Sociology 63993
Exam 1 Answer Key
February 15, 2013

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

1. The independent variables in a model include X1, X2, and X1*X2. X1 and X2 both have missing values. If multiple imputation is used for X1 and X2, then passive imputation should be used to impute values for X1*X2.

False. This can result in a downward bias of the relationship between X1*X2 and other variables in the model. Instead it is better to treat X1*X2 as "just another variable," i.e. compute it first and then impute it just like you do X1 and X2.

2. A researcher runs the following analysis:

```
. alpha v1 v2 v3, i

Test scale = mean(unstandardized items)

                                                      average
                            item-test     item-rest       interitem
Item          |  Obs  Sign   correlation   correlation     covariance      alpha
--------------+------------------------------------------------------------------
v1            | 2500    +       0.7296        0.3393        .0357863      0.2613
v2            | 2500    +       0.6693        0.2537        .0634239      0.4150
v3            | 2500    +       0.6820        0.2610         .060012      0.4036
--------------+------------------------------------------------------------------
Test scale    |                                            .0530741      0.4610
-------------------------------------------------------------------------------
```

Based on these results, v1 should be dropped from the scale.

False. As the last column shows, the Cronbach's Alpha would decline if any of the variables were dropped from the scale, and this is especially true for v1.

3. In a bivariate regression, if a case is an extreme outlier on Y, then the closer its value on X is to the mean of X, the more impact the case will have on the slope coefficient.

False. The closer the X value is to the mean of X, the less leverage the case will have, and hence the less influence the case will have on the slope coefficient.

4. While random measurement error in the independent variables is problematic, random measurement error in the dependent variable has no adverse consequences.

False. Random measurement error in the dependent variable leads to increases in its variance, attenuated correlations with other variables, and larger standard error estimates.

5. Marital satisfaction is a key independent variable in the analysis. However, some subjects are not married. The Cohen and Cohen dummy variable adjustment technique is one way of dealing with this problem.

True. It isn't that subjects have failed to report their marital satisfaction; rather, for those who are not married, the value does not exist. Cohen and Cohen's approach can be useful in such cases.

II. *Short answer.* Discuss all <u>three</u> 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.

```
. logit diabetes wgt female black age, nolog

Logistic regression                                     Number of obs   =       8316
                                                        LR chi2(4)      =     347.55
                                                        Prob > chi2     =     0.0000
Log likelihood = -1201.6762                             Pseudo R2       =     0.1263

------------------------------------------------------------------------------
    diabetes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         wgt |   .0242748   .0035619     6.82   0.000     .0172936    .0312561
      female |   .4823663   .1217499     3.96   0.000     .2437409    .7209917
       black |   .8426339    .151018     5.58   0.000     .5466441    1.138624
         age |   .0623913   .0042484    14.69   0.000     .0540645    .0707181
       _cons |  -8.586301     .40814   -21.04   0.000    -9.386241   -7.786361
------------------------------------------------------------------------------

. sum diabetes wgt female black age

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    diabetes |     10335    .0482825     .214373          0          1
         wgt |      8316    71.83235    15.51516      30.84     175.88
      female |     10335    .5250121    .4993982          0          1
       black |     10335    .1050798    .3066711          0          1
         age |     10335    47.56584    17.21752         20         74

. mi set mlong
. mi register imputed wgt
(2019 m=0 obs. now marked as incomplete)
. mi impute pmm wgt diabetes female black age, add(10) knn(5) rseed(2232)

Univariate imputation                       Imputations =       10
Predictive mean matching                           added =       10
Imputed: m=1 through m=10                        updated =        0

                                       Nearest neighbors =        5

------------------------------------------------------------------
                  |               Observations per m
                  |-----------------------------------------------
         Variable |   Complete   Incomplete   Imputed |     Total
------------------+-----------------------------------+----------
              wgt |       8316         2019      2019 |     10335
------------------------------------------------------------------
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

```
. mi estimate, dots: logit diabetes wgt female black age

Imputations (10):
  .........10 done

Multiple-imputation estimates                     Imputations     =         10
Logistic regression                               Number of obs   =      10335
                                                  Average RVI     =     0.0374
                                                  Largest FMI     =     0.1465
DF adjustment:   Large sample                     DF:     min     =     442.32
                                                          avg     =  369971.66
                                                          max     = 1543449.26
Model F test:       Equal FMI                     F(   4,15282.4) =      78.10
Within VCE type:          OIM                     Prob > F        =     0.0000

