Sociology 63993
Exam 1 Answer Key
February 15, 2008

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 technique has been totally discredited and should not be used under any circumstances.

False. The method can be useful when the missing value simply simply does not exist, e.g. the question is on spouse's attitudes and there is no spouse.

2. There is an inherent downward bias in the R^2 statistic, i.e. $E(R^2) < \rho^2$.

False. There is an inherent upward bias. Because of sampling variability, even effects that are truly zero will be estimated as non-zero and cause R^2 to increase.

3. A researcher runs the following analysis:

```
. alpha v1 v2 v3, i

Test scale = mean(unstandardized items)

                                                    average
                          item-test     item-rest     inter-item
Item          |  Obs  Sign   correlation   correlation     covariance      alpha
--------------+-----------------------------------------------------------------
v1            | 3975    +       0.7493        0.5546        .2940328      0.8210
v2            | 3975    +       0.7853        0.5922        .2614789      0.7834
v3            | 3975    +       0.9918        0.9660        .0459916      0.3323
--------------+-----------------------------------------------------------------
Test scale    |                                           .2005011      0.7977
-------------------------------------------------------------------------------
```

Based on these results, she should drop v3 from her scale.

False. That would be the worst thing to do, since the scale's reliability would drop to .3323. If anything, drop v1, as that will make the scale slightly more reliable.

4. Robust standard errors are one means for dealing with the problem of multicollinearity.

False. Robust standard errors are a way of dealing with errors that are not iid.

5. A researcher has collected earnings data on a firm for each of the past 60 months. When she computes the Durbin-Watson statistic, she gets a value of 2.0. This indicates that first-order serial correlation is a problem in her data.

False. A value of 2.0 indicates that there is no first order 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.

```
. reg  psyscore workatt qscale01

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  2,     7) =    9.53
       Model |  1775.55796     2  887.778982           Prob > F      =  0.0100
    Residual |  652.122126     7  93.1603037           R-squared     =  0.7314
-------------+------------------------------           Adj R-squared =  0.6546
       Total |  2427.68009     9  269.742232           Root MSE      =   9.652

------------------------------------------------------------------------------
    psyscore |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     workatt |   1.414823   .6474377     2.19   0.065    -.1161239     2.94577
    qscale01 |    3.57697   .9083323     3.94   0.006     1.429106    5.724835
       _cons |  -43.93438    8.64232    -5.08   0.001    -64.37022   -23.49854
------------------------------------------------------------------------------

. list

     +---------------------------------------------------+
     |  psyscore   female   workatt   qscale01   qscale02 |
     |---------------------------------------------------|
  1. |    -38.83   Female     -3.65          2          . |
  2. |    -29.43   Female      5.35          4          . |
  3. |  7.969999   Female      8.35          8          . |
  4. |    -31.23   Female      -.65          8          . |
  5. |     -6.83   Female      1.35          8          . |
     |---------------------------------------------------|
  6. |  4.370001   Female      3.35         10          . |
  7. |  1.969999   Female      3.35         12          . |
  8. | -2.629999   Female     -3.65         12          . |
  9. |     -3.83   Female      5.35         12          . |
 10. |     -9.83   Female     -7.65         12          . |
     |---------------------------------------------------|
 11. | -5.429998     Male     -5.65          .         12 |
 12. |  .7699985     Male     -3.65          .         13 |
 13. |     11.37     Male      -.65          .         14 |
 14. |  .7699985     Male      4.35          .         14 |
 15. |      8.17     Male      6.35          .         15 |
     |---------------------------------------------------|
 16. |      3.17     Male     -6.65          .         15 |
 17. |     28.97     Male      4.35          .         16 |
 18. | -4.629999     Male    -11.65          .         16 |
 19. |     17.37     Male     -2.65          .         17 |
 20. |     47.77     Male      4.35          .         21 |
     +---------------------------------------------------+

. gen qscale =  qscale01
(10 missing values generated)

. replace  qscale = qscale02 if missing(qscale)
(10 real changes made)
```

```
. reg  psyscore workatt qscale

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  2,    17) =   46.33
       Model |  6152.90086     2  3076.45043           Prob > F      =  0.0000
    Residual |    1128.801    17  66.4000591           R-squared     =  0.8450
-------------+------------------------------           Adj R-squared =  0.8267
       Total |  7281.70187    19  383.247467           Root MSE      =  8.1486

------------------------------------------------------------------------------
    psyscore |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     workatt |    1.298731   .3443179     3.77   0.002     .5722835    2.025178
      qscale |    3.866786   .4198989     9.21   0.000     2.980876    4.752695
       _cons |   -46.59477   5.377861    -8.66   0.000    -57.94106   -35.24847
------------------------------------------------------------------------------
```

