

Lecture Note 12

Qualitative Response Models I

Usually regression methods are designed for a continuous variable. In practice we often have to analyze qualitative response, i.e., a discrete dependent variable. For example, decide whether or not to quit a job, whether or not to go for post-graduate studies, etc. Sometime we have to choose among many alternatives, such as how to commute to works, or choose sequentially among various options.

Now, we cannot simply assume that the error term ε has a continuous pdf, because y does not have a continuous pdf.

Normally we have to write the likelihood function, which make the estimation of IV and panel data models quite complicated, since the likelihood function is, normally, non-linear.

Typically, these models are estimated using either Non-Linear Least Squares (NLLS), GMM, or MLE. We will be making some distributional assumptions and therefore need to be aware of the fact that the parameter estimates will be sensitive to the choices we make.

Typically, our models will be of the form

$$\begin{aligned}y_i &= E[y_i|x_i] + \varepsilon_i \\ &= g(x_i; \beta) + \varepsilon_i \\ &= g(x_i'\beta) + \varepsilon_i,\end{aligned}$$

where $g(\cdot)$ is a known non-linear function. A NLLS is an estimator for β that solves

$$\hat{\beta}_n = \arg \min_{\beta \in B} \frac{1}{n} \sum (y_i - g(x_i'\beta))^2.$$

Binary Choice Models:

Linear probability model (LPM):

Here we assume that we have the usual linear regression model, even though $y_i \in \{0, 1\}$. Consequently, the interpretation of

$$E[y_i|x_i] = x_i'\beta,$$

being the probability that an event occurs breaks down when $x'_i \hat{\beta} \notin [0, 1]$.

The advantage of the LMP is that it allows one to easy handle IV estimation easily. However, the application of the LPM is valid only when $y_i = x'_i \hat{\beta}$ is not close to either 0 or 1. A way to allow for possible non-linearity of the conditional probabilities is by including a number of polynomials in the x 's.

Now we write the model as

$$y_i = x'_i \beta + \varepsilon_i, \quad i = 1, \dots, n.$$

Note that in this setup, since $y_i \in \{0, 1\}$, it must be that the error follow a very restrictive distribution, that is,

$$\varepsilon_i = \begin{cases} 1 - x'_i \beta & \text{if } y_i = 1, \\ x'_i \beta & \text{if } y_i = 0. \end{cases}$$

In fact, y_i is a Bernoulli random variable, with a pdf given by

$$f(y_i; x_i, \beta) = \Pr(y_i = 1 | x_i, \beta)^{y_i} (1 - \Pr(y_i = 1 | x_i, \beta))^{1-y_i}.$$

Hence, the conditional variance of ε_i , conditional on x_i , is given by

$$\begin{aligned} \text{Var}(y_i | x_i) &= \text{Var}(\varepsilon_i | x_i), \\ &= \Pr(y_i = 1 | x_i, \beta) (1 - \Pr(y_i = 1 | x_i, \beta)). \end{aligned}$$

This implies that for the linear probability model we have

$$\text{Var}(\varepsilon_i | x_i) = x'_i \beta (1 - x'_i \beta),$$

That is, the linear probability model implies that the variance of ε_i , conditional on x_i , is heteroskedastic.

Logit, Probit and MLE:

The MLE method translate the discrete dependent variable into a continuous domain, using cdf's, since it is always the case that any cdf $F(\cdot)$, satisfies $F(\cdot) \in [0, 1]$.

Often, the binary choice model is derived from an underlying behavioral assumptions (e.g. a woman will choose to work for pay if the utility she derives from working is larger than the utility from not working). This leads to a latent variable representation of the model. Assuming a linear additive relationship we obtain for the utility difference, denoted y_i^* , that

$$y_i^* = x'_i \beta + u_i, \tag{12.1}$$

and we only get to observe whether or not a woman participate in the labor force, i.e.,

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Then, for a symmetric $F(\cdot)$ we have

$$\begin{aligned} \Pr(y_i = 1|x_i) &= \Pr(u_i > -x_i'\beta), \\ &= 1 - F(-x_i'\beta), \\ &= F(x_i'\beta). \end{aligned}$$

The two most common choices are the logit distribution, i.e.,

$$F(x_i'\beta) = \Lambda(x_i'\beta) = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}},$$

and the standard normal (probit model) distribution, i.e.,

$$F(x_i'\beta) = \Phi(x_i'\beta).$$

The sample likelihood is built then under random sampling.

