ECON6300/7320: Advanced Microeconometrics

Problem Set 1

Fu Ouyang

March 14, 2024

Instruction

Answer all questions and clearly label your answers. For empirical questions, you should show your R script(s) and outputs (e.g., screenshots for commands, tables, and figures, etc.). You will lose *2 points* whenever you fail to provide R commands and outputs. When you are asked to explain or discuss something, your response should be brief and compact. You should upload your assignment (in PDF or Word format) via the "Turnitin" submission link (in the "Problem Set 1" folder under "Assessment") by 11:59 AM on the due date March 28, 2024. Do not hand in a hard copy. You are allowed to work on this assignment in groups; that is, you can discuss how to answer these questions with your group members. However, this is *not* a group assignment, which means that you must answer all the questions in your own words and submit your work separately. The marking system will check the similarity, and UQ's student integrity and misconduct policies on plagiarism apply.

Assignment Project Exam Help
https://tutorcs.com
WeChat: cstutorcs

OLS (30 points)

Use the `cps09mar` dataset described in Tutorial 1. Take the sub-sample of non-Hispanic women to estimate the following wage equation:

$$\log(\texttt{wage}) = \beta_0 + \beta_1 \texttt{education} + \beta_2 \texttt{experience} + \beta_3 \texttt{experience}^2/100 + u, \tag{1}$$

where $\texttt{wage} = \texttt{earnings}/(\texttt{hours} \times \texttt{week})$ and $\texttt{experience} = \texttt{age} - \texttt{education} - 6$.

(a) Estimate equation (1) using the described sub-sample and compute the R^2 (5 points).

(b) Include a set of dummy variables for regions and marital status in (1) and estimate the extended model. For regions, create dummy variables for Northeast, South, and West so that Midwest is the excluded group. For marital status, create variables for married ($\texttt{marital} \leq 3$), widowed or divorced, and separated, so that single (never married) is the excluded group (6 points). Calculate standard errors using the HC0, HC1, and HC3 methods (3 points). Are they very different?

(c) In what follows, use the estimation results obtained from (a). Let θ be the ratio of the return to one year of education to the return to one year of experience for $\texttt{experience} = 10$. Write θ as a function of the regression coefficients and compute $\hat{\theta}$ from the estimated model (3 points). Suppose the OLS estimator for model (1) is consistent. Is $\hat{\theta}$ consistent (3 points)? Hint: Apply continuous mapping theorem.

(d) Compute the regression function at $\texttt{education} = 16$ and $\texttt{experience} = 10$ (2 points). Compute a 95% confidence interval for the regression function at this point (3 points). Hint: You can use the `glht()` function in the `multcomp` package.

(e) Consider the same out-of-sample individual as in (d) (`education` = 16, `experience` = 10). Construct an 90% forecast interval for their wage (5 points). Hint: To obtain the forecast interval for the wage, apply the exponential function to both endpoints.

MLE: Tobit Regression (25 points)

Consider the Tobin (1958) regression model:

$$Y^* = X^T\beta + e \text{ with } e|X \sim N(0, \sigma^2), \tag{2}$$

$$Y = \max\{Y^*, 0\}. \tag{3}$$

Tobin (1958) used this model to study household consumption Y of durable goods. He observed that in survey data, Y is zero for a positive fraction of households. He proposed treating the observed Y as a *censored* realization from a latent continuous variable Y^* (like "willingness to pay"); that is, $Y = Y^*$ when $Y^* > 0$ and $Y = 0$ when $Y^* \leq 0$. Here the observed variable Y is censored Y^* from below at zero. After all, negative pay is infeasible. This model is known as *Tobit regression*, *censored regression*, or *Type I Tobit model*.

Since $e|X \sim N(0, \sigma^2)$ is assumed in (2), Tobit model (2)–(3) is parametric and its unknown parameters (β, σ) can be estimated using the maximum likelihood method. By definition, it is easy to write out the distribution function of Y conditional on X:

$$\begin{aligned} F(y|x) &= P(Y \leq y|X = x) = P(Y^* \leq 0|X = x)^{\mathbf{1}[y\leq 0]} \cdot P(Y^* \leq y|X = x)^{\mathbf{1}[y>0]} \\ &= P(x^T\beta + e \leq 0)^{\mathbf{1}[y\leq 0]} \cdot P(x^T\beta + e \leq y)^{\mathbf{1}[y>0]} \\ &= \Phi(-x^T\beta/\sigma)^{\mathbf{1}[y\leq 0]} \cdot \Phi((y - x^T\beta)/\sigma)^{\mathbf{1}[y>0]}, \end{aligned} \tag{4}$$

where $\mathbf{1}[\cdot]$ is the indicator function[1] and $\Phi(\cdot)$ is the CDF of $N(0, 1)$. Taking derivative of (4) with respect to y yields the likelihood function:

$$f_{Y|X}(y|x) = \Phi(-x^T\beta/\sigma)^{\mathbf{1}[y\leq 0]} \cdot [\sigma^{-1}\phi((y - x^T\beta)/\sigma)]^{\mathbf{1}[y>0]}, \tag{5}$$

where $\phi(\cdot)$ is the PDF of $N(0, 1)$.