The researcher is worried about missing data. Half his cases are missing in the original regression. After listing the values, he realizes that only females have scores on qscale01, and only males have scores on qscale02. He therefore decides to combine the items into a single scale. This may be a great strategy if the two scales really are the same questions but asked at different points in the questionnaire. Skip patterns might produce such a result. If the items aren't the same though, this could be a terrible strategy and it may be better just to stick with listwise deletion, keeping in mind that you would then only be analyzing females.

II-2.

```
. reg hscale age black female

      Source |       SS       df       MS              Number of obs =   10335
-------------+------------------------------           F(  3, 10331) =  533.44
       Model |  95636.8263     3  31878.9421           Prob > F      =  0.0000
    Residual |  617392.308 10331  59.7611372           R-squared     =  0.1341
-------------+------------------------------           Adj R-squared =  0.1339
       Total |  713029.135 10334   68.998368           Root MSE      =  7.7305

------------------------------------------------------------------------------
      hscale |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.1662599   .0044192   -37.62   0.000    -.1749225   -.1575973
       black |  -3.521926    .248113   -14.19   0.000    -4.008275   -3.035576
      female |  -.5391638   .1522892    -3.54   0.000    -.8376801   -.2406475
       _cons |   21.69226   .2388228    90.83   0.000     21.22412     22.1604
------------------------------------------------------------------------------

. predict rstandard, rstandard
(2 missing values generated)

. extremes rstandard hscale age black female

  +-------------------------------------------------+
  |  obs:   rstandard   hscale   age   black   female |
  |-------------------------------------------------|
  | 3446.   -2.225543        1    21       0        0 |
  | 6078.   -2.225543        1    21       0        0 |
  |  174.   -2.204013        1    22       0        0 |
  |  503.   -2.182483        1    23       0        0 |
  | 8122.    -2.15577        1    21       0        1 |
  +-------------------------------------------------+
```

```
  +-------------------------------------------------+
  | 7299.   2.433496       25    72       1        0 |
  | 8187.   2.460159       25    70       1        1 |
  |  110.   2.503239       25    72       1        1 |
  |  378.   2.546324       25    74       1        1 |
  |    8.   30.83233      250    57       0        1 |
  +-------------------------------------------------+
```

```
. qreg hscale age black female, nolog

Median regression                                   Number of obs =     10335
  Raw sum of deviations    71056 (about 9)
  Min sum of deviations 62710.56                   Pseudo R2     =    0.1174

------------------------------------------------------------------------------
      hscale |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.1794872   .0052734   -34.04   0.000     -.189824   -.1691503
       black |  -4.282051   .2978292   -14.38   0.000    -4.865854   -3.698248
      female |  -.3589744   .1820799    -1.97   0.049    -.7158863   -.0020625
       _cons |   21.02564   .2853329    73.69   0.000     20.46633    21.58495
------------------------------------------------------------------------------
```

Outliers are a concern. Case 8 has a score on hscale that is very extreme compared to other values. The researcher therefore decides to use median regression, which is less sensitive to the pull of outliers. If I were the researcher, though, I would check to see if an extra zero accidentally got added to case 8. I would also try out other methods, e.g. robust regression, or dropping the outlier, to see if it made much difference how the outlier was handled.

