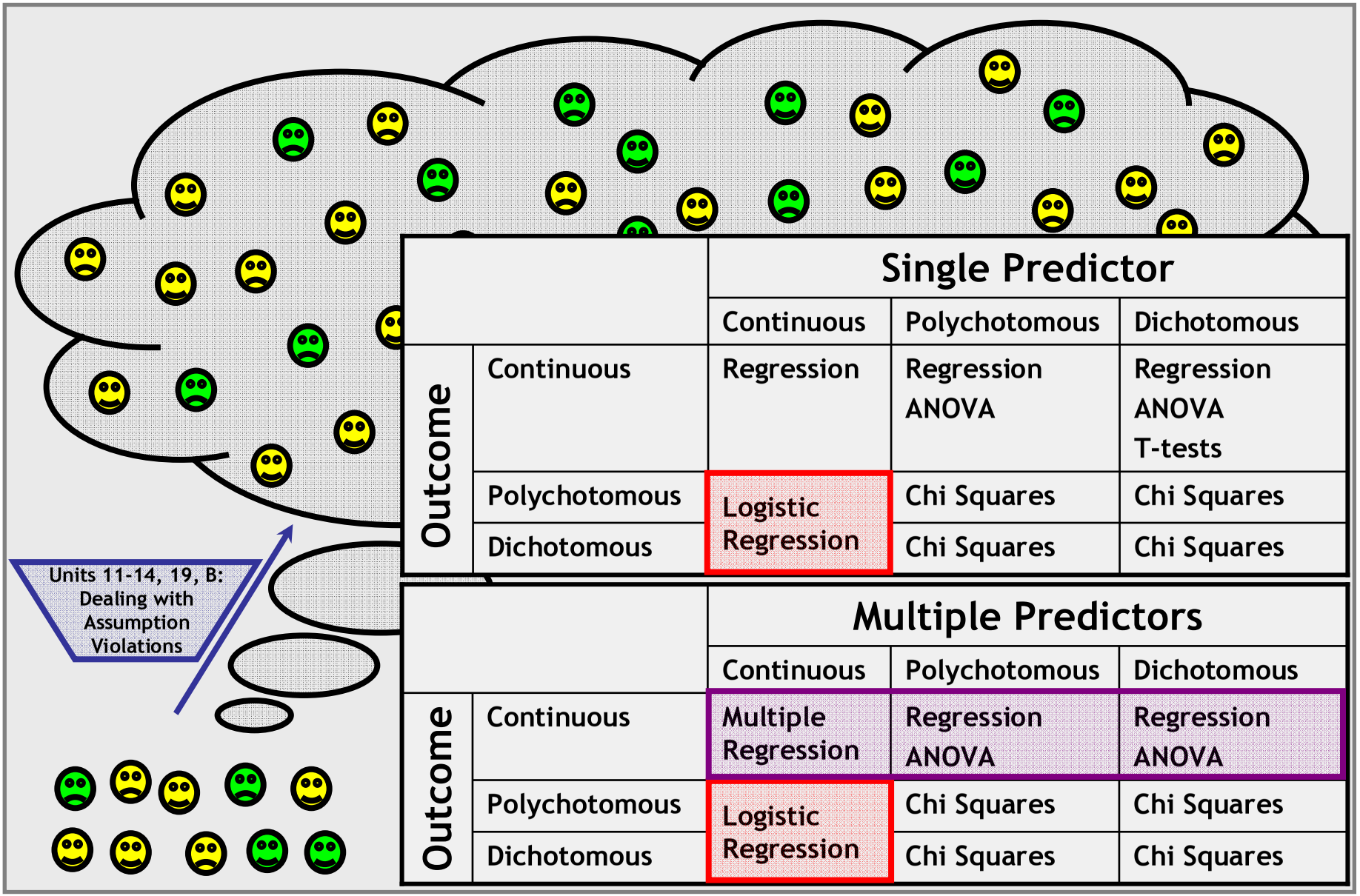
Unit 13: Road Map (VERBAL)

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 13: Road Map (Schematic)



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

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

Unit 13: 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 18

Unit 16

Unit 12

Unit 11

Unit 13

Unit 17

Unit 14

Unit 15

Unit 13: Non-Linear Transformations

Unit 13 Post Hole:

Propose a non-linear transformation, if necessary, to meet the normality and linearity assumptions of the general linear model.

Unit 13 Technical Memo and School Board Memo:

Use simple linear regression to describe a non-linear relationship between two variables (from a provided data set), and graph your results using spreadsheet software.

Unit 13 Review:

Review Unit 3.

Unit 13 Reading:

Meyers et al., Chapters 4a and 4b.

Unit 13: 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 13: 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)
    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)

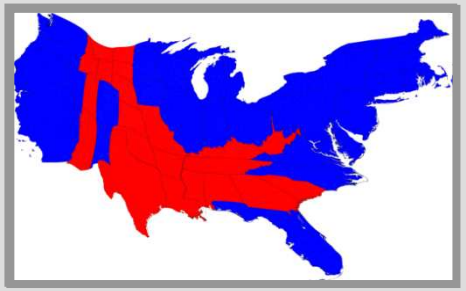
III. Memo Metacognitive

Unit 13: Research Question

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

Research Question: Controlling for percentage of SAT takers, which states perform best on the SAT?

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



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

Variables:

Outcome—State Average SAT Score (*SAT*)

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

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

SAT.sav Codebook

SAT Scores by State

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

Dataset entered on: 2005-09-07

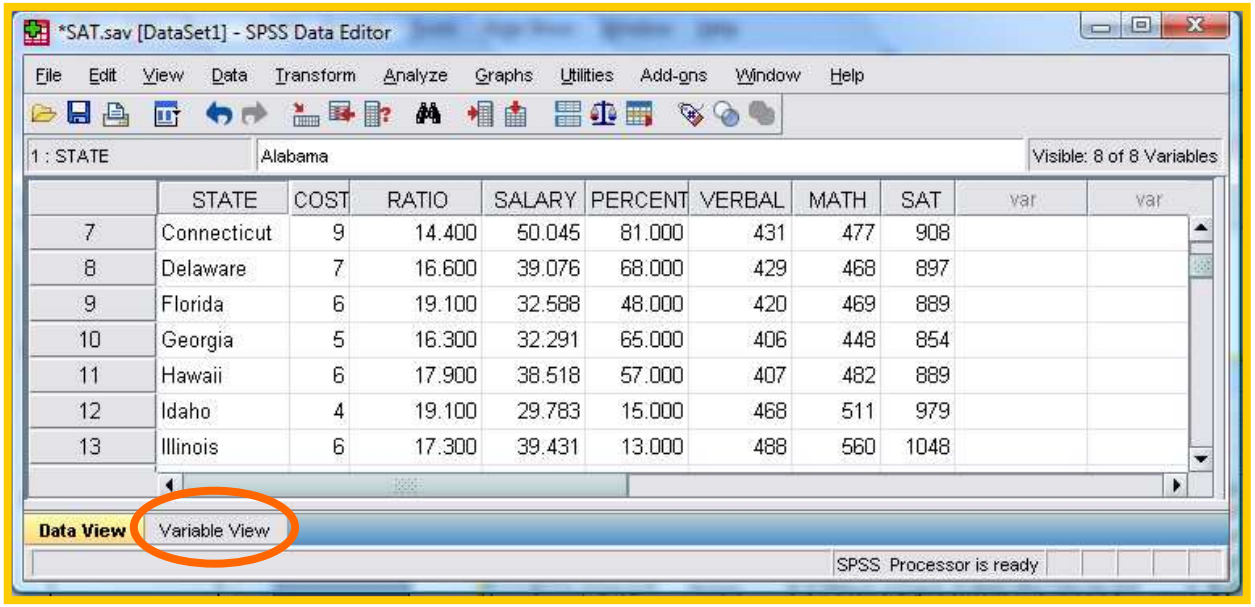
Summary
Is School Performance Related to Spending? This data set provides an example of the types of data that public policy makers consider when making decisions and crafting arguments.

Sample: The 50 United States, 1994-95.

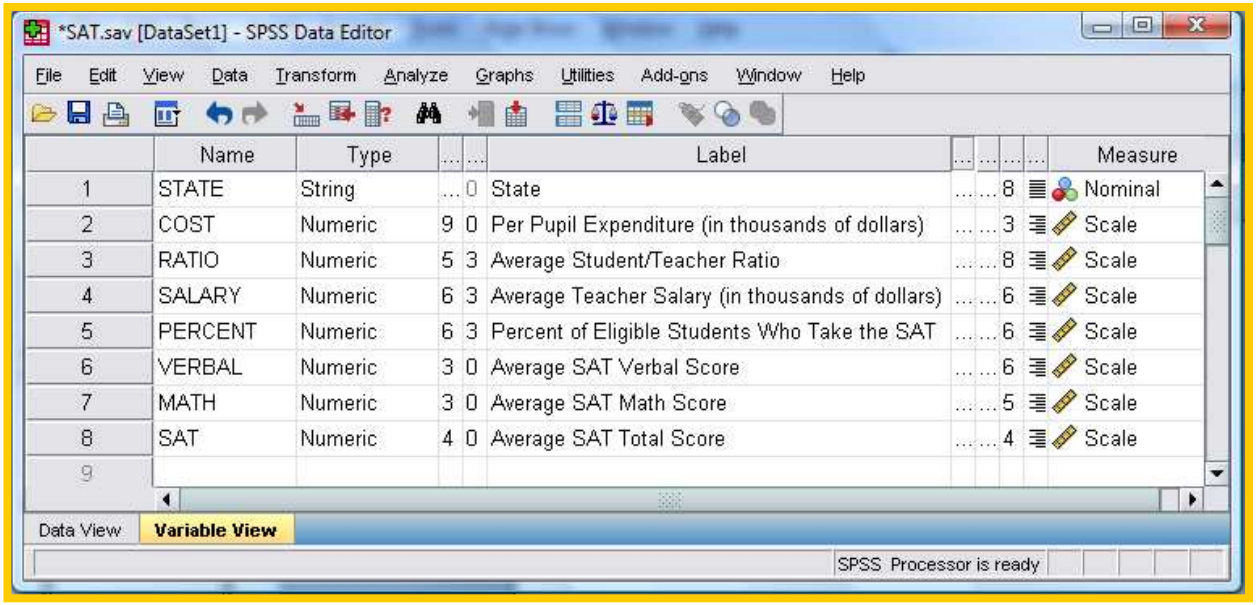
Documentation
This data set includes eight variables:

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

The SAT Data Set

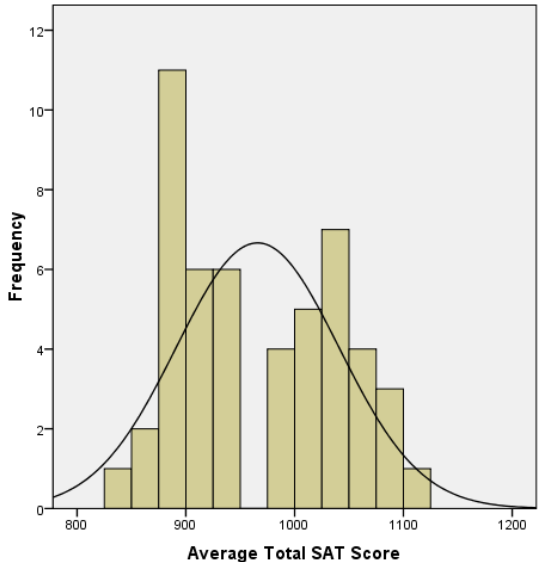


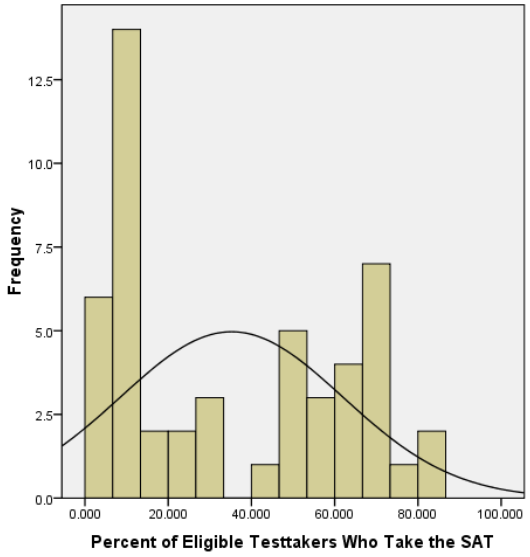
	STATE	COST	RATIO	SALARY	PERCENT	VERBAL	MATH	SAT	var	var
7	Connecticut	9	14.400	50.045	81.000	431	477	908		
8	Delaware	7	16.600	39.076	68.000	429	468	897		
9	Florida	6	19.100	32.588	48.000	420	469	889		
10	Georgia	5	16.300	32.291	65.000	406	448	854		
11	Hawaii	6	17.900	38.518	57.000	407	482	889		
12	Idaho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

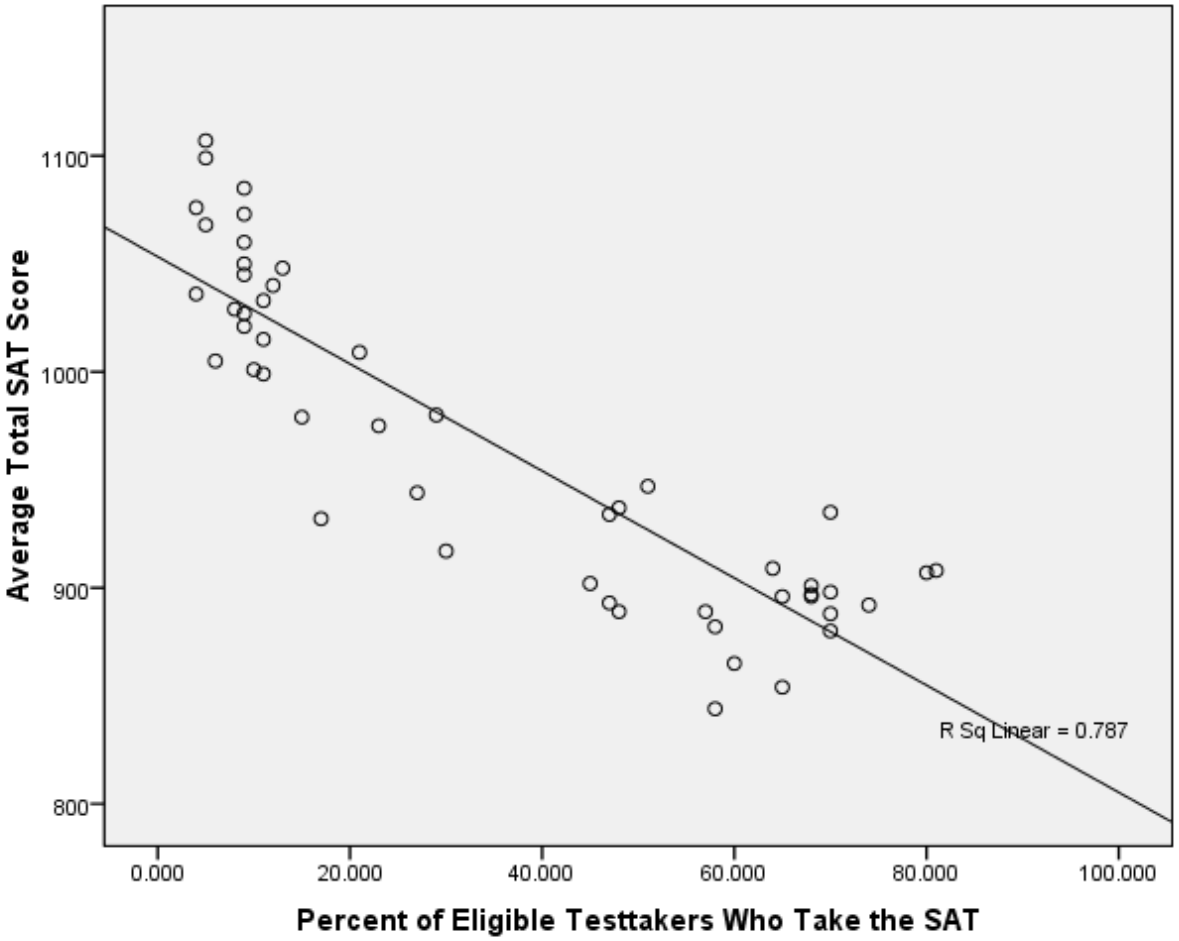


	Name	Type			Label				Measure
1	STATE	String	...	0	State	...	...	8	Nominal
2	COST	Numeric	9	0	Per Pupil Expenditure (in thousands of dollars)	...	...	3	Scale
3	RATIO	Numeric	5	3	Average Student/Teacher Ratio	...	...	8	Scale
4	SALARY	Numeric	6	3	Average Teacher Salary (in thousands of dollars)	...	...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

Exploratory Graphs





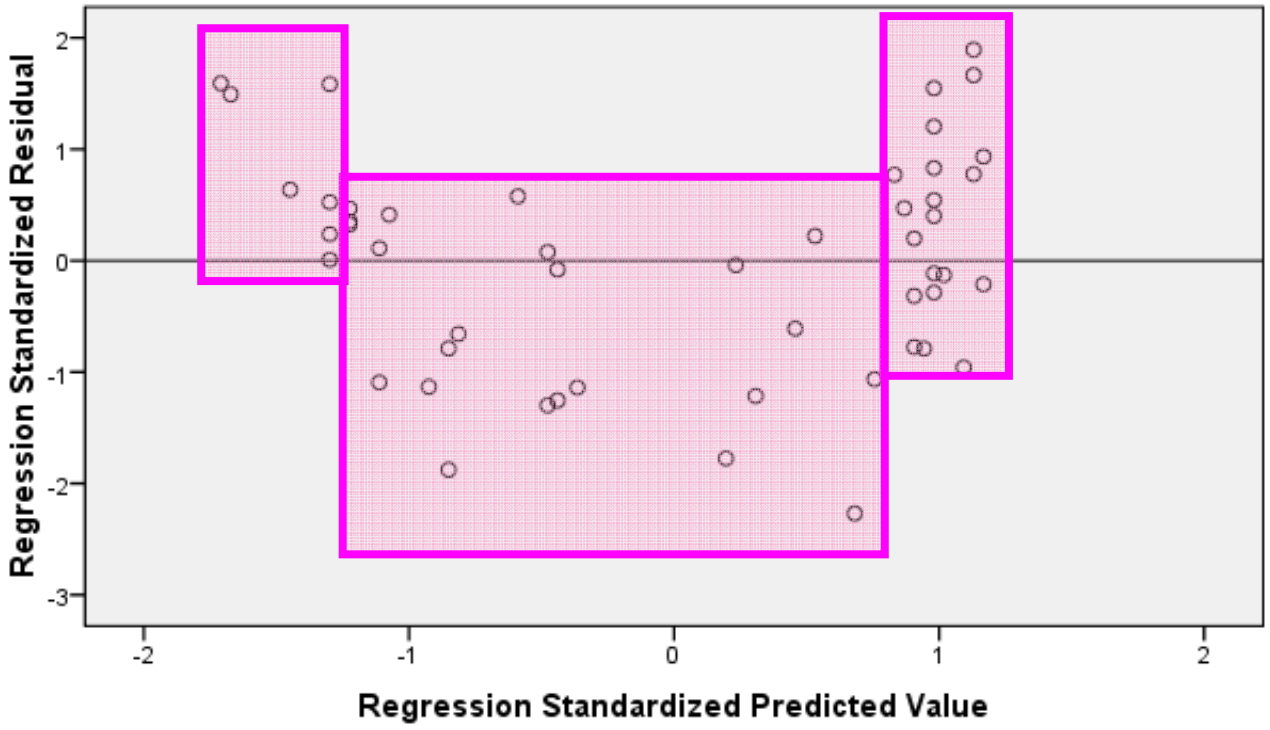


A skewed distribution in the outcome and/or predictor sometimes (but not always) foreshadows a non-linear relationship.

Fitting the Wrong Model

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	1053.320	8.211		128.278	.000	1036.811	1069.830
	Percent of Eligible Testtakers Who Take the SAT	-2.480	.186	-.887	-13.317	.000	-2.855	-2.106

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



Just because a relationship is statistically significant does not mean that you are modeling the right relationship.

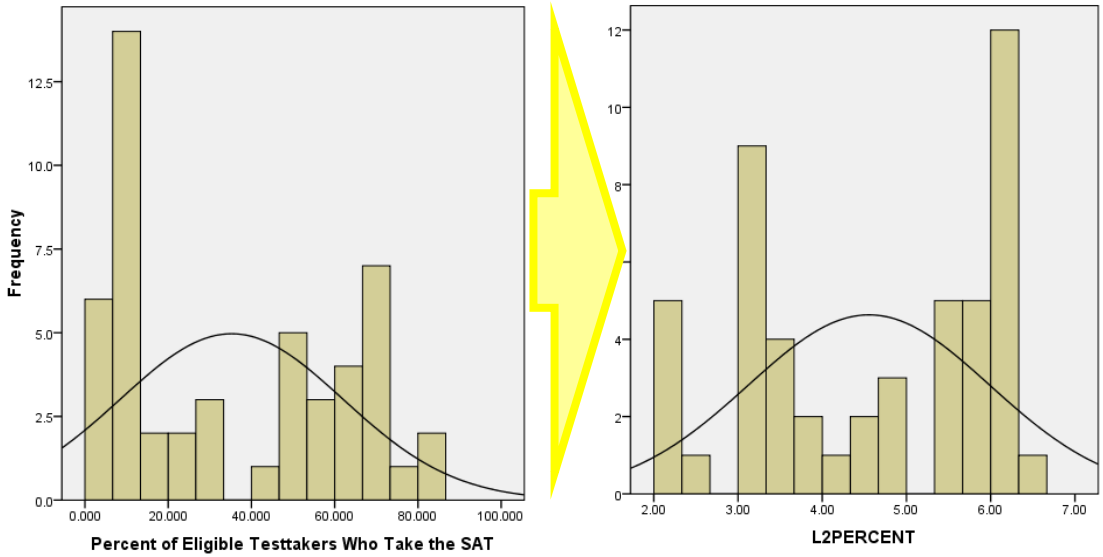
Just because a correlation is high does not mean the correlation is right.

A horseshoe pattern in the residual versus fitted (RVF) plot indicates a violation of the GLM linearity assumption. Our goal is to have a patternless cloud, but a concave up horsehoe pattern indicates that for our low predictions we are underestimating, for our middling predictions we are overestimating, and for our high predictions we are underestimating.

Logarithmic Transforming to Achieve Linearity

$PERCENT$

$$L2PERCENT = Log_2(PERCENT) = \frac{Log_{10}(PERCENT)}{Log_{10}(2)}$$



*Annoyingly, SPSS only does logarithmic functions to the base 10 (i.e., orders of magnitude) and the base *e* (i.e., the base 2.71828…, or the "natural" log).

*Fortunately, we can use properties of logarithms to get base 2.

COMPUT E L2PERCENT = LG10(PERCENT)/LG10(2).

EXECUTE.

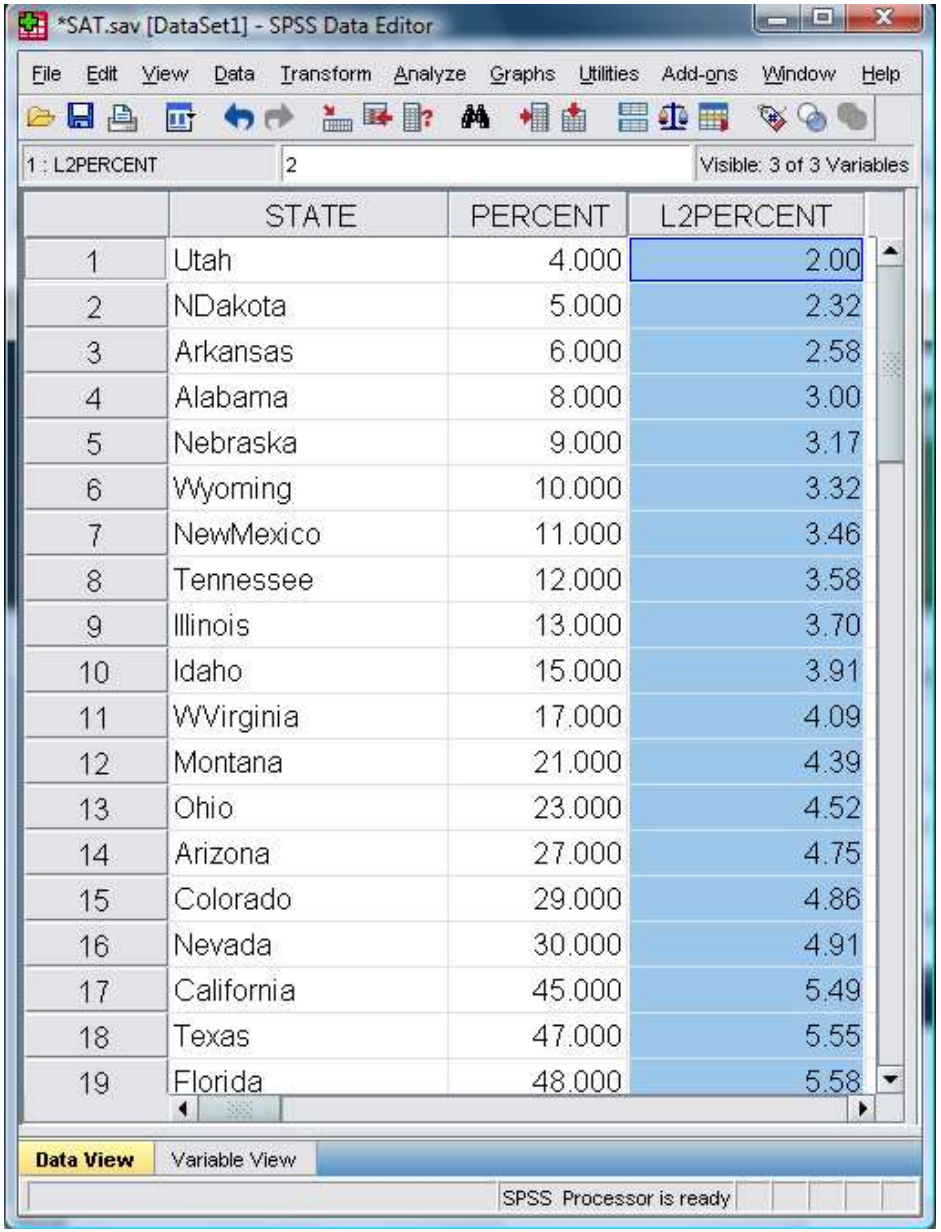
When we get to interpretations, you will see why I want to use base 2 rather than base 10 or base *e* for my logarithmic transformation.

