Sociology 592 - Research Statistics I
Final Exam Answer Key - DRAFT
December 16, 2003

Where appropriate, show your work - partial credit may be given. (On the other hand, don't waste a lot of time on excess verbiage.) Do not spend too much time on any one problem. You are free to refer to anything that was demonstrated in the homework or handouts.

1. (5 points each, 20 points total). For <u>each</u> of the following, indicate whether the statement is true or false. If you think the statement is false, indicate how the statement could be corrected.

NOTE: These are all pretty easy, but you could waste a great deal of time on some of them or make stupid mistakes if you don't happen to see what the easiest way to approach each problem is.

a. The correlation between Y and X1 is positive. Therefore, if Y is regressed on X1, X2, and X3, the unstandardized coefficient for X1 will also always be positive (or else zero).

False. If suppressor effects are present, a variable can have a positive correlation with Y and yet still have a negative once other variables are controlled for (and vice-versa). This can occur, for example, with compensatory programs that are designed to offset some sort of disadvantage. Participants in the programs still have lower scores than non-participants (leading to a negative correlation) but if the programs are beneficial (i.e. have a positive effect) the gap will not be as large as it would be in the absence of the program.

b. A researcher has estimated the following two models:

$$\text{Model 1: } \hat{y} = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$
$$\text{Model 2: } \hat{y} = a + b_1X_1 + b_2X_2 + b_5X_5$$

Assuming that the same cases are analyzed in both models, the R^2 for Model 1 will be at least as large as the R^2 for Model 2.

False. The two models are not nested, i.e. you do not get one model just by adding additional variables to the other. While Model 1 has more variables, it can still have a lower R^2 if X5 has more impact than X3 and X4 do.

c. A researcher obtains the following results:

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

Variable	B	SE B	Beta	Correl	Part Cor	Partial	T	Sig T
STATUS	8.351967	1.447318	.278956	.400000	.222883	.250827	5.771	.0000
MALE	10.698092	1.450267	.286709	.330000	.284911	.314422	7.377	.0000
ANOMIE	1.282964	.402837	.153956	.350000	.123009	.141563	3.185	.0015
(Constant)	37.213275	3.878287					9.595	.0000

If she is using backwards stepwise regression with $\alpha = .05$, on the next step R^2 will decline by $.123009^2$.

False. There won't be a next step. All variables are statistically significant so do not go any further.

d. In a bivariate regression with N = 500, the null and alternative hypotheses are

H_0: $\beta = 0$
H_A: $\beta > 0$

It is found that b = 37 and F = 36. The null hypothesis should be rejected.

True. In a bivariate regression, $F = T^2$. So, in this case, T must equal 6 or -6. Because b is positive, we know T is positive, and a T value of 6 is highly significant and in the correct direction.

2. Short answer problems. (10 points each, 30 points total, up to 5 points extra credit). Answer three of the following. You will get up to five points extra credit if you can solve all four problems.

a. In a multivariate regression, n = 150, k = 12, R^2 = .25, SSE = 144. Construct the ANOVA table.

Note that $R^2 = .25 = SSR/(SSR + SSE) = SSR/(SSR + 144)$ which implies that SSR = 48 and SST = 192. The rest follows easily.

Source	SS	d.f.	MS	F
Regression (or explained)	SSR = 48	K = 12	SSR / K = 4	MSR/MSE = 3.806
Error (or residual)	SSE = 144	N - K - 1 = 137	SSE / (N-K-1) = 1.05	
Total	SST = 192	N - 1 = 149	SST / (N - 1) = 1.29	

b. Y = income (in thousands of dollars). X1 = 1 if black, 0 otherwise. X2 = 1 if female, 0 otherwise. X3 = X1 * X2. (NOTE: X3 is referred to as an interaction term.) Suppose a = 18, b1 = -4, b2 = -3, b3 = -2. What are the average incomes for black males, black females, nonblack males, and nonblack females?

The model is E(Income) = 18 -4*X1 – 3*X2 – 2*X3. Note that Black Females are coded 1 on X3, all others are coded 0. Plugging in the appropriate values of X1, X2, and X3 for each of of the groups, we get:

Black Males	18 – 4 – 0 – 0 = 14
Black Females	18 – 4 – 3 – 2 = 9
Nonblack Males	18 – 0 – 0 – 0 = 18
NonBlack Females	18 – 3 – 0 – 0 = 15

c. When Y is regressed on X1 and X2, $r_{y1} = .5$, $r_{y2} = .5$, $TOL_{X1} = .75$. X1 and X2 are positively correlated. Compute the semipartial correlations.

