Unit 10: Road Map (VERBAL)

Nationally Representative Sample of 7,800 8th Graders Surveyed in 1988 (NELS 88).
Outcome Variable (aka Dependent Variable):

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

Predictor Variables (aka Independent Variables):

FREELUNCH, a dichotomous variable, 1 = Eligible for Free/Reduced Lunch and 0 = Not
RACE, a polychotomous variable, 1 = Asian, 2 = Latino, 3 = Black and 4 = White

- Unit 1: In our sample, is there a relationship between reading achievement and free lunch?
- Unit 2: In our sample, what does reading achievement look like (from an outlier resistant perspective)?
- Unit 3: In our sample, what does reading achievement look like (from an outlier sensitive perspective)?
- Unit 4: In our sample, how strong is the relationship between reading achievement and free lunch?
- Unit 5: In our sample, free lunch predicts what proportion of variation in reading achievement?
- Unit 6: In the population, is there a relationship between reading achievement and free lunch?
- Unit 7: In the population, what is the magnitude of the relationship between reading and free lunch?
- Unit 8: What assumptions underlie our inference from the sample to the population?
- Unit 9: In the population, is there a relationship between reading and race?
- Unit 10: In the population, is there a relationship between reading and race controlling for free lunch?
- Appendix A: In the population, is there a relationship between race and free lunch?

Unit I0: Roadmap (SPSS Output)

Model Summary

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

a. Predictors: (Constant), FREELUNCH

Unit 5

Statistics

		READING	FREELUNCH
N	Valid	7800	7800
	Missing	0	0
Mean		47.4940	.3354
Std. Deviation		8.56944	.47216
Minimum		23.96	.00
Maximum		63.49	1.00
Percentiles	25	41.2400	.0000
	50	47.4300	.0000
	75	53.9300	1.0000

Unit 3

Unit 2

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	40744.322	1	40744.322	597.251	.000[a]
	Residual	531977.541	7798	68.220		
	Total	572721.864	7799			

a. Predictors: (Constant), FREELUNCH

b. Dependent Variable: READING

Unit 9

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	49.118	.115		428.169	.000	48.893	49.342
	FREELUNCH	-4.841	.198	-.267	-24.439	.000	-5.229	-4.453

a. Dependent Variable: READING

Unit 1 Unit 8 Unit 4 Unit 6 Unit 7

Unit I0: Roadmap (SPSS Output)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.221[a]	.049	.049	8.35882

a. Predictors: (Constant), BLACK, ASIAN, LATINO

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	28016.721	3	9338.907	133.662	.000[a]
	Residual	544705.143	7796	69.870		
	Total	572721.864	7799			

a. Predictors: (Constant), BLACK, ASIAN, LATINO

b. Dependent Variable: READING

Unit 9

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
	LATINO	-4.418	.306	-.161	-14.447	.000	-5.017	-3.818
	BLACK	-4.889	.339	-.161	-14.423	.000	-5.554	-4.225

a. Dependent Variable: READING

Unit I0: Roadmap (SPSS Output)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.314[a]	.099	.098	8.13954

a. Predictors: (Constant), FREELUNCHxBLACK, FREELUNCHxASIAN, FREELUNCHxLATINO, ASIAN, FREELUNCH, LATINO, BLACK

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	56485.879	7	8069.411	121.799	.000[a]
	Residual	516235.985	7792	66.252		
	Total	572721.864	7799			

a. Predictors: (Constant), FREELUNCHxBLACK, FREELUNCHxASIAN, FREELUNCHxLATINO, ASIAN, FREELUNCH, LATINO, BLACK

b. Dependent Variable: READING

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	49.439	.127		389.587	.000	49.191	49.688
	FREELUNCH	-3.882	.238	-.214	-16.293	.000	-4.349	-3.415
	ASIAN	1.491	.431	.043	3.461	.001	.646	2.335
	LATINO	-3.250	.426	-.119	-7.624	.000	-4.086	-2.415
	BLACK	-3.406	.504	-.112	-6.764	.000	-4.393	-2.419
	FREELUNCHxASIAN	-2.472	.865	-.037	-2.858	.004	-4.167	-.776
	FREELUNCHxLATINO	-.363	.606	-.010	-.600	.549	-1.550	.824
	FREELUNCHxBLACK	-.501	.678	-.013	-.739	.460	-1.829	.828

a. Dependent Variable: READING

Unit 10 Multiple Regression

Unit I0: Roadmap (SPSS Output)

Tests of Between-Subjects Effects

Dependent Variable:READING

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	56485.879[a]	7	8069.411	121.799	.000
Intercept	6087049.770	1	6087049.770	91877.151	.000
RACE	14705.701	3	4901.900	73.989	.000
FREELUNCH	16141.887	1	16141.887	243.644	.000
RACE * FREELUNCH	559.039	3	186.346	2.813	.038
Error	516235.985	7792	66.252		
Total	1.817E7	7800			
Corrected Total	572721.864	7799			

a. R Squared = .099 (Adjusted R Squared = .098)

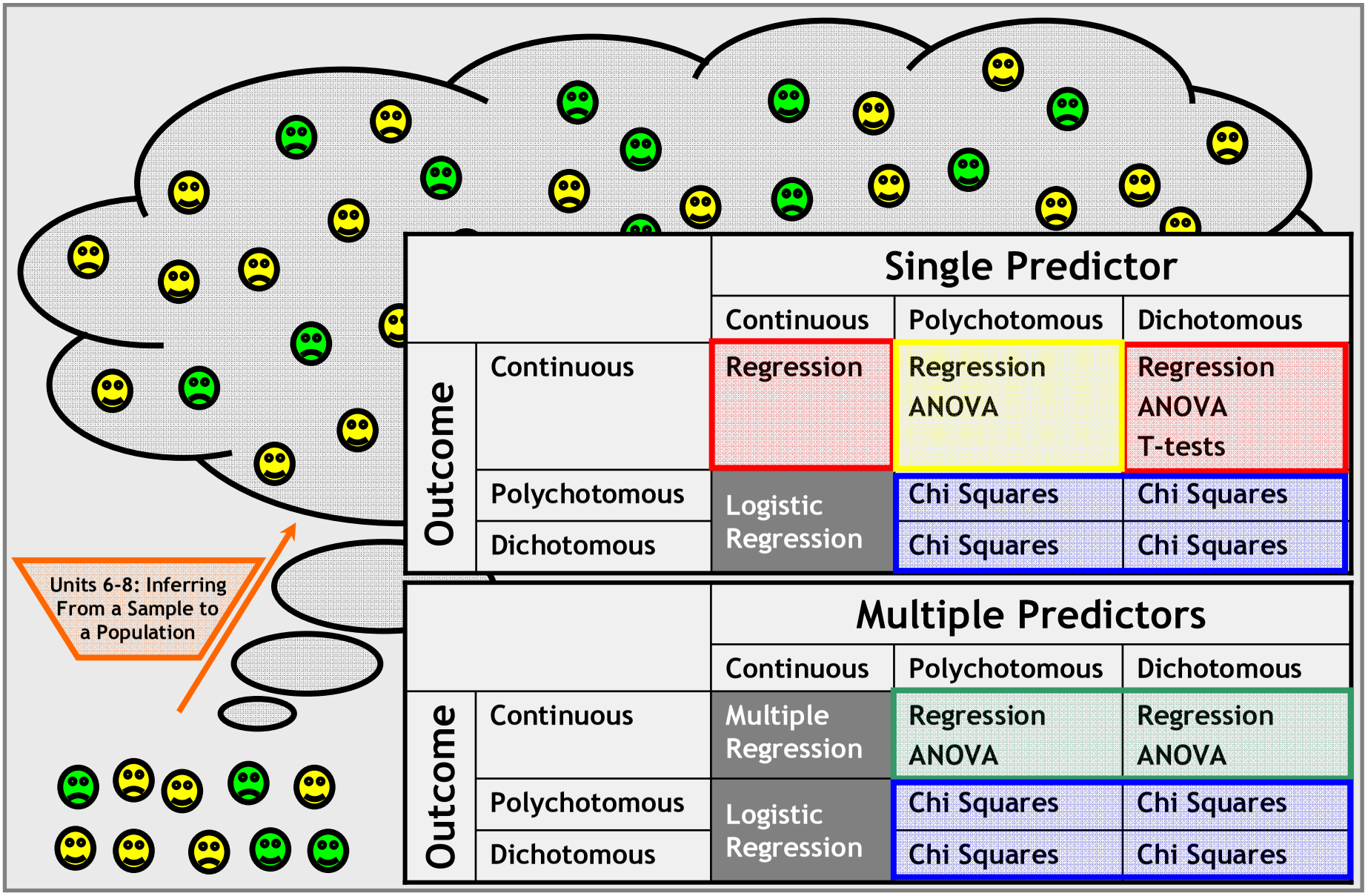
Unit 10
ANOVA

FREELUNCH * R^S RACE/ETHNIC BACKGROUND Crosstabulation

Appendix A

			R^S RACE/ETHNIC BACKGROUND				
			Asian	Latino	Black	White	Total
FREELUNCH	0	Count	391	400	279	4114	5184
		Expected Count	344.3	570.9	451.9	3816.9	5184.0
		Std. Residual	2.5	-7.2	-8.1	4.8	
	1	Count	127	459	401	1629	2616
		Expected Count	173.7	288.1	228.1	1926.1	2616.0
		Std. Residual	-3.5	10.1	11.5	-6.8	
	Total	Count	518	859	680	5743	7800
		Expected Count	518.0	859.0	680.0	5743.0	7800.0

Unit 10: Road Map (Schematic)



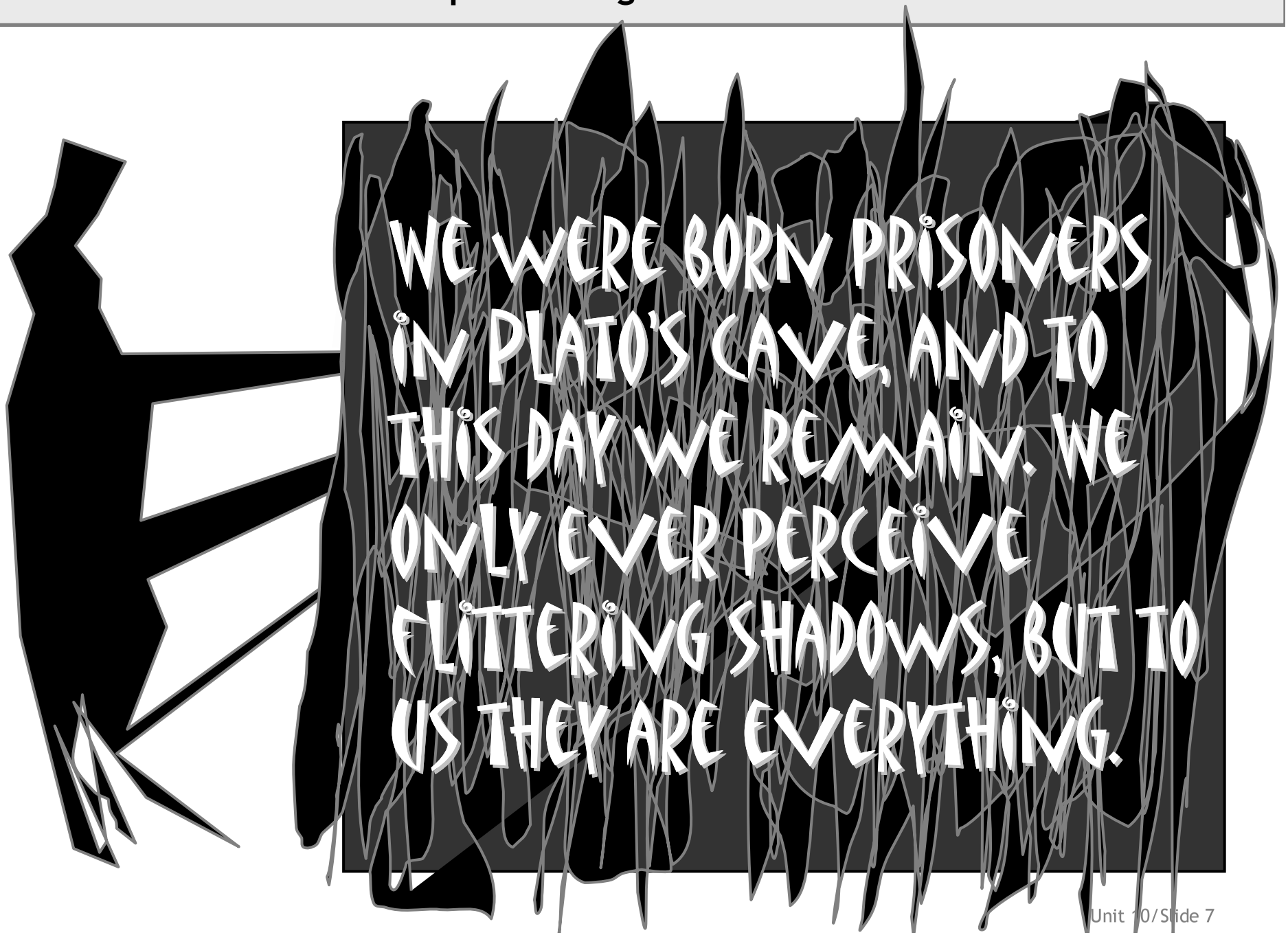
Single Predictor

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

Multiple Predictors

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

Epistemological Minute

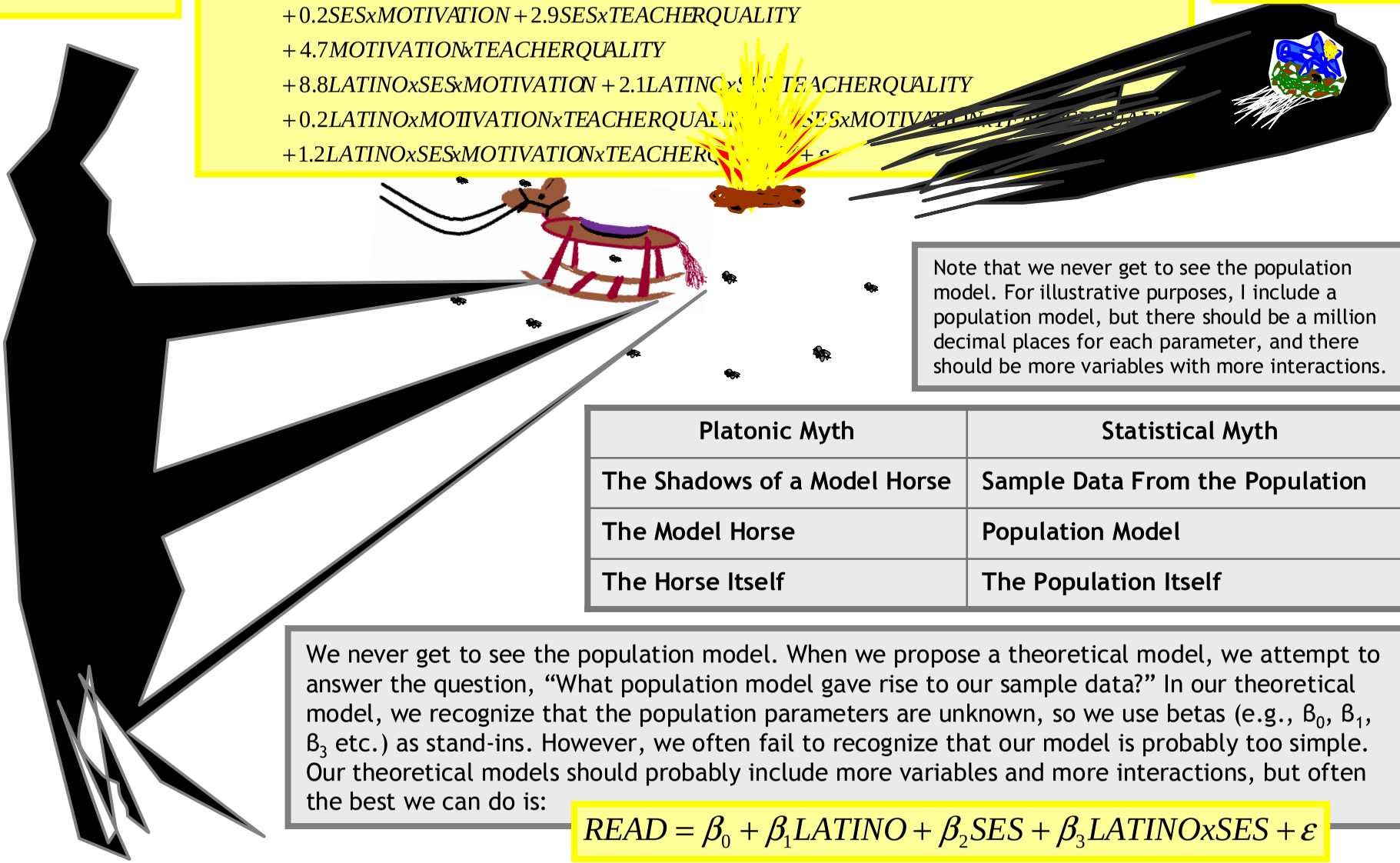


Epistemological Minute

sample data

$$READ = 3.4 + 6.7LATINO + 1.4SES + 9.3MOTIVATION + 8.4TEACHERQUALITY$$
$$+ 3.7LATINOxSES + 6.2LATINOxMOTIVATION + 6.1LATINOxTEACHERQUALITY$$
$$+ 0.2SESxMOTIVATION + 2.9SESxTEACHERQUALITY$$
$$+ 4.7MOTIVATIONxTEACHERQUALITY$$
$$+ 8.8LATINOxSESxMOTIVATION + 2.1LATINOxSESxTEACHERQUALITY$$
$$+ 0.2LATINOxMOTIVATIONxTEACHERQUALITY + SESxMOTIVATIONxTEACHERQUALITY$$
$$+ 1.2LATINOxSESxMOTIVATIONxTEACHERQUALITY + \varepsilon$$

Students
with a capital S



Note that we never get to see the population model. For illustrative purposes, I include a population model, but there should be a million decimal places for each parameter, and there should be more variables with more interactions.

Platonic Myth	Statistical Myth
The Shadows of a Model Horse	Sample Data From the Population
The Model Horse	Population Model
The Horse Itself	The Population Itself

We never get to see the population model. When we propose a theoretical model, we attempt to answer the question, "What population model gave rise to our sample data?" In our theoretical model, we recognize that the population parameters are unknown, so we use betas (e.g., β_0, β_1, β_3 etc.) as stand-ins. However, we often fail to recognize that our model is probably too simple. Our theoretical models should probably include more variables and more interactions, but often the best we can do is:

$$READ = \beta_0 + \beta_1 LATINO + \beta_2 SES + \beta_3 LATINOxSES + \varepsilon$$

Unit 10 Post Hole:

Interpret a two-way analysis of variance using F-tests and graphs.

Unit 10 Technical Memo and School Board Memo:

Conduct a two-way analysis of variance, produce an appropriate table and graph, fit the equivalent regression model, and discuss your results.

Unit 10 (and Unit 9) Reading:

http://onlinestatbook.com/
Chapter 8, ANOVA

Unit 10: Technical Memo and School Board Memo

Conduct one analysis (but from two perspectives) using the Sport.sav data set.

Answer the following research question:

* Given that boys' self-perceptions of athletic ability tend to be greater than girls' self-perceptions in the population of U.S. students, does the boy/girl difference vary from the third grade to the sixth grade to the ninth grade?

Conduct the analysis from a regression perspective.

Conduct the analysis from an ANOVA perspective.

Unit 10: Technical Memo and School Board Memo

Work Products (Part I of II):

I. Technical Memo: Have one section per biviariate analysis. For each section, follow this outline. (2 Sections)

- A. Introduction
  - i. State a theory (or perhaps hunch) for the relationship—think causally, be creative. (1 Sentence)
  - ii. State a research question for each theory (or hunch)—think correlationally, be formal. Now that you know the statistical machinery that justifies an inference from a sample to a population, begin each research question, "In the population,…" (1 Sentence)
  - iii. List the two variables, and label them "outcome" and "predictor," respectively.
  - iv. Include your theoretical model.
- B. Univariate Statistics. Describe your variables, using descriptive statistics. What do they represent or measure?
  - i. Describe the data set. (1 Sentence)
  - ii. Describe your variables. (1 Short Paragraph Each)
    - a. Define the variable (parenthetically noting the mean and s.d. as descriptive statistics).
    - b. Interpret the mean and standard deviation in such a way that your audience begins to form a picture of the way the world is. Never lose sight of the substantive meaning of the numbers.
    - c. Polish off the interpretation by discussing whether the mean and standard deviation can be misleading, referencing the median, outliers and/or skew as appropriate.
- C. Correlations. Provide an overview of the relationships between your variables using descriptive statistics.
  - i. Interpret all the correlations with your outcome variable. Compare and contrast the correlations in order to ground your analysis in substance. (1 Paragraph)
  - ii. Interpret the correlations among your predictors. Discuss the implications for your theory. As much as possible, tell a coherent story. (1 Paragraph)
  - iii. As you narrate, note any concerns regarding assumptions (e.g., outliers or non-linearity), and, if a correlation is uninterpretable because of an assumption violation, then do not interpret it.

Unit 10: Technical Memo and School Board Memo

Work Products (Part II of II):

I. Technical Memo (continued)

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

i. Include your fitted model.

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

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

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

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

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

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

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

ii. For 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)

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 10: Research Question I (Regression Perspective)

Theory: The Anglo/Latino reading gap is an artifact of Anglo/Latino differences in socioeconomic status. Once we statistically control for socioeconomic status, the gap will disappear. This will be true for four-year-college bound boys.

Research Question: Controlling for socioeconomic status, is there a statistically significant difference in reading ability between Anglo students and Latino students in our sample of four-year-college bound boys? Notice that this research question is about the sample, not the population, because I am theorizing zero relationship in the population (i.e., the null hypothesis), but we can never confirm the null hypothesis.

Unit 10: Research Question I

Data Set: NELSBoys.sav National Education Longitudinal Survey (1988), a subsample of 1820 four-year-college bound boys, of whom 182 are Latino and the rest are Anglo.

