Sociology 593
Exam 3 Answer Key
May 2, 2003

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

1. If a model is underidentified, it must be theoretically incorrect.

False. It may be impossible to estimate but it could still be theoretically correct.

2. Unlike OLS, multicollinearity is not a problem in logistic regression

False. The problems of multicollinearity are basically the same with both methods.

3. Event Count Models should be used for dependent variables that can take on extremely large values, e.g. in the millions.

False. Event count models should be used with variables where the counts are fairly small, i.e. rare events.

4. For ordinal dependent variables, the key problem with using multinomial logistic regression is a loss of efficiency.

True. Assuming other assumptions are met, ordinal regression will be more efficient because fewer parameters are estimated. Multinomial logistic regression fails to take advantage of the fact that the categories are ordered.

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

II-1. A researcher is examining home mortgage lending in the state of South Dakota during the year 2001. She wants to examine how characteristics of applicants, neighborhoods and lenders affect the likelihood of a loan application being denied. She is particularly interested in determining whether subprime lenders (lenders who specialize in higher interest loans) have higher denial rates than do other types of lenders.

For each home mortgage application made to a lender in the state during that year, she has measured the following variables: DENIAL (coded 1 if the application was denied, 0 if it was approved), BLACK (coded 1 if the applicant was black, 0 otherwise), MINPCT90 (the percentage of neighborhood residents that were minority group members, as measured in the 1990 census) and SUBPRIME (coded 1 if the application was made to a subprime lender, 0 otherwise).

Based on the following printout, answer the following.

a. What do DEV_M, G_M, and DEV_0 equal?

From the printout, you can see that DEV_M (aka -2ll) = 27176.731. G_M (aka the Model chi square) is 9169.258. You can compute $DEV_0 = DEV_M + G_M = 36345.989$.

b. What does McFadden's Pseudo R^2 equal?

McFadden's Pseudo $R^2 = G_M/DEV_0 = 9169.258/36345.989 = .25228$.

c. What is the probability that a Black applying for a loan from a subprime lender in a neighborhood that is 10% minority (i.e., MINPCT90 = 10) will have their application denied?

According to the parameter estimates,

Log odds = -2.440 + 2.527 * Subprime + .907 * Black + .052 * MinPct90.

In this specific case, subprime = 1, black = 1, minpct90 = 10, so

Log odds = -2.440 + 2.527 + .907 + .52 = 1.514

Odds = exp (log odds) = exp(1.514) = 4.54487

P(Denial) = Odds/ (1 + Odds) = 4.54487/5.54487 = .8197

So, there is almost an 82% chance that such an individual will have their application denied.

d. Briefly interpret the results. Overall, what percentage of the respondents had their applications denied? How do characteristics of individuals, neighborhoods and lenders affect the likelihood of a loan being denied? Be sure to indicate the parts of the printout that support your conclusions, and why.

From the classification table, you can tell that the observed number of acceptances = 21,567 + 3,374 = 24,941. The observed number of denials = 2,477 + 5,461 = 7,938. The total number of applicationsis 24,941 + 7,938 = 32,879. The denial rate is therefore 7,938/32,879 = 24.14%.

The positive coefficients tell you that being black and applying for a loan from a subprime lender all increase the likelihood that the application will be denied. Also, the more heavily minority the neighborhood is, the less likely it is that the loan will be approved.

Logistic Regression

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	9169.258	3	.000
	Block	9169.258	3	.000
	Model	9169.258	3	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	27176.731	.243	.364

Classification Table[a]

			Predicted		
			DENIAL		Percentage Correct
	Observed		.00	1.00	
Step 1	DENIAL	.00	21567	3374	86.5
		1.00	2477	5461	68.8
	Overall Percentage				82.2

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	SUBPRIME	2.527	.031	6669.091	1	.000	12.511
	BLACK	.907	.205	19.617	1	.000	2.478
	MINPCT90	.052	.002	511.724	1	.000	1.053
	Constant	-2.440	.025	9280.539	1	.000	.087

a. Variable(s) entered on step 1: SUBPRIME, BLACK, MINPCT90.

