# Sociology 593
# Exam 1 Answer Key
# February 11, 2005

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

1. A researcher is trying to construct a scale that measures political liberalism. She obtains the following results:

```
. alpha lib*, c i

Test scale = mean(unstandardized items)

                                                            average
                                  item-test     item-rest     inter-item
Item         |  Obs  Sign   correlation   correlation     covariance      alpha
-------------+-----------------------------------------------------------------
liberal1     | 3975    +       0.8436        0.3092        .0592051      0.5014
liberal2     | 3975    +       0.5812        0.4236        .0544034      0.7169
liberal3     | 3975    +       0.5262        0.3809        .0575775      0.7238
liberal4     | 3975    +       0.5010        0.3260        .0578235      0.7337
liberal5     | 3975    +       0.6197        0.4707        .0527308      0.7086
liberal6     | 3975    +       0.6542        0.5158        .0513626      0.7007
liberal7     | 3975    +       0.5306        0.3668        .0567058      0.7264
liberal8     | 3975    +       0.6010        0.4630        .0543518      0.7108
liberal9     | 3975    +       0.3012        0.5006        .0515661      0.8392
-------------+-----------------------------------------------------------------
Test scale   |                                           .0550807      0.7459
-------------------------------------------------------------------------------
```

Based on these results, if she wants to increase the reliability of her scale, she should drop liberal1.

False. The last column shows you what the Cronbach's Alpha would be if the item was deleted. Thus, if you dropped liberal1, the reliability of the scale would go down, from .7459 to .5014. If she wants to increase the reliability of the scale she should drop liberal9.

2. $sr_1^2 = .23, sr_2^2 = .15$. Therefore, dropping both x1 and x2 from the equation will reduce $R^2$ by .38.

False (unless x1 and x2 are uncorrelated). Semipartial correlations have to be recomputed after each variable is dropped.

3. In a bivariate regression, random measurement error in X causes the slope coefficient to be attenuated. Unfortunately, increasing the sample size will not alleviate this problem.

True. A larger sample size will not affect the attenuation bias that is caused by random measurement error.

4. Religion is coded 1 = Catholic, 2 = Protestant, 3 = Other. However, information on religion is missing for several respondents. According to Allison and others, the best way to deal with this problem is to treat missing as another category of religion and then construct 3 dummy variables from religion.

False. As pointed out in the notes on missing data, Allison says that this procedure will produce biased estimates.

5. Outlying values on Y will have the greatest influence on regression coefficients when (a) their corresponding X values are close to the mean of X, and (b) the Y value is out of line with the rest of the Y values, i.e. it does not fall on the same line that the other cases do.

False (or only half-true). Part (b) is right, but for (a) outliers on Y that are paired with average values of X will have less influence on parameter estimates than outliers on Y that are paired with above or below-average values on X. Influence on coefficients = Leverage * Discrepancy, and cases that are further from the mean of X have greater leverage.

*II.* *Short answer.* Discuss <u>three</u> of the following five problems. (15 points each, 45 points total, up to 5 points extra credit for each additional problem.) 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 =      80
-------------+------------------------------           F(  1,    78) =    0.64
       Model |  8161.29461     1  8161.29461           Prob > F      =  0.4245
    Residual |  987754.387    78  12663.5178           R-squared     =  0.0082
-------------+------------------------------           Adj R-squared = -0.0045
       Total |  995915.682    79  12606.5276           Root MSE      =  112.53

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -2.032806   2.532175    -0.80   0.425    -7.073978    3.008366
       _cons |  -12.75283   12.58149    -1.01   0.314    -37.80066      12.295
------------------------------------------------------------------------------

. predict rstudent, rstudent
. predict cooksd, cooksd
. extremes  rstudent cooksd y x

  +------------------------------------------------------+
  | obs:    rstudent     cooksd           y            x |
  |------------------------------------------------------|
  |  75.   -188.0042   1.870291        -999     8.116624 |
  |   1.   -.3794509   .0082487   -26.45146    -13.27932 |
  |   2.   -.2554911   .0021359   -20.94568    -9.759867 |
  |  10.   -.1944049   .0005867   -22.56966    -5.837373 |
  |   3.   -.1862795    .000874   -16.46544    -8.301276 |
  +------------------------------------------------------+