Recall that we construct the sample analog of $E(\ln f(x_i, \theta))$, where $f(x_i, \theta)$ is the individual likelihood contribution, and maximize it w.r.t. θ .

Random sampling guarantees that

$$\frac{1}{n} \sum_{i=1}^n \ln f(x_i, \theta) \xrightarrow{p} E(\ln f(x_i, \theta)).$$

If we have to explain these types of variables a linear regression model is generally inappropriate, our models are intrinsically non-linear models.

Identification:

Suppose now that for the error term u_i in (12.1) we have

$$u_i \sim N(0, \sigma_\varepsilon^2).$$

Then

$$F(x_i'\beta) = \Phi\left(x_i' \frac{\beta}{\sigma_\varepsilon}\right),$$

that is, we can only identify is β/σ_ε . This is due to the fact that we observe a limited set of the latent variable: $y_i = \tau(y_i^*)$. In consequence, in the binary choice example σ_ε is not identified. The

rationale for this is that by observing y_i we only know whether y_i^* exceeds the threshold or not, but there is no way to find the scale of y_i^* . In the sequel we assume that $\sigma_\varepsilon^2 = 1$.

Estimation:

If we assume that $F(\cdot)$ is known, then the optimal parametric estimator for this problem is the ML, i.e.,

$$\hat{\beta}_n = \arg \min_{\beta \in B} \frac{1}{n} \sum_{i=1}^n \ln f(y_i | x_i, \beta),$$

where B denotes the parameter space.

As mentioned above, y_i is a Bernoulli random variable, with the pdf

$$\begin{aligned} f(y_i; x_i, \beta) &= \Pr(y_i = 1 | x_i, \beta)^{y_i} (1 - \Pr(y_i = 1 | x_i, \beta))^{1-y_i} \\ &= F(x_i' \beta)^{y_i} (1 - F(x_i' \beta))^{1-y_i}, \end{aligned}$$

if $F(\cdot)$ is a symmetric distribution.

Therefore, the log-likelihood is

$$l(\beta) = \sum_{i=1}^n \{y_i \ln F(x_i' \beta) + (1 - y_i) \ln (1 - F(x_i' \beta))\}.$$

In the probit model we have

$$l(\beta) = \sum_{i=1}^n \{y_i \ln \Phi(x_i' \beta) + (1 - y_i) \ln (1 - \Phi(x_i' \beta))\},$$

while in the logit model we have

$$l(\beta) = \sum_{i=1}^n \left\{ y_i \ln \frac{e^{x_i' \beta}}{1 + e^{x_i' \beta}} + (1 - y_i) \ln \frac{1}{1 + e^{x_i' \beta}} \right\}.$$

Probit or Logit MLE estimation will involve solving the FOC using numerical optimization procedures. Both problems are globally concave in the parameters, so we will always obtain the global optima (convergence usually is achieved quite fast). From computational aspect, the probit estimation is more demanding, since in the probit setting the probabilities are one dimensional integrals, whereas in the logit the probabilities have very simple expressions.

Remarks:

1. The (0,1) labelling is innocuous.
2. Because the logistic distribution has a variation of $\pi^2/3$, the estimates for β obtained from the logit model have to be multiplied by $\sqrt{3}/\pi$ to make them comparable the probit estimates (where σ_ε^2 is normalized to 1).

In our binary choice model (also called the univariate *dichotomous* model) it does not matter much whether one uses a probit or logit model, except in cases where the data are heavily concentrated in the tails. This is because the two distributions normal/logistic are fairly similar, except for the fatter tails of the logistic distribution. In *multi-response of multivariate models* the logit and probit models differ substantially.

Interpretation:

Apart from the sign of the coefficients, the coefficients in these binary choice models are not easily interpretable. Except maybe in the logit model, where one can consider the β 's to represent the marginal effect of x_{ki} on the log of the odds ratio (*OR*):

$$OR = \frac{p_{1i}}{1 - p_{1i}}, \quad \text{where } p_{1i} = \frac{e^{x'_i\beta}}{1 + e^{x'_i\beta}}.$$

Hence,

$$\begin{aligned} \ln OR &= \ln \left[\frac{e^{x'_i\beta} / 1 + e^{x'_i\beta}}{1 / 1 + e^{x'_i\beta}} \right], \\ &= \ln \left[e^{x'_i\beta} \right], \\ &\quad x'_i\beta. \end{aligned}$$

One way to interpret the parameters (and to ease comparisons across different models) is to look at the derivatives of the probabilities with respect to a particular independent variable (continuous variables). We can analyze the effect of a dummy variable by comparing the probabilities that result when the variables takes its two, while evaluating the probabilities at the sample means for all other variables.

