Sociology 63993, Exam 2
March 27, 2015

Richard Williams, University of Notre Dame, http://www3.nd.edu/~rwilliam/

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

1. A researcher has inadvertently included an extraneous variable in her model. Unfortunately, increasing the sample size will not help to reduce the problems this creates.

2. A researcher regresses income on the respondent's gender, years of education, IQ, and mother's education (i.e. the number of years of education the respondent's mother had). The estimated effect of mother's education is 0 and is statistically insignificant. This means that, in terms of their own income, respondents gain no benefit from having a better educated mother.

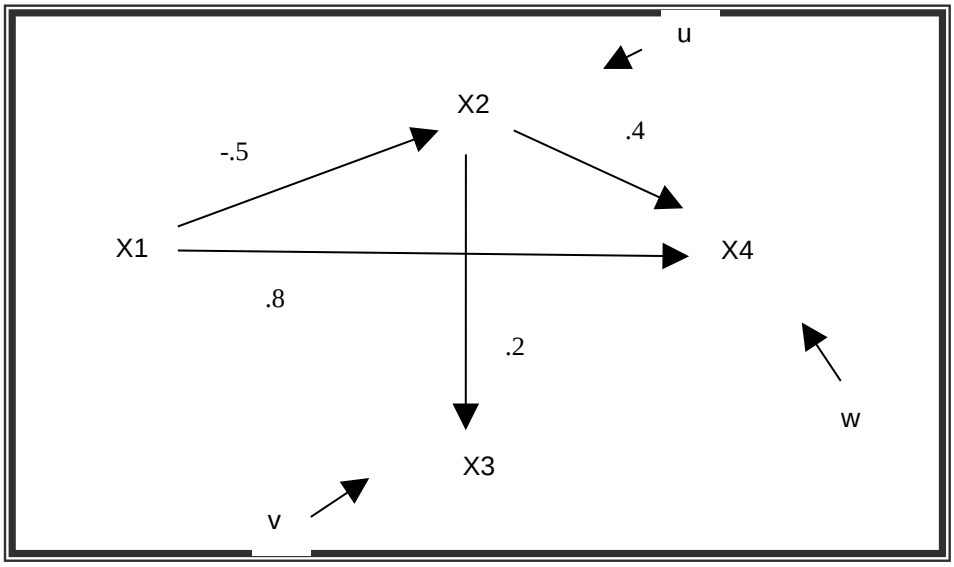
3. Personal Fulfillment (measured on a 200 point scale) is regressed on Income, Female, and Female*Income. All terms are positive and statistically significant. The coefficient for Female is +12. This means that, whenever a man and a woman have equal incomes, the woman is expected to score 12 points higher than the man on Personal Fulfillment.

4. A researcher believes that X2 and X3 are positively correlated only because X1 is a common cause of both, i.e. X2 does not directly or indirectly affect X3, nor does X3 directly or indirectly affect X2. Therefore knowledge of X2 will be of no use to her for predicting X3.

5. A researcher hypothesizes that income positively affects the self-image of men but has a negative effect on the self-image of women. She gets

$$\hat{\beta}_{Income} = 4$$
$$\hat{\beta}_{Female} = 0$$
$$\hat{\beta}_{Income\ *\ Female} = -4$$

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



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.

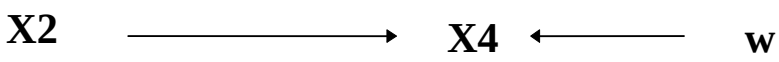
b. (10 pts) Part of the correlation matrix is shown below. Determine the complete correlation matrix. (Remember, variables are standardized. You can use either normal equations or Sewell Wright, but you might want to use both as a double-check.)

	x1	x2	x3	x4
x1	1.0000			
x2	-0.5000	1.0000		
x3	?	?	1.0000	
x4	.?	?	?	1.0000

c. (5 pts) Decompose the correlation between X1 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 X2 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?

