Sociology 63993
Exam 1 Answer Key
February 17, 2012

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

1. Cohen and Cohen's dummy variable adjustment method is useful when variables like gender or age have missing values.

False. The method should not be used when values exist but are not known (and values for gender and age surely exist). The method can be useful when values don't exist, e.g. father's education is missing because there is no father in the family.

2. R^2 is biased downwards.

False. It is biased upwards. Sampling error will always cause R^2 to be greater than zero, i.e. even if no variable has an effect R^2 will be positive in a sample. When there are no effects, across multiple samples you will see estimated coefficients sometimes positive, sometimes negative, but either way you are going to get a non-zero positive R^2. Further, when there are many Xs for a given sample size, there is more opportunity for R^2 to increase by chance. Adjusted R^2 corrects for this bias.

3. The more "tolerant" a variable is (i.e. the less highly correlated it is with the other IVs), the smaller its unique contribution to R^2 will be.

False. The more tolerant a variable is, the more unique (and higher) its contribution to R^2 will be. You can see this via such formulas as

$$sr_k^2 = R_{YH}^2 - R_{YG_k}^2 = b_k'^2 * (1 - R_{X_k G_k}^2) = b_k'^2 * Tol_k$$

4. When you have more than one independent variable, random measurement error can cause coefficients to be biased either upward or downward.

True. In bivariate regression, the bias will be downward, but once you have more than one independent variable the bias can go in either direction.

5. A Durbin-Watson statistic of 4 or greater indicates that the case is an extreme outlier.

False. The Durbin-Watson statistic checks for serial correlation.

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.

```
. use "http://www.nd.edu/~rwilliam/stats3/statafiles/rwm11.dta", clear
(German Health Care Panel Data, Riphahn Wambach Million (2003), Greene (2007))

. reg newhsat female age handdum educ married working if year==1984

      Source |       SS       df       MS              Number of obs =    3874
-------------+------------------------------           F(  6,  3867) =  141.19
       Model |   4483.9589     6  747.326483            Prob > F      =  0.0000
    Residual |  20468.7796  3867  5.29319359            R-squared     =  0.1797
-------------+------------------------------           Adj R-squared =  0.1784
       Total |  24952.7385  3873  6.44274168            Root MSE      =  2.3007

------------------------------------------------------------------------------
     newhsat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -.186618   .0870046    -2.14   0.032    -.3571973   -.0160387
         age |  -.0388289   .0035637   -10.90   0.000    -.0458159    -.031842
     handdum |  -2.341489   .1236374   -18.94   0.000     -2.58389   -2.099088
        educ |   .1089876   .0173546     6.28   0.000     .0749626    .1430125
     married |   .2048268   .0918166     2.23   0.026     .0248133    .3848402
     working |   .2955985   .0912629     3.24   0.001     .1166704    .4745265
       _cons |   7.409179   .2928951    25.30   0.000     6.834935    7.983422
------------------------------------------------------------------------------

. estat hettest

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

         chi2(1)      =    55.33
         Prob > chi2  =   0.0000

. reg newhsat female age handdum educ married working if year==1984, robust

Linear regression                                       Number of obs =    3874
                                                        F(  6,  3867) =  129.51
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.1797
                                                        Root MSE      =  2.3007

------------------------------------------------------------------------------
             |               Robust
     newhsat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -.186618   .0903374    -2.07   0.039    -.3637316   -.0095044
         age |  -.0388289   .0036116   -10.75   0.000    -.0459097   -.0317482
     handdum |  -2.341489   .1363509   -17.17   0.000    -2.608816   -2.074163
        educ |   .1089876   .0161526     6.75   0.000     .0773191    .1406561
     married |   .2048268   .0946607     2.16   0.031      .019237    .3904165
     working |   .2955985   .0974066     3.03   0.002     .1046252    .4865717
       _cons |   7.409179   .2968794    24.96   0.000     6.827124    7.991234
------------------------------------------------------------------------------
```

The hettest command revealed that the data were heteroskedastic, i.e. errors were not iid. The researcher therefore used robust standard errors, which relax the assumption that the errors are identically distributed. This may be ok, but the researcher should probably check out some other options too. For example, maybe the slopes differ by gender. Or, some important variable may have been left out. Correcting either of these problems might eliminate the heteroskedasticity. If the researcher had a clear enough theory, weighted least squares might be another way of dealing with the problem.

