Sociology 63993, Exam1
February 12, 2015

Richard Williams, University of Notre Dame, http://www3.nd.edu/~rwilliam/

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

1. When working with complex survey data (and using the svy: prefix in Stata) nested models should be tested via the use of incremental F tests (i.e. you should estimate the constrained and unconstrained models separately and then use an incremental F test to contrast them).

2. The closer the tolerance of a variable is to 1, the more likely it is that you will have problems with multicollinearity.

3. The most extreme outliers on Y (i.e. the cases where Y is furthest from the mean) will always have the most influence on the regression line.

4. Cohen and Cohen's Dummy Variable Adjustment technique has been discredited and should not be used under any circumstances.

5. A researcher runs the following analysis:

```
. alpha v1 v2 v3, i

Test scale = mean(unstandardized items)

                                                            average
                              item-test     item-rest       interitem
Item         |  Obs  Sign    correlation   correlation     covariance      alpha
-------------+-----------------------------------------------------------------
v1           | 3975    +       0.4842        0.1522         .2360952      0.7907
v2           | 3975    +       0.8448        0.6473         .0374987      0.1997
v3           | 3975    +       0.8836        0.5602         .0342703      0.2815
-------------+-----------------------------------------------------------------
Test scale   |                                              .1026214      0.6060
-------------------------------------------------------------------------------
```

Based on these results, she should drop v2 from her scale.

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.

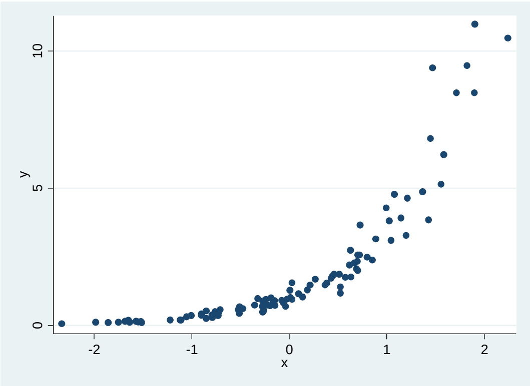
```
. reg health age weight height i.female i.race

      Source |       SS       df       MS              Number of obs =     800
-------------+------------------------------           F(  6,   793) =   28.89
       Model |  193.406808     6  32.234468           Prob > F      =  0.0000
    Residual |  884.811942   793  1.11577798           R-squared     =  0.1794
-------------+------------------------------           Adj R-squared =  0.1732
       Total |  1078.21875   799  1.34946026           Root MSE      =  1.0563

------------------------------------------------------------------------------
      health |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0211716   .0022829    -9.27   0.000    -.0256528   -.0166904
      weight |  -.0039733   .0028427    -1.40   0.163    -.0095535    .0016068
      height |   .0127136   .0058336     2.18   0.030     .0012625    .0241647
    1.female |   .1804978   .1031217     1.75   0.080    -.0219259    .3829216
             |
        race |
      Black  |  -.7038075   .0977277    -7.20   0.000     -.895643   -.5119719
      Other  |  -.0574042   .1235336    -0.46   0.642    -.2998958    .1850873
             |
       _cons |   2.645635   1.012356     2.61   0.009     .6584201    4.632851
------------------------------------------------------------------------------

. sum, sep(6)

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        race |       800       1.445    .7071776          1          3
         age |      1100    48.88364    17.46024         20         74
      height |      1100    167.2078    10.19798      138.5        200
      weight |      1100     71.0562    15.31384      30.84     149.69
      health |      1100    3.401818    1.172321          1          5
      female |      1100    .5218182     .499751          0          1

. tab1 race

-> tabulation of race

   1=white, |
   2=black, |
    3=other |      Freq.     Percent        Cum.
------------+-----------------------------------
      White |        545       68.13       68.13
      Black |        154       19.25       87.38
      Other |        101       12.63      100.00
------------+-----------------------------------
      Total |        800      100.00

. mi set mlong

. mi register imputed race
(300 m=0 obs. now marked as incomplete)
```

```
. mi impute mlogit race health age height weight female, add(50) rseed(2232)

Univariate imputation                        Imputations =       50
Multinomial logistic regression                    added =       50
Imputed: m=1 through m=50                        updated =        0

------------------------------------------------------------------
                   |               Observations per m
                   |----------------------------------------------
          Variable |   Complete   Incomplete   Imputed |     Total
-------------------+-----------------------------------+----------
              race |        800          300       300 |      1100
------------------------------------------------------------------
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)

. mi estimate: reg health age weight height i.female i.race

Multiple-imputation estimates                  Imputations     =         50
Linear regression                              Number of obs   =       1100
                                               Average RVI     =     0.1216
                                               Largest FMI     =     0.2769
                                               Complete DF     =       1093
DF adjustment:   Small sample                  DF:     min     =     357.50
                                                       avg     =     823.32
                                                       max     =    1052.35
Model F test:       Equal FMI                  F(   6, 1032.2) =      33.72
Within VCE type:          OLS                  Prob > F        =     0.0000

------------------------------------------------------------------------------
      health |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0200064   .0019955   -10.03   0.000    -.0239221   -.0160908
      weight |  -.0051971   .0024855    -2.09   0.037    -.0100745   -.0003198
      height |    .015661   .0050562     3.10   0.002     .0057391    .0255829
    1.female |   .2240806   .0902576     2.48   0.013     .0469683    .4011929
             |
        race |
          2  |  -.7150266     .09896    -7.23   0.000    -.9096434   -.5204098
          3  |  -.0666072   .1224521    -0.54   0.587    -.3074037    .1741892
             |
       _cons |   2.161208   .8808568     2.45   0.014     .4326168    3.889799
------------------------------------------------------------------------------
```

