Sociology 593
Exam 2 Answer Key [DRAFT]
March 26, 2004

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

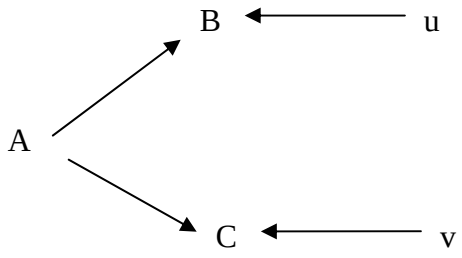
1. X1 and X2 are negatively correlated. This means that, if something causes X1 to increase, X2 will be expected to decrease.

False. This will only be true if X1 is a direct or indirect cause of X2. It need not be true if, say, they are correlated because they share a common cause. Further, if suppressor effects are present, increases in X1 could produce increases in X2.

2. A researcher believes that the relationship between Y and X is nonlinear. She is not sure whether she should use the log of Y, or whether she should leave Y as is and add an X^2 term to her model. She can use an incremental F test to help her make this decision.

False. Unfortunately, these models are not nested. You'll have to rely on theory or inspection of the data to decide on what is best.

3. A researcher believes that A is a common cause of B and C, and that neither B nor C is a direct or indirect cause of each other. Hence, knowledge of B will be of no use to her in predicting the value of C.



False. B and C will be correlated, so knowledge of B can improve your prediction of C, even though the relationship between them is non-causal.

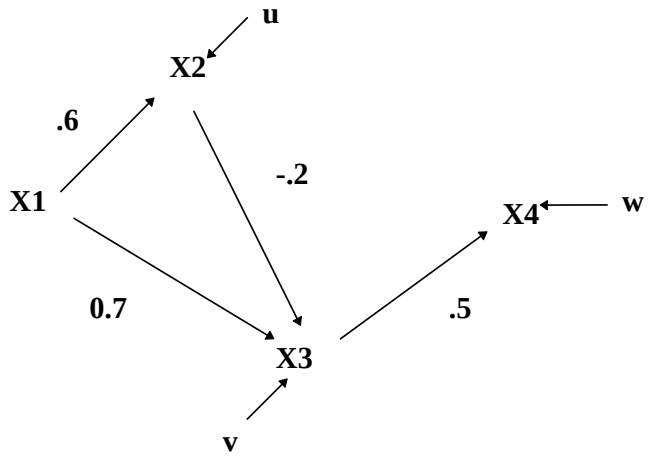
4. X4 and X1 are positively correlated. However, when X4 is regressed on X1, X2, and X3, the effect of X1 is zero. This means that suppressor effects must be present in the model.

False. The effect of X1 on X4 could be indirect; or they might be correlated because they share a common cause.

5. X1 and X2 are uncorrelated. Both affect Y. If X1 is omitted from the model, the estimated effect of X2 on Y will be unbiased.

True. Omitted variable bias is not problematic when the Xs are uncorrelated.

II. II. Path Analysis/Model specification (30 pts). A sociologist believes that the following model describes the relationships between X1, X2, X3 and X4. All variables are in standardized form. The hypothesized value of each path is included in the diagram.



a. Write out the structural equation for each endogenous variable.

$$X_2 = \beta_{21}X_1 + u = .6X_1 + u$$
$$X_3 = \beta_{31}X_1 + \beta_{32}X_2 + v = .7X_1 - .2X_2 + v$$
$$X_4 = \beta_{43}X_3 + w = .5X_3 + w$$

b. 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	?	1.0000		
x3	?	?	1.0000	
x4	0.2900	0.1100	?	1.0000

The complete correlation matrix is

	x1	x2	x3	x4
x1	1.0000			
x2	0.6000	1.0000		
x3	0.5800	0.2200	1.0000	
x4	0.2900	0.1100	0.5000	1.0000

c. Decompose the correlation between X2 and X3 into

- Correlation due to direct effects

-.2

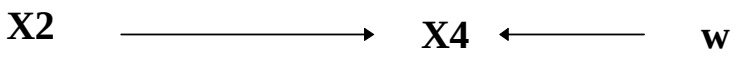
- Correlation due to indirect effects

0

- Correlation due to common causes

.42

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

The researcher would erroneously conclude that X2 has a positive direct effect on X4 when in reality it has a negative indirect effect.

III. Group comparisons (30 points).

Notre Dame is considering offering more distance learning courses to graduate students. These courses are taken over the internet and make it possible for students all across the world to take classes at Notre Dame. Before Notre Dame does so, however, it wants to gain a better understanding of what affects student satisfaction with distance learning courses. In particular, it fears that, the more time students spend working at paid jobs outside of school, the less satisfied they will be with their distance learning courses. The University also suspects that gender may be related to course satisfaction, although it is not sure how.

