Sociology 63993
Exam 2 Answer Key
April 5, 2013

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

1. When analyzing data sets with complicated sampling schemes (e.g. `svyset` data in Stata) incremental F tests, rather than Wald tests, should be used to make comparisons of nested models.

False. With survey data, assumptions that cases are independent of each other are violated. You should use Wald tests instead of incremental F tests.

2. An exponential/growth model can be appropriate if it is thought that the slope of the effect of X on E(Y) changes sign as X increases.

False. While the amount of change produced by X can get bigger or smaller as X increases, the effect can't change signs. Use polynomial models instead.

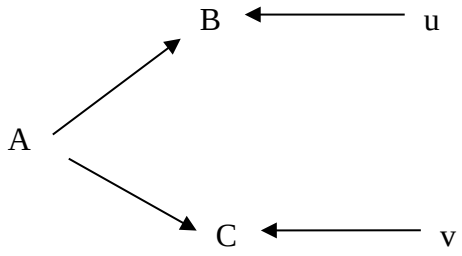
3. A researcher is not sure whether certain variables should be included in her model or not. She might as well include them because there are no adverse consequences to including extraneous variables in a model.

False. Extraneous variables increase standard errors, causing estimates to be less precise and increasing the likelihood that non-zero effects will be judged as insignificant.

4. A researcher regresses income on education, race (coded 1 = white, 0 otherwise), and occupational prestige. If the effect of race is 0, this means that whites and non-whites have the same mean levels of income.

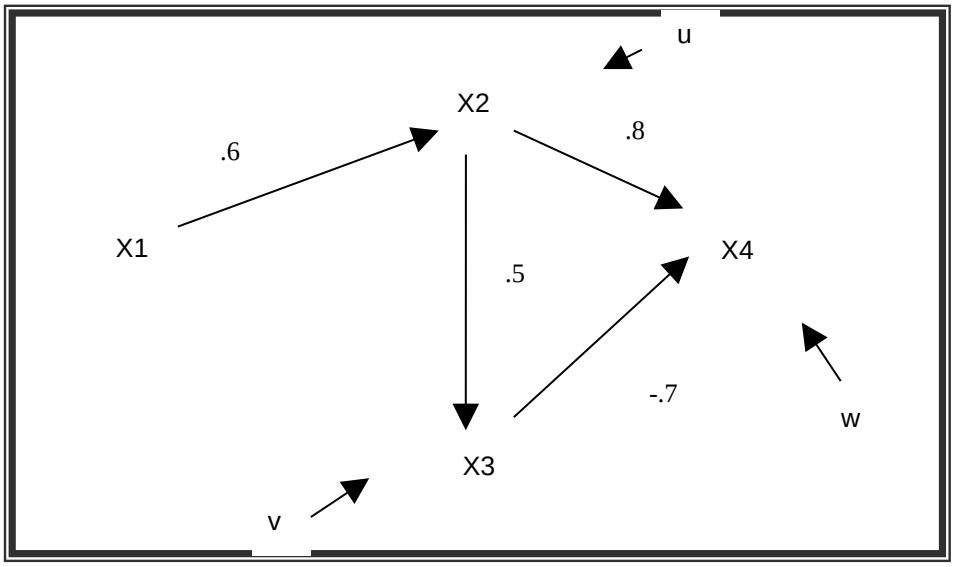
False. If blacks and whites have different mean levels of education and occupational prestige, they will also probably have different mean levels of income.

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



False. B and C will be correlated because of the common cause A. If A has positive effects on both, then those with higher scores on B will tend to have higher scores on C.

