

Sociology 593
Exam 2 Answer Key
March 28, 2002

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

1. A variable is called CATHOLIC. This probably means that only Catholics have a score on this variable.

FALSE. More likely is that Catholics are coded 1 on this variable, non-Catholics are coded 0.

2. A researcher believes that Education is a cause of earnings: the more educated you are, the higher your earnings tend to be. Further, she knows that blacks make less, on average, than whites do. This means that the effect of education on earnings must be greater for whites than it is for blacks.

FALSE. The effect of education could be the same for both groups. But, if whites tend to have more years of education, they will also tend to have higher incomes.

3. The correlation between X2 and X4 is .6. Hence, in a path model using standardized variables, the effect of X2 on X4 will be .6.

FALSE. Variables can be correlated for many reasons, both causal and non-causal, e.g. they can be correlated because of direct effects, indirect effects and common causes. Hence, the direct effect need not equal the correlation.

4. A researcher regresses Y on X1, X2 and X3. In reality, X3 is irrelevant and does not belong in the model. The estimates of the betas will therefore be biased.

FALSE. Including extraneous variables does not produce biased estimates, although they can cause standard errors to be higher than is necessary.

5. A researcher computes $SUMX1X2 = X1 + X2$. She regresses Y on X1 and SUMX1X2. She obtains the following. She should conclude that X1 does not affect Y.

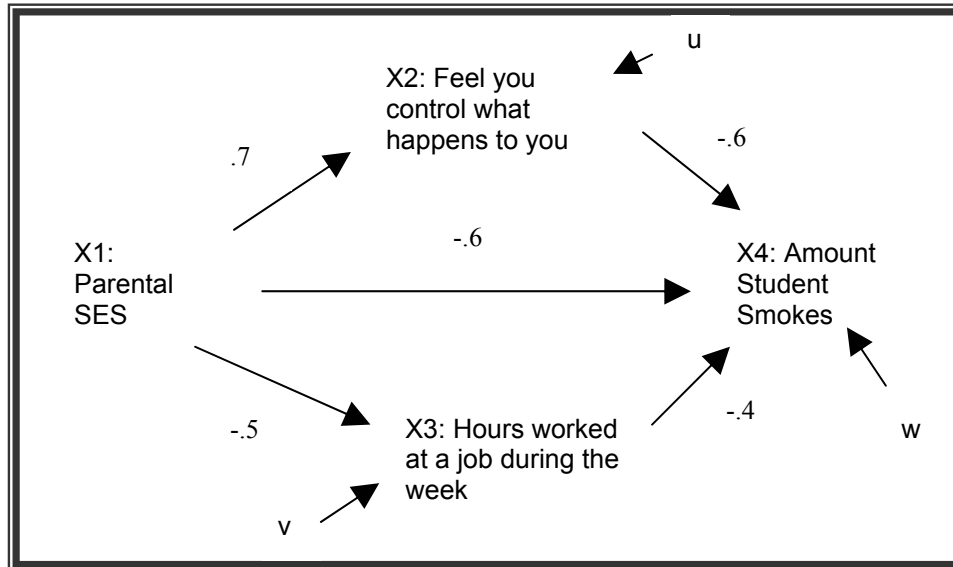
Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	1.292	2.924		.442	.660
X1	-.212	.871	-.052	-.243	.808
SUMX1X2	2.391	.658	.775	3.634	.000

a. Dependent Variable: Y

FALSE. The correct conclusion is that the effect of X1 does not significantly differ from the effect of X2.

II. Path Analysis/Model specification. (30 points). A researcher is interested in the determinants of teenage smoking. She has measures of Parental Socio-economic status, hours worked at paying jobs during the school week, feelings of control over your destiny (“Do you feel you control what happens to you in life, or do you think things just happen?”), and the number of cigarettes smoked per week. All her variables are in standardized form. The estimated value of each path in her model is included in the diagram.



a. Write out the structural equation for each endogenous variable.

