Econometrics II, ECO351, Semester I, 2025-26
Homework III (125 points)

Instructor: M.A. Rahman

Deadline: 6:00 pm, October 30, 2025 (in my office: Room 409, ESB-2).

Please read the instructions carefully and follow them while writing answers.

- *Solutions to homework should be written in A4 size loose sheets. If you are not comfortable writing on white sheets, please ask for biology paper in Tarun Book Store.*
- *Questions should be answered in order as they appear in the homework. Every new question should begin in a new page. Please number all the pages of your homework solution. Please leave a margin of one inch from top and one inch from left. Staple the sheets on the top-left.*
- *Please submit computational assignments (if any) and written answers together and in the correct order. Your answer script should directly address the questions, with all code included in the appendix. All questions in this assignment must be completed using MATLAB. Work done in any other software will NOT be accepted.*

---

1. (5+5+5+5 = 20 points) `Truncation and Censoring`: Suppose that the length of employment of a given individual (say in months) at a large corporation follows an exponential distribution, with *pdf* and *cdf* given as,

$$\begin{aligned} f(y_i) &= \mu^{-1}\exp(-y_i/\mu), \qquad y \geq 0,\ \mu > 0, \\ F(y_i) &= 1 - \exp(-y_i/\mu). \end{aligned} \tag{1}$$

Based on the above information, answer the following.

(a) Assume you have a sample of employees for who you record their length of employment (y_i) and a set of socio-demographic characteristics, x_i ($k \times 1$ vector). To allow these attributes to enter the model you decide to parameterize μ as $\exp(x_i'\beta)$. Write down the *pdf* and *cdf* for this parameterized model.

(b) For this parameterized model, what is the expression for the probability that a worker is employed for more than 36 months?

(c) Assume your sample of $\tilde{n}$ workers includes only those who have been with the company for over 36 months. To be specific, you have neither information on y_i nor x_i for workers with $y_i \leq 36$. Write down the individual likelihood function, $l(\beta)$; individual log-likelihood, $\ln l(\beta)$; and sample log-likelihood function $\ln L(\beta)$ corresponding to this situation. Is this sample truncated or censored? Explain.

(d) Same as Part (c) but now assume that you know x_i for *all* n workers in the company. However, as before you only observe y_i for those $\tilde{n}$ workers who have been around for over 36 months. Write down the individual likelihood contribution *for each case* (i.e., ≤ 36, > 36), the individual log-likelihood *for each case*, and the sample log-likelihood. Is this sample truncated or censored? Explain

2. **(3+2+3+3+8+3+3 = 25 points)** `Tobit Type I Model`: The data file "`Aptitude.xlsx`" contains 200 observations on four variables: reading score (`read`), maths score (`math`), type of program (`prog`) categorized as academic, general or vocational, and academic aptitude (`apt`). Based on the data answer the following questions.

(a) Create an indicator variable for each program type as follows: `academic=1` if program is academic, `general=2` if program type is general, and `vocational=1` if program type is vocational. What are the number and percentage of observations in each category?

(b) Present a descriptive summary for the three variables i.e., `read`, `math`, and `apt`. Report the mean, median, standard deviation, maxmimum, and minimum of the variables.

(c) Present a histogram of `apt` classified according to the different types of academic program. What do yo observe in the histogram?

(d) Present a correlation table for the three continuous variables i.e., `read`, `math`, and `apt`. Also, present a scatter plot for (`read`, `math`), (`read`, `apt`) and (`math`, `apt`). Do you observe any censoring in the scatter plots?

(e) Estimate a Tobit Type I model (CRM with normal errors) where `apt` is the dependent variable and the covariates are `read`, `math`, and `prog` types. Use academic as the base category for the program type. Report the coefficient estimates, standard errors, and the t-statistics. Are all the coefficients significant?

(f) What is the expected change in `apt` for a one unit increase in `read`? What is the expected change in `apt` for a one unit increase in `math`? What is the expected change in `apt` for a student taking vocational program?

(g) Create a scatter plot of residuals and fitted (or predicted) values of `apt`. Similarly present a scatter plot of `apt` and fitted values of `apt`. What do you observe?

3. **(2+3+5+5+5+5+5+10 = 40 points)** Consider using data file "`mroz.xlsx`", where the sheet "`Data`" contains the data and sheet "`Desc`" contains the definition of the variables. We will only use a subset of the variables present in the file.

Our goal is to estimate a model explaining married woman's hours of work (`hours`), as a function of her education (`educ`), her experience (`exper`), and her husband's hours of work (`hhours`). Based on the above data file, answer the following.

(a) Use all observations to estimate the regression model,

$$\texttt{hours} = \beta_1 + \beta_2\,\texttt{educ} + \beta_3\,\texttt{exper} + \beta_4\,\texttt{hhours} + \varepsilon, \tag{2}$$

and report the coefficient estimates, standard errors, and t-statistics. Is ordinary least squares (OLS) a consistent estimator in this case?

(b) Use only the observations for which $\texttt{hours} > 0$ to estimate the regression model in Part (a). Report the coefficient estimates, standard errors, and t-statistics. Is OLS a consistent estimator in this case?

(c) Estimate a probit model for the women's decision to be in the labor force, `LFP`=1, where

$$\Pr(\texttt{LFP} = 1) = \Phi\big(\gamma_1 + \gamma_2\,\texttt{exper} + \gamma_3\,\texttt{kidl6} + \gamma_4\,\texttt{kids618} + \gamma_5\,\texttt{MTR} + \gamma_6\,\texttt{largecity}\big). \tag{3}$$

Report the coefficient estimates, standard errors, and t-statistics. Which if any of the variables help explain the woman's labor force participation decision?

