

# Sociology 592 - Research Statistics I

## Final Exam Answer Key

### December 15, 2004

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.** (4 points each, 20 points total). For each 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.** If Y is regressed on X1 and X2, then it must be the case that  $R^2_{Y|X_1, X_2} \geq sr^2_1 + sr^2_2$

False. This need not be true when suppressor effects are present.

**b.** Because groups have different means and standard deviations, it is important to use standardized coefficients when comparing them.

False. It is because groups have different means and standard deviations that standardized coefficients are problematic; each group gets standardized in a different way.

**c.** In a regression analysis, the null and alternative hypotheses are

$$H_0: \beta_{educ} = 2$$

$$H_A: \beta_{educ} \neq 2$$

The following output is obtained:

```
. use "D:\SOC593\Statafiles\blwh.dta"
```

```
. reg income educ jobexp
```

Source	SS	df	MS			
Model	32798.4018	2	16399.2009	Number of obs =	500	
Residual	7382.84742	497	14.8548238	F( 2, 497) =	1103.96	
Total	40181.2493	499	80.5235456	Prob > F =	0.0000	
				R-squared =	0.8163	
				Adj R-squared =	0.8155	
				Root MSE =	3.8542	

  

income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	1.94512	.0436998	44.51	0.000	1.859261	2.03098
jobexp	.7082212	.0343672	20.61	0.000	.6406983	.775744
_cons	-7.382935	.8027781	-9.20	0.000	-8.960192	-5.805678

Using the .05 level of significance, the null hypothesis should be rejected.

False. The t-test given tests the hypothesis that  $\beta_{educ} = 2$ , not 0. The test we want is

$$T_{N-K-1} = \frac{b_k - \beta_{k0}}{s_{b_k}} = \frac{b_k - 2}{s_{b_k}} = \frac{1.94512 - 2}{.0436998} = \frac{-.05488}{.0436998} = -1.2558$$

which is not significant. Confirming with Stata,

```
. test educ = 2
( 1) educ = 2
      F( 1, 497) = 1.58
      Prob > F = 0.2098
```

Alternatively and more simply, you can just note that 2 falls within the 95% confidence interval so do not reject the null.

- d. The larger the sample size, the smaller the difference will be between  $R^2$  and Adjusted  $R^2$ .

True.

- e. A researcher obtains the following:

```
. use "D:\SOC593\Statafiles\md.dta", clear
. hireg income (jobexp) (educ black)
```

Model 1:

```
Variables in Model:
Adding          : jobexp
```

Source	SS	df	MS	Number of obs =	500
Model	3367.63913	1	3367.63913	F( 1, 498) =	45.56
Residual	36813.6101	498	73.9229119	Prob > F =	0.0000
Total	40181.2493	499	80.5235456	R-squared =	0.0838
				Adj R-squared =	0.0820
				Root MSE =	8.5978

  

income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
jobexp	.5132345	.0760401	6.75	0.000	.3638355 .6626334
_cons	20.85107	1.097614	19.00	0.000	18.69454 23.0076

Model 2:  
 Variables in Model: jobexp  
 Adding : educ black

Source	SS	df	MS	Number of obs =	405
Model	26848.0425	3	8949.34751	F( 3, 401) =	650.82
Residual	5514.07625	401	13.7508136	Prob > F =	0.0000
				R-squared =	0.8296
				Adj R-squared =	0.8283
Total	32362.1188	404	80.1042544	Root MSE =	3.7082

income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
jobexp	.637444	.0394806	16.15	0.000	.5598292 .7150589
educ	1.829589	.0523698	34.94	0.000	1.726635 1.932542
black	-2.706644	.5186577	-5.22	0.000	-3.726272 -1.687016
_cons	-4.328552	1.03952	-4.16	0.000	-6.372142 -2.284963

R-Square Diff. Model 2 - Model 1 = 0.746 F(2,401) = 853.782 p = 0.000

Model	R2	F(df)	p	R2 change	F(df) change	p
1:	0.084	45.556(1,498)	0.000			
2:	0.830	650.823(3,401)	0.000	0.746	853.782(2,401)	0.000

Based on the incremental F test that is reported, she should conclude that the effects of educ and/or black significantly differ from zero.

**False.** The incremental F test that is reported is computed incorrectly. Missing data causes the number of cases to differ in the 2 models. The same cases should be analyzed in both models if you want to use an incremental F test.

**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,  $n = 200$ ,  $k = 9$ ,  $F = 4$ ,  $SSR = 72$ . Construct the ANOVA table.

From the information given, the rest of the numbers can be easily determined.

