Sociology 63993
Exam 1
February 22, 2007

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

1. An outlier on Y will have the most effect on the regression line when its value for X is equal to the mean of X.

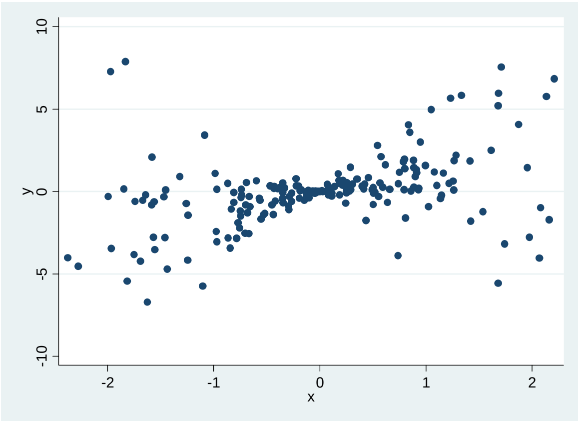
2. Serial correlation will not affect the unbiasedness or consistency of OLS estimators, but it does affect their efficiency.

3. Increasing the number of items in a scale will always increase the value of Cronbach's Alpha.

4. Religion has four categories: Catholic, Protestant, Jewish and Other. The researcher wants to use Catholic as her reference category. Therefore, she could compute 3 dummy variables: Protestant, Jewish and Other. On each of these variables, Catholics should be coded as missing.

5. A researcher conducts the following analysis:

```
. scatter y x
```




```
. quietly reg y x

. hettest

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

         chi2(1)      =     0.18
         Prob > chi2  =   0.6721
```

Based on the above, she can conclude that heteroskedasticity is probably not a problem with her data.

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

II-1.

```
. reg y x1 x2 x3

      Source |       SS       df       MS              Number of obs =    3975
-------------+------------------------------           F(  3,  3971) =   24.14
       Model |  7240.36972     3  2413.45657           Prob > F      =  0.0000
    Residual |  397049.865  3971  99.9873746           R-squared     =  0.0179
-------------+------------------------------           Adj R-squared =  0.0172
       Total |  404290.234  3974  101.733828           Root MSE      =  9.9994

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .4211835   .8683167     0.49   0.628    -1.281205    2.123572
          x2 |   .8435702   .8633128     0.98   0.329    -.8490077    2.536148
          x3 |   1.025274   .7932184     1.29   0.196    -.5298791    2.580428
       _cons |   .2881428   .2493349     1.16   0.248    -.2006937    .7769793
------------------------------------------------------------------------------

. alpha x1 x2 x3, i gen(xscale)

Test scale = mean(unstandardized items)

                                                            average
                              item-test     item-rest     inter-item
Item         |  Obs  Sign    correlation   correlation     covariance      alpha
-------------+-----------------------------------------------------------------
x1           | 3975    +       0.8006        0.6392         .3240937      0.8219
x2           | 3975    +       0.8079        0.6466         .3153546      0.8135
x3           | 3975    +       0.9928        0.9702         .0760545      0.4777
-------------+-----------------------------------------------------------------
Test scale   |                                              .2385009      0.8263
-------------------------------------------------------------------------------

. reg y xscale

      Source |       SS       df       MS              Number of obs =    3975
-------------+------------------------------           F(  1,  3973) =   71.79
       Model |  7175.70039     1  7175.70039           Prob > F      =  0.0000
    Residual |  397114.534  3973  99.9533184           R-squared     =  0.0177
-------------+------------------------------           Adj R-squared =  0.0175
       Total |  404290.234  3974  101.733828           Root MSE      =  9.9977

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      xscale |   2.501135   .2951914     8.47   0.000     1.922395    3.079876
       _cons |   .3140862   .2443089     1.29   0.199    -.1648963    .7930686
------------------------------------------------------------------------------
```

