International School of Economics at TSU

Econometrics 2

Lasha Chochua

Problem Set 10

Instructions: You are encouraged to solve the problems before the recitation. Additionally, you are encouraged to work in groups. It is **not mandatory** to submit solutions unless stated otherwise. However, if you would like to share your solution, I would be happy to review it.

Problem 1: Let z_1 be a vector of variables, let z_2 be a continuous variable, and let d_1 be a dummy variable.

a. In the model

$$\mathbb{P}(y = 1 \mid \mathbf{z}_1, z_2) = \Phi(\mathbf{z}_1\delta_1 + \gamma_1 z_2 + \gamma_2 z_2^2),$$

find the partial effect of z_2 on the response probability. How would you estimate this partial effect?

b. In the model

$$\mathbb{P}(y = 1 \mid \mathbf{z}_1, z_2, d_1) = \Phi(\mathbf{z}_1\delta_1 + \gamma_1 z_2 + \gamma_2 d_1 + \gamma_3 z_2 d_1),$$

find the partial effect of z_2. How would you measure the effect of d_1 on the response probability? How would you estimate these effects?

Solution:

a. If $\mathbb{P}(y_i = 1 \mid \mathbf{z}_1, z_2) = \Phi(\mathbf{z}_1\delta_1 + \gamma_1 z_2 + \gamma_2 z_2^2)$ then

$$\frac{\partial \mathbb{P}(y = 1 \mid \mathbf{z}_1, z_2)}{\partial z_2} = (\gamma_1 + 2\gamma_2 z_2) \cdot \phi(\mathbf{z}_1\delta_1 + \gamma_1 z_2 + \gamma_2 z_2^2);$$

for given $\mathbf{z}$, the partial effect is estimated as

$$(\hat{\gamma}_1 + 2\hat{\gamma}_2 z_2) \cdot \phi(\mathbf{z}_1\hat{\delta}_1 + \hat{\gamma}_1 z_2 + \hat{\gamma}_2 z_2^2),$$

where, of course, the estimates are the probit estimates.

b. In the model

$$\mathbb{P}(y_i = 1 \mid \mathbf{z}_1, z_2, d_1) = \Phi(\mathbf{z}_1\delta_1 + \gamma_1 z_2 + \gamma_2 d_1 + \gamma_3 z_2 d_1),$$

the partial effect of z_2 is

$$\frac{\partial \mathbb{P}(y = 1 \mid \mathbf{z}_1, z_2, d_1)}{\partial z_2} = (\gamma_1 + \gamma_3 d_1) \cdot \phi(\mathbf{z}_1\delta_1 + \gamma_1 z_2 + \gamma_2 d_1 + \gamma_3 z_2 d_1).$$

The effect of d_1 is measured as the difference in the probabilities at $d_1 = 1$ and $d_1 = 0$:

$$\mathbb{P}(y = 1 \mid \mathbf{z}, d_1 = 1) - \mathbb{P}(y = 1 \mid \mathbf{z}, d_1 = 0) = \Phi(\mathbf{z}_1\delta_1 + \gamma_2 + (\gamma_1 + \gamma_3)z_2) - \Phi(\mathbf{z}_1\delta_1 + \gamma_1 z_2).$$

Again, to estimate these effects at given $\mathbf{z}$ and — in the first case — d_1, we just replace the parameters with their probit estimates, and use average or other interesting values of $\mathbf{z}$.

Problem 2: Consider the probit model

$$\mathbb{P}(y = 1 \mid \mathbf{z}, q) = \Phi(\mathbf{z}_1\delta_1 + \gamma_1 z_2 q),$$

where q is independent of $\mathbf{z}$ and distributed as Normal$(0, 1)$; the vector $\mathbf{z}$ is observed but the scalar q is not.

a. Find the partial effect of z_2 on the response probability, namely,

$$\frac{\partial \mathbb{P}(y = 1 \mid \mathbf{z}, q)}{\partial z_2}.$$

b. Show that $\mathbb{P}(y = 1 \mid \mathbf{z}) = \Phi[\mathbf{z}_1\delta_1/(1 + \gamma_1^2 z_2^2)^{1/2}]$.