Variables:

Outcome—Reading Achievement Score (*READ*)

Predictors—Latino = 1, Anglo = 0 (*LATINO*)

—Low SES=1, Mid SES=2, High SES=3 (*SocioeconomicStatus*)

Model:

$$READ = \beta_0 + \beta_1 LATINO + \beta_2 LowSES + \beta_3 HighSES + \beta_4 LowSESxLATINO + \beta_5 HighSESxLATINO + \varepsilon$$

$$\begin{aligned} READ = \beta_0 &+ \beta_1 LATINO \\ &+ \beta_2 LowSES + \beta_3 HighSES \\ &+ \beta_4 LowSESxLATINO + \beta_5 HighSESxLATINO \\ &+ \varepsilon \end{aligned}$$

Multiple Regression with Dummies and Interactions

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.251[a]	.063	.061	9.06696

a. Predictors: (Constant), HighSESXLatino, LowSESXLatino, HighSES, LowSES, Latino

b. Dependent Variable: Reading score

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10067.229	5	2013.446	24.492	.000[a]
	Residual	149128.430	1814	82.210		
	Total	159195.659	1819			

a. Predictors: (Constant), HighSESXLatino, LowSESXLatino, HighSES, LowSES, Latino

b. Dependent Variable: Reading score

There is a statistically significant relationship between our outcome and our predictors, $F(5,1814)=24.492$, $p < 0.05$. Ethnicity and socioeconomic status (and their interaction) predict about 6% of the variation in reading scores.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Read
  /METHOD=ENTER Latino LowSES HighSES LowSESXLatino HighSESXLatino
  /SCATTERPLOT=(*ZRESID ,*ZPRED).
```

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	54.507	.413		131.982	.000	53.697	55.317
	Latino	-.840	1.270	-.027	-.661	.508	-3.331	1.651
	LowSES	-.975	.768	-.037	-1.269	.205	-2.481	.532
	HighSES	3.001	.506	.159	5.930	.000	2.009	3.994
	LowSESXLatino	-5.297	1.800	-.107	-2.942	.003	-8.828	-1.766
	HighSESXLatino	-3.621	1.772	-.067	-2.043	.041	-7.097	-.145

a. Dependent Variable: Reading score

Interpreting The Mess: Plug and Play (Part I of II)

$$READ = \beta_0 + \beta_1 LATINO + \beta_2 LowSES + \beta_3 HighSES + \beta_4 LowSESxLATINO + \beta_5 HighSESxLATINO + \varepsilon$$

$$\hat{READ} = 54.5 - 0.8(LATINO) - 1.0(LowSES) + 3.0(HighSES) - 5.3(LowSES * LATINO) - 3.6(HighSES * LATINO)$$

Let's first look at our predictions for Anglo students (LATINO = 0):

$$\hat{READ} = 54.5 - 0.8(0) - 1.0(LowSES) + 3.0(HighSES) - 5.3(LowSES * 0) - 3.6(HighSES * 0)$$

$$\hat{READ} = 54.5 - 1.0(LowSES) + 3.0(HighSES)$$

Let's then look at our predictions for Latino students (LATINO = 1):

$$\hat{READ} = 54.5 - 0.8(1) - 1.0(LowSES) + 3.0(HighSES) - 5.3(LowSES * 1) - 3.6(HighSES * 1)$$

$$\hat{READ} = 53.7 - 6.3(LowSES) - 0.6(HighSES)$$

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	54.507	.413		131.982	.000	53.697	55.317
	Latino	-.840	1.270	-.027	-.661	.508	-3.331	1.651
	LowSES	-.975	.768	-.037	-1.269	.205	-2.481	.532
	HighSES	3.001	.506	.159	5.930	.000	2.009	3.994
	LowSESXLatino	-5.297	1.800	-.107	-2.942	.003	-8.828	-1.766
	HighSESXLatino	-3.621	1.772	-.067	-2.043	.041	-7.097	-.145

a. Dependent Variable: Reading score

Interpreting The Mess: Plug and Play (Part II of II)

Anglo students (LATINO = 0):

$$\hat{READ} = 54.5 - 1.0(LowSES) + 3.0(HighSES)$$

Latino students (LATINO = 1):

$$\hat{READ} = 53.7 - 6.3(LowSES) - 0.6(HighSES)$$

Low SES students (SocioEconomicStatus=1: LowSES = 1 and HighSES = 0):

$$\hat{READ} = 54.5 - 1.0(1) + 3.0(0) = 53.5$$

$$\hat{READ} = 53.7 - 6.3(1) - 0.6(0) = 47.4$$

Mid SES students (SocioEconomicStatus=2: LowSES = 0 and HighSES = 0):

$$\hat{READ} = 54.5 - 1.0(0) + 3.0(0) = 54.5$$

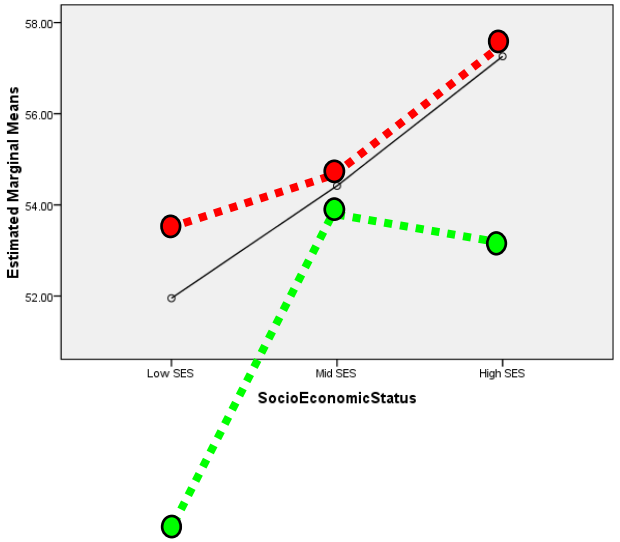
$$\hat{READ} = 53.7 - 6.3(0) - 0.6(0) = 53.7$$

High SES students (SocioEconomicStatus=3: LowSES = 0 and HighSES = 1):

$$\hat{READ} = 54.5 - 1.0(0) + 3.0(1) = 57.5$$

$$\hat{READ} = 53.7 - 6.3(0) - 0.6(1) = 53.1$$

The Anglo/Latino reading gap differs by socioeconomic status. There appears to be little or no gap for four-year-college bound boys of middle SES. However, there are large gaps of 6.1 and 4.4 points for students of low and high SES, respectively. (Note, we can also write about how the SES/Reading relationship differs for Anglo students and Latino students.)



This is our graph from Unit 9 that shows a relationship between socioeconomic status and reading scores. We'll add to it in light of our new information.

Interpreting the Parameter Estimates

1. Write out your fitted model.
2. Create the first branches of your tree.
   - -Choose a factor (e.g., ethnicity).
   - -For each level (e.g., Anglo or Latino) create a branch.
   - -Each branch is a fitted model, partially instantiated.
3. Create the next branches of your tree.
   - -For each level of the other factor (e.g., high, mid, low SES) create a branch.
   - -Each branch is a fitted model, fully instantiated.
4. Graph it.
   - -Each point is a mean for a subgroup.
   - -Connect the points with colored dotted lines.
5. Make sense of the graph in real-world terms.
   - -Consider which mean differences might be statistically significant.
   - -Consider which interactions (non-parallelisms) might be statistically significant.

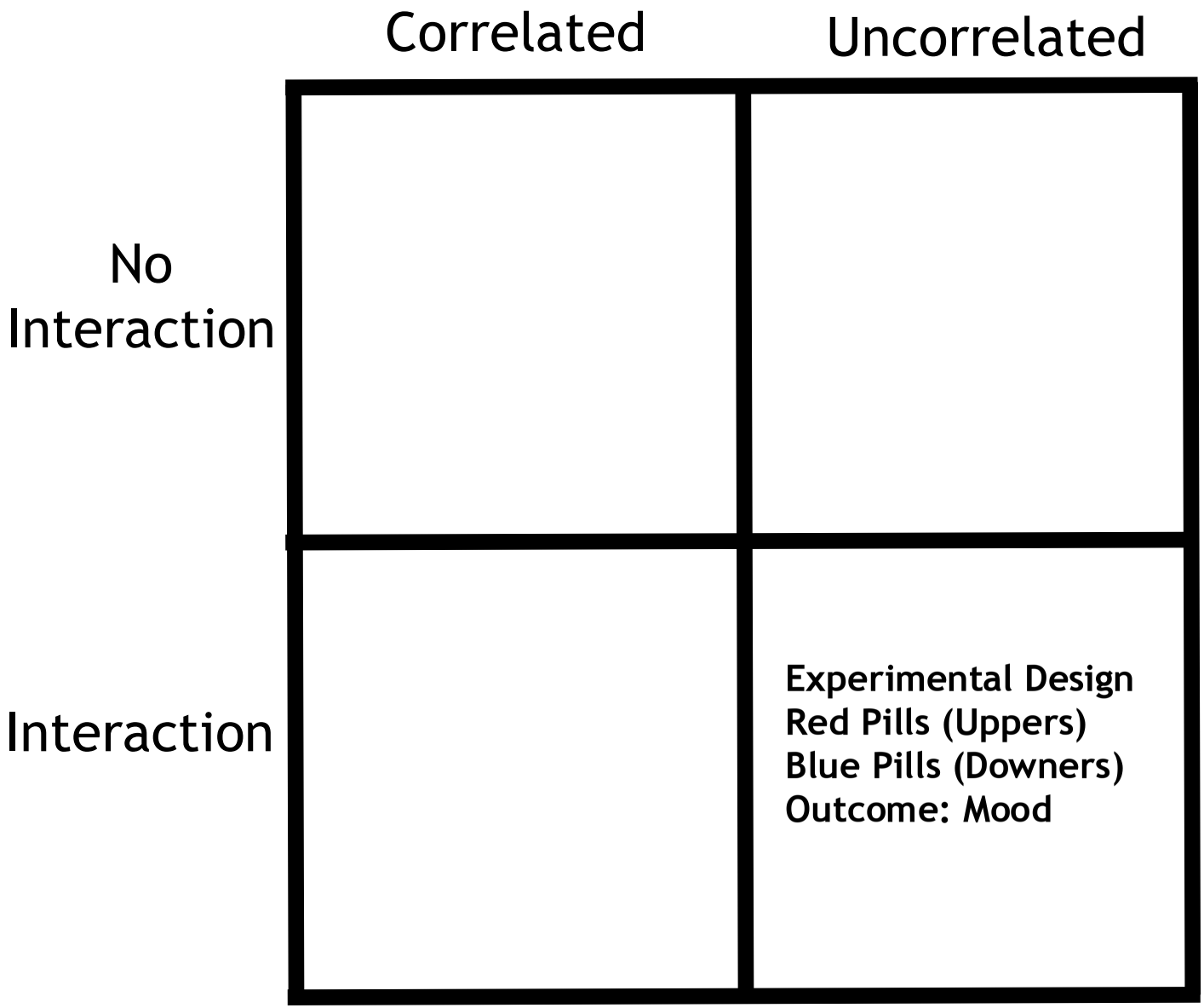
Note, some people find it easier to collapse steps 2 and 3. Determine your subgroups (e.g., low SES Anglo students etc.) and solve for each subgroup (LowSES = 1, HighSES = 0, and Latino = 0).

This probably should be a post hole, but it's not!

Filling in the Gaps

- **What is statistical interaction?**
  - Abstractly: Sometimes the relationship between your outcome and one predictor differs by the level of another predictor.
  - Geometrically: Sometimes your trend lines are not parallel.
  - Practically: Sometimes the effectiveness of your intervention or program differs by gender, SES, age, proficiency etc. Or, sometimes the effectiveness of a drug is helped or hindered by another drug.
- **What is statistical control?**
  - A short introduction: Sometimes we include a predictor in our model not because we are interested in the relationship between that predictor and the outcome, but rather because we are uninterested. Everybody knows SES is correlated with academic achievement, who cares? If you are interested in the Anglo/Latino achievement gap, you want to include SES in your model exactly because you do not care about SES. By including SES in your model, you get to compare Anglo students and Latino students of equal SES—you are statistically controlling for SES.
- **How do you check assumptions when you have multiple predictors?**
  - Create a residual vs. fitted scatterplot.
  - Residual values identify how wrong your prediction is.
  - Fitted values identify your prediction.
  - A residuals vs. fitted plot tells us how wrong our prediction is for each prediction.
  - You can use a residual vs. fitted plot to search HI-N-LO.

Conceptually Distinct: Interactions and Correlations



Conceptually Distinct: Interactions and Correlations

- **Experimental Design**

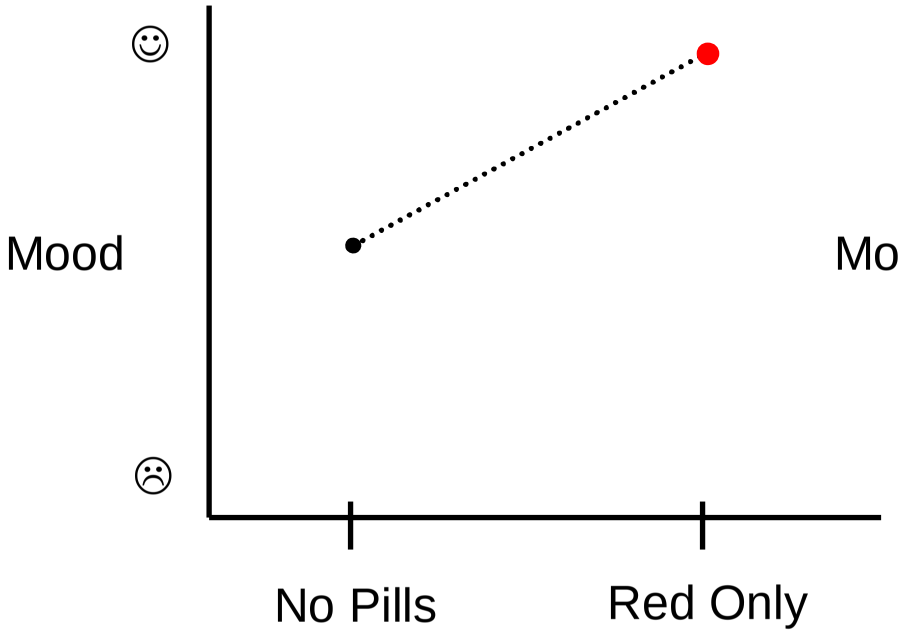
Review: Two variables are uncorrelated if, and only if, knowing one does not help you predict the other.

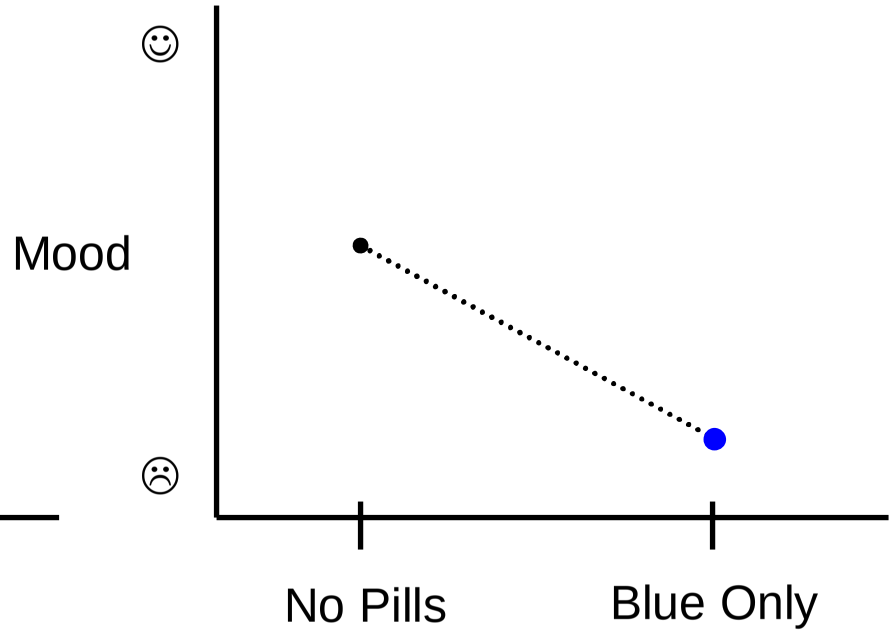
Interaction is about three or more variables: one outcome and at least two predictors.

10 7th Graders Take Red Pills (Uppers) Only
10 7th Graders Take Blue Pills (Downers) Only
10 7th Graders Take Both Pills
10 7th Graders Take None

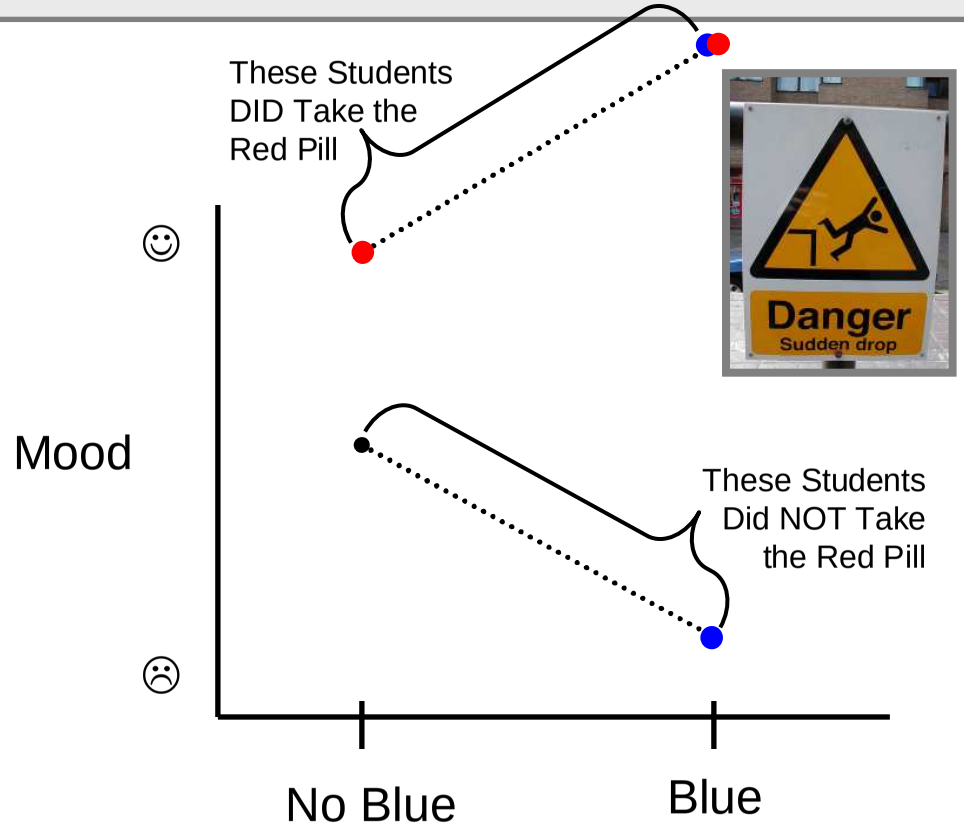
By Design, Red and Blue Conditions are Uncorrelated!

By Mrs. Dimitri, my 7th grade guidance counselor, Red Pills Interact with Blue Pills! In fact, if you take both pills at the same time, they will make you want to jump off the roof.





Conceptually Distinct: Interactions and Correlations

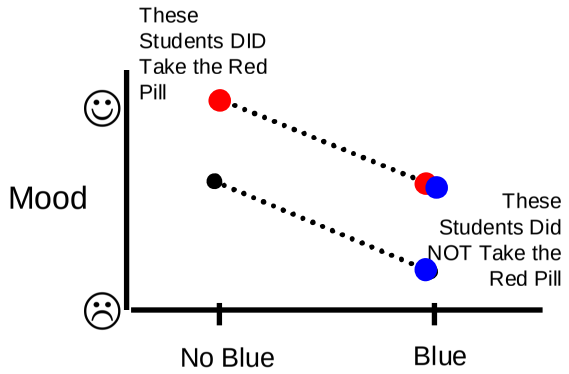


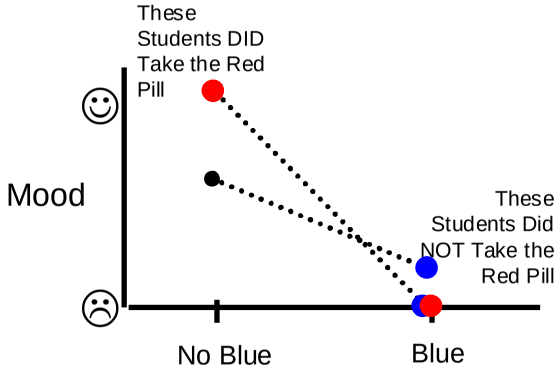
When the lines are parallel, there is no interaction.

When the lines are non-parallel there is an interaction.

When the non-parallel lines do not cross, the interaction is ordinal.

When the non-parallel lines cross, the interaction is disordinal.





Including Interaction Terms in Your Regression Model

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.242[a]	.059	.057	9.08459

a. Predictors: (Constant), HighSES, Latino, LowSES

b. Dependent Variable: Reading score

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Read
  /METHOD=ENTER Latino LowSES HighSES
  /SCATTERPLOT=(*ZRESID ,*ZPRED).
```

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9321.513	3	3107.171	37.649	.000[a]
	Residual	149874.145	1816	82.530		
	Total	159195.659	1819			

a. Predictors: (Constant), HighSES, Latino, LowSES

b. Dependent Variable: Reading score

A main effects model has no interaction terms. The trend lines are constrained to be parallel.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	54.822	.399		137.458	.000	54.040	55.604
	Latino	-3.817	.729	-.122	-5.236	.000	-5.247	-2.387
	LowSES	-1.887	.691	-.071	-2.729	.006	-3.243	-.531
	HighSES	2.650	.485	.141	5.461	.000	1.699	3.602

a. Dependent Variable: Reading score

Including Interaction Terms in Your Regression Model

```
COMPUTE LowSESXLatino=(LowSES*Latino).
COMPUTE HighSESXLatino=(HighSES*Latino).
Execute.
```