If $x = b^y$ then, $y = \log_b(x)$

Logarithmic transformations pull in long upper tails.

When I logarithmically transform a variable to the base 2, I ask of each value of the variable, "By what power must I raise 2 in order to equal you?" The power by which I must raise 2 becomes the transformed value of the variable. Take Alabama, for example. Of the eligible students in Alabama, 8% take the SAT, so in the variable called *PERCENT*, Alabama has a value of 8. I ask, "What power do I need to raise 2 by in order to equal 8?" $2^1=2$. $2^2=4$. $\underline{2^3=8}$. $2^4=16$. $2^5=32$. $2^6=64$. And so on… Notice that I must raise 2 to the power of 3 to get the 8 for which I was looking. Thus, 3 is the transformed value of 8. Three is the new eight.

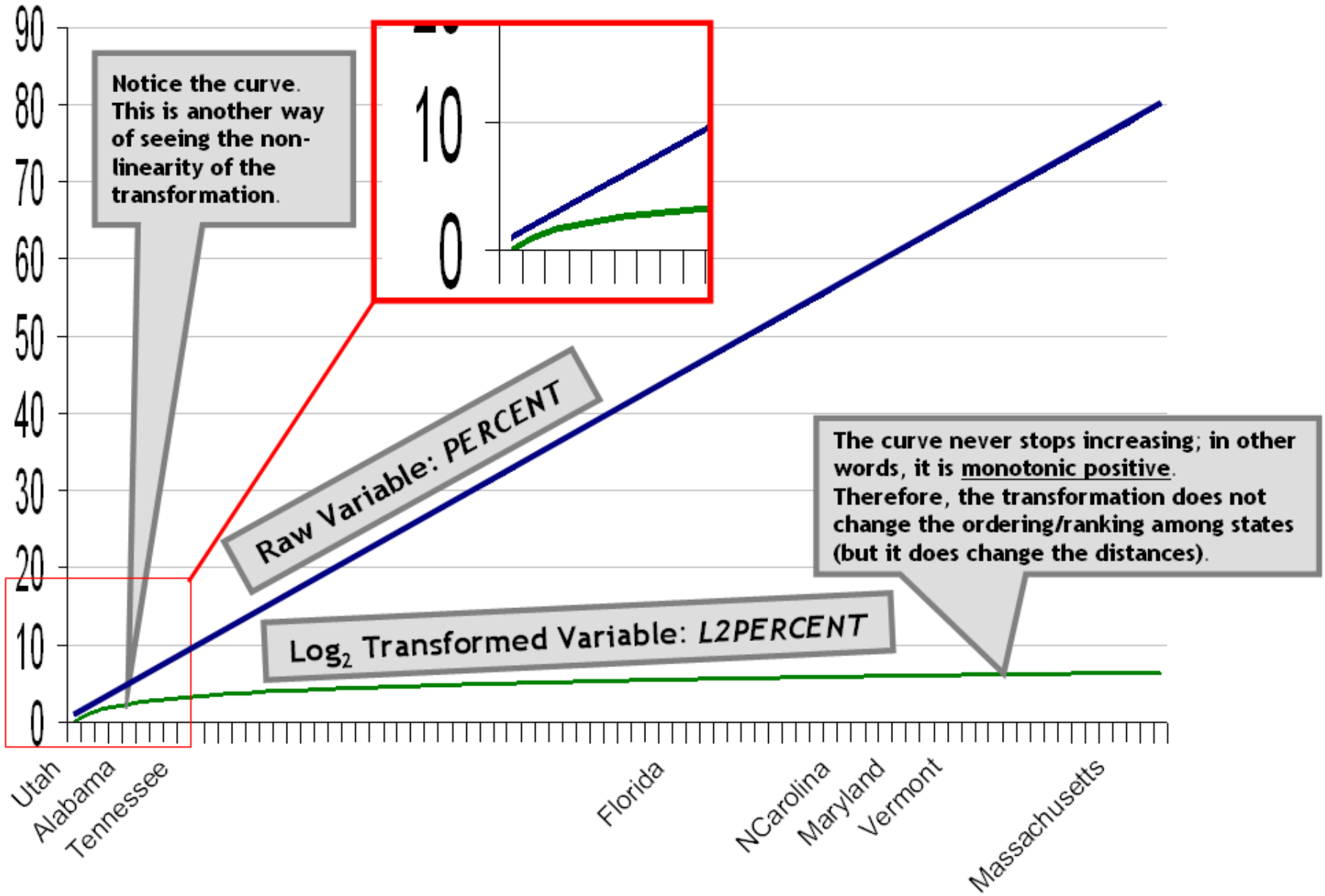
Comparing Our Raw Variable with our Transformed Variable



	STATE	PERCENT	L2PERCENT
1	Utah	4.000	2.00
2	NDakota	5.000	2.32
3	Arkansas	6.000	2.58
4	Alabama	8.000	3.00
5	Nebraska	9.000	3.17
6	Wyoming	10.000	3.32
7	NewMexico	11.000	3.46
8	Tennessee	12.000	3.58
9	Illinois	13.000	3.70
10	Idaho	15.000	3.91
11	WVirginia	17.000	4.09
12	Montana	21.000	4.39
13	Ohio	23.000	4.52
14	Arizona	27.000	4.75
15	Colorado	29.000	4.86
16	Nevada	30.000	4.91
17	California	45.000	5.49
18	Texas	47.000	5.55
19	Florida	48.000	5.58

Logarithmic transformations are nonlinear because they hit bigger numbers harder. Notice that Alabama's 8 becomes a 3, because $8 = 2^3$. However, California's 45 becomes 5.49, because $45 = 2^{5.49}$. With logarithmic transformations, the bigger they are, the harder they fall.

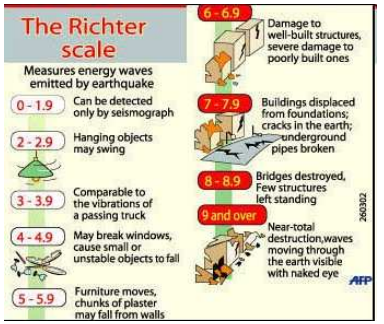
The Bigger They Are…

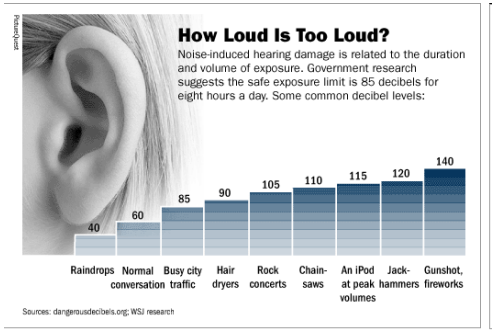


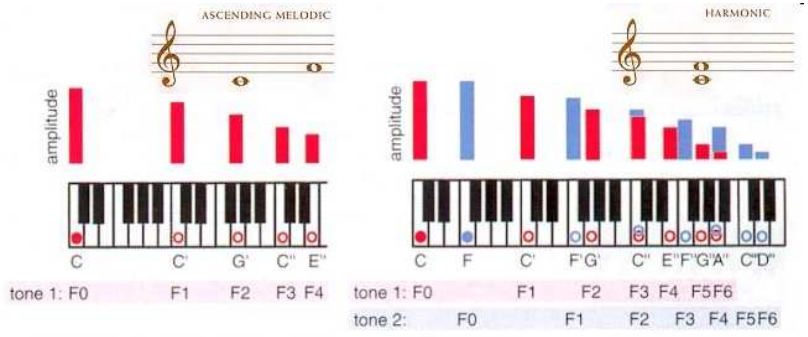
Logarithms In Everyday Life

Various quantities in science are expressed as logarithms of other quantities; see logarithmic scale for an explanation and a more complete list. (From Wikipedia)

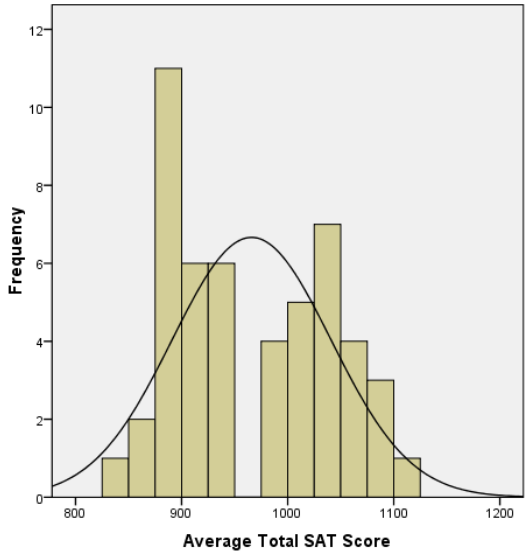
- The Richter scale measures earthquake intensity on a base-10 logarithmic scale.
- In astronomy, the apparent magnitude measures the brightness of stars logarithmically, since the eye responds approximately logarithmically to brightness.
- In psychophysics, the Weber-Fechner law proposes a logarithmic relationship between stimulus and sensation. (Thus, the magnitude of sound is measured logarithmically in decibels.)
- Musical intervals are measured logarithmically as semitones. The interval between two notes in semitones is the base-21/12 logarithm of the frequency ratio (or equivalently, 12 times the base-2 logarithm).
- Computer storage capacity (e.g., thumb drives and MP3 players) "grows" logarithmically from 2MB to 4MB to 8MB to 32MB to 64MB 128MB to 256MB to 512MB to 1024MB (or 1GB) to 2048MB (or 2GB) to 4096MB (or 4GB) to 8192MB (or 8GB) to 16384MB (or 16GB) etc.

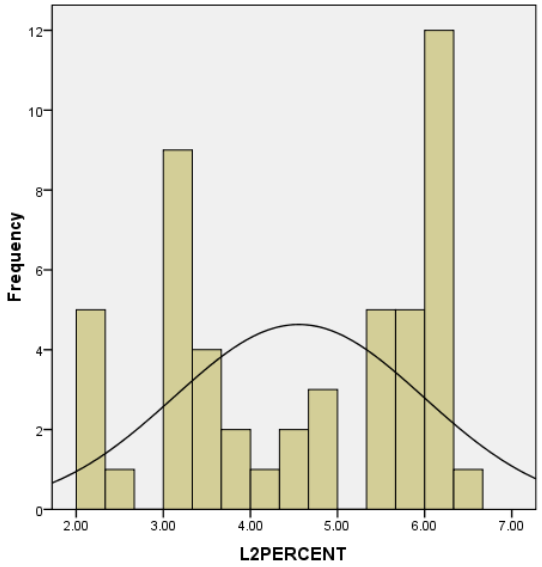


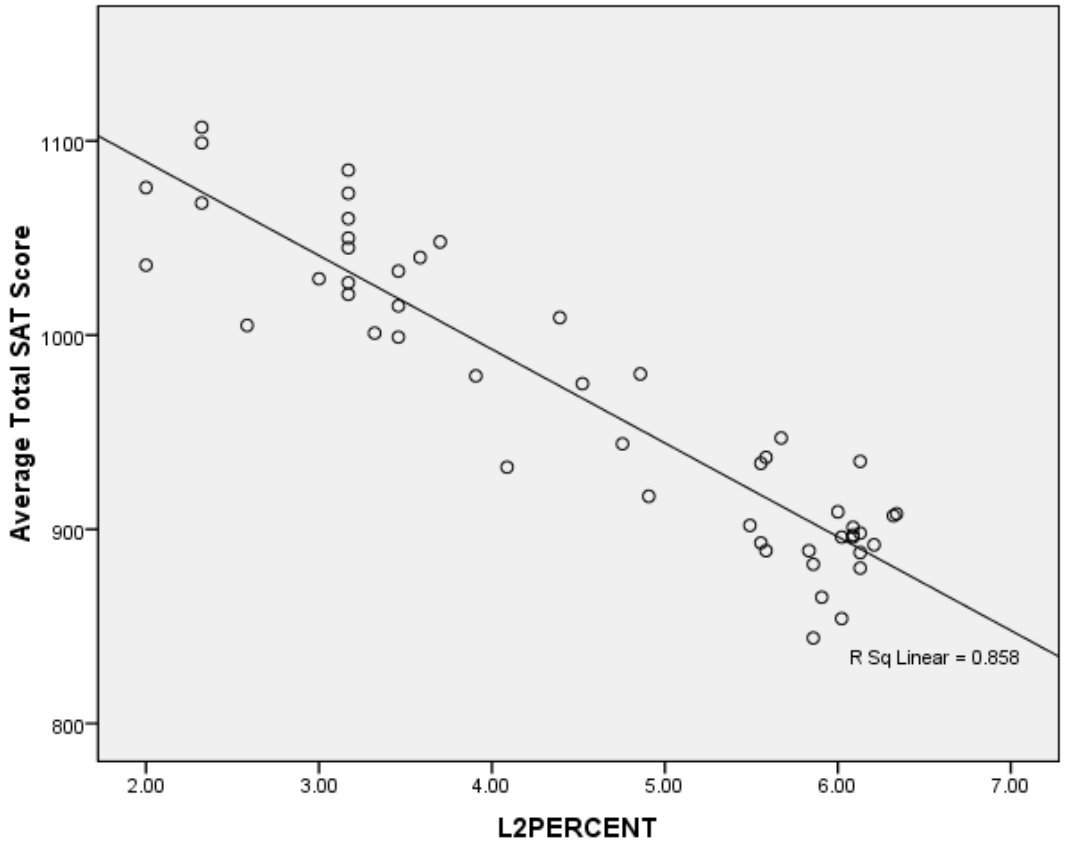




Exploratory Graphs (Second Iteration)





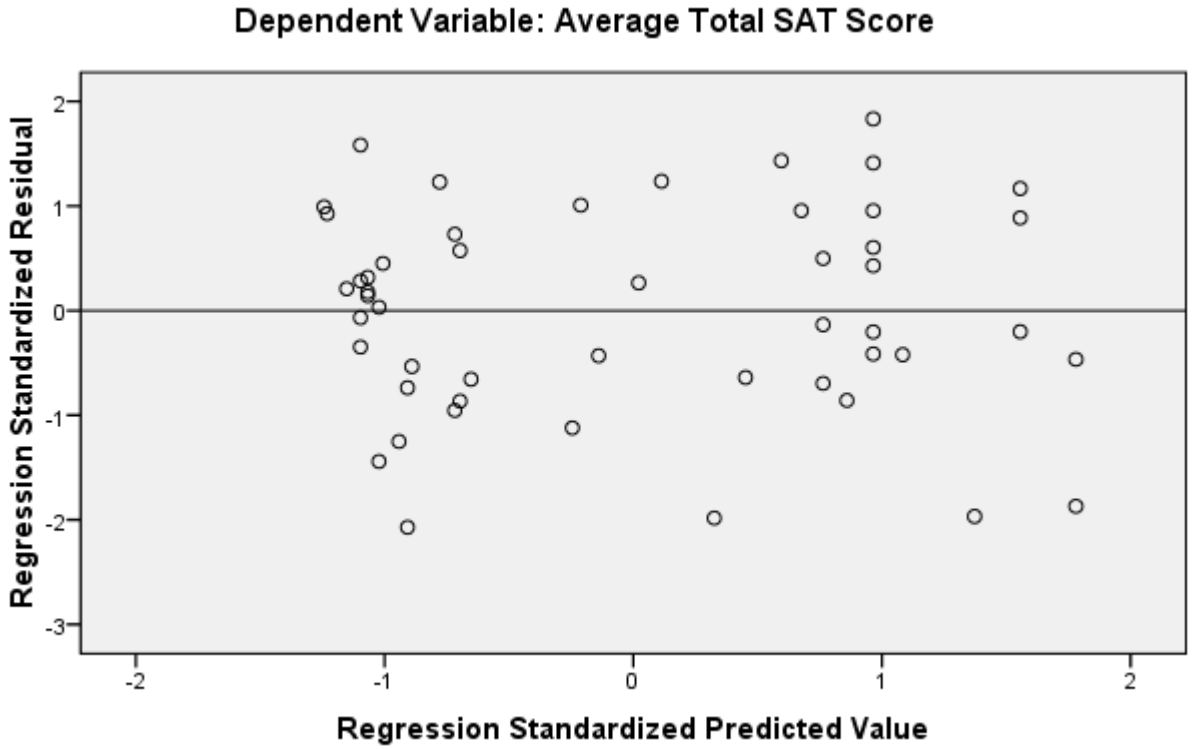


Fitting the Right Model

$$SAT = \beta_0 + \beta_1 L2PERCENT + \varepsilon$$

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	1185.841	13.530		87.645	.000	1158.637	1213.045
	L2PERCENT	-48.281	2.836	-.926	-17.027	.000	-53.983	-42.580

a. Dependent Variable: Average Total SAT Score



Top Ten States (Controlling for Percent Test Takers)

Case Summaries			
	STATE	Average Total SAT Score	Unstandardized Residual
1	Minnesota	1085	52.20646
2	NewHampshire	935	45.08787
3	Illinois	1048	40.82035
4	Wisconsin	1073	40.20646
5	Montana	1009	35.22505
6	Oregon	947	35.03022
7	NDakota	1107	33.26414
8	Colorado	980	28.70786
9	Connecticut	908	28.25430
10	Tennessee	1040	27.24497

Bottom Ten States (Controlling for Percent Test Takers)

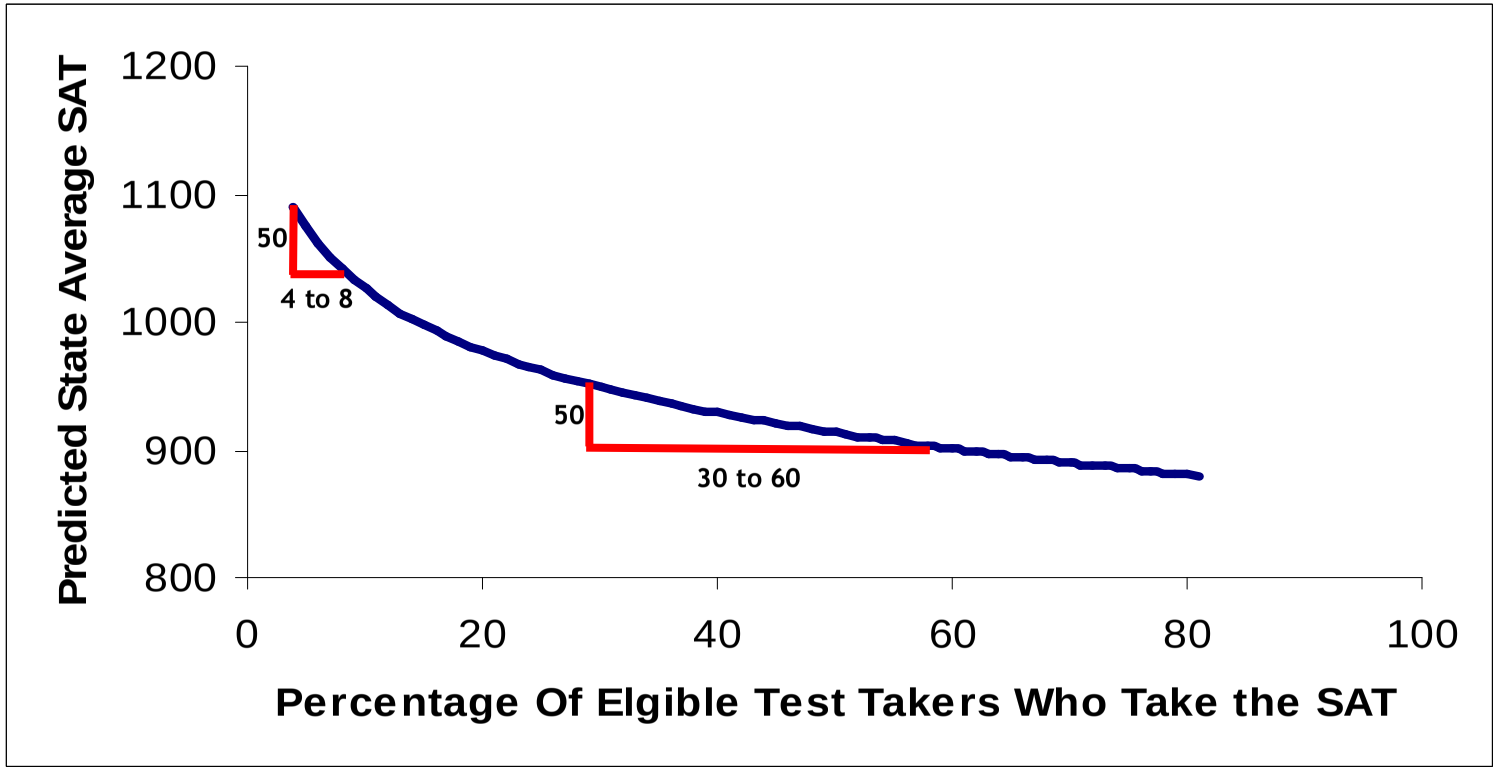
Case Summaries			
	STATE	Average Total SAT Score	Unstandardized Residual
41	Wyoming	1001	-24.45465
42	Texas	893	-24.65907
43	Florida	889	-27.19259
44	Nevada	917	-31.93073
45	NCarolina	865	-35.64951
46	Georgia	854	-41.07413
47	Mississippi	1036	-53.27894
48	Arkansas	1005	-56.03624
49	WVirginia	932	-56.49370
50	SCarolina	844	-59.01093

Middling Ten States (Controlling for Percent Test Takers)

Case Summaries			
	STATE	Average Total SAT Score	Unstandardized Residual
21	NewJersey	898	8.08787
22	Ohio	975	7.56170
23	NewYork	892	5.95859
24	Delaware	897	5.06874
25	Maine	896	4.06874
26	Virginia	896	.92587
27	RhodeIsland	888	-1.91213
28	NewMexico	1015	-3.81581
29	SDakota	1068	-5.73586
30	Oklahoma	1027	-5.79354

Interpreting Our Results

Figure 13.1: A non-linear trend line depicting the relationship between predicted state average SAT scores and the percentage of eligible test takers who take the SAT (n=50).

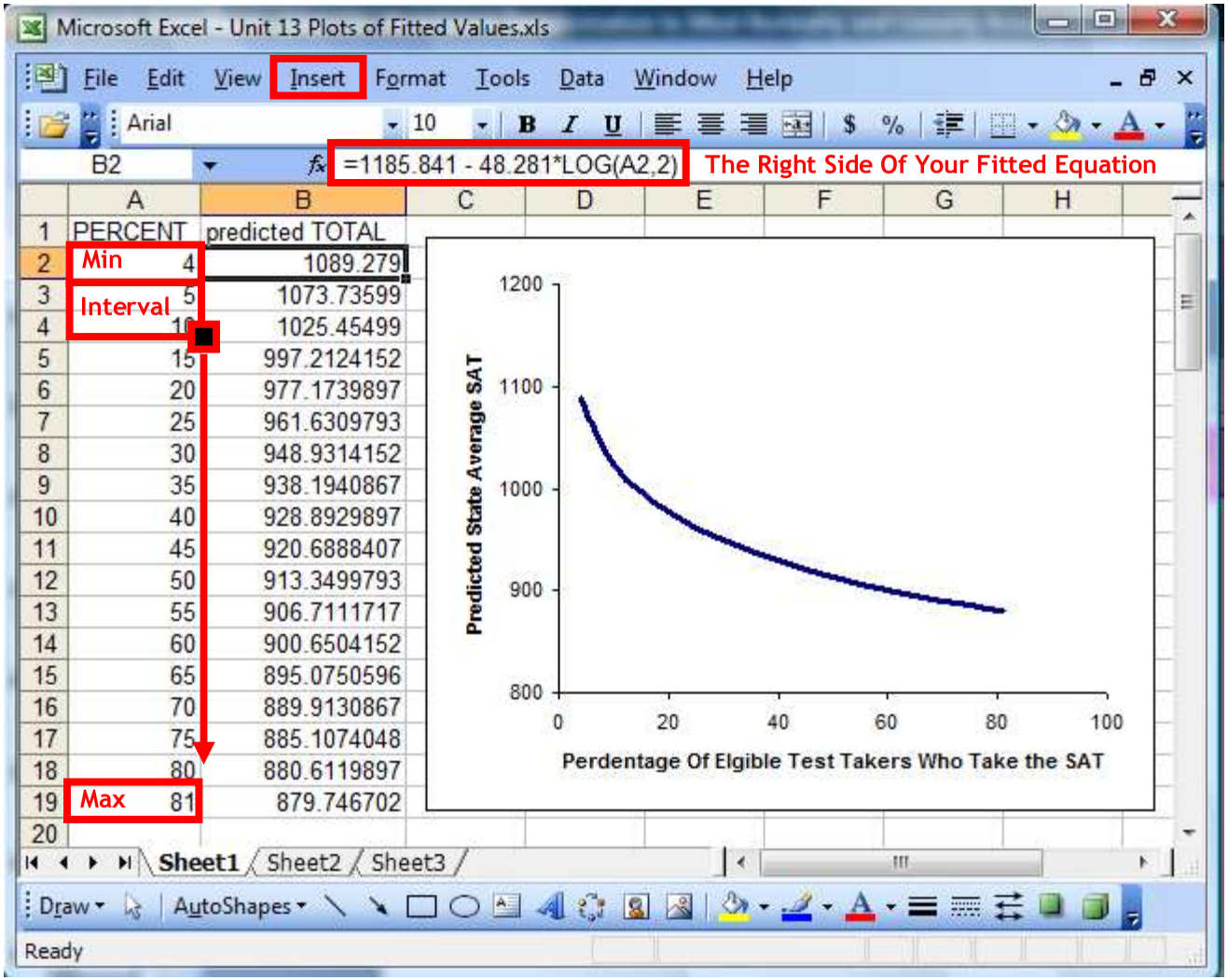


When we log transform our predictor using base 2, the slope coefficient is no longer the difference in our outcome associated with a one unit difference in our predictor, but rather the slope coefficient is now the difference in our outcome associated with a doubling of our predictor.