Source	SS	d.f.	MS	F
Regression (or explained)	SSR = 72	K = 9	SSR / K = 8	MSR/MSE = 4
Error (or residual)	SSE = 380	N - K - 1 = 190	SSE / (N-K-1) = 2	
Total	SST = 452	N - 1 = 199	SST / (N - 1) = 2.27	

**b.**  $Y = \text{Religiosity}$ , measured on a continuous scale.  $X_1 = 1$  if Catholic, 0 otherwise.  $X_2 = 1$  if male, 0 otherwise.  $X_3 = X_1 * X_2$ . (NOTE:  $X_3$  is referred to as an interaction term.) Suppose  $a = 30$ ,  $b_1 = 5$ ,  $b_2 = 4$ ,  $b_3 = -5$ . What are the average levels of religiosity for Catholic males, Catholic females, NonCatholic males, and NonCatholic females?

Note that Catholic Males are coded 1 on X3, everyone else is coded 0.

Entire Population:  $E(Y) = 30 + 5*X1 + 4*X2 - 5*X3$

Catholic Males:  $E(Y) = 30 + 5*1 + 4*1 - 5*1 = 34$

Catholic Females:  $E(Y) = 30 + 5*1 + 4*0 - 5*0 = 35$

NonCatholic Males:  $E(Y) = 30 + 5*0 + 4*1 - 5*0 = 34$

NonCatholic Females:  $E(Y) = 30 + 5*0 + 4*0 - 5*0 = 30$

c. Using the following information, compute the semipartial correlations.

```
. use "D:\SOC593\Statafiles\drinking.dta"
```

```
. reg drink gpa male
```

Source	SS	df	MS	Number of obs = 218		
Model	1437.71088	2	718.855442	F( 2, 215)	=	18.36
Residual	8416.31205	215	39.1456374	Prob > F	=	0.0000
Total	9854.02294	217	45.4102439	R-squared	=	0.1459
				Adj R-squared	=	0.1380
				Root MSE	=	6.2566

  

drink	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
gpa	-3.4529	.9400734	-3.67	0.000	-5.30584	-1.59996
male	3.535818	.8649733	4.09	0.000	1.830904	5.240732
_cons	26.91249	2.7702	9.71	0.000	21.45226	32.37272

You should use the same formula `pcorr2` uses. Don't forget the negative sign.

$$sr_1 = \frac{T_1 * \sqrt{1 - R_{YH}^2}}{\sqrt{N - K - 1}} = \frac{-3.67 * \sqrt{1 - .1459}}{\sqrt{215}} = \frac{-3.3917}{14.663} = -.231$$

$$sr_2 = \frac{T_2 * \sqrt{1 - R_{YH}^2}}{\sqrt{N - K - 1}} = \frac{4.09 * \sqrt{1 - .1459}}{\sqrt{215}} = \frac{3.7799}{14.663} = .258$$

Confirming with Stata,

```
. pcorr2 drink gpa male
```

```
(obs=218)
```

Partial and Semipartial correlations of drink with

Variable	Partial	SemiP	Partial^2	SemiP^2	Sig.
gpa	-0.2430	-0.2315	0.0590	0.0536	0.000
male	0.2685	0.2576	0.0721	0.0664	0.000

d. Using the following information, computed the standardized coefficients for when Y is regressed on X1 and X2.

```
. corr
(obs=500)
```

	y	x1	x2
y	1.0000		
x1	0.4000	1.0000	
x2	0.4000	-0.4000	1.0000

$$b_1' = (r_{y1} - r_{12} * r_{y2}) / (1 - r_{12}^2) = (.4 + .4 * .4) / (1 - .4^2) = .56 / .84 = .667$$

$$b_2' = (r_{y2} - r_{12} * r_{y1}) / (1 - r_{12}^2) = (.4 + .4 * .4) / (1 - .4^2) = .56 / .84 = .667$$

Confirming with Stata,

```
. reg y x1 x2, beta
```

Source	SS	df	MS	
Model	266.133336	2	133.066668	Number of obs = 500
Residual	232.866665	497	.468544597	F( 2, 497) = 284.00
Total	499.000001	499	1	Prob > F = 0.0000
				R-squared = 0.5333
				Adj R-squared = 0.5315
				Root MSE = .6845

  

	Coef.	Std. Err.	t	P> t	Beta
x1	.6666667	.0334338	19.94	0.000	.6666667
x2	.6666667	.0334338	19.94	0.000	.6666667
_cons	-8.11e-09	.0306119	-0.00	1.000	.

