Geneva Graduate Institute (IHEID)

Econometrics I (EI035), Fall 2024

Marko Mlikota

Problem Set 5

Due: Sunday, 24 November, 23:59

- Prepare concise answers.
- State clearly any additional assumptions, if needed.
- 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 have a dataset containing a shop's sales, which includes the date, some characteristics of the customer (like income, age), some characteristics of the transaction (like type of good sold, price, and whether cash or a card was used). You are interested in shedding light on the determinants of cash vs card payment.

(a) How could you use the probit model for your research question? What is your y_i variable? How can we interpret the underlying latent variable y_i^*?

(b) In your probit model, derive the effect of age increasing by 5 years on the probability of using cash. Does the effect depend on the current age of the customer? Does it depend on the values of the other variables?

(c) Could you use a standard linear regression, estimated via OLS, to answer your question?

(d) Derive the same effect as in (b) in your linear regression model. Does it depend on the current age of the customer? Does it depend on the values of the other variables?

(e) Based on your reasoning so far, for which customers would you expect the predicted effect under the linear regression to be close to the one under probit? For what type of customers will the two differ more? As a result, for what kind of research questions is the linear regression a good/bad specification?
Hint: Besides comparing (partial) effects under the two models, you might want to compare the functional form of $\mathbb{E}[y_i|x_i]$ *under the two models.*

Problem 2

Suppose you are interested in relating air quality in different cities – measured by the concentration of carbon monoxide in the air, y_i – to possible determinants x_i. The measurement device used in your data cannot detect concentrations below a certain value, δ, but simply codes them as zero. For this purpose, you set up a Tobit model for observations y_i with a lower-censoring at δ:

$$\begin{aligned} y_i^* &= x_i'\beta + u_i \ , \quad u_i \sim N(0, \sigma^2) \ , \\ y_i &= y_i^* \ \mathbf{1}\{y_i^* > \delta\} \ . \end{aligned} \tag{1}$$

(a) Derive the probability of measuring a concentration of carbon monoxide of zero as a function of determinants x_i (and parameters β and the censoring point δ), $\mathbb{P}[y_i = 0|x_i]$.

(b) Derive the conditional mean $\mathbb{E}[y_i^*|x_i]$, i.e. the expected air quality (true concentration of carbon monoxide) for generic a city i with characteristics x_i.

(c) Derive the conditional mean $\mathbb{E}[y_i|x_i]$, i.e. the expected (measurable) concentration of carbon monoxide for generic a city i with characteristics x_i.
Hint: recall that for $z_i \sim N(0,1)$, $\mathbb{E}[z_i|z_i > -c] = \phi(c)/\Phi(c)$ (Inverse-Mills ratio).

(d) Suppose one of your variables in x_i is the cost of public transport as a fraction of the average hourly wage in the city, c_i. Using your result from the previous two exercises, derive the predicted effect of decreasing this ratio by 10 percentage points on air quality y_i^* and measured carbon monoxide concentration y_i.

(e) Instead, suppose you simply use a linear regression to relate y_i to x_i for the cities for whom the concentration was measured precisely, i.e. for cities $i \in \mathcal{U} \equiv \{i : y_i > \delta\}$:

$$y_i = x_i'\gamma + v_i \ , \quad i \in \mathcal{U} \ . \tag{2}$$

What is the effect of decreasing c_i on y_i in this specification? Presuming for a moment that γ and β are the same thing, for which cities is the predicted effect under the linear regression close to/far from the one under the above tobit model?

(f) (Bonus question) You are in fact not interested in relating y_i to x_i, but in relating the true air quality y_i^* – of which y_i is an imperfect measure – to x_i, i.e. you are interested in β, not γ. Supposing that Eq. (1) is the true model generating the data, can you use the OLS estimator for γ from Eq. (2) to consistently estimate β? Under which circumstances will OLS work better/worse?
Hint: For $i \in \mathcal{U}$, we simply have $y_i = y_i^* = x_i'\beta + u_i$. Also, for a generic random variable z_i,

$$\frac{1}{n_u}\sum_{i\in\mathcal{U}} z_i \xrightarrow{p} \mathbb{E}[z_i|y_i > \delta] = \mathbb{E}[z_i|y_i^* > \delta] = \mathbb{E}[z_i|u_i > \delta - x_i'\beta] \ ,$$

where $n_u = |\mathcal{U}|$ *is the number of observations* i *in* $\mathcal{U}$.