Matematické Metody v Ekonometrii 10.

Limited Dependent Variable Models

Blanka Šedivá

KMA

zimní semestr 2016/2017

Model with Limited Dependent Variables

- in practical applications one often has to cope with phenomena that are of a discrete or mixed discrete continuous nature;
- the regressand Y is a binary, or dichotomous, variable;
- in a model where Y is quantitative, our objective is to estimate its expected, or mean, value given the values of the regressors

$$\mathrm{E}\left(Y_i | X_{1i}, X_{2i}, \ldots, X_{ki}\right)$$

  where the X's are regressors, both quantitative and qualitative
- in models where Y is qualitative, our objective is to find the probability of something happening, such as voting for a Democratic candidate, or owning a house, or belonging to a union, or participating in a sport etc.
- qualitative response model, probability models, discrete choice models, binary response regression model, ...

Examples

- Marketing researchers use discrete choice models to study consumer demand and to predict competitive business responses, enabling choice modelers to solve a range of business problems, such as pricing, product development, and demand estimation problems.
- Transportation planners use discrete choice models to predict demand for planned transportation systems, such as highway routings and rapid transit systems.
- Energy forecasters and policymakers use discrete choice models for households' and firms' choice of heating system, appliance efficiency levels, and fuel efficiency level of vehicles.
- Environmental studies utilize discrete choice models to examine the recreators' choice of, e.g., fishing or skiing site and to infer the value of amenities, such as campgrounds, fish stock, and warming huts, and to estimate the value of water quality improvements.
- Labour economists use discrete choice models to examine participation in the work force, occupation choice, and choice of college and training programs.

Binary response model

There are three approaches to developing a probability model for a binary response variable:

- The linear probability model (LPM)
- The logit model
- The probit model

consider the following regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon,$$

where $Y_1, Y_2, \ldots Y_n$ are binary variables, defined as

$$Y_i = \begin{cases} 1 & \text{with probability } \pi_i \\ 0 & \text{with probability } 1 - \pi_i, \end{cases}$$

The conditional expectation of Y_i given X_i, $E(Y_i|X_i)$, can be interpreted as the conditional probability that the event will occur given X_i , that is, $Pr(Y_i = 1|X_i)$.

Properties of linear probability model (LPM) model

- assuming $E(\varepsilon_i) = 0$, we obtain $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$
- $E(Y_i|X_i) = 1 \cdot Pr(Y_1 = 1|X_i) + 0 \cdot Pr(Y_1 = 0|X_i) = Pr(Y_1 = 1|X_i)$
- thus, the linear model implies that $\beta_0 + \beta_1 X_i$ is a probability and should therefore lie between 0 and 1.
  This is only possible if the X_i values are bounded and if certain restrictions on β_0, β_1 are satisfied. Usually this is hard to achieve in practice.
- the error term in has a highly non-normal distribution and suffers from heteroskedasticity
  $Pr(\varepsilon_i = -\beta_0 - \beta_1 X_i|X_i) = Pr(Y_i = 0|X_i) = 1 - \beta_0 - \beta_1 X_i$
  $Pr(\varepsilon_i = 1 - \beta_0 - \beta_1 X_i|X_i) = Pr(Y_i = 1|X_i) = \beta_0 + \beta_1 X_i$
  $var(\varepsilon_i|X_i) = (\beta_0 + \beta_1 X_i) \cdot (1 - \beta_0 - \beta_1 X_i)$
  var is not constant and there is depend on X_i
  we can use weighted least squares, where
  $\sqrt{w_i} = \sqrt{E(Y_i|X_i)(1 - E(Y_i|X_i))}$ and model
  $\frac{Y_i}{\sqrt{w_i}} = \frac{\beta_0}{\sqrt{w_i}} + \beta_1 \frac{X_i}{\sqrt{w_i}} + \frac{\varepsilon_i}{\sqrt{w_i}}$

Logit model

- For estimation we can used maximum likelihood principle
- $P(Y_i = j) = \pi_i^j (1-\pi_i)^{1-j} \qquad j = 0, 1$
- if $Y_1, Y_2, \ldots$ are independent, than $L = \prod_i P(Y_i) = \prod_{i=1}^n \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$
-

$$
\begin{aligned}
\ell(\pi_1, \pi_2, \ldots, \pi_n, Y_1, Y_2, \ldots, Y_n) &= \ln \prod \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \\
&= \sum_{i=1}^{n} (Y_i \ln \pi_i + (1-Y_i)\ln(1-\pi_i)) \\
&= \sum Y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \sum \ln(1-\pi_i)
\end{aligned}
$$

