PROBLEM SET 7

Problem 1 (Probit Selection Model)

Consider the following Probit selection model:

$$\begin{aligned} Y_1 &= Y_2 \cdot Y_1^* \\ Y_1^* &= X_1\beta_1 + U_1, \\ Y_2 &= 1\left[Y_2^* \geq 0\right], \\ Y_2^* &= X_2\beta_2 + U_2, \end{aligned}$$

from which one obtains a random sample of observations of (Y_1, Y_2, X_1, X_2), denoted $\{(y_{1i}, y_{2i}, x_{1i}, x_{2i}) : i$ where Y_1^* is the outcome variable of interest and the realization of Y_2 determines whether $Y_1 = Y_1^*$ or $Y_1 = 0$. Assume that (U_1, U_2) and $X = (X_1, X_2)$ are independent, and that (U_1, U_2) are bivariate normally distributed, each with mean zero, variances $var(U_1) = var(U_2) = 1$, and covariance $cov(U_1, U_2) = \rho$

1. What is the selection probability $\Pr\{Y_2 = 1|X = x\}$? How can you consistently estimate the parameters of this equation? [5]

2. Solve for the conditional density for Y_1 given $Y_2 = 1$ and X. What is $E[Y_1|Y_2 = 1, X = x]$? [5]

   Hint: By joint normality of U_1 and U_2, the conditional distribution of U_1 given $U_2 > c$ can be deduced as a function of c and ρ as

   $$g(u_1; c, \rho) = f(u_1|U_2 > c) = \frac{1}{1 - \Phi(c)} \int_c^{\infty} f(u_1|U_2 = u_2)\, \phi(u_2)\, du_2,$$

   where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal pdf and cdf, respectively. You may use the function $g(u_1; c, \rho)$ in your answer.

   Furthermore, recall that the distribution of U_1 conditional on $U_2 = c$ is

   $$U_1|(U_2 = c) \sim \mathcal{N}\left(\rho c, 1 - \rho^2\right),$$

   with conditional pdf

   $$f_{U_1|U_2}(u_1|U_2 = c) = \frac{1}{\sqrt{1 - \rho^2}} \phi\left(\frac{u_1 - \rho c}{\sqrt{1 - \rho^2}}\right),$$

i.e. the pdf of a normal random variable with mean ρc and variance $1-\rho^2$. Because U_1 is normally distributed with mean zero and unit variance,

$$E\left[U_1|U_1 > c\right] = \frac{\phi\left(c\right)}{1-\Phi\left(c\right)}.$$

3. What is the log likelihood for for this model? [5]

   Hint: You may use the function $g\left(u_1; c, \rho\right)$ defined in the previous hint in your answer.

4. Propose two different ways to consistently estimate β_1. Which provides a more efficient estimator asymptotically? [5]

5. Are the parameters β_1, β_2, ρ identified? Do you need to impose any additional conditions for identification? [5]

6. Why might one wish to test the hypothesis that $\rho = 0$? What implication would this have? [5]

Problem 2 (Applied Probit Selection Model)

We use the dataset RandHIE from package *sampleSelection* coming from the RAND Health Insurance Experiment (RHIE). For more details read the R and Wikipedia help on the experiement. The data extract comes from Deb and Trivedi (2002), who modeled the number of outpatient visits to a medical doctor and to all providers using count data models.

Here instead we model annual health expenditures. The regressors are can be broken down into health insurance variables (logc, idp, lpi, and fmde), socioeconomic characteristics (linc, lfam, xage, female, child, fchild, black and educdec) and health status variables (physlm, disea, hlthg, hlthf and hlthp). The analysis is using only the second year of data.

The dependent variable y is annual individual health expenditures (*meddol*). We are especially interested in the effect of coinsurance rate *logc* on the individual expenditure (http://en.wikipedia.org/wiki/Co-insurance). An econometric model needs to take account of two complications: (1) Health expenditures are zero for 23.2% of the sample and (2) the positive health expenditures are very right-skewed with a mean of $221 that is much larger than the median of 53$. The logarithmic transformation eliminates this skewness, with a mean of 4.07 close to the median of 3.96 and the skewness statistic falls from 24.0 to 0.3. The kurtosis is 3.29, close to the normal value of 3.

We focus on modeling $\ln y$ for those with positive medical expenditures. We model the data with a Tobit II model where the selection is given by y_2 is the indicator of positive expenditure (*binexp*), and y_1 is *lnmeddol*. Note that it is not meaningful to consider the value of y_1 when $y_2 = 0$. In that case the annual individual helth expenditure is 0 with no defined logarithm.

1. Explain why we believe there might be behavioural selection in this data set.
2. Read the data in your memory and understand what each variable mean
3. Create a subsample using only the variables from year two and make sure that your subsample does not have NA in variable *educdec*
4. Assumme $X = X_1$ and run Procedure 19.1
5. Estimate the model with Maximum likelihood
6. Interpret your results