II-2.

```
. reg y2 x11

      Source |       SS       df       MS              Number of obs =    3975
-------------+------------------------------           F(  1,  3973) =   11.39
       Model |  21109.1326     1  21109.1326           Prob > F      =  0.0007
    Residual |  7361172.91  3973  1852.79963           R-squared     =  0.0029
-------------+------------------------------           Adj R-squared =  0.0026
       Total |  7382282.04  3974   1857.6452           Root MSE      =  43.044

------------------------------------------------------------------------------
          y2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         x11 |   2.790641   .8267667     3.38   0.001     1.169714    4.411567
       _cons |  -.0881977    1.03719    -0.09   0.932    -2.121672    1.945276
------------------------------------------------------------------------------

. dfbeta
                           DFx11:  DFbeta(x11)

. extremes DFx11 y2 x11

  +---------------------------------------------+
  |  obs:       DFx11          y2          x11 |
  |---------------------------------------------|
  | 2100.   -.0235065   -25.31333    2.568708 |
  |  619.   -.0191516     33.0916   -.3147684 |
  | 3124.   -.0170745   -23.67617    2.227267 |
  | 3828.   -.0157568   -15.25209    2.531287 |
  | 3739.   -.0153048   -20.67157    2.223721 |
  +---------------------------------------------+

  +-------------------------------------------+
  | 2008.   .0140591    35.20511     2.00463 |
  | 3950.   .0141953     -30.203   -.1235355 |
  |  906.   .0145061   -26.45573   -.3324702 |
  | 2546.   .0156778   -29.41599    -.286212 |
  |   10.   6.147187    2643.918     2.15582 |
  +-------------------------------------------+

. replace y2 = y2/100 in 10
(1 real change made)

. reg y2 x11

      Source |       SS       df       MS              Number of obs =    3975
-------------+------------------------------           F(  1,  3973) =   71.23
       Model |  7120.96787     1  7120.96787           Prob > F      =  0.0000
    Residual |  397169.267  3973  99.9670945           R-squared     =  0.0176
-------------+------------------------------           Adj R-squared =  0.0174
       Total |  404290.234  3974  101.733828           Root MSE      =  9.9984

------------------------------------------------------------------------------
          y2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         x11 |   1.620833   .1920425     8.44   0.000     1.244322    1.997344
       _cons |   .3580865   .2409198     1.49   0.137    -.1142516    .8304245
------------------------------------------------------------------------------
```

II-3.

```
. reg  income skills scale01

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  2,     7) =    9.55
       Model |    444.0924     2    222.0462           Prob > F      =  0.0100
    Residual |  162.827596     7  23.2610852           R-squared     =  0.7317
-------------+------------------------------           Adj R-squared =  0.6551
       Total |  606.919997     9  67.4355552           Root MSE      =   4.823

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      skills |   .6777237   .3230182     2.10   0.074     -.086093     1.44154
     scale01 |   3.271439   .8299957     3.94   0.006     1.308811    5.234067
       _cons |   -8.71115   6.793964    -1.28   0.241    -24.77632    7.354021
------------------------------------------------------------------------------

. list

     +---------------------------------------------------+
     | income   black   skills    scale01    scale02 |
     |---------------------------------------------------|
  1. |      5   black        9   1.741101          . |
  2. |    9.7   black       18   3.031433          . |
  3. |   28.4   black       21   5.278032          . |
  4. |    8.8   black       12   5.278032          . |
  5. |     21   black       14   5.278032          . |
     |---------------------------------------------------|
  6. |   26.6   black       16   6.309574          . |
  7. |   25.4   black       16   7.300372          . |
  8. |   23.1   black        9   7.300372          . |
  9. |   22.5   black       18   7.300372          . |
 10. |   19.5   black        5   7.300372          . |
     |---------------------------------------------------|
 11. |   21.7   white        7          .   7.300372 |
 12. |   24.8   white        9          .   7.783137 |
 13. |   30.1   white       12          .   8.258524 |
 14. |   24.8   white       17          .   8.258524 |
 15. |   28.5   white       19          .   8.727161 |
     |---------------------------------------------------|
 16. |     26   white        6          .   8.727161 |
 17. |   38.9   white       17          .   9.189587 |
 18. |   22.1   white        1          .   9.189587 |
 19. |   33.1   white       10          .   9.646264 |
 20. |   48.3   white       17          .   11.42288 |
     +---------------------------------------------------+

. sum

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      income |        20      24.415    9.788354          5       48.3
       black |        20          .5    .5129892          0          1
      skills |        20       12.65    5.460625          1         21
     scale01 |        10    5.611769     1.93984   1.741101   7.300372
     scale02 |        10    8.850319    1.142792   7.300372   11.42288

. gen xscale01 =  scale01
(10 missing values generated)

. replace  xscale01 = 5.611769 if missing(scale01)
(10 real changes made)

. gen md = 0

. replace md = 1 if missing(scale01)
(10 real changes made)
```

```
. reg  income skills xscale01 md

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  3,    16) =   11.45
       Model |  1242.10592     3  414.035306           Prob > F      =  0.0003
    Residual |  578.319556    16  36.1449723           R-squared     =  0.6823
-------------+------------------------------           Adj R-squared =  0.6228
       Total |  1820.42548    19  95.8118671           Root MSE      =  6.0121

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      skills |   .7629064    .258854     2.95   0.009     .2141605    1.311652
    xscale01 |   3.283386   1.033725     3.18   0.006     1.091988    5.474785
          md |   12.58468   2.753807     4.57   0.000     6.746875    18.42249
       _cons |  -9.953716   7.175127    -1.39   0.184     -25.1643    5.256874
------------------------------------------------------------------------------
```

III. Computation and interpretation. (35 points total)

Both Sociologists and Public Health researchers are interested in the determinants of self-reported health. The NHANES2F data, available from Stata's web site, includes information on the following.

