Microeconometrics

Problem set 3 - Solutions

Note. Upload a copy of your solutions on Moodle. The solutions should include a .pdf document with the answers to each one of the questions **including the screenshot of command used to generate the output and the screenshot of results or the output** generated from the statistical software (feel free to use STATA or R, or other softwares). Please check the submission deadline on moodle, late problem sets will not be accepted.

1. (50%) **Discrete choice models.** Suppose you wish to predict productivity of new workers in a large manufacturing firm. Consider the dataset **weco** that contains the following variables: y_i is a physical productivity measure for worker i after the initial training period, sex_i is a dummy variable equal to 1 for male workers, dex_i is a score in a physical dexterity exam administered before the worker was hired, lex_i is the number of years of education of the worker, and $quit_i$ is a dummy variable equal to 1 if the worker quit within the first six months. The variables *job_tenure* and *censored* provide the actual duration of employment and a censoring indicator, respectively. If the censored indicator is 0 then the corresponding duration is censored. You won't need *job_tenure* and *censored* in this exercise. Consider the following model for quits:

$$quit = \alpha_0 + \alpha_1 sex + \alpha_2 dex + \alpha_3 lex + \alpha_4 lex^2 + u$$

   (a) Estimate the model by OLS. Interpret the regression coefficients.

   SOLUTION: in STATA, you can use the following commands:

   ```
   gen lexsq=lex×lex
   reg quit sex dex lex lexsq, robust
   ```

   OLS estimation yields the following estimates:

$$\widehat{quit} = \underset{(0.801)}{1.601} + \underset{(0.0333)}{0.0892} sex - \underset{(0.002)}{0.0168} dex - \underset{(0.125)}{0.101} lex + \underset{(0.005)}{0.004} lex^2$$

   Male workers are 8.9 percentage points more likely to quit within the first 6 months than female workers, on average, ceteris paribus. An additional point in the score of the dexterity exam decreases the likelihood of quitting by 1.68 percentage points, on average, ceteris paribus. An additional year of education changes the probability of quitting by -0.101-2*0.004*lex, on average, ceteris paribus. This means that the marginal effect of education depends on the number of years of education of the worker.

   (b) Estimate the model by probit or logit (motivating your choice) and interpret the results.

   SOLUTION: In case you estimate a probit model, in STATA, you can use the following command: `probit quit sex dex lex lexsq`. Probit estimation yields the following estimates:

$$\widehat{quit} = \underset{(2.449)}{3.549} + \underset{(0.109)}{0.268} sex - \underset{(0.008)}{0.053} dex - \underset{(0.388)}{0.313} lex + \underset{(0.016)}{0.012} lex^2$$

   In a probit model we define the probability of a success as: $\Pr[y_i = 1] = F(x_i'\beta)$, where $F(.)$ is the cumulative normal distribution function.

Assuming that $F(.)$ is differentiable with derivative $f(.)$, this implies that the marginal effect of the jth explanatory variable is given by:

$$\frac{\partial \Pr[y_i = 1]}{\partial x_{ij}} = f(x_i'\beta)\beta_j$$

where $f(.)$ is the normal density function in the case of the probit model. This result means that the marginal effect of changes in the explanatory variables depends on the level of these variables and that the estimated coefficients $\widehat{\beta}$ give the sign of the impact because $F(.) > 0$. Then, according to the probit estimates, the probability of quitting within 6 months is higher for men than for women, decreasing in the score of the dexterity exam, and convex in the number of years of education.

In case you estimate a logit model, in STATA, you can use the following command: `logit quit sex dex lex lexsq`. Logit estimation yields the following estimates:

$$\widehat{quit} = \underset{(4.105)}{6.220} + \underset{(0.186)}{0.479}\,sex - \underset{(0.014)}{0.094}\,dex - \underset{(0.648)}{0.539}\,lex + \underset{(0.026)}{0.021}\,lex^2$$

In a logit model we define the probability of a success as: $\Pr[y_i = 1] = F(x_i'\beta)$, where $F(.)$ is the cumulative distribution function of the logistic distribution. Assuming that $F(.)$ is differentiable with derivative $f(.)$, this implies that the marginal effect of the jth explanatory variable is given by:

$$\frac{\partial \Pr[y_i = 1]}{\partial x_{ij}} = f(x_i'\beta)\beta_j$$

where $f(.)$ is the logistic density in the case of the logit model. This result means that the marginal effect of changes in the explanatory variables depends on the level of these variables and that the estimated coefficients $\widehat{\beta}$ give the sign of the impact because $F(.) > 0$. Then, according to the logit estimates, the probability of quitting within 6 months is higher for men than for women, decreasing in the score of the dexterity exam, and convex in the number of years of education.