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.

$$X_2 = \beta_{21}X_1 + u = .6X_1 + u$$
$$X_3 = \beta_{32}X_2 + v = .5X_2 + v$$
$$X_4 = \beta_{42}X_2 + \beta_{43}X_3 + w = .8X_2 - .7X_3 + w$$

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.)

```
. corr
(obs=100)

             |       x1       x2       x3       x4
-------------+------------------------------------
          x1 |   1.0000
          x2 |   0.6000   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 |   0.6000   1.0000
          x3 |   0.3000   0.5000   1.0000
          x4 |   0.2700   0.4500  -0.3000   1.0000
```

To confirm by hand,

$$\rho_{31} = \beta_{31} + \beta_{21}\beta_{32} = 0 + .6*.5 = .30$$

$$\rho_{32} = \beta_{32} + \beta_{31}\beta_{21} = .5 + 0*.6 = .5$$

$$\rho_{41} = \beta_{41} + \beta_{21}\beta_{42} + \beta_{31}\beta_{43} + \beta_{21}\beta_{32}\beta_{43} = 0 + .6*.8 + .6*.5*-.7 = .27$$

$$\rho_{42} = \beta_{42} + \beta_{32}\beta_{43} + \beta_{41}\beta_{21} + \beta_{43}\beta_{31}\beta_{21} = .8 + .5*-.7 + 0*.6 + -.7*0*.6 = .45$$

$$\rho_{43} = \beta_{43} + \beta_{41}\beta_{31} + \beta_{42}\beta_{32} + \beta_{41}\beta_{21}\beta_{32} + \beta_{42}\beta_{21}\beta_{31} = -.7 + 0*0 + .8*.5 + 0*.6*.5 + .8*.6*0 = -.30$$

To confirm using pathreg and sem,

```
. matrix input corr = (1,.6,.3,.27\.6,1,.5,.45\.3,.5,1,-.3\.27,.45,-.3,1)
. corr2data x1 x2 x3 x4, corr(corr) n(100)
(obs 100)
. * confirm with pathreg. Must be installed

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

------------------------------------------------------------------------------
          x2 |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
          x1 |         .6   .0808122     7.42   0.000                       .6
       _cons |   4.55e-09   .0804071     0.00   1.000                        .
------------------------------------------------------------------------------
                n = 100  R2 = 0.3600  sqrt(1 - R2) = 0.8000

------------------------------------------------------------------------------
          x3 |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
          x1 |   6.33e-09   .1099144     0.00   1.000                 6.33e-09
          x2 |         .5   .1099144     4.55   0.000                       .5
       _cons |  -6.83e-09   .0874908    -0.00   1.000                        .
------------------------------------------------------------------------------
                n = 100  R2 = 0.2500  sqrt(1 - R2) = 0.8660

------------------------------------------------------------------------------
          x4 |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
          x1 |   1.15e-08   .0836582     0.00   1.000                 1.15e-08
          x2 |         .8   .0921507     8.68   0.000                       .8
          x3 |        -.7   .0772802    -9.06   0.000                      -.7
       _cons |  -9.48e-09   .0665911    -0.00   1.000                        .
------------------------------------------------------------------------------
                n = 100  R2 = 0.5700  sqrt(1 - R2) = 0.6557
```

```
. * Confirm with sem
. sem (x1 -> x2) (x1 x2 -> x3) (x1 x2 x3 -> x4)

Endogenous variables

Observed:  x2 x3 x4

Exogenous variables

Observed:  x1

Fitting target model:

Iteration 0:   log likelihood = -486.66839
Iteration 1:   log likelihood = -486.66839

Structural equation model                       Number of obs      =       100
Estimation method  = ml
Log likelihood     = -486.66839

------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  x2 <-      |
          x1 |         .6        .08     7.50   0.000     .4432029    .7567971
       _cons |   4.55e-09    .079599     0.00   1.000    -.1560112    .1560112
  -----------+----------------------------------------------------------------
  x3 <-      |
          x2 |         .5   .1082532     4.62   0.000     .2878277    .7121723
          x1 |   6.33e-09   .1082532     0.00   1.000    -.2121723    .2121723
       _cons |  -6.83e-09   .0861684    -0.00   1.000     -.168887     .168887
  -----------+----------------------------------------------------------------
  x4 <-      |
          x2 |         .8   .0902889     8.86   0.000      .623037     .976963
          x3 |        -.7   .0757188    -9.24   0.000    -.8484061   -.5515939
          x1 |   1.15e-08    .081968     0.00   1.000    -.1606543    .1606543
       _cons |  -9.48e-09   .0652457    -0.00   1.000    -.1278792    .1278792
-------------+----------------------------------------------------------------
Variance     |
        e.x2 |      .6336   .0896046                      .4802165    .8359749
        e.x3 |      .7425   .1050054                      .5627537    .9796581
        e.x4 |      .4257   .0602031                      .3226455    .5616707
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(0)   =      0.00, Prob > chi2 =      .
```