Predicted marginal effects:

The predicted marginal effect for the probit model is given by

$$\frac{\partial}{\partial x_{ki}} \widehat{\Pr}(y_i = 1 | x_i, \beta) = \phi(x'_i \widehat{\beta}_n) \widehat{\beta}_{nk}, \quad \text{for } k = 1, \dots, K,$$

while for the logit model it is given by

$$\frac{\partial}{\partial x_{ki}} \widehat{\Pr}(y_i = 1 | x_i, \beta) = \frac{e^{x'_i \widehat{\beta}_n}}{(1 + e^{x'_i \widehat{\beta}_n})^2} \widehat{\beta}_{nk}, \quad \text{for } k = 1, \dots, K.$$

Typically, these marginal effect are computed at the mean of the independent variables. The delta method can be used then to obtain standard errors for the marginal effects.

Recall that the estimator $\widehat{\beta}_n$ is a consistent estimator for β , with

$$\sqrt{n} (\widehat{\beta}_n - \beta) \xrightarrow{D} N(0, I^{-1}(\beta)).$$

Hence, for any continuously differentiable function $h(\cdot)$ at β we have

$$\sqrt{n} (h(\widehat{\beta}_n) - h(\beta)) \xrightarrow{D} N(0, H(\beta)' I^{-1}(\beta) H(\beta)),$$

where

$$H(\beta)' \equiv \frac{\partial h(\beta)}{\partial \beta'}.$$

Note that one can also evaluate the elasticities in the same fashion.

Multi-response Models:

For many applications, the number of alternatives that can be chosen is larger than two. For example, an individual can choose different employment levels: full-time, part-time, or not working. Another example is a choice of mode of transportation: bus, train, car, cycle, or walking. Within multi-response model we distinguish between *ordered* and *unordered* response models. An ordered response model requires a logical ordering of the alternatives, such as fertility outcomes, level of insurance etc.

Ordered Response Model:

You have already seen one approach to ordered outcomes: The Poisson Regression Model. An alternative approach is an Ordered Probit or Ordered Logit Models.

The model specification is as follows:

$$y_i^* = x_i' \beta + \varepsilon_i$$

is a latent variable. Let

$$-\infty < \mu_1 < \mu_2 < \dots < \mu_{J-1} < \infty$$

be $J - 1$ thresholds. These thresholds are *unknown parameter* that need to be estimated along with β . Let $\mu' = (\mu_1, \dots, \mu_{J-1})$.

Now define

$$y_i = j \quad \text{if } \mu_j < y_i^* < \mu_{j+1}, \quad \text{for } j = 0, 1, \dots, J.$$

Then,

$$\begin{aligned} \Pr(y_i = j | x_i) &= \Pr(\mu_j < y_i^* < \mu_{j+1} | x_i) \\ &= \Pr(\varepsilon_i < \mu_{j+1} - x_i' \beta | x_i) - \Pr(\varepsilon_i \leq \mu_j - x_i' \beta | x_i). \end{aligned}$$

If we assume that

$$\varepsilon_i | x_i \sim N(0, \sigma_\varepsilon^2),$$

then the model is called the *ordered probit* model. Since the latent variable y_i^* is only partially observed, we normalized σ_ε^2 to equal 1.

Define now

$$z_{ji} = \begin{cases} 1 & \text{if } y_i = j, \\ 0 & \text{otherwise.} \end{cases}$$

Then the log-likelihood function is given by

$$l(\beta, \mu) = \sum_{i=1}^n \left\{ \sum_{j=0}^J z_{ji} \ln [\Phi(\mu_{j+1} - x_i' \beta) - \Phi(\mu_j - x_i' \beta)] \right\}.$$

Unordered Response Model:

The natural framework within which to consider choice models is that of *Random Utility maximization* (RUM). The utility is random only for the econometrician, but it is not random from the point of view of the individual decision maker. This is in order to accommodate internal inconsistencies whereby two individuals with the same observed variables make different choices.

There is a choice set with J *mutually exclusive and exhaustive* alternatives, where each alternative j yields the utility

$$y_{ji}^* = x_{ji}' \beta_j + \varepsilon_{ji}.$$

The RUM hypothesis: Choose the k th alternative if

$$k = \arg \max \{y_{1i}^*, y_{2i}^*, \dots, y_{Ji}^*\}.$$

For convenience, denote $y_i^* = (y_{1i}^*, y_{2i}^*, \dots, y_{Ji}^*)'$.

