Sociology 63993
Exam 3 Answer Key - DRAFT
May 8, 2007

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

1. The odds of an event occurring range from negative infinity to positive infinity.

False. Odds range from 0 to positive infinity.

2. If a model is not identified, then OLS regression should be used to estimate its parameters.

False. If you want to estimate the model, you need to figure out how to get it identified, and then use a technique like 2sls.

3. McFadden's Pseudo R^2 is popular because it uses the exact same formula for R^2 that is used in OLS regression.

False. McFadden's Pseudo R^2 is a logical analog to OLS R^2, but the formulas are not identical.

4. A Brant test can be used to test the assumptions of the ordered logit model.

True.

5. The main advantage of a multinomial logit model over an ordered logit model is that the multinomial logit model has fewer parameters and is hence easier to interpret.

False. Just the opposite is true; the ordered logit model is more parsimonious (but its assumptions are not always met.)

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

II-1. (25 points) There is growing concern about diabetes in the United States. A demographer and a public health researcher have joined together to examine how demographic characteristics are related to the risk of having diabetes. They have gathered information from more than 20,000 individuals on the following:

Variable	*Description*
diabetes	Coded 1 if respondent has diabetes, 0 otherwise
black	Coded 1 if black, 0 otherwise
female	Coded 1 if female, 0 if male
age	Age in years

They obtain the following results:

```
. nestreg, lr: logit diabetes black female age, nolog

Block  1: black

Logistic regression                               Number of obs   =      20670
                                                  LR chi2(1)      =      43.58
                                                  Prob > chi2     =     0.0000
Log likelihood =  -3976.344                       Pseudo R2       =     0.0054

------------------------------------------------------------------------------
    diabetes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |    .609734   .0870692     7.00   0.000     .4390816    .7803864
       _cons |  -3.063142   .0355983   -86.05   0.000    -3.132913   -2.993371
------------------------------------------------------------------------------

Block  2: female

Logistic regression                               Number of obs   =      20670
                                                  LR chi2(2)      =      50.02
                                                  Prob > chi2     =     0.0000
Log likelihood = -3973.1221                       Pseudo R2       =     0.0063

------------------------------------------------------------------------------
    diabetes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |   .6072815   .0870934     6.97   0.000     .4365816    .7779813
      female |   .1658387   .0655102     2.53   0.011     .0374412    .2942363
       _cons |   -3.15304   .0511592   -61.63   0.000     -3.25331    -3.05277
------------------------------------------------------------------------------

Block  3: age

Logistic regression                               Number of obs   =      20670
                                                  LR chi2(3)      =     748.34
                                                  Prob > chi2     =     0.0000
Log likelihood = -3623.9656                       Pseudo R2       =     0.0936

------------------------------------------------------------------------------
    diabetes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |   .7179046    .089665     8.01   0.000     .5421644    .8936447
      female |   .1545569   .0666786     2.32   0.020     .0238692    .2852445
         age |   .0594654   .0026398    22.53   0.000     .0542916    .0646393
       _cons |  -6.405437   .1677385   -38.19   0.000    -6.734198   -6.076675
------------------------------------------------------------------------------


  +--------------------------------------------------------------------+
  | Block |        LL       LR     df  Pr > LR       AIC       BIC |
  |-------+------------------------------------------------------------|
  |     1 | -3976.344    43.58      1   0.0000  7956.688  7972.561 |
  |     2 | -3973.122     6.44      1   0.0111  7952.244  7976.054 |
  |     3 | -3623.966   698.31      1   0.0000  7255.931  7287.677 |
  +--------------------------------------------------------------------+
```

Based on the printout above, answer the following.

a. In Model 1 (i.e. Block 1), what do DEV_M, G_M, DEV_0, and McFadden's Pseudo R^2 equal?

