Sociology 63993
Exam 2 Answer Key
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.

False. While the direct effect of race on income does not significantly differ from 0, race could have an indirect effect, e.g. race affects education which in turn affects income. Remember that a simple regression model like this is only telling you the estimated direct effect, not any possible indirect effects.

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.

True. While the effect of education is smaller for men than for women, it is still positive for both (.0776 for women, .0776 - .0324 = .0452 for men).

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.

False (unless, say, x4 = x2 + x3, but nothing in the problem indicates that this is the case). The second model is not a special/constrained case of the first model (i.e. the models are not nested), so it is not appropriate to use incremental F tests or Likelihood Ratio tests to compare them.

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

False. The variables are measured in totally different metrics, so it is kind of silly to test whether their slope coefficients are equal. Instead, she might want to look at something like the standardized coefficients or the squared semipartials.

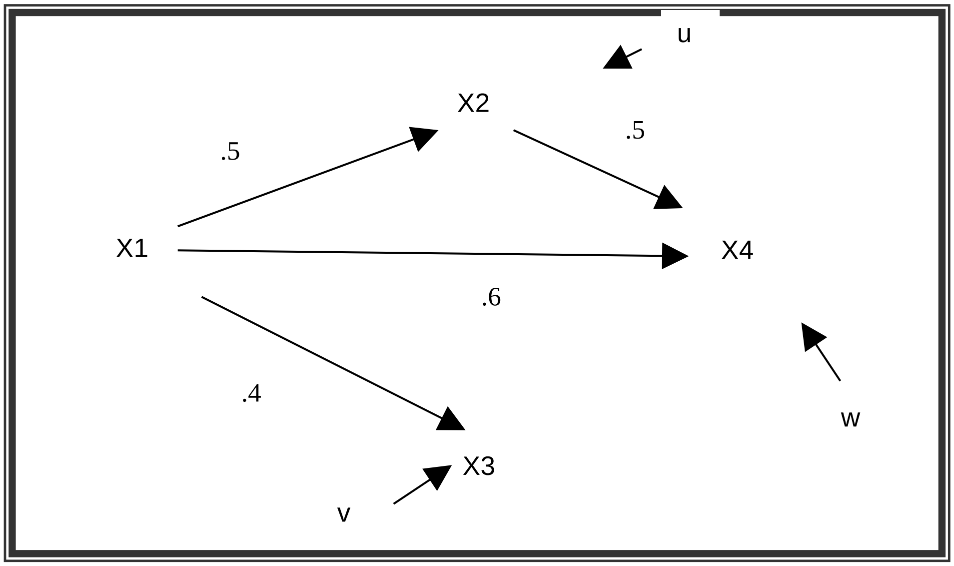
5. A researcher has inadvertently omitted an important variable from her model. Fortunately, as the sample size gets bigger and bigger, the omitted variable bias will diminish and eventually disappear.

False. The formula for omitted variable bias does not include sample size, so the bias is the same regardless of the sample size:

$$E(b_1*) = \beta_1 + \beta_2 \frac{\sigma_{12}}{\sigma_1^2}$$

A larger sample size can help if the model includes extraneous variable. Extraneous variables increase standard errors while larger sample sizes reduce them.

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.

$$X_2 = \beta_{21}X_1 + u = .5X_1 + u$$
$$X_3 = \beta_{31}X_1 + v = .4X_1 + v$$
$$X_4 = \beta_{41}X_1 + \beta_{42}X_2 + w = .6X_1 + .5X_2 + w$$

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

Here is the complete correlation matrix:

```
. corr
(obs=100)
```

	x1	x2	x3	x4
x1	1.0000			
x2	0.5000	1.0000		
x3	***0.4000***	***0.2000***	1.0000	
x4	***0.8500***	***0.8000***	***0.3400***	1.0000

To compute by hand,

$$\rho_{31} = \beta_{31} + \beta_{21}\beta_{32} = .4 + (.5 * .0) = .4$$
$$\rho_{32} = \beta_{32} + \beta_{31}\beta_{21} = 0 + (.4 * .5) = .2$$
$$\rho_{41} = \beta_{41} + \beta_{21}\beta_{42} + \beta_{31}\beta_{43} + \beta_{21}\beta_{32}\beta_{43} = .6 + (.5 * .5) + (.5 * 0 * 0) = .85$$
$$\rho_{42} = \beta_{42} + \beta_{32}\beta_{43} + \beta_{41}\beta_{21} + \beta_{43}\beta_{31}\beta_{21} = .5 + (0 * 0) + (.6 * .5) + (0 * .4 * .5) = .80$$
$$\rho_{43} = \beta_{43} + \beta_{41}\beta_{31} + \beta_{42}\beta_{32} + \beta_{41}\beta_{21}\beta_{32} + \beta_{42}\beta_{21}\beta_{31} = 0 + (.6 * .4) + (.5 * .0) + (.6 * .4 * .0) + (.5 * .4 * .4) = .34$$