II-3.

```
. reg health height weight female

      Source |       SS       df       MS              Number of obs =   10335
-------------+------------------------------           F(  3, 10331) =  145.08
       Model |  1227409.22     3  409136.406           Prob > F      =  0.0000
    Residual |  29134953.1 10331   2820.1484           R-squared     =  0.0404
-------------+------------------------------           Adj R-squared =  0.0401
       Total |  30362362.4 10334  2938.10358           Root MSE      =  53.105

------------------------------------------------------------------------------
      health |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   1.541938   .0798774    19.30   0.000     1.385363    1.698513
      weight |  -.4657993   .0388145   -12.00   0.000    -.5418833   -.3897153
      female |   10.15199   1.465599     6.93   0.000     7.279136    13.02485
       _cons |  -174.9994   13.31385   -13.14   0.000    -201.0972   -148.9017
------------------------------------------------------------------------------

. hettest

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

         chi2(1)      =   114.87
         Prob > chi2  =   0.0000

. * Compute the natural log of health and use it instead
. gen lnhealth = ln(health)
(2 missing values generated)
```

```
. reg lnhealth height weight female

      Source |       SS       df       MS              Number of obs =   10335
-------------+------------------------------           F(  3, 10331) =  168.05
       Model |  699.558344     3  233.186115            Prob > F      =  0.0000
    Residual |  14335.4632 10331  1.38761623            R-squared     =  0.0465
-------------+------------------------------           Adj R-squared =  0.0463
       Total |  15035.0216 10334  1.45490822           Root MSE      =   1.178

------------------------------------------------------------------------------
    lnhealth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .0378505   .0017718    21.36   0.000     .0343773    .0413236
      weight |   -.010879    .000861   -12.64   0.000    -.0125667   -.0091913
      female |   .3079935   .0325098     9.47   0.000     .2442681     .371719
       _cons |  -2.311375   .2953265    -7.83   0.000    -2.890272   -1.732478
------------------------------------------------------------------------------

. hettest

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

         chi2(1)      =     1.02
         Prob > chi2  =   0.3134
```

The initial Breusch-pagan test indicates that heteroskedasticity is a problem with these data. The researcher decides to address the problem by transforming the dependent variable, i.e. he takes the log of it. After doing this, heteroskedasticity is no longer a problem. This is often a good approach, but the researcher needs to think about whether the variable transformation makes sense or not, i.e. from a theoretical standpoint, does it make more sense to use logged or non-logged health? We'll talk more later about the rationales behind different kinds of variable transformations.

III. Computation and interpretation. (35 points total)

Hillary Clinton's presidential campaign is reeling after a series of losses to Barack Obama. Clinton's new campaign manager, Maggie Williams, is confident that victory is still possible. But, she feels the campaign must better identify the issues that have the strongest impact on voters' opinions of Clinton and deal with them accordingly. She has therefore commissioned a survey of 5,000 <u>likely</u> voters in the upcoming primary states of Wisconsin, Texas, Ohio and Pennsylvania. All attitudinal items are measured on scales that range from 0 to 200. The variables include

Variable	Description
hillary	Attitudes toward Hillary. The higher the score, the more favorable the impression. This is the dependent variable in the analysls.
security	Attitudes toward national security. The higher the score, the more important strong national security is to the respondent.
healthcare	Attitudes towards national health care. The higher the score, the more important national health care is to the respondent.
economy	Attitudes toward the economy. The higher the score, the more important economic issues are to the respondent.
female	Coded 1 if the respondent is female, 0 if male.

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.]

```
. corr , means
(obs=5000)

    Variable |         Mean    Std. Dev.          Min          Max
-------------+--------------------------------------------------
     hillary |     166.3653       9.3227      136.797          200
    security |      70.4279      15.2894        30.84       159.44
  healthcare |     62.45453     21.84582     10.66667          196
     economy |     152.6836     15.24664           86          174
      female |        .4002     .4899877            0            1


             |  hillary security health~e  economy   female
-------------+---------------------------------------------
     hillary |   1.0000
    security |   0.4627   1.0000
  healthcare |   0.0741  -0.0707   1.0000
     economy |   0.0782   0.0813   0.1120   1.0000
      female |   0.6769   0.3592   0.0423   0.0608   1.0000


. pcorr2  hillary security healthcare economy female

(obs=5000)

Partial and Semipartial correlations of hillary with

    Variable |     Partial      SemiP    Partial^2     SemiP^2       Sig.
-------------+------------------------------------------------------------
    security |      0.3253      0.2388      0.1058      0.0570       0.000
  healthcare |      0.0937      0.0653      0.0088      0.0043       0.000
     economy |      0.0207      0.0143      0.0004      0.0002       0.144
      female |      0.6143      0.5403      0.3773      0.2919       0.000


. reg  hillary security healthcare economy female, beta

      Source |       SS       df       MS              Number of obs =    5000
-------------+------------------------------           F(  4,  4995) =    [1]
       Model |  225188.625     4  56297.1563           Prob > F      =  0.0000
    Residual |  209288.162  4995      [2]              R-squared     =  0.5183
-------------+------------------------------           Adj R-squared =  0.5179
       Total |  434476.787  4999    86.91274           Root MSE      =   6.473

------------------------------------------------------------------------------
     hillary |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
    security |   .1571209   .0064615    24.32   0.000                .2576811
  healthcare |       [3]    .0042413     6.65   0.000                .0660868
     economy |   .0088624   .0060693     1.46   0.144                .0144938
      female |   11.04764   .2008013      [4]   0.000                .5806478
       _cons |   147.7639   1.010788   146.19   0.000                       .
------------------------------------------------------------------------------
```