  +---------------------------------------------------+
  |  76.   .4368796   .0052328   17.54447   8.773149 |
  |  77.   .4445563   .0059247   17.25927   9.270648 |
  |  78.   .4700404   .0067283   19.84505   9.361078 |
  |  74.   .4758831   .0050422   24.34256   7.662218 |
  |  80.    .511474   .0102893   20.86036   10.86633 |
  +---------------------------------------------------+

. preserve
. drop if y == -999
(1 observation deleted)

. * [CONTINUED NEXT PAGE]
```

```
. reg y x

      Source |       SS       df       MS              Number of obs =      79
-------------+------------------------------           F(  1,    77) =  314.60
       Model |  8772.60996     1  8772.60996           Prob > F      =  0.0000
    Residual |  2147.13658    77  27.8848907           R-squared     =  0.8034
-------------+------------------------------           Adj R-squared =  0.8008
       Total |  10919.7465    78  139.996751           Root MSE      =  5.2806

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   2.144088   .1208823    17.74   0.000     1.903381    2.384795
       _cons |  -.0483975   .5942454    -0.08   0.935    -1.231691    1.134896
------------------------------------------------------------------------------

. restore
```

The researcher is concerned about outliers. The studentized residual measures discrepancy while the Cook statistic measures influence. By listing extreme values, the researcher finds that the residual statistics for case 75 are enormously large and that the y value for that case is -999. The researcher opts to simply drop the outlying case from the analysis. Notice that the regression coefficients change dramatically when she does this. This is a reasonable strategy if -999 is a known data error or is supposed to be a missing data code. Possibly better would be to double-check the coding to make sure the wrong number wasn't entered. If, by some chance, -999 is a legitimate code, the researcher may want to look for additional explanatory variables, or try a technique like `qreg` or `rreg` that is designed to deal with outliers.

*II-2.*

```
. reg y x

      Source |       SS       df       MS              Number of obs =     240
-------------+------------------------------           F(  1,   238) =   14.98
       Model |  2611.95343     1  2611.95343           Prob > F      =  0.0001
    Residual |  41486.6545   238  174.313674           R-squared     =  0.0592
-------------+------------------------------           Adj R-squared =  0.0553
       Total |  44098.6079   239  184.513004           Root MSE      =  13.203

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.148551   .2967106     3.87   0.000     .5640361    1.733065
       _cons |  -.3421508   1.841041    -0.19   0.853    -3.968968    3.284666
------------------------------------------------------------------------------

. hettest

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

         chi2(1)      =    81.80
         Prob > chi2  =   0.0000
```

```
. reg y x, robust

Regression with robust standard errors                 Number of obs =     240
                                                       F(  1,   238) =   13.00
                                                       Prob > F      =  0.0004
                                                       R-squared     =  0.0592
                                                       Root MSE      =  13.203

------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.148551   .3185281     3.61   0.000      .521056    1.776045
       _cons |  -.3421508   1.224297    -0.28   0.780    -2.753993    2.069692
------------------------------------------------------------------------------
```

The researcher is concerned that errors are heteroscedastic, and the significant value for the Breusch-Pagan test suggests that this is indeed the case. To address the problem, she opts to use robust standard errors, which relax the assumption that errors are independent and identically distributed. This makes the standard errors more accurate but does not change the parameter estimates themselves. The use of robust standard errors has a very modest effect in this case, changing the estimates of the standard errors only slightly. Another alternative would be to try weighted least squares, if she thinks she knows enough to specify what the weights are. Transformations of the variables might also be warranted in some cases, e.g. take logs; or, the heteroscedasticity might go away if additional variables were added to the model or if subgroups were analyzed separately.

*II-3.*

```
. reg  activism ses liberalism black  white

      Source |       SS       df       MS              Number of obs =     400
-------------+------------------------------           F(  4,   395) =  589.72
       Model |  27155.3953     4  6788.84883           Prob > F      =  0.0000
    Residual |  4547.19637   395  11.5118895           R-squared     =  0.8566
-------------+------------------------------           Adj R-squared =  0.8551
       Total |  31702.5917   399   79.455117           Root MSE      =  3.3929

