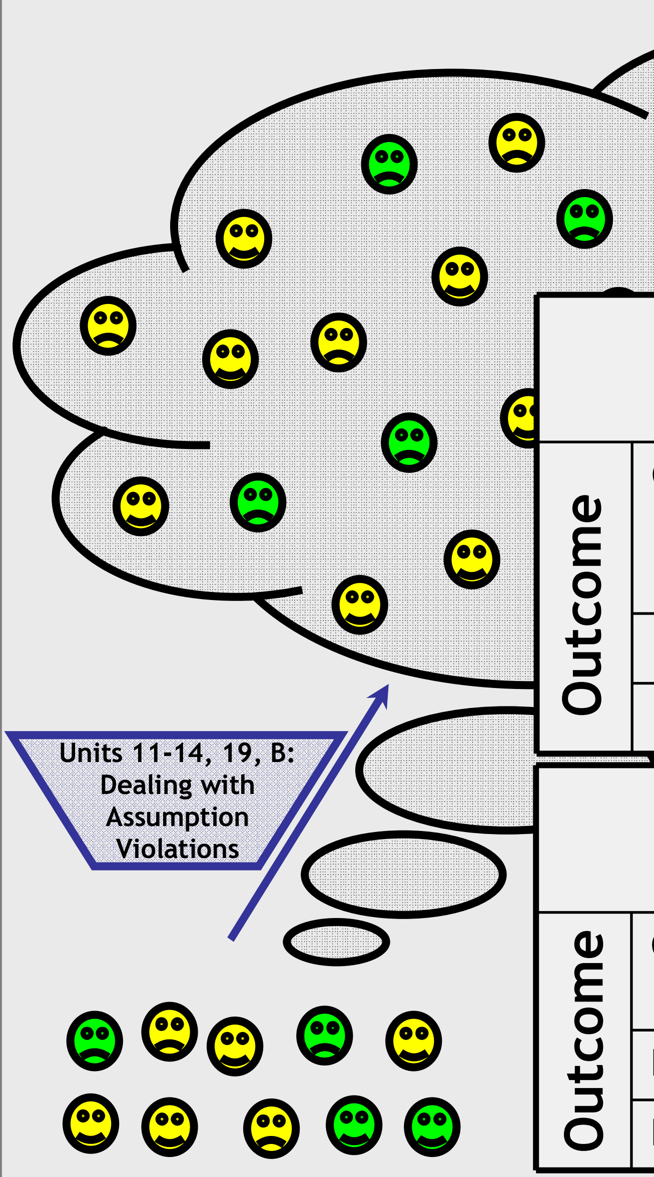
Unit 16: Road Map (VERBAL)

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 16: 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 16: Roadmap (SPSS Output)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	48.338	.110		438.242	.000	48.122	48.554
	ASIAN	1.034	.383	.030	2.697	.007	.283	1.786
	BLACK	-4.889	.339	-.161	-14.423	.000	-5.554	-4.225
	LATINO	-4.418	.306	-.161	-14.447	.000	-5.017	-3.818
2	(Constant)	43.878	.280		156.558	.000	43.328	44.427
	ASIAN	.727	.377	.021	1.929	.054	-.012	1.465
	BLACK	-4.796	.333	-.158	-14.412	.000	-5.448	-4.144
	LATINO	-4.123	.301	-.151	-13.715	.000	-4.712	-3.534
	L2HOMEWORKP1	1.766	.102	.188	17.254	.000	1.565	1.967
3	(Constant)	45.381	.284		159.528	.000	44.823	45.938
	ASIAN	.461	.441	.013	1.045	.296	-.404	1.325
	BLACK	-3.622	.331	-.119	-10.956	.000	-4.270	-2.974
	LATINO	-3.311	.366	-.121	-9.035	.000	-4.029	-2.592
	L2HOMEWORKP1	1.603	.100	.170	15.974	.000	1.406	1.799
	ESL	.218	.363	.009	.600	.548	-.494	.930
	FREELUNCH	-3.867	.199	-.213	-19.452	.000	-4.256	-3.477
4	(Constant)	45.358	.288		157.560	.000	44.794	45.923
	ASIAN	-.377	.668	-.011	-.564	.573	-1.687	.933
	BLACK	-3.447	.498	-.113	-6.922	.000	-4.423	-2.471
	LATINO	-2.779	.517	-.102	-5.371	.000	-3.793	-1.765
	L2HOMEWORKP1	1.591	.100	.169	15.866	.000	1.394	1.788
	ESL	-.876	.638	-.035	-1.373	.170	-2.126	.374
	FREELUNCH	-3.574	.235	-.197	-15.208	.000	-4.035	-3.113
	ESLxASIAN	3.245	.999	.080	3.249	.001	1.287	5.202
	ESLxBLACK	5.872	1.885	.036	3.115	.002	2.177	9.568
	ESLxLATINO	.446	.858	.013	.520	.603	-1.235	2.127
	FREELUNCHxASIAN	-2.769	.853	-.041	-3.245	.001	-4.442	-1.096
	FREELUNCHxBLACK	-.751	.666	-.019	-1.127	.260	-2.058	.555
	FREELUNCHxLATINO	-.437	.604	-.012	-.724	.469	-1.622	.747

a. Dependent Variable: READING

Unit 18
Unit 9
Unit 16
Unit 17
Unit 8
Unit 11
Unit 12
Unit 13
Unit 14

Unit 16: Multiple Regression

Unit 16 Post Hole:

Interpret a fitted multiple regression model.

Unit 16 Technical Memo and School Board Memo:

Fit and interpret a multiple regression model with your variables from Memo 15.

Unit 16 Review:

Review Unit 9.

Unit 16 Supplementary Reading:

Meyers et al. Chapters 5a and 5b.

Unit 16: 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 <u>descriptive statistics</u>. 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 <u>descriptive statistics</u>. 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 16: 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 I6: Research Question

Theory: Head Start programs provide educationally disadvantaged preschoolers the skills and knowledge to start kindergarten on a level playing field.

Research Question: Controlling for ***SES***, ***ESL*** and ***AGE***, is ***GENERALKNOWLEDGE*** positively correlated with ***HEADSTARTHOURS*** for Latina kindergarteners.

Data Set: ECLS (Early Childhood Longitudinal Study) subset of Latinas with no missing data for the variables below (n = 816)

Variables:

Outcome: (***GENERALKNOWLEDGE***) IRT Scaled Score on a Standardized Test of General Knowledge in Kindergarten

Question Predictor: (***HEADSTARTHOURS***) Hours Per Week of Head Start in the Year Before Kindergarten

Control Predictors:

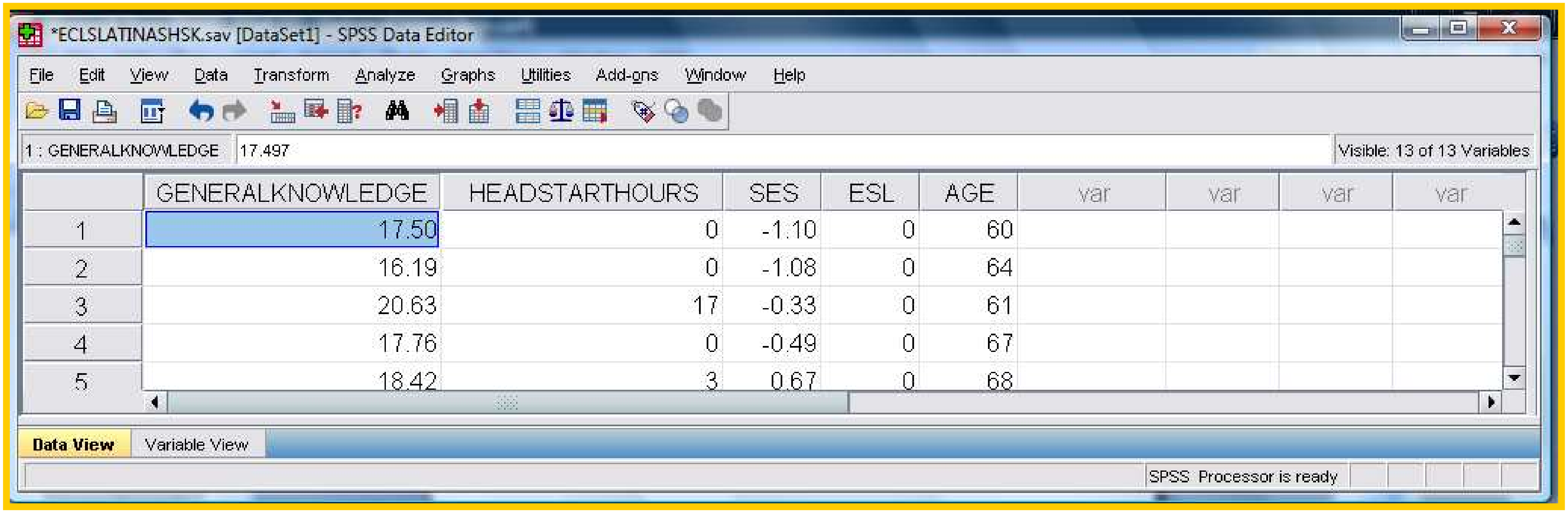
(***SES***) A Composite Measure of the Family's Socioeconomic Status

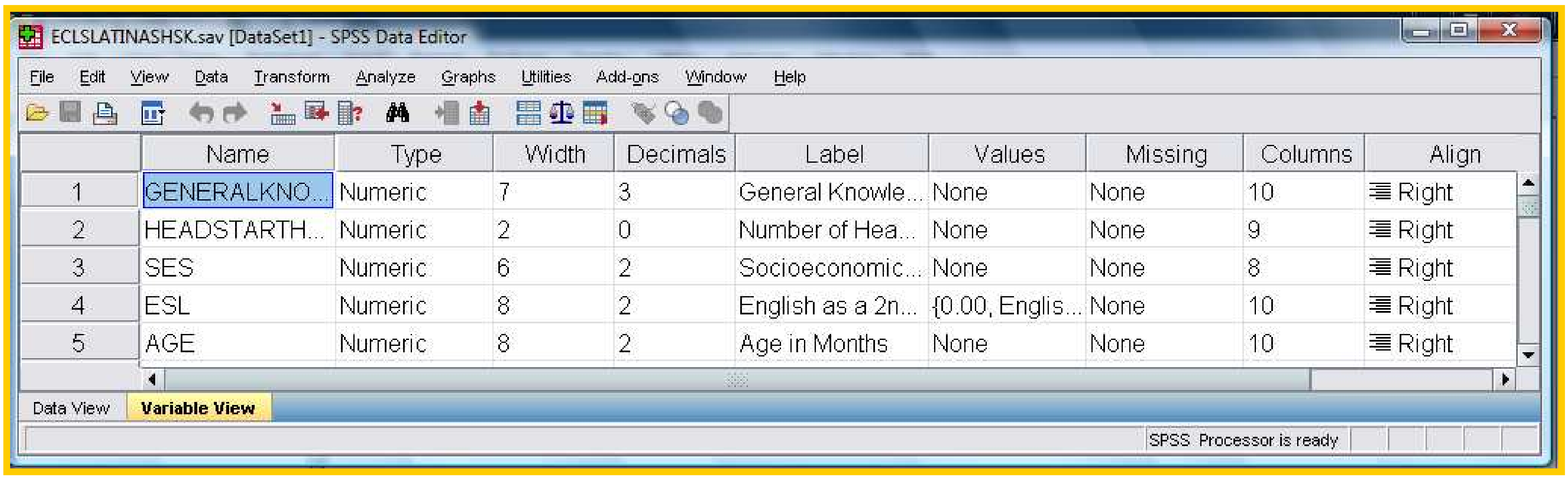
(***ESL***) A Dichotomy for which 1 Denotes that English is a 2^{nd} Language (0 = Not)