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

- Correlation due to direct effects

.8

- Correlation due to indirect effects

-.35

- Correlation due to common causes

0

Note that we can confirm the estimates of the direct and indirect effects with the following post-estimation command after the above sem command:

```
. estat teffects
```

Direct effects

	OIM				
	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]
Structural					
x2 <-					
x1	.6	.08	7.50	0.000	.4432029 .7567971
x3 <-					
x2	.5	.1082532	4.62	0.000	.2878277 .7121723
x1	6.33e-09	.1082532	0.00	1.000	-.2121723 .2121723
x4 <-					
x2	***.8***	.0902889	8.86	0.000	.623037 .976963
x3	-.7	.0757188	-9.24	0.000	-.8484061 -.5515939
x1	1.15e-08	.081968	0.00	1.000	-.1606543 .1606543

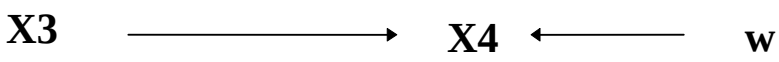
Indirect effects

	OIM				
	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]
Structural					
x2 <-					
x1	0	(no path)			
x3 <-					
x2	0	(no path)			
x1	.3	.0762807	3.93	0.000	.1504925 .4495075
x4 <-					
x2	***-.35***	.0757772	-4.62	0.000	-.4985206 -.2014794
x3	0	(no path)			
x1	.27	.0859637	3.14	0.002	.1015143 .4384857

Total effects

	OIM				
	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]
Structural					
x2 <-					
x1	.6	.08	7.50	0.000	.4432029 .7567971
x3 <-					
x2	.5	.1082532	4.62	0.000	.2878277 .7121723
x1	.3	.0953939	3.14	0.002	.1130314 .4869687
x4 <-					
x2	***.45***	.117874	3.82	0.000	.2189713 .6810287
x3	-.7	.0757188	-9.24	0.000	-.8484061 -.5515939
x1	.27	.096286	2.80	0.005	.0812828 .4587172

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 less than half as large as the effect found in the correct model of -.7. 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). Opponents of gay marriage are disheartened by recent events. An ABC News/Washington Post poll shows that 58% of Americans now believe gay marriage should be legal, up from 32% less than a decade ago. Hundreds of prominent figures from across the political spectrum, including former Republican Presidential candidate Jon Huntsman, Mitt Romney advisor Senator Rob Portman, former Secretary of State Hillary Clinton, and actor Clint Eastwood have announced their support for legalizing gay marriage. Even Barbara Bush, daughter of former president George W. Bush, has made an ad in support of marriage equality. New York Times columnist Frank Bruni has even gone so far as to say that the question isn't whether gay rights advocates will have a happy ending, the question is when.

The opponents are not giving up however. They want to better identify where their support is and what determines attitudes toward gay marriage. They have collected data from a representative sample of 7,000 American adults. The study measured the following variables.

Variable	Description
gaymarr	Scale that measures support for gay marriage. Ranges from strong opposition (-100) to strong support (+100)
relig	Religiosity/ Traditional religious values scale, centered to have a mean of zero. The centered scale ranges from a low of -57 to a high of 102.
older	Equals 1 if the respondent is older than the average age, 0 otherwise
oldrelig	older * relig