Recall that $Tol_{X1} = 1 - R_{12}^2 = .75$. Since X1 and X2 are positively correlated, $r_{12} = .5$. We therefore compute

$$sr_1 = \frac{r_{Y1} - r_{Y2} r_{12}}{\sqrt{1 - r_{12}^2}} = \frac{r_{Y1} - r_{Y2} r_{12}}{\sqrt{Tol_1}} = \frac{.5 - .5 * .5}{\sqrt{.75}} = \frac{.25}{\sqrt{.75}} = .289$$

$$sr_2 = \frac{r_{Y2} - r_{Y1} r_{12}}{\sqrt{1 - r_{12}^2}} = \frac{r_{Y2} - r_{Y1} r_{12}}{\sqrt{Tol_2}} = \frac{.5 - .5 * .5}{\sqrt{.75}} = \frac{.25}{\sqrt{.75}} = .289$$

d. A researcher obtains the following results. Are there any logical inconsistencies in the results of the significance tests? If so, discuss what might be responsible. Cite evidence from the printouts to support your points.

Regression

Correlations

		Y	X1	X2
Pearson Correlation	Y	1.000	.240	.250
	X1	.240	1.000	.950
	X2	.250	.950	1.000

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.194	2	3.097	3.237	.044[a]
	Residual	92.806	97	.957		
	Total	99.000	99			

a. Predictors: (Constant), X2, X1

b. Dependent Variable: Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	.0000	.0978		.0000	1.0000		
	X1	.0256	.3148	.0256	.0814	.9353	.0975	10.2564
	X2	.2256	.3148	.2256	.7167	.4753	.0975	10.2564

a. Dependent Variable: Y

The global F is significant, implying that at least one β does not equal zero. However, all of the individual T tests are insignificant, implying that all βs do equal zero. This seeming paradox is likely produced by multicollinearity. X1 and X2 have a correlation of .95 and hence have very low tolerances. Because they are so highly correlated, their standard errors are very large, and hence their confidence intervals include zero.

Hence, while it appears that at least one of the βs does not equal zero, we can't tell which one it is because of the multicollinearity.

3. (50 points total; up to 5 points extra credit.) Credit Scores are increasingly being used in the lending industry. These scores supposedly measure how credit-worthy a person is, by taking into account such things as their income, net worth, how well they have handled credit in the past, the stability of their employment record, etc. Proponents of these scores argue that they eliminate racial discrimination in lending: loan applicants are judged only on their credit-worthiness and not on their race. Others, however, express concern because the formulas used to compute credit scores are not public knowledge; therefore it could be that the scores themselves contain a hidden racial bias.

A researcher has therefore collected data from a random sample of 1,605 recent loan applicants at a major bank. Her variables are as follows:

Variable	**Description**
Black	Coded 1 if the applicant is black, 0 otherwise
Applinc	Applicant Income, in thousands of dollars
Networth	Applicants net worth (assets – debts), measured in thousands of dollars. When debts exceed assets, the value is negative.
CrScore	The applicants credit score, which can range from a low of 0 to a high of 1000. Higher scores indicate the applicant is more credit-worthy.

She obtains the following results:

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
CRSCORE Credit Score	455.18	104.965	1605
BLACK Is Applicant black?	.0829	.27577	1605
APPLINC Applicant income	$71.01	$46.281	1605
NETWORTH Net worth, in thousands of dollars	$33.73	$22.785	1605

Correlations

		CRSCORE Credit Score	BLACK Is Applicant black?	APPLINC Applicant income	NETWORTH Net worth, in thousands of dollars
Pearson Correlation	CRSCORE Credit Score	1.000	-.198	.620	.507
	BLACK Is Applicant black?	-.198	1.000	-.123	-.447
	APPLINC Applicant income	.620	-.123	1.000	.653
	NETWORTH Net worth, in thousands of dollars	.507	-.447	.653	1.000

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.198[a]	.039	.039	102.919
2	[b]	**[1]**		80.893

a. Predictors: (Constant), BLACK Is Applicant black?

