Introduction to Econometrics (Econ 107)
Drake University, Spring 2006
William M. Boal

Signature: _____

Printed name: _____

MIDTERM EXAMINATION #3 VERSION B
"Multiple Regression with Cross-Sectional Data"
March 30, 2006

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 and F 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 [2 pts each—28 pts total]

(1) Suppose we want to estimate the effect of average income on average house size. Local average house price also has an effect on house size, but we omit price from the equation for lack of data. Suppose income has a positive effect on size, price has a negative effect, and income and price are positively correlated with each other (that is, areas with high incomes also have high house prices). Then omitting price will cause the least-squares estimator of the coefficient of income

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

(2) If we estimate the following by ordinary least squares:

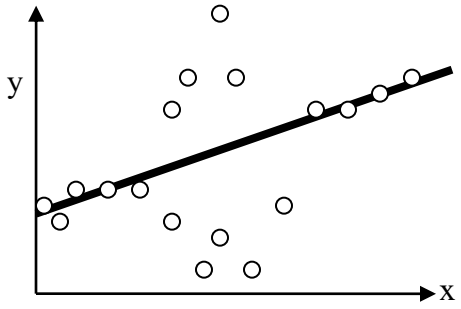
$$y_i = \beta_1 + \beta_2 x_{i2}, + \beta_3 x_{i3} + \varepsilon_i ,$$

which equation holds necessarily, regardless of the data?

a. $\sum \hat{y}_i \hat{\varepsilon}_i = 0$.
b. $\sum (\hat{y}_i - \bar{y})^2 = \sum (y_i - \bar{y})^2 + \sum \hat{\varepsilon}_i^2$.
c. $\sum x_{i2} y_i = 0$.
d. $\sum x_{i3} \hat{y}_i = 0$.
e. $\sum x_{i2} x_{i3} = 0$.

(3) 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. No autocorrelation: $Cov(\varepsilon_i,\varepsilon_j) = 0$ for $i \neq j$.
b. $E(\varepsilon_i|x_{i1},x_{i2},\ldots,x_{iK}) = 0$.
c. Homoskedasticity: $Var(\varepsilon_i) = \sigma^2$, a constant.
d. All of the above.
e. None of the above.



(4) The variance of the least-squares slope estimator $\hat{\beta}_j$ is larger, and thus the true value of β_j is estimated less precisely,

a. the larger the sample size.
b. the smaller the variance of the error term σ^2.
c. the larger 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.
e. All of the above.

(5) Adding another regressor to a regression equation will necessarily decrease

a. the t statistics of the regressors.
b. the R^2 value.
c. Theil's adjusted R^2 value.
d. the sum of squared residuals.
e. the estimated coefficients.

(6) If two regressors x_{i2} and x_{i3} are *perfectly* correlated, then the least-squares estimators of their coefficients

a. will be biased.
b. will be inconsistent.
c. will have large standard errors.
d. will be zero.
e. cannot be computed.

(7) The equation

$$y_i = 5.6 + 0.7\, x_{i2} + 1.2\, x_{i3}$$

implies that, holding x_{i2} constant, a one-unit increase in x_{i3} will cause y_i to increase by about

a. 5.6 units.
b. 0.7 units.
c. 1.2 units.
d. 6.8 units.
e. 1.9 units.

(8) The equation

$$\ln(wage) = 1.7 + 0.11\, educ$$

implies that if *educ* increases by one unit, then *wage* will increase by about

a. 0.11 percent.
b. 11.0 percent.
c. \$0.11.
d. \$1.70.
e. \$11.00.

(9) The equation

$$y_i = 3.0 + 1.2\, x_{i2} + 2.1\, x_{i3} + 0.2\, x_{i2}\, x_{i3}$$

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. 3.3 units.
c. $(1.2 + 0.2x_{i3})$ units.
d. $(1.2 + 2.1x_{i2})$ units.
e. $(1.2 + 4.2x_{i2})$ units.

(10) Suppose Q = quantity of output, L = labor input, and K = capital input. Which specification is a Cobb-Douglas production function?

a. $Q_i = 5.7 + 0.7 L_i + 0.3 K_i$.
b. $\ln(Q_i) = 4.5 + 0.4 L_i + 0.2 K_i$.
c. $\ln(Q_i) = 3.6 + 0.6 \ln(L_i) + 0.4 \ln(K_i)$.
d. $Q_i = 9.7 + 7.1 L_i + 2.8 K_i + 0.3 (L_i K_i)$.
e. $Q_i = 8.3 + 4.1 L_i - 0.2 L_i^2 + 3.2 K_i - 0.1 K_i^2$.

(11) Which of the following issues can cause the error term to be correlated with a regressor?

a. An independent (x) variable is measured with error.
b. The wrong variable is used as the dependent variable.
c. A regressor, which happens to be correlated with the included regressors, is omitted from the equation.
d. All of the above.

(12) Suppose we wish to estimate the effect of income on automobile purchases using data on states:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

If y_i is defined as *per-capita* automobile purchases and x_i is defined as *per-capita* income in the state, then we should suspect that the variance of the error term ε_i may 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.

(13) If the error term in a regression is heteroskedastic, the least-squares estimators for the coefficients will no longer be

a. computable.
b. unbiased.
c. consistent.
d. best linear unbiased estimators.
e. The least-squares estimators will no longer be any of the above.

(14) Under the null hypothesis of no heteroskedasticity, the Breusch-Pagan test statistic is close to

a. minus one.
b. zero.
c. one.
d. two.
e. four.

II. PROBLEMS: Please write your answers in the boxes on this question sheet.

(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")	5	260	52
Residual (or "Error")	120	240	2
Total	125	500	4

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").

e. Compute the value of Theil's adjusted R^2 (sometimes called "$\overline{R}^2$").

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 tests: 6 pts] Suppose we want to test whether a year at a junior college has the same effect on a person's wage as a year at a university. We have data on 5000 individual workers. Let *yearsjc* denote years of education at a junior college, let *yearsuniv* denote years of education at a university and let *yearstot = yearsjc + yearsuniv*. We have estimated the following equation (standard errors in parentheses).

$$\underset{}{ln(wage_i)} = \underset{(0.025)}{1.75} + \underset{(0.005)}{0.012}\ yearsjc + \underset{(0.009)}{0.072}\ yearstot$$

Test the null hypothesis that a year at a junior college has the same effect on a person's wage as a year at a university, a two-tailed test, at 5% significance. Give the value of the test statistic, the critical point, and your conclusion (whether you can reject the null hypothesis).

Value of test statistic = _____. Critical point(s) = _____.

Reject null hypothesis? _____.

(3) [Dummy variables and structural change: 22 pts] Suppose we wish to estimate the effect of tax rates on economic growth, using cross-sectional data for n=50 states. The following variables are to be used.

y_i = economic growth rate in state i.
x_i = tax rate in state i.
ds_i = 1 if state i is in the South, and 0 otherwise.
dm_i = 1 if state i is in the Midwest, and 0 otherwise.
dw_i = 1 if state i is in the West, and 0 otherwise.

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

[1] $y_i = 0.025 - 0.007\ x_i$ SSR=360

[2] $y_i = 0.021 + 0.002\ ds_i - 0.003\ dm_i + 0.001\ dw_i - 0.0068\ x_i$ SSR=270

[3] $y_i = 0.019 - 0.0068\ x_i - 0.0001\ (ds_i\ x_i) + 0.0003\ (dm_i\ x_i) - 0.0005\ (dw_i\ x_i)$ SSR=280

[4] $y_i = 0.019 - 0.0068\ x_i + 0.002\ ds_i - 0.003\ dm_i + 0.001\ dw_i + 0.0001\ (ds_i\ x_i) + 0.0002\ (dm_i\ x_i) - 0.0003\ (dw_i\ x_i)$ SSR=168

a. Although there are four official Census regions, only three dummy variables are used. If a fourth dummy variable were created for the remaining Northeast region and all four regional 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 Northeast?

c. According to equation [4], what is the intercept for the Midwest?

d. According to equation [4], what is the slope for the South?

We wish to test the null hypothesis that all states also have the same intercept and slope, against the alternative hypothesis that they have different intercepts and slopes by region, 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: 18 pts] The regression equation $y_i = \beta_1 + \beta_2 x_{i2} + \varepsilon_i$ was estimated using 30 cross-sectional observations on countries, by ordinary least squares. To check for heteroskedasticity related to population, separate regressions were run for the 12 countries with the lowest populations and the 12 countries with the highest populations. The sum of squared residuals for the small countries was **25.0**. The sum of squared residuals for the large countries was **125.0**.

a. [4 pts] Compute unbiased estimates of the variance of the error term in the two subsamples.

Estimated error variance, small countries = _____

Estimated error variance, large countries = _____

b. [4 pts] Given these results, which subsample appears to lie closer to the true regression line: the small countries or the large countries? Explain your answer.

c. [6 pts] Test the null hypothesis of homoskedasticity, against the (one-sided) alternative hypothesis that large countries have higher error variance, at 5% significance using a Goldfeld-Quandt test. Give the value of the test statistic, the critical point, and your conclusion (whether you reject the null hypothesis of homoskedasticity).

Value of test statistic = _____. Critical point = _____.

Reject null hypothesis? _____.

d. [6 pts] Regardless of your conclusion for part (c), suppose you believe that heteroskedasticity is indeed present and that the variance of the error term is proportional to country population:

$$\text{Var}(\varepsilon_i) = \alpha \cdot \text{pop}_i,$$

where α = an unknown constant and pop_i = population of country i. Show how you would transform the equation to restore homoskedasticity by computing the transformed data for the first two observations in the table below.

Obs.	Raw data			Transformed data		
	y	x	pop	y	intercept	x
#1	77	42	49			
#2	54	24	36			

III. CRITICAL THINKING [4 pts] Suppose we wish to estimate the demand for electricity using a sample of households. We have data on household electricity consumption, household income, and the price of electricity faced by that household. All of these variables vary substantially over the sample. Now which variable should be on the left-hand side of the regression equation—that is, which variable should be the "y" variable? Why?

[end of exam]