III. Group comparisons (25 points). Yik Yak is an anonymous social media app. It allows users to share posts with others who are within a few miles of them, and has become very popular on college campuses, including Notre Dame. However, the presence of racist and sexist posts on Yik Yak has generated great controversy, leading, for example, to 150 Notre Dame faculty signing a letter denouncing some of the postings (http://www.southbendtribune.com/news/notre-dame-profs-students-respond-to-racist-yik-yak-posts/article_b1f2d81a-a311-11e4-ae3a-1b3bd646ee0d.html?_dc=479429358849.30194). Potential investors in Yik Yak are worried that such controversies could undermine the financial prospects for the application. They want to assess how serious the concerns about offensive posts are. They have therefore conducted a study of 2000 randomly selected college students. Participants were asked to use Yik Yak for a month, and then provide answers to the following questions:

Variable	Description
yikyak	How likely is the respondent to use yikyak in the future? Scores range from a low of 1 (definitely will not use) to a high of 150 (certain to use).
offensive	How offensive did the user find yikyak? The original scale ranged from a low of 1(not offensive at all) to a high of 100 (extremely offensive). The scale has been centered so that a score of zero corresponds to an average score on the original measure.
female	Coded 1 if female, 0 if male
femXoffensive	female * offensive.

The results of the analysis are as follows:

. ttest yikyak, by(female)

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     924    63.28586    1.039956    31.61193    61.24491    65.32681
       1 |    1076    44.10861    .9983174    32.74725    42.14974    46.06748
---------+--------------------------------------------------------------------
combined |    2000     52.9685    .7515199    33.60899    51.49465    54.44234
---------+--------------------------------------------------------------------
    diff |            19.17725    1.445449                16.34251      22.012
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  13.2673
Ho: diff = 0                     degrees of freedom =     1998

    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 yikyak offensive female femXoffensive

Block 1: offensive

```
      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F(  1,  1998) =  358.06
       Model |   343159.58     1   343159.58           Prob > F      =  0.0000
    Residual |  1914839.55  1998  958.378155           R-squared     =  0.1520
-------------+------------------------------           Adj R-squared =  0.1516
       Total |  2257999.13  1999  1129.56435           Root MSE      =  30.958

------------------------------------------------------------------------------
      yikyak |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   offensive |  -.8789011   .0464473   -18.92   0.000    -.9699913   -.7878109
       _cons |    52.9685   .6922348    76.52   0.000     51.61092    54.32608
------------------------------------------------------------------------------
```

Block 2: female

```
      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F(  2,  1997) =  250.27
       Model |  452528.792     2  226264.396           Prob > F      =  0.0000
    Residual |  1805470.34  1997  904.091308           R-squared     =  0.2004
-------------+------------------------------           Adj R-squared =  0.1996
       Total |  2257999.13  1999  1129.56435           Root MSE      =  30.068

------------------------------------------------------------------------------
      yikyak |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   offensive |  -.7912462   .0458112   -17.27   0.000    -.8810889   -.7014035
      female |  -15.06238   1.369469   -11.00   0.000    -17.74812   -12.37664
       _cons |   61.07206   .9974378    61.23   0.000     59.11593    63.02819
------------------------------------------------------------------------------
```

Block 3: femXoffensive

```
      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F(  3,  1996) =  172.87
       Model |  465693.904     3  155231.301           Prob > F      =  0.0000
    Residual |  1792305.23  1996  897.948512           R-squared     =  0.2062
-------------+------------------------------           Adj R-squared =  0.2050
       Total |  2257999.13  1999  1129.56435           Root MSE      =  29.966

------------------------------------------------------------------------------
       yikyak |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
    offensive |  -.5990285   .0678563    -8.83   0.000    -.7321051   -.4659519
       female |  -15.21819   1.365415   -11.15   0.000    -17.89598   -12.54041
femXoffensive |  -.3512047    .091722    -3.83   0.000    -.5310857   -.1713237
        _cons |   61.60986   1.003917    61.37   0.000     59.64103     63.5787
------------------------------------------------------------------------------
```

Block	F	Block df	Residual df	Pr > F	R2	Change in R2
1	358.06	1	1998	0.0000	0.1520	
2	120.97	1	1997	0.0000	0.2004	0.0484
3	14.66	1	1996	0.0001	0.2062	0.0058

. ttest offensive, by(female)

