Geneva Graduate Institute (IHEID)

Econometrics I (EI035), Fall 2024

Marko Mlikota

Problem Set 6

Due: Saturday, 7 December, 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

The dataset `dat_SalesCustomers.csv` contains data on sales of shopping malls in Instanbul. It includes the following variables: *invoice_no* (identifier of transaction/invoice), *customer_id* (identifier of customer), *category* (type of goods sold), *price* (in TRY, Turkish Lira), *invoice_date*, *shopping_mall*, *gender*, *age* and *payment_method* (cash- vs. credit-card- vs. debit-card-payment).

You are interested in shedding light on the determinants of cash- vs card-payment. For this purpose, you set up a probit model:

$$y_i^* = x_i'\beta + u_i \quad , \quad u_i|x_i \sim N(0,1) \; , \tag{1}$$

whereby we observe $y_i = \mathbf{1}\{y_i^* \geqslant 0\}$, a dummy variable for cash payment. Recall that the Maximum Likelihood (ML) estimator for β solves

$$\hat{\beta} = \arg\min_{\beta} Q_n(\beta; Z_n) \quad \text{for} \quad Q_n(\beta; Z_n) = -\frac{1}{n}\ell(\beta; Z_n) \; , \tag{2}$$

where

$$\ell(\beta; Z_n) = \sum_{i=1}^{n} y_i \log(\Phi(x_i'\beta)) + (1 - y_i)\log(\Phi(-x_i'\beta))$$

is the log-likelihood and $Z_n = \{y_i, x_i\}_{i=1}^n$ comprises all of the data you have available (outcome-variables and covariates for the n observations in your sample).

(a) Are there missing values in your data? Delete all observations with a missing value in the variables *category*, *price*, *gender*, *age* or *payment_method*. How many observations do you have left?

(b) Based on the variable *payment_method*, generate a dummy variable for cash payment and call it *paid_in_cash*. Also, based on *gender*, create a dummy for males, *male*. What fraction of transactions were carried out in cash? What fraction of the overall sales (in TRY) were carried out in cash?

(c) Based on the variable *category*, create a dummy for each category of goods sold. How are the transactions split across goods categories? How are the sales split across goods categories?

(d) Taking *paid_in_cash* as your outcome variable y_i and *price*, *male*, *age* and all category-dummies but one as your covariates x_i, use a numerical optimization-command from the software of your choice to solve the optimization problem in Eq. (2) and obtain $\hat{\beta}$ for your sample.[1]

[1] As part of your derivations for exercise (f), you have to find the score and the Hessian of the objective function in Eq. (2),

$$s_n(\beta) \equiv Q_n^{(1)}(\beta; Z_n) \equiv \frac{\partial Q_n(\beta; Z_n)}{\partial \beta} \quad \text{and} \quad H_n(\beta) \equiv Q_n^{(2)}(\beta; Z_n) \equiv \frac{\partial^2 Q_n(\beta; Z_n)}{\partial \beta \partial \beta'} = \frac{\partial s_n(\beta)}{\partial \beta'} \; .$$

Hint: instead of computing first $\Phi(x)$ using a software-command for the cdf of a $N(0,1)$ RV (`pnorm(x)` in R) and then taking logs, it's better to directly use a software-command for the log of the cdf of a $N(0,1)$ RV (`pnorm(x,log.p=TRUE)` in R). This way, you avoid having to compute the log of a number very close to zero, which can result in `-Inf`.[2]

(e) Based on your estimate, compute the effect of age increasing by 5 years on the expected probability of using cash for a 30 year-old male who bought clothes for 500 TRY, i.e. for an observation with $x_i = x_i^* \equiv [500, 1, 30, 0, ..., 0, 1, 0, ..., 0]$.[3] We will call this quantity $\gamma_1(\hat{\beta})$. Also, compute the same effect without conditioning on the category of goods sold in two steps: (i) compute the effect for each category and (ii) take a weighted average of them, with weights given by the proportions of these goods-categories in overall sales (see your answer to (c)). We will call this quantity $\gamma_2(\hat{\beta})$.

(f) Suppose that your probit model in Eq. (1) is correctly specified. Is your estimator $\hat{\beta}$ consistent? Use the simplified version of the extremum estimation theory we discussed in class to answer this question.

