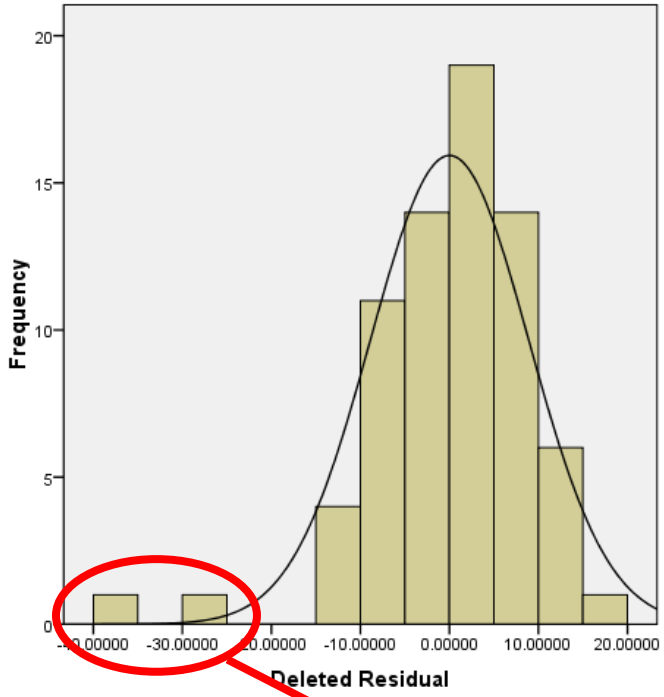
- *PRE_#*: A predicted/fitted value for each observation.
- *RES_#*: A residual for each observation.
- *DRE_#*: A deleted residual for each observation.
- *LEV_#*: A leverage statistic for each observation.
- *COO_#*: An influence statistic (Cook's D) for each obs.

```
* We do not have calculate deleted residual "by hand," we can have the computer do it automatically for every case, and, along the way, we can have the computer do a whole bunch of other things.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT READINGIMPROVEMENT
  /METHOD=ENTER ZEROCENTEREDREADING88
  /SCATTERPLOT=(*RESID ,*PRED)
  /RESIDUALS HIST(RESID) NORM(RESID)
  /SAVE PRED RESID DRESID LEVER COOK.
```

```
* Once we produce our variables, we can examine them.
EXAMINE VARIABLES=DRE_1 LEV_1 COO_1
  /COMPARE GROUP
  /STATISTICS DESCRIPTIVES EXTREME
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

