EEP 118 / IAS 118 | Elisabeth Sadoulet and Kelly Jones
University of California at Berkeley | Fall 2008

Introductory Applied Econometrics
Final examination

Scores add up to 125 points

Your name:____ SID:__________________________

SID ______________________________

1. (25 points) A city has been publicizing its new compost collection service with flyers and billboards in order to increase use of the service. The partial results below show a regression of 100 pounds of compost collected per neighborhood (*compost*) on expenditures for *flyers* and *billboards* in $100's.

```
      Source |  SS         df           MS           Number of obs =    123
-------------+------------------------------          F(  3,  122)  =
       Model | 186.        2            83.           Prob > F      =
    Residual | 558.        120          4.65          R-squared     =
-------------+------------------------------          Adj R-squared =
       Total | 744.        122          6.09          Root MSE      =

------------------------------------------------------------------------------
     compost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      flyers |      1.60         0.40
  billboards |      0.23         0.15
       _cons |     10.00          2.0
------------------------------------------------------------------------------
```

a) Formally test whether the effect of billboards on pounds of compost collected is different from zero.

b) Fully interpret the coefficient on flyers.

c) Calculate the 95% confidence interval for the effect of flyer spending.

d) Calculate and interpret the R-squared for this regression.

e) If the city had not spent anything on publicity, how much compost is expected to be collected per neighborhood?

2. (15 points) From a sample of 200 households, we estimated the following two models of gasoline consumption (t-statistics in parentheses):

$$gas = \underset{(2.3)}{34.2} + \underset{(3.1)}{10.5} suv + \underset{(1.7)}{0.25} inc - \underset{(1.8)}{0.00005} inc^2 \qquad R^2 = 0.356$$

$$gas = \underset{(2.3)}{22.2} + \underset{(3.1)}{15.3} suv \qquad R^2 = 0.323$$

where *gas* gives the number of gallons per month, *suv* is a dummy variable for whether the household owns an SUV, and *inc* is the annual household income in thousands of $.

a) Using the estimated parameters in the first equation, how does gasoline consumption vary with income?

b) Are the two income variables jointly significant at the 5% level?

c) Comparing the *suv* parameter in the two equations, what can you infer about the correlation between income and SUV ownership?

3.(20 points) Consider the basic wage model where experience and education are expressed in years:

$$wage = \beta_0 + \beta_1 exper + \beta_2\ educ + u \qquad \text{model (1)}$$

a) What equation would you estimate to check whether the effect of experience depends on the level of education? What test would you perform?

b) Suppose now that the effect of experience does not depend on education, but education is specified in 3 levels only, “no diploma”, “primary diploma”, “secondary diploma and above”. How would you re-specify model (1)? How would you test that education has no influence on wages?

c) Considering the original model (1), how would you proceed to test whether the wage equations for men and women are the same?

d) Now suppose you estimated the following two equations

$$wage = \hat{\beta}_0 + \hat{\beta}_1 exper + \hat{\beta}_2\ educ \qquad R^2 = .32 \qquad (1)$$

$$\log(wage) = \hat{\beta}_0^* + \hat{\beta}_1^* exper + \hat{\beta}_2^*\ educ \qquad R^2 = .30 \qquad (2)$$

Explain how you would decide which does a better job of predicting wages.

4. (10 points) You have a data set consisting of observations on individuals' health indicators and outcomes for persons over 50 years old:

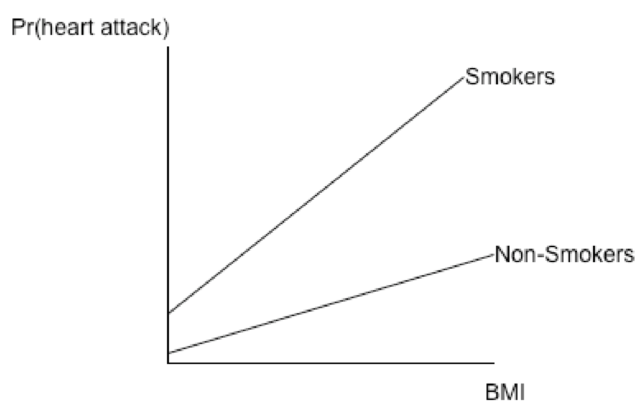
h_attack	=1 if the person has had a heart attack, 0 otherwise.
smoker	=1 if the person smokes, 0 otherwise.
bmi	individual's body mass index (higher implies more overweight)

You estimate the following linear probability model to predict the probability of a heart attack

$$\mathrm{P}\left(h\_attack = 1\right) = \beta_0 + \beta_1 bmi + \beta_2 smoker + \beta_3 bmi * smoker$$

a) Give the expression for the marginal effect of *bmi* on probability of having a heart attack for smokers. Give the expression for the marginal effect of *bmi* on probability of having a heart attack for non-smokers.

b) Based on the graph below, what sign and significance do you expect for β_0 ? Which parameter in the equation identifies the difference in the slopes of the "smokers" line and the "non-smokers" line below?



5. (10 points) To evaluate the effect of the presence of refugee camps (RC) on the price of staple foods in Kenya, data were collected on prices in markets that are close to refugee camps and markets that are far from refugee camps, at two points in time, one before the installation of the refugee camps, and one after their installation. Results are as follows:

	Average price of maize (in Shilling per lb)	
	Markets far from RC	Markets close to RC
Before the installation of RC	15.2	12.1
After the installation of RC	18.4	17.5

a) In the markets close to the refugee camps, by how much did the price of maize increase between the two periods? Can this be attributed to the presence of the refugee camps? Why or why not?

b) Compute the double-difference estimate of the impact of refugee camps on the maize price in markets close to them. Is this a better estimate of the effect of refugee camps on food prices than the one you calculated in part (a)? Why or why not?

6. (10 points) Let $gGDP_t$ denote the annual percentage change in gross domestic product and let int_t denote the short-term interest rate (in %). Suppose that we estimated the following relationship between $gGDP_t$ and interest rates:

$$gGDP_t = 2.1 - 0.4\ int_t + 0.1\ int_{t-1}$$

a) Suppose that in 2006, there was a one-time increase in the interest rate by 1 percentage point. What is the effect of this increase on GDP growth in 2006? What is the effect on GDP growth in 2007? In 2008?

b) Suppose the 2006 increase in the interest rate were permanent. What would be the effect of this increase on GDP growth in 2006? In 2007? In 2008?

7. (10 points) A sample of 17,394 children 10-15 years old in rural Mexico includes the following variables:

enroll	=1 if the child was enrolled in school
male	=1 if the child is a boy
age	= age of the child
poor	=1 if the child lives in a poor household

```
. dprobit enroll male age poor

Probit regression, reporting marginal effects           Number of obs =  17394
                                                        LR chi2(3)    =3464.00
                                                        Prob > chi2   = 0.0000
Log likelihood = -6882.8172                             Pseudo R2     = 0.2010

---------------------------------------------------------------------------------
enroll97 |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+-----------------------------------------------------------------------
   male*|   .0538895    .005481     9.87   0.000    .51742   .043147  .064632
    age |  -.0942691   .0016648   -52.19   0.000    12.381  -.097532 -.091006
   poor*|  -.0415562   .0053894    -7.46   0.000   .648787  -.052119 -.030993
---------+-----------------------------------------------------------------------
  obs. P |   .8036679
 pred. P |   .8595585  (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 reported results, interpret the estimated role of the variable *male.*

b) Using the reported results, interpret the estimated role of the variable *age.*

8. (20 points) Using state level data on murder rates (*mrdrte*) and unemployment rates (*unem*) in 1987, 1990, and 1993, we want to estimate the effect of unemployment on murder rate. Two estimations are reported below in which *d90* and *d93* represent dummy variables for the years 1990 and 1993, and state is a variable that takes the values 1,2, …, 51 for the states.

a) Write the equation of the model corresponding to the first estimation [be very careful with indices].

b) Does the coefficient on the unemployment variable correctly identify the effect of unemployment on crime? If yes, justify. If not, give a concrete example that illustrates the reason for a biased estimate and the direction of the bias.

c) Write the equation corresponding to the second estimation [be careful with indices]. Describe precisely each new variable you introduce and explain what it represents.

d) Interpret the coefficient on the year 1990 dummy variable in the second estimation.

e) From the second estimation, what do you conclude about the effect of unemployment on crime? Are there still potential sources of bias in this estimation? Justify your response.

```
. reg mrdrte unem d90 d93

      Source |       SS       df       MS              Number of obs =     153
-------------+------------------------------           F(  3,   149) =    3.84
       Model |  920.910876     3  306.970292           Prob > F      =  0.0111
    Residual |  11924.4272   149   80.029713           R-squared     =  0.0717
-------------+------------------------------           Adj R-squared =  0.0530
       Total |  12845.3381   152  84.5088034           Root MSE      =  8.9459

------------------------------------------------------------------------------
      mrdrte |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        unem |   1.443404   .4460414     3.24   0.001     .5620206    2.324788
         d90 |   2.677292   1.815129     1.47   0.142    -.9094277    6.264011
         d93 |   1.667332   1.771566     0.94   0.348    -1.833306    5.167971
       _cons |  -1.999366   3.062258    -0.65   0.515    -8.050428    4.051696
------------------------------------------------------------------------------

. xtreg mrdrte unem d90 d93, i(state) fe

Fixed-effects (within) regression               Number of obs      =       153
Group variable (i): state                       Number of groups   =        51

R-sq:  within  = 0.0676                         Obs per group: min =         3
       between = 0.1015                                        avg =       3.0
       overall = 0.0314                                        max =         3

                                                F(3,99)            =      2.39
corr(u_i, Xb)  = 0.0951                         Prob > F           =    0.0731

------------------------------------------------------------------------------
      mrdrte |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        unem |   .2019432   .2947557     0.69   0.495    -.3829162    .7868025
         d90 |   1.577016   .7433858     2.12   0.036     .1019775    3.052055
         d93 |   1.681938   .6959821     2.42   0.017     .3009584    3.062917
       _cons |   5.778023   1.911012     3.02   0.003     1.986161    9.569885
-------------+----------------------------------------------------------------
     sigma_u |  8.6877605
     sigma_e |  3.5144936
         rho |  .85936665   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(50, 99) =    17.33             Prob > F = 0.0000
```

SID ______________________________

Formulae

Statistics and miscellaneous

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

$\operatorname{cov}(a_1x+b_1, a_2y+b_2)=a_1a_2\operatorname{cov}(x,y)$

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

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: $LR=2\left(Loglikelihood_{UR}-Loglikelihood_R\right)\sim\chi_q^2$

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

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

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