

Limited Dependent Variables & Selection: PS #2

Francis DiTraglia

HT 2021

This problem set is due on *Friday in Week 3 of HT 2021*. You need only submit solutions to questions 1–4, as question #5 will not be marked. See the explanation immediately preceding question #5 for further information.

1. Suppose that we observe N iid draws (y_i, \mathbf{x}_i) from a population of interest where $y_i \in \{0, 1\}$ and \mathbf{x}_i is a $(k \times 1)$ vector of dummy variables indicating which of k mutually exclusive “bins” person i falls into. For example, suppose that $k = 2$ and we defined the bins to be “female” and “male.” Then $\mathbf{x}'_i = [1 \ 0]$ would indicate that person i is female while $\mathbf{x}'_i = [0 \ 1]$ would indicate that person i is male. Note that \mathbf{x}_i does not include an intercept to avoid the dummy variable trap. The following parts explore the results of fitting the linear probability model $\mathbb{P}(y_i|\mathbf{x}_i) = \mathbf{x}'_i\boldsymbol{\beta}$ by running an OLS regression of y_i on \mathbf{x}_i . Following the usual conventions, define

$$\mathbf{X}' = [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_N], \quad \mathbf{y}' = [y_1 \ y_2 \ \cdots \ y_N]$$

- (a) Let N_j denote the number of individuals in the sample who fall into category j . In other words, if $x_i^{(j)}$ is the j th element of \mathbf{x}_i , then $N_j \equiv \sum_{i=1}^N x_i^{(j)}$. Show that

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} N_1 & & & 0 \\ & N_2 & & \\ & & \ddots & \\ 0 & & & N_k \end{bmatrix}$$

i.e. that $\mathbf{X}'\mathbf{X}$ is a $(k \times k)$ diagonal matrix with j th diagonal element N_j .

- (b) Substitute the preceding part into $\widehat{\boldsymbol{\beta}} \equiv (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ to obtain a simple, closed-form expression for $\widehat{\beta}_j$. Interpret your result.
 - (c) A critique of the LPM is that it can yield predicted probabilities that are greater than one or less than zero. Is this a problem in the present example?
2. This question concerns the Probit regression model $\mathbb{P}(y = 1|\mathbf{x}) = \Phi(\mathbf{x}'\boldsymbol{\beta})$ where Φ is the standard normal CDF.
 - (a) Derive the first order conditions for the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ based on an iid sample $(y_1, \mathbf{x}), \dots, (y_N, \mathbf{x}_N)$.

- (b) Suppose that $y = \mathbb{1}\{\mathbf{x}'\boldsymbol{\beta} + u > 0\}$ where $u \sim \mathcal{N}(0, 1)$ independently of \mathbf{x} and $\mathbb{1}(\cdot)$ is the indicator function. Show that this model is in fact *exactly equivalent* to the Probit regression model.
3. Consider a logit-Family model with $P_{ni} = \exp(V_{ni}) / \sum_{j=1}^J \exp(V_{nj})$ and $V_{nj} = \mathbf{x}'_{nj}\boldsymbol{\beta}$.
- (a) What *variety* of Logit-family model is this? How can you tell?
- (b) Show that the partial effects for this model are given by

$$\frac{\partial P_{ni}}{\partial \mathbf{x}_{ni}} = P_{ni}(1 - P_{ni})\boldsymbol{\beta}, \quad \text{and} \quad \frac{\partial P_{ni}}{\partial \mathbf{x}_{nk}} = -P_{ni}P_{nk}\boldsymbol{\beta} \quad \text{for } i \neq k$$

4. *This question is adapted from Wooldridge (2010).* Consider the Heckman selection model from the lecture slides. Assumption (d) of this model states that the conditional mean of u_1 given v_2 is linear: $\mathbb{E}(u_1|v_2) = \gamma_1 v_2$. In this question, you will explore the consequences of replacing Assumption (d) with a *quadratic* conditional mean function, in particular

$$\text{Assumption (d*)} \quad \mathbb{E}(u_1|v_2) = \gamma_1 v_2 + \gamma_2(v_2^2 - 1).$$

In your answers to the following parts, assume that all assumptions other than (d) of the Heckman Selection model continue to apply.

- (a) Show that Assumption (c) and (d*) imply $\mathbb{E}(u_1) = 0$. Using your answer, explain why the RHS of Assumption (d*) does *not* take the form $\gamma_1 v_2 + \gamma_2 v_2^2$.
- (b) Let a be a constant, $z \sim N(0, 1)$ and $\lambda(\cdot)$ be the inverse Mills ratio defined in the lecture slides. It can be shown that:

$$\text{Var}(z|z > -a) = 1 - \lambda(a) [\lambda(a) + a].$$

Use this result to prove that

$$\mathbb{E}(y_1|\mathbf{x}, y_2 = 1) = \mathbf{x}'_1\boldsymbol{\beta}_1 + \gamma_1\lambda(\mathbf{x}'\boldsymbol{\delta}_2) - \gamma_2\lambda(\mathbf{x}'\boldsymbol{\delta}_2)\mathbf{x}'\boldsymbol{\delta}_2.$$

$$\text{Hint: } \mathbb{E}(v_2^2|v_2 > -a) = \text{Var}(v_2|v_2 > -a) + [\mathbb{E}(v_2|v_2 > -a)]^2.$$

- (c) Using the expression for $\mathbb{E}(y_1|\mathbf{x}, y_2 = 1)$ from the preceding part, explain how to carry out the Heckman Two-step procedure under assumption (d*).
- (d) Consider a “naïve” OLS regression of y_1 on \mathbf{x}_1 for the subset of individuals with $y_2 = 1$. Without actually running the naïve regression, explain how you could use the estimates from your Heckman Two-step procedure in the preceding part to determine whether or not the naïve OLS of β_1 would be biased.

The following applied question will *not be marked*, but you encouraged to complete it nonetheless as it will build your understanding of the material from the lectures. Solving this problem will requires some of the R material from Lecture #6.

5. *This question is adapted from Wooldridge (2010).* To answer it you will need to use the dataset `BWGHT.RAW`, which can either be downloaded from the MIT Press website for the text, or loaded directly into R using the package `Wooldridge`. Documentation for the dataset is available in the R package or alternatively at <http://fmwww.bc.edu/ec-p/data/wooldridge/bwght.des>
- (a) Create a binary variable called *smokes* that equals one if a woman smokes during pregnancy, zero otherwise. Then estimate a probit regression that uses *motheduc*, *white*, and $\log(\textit{faminc})$ to predict *smokes*. Summarize your results.
 - (b) Consider two white women with family income equal to the sample mean: Alice has 12 years of education while Beth has 16. What is the estimated difference in the probability of smoking during pregnancy for Alice compared to Beth?
 - (c) Calculate the average partial effect of $\log(\textit{faminc})$ in your estimated model.
 - (d) Calculate the pseudo-R-squared of your model.