Sociology 63993
Exam 3 Answer Key - DRAFT
May 6, 2009

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

1. One possible use for LISREL is to make multiple-group comparisons.

True.

2. If a model fails the Brant test, an ordered logit model should be used instead.

False. Brant tests are used to test the assumptions of ordered logit models. If an ordered logit model fails the Brant test, you may want to consider using some other ordinal regression model.

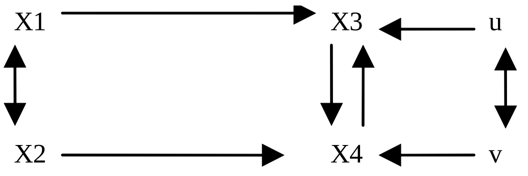
3. When comparing two populations, if the structural effect (Beta) is smaller in one population, the R^2 in that population will also be smaller.

False. Cross-population differences in the exogenous and/or residual variances could cause the R^2 to be equal or greater even if the beta is smaller.

4. A key advantage of mulitinomial logit over ordered logit is that the models tend to be more parsimonious and easier to interpret.

False. Multinomial logit models are less parsimonious (have more parameters) and hence may be harder to interpret.

5. A researcher is interested in the following model:



In order for this model to be identified, it must be the case that X1 is uncorrelated with X4.

False. X1 and X4 will be correlated because X1 is a cause of X3 which in turn is a cause of X4. Also, X1 is correlated with X2 which is a cause of X4.

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

II-1. (25 points): Public health researchers are concerned about the growing number of parents who are choosing not to have their children vaccinated against childhood diseases. They suspect that several factors may be responsible.

- They think that parents who are more concerned about autism will be less likely to get their children vaccinated, because they may (mistakenly) believe that vaccines lead to autism.
- Because of research in other areas that shows that blacks are less trusting of the health care system, they believe that blacks may also have less faith in the desirability of vaccination.

- Finally, the researchers are unsure about the effect of socio-economic status. They think that higher ses individuals will be more likely to get their children vaccinated; but after SES reaches a certain point, increases in SES may actually reduce the likelihood of getting vaccinated, as the more “elite” parents are more willing or able to work outside the traditional health care system.

They therefore collect data from more than 10,000 parents of young children on the following:

Variable	*Description*
vaccine	1 = children have been vaccinated, 0 = not vaccinated
autism	Concern about autism. Coded 0 if the parent is not worried about their child being autistic, 1 if they are
black	1 = black, 0 = not black
ses	Socio-Economic Status scale. The scale has been centered to have a mean of 0 and ranges from about -32 to 32
ses2	ses squared

They obtain the following results (some extraneous output is deleted):

```
. nestreg, lr: logit vaccine autism black ses ses2

Block  1: autism

Logistic regression                               Number of obs   =      10337
                                                  LR chi2(1)      =       4.56
                                                  Prob > chi2     =     0.0328
Log likelihood = -2633.5599                       Pseudo R2       =     0.0009

------------------------------------------------------------------------------
     vaccine |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      autism |  -.1985889   .0916108    -2.17   0.030    -.3781427   -.0190352
       _cons |   2.621071   .0437283    59.94   0.000     2.535365    2.706777
------------------------------------------------------------------------------

Block  2: black


Logistic regression                               Number of obs   =      10337
                                                  LR chi2(2)      =      62.41
                                                  Prob > chi2     =     0.0000
Log likelihood = -2604.6352                       Pseudo R2       =     0.0118

------------------------------------------------------------------------------
     vaccine |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      autism |  -.1627566    .092139    -1.77   0.077    -.3433458    .0178326
       black |  -.8073986   .0993375    -8.13   0.000    -1.002097   -.6127006
       _cons |   2.728819    .047132    57.90   0.000     2.636442    2.821196
------------------------------------------------------------------------------
```

```
Block  3: ses

Iteration 0:   log likelihood = -2635.8381
Iteration 1:   log likelihood = -2597.5867
Iteration 2:   log likelihood = -2592.8537
Iteration 3:   log likelihood = -2592.8225
Iteration 4:   log likelihood = -2592.8225

Logistic regression                               Number of obs   =      10337
                                                  LR chi2(3)      =       [1]
                                                  Prob > chi2     =     0.0000
Log likelihood = -2592.8225                       Pseudo R2       =     0.0163

------------------------------------------------------------------------------
     vaccine |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      autism |  -.3122938   .0972726    -3.21   0.001    -.5029446   -.1216431
       black |  -.8018448   .0996126      [2]   0.000    -.9970819   -.6066077
         ses |   .0204479   .0042197     4.85   0.000     .0121775    .0287184
       _cons |   2.772862   .0486335    57.02   0.000     2.677542    2.868182
------------------------------------------------------------------------------

Block  4: ses2

Logistic regression                               Number of obs   =      10337
                                                  LR chi2(4)      =      86.37
                                                  Prob > chi2     =     0.0000
Log likelihood = -2592.6509                       Pseudo R2       =     0.0164

------------------------------------------------------------------------------
     vaccine |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      autism |  -.3198695    .098012    -3.26   0.001    -.5119695   -.1277694
       black |  -.8006079   .0996401    -8.03   0.000    -.9958989   -.6053169
         ses |   .0207327   .0042933     4.83   0.000     .0123179    .0291474
        ses2 |   .0002002   .0003441     0.58   0.561    -.0004743    .0008746
       _cons |   2.756034   .0564051    48.86   0.000     2.645482    2.866586
------------------------------------------------------------------------------


    +----------------------------------------------------------------+
    | Block |        LL       LR     df   Pr > LR       AIC       BIC |
    |-------+--------------------------------------------------------|
    |     1 |  -2633.56     4.56      1    0.0328   5271.12  5285.607 |
    |     2 | -2604.635      [3]      1    0.0000   5215.27  5237.001 |
    |     3 | -2592.822    23.63      1    0.0000  5193.645  5222.619 |
    |     4 | -2592.651     0.34      1    0.5580  5195.302  5231.519 |
    +----------------------------------------------------------------+
```

