Sociology 63993
Exam 2
March 28, 2014

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

1. A researcher runs the following regression:

```
. reg income black educ

      Source |       SS       df       MS              Number of obs =     534
-------------+------------------------------           F(  2,   531) =  255.09
       Model |  66859.5212     2  33429.7606           Prob > F      =  0.0000
    Residual |  69588.4788   531  131.051749           R-squared     =  0.4900
-------------+------------------------------           Adj R-squared =  0.4881
       Total |      136448   533         256           Root MSE      =  11.448

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |   .0175821    5.19801     0.00   0.997    -10.1936    10.22877
        educ |   3.499835   .1624378    21.55   0.000     3.180736    3.818935
       _cons |  -1.23e-08    .495394    -0.00   1.000    -.9731727    .9731726
------------------------------------------------------------------------------
```

Based on these results, the researcher should conclude that a person’s race has no effect on his or her income.

2. A researcher runs the following:

```
. gen edmale = ed * male
. reg warm male ed edmale

      Source |       SS       df       MS              Number of obs =    2293
-------------+------------------------------           F(  3,  2289) =   60.35
       Model |  144.755012     3  48.2516706           Prob > F      =  0.0000
    Residual |  1829.99597  2289  .799473993           R-squared     =  0.0733
-------------+------------------------------           Adj R-squared =  0.0721
       Total |  1974.75098  2292  .861584198           Root MSE      =  .89413

------------------------------------------------------------------------------
        warm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .0589486   .1509281     0.39   0.696    -.2370216    .3549188
          ed |   .0776066   .0091143     8.51   0.000     .0597334    .0954797
      edmale |   -.032414    .011976    -2.71   0.007    -.0558989   -.0089291
       _cons |   1.817813   .1133239    16.04   0.000     1.595584    2.040041
------------------------------------------------------------------------------
```

This means that the estimated effect of education is positive for both men and women.

3. A researcher has run the following commands:

```
reg y x1 x2 x3
est store m1
reg y x1 x4
est store m2
```

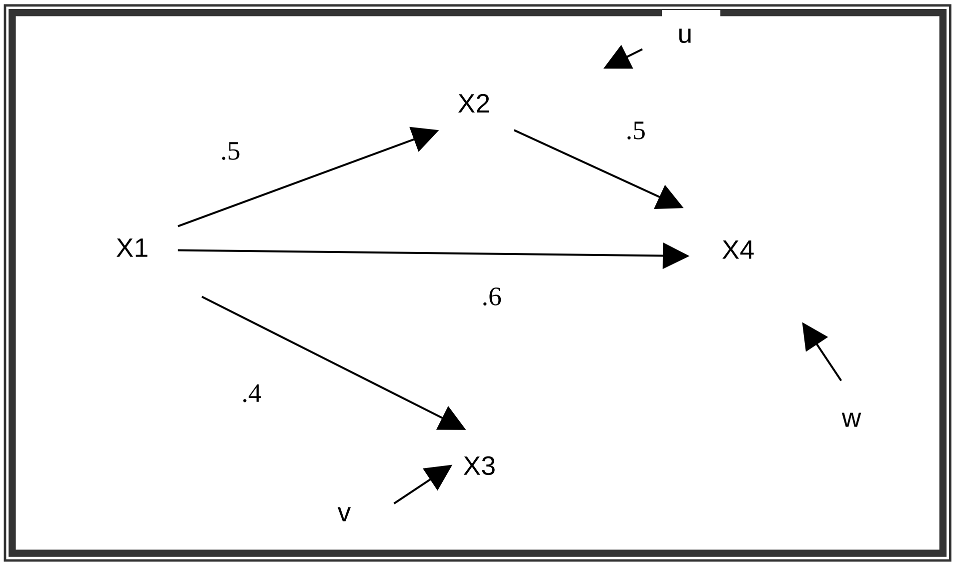
She can now use an incremental F test or a Likelihood Ratio test to determine which of her two regression models is better.