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

- Correlation due to direct effects

  .5

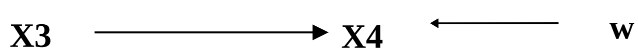
- Correlation due to indirect effects

  0

- Correlation due to common causes

  .30

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.

The researcher would conclude that the direct effect of X3 on X4 is .34 (the same as their correlation). In reality, the model shows that the direct effect of X3 on X4 is zero. There is omitted variable bias because X1 and X2 should be in the model but are not. The correlation between X3 and X4 is due to the fact that X1 is a common cause of both of them. The researcher will therefore believe that increasing X3 will lead to increases in X4, when in reality X3 has neither a direct nor indirect effect on X4.

To confirm above results using Stata commands,

```
. * Problem II, Path analysis
. clear all
. matrix input corr = (1,.5,.4,.85\.5,1,.2,.80\.4,.2,1,.34\.85,.80,.34,1)
. corr2data x1 x2 x3 x4, corr(corr) n(100) clear
(obs 100)
```

```
. corr
(obs=100)

             |       x1       x2       x3       x4
-------------+------------------------------------
          x1 |   1.0000
          x2 |   0.5000   1.0000
          x3 |   0.4000   0.2000   1.0000
          x4 |   0.8500   0.8000   0.3400   1.0000

. pathreg (x2 x1) (x3 x1 x2) (x4 x1 x2 x3)

------------------------------------------------------------------------------
          x2 |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
          x1 |         .5   .0874818     5.72   0.000                       .5
       _cons |   1.42e-09   .0870433     0.00   1.000                        .
------------------------------------------------------------------------------
                 n = 100  R2 = 0.2500  sqrt(1 - R2) = 0.8660

------------------------------------------------------------------------------
          x3 |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
          x1 |         .4   .1074541     3.72   0.000                       .4
          x2 |   8.48e-10   .1074541     0.00   1.000                 8.48e-10
       _cons |  -1.14e-09   .0925916    -0.00   1.000                        .
------------------------------------------------------------------------------
                 n = 100  R2 = 0.1600  sqrt(1 - R2) = 0.9165

------------------------------------------------------------------------------
          x4 |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
          x1 |         .6   .0377964    15.87   0.000                       .6
          x2 |         .5   .0353553    14.14   0.000                       .5
          x3 |   2.75e-09   .0334077     0.00   1.000                 2.75e-09
       _cons |  -4.87e-09   .0304651    -0.00   1.000                        .
------------------------------------------------------------------------------
                 n = 100  R2 = 0.9100  sqrt(1 - R2) = 0.3000

. sem (x2 <- x1) (x3 <- x1 x2) (x4 <- x1 x2 x3)

Endogenous variables

Observed:  x2 x3 x4

Exogenous variables

Observed:  x1

Fitting target model:

Iteration 0:   log likelihood = -422.06629
Iteration 1:   log likelihood = -422.06629

Structural equation model                       Number of obs      =       100
Estimation method  = ml
Log likelihood     = -422.06629
```

```
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  x2 <-      |
          x1 |         .5   .0866025     5.77   0.000     .3302621    .6697379
       _cons |   1.42e-09   .0861684     0.00   1.000     -.168887     .168887
  -----------+----------------------------------------------------------------
  x3 <-      |
          x2 |   8.48e-10   .1058301     0.00   1.000    -.2074231    .2074231
          x1 |         .4   .1058301     3.78   0.000     .1925769    .6074231
       _cons |  -1.14e-09   .0911921    -0.00   1.000    -.1787332    .1787332
  -----------+----------------------------------------------------------------
  x4 <-      |
          x2 |         .5    .034641    14.43   0.000     .4321049    .5678952
          x3 |   2.75e-09   .0327327     0.00   1.000    -.0641549    .0641549
          x1 |         .6   .0370328    16.20   0.000      .527417     .672583
       _cons |  -4.87e-09   .0298496    -0.00   1.000    -.0585042    .0585042
-------------+----------------------------------------------------------------
    var(e.x2)|      .7425   .1050054                      .5627537    .9796581
    var(e.x3)|      .8316    .117606                      .6302842    1.097217
    var(e.x4)|      .0891   .0126006                      .0675304     .117559
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(0)   =      0.00, Prob > chi2 =      .

. estat teffects


Direct effects
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  x2 <-      |
          x1 |         .5   .0866025     5.77   0.000     .3302621    .6697379
  -----------+----------------------------------------------------------------
  x3 <-      |
          x2 |   8.48e-10   .1058301     0.00   1.000    -.2074231    .2074231
          x1 |         .4   .1058301     3.78   0.000     .1925769    .6074231
  -----------+----------------------------------------------------------------
  x4 <-      |
          x2 |         .5    .034641    14.43   0.000     .4321049    .5678952
          x3 |   2.75e-09   .0327327     0.00   1.000    -.0641549    .0641549
          x1 |         .6   .0370328    16.20   0.000      .527417     .672583
------------------------------------------------------------------------------
```

```
Indirect effects
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  x2 <-      |
          x1 |          0  (no path)
  -----------+----------------------------------------------------------------
  x3 <-      |
          x2 |          0  (no path)
          x1 |   4.24e-10    .052915     0.00   1.000    -.1037115    .1037115
  -----------+----------------------------------------------------------------
  x4 <-      |
          x2 |   1.11e-16  (constrained)
          x3 |          0  (no path)
          x1 |        .25   .0484399     5.16   0.000     .1550594    .3449406
------------------------------------------------------------------------------
```

```
Total effects
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  x2 <-      |
          x1 |         .5   .0866025     5.77   0.000     .3302621    .6697379
  -----------+----------------------------------------------------------------
  x3 <-      |
          x2 |   8.48e-10   .1058301     0.00   1.000    -.2074231    .2074231
          x1 |         .4   .0916515     4.36   0.000     .2203663    .5796337
  -----------+----------------------------------------------------------------
  x4 <-      |
          x2 |         .5    .034641    14.43   0.000     .4321049    .5678952
          x3 |   2.75e-09   .0327327     0.00   1.000    -.0641549    .0641549
          x1 |        .85   .0526783    16.14   0.000     .7467525    .9532475
------------------------------------------------------------------------------
```

