Sociology 63993
Exam 3 Answer Key
May 5, 2008

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

1. $R^2 = .5$ in both populations A and B. This means that the structural effect of X is the same in both populations.

False. R^2 is a function of the exogenous variance, the residual variance and the structural effect. It could therefore just be coincidence that it has the same value in two populations.

2. The dependent variable Y suffers from random measurement error. Therefore, when doing cross-population comparisons, it is best to focus on the standardized coefficients.

False. Look at the non-standardized coefficients, as they will not be biased by random measurement error.

3. A researcher obtains the following:

```
. logit  warmlt2  age , or nolog

Logistic regression                               Number of obs   =       2293
                                                  LR chi2(1)      =      36.17
                                                  Prob > chi2     =     0.0000
Log likelihood = -865.82744                       Pseudo R2       =     0.0205

------------------------------------------------------------------------------
     warmlt2 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.022106   .0037148     6.02   0.000     1.014851    1.029412
------------------------------------------------------------------------------
```

This means that, with each additional year of age, the log odds increase by slightly over 1.

False. Note that the odds ratio, not the coefficient, is being reported. Each one year increase in age will multiply the odds by a little over 1.

4. In logistic regression, as in OLS regression, Stata commands like `collin` can be used to test for multicollinearity.

True. Most of the same techniques for assessing collinearity can be used for both logistic and OLS regression.

5. A researcher wants to test the hypothesis

H_0: $\mu_{11} = \mu_{21}$
$\mu_{12} = \mu_{22}$
$\mu_{13} = \mu_{23}$

A nonrecursive model is called for.

False. This is the sort of hypothesis that can be tested via MANOVA.

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

II-1. (25 points) A medical researcher and a sociologist have teamed up to do work on the relationship between health, race, residence, and concentrations of lead in the body. The sociologist stresses that race and residence are very important. The medical researcher agrees and adds that the amount of lead in the body is very important. They therefore collect information on the following variables:

Variable	*Description*
poorhealth	Coded 1 if respondent has poor health, 0 otherwise
black	Coded 1 if black, 0 otherwise
rural	Coded 1 if respondent lives in a rural area, 0 otherwise
highlead	Coded 1 if the respondent has a high lead concentration in his or her body, 0 otherwise

They obtain the following results:

```
. nestreg, lr: logit poorhealth black rural highlead

Block  1: black

Iteration 0:   log likelihood = -12353.146
Iteration 1:   log likelihood = -12245.327
Iteration 2:   log likelihood = -12229.778
Iteration 3:   log likelihood =  -12229.69
Iteration 4:   log likelihood =  -12229.69

Logistic regression                               Number of obs   =      49400
                                                  LR chi2(1)      =     246.91
                                                  Prob > chi2     =     0.0000
Log likelihood =  -12229.69                       Pseudo R2       =     0.0100

------------------------------------------------------------------------------
  poorhealth |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |    .783832   .0466522    16.80   0.000     .6923955    .8752686
       _cons |  -2.716259   .0196896  -137.95   0.000     -2.75485   -2.677668
------------------------------------------------------------------------------

Block  2: rural

Iteration 0:   log likelihood = -12353.146
Iteration 1:   log likelihood =  -12079.44
Iteration 2:   log likelihood = -12057.058
Iteration 3:   log likelihood = -12056.937
Iteration 4:   log likelihood = -12056.937

Logistic regression                               Number of obs   =      49400
                                                  LR chi2(2)      =     592.42
                                                  Prob > chi2     =     0.0000
Log likelihood = -12056.937                       Pseudo R2       =     0.0240

------------------------------------------------------------------------------
  poorhealth |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |   1.007662    .048794    20.65   0.000     .9120278    1.103297
       rural |   .6951541   .0373445    18.61   0.000     .6219603    .7683478
       _cons |  -3.045964   .0284908  -106.91   0.000    -3.101805   -2.990124
------------------------------------------------------------------------------
```

```
Block  3: highlead

Iteration 0:   log likelihood = [1]
Iteration 1:   log likelihood = -12072.715
Iteration 2:   log likelihood = -12049.038
Iteration 3:   log likelihood = -12048.908
Iteration 4:   log likelihood = -12048.908

Logistic regression                               Number of obs   =      49400
                                                  LR chi2(3)      =       [2]
                                                  Prob > chi2     =     0.0000
Log likelihood = -12048.908                       Pseudo R2       =     0.0246

------------------------------------------------------------------------------
  poorhealth |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |   .9879906   .0490653     [3]    0.000     .8918244    1.084157
       rural |   .6990782   .0373575    18.71   0.000     .6258589    .7722975
    highlead |   .2798416   .0677367     4.13   0.000     .1470801    .4126031
       _cons |  -3.063828   .0288938  -106.04   0.000    -3.120459   -3.007197
------------------------------------------------------------------------------


  +----------------------------------------------------------------+
  | Block |        LL        LR     df  Pr > LR        AIC       BIC |
  |-------+--------------------------------------------------------|
  |     1 | -12229.69   246.91      1   0.0000  24463.38  24480.99 |
  |     2 | -12056.94   345.50      1   0.0000  24119.87   24146.3 |
  |     3 | -12048.91    16.06      1   0.0001  24105.82  24141.05 |
  +----------------------------------------------------------------+
```