Metaphorically, the crossproduct terms (i.e., interaction terms) allow the predictors to talk to one another.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Read
  /METHOD=ENTER Latino LowSES HighSES LowSESXLatino HighSESXLatino
  /SCATTERPLOT=(*ZRESID ,*ZPRED).
```

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	54.507	.413		131.982	.000	53.697	55.317
	Latino	-.840	1.270	-.027	-.661	.508	-3.331	1.651
	LowSES	-.975	.768	-.037	-1.269	.205	-2.481	.532
	HighSES	3.001	.506	.159	5.930	.000	2.009	3.994
	LowSESXLatino	-5.297	1.800	-.107	-2.942	.003	-8.828	-1.766
	HighSESXLatino	-3.621	1.772	-.067	-2.043	.041	-7.097	-.145

a. Dependent Variable: Reading score

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

1 : ID | 6898594 | Visible: 2

	ID	Read	Black	Latino	SocioEconomicStatus	LowSES	HighSES	LowSESXLatino	HighSESXLatino
1	6898594.00	37.57	0.00	1.00	1.00	1.00	0.00	1.00	0.00
2	2582532.00	65.29	0.00	1.00	3.00	0.00	1.00	0.00	1.00
3	2564477.00	46.71	0.00	1.00	1.00	1.00	0.00	1.00	0.00
4	7957674.00	48.16	0.00	1.00	3.00	0.00	1.00	0.00	1.00
5	7955910.00	66.19	0.00	1.00	3.00	0.00	1.00	0.00	1.00

Filling in the Gaps

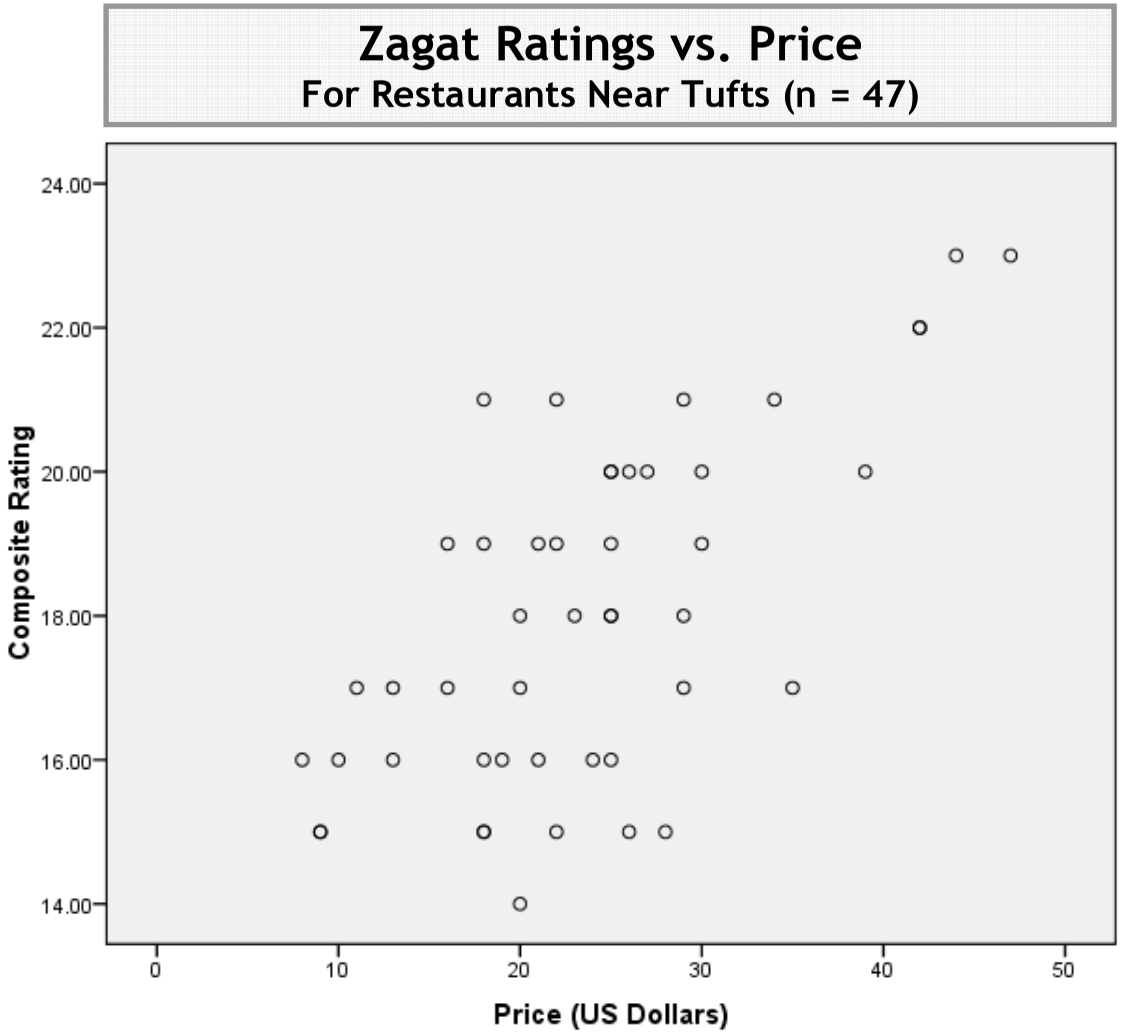
- **What is statistical interaction?**
  - Abstractly: Sometimes the relationship between your outcome and one predictor differs by the level of another predictor.
  - Geometrically: Sometimes your trend lines are not parallel.
  - Practically: Sometimes the effectiveness of your intervention or program differs by gender, SES, age, proficiency etc. Or, sometimes the effectiveness of a drug is helped or hindered by another drug.
- **What is statistical control?**
  - A short introduction: Sometimes we include a predictor in our model not because we are interested in the relationship between that predictor and the outcome, but rather because we are uninterested. Everybody knows SES is correlated with academic achievement, who cares? If you are interested in the Anglo/Latino achievement gap, you want to include SES in your model exactly because you do not care about SES. By including SES in your model, you get to compare Anglo students and Latino students of equal SES—you are statistically controlling for SES.
- **How do you check assumptions when you have multiple predictors?**
  - Create a residual vs. fitted scatterplot.
  - Residual values identify how wrong your prediction is.
  - Fitted values identify your prediction.
  - A residuals vs. fitted plot tells us how wrong our prediction is for each prediction.
  - You can use a residual vs. fitted plot to search HI-N-LO.

Introduction to Statistical Control

If money is no object, you go to the best restaurant regardless of the price. You know that the best restaurant will be expensive, but you don't care.

If you are budget conscious, however, then you want to maximize the value of your dining dollar. You want to patronize restaurants that are good for the price. You want to compare cheap eats to cheap eats and determine the best. And, you want to compare fine dining to fine dining and determine the best. In statistics, we compare apple to apples and oranges to oranges by including a control predictor in our model and analyzing the residuals.

Look at the scatterplot. Look for good deals. I bet you naturally conduct a residual analysis.



Introduction to Statistical Control

A Control Model:

$$RATING = \beta_0 + \beta_1 PRICE + \varepsilon$$

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.668[a]	.447	.435	1.80472

a. Predictors: (Constant), Price (US Dollars)

b. Dependent Variable: Composite Rating

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	118.413	1	118.413	36.356	.000[a]
	Residual	146.566	45	3.257		
	Total	264.979	46			

a. Predictors: (Constant), Price (US Dollars)

b. Dependent Variable: Composite Rating

Coefficients[a]

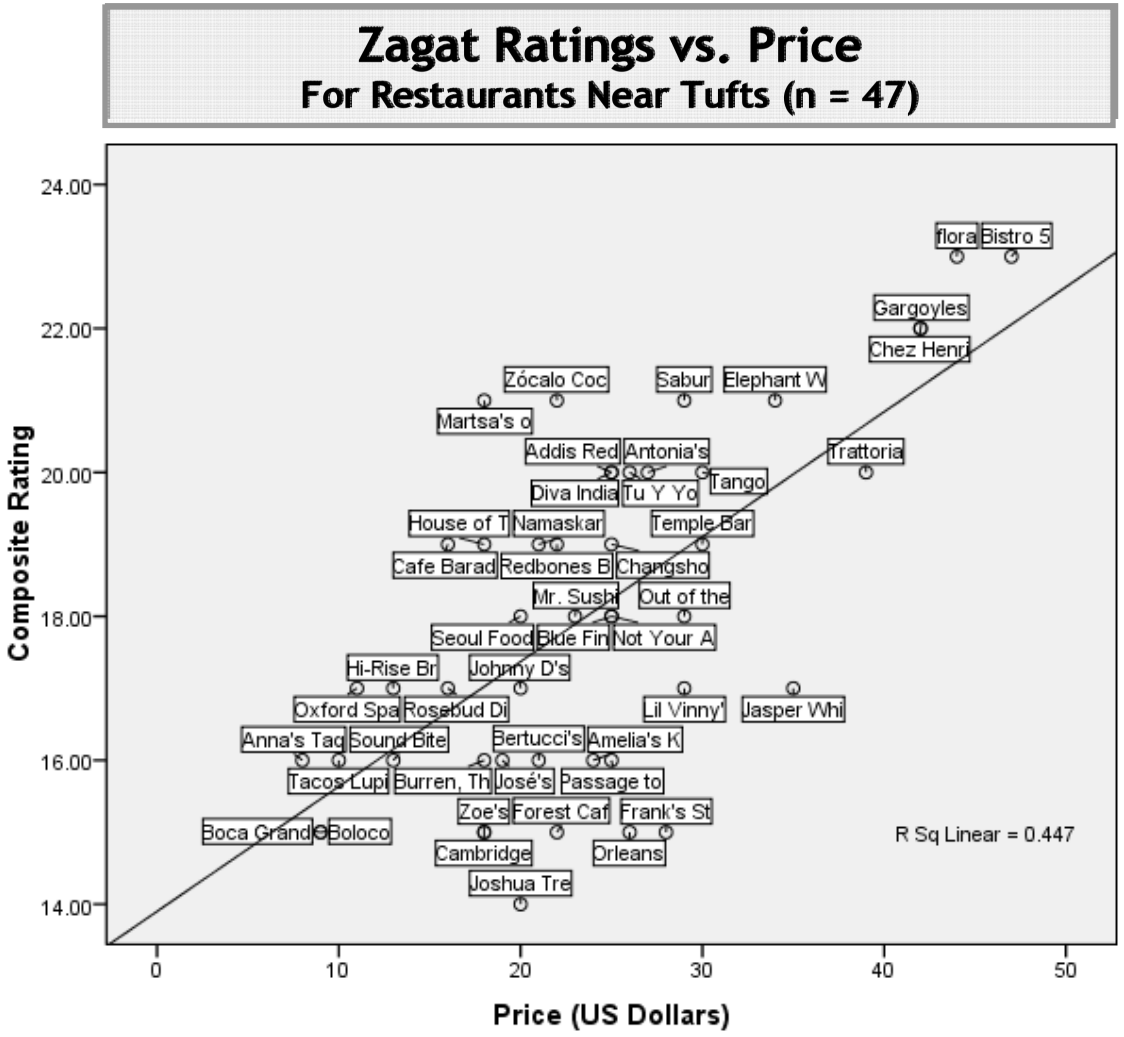
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	13.900	.732		18.978	.000
	Price (US Dollars)	.174	.029	.668	6.030	.000

a. Dependent Variable: Composite Rating

Introduction to Statistical Control

Once we fit the model, we produce residuals. Every observation has an associated residual. You can think of a residual as a controlled observation. In other words, the residuals are a measure of your outcome with the statistical effect of price removed.

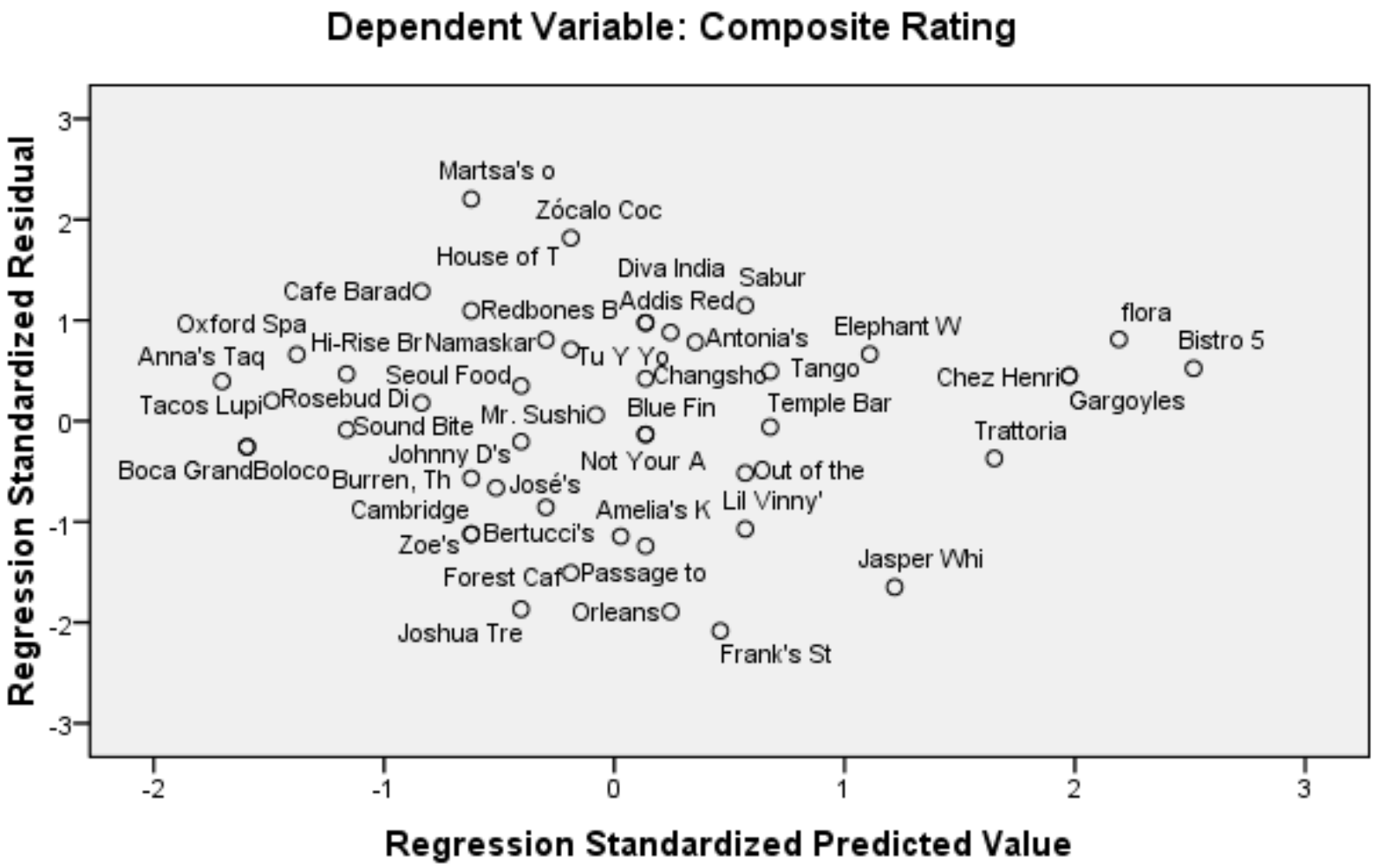
By including price in our model and examining the residuals, we statistically remove the relationship between rating and price. Restaurants with a large positive residuals are good values! We are statistically controlling for price. Now, if we include another variable in our model, say a dichotomous variable indicating town, Somerville or Medford, we will have controlled for price, and the controlled relationship between rating and town will tell us which town has the better *dining values* on average. If we wanted to know which town has the best restaurants, we would simply regress rating on town. But, we want to level the playing field in terms of price, so we regress rating on town and price simultaneously.



Likewise, if we want to consider the Latino/Anglo achievement gap, we may want to statistically level the SES playing field. We level the playing field by including in our model SES as a predictor along with ethnicity. Thus, we compare students of low SES to low SES, mid to mid, and high to high.

Introduction to Statistical Control

Residual vs. Fitted Plot
For Restaurants Near Tufts (n = 47)



Filling in the Gaps

- **What is statistical interaction?**
  - Abstractly: Sometimes the relationship between your outcome and one predictor differs by the level of another predictor.
  - Geometrically: Sometimes your trend lines are not parallel.
  - Practically: Sometimes the effectiveness of your intervention or program differs by gender, SES, age, proficiency etc. Or, sometimes the effectiveness of a drug is helped or hindered by another drug.
- **What is statistical control?**
  - A short introduction: Sometimes we include a predictor in our model not because we are interested in the relationship between that predictor and the outcome, but rather because we are uninterested. Everybody knows SES is correlated with academic achievement, who cares? If you are interested in the Anglo/Latino achievement gap, you want to include SES in your model exactly because you do not care about SES. By including SES in your model, you get to compare Anglo students and Latino students of equal SES—you are statistically controlling for SES.
- **How do you check assumptions when you have multiple predictors?**
  - Create a residual vs. fitted scatterplot.
  - Residual values identify how wrong your prediction is.
  - Fitted values identify your prediction.
  - A residuals vs. fitted plot tells us how wrong our prediction is for each prediction.
  - You can use a residual vs. fitted plot to search HI-N-LO.

Checking Our Regressions Assumptions: Searching HI-N-LO

Homoscedasticity: The variances are roughly equal for each prediction.

Independence: We cannot tell if the students are clustered in, for example, schools.

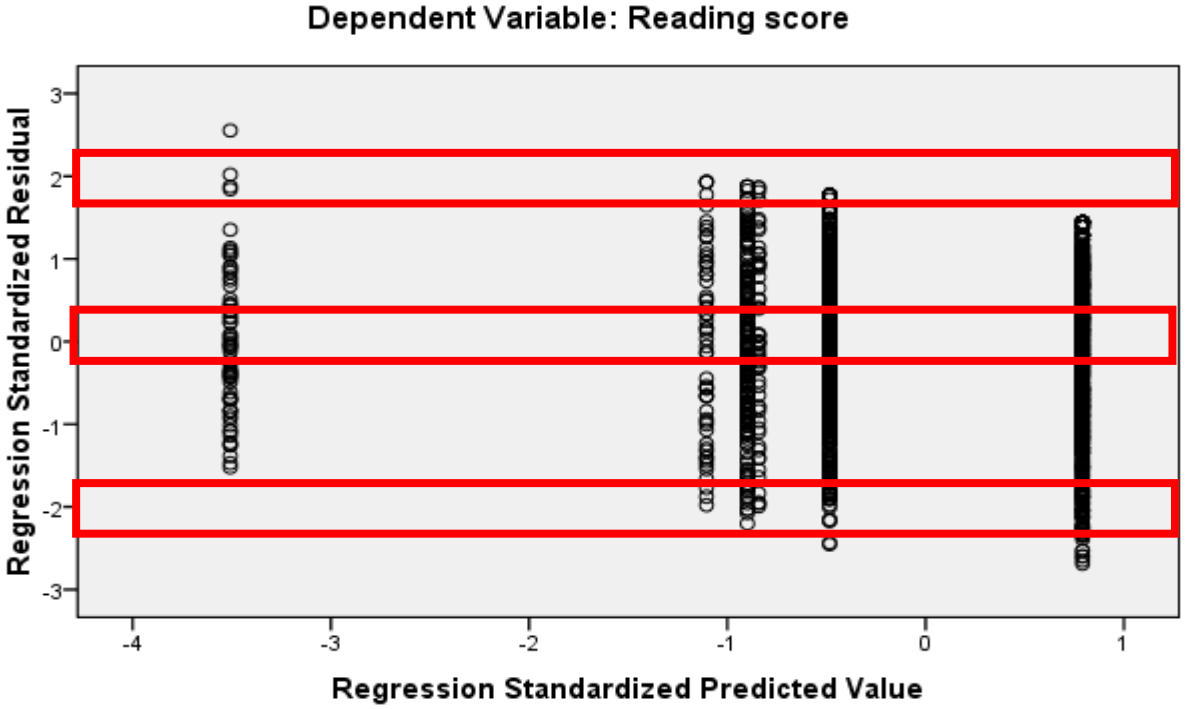
Normality: For our lowest prediction, the conditional distribution is positively skewed. For our highest predictions, the conditional distributions are negatively skewed. This is related, at least in part, to the ceiling effect of the test.

Linearity: No horseshoe, no problem.

Outliers: No outliers appear to be driving the conclusion.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Read
  /METHOD=ENTER Latino LowSES HighSES LowSESXLatino HighSESXLatino
  /SCATTERPLOT=(*ZRESID ,*ZPRED).
```

Residual vs. Fitted Plot



We are underpredicting badly for these students.

We are predicting nearly perfectly for these students.

We are overpredicting badly for these students.

Answering our Roadmap Question (Regression Perspective)

Unit 10: In the population, is there a relationship between reading and race controlling for free lunch?

$$Reading = \beta_0 + \beta_1 FreeLunch + \beta_2 Asian + \beta_3 Latino + \beta_4 Black + \beta_5 Free*Asian + \beta_6 Free*Latino + \beta_7 Free*Black + \varepsilon$$

$$Reading = 49.4 - 3.9 FreeLunch + 1.5 Asian - 3.3 Latino - 3.4 Black - 2.5 Free*Asian + -0.4 Free*Latino + -0.5 Free*Black + \varepsilon$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.314[a]	.099	.098	8.13954

a. Predictors: (Constant), FREELUNCHxBLACK, FREELUNCHxASIAN, FREELUNCHxLATINO, ASIAN, FREELUNCH, LATINO, BLACK

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	56485.879	7	8069.411	121.799	.000[a]
	Residual	516235.985	7792	66.252		
	Total	572721.864	7799			

a. Predictors: (Constant), FREELUNCHxBLACK, FREELUNCHxASIAN, FREELUNCHxLATINO, ASIAN, FREELUNCH, LATINO, BLACK

b. Dependent Variable: READING

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	49.439	.127		389.587	.000	49.191	49.688
	FREELUNCH	-3.882	.238	-.214	-16.293	.000	-4.349	-3.415
	ASIAN	1.491	.431	.043	3.461	.001	.646	2.335
	LATINO	-3.250	.426	-.119	-7.624	.000	-4.086	-2.415
	BLACK	-3.406	.504	-.112	-6.764	.000	-4.393	-2.419
	FREELUNCHxASIAN	-2.472	.865	-.037	-2.858	.004	-4.167	-.776
	FREELUNCHxLATINO	-.363	.606	-.010	-.600	.549	-1.550	.824
	FREELUNCHxBLACK	-.501	.678	-.013	-.739	.460	-1.829	.828

a. Dependent Variable: READING

In our nationally representative sample of 7,800 8th graders, there is a statistically significant relationship between reading achievement and our multiple predictors: race, SES and their interactions, $F(7, 7792) = 121.8$, $p < .001$. In our sample, among students who are eligible for free lunch, minority groups on average score lower than their White counterparts. Among students ineligible for free lunch, however, only Black and Latino students score lower on average than their White counterparts, and Asian students score higher on average. Our model predicts 10% of the variation in reading scores.

Unit 10: Research Question II (ANOVA Perspective)

Theory: Talking about a Latino/Anglo reading gap is an oversimplification. When we look across socioeconomic strata, we expect a gap; however, when we look within socioeconomic strata we expect the gaps to differ. Because risk factors are multiplicative, we expect greater Latino/Anglo reading gaps within lower socioeconomic strata. We expect this to hold true of 4-year-college bound boys.

Research Question: In the population of U.S. 4-year-college bound Latino and Anglo boys, is there an interaction between SES and ethnicity such that Anglo/Latino differences in reading scores are greatest for low SES students and least for high SES students?

*** Note that this is a different theory and research question from Question I. It need not be. I could have switched them, or I could have picked either and used it for both. Also note, my theory is wrong again! Ah, well.**

*** Our design is a 2x3 factorial design. We have two factors, Latino and SocioeconomicStatus. Latino has 2 levels—Anglo and Latino. SocioeconomicStatus has 3 levels—High, Mid, Low. If we added a third factor, Male, we would have a 2x3x2 factorial design. Please stop me if you don't see why.**

Unit 10: Research Question II (ANOVA Perspective)

Data Set: NELSBoys.sav National Education Longitudinal Survey (1988), a subsample of 1820 four-year-college bound boys, of whom 182 are Latino and the rest are Anglo.

Variables:

Outcome—Reading Achievement Score (*READ*)

Predictors—Latino = 1, Anglo = 0 (*LATINO*)

—Low SES=1, Mid SES=2, High SES=3 (*SocioeconomicStatus*)

ANOVA Model:

$$\text{READ}_{ijk} = \mu + \text{Latino}_i + \text{SocioeconomicStatus}_j + (\text{Latino} * \text{SocioeconomicStatus})_{ij} + \varepsilon_{ijk}$$

READ_{ijk} = The reading score of the *k*th student within the *ij*th group

μ = The grand mean.

Latino_i = The main effect of the "Latino" factor with two levels (i = 0,1)

$\text{SocioeconomicStatus}_j$ = The main effect of the SocioeconomicStatus factor with three levels (j = 1,2,3)

$(\text{Latino}*\text{SocioeconomicStatus})_{ij}$ = The interaction effect.

$\boldsymbol{\varepsilon}_{ijk}$ = The error associated with the *k*th student within the *ij*th group.

NELSBoys.sav After Creating Dummies and Crossproducts

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

1 : ID 6898594 Visible: 2

	ID	Read	Black	Latino	SocioEconomicStatus	LowSES	HighSES	LowSESXLatino	HighSESXLatino
1	6898594.00	37.57	0.00	1.00	1.00	1.00	0.00	1.00	0.00
2	2582532.00	65.29	0.00	1.00	3.00	0.00	1.00	0.00	1.00
3	2564477.00	46.71	0.00	1.00	1.00	1.00	0.00	1.00	0.00
4	7957674.00	48.16	0.00	1.00	3.00	0.00	1.00	0.00	1.00
5	7955910.00	66.19	0.00	1.00	3.00	0.00	1.00	0.00	1.00

Note that for ANOVA, we do not need to create dummies or crossproducts. Yippy skippy!

Sifting Through ANOVA Output

Univariate Analysis of Variance

Between-Subjects Factors

		Value Label	N
SocioEconomicStatus	1	Low SES	264
	2	Mid SES	539
	3	High SES	1017
Latino	0		1638
	1		182

```
UNIANOVA Read BY SocioEconomicStatus Latino
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PLOT=PROFILE(SocioEconomicStatus*Latino Latino*SocioEconomicStatus)
  /CRITERIA=ALPHA(0.05)
  /DESIGN=SocioEconomicStatus Latino SocioEconomicStatus*Latino.
```

Ignore the lines for intercept and total. Everybody else does.

We covered Corrected Model, Error and Corrected Total (by other names) in Unit 5 and Unit 9.

Tests of Between-Subjects Effects

Dependent Variable:Reading score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	10067.229[a]	5	2013.446	24.492	.000
Intercept	1761374.230	1	1761374.230	21425.377	.000
SocioEconomicStatus	2592.035	2	1296.017	15.765	.000
Latino	2255.099	1	2255.099	27.431	.000
SocioEconomicStatus * Latino	745.715	2	372.858	4.535	.011
Error	149128.430	1814	82.210		
Total	5795063.720	1820			
Corrected Total	159195.659	1819			

a. R Squared = .063 (Adjusted R Squared = .061)

Trivia : What is $\frac{10067.229}{159195.659}$?

Sifting Through ANOVA Output

Univariate Analysis of Variance

Between-Subjects Factors

		Value Label	N
SocioEconomicStatus	1	Low SES	264
	2	Mid SES	539
	3	High SES	1017
Latino	0		1638
	1		182

We can break down the corrected model sum of squares into its component sums of squares.

After glancing at the omnibus F-test to see if anything is happening, look at the interaction F-test.

Then look at the F-tests associated with the main effects.

Tests of Between-Subjects Effects

Dependent Variable:Reading score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	10067.229[a]	5	2013.446	24.492	.000
Intercept	1761374.230	1	1761374.230	21425.377	.000
SocioEconomicStatus	2592.035	2	1296.017	15.765	.000
Latino	2255.099	1	2255.099	27.431	.000
SocioEconomicStatus * Latino	745.715	2	372.858	4.535	.011
Error	149128.430	1814	82.210		
Total	5795063.720	1820			
Corrected Total	159195.659	1819			

a. R Squared = .063 (Adjusted R Squared = .061)

Sifting Through ANOVA Output

A stat sig interaction tells us that the effect of one factor varies by the levels of another factor (enough to warrant an inference from the sample to the population).

Alert! "Effect" here has nothing to do with cause and effect.

There is a statistically significant interaction such that the relationship between SES and reading differs for Latinos and Anglos, F(2,1814) = 4.535, p = 0.011.

OR

There is a statistically significant interaction such that the relationship between ethnicity and reading differs by level of SES, F(2,1814) = 4.535, p = 0.011.

Always interpret the interaction first! If it's stat sig, the main effects are less important.

Dependent Variable:Reading score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	10067.229[a]	5	2013.446	24.492	.000
Intercept	1761374.230	1	1761374.230	21425.377	.000
SocioEconomicStatus	2592.035	2	1296.017	15.765	.000
Latino	2255.099	1	2255.099	27.431	.000
SocioEconomicStatus * Latino	745.715	2	372.858	4.535	.011
Error	149128.430	1814	82.210		
Total	5795063.720	1820			
Corrected Total	159195.659	1819			

a. R Squared = .063 (Adjusted R Squared = .061)

Sifting Through ANOVA Output

A stat sig main effect tells us that the averages within levels of the factor differ from the mean (enough to warrant an inference from the sample to the population).

There is a statistically significant main effect of SES such that average reading ability differs for students of low, medium, and high SES, $F(2,1814) = 15.765$, $p < 0.001$. On average, students from higher socioeconomic strata tend to read better. (I peeked at the graph to get the second sentence!)

There is a statistically significant main effect of ethnicity such that Anglo students, on average, tend to read better than Latino students, $F(1,1814) = 27.431$, $p < 0.001$. (I peeked at the graph!)

We need graphs, planned contrasts and/or post hoc comparisons to explore the relationships more deeply.

Dependent Variable:Reading score

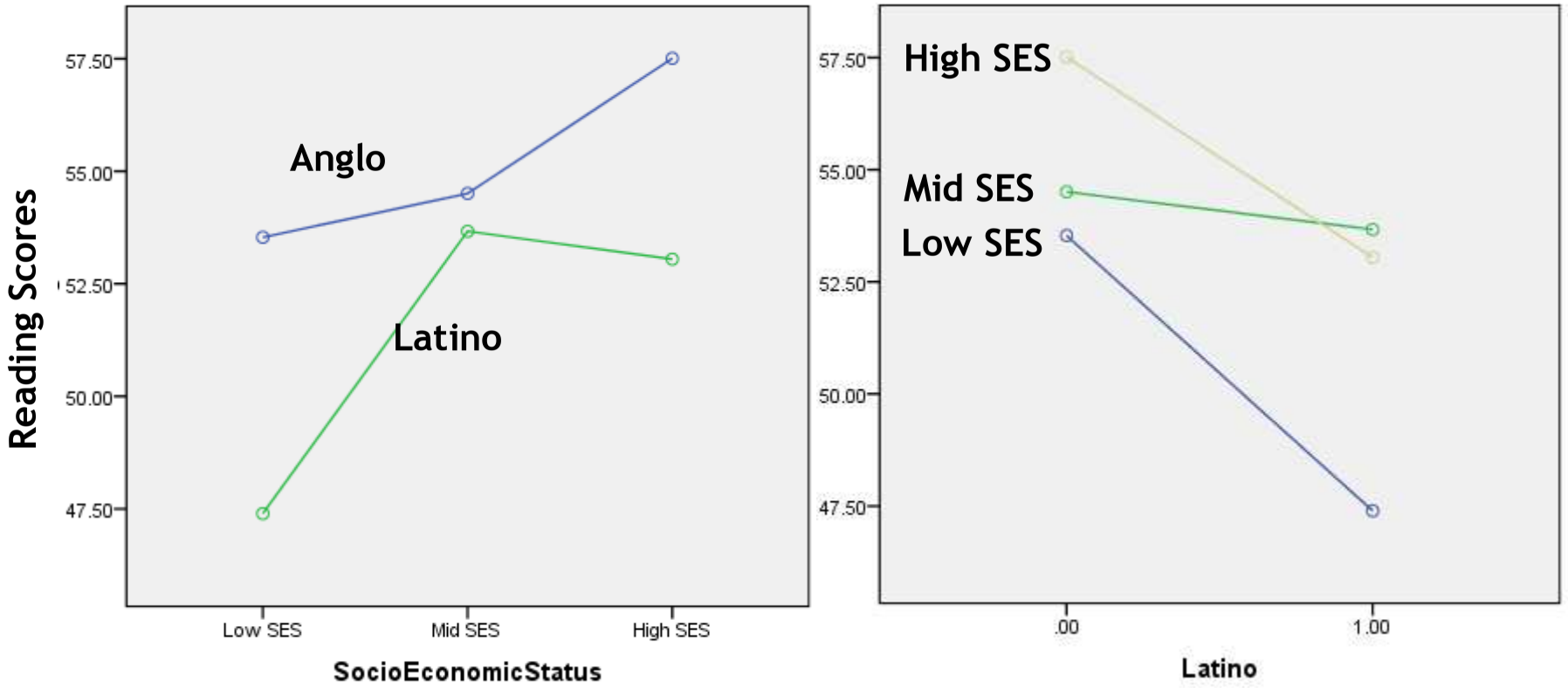
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	10067.229[a]	5	2013.446	24.492	.000
Intercept	1761374.230	1	1761374.230	21425.377	.000
SocioEconomicStatus	2592.035	2	1296.017	15.765	.000
Latino	2255.099	1	2255.099	27.431	.000
SocioEconomicStatus * Latino	745.715	2	372.858	4.535	.011
Error	149128.430	1814	82.210		
Total	5796063.720	1820			
Corrected Total	159195.659	1819			

a. R Squared = .063 (Adjusted R Squared = .061)

Two Snapshots of the Same Thing

See the interaction. The lines are not parallel.

