Chapter 2: Discrete Choice

Joan Llull

Advanced Econometric Methods II
Master in Economics and Finance
Barcelona GSE

I. Binary Outcome Models

A. *Introduction*

In this chapter we analyze several models to deal with discrete outcome variables. These models that predict in which of m mutually exclusive categories the outcome of interest falls. In this section $m = 2$, this is, we consider binary or dichotomic variables (e.g. whether to participate or not in the labor market, whether to have a child or not, whether to buy a good or not, whether to take our kid to a private or to a public school,...). In the next section we generalize the results to multiple outcomes.

It is convenient to assume that the outcome y takes the value of 1 for category A, and 0 for category B. This is very convenient because, as a result, $N^{-1}\sum_{\mathrm{i=1}}^{\mathrm{N}} y_i = \widehat{\Pr}[A \text{ is selected}]$. As a consequence of this property, the coefficients of a linear regression model can be interpreted as marginal effects of the regressors on the probability of choosing alternative A. And, in the non-linear models, it allows us to write the likelihood function in a very compact way.

B. *The Linear Probability Model*

A simple approach to estimate the effect of $\boldsymbol{x}$ on the probability of choosing alternative A is the linear regression model. The OLS regression of y on $\boldsymbol{x}$ provides consistent estimates of the sample-average marginal effects of regressors $\boldsymbol{x}$ on the probability of choosing alternative A. As a result, the linear model is very useful for exploratory purposes. For example, it provides a good guide to which variables are statistically significant, and on which sign is its effect. The treatment effects have turned the popularity of the linear regression model up, as the interest in that case is on the treatment effect of a variable, and not the predicted probability. The linear probability model provides consistent estimates of the difference in the expected outcome with and without a treatment under the relevant assumptions.

However, the linear probability model has two important drawbacks for the analysis of binary outcomes. The first one is that predicted probabilities $\hat{p}(\boldsymbol{x}) = \boldsymbol{x}'\hat{\boldsymbol{\beta}}$ are not bounded between zero and one. The second drawback is that the error term is heteroscedastic and (given $\boldsymbol{x}$) has a discrete support. In particular, $u = -\boldsymbol{x}'\boldsymbol{\beta}$ if $y = 0$ and $u = 1-\boldsymbol{x}'\boldsymbol{\beta}$ if $y = 1$; and its variance is $\boldsymbol{x}'\boldsymbol{\beta}(1-\boldsymbol{x}'\boldsymbol{\beta})$, which depends on $\boldsymbol{x}$. For this reason, in this chapter we review different alternatives to the linear probability model that circumvent these drawbacks.

C. The General Binary Outcome Model

In this section we present a very general framework that nests all models that are covered in this section as special cases. This general notation is useful as many of the results are general across models.

The conditional probability of choosing alternative A given $\boldsymbol{x}$ is given by:

$$p(\boldsymbol{x}) \equiv \Pr[y = 1|\boldsymbol{x}] = F(\boldsymbol{x}'\boldsymbol{\beta}). \tag{1}$$

$F(\boldsymbol{x}'\boldsymbol{\beta})$ is a specified function of $\boldsymbol{x}'\boldsymbol{\beta}$. The models of this class are called *single-index* models, as the argument of the conditional probability function $F(\cdot)$ is a single index of regressors, $\boldsymbol{x}'\boldsymbol{\beta}$. In the linear probability model, the specified function $F(\cdot)$ is the identity, i.e. $F(\boldsymbol{x}'\boldsymbol{\beta}) = \boldsymbol{x}'\boldsymbol{\beta}$. In the other cases, However, it is natural to specify $F(\cdot)$ to be a cumulative distribution function (*cdf*) to ensure that $0 \leq p \leq 1$. This is the case of the *Logit* and the *Probit*, that assume respectively the logistic and the standard normal cdfs:

$$\text{Logit:} \quad F(\boldsymbol{x}'\boldsymbol{\beta}) = \Lambda(\boldsymbol{x}'\boldsymbol{\beta}) = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}'\boldsymbol{\beta}}}. \tag{2}$$

$$\text{Probit:} \quad F(\boldsymbol{x}'\boldsymbol{\beta}) = \Phi(\boldsymbol{x}'\boldsymbol{\beta}) = \int_{-\infty}^{\boldsymbol{x}'\boldsymbol{\beta}} \phi(z)dz. \tag{3}$$

