Empirical Project + Theoretical Exercise

Econometric Methods for Empirical Economics, 2021-2022

## 1 Instructions

For Exercises 1-2: Prepare: (1) pdf file containing the answers to the questions below; (2) Stata/R do file. Save the two files in a zip folder named with name and surname of each group member (max 2 people). Email the zip folder as an attachment to vishal.kamat@tse-fr.eu by December 17 at 9pm.

For Exercise 3: Prepare a pdf file containing the answer with your name and surname (**cannot work in groups for this exercise**). Email pdf to vishal.kamat@tse-fr.eu by December 17 at 9pm.

## 2 Exercise 1: Linear Probability Model, Probit Model, Logit Model (10 points)

(1) Download the data bwght.dta from *here* and import the dataset.

(2) Create a dummy variable named *smokes* equal to 1 if a woman smokes during pregnancy and zero otherwise. (0.25 points)

(3) Study the dataset

(a) Briefly describe each variable (remember to indicate units of measure). Can you tell what we could study using this dataset? (0.25 points)

(b) Compute the main descriptive statistics of each variable. (0.25 points)

(c) Find the number of smoking women. (0.25 points)

(d) Find the number of white smoking women. (0.25 points)

(e) Find the number of smoking women with family income above the sample average. (0.25 points)

(f) Find the number of smoking women with family income above the sample median. (0.25 points)

(g) Find the number of observations with at least one missing value among

$$smokes, motheduc, white, lfaminc$$

(0.25 points)

(4) Linear Probability Model

  (a) Estimate the impact of $motheduc, white, lfaminc$ on $smokes$ using the Linear Probability Model. (0.25 points)

  (b) Which coefficients are statistically significant at 5%? (0.25 points)

  (c) Interpret the coefficients of $motheduc, white, lfaminc$. (0.25 points)

  (d) Compute the proportion of fitted probabilities outside the unit interval. (0.25 points)

(5) Logit Model

  (a) Estimate the impact of $motheduc, white, lfaminc$ on $smokes$ using the Logit Model. (0.25 points)

  (b) Compute the estimated marginal effects of $motheduc, white, lfaminc$ on the probability of smoking. If needed, use the sample average of the covariates to evaluate the estimated marginal effects. (0.25 points)

  (c) Discuss differences (sign, magnitude) between the estimated marginal effects from the Linear Probability Model and the estimated marginal effects from the Logit Model. (0.25 points)

  (d) Compare the results from the Linear Probability Model with the results from the Logit Model in terms of proportions of correct predictions. (0.25 point)

  (e) Are your results (estimated coefficient and marginal effects) from the Logit Model in line with the existing empirical literature on the same (or similar) topic? [max 1 page of discussion] (4 points)

  - Look for papers published in top economic journals discussing the determinants of smoking behaviour. Remember to add a "References" section where you provide full references of the cited papers.
  - Compare your results with the results in the papers you found.
  - Discuss differences (if any) between your methodology and the methodologies used in the papers you found.
  - Which important variables do you think your estimation does not take into account? Are peer effects relevant for smoking behaviour?

(f) At $lfaminc$ evaluated at the sample average, what is the estimated difference between the probability of smoking for a black woman with 16 years of education and the probability of smoking for a black woman with 12 years of education? (0.25 point)

(g) Do you think that $lfaminc$ is exogenous in the smoking equation? What about $motheduc$? Why yes or why not? (0.5 points)

(h) Test the null hypothesis that $lfaminc$ is exogenous. If you reject the null hypothesis, explain in words how you would address such endogeneity issue. (1 point)

(6) Repeat (5) (a) using the Probit Model in place of the Logit Model. Are the estimated coefficients and marginal effects from the Logit and Probit Models noticeably different? Why? (0.25 point)

## 3 Exercise 2: Tobit Model (10 points)

(1) Download the data fringe.dta from *here* and import the dataset in Stata.

(2) Keep the variables

$$hrbens, exper, age, educ, tenure, married, male, white, nrtheast, nrthcen, south, union$$

(0.25 points)

(3) Study the dataset

(a) Briefly describe each variable (remember to indicate units of measure). Can you tell what we could study using this dataset? (you should interpret $hrbens$ as hourly dollar value of fringe benefits, e.g., medical and dental insurance, use of a company car, housing allowance, educational assistance, etc.) (0.25 points)

(b) Compute the descriptive statistics of each variable. (0.25 points)

(c) Find the number of women. (0.25 points)

(d) Find the number of married women with $tenure$ above the the sample average. (0.25 points)

(e) Find the mean value of $hrbens$ for women and for men. (0.25 points)

