Sociology 63993
Exam 3
May 7, 2014

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

1. The use of standardized coefficients is one way to effectively deal with problems of random measurement error in variables.
2. Stepwise regression, pairwise deletion of missing data, and analysis of residuals can be done in both OLS and logistic regression.
3. A researcher has a sample of blacks and a sample of whites. For both samples separately, he regresses political liberalism on education. The R^2 value is larger for whites than it is for blacks. This means that the structural effect of education on liberalism is larger for whites than it is for blacks (i.e. $\beta^{White} > \beta^{Black}$).
4. Fixed effects models provide a means of controlling for omitted variables whose effects vary across time.
5. When analyzing svyset data (i.e. data where cases have unequal weights and/or stratification and clustering are used) Wald tests rather than likelihood ratio tests should be used.

II. Short answer. (25 pts each, 50 pts total). Answer *both* of the following.

II-1. (25 points): Donald Sterling, the owner of the Los Angeles Clippers basketball team, provoked outrage when a recording revealed he made the following statements to a woman associate: “It bothers me a lot that you want to broadcast that you're associating with black people”, and, “You can sleep with [black people]. You can bring them in, you can do whatever you want”, but “the little I ask you is ... not to bring them to my games.” The National Basketball Association has responded by imposing a lifetime ban on Sterling and demanding that he sell his team. Sterling, however, may want to fight back. He believes, among other things, that a majority of Americans actually support him in opposing the ban. He also believes that African Americans support him as much as other racial groups do; that sports fans are especially supportive of him; and that younger people are more upset by his comments than older people are.

He has therefore commissioned a study to determine where support for him is strongest and weakest. A survey firm has collected data from 7,200 adults on the following:

Variable	*Description*
sterling	Coded 1 if the person supports Sterling and opposes the ban, 0 otherwise
black	Coded 1 if black, 0 otherwise
fan	Coded 1 if the person considers himself or herself a sports fan, 0 otherwise
c_age	Age in years of the respondent, centered to have a mean of 0.

The study obtains the following results (parts of the output have been deleted):

```
. fre sterling

sterling
-----------------------------------------------------------------------
                            |      Freq.    Percent      Valid       Cum.
----------------------------+------------------------------------------
Valid   0 Opposes Sterling  |       5900      81.94      81.94      81.94
        1 Supports Sterling |       1300      18.06      18.06     100.00
        Total               |       7200     100.00     100.00
-----------------------------------------------------------------------

. nestreg, lr: logit sterling black fan c_age

Block  1: black

Iteration 0:   log likelihood =  -3400.091
Iteration 1:   log likelihood = -3337.2523
Iteration 2:   log likelihood = -3336.4485
Iteration 3:   log likelihood = -3336.4484

Logistic regression                               Number of obs   =       7200
                                                  LR chi2(1)      =        [1]
                                                  Prob > chi2     =     0.0000
Log likelihood = -3336.4484                       Pseudo R2       =     0.0187

------------------------------------------------------------------------------
    sterling |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |  -.6954329    .061942   -11.23   0.000    -.8168369   -.5740288
       _cons |   -1.14762   .0424811   -27.01   0.000    -1.230881   -1.064358
------------------------------------------------------------------------------

Block  2: fan

Logistic regression                               Number of obs   =       7200
                                                  LR chi2(2)      =     469.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -3165.2316                       Pseudo R2       =     0.0691

------------------------------------------------------------------------------
    sterling |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |  -1.332073   .0737807   -18.05    [2]     -1.47668   -1.187465
         fan |   -1.34692    .076113   -17.70   0.000    -1.496098   -1.197741
       _cons |  -.2418044   .0643669    -3.76   0.000    -.3679612   -.1156476
------------------------------------------------------------------------------
```

Block 3: c_age

```
Logistic regression                               Number of obs   =       7200
                                                  LR chi2(3)      =     520.29
                                                  Prob > chi2     =     0.0000
Log likelihood = -3139.9479                       Pseudo R2       =        [3]

------------------------------------------------------------------------------
    sterling |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |  -1.248073   .0746598   -16.72   0.000    -1.394404   -1.101743
         fan |  -1.335674   .0762328   -17.52   0.000    -1.485088    -1.18626
       c_age |   .0478603   .0067525     7.09   0.000     .0346256    .0610949
       _cons |  -.3076225   .0654201    -4.70   0.000    -.4358435   -.1794015
------------------------------------------------------------------------------
```

```
  +----------------------------------------------------------------+
  | Block |        LL       LR     df  Pr > LR       AIC       BIC |
  |-------+--------------------------------------------------------|
  |     1 | -3336.448   127.29      1   0.0000  6676.897   6690.66 |
  |     2 | -3165.232   342.43      1   0.0000  6336.463  6357.109 |
  |     3 | -3139.948    50.57      1   0.0000  6287.896  6315.423 |
  +----------------------------------------------------------------+
```

Based on the printout above, answer the following.

a. (6 points) Fill in the missing items [1], [2] and [3]. (HINT: The calculations are pretty simple.)

b. (6 pts) Using Model 3 (i.e. Block 3), complete the following table:

black	*fan*	*c_age*	*Log odds*	*Odds*	*P(priority = 1)*
0	0	0			
0	1	0			

c. (9 points) Explain which of the models you think is best, and why. Explain what the model tells us about the effects (or non-effects) of the three independent variables included in the analysis. Be sure to indicate whether and how the results support or refute Sterling's beliefs that a majority of Americans actually support him in opposing the ban; African Americans support him as much as other racial groups do; sports fans are especially supportive of him; and that younger people are more upset by his comments than older people are. [Hint: think about what his beliefs imply about the sign and/or significance of each variable in the model.]