------------------------------------------------------------------------------
    diabetes |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         wgt |   .0241408   .0031895     7.57   0.000     .0178724    .0304092
      female |   .3806918   .0993647     3.83   0.000     .1859312    .5754525
       black |   .6186384   .1288599     4.80   0.000     .3660765    .8712002
         age |   .0615972   .0038579    15.97   0.000     .0540359    .0691585
       _cons |  -8.426041   .3708681   -22.72   0.000     -9.15351   -7.698571
------------------------------------------------------------------------------
```

(a) The researcher was checking for missing data and observed that about 20% of the cases had missing values on wgt.

(b) She therefore decided to use multiple imputation to create estimates of the missing values. The specific imputation method used was Predictive Mean Matching (pmm). PMM can be used when imputing values for continuous variables. It may be preferable to linear regression when the normality of the variable is suspect (perhaps the researcher thought that would be the case for wgt). The basic idea is that you use regression methods to come up with an estimate of the missing value for variable X. However, rather than use that estimate, you identify one or more neighbors (in this case five) who have similar estimated values. (Note that it is the estimated value for the neighbor, not the neighbor's observed value.) The observed value of the nearest neighbor (or the randomly chosen nearest neighbor) is then used for the imputed value for the case with missing data on X. More generally, multiple imputation techniques lead to better estimates of standard errors because they do not treat the imputed values as though they were perfectly measured.

(c) The researcher has already tried one alternative, listwise deletion of missing data. But, that cost her 20% of her data, and produced somewhat different estimates (especially for black) than MI did. She could have used a single imputation technique, but those tend to produce inaccurate estimates of standard errors and can also produce biased coefficients. Probably her best alternative would be to use the regress imputation method rather than PMM, but again she may have had reasons for thinking PMM was better in this case.

NOTE: As a look at the Stata code in the appendix shows, the data were created to be MCAR – older people were more likely to have missing data, but among the old the data were missing at random. When data are MCAR, listwise deletion can produce biased estimates, which may explain why some coefficients were noticeably different after imputation.

II-2.

```
. reg hscore weight

      Source |       SS       df       MS              Number of obs =    6000
-------------+------------------------------           F(  1,  5998) = 5546.54
       Model |   789854.23     1   789854.23           Prob > F      =  0.0000
    Residual |   854144.85  5998  142.404943           R-squared     =  0.4804
-------------+------------------------------           Adj R-squared =  0.4804
       Total |  1643999.08  5999  274.045521           Root MSE      =  11.933

------------------------------------------------------------------------------
      hscore |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   .7383213   .0099137    74.48   0.000      .718887    .7577557
       _cons |    .197513   .7323928     0.27   0.787     -1.23824    1.633266
------------------------------------------------------------------------------

. dfbeta
                      _dfbeta_1: dfbeta(weight)
. extremes _dfbeta_1 hscore weight

  +--------------------------------------+
  |  obs:   _dfbeta_1     hscore   weight |
  |--------------------------------------|
  | 4652.   -1.002746        382    34.93 |
  | 4906.   -.1199883   51.05366   118.84 |
  | 3367.   -.0904366   79.91132   135.63 |
  | 4616.   -.0789586   59.08268   115.33 |
  | 5076.   -.0768352   73.24126   126.44 |
  +--------------------------------------+

  +--------------------------------------+
  | 5164.   .0792918    112.779   120.77 |
  | 4139.   .1081116   121.6261   123.72 |
  | 2123.   .1172843   136.6578   158.53 |
  | 1732.   .1235975   131.2471   144.24 |
  | 5611.   .1358087   140.1575   159.44 |
  +--------------------------------------+

. drop _dfbeta_1
. drop in 4652
(1 observation deleted)
```

```
. reg hscore weight

      Source |       SS       df       MS              Number of obs =    5999
-------------+------------------------------           F(  1,  5997) = 6669.63
       Model |  808826.687     1  808826.687           Prob > F      =  0.0000
    Residual |  727256.701  5997  121.270085           R-squared     =  0.5266
-------------+------------------------------           Adj R-squared =  0.5265
       Total |  1536083.39  5998  256.099264           Root MSE      =  11.012