Then,

$$\begin{aligned} \Pr(y_i = k) &= \Pr(y_{ki}^* > y_{ji}^* \text{ for all } j \neq k) \\ &= \Pr(x_{ji}' \beta_k + \varepsilon_{ji} > x_{ji}' \beta_j + \varepsilon_{ji} \text{ for all } j \neq k). \end{aligned}$$

Remarks:

1. We need not worry about ties, because they have zero probability of occurring assuming that $\varepsilon_i' = (\varepsilon_{1i}, \dots, \varepsilon_{Ji})$ has a continuous distribution.

2. Depending on the specific distributional assumptions about ε_i we will get quite different implied properties for the discrete choices.
3. For a discrete choice problem of dimension J there is always an equivalent formulation of order $J - 1$.

An example for $j = 2$:

$$\begin{aligned} y_{1i}^* &= x'_{1i}\beta_1 + \varepsilon_{1i}, & \text{the utility from choice 1,} \\ y_{2i} &= x'_{2i}\beta_2 + \varepsilon_{2i}, & \text{the utility from choice 2.} \end{aligned}$$

Hence,

$$\begin{aligned} y_i &= 1 & \text{if } y_{1i}^* \geq y_{2i}^*, \\ y_i &= 2 & \text{if } y_{1i}^* < y_{2i}^*. \end{aligned}$$

The difference in the utility is given then by

$$\begin{aligned} w_i^* &= y_{1i}^* - y_{2i}^*, \\ &= x'_{1i}\beta_1 - x'_{2i}\beta_2 + \varepsilon_{1i} - \varepsilon_{2i}, \\ &= z'_i\gamma + u_i, \end{aligned}$$

where

$$u_i = \varepsilon_{1i} - \varepsilon_{2i},$$

and z_i contains all the variables that are in x_{1i} and x_{2i} (without duplications).

Define

$$w_i = \begin{cases} 1 & \text{if } w_i^* \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Note that here, y_i^* and ε_i are J -dimensional vectors, while w_i^* and u_i are only $(J - 1)$ -dimensional vectors. Also, what matters here is which of the two alternative gives the highest utility, but the exact level of the utility does not matter.

Hence,

$$\begin{aligned} \Pr(y_i = k) &= \Pr(\varepsilon_{ki} - \varepsilon_{ji} > x'_{ji}\beta_j - x'_{ki}\beta_k, \text{ for all } j \neq k) \\ &= \Pr(u_{ji} > z'_{ji}\gamma_j, \text{ for all } j \neq k). \end{aligned}$$

Distributional assumptions:

There are typically two leading distributional assumptions, that give rise to the following two models:

1. *Multinomial probit* (MNP) model, where

$$\varepsilon_i \sim N(0, \Omega).$$

2. *Multinomial logit* (MNL) model, where

$$\varepsilon_i \sim \text{i.i.d. extreme value (Type I) distribution.}$$

The MNL model has the nice feature that the choice probabilities are given by the simple form

$$p_j(x_i) = \Pr(y_i = j | x_{1i}, \dots, x_{Ji}) = \frac{\exp\{x'_{ji}\beta_j\}}{\sum_{l=1}^J \exp\{x'_{li}\beta_l\}}.$$

For the MNP, in contrast, $p_j(x_i)$ is a $J - 1$ integral, over the joint distribution of u_i . Typically, for $J > 3$, the MNP becomes difficult to estimate, unless one is willing to restrict the covariance matrix. A typical estimation would require the use of some simulation methods for calculating the $J - 1$ integrals needed for obtaining the conditional probabilities.

However, the MNL model exhibits the *Independence of Irrelevant Alternatives* (IIA) property, which amounts to the assumption of equally substitutable alternatives, conditional on the x 's. The “relative odds” of choosing alternative j over alternative l only depends on the characteristics of the j th and l th alternatives, but does not depend on the characteristics of the other “irrelevant” alternatives. That is,

$$\begin{aligned} \frac{p_j(x_i)}{p_l(x_i)} &= \frac{\exp\{x'_{ji}\beta_j\} / \sum_{m=1}^J \exp\{x'_{mi}\beta_m\}}{\exp\{x'_{li}\beta_l\} / \sum_{m=1}^J \exp\{x'_{mi}\beta_m\}}, \\ &= \frac{\exp\{x'_{ji}\beta_j\}}{\exp\{x'_{li}\beta_l\}}. \end{aligned}$$

Note that if an individual is faced with a new alternative relative odds do not change. Typically, this model predicts too high a joint probability of selection for 2 alternatives that are, in fact, perceived as similar, rather than independent. Classic illustration: McFadden blue bus/red bus problem. It is possible to relax the IIA property but this generally leads to more complicated models (see, for example, Amemiya (1981) or Maddala (1983)).