Given two states where one state has twice the other's percentage of eligible test takers taking the SAT, we expect the state with twice the percentage of eligible test takers taking the SAT to have an average SAT score of about 50 points less. For example, take Mississippi (*PERCENT* = 4) and Alabama (*PERCENT* = 8). Alabama has twice the percentage of Mississippi; therefore, we predict that Alabama's average SAT will be about 50 points less than Mississippi's. (In fact, it is 47 points less.) For another example, take Nevada (*PERCENT* = 30) and North Carolina (*PERCENT* = 60). North Carolina has twice the percentage of Nevada; therefore, we predict that North Carolina's average SAT will be about 50 points less than Nevada's. (In fact, it is 52 points less.)

Making The Excel Graph

In order to create the Excel plot of prototypical fitted values, you only need your fitted equation and min/max information about your predictor(s). <u>You do not need any raw data.</u>



Step1: Create a column for your untransformed predictor. Fill it with values beginning with the minimum and ending with the maximum. You can choose any intervals to get you from the min to the max. Fill out the column by selecting the two cells that define the interval, then click and drag down the lower right hand corner.

Step 2: Create a column for your predicted outcome. Make the first entry in your column equal to the right side of your fitted equation. Instead of writing out your predictor variable name, however, use the predictor variable value to the left. If you transformed the predictor, work that into your equation in the function bar. E.g., LOG (A2, 2) says take the log of cell A2 to the base 2.

Step 3: Create a scatterplot of the two columns and format to your heart's content. Start by going to Insert > Chart > XY (Scatter)

Using Your Fitted Regression Equations In The Spreadsheet

Note that I use "A2" to represent the appropriate cell value from the raw predictor column.

Detransforming X:

- When we go down in X, e.g., $LOG_2(X)$, we work our transformation into our fitted equation:

  = yintercept+slope*LOG(A2,2))

- When we go up in X, e.g., X^2, we work our transformation into our fitted equation:

  = yintercept+slope*(A2)^2

On the next slide I will discuss LN, natural logs, or logs to the base *e*.

Detransforming Y:

- When we go down in Y, e.g, LN(Y), we can use our good old linear equation:

  = yintercept+slope*(A2)

but that just gives us predicted LN(Y), so we then must take those numbers and antilog them with EXP( ), or we can antilog all in one step:

  = EXP(yintercept+slope*(A2))

- When we go up in Y, e.g, Y2, we can use our good old linear equation:

  = y-intercept+slope*(A2)

but that just gives us predicted Y2, so we then must take those numbers and "antisquare" them (i.e., take their square root) with SQRT( ), or we can square root all in one step:

  = SQRT(yintercept+slope*(A2))

Detransforming X and Y:

- When we went down in X, e.g., LN(X), and down in Y, e.g., LN(Y), we combine strategies above:

  = EXP(yintercept+slope*LN(A2))

The Natural Logarithm, LN, Log_e, or Log Base e

Log_2 is great for pedagogical purposes because we have an intuitive understanding of the concept of doubling. For the same reason, it is great for data analytic reporting (e.g., school board memos) when you must transform your predictor. However, base 2 is not the most common log base used in research. The most common log base is e, Euler's number, approximately 2.71828. There is no great reason why researchers use e instead of 2 (or any other base). Granted, e is perhaps the coolest number. It is as at least as helpful in calculus as π is in geometry. But, the (small) payoff in data analysis is this:

For : $\ln(OUTCOME) = \beta_0 + \beta_1 PREDICTOR + \varepsilon$

If : $\beta_1 < 0.25,$

Then : $100 * \beta_1 \approx$ The Percentage Difference in the Outcome (Y) Associated with a 1 Unit Difference in the Predictor (X)

Note that, whatever base we use, and whatever our β_1, when we log our outcome, we can interpret the model in terms of percentage differences in Y. We simply need to run our β_1 through the following formula:

$100 * a^{\beta_1} - 100 =$ The Percentage Difference in the Outcome (Y) Associated with a 1 Unit Difference in the Predictor (X)

Where $a =$ the log base (e.g., e or 2 or 10)

See the Math Appendix for an extended discussion.

Thank you, Judy Singer!

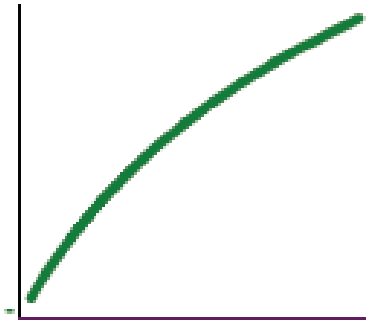
In every unit of these courses, I am greatly indebted to Judy Singer as well my other mentors at the Harvard Graduate School of Education including John Willett, Kate Elgin, Terry Tivnan, Dick Murnane and Dan Koretz. This unit in particular, however, I find myself using Professor Singer's data sets and insights more than in most units. In fact, the next slide is a cut-and-paste from her slides, because I can't get a handle on how to do it any other way…

As the next slide demonstrates, logarithmic transformations admit of very useful interpretations. Be careful, however, because log transformations cannot handle zeroes or negatives. There is no power to which you can raise a number to get a zero or negative number. For distributions with zeroes or negatives, first linearily transform to make all values positive, then log transform.

http://isites.harvard.edu/icb/icb.do?keyword=k18618&

Review: How to fit and interpret models using log-transformed variables

Learning Curve

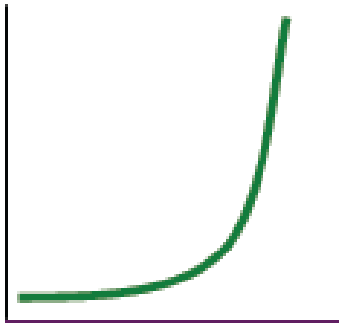


Regress Y on $\log_2(X)$

$$Y = \hat{\beta}_0 + \hat{\beta}_1 \log_2(X)$$

Every doubling of X (100% difference) is associated with a $\hat{\beta}_1$ difference in Y

Exponential Growth Model

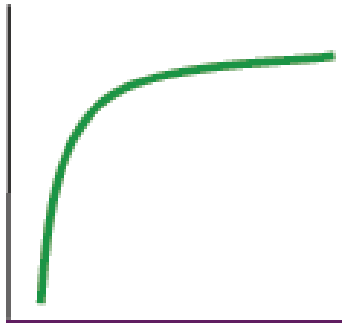


Regress $\log_e(Y)$ on X

$$\log_e(Y) = \hat{\beta}_0 + \hat{\beta}_1 X$$

Every 1 unit difference in X is associated with a $100(e^{\hat{\beta}_1} - 1)$% difference in Y (often interpreted as a %age growth rate)

Proportional Growth Model



Regress $\log_e(Y)$ on $\log_e(X)$

$$\log_e(Y) = \hat{\beta}_0 + \hat{\beta}_1 \log_e(X)$$

Every 1% difference in X is associated with a $\hat{\beta}_1$% difference in Y

Helpful mnemonic device: If you've logarithmically transformed a variable, you'll be modifying the interpretation of an effect by expressing differences for that variable in percentage, not unit, terms

Tukey's Ladder and Tukey's Rule of the Bulge

Tukey's Ladder

▲ MANY MORE ▲

$\text{antilog}(VAR)$

VAR^3

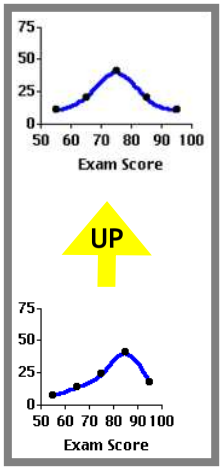
VAR^2

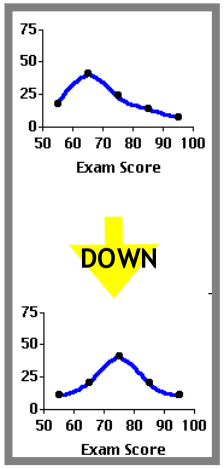
$\sqrt{VAR}$

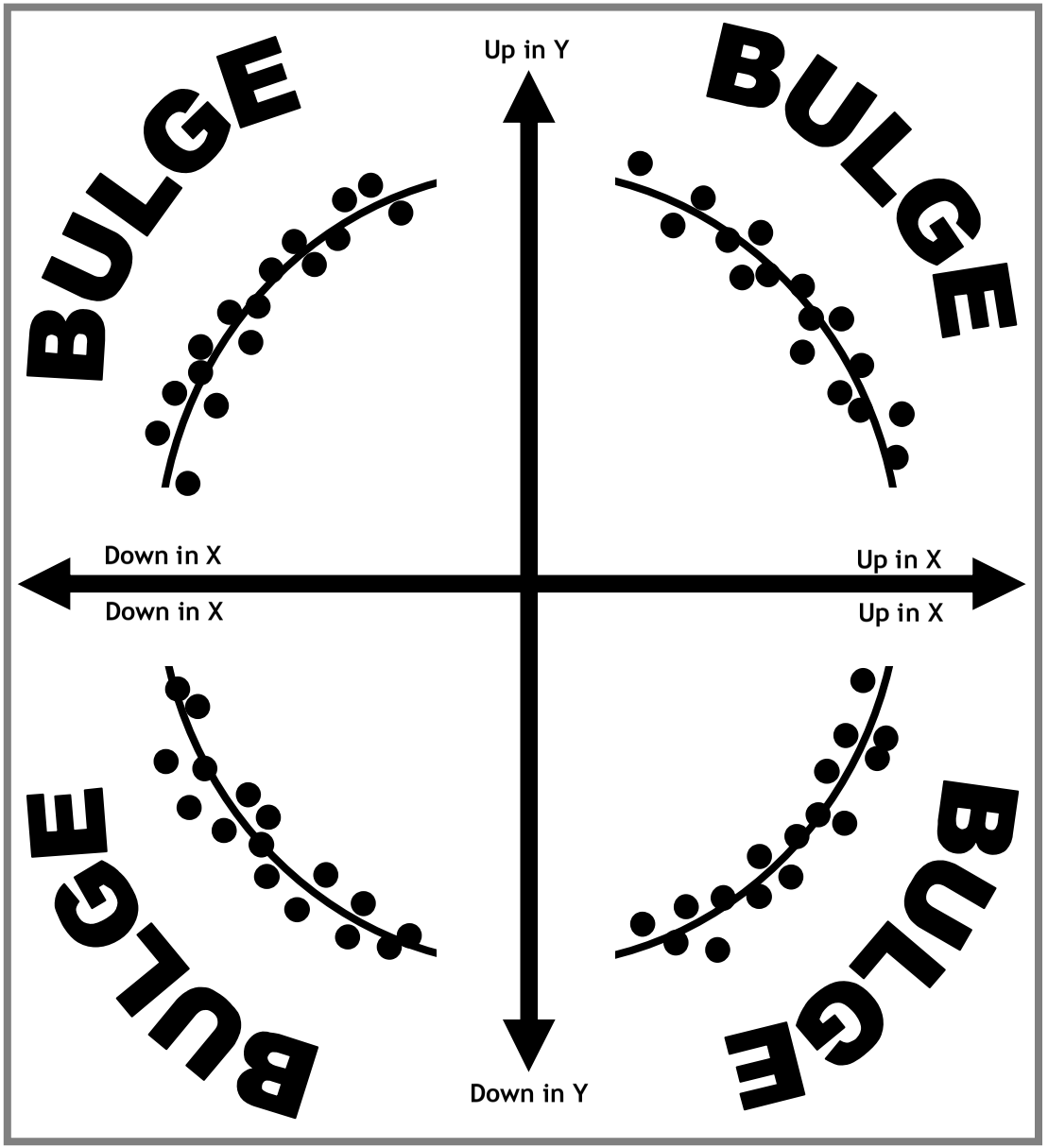
$\ln(VAR)$

$-\frac{1}{VAR}$

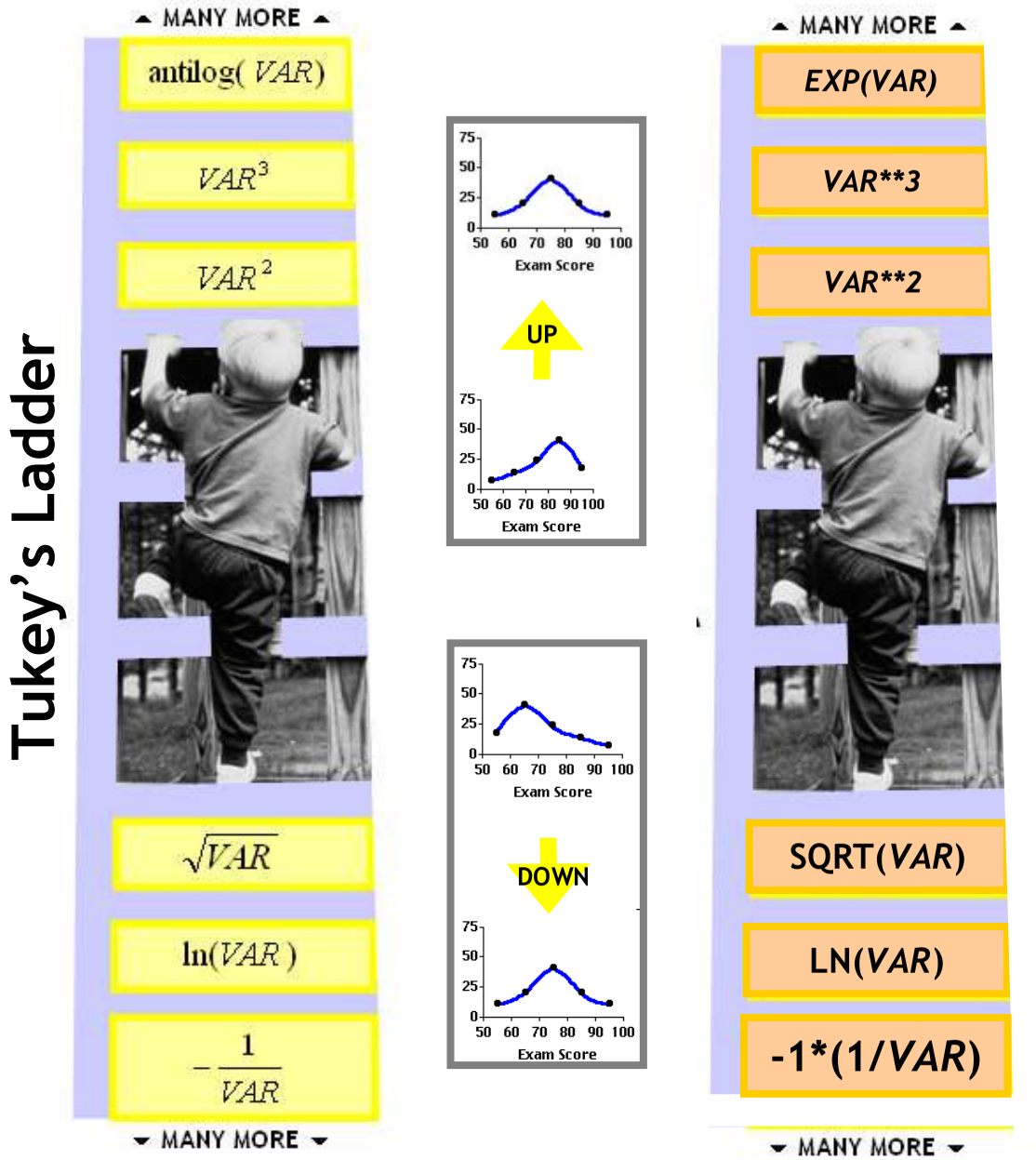
▼ MANY MORE ▼







Tukey's Ladder in SPSS




```
COMPUTE ANTILOGVAR = EXP(VAR).

COMPUTE VARCUBED = VAR**3.

COMPUTE VARSQ = VAR**2.

COMPUTE VARROOT = VAR**(1/2).
COMPUTE VARROOT = SQRT(VAR).
COMPUTE VARROOTP7 = SQRT(VAR+7).

COMPUTE L2VAR = LG10(VAR)/LG10(2).
COMPUTE LNVAR = LN(VAR).
COMPUTE LNVARP16 = LN(VAR+0.16).

COMPUTE INVPWRVAR = -1*(1/VAR).

EXECUTE.
```

Transformations Cheat Sheet

Linear Transformation

- add/subtract and/or multiply/divide by a constant
- does not change the shape of the distribution
- Common examples include converting to z-scores (i.e., standardizing) and converting to percentages.

Non-Linear Transformation

- changes the shape of the distribution
- "going up": squaring, cubing, etc. contracts left tails and expands right tails
- "going down": logs, roots, inverse powers, etc. expands left tails and contracts right tails

Think of non-linearly transforming as not "making the data fit your model," but rather as making your model fit the data. You legitimize this perspective when you do the hard work of reporting your results through plots of prototypical fitted values where the data are de-transformed.

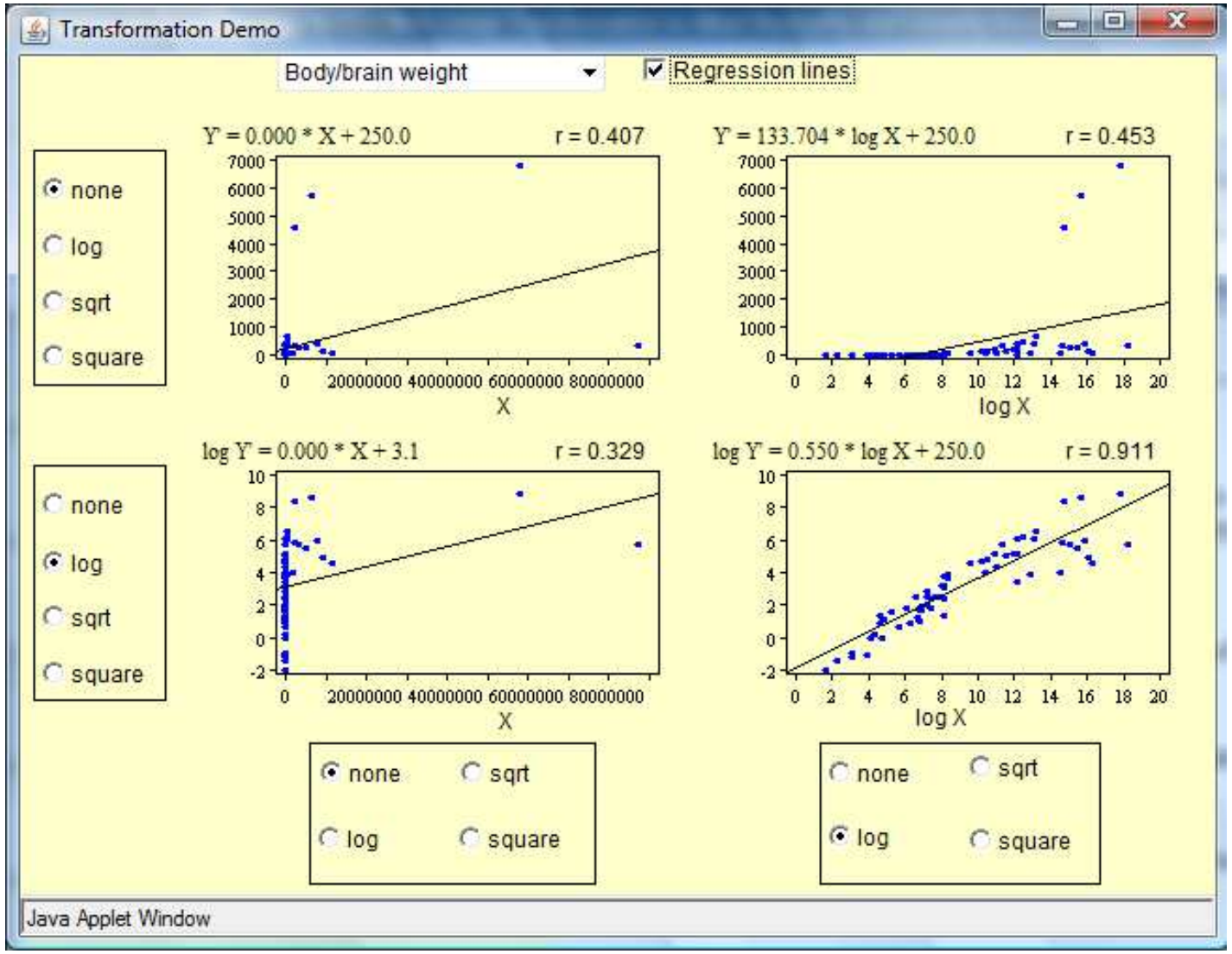
Everything in this unit is still linear regression. We know that we've non-linearily transformed our variables, but SPSS does not know. The burden is on us to interpret.

If you log transform Y, you will want to anti-log your predictions for graphing purposes. In such a case, remember what a log transformation is, and undo it. If you log transformed Y, then you used the base *e* (or 2.718281…), so you want to make each predicted value the power of 2.718281, and Excel is set up to handle it with the expression EXP(*VAR*).

A Recipe

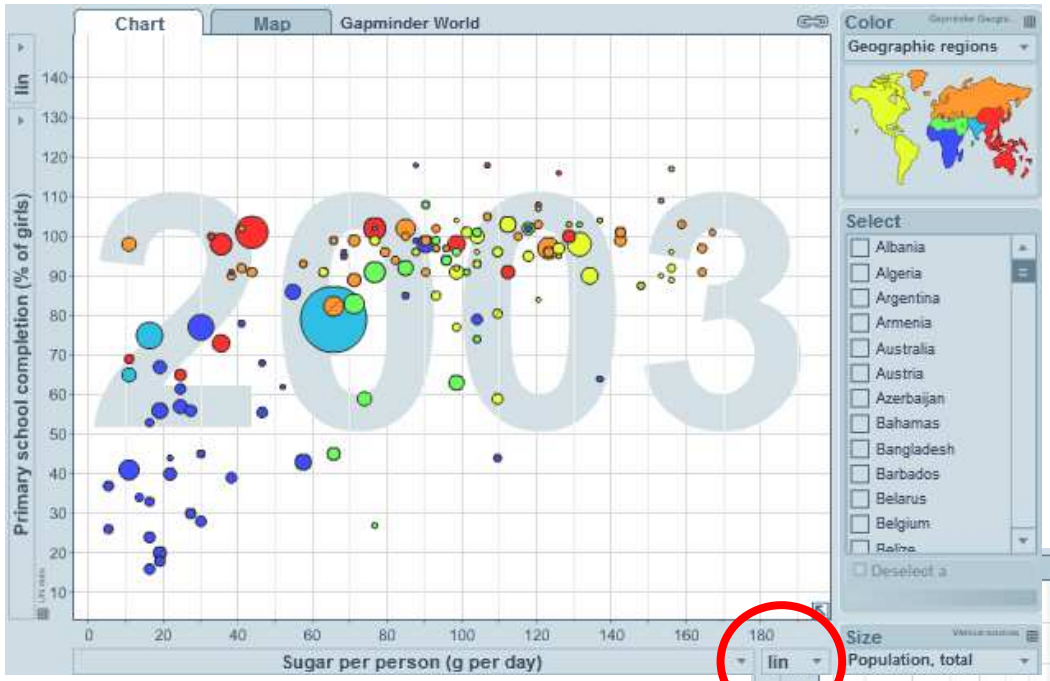
- Spot non-linearity in your exploratory data analysis.
- Use Tukey's Ladder and Tukey's Rule of the Bulge to choose a reasonable starting point for transformation.
- You may need several steps of trial and error, experimenting with different non-linear transformations and linear-then-non-linear transformations.
- When you have found a suitable transformation, use spreadsheet software to create a plot of prototypical fitted values.
- Write a good caption for your plot, and interpret it for your audience. When possible, use logarithmic transformations because of their nifty interpretations.