Maximum Likelihood Estimation We assume a sample of N independent observations of $\{y_i, \boldsymbol{x}_i\}_{i=1}^{N}$. Given the binomial nature of the data and the independence assumption across observations, a very convenient feature of the binomial model is that the distribution of the outcome is known: the *Bernoulli* distribution. The choice of 0 and 1 for the values of the outcome variable allow us to write the probability mass function in a very compact way:

$$g(y|\boldsymbol{x}) = p^y(1-p)^{1-y} = \begin{cases} p & \text{if } y = 1, \\ 1-p & \text{if } y = 0, \end{cases} \tag{4}$$

where $p \equiv F(\boldsymbol{x}'\boldsymbol{\beta})$. Therefore, the log-likelihood is given by:

$$\mathcal{L}_{\mathrm{N}}(\boldsymbol{\beta}) = \sum_{\mathrm{i=1}}^{\mathrm{N}} \{y_i \ln F(\boldsymbol{x}_i'\boldsymbol{\beta}) + (1-y_i)\ln(1-F(\boldsymbol{x}_i'\boldsymbol{\beta}))\}. \tag{5}$$

The ML estimator is obtained from the first order condition of the maximization of the previous expression:

$$\begin{aligned}\frac{\partial \mathcal{L}_{\mathrm{N}}}{\partial \boldsymbol{\beta}} &= \sum_{\mathrm{i=1}}^{\mathrm{N}} \left\{ \frac{y_i}{F(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})} f(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})\boldsymbol{x}_i - \frac{1-y_i}{1-F(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})} f(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})\boldsymbol{x}_i \right\} \\ &= \sum_{\mathrm{i=1}}^{\mathrm{N}} \frac{y_i - F(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})}{F(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})(1-F(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}))} f(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})\boldsymbol{x}_i = \mathbf{0}, \qquad (6)\end{aligned}$$

where $f(\cdot) \equiv \frac{\partial F(z)}{\partial z}$.

There is no explicit solution in general for $\hat{\boldsymbol{\beta}}_{MLE}$, so numerical algorithms are needed. Newton-Raphson procedure usually converges very quickly because the log-likelihood function is globally concave for probit and logit models.

Asymptotic properties The general ML results reviewed in Chapter 1 are of application in this context. We already know that the distribution of y is a Bernoulli distribution, so for consistence, we additionally need $F(\boldsymbol{x}'\boldsymbol{\beta}_0)$ to be the correct specification of p. The true parameter vector needs to be the maximand of the population likelihood $\mathbb{E}[y \ln F(\boldsymbol{x}'\boldsymbol{\beta}) + (1-y)\ln(1-F(\boldsymbol{x}'\boldsymbol{\beta}))]$:

$$\mathbb{E}\left[\frac{y - F(\boldsymbol{x}'\boldsymbol{\beta})}{F(\boldsymbol{x}'\boldsymbol{\beta})(1-F(\boldsymbol{x}'\boldsymbol{\beta}))} f(\boldsymbol{x}'\boldsymbol{\beta})\boldsymbol{x}\right] = \mathbb{E}\left[\frac{\mathbb{E}[y|\boldsymbol{x}] - F(\boldsymbol{x}'\boldsymbol{\beta})}{F(\boldsymbol{x}'\boldsymbol{\beta})(1-F(\boldsymbol{x}'\boldsymbol{\beta}))} f(\boldsymbol{x}'\boldsymbol{\beta})\boldsymbol{x}\right] = \mathbf{0}, \qquad (7)$$

where the first equality is obtained by applying the law of iterated expectations. We can easily see that this expression equals $\mathbf{0}$ if $\mathbb{E}[y|\boldsymbol{x}] = F(\boldsymbol{x}'\boldsymbol{\beta}_0)$.