Variable	Description
health	Self-reported health. Values range from 1 (poor health) to 5 (excellent health)
age	Age in years
female	Coded 1 if female, 0 otherwise
black	Coded 1 if black, 0 otherwise
rural	Coded 1 if respondent lives in a rural area, 0 otherwise

[NOTE: These data are weighted and proper analysis should take that into account. In addition, the dependent variable is ordinal and would probably be better analyzed by ordinal regression methods that we will talk about later. For simplicity, we will ignore such details for now.]

An analysis of the data yields the following results.

```
. webuse nhanes2f, clear

. keep  health age female black rural
```

```
. corr , means
(obs=10335)

    Variable |         Mean    Std. Dev.          Min          Max
-------------+----------------------------------------------------
      health |     3.413836     1.206196            1            5
         age |     47.56584     17.21752           20           74
      female |     .5250121     .4993982            0            1
       black |     .1050798     .3066711            0            1
       rural |     .3672956     .4820913            0            1

             |   health      age   female    black    rural
-------------+---------------------------------------------
      health |   1.0000
         age |  -0.3686   1.0000
      female |  -0.0320   0.0090   1.0000
       black |  -0.1286  -0.0321   0.0100   1.0000
       rural |  -0.0827   0.0565  -0.0341  -0.1838   1.0000


. alpha  female black rural, i gen(demscale)

Test scale = mean(unstandardized items)

                                                            average
                              item-test     item-rest     inter-item
Item         |  Obs  Sign   correlation   correlation     covariance      alpha
-------------+-----------------------------------------------------------------
female       | 10337   +       0.6438        0.0316        .0271779      0.2855
black        | 10337   +       0.4966        0.1326        .0082063      0.0659
rural        | 10337   -       0.6892        0.1247        .0015225      0.0176
-------------+-----------------------------------------------------------------
Test scale   |                                             .0123022      0.1704
-------------------------------------------------------------------------------

. drop demscale

. pcorr2  health age female black rural

(obs=10335)

Partial and Semipartial correlations of health with

    Variable |     Partial      SemiP    Partial^2     SemiP^2       Sig.
-------------+-------------------------------------------------------------
         age |     -0.3730     -0.3675      0.1391      0.1350      0.000
      female |     -0.0331     -0.0303      0.0011      0.0009      0.001
       black |     -0.1664     -0.1542      0.0277      0.0238      0.000
       rural |     -0.0981     -0.0901      0.0096      0.0081      0.000
```

```
. reg  health age female black rural

      Source |       SS       df       MS              Number of obs =   10335
-------------+------------------------------           F(  4, 10330) =  508.24
       Model |  2472.36904 [ 1 ]  618.092259           Prob > F      =  0.0000
    Residual |  12562.6523 10330  1.21613285           R-squared     =  [  2 ]
-------------+------------------------------           Adj R-squared =  0.1641
       Total |  15035.0214 10334   [  3    ]           Root MSE      =  1.1028

------------------------------------------------------------------------------
      health |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0257926   .0006313   [  4 ]   0.000      -.02703   -.0245552
      female |  -.0731412   .0217366    -3.36   0.001    -.1157492   -.0305331
       black |  -.6173391   .0359965   -17.15   0.000    -.6878992   -.5467791
       rural |  -.2296753   .0229357   -10.01   0.000    -.2746336   -.1847169
       _cons |    4.82831   .0349816   138.02   0.000     4.759739     4.89688
------------------------------------------------------------------------------

. test rural

 ( 1)  rural = 0

       F(  1, 10330) =  [5]
            Prob > F =    0.0000
```

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

b) (25 points) Answer the following questions about the analysis and the results, explaining how the printout supports your conclusions.

1. Briefly interpret the results, i.e. explain how race, gender, age and geographic location affect self-reported health. What types of people report the highest levels of health and which types of people report the lowest?

