

LECTURE 9

QUALITATIVE RESPONSE (QR) MODELS

In many models that arise in practice the dependent variable is a categorical variable that takes discrete values. Such models are called **qualitative response (QR)** or **discrete choice** or **quantal** or **categorical models**. A QR model is called a **binary response** (or **binary choice** or **threshold crossing**) **model** if the dependent variable can take on only two values (which can be taken to 0 and 1 without loss of generality). A QR model is called a **multinomial response** model if the dependent variable can take on more than two values. See Amemiya (1985) for a review of these models.

1. THE BINARY CHOICE MODEL

In this model the dependent variable Y_i takes on two values (0 and 1) i.e. it is a **Bernoulli** variable. We typically parameterize the probability of “success”, i.e. the probability of obtaining Y_i equal to 1, by a certain *known* function that depends on a $(1 \times K)$ vector of conditioning variables (or independent variables or regressors) X_i through a finite-dimensional vector of unknown parameters β_0 . There are at least 3 well-known and much used models:

1. Linear Probability Model (LPM):

$$\Pr(Y_i = 1|X_i) = X_i\beta_0$$

2. Probit Model:

$$\Pr(Y_i = 1|X_i) = \Phi(X_i\beta_0) = \int_{-\infty}^{X_i\beta_0} \phi(u) du = \int_{-\infty}^{X_i\beta_0} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) du$$

where Φ and ϕ are the cdf and pdf of the standard normal distribution.

3. Logit Model:

$$\Pr(Y_i = 1|X_i) = \Lambda(X_i\beta_0) = \frac{\exp(X_i\beta_0)}{1 + \exp(X_i\beta_0)}$$

where Λ is the cdf of the logistic distribution. Recall that the logistic distribution is a symmetric around 0 distribution that has variance equal to $\pi^2/3$.

4. Log-Weibull Model:

$$\Pr(Y_i = 1|X_i) = \exp(-\exp(X_i\beta_0))$$

Note that although the last three models constrain the probabilities to be between 0 and 1 this is not the case with the LPM. Also note that both the probit and the logit model utilize distributions that are symmetric around 0 whereas the Weibull model uses a non-symmetric distribution.

1.1. Latent Variable Threshold Crossing Interpretation of Binary Choice Model

Let y_i^* denote an unobservable (**latent variable**) generated as

$$y_i^* = x_i\beta_0 + \varepsilon_i \tag{1}$$

Typically we will assume that $-\varepsilon_i$ is distributed according to a *known* cdf F conditional on x_i . Usually F will be a continuous and symmetric (so that ε_i will have the same distribution as $-\varepsilon_i$) around zero.¹ Frequently, ε_i and x_i are assumed independent (so that F is the conditional and unconditional distribution of ε_i), as in the case of the logit and probit models.

Although we observe x_i for all i , we do not observe y_i^* but only whether it crosses a certain threshold, which is known and fixed for all individuals, say c . Since x_i will typically contain the constant regressor, we may without loss of generality take c to be zero. In other words we observe x_i and

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

Defining the **indicator function** $1\{A\}$ to be equal to 1 if event A happens and zero otherwise, the observed dependent variable is

$$y_i = 1\{y_i^* \geq 0\} = 1\{x_i\beta_0 + \varepsilon_i \geq 0\}$$

Then

$$\begin{aligned} \Pr(y_i = 1|x_i) &= \Pr(y_i^* \geq 0|x_i) \\ &= \Pr(x_i\beta_0 + \varepsilon_i \geq 0|x_i) \\ &= \Pr(-\varepsilon_i \leq x_i\beta_0|x_i) \\ &= F(x_i\beta_0) \end{aligned}$$

F can be for example either the standard normal or the logistic cdf.

Identification

The intercept in the latent regression model, say β_{01} , is identified only if the common threshold c is known to the econometrician. If c is not known, then *the intercept cannot be separately identified from the threshold*. In other words, in this case we can only identify $\beta_{01} - c$.

