

**Sociology 63993**  
**Exam 2 Answer Key [Draft]**  
**March 26, 2010**

I. True-False. (20 points) Indicate whether the following statements are true or false. If false, briefly explain why.

1. A researcher has included extraneous variables in her model. Increasing her sample size will help with the problems that are created by doing this.

True. Extraneous variables lead to larger standard errors. Increasing sample size will help to reduce the standard errors.

2. A researcher obtains the following:

```
. estat ovtest
```

```
Ramsey RESET test using powers of the fitted values of warm
Ho: model has no omitted variables
      F(3, 2288) =      3.81
      Prob > F =      0.0098
```

This suggests that she should add interaction terms to her model.

False. She should add powers of X, e.g. X<sup>2</sup>, X<sup>3</sup>, X<sup>4</sup>.

3. A researcher regresses income on the respondent's race, years of education, IQ, and father's education (i.e. the number of years of education the respondent's father had). The estimated effect of father's education is 0 and is statistically insignificant. This means that, in terms of their own income, respondents gain no benefit from having a better educated father.

False (or at least not necessarily true). Father's education may have an indirect effect on child's income, e.g. having a better educated father may lead to a better educated child which in turn leads to a higher income. This is a slight variation of the status attainment example we went over in the logic of causal order.

4. In order to make interaction effects more interpretable, the dependent variable should be centered about its mean.

False. You should center the continuous independent variables. Centering the dependent variable will only cause the intercept to shift.

5. A researcher hypothesizes that informedness (measured on a 100 point scale) positively affects feelings of political efficacy for whites but has a negative effect for blacks. She obtains the following:

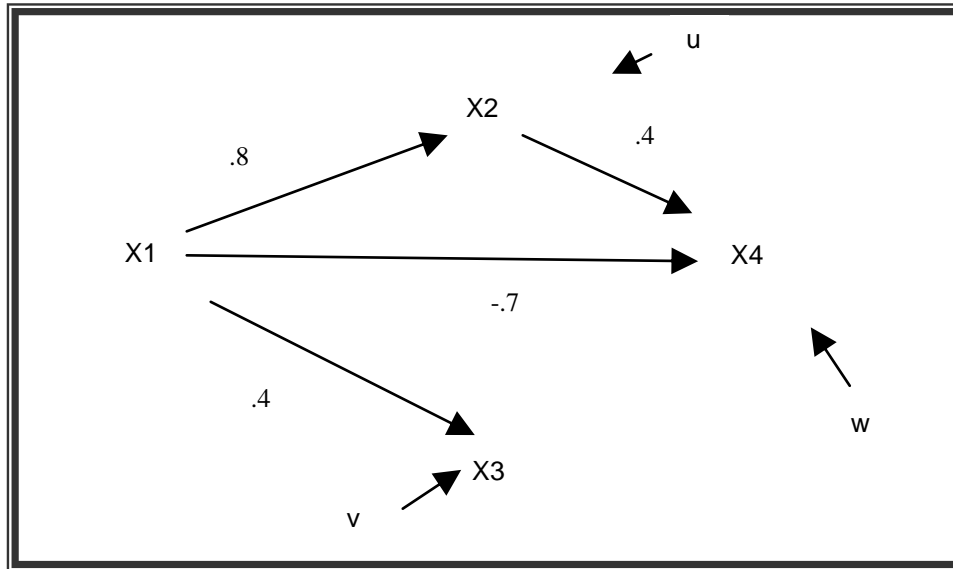
$$\begin{aligned}\beta_{\text{White}} &= 0 \\ \beta_{\text{Inf}} &= 7 \\ \beta_{\text{Inf*White}} &= -5\end{aligned}$$

$\beta_{\text{Inf}}$  and  $\beta_{\text{Inf*White}}$  are both highly significant. The results support the researcher's hypothesis.

False. For both whites and blacks, the effect of informedness is positive (7 for blacks and  $7 - 5 = 2$  for whites). To add insult to injury, the effect is actually more positive for blacks than it is for whites.

