Sociology 63993
Exam 2 Answer Key - DRAFT
April 4, 2008

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

1. When a model has two independent variables, e.g. X1 and X2, it is usually a good idea to test whether their effects are equal.

False. It usually only makes sense when X1 and X2 are measured in the same metric, e.g. years, dollars. Even then it may or may not make substantive sense.

2. A Chow test is used to examine whether or not data are missing at random.

False. A Chow test is used to test whether there are differences in coefficients across groups.

3. A researcher regresses Y on X1, X2, X3 and X4. The estimated effect of X1 is zero. We can therefore be confident that, if something is done that causes the value of X1 to increase, Y will be unaffected.

False. X1 could have indirect effects, e.g. X1 affects X2 and X3 which in turn affect X4.

4. A larger sample size will help to reduce the problems caused by omitted variable bias.

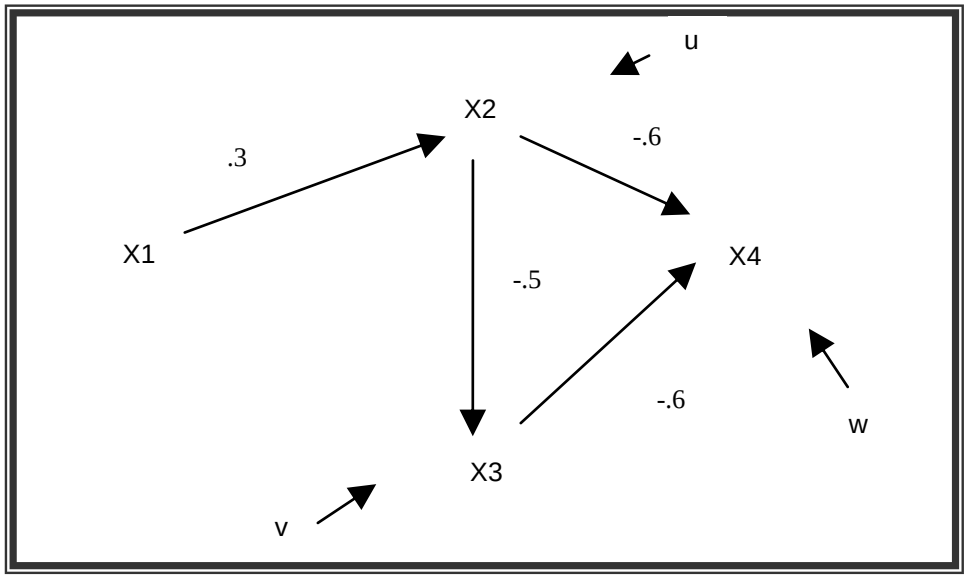
False. The estimates will continue to be biased.

5. A researcher believes that the effect of age is greater (larger in magnitude) for those older than 50 than for those who are younger. The following results contradict her hypothesis:

```
. use "http://www.indiana.edu/~jslsoc/stata/spex_data/ordwarm2.dta", clear
(77 & 89 General Social Survey)

. mkspline age1 50 age2=age

. reg warm  age1 age2

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

------------------------------------------------------------------------------
        warm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        age1 |  -.0106279   .0022345    -4.76   0.000    -.0150098    -.006246
        age2 |  -.0126036   .0026858    -4.69   0.000    -.0178705   -.0073368
       _cons |   3.094973   .0844513    36.65   0.000     2.929364    3.260582
------------------------------------------------------------------------------

. test age1 = age2

 ( 1)  age1 - age2 = 0

       F(  1,  2290) =    0.21
            Prob > F =    0.6507
```

True. The marginal option was not used, so the effects of age1 and age2 correspond to the effects of age for each age group. The test command shows that, counter to the researcher's hypothesis, the effects of age are not grater for older people.