Furthermore, the scale of β_0 is not identified separately from the scale (variance) of the error term. This is due to the fact that the sign of y_i^* and hence y_i do not change when we multiply the model by a positive constant. Consider for example two probit models,

$$\begin{aligned} y_i^* &= x_i\gamma_0 + \varepsilon_{i1} \\ y_i^* &= x_i\delta_0 + \varepsilon_{i2} \end{aligned}$$

¹Note that the zero mean assumption is innocuous as long as X_i contains the constant regressor 1.

where ε_{i1} (and hence $-\varepsilon_{i1}$) is distributed $N(0, \sigma_1^2)$ and ε_{i2} (and hence $-\varepsilon_{i2}$) is distributed $N(0, \sigma_2^2)$, and where $\gamma_0/\sigma_1 = \delta_0/\sigma_2$. These models are easily seen to be **observationally equivalent**, since they both imply the same (conditional) probability of success. Indeed, in the first model

$$\begin{aligned} \Pr(y_i = 1|x_i) &= \Pr(x_i\gamma_0 \geq -\varepsilon_{i1}|x_i) \\ &= \Pr(x_i\gamma_0 \geq -\varepsilon_{i1}) \\ &= \Pr\left(\frac{x_i\gamma_0}{\sigma_1} \geq -\frac{\varepsilon_{i1}}{\sigma_1}\right) \\ &= \Phi\left(x_i\frac{\gamma_0}{\sigma_1}\right) \end{aligned}$$

which is obviously the same as the conditional probability in the second model

$$\begin{aligned} \Pr(y_i = 1|x_i) &= \Pr(x_i\delta_0 \geq -\varepsilon_{i2}|x_i) \\ &= \Pr(x_i\delta_0 \geq -\varepsilon_{i2}) \\ &= \Pr\left(\frac{x_i\delta_0}{\sigma_2} \geq -\frac{\varepsilon_{i2}}{\sigma_2}\right) \\ &= \Phi\left(x_i\frac{\delta_0}{\sigma_2}\right) \end{aligned}$$

In other words, *in the binary choice model β_0 is only identified up to scale*. This is why we typically fix the scale of ε_i ; in the probit model we typically assume that ε_i is a standard normal, while in the logit model the variance is fixed at $\pi^2/3$.

1.2. Random Utility Interpretation of Binary Choice Model

Let the utility of person i from choosing choice j ($j = 0, 1$) be:

$$U_{ij} = x_{ij}\beta_0 + \varepsilon_{ij}$$

where ε_{ij} is a continuous scalar RV. Then

$$\begin{aligned} \Pr(y_i = 1|x_{i1}, x_{i0}) &= \Pr(U_{i1} > U_{i0}|x_{i1}, x_{i0}) \\ &= \Pr(x_{i1}\beta_0 + \varepsilon_{i1} > x_{i0}\beta_0 + \varepsilon_{i0}|x_{i1}, x_{i0}) \\ &= \Pr(-(\varepsilon_{i1} - \varepsilon_{i0}) < (x_{i1} - x_{i0})\beta_0|x_{i1}, x_{i0}) \end{aligned}$$

Let F denote the distribution of $-(\varepsilon_{i1} - \varepsilon_{i0})$ conditional on x_{i1}, x_{i0} . Then

$$\begin{aligned} \Pr(y_i = 1|x_{i1}, x_{i0}) &= F(x_i\beta_0) \\ \Pr(y_i = 0|x_{i1}, x_{i0}) &= 1 - F(x_i\beta_0) \end{aligned}$$

It is clear that we can recast this model as a latent variable model where $y_i^* = U_{i1} - U_{i0}$, $x_i = x_{i1} - x_{i0}$ and $\varepsilon_i = \varepsilon_{i1} - \varepsilon_{i0}$. Note that if ε_{ij} 's are jointly normal (independent or not) then their difference is also normal. Thus we obtain the probit model. If the ε_{ij} 's are independent Type I Extreme Value

(or log Weibull) distributed² RV's then their difference is distributed as logistic. Thus we obtain the logit model.