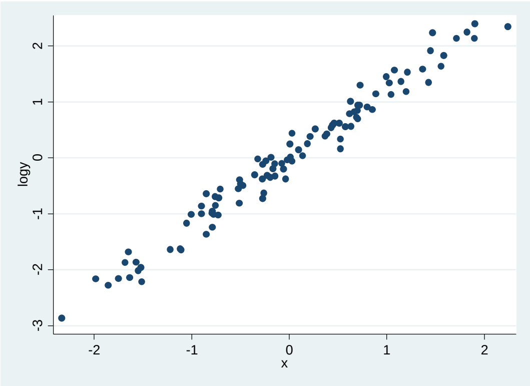
II-2.

```
. reg y x

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =  204.90
       Model |   387.15257     1   387.15257           Prob > F      =  0.0000
    Residual |   185.17243    98  1.88951459           R-squared     =  0.6765
-------------+------------------------------           Adj R-squared =  0.6732
       Total |     572.325    99  5.78106061           Root MSE      =  1.3746

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.977532   .1381521    14.31   0.000     1.703373     2.25169
       _cons |    1.94407   .1374596    14.14   0.000     1.671286    2.216854
------------------------------------------------------------------------------
```

```
. estat hettest

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

         chi2(1)      =    32.27
         Prob > chi2  =   0.0000

. twoway scatter y x, name(g1)
```




```
. gen logy = log(y)

. twoway scatter logy x, name(g2)
```




```
. reg logy x

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) = 3528.00
       Model |      142.56     1      142.56           Prob > F      =  0.0000
    Residual |  3.96000021    98  .040408165           R-squared     =  0.9730
-------------+------------------------------           Adj R-squared =  0.9727
       Total |  146.520001    99  1.48000001           Root MSE      =  .20102

------------------------------------------------------------------------------
        logy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |        1.2   .0202031    59.40   0.000     1.159908    1.240092
       _cons |   2.14e-09   .0201018     0.00   1.000    -.0398913    .0398913
------------------------------------------------------------------------------
```

```
. estat hettest

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

         chi2(1)      =     0.02
         Prob > chi2  =   0.8781
```

