Geneva Graduate Institute, Econometrics I

Problem Set 5 Solutions

Francesco Casalena

Fall 2024

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^*?

Solution: The observed outcome variable y_i is a binary variable taking value one when a payment is settled in cash and zero otherwise:

$$y_i = \mathbf{1}\{y_i^* > 0\}$$

The underlying latent (i.e. unobserved) variable y_i^* can be interpreted as the customer's willingness to pay in cash; if it is above a certain threshold (normalized to zero without loss of generality) the payment actually takes place in cash.

(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?

Solution: In the probit model

$$\mathbb{E}[y_i|x_i] = \Phi(x_i'\beta) \ .$$

When the explanatory variable changes from x_1 to x_2 by $\Delta x_i = x_2 - x_1$, the corresponding change in the dependent variable is given by:

$$\mathbb{E}[y_i|x_i = x_2] - \mathbb{E}[y_i|x_i = x_2] = \Phi(x_2'\beta) - \Phi(x_1'\beta) \ .$$

For small changes in the explanatory variable, this effect can be approximated by the first derivative:

$$\frac{\partial \mathbb{E}[y_i|x_i]}{\partial x_{i,c}} = \frac{\partial \Phi(x_i'\beta)}{\partial x_{i,c}} = \phi(x_i'\beta)\beta_c$$

for any covariate $c = 1, 2, ...k$. Therefore, for a discrete change in x_c by 5 units ($\Delta x_c = 5$), the approximate expected change in the dependent variable is:

$$\Delta\mathbb{E}[y_i|x_i] = \Phi(x_2'\beta) - \Phi(x_1'\beta) \approx \Delta x_c \phi(x_i'\beta)\beta = 5 \times \phi(x_i'\beta)\beta_c \ ,$$

whereby x_1 is the vector of regressors that includes the current age of the customer and x_2 is the same vector of regressors (same values), but with age increased by 5 years.

Notice that the marginal effect of a change in the regressor $x_{i,c}$ depends not only on the level of $x_{i,c}$ itself, but also on the level of all other regressors in the vector of explanatory variables x_i.

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

Solution: Yes, in principle OLS could be used to estimate the probability of a purchase being made in cash. The problem is that the predicted values of the dependent variable under OLS, $\hat{y_i} = x_i'\hat{\beta}_{OLS}$ would not be bounded between zero and one. Therefore, they cannot be interpreted as probabilities of a payment being settled in cash.

(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?

Solution: When estimating the OLS model $y_i = x_i'\beta + u_i$, the marginal effect of an increase in x would be:

$$\frac{\partial\mathbb{E}[y_i|x_i]}{\partial x_{i,c}} = \frac{\partial x_i'\beta}{\partial x_{i,c}} = \beta_c$$

For any covariate $c = 1, 2, ...k$. Hence, the marginal effect of a change in the regressor x_c is constant and does not depend on the level of $x_{i,c}$, as the OLS model is linear.

(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.*

Solution: The predicted effects under any model are simply given by the difference in the conditional expectation $\mathbb{E}[y_i|x_i]$ between $x_i = x_1$ and $x_i = x_2$ (e.g. age going up 5 years while all other covariates stay constant). In the probit model, this condictional expectation is a non-linear function of covariates x_i: $\mathbb{E}[y_i|x_i] = \Phi(x_i'\beta)$. By properties of Φ, it goes from 0 to 1, whereby it starts pretty flat for small values of $x_i'\beta$, increases gradually and becomes pretty linear around $x_i'\beta = 0$, after which it starts to flatten out. In contrast, under the linear regression, $\mathbb{E}[y_i|x_i] = x_i'\beta$ is always a linear function of covariates x_i.

As a result, the effects under the two models will be the closest when $\hat{y}_i = x_i'\hat{\beta}$ is around zero, i.e. when the predicted probability of cash-payment is around 50%. Hence, for the customers who are at the margin between cash- and card-payment, the marginal effect estimated with OLS will be closest to the marginal effect estimated under probit. Conversely, for the more extreme values of $x_i'\hat{\beta}$ (i.e. for the more "extreme" customers; the (predictedly) convinced card- or cash-users), the marginal effects estimated under OLS will differ the most from those estimated under probit. Hence, if our research question explores customers "at the margin", we could also use OLS, while to get sensible results for the customers "at the extremes", we should use probit.

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), $\mathbb{P}[y_i = 0|x_i]$.

Solution:

$$\begin{aligned} \mathbb{P}[y_i = 0] &= \mathbb{P}[y_i^* \leq \delta] \\ &= \mathbb{P}[x_i'\beta + u_i \leq \delta] \\ &= \mathbb{P}[u_i \leq \delta - x_i'\beta] \\ &= \Phi\left(\frac{\delta - x_i'\beta}{\sigma}\right) \end{aligned}$$

(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.

