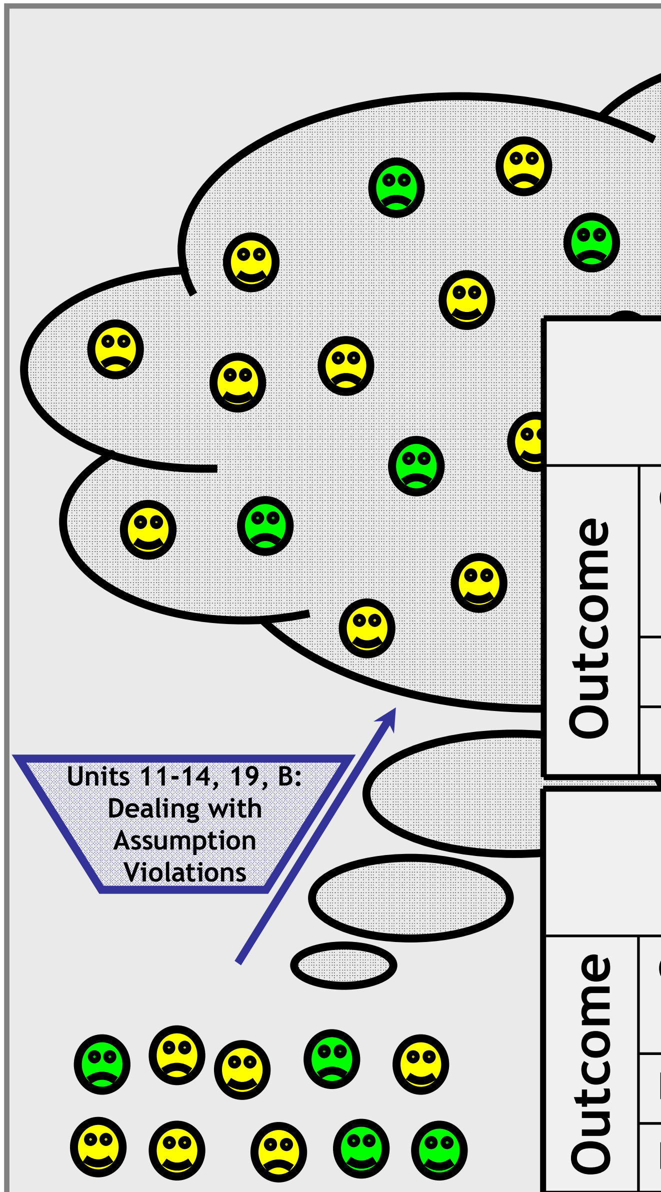
Unit 18: Road Map (VERBAL)

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

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

Unit 18: Road Map (Schematic)



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

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

Unit 18: Roadmap (SPSS Output)

Coefficients[a]

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

a. Dependent Variable: READING

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

Unit 18 Post Hole:

Sketch two model building strategies: a baseline-control strategy and a question-centered strategy.

Unit 18 Technical Memo and School Board Memo:

Create a table of hierarchical fitted models that tells a logical and coherent story of your final model, and then use words to tell that logical and coherent story.

Unit 18: Technical Memo and School Board Memo

Work Products (Part I of II):

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

  A. Introduction

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

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

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

    iv. Include your theoretical model.

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

    i. Describe the data set. (1 Sentence)

    ii. Describe your variables. (1 Paragraph Each)

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

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

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

      d. Note validity threats due to measurement error.

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

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

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

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

Unit 18: Technical Memo and School Board Memo

Work Products (Part II of II):

I. Technical Memo (continued)

D. Regression Analysis. Answer your research question using <u>inferential statistics</u>. Weave your strategy into a coherent story.

i. Include your fitted model.

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

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

iv. Create, display and discuss a table with a taxonomy of fitted regression models.

v. Use spreadsheet software to graph the relationship(s), and include a well-written caption.

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

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

viii. If regression diagnostics reveal a problem, describe the problem and the implications for your analysis and, if possible, correct the problem.

i. Primarily, check your residual-versus-fitted (RVF) plot. (Glance at the residual histogram and P-P plot.)

ii. Check your residual-versus-predictor plots.

iii. Check for influential outliers using leverage, residual and influence statistics.

iv. Check your main effects assumptions by checking for interactions before you finalize your model.

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

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

1. Note if the shape foreshadows a need to nonlinearly transform and, if so, which transformation might do the trick.

ii. For each relationship between your outcome and predictor, use a coherent narrative to convey the results of your exploratory bivariate analysis of the data. (1 Paragraph Each)

1. If a relationship is non-linear, transform the outcome and/or predictor to make it linear.

2. If a relationship is heteroskedastic, consider using robust standard errors.

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

III. Memo Metacognitive

Unit 18: Research Question

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

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

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

Variables:

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

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

Control Predictors:

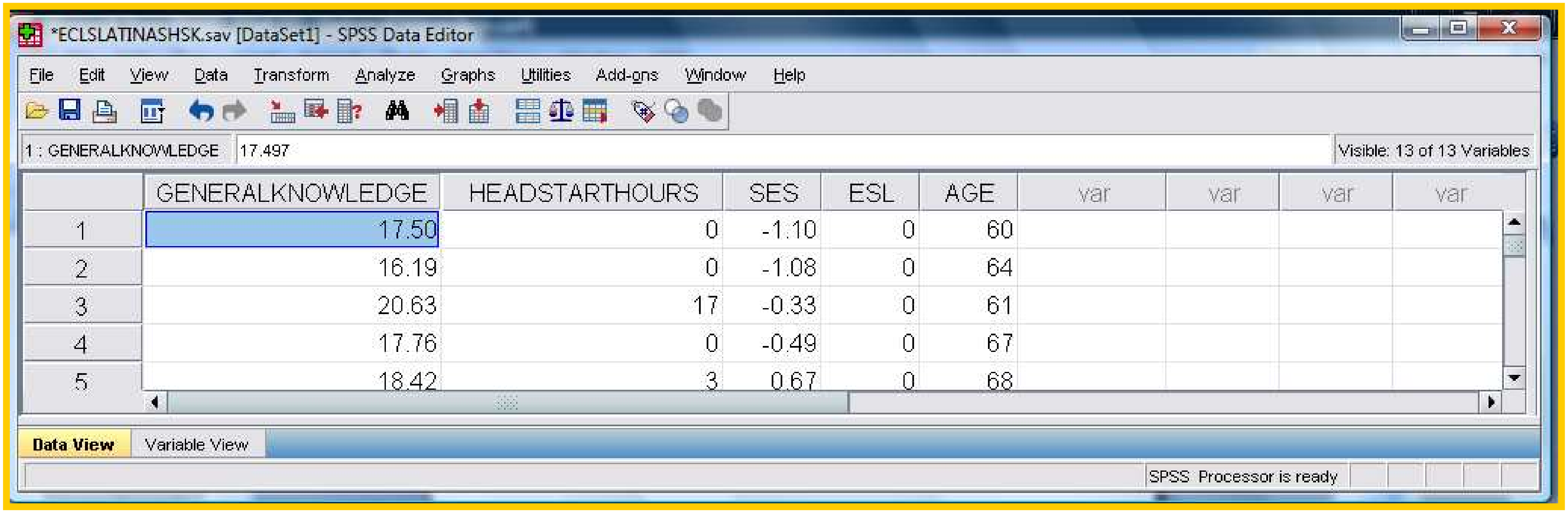
(***SES***) A Composite Measure of the Family's Socioeconomic Status

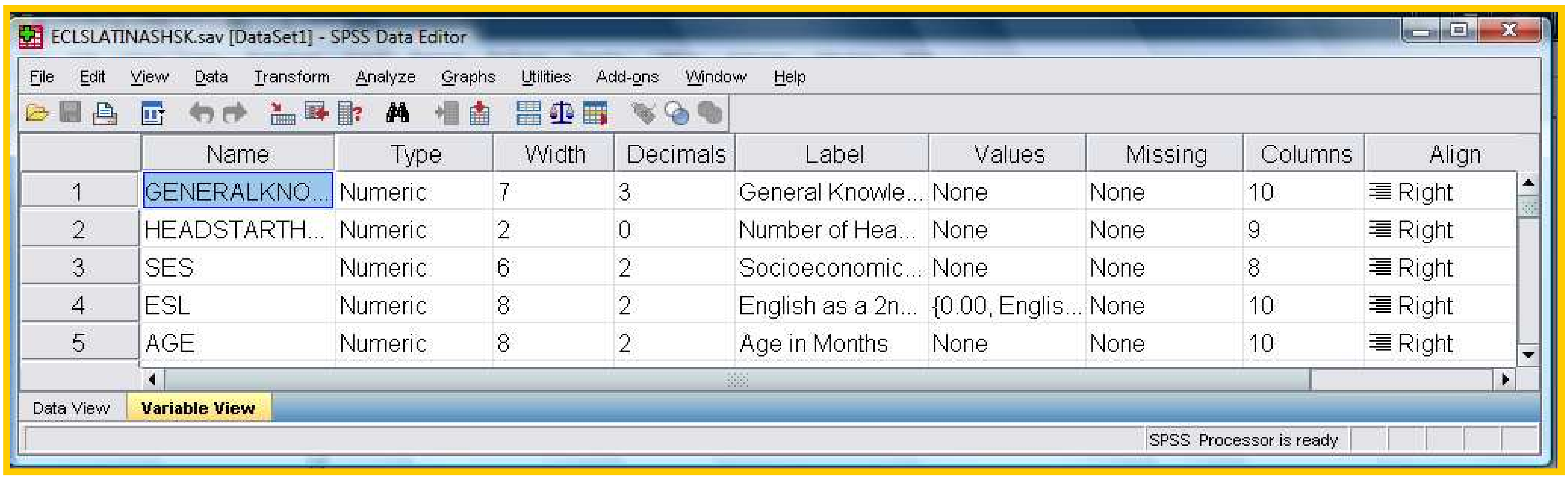
(***ESL***) A Dichotomy for which 1 Denotes that English is a 2^{nd} Language (0 = Not)

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

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

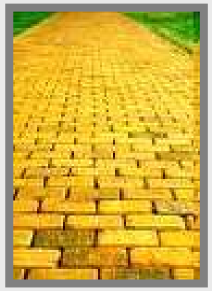
SPSS DATA


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

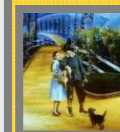

ECLSLATINASHSK.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Name Type Width Decimals Label Values Missing Columns Align
1 GENERALKNO... Numeric 7 3 General Knowle... None None 10 Right
2 HEADSTARTH... Numeric 2 0 Number of Hea... None None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

Four Multiple Regression Model Building Strategies: Two Good, Two Bad

- Question Centered Strategy
- Baseline Control Strategy
- Regression for the Brain Dead
- Data Analytic Fishing Expedition



There is a yellow brick road of data analysis based on a conscientious synergy of theories, questions, variables and models.

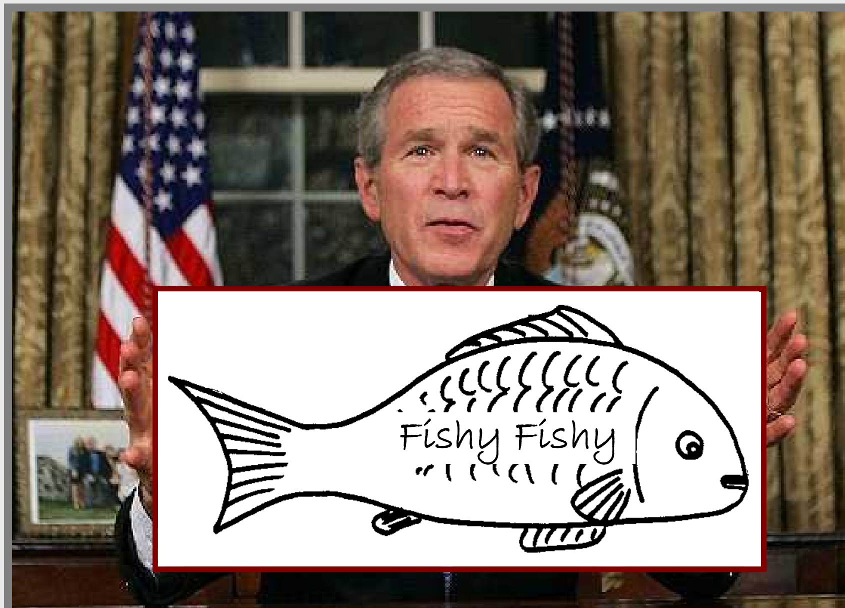


Stay on it.

In the back half of this unit, I will introduce tools for more advanced hypothesis testing, once you've strategically built your final model.

As soon as you stop fitting models and start "running regressions," you can run all you want, but you can't hide. In the end, the runaround will get you nowhere.

… and your little dog, too!



Question Centered Strategy

Table 1. Parameter estimates, (type III heteroskedasticity consistent standard errors), approximate p values and goodness-of-fit tests for a nested taxonomy of regression models that describe the relationship between scores on a kindergarten IRT scaled general knowledge test and the number of hours per week the child previously spent in Head Start controlling for age, SES and ESL among a nationally representative sample of female Latinas (n = 816)

	Models			
	M1	M2	M3	M4
Intercept	20.15***	20.29***	-2.91	-2.49
	(.27)	(.27)	(3.35)	(3.34)
Head Start Hours	-.10***	-.35***	-.10~	-.22**
	(.02)	(.06)	(.06)	(.08)
HSHxHSH		.008***	.003~	.007**
		(.001)	(.002)	(.003)
Age			.38***	.37***
			(.05)	(.05)
SES			4.01***	-4.43***
			(.35)	(.52)
ESL			-3.77***	-4.43***
			(.42)	(.52)
SESxESL				-1.66*
				(.74)
ESLxHSH				.26*
				(.12)
ESLxHSHxHSH				-.008*
				(.004)
R^2	.015***	.028***	.302***	.310***
df(Residual)	814	813	810	807
ΔR^2		.013**	.275***	.008*
df(ΔR^2)		1	3	3

Key: ~p < .10; *p < .05; **p < .01; ***p < .001

A question centered strategy for multiple regression model building always includes the question predictor in each model.

(1) We begin with a simple linear regression of the outcome on the question predictor.

(2) We address any non-linearity concerns.

(3) We add judiciously chosen control variables to our model, and we may want to add them in waves, for example, a wave of demographic variables, a wave of personal characteristics, and a wave of test scores. (Here we have displayed all the controls in one wave.)

(4) We check for interactions.

(5) We trim statistically insignificant controls from the model, unless they are of strong theoretical importance to us or our audience.

We follow rounds (1) and (5) with comprehensive assumption checks. Ideally, we would check our assumptions every step of the way, but there are only so many hours in the day.

Nobody wants to read your data analytic diary of everything you did during winter break.* Make your data analytic story to the point, even if the actual data analysis was meandering. Do not describe every assumption check or every interaction check. The astute reader is looking for clues in your writing not only that you made all the right moves but also that you know which moves are noteworthy. *This is a Willett-icism (more to follow).

Question Centered Presentation

Table 1. Parameter estimates, (type III heteroskedasticity consistent standard errors), approximate p values and goodness-of-fit tests for a nested taxonomy of regression models that describe the relationship between scores on a kindergarten IRT scaled general knowledge test and the number of hours per week the child previously spent in Head Start controlling for age, SES and ESL among a nationally representative sample of female Latinas (n = 816)

	Models			
	M1	M2	M3	M4
Intercept	20.15***	20.29***	-2.91	-2.49
	(.27)	(.27)	(3.35)	(3.34)
Head Start Hours	-.10***	-.35***	-.10~	-.22**
	(.02)	(.06)	(.06)	(.08)
HSHxHSH		.008***	.003~	.007**
		(.001)	(.002)	(.003)
Age			.38***	.37***
			(.05)	(.05)
SES			4.01***	-4.43***
			(.35)	(.52)
ESL			-3.77***	-4.43***
			(.42)	(.52)
SESxESL				-1.66*
				(.74)
ESLxHSH				.26*
				(.12)
ESLxHSHxHSH				-.008*
				(.004)
R^2	.015***	.028***	.302***	.310***
df(Residual)	814	813	810	807
ΔR^2		.013**	.275***	.008*
df(ΔR^2)		1	3	3

Key: ~p < .10; *p < .05; **p < .01; ***p < .001

The strength of the question centered presentation of your models is that the audience can follow the story of your question predictor as its relationship with the outcome changes in strength and/or magnitude upon various statistical controls.