The results of the analysis are as follows:

```
. ttest gaymarr, by(older)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |    3315     66.9771    .5401228    31.09813    65.91809    68.03611
       1 |    3685    39.85642    .5184244    31.47054    38.83999    40.87284
---------+--------------------------------------------------------------------
combined |    7000        52.7    .4075389    34.09715     51.9011     53.4989
---------+--------------------------------------------------------------------
    diff |            27.12068    .7491344                25.65215    28.58922
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  36.2027
Ho: diff = 0                                     degrees of freedom =     6998

    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 gaymarr relig older oldrelig

Block  1: relig

      Source |       SS       df       MS              Number of obs =    7000
-------------+------------------------------           F(  1,  6998) = 1631.19
       Model |  1538177.77     1  1538177.77           Prob > F      =  0.0000
    Residual |  6598969.61  6998  942.979367           R-squared     =  0.1890
-------------+------------------------------           Adj R-squared =  0.1889
       Total |  8137147.38  6999  1162.61571           Root MSE      =  30.708

------------------------------------------------------------------------------
     gaymarr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       relig |  -.8720741   .0215924   -40.39   0.000    -.9144018   -.8297464
       _cons |       52.7   .3670304   143.58   0.000     51.98051    53.41949
------------------------------------------------------------------------------
```

Block 2: older

```
      Source |       SS       df       MS              Number of obs =    7000
-------------+------------------------------           F(  2,  6997) = 1015.70
       Model |  1830865.88     2  915432.938           Prob > F      =  0.0000
    Residual |   6306281.5  6997  901.283622           R-squared     =  0.2250
-------------+------------------------------           Adj R-squared =  0.2248
       Total |  8137147.38  6999  1162.61571           Root MSE      =  30.021

------------------------------------------------------------------------------
     gaymarr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       relig |  -.6222939   .0252535   -24.64   0.000    -.6717984   -.5727895
       older |  -15.49284   .8597242   -18.02   0.000    -17.17816   -13.80752
       _cons |   60.85587   .5775695   105.37   0.000     59.72366    61.98808
------------------------------------------------------------------------------
```

Block 3: oldrelig

```
      Source |       SS       df       MS              Number of obs =    7000
-------------+------------------------------           F(  3,  6996) =  699.70
       Model |  1878017.47     3  626005.824           Prob > F      =  0.0000
    Residual |  6259129.9  6996  894.672656           R-squared     =  0.2308
-------------+------------------------------           Adj R-squared =  0.2305
       Total |  8137147.38  6999  1162.61571           Root MSE      =  29.911

------------------------------------------------------------------------------
     gaymarr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       relig |  -.4177053   .0377791   -11.06   0.000    -.4917638   -.3436468
       older |  -16.06223   .8601487   -18.67   0.000    -17.74838   -14.37607
    oldrelig |  -.3676665   .0506452    -7.26   0.000    -.4669463   -.2683866
       _cons |   62.86832    .638737    98.43   0.000      61.6162    64.12044
------------------------------------------------------------------------------
```

Block	F	Block df	Residual df	Pr > F	R2	Change in R2
1	1631.19	1	6998	0.0000	0.1890	
2	324.75	1	6997	0.0000	0.2250	0.0360
3	52.70	1	6996	0.0000	0.2308	0.0058

. ttest relig, by(older)

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |    3315   -9.836558    .2388704    13.75321   -10.30491    -9.36821
       1 |    3685    8.848901    .2406829    14.61046    8.377016    9.320786
---------+--------------------------------------------------------------------
combined |    7000    1.46e-06    .2031809    16.99933   -.3982946    .3982976
---------+--------------------------------------------------------------------
    diff |           -18.68546    .3401821               -19.35232    -18.0186
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t = -54.9278
Ho: diff = 0                                     degrees of freedom =     6998

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

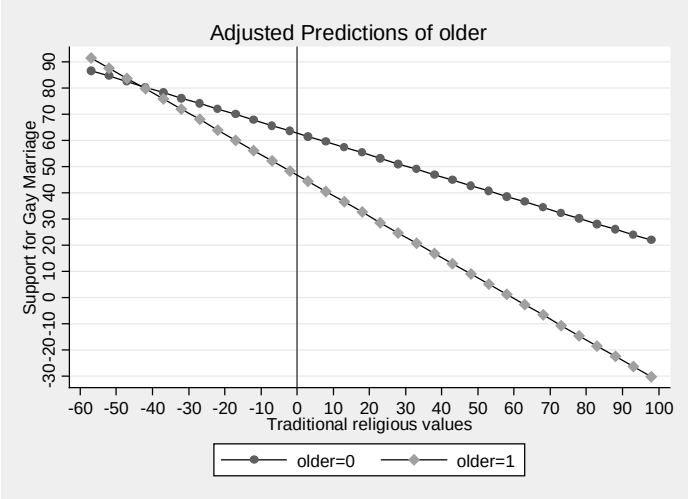
The initial t-test shows that older people have significantly lower levels of support for legalizing gay marriage. Based on the remaining results, explain to the opponents of gay marriage why that is the case. When thinking about your answers, keep in mind the various reasons that two groups can differ 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 age and traditional religious values affect the amount of support for gay marriage? What ways (if any) do the determinants of support for gay marriage differ by age?