```
GRAPH
  /HISTOGRAM (NORMAL)=DRE_1.
GRAPH
  /HISTOGRAM=LEV_1.
GRAPH
  /HISTOGRAM=COO_1.
```

Outlier Detection: Deleted Residuals (Part 3 of 3)



Extreme Values

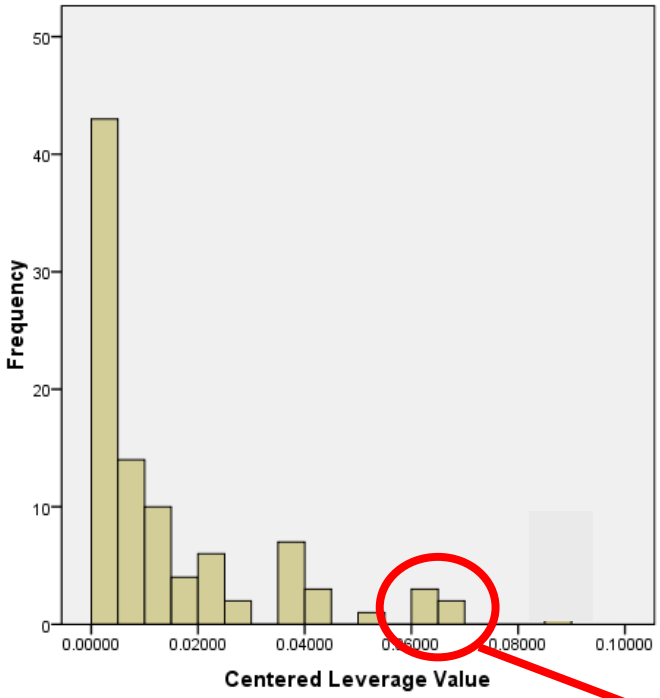
			Case Number	Value
Deleted Residual	Highest	1	37	15.69550
		2	87	14.10943
		3	47	13.79350
		4	61	11.49472
		5	16	10.74174
	Lowest	1	27	-36.89485
		2	8	-27.72992
		3	53	-14.28699
		4	91	-12.24847
		5	6	-10.54415

Two obvious outliers.

Who are they?

A deleted residual is a residual based on subtracting the predicted value from the observed value, just like a typical, raw residual, except that the predicted value is calculated with the observation removed in order to avoid the part/whole problem in which we are looking for outliers from the trend but the outlier is part of the trend.

Outlier Detection: The Leverage Statistic



Extreme Values

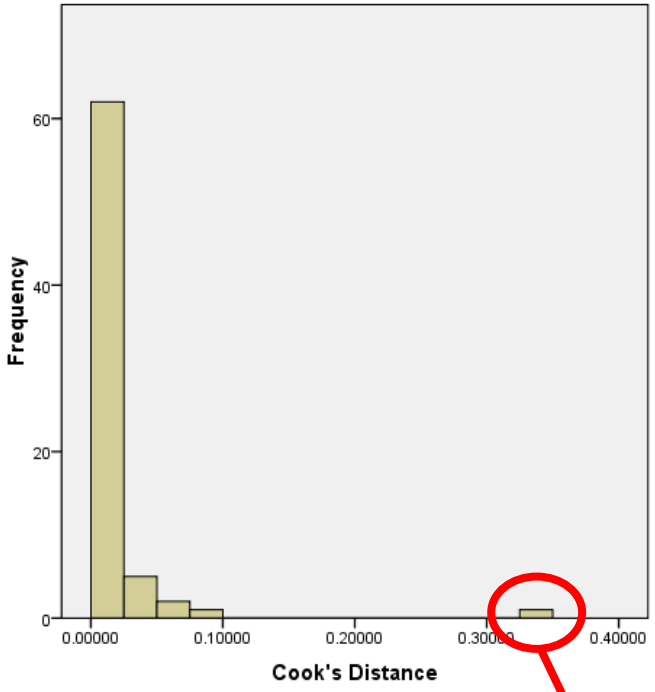
Centered Leverage Value	Highest	1	24	.06740
		2	84	.06401
		3	31	.06265
		4	16	.06085
		5	18	.05200
	Lowest	1	58	.00007
		2	22	.00009
		3	33	.00012
		4	56	.00016
		5	85	.00018

Some high leverage observations.

Who are they?

A leverage statistic is a measure of the extremity of an observation based on the value(s) of its predictor(s). When we have one predictor, we can easily see who is extreme on that predictor, but when we have 12 predictors, it can be impossible to see who is generally extreme on all predictors.

Outlier Detection: The Cook's D Statistic



Extreme Values

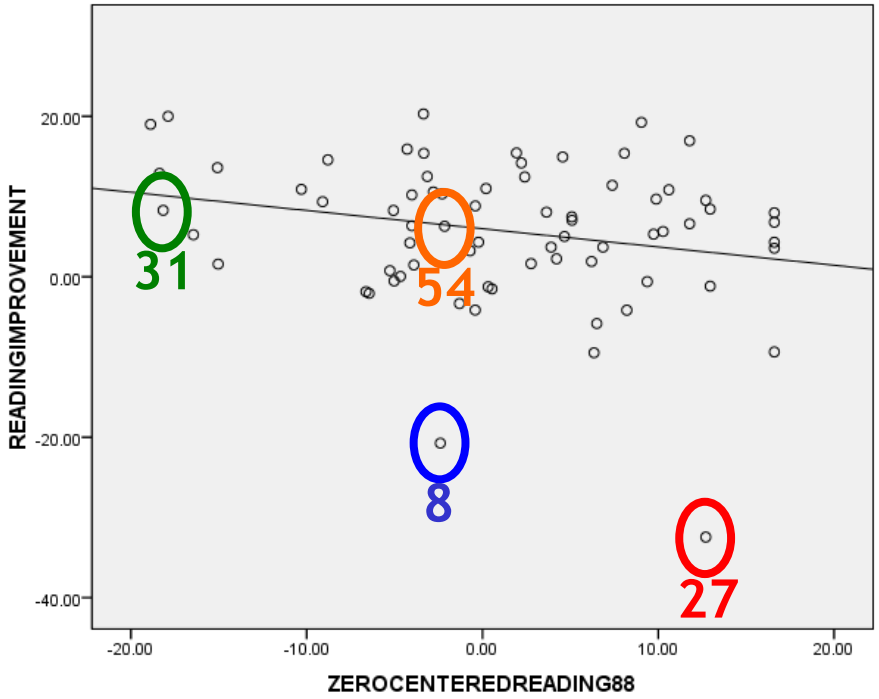
Cook's Distance	Highest	1	27	.32647
		2	8	.08243
		3	16	.05712
		4	91	.05341
		5	24	.04840
	Lowest	1	32	.00000
		2	75	.00000
		3	65	.00004
		4	60	.00007
		5	74	.00018

A high influence observation.

Who is it?

An influence statistic compares the trend line (calculated from all the data, including the observation) with a hypothetical trend line (calculated from all the data except the observation). The bigger the difference between the two trend lines, the greater the influence. Cook's D statistic is the influence statistic that we will use, but there are others.

Outlier Detection: Residuals, Leverage, Influence

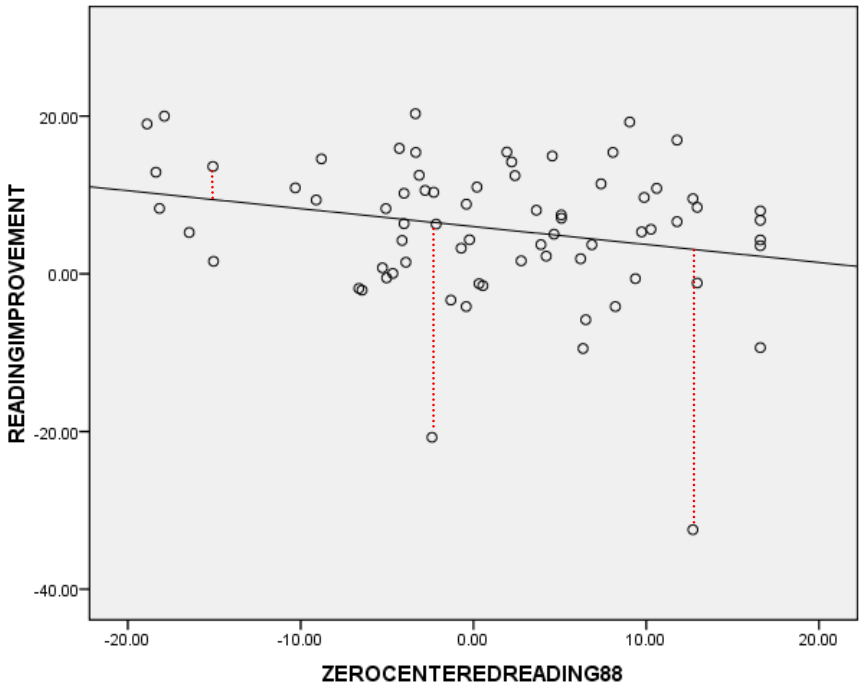


Case Number	Deleted Residual	Leverage	Cook's Distance	Result
8	Extreme	Minimal	Moderate	Extreme in Y, Not in X: Influence Y-Intercept
27	Extreme	Extreme	Extreme	Extreme in Y And in X: Influence Slope
31	Minimal	Extreme	Minimal	Not Extreme in Y, But in X: Little Influence
54	Minimal	Minimal	Minimal	Neither Extreme in Y Nor in X: Little Influence

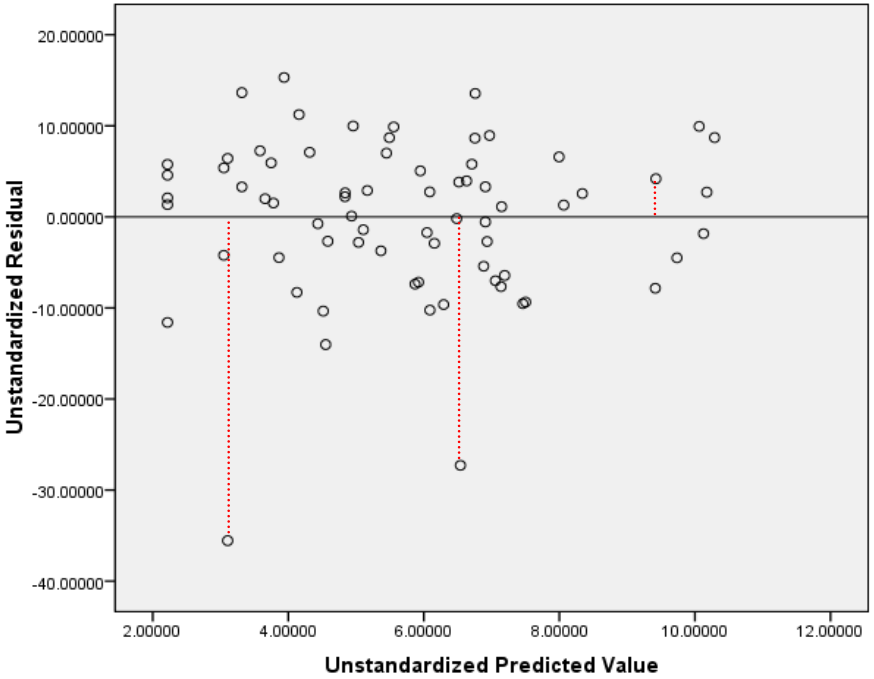
Non-Linearity Detection: RVF Plot

A residual versus fitted plot (RVF plot), also known as a residual versus predicted plot, is just what it says it is: a scatterplot of residual values versus fitted/predicted values.

Good Old Outcome Vs. Predictor Scatterplot:



Shiny New Residual Vs. Fitted Scatterplot:

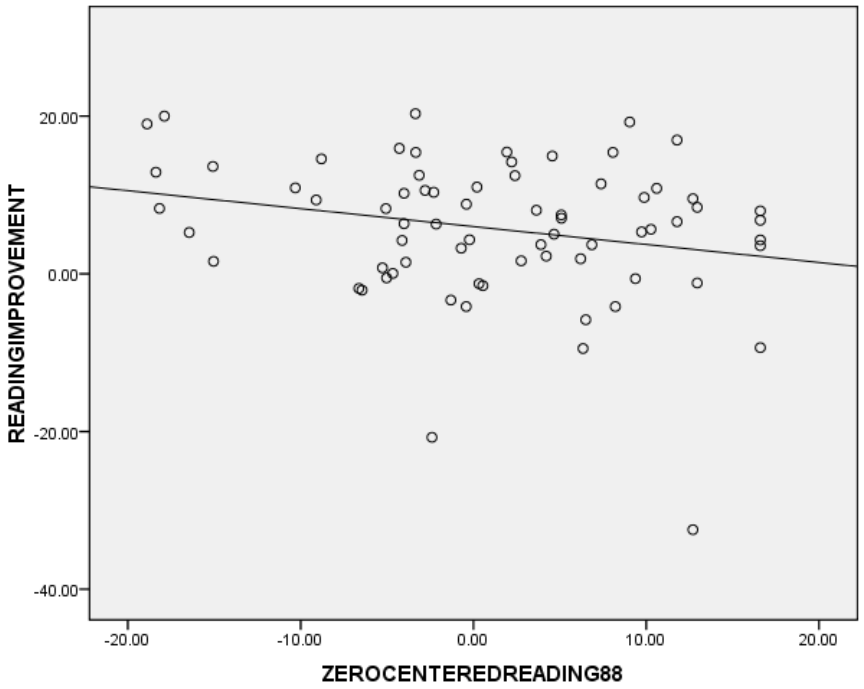


Horseshoe shapes indicate non-linearity. If there were a horseshoe shape in our outcome versus predictor plot, it would be magnified in the residual versus fitted plot, but everything looks okay here. In Unit 13, we'll see examples of non-linear relationships (and attendant horseshoes). If you are wondering "what's the big deal?" wait until we have 7 predictors. No matter how many predictors, we will still have only one predicted value and only one fitted value for each observation, so we can still use an RVF plot for the multiple regression model, whereas we would need not a two dimensional scatterplot of the outcome versus predictors but an 8 dimensional scatterplot!

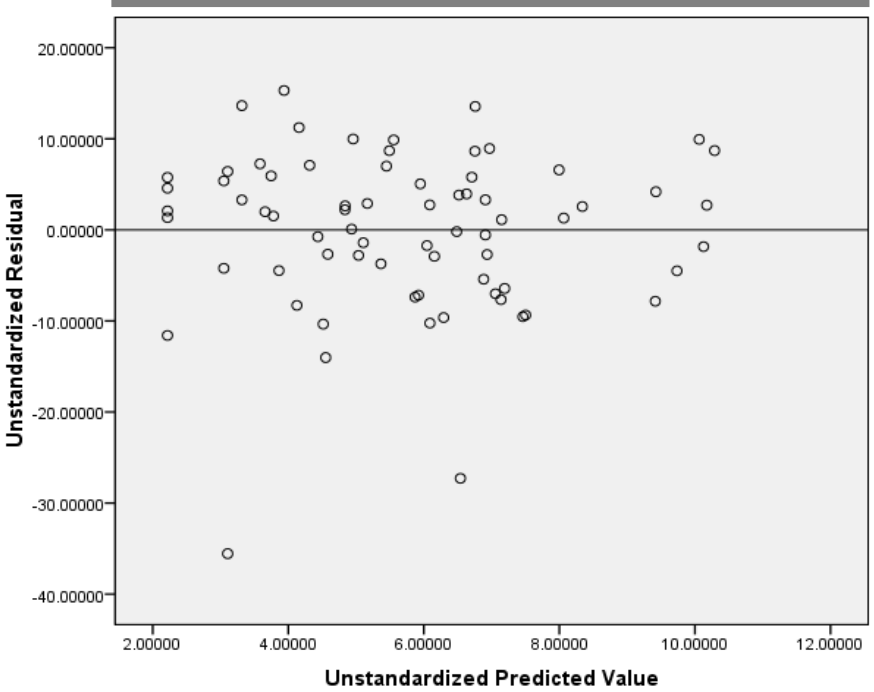
Heteroskedasticity Detection: RVF Plot

A residual versus fitted plot (RVF plot), also known as a residual versus predicted plot, is just what it says it is: a scatterplot of residual values versus fitted/predicted values.

Good Old Outcome Vs. Predictor Scatterplot:

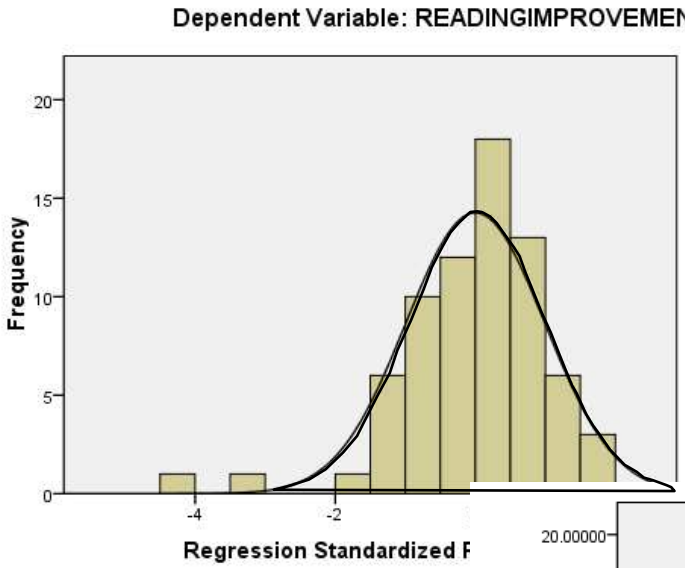


Shiny New Residual Vs. Fitted Scatterplot:

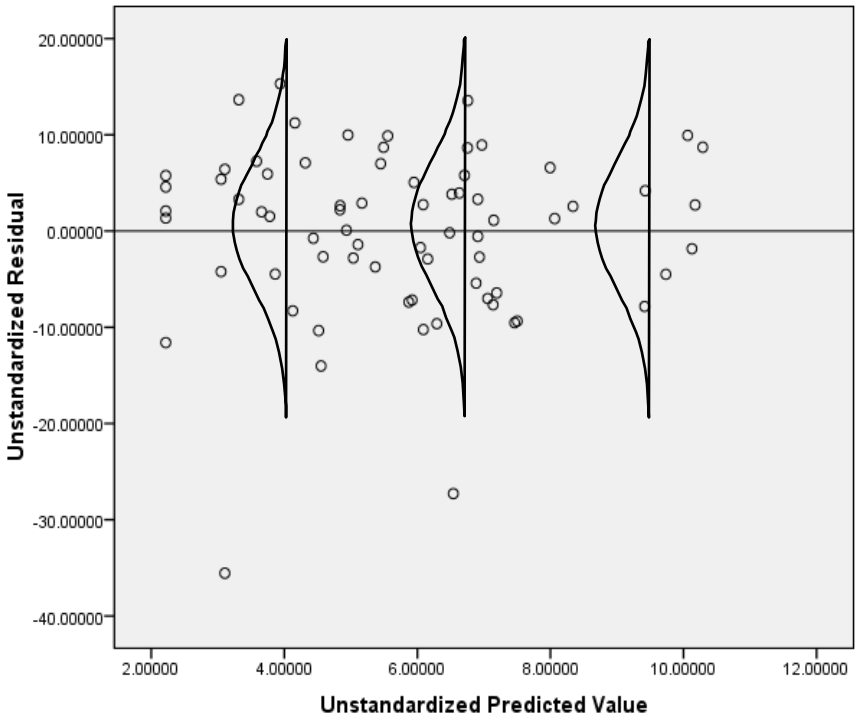


Funnel shapes indicate heterskedasticity. If there were a funnel shape in our outcome versus predictor plot, it would be magnified in the residual versus fitted plot, but everything looks okay here. In Unit 14, we'll see examples of heteroskedastic relationships (and attendant funnels).

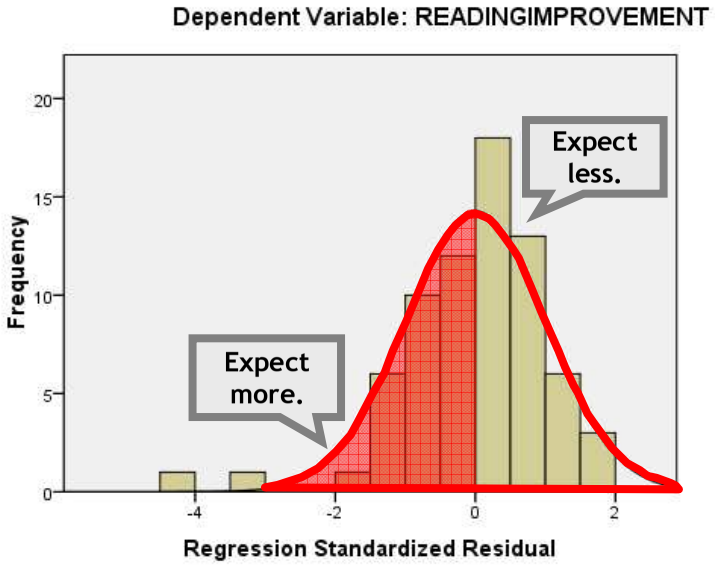
Non-Normality Detection: Residual Histograms



A histogram of residuals can give an indication whether or not the residuals are normally distributed; however, use with caution, because histograms of residuals show an unconditional distribution (i.e., they don't think vertically). We are ultimately concerned with normality (and homoskedasticity) conditional on X. Nevertheless, such histograms can be useful, especially when supplemented with an RVF plot which allows you to think in terms of vertical slices and consequently think about conditional distributions.

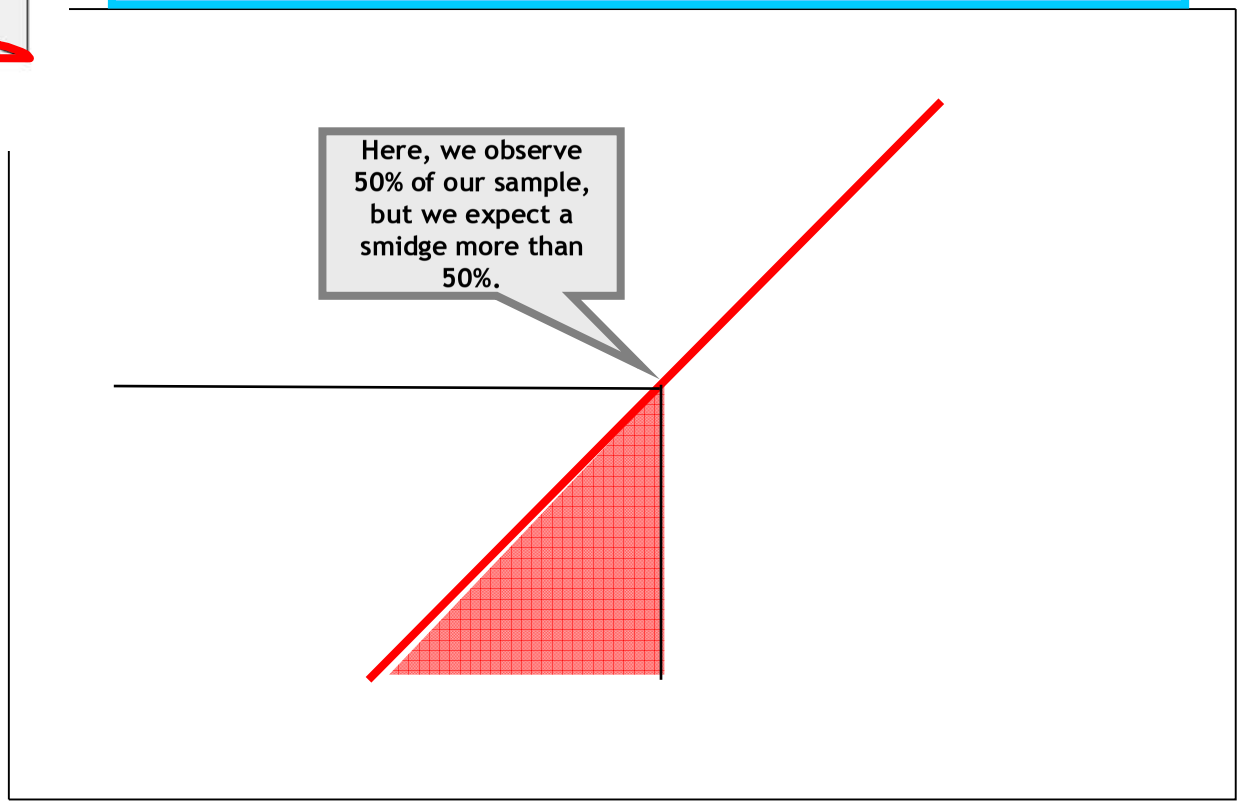


Non-Normality Detection: P-P Plots



A probability-probability plot (P-P plot) is another way of looking at a residual histogram, with a focus on normality. In a normal distribution we expect 50% of the observations to be below average, and, because it's a mathematical construct, we observe 50% of the observations to be below average. This simple truth forms our baseline of comparison (the red line below). In a sample distribution from a population with a normal distribution, we expect 50% of the observations to be below average, but due to sampling error (or perhaps due to a non-normal population distribution, we may observe more or fewer than 50% of observations to be below average.

The take-home message for P-P plots is that we want the dotted line to lie on top of the straight line, and where the dotted line deviates, we have non-normality in our sample, which may indicate non-normality in our population.



Reflecting on our Unit 12 Research Questions

Q1: Which students improve their reading through high school more than we would expect based on their 8th grade reading?

Q2: Is the relationship between reading improvement and 8th grade reading non-linear, in particular exponential? Is it heteroskedastic, with more variance in improvement for low scorers?

To answer the first question, we can sort our data by residuals and find the largest positive residuals:

*ReducedNELS Unit 12.sav [DataSet1] - SPSS Data Editor

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

1 : RES_1 | 15.312068006371483 | Visible: 25 of 25 Variables

	ID	READING 88	READING 92	ZEROCENTEREDREADING88	READINGIMPROVE...	PRE_1	RES	LEV_1	var	var	var	var
1	4574650	55.93	75.18	9.04	19.25	3.93793	15.31	0.01034				
2	7831166	58.67	75.63	11.78	16.96	3.31472	13.64	0.01881				
3	4686977	43.53	63.84	-3.36	20.31	6.75831	13.55	0.00345				
4	6884340	54.96	70.36	8.07	15.40	4.15856	11.24	0.00795				
5	7274602	51.45	66.39	4.56	14.94	4.95691	9.98	0.00191				
6	1389803	29.00	49.00	-17.89	20.00	10.06315	9.93	0.06085				
7	2527062	48.82	64.26	1.93	15.44	5.55510	9.88490	0.00009				

Cut
Copy
Paste
Clear
Insert Variable
Sort Ascending
Sort Descending

Data View | Variable View

SPSS Processor is ready

To answer our second question, we see from our RVF plot that the relationship appears linear (no horseshoe) and homoskedastic (no funnel).

Checking Regression Assumptions With Regression Diagnostics

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

Homoscedasticity: Use an RVF Plot to look for funnels.

Independence

Normality: Use an RVF Plot Residual Histogram and P-P Plot.

Linearity: Use an RVF plot to look for horseshoes.

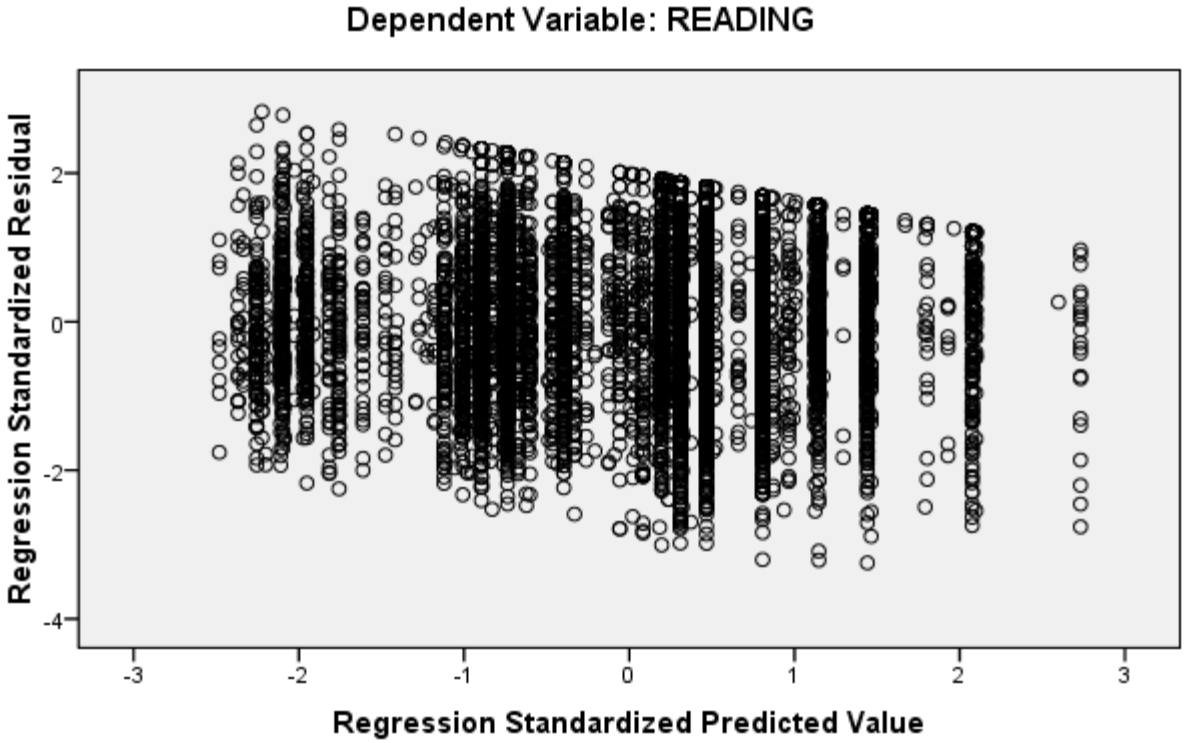
Outliers: Use deleted residuals, leverage and influence statistics.

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.

Answering our Roadmap Question

Unit 12: What tools can we use to detect assumption violations (e.g, outliers)?

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

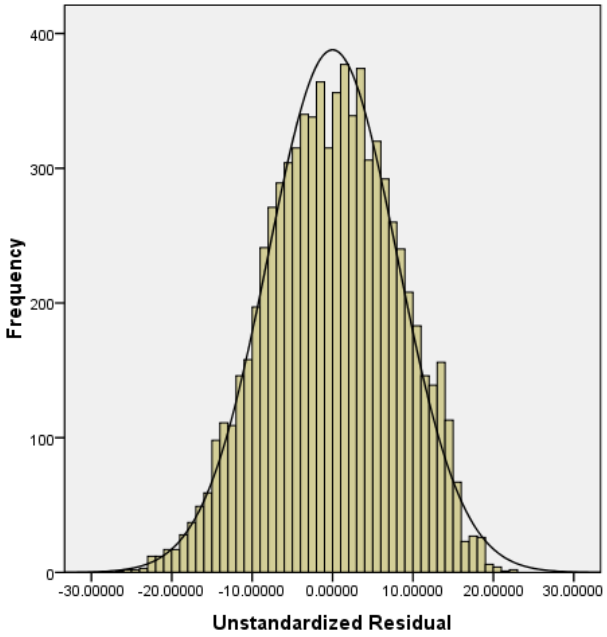


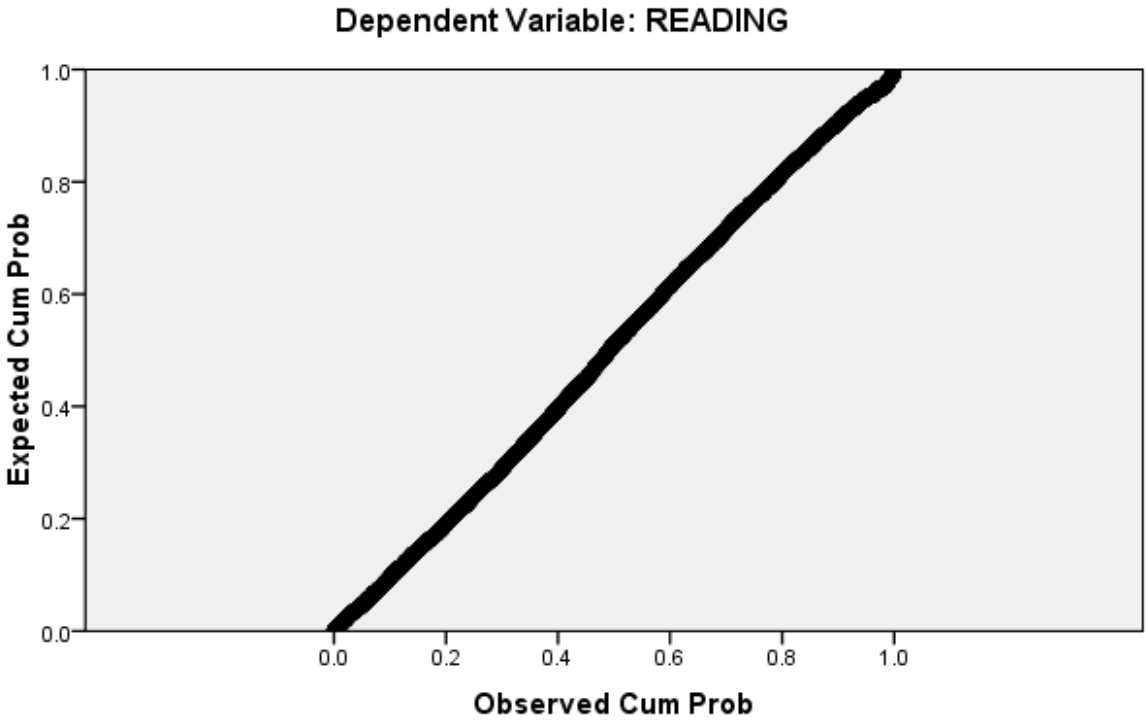
From the RVF plot, we do not appear to have a problem with meeting the linearity assumption. However, due to a ceiling effect, the homoskedasticity and normality assumptions are questionably met. Other than the ceiling effect, the conditional variances appear roughly equal. We are concerned that the high-end predictions are negatively skewed because of the ceiling effect

Use the RVF plot to look for homoskedasticity, normality, linearity and outliers.

Looking at Normality

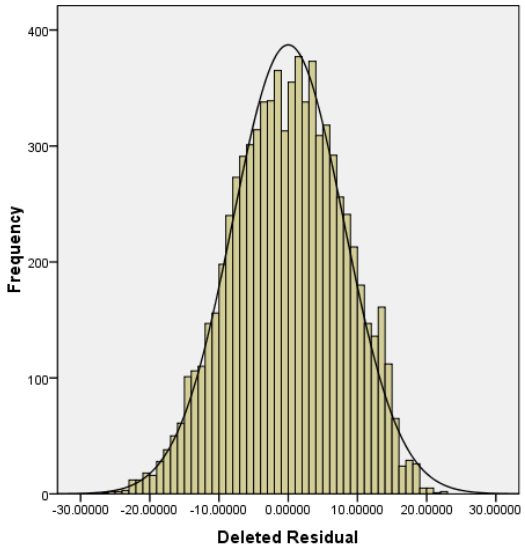
Unit 12: What tools can we use to detect assumption violations (e.g, outliers)?





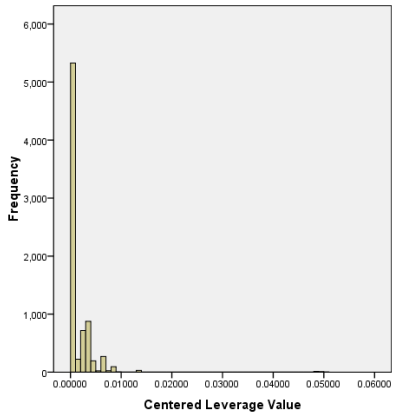
From a histogram of residuals and P-P plot, we see a slight negative skew of the residuals that we attribute to the ceiling effect of our reading measure.

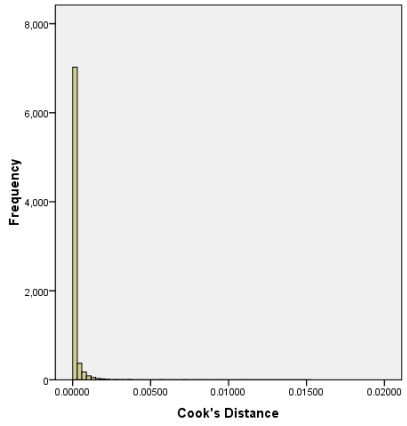
Looking for Outliers



There are no outliers of concern, in part because the large sample size minimizes the influence of any one datum.

Because of the large sample size, the histograms below are fairly useless, so I will turn the distribution of Cook's D statistics into a scatterplot….



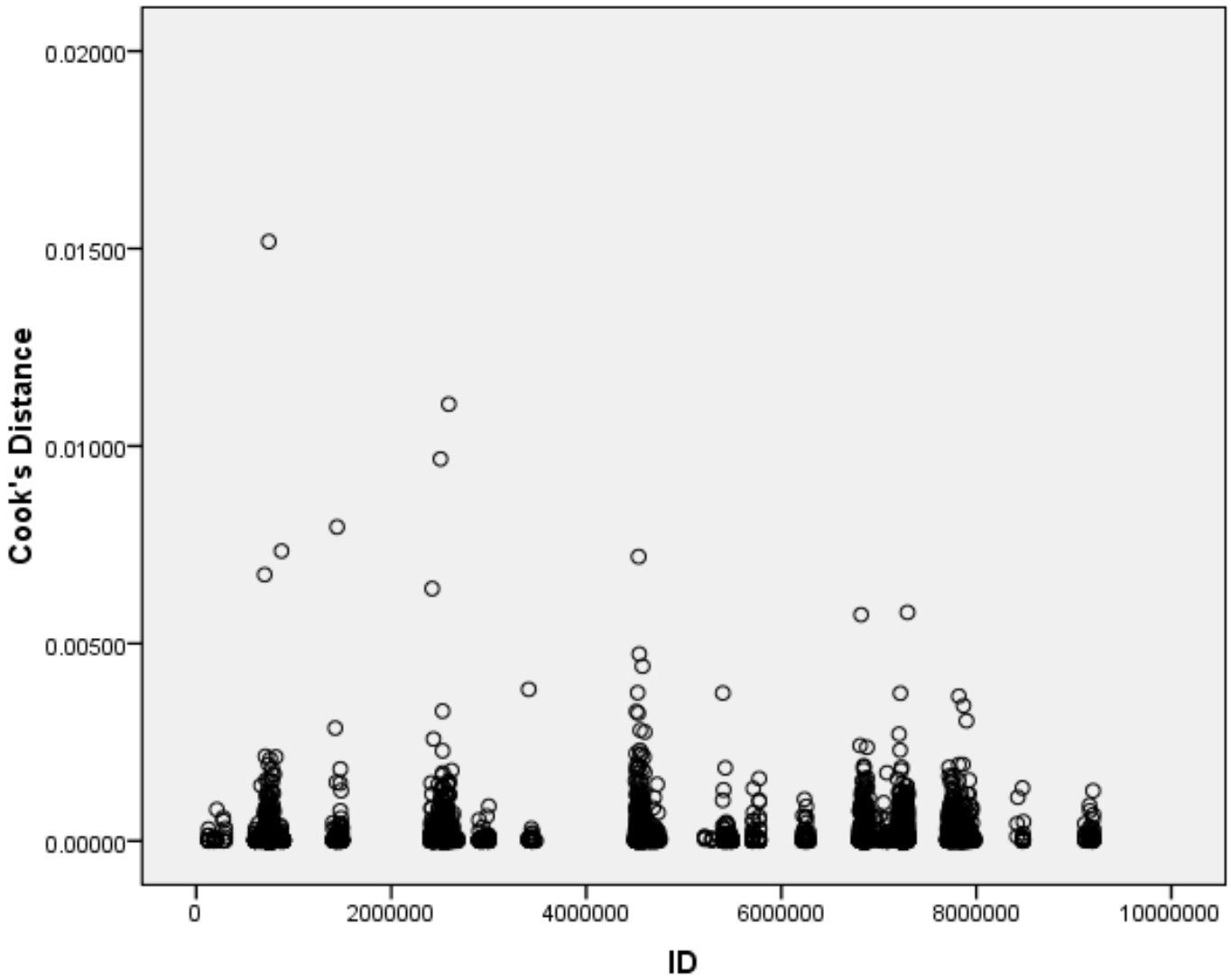


Extreme Values

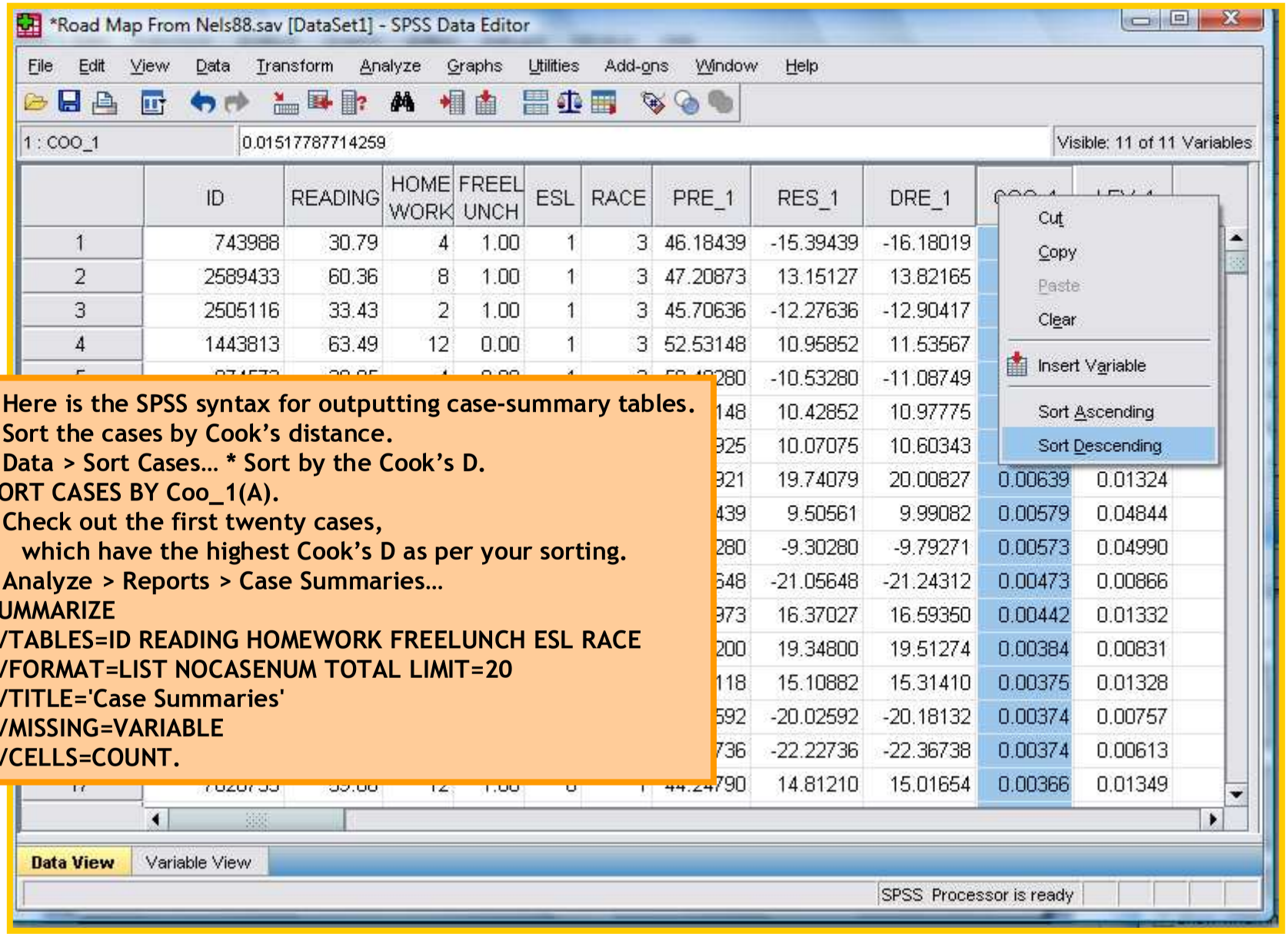
			Case Number	Value
Deleted Residual	Highest	1	276	22.76213
		2	3437	22.37566
		3	1833	21.30180
		4	6603	20.81806
		5	3178	2.03847E1
	Lowest	1	7487	-26.07928
		2	6956	-25.78442
		3	6441	-25.71603
		4	6690	-24.80400
		5	608	-24.13497
Centered Leverage Value	Highest	1	7666	.05097
		2	6989	.05011
		3	2085	.04999
		4	6680	.04990
		5	6752	.04990
	Lowest	1	6612	.00014
		2	6611	.00014
		3	6610	.00014
		4	6609	.00014
		5	6608	.00014[b]
Cook's Distance	Highest	1	2708	.01518
		2	5637	.01106
		3	1215	.00967
		4	6680	.00795
		5	2923	.00734
	Lowest	1	4351	.00000
		2	3561	.00000
		3	1081	.00000
		4	6268	.00000
		5	7058	.00000

A Better Look at The Influential Outliers

When we plot the Cook's D statistics versus an arbitrary x-variable, we see about 10 students that stand out from the pack. We will inspect those 10 students more closely to see if there is a further pattern.



Looking For Patterns in the Influential Outliers