Interpretation of $\ln\left(\frac{\pi_i}{1-\pi_i}\right)$ I

- odds for an event is $P/(1-P)$ such that P is the probability of the event
$$\omega_i = \frac{\pi_i}{1-\pi_i} = \frac{P(y_i = 1)}{P(y_i = 0)}$$
- $\omega_i \in \langle 0; +\infty)$
- after logarithmic transformation $\eta_i = \ln(\omega_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) \in \mathbf{R}$
- the new form of model
$$\eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} \qquad i = 1, \ldots, n$$
in matrix form $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$,

Interpretation of $\ln\left(\frac{\pi_i}{1-\pi_i}\right)$ II

- **odds ratio** (which we will write as θ) between the odds for two sets of predictors (say $\mathbf{X}_{(1)}$ and $\mathbf{X}_{(2)}$) is given by

$$\theta = \frac{(\pi/(1-\pi))|_{\mathbf{x}=\mathbf{x}_{(1)}}}{(\pi/(1-\pi))|_{\mathbf{x}=\mathbf{x}_{(2)}}}$$

- An odds ratio of 1 serves as the baseline for comparison and indicates there is no association between the response and predictor. If the odds ratio is greater than 1, then the odds of success are higher for higher levels of a continuous predictor (or for the indicated level of a factor).
- for just a single predictor, X, the odds of success are $\frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_1 X)$
- If we increase X by one unit, the odds ratio is

$$\theta = \frac{\exp(\beta_0+\beta_1(X+1))}{\exp(\beta_0+\beta_1 X)} = \exp(\beta_1)$$

Logit model

- $$\begin{aligned} \ln\left(\frac{\pi_i}{1-\pi_i}\right) &= \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} \\ \frac{\pi_i}{1-\pi_i} &= \exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}) \\ \pi_i &= \frac{\exp(\ldots)}{1+\exp(\ldots)} = \frac{1}{1+\exp(-\beta_0 - \beta_1 x_{1i} - \ldots - \beta_k x_{ki})} \end{aligned}$$
- for one variable $\pi(x) = \frac{1}{1+e^{-\beta_0-\beta_1 x}}$
- function with inflexion point $x = -\frac{\beta_0}{\beta_1}$

Estimation of parameters I

- likelihood function

$$
\begin{aligned}
\ell(\beta_0, \beta_1, ..., \beta_k) = \quad &= \sum_{i=1}^{n} y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{n} \ln(1-\pi_i) \\
&= \beta_0 \sum_{i=1}^{n} y_i + \sum_{j=1}^{k} \beta_j \sum_{i=1}^{k} y_i x_{ji} + \sum_{i=1}^{n} \ln(1-\pi_i) \\
&= \sum_{i=1}^{n} (y_i \boldsymbol{X}_i \boldsymbol{\beta} - \ln(1+\exp(\boldsymbol{X}_i \boldsymbol{\beta}))
\end{aligned}
$$

- non-linear function where, $\pi_i = \pi_i(\beta_0, \beta_1, \cdots, \beta_k)$.

Estimation of parameters II

- for $1 - \pi_i = (1 - e^{\eta_i})^{-1}$ we can show

$$\begin{aligned} \frac{\partial}{\partial \eta_i} \ln(1 - \pi_i) &= -\frac{\partial}{\partial \eta_i} \ln(1 + e^{\eta_i}) \\ &= -\frac{e^{\eta_i}}{1 + e^{\eta_i}} = -\pi_i \end{aligned}$$

-

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_0} &= \sum_{i=1}^{n} \frac{\partial l}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi_i \\ \frac{\partial \ell}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{\partial l}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ji} - \sum_{i=1}^{n} \pi_i x_{ji}, \end{aligned}$$

Estimation of parameters III

- $$\begin{aligned} \sum y_i &= \sum \pi_i \\ \sum y_i x_{ji} &= \sum x_{ji} \pi_i \end{aligned}$$
- non-linear system of equations

$$(1, \mathbf{x})^T (\mathbf{y} - \pi(\beta_0, \beta_1, \ldots, \beta_k)) = 0$$

- we can solve it numerical and obtain $\hat{\pi} = \pi(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$
- $\hat{\beta}_j = b_j$ and $\hat{\pi}_i = p_i$