(***AGE***) Age in Months at Kindergarten Entry

Model: $GENERALKNOWLEDGE = \beta_0 + \beta_1 HEADSTARTHOURS + \beta_2 SES + \beta_3 ESL + \beta_4 AGE + \varepsilon$

SPSS DATA


*ECLSLATINASHSK.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
1 : GENERALKNOWLEDGE 17.497
Visible: 13 of 13 Variables
GENERALKNOWLEDGE HEADSTARTHOURS SES ESL AGE var var var var
1 17.50 0 -1.10 0 60
2 16.19 0 -1.08 0 64
3 20.63 17 -0.33 0 61
4 17.76 0 -0.49 0 67
5 18.42 3 0.67 0 68
Data View
Variable View
SPSS Processor is ready


ECLSLATINASHSK.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Name Type Width Decimals Label Values Missing Columns Align
1 GENERALKNO... Numeric 7 3 General Knowle... None None 10 Right
2 HEADSTARTH... Numeric 2 0 Number of Hea... None None 9 Right
3 SES Numeric 6 2 Socioeconomic... None None 8 Right
4 ESL Numeric 8 2 English as a 2n... {0.00, Englis... None 10 Right
5 AGE Numeric 8 2 Age in Months None None 10 Right
Data View
Variable View
SPSS Processor is ready

A Nested Hierarchy of Multiple Regression Models

We are going to fit and interpret a nested hierarchy of regression models.

$$GENERALKNOWLEDGE = \beta_0 + \beta_1 HEADSTARTHOURS + \varepsilon$$

$$GENERALKNOWLEDGE = \beta_0 + \beta_1 HEADSTARTHOURS + \beta_2 SES + \varepsilon$$

$$GENERALKNOWLEDGE = \beta_0 + \beta_1 HEADSTARTHOURS + \beta_2 SES + \beta_3 ESL + \varepsilon$$

$$GENERALKNOWLEDGE = \beta_0 + \beta_1 HEADSTARTHOURS + \beta_2 SES + \beta_3 ESL + \beta_4 AGE + \varepsilon$$

Pedagogical Strategy:

1. Interpret the fitted models of increasing complexity.
    1. Verbally.
    2. Graphically, with spreadsheet software.
2. Explain the meaning of statistical significance in multiple regression.
    1. Easy, due to the hard work of Unit 15!
3. Explain how we fit multiple regression models.
    1. Mathematically, we are minimizing the sum of squared residuals as always.
    2. Graphically, let's go 3-D!
4. Checking Assumptions
    1. Established Tools
    2. Residual vs. Predictor Plots

A One-Predictor Model

$$GENERALKNOWLEDGE = \beta_0 + \beta_1 HEADSTARTHOURS + \varepsilon$$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	20.145	.259		77.913	.000	19.638	20.653
	Number of Head Start Hours Per Week	-.096	.027	-.122	-3.499	.000	-.150	-.042

a. Dependent Variable: General Knowledge IRT Scaled Score

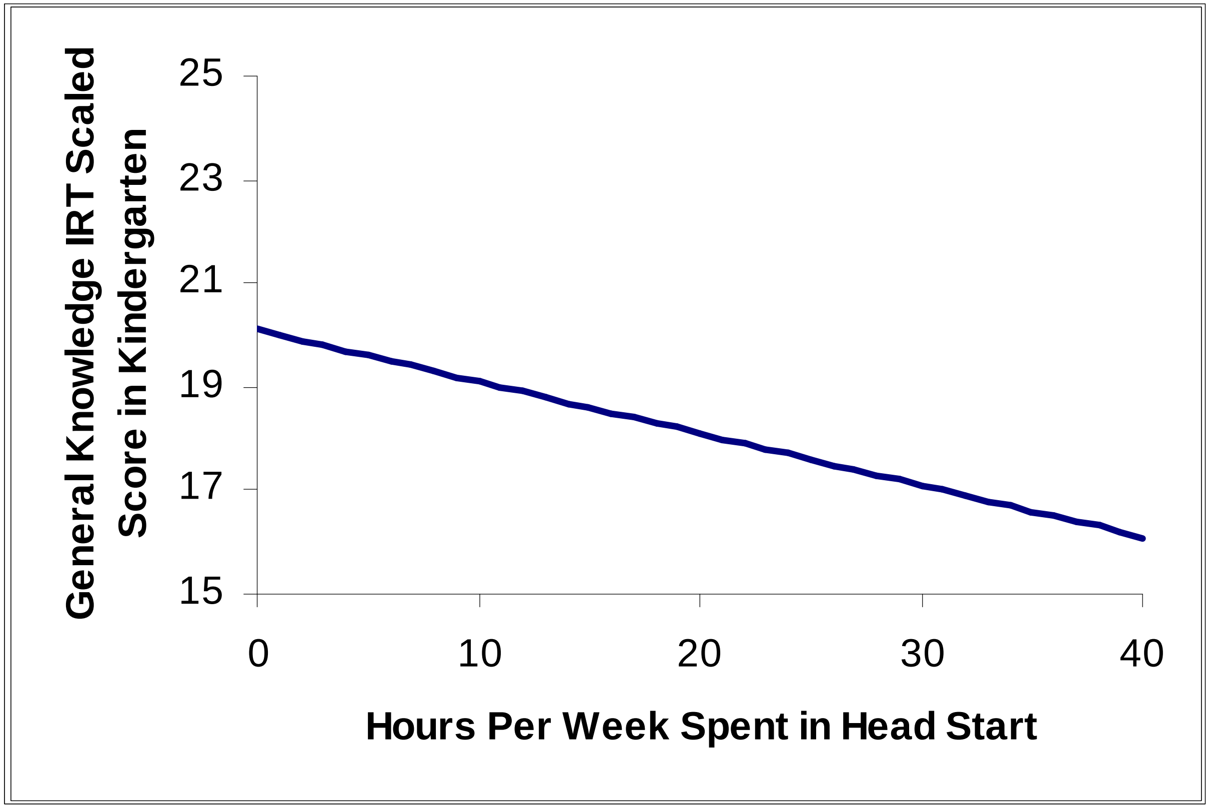
$$GENERAL\hat{K}NOWLEDGE = 20.1 + 0.1 HEADSTARTHOURS$$

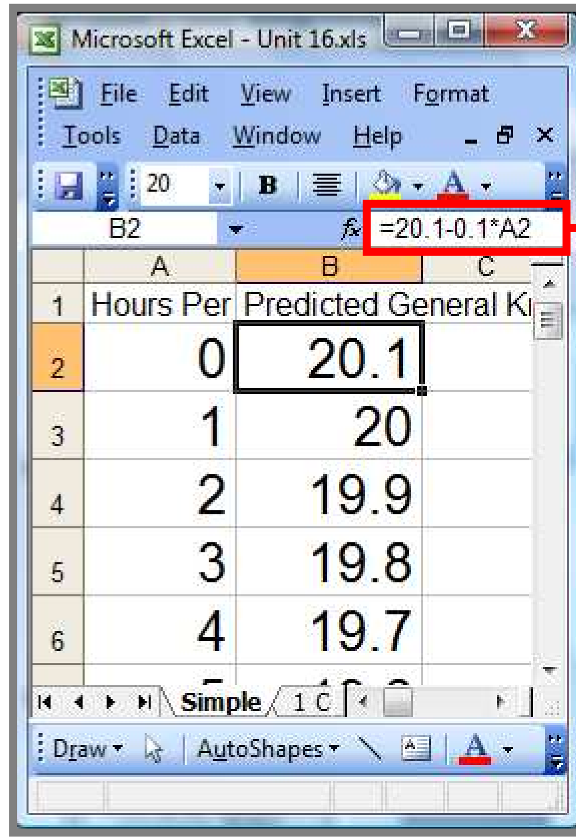
Hours of Head Start have a statistically significant negative correlation with scores on the kindergarten general knowledge test ($p < .001$). Using 95% confidence intervals, we can say that, in the population, a difference of ten Head Start hours is associated with an average difference of between 1.5 to 0.4 points on the general knowledge test, where the children with more head start hours tend to score lower.

Note that the constant (i.e., the y-intercept) is our prediction for GENERALKNOWLEDGE when HEADSTARTHOURS is zero. It is meaningful here, because zero falls within our observed range of HEADSTARTHOURS. For kindergarten Latinas who did not attend head start, we predict a score of about 20 points on the general knowledge test.

Graphing A One-Predictor Model

Figure 16.1. A plot of prototypical fitted values depicting the relationship between HEADSTARTHOURS and GENERALKNOWLEDGE for Latina kindergarteners (n = 816).





$$GENERAL\hat{K}NOWLEDGE = 20.1 - 0.1HEADSTARTHOURS$$

We learned how to graph lines more complex than this in Unit 13.

A Two-Predictor Model

$$GENERALKNOWLEDGE = \beta_0 + \beta_1 HEADSTARTHOURS + \beta_2 SES + \varepsilon$$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	20.603	.238		86.748	.000	20.137	21.070
	Number of Head Start Hours Per Week	-.014	.026	-.018	-.559	.576	-.065	.036
	Socioeconomic Status Composite Score	4.633	.352	.428	13.150	.000	3.941	5.324

a. Dependent Variable: General Knowledge IRT Scaled Score

$$GENERAL\hat{K}NOWLEDGE = 20.6 - .014 HEADSTARTHOURS + 4.6 SES$$

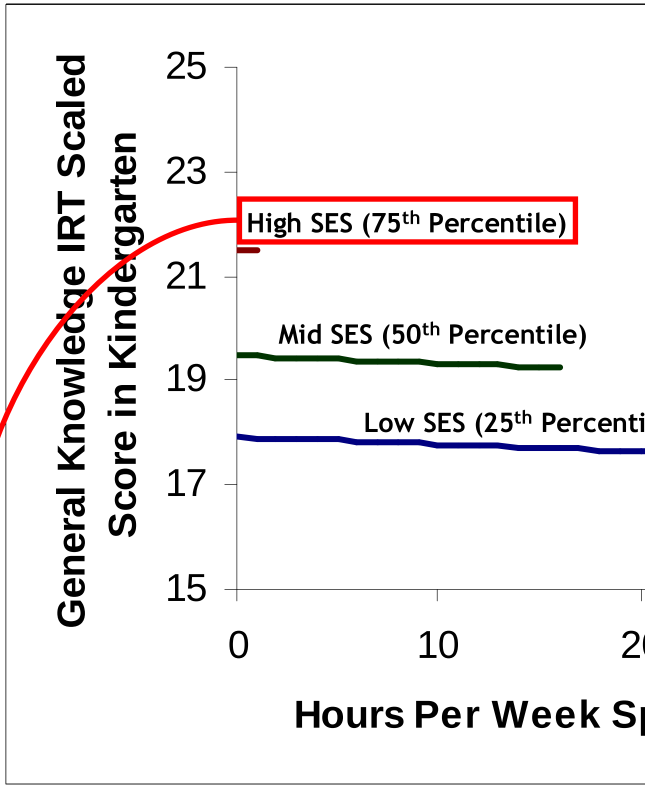
Controlling for SES, hours of Head Start have a statistically insignificant negative correlation with scores on the kindergarten general knowledge test (p = .576). In our sample, when we make comparisons among students of equal SES, we find that a difference of ten hours of Head Start is associated with an average difference of .14 points on the general knowledge test, where the children with more head start hours tend to score lower.

Controlling for hours of Head Start, SES has a statistically significant positive correlation with scores on the kindergarten general knowledge test (p < .001). In our sample, when we make comparisons among students of equal Head Start attendance, we find that a difference of one standard deviation of SES is associated with an average difference of 4.6 points on the general knowledge test, where the children of higher SES tend to score higher.

