Econometric Methods: HUL315

Indian Institute of Technology Delhi

Assignment 1

Instructions

- Please make your answers brief and to the point.
- Your answers must be typed and compiled using LaTeX and you must submit a single pdf on Moodle using the link 'Submit Assignment 1'. Handwritten assignments will not be accepted. IIT Delhi provides Overleaf Professional features for all students who would like to use their online LaTeX editor for their projects. Overleaf Professional features include real-time track changes, unlimited collaborators, and full document history with a guide on how to start a project and answers to FAQs. But this requires access to the internet connection. If you do not have a stable internet, you can use other freely available offline tex editors.
- You can work individually or in a group (maximum group size is five) on the assignment. Please mention the names of the group members and their IDs clearly at the first page of the submission. Even if you are working in a group, all member of the group should upload their individual submission. A single submission for a group is not allowed.
- Submit your completed assignment only on Moodle. Please do not email it to the instructor or the TAs. The last date and time of submission is 11th September 11:59 pm. This is a strict deadline, and the submission link will stop working after the deadline.
- This assignment carries **10%** weight in the course.
- For data-related exercises, you can use any statistical software (such as STATA, Python, or R), but you must attach your code, the log file, and relevant outputs, such as figures, tables, etc.

1. (5 points) Consider the following equation, $x + y + z = 100$ where $x, y, z \in \mathbb{Z}_{\geq 0}$. $\mathbb{Z}_{\geq 0}$ is the set of non-negative integers including zero. How many solutions exist for the above equation?

2. (5 points) n friends place n pieces of paper with their names in a hat. After shuffling them each one of them takes one piece. How many possible outcomes exist such that no one takes their own name out of the hat?

3. (5 points) Three numbers x, y, z are randomly selected (with replacement) from the set $\mathbb{Z}_{\geq 0}$ such that all of them are less than 100. What is the probability $x < y < z$.

4. (5 points) A contestant in a TV game show is told that a valuable prize lies behind one of the closed doors marked A, B, and C. There is a "booby prize" in the form of a goat is behind another door, and the third door has nothing at all. The contestant chooses one door. The game master then walks to one of the other doors not chosen and opens it to reveal the goat (placed there by an assistant behind the scenes after the contestant's choice was announced). The contestant, now able to eliminate one door as the prize's location, is given the option to switch from the first choice to the remaining unopened door. For example, if A was chosen first and the goat revealed at C, the contestant could either stick with A or switch to B. Does making the switch increase the probability of winning? Show your work.

5. (5 points) The Magenta line (connects the IIT Delhi campus) metro station has 25 stations. It arrives at the Botanical Garden stop with no passengers on board. At every station, 0, 1, or 2 passengers could get on the metro with a probability of 0.3, 0.5, and 0.2, respectively. An onboard passenger could alight with a probability of 0.1 at any station. Find the expected value and the variance of the number of passengers on the train after leaving the second station. The number of passengers boarding the train at a station is independent of the other stations, and the metro coaches have an infinite capacity.

6. (5 points) Na-real, a company that packages coconut water, is competing with the market leader Real beverages. To remain competitive, Na-real regularly assesses the quality of their product. Unfortunately, the assessment procedure is somewhat destructive, as they must open the juice box. Assume they know that on an average in a batch of 1000 boxes there are 32 that have a problem that would be detected in their assessment. Suppose they take 100 boxes at random for testing. Let X be the number of defective boxes found during the testing. Find the p.m.f. for X.

7. (10 points) The p.m.f. for a discrete random variable X, which can only take natural number values, is given by $f(X = x) = \begin{cases} \dfrac{1}{x(x+1)}, & \text{if } x \in \{1, 2, 3, \ldots\}, \\ 0, & \text{Otherwise.} \end{cases}$

   Find $\mathbb{E}[X]$?

8. A factory is believed to produce a finite number N of consecutively numbered tanks each month. You draw a simple random sample without replacement of size n and observe serial numbers $X_1, X_2, \cdots, X_n$. Consider the following two estimators of N

(i) $\hat{N}_{mn} = 2\bar{X} - 1$

(ii) $\hat{N}_{mx} = \frac{n+1}{n} \times max\{X_1, X_2, \cdots, X_n\} - 1$

(a) (10 points) Compute the mean and variance of the estimators outlined in (i) and (ii).

(b) (10 points) Historically, 220 is the monthly maximum of German tank production. Set $n = 10$ and fix the true $N = 220$. Write a program using your favorite statistical software that plots the empirical distribution of $\hat{N}_{mn}$ and $\hat{N}_{mx}$ after computing the point estimates repeatedly 10,000 times with a random sample of size $n = 10$.

Guidance for part (b) — to standardize outputs (no change to marks):

- Draw *without replacement* from $\{1, \ldots, 220\}$ on each repetition; use $R = 10{,}000$ and set a random seed; include your code.
- Plot both estimators' empirical distributions on the *same axes* (overlaid histograms or kernel densities).
- Add a vertical line at the true value $N = 220$.
- Report the empirical mean, variance, bias (mean $- N$), and MSE (variance + bias^2) for each estimator.

(c) (3 points) Comment on the distributions in part (b).

Data Exercises

9. (a) (7 points) Write a code using either Stata / R / Python to create a vector $\mathbf{x}$ of dimension (1000×1) with realizations of a random variable distributed uniformly over the unit interval $[0, 1]$. Now draw a random sample of 100 elements from the vector $\mathbf{x}$ and calculate the sample mean. Repeat the exercise 1000 times and store the sample means in another vector $\mathbf{y}$ of dimension (1000×1).
Task: Plot the distribution of $\mathbf{x}$ and $\mathbf{y}$ in the same diagram.

