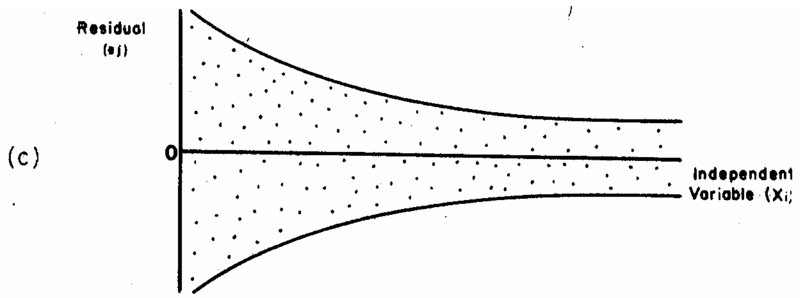
Sociology 593
Exam 1 Answer Key
February 14, 2003

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

1. A key advantage of Cohen and Cohen's dummy variable adjustment for missing data is that it uses as much data as possible while still producing unbiased estimates of parameter coefficients.

False. As Allison points out, the technique produces biased estimates of parameter coefficients.

2. A researcher plots his data and observes the following:



Assuming the researcher is confident her model is correctly specified, a GQ test is appropriate.

True.

3. When making comparisons across groups, metric (nonstandardized) coefficients are generally superior to standardized coefficients.

True.

4. With a dummy variable, the value of the coefficient indicates how much that group differs from the overall mean.

False. It indicates how much that group differs from the reference group, i.e. the group that is coded 0 on all the dummies.

5. When multicollinearity is present, one or more independent variables should be dropped from the analysis.

False. You may want to do this, but there are several other options, e.g. compute scales, test groups of variables rather than individual variables.

II. *Short answer.* Answer two of the following. (20 points each, 40 points total, up to 10 points extra credit.)

1. Here is part of the data obtained by a researcher:

X	Y
5	14
7	16
8	15
9	17
9	18
10	19
11	200
12	22
13	23
14	25

a. What problem may be present? What effect might this problem have on her analysis?

The 7th case is an extreme outlier. The slope, intercept, and other statistics may be heavily affected by this single case.

b. The entire data set is quite large. Discuss two or more procedures she could use in SPSS to reduce the likelihood of such a problem being missed.

She could use the Frequency or Examine procedures to identify extreme outliers. The Casewise diagnostics option on the regression command could also identify which cases are outliers and how much effect they are having on parameter estimates.

c. What solutions should the researcher consider using to solve this problem? Be sure to explain when and why a particular solution would be appropriate.

Begin by checking the coding. My first guess would be that an extra zero has been added by mistake. Also, make sure 200 is not supposed to be an MD code. If 200 is a legitimate value, you might want to reconsider whether this case falls within the population of interest and if not, it could be dropped. If possible, an ideal solution would be to include additional variables which cause the case to not be such an outlier.

2. A researcher obtains the following data:

Gender	X1	X2	Y
Male	17	Missing	31
Male	13	Missing	28
Male	15	Missing	23
Male	20	Missing	32
Female	Missing	38	17
Female	Missing	40	16
Female	Missing	46	20
Female	Missing	53	24

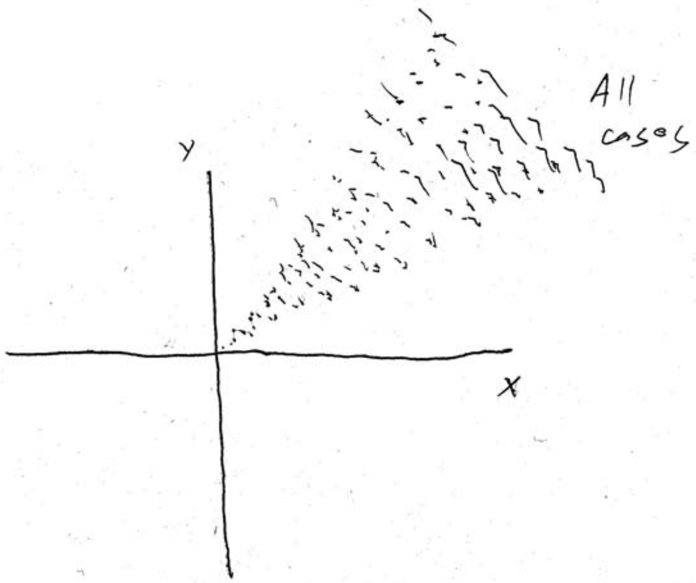
Explain what you think is likely causing the pattern of missing data. Discuss at least two methods (other than Cohen and Cohen's dummy variable adjustment) that you think would be inappropriate for dealing with the problem. What would you recommend that the researcher do in this case?