$$X2 = .7X1 + u$$

$$X3 = -.5X1 + v$$

$$X4 = -.6X1 - .6X2 - .4X3 + w$$

b. The correlation between X3 (hours worked) and X4 (# of cigarettes smoked) is

$$\begin{aligned}
 E(X_3 X_4) &= \rho_{43} \\
 &= \beta_{41}\beta_{31} + \beta_{42}\beta_{21}\beta_{31} + \beta_{43} \\
 &= (-.6 * -.5) + (-.6 * .7 * -.5) + (-.4) \\
 &= .30 + .21 - .4 = .11
 \end{aligned}$$

Decompose the correlation between X3 and X4 into

- Correlation due to direct effects
-4 (as shown by the path from X3 to X4)

- Correlation due to indirect effects
0. There is no indirect effect of X3 on X4.

- Correlation due to common causes
.51. There are two sources of association due to common causes. X1 is a common cause of both X3 and X4 (-.6 * -.5 = .30). Also, X1 is a cause of X3, and X1 is a cause of X2 which is a cause of X4 (-.5 * .7 * -.6 = .21).

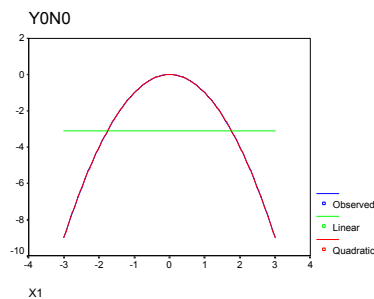
c. There is controversy over the harms and benefits of students working at jobs while attending school. Based on these results, do you think that working at a job tends to increase smoking, decrease smoking, or have little or no effect? Explain how, if variables were erroneously omitted from the above model, a researcher might reach a very different, and incorrect, conclusion.

If you simply regressed X4 on X1, the estimated coefficient would be the same as the correlation between X1 and X4, i.e. .11. You would therefore conclude that working at a job tends to increase the amount you smoke. As the full model shows us (specifically, the effect of X3 on X4), this conclusion is wrong: in reality, working at a job is beneficial in that it reduces the amount a teenager smokes. Suppressor effects are present. Those who work more tend to come from lower SES families. Lower SES causes individuals to smoke more. Also, lower SES results in lower feelings of control, which in turn leads to more smoking. In short, those who work more have characteristics that increase the amount they smoke; but they would smoke even more if they weren't working. Perhaps working reduces the amount of time available for smoking, or produces some sort of incentive to not smoke so much.

III. Short answer. Answer *two* of the following three questions. (25 points each; up to 10 points extra credit if you do all 3).

1. Each of the following describes a nonlinear or nonadditive relationship between variables. Draw a scatterplot that illustrates the relationship. Describe the harms that might result if you simply regressed Y on X, e.g. would values be over-estimated, under-estimated, or what? Indicate the model you think should be estimated, e.g. $E(Y) = \alpha + \beta_1X + \beta_2X^2$. Explain what variables you would need to compute in order to actually estimate the model, e.g. logs of variables, interaction terms.

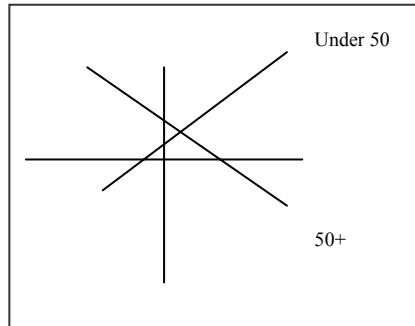
a. Studies have found that moderate amounts of alcohol consumption actually tend to increase life expectancy. However, after a certain point, the more alcohol you consume, the more your life expectancy goes down.



Relationship is curvilinear. A linear regression might lead you to believe there was no relationship whatsoever, and in any even would miss the fact that the relationship changes direction after a certain point. You should therefore estimate a polynomial model. Compute X^2 and then estimate $E(Y) = \alpha + \beta_1X + \beta_2X^2$.

A piecewise model might also work, but I think it is less plausible, i.e. I don't think the effect of drinking would dramatically change all of a sudden. Plus, with a piecewise model, you have to figure out where the break point is.