1.3. Estimation

1.3.1. Nonlinear Least Squares and Weighted Nonlinear Least Squares

Since y_i is Bernoulli with probability of success $F(x_i\beta_0)$, its CEF given x_i is:

$$E(y_i|x_i) = 1 \cdot F(x_i\beta_0) + 0 \cdot (1 - F(x_i\beta_0)) = F(x_i\beta_0)$$

Defining

$$u_i \equiv y_i - E(y_i|x_i)$$

we obtain the nonlinear regression model

$$y_i = F(x_i\beta_0) + u_i$$

where u_i satisfies by construction the mean independence restriction

$$\begin{aligned} E(u_i|x_i) &= E(y_i - E(y_i|x_i)) \\ &= E(E[y_i - E(y_i|x_i) | x_i]) \\ &= E(E(y_i|x_i) - E(y_i|x_i)) \\ &= 0 \end{aligned}$$

The above discussion leads naturally to the NLS estimator of β_0 , defined as

$$\hat{\beta}_{NLS} = \arg \min_{\beta} \sum_i (y_i - F(x_i\beta))^2$$

Under appropriate regularity conditions, $\hat{\beta}_{NLS}$ can be shown to be consistent and asymptotically normal.

Notice that the new error term u_i in the model above has a binomial distribution conditional on x_i :

$$u_i = \begin{cases} 1 - F(x_i\beta_0) & (\text{if } y_i = 1) & \text{with probability } F(x_i\beta_0) \\ -F(x_i\beta_0) & (\text{if } y_i = 0) & \text{with probability } 1 - F(x_i\beta_0) \end{cases}$$

and hence its conditional variance is

$$\begin{aligned} \text{Var}(u_i|x_i) &= [1 - F(x_i\beta_0)]^2 \cdot F(x_i\beta_0) + [-F(x_i\beta_0)]^2 (1 - F(x_i\beta_0)) \\ &= F(x_i\beta_0) \cdot (1 - F(x_i\beta_0)) \end{aligned}$$

²The cdf of ε that is distributed as a log Weibull is:

$$F(c) = \Pr(\varepsilon \leq c) = \exp(-\exp(-c))$$

i.e. u_i is conditionally heteroskedastic. Therefore, the WNLS estimator defined as

$$\hat{\beta}_{WNLS} = \arg \min_{\beta} \sum_i \frac{(y_i - F(x_i\beta))^2}{F(x_i\beta_0) \cdot (1 - F(x_i\beta_0))}$$

will be asymptotically more efficient than NLS. The WNLS estimator solves the FOC:

$$-\sum_i \frac{2(y_i - F(x_i\beta))}{F(x_i\beta_0) \cdot (1 - F(x_i\beta_0))} f(x_i\beta) x'_i = 0$$

where $f(\cdot)$ is the density corresponding to F .

In order to implement WNLS we need to estimate the weights $F(x_i\beta_0) \cdot (1 - F(x_i\beta_0))$ in a first stage consistently. Since NLS is consistent for β_0 we may for example use it to construct consistent estimates of the weights.

1.3.2. Maximum Likelihood Estimation of the Binary Choice Model

Since y_i is Bernoulli with probability of success $F(x_i\beta_0)$ its density conditional on x_i is

$$f(y_i|x_i) = F(x_i\beta_0)^{y_i} (1 - F(x_i\beta_0))^{1-y_i}$$

The likelihood function for a random sample of n observations is

$$L(\beta) = \prod_{i=1}^n F(x_i\beta)^{y_i} (1 - F(x_i\beta))^{1-y_i}$$

and the log-likelihood is

$$\mathcal{L}(\beta) = \ln L(\beta) = \sum_i [y_i \ln F(x_i\beta) + (1 - y_i) \ln (1 - F(x_i\beta))]$$