My guess is that X1 was asked only of men and X2 was asked only of women. If they are actually the same question, just asked in different places or in slightly different ways, X1 and X2 could be combined into a single item. Alternatively, it may just be necessary to run different models for men and women.

In terms of other alternatives – listwise deletion would clearly be bad because all cases would be excluded. Pairwise deletion would also be problematic since no case has data on both X1 and X2. Substituting the mean or other imputation techniques would almost certainly be problematic since it is highly unlikely the data are simply missing at random.

Whatever strategy you choose, you should look at the original questions and codebook first, so you understand why the data are missing.

3. A researcher obtains the following results when she analyzes her data. Indicate (i) what problem appears to be present (and how you can tell that from the information given) (ii) why you should be concerned about the problem, i.e. what harmful effects might it have when estimating regression models, and (iii) possible solutions. When discussing solutions, be sure to look carefully at the information presented; if, in this particular case, some solutions appear to be better than others, explain why.



Regression Line for Men: Y = 0 + 5X

Regression Line for Women: Y = 0 + 3X

At first glance, the data appear to be heteroskedastic. This can cause standard errors to be off and significance tests and confidence intervals inaccurate. A careless researcher might be tempted to use Weighted Least Squares.

Upon closer examination, however, you see that the slope coefficients are very different for men and women. Hence, for low values of X, the observed values are fairly close to each other, but as X gets bigger and bigger the male and female cases get spread more and more apart. If these gender differences are ignored, your predictions for men will be too low and your predictions for women too high. Rather than use WLS, you should estimate separate models for men and women, or else include interaction terms that capture the gender differences. Most or all of the apparent heteroskedasticity will likely then go away.

III. Computation and interpretation. (40 points total)

Credit scores play a major role in determining whether applications for loans are approved or denied. A researcher has collected data from 900 individuals on the following variables:

Variable	Description
CrScore	Credit score, measured on a scale ranging from 300 to 1000. Higher scores indicate a stronger credit rating
Black	Coded 1 if the applicant is black, 0 otherwise
Income	Annual income, measured in thousands of dollars
Debt	Total debt (excluding home mortgage loans) owed by the individual, measured in thousands of dollars (includes credit card debt and car loans)
YrEmploy	Number of years employed at current job

She obtains the following results.

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
CRSCORE	642.0000	77.00000	900
BLACK	.1000	.09000	900
INCOME	37.0000	11.00000	900
DEBT	14.0000	4.00000	900
YREMPLOY	11.0000	3.60000	900

Correlations

		CRSCORE	BLACK	INCOME	DEBT	YREMPLOY
Pearson Correlation	CRSCORE	1.000	-.430	.600	-.300	.500
	BLACK	-.430	1.000	-.400	.400	-.200
	INCOME	.600	-.400	1.000	.100	.400
	DEBT	-.300	.400	.100	1.000	-.200
	YREMPLOY	.500	-.200	.400	-.200	1.000
Sig. (1-tailed)	CRSCORE	.	.000	.000	.000	.000
	BLACK	.000	.	.000	.000	.000
	INCOME	.000	.000	.	.001	.000
	DEBT	.000	.000	.001	.	.000
	YREMPLOY	.000	.000	.000	.000	.
N	CRSCORE	900	900	900	900	900
	BLACK	900	900	900	900	900
	INCOME	900	900	900	900	900
	DEBT	900	900	900	900	900
	YREMPLOY	900	900	900	900	900

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.430[a]	.185	.184	69.55652	.185	203.705	1	898	.000	
2	.730[b]	.532	.530	52.78230	.347	221.488	3	895	.000	1.979

a. Predictors: (Constant), BLACK

b. Predictors: (Constant), BLACK, YREMPLOY, DEBT, INCOME

c. Dependent Variable: CRSCORE

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	985548.6	1	985548.569	203.705	.000[a]
	Residual	4344622	898	4838.109		
	Total	5330171	899			
2	Regression	2836727	4	709181.718	**[1]**	.000[b]
	Residual	2493444	895	2785.971		
	Total	5330171	899			

a. Predictors: (Constant), BLACK

b. Predictors: (Constant), BLACK, YREMPLOY, DEBT, INCOME

c. Dependent Variable: CRSCORE

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	678.789	3.467		**[2]**	.000			
	BLACK	-367.889	25.776	-.430	-14.273	.000	-.430	-.430	-.430
2	(Constant)	537.569	9.860		54.519	.000			
	BLACK	-59.910	24.474	-.070	-2.448	.015	-.430	-.082	-.056
	INCOME	3.563	.203	.509	17.590	.000	**[3]**	.507	.402
	DEBT	**[4]**	.525	-.278	-10.172	.000	-.300	-.322	-.233
	YREMPLOY	4.853	.555	**[5]**	8.737	.000	.500	.280	.200

a. Dependent Variable: CRSCORE

Excluded Variables[b]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	INCOME	.510[a]	18.101	.000	.517	.840
	DEBT	-.152[a]	-4.689	.000	-.155	.840
	YREMPLOY	.431[a]	15.861	.000	.468	.960

a. Predictors in the Model: (Constant), BLACK

b. Dependent Variable: CRSCORE

a) (10 pts) Fill in the missing quantities [1] – [5].