------------------------------------------------------------------------------
    activism |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ses |   1.756392   .0488747    35.94   0.000     1.660305    1.852479
  liberalism |   .6095345   .0363666    16.76   0.000     .5380382    .6810308
       black |   1.420534   .6593005     2.15   0.032     .1243567     2.71671
       white |   5.159183    .569313     9.06   0.000     4.039921    6.278446
       _cons |  -7.413753   1.019549    -7.27   0.000    -9.418173   -5.409332
------------------------------------------------------------------------------

. sum

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    activism |       500       27.79    8.973491          5       48.3
         ses |       400       12.96    3.961393          2         21
  liberalism |       500       13.52    5.061703          1         21
       black |       500          .2    .4004006          0          1
       other |       500          .1    .3003005          0          1
-------------+--------------------------------------------------------
       white |       500          .7    .4587165          0          1
        race |       500         1.4    .6639893          1          3
```

```
. impute ses  liberalism black white, gen(xses)
 20.00% (100) observations imputed

. sum ses xses

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         ses |       400       12.96    3.961393          2         21
        xses |       500    13.04019    3.627564          2         21

. reg  activism xses liberalism black white

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  4,   495) =  330.18
       Model |   29226.966     4   7306.7415           Prob > F      =  0.0000
    Residual |  10954.2833   495  22.1298652           R-squared     =  0.7274
-------------+------------------------------           Adj R-squared =  0.7252
       Total |  40181.2493   499  80.5235456           Root MSE      =  4.7042

------------------------------------------------------------------------------
    activism |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        xses |   1.756392   .0677641    25.92   0.000     1.623251    1.889533
  liberalism |   .6381163     .04471    14.27   0.000     .5502715    .7259611
       black |   1.901355   .8433631     2.25   0.025      .244342    3.558368
       white |   5.155422   .7217568     7.14   0.000     3.737337    6.573507
       _cons |  -7.730077   1.342503    -5.76   0.000    -10.36778    -5.09237
------------------------------------------------------------------------------
```

The researcher was perhaps surprised that only 400 cases showed up in the regression analysis. A look at the summary statistics revealed that 100 cases were missing on ses. The researcher opted to replace those missing values with regression estimates. That is, by using the `impute` command, for the 400 cases where data was not missing, ses was regressed on liberalism, black and white. The resulting regression coefficients were then used to compute estimates of ses for the other 100 cases. The researcher apparently believes that the regression estimates are the "best guess" as to the values of the missing cases, but there are various problems with this strategy. The significance tests will not reflect the uncertainty that is created by using estimates rather than real values for some cases. The impute approach assumes that the 100 missing cases are typical; they may be missing precisely because they are not typical and a regression estimate may therefore be a bad estimate. Assuming you can't find out what the true values for the missing cases are, sticking with listwise deletion, or using more advanced techniques, like multiple imputation, may be better. You need to know why the data are missing to decide on the best strategy.

*II-4.*

```
. reg y x1 x2 x3

      Source |       SS       df       MS              Number of obs =     142
-------------+------------------------------           F(  3,   138) =    4.14
       Model |  2487.49068     3  829.163558           Prob > F      =  0.0076
    Residual |  27620.5083   138  200.148611           R-squared     =  0.0826
-------------+------------------------------           Adj R-squared =  0.0627
       Total |   30107.999   141  213.531908           Root MSE      =  14.147

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -1.264819   1.710862    -0.74   0.461    -4.647712    2.118074
          x2 |   .4646403   1.195707     0.39   0.698    -1.899635    2.828915
          x3 |   1.547936   1.196244     1.29   0.198     -.817401    3.913273
       _cons |   20.29428   1.376812    14.74   0.000      17.5719    23.01666
------------------------------------------------------------------------------

. test x1=x2=x3

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

       F(  2,   138) =    0.60
            Prob > F =    0.5504

. alpha x1 x2 x3, c i gen(xscale)

Test scale = mean(unstandardized items)

                                                            average
                              item-test     item-rest      inter-item
Item         |  Obs  Sign   correlation   correlation     covariance      alpha
-------------+-----------------------------------------------------------------
x1           |  142    +       0.9958        0.9907        26.42669      0.9816
x2           |  142    +       0.9898        0.9770        26.46523      0.9907
x3           |  142    +       0.9900        0.9773        26.26876      0.9906
-------------+-----------------------------------------------------------------
Test scale   |                                             26.38689      0.9917
-------------------------------------------------------------------------------