You may or may not want to use a question centered presentation if you used a question centered strategy.

You may notice that the quadratic relationship between Head Start hours and general knowledge IRT scaled scores all but disappears when we control for Age, SES, and ESL.

However, when we allow the relationship to vary between speakers of English as a first and second language, we see the quadratic relationship reappear. As it turns out, there is a strong quadratic relationship for speakers of English as a first language and virtually no relationship for speakers of English as a second language.

With our final model, we predict 31% of the variance in general knowledge IRT scaled scores, but head start hours is only a small contributor to that percentage as evidenced by the 3% of variance that our model predicts without the controls.

The R^2 statistic never justifies a model.
What justifies a model? 1st, the model answers your research questions. 2nd, it meets third-variable objections. 3rd, it meets the GLM assumptions.

Baseline Control Strategy

Table 2. Parameter estimates, (type III heteroskedasticity consistent standard errors), approximate p values and goodness-of-fit tests for a nested taxonomy of regression models that describe the relationship between scores on a kindergarten IRT scaled general knowledge test and the number of hours per week the child previously spent in Head Start controlling for age, SES and ESL among a nationally representative sample of female Latinas (n = 816)

	Models		
	M1	M2	M3
Intercept	-2.99***	-2.99	-2.49
	(3.34)	(3.35)	(3.34)
Age	.38***	.38***	.37***
	(.05)	(.051)	(.05)
SES	4.56***	4.56***	-4.43***
	(.41)	(.41)	(.52)
ESL	-4.30***	-4.30***	-4.43***
	(.50)	(.50)	(.52)
SESxESL	-1.82*	-1.82*	-1.66*
	(.72)	(.72)	(.74)
Head Start Hours		.001	-.22**
		(.020)	(.08)
HSHxHSH			.007**
			(.003)
ESLxHSH			.26*
			(.12)
ESLxHSHxHSH			-.008*
			(.004)
R^2	.306***	.306***	.310***
df(Residual)	811	810	807
ΔR^2		.000	.005
df(ΔR^2)		1	3

Key: ~p < .10; *p < .05; **p < .01; ***p < .001

A baseline control strategy for multiple regression model building includes your predictor only after you have established your control model.

(1) We begin by building a control model. In building your control model, you may begin with a strong theory about the variables that influence your outcome but that you are not interesting (for your purposes) per se. Be sure to check for interactions among your control variables.

(2) We include our question predictor.

(3) We check for interactions.

Personally, I prefer the question centered approach because, when I have a question, I obsess about it, and it's easier for me to think about the control predictors by way of my question predictor (i.e., by way of my obsession). However, there are strengths to the baseline control strategy. It may force us to think more deeply about our control predictors which, for example, may lead us to check for interactions that we might not otherwise check. It may help us resist the temptation to fidget with our control predictors with an eye towards "making" our question predictor statistically significant, which may lead us to ignore (or otherwise mistreat) important control variables.

A combination of the question centered strategy and the baseline control strategy is often the best route. The key is *strategy*! You want a model building strategy that tightly unifies your theory, question and variables.

Baseline Control Presentation

Table 2. Parameter estimates, (type III heteroskedasticity consistent standard errors), approximate p values and goodness-of-fit tests for a nested taxonomy of regression models that describe the relationship between scores on a kindergarten IRT scaled general knowledge test and the number of hours per week the child previously spent in Head Start controlling for age, SES and ESL among a nationally representative sample of female Latinas (n = 816)

	Models		
	M1	M2	M3
Intercept	-2.99***	-2.99	-2.49
	(3.34)	(3.35)	(3.34)
Age	.38***	.38***	.37***
	(.05)	(.051)	(.05)
SES	4.56***	4.56***	-4.43***
	(.41)	(.41)	(.52)
ESL	-4.30***	-4.30***	-4.43***
	(.50)	(.50)	(.52)
SESxESL	-1.82*	-1.82*	-1.66*
	(.72)	(.72)	(.74)
Head Start Hours		.001	-.22**
		(.020)	(.08)
HSHxHSH			.007**
			(.003)
ESLxHSH			.26*
			(.12)
ESLxHSHxHSH			-.008*
			(.004)
R^2	.306***	.306***	.310***
df(Residual)	811	810	807
ΔR^2		.000	.005
df(ΔR^2)		1	3

Key: ~p < .10; *p < .05; **p < .01; ***p < .001

The strength of the baseline control presentation of your models is that the audience can easily see the predictive value added of your question predictor by noting the change in R^2 statistic from your control model to your final model. Recall from Unit 15 that the change in R^2 statistic from a tightly nested control model is a measure of the uniquely predicted variance of the question predictor. This change statistic is called the "partial R^2 statistic" or, in ANOVA language, the "partial η^2 statistic" or "partial eta-squared statistic."

In Model 2, for which we assume only a main effect, Head Start hours predicts a statistically insignificant proportion of the variation in general knowledge scores (partial $\eta^2 < .001$, $p = .959$). (Note for students: "partial η^2" (pronounced, "partial eta-squared") is just a more common way to write the partial R^2 statistic from Unit 15. It is the change in R^2 statistic associated with a variable when we add that variable (and only that variable) to the model that includes all the other variables.)

However, in Model 3, for which we relax the main effects assumption, we find that Head Start hours statistically significantly interacts with itself and ESL. Head Start hours predicts general knowledge scores, but the prediction differs by number of Head Start hours and ESL status. The reader may note that inclusion of the interaction effects increases our R^2 statistic by only .005.

For the sake of a crisp presentation, make your models as simple as possible, but no simpler. Remove a statistically insignificant predictor unless:

It is your question predictor.
It is an important control predictor.
It is part of a stat sig interaction.
It is part of a stat sig set of dummies.

Note that this is not stat sig because the F-test is not robust to heteroskedasticity.

R^2 Change Statistics

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.553[a]	.306	.302	5.77372	.306	89.284	4	811	.000
2	.553[b]	.306	.301	5.77728	.000	.003	1	810	.959
3	.557[c]	.310	.304	5.76826	.005	1.845	3	807	.137

You can get R^2 change statistics (and tests for their statistical significance) through SPSS dropdown menus by going to “Analyze > Regression > Linear” (as you would for any regression analysis). From there, choose “Statistics” and select “R squared change.” Note that this only makes sense when you fit several model consecutively. Do this in drop downs by using the “Next” button in the Block section.

Where did we get the p-values and degrees of freedom for R^2 statistics in the first place? Recall that the omnibus F-test (Units 9 and 10) is for the significance of the R^2 statistic. Get the p-values and degrees of freedom from your ANOVA table.

ANOVA[d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11905.428	4	2976.357	89.284	.000[a]
	Residual	27035.414	811	33.336		
	Total	38940.842	815			
2	Regression	11905.516	5	2381.103	71.340	.000[b]
	Residual	27035.326	810	33.377		
	Total	38940.842	815			
3	Regression	12089.701	8	1511.213	45.419	.000[c]
	Residual	26851.141	807	33.273		
	Total	38940.842	815			

a. Predictors: (Constant), ESLxSES, Age in Months, English as a 2nd Language, Socioeconomic Status Composite Score

b. Predictors: (Constant), ESLxSES, Age in Months, English as a 2nd Language, Socioeconomic Status Composite Score, Number of Head Start Hours Per Week

c. Predictors: (Constant), ESLxSES, Age in Months, English as a 2nd Language, Socioeconomic Status Composite Score, Number of Head Start Hours Per Week, ESLxHSHSQ, HSHSQ, ESLxHSH