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. β_{42}) and the estimated value of the path coefficient.

$$X_2 = \beta_{21}X_1 + u = .3X_1 + u$$
$$X_3 = \beta_{32}X_2 + v = -.5X_2 + v$$
$$X_4 = \beta_{42}X_2 + \beta_{43}X_3 + w = -.6X_2 - .6X_3 + w$$

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	0.3000	1.0000		
x3	?	?	1.0000	
x4	?	?	?	1.0000

Complete matrix:

	x1	x2	x3	x4
x1	1.0000			
x2	0.3000	1.0000		
x3	-0.1500	-0.5000	1.0000	
x4	-0.0900	-0.3000	-0.3000	1.0000

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

- Correlation due to direct effects

-.6

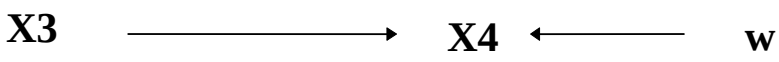
- Correlation due to indirect effects

.3

- 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 X3 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 equal to the correlation between X3 and X4, -.3. This is only half as large as the effect found in the correct model of -.6. Thus, the researcher would greatly underestimate the impact of X3 on X4. The smaller effect would also increase the likelihood that the researcher would conclude that the effect did not significantly differ from 0. This mistake would occur because of omitted variable bias; the correlation between X3 and X4 that is due to the common cause of X2 would instead be attributed to the direct effect of X3 on X4.

III. Group comparisons (25 points). It is April 23, 2008. To the dismay of her critics, Hillary Clinton continues to fight fiercely for the presidency – and her landslide victory in Pennsylvania yesterday has the Obama camp worried. With Clinton surging, everyone agrees that, if she can repeat her success in the key battleground state of Indiana, the party convention could well become hopelessly deadlocked in August. Obama's staff therefore feels it must get a better understanding of the reasons for Clinton's popularity. In particular, the staff feels that it has to know how people's gender and their concerns about health care are related to their attitudes towards Clinton. Pollsters have therefore collected information on the following variables:

Variable	Description
clinton	Liking for Clinton, measured on a scale that ranges from a low of 0 to a high of 100
female	Coded 1 if female, 0 otherwise
hlthcare	How concerned the respondent is with health care. Scores can range from a low of 0 (not concerned at all) to a high of 30 (extremely concerned)
femed	female * hlthcare