II-2.

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

      Source |       SS       df       MS              Number of obs =    1290
-------------+------------------------------           F(  6,  1283) =   27.43
       Model |  124.537637     6  20.7562729           Prob > F      =  0.0000
    Residual |  970.982518  1283  .756806327           R-squared     =  0.1137
-------------+------------------------------           Adj R-squared =  0.1095
       Total |  1095.52016  1289  .849899267           Root MSE      =  .86995

------------------------------------------------------------------------------
        warm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        yr89 |   .2238435   .0502656     4.45   0.000     .1252318    .3224553
        male |  -.2846409    .048767    -5.84   0.000    -.3803127   -.1889691
       white |  -.2322106    .074535    -3.12   0.002    -.3784345   -.0859867
         age |  -.0086944    .001544    -5.63   0.000    -.0117234   -.0056654
          ed |   .0399421   .0098042     4.07   0.000     .0207081    .0591761
        prst |   .0019401   .0020726     0.94   0.349     -.002126    .0060063
       _cons |   2.688483   .1437892    18.70   0.000     2.406395    2.970571
------------------------------------------------------------------------------

. sum warm yr89 male white age ed prst

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        warm |      2293    2.607501    .9282156          1          4
        yr89 |      2293    .3986044    .4897178          0          1
        male |      2293    .4648932    .4988748          0          1
       white |      1712    .8785047    .3267975          0          1
         age |      2293    44.93546    16.77903         18         89
-------------+--------------------------------------------------------
          ed |      1709     12.1849    3.179042          0         20
        prst |      2293    39.58526    14.49226         12         82

. mi set mlong
. mi register imputed white ed
(1003 m=0 obs. now marked as incomplete)

. mi register regular warm yr89 male age prst
. mi impute chained (logit) white (regress) ed = warm yr89 male age prst, add(50) rseed(1234)

Conditional models:
             white: logit white ed warm yr89 male age prst
                ed: regress ed i.white warm yr89 male age prst

Performing chained iterations ...

Multivariate imputation                     Imputations =       50
Chained equations                                 added =       50
Imputed: m=1 through m=50                       updated =        0

Initialization: monotone                     Iterations =      500
                                                burn-in =       10

             white: logistic regression
                ed: linear regression

------------------------------------------------------------------
                  |              Observations per m
                  |----------------------------------------------
         Variable |   Complete   Incomplete   Imputed |     Total
------------------+-----------------------------------+----------
            white |       1712          581       581 |      2293
               ed |       1709          584       584 |      2293
------------------------------------------------------------------
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

```
. mi estimate, dots: reg warm yr89 male white age ed prst

Imputations (50):
  .........10.........20.........30.........40.........50 done

Multiple-imputation estimates                     Imputations     =         50
Linear regression                                 Number of obs   =       2293
                                                  Average RVI     =     0.1214
                                                  Largest FMI     =     0.3154
                                                  Complete DF     =       2286
DF adjustment:   Small sample                     DF:     min     =     380.43
                                                          avg     =    1356.48
                                                          max     =    2270.39
Model F test:       Equal FMI                     F(   6, 2035.1) =      46.46
Within VCE type:          OLS                     Prob > F        =     0.0000

------------------------------------------------------------------------------
        warm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        yr89 |   .2607513   .0380632     6.85   0.000     .1861084    .3353942
        male |  -.3335333   .0366975    -9.09   0.000    -.4054975   -.2615691
       white |  -.1599008    .068205    -2.34   0.020    -.2940068   -.0257948
         age |  -.0098668   .0011996    -8.22   0.000    -.0122195   -.0075141
          ed |   .0340434   .0087675     3.88   0.000     .0168149    .0512719
        prst |   .0023128   .0016503     1.40   0.161    -.0009248    .0055503
       _cons |   2.735879   .1198425    22.83   0.000     2.500675    2.971083
------------------------------------------------------------------------------
```

The initial regression only had 1290 cases even though there are 2293 cases in the data. The sum command showed that both white and ed are missing about 25% of their cases. Rather than lose more than 40% of her data, the researcher decided to use multivariate Imputation using Chained Equations (ICE). MI [multiple imputation] is a procedure by which missing data are imputed several times to produce several different complete-data estimates of the parameters. The parameter estimates from each imputation are then combined to give an overall estimate of the complete-data parameters as well as reasonable estimates of the standard errors. (If you only did one imputation, the estimates and standard errors would not reflect the fact that the imputed values are themselves uncertain and subject to error.)