Solution:

$$\mathbb{E}[y_i^*|x_i] = \mathbb{E}[x_i'\beta + u_i|x_i] = x_i'\beta$$

(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).

Solution: For simplicity, we omit the conditioning on x_i. We have

$$\begin{aligned}
\mathbb{E}[y_i] &= \mathbb{E}[y_i|y_i^* \leq \delta]\mathbb{P}[y_i^* \leq \delta] + \mathbb{E}[y_i|y_i^* > \delta]\mathbb{P}[y_i^* > \delta] \\
&= \mathbb{E}[y_i|y_i^* > \delta](1 - \mathbb{P}[y_i^* \leq \delta]) \\
&= \mathbb{E}[y_i|u_i > \delta - x_i'\beta](1 - \mathbb{P}[u_i \leq \delta - x_i'\beta]) \\
&= \left(x_i'\beta + \mathbb{E}[u_i|u_i > \delta - x_i'\beta]\right)(1 - \mathbb{P}[u_i \leq \delta - x_i'\beta]) \\
&= \left(x_i'\beta + \mathbb{E}\left[z_i|z_i > \frac{\delta - x_i'\beta}{\sigma}\right]\right)\left(1 - \mathbb{P}\left[z_i \leq \frac{\delta - x_i'\beta}{\sigma}\right]\right) \\
&= \left(x_i'\beta + \sigma\frac{\phi\left(\frac{x_i'\beta - \delta}{\sigma}\right)}{\Phi\left(\frac{x_i'\beta - \delta}{\sigma}\right)}\right)\Phi\left(\frac{x_i'\beta - \delta}{\sigma}\right) \\
&= \Phi\left(\frac{x_i'\beta - \delta}{\sigma}\right)x_i'\beta + \sigma\phi\left(\frac{x_i'\beta - \delta}{\sigma}\right) ,
\end{aligned}$$

where $z_i = u_i/\sigma \sim N(0,1)$.

(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.

Solution: We can compute the predicted change in y_i when we change x_i from x_1 to x_2 by by taking the difference in the conditional expectation function:

$$\mathbb{E}[y_i|x_i = x_2] - \mathbb{E}[y_i|x_i = x_1] = \Phi\left(\frac{x_i'\beta - \delta}{\sigma}\right)x_i'\beta + \sigma\phi\left(\frac{x_i'\beta - \delta}{\sigma}\right) .$$

In our case, x_1 includes c_i and x_2 includes $c_i - 10$ (or $c_i - 0.1$ if c_i is measured as a decimal rather than in percent).

Sidenote: for small changes $\Delta x_i = x_2 - x_1$, the above is approximately equal to Δx_i (a vector of dimension k) multiplied by the first derivative of $\mathbb{E}[y_i|x_i]$ w.r.t. x_i (also a vector of dimension k),

$$\frac{\partial \mathbb{E}[y_i|x_i]}{\partial x_i} = \frac{\beta}{\sigma}\phi\left(\frac{x_i'\beta - \delta}{\sigma}\right)x_i'\beta + \beta\Phi\left(\frac{x_i'\beta - \delta}{\sigma}\right) + \beta\phi'\left(\frac{x_i'\beta - \delta}{\sigma}\right) .$$

If we only change one of the covariates $c = 1, 2, ..., k$ in x_i, we get the marginal effect

$$\Delta\mathbb{E}[y_i|x_i] \approx \Delta x_{ic} \times \left(\frac{\beta_c}{\sigma}\phi\left(\frac{x_i'\beta - \delta}{\sigma}\right)x_i'\beta + \Phi\left(\frac{x_i'\beta - \delta}{\sigma}\right)\beta_c + \beta_c\phi'\left(\frac{x_i'\beta - \delta}{\sigma}\right)\right) ,$$

whereby $\Delta x_{ic} = -10$ (or -0.1) in our case.

Notice that this marginal effect depends on the level of $x_{i,c}$, as well as on the levels of all the other elements of the vector of covariates x_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?

Solution: In the case of a linear regression, we have

$$\mathbb{E}[y_i|x_i = x_2] - \mathbb{E}[y_i|x_i = x_1] = x_2'\gamma - x_1'\gamma = \Delta x_i'\gamma \ .$$

When we change a single regressor c, we get $\gamma_c \Delta x_{ic}$. (Sidenote: since the conditional mean is linear in x_i, the exact partial effect and the approximate one that uses first derivatives coincide.)

