Limited Dependent Variables & Selection: PS #1

Francis DiTraglia

HT 2021

This problem set is due on *Monday of HT Week 6 at noon*. You do not have to submit solution to questions 1–2; they will be discussed in class but will not be marked.

Question #1 will not be marked; you do not have to submit a solution.

1. Let $y \sim$ Poisson(θ).

   (a) Using steps similar to the derivation of $\mathbb{E}[y]$ from the lecture slides, show that $\mathbb{E}[y(y-1)] = \theta^2$.

   (b) Use your answer to the preceding part, along with the result $\mathbb{E}[y] = \theta$, to show that $\text{Var}(y) = \theta$.

Question # 2 will not be marked; you do not have to submit a solution.

2. Suppose that we observe count data $y_1, \ldots, y_N \sim$ iid p_o and our model $f(y_i|\theta)$ is a Poisson(θ) probability mass function. Show that $\widehat{K} = s_y^2/(\bar{y})^2$ where we define $s_y^2 = \frac{1}{N}\sum_{i=1}^N (y_i - \bar{y})^2$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^N y_i$.

3. Let $\widehat{\boldsymbol{\beta}}$ be the conditional maximum likelihood estimator of $\boldsymbol{\beta}_o$ in a Poisson regression model with conditional mean function $\mathbb{E}(y_i|\mathbf{x}_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta}_o)$, based on a sample of iid observations $(y_1, \mathbf{x}_1), \ldots, (y_N, \mathbf{x}_N)$.

   (a) Derive the first-order conditions for $\widehat{\boldsymbol{\beta}}$.

   (b) Using your answer to the previous part show that, so long as $\mathbf{x}_i$ includes a constant, the residuals $\widehat{u}_i \equiv y_i - \exp(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})$ sum to zero, as in OLS regression.

   (c) Using your answer to the preceding part, show that $\left[\frac{1}{N}\sum_{i=1}^N \exp(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})\right] = \bar{y}$, where $\bar{y}$ is the sample mean of y, so that $\bar{y}\widehat{\beta}_j$ equals the estimated average partial effect of x_j in this model.

   (d) Explain why multiplying the estimated coefficients from this model by $\bar{y}$ makes them roughly comparable to the corresponding OLS estimates from the model $y_i = \mathbf{x}_i'\boldsymbol{\theta} + \varepsilon_i$.

4. 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 $\mathbf{x}_i' = \begin{bmatrix} 1 & 0 \end{bmatrix}$ would indicate that person i is female while $\mathbf{x}_1' = \begin{bmatrix} 0 & 1 \end{bmatrix}$ 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

$$\mathbf{X}' = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \end{bmatrix}, \quad \mathbf{y}' = \begin{bmatrix} y_1 & y_2 & \cdots & y_N \end{bmatrix}$$

(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 jth 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 jth 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?