Note that LL_M = -3976.344, so $DEV_M = -2LL_M$ = 7952.688. G_M = The Model Chi-Square (labeled by Stata as LR chi2(3)) = 43.58. $DEV_0 = DEV_M + G_M$ = 7952.688 + 43.58 = 7996.268. McFadden's Pseudo R^2 is included in the printout and equals .0054. To confirm, Pseudo $R^2 = G_M/DEV_0$ = 43.58/7996.268 = .00545.

b. Using Model 2 (i.e. Block 2), complete the following table:

Race	*Gender*	*Log odds*	*Odds*	*P(Diabetes)*
Black	Female			
Not Black	Male			

Here are the answers:

Race	*Gender*	*Log odds = a + XB*	*Odds = exp(Log Odds)*	*P(Diabetes) = Odds/(1 + Odds)*
Black	Female	.6072815 + .1658387 - 3.15304 = -2.37992	.092558	.084717
Not Black	Male	-3.15304	.042722	.040972

We can use Stata to confirm the results:

```
. adjust black=1 female=1, xb

------------------------------------------------------------------------------
     Dependent variable: diabetes     Command: logit
Covariates set to value: black = 1, female = 1
------------------------------------------------------------------------------

----------------------
      All |         xb
----------+-----------
          |   -2.37992
----------------------
     Key:  xb  =  Linear Prediction

. adjust black=1 female=1, exp

------------------------------------------------------------------------------
     Dependent variable: diabetes     Command: logit
Covariates set to value: black = 1, female = 1
------------------------------------------------------------------------------

----------------------
      All |    exp(xb)
----------+-----------
          |    .092558
----------------------
     Key:  exp(xb)  =  exp(xb)

. adjust black=1 female=1, pr

------------------------------------------------------------------------------
     Dependent variable: diabetes     Command: logit
Covariates set to value: black = 1, female = 1
------------------------------------------------------------------------------

----------------------
      All |         pr
----------+-----------
          |    .084717
----------------------
     Key:  pr  =  Probability
```

```
. adjust black=0 female=0, xb

------------------------------------------------------------------------------
     Dependent variable: diabetes     Command: logit
Covariates set to value: black = 0, female = 0
------------------------------------------------------------------------------

---------------------
      All |         xb
----------+----------
          |   -3.15304
---------------------
     Key:  xb  =  Linear Prediction

. adjust black=0 female=0, exp

------------------------------------------------------------------------------
     Dependent variable: diabetes     Command: logit
Covariates set to value: black = 0, female = 0
------------------------------------------------------------------------------

---------------------
      All |    exp(xb)
----------+----------
          |    .042722
---------------------
     Key:  exp(xb)  =  exp(xb)

. adjust black=0 female=0, pr

------------------------------------------------------------------------------
     Dependent variable: diabetes     Command: logit
Covariates set to value: black = 0, female = 0
------------------------------------------------------------------------------

---------------------
      All |         pr
----------+----------
          |    .040972
---------------------
     Key:  pr  =  Probability
```

c. Three models are estimated. Which model do you think is best, and why? What does this model say about the effect of race, gender and age on diabetes?

Model 3 is best. Both Wald Tests and LR chi-square tests indicate that all three variables are statistically significant and should be included in the model. The results show that blacks, women, and older people are all more likely to have diabetes.

NOTE: This problem uses a modified version of the nhanes2f data, available from the Stata website. From within Stata, type

```
webuse nhanes2f
expand 2
```

The other Stata Commands used are included in the exam.

II-2. (25 points) For <u>each</u> 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. There is ongoing controversy over whether homework helps or hinders the academic and social progress of children in grades 1-4. To address these issues, students are randomly assigned to two classrooms. One class has homework every day, the other class never has any homework. Otherwise the style of teaching and the material covered is identical in the

two classes. After 12 weeks, both classes will take the same standardized tests to measure how much they have learned, how much they enjoy school, and their overall psychological well-being.

There is one treatment and multiple dependent variables. Manova (or a LISREL-type model) would be appropriate.

b. It is summer 2008. A brutal primary campaign has left the Democratic Party bitterly divided, and Rush Limbaugh and Ann Coulter are already cackling over what appears to be an all but certain Republican victory in November. Suddenly, however, after an impassioned call for unity by former presidents Clinton and Carter, the party convention drafts the one man it hopes can bring its warring factions together: the Academy Award winning former vice-president, Al Gore. But, even as Melissa Etheridge leads the delegates in joyous song, Gore knows that the most critical decision of his campaign is only hours away: his choice of a running mate. His instincts tell him that Hillary Clinton is the best choice. But, his instincts once told him that Joe Lieberman would be a great pick too. He has therefore commissioned an overnight telephone survey. A group of 1,000 likely voters will be asked if having Clinton as the vice-presidential nominee would make them more likely to vote for Gore, have no effect on their likelihood of voting for Gore, or make them less likely to vote for Gore. The study will further examine how voter preferences are affected by their gender, race, and party affiliation.