------------------------------------------------------------------------------
      hscore |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   .7474949   .0091529    81.67   0.000      .729552    .7654378
       _cons |  -.5244449    .676231    -0.78   0.438    -1.850101     .801211
------------------------------------------------------------------------------

. dfbeta
                      _dfbeta_1: dfbeta(weight)

. extremes _dfbeta_1 hscore weight

  +--------------------------------------+
  |  obs:   _dfbeta_1     hscore   weight |
  |--------------------------------------|
  | 4905.   -.1313889   51.05366   118.84 |
  | 3367.   -.1005512   79.91132   135.63 |
  | 4616.   -.0866939   59.08268   115.33 |
  | 5075.   -.0850932   73.24126   126.44 |
  | 4930.   -.0845716   76.26996   129.05 |
  +--------------------------------------+

  +--------------------------------------+
  | 4654.   .0847623   111.6935   115.33 |
  | 4139.   .1155964   121.6261   123.72 |
  | 2123.   .1223528   136.6578   158.53 |
  | 1732.    .130713   131.2471   144.24 |
  | 5610.   .1423299   140.1575   159.44 |
  +--------------------------------------+
```

(a) The researcher was checking to see if outliers might be a problem in her data. She therefore computed the dfbeta value for each case in her analysis. Dfbeta shows how much a coefficient would change if that case were dropped from the data. According to the Stata 12 manual, “DFBETAs are perhaps the most direct influence measure of interest to model builders. DFBETAs focus on one coefficient and measure the difference between the regression coefficient when the ith observation is included and excluded, the difference being scaled by the estimated standard error of the coefficient. Belsley, Kuh, and Welsch (1980, 28) suggest observations with dfbetas > 2/Sqrt(N) should be checked as deserving special attention, but it is also common practice to use 1 (Bollen and Jackman 1990, 267), meaning that the observation shifted the estimate at least one standard error.”

Case 4652 had a dfbeta value of -1, much larger than any other case had. Further, its value on hscore (382) was extremely large compared to other cases, especially given its low value on weight.

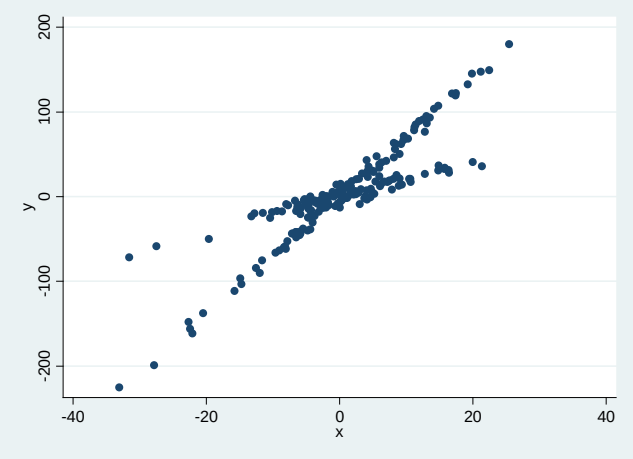
(b) The researcher decided to solve the problem simply by dropping case 4652 from the analysis. Perhaps she was convinced that the reported values for the case were wrong but she had no way of fixing them. Or, maybe she decided the case did not belong to her population of interest. Or, maybe she just didn't want to bother with a problematic case! In any event, dropping the case seemed to greatly reduce any concersna bout outliers.

(c) If the data were miscoded maybe she could have corrected the miscoding e.g. maybe hscore was supposed to be coded 38.2 rather than 382. Perhaps there are omitted variables that, if added to the model, could have accounted for the outlier. Robust regression techniques designed to deal with outliers (by placing less emphasis on them when computing estimates) might have been used, e.g. `qreg` could have been used for median regression. Luckily, the sample is large enough that if the researcher was making a mistake by excluding the outlier, it doesn't seem to have affected the regression estimates very much.

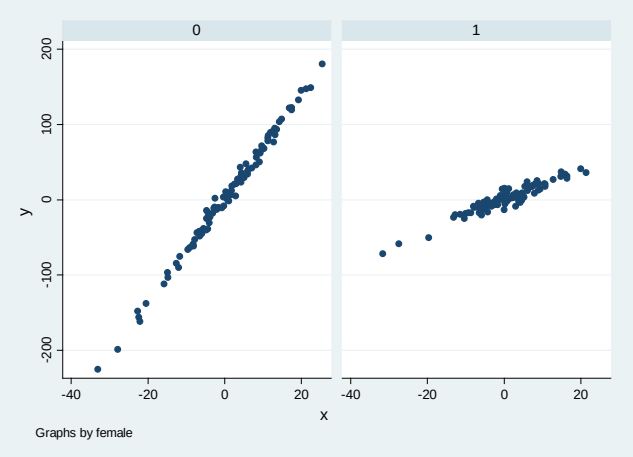
II-3.