```
* Here is the SPSS syntax for outputting case-summary tables.
* Sort the cases by Cook's distance.
* Data > Sort Cases… * Sort by the Cook's D.
SORT CASES BY Coo_1(A).
* Check out the first twenty cases,
    which have the highest Cook's D as per your sorting.
* Analyze > Reports > Case Summaries…
SUMMARIZE
  /TABLES=ID READING HOMEWORK FREELUNCH ESL RACE
  /FORMAT=LIST NOCASENUM TOTAL LIMIT=20
  /TITLE='Case Summaries'
  /MISSING=VARIABLE
  /CELLS=COUNT.
```

Unit 12 Appendix: Key Concepts

Notice that I use “increase.” The longitudinal data warrant the developmental conclusion!

READING92 has measurement error, and READING88 has measurement error, when I take their difference, their difference necessarily has more measurement error than either! Ah, well.

Whenever there is an element of randomness in the outcome, we expect regression to the mean. Measurement error is one possible source of randomness, but not the only possible source of randomness. If we predict adult height by mother’s height, we will get regression to the mean, even though there is only trivial measurement error with height.

Unit 12 Appendix: Key Interpretations

For the average student, we expect an increase of 6 reading points from the 8th grade to the 12th grade, from 48 points to 54 points.

From the RVF plot, we do not appear to have a problem with meeting the linearity assumption. However, due to a ceiling effect, the homoskedasticity and normality assumptions are questionably met. Other than the ceiling effect, the conditional variances appear roughly equal. We are concerned that the high-end predictions are negatively skewed because of the ceiling effect.

From a histogram of residuals and P-P plot, we see a slight negative skew of the residuals that we attribute to the ceiling effect of our reading measure.

There are no outliers of concern, in part because the large sample size minimizes the influence of any one datum.

When we plot the Cook's D statistics versus an arbitrary x-variable, we see about 10 students that stand out from the pack. We will inspect those 10 students more closely to see if there is a further pattern.

Unit 12 Appendix: Key Terminology

At least in simple linear regression, diagnostics provide information that we could conceivably glean from a bivariate scatterplot of the outome versus predictor; nevertheless they can provide a helpfully detailed view. In multiple regression, however, diagnostics provide information that we could never gather by eye.

A residual (aka error) is the difference between our observed outcome and our predicted outcome. If the residual is negative that means we should have predicted lower (i.e., we overpredicted). If the residual is positive, we should have predicted higher (i.e., we underpredicted). Of course, we expect residuals because of individual variation, hidden variables, and measurement error.

Every datum has an associated residual, and we can graph the residuals with a histogram.

A deleted residual is a residual based on subtracting the predicted value from the observed value, just like a typical, raw residual, except that the predicted value is calculated with the observation removed in order to avoid the part/whole problem in which we are looking for outliers from the trend but the outlier is part of the trend.

A leverage statistic is a measure of the extremity of an observation based on the value(s) of its predictor(s). When we have one predictor, we can easily see who is extreme on that predictor, but when we have 12 predictors, it can be impossible to see who is generally extreme on all predictors.

An influence statistic compares the trend line (calculated from all the data, including the observation) with a hypothetical trend line (calculated from all the data except the observation). The bigger the difference between the two trend lines, the great the influence. Cook's D statistic is the influence statistic that we will use, but there are others.

A residual versus fitted plot (RVF plot), also known as a residual versus predicted plot, is just what it says it is: a scatterplot of residual values versus fitted/predicted values.

A histogram of residuals can give an indication whether or not the residuals are normally distributed; however, use with caution, because histograms of residuals show an unconditional distribution (i.e., they don't think vertically). We are ultimately concerned with normality (and homoskedasticity) conditional on X. Nevertheless, such histograms can be useful, especially when supplemented with an RVF plot which allows you to think in terms of vertical slices and consequently think about conditional distributions.

A probability-probability plot (P-P plot) is another way of looking at a residual histogram, with a focus on normality. In a normal distribution we expect 50% of the observations to be below average and, because it's a mathematical construct, we observe 50% of the observations to be below average. This simple truth forms our baseline of comparison (in red below). In a sample distribution from a population with a normal distribution, we expect 50% of the observations to be below average, but due to sampling error, we may observe more or fewer than 50% of observations to be below average.

Unit 12 Appendix: Formulas

Reliability Notation

The reliability of measurement X is denoted: $\rho_{XX'}$, where the Greek letter rho, stands for the population correlation, the subscript X stands for one form of measurement X, and the subscript X' stands for a parallel form of the measurement.

The Reliability of a Difference

$$\rho_{DD'} = \frac{\rho_{XX'}\sigma_X^2 + \rho_{YY'}\sigma_Y^2 - 2\rho_{XY}\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2 - 2\rho_{XY}\sigma_X\sigma_Y}$$

The population variance of measurement X is denoted: σ_X^2 .

The population standard deviation of measurement X is denoted: σ_X .

The population correlation between measurements X and Y is denoted: ρ_{XY} .

Unit 12 Appendix: Formulas

The Reliability of a Difference

$$\rho_{DD'} = \frac{\rho_{XX'}\sigma_X^2 + \rho_{YY'}\sigma_Y^2 - 2\rho_{XY}\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2 - 2\rho_{XY}\sigma_X\sigma_Y}$$

$$\rho_{DD'} = \text{WantBig} = \frac{\text{WantBig}}{\text{WantSmall}} = \frac{\text{WantBig} + \text{WantBig} - \text{WantSmall}}{\text{BaselineOf PerfectReliability}: \rho_{XX'} = \rho_{YY'} = 1}$$

For the sake of simplicity, let us assume that measurements X and Y are standardized such that their standard deviations (and consequently variances) equal one:

$$\rho_{DD'} = \frac{\rho_{XX'} + \rho_{YY'} - 2\rho_{XY}}{2 - 2\rho_{XY}}$$

- We want the reliable variance in measurement X to be big.
- We want the reliable variance in measurement Y to be big.
- We want the correlation between measurements X and Y to be small.
  - If the correlation is negative, then the reliability of the difference can actually exceed the reliability of the individual tests!
- What happens when measurement X and Y are perfectly reliable?
- What happens when measurement X and Y are perfectly unreliable?
  - Note that, if measurements X and Y are perfectly unreliable, then they must be perfectly uncorrelated as well.
- What happens when measurement X and Y are perfectly correlated?
  - Note that, if measurements X and Y are perfectly correlated (and they have the same standard deviation), then everybody has the same exact difference score.

Unit 12 Appendix: SPSS Syntax

```
*Example SPSS syntax for computing transformed variables.

*This linear transformation is not a z-transformation because I did not divide the difference by the standard
deviation.

COMPUTE ZEROCENTEREDREADING88 = READING88 - 48.0155.

EXECUTE.

*This (goofy) transformation is non-linear because I do more than add/subtract and/or multiply/divide by a constant. I
use powers and logs.

COMPUTE Sean_Is_A_Great_SPSS_Programmer = READING88 * 48.0155 - 1975/27 + FREELUNCH**(1/2) +
LN(HOMEWORK+1).

EXECUTE.

*If we are interested in changes, let's compute a change score and use it as our outcome. This is not a linear
transformation because I add/subtract and/or multiply/divide by a variable, not a constant.

COMPUTE READINGIMPROVEMENT = READING92 - READING88.

EXECUTE.

* Identify the residual, temporarily remove it, and refit the line.

TEMPORARY.

SELECT IF NOT (ID = 2999973).

REGRESSION

  /MISSING LISTWISE

  /STATISTICS COEFF OUTS CI R ANOVA

  /CRITERIA=PIN(.05) POUT(.10)

