BOSTON COLLEGE
Department of Economics

EC 228 02
Econometrics
Spring 2003
Prof. Baum, Mr. Barbato

Midterm Exam
10 March 2003

Exam ends at 2:55 P.M. **sharp**. Answer all questions. Total of 95 points. Partial credit given for partial answers.

1. (35 pts) Briefly explain each term. Use examples to illustrate your explanation.
a. bias of an estimator
b. $p-$value
c. "OLS is BLUE"
d. homoskedasticity
e. asymmetry of specification error (see pp.8–10, Chapter 3 notes)
f. joint hypothesis test
g. double–log model

2. (12 pts) Compare the simple regression $y = \beta_0 + \beta_1 X_1$, in which you estimate parameters b_0 and b_1, with the multiple regression $y = \delta_0 + \delta_1 X_1 + \delta_2 X_2$ in which you estimate parameters d_0, d_1 and d_2. Describe two circumstances in which $b_1 = d_1$ and explain each of these circumstances.

(p.79) If the partial effect of X_2 is zero in the sample, so that d_2 is zero, the two coefficients will be equal. Likewise, if X_1 and X_2 are uncorrelated in the sample, they will be equal. In all other circumstances—in which X_2 has some explanatory power, and has a nonzero correlation with X_1—they will not be equal.

3. (24 pts) The following models of $narr86$, the number of times a sample of men were arrested in 1986, yields:

	(1) narr86	(2) narr86	(3) narr86
pcnv	-0.150	-0.151	-0.130
	(3.67)**	(3.69)**	(3.22)**
ptime86	-0.034	-0.037	-0.041
	(4.01)**	(4.25)**	(4.72)**
qemp86	-0.104	-0.103	-0.095
	(10.02)**	(9.94)**	(9.13)**
avgsen		0.007	0.004
		(1.57)	(0.80)
black			0.338
			(7.43)**
hispan			0.203
			(5.10)**
Constant	0.712	0.707	0.585
	(21.56)**	(21.32)**	(16.25)**
Observations	2725	2725	2725
Adjusted R^2	0.04	0.04	0.06
F-test:B=H=0			33.59
Pr> F			0.00

Absolute value of t *statistics in parentheses*

* significant at 5%; ** significant at 1%

pcnv measures the proportion of arrests prior to 1986 that led to convictions, *avgsen* is the average sentence length served for prior convictions, *ptime*86 is months spent in prison in 1986, *qemp*86 is quarters during which the man was employed in 1986, *black* and *hispan* are "dummy variables" equal to one if the individual is, respectively, black or hispanic.

a. In the model in column (1), do the coefficients have the expected signs? Explain how you would interpret each of the coefficients, and whether they make sense in this context.

Each of the slope coefficients have the expected sign. An increase in the proportion of convictions lowers the predicted number of arrests. Similarly, those who spent more time in prison are less likely to reoffend, and an increase in legal employment leads to a lower predicted number of arrests.

b. Compare the models in columns (1) and (2). Which do you prefer, and why? How do you interpret the coefficient on *avgsen*?

Adding the *avgsen* variable has very little effect on the model, and its coefficient is not significantly different from zero. The positive coefficient is unexpected on theoretical grounds. Given these weaknesses, Model 1 is to be preferred on the basis of parsimony.

c. Does the model of column (3) improve upon the model of column (1)? On what basis do you make that judgment?

The latter model is substantially more informative; it has a higher adjusted R^2, and each of the explanatory factors in the prior model retains sign and significance.

d. How do you interpret the $F-$test for the coefficients of Black and Hispan? What conclusion do you draw from these coefficients about the likelihood of being arrested?

The joint test of coefficients *black* and *hispan* being zero is decisively rejected, and they have the expected signs, indicating that members of minority groups may be arrested more often.

4. (24 pts) The following models were estimated for $stndfl$, the standardized outcome on a final exam (that is, demeaned score divided by the standard deviation) in Micro Principles achieved by a sample of college students.

atndrte is the percent of classes attended; *priGPA* is the student's GPA prior to this term; ACT is their college–entry test score; *hwrte* is the percent of homeworks turned in; $frosh$ and *soph* are "dummy variables" equal to one if the student is a freshman or sophomore, respectively. The squares of *priGPA* and ACT were computed as $pGPA2$ and $ACT2$, respectively, while *pa* is the interaction $priGPA \bullet atndrte$.

	(1)	(2)	(3)
	stndfnl	stndfnl	stndfnl
atndrte	0.005	-0.007	-0.002
	(2.25)**	(0.66)	(0.56)
priGPA	0.402	-1.629	-1.922
	(5.14)***	(3.39)***	(3.93)***
ACT	0.084	-0.128	
	(7.54)***	(1.30)	
pGPA2		0.296	0.492
		(2.93)***	(5.43)***
ACT2		0.005	
		(2.08)**	
pa		0.006	
		(1.29)	
hwrte			0.004
			(1.72)*
frosh			-0.120
			(1.07)
soph			-0.196
			(2.12)**
Constant	-3.344	2.050	1.484
	(11.18)***	(1.51)	(2.34)**
Observations	680	680	674
Adjusted R^2	0.20	0.22	0.17
F-test:sqrs=0		7.34	
Prob$>F$		0.00	

Absolute value of *t statistics in parentheses*

* significant at 10%; ** significant at 5%; *** significant at 1%

a. How successful is model 1 in explaining students' test scores? How do you interpret each of the coefficients?

The simple model is quite successful, with an adjusted R^2 of 0.20. Students who attend classes more frequently, have a higher GPA or scored higher on the ACT are all predicted to do better on the exam.

b. What is the purpose of including the squared terms in model 2?

Interpret the coefficients on $priGPA$ and ACT in model 2 vis-à-vis their counterparts in model 1. The $F-$test in column 2 tests the squared terms. What do you conclude from that test?

The squared terms allow for nonlinearities in the relations between these two variables and the dependent variable. In this context, $\frac{\partial stndfl}{\partial ACT} = -0.128 + 0.005ACT$ and $\frac{\partial stndfl}{\partial priGPA} = -1.629 + 0.296GPA + 0.006$. This does not mean the effects of these variables on the exam score are negative; we must evaluate these derivatives at reasonable values of ACT and $priGPA$ to calculate their effects. The $F-$test indicates that the squared terms are jointly significant, and rejects linearity in each case in favor of a nonlinear (curvilinear) relationship.

c. How do you interpret the coefficient of the interaction term pa? Ignoring its lack of statistical significance, what does it add to the model from a conceptual standpoint?

The interaction term allows the effects of $priGPA$ and $atndrte$ to be nonlinear, depending on the level of the other variable. As above, the coefficient on pa enters the expression for the effect of $priGPA$ on $stndrte$.

d. Compare model 1 and model 3. Which do you prefer? Why? How do you interpret each of the added terms in model 3?

Model 3 is worse in the sense that adjusted R^2 has fallen with the inclusion of the three "dummy variables" and the removal of the ACT variables. Since the squared ACT term was significant, these variables should be left in the equation. Perhaps the *hwrte* and *soph* dummies should be included as well in another specification. The *hwrte* coefficient indicates that doing the homeworks improves test scores; both freshman and sophomore students do worse than upperclass students, although the effect for freshmen is not significant.