EEP 118 / IAS 118 Elisabeth Sadoulet and Clair Null
University of California at Berkeley Fall 2007

Introductory Applied Econometrics
Final examination

Scores add up to 130 points

Your name:_____ SID:_____

SID _____

1. (15 points) In order to estimate the following model of emission by power plants:

$$\log(CO2) = \beta_0 + \beta_1 \log(Q) + \beta_2 OLD + u$$

you collect information from 25 firms on carbon dioxide emission (CO2), their level of production (Q), and a dummy variable that is equal to 1 for firms whose generator is older than 15 years (OLD) and zero otherwise. The estimated equation is the following:

$$\widehat{\log}(CO2) = 2.54 + 0.924 \log(Q) + 0.210 OLD \qquad R^2 = 0.253$$
$$(0.52)\ (0.042) \qquad (0.097)$$

(standard errors in parentheses)

a. What is the economic interpretation of the true parameters β_1 and β_2 ?

b. Construct a 95% confidence interval for β_1 . Give an interpretation.

c. Test the hypothesis $\beta_1 = 1$ against $\beta_1 \neq 1$ at the 95% significance level. Follow the 4 steps of hypothesis testing that we have used in class. What is the ***economic*** interpretation of your result?

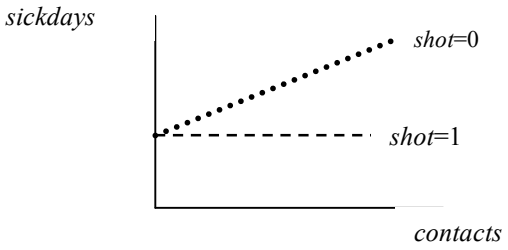
2. (15 points) You want to estimate a simple equation that relates log(wage) to education. Suppose that education is given in 4 categories: high school or less, some college, complete college, post-graduate studies.

a. Write a model that will allow you to estimate the return to education using these data. Be sure to explain what values your variables take on.

b. We have seen two methods that you could use with these data to test the hypothesis that the return to education is the same for men and women. Describe one of these methods in detail (you can choose which one). If you plan to modify the model in some way, write down the new equation that you will estimate. Follow the 4 steps of hypothesis testing that we have used in class and be sure to explain where all the elements of your test statistic come from.

SID _____

3. (10 points) Suppose you have data for a sample of employees of a firm on the number of sick leave days taken in one year (*sickdays*), a rough estimate of how many people the employee comes in contact during his/her workday (*contacts*), and whether or not the they got a flu shot (*shot*=1 for those who got a flu shot and 0 for those who didn't). The data give you the following plot:



a. Reading from the graph, explain in words the effect of flu shot on sick leave days.

b. You estimate the following model:

$$\hat{sickdays} = \hat{\beta}_0 + \hat{\beta}_1 contacts + \hat{\beta}_2 shot + \hat{\beta}_3 (contacts \times shot)$$

What do you expect to find for the sign and significance of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$, and $\hat{\beta}_3$ given the graph of your data above?

SID _____

4. (10 points) Consider the following two models, in which:

```
       bwght  is the child birth weight (in ounces)
       cigs   is the average number of cigarettes smoked by the mother each day
       parity is the rank of the child in the family (in terms of birth order)
       faminc is the household income
and    mothedu and fathedu are the mother's and father's education level, in years.
```

```
. reg bwght cigs parity faminc motheduc fatheduc

      Source |       SS       df       MS              Number of obs =    1191
-------------+------------------------------           F(  5,  1185) =    9.55
       Model |  18705.5567     5  3741.11135           Prob > F      =  0.0000
    Residual |  464041.135  1185  391.595895           R-squared     =  0.0387
-------------+------------------------------           Adj R-squared =  0.0347
       Total |  482746.692  1190  405.669489           Root MSE      =  19.789

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.5959362   .1103479    -5.40   0.000    -.8124352   -.3794373
      parity |   1.787603   .6594055     2.71   0.007     .4938709    3.081336
      faminc |   .0560414   .0365616     1.53   0.126    -.0156913    .1277742
    motheduc |   .3704503   .3198551     1.16   0.247    -.2570951    .9979957
    fatheduc |   .4723944   .2826433     1.67   0.095    -.0821426    1.026931
       _cons |   114.5243   3.728453    30.72   0.000     107.2092    121.8394
------------------------------------------------------------------------------

. reg bwght cigs parity faminc

      Source |       SS       df       MS              Number of obs =    1191
-------------+------------------------------           F(  3,  1187) =   14.95
       Model |  17579.8997     3  5859.96658           Prob > F      =  0.0000
    Residual |  465166.792  1187  391.884408           R-squared     =  0.0364
-------------+------------------------------           Adj R-squared =  0.0340
       Total |  482746.692  1190  405.669489           Root MSE      =  19.796

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.5978519   .1087701    -5.50   0.000    -.8112549   -.3844489
      parity |   1.832274   .6575402     2.79   0.005     .5422035    3.122345
      faminc |   .0670618   .0323938     2.07   0.039     .0035063    .1306173
       _cons |   115.4699   1.655898    69.73   0.000     112.2211    118.7187
------------------------------------------------------------------------------
```

a. Test for the joint significance of the parent education variables in the first model. Use the 4 steps of hypothesis testing that we have used in class.

b. Explain why the parameter on household income might have changed from the first to the second model.

5. (5 points) Under what conditions are the OLS estimators on time series data unbiased? Be very precise in your notations and write a sentence interpreting each condition.

6. (15 points) For a sample of 3791 Mexican children of 11-13 years old, the two following probit models of school enrollment in 1998 have been estimated:

```
enroll         =1 if child is enrolled in school, 0 otherwise
male           =1 if child is male, 0 if female
poor           =1 if household is classified as poor, 0 otherwise
h_edu          household head education (in years)
ownland        total land owned (in hectares)
cattle         number of heads of cattle owned

. dprobit enroll male poor h_edu ownland cattle

Probit regression, reporting marginal effects           Number of obs =   3791
                                                        LR chi2(5)    =  40.89
                                                        Prob > chi2   = 0.0000
Log likelihood = -1498.8221                             Pseudo R2     = 0.0135

------------------------------------------------------------------------------
enroll   |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
    male*|   .0329786   .0111125     2.97   0.003   .509628   .011199  .054759
    poor*|  -.0288193   .0117069    -2.39   0.017   .683725  -.051764 -.005874
   h_edu |   .0113324   .0023733     4.74   0.000   2.72382   .006681  .015984
 ownland |  -.0002177   .0015425    -0.14   0.888   1.64465  -.003241  .002805
  cattle |   .0015823     .00247     0.64   0.522   .719863  -.003259  .006423
---------+--------------------------------------------------------------------
  obs. P |   .8623055
 pred. P |   .8657898  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| correspond to the test of the underlying coefficient being 0

. dprobit enroll male poor h_edu

Probit regression, reporting marginal effects           Number of obs =   3791
                                                        LR chi2(3)    =  40.45
                                                        Prob > chi2   = 0.0000
Log likelihood = -1499.0412                             Pseudo R2     = 0.0133

------------------------------------------------------------------------------
enroll   |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
    male*|   .0329302   .0111136     2.96   0.003   .509628   .011148  .054712
    poor*|  -.0295152   .0115888    -2.47   0.014   .683725  -.052229 -.006802
   h_edu |   .0113046   .0023734     4.73   0.000   2.72382   .006653  .015956
---------+--------------------------------------------------------------------
  obs. P |   .8623055
 pred. P |   .8657429  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

a. Using the first model, use the SSS method to interpret the coefficient on `h_edu`.

b. Using the first model, use the SSS method to interpret the coefficient on `male`.

c. Test for the joint significance of the two variables `ownland` and `cattle` that capture the fact that the household is an agricultural household. Use the 4 steps of hypothesis testing that we have used in class. What do you conclude?

7. (25 points) Using panel data for 5 countries (Morocco, Algeria, Tunisia, Libya, and Egypt) for 5 years 1996-2000, you want to estimate the role of foreign aid on GDP growth (GROWTH). The foreign aid variable AID measures aid in percent of GDP and the *it* subscript refers to country *i* in year *t*.

a. Does the following regression allow you to identify a causal effect of foreign aid on growth? Why or why not?

$$\hat{GROWTH}_{it} = 0.03 - 0.028 AID_{it} \qquad R^2 = 0.09$$
$$(0.01)\ (0.012)$$

b. Now consider this following model: $GROWTH_{it} = a_i + \beta AID_{it} + u_{it}$

Interpret the *a* parameters. What do they represent? What is controlled for by including them in the model?

c. Let *Y*96, *Y*97, … , *Y*00 be dummy variables for the years 1996 to 2000. Consider now the following regression:

$$GROWTH_{it} = \beta_0 + b_{97}Y97 + b_{98}Y98 + b_{99}Y99 + b_{00}Y00 + \beta AID_{it} + u_{it}$$

Interpret the *b* parameters: what do they represent? What is controlled for by including them in the model?

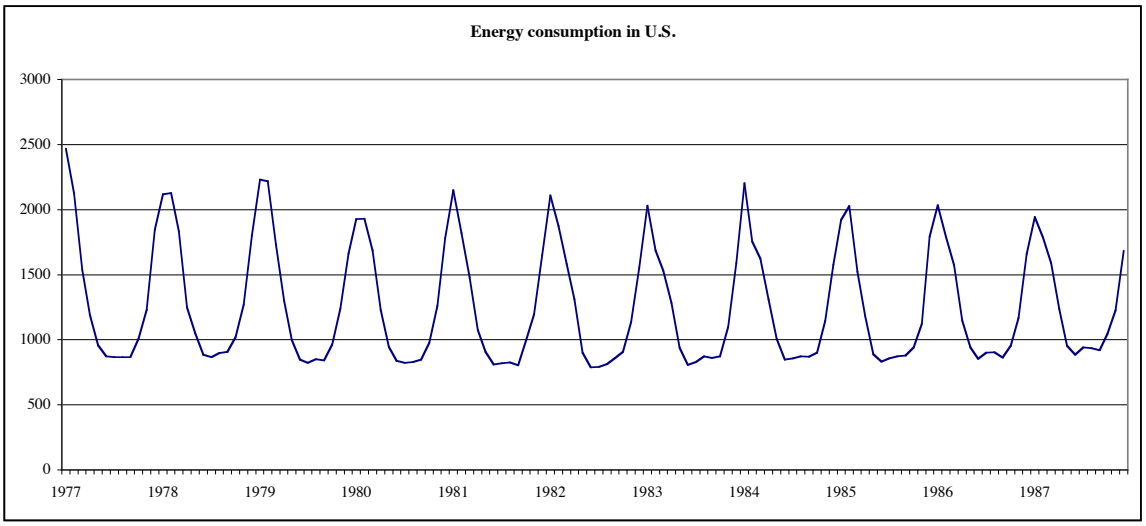
d. Following is the result of the panel regression estimation for the 5 countries, where *Mo*, *Al*, *Tu*, and *Li* are dummy variables for Morocco, Algeria, Tunisia, and Libya. (Numbers in parentheses are standard errors)

$$\begin{array}{cccccccccc} \hat{GROWTH} = & 0.002 & + 0.015Mo & + 0.011Al & + 0.021Tu & - 0.032Li & + 0.012Y97 & - 0.041Y98 & + 0.012Y99 & \\ & (0.001) & (0.010) & (0.001) & (0.010) & (0.001) & (0.001) & (0.010) & (0.001) & \\ & & & & + 0.023Y00 & + 0.026AID & & & & \\ & & & & (0.010) & (0.012) & & & & \end{array}$$

Interpret the sign and size of the parameters on *Mo*, *Li*, and *Y*98 (you do not need to interpret the significance of these parameters for this problem).

e. Interpret the sign and size of the parameter on *AID* in the model of part d (you do not need to interpret the significance of the parameter for this problem).

8. (5 points) The following graph represents monthly consumption of energy by the residential and commercial sectors in the U.S. from 1977 to 1987



Source: http://www.economagic.com

Write the model that would allow you to estimate the average growth rate and the seasonal pattern of energy consumption during this period. Be sure to explain what values your variables take on.

9. (15 points) Suppose that you have estimated this typical supply response model of cereal production:

$$\widehat{\log}(Q_t) = \hat{\beta}_0 + \underset{(0.02)}{0.08}\log(p_t) + \underset{(0.04)}{0.15}\log(p_{t-1}) + \underset{(0.05)}{0.10}\log(p_{t-2})$$

where Q_t is the quantity produced in million tons and p_t the average cereal price in year *t*.

a. What is the short-term effect of a permanent cereal price increase by 20%?

b. What will be the long-term effect of a permanent price increase by 20%?

c. Explain why it is important to include the lagged variables in this model.

SID _____

10. (15 points) To evaluate the effect of a new teaching method, the program is introduced as a pilot in a few districts in 1996. Test scores are collected in these districts and in some comparison districts that are not in the pilot the year before (1995) and two years later (1997). Results are as follows:

Percentage of children that successfully pass the standardized test

	Districts not in the pilot	Districts in the pilot
1995	0.45	0.48
1997	0.46	0.52

a. Can the comparison of the test results in 1997 give a measure of the impact of the teaching method? Why or why not?

b. Compute the double-difference estimate of the impact of the teaching method. Explain the logic behind this estimator in words.

c. Write an equation for a regression that you could estimate that will allow you to get a standard error on this estimate. Explain what values your variables take on.

SID _____

Formulae

Statistics and miscellaneous

Covariance between two variables in a population: $\operatorname{cov}(x,y) = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})$

$\operatorname{cov}(a_1 x + b_1, a_2 y + b_2) = a_1 a_2 \operatorname{cov}(x,y)$

$\operatorname{var}(x+y) = \operatorname{var} x + \operatorname{var} y + 2\operatorname{cov}(x,y)$

When y is a binary variable with probability prob(y = 1) = $p(x)$, the variance conditional on x is $p(x)(1-p(x))$

For small values of x: $e^{ax} \approx 1 + ax$

OLS estimator

$\hat{\beta}_1 = \frac{\operatorname{cov}(x,y)}{\operatorname{var} x}$ with $\operatorname{var}(\hat{\beta}_1) = \frac{\sigma^2}{SST_x}$

For multiple regression: $\operatorname{var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j^2)}$

Adjusted R square: $\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)} = 1 - \frac{\hat{\sigma}^2}{SST/(n-1)}$

Test statistics:

Loglikelihood ratio statistic for q restrictions: $LR = 2(Loglikelihood_{UR} - Loglikelihood_R) \sim \chi_q^2$

F statistic for q restrictions in a regression done with n observations and k exogenous variables:

$$\frac{(R_{UR}^2 - R_R^2)/q}{(1-R_{UR}^2)/(n-k-1)} \sim F(n-k-1, q)$$

Chow statistic: $F = \frac{[SSR_P - (SSR_1 + SSR_2)]/k+1}{SSR_1 + SSR_2/[n-2(k+1)]} \sim F(k+1, n-2(k+1))$