

Sociology 63993
Exam 2 Answer Key [DRAFT]
March 30, 2007

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

1. A researcher hypothesizes that education positively affects the self-esteem of men but has no effect on the self-esteem of women. She gets

$$\hat{\beta}_{Education} = 5$$

$$\hat{\beta}_{Female} = 0$$

$$\hat{\beta}_{Education * Female} = -5$$

Female = 1 if female, 0 if male. The T values for Education and for the interaction term are both highly significant. The evidence supports the researcher's hypothesis.

True. The effect of education for women is $5 - 5 = 0$.

2. A researcher believes that, for those older than 50, age has no effect on their attitudes toward working mothers. The following analysis confirms her hypothesis.

```
. mkspline age1 50 age2=age, marginal
. reg warm age1 age2
```

Source	SS	df	MS			
Model	85.5158632	2	42.7579316	Number of obs =	2293	
Residual	1889.23512	2290	.824993501	F(2, 2290) =	51.83	
				Prob > F =	0.0000	
				R-squared =	0.0433	
				Adj R-squared =	0.0425	
				Root MSE =	.90829	
Total	1974.75098	2292	.861584198			

warm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age1	-.0106279	.0022345	-4.76	0.000	-.0150098	-.006246
age2	-.0019757	.0043635	-0.45	0.651	-.0105326	.0065811
_cons	3.094973	.0844513	36.65	0.000	2.929364	3.260582

False. Note that the marginal option is being used. Because the effect of age2 is insignificant, this means that the effect of age is the same after age 50 as it is before. You can therefore just use age instead of age1 and age2.

3. A researcher has collected data from respondents in both the United States and Canada. Her measures include a 100 point scale that measures socio-economic status (SES) and a 50 point scale that measures Ambition. She regresses SES on Ambition, but does NOT include dummy variables or interaction terms for nationality. If this model is correct, it implies that the average level of SES in both countries is the same.

False. If the levels of ambition differ across nations (e.g. Americans tend to be more ambitious than Canadians) then the mean SES will differ across nations as well.

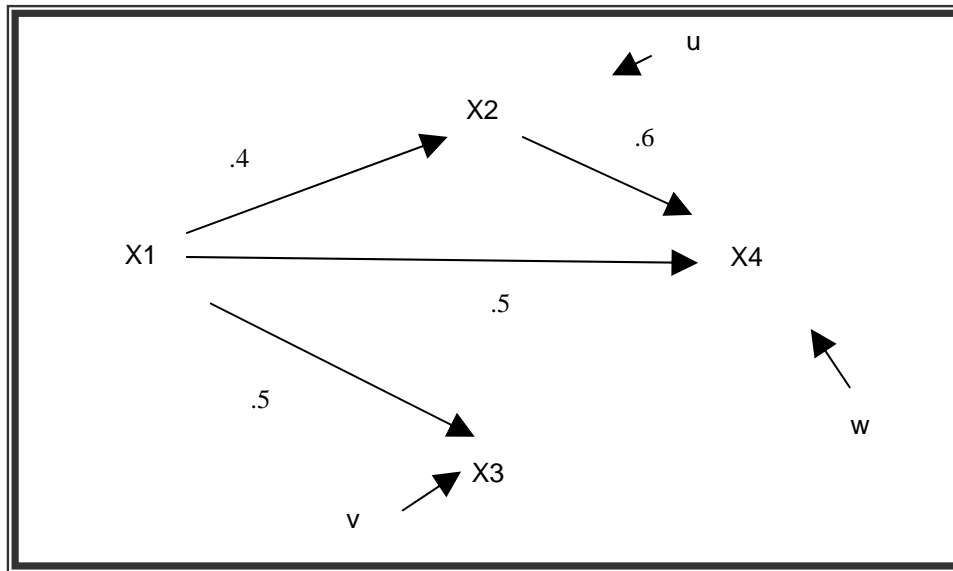
4. One reason people sometimes prefer an incremental F test over a Wald test is that the incremental F test does not require that the constrained and unconstrained models have the same number of cases.

False. You should be analyzing the same cases throughout if you are going to use an incremental F test.

5. A researcher has included dummy variables called Catholic, Protestant and Jewish in her model. This likely means that members of Other religions are not included in her analysis.

False. It likely means that “Others” are coded 0 on the other dummy variables and hence are being used as the reference category.