Notre Dame has therefore collected data from a random sample of 1,000 graduate students taking distance learning courses at other universities. The variables are "female" (coded 1 if female, 0 otherwise), "work" (hours worked per week in paid employment) and "satisfaction" (a scale that theoretically ranges from a low of 0 to a high of 100, where higher scores indicate greater satisfaction.) It conducts the following analyses with these data.

```
. * Model 1
. reg  satisfaction work

      Source |       SS       df       MS              Number of obs =    1000
-------------+------------------------------           F(  1,   998) =  246.84
       Model |  23969.7024     1  23969.7024           Prob > F      =  0.0000
    Residual |  96911.0852   998  97.1052958           R-squared     =  0.1983
-------------+------------------------------           Adj R-squared =  0.1975
       Total |  120880.788   999  121.001789           Root MSE      =  9.8542

------------------------------------------------------------------------------
satisfaction |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        work |  -.8378042   .0533252   -15.71   0.000    -.9424466   -.7331617
       _cons |   79.28226   1.120564    70.75   0.000     77.08332    81.48119
------------------------------------------------------------------------------

. * Model 2
```

```
. reg  satisfaction work female

      Source |       SS       df       MS              Number of obs =    1000
-------------+------------------------------           F(  2,   997) =  123.77
       Model |  24043.7468     2  12021.8734           Prob > F      =  0.0000
    Residual |  96837.0408   997   97.128426           R-squared     =  0.1989
-------------+------------------------------           Adj R-squared =  0.1973
       Total |  120880.788   999  121.001789           Root MSE      =  9.8554

------------------------------------------------------------------------------
satisfaction |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        work |  -.8668814   .0628755   -13.79   0.000     -.990265   -.7434979
      female |  -.7000576   .8017903    -0.87   0.383    -2.273448    .8733326
       _cons |   80.35922   1.666553    48.22   0.000     77.08886    83.62957
------------------------------------------------------------------------------

. test female

 ( 1)  female = 0

       F(  1,   997) =    0.76
            Prob > F =    0.3828
. * Model 3
. gen femwork = female * work

. reg  satisfaction work female femwork

      Source |       SS       df       MS              Number of obs =    1000
-------------+------------------------------           F(  3,   996) =   82.73
       Model |  24113.4109     3  8037.80364           Prob > F      =  0.0000
    Residual |  96767.3767   996  97.1560007           R-squared     =  0.1995
-------------+------------------------------           Adj R-squared =  0.1971
       Total |  120880.788   999  121.001789           Root MSE      =  9.8568

------------------------------------------------------------------------------
satisfaction |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        work |  -.9520935   .1186636    -8.02   0.000    -1.184953   -.7192342
      female |  -3.427172   3.318911    -1.03   0.302    -9.940032    3.085688
     femwork |   .1184876   .1399276     0.85   0.397    -.1560991    .3930743
       _cons |    82.4821   3.010527    27.40   0.000     76.57439     88.3898
------------------------------------------------------------------------------

. test female femwork

 ( 1)  female = 0
 ( 2)  femwork = 0

       F(  2,   996) =    0.74
            Prob > F =    0.4776
```

```
. * Model 3b -- try to make these results easier to interpret
. egen meanwork = mean(work)

. gen xwork = work - meanwork

. gen femxwork = female * xwork

. reg  satisfaction xwork female femxwork

      Source |       SS       df       MS              Number of obs =    1000
-------------+------------------------------           F(  3,   996) =   82.73
       Model |  24113.4109     3  8037.80365           Prob > F      =  0.0000
    Residual |  96767.3767   996  97.1560007           R-squared     =  0.1995
-------------+------------------------------           Adj R-squared =  0.1971
       Total |  120880.788   999  121.001789           Root MSE      =  9.8568

------------------------------------------------------------------------------
satisfaction |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       xwork |  -.9520935   .1186636    -8.02   0.000    -1.184953   -.7192342
      female |  -1.035514   .8944215    -1.16   0.247    -2.790681    .7196529
    femxwork |   .1184876   .1399276     0.85   0.397    -.1560991    .3930743
       _cons |    63.2642   .7991364    79.17   0.000     61.69602    64.83239
------------------------------------------------------------------------------

. test  female femxwork

 ( 1)  female = 0
 ( 2)  femxwork = 0

       F(  2,   996) =    0.74
            Prob > F =    0.4776
* Check for gender differences.

. ttest satisfaction, by(female)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    Male |     300    58.76272    .6168699     10.6845    57.54876    59.97667
  Female |     700    63.91781    .4072876    10.77582    63.11815    64.71746
---------+--------------------------------------------------------------------
combined |    1000    62.37128    .3478531    11.00008    61.68867    63.05389
---------+--------------------------------------------------------------------
    diff |           -5.155089    .7417197               -6.610598    -3.69958
------------------------------------------------------------------------------
Degrees of freedom: 998

                 Ho: mean(Male) - mean(Female) = diff = 0

     Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
       t =  -6.9502                 t =  -6.9502                 t =  -6.9502
   P < t =   0.0000           P > |t| =   0.0000           P > t =   1.0000
```

```
. ttest  work, by(female)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    Male |     300    24.91286    .2773454    4.803764    24.36707    25.45866
  Female |     700     18.1586    .1900283    5.027675    17.78551    18.53169
---------+--------------------------------------------------------------------
combined |    1000    20.18488    .1848869    5.846636    19.82207    20.54769
---------+--------------------------------------------------------------------
    diff |            6.754265    .3423865                6.082385    7.426145
------------------------------------------------------------------------------
Degrees of freedom: 998

                 Ho: mean(Male) - mean(Female) = diff = 0

     Ha: diff < 0               Ha: diff != 0                 Ha: diff > 0
       t =  19.7270                 t =  19.7270                  t =  19.7270
   P < t =   1.0000           P > |t| =   0.0000           P > t =   0.0000
```