Note that the constant (i.e., the y-intercept) is our prediction for GENERALKNOWLEDGE when HEADSTARTHOURS is zero and SES is zero. It is meaningful here, because zero falls within our observed ranges of HEADSTARTHOURS and SES. For kindergarten Latinas who did not attend head start and had about an average SES (since SES is standardized), we predict a score of about 21 points on the general knowledge test.

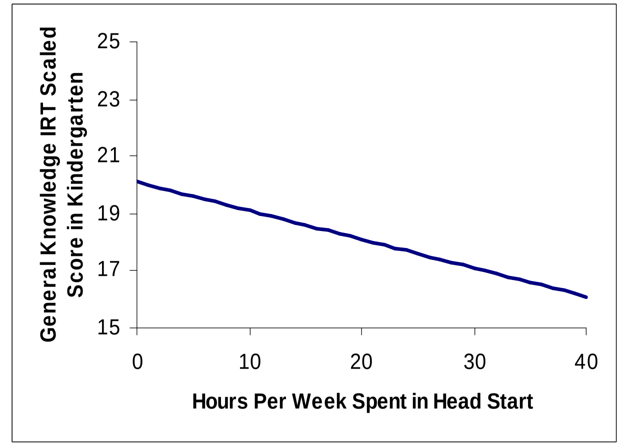
Graphing A Two-Predictor Model

Figure 16.2. A plot of prototypical fitted values depicting the relationship between HEADSTARTHOURS and GENERALKNOWLEDGE for Latina kindergarteners, controlling for SES (n = 816).



We can trim our lines so that we do not graph beyond the range of our data. When we do, we can see the vestiges of our simple linear regression.

Compare with earlier:



In order to graph three or more variables in two dimensions, we can (1) choose prototypical values for the extra variable(s) and/or (2) we can hold the extra variable(s) constant at their means (or medians or wherever).

I chose three prototypical values for SES:

25^{th} Percentile: -.5800

50^{th} Percentile: -.2400

75^{th} Percentile: .1975

This is hard to graph. ➔

$$GENERAL\hat{K}NOWLEDGE = 20.6 - .014HEADSTARTHOURS + 4.6SES$$

This is not. ↓

$$GENERAL\hat{K}NOWLEDGE \mid \text{High SES} = 20.6 - .014HEADSTARTHOURS + 4.6*(.1975)$$

Constructing Plots Of Prototypical Fitted Values (Part I of III)

1. Sketch the graph with paper and pencil before you even begin to play with the spreadsheet software.

   A. Your Y-axis is a no-brainer; it's your outcome.
      - General Knowledge Scores

   B. Your X-axis should be continuous, and, if your question predictor is continuous, then your X-axis should be your question predictor, because your X-axis variable will jump out most.
      - Head Start Hours

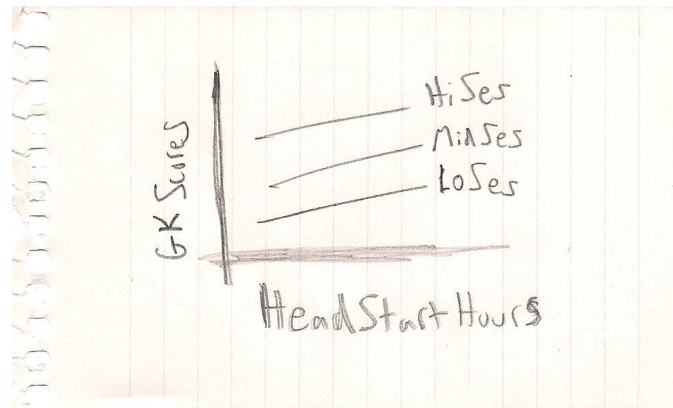
   C. Create separate trend lines by prototypical levels of a predictor of your choice. If your question predictor is categorical, then choose your question predictor and its categories. Otherwise, use the most interesting control.
      - SES
        - Let "High SES" = 0.1975
        - Let "Medium SES" = -0.2400
        - Let "Low SES" = -0.5800

   D. For the predictors not on your X-axis and without their own trend lines, set them constant (probably at their means).

2. The first column of your spreadsheet will be values that define your X-axis, running from the min to the max at generally equal intervals. Do not use raw data here, or anywhere in this process.
   - Head Start Hours Per Week Range from 0 to 40

3. For each line in your sketch, you will have one additional column. Each of the additional columns will be filled with predicted-outcome values from your fitted model.
   - Predicted-Outcome|High SES = 20.6 - 0.014*(A2) + 4.6*(0.1975)
   - Predicted-Outcome|Med SES = 20.6 - 0.014*(A2) + 4.6*(-0.2400)
   - Predicted-Outcome|Low SES = 20.6 - 0.014*(A2) + 4.6*(-0.5800)

Notice that in my hand sketch, I get the slopes wrong. That's fine. It's just a sketch…



$$\widehat{GENERAL\ KNOWLEDGE} = 20.6 - .014\,HEADSTARTHOURS + 4.6\,SES$$

See that for our "High SES" trend line, we lock in our chosen prototypical value for High SES. In general, we lock in *ALL* the variable values except the values for the X-axis variable.

Constructing Plots of Prototypical Fitted Values (Part II of III)

Microsoft Excel - Unit 16.xls

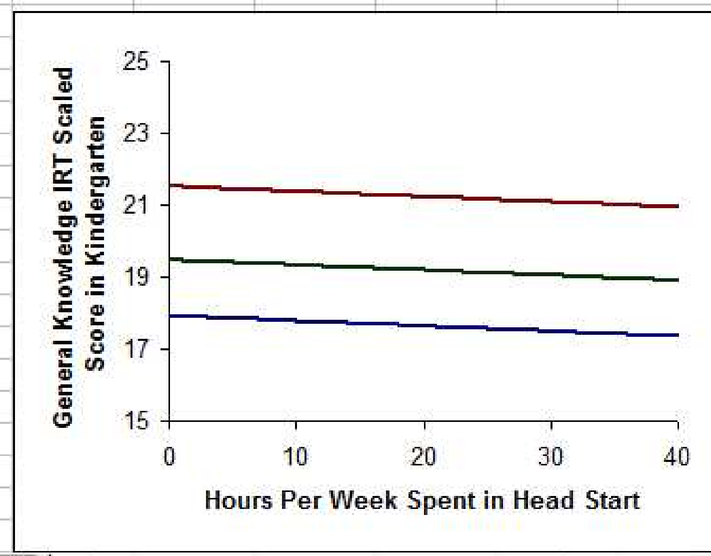
D2 =20.60346-0.01434*(A2)+4.632518*(0.1975)

	A	B	C	D	E	F	G	H	I	J
1	Hours Per	LoSES Pre	MidSES P	HiSES Pred Know					Fitted Model (Right Side):	
2	0	17.9166	19.47313	21.51838					20.6-.014*HEADSTARTHOURS+4.6*SES	
3	1	17.90226	19.45879	21.50404					Prototypical Values:	
4	2	17.88792	19.44445	21.4897					LoSES	-0.58
5	3	17.87358	19.43011	21.47536					MidSES	-0.24
6	4	17.85924	19.41577	21.46102					HiSES	0.1975
7	5	17.8449	19.40143	21.44668						
8	6	17.83056	19.38709	21.43234						
9	7	17.81622	19.37275	21.418						
10	8	17.80188	19.35841	21.40366						
11	9	17.78754	19.34407	21.38932						
12	10	17.7732	19.32973	21.37498						
13	11	17.75886	19.31539	21.36064						
14	12	17.74452	19.30105	21.3463						
15	13	17.73018	19.28671	21.33196						
16	14	17.71584	19.27237	21.31762						
17	15	17.7015	19.25803	21.30328						
18	16	17.68716	19.24369	21.28894						
19	17	17.67282	19.22935	21.2746						
20	18	17.65848	19.21501	21.26026						
21	19	17.64414	19.20067	21.24592						
22	20	17.6298	19.18633	21.23158						
23	21	17.61546	19.17199	21.21724						
24	22	17.60112	19.15765	21.2029						

Sheets: Simple, 1 Control, 2 Controls, 3 Controls, Interactions (4), Interactions (3)

Include notes in your spread-sheet file to help you keep track of things.

Graph with a scatterplot, not with a line graph (or any other sort of graph).

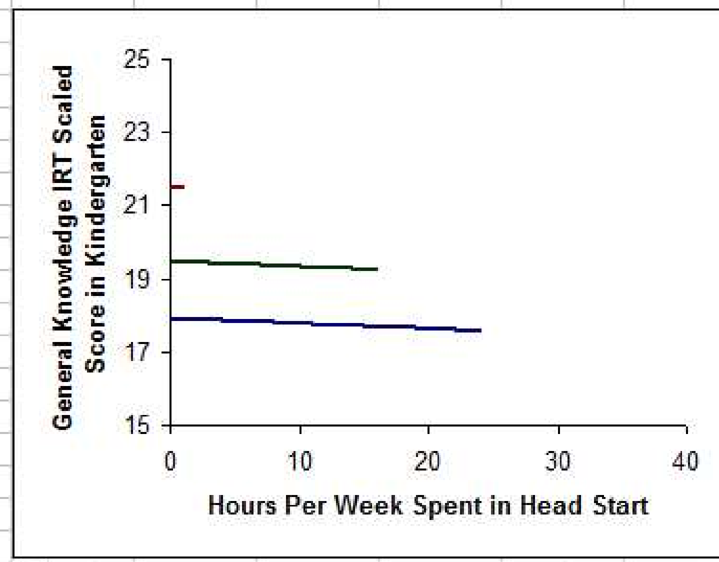


Constructing Plots of Prototypical Fitted Values (Part III of III)

Microsoft Excel - unit16a.xls

D2 =20.60346-0.01434*(A2)+4.632518*(0.1975)

	A	B	C	D	E	F	G	H	I	J
1	Hours Per	LoSES Pre	MidSES P	HiSES Pred Know					Fitted Model (Right Side):	
2	0	17.9166	19.47313	21.51838					20.6-.014*HEADSTARTHOURS+4.6*SES	
3	1	17.90226	19.45879	21.50404					Prototypical Values:	
4	2	17.88792	19.44445						LoSES	-0.58
5	3	17.87358	19.43011						MidSES	-0.24
6	4	17.85924	19.41577						HiSES	0.1975
7	5	17.8449	19.40143							
8	6	17.83056	19.38709							
9	7	17.81622	19.37275							
10	8	17.80188	19.35841							
11	9	17.78754	19.34407							
12	10	17.7732	19.32973							
13	11	17.75886	19.31539							
14	12	17.74452	19.30105							
15	13	17.73018	19.28671							
16	14	17.71584	19.27237							
17	15	17.7015	19.25803							
18	16	17.68716	19.24369							
19	17	17.67282	19.22935							
20	18	17.65848								
21	19	17.64414								
22	20	17.6298								
23	21	17.61546								
24	22	17.60112								



You can trim back the trend lines to stay within the range of your data by deleting the appropriate cells. There is no exact science to determining the range of your data, because the prototypical values (e.g., for High SES) themselves represent a range of data. Try it. You'll see what I mean.