d. Dependent Variable: General Knowledge IRT Scaled Score

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA CHANGE
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT GENERALKNOWLEDGE
  /METHOD=ENTER AGE SES ESL ESLxSES
  /METHOD=ENTER HEADSTARTHOURS
  /METHOD=ENTER  HSHSQ ESLxHSH ESLxHSHSQ
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /SAVE PRED COOK LEVER RESID DRESID.
```

Regression for the Brain Dead

John Willett-icisms (on the snide side):

"You can't fix by analysis what you bungled by design."

"Put garbage in; get garbage out."

"'It's too COMPLICATED not to work!'"

Automated model building is "regression for the brain dead."

Unless you have a Bat Computer, ***DO NOT*** use automated model builders:

http://www.youtube.com/watch?v=wIPHWANYqUQ.

Automated Model Builders ***DO NOT***:

Consider the effects of measurement error. When you have a choice between two variables for the same latent trait, but one has less measurement error, you choose it. The computer chooses strictly based on correlation.

Detect and account for outliers. Outliers can drive correlations, and correlations drive automated model builders.

Detect and correct non-linearity. When we transform to achieve linearity, we "trick" the computer. The computer itself does not know the trick.

Detect and account for heteroskedasticity. The computer does not care about trustworthy standard errors.

Care about your research question. An automated model builder will toss your question predictor from the model as though it were a peripheral control predictor.

Check for interactions. If you have crossproducts among your variables, an automated builder will treat them like any other variable, so if it keeps an interaction term, it may or may not keep the constituent main effects.

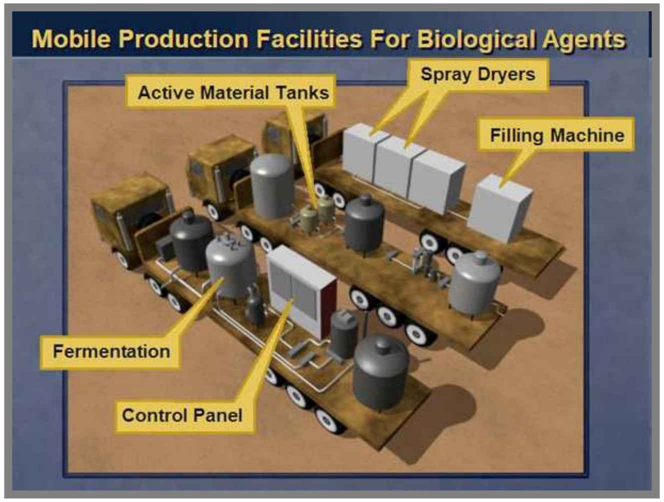
Automated model builders come in different flavors: stepwise, backward forward. (The following descriptions are from the SPSS help menus.)

Stepwise selection. If there are independent variables already in the equation, the variable with the largest probability of F is removed if the value is larger than POUT. The equation is recomputed without the variable and the process is repeated until no more independent variables can be removed. Then, the independent variable not in the equation with the smallest probability of F is entered if the value is smaller than PIN. All variables in the equation are again examined for removal. This process continues until no variables in the equation can be removed and no variables not in the equation are eligible for entry, or until the maximum number of steps has been reached.

Backward elimination. Variables in the block are considered for removal. At each step, the variable with the largest probability-of-F value is removed, provided that the value is larger than POUT.

Forward entry. Variables in the block are added to the equation one at a time. At each step, the variable not in the equation with the smallest probability of F is entered if the value is smaller than PIN.

Data Analytic Fishing Expedition



http://en.wikipedia.org/wiki/Mobile_weapons_laboratory

If we look for anything to support our theory we will always find something (e.g., "Winnebegos of Death"). Therefore, it is not epistemologically valuable to look for anything and find something, because it is a foregone conclusion that we will find something-anything. It is epistemological valuable, however, when we look for something (guided by our theory) and find it. The difference between finding something in general and finding something in particular is all the difference. When we go on a data analytic fishing expedition, we look for any statistically significant relationship with any predictors. The converse is a thoughtful research question that follows from our theory and that we can answer by allowing the data to fall where they may.

The problem with data analytic fishing expeditions lies in our true alpha level. If you will recall, our alpha level is a measure of our tolerance for false positives, and it is customarily < .05. The .05 means that, in expectation, we will reject the null hypothesis exactly 5% of the time when it is true. Thus, if we conduct 100 statistical tests (where unbeknownst to us the null hypothesis is true for each), we will (erringly) conclude that there is a relationship in the population about 5 times (which can mean five published papers—"Tenure, here we come!"). If those 100 statistical tests (where unbeknownst to us the null hypothesis is true) are spread out over an entire research career, then maybe 5 erroneous findings are acceptable. However, if those 100 statistical tests are spread out over a two-hour SPSS jam session between us and our data set, then 5 erroneous findings are 5 too many.

One major assumption for all statistical hypothesis tests is that the sample is a random draw from the population. This is easy to forget, because there is no statistical or graphical check on random sampling. A random sample is representative of the population in expectation, but there will always be idiosyncrasies to any sample—sampling error. If we stare at a sample long enough, or we fit enough models with the sample, we will effectively de-randomize it. We will discover the idiosynchrocies and, inevitably, fixate on them. (Note that the randomness of a sample is an epistemic property, not a metaphysical property. I.e., the random sample is random ***to us***, not random in and of itself. Go ahead and cherry pick the most biased of samples; that sample could have been drawn as a random sample; in fact, we are as likely to randomly draw that biased sample as we are likely to randomly draw any other sample!) We can make a sample non-random just by thinking about it too much; that is, in part, why we are skeptical sometimes of personal annecdotes even when they are based on experiences with many subjects. We are pitter patterned style thinkers, idea tinkers, null sinkers, stat seekers, number tweakers. We see it when we believe it.

Fishing in Ontario

ELSEVIER

Journal of Clinical Epidemiology 59 (2006) 964–969

Journal of Clinical Epidemiology

Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health

Peter C. Austin[a,b,c,*], Muhammad M. Mamdani[a,d], David N. Juurlink[a,e,f], Janet E. Hux[a,c,e,f]

[a]Institute for Clinical Evaluative Sciences, G1 06, 2075 Bayview Avenue, Toronto, Ontario, M4N 3M5 Canada
[b]Department of Public Health Sciences, University of Toronto, Toronto, Ontario, Canada
[c]Department of Health Policy, Management and Evaluation, University of Toronto, Canada
[d]Faculty of Pharmacy, University of Toronto, Canada
[e]Clinical Epidemiology and Health Care Research Program (Sunnybrook & Women's College Site), Canada
[f]Division of General Internal Medicine, Sunnybrook & Women's College Health Sciences Centre and the University of Toronto, Canada

Accepted 19 January 2006

Abstract

Objectives: To illustrate how multiple hypotheses testing can produce associations with no clinical plausibility.

Study Design and Setting: We conducted a study of all 10,674,945 residents of Ontario aged between 18 and 100 years in 2000. Residents were randomly assigned to equally sized derivation and validation cohorts and classified according to their astrological sign. Using the derivation cohort, we searched through 223 of the most common diagnoses for hospitalization until we identified two for which subjects born under one astrological sign had a significantly higher probability of hospitalization compared to subjects born under the remaining signs combined ($P < 0.05$).

Results: We tested these 24 associations in the independent validation cohort. Residents born under Leo had a higher probability of gastrointestinal hemorrhage ($P = 0.0447$), while Sagittarians had a higher probability of humerus fracture ($P = 0.0123$) compared to all other signs combined. After adjusting the significance level to account for multiple comparisons, none of the identified associations remained significant in either the derivation or validation cohort.

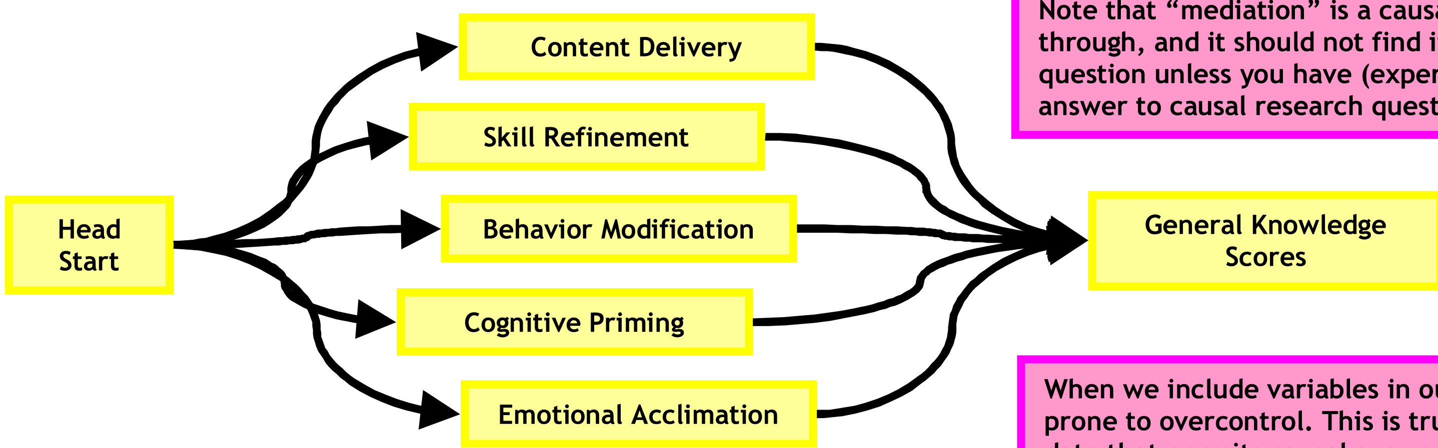
Conclusions: Our analyses illustrate how the testing of multiple, non-prespecified hypotheses increases the likelihood of detecting implausible associations. Our findings have important implications for the analysis and interpretation of clinical studies.

Keywords: Subgroup analyses; Multiple comparisons; Hypothesis testing; Astrology; Data mining; Statistical methods

http://www.centrochranedobrasil.org.br/mestradoProfissional/aulas/29junho/Testing_multiple_statistical_hypotheses_resulted_in_spurious_association_-_a_study_of_astrological_signs_and_health.pdf

Overcontrolling and Fishing (and Mediating)

We overcontrol when we want to observe a causal relationship between a predictor and an outcome but we unwittingly include a mediating variable as a control. Causality can be thought of as a chain, and a mediator is a middle link in that chain. If A causes B causes C causes D causes E causes F, it is perfectly appropriate to say that A causes F, but it is mediated by B, C, D and E. For example, if Head Start causes higher general knowledge scores it does so though mediators such as content delivery, skill refinement, behavior modification, cognitive priming, emotional acclimation etc. If we are using a statistical model to estimate the causal effect of Head Start on general knowledge scores, we will drive our effect size to zero if we (over)control for all the moderating variables.



Note that "mediation" is a causal concept through and through, and it should not find its way into your research question unless you have (experimental) data that permits an answer to causal research questions.

When we include variables in our model willy nilly, we are prone to overcontrol. This is true even when we do not have data that permits causal answers to causal research questions, because even though our research questions are not causal, they almost always address causal theories.

Aristotle's Four Causes, Where a Cause is Just an Answer to A "Why?" Question

"Why does the bathroom-light switch turn on the bathroom light?"

Final Cause: The bathroom-light switch turns on the bathroom light because we need to see in order to do our bathroom business tidily, so the light switch was designed by us so that we can turn on the bathroom light when it is dark.

Formal Cause: The bathroom light switch turns on the bathroom-light because, if it did not, it would not, by definition, be the bathroom light switch.

Material Cause: The bathroom-light switch turns on the bathroom light because it is part of an electrical circuit composed of a control toggle, conductive copper, and lighting filaments.

Efficient Cause: The bathroom-light switch turns on the bathroom light because it sets off a chain of events that begins with toggling the switch which allows electricity to circulate through the copper wire which heats the lighting filaments which makes a hot glow.

BTW: Moderation is (according to Preacher) simply a statistical interaction, although (he qualifies) that some researchers have a somewhat tighter definition.

Model Building Strategy From The Start

Before you look at your data…

If you are collecting your own data, before you begin data collection…

Sketch out a model building strategy, either question centered or baseline control or something in between, and let that be your yellow brick road.

In reality, you are going to deviate from your yellow brick road, but as is quintessentially the case in statistics, *how far* you deviate matters.

Thus the struggle: you can deviate pretty far from the yellow brick road and nobody will know. Nobody is going to read your data analytic diary to see your original strategic sketch. This is an ethical struggle. It is not an ethical dilemma. An ethical dilemma is when the two ethical choices are difficult to decide between because it is difficult to determine which is (more) right. In our case, it is obvious which is right, but sometimes it is a difficult decision nonetheless because of all the people breathing down our backs who worship the false idol of "$p < .05$" whether "$p < .05$" means $p < .05$ or not.

Dig the Post Hole

Unit 18 Post Hole:

Sketch two model building strategies: a baseline-control strategy and a question-centered strategy.

Baseline-Control Strategy:

Outcome: *GENERALKNOWLEDGE*

Model 1	Model 2	Model 3	Model 4
SES	*SES*	*SES*	*SES*
	ESL	*ESL*	*ESL*
	AGE	*AGE*	*AGE*
		HEADSTARTH	*HEADSTARTH*
			interactions

Question-Centered Strategy:

Outcome: *GENERALKNOWLEDGE*

Model 1	Model 2	Model 3	Model 4
HEADSTARTH	*HEADSTARTH*	*HEADSTARTH*	*HEADSTARTH*
	SES	*SES*	*SES*
		ESL	*ESL*
		AGE	*AGE*
			interactions

Baseline-Control Strategy

1. Start with a primary control variable or a set of primary control variables.
2. Add your secondary control variables or a set thereof. (As secondary control variables, you may end up removing them from your final model if they are not stat. sig.)
3. Add your question predictor or set of question predictors or question interaction.
4. Check for interactions in order to check your main effects assumptions.

Question-Centered Strategy