Based on the printout above, answer the following.

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

Here are the uncensored sections:

```
Block  3: ses

Iteration 0:   log likelihood = -2635.8381
Iteration 1:   log likelihood = -2597.5867
Iteration 2:   log likelihood = -2592.8537
Iteration 3:   log likelihood = -2592.8225
Iteration 4:   log likelihood = -2592.8225
```

```
Logistic regression                               Number of obs   =      10337
                                                  LR chi2(3)      =      86.03
                                                  Prob > chi2     =     0.0000
Log likelihood = -2592.8225                       Pseudo R2       =     0.0163

------------------------------------------------------------------------------
     vaccine |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      autism |  -.3122938   .0972726    -3.21   0.001    -.5029446   -.1216431
       black |  -.8018448   .0996126    -8.05   0.000    -.9970819   -.6066077
         ses |   .0204479   .0042197     4.85   0.000     .0121775    .0287184
       _cons |   2.772862   .0486335    57.02   0.000     2.677542    2.868182
------------------------------------------------------------------------------

  +-------------------------------------------------------------------+
  | Block |        LL        LR     df   Pr > LR       AIC        BIC |
  |-------+-----------------------------------------------------------|
  |     1 |  -2633.56      4.56      1    0.0328   5271.12   5285.607 |
  |     2 | -2604.635     57.85      1    0.0000   5215.27   5237.001 |
  |     3 | -2592.822     23.63      1    0.0000  5193.645   5222.619 |
  |     4 | -2592.651      0.34      1    0.5580  5195.302   5231.519 |
  +-------------------------------------------------------------------+
```

To confirm,

[1] $L^2 = DEV_0 – DEV_M = (-2 * LL_0) – (-2 * LL_M) = (-2 * -2635.8381) – (-2 * -2592.8225) = 86.03$. (Other solutions are possible)

[2] $Z_{black} = -.8018448/ .0996126 = -8.0496$

[3] Incremental L^2 = Model L^2 $_{Unconstrained}$ – Model L^2 $_{Constrained}$ = 62.41 – 4.56 = 57.85

b. (7 points) The researchers decided that the third model (Block 3) was best. Explain why, and explain what this model tells you about the effects of race, SES and concern about autism on the likelihood of getting vaccinated. Make clear the extent to which the researchers' original ideas are supported or are not supported by the actual results.

As the chi-square contrasts at the end show, Model 2 is a significant improvement over Model 1, and Model 3 is a significant improvement over Model 2, but Model 4 is not significantly better than Model 3. Most, but not all, of the researchers' hypotheses are confirmed. Blacks, and those who are concerned about autism, are less likely to get vaccinated. Higher ses people are more likely to get vaccinated, but counter to expectations, the relationship is not curvilinear, i.e. the ses2 effect is not significant.

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

Autism	*Race*	*SES*	*Log odds*	*Odds*	*P(Vaccination))*
Concerned about autism	Black	0			
Not Concerned about autism	Not Black	0			

Note that the coefficient for autism is -.3122938, the coefficient for black is -.8018448, and the constant is 2.772862. Ergo,

Autism	*Race*	*SES*	*Log odds*	*Odds*	*P(Vaccination))*
Concerned about autism	Black	0	1.65872	5.2526	.840067
Not Concerned about autism	Not Black	0	2.772862	16.0044	.941192

d. (6 points) The researchers also ran the following:

```
. quietly logit vaccine autism black ses

. tab1 vaccine if e(sample)

-> tabulation of vaccine if e(sample)

    vaccine |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        729        7.05        7.05
          1 |      9,608       92.95      100.00
------------+-----------------------------------
      Total |     10,337      100.00

. estat clas

Logistic model for vaccine

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |      9608           729  |      10337
     -     |         0             0  |          0
-----------+--------------------------+-----------
   Total   |      9608           729  |      10337

Classified + if predicted Pr(D) >= .5
True D defined as vaccine != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)  100.00%
Specificity                     Pr( -|~D)    0.00%
Positive predictive value       Pr( D| +)   92.95%
Negative predictive value       Pr(~D| -)       .%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)  100.00%
False - rate for true D         Pr( -| D)    0.00%
False + rate for classified +   Pr(~D| +)    7.05%
False - rate for classified -   Pr( D| -)       .%
--------------------------------------------------
Correctly classified                        92.95%
--------------------------------------------------
```

Are you impressed by these results of the classification analysis? Why or why not?