b. Predictors: (Constant), BLACK Is Applicant black?, APPLINC Applicant income, NETWORTH Net worth, in thousands of dollars

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	692774.580	1	692774.580	65.404	.000[a]
	Residual	16979434.842	1603	10592.286		
	Total	17672209.422	1604			
2	Regression	7195855.943	3	2398618.648	366.558	.000[b]
	Residual	10476353.479	1601	6543.631		
	Total	17672209.422	1604			

a. Predictors: (Constant), BLACK Is Applicant black?

b. Predictors: (Constant), BLACK Is Applicant black?, APPLINC Applicant income, NETWORTH Net worth, in thousands of dollars

c. Dependent Variable: CRSCORE Credit Score

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	461.421	2.683		172.011	.000	456.160	466.683					
	BLACK Is Applicant black?	-75.362	9.319	**[2]**	-8.087	.000	-93.640	-57.084	-.198	-.198	-.198	1.000	1.000
2	(Constant)	352.655	4.331		81.432	.000	344.160	361.149					
	BLACK Is Applicant black?	-28.715	**[3]**	-.075	-3.395	.001	-45.303	**[4]**	-.198	-.085	-.065	.750	1.333
	APPLINC Applicant income	**[5]**	.059	.527	20.089	.000	1.078	1.312	.620	.449	.387	.538	1.858
	NETWORTH Net worth, in thousands of dollars	.594	.134	.129	4.433	.000	.331	.857	**[6]**	.110	.085	.437	**[7]**

a. Dependent Variable: CRSCORE Credit Score

a. (21 points) Fill in the missing items [1] – [7].

Here are non-censored parts of the printout, with information on the change statistics also added.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.198[a]	.039	.039	102.919	.039	65.404	1	1603	.000
2	.638[b]	.407	.406	80.893	.368	496.902	2	1601	.000

a. Predictors: (Constant), BLACK Is Applicant black?

b. Predictors: (Constant), BLACK Is Applicant black?, APPLINC Applicant income, NETWORTH Net worth, in thousands of dollars

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	461.421	2.683		172.011	.000	456.160	466.683					
	BLACK Is Applicant black?	-75.362	9.319	-.198	-8.087	.000	-93.640	-57.084	-.198	-.198	-.198	1.000	1.000
2	(Constant)	352.655	4.331		81.432	.000	344.160	361.149					
	BLACK Is Applicant black?	-28.715	8.457	-.075	-3.395	.001	-45.303	-12.127	-.198	-.085	-.065	.750	1.333
	APPLINC Applicant income	1.195	.059	.527	20.089	.000	1.078	1.312	.620	.449	.387	.538	1.858
	NETWORTH Net worth, in thousands of dollars	.594	.134	.129	4.433	.000	.331	.857	.507	.110	.085	.437	2.288

a. Dependent Variable: CRSCORE Credit Score

To confirm that SPSS got it right:

[1] = R^2_{Y123} = SSR/SST = 7195855.943/17672209.422 = .407
[2] = $b'_{black} = r_{CrScore,Black}$ = -.198 (because it is a bivariate regression)
[3] = $s_{b\text{-}Black} = b_{black}/t_{black}$ = -28.715/-3.395 = 8.46
[4] = upper bound of black ci = -28.715 + 1.96*8.46 = -12.13
[5] = $b_{Applinc} = s_{b\text{-}Applinc} * t_{Applinc}$ = .059 * 20.089 = 1.185. Or, alternatively (and more accurately), $b_{Applinc} = b'_{Applinc} * s_{CrScore}/s_{Applinc}$ = .527 * 104.965/46.281 = 1.195
[6] = $r_{CrScore,Networth}$ = .507 (from the correlation matrix)
[7] = $VIF_{Networth} = 1/Tol_{Networth}$ = 1/.437 = 2.288

b. (4 points) Two models are estimated. Do you think the researcher used hierarchical model selection or stepwise selection? Point to evidence from the printout to support your argument.