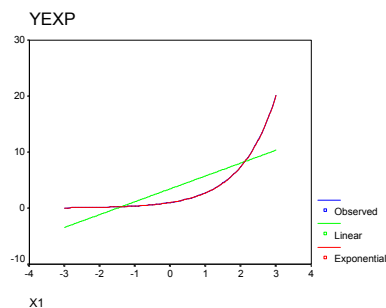
b. A political analyst wants to know how TV ads affect the opinions of voters. She finds that, for voters aged 50 and above, the more TV ads they see, the less they like her candidate. For voters younger than 50, just the opposite is true: the more TV ads they see, the more they like the candidate.



Interaction effects are present. Watching TV ads increases candidate popularity for people under age 50 but has the opposite effect for those over 50. If you estimated a linear regression, you might find no effect at all, or effects that are much weaker than the above suggests (and which were in the wrong direction for one of the groups.) Compute a dummy variable coded 1 = 50+, 0 = Under 50. Then, compute the interaction term $DUMMY_{TV} = DUMMY * TV$. Then, estimate

$$E(Y) = \alpha + \beta_1 X + \beta_{Dummy} Dummy + \beta_{DummyTV} DummyTV$$

c. An employer believes that workers become more skillful and productive the longer they work on the job. Specifically, she believes that each year a worker is on the job, he or she will be 5% more productive than they were the year before.



A growth model is called for. A linear regression would cause you to overestimate the amount of growth early on and then underestimate it later. Compute $\ln(y)$, where y = productivity. Estimate $E(\ln Y) = \alpha + \beta_1 X$.

2. [This problem is modified from Hamilton's *Statistics with Stata 5*.] A researcher believes that better students (as measured by Grade Point Average) drink less than do weaker students. However, she suspects that this is less true for men than for women, i.e. the effect of GPA on drinking is smaller for males than it is for females. She tests her ideas with a survey of undergraduate students collected by Ward and Ault (1990). DRINKING is measured on a 33 point scale, where higher values indicate higher levels of drinking. GPA is the student's Grade Point Average, centered so as to have a mean of zero (higher values indicate better grades). MALE is coded 1 if the student is male, 0 if Female. $MALEGPA = MALE * GPA$. She obtains the following results:

```

REGRESSION
/MISSING LISTWISE/ descriptives
/STATISTICS COEFF OUTS R ANOVA CHA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT drink
/METHOD=ENTER GPA /METHOD=ENTER male /METHOD=ENTER MaleGpa .

```

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
DRINK 33-point drinking scale	18.82	6.739	218
GPA Grade Point Average	.0000	.45917	218
MALE	.4541	.49904	218
MALEGPA	-.0407	.30988	218

Correlations

		DRINK 33-point drinking scale	GPA Grade Point Average	MALE	MALEGPA
Pearson Correlation	DRINK 33-point drinking scale	1.000	-.282	.304	-.170
	GPA Grade Point Average	-.282	1.000	-.178	.687
	MALE	.304	-.178	1.000	-.144
	MALEGPA	-.170	.687	-.144	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	.282 ^a	.080	.075	6.480	.080	18.660	1	216	.000
2	.382 ^b	.146	.138	6.257	.066	16.710	1	215	.000
3	.384 ^c	.148	.136	6.265	.002	.412	1	214	.522

a. Predictors: (Constant), GPA Grade Point Average

b. Predictors: (Constant), GPA Grade Point Average, MALE

c. Predictors: (Constant), GPA Grade Point Average, MALE, MALEGPA

ANOVA^d

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	783.590	1	783.590	18.660	.000 ^a
	Residual	9070.432	216	41.993		
	Total	9854.023	217			
2	Regression	1437.711	2	718.855	18.364	.000 ^b
	Residual	8416.312	215	39.146		
	Total	9854.023	217			
3	Regression	1453.879	3	484.626	12.346	.000 ^c
	Residual	8400.144	214	39.253		
	Total	9854.023	217			

- a. Predictors: (Constant), GPA Grade Point Average
- b. Predictors: (Constant), GPA Grade Point Average, MALE
- c. Predictors: (Constant), GPA Grade Point Average, MALE, MALEGPA
- d. Dependent Variable: DRINK 33-point drinking scale