Again, from the general MLE results reviewed in Chapter 1, $\hat{\boldsymbol{\beta}} \underset{d}{\rightarrow} \mathcal{N}(\boldsymbol{\beta}, \Omega_0/N))$, where $\Omega_0 = -\mathbb{E}\left[\partial^2\mathcal{L}/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'\right]^{-1}$. It is very easy to compute this expression using equation (6) and the information matrix equality:

$$\begin{aligned}-\mathbb{E}\left[\frac{\partial^2\mathcal{L}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right]^{-1} &= \mathbb{E}\left[\frac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}}\frac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}'}\right]^{-1} = \mathbb{E}\left[\left(\frac{y - F(\boldsymbol{x}'\boldsymbol{\beta})}{F(\boldsymbol{x}'\boldsymbol{\beta})\,(1-F(\boldsymbol{x}'\boldsymbol{\beta}))} f(\boldsymbol{x}'\boldsymbol{\beta})\right)^2 \boldsymbol{x}\boldsymbol{x}'\right]^{-1} \\ &= \mathbb{E}\left[\frac{y^2 + F(\boldsymbol{x}'\boldsymbol{\beta})^2 - 2yF(\boldsymbol{x}'\boldsymbol{\beta})}{\left(F(\boldsymbol{x}'\boldsymbol{\beta})\,(1-F(\boldsymbol{x}'\boldsymbol{\beta}))\right)^2} f(\boldsymbol{x}'\boldsymbol{\beta})^2\boldsymbol{x}\boldsymbol{x}'\right]^{-1} \\ &= \mathbb{E}\left[\frac{\mathbb{E}[y|x] + F(\boldsymbol{x}'\boldsymbol{\beta})^2 - 2\,\mathbb{E}[y|x]F(\boldsymbol{x}'\boldsymbol{\beta})}{\left(F(\boldsymbol{x}'\boldsymbol{\beta})\,(1-F(\boldsymbol{x}'\boldsymbol{\beta}))\right)^2} f(\boldsymbol{x}'\boldsymbol{\beta})^2\boldsymbol{x}\boldsymbol{x}'\right]^{-1} \\ &= \mathbb{E}\left[\frac{F(\boldsymbol{x}'\boldsymbol{\beta}) + F(\boldsymbol{x}'\boldsymbol{\beta})^2 - 2F(\boldsymbol{x}'\boldsymbol{\beta})^2}{\left(F(\boldsymbol{x}'\boldsymbol{\beta})\,(1-F(\boldsymbol{x}'\boldsymbol{\beta}))\right)^2} f(\boldsymbol{x}'\boldsymbol{\beta})^2\boldsymbol{x}\boldsymbol{x}'\right]^{-1} \\ &= \mathbb{E}\left[\frac{1}{F(\boldsymbol{x}'\boldsymbol{\beta})\,(1-F(\boldsymbol{x}'\boldsymbol{\beta}))} f(\boldsymbol{x}'\boldsymbol{\beta})^2\boldsymbol{x}\boldsymbol{x}'\right]^{-1}, \qquad (8)\end{aligned}$$

where we go from the second to the third line by applying the law of iterated expectations, and the expression in the fourth line is obtained under the condition

$\mathbb{E}[y|\boldsymbol{x}] = F(\boldsymbol{x}'\boldsymbol{\beta})$. It is noteworthy that the previous expression for the variance of the estimator has the form $\mathbb{E}[\omega\boldsymbol{x}\boldsymbol{x}']^{-1}$, where ω denotes the weights implicit in the previous expression.