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
```

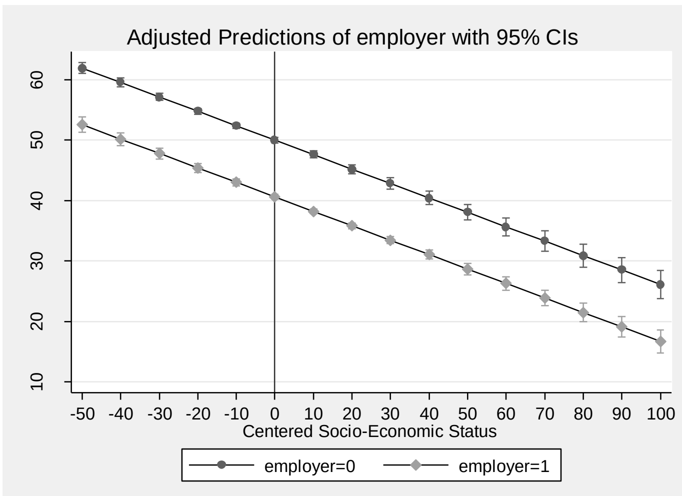
The initial t-test shows that those with employer-provided health insurance have significantly lower levels of support for the Affordable Care Act. Based on the remaining results, explain to the Democratic Party officials why that is the case. When thinking about your answers, keep in mind the various 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?

The interaction term in Model 3 is statistically insignificant so there is no need to include it. However, Model 2 is a statistically significant improvement over Model 1, so we should prefer it. Model 2 says that the intercepts differ across the two groups (insured by employer, and not insured) but the effect of SES does not. According to Model 2, on an all other things equal basis those with higher levels of SES tend to be less supportive of the ACA. Also, on an all other things equal basis, those with insurance from their employer also tend to be less supportive. These results would not be hard to believe. Those with higher SES, and those with insurance through their employers, are probably less likely to need the benefits provided by the ACA and may also have to bear some of the costs of insuring others.

The following graph will also help to show the relationships. It plots the predicted lines separately for those with employer insurance and those without. The line at x = 0 helps with the next question.