For ease of exposition, we compare this to the approximate marginal effect under Tobit (rather than the exact one),

$$\frac{\partial \mathbb{E}[y_i|x_i]}{\partial x_{i,c}} = \beta_c \left(\frac{1}{\sigma}\phi\left(\frac{x_i'\beta - \delta}{\sigma}\right) x_i'\beta + \Phi\left(\frac{x_i'\beta - \delta}{\sigma}\right) + \phi'\left(\frac{x_i'\beta - \delta}{\sigma}\right) \right) \ .$$

Assuming that $\beta = \gamma$, then the two marginal effects will be closer when the term in brackets is close to one. By recalling the properties of Normal pdf's and cdf's, we know that:

- $\phi(z) \in (0, 1)$ for $z \in \mathbb{R}$ and that $\phi(z) \to 0$ for $z \to \pm\infty$
- $\phi'(z) \to 0$ for $z \to \pm\infty$
- $\Phi(z) \to 1$ as $z \to +\infty$ and that $\Phi(z) \to 0$ as $z \to -\infty$

By putting the three pieces together, we know that the expression in brackets will be closer to 1 for the highest values of the distribution of x_i and it will be closer to zero for the lowest values of x_i. Therefore, we know that the marginal effects estimated under OLS will be the closest to those estimated under tobit for the most polluted cities (i.e. the cities farthest away from the censoring point of δ) and they will be the most distant from those estimated under tobit for the least polluted cities.

(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 $\mathcal{U}$.

Solution: The OLS estimator is given by

$$\begin{aligned}
\hat{\gamma} &= \left[\frac{1}{n_u}\sum_{i\in\mathcal{U}} x_i x_i'\right]^{-1} \frac{1}{n_u}\sum_{i\in\mathcal{U}} x_i y_i \\
&= \left[\frac{1}{n_u}\sum_{i\in\mathcal{U}} x_i x_i'\right]^{-1} \frac{1}{n_u}\sum_{i\in\mathcal{U}} x_i (x_i'\beta + u_i) \\
&= \beta + \left[\frac{1}{n_u}\sum_{i\in\mathcal{U}} x_i x_i'\right]^{-1} \frac{1}{n_u}\sum_{i\in\mathcal{U}} x_i u_i \\
&= \beta + \left[\frac{1}{n}\sum_{i} x_i x_i' \, \mathbf{1}\left\{y_i^* > \delta\right\}\right]^{-1} \frac{1}{n}\sum_{i} x_i u_i \, \mathbf{1}\left\{y_i^* > \delta\right\} \\
&\xrightarrow{p} \beta + \mathbb{E}[x_i x_i|y_i^* > \delta]^{-1}\mathbb{E}[x_i u_i|y_i^* > \delta] \\
&= \beta + \mathbb{E}[x_i x_i|u_i > \delta - x_i'\beta]^{-1}\mathbb{E}[x_i\mathbb{E}[u_i|u_i > \delta - x_i'\beta, x_i]] \\
&= \beta + \mathbb{E}[x_i x_i|u_i > \delta - x_i'\beta]^{-1}\mathbb{E}\left(x_i\sigma\frac{\phi\left(\frac{x_i'\beta-\delta}{\sigma}\right)}{\Phi\left(\frac{x_i'\beta-\delta}{\sigma}\right)}\right) \\
&\neq \beta \ .
\end{aligned}$$

The OLS estimator is not consistent. By the LIE, the asymptotic bias equals

$$\begin{aligned}
\mathbb{E}[x_i x_i|y_i^* > \delta]^{-1}\mathbb{E}[x_i u_i|y_i^* > \delta] &= \mathbb{E}[x_i x_i|u_i > \delta - x_i'\beta]^{-1}\mathbb{E}[x_i\mathbb{E}[u_i|u_i > \delta - x_i'\beta, x_i]] \\
&= \mathbb{E}[x_i x_i|u_i > \delta - x_i'\beta]^{-1}\mathbb{E}\left(x_i\sigma\frac{\phi\left(\frac{x_i'\beta-\delta}{\sigma}\right)}{\Phi\left(\frac{x_i'\beta-\delta}{\sigma}\right)}\right) \ .
\end{aligned}$$

Its size depends on the magnitude of the ratio $\phi\left(\frac{x_i'\beta-\delta}{\sigma}\right)/\Phi\left(\frac{x_i'\beta-\delta}{\sigma}\right)$. By the properties of ϕ and Φ (see solution to previous exercise), the asymptotic bias is expected to be high (low) if the "average" $x_i'\beta$ is close to (far above of) the censoring point δ as then $x'\beta - \delta$ is close to (far above of) zero and, therefore, $\phi(x'\beta - \delta)$ is high (low). In our application, we get a low bias if we have lots of polluted cities in our sample. If the average $x_i'\beta$ is far below δ, it's not immediately clear what happens, as not only the numerator $\phi(x'\beta - \delta)$ is close to zero, but also the denominator, $\Phi(x'\beta - \delta)$.