Simple | 1 Control | 2 Controls | 3 Controls | Interactions (4) | Interactions (3)

A Three-Predictor Model

$$GENERALKNOWLEDGE = \beta_0 + \beta_1 HEADSTARTHOURS + \beta_2 SES + \beta_3 ESL + \varepsilon$$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	21.574	.257		83.885	.000	21.070	22.079
	Number of Head Start Hours Per Week	.008	.025	.010	.317	.752	-.041	.057
	Socioeconomic Status Composite Score	4.149	.344	.384	12.069	.000	3.475	4.824
	English as a 2nd Language	-3.908	.476	-.256	-8.208	.000	-4.843	-2.974

a. Dependent Variable: General Knowledge IRT Scaled Score

$$\widehat{GENERALKNOWLEDGE} = 21.6 + .008 HEADSTARTHOURS + 4.1 SES - 3.9 ESL$$

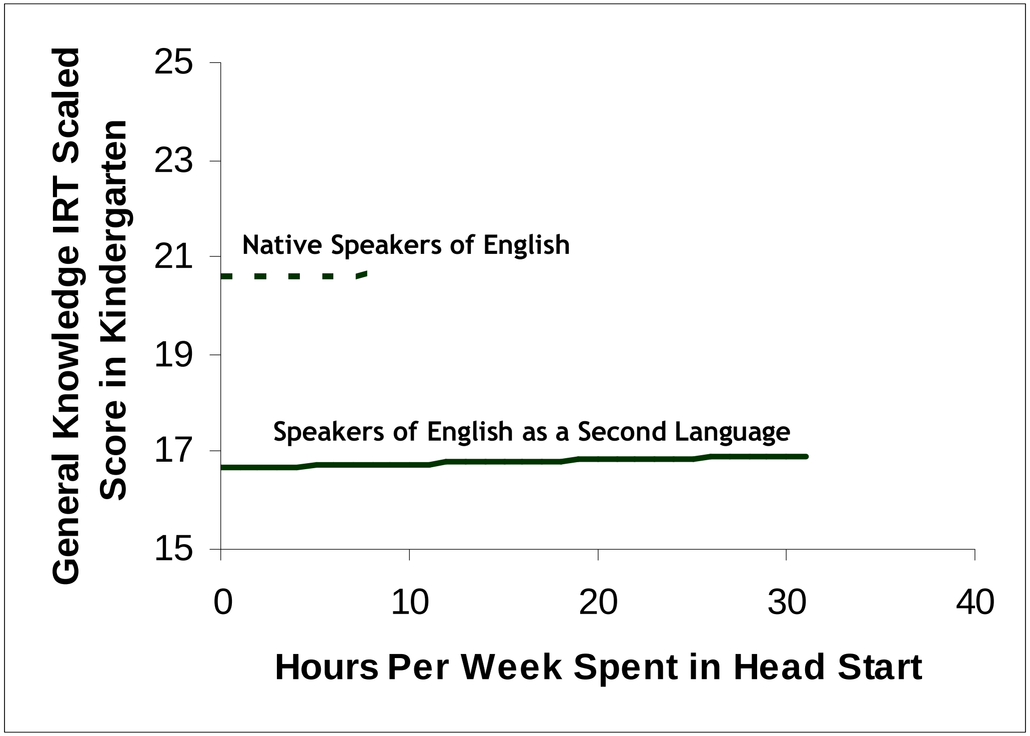
Controlling for SES and ESL, hours of Head Start have a statistically insignificant positive correlation with scores on the kindergarten general knowledge test (p = .752). In our sample, when we make comparisons among students of equal SES and fluency, we find that a difference of ten hours of Head Start is associated with an average difference of .08 points on the general knowledge test, where the children with more head start hours tend to score higher.

Controlling for hours of Head Start and ESL, SES has a statistically significant positive correlation with scores on the kindergarten general knowledge test (p < .001). Controlling for head start hours and SES, ESL has a statistically significant negative relationship with scores on the general knowledge test such that students for whom English is a second language, on average, score lower than native speakers of English (p < .001).

Note that the constant (i.e., the y-intercept) is our prediction for GENERALKNOWLEDGE when HEADSTARTHOURS is zero, SES is zero, and ESL is zero. It is meaningful here, because zero falls within our observed ranges of HEADSTARTHOURS and SES. For kindergarten Latinas who speak English as a first language, did not attend head start and had about an average SES (since SES is standardized), we predict a score of about 21 points on the general knowledge test.

Graphing A Three-Predictor Model

Figure 16.3. A plot of prototypical fitted values depicting the relationship between HEADSTARTHOURS and GENERALKNOWLEDGE for Latina kindergarteners, controlling for SES (set at the median) and ESL (n = 816).



In order to graph three or more variables in two dimensions, we can (1) choose prototypical values for the extra variable(s) and/or (2) we can hold the extra variable(s) constant at their means (or medians).

I chose to hold SES constant at its median:
50^{th} Percentile of SES = -.24
I "chose" prototypical values for ESL:
ESL = 0 and ESL = 1

Why hold a control variable constant? If we plot prototypical values for every variable, the plot may become too cluttered with trend lines, because the total number of trend lines will equal the number of prototypical values for control1 times the number of prototypical values for control2.

Why NOT hold a control variable constant? Because it will disappear from sight! (This is sad, but often unavoidable.) Where's SES?

$$\widehat{GENERALKNOWLEDGE} = 21.6 + .008HEADSTARTHOURS + 4.1SES - 3.9ESL$$

$$[\widehat{GENERALKNOWLEDGE} \mid SES = -.24, ESL = 1] = 21.6 + .008HEADSTARTHOURS + 4.1*(-.24) - 3.9*(1)$$

Which trend line is associated with this equation?

A Four-Predictor Model

$$GENERALKNOWLEDGE = \beta_0 + \beta_1 HEADSTARTHOURS + \beta_2 SES + \beta_3 ESL + \beta_4 AGE + \varepsilon$$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-3.207	3.247		-.988	.324	-9.580	3.167
	Number of Head Start Hours Per Week	.002	.024	.003	.084	.933	-.045	.049
	Socioeconomic Status Composite Score	4.069	.332	.376	12.240	.000	3.416	4.721
	English as a 2nd Language	-3.784	.460	-.248	-8.217	.000	-4.687	-2.880
	Age in Months	.380	.050	.225	7.655	.000	.283	.478

a. Dependent Variable: General Knowledge IRT Scaled Score

$$GENERAL\hat{K}NOWLEDGE = -3.2 + .002HEADSTARTHOURS + 4.1SES - 3.9ESL + .4AGE$$

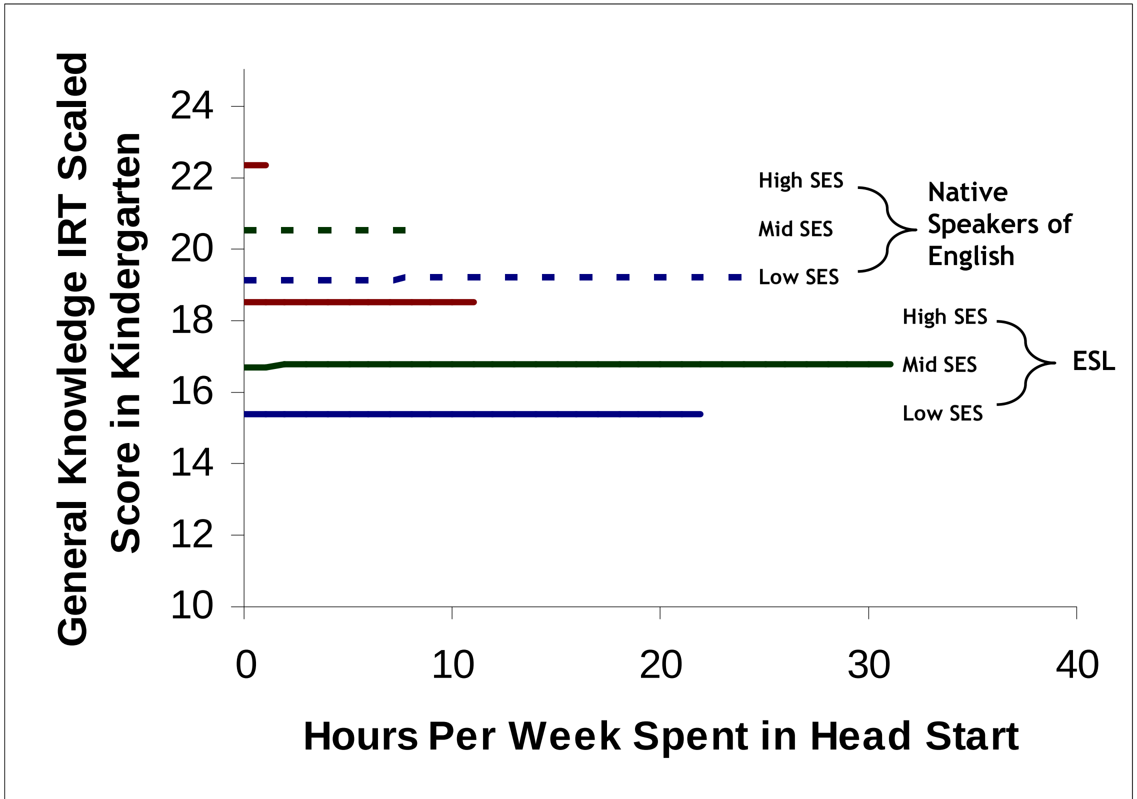
Controlling for SES, ESL, and age, hours of Head Start have a statistically insignificant positive correlation with scores on the kindergarten general knowledge test (p = .933). In our sample, when we compare kindergarten Latinas who spent 40 hours per week in Head Start versus kindergarten Latinas who did not attend Head Start, we observe a difference of about .08 points on the general knowledge test favoring Head Starters.

Controlling for Head Start hours, ESL, and age, SES has a statistically significant positive correlation with scores on the kindergarten general knowledge test (p < .001). Controlling for Head Start hours, SES, and age, ESL has a statistically significant negative correlation with scores on the kindergarten general knowledge test (p < .001). Controlling for Head Start hours, SES, and ESL, age has a statistically significant positive correlation with scores on the kindergarten general knowledge test (p < .001).

Note that the constant (i.e., the y-intercept) is our prediction for GENERALKNOWLEDGE when HEADSTARTHOURS is zero, SES is zero, ESL is zero and AGE is zero. Thus, it is meaningless in and of itself, because AGE is never zero.

Graphing A Four-Predictor Model

Figure 16.4. A plot of prototypical fitted values depicting the relationship between HEADSTARTHOURS and GENERALKNOWLEDGE for kindergarten Latinas, controlling for ESL and SES, and holding AGE constant at its median of 64 (n = 816).



Note: I think this is too cluttered, but sometime six lines works.

I chose to hold *AGE* constant at its median:
50th Percentile of *AGE* = 64
I "chose" prototypical values for *ESL*:
ESL = 0 and *ESL* = 1
I chose three prototypical values for *SES*:
25th Percentile: -.58
50th Percentile: -.24
75th Percentile: .20

Notice that all our lines have been parallel. That is because we have *assumed* them to be parallel—the main effects assumption. In Unit 17, we will relax the main effects assumption when we learn to model statistical interactions.

$$GENERAL\hat{K}NOWLEDGE = -3.2 + .002HEADSTARTHOURS + 4.1SES - 3.8ESL + 0.4AGE$$

$$[\,GENERAL\hat{K}NOWLEDGE \mid \text{Low SES}\,(SES = -.58), \text{ESL Student}\,(ESL = 1), AGE = 64] =$$
$$-3.2 + .002HEADSTARTHOURS + 4.1*(-.58) - 3.8*(1) + 0.4*(64)$$

Which trend line is associated with this equation?

Dig the Post Hole

Unit 16 Post Hole:

Interpret a fitted multiple regression model.

- Interpret each slope coefficient as though it were from a simple linear regression, but note that you are controlling for all the other predictors in the model.
  - ✓ Avoid unwarranted causal and developmental language! (*If* you were controlling for every variable under the sun, you could draw causal conclusions. But, you're not.)
- If the y-intercept coefficient is interesting, interpret that as well. The y-intercept, as always, is our prediction for an observation with a zero for every predictor in the model.

What Does Statistical Significance Mean In Multiple Regression?