Solution:

a. If $\mathbb{P}(y = 1 \mid \mathbf{z}, q) = \Phi(\mathbf{z}_1\delta_1 + \gamma_1 z_2 q)$ then

$$\frac{\partial \mathbb{P}(y = 1 \mid \mathbf{z}, q)}{\partial z_2} = \gamma_1 q \cdot \phi(\mathbf{z}_1\delta_1 + \gamma_1 z_2 q),$$

assuming that z_2 is not functionally related to $\mathbf{z}_1$.

b. Write $y^* = \mathbf{z}_1\delta_1 + r$, where $r = \gamma_1 z_2 q + e$, and e is independent of $(\mathbf{z}, q)$ with a standard normal distribution. Because q is assumed independent of $\mathbf{z}$,

$$q \mid \mathbf{z} \sim \text{Normal}(0, \gamma_1^2 z_2^2 + 1);$$

this follows because $\mathbb{E}(r \mid \mathbf{z}) = \gamma_1 z_2 \mathbb{E}(q \mid \mathbf{z}) + \mathbb{E}(e \mid \mathbf{z}) = 0$. Also,

$$\begin{aligned} \text{Var}(r \mid \mathbf{z}) &= \gamma_1^2 z_2^2 \text{Var}(q \mid \mathbf{z}) + \text{Var}(e \mid \mathbf{z}) + 2\gamma_1 z_2 \text{Cov}(q, e \mid \mathbf{z}) \\ &= \gamma_1^2 z_2^2 + 1 \end{aligned}$$

because $\text{Cov}(q, e \mid \mathbf{z}) = 0$ by independence between e and $(\mathbf{z}, q)$. Thus, $r/\sqrt{\gamma_1^2 z_2^2 + 1}$ has a standard normal distribution independent of $\mathbf{z}$. It follows that

$$\mathbb{P}(y = 1 \mid \mathbf{z}) = \Phi\left(\frac{\mathbf{z}_1\delta_1}{\sqrt{\gamma_1^2 z_2^2 + 1}}\right). \tag{1}$$

Problem 3: Consider taking a large random sample of workers at a given point in time. Let $sick_i = 1$ if person i called in sick during the last 90 days, and zero otherwise. Let $\mathbf{z}_i$ be a vector of individual and employer characteristics. Let $cigs_i$ be the number of cigarettes individual i smokes per day (on average).

a. Explain the underlying experiment of interest when we want to examine the effects of cigarette smoking on workdays lost.

b. Why might $cigs_i$ be correlated with unobservables affecting $sick_i$?

c. One way to write the model of interest is

$$\mathbb{P}(sick = 1 \mid \mathbf{z}, cigs, q_1) = \Phi(\mathbf{z}_1\delta_1 + \gamma_1 cigs + q_1),$$

where $\mathbf{z}_1$ is a subset of $\mathbf{z}$ and q_1 is an unobservable variable that is possibly correlated with $cigs$. What happens if q_1 is ignored and you estimate the probit of $sick$ on $\mathbf{z}_1$, $cigs$?

d. Can $cigs$ have a conditional normal distribution in the population? Explain.

e. Explain how to test whether $cigs$ is exogenous. Does this test rely on $cigs$ having a conditional normal distribution?

f. Suppose that some of the workers live in states that recently implemented no-smoking laws in the workplace. Does the presence of the new laws suggest a good IV candidate for $cigs$?