II. Path Analysis/Model specification (30 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. (10 pts) Write out the structural equation for each endogenous variable, using both the names for the paths (e.g. β_{42}) and the estimated value of the path coefficient.

$$X_2 = \beta_{21}X_1 + u = .4X_1 + u$$

$$X_3 = \beta_{31}X_1 + v = .5X_1 + v$$

$$X_4 = \beta_{41}X_1 + \beta_{42}X_2 + w = .5X_1 + .6X_2 + w$$

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	0.4000	1.0000		
x3	?	?	1.0000	
x4	?	?	?	1.0000

Here is the complete correlation matrix:

	x1	x2	x3	x4
x1	1.0000			
x2	0.4000	1.0000		
x3	0.5000	0.2000	1.0000	
x4	0.7400	0.8000	0.3700	1.0000

c. (5 pts) Decompose the correlation between X2 and X4 into

- Correlation due to direct effects

.6

- Correlation due to indirect effects

0

- Correlation due to common causes

.2

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 X3 on X4? Why would he make these mistakes? Discuss the consequences of this mis-specification.

In the mis-specified model, the standardized coefficient will be the same as the bivariate correlation, .37. The researcher will therefore conclude that increases in X3 lead to increases in X4. However, in the correctly specified model, we see that the effect of X3 on X4 is 0, i.e. increases in X3 have no effect on X4. The correlation between X3 and X4 is due entirely to their shared common cause, X1. Put another way, because X1 and X2 are not included in the model, the parameter estimates will suffer from omitted variable bias. Misguided efforts to increase X3 (other than by increasing X1) will be ineffective.

III. Group comparisons (30 points). A researcher is interested in the relationship between self-reported health and weight. In particular, she suspects that men feel their weight has relatively little effect on their health (indeed, being heavier might even make them feel healthier). Women, on the other hand, may be more likely to feel that being heavier hurts their health. To test her hypotheses, she uses the NHANES2F data, available from Stata's web site. The data includes information on the following:

Variable	Description
health	Self-reported health. Values range from 1 (poor health) to 5 (excellent health)
female	Coded 1 if female, 0 otherwise

weight	Weight, in kilograms
femweight	Computed variable; equals female * weight

[NOTE: These data are weighted and proper analysis should take that into account. In addition, the dependent variable is ordinal and would probably be better analyzed by ordinal regression methods that we will talk about later. For simplicity, we will ignore such details for now.]

An analysis of the data yields the following results.

```
. webuse nhanes2f, clear
. * Estimate models 1-3
. nestreg: reg health weight female femweight
```

Block 1: weight

Source	SS	df	MS			
Model	26.2659433	1	26.2659433	Number of obs =	10335	
Residual	15008.7554	10333	1.45250706	F(1, 10333) =	18.08	
Total	15035.0214	10334	1.4549082	Prob > F =	0.0000	
				R-squared =	0.0017	
				Adj R-squared =	0.0017	
				Root MSE =	1.2052	

health	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
weight	-.0032831	.0007721	-4.25	0.000	-.0047965	-.0017698
_cons	3.649905	.0567655	64.30	0.000	3.538634	3.761176

Block 2: female

Source	SS	df	MS			
Model	66.3199383	2	33.1599691	Number of obs =	10335	
Residual	14968.7014	10332	1.44877095	F(2, 10332) =	22.89	
Total	15035.0214	10334	1.4549082	Prob > F =	0.0000	
				R-squared =	0.0044	
				Adj R-squared =	0.0042	
				Root MSE =	1.2036	

health	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
weight	-.0049337	.0008325	-5.93	0.000	-.0065656	-.0033018
female	-.1345989	.0255987	-5.26	0.000	-.1847774	-.0844205
_cons	3.83925	.0671624	57.16	0.000	3.707598	3.970901

Block 3: femweight

Source	SS	df	MS			
Model	180.099242	3	60.0330806	Number of obs =	10335	
Residual	14854.9221	10331	1.4378978	F(3, 10331) =	41.75	
Total	15035.0214	10334	1.4549082	Prob > F =	0.0000	
				R-squared =	0.0120	
				Adj R-squared =	0.0117	
				Root MSE =	1.1991	

health	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
weight	.0034455	.0012551	2.75	0.006	.0009853	.0059057
female	.950207	.1245888	7.63	0.000	.7059889	1.194425
femweight	-.0148752	.0016722	-8.90	0.000	-.0181531	-.0115973
_cons	3.185756	.0993674	32.06	0.000	2.990977	3.380535

Block	F	Block df	Residual df	Pr > F	R2	Change in R2
1	18.08	1	10333	0.0000	0.0017	
2	27.65	1	10332	0.0000	0.0044	0.0027
3	79.13	1	10331	0.0000	0.0120	0.0076

```
. * Test whether the effect of weight is significant for women
. lincom weight + femweight
```

```
( 1) weight + femweight = 0
```

health	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
(1)	-.0114297	.0011051	-10.34	0.000	-.0135959 - .0092636

```
. * Now estimate Model 4
. egen meanweight = mean(weight) if e(sample)
(2 missing values generated)
. gen xweight = weight - meanweight
(2 missing values generated)
. gen femxweight = female * xweight
(2 missing values generated)
```

```
. reg health xweight female femxweight
```