Based on the above results, answer the following questions:

a) The researcher begins by estimating three models. Which of these three models do you think is best, and why?

Model 1, which includes the direct effect of work and nothing else. The subsequent models show that neither the direct nor interactive effects of gender are significant.

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

In model 1, the coefficient for female tells us the expected difference in satisfaction between a male and female who work 0 hours a week. In model 2, after centering, the coefficient for female tells us the expected difference in satisfaction between a man and woman who both work the average number of hours per week. In this base, both approaches yield substantively interesting results; but in neither case is the difference between a man and a woman significant. Model 3B does not change our conclusion about what model is best.

c) Based on these results, the researchers analyzing these data come to two major conclusions:

1. The more students work at paying jobs outside of class, the less satisfied they are with their courses.
2. Gender is unrelated to course satisfaction. Men and women are equally satisfied with their distance learning courses.

Indicate what you think the researchers are basing each conclusion on. Also indicate whether or not you agree with each conclusion, and why. If you disagree, suggest a simple path model that might explain how gender, work, and satisfaction are interrelated.

I agree with conclusion 1 but not conclusion 2. The preferred model does indeed show that hours worked negatively affects satisfaction. The conclusion about gender is apparently based on the finding that gender terms do not enter into the model. However, the t-tests show us that women score significantly higher on the satisfaction measure, and they also show us that women work significantly fewer hours than men. Hence, because women work less, they tend to be more satisfied than men are. One plausible explanation is that the effect of gender on satisfaction is indirect: women work fewer hours, and those who work fewer hours tend to be more satisfied.

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 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. An educator is interested in the extent to which students' attention span (i.e. the extent to which students focus on what their teachers are saying) varies across grades 1-12. She believes that, between grades 1-6, attention span gradually increases. However, between grades 7-10, attention span actually declines, as students become more and more preoccupied with social matters rather than their studies. In grades 11-12, however, attention span once again increases, as students start to get concerned with getting into college and with their careers after high school.

This is a cubic model. Include terms for X, X^2 and X^3.

b. A business is trying to figure out how much impact advertising has on sales. It reviews its advertising expenditures and sales revenues for the past several years. It finds that, for the first $100,000 it spends on an advertising campaign, sales increase on average by $10 for each $1 spent on advertising. However, for each additional dollar above $100,000 spent on an advertising campaign, sales only increase by $5 for each dollar spent on advertising.

Piecewise regression model. Between $0 and $100,000, the slope is 10; after $100,000, the slope drops to 5, i.e. advertising dollars become less effective once you spend more than $100,000.

2. The President of Notre Dame is concerned about low faculty morale. He believes that more information and closer contact with top administrators will help faculty to feel better about the university. He is therefore holding a series of public forums open to any faculty who wish to attend. At these forums, the President, Deans and Provost discuss the major issues facing the University and seek faculty input.

At the end of the year, the President is dismayed to learn that surveys show that faculty who attended these sessions actually have lower morale than faculty who did not attend. Some of his advisors tell him that they think these sessions are doing more harm than good, while others maintain that the forums are helping to achieve their intended purpose and should be continued.

The President has therefore called on you, a professionally trained social scientist, to explain why these relationships might exist. 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 what should be done about the program. To be fair, you will want to present one or more models that suggest that the forums help morale, one or more models which imply that the forums do more harm than good, and one or two models which suggest that the forums are not achieving what the President wants but the problems are correctable (i.e. you don't have to completely scrap the forums to solve the problem). When presenting your answer, keep in mind that the President does not know very much about the logic of causal order, so you will have to make things very clear for him.

Forums are bad. Simplest model: The forums decrease morale. Perhaps because they are poorly run, the forums make faculty feel worse instead of better.

Forums are good. Suppressor effects: Those who have lower morale are more likely to attend the forums. The forums boost their morale, but it is still lower than those who do not attend. Eliminating the forums would therefore make more problems even worse.

Fixable problems: The forums have good and bad points. Listening to faculty is good, but lecturing/informing the faculty just irritates them. Change the format of the forums so that only the positive and beneficial features are kept.