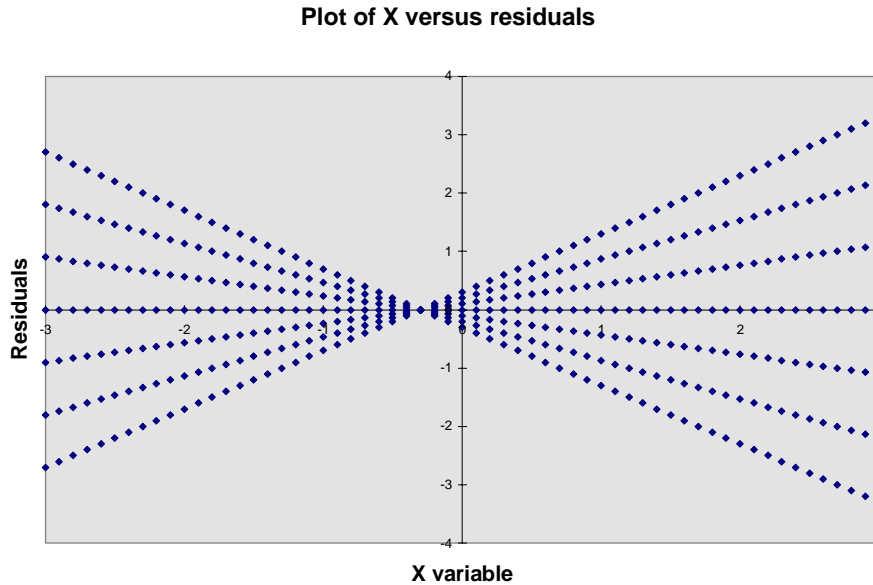


**Sociology 593**  
**Exam 1 Answer Key**  
**February 17, 1995**

I. *True-False.* (25 points) Indicate whether the following statements are true or false. If false, briefly explain why.

1. A researcher regressed Y on X. When he plotted the residuals against X, he got the following. He should now use the GQ test to determine whether heteroscedasticity is present.



**False.** GQ should only be used when error variances increase monotonically with X, i.e. form a funnel rather than hourglass shape.

2. In a regression, if the alternative hypothesis is two-tailed *and* there is only one IV, either an F test or a T test is appropriate.

**True.** Under such conditions, F and T test the same hypothesis, and  $F = T^2$ .

3. Random measurement error results in biased estimates of means, variances and covariances.

**False.** Estimates of means and variances are not biased, although variances are inflated.

4. If more variables are added to an equation, the F value will always either stay the same or increase.

**False.** Adding extraneous variables that do not increase  $R^2$  can actually cause F to decline, as the following formula indicates:

$$F = \frac{R^2 * (N - K - 1)}{(1 - R^2) * K}$$

As the formula shows, increases in K which are not accompanied by increases in  $R^2$  cause the numerator to decline and the denominator to increase, leading to smaller F values.

5. A researcher found that the reliability of a measuring instrument was higher for women than it was for men. This must mean that there is less error variance in women's responses, i.e. women provide more accurate answers than do men.

**False.** Remember that reliability is a function of both True variance and Error variance. Hence, differing reliabilities could mean that men are more "error-prone," but it could also mean that there is more true variability among women than there is among men.

$$\text{Rel}(X) = \frac{\text{True Variance}}{\text{True Variance} + \text{Error Variance}}$$

II. *Short answer.* Answer *three* of the following four questions. (25 points each; up to 10 points extra credit if you do all 4).

1. A researcher collected the following data:

Case #	Y	X1	X2	X3
1	30	2	Missing	12
2	37	2	1	Missing
3	41	3	1	20
4	42	1	Missing	16
5	45	3	2	Missing
6	49	1	2	27
7	51	Missing	1	30
8	55	3	2	33
9	58	Missing	2	19
10	60	2	Missing	24

a. Suppose the researcher believes that data are missing on a *random* basis, i.e. those who did not respond are no different than those who did. What would you recommend for her—pairwise deletion of missing data, or listwise deletion? Why?

Listwise deletion would result in 70% of the cases being deleted. Because data are missing randomly and MD is spread across several variables, pairwise deletion seems a much more reasonable option in this case.

b. Suppose the researcher believes that data may be missing on a *non-random* basis. What would you recommend for her—substitution of the mean for MD cases, or substitution of the mean plus including missing data dichotomies. Why?

