Sociology 593
Exam 3 Answer Key (DRAFT)
May 4, 2005

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 are 1. This means that the event will definitely happen.

False. If the odds of an event occurring are 1, the probability of the event occurring is .5.

2. A key limitation of the LISREL method is that it cannot estimate nonrecursive models.

False. As the class handout showed, LISREL can estimate many types of models, including nonrecursive.

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. $\beta^{White} > \beta^{Black}$).

False. This is one possibility. But, it could also be that the exogenous or residual variances differ across populations.

4. Stepwise regression, analysis of outliers, and the centering of continuous independent variables can all be done in both OLS and logistic regression.

True.

5. If a model is under-identified, 2SLS should be used to estimate it.

False. You need to find a way to identify the model before you can estimate it.

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

II-1. (25 points) Long and Freese (2003) present data from the 1977/1989 General Social Surveys. Respondents are asked to evaluate the following statement: "A working mother can establish just as warm and secure a relationship with her child as a mother who does not work." The variables are

Variable	*Description*
workmom	Coded 1 if respondent agreed or strongly agreed, 0 otherwise
male	Coded 1 if male, 0 if female
yr89	Coded 1 if 1989, 0 if 1977
yr89male	= yr89 * male

Based on the printout below, answer the following.

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

DEV_M = log likelihood$_M$ * -2 = -1550.4116 * -2 = 3100.8232
G_M = 49.98 (printout labels it as LR chi2(1)

$DEV_0 = DEV_M + G_M = 3100.8232 + 49.98 = 3150.8032$
McFadden’s Pseudo $R^2 = .0159$ (printout labels it as Pseudo R2). To confirm,

McFadden’s Pseudo $R^2 = G_M/DEV_0 = 49.98/3150.8032 = .0159$

b. Using Model 2, complete the following table:

Male	*Yr89*	*Log odds*	*Odds*	*P(Agree)*
Female	1977			
Male	1989			

Here is how we can let Stata do the work:

```
. * preserve real data
. preserve

. * Temporarily delete real data
. drop in 1/l
(2293 observations deleted)

. * Interactively enter hypothetical data
. edit
- preserve

. list yr89 male

     +--------------+
     | yr89    male |
     |--------------|
  1. | 1977   Women |
  2. | 1989     Men |
     +--------------+

. * log odds
. predict logodds, xb

. * odds
. gen odds = exp(logodds)

. * Predicted probabilities
. predict p, p

. list male yr89 logodds odds p

     +---------------------------------------------+
     |  male   yr89    logodds       odds          p |
     |---------------------------------------------|
  1. | Women   1977   .2632326   1.301129   .5654308 |
  2. |   Men   1989   .2831538   1.327309   .5703192 |
     +---------------------------------------------+

. * restore original data
. restore
```

Incidentally, this shows that, by 1989, men were about where women were in 1977 in their support for working mothers.

c. Three models are estimated. Which model do you think is best, and why? What does this model say about the effect of gender on support for working mothers? What does this model tell you about differences across time in the determinants of support for working mothers?

Using either the LR chi-square tests or the Wald tests, we see that Model II is a significant improvement over Model I, but Model III is not a significant improvement over Model II. This model shows us that men are less supportive of working mothers than are women (see the negative coefficient for male). Support for working mothers increased between 1977 and 1989 (see the positive coefficient for yr89); however, the effect of gender on support did not change across time, i.e. there was no interaction effect.

```
. * Model 1
. logit  workmom  male, nolog

Logit estimates                                   Number of obs   =       2293
                                                  LR chi2(1)      =      49.98
                                                  Prob > chi2     =     0.0000
Log likelihood = -1550.4116                       Pseudo R2       =     0.0159

------------------------------------------------------------------------------
     workmom |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |  -.5981885    .085043    -7.03   0.000    -.7648696   -.4315073
       _cons |   .5043109    .058921     8.56   0.000     .3888278     .619794
------------------------------------------------------------------------------

. est store m1

. * Model 2
. logit  workmom  male yr89, nolog

Logit estimates                                   Number of obs   =       2293
                                                  LR chi2(2)      =      98.22
                                                  Prob > chi2     =     0.0000
Log likelihood = -1526.2886                       Pseudo R2       =     0.0312

------------------------------------------------------------------------------
     workmom |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   -.589729     .08595    -6.86   0.000    -.7581878   -.4212702
        yr89 |   .6096502   .0885544     6.88   0.000     .4360868    .7832136
       _cons |   .2632326   .0681564     3.86   0.000     .1296486    .3968167
------------------------------------------------------------------------------

. est store m2

. lrtest m2 m1

likelihood-ratio test                                  LR chi2(1)  =     48.25
(Assumption: m1 nested in m2)                          Prob > chi2 =    0.0000
```

```
. * Model 3
. logit  workmom  male yr89  yr89male, nolog

Logit estimates                                   Number of obs   =       2293
                                                  LR chi2(3)      =      99.36
                                                  Prob > chi2     =     0.0000
Log likelihood = -1525.7197                       Pseudo R2       =     0.0315

------------------------------------------------------------------------------
     workmom |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |  -.6614662   .1093128    -6.05   0.000    -.8757154   -.4472169
        yr89 |   .5191738   .1222222     4.25   0.000     .2796226     .758725
    yr89male |   .1889503   .1770994     1.07   0.286    -.1581582    .5360588
       _cons |   .2974382   .0754663     3.94   0.000     .1495271    .4453494
------------------------------------------------------------------------------

. est store m3

. lrtest m3 m2

likelihood-ratio test                                  LR chi2(1)  =      1.14
(Assumption: m2 nested in m3)                          Prob > chi2 =    0.2861
```

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. President Bush wants to know what impact his press conference had on support for his social security plan, his personal popularity, and support for his judicial appointments. All three of these variables are measured on continuous scales that range from 0 to 100. Five hundred American adults will be asked whether or not they saw the press conference and how they feel about each of these three issues.

