MOEC0021
Empirical Methods
University of Zurich — HS 2018
PROF: Greg Crawford / TA: E. Dicarlo, M.R. Greco, S. Bagagli — Handout 1

Problem Set 1

This problem set is due on the **15th of October** at **23:59**.
Solutions should be turned in via email to **emanuele.dicarlo@econ.uzh.ch** in PDF form.
Please follow the following steps when submitting your solution:

1. Email Title: MOEC0021 Problem Set 1 Solutions
2. Attachment Title: GroupName_PS1.pdf
   For example, if my group was called 'DataJedi' I would name the attachment
   DataJedi_PS1.pdf

Remember, your goal is to communicate. Full credit will be given only to the correct solution which is described clearly. Convoluted and obtuse descriptions might receive low marks, even when they are correct. Also, aim for concise solutions, as it will save you time spent on write-ups, and also help you conceptualize the key idea of the problem.

1 Pencil and Paper Questions

1. Suppose you *knew* the process generating the data in a population of interest was of the form

$$Y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

with $\beta_1 = 3$ and $\beta_2 = 1.5$.

(a) Write down the population regression function. Draw a picture of $E(Y_i|X_i)$, the non-random part of the PRF.

(b) You are given the following 4 observations drawn independently from this population:

	X	Y
1	1	4
2	4	10
3	3	9
4	2	7

Construct a table with the following values: the mean of X, $\bar{X}$; the mean of Y, $\bar{Y}$; X in mean-deviation form, $x_i = (X_i - \bar{X})$; Y in mean-deviation form, $y_i = (Y_i - \bar{Y})$, and the cross-product $x_i y_i$, and square of x_i, x_i^2, both in mean-deviation form.

(c) Calculate the OLS estimates of β_1 and β_2.

(d) Plot the 4 points and the OLS line. Note this is the non-residual part of your Sample Regression Function.

(e) How does the OLS line compare to the line you drew from your Population Regression Function?

(f) Does your sample regression function cross the population regression function? Suppose you were to select another sample from the same population. Is it possible that the two lines would *not* cross? Why or why not?

(g) Calculate the error, ϵ_i, for the data points in your sample. Also calculate the residuals, e_i, for these data points. Do the errors sum to zero? Do the residuals? Do your answers differ? If so, explain why in your own words.

(h) Show that $\sum_{i=1}^{4}(X_i - \bar{X}) = 0$. Is this an idiosyncratic feature of this sample or would you expect it to hold in every sample?

(i) Show that $\sum x_i y_i = \sum x_i Y_i$. Also show that this equals $\sum X_i y_i$. Is this an idiosyncratic feature of this sample or would you expect it to hold in every sample?

2. Consider the simple linear regression model

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

(a) Suppose that the *unconditional* expectation $E(\epsilon_i) = \mu_\epsilon \neq 0$. Using the formulas for $\hat{\beta}_1$ and $\hat{\beta}_2$ on slide 4 of the lecture notes, evaluate $E(\hat{\beta}_1)$ and $E(\hat{\beta}_2)$. What does your answer tell you about the robustness of OLS estimation to ϵ_i having a non-zero error?

Consider now the general CLRM with multiple regressors

$$y = X\beta + \epsilon \tag{1}$$

where β includes the constant β_1

(b) Suppose you decided to measure all of your X variables in different units such that your new X variable, call it $\tilde{X}$, is exactly double your old one, i.e. $\tilde{X} = 2X$. Suppose you run the regression of y on $\tilde{X}$; call the resulting estimate $\tilde{\beta}$. What is the relationship between $\tilde{\beta}$ and $\hat{\beta}$, the regular OLS estimator from the regression of y on X?

(c) Suppose you decided now to measure y in different units such that your new y variable, call it y^*, is exactly double your old one, i.e. $y^* = 2 * y$. Suppose you run the regression of $y*$ on X; call the resulting estimate β^*. What is the relationship between β^* and $\hat{\beta}$?

(d) Given your answers to the last two questions, how meaningful are the units in which X and y are measured for the conclusions you draw from an OLS regression?

(e) Return to the scenario in part (2b) above with $\tilde{X} = 2X$. Calculate $V(\tilde{\beta})$. What is the relationship between $V(\tilde{\beta})$ and $V(\hat{\beta})$?

2 Computer Questions

1. In this empirical exercise, we will illustrate the impact of sample size on the variance of the sample mean using what are called "Monte carlo methods". In monte carlo methods, you *create your own data* and then evaluate the properties of functions of that data. While the concepts at play in this question are (fairly) easy, it is not necessarily as easy to program the computer to have it do exactly what you want it to. Thus this question is about having you develop some of your programming skills.

   In this question, we will work with data that are drawn from an *exponential* distribution. If you are not familiar with the exponential distribution, look it up on Wikipedia or Wolfram MathWorld. If a random variable, x_i, is distributed as an exponential, we denote this, $x_i \sim \exp(\lambda)$, where λ is the parameter governing the shape of the distribution. For $x_i \sim \exp(\lambda)$, you can show (or look up) that $E(x_i) = \frac{1}{\lambda}$ and $V(x_i) = \frac{1}{\lambda^2}$. For the rest of this question, we will assume $x_i \sim \exp(1)$.