   (b) (3 points) Interpret your results.

10. Read the data in the file `datadino.csv` in STATA (or any other statistical software of your choice).

   (a) (5 points) Plot the variables x and y (in their respective axes) using a scatter plot for each distinct string values of the variable `dataset`.

   (b) (3 points) Compute the average values of x and y, the standard deviations of x and y, and the correlation coefficient between x and y for each distinct string values of the variable `dataset`.

   (c) (5 points) Plot the regression line of y on x and a constant on the same plots in part (a).

   (d) (2 points) Summarize your findings in two sentences.

11. Read the data in the file `energy_gdp.csv` in STATA (or any other statistical software of your choice).

    (a) (5 points) Estimate the following simple linear regression and report the point estimates of the parameters α and β:

    $$\log(\text{Energy}) = \alpha + \beta \log(\text{Real GDP}) + \epsilon$$

    (b) (5 points) Plot the residuals. Does anything strike you?

    (c) (5 points) Test the hypothesis that the coefficient on log(Real GDP) is equal to one.

    (d) (5 points) For one country the energy data is recorded wrongly. Multiply it with 1000 and repeat the exercise in (a), (b), and (c).

    (e) (5 points) Now estimate the following equation instead and report the point estimates of the parameters δ and γ:

    $$\log(\text{Real GDP}) = \delta + \gamma \log(\text{Energy}) + \epsilon.$$

    (f) (2 points) Compute the R^2 values from both the models.

    (g) (5 points) Prove that $\hat{\beta} \times \hat{\gamma} = R^2$ as a general result. Confirm that it holds numerically from the models that you have estimated.

    (h) (5 points) Suppose energy is measured in BTU's instead of kilograms of coal. This implies that the original series must be multiplied by 60. How does it change the estimates for the intercept and slope in the following equation?

    $$\log(\text{Energy}) = \alpha + \beta \log(\text{Real GDP}) + \epsilon.$$

    Explain why $\hat{\alpha}$ is not the same as earlier but $\hat{\beta}$ is, in the log-log equation?

12. The data, `Earnings_and_Height.csv`, consists of earnings, height, and other characteristics of a random sample of U.S. workers, taken from the US National Health Interview Survey for 1994. It is a subset of the data used in Anne Case and Christina Paxson's paper "Stature and Status: Height, Ability, and Labor Market Outcomes," *Journal of Political Economy*, 2008,116(3): 499–532. The dataset contains information on 17,870 workers. Using the data description provided on the next page,...

    (a) (1 points) What is the median value of height in the sample?

    (b) Do the following –

        i. (1 points) Estimate average earnings for workers whose height is at most 67 inches.

        ii. (1 points) Estimate average earnings for workers whose height is greater than 67 inches.

        iii. (2 points) On average, do taller workers earn more than shorter workers? How much more? What is a 95% confidence interval for the difference in average earnings?

(c) (3 points) Construct a scatterplot of annual earnings (`Earnings`) on height (`Height`). Notice that the points on the plot fall along horizontal lines. Why?

(d) Run a regression of Earnings on Height.

i. (1 points) What is the estimated slope?

ii. (1 points) Use the estimated regression to predict earnings for a worker who is 67 inches tall, for a worker who is 70 inches tall, and for a worker who is 65 inches tall.

(e) Suppose height were measured in centimetres instead of inches. Answer the following questions about the Earnings on Height (in cm) regression.

i. (2 points) What is the estimated slope of the regression?

ii. (1 points) What is the estimated intercept?

iii. (1 points) What is the R^2?

iv. (1 points) What is the standard error of the regression?

(f) Run a regression of Earnings on Height, using data for female workers only.

i. (2 points) What is the estimated slope?

ii. (3 points) A randomly selected woman is 1 inch taller than the average woman in the sample. Would you predict her earnings to be higher or lower than the average earnings for women in the sample? By how much?

(g) (5 points) Do you think that height is uncorrelated with other factors that cause earning? That is, do you think that the regression error term, u_i has a conditional mean of 0 given Height (X_i)?

Variable Name Description

Variable Name	**Description**
age	Age, in years
cworker	Class of Worker: 1 = Private company Employee 2 = Federal Government Employee 3 = State Government Employee 4 = Local Government Employee 5 = Incorporated Business Employee 6 = Self Employed
earnings	annual labor earnings, expressed in \$2012 (see Table notes)
educ	years of education
height	height without shoes (in inches)

Variable Name	Description
mrd	Marital Status 1 = Married, Spouse in household 2 = Married, Spouse not in household 3 = Widowed 4 = Divorced 5 = Separated 6 = Never Married
occupation	Occupations in 15 categories: 1 = Exec/Manager 2 = Professionals 3 = Technicians 4 = Sales 5 = Administrat 6 = Household service 7 = Protective service 8 = Other Service 9 = Farming 10 = Mechanics 11 = Construction/Mining 12 = Precision production 13 = Machine Operator 14 = Transport 15 = Laborer
Race	race/ethnicity 1 = non-Hispanic white 2 = non-Hispanic black 3 = Hispanic 4 = other
region	Region of the U.S. 1 = Northeast 2 = Midwest 3 = South 4 = West
sex	Sex, 1=Male, 0 = Female
Weight	weight without shoes (in pounds)

Notes: In the survey, labor earnings are reported in 23 brackets (for example, \$26,000–\$30,00). For each of these brackets, a value of average earnings based on information in the Current Population was estimated, and these average values were assigned to all workers with incomes in the corresponding bracket. The earnings values for 1994 were converted to \$2012 using the consumer price index.