Trick question: Is the interaction ordinal or disordinal?



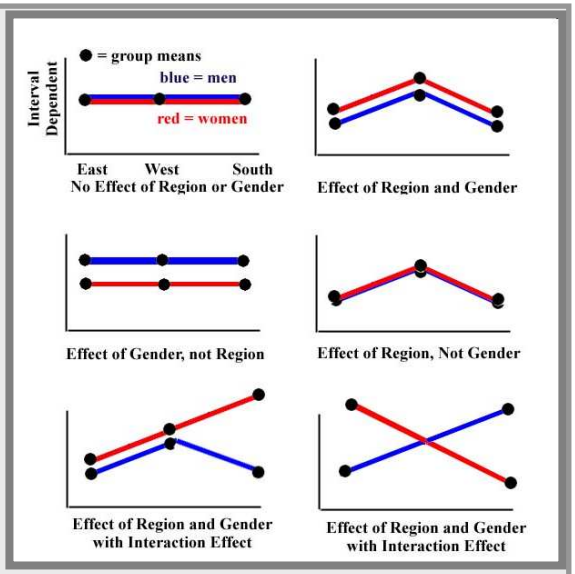
See the main effects. The average Anglo student performs differently from the average Latino student. The average High SES student differs from the average Mid SES student who, in turn, differs from the average Low SES student. (You only need one difference to make a main effect stat sig, so you must consult graphs, planned contrasts and/or post hoc comparisons to find where the action is.)

Interpreting Two-Way ANOVAS

Interpret your omnibus F-test.
Interpret your interaction F-test.
Interpret your main effects F-tests.

*** Conduct your analysis in this order, although you may want to report in a different order.**

http://www.psm.upm.edu.my/5950/StatisticsNotes/PA%20765%20ANOVA.htm



Make sense of the graph in real-world terms.

- Consider which mean differences might be statistically significant.
- Consider which interactions (non-parallelisms) might be statistically significant.

In the future, use planned contrasts and/or post hoc comparisons to dig deeper.

This is the key to Post Hole 10. Practice is in back.

Calculating a Two-Way ANOVA by Hand

Group Size: 25 MSE: 25

Enter data below or drag the points in the graph to change their values:

	Lo SES	Mid SES	Hi SES	Marginal Mean
Anglo	53.5	54.5	57.5	55.17
Latino	47.4	53.7	53.1	51.40
Marginal Mean	50.45	54.10	55.30	53.28

Source	SSQ	df	MS	F	p
A	638.08	2	319.04	12.76	0.00
B	532.04	1	532.04	21.28	0.00
AB	183.08	2	91.54	3.66	0.03
Error	3600.00	144	25.00		
Total	4953.21	149			

Six groups of 25 each.

Each group has an average reading score.

Each group has variation around its average reading score, which can be measured by taking the mean square error. This is "bad" variation.

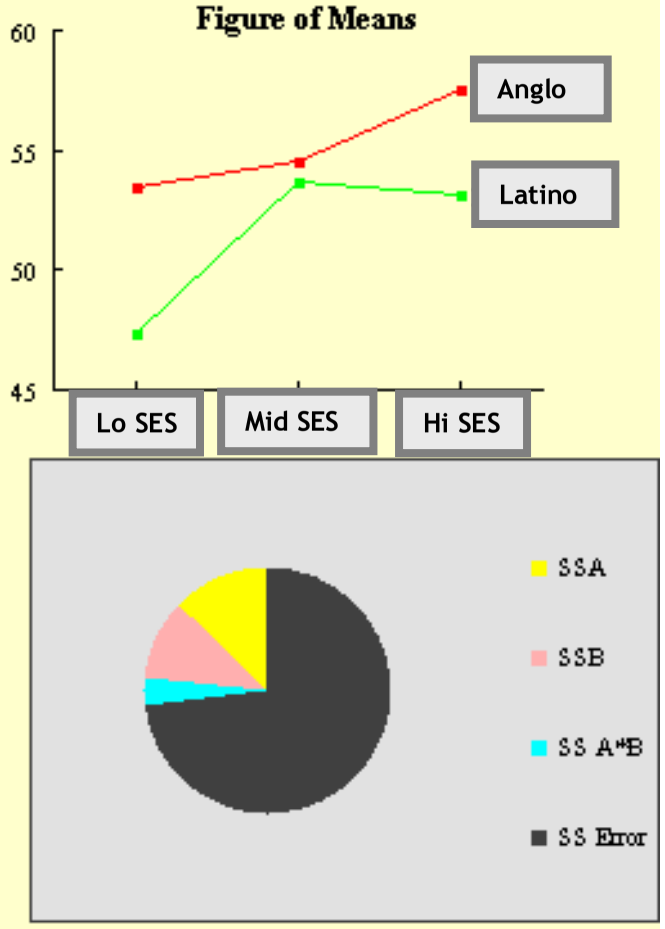
Notice the means in the margins; these are "marginal means."

Notice the grand mean.

With the above information, we can calculate by hand the ANOVA table…

Default Means Clear All

You can draw the mean positions on the figure below to change the mean values:



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

Calculating a Two-Way ANOVA by Hand

Group Size: 25 MSE: 25

Enter data below or drag the points in the graph to change their values:

	A1	A2	A3	Marginal Mean
B1	53.5	54.5	57.5	55.17
B2	47.4	53.7	53.1	51.40
Marginal Mean	50.45	54.10	55.30	53.28

Source	SSQ	df	MS	F	p
A	638.08	2	319.04	12.76	0.00
B	532.04	1	532.04	21.28	0.00
AB	183.08	2	91.54	3.66	0.03
Error	3600.00	144	25.00		
Total	4953.21	149			

Before we begin, let's think backwards. It will help to root for statistically significant results.

We want a small p-value, therefore we want a big F-value.

The F-value is the mean square (for the main effect or interaction) divided by the mean square error.* You can think of the F-value as the ratio of good mean squares to bad mean squares. Some people think of it as the ratio of signal to noise.

$$F_{Latino} = \frac{MS_{Latino}}{MS_{Error}} = \frac{Good}{Bad} = \frac{Signal}{Noise} = \frac{WantBig}{WantSmall}$$

Mean squares are sums of squares divided by degrees of freedom (df).

*This is true for fixed effects models. Random effects models are different, but we are not going to go there. Just kind of have in the back of your head that there's a funky sort of ANOVA called "random effects ANOVA" that you have not studied.

Calculating a Two-Way ANOVA by Hand

Group Size: 25 MSE: 25

Enter data below or drag the points in the graph to change their values:

	A1	A2	A3	Marginal Mean
B1	53.5	54.5	57.5	55.17
B2	47.4	53.7	53.1	51.40
Marginal Mean	50.45	54.10	55.30	53.28

Source	SSQ	df	MS	F	p
A	638.08	2	319.04	12.76	0.00
B	532.04	1	532.04	21.28	0.00
AB	183.08	2	91.54	3.66	0.03
Error	3600.00	144	25.00		
Total	4953.21	149			

Now, let's think forwards.

We can calculate the sum of squares for each main effect. We will continue looking at the main effect of ethnicity.

Each of the 75 Anglo students has a "good" square: $(55.17-53.28)^2$.

Each of the 75 Latino students has a "good" square: $(51.40-53.28)^2$.

We add up all 150 squares to get $SS_{Latino}=532.04$.

To get the SS_{error} we would calculate the squared deviation from the group mean (there are six groups) for each student. These squared deviations are "bad" squares. We would add the squared errors to get the SS_{error}, but we need the observed scores to do the math, and this little applet does not provide that information.

The $SS_{SocioEconomicStatus}$ is basically the same. Each of the 50 low SES students has a good square: $(55.17-53.28)^2$…

Calculating a Two-Way ANOVA by Hand

Group Size: 25 MSE: 25

Enter data below or drag the points in the graph to change their values:

	A1	A2	A3	Marginal Mean
B1	53.5	54.5	57.5	55.17
B2	47.4	53.7	53.1	51.40
Marginal Mean	50.45	54.10	55.30	53.28

Source	SSQ	df	MS	F	p
A	638.08	2	319.04	12.76	0.00
B	532.04	1	532.04	21.28	0.00
AB	183.08	2	91.54	3.66	0.03
Error	3600.00	144	25.00		
Total	4953.21	149			

We can calculate the sum of squares for the interaction. The key to an interaction is that it is what's not predicted by the main effects. We have six predictions (i.e., group means). A main effect is a deviation between the marginal mean and the grand mean. We will subtract out the main effects from the group mean before we square the group mean deviation from the grand mean. We will use mid SES Latino students as an example

For the 25 mid SES Latino students:

Main effect of being Latino = (51.40-53.28).

Main effect of being mid SES = (54.10-53.28).

Each of the 25 mid SES Latino has a "good" square equal to 2.19: $(53.7-(51.40-53.28)-(54.10-53.28)-53.28)^2$.

183.08 is the sum of the 150 "good" squares.

Note: When the main effects (i.e., marginal means) alone perfectly predict the group means, there is no interaction. Here, however, we find an interaction.

Calculating a Two-Way ANOVA by Hand

Group Size: 25 MSE: 25

Enter data below or drag the points in the graph to change their values:

	A1	A2	A3	Marginal Mean
B1	53.5	54.5	57.5	55.17
B2	47.4	53.7	53.1	51.40
Marginal Mean	50.45	54.10	55.30	53.28

Source	SSQ	df	MS	F	p
A	638.08	2	319.04	12.76	0.00
B	532.04 ÷	1 =	532.04	21.28	0.00
AB	183.08	2	91.54	3.66	0.03
Error	3600.00 ÷	144 =	25.00		
Total	4953.21	149			

In order to get our mean squares, we divide by the degrees of freedom.

We want to divide our "bad" sum of squares (SS_{error}) by a big number. We get our wish when we have a large sample, because we divide by (basically) the sample size (really the degrees of freedom or "df" for short).

We want to divide a "good" sum of squares (e.g., SS_{Latino}) by a small number. We divide by (basically) the number of levels of our factor, so we want to keep the levels as few as possible by excluding non-predictive variables from our model. I think of this step as a penalty for crappy variables.

Calculating a Two-Way ANOVA by Hand

Group Size: 25 MSE: 25

Enter data below or drag the points in the graph to change their values:

	A1	A2	A3	Marginal Mean
B1	53.5	54.5	57.5	55.17
B2	47.4	53.7	53.1	51.40
Marginal Mean	50.45	54.10	55.30	53.28

Source	SSQ	df	MS	F	p
A	638.08	2	319.04	12.76	0.00
B	532.04	1	532.04	21.28	0.00
AB	183.08	2	91.54	3.66	0.03
Error	3600.00	144	25.00		
Total	4953.21	149			

As you may note, we are coming full circle.

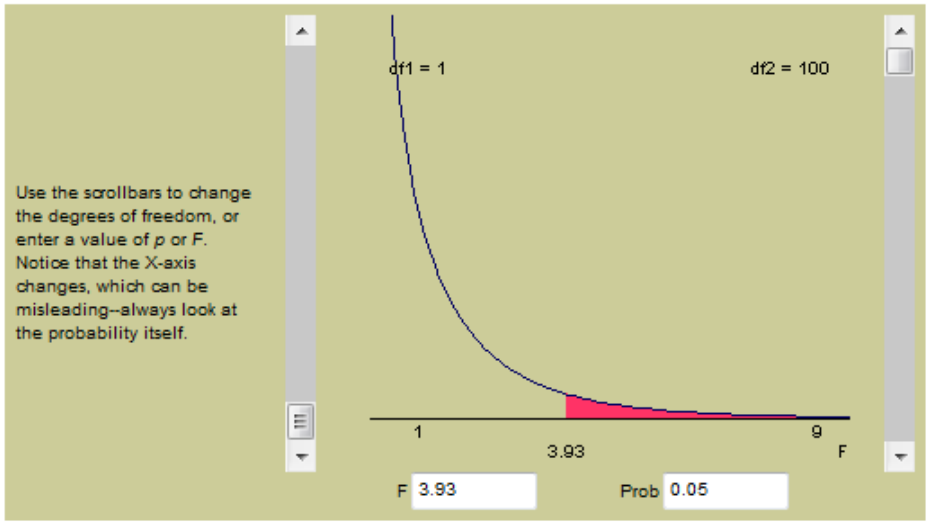
The F-value is the mean square (for the main effect or interaction) divided by the mean square error.* You can think of the F-value as the ratio of good mean squares to bad mean squares. Some people think of it as the ratio of signal to noise.

$$F_{Latino} = \frac{MS_{Latino}}{MS_{Error}} = \frac{Good}{Bad} = \frac{Signal}{Noise} = \frac{WantBig}{WantSmall}$$

Once we have our F-statistic, we consult an F-distribution. An F-distribution is a theoretical sampling distribution derived from the Central Limit Theorem, closely related to the t-distribution. If our F-statistic is far enough away from zero, we reject the null hypothesis that there is no relationship in the population.

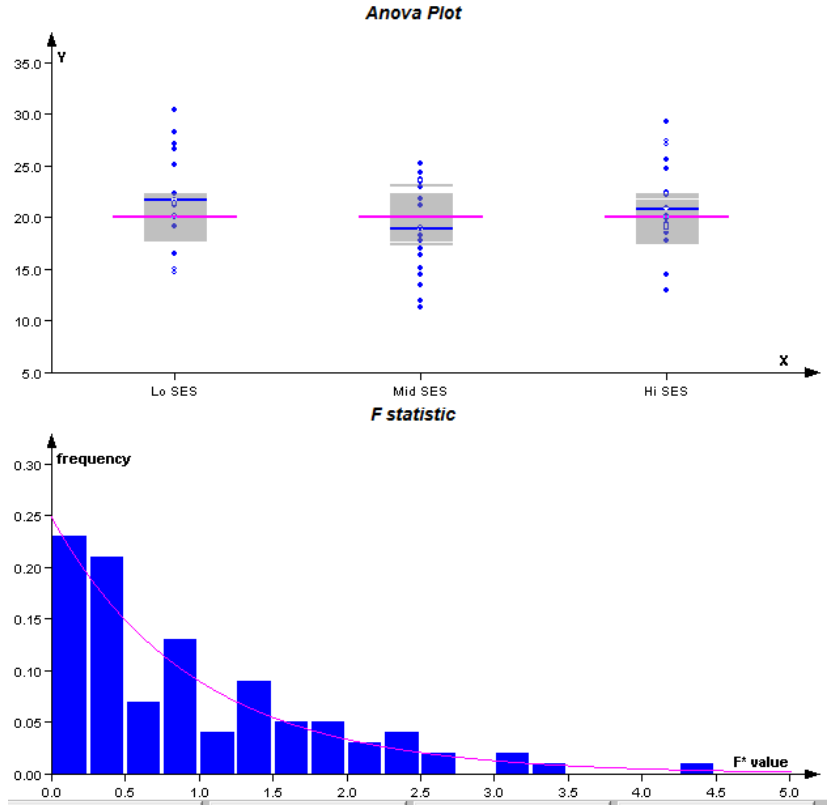
Based on our p value of less than 0.05, we reject the null hypothesis and conclude there is a main effect of ethnicity in the population. If there were no relationship in the population, it is very unlikely we would randomly draw a sample with an F-statistic as extreme or more extreme than 21.28.

The F-Distribution



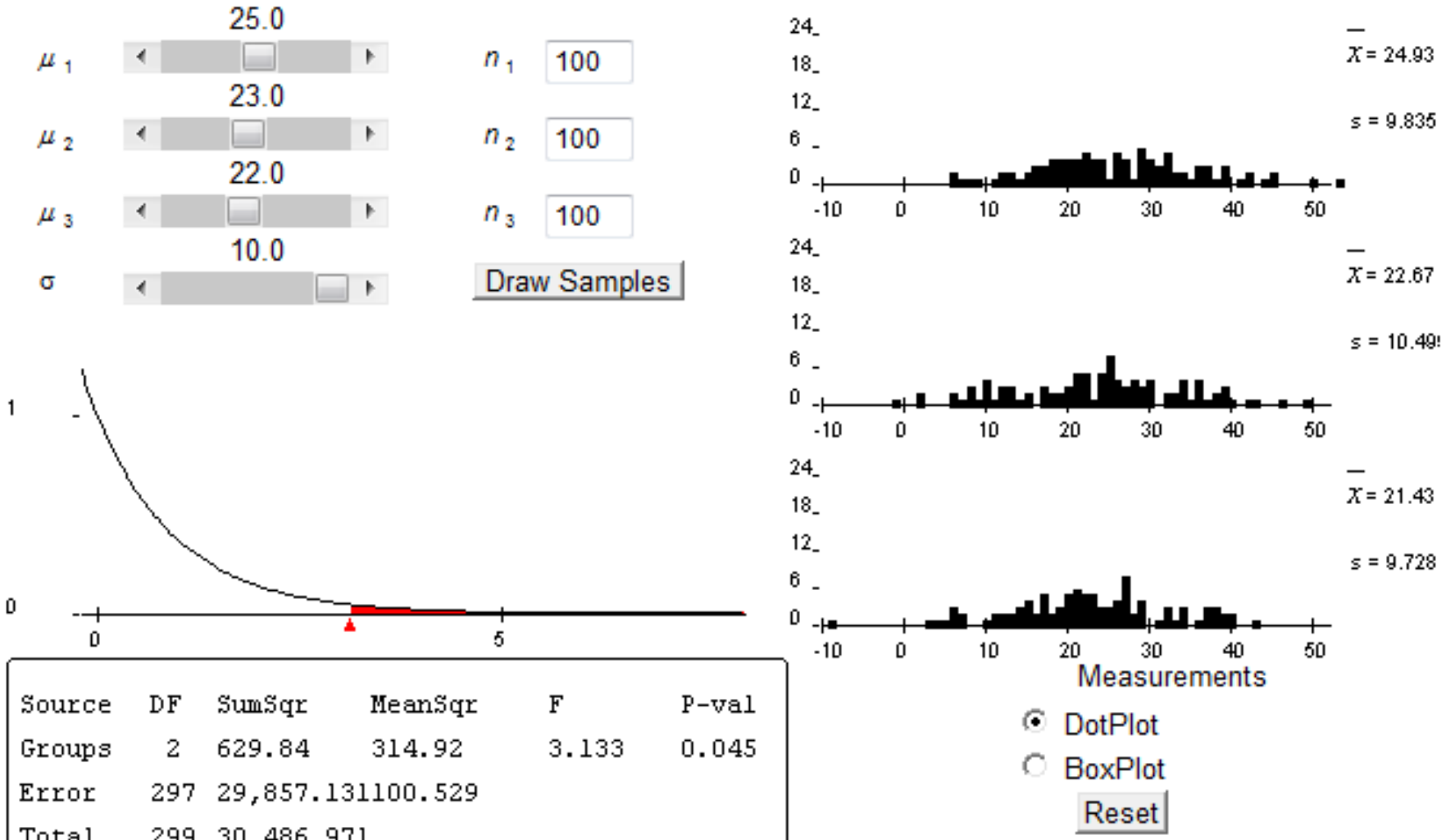
http://www.uvm.edu/~dhowell/SeeingStatisticsApplets/FProb.html

The F-distribution is a sampling distribution derived from the Central Limit Theorem. It takes different shapes depending on the degrees of freedom in the numerator and denominator. This is why we report the degrees of freedom whenever we report the F-statistic, $F(1,144)=21.28$, $p < 0.001$. The t-distribution also takes different shapes depending on degrees of freedom, but it never drastically diverges from normal, so it is not so important to report the degrees of freedom.



http://serc.carleton.edu/sp/cause/interactive/examples/17734.html

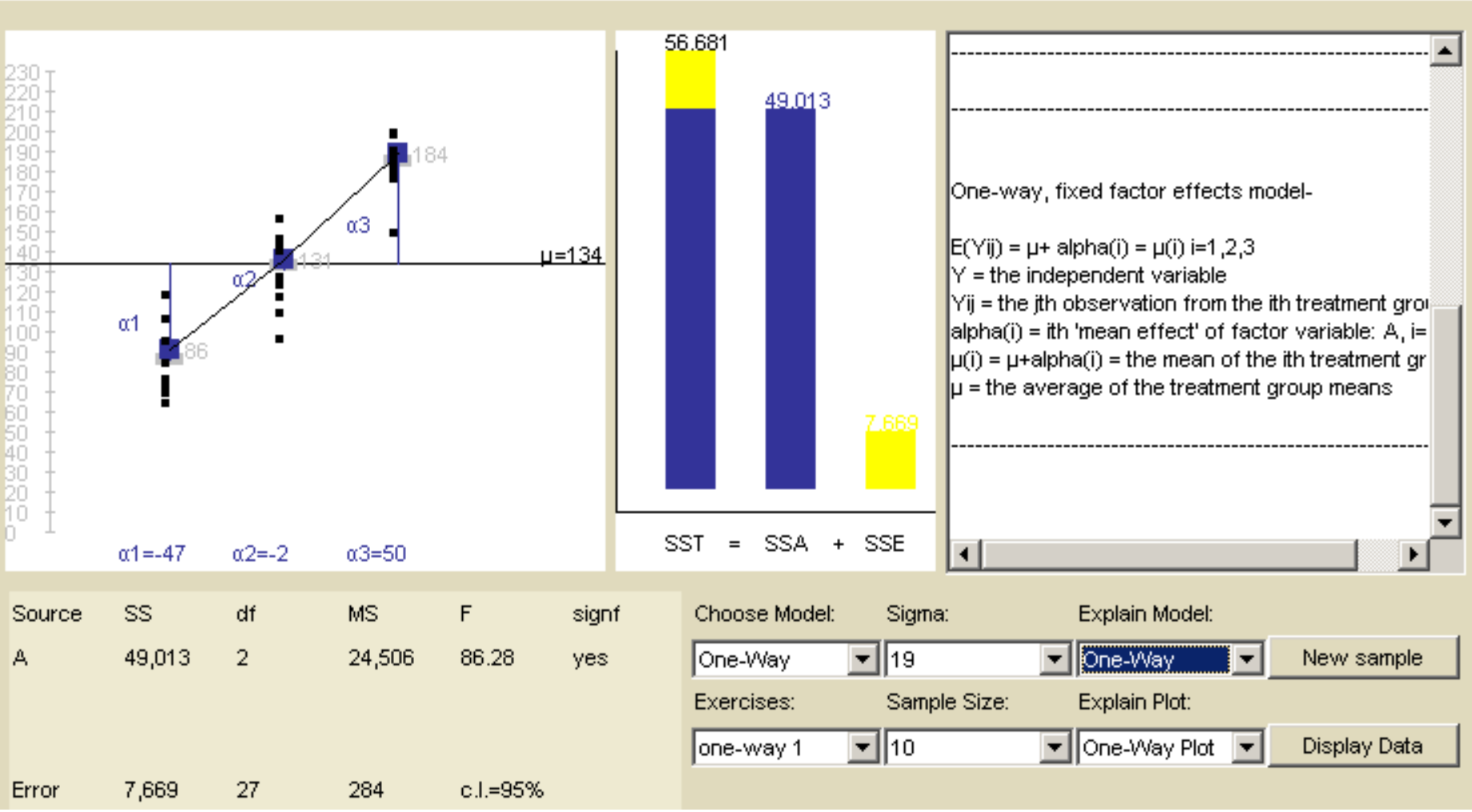
One-Way ANOVA (Horizontal Groups)



Source	DF	SumSqr	MeanSqr	F	P-val
Groups	2	629.84	314.92	3.133	0.045
Error	297	29,857.131	100.529		
Total	299	30,486.971			

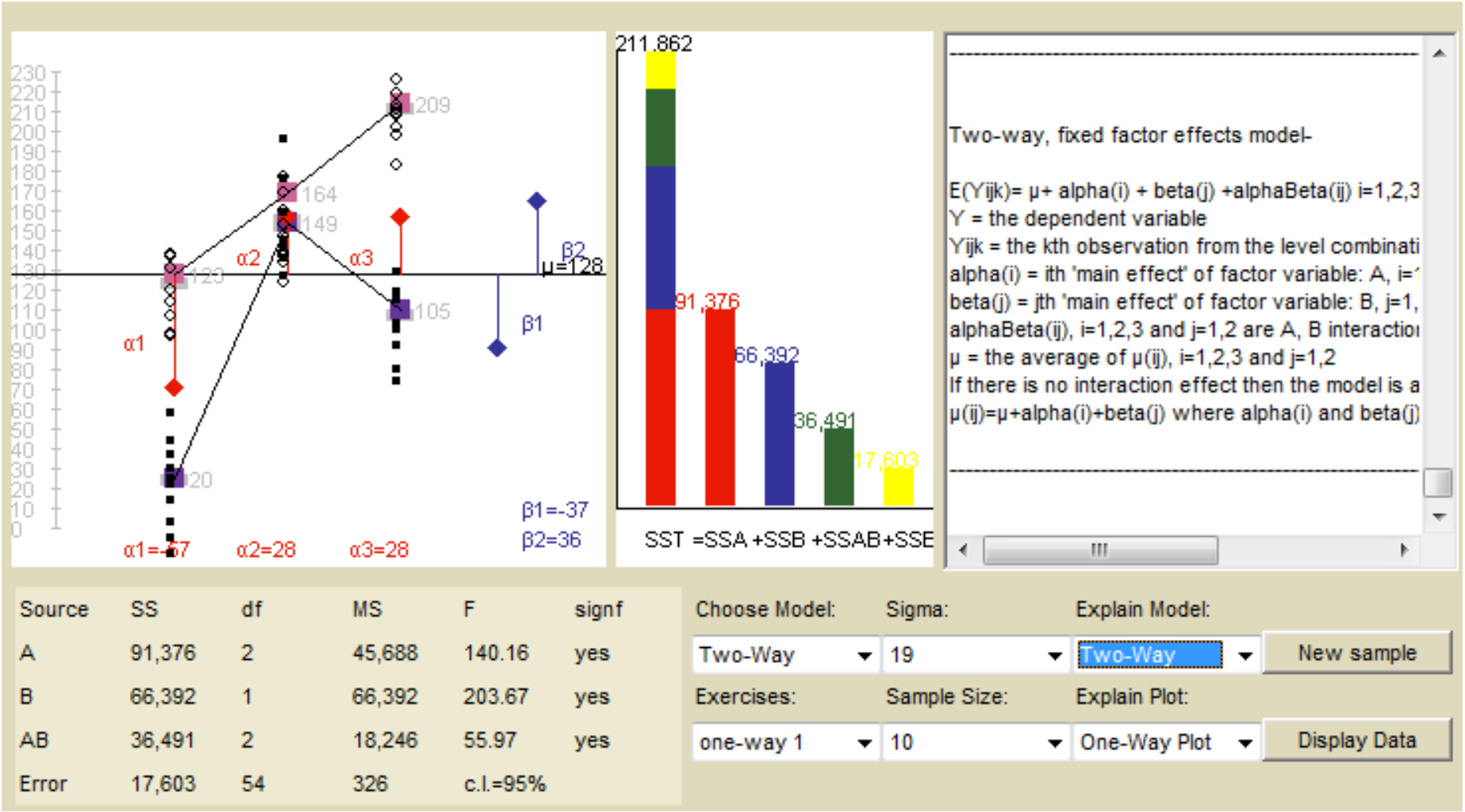
http://www.rossmanchance.com/applets/Anova/Anova.html

One-Way ANOVA (Vertical Groups)



http://www.kingsborough.edu/academicDepartments/math/faculty/rsturm/anova/Anova0126.html

Two-Way ANOVA (Vertical Groups)



Source	SS	df	MS	F	signf
A	91,376	2	45,688	140.16	yes
B	66,392	1	66,392	203.67	yes
AB	36,491	2	18,246	55.97	yes
Error	17,603	54	326	c.l.=95%	

Choose Model: Two-Way | Sigma: 19 | Explain Model: Two-Way | New sample

Exercises: one-way 1 | Sample Size: 10 | Explain Plot: One-Way Plot | Display Data

http://www.kingsborough.edu/academicDepartments/math/faculty/rsturm/anova/Anova0126.html

Answering our Roadmap Question (ANOVA Perspective)

Unit 10: In the population, is there a relationship between reading and race controlling for free lunch?