II. Path Analysis/Model specification (25 pts).

A sociologist believes that the following model describes the relationship between X1, X2, X3, and X4. All her variables are in standardized form. The estimated value of each path in her model is included in the diagram.



a. (5 pts) Write out the structural equation for each endogenous variable, using both the names for the paths (e.g.  $\beta_{42}$ ) and the estimated value of the path coefficient.

$$X_2 = \beta_{21}X_1 + u = .8X_1 + u$$

$$X_3 = \beta_{31}X_1 + v = .4X_1 + v$$

$$X_4 = \beta_{41}X_1 + \beta_{42}X_2 + w = -.7X_1 + .4X_2 + w$$

b. (10 pts) Part of the correlation matrix is shown below. Determine the complete correlation matrix. (Remember, variables are standardized. You can use either normal equations or Sewell Wright, but you might want to use both as a double-check.)

	x1	x2	x3	x4
x1	1.0000			
x2	0.8000	1.0000		
x3	?	?	1.0000	
x4	?	?	?	1.0000

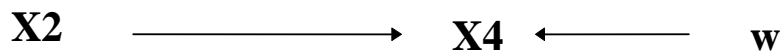
Here is the complete correlation matrix:

	x1	x2	x3	x4
x1	1.0000			
x2	0.8000	1.0000		
x3	0.4000	0.3200	1.0000	
x4	-0.3800	-0.1600	-0.1520	1.0000

c. (5 pts) Decompose the correlation between X1 and X4 into

- Correlation due to direct effects  
-.7
- Correlation due to indirect effects  
.32
- Correlation due to common causes  
0

d. (5 pts) Suppose the above model is correct, but instead the researcher believed in and estimated the following model:



What conclusions would the researcher likely draw? In particular, what would the researcher conclude about the effect of changes in X2 on X4? Discuss the consequences of this mis-specification, and in what ways, if any, the results would be misleading. Why would she make these mistakes?

The estimated effect would be the same as the correlation between the two variables, i.e. -.16, which is both smaller and opposite in sign to the true effect of .4. Correlation due to the common cause of X1 would be falsely attributed to the direct effect of X2 on X4. If X4 is a policy-related variable, the researcher might wind up doing the exact opposite of what would be effective.

III. Group comparisons (25 points). The Republican Party is uncertain about its prospects in the November Congressional elections. On the one hand, it is very excited by polls that show that, in a race between a Republican and a Democrat, 44 percent favor the Republican compared to only 39 percent for the Democrat. On the other hand, those same polls show that, if there is a third party candidate on the ballot (specifically, a Tea Party candidate), 36 percent favor the Democrat, 25 percent prefer the Republican and 15 percent say they would vote for the third party candidate. It therefore feels it needs to get a better understanding of support for third party candidacies. It has collected data from 5000 people on the following:

Variable	Description
thirdparty	Support for third party candidates, measured on a scale that ranges from -1500 to 1500. (Higher values indicate more support for a third party candidate.)
socialconservative	Scale that measures conservatism on various social issues, e.g. abortion, gay marriage. Ranges from -100 (very liberal) to 100 (very conservative). The variable has been centered to have a mean of zero.
teaparty	Coded 1 if the respondent says s/he is a supporter of the Tea Party, 0 otherwise
teasocial	teaparty * socialconservative

The results of the analysis are as follows:

```
. * Descriptive statistics
. sum thirdparty socialconservative teaparty teasocial
```

Variable	Obs	Mean	Std. Dev.	Min	Max
thirdparty	5000	-3.07e-06	69.06038	-319.041	1152.123
socialconservative	5000	-2.08e-07	13.826	-33.08312	93.46588
teaparty	5000	.1056	.3073557	0	1
teasocial	5000	.3378557	5.075021	-28.39412	93.46588

```
. * See if there are differences in 3rd party support by tea party affiliation
. ttest thirdparty, by(teaparty)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	4472	-7.555341	.4181107	27.96032	-8.375045	-6.735638
1	528	63.99142	8.027088	184.4484	48.2224	79.76044
combined	5000	-3.07e-06	.9766612	69.06038	-1.914687	1.914681
diff		-71.54676	3.01283		-77.45323	-65.6403

diff = mean(0) - mean(1) t = -23.7474  
 Ho: diff = 0 degrees of freedom = 4998

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0  
 Pr(T < t) = 0.0000 Pr(|T| > |t|) = 0.0000 Pr(T > t) = 1.0000