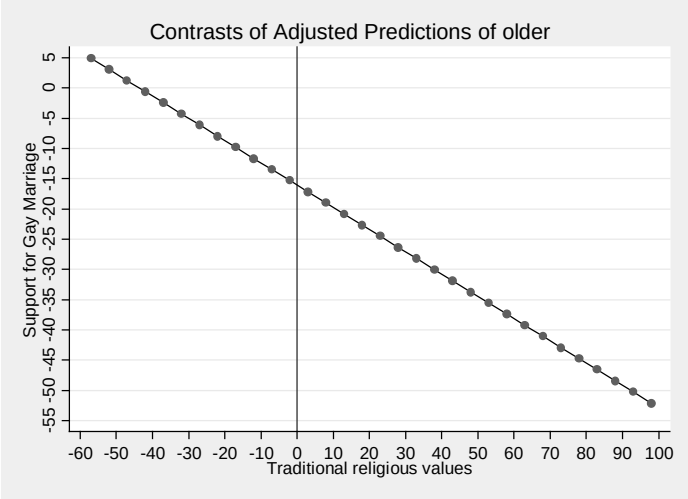
All of the terms in the third model are highly significant, and, from a purely empirical standpoint, it can be considered the best. It makes substantive sense as well. According to the third model, both the slopes and intercepts differ by age. Those with more traditional religious values are less supportive of gay marriage. Older people are also less supportive. Further, as the interaction term shows, the negative effect of religiosity is almost twice as large for older people as it is for younger people.

The following graphs will also help to show the relationships. The first plots the predicted lines separately for older and younger people. The second plots the predicted difference between the two groups. Both make clear that, the higher the religiosity score, the less support there is for gay marriage, and the greater the gap is between the older and younger groups.

```
. quietly reg gaymarr relig i.older i.older#c.relig
. quietly margins older, at(relig = (-57(5)102))
. marginsplot, noci scheme(sj) ylabel(#20) xlabel(#20) ytitle("Support for Gay Marriage") xline(0)
```




```
. quietly margins r.older, at(relig = (-57(5)102))
. marginsplot, noci scheme(sj) ylabel(#20) xlabel(#20) ytitle("Support for Gay Marriage") xline(0)
```



b) (5 pts) According to your preferred model, how does the gay marriage score of the "average" (on relig) older person compare to that of the "average" younger person?

Since relig is centered, the coefficient for older tells us the difference between the average older and younger person (where average is defined as having the mean value on relig). So, the average older person scores about 16 points lower on the gay marriage support scale than the average younger person. The two graphs show this as well. Note, however, that the amount of difference depends on the degree of religiosity: the higher the score on the traditional religious values scale, the greater the gap between the young and the old.

