Introduction to Econometrics (Econ 107)
Drake University, Fall 2007
William M. Boal

Signature: _____

Printed name: _____

MIDTERM EXAMINATION #3 VERSION A
"Multiple Regression With Cross-Sectional Data"
November 13, 2007

INSTRUCTIONS: This exam is closed-book, closed-notes. You may use a calculator for this exam, but not a graphing calculator or a calculator with alphabetical keys. Point values for each question are noted in brackets. Tables of the t distribution, F distribution, and chi-square distribution are attached.

NOTATION: In this exam, $\hat{\beta}_j$ denotes the least-squares coefficient estimators of the line $y_i = \beta_1 + \beta_2 x_{i2}, +\ldots+ \beta_K x_{iK} + \varepsilon_i$. The least-squares fitted value is denoted $\hat{y}_i$. The least-squares residual is denoted $\hat{\varepsilon}_i$. The sample size is denoted n. The true or population value of the variance of the unobserved error term ε_i is denoted σ^2. The (unbiased) least-squares estimator of σ^2 is denoted $\hat{\sigma}^2$. The sample mean of y is denoted $\bar{y}$.

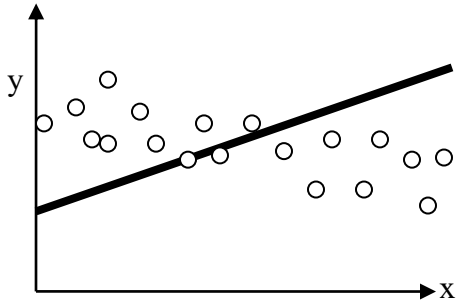
I. MULTIPLE CHOICE: Circle the one best answer to each question. Feel free to use margins for scratch work [1 pt each—10 pts total]

(1) Suppose we want to estimate the effect of the number of police officers per population on the crime rate. Poverty also has an effect on the crime rate, but we omit poverty from the equation for lack of data. Suppose police officers have a negative effect and poverty has a positive effect on crime, and police officers and poverty are positively correlated with each other in our data. Then omitting poverty from the equation will cause the least-squares estimator of the coefficient of the number of police officers

a. to be biased up (toward zero).
b. to be biased down (away from zero).
c. to be unbiased.
d. Cannot be determined from information given.

(2) In the graph below, the solid line is the true population regression line and the circles are observations in the sample. Which assumption appears to be violated in this sample?

a. $E(\varepsilon_i|x_i) = 0$.
b. Homoskedasticity: $Var(\varepsilon_i) = \sigma^2$, a constant.
c. No autocorrelation: $Cov(\varepsilon_i,\varepsilon_j)=0$ for $i \neq j$.
d. All of the above.
e. None of the above.



(3) Adding another regressor to a regression equation will necessarily *increase*
a. the estimated coefficients.
b. the t statistics of the regressors.
c. the R^2 value.
d. Theil's adjusted R^2 value.
e. the sum of squared residuals.

(4) If two regressors x_{i2} and x_{i3} are closely *but not perfectly* correlated, then the least-squares estimators of their coefficients
a. will have large standard errors.
b. will be zero.
c. cannot be computed.
d. will be biased.
e. will be inconsistent.

(5) The equation
$$y_i = 3.2 + 1.5\, x_{i2} + 1.2\, x_{i3}$$
implies that, holding x_{i3} constant, a one-unit increase in x_{i2} will cause y_i to increase by about
a. 1.2 units.
b. 1.5. units.
c. 3.2 units.
d. 2.7 units.
e. 5.9 units.

(6) The equation
$$\ln(wage) = 2.8 + 0.09\, educ$$
implies that if *educ* increases by one unit, then *wage* will increase by about
a. \$9.00.
b. \$1.09.
c. \$0.09.
d. 9.0 percent.
e. 0.09 percent.

(7) The equation
$$y_i = 2.0 + 2.5\, x_{i2} + 0.07\, x_{i2}^2$$
implies that a one-unit increase in x_{i2} will cause y_i to increase by about
a. 0.2 units.
b. 2.0 units.
c. 2.5 units.
d. $(2.5 + 0.14x_{i2})$ units.
e. $(4.5 + 0.07x_{i2})$ units.

(8) Suppose Q = quantity demanded, P = price of the good, and I = consumer income. In which specification does 0.8 equal the price elasticity of demand?
a. $\ln(Q_i) = 4.5 - 0.8\, P_i + 1.1\, I_i$.
b. $\ln(Q_i) = 6.7 - 0.8 \ln(P_i) + 1.1 \ln(I_i)$.
c. $Q_i = 85.3 - 0.8 \ln(P_i) + 1.1 \ln(I_i)$.
d. $Q_i = 166.1 - 0.8\, P_i + 1.1\, I_i$.
e. $Q_i = 94.5 - 0.8\, (P_i/I_i)$.