Mean substitution alone can be preferable if data are thought to be missing randomly. In such a case, the MD indicator would simply be an extraneous variable. Since data

are thought to be missing non-randomly, MD indicators should be used. The coefficients and T-values for these variables will indicate whether or not the MD cases differ significantly from the non-MD cases.

2. A researcher obtained the following printout:

Listwise Deletion of Missing Data

	Mean	Std Dev	Label
Y	79.000	9.400	
V1	14.000	2.700	
V2	32.000	5.600	
V3	42.000	7.100	

N of Cases = 200

Correlation:

	Y	V1	V2	V3
Y	1.000	.240	.250	.270
V1	.240	1.000	.810	.850
V2	.250	.810	1.000	.900
V3	.270	.850	.900	1.000

\* \* \* \* MULTIPLE REGRESSION \* \* \* \*

Equation Number 1 Dependent Variable.. Y

	Multiple R	R Square	Adjusted R Square	Standard Error	Analysis of Variance	DF	Sum of Squares	Mean Square
	.27102	.07345	.05927	9.11718	Regression	3	1291.53396	430.51132
					Residual	196	16292.10604	83.12299

F = 5.17921 Signif F = .0018

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

Variable	B	SE B	Beta	Tolerance	VIF	T	Sig T
V1	.115363	.463386	.033136	.266842	3.748	.249	.8037
V2	.048669	.270006	.028994	.182703	5.473	.180	.8571
V3	.285627	.237076	.215740	.147427	6.783	1.205	.2297
(Constant)	63.831186	3.920948				16.280	.0000

Much to her dismay, none of the T values for the beta coefficients are statistically significant.

a. What problem might account for this? Point to at least two things in the printout that support your argument.

Multicollinearity appears to be a threat. The correlations among the IVs are fairly large ( $r_{12} = .81$ ,  $r_{13} = .85$ ,  $r_{23} = .90$ ) and the sample size is small ( $N = 200$ ). The overall F is very significant, yet none of the T values for the individual betas are. The three tolerances are also very low.

b. Does this problem cause parameter estimates to be biased? If not, then why should you be concerned about it?

The coefficients are not biased. However, because of the large standard errors, they are imprecise, and subject to great sampling variability. The large standard errors also increase the likelihood that you will accept the null hypothesis when you should reject it, i.e. you will conclude that effects are not significant when they really are.

c. Briefly discuss at least three possible ways for dealing with the problem.

- Make sure you haven't made any computational errors, i.e. you haven't accidentally included variables which were computed from each other
- Increase the sample size. This will generally make estimates more precise and reduce standard errors.
- Drop one or more variables—but this would be bad if the variables actually belong in the model
- Use information from prior research to place justifiable constraints on parameters. For example, we might know from prior research that  $b_1 = 2b_2$ .
- If the three X's are all indicators of the same concept, create some sort of composite scale and use it instead.
- Use joint hypothesis tests—test all the relevant coefficients simultaneously rather than individually.

3. A researcher fears that heteroscedasticity may be a problem in her data (N = 210). She therefore runs two regressions. Following are part of her results:

Regression 1:

Selecting only Cases for which X LE 1.20

Equation Number 1 Dependent Variable.. Y

Block Number 1. Method: Enter X

Variable(s) Entered on Step Number 1.. X

Multiple R	.45594	Analysis of Variance			
R Square	.20788		DF	Sum of Squares	Mean Square
Adjusted R Square	.19822	Regression	1	10.01000	10.01000
Standard Error	.68202	Residual	82	38.14205	.46515
		F =	21.52008	Signif F =	.0000

Regression 2:

Selecting only Cases for which X GE 1.90

Equation Number 1 Dependent Variable.. Y

Block Number 1. Method: Enter X

Variable(s) Entered on Step Number 1.. X

Multiple R	.18365	Analysis of Variance			
R Square	.03373		DF	Sum of Squares	Mean Square
Adjusted R Square	.02194	Regression	1	10.01000	10.01000
Standard Error	1.87012	Residual	82	286.78205	3.49734
		F =	2.86217	Signif F =	.0945

a. Explain, in your own words, the logic behind the researcher's strategy, i.e. why is this a good way for testing for heteroscedasticity?