```
. * Estimate Models
. nestreg: reg thirdparty socialconservative teaparty teasocial
```

Block 1: socialconservative

Source	SS	df	MS	Number of obs = 5000		
Model	11190838.7	1	11190838.7	F( 1, 4998) = 4421.11		
Residual	12651070.1	4998	2531.22651	Prob > F = 0.0000		
Total	23841908.8	4999	4769.33562	R-squared = 0.4694		
				Adj R-squared = 0.4693		
				Root MSE = 50.311		

thirdparty	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
socialconservative	3.422105	.0514668	66.49	0.000	3.321207	3.523003
_cons	-2.35e-06	.7115092	-0.00	1.000	-1.394872	1.394868

Block 2: teaparty

Source	SS	df	MS	Number of obs =	5000
Model	12862350.9	2	6431175.47	F( 2, 4997)	= 2926.95
Residual	10979557.8	4997	2197.2299	Prob > F	= 0.0000
Total	23841908.8	4999	4769.33562	R-squared	= 0.5395
				Adj R-squared	= 0.5393
				Root MSE	= 46.875

thirdparty	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
socialcons~e	3.3166	.0481036	68.95	0.000	3.222296	3.410904
teaparty	59.68283	2.163876	27.58	0.000	55.44069	63.92498
_cons	-6.30251	.7011852	-8.99	0.000	-7.67714	-4.927879

Block 3: teasocial

Source	SS	df	MS	Number of obs =	5000
Model	23663705.5	3	7887901.82	F( 3, 4996)	= .
Residual	178203.299	4996	35.6691952	Prob > F	= 0.0000
Total	23841908.8	4999	4769.33562	R-squared	= 0.9925
				Adj R-squared	= 0.9925
				Root MSE	= 5.9724

thirdparty	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
socialcons~e	2.009966	.0065728	305.80	0.000	1.99708	2.022852
teaparty	32.32348	.2801499	115.38	0.000	31.77427	32.8727
teasocial	10.01233	.0181946	550.29	0.000	9.976659	10.048
_cons	-6.796085	.0893436	-76.07	0.000	-6.971238	-6.620932

Block	F	df	Residual df	Pr > F	R2	Change in R2
1	4421.11	1	4998	0.0000	0.4694	
2	760.74	1	4997	0.0000	0.5395	0.0701
3	3.0e+05	1	4996	0.0000	0.9925	0.4530

. \* See if there are differences in social conservatism by tea party affiliation  
 . ttest socialconservative, by(teaparty)

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	4472	-.3777459	.2032083	13.58915	-.7761347	.0206429
1	528	3.199391	.6673407	15.33432	1.888417	4.510366
combined	5000	-2.08e-07	.1955291	13.826	-.383323	.3833226
diff		-3.577137	.6342778		-4.8206	-2.333675

diff = mean(0) - mean(1) t = -5.6397  
 Ho: diff = 0 degrees of freedom = 4998

Ha: diff < 0 Pr(T < t) = 0.0000  
 Ha: diff != 0 Pr(|T| > |t|) = 0.0000  
 Ha: diff > 0 Pr(T > t) = 1.0000

The initial t-test shows that Tea Party members have much higher levels of support for third party candidates. Based on the remaining results, explain to the Republican Party why that is the case. When thinking about your answers, keep in mind the various reasons that two groups can differ on some outcome measure. Specifically, answer the following:

- a) (15 pts) The researchers estimate a series of models. Which of the models do you think is best, and why? What do these models tell us about how social conservatism and tea party membership affect the amount of support for third parties? What ways (if any) do the determinants of third party support differ by tea party membership? According to your preferred model, how does the thirdparty score of the “average” (on social conservatism) tea party member compare to the “average” non-member?

The 3<sup>rd</sup> model is best. It tells us that both social conservatives and tea party members are more supportive of 3<sup>rd</sup> parties. What is more, the effect of social conservatism is especially strong for tea party members. Since social conservatism is centered, the difference in 3<sup>rd</sup> party support between the “average” tea party member and non-member is 32.32 points.

- b) (10 pts) The researchers then do one last t-test. What does this test tell us about how social conservatism differs by tea party membership? What additional insights, if any, does this test give us as to why Tea party members are more supportive of third parties?

Tea party members tend to be more socially conservative than non-members. Since social conservatives are more supportive of 3<sup>rd</sup> parties, this compositional difference further contributes to the differences in 3<sup>rd</sup> party support across groups.

IV. Short answer. Answer *both* of the following questions. (15 points each, 30 points total.) Each of the following describes a nonlinear or nonadditive relationship between variables. Draw a scatterplot that illustrates the relationship. Describe the harms that might result if you simply regressed Y on X, e.g. would values be over-estimated, under-estimated, or what? Indicate the model you think should be estimated, e.g.  $E(Y) = \alpha + \beta_1 X + \beta_2 X^2$ . Explain what variables you would need to compute in order to actually estimate the model, e.g. logs of variables, interaction terms. Finally, indicate how you would actually test whether or not nonlinearity or nonadditivity actually was a problem. If you find it helpful, you are welcome to present the Stata commands you would use, but the statistical rationale behind the command still needs to be clear.

a. Our Lady of the Angels Catholic Grade School has many Hispanic immigrant students, most of whom enter first grade speaking little or no English. The Principal suspects that students learn English slowly the first few years, and then start to pick up the language much more quickly. A standardized test is used to measure the English proficiency of all students. The school finds that, for grades 1-3, each additional year of schooling leads to an average gain of 5 points on the test. For grades 4-8, each additional year of schooling leads to an average gain of 20 points on the test.

The results suggest that a spline model would be good. The effect of education gets greater once children reach 4<sup>th</sup> grade. If you simply ignored this change you would probably over-estimate the effect of schooling in the early grades and then under-estimate it later. You could test this model by testing whether or not the two slopes were the same.

b. Notre Dame academic advisors have observed that student satisfaction seems to go up and down throughout the course of a student’s four years here. They suspect that students are at first happy to be in college, then start to get tired of it, and then regain their enthusiasm once graduation is in sight. To examine this further, Notre Dame is conducting monthly studies of student satisfaction. It finds that, throughout the first three semesters of school, student satisfaction gradually rises. However, starting around second semester sophomore year, satisfaction steadily declines. However, in the last few months of Senior year, satisfaction once again starts to go up, and at a rapid pace.

This suggests a polynomial model that includes  $X$ ,  $X^2$  and  $X^3$  ( $X$  = month in school). Note that there are two “bends” (changes in direction). If you ignored these shifts, you would likely estimate a straight linear positive or negative relationship between time in school and satisfaction, and you would alternate between over-estimating levels of satisfaction and under-estimating them. You could use a Wald test, or an incremental F test; Stata’s `ovtest` command could do this for you. A spline model that allowed for three different slopes at the different times would also be a possibility.

### *APPENDIX: Stata Code for Selected Problems*