II-2. 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 company wants to determine whether or not providing daycare to it employees will increase productivity and improve morale. A daycare center will be built at its downtown office, while the Western branch office will continue to be without daycare. After a year, the company will use surveys and personnel records to determine, for each employee at both sites, the number of days missed from work during the year, and the level of job satisfaction.

Manova or Lisrel would be appropriate. There is a single independent variable (whether the site has daycare or not) whose effect on two possible outcomes is being examined (productivity as measured by absenteeism and level of job satisfaction.)

b. A psychologist wants to get unbiased estimates of the effect of leadership ability on personal popularity. Data are collected from 300 respondents. Subjects are asked eight questions that pertain to leadership ability and another five questions that tap the popularity of the respondent.

Lisrel. You have multiple measures of two underlying latent variables. You can use Lisrel to get unbiased estimates of the structural effects.

c. College football continues to be rocked by scandals. A researcher wants to know why some schools, such as the University of Alabama, seem to have one scandal after another, while many other schools have no scandals at all. She therefore collects data on the top 100 college football programs from across the country. She adds up the number of major scandals that have occurred at each school during the past 20 years. Using interval-level scales that she has developed herself or gotten from elsewhere, she also collects data on the degree of institutional oversight of athletics, aptitude test scores of freshman athletes, and alumni pressure to win. Among other things, she finds that the number of major scandals ranges from as few as 0 at a majority of universities to as many as 5 at others.

Event count models. We are trying to determine what affects the number of times scandals occur at an institution.

d. The graduate school is concerned about the amount of time it takes students to complete their degrees. It wants to find out what it is that causes some students to take longer than others. It randomly selects 100 students who began their studies 6 years ago. For each student, the graduate school collects data on GRE Scores, College grade point average, country of origin (coded 1 = foreign student, 0 = United States), and when and if the Ph.D. has been completed by the student.

Event history Analysis. We are looking at what affects the timing of an event, in this case getting a Ph.D. Also, some cases will be censored, i.e. some students won't have achieved their Ph.D. yet but they will sometime in the future.

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. Event count models
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 the course notes for discussions of each question.

IV. *Extra Credit.* (10 points.) The same researcher as in problem II-1 also wants to examine whether the effect of income on denial rates is different for blacks than for whites. She uses the variables DENIAL (coded 1 if the application was denied, 0 if it was approved), BLACK (coded 1 if the applicant was black, 0 otherwise), and APPLINC (applicant income measured in thousands of dollars). Based on the three models that follow,

1. Indicate whether there are significant racial differences in the determinants of home mortgage lending.

We proceed as we would in an OLS problem, substituting chi-square statistics for residual sums of squares and incremental F tests.

Model I is the constrained model, in that the intercept and the slope must be the same for both racial groups, since only one model is estimated for both populations. In Model 1, the deviance (-2LL) is 34310.677. This is the constrained chi-square, i.e. it is the equivalent of the constrained error sum of squares in OLS.

Models II and III are unconstrained in that the slope and intercept are free to differ across groups. Their respective deviances are 34104.914 and 181.422. Similar to OLS, you add these up to get the unconstrained deviance and get 34286.336.

To test whether the constrained and unconstained models differ, you subtract the unconstrained deviance from the constrained deviance and get 34310.677 – 34286.336

and get 24.341. D.F. = 2, as 2 parameters are estimated in the constrained model compared to 4 parameters in the unconstrained model. This is highly significant.

As is the case with OLS, this is a global test of whether ANY parameters (slopes or intercepts) differ across groups.

2. If so, offer a substantive explanation of those differences.

The results tell us that the models are not identical across racial groups. Unfortunately, they do not tell us whether the differences are in the intercepts, slopes, or both. Just by eyeballing the numbers, it would appear that the constants are very different across groups, with blacks being much more likely to be denied than others are. The estimated effect of income is a little larger for blacks than for others, which may or may not be a good thing. It could, for example, mean that poor blacks will have more trouble getting a loan than poor whites will.