1. Start with your question predictor or set of question predictors or question interaction.
2. Add your primary control variable or a set of primary control variables.
3. Add your secondary control variables or a set thereof. (As secondary control variables, you may end up removing them from your final model if they are not stat. sig.)
4. Check for interactions in order to check your main effects assumptions.

http://calamitykim.typepad.com/calamity_kim/2009/01/follow-the-yellow-brick-road.html

A Hypothesized Interaction (Part I of V)

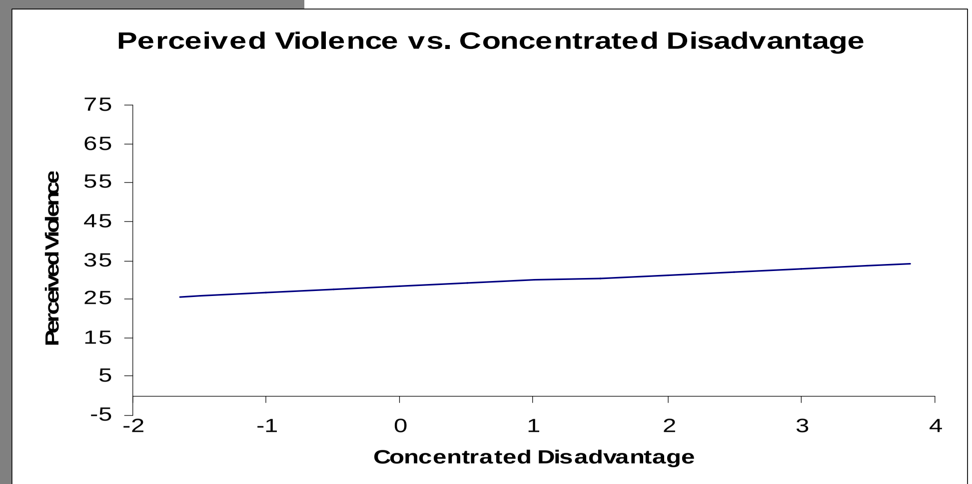
Research Question: Does collective efficacy moderate the relationship between violence and disadvantage in Chicago neighborhoods? Specifically, does the positive relationship between violence and disadvantage disappear when collective efficacy is high?

Table 1. Parameter estimates, (standard errors), approximate p values and goodness-of-fit tests for a nested taxonomy of regression models that describe the relationship between perceived violence and concentrated disadvantage controlling for collective efficacy in Chicago neighborhoods (n=342)

	Models: M1
Intercept	28.2*** (0.7)
Concentrated Disadvantage	1.6* (0.7)
Collective Efficacy	
Disadvan/Efficacy Interaction	
Disadvan/Disadvan Interaction	
R^2	0.01*
df(Residual)	340
ΔR^2	
df (ΔR^2)	

Key: ~p <.10 *p<.05; **p<.01; ***p<.001

Notice that I build the direction into my relationship. Because of the hypothesized interaction, it's a complex direction, but that is all the more interesting! I am looking for something in (very) particular, so if I find it, I will have given my theory great support.



$$P\hat{erc}Viol = 28.2 + 1.6ZDisadv$$

A Hypothesized Interaction (Part II of V)

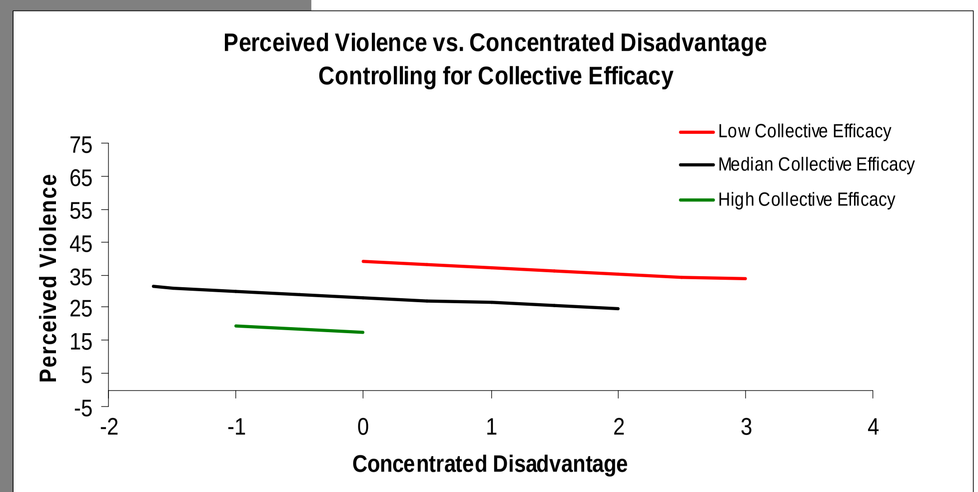
Research Question: Does collective efficacy moderate the relationship between violence and disadvantage in Chicago neighborhoods? Specifically, does the positive relationship between violence and disadvantage disappear when collective efficacy is high?

Table 1. Parameter estimates, (standard errors), approximate p values and goodness-of-fit tests for a nested taxonomy of regression models that describe the relationship between perceived violence and concentrated disadvantage controlling for collective efficacy in Chicago neighborhoods (n=342)

	Models	
	M1	**M2**
Intercept	28.2*** (0.7)	28.2*** (0.7)
Concentrated Disadvantage	1.6* (0.7)	-1.8* (0.9)
Collective Efficacy		-5.4*** (0.9)
Disadvan/Efficacy Interaction		
Disadvan/Disadvan Interaction		
R^2	0.01*	0.11***
df(Residual)	340	339
ΔR^2		0.10***
df (ΔR^2)		1

Key: ~p <.10 *p<.05; **p<.01; ***p<.001

Model 2 is interesting because, upon statistical control of collective efficacy, the direction of the violence/disadvantage relationship gets reversed, statistically significantly.



$$P\hat{erc}Viol = 28.2 - 1.8ZDisadv - 5.4ZCollEff$$

A Hypothesized Interaction (Part III of IV)

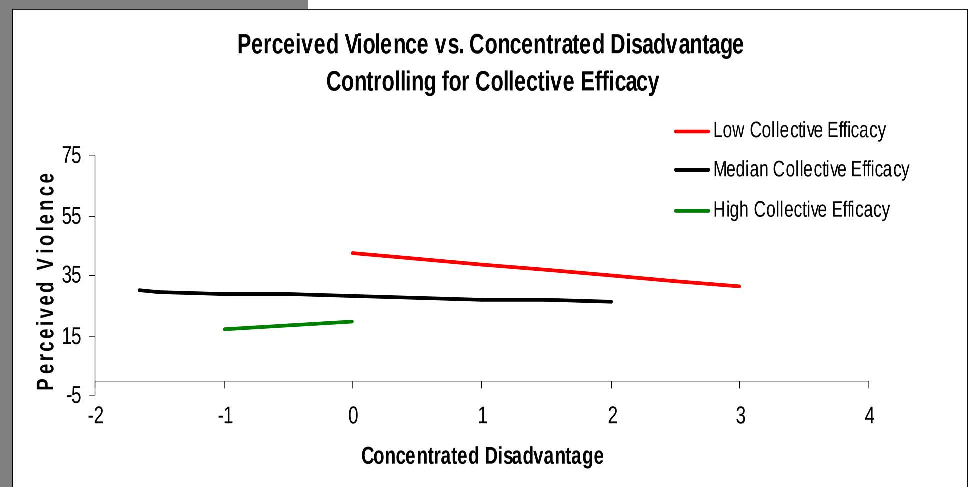
Research Question: Does collective efficacy moderate the relationship between violence and disadvantage in Chicago neighborhoods? Specifically, does the positive relationship between violence and disadvantage disappear when collective efficacy is high?

Table 1. Parameter estimates, (standard errors), approximate p values and goodness-of-fit tests for a nested taxonomy of regression models that describe the relationship between perceived violence and concentrated disadvantage controlling for collective efficacy in Chicago neighborhoods (n=342)

	Models		
	M1	M2	M3
Intercept	28.2*** (0.7)	28.2*** (0.7)	29.1*** (0.8)
Concentrated Disadvantage	1.6* (0.7)	-1.8* (0.9)	-0.9 (1.0)
Collective Efficacy		-5.4*** (0.9)	-4.9*** (0.9)
Disadvan/Efficacy Interaction			1.4~ (0.7)
Disadvan/Disadvan Interaction			
R^2	0.01*	0.11***	0.12***
df(Residual)	340	339	338
ΔR^2		0.10***	0.01~
df (ΔR^2)		1	1

Key: ~p <.10 *p<.05; **p<.01; ***p<.001

Holy crap! This is the exact opposite relationship from the one I predicted. If I reverse my hypothesis now and submit for publication as though my theory predicted this, then shame on me. This is nevertheless really interesting. Maybe somebody will let me publish it anyway.



$$Per\hat{c}Viol = 29.1 - 1.0ZDisadv - 4.8ZCollEff + 1.4ZDisadv * ZCollEff$$

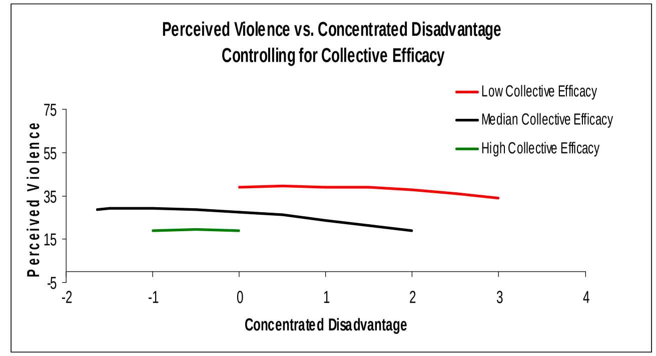
A Hypothesized Interaction (Part IV of V)

Research Question: Does collective efficacy moderate the relationship between violence and disadvantage in Chicago neighborhoods? Specifically, does the positive relationship between violence and disadvantage disappear when collective efficacy is high?

Table 1. Parameter estimates, (standard errors), approximate p values and goodness-of-fit tests for a nested taxonomy of regression models that describe the relationship between perceived violence and concentrated disadvantage controlling for collective efficacy in Chicago neighborhoods (n=342)

	Models			
	M1	M2	M3	M4
Intercept	28.2*** (0.7)	28.2*** (0.7)	29.1*** (0.8)	29.2*** (0.8)
Concentrated Disadvantage	1.6* (0.7)	-1.8* (0.9)	-0.9 (1.0)	-0.7 (1.1)
Collective Efficacy		-5.4*** (0.9)	-4.9*** (0.9)	-5.1*** (0.9)
Disadvan/Efficacy Interaction			1.4~ (0.7)	(Note that comparison statistics are with M2.)
Disadvan/Disadvan Interaction				-1.0~ (0.5)
R^2	0.01*	0.11***	0.12***	0.12***
df(Residual)	340	339	338	338
ΔR^2		0.10***	0.01~	0.01~
df (ΔR^2)		1	1	1

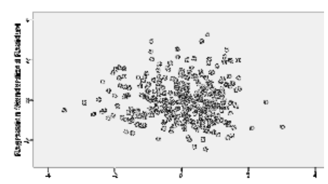
Key: ~p<.10 *p<.05; **p<.01; ***p<.001

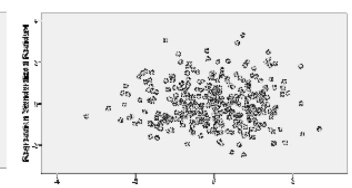


$\hat{PercViol} = 29.2 - 0.7ZDisadv - 5.1ZCollEff - 1.0ZDisadv * ZDisadv$

From a purely statistical perspective, Model 3 is just as good as Model 4. (Compare the residuals below.) From a data analytic perspective, however, Model 3 is better than Model 4 because it answers our research question!

Nearly Identical RVF Plots for M3 and M4





A Hypothesized Interaction (Part V of V)

Research Question: Does collective efficacy moderate the relationship between violence and disadvantage in Chicago neighborhoods? Specifically, does the positive relationship between violence and disadvantage disappear when collective efficacy is high?

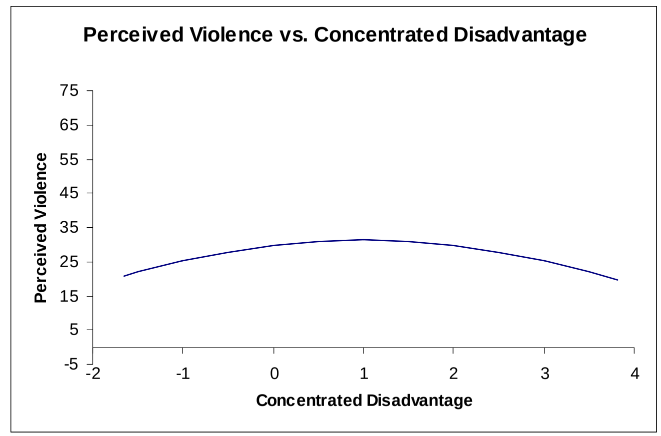
Table 1. Parameter estimates, (standard errors), approximate p values and goodness-of-fit tests for a nested taxonomy of regression models that describe the relationship between perceived violence and concentrated disadvantage controlling for collective efficacy in Chicago neighborhoods (n=342)

	Models				
	M1	M2	M3	M4	M5
Intercept	28.2*** (0.7)	28.2*** (0.7)	29.1*** (0.8)	29.2*** (0.8)	29.0*** (0.9)
Concentrated Disadvantage	1.6* (0.7)	-1.8* (0.9)	-0.9 (1.0)	-0.7 (1.1)	-0.6 (1.1)
Collective Efficacy		-5.4*** (0.9)	-4.9*** (0.9)	-5.1*** (0.9)	-4.9*** (0.9)
Disadvan/Efficacy Interaction			1.4~ (0.7)	(Note that comparison statistics are with M2.)	0.9 (1.0)
Disadvan/Disadvan Interaction				-1.0~ (0.5)	-0.6 (0.6)
R^2	0.01*	0.11***	0.12***	0.12***	0.12***
df(Residual)	340	339	338	338	337
$\bar{D}R^2$		0.10***	0.01~	0.01~	0.00
df ($\bar{D}R^2$)		1	1	1	1

Key: ~p <.10 *p<.05; **p<.01; ***p<.001

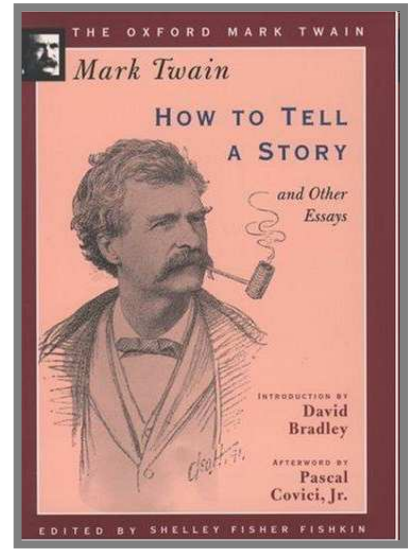
The interaction term and the quadratic term appear to be collinear. They do basically the same predictive work.

For the final presentation, we should boot Models 4 and 5 from our taxonomy. They are pedagogically interesting, but they are not going to help us tell our (sad, but interesting) story.



Even the graph is sad.

Once Upon a Time…



Here is my one story-telling suggestion: find a punch line to guide your discussion of your table of fitted models. The punch line will probably be the insight that lead you to adopt your model-building strategy in the first place. However, (another John Willetticism) nobody wants to read your data analytic diary of what you did last summer! Condense your story. Make it snappy. Hence, the punch line analogy. Except, where punch lines go at the end of jokes, lead off with your punch line and hammer it home repeatedly. Yes, repeat your joke early and often, and feel free to laugh at your own joke. It's your take-home message. Your elevator speech. It's your story.

- Possible Punch Line #1: "Our final model is complex, but it would be inadequate if it were any simpler," and then point out the inadequacies of the earlier models.
- Possible Punch Line #2: "Our question predictor is doing most/little of the predictive work," and provide evidence by tracking the R-square statistic to convey a sense of how much predictive work your question predictor is accomplishing relative to the control predictors.
- Possible Punch Line #3: "Our question predictor's magnitude changes upon statistical control," and provide evidence by tracking the magnitude of the question predictor, which is a great option when it changes drastically.
- Possible Punch Line #4: "Our models systematically address third-variable objections."

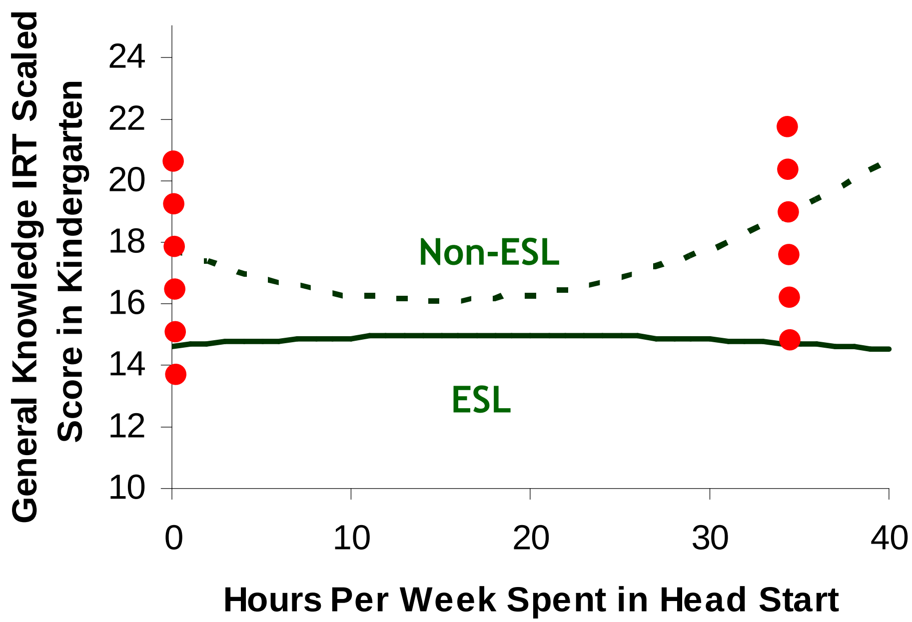
➢ Tip: Use your correlation matrices (simple and partial) to set up your story. This is where foreshadowing comes into play.

So We Have A Responsibly Built Model… Custom GLM Testing

$$GENERALKNOWLEDGE = \beta_0 + \beta_1 HEADSTARTHOURS + \beta_2 SES + \beta_3 ESL + \beta_4 AGE + \beta_5 ESLxSES + \beta_6 HSHSQ + \beta_7 ESLxHSH + \beta_8 ESLxHSHSQ + \varepsilon$$

$$GENERAL\hat{K}NOWLEDGE = -2.5 - .2HEADSTARTHOURS + 4.4SES - 4.4ESL + .4AGE - 1.7ESLxSES + .007HSHSQ + .26ESLxHSH - .008ESLxHSHSQ$$

Figure 17.3. A plot of prototypical fitted values depicting the quadratic effect of *HEADSTARTHOURS* in predicting *GENERALKNOWLEDGE* with trend lines for *ESL*, controlling for *AGE* (held constant at the mean, 64 months) and *SES* (held constant 1 standard deviation below the mean to better represent the Head Start population) (n = 816)



Once we have a model, we can use it to test our custom hypotheses. Up to this point, we have had very limited control over our hypothesis tests. In simple regression models, our usual hypothesis test was

H_0: B_1 = 0

In MR models, our usual hypothesis test was

H_0: B_1 = 0, controlling for all other predictors

Now that we have a final model, however, the sky is the limit with hypothesis testing. Two of the many types of testable hypotheses are

H_0: B_1 = B_2

H_0: B_1 = 4.2

H_0: B_1 = B_2 and B_1 = 4.2, simultaneously

In the following link, UCLA tells us how:
http://www.ats.ucla.edu/stat/Spss/faq/glm_hypothesis_testing.htm.

We may need help from a statistician to get the programming right (there's no simple dropdown menu), but from this course we'll know what to ask.

Simply put, we can isolate any two points or lines on our plot of prototypical fitted values and ask, "Are they statistically significantly different from one another?"

Controlling for Age and SES, do native English speakers who attended Head Start thirty-five hours per week score higher than native English speakers who did not attend Head Start at all? (Based on our sample estimates, "yes," but there are confidence intervals we need to consider if we want to draw conclusions about the population!)

Custom GLM Testing (Example I, Part I of II)

$$GENERALKNOWLEDGE = \beta_0 + \beta_1 HEADSTARTHOURS + \beta_2 SES + \beta_3 ESL + \beta_4 AGE + \beta_5 ESLxSES + \beta_6 HSHSQ + \beta_7 ESLxHSH + \beta_8 ESLxHSHSQ + \varepsilon$$

Our null hypothesis is that, in the population, controlling for age and SES, native English speakers who attended Head Start for thirty-five hours per week score the same as native English speakers who did not attend Head Start at all. We can put this in the language of our equation:

$$[GENERALKNOWLEDGE \mid HEADSTARTHOURS = 35, ESL = 0, SES, AGE] = [GENERALKNOWLEDGE \mid HEADSTARTHOURS = 0, ESL = 0, SES, AGE]$$

$$\beta_0 ?+ \beta_1 ?+ \beta_2 ?+ \beta_3 ?\ \beta_4 ?+ \beta_5 ?+ \beta_6 ?+ \beta_7 ?+ \beta_8 ?+ \varepsilon = \beta_0 ?+ \beta_1 ?+ \beta_2 ?+ \beta_3 ?\ \beta_4 ?+ \beta_5 ?+ \beta_6 ?+ \beta_7 ?+ \beta_8 ?+ \varepsilon$$

By our equation, ***GENERALKNOWLEDGE*** is just a bunch of betas (i.e. parameter estimate) each times a something (i.e., predictor). The trick is to find the right somethings:

$$GENERALKNOWLEDGE = \boxed{\beta_0} + \beta_1 \boxed{HEADSTARTHOURS} + \boxed{\beta_2 SES} + \beta_3 \boxed{ESL} + \boxed{\beta_4 AGE} + \beta_5 \boxed{ESL}xSES + \beta_6 \boxed{HSHSQ} + \beta_7 \boxed{\boxed{ESL}xHSH} + \beta_8 \boxed{\boxed{ESL}xHSHSQ} + \varepsilon$$

First, forget about age and SES (and the constant), because we are holding them constant

ESL = 0

We want to set up a null hypothesis where 35 *HEADSTARTHOURS* equals 0 *HEADSTARTHOURS*.

$$\beta_1 \boxed{35} + \beta_3 \boxed{0} + \beta_5 \boxed{0} xSES + \beta_6 \boxed{35x35} + \beta_7 \boxed{\boxed{0}x35} + \beta_8 \boxed{\boxed{0}x35x35} = \beta_1 \boxed{0} + \beta_3 \boxed{0} + \beta_5 \boxed{0} xSES + \beta_6 \boxed{0} + \beta_7 \boxed{\boxed{0}x0} + \beta_8 \boxed{\boxed{0}x0}$$

$$\beta_1 35 + \beta_3 0 + \beta_5 0xSES + \beta_6 35x35 + \beta_7 0x35 + \beta_8 0x35x35 = \beta_1 0 + \beta_3 0 + \beta_5 0xSES + \beta_6 0 + \beta_7 0x0 + \beta_8 0x0$$

$$H_0 : \beta_1 35 + \beta_6 1225 = 0$$

Because we have zero on the right-hand side of our equation, our k matrix is a zero matrix, which means we only have to worry about our l matrix. Nice.

Custom GLM Testing (Example I, Part II of II)