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	-3.207	3.247		-.988	.324	-9.580	3.167
	Number of Head Start Hours Per Week	.002	.024	.003	.084	.933	-.045	.049
	Socioeconomic Status Composite Score	4.069	.332	.376	12.240	.000	3.416	4.721
	English as a 2nd Language	-3.784	.460	-.248	-8.217	.000	-4.687	-2.880
	Age in Months	.380	.050	.225	7.655	.000	.283	.478

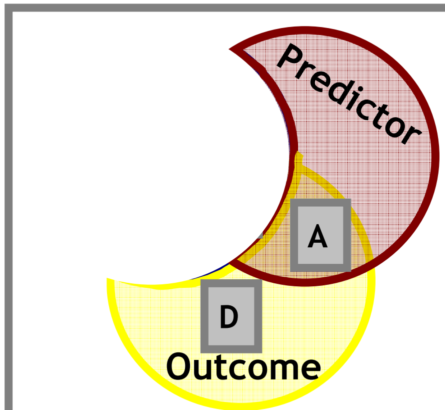
a. Dependent Variable: General Knowledge IRT Scaled Score

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-3.207	3.247		-.988	.324
	Socioeconomic Status Composite Score	4.069	.332	.376	12.240	.000
	Number of Head Start Hours Per Week	.002	.024	.003	.084	.933
	Age in Months	.380	.050	.225	7.655	.000
	English as a 2nd Language	-3.784	.460	-.248	-8.217	.000

a. Dependent Variable: General Knowledge IRT Scaled Score

*** Note that "Controls" stands for *ALL* other predictors in the model!**

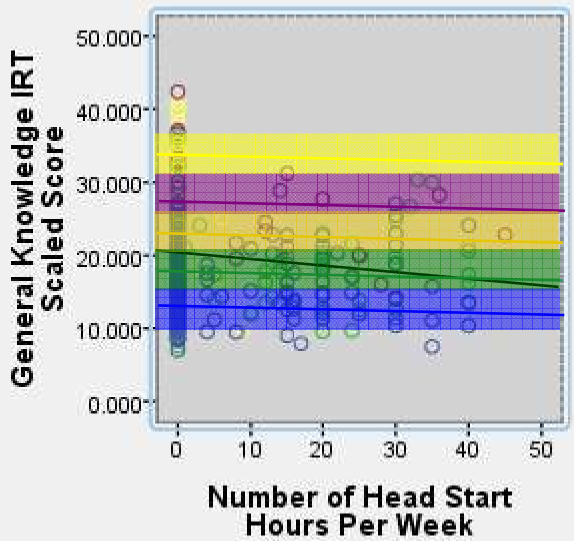


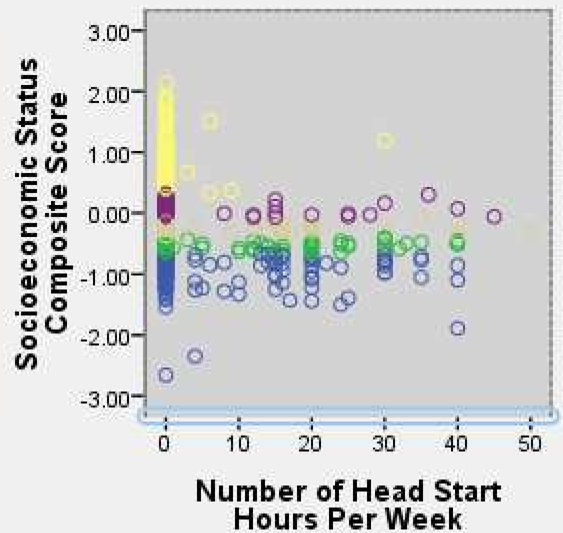
Order does not matter! SPSS and R do not care about the order in which you input your variables. SPSS and R do not care about the order in which you fit your models.

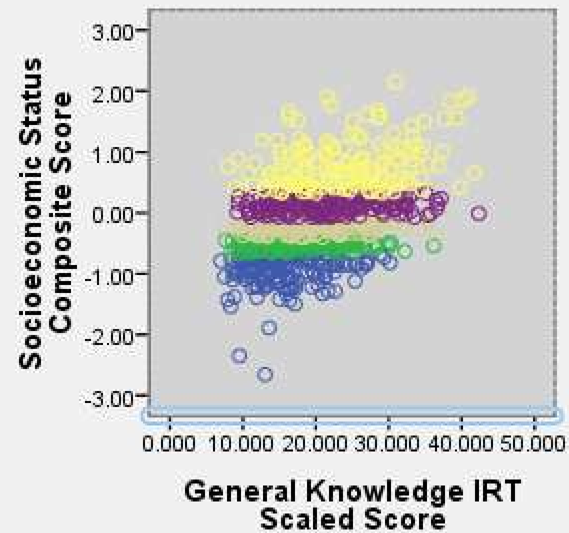
A parameter estimate (aka, regression coefficient or slope estimate) is statistically significant when the associated predictor has a statistically significant partial correlation with the outcome controlling for every other predictor in the model.

E.g., HEADSTART has a stat insig partial correlation with GENERALKNOWLEDGE controlling for SES, ESL, and AGE. SES has a stat sig partial correlation with GENERALKNOWLEDGE controlling for HEADSTART, ESL, and AGE.

Fitting the Model





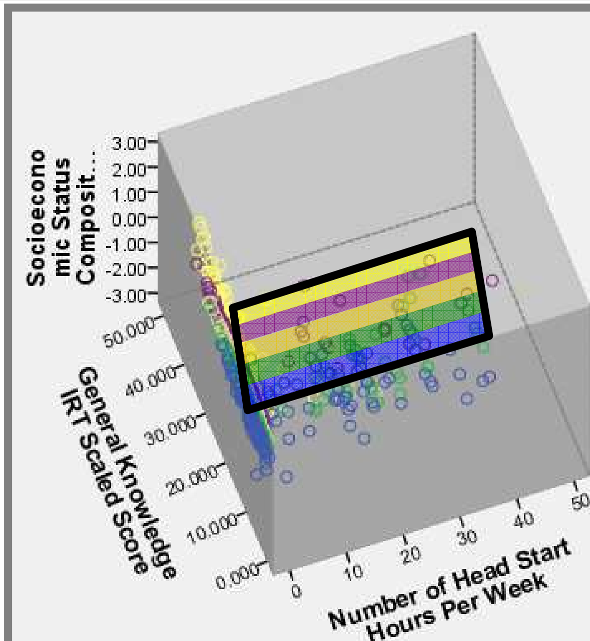


We can color code SES such that the lowest 20% are blue, the next 20% are green, the next 20% are tan, the next 20% are violet, and the highest 20% are yellow.

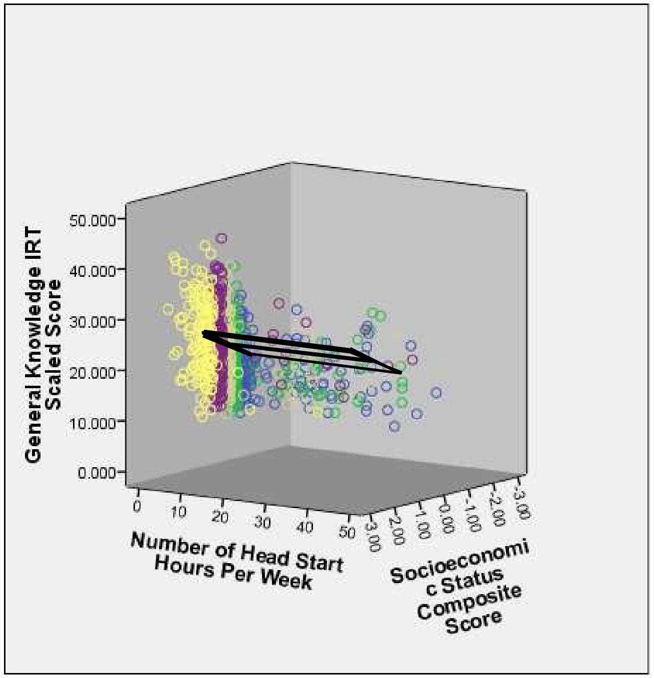
Instead of dropping a line across all the observations…

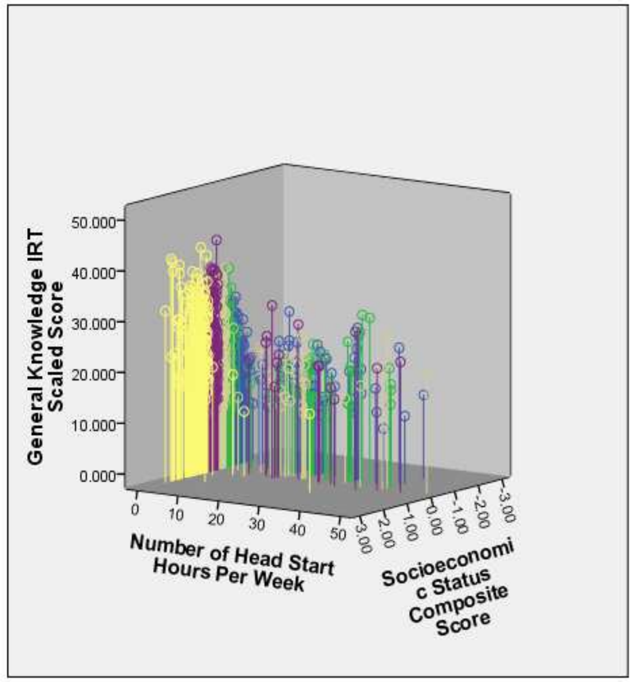
We can drop a line for each layer of SES. (This is an over simple, but often heuristically useful, way to think about multiple regression.)

Even better, we can go three dimensional and drop a plane. Now, we are talking multiple regression!

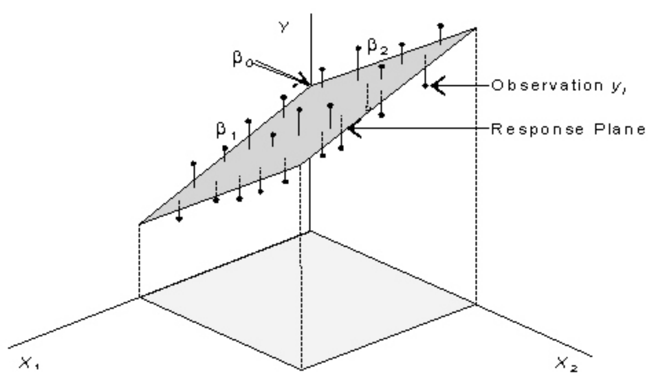


Fitting a Plane



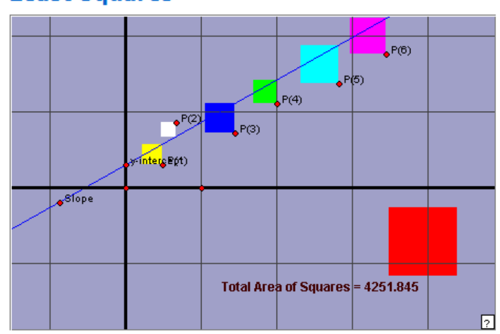


http://www.sjsu.edu/faculty/gerstman/EpiInfo/cont-mult1.jpg



http://www.jerrydallal.com/LHSP/regpix.htm

Least Squares



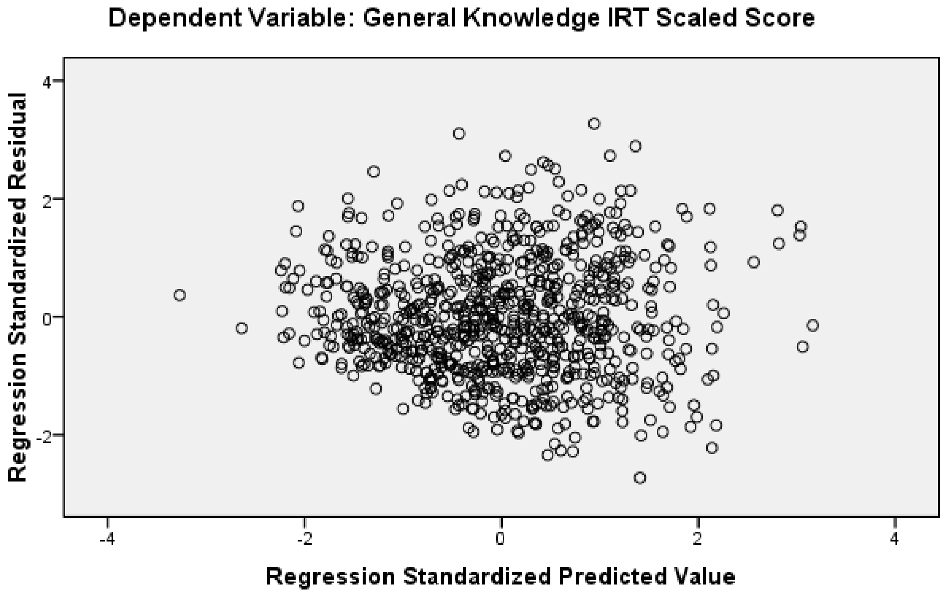
http://www.dynamicgeometry.com/JavaSketchpad/Gallery/Other_Explorations_and_Amusements/Least_Squares.html