The specific MI method employed here, ICE, uses iterative procedures to impute missing values when more than one variable is missing. These variables can be of different types, e.g. they might be binary, ordinal or continuous. In this case, white is dichotomous, so logit is used to impute its values. Ed is continuous, so regress is used. (Note how logit and regress are both specified on the mi impute command.) She creates 50 imputed data sets, and all of the missing data are replaced with imputed values. Note that the imputation models are congenial, i.e. the imputation models include the same variables (including the dependent variable) that are in the analytic model; otherwise relationships with the variables that had been omitted would be biased toward 0.

The researcher has already tried one alternative, listwise deletion. This is often ok, but in this case it loses her more than 40% of her cases. The coefficients are a bit different between the two approaches, but probably the most striking difference is that the standard errors are smaller with multiple imputation, reflecting the larger sample size.

The researcher may also wish to make sure the data really are missing, e.g. sometimes the same questions get asked for different people at different points in the interview.

II-3.

```
. reg docvis educ ses female

      Source |       SS       df       MS              Number of obs =   27326
-------------+------------------------------           F(  3, 27322) =  138.65
       Model |  13264.8729     3   4421.6243           Prob > F      =  0.0000
    Residual |   871315.75 27322  31.8906284           R-squared     =  0.0150
-------------+------------------------------           Adj R-squared =  0.0149
       Total |  884580.623 27325  32.3725754           Root MSE      =  5.6472

------------------------------------------------------------------------------
      docvis |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   -.366295   .2382335    -1.54   0.124    -.8332447    .1006548
         ses |   .1992199   .2378011     0.84   0.402    -.2668824    .6653222
      female |   1.023064   .0695639    14.71   0.000     .8867151    1.159413
       _cons |   4.535222   .1913001    23.71   0.000     4.160265     4.91018
------------------------------------------------------------------------------

. corr docvis educ ses female
(obs=27326)

             |   docvis     educ      ses   female
-------------+------------------------------------
      docvis |   1.0000
        educ |  -0.0847   1.0000
         ses |  -0.0843   0.9981   1.0000
      female |   0.1023  -0.1831  -0.1832   1.0000


. reg docvis educ female

      Source |       SS       df       MS              Number of obs =   27326
-------------+------------------------------           F(  2, 27323) =  207.63
       Model |  13242.4908     2  6621.24541           Prob > F      =  0.0000
    Residual |  871338.132 27323  31.8902804           R-squared     =  0.0150
-------------+------------------------------           Adj R-squared =  0.0149
       Total |  884580.623 27325  32.3725754           Root MSE      =  5.6471

------------------------------------------------------------------------------
      docvis |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.1671061   .0149471   -11.18   0.000    -.1964031   -.1378091
      female |   1.022659   .0695618    14.70   0.000      .886314    1.159003
       _cons |   4.585648   .1815832    25.25   0.000     4.229735     4.94156
------------------------------------------------------------------------------
```

The researcher may have been surprised that both educ and ses had insignificant effects. The correlation command revealed that the two variable were very highly correlated, .9981. To deal with this multicollinearity, the researcher dropped ses, which had a slightly lower correlation with docvis. After doing so, educ had a highly significant effect.

This may be a very questionable strategy. It could result in omitted variable bias. In a slightly different sample, ses might have done slightly better than educ. One wonders if educ was used to compute ses, in which case keeping ses might be the better idea. If educ was not used to compute ses, perhaps the two could be combined in a scale.

III. Computation and interpretation. (35 points total) President Obama's plan to provide free birth control to most women has proven to be far more controversial than he expected. The President has therefore commissioned a study of 7,500 Americans to see where the public stands. The variables are

Variable	Description
bcontrol	Support for Obama's birth control policy. Ranges from a low of 0 (strongly oppose the policy) to a high of 100 (strongly favor)
catholic	Coded 1 if the respondent is Catholic, 0 otherwise
female	Coded 1 if the respondent is female, 0 otherwise
health	Overall health of the respondent. Ranges from 0 (very poor health) to 100 (very good health).
liberal	How liberal is the respondent? Ranges from 0 (very conservative) to 100 (very liberal).