Here are uncensored parts of the printout:

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	985548.569	1	985548.569	203.705	.000[a]
	Residual	4344622.315	898	4838.109		
	Total	5330170.884	899			
2	Regression	2836726.874	4	709181.718	254.555	.000[b]
	Residual	2493444.010	895	2785.971		
	Total	5330170.884	899			

a. Predictors: (Constant), BLACK

b. Predictors: (Constant), BLACK, YREMPLOY, DEBT, INCOME

c. Dependent Variable: CRSCORE

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			Correlations		
Model		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part
1	(Constant)	678.789	3.467		195.789	.000			
	BLACK	-367.889	25.776	-.430	-14.273	.000	-.430	-.430	-.430
2	(Constant)	537.569	9.860		54.519	.000			
	BLACK	-59.910	24.474	-.070	-2.448	.015	-.430	-.082	-.056
	INCOME	3.563	.203	.509	17.590	.000	.600	.507	.402
	DEBT	-5.342	.525	-.278	-10.172	.000	-.300	-.322	-.233
	YREMPLOY	4.853	.555	.227	8.737	.000	.500	.280	.200

a. Dependent Variable: CRSCORE

To confirm that SPSS got it right:

[1] F = MSR/MSE = 709181.718/2785.971 = 254.555
[2] $t_{Constant}$ = Constant/Standard error of Constant = 678.789/3.467 = 195.789
[3] $r_{Income, Crscore}$ = .600 (see the correlations table)
[4] $b_{Debt} = t_{debt} * s_{bDebt}$ = -10.172 * .525 = -5.3403; or,
$b_{Debt} = b'_{Debt} * s_{Crscore}/s_{Debt}$ = -.278 * 77/4 = -5.352
[5] $b'_{Yremploy} = b_{Yremploy} * s_{Yremploy}/s_{Crscore}$ = 4.853 * 3.6/77 = .227

b) (5 pts) The Durbin-Watson statistic is reported (see the model summary). Is it appropriate given the nature of this analysis? Does the value of the statistic give the researcher reason for concern?

Durbin-Watson is used with time series data, e.g. repeated observations of manufacturing activity. These data are not time series so DW is not appropriate. Even if it was, the DW statistic is close to 2, which indicates there is probably no problem.

c) (5 pts) If the researcher had used forward stepwise regression, what variable would have been entered first? If she were to now use backward stepwise regression, what variable would be deleted? Explain your answer.

Income would have been entered first because it has the largest bivariate correlation with CrScore. No variables would now be deleted, because all of the variables that are in the equation are highly significant.

d) (5 pts) Prior to her analysis, the researcher was concerned that multicollinearity might be a problem with the data. Based on her results, do you think she has reason to be worried? Explain why or why not.

There are no apparent indications that multicollinearity is a problem. The F and the T values are all significant. The bivariate correlations of the Xs with each other are not that high. The N, 900, is fairly large.

e) (5 pts) The researcher believes that Debt is poorly measured in her data, i.e. there is a great deal of random measurement error in this item. Explain to her what impact this may have on her analyses.

Poor measurement will cause parameter estimates to be biased. If this were a bivariate regression, the effect of debt would be biased downward. Since the analysis is multivariate, all we can safely say is that the coefficients will somehow be biased.

f) (10 pts) Interpret the results. What percentage of the respondents are black? Who has better credit scores – blacks or whites? What might account for the fact that the effect of black diminishes once income, debt, and years employed are added to the model? Briefly explain what the results imply about the effects of each independent variable on credit scores.

Ten percent of the respondents are black (see the mean of black in the descriptive statistics. As the negative correlation between black and crscore indicates, whites have better credit scores. Part of the reason blacks have lower credit scores is probably because they also have lower incomes, more debt, and have been employed fewer years; hence, once these variables are controlled, the direct effect of race diminishes (but the fact that it doesn’t disappear completely means there is some other reason blacks still score lower. Perhaps this reflects discrimination, biases in the way credit scores are computed, or other unmeasured variables.)

The results suggest that being black, having a lower income, being more in debt, and having fewer years of employment all cause your credit score to be lower.