Solution:

a. What we would like to know is that, if we exogenously change the number of cigarettes that someone smokes per day, what effect would this have on the probability of missing work over a three-month period? In other words, we want to infer causality, not just find a correlation between missing work and cigarette smoking.

b. Since people choose whether and how much to smoke, we certainly cannot treat the data as coming from the experiment we have in mind in part a. (That is, we cannot randomly assign people a daily cigarette consumption.) It is possible that smokers are less healthy to begin with, or have other attributes that cause them to miss work more often. Or, it could go the other way: cigarette consumption may be related to personality traits that make people harder workers. In any case, *cigs* might be correlated with the unobservables in the equation.

c. If we start with the model

$$\mathbb{P}(y = 1 \mid \mathbf{z}, cigs, q_1) = \Phi(\mathbf{z}_1\delta_1 + \gamma_1 cigs + q_1), \tag{2}$$

but ignore q_1 when it is correlated with *cigs*, we will not consistently estimate anything of interest, whether the model is linear or nonlinear. Thus, we would not be estimating a causal effect. If q_1 is independent of *cigs*, the probit ignoring q_1 does estimate the average partial effect of another cigarette.

d. No. There are many people in the working population who do not smoke. Thus, the distribution (conditional or unconditional) of *cigs* piles up at zero. Also, since *cigs* takes on integer values, it cannot be normally distributed. But it is really the pile up at zero that is the most serious issue.

e. Use the Rivers-Vuong test. Obtain the residuals, $\hat{r}_2$, from the regression of *cigs* on $\mathbf{z}$. Then, estimate the probit of y on $\mathbf{z}_1, cigs, \hat{r}_2$ and use a standard t-test on $\hat{r}_2$. This does not rely on normality of r_2 (or *cigs*). It does, of course, rely on the probit model being correct for y under H_0.

f. Assuming people will not immediately move out of their state of residence when the state implements no smoking laws in the workplace, and that state of residence is roughly independent of general health in the population, a dummy indicator for whether the person works in a state with a new law can be treated as exogenous and excluded from (2). (These situations are often called "natural experiments.") Further, *cigs* is likely to be correlated with the state law indicator because since people will not be able to smoke as much as they otherwise would. Thus, it seems to be a reasonable instrument for *cigs*.

Problem 4: *Suppose we have a distribution with the following pdf (called a gamma distribution)*

$$f(x \mid a) = \frac{a^5}{(4)!}x^4 e^{-ax}.$$

Furthermore, Suppose we have independent data $x_1, x_2, ..., x_m$ drawn from this distribution. Find the maximum likelihood estimate (MLE) for a.

Solution:

The likelihood for x_i is

$$f(x_i \mid a) = \frac{a^5}{4!} x_i^4 e^{-ax_i}.$$

So, the likelihood of the data is

$$f(\text{data} \mid a) = \prod_{i=1}^{m} f(x_i \mid a) = \frac{a^{5m}}{(4!)^m} P^4 e^{-aS},$$

where $P = \prod x_i$ (product of data) and $S = \sum x_i$ (sum of data).

So, the log likelihood is

$$l(a) = 5m \ln(a) + 4 \ln(P) - aS - m \ln(4!).$$

Taking the derivative and setting it to 0, we get

$$l'(a) = \frac{5m}{a} - S = 0 \Rightarrow \boxed{\text{The MLE } \hat{a} = \frac{5m}{S}}.$$

Note: $\hat{a} = 5/(S/m) = 5/\bar{x}$, where $\bar{x}$ is the data mean.

Problem 5: *In this problem we will use maximum likelihood estimates to develop Gauss' method of least squares for fitting lines to data.*

Bivariate data means data of the form

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n).$$

For bivariate data the simple linear regression model assumes that, for some values of the parameters a and b, we have

$$y_i = ax_i + b + \text{ random measurement error.}$$

The model assumes the measurement errors are independent and identically distributed and follow a $N(0, \sigma^2)$ distribution. (The values x_i may or may not be random.)

It turns out that, under some assumptions about random variation of measurement error, one way to find a "best" line is by solving a maximum likelihood problem.

The goal is to find the values of the model parameters a and b that give the MLE for this model. To guide you, we note that the model says that

$$y_i \sim N(ax_i + b, \sigma^2).$$

Also remember that you know the density function for this distribution.