4. A model includes two independent variables: education, measured in years, and income, measured in thousands of dollars. If the researcher wishes to compare the effects of these two variables, she should test the hypotheses

   H_0: $\beta_{\text{education}} = \beta_{\text{income}}$
   H_A: $\beta_{\text{education}} \neq \beta_{\text{income}}$

5. A researcher has inadvertently omitted an important variable from her model. Fortunately, as the sample size gets bigger and bigger, the omitted variable bias will diminish and eventually disappear.

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.

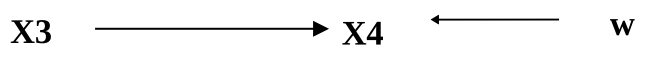
b. (10 pts) Part of the correlation matrix is shown below. Determine the complete correlation matrix. Show your work. (Remember, variables are standardized.)

```
             |       x1       x2       x3       x4
-------------+------------------------------------
          x1 |   1.0000
          x2 |   0.5000   1.0000
          x3 |      ?        ?     1.0000
          x4 |      ?        ?        ?     1.0000
```

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

- Correlation due to direct effects
- Correlation due to indirect effects
- Correlation due to common causes

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.

III. Group comparisons (25 points). The signup period for the Affordable Care Act will end in a few days. Democratic Party officials are worried that opposition to the act will hurt the party in the mid-term elections. They are therefore trying to identify factors that are related to support for the ACA. In particular, They fear that people who already have insurance through their employers will be less favorable toward the Act. A random sample of more than 4,400 American adults has therefore been asked about the following:

Variable	Description
aca	Support for the Affordable Care Act. Scores potentially range from a low of 0 to a high of 100.
ses	Socio-Economic Scale. The scale has been centered to have a mean of zero. Observed values on the centered scale range from about -50 to +100.
employer	Does the respondent already have insurance provided by an employer? 1 = yes, 0 = no
empses	Interaction term; employer * ses

The results of the analysis are as follows:

```
. ttest aca, by(employer)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |    2112    52.27996    .2252155    10.35011     51.8383    52.72163
       1 |    2320    38.47903    .2224307    10.71368    38.04284    38.91521
---------+--------------------------------------------------------------------
combined |    4432    45.05565    .1891882    12.59488    44.68474    45.42655
---------+--------------------------------------------------------------------
    diff |            13.80094    .3170529                13.17936    14.42252
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  43.5288
Ho: diff = 0                                     degrees of freedom =     4430

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

```
. nestreg: reg aca ses employer empses

Block  1: ses

      Source |       SS       df       MS              Number of obs =    4432
-------------+------------------------------           F(  1,  4430) = 1687.72
       Model |  193909.975     1  193909.975           Prob > F      =  0.0000
    Residual |  508983.622  4430  114.894723           R-squared     =  0.2759
-------------+------------------------------           Adj R-squared =  0.2757
       Total |  702893.598  4431  158.630918           Root MSE      =  10.719

------------------------------------------------------------------------------
         aca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ses |  -.3873433   .0094286   -41.08   0.000     -.405828   -.3688586
       _cons |   45.05565    .161009   279.83   0.000     44.73999    45.37131
------------------------------------------------------------------------------

Block  2: employer

      Source |       SS       df       MS              Number of obs =    4432
-------------+------------------------------           F(  2,  4429) = 1321.00
       Model |  262628.413     2  131314.206           Prob > F      =  0.0000
    Residual |  440265.185  4429  99.4050993           R-squared     =  0.3736
-------------+------------------------------           Adj R-squared =  0.3734
       Total |  702893.598  4431  158.630918           Root MSE      =  9.9702

------------------------------------------------------------------------------
         aca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ses |  -.2387547   .0104332   -22.88   0.000    -.2592089   -.2183004
    employer |   -9.37911   .3567215   -26.29   0.000    -10.07846   -8.679758
       _cons |   49.96529   .2393692   208.74   0.000     49.49601    50.43457
------------------------------------------------------------------------------