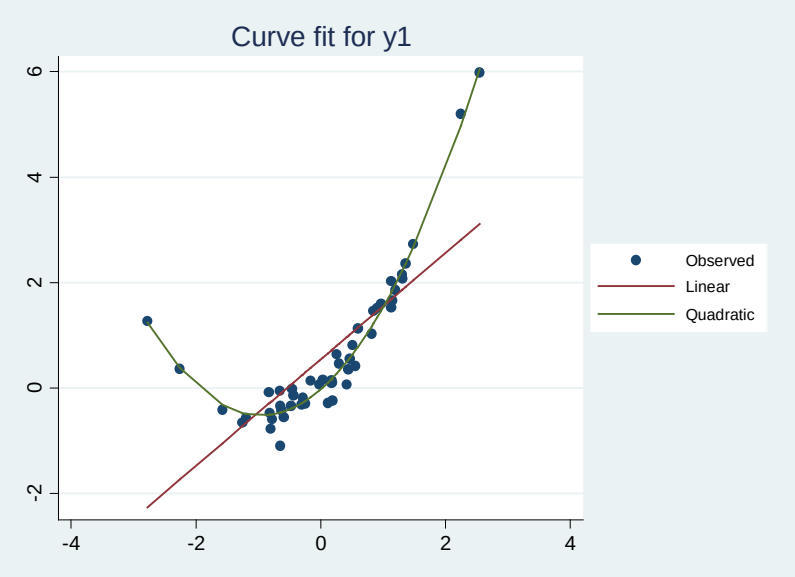
c) (10 pts) The researchers then do one last t-test. What does this test tell us about how religiosity differs by age? What additional insights, if any, does this test give us as to why older people are less supportive of gay marriage?

Older people score significantly higher on the traditional religious values scale, almost 19 points. This further contributes to the overall 27 point gap on the gay marriage measure between the young and the old. Older people have stronger traditional values (which in and of itself would lower their scores on gay marriage) and the effect of those values is stronger for them than it is for the young. In other words, both compositional differences on relig and differences in the effects of relig contribute to the younger/older differences in support for gay marriage.

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 y1 x

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =   62.99
       Model |  53.5506531     1  53.5506531           Prob > F      =  0.0000
    Residual |  40.8061587    48  .850128306           R-squared     =  0.5675
-------------+------------------------------           Adj R-squared =  0.5585
       Total |  94.3568118    49  1.92564922           Root MSE      =  .92202

------------------------------------------------------------------------------
          y1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.008769   .1271018     7.94   0.000     .7532139    1.264325
       _cons |   .5421014    .130987     4.14   0.000     .2787343    .8054685
------------------------------------------------------------------------------

. curvefit y1 x, f(1 4 )

Curve Estimation between y1 and x

-----------------------------------------
    Variable |   Linear       Quadratic
-------------+---------------------------
b0           |
       _cons |   .54210142     -.0255622
             |        4.14         -0.64
             |      0.0001        0.5222
-------------+---------------------------
b1           |
       _cons |   1.0087692     1.0263535
             |        7.94         31.76
             |      0.0000        0.0000
-------------+---------------------------
b2           |
       _cons |                 .53286587
             |                     26.38
             |                    0.0000
-------------+---------------------------
Statistics   |
           N |          50           50
        r2_a |    .5585238     .97147575
-----------------------------------------
                           legend: b/t/p
```




```
. reg y1 x c.x#c.x

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =  835.42
       Model |  91.7752102     2  45.8876051           Prob > F      =  0.0000
    Residual |  2.58160158    47  .054927693           R-squared     =  0.9726
-------------+------------------------------           Adj R-squared =  0.9715
       Total |  94.3568118    49  1.92564922           Root MSE      =  .23437