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
-------------+--------------------------------------------------------
    bcontrol |      7500    43.13119    9.173105     18.504     95.118
    catholic |      7500    .5262667    .4993429          0          1
      female |      7500    .1141333    .3179943          0          1
      health |      7500    57.41967    9.648723         25         87
     liberal |      7500    62.04155    22.21342         26       96.2

. reg bcontrol catholic female health liberal

      Source |       SS       df       MS              Number of obs =    7500
-------------+------------------------------           F(  4,  7495) =     [1]
       Model |  153728.687     4  38432.1718           Prob > F      =  0.0000
    Residual |  477281.098  7495  63.6799331           R-squared     =     [2]
-------------+------------------------------           Adj R-squared =
       Total |  631009.786  7499  84.1458575           Root MSE      =    7.98

------------------------------------------------------------------------------
    bcontrol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    catholic |  -.8740108   .2610363    -3.35   0.001    -1.385715   -.3623065
      female |   2.189993    .289954     [3]    0.000     1.621602    2.758385
      health |  -.4399353   .0138063   -31.86   0.000    -.4669995   -.4128712
     liberal |   .0540596   .0043219    12.51   0.000     .0455874    .0625318
       _cons |    65.2482   .6861085    95.10   0.000     63.90324    66.59317
------------------------------------------------------------------------------

. collin catholic female health liberal
(obs=7500)

  Collinearity Diagnostics

                       SQRT                   R-
  Variable      VIF     VIF    Tolerance    Squared
----------------------------------------------------
  catholic      2.00    1.41     [4]        0.5002
    female      1.00    1.00    0.9989      0.0011
    health      2.09    1.45    0.4785      0.5215
   liberal      1.09    1.04    0.9213      0.0787
----------------------------------------------------
  Mean VIF      1.54
```

```
                        Cond
       Eigenval         Index
--------------------------------
    1     3.6921         1.0000
    2     0.8578         2.0746
    3     0.3751         3.1373
    4     0.0671         7.4161
    5     0.0078        21.6930
--------------------------------
 Condition Number       21.6930
 Eigenvalues & Cond Index computed from scaled raw sscp (w/ intercept)
 Det(correlation matrix)    0.4780

. test liberal

 ( 1)  liberal = 0

       F(  1,  7495) =    [5]
            Prob > F =    0.0000

. test female = -catholic

 ( 1)  catholic + female = 0

       F(  1,  7495) =   11.55
            Prob > F =    0.0007

. test catholic female health liberal

 ( 1)  catholic = 0
 ( 2)  female = 0
 ( 3)  health = 0
 ( 4)  liberal = 0

       F(  4,  7495) =  603.52
            Prob > F =    0.0000

. alpha catholic female health liberal

Test scale = mean(unstandardized items)

Average interitem covariance:     7.936035
Number of items in the scale:            4
Scale reliability coefficient:      0.1862

. alpha catholic female health liberal, s

Test scale = mean(standardized items)

Average interitem correlation:      0.1506
Number of items in the scale:            4
Scale reliability coefficient:      0.4149

. predict rstandard, rstandard
. extremes rstandard rstandard bcontrol catholic female health liberal

  +-------------------------------------------------------------------------------------+
  |  obs:   rstandard   rstandard   bcontrol   catholic   female     health   liberal |
  |-------------------------------------------------------------------------------------|
  | 4166.   -2.774446   -2.774446      28.44          0        1     49.703      92.3 |
  | 5909.   -2.773224   -2.773224     27.966          0        0         45      85.8 |
  |  710.   -2.753186   -2.753186     18.504          0        0     67.203      88.4 |
  | 4022.   -2.747619   -2.747619      29.94          0        1   46.30099      88.4 |
  | 2839.   -2.601393   -2.601393       24.9          0        0   55.40199      88.4 |
  +-------------------------------------------------------------------------------------+

  +-------------------------------------------------------------------------------------+
  | 4213.   5.357062   5.357062     78.924          1        0     68.703      37.7 |
  |  742.   5.520198   5.520198     83.346          1        1     66.703        39 |
  | 2097.   5.758528   5.758528     89.742          1        0   57.30099      85.8 |
  | 1592.   5.899137   5.899137     95.118          0        1   48.80099        39 |
  | 6511.   5.932644   5.932644     88.926          1        1   63.90199      58.5 |
  +-------------------------------------------------------------------------------------+
```

```
. pcorr bcontrol catholic female health liberal
(obs=7500)

Partial and semipartial correlations of bcontrol with

              Partial   Semipartial      Partial   Semipartial   Significance
   Variable |    Corr.         Corr.      Corr.^2       Corr.^2         Value
------------+------------------------------------------------------------------
   catholic |  -0.0386       -0.0336       0.0015        0.0011         0.0008
     female |   0.0869        0.0759       0.0076        0.0058         0.0000
     health |  -0.3454       -0.3201       0.1193        0.1025         0.0000
    liberal |   0.1430        0.1257       0.0204        0.0158         0.0000
```

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

First, here are the uncensored parts of the printout.