Playing Around With Transformations



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

GapMinder.Org



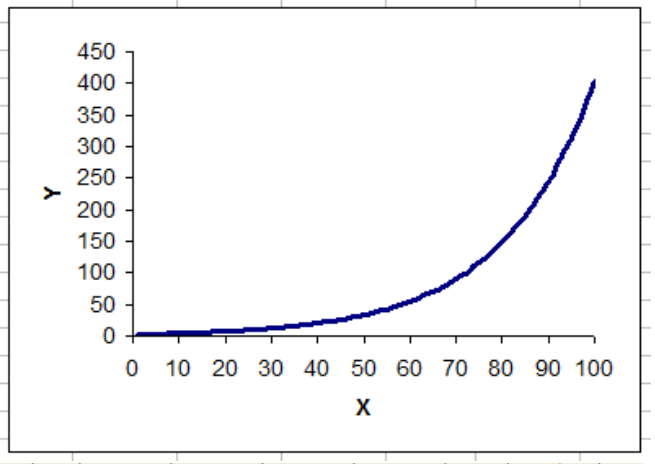
[GapMinder.Org]

Manipulating the Parameter Estimates in Excel

$$\log_e Y = \beta_0 + \beta_1 X + \varepsilon$$

β0	β1	ε
1	0.05	0

$\log_e Y$	β_0	β_1	X	ε	Y	$100(e^{\beta_1}-1)$ = percentage growth rate
1.05	1	0.05	1.00000	0	2.86	5.13
1.10	1	0.05	2.00000	0	3	5.13
1.15	1	0.05	3.00000	0	3.16	5.13
1.20	1	0.05	4.00000	0	3.32	5.13
1.25	1	0.05	5.00000	0	3.49	5.13
1.30	1	0.05	6.00000	0	3.67	5.13
1.35	1	0.05	7.00000	0	3.86	5.13
1.40	1	0.05	8.00000	0	4.06	5.13
1.45	1	0.05	9.00000	0	4.26	5.13
1.50	1	0.05	10.00000	0	4.48	5.13
1.55	1	0.05	11.00000	0	4.71	5.13
1.60	1	0.05	12.00000	0	4.95	5.13
1.65	1	0.05	13.00000	0	5.21	5.13
1.70	1	0.05	14.00000	0	5.47	5.13
1.75	1	0.05	15.00000	0	5.75	5.13
1.80	1	0.05	16.00000	0	6.05	5.13
1.85	1	0.05	17.00000	0	6.36	5.13
1.90	1	0.05	18.00000	0	6.69	5.13
1.95	1	0.05	19.00000	0	7.03	5.13

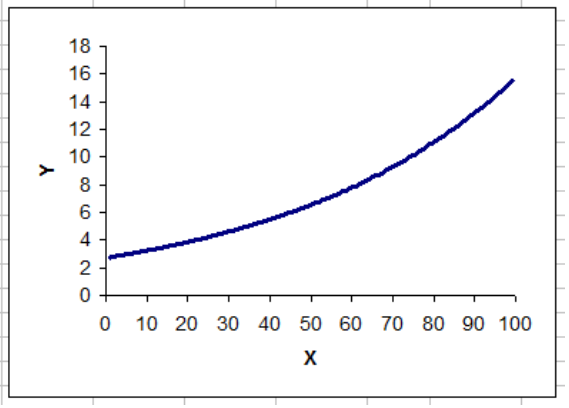


$$\log_e Y = \beta_0 + \beta_1 \log_e X + \varepsilon$$

HIDE

β0	β1	ε
1	0.05	0

$\log_e Y$	β_0	β_1	$\log_e X$	ε	X	Y	Percentage Difference In X		Percentage Difference in Y	
1.00	1	0.05	0.00000	0	1	2.718282	X			
1.00	1	0.05	0.00995	0	1.01	2.719635	1.00		0.05	####
1.00	1	0.05	0.01990	0	1.0201	2.720988	1.00		0.05	####
1.00	1	0.05	0.02985	0	1.030301	2.722342	1.00		0.05	####
1.00	1	0.05	0.03980	0	1.040604	2.723697	1.00		0.05	####
1.00	1	0.05	0.04975	0	1.05101	2.725052	1.00		0.05	####
1.00	1	0.05	0.05970	0	1.06152	2.726408	1.00		0.05	####
1.00	1	0.05	0.06965	0	1.072135	2.727765	1.00		0.05	####
1.00	1	0.05	0.07960	0	1.082857	2.729123	1.00		0.05	####
1.00	1	0.05	0.08955	0	1.093685	2.730481	1.00		0.05	####
1.00	1	0.05	0.09950	0	1.104622	2.731839	1.00		0.05	####
1.01	1	0.05	0.10945	0	1.115668	2.733199	1.00		0.05	####
1.01	1	0.05	0.11940	0	1.126825	2.734559	1.00		0.05	####
1.01	1	0.05	0.12935	0	1.138093	2.73592	1.00		0.05	####
1.01	1	0.05	0.13930	0	1.149474	2.737281	1.00		0.05	####
1.01	1	0.05	0.14925	0	1.160969	2.738644	1.00		0.05	####
1.01	1	0.05	0.15921	0	1.172579	2.740006	1.00		0.05	####
1.01	1	0.05	0.16916	0	1.184304	2.74137	1.00		0.05	####



Go To.

Answering our Roadmap Question

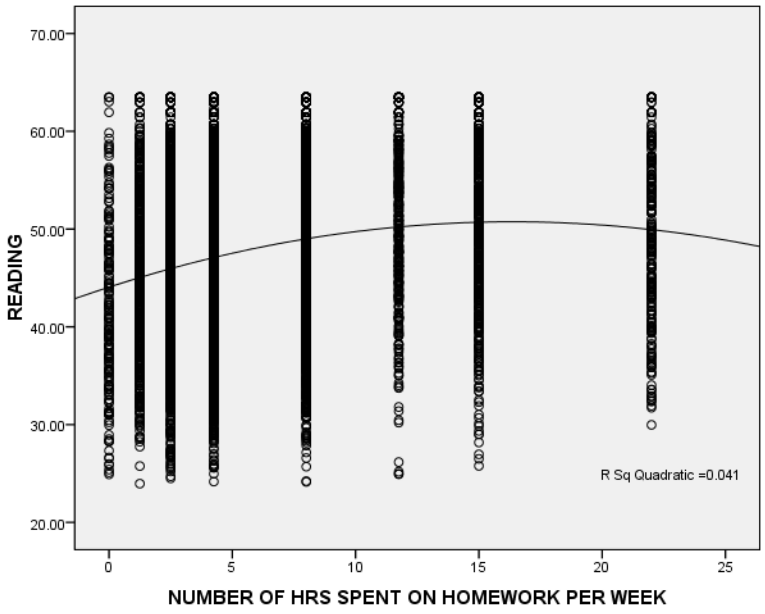
Unit 13: How do we deal with violations of the ***linearity*** and ***normality*** assumptions?

$$READING = \beta_0 + \beta_1 HOMEWORK + \varepsilon$$

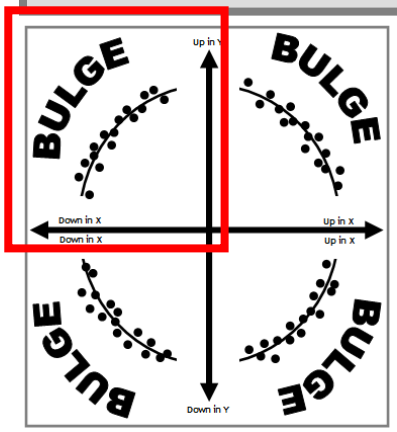
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	45.514	.154		296.359	.000	45.213	45.815
	NUMBER OF HRS SPENT ON HOMEWORK PER WEEK	.332	.020	.183	16.451	.000	.292	.371

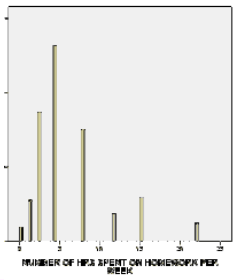
a. Dependent Variable: READING



Intuitively, it makes sense that there is a law of diminishing returns for time on homework.



We can try going up in Y, or going down in X.



Let's go down in X and pull in that long upper tail.

As a general rule, when we have choice of transforming X or Y, we should lean toward X, because if we transform Y, the transformation effects not only X but all the other predictors in our model. It is "inexpensive" to transform X.

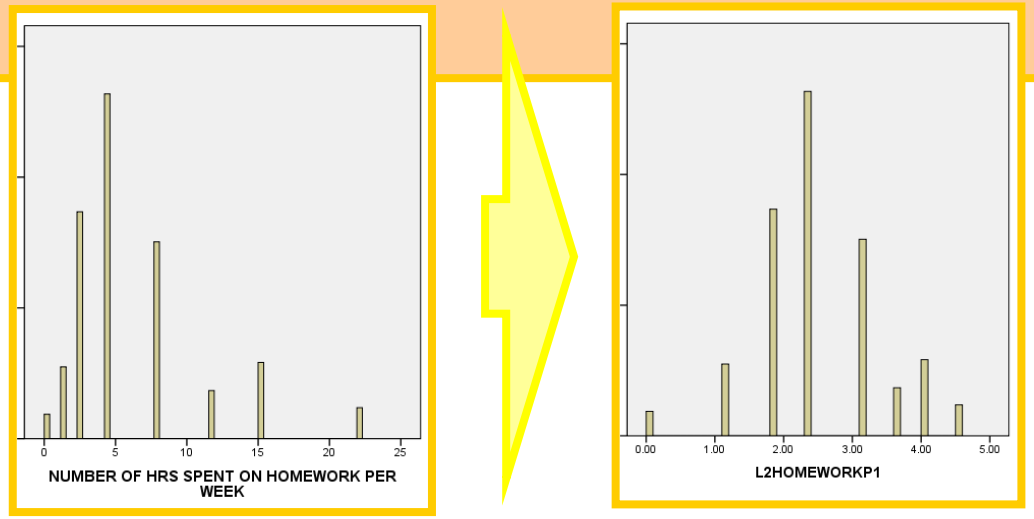
Answering our Roadmap Question

Unit 13: How do we deal with violations of the ***linearity*** and ***normality*** assumptions?

Since *HOMEWORK* takes on values of zero (i.e., some students fess up to spending no time on HW), we cannot log transform unless we first add a smidge. I'll add 1 to each value. I will take the log base-2 after adding 1 to each value of *HOMEWORK*, and I will call the new variable *L2HOMEWORKP1*, where the "L2" reminds me that I took the log base-2, and the "P1" reminds me that I added one ("P" for plus).

In order to use Professor Singer's nifty interpretations. When you log only the predictor, use base-2. When you log the outcome (and perhaps also the predictor), use base-e, i.e., the natural log.

```
COMPUTE L2HOMEWORKP1=(LG10(HOMEWORK+1))/(LG10(2)).
EXECUTE.
COMPUTE LNHOMEWORKP1=LN(HOMEWORK+1).
EXECUTE.
```



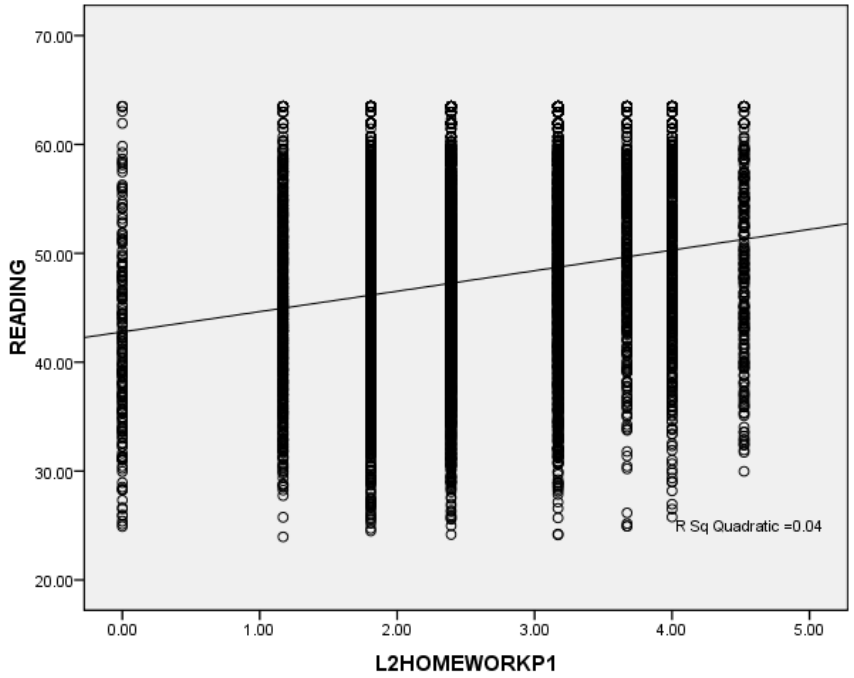
Answering our Roadmap Question

Unit 13: How do we deal with violations of the ***linearity*** and ***normality*** assumptions?

$$READING = \beta_0 + \beta_1 L2HOMEWORKP1 + \varepsilon$$

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	42.760	.279		153.130	.000	42.213	43.307
	L2HOMEWORKP1	1.883	.104	.200	18.030	.000	1.678	2.088

a. Dependent Variable: READING

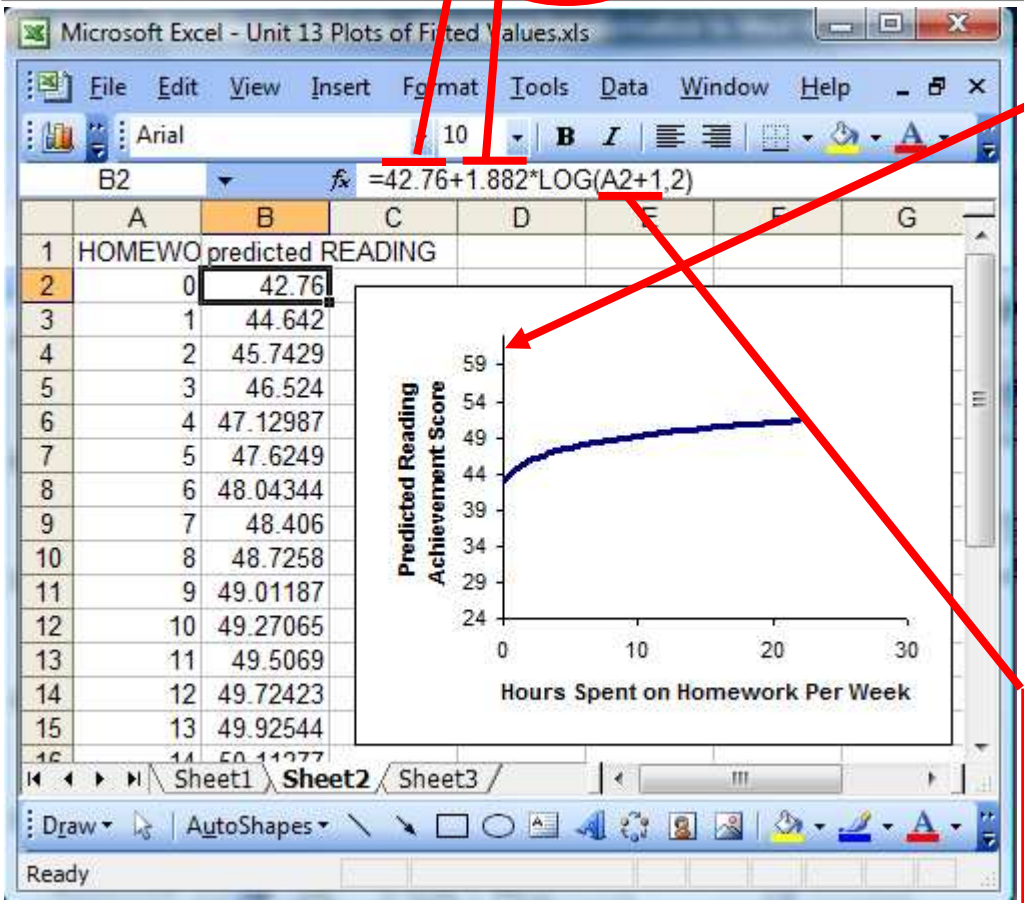


With so much data, it's hard to see what's going on, but from Unit 11, I knew something is going on. Just like two slides ago, I fit a quadratic curve to pick up non-linearity, but unlike two slides ago, I get a straight line—success!

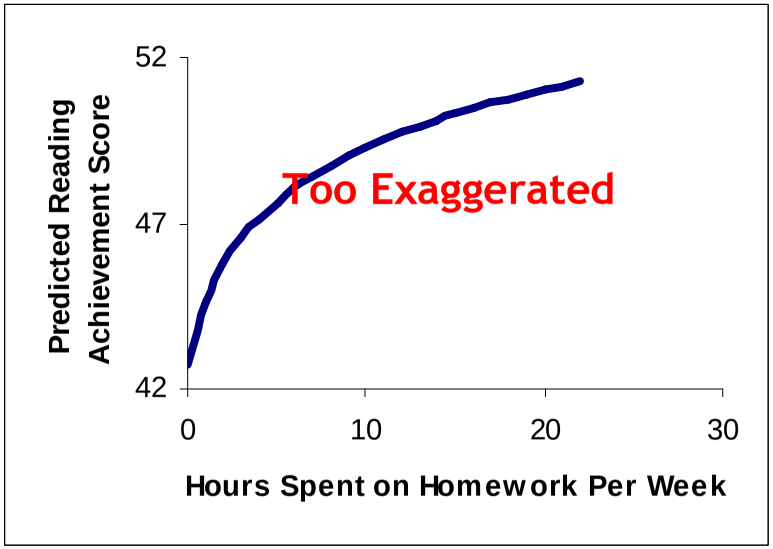
Answering our Roadmap Question

Unit 13: How do we deal with violations of the ***linearity*** and ***normality*** assumptions?

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	42.760	.279		153.130	.000	42.213	43.307
	L2HOMEWORKP1	1.883	.104	.200	18.030	.000	1.678	2.088



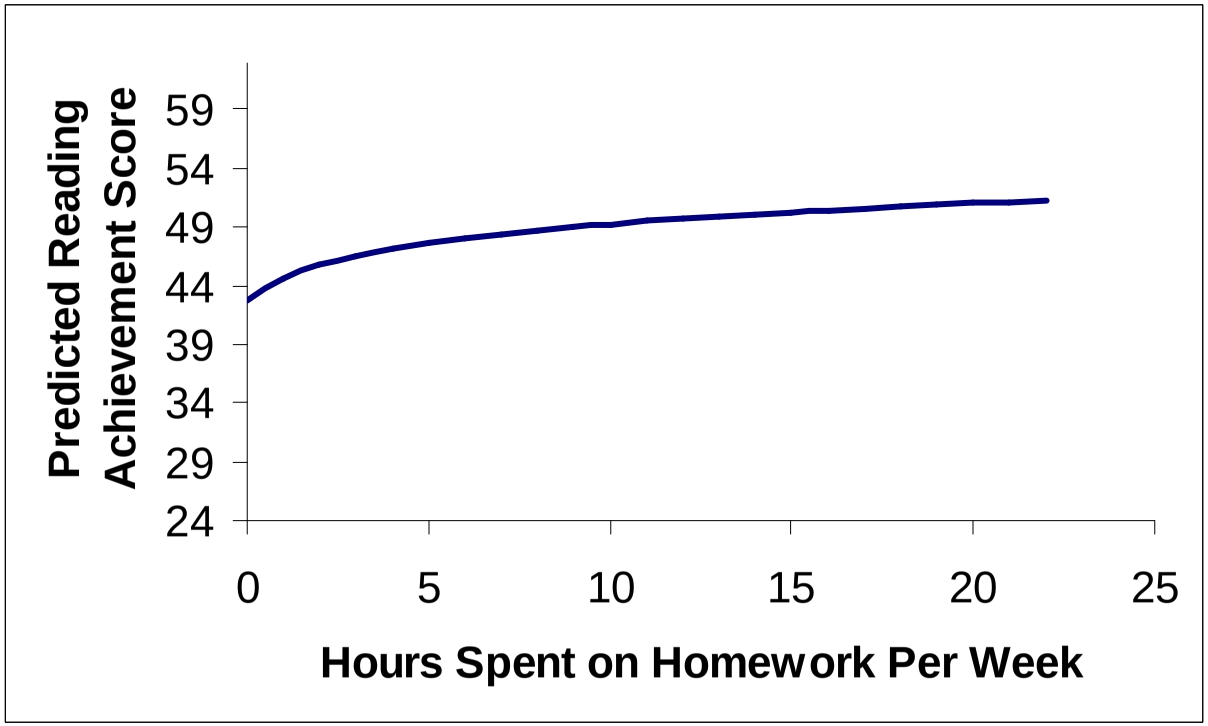
I made some judgment calls regarding the scales graph. They are little decisions with big consequences.

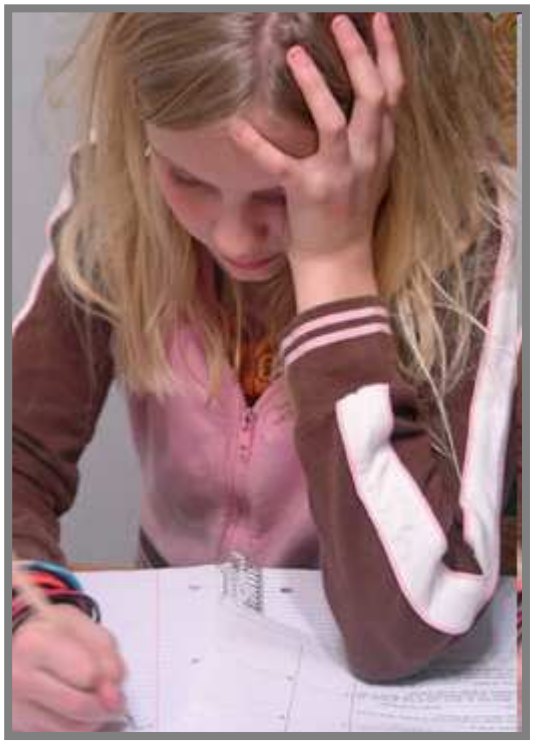


I use A2+1, because I added 1 to each value of *HOMEWORK* before base-2 logging it.

Answering our Roadmap Question

Figure 13.Y. A plot of fitted values depicting the nonlinear relationship between reading score and hours spent on homework per week for a nationally representative sample of 8th graders (n = 7,800).





In our representative sample of 7,800 8th graders, we found a statistically significant positive relationship between reading achievement score and hours spent on homework per week ($p < 0.001$). The relationship was non-linear such that the magnitude of the relationship was greatest from 0-2 hours and least from 20-22 hours. Twice as many homework hours per week is associated with a difference of about 2 points on the reading achievement test. Take for example one student who studies 1 hour per week and another student who studies twice as much, 2 hours per week, we expect the more diligent student to score about 2 points higher. We also expect this 2-point difference when we compare two students who study 10 hours per week and 20 hours per week.

Unit 13 Appendix: Key Concepts

A skewed distribution in the outcome and/or predictor sometimes (but not always) foreshadows a non-linear relationship.

Just because a relationship is statistically significant does not mean that you are modeling the right relationship.

Just because a correlation is high does not mean the correlation is right.

A horseshoe pattern in the residual versus fitted (RVF) plot indicates a violation of the GLM linearity assumption. Our goal is to have a patternless cloud, but a concave up horsehoe pattern indicates that for our low predictions we are underestimating, for our middling predictions we are overestimating, and for our high predictions we are underestimating.

Logarithmic transformations pull in long upper tails.

When we log transform our predictor using base 2, the slope coefficient is no longer the difference in our outcome associated with a one unit difference in our predictor, but rather the slope coefficient is now the difference in our outcome associated with a doubling of our predictor.

In order to create the Excel plot of prototypical fitted values, you only need your fitted equation and information about your predictor(s). You do not need any raw data.

Logarithmic transformations admit of very useful interpretations. Be careful, however, because log transformations cannot handle zeroes or negatives. There is no power to which you can raise a number to get a zero or negative number. For distributions with zeroes or negatives, first linearily transform to make all values positive, then log transform.

Think of non-linearly transforming as not "making the data fit your model," but rather as making your model fit the data. You legitimize this perspective when you do the hard work of reporting your results through plots of prototypical fitted values where the data are de-transformed.

Everything in this unit is still linear regression. We know that we've non-linearily transformed our variables, but SPSS does not know. The burden is on us to interpret.

If you log transform Y, you will want to anti-log your predictions for graphing purposes. In such a case, remember what a log transformation is, and undo it. If you log transformed Y, then you used the base *e* (or 2.718281…), so you want to make each predicted value the power of 2.718281, and Excel is set up to handle it with the expression EXP(*VAR*).

As a general rule, when we have choice of transforming X or Y, we should lean toward X, because if we transform Y, the transformation effects not only X but all the other predictors in our model. It is "inexpensive" to transform X.

In order to use Professor Singer's nifty interpretations. When you log only the predictor, use base-2. When you log the outcome (and perhaps also the predictor), use base-e, i.e., the natural log.

Unit 13 Appendix: Key Interpretations

Given two states where one state has twice the other's percentage of eligible test takers taking the SAT, we expect the state with twice the percentage of eligible test takers taking the SAT to have an average SAT score of about 50 points less. For example, take Mississippi (*PERCENT* = 4) and Alabama (*PERCENT* = 8). Alabama has twice the percentage of Mississippi; therefore, we predict that Alabama's average SAT will be about 50 points less than Mississippi's. (In fact, it is 47 points less.) For another example, take Nevada (*PERCENT* = 30) and North Carolina (*PERCENT* = 60). North Carolina has twice the percentage of Nevada therefore, we predict that North Carolina's average SAT will be about 50 points less than Nevada's. (In fact, it is 52 points less.)

In our representative sample of 7,800 8th graders, we found a statistically significant positive relationship between reading achievement score and hours spent on homework per week ($p < 0.001$). The relationship was non-linear such that the magnitude of the relationship was greatest from 0-2 hours and least from 20-22 hours. Twice as many homework hours per week is associated with a difference of about 2 points on the reading achievement test. Take for example one student who studies 1 hour per week and another student who studies twice as much, 2 hours per week, we expect the more diligent student to score about 2 points higher. We also expect this 2-point difference when we compare two students who study 10 hours per week and 20 hours per week.

Unit 13 Appendix: Key Terminology

Linear Transformation

- add/subtract and/or multiply/divide by a constant
- does not change the shape of the distribution
- Common examples include converting to z-scores (i.e., standardizing) and converting to percentages.