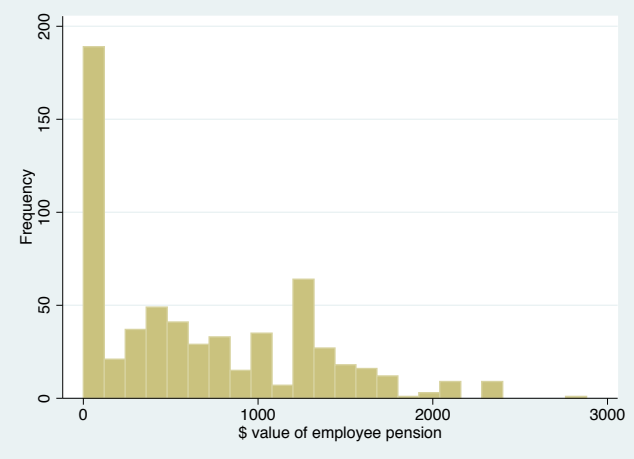
2. (50%) **Tobit model.** Consider the dataset **PENSION** containing cross-sectional family data on pension benefits.

   (a) Create a histogram for the variable pension. This variable is the value in dollars of an employee's pension. Why a Tobit model is appropriate for modeling pension benefits? Provide a detailed description of the model including the distributional assumptions and the corresponding likelihood function.

   SOLUTION: in STATA, you can use the following commands:

```
hist pension
hist pension, freq
sum pension, d
```

   We can see that out of 616 workers, 172, or about 28%, have zero pension benefits. For the 444 workers reporting positive pension benefits, the range is from 7.28 to 2,880.27. Therefore, we have a nontrivial fraction of the sample with pensions equal to 0, and the range of positive pension benefits is fairly wide. The Tobit model is well-suited to this kind of dependent variable.



In the case of a censored regression, we can write

$$y = \begin{cases} y^* \text{ if } y^* > 0 \\ 0 \text{ if } y* \leq 0 \end{cases}$$

Consistency requires the density $f(y^*|x)$ is correctly specified and model errors have to be normal and homoskedastic. In the case of the censored Tobit model the density varies according to whether $y > 0$ or $y = 0$. For $y > 0$, y$\sim \mathcal{N}[x'\beta, \sigma^2]$, the density is equal to:

$$\begin{aligned} f(y) &= f(y^*) \\ &= \left(\frac{1}{\sqrt{(2\pi\sigma^2)}}\right) \times \exp\left(\frac{-(y - x'\beta)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sigma}\phi\left(\frac{y - x'\beta}{\sigma}\right) \end{aligned}$$

where $\phi(.)$ is the $\mathcal{N}(0, 1)$ density function. Instead, for $y = 0$, $y^* \leq 0$, $y^* \sim \mathcal{N}[x'\beta, \sigma^2]$, the density is equal to:

$$\begin{aligned} f(0) &= \Pr[y = 0] = \Pr[y^* \leq 0] \\ &= \Pr[x'\beta + \varepsilon \leq 0] = \Pr\left[\frac{\varepsilon}{\sigma} \leq \frac{x'\beta}{\sigma}\right] \\ &= \Phi\left(\frac{-x'\beta}{\sigma}\right) = 1 - \Phi\left(\frac{x'\beta}{\sigma}\right) \end{aligned}$$

where $\Phi(z)$ is the $\mathcal{N}(0, 1)$ cumulative distribution function. Define an indicator variable, such that:

$$d_i = \begin{cases} 1 \text{ if } y_i > 0 \\ 0 \text{ if } y_i = 0 \end{cases}$$