The ML estimator of β_0 is defined as

$$\hat{\beta}_{ML} = \arg \max_{\beta} \sum_i [y_i \ln F(x_i\beta) + (1 - y_i) \ln (1 - F(x_i\beta))]$$

and therefore solves the FOC

$$\begin{aligned} \sum_i \left[\frac{y_i}{F(x_i\beta)} f(x_i\beta) x'_i - \frac{(1 - y_i)}{(1 - F(x_i\beta))} f(x_i\beta) x'_i \right] &= 0 \Leftrightarrow \\ \sum_i \frac{y_i - y_i \cdot F(x_i\beta) - F(x_i\beta) + y_i \cdot F(x_i\beta)}{F(x_i\beta) (1 - F(x_i\beta))} f(x_i\beta) x'_i &= 0 \Leftrightarrow \\ \sum_i \frac{(y_i - F(x_i\beta))}{F(x_i\beta) \cdot (1 - F(x_i\beta))} f(x_i\beta) x'_i &= 0 \end{aligned}$$

Comparing these to the FOC of the WNLS estimator we see that they are asymptotically equivalent; hence WNLS is as efficient as ML in the binary choice (subject of course to regularity conditions).

1.3.3. Interpretation of the Estimates

Apart from their sign, the coefficients in these binary models are not easily interpretable, except maybe in the logit model, where one can consider the $\beta_{0,k}$ to represent the marginal effect of x_{ik} on the log of the **odds ratio (OR)**:

$$OR = \frac{\Pr(y_i = 1|x_i)}{1 - \Pr(y_i = 1|x_i)}$$

In the logistic case where

$$\Pr(y_i = 1|x_i) = \frac{\exp(x_i\beta_0)}{1 + \exp(x_i\beta_0)}$$

we have that

$$\begin{aligned} \ln OR &= \ln \left(\frac{\frac{\exp(x_i\beta_0)}{1+\exp(x_i\beta_0)}}{\frac{1}{1+\exp(x_i\beta_0)}} \right) \\ &= \ln(\exp(x_i\beta_0)) \\ &= x_i\beta_0 \end{aligned}$$

One way to put the estimated parameters in use (and to ease comparisons across different models) is to look at the derivatives of the probabilities with respect to a particular regressor (in the case of a continuous x_{ik}) or to the difference of the probabilities at two values of a discrete regressor. In either case the probabilities are typically evaluated at the sample averages of the regressors. Alternatively, these marginal effects are evaluated for each individual in the sample and then averaged.

In the case of a continuous regressor x_{ik} , the predicted marginal effect of x_{ik} for individual i is typically computed as

$$\frac{\partial \widehat{\Pr}(y_i = 1|x_i)}{\partial x_{ik}} = f(\bar{x}\hat{\beta})\hat{\beta}_k$$

where \bar{x} denotes the vector of sample averages of the regressors, while the effect of a binary regressor taking on two values 0 and 1 is

$$\widehat{\Pr}(y_i = 1|\tilde{x}_i = \bar{x}, x_{ik} = 1) - \widehat{\Pr}(y_i = 1|\tilde{x}_i = \bar{x}, x_{ik} = 0)$$

where the vector \tilde{x}_i contains all but the k 'th regressor.

Elasticities may be computed similarly. Standard errors for either the predicted probabilities, marginal effects or elasticities can be computed using the delta-method.

1.3.4. Computational Aspects

1. The (0, 1) labelling of choices is innocuous.

2. Because the logistic distribution has variance equal to $\pi^2/3$ the logit estimates have to be multiplied by $\sqrt{3}/\pi (= 0.5513)$ to be comparable in scale to those resulting from a probit model (where the variance of the error term ε_i is assumed to be 1). The estimates from the two models should be close unless the data are heavily concentrated in the tails.³
3. The log-likelihood functions in both the logistic and the probit model are globally concave with respect to β and therefore maximizing them is computationally easy.

2. MULTINOMIAL CHOICE MODELS

Suppose the dependent variable Y_i takes $(J_i + 1)$ values, say $0, 1, 2, \dots, J_i$. The general **multinomial QR model** is defined as

$$\Pr(Y_i = j|X_i) = F_{ij}(X_i, \theta_0) \quad i = 1, \dots, n; \quad j = 1, \dots, J$$

(Strictly speaking we should write j as j_i but we shall suppress the subscript i .) Note that (a) we allow the possibility that not all the independent variables X_i and not all the parameters are included in the argument of every $F_{ij} (\equiv F_{ij}(X_i, \theta_0))$. (b) $\Pr(Y_i = 0|X_i)$ is not specified above because it must equal to one minus the sum of the J_i probabilities defined above. (c) It is important to let J_i to depend on i because in many applications individuals face different choice sets.