Almost 2300 likely voters are surveyed. The results of the analysis are as follows:

```
. * Estimate Models
. nestreg: reg  clinton  hlthcare female femed

Block  1: hlthcare

      Source |       SS       df       MS              Number of obs =    2293
-------------+------------------------------           F(  1,  2291) =  547.23
       Model |  319218.141     1  319218.141           Prob > F      =  0.0000
    Residual |  1336427.49  2291   583.33806           R-squared     =  0.1928
-------------+------------------------------           Adj R-squared =  0.1925
       Total |  1655645.64  2292   722.35848           Root MSE      =  24.152

------------------------------------------------------------------------------
     clinton |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    hlthcare |   2.971744   .1270363    23.39   0.000     2.722626    3.220862
       _cons |   7.295779   1.868545     3.90   0.000     3.631561       10.96
------------------------------------------------------------------------------

Block  2: female

      Source |       SS       df       MS              Number of obs =    2293
-------------+------------------------------           F(  2,  2290) = 1325.93
       Model |  888439.231     2  444219.616           Prob > F      =  0.0000
    Residual |  767206.404  2290  335.024631           R-squared     =  0.5366
-------------+------------------------------           Adj R-squared =  0.5362
       Total |  1655645.64  2292   722.35848           Root MSE      =  18.304

------------------------------------------------------------------------------
     clinton |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    hlthcare |   1.090055   .1065482    10.23   0.000     .8811136    1.298996
      female |   34.96084   .8481638    41.22   0.000     33.29759    36.62409
       _cons |    15.2379   1.429109    10.66   0.000     12.43542    18.04038
------------------------------------------------------------------------------

Block  3: femed

      Source |       SS       df       MS              Number of obs =    2293
-------------+------------------------------           F(  3,  2289) =  885.36
       Model |  889272.558     3  296424.186           Prob > F      =  0.0000
    Residual |  766373.078  2289  334.806936           R-squared     =  0.5371
-------------+------------------------------           Adj R-squared =  0.5365
       Total |  1655645.64  2292   722.35848           Root MSE      =  18.298

------------------------------------------------------------------------------
     clinton |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    hlthcare |   .9038511    .158982     5.69   0.000     .5920873    1.215615
      female |   30.27528   3.088626     9.80   0.000     24.21848    36.33208
       femed |   .3378538   .2141501     1.58   0.115    -.0820948    .7578023
       _cons |   17.53522   2.039965     8.60   0.000     13.53485     21.5356
------------------------------------------------------------------------------


  +-------------------------------------------------------------+
  |       |          Block  Residual                    Change |
  | Block |       F     df        df   Pr > F       R2    in R2 |
  |-------+-----------------------------------------------------|
  |     1 |  547.23      1      2291   0.0000   0.1928          |
  |     2 | 1699.04      1      2290   0.0000   0.5366   0.3438 |
  |     3 |    2.49      1      2289   0.1148   0.5371   0.0005 |
  +-------------------------------------------------------------+
```

```
. * Differences by gender
. ttest clinton, by(female)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    Male |    1066    28.68668    .5322045    17.37629    27.64239    29.73097
  Female |    1227    67.36528     .565294    19.80143    66.25623    68.47433
---------+--------------------------------------------------------------------
combined |    2293    49.38386    .5612733    26.87673    48.28321    50.48452
---------+--------------------------------------------------------------------
    diff |            -38.6786    .7835188               -40.21508   -37.14212
------------------------------------------------------------------------------
    diff = mean(Male) - mean(Female)                              t = -49.3653
Ho: diff = 0                                     degrees of freedom =     2291

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

. ttest hlthcare, by(female)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    Male |    1066    12.33771    .1080179    3.526749    12.12576    12.54966
  Female |    1227    15.74833    .1039811    3.642307    15.54433    15.95233
---------+--------------------------------------------------------------------
combined |    2293    14.16276    .0829322    3.971231    14.00013    14.32539
---------+--------------------------------------------------------------------
    diff |           -3.410618    .1502729               -3.705304   -3.115933
------------------------------------------------------------------------------
    diff = mean(Male) - mean(Female)                              t = -22.6962
Ho: diff = 0                                     degrees of freedom =     2291

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

Based on the above results, advise the Obama team on the following. When thinking about your answers, keep in mind the various reasons that two groups can differ on some outcome measure.

a) (15 pts) The researchers begin by estimating a series of models. Which of the models do you think is best, and why? What do these models tell us about how concern about healthcare affects support for Clinton? What ways (if any) do the determinants of support for Clinton differ by gender?

Model 2 is best. It is a significant improvement over Model 2, while model 3 is not a significant improvement over model 2, i.e. the interaction term is not significant. This model says that the intercepts differ for men and women, but the effects of hlthcare do not. People who are more concerned about health care, and also women, tend to have higher opinions of Hillary (after controlling for the other variable in the model). Put another way, when men and women have the same attitudes on hlthcare, the women tend to like Hillary more.

b) (10 pts) The researchers then run a series of t-tests. What do these t-tests tell us about how attitudes differ by gender? What additional insights, if any, do these tests give us as to why support for Clinton differs by gender?

Hillary is much more popular with women than she is with men. Women are also more concerned about health care. Because hlthcare positively affects attitudes toward Clinton, women's greater concern for health care (a compositional difference) adds to Hillary's greater popularity among women.

In short, gender is important for two reasons. First, the intercept is greater for women than it is for men. Second, women tend to be more concerned about health care, which in turn causes them to like Hillary more.

IV. Short answer. Answer *both* of the following questions. (15 points each, 30 points total.) In each of the following problems, a researcher runs through a sequence of commands. Explain why she didn't stop after the first command, i.e. explain what the purpose of each subsequent command was, what it told her, and why she did not run additional commands after the last one. If she had stopped after the first command, what would the consequences have been, i.e. in what ways would her conclusions have been incorrect or misleading?

1.