(g) Use bootstrapping to find a numerical approximation of the finite sample distribution of $\hat{\beta}$ as well as the two marginal effects $\gamma_1(\hat{\beta})$ and $\gamma_2(\hat{\beta})$: draw $M = 100$ different samples of n observations with replacement from your dataset and compute (numerically) $\hat{\beta}$, $\gamma_1(\hat{\beta})$ and $\gamma_2(\hat{\beta})$ for each of them. Plot the finite sample distributions you obtained (regarding $\hat{\beta}$, you can limit yourself to the coefficient on *age*).

(h) Another approach to approximate the finite sample distribution of $\hat{\beta}$ and functions of it like the marginal effects is to use their asymptotic distribution. Use the simplified version of the extremum estimation theory we discussed in class to show that the asymptotic distribution of $\hat{\beta}$ is given by

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N\left(0, H^{-1}\right) \quad \text{with} \quad H = \mathbb{E}\left[\frac{\phi(x_i'\beta_0)^2}{\Phi(x_i'\beta_0)\Phi(-x_i'\beta_0)} x_i x_i'\right] . \tag{3}$$

Then, use the asymptotic distribution in Eq. (3) to approximate the finite sample distribution of $\hat{\beta}$ in your sample. How does this approximate finite sample distribution of the estimated coefficient on *age* compare to the one obtained via bootstrapping?

Hint: The numerator and the denominator in the fraction that appears in H are often both very close to zero. Rather than computing it as-is, first compute the log of it and then take the exponential, i.e. compute

$$\frac{\phi(x_i'\beta_0)^2}{\Phi(x_i'\beta_0)\Phi(-x_i'\beta_0)} \quad as \quad exp\left\{2log\phi(x_i'\beta_0) - log\Phi(x_i'\beta_0) - log\Phi(-x_i'\beta_0)\right\} .$$

You can also use them to construct your own numerical optimization algorithm to find $\hat{\beta}$.

[2] The alternative is to do manual adjustments, coding -Inf as a very large negative number, but this can be imprecise.

[3] The zeros after 30 are for all category-dummies except the dummy for clothing, which contains a one

To compute log $\phi(x)$ and log $\Phi(x)$, as before in exercise (b), it's better practice to use the log-pdf/cdf software-commands than to compute first the pdf/cdf and then take logs manually (i.e. in R, use `dnorm(x,log=TRUE)` *and* `pnorm(x,log.p=TRUE)`*).*

(i) Use the asymptotic distribution of $\hat{\beta}$ from Eq. (3) and the Delta method to find the asymptotic distribution of $\gamma_1(\hat{\beta})$. Then, use it to approximate the finite sample distribution of $\gamma_1(\hat{\beta})$ in your sample. How does this approximate finite sample distribution compare to the one obtained via bootstrapping?

(j) Now let's test whether the true partial effect $\gamma_1(\hat{\beta})$ (i.e. the true change in the expected probability of cash payment for a 30 year-old male buying clothes for 500 TRY when this individual becomes 5 years older) is significantly different from 0 at the $\alpha = 0.05$ level:

$$\mathcal{H}_0 : \gamma_1(\beta) = 0 \quad \text{vs.} \quad \mathcal{H}_1 : \gamma_1(\beta) \neq 0 \, .$$

(In other words, we are testing whether the expected probabilities of cash payment for a 30 year-old and a 35 year-old male buying clothes for 500 TRY are different.) One approach to do so uses the finite sample distribution of $\gamma_1(\hat{\beta})$ approximated via its asymptotic distribution, which you found in the exercise before:

$$\gamma_1(\hat{\beta}) \overset{approx.}{\sim} N\left(\gamma_1(\beta), \frac{1}{n}\hat{V}\right) \, ,$$

for some $\hat{V}$ you had to find. Use this expression to construct a t-test. What do you conclude? Also, use the above expression to construct a 95% confidence interval for $\gamma_1(\beta)$.[4] (If you couldn't find $\hat{V}$, just state the test statistic and critical value for a general $\hat{V}$.)

[4]Note that in general, we would use the Wald-test. Here we can use the t-test because we are testing a single thing, i.e. our testing function $g(\beta) = \gamma_1(\beta) = 0$ is a scalar. Our t-test will give the same result as the Wald test, because the asymptotic distribution of the Wald-test-statistic is derived in the same way as that of our t-test statistic here (i.e. it also uses the Delta method), except that it squares things in the end to go from a Normal to a Chi-Squared distribution.