PS5 Solutions

Jingle Fu

Problem 1

The dataset `dat_SalesCustomers.csv` contains data on sales of shopping malls in Istanbul. It includes the following variables: *invoice no* (identifier of transaction or 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:

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

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

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

where

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

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)

To decrease computational costs, consider only the first $n = 1000$ observations for the following questions. Based on the variable category, create a dummy for each of the following four categories: i) clothes and shoes, ii) cosmetics, iii) food, iv) technology. In this way, we divide the categories into five groups, whereby the fifth is made up by the rest, i.e. goods that do not belong to either of the four categories. How are the transactions split across these five categories? How are the sales split across these five 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. If manual optimization does not work, you can use a pre-programmed command to estimate the probit model.

(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 $x_i = x_i^* \equiv [500, 1, 30, 0, \ldots, 0, 1, 0, \ldots, 0]$. Put differently, this is the difference in expected probabilities of cash payment between a 60 year-old and a 30 year-old male who bought clothes/shoes for 500 TRY. 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 of the five categories 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

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N\left(0, H^{-1}\right),$$

with

$$H = E\left[\frac{\phi(x_i'\beta_0)^2}{\Phi(x_i'\beta_0)\Phi(-x_i'\beta_0)} x_i x_i'\right]. \tag{3}$$

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

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

(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:

$$H_0 : \gamma_1(\beta) = 0 \quad \text{vs.} \quad H_1 : \gamma_1(\beta) \neq 0.$$

(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,

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

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 $\gamma_1(\beta)$. (If you couldn't find $\hat{V}$, just state the test statistic and critical value for a general $\hat{V}$.)

Solution (a).

Yes, here missing values in the data set. The initial number of observations is 99457, The dataset's reported missing values indicate that only *age* has missing observations (119 missing values). After dropping these, we end up with 99338 observations.

Solution (b).

Define

$$\text{paid\_in\_cash}_i = \mathbf{1}\{\text{payment\_method}_i = \text{Cash}\}$$

$$\text{male}_i = \mathbf{1}\{\text{gender}_i = \text{Male}\}$$

The fraction of transactions carried out in cash is

$$\frac{1}{n}\sum_{i=1}^{n}\text{paid\_in\_cash}_i.$$

Empirically, this is about 44.69%.

The fraction of overall sales carried out in cash is

$$\frac{\sum_{i=1}^{n}\text{paid\_in\_cash}_i \cdot price_i}{\sum_{i=1}^{n} price_i}.$$

Empirically, this fraction is about 44.79%.

These results indicate that cash payments represent nearly half of all transactions and sales value.

Solution (c).

We now consider only the first $n = 1000$ observations. Let the categories be divided into five mutually exclusive groups: Clothes and Shoes (C), Cosmetics (Cos), Food (F), Technology (T), and Other (O). Define indicator variables:

$$\begin{aligned}
d_{C,i} &= \mathbf{1}\{\text{category}_i = \text{Clothes and Shoes}\} \\
d_{Cos,i} &= \mathbf{1}\{\text{category}_i = \text{Cosmetics}\} \\
d_{F,i} &= \mathbf{1}\{\text{category}_i = \text{Food}\} \\
d_{T,i} &= \mathbf{1}\{\text{category}_i = \text{Technology}\} \\
d_{O,i} &= 1 - (d_{C,i} + d_{Cos,i} + d_{F,i} + d_{T,i}).
\end{aligned}$$

The fraction of transactions in category j is

$$\frac{1}{1000}\sum_{i=1}^{1000} d_{j,i}.$$

The fraction of sales in category j is

$$\frac{\sum_{i=1}^{1000} d_{j,i} \cdot price_i}{\sum_{i=1}^{1000} price_i}.$$

Empirically:

- Transactions fraction: Clothes/Shoes: 43.8%, Cosmetics: 14.8%, Food: 14.0%,

Technology: 5.0%, Other: 22.4%.

- Sales fraction: Clothes/Shoes: 70.58%, Cosmetics: 2.72%, Food: 0.32%, Technology: 23.9%, Other: 2.49%.

The result shows that most transactions and sales are in the Clothes/Shoes category. Technology, though having the lowest transaction fraction, has the second-highest sales fraction, meaning that it has the highest average price.

Solution (d).