II-3.

```
. reg y x

      Source |       SS       df       MS              Number of obs =    3975
-------------+------------------------------           F(  1,  3973) =    0.04
       Model |  559.504598     1  559.504598           Prob > F      =  0.8437
    Residual |  57188892.5  3973  14394.3852           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared = -0.0002
       Total |  57189452.1  3974  14390.9039           Root MSE      =  119.98

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .3505943   1.778278     0.20   0.844    -3.135828    3.837017
       _cons |    3.50558   2.474016     1.42   0.157    -1.344881    8.356041
------------------------------------------------------------------------------

. dfbeta
                       _dfbeta_1: dfbeta(x)

. extremes _df* y x

  +-------------------------------------------+
  |  obs:   _dfbeta_1           y           x |
  |-------------------------------------------|
  | 2846.   -20.48698    7560.241    .1534038 |
  | 2100.   -.0031762   -6.815401    3.137415 |
  | 3828.   -.0019974   -2.850775    3.062574 |
  |   70.   -.0019538    -3.89073    2.776838 |
  | 3739.   -.0019023   -5.510675    2.447441 |
  +-------------------------------------------+

  +-------------------------------------------+
  | 2439.   .0025574   -8.791584   -.8340001 |
  | 2444.    .002686    -7.56208   -1.147358 |
  |  171.   .0027336   -11.03818   -.6557877 |
  | 1442.   .0027977   -7.366304   -1.281223 |
  | 2546.   .0028055   -12.22433   -.5724241 |
  +-------------------------------------------+

. drop  _df*

. replace y = y/1000 in 2846
(1 real change made)
```

```
. reg y x

      Source |       SS       df       MS              Number of obs =    3975
-------------+------------------------------           F(  1,  3973) =  693.88
       Model |  11237.3545     1  11237.3545           Prob > F      =  0.0000
    Residual |  64343.0002  3973  16.1950668           R-squared     =  0.1487
-------------+------------------------------           Adj R-squared =  0.1485
       Total |  75580.3548  3974  19.0187103           Root MSE      =  4.0243

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.571213   .0596478    26.34   0.000      1.45427    1.688156
       _cons |   .5203355   .0829846     6.27   0.000     .3576392    .6830318
------------------------------------------------------------------------------

. dfbeta
                       _dfbeta_1: dfbeta(x)

. extremes _df* y x

  +--------------------------------------------+
  |  obs:   _dfbeta_1           y           x  |
  |--------------------------------------------|
  | 2100.   -.1018125   -6.815401    3.137415  |
  |  619.   -.0739862    12.73301   -.6295367  |
  | 2606.   -.0678742     12.7148   -.5239393  |
  | 3828.    -.065618   -2.850775    3.062574  |
  | 2574.   -.0643436    10.71742   -.6628681  |
  +--------------------------------------------+

  +------------------------------------------+
  | 1111.   .0597705    13.48404    2.771187 |
  |  171.   .0599936   -11.03818   -.6557877 |
  | 2724.   .0613352    13.84974    2.727732 |
  |  665.   .0637711    13.70689    2.887728 |
  | 2546.   .0638679   -12.22433   -.5724241 |
  +------------------------------------------+
```

III. *Computation and interpretation.* (35 points total) The Center for Disease Control is very concerned about the anti-vaccination movement in the United States. According to the World Health Organization (http://www.who.int/mediacentre/factsheets/fs286/en/), measles is one of the leading causes of death among young children worldwide even though a safe and cost-effective vaccine is available. In the United States, the number of measles cases has skyrocketed in recent years, largely because growing numbers of parents are choosing not to vaccinate their children. Various explanations have been offered.

- Due to a now discredited study (http://www.newsweek.com/autism-how-childhood-vaccines-became-villains-82273) , some parents fear that the measles vaccine can cause autism.
- Another recent article claimed that vaccination refusal was a “white privilege” problem: it takes money and time to refuse vaccinations, and whites are more likely to have that money and time than are minorities (http://www.xojane.com/issues/vaccination-refusal-white-privilege).
- Finally, another study, recently reported on NPR (http://www.npr.org/blogs/health/2015/02/06/384322665/to-get-parents-to-vaccinate-their-kids-dont-ask-just-tell), claims that a doctor’s approach has a major impact on

whether or not parents vaccinate their children. When doctors just simply presumed that the parent was going to be fine with the vaccines that the doctor was going to recommend (e.g. "Johnny is due for his DTaP shot today"), parents were much more likely to get their child vaccinated than they were when the doctor asked them how they felt about vaccination.

To assess the validity and importance of these different claims, the CDC has collected complete data from 2000 parents of young children. The items included in the survey are:

Variable	Description
vaccination	Scale that measures feelings about vaccination. Ranges from 0 = extremely negative about vaccinations to 100 = extremely positive. This is the dependent variable.
white	Coded 1 if white, 0 if non-white
autism	Scale that measures beliefs about whether vaccines can lead to autism. 0 = no chance that vaccinations can cause autism to 100 = extremely likely that vaccinations can cause autism.
approach	Scale that measures how forceful the children's doctor is in pushing vaccinations. 0 = not forceful at all, 100 = just assumes the parent will want their child vaccinated.

