Answer Key

Sociology 593
Exam 1
February 13, 1998

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

1. When outliers are eliminated from a data set, correlations between variables will always go up. F
Pairs of outliers (e.g. miscoded MD) can ↑ correlations

2. The null and alternative hypotheses are

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

F. b is the wrong sign, i.e 7 is NOT less than 0

In the sample, N = 500, b = 7, $s_b = 3.00$. If you are using the .05 level of significance, you should reject the null hypothesis.

3. When doing an incremental F test, the sample size for the unconstrained model will be greater than the sample size for the constrained model. F should analyze same sample throughout. # of vars will differ.

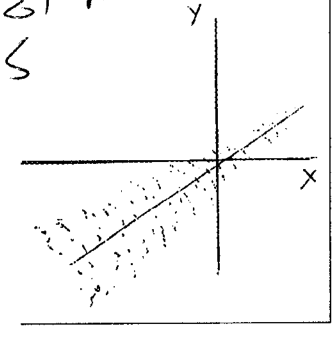
4. Because of skip patterns, women are not asked questions 15-20 in a survey. The researcher should therefore use pairwise deletion of missing data when working with those questions.
F. Clearly, data are not missing at random here.

5. As we saw, larger samples tend to produce larger F values. This is because, the larger the sample, the larger R^2 tends to be. F. A larger N in the F formula numerator leads to bigger F - can be more certain value differs from 0

II. *Short answer.* (15 pts. Each, 45 points total). For each of the following, 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.

1. A researcher wants to regress Y on X. When she plots the data, she gets the following (the diagonal line is the estimated regression line):

Estimates are not BLUE - less stable across samples.



Hetero, use GQ, with $\frac{SSE\ low}{SSE\ high}$ (because σ ↓ when X ↑). Use WLS estimation.

2. X1 and X2 are both measured on scales that run from -7 to +7. Y is the dependent variable.

Multicollinearity

① Nearly identical corr of X_1-X_2 with Y, yet b's are way different

② High X_1-X_2 Corr.

③ F is highly sign., but the individual T's are not

Descriptive Statistics

	Mean	Std. Deviation	N
Y	5.0096	46.5603	100
X1	-.0803	6.4982	100
X2	.0019	7.0156	100

Correlations

		Y	X1	X2
Pearson Correlation	Y	1.000	.423	.406
	X1	.423	1.000	.952
	X2	.406	.952	1.000
Sig. (1-tailed)	Y	.	.000	.000
	X1	.000	.	.000
	X2	.000	.000	.
N	Y	100	100	100
	X1	100	100	100
	X2	100	100	100

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.423[a]	.179	.162	42.6225

a. Predictors: (Constant), X2, X1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	38400.99	2	19200.50	10.569	.000[a]
	Residual	176217.6	97	1816.676		
	Total	214618.6	99			

a. Predictors: (Constant), X2, X1

b. Dependent Variable: Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.232	4.266		1.226	.223
	X1	2.774	2.150	.387	1.290	.200
	X2	.249	1.992	.038	.125	.901

a. Dependent Variable: Y

Multic produces high standard errors & volatile estimates / Could maybe ① Add vars together ② Just use F ③ Drop a var, but be careful about doing this.

Have measurement error problem. Male responses are of lower quality, hence relationships might be attenuated or otherwise distorted. She might appear to be right when she isn't,

3. A psychologist is doing a study of 6th grade students. She believes that the attitudes of young girls are strongly affected by their friends' beliefs. However, for young boys, she thinks that attitudes of friends have little effect on beliefs. She is therefore interesting in making comparisons between boys and girls. When she administers her questionnaire, she notices that girls think very carefully when answering, while boys tend to just rush through items quickly.

Collect better data - maybe use ~~better~~ multiple indicators of concepts

---

III. Computation and interpretation. (35 points; up to 10 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 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:

Model I: Bivariate Regression.

```
                Mean   Std Dev  Label

AIDSIQ     57.000   15.000
USERDRIV     .500     .501
FEMALE       .200     .402
EDUC       10.500    1.800

N of Cases =    500

Correlation:

              AIDSIQ   USERDRIV      FEMALE        EDUC

AIDSIQ         1.000       .400        .330        .350
USERDRIV        .400      1.000        .100        .600
FEMALE          .330       .100       1.000        .100
EDUC            .350       .600        .100       1.000

Equation Number 1    Dependent Variable..    AIDSIQ


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

Variable                B        SE B       Beta          T  Sig T

USERDRIV        11.976048    1.229641    .400000      9.739  .0000
(Constant)      51.011976     .869922                58.640  .0000


------------------------ Variables not in the Equation ------------------------

Variable      Beta In  Partial  Tolerance        VIF  Min Toler        T  Sig T

FEMALE        .292929  .318010    .990000      1.010    .990000    7.478  .0000
EDUC          .171875  .150025    .640000      1.563    .640000    3.383  .0008
-------------------------------------------------------------------------------
```

It has the biggest bivar corr, hence it produces ↑the biggest initial R^2

a. (5 points) Since forward stepwise selection is being used, why was USERDRIV the first variable selected? What variable should be added to the equation next? [HINT. Variables not in the equation tells you what the parameter estimates for a variable would be if it were entered into the model next.]

Female, because it has the biggest + sign T / ↑in R2 if added next.

b. (10 points) Interpret the results from Model I. 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 the regression model? Why or why not?

20%

10.5

So far so good - Those in the user drives program score 12 pts more than those not in - userdriv has the strongest bivar relationship.

Female has biggest effect. Part of the original relationship better userdriv & AidsIQ was due to the fact that those in the user driven program were better educated (as is evidenced by the .6 corr). once educ is controlled for, userdriv still has a high sign. effect though.

Model III: Multivariate regression

```
Multiple R              .50998
R Square                .26008
Adjusted R Square       .25561
Standard Error        12.94175


F =       [1]

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

Variable              B           SE B        Beta    Correl      Tolerance         VIF          T

USERDRIV       8.351967       1.447318     .278956   .400000        .638384       1.566        [2]
FEMALE        10.698092       1.450267     .286709     [3]          .987500       1.013      7.377
EDUC           1.282964         [4]        .153956   .350000        .638384        [5]       3.185
(Constant)    37.213275       3.878287                                                       9.595
```

c. (15 points; up to 10 points extra credit if you get all 5 right) Fill in <u>three</u> of the missing items [1] - [5]. [HINT: If one formula does not seem to be working, try using an alternative formula. Remember that the means and correlations were already presented with Model I.]

d. (5 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 a? What do you think accounts for any seeming discrepancy? [HINT: What does the correlation between USERDRIV and EDUC imply about how similar the two cities were before the study was conducted?]

(1) $F = \frac{R^2 * (N-k-1)}{(1-R^2) * k} = \frac{.26008 * (500-3-1)}{(1-.26008) * 3} = \frac{129}{2.22} = 58.11$

(2) $t = \frac{b_1}{s_{b_1}} = \frac{8.35}{1.44} = 5.86$

(3) .33 (As shown in the corr matrix on last page.

(4) $b_3 = \cancel{T_3 s_{b_3} = 3.185 * 1.282}$

$s_{b_3} = \frac{b_3}{T_3} = \frac{1.28}{3.19} = .40$

(5) $VIF_3 = \frac{1}{Tol_3} = \frac{1}{.638} = 1.6$