```
. reg y x

      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  1,   198) =  824.88
       Model |  499764.343     1  499764.343           Prob > F      =  0.0000
    Residual |  119960.988   198  605.863576           R-squared     =  0.8064
-------------+------------------------------           Adj R-squared =  0.8055
       Total |  619725.331   199  3114.19765           Root MSE      =  24.614

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   5.027104   .1750341    28.72   0.000     4.681934    5.372274
       _cons |   .7003925    1.74319     0.40   0.688    -2.737208    4.137993
------------------------------------------------------------------------------

. estat hettest

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

         chi2(1)      =     3.36
         Prob > chi2  =   0.0670

. estat imtest

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |     149.71      2    0.0000
            Skewness |      18.73      1    0.0000
            Kurtosis |       2.19      1    0.1385
---------------------+-----------------------------
               Total |     170.64      4    0.0000
---------------------------------------------------
```

```
. scatter y x
```




```
. scatter y x, by(female)
```




```
. reg y x i.female i.female#c.x
```

```
      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  3,   196) = 5549.78
       Model |  612514.665     3  204171.555           Prob > F      =  0.0000
    Residual |  7210.66696   196  36.7891171           R-squared     =  0.9884
-------------+------------------------------           Adj R-squared =  0.9882
       Total |  619725.331   199  3114.19765           Root MSE      =  6.0654

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   6.909324   .0549867   125.65   0.000     6.800882    7.017765
    1.female |   .1125837   .8591587     0.13   0.896    -1.581799    1.806966
             |
  female#c.x |
          1  |  -4.904023    .088691   -55.29   0.000    -5.078934   -4.729111
             |
       _cons |   .3852018   .6080858     0.63   0.527    -.8140293    1.584433
------------------------------------------------------------------------------
```

```
. estat hettest

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

         chi2(1)      =     0.17
         Prob > chi2  =   0.6787

. estat imtest

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |       2.43      5    0.7871
            Skewness |       2.60      3    0.4576
            Kurtosis |       1.27      1    0.2601
---------------------+-----------------------------
               Total |       6.30      9    0.7099
---------------------------------------------------
```

(a) The researcher was concerned about heteroskedasticity – or, perhaps, a model misspecification that could cause the data to appear to be heteroskedastic. The initial `hettest` command did not show siginificant linear heteroskedasticity (where errors get larger as x gets larger) but the `imtest` suggested nonlinear heteroskedasticity might be present, e.g. something along the lines of the hourglass shape shown in the first scatter plot. But, when the researcher generated separate scatterplots by gender, she found that the slope for men appeared to be much steeper than the slope for women – suggesting that the real problem was not heteroskedasticity, but the pooling together of two populations that ought to somehow be treated separately. That is, it appears that the errors are heteroskedastic, but the real problem is that the two regression lines get further and further apart as the values of X get more extreme in either direction.

(b) The researcher addressed the problem by adding interaction terms to the model. This made it possible for the effects of x to differ by gender. When she did this, the heteroskedasticity tests no longer showed any problems. Basically, this means that, for each gender, the errors are homoscedastic.

(c) It might have been tempting to take the easy way out and just use robust standard errors, which relax the assumption of homoskedasticity. Or, she could have tried a complicated weighted least squares approach that would weight cases with large residuals less heavily. But, by examining her data more carefully and determining that the real problem was model misspecification rather than heteroskedasticity, the researcher probably came up with the best solution.

III. Computation and interpretation. (35 points total) The Republican Party is dismayed that it has lost the popular vote in 5 of the last 6 presidential elections. The party leadership strongly suspects that Vice-President Joe Biden will be the Democratic Party nominee in 2016. It has therefore commissioned political consultant Dick Morris to assess Biden's strengths and weaknesses. Morris has conducted a random survey of more than 5,000 registered voters. The variables he has collected data on are

Variable	Description
biden	How much the respondent likes Biden. Scores potentially range from a low of 0 to a high of 200
m47	Is the respondent a member of the 47%, i.e. the proportion of the population that does not pay federal income taxes (although most pay other types of taxes)? Coded 1 if yes, 0 otherwise
obamacare	How much does the respondent support Obama's health care program? Scores range from a low of 0 to a high of 10.
teaparty	Does the respondent consider himself or herself a member of the Tea Party? 1 = Tea Party, 0 = not Tea Party
black	Respondent's race (1 = black, 0 = not black)

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.]