```
* Part I-2
use "http://www.indiana.edu/~jlsoc/stata/spex_data/ordwarm2.dta", clear
reg warm age
estat ovtest

* Part II - Path Analysis
clear
matrix input corr = (1, .8, 0.4, -.38 \ .8, 1, .32, -.16 \ .4, .32, 1, -.152 \ -.38, -.16, -.152, 1)
corr2data x1 x2 x3 x4, corr(corr) n(100)
corr x1 x2 x3 x4
reg x2 x1
reg x3 x2 x1
reg x4 x1 x2 x3

* Part III - Interaction effects
* Generate the variables by manipulating nhanes2f
webuse nhanes2f, clear
keep health weight black
keep if !missing(health, weight, black)
set seed 123456
sample 5000, count
center weight, gen(socialconservative)
replace socialconservative = socialconservative * .9
clonevar teaparty = black
gen teasocial = teaparty * socialconservative
gen depvar = health * -5 + 2*socialconservative + 30 * teaparty + 10 * teasocial + 85
center depvar, gen(thirdparty)
drop weight health black depvar
* Descriptive statistics
sum thirdparty socialconservative teaparty teasocial
* See if there are differences in 3rd party support by tea party affiliation
ttest thirdparty, by(teaparty)
* Estimate Models
nestreg: reg thirdparty socialconservative teaparty teasocial
* See if there are differences in social conservatism by tea party affiliation
ttest socialconservative, by(teaparty)
```