Marginal effects As parameters are, in general, not interpretable in this model given the implicit normalizations discussed below, our objects of interest are the *marginal effects*: the effect of a marginal change in regressor k on the probability of choosing alternative A:

$$\frac{\partial \Pr[y=1|\boldsymbol{x}]}{\partial x_k} = f(\boldsymbol{x}'\boldsymbol{\beta})\beta_k. \tag{9}$$

Note that, unlike in the linear probability model, where $f(\boldsymbol{x}'\boldsymbol{\beta}) = 1$, in any nonlinear model, marginal effects depend on $\boldsymbol{x}$, and we may either compute the sample average marginal effect (i.e. the average over the evaluation of equation (9) over all observations), the marginal effect for the average individual (i.e. the evaluation of equation (9) at sample average values $\bar{\boldsymbol{x}}$), or the marginal effect for any other "representative" individual with $\boldsymbol{x} = \boldsymbol{x}^*$.

Estimated coefficients are still carry relevant information. Given that $f(\boldsymbol{x}'\boldsymbol{\beta})$ is a pdf, it is positive for all values of $\boldsymbol{x}$; therefore, the signs of the marginal effects are determined by the sign of the estimated coefficients. Additionally, the ratio of marginal effects for two different regressors are constant across individuals, and equal to the ratio of the two coefficients:

$$\frac{\partial \Pr[y=1|\boldsymbol{x}]/\partial x_k}{\partial \Pr[y=1|\boldsymbol{x}]/\partial x_l} = \frac{f(\boldsymbol{x}'\boldsymbol{\beta})\beta_k}{f(\boldsymbol{x}'\boldsymbol{\beta})\beta_l} = \frac{\beta_k}{\beta_l}. \tag{10}$$

The marginal effect for discrete regressors is computed as the difference in predicted probabilities. In particular, in the case of a dichotomic regressor, the marginal effect would be computed as:

$$\Pr[y=1|\boldsymbol{x}_{-k}, x_k = 1] - \Pr[y=1|\boldsymbol{x}_{-k}, x_k = 0] = F(\boldsymbol{x}'_{-k}\boldsymbol{\beta}_{-k} + \beta_k) - F(\boldsymbol{x}'_{-k}\boldsymbol{\beta}_{-k}), \tag{11}$$

where $\boldsymbol{x}_{-k}$ denote a vector with all regressors in $\boldsymbol{x}$ but x_k, and $\boldsymbol{\beta}_{-k}$ denote a vector with all coefficients in $\boldsymbol{\beta}$ but β_k.

D. The Logit Model

The *logit* model is a model of the general class seen in Section I.C in which we specify the conditional probabilities to be given by the logistic cdf:

$$F(\boldsymbol{x}'\boldsymbol{\beta}) = \Lambda(\boldsymbol{x}'\boldsymbol{\beta}) = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}'\boldsymbol{\beta}}}. \tag{12}$$

This parametrization is very convenient because the general algebra described in the previous section is simplified considerably. In particular, a very nice property of this cdf is that, given that $\partial\Lambda(z)/\partial(z) = e^z/(1+e^z)^2$ and $1-\Lambda(z) = 1/(1+e^z)$, then $\partial\Lambda(z)/\partial(z) = \Lambda(z)(1-\Lambda(z))$. The expression for the first order condition of the ML estimator reduces to:

$$\sum_{\mathrm{i=1}}^{\mathrm{N}} \left(y_i - \Lambda(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})\right) \boldsymbol{x}_i = \mathbf{0}. \tag{13}$$

Interestingly, this expression implies that the residual, $y - \Lambda(\boldsymbol{x}'\hat{\boldsymbol{\beta}})$, is orthogonal to $\boldsymbol{x}$, as in the case of OLS.