```
. reg y x

      Source |       SS       df       MS              Number of obs =    2293
-------------+------------------------------           F(  1,  2291) =  223.47
       Model |  94109363.8     1  94109363.8           Prob > F      =  0.0000
    Residual |   964794055  2291  421123.551           R-squared     =  0.0889
-------------+------------------------------           Adj R-squared =  0.0885
       Total |  1.0589e+09  2292  461999.746           Root MSE      =  648.94

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   12.07653   .8078496    14.95   0.000     10.49234    13.66072
       _cons |   422.1198   13.55198    31.15   0.000     395.5444    448.6952
------------------------------------------------------------------------------

. estat ovtest

Ramsey RESET test using powers of the fitted values of y
       Ho:  model has no omitted variables
                F(3, 2288) =    526.10
                  Prob > F =      0.0000

. gen x2 = x^2

. reg y x x2

      Source |       SS       df       MS              Number of obs =    2293
-------------+------------------------------           F(  2,  2290) =  977.72
       Model |   487727633     2   243863816           Prob > F      =  0.0000
    Residual |   571175786  2290  249421.741           R-squared     =  0.4606
-------------+------------------------------           Adj R-squared =  0.4601
       Total |  1.0589e+09  2292  461999.746           Root MSE      =  499.42

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   2.012126    .671355     3.00   0.003     .6955989    3.328654
          x2 |   1.443811   .0363446    39.73   0.000      1.37254    1.515083
       _cons |    15.8122   14.60768     1.08   0.279    -12.83346    44.45787
------------------------------------------------------------------------------

. estat ovtest

Ramsey RESET test using powers of the fitted values of y
       Ho:  model has no omitted variables
                F(3, 2287) =      0.08
                  Prob > F =      0.9714
```

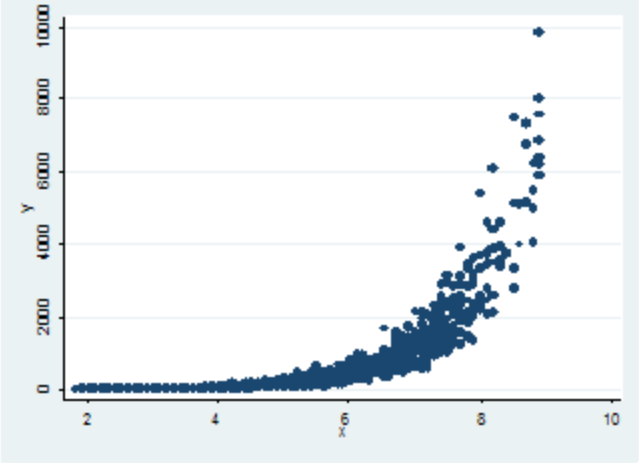
The researcher suspects that x may have a curvilinear relationship with y. The estat ovtest command confirms that adding one or more higher powers of x (x^2, x^3, x^4) would significantly improve the fit of the model. She therefore generates x^2 and adds it to the model. The effect of x^2 is highly significant, and the subsequent estat ovtest command shows that there is now no need to add any more higher powers. If she stuck with the original model, she would overestimate y in some parts of the x range and underestimate it in others. Further, she would miss the curvilinear relationship, and erroneously conclude that the effect of x is always positive when in fact it switches to being negative after the bend.

2.