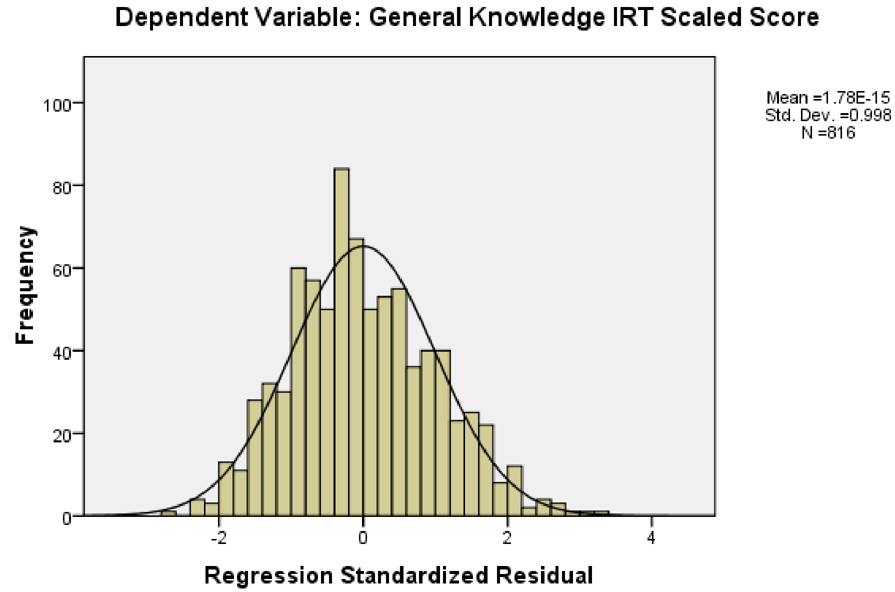
In simple linear regression, with one predictor, we fit a one dimensional object—a line—in a two dimensional space. In multiple regression with two predictors we fit a two dimensional object—a plane—in a three dimensional space. In both, we use ordinary least squares (OLS) regression, in which we fit our object with an eye toward the least sum of squared residuals. When we get beyond two or three predictors, however, our eyes begin to fail us, but the math does not.

For residuals, no matter how many dimensions, the math never stops thinking vertically, whatever vertical may mean in, say, seven dimensional space. A residuals is, and always will be, the difference between the observed value and the predicted value, and the squared residual will always be the difference times itself.

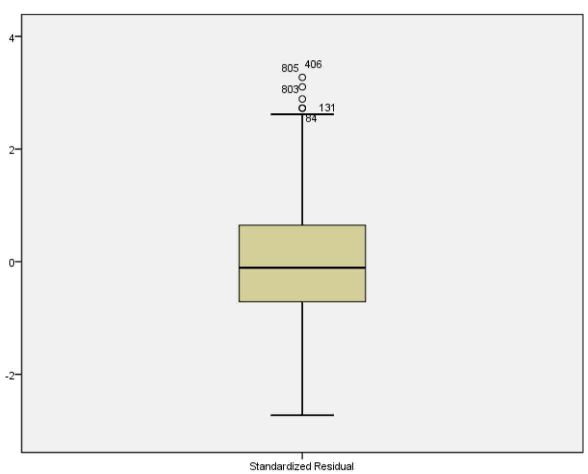
See the Math Appendix for the OLS formula for model fitting (i.e., estimating parameters).

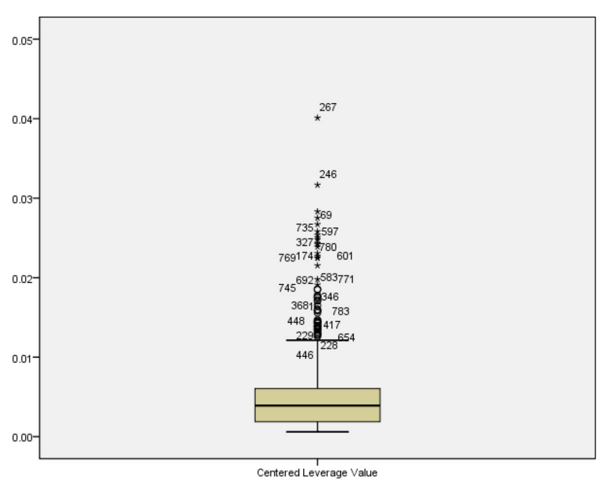
Checking Assumptions: Using The Tools From Unit 12

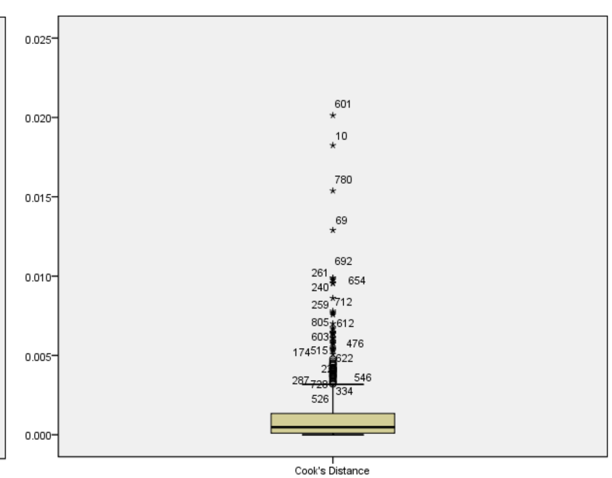




Note that I am using box-and-whisker plots to begin my examination of residuals, leverage and influence. I could use histograms or scatterplots (vs. ID), but I felt like starting with box-and-whisker plots. So there.







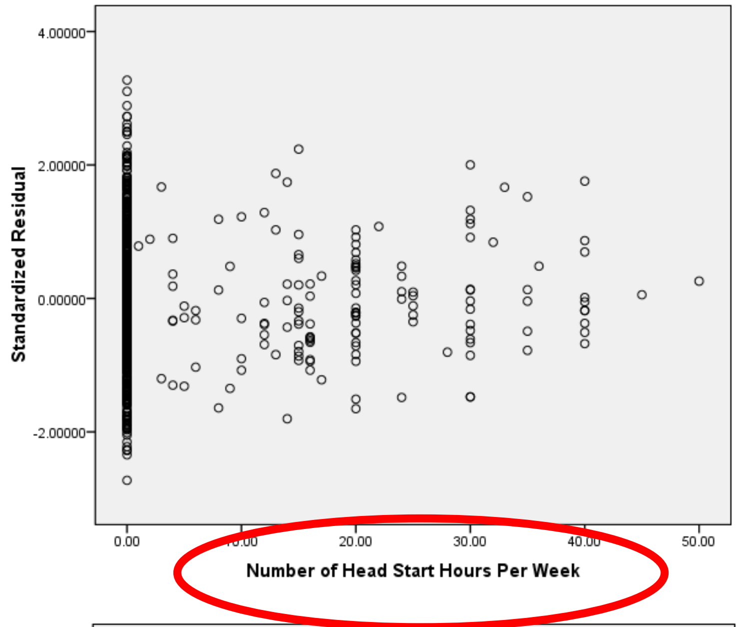
Residual vs. Predictor Plots

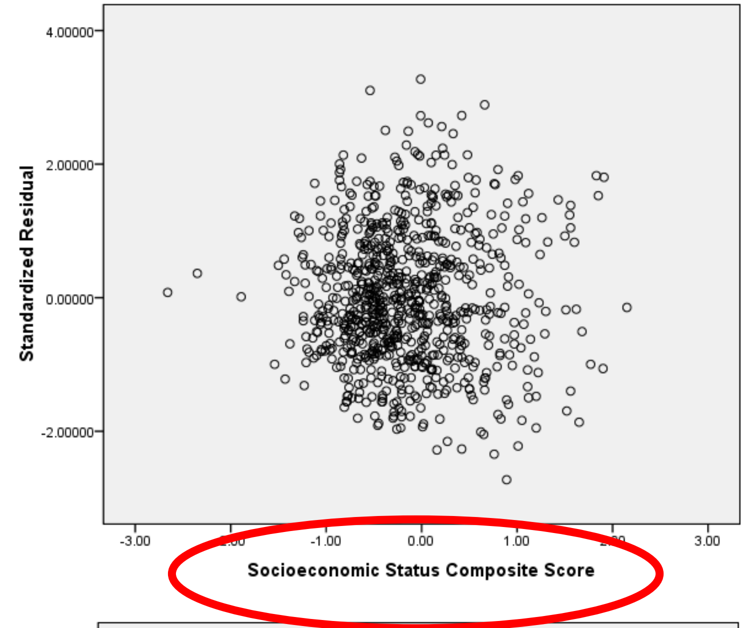
A residual vs. predictor plot is similar to a residual vs. predicted plot. The former uses predictor values on the X-axis; the latter used predicted values on the X-axis. In effect, the residual vs. predictor plots dissect the residual vs. predicted plot.

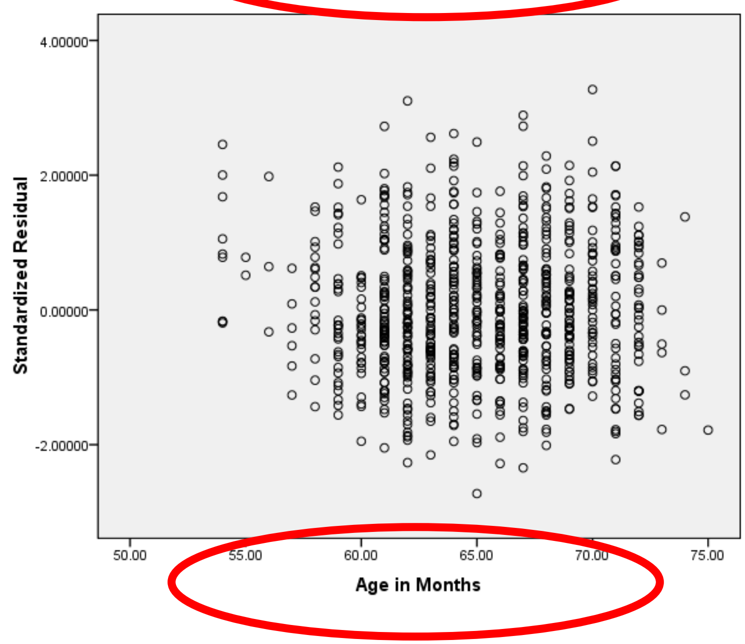
Read these plots just like you would a residual vs. predicted plot (i.e., residual vs. fitted (RVF) plot). Look HI-N-LO for assumption violations.