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 vaccinate white autism approach

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   vaccinate |      2000     47.3155     26.8503          1        100
       white |      2000        .885    .3191017          0          1
      autism |      2000     27.4605    17.03163          0         71
    approach |      2000      60.605    15.83222          0        100

. reg vaccinate i.white autism approach, vce(robust)

Linear regression                                      Number of obs =    2000
                                                       F(  3,  1996) =   47.22
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0648
                                                       Root MSE      =  25.985

------------------------------------------------------------------------------
             |               Robust
   vaccinate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |
      White  |  -4.363207   1.798159    -2.43   0.015    -7.889673   -.8367423
      autism | -.2537355   .0352135    -7.21   0.000    -.3227946   -.1846764
    approach |   .2573966   .0386828     6.65   0.000     .1815336    .3332596
       _cons |   42.54512   3.013613    14.12   0.000     36.63497    48.45528
------------------------------------------------------------------------------
```

```
. reg vaccinate i.white autism approach

      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F(  3,  1996) =     [2]
       Model |  93383.6877   [1]  31127.8959           Prob > F      =  0.0000
    Residual |  1347772.23  1996  675.236589           R-squared     =     [3]
-------------+------------------------------           Adj R-squared =  0.0634
       Total |  1441155.92  1999      [4]              Root MSE      =  25.985

------------------------------------------------------------------------------
   vaccinate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |
      White  |  -4.363207   1.838179     [5]   0.018    -7.968157   -.7582576
      autism |  -.2537355   .0357901    -7.09   0.000    -.3239253   -.1835457
    approach |   .2573966   .0387209     6.65   0.000     .1814591    .3333342
       _cons |   42.54512   3.099787    13.73   0.000     36.46596    48.62428
------------------------------------------------------------------------------


. estat hettest

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

         chi2(1)      =     2.40
         Prob > chi2  =   0.1212

. estat imtest

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      14.88      8    0.0614
            Skewness |      44.28      3    0.0000
            Kurtosis |      56.84      1    0.0000
---------------------+-----------------------------
               Total |     116.00     12    0.0000
---------------------------------------------------

. testparm i.white autism approach

 ( 1)  1.white = 0
 ( 2)  autism = 0
 ( 3)  approach = 0

       F(  3,  1996) =   46.10
            Prob > F =    0.0000

. test approach = -autism

 ( 1)  autism + approach = 0

       F(  1,  1996) =    0.00
            Prob > F =    0.9514
```

```
. pcorr vaccinate white autism approach
(obs=2000)

Partial and semipartial correlations of vaccinate with

               Partial   Semipartial      Partial   Semipartial   Significance
   Variable |    Corr.         Corr.      Corr.^2       Corr.^2          Value
------------+-----------------------------------------------------------------
      white |  -0.0531       -0.0514       0.0028        0.0026         0.0177
     autism |  -0.1567       -0.1535       0.0246        0.0235         0.0000
   approach |   0.1472        0.1439       0.0217        0.0207         0.0000

. reg vaccinate i.white autism approach i.white#i.white

      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F(  3,  1996) =   46.10
       Model |  93383.6877     3  31127.8959           Prob > F      =  0.0000
    Residual |  1347772.23  1996  675.236589           R-squared     =  0.0648
-------------+------------------------------           Adj R-squared =  0.0634
       Total |  1441155.92  1999  720.938429           Root MSE      =  25.985

------------------------------------------------------------------------------
   vaccinate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |
      White  |  -4.363207   1.838179    -2.37   0.018    -7.968157   -.7582576
      autism |  -.2537355   .0357901    -7.09   0.000    -.3239253   -.1835457
    approach |   .2573966   .0387209     6.65   0.000     .1814591    .3333342
       _cons |   42.54512   3.099787    13.73   0.000     36.46596    48.62428
------------------------------------------------------------------------------
```

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

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. In your discussion, indicate whether or not the beliefs that caused the CDC to examine the variables in the first place were borne out by the results.

2. An additional 227 cases were dropped from the analysis because they were missing data on race and/or approach. If you wanted to keep those cases in the analysis, what multiple imputation method or methods would you recommend using (e.g. logit, mlogit, regress, ologit, pmm, poisson, or something else)? Briefly explain why.

3. The researchers ran the regression with vce(robust) and then again without vce(robust). They noticed that the coefficients did not change, so they decided to not use vce(robust). Do you think this was sound reasoning on their part? Whether it was or was not sound reasoning is there other evidence from the printout that supports or challenges their decision to not use robust standard errors?

4. Some of the researchers believe that beliefs about autism have the greatest impact on support for vaccinations. Others say that it is the doctor's approach that matters the most. Still

others contend that both variables are about equally important and that the differences in their effects are either trivial or non-existent. What is your own position on this, and why? Be sure to cite multiple pieces of information from the printout to support your position.

5. An undergraduate intern has been told that it is often important to include squared terms in models, so he added white^2 to the final regression. To his surprise, none of the results changed. Indeed the squared term didn't even show up in the output. Explain to him why this was the case. [Note: You can draw on your vast sociological expertise in offering a theoretical explanation for this. Or, if that student happens to be visiting us this weekend, you can explain why Notre Dame Sociology may want to think twice before admitting him.]