Wald test

- the test of significance for individual regression coefficients in logistic regression
- matrix of variance for model

$$\boldsymbol{W}_\pi(\boldsymbol{\beta}) = \mathrm{diag}\left(\pi_1(1-\pi_1), \pi_2(1-\pi_2), \ldots, \pi_\mathrm{n}(1-\pi_\mathrm{n})\right)$$

- estimation of π is given by $\hat{\pi}_i = \frac{\exp(\boldsymbol{X}_i\boldsymbol{b})}{1+\exp(\boldsymbol{X}_i\boldsymbol{b})}, \qquad i = 1, 2, \ldots, n$
- Fisher matrix $\boldsymbol{J}(\boldsymbol{b}) = \boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{b})\boldsymbol{X}$
- matrix of variance for $\boldsymbol{b}$ is $\boldsymbol{V} = \boldsymbol{J}^{-1}$
- vector of standard errors of the coefficient estimates $\boldsymbol{b}$ is $se_j = \sqrt{v_{jj}}, \qquad j = 1, 2, \ldots, k$
- $Z = \frac{b_j}{\sqrt{v_{jj}}} \sim N(0,1)|_{H_0:\beta_j=0}$
- interval for $b_j \pm u_{1-\alpha/2} \cdot se_j$

Goodness of fit

- A goodness-of-fit measure is a summary statistic indicating the accuracy with which the model approximates the observed data, like the R^2 measure in the linear regression model.
- Let ℓ_1 denote the maximum loglikelihood value of the model of interest and let ℓ_0 denote the maximum value of the loglikelihood function when all parameters, except the intercept, are set to zero.
- To compute ℓ_0 it is not necessary to estimate a probit or logit model with an intercept term only
  $\ell_0 = \sum y_i \ln(N_1/N) + \sum(1 - y_i)\ln(1 - N_1/N) =$
  $N_1 \ln(N_1/N) + N_0 \ln(N_0/N)$, where $N_1 = \sum y_i$ and $N_0 = N - N_1$
- *pseudo* $R^2 = 1 - \frac{1}{1+2(\ell_1-\ell_0)/N}$
- *McFadden* $R^2 = 1 - \frac{\ell_1}{\ell_0}$
- deviance $D = 2\left(\ell_0 - \ell_1\right)$

Residual I

- **raw residual** is the difference between the actual response and the estimated probability from the mode

$$r_i = y_i - \hat{\pi}_i$$

- **Pearson residual** corrects for the unequal variance in the raw residuals by dividing by the standard deviation

$$p_i = \frac{r_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}$$

- **deviance residuals** are also popular because the sum of squares of these residuals is the deviance statistic

$$d_i = \pm\sqrt{2\left[y_i \ln\left(\frac{y_i}{\hat{\pi}_i}\right) + (1-y_i)\ln\left(\frac{1-y_i}{1-\hat{\pi}_i}\right)\right]}$$

Residual II

- hat matrix $\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}^{1/2}$, where $\boldsymbol{W}^{1/2}$ is diagonal matrix with $\boldsymbol{W}_{ii}^{1/2} = \sqrt{w_{ii}}$ and studentized residuals
$$sp_i = \frac{p_i}{\sqrt{1-h_{i,i}}}$$
- C and $\overline{C}$ are extensions of Cook's distance for logistic regression. C measures the overall change in fitted logits due to deleting the i-th observation for all points including the one deleted, while $\overline{C}$ excludes the deleted point.
$$C_i = \frac{p_i^2 h_{i,i}}{p(1-h_{i,i})^2}, \qquad \overline{C}_i = \frac{p_i^2 h_{i,i}}{p(1-h_{i,i})}$$
- *DFDEV* and *DFCHI* are statistics that measure the change in deviance and in Pearson's chi-square, respectively, that occurs when an observation is deleted from the data set. Large values of these statistics indicate observations that have not been fitted well.
$$DFDEV_i = d_i^2 + \bar{\mathrm{C}}_i, \qquad DFCHI_i = \frac{\bar{\mathrm{C}}_i}{h_{i,i}}$$