Tests of Between-Subjects Effects

Dependent Variable:READING

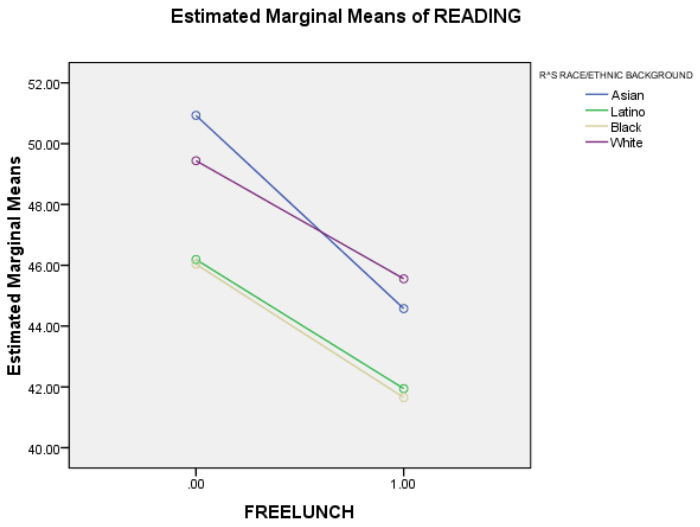
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	56485.879[a]	7	8069.411	121.799	.000
Intercept	6087049.770	1	6087049.770	91877.151	.000
RACE	14705.701	3	4901.900	73.989	.000
FREELUNCH	16141.887	1	16141.887	243.644	.000
RACE * FREELUNCH	559.039	3	186.346	2.813	.038
Error	516235.985	7792	66.252		
Total	1.817E7	7800			
Corrected Total	572721.864	7799			

a. R Squared = .099 (Adjusted R Squared = .098)

See the interaction in the graph. Not all the lines are parallel. It looks like the action is a disordinal interaction among White students and Asian students of varying SES.

See the main effect of race. The lines differ vertically with Black and Latino students scoring lower on average (disregarding SES) than White and Asian students.

See the main effect of SES. The lines are sloping downward indicating that students eligible for free lunch score lower on average (disregarding race) than their ineligible counterparts.



In our nationally representative sample of 7,800 8th graders, there is a statistically significant relationship between reading achievement and our multiple predictors: race, SES and their interactions, F (7, 7792) = 121.8, p < .001. There is a statistically significant interaction such that the relationship between reading achievement and race differs by level of SES, F(3, 7792) = 2.8, p = 0.038. There are also statistically significant main effects of race and SES, F(3, 7792) = 74.0, p < 0.001 and F(1, 7792) = 243.6, p < 0.001, respectively. On average, students eligible for free lunch score lower than ineligible students, and Black and Latino students score lower than Asian and White students. There is a disordinal interaction between race and SES for White and Asian students such that among students eligible for free lunch, Asian students score lower on average than White students, whereas the opposite is true for ineligible students.

Unit 10 Appendix: Key Concepts

Review: Two variables are uncorrelated if, and only if, knowing one does not help you predict the other.

Interaction is about three or more variables: one outcome and at least two predictors.

Unit 10 Appendix: Key Concepts

When we talk about "effect sizes," "main effects" and "effects," we are not implying cause and effect. It's really just unfortunate statistical nomenclature.

Predictions are more informative the more they differ from the mean. I call it "added value."

A two-way ANOVA breaks down the corrected model sum of squares into its component sums of squares.

Attacking a Two-Way ANOVA table:

(1) After glancing at the omnibus F-test to see if anything is going on, look at the interaction F-test.
(2) Then look at the F-tests associated with the main effects.

Always interpret the interaction first! If it's stat sig, the main effects are less important. We need graphs, planned contrasts and/or post hoc comparisons to explore the relationships more deeply.

The F-distribution is a sampling distribution derived from the Central Limit Theorem. It takes different shapes depending on the degrees of freedom in the numerator and denominator. This is why we report the degrees of freedom whenever we report the F-statistic, $F(1,144)=21.28$, $p < 0.001$. The t-distribution also takes different shapes depending on degrees of freedom but it never drastically diverges from normal, so it is not so important to report the degrees of freedom.

Unit 10 Appendix: Key Interpretations (Regression)

The Anglo/Latino reading gap differs by socioeconomic status. There appears to be little or no gap for four-year-college bound boys of middle SES. However, there are large gaps of 6.1 and 4.4 points for students of low and high SES, respectively. (Note, we can also write about how the SES/Reading relationship differs for Anglo students and Latino students.)

There is a statistically significant relationship between our outcome and our predictors, $F(5,1814)=24.492$, $p < 0.05$. Ethnicity and socioeconomic status (and their interaction) predict about 6% of the variation in reading scores.

Homoscedasticity: The variances are roughly equal for each prediction.
Independence: We cannot tell if the students are clustered in, for example, schools.
Normality: For our lowest prediction, the conditional distribution is positively skewed. For our highest predictions, the conditional distributions are negatively skewed. This is related, at least in part, to the ceiling effect of the test.
Linearity: No horseshoe, no problem.
Outliers: No outliers appear to be driving the conclusion.

Unit 10 Appendix: Key Interpretations (ANOVA)

Interaction:

There is a statistically significant interaction such that the relationship between SES and reading differs for Latinos and Anglos, $F(2,1814) = 4.535$, $p = 0.011$.

OR

There is a statistically significant interaction such that the relationship between ethnicity and reading differs by level of SES, $F(2,1814) = 4.535$, $p = 0.011$.

Main Effects:

There is a statistically significant main effect of SES such that average reading ability differs for students of low, medium, and high SES, $F(2,1814) = 15.765$, $p < 0.001$. On average, students from higher socioeconomic strata tend to read better. (I peeked at the graph to get the second sentence!)

There is a statistically significant main effect of ethnicity such that Anglo students, on average, tend to read better than Latino students, $F(1,1814) = 27.431$, $p < 0.001$. (I peeked at the graph.)

Unit 10 Appendix: Key Terminology (Regression)

See Slide 19 for:

Statistical Interaction

Statistical Control

Residual vs. Fitted Scatterplot

Interactions appear in plots:

- **When the lines are parallel, there is no interaction.**
- **When the lines are non-parallel there is an interaction.**
- **When the non-parallel lines do not cross, the interaction is ordinal.**
- **When the non-parallel lines cross, the interaction is disordinal.**

A main effects model has no interaction terms. The trend lines are constrained to be parallel.

Unit 10 Appendix: Key Terminology (ANOVA)

Our design is a 2x3 factorial design. We have two factors, Latino and SocioeconomicStatus. Latino has 2 levels—Anglo and Latino. SocioeconomicStatus has 3 levels—High, Mid, Low. If we added a third factor, Male, we would have a 2x3x2 factorial design. See the pattern?

A stat sig interaction tells us that the effect of one factor varies by the levels of another factor (enough to warrant an inference from the sample to the population).

A stat sig main effect tells us that the averages within levels of the factor differ from the mean (enough to warrant an inference from the sample to the population).

Unit 10 Appendix: SPSS Syntax (Regression)

```
*****************************************.
*Create interaction terms (or crossproduct terms).
*****************************************.

COMPUTE LowSESXLatino=(LowSES*Latino).
COMPUTE HighSESXLatino=(HighSES*Latino).
Execute.


*****************************************.
*Regress Read on Ethnicity and Socioeconomic Status (with interaction).
*****************************************.

REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Read
  /METHOD=ENTER Latino LowSES HighSES LowSESXLatino HighSESXLatino
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /SAVE SDRESID.
```

Unit 10 Appendix: SPSS Syntax (ANOVA)

```
*****************************************************************.
*Two-Way ANOVA.
*****************************************************************.
UNIANOVA Read BY SocioEconomicStatus Latino
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PLOT=PROFILE(SocioEconomicStatus*Latino Latino*SocioEconomicStatus)
  /CRITERIA=ALPHA(0.05)
  /DESIGN=SocioEconomicStatus Latino SocioEconomicStatus*Latino.
```

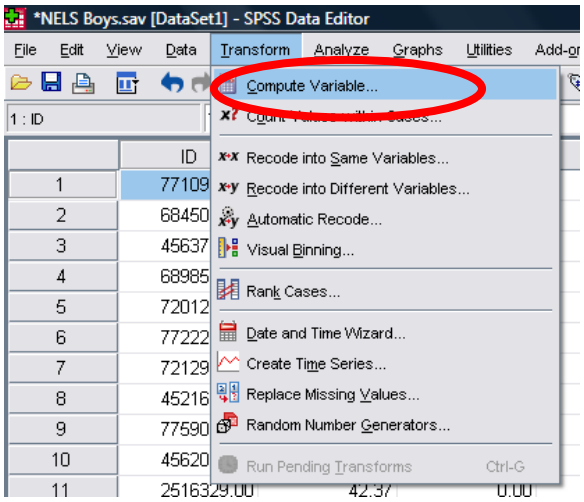
SPSS Menu Navigation

In addition to the dummies from Unit 9, you need to create interaction terms (i.e., crossproducts). You can do this easily through code or dropdown menus.

Code:

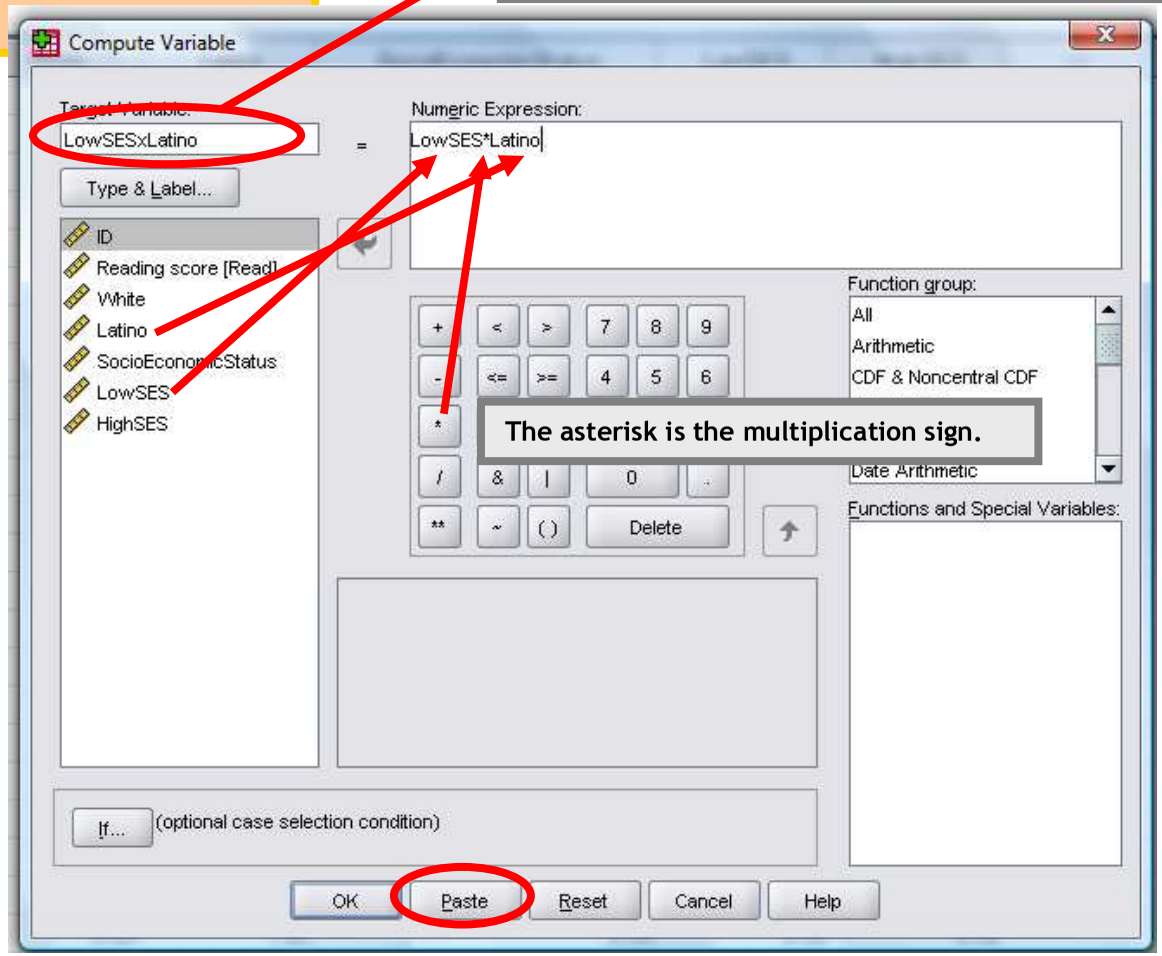
```
COMPUTE LowSESXLatino=(LowSES*Latino).
COMPUTE HighSESXLatino=(HighSES*Latino).
Execute.
```

Menus:



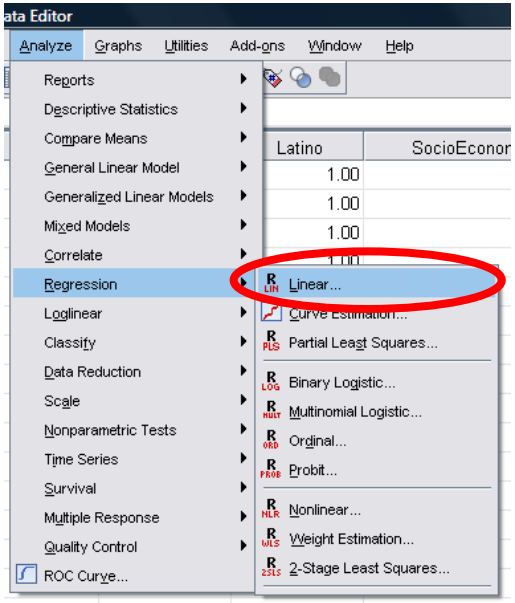
Go to Transform > Compute Variable…

With "Target Variable" you choose any name you want. I use the letter "x" to connect the two multipliers.



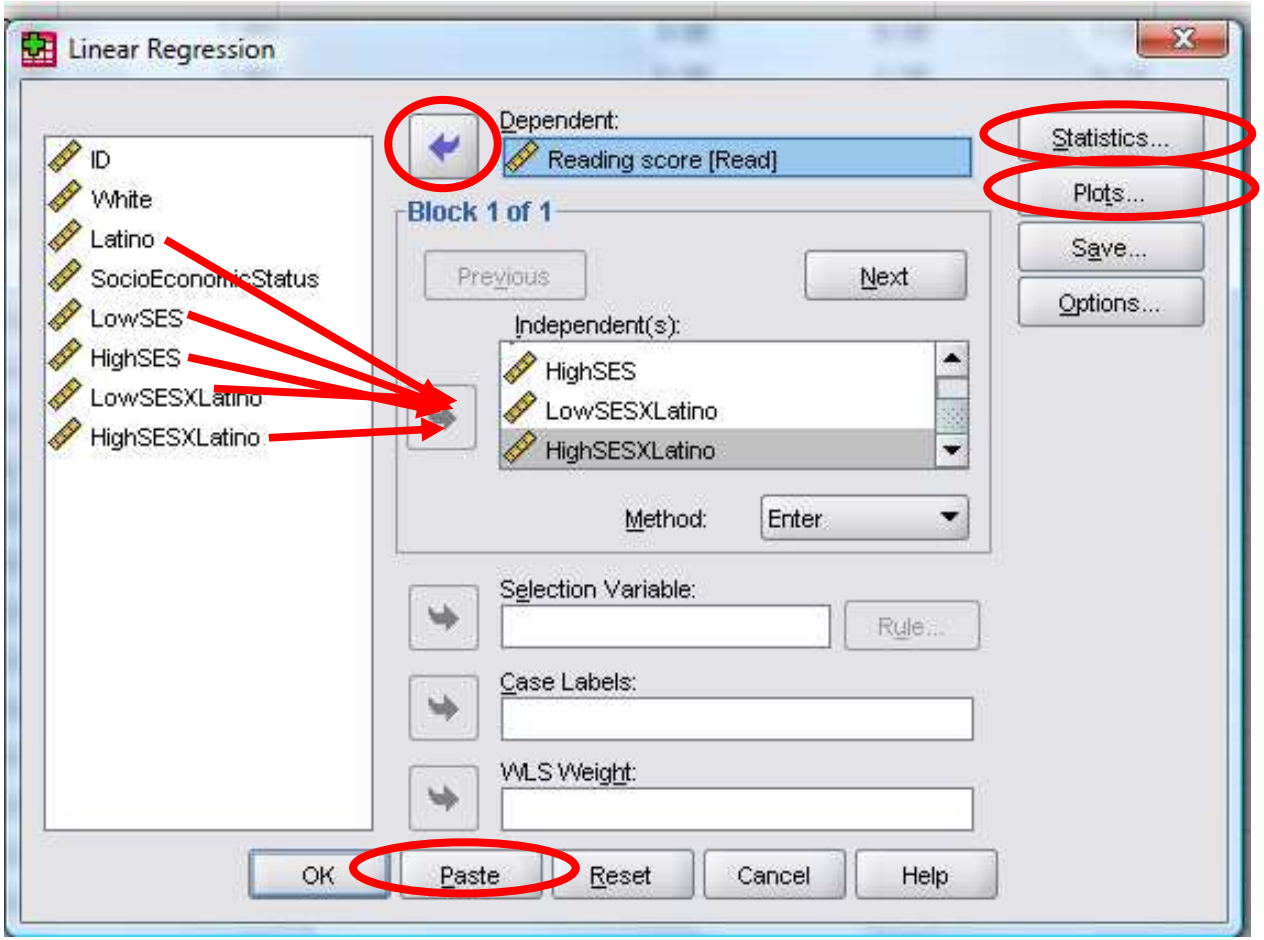
The asterisk is the multiplication sign.

SPSS Menu Navigation



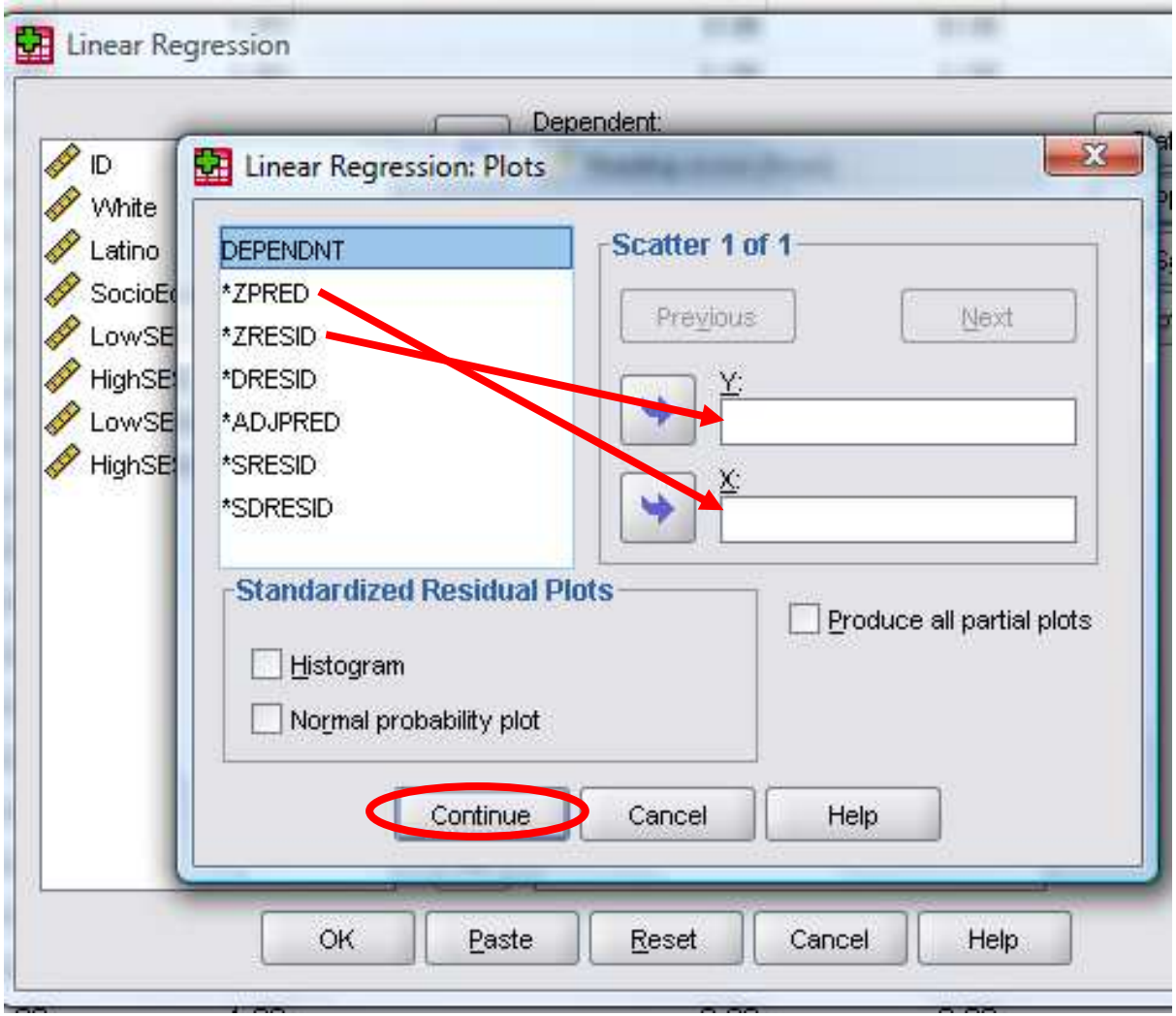
As you would with any regression:

Go to Analyze > Regression > Linear…



As usual, choose your outcome (i.e., dependent variable) and your predictor(s) (i.e., independent variable(s)): Note that you have five. As usual, got the "Statistics…" and choose confidence intervals. The <u>new step</u> is "Plots…" (next slide).

SPSS Menu Navigation



Create a residual vs. fitted plot (ZRESID vs. ZPRED plot).