The table is not very useful in this case. Yes, 92.95% get classified correctly, but that is because everybody gets classified as getting vaccinated. A substantial majority of people (92.95%) do get their children vaccinated, and even the people least likely to get vaccinated still have a predicted probability of more than 50% of doing so.

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. A researcher wants to obtain an unbiased estimate of the effect of religiosity on altruistic behavior. She has six items that measure each of the two concepts, all of which suffer from some degree of random measurement error.

This would be a good candidate for a LISREL model, which allows you to specify a measurement model and a structural equation model simultaneously. Probably less optimal but still worth considering would be to create scales from the two sets of items first and then use OLS regression.

b. A statistics professor is unsure whether using statistical software in class actually affects student performance and satisfaction. Students are therefore randomly divided into two sections. In one class, she will use statistical software, in the other she will not. In both classes, she will measure student satisfaction with the course, scores on exams (exams will be the same in both sections), and the number of days that each student attends class.

There is one binary independent variable (coded uses software/ does not use software) and multiple dependent variables. Manova would be a good choice, or perhaps an equivalent model estimated using LISREL.

c. Father Jenkins is trying to figure out how best to bolster support for his decision to bring Barack Obama to Notre Dame. Two letters have been prepared. The first stresses Notre Dame's role as a University and its openness to diverse lines of thought. The second stresses Barack Obama's accomplishments in many areas and why he deserves to be honored. One hundred randomly selected alumni will receive the first letter while another hundred randomly selected alumni receive the other. Recipients of the letters will then be asked whether they approve or disapprove of the decision to honor Obama at commencement.

There are lots of ways to approach this. Note that two binary variables are involved: type of letter (academic freedom/ praise Obama) and feelings about the decision (approve/ disapprove). You could therefore simply do a 2 x 2 crosstab and estimate the chi-square for the model of independence. If you want to be fancier, you could run a logistic regression with feelings about the decision as the DV and type of letter as the IV. You should NOT do a t-test or an OLS regression because the dependent variable is a dichotomy.

d. Ten years ago, a community group started a credit education program designed to teach new low income home owners how to manage their finances. Because more people wanted into the program than could be accommodated, participants were randomly selected from all applicants. Today, the group wants to see how effective the program has been in preventing or at least delaying home mortgage foreclosures. For each of the 500 people who applied to be in the program, the group knows (a) whether or not they actually participated in the program, and (b) the exact date their home was foreclosed on (if ever).

This is a good choice for event history analysis. You examine how participation in the program slows down the pace at which people default. Note too that the dependent variable is censored; just because people haven't defaulted after 10 years does not mean that they won't default eventually. When you are interested in determinants of the timing of events and when outcomes are censored, EHA is a good choice.

e. A politician supports gay marriage, but she isn't sure whether she should mention that in her campaigning. She will be giving talks in front of twenty fairly similar neighborhood groups over the next few days. In half her talks, she will mention her support for gay marriage, but in the other half she will ignore the issue. After each talk, those attending will be asked if they strongly support the candidate, somewhat support the candidate, or do not support the candidate.

The dependent variable is ordinal, so some type of ordinal regression model (e.g. ordered logit) is probably called for. If the assumptions of the ordinal model are violated, a less restrictive multinomial logit model could be used instead.

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

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. Present a substantive problem, real or hypothetical, where a nonrecursive model might be appropriate. Explain why you think the model should be nonrecursive. What problems might you encounter if you tried to use OLS regression to estimate this model? Even if you are correct in saying the model is nonrecursive, explain why it might be difficult for you to estimate your model.

See the course notes for ideas on each essay.

Appendix: Stata Code used in the exam

```
* Problem II-1.
webuse nhanes2f, clear
* Create the variables
gen vaccine = health >=2
gen autism = weight > 84
center height, gen(ses)
gen ses2 = ses^2
* Do the analysis presented in the exam
nestreg, lr: logit vaccine autism black ses ses2
quietly logit vaccine autism black ses
tab1 vaccine if e(sample)
estat clas
* Calculations for II-1 part c.
adjust black = 1 autism = 1 ses = 0, xb
adjust black = 1 autism = 1 ses = 0, exp
adjust black = 1 autism = 1 ses = 0, pr
adjust black = 0 autism = 0 ses = 0, xb
adjust black = 0 autism = 0 ses = 0, exp
adjust black = 0 autism = 0 ses = 0, pr
```