d. (4 points) The researchers also ran the following:

```
. estat class

Logistic model for sterling

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |       150           100  |        250
     -     |      1150          5800  |       6950
-----------+--------------------------+-----------
   Total   |      1300          5900  |       7200

Classified + if predicted Pr(D) >= .5
True D defined as sterling != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   11.54%
Specificity                     Pr( -|~D)   98.31%
Positive predictive value       Pr( D| +)   60.00%
Negative predictive value       Pr(~D| -)   83.45%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    1.69%
False - rate for true D         Pr( -| D)   88.46%
False + rate for classified +   Pr(~D| +)   40.00%
False - rate for classified -   Pr( D| -)   16.55%
--------------------------------------------------
Correctly classified                        82.64%
--------------------------------------------------

. bitesti 7200 5950 0.8194, detail

        N   Observed k   Expected k   Assumed p   Observed p
------------------------------------------------------------
     7200       5950      5899.68       0.81940      0.82639

  Pr(k >= 5950)              = 0.062921  (one-sided test)
  Pr(k <= 5950)              = 0.940816  (one-sided test)
  Pr(k <= 5849 or k >= 5950) = 0.125565  (two-sided test)

  Pr(k == 5950)              = 0.003737  (observed)
  Pr(k == 5850)              = 0.003826
  Pr(k == 5849)              = 0.003651  (opposite extreme)
```

Are you impressed by these results of the classification analysis? Do you think you could have done just as well even without running the logistic regressions? Put another way, are more cases correctly classified by the logistic regression than you likely would have correctly classified yourself? (Remember that the earlier output showed the frequency counts for sterling. It will help if you can figure out what the `bitesti` command is telling you but it isn't essential.)

II-2. (25 points) For each of the following circumstances describe the statistical technique you would use for revealing the relationship between the dependent and independent variables. Write a few sentences explaining and justifying your answer. In some instances more than one technique may be reasonable. Some problems may require the use of advanced techniques while in other instances the required technique may be simple and basic.

a. A physician has developed a new exercise program. She believes that those who participate in the program will be happier, more physically fit, and will work better on the job than those who do not participate. Happiness, physical fitness, and job productivity are all measured on interval-level scales. Participation in the program is coded 0 or 1.

b. A college is offering an online course. Students have up to 12 weeks to finish the course, but some will finish it sooner while some may never finish at all. The college wants to know what affects how quickly someone finishes. Every week, data is collected on the student's gender, whether or not they were employed full-time that week, and whether or not they completed the course that week. Data collection is in week 6 now and half the students still have not completed the course.

c. A researcher is frustrated because several of her variables have small amounts of missing data. Even though no one variable has more than 5% of its data missing, if she uses listwise deletion she will lose 40% of the cases when she estimates her final model.

d. Thanks to Notre Dame Professor Terry McDonnell, Notre Dame now has its very own "Happy" video (check it out at http://youtu.be/ofGZlmcvFkw if you haven't already). In order to see whether the video really does make people happy, 100 students will have their sadness/happiness measured on a 100 point scale (ranging from very sad to very happy). They will then see the video, and their sadness/happiness will again be measured using the same scale.

e. It is October 2016. Conservative Republicans are dismayed by the possibility that Hillary Clinton and Elizabeth Warren may become the first female President and Vice-President of the United States. They need to understand the determinants of Clinton/Warren support better. They have therefore gathered polling data on support for the Democratic Ticket (measured on a 100 point scale), gender (1 = female, 0 = male) and political ideology (a 100 point scale that ranges from very conservative to very liberal). They believe that the effect of ideology will be greater for women than it is for men.

III. *Essay.* (30 points) Answer *one* of the following questions.

1. Several assumptions are made when using OLS regression. Discuss TWO of the following in depth. What does the assumption mean? When might the assumption be violated? What effects do violations of the assumption have on OLS estimates? How can violations of the assumption be avoided or dealt with? Be sure to talk about techniques such as 2SLS and logistic regression where appropriate. [NOTE: While the material from the last third of the course is especially relevant here, you should try to tie in earlier material as much as possible too. Also, keep in mind that there are often different ways an assumption can be violated, and the appropriate solutions will therefore often differ too.]

a. The effects of the independent variables are linear and additive
b. Errors are homoskedastic
c. Variables are measured without error
d. All relevant variables are included in the model

2. We've talked about several ways that OLS regression can be modified to deal with violations of its assumptions. Some problems, however, require the use of techniques besides OLS. For three of the following, explain why and when the method would be used instead of OLS. Be sure to make clear what assumptions would be violated if OLS was used instead.

a. 2 stage least squares
b. Logistic regression
c. Robust regression techniques (e.g. rreg, qreg, robust standard errors)
d. Event History Analysis
e. Fixed effects regression models
f. Structural Equation Modeling using multiple indicators of variables

3. Your psychology professor has told you that you should almost always focus on standardized, rather than unstandardized (metric) coefficients. Explain to your professor (as politely as possible) why he is wrong. Among other things, you may want to discuss the relative strengths and weaknesses of standardized vs. unstandardized coefficients with regard to:

a. Variables with arbitrary metrics (e.g. attitudinal scales)
b. Structural equation models
c. Multiple-group comparisons
d. Interpretability of coefficients
e. Effect of random measurement error on coefficients