```
glm GENERALKNOWLEDGE with HEADSTARTHOURS SES ESL AGE  ESLxSES  HSHSQ ESLxHSH ESLxHSHSQ
/print=parameter
/design=HEADSTARTHOURS SES ESL AGE  ESLxSES  HSHSQ ESLxHSH ESLxHSHSQ
/lmatrix = '35-0 Hours Difference, EFL' HEADSTARTHOURS 35 HSHSQ 1225.
```

$$H_0 : \beta_1 35 + \beta_6 1225 = 0$$

Contrast Results (K Matrix)[a]

Contrast			Dependent Variable: General Knowledge IRT Scaled Score
L1	Contrast Estimate		1.295
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		1.295
	Std. Error		1.269
	Sig.		.308
	95% Confidence Interval for Difference	Lower Bound	-1.195
		Upper Bound	3.785

a. Based on the user-specified contrast coefficients (L') matrix: 35-0 Hours Difference, EFL

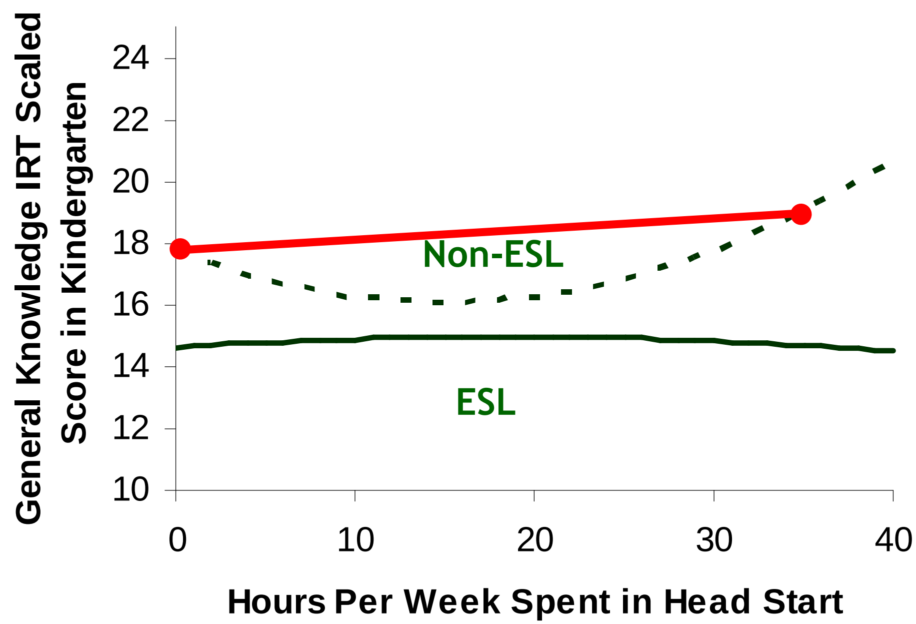
Test Results

Dependent Variable:General Knowledge IRT Scaled Score

Source	Sum of Squares	df	Mean Square	F	Sig.
Contrast	34.652	1	34.652	1.041	.308
Error	26851.141	807	33.273		

From our sample of 816 Latina kindergartners, we estimate a statistically insignificant difference of 1.3 general knowledge points between students with 35 hours of Head Start and students with zero hours of Head Start (p = .308).

Figure 17.3. A plot of prototypical fitted values depicting the quadratic effect of *HEADSTARTHOURS* in predicting *GENERALKNOWLEDGE* with trend lines for *ESL*, controlling for *AGE* (held constant at the mean, 64 months) and *SES* (held constant 1 standard deviation below the mean to better represent the Head Start population) (n = 816)



Custom GLM Testing (Example II, Fishing Expedition)