(9) Suppose we wish to estimate the effect of income on health care spending using data on states:
$$y_i = \beta_1 + \beta_2\, x_i + \varepsilon_i$$
If y_i is defined as *total* health care spending and x_i is defined as *total* income in the state, then we should suspect that the variance of the error term ε_i might be
a. proportional to state population.
b. inversely proportional to state population.
c. constant and unrelated to state population.
d. zero for all observations.
e. infinite.

(10) Under the null hypothesis of no heteroskedasticity, the Goldfeld-Quandt test statistic is close to
a. minus one.
b. zero.
c. one.
d. two.
e. four.

II. MULTIPLE ANSWER: The questions below may have more than one correct answer. Write "YES" next to all correct answers and "NO" next to all incorrect answers.

(1) [6 pts] If we estimate the equation $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$ by ordinary least squares, then which of the following sums are necessarily zero, regardless of the data?

a. $\sum x_{i2} x_{i3}$ ☐
b. $\sum x_{i2} \hat{\varepsilon}_i$ ☐
c. $\sum \hat{\varepsilon}_i$ ☐
d. $\sum \hat{\varepsilon}_i^2$ ☐
e. $\sum x_{i2} y_i$ ☐
f. $\sum x_{i3} \hat{y}_i$ ☐

(2) [4 pts] The variance of the least-squares slope estimator $\hat{\beta}_j$ is smaller, and thus the true value of β_j is estimated more precisely,

a. the larger the sample size. ☐
b. the larger the variance of the error term σ^2. ☐
c. the greater the variation of the x_{ij} values around the sample mean $\bar{x}_j$. ☐
d. the more closely correlated x_{ij} is with the other regressors. ☐

(3) [4 pts] The error term in a regression equation (ε_i) will be correlated with a regressor (x_i) if

a. the dependent variable (y) is measured with error. ☐
b. the regressor (x) is measured with error. ☐
c. an important regressor is omitted from the equation that happens to be correlated with the included regressors (x). ☐
d. the error term (ε_i) is heteroskedastic. ☐

(4) [4 pts] If the error term in a regression equation (ε_i) is heteroskedastic, then the

a. least squares estimators ($\hat{\beta}_j$) will be biased. ☐
b. least squares estimators ($\hat{\beta}_j$) will be inconsistent. ☐
c. the standard errors will be incorrect. ☐
d. the t and F statistics will be invalid. ☐

III. PROBLEMS: Please write your answers in the boxes on this question sheet. Show your work and circle your final answers.

(1) [Analysis of variance table, R^2, F-test: 20 pts] A regression program computed the following analysis-of-variance (ANOVA) table:

	Degrees of freedom ("DF")	Sums of squares ("SS")	Mean squares ("MS")
Regression (or "Model" or "Explained")	3	75	25
Residual (or "Error")	18	72	4
Total	21	147	7

a. What is the sample size?

b. How many β coefficients were estimated, including the intercept?

c. What is the unbiased estimate of the variance of the error term?

d. Compute the value of R^2 (sometimes called the "coefficient of determination") to at least three significant digits.

e. Compute the value of Theil's adjusted R^2 (sometimes called "$\overline{R}^2$") to at least three significant digits.

f. [10 pts] Test the joint null hypothesis that all the coefficients except the intercept are zero (against the alternative hypothesis that at least one of these coefficients is not zero) at 5% significance. Give the value of the test statistic, its degrees of freedom, the critical point, and your conclusion (whether you can reject the null hypothesis).

Degrees of freedom in numerator = _____ Degrees of freedom in denominator = _____

Value of F statistic = _____ Critical point = _____

Reject null hypothesis? _____.

(2) [LS prediction: 12 pts] Using a sample of 400 houses, we have estimated an equation relating the selling price of a house (in thousands of dollars) to its size in square feet, its number of bathrooms, and whether the house has an attached two-car garage:

price	=	66.2	+	0.050 *size*	+	5.1 *baths*	+	22.4	*garage*
		(12.3)		(0.017)		(1.2)		(3.3)	

Here "garage" is a dummy variable equal to 1 if the house has an attached two-car garage, and equal to 0 otherwise.

a. Everything else equal, an increase in the size of the house by 500 square feet causes the price to increase by how much? $ thousand

b. Everything else equal, an attached two-car garage causes the price to increase by how much? $ thousand

Suppose we wish to predict the selling price of a house with size = 2000 square feet, three bathrooms, and an attached garage. So to simplify calculations, we first transform the data and estimate the same equation on the transformed data.

c. Which variables should be transformed? How?

