Geneva Graduate Institute (IHEID)

Econometrics II (EI062), Spring 2025

Marko Mlikota

Problem Set 1

Due: Sunday, 9 March, 23:59

- Prepare concise answers.
- State clearly any additional assumptions, if needed.
- You are encouraged to collaborate in groups but the final write-up should be individual.
- Submit your solutions, along with any code (if applicable), in a **single pdf file** through **Moodle**. If you choose to write your solutions by hand, please make sure your scanned answers are legible.
- Grading scale:

5.5	default grade
6	absolutely no mistakes and particularly appealing write-up (clear and concise answers, decent formatting, etc.)
5	more than a few mistakes, or single mistake and particularly long, wordy answers
4	numerous mistakes, or clear lack of effort (e.g. parts not solved or not really attempted)
1	no submission by due date

Problem 1

Suppose you are interested in estimating the effect of fertilizer on crop yields. Let $y_i > 0$ denote crop yields in USD per acre (realized in one agricultural season), and let $x_i^* > 0$ denote the amount of fertilizer applied (in liters per square meter). The unit of observation i refers to a plot of land of size one acre. Suppose y_i is determined by the following linear function:

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \beta_3 g_i + u_i \ ,$$

where $r_i \in \{0, 1\}$ is an indicator for whether a plot of land is of high quality, and $g_i > 0$ is the precipitation (rainfall) (measured in liters per cubic meter).

(a) Simulate a dataset of size $n = 100$ using the following Data Generating Process (DGP):

1. $u_i \sim N(0, 5)$[1]
2. $g_i \sim \text{Gamma}(2, 2)$[2]
3. $r_i = 1$ and $r_i = 0$ with equal probability
4. $x_i|(r_i = 1) \sim \text{Gamma}(3, 1)$ and $x_i|(r_i = 0) \sim \text{Gamma}(7, 1)$
5. Generate y_i by the equation above, using $\beta_0 = 400$, $\beta_1 = 5$, $\beta_2 = 200$ and $\beta_3 = 10$.

In addition, simulate two further variables: $n_i^1 \sim N(10, 3)$ and $n_i^2 \sim N(5 + \sqrt{x_i}, 3)$.

(b) Using your simulated data, run the following five regressions. For each of them, report your estimate of β_1, compare it to the true value, report its standard error, and discuss your results more generally.

1. regress y_i on x_i^* and a constant (intercept):

$$y_i = \beta_0 + \beta_1 x_i^* + \text{error}_i \ .$$

2. regress y_i on x_i^*, r_i and a constant (intercept):

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \text{error}_i \ .$$

3. regress y_i on x_i^*, r_i, g_i and a constant (intercept):

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \beta_3 g_i + \text{error}_i \ .$$

[1] The first parameter denotes the mean, the second the variance (not the standard deviation!).

[2] The first parameter denotes the shape, the second the scale. See the following wikipedia article.

4. regress y_i on x_i^*, r_i, g_i, n_i^1 and a constant (intercept):

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \beta_3 g_i + \text{error}_i \ .$$

5. regress y_i on x_i^*, r_i, g_i, n_i^1 and a constant (intercept):

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 r_i + \beta_3 g_i + \text{error}_i \ .$$

(c) Repeat the previous questions for $M = 100$ different samples of size $n = 100$. (Concretely, simulate one dataset, run all five regressions and store their output of interest, and proceed in that way $M = 100$ times.) Show histograms of the estimators of β_1 under the five different regressions. (No need to compute its standard error.) Comment on your results.

(d) Repeat your analysis (for $M = 100$ repeated samples) by changing the following elements (one at a time) in the DGP:

- Let $x_i|(r_i = 1) = x_i|(r_i = 0) \sim \text{Gamma}(5, 1)$.
- Let $\beta_2 = 0$.
- Let $r_i = 1$ with probability 0.1.
- Let $\beta_3 = 50$.

You may restrict yourself to the first three regressions.