This sounds like a problem for Manova or possibly LISREL. There is one independent variable (watched/did not watch press conference) and three dependent variables (support for SS plan, Bush's personal popularity, support for judicial appts.)

b. A medical sociologist believes that social psychological factors play a key role in self-perceptions of health. Respondents are asked how their health is, with the possible responses being poor, fair, good, and excellent. They are also asked their gender, income, and the number of close friends they have.

The dependent variable is ordinal so an ordinal regression technique sounds best (at least if the assumptions of the method are met; if not, you may have to use multinomial logit).

c. A professor has repeatedly been told by students that essay exams are fairer than multiple choice exams and that students perform better on them. The professor has decided to determine whether exam format does affect student performance. The first exam will be all multiple choice. The second exam will be all essay. Both will be graded on 100 point scales. Students will use id numbers that keep them anonymous from the professor but which make it possible to record, for each student individually, their score on the first and second exam. The professor will then test whether exam format affects student grades.

Matched pairs t-test. Subjects receive two treatments (essay and multiple choice) and you want to see if their average scores are the same or not.

d. A researcher believes that unreliable measurement has been a key factor in the failure of 200 previous studies to support her hypothesis that enthusiasm for college football is a major determinant of support for the Republican Party. She has therefore written six questions that measure enthusiasm for college football and another five questions that measure support for the Republican Party. All items are measured on continuous scales.

A LISREL model with multiple indicators for each of the two underlying variables (enthusiasm for college football and support for the Republican Party) would probably be good. A less high-tech approach would be to just create factor scores for the two latent variables and use those in a regular OLS regression analysis.

e. With summer approaching, a fast food company is worried about attrition rates among its staff. It wants to hold on to employees for as long as possible. It has therefore drawn a random sample of employee records from the last five years. For each employee, it has recorded (a) whether the employee is still on the job (b) how long the employee was or has been employed with the company (c) the age of the employee when hired and (d) the employee's score on an attitudinal test taken at the time of hiring.

Event history analysis. You are interested in how long employees last.

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 <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. 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

See class notes and readings for the essays. I never write out detailed answers for these because then people could just copy them if I used the essays again!

IV. Extra Credit. (10 points)

Following are the results from an ordinal regression.

```
. des health female black age

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
health          byte   %9.0g       junk       1=poor,..., 5=excellent
female          byte   %8.0g                  1=female, 0=male
black           byte   %8.0g                  1 if race=black, 0 otherwise
age             byte   %9.0g                  age in years
```

```
. ologit health female black age

Iteration 0:   log likelihood = -15764.397
Iteration 1:   log likelihood = -14931.648
Iteration 2:   log likelihood = -14923.357
Iteration 3:   log likelihood = -14923.345

Ordered logit estimates                           Number of obs   =      10335
                                                  LR chi2(3)      =    1682.10
                                                  Prob > chi2     =     0.0000
Log likelihood = -14923.345                       Pseudo R2       =     0.0534

------------------------------------------------------------------------------
      health |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.1170992   .0355732    -3.29   0.001    -.1868214   -.0473769
       black |  -.8845093   .0583105   -15.17   0.000    -.9987959   -.7702228
         age |  -.0410673   .0010907   -37.65   0.000     -.043205   -.0389295
-------------+----------------------------------------------------------------
       _cut1 |  -4.910859    .074328          (Ancillary parameters)
       _cut2 |  -3.428162   .0648868
       _cut3 |  -2.004318   .0586633
       _cut4 |  -.7512595   .0561221
------------------------------------------------------------------------------
```

Briefly interpret the results. Then compute the probability that a 50 year old black female will report being in poor health.

Women, blacks and older people all tend to report worse health than do others. The sample estimate of Y* for a 50 year old black female is

$$Z_i = \sum_{k=1}^{K} \beta_k X_k = -.1170992 - .8845093 - 50 * .0410673 = -3.0549735$$

The probability that such an individual will be in poor health is

$$P(Y=1) = \frac{1}{1+\exp(Z_i - \delta_1)} = \frac{1}{1+\exp(-3.0549735+4.910859} = .135$$

Confirming with Stata (`prvalue` is part of Long and Freese's `spost` package of routines)

```
. prvalue, x(female=1 black=1 age=50)

ologit: Predictions for health

  Pr(y=poor|x):       0.1352
  Pr(y=fair|x):       0.2726
  Pr(y=average|x):    0.3331
  Pr(y=good|x):       0.1683
  Pr(y=excellen|x):   0.0908

    female   black     age
x=       1       1      50
```