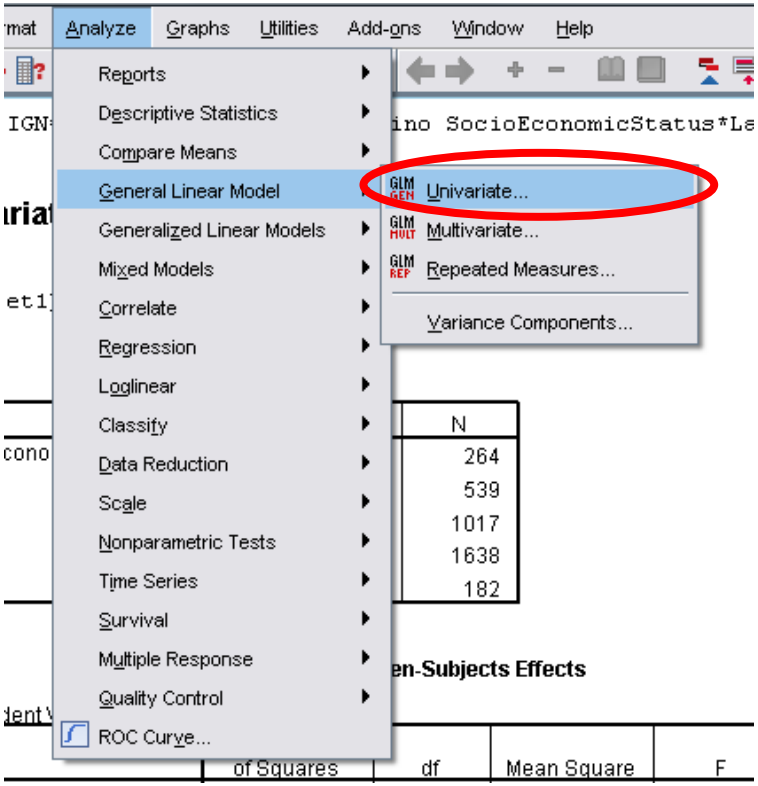
FYI:

"Fitted" and "predicted" are synonymous.

FYI:

When we talk about scatterplots, we talk about plotting Y vs. X (not X vs. Y). When we talk about regression, we talk about regressing Y on X. It helps me to think of reading the plots and models from left to right.

Unit 11 Appendix: SPSS Tutorial

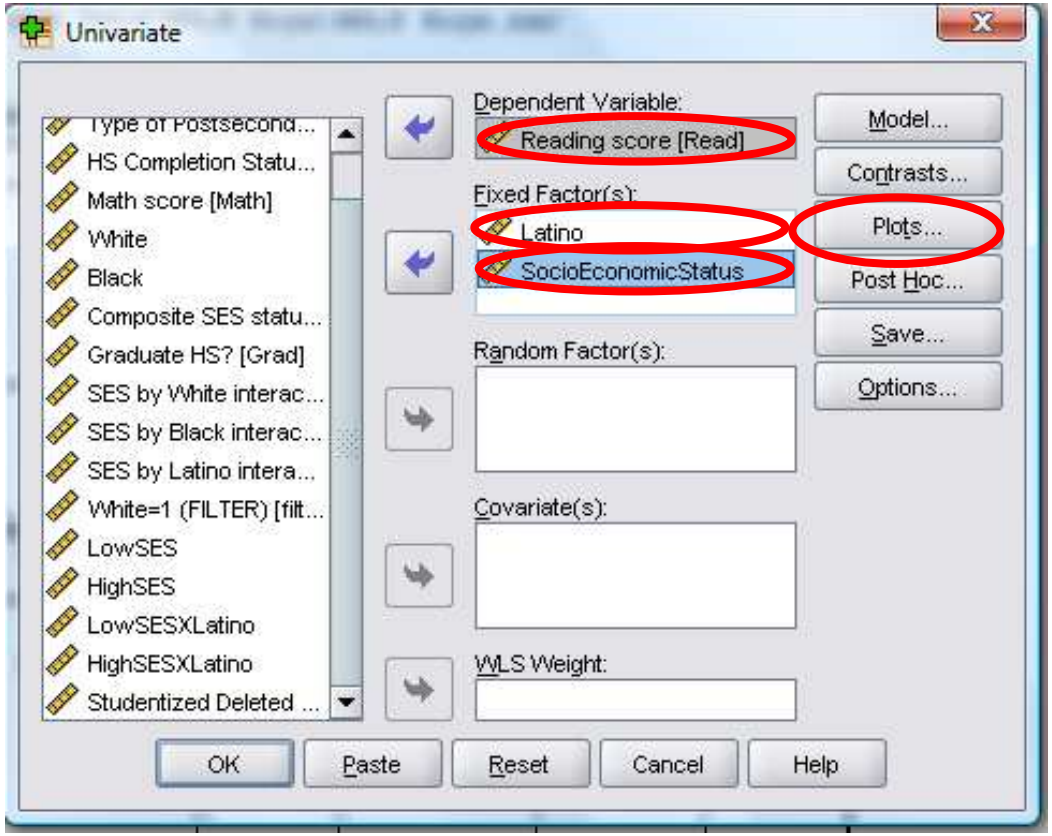


Go to:

Analyze > General Linear Model > Univariate

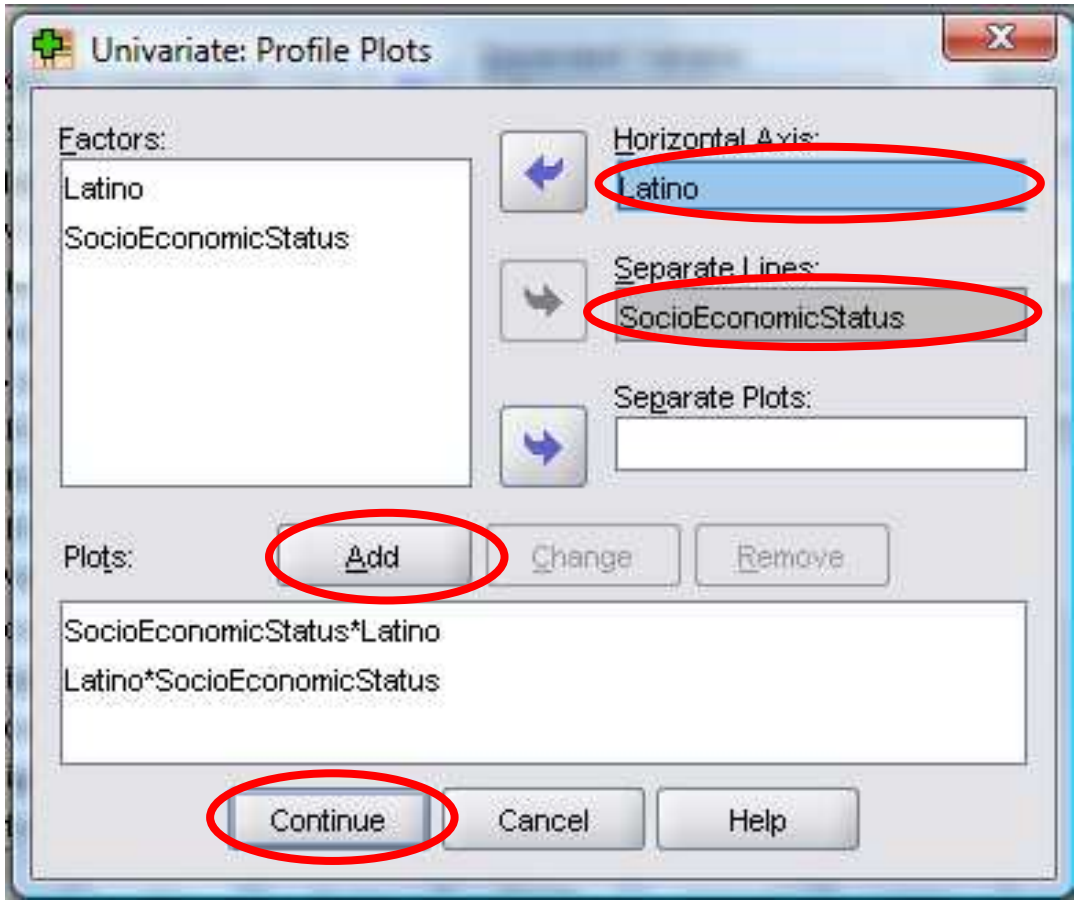
Add your outcome (i.e., dependent variable).
Add your <u>two</u> factors as fixed factors.

Go to "Plots…"



Unit 11 Appendix: SPSS Tutorial

Feel free to create a variety of graphs. Make sure you add each combination. When you've added a plot, you will see it appear below. In my example, I have added the two plots that appear in these slides.



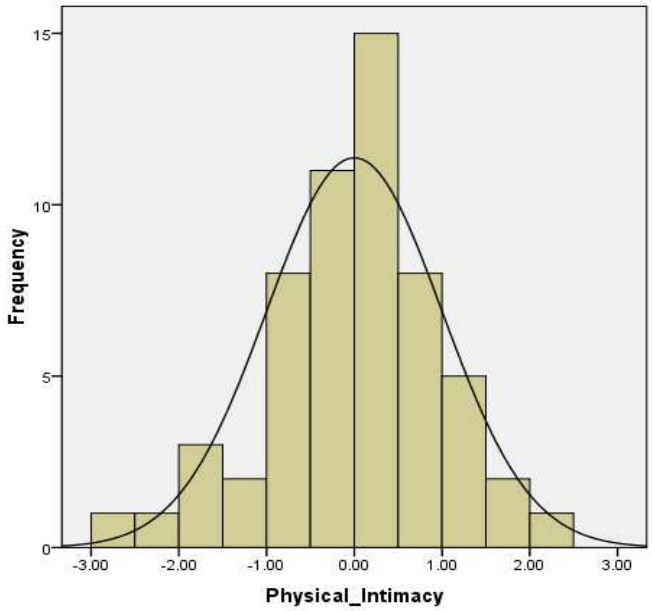
Perceived Intimacy of Adolescent Girls (Intimacy.sav)

- **Overview: Dataset contains self-ratings of the intimacy that adolescent girls perceive themselves as having with: (a) their mother and (b) their boyfriend.**
- **Source: HGSE thesis by Dr. Linda Kilner entitled Intimacy in Female Adolescent's Relationships with Parents and Friends (1991). Kilner collected the ratings using the Adolescent Intimacy Scale.**
- **Sample: 64 adolescent girls in the sophomore, junior and senior classes of a local suburban public school system.**
- **Note on Physical_Intimacy (with boyfriend): This is a composite variable based on a principle components analysis. Girls who score high on Physical_Intimacy scored high on (1) Physical Affection and (2) Mutual Caring, but low on (3) Risk Vulnerability and (4) Resolve Conflicts, regardless of (5) Trust and (6) Self Disclosure.**
- **Variables:**

(Physical_Intimacy)
Physical Intimacy With Boyfriend—see above

(RiskVulnerabilityWMom)
1=Tend to Risk Vulnerability with Mom, 0=Not

(ResolveConflictWMom)
1=Tend to Resolve Conflict with Mom, 0=Not



Perceived Intimacy of Adolescent Girls (Intimacy.sav)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.290[a]	.084	.032	.98378419

a. Predictors: (Constant), RVXRC, ResConflictWMom, RiskVulnerabilityWMom

b. Dependent Variable: REGR factor score 2 for analysis 1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.705	3	1.568	1.620	.196[a]
	Residual	51.295	53	.968		
	Total	56.000	56			

a. Predictors: (Constant), RVXRC, ResConflictWMom, RiskVulnerabilityWMom

b. Dependent Variable: REGR factor score 2 for analysis 1

Use the fitted model to generate two predictions:

One for girls who risk vulnerability with mom but who do not resolve conflicts with mom.

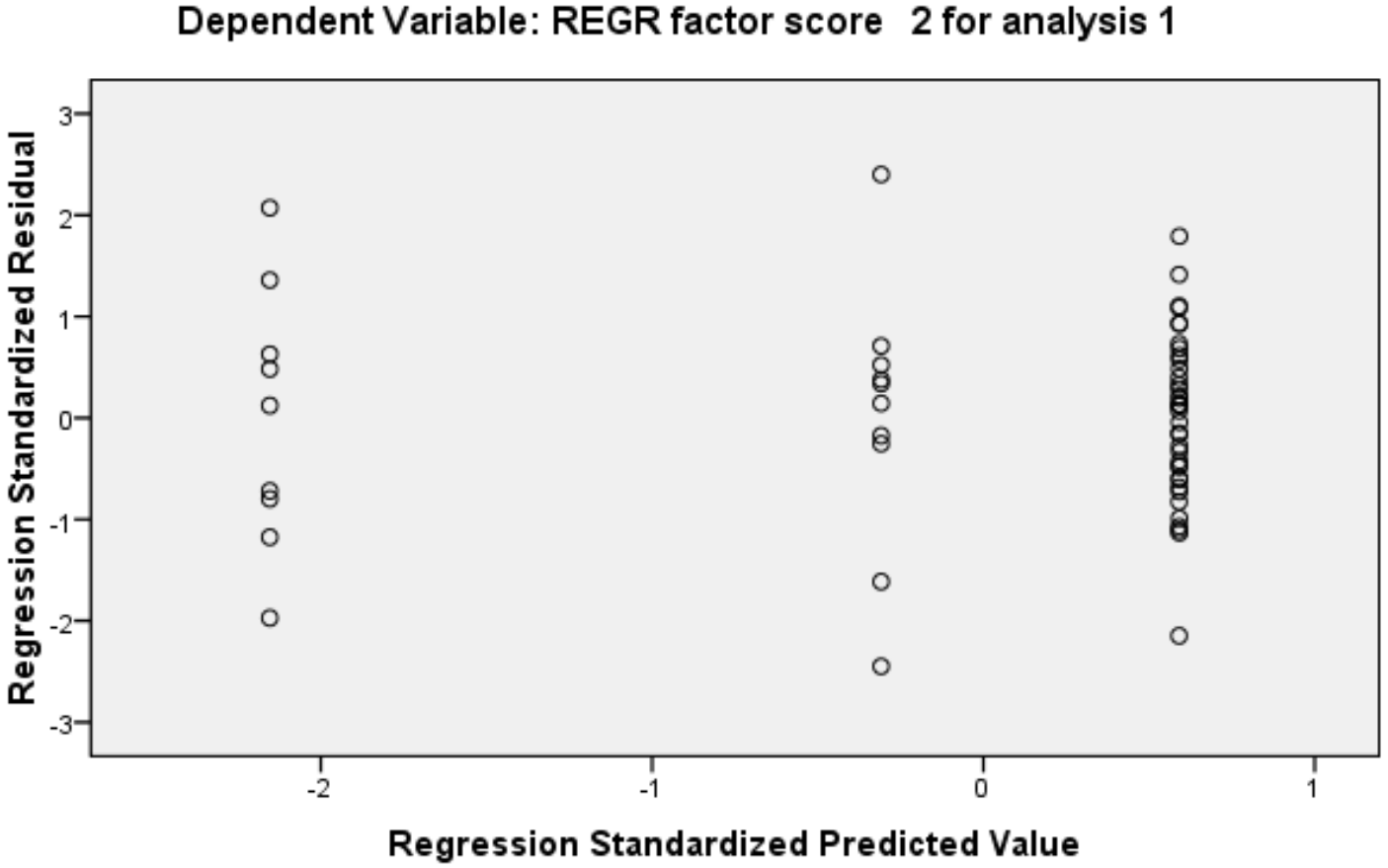
One for girls who risk vulnerability with mom and who resolve conflicts with mom..

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.171	.239		.718	.476	-.307	.650
	RiskVulnerabilityWMom	-.796	.406	-.401	-1.962	.055	-1.609	.018
	ResConflictWMom	-.261	.392	-.131	-.665	.509	-1.047	.526
	RVXRC	1.057	.554	.514	1.906	.062	-.055	2.169

a. Dependent Variable: REGR factor score 2 for analysis 1

Perceived Intimacy of Adolescent Girls (Intimacy.sav)



Perceived Intimacy of Adolescent Girls (Intimacy.sav)

Univariate Analysis of Variance

Between-Subjects Factors

		N
RiskVulnerabilityWMom	0	27
	1	30
ResConflictWMom	0	26
	1	31

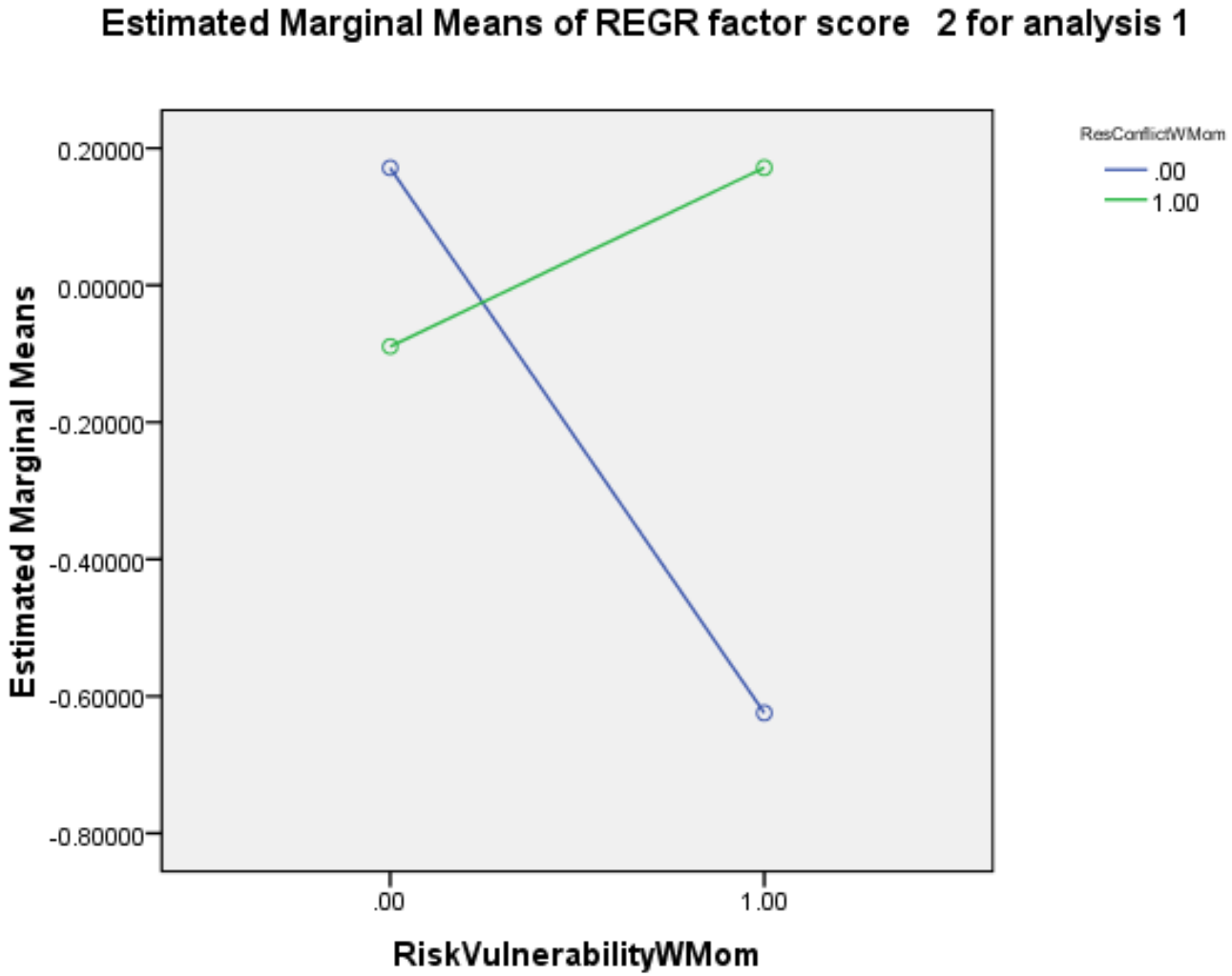
Tests of Between-Subjects Effects

Dependent Variable:REGR factor score 2 for analysis 1

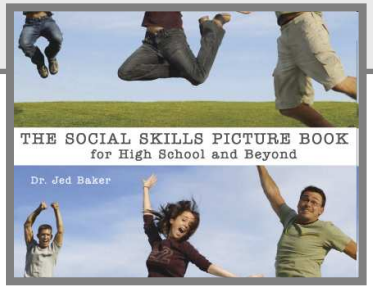
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	4.705[a]	3	1.568	1.620	.196
Intercept	.433	1	.433	.448	.506
RiskVulnerabilityWMom	.900	1	.900	.930	.339
ResConflictWMom	.902	1	.902	.932	.339
RiskVulnerabilityWMom * ResConflictWMom	3.516	1	3.516	3.633	.062
Error	51.295	53	.968		
Total	56.000	57			
Corrected Total	56.000	56			

a. R Squared = .084 (Adjusted R Squared = .032)

Perceived Intimacy of Adolescent Girls (Intimacy.sav)



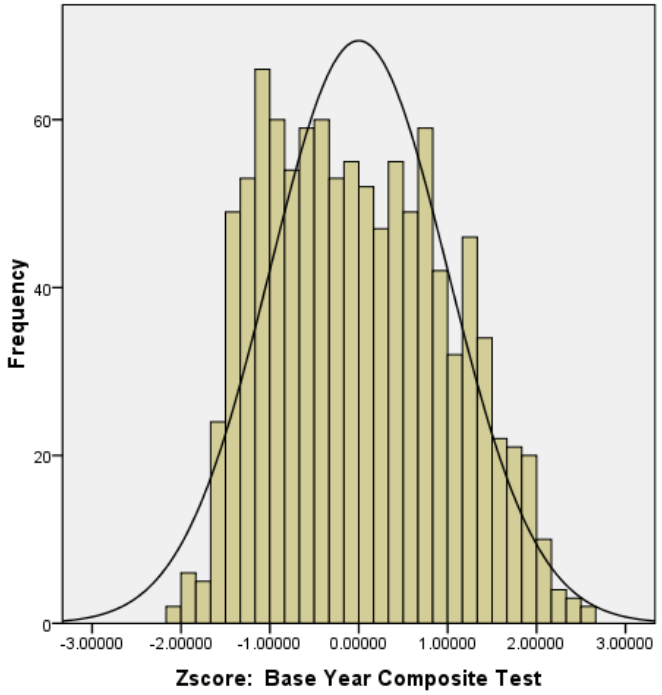
High School and Beyond (HSB.sav)



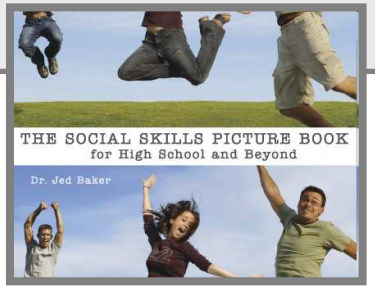
- **Overview: High School & Beyond – Subset of data focused on selected student and school characteristics as predictors of academic achievement.**
- **Source: Subset of data graciously provided by Valerie Lee, University of Michigan.**
- **Sample: This subsample has 1044 students in 205 schools. Missing data on the outcome test score and family SES were eliminated. In addition, schools with fewer than 3 students included in this subset of data were excluded.**
- **Variables:**

(ZBYTest) Standardized Base Year Composite Test Score
(Sex) 1=Female, 0=Male
(RaceEthnicity) Students Self-Identified Race/Ethnicity
1=White/Asian/Other, 2=Black, 3=Latino/a

Dummy Variables for RaceEthnicity:
(Black) 1=Black, 0=Else
(Latin) 1=Latino/a, 0=Else
*Note that we will use RaceEthnicity=1, White/Asian/Other, as our reference category.



High School and Beyond (HSB.sav)



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.454[a]	.206	.202	.89321776

a. Predictors: (Constant), SexXLatin, SexXBlack, 1 = Female, 0 = Other, 1 = Latino/a, 0 = Other, 1 = Black, 0 = Other

b. Dependent Variable: Zscore: Base Year Composite Test

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	214.844	5	42.969	53.857	.000[a]
	Residual	828.156	1038	.798		
	Total	1043.000	1043			

a. Predictors: (Constant), SexXLatin, SexXBlack, 1 = Female, 0 = Other, 1 = Latino/a, 0 = Other, 1 = Black, 0 = Other

b. Dependent Variable: Zscore: Base Year Composite Test

Use the fitted model to generate two predictions:

One for Latinas.

One for Latinos.