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     924   -2.797863     .478188    14.53566   -3.736325   -1.859401
       1 |    1076    2.402631    .4514825    14.80973    1.516744    3.288518
---------+--------------------------------------------------------------------
combined |    2000    2.53e-06    .3333394    14.90739   -.6537265    .6537316
---------+--------------------------------------------------------------------
    diff |           -5.200494    .6585821               -6.492073   -3.908914
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -7.8965
Ho: diff = 0                                     degrees of freedom =     1998

    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 women are significantly less likely to keep using Yik Yak in the future. Based on the remaining results, explain to Yik Yak's backers 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 gender and the perceived offensiveness of Yik Yak affect the likelihood of using Yik Yak in the future? What ways (if any) do the determinants of support for Yik Yak differ by gender?

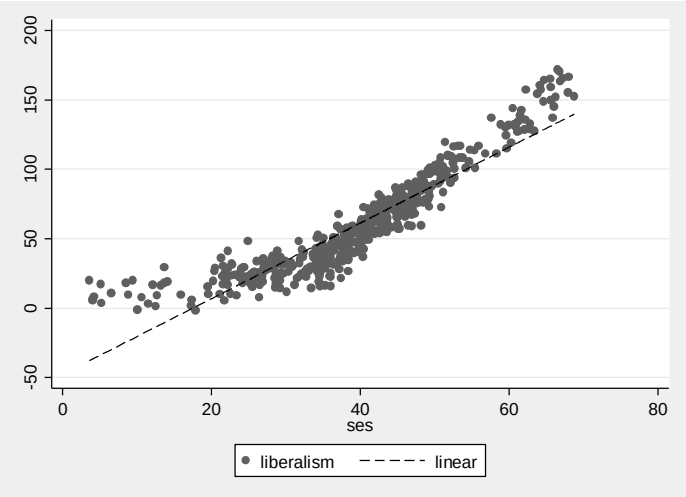
b) (5 pts) According to your preferred model, how does the yikyak score of the "average" male compare to that of the "average" female?

c) (10 pts) The researchers then do one last t-test. What does this test tell us about how feelings on offensiveness differ by gender? What additional insights, if any, does this test give us as to why women are less supportive of Yik Yak?

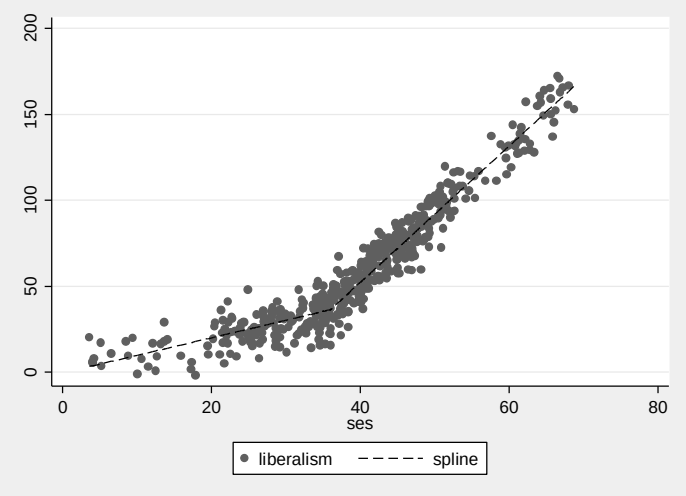
IV. Short answer. Answer *both* of the following questions. (15 points each, 30 points total.) In each of the following problems, a researcher runs through a sequence of commands. Explain why she didn't stop after the first command, i.e. explain what the purpose of each subsequent command was, what it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

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) = 2764.60
       Model |  567994.874     1  567994.874           Prob > F      =  0.0000
    Residual |   102315.49   498  205.452792           R-squared     =  0.8474
-------------+------------------------------           Adj R-squared =  0.8471
       Total |  670310.364   499  1343.30734           Root MSE      =  14.334

------------------------------------------------------------------------------
  liberalism |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ses |   2.729271   .0519075    52.58   0.000     2.627286    2.831256
       _cons |  -47.85882   2.156255   -22.20   0.000     -52.0953   -43.62234
------------------------------------------------------------------------------