The Censored Tobit density is given by:

$$f(y_i|x_i) = \left[\frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)\right]^{d_i} \times \left[1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)\right]^{1-d_i}$$

The log-likelihood function is given by:

$$\ln L(y_i|x_i,\beta,\sigma^2) = \sum_{i=1}^{N}\left[d_i \ln \frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right) + (1-d_i)\ln\left[1-\Phi\left(\frac{x_i'\beta}{\sigma}\right)\right]\right]$$

(b) Estimate a Tobit model explaining pension in terms of exper, age, tenure, educ, depends, married, white, and male. Do white and male individuals have statistically significant higher expected pension benefits?

SOLUTION: in STATA, you can use the following commands:

```
tobit pension exper age tenure educ depends married white male, ll(0)
```

Being white or male (or both) increases the predicted pension benefits, although only male is statistically significant with the corresponding t-statistic approximately equal to 4.41 and p-value of 0. For white, the p-value of 0.159 does not allow us to reject the null hypothesis that its coefficient equals 0 even at 10% significance level.

(c) Compute the marginal effects from the Tobit model. Compute also the marginal effects for the censored conditional mean evaluated at the mean. How does the pension (evaluated at the mean) change with one more year of experience? [HINT: the option `atmeans` used with the command `margins` in STATA allows to evaluate partial effects at the mean]

SOLUTION: in STATA, you can use the following commands:

```
mfx
margins, dydx(*) predict (ystar(0,.))  atmeans
```

According to the estimates, the actual pension increases 4.06$ for individuals receiving and not receiving pensions with an additional year of experience.

(d) Write the general expression for the expected value of y conditional on the covariate x, $E(y|x)$, in the Tobit model. Use the results from part (b) to estimate the difference in expected pension benefits for a white male and a non-white female, both of whom are 35 years old, are single with no dependents, have 16 years of education, and have 10 years of experience. [HINT: you don't need any command for this, you can compute it using the expression for $E(y|x)$ and estimates in point b)]

SOLUTION: in the Tobit model $E(y|x)$ can be written as

$$E(y|x) = \Phi\left(\frac{x'\beta}{\sigma}\right)x'\beta + \sigma\phi\left(\frac{x'\beta}{\sigma}\right)$$

First, we consider $\mathbf{x_A}$ with white = 1, male = 1, age = 35, married = 0, depends = 0, educ = 16 and exper = tenure = 10. The linear index $x'\widehat{\beta}$ is equal to $-1252.43 + 5.20 \times 10 - 4.64 \times 35 + 36.02 \times 10 + 93.21 \times 16 + 35.28 \times 0 + 53.69 \times 0 + 144.09 \times 1 + 308.15 \times 1 = 940.97$. Second, we consider $\mathbf{x_B}$ with white = 0, male = 0, age = 35, married = 0, depends = 0, educ = 16 and exper = tenure = 10. The linear index $x'\widehat{\beta}$ is equal to $-1252.43 + 5.20 \times 10 - 4.64 \times 35 + 36.02 \times 10 + 93.21 \times 16 + 35.28 \times 0 + 53.69 \times 0 + 144.09 \times 0 + 308.15 \times 0 = 488.73$. Since the estimated standard deviation of the error term ε_i is equal to $\widehat{\sigma} = 677.74$ we can

then write:

$$\begin{aligned} E(y|\mathbf{x_A}) &= \Phi\left(\frac{940.97}{677.74}\right)940.97 + 677.74\phi\left(\frac{940.97}{677.74}\right) \\ &= 0.92 \times 940.97 + 677.74 \times 0.15 \\ &= 966.49 \end{aligned}$$

and

$$\begin{aligned} E(y|\mathbf{x_B}) &= \Phi\left(\frac{488.73}{677.74}\right)488.73 + 677.74\phi\left(\frac{488.73}{677.74}\right) \\ &= 0.76 \times 488.73 + 677.74 \times 0.31 \\ &= 582.16 \end{aligned}$$

respectively. The difference in the expected pension value for a white male and for a nonwhite female with the same all other characteristics is thus 966.49 - 582.16 = 384.33.