MOEC0021
Empirical Methods
University of Zurich — HS 2019
PROF: Greg Crawford / TA: E. Dicarlo, M.R. Greco, S. Bagagli, A. Jenni — Handout 1

Exercise 1

1 Theory

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 (sample) 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 Empirical Application

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?