Econometrics 3 PPD: Maximum Likelihood and Binary outcomes

Problem Set 1

Send one single solution per group by the 27th of January, 10PM, to anubhav.agarwal@sciencespo.fr.
Exercises labeled with (*) are optional

1 CMLE and OLS

Let $x_i = (x_{i1}, \ldots, x_{iK})'$ be a $(K \times 1)$-dimensional vector of regressors. Let y_i be a (scalar) continuous dependent variable, such that

$$y_i = x_i'\beta + u_i,$$

where $\beta = (\beta_1, \ldots, \beta_K)'$ and where

$$u_i | x_i \sim N(0, \sigma^2).$$

1. What is the average marginal effect of x_{i1}, the first component of x_i, on y_i?
2. Given the above model, does OLS provide a consistent estimator of β? Why or why not?
3. Write down the (conditional) log-likelihood function for a sample of size n.
4. Based on the (conditional) log-likelihood function derived in (3), derive the (conditional) Maximum Likelihood estimator of β and σ^2. How does it compare to the OLS estimator?

2 Anonymous resumes

The paper "Unintended Effects of Anonymous Résumés" and its online appendix are on the course webpage. Read the introduction and Sections I, II and IIIA.

Balancing of candidates' characteristics

1. Given the experimental design, do you expect the contents of the résumés of minority candidates in the experimental sample to differ statistically from those of the majority candidates? Why does this make it quite different from correspondence studies that are used to test for discrimination?

2. Do you expect candidates in the treatment group (with anonymous résumés) to be statistically different from candidates in the control group (with standard résumés)?

3. Table A-4 runs a balancing test of treatment and control candidates. Are there statistically different characteristics? What is the limitation of testing balance characteristic by characteristic?

4. As an alternative, you write a logit model relating the probability that a candidate is treated to the same covariates $X_1, ..., X_K$ as in Table A-4:

$$\Pr(Y = 1|X) = \Lambda(\beta_0 + X\beta)$$

where X is the row vector $(X_1...X_K)$ and β if the column vector such that $\beta' = (\beta_1...\beta_K)$. What is the null hypothesis that you need to test?

5. Perform that test including in X the variables from the candidates' survey only. Can you reject balance? Compare with the same test using a probit model. Hint: Perform a likelihood ratio test using the lmtest package on R.

6. Perform that test including in X the variables from the candidates' survey and the variables from coding the résumés. Can you reject balance? Compare with the same test using a probit model. Hint: same as the previous question.

7. Why is footnote 25 important?

Impact of anonymization on interviews

1. In Table 4 panel A, explain in what sense the estimate $-0.024(0.031)$ has no causal interpretation, and why $0.046^{**}(0.020)$ does.

2. δ in equation (1) can be estimated as the coefficient on $D_i \times An_j$ in the OLS regression of Y_{ij} on D_i, An_j and $D_i \times An_j$, where D_i is equal to 1 if the candidate belongs to the minority. Run that regression and compare the results to Table 4 (*hint*: use the survey sampling weights POIDS_SEL. You may need to use the svyglm() function of the 'survey' package on R. Note that the 'weights' argument within the lm() function is meant for precision weights as opposed to survey weights. To understand the difference between the two, you may refer to the following link: `https://notstatschat.rbind.io/2020/08/04/weights-in-statistics/`).

3. Noting that Y_{ij} is binary, you decide to estimate a probit model instead (dropping indexes for convenience):

$$Y^* = \alpha_0 + \alpha_1 D + \alpha_2 An + \alpha_3 D \times An + \varepsilon,$$

$$Y = \begin{cases} 1 \text{ if } Y^* > 0 \\ 0 \text{ if } Y^* \leq 0 \end{cases} .$$

and

$$\varepsilon | D, An \sim N(0,1).$$

You interpret this model as a "structural" model: Y^* is the the latent net profit expected by the firm from a candidate. What are the strong assumptions imposed by this model?

4. Write down the causal effect of anonymization on minority candidates' interview rate (resp. majority candidates) as a function of $\alpha_0, \alpha_1, \alpha_2$ and α_3.

5. Write down the causal effect of anonymization on the interview gap between minority and majority candidates (δ) as a function of $\alpha_0, \alpha_1, \alpha_2$ and α_3. Estimate it (*hint*: use the survey sampling weights POIDS_SEL and again use the 'survey' package on R). Compare to the OLS model and conclude.

3 Normalization of variance in probit (*)

The probit can be motivated by a latent variable model, where

$$y_i^* = x_i'\beta + u_i$$

and

$$y_i = \begin{cases} 1 \text{ if } y_i^* > 0 \\ 0 \text{ if } y_i^* \leq 0 \end{cases} .$$

The probit is obtained by assuming that $u_i | x_i \sim N(0,1)$. Now, assume that $u_i | x_i \sim N(0, 1/4)$ (i.e., $\text{Var}(u_i | x_i) = 1/4$). How will this affect your estimator of β? How will this affect your estimator of marginal effects? What do you conclude with respect to the "standard" choice $u_i | x_i \sim N(0,1)$?

4 Structural model of charitable contributions (*)

Family i chooses annual consumption c_i and charitable contributions q_i to solve the problem

$$\begin{cases} \max_{c,q} \; c + a_i \log(1+q) \\ s.t. \; c + p_i q \leq m_i \\ q \geq 0 \end{cases}$$

where m_i is income of family i, p_i is the price of one dollar of charitable contributions – where $p_i < 1$ because of the tax deductability of charitable contributions, and this price differs across families because of different marginal tax rates and different state tax codes – and $a_i \geq 0$ determines the marginal utility of charitable contributions. Take m_i and a_i as exogenous to the family in this problem.

1. Show that the optimal solution is $q_i = 0$ if $a_i \leq p_i$, and $q_i = a_i/p_i - 1$ if $a_i > p_i$.

2. Define $y_i = 1$ if the family makes charitable contributions, 0 otherwise. Suppose that $a_i = \exp(z_i\gamma + v_i)$, where z_i is a vector of observable family traits and v_i is unobservable. Note $\sigma^2 = Var(v_i)$ the variance of v_i. Assume that v_i/σ is independent of (z_i, m_i, p_i) and has symmetric distribution with c.d.f. $G(.)$. Show that

$$P(y_i = 1|z_i, m_i, p_i) = G[(z_i\gamma - \log p_i)/\sigma].$$

3. Which estimator would you use to estimate γ and σ? Write the (conditional) log-likelihood. Which parametric assumptions have you made?

4. Where does the assumption that v_i/σ is independent of (z_i, m_i, p_i) matter?

5. Assume that family traits in z_i are the age of the household head and the number of kids. Do you think that your estimator yields a valid estimate of the causal impact of family size on charitable contributions? Why?