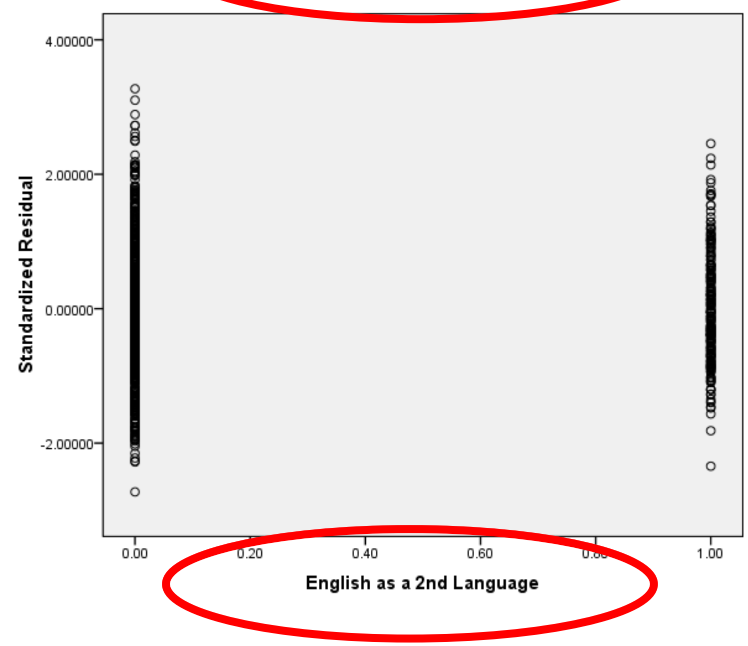
If you turn up a problem in the RVF plot, then these plots may help you find the source.

It's also possible that a problem is masked in the RVF plot that only turns up in the residual vs. predictor plots.









Answering Our Road Map Question

Unit 16: Is there a relationship between reading and race, controlling for SES, ESL and homework?

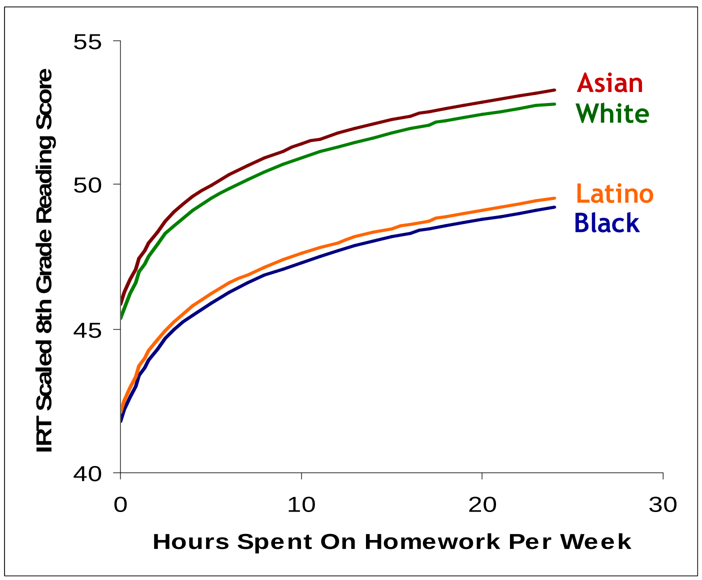
Coefficients[a]

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

a. Dependent Variable: READING

See Unit 9 for why I did not include ***WHITE*** in my model as a predictor. In short, it is my reference category, where ***WHITE***, ***ASIAN***, ***BLACK*** and ***LATINO*** constitute one conceptual, predictor, ***RACE***.

Figure 16.5. A plot of prototypical fitted values depicting the relationship between RACE, HOMEWORK and READING holding ESL and FREELUNCH constant at the mode, i.e., not ESL, not eligible for free lunch (n = 7,800).



Controlling for *HOMEWORK*, *ESL* and *FREELUNCH*, the difference in *READING* performance between Asian students and White students is not statistically significant ($p = 0.298$); however, Black students and Latino students score on average 3 to 4 points lower than White students ($p < .001$).

Unit 16 Appendix: Key Concepts

In order to graph three or more variables in two dimensions, we can (1) choose prototypical values for the extra variable(s) and/or (2) we can hold the extra variable(s) constant at their means (or medians or wherever).

We can trim our lines so that we do not graph beyond the range of our data. When we do, we can see the vestiges of our simple linear regression.

Notice that all our lines have been parallel. That is because we have *assumed* them to be parallel—the main effects assumption. In Unit 17, we will relax the main effects assumption when we learn to model statistical interactions.

Order does not matter! SPSS and R do not care about the order in which you input your variables. SPSS and R do not care about the order in which you fit your models.

For residuals, no matter how many dimensions, the math never stops thinking vertically, whatever vertical may mean in, say, eight dimensional space. A residuals is, and always will be, the difference between the observed value and the predicted value, and the squared residual will always be the difference times itself.

Unit 16 Appendix: Key Interpretations

Controlling for SES, hours of Head Start have a statistically insignificant negative correlation with scores on the kindergarten general knowledge test ($p = .576$). In our sample, when we make comparisons among students of equal SES, we find that a difference of ten hours of Head Start is associated with an average difference of .14 points on the general knowledge test, where the children with more head start hours tend to score lower.

Controlling for hours of Head Start, SES has a statistically significant positive correlation with scores on the kindergarten general knowledge test ($p < .001$). In our sample, when we make comparisons among students of equal Head Start attendance, we find that a difference of one standard deviation of SES is associated with an average difference of 4.6 points on the general knowledge test, where the children of higher SES tend to score higher.

Controlling for SES and ESL, hours of Head Start have a statistically insignificant positive correlation with scores on the kindergarten general knowledge test ($p = .752$). In our sample, when we make comparisons among students of equal SES and fluency, we find that a difference of ten hours of Head Start is associated with an average difference of .08 points on the general knowledge test, where the children with more head start hours tend to score higher.

Controlling for hours of Head Start and ESL, SES has a statistically significant positive correlation with scores on the kindergarten general knowledge test ($p < .001$). Controlling for head start hours and SES, ESL has a statistically significant negative relationship with scores on the general knowledge test such that students for whom English is a second language, on average, score lower than native speakers of English ($p < .001$).

Controlling for Head Start hours, ESL, and age, SES has a statistically significant positive correlation with scores on the kindergarten general knowledge test ($p < .001$). Controlling for Head Start hours, SES, and age, ESL has a statistically significant negative correlation with scores on the kindergarten general knowledge test ($p < .001$). Controlling for Head Start hours, SES, and ESL, age has a statistically significant positive correlation with scores on the kindergarten general knowledge test ($p < .001$).

Controlling for SES, ESL, and age, hours of Head Start have a statistically insignificant positive correlation with scores on the kindergarten general knowledge test ($p = .933$). In our sample, when we compare kindergarten Latinas who spent 40 hours per week in Head Start versus kindergarten Latinas who did not attend Head Start, we observe a difference of about .08 points on the general knowledge test favoring Head Starters.

Controlling for *HOMEWORK*, *ESL* and *FREELUNCH*, the difference in *READING* performance between Asian students and White students is not statistically significant ($p = 0.298$); however, Black students and Latino students score on average 3 to 4 points lower than White students ($p < .001$).

Unit 16 Appendix: Key Terminology

A parameter estimate (aka, regression coefficient or slope estimate) is statistically significant when the associated predictor has a statistically significant partial correlation with the outcome controlling for every other predictor in the model.

A residual vs. predict<u>or</u> plot is similar to a residual vs. predict<u>ed</u> plot. The former uses predictor values on the X-axis; the latter used predicted values on the X-axis. In effect, the residual vs. predictor plots dissect the residual vs. predicted plot.

Unit 16 Appendix: Math (Very Optional)

If you want to fit by hand a ***simple*** linear model using ordinary least squares (OLS) regression, you'll need multivariable calculus. Calculus is very good at finding minimums and maximums. When we do OLS regression, we want to find a y-intercept (β_0) and slope (β_1) that minimizes the sum of squared errors (i.e., sum of squared residuals). A statistical error (i.e., residual) is the difference between our observation and prediction. Say that we have three observations:

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

We propose a model:

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

Thus:

$$READING - \beta_0 - \beta_1 FREELUNCH = \varepsilon$$

Thus:

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

Each subject has a squared error:

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

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

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

The sum of squared errors (SSE) is a function of two variables, β_0 and β_1:

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

Unit 16 Appendix: Math (Very Optional)

If you want to fit by hand a ***multiple*** linear model using ordinary least squares (OLS) regression, you'll follow the same logic as with the simple linear model.

NAME	READING	FREELUNCH	HOMEWORK
Sean	90	0	0
Betsy	100	0	20
Waverly	80	1	10

We propose a model:

$$READING = \beta_0 + \beta_1 FREELUNCH + \beta_2 HOMEWORK + \varepsilon$$

Thus:

$$READING - \beta_0 - \beta_1 FREELUNCH - \beta_2 HOMEWORK = \varepsilon$$

Thus:

$$(READING - \beta_0 - \beta_1 FREELUNCH - \beta_2 HOMEWORK)^2 = (\varepsilon)^2$$

Each subject has a squared error:

$$(90 - \beta_0 - \beta_1 0 - \beta_2 0)^2 = (\varepsilon_{Sean})^2$$

$$(100 - \beta_0 - \beta_1 0 - \beta_2 20)^2 = (\varepsilon_{Betsy})^2$$

$$(80 - \beta_0 - \beta_1 1 - \beta_2 10)^2 = (\varepsilon_{Wavy})^2$$

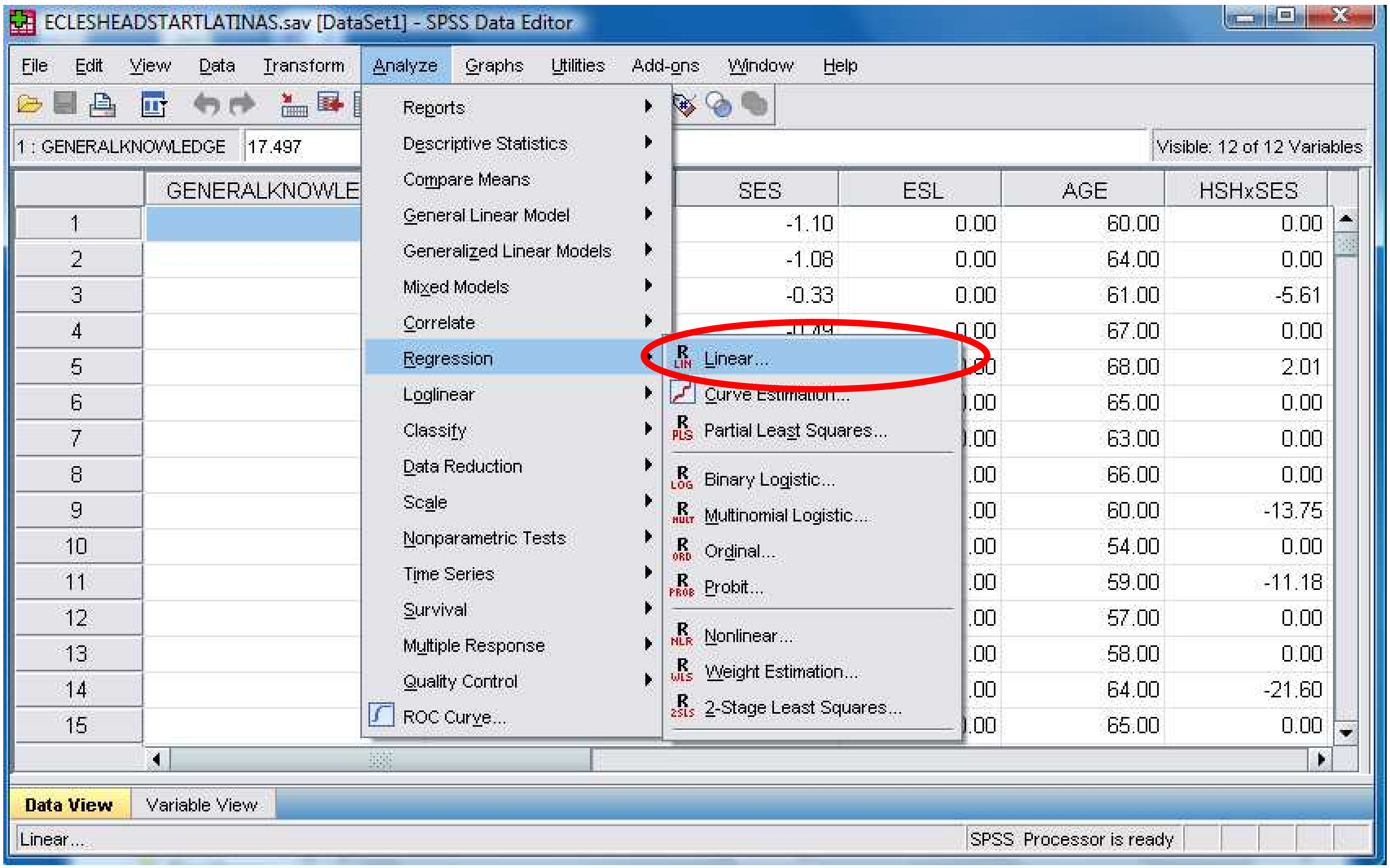
The sum of squared errors (SSE) is a function of THREE variables, β_0, β_1 and β_2:

$$SSE(\beta_0, \beta_1, \beta_2) = (90 - \beta_0 - \beta_1 0 - \beta_2 0)^2 + (100 - \beta_0 - \beta_1 0 - \beta_2 20)^2 + (80 - \beta_0 - \beta_1 1 - \beta_2 10)^2$$

Unit 16 Appendix: SPSS Syntax

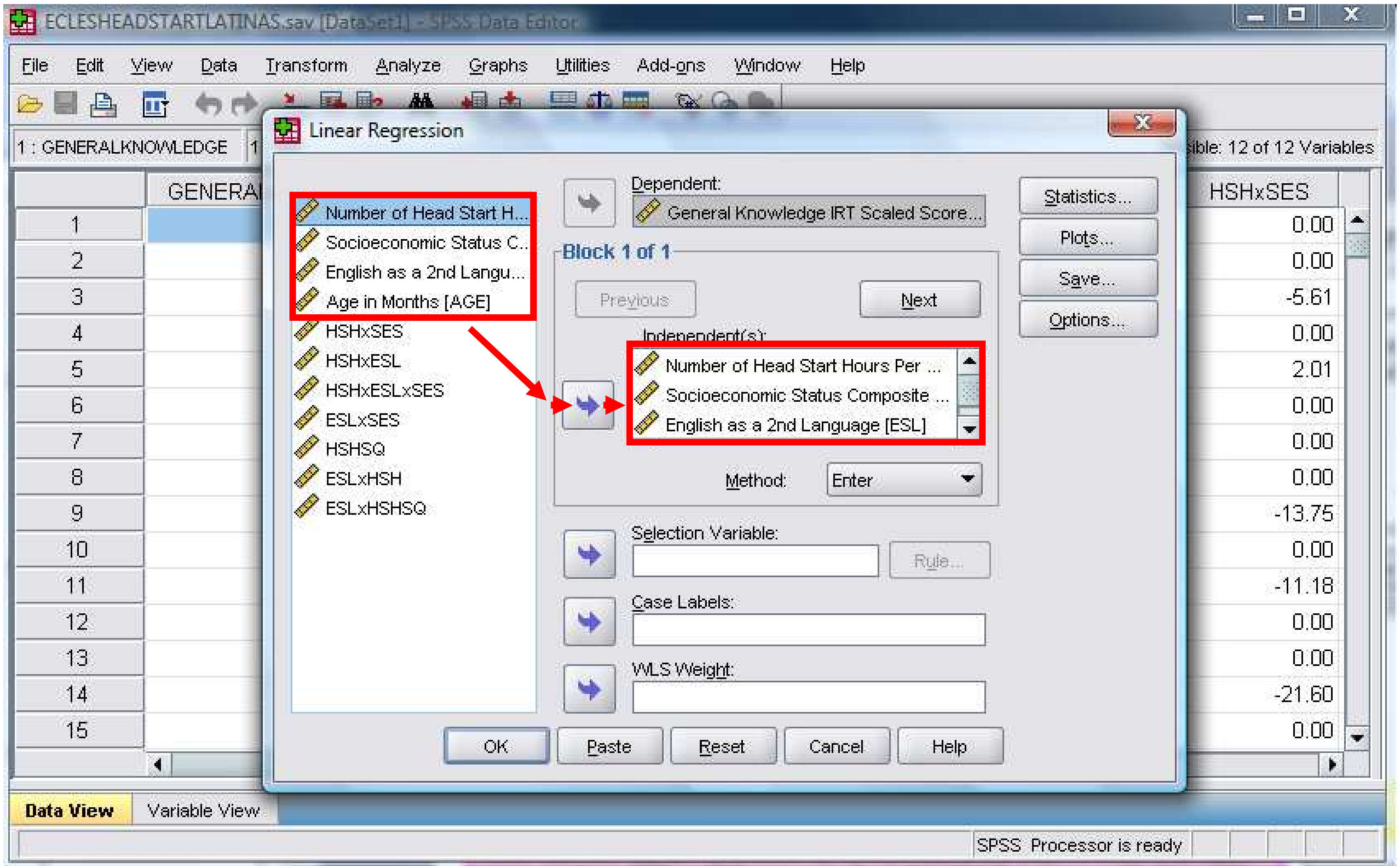
```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA
  /CRITERIA=PIN(.05)POUT(.10)
  /NOORIGIN
  /DEPENDENT GENERALKNOWLEDGE
  /METHOD=ENTER HEADSTARTHOURS SES ESL AGE
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HIST(ZRESID) NORM(ZRESID)
  /SAVE ZPRED COOK LEVER ZRESID.
GRAPH
  /SCATTERPLOT(BIVAR)=HEADSTARTHOURS WITH ZRE_1
  /MISSING=LISTWISE.
GRAPH
  /SCATTERPLOT(BIVAR)=SES WITH ZRE_1
  /MISSING=LISTWISE.
GRAPH
  /SCATTERPLOT(BIVAR)=ESL WITH ZRE_1
  /MISSING=LISTWISE.
GRAPH
  /SCATTERPLOT(BIVAR)=AGE WITH ZRE_1
  /MISSING=LISTWISE.
```

Go to Analyze > Regression > Linear…

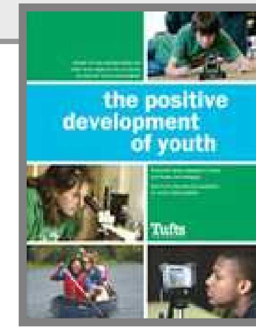


Obtaining SPSS Output

Do everything as usual, but include ALL your predictors, not just your question predictor, as an "Independent (Variable)."



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



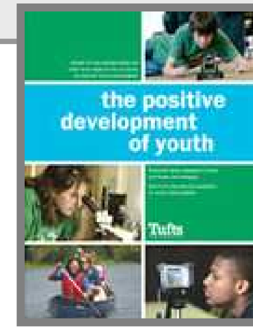
- **4-H Study of Positive Youth Development**
- **Source: Subset of data from IARYD, Tufts University**

- **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

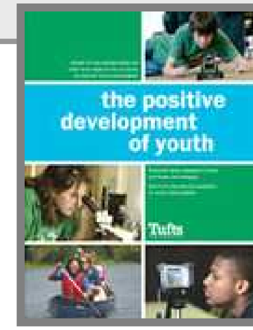
4-H Study of Positive Youth Development (4H.sav)



Statistics

		Grades in School	Female = 1, Male = 0	Self-Worth	Self-Percieved Academic Competence	Birth Mother Education
N	Valid	409	409	409	409	409
	Missing	0	0	0	0	0
Mean		3.3802	.60	3.1209	3.0292	13.86
Std. Deviation		.75184	.491	.60645	.65793	2.289
Minimum		.50	0	1.00	1.00	8
Maximum		4.00	1	4.00	4.00	20
Percentiles	25	3.0000	.00	2.6667	2.5000	12.00
	50	3.5000	1.00	3.1667	3.0000	13.00
	75	4.0000	1.00	3.6667	3.5000	16.00

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



Model Summary[e]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.058[a]	.003	.001	.75150	.003	1.372	1	407	.242
2	.352[b]	.124	.120	.70546	.121	55.865	1	406	.000
3	.568[c]	.323	.318	.62091	.199	119.090	1	405	.000
4	.577[d]	.333	.327	.61692	.010	6.261	1	404	.013

a. Predictors: (Constant), Female = 1, Male = 0

b. Predictors: (Constant), Female = 1, Male = 0, Self-Worth

c. Predictors: (Constant), Female = 1, Male = 0, Self-Worth, Self-Percieved Academic Competence

d. Predictors: (Constant), Female = 1, Male = 0, Self-Worth, Self-Percieved Academic Competence, Birth Mother Education

e. Dependent Variable: Grades in School

ANOVA[e]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.775	1	.775	1.372	.242[a]
	Residual	229.855	407	.565		
	Total	230.630	408			
2	Regression	28.577	2	14.288	28.711	.000[b]
	Residual	202.053	406	.498		
	Total	230.630	408			
3	Regression	74.490	3	24.830	64.405	.000[c]
	Residual	156.140	405	.386		
	Total	230.630	408			
4	Regression	76.873	4	19.218	50.496	.000[d]
	Residual	153.757	404	.381		
	Total	230.630	408			

a. Predictors: (Constant), Female = 1, Male = 0

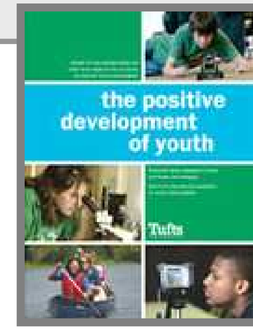
b. Predictors: (Constant), Female = 1, Male = 0, Self-Worth

c. Predictors: (Constant), Female = 1, Male = 0, Self-Worth, Self-Percieved Academic Competence

d. Predictors: (Constant), Female = 1, Male = 0, Self-Worth, Self-Percieved Academic Competence, Birth Mother Education

e. Dependent Variable: Grades in School

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



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	3.327	.059		56.872	.000	3.212	3.442
	Female = 1, Male = 0	.089	.076	.058	1.171	.242	-.060	.238
2	(Constant)	1.970	.190		10.386	.000	1.597	2.343
	Female = 1, Male = 0	.110	.071	.072	1.545	.123	-.030	.250
	Self-Worth	.431	.058	.347	7.474	.000	.317	.544
3	(Constant)	1.218	.181		6.743	.000	.863	1.573
	Female = 1, Male = 0	.126	.063	.082	2.007	.045	.003	.249
	Self-Worth	.085	.060	.068	1.414	.158	-.033	.202
	Self-Percieved Academic Competence	.602	.055	.527	10.913	.000	.493	.710
4	(Constant)	.835	.236		3.536	.000	.371	1.298
	Female = 1, Male = 0	.143	.063	.093	2.283	.023	.020	.266
	Self-Worth	.085	.059	.068	1.425	.155	-.032	.202
	Self-Percieved Academic Competence	.563	.057	.492	9.869	.000	.450	.675
	Birth Mother Education	.035	.014	.108	2.502	.013	.008	.063

a. Dependent Variable: Grades in School

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