  /NOORIGIN

  /DEPENDENT READINGIMPROVEMENT

  /METHOD=ENTER ZEROCENTEREDREADING88.
```

Unit 12 Appendix: SPSS Syntax

```
* We do not have calculate deleted residual "by hand," we can have the computer do it automatically for every case,
and, along the way, we can have the computer do a whole bunch of other things.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT READINGIMPROVEMENT
  /METHOD=ENTER ZEROCENTEREDREADING88
  /SCATTERPLOT=(*RESID ,*PRED)
  /RESIDUALS HIST(RESID) NORM(RESID)
  /SAVE PRED RESID DRESID LEVER COOK.
* Once we produce our variables, we can examine them.
EXAMINE VARIABLES=DRE_1 LEV_1 COO_1
  /COMPARE GROUP
  /STATISTICS DESCRIPTIVES EXTREME
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
GRAPH
  /HISTOGRAM (NORMAL)=DRE_1.
GRAPH
  /HISTOGRAM=LEV_1.
GRAPH
  /HISTOGRAM=COO_1.
```

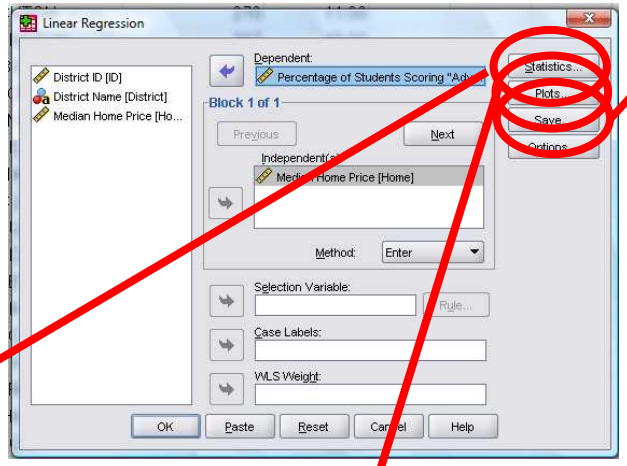
Output Your New Variables and Nifty Plots

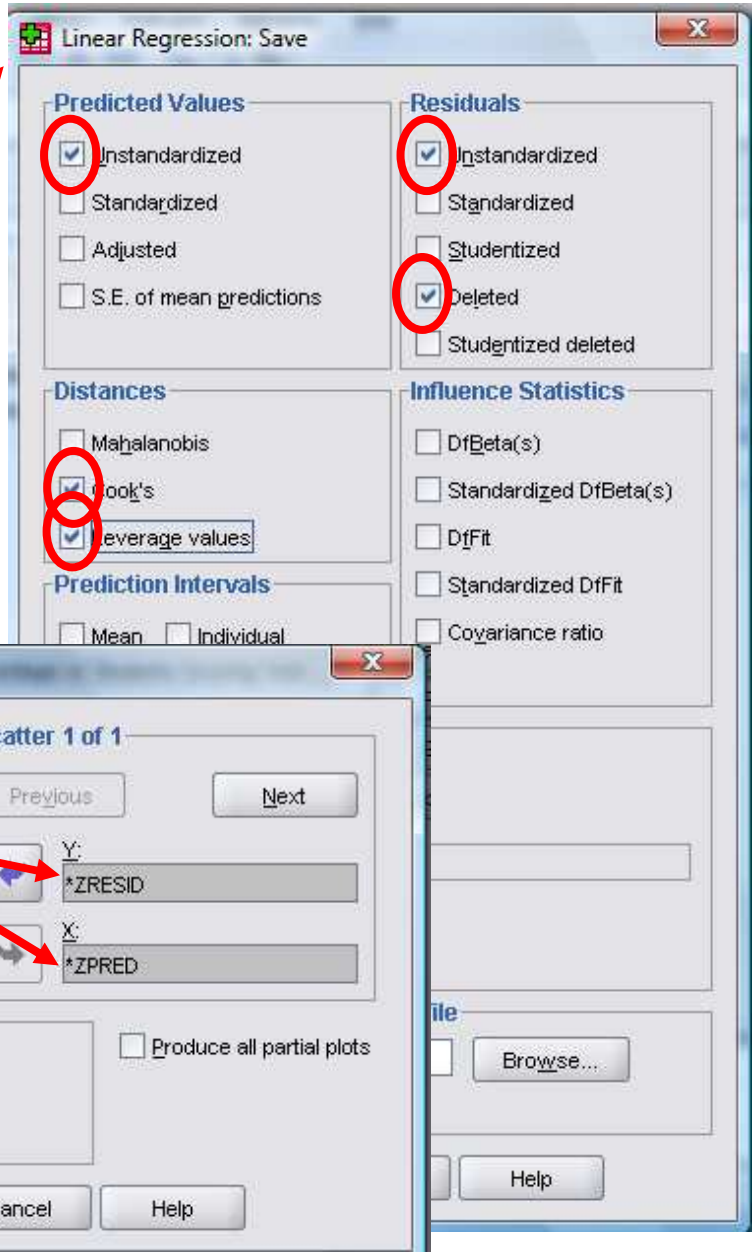
Start from:

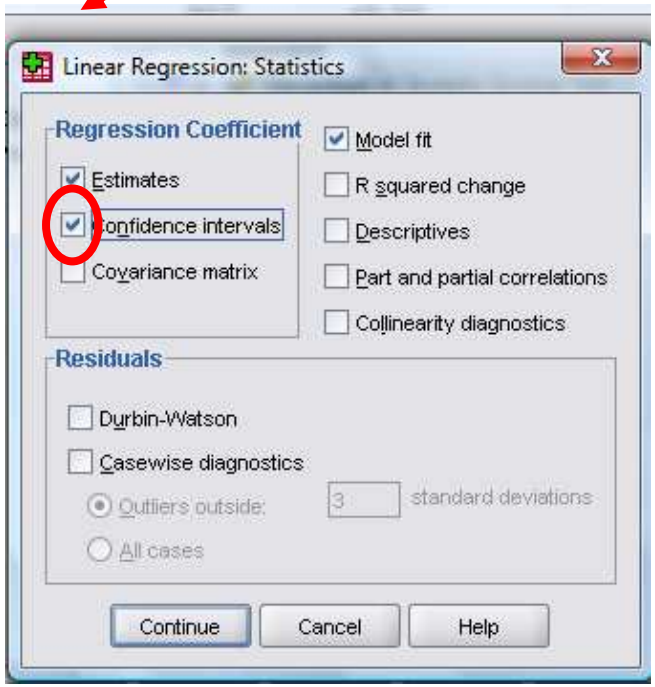
Analyze >

Regression >

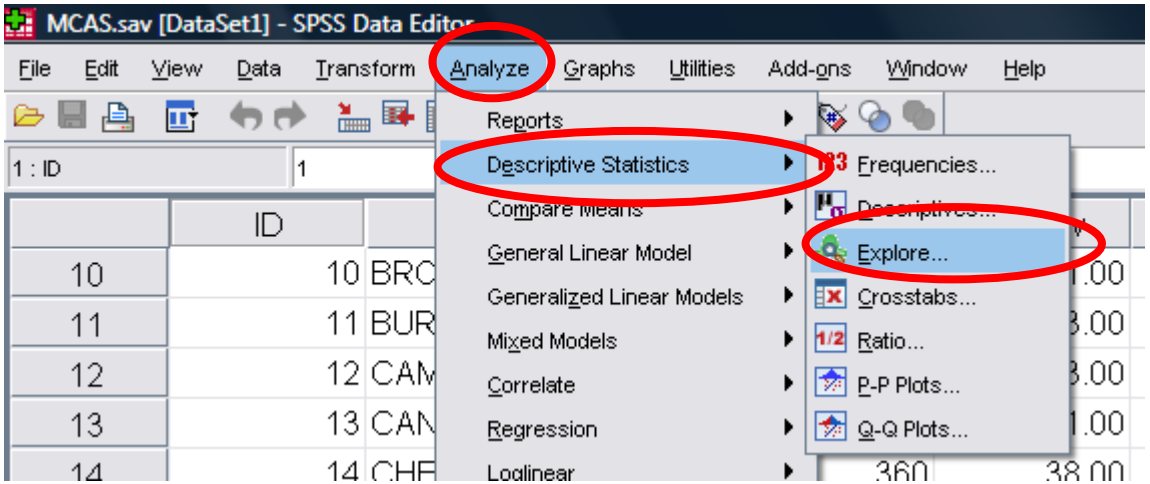
Linear



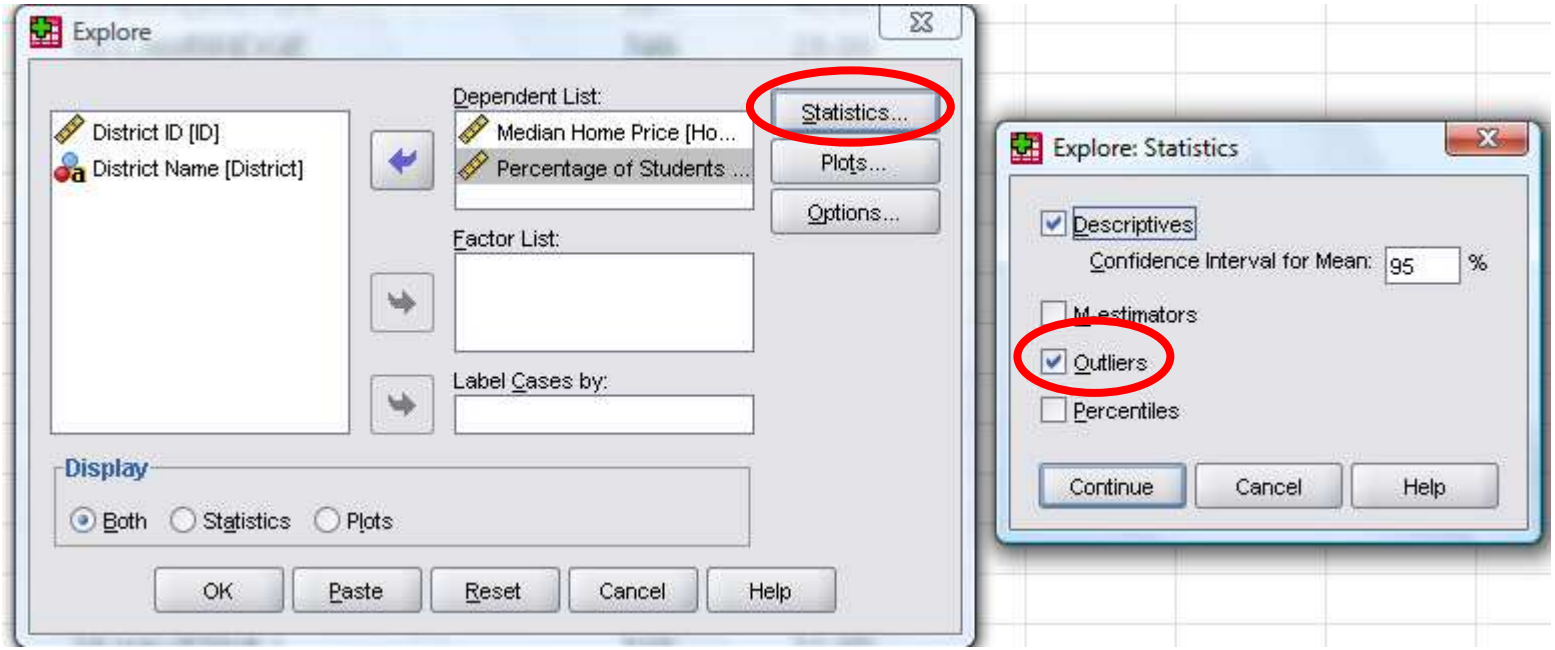




Examine Your New Variables



Look around and check out your options.



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

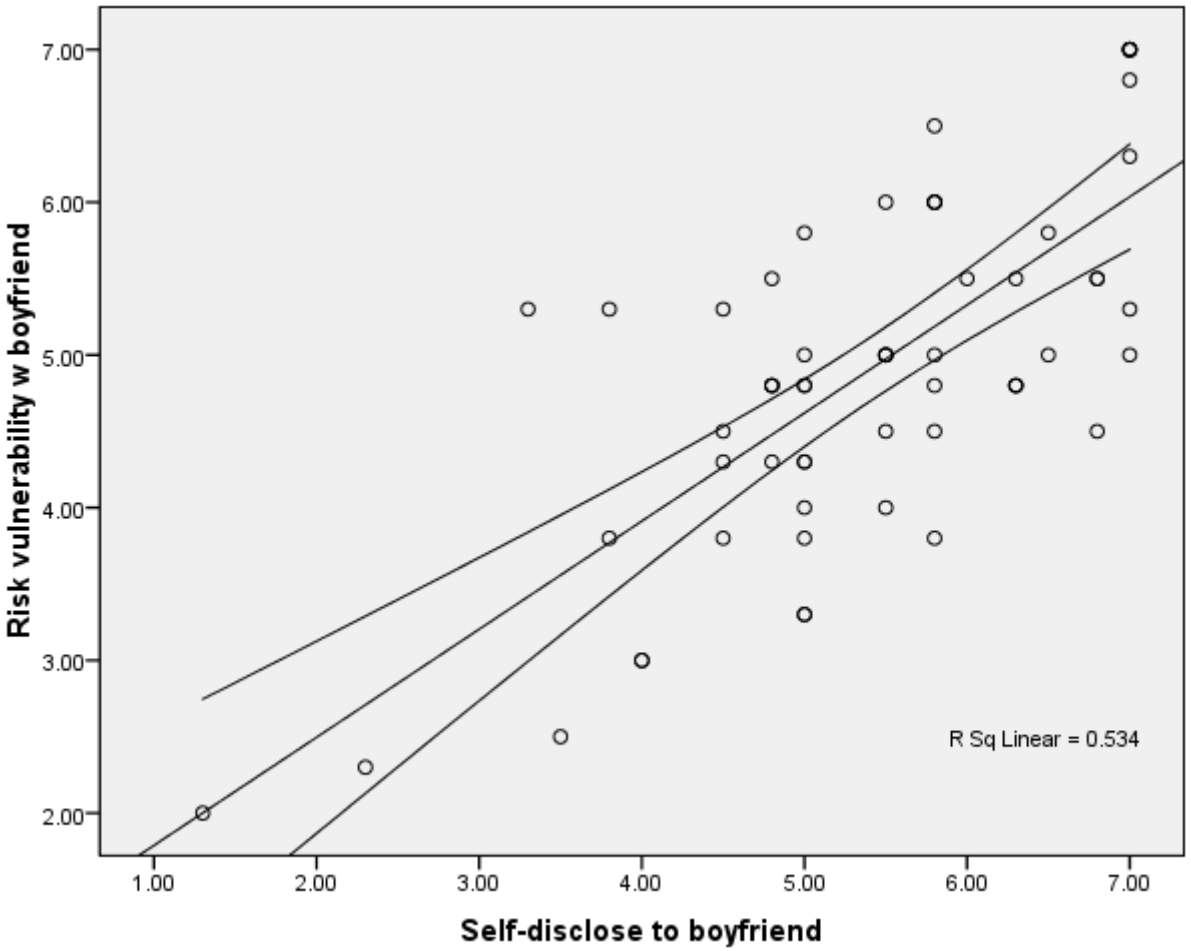
b. Dependent Variable: Risk vulnerability w boyfriend

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

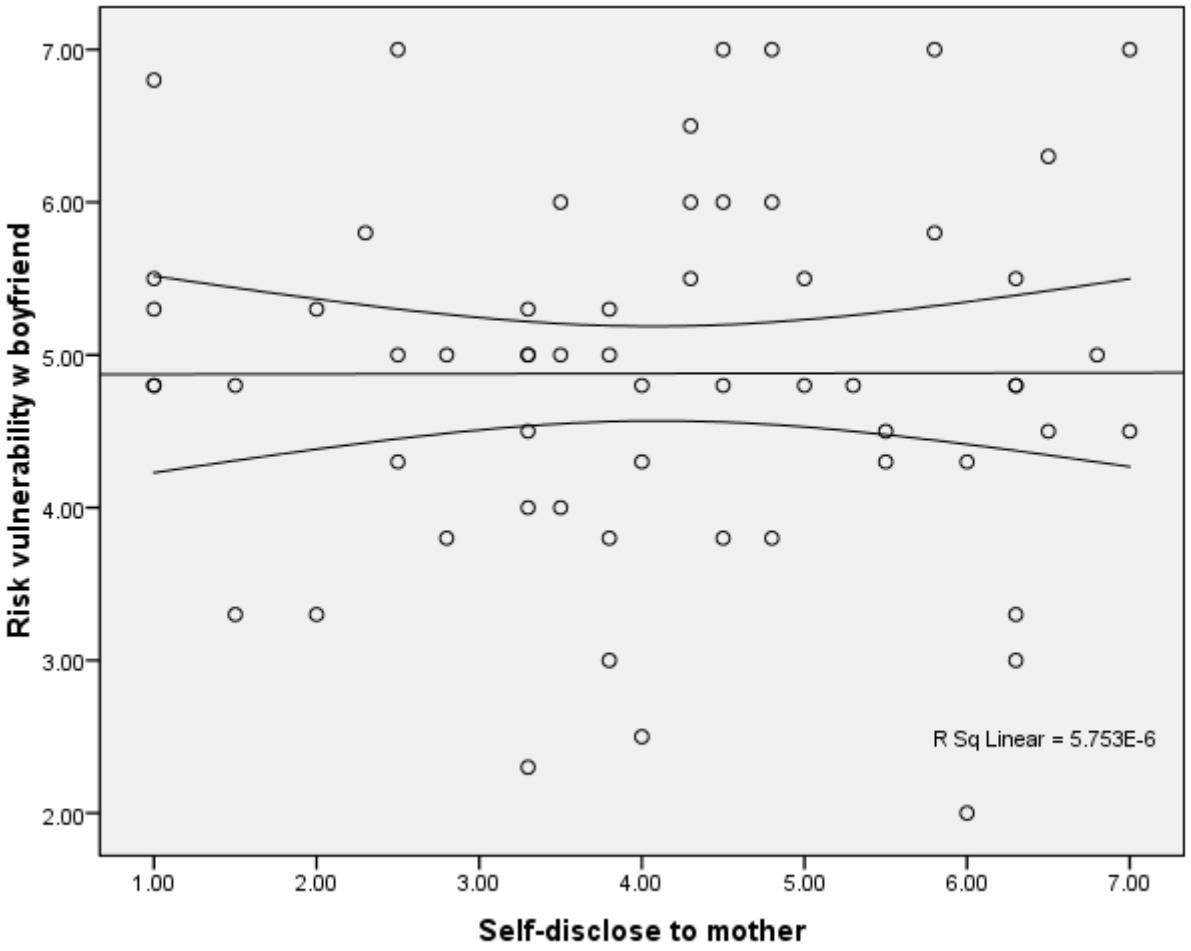
b. Dependent Variable: Risk vulnerability w boyfriend

Coefficients[a]

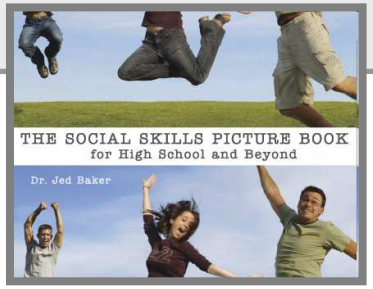
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	4.872	.404		12.050	.000	4.062	5.681
	Self-disclose to mother	.002	.091	.002	.018	.985	-.181	.184

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:**

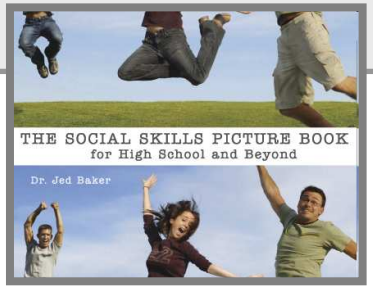
Variables about the student—

(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

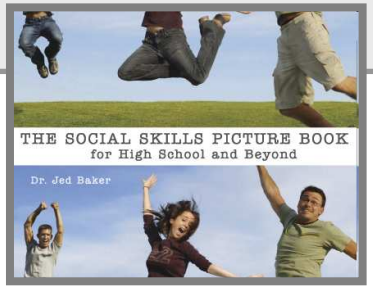
b. Dependent Variable: Base Year Composite Test

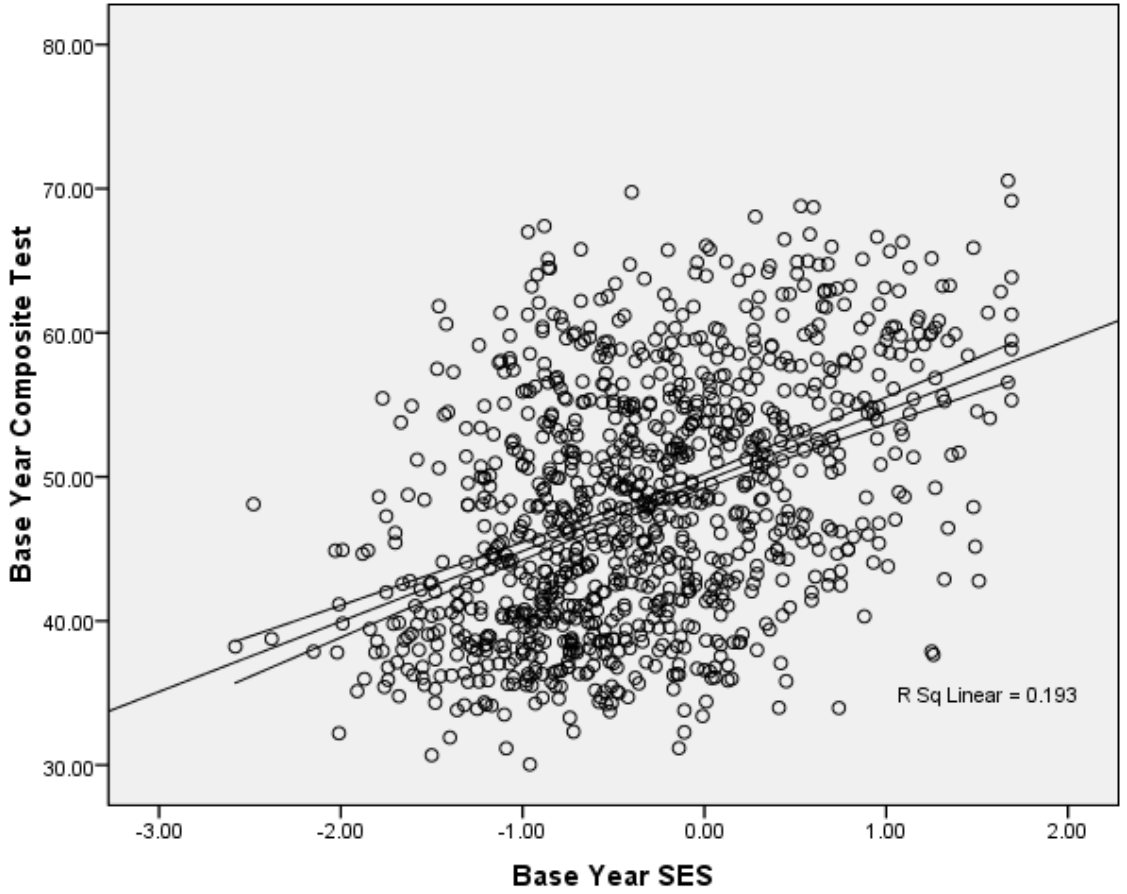
Coefficients[a]

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