3. Explain why the analysis, as presented, does not allow the researcher to answer the above question as completely as would be desirable. Tell her what she should do instead to be able to get a more detailed answer.

[HINT: You'll have to do some improvising here. You know how to approach this sort of problem in regular OLS regression. To do it with Logistic regression, think about the similarities between the Deviance and the OLS residual sums of squares, and between an incremental F test and a chi-square test. Also, make sure you get the degrees of freedom right and are clear as to what hypotheses these models are testing.]

As implied above, the key problem is that you know there are differences between blacks and whites, but she doesn't know where they are. A better approach would probably be to run a single model for both groups, adding a dummy variable for BLACK and an interaction term for BLACK * APPLINC. As the following shows, once you do this, you find that the intercepts differ significantly across groups but the slopes do not. Ergo, there is a penalty for being black, but the effect of income for different racial groups does not significantly differ. Of course, there are fairly few blacks applying for loans in South Dakota, which makes it more difficult for results to be significant.

Logistic Regression – Model IV – Interaction terms

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	24.341	2	.000
	Block	24.341	2	.000
	Model	3764.797	3	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	34286.336	.102	.154

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	APPLINC	-.040	.001	2450.214	1	.000	.961
	BLACK	1.326	.494	7.210	1	.007	3.767
	BLACK by APPLINC	-.016	.015	1.074	1	.300	.984
	Constant	.394	.031	157.984	1	.000	1.482

a. Variable(s) entered on step 1: BLACK, BLACK * APPLINC .

Logistic Regression – Model I – Full Sample

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	35153	100.0
	Missing Cases	0	.0
	Total	35153	100.0
Unselected Cases		0	.0
Total		35153	100.0

a. If weight is in effect, see classification table for the total number of cases.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	3740.456	1	.000
	Block	3740.456	1	.000
	Model	3740.456	1	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	34310.677	.101	.153

Classification Table[a]

	Observed		Predicted DENIAL .00	Predicted DENIAL 1.00	Predicted Percentage Correct
Step 1	DENIAL	.00	26938	73	99.7
		1.00	7996	146	1.8
	Overall Percentage				77.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	APPLINC	-.040	.001	2475.228	1	.000	.961
	Constant	.403	.031	166.871	1	.000	1.497

a. Variable(s) entered on step 1: APPLINC.

Logistic Regression – Model II – NonBlacks Only

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	35010	100.0
	Missing Cases	0	.0
	Total	35010	100.0
Unselected Cases		0	.0
Total		35010	100.0

a. If weight is in effect, see classification table for the total number of cases.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	3700.350	1	.000
	Block	3700.350	1	.000
	Model	3700.350	1	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	34104.914	.100	.152

Classification Table[a]

			Predicted		
			DENIAL		Percentage Correct
	Observed		.00	1.00	
Step 1	DENIAL	.00	26895	44	99.8
		1.00	7976	95	1.2
	Overall Percentage				77.1

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	APPLINC	-.040	.001	2450.214	1	.000	.961
	Constant	.394	.031	157.984	1	.000	1.482

a. Variable(s) entered on step 1: APPLINC.

Logistic Regression – Model III – Blacks Only

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	143	100.0
	Missing Cases	0	.0
	Total	143	100.0
Unselected Cases		0	.0
Total		143	100.0

a. If weight is in effect, see classification table for the total number of cases.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	16.811	1	.000
	Block	16.811	1	.000
	Model	16.811	1	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	181.422	.111	.148

Classification Table[a]

	Observed		Predicted DENIAL .00	Predicted DENIAL 1.00	Predicted Percentage Correct
Step 1	DENIAL	.00	40	32	55.6
		1.00	20	51	71.8
	Overall Percentage				63.6

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	APPLINC	-.055	.015	13.423	1	.000	.946
	Constant	1.720	.493	12.173	1	.000	5.585

a. Variable(s) entered on step 1: APPLINC.