```
. sum

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       biden |      5032    72.01731    15.41968      30.84     158.53
   obamacare |      5032    4.684237    1.383346   .5998579   9.042428
         m47 |      5032    .4789348    .4996057          0          1
    teaparty |      5032    .1405008    .3475404          0          1
       black |      5032    .1065183    .3085305          0          1

. alpha m47 teaparty black

Test scale = mean(unstandardized items)

Average interitem covariance:     .0011512
Number of items in the scale:            3
Scale reliability coefficient:      0.0219

. collin obamacare m47 teaparty black
(obs=5032)

  Collinearity Diagnostics

                       SQRT                   R-
  Variable      VIF     VIF    Tolerance    Squared
----------------------------------------------------
 obamacare      1.97    1.40    [1]          0.4927
       m47      1.95    1.40    0.5118       0.4882
  teaparty      1.02    1.01    0.9837       0.0163
     black      1.00    1.00    0.9986       0.0014
----------------------------------------------------
  Mean VIF      1.49
```

```
. reg biden m47 obamacare teaparty black, l(99)

      Source |       SS       df       MS              Number of obs =    5032
-------------+------------------------------           F(  4,  5027) =  [2]
       Model |  314848.283     4  78712.0707           Prob > F      =  0.0000
    Residual |  881355.261  5027  175.324301           R-squared     =  0.2632
-------------+------------------------------           Adj R-squared =  0.2626
       Total |  1196203.54  5031     [3]               Root MSE      =  13.241

------------------------------------------------------------------------------
       biden |      Coef.   Std. Err.      t    P>|t|     [99% Conf. Interval]
-------------+----------------------------------------------------------------
         m47 |   1.491152   .5222973     2.85   0.004     .1452928    2.837012
   obamacare |   5.184225   .1894643    [4]     0.000     4.696012    5.672438
    teaparty |  -6.736499   .5415667   -12.44   0.000    -8.132012   -5.340986
       black |     2.9218   .6054786     4.83   0.000     1.361598    4.482001
       _cons |   47.65426   .7544787    63.16   0.000     45.71011    49.59841
------------------------------------------------------------------------------

. test teaparty

 ( 1)  teaparty = 0

       F(  1,  5027) =  [5]
            Prob > F =    0.0000

. test obamacare = 5

 ( 1)  obamacare = 5

       F(  1,  5027) =    0.95
            Prob > F =    0.3309


. pcorr biden obamacare m47 teaparty black
(obs=5032)

Partial and semipartial correlations of biden with

               Partial   Semipartial      Partial   Semipartial   Significance
   Variable |    Corr.         Corr.      Corr.^2       Corr.^2          Value
------------+------------------------------------------------------------------
  obamacare |   0.3600        0.3313       0.1296        0.1097         0.0000
        m47 |   0.0402        0.0346       0.0016        0.0012         0.0043
   teaparty |  -0.1728       -0.1506       0.0299        0.0227         0.0000
      black |   0.0679        0.0584       0.0046        0.0034         0.0000

. estat hettest

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

         chi2(1)      =     0.22
         Prob > chi2  =   0.6421
```

```
. estat imtest

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      82.26     11    0.0000
            Skewness |     159.73      4    0.0000
            Kurtosis |      34.13      1    0.0000
---------------------+-----------------------------
               Total |     276.12     16    0.0000
---------------------------------------------------
```

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 are the uncensored parts of the printout.

