Answer Key

Sociology 593
Exam 3
May 6, 1997

I. True-False. (30 points) Indicate whether the following statements are true or false. If false, briefly explain why.

1. A school district has set up a program to encourage reading. It believes that those who participate in the program will read more books and will spend less time watching TV than those who do not participate. Two-way Analysis of Variance is the appropriate statistical technique for testing hypotheses. F. Ns. Manova or Lisrel, because there are multiple DVs.
2. In a nonrecursive model, effects do not need to be unidirectional.
   T.
3. Increasing the sample size is one way of reducing the problem of under-identification.
   F. Greater sample size won't help.
4. A logistic regression is run, where X is the sole independent variable. The coefficient for X is 5. This means that each 1 unit increase in X produces a 5 percent increase in the probability of Y occurring. F. It will produce a 5 unit increase in the Log odds.
5. The logistic regression classification table is especially useful when dealing with rare events.
   F. The same prediction tends to get made for most or all cases, i.e. you always predict it won't happen.

II. Short answer. (15 pts each; 45 pts total; up to 10 points extra credit). Answer *three* of the following (up to 10 pts. extra credit for getting all 4 right).

1. A researcher has collected data on the following variables: CATHOLIC (1 = Catholic, 0 = not Catholic), FEMALE (1 = female, 0 = male) and EDUCATION (measured in years). Her dependent variable is CHURCH ATTENDANCE (1 = Attends church regularly, 0 = Does not attend regularly). When she runs her logistic regression, she gets

$b_{Catholic} = -1.5$ $b_{Female} = 1.0$ $b_{Education} = -0.25$ $a = 2.5$

Complete the following table:

Religion	Gender	Education	Log odds	Odds	P(Attend Church Regularly)
Catholic	Female	8	0	1.000	50.00%
Not Catholic	Female	8	1.5	4.482	81.76%
Catholic	Male	16	-3.0	0.050	4.74%
Not Catholic	Male	16	-1.5	0.223	18.24%

2. A researcher has collected data on home mortgage lending in St. Joseph County, IN. The variable FEMALE is coded 1 if the applicant and co-applicant (if any) are both female, 0 if either the applicant or co-applicant is male. APPLINC is income measured in $1000s, ~~FEMINC = FEMALE * APPLINC~~. DENIAL is coded 1 if the application was denied, 0 otherwise. Based on the 3 models that follow,

$\frac{958}{958+8170} = 10.50\%$

(1) indicate what percentage of applications were denied

(2) indicate whether there are significant gender differences in the determinants of home mortgage lending, $L^2 = 22.031$, 2 d.f. – Highly Sign.

(3) if so, offer a substantive discussion of what those differences are.

[HINT: Remember the parallels between -2LL and the Residual Sum of Squares.]

Logistic Regression - Model 1 - Full Sample

```
Number of cases included in the analysis: 9128

Dependent Variable..    DENIAL

Beginning Block Number  1.  Method: Enter

Variable(s) Entered on Step Number
1..        APPLINC    Applicant income

-2 Log Likelihood          5842.157
 Goodness of Fit         6.400E+10
 Cox & Snell - R^2            .031
 Nagelkerke - R^2             .031
```

Constrained χ^2

```
Classification Table for DENIAL
The Cut Value is .50
                    Predicted
                   .00     1.00      Percent Correct
                     0  I    1
Observed          +-------+-------+
   .00       0    I  8170 I     0 I   100.00%
                  +-------+-------+
  1.00       1    I   958 I     0 I     .00%
                  +-------+-------+
                            Overall    89.50%

---------------------- Variables in the Equation -----------------------

Variable            B        S.E.      Wald     df       Sig        R      Exp(B)

APPLINC          -.0300      .0021  202.0886      1     .0000   -.1807      .9704
Constant        -1.0392      .0760  187.0006      1     .0000
```

You can't tell whether the differences are in the intercept, slope, or both. But, it appears Applinc has a larger effect on women than men, Intercept appears smaller for women.