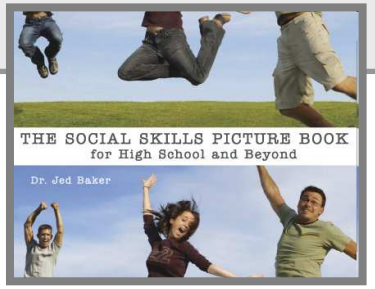
a. Dependent Variable: Base Year Composite Test

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

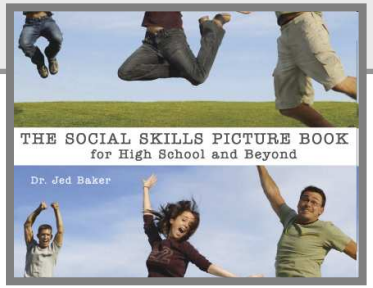
b. Dependent Variable: Base Year Composite Test

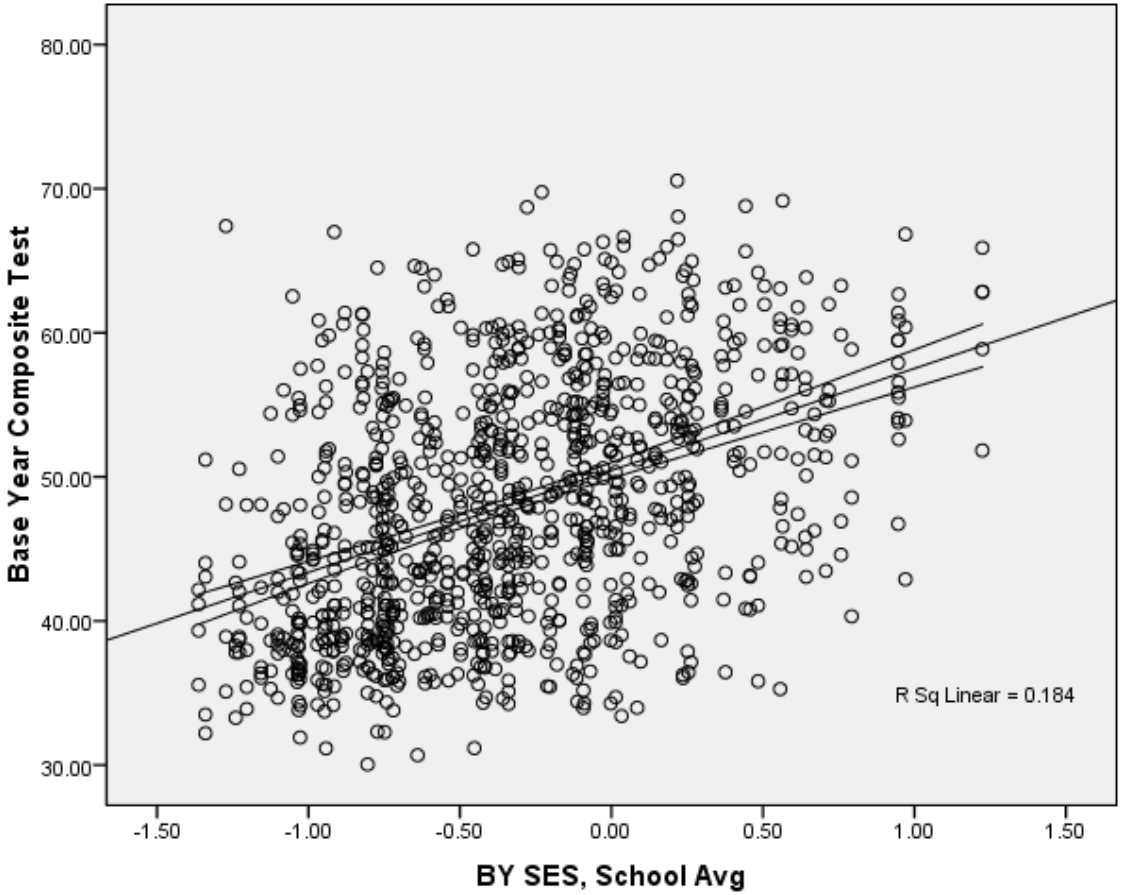
Coefficients[a]

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

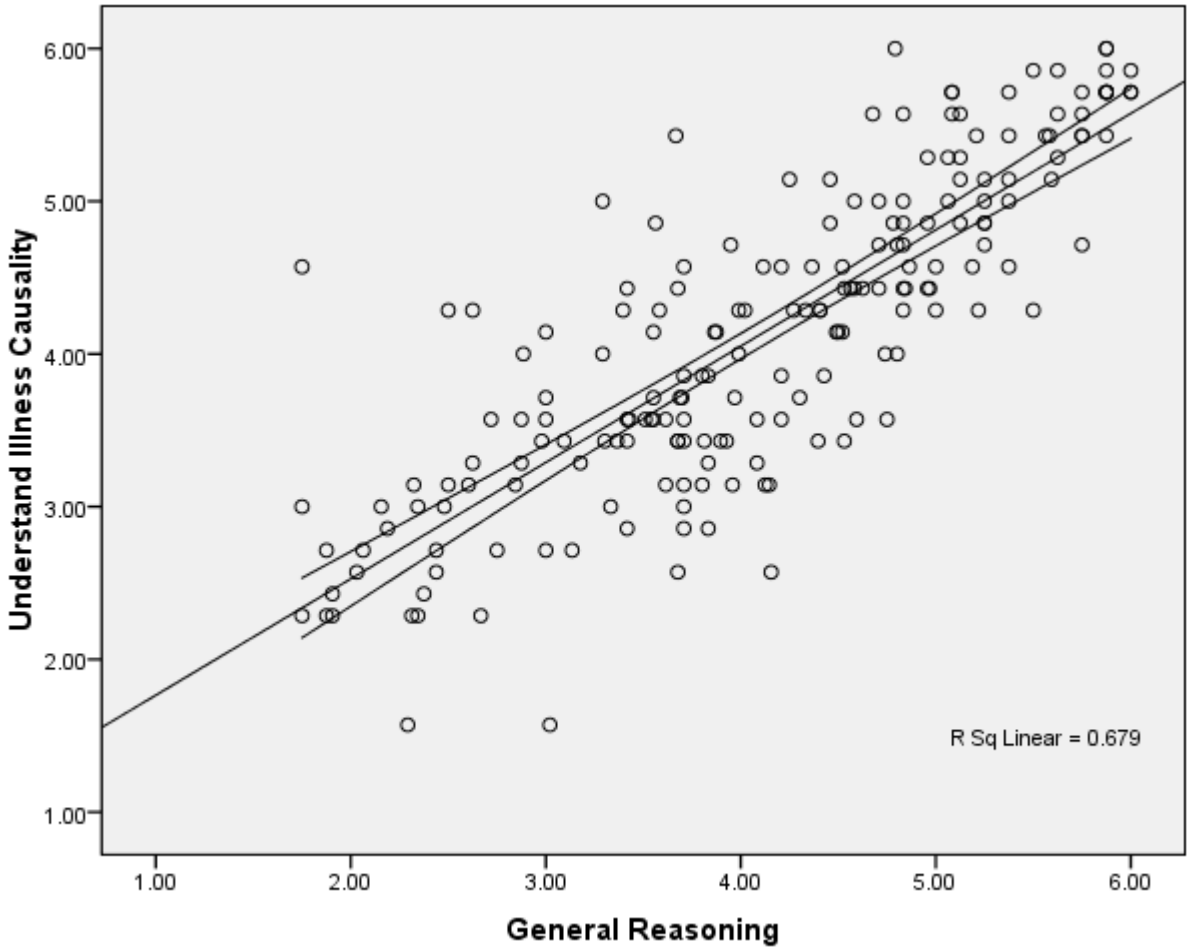
b. Dependent Variable: Understand Illness Causality

Coefficients[a]

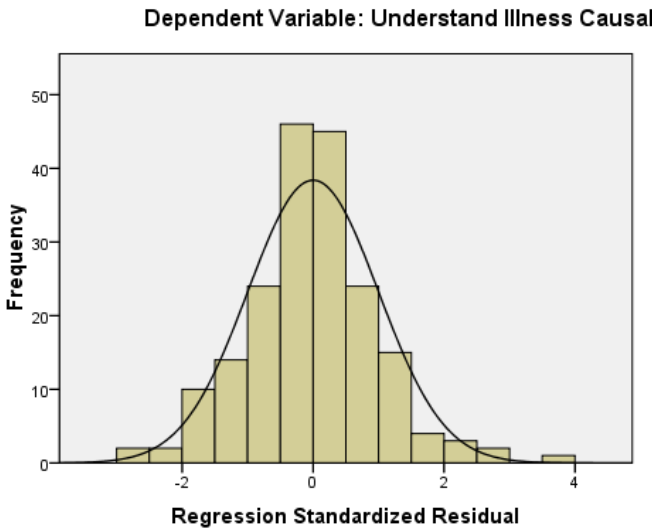
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	1.004	.162		6.204	.000	.685	1.323
	General Reasoning	.762	.038	.824	20.061	.000	.687	.837

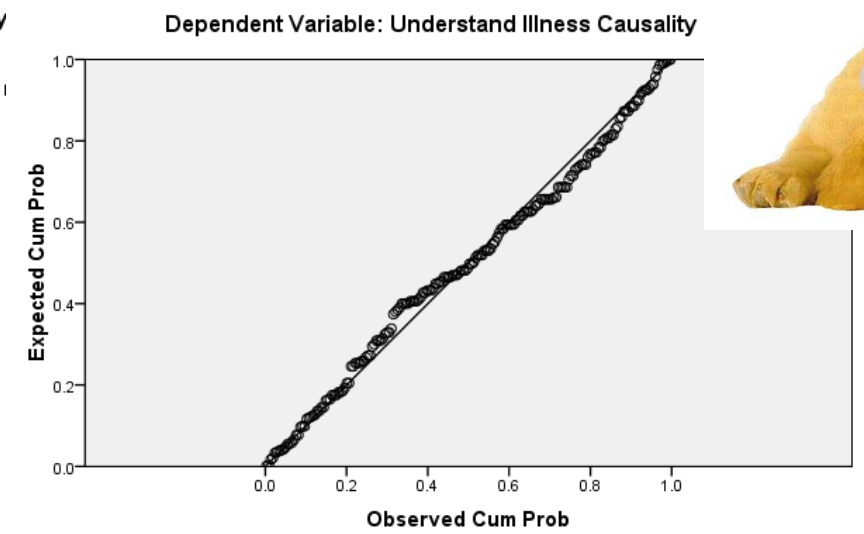
a. Dependent Variable: Understand Illness Causality

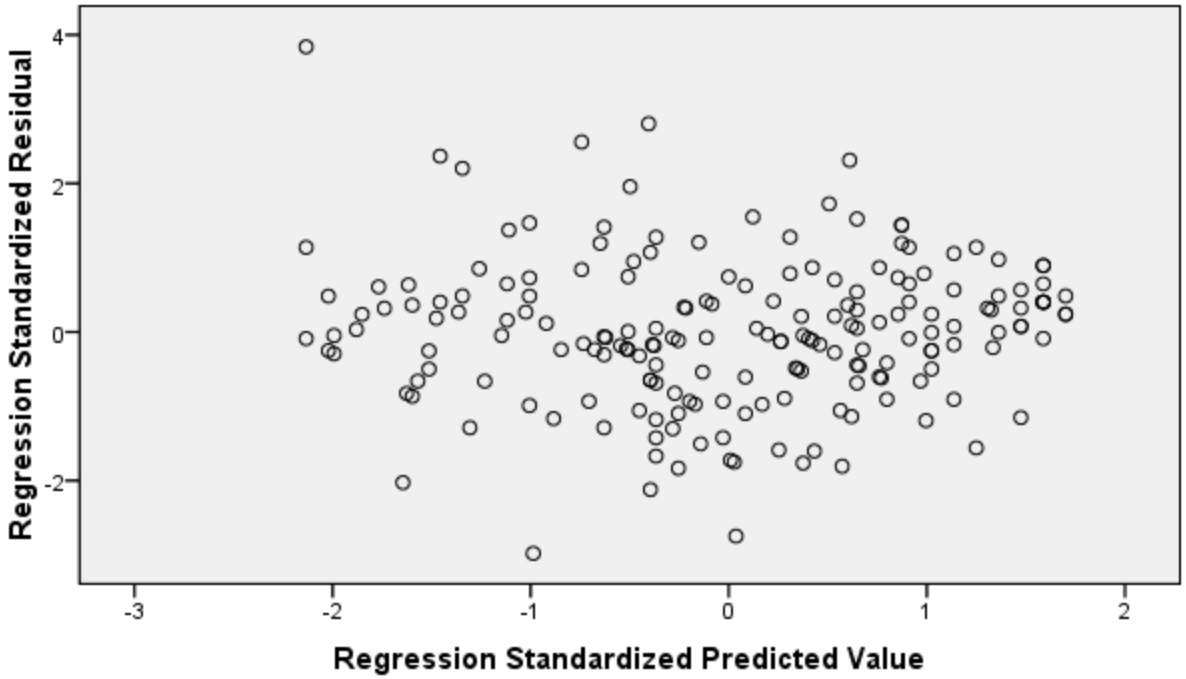
Understanding Causes of Illness (ILLCAUSE.sav)



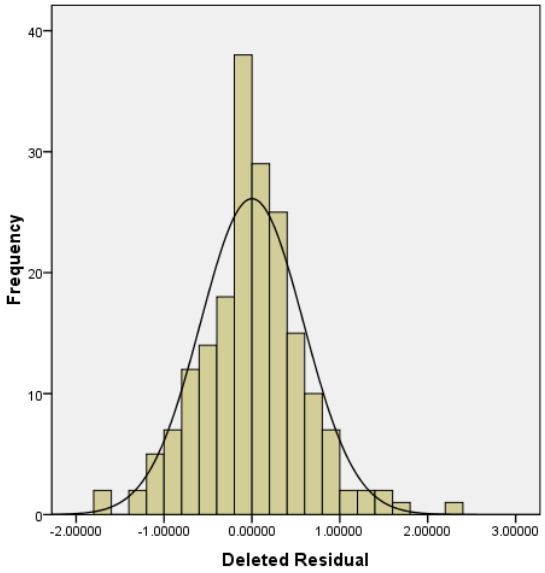
Understanding Causes of Illness (ILLCAUSE.sav)

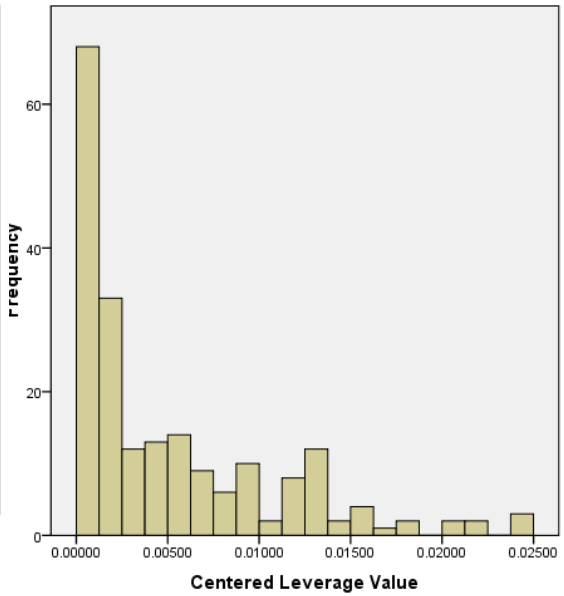


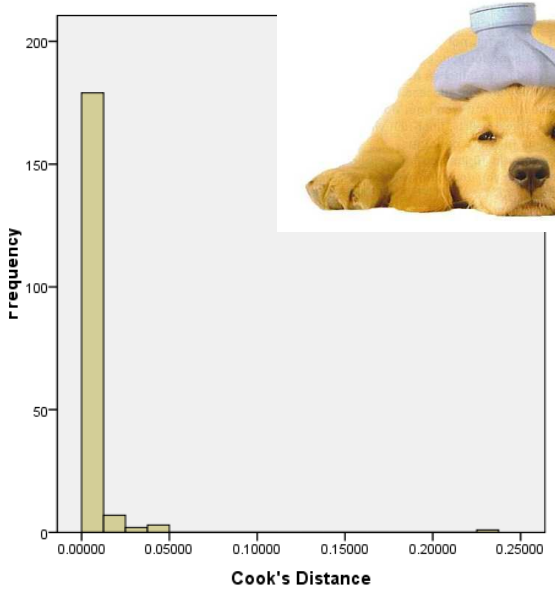




Understanding Causes of Illness (ILLCAUSE.sav)







Extreme Values

			Case Number	Value
Deleted Residual	Highest	1	86	2.30067
		2	116	1.64138
		3	183	1.50026
		4	5	1.40034
		5	158	1.35507
	Lowest	1	40	-1.75250
		2	49	-1.60749
		3	48	-1.24168
		4	46	-1.20236
		5	7	-1.07299

Centered Leverage Value	Highest	1	77	.02384
		2	86	.02384
		3	106	.02384
		4	72	.02138
		5	94	.02138
	Lowest	1	100	.00000

Cook's Distance	Highest	1	86	.22708
		2	5	.04729
		3	40	.04677
		4	46	.04137
		5	113	.03670
	Lowest	1	182	.00000

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

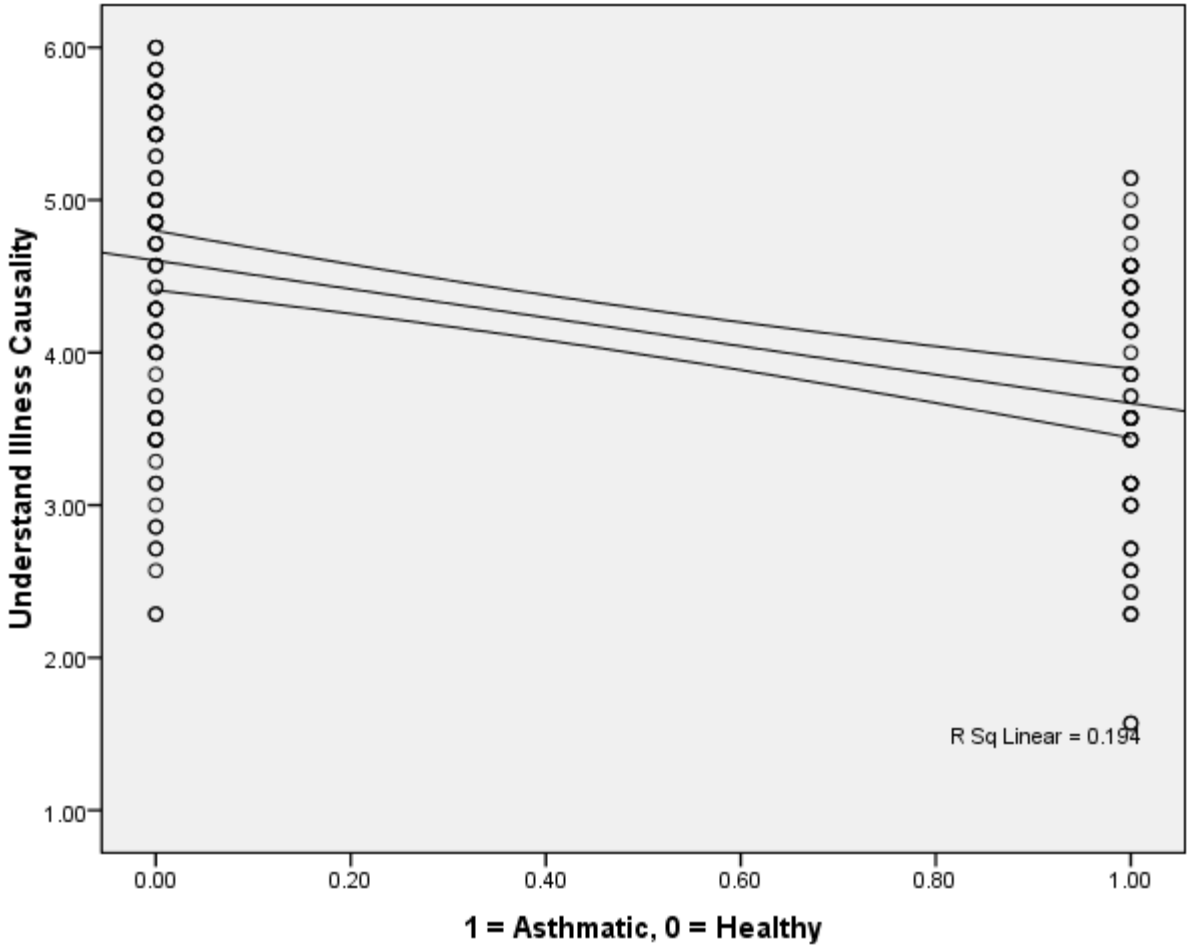
b. Dependent Variable: Understand Illness Causality

Coefficients[a]

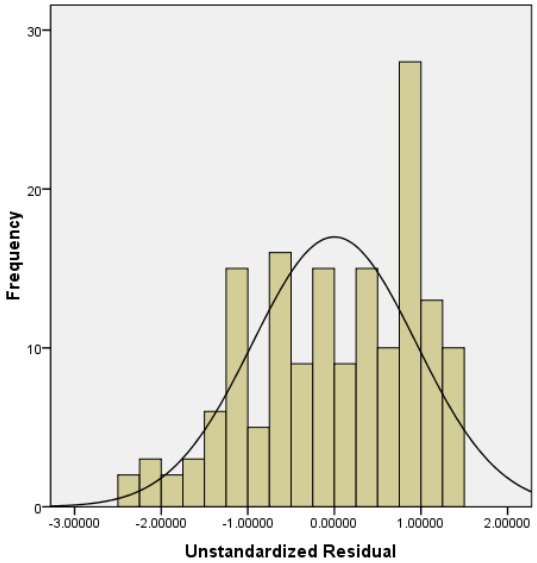
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	4.604	.098		46.807	.000	4.409	4.798
	1 = Asthmatic, 0 = Healthy	-.936	.151	-.440	-6.182	.000	-1.234	-.637

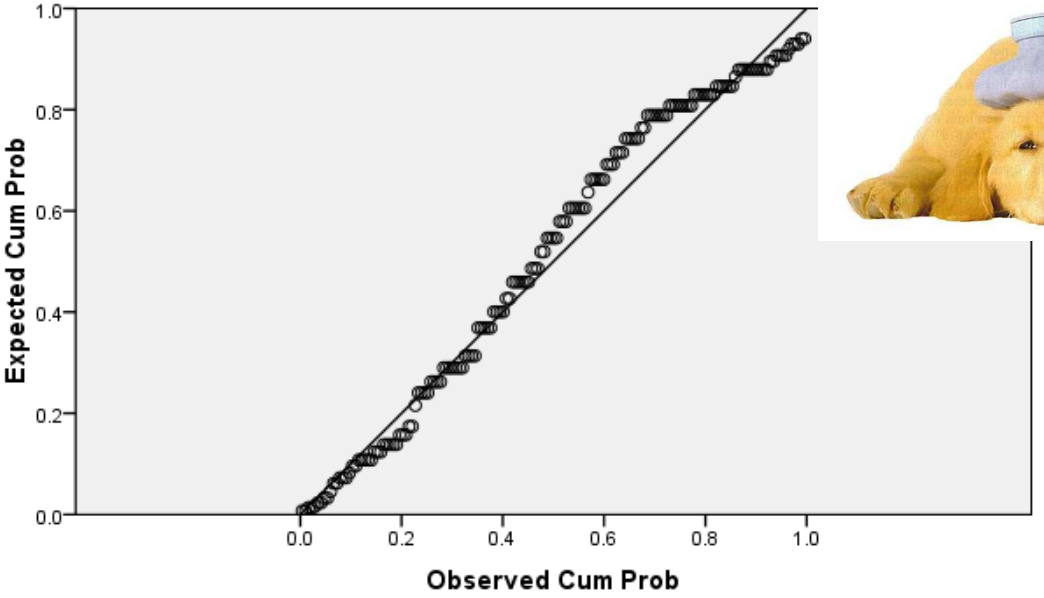
a. Dependent Variable: Understand Illness Causality

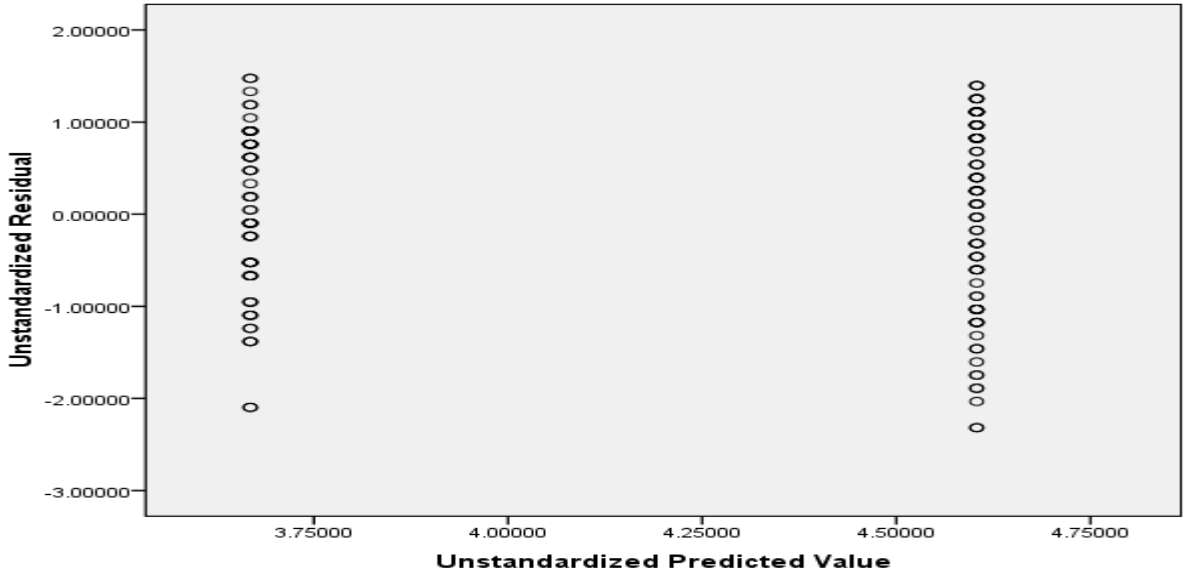
Understanding Causes of Illness (ILLCAUSE.sav)



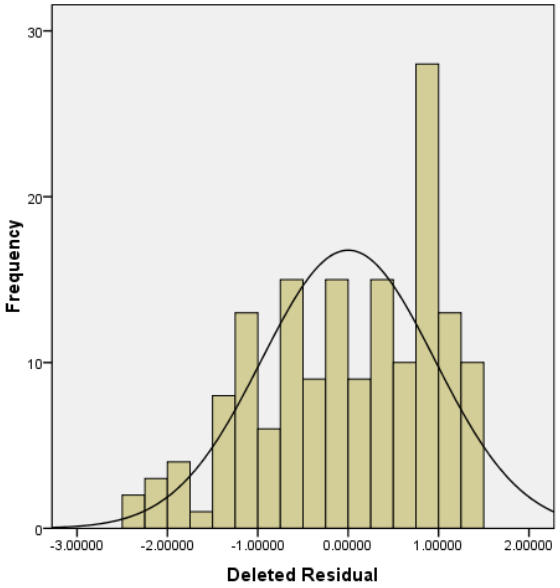