Hierarchical selection must have been used, i.e. the researcher chose the models rather than let the computer do it for her via stepwise regression. If stepwise had been used, then the variable that had the largest correlation with CrScore, Applinc, would have been entered first, rather than Black. Also, with stepwise, only one variable would be added at the second step, rather than two. The researcher probably entered Black first because it is temporally prior to the other variables and because she wanted to see how the effects of Black changed once other variables were controlled for.

c. (5 points) Do an incremental F test of the hypothesis

H_0: $\beta_{Applinc} = \beta_{Networth} = 0$

As the F change statistic in the uncensored printout shows, F = 496.902 with df = 2, 1601, which is highly significant. We reject the null. To confirm that SPSS got it right:

$$F_{J,N-K-1} = \frac{(SSE_c - SSE_u)/J}{SSE_u/(N-K-1)} = \frac{MSE_{c-u}}{MSE_u} = \frac{(SSE_c - SSE_u)*(N-K-1)}{SSE_u * J}$$

$$= \frac{(16979434.842 - 10476353.479)*(1605-3-1)}{10476353.479*2} = \frac{10411433262.163}{20952706.958} = 496.902$$

Or, slightly less accurately because of greater rounding,

$$F_{J,N-K-1} = \frac{(R_u^2 - R_c^2)*(N-K-1)}{(1-R_u^2)*J} = \frac{(.407-.039)*1601}{(1-.407)*2} = 496.769$$

d. (5 points) When Applinc and Networth are added to the model, the effect of Black becomes much smaller. Offer a substantive explanation as to why this might occur. What evidence from the printout might support your argument?

Note that Black is negatively correlated with Applinc and Networth, which in turn have positive effects on credit scores. This implies that part of the reason blacks have lower credit scores is because they tend to have lower incomes and less net worth. So, while blacks score 75.362 points lower on average, the expected difference between a black and a non-black with comparable incomes and net worth would only be 28.715 points.

To put it another way, Model 1 likely suffers from omitted variable bias. In reality, race may have both direct and indirect effects on credit scores. Race affects both income and net worth, which in turn affect credit scores. When income and net worth are excluded from the model, the direct effect of race gets over-estimated. We will talk more about direct and indirect effects, and omitted variable bias, next semester.

e. (15 points) Interpret the results. Be sure to answer the following questions. Explain what information from the printout supports your conclusions.

1. What proportion of the sample is black?

The mean of Black is .0829, implying that 8.29% of the sample is black.

2. What is the mean credit score for whites? What is the mean credit score for blacks?

The constant in Model 1 gives the mean score for whites, 461.421. Model 1 also shows that blacks score 75.362 points lower, for an average score of 386.059.

3. What would you say is the most important determinant of credit scores? Cite at least two pieces of evidence from the printout to support your argument.

Applicant income would seem to be the most important determinant of credit scores. It has the largest bivariate, partial and semipartial correlations with credit score. It has the largest standardized effect and the largest T value. Also, both applicant income and net

worth are measured in thousands of dollars, and the effect of applicant income is more than twice as large.

4. Is there evidence to suggest that there may be racial bias in the credit scores? If so, what is that evidence?

While the effects of race decline once income and net worth are controlled for, the effect of Black continues to be negative and statistically significant. One possible explanation for this is that the credit scores have a racial bias in the way they are computed.

5. If you were an industry representative who had to defend credit scores against charges of racial bias, how might you respond to whatever arguments you raised in pt. 4? That is, how would you argue that the analysis has not been done correctly or else that alternative explanations of the results are possible?

As was pointed out in the original question, "These scores supposedly measure how credit-worthy a person is, by taking into account such things as their income, net worth, how well they have handled credit in the past, the stability of their employment record, etc." Only two of the variables mentioned, income and net worth, are included in the model. It could be that remaining racial differences reflect the fact that blacks may have had more credit problems in the past or might tend to have less stable employment records. Note how much the effect of race went down once income and net worth were controlled for; it might go down and disappear completely once these other variables were considered. In effect, the industry representative would argue that the models suffer from omitted variable bias. The remaining racial differences are due to important variables that have been excluded from the analysis.

f. (5 points extra credit) Suppose you wanted to test

H_0: $\beta_{Applinc} = \beta_{Networth}$

i.e. income and wealth have equal effects on credit scores. How would you go about doing it? [Hint: You could do it with an incremental F test. The unconstrained model is Model 2 above. What would the constrained model be? In practice, how would you estimate the constrained model, using SPSS?] You don't have to give me a value for the test statistic, just describe what additional analyses you would run and how you would then compute the value of the incremental F statistic.