The expression for the asymptotic variance of $\hat{\boldsymbol{\beta}}_{MLE}$ is also very simple:

$$\Omega_0 = \mathbb{E}\left[\Lambda(\boldsymbol{x}'\boldsymbol{\beta})\left(1-\Lambda(\boldsymbol{x}'\boldsymbol{\beta})\right)\boldsymbol{x}\boldsymbol{x}'\right]^{-1}. \tag{14}$$

And the marginal effects are given by:

$$\frac{\partial \Pr[y=1|\boldsymbol{x}]}{\partial x_k} = \Lambda(\boldsymbol{x}'\boldsymbol{\beta})(1-\Lambda(\boldsymbol{x}'\boldsymbol{\beta}))\beta_k. \tag{15}$$

An additional interesting feature of the logit model is that the log-odds ratio is linear in the regressors:

$$p = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}'\boldsymbol{\beta}}} \Leftrightarrow \ln\frac{p}{1-p} = \boldsymbol{x}'\boldsymbol{\beta}. \tag{16}$$

This expression is interesting because it allows us to interpret $\boldsymbol{\beta}$ as a semi-elasticity.[1] Additionally, it allows to estimate $\boldsymbol{\beta}$ with aggregate data under certain assumptions. This transformation is used in chapter 6.

E. The Probit Model

The *probit* model is a model of the general class described in Section I.C in which we specify the conditional probabilities to be given by the standard normal cdf:

$$F(\boldsymbol{x}'\boldsymbol{\beta}) = \Phi(\boldsymbol{x}'\boldsymbol{\beta}) = \int_{-\infty}^{\boldsymbol{x}'\boldsymbol{\beta}} \phi(z)dz. \tag{17}$$

The first order conditions for the MLE in this case are:

$$\sum_{\mathrm{i=1}}^{\mathrm{N}} \frac{y_i - \Phi(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})}{\Phi(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})(1-\Phi(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}))}\phi(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})\boldsymbol{x}_i = \mathbf{0}. \tag{18}$$

[1] Given a function $h(z)$, a semi-elasticity gives the percentage change in $h(z)$ in terms of a change (not percentage-wise) of z.

The marginal effects are given by:

$$\frac{\partial \Pr[y=1|\boldsymbol{x}]}{\partial x_k} = \phi(\boldsymbol{x}'\boldsymbol{\beta})\beta_k. \tag{19}$$

Finally, the asymptotic variance of the estimator is:

$$\Omega_0 = \mathbb{E}\left[\frac{\phi(\boldsymbol{x}'\boldsymbol{\beta})^2}{\Phi(\boldsymbol{x}'\boldsymbol{\beta})\,(1-\Phi(\boldsymbol{x}'\boldsymbol{\beta}))}\boldsymbol{x}\boldsymbol{x}'\right]^{-1}. \tag{20}$$

F. Latent Variable Representation

One way to give a more structural interpretation to the model is to conceive it in terms of a latent measure of utility. A latent variable is a variable that we do not observe completely. For instance, we can observe the final decision of an individual but not the intrinsic utility experienced by her. Individual decisions can be modeled based on their utility function, whose parameters can be estimated thanks to the revealed preference.

There are two alternative ways of modeling the binary outcome model in terms of a latent variable. The first one is called the *index function model*, in which a threshold on the latent variable determines whether the individual chooses one alternative or the other (e.g. I buy a product if my utility from buying it is positive). The second one is called the *(additive) random utility model*, in which the individual compares the latent utility associated with the two alternatives and chooses the one that provides the largest utility. In the binary case, these interpretations are somewhat equivalent, but the distinction between them becomes relevant in the multinomial case.

Index function model Let y^* be the latent variable of interest, such that:

$$y^* = \boldsymbol{x}'\boldsymbol{\beta} + u, \tag{21}$$

where $\boldsymbol{x}$ is a vector of regressors, and u is an unobserved error component with cdf $F(\cdot)$. We only observe which alternative is chosen by the individual, this is:

$$y = \begin{cases} 1 \text{ if } y^* > 0, \\ 0 \text{ if } y^* \leq 0, \end{cases} \tag{22}$$

where 0 is a normalization of a threshold c, not identified if the model includes an intercept term.