Define the following indicator variables

$$Y_{ij} = \begin{cases} 1 & \text{if } Y_i = j \\ 0 & \text{otherwise} \end{cases}$$

for any $j = 0, \dots, J_i$. Then the log-likelihood function is

$$\mathcal{L}(\theta) = \sum_{i=1}^n \sum_{j=1}^{J_i} Y_{ij} \ln(F_{ij}(X_i, \theta))$$

Multinomial models can be classified into **ordered** and **unordered models**. In the ordered model the values that Y_i takes correspond to a partition of the real line whereas in the unordered model they correspond either to a nonsuccessive partition of the real line or to a partition of a higher dimensional space.

³Amemiya (1981) argues that a better approximation emerges if one multiplies the logit estimates by 0.625. Furthermore, he argues that the OLS slope estimates should be approximately equal to 0.25 times the corresponding logit slope parameter estimates, while the intercept and dummy variable coefficient estimates should be approximately equal 0.25 times the corresponding logit estimates plus 0.5.

2.1. Ordered Models

In most applications the **ordered model** takes the form

$$\Pr(Y_i = j|X_i) = F(\alpha_{j+1} - X_i\beta_0) - F(\alpha_j - X_i\beta_0) \quad i = 1, \dots, n; \quad j = 1, \dots, J_i$$

where

$$j = 0, \dots, J; \quad \alpha_0 = -\infty \quad \alpha_j \leq \alpha_{j+1} \quad \alpha_{J+1} = \infty$$

for some distribution function F . If F is the standard normal Φ then the model is called the **ordered probit** model. If F is the logistic cdf then we obtain the **ordered logit** model. It is known that the log likelihood corresponding to the specification above is globally concave if the density corresponding to F , say f , is positive and $\ln f$ is concave.

The ordered model above can be motivated by a latent (i.e. unobservable) continuous random variable Y_i^* given by a linear regression model

$$Y_i^* = X_i\beta_0 + \varepsilon_i$$

that determines the outcome (observed) variable Y_i by the rule

$$Y_i = j \text{ if and only if } \alpha_j < Y_i^* < \alpha_{j+1}$$

for $j = 0, 1, \dots, J$.

We should be cautious in using an ordered model because if the true model is in fact unordered, the ordered model specification can lead to serious biases in the estimation of probabilities. On the other hand, the cost of using an unordered model when the true model is ordered is a loss of efficiency rather consistency.

2.2. Unordered Models

Unordered models may be motivated within the **Random Utility Maximization** framework. Suppose that the utility of each of the $J + 1$ choices is given by a latent regression model

$$U_{ij} = X_i\beta_{0j} + \varepsilon_{ij}$$

where the unobservable errors $\{\varepsilon_{ij}\}_{j=0}^J$ have some joint distribution. Note that here we allow X_i to be different across different j by allowing β_{0j} to be different across j and to contain zeros for whichever alternative the corresponding X_i does not enter. We observe

$$Y_{ij} = 1 \text{ if and only if } U_{ij} = \max\{U_{i0}, U_{i1}, \dots, U_{iJ}\}$$

Note that if the U 's are continuous RV's then the probability of a tie is zero.

Below we discuss two well known and widely used multinomial models.

2.3. The Multinomial Logit (MNL) Model

The **multinomial logit (MNL)** model arises if and only if the underlying errors are *independent and identically distributed* with the Type I extreme value (or log-Weibull) distribution (McFadden (1974)). In this case it can be shown that

$$\Pr(Y_i = j|X_i) = \frac{\exp(X_i\beta_{0j})}{\sum_{j=0}^J \exp(X_i\beta_{0j})}$$

which leads to a globally concave log-likelihood function.