```
glm GENERALKNOWLEDGE with HEADSTARTHOURS SES ESL AGE  ESLxSES  HSHSQ ESLxHSH ESLxHSHSQ
/print=parameter
/design=HEADSTARTHOURS SES ESL AGE  ESLxSES  HSHSQ ESLxHSH ESLxHSHSQ
/lmatrix = '35-18 Hours Difference, EFL' HEADSTARTHOURS 17 HSHSQ 901.
```

Contrast Results (K Matrix)[a]

Contrast			Dependent ... General Knowledge IRT Scaled Score
L1	Contrast Estimate		2.870
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		2.870
	Std. Error		1.359
	Sig.		.035
	95% Confidence Interval for Difference	Lower Bound	.202
		Upper Bound	5.538

a. Based on the user-specified contrast coefficients (L') matrix: 35-18 Hours Difference, EFL

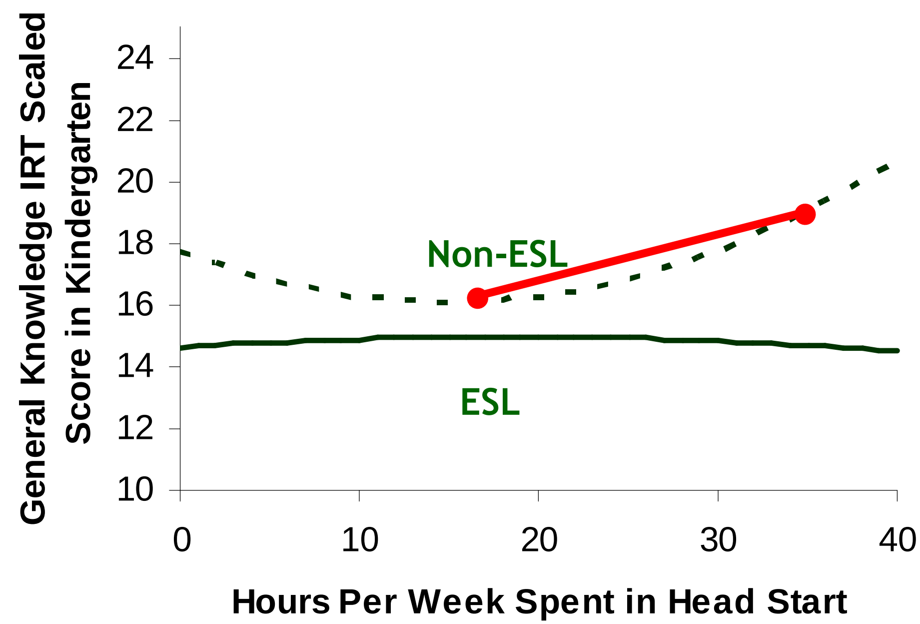
Test Results

Dependent Variable:General Knowledge IRT Scaled Score

Source	Sum of Squares	df	Mean Square	F	Sig.
Contrast	148.324	1	148.324	4.458	.035
Error	26851.141	807	33.273		

We estimate a statistically significant 2.9 point difference from our sample of 816 Latina kindergartners between students with 35 hours of Head Start and 18 hours of Head Start (p = .035).

Figure 17.3. A plot of prototypical fitted values depicting the quadratic effect of *HEADSTARTHOURS* in predicting *GENERALKNOWLEDGE* with trend lines for *ESL*, controlling for *AGE* (held constant at the mean, 64 months) and *SES* (held constant 1 standard deviation below the mean to better represent the Head Start population) (n = 816)



Answering our Roadmap Question

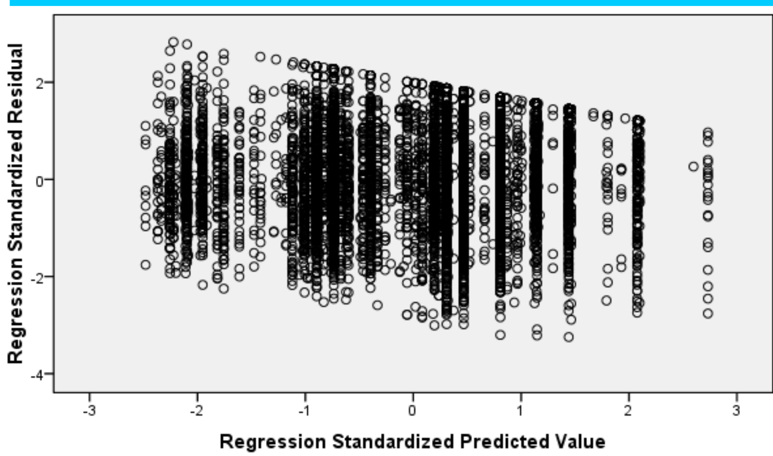
Unit 18: What are sensible strategies for building complex statistical models from scratch?

Final Model:

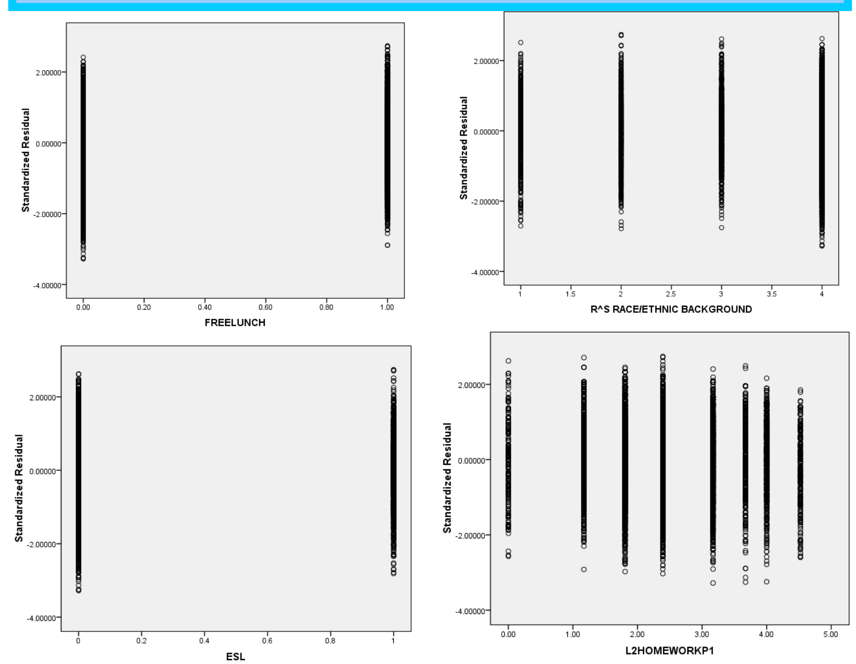
Whenever we write "final model" there is always an implicit "???" after it.

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

A residual versus fitted (RVF) plot is also known as a <u>residual versus predicted plot</u>.



When we plot our residuals versus a predictor, it is knows as a <u>residual versus predictor plot</u>.



Before you finalize your model, check your residual versus predict<u>or</u> plots in addition to your residual versus predict<u>ed</u> plot. You never know.

Comparing Regression Models

Table 3. Comparison of regression models predicting 8[TH] grade IRT scaled reading scores for a nationally representative sample of students (n = 7,800)

	Models			
	M1	M2	M3	M4
Intercept	48.33*** (0.11) 438.24	45.38*** (0.28) 156.56	45.36*** (0.28) 159.53	45.36*** (0.29) 157.57
ASIAN	1.03** (0.38) 2.70	0.73~ (0.38) 1.93	0.46 (0.44) 1.05	-0.38 (0.67) -0.56
BLACK	-4.89*** (0.34) -14.42	-4.80*** (0.33) -14.41	-3.62*** (0.33) -10.96	-3.44*** (0.50) -6.92
LATINO	-4.42*** (0.30) -14.45	-4.12*** (0.30) -13.72	-3.31*** (0.37) -9.04	-2.78*** (0.52) -5.37
Log2 (HOMEWORK HOURS + 1)		1.77*** (0.10) 17.25	1.60*** (0.10) 15.97	1.59*** (0.10) 15.87
ESL			-0.22 (0.36) 0.60	-0.88 (0.64) -1.37
FREELUNCH			-3.87*** (0.20) -19.45	-3.57*** (0.24) -15.21
ESL x ASIAN				3.25** (1.00) 3.25
ESL x BLACK				5.87** (1.89) 3.12
ESL x LATINO				0.45 (0.86) 0.52
FREELUNCH x ASIAN				-2.77** (0.85) -3.25
FREELUNCH x BLACK				-0.75 (0.67) -1.13
FREELUNCH x LATINO				-0.44 (0.60) -0.72
R^2	0.045	0.084	0.126	0.130
F	133.66	178.48	187.83	96.64
(df)	(3, 7796)	(4, 7795)	(6, 7793)	(12, 7787)
p	< 0.001	< 0.001	< 0.001	< 0.001

Cell entries are estimated regression coefficients, (standard errors) and t-statistics.

Story?

Story?

Story?

Story?

Story?

What I really want to do is to detect reading gaps between White students and minority students. I can do that in Models 1 through 3 fairly easily, but in Model 4 it gets tricky because of the interactions. In Model 4, the gaps differ by ESL and free lunch status.

As soon as I start comparing White students to Asian, Black and Latino students depending on their ESL or free lunch status, I jump from making 3 comparisons to 12 comparisons:

No ESL, No FL: WvA WvB WvL

ESL, No FL: WvA WvB WvL

No ESL, FL: WvA WvB WvL

ESL, FL: WvA WvB WvL

It's Bonferroni time! But, how do I make post hoc adjustments for multiple comparisons in multiple regression? The answer is in generalized linear models.

Note: If you want to look at other gaps, that's fine. Just define which (before you look!), and prepare to make a Bonferroni adjustment according to how many.

What Do Achievement Gaps Mean, Controlling for Homework?

Estimates

FRE ELU NCH	ESL	R^S RACE /ET...	Mean	Std. Error	95% Wald Confidence Interval	
					Lower	Upper
0	0	Asian	48.7937	.08839	48.6204	48.9669
		Latino	46.8447	.07274	46.7021	46.9873
		Black	45.8439	.06075	45.7249	45.9630
		White	49.3504	.01584	49.3193	49.3814
	YES	Asian	51.4432	.06178	51.3221	51.5643
		Latino	45.9113	.06887	45.7763	46.0463
		Black	53.1858	.35363	52.4927	53.8789
		White	48.7675	.09246	48.5863	48.9488
1	0	Asian	43.4677	.18570	43.1037	43.8316
		Latino	42.1925	.08645	42.0231	42.3619
		Black	41.6325	.05077	41.5329	41.7320
		White	45.8048	.02517	45.7555	45.8542
	YES	Asian	44.7619	.10102	44.5639	44.9599
		Latino	42.2920	.05551	42.1832	42.4008
		Black	45.1825	.27736	44.6388	45.7261
		White	44.1680	.14909	43.8758	44.4602

Covariates appearing in the model are fixed at the following values: L2HOMEWORKP1=2.5142

I added a three-way interaction in my model so that the computer could compute these. In regression I shied away because of the gazillion dummies, but in GLM, I simply added factors, and it did the work for me.

These are not simple averages for the subgroups. Rather, they are estimated means, controlling for HOMEWORK.

Recall that it only makes sense to hold *HOMEWORK* constant because it does not interact with any of the "gap" variables!!!

Sifting Through 240 Pairwise Differences (i.e., Achievement Gaps)

Here are all the pairwise differences between Asian Non-Free Lunch Non-ESL students and each of the other 15 subgroups. We are only interested in 12 comparisons of which only 1 is included here. We are interested in the achievement gap between Asian Non-Free Lunch Non-ESL students and White Non-Free Lunch Non-ESL students.

Pairwise Comparisons

(I) FREELUNCH*ESL*RACE	(J) FREELUNCH*ESL*RACE	Mean Difference (I-J)	Std. Error	df	Sig.	95% Wald Confidence Interval for Difference Lower	Upper
[FREELUNCH=.00]*[ESL=0]*[RACE=1]	[FREELUNCH=.00]*[ESL=0]*[RACE=2]	1.9490[a]	.11448	1	.000	1.7246	2.1734
	[FREELUNCH=.00]*[ESL=0]*[RACE=3]	2.9498[a]	.10725	1	.000	2.7396	3.1600
	[FREELUNCH=.00]*[ESL=0]*[RACE=4]	-.5567[a]	.08979	1	.000	-.7327	-.3807
	[FREELUNCH=.00]*[ESL=1]*[RACE=1]	-2.6495[a]	.10782	1	.000	-2.8608	-2.4381
	[FREELUNCH=.00]*[ESL=1]*[RACE=2]	2.8824[a]	.11206	1	.000	2.6628	3.1021
	[FREELUNCH=.00]*[ESL=1]*[RACE=3]	-4.3922[a]	.36450	1	.000	-5.1066	-3.6777
	[FREELUNCH=.00]*[ESL=1]*[RACE=4]	.0261	.12790	1	.838	-.2245	.2768
	[FREELUNCH=1.00]*[ESL=0]*[RACE=1]	5.3260[a]	.20566	1	.000	4.9229	5.7291
	[FREELUNCH=1.00]*[ESL=0]*[RACE=2]	6.6012[a]	.12365	1	.000	6.3588	6.8435
	[FREELUNCH=1.00]*[ESL=0]*[RACE=3]	7.1612[a]	.10194	1	.000	6.9614	7.3610
	[FREELUNCH=1.00]*[ESL=0]*[RACE=4]	2.9889[a]	.09192	1	.000	2.8087	3.1690
	[FREELUNCH=1.00]*[ESL=1]*[RACE=1]	4.0318[a]	.13423	1	.000	3.7687	4.2949
	[FREELUNCH=1.00]*[ESL=1]*[RACE=2]	6.5017[a]	.10439	1	.000	6.2971	6.7063
	[FREELUNCH=1.00]*[ESL=1]*[RACE=3]	3.6112[a]	.29110	1	.000	3.0407	4.1818
	[FREELUNCH=1.00]*[ESL=1]*[RACE=4]	4.6257[a]	.17333	1	.000	4.2859	4.9654

Notice that the difference between Asian Non-Free Lunch Non-ESL students and White Non-Free Lunch ESL students has a p-value of .838 (even before we Bonferroni adjust it). All the other differences have apparent p-values of less than .001, but we aren't looking at all the other differences. We are making a bee-line only to one of the differences. If we are going to pay attention to other p-values (other than our pre-chosen 12) then we have to change our Bonferroni adjustment. We must be honest with ourselves. How many differences are we *really* looking at?

This is 1/16th of the (enormous) output! There are 15 more blocks, one for each of the RACE/LUNCH/ESL combinations. We can expedite the process by jumping to the bocks for White students…

The 12 Achievement Gaps For Our Story

We are looking at twelve differences. All but two of the differences have an unadjusted p-value of ***less than*** .001. The other two have unadjusted p-values of .001. When we Bonferroni adjust for multiple comparisons, we multiply each p-value by 12. Since 0.001*12 = 0.012 < .05, we reject each null hypothesis (that there is no difference in the population) for each difference. Now, if we were to look at all 120 differences, then we would Bonferroni adjust the .001 p-values to .120, and we could not reject the null hypotheses! Hypotheses are like dates. You dance with the belle (or beau) who brought you to the ball. Don't be ogling other hypotheses! Keep your eyes on your hypothesis.

Non-ESL

Non-Free Lunch

[FREELUNCH=.00]*[ESL=0]*[RACE=4]	[FREELUNCH=.00]*[ESL=0]*[RACE=1]	.5567^a	.08979	1	.000	.3807	.7327
	[FREELUNCH=.00]*[ESL=0]*[RACE=2]	2.5057^a	.07445	1	.000	2.3597	2.6516
	[FREELUNCH=.00]*[ESL=0]*[RACE=3]	3.5064^a	.06278	1	.000	3.3834	3.6295
	[FREELUNCH=.00]*[ESL=1]*[RACE=1]	-2.0928^a	.06373	1	.000	-2.2177	-1.9679
	[FREELUNCH=.00]*[ESL=1]*[RACE=2]	3.4391^a	.07069	1	.000	3.3005	3.5776
	[FREELUNCH=.00]*[ESL=1]*[RACE=3]	-3.8355^a	.35397	1	.000	-4.5293	-3.1417
	[FREELUNCH=.00]*[ESL=1]*[RACE=4]	.5828^a	.09379	1	.000	.3990	.7666
	[FREELUNCH=1.00]*[ESL=0]*[RACE=1]	5.8827^a	.18637	1	.000	5.5174	6.2480
	[FREELUNCH=1.00]*[ESL=0]*[RACE=2]	7.1579^a	.08792	1	.000	6.9855	7.3302
	[FREELUNCH=1.00]*[ESL=0]*[RACE=3]	7.7179^a	.05320	1	.000	7.6136	7.8222
	[FREELUNCH=1.00]*[ESL=0]*[RACE=4]	3.5455^a	.02979	1	.000	3.4872	3.6039
	[FREELUNCH=1.00]*[ESL=1]*[RACE=1]	4.5885^a	.10225	1	.000	4.3881	4.7889
	[FREELUNCH=1.00]*[ESL=1]*[RACE=2]	7.0584^a	.05776	1	.000	6.9452	7.1716
	[FREELUNCH=1.00]*[ESL=1]*[RACE=3]	4.1679^a	.27780	1	.000	3.6234	4.7124
	[FREELUNCH=1.00]*[ESL=1]*[RACE=4]	5.1823^a	.14995	1	.000	4.8884	5.4762

Free Lunch

[FREELUNCH=1.00]*[ESL=0]*[RACE=4]	[FREELUNCH=.00]*[ESL=0]*[RACE=1]	-2.9889^a	.09192	1	.000	-3.1690	-2.8087
	[FREELUNCH=.00]*[ESL=0]*[RACE=2]	-1.0399^a	.07696	1	.000	-1.1907	-.8890
	[FREELUNCH=.00]*[ESL=0]*[RACE=3]	-.0391	.06575	1	.552	-.1680	.0898
	[FREELUNCH=.00]*[ESL=0]*[RACE=4]	-3.5455^a	.02979	1	.000	-3.6039	-3.4872
	[FREELUNCH=.00]*[ESL=1]*[RACE=1]	-5.6384^a	.06681	1	.000	-5.7693	-5.5074
	[FREELUNCH=.00]*[ESL=1]*[RACE=2]	-.1065	.07329	1	.146	-.2501	.0372
	[FREELUNCH=.00]*[ESL=1]*[RACE=3]	-7.3810^a	.35456	1	.000	-8.0760	-6.6861
	[FREELUNCH=.00]*[ESL=1]*[RACE=4]	-2.9627^a	.09584	1	.000	-3.1506	-2.7749
	[FREELUNCH=1.00]*[ESL=0]*[RACE=1]	2.3371^a	.18739	1	.000	1.9699	2.7044
	[FREELUNCH=1.00]*[ESL=0]*[RACE=2]	3.6123^a	.08998	1	.000	3.4360	3.7887
	[FREELUNCH=1.00]*[ESL=0]*[RACE=3]	4.1724^a	.05665	1	.000	4.0613	4.2834
	[FREELUNCH=1.00]*[ESL=1]*[RACE=1]	1.0430^a	.10413	1	.000	.8389	1.2470
	[FREELUNCH=1.00]*[ESL=1]*[RACE=2]	3.5128^a	.06090	1	.000	3.3935	3.6322
	[FREELUNCH=1.00]*[ESL=1]*[RACE=3]	.6224^a	.27851	1	.025	.0765	1.1682
	[FREELUNCH=1.00]*[ESL=1]*[RACE=4]	1.6368^a	.15118	1	.000	1.3405	1.9331

ESL

Non-Free Lunch

[FREELUNCH=.00]*[ESL=1]*[RACE=4]	[FREELUNCH=.00]*[ESL=0]*[RACE=1]	-.0261	.12790	1	.838	-.2768	.2245
	[FREELUNCH=.00]*[ESL=0]*[RACE=2]	1.9228^a	.11764	1	.000	1.6923	2.1534
	[FREELUNCH=.00]*[ESL=0]*[RACE=3]	2.9236^a	.11063	1	.000	2.7068	3.1405
	[FREELUNCH=.00]*[ESL=0]*[RACE=4]	-.5828^a	.09379	1	.000	-.7666	-.3990
	[FREELUNCH=.00]*[ESL=1]*[RACE=1]	-2.6756^a	.11117	1	.000	-2.8935	-2.4577
	[FREELUNCH=.00]*[ESL=1]*[RACE=2]	2.8563^a	.11531	1	.000	2.6303	3.0823
	[FREELUNCH=.00]*[ESL=1]*[RACE=3]	-4.4183^a	.36550	1	.000	-5.1347	-3.7019
	[FREELUNCH=1.00]*[ESL=0]*[RACE=1]	5.2999^a	.20744	1	.000	4.8933	5.7065
	[FREELUNCH=1.00]*[ESL=0]*[RACE=2]	6.5750^a	.12660	1	.000	6.3269	6.8232
	[FREELUNCH=1.00]*[ESL=0]*[RACE=3]	7.1351^a	.10549	1	.000	6.9283	7.3418
	[FREELUNCH=1.00]*[ESL=0]*[RACE=4]	2.9627^a	.09584	1	.000	2.7749	3.1506
	[FREELUNCH=1.00]*[ESL=1]*[RACE=1]	4.0057^a	.13693	1	.000	3.7373	4.2741
	[FREELUNCH=1.00]*[ESL=1]*[RACE=2]	6.4756^a	.10786	1	.000	6.2642	6.6870
	[FREELUNCH=1.00]*[ESL=1]*[RACE=3]	3.5851^a	.29236	1	.000	3.0121	4.1581
	[FREELUNCH=1.00]*[ESL=1]*[RACE=4]	4.5995^a	.17545	1	.000	4.2557	4.9434

Free Lunch

[FREELUNCH=1.00]*[ESL=1]*[RACE=4]	[FREELUNCH=.00]*[ESL=0]*[RACE=1]	-4.6257^a	.17333	1	.000	-4.9654	-4.2859
	[FREELUNCH=.00]*[ESL=0]*[RACE=2]	-2.6767^a	.16588	1	.000	-3.0018	-2.3516
	[FREELUNCH=.00]*[ESL=0]*[RACE=3]	-1.6759^a	.16099	1	.000	-1.9914	-1.3604
	[FREELUNCH=.00]*[ESL=0]*[RACE=4]	-5.1823^a	.14995	1	.000	-5.4762	-4.8884
	[FREELUNCH=.00]*[ESL=1]*[RACE=1]	-7.2751^a	.16145	1	.000	-7.5916	-6.9587
	[FREELUNCH=.00]*[ESL=1]*[RACE=2]	-1.7432^a	.16420	1	.000	-2.0651	-1.4214
	[FREELUNCH=.00]*[ESL=1]*[RACE=3]	-9.0178^a	.38383	1	.000	-9.7701	-8.2655
	[FREELUNCH=.00]*[ESL=1]*[RACE=4]	-4.5995^a	.17545	1	.000	-4.9434	-4.2557
	[FREELUNCH=1.00]*[ESL=0]*[RACE=1]	.7003^a	.23814	1	.003	.2336	1.1671
	[FREELUNCH=1.00]*[ESL=0]*[RACE=2]	1.9755^a	.17229	1	.000	1.6378	2.3132
	[FREELUNCH=1.00]*[ESL=0]*[RACE=3]	2.5356^a	.15749	1	.000	2.2269	2.8442
	[FREELUNCH=1.00]*[ESL=0]*[RACE=4]	-1.6368^a	.15118	1	.000	-1.9331	-1.3405
	[FREELUNCH=1.00]*[ESL=1]*[RACE=1]	-.5938^a	.18012	1	.001	-.9469	-.2408
	[FREELUNCH=1.00]*[ESL=1]*[RACE=2]	1.8760^a	.15906	1	.000	1.5643	2.1878
	[FREELUNCH=1.00]*[ESL=1]*[RACE=3]	-1.0144^a	.31491	1	.001	-1.6317	-.3972

Unit 18 Appendix: Key Concepts

Nobody wants to read your data analytic diary of everything you did during winter break.* Make your data analytic story to the point, even if the actual data analysis was meandering. Do not describe every assumption check or every interaction check. The astute reader is looking for clues in your writing not only that you made all the right moves but also that you know which moves are noteworthy. *This is a Willett-icism (more to follow).

- The R2 statistic never justifies a model.
- What justifies a model? 1st, the model answers your research questions. 2nd, it meets third-variable objections. 3rd, it meets the GLM assumptions.

A combination of the question centered strategy and the baseline control strategy is often the best route. The key is *strategy*! You want a model building strategy that tightly unifies your theory, question and variables.

- For the sake of a crisp presentation, make your models as simple as possible, but no simpler. Remove a statistically insignificant predictor unless:
- It is your question predictor.
- It is an important control predictor.
- It is part of stat sig interaction.
- It is part of a stat sig set of dummies.

Note that "mediation" is a causal concept through and through, and it should not find its way into your research question unless you have (experimental) data that permits an answer to causal research questions.

When we include variables in our model willy nilly, we are prone to overcontrol. This is true even when we do not have data that permits causal answers to causal research questions, because even though our research questions are not causal, they almost always address causal theories.

Unit 18 Appendix: Key Concepts

Once we have a model, we can use it to test our custom hypotheses. Up to this point, we have had very limited control over our hypothesis tests. In simple regression models, our usual hypothesis test was

$$H_0: \beta_1 = 0$$

In MR models, our usual hypothesis test was

$$H_0: \beta_1 = 0\text{, controlling for all other predictors}$$

Now that we have a final model, however, the sky is the limit with hypothesis testing. Two of the many types of testable hypotheses are

$$H_0: \beta_1 = \beta_2$$

$$H_0: \beta_1 = 4.2$$

$$H_0: \beta_1 = \beta_2 \text{ and } \beta_1 = 4.2\text{, simultaneously}$$

In the following link, UCLA tells us how: http://www.ats.ucla.edu/stat/Spss/faq/glm_hypothesis_testing.htm.

We may need help from a statistician to get the programming right (there's no simple dropdown menu), but from this course we'll know what to ask.

Simply put, we can isolate any two points or lines on our plot of prototypical fitted values and ask, "Are they statistically significantly different from one another?"

Whenever we write "final model" there is always an implicit "???" after it.

Before you finalize your model, check your residual versus predict<u>or</u> plots in addition to your residual versus predict<u>ed</u> plot. You never know.

Unit 18 Appendix: Key Interpretations

You may notice that the quadratic relationship between Head Start hours and general knowledge IRT scaled scores all but disappears when we control for Age, SES, and ESL.However, when we allow the relationship to vary between speakers of English as a first and second language, we see the quadratic relationship reappear. As it turns out, there is a strong quadratic relationship for speakers of English as a first language and virtually no relationship for speakers of English as a second language.

With our final model, we predict 31% of the variance in general knowledge IRT scaled scores, but head start hours is only a small contributor to that percentage as evidenced by the 3% of variance that our model predicts without the controls.

In Model 2, for which we assume only a main effect, Head Start hours predicts a statistically insignificant proportion of the variation in general knowledge scores (partial $\eta 2 < .001$, $p = .959$).

However, in Model 3, for which we relax the main effects assumption, we find that Head Start hours statistically significantly interacts with itself and ESL. Head Start hours predicts general knowledge scores, but the prediction differs by number of Head Start hours and ESL status. The reader may note that inclusion of the interaction effects increases our R2 statistic by only .005.

From our sample of 816 Latina kindergartners, we estimate a statistically insignificant difference of 1.3 general knowledge points between students with 35 hours of Head Start and students with zero hours of Head Start ($p = .308$).

We estimate a statistically significant 2.9 point difference from our sample of 816 Latina kindergartners between students with 35 hours of Head Start and 18 hours of Head Start ($p = .035$).

Unit 18 Appendix: Key Terminology

- A question centered strategy for multiple regression model building always includes the question predictor in each model.

(1) We begin with a simple linear regression of the outcome on the question predictor.

(2) We address any non-linearity concerns.

(3) We add judiciously chosen control variables to our model, and we may want to add them in waves, for example, a wave of demographic variables, a wave of personal characteristics, and a wave of test scores. (Here we have displayed all the controls in one wave.)

(4) We check for interactions.

(5) We trim statistically insignificant controls from the model, unless they are of strong theoretical importance to us or our audience. We follow rounds (1) and (5) with comprehensive assumption checks. Ideally, we would check our assumptions every step of the way, but there are only so many hours in the day.

- A baseline control strategy for multiple regression model building includes your predictor only after you have established your control model.

(1) We begin by building a control model. In building your control model, you may begin with a strong theory about the variables that influence your outcome but that you are not interesting (for your purposes) per se. Be sure to check for interactions among your control variables.

(2) We include our question predictor.

(3) We check for interactions.

We overcontrol when we want to observe a causal relationship between a predictor and an outcome but we unwittingly include a mediating variable as a control. Causality can be thought of as a chain, and a mediator is a middle link in that chain. If A causes B causes C causes D causes E causes F, it is perfectly appropriate to say that A causes F, but it is mediated by B, C, D and E. For example, if Head Start causes higher general knowledge scores it does so though mediators such as content delivery, skill refinement, behavior modification, cognitive priming, emotional acclimation etc. If we are using a statistical model to estimate the causal effect of Head Start on general knowledge scores, we will drive our effect size to zero if we (over)control for all the moderating variables.

BTW: Moderation is (according to Preacher) simply a statistical interaction, although (he qualifies) that some researchers have a somewhat tighter definition.

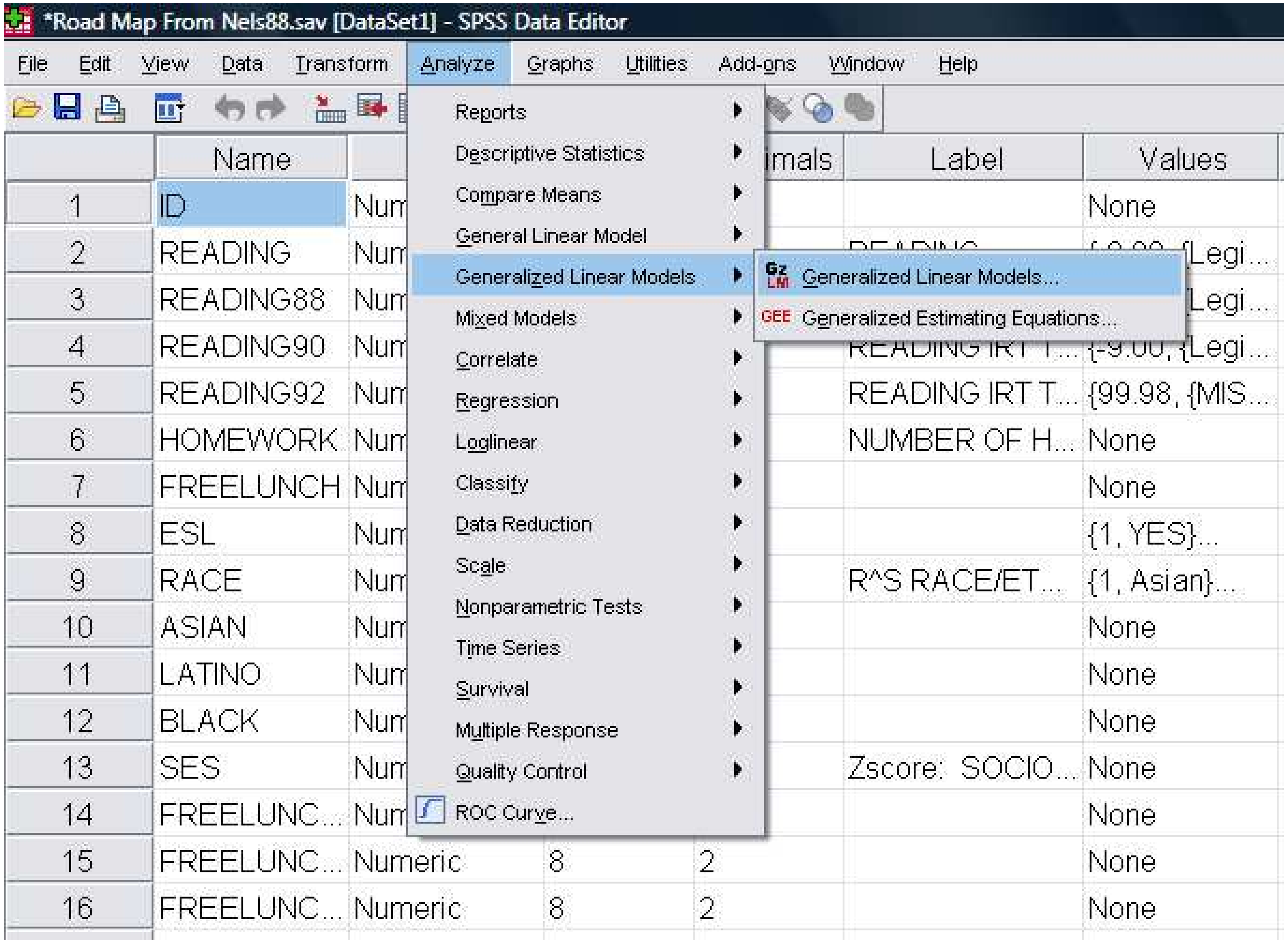
A residual versus fitted (RVF) plot is also known as a residual versus predicted plot.

When we plot our residuals versus a predictor, it is knows as a residual versus predictor plot.

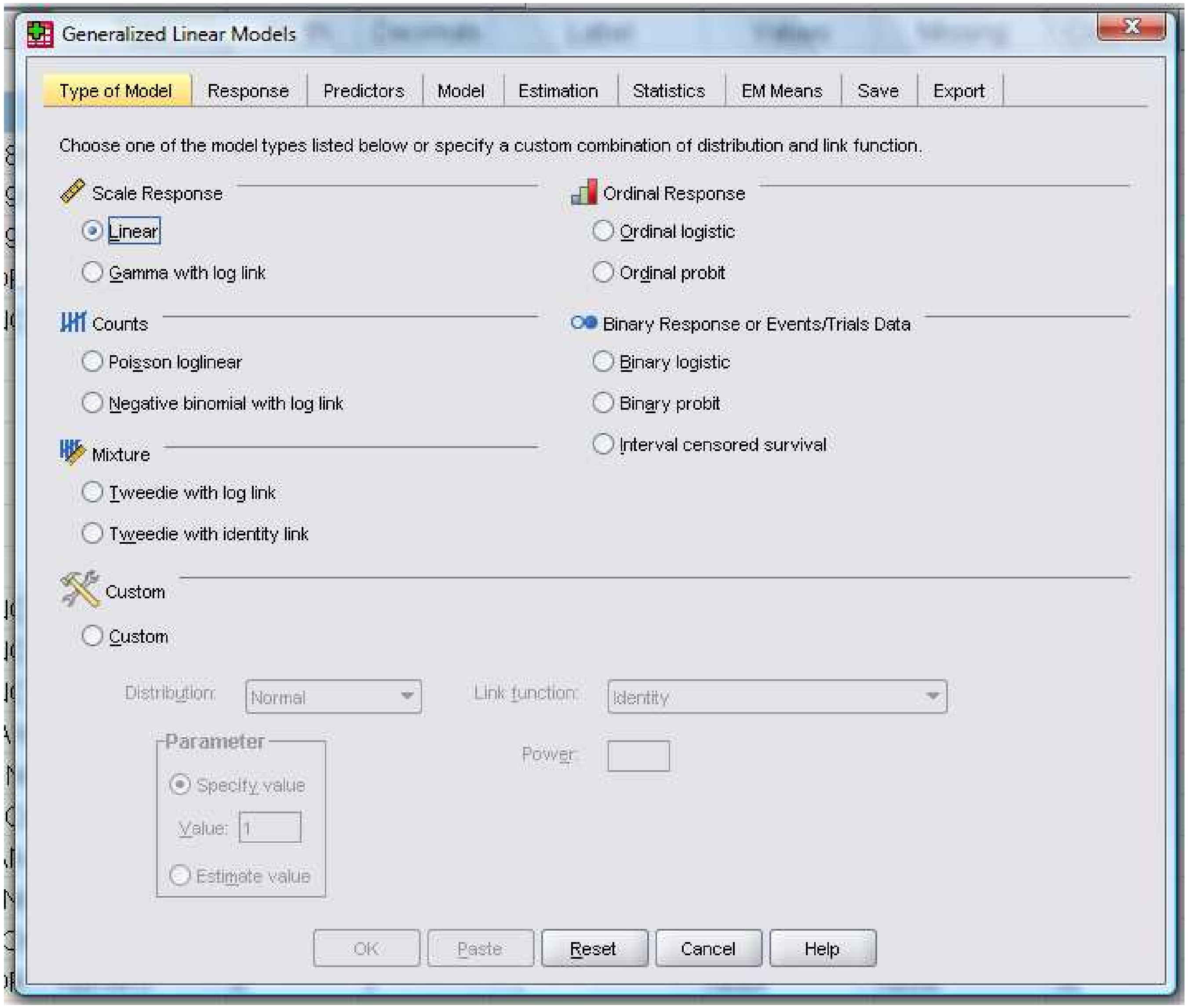
Unit 18 Appendix: SPSS Syntax