. reg y xscale

      Source |       SS       df       MS              Number of obs =     142
-------------+------------------------------           F(  1,   140) =   11.29
       Model |  2247.45567     1  2247.45567           Prob > F      =  0.0010
    Residual |  27860.5433   140  199.003881           R-squared     =  0.0746
-------------+------------------------------           Adj R-squared =  0.0680
       Total |   30107.999   141  213.531908           Root MSE      =  14.107

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      xscale |    .773987   .2303132     3.36   0.001     .3186454    1.229329
       _cons |   20.21559   1.370705    14.75   0.000     17.50563    22.92555
------------------------------------------------------------------------------
```

The researcher is concerned about multicollinearity. The Global F is significant but the individual T values are not. The researcher apparently thinks that it may be legitimate to combine the 3 Xs into a single scale. The `test` command shows that indeed, their effects do not significantly differ from each other (hence they can be added together) and the `alpha` command further shows that, if added together, they would form a highly reliable scale. Once she creates the scale, the problem of multicollinearity obviously goes away. Incidentally, note that, although $R^2$ declines when she does this, adjusted $R^2$ goes up, which seems to further validate her decision. Even though this seems to work, simply adding the items together could be a questionable choice if they are measured in very different ways. Other options could include dropping some of the variables (but this could lead to specification error) or just using an incremental F test for all 3 X coefficients together rather than separate T tests.

*II-5.*

```
. reg  activism anomia1 anomia2 anomia3 anomia4 anomia5

      Source |       SS       df       MS              Number of obs =    3975
-------------+------------------------------           F(  5,  3969) =   29.11
       Model |   14532.664     5   2906.5328           Prob > F      =  0.0000
    Residual |  396311.025  3969  99.8516062           R-squared     =  0.0354
-------------+------------------------------           Adj R-squared =  0.0342
       Total |  410843.689  3974  103.382911           Root MSE      =  9.9926

------------------------------------------------------------------------------
    activism |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     anomia1 |   4.374225   3.002524     1.46   0.145    -1.512409    10.26086
     anomia2 |  -1.658814   1.585215    -1.05   0.295    -4.766726    1.449098
     anomia3 |  -2.497017   1.585327    -1.58   0.115    -5.605148    .6111143
     anomia4 |   1.868915   1.129336     1.65   0.098    -.3452169    4.083047
     anomia5 |   1.932719    1.58515     1.22   0.223    -1.175066    5.040504
       _cons |  -.0062429   .1910416    -0.03   0.974    -.3807918    .3683061
------------------------------------------------------------------------------

. corr

(obs=3975)

             |  anomia1  anomia2  anomia3  anomia4  anomia5 activism
-------------+------------------------------------------------------
     anomia1 |   1.0000
     anomia2 |   0.9776   1.0000
     anomia3 |   0.9775   0.9556   1.0000
     anomia4 |   0.9571   0.9353   0.9345   1.0000
     anomia5 |   0.9774   0.9555   0.9554   0.9354   1.0000
    activism |   0.1829   0.1753   0.1735   0.1827   0.1828   1.0000

. sw reg  activism anomia1 anomia2 anomia3 anomia4 anomia5, pe(.05)

                      begin with empty model
p = 0.0000 <  0.0500  adding   anomia1

      Source |       SS       df       MS              Number of obs =    3975
-------------+------------------------------           F(  1,  3973) =  137.51
       Model |  13744.2243     1  13744.2243           Prob > F      =  0.0000
    Residual |  397099.465  3973  99.9495254           R-squared     =  0.0335
-------------+------------------------------           Adj R-squared =  0.0332
       Total |  410843.689  3974  103.382911           Root MSE      =  9.9975