Logistic Regression - Model 2 - Males Only

```
FEMALE:        .00

     Number of cases included in the analysis: 7771

Dependent Variable..    DENIAL
```

```
Beginning Block Number  1.  Method: Enter

Variable(s) Entered on Step Number
1..        APPLINC    Applicant income

-2 Log Likelihood          4842.885
 Goodness of Fit           2.272E+10
 Cox & Snell - R^2              .029
 Nagelkerke - R^2               .029

---------------------- Variables in the Equation ----------------------

Variable            B        S.E.      Wald     df      Sig        R      Exp(B)

APPLINC          -.0287      .0023  162.3676      1     .0000   -.1778      .9717
Constant        -1.0545      .0870  146.9281      1     .0000
```

Male χ^2

Logistic Regression - Model 3 - Females Only

```
FEMALE:       1.00

Number of cases included in the analysis: 1357

Dependent Variable..    DENIAL

Beginning Block Number  1.  Method: Enter

-2 Log Likelihood          977.241
 Goodness of Fit           1582.996
 Cox & Snell - R^2              .050
 Nagelkerke - R^2               .093
```

Female χ^2

Male + Female L^2

4842.885
+ 977.241
5820.126 (unconstrained)

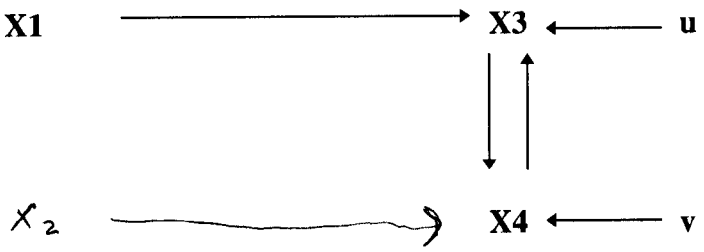
constrained − unconstrained = 22.031, 2 d.f.

```
---------------------- Variables in the Equation ----------------------

Variable            B        S.E.      Wald     df      Sig        R      Exp(B)

APPLINC          -.0688      .0098   48.9243      1     .0000   -.2117      .9335
Constant         -.3587      .2117    2.8691      1     .0903
```

3. Consider the following model:



X_3: G = 2, H = 1. H < G, so not identified

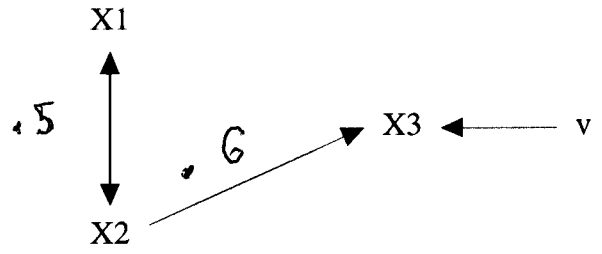
H_4: G = 1, H = 1. H ≥ G, so identified

Explain why you agree or disagree with the following statement: The X4 equation is identified, but the X3 equation is under-identified.

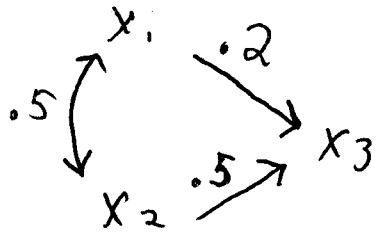
$X_3 = \beta_{31} X_1 + \beta_{34} X_4 + u$

$X_4 = \beta_{43} X_3 + v$

4. A researcher believes in the following model:



Estimated model:



implies $r_{13} = .3$, when it is really .45

A sample of 100 cases is collected. When she regresses X3 on X1 and X2, she gets $b_{31} = .2$, $b_{32} = .5$, $r_{12} = .5$. All variables are in standardized form. Test whether the over-identifying restriction in her preferred model appears reasonable.

$r_{12} = .5 \quad r_{13} = .45 \quad r_{23} = .6$

III. *Essay.* (25 points) Answer *one* of the following questions.

1. Often the dependent variable of interest is a dichotomy (such as whether a baby died within the first year of life). What general problems are created when we have such a dependent variable and attempt to apply OLS multiple regression to predict its value? Discuss the strengths and weaknesses of WLS and Logistic Regression as means for dealing with such variables.

2. Several assumptions are made when using OLS regression. Discuss TWO of the following. What does the assumption mean? When might the assumption be violated? What effects do violation 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.]

a. The effects of the independent variables are linear
b. Errors are homoskedastic
c. Variables are measured without error
d. The X's (independent variables) are uncorrelated with the residuals

$R^2_c = .6^2 = .36$

$R^2_u = b_1'^2 + b_2'^2 + 2b_1'b_2'r_{12} = .04 + .25 + .10 = .39$

$$F = \frac{(R^2_u - R^2_c)(N-K-1)}{(1-R^2_u)*J} = \frac{.03 * 97}{.61} = \frac{2.91}{.61} = 4.77$$

which is sign.