Sociology 63993
Exam 3
May 9, 2013

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

1. Bivariate regressions are run in two separate populations. The R^2 value is twice as large in population 1 as it is in population 2. This means that the exogenous variances must be different in the two populations.

2. In a logistic regression the Pseudo R^2 is .999. This means that almost all the subjects experience the event.

3. If a model is underidentified, we should try adding a variable that is uncorrelated with any of the variables that are already in the model.

4. In logistic regression, the odds ratio and the odds are two different names for the same thing.

5. Similar or even identical techniques can be used to assess multicollinearity in both logistic and OLS regression.

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

II-1.   (25 points): The nation is rejoicing as three long-time kidnap victims have finally been found and freed in Ohio. However, the case is also re-igniting concerns about how missing person cases are handled in this country; in particular, do cases involving non-Hispanic whites receive higher priority from police and the public than do cases involving minorities? As Joan Walsh writes in Salon (http://www.salon.com/2013/05/08/cleveland%E2%80%99s_lost_girls/),

> I wonder if any of the missing girls were considered "white" by authorities — or at least white enough to be part of the "missing white woman syndrome," in which the disappearance of pretty, upper-middle-class white girls and women becomes a police priority and a national scandal. Think Chandra Levy, Natalee Holloway or Laci Peterson.

A criminologist is interested in determining what causes a missing persons case to become a police priority. She has drawn a random sample of 7,932 missing person reports drawn from police files across the country. She has developed a way to determine whether the case received high priority from the police. Her measures include

Variable	Description
priority	Coded 1 if the case was treated as a high priority by the police, 0 otherwise
white	Coded 1 if the missing person was a non-Hispanic white, 0 otherwise
minor	Coded 1 if the missing person was under the age of 18, 0 otherwise
highses	Coded 1 if the missing person was from a wealthy family, 0 otherwise

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

```
. nestreg, lr: logit priority white minor highses

Block  1: white

Iteration 0:   log likelihood = -5494.0397
Iteration 1:   log likelihood = -5440.7181
Iteration 2:   log likelihood = -5440.6037
Iteration 3:   log likelihood = -5440.6037

Logistic regression                             Number of obs   =       7932
                                                LR chi2(1)      =     106.87
                                                Prob > chi2     =     0.0000
Log likelihood = -5440.6037                     Pseudo R2       =     0.0097


------------------------------------------------------------------------------
    priority |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |    .789018   .0787585     10.02   0.000     .6346542    .9433818
       _cons |  -.0143442   .0237179     -0.60   0.545    -.0608304     .032142
------------------------------------------------------------------------------

Block  2: minor

Logistic regression                             Number of obs   =       7932
                                                LR chi2(2)      =     157.78
                                                Prob > chi2     =     0.0000
Log likelihood = -5415.1516                     Pseudo R2       =     0.0144


------------------------------------------------------------------------------
    priority |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |   .8896065   .0801685     11.10   0.000     .7324792    1.046734
       minor |   .3403085   .0478237      7.12   0.000     .2465758    .4340411
       _cons |  -.1492367   .0304071     -4.91   0.000    -.2088335     -.08964
------------------------------------------------------------------------------
```

Block 3: highses

```
Logistic regression                              Number of obs   =       7932
                                                 LR chi2(3)      =     173.18
                                                 Prob > chi2     =     0.0000
Log likelihood = -5407.4493                      Pseudo R2       =     [1]

------------------------------------------------------------------------------
    priority |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |   .8912017   .0802451    11.11   0.000     .7339242    1.048479
       minor |   .3494764   .0479373     [2]    0.000      .255521    .4434318
     highses |   .1786599   .0455461     3.92   0.000     .0893911    .2679287
       _cons |  -.2467203   .0393447    -6.27   0.000    -.3238345    -.169606
------------------------------------------------------------------------------
```

```
+----------------------------------------------------------------------+
| Block |        LL        LR     df   Pr > LR       AIC         BIC  |
|-------+--------------------------------------------------------------|
|     1 |  -5440.604    106.87     1    0.0000   10885.21    10899.16  |
|     2 |  -5415.152     50.90     1    0.0000    10836.3    10857.24  |
|     3 |  -5407.449       [3]     1    0.0001    10822.9    10850.81  |
+----------------------------------------------------------------------+
```

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:

white	minor	highses	Log odds	Odds	P(priority = 1)
1	0	0			
0	0	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 make clear what your preferred model says about the relationship between race, SES and age with the likelihood of a case being a high priority.