------------------------------------------------------------------------------
    activism |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     anomia1 |   4.014533    .342346    11.73   0.000     3.343342    4.685723
       _cons |  -.0045298   .1911315    -0.02   0.981    -.3792549    .3701952
------------------------------------------------------------------------------
```

Once again, the researcher is concerned about multicollinearity. Again, the Global F is significant and the individual T values are not; further the correlation matrix shows that the anomia measures are very highly correlated with each other. However, rather than use the brain that God gave him to try to find a solution for the problem, he uses mindless atheoretical stepwise regression to pick the variables. Stepwise regression chooses the variables that will most increase $R^2$, i.e. it just uses blind empiricism to select the variables for the model. This is potentially problematic. Note that anomia1 just barely edges out the other measures; a slightly different sample could easily produce different results. This could also result in specification error if one or more of the other variables is supposed to be in the model. If it is believed that the items all measure exactly the same thing and are more or less interchangeable, then it might make more sense for the researcher to choose whichever item he feels has the most face validity. Constructing a scale from the anomia items might be another good choice.

## *III. Computation and interpretation.* (35 points total)

President Bush realizes that he faces a tough battle in getting the American people to back his plan for reforming Social Security. He wants to know what factors currently affect support for his plan. His pollsters have collected data from 3000 individuals on the following variables:

| Variable | Description |
|---|---|
| ssplan | Support for Bush's plan, measured on a scale that ranges from 0 (strongly opposes plan) to 100 (strongly supports plan). |
| bush | Coded 1 if the respondent voted for Bush in 2004, 0 otherwise |
| female | Coded 1 if the respondent is female, 0 otherwise |
| age | Age in years |

The study obtains the following results.

```
. corr, means

(obs=3000)

    Variable |         Mean    Std. Dev.          Min          Max
-------------+----------------------------------------------------
      ssplan |     54.95611     18.35886     5.713294     97.19326
        bush |     .5166667     .4998055            0            1
      female |     .5333333     .4989708            0            1
         age |        47.15     14.16359     13.08336     82.43011


             |   ssplan     bush   female      age
-------------+------------------------------------
      ssplan |   1.0000
        bush |   0.7023   1.0000
      female |  -0.1564  -0.1025   1.0000
         age |  -0.1319   0.1359   0.0814   1.0000
```

```
. reg  ssplan bush female age, beta

      Source |       SS       df       MS              Number of obs =    3000
-------------+------------------------------           F(  3,  2996) = 1219.73
       Model |  555765.846     3  185255.282           Prob > F      =  0.0000
    Residual |  455039.916  2996  151.882482           R-squared     =  [1]
-------------+------------------------------
       Total |  1010805.76  2999  337.047603           Root MSE      =  12.324

------------------------------------------------------------------------------
      ssplan |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        bush |   26.68239   .4575126    58.32   0.000                  .726407
      female |  -2.338038   .4555404     [2]    0.000                -.0635449
         age |   -.292231   .0161132   -18.14   0.000                    [3]
       _cons |   56.19586   .8172954    68.76   0.000                        .
------------------------------------------------------------------------------

. vif

    Variable |       VIF       1/VIF
-------------+----------------------
        bush |       [4]    0.968552
         age |      1.03    0.972347
      female |      1.02    0.980228
-------------+----------------------
    Mean VIF |      1.03

. test bush

 ( 1)  bush = 0

       F(  1,  2996) = [5]
            Prob > F =    0.0000

. test bush = female

 ( 1)  bush - female = 0

       F(  1,  2996) = 2283.01
            Prob > F =    0.0000

. test female age

 ( 1)  female = 0
 ( 2)  age = 0

       F(  2,  2996) =  188.40
            Prob > F =    0.0000
```

```
. pcorr2  ssplan bush female age

(obs=3000)

Partial and Semipartial correlations of ssplan with

    Variable |     Partial      SemiP    Partial^2     SemiP^2       Sig.
-------------+-----------------------------------------------------------
        bush |      0.7292     0.7149      0.5317      0.5111       0.000
      female |     -0.0934    -0.0629      0.0087      0.0040       0.000
         age |     -0.3145    -0.2223      0.0989      0.0494       0.000

. hettest  age

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: age

         chi2(1)      =     0.75
         Prob > chi2  =   0.3872
```

a) (10 pts) Fill in the missing quantities [1] – [5].

Here are the uncensored parts of the printout:

```
. reg  ssplan bush female age, beta

      Source |       SS       df       MS              Number of obs =    3000
-------------+------------------------------           F(  3,  2996) = 1219.73
       Model |  555765.846     3  185255.282           Prob > F      =  0.0000
    Residual |  455039.916  2996  151.882482           R-squared     =  0.5498
-------------+------------------------------           Adj R-squared =  0.5494
       Total |  1010805.76  2999  337.047603           Root MSE      =  12.324