```
. quietly reg aca ses i.employer
. quietly margins employer, at(ses = (-50(10)100))
. marginsplot, scheme(sj) xline(0)

  Variables that uniquely identify margins: ses employer
```



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?

Because SES is centered, an average person has a score of zero on SES. Hence, SES drops out of the calculations and we just need to look at the constant and the coefficient for employer. Those without employer insurance have a value of zero on employer, so their predicted score on ACA is just the value of the constant, 49.97. Those with employer insurance have a value of 1 on employer, so their predicted score on ACA is constant + $b_{employer}$ = 49.97 – 9.38 = 40.59. The above graph (see the line where ses = 0) also shows this. We can further check our calculations via

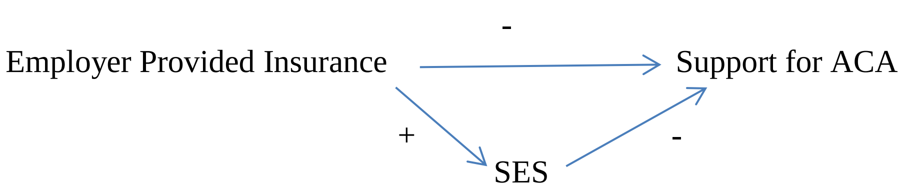
```
. quietly reg aca ses i.employer
. margins employer, at(ses = 0)

Adjusted predictions                              Number of obs   =       4432
Model VCE    : OLS

Expression   : Linear prediction, predict()
at           : ses             =           0

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    employer |
          0  |   49.96529   .2393692   208.74   0.000     49.49601    50.43457
          1  |   40.58618   .2265515   179.15   0.000     40.14203    41.03033
------------------------------------------------------------------------------
```

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?

Those with employer provided insurance also have a significantly higher average SES score (18.52 points) than those who do not have such insurance. Their higher SES, in turn, lowers their support for the ACA. Hence, even though the effect of SES is the same for both groups, the differences in their levels of SES further adds to their differences in ACA support.

That is, those with insurance through their employers are less supportive of the ACA because (a) the variable employer has a negative direct effect on ACA support (a difference in effects, specifically, a difference in the intercepts for the two groups), and (b) ses also has a negative direct effect, and those with employer insurance have higher average levels of ses (a difference in composition; those with employer insurance have more of the things that tend to lower support for the ACA).

One other way of thinking about it: Employer provided insurance has a negative direct effect on support for the ACA. It may also have a negative indirect effect: Those with employer provided insurance tend to have higher levels of SES, while those with higher levels of SES have lower levels of support for the ACA.



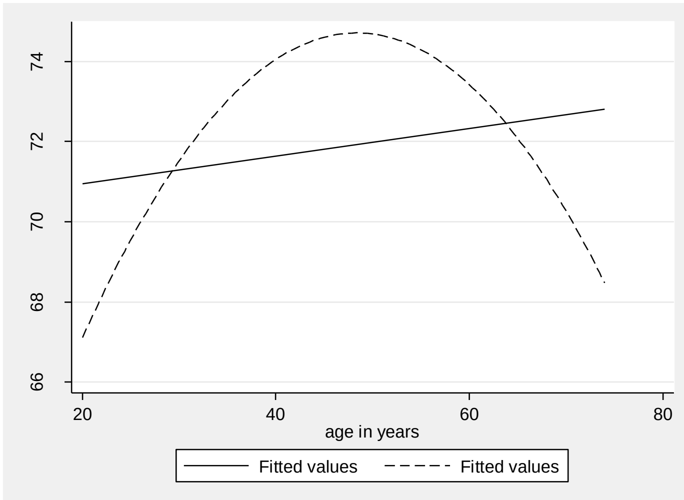
IV. Short answer. Answer *both* of the following questions. (15 points each, 30 points total.) In each of the following problems, a researcher runs through a sequence of commands. Explain why she didn't stop after the first command, i.e. explain what the purpose of each subsequent command was, what it 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
```

The researcher started by estimating a model in which age has a linear effect on y. However, she apparently suspected that the effect might be curvilinear, e.g. maybe y initially increases with increases in age but, after some point, additional increases in age actually cause y to decrease. The ovtest command basically tested whether model fit would be improved by adding age^2, age^3, and age^4 to the model. The test statistic was highly significant, so she decided to add age^2. The subsequent ovtest indicated that no additional polynomial terms were needed, so she stopped. She may have also thought that her theory justified a squared term but higher order polynomials made no sense.