(d) Using the estimates from the probit model, obtain,

$$\tilde{w} = \tilde{\gamma}_1 + \tilde{\gamma}_2\,\texttt{exper} + \tilde{\gamma}_3\,\texttt{kidl6} + \tilde{\gamma}_4\,\texttt{kids618} + \tilde{\gamma}_5\,\texttt{MTR} + \tilde{\gamma}_6\,\texttt{largecity}.$$

Create the inverse Mills ratio $\tilde{\lambda} = \phi(\tilde{w})/\Phi(\tilde{w})$. What are the sample mean and variance of $\tilde{\lambda}$?

(e) Estimate the model,

$$\texttt{hours} = \beta_1 + \beta_2\,\texttt{educ} + \beta_3\,\texttt{exper} + \beta_4\,\texttt{hhours} + \beta_\lambda \tilde{\lambda} + \varepsilon,$$

using the observations for which $\texttt{hours} > 0$ and report the coefficient estimates, standard errors, and t-statistics. Compare these estimates to those in parts (a) and (b). Are the standard errors from this estimation correct?

(f) Estimate the model in part (e) using heteroscedasticity robust standard errors. These standard errors are not absolutely correct but an improvement over the ones in part (e). Once again, report the coefficient estimates, standard errors, and t-statistics.

[`Hint`: To obtain heteroscedastic consistent (or robust) estimates and standard errors, estimate a linear regression model and collect the residuals $\hat{\varepsilon}' = (\hat{\varepsilon}_1, \hat{\varepsilon}_2, \cdots, \hat{\varepsilon}_n)$ and estimate the covariance matrix $\hat{\Omega}_{n\times n} = \text{diag}[\hat{\varepsilon}_1^2,\, \hat{\varepsilon}_2^2,\, \cdots\,,\, \hat{\varepsilon}_n^2]$. Then estimate $\hat{\beta}_{FGLS} = (X'\hat{\Omega}^{-1}X)(X'\hat{\Omega}^{-1}y)$ and $\widehat{V}(\hat{\beta}) = (X'X)^{-1}(X'\hat{\Omega}X)(X'X)^{-1}$.]

(g) Estimate the model in part (e) using bootstrap standard errors, with $B = 400$ bootstrap replications. Compare the standard errors to those in parts (e) and (f).

(h) Estimate the Heckman's sample selection model, where (2) is the outcome equation and (3) is the selection equation, using the full information maximum likelihood (FIML). That is, maximize the log-likelihood with respect to the model parameters using the maximum likelihood method and report the coefficient estimates, standard errors, and t-statistics. Compare the results from two-stage estimator in part (e) to those obtained from the FIML approach.

4. **(2+3+5+5+5+15+5= 40 points)** Consider using data file "`jtrain2.xlsx`", where the sheet "`Data`" contains the data and sheet "`Desc`" contains the definition of the variables. We will only use a subset of the variables present in the file "`jtrain2.xlsx`".

The data comprise 445 observations on low-income male workers, 118 of whom underwent some job training in the late 1970s. The outcome of interest are real earnings in 1978 (training occurred in 1975-1977). The data set is sorted by treatment (treated first, then untreated), and has information on 19 variables.

Considering the treatment effect model,

$$T_i^* = x_{1i}'\beta_1 + \varepsilon_{1i} \tag{4}$$

$$y_i = x_{2i}'\beta_2 + \gamma T_i + \varepsilon_{2i} \tag{5}$$

$$T_i = \begin{cases} 1 & \text{if } T_i^* > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right). \tag{7}$$

and information on the above data, answer the following questions.

(a) Use the indicator variable `train` to determine the fraction of men receiving job training.

(b) The variable `re78` is earnings from 1978, measured in thousands of 1982 dollars. Find the averages of `re78` for the sample of men receiving job training and the sample not receiving job training. Is the difference economically large?

(c) The variable `unem78` is an indicator of whether a man is unemployed or not in 1978. What fraction of the men who received job training are unemployed? What about for men who did not receive job training? Comment on the difference.

(d) From parts (b) and (c), does it appear that the job training program was effective? What would make our conclusions more convincing?

(e) Write down the likelihood function of the treatment effect model as presented in equations (4)-(7).

(f) Now, consider the data present in the file "`jtrain2.xlsx`". Suppose, for the treatment equation, the dependent variable `train` is a function of the following covariates: `Intercept`, `age`, `educ`, `black`, `hisp`, `married`, `nodegree`, `re74`, and `unemp74`.

For the outcome equation, the dependent variable re78 is function of the following covariates: Intercept, age, educ, black, hisp, married, re74, re75, and train.

Note that nodegree and unem74 are the identification variables that are not present in the outcome equation.

Based on the above information estimate the treatment effect model by maximizing the log-likelihood with respect to model parameters. Report the coefficient estimates, standard errors, and t-statistics of the parameters from treatment and outcome equations. Report the estimates, standard errors, and t-statistics for the components of covariance matrix Σ. Is σ_{12} significant?

(g) What is the coefficient estimate for γ, the parameter for the indicator variable train in the outcome (or wage) equation? Is it statistically significant? Based on this result, comment on the effectiveness of job training program.