Suppose the transformed data yield the following estimates:

price	=	203.9	+	0.050 *size*	+	5.1 *baths*	+	22.4	*garage*
		(4.0)		(0.017)		(1.2)		(3.3)	

The estimated variance of the error term is $\hat{\sigma}^2 = 9.0$.

d. Compute the predicted selling price of a house with size = 2000 square feet, three bathrooms, and an attached garage. $ thousand

e. Compute the standard error of the prediction error.

f. Compute a 95% prediction interval for the selling price.

(3) [Dummy variables and structural change: 22 pts] Suppose we wish to estimate the effect of high school mathematics courses on test scores, using a random sample of 300 students.

$score_i$ = test score of student i.
$courses_i$ = semesters of math courses taken by student i.
d_i = 1 if student i attends a private school, and
= 0 if student attends a public school.

The following four equations were estimated, with the sums of squared residuals (SSR) as shown.

[1] $score_i = 65.3 + 1.7\ courses_i$ SSR=2412
[2] $score_i = 60.1 + 1.5\ courses_i + 2.1\ d_i$ SSR=2376
[3] $score_i = 59.2 + 1.4\ courses_i + 0.2\ (d_i\ courses_i)$ SSR=2405
[4] $score_i = 58.7 + 1.3\ courses_i + 3.2\ d_i - 0.1\ (d_i\ courses_i)$ SSR=2220

a. Although there are two categories of students—public-school students and private-school students—only one dummy variable is used. If a second dummy variable were created for public-school students and both dummy variables were included in the same regression, then what econometric problem would arise?
b. According to equation [4], what is the intercept for the private-school students?
c. According to equation [4], what is the slope for the public-school students?
d. According to equation [4], what is the slope for the private-school students?

Assume that public-school students and private-school students have different intercepts. We wish to test the null hypothesis that all students also have the same slope, against the alternative hypothesis that they have different slopes (public school versus private school) at 5% significance.

e. Which equation, [1], [2], [3], or [4], is the *restricted* equation?

f. Which equation, [1], [2], [3], or [4], is the *unrestricted* equation?

g. [10 pts] Give the value of the test statistic, its degrees of freedom, the critical point, and your conclusion (whether you can reject the null hypothesis).

Degrees of freedom in numerator = _____ Degrees of freedom in denominator = _____

Value of F statistic = _____ Critical point = _____

Reject null hypothesis? _____.

(4) [Heteroskedasticity: 12 pts] We have estimated the following equation by ordinary least squares, using average data for **50** states:

$$\text{average test score} = \beta_1 + \beta_2 \text{ per-pupil spending} + \beta_3 \text{ per-capital income} + \varepsilon .$$

We believe that all the Gauss-Markov assumptions are satisfied, except that we fear that the error term (ε) might be heteroskedastic, with variance related to state population.

a. If the error term (ε) is heteroskedastic, are the least squares estimators $\hat{\beta}_1$, $\hat{\beta}_1$, and $\hat{\beta}_3$ still unbiased? (Answer *yes* or *no*.)

b. If the error term (ε) is heteroskedastic, are usual standard errors for the least squares estimators valid? (Answer *yes* or *no*.)

c. Given that the dependent variable is an average, is the variance of the error term (ε) more likely to be *positively* or *negatively* related to population?

To test for heteroskedasticity, we save the least-squares residuals from the above equation and estimate the following auxiliary regression by least squares:

$$\hat{\varepsilon}^2 = \alpha_1 + \alpha_2 \text{ population } + \nu ,$$

where υ is a new error term. The R^2 value from this auxiliary equation is 0.09.

d. Compute the value of the Breusch-Pagan test statistic.

e. Find the critical point in the appropriate table at 5% significance.

f. Can you reject the null hypothesis of no heteroskedasticity at 5% significance?

IV. CRITICAL THINKING: [8 pts] Suppose you want to estimate the price elasticity of demand for electricity. Using data from 50 countries, you plan to estimate the following equation:

$$\ln(elect)_i = \beta_1 + \beta_2 \ln(price)_i + \varepsilon_i ,$$

where *elect* = electricity consumption per capita and *price* = average price of electricity. To ensure that your estimator for β_2 is unbiased, what other regressors do you think should be included? **WHY?**

[end of exam]