------------------------------------------------------------------------------
          y1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.026353   .0323145    31.76   0.000     .9613451    1.091362
             |
     c.x#c.x |   .5328659   .0201996    26.38   0.000     .4922296    .5735022
             |
       _cons |  -.0255622   .0396437    -0.64   0.522    -.1053151    .0541907
------------------------------------------------------------------------------
```

The first model produced a fairly large R^2 value. However, when she actually plotted the observed values, she saw that there seemed to be a U-shaped curvilinear relationship between y1 and x (or actually, kind of a j-shaped relationship in this case). Further, the curvefit graph suggested that a quadratic model would fit the data very well. Quadratic models can be good when you believe that the effect of X on Y will change sign at some point. In this case, increases in X initially produce decreases in y, but then subsequent increases in X produce increases in Y. By running a new model with an X^2 term, she got a near perfect fit to the data.

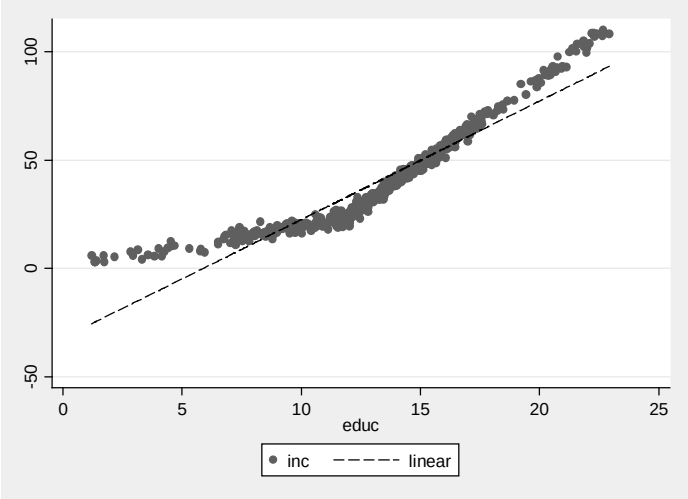
If she had simply estimated the linear model, she would have missed the curvilinear relationship. She would have thought that increases in X always produce increases in Y. She would have initially underestimated the predicted values of Y, then overestimated them, and then gone back to underestimating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

If she wanted to be a little more thorough, she could also have tested higher order polynomials, e.g. X^3 and X^4. Given that the model fit very well as it was, and perhaps

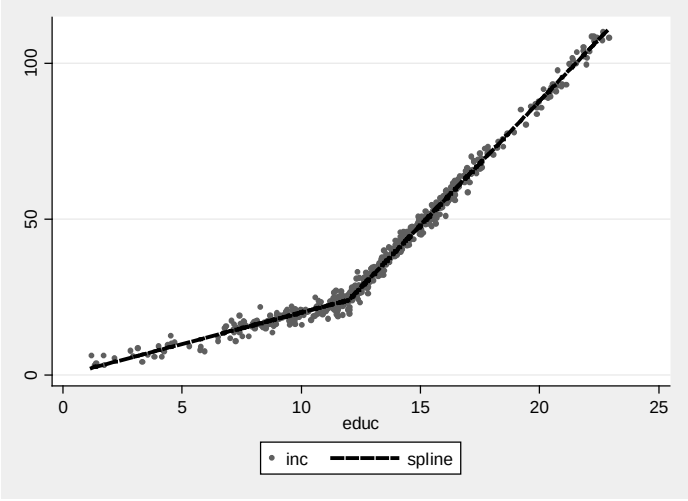
because her theory justified a curvilinear relationship, she apparently didn't feel the need to do that.

2.

```
. reg inc educ

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) = 4021.97
       Model |  253767.956     1  253767.956           Prob > F      =  0.0000
    Residual |  31421.5231   498  63.0954279           R-squared     =  0.8898
-------------+------------------------------           Adj R-squared =  0.8896
       Total |  285189.479   499  571.522002           Root MSE      =  7.9433

------------------------------------------------------------------------------
         inc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   5.472857   .0862968    63.42   0.000     5.303306    5.642407
       _cons |  -32.12951   1.194932   -26.89   0.000    -34.47724   -29.78178
------------------------------------------------------------------------------

. predict linear
(option xb assumed; fitted values)

. label variable linear "linear"
. scatter inc educ || line linear educ, scheme(sj) sort
```



```
. mkspline edlow 12 edhi = educ
. reg inc edlow edhi

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  2,   497) =38038.93
       Model |  283338.491     2  141669.245           Prob > F      =  0.0000
    Residual |  1850.98836   497  3.72432266           R-squared     =  0.9935
-------------+------------------------------           Adj R-squared =  0.9935
       Total |  285189.479   499  571.522002           Root MSE      =  1.9299

------------------------------------------------------------------------------
         inc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       edlow |   2.015082    .044107    45.69   0.000     1.928423    2.101742
        edhi |   7.976738   .0350599   227.52   0.000     7.907854    8.045622
       _cons |  -.0942312   .4621001    -0.20   0.838    -1.002142    .8136793
------------------------------------------------------------------------------