Source	SS	df	MS	Number of obs =	10335
Model	180.099242	3	60.0330806	F(3, 10331) =	41.75
Residual	14854.9221	10331	1.4378978	Prob > F =	0.0000
Total	15035.0214	10334	1.4549082	R-squared =	0.0120
				Adj R-squared =	0.0117
				Root MSE =	1.1991

health	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
xweight	.0034455	.0012551	2.75	0.006	.0009853 .0059057
female	-.1193687	.0255599	-4.67	0.000	-.1694711 -.0692664
femxweight	-.0148752	.0016722	-8.90	0.000	-.0181531 -.0115973
_cons	3.433499	.0187423	183.20	0.000	3.39676 3.470237

Based on the above results, answer the following questions:

a) (15 pts) The researcher begins by estimating three models. Which of these three models do you think is best, and why? Summarize what your preferred model says about the effect of weight on self-reported health, and what it tells you about differences between men and women. Be sure to make clear whether the effect of weight differs by gender, and if so what do those differences tell us?

The incremental F tests show us that Model is a significant improvement over model 1, and Model 3 is a statistically significant improvement over Model 2. We should therefore go with model 3. Model 3 shows that both the intercept and slope differs by gender. Indeed, for men, increases in weight result in higher average scores on self-reported health (see the .0034455 coefficient for weight). But, for women, heavier weights result in lower levels of self-reported health (the negative interaction term is so large that it makes the overall effect for women negative; and the lincom test shows that the negative effect of weight for women is statistically significant.)

b) (10 pts) The researcher then estimates a fourth model. What is the rationale for doing this? While most results are the same between models 3 and 4, the constant and the coefficient for female change. Explain how the interpretation of these two coefficient differs between the two models. Does Model 4 change your decision about what model is best?

In model 4, the researcher centers weight to make the coefficients more interpretable. In model 3, the constant represents the predicted score for a man who weighs 0 kilograms. No such man exists (either in the sample or anywhere else!) In model 4, the constant represents the predicted score for a man of average weight. This average man has a predicted score of 3.43 on the health scale.

In model 3, the coefficient for female, .950207, represents the predicted difference in health between a man and women who both weigh zero kilograms. Again, no such people existed, so this isn't a particularly useful comparison. In model 4, the coefficients for female, -.1193687, is the predicted difference in health between a man and woman of average weight. Such a woman is predicted to consider herself less healthy than her male counterpart. Even apart from the researcher's theory, this may not be too surprising, since a woman of "average" overall weight would be above the average for women while the average man would be below the average for men (i.e. the average weight is computed for men and women together, not for each gender separately.)

Model 4 does not change our decision about which model is best (Models 3 and 4 are basically the same), although we might prefer the way things are parameterized in Model 4.

c) (5 pts) The researcher is now going to analyze men separately. Do you think it is wise for her to continue simply regressing health on weight, i.e. do you think her models and results will be plausible if she does this? If not, what might she try instead?

I don't find it surprising that the effect of weight differs by gender, but I do find it questionable to think that the effect of weight would be strictly linear for either gender, e.g. I doubt that a 400 pound man really thinks he is healthier than a 200 pound man. My guess is that the relationship is nonlinear in some way, e.g. the relationship is curvilinear; or, at the very least, the effect of weight on self-reported health declines as weight gets greater and greater. She might try adding weight^2 to the model; or she might try a piecewise model that allows different weights to have different effects.

IV. **Short answer.** Answer *one* of the following two questions. (20 points; up to 10 points extra credit if you do both).

1. Both of the following suggest or describe a nonlinear or nonadditive relationship between variables. Draw a scatterplot that illustrates each 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.

a. A researcher believes that Socio-Economic Status (SES) affects political liberalism, but she is unclear as to what the relationship is. A study of lower-income working class families found that the higher the SES, the more liberal

individuals tended to be. But, a study of upper-income individuals found just the opposite: the higher the income, the less liberal people tended to be. To give her some additional insights into how to model the relationship, she runs the following analysis:

```
. quietly reg liberalism ses
. estat ovtest
```

```
Ramsey RESET test using powers of the fitted values of warm
Ho: model has no omitted variables
      F(3, 2288) =      13.81
      Prob > F =      0.0000
```

Based on the test she ran, the researcher apparently suspects that the relationship between liberalism and ses is curvilinear. The ovtest is highly significant, so she should probably add ses^2 to her model. She may also want to add and test ses^3 and ses^4 to her model, but based on past results and what seems plausible ses^2 is probably sufficient. Stata statements like `gen ses2 = ses^2` will create these variables.