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
    security |      1.16    0.858783
      female |       [5]    0.865810
  healthcare |      1.02    0.976340
     economy |      1.02    0.978808
-------------+----------------------
    Mean VIF |      1.09

. test  security healthcare economy female

 ( 1)  security = 0
 ( 2)  healthcare = 0
 ( 3)  economy = 0
 ( 4)  female = 0

       F(  4,  4995) = 1343.62
            Prob > F =    0.0000

. test economy = healthcare

 ( 1) - healthcare + economy = 0

       F(  1,  4995) =    6.15
            Prob > F =    0.0132

. test female = 10

 ( 1)  female = 10

       F(  1,  4995) =   27.22
            Prob > F =    0.0000
```

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

First off, here are the uncensored parts of the printout:

```
. reg  hillary security healthcare economy female, beta

      Source |       SS       df       MS              Number of obs =    5000
-------------+------------------------------           F(  4,  4995) = 1343.62
       Model |  225188.625     4  56297.1563           Prob > F      =  0.0000
    Residual |  209288.162  4995  41.8995319           R-squared     =  0.5183
-------------+------------------------------           Adj R-squared =  0.5179
       Total |  434476.787  4999    86.91274           Root MSE      =   6.473

------------------------------------------------------------------------------
     hillary |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
    security |   .1571209   .0064615    24.32   0.000                .2576811
  healthcare |   .0282025   .0042413     6.65   0.000                .0660868
     economy |   .0088624   .0060693     1.46   0.144                .0144938
      female |   11.04764   .2008013    55.02   0.000                .5806478
       _cons |   147.7639   1.010788   146.19   0.000                       .
------------------------------------------------------------------------------
```

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
    security |      1.16    0.858783
      female |      1.15    0.865810
  healthcare |      1.02    0.976340
     economy |      1.02    0.978808
-------------+----------------------
    Mean VIF |      1.09
```

To confirm the results,

[1] F = 1343.62. The easiest way to solve this is to simply note that the first test command does the calculation for you. Other formulas also work.

[2] MSE = $(\text{Root MSE})^2 = 6.473^2 = 41.9$. Or, MSE = SSE/DFE = 209288.162/4995 = 41.9

[3] $b_{healthcare} = t_{healthcare} * se_{healthcare} = 6.65 * .0042413 = .0282$.

[4] $t_{female} = b_{female} / se_{female} = 11.04764/.2008013 = 55.02$

[5] $tol_{female} = 1/vif_{female} = 1/0.865810 = 1.15$

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

1. Based on these results, the Clinton campaign is very concerned about turnout by women voters, i.e. it is worried that not enough women are likely to vote. What is the basis for this concern?

The regressions indicated that women like Hillary, so she wants them to vote. But, the means show that only 40% of the likely voters are women. Perhaps women have become demoralized over the recent losses and hence are not planning to vote. Hillary needs to get those women voters to the polls.

2. If you were Clinton's campaign manager, what issue would you tell her to emphasize most, i.e. what issue is most important for people liking her? Cite several items from the printout that support your argument.

National security issues seem to have the strongest impact on how much people like her. Since the attitude items are all measured on 200 point scales, it is legitimate to compare the metric coefficients, and security has the largest effect. Of the different attitudes, it also has the largest t-value, the largest standardized coefficient, and the largest partial and semipartial correlations. If Hillary can successfully hammer away at national security issues and make people more concerned about them, it may drive her popularity ratings higher.

3. For months, Bill Clinton has been telling his wife's campaign staff that "It is the economy, stupid." He thinks Hillary should be paying far more attention to economic issues. He had to fight with the pollsters to include the economy questions, and even then they only got added at the end of the questionnaire when respondents were tired and rushing to get finished. Do you think the results support the former President's claims? If not, can you make an argument as to why he might be right anyway?