The dependent variable is ordinal, so an ordered logit model (or perhaps some other type of ordinal regression) is appropriate.

Those who would like to sing along with Melissa can do so at

http://www.melissaetheridge.com/

Make sure your computer's audio is turned on!

c. The management of a large company is interested in examining how a team-oriented approach to problem solving works. Individuals work with a partner to solve a problem. It is believed that, the harder one partner works to solve the problem, the harder the other partner works, and vice versa. Information on the IQ, income and education of each partner is also available.

Partners influence each other, so a nonrecursive model (possibly estimated by 2sls or via a maximum likelihood technique such as LISREL provides) seems called for. Fortunately, the background variables for each partner should make the model identified.

d. A researcher is interested in the relationship between popularity and grades. She believes that students with very low grades, and students with very high grades, will be less popular than students whose grades are more in the middle.

This sounds like a nonlinear relationship. It can be estimated via OLS by including a term for grades^2. Alternatively, you might try a piecewise regression model.

e. A gun control group feels that the Virginia Tech shootings provide yet another reason for greater regulation of firearms. However, it fears that if it mentions the Tech shootings in its ads, it will be seen as exploitive and the ads will be ineffective. Subjects will therefore see each of two gun control ads, one which mentions Virginia Tech and one which does not. The perceived effectiveness of each ad will be measured on 100 point scales.

Since the same respondents see both ads, this calls for a matched pairs T-Test.

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

1. 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. Ordered Logit models
d. Robust regression techniques (e.g. rreg, qreg, robust standard errors)
e. Event History Analysis
f. Hierarchical Linear Modeling

2. 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

3. 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. The data are a random and representative sample of the larger population.

See the course notes and readings for information pertaining to each of these subjects.

IV. Extra Credit. (10 points)

Following are additional results related to the analysis of Model III in part II-1.

```
. quietly logit diabetes black female age
. tab1 diabetes if e(sample)

-> tabulation of diabetes if e(sample)

  diabetes, |
1=yes, 0=no |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |     19,672       95.17       95.17
          1 |        998        4.83      100.00
------------+-----------------------------------
      Total |     20,670      100.00
```

```
. predict prob, pr

. sum prob

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        prob |     20674    .0482766    .0417819    .005399   .2436938

. estat clas

Logistic model for diabetes

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |         0             0  |          0
     -     |       998         19672  |      20670
-----------+--------------------------+-----------
   Total   |       998         19672  |      20670

Classified + if predicted Pr(D) >= .5
True D defined as diabetes != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)    0.00%
Specificity                     Pr( -|~D)  100.00%
Positive predictive value       Pr( D| +)       .%
Negative predictive value       Pr(~D| -)   95.17%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    0.00%
False - rate for true D         Pr( -| D)  100.00%
False + rate for classified +   Pr(~D| +)       .%
False - rate for classified -   Pr( D| -)    4.83%
--------------------------------------------------
Correctly classified                        95.17%
--------------------------------------------------

. test age = female

 ( 1) - female + age = 0

           chi2(  1) =    2.03
         Prob > chi2 =    0.1542
```

a) According to the classification table, what percentage of the cases were correctly classified? Why were so many cases classified correctly? Do you think this indicates that the model is outstanding? Explain whether you think the classification table is useful or not very useful in this case. Other information in the printout may make this question easier to answer.

Note that less than 5% of the respondents have diabetes, and the highest predicted probability of diabetes is only 24.37%. The classification table therefore predicts that no one will have diabetes. It is always right for the 95% of the subjects who do not have diabetes, but it is always wrong for the 5% who do. Classification tables tend not to be very helpful when you have extreme splits like this; you could easily do just as well yourself once you knew what the frequencies were for diabetes.

b) Explain what the `test` command is testing. Indicate whether or not you would conduct the same test, and why.

The test command is testing whether the effect of one year of age is the same as the effect of being female rather than male. Given how different the measurement of age and gender is, it would take a rather esoteric theory to make such a hypothesis substantively plausible and interesting. It is not something I would test. Perhaps the researcher just looked at the results and decided to see if he could trick students into thinking this was a smart thing to test. ☺