Non-Linear Transformation

- changes the shape of the distribution
- "going up": squaring, cubing, etc. contracts left tails and expands right tails
- "going down": logs, roots, inverse powers, etc. expands left tails and contracts right tails

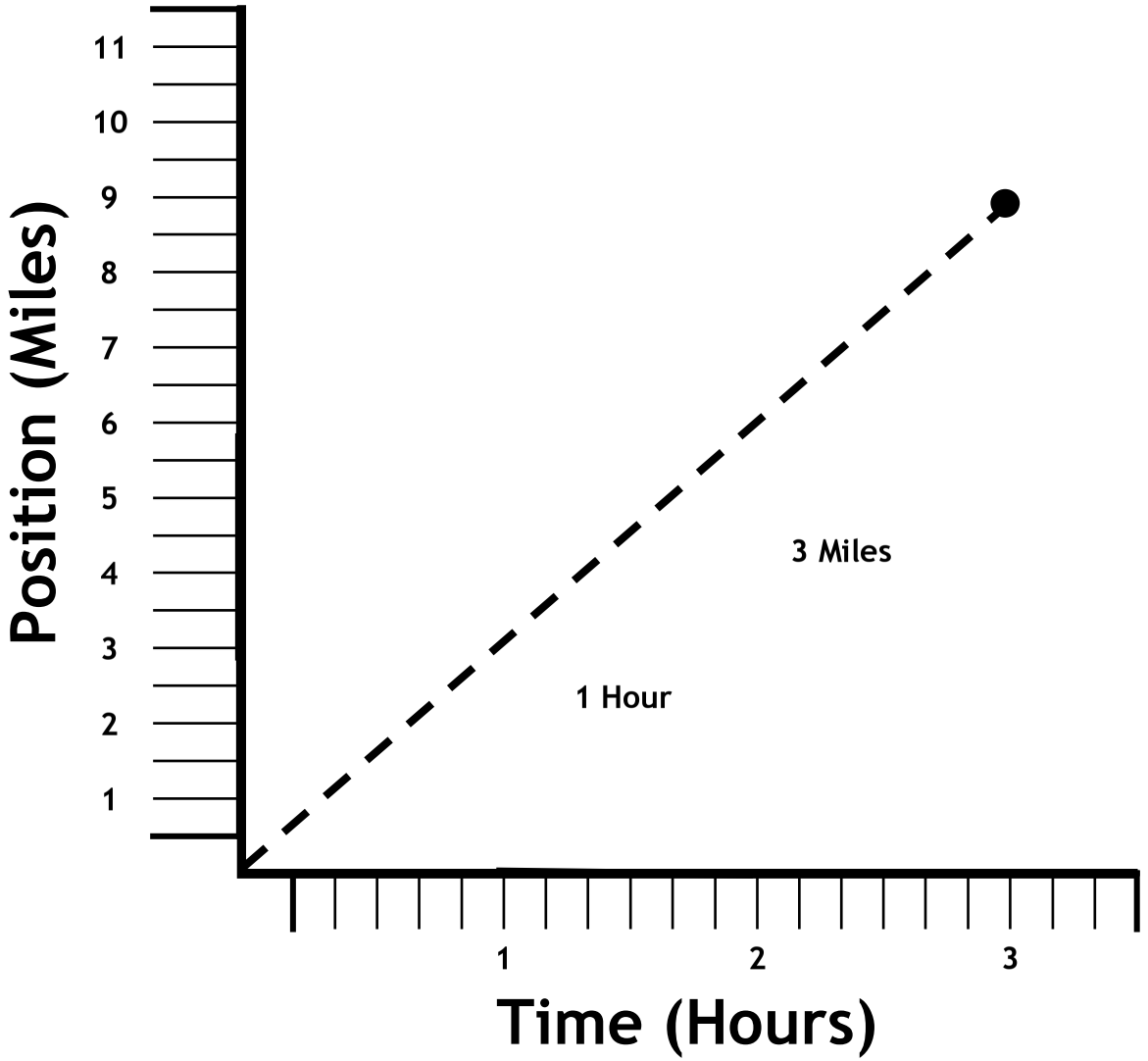
Unit 13 Appendix: Formulas

$$L2PERCENT = Log_2(PERCENT) = \frac{Log_{10}(PERCENT)}{Log_{10}(2)}$$

$$\text{If } x = b^y \text{ then, } y = \log_b(x)$$

When I logarithmically transform a variable to the base 2, I ask of each value of the variable, "By what power must I raise 2 in order to equal you?" The power by which I must raise 2 becomes the transformed value of the variable. Take Alabama, for example. Of the eligible students in Alabama, 8% take the SAT, so in the variable called *PERCENT*, Alabama has a value of 8. I ask, "What power do I need to raise 2 by in order to equal 8?" $2^1=2$. $2^2=4$. $2^3=8$. $2^4=16$. $2^5=32$. $2^6=64$. And so on… Notice that I must raise 2 to the power of 3 to get the 8 for which I was looking. Thus, 3 is the transformed value of 8. Three is the new eight.

Math Appendix: Calculus, Slopes, Non-Linearity, Change, e



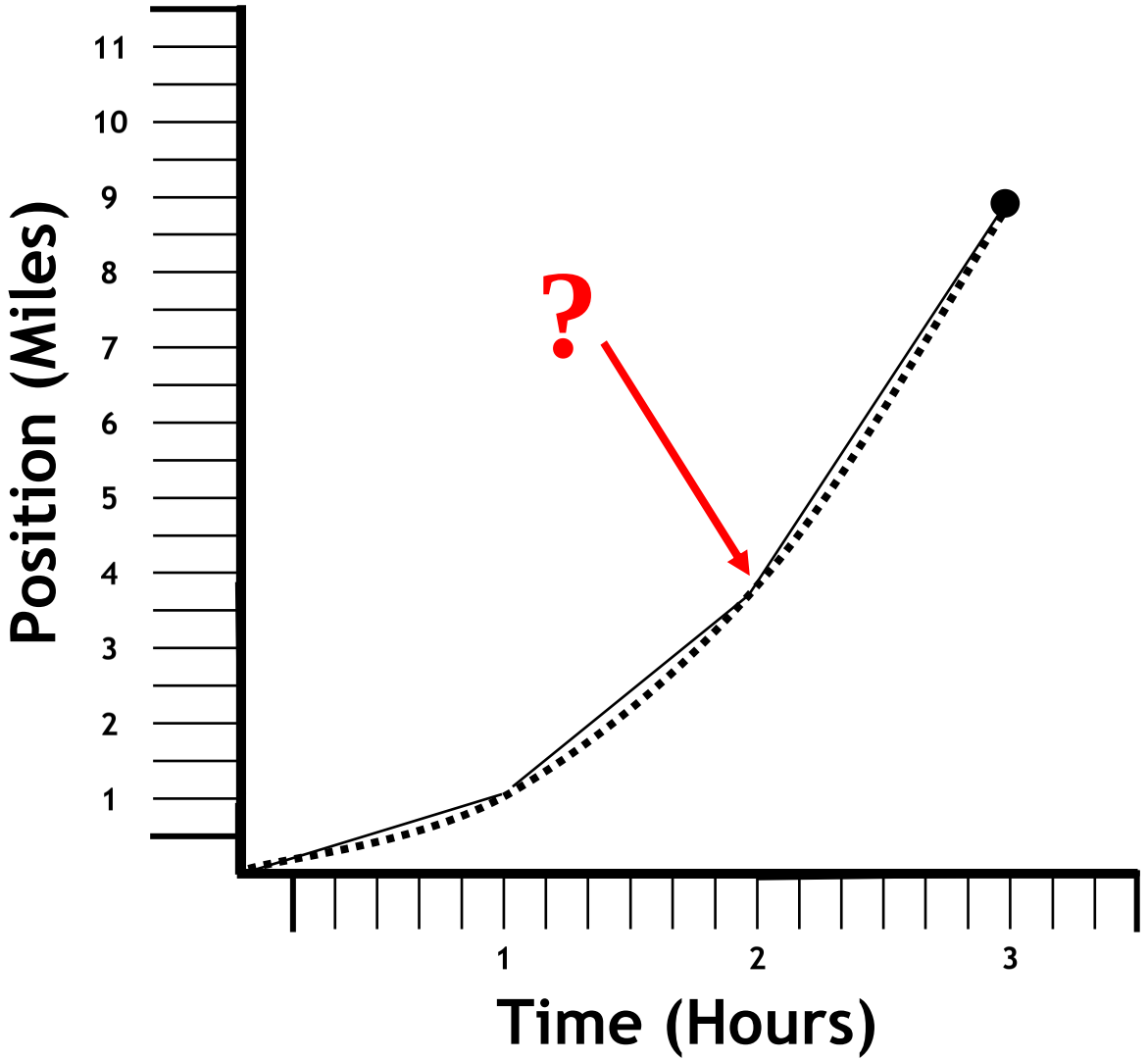
If, after 3 hours of hiking, you hiked 9 miles, how fast did you go? I.e., what was your speed? I.e., what was your rate of change?

3 MPH (Miles/Hour)

You've calculated a slope!

But… You assumed the rate of change was constant. (You assumed the relationship between position and time was linear.) Yes, you did calculate the average rate of change, but you did not necessarily calculate the instantaneous rate of change at all instants.

Math Appendix: Calculus, Slopes, Non-Linearity, Change, *e*



What was your ***average*** rate of change in the 1st hour?

1 MPH (Mile/Hour)

What was your ***average*** rate of change in the 2nd hour?

3 MPH (Miles/Hour)

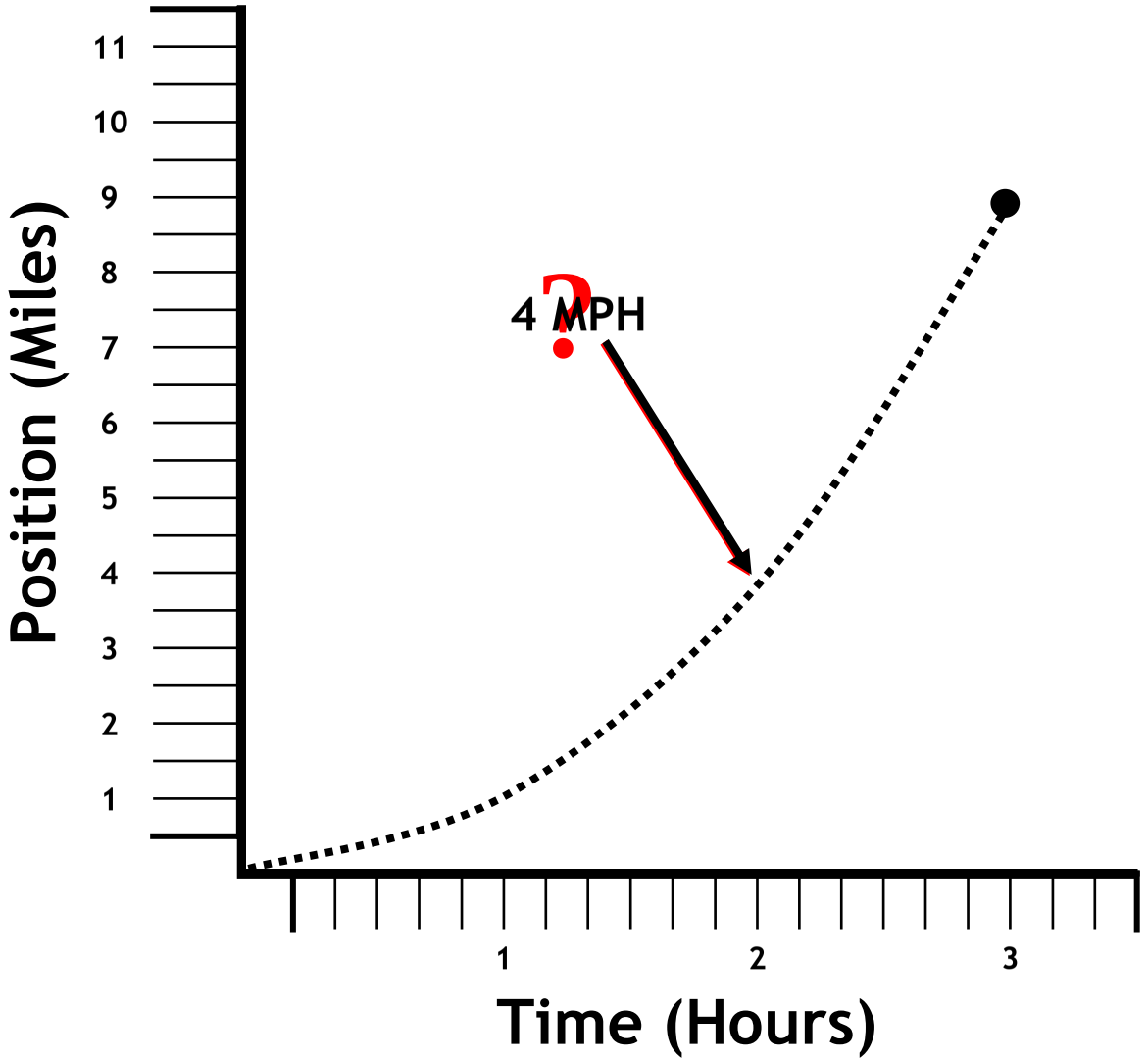
What was your ***average*** rate of change in the 3rd hour?

5 MPH (Miles/Hour)

What was the ***instantaneous*** rate of change at the 2-hour mark?

You can't calculate the slope, because you don't have a chunk of time (e.g., 1 hour or 1 minute) for the denominator/run.

Math Appendix: Calculus, Slopes, Non-Linearity, Change, *e*



You can use calculus to calculate the ***instantaneous*** rate of change if the relationship's functional form is well-behaved.

This functional form is well-behaved:

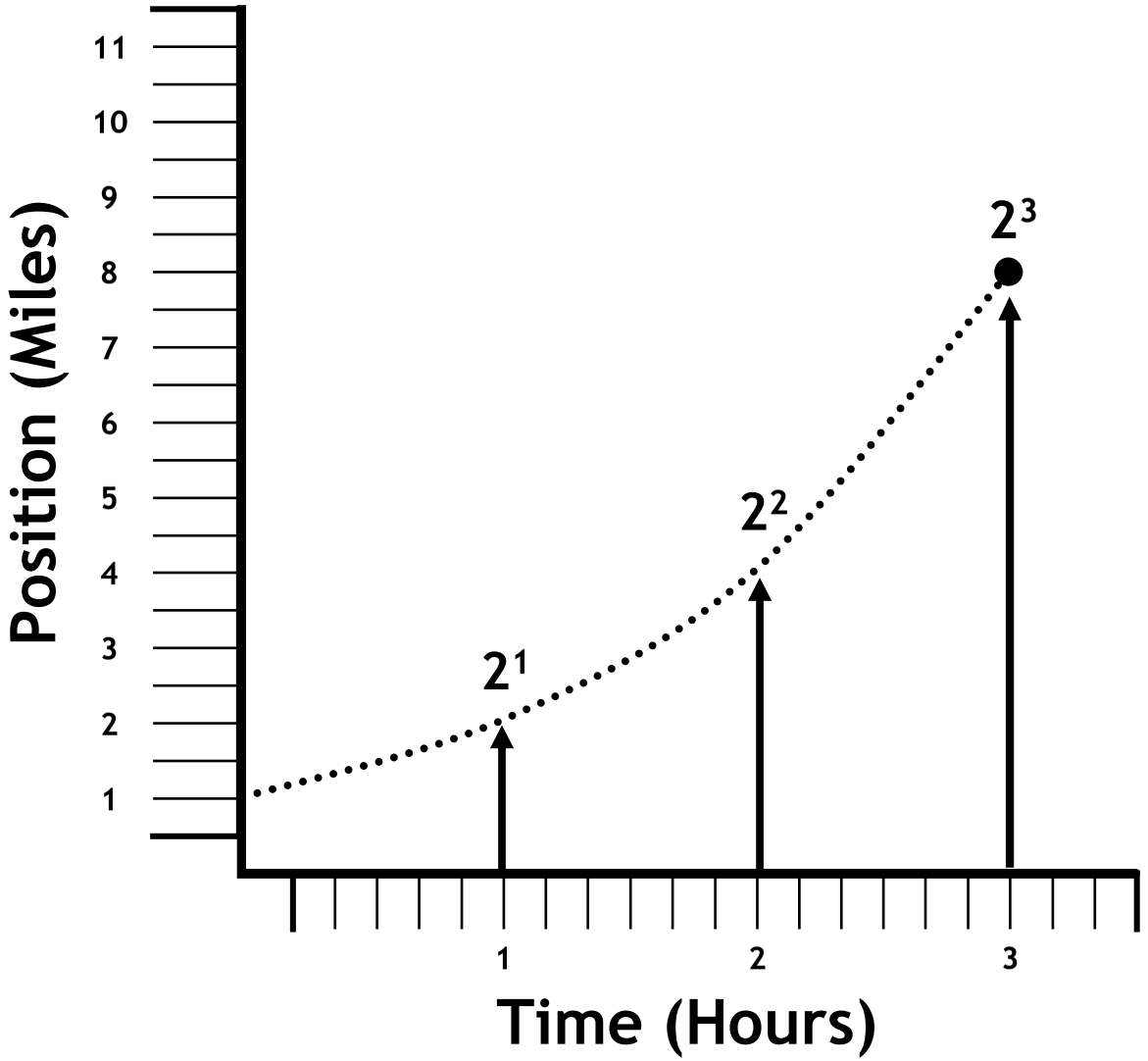
$$\text{Position} = \text{Time}^2$$

Therefore, using the power rule in calculus to calculate speed (i.e., the first derivative of position):

$$\text{Speed} = 2*\text{Time}$$

Thus, at the 2-hour mark, your instantaneous rate of change (i.e., speed) is 4 MPH.

Math Appendix: Calculus, Slopes, Non-Linearity, Change, e



Functions with constant exponents (e.g., squares and square roots) are generally well-behaved. Also, logarithmic functions and functions with variable exponents are generally well-behaved.

This exponential functional form is well- behaved:

$$\text{Position} = 2^{\text{Time}}$$

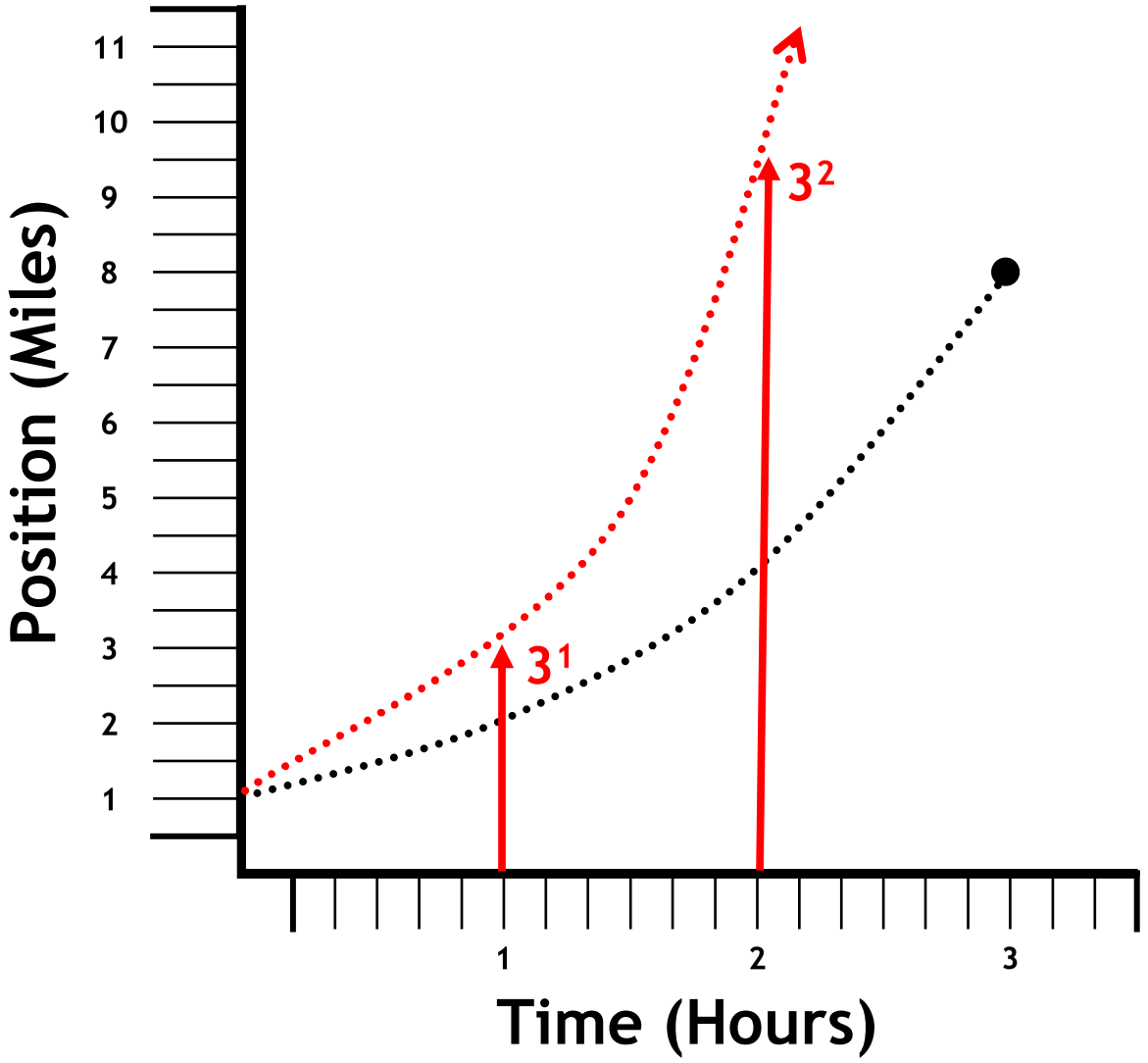
Where Position equals 2 to the power of Time.

From calculus, we know:

$$\text{Speed} = \text{Something} * 2^{\text{Time}}$$

And, that "Something" is about 0.6931. For each a^{Time}, a has an associated "Something," and when a = 2, then the "Something" is about 0.6931.

Something More About "Something"



Suppose:

$$\text{Position} = 3^{\text{Time}}$$

Where Position equals 3 to the power of Time.

Then:

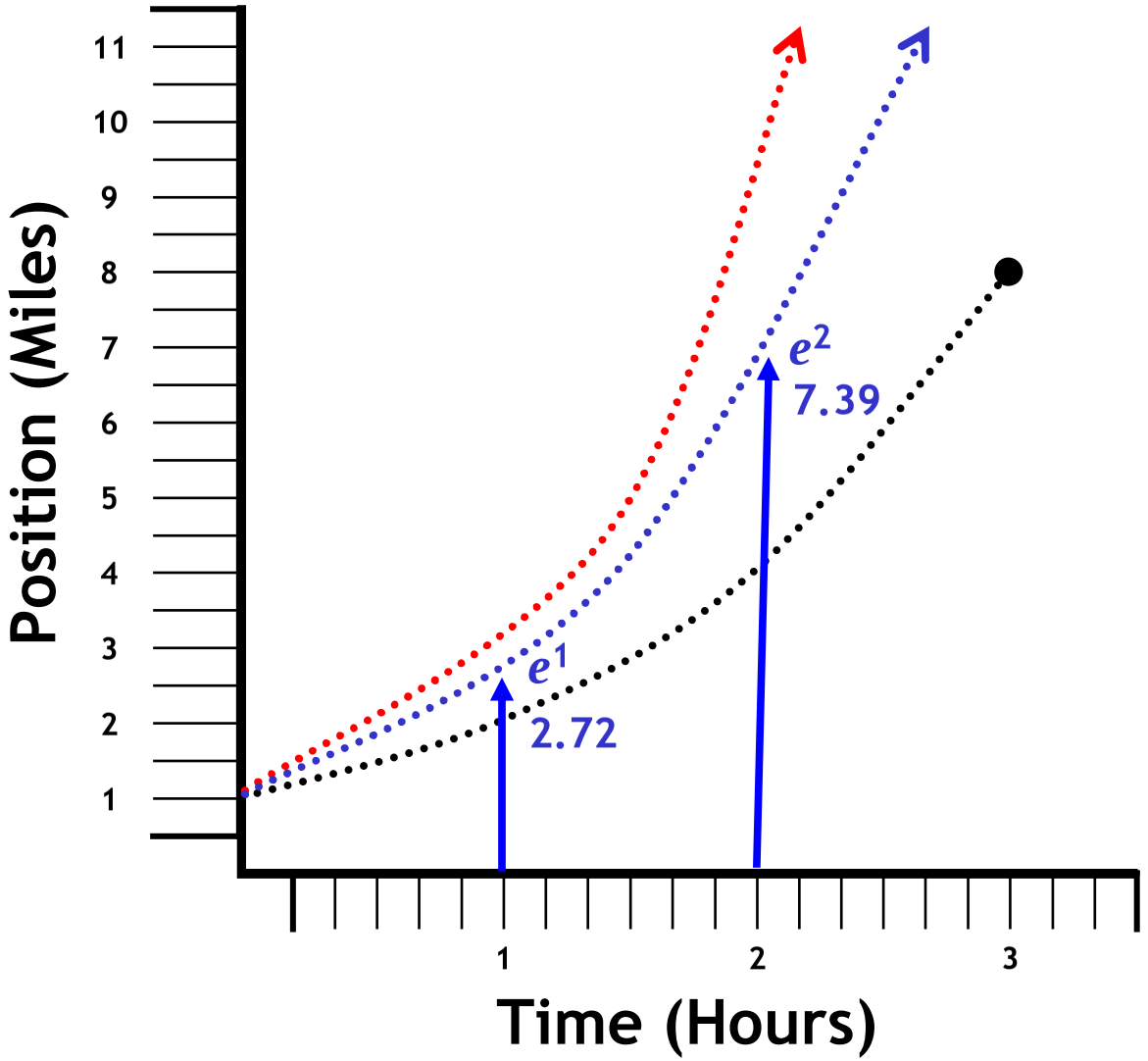
$$\text{Speed} = \text{Something} * 3^{\text{Time}}$$

And, for a=3, "Something" is about 1.0986.

a	"Something"
1	0.000000
2	0.693147
3	1.098612
4	1.386294
5	1.609438
6	1.791759

Notice that, for some value of a, between 2 and 3, "Something" equals 1.

Something About e



The value of a for which "Something" equals 1 is called "e"! And, it is roughly equal to 2.71828. "e" stands for "exponential" or "Euler's number" or "easy."

Suppose:

$$\text{Position} = e^{\text{Time}}$$

Where Position equals e to the power of Time.