2. Suppose that, 20 years from now, your friend decides to finally move away from the big city and live in a small farm out in the country. His mother thinks it is a terrible idea but at least she manages to talk him out of having a sex change operation too. According to the above model, how much higher/lower can your friend expect his health score to be then than it is now?

3. The researcher created a variable called demscale but then immediately deleted it. Why did he do this?

4. Suppose the researcher now ran backwards stepwise regression using the .05 level of significance, i.e. gave the command

```
. sw, pr(.05): reg  health age female black rural
```

How would the results differ from the regression reported above?

5. Suppose that in previous studies it has been found that, after controlling for other variables, on average women score a tenth of a point lower on health than do men. The researcher therefore decides to test

$$H_0: \beta_{female} = -.10$$
$$H_A: \beta_{female} \neq -.10$$

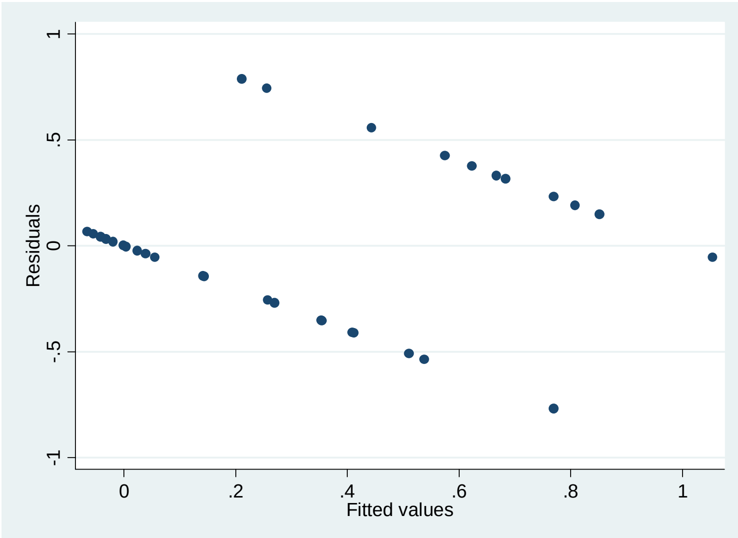
Based on the results presented above and using the .05 level of significance, should the researcher reject or not reject the null hypothesis?

IV. Extra Credit. (Up to 10 points.)

Y is coded 0 if failed, 1 if succeeded. X1, X2 and X3 are explanatory variables. An analysis of these data reveals the following:

```
. reg y x1 x2 x3

      Source |       SS       df       MS              Number of obs =      32
-------------+------------------------------           F(  3,    28) =    7.26
       Model |  3.1589065     3  1.05296883            Prob > F      =  0.0009
    Residual |  4.0598435    28  .144994411            R-squared     =  0.4376
-------------+------------------------------           Adj R-squared =  0.3773
       Total |    7.21875    31  .232862903            Root MSE      =  .38078

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .0348776    .011257     3.10   0.004     .0118187    .0579365
          x2 |   .0088082   .0191632     0.46   0.649    -.0304459    .0480622
          x3 |   .3690164   .1365854     2.70   0.012      .089234    .6487988
       _cons |  -.6336808   .3923563    -1.62   0.118    -1.437386    .1700247
------------------------------------------------------------------------------

. predict yhat
(option xb assumed; fitted values)

. predict e, resid

. extremes yhat e y

  +--------------------------------+
  | obs:        yhat          e   y |
  |--------------------------------|
  |   3.   -.0662836   .0662836   0 |
  |   6.   -.0550311   .0550311   0 |
  |   5.   -.0425547   .0425547   0 |
  |   8.   -.0328912   .0328912   0 |
  |   4.   -.0198185   .0198185   0 |
  +--------------------------------+

  +--------------------------------+
  |  25.   .7690787    .2309213   1 |
  |  27.    .769383    -.769383   0 |
  |  26.   .8080336    .1919664   1 |
  |  29.    .851562     .148438   1 |
  |  31.   1.054008   -.0540077   1 |
  +--------------------------------+
```

```
. rvfplot
```



a) Why does the plot of residuals versus fitted values (i.e. yhat versus e) look the way it does? [HINT: Any OLS regression using a binary dependent variable is going to look more or less like the above. Think about why that has to be.]

b) Does anything in the above analysis suggest any problems with the use of OLS regression with binary dependent variables, e.g. are any OLS assumptions being violated, are the estimates sensible? [HINT: the yhat values can be interpreted as the predicted probability of success given the values of the x variables, e.g. a yhat value of .73 implies a 73% chance of succes.]