We should point out a restrictive property of the MNL model: The assumption of independence of the ε_{ij} 's implies that the alternative choices are dissimilar. Using McFadden's famous example, suppose that the individual chooses between three alternatives: car, red bus, or blue bus. In such a case the independence between the ε_{blue} and ε_{red} is clearly unreasonable because a high (low) utility for the red bus should generally imply a high (low) utility for the blue bus. The probability of choosing a car, $\Pr(Y_i = car) = \Pr(U_{car} > U_{blue}, U_{car} > U_{red})$, calculated under the independence assumption would underestimate the true probability in this case because the assumption ignores the fact that the event $U_{car} > U_{blue}$ makes the event $U_{car} > U_{red}$ more likely. Furthermore, note that in the multinomial logit model the relative probabilities between a pair of alternatives are specified ignoring the other alternatives. For example, the relative probabilities between car and red bus are specified the same way regardless of whether the third alternative is blue bus or, say, train. In other words

$$\Pr(Y_i = j|Y_i = j \text{ or } k) = \frac{\exp(X_i\beta_{0j})}{\exp(X_i\beta_{0j}) + \exp(X_i\beta_{0k})}$$

McFadden has called this feature of the MNL model the **Independence from Irrelevant Alternatives (IIA)**.

A generalization of the MNL model that alleviates the above shortcoming of the MNL model and which is attributed to McFadden (1977, 1981) is the so-called **Nested Logit Model**. In the example above where there are three alternatives, we may allow for correlation between the two of the alternatives, the blue and the red bus, assuming that they follow the so-called Gumbel's Type B bivariate extreme-valued distribution (see Johnson and Kotz (1972)), which is defined by

$$F(\varepsilon_1, \varepsilon_2) = \exp\left\{-\left[\exp(-\rho^{-1}\varepsilon_1) + \exp(-\rho^{-1}\varepsilon_2)\right]^\rho\right\}$$

and where $corr(\varepsilon_1, \varepsilon_2) = 1 - \rho^2$. If $\rho = 1$ then $corr(\varepsilon_1, \varepsilon_2) = 0$ and the distribution above becomes the product of two independent Type I extreme value distributions. If ε_0 follows the Type I extreme value distribution then we can show that

$$\begin{aligned} \Pr(Y_i = 0|X_i) &= \frac{\exp(X_i\beta_{00})}{\exp(X_i\beta_{00}) + [\exp(\rho^{-1}X_i\beta_{01}) + \exp(\rho^{-1}X_i\beta_{02})]^\rho} \\ \Pr(Y_i = 1|X_i, Y_i \neq 0) &= \frac{\exp(\rho^{-1}X_i\beta_{01})}{\exp(\rho^{-1}X_i\beta_{01}) + \exp(\rho^{-1}X_i\beta_{02})} \end{aligned}$$

The other probabilities can be deduced from the two previous probabilities.

The nested logit model described above can be regarded as implying *two levels of nesting* because the responses are classified into 2 groups (car, bus) and each group is further classified into the individual elements (blue bus, red bus). It is possible to generalize the model for higher levels of nesting (see McFadden (1981)).

2.4. The Multinomial Probit (MNP) Model

The **multinomial probit (MNP)** model arises when the underlying errors are jointly normal. The model has been used little in practice until recently due to its computational difficulty. To demonstrate the complexity of the problem consider the case where $J = 2$, i.e. there are three alternatives. Then, to evaluate $\Pr(Y_i = 2)$ for example, one must calculate the trivariate integral

$$\begin{aligned}\Pr(Y_i = 2) &= \Pr(U_{i2} > U_{i1}, Y_{i2} > U_{i0}) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{U_{i2}} \int_{-\infty}^{U_{i2}} f(v_1, v_2, v_3) dv_1 dv_2 dv_3\end{aligned}$$

where f is the trivariate normal density. The calculation of multiple integrals is in general a difficult task unless one is willing to restrict the covariance matrix.

REFERENCES

- Amemiya, T. (1985): *Advanced Econometrics*. Harvard University Press.
- Johnson, N.L. and S. Kotz (1972): *Distributions in Statistics: Continuous Multivariate Distributions*. New York: John Wiley.
- McFadden, D. (1974): "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka, ed. *Frontiers in Econometrics*. New York: Academic Press.