Then:

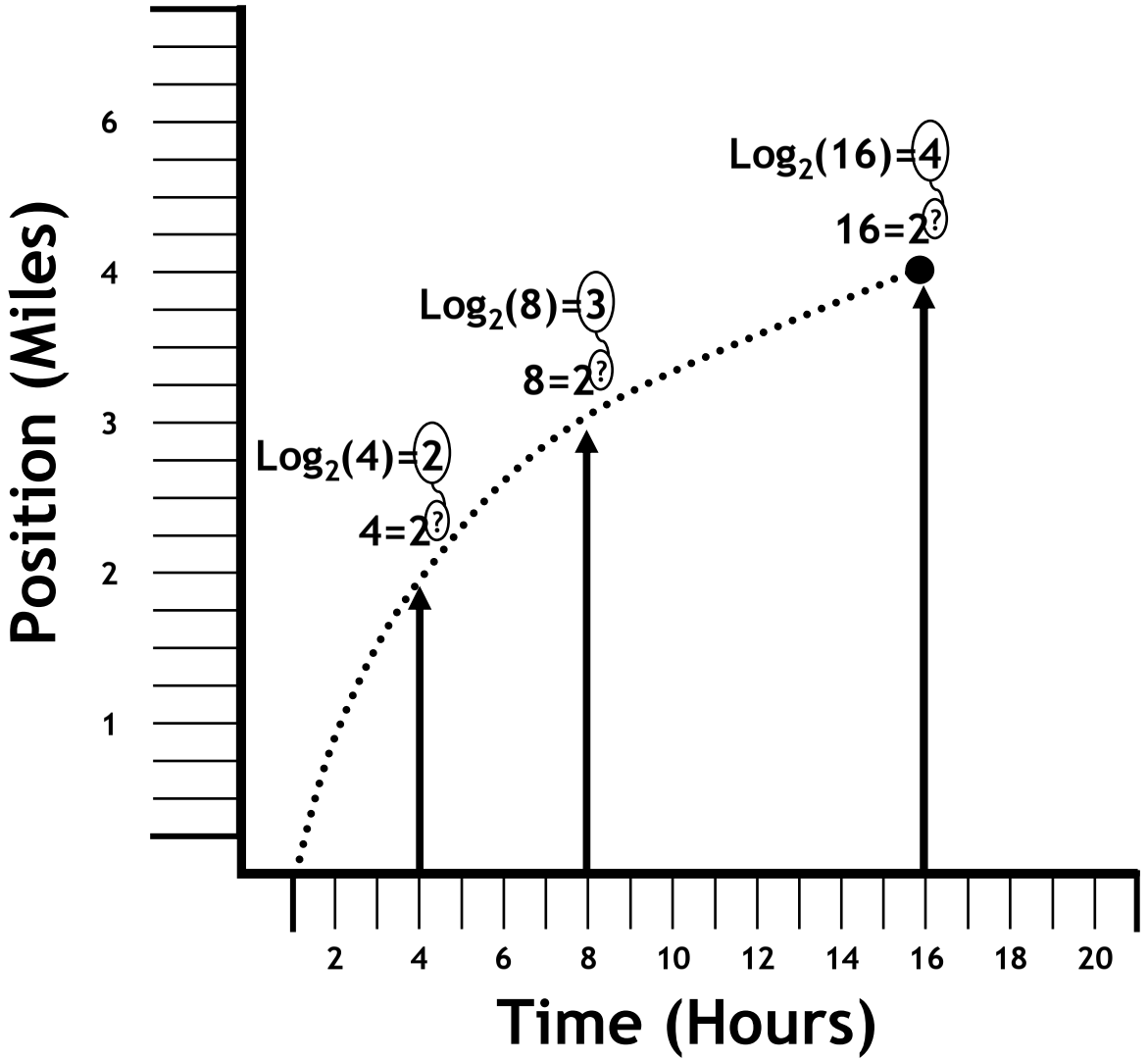
$$\text{Speed} = \text{Something} * e^{\text{Time}}$$

And, for $a=e$, "Something" is exactly 1. So,

$$\text{Speed} = e^{\text{Time}} = \text{Position}$$

When you've gone 1 mile, you are going 1 MPH, and when you've gone 2 miles, you are going 2 MPH and when you've gone 3 miles…

Log_2 or "Log to the Base 2"



This functional form is well-behaved:

Position = $\log_2$(Time)

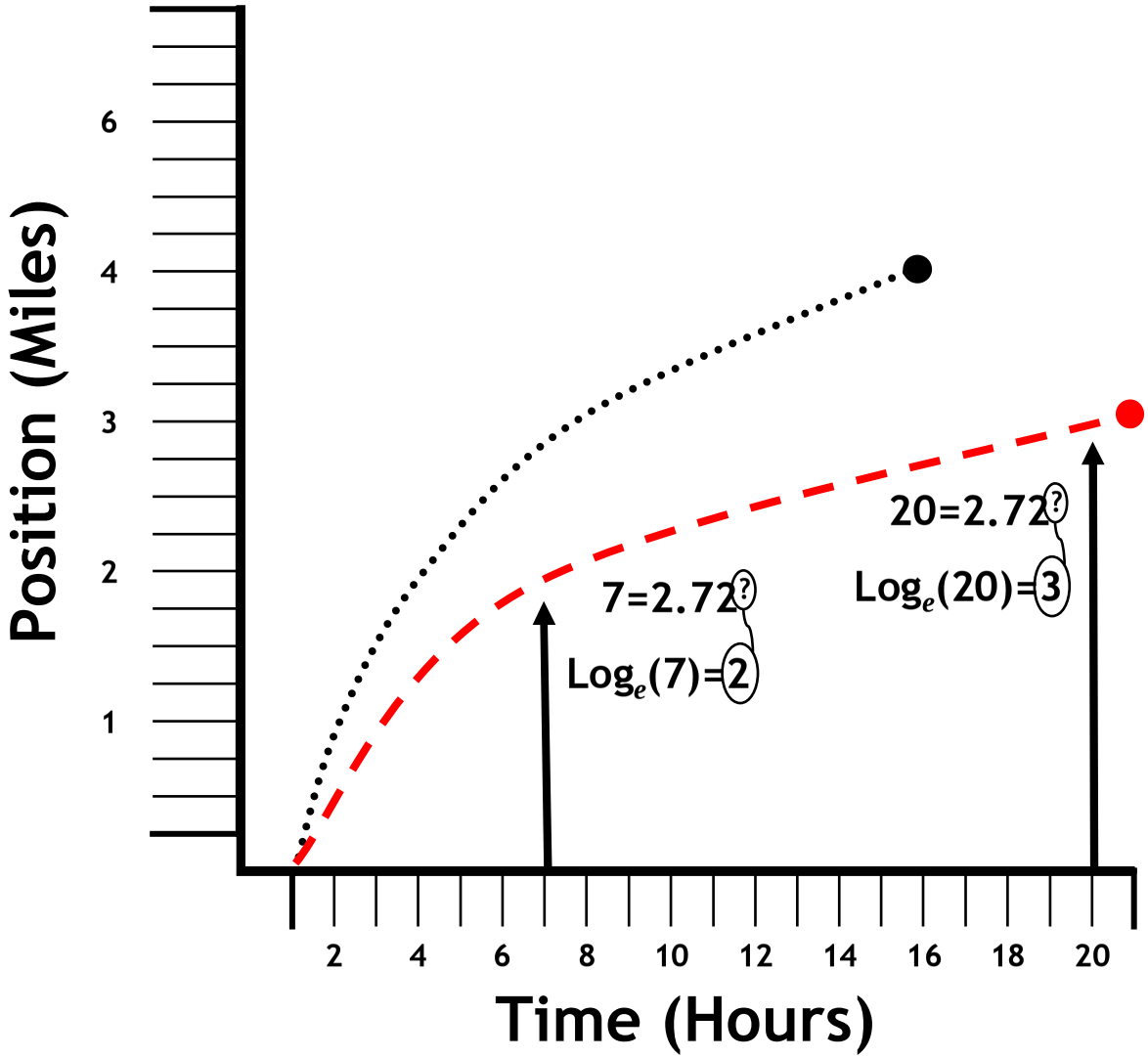
It says that in the first ***hour***, you go one mile. But, it takes you another ***two hours*** to go another mile. Then, it takes you another ***four hours*** to go another mile. Then, it takes you another ***eight hours*** to go another mile… Each additional mile takes you a doubling of hours!

From calculus, we know:

Speed = 1/(Something*Time)

The "Something" is the same exact something as before.

Ln or Log_e or "Log to the Base e" or "The Natural Log"



Suppose:

$\text{Position} = \log_e(\text{Time})$

$\text{Position} = \ln(\text{Time})$

Then:

$\text{Speed} = 1/\text{Time}$

When you've gone 2 miles, you are going 1/2 MPH, and when you've gone 3 miles, you are going 1/3 MPH, and when you've gone 4 miles…

Remember, e is just a number, a particularly snazzy number, much like π is a particularly snazzy number.

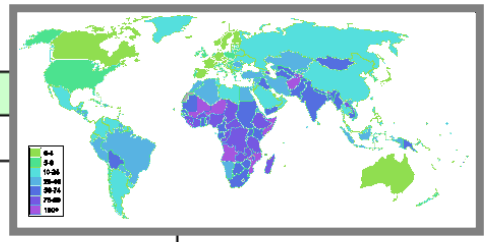
$$e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n \approx 2.7$$

Unit 13 Appendix: SPSS Syntax

```
TEMPORARY.
SELECT IF NOT (ID = 12).
FREQUENCIES VARIABLES=INFMORT PCI
  /FORMAT=NOTABLE
  /STATISTICS=MINIMUM MAXIMUM
  /ORDER=ANALYSIS.
GRAPH
  /HISTOGRAM(NORMAL)=INFMORT.
GRAPH
  /HISTOGRAM(NORMAL)=PCI.
GRAPH
  /SCATTERPLOT(BIVAR)=PCI WITH INFMORT
  /MISSING=LISTWISE.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT INFMORT
  /METHOD=ENTER PCI
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HIST(ZRESID) NORM(ZRESID)
  /SAVE PRED COOK LEVER RESID DRESID.
```

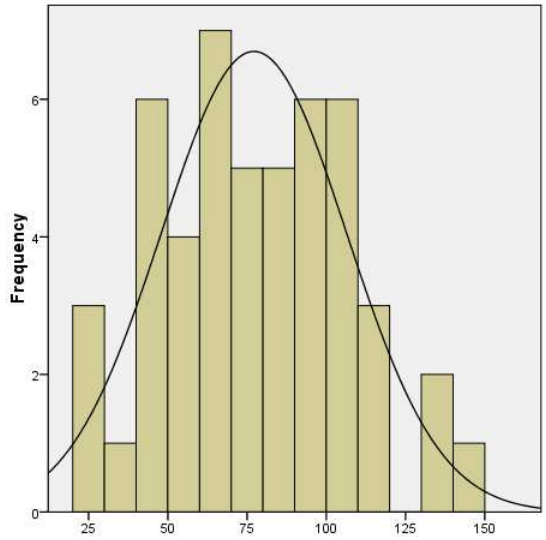
```
*Different Transformations.
COMPUTE L2PCI = LG10(PCI)/LG10(2).
COMPUTE LNPCI = LN(PCI).
COMPUTE LNPCIP16 = LN(PCI+0.16).
COMPUTE PCISQ = PCI**2.
COMPUTE PCICUBED = PCI**3.
COMPUTE PCIROOT = PCI**(1/2).
*Or, equivalently.
COMPUTE PCIROOT = SQRT(PCI).
COMPUTE PCIROOTP7 = SQRT(PCI+7).
EXECUTE.
```

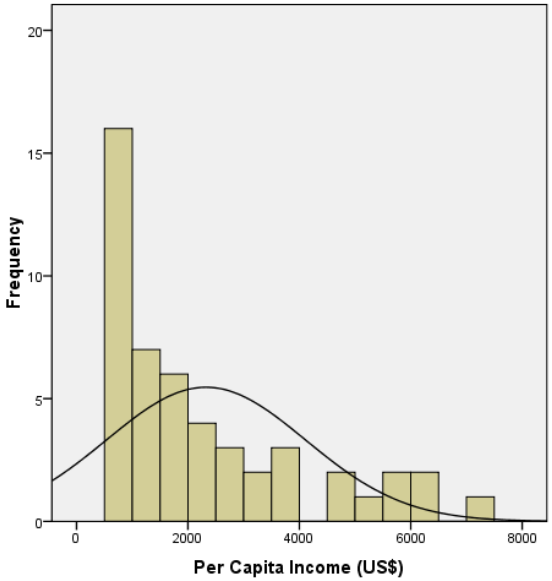
Infant Mortality Rate and Per Capita Income (InfMort.sav)



Infant Mortality Rates in Developing Countries			
Filename	InfMort.txt		
Overview	Is a country's wealth and its prevalence of "high risk" births associated with infant mortality rates?		
Source	Alan Guttmacher Institute (2002). Family planning can reduce high infant mortality levels. *Issues in Brief, 2002 Series, 2*. Retrieved February 28, 2008, from http://www.guttmacher.org/pubs/ib_2-02.html		
Sample Size	49 developing countries		
Updated	28 February, 2008, Tara Chiatovich		
Contents			
Col #s	Var Name	Description	Metric / Comments
1-2	ID	ID Number	Integer
4-23	COUNTRY	Name of country	Alphanumeric
27-29	INFMORT	Infant morality rate	Number of deaths of infants under the age of 1 year per 1,000 births.
31-34	PCI	Per capita income	Per capita income in U.S. dollars.
36-37	YOUNGMOM	Births to young mothers	Percentage of live births to mothers under the age of 20.
39-40	OLDMOM	Births to older mothers	Percentage of live births to mothers aged 40 or older.
42-43	CLOSE	Close births	Percentage of births less than two years apart.

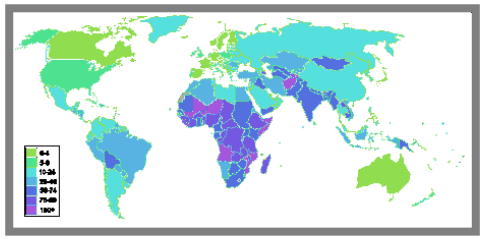
Infant Mortality Rate and Per Capita Income (InfMort.sav)

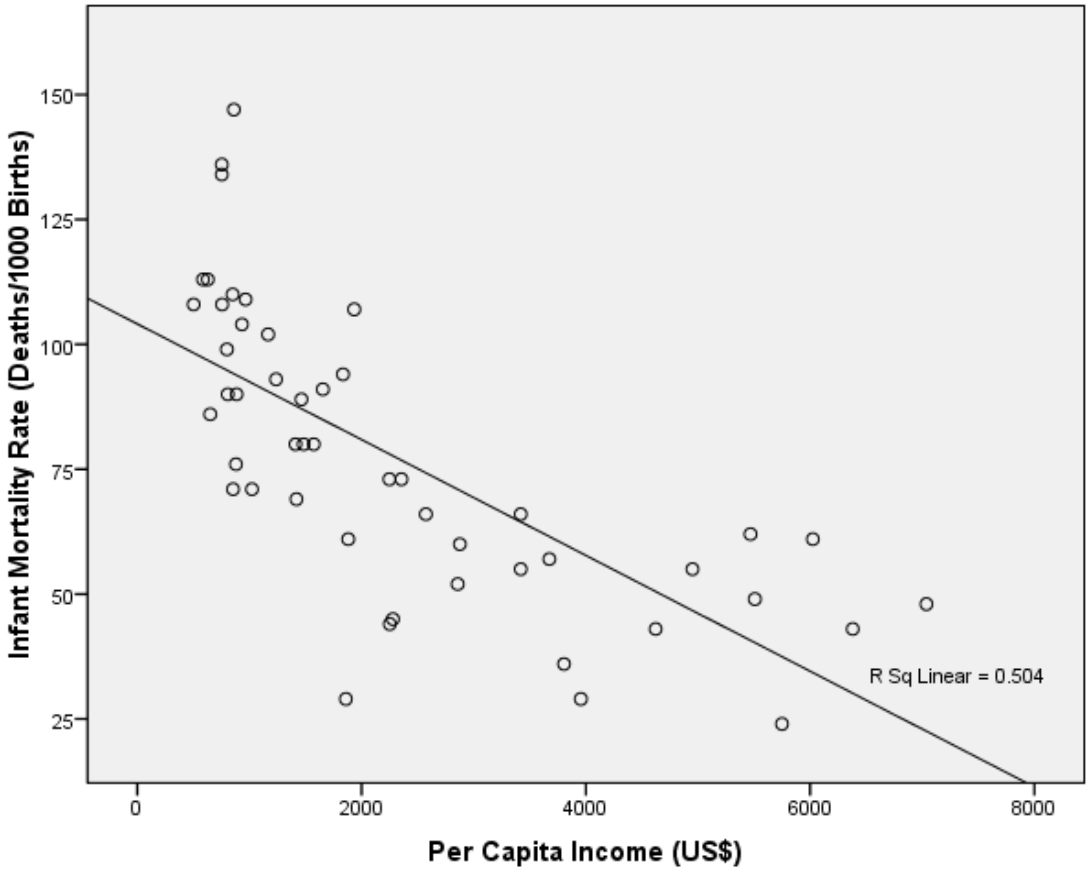




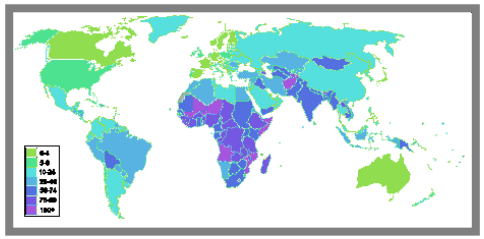
Statistics

		Infant Mortality Rate (Deaths/1000 Births)	Per Capita Income (US$)
N	Valid	49	49
	Missing	0	0
	Minimum	24	501
	Maximum	147	7037





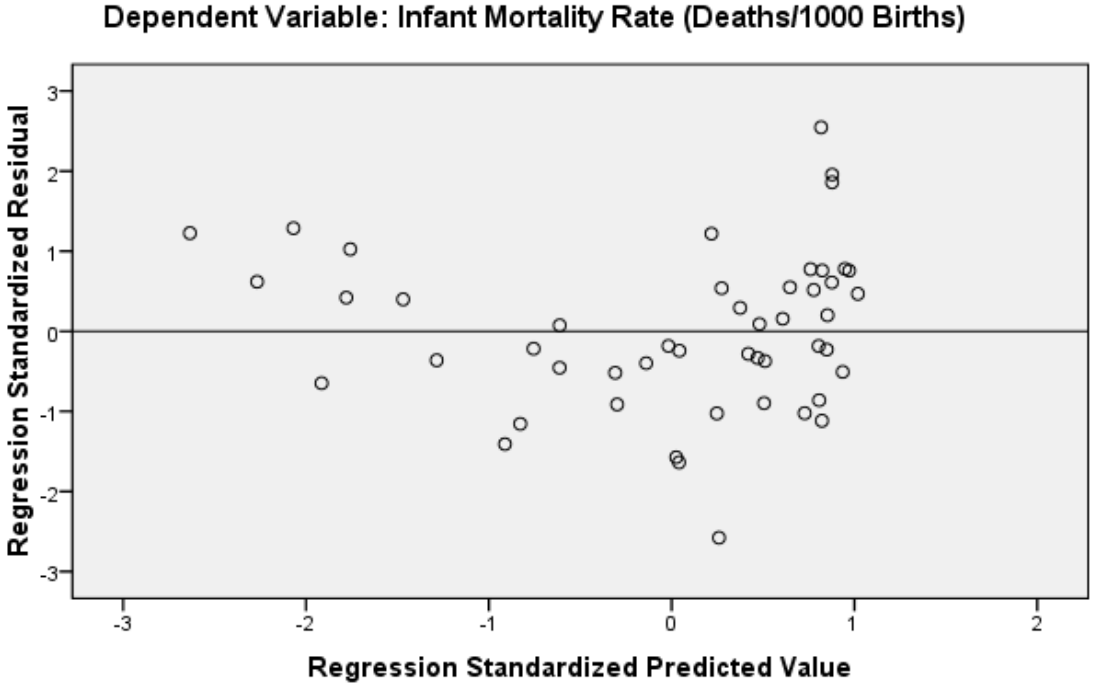
Infant Mortality Rate and Per Capita Income (InfMort.sav)



Coefficients[a]

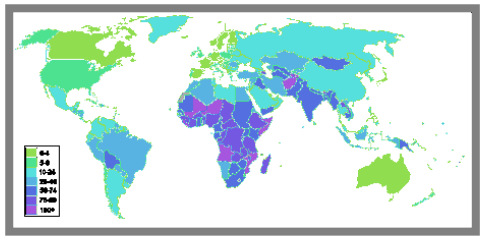
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	104.086	4.895		21.263	.000	94.239	113.934
	Per Capita Income (US$)	-.012	.002	-.710	-6.915	.000	-.015	-.008

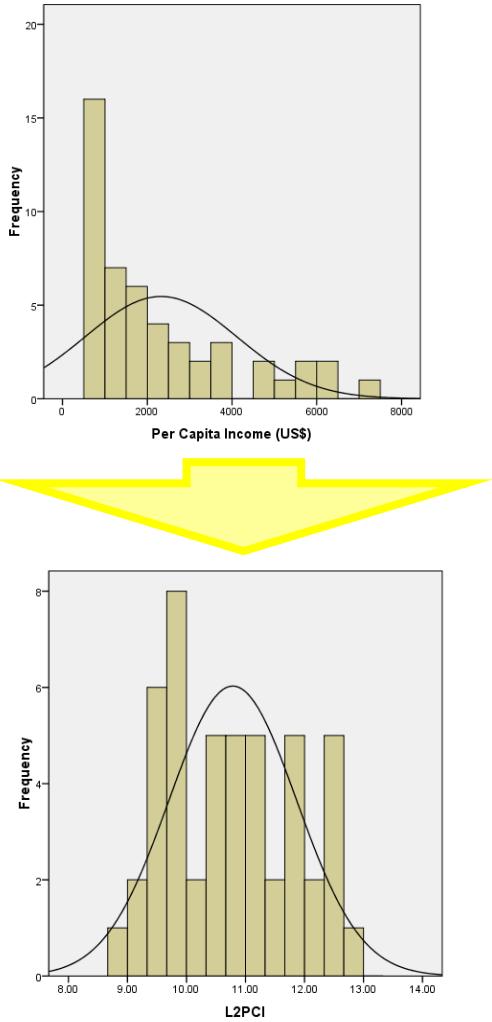
a. Dependent Variable: Infant Mortality Rate (Deaths/1000 Births)

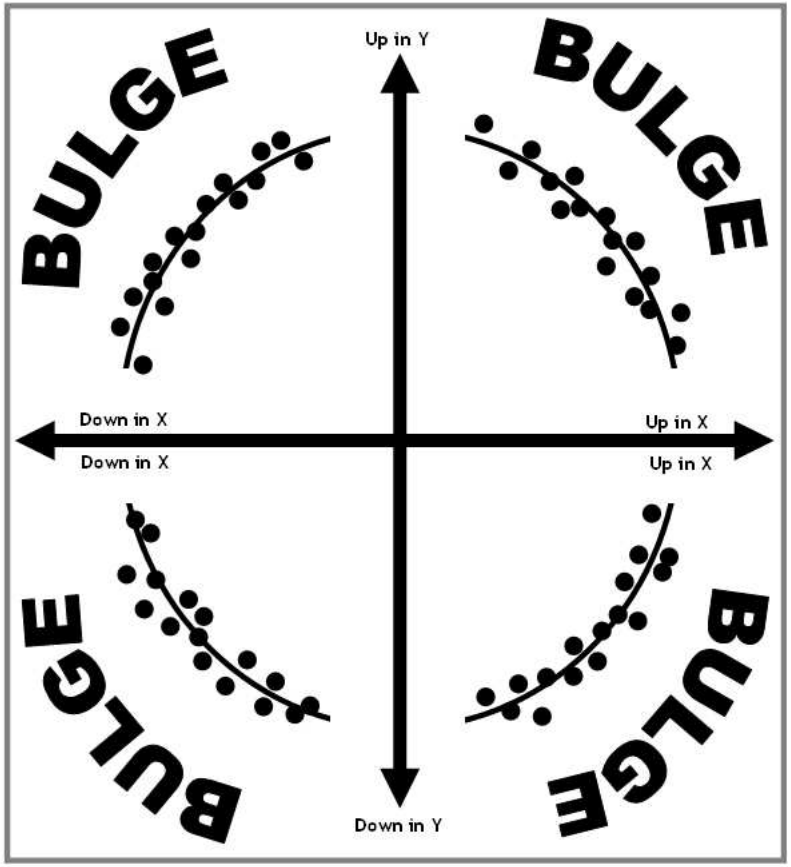


Infant Mortality Rate and Per Capita Income (InfMort.sav)

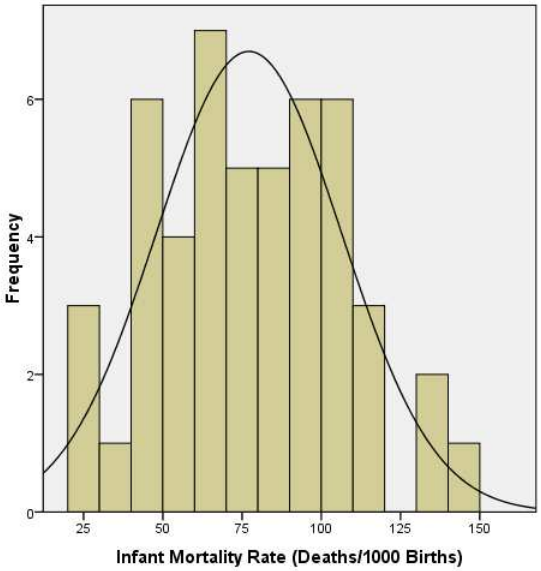
```
COMPUTE L2PCI = LG10(PCI)/LG10(2).
EXECUTE.
```