Understanding Causes of Illness (ILLCAUSE.sav)

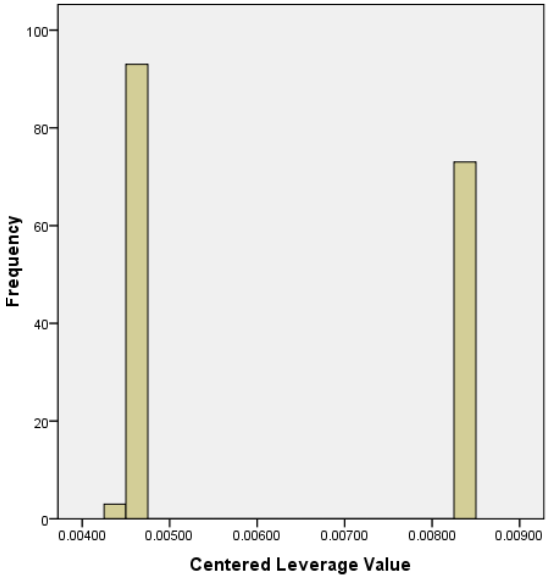


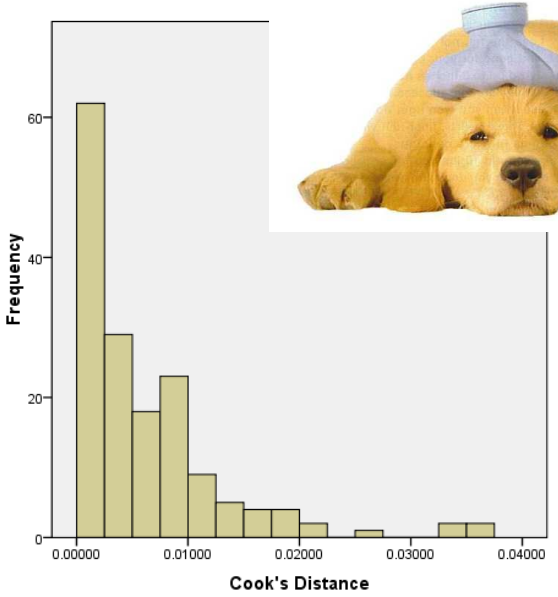




Understanding Causes of Illness (ILLCAUSE.sav)







Extreme Values

			Case Number	Value
Deleted Residual	Highest	1	39	1.49696
		2	51	1.49696
		3	124	1.41152
		4	151	1.41152
		5	158	1.41152
	Lowest	1	142	-2.34285
		2	125	-2.34285
		3	46	-2.12836
		4	40	-2.12836
		5	145	-2.05475
Centered Leverage Value	Highest	1	37	.00849
		2	38	.00849
		3	39	.00849
		4	40	.00849
		5	41	.00849[a]
	Lowest	1	205	.00454
Cook's Distance	Highest	1	40	.03702
		2	46	.03702
		3	125	.03280
		4	142	.03280
		5	145	.02523
	Lowest	1	180	.00001

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

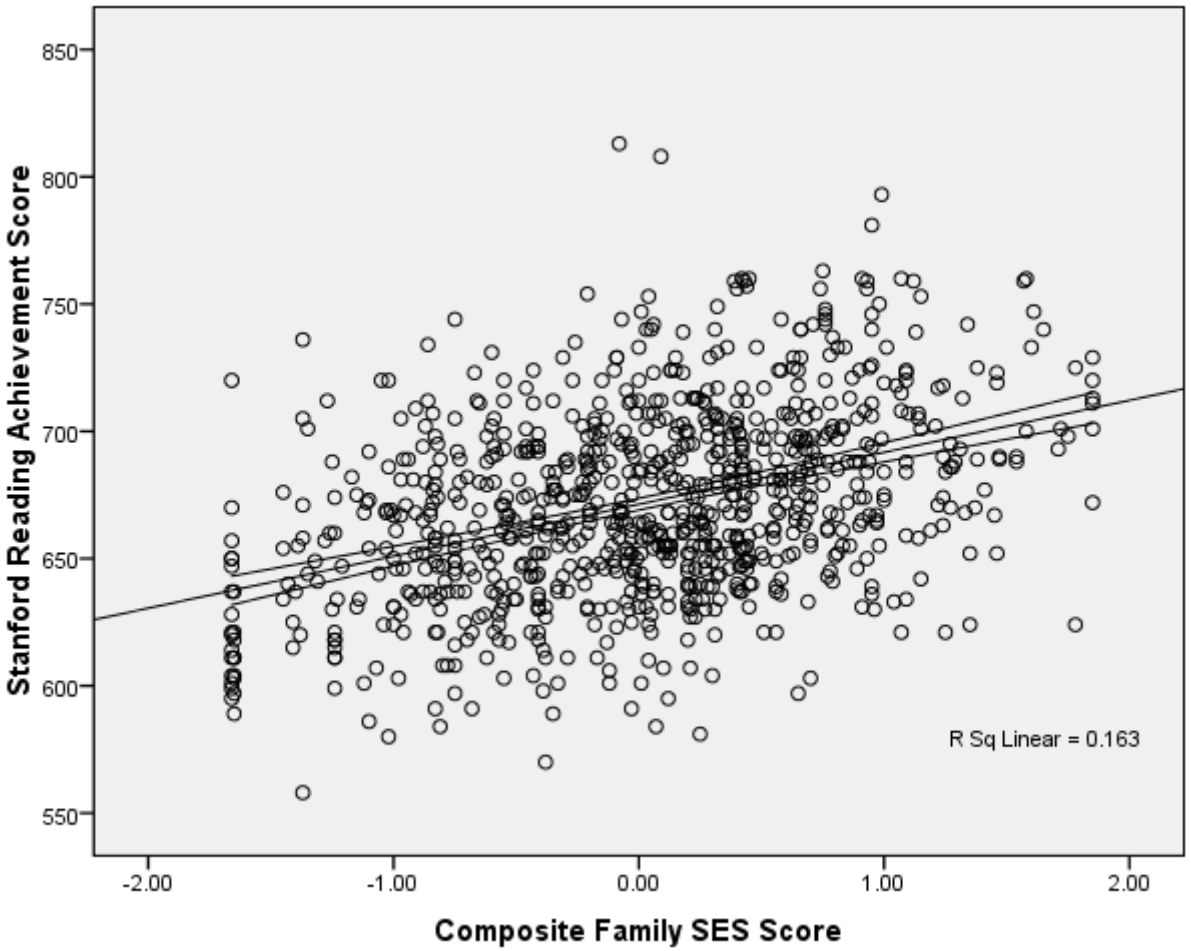
b. Dependent Variable: Stanford Reading Achievement Score

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

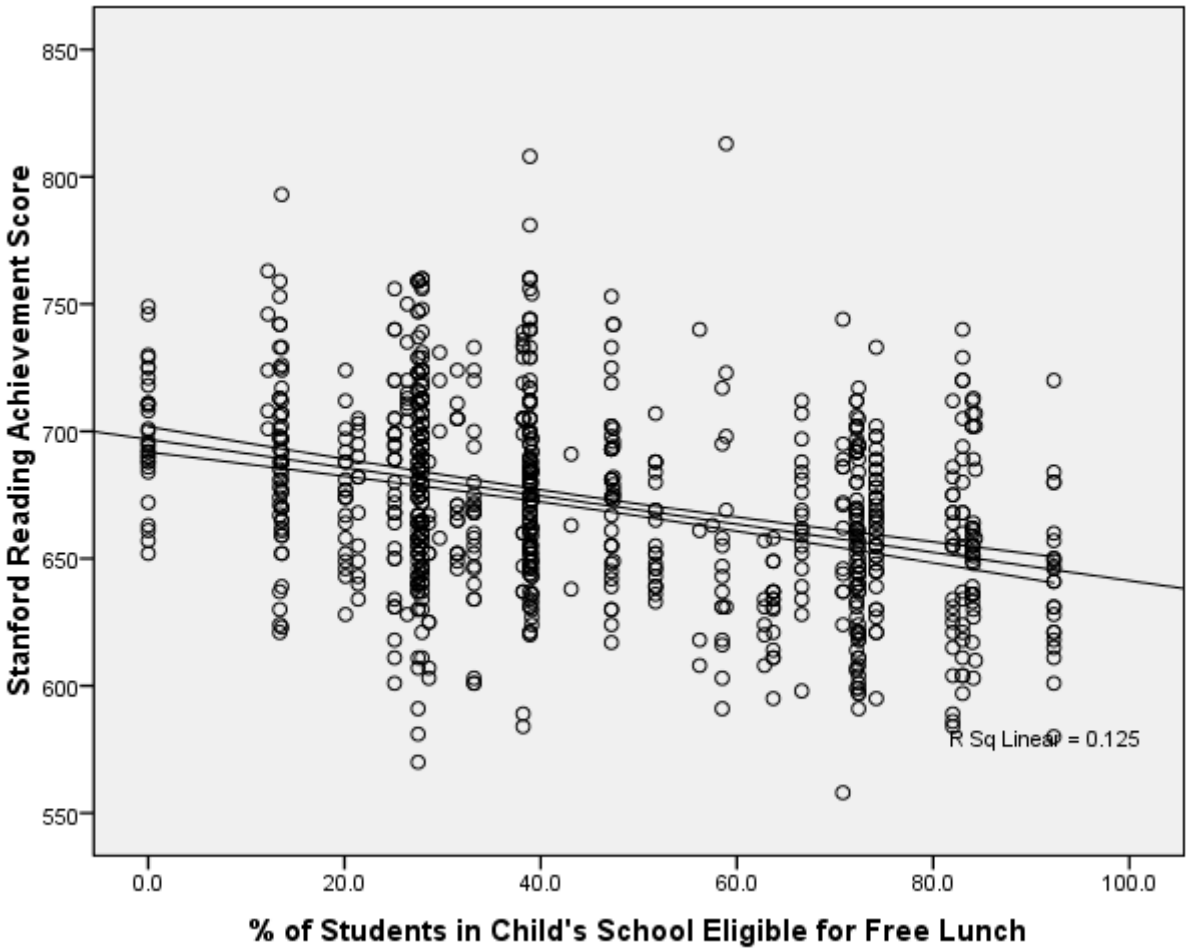
b. Dependent Variable: Stanford Reading Achievement Score

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

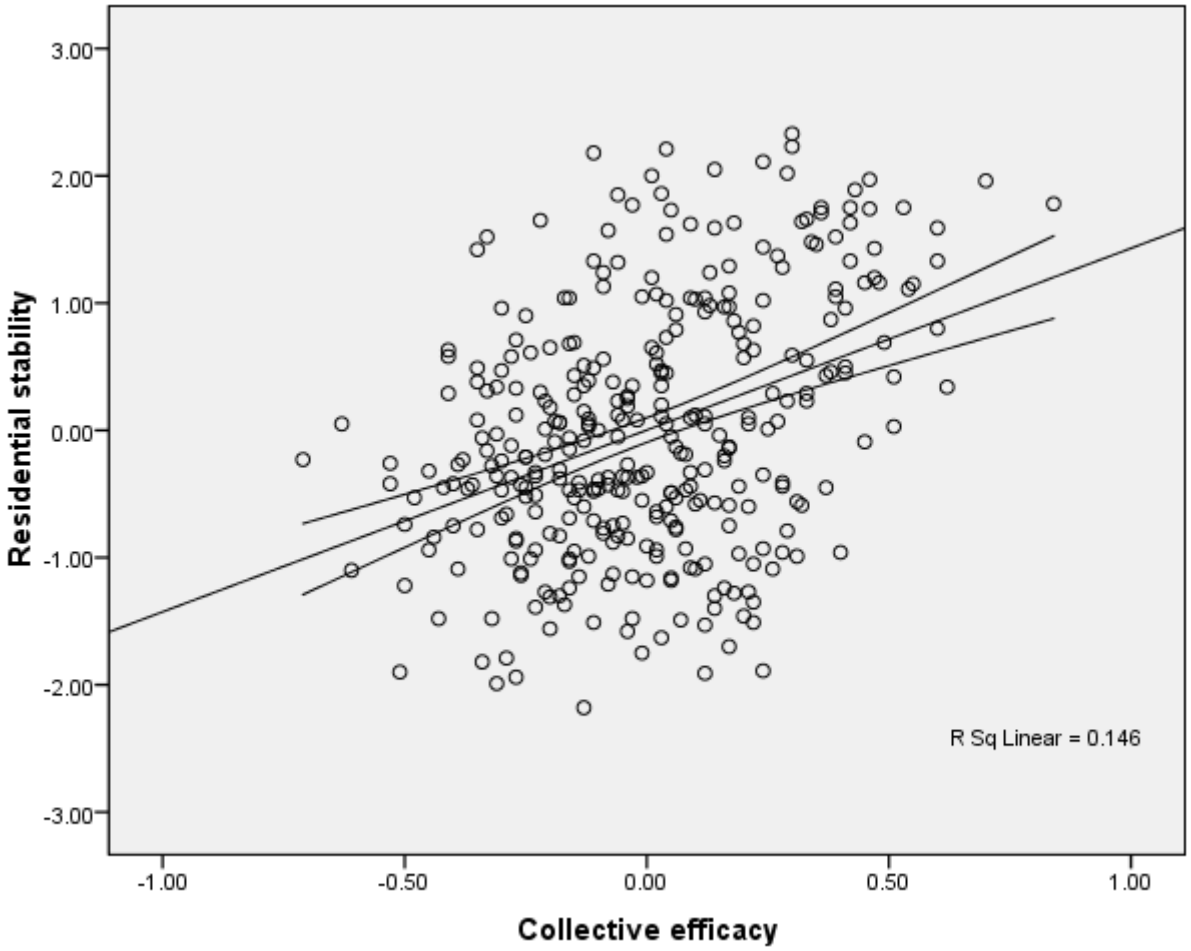
b. Dependent Variable: Residential stability

Coefficients[a]

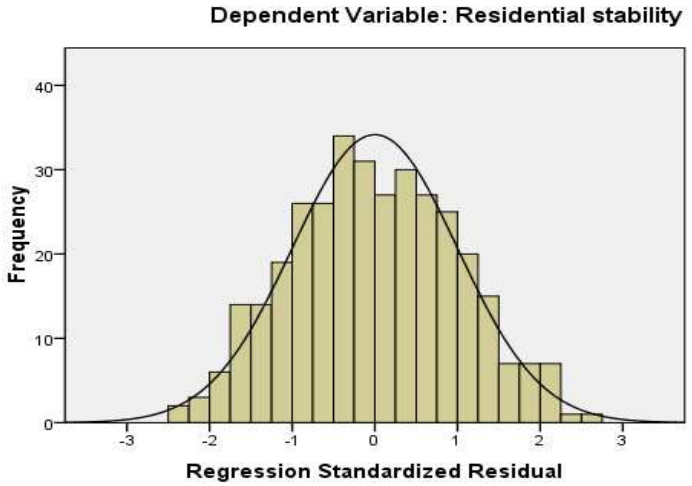
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.002	.049		.050	.961	-.094	.099
	Collective efficacy	1.429	.187	.382	7.620	.000	1.060	1.797

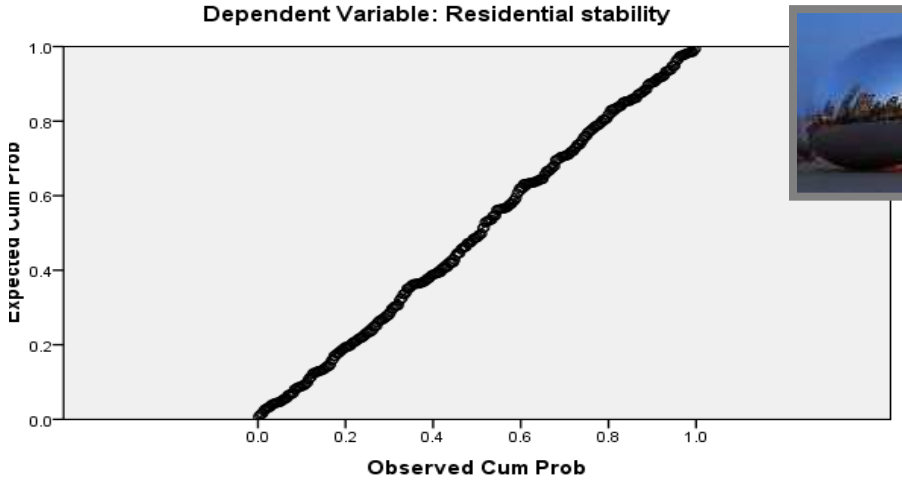
a. Dependent Variable: Residential stability

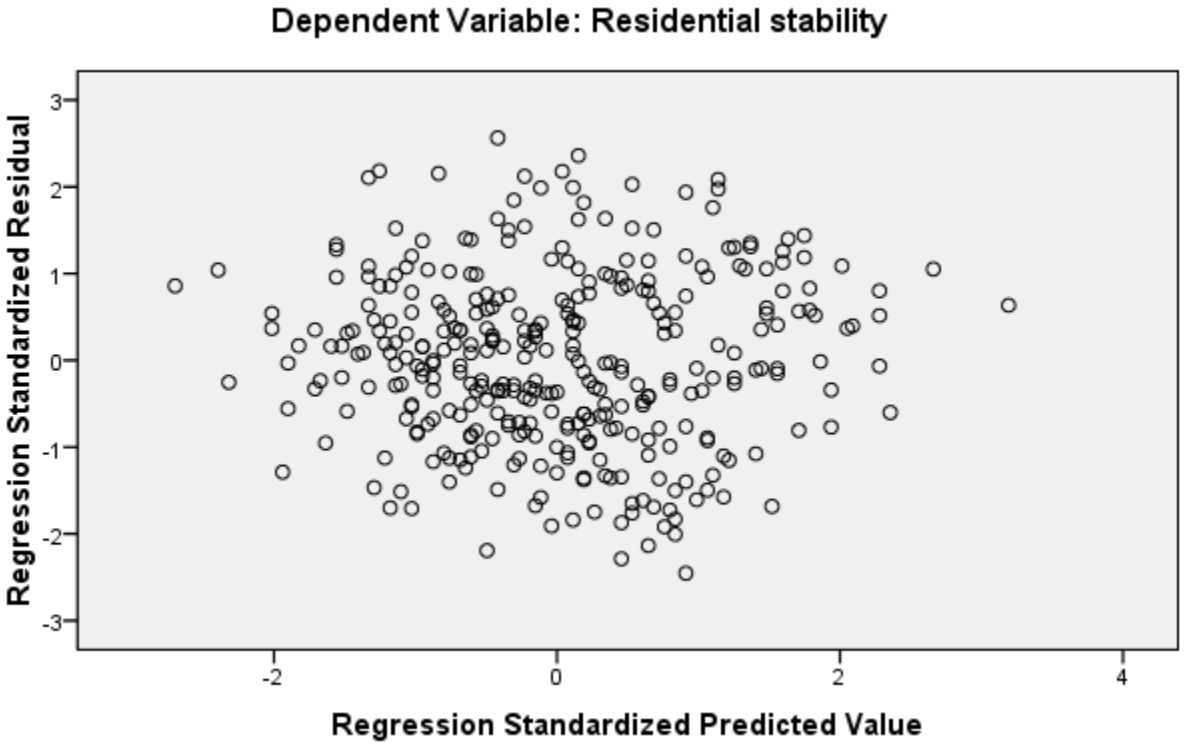
Human Development in Chicago Neighborhoods (Neighborhoods.sav)



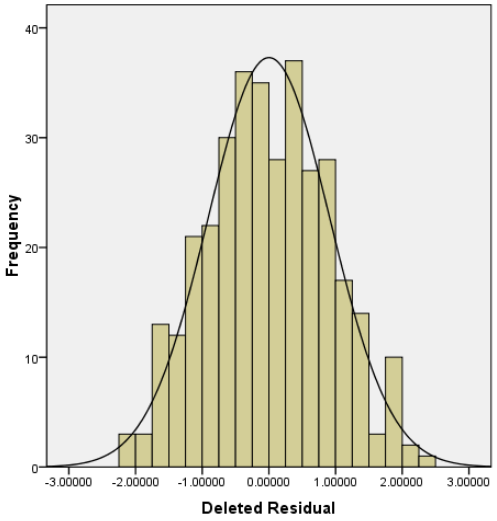
Human Development in Chicago Neighborhoods (Neighborhoods.sav)

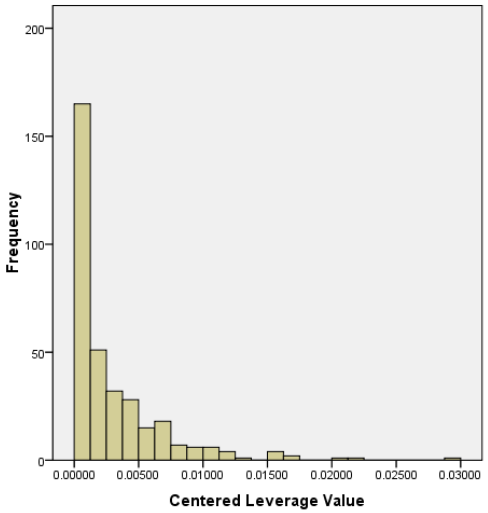


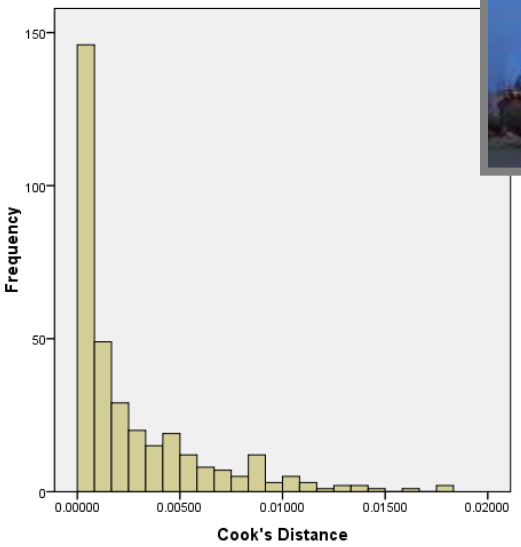




Human Development in Chicago Neighborhoods (Neighborhoods.sav)







Extreme Values

			Case Number	Value
Deleted Residual	Highest	1	246	2.34277
		2	303	2.15686
		3	252	2.00414
		4	237	1.98910
		5	239	1.97168
	Lowest	1	194	-2.24737
		2	42	-2.09127
		3	53	-2.00401
		4	162	-1.95341
		5	41	-1.83587

Centered Leverage Value	Highest	1	333	.02987
		2	204	.02136
		3	331	.02074
		4	277	.01682
		5	271	.01627
	Lowest	1	235	.00000