. predict linear
(option xb assumed; fitted values)

. label variable linear "linear"

. scatter liberalism ses || line linear ses, scheme(sj) sort
```



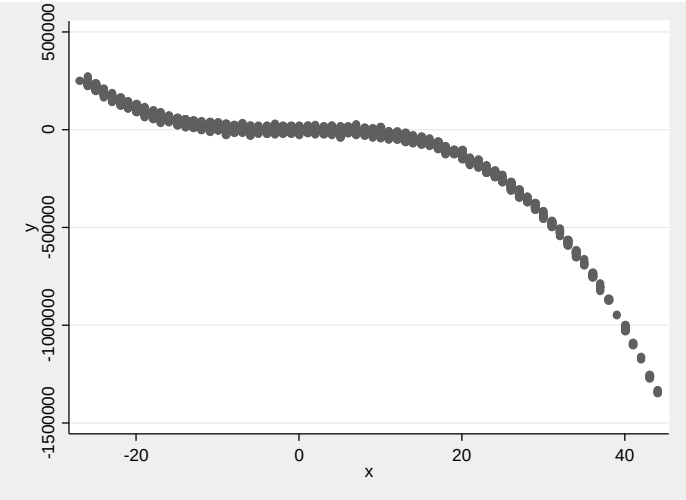
```
. mkspline seslow 36 seshi = ses

. reg liberalism seslow seshi

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  2,   497) = 4195.50
       Model |  632827.848     2  316413.924           Prob > F      =  0.0000
    Residual |  37482.5161   497  75.4175374           R-squared     =  0.9441
-------------+------------------------------           Adj R-squared =  0.9439
       Total |  670310.364   499  1343.30734           Root MSE      =  8.6843

------------------------------------------------------------------------------
  liberalism |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      seslow |   1.022624   .0661605    15.46   0.000     .8926348    1.152612
       seshi |   3.965107   .0525898    75.40   0.000     3.861781    4.068433
       _cons |   -.424041   2.079451    -0.20   0.838    -4.509639    3.661557
------------------------------------------------------------------------------

. predict spline
(option xb assumed; fitted values)

. label variable spline "spline"

. scatter liberalism ses || line spline ses, scheme(sj) sort
```



2.

```
. reg y x

      Source |       SS       df       MS              Number of obs =    2293
-------------+------------------------------           F(  1,  2291) = 4666.20
       Model |  5.9161e+13     1  5.9161e+13           Prob > F      =  0.0000
    Residual |  2.9047e+13  2291  1.2679e+10           R-squared     =  0.6707
-------------+------------------------------           Adj R-squared =  0.6706
       Total |  8.8208e+13  2292  3.8485e+10           Root MSE      =  1.1e+05

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -9575.092   140.1721   -68.31   0.000     -9849.97   -9300.215
       _cons |  -37867.18   2351.439   -16.10   0.000    -42478.35   -33256.01
------------------------------------------------------------------------------
```

```
. scatter y x, scheme(sj)
```




```
. reg y x c.x#c.x c.x#c.x#c.x

      Source |       SS       df       MS              Number of obs =    2293
-------------+------------------------------           F(  3,  2289) =       .
       Model |  8.7979e+13     3  2.9326e+13           Prob > F      =  0.0000
    Residual |  2.2843e+11  2289  99793436.9           R-squared     =  0.9974
-------------+------------------------------           Adj R-squared =  0.9974
       Total |  8.8208e+13  2292  3.8485e+10           Root MSE      =  9989.7

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -5.618982   24.07808    -0.23   0.815    -52.83611    41.59815
             |
     c.x#c.x |  -30.70687   .9647104   -31.83   0.000    -32.59867   -28.81507
             |
 c.x#c.x#c.x |  -15.02577   .0391987  -383.32   0.000    -15.10264    -14.9489
             |
       _cons |   249.4734   309.3394     0.81   0.420    -357.1414    856.0882
------------------------------------------------------------------------------

. ovtest

Ramsey RESET test using powers of the fitted values of y
       Ho:  model has no omitted variables
                F(3, 2286) =      0.14
                  Prob > F =      0.9390
```