------------------------------------------------------------------------------
      ssplan |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        bush |   26.68239   .4575126    58.32   0.000                  .726407
      female |  -2.338038   .4555404    -5.13   0.000                -.0635449
         age |   -.292231   .0161132   -18.14   0.000                -.2254519
       _cons |   56.19586   .8172954    68.76   0.000                        .
------------------------------------------------------------------------------
```

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
        bush |      1.03    0.968552
         age |      1.03    0.972347
      female |      1.02    0.980228
-------------+----------------------
    Mean VIF |      1.03

. test bush

 ( 1)  bush = 0

       F(  1,  2996) = 3401.29
            Prob > F =    0.0000
```

To confirm that Stata got it right:

[1] = $R^2$ = SSR/SST = 555765.846/1010805.76 = .5498

[2] = $t_{female} = b_{female}/ s_{b\text{-}female}$ = -2.338038/ .4555404 = -5.13

[3] = $b'_{age} = b_{age} * s_{age} / s_{ssplan}$ = -.292231 * 14.16359 / 18.35886 = -.2254519

[4] = $vif_{bush} = 1/Tol_{bush}$ = 1/ 0.968552 = 1.03

[5] = Wald test of ($H_0$: $\beta_{bush} = 0$) = $T_{bush}^2 = 58.32^2$ = 3401.22

b) (20 points) Interpret the results. Be sure to answer the following questions, explaining how the printout supports your conclusions.

1. What percentage of the sample is female? What percentage voted for Bush?

As you can see from the means in the descriptive statistics, 53.33% of the sample is female (i.e. 1600 people), while 51.67% voted for Bush (i.e. 1550 people).

2. Who was more likely to have voted for Bush – men or women? Younger people or older people?

The correlation matrix shows that female is negatively correlated with bush, which means that women were less likely to vote for Bush. On the other hand, age is positively correlated with Bush, which means that older people were more likely to vote for him than were younger people.

3. Which variable has the strongest impact on support for Bush's plan? Cite at least two pieces of evidence from the printout to support your conclusion on this point.

The variable bush (voted for Bush) has the largest T value, the largest standardized beta, and the largest partial and semipartial correlations. Hence, voting for Bush seems to be the strongest determinant of supporting Bush's social security plan. Of course, the direction of the causality may be debateable here; perhaps people voted for Bush at least partly because they supported his social security plan.

4. According to the model, which types of individuals are most likely to support Bush's plan?

The signs of the regression coefficients imply that people who voted for Bush, men, and younger individuals are more likely to support his social security plan.

5. Bush's statistical advisors were worried that there would be heteroscedasticity associated with age, i.e. the older the respondents were, the more variability there would be in their responses. Were these fears warranted?

The chi-square value of the `hettest` for age is insignificant, suggesting that this fear was not borne out by the data.

c) (5 points) The advisor preparing the report for Bush is very annoyed with his assistant who did the computer runs. He specifically told her that he wanted to be able to do an incremental F test of the hypothesis that neither age nor gender affected support for Bush's plan; but since only one model was estimated, he says he cannot do that. Explain why you either agree or disagree with him; if you disagree, give him the information he wants.

Perhaps the advisor is so used to doing things with SPSS that he does not realize that Stata has an equivalent way of doing things. The command `test female age` does a Wald test of the hypothesis that neither age nor gender affects support for Bush's plan; the highly significant F value indicates that this hypothesis should be rejected, i.e. one or the other or both affects support. In the case of OLS regression, the Wald test and the incremental F test will yield identical results (although this need not be true for other techniques like logistic regression.) To confirm this we will use the `hireg` command to compute the incremental F value for us and then repeat the Wald test presented earlier:

```
. hireg ssplan (bush) (female age), nomiss

[UNNECESSARY PARTS OF OUTPUT DELETED]…

Model  R2      F(df)              p         R2 change  F(df) change       p
   1:  0.493  2917.627(1,2998)    0.000
   2:  0.550  1219.728(3,2996)    0.000     0.057     188.401(2,2996)     0.000


. test female age

 ( 1)  female = 0
 ( 2)  age = 0

       F(  2,  2996) =  188.40
            Prob > F =    0.0000
```

As you see, both the Wald test and the incremental F test produce a highly significant F value of 188.40.