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

1. A researcher has written her own computer program to compute regression estimates. She gets $F = 17$, $R^2 = .25$, Adjusted $R^2 = .27$. As far as we can tell, her program is working correctly.

False. There is an upward bias in $R^2$ that Adjusted $R^2$ corrects for so Adjusted $R^2$ should be smaller.

2. Cook’s distance is used to test for serial correlation.

False. Cook’s distance is used to measure the influence of outliers. Use the Durbin-Watson statistic for serial correlation.

3. One of the rare times when pairwise deletion of missing data is desirable is when skip patterns have caused data for some cases to be missing.

False. If anything, this could be one of the worst times to use pairwise deletion. Pairwise deletion might make sense when data are missing on a totally random basis, e.g. only a random subsample of the total sample was asked some questions. But with skip patterns, the people who aren’t asked questions may be qualitatively different from those who are, e.g. a question might only be asked of women or married people. Further, the question might make no sense for those not asked it, e.g. asking a man how many times have you been pregnant?

4. Random measurement error results in biased estimates of means, correlations and covariances.

False. Correlations are attenuated but means and covariances remain unbiased.

5. Robust regression routines work best when it is the DVs that have outliers rather than the IVs.

True. This is straight from the notes on outliers.

II. **Short answer.** Discuss all three of the following problems. (15 points each, 45 points total.) In each case, the researcher has used Stata to test for a possible problem, concluded that there is a problem, and then adopted a strategy to address that problem. Explain (a) what problem the researcher was testing for, and why she concluded that there was a problem, (b) the rationale behind the solution she chose, i.e. how does it try to address the problem, and (c) one alternative solution she could have tried, and why. (NOTE: a few sentences on each point will probably suffice – you don’t have to repeat everything that was in the lecture notes.)

II-1.

```
  . reg  warmlt2 yr89 male white age ed prst

Source |       SS       df       MS              Number of obs =    2293
-------------+--------------------------------------------------
Model | 14.1569236     6  2.35948727           Prob > F      =  0.0000
Residual | 244.374258  2286  .106900375           R-squared     =  0.0548
          |-----------------------------------------------
          |-----------------------------------------------
          |-----------------------------------------------
          |-----------------------------------------------
          |-----------------------------------------------
Total | 258.531182  2292  .11279722           Root MSE      =  .32696
```
The researcher used the Breusch-Pagan test to test for heteroskedasticity. Because the test statistic was significant, she decided to use robust standard errors, which relax the assumption that errors are independent and identically distributed. She might have also used weighted least squares. As we'll see later on though, either of these approaches is wrong in this case. As the tab1 command shows, her dependent variable is a dichotomy. In such cases, you should quit trying to “fix” OLS and switch to a technique like logistic regression instead.
II-2.

. reg y x1 x2 x3 x4

<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>81.42737</td>
<td>4</td>
<td>20.3568442</td>
<td>F( 4, 2288) = 24.60</td>
</tr>
<tr>
<td>Residual</td>
<td>1893.3236</td>
<td>2288</td>
<td>.827501575</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>1974.75098</td>
<td>2292</td>
<td>.861584198</td>
<td>Adj R-squared = 0.0396</td>
</tr>
</tbody>
</table>

| y | Coef. | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|---|-------|-----------|------|------|-------------------|
| x1 | .0001393 | .003306 | 0.04 | 0.966 | -.0063436 -.0066223 |
| x2 | -.0043145 | .0033019 | -1.31 | 0.191 | -.0107895 .0021605 |
| x3 | -.0025131 | .0032995 | -0.76 | 0.446 | -.0089835 .0039573 |
| x4 | -.0044104 | .0033055 | -1.33 | 0.182 | -.0108925 .0020716 |
| _cons | 3.106225 | .0539982 | 57.52 | 0.000 | 3.000334 3.212116 |

. test x1 = x2 = x3 = x4

( 1) x1 - x2 = 0
( 2) x1 - x3 = 0
( 3) x1 - x4 = 0

F( 3, 2288) = 0.31
Prob > F = 0.8152

. gen x1234 = x1 + x2 + x3 + x4

. reg y x1234

<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>80.647724</td>
<td>1</td>
<td>80.647724</td>
<td>F( 1, 2291) = 97.55</td>
</tr>
<tr>
<td>Residual</td>
<td>1894.10326</td>
<td>2291</td>
<td>.826758296</td>
<td>R-squared = 0.0408</td>
</tr>
<tr>
<td>Total</td>
<td>1974.75098</td>
<td>2292</td>
<td>.861584198</td>
<td>Root MSE = .90926</td>
</tr>
</tbody>
</table>

| y | Coef. | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|---|-------|-----------|------|------|-------------------|
| x1234 | -.0027758 | .0002811 | -9.88 | 0.000 | -.003327 -.0022247 |
| _cons | 3.106433 | .0539674 | 57.56 | 0.000 | 3.000602 3.212263 |