3. (50 points; up to 5 points extra credit) Intravenous drug use has been cited as a major factor in the spread of AIDS. When drug addicts share dirty needles and/or engage in unsafe sex, the AIDS virus can easily be spread from one user to another. As a result, a number of programs have recently been launched which aim to get drug users to follow safer practices. Most programs rely on outreach workers who contact drug users, try to educate them about safe practices, give them bleach for cleaning needles, etc. Such programs are very expensive and have had only limited effectiveness. A new proposal calls for a user-driven approach. Under this system, addicts will be paid small stipends for recruiting other users into the program, for distributing bleach and condoms, and for assisting in educational efforts. In addition, attempts will be made to develop group norms among drug users which encourage safe practices. If successful, the new program may be much less expensive than current approaches and also more effective, because users will internalize the attitudes needed to sustain the safe practices.

To test this idea, two communities have been selected for study. In one community, a conventional outreach program using social workers will be set up. In the other community, the user-driven approach will be tried. The two communities are similar to each other in many ways but, as is so often the case in real-world experiments, there is no guarantee that there are not some important differences between them.

Some of the variables that might be examined in this analysis are:

AidsIQ	Participants will be asked a number of questions about "safe" practices (safe insofar as they reduce the chance of getting or transmitting Aids). The more questions right, the higher the score. Both programs hope that their educational efforts will raise the AidsIQ of the drug user population; hence, this will be the dependent variable in the current analysis
UserDriv	This variable is coded 1 if the subject is participating in the user-driven program, 0 if participating in the conventional program. Obviously, the researchers are hypothesizing and hoping that participants in the user-driven program get higher AidsIQ scores than those in the conventional program.
Female	This variable is coded 1 if the subject is female, 0 if male
Educ	Years of education.

The latter two variables (Female and Educ) are included because (1) they may be related to AidsIQ, e.g. women and/or better educated subjects may know more about safe practices, and (2) the two communities chosen for the study may not be completely comparable on these variables - e.g. one community might have more women or better-educated drug users than the other - hence the researchers want to make sure that apparent differences between the two programs are not actually due to community differences in education and gender.

Following are hypothetical results from this proposed study. Stepwise regression was used to estimate three models, the first and last of which are presented here:

## Regression

### Descriptive Statistics

	Mean	Std. Deviation	N
AidsIQ	60.000	15.000	400
UserDriv	.500	.501	400
Female	.200	.402	400
Educ	10.000	2.000	400

### Correlations

		AidsIQ	UserDriv	Female	Educ
Pearson Correlation	AidsIQ	1.000	.500	.400	.400
	UserDriv	.500	1.000	.500	.800
	Female	.400	.500	1.000	.100
	Educ	.400	.800	.100	1.000

### 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	.500 <sup>a</sup>	.250	.248	13.0066904	.250	132.667	1	398	.000
3	.543 <sup>c</sup>	<b>[1]</b>	.290	12.6422587	.015	8.426	1	396	.004

a. Predictors: (Constant), UserDriv

c. Predictors: (Constant), UserDriv, Female, Educ

**ANOVA<sup>d</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	22443.750	1	22443.750	132.667	.000 <sup>a</sup>
	Residual	67331.250	398	169.174		
	Total	89775.000	399			
3	Regression	26483.625	3	8827.875	55.234	.000 <sup>c</sup>
	Residual	63291.375	396	159.827		
	Total	89775.000	399			

a. Predictors: (Constant), UserDriv

c. Predictors: (Constant), UserDriv, Female, Educ

d. Dependent Variable: AidsIQ

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	52.515	.919		57.121	.000					
	UserDriv	14.970	1.300	.500	11.518	.000	.500	.500	.500	1.000	1.000
3	(Constant)	36.766	5.500		6.684	.000					
	UserDriv	4.491	2.963	.150	<b>[2]</b>	.130	.500	.076	.064	.182	5.500
	Female	<b>[3]</b>	2.227	.300	5.028	.000	.400	.245	.212	.500	2.000
	Educ	1.875	.646	<b>[4]</b>	2.903	.004	.400	.144	.122	<b>[5]</b>	4.167

a. Dependent Variable: AidsIQ

**Excluded Variables<sup>c</sup>**

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	Female	.200 <sup>a</sup>	4.067	.000	.200	.750	1.333	.750
	Educ	.000 <sup>a</sup>	.000	1.000	.000	.360	2.778	.360

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

c. Dependent Variable: AidsIQ

a. (10 points) Since forward stepwise selection is being used, why was USERDRIV the first variable selected? What variable was added in the second model (not shown)?

UserDriv has the largest bivariate correlation with AidsIQ and its effect is statistically significant. Excluded Variables shows you Female entered on Step 2. [NOTE: Just because Female has the biggest T value in the coefficients table does not mean it had to enter 2<sup>nd</sup>. You should use the Excluded Variables table to determine what entered 2<sup>nd</sup>.]

b. (10 points) Interpret the results from the descriptive statistics (means, correlations and standard deviations) and from Model I (the model with USERDRIV only). What proportion of the sample is female? How many years of education does the average drug user have? Do you think the researchers would be happy with the results from Model I? Why or why not?