Block  3: empses

      Source |       SS       df       MS              Number of obs =    4432
-------------+------------------------------           F(  3,  4428) =  880.52
       Model |  262637.684     3  87545.8948           Prob > F      =  0.0000
    Residual |  440255.913  4428  99.4254546           R-squared     =  0.3737
-------------+------------------------------           Adj R-squared =  0.3732
       Total |  702893.598  4431  158.630918           Root MSE      =  9.9712

------------------------------------------------------------------------------
         aca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ses |  -.2352496   .0155117   -15.17   0.000    -.2656603    -.204839
    employer |  -9.387526   .3578209   -26.24   0.000    -10.08903   -8.686018
      empses |  -.0064017   .0209634    -0.31   0.760    -.0475003     .034697
       _cons |   49.99927   .2639912   189.40   0.000     49.48172    50.51682
------------------------------------------------------------------------------


  +-------------------------------------------------------------+
  |       |          Block  Residual                     Change |
  | Block |       F     df        df   Pr > F       R2    in R2 |
  |-------+-----------------------------------------------------|
  |     1 | 1687.72      1      4430   0.0000   0.2759          |
  |     2 |  691.30      1      4429   0.0000   0.3736   0.0978 |
  |     3 |    0.09      1      4428   0.7601   0.3737   0.0000 |
  +-------------------------------------------------------------+
. ttest ses, by(employer)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |    2112   -9.694785    .3044379     13.9909   -10.29181   -9.097755
       1 |    2320    8.825596    .3048539    14.68371    8.227782    9.423411
---------+--------------------------------------------------------------------
combined |    4432   -4.62e-07    .2565389    17.07863   -.5029449    .5029439
---------+--------------------------------------------------------------------
    diff |           -18.52038    .4318123               -19.36695   -17.67381
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t = -42.8899
Ho: diff = 0                                     degrees of freedom =     4430

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

The initial t-test shows that those with employer-provided health insurance have significantly lower levels of support for the Affordable Care Act. Based on the remaining results, explain to the Democratic Party officials 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) (10 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 SES and employer-provided insurance affect the amount of support for the ACA? What ways (if any) do the determinants of support for the ACA differ by those who have and do not have employer-provided insurance?

b) (5 pts) Suppose you had two people with average SES scores, one of whom had insurance through their employer while the other did not. According to your preferred model, what would be the predicted ACA score for each person?

c) (10 pts) The researchers then do one last t-test. What does this test tell us about how SES differs between those who have and do not have employer-provided insurance? What additional insights, if any, does this test give us as to why those with insurance from their employers are less supportive of the ACA?

IV. Short answer.

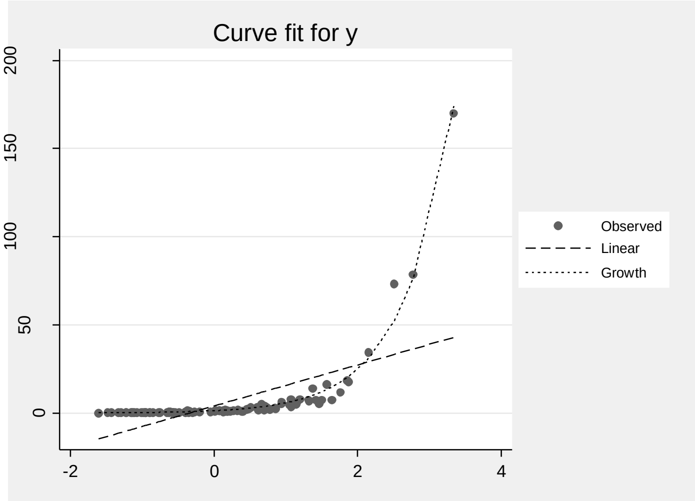
Answer *both* of the following questions. (15 points each, 30 points total.) In each of the following problems, a researcher runs through a sequence of commands. Explain why she didn't stop after the first command, i.e. explain what the purpose of each subsequent command was, what it told her, and why 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? Include diagrams or scatterplots that describe the relationships if they have not already been provided in the problem.

