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

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

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

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

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

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

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

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

False. The estimates will continue to be biased.

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

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

. mkspline age1 50 age2=age

. reg warm age1 age2
```

```
Source |       SS       df       MS              Number of obs =    2293
-------------+--------------------------------------------------
Model |  85.5158632     2  42.7579316           Prob > F      =  0.0000
Residual |  1889.23512  2290  .824993501           R-squared     =  0.0433
-------------+--------------------------------------------------
     Total |  1974.75098  2292  .861584198           Root MSE      =  .90829
-------------+--------------------------------------------------
          |      Coef.   Std. Err.     t    P>|t|     [95% Conf. Interval]
-------------+--------------------------------------------------
age1 |  -.0106279   .0022345  -4.76   0.000    -.0150098    -.006246
age2 |  -.0126036   .0026858  -4.69   0.000   -.0178705   -.0073368
_cons |   3.094973   .0844513   36.65   0.000    2.929364    3.260582
-------------+--------------------------------------------------
          |    test age1 = age2
```

```
( 1)  age1 - age2 = 0

F(  1,  2290) =  0.21
Prob > F =  0.6507
```
True. The marginal option was not used, so the effects of age1 and age2 correspond to the effects of age for each age group. The test command shows that, counter to the researcher’s hypothesis, the effects of age are not greater for older people.

II. Path Analysis/Model specification (25 pts). A sociologist believes that the following model describes the relationship between X1, X2, X3, and X4. All her variables are in standardized form. The estimated value of each path in her model is included in the diagram.

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

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

<table>
<thead>
<tr>
<th></th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x2</td>
<td>0.3000</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x3</td>
<td></td>
<td></td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>x4</td>
<td></td>
<td></td>
<td></td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Complete matrix:

<table>
<thead>
<tr>
<th></th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x2</td>
<td>0.3000</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x3</td>
<td>-0.1500</td>
<td>-0.5000</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>x4</td>
<td>-0.0900</td>
<td>-0.3000</td>
<td>-0.3000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
c. (5 pts) Decompose the correlation between X2 and X4 into

- Correlation due to direct effects
  - .6
- Correlation due to indirect effects
  - .3
- Correlation due to common causes
  - 0

d. (5 pts) Suppose the above model is correct, but instead the researcher believed in and estimated the following model:

\[
X_3 \rightarrow X_4 \leftarrow w
\]

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 misspecification, 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, -0.3. This is only half as large as the effect found in the correct model of -0.6. Thus, the researcher would greatly underestimate the impact of X3 on X4. The smaller effect would also increase the likelihood that the researcher would conclude that the effect did not significantly differ from 0. This mistake would occur because of omitted variable bias; the correlation between X3 and X4 that is due to the common cause of X2 would instead be attributed to the direct effect of X3 on X4.

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

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

Almost 2300 likely voters are surveyed. The results of the analysis are as follows:
## *Estimate Models*

```
.nestreg: reg clinton hlthcare female femed
```

### Block 1: `hlthcare`

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 2293</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>319218.141</td>
<td>1</td>
<td>319218.141</td>
<td>F( 1, 2291) = 547.23</td>
</tr>
<tr>
<td>Residual</td>
<td>1336427.49</td>
<td>2291</td>
<td>583.33806</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>1655645.64</td>
<td>2292</td>
<td>722.35848</td>
<td>R-squared = 0.1925</td>
</tr>
</tbody>
</table>

| hlthcare | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|----------|-------|-----------|---|-----|-----------------------|
| hlthcare | 2.971744 | 0.1270363 | 23.39 | 0.000 | 2.722626 3.220862 |

| _cons | 7.295779 | 1.868545 | 3.90 | 0.000 | 3.631561 10.96 |

### Block 2: `female`

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 2293</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>888439.231</td>
<td>2</td>
<td>444219.616</td>
<td>F( 2, 2290) = 1325.93</td>
</tr>
<tr>
<td>Residual</td>
<td>767206.404</td>
<td>2290</td>
<td>335.024631</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>1655645.64</td>
<td>2292</td>
<td>722.35848</td>
<td>R-squared = 0.5366</td>
</tr>
</tbody>
</table>

| clinton | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|---------|-------|-----------|---|-----|-----------------------|
| hlthcare | 1.090055 | 0.1065482 | 10.23 | 0.000 | 0.8811136 1.298996 |
| female | 34.96084 | 0.8481638 | 41.22 | 0.000 | 33.29759 36.62409 |

| _cons | 15.2379 | 1.429109 | 10.66 | 0.000 | 12.43542 18.04038 |

### Block 3: `femed`

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 2293</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>889272.558</td>
<td>3</td>
<td>296424.186</td>
<td>F( 3, 2289) = 885.36</td>
</tr>
<tr>
<td>Residual</td>
<td>766373.078</td>
<td>2289</td>
<td>334.806393</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>1655645.64</td>
<td>2292</td>
<td>722.35848</td>
<td>R-squared = 0.5371</td>
</tr>
</tbody>
</table>

| clinton | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|---------|-------|-----------|---|-----|-----------------------|
| hlthcare | 0.903851 | 0.158982 | 5.69 | 0.000 | 0.5920873 1.215615 |
| female | 30.27528 | 0.888626 | 9.80 | 0.000 | 24.21848 36.33208 |
| femed | 3378538 | 0.2141501 | 1.58 | 0.115 | -0.0820948 0.7578023 |

| _cons | 17.53522 | 2.039965 | 8.60 | 0.000 | 13.53485 21.5356 |

<table>
<thead>
<tr>
<th>Block</th>
<th>Residual</th>
<th>Change</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>df</td>
<td>df</td>
<td>Pr &gt; F</td>
</tr>
<tr>
<td>1</td>
<td>547.23</td>
<td>1</td>
<td>2291</td>
</tr>
<tr>
<td>2</td>
<td>1699.04</td>
<td>1</td>
<td>2290</td>
</tr>
<tr>
<td>3</td>
<td>2.49</td>
<td>1</td>
<td>2289</td>
</tr>
</tbody>
</table>

---

* Sociology 63993—Exam 2 Answer Key—Page 4
. * Differences by gender
. ttest clinton, by(female)

Two-sample t test with equal variances

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>1066</td>
<td>28.68668</td>
<td>.5322045</td>
<td>17.37629</td>
<td>27.64239  29.73097</td>
</tr>
<tr>
<td>Female</td>
<td>1227</td>
<td>67.36528</td>
<td>.565294</td>
<td>19.80143</td>
<td>66.25623  68.47433</td>
</tr>
<tr>
<td>combined</td>
<td>2293</td>
<td>49.38386</td>
<td>.5612733</td>
<td>26.87673</td>
<td>48.28321  50.48452</td>
</tr>
</tbody>
</table>

diff = mean(Male) - mean(Female)
Ho: diff = 0
degrees of freedom = 2291

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

. ttest hlthcare, by(female)

Two-sample t test with equal variances

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>1066</td>
<td>12.33771</td>
<td>.1080179</td>
<td>3.526749</td>
<td>12.12576  12.54966</td>
</tr>
<tr>
<td>Female</td>
<td>1227</td>
<td>15.74833</td>
<td>.1039811</td>
<td>3.642307</td>
<td>15.54433  15.95233</td>
</tr>
<tr>
<td>combined</td>
<td>2293</td>
<td>14.16276</td>
<td>.0829322</td>
<td>3.971231</td>
<td>14.00013  14.32539</td>
</tr>
</tbody>
</table>

diff = mean(Male) - mean(Female)
Ho: diff = 0
degrees of freedom = 2291

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

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

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

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

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

Hillary is much more popular with women than she is with men. Women are also more concerned about health care. Because hlthcare positively affects attitudes toward Clinton, women’s greater concern for health care (a compositional difference) adds to Hillary's greater popularity among women.
In short, gender is important for two reasons. First, the intercept is greater for women than it is for men. Second, women tend to be more concerned about health care, which in turn causes them to like Hillary more.

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

1. 