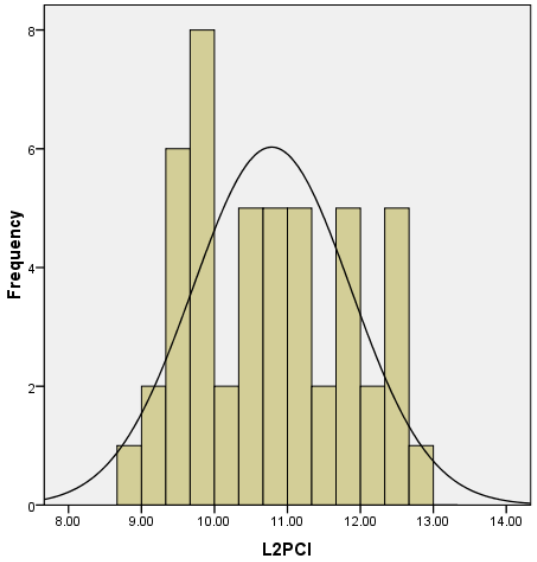


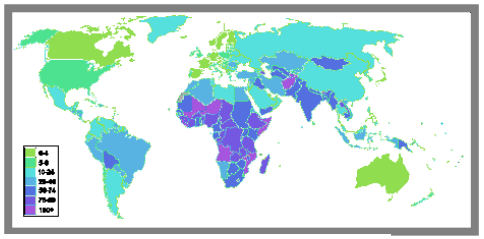


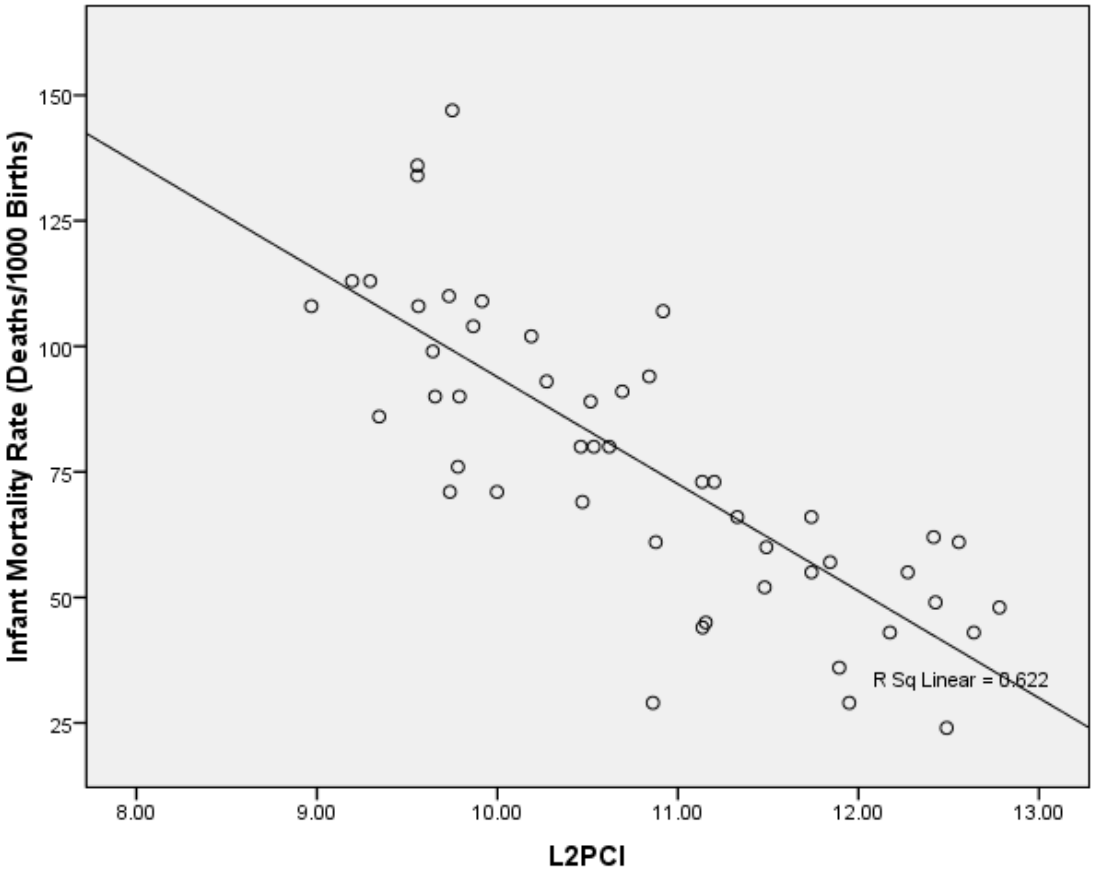


Infant Mortality Rate and Per Capita Income (InfMort.sav)

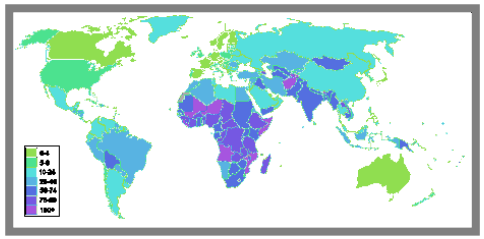








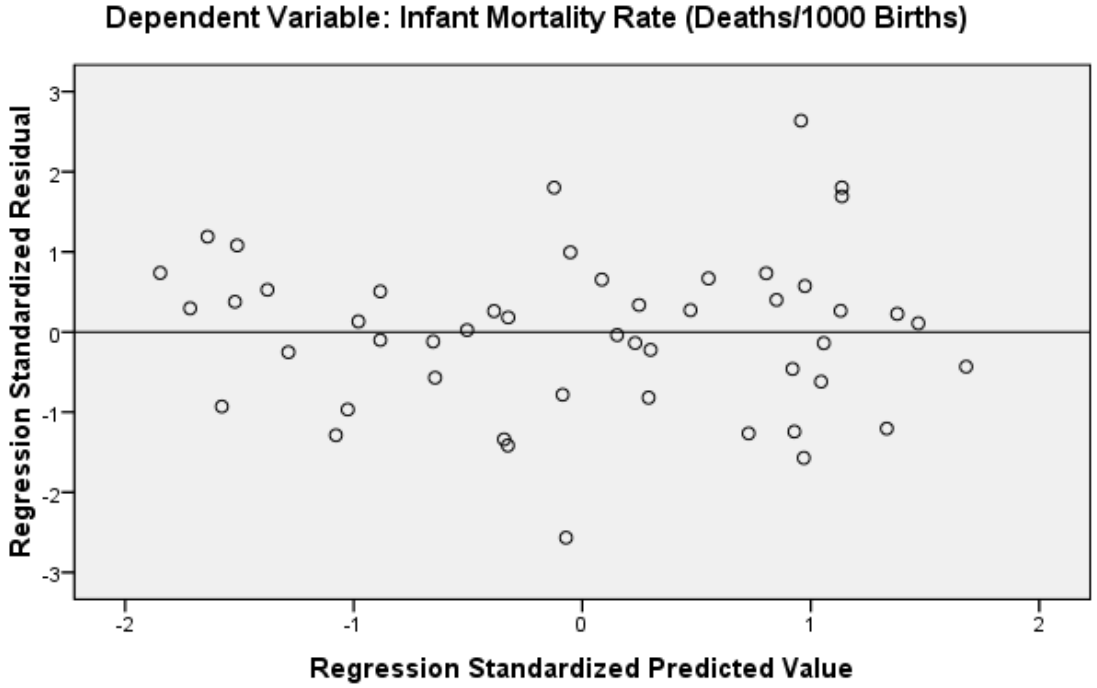
Infant Mortality Rate and Per Capita Income (InfMort.sav)



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	306.899	26.240		11.696	.000	254.111	359.687
	L2PCI	-21.302	2.421	-.789	-8.798	.000	-26.173	-16.432

a. Dependent Variable: Infant Mortality Rate (Deaths/1000 Births)

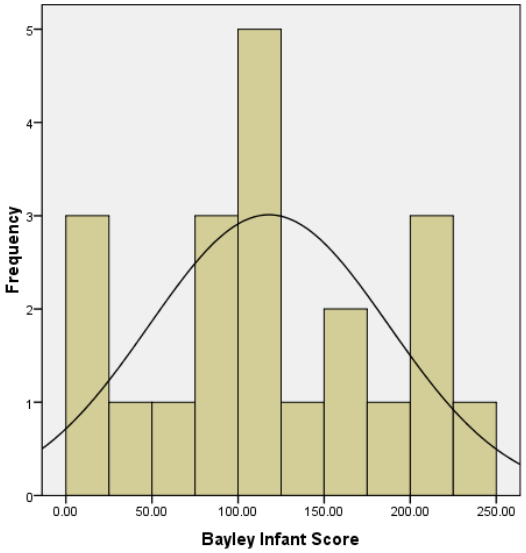


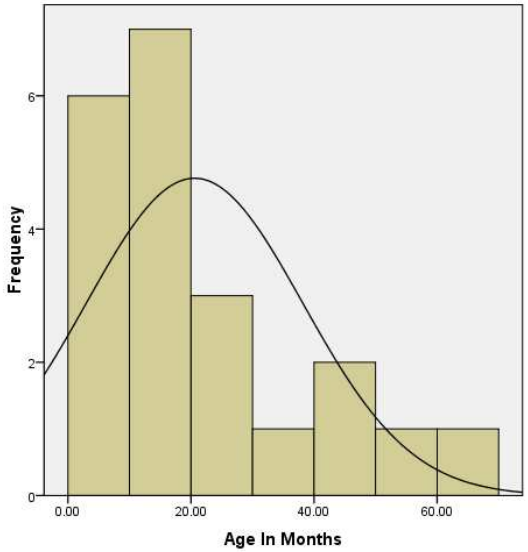
Bayley's Infant IQ (Bayley.sav)

Dataset	BAYLEY.txt
Overview	IQ as a function of age for a female infant, from birth to age 60 months.
Source	Target child is a female infant (infant #8) from the *Berkeley Growth and Guidance Study*.
More Info	To learn more about the data, consult: • The overview of the *Oakland and Berkeley Growth and Guidance Studies* at the *Carolina Population Center*. • Glen Elder's presentation on "Longitudinal Studies and the Life Course, the 1960s and 1970s," prepared for the anniversary of the *Institute of Human Development*, UC Berkeley (2003).
Sample size	One infant, over 21 occasions of measurement.
Last updated	August 28, 2003

Structure of Dataset			
Col. #	**Variable Name**	**Variable Description**	**Variable Metric/Labels**
1	Score	Infant's score on the *Bayley Scales of Infant Development*	Continuous raw score
2	Age	Age of infant	Months

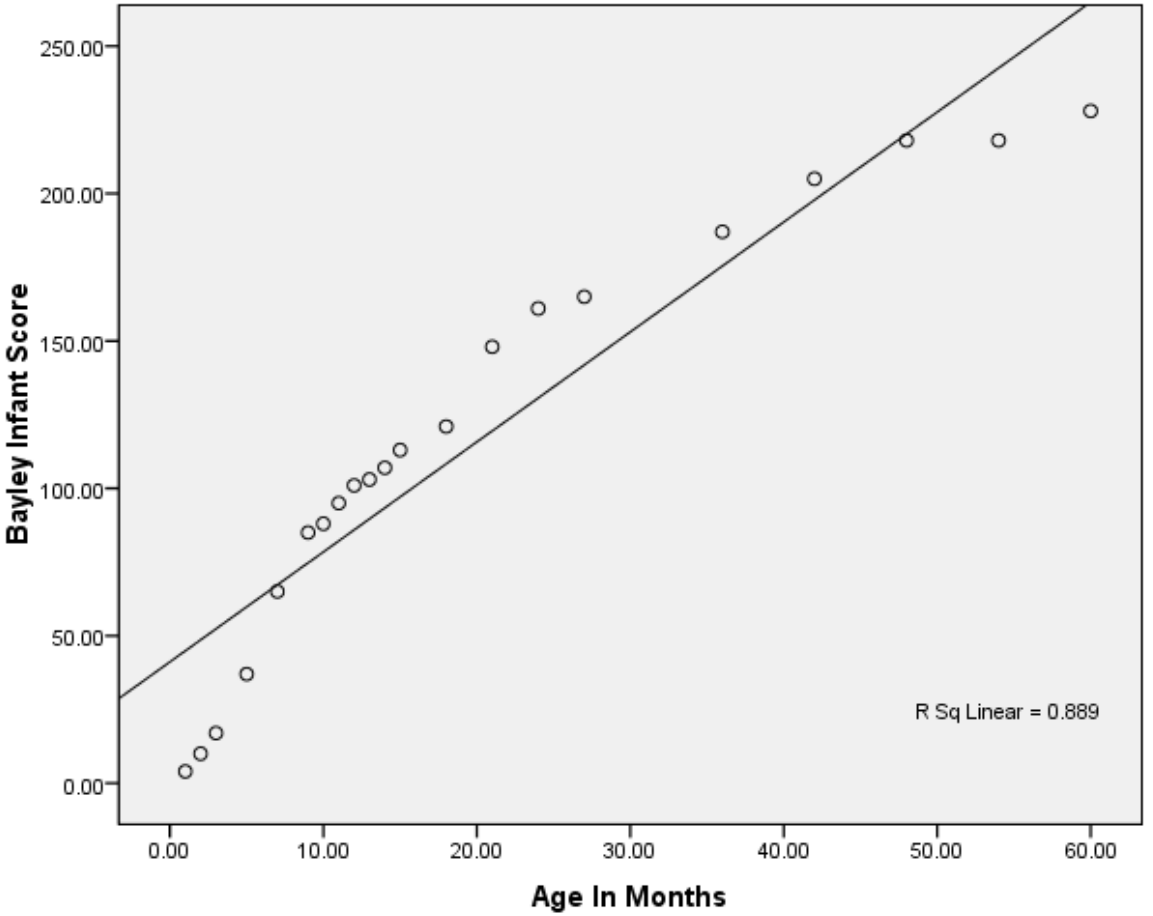
Bayley's Infant IQ (Bayley.sav)





Statistics

		Bayley Infant Score	Age In Months
N	Valid	21	21
	Missing	0	0
	Minimum	4.00	1.00
	Maximum	228.00	60.00

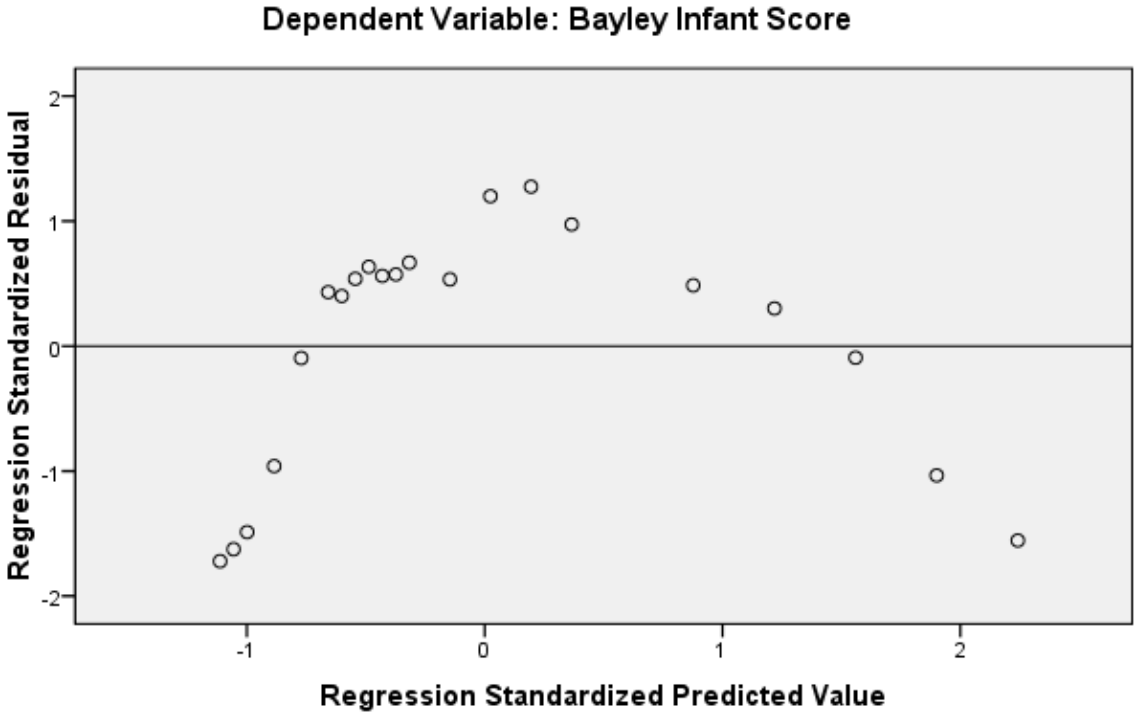


Bayley's Infant IQ (Bayley.sav)

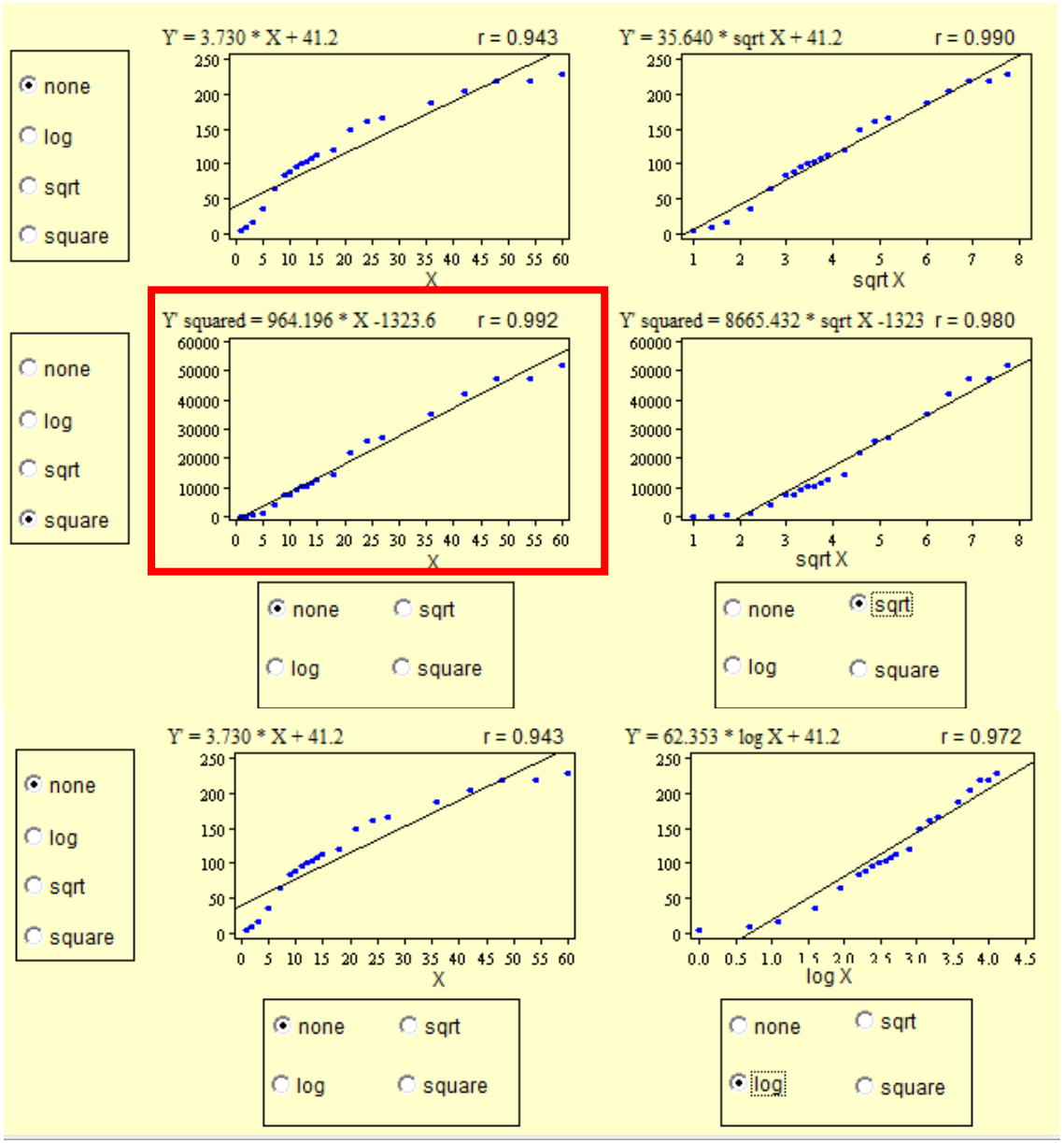
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	41.176	8.095		5.087	.000	24.234	58.119
	Age In Months	3.730	.302	.943	12.345	.000	3.097	4.362

a. Dependent Variable: Bayley Infant Score



Bayley's Infant IQ (Bayley.sav)



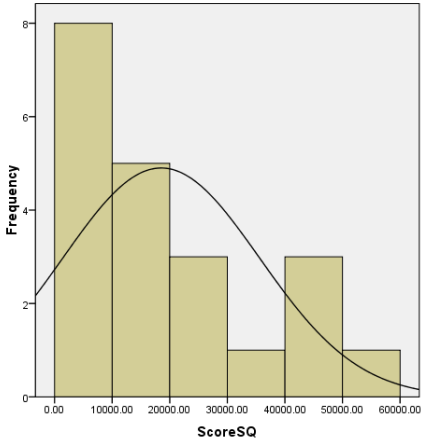
http://onlinestatbook.com/stat_sim/transformations/index.html

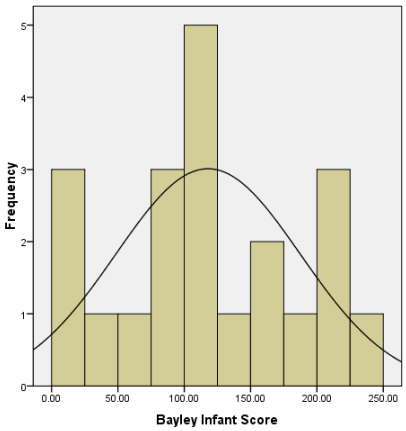
Which is the best fit?
I'm thinking Y^2.

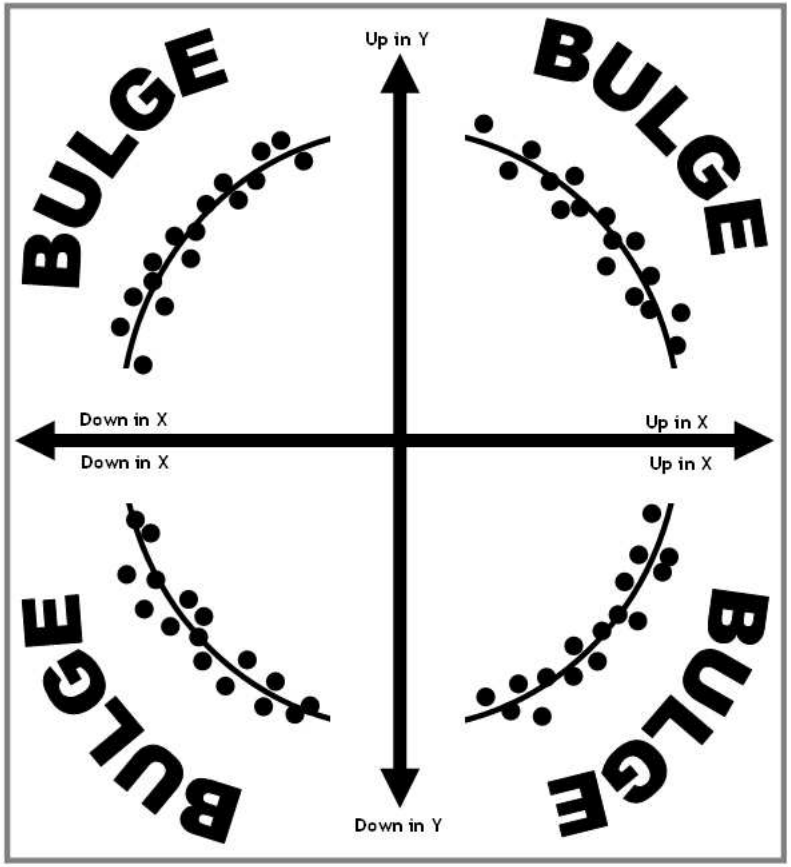
1	4
2	10
3	17
5	37
7	65
9	85
10	88
11	95
12	101
13	103
14	107
15	113
18	121
21	148
24	161
27	165
36	187
42	205
48	218
54	218
60	228

Bayley's Infant IQ (Bayley.sav)

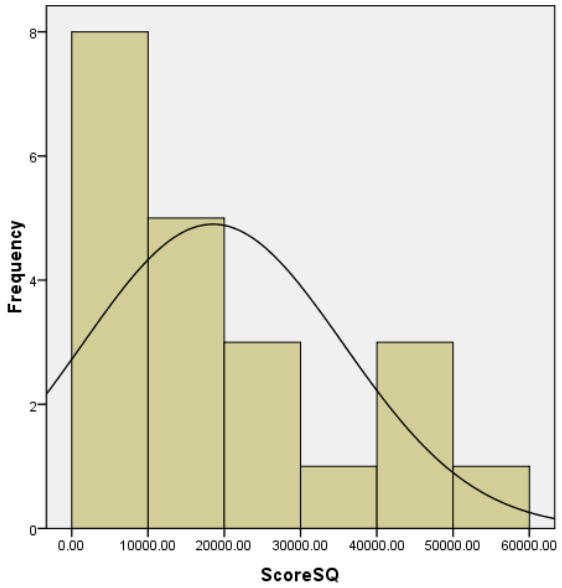
```
COMPUTE ScoreSQ = Score**2.
EXECUTE.
```