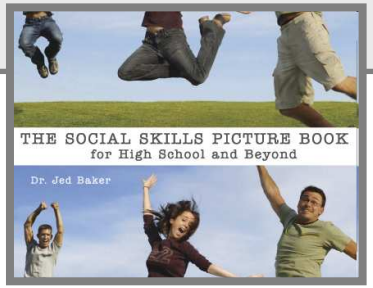
(Note that "SexXLatin" would be better named "FemalexLatin")

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.663	.063		10.553	.000	.540	.787
	1 = Female, 0 = Other	-.303	.086	-.151	-3.530	.000	-.472	-.135
	1 = Black, 0 = Other	-.990	.105	-.448	-9.437	.000	-1.196	-.784
	1 = Latino/a, 0 = Other	-.762	.096	-.349	-7.919	.000	-.951	-.573
	SexXBlack	.062	.137	.024	.452	.652	-.207	.330
	SexXLatin	.031	.133	.011	.231	.817	-.230	.291

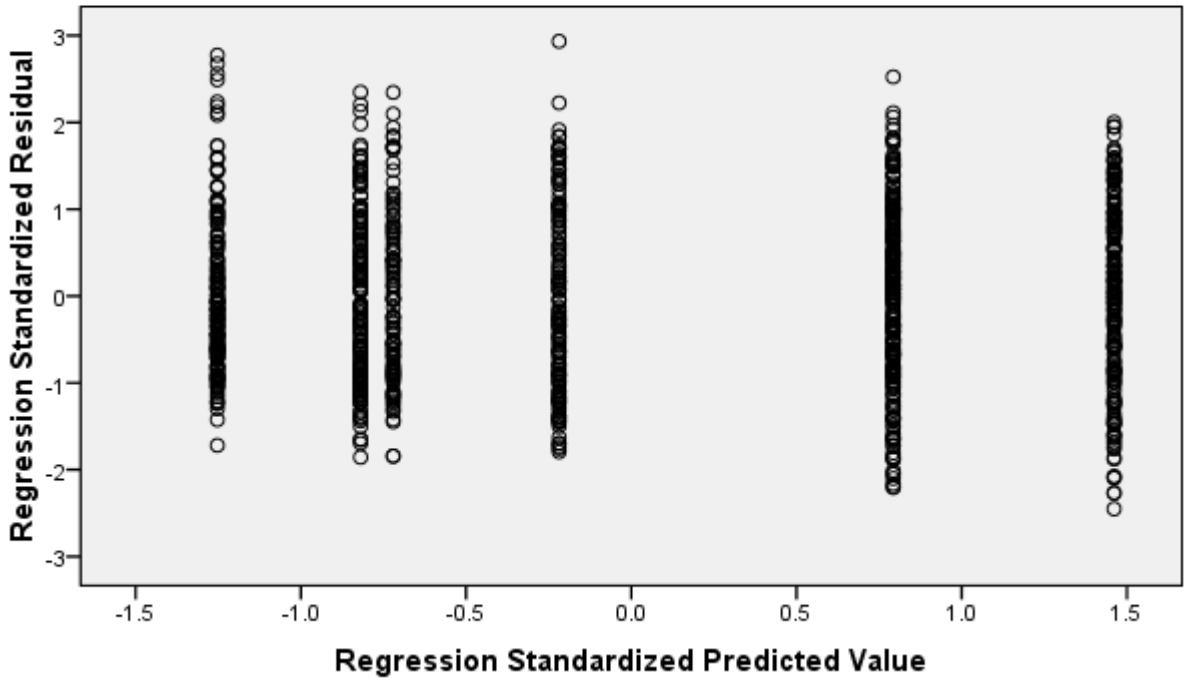
a. Dependent Variable: Zscore: Base Year Composite Test

High School and Beyond (HSB.sav)

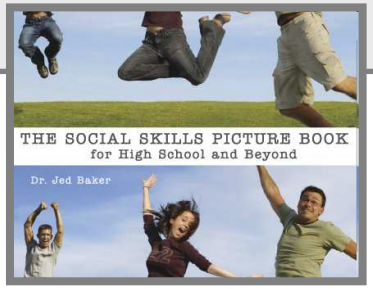


Scatterplot

Dependent Variable: Zscore: Base Year Composite Test



High School and Beyond (HSB.sav)



Between-Subjects Factors

		Value Label	N
1 = Female, 0 = Other	0	Male	465
	1	Female	579
RaceEthnicity	1	White/Asian/Other	434
	2	Black	299
	3	Latino/a	311

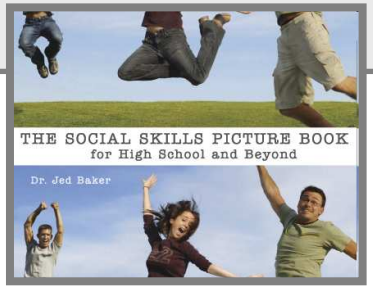
Tests of Between-Subjects Effects

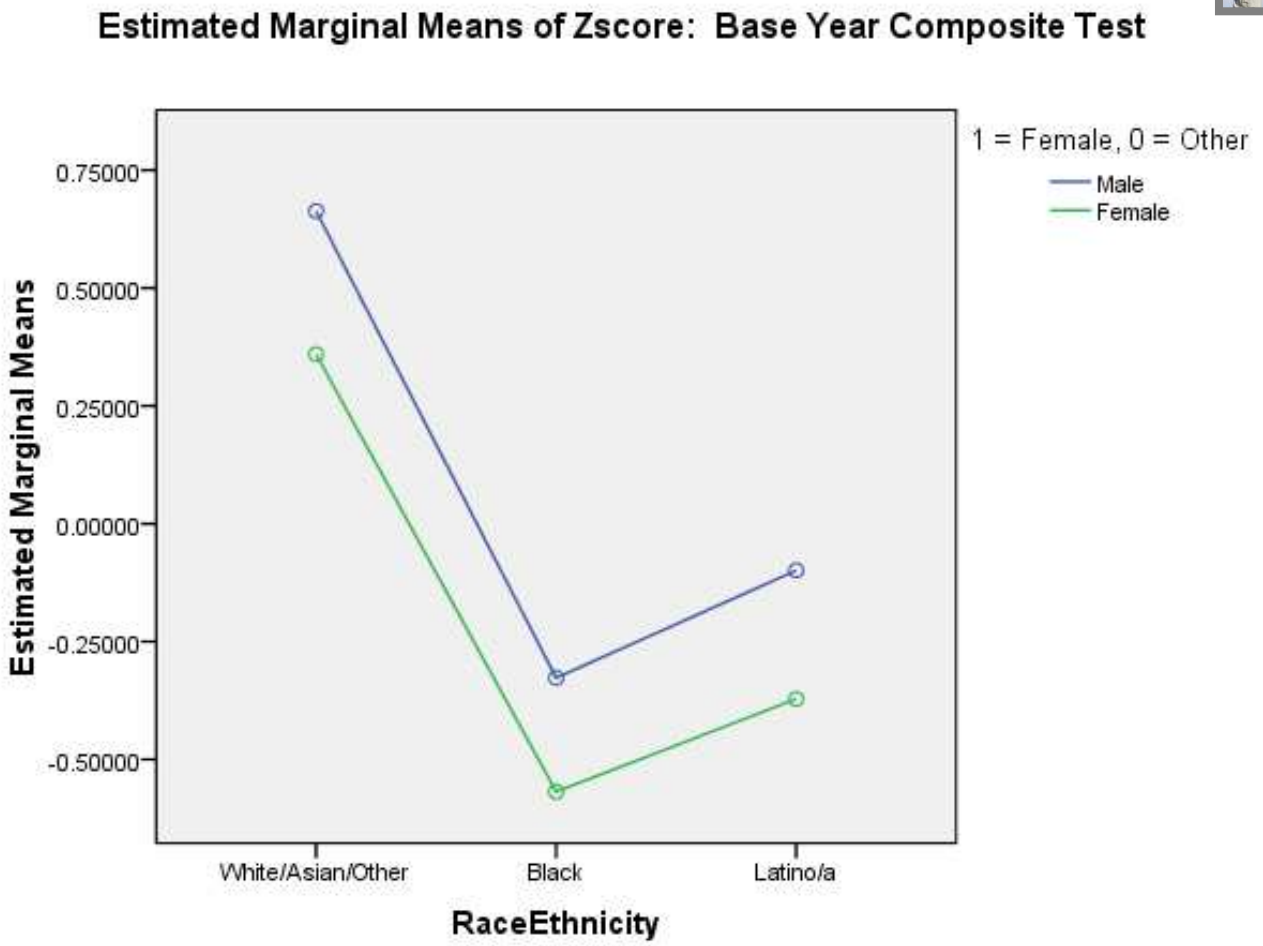
Dependent Variable:Zscore: Base Year Composite Test

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	214.844[a]	5	42.969	53.857	.000
Intercept	3.246	1	3.246	4.068	.044
Sex	18.385	1	18.385	23.044	.000
RaceEthnicity	186.160	2	93.080	116.665	.000
Sex * RaceEthnicity	.165	2	.082	.103	.902
Error	828.156	1038	.798		
Total	1043.000	1044			
Corrected Total	1043.000	1043			

a. R Squared = .206 (Adjusted R Squared = .202)

High School and Beyond (HSB.sav)





Understanding Causes of Illness (ILLCAUSE.sav)

- **Overview: Data for investigating differences in children's understanding of the causes of illness, by their health status.**
- **Source: Perrin E.C., Sayer A.G., and Willett J.B. (1991). Sticks And Stones May Break My Bones: Reasoning About Illness Causality And Body Functioning In Children Who Have A Chronic Illness, *Pediatrics*, 88(3), 608-19.**
- **Sample: 301 children, including a sub-sample of 205 who were described as asthmatic, diabetic,or healthy. After further reductions due to the *list-wise deletion* of cases with missing data on one or more variables, the analytic sub-sample used in class ends up containing: 33 diabetic children, 68 asthmatic children and 93 healthy children.**
- **Variables:**

(IllCause)	A Measure of Understanding of Illness Causality
(SocioEconomicStatus)	1=Low SES, 2=Lower Middle, 3=Upper Middle 4 = High SES
(HealthStatus)	1=Healthy, 2=Asthmatic 3=Diabetic

Dummy Variables for SocioEconomicStatus:

(LowSES)	1=Low SES, 0=Else
(LowerMiddleSES)	1=Lower MiddleSES, 0=Else
(HighSES)	1=High SES, 0=Else

*Note that we will use SocioEconomicStatus=3, Upper Middle SES, as our reference category.

Dummy Variables for HealthStatus:

(Asthmatic)	1=Asthmatic, 0=Else
(Diabetic)	1=Diabetic, 0=Else

*Note that we will use HealthStatus=1, Healthy, as our reference category.

Understanding Causes of Illness (ILLCAUSE.sav)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.488[a]	.238	.192	.91832

a. Predictors: (Constant), HighSESXDiabetic, HighSESXAsthmatic, LowSESXDiabetic, LowerMiddleSESXDiabetic, LowSESXAsthmatic, LowerMiddleSESXAsthmatic, HighSES, Asthmatic, Diabetic, LowerMiddleSES, LowSES

b. Dependent Variable: Understand Illness Causality

Use the fitted model to generate two predictions:

One for low SES asthmatic children.

One for lowmiddle SES asthmatic children.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	48.063	11	4.369	5.181	.000[a]
	Residual	153.484	182	.843		
	Total	201.547	193			

a. Predictors: (Constant), HighSESXDiabetic, HighSESXAsthmatic, LowSESXDiabetic, LowerMiddleSESXDiabetic, LowSESXAsthmatic, LowerMiddleSESXAsthmatic, HighSES, Asthmatic, Diabetic, LowerMiddleSES, LowSES

b. Dependent Variable: Understand Illness Causality

Coefficients[a]

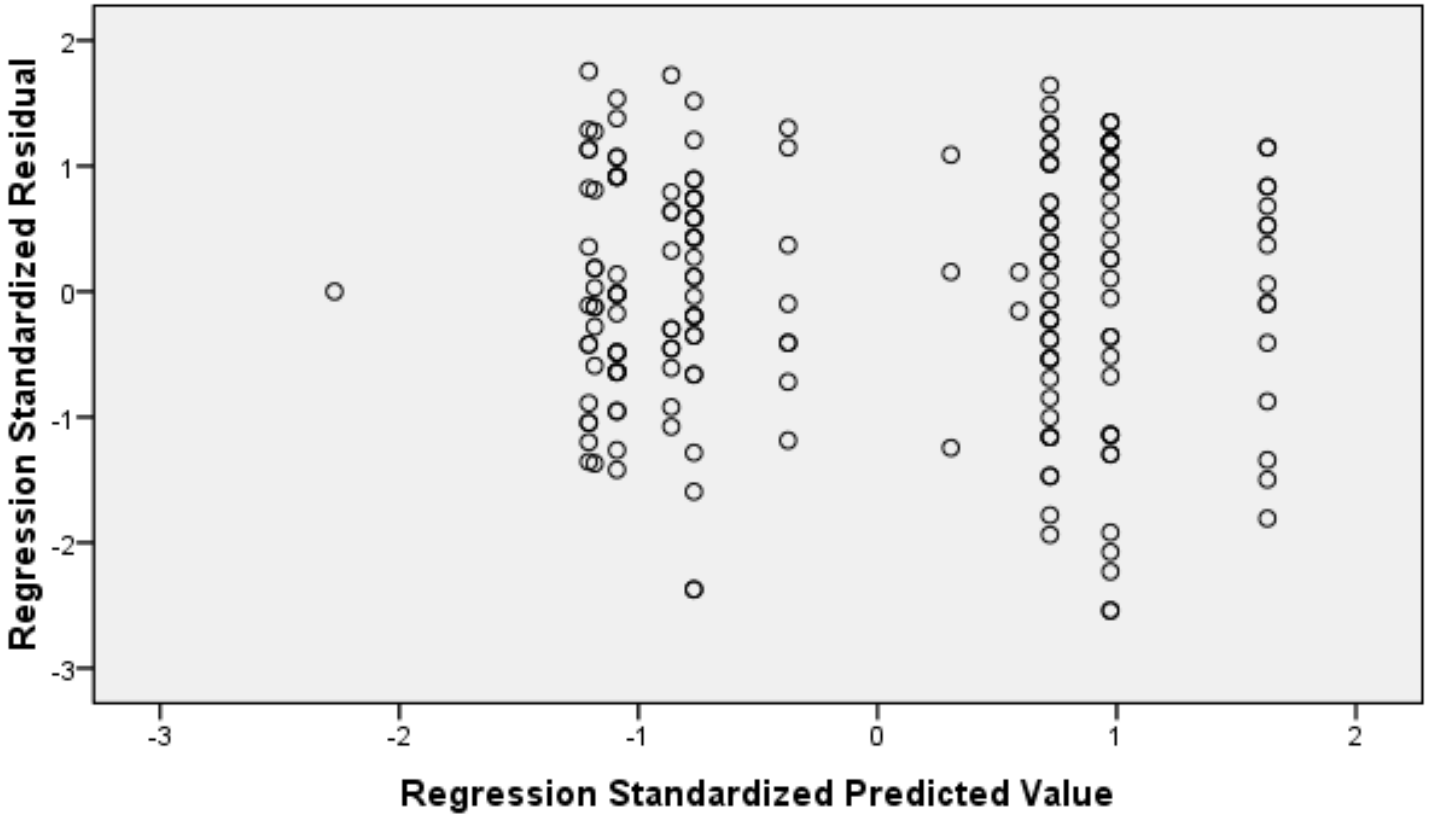
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	4.493	.145		30.943	.000	4.206	4.779
	LowSES	-1.493	.930	-.473	-1.606	.110	-3.327	.342
	LowerMiddleSES	.453	.272	.195	1.669	.097	-.082	.989
	HighSES	.126	.211	.051	.598	.551	-.290	.542
	Asthmatic	-.743	.226	-.348	-3.283	.001	-1.189	-.296
	Diabetic	-.790	.302	-.291	-2.615	.010	-1.387	-.194
	LowSESXAsthmatic	1.273	.977	.323	1.303	.194	-.654	3.201
	LowSESXDiabetic	1.737	1.020	.339	1.703	.090	-.276	3.749
	LowerMiddle SESXAsthmatic	-.614	.373	-.198	-1.647	.101	-1.350	.122
	LowerMiddle SESXDiabetic	-.613	.478	-.133	-1.283	.201	-1.556	.330
	HighSESXAsthmatic	.552	.704	.055	.784	.434	-.838	1.942
	HighSESXDiabetic	.457	.629	.055	.727	.468	-.784	1.699

a. Dependent Variable: Understand Illness Causality

Understanding Causes of Illness (ILLCAUSE.sav)

Scatterplot

Dependent Variable: Understand Illness Causality



Understanding Causes of Illness (ILLCAUSE.sav)

Univariate Analysis of Variance

Between-Subjects Factors

		Value Label	N
SocioEconomicStatus	1	Low SES	23
	2	Lower Middle SES	50
	3	Upper Midde SES	80
	4	High SES	41
HealthStatus	1	Healthy	93
	2	Asthmatic	68
	3	Diabetic	33

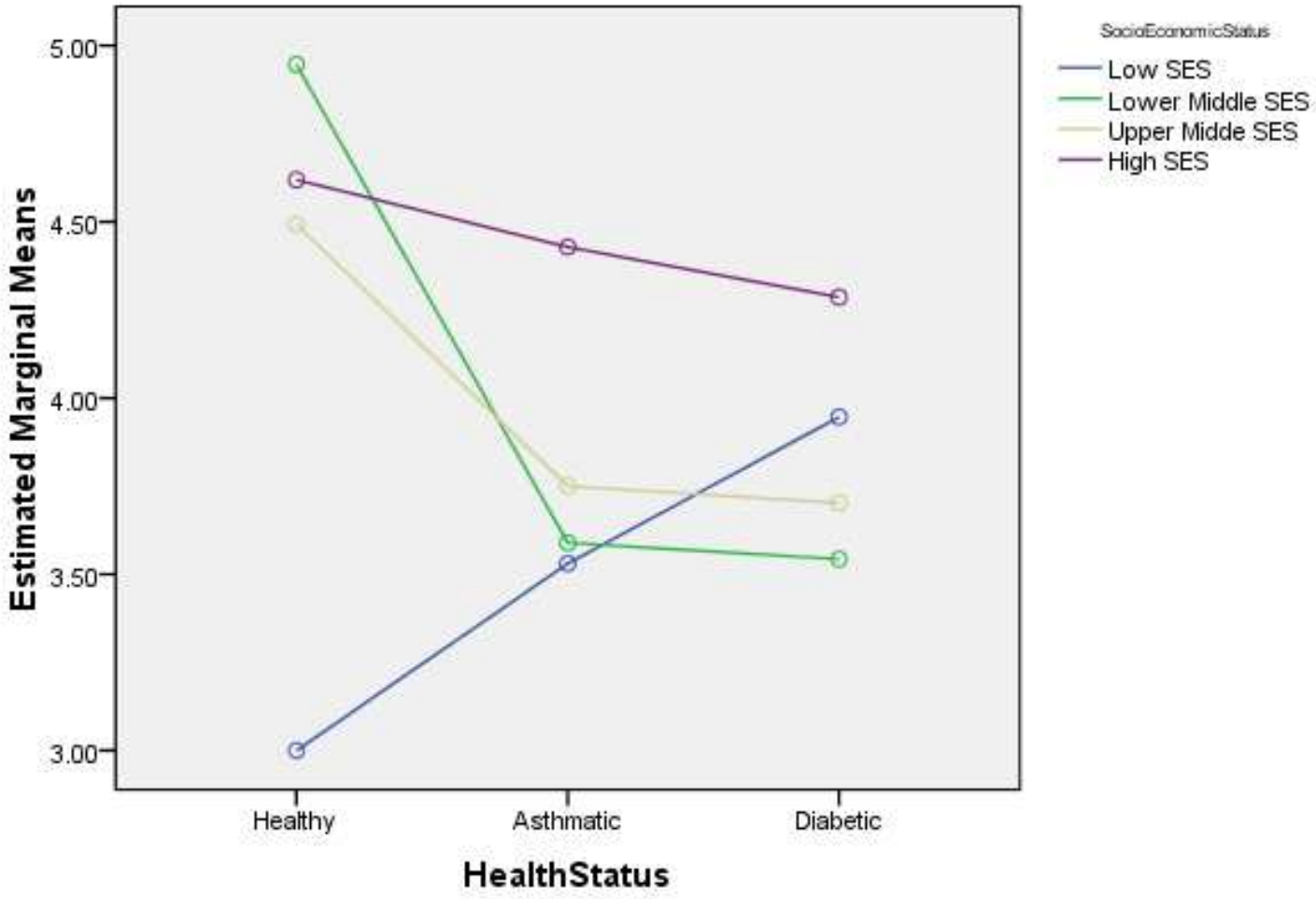
Tests of Between-Subjects Effects

Dependent Variable:Understand Illness Causality

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	48.063[a]	11	4.369	5.181	.000
Intercept	951.109	1	951.109	1127.817	.000
SocioEconomicStatus	4.049	3	1.350	1.600	.191
HealthStatus	1.964	2	.982	1.164	.314
SocioEconomicStatus * HealthStatus	7.553	6	1.259	1.493	.183
Error	153.484	182	.843		
Total	3515.849	194			
Corrected Total	201.547	193			

a. R Squared = .238 (Adjusted R Squared = .192)

Understanding Causes of Illness (ILLCAUSE.sav)

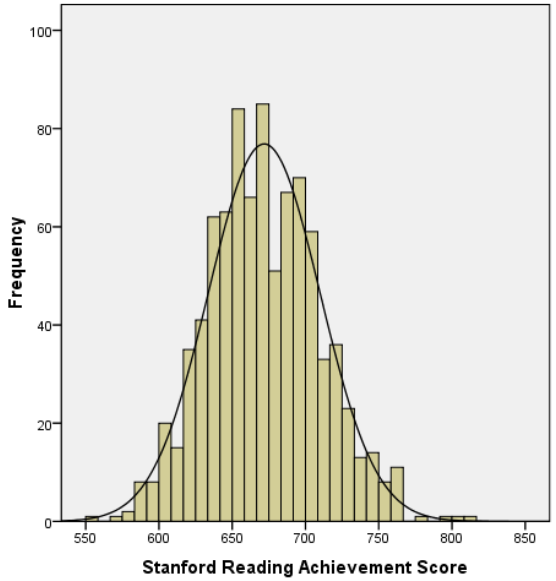


Children of Immigrants (ChildrenOfImmigrants.sav)

- **Overview: "CILS is a longitudinal study designed to study the adaptation process of the immigrant second generation which is defined broadly as U.S.-born children with at least one foreign-born parent or children born abroad but brought at an early age to the United States. The original survey was conducted with large samples of second-generation children attending the 8th and 9th grades in public and private schools in the metropolitan areas of Miami/Ft. Lauderdale in Florida and San Diego, California" (from the website description of the data set).**
- **Source: Portes, Alejandro, & Ruben G. Rumbaut (2001). *Legacies: The Story of the Immigrant SecondGeneration*. Berkeley CA: University of California Press.**
- **Sample: Random sample of 880 participants obtained through the website.**
- **Variables:**

(Reading) Stanford Reading Achievement Scores
(Depressed) 1=The Student is Depressed, 0=Not Depressed
(SESCat) A Relative Measure Of Socio-Economic Status
1=Low SES, 2=Mid SES, 3=High SES

Dummy Variables for SESCat:
(LowSES) 1=Low SES, 0=Else
(MidSES) 1=Mid SES, 0=Else
(HighSES) 1=High SES, 0=Else



Children of Immigrants (ChildrenOfImmigrants.sav)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.345[a]	.119	.114	35.814

a. Predictors: (Constant), DepressedXHighSES, DepressedXLowSES, HighSES, LowSES, Depressed

b. Dependent Variable: Stanford Reading Achievement Score

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	151860.209	5	30372.042	23.679	.000[a]
	Residual	1121033.968	874	1282.648		
	Total	1272894.177	879			

a. Predictors: (Constant), DepressedXHighSES, DepressedXLowSES, HighSES, LowSES, Depressed

b. Dependent Variable: Stanford Reading Achievement Score

Use the fitted model to generate two predictions:

One for high SES depressed students.

One for low SES depressed students.

Coefficients[a]

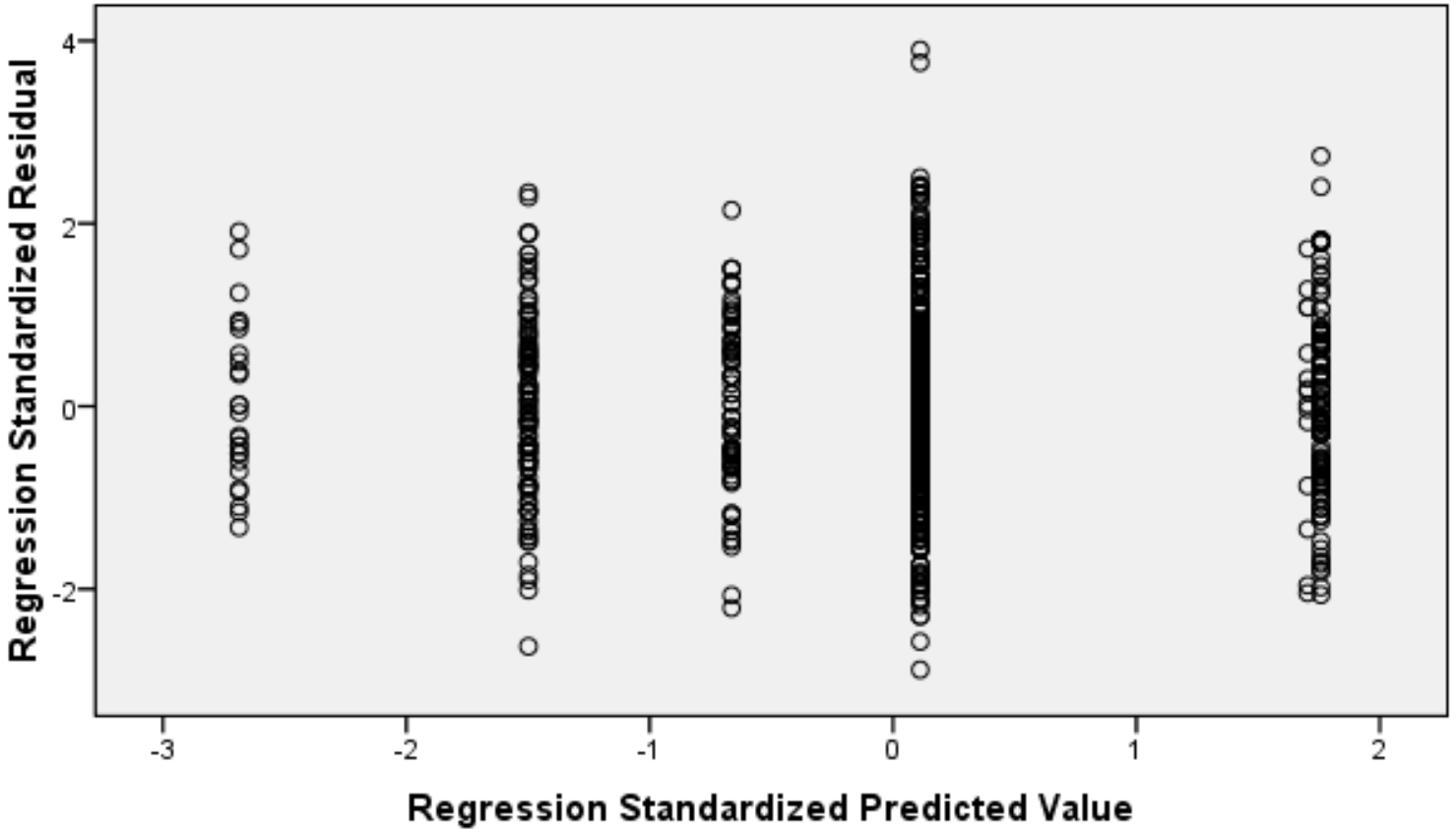
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	673.279	1.544		436.046	.000	670.248	676.309
	Depressed	-10.183	4.467	-.090	-2.280	.023	-18.950	-1.416
	LowSES	-21.155	3.706	-.203	-5.708	.000	-28.429	-13.881
	HighSES	21.643	3.679	.202	5.882	.000	14.422	28.864
	DepressedXLowSES	-5.441	8.980	-.024	-.606	.545	-23.066	12.184
	DepressedXHighSES	9.461	10.799	.032	.876	.381	-11.734	30.656

a. Dependent Variable: Stanford Reading Achievement Score

Children of Immigrants (ChildrenOfImmigrants.sav)