```
. collin obamacare m47 teaparty black
(obs=5032)

  Collinearity Diagnostics

                       SQRT                   R-
  Variable      VIF     VIF    Tolerance    Squared
----------------------------------------------------
 obamacare      1.97    1.40    0.5073      0.4927
       m47      1.95    1.40    0.5118      0.4882
  teaparty      1.02    1.01    0.9837      0.0163
     black      1.00    1.00    0.9986      0.0014
----------------------------------------------------
  Mean VIF      1.49
```

```
. reg biden m47 obamacare teaparty black, l(99)

      Source |       SS       df       MS              Number of obs =    5032
-------------+------------------------------           F(  4,  5027) =  448.95
       Model |  314848.283     4  78712.0707           Prob > F      =  0.0000
    Residual |  881355.261  5027  175.324301           R-squared     =  0.2632
-------------+------------------------------           Adj R-squared =  0.2626
       Total |  1196203.54  5031  237.766556           Root MSE      =  13.241

------------------------------------------------------------------------------
       biden |      Coef.   Std. Err.      t    P>|t|     [99% Conf. Interval]
-------------+----------------------------------------------------------------
         m47 |   1.491152   .5222973     2.85   0.004     .1452928    2.837012
   obamacare |   5.184225   .1894643    27.36   0.000     4.696012    5.672438
    teaparty |  -6.736499   .5415667   -12.44   0.000    -8.132012   -5.340986
       black |     2.9218   .6054786     4.83   0.000     1.361598    4.482001
       _cons |   47.65426   .7544787    63.16   0.000     45.71011    49.59841
------------------------------------------------------------------------------
```

```
. test teaparty

 ( 1)  teaparty = 0

       F(  1,  5027) =  154.73
            Prob > F =    0.0000
```

To confirm that Stata got it right,

[1] $tol_{obamacare} = 1 - R^2_{obamacare.gobamare} = 1 - .4927 = .5073$. Less precisely, it equals $1/vif_{obamacare} = 1/1.97 = .5076$

[2] Global F test = MSR/MSE = MSR/MSE = 78712.0707/175.32 = 448.96. Those who prefer more of a challenge could do

$$F = \frac{R^2 * (N - K - 1)}{(1 - R^2) * K} = \frac{.2632 * (5032 - 4 - 1)}{(1 - .2632) * 4} = \frac{.2632 * (5032 - 4 - 1)}{(1 - .2632) * 4} = \frac{1323.11}{2.9472} = 448.94$$

[3] MST = SST/DFT = 1196203.54/5031 = 237.77. Or, if you prefer, remember that MST = Variance(biden) = SD(biden)^2 = 15.41968^2 = 237.77.

[4] $t_{obamacare} = b_{obamacare}/se_{obamacare} = 5.184225/.1894643 = 27.36$.

[5] When testing a single variable, $F_{teaparty} = T_{teaparty}^2 = -12.44^2 = 154.75$.

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 supportive of Biden and which are least supportive?

As the regression coefficients show, Biden gets greater support from members of the 47 percent and from blacks than he does from those who are not members of the 47 percent and who are not black. The more people like Obamacare, the more they tend to like Biden. However, members of the Tea Party are much less supportive of him than are those not in the Tea Party.

2. Based on the results of the `pcorr` command, one analyst thinks m47 should be dropped from the model. Explain what you think her reasoning is and why you agree or disagree.

The analyst probably noticed that m47 was the least statistically significant variable in the model, and that (as the squared semipartials show) dropping m47 would only decrease R^2 by .0012. Nonetheless, I personally would disagree with the decision to drop it. The effect is easily significant at even the .01 level; and, given the tremendous emphasis Mitt Romney placed on this variable in 2012, it probably needs to be included in the model even if its effects are substantively trivial.

3. There was concern that the variables teaparty, m47 and black would be highly collinear. Do you think that fear was justified? Would you recommend combining the items into a scale?

The `collin` command gives no indication of a collinearity problem with these variables. Both teaparty and black have near perfect tolerances, meaning they are largely uncorrelated with each other and with the other variables in the model. (Incidentally, somebody with a sharp eye might be surprised to find that race had nothing to do with Tea Party membership; I know I certainly would be if I didn't know the data were fake.) The tolerance for m47 is smaller but not small enough that it violates

any rules of thumb for concern. In any event, creating a scale out of the items would be a disaster, since the Scale reliability coefficient is only 0.0219. (Incidentally, the `alpha` command is smart enough to reverse the scoring when variables have negative rather than positive relationships with the other variables in the scale.)

4. The party leaders are upset because they thought the analysis revealed a clear violation of OLS assumptions but nothing was done about it. Why did they feel that way?

The Cameron and Trivedi IM-test strongly suggested that the data were heteroskedastic. There is no indication that anything was done about this. At a minimum robust standard errors could have been used. Better still would have been to check to see if there were problems with model misspecification, e.g. were variables omitted, should interaction terms have been included?

5. Previous studies had found that the slope coefficient for obamacare was 6. The Republican leaders wanted to see if that had changed, so, using the .01 level of significance, they wanted to test the hypothesis

$$H_0: \beta_{obamacare} = 6$$
$$H_A: \beta_{obamacare} \neq 6$$

Unfortunately Morris thought they said 5, not 6, so the wrong test was conducted. Explain to the party leaders why they still have the information they need to reject the null hypothesis.

There are at least two ways to do this. First, you can just look at the 99% confidence interval for obamacare. Any value specified in the null hypothesis that does not fall within the CI will cause the null hypothesis to be rejected at the .01 level if the alternative is 2 tailed. The upper limit of the CI is 5.67, which is less than 6, so reject the null; the effect of obamacare significantly differs from 6.

You also have enough information to compute the T statistic yourself:

$$T_{N-K-1} = \frac{b_k - \beta_{k0}}{s_{b_k}} = \frac{5.184225 - 6}{.1894643} = \frac{-.815775}{.1894643} = -4.30569$$

Given the large N, that is easily significant. If you prefer an F statistic, just square the above to get an F value of 18.54. But, if you just don't feel comfortable doing it yourself, the correct command in Stata is