The unconstrained model (Model 2) can be written as

$$E(Creditscore) = \alpha + \beta_{Black} Black + \beta_{Applinc} Applinc + \beta_{Networth} Networth$$

If the null hypothesis is true, then the effects of Applinc and Networth are equal, and the model becomes

$$E(Creditscore) = \alpha + \beta_{Black} Black + \beta_{Applinc} (Applinc + Networth)$$

Now, the question is, how would you estimate this constrained model? Some programs, such as Stata, make it easy to test equality constraints on parameters. SPSS, alas, does not. HOWEVER, as the above model implies, if you add Applinc and Networth together and then use that variable in the model, their effects will be constrained to be equal. So, in SPSS, do something like

Compute IncWorth = Applinc + Networth

Then, run a regression where CrScore is dependent and Black and IncWorth are independent. This gives you the results for the constrained model, and you can then use our useful incremental F procedures to determine whether the constrained model is adequate (i.e. effects are equal) or whether the unconstrained model is significantly better. Note that K = 3 and J = 1 (there is one constraint, in that instead of estimating 2 parameters for Applinc and Networth, we only estimate 1). We discuss how to test such hypotheses in more detail 2nd semester, but here are ways you could do it in SPSS and Stata.

SPSS Solution.

```
Compute IncWorth = Applinc + Networth.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT crscore
  /METHOD=ENTER black incworth  .
```

Regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.635[a]	.403	.402	81.149

a. Predictors: (Constant), INCWORTH, BLACK Is Applicant black?

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7122666	2	3561333.235	540.806	.000[a]
	Residual	10549543	1602	6585.233		
	Total	17672209	1604			

a. Predictors: (Constant), INCWORTH, BLACK Is Applicant black?

b. Dependent Variable: CRSCORE Credit Score

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	348.721	4.181		83.404	.000
	BLACK Is Applicant black?	-16.070	7.589	-.042	-2.118	.034
	INCWORTH	1.029	.033	.623	31.248	.000

a. Dependent Variable: CRSCORE Credit Score

To compute the incremental F statistic,

$$F_{J,N-K-1} = \frac{(SSE_c - SSE_u)/J}{SSE_u/(N-K-1)} = \frac{MSE_{c-u}}{MSE_u} = \frac{(SSE_c - SSE_u)*(N-K-1)}{SSE_u * J}$$

$$= \frac{(10549542.95266 - 10476353.479)*(1605-3-1)}{10476353.479*1} = \frac{117176347.2336}{10476353.479} = 11.18$$

Or, alternatively,

$$F_{J,N-K-1} = \frac{(R_u^2 - R_c^2)*(N-K-1)}{(1-R_u^2)*J} = \frac{(.407-.403)*1601}{(1-.407)*1} = 10.799$$

Differences are due to the greater rounding with R^2. The first F is more accurate.

Stata Solution. Stata makes it easy to test hypotheses such as the above. We would use the Regress and Test commands. The first test command tests

H_0: $\beta_{Applinc} = \beta_{Networth} = 0$

The second test command tests

H_0: $\beta_{Applinc} = \beta_{Networth}$

```
. regress  crscore black applinc networth

      Source |       SS       df       MS              Number of obs =    1605
-------------+------------------------------           F(  3,  1601) =  366.56
       Model |  7195856.01     3  2398618.67           Prob > F      =  0.0000
    Residual |  10476353.6  1601  6543.63123           R-squared     =  0.4072
-------------+------------------------------           Adj R-squared =  0.4061
       Total |  17672209.6  1604   11017.587           Root MSE      =  80.893

------------------------------------------------------------------------------
     crscore |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |  -28.71507   8.456968    -3.40   0.001    -45.30296   -12.12718
     applinc |   1.194999   .0594866    20.09   0.000     1.078319    1.311679
    networth |   .5943289   .1340799     4.43   0.000     .3313382    .8573196
       _cons |   352.6545    4.33065    81.43   0.000     344.1602    361.1489
------------------------------------------------------------------------------

. test applinc networth

 ( 1)  applinc = 0
 ( 2)  networth = 0

       F(  2,  1601) =  496.90
            Prob > F =    0.0000

. test applinc=networth

 ( 1)  applinc - networth = 0

       F(  1,  1601) =   11.18
            Prob > F =    0.0008
```