```
. reg bcontrol catholic female health liberal

      Source |       SS       df       MS              Number of obs =    7500
-------------+------------------------------           F(  4,  7495) =  603.52
       Model |  153728.687     4  38432.1718           Prob > F      =  0.0000
    Residual |  477281.098  7495  63.6799331           R-squared     =  0.2436
-------------+------------------------------           Adj R-squared =  0.2432
       Total |  631009.786  7499  84.1458575           Root MSE      =    7.98

------------------------------------------------------------------------------
    bcontrol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    catholic |  -.8740108   .2610363    -3.35   0.001    -1.385715   -.3623065
      female |   2.189993    .289954     7.55   0.000     1.621602    2.758385
      health |  -.4399353   .0138063   -31.86   0.000    -.4669995   -.4128712
     liberal |   .0540596   .0043219    12.51   0.000     .0455874    .0625318
       _cons |    65.2482   .6861085    95.10   0.000     63.90324    66.59317
------------------------------------------------------------------------------

. collin catholic female health liberal
(obs=7500)

  Collinearity Diagnostics

                       SQRT                   R-
  Variable      VIF     VIF    Tolerance    Squared
----------------------------------------------------
  catholic      2.00    1.41    0.4998      0.5002
    female      1.00    1.00    0.9989      0.0011
    health      2.09    1.45    0.4785      0.5215
   liberal      1.09    1.04    0.9213      0.0787
----------------------------------------------------
  Mean VIF      1.54 [Rest of Collin output appears above]

. test liberal

 ( 1)  liberal = 0

       F(  1,  7495) =  156.46
            Prob > F =    0.0000
```

To confirm that Stata got it right,

[1] = Global F test = MSR/MSE = MSR/MSE = 38432.17/63.68 = 603.52. Of course, those who prefer to do things the easy way can simply note that the third test command already computed this value for you.

[2] = R^2 = SSR/SST = 153728.687/631009.786 = .2436

[3] $t_{female} = b_{female}/se_{female} = 2.189993/.289954 = 7.55$

[4] $tol_{catholic} = 1 - R^2_{catholic.gcatholic} = 1 - .5002 = .4998$. Less precisely, it equals $1/vif_{catholic} = 1/2 = .5$

[5] When testing a single variable, $F_{liberal} = T_{liberal}^2 = 12.51^2 = 156.5$

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. What groups or types of individuals are most supportive of the President's policy and which are least supportive?

The regression results show you that Catholics and those who are healthier tend to be less supportive of the President's policy. Women and liberals tend to be more supportive. All of these variable have significant effects but the health variable is the most significant.

2. The researchers were worried that outliers might be problematic. Based on the results, do you see any reasons to be concerned?

There are indeed some large outliers of 5 or greater. Of course, it is a large sample, so some large outliers are to be expected, but probably not this many this large. There are no obvious coding mistakes. Assuming everything is coded correctly, the researchers may wish to examine whether adding some variables or otherwise modifying the model could reduce the magnitude of the outliers.

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 items in order to deal with the problem?

Two different scaling commands are used, and both result in very low values for Cronbach's Alpha. If measurement error is a problem, the researcher needs to find some other way to deal with it.

4. How would the R^2 value change if the variable liberal were dropped from the model? Do you think that would be a good idea?

The squared semipartial value for liberal in the pcorr command shows us that R^2 would decline by .0158 (from .2436 to .2278) if liberal were dropped. The effect of liberal is highly significant (in fact more significant than any other variable except health) so dropping it would probably be a bad idea. But if you just love to do things the hard way (or don't read printouts very thoroughly) you could do

$$sr_{liberal} = \frac{T_{liberal} * \sqrt{1 - R^2_{YH}}}{\sqrt{N - K - 1}} = \frac{12.51 * \sqrt{1 - .2436}}{\sqrt{7500 - 4 - 1}} = \frac{10.88}{86.574} = .12567,$$

$$sr^2_{liberal} = .12567^2 = .0158$$

To confirm,