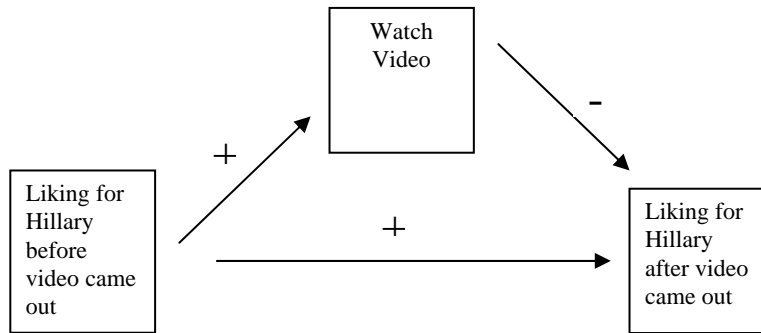
b. A government investigator is researching complaints that a company discriminates against its women employees. She finds that, for both men and women, each year the person is employed by the company his or her salary goes up an average of \$2,500. However, on average, women make \$5,000 less than men who have been at the company the same length of time.

These results suggest that the correct model is something like $E(Y) = \alpha + \beta_1 \text{YearsEmployed} + \beta_2 \text{Female}$. That is, there is no interaction effect involving gender (since for both genders each year of employment raises salary by an average of \$2,500) but the intercepts do differ by gender (women make \$5,000 less than do otherwise comparable men). If you simply ignored these gender differences, you would underestimate the income of men while overestimating it for women.)

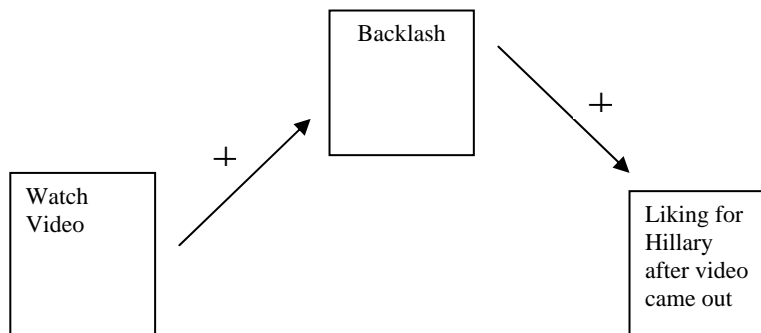
2. A supporter of Presidential candidate Barack Obama has created a video that has become a sensation on YouTube, with millions of people viewing it. The video uses clips from a famous 1984 Apple Macintosh ad and video from her own website to attack Hillary Clinton. (Go to <http://www.slate.com/id/2162286/> if you'd like to see it.) Although it did not authorize the video, the Obama campaign has commissioned a nationwide survey to determine what effect it is having on popular opinion. It wants to know whether it should encourage or discourage the creation of such videos in the future. Much to its surprise, it finds that people who have viewed the video are more supportive of Hillary Clinton than those who have not.

Drawing on your knowledge of the logic of causal order, present different models that could account for the observed relationships. Indicate what implications the different models have for encouraging or discouraging such efforts in the future. To be fair, you will want to present one or more models that suggest that such videos actually lessen support for Clinton, one or more models which imply that such videos do more harm than good for Obama (i.e. they increase support for Clinton), and one or two models which suggest that such videos do not achieve what Obama wants but the problems are correctable (i.e. you don't have to completely do away with such videos to solve the problem). When presenting your answer, keep in mind that, while the Obama campaign may have some very smart people working for it, they aren't that familiar with the logic of causal order, so you will have to make things very clear for them. Don't just draw diagrams; explain substantively what the models mean and why they might be plausible.

It could be that people who like Hillary are more likely to watch the video. After seeing the video, they like her less than they did before, although they still like her more than did the people who did not watch the video. Put another way, Hillary would be even more popular if not for this video.



It could also be that the video is counter-productive. Watching the video may create a backlash that makes people like Hillary more than before, i.e. an attack that is perceived as unfair may have exactly the opposite effect of what is occurring.



It could also be that the direct effect of watching the video is negative (it makes people like Hillary less) but there is also a positive indirect effect (the video produces feelings of backlash which in turn make people like Hillary more). If so, you might try to figure out how to avoid the backlash. Perhaps videos which had a few more facts behind them would avoid backlash. Or, maybe some types of images are more prone to producing feelings of backlash (e.g. the 1984 imagery of the ad might seem a bit much) so you could avoid such imagery in the future.

