

Sociology 63993

Exam 2 Answer Key

April 1, 2011

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

1. A researcher computes a variable $X_4 = X_2 + X_3$. She then estimates the following two models using OLS regression:

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

$$Y = \beta_1 X_1 + \beta_4 X_4 + \varepsilon$$

She can use an incremental F test to determine which of these two models is better.

True. Since $X_4 = X_2 + X_3$, the incremental F is a test of whether or not $\beta_2 = \beta_3$.

2. A researcher runs the following:

```
. webuse nhanes2f, clear
. gen femage = female * age
. reg health female age femage
```

Source	SS	df	MS			
Model	2069.28161	3	689.760537	Number of obs =	10335	
Residual	12965.7398	10331	1.2550324	F(3, 10331) =	549.60	
				Prob > F	= 0.0000	
				R-squared	= 0.1376	
				Adj R-squared	= 0.1374	
Total	15035.0214	10334	1.4549082	Root MSE	= 1.1203	

health	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	-.2752255	.0648373	-4.24	0.000	-.4023191	-.1481319
age	-.0280887	.0009315	-30.15	0.000	-.0299146	-.0262627
femage	.0043295	.0012822	3.38	0.001	.0018162	.0068428
_cons	4.78594	.0469616	101.91	0.000	4.693886	4.877994

These results show that age has a negative effect on the health of males and a positive effect on the health of females.

False. The effect of age is less negative for females ($-.0280887 + .0043295 = -.0237592$) but it is still negative.

3. A researcher has included several extraneous variables in her model. The larger her sample, the more serious this problem will be.

False. Adding extraneous variables increases standard errors. Larger sample sizes decrease standard errors.

4. A researcher regresses income on education. She does not include any dummy variables or interaction terms involving gender. One implication of this model is that, if it is true, the mean income for men will be the same as the mean income for women.

False. If men and women differ in their mean levels of education, they will also differ in their mean incomes.

5. A researcher is interested in the relationship between bmi (Body Mass Index) and health. She does the following:

```
. webuse nhanes2f, clear
. gen bmi = weight/ (height/100)^2
. gen bmi2 = bmi * bmi
. reg health bmi bmi2
```

Source	SS	df	MS			
Model	316.928298	2	158.464149	Number of obs =	10335	
Residual	14718.0931	10332	1.4245154	F(2, 10332) =	111.24	
				Prob > F	= 0.0000	
				R-squared	= 0.0211	
				Adj R-squared	= 0.0209	
				Root MSE	= 1.1935	

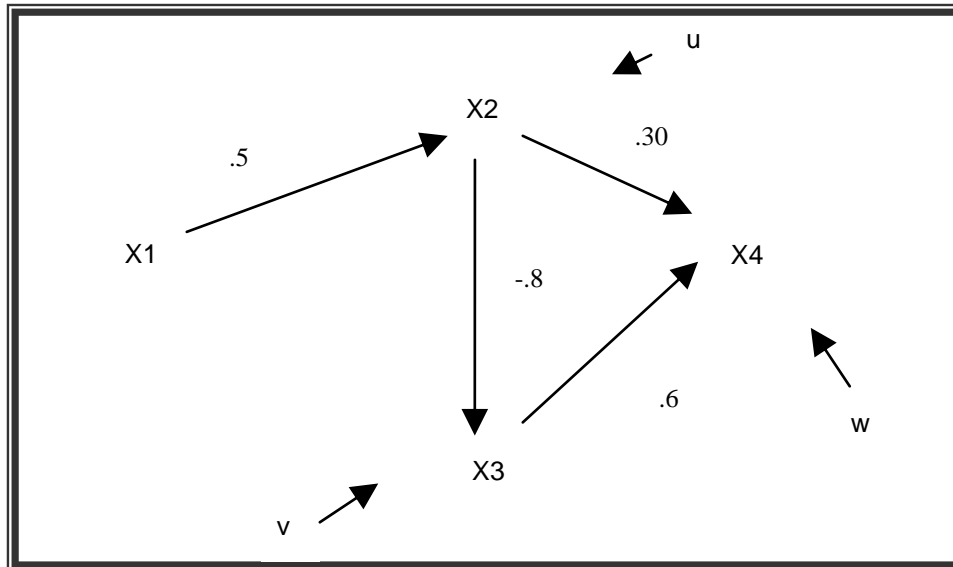
health	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
bmi	.0072049	.0152456	0.47	0.637	-.0226794	.0370892
bmi2	-.0007416	.0002646	-2.80	0.005	-.0012601	-.000223
_cons	3.731409	.2147848	17.37	0.000	3.31039	4.152429

Based on these results, she should conclude that bmi is not related to health.