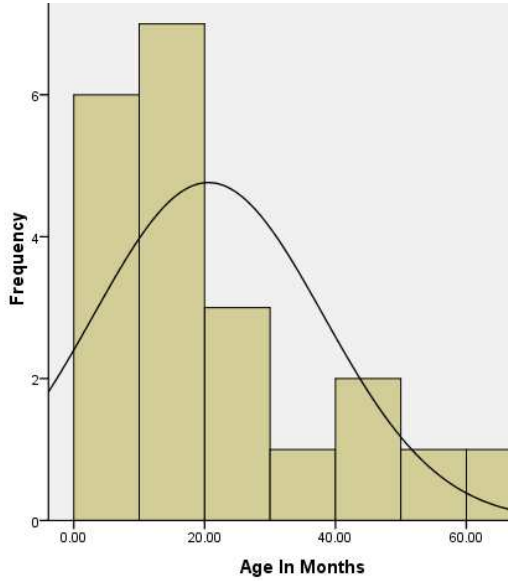


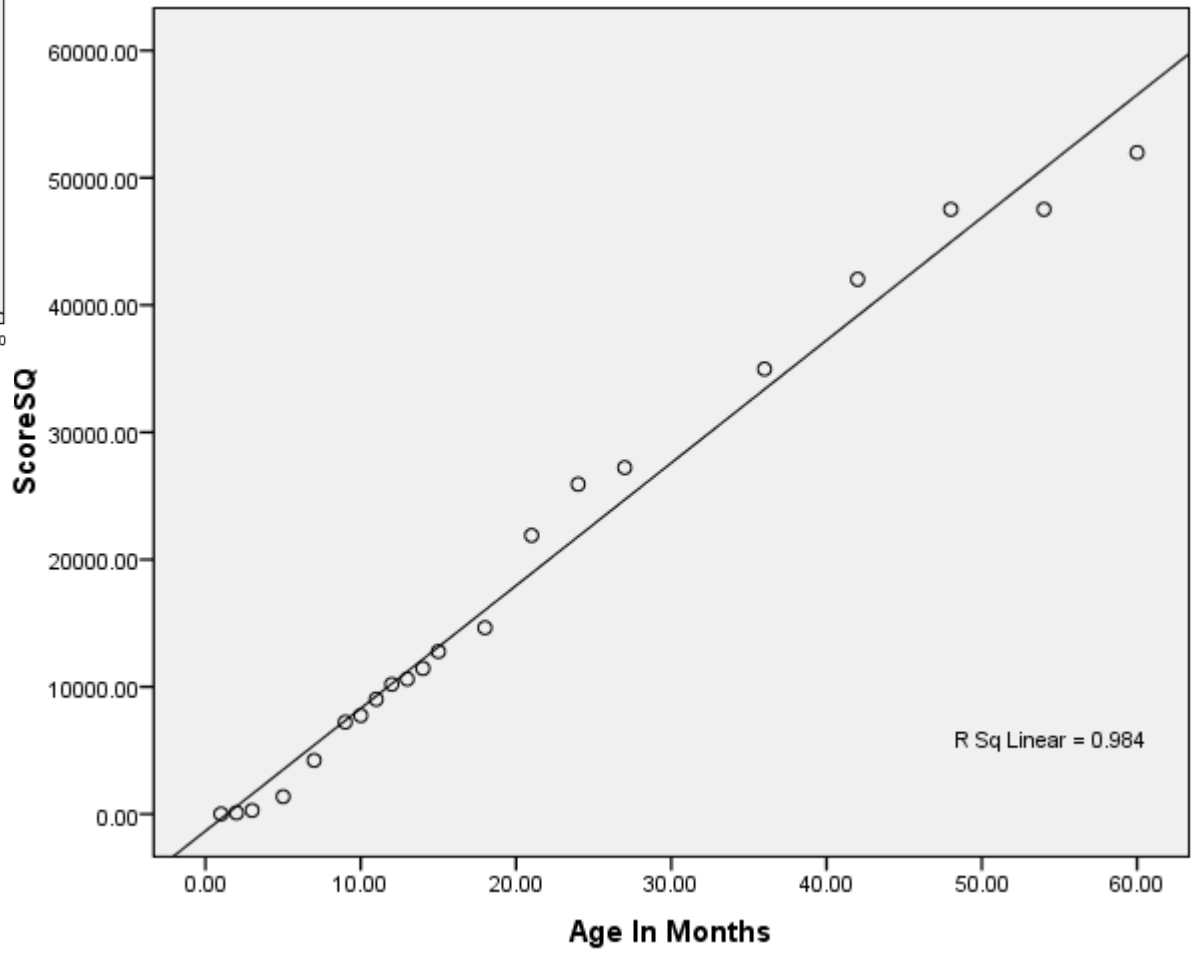




Bayley's Infant IQ (Bayley.sav)





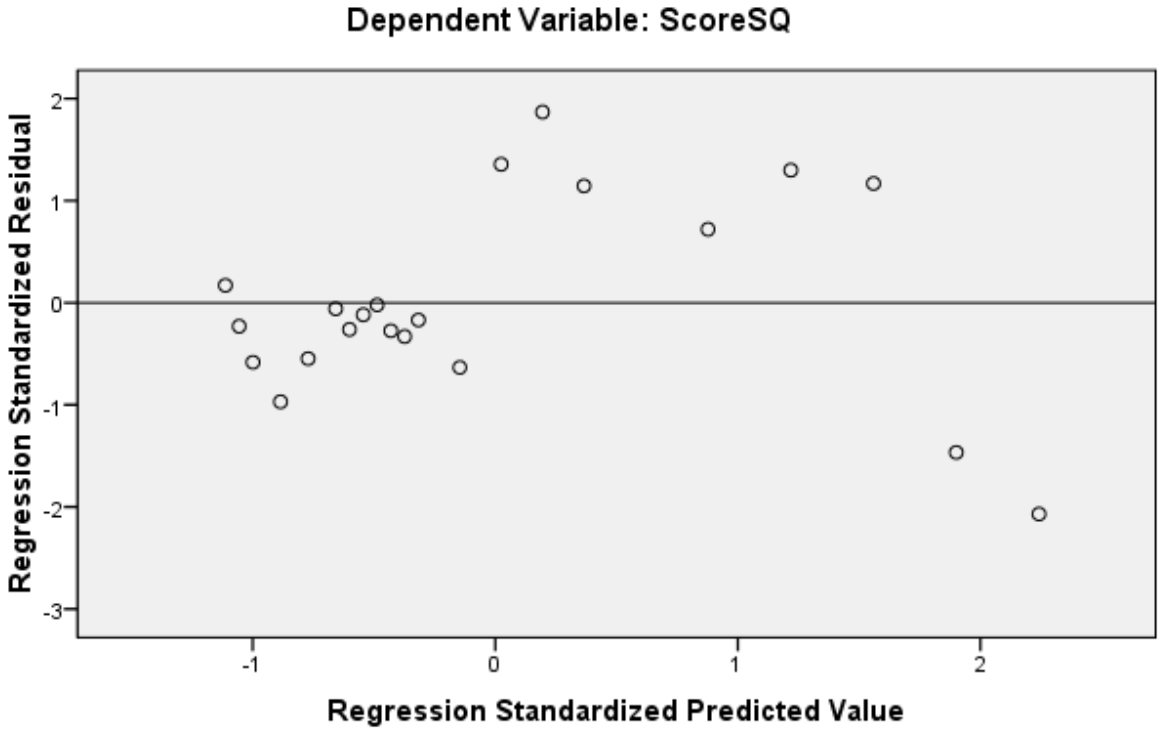


Bayley's Infant IQ (Bayley.sav)

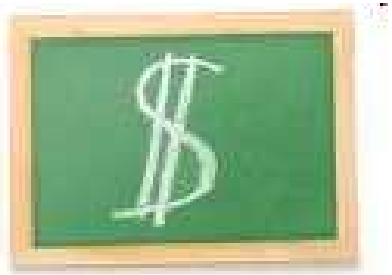
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-1323.561	748.132		-1.769	.093	-2889.420	242.297
	Age In Months	964.196	27.924	.992	34.530	.000	905.751	1022.641

a. Dependent Variable: ScoreSQ



GDP and PPE (OECD.sav)

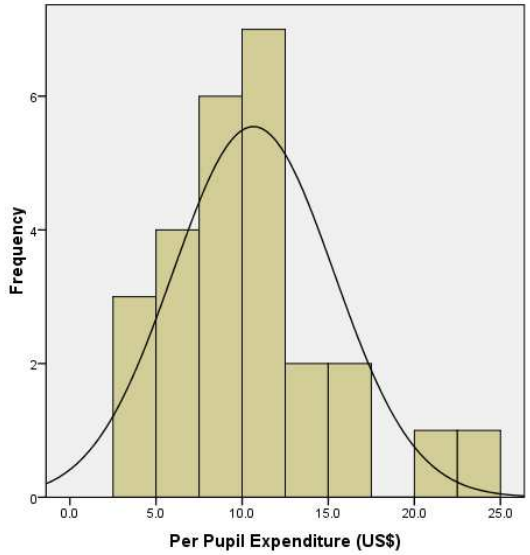


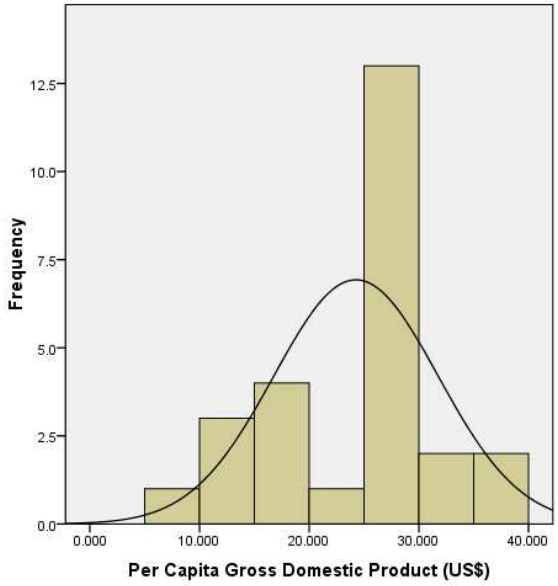
OECD Tertiary PPE and Per Capita GDP	
Filename	OECD.txt
Overview	OECD issues an annual report, *Education at a Glance*, which profiles the 26 OECD countries across a variety of education and economic measures. This dataset is described in their Chart B1.6, page 166, which examines the relationship between tertiary per pupil expenditures (PPE) and per capita gross domestic product (GDP)
Source	http://www.oecd.org
Sample size	26 countries
Updated	4 January 2006, Judith D. Singer

Contents

Col #s	Var Name	Description	Metric/Comments
1-2	ID	Country ID	Integer
4-24	Country	Country name	Alphanumeric
27-33	GDP	Per capita gross domestic product	US Dollars
38-50	PPE	Expenditure per student	US Dollars

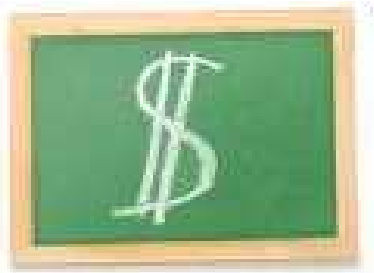
GDP and PPE (OECD.sav)

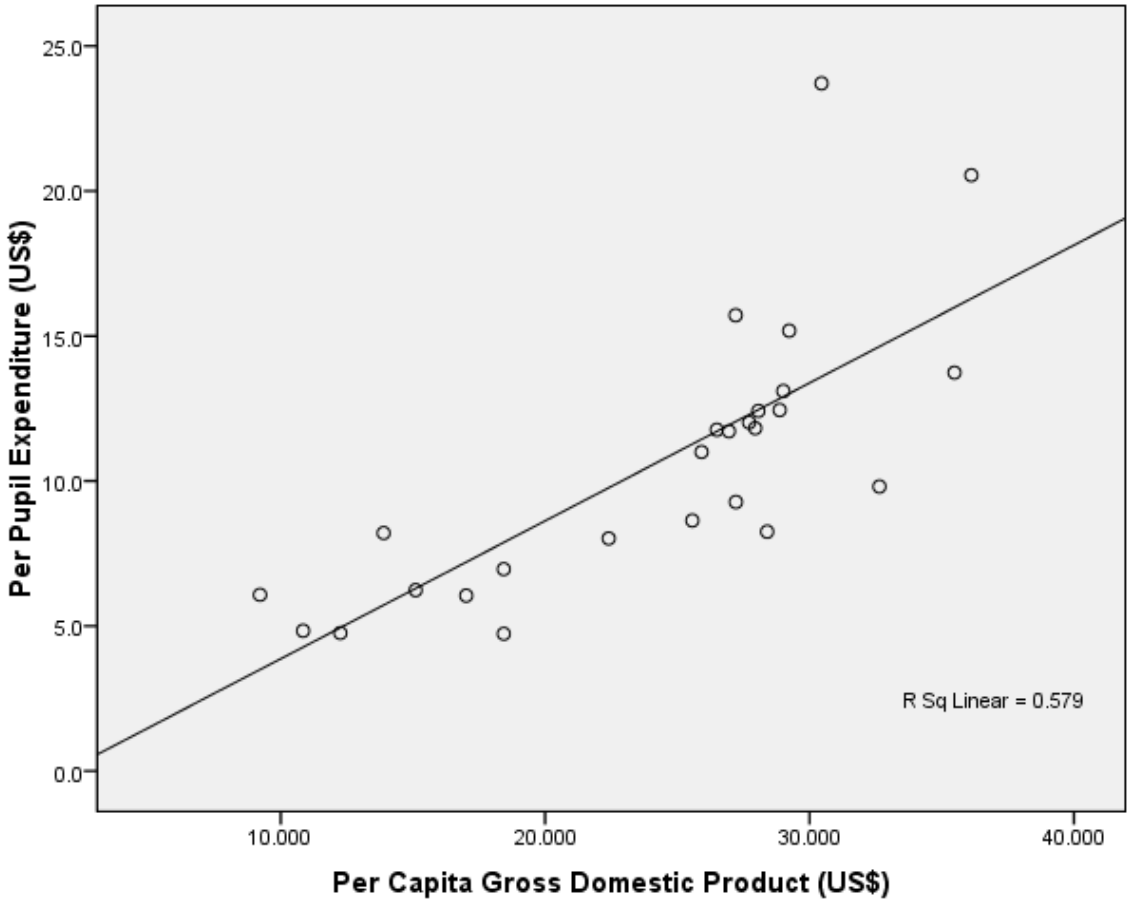




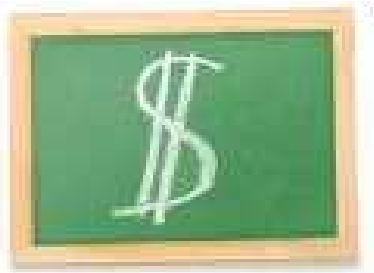
Statistics

		Per Pupil Expenditure (US$)	Per Capita Gross Domestic Product (US$)
N	Valid	26	26
	Missing	0	0
	Minimum	4.7	9.215
	Maximum	23.7	36.121





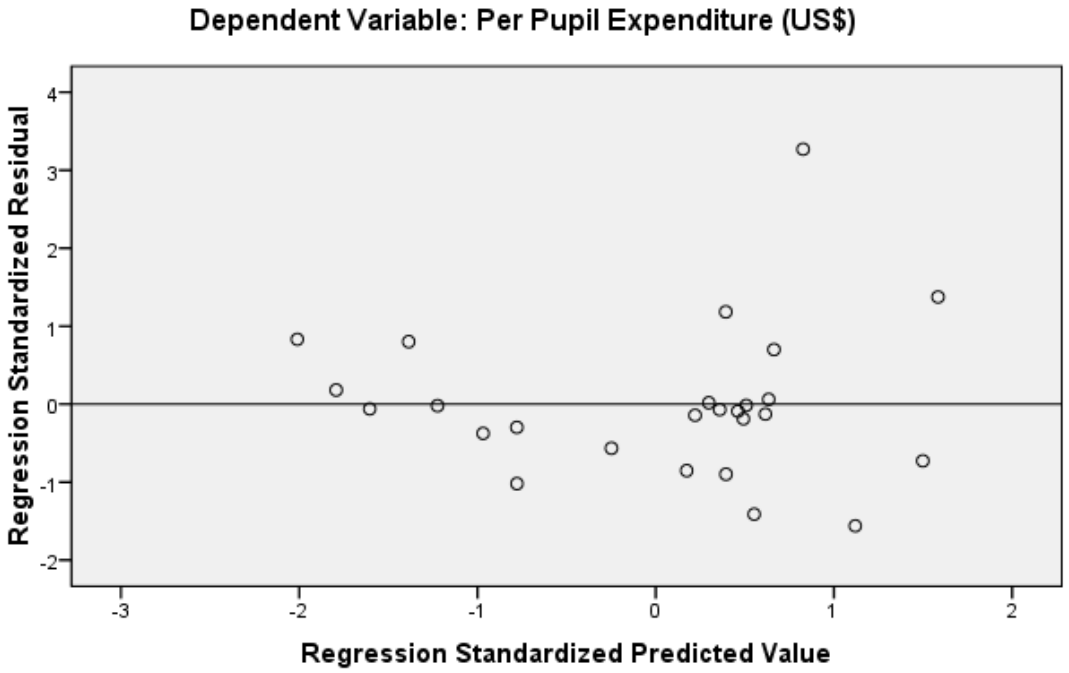
GDP and PPE (OECD.sav)



Coefficients[a]

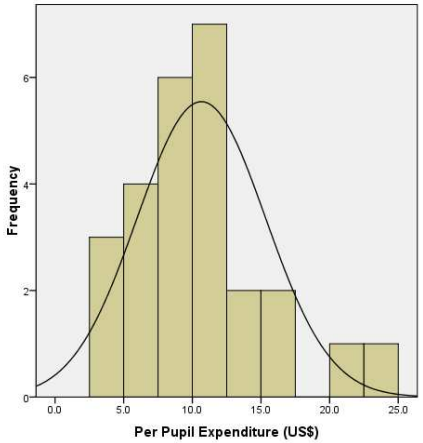
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-.883	2.097		-.421	.677	-5.211	3.444
	Per Capita Gross Domestic Product (US$)	.475	.083	.761	5.749	.000	.305	.646

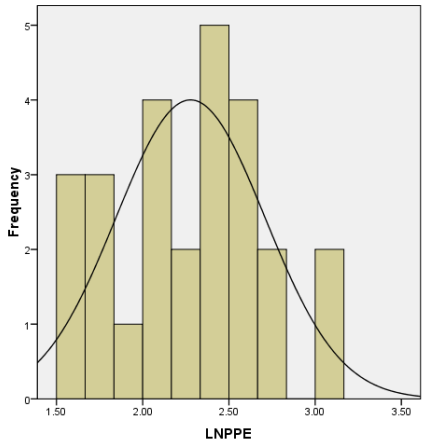
a. Dependent Variable: Per Pupil Expenditure (US$)

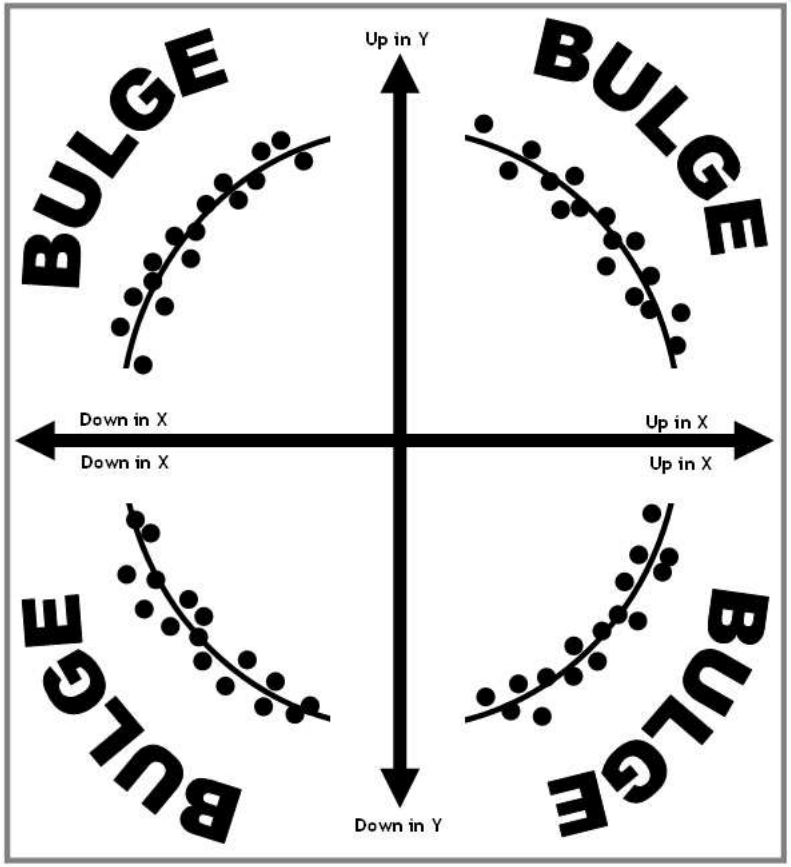


GDP and PPE (OECD.sav)

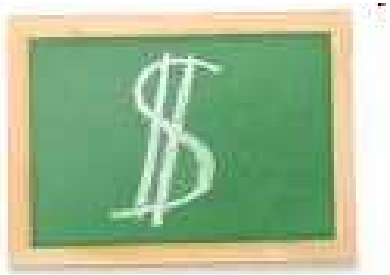
```
COMPUTE LNPPE = LN(PPE).
EXECUTE.
```

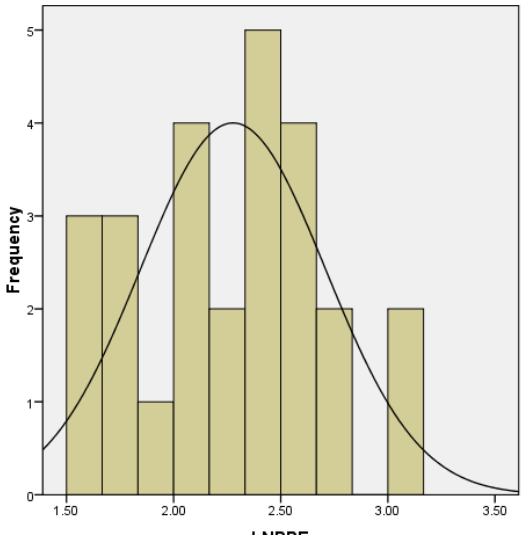


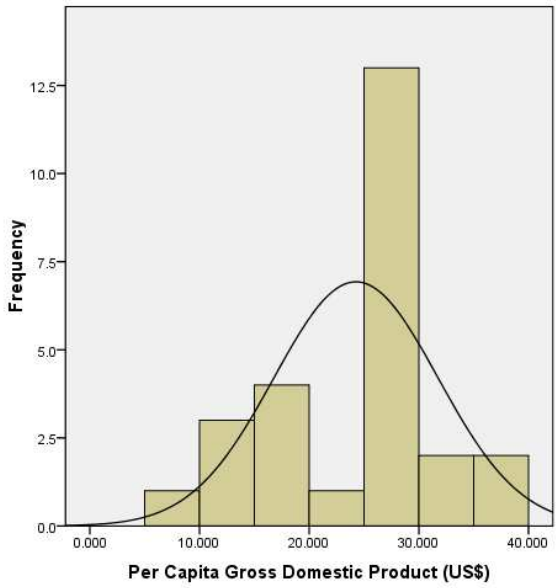


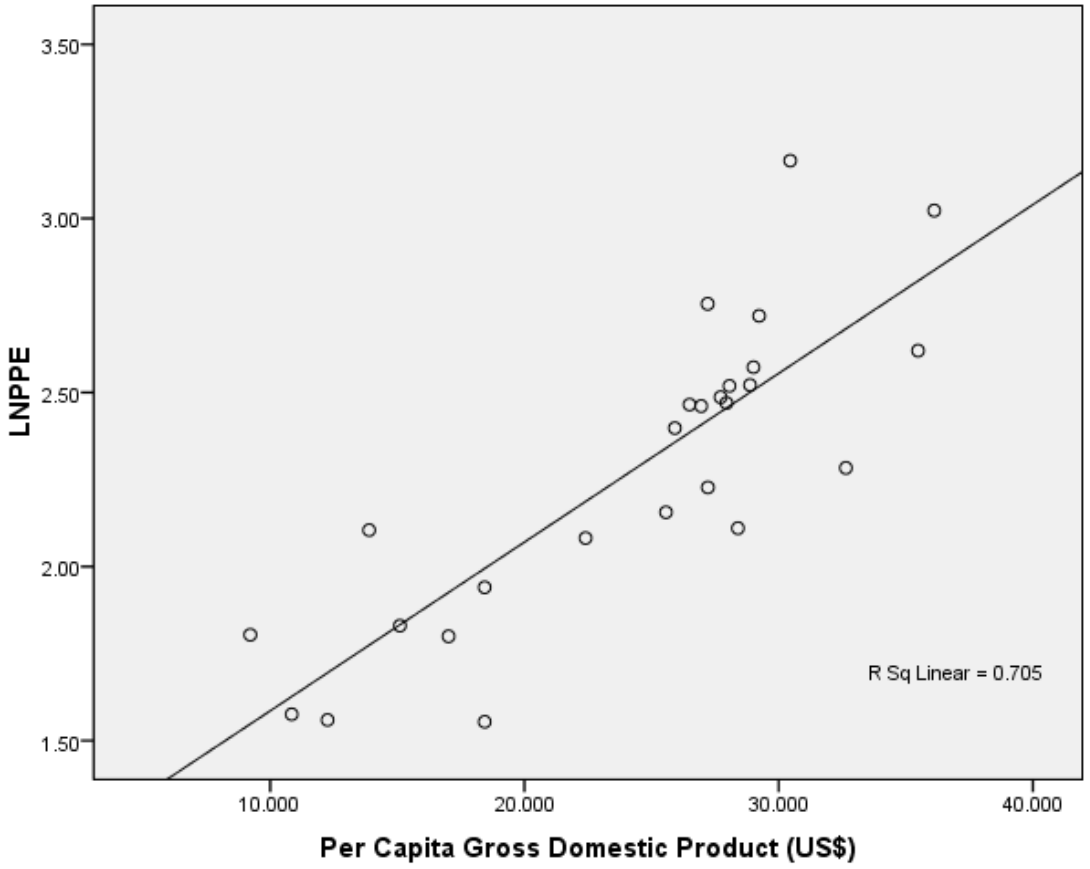


GDP and PPE (OECD.sav)

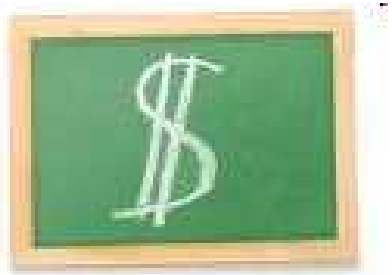








GDP and PPE (OECD.sav)



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	1.101	.162		6.787	.000	.766	1.436
	Per Capita Gross Domestic Product (US$)	.048	.006	.840	7.575	.000	.035	.062

a. Dependent Variable: LNPPE