```
. reg y x
```

```
      Source |       SS       df       MS              Number of obs =    2293
-------------+------------------------------           F(  1,  2291) = 2024.47
       Model |   728202953     1   728202953           Prob > F      =  0.0000
    Residual |   824073563  2291  359700.377           R-squared     =  0.4691
-------------+------------------------------           Adj R-squared =  0.4689
       Total |  1.5523e+09  2292  677258.515           Root MSE      =  599.75

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   335.9325   7.466141    44.99   0.000     321.2914    350.5736
       _cons |  -1139.057   35.81109   -31.81   0.000    -1209.282   -1068.831
------------------------------------------------------------------------------
```

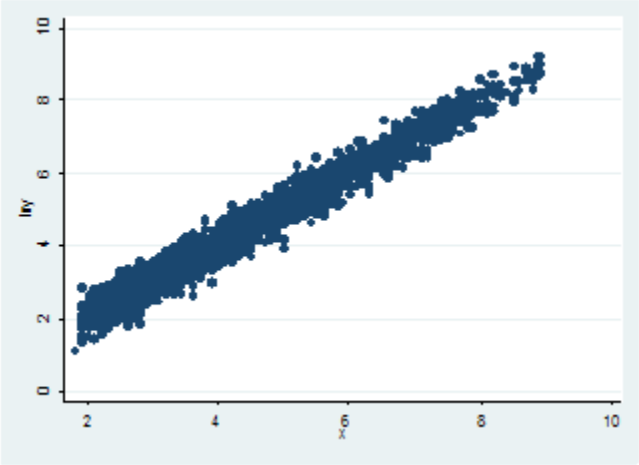
```
. scatter y x
```




```
. gen lny = ln(y)
```

```
. reg lny x

      Source |       SS       df       MS              Number of obs =    2293
-------------+------------------------------           F(  1,  2291) =70636.43
       Model |  6346.42315     1  6346.42315           Prob > F      =  0.0000
    Residual |    205.8379  2291  .089846312           R-squared     =  0.9686
-------------+------------------------------           Adj R-squared =  0.9686
       Total |  6552.26105  2292  2.85875264           Root MSE      =  .29974

------------------------------------------------------------------------------
         lny |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .9917227   .0037314   265.78   0.000     .9844054    .9990401
       _cons |   .0371944   .0178977     2.08   0.038      .002097    .0722918
------------------------------------------------------------------------------
```

```
. scatter lny x
```



The researcher suspects that the relationship between x and y may not be linear. The scatterplot suggests an exponential relationship. The researcher therefore computes the log of y, and regresses it on x. This produces a much larger R^2 value. The subsequent scatterplot suggests that the relationship between lny and x is indeed linear, so the researcher decides that no additional analysis is necessary. Failure to make this transformation would cause y to alternate between being overestimated and underestimated and would miss the exponential growth that is truly going on.

Appendix: Stata code used in this exam

Problem II:

```
clear
matrix input corr = (1,.3,-.15,-.09\.3,1,-.5,-.30\-.15,-.5,1,-.30\-.09,-.30,-.30,1)
corr2data x1 x2 x3 x4, n(100) corr(corr) double
reg x2 x1
reg x3 x1 x2
reg x4 x1 x2 x3
```

Problem III:

```
use "http://www.indiana.edu/~jslsoc/stata/spex_data/ordwarm2.dta", clear
gen female = male==0
label define female 0 "Male" 1 "Female"
label values female female
gen hlthcare = female * .3 * ed + ed
gen femed = female * hlthcare
gen clinton = ((female * .04 * warm + warm + female*1.5) - 1) * 20

* Results for exam start below

* Estimate Models
nestreg: reg  clinton  hlthcare female femed

* Differences by gender
ttest clinton, by(female)
ttest hlthcare, by(female)
```

Problem IV-a:

```
use "http://www.indiana.edu/~jslsoc/stata/spex_data/ordwarm2.dta", clear
corr2data e, sd(500)
center age
ren c_age x
gen y = 3*x + 1.5*x^2 + e
reg y x
estat ovtest
gen x2 = x^2
reg y x x2
estat ovtest
```

Problem IV-b:

```
use "http://www.indiana.edu/~jslsoc/stata/spex_data/ordwarm2.dta", clear
gen x = age/10
corr2data e, sd(.3)
gen y = exp(x + e)
reg y x
scatter y x
gen lny = ln(y)
reg lny x
scatter lny x
```