Cook's Distance	Highest	1	231	.01828
		2	252	.01825
		3	194	.01631
		4	334	.01482
		5	49	.01401
	Lowest	1	193	.00000

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

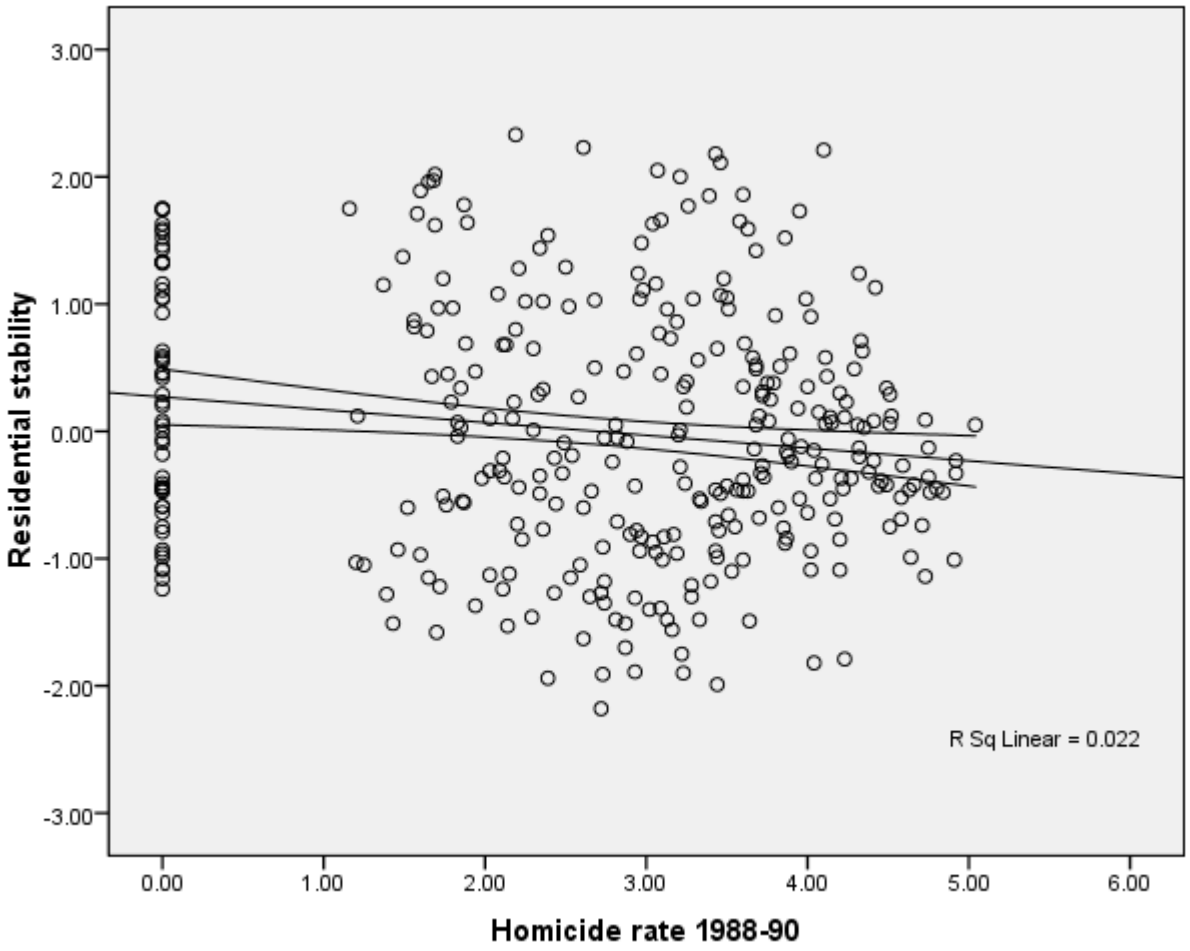
b. Dependent Variable: Residential stability

Coefficients[a]

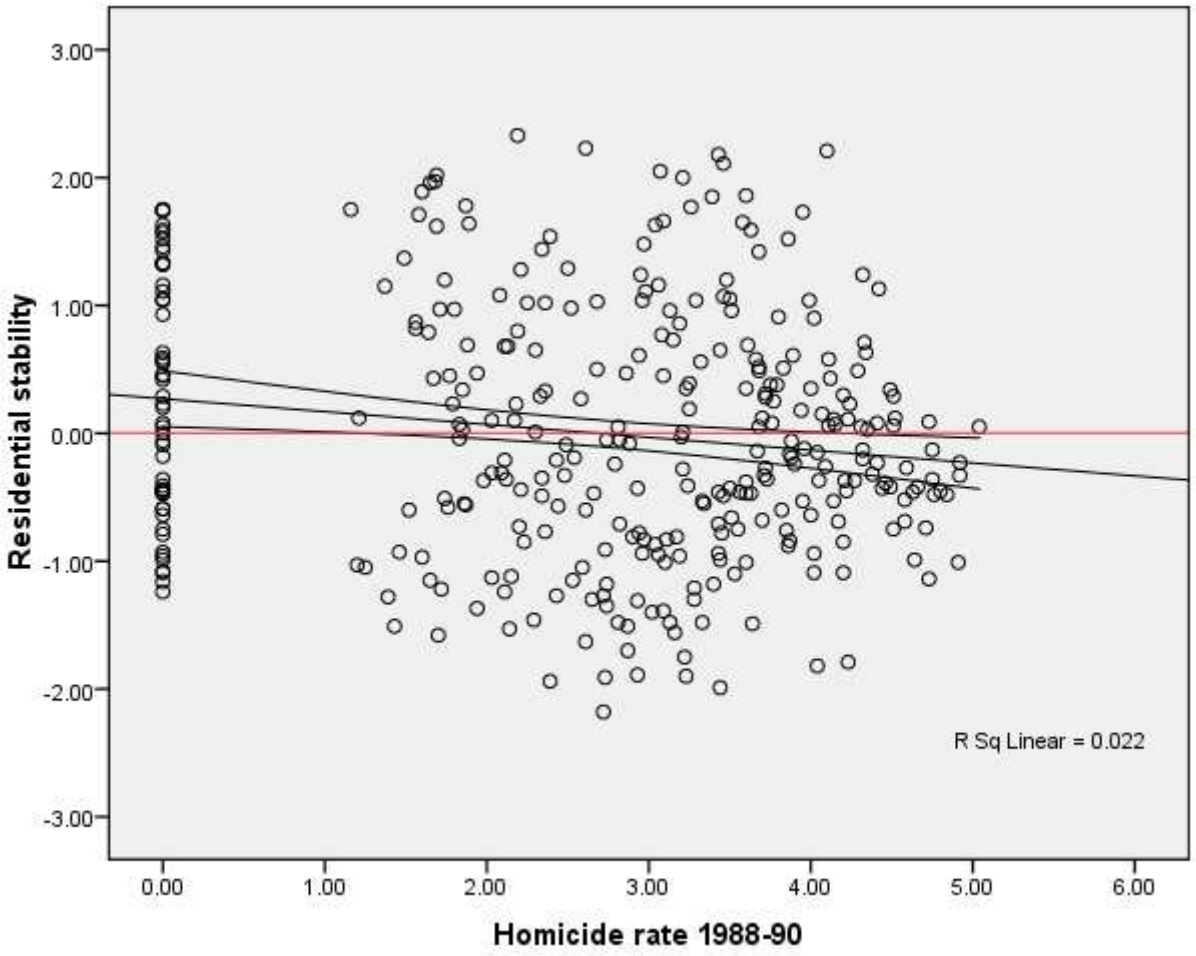
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.270	.111		2.432	.016	.052	.489
	Homicide rate 1988-90	-.100	.037	-.147	-2.735	.007	-.173	-.028

a. Dependent Variable: Residential stability

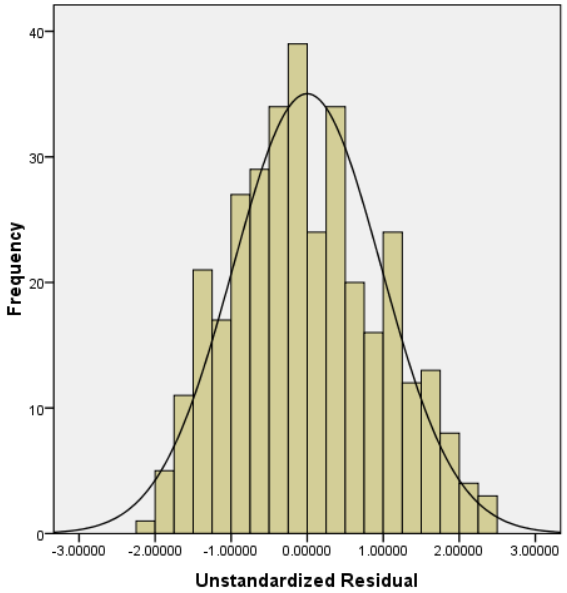
Human Development in Chicago Neighborhoods (Neighborhoods.sav)

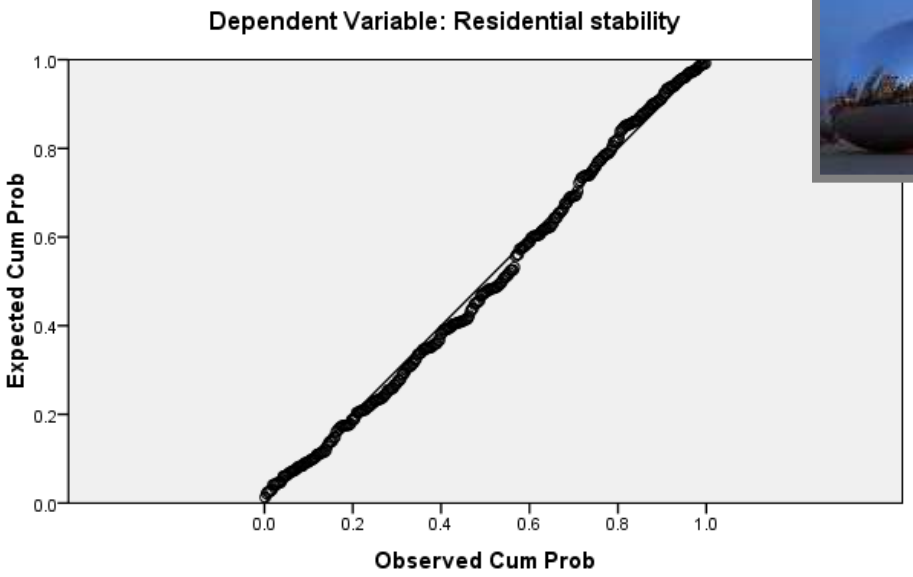


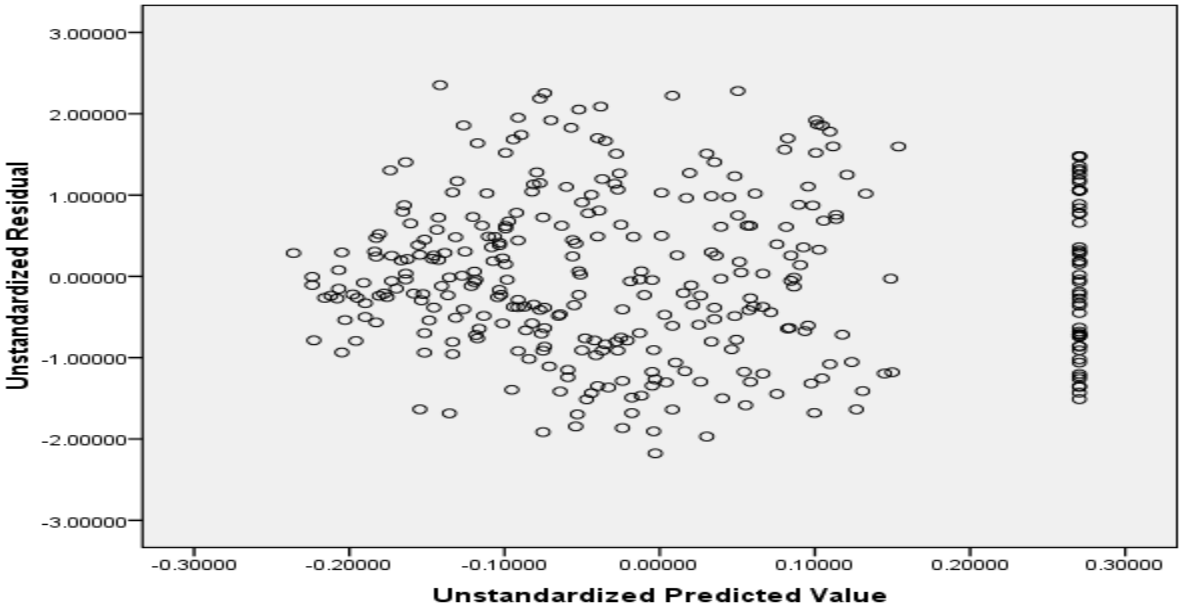
Human Development in Chicago Neighborhoods (Neighborhoods.sav)



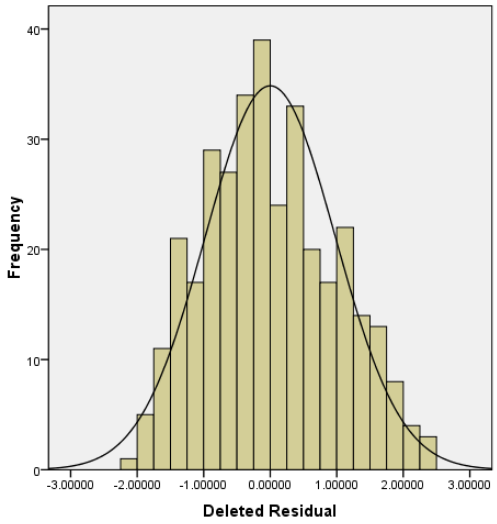
Human Development in Chicago Neighborhoods (Neighborhoods.sav)

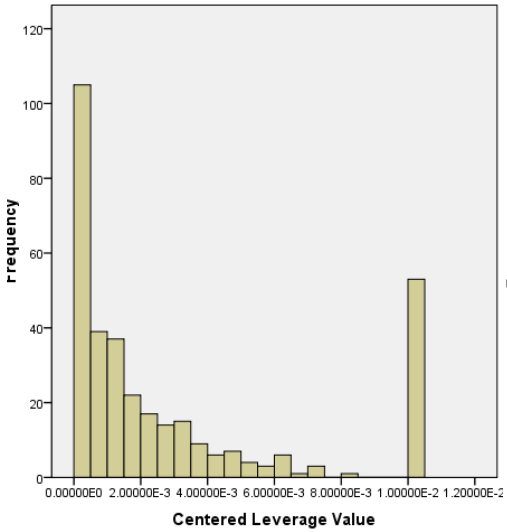


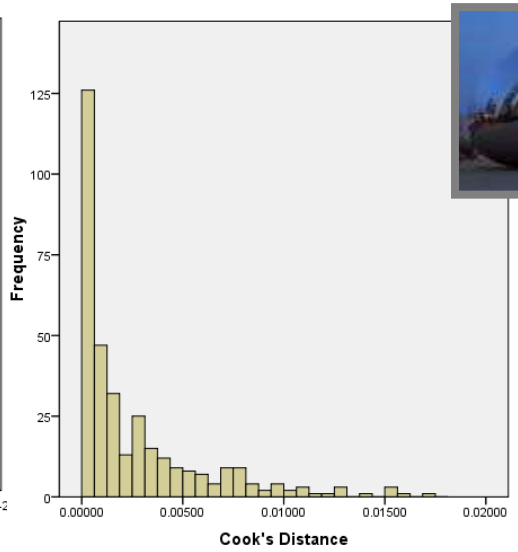




Human Development in Chicago Neighborhoods (Neighborhoods.sav)



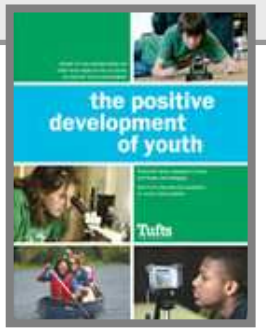




Extreme Values

			Case Number	Value
Deleted Residual	Highest	1	303	2.36537
		2	334	2.28704
		3	246	2.26271
		4	253	2.22835
		5	240	2.19562
	Lowest	1	53	-2.18352
		2	4	-1.97627
		3	28	-1.92204
		4	42	-1.91172
		5	194	-1.87168
Centered Leverage Value	Highest	1	10	.01008
		2	11	.01008
		3	12	.01008
		4	14	.01008
		5	17	.01008[a]
	Lowest	1	110	.00000
Cook's Distance	Highest	1	303	.01721
		2	11	.01601
		3	321	.01537
		4	337	.01537
		5	319	.01516
	Lowest	1	218	.00000

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



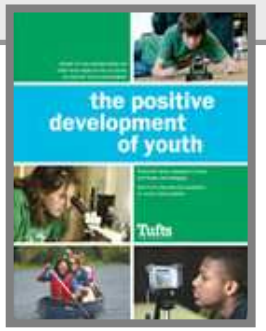
- 4-H Study of Positive Youth Development
- Source: Subset of data from IARYD, Tufts University

- 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

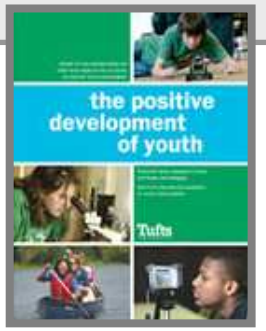
b. Dependent Variable: Self-Worth

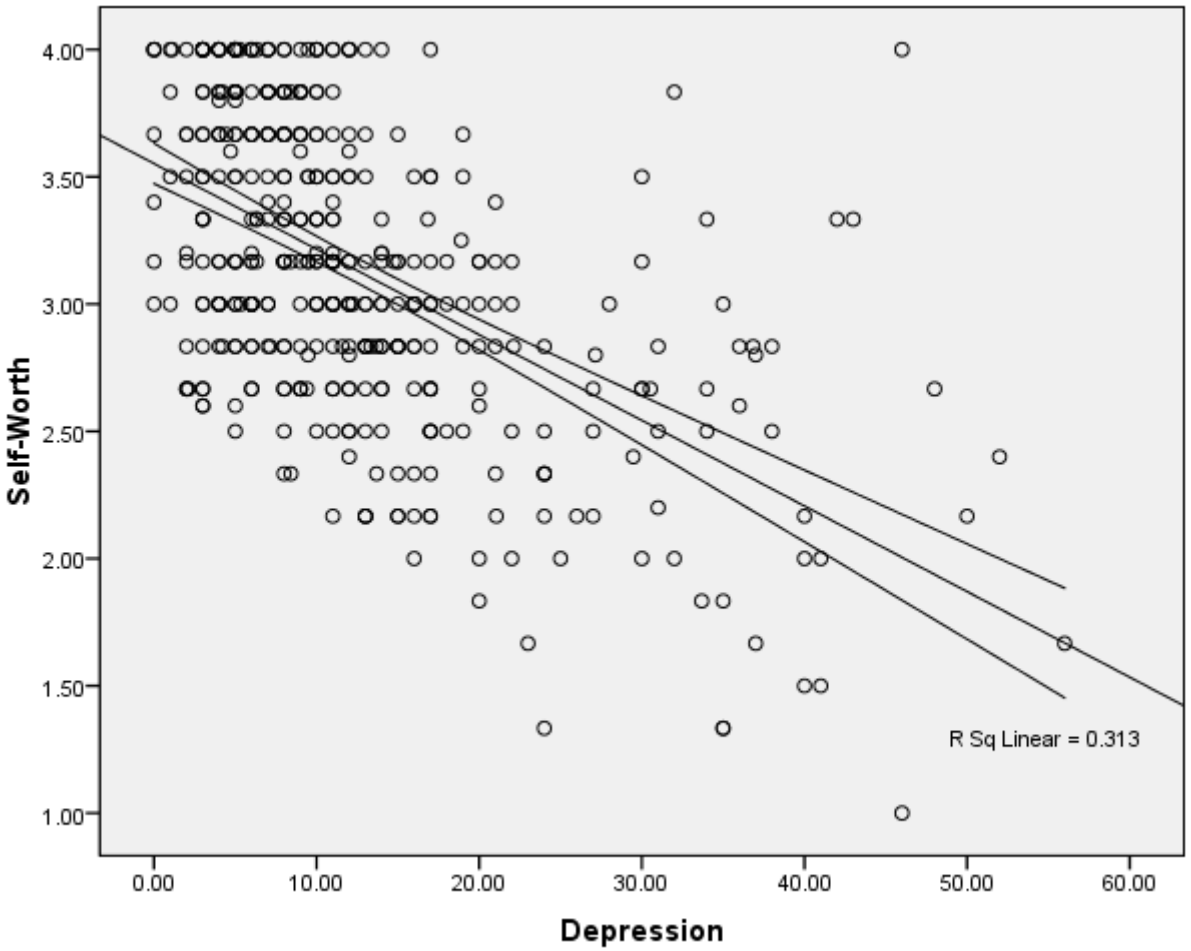
Coefficients[a]

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

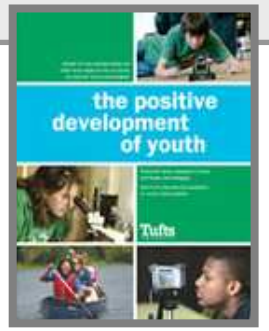
a. Dependent Variable: Self-Worth

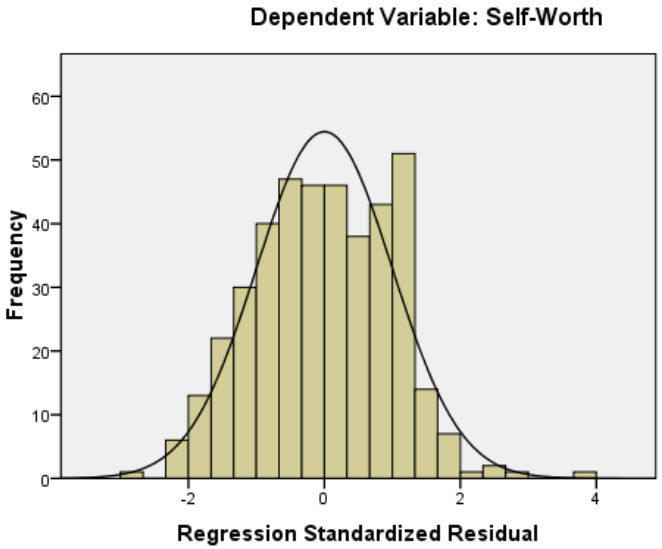
4-H Study of Positive Youth Development (4H.sav)