Example: *Nested Multinomial Logit* Model (McFadden 1981)—Housing Choice.

A hierarchical elimination model based on generalized extreme value structure generalizes the MNL model to a nested multinomial logit structure.

The structure:

Decision 1	\implies	Decision 2	\implies	Housing outcome
Own a house	\implies		\implies	1 = Own a house
Rent a house	\implies	Head	\implies	2 = Rent alone
		Share	\implies	2 = Rent shared house

The implied probabilities are calculated backward, as follows: (i) start at end and calculate the binary logit conditional on renting; (ii) compute an expected utility measure (inclusive value); and (iii) use that value to compute the probabilities of owning and renting a house.

The probabilities:

$$\begin{aligned}
 \Pr(\text{Head}|\text{Rent}) &= \frac{\exp\{x'_2\beta_2\}}{\exp\{x'_2\beta_2\} + \exp\{x'_3\beta_3\}}, \\
 \Pr(\text{Share}|\text{Rent}) &= \frac{\exp\{x'_3\beta_3\}}{\exp\{x'_2\beta_2\} + \exp\{x'_3\beta_3\}}, \\
 \Pr(\text{Own a house}) &= \frac{\exp\{x'_1\beta_1\}}{\exp\{x'_1\beta_1\} + \exp\{I_{23}\lambda_{23}\}}, \\
 \Pr(\text{Rent}) &= \frac{\exp\{I_{23}\lambda_{23}\}}{\exp\{x'_1\beta_1\} + \exp\{I_{23}\lambda_{23}\}},
 \end{aligned}$$

where I_{23} is the *inclusive value*, satisfying

$$I_{23} = \ln[\exp\{x'_2\beta_2\} + \exp\{x'_3\beta_3\}].$$

The contribution to the likelihood function of the i th individual is obtained then by

$$\begin{aligned}
 \Pr(y_i = 1|x_i) &= \Pr(\text{Own a house}|x_i), \\
 &= \frac{\exp\{x'_1\beta_1\}}{\exp\{x'_1\beta_1\} + \exp\{I_{23}\lambda_{23}\}}, \\
 \Pr(y_i = 2|x_i) &= \Pr(\text{Rent}|x_i) \Pr(\text{Head}|\text{Rent}, x_i), \\
 &= \frac{\exp\{I_{23}\lambda_{23}\}}{\exp\{x'_1\beta_1\} + \exp\{I_{23}\lambda_{23}\}} \cdot \frac{\exp\{x'_2\beta_2\}}{\exp\{x'_2\beta_2\} + \exp\{x'_3\beta_3\}}, \\
 \Pr(y_i = 2|x_i) &= \Pr(\text{Rent}|x_i) \Pr(\text{Share}|\text{Rent}, x_i), \\
 &= \frac{\exp\{I_{23}\lambda_{23}\}}{\exp\{x'_1\beta_1\} + \exp\{I_{23}\lambda_{23}\}} \cdot \frac{\exp\{x'_3\beta_3\}}{\exp\{x'_2\beta_2\} + \exp\{x'_3\beta_3\}}.
 \end{aligned}$$

Remarks:

1. Note that if $\lambda_{23} = 1$ then

$$p_j(x_i) = \frac{\exp \{x'_{ji}\beta_j\}}{\sum_{l=1}^J \exp \{x'_{li}\beta_l\}},$$

that is, all the alternative are equally substitutable. The nesting does not matter, and we get the simple MNL model.

2. If $\lambda_{23} \neq 1$ then the IIA property holds only within a node. That is, there is higher substitution within a node than across nodes.

Estimation:

1. *Full MLE*—Quite hard numerically, because the likelihood is not concave in all the parameters and is highly non-linear in the inclusive value coefficient vectors.
2. *Sequential method “Folding Back”*—In the above example, only use data on renters and perform binary logit to get \hat{I}_{23} . Then use the whole sample and \hat{I}_{23} to estimate β_1 and λ_{23} . This method is less efficient than the MLE. Also, one need to correct for the standard errors beyond the first stage, because we use \hat{I}_{23} , an estimate of I_{23} , rather than I_{23} itself, in the second stage.