The probability that we observe $y = 1$ is given by:

$$\Pr[y=1|\boldsymbol{x}] = \Pr[y^* > 0|\boldsymbol{x}] = \Pr[\boldsymbol{x}'\boldsymbol{\beta} + u > 0] = \Pr[u > -\boldsymbol{x}'\boldsymbol{\beta}] = F(\boldsymbol{x}'\boldsymbol{\beta}), \tag{23}$$

where the last equality comes from assuming that the pdf of u is symmetric around zero (e.g. the logistic and the standard normal distributions). This model delivers the logit model if $F(\cdot) = \Lambda(\cdot)$, and the probit if $F(\cdot) = \Phi(\cdot)$.

In the previous expression, it emerges that a threshold c would not be identified if the model includes an intercept because we could increase the intercept and reduce the threshold by the same amount a and the likelihood would not change. Similarly, the parameters are only identified up to a scale:

$$\Pr[u > -\boldsymbol{x}'\boldsymbol{\beta}] = \Pr[ua > -\boldsymbol{x}'\boldsymbol{\beta}a]. \tag{24}$$

Therefore, we have to impose a restriction on the variance of the error term so that the parameters can be uniquely identified. This restriction is implicit in the logistic distribution (we impose that the variance is $\pi^2/3$), and in the case of normal errors, we typically impose that the variance is 1 so that we work with the standard normal distribution and the problem reduces to the probit model. These restrictions have to be taken into account when interpreting the coefficients. In fact, this is the reason why coefficients themselves are hard to be interpreted directly and we focus on the marginal effects of each regressor (or the ratio between two coefficients, which delivers the ratio of marginal effects).

(Additive) Random utility model Consider the utility of the two alternatives:

$$U_0 = V_0 + \varepsilon_0, \tag{25}$$

$$U_1 = V_1 + \varepsilon_1, \tag{26}$$

where V_0 and V_1 are deterministic components of the utility, and ε_0 and ε_1 are random components of utility. Instead of observing the utility associated to the alternatives, we observe:

$$y = \begin{cases} 1 \text{ if } U_1 > U_0, \\ 0 \text{ if } U_1 \leq U_0. \end{cases} \tag{27}$$

The probability that we observe $y = 1$ is given by:

$$\begin{aligned} \Pr[y = 1|\boldsymbol{x}] = \Pr[U_1 > U_0|\boldsymbol{x}] &= \Pr[V_1 + \varepsilon_1 > V_0 + \varepsilon_0|\boldsymbol{x}] \\ &= \Pr[\varepsilon_0 - \varepsilon_1 < V_1 - V_0|\boldsymbol{x}] = F(V_1 - V_0), \end{aligned} \tag{28}$$

where $F(\cdot)$ is the distribution of $\varepsilon_0 - \varepsilon_1$ given $\boldsymbol{x}$.

We typically express $V_1 - V_0$ as a single-index ($V_1 - V_0 = \boldsymbol{x}'\boldsymbol{\beta}$); anyway, if we were interested in modeling $V_1 = \boldsymbol{x}'\boldsymbol{\beta}_1$ and $V_0 = \boldsymbol{x}'\boldsymbol{\beta}_0$, then $V_1 - V_0 = \boldsymbol{x}'(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$, and only $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$ would be identified. The case in which each utility is affected by different regressors is the particular case of the previous expression in which some of the $\boldsymbol{\beta}$s take value 0. The case of alternative-varying regressors can be seen also as a special case, but it is a bit more sophisticated:

$$V_j = \boldsymbol{z}_j'\boldsymbol{\alpha}+\boldsymbol{x}'\boldsymbol{\beta}_j \quad j=0,1 \quad \Rightarrow \quad \Pr[y=1|\boldsymbol{z},\boldsymbol{x}] = F\left((\boldsymbol{z}_1-\boldsymbol{z}_0)'\boldsymbol{\alpha}+\boldsymbol{x}'(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)\right). \tag{29}$$

The parameter vector $\boldsymbol{\alpha}$ is usually assumed to be constant across alternatives, but it could also be allowed to vary.