The researcher saw that multicollinearity appeared to be a problem in her data. The global F statistic was significant but none of the individual T values were. The test command showed her that the coefficients for the four X's did not significantly differ from each other. She therefore just added the four items together and used the resulting scale in the regression. Since there is only one variable in the regression, there is no multicollinearity problem. This would especially make sense if the items are measured the same way (e.g. 5 point scales) and are thought to tap the same concept. Alternatively she might have considered dropping one or more items if she felt they were not important to the model, or she could have created a scale using some other means. Or, she could have been content just using the global F test and saying that one or more effects differed from zero.
II-3.

```
. reg price mpg weight length foreign

Source |       SS       df       MS              Number of obs =     875
-------------+----------------------------------------------------------------
Model |  1.0147e+09     4   253674918           Prob > F      =  0.0000
Residual |  1.2653e+09   870     1454327           R-squared     =  0.4451
-------------+----------------------------------------------------------------
Total |  2.2800e+09   874  2608654.65           Root MSE      =    1206
-------------+----------------------------------------------------------------

price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
mpg |  -38.37705   10.34107   -3.71   0.000    -58.67342    -18.08068
weight |  -16.10697   .9834499   -1.63   0.104    -3.976626    1.944903
length |   61.42098   7.731232    7.94   0.000     46.24694    76.59503
foreign |   1893.053   89.09917   21.25   0.000     1718.179    2067.928
_cons |  -4470.567   943.7682   -4.74   0.000    -6322.895    -2618.238
-------------+----------------------------------------------------------------

. sum

Variable |       Obs        Mean    Std. Dev.       Min        Max
--------+----------------------------------------------------------
price |    1850    6165.257    2930.291       3291      15906
mpg |    1850     21.2973    5.747833         12         41
weight |     875    2312.571     342.109     1760       2930
length |    1850    187.9324    22.12136        142        233
foreign |    1850    .2972973    .4571921          0          1

. impute weight mpg length foreign, gen(xweight)
52.70% (975) observations imputed

. reg price mpg xweight length foreign

Source |       SS       df       MS              Number of obs =    1850
-------------+----------------------------------------------------------------
Model |  5.4367e+09     4  1.3592e+09           Prob > F =  0.0000
Residual |  1.0440e+10  1845  5658506.19           R-squared =  0.3410
-------------+----------------------------------------------------------------
Total |  1.5877e+10  1849  8586606.22           Root MSE =  2378.8
-------------+----------------------------------------------------------------

price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
mpg |  -143.850   17.54604   -8.20   0.000   -178.2628   -109.4384
xweight |  -391066   .5884892   -6.66   0.506   -1.545241    .7631088
length |    68.06994   13.55269    5.02   0.000     41.48971    94.65017
foreign |   2611.786   156.4679    16.69   0.000     2304.913    2918.658
_cons |  -3244.343   1307.184   -2.48   0.013   -5808.064    680.6273
-------------+----------------------------------------------------------------
```

The researcher noticed that she only had 875 cases in her first regression, even though there are 1850 cases in her data set. The summarize command showed her that all of the missing data was in one variable, weight. She therefore used the impute command to substitute regression estimates for the missing values. The idea is that this is her “best guess” of what the missing values really equal. This practice has various problems; if nothing else, the significance tests are misleading because the imputed values are treated the same as the real values, rather than as estimates that are themselves subject to uncertainty. Further, the cases that are missing may be qualitatively different from the ones that aren’t, e.g. maybe weight was not measured for
foreign automobiles. As an alternative, she might have simply used listwise deletion; or she could have used a more advanced technique like multiple imputation whose standard errors and significance tests would have been more correct. Also, unless it is vitally important to the theory behind the model, I would seriously consider just dropping the weight variable since it is not significant either before or after imputation. I would especially consider dropping it if it is problems in the data collection process that caused so much data to be missing; it may just be that it isn’t well-enough measured to be useful.