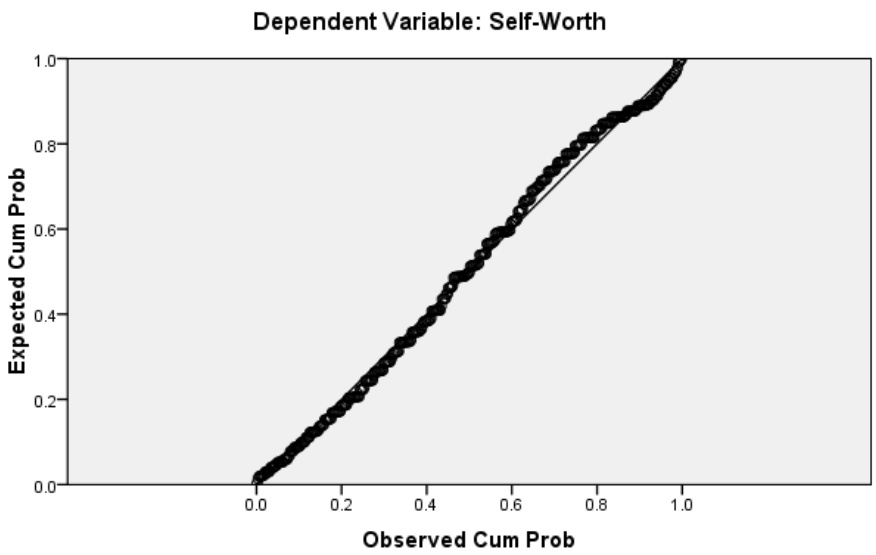


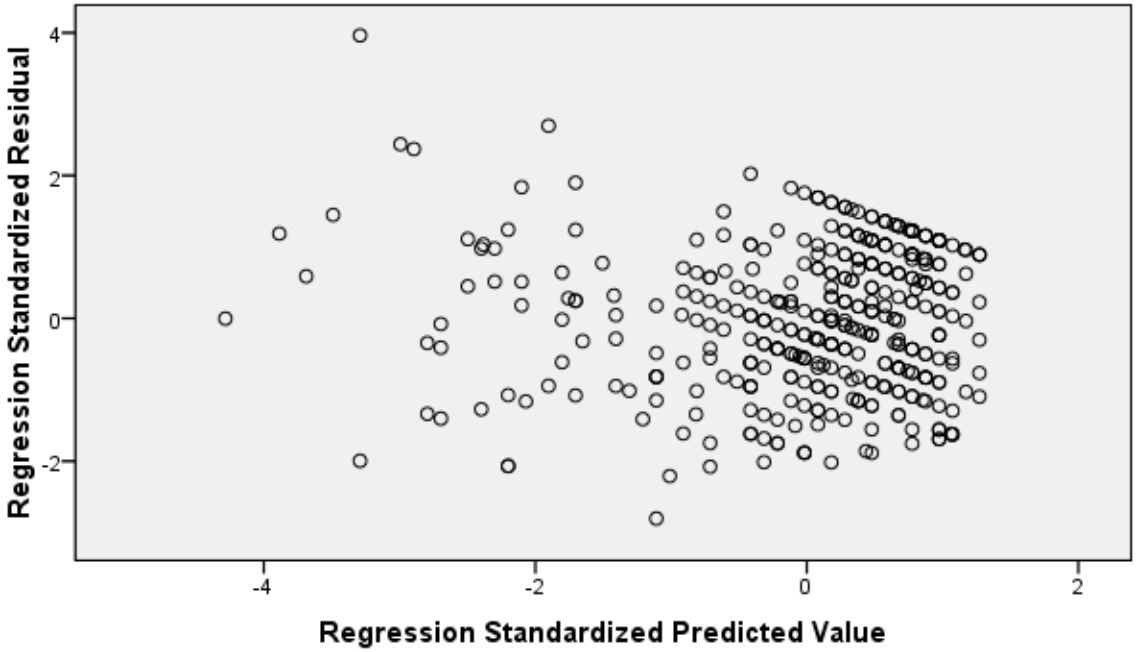


4-H Study of Positive Youth Development (4H.sav)

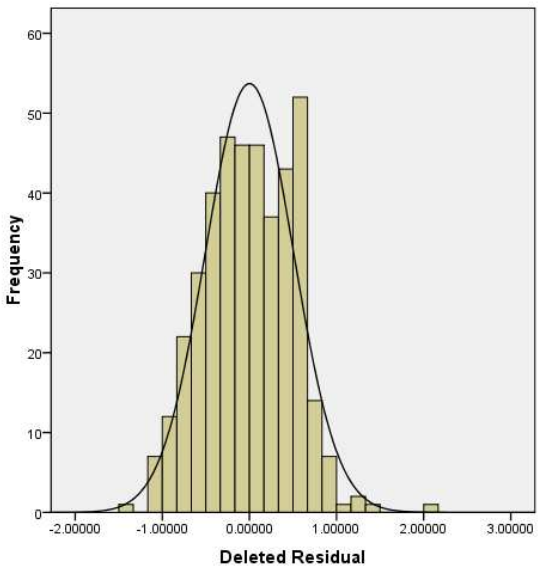


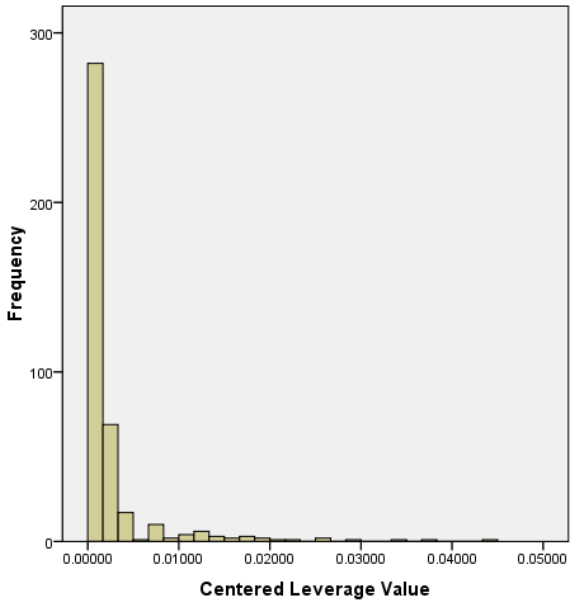


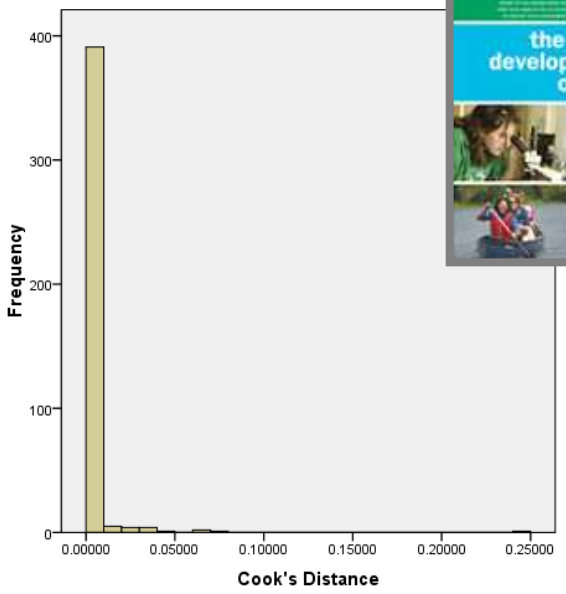


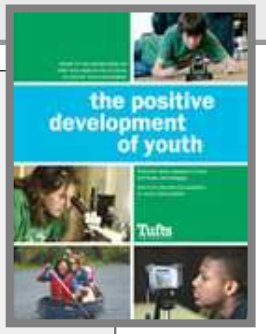


4-H Study of Positive Youth Development (4H.sav)









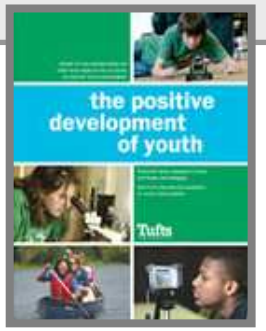
Extreme Values

			Case Number	Value
Deleted Residual	Highest	1	286	2.05470
		2	259	1.37311
		3	404	1.25826
		4	368	1.22200
		5	232	1.02264
	Lowest	1	388	-1.41926
		2	350	-1.11734
		3	300	-1.05665
		4	138	-1.05665
		5	56	-1.04992

Centered Leverage Value	Highest	1	359	.04497
		2	365	.03702
		3	401	.03334
		4	351	.02985
		5	286	.02655[a]
	Lowest	1	398	.00000

Cook's Distance	Highest	1	286	.24152
		2	404	.07625
		3	368	.06770
		4	370	.06127
		5	259	.04210
	Lowest	1	205	.00000

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

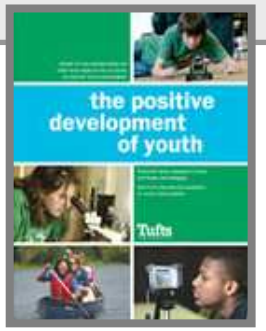
b. Dependent Variable: Self-Worth

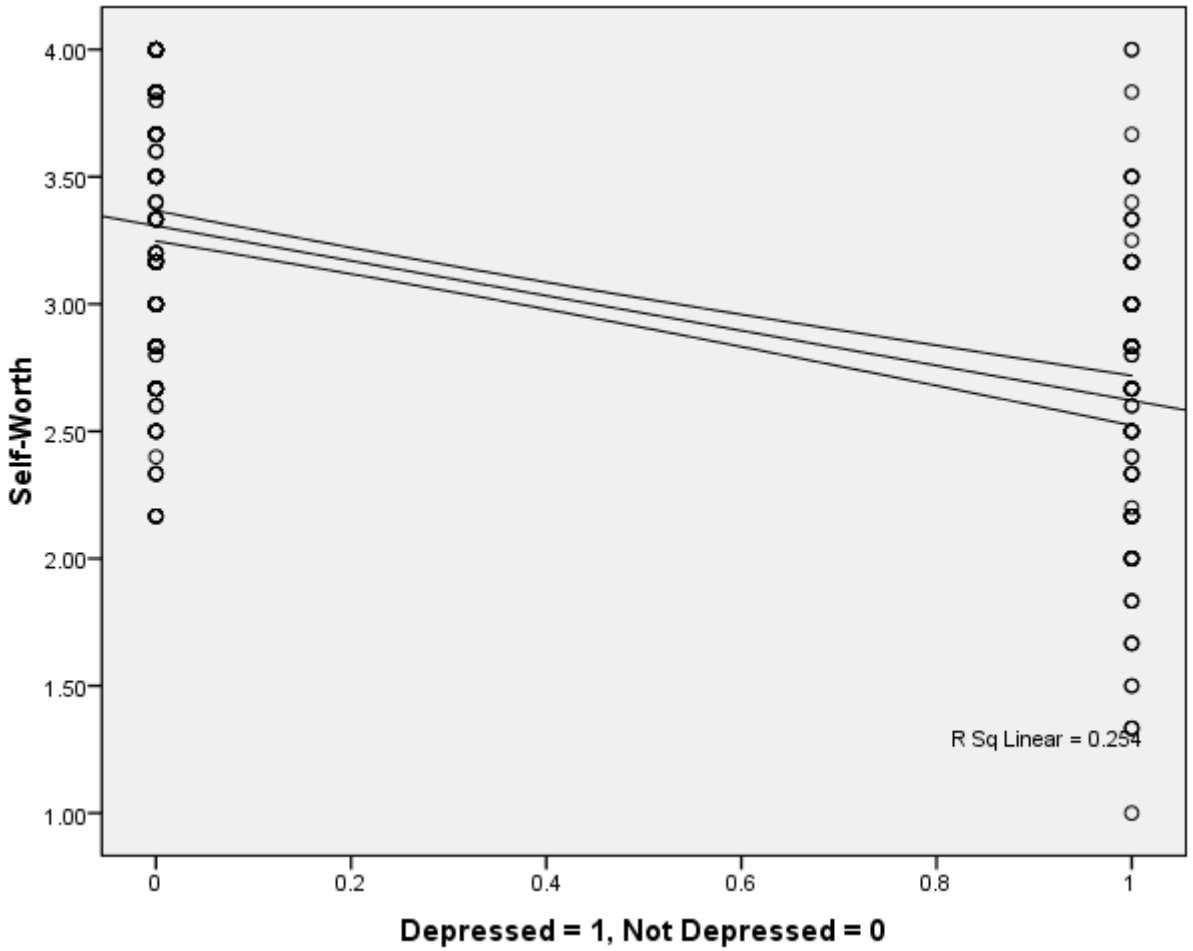
Coefficients[a]

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

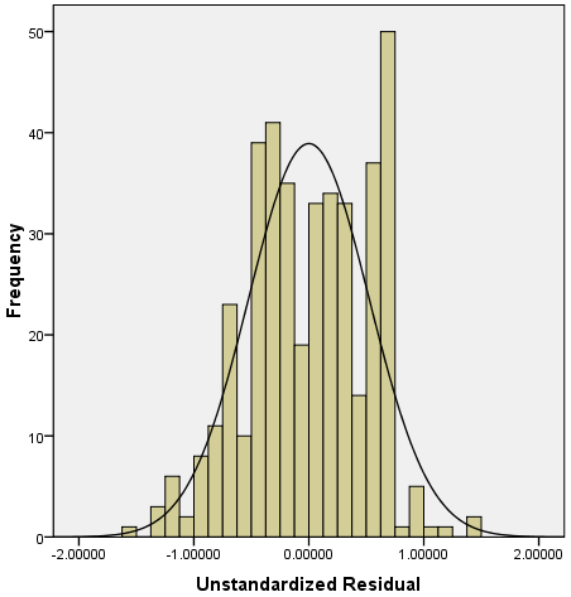
a. Dependent Variable: Self-Worth

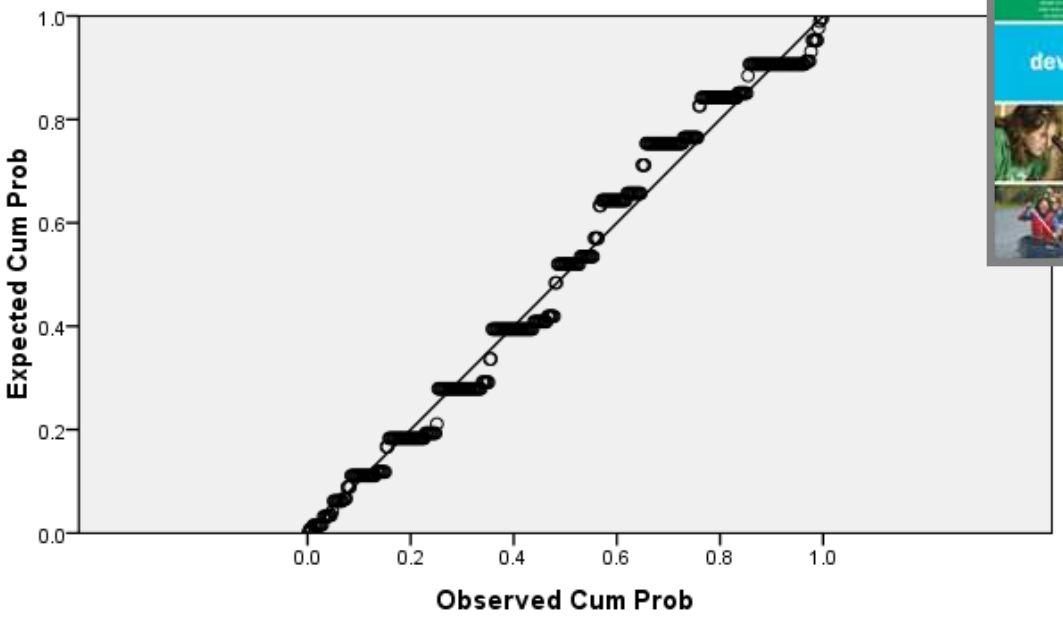
4-H Study of Positive Youth Development (4H.sav)

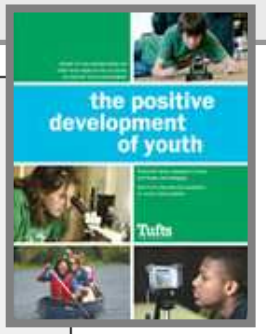




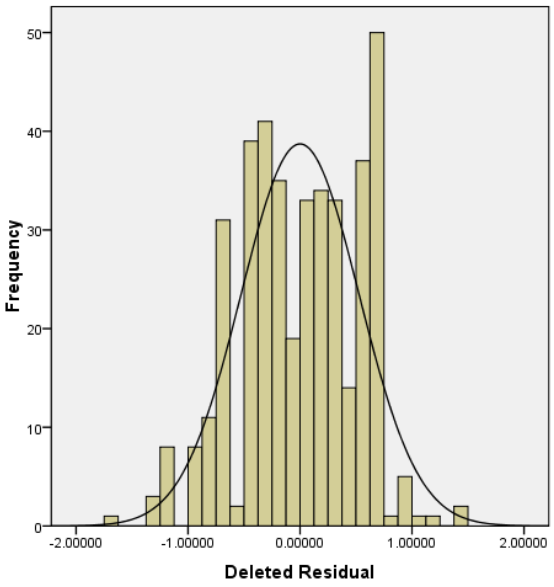
4-H Study of Positive Youth Development (4H.sav)

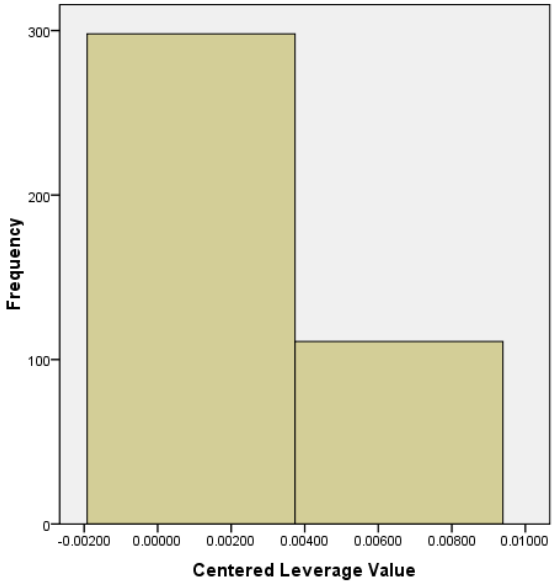


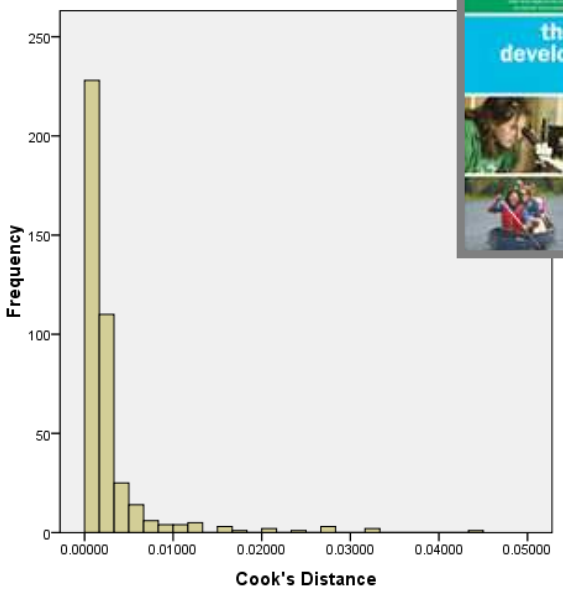


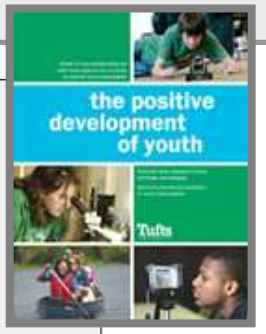


4-H Study of Positive Youth Development (4H.sav)









Extreme Values

			Case Number	Value
Deleted Residual	Highest	1	232	1.39136
		2	286	1.39136
		3	259	1.22318
		4	343	1.05500
		5	130	.88682[a]
	Lowest	1	370	-1.63591
		2	388	-1.29955
		3	300	-1.29955
		4	138	-1.29955
		5	398	-1.1442E0

Centered Leverage Value	Highest	1	11	.00656
		2	17	.00656
		3	18	.00656
		4	19	.00656
		5	21	.00656[c]
	Lowest	1	409	.00091

Cook's Distance	Highest	1	370	.04380
		2	232	.03169
		3	286	.03169
		4	138	.02764
		5	300	.02764[e]
	Lowest	1	383	.00000