False. The results indicate that there is a curvilinear relationship between bmi and health. Increases in body mass index are good up to point, but after that further increases are harmful. (In other words, it isn't good to be obese.)

II. Path Analysis/Model specification (25 pts).

A sociologist believes that the following model describes the relationship between X1, X2, X3, and X4. All her variables are in standardized form. The estimated value of each path in her model is included in the diagram.



a. (5 pts) Write out the structural equation for each endogenous variable, using both the names for the paths (e.g. β_{42}) and the estimated value of the path coefficient.

$$X_2 = \beta_{21}X_1 + u = .5X_1 + u$$

$$X_3 = \beta_{32}X_2 + v = -.8X_2 + v$$

$$X_4 = \beta_{42}X_2 + \beta_{43}X_3 + w = .3X_2 + .6X_3 + w$$

b. (10 pts) Part of the correlation matrix is shown below. Determine the complete correlation matrix. (Remember, variables are standardized. You can use either normal equations or Sewell Wright, but you might want to use both as a double-check.)

	x1	x2	x3	x4
x1	1.0000			
x2	0.5000	1.0000		
x3	?	?	1.0000	
x4	?	?	?	1.0000

Here is the uncensored output:

	x1	x2	x3	x4
x1	1.0000			
x2	0.5000	1.0000		
x3	-0.4000	-0.8000	1.0000	
x4	-0.0900	-0.1800	0.3600	1.0000

To confirm that this reproduces the estimated path coefficients:

```
. pathreg (x2 x1) (x3 x2 x1) (x4 x3 x2 x1)
```

x2	Coef.	Std. Err.	t	P> t	Beta
x1	.5	.0874818	5.72	0.000	.5
_cons	8.90e-09	.0870433	0.00	1.000	.

n = 100 R2 = 0.2500 sqrt(1 - R2) = 0.8660

x3	Coef.	Std. Err.	t	P> t	Beta
x2	-.8	.0703452	-11.37	0.000	-.8
x1	3.08e-09	.0703452	0.00	1.000	3.08e-09
_cons	3.66e-09	.0606154	0.00	1.000	.

n = 100 R2 = 0.6400 sqrt(1 - R2) = 0.6000

x4	Coef.	Std. Err.	t	P> t	Beta
x3	.6	.1557167	3.85	0.000	.6
x2	.3	.164795	1.82	0.072	.3
x1	-8.07e-09	.1078837	-0.00	1.000	-8.07e-09
_cons	-7.66e-09	.0929617	-0.00	1.000	.

n = 100 R2 = 0.1620 sqrt(1 - R2) = 0.9154

c. (5 pts) Decompose the correlation between X3 and X4 into

- Correlation due to direct effects

.6

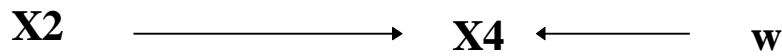
- Correlation due to indirect effects

0

- Correlation due to common causes

-.24

d. (5 pts) Suppose the above model is correct, but instead the researcher believed in and estimated the following model:



What conclusions would the researcher likely draw? In particular, what would the researcher conclude about the effect of changes in X2 on X4? Discuss the consequences of this mis-specification, and in what ways, if any, the results would be misleading. Why would she make these mistakes?

In the correctly specified model the effect is .3, but in the incorrectly specified model the effect is -.18 (the same as the correlation between the variables). This occurs because the correlation that is due to the common cause is incorrectly attributed to the direct effect of X2 on X4. The researchers would think that increasing X2 will tend to increase X4, when in reality just the opposite is true.