. test edlow = edhi

 ( 1)  edlow - edhi = 0

       F(  1,   497) = 7939.84
            Prob > F =    0.0000

. predict spline
(option xb assumed; fitted values)

. label variable spline "spline"
. scatter inc educ || line spline educ, scheme(sj) sort
```



The linear model showed a strong relationship between income and education, and it might have been tempting to stop there. However, when she actually plotted the observed vs predicted values, she saw that the effect of education appeared to be much smaller for grades 0-12 than it was for grades 13 and higher. This probably made substantive sense to her as well, as it implies that each year of college education produces more gains than each year of elementary school education.

She therefore decided to run a piecewise regression model. Such a model allows the effect of X to differ across its range. Specifically, she allowed grades 1-12 to have one

effect, while grades 13+ had a different effect. Because of the way she ran the mkspline command, the two coefficients showed the effects for each of the two levels of education rather than the difference in their effects, e.g. the results showed that grades 1-12 each had an effect of two while grades 13+ had an effect of 8. Just to make sure that the estimated effects really were different, she ran a test command that showed the differences in the estimated effects were highly significant.

Her final regression showed a near perfect relationship between the Xs and Y, and the graph also showed a strong correspondence between the predicted values and the observed values. (You can tell these data are fake, can't you?) Given that the model also made good theoretical sense she didn't feel the need to do anything else.

Appendix: Stata Code used in this exam

```
* Exam 2, Soc 63993, April 5, 2013
version 12.1

* Part II - Path analysis
clear all
matrix input corr = (1,.6,.3,.27\.6,1,.5,.45\.3,.5,1,-.3\.27,.45,-.3,1)
corr2data x1 x2 x3 x4, corr(corr) n(100)
corr
* confirm with pathreg. Must be installed
pathreg (x2 x1) (x3 x1 x2) (x4 x1 x2 x3)
* Confirm with sem
sem (x1 -> x2) (x1 x2 -> x3) (x1 x2 x3 -> x4)
* teffects will give us the direct, indirect and total effects.
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 7000, count
gen older = female
replace weight = weight + (30 * older)
center weight, gen(relig)
label variable relig "Traditional religious values"
gen gaymarr = (health-1) * 25 - .5*relig - 15*older
label variable gaymarr "Support for Gay Marriage"
gen oldrelig = older * relig
* Do analyses
ttest gaymarr, by(older)
nestreg: reg gaymarr relig older oldrelig
ttest relig, by(older)
* Additional analysis. This will plot the relationships
* and show differences in effects between the young and the old.
quietly reg gaymarr relig i.older i.older#c.relig
quietly margins older, at(relig = (-57(5)102))
marginsplot, noci scheme(sj) ylabel(#20) xlabel(#20) ytitle("Support for Gay Marriage") xline(0)
quietly margins r.older, at(relig = (-57(5)102))
marginsplot, noci scheme(sj) ylabel(#20) xlabel(#20) ytitle("Support for Gay Marriage") xline(0)
```

```
* Part IV-1: Quadratic Model
* Manipulate data. By construction the relationship is strongly quadratic.
clear all
set seed 123456
set obs 50
gen x = rnormal()
gen y1 = x + .5*x^2 + rnormal(.0, .2)
* Do analysis
reg y1 x
curvefit y1 x, f(1 4 )
reg y1 x c.x#c.x

* Part IV-2: Piecewise regression
* Manipulate data. By construction, the effect of education is very different
* for lower grades than it is for higher.
clear all
use "http://www3.nd.edu/~rwilliam/stats2/statafiles/blwh.dta", clear
set seed 123456
replace educ = educ + rnormal()
gen inc = 2 * educ if educ <=12
replace inc = 8 * educ - 72 if educ > 12
replace inc = inc + rnormal(0, 2)
* Do analysis
reg inc educ
predict linear
label variable linear "linear"
scatter inc educ || line linear educ, scheme(sj) sort
mkspline edlow 12 edhi = educ
reg inc edlow edhi
predict spline
label variable spline "spline"
scatter inc educ || line spline educ, scheme(sj) sort
```

Note: Don't just run all the code at once, as graphs will get overwritten as new graphs are generated.