```
glm GENERALKNOWLEDGE with HEADSTARTHOURS SES ESL AGE  ESLxSES
  HSHSQ ESLxHSH ESLxHSHSQ
 /print=parameter
 /design=HEADSTARTHOURS SES ESL AGE  ESLxSES  HSHSQ ESLxHSH
  ESLxHSHSQ
 /lmatrix = '35-0 Hours Difference, EFL' HEADSTARTHOURS 35 HSHSQ
  1225.
```

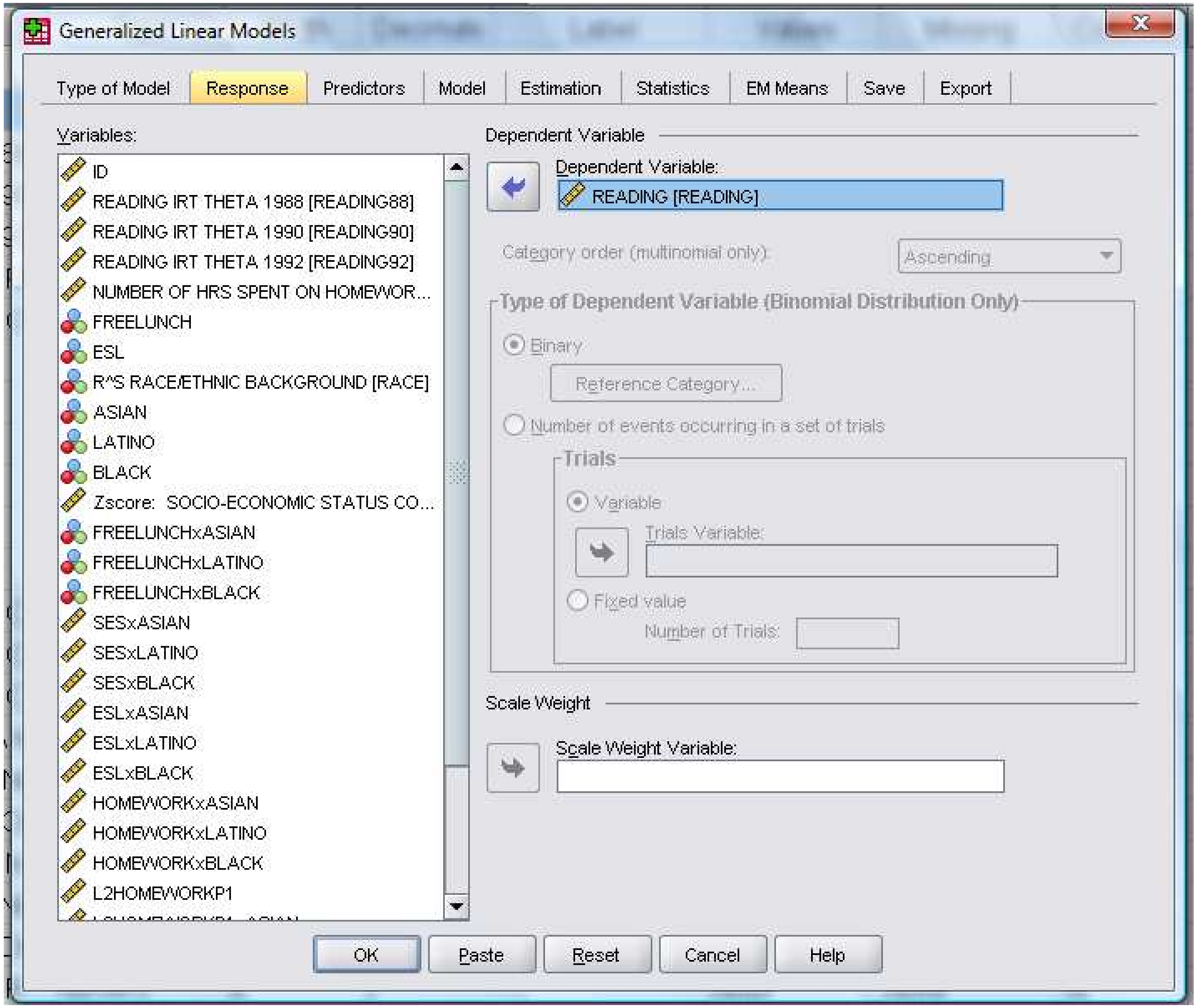
SPSS Walkthrough of Using Generalized Linear Models for Post Hoc Tests



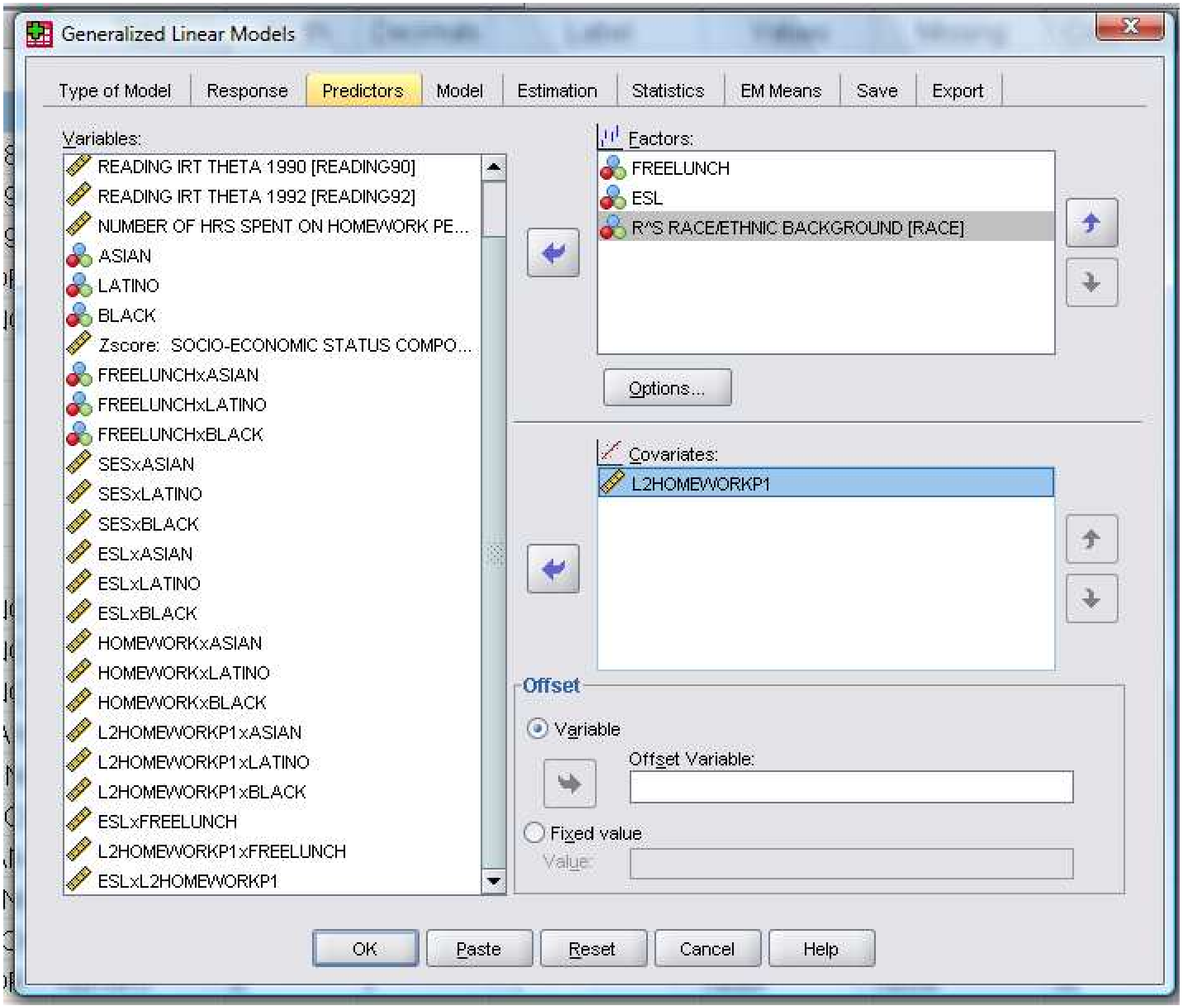
SPSS Walkthrough of Using Generalized Linear Models for Post Hoc Tests



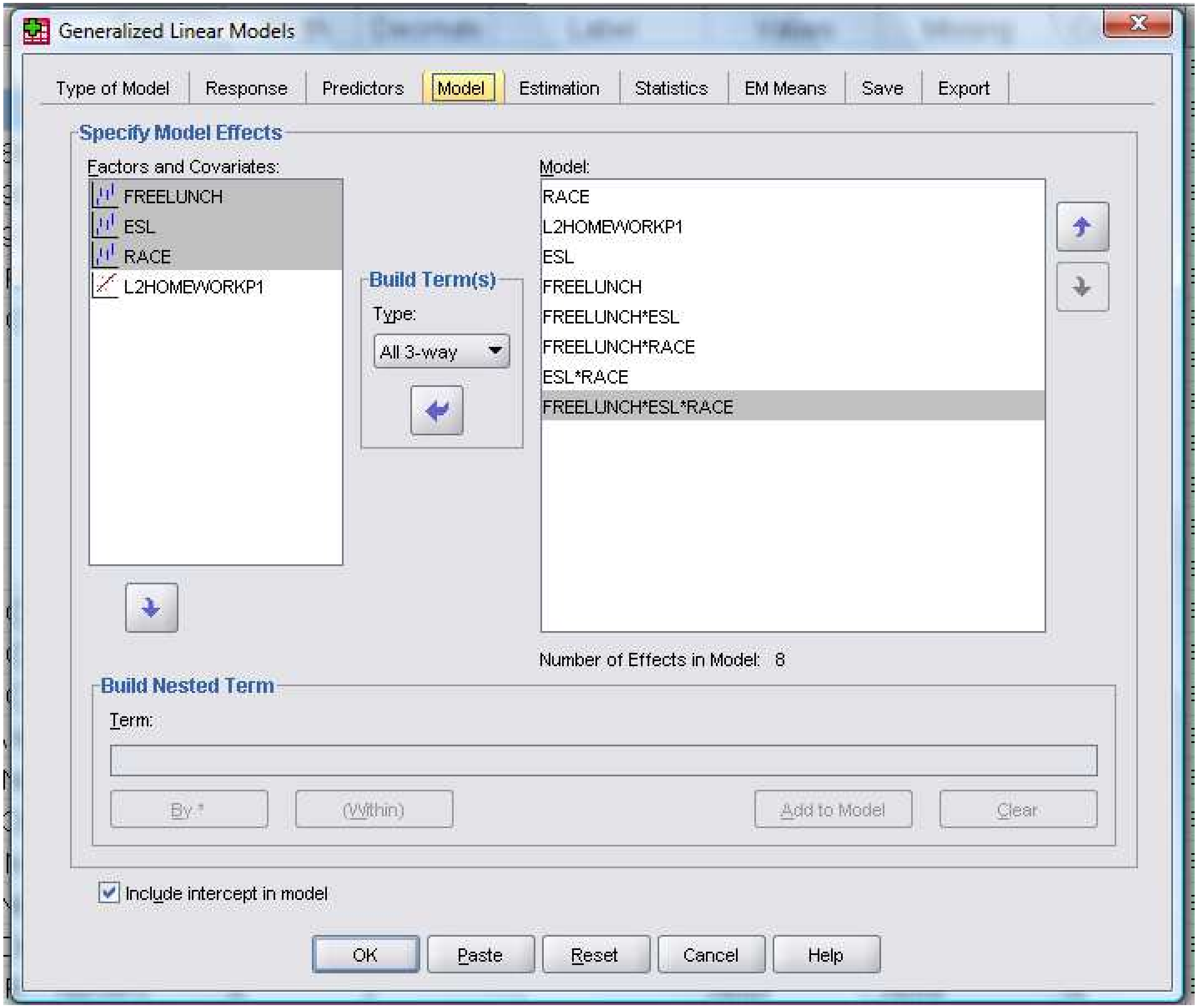
SPSS Walkthrough of Using Generalized Linear Models for Post Hoc Tests



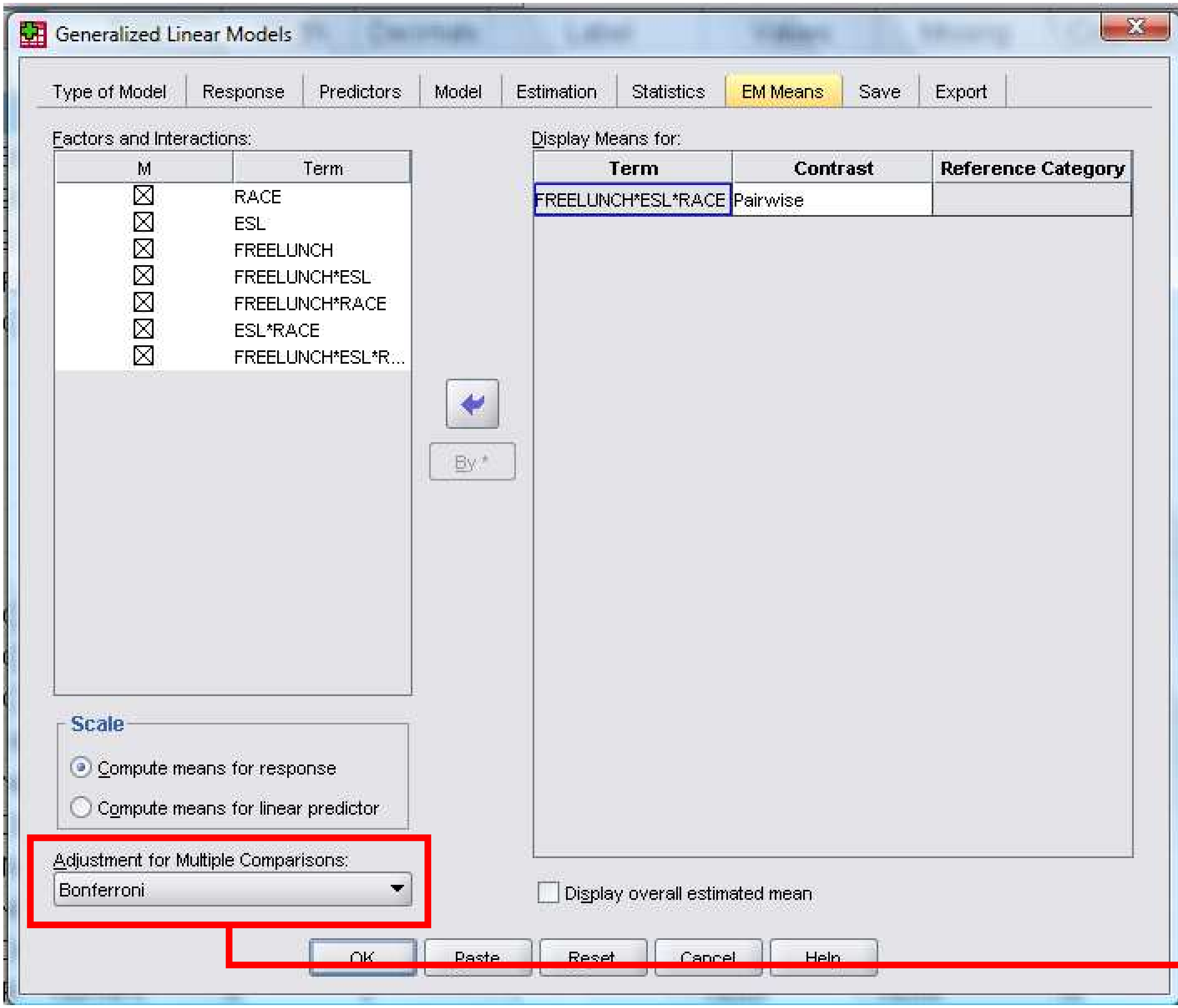
SPSS Walkthrough of Using Generalized Linear Models for Post Hoc Tests



SPSS Walkthrough of Using Generalized Linear Models for Post Hoc Tests



SPSS Walkthrough of Using Generalized Linear Models for Post Hoc Tests



Oh no, it's making me choose a post hoc adjustment! That's actually a bad thing because it will assume I'm comparing every group to every group when I only want to make 12 comparisons, not 120.

http://www.mathsisfun.com/combinatorics/combinations-permutations-calculator.html

I want the p-values so I can adjust them by hand, multiplying them by 12.

SPSS Walkthrough of Using Generalized Linear Models for Post Hoc Tests