III. Computation and interpretation. (35 points total)

A graduate student wants to do her dissertation on the determinants of women’s socio-economic status (SES). To see whether the idea is worth pursuing, she is analyzing a few key variables that were collected as part of a nationwide study of 488 women. Her measures include the following:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ses</td>
<td>Socio-Economic Status scale. Ranges from a low of 0 to a high of 100.</td>
</tr>
<tr>
<td>nev_mar</td>
<td>Coded 1 if the woman has never been married, 0 otherwise</td>
</tr>
<tr>
<td>rural</td>
<td>Coded 1 if the respondent lives in a rural area, 0 otherwise</td>
</tr>
<tr>
<td>school</td>
<td>Number of years of schooling respondent has completed</td>
</tr>
<tr>
<td>tenure</td>
<td>Number of years respondent has worked in her current job</td>
</tr>
</tbody>
</table>

An analysis of the data yields the following results. [NOTE: You’ll need some parts of the following to answer the questions, but other parts are extraneous. You’ll have to figure out which is which.]

```
. reg ses nev_mar rural school tenure
Source |        SS       df       MS                  Number of obs = 488
---------|-------------|-----------|-------------|------------------|-----------|
Model | 29626.8441    4     7406.7104          Prob > F =  [ 1 ]
Residual | 47422.5089  483  98.1832482               R-squared =  [ 2 ]
---------|-------------|-----------|-------------|------------------|-----------|
Total | 77049.353   487 158.212224            Root MSE = 9.9087
---------|-------------|-----------|-------------|------------------|-----------|
ses | Coef. Std. Err.     t   P>|t|     [95% Conf. Interval] |
---------|-------------|-----------|-------------|------------------|-----------|
nev_mar | -0.1388159  1.001324 -0.14 0.890 -2.106304  1.828673 |
rural | -4.743383   1.025829 -4.66 0.000 -6.759023 -2.727744 |
school | 1.943179    .1719365 11.30 0.000 1.605343  2.281015 |
tenure | [ 4 ] 1.232743  8.16 0.000 .7639161  1.248356 |
_cons | 17.19019    2.273869  7.56 0.000 12.72229  21.65808 |
---------|-------------|-----------|-------------|------------------|-----------|
. pcorr2 ses nev_mar rural school tenure
(obs=488)
Partial and Semipartial correlations of ses with
Variable | Partial | SemiP | Partial^2 | SemiP^2 | Sig. |
---------|---------|-------|-----------|---------|-----|
nev_mar | -0.0063 | -0.0049 | 0.00000 | 0.00000 | 0.890 |
rural | -0.2059 | -0.1651 | 0.04244 | 0.02720 | 0.000 |
school | 0.4573 | 0.4034 | 0.20911 | 0.16280 | 0.000 |
tenure | 0.3481 | 0.2914 | 0.12122 | 0.08489 | 0.000 |
Variable |       Obs  | Mean   | Std. Dev. | Min   | Max
---------|-----------|--------|-----------|-------|------
        |           |        |           |       |      |
ses      |       488  | 43.32709| 12.57824  | 2.465307 | 84.2362 |
nev_mar  |       488  | .2868852| .4527717  | 0      | 1    |
rural    |       488  | .272541 | .4457236  | 0      | 1    |
school   |       488  | 12.71107| 2.70533   | 0      | 18   |
tenure   |       488  | 2.752732| 3.776793  | 0      | 21.75|

.test nev_mar rural school tenure
( 1) nev_mar = 0
( 2) rural = 0
( 3) school = 0
( 4) tenure = 0
F( 4, 483) = 75.44
Prob > F = 0.0000

collin nev_mar rural school tenure

Collinearity Diagnostics

<table>
<thead>
<tr>
<th>Variable</th>
<th>VIF</th>
<th>VIF</th>
<th>Tolerance</th>
<th>Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>nev_mar</td>
<td>1.02</td>
<td>1.01</td>
<td>0.9808</td>
<td>0.0192</td>
</tr>
<tr>
<td>rural</td>
<td>1.04</td>
<td>1.02</td>
<td>0.9643</td>
<td>0.0357</td>
</tr>
<tr>
<td>school</td>
<td>1.07</td>
<td>1.04</td>
<td>[ 5 ]</td>
<td>0.0682</td>
</tr>
<tr>
<td>tenure</td>
<td>1.08</td>
<td>1.04</td>
<td>0.9301</td>
<td>0.0699</td>
</tr>
</tbody>
</table>

Mean VIF 1.05

.estat imtest, white

White's test for Ho: homoskedasticity
against Ha: unrestricted heteroskedasticity
chi2(12) = 6.91
Prob > chi2 = 0.8637

Cameron & Trivedi's decomposition of IM-test

<table>
<thead>
<tr>
<th>Source</th>
<th>chi2</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heteroskedasticity</td>
<td>6.91</td>
<td>12</td>
<td>0.8637</td>
</tr>
<tr>
<td>Skewness</td>
<td>1.50</td>
<td>4</td>
<td>0.8272</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>6.72</td>
<td>1</td>
<td>0.0096</td>
</tr>
<tr>
<td>Total</td>
<td>15.12</td>
<td>17</td>
<td>0.5868</td>
</tr>
</tbody>
</table>

.test school = tenure
( 1) school - tenure = 0
F( 1, 483) = 16.28
Prob > F = 0.0001
a) (10 pts) Fill in the missing quantities [1] – [5].

First off, here is the uncensored printout:

```
. reg ses nev_mar rural school tenure
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 488</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>29626.8441</td>
<td>4</td>
<td>7406.71104</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>47422.5089</td>
<td>483</td>
<td>98.1832482</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>77049.353</td>
<td>487</td>
<td>158.212224</td>
<td></td>
</tr>
</tbody>
</table>

F( 4, 483) = 75.44

Model | Prob > F = 0.0000
Residual | R-squared = 0.3845
Total | Adj R-squared = 0.3794

```

| ses | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|-----|-------|-----------|---|-----|-----------------------|
| nev_mar | -.1388159 | 1.001324 | -0.14 | 0.890 | -2.106304 - 1.828673 |
| rural | -4.743383 | 1.025829 | -4.62 | 0.000 | -6.759023 - 2.727744 |
| school | 1.943179 | .1719365 | 11.30 | 0.000 | 1.605343 2.281015 |
| tenure | 1.006136 | .1719365 | 11.30 | 0.000 | 1.605343 2.281015 |
| _cons | 17.19019 | 2.273869 | 7.56 | 0.000 | 12.72229 21.65808 |

```

. collin nev_mar rural school tenure

Collinearity Diagnostics

<table>
<thead>
<tr>
<th>Variable</th>
<th>VIF</th>
<th>SQRT VIF</th>
<th>Tolerance</th>
<th>R-Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>nev_mar</td>
<td>1.02</td>
<td>1.01</td>
<td>0.9808</td>
<td>0.0192</td>
</tr>
<tr>
<td>rural</td>
<td>1.04</td>
<td>1.02</td>
<td>0.9643</td>
<td>0.0357</td>
</tr>
<tr>
<td>school</td>
<td>1.07</td>
<td>1.04</td>
<td>0.9318</td>
<td>0.0682</td>
</tr>
<tr>
<td>tenure</td>
<td>1.08</td>
<td>1.04</td>
<td>0.9301</td>
<td>0.0699</td>
</tr>
</tbody>
</table>

Mean VIF 1.05

To confirm that Stata got it right:

[1] = P value for global F = 0.0000. You can tell because the command “test nev_mar rural school tenure” tests the same hypothesis that the global F does.


[3] = t_{rural} = b_{rural} / s_{rural} = -4.743383 / 1.025829 = -4.62

[4] = b_{tenure} = s_{tenure} * t_{tenure} = .1232743 * 8.16 = 1.0059. Or, to be more precise, compute the midpoint of the confidence interval: (.7639161 + 1.248356) / 2 = 1.00613605.

[5] = tol_{school} = 1/vif_{school} = 1/1.07 = .9346. Or, if you want to be really precise, tol_{school} = 1 - R^2_{xGk} = 1 - .0682 = .9318.
b) (25 points) Answer the following questions about the analysis and the results, explaining how the printout supports your conclusions.

1. Summarize the key results. What percentage of the women have never been married? How many live in rural areas? What types of women have the highest SES scores, and which types of women have the lowest?

The means from the summarize command show us that 28.69% of the women have never been married and 27.25% live in rural areas. The regression coefficients show us that women with the highest SES levels live in non-rural areas and have more years of schooling and longer tenure in their current job. Conversely, the women with the lowest levels of SES live in rural areas and have fewer years of schooling and job tenure. It may also help your SES to have been married (or hurt to have never been married) but the effect is small and statistically insignificant.