From the means you can see that 20% are female and the average drug user has 10 years of education. They should be happy so far because Model 1 shows that those in the user-driven program have significantly higher AIDS IQs.

c. (10 points) Fill in the missing items [1] - [5].

Here are the uncensored parts of the printout:

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	.500 <sup>a</sup>	.250	.248	13.0066904	.250	132.667	1	398	.000
2	.529 <sup>b</sup>	.280	.276	12.7599420	.030	16.542	1	397	.000
3	.543 <sup>c</sup>	.295	.290	12.6422587	.015	8.426	1	396	.004

a. Predictors: (Constant), UserDriv

b. Predictors: (Constant), UserDriv, Female

c. Predictors: (Constant), UserDriv, Female, Educ

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	52.515	.919		57.121	.000					
	UserDriv	14.970	1.300	.500	11.518	.000	.500	.500	.500	1.000	1.000
2	(Constant)	52.519	.902		58.230	.000					
	UserDriv	11.976	1.472	.400	8.134	.000	.500	.378	.346	.750	1.333
	Female	7.463	1.835	.200	4.067	.000	.400	.200	.173	.750	1.333
3	(Constant)	36.766	5.500		6.684	.000					
	UserDriv	4.491	2.963	.150	1.516	.130	.500	.076	.064	.182	5.500
	Female	11.194	2.227	.300	5.028	.000	.400	.245	.212	.500	2.000
	Educ	1.875	.646	.250	2.903	.004	.400	.144	.122	.240	4.167

a. Dependent Variable: AidsIQ

To confirm that SPSS got it right:

[1] =  $R^2 = SSR/SST = 26483/89775 = .295$ . Or, easier still,  $R^2 = .543^2 = .295$  (I meant to blank out the value of R and I forgot – but most people ignored that and found a harder way to do it!)

[2] =  $t_{UserDriv} = b_{Userdriv}/s_{b-Userdriv} = 4.491/ 2.963 = 1.516$

[3] =  $b_{Female} = t_{Female} * s_{b-Female} = 5.028 * 2.227 = 11.197$ ; or,  
 $b'_{Female} * s_{AidsIQ} / s_{Female} = .3 * 15/.402 = 11.194$

[4] =  $b'_{Educ} = b_{Educ} * s_{Educ} / s_{AidsIQ} = 1.875 * 2/ 15 = .25$

[5] =  $Tol_{Educ} = 1/VIF_{Educ} = 1/4.167 = .240$

**d.** (10 points) Do an incremental F test of the hypothesis  $H_0: \beta_{\text{female}} = \beta_{\text{educ}} = 0$ . (Remember that the results for Model II have been deleted; the calculation I am asking you to do is NOT already included in the printout.)

Model III is the unconstrained model, Model I is the constrained model.

$$F_{J, N-K-1} = \frac{(SSE_c - SSE_u) * (N - K - 1)}{SSE_u * J} = \frac{(67331.250 - 63291.375) * (400 - 3 - 1)}{63291.375 * 2} = \frac{1,599,790.5}{126,582.75} = 12.64$$

$$= \frac{(R_u^2 - R_c^2) * (N - K - 1)}{(1 - R_u^2) * J} = \frac{(.295 - .250) * (400 - 3 - 1)}{(1 - .295) * 2} = \frac{17.82}{1.41} = 12.64$$

**e.** (10 points) Based on Model III, what would you say is the most important determinant of AidsIQ? Is this consistent with Model I and your answer in part b? Do you think the researchers will be happy with this final model? Cite evidence from the printout to support your arguments.

Female has the largest standardized coefficient, the largest T value and the largest semipartial correlation, so it is arguably the most important. The effects of UserDriv are not statistically significant, so the researchers are probably not too happy.

**f.** (5 points extra credit) Explain how pre-existing differences between the two communities may account for apparent discrepancies between Model I and Model III and caused the user-driven approach to appear more effective than it actually was. Cite evidence from the printout to support your arguments.

As the correlation matrix shows, those in the User-Driven program are more likely to be female and also tend to be better-educated. Further, both the correlations and the regression analyses show that females and better-educated people tend to have higher Aids IQs. Thus, it looks like those in the User-driven program did better not so much because of the program but because of their pre-existing gender and educational differences from the traditional outreach group. The researchers might want to try to find communities that are more similar when conducting future analyses. [NOTE: Several people said there may have been gender and educational differences between the two communities. There is no need to be so tentative – you can tell from the correlation matrix that this is the case.]