To find the Maximum Likelihood Estimator (MLE) $\hat{\beta}$, we differentiate the log-likelihood with respect to β. Let $\phi(\cdot)$ denote the standard normal PDF. We use:

$$\frac{d}{dt}\log(\Phi(t)) = \frac{\phi(t)}{\Phi(t)}, \quad \text{and} \quad \frac{d}{dt}\log(1-\Phi(t)) = -\frac{\phi(t)}{1-\Phi(t)}.$$

For each element β_j of β, the derivative of the log-likelihood is:

$$\frac{\partial \ell(\beta; Z_n)}{\partial \beta_j} = \sum_{i=1}^{n} \left[y_i \frac{\phi(x_i'\beta)}{\Phi(x_i'\beta)} - (1-y_i)\frac{\phi(x_i'\beta)}{1-\Phi(x_i'\beta)} \right] x_{ij}.$$

Stacking all partial derivatives together, the gradient (score vector) is:

$$\nabla_\beta \ell(\beta; Z_n) = \sum_{i=1}^{n} \left[\frac{y_i - \Phi(x_i'\beta)}{\Phi(x_i'\beta)(1-\Phi(x_i'\beta))}\phi(x_i'\beta) \right] x_i.$$

Often written more simply as:

$$\nabla_\beta \ell(\beta; Z_n) = \sum_{i=1}^{n} \left[y_i \frac{\phi(x_i'\beta)}{\Phi(x_i'\beta)} - (1-y_i)\frac{\phi(x_i'\beta)}{1-\Phi(x_i'\beta)} \right] x_i.$$

Step 1: Characterizing the MLE $\hat{\beta}$ The MLE $\hat{\beta}$ sets the gradient to zero:

$$\nabla_\beta \ell(\hat{\beta}; Z_n) = 0.$$

Substituting back:

$$\sum_{i=1}^{n} \left[y_i \frac{\phi(x_i'\hat{\beta})}{\Phi(x_i'\hat{\beta})} - (1-y_i)\frac{\phi(x_i'\hat{\beta})}{1-\Phi(x_i'\hat{\beta})} \right] x_i = 0.$$

This is a system of k nonlinear equations in the k unknowns $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_k)'$. **Step 2: No Closed-Form Solution**

Unlike in linear regression or the logit model (even the logit doesn't have a closed form), the Probit model does not admit a closed-form solution for $\hat{\beta}$. The equation above must be solved using numerical optimization techniques such as the Newton-Raphson algorithm or other iterative methods.

Step 3: Numerical Optimization

A common iterative procedure is:

Algorithm 1: Newton-Raphson Method

Input: Initialize β_0, tolerence level $\varepsilon > 0$

for $m = 1$ **to** M **do**
  Given β^m, compute $\nabla_\beta \ell(\beta^{(m)}; Z_n)$ and $[H(\beta^{(m)}; Z_n)]$;
  Set $\beta^{(m+1)} = \beta^{(m)} - [H(\beta^{(m)}; Z_n)]^{-1} \nabla_\beta \ell(\beta^{(m)}; Z_n),$;
  if $\|\beta^{m+1} - \beta^m\| < \varepsilon$ **then**
    $\hat{\beta} = \beta^{m+1}$;
  else
    Proceed to the next iteration;
  end
end

where $H(\beta; Z_n)$ is the Hessian matrix of second derivatives evaluated at β. Convergence is achieved when changes in β or the norm of the gradient are below a given tolerance.

The regression result is as follows:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_{price} \\ \hat{\beta}_{male} \\ \hat{\beta}_{age} \\ \hat{\beta}_{cosmetics} \\ \hat{\beta}_{food} \\ \hat{\beta}_{technology} \end{bmatrix} = \begin{bmatrix} 0.0682 \\ 0.000112 \\ -0.0502 \\ -0.00183 \\ -0.2879 \\ -0.1195 \\ 0.0640 \\ -0.4195 \end{bmatrix}.$$

Interpretation:

- The coefficient on price is positive but very small, suggesting a tiny positive association of price with the probability of cash payment (not statistically significant).
- male is negative, but not significant, suggesting no strong gender effect on the probability of cash usage.
- age coefficient is negative and small, not statistically significant either.
- Some category dummies (like Clothes/Shoes) are significantly different from zero, indicating that the reference category (likely"Other") differs in payment method probability.

Solution (e).

γ_1 (effect of age increasing by 30 years): -0.02096

γ_2 (weighted effect over categories): -0.020774

Solution (f).

Solution (g).

Solution (h).

Solution (i).

Solution (j).

t-statistic: -0.68,

We cannot reject the null hypothesis at the 5% significance level.

So, we conclude that the expected probabilities of cash payment for a 30 year-old and a 60 year-old male buying clothes for 500 TRY are not significantly different.

95% Confidence Interval for $\gamma_1(\beta)$: -0.0815 to 0.0396

Table 1: Optimization model

	Dependent variable:
	paid_in_cash
price	0.0001
	(0.0001)
male	−0.050
	(0.081)
age	−0.002
	(0.003)
clothes_shoes	−0.288**
	(0.130)
cosmetics	−0.120
	(0.133)
food	0.064
	(0.135)
technology	−0.420
	(0.314)
Constant	0.068
	(0.148)
Observations	1,000
Log Likelihood	−685.217
Akaike Inf. Crit.	1,386.434
Note:	*p<0.1; **p<0.05; ***p<0.01