III. Group comparisons (25 points). This week, the Supreme Court heard a landmark gender discrimination case against retail giant Wal-Mart. The plaintiffs based their case, in part, on work done by Sociologist William Bielby. Bielby's devastating arguments have put the fear of God into another company making it wonder if it, too, might face such a lawsuit. It has therefore conducted its own study of gender equity within its work force, collecting data from a random sample of 7500 of its employees on the following variables:

Variable	Description
pay	Annual Salary (in thousands of dollars)
qual	A qualifications scale that the company has constructed and believes to be very valid. It takes into account such things as past performance, aptitude test scores, education, and years of experience. The scale ranges from -40 to 40 and has been centered to have a mean of 0 (i.e. 0 means average qualifications; and the higher the score, the more qualified the person is)
female	Coded 1 if female, 0 if male
femqual	female * qual

The results of the analysis are as follows:

. ttest pay, by(female)

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	3572	78.1415	.2298254	13.73579	77.6909	78.5921
1	3928	47.23287	.2309217	14.47273	46.78013	47.68561
combined	7500	61.95362	.241623	20.92516	61.47997	62.42727
diff		30.90863	.3266069		30.26839	31.54887

diff = mean(0) - mean(1) t = 94.6356
 Ho: diff = 0 degrees of freedom = 7498

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
 Pr(T < t) = 1.0000 Pr(|T| > |t|) = 0.0000 Pr(T > t) = 0.0000

. nestreg: reg pay qual female femqual

Block 1: qual

Source	SS	df	MS	Number of obs = 7500		
Model	1453171.17	1	1453171.17	F(1, 7498)	= 5952.86	
Residual	1830359.2	7498	244.112991	Prob > F	= 0.0000	
Total	3283530.37	7499	437.862431	R-squared	= 0.4426	
				Adj R-squared	= 0.4425	
				Root MSE	= 15.624	

pay	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
qual	1.438618	.0186459	77.15	0.000	1.402067	1.475169
_cons	61.95362	.1804117	343.40	0.000	61.59996	62.30728

Block 2: female

Source	SS	df	MS	Number of obs = 7500		
Model	1922795.12	2	961397.559	F(2, 7497)	= 5296.84	
Residual	1360735.25	7497	181.503969	Prob > F	= 0.0000	
Total	3283530.37	7499	437.862431	R-squared	= 0.5856	
				Adj R-squared	= 0.5855	
				Root MSE	= 13.472	

pay	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
qual	.6212315	.0227315	27.33	0.000	.5766715	.6657916
female	-22.40072	.4403823	-50.87	0.000	-23.26399	-21.53744
_cons	73.68562	.2782026	264.86	0.000	73.14027	74.23098

Block 3: femqual

Source	SS	df	MS	Number of obs =	7500
Model	1939531.43	3	646510.478	F(3, 7496) =	3605.84
Residual	1343998.94	7496	179.295483	Prob > F =	0.0000
				R-squared =	0.5907
				Adj R-squared =	0.5905
Total	3283530.37	7499	437.862431	Root MSE =	13.39

	pay	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
	qual	.8329084	.0314714	26.47	0.000	.7712156 .8946012
	female	-22.3506	.4377256	-51.06	0.000	-23.20866 -21.49253
	femqual	-.4367667	.0452069	-9.66	0.000	-.5253848 -.3481486
	_cons	72.16734	.3180414	226.91	0.000	71.54389 72.79079

Block	F	Block df	Residual df	Pr > F	R2	Change in R2
1	5952.86	1	7498	0.0000	0.4426	
2	2587.40	1	7497	0.0000	0.5856	0.1430
3	93.34	1	7496	0.0000	0.5907	0.0051

. ttest qual, by(female)

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]
0	3572	7.172654	.1191291	7.119893	6.939086 7.406222
1	3928	-6.522586	.1050536	6.584106	-6.728551 -6.316621
combined	7500	3.32e-08	.1117328	9.676342	-.2190275 .2190276
diff		13.69524	.1582456		13.38503 14.00545

diff = mean(0) - mean(1) t = 86.5442
 Ho: diff = 0 degrees of freedom = 7498

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
 Pr(T < t) = 1.0000 Pr(|T| > |t|) = 0.0000 Pr(T > t) = 0.0000

The initial t-test shows that men make substantially more than women. The company then does additional analyses to find out why. It wants your help in answering the following:

- a) (15 pts) The researchers estimate a series of models. Which of the models do you think is best, and why? What do these models tell us about how qualifications and gender affect pay?

The third and final model provides the best fit. It says that both the intercepts and the slopes differ by gender. Because qual is centered, we know that the average woman makes \$22,000 less than the average man, even after controlling for qualifications. Further, for women, qualifications have an effect that is less than half as large as it is for men (each qualification point is worth, on average, about \$833 for men, but only about \$396 for women).

- b) (10 pts) Suppose the company was sued on the basis that it discriminated against women. What evidence, if any, do you think the company would cite in its defense? What evidence, if any, would its critics cite? Consider both the t-tests and the regression analyses in your answer. If you were the president of the company, would these results make you be worried about a lawsuit?

The company would no doubt note that, on average, women are less qualified than men (by about 13.7 points, as the last t-test shows). Critics will no doubt note the evidence raised in point A, namely that a woman with average qualifications earns \$22,000 less than a similarly qualified man, and women are only rewarded half as much for their qualifications as men are. If I were the president, I would be very worried about a lawsuit.

IV. Short answer. Answer *both* of the following questions. (15 points each, 30 points total.) In each of the following problems, a researcher runs through a sequence of commands. Explain why she didn't stop after the first command, i.e. explain what the purpose of each subsequent command was, what it told her, and why she did not run additional commands after the last one. If she had stopped after the first command, what would the consequences have been, i.e. in what ways would her conclusions have been incorrect or misleading?

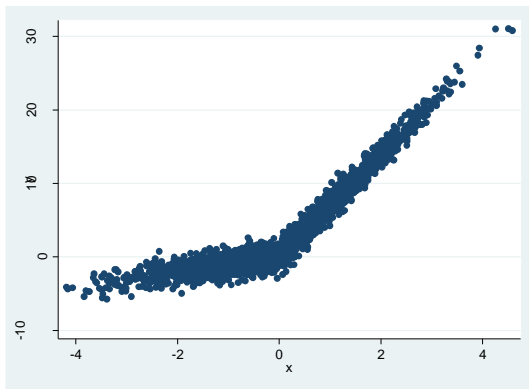
1.

```
. reg y x
```

Source	SS	df	MS	Number of obs =	2293
Model	68744.4388	1	68744.4388	F(1, 2291) =	9754.77
Residual	16145.2885	2291	7.04726691	Prob > F =	0.0000
				R-squared =	0.8098
				Adj R-squared =	0.8097
Total	84889.7273	2292	37.0374028	Root MSE =	2.6547

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
x	3.94874	.0399807	98.77	0.000	3.870337 4.027142
_cons	3.328859	.0554381	60.05	0.000	3.220145 3.437573

```
. scatter y x
```



```
. mkspline xlow 0 xhigh = x
```

. reg y xlow xhigh

Source	SS	df	MS			
Model	82602.9569	2	41301.4785	Number of obs =	2293	
Residual	2286.77032	2290	.998589661	F(2, 2290) =	41359.81	
Total	84889.7273	2292	37.0374028	Prob > F =	0.0000	
				R-squared =	0.9731	
				Adj R-squared =	0.9730	
				Root MSE =	.99929	

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
xlow	1.02005	.029061	35.10	0.000	.9630619	1.077039
xhigh	6.933698	.0294706	235.28	0.000	6.875907	6.99149
_cons	.0479089	.0348016	1.38	0.169	-.020337	.1161549

The scatterplot strongly suggests that the effect of X is not the same across the range of X. In particular, the effect of X becomes much greater once X goes past 0. The mkspline computation and the subsequent regression shows that between -4 and 0, the slope of X is 1, and after that the slope of X is about 7. The R² is extremely high and the results are consistent with the scatterplot so the researcher probably thought it was ok to stop at that point. If the researcher had not done the 2nd regression, the researcher would have concluded that the effect of X was about 4 throughout its range, when in reality the effect of X is sometimes much less than that and sometimes much more.

2.

. reg y x

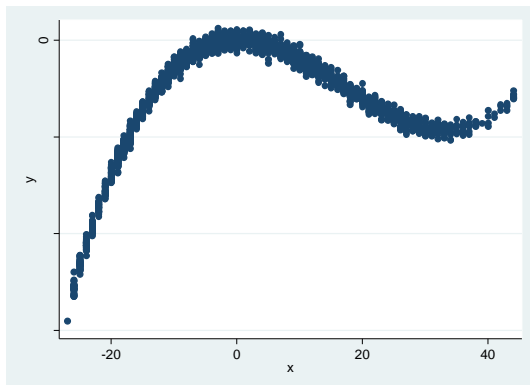
Source	SS	df	MS			
Model	4.7856e+10	1	4.7856e+10	Number of obs =	2293	
Residual	2.8433e+11	2291	124108295	F(1, 2291) =	385.60	
Total	3.3219e+11	2292	144933916	Prob > F =	0.0000	
				R-squared =	0.1441	
				Adj R-squared =	0.1437	
				Root MSE =	11140	

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	272.3303	13.86839	19.64	0.000	245.1343	299.5262
_cons	-12109.01	232.6475	-52.05	0.000	-12565.23	-11652.79

. ovtest

Ramsey RESET test using powers of the fitted values of y
 Ho: model has no omitted variables
 F(3, 2288) = 94189.71
 Prob > F = 0.0000

. scatter y x



```
. gen x2 = x^2
. gen x3 = x^3
. reg y x x2 x3
```

Source	SS	df	MS			
Model	3.2990e+11	3	1.0997e+11	Number of obs =	2293	
Residual	2.2843e+09	2289	997934.435	F(3, 2289) =	.	
Total	3.3219e+11	2292	144933916	Prob > F =	0.0000	
				R-squared =	0.9931	
				Adj R-squared =	0.9931	
				Root MSE =	998.97	

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	.3380926	2.407808	0.14	0.888	-4.383621	5.059806
x2	-50.07069	.096471	-519.02	0.000	-50.25987	-49.88151
x3	.9974231	.0039199	254.45	0.000	.9897362	1.00511
_cons	24.94746	30.93394	0.81	0.420	-35.71402	85.60894

```
. ovtest
```

```
Ramsey RESET test using powers of the fitted values of y
Ho: model has no omitted variables
F(3, 2286) = 0.87
Prob > F = 0.4561
```

The ovtest command indicated that higher powers of X should be included in the model. The subsequent scatterplot indicated that there were two bends in the data, suggesting that X^2 and X^3 should be added to the model. The final ovtest indicated that no more higher powers were needed so the researcher stopped. If the researcher had not done the initial analyses she would have erroneously concluded that the effect of X was linear and positive when in fact the relationship is curvilinear.

Appendix: Stata Code

```
version 11.1

* I-2 - T/F
webuse nhanes2f, clear
gen femage = female * age
reg health female age femage

* I-5 - T/F
webuse nhanes2f, clear
gen bmi = weight/ (height/100)^2
gen bmi2 = bmi * bmi
reg health bmi bmi2

* II - Path Analysis
clear all
matrix input corr = (1, .5, -.4, -.09 \ .5, 1, -.8, -.18 \ -.4, -.8, 1, .36 \ -.09, -.18, .36, 1)
corr2data x1 x2 x3 x4, n(100) corr(corr) double
corr
*** Double-check results
pathreg (x2 x1) (x3 x2 x1) (x4 x3 x2 x1)

* III - Interaction Effects, Group differences
*** Set up data
webuse nhanes2f, clear
set seed 123
sample 7500, count
gen pay = weight - 3*female - .1*female*height
sum height
gen qual = height - r(mean)
gen femqual = female * qual
*** Do analyses
ttest pay, by(female)
nestreg: reg pay qual female femqual
ttest qual, by(female)

* IV-1 - Nonlinear relationships
*** Set up data
use "http://www.indiana.edu/~jlsoc/stata/spex_data/ordwarm2.dta", clear
corr2data e1 e2
gen x = warm + e1
sum x
replace x = x - r(mean)
gen y = x if x < 0
replace y = 7*x if x > 0
replace y = y + e2
*** Do analyses
reg y x
scatter y x
mkspline xlow 0 xhigh = x
reg y xlow xhigh

* IV-2 - Nonlinear relationships
*** Set up data
use "http://www.indiana.edu/~jlsoc/stata/spex_data/ordwarm2.dta", clear
corr2data e, sd(1000)
sum age
gen x = age - r(mean)
gen y = x - (50 * x^2) + (x^3) + e
*** Do analyses
reg y x
scatter y x
gen x2 = x^2
gen x3 = x^3
reg y x x2 x3
ovtest
```