2. The researcher was worried that missing data, heteroskedasticity, and/or multicollinearity might be problematic. Based on the results, are they?

All 488 cases are showing up in all parts of the analysis, so there is no missing data. White’s test shows no heteroskedasticity of any sort. The collin command shows very high tolerances so multicollinearity does not appear to be a problem either. If only all dissertations could be so trouble-free…

3. The researcher had hypothesized that years in current job (tenure) would have a significantly larger effect on ses than would years in school (school). Do the results support her hypothesis?

The “test school = tenure” command does show that the effects of schooling and tenure significantly differ. But, the regression coefficients show that the difference is in the opposite direction of what she hypothesized: the estimated effect of years of schooling is almost double the estimated effect of tenure. Therefore her hypothesis is not supported. (Hopefully this wasn’t the most critical element of her theory.)

4. The researcher debated whether or not to include the variable rural in her model. If she had not included it, how would the R² have been affected?

As the squared semipartial for rural shows (see the pcorr2 command output), R² would drop by .0272 if rural was dropped, i.e. R² would go from .3845 to .3573. To confirm,
5. The researcher’s daughter has just graduated from high school. She wants to spend the next four years living on a farm taking a richly deserved vacation from school and work. According to the researcher’s model, if her daughter instead spends those years going to college at UCLA in Los Angeles, what will be the expected impact on her socio-economic status?

Four additional years of schooling would be expected to increase her SES score by 4 * 1.943179 = 7.772716. In addition, living on a farm (i.e. in a rural area) instead of living in an urban area like Los Angeles would lower her SES by 4.743383. So, her SES score would be expected to be 12.516099 points higher if she went to school for four years in LA rather than taking the nice little break on the farm. I suspect mom may not go along with her daughter on this one.

Incidentally, we can confirm our answer in Stata by using the adjust command:

```
. adjust rural = 1 school = 12 tenure = 0 nev_mar = 1
```

---

```
Dependent variable: ses  Command: regress
Covariates set to value: rural = 1, school = 12, tenure = 0, nev_mar = 1

----------------------------------
  |        xb
---+----------------------------------
  | 35.6261
----------------------------------
Key:  xb  =  Linear Prediction
```

```
. adjust rural = 0 school = 16 tenure = 0 nev_mar = 1
```

---

```
Dependent variable: ses  Command: regress
Covariates set to value: rural = 0, school = 16, tenure = 0, nev_mar = 1

----------------------------------
  |        xb
---+----------------------------------
  | 48.1422
----------------------------------
Key:  xb  =  Linear Prediction
```

```
. display 48.1422 - 35.6261
12.5161
```
Appendix: Stata Commands for Exam 1. Here are the commands I used to generate the Stata output on the exam. Alas, I haven’t really conducted any new nationwide studies, but I have manipulated and sometimes disguised other data sets I have sitting around.

* Problem II-1
use "http://www.indiana.edu/~jslsoc/stata/spex_data/ordwarm2.dta", clear
reg warm yr89 male white age ed prst
hettest
tabl warmlt2, nolabel
reg warmlt2 yr89 male white age ed prst, robust

* Problem II-2
use "http://www.indiana.edu/~jslsoc/stata/spex_data/ordwarm2.dta", clear
corr2data e1 e2 e3 e4, seed(1234) sd(5 5 5 5)
gen x1 = age + e1
gen x2 = age + e2
gen x3 = age + e3
gen x4 = age + e4
clonedvar y = warm
reg y x1 x2 x3 x4
test x1 = x2 = x3 = x4
gen x1234 = x1 + x2 + x3 + x4
reg y x1234

* Problem II-3
webuse auto, clear
keep price mpg weight length foreign
replace weight = . if weight >= 3000
expand 25
reg price mpg weight length foreign
sum
impute weight mpg length foreign, gen(xweight)
reg price mpg xweight length foreign

* Problem III
webuse womenwage, clear
gen ses = ln(wage) * 25 - 25
drop age age2 wage wagecat r
order ses
reg ses nev_mar rural school tenure
pcorr2 ses nev_mar rural school tenure
sum
test nev_mar rural school tenure
collin nev_mar rural school tenure
estat imtest, white
test school = tenure
reg ses nev_mar school tenure
reg ses nev_mar rural school tenure
adjust rural = 1 school = 12 tenure = 0 nev_mar = 1
adjust rural = 0 school = 16 tenure = 0 nev_mar = 1
display 48.1422 - 35.6261