Based on the printout above, answer the following.

a. (6 points) Fill in the missing items [1], [2] and [3]. (HINT: You either need no calculations, or very easy ones.)

[1] = LL0 = -12353.146 (it is the same as in blocks 1 and 2, where it is reported)

[2] = L^2 = 592.42 + 16.06 = 608.48 (Just add the L^2 for block 2 to the LR contrast for block 3 that is reported in the summary table)

[3] = z_{black} = .9879906/.0490653 = 20.136 (Just divide the coefficient by its standard error)

b. (7 points) The researchers decided that the last model (Block 3) was best. Explain why, and explain what this model tells you about the effects of race, residence and body lead on health.

Both the LR contrast and the z values in the final model confirm that each effect is statistically significant. The model says that blacks, people in rural areas, and those with high concentrations of lead are more likely to have poor health.

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

Race	*Residence*	*Lead*	*Log odds*	*Odds*	*P(Poorhealth)*
Black	Rural	Low Lead			
Not Black	Non Rural	Low lead			

Race	*Residence*	*Lead*	*Log odds = a + XB*	*Odds = exp(Log Odds)*	*P(Poorhealth) = Odds/(1 + Odds)*
Black	Rural	Low Lead	-1.37676	.252395	.20153
Not Black	Non Rural	Low lead	-3.06383	.046709	.044624

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

```
. tab1 poorhealth if e(sample)

-> tabulation of poorhealth if e(sample)

  RECODE of |
     health |
(1=excellen |
     t,...,  |
    5=poor) |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |     46,010       93.14       93.14
          1 |      3,390        6.86      100.00
------------+-----------------------------------
      Total |     49,400      100.00
```

```
. estat clas

Logistic model for poorhealth

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |         0             0  |          0
     -     |      3390         46010  |      49400
-----------+--------------------------+-----------
   Total   |      3390         46010  |      49400

Classified + if predicted Pr(D) >= .5
True D defined as poorhealth != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)    0.00%
Specificity                     Pr( -|~D)  100.00%
Positive predictive value       Pr( D| +)       .%
Negative predictive value       Pr(~D| -)   93.14%
--------------------------------------------------
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| -)    6.86%
--------------------------------------------------
Correctly classified                        93.14%
```

The medical researcher, who has little statistical background, was very impressed when he saw that 93.14% of the cases were correctly classified by the model. The sociologist, who has extensive training in statistics, was not very impressed. Explain why.

The model classified everyone as being in good health. Since only 6.86% of the sample had poor health, the model was bound to have a high success rate. Such results are not unusual when dealing with relatively rare events.

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. It is January 21, 2009. After Al Gore’s stunning decision to decline the nomination, Hillary Clinton successfully united her party and went on to win the Presidency in the largest Democratic landslide in history. But now, only hours after her inauguration, it is 3 a.m., the phone in the Oval Office is ringing – and Clinton must face the challenge she has spent a lifetime preparing for.

Clinton’s economic advisors have just learned that five of the largest banks in the country, along with mortgage lending giants Fannie Mae and Freddie Mac, are planning to declare bankruptcy at noon. Their poor decisions in the subprime loan market, combined with the disastrous consequences of the predatory and exploitive practices that some of them had engaged in, have left these institutions teetering on collapse. Clinton knows that many Americans will feel the government should do nothing to keep these companies afloat, and she shares the scorn for the corporate greed that has produced this crisis. At the same time, there is no doubt in her mind that inaction will result in millions of innocent people losing their homes, the further collapse of financial markets, and possibly even a depression the likes of which has not been seen since 1929.

She therefore is convinced that bold and decisive government action is needed. Anticipating this crisis, she prepared a plan months ago - but what is the best way to rally the American people to her side? By 6 a.m., she and her advisors have already crafted two radically different speeches in support of her plan – and by 8 a.m. 200 randomly selected Americans have been recruited for their feedback. Half of them will hear one of the speeches while the other half will hear the other. After hearing the talks, respondents will be asked to rate their support for the President’s plan on a five point scale, ranging from strongly agree to strongly disagree.