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	18.821	.439		42.883	.000
	GPA Grade Point Average	-4.138	.958	-.282	-4.320	.000
2	(Constant)	17.215	.578		29.794	.000
	GPA Grade Point Average	-3.453	.940	-.235	-3.673	.000
	MALE	3.536	.865	.262	4.088	.000
3	(Constant)	17.257	.582		29.640	.000
	GPA Grade Point Average	-4.011	1.282	-.273	-3.129	.002
	MALE	3.553	.867	.263	4.100	.000
	MALEGPA	1.212	1.889	.056	.642	.522

- a. Dependent Variable: DRINK 33-point drinking scale

Based on the above results, answer the following questions. Be sure to indicate how the printout supports your arguments.

- a) Is the researcher correct in hypothesizing that better students tend to drink less?

Yes. GPA has a significant negative effect on drinking in all models and the GPA-drinking correlation is also negative

- b) Are there significant differences in the determinants of drinking for men and women? If so, are these differences limited to differences in the intercepts, or does the effect of GPA differ by gender? If differences are found, be specific as to what they are, e.g. how much greater (or weaker) is the effect of GPA on men than it is for women.

There are differences in the intercepts, as you can tell from the T value for male (or else the F change statistic) in Model 2. However, in Model 3, the interaction term is not significant; ergo, counter to what the researcher hypothesized, it appears the effect of GPA is the same for both men and women. (If the interaction term were significant, it would support the researcher's hypothesis: a positive interaction term implies that the effect of GPA is less for men than it is for women. The problem may be that the sample isn't large enough to detect the predicted differences.)

c) Of the three models presented, which one would you say is best, and why? Briefly discuss the substantive implications of your preferred model.

Model 2 is best; all the variables in it are statistically significant. The model says that those with higher GPAs tend to drink less. Also, men tend to drink more than do comparable women. The effect of GPA on drinking is the same for both men and women.

3. A researcher is interested in the following model:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

Using OLS regression, briefly tell her how to go about testing each of the following hypotheses. You don't have to give exact formulas but you should describe the general procedures (e.g. what variables would you need to compute, what models would you need to estimate) and the types of statistics (e.g. T-Test, Global F-test, Incremental F-test) that are appropriate. For example, you might describe the unconstrained and constrained models that need to be estimated, and then say that an incremental F-test should be used to tell whether the two models significantly differ from each other. If necessary, describe any additional variables you would need to compute, such as interaction terms.

- $\beta_1 = 100$

The simplest strategy is just to look at the confidence interval for B_1 . If it includes 100, do not reject the null hypothesis that $B_1 = 100$; otherwise you do reject it. Alternatively, you can do a T-Test: $t = (b - 100)/s_b$. SPSS won't give you this T-value (it tests the hypotheses that $b = 0$) but you can compute it easily enough yourself.

- $\beta_1 = \beta_2$

Estimate a constrained and unconstrained model and then compute an incremental F. The unconstrained model is the model given above, i.e.

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

This model is unconstrained in that B_1 and B_2 need not equal each other. To estimate the constrained model, we first compute $\text{Sum}X_1X_2 = X_1 + X_2$. We then estimate

$$Y = \alpha + \beta_1 \text{Sum}X_1X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

The model is constrained in that the effects of X_1 and X_2 are not allowed to differ. An incremental F test is then used to see whether this constraint is justified. If the incremental F is not statistically significant, we conclude that $B_1 = B_2$.

- $\beta_3 = \beta_4 = 0$

The unconstrained model is again

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

In the constrained model, we drop X3 and X4, which in effect means we are constraining their effects to be zero:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

If the incremental F value is not significant, we conclude that the constraint is legitimate, i.e. $B_3 = B_4 = 0$

- Suppose X4 is a dummy variable coded 1 if Black, 0 Otherwise. How would you test the hypothesis that the value of β_1 is different for blacks and whites? Assume that the effects of X2 and X3 are the same for both groups.

The simplest approach is to compute the interaction term $X_1 X_4 = X_1 * X_4$. We then estimate

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_{14} X_{14} + \varepsilon$$

If the T value for B14 is statistically significant, we conclude that the effect of X1 is different for blacks and whites; otherwise we conclude the effect is not different.