Different assumptions on the distribution of the random utility components deliver different models. A natural choice of error distribution is that ε_1 and ε_0 are normally distributed; in this case, $(\varepsilon_0 - \varepsilon_1)$ is also normally distributed, and the normalization such that the variance of this difference is the unity delivers the probit model. Alternatively, if ε_1 and ε_0 are assumed to be independently distributed as Type I extreme value (a.k.a. log Weibull), i.e.:

$$f(\varepsilon_j) = e^{-\varepsilon_j}\exp\{e^{-\varepsilon_j}\}, \quad j=0,1, \tag{30}$$

then $(\varepsilon_0 - \varepsilon_1)$ follows a logistic distribution, and the logit model emerges.[2]

II. Multinomial Models

A. Multinomial Outcomes

In this section we generalize the estimation of models for discrete outcomes to the case in which individuals choose among $m > 2$ mutually exclusive alternatives. Now we have to distinguish between the case in which the available alternatives are ordered (having zero, one or two/more children; not liking, being indifferent or loving something;...) or unordered (going to work by bus, train or car; working, going to school or staying home;...). Most of this section covers the case of unordered data. The only exception is in Section II.F, where we cover the models for ordered data.

For notational convenience, we define m binary variables y_j for $j = 1, ..., m$, each of which take the value of 1 if the category j is selected and zero otherwise. This is very convenient because, as a result, $N^{-1}\sum_{\mathrm{i}=1}^{\mathrm{N}} y_{ij} = \widehat{\Pr}[y=j]$. The fact that alternatives are mutually exclusive imply that one and only one of $y_{i1}, ..., y_{im}$ is equal to one, and the rest are zero.

B. The General Multinomial Model

As in the case of binary outcome models, in this section we present a very general model that nests the models that are covered below as special cases. This

[2] The proof is available at Cameron and Trivedi (2005), Section 14.8.

general notation is useful because many of the results are general across models.

The conditional probability of choosing alternative j given $\boldsymbol{x}$ is given by:

$$p_j(\boldsymbol{x}) \equiv \Pr[y = j|\boldsymbol{x}] = F_j(\boldsymbol{x}'\boldsymbol{\beta}), \quad j = 1, ..., m. \tag{31}$$

$F_j(\boldsymbol{x}'\boldsymbol{\beta})$ is a specified function of $\boldsymbol{x}'\boldsymbol{\beta}$ such that the predicted probabilities lie between 0 and 1, and such that $\sum_j p_j = 1$. Different specifications of $F_j(\cdot)$ correspond to the different models discussed below. The binary model is a special case in this notation.

Maximum Likelihood estimation We assume a sample of N independent observations $\{y_i, \boldsymbol{x}_i\}_{i=1}^N$. In this case, y given $\boldsymbol{x}$ follows a *multinomial* distribution, which can be written in a very compact way as:

$$g(y|\boldsymbol{x}) = p_1^{y_1} \times p_2^{y_2} \times \cdots \times p_m^{y_m} = \prod_{j=1}^{m} p_j^{y_j} = \begin{cases} p_1 & \text{if } y = 1, \\ p_2 & \text{if } y = 2, \\ \vdots & \\ p_m & \text{if } y = m, \end{cases} \tag{32}$$

where we omitted that p_j's are functions of $\boldsymbol{x}$ for notational simplicity.