Armed with this knowledge, at 11 a.m. a calm, collected and determined Hillary Clinton will give the most important speech of her life before a worried nation.

Ordinal regression. The dependent variable is ordinal and the independent variable is a dichotomy.

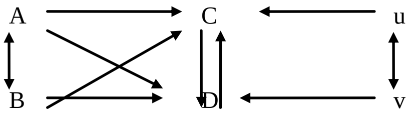
b. President Bush is discouraged by polls that show he is one of the least popular presidents in recent American history. He thinks the American people are not aware of many of the good things he has done, such as his contributions to the fight against HIV/AIDs in Africa. He wants to know whether efforts to publicize these accomplishments would help his popularity. Two hundred randomly selected Americans will therefore be asked to rate Bush on a scale that ranges from 0 to 100. They will then see a film about Bush's efforts in the war against AIDS. After the film, respondents will once again be asked to rate Bush.

Matched pairs T-Test. The same subjects are tested both before and after seeing the film and the dependent variable is continuous. An independent samples t-test is an inferior alternative because the DV is not continuous, but many would nonetheless use it in this case.

c. The Democratic primary campaign has lasted far longer than anyone expected. It has taken a toll on the staff of both campaigns, with many quitting along the way while others persisted. For future reference, the Obama campaign wants to know the factors that affect how long someone stays with the campaign. It suspects that factors such as age, marital status, and past experience may all be important. For 300 randomly selected staffers, it has therefore collected information on (a) whether the staffer is still with the campaign (b) how long the staffer was or has been with the campaign (c) the marital status of the staffer when hired (d) whether the staffer had previously worked on a presidential campaign, and (e) the age of the staffer when hired.

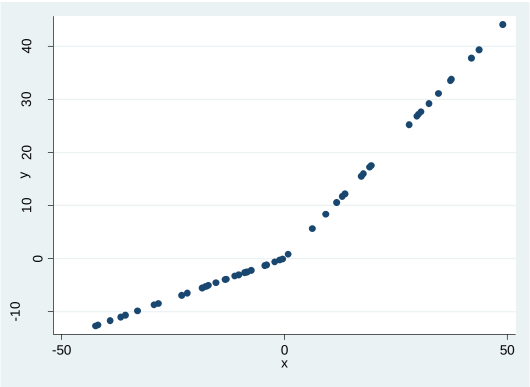
Event history analysis. The Obama campaign wants to know what factors affect how long staffers stay with the campaign.

d. A researcher believes that the following model describes her data. All variables are continuous.



God's help will be needed with this one, because the model is not identified. If the researchers can modify the model so it is identified, then something like 2SLS or LISREL could be used.

e. A researcher has collected data on two scales – x (interest in politics) and y (participation in political activities). She is puzzled that the regression of y on x does not show as clear a relationship as she expected. To clarify what is going on, she creates the following graph:



The effect of X on Y suddenly becomes much greater once X equals 0 or more. A piecewise regression model appears to be appropriate. The researcher might also consider adding X2 to the model if she thinks that makes more theoretical sense.

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. Path analysis first became popular in Sociology during the 1960s, and has evolved considerably since then.

a. In the early days of path analysis, standardized coefficients were widely used. Give two or three reasons why, in Sociology at least, that practice fell out of favor.

b. In the 1970s, the development of the LISREL program gave new life to path analysis. Discuss some of the key strengths of the LISREL method. Explain how LISREL made it possible to estimate important new sorts of models and how it provided an alternative means for estimating models that could also be approached via other methods.

See the lecture notes for ideas pertaining to each question.

APPENDIX: Stata code used in this exam

```
* II-1
webuse nhanes2f, clear
expand 10
recode health (2 3 4 5 = 0)(1 = 1), gen(poorhealth)
nestreg, lr: logit poorhealth black rural highlead
tab1 poorhealth if e(sample)
estat clas
* To double-check answers for part c:
adjust black = 1 rural = 1 highlead = 0, xb
adjust black = 1 rural = 1 highlead = 0, exp
adjust black = 1 rural = 1 highlead = 0, pr
adjust black = 0 rural = 0 highlead = 0, xb
adjust black = 0 rural = 0 highlead = 0, exp
adjust black = 0 rural = 0 highlead = 0, pr


* II-2e
clear
corr2data x, mean(0) sd(25) n(50)
corr2data e, sd(5)
gen y = .1*x + e if x <= 0
replace y = .7*x + e if x > 0
scatter y x
```