Sociology 593
Exam 3
May 6, 2004

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

1. If a model is recursive, the use of OLS regression will result in biased parameter estimates.
2. One unfortunate difference between logistic regression and OLS regression is that, with logistic regression, it is not possible to identify extreme outliers that may be affecting the results.
3. The log odds of an event occurring are 0. This means that there is no chance the event will happen.
4. In a multinomial logit model, if the number of cases in some categories is small, you may wish to combine two or more categories.
5. In population 1, R^2 = .6. In population 2, R^2 = .4. This means that the structural effect of X on Y is larger in population 1 than it is in population 2.

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

II-1. (25 points) The data used here are described in Hosmer, D. W. Jr. and S. Lemeshow. 2000. Applied Logistic Regression. 2nd ed. New York: John Riley and Sons. This example is adapted from the Stata 8 Reference Manual documentation on the `logistic` command. You can access these data from within Stata with the command

```
use http://www.stata-press.com/data/r8/lbw.dta
```

The following results are based on a study of risk factors associated with low birth weight. The variables are:

Variable	*Description*
low	Coded 1 if birth weight was low (less than 2500g), 0 otherwise
smoke	Coded 1 if the mother smoked during pregnancy, 0 otherwise
white	Coded 1 if white, 0 if black or other
whsmoke	= white * smoke

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?

b. Using Model 2, complete the following table:

Smoke	*White*	*Log odds*	*Odds*	*P(Low birth weight)*
Did not smoke	White			
Did not smoke	Black or Other			
Did smoke	White			
Did smoke	Black or Other			

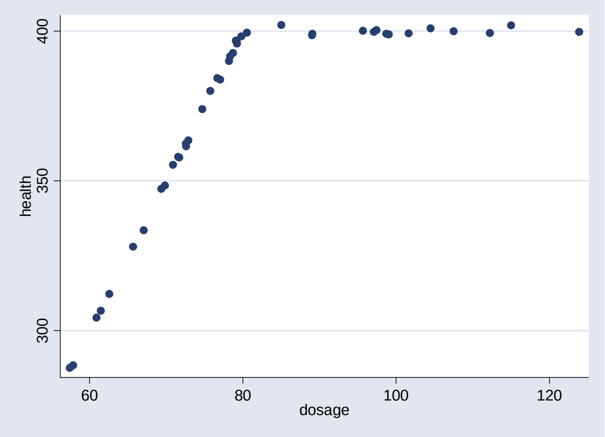
c. Three models are estimated. Which model do you think is best, and why? What does this model say about the effect of smoking on low birth weight? What does this model tell you about racial differences in the determinants of low birth weight?

```
. * Model 1
. logit low smoke

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood =  -114.9123
Iteration 2:   log likelihood =  -114.9023

Logit estimates                                   Number of obs   =        189
                                                  LR chi2(1)      =       4.87
                                                  Prob > chi2     =     0.0274
Log likelihood =  -114.9023                       Pseudo R2       =     0.0207

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |   .7040592   .3196386     2.20   0.028     .0775791    1.330539
       _cons |  -1.087051   .2147299    -5.06   0.000    -1.507914   -.6661886
------------------------------------------------------------------------------

. est store m1
. * Model 2
. logit low smoke white

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -110.10218
Iteration 2:   log likelihood = -109.98872
Iteration 3:   log likelihood = -109.98859

Logit estimates                                   Number of obs   =        189
                                                  LR chi2(2)      =      14.69
                                                  Prob > chi2     =     0.0006
Log likelihood = -109.98859                       Pseudo R2       =     0.0626

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |    1.11304   .3642634     3.06   0.002     .3990969    1.826983
       white |  -1.100347   .3644827    -3.02   0.003     -1.81472   -.3859743
       _cons |  -.7381552   .2378671    -3.10   0.002    -1.204366   -.2719442
------------------------------------------------------------------------------

. est store m2

. lrtest m2 m1

likelihood-ratio test                                  LR chi2(1)  =      9.83
(Assumption: m1 nested in m2)                          Prob > chi2 =    0.0017
```

```
. * Model 3
. logit low smoke white whsmoke

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -109.35259
Iteration 2:   log likelihood =   -108.861
Iteration 3:   log likelihood = -108.84969
Iteration 4:   log likelihood = -108.84968

Logit estimates                                   Number of obs   =        189
                                                  LR chi2(3)      =      16.97
                                                  Prob > chi2     =     0.0007
Log likelihood = -108.84968                       Pseudo R2       =     0.0723

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |   .6097656   .4935153     1.24   0.217    -.3575066    1.577038
       white |   -1.69282   .5802918    -2.92   0.004    -2.830171   -.5554685
     whsmoke |   1.140751   .7755588     1.47   0.141    -.3793163    2.660818
       _cons |  -.6097656   .2484736    -2.45   0.014    -1.096765   -.1227663
------------------------------------------------------------------------------

. est store m3

. lrtest m3 m2

likelihood-ratio test                                  LR chi2(1)  =      2.28
(Assumption: m2 nested in m3)                          Prob > chi2 =    0.1312
```

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. A pharmaceutical firm has invented a new drug designed to improve general health (measured on a scale that ranges from 200 to 500). It is trying to determine what the optimal dosage is; in particular, it wants to know whether, after some point, increases in dosage produce no additional benefit or even become harmful. Since it is unsure how to model the relationship between dosage and health, it has administered varying dosages to 41 test subjects (all of whom were equally healthy before the study) and constructed the following graph:



b. A researcher is interested in how husbands and wives influence each other's opinions. He believes that the husband's opinion on a subject is affected by his level of education, the socio-economic status of his parents when he was growing up, and by his wife's opinion on the subject. Similarly, he thinks the wife's opinion is determined by her level of education, the socio-economic status of her parents when she was growing up, and her husband's opinion. A random sample of 500 married couples will be interviewed for this study.

c. A psychologist is examining the factors that influence depression. Subjects are asked how depressed they are, with the possible responses being (1) not at all depressed (2) somewhat depressed (3) very depressed. The independent variables in the analysis include annual income in thousands of dollars, feelings of personal efficacy (measured on a scale that ranges from 1 to 50) and age in years.

d. A professor wonders whether students learn anything in her course. On the first day of class, students will complete a test that measures knowledge of the course's subject matter. On the last stay, students will take an equivalent test that measures their knowledge.

e. John Kerry is very concerned about the effectiveness of his campaign ads. To help him decide which strategies are most effective, he has had two sets of ads prepared. In one set of ads he appears with General Wesley Clark and stresses the accomplishments of his military and political career. In the other set, he appears with Governor Howard Dean and focuses on his positions on key issues. A group of randomly selected individuals will see the first set of ads while another randomly selected group will view the second set. Afterwards, respondents will be asked to rate, on scales ranging from 1 to 100, (1) how likely they think they are to vote for John Kerry, (2) How much they approve or disapprove of the job George Bush is doing as president, and (3) how much they liked the ads.

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

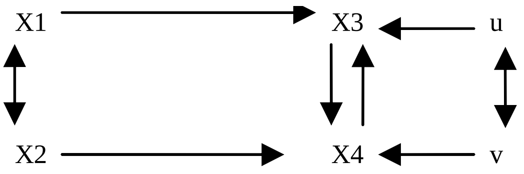
2. Path analysis first became popular in Sociology during the 1960s, and has evolved considerably since then.

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.

IV. *Extra Credit.* (10 points)

A researcher is interested in the following model:



Explain why the following situations would likely be problematic:

a. X2 is uncorrelated with X3.

b. You do not have a 2SLS program available to you, so instead you must use an OLS program for both stages.