If error variances increase as X increases (a common form of heteroscedasticity), then larger values of X will produce larger residual sums of squares than do smaller values of X. Hence, by comparing the residual sums of squares, we can see whether this form of heteroscedasticity appears to be present.

b. Compute the appropriate test statistic. Based on the test statistic, what should the researcher conclude about heteroscedasticity? [HINT: In case you don't have your tables handy, the critical value for the test statistic is about 1.5]

$$F_{(N-d-4)/2, (N-d-4)/2} = \frac{SSE_{high}}{SSE_{low}} = \frac{286.78}{38.14} = 7.52$$

This value is highly significant, suggesting heteroscedasticity is present.

c. If a problem appears to exist, suggest two or more ways the researcher might try to deal with it.

- She could try to specify a better model. Heteroscedasticity often results from other errors in model specification, e.g. omitted variables, nonlinear effects.
- She could use GLS or WLS. Instead of weighting all cases equally, like OLS does, WLS gives greater weight to cases with smaller residual variances. This is because the observations with the smallest error variances give the best information about the true position of the regression line.

4. A researcher has collected information on the following variables:

Y	Depression (where high values indicate high levels of depression)
X1	Job dissatisfaction (high values indicate high dissatisfaction)
X2	Physical health (high values indicate good physical health)
X3	Income (measured in thousands of dollars)

She obtains the following results:

	Mean	Std Dev
Y	.000	2.128
X1	.000	10.726
X2	.000	79.477
X3	.000	.939

N of Cases = 427

\*\*\*\*\* MULTIPLE REGRESSION \*\*\*\*\*

Equation Number 1 Dependent Variable.. Y

Block Number 1. Method: Enter X1 X2 X3

Variable(s) Entered on Step Number  
 1.. X3  
 2.. X2  
 3.. X1

R Square	(1)	Analysis of Variance				
Standard Error	1.19664	Regression	Residual	DF	Sum of Squares	Mean Square
				3	1323.37441	441.12480
			(2)		605.71583	1.43195
		F =	308.05831	Signif F =	.0000	

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

Variable	B	SE B	Beta	Correl	Part Cor	Partial	Tolerance	VIF	T
X1	(3)	.046160	.590600	.759340	.069157	.122489	.013712	72.931	2.538
X2	-.006737	.005013	(4)	.680281	-.036614	-.065203	.021178	47.218	-1.344
X3	1.141410	.165034	.503680	.811429	.188433	.318738	(5)	7.145	6.916
(Constant)	-1.18480E-15	.057910							.000

a. Fill in the missing quantities (1)-(5)

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE} = \frac{1323.37441}{1323.37441 + 605.71583} = .686$$

$$DFE = N - K - 1 = 427 - 3 - 1 = 423$$

$$b_1 = T_{b_1} * s_{b_1} = 2.538 * .046160 = .117154$$

$$b'_2 = b_2 * \frac{s_{x_2}}{s_y} = -.006737 * \frac{79.477}{2.128} = -.251615$$

$$Tol_{x_3} = \frac{1}{VIF_{x_3}} = \frac{1}{7.145} = .139958$$

b. The researcher believes that higher incomes lead to lower levels of depression. Do the results support her belief?

If she were correct, then income (X3) would have a negative effect on depression. Since the estimated coefficient for X3 is positive, the results obviously do not support her.

c. If you were doing stepwise regression, what variable would be removed from the equation next?

X2 has the smallest semipartial correlation, and its effect is not statistically significant. Hence, it would be eliminated next.

d. Do an F test of the hypothesis

$$H_0: \beta_2 = \beta_3 = 0$$

$$H_A: \beta_2 \text{ and/or } \beta_3 \neq 0$$

Note that the unconstrained model (the model in which all coefficients are free to differ from 0) is presented. As shown above,  $R^2_u = .686$ . Note that the constrained model only has one IV, X1. Further, in a bivariate regression, recall that  $R^2 = r^2_{yx}$ . As the column labeled "correlation" shows you,  $r_{y1} = .759340$ , hence  $R^2_c = .759340^2 = .577$ . Note further that  $N = 427$ ,  $K = 3$  (the number of IVs in the unconstrained model),  $J = 2$  (the number of parameters constrained to equal 0). Hence, we get

$$F_{J, N-K-1} = \frac{(R_u^2 - R_c^2) * (N - K - 1)}{(1 - R_u^2) * J} = \frac{(.686 - .577) * (427 - 3 - 1)}{(1 - .686) * 2} = 73.42$$

which is highly significant, ergo we reject the null.