(4) Estimate the impact of

$$exper, age, educ, tenure, married, male, white, nrtheast, nrthcen, south, union$$

on $hrbens$ using the Linear Regression Model. Which coefficients are statistically significant at 5%? Interpret the estimated coefficients of $exper$, $age$, $educ$, $tenure$, $married$, $male$, $white$, $nrtheast$, $nrthcen$, $south$, $union$ from the Linear Probability Model. (0.25 points)

(5) Can we classify $hrbens$ as a corner solution dependent variable? Can we classify $hrbens$ as a censored dependent variable? Why? (1 point)

(6) Estimate the impact of

$$exper, age, educ, tenure, married, male, white, nrtheast, nrthcen, south, union$$

on $hrbens$ using the Tobit Model. (0.25 points)

(7) Why are the results from the Tobit Model so similar to the results from the Linear Regression Model? (0.5 points)

(8) Add the square of both experience and tenure to the Tobit Model from the previous part and re-run the estimation. Do you think these squared terms should be included? Why? How do the results change? (1 point)

(9) What is the estimated marginal impact of one more year of tenure on $\mathbb{E}(hrbens|hrbens > 0, X = \bar{x}_n)$ using the Tobit Model? (careful: you have to tell Stata that $tenuresq$ is the square of $tenure$ so that the marginal effects computed using the command $margin$ are correct). (0.5 point)

(10) Test the null hypothesis that $educ$ is exogenous. If you reject the null hypothesis, explain in words how you would address such endogeneity issue. (1 point)

(11) Are your results (estimated coefficient and marginal effects) from the Tobit Model about the impact of gender on fringe benefits in line with the existing empirical literature on the same (or similar) topic? [max 1 page of discussion] (4 points)

## 4 Exercise 3: Theoretical Exercise (15 points)

A latent variable $y_{1i}^*$ determines whether $y_{1i}$ and $y_{2i}$ are observed or censored at 0. In particular, suppose that we have

$$y_{1i} = y_{1i}^* 1\{y_{1i}^* > 0\} \ , \tag{1}$$

$$y_{2i} = y_{2i}^* 1\{y_{1i}^* \leq 0\} \ , \tag{2}$$

and that

$$y_{1i}^* = x_{1i}'\beta_1 + \epsilon_{1i} \ ,$$

$$y_{2i}^* = x_{2i}'\beta_2 + \epsilon_{2i} \ ,$$

where $(y_{1i}, y_{2i}, x_{1i}, x_{2i})$ is observed. Moreover, assume that $(\epsilon_{1i}, \epsilon_{2i}) \perp (x_{1i}, x_{2i})$ and that

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right).$$

Our objective is to estimate the unknown parameters $(\beta_1, \beta_2, \sigma_1, \sigma_2, \sigma_{12})$.

(a) Derive expressions for: (i) $P(y_{1i} > 0|x_i)$; (ii) $E[y_{1i}|x_i, y_{1i} > 0]$; and (iii) $E[y_{2i}|x_i, y_{1i} = 0]$. (4 points)

(b) Derive an expression for $E[y_{2i}|x_i]$. (1 point)

(c) Describe a two-step procedure using the expressions in (a) to consistently estimate $(\beta_1, \beta_2, \sigma_1, \sigma_2, \sigma_{12})$ or a subvector of it. Clearly explain how each parameter can be estimated, and also why not if it cannot be estimated or if we have to make any type of normalizations. (3 points)

(d) Write a log likelihood function to consistently estimate $(\beta_1, \beta_2, \sigma_1, \sigma_2, \sigma_{12})$ or a subvector of it using maximum likelihood, and also state if any parameter cannot be estimated or if we have to make any type of normalizations. (2 points)

(e) Instead of the above, suppose we simply ran a regression of $y_{2i}$ on a constant and $x_{2i}$ to estimate $\beta_2$. Derive an expression for the bias between the regression estimate with respect to $\beta_2$. (1 point)

(f) Suggest a way to test that $\sigma_{12} = 0$. (1 point)

(g) Instead of the equations in (1)-(2), suppose we have

$$y_{1i} = 1\{y_{1i}^* > 0\} \ , \tag{3}$$

$$y_{2i} = y_{2i}^* 1\{y_{1i}^* \leq 0\} \ , \tag{4}$$

i.e. we continue to have $y_{2i}$ to be censored at 0, but now only have a binary variable $y_{1i}$ determining whether $y_{1i}^*$ is strictly positive or not. In this case, similar to (c) above, clearly explain how we can consistently estimate $(\beta_1, \beta_2, \sigma_1, \sigma_2, \sigma_{12})$ or a subvector of it, highlighting if we cannot estimate certain parameters and if we have to make any type of normalizations. (3 points)