Market Analysis
Prof. Fang-Chang Kuo

Problem Set 1

Due Date:
Tue. April 7

In this problem set we will apply some of the techniques introduced in class. Questions will require the use of Python and some other statistical/matrix package.

Q 1 Linear Model

Consider the following linear model with two parameters:

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

The file **data1.dta** contains data generated by this model. The first column is y_i, and the second column is x_i. β_1 and β_2 are the parameters of interest.

a) Visualize the relationship between y and x. Comment on your figure.

b) Estimate this model using OLS estimation formula. Specifically,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

(Hint: You need to figure out numpy methods for transpose, inverse, and matrix multiplication.)

c) Estimate this model by minimizing the sum of squared errors. Specifically,

$$S(\mathbf{b}) = \sum_{i=1}^{n}(y_i - b_1 - b_2 x_i)^2$$

$$\hat{\boldsymbol{\beta}} \equiv \arg\min_{\mathbf{b}} S(\mathbf{b})$$

This will require use of an optimization procedure in Python.

d) (Bonus Question) Under the assumption that the distribution of ϵ_i is i.i.d. (across observations) $N(0, \sigma^2)$, estimate this model using Maximum Likelihood. Note that there are now three parameters to estimate, β_1, β_2, and σ. Also note that the parameter σ is restricted to be greater than zero.

Q 2 Panel Data

The data set **chile.dta** is a Stata file containing a panel of Chilean firms from the annual manufacturing census Encuesta Nacional Industrial Anual (ENIA) run by Chile's national statistical institute, Instituto Nacional de Estadisticas (INE). The panel covers the period 1979–1996 and only includes data for firms in the food industry (CIIU code 311). Each observation is a year-firm combination. The data have already been cleaned up and the relevant variables constructed.

The data includes: gross output, labor, capital stock, energy inputs, material inputs, and investment. All variables except labor are measured in 1985 Chilean pesos. Labor is a wage-weighted aggregate of white- and blue-collar employees measured in blue-collar person-year equivalents. In addition, a firm id and year are included.

Problem Set 1

a) How many unique firms are included in this dataset?

b) On average, how many years of observation do we have for all the firms? Also, what is the longest and shortest duration?

c) Show me the correlation matrix for gross output, labor, capital stock, energy inputs, material inputs, and investment.

d) Calculating the industry average for each year, and plot time series plots for gross output, labor, capital stock, energy inputs, material inputs, and investment. Comment on these time series plots.

e) Report the estimates of the parameters of a Cobb-Douglas production function with labor, capital, energy, and materials as inputs from OLS estimation. Essentially, you are going to estimate the following econometric model:

$$y_{it} = \beta_0 + \beta_1 k_{it} + \beta_2 l_{it} + \beta_3 e_{it} + \beta_4 m_{it} + \epsilon_{it}$$

where y_{it} is the log of output, k_{it} is the log of capital, l_{it} is the log of labor, e_{it} is the log of energy, and m_{it} is the log of materials. (You need to take logs for your dependent variable and independent variables before running OLS estimation in Python.)

f) (Bonus Question) Do you see entries or exits in the dataset? Calculate the numbers of entries and exits for each year and draw time series plots. Comment on your results. (Hint: It would be helpful to define starting year and ending year for each firm.)

g) (Bonus Question) The data is actually a panel. As you might have learned in Econometrics course, it is useful to use a fixed-effect model for panel data. The fixed-effect model can control for time-invariant unobserved heterogeneity at individual level. Specifically, consider a regression model with individual fixed-effect:

$$y_{it} = \beta_0 + \beta_1 k_{it} + \beta_2 l_{it} + \beta_3 e_{it} + \beta_4 m_{it} + \alpha_i + \epsilon_{it}$$

where α_i is the fixed-effect for each individual. Do you think this type of model is useful in terms of **prediction**? Why, or why not?