1.

```
. reg y c.age

      Source |       SS       df       MS              Number of obs =   10337
-------------+------------------------------           F(  1, 10335) =   15.53
       Model |  3656.60319     1  3656.60319           Prob > F      =  0.0001
    Residual |  2433370.65 10335  235.449506           R-squared     =  0.0015
-------------+------------------------------           Adj R-squared =  0.0014
       Total |  2437027.25 10336    235.7805           Root MSE      =  15.344

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .034547   .0087664     3.94   0.000     .0173632    .0517309
       _cons |    70.2577    .443435   158.44   0.000     69.38848    71.12691
------------------------------------------------------------------------------

. estat ovtest

Ramsey RESET test using powers of the fitted values of y
       Ho:  model has no omitted variables
               F(3, 10332) =     65.30
                  Prob > F =      0.0000

. reg y c.age c.age#c.age

      Source |       SS       df       MS              Number of obs =   10337
-------------+------------------------------           F(  2, 10334) =  104.31
       Model |  48224.7286     2  24112.3643           Prob > F      =  0.0000
    Residual |  2388802.52 10334  231.159524           R-squared     =  0.0198
-------------+------------------------------           Adj R-squared =  0.0196
       Total |  2437027.25 10336    235.7805           Root MSE      =  15.204

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .9165035   .0641083    14.30   0.000     .7908388    1.042168
             |
 c.age#c.age |  -.0094794   .0006827   -13.89   0.000    -.0108176   -.0081412
             |
       _cons |   52.56348   1.347931    39.00   0.000     49.92127    55.20568
------------------------------------------------------------------------------

. estat ovtest

Ramsey RESET test using powers of the fitted values of y
       Ho:  model has no omitted variables
               F(3, 10331) =      1.09
                  Prob > F =      0.3523
```

2.

```
. reg y x

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =   53.34
       Model |  14049.5785     1  14049.5785           Prob > F      =  0.0000
    Residual |  25810.4821    98  263.372267           R-squared     =  0.3525
-------------+------------------------------           Adj R-squared =  0.3459
       Total |  39860.0606    99  402.626875           Root MSE      =  16.229

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   11.65543   1.595811     7.30   0.000     8.488591    14.82226
       _cons |   4.036725   1.644938     2.45   0.016     .7723995    7.301051
------------------------------------------------------------------------------

. curvefit y x, f(1 0)

Curve Estimation between y and x

------------------------------------------
    Variable |   Linear        Growth     
-------------+----------------------------
b0           |
       _cons |   4.0367252     .31302195  
             |        2.45          4.04  
             |      0.0159        0.0001  
-------------+----------------------------
b1           |
       _cons |   11.655426     1.4498163  
             |        7.30         58.10  
             |      0.0000        0.0000  
-------------+----------------------------
Statistics   |
           N |         100           100  
        r2_a |   .34586516      .9826695  
------------------------------------------
                             legend: b/t/p
```



```
. glm y x, link(log)

Generalized linear models                          No. of obs      =       100
Optimization     : ML                              Residual df     =        98
                                                   Scale parameter =  7.531402
Deviance         =  738.0774104                    (1/df) Deviance =  7.531402
Pearson          =  738.0774104                    (1/df) Pearson  =  7.531402

Variance function: V(u) = 1                        [Gaussian]
Link function    : g(u) = ln(u)                    [Log]

                                                   AIC             =  4.876756
Log likelihood   = -241.8377796                    BIC             =  286.7707

------------------------------------------------------------------------------
             |                 OIM
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.449816   .0237301    61.10   0.000     1.403306    1.496327
       _cons |   .3130218   .0738521     4.24   0.000     .1682745    .4577692
------------------------------------------------------------------------------
```