Scatterplot

Dependent Variable: Stanford Reading Achievement Score



Children of Immigrants (ChildrenOfImmigrants.sav)

Univariate Analysis of Variance

Between-Subjects Factors

		N
Depressed	0	766
	1	114
SESCat	1	139
	2	611
	3	130

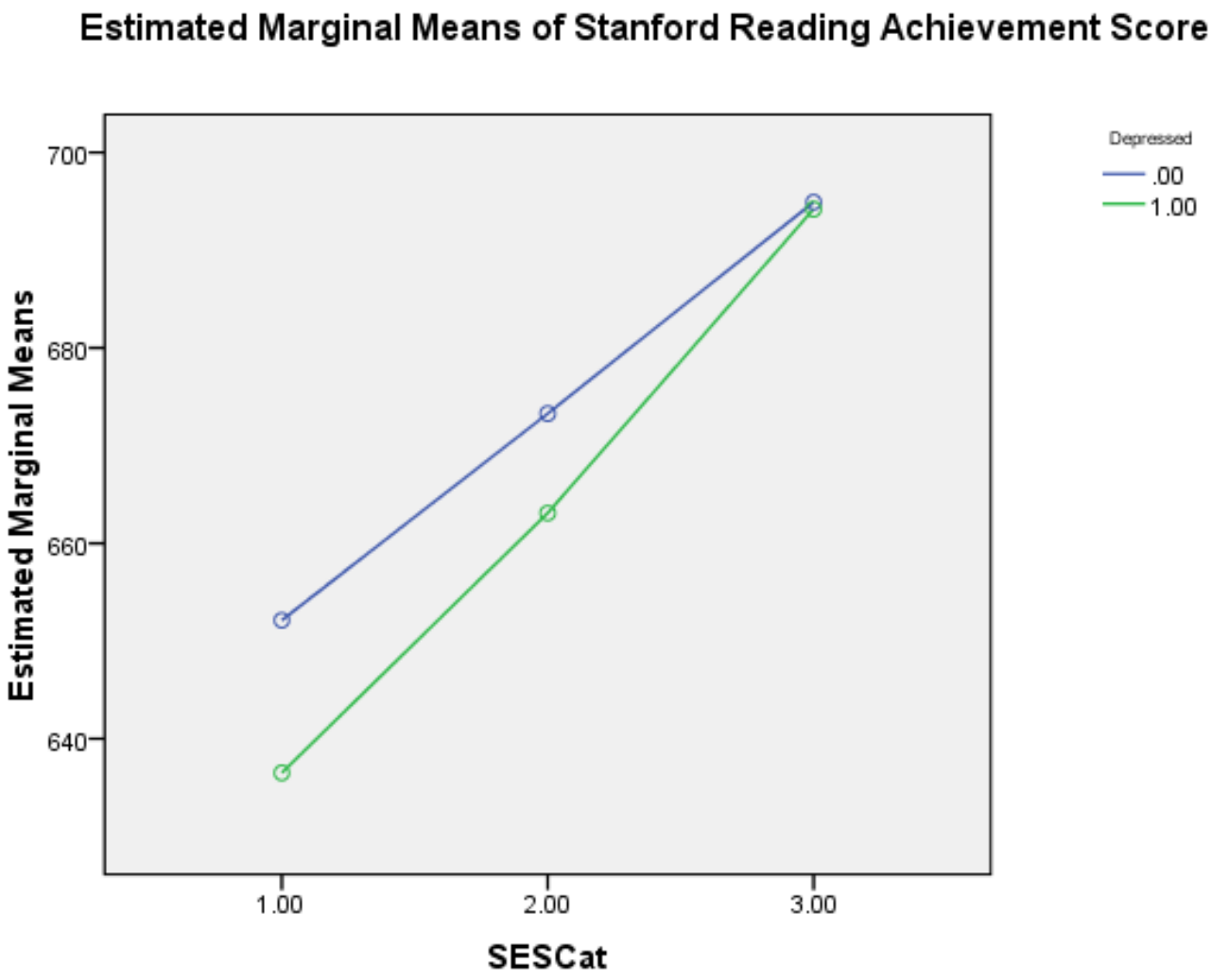
Tests of Between-Subjects Effects

Dependent Variable:Stanford Reading Achievement Score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	151860.209[a]	5	30372.042	23.679	.000
Intercept	1.166E8	1	1.166E8	90880.074	.000
Depressed	5091.226	1	5091.226	3.969	.047
SESCat	84142.991	2	42071.496	32.801	.000
Depressed * SESCat	1812.391	2	906.195	.707	.494
Error	1121033.968	874	1282.648		
Total	3.984E8	880			
Corrected Total	1272894.177	879			

a. R Squared = .119 (Adjusted R Squared = .114)

Children of Immigrants (ChildrenOfImmigrants.sav)

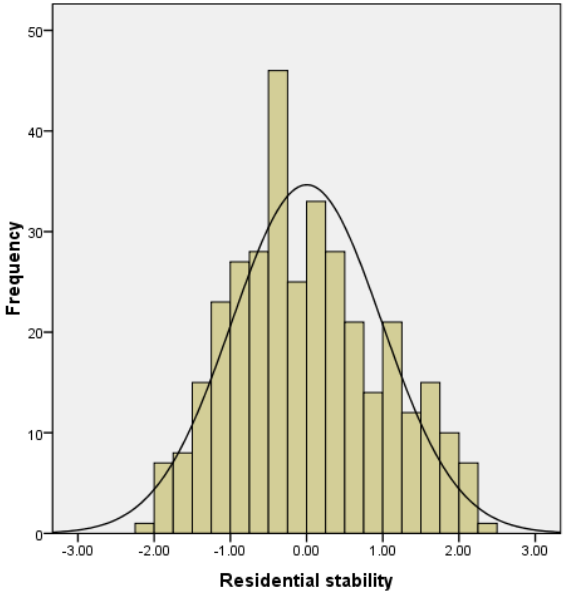


Human Development in Chicago Neighborhoods (Neighborhoods.sav)

- **These data were collected as part of the Project on Human Development in Chicago Neighborhoods in 1995.**
- **Source: Sampson, R.J., Raudenbush, S.W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, 277, 918-924.**
- **Sample: The data described here consist of information from 343 Neighborhood Clusters in Chicago Illinois. Some of the variables were obtained by project staff from the 1990 Census and city records. Other variables were obtained through questionnaire interviews with 8782 Chicago residents who were interviewed in their homes.**
- **Variables:**

(ResStab) Residential Stability, A Measure Of Neighborhood Flux
(NoMurder95) 1=No Murders in 1995, 0=At Least One Murder in 1995
(SES) A Relative Measure Of Socio-Economic Status
1=Low SES, 2=Mid SES, 3=High SES

Dummy Variables for MothEdCat:
(LowSES) 1=Low SES, 0=Else
(MidSES) 1=Mid SES, 0=Else
(HighSES) 1=High SES, 0=Else



Human Development in Chicago Neighborhoods (Neighborhoods.sav)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.160[a]	.026	.011	.97878

a. Predictors: (Constant), NoMurder95XHighSES, NoMurder95XLowSES, LowSES, HighSES, NoMurder95

b. Dependent Variable: Residential stability

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.472	5	1.694	1.769	.119[a]
	Residual	321.889	336	.958		
	Total	330.361	341			

a. Predictors: (Constant), NoMurder95XHighSES, NoMurder95XLowSES, LowSES, HighSES, NoMurder95

b. Dependent Variable: Residential stability

Use the fitted model to generate two predictions:

One for high SES murderless neighborhoods.

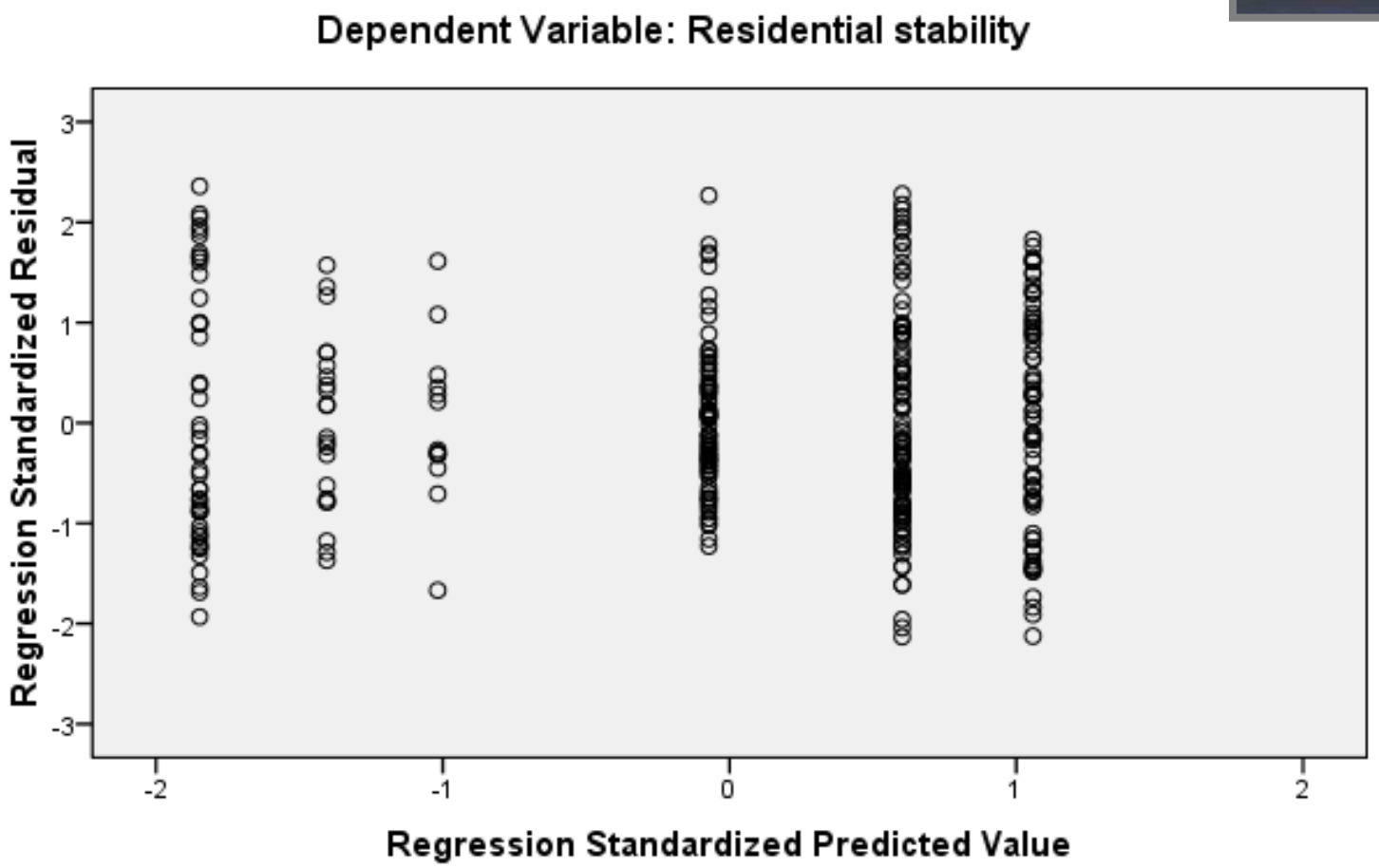
One for low SES murderless neighborhoods.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.098	.098		.992	.322	-.096	.291
	NoMurder95	-.316	.231	-.150	-1.371	.171	-.770	.138
	LowSES	-.106	.145	-.049	-.734	.463	-.391	.178
	HighSES	-.386	.172	-.189	-2.243	.026	-.725	-.047
	NoMurder95XLowSES	.167	.372	.033	.450	.653	-.564	.898
	NoMurder95XHighSES	.774	.293	.326	2.641	.009	.198	1.351

a. Dependent Variable: Residential stability

Human Development in Chicago Neighborhoods (Neighborhoods.sav)



Human Development in Chicago Neighborhoods (Neighborhoods.sav)

Between-Subjects Factors

		Value Label	N
SES	1	Low SES	98
	2	Mid SES	121
	3	High SES	123
NoMurder95	0	At Least One Murder in 1995	232
	1	No Murders in 1995	110

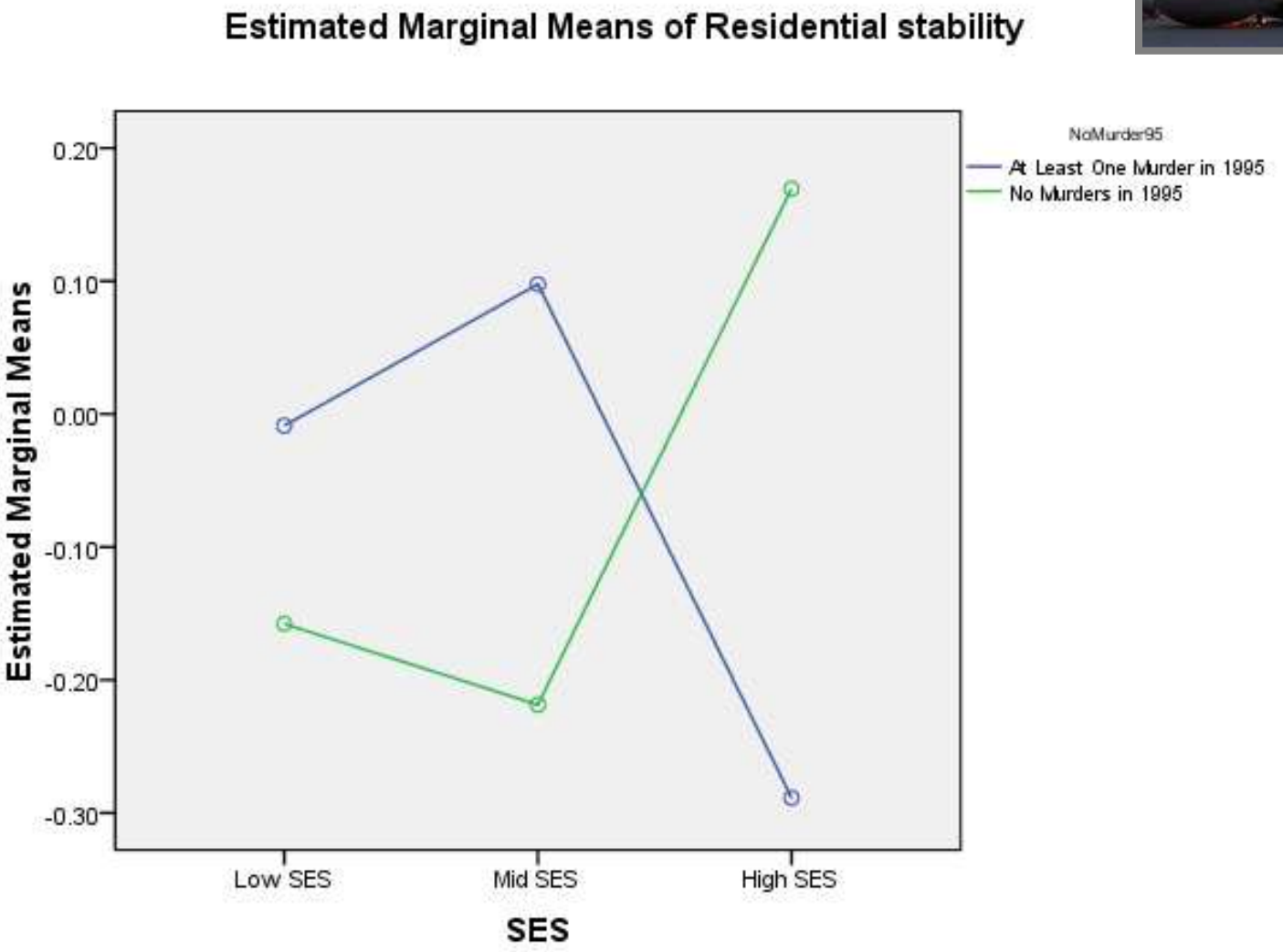
Tests of Between-Subjects Effects

Dependent Variable:Residential stability

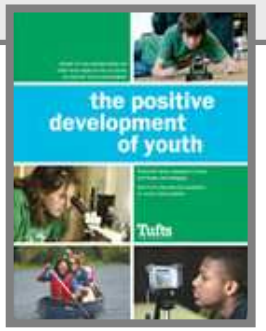
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	8.472[a]	5	1.694	1.769	.119
Intercept	.926	1	.926	.967	.326
SES	.020	2	.010	.010	.990
NoMurder95	.000	1	.000	.000	.986
SES * NoMurder95	7.568	2	3.784	3.950	.020
Error	321.889	336	.958		
Total	330.363	342			
Corrected Total	330.361	341			

a. R Squared = .026 (Adjusted R Squared = .011)

Human Development in Chicago Neighborhoods (Neighborhoods.sav)



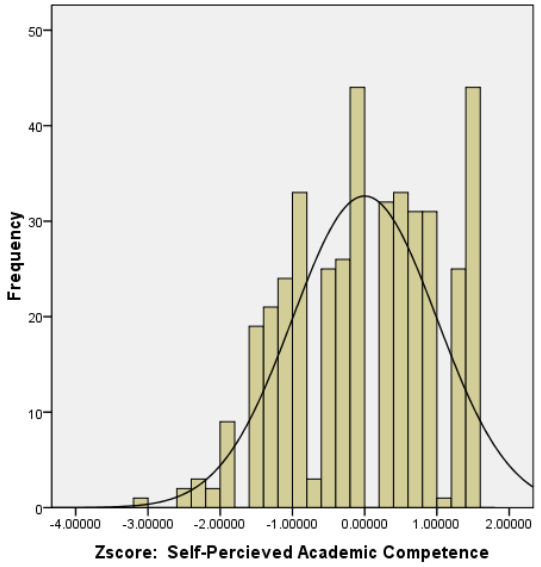
4-H Study of Positive Youth Development (4H.sav)



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

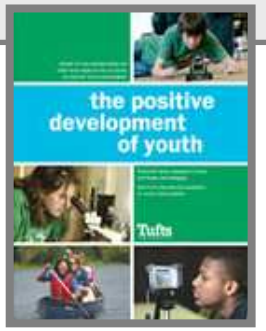
- Sample: These data consist of seventh graders who participated in Wave 3 of the 4-H Study of Positive Youth Development at Tufts University. This subfile is a substantially sampled-down version of the original file, as all the cases with any missing data on these selected variables were eliminated.
- Variables:

(ZAcadComp) Standardized Self-Perceived Academic Competence
(SexFem) 1=Female, 0=Male
(MothEdCat) Mother's Educational Attainment Category
1=High School Dropout, 2=High School Graduate,
3 =Up To 3 Years of College, 4 = 4-Plus Years of College



Dummy Variables for MothEdCat:
(MomHSDropout) 1=High School Dropout, 0=Else
(MomHSGrad) 1=High School Graduate, 0=Else
(MomUpTo3YRSCollege) 1=Up To 3 Years of College, 0=Else
(Mom4plusYRSCollege) 1=4-Plus Years of College, 0=Else

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



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.337[a]	.114	.098	.94956329

a. Predictors: (Constant), SexFemXMom4plusYRSCollege, SexFemXMomHSDropout, SexFemXMomUpTo3YRSCollege, Female = 1, Male = 0, Mom4plusYRSCollege, MomUpTo3YRSCollege, MomHSDropout

b. Dependent Variable: Zscore: Self-Percieved Academic Competence

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	46.430	7	6.633	7.356	.000[a]
	Residual	361.570	401	.902		
	Total	408.000	408			

a. Predictors: (Constant), SexFemXMom4plusYRSCollege, SexFemXMomHSDropout, SexFemXMomUpTo3YRSCollege, Female = 1, Male = 0, Mom4plusYRSCollege, MomUpTo3YRSCollege, MomHSDropout

b. Dependent Variable: Zscore: Self-Percieved Academic Competence

Use the fitted model to generate two predictions:

One for girls whose moms dropped out from HS.

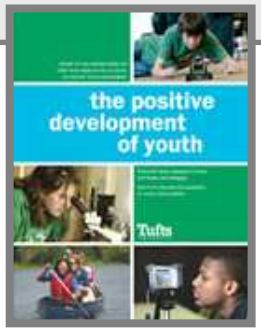
One for boys whose moms dropped out from HS.

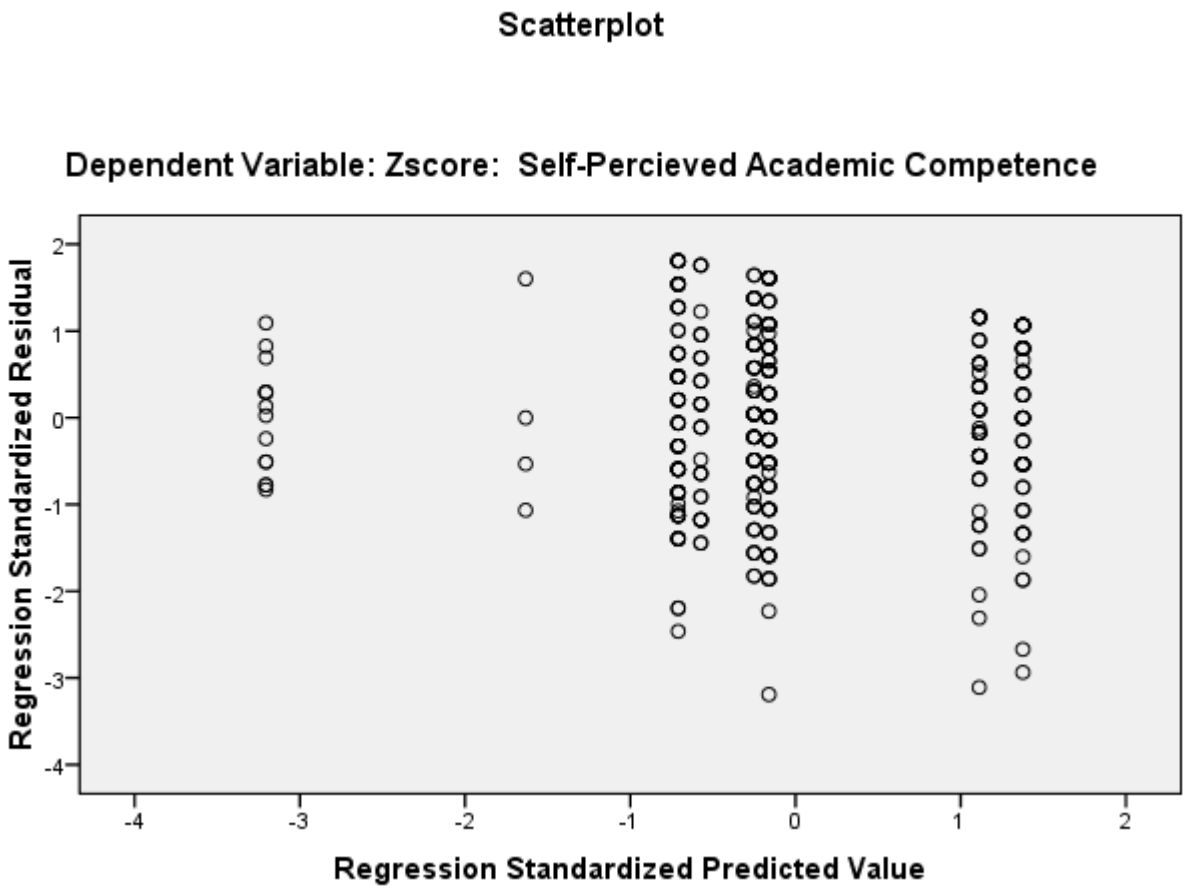
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-.193	.165		-1.170	.243	-.518	.132
	Female = 1, Male = 0	-.046	.198	-.023	-.234	.815	-.435	.342
	MomHSDropout	-.358	.503	-.073	-.712	.477	-1.346	.631
	MomUpTo3YRSCollege	.109	.201	.053	.540	.590	-.287	.505
	Mom4plusYRSCollege	.569	.206	.262	2.763	.006	.164	.973
	SexFemXMomHSDropout	-.484	.573	-.088	-.844	.399	-1.611	.643
	SexFemXMomUp To3YRSCollege	.077	.250	.032	.306	.759	-.415	.568
	SexFemXMom4plus YRSCollege	.135	.261	.049	.518	.605	-.377	.647

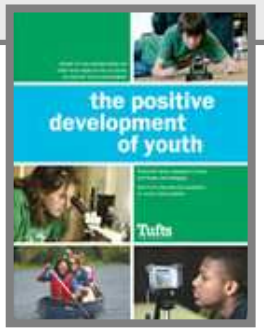
a. Dependent Variable: Zscore: Self-Percieved Academic Competence

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





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



Between-Subjects Factors

		Value Label	N
Female = 1, Male = 0	0	Boy	165
	1	Girl	244
MothEdCat	1	Mom HS Dropout	18
	2	Mom HS Grad	110
	3	Mom Up to 3 YRS College	156
	4	Mom 4+ YRS College	125

Tests of Between-Subjects Effects

Dependent Variable:Zscore: Self-Percieved Academic Competence

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	46.430[a]	7	6.633	7.356	.000
Intercept	4.403	1	4.403	4.883	.028
SexFem	.494	1	.494	.548	.460
MothEdCat	34.991	3	11.664	12.936	.000
SexFem * MothEdCat	1.190	3	.397	.440	.725
Error	361.570	401	.902		
Total	408.000	409			
Corrected Total	408.000	408			

a. R Squared = .114 (Adjusted R Squared = .098)

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