The log-likelihood function is:

$$\mathcal{L}_{\mathrm{N}}(\boldsymbol{\beta}) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{j=1}^{m} y_{ij} \ln F_j(\boldsymbol{x}_i'\boldsymbol{\beta}). \tag{33}$$

The ML estimator $\hat{\boldsymbol{\beta}}_{MLE}$ is given by the first order conditions:

$$\frac{\partial \mathcal{L}_{\mathrm{N}}}{\partial \boldsymbol{\beta}} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{j=1}^{m} \frac{y_{ij}}{F_j(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})} f_j(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})\boldsymbol{x}_i = \boldsymbol{0}, \tag{34}$$

where $f_j(z) \equiv \partial F_j(z) \partial(z)$.

Asymptotic properties We check consistency by showing that the true parameter is the solution of the population counterpart of our maximization problem which is $\mathbb{E}\left[\sum_{j=1}^{m} y_j \ln F_j(\boldsymbol{x}'\boldsymbol{\beta})\right]$. Given that $\mathbb{E}[y_j|\boldsymbol{x}] = p_j$ and $\sum_j p_j = 1$, consistency is achieved if $F_1(\boldsymbol{x}'\boldsymbol{\beta}_0), ..., F_m(\boldsymbol{x}'\boldsymbol{\beta}_0)$ are the correct specification of $p_1, ..., p_m$:

$$\mathbb{E}\left[\sum_{j=1}^{m} \frac{y_j}{F_j(\boldsymbol{x}'\boldsymbol{\beta})} f_j(\boldsymbol{x}'\boldsymbol{\beta})\boldsymbol{x}\right] = \mathbb{E}\left[\sum_{j=1}^{m} \frac{\mathbb{E}[y_j|\boldsymbol{x}]}{F_j(\boldsymbol{x}'\boldsymbol{\beta})} f_j(\boldsymbol{x}'\boldsymbol{\beta})\boldsymbol{x}\right] = \mathbb{E}\left[\sum_{j=1}^{m} f_j(\boldsymbol{x}'\boldsymbol{\beta})\boldsymbol{x}\right] = \boldsymbol{0}, \tag{35}$$

where the first equality is obtained by applying the law of iterated expectations, the second is provided by assuming $\mathbb{E}[y_j|\boldsymbol{x}] = F_j(\boldsymbol{x}'\boldsymbol{\beta}_0)$ is satisfied, and the third one is obtained by taking partial derivatives to $\sum_j p_j = 1$.

We can also apply the results from Chapter 1 to obtain the asymptotic distribution of the ML estimator. We know that $\hat{\boldsymbol{\beta}}_{MLE} \underset{d}{\rightarrow} \mathcal{N}(\boldsymbol{\beta}, \Omega_0/N)$, where:

$$\Omega_0 = -\,\mathbb{E}\left[\frac{\partial^2 \mathcal{L}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right]^{-1} = \mathbb{E}\left[\sum_{j=1}^{m}\left(\frac{y_j}{p_j^2}\frac{\partial p_j}{\partial\boldsymbol{\beta}}\frac{\partial p_j}{\partial\boldsymbol{\beta}'} - \frac{y_j}{p_j}\frac{\partial^2 p_j}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right)\right]^{-1}$$
$$= \mathbb{E}\left[\sum_{j=1}^{m}\left(\frac{1}{p_j}\frac{\partial p_j}{\partial\boldsymbol{\beta}}\frac{\partial p_j}{\partial\boldsymbol{\beta}'} - \frac{\partial^2 p_j}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right)\right]^{-1}, \qquad (36)$$

which is obtained making use of the law of iterated expectations once again.

Marginal effects Marginal effects are calculated analogously to what we have seen for the binomial case. However, two remarks are noteworthy here. The first one, is that the sign of the coefficient is not necessarily the same as the sign of the corresponding marginal effect anymore. The second is that they differ depending on whether we are talking about an alternative-varying or an alternative-invariant variable. The marginal effect for the former indicates by how much the probabilities change when the regressor is changed for one alternative leaving constant its value for the other alternatives. In the case of alternative-invariant regressors, this *ceteris paribus* assumption does not hold anymore, as when a regressor is increased, it is increased for all alternatives equally.

C. The Logit Model

In the case of the logit model, whether regