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

1. Adjusted $R^2$ is used to correct for heteroskedasticity.
   False. Adjusted $R^2$ is used to correct for the inherent upward bias in $R^2$ caused by the fact that, thanks to sampling variability, estimates of $R^2$ will always be positive even if the population value is 0.

2. Passive imputation should be used when one of the variables in the model is an interaction term that is the product of variables that have missing data.
   False. Or at least Allison says so. He recommends using the “just another variable” approach. You compute the interaction first and then impute its values like you do any other variable. If you don’t do that, the effect of the interaction is biased toward zero.

3. Studentized residuals, Predictive Mean Matching (PMM), and Cook’s distance are among the measures that can be used to detect outliers.
   False. Predictive Mean Matching is a technique sometimes used in multiple imputation to impute values for continuous variables that have missing data.

4. Pairwise deletion of missing data is commonly used when skip patterns have caused data for some cases to be missing.
   False. Pairwise deletion may be ok when data are missing totally at random. But data are usually not missing at random when skip patterns are used, e.g. men will not be asked how many times they have been pregnant.

5. If possible, increasing the sample size is one way to reduce problems of multicollinearity.
   True. Multicollinearity increases standard errors. Larger sample sizes reduce them.

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 y x

Source |       SS       df       MS              Number of obs =  3975
---------+------------------------------           F(  1,  3973) =    0.33
Model |  7588.90836     1  7588.90836           Prob > F      =  0.5681
Residual |    92495852  3973  23281.1105           R-squared     =  0.0001
---------+------------------------------           Adj R-squared = -0.0002
Total |    92503441  3974  23277.1618           Root MSE      =  152.58

------------------------------------------------------------------------------
        y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+------------------------------------------------------------------
        x |  -1.606451   2.813712    -0.57   0.568    -7.122906    3.910004
     _cons |   2.176363   4.323281     0.50   0.615    -6.299693    10.65242
------------------------------------------------------------------------------

. predict rstudent, rstudent
. extremes rstudent y x

+-------------------------------------------+
<table>
<thead>
<tr>
<th>obs:    rstudent           y           x</th>
</tr>
</thead>
<tbody>
<tr>
<td>3124.   -1920.873   -9610.815    2.340901</td>
</tr>
<tr>
<td>2546.   -.1171118   -14.99421   -.4293181</td>
</tr>
<tr>
<td>3950.   -.1160397   -15.22504   -.1853033</td>
</tr>
<tr>
<td>3417.   -.1099706   -14.32139   -.1716456</td>
</tr>
<tr>
<td>2120.    -.108885   -14.19914   -.1448071</td>
</tr>
</tbody>
</table>
+-------------------------------------------+

+----------------------------------------+
| 3082.   .1290129   18.32841   2.197414 |
| 3914.   .1295774   18.58599   2.09107 |
|  701.   .1302823   18.6424   2.122781 |
| 1613.   .1351598   18.73533   2.526286 |
| 2008.   .1353822   19.60719   2.006944 |
+----------------------------------------+

. drop  rstudent
. reg y x if y > -9600

Source |       SS       df       MS              Number of obs =  3974
---------+------------------------------           F(  1,  3972) =  416.97
Model |   10441.431     1   10441.431           Prob > F      =  0.0000
Residual |   99464.251  3972  25.0413522           R-squared     =  0.0950
---------+------------------------------           Adj R-squared =  0.0948
Total |  109905.682  3973  27.6631467           Root MSE      =  5.0041

------------------------------------------------------------------------------
        y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+------------------------------------------------------------------
        x |   1.884699   .0922977    20.42   0.000     1.703744    2.065655
     _cons |   .1503674   .1417923     1.06   0.289    -.127625    .4283598
------------------------------------------------------------------------------

. predict rstudent if e(sample), rstudent
(1 missing value generated)
```
The researcher was concerned about outliers. She computed the studentized residual for each case, which is a way of assessing the discrepancy between the predicted value for the case and its observed value, after adjusting for the standard errors. Depending on the sample size, studentized residuals greater than 3 or less than -3 may be cause for concern. She then ran the extremes command to identify the outliers in both the positive and negative directions.

The studentized residuals indicated that case 3124 was a huge outlier – the studentized residual was -1910 and the observed Y value for the case was -9610.815, which was well outside the range of the other observed Y values. The researcher dealt with the problem simply by deleting that case from the analysis. That might be ok if the researcher is convinced that the reported value of -9610.815 for Y is wrong and there is no way to retrieve the correct value. If I could check the data, it wouldn’t surprise me if the reported value was off by three decimal places and was supposed to be -9.61. You could also use techniques like rreg or qreg, which give outliers less influence on the regression estimates.

II-2.

```
. reg y x

Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  1,   198) =    0.50
Model |  54113.6335     1  54113.6335           Prob > F      =  0.4815
Residual |  21545981.1   198  108818.086           R-squared     =  0.0025
-------------+------------------------------           Adj R-squared = -0.0025
Total |  21600094.7   199  108543.189           Root MSE      =  329.88

------------------------------------------------------------------------------
      y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      x |   1.654202   2.345772     0.71   0.482    -2.971701    6.280105
   _cons |   202.0068   23.36188     8.65   0.000     155.9367    248.0768
------------------------------------------------------------------------------
```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity  
Ho: Constant variance  
Variables: fitted values of y  

\[
\text{chi}^2(1) = 69.81 \\
\text{Prob > chi}^2 = 0.0000
\]

. estat imtest

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>153.96</td>
<td>2</td>
<td>0.0000</td>
</tr>
<tr>
<td>Skewness</td>
<td>30.99</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>3.81</td>
<td>1</td>
<td>0.0510</td>
</tr>
<tr>
<td>Total</td>
<td>188.76</td>
<td>4</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

. scatter y x

. gen xsquare = x * x  
. reg y x xsquare

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>21580153.5</td>
<td>2</td>
<td>10790076.8</td>
<td>F( 2, 197) = .</td>
</tr>
<tr>
<td>Residual</td>
<td>19941.1507</td>
<td>197</td>
<td>101.224115</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>21600094.7</td>
<td>199</td>
<td>108543.189</td>
<td>R-squared = 0.9991</td>
</tr>
</tbody>
</table>

| y    | Coef.   | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|------|---------|-----------|---|-----|---------------------|
| x    | 6.893609 | .0724412  | 95.16 | 0.000 | 6.75075 7.036469 |
| xsquare | 1.993747 | .0043234  | 461.15 | 0.000 | 1.985221 2.002273 |
| _cons | 1.356793 | .8348716  | 1.63 | 0.106 | -.28964 3.003225 |
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of y

       chi2(1)     =     1.22
Prob > chi2  =   0.2686

. estat imtest

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>4.75</td>
<td>4</td>
<td>0.3140</td>
</tr>
<tr>
<td>Skewness</td>
<td>1.57</td>
<td>2</td>
<td>0.4553</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>1.36</td>
<td>1</td>
<td>0.2435</td>
</tr>
<tr>
<td>Total</td>
<td>7.68</td>
<td>7</td>
<td>0.3613</td>
</tr>
</tbody>
</table>

The hettest and imtests yielded highly significant results, indicating the presence of heteroskedasticity. Rather than immediately going to robust standard errors or weighted least squares, the researcher decided to take a closer look at her model specification. The scatterplot clearly indicated that the relationship between x and y was not linear, i.e. her model was mis-specified. She therefore added an x^2 term to her model. When she did that, model fit was greatly improved, and the heteroskedasticity tests no longer yielded significant results. Robust standard errors would have relaxed assumptions about the errors being iid, and wls would have allowed cases to be weighted differently, but the approach chosen clearly seems best. When hetero tests indicate problems, you should first consider possible problems with model specification, and fix them if necessary. If the model seems ok, then consider ways to correct for the hetero.

II-3.

. reg iq fatheriq motheriq

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 8708</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>554164.753</td>
<td>2</td>
<td>277082.376</td>
<td>F( 2, 8705) = 1636.98</td>
</tr>
<tr>
<td>Residual</td>
<td>1473444.77</td>
<td>8705</td>
<td>169.264189</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>2027609.52</td>
<td>8707</td>
<td>232.871198</td>
<td>R-squared = 0.2731</td>
</tr>
<tr>
<td></td>
<td>Adj R-squared = 0.2731</td>
<td>Root MSE = 13.01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|          | Coef. | Std. Err. | t     | P>|t|   | 95% Conf. Interval |
|----------|-------|-----------|-------|-------|-------------------|
| fatheriq | .654312 | .0299545  | 21.84 | 0.000 | .5955941-.71303   |
| motheriq | .6981873 | .0148528  | 47.01 | 0.000 | .6690722-.7273023 |
| _cons    | -34.6617 | 2.644898  | -13.11| 0.000 | -39.84633-.29.47708 |
The researcher observed that many cases were missing data on fatheriq. She therefore used the Cohen and Cohen dummy variable adjustment technique. Whether or not this is a good idea depends on why the data are missing. If, in effect, the respondent did not have a father (e.g. died when the respondent was young or was never involved with the family) this might be a wise choice. However, if it is simply a matter of the fatheriq being unknown, this would be a poor choice that would produce biased estimates. By examining the codebook or via other means, the researcher should try to determine why the cases are missing before choosing this approach (maybe she did this and decided Cohen and Cohen was justified, but then again maybe she just didn’t know any better.) If father’s IQ is unknown rather than non-existent, just sticking with listwise deletion (i.e. deleting all cases with any missing data, as was done in the original regression) or using multiple imputation (using methods regress or pmm) would probably be better.

III. Computation and interpretation. (35 points total) A recent unpublished study by Princeton researchers claims that Facebook will lose 80% of its users by 2015-2017. Although Facebook claims that the study is ridiculous, it is concerned about its long-term popularity and growth. It
has therefore collected data on the following variables from a random sample of 7,236 active Facebook members.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>stayfb</td>
<td>Scale that measures the likelihood the respondent will still be a Facebook user in five years. Potentially ranges from a low of 0 to a high of 100</td>
</tr>
<tr>
<td>frdties</td>
<td>Strength of the respondent’s current friendship ties on Facebook. Potentially ranges from a low of 0 to a high of 100</td>
</tr>
<tr>
<td>age</td>
<td>Age in years</td>
</tr>
<tr>
<td>female</td>
<td>1 = female, 0 = male</td>
</tr>
<tr>
<td>white</td>
<td>1 = white, 0 = not white</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 stayfb frdties age female white, beta
```

```
Source |       SS       df       MS              Number of obs = 7236
       |------------------+-----------------------------------
Model  | 70915.1394       1         70915.1394  F(  4, 7231) = [1]
Residual | 204149.201 7231  28.2324992  Prob > F = 0.0000
Total  | 275064.341 7235  38.0185682  R-squared = [3]
       |------------------+-----------------------------------
-------------+------------------------------
-------------+------------------------------
stayfb |       Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
frdties |   .2951861    .009427  0.000                 .4638028
age |  -.0458587   .0037968   -12.08   0.000                -.1278213
female |   .7808451   .1787656     4.37   0.000                 .0632679
white |   1.641666   .2021571     8.12   0.000                 .0823415
_cons |   67.02458   .3835846   174.73   0.000                        
-------------+----------------------------------------------------------------
```

```
. vif
```

```
Variable |       VIF       1/VIF
-------------+----------------------
frdties |      2.14    0.467839
female |      2.04    0.497725
age |      1.09    0.916446
white |      1.00    0.998317
-------------+----------------------
Mean VIF |      1.57
```

```
. test frdties age female white
```

```
( 1)  frdties = 0
( 2)  age = 0
( 3)  female = 0
( 4)  white = 0
F(  4, 7231) = 627.96
Prob > F =  0.0000
```

```
. `sum

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>stayfb</td>
<td>7236</td>
<td>79.21281</td>
<td>6.16592</td>
<td>37.648</td>
<td>94.028</td>
</tr>
<tr>
<td>frdties</td>
<td>7236</td>
<td>42.35774</td>
<td>9.688028</td>
<td>10</td>
<td>74.5</td>
</tr>
<tr>
<td>age</td>
<td>7236</td>
<td>47.70854</td>
<td>17.18618</td>
<td>20</td>
<td>74</td>
</tr>
<tr>
<td>female</td>
<td>7236</td>
<td>.5210061</td>
<td>.4995931</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>white</td>
<td>7236</td>
<td>.8928966</td>
<td>.309266</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>black</td>
<td>7236</td>
<td>.1071034</td>
<td>.309266</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

. test white = female

( 1)  - female + white = 0

F( 1, 7231) = 10.23
Prob > F = 0.0014

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of stayfb

chi2(1) = 0.55
Prob > chi2 = 0.4591

. estat imtest, white

White's test for Ho: homoskedasticity
against Ha: unrestricted heteroskedasticity

chi2(12) = 107.11
Prob > chi2 = 0.0000

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>107.11</td>
<td>12</td>
<td>0.0000</td>
</tr>
<tr>
<td>Skewness</td>
<td>178.07</td>
<td>4</td>
<td>0.0000</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>27.67</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>312.86</td>
<td>17</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

. pcorr2 stayfb frdties age female white

(obs=7236)

Partial and Semipartial correlations of stayfb with

<table>
<thead>
<tr>
<th>Variable</th>
<th>Partial</th>
<th>SemiP</th>
<th>Partial^2</th>
<th>SemiP^2</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>frdties</td>
<td>0.3456</td>
<td>0.3172</td>
<td>0.1194</td>
<td>0.1006</td>
<td>0.000</td>
</tr>
<tr>
<td>age</td>
<td>-0.1406</td>
<td>-0.1224</td>
<td>0.0198</td>
<td>0.0150</td>
<td>0.000</td>
</tr>
<tr>
<td>female</td>
<td>0.0513</td>
<td>0.0443</td>
<td>0.0026</td>
<td>0.0020</td>
<td>0.000</td>
</tr>
<tr>
<td>white</td>
<td>0.0951</td>
<td>0.0823</td>
<td>0.0090</td>
<td>0.0068</td>
<td>0.000</td>
</tr>
</tbody>
</table>
. alpha frdties age female white, s

Test scale = mean(standardized items)

Average interitem correlation: 0.1590
Number of items in the scale: 4
Scale reliability coefficient: 0.4307

a) (10 pts) Fill in the missing quantities [1] – [5]. (A few other values may have also been blanked out, but you don’t need to fill them in.)

Here is the uncensored printout.

. reg stayfb frdties age female white, beta

Source |       SS       df       MS              Number of obs = 7236
-------------+------------------------------           F(  4,  7231) = 627.96
Model |  70915.1394       4  17728.7848           Prob > F      =  0.0000
Residual |  204149.201  7231  28.2324992           R-squared     =  0.2578
-------------+------------------------------           Adj R-squared =  0.2574
Total |  275064.341  7235  38.0185682           Root MSE      =  5.3134

------------------------------------------------------------------------------
     stayfb |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
     frdties |   .2951861    .009427  31.31    0.000                 .4638028
     age |  -.0458587   .0037968  -12.08    0.000                -.1278213
     female |   .7808451   .1787656     4.37    0.000                 .0632679
     white |   1.641666   .2021571     8.12    0.000                 .0823415
      _cons |   67.02458   .3835846   174.73    0.000                        .
------------------------------------------------------------------------------

. vif

Variable |       VIF       1/VIF
-------------+----------------------
     frdties |      2.14    0.467839
     female |      2.04   0.489227
     age |      1.09    0.916446
     white |      1.00    0.998317
-------------+----------------------
Mean VIF |      1.57

To confirm that Stata got it right,

[1] Global F test = MSR/MSE = 17728.7848 / 28.2324992 = 627.96. Those who prefer to have more leisure time in their lives can simply note that the command test frdties age female white already gave the answer.

[2] DF\text{model} = K = Number of Independent Variables = 4. Masochistic types can also compute DF\text{model} = \text{SS}_{\text{model}} / \text{MS}_{\text{model}} = 70915.1394/ 17728.7848 = 4.


[4] t_{frdties} = b_{frdties}/se_{frdties} = .2951861/.009427 = 31.31
b)  (25 points) Answer the following questions about the analysis and the results, explaining how the printout supports your conclusions.

1. Summarize the key findings. Which groups or types of individuals are most likely to be on Facebook 5 years from now and which are least likely?

All the variables have statistically significant effects. The signs of the coefficients indicate that younger people, those with stronger friendship ties, women, and whites are more likely to be on Facebook five years from now. Older people, those with weaker friendship ties, men, and nonwhites are less likely.

2. The researchers originally planned to use multiple imputation to deal with missing data, but then it turned out there was no missing data (Facebook knows everything about you). If there had been missing data on frdties and female, what imputation method or methods would you recommend using (e.g. logit, mlogit, regress, ologit, pmm, poisson, or something else)? Briefly explain why.

Frdties is a continuous variable so regress (or possibly pmm, Predictive Mean Matching) would be appropriate. Female is a binary variable so use logit for it. Since multiple variables are missing I would also use Imputation using Chained Equations (ICE).

3. The researchers were concerned that the items may suffer from random measurement error. Would you encourage them to create a scale out of the independent variables in order to deal with the problem?

Conceptually, it is hard to make an argument that the variables all measure the same thing. Empirically, the alpha command only yielded a Cronbach’s Alpha of .4307, well below the .80 that is typically recommended.

4. The Breusch-Pagan / Cook-Weisberg test for heteroskedasticity is insignificant. Would you therefore agree that there is no need to be worried about heteroskedasticity?

There may be no need to worry about linear heteroskedasticity. But the White test for any type of heteroskedasticity is highly significant. The model might need to be modified in some way, e.g. include squared terms, or include interactions between gender or race with friendship ties. If the model is correctly specified, you might need to use robust standard errors or weighted least squares.

5. Facebook wants to know which variable has the greatest impact on whether or not somebody is likely to stay with Facebook. What would you say? Cite at least two or three results from the printout to support your argument.

Friendship ties has the biggest standardized beta, the biggest T value, and the largest squared semipartial value. The squared semipartial shows that dropping frdties from the model would reduce R^2 by .10, a far larger impact than dropping any other variable would have. Conceptually, that seems plausible. Those who have the strongest friendship ties on Facebook should be less inclined to quit Facebook and leave them behind.
version 12.1
* II-1
use http://www3.nd.edu/~rwilliam/statafiles/anomia.dta, clear
* Create/ manipulate data
corr2data e1 e2
gen x = anomia8 + anomia9 + e1*.30
gen y = anomia8 + anomia9 + x + 5*e2
replace y = y*1000 in 3124
* Now do the analysis
reg y x
predict rstudent, rstudent
extremes rstudent y x
drop rstudent
reg y x if y > -9600
predict rstudent if e(sample), rstudent
extremes rstudent y x

* II-2
clear all
set obs 200
set seed 123456
gen x = rnormal(0, 10)
gen y = 7*x + 2*x^2 + rnormal(0,10)
reg y x
estat hettest
estat imtest
scatter y x
gen xsquare = x * x
reg y x xsquare
estat hettest
estat imtest

* II-3
webuse nhanes2f, clear
* Create data
clonedvar iq = weight
clonedvar fatheriq = hdresult
clonedvar motheriq = height
replace fatheriq = fatheriq * -1 /3
* Results
reg iq fatheriq motheriq
sum iq fatheriq motheriq
gen one = 1
impute fatheriq one, gen(fatheriq2)
gen mdfatheriq = missing(fatheriq)
sum iq fatheriq motheriq fatheriq2 mdfatheriq
reg iq fatheriq2 motheriq mdfatheriq

* III
webuse nhanes2f, clear
version 12.1
set seed 1234567
sample 70 if !missing(weight, height, age, female, black)
keep weight height age female black
* Cleverly disguise the data
gen stayfb = (weight * -1 + 270) * .4
gen frdties = 210 - height
gen white = black==0
drop height weight
order stayfb frdties age female white
* Run analyses
reg stayfb frdties age female white, beta
vif
test frdties age female white
sum
test white = female
estat hettest
estat imtest, white
pcorr2 stayfb frdties age female white
alpha frdties age female white, s