```
* Generalized Linear Models.
GENLIN READING BY FREELUNCH ESL RACE (ORDER=ASCENDING) WITH L2HOMEWORKP1
  /MODEL RACE L2HOMEWORKP1 ESL FREELUNCH FREELUNCH*ESL FREELUNCH*RACE ESL*RACE FREELUNCH*ESL*RACE
    INTERCEPT=YES
 DISTRIBUTION=NORMAL LINK=IDENTITY
  /CRITERIA SCALE=1 COVB=MODEL PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD)
    CILEVEL=95 CITYPE=WALD LIKELIHOOD=FULL
  /EMMEANS TABLES=FREELUNCH*ESL*RACE SCALE=ORIGINAL COMPARE=FREELUNCH*ESL*RACE CONTRAST=PAIRWISE
    *PADJUST=BONFERRONI.     ********************Erase this because we'll compute our own.
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
```

Perceived Intimacy of Adolescent Girls (Intimacy.sav)

- **Overview: Dataset contains self-ratings of the intimacy that adolescent girls perceive themselves as having with: (a) their mother and (b) their boyfriend.**
- **Source: HGSE thesis by Dr. Linda Kilner entitled Intimacy in Female Adolescent's Relationships with Parents and Friends (1991). Kilner collected the ratings using the Adolescent Intimacy Scale.**
- **Sample: 64 adolescent girls in the sophomore, junior and senior classes of a local suburban public school system.**
- **Variables:**

Self Disclosure to Mother (M_Seldis)
Trusts Mother (M_Trust)
Mutual Caring with Mother (M_Care)
Risk Vulnerability with Mother (M_Vuln)
Physical Affection with Mother (M_Phys)
Resolves Conflicts with Mother (M_Cres)

Self Disclosure to Boyfriend (B_Seldis)
Trusts Boyfriend (B_Trust)
Mutual Caring with Boyfriend (B_Care)
Risk Vulnerability with Boyfriend (B_Vuln)
Physical Affection with Boyfriend (B_Phys)
Resolves Conflicts with Boyfriend (B_Cres)

Sketch a Question-Centered Strategy

Outcome:

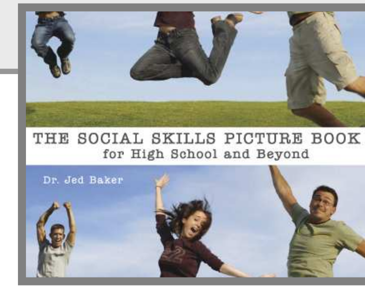
Model 1	Model 2	Model 3	Model 4

Sketch a Baseline Control Strategy

Outcome:

Model 1	Model 2	Model 3	Model 4

High School and Beyond (HSB.sav)



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

Variables about the student—

(Black) 1=Black, 0=Other
(Latin) 1=Latino/a, 0=Other
(Sex) 1=Female, 0=Male
(BYSES) Base year SES
(GPA80) HS GPA in 1980
(GPS82) HS GPA in 1982
(BYTest) Base year composite of reading and math tests
(BBConc) Base year self concept
(FEConc) First Follow-up self concept

Variables about the student's school—

(PctMin) % HS that is minority students Percentage
(HSSize) HS Size
(PctDrop) % dropouts in HS Percentage
(BYSES_S) Average SES in HS sample
(GPA80_S) Average GPA80 in HS sample
(GPA82_S) Average GPA82 in HS sample
(BYTest_S) Average test score in HS sample
(BBConc_S) Average base year self concept in HS sample
(FEConc_S) Average follow-up self concept in HS sample

Sketch a Question-Centered Strategy

Outcome:

Model 1	Model 2	Model 3	Model 4

Sketch a Baseline Control Strategy

Outcome:

Model 1	Model 2	Model 3	Model 4

Understanding Causes of Illness (ILLCAUSE.sav)

- **Overview: Data for investigating differences in children's understanding of the causes of illness, by their health status.**
- **Source: Perrin E.C., Sayer A.G., and Willett J.B. (1991).**

  Sticks And Stones May Break My Bones: Reasoning About Illness Causality And Body Functioning In Children Who Have A Chronic Illness, *Pediatrics*, 88(3), 608-19.
- **Sample: 301 children, including a sub-sample of 205 who were described as asthmatic, diabetic,or healthy. After further reductions due to the *list-wise deletion* of cases with missing data on one or more variables, the analytic sub-sample used in class ends up containing: 33 diabetic children, 68 asthmatic children and 93 healthy children.**
- **Variables:**

(ILLCAUSE)	Child's Understanding of Illness Causality
(SES)	Child's SES (Note that a high score means low SES.)
(PPVT)	Child's Score on the Peabody Picture Vocabulary Test
(AGE)	Child's Age, In Months
(GENREAS)	Child's Score on a General Reasoning Test
(ChronicallyIll)	1 = Asthmatic or Diabetic, 0 = Healthy
(Asthmatic)	1 = Asthmatic, 0 = Healthy
(Diabetic)	1 = Diabetic, 0 = Healthy

Sketch a Question-Centered Strategy

Outcome:

Model 1	Model 2	Model 3	Model 4

Sketch a Baseline Control Strategy

Outcome:

Model 1	Model 2	Model 3	Model 4

Children of Immigrants (ChildrenOfImmigrants.sav)

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

(Reading)	Stanford Reading Achievement Score
(Freelunch)	% students in school who are eligible for free lunch program
(Male)	1=Male 0=Female
(Depress)	Depression scale (Higher score means more depressed)
(SES)	Composite family SES score

Children of Immigrants (ChildrenOfImmigrants.sav)

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

(Reading)	Stanford Reading Achievement Score
(Freelunch)	% students in school who are eligible for free lunch program
(Male)	1=Male 0=Female
(Depress)	Depression scale (Higher score means more depressed)
(SES)	Composite family SES score

Sketch a Question-Centered Strategy

Outcome:Depress

Model 1	Model 2	Model 3	Model 4
Reading	**Reading**	**Reading**	**Reading**
	Male	**Male**	**Male**
		MxR	**MxR**
			SES
			FreeL
			Check other interactions!

Sketch a Baseline Control Strategy

Outcome:

Model 1	Model 2	Model 3	Model 4

Human Development in Chicago Neighborhoods (Neighborhoods.sav)

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

(Homr90)	Homicide Rate c. 1990
(Murder95)	Homicide Rate 1995
(Disadvan)	Concentrated Disadvantage
(Imm_Conc)	Immigrant
(ResStab)	Residential Stability
(Popul)	Population in 1000s
(CollEff)	Collective Efficacy
(Victim)	% Respondents Who Were Victims of Violence
(PercViol)	% Respondents Who Perceived Violence

Sketch a Question-Centered Strategy

Outcome:

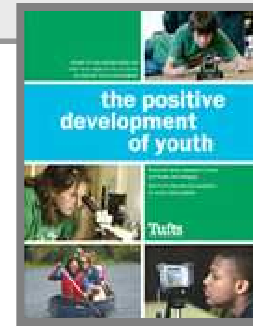
Model 1	Model 2	Model 3	Model 4

Sketch a Baseline Control Strategy

Outcome:

Model 1	Model 2	Model 3	Model 4

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



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

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

(SexFem)	1=Female, 0=Male
(MothEd)	Years of Mother's Education
(Grades)	Self-Reported Grades
(Depression)	Depression (Continuous)
(FrInfl)	Friends' Positive Influences
(PeerSupp)	Peer Support
(Depressed)	0 = (1-15 on Depression)
	1 = Yes (16+ on Depression)

(AcadComp)	Self-Perceived Academic Competence
(SocComp)	Self-Perceived Social Competence
(PhysComp)	Self-Perceived Physical Competence
(PhysApp)	Self-Perceived Physical Appearance
(CondBeh)	Self-Perceived Conduct Behavior
(SelfWorth)	Self-Worth

Sketch a Question-Centered Strategy

Outcome:

Model 1	Model 2	Model 3	Model 4

Sketch a Baseline Control Strategy

Outcome:

Model 1	Model 2	Model 3	Model 4