Use the `CHJ2004` dataset (see the data description file). The variables `tinkind` and `income` are household transfers received in-kind and household income, respectively. Divide both variables by 1000 to standardize.

(a) Estimate a linear regression of `tinkind` on `income` and `income`^{2} (5 points).

(b) Calculate the percentage of censored observations (`tinkind` = 0) (3 points). Estimate a linear regression of `tinkind` on `income` and `income`^{2} by omitting the censored observations (5 points).

(c) Estimate a Tobit regression of `tinkind` on `income` and `income`^{2} (6 points). Explain the differences between your results in (a)–(c) (6 points). Hint: You can use the `tobit`() function in the `AER` package. Alternatively, you can also use (5) to code up the maximum likelihood estimation and inference by yourself and check if you can replicate the results returned by `tobit`(). But this is optional.

[1] $\mathbf{1}[A] = 1$ if event A occurs and $\mathbf{1}[A] = 0$, otherwise.

Assignment Project Exam Help
https://tutorcs.com
WeChat: cstutorcs

2SLS and GMM (25 points)

In an influential paper, David Card (1995) suggested that if a potential student lives close to a college, this reduces the cost of attendance and raises the likelihood that the student will attend college. However, college proximity does not directly affect a student's skills or abilities, so it should not affect their wage. These considerations suggest that college proximity can be a valid IV for education in a wage regression.

Use the `Card1995` dataset to replicate the baseline analysis conducted in Card (1995). Consider the following model:

$$\begin{aligned}\log(\texttt{wage}) =& \beta_0 + \beta_1 \texttt{education} + \beta_2 \texttt{experience} + \beta_3 \texttt{experience}^2/100 \\ &+ \beta_4 \texttt{black} + \beta_5 \texttt{south} + \beta_6 \texttt{smsa} + u, \end{aligned} \tag{6}$$

where `education` is years of schooling and `experience` = `age` − `education` − 6. For all the questions below, only use data for 1976. See the data description file for variable definitions.

(a) Estimate model (6) by OLS and 2SLS (using `nearc4` as the instrument for `education`) (8 points).

(b) Estimate model (6) by 2SLS using instruments {`nearc4a`, `nearc4b`} (3 points). What is the impact on the structural estimate of model (6) (2 points)?

(c) Are the instruments used in (b) strong or weak? Test it (4 points).

(d) Use the 2SLS regression in (b) to test the exogeneity of `education` (4 points).

(e) Use the 2SLS regression in (b) to test the exogeneity of instruments (4 points).

(f) (Optional) Re-estimate model (6) using the same set of instruments as in (a) by efficient GMM. Do the results change meaningfully? Hint: Use the `gmm()` function in the `gmm` package. This question has no points; it is here just because we don't have room for it in Tutorial 4.

Assignment Project Exam Help
https://tutorcs.com
WeChat: cstutorcs

Theoretical Questions (20 points)

1. Prove that the regression errors of LPM must be heteroskedastic (5 points).

2. Prove that $E[Y|X] = \arg\min_g E[(Y - g(X))^2]$; i.e., $E[Y|X]$ has the smallest *mean squared error* (MSE) among all functions of X. Hint: Consider $(Y - g(X))^2 = [(Y - E[Y|X]) + (E[Y|X] - g(X))]^2$ and apply the law of iterated expectation (LIE) (5 points).

3. Consider the following simple linear regression model:

$$y_i = \beta x_i + \varepsilon_i, \ i = 1, ..., n, \tag{7}$$

where $\{x_i, y_i\}_{i=1}^n$ are i.i.d., $E[\varepsilon_i|x_i] = 0$, and $\mathrm{V}[\varepsilon_i|x_i] = \sigma^2$. Suppose $x_i > 1$ and $E[x_i^2] < \infty$ for all $i = 1, \ldots, n$. We have the following three estimators of β.

$$\hat{\beta}_A = \sum_{i=1}^n x_i y_i / \sum_{i=1}^n x_i^2,$$

$$\hat{\beta}_B = \sum_{i=1}^n y_i / \sum_{i=1}^n x_i,$$

$$\hat{\beta}_C = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i}.$$

(a) Show that $\hat{\beta}_A$, $\hat{\beta}_B$ and $\hat{\beta}_C$ are all unbiased (6 points). Hint: Use (7) and LIE.

(b) Among $\hat{\beta}_A$, $\hat{\beta}_B$ and $\hat{\beta}_C$, which one has the smallest variance? Why? (2 points) Hint: Review the Gauss-Markov Theorem.[2]

(c) Show that $\hat{\beta}_C$ is consistent (2 points). Hint: Apply WLLN and LIE.

Assignment Project Exam Help

https://tutorcs.com

WeChat: cstutorcs

[2]When we say an estimator is linear, we mean the estimator is a linear function of y.