   This question asks you to draw many samples of data from the distribution of x_i. Each sample is distinguished by its number of observations, which we denote (as usual) N. But in each question below, I will ask you to draw samples of size N many times. We will call these different samples *replications* and index them by the letter $r = 1, \ldots, R$. Thus the i^{th} draw of x from the r^{th} replication can be denoted x_i^r. And the sample average of the N values of x_i^r in the r^{th} replication can be denoted $\bar{x}^r$. We can also take the sample average and variance across the R replications of $\bar{x}^r$, which we will denote $\bar{x}$ (note *no* r) and $s_{\bar{x}}$. $s_{\bar{x}}$ is our estimate of the variance of the sample mean from a sample of size N discussed extensively in lecture.

   (a) Let $N = 1$ and $R = 200$. Calculate $\bar{x}^r$ for $r = 1, \ldots, 200$ show them in a histogram. Also calculate the across-replication average, $\bar{x}$, and sample variance, s_{barx}.

   (b) Let $N = 5$ and $R = 200$. Calculate $\bar{x}^r$ for $r = 1, \ldots, 200$ show them in a histogram. Also calculate the across-replication average, $\bar{x}$, and sample variance, s_{barx}.

   (c) Let $N = 20$ and $R = 200$. Calculate $\bar{x}^r$ for $r = 1, \ldots, 200$ show them in a histogram. Also calculate the across-replication average, $\bar{x}$, and sample variance, s_{barx}.

   (d) Let $N = 1,000$ and $R = 200$. Calculate $\bar{x}^r$ for $r = 1, \ldots, 200$ show them in a histogram. Also calculate the across-replication average, $\bar{x}$, and sample variance, s_{barx}.

   (e) Based on your answers to the previous parts of this question,

      i. For each of $N = 1$, $N = 5$, $N = 20$, and $N = 1,000$: Does the distribution of $\bar{x}^r$ look more like an exponential distribution or a normal distribution?

      ii. Is your estimate of $\bar{x}$ close to $E(x_i) = 1$ in each experiment? If not, why not?

      iii. Is your estimate of $s_{\bar{x}}$ close to $V(x_i) = 1$ in each experiment? If not, why not?

2. In this empirical exercise, we will try to understand more about people's attitude toward smoking. In particular we will try to understand how smoking relates to a person's education and age.

   (a) Download the data from this link [1] and import them into Stata or R. How many observations are there? For each observation, i, there are 10 variables in the dataset, listed here:

   - *educ*: i's years of schooling
   - *cigpric*: the average cigarette price (in cents/pack) in i's state
   - *white*: a dummy variable=1 if i is white
   - *age*: i's age, measured in years
   - *income*: i's annual income,
   - *cigs*: the number of cigarettes smoked by i per day
   - *restaurn*: a dummy variable =1 if the restaurants in i's state restrict smoking
   - *lincome*: the log of income
   - *agesq*: the square of age
   - *lcigpric*: the log of the average cigarette price

   (b) Provide a table of summary statistics for the variables *cigs*, *educ*, *age*, *income*, *white*, *restaurn*. Briefly describe patterns you find particularly interesting (if any).

   (c) We want to estimate the relationship between number of cigarettes smoked and education, measured as i's years of schooling.

   $$cigs_i = \beta_1 + \beta_2 educ_i + \epsilon_i$$

   i. Compute β_1 and β_0 (easier in this order) using the formulas on either slide 4 and/or slide 16 of the lecture notes.

   ii. Run the regression in equation (2c) How do the computer's estimates of β_0 and β_1 relate to the ones you have just computed?

   iii. Suppose that Assumption 2 (Mean-zero error) is satisfied. How do you interpret the coefficient of *educ*? Is this a big or small effect?

   iv. Using your estimates, predict the number of cigarettes consumed by i and denote this $\widehat{cigs}$. In a graph, display both the scatterplot of cigarettes smoked against education and your regression line.

   v. Now regress cigarettes on education **without** including a constant. Generate predicted values and add the new regression line to the previous graph. What changes compared to the earlier regression line? Do you think you should include a constant or not?

[1] http://fmwww.bc.edu/ec-p/data/wooldridge/smoke.dta

(d) Now regress $cigs$ on $educ$, age, age^2, $white$ and $restaurant$ and assume again that Assumption 2 (Mean-zero error) is satisfied

i. What are the coefficients of race (i.e. white) and of the dummy restaurant? How would you interpret them?

ii. Calculate the marginal effect of age on cigarette consumption. What is the value of this marginal effect at age 20? At age 40? At age 60?

iii. Predict the residuals from your model, $e_i = cigs_i - \hat{cigs}_i$, where $\hat{cigs}_i$ is the fitted value of $cigs_i$ from your regression.

A. Construct a scatter plot of these residuals against age. What does this tell you about the likely validity of our Assumption 3?

B. Calculate the correlation of these residuals across individuals. What does this tell you about the likely validity of our Assumption 4?

C. Plot the density of the residuals together with the density of a normal distribution. What does this tell you about the likely validity of our Assumption 5?