The effect of the economy variable is small and insignificant, which undercuts Bill Clinton's argument. However, if people were rushed when they answered the economy questions, the resulting scale may suffer from random measurement error, which could cause the effect of economic attitudes to be understated. The fact that the economy variable has the highest mean might also indicate that it is important to voters. Of course, if voters don't feel that Hillary can do much about the economy or that both candidates are equally good in this area, that could explain why the variable has so little effect. (Bill could also argue that if Hillary handled the issue differently the issue would have more effect, i.e. the coefficient would change and the economy would have more of an impact if the topic were handled better.)

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

```
. sw, pr(.05): reg  hillary security healthcare economy female
```

How would the results differ from the regression reported above? i.e. what variables, if any, would be dropped, and what would the new value of R^2 be?

economy would be dropped. It has the smallest squared semipartial and is not statistically significant. The squared semipartial for economy is .0002 so R^2 will go from .5183 down to .5181. To confirm,

```
. sw, pr(.05): reg  hillary security healthcare economy female
                    begin with full model
p = 0.1443 >= 0.0500  removing economy

      Source |       SS       df       MS              Number of obs =    5000
-------------+------------------------------           F(  3,  4996) = 1790.38
       Model |  225099.288     3  75033.0961           Prob > F      =  0.0000
    Residual |  209377.499  4996   41.909027           R-squared     =  0.5181
-------------+------------------------------           Adj R-squared =  0.5178
       Total |  434476.787  4999    86.91274           Root MSE      =  6.4737

------------------------------------------------------------------------------
     hillary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    security |   .1578254   .0064442    24.49   0.000      .145192    .1704587
  healthcare |   .0289232   .0042129     6.87   0.000      .020664    .0371824
      female |   11.05516   .2007579    55.07   0.000     10.66158    11.44873
       _cons |   149.0194   .5314737   280.39   0.000     147.9774    150.0613
------------------------------------------------------------------------------
```

5. An earlier and much larger survey found that the coefficient for female was 10. There is concern in the Clinton camp that her support among females has eroded since then. Clinton's researchers therefore decide to test

$$H_0: \beta_{female} = 10$$
$$H_A: \beta_{female} < 10$$

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

Do not reject the null! Remember, this is a one-tailed alternative, and the actual result (b_{female} = 11.04764) is in the opposite direction of what was predicted, i.e. the effect of female is actually greater than it used to be (although again, it may not do much good if

females don't go out and vote). The `test female = 10` statement above is misleading because it does a two-tailed test when a one-tailed test is called for.

Appendix: Stata Commands for Exam 1. Here are the commands I used to generate the Stata output on the exam. Alas, I haven't really conducted any new nationwide studies, but I have manipulated and disguised other data sets I have sitting around.

```
* Problem I-3.
use http://www.nd.edu/~rwilliam/xsoc63993/statafiles/anomia.dta, clear
clonevar v1 = anomia1
clonevar v2 = anomia9
corr2data e
gen v3 = v1 + v2 + e*.20
alpha v1 v2 v3, i


*** Problem II-1.
use http://www.nd.edu/~rwilliam/xsoc63993/statafiles/reg01.dta, clear
gen psyscore = (income * 2 - 48.83)
gen female = race
label define female 1 "Female" 0 "Male"
label values female female
gen workatt = jobexp - 12.65
gen qscale01 = educ if female
gen qscale02 = educ if !female
keep  psyscore female workatt qscale01 qscale02
reg  psyscore workatt qscale01
list
gen qscale =  qscale01
replace  qscale = qscale02 if missing(qscale)
reg  psyscore workatt qscale


*** Problem II-2.
webuse nhanes2f, clear
gen hscale = health ^2
replace hscale = 250 in 8
reg hscale age black female
predict rstandard, rstandard
extremes rstandard hscale age black female
qreg hscale age black female, nolog


*** Problem II-3.
webuse nhanes2f, clear
replace health = exp(health)
reg health height weight female
hettest
* Compute the natural log of health
gen lnhealth = ln(health)
reg lnhealth height weight female
hettest


*** Problem III.
webuse nhanes2f, clear
* Cleverly disguise the data!
gen hillary = height
gen security = weight
gen healthcare = iron
gen economy = zinc
recode female (0=1)(1=0)
replace healthcare = healthcare * 2/3
replace economy = 2 * economy
```

```
sort economy, stable
keep in 1/5000
order hillary security healthcare economy female
keep  hillary security healthcare economy female
corr, means
pcorr2  hillary security healthcare economy female
reg  hillary security healthcare economy female, beta
vif
test  security healthcare economy female
test economy = healthcare
test female = 10
sw, pr(.05): reg  hillary security healthcare economy female
```