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

```
. fre priority
```

priority -- RECODE of health (1=poor,..., 5=excellent)

```
--------------------------------------------------------------
              |     Freq.     Percent      Valid       Cum.
--------------+-----------------------------------------------
Valid   0     |     3840       48.41       48.41      48.41
        1     |     4092       51.59       51.59     100.00
      Total   |     7932      100.00      100.00
--------------------------------------------------------------
```

```
. estat clas
```

Logistic model for priority

```
                -------- True --------
Classified |         D              ~D  |     Total
-----------+----------------------------+-----------
     +     |      2114            1527   |      3641
     -     |      1978            2313   |      4291
-----------+----------------------------+-----------
   Total   |      4092            3840   |      7932
```

Classified + if predicted Pr(D) >= .5
True D defined as priority != 0

```
--------------------------------------------------
Sensitivity                    Pr( +| D)    51.66%
Specificity                    Pr( -|~D)    60.23%
Positive predictive value      Pr( D| +)    58.06%
Negative predictive value      Pr(~D| -)    53.90%
--------------------------------------------------
False + rate for true ~D       Pr( +|~D)    39.77%
False - rate for true D        Pr( -| D)    48.34%
False + rate for classified +  Pr(~D| +)    41.94%
False - rate for classified -  Pr( D| -)    46.10%
--------------------------------------------------
Correctly classified                        55.81%
--------------------------------------------------
```

```
. bitesti 7921 4427 0.5159, detail
```

```
        N    Observed k    Expected k    Assumed p    Observed p
------------------------------------------------------------------
      7921         4427      4086.444      0.51590       0.55889

  Pr(k >= 4427)              = 0.000000  (one-sided test)
  Pr(k <= 4427)              = 1.000000  (one-sided test)
  Pr(k <= 3745 or k >= 4427) = 0.000000  (two-sided test)

  Pr(k == 4427)              = 0.000000  (observed)
  Pr(k == 3746)              = 0.000000
  Pr(k == 3745)              = 0.000000  (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? (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.

a.   Microsoft is concerned because its new Windows 8 operating system has not received a more enthusiastic reception. It believes that a big part of the problem is that people do not understand how to use the new interface. Therefore, 200 randomly selected respondents will see a 15 minute video on how to use Windows 8. A different 200 randomly selected subjects will not. All 400 subjects will then be asked to rate (on 100 point scales) how difficult they think Windows 8 is to use, how enjoyable they think it would be to use Windows 8, and how likely they are to purchase Windows 8 in the future.

b.   A computer problem caused a research firm to lose data on gender for 23% of the respondents to a survey. The firm is confident that data were lost at random. Still, gender needs to be included in its statistical models, and the firm is very reluctant to simply discard all the other usable data it has collected for those cases where gender is missing.

c.   Starting at age 30, a group of respondents was interviewed annually for 10 years. No more data will be collected. The researchers are just now realizing that some critical variables were never measured. These include whether or not the respondent's parents were married when the respondent was born, and whether or not the respondent ever spent a year or more overseas before reaching the age of 12.

d.   Facebook is concerned because some people are starting to drop the service. It has therefore selected a random sample of 1 million Facebook users. For each user it has information on age, gender, number of friends, and other demographic variables. Over the course of 3 years, it will examine how these variables are related to how long somebody remains a Facebook user.

e.   A researcher wants to get an unbiased estimate of the effect of self-confidence on academic performance. She has 4 items that measure self-confidence and another 5 items that tap academic performance. Unfortunately, she is sure that all of these items suffer from random measurement error.

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 <u>three</u> 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

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