Here is a graph of what the linear and quadratic relationships looks like.

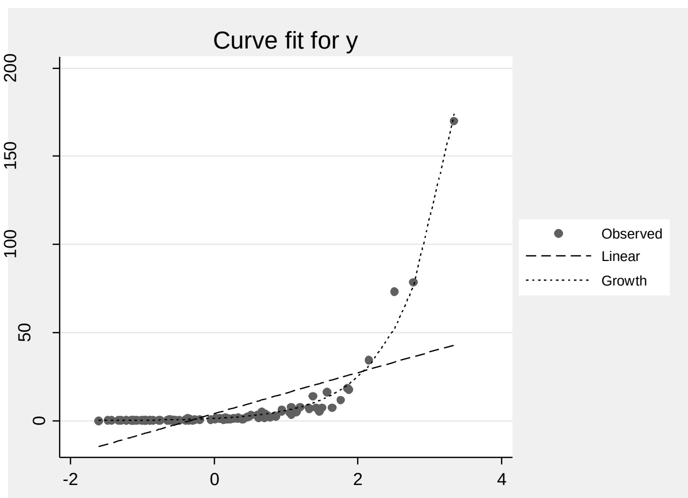


If she had simply estimated the linear model, she would have missed the curvilinear relationship. She would have thought that increases in age always produce increases in Y. She would have initially overestimated the predicted values of Y, then underestimated them, and then gone back to overestimating again.

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
------------------------------------------------------------------------------
```

The researcher initially estimated a model where x had a linear effect on y. However, she then used curvefit to also estimate an exponential growth model and plotted the observed points and the lines for the linear and the growth models. The observed points corresponded much more closely to the growth model than to the linear model, so she went with it. Specifically, she estimated a generalized linear model with link log. As the graph shows, had she stuck with the linear model, she would initially underestimate the values for y, then overestimate them, then go back to underestimating them.

Appendix: Stata Code used in the exam

```
version 12.1
* Problem I - 1
clear all
matrix input corr = (1,.3,.2101\.3,1,.7\.2101,.7,1)
corr2data black educ income, corr(corr) n(534) sd(.1 3.2 16) clear
pathreg (educ black) (income black educ)
reg income black educ

* Problem I - 2
sysuse ordwarm2,clear
gen edmale = ed * male
reg warm male ed edmale


* Problem II, Path analysis
clear all
matrix input corr = (1,.5,.4,.85\.5,1,.2,.80\.4,.2,1,.34\.85,.80,.34,1)
corr2data x1 x2 x3 x4, corr(corr) n(100) clear
corr
pathreg (x2 x1) (x3 x1 x2) (x4 x1 x2 x3)
sem (x2 <- x1) (x3 <- x1 x2) (x4 <- x1 x2 x3)
estat teffects

* Part III – Interaction effects
* Generate the variables by manipulating nhanes2f
* The manipulations produce the kind of relationships desired for the problem!
```

```
clear all
webuse nhanes2f, clear
keep health weight female
keep if !missing(health, weight, female)
set seed 123456
sample 4432, count
gen employer = female
replace weight = weight + (30 * employer)
center weight, gen(ses)
label variable ses "Centered Socio-Economic Status"
gen empses = employer * ses
gen aca = (rnormal(0, 30) - .7*ses - 30*employer - .01* empses + 150) / 3
label variable aca "Support for Affordable Care Act"

* Do analyses
ttest aca, by(employer)
nestreg: reg aca ses employer empses
ttest ses, by(employer)
* Additional analysis. This will plot the relationships
quietly reg aca ses i.employer
quietly margins employer, at(ses = (-50(10)100))
marginsplot, scheme(sj) xline(0)
quietly reg aca ses i.employer
margins employer, at(ses = 0)


* Problem IV - 1
webuse nhanes2f, clear
clonevar y = weight
reg y c.age
estat ovtest
reg y c.age c.age#c.age
estat ovtest
twoway lfit y age || qfit y age , sort scheme(sj)

* Problem IV - 2
clear all
set obs 100
set seed 12345
gen x = rnormal()
gen e = rnormal()
gen y = exp(1.5*x+.3*e)
reg y x
curvefit y x, f(1 0)
* Graph was manually converted to SJ scheme
glm y x, link(log) nolog
```