(a) *For a general datum (x_1, y_1) give the likelihood and log likelihood functions (these will be functions of y_1, x_1, a, b, and σ.)*

Solution: Since $y_i \sim N(ax_i + b, \sigma^2)$ the likelihood with data (x_1, y_1) is

$$f(x_1, y_1 \mid a, b, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-(y_1 - ax_1 - b)^2/(2\sigma^2)}.$$

The log likelihood is

$$\ln(f(x_1, y_1 \mid a, b, \sigma)) = -\ln(\sqrt{2\pi}\,\sigma) - \frac{(y_1 - ax_1 - b)^2}{2\sigma^2}.$$

(b) *Consider the data (1,8), (3,2), (5,1). Assume that $\sigma = 3$ is a known constant and find the maximum likelihood estimate for a and b.*

Note: *We gave you a specific value of σ, to avoid the distraction of one more symbol. If you look at your calculations, you should see that the value of σ plays no role in finding the MLE for a and b. We get the same answer no matter what the value.*

Solution:

The likelihood for all the data is the product of the individual likelihoods. So,

$$f((1,8),\ (3,2),\ (5,1) \mid a, b, \sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^3 e^{-[(8-a-b)^2 + (2-3a-b)^2 + (1-5a-b)^2]/(2\sigma^2)}.$$

Taking the natural log (and replacing the list of data by the word 'data') we get

$$\ln(f(\text{data} \mid a, b, \sigma)) = -3\ln(\sqrt{2\pi}\,\sigma) - \frac{(8-a-b)^2 + (2-3a-b)^2 + (1-5a-b)^2}{2\sigma^2}.$$

Since we want to find a and b that maximize the likelihood, we take the partial derivatives and set them to 0.

$$\frac{\partial \ln(f(\text{data} \mid a, b, \sigma))}{\partial a} = \frac{2}{2\sigma^2}\left((8 - a - b) + 3(2 - 3a - b) + 5(1 - 5a - b)\right) = 0$$
$$\frac{\partial \ln(f(\text{data} \mid a, b, \sigma))}{\partial b} = \frac{2}{2\sigma^2}\left((8 - a - b) + (2 - 3a - b) + (1 - 5a - b)\right) = 0$$

These are two equations in the unknowns a and b. We simplify and solve:

$$\begin{aligned} 35a + 9b &= 19 \\ 9a + 3b &= 11 \end{aligned}$$

which gives

$$a = -7/4 = -1.75, \quad b = 107/12 \approx 8.917.$$

The linear regression fit of a line to the data is

$$\boxed{y = ax + b = -\frac{7}{4}x + \frac{107}{12}}.$$

Problem 6:

(a) *Suppose we have data 1.2, 2.1, 1.3, 10.5, 5 which we know is drawn independently from a uniform(a, b) distribution. Give the maximum likelihood estimate for the parameters a and b.*

Solution: The pdf for uniform(a, b) distribution takes two values

$$f(x \mid a, b) = \begin{cases} 1/(b-a) & \text{if } x \text{ is in } [a, b] \\ 0 & \text{otherwise} \end{cases}$$

Since the likelihood is the product of the likelihoods of each data point, the likelihood function is

$$f(\text{data} \mid a, b) = \begin{cases} 1/(b-a)^5 & \text{if all data is in } [a, b] \\ 0 & \text{if not} \end{cases}$$

This is maximized when $(b-a)$ is as small as possible. Since all the data has to be in the interval $[a,b]$, we minimize $(b-a)$ by taking $a = \min(\text{data})$ and $b = \max(\text{data})$.

So,

$$\boxed{a = 1.2, \quad b = 10.5}.$$

(b) *Suppose we have data $x_1, x_2, \ldots, x_n$ which we know is drawn independently from a uniform(a,b) distribution. Give the maximum likelihood estimate for the parameters a and b.*

Solution: The same logic as in part (a) shows

$$\boxed{a = \min(x_1, \ldots, x_n) \quad \text{and} \quad b = \max(x_1, \ldots, x_n)}.$$