```
. reg y x
Source |       SS       df       MS              Number of obs =    2293
-------------+------------------------------------------------------------------
Model |  94109363.8     1  94109363.8            F(  1,  2291) =  223.47
Residual |  964794055  2291  421123.551          Prob > F      =  0.0000
-------------+------------------------------------------------------------------
Total |  1.0589e+09  2292  461999.746          Root MSE      =  648.94
-------------+---------------------------------------------------------------
        y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+---------------------------------------------------------------
x |   12.07653   .8078496    14.95   0.000     10.49234    13.66072
_cons |   422.1198   13.55198    31.15   0.000     395.5444    448.6952
-------------+---------------------------------------------------------------
```

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

```
. gen x2 = x^2
. reg y x x2
Source |       SS       df       MS              Number of obs =    2293
-------------+------------------------------------------------------------------
Model |   487727633     2   243863816            F(  2,  2290) =  977.72
Residual |   571175786  2290  249421.741          Prob > F      =  0.0000
-------------+------------------------------------------------------------------
Total |  1.0589e+09  2292  461999.746          Root MSE      =  499.42
-------------+---------------------------------------------------------------
        y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+---------------------------------------------------------------
x |   2.012126    .671355     3.00   0.003     .6955989    3.328654
  x2 |  1.443811   .0363446    39.73   0.000     1.37254     1.515083
_cons |   15.8122     14.60768     1.08   0.279     -12.83346    44.45787
-------------+---------------------------------------------------------------
```

```
. estat ovtest
Ramsey RESET test using powers of the fitted values of y
Ho:  model has no omitted variables
   F(3, 2287) =      0.08
             Prob > F =  0.9714
```
The researcher suspects that \( x \) may have a curvilinear relationship with \( y \). The `estat ovtest` command confirms that adding one or more higher powers of \( x \) (\( x^2, x^3, x^4 \)) would significantly improve the fit of the model. She therefore generates \( x^2 \) and adds it to the model. The effect of \( x^2 \) is highly significant, and the subsequent `estat ovtest` command shows that there is now no need to add any more higher powers. If she stuck with the original model, she would overestimate \( y \) in some parts of the \( x \) range and underestimate it in others. Further, she would miss the curvilinear relationship, and erroneously conclude that the effect of \( x \) is always positive when in fact it switches to being negative after the bend.

2.

```
reg y x
```

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

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

```
scatter y x
```

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

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 2293</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6346.42315</td>
<td>1</td>
<td>6346.42315</td>
<td>F( 1, 2291) =70636.43</td>
</tr>
<tr>
<td>Residual</td>
<td>205.8379</td>
<td>2291</td>
<td>0.089846312</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>6552.26105</td>
<td>2292</td>
<td>2.85875264</td>
<td>R-squared = 0.9686</td>
</tr>
</tbody>
</table>

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

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

Problem II:

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

Problem III:

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

* Results for exam start below
* Estimate Models
nestreg: reg clinton hlthcare female femed
* Differences by gender
ttest clinton, by(female)
ttest hlthcare, by(female)
```

Problem IV-a:

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

Problem IV-b:

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