Sociology 63993
Exam 3 Answer Key [Draft]
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.

False. The exogenous variances could be different, but it could also be that the structural effects and/or the residual variances differ across populations

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

False. It means that the model is pretty good at predicting the individual outcomes, whether they are 0 or 1. The statement would be true if it instead said p = .999 or odds = 999.

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.

False. Adding a variable that is uncorrelated with anything else has the same effect as adding no variable at all. In the case of reciprocal causation, you need variables that directly affect one of the dependent variables while only indirectly affecting the other. Because of the direct and indirect effects the added variable must be correlated with the already existing variables.

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

False. The odds ratio is a ratio of odds. For example it may tell you that the odds for men are twice as great as the odds for comparable women, no matter what the odds for women are.

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

True. For example, you could use Stata's `collin` command.

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.)

Here is the uncensored printout for the last parts of the output:

```
Block  3: highses

Iteration 0:   log likelihood = -5494.0397
Iteration 1:   log likelihood = -5407.5937
Iteration 2:   log likelihood = -5407.4494
Iteration 3:   log likelihood = -5407.4493

Logistic regression                                     Number of obs   =       7932
                                                        LR chi2(3)      =     173.18
                                                        Prob > chi2     =     0.0000
Log likelihood = -5407.4493                             Pseudo R2       =     0.0158

------------------------------------------------------------------------------
    priority |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |   .8912017   .0802451    11.11   0.000     .7339242    1.048479
       minor |   .3494764   .0479373     7.29   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    15.40      1   0.0001   10822.9  10850.81 |
  +--------------------------------------------------------------------+
```

To confirm,

[1] = Pseudo R^2 = G_M/Dev_0 = $G_M/(-2 * LL_0)$ =173.18/ (-2*-5494.0397) = .0157598

[2] = z_{minor} = b_{minor}/se_{minor} = .3494764/ .0479373 = 7.29

[3] = incremental chi square = $G_{m3} - G_{m2}$ = 173.18 – 157.78 = 15.40

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			

Note that the coefficient for white is .8912017 and the constant is -.2467203. For the purposes of this problem the other coefficients do not matter because the values of the variables are 0. Ergo,

White	*Minor*	*highses*	*Log odds = a + Xb*	*Odds = exp(LogOdds)*	*P(priority = 1) = Odds/(1 + Odds)*
1	0	0	.6444814	1.904998841	.655765783
0	0	0	-.2467203	.781359209	.438630909

The margins command in Stata can do this easily, especially if we redo the logit command using factor variables:

```
. * Confirm answers
. quietly logit priority i.white i.minor i.highses
. * Log Odds
. margins white, at(minor = 0 highses = 0) predict(xb)

Adjusted predictions                              Number of obs   =       7932
Model VCE    : OIM

Expression   : Linear prediction (log odds), predict(xb)
at           : minor           =           0
               highses         =           0

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |
           0 |  -.2467203   .0393447    -6.27   0.000    -.3238345    -.169606
           1 |   .6444814   .0791399     8.14   0.000       .48937    .7995927
------------------------------------------------------------------------------
```

```
. * Odds
. margins white, at(minor = 0 highses = 0) expression(exp(predict(xb)))

Adjusted predictions                              Number of obs   =       7932
Model VCE    : OIM

Expression   : exp(predict(xb))
at           : minor           =           0
               highses         =           0

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |
          0  |   .7813592   .0307424    25.42   0.000     .7211053    .8416132
          1  |   1.904999   .1507614    12.64   0.000     1.609512    2.200486
------------------------------------------------------------------------------

. * Probabilities
. margins white, at(minor = 0 highses = 0)

Adjusted predictions                              Number of obs   =       7932
Model VCE    : OIM

Expression   : Pr(priority), predict()
at           : minor           =           0
               highses         =           0

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |
          0  |   .4386309    .009688    45.28   0.000     .4196428    .4576191
          1  |   .6557658   .0178648    36.71   0.000     .6207514    .6907802
------------------------------------------------------------------------------
```

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.

Model 3 is best – all terms in it are statistically significant, regardless of whether you use Wald tests or Likelihood Ratio tests. The model says that whites, minors, and individuals from wealthier families are all more likely to have their cases given high priority by the police. For example, as the table showed, an adult non-wealthy non-Hispanic white missing person is almost 22 percentage points more likely to have their case given a high priority than is a comparable minority group member. Of course, there might be other variables currently omitted from the model that might account for this, e.g. maybe there are other reasons that these cases actually do deserve higher priority. But, as it stands, if the data were real (as opposed to being totally fictitious!) the results would seem to support Walsh's concerns.

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.)

If you picked yourself, you would be right 51.59% of the time if you always picked the case to have high priority. Or, you would expect to be right about 51.59% of the time if you randomly picked 51.59% of the cases to be high priority. The classification table was right 55.81% of the time, which is better than that. That may not seem like much, but the `bitesti` command shows that your odds of getting that many right just by guessing are infinitesimal. That is, as Pr(k >= 4427) shows, the probability of correctly classifying 4,427 or more cases (which is what the classification table did) by chance alone is incredibly small (less than 1 in a million).

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.

There is one binary treatment and multiple continuous dependent variables. MANOVA would be a good choice. MANOVA takes into account that there are multiple dependent variables that may themselves be correlated with each other, thus making the significance tests more accurate. You could also set this up as a structural equation model.

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.

Multiple imputation would be good. You can impute values for the missing cases, hence keeping all of the cases in the analysis. You should use logit for the imputation since gender is a dichotomous variable. Since data are missing at random it might not be too horrible to use pairwise deletion with techniques that allow you to analyze a correlation matrix (e.g. OLS regression; but not logistic regression), but multiple imputation is probably better.

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.

When panel data are available a fixed effects regression model can be used to control for omitted variables that have time invariant values with time invariant effects. The variables that are specifically mentioned are time-invariant, i.e. their values were permanently fixed long before the study began. So, if their effects are also time-invariant, a fixed effects model will allow you to control for them, since each variable will have the same effect on the respondents at each time period and hence can be partialled out. Put another way, you can control for the effects of the omitted variables but you can't estimate what their effects are.

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.

We are interested in how long somebody stays a user, not just whether they stay. Further, the outcome is censored (people could drop out after the three years are up) and the values of the independent variables could change across time. All of these characteristics make the problem a good candidate for Event History Analysis.

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.

A structural equation modeling approach that combines both a measurement model with multiple indicators and a two variable structural model would be a good option. Having multiple indicators of each concept allows you to control for random measurement error and get unbiased estimates of effects. A less advanced approach that simply combined the items into scales might also be ok.

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

See the course notes for ideas on each of these.

Appendix: Stata Code used in the exam

```
* II-1
* Manipulate the data
webuse nhanes2f, clear
set seed 123456
sample 7932, count
recode health (1 2 3  = 1)(else = 0), gen(priority)
gen white = black
gen minor = rural
gen highses = female

* Generate the output
nestreg, lr: logit priority white minor highses
fre priority
estat clas
bitesti 7921 4427 0.5159, detail

* Confirm the answers
quietly logit priority i.white i.minor i.highses
* Log Odds
margins white, at(minor = 0 highses = 0) predict(xb)
* Odds
margins white, at(minor = 0 highses = 0) expression(exp(predict(xb)))
* Probabilities
margins white, at(minor = 0 highses = 0)
```