```
. reg bcontrol catholic female health

      Source |       SS       df       MS              Number of obs =    7500
-------------+------------------------------           F(  3,  7496) =  737.25
       Model |  143765.611     3  47921.8703           Prob > F      =  0.0000
    Residual |  487244.175  7496  65.0005569           R-squared     =  0.2278
-------------+------------------------------           Adj R-squared =  0.2275
       Total |  631009.786  7499  84.1458575           Root MSE      =  8.0623

[Rest of output omitted]
```

Or, to make the calculations even easier, i.e. let Stata do the work,

```
. nestreg, quietly: reg bcontrol (catholic female health) liberal

Block  1: catholic female health
Block  2: liberal

  +-------------------------------------------------------------------+
  |       |          Block  Residual                        Change |
  | Block |       F     df        df   Pr > F       R2     in R2 |
  |-------+-----------------------------------------------------------|
  |     1 |  737.25      3      7496   0.0000   0.2278           |
  |     2 |  156.46      1      7495   0.0000   0.2436   0.0158 |
  +-------------------------------------------------------------------+
```

5. The President's advisors believe that Female support for health care reform is stronger than Catholic opposition to it. Do you think they are right?

The estimated effect of female is more than twice as large in magnitude as the estimated effect of catholic. The command test female = -catholic shows that this difference is very statistically significant. So, it looks like the advisors are right.

c) (1 point extra credit) As soon as the President started reading the results, he became concerned that something might be seriously wrong with the data. Why?

The President was immediately suspicious when he saw that 52.6% of the sample is Catholic, since that is about double what it is in the population. But, even somebody who missed that would realize that 11.4% female is way too low. The sampling procedure may be seriously flawed, or maybe some variables have been mislabeled. Further investigation revealed that the person writing the exam was not as observant as the President was and, rather than rewrite the whole problem, thought this question would be a clever way of making it look like he had planned it this way all along.

Appendix: Stata Code

```
use "http://www.indiana.edu/~jslsoc/stata/spex_data/ordwarm2.dta", clear
* Exam 1, 2012
version 12.1

* Problem II-1
use "http://www.nd.edu/~rwilliam/stats3/statafiles/rwm11.dta", clear
reg newhsat female age handdum educ married working if year==1984
estat hettest
reg newhsat female age handdum educ married working if year==1984, robust

* Problem II-2
use "http://www.indiana.edu/~jslsoc/stata/spex_data/ordwarm2.dta", clear
*** Several values will be changed to missing ***
set seed 123456
gen mdwhite = uniform() < .25
gen mded = uniform() < .25
replace white = . if mdwhite
replace ed = . if mded
reg warm yr89 male white age ed prst
sum warm yr89 male white age ed prst
mi set mlong
mi register imputed white ed
mi register regular warm yr89 male age prst
mi impute chained (logit) white (regress) ed = warm yr89 male age prst, add(50) rseed(1234)
mi estimate, dots: reg warm yr89 male white age ed prst

* Problem II-3
use "http://www.nd.edu/~rwilliam/stats3/statafiles/rwm11.dta", clear
*** ses is constructed so it will be very highly correlated with educ ***
set seed 123456
gen ses = educ + runiform() * .5
reg docvis educ ses female
corr docvis educ ses female
reg docvis educ female

* Probem III
webuse nhanes2f, clear
keep in 1/7500
keep weight height age female black
*** Cleverly disguise the data ***
gen bcontrol = weight * .6
drop weight
gen health = (225 - height)
drop height
gen catholic = female
drop female
gen female = black
drop black
gen liberal = age * 1.3
drop age
order bcontrol catholic female health liberal
*** Run analyses ***
sum
reg bcontrol catholic female health liberal
collin catholic female health liberal
test liberal
test female = -catholic
test catholic female health liberal
alpha catholic female health liberal
alpha catholic female health liberal, s
predict rstandard, rstandard
extremes rstandard rstandard bcontrol catholic female health liberal
pcorr bcontrol catholic female health liberal
*** Extra runs ***
reg bcontrol catholic female health
nestreg, quietly: reg bcontrol (catholic female health) liberal
```