```
. test obamacare = 6

 ( 1)  obamacare = 6

       F(  1,  5027) =   18.54
            Prob > F =    0.0000
```

Appendix: Stata Code

```
version 12.1
* Problem I-3.
use http://www3.nd.edu/~rwilliam/xsoc63993/statafiles/anomia.dta, clear
sample 2500, count
clonevar v1 = anomia5
clonevar v2 = anomia7
clonevar v3 = anomia2
alpha v1 v2 v3, i

* Problem II-1
webuse nhanes2f, clear
drop if missing(diabetes, weight, female, black, age)
clonevar wgt = weight
replace wgt = . if uniform() < .40 & age > 50
sum diabetes wgt female black age
logit diabetes wgt female black age, nolog
mi set mlong
mi register imputed wgt
mi impute pmm wgt diabetes female black age, add(10) knn(5) rseed(2232)
mi estimate, dots: logit diabetes wgt female black age

* Problem II-2
webuse nhanes2f, clear
set seed 12345
sample 6000, count
gen hscore = .74*weight + rnormal(0, 11)
replace hscore = 382 in 4652
reg hscore weight
dfbeta
extremes _dfbeta_1 hscore weight
drop _dfbeta_1
drop in 4652
reg hscore weight
dfbeta
extremes _dfbeta_1 hscore weight

* Problem II-3
clear all
set obs 200
gen female = _n > 100
label define gender 0 "Male" 1 "Female"
set seed 123456
gen x = rnormal(0, 10)
gen y = 7*x + rnormal(0,6) if !female
replace y = 2*x + rnormal(0,6) if female
reg y x
estat hettest
estat imtest
scatter y x
scatter y x, by(female)
reg y x i.female i.female#c.x
estat hettest
estat imtest

* Problem III
* Cleverly disguise the data
webuse nhanes2f, clear
set seed 56789
sample 5032, count
gen biden = weight
gen obamacare = (height - 135)/7
```

```
gen m47 = female==0
gen teaparty = age <= 25
keep biden obamacare m47 teaparty black
order biden obamacare m47 teaparty black
* Start analyses
sum
alpha m47 teaparty black
collin obamacare m47 teaparty black
reg biden m47 obamacare teaparty black, l(99)
test teaparty
test obamacare = 5
pcorr biden obamacare m47 teaparty black
estat hettest
estat imtest
* Correct test command
test obamacare = 6
```