Econometric Methods: HUL315

Indian Institute of Technology Delhi

Assignment 2

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 2'. 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 **17th November 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. (6 points) Consider the following simple bivariate linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i.$$

   (a) (3 points) Derive the OLS estimators for β_0 and β_1.

   (b) (3 points) Show that the OLS estimators in (a) are the same as the vector expression below:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

   where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the regressor and dependent variable matrices respectively.

   The matrices and vectors are defined as:

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \qquad \boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}.$$

2. (20 points) A researcher is using the following regression specifications to estimate the relationship between variables X and Y using a random sample of size n.

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i \quad (1)$$

$$\ln\left(\frac{Y_i}{X_i}\right) = \alpha_1 + \alpha_2 \ln X_i + \nu_i \quad (2)$$

   (a) (2 points) Determine whether equation (2) can be expressed as a restricted version of equation (1).

   (b) (2 points) Find the relationship between the OLS estimators of the parameters for equation (1) and (2).

   (c) (4 points) Define $y = \ln(Y)$; $x = \ln(X)$; and $z = \ln\left(\frac{Y}{X}\right)$. Show $\hat{z} = \hat{u} - x$.

   (d) (4 points) Show that the residuals from (2) are identical to those from (1).

   (e) (5 points) Show that the standard errors of equation (1) and (2) are identical.

   (f) (3 points) Argue whether R^2 would be the same for the two regressions.

3. (15 points) Use the aatached German health care usage dataset and estimate the following regression:

$$\text{hhninc} = \beta_0 + \beta_1 \text{educ} + \varepsilon$$

   (a) (3 points) Estimate the model above and plot $\hat{\beta}_1$ as a vertical line on the Cartesian plane (with estimates of β_1 measured along the horizontal axis).

   (b) (10 points) Draw a random sample of size 30 from the data and estimate β_1 on this restricted sample. Repeat this 1000 times and store the estimates in a vector $\mathbf{b}_1$. Plot the empirical distribution of $\mathbf{b}_1$ on the same graph as in part (a).

   (c) (2 points) Interpret your results briefly.

4. (15 points) Use the German health care usage dataset and estimate the following model:

$$\text{hhninc} = \beta_0 + \beta_1\text{educ} + \beta_2\text{univ} + \beta_3\text{working} + \beta_4\text{bluec} + \beta_5\text{selfemp} + \varepsilon \qquad (3)$$

   (a) (3 points) Estimate the model in equation (3).

   (b) (5 points) Estimate the models in equations (4) and (5) and collect the residuals in variables `hhninc_res` and `educ_res`:

$$\text{hhninc} = \gamma_0 + \gamma_1\text{univ} + \gamma_2\text{working} + \gamma_3\text{bluec} + \gamma_4\text{selfemp} + \varepsilon \qquad (4)$$

$$\text{educ} = \delta_0 + \delta_1\text{univ} + \delta_2\text{working} + \delta_3\text{bluec} + \delta_4\text{selfemp} + \varepsilon \qquad (5)$$

   (c) (2 points) Estimate the model in equation (6):

$$\text{hhninc\_res} = \beta_0 + \beta_1\text{educ\_res} + \varepsilon \qquad (6)$$

   (d) (5 points) Compare the point estimates on `educ` and `educ_res` from regressions (3) and (6). Provide a theoretical explanation for your findings.

5. (10 points) You are analyzing a balanced panel dataset of n individuals observed over T periods. Consider the model below:

$$\ln(\text{wage})_{it} = \beta_0 + \beta_1\text{educ}_i + \beta_2\text{exper}_{it} + \beta_3\text{exper}^2_{it} + c_i + u_{it},$$

   where c_i denotes unobserved, time-invariant individual heterogeneity and u_{it} is an idiosyncratic error term.

   (a) (3 points) Explain why estimating this equation using pooled OLS can lead to a biased estimate of β_1. What kind of omitted variable problem does c_i introduce?

   (b) (3 points) Show how the *within* (individual fixed-effects) transformation eliminates c_i. Derive the transformed regression using deviations from individual means.

   (c) (2 points) Briefly discuss how including individual-specific intercepts in the regression achieves the same result as the within transformation.

   (d) (2 points) Suppose you also include a time dummy for each year in the panel. Explain what additional source of heterogeneity this controls for and why this might be useful in empirical work.

6. (10 points) In the 1960s and early 1970s, young American men were drafted for military service to serve in Vietnam. Concerns about the fairness of the conscription policy led to the introduction of a draft lottery in 1970. Angrist (1990) estimates the causal effect of veteran status on earnings using draft eligibility (d_i), a binary variable, as an instrument for veteran status (s_i). The structural equation is: $y_i = \alpha_0 + \alpha_s s_i + u_i$ where y_i denotes the earnings of individual i and s_i indicates veteran status.

   The first stage of the model is given by: $s_i = \beta_0 + \beta_s d_i + \nu_i$ where d_i indicates draft eligibility. u_i and ν_i are white noise error terms. Show that the instrumental variables estimate for α_s is given by the ratio of the differences in average earnings for the draft-eligible ($\bar{y}_e$) and ineligible ($\bar{y}_n$) groups and the difference in the proportion of individuals actually entering military service among the draft-eligible ($\bar{s}_e$) and ineligible ($\bar{s}_n$) groups, i.e.,

$$\hat{\alpha}_s = \frac{\bar{y}_e - \bar{y}_n}{\bar{s}_e - \bar{s}_n}.$$

7. (24 points) Concerned about declining attendance in late afternoon lectures, the Dean Academics at IIT Delhi has asked instructors to analyze the determinants of attendance using data rather than anecdotes. As part of this initiative, an HUL315 instructor designed a simple experiment. In pre-midterm lectures, short unannounced quizzes were conducted in class, whereas in post-midterm lectures, quizzes were pre-announced. For each student i and lecture t, attendance is recorded as a binary variable

$$y_{it} = \begin{cases} 1, & \text{if student } i \text{ attended lecture } t, \\ 0, & \text{otherwise.} \end{cases}$$

Let PreMidterm_t be a dummy variable that equals 1 for lectures held before the midterm exam, and AnnouncedQuiz_t be a dummy that equals 1 for lectures where a quiz was pre-announced. The distance of each student from their hostel to the Lecture Hall Complex (LHC) is indicated by Dist_LHC_i (in kilometers).

The attendance decision is modeled as a latent variable:

$$y^*_{it} = \beta_0 + \beta_1 \, \text{PreMidterm}_t + \beta_2 \, \text{AnnouncedQuiz}_t + \beta_3 \, \text{Dist\_LHC}_i + u_{it},$$

where $u_{it} \sim \text{Logistic}(0, 1)$ with cumulative distribution function $\Lambda(z) = \dfrac{1}{1 + e^{-z}}$. The observed outcome is $y_{it} = \mathbf{1}\{y^*_{it} > 0\}$. Define the vector of controls as

$$\mathbf{x}_{it} = \big(1, \ \text{PreMidterm}_t, \ \text{AnnouncedQuiz}_t, \ \text{Dist\_LHC}_i\big)'$$

and the corresponding parameter vector as $\boldsymbol{\beta} = \big(\beta_0, \ \beta_1, \ \beta_2, \ \beta_3\big)'$. Thus, the model can be compactly written as $y^*_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + u_{it}$.

(a) (3 points) Using the distributional assumption on u_{it}, write down the probability that a student attends a lecture given the observed covariates, i.e., $\Pr(y_{it} = 1 \mid \mathbf{x}_{it})$. Express your answer using the logistic CDF $\Lambda(\cdot)$.

(b) (3 points) Derive the expression for the *log-odds ratio* (or *logit*) of attending a lecture, defined as

$$\log\left(\frac{\Pr(y_{it} = 1 \mid \mathbf{x}_{it})}{1 - \Pr(y_{it} = 1 \mid \mathbf{x}_{it})}\right).$$

Show that it is linear in the covariates and equals $\mathbf{x}'_{it}\boldsymbol{\beta}$.

(c) (2 points) Interpret the coefficients β_1 and β_2 in the context of this model. What do positive values of these coefficients imply about attendance behavior before the midterm and on preannounced quiz days?

(d) (3 points) Suppose you do not observe Dist_LHC_i in the data. Explain how omitting this variable could bias the estimated coefficients on PreMidterm_t and AnnouncedQuiz_t. Provide a potential solution to the problem of unobserved data that do not change over time.

(e) (3 points) Write the expression for the marginal effect of Dist_LHC_i on the probability of attendance, $\dfrac{\partial \Pr(y_{it} = 1 \mid \mathbf{x}_{it})}{\partial\, \text{Dist\_LHC}_i}$, and discuss why its magnitude depends on $\mathbf{x}_{it}'\boldsymbol{\beta}$. Can you compute the average marginal effect (AME) of distance from LHC from the sample data?

(f) (10 points) Estimate the model using the attendance data from rollcall. Since Dist_LHC_i is unobserved, use the approach proposed in part (d) to address the resulting unobserved heterogeneity. Convert the data into a panel format with one observation per student-lecture pair before estimation. Based on your results, prepare a brief note summarizing your key findings and their implications for the Dean Academics.