BOSTON COLLEGE

Department of Economics

EC 228 02
Econometric Methods
Prof. Baum
Fall, 2009

MIDTERM EXAMINATION
3 November 2009

Answer all questions. Total 87 points. Exam ends at 10:15 AM sharp. Partial credit given for partial answers.

1. (40 pts) Indicate clearly whether each of the following statements are TRUE or FALSE, and EXPLAIN your answer. No credit without explanation!

a) Simple regressions of y on x and y on z will yield different slope coefficients than the multiple regression of y on (x, z) unless x and z are correlated. *F: different slopes will result unless x and z are perfectly UNcorrelated.*

b) The best linear unbiased estimator (BLUE) of the population mean of x is the sample mean of x. *T: see the sample exam!*

c) OLS estimators are 'BLUE' as long as the errors in the equation follow a Normal distribution. *F: OLS→BLUE (Gauss–Markov proof) does not depend on Normality.*

d) Two-way ANOVA models allow us to consider the effects of two qualitative factors (and nothing else) on the dependent variable. *T: by definition, one-way ANOVA contains nothing but dummies for one qualitative factor; two-way ANOVA contains nothing but dummies for two qualitative factors (e.g., race and gender). This is a different use of "ANOVA" than the "ANOVA F" printed on every regression.*

e) Adding a variable to a regression model cannot decrease the sum of squared residuals, no matter what that variable contains. *F: it cannot INcrease the sum of squared residuals (nor decrease the R^2).*

f) A test of gender-based statistical discrimination in salary levels involves examining the residuals from a regression of salary on education, job class, job tenure and gender. *F: gender should not be included in the model, as it is not a valid determinant of salary. The residuals should be categorized by gender.*

g) Unbiasedness of an estimator is not sufficient to make it a useful tool in statistical inference. *T:* $\tilde{\mu} = (x_1 + x_2)/2$ *is an unbiased estimator of population mean, but horribly inefficient if* $N > 2$. *Efficiency must also be considered.*

h) In a one-tailed test, we must double the p-value reported for a two-tailed test. *F: you must HALVE the reported p-value.*

i) The single-log transformation is often used to express a model in constant-growth form. *T: constant growth is a constant percent change per unit time, which is approximated by* $\partial \log y / \partial t$. *The double-log transformation is a constant-elasticity formulation, not constant-growth.*

j) The ANOVA F test reported for every regression evaluates the null hypothesis that all the estimated coefficients are jointly zero. *F: The ANOVA F tests that all slope coefficients are jointly zero, contrasting the estimated model with the naïve model* $y_i = \gamma + \epsilon_i$. *It does not restrict the constant term.*

2. (20 pts) In a renowned study of the automobile market, Griliches reported the following results from a large sample of automobile purchases:

$\log(Price) = 6.4 + 0.056$ H $+ 0.249$ W $+ 0.023$ L $+ 0.010$ V $+ 0.023$ T $+ 0.090$ A $+$ 0.088 P $+ 0.109$ B $+ 0.157$ C $- 0.044$ D55 $- 0.015$ D56 $+ 0.019$ D57 $+ 0.440$ D58 $+ 0.044$ D59 $+ 0.023$ D60 $+ \epsilon$

where:
H = hundreds of horsepower
W = weight, thousand lb
L = length, tens of inches
V = 1 if engine is V8
T = 1 if hard top, 0 if convertible
A = 1 if automatic transmission
P = 1 if power steering standard
B = 1 if power brakes standard
C = 1 if a compact car
D55 ... D60 are dummies for model years 1955, ... , 1960; 1954 is the base year

a) How might we interpret the coefficients in this regression? For instance, what does the model predict would be the effect on price of an additional 100 horsepower? *The coefficients are semi-elasticities, or percentage changes in price per unit of the regressor. An auto trans adds about 9.0% to the price of the car. 100 more Hp adds 5.6% to the price. A common error was to multiply this by 100!*

b) Cet. par., what does the model predict happened to the price of cars between 1956 and 1957? Between 1956 and 1959? Between 1954 and 1957? *Values expected! 1956–1957: 3.4%. 1956–1959: 5.9% (intermediate years not relevant). 1954–1957: 1.9%.*

c) Cet. par., how much more would it cost to buy a car with a V-8 and power steering in 1955 than one without these two features in 1954? *-0.044 + 0.010 + 0.088 = 0.054, or about 5.4% more.*

d) The R^2 in this regression, using 570 observations, was 0.922. Suppose that if the model were run without the set of D dummies the R^2 was 0.919. Would this support the hypothesis that variation in the mix of features and characteristics (weight, horsepower, etc.) was entirely responsible for the variation in car prices over the seven-year period? *The adjusted R^2 ($\bar{R}^2$) is not relevant here. The R^2 form of the subset F statistic (Wooldridge p. 150) can be used:*

$$\frac{(R_U^2 - R_R^2)/6}{(1 - R_U^2)/(570 - 16)}$$

where U and R refer to the unrestricted and restricted model, respectively. The d.f. refer to the number of constraints (dummy coefficients) and $N - k - 1$ in the unrestricted regression. Plugging in values yields $F_{554}^6 = 3.55$. Although you did not have a F-table handy, you could comment that this value is fairly large (the 95% critical value is 2.10), and might lead to a rejection of the null hypothesis that the year dummies are jointly insignificant.

3. (15 pts) In the NLSW88 dataset, 2,246 individuals are classified by `race` (white/black/other, coded 1,2,3), `industry` (5 classifications, coded 1,2,...,5) and whether they are a `union` member (0/1).

a) Specify how you would test the hypothesis that these workers' `wage` is influenced by their `race` and `union` status. Indicate the transformations you would apply to the data, the regression equation you would use, and the tests you would apply after that equation. *Create dummies for two races (say, black and other) and regress* `wage` *on* `black other union`. *Test for relevance of the qualitative factor* `race` *with an F-test of the* `black` *and* `other` *coefficients jointly zero.*

b) The model of part (a) assumes that `race` and `union` status have independent effects on `wage`. Indicate how you would test for non-independent effects of those factors on `wage`. *Create interactions* `black`×`union`, `other`×`union` *and include them in the equation. (Note that you cannot also include an interaction with* `white`*). Do a joint (F) test for the two interaction coefficients being zero.*

c) Returning to the model of part (a): how would you include the effects of `industry` on `wage`? Write down the equation you would use, and indicate the test that you would apply to consider the importance of `industry` on `wage`. *Create dummies for four of the five industries and include them in the equation. Do a joint F test for each of those four coefficients being zero; this tests the significance of the qualitative factor* `industry`.

4. (12 pts) Four variables from the Anscombe dataset, with their descriptive statistics:

```
. l x1 y1 x4 y4, sep(0) noobs
```

x1	y1	x4	y4
10	8.04	8	6.58
8	6.95	8	5.76
13	7.58	8	7.71
9	8.81	8	8.84
11	8.33	8	8.47
14	9.96	8	7.04
6	7.24	8	5.25
4	4.26	19	12.5
12	10.84	8	5.56
7	4.82	8	7.91
5	5.68	8	6.89

```
. su x1 y1 x4 y4
    Variable |       Obs        Mean    Std. Dev.       Min        Max
```

x1	11	9	3.316625	4	14
y1	11	7.500909	2.031568	4.26	10.84
x4	11	9	3.316625	8	19
y4	11	7.500909	2.030579	5.25	12.5

a) What do you note about the pair of x variables? About the pair of y variables? *They have identical means and variances, but different ranges.*

Consider two simple regressions:

	(1)	(2)
	y1	y4
x1	0.500**	
	(4.24)	
x4		0.500**
		(4.24)
_cons	3.000*	3.002*
	(2.67)	(2.67)
N	11	11
r2	0.667	0.667
rmse	1.24	1.24

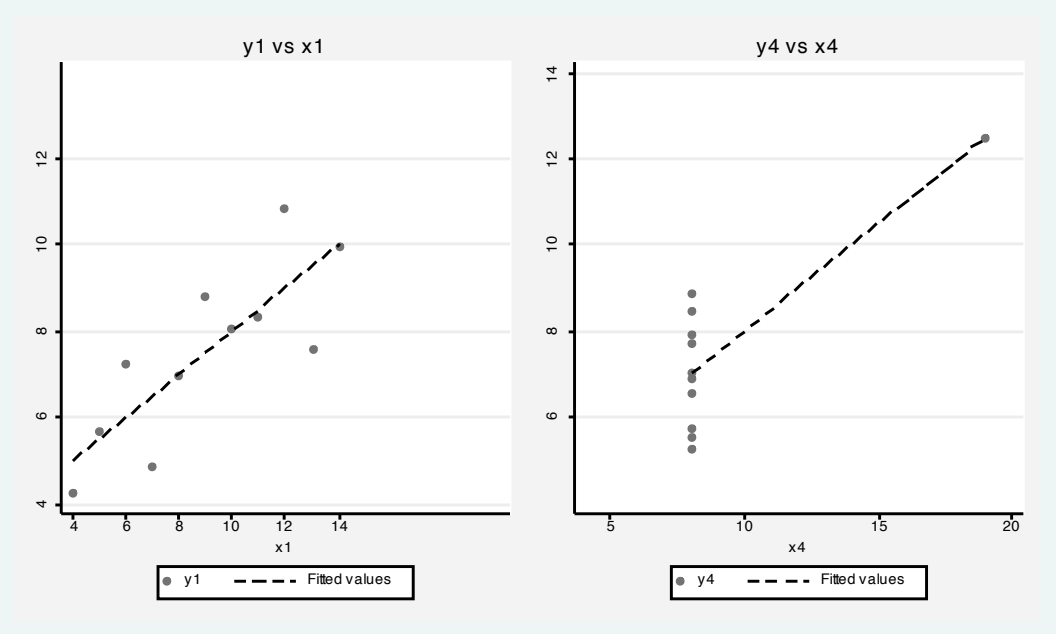
t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

b) What must be true about the correlations of $(x1, y1)$ and $(x4, y4)$? *They must be identical, as the two regressions have the same* $\sqrt{R^2}$*. Note, however, that the correlations do not equal 0.500!*

c) Does one of these regressions do a better job than the other in explaining the relationship? Why or why not? *From the regression table, they do equally well; they have the same* R^2 *values and same RMS errors.*

Consider the scatterplots and estimated regression lines:



d) Does your answer to part c) change? Why or why not? *In data set 4, 10 of the 11 people have the same X value. Drawing a regression line in that context makes little sense, as its slope is wholly dependent on the outlying observation. Thus, the regression from data set 1 is a more sensible representation of the relationship between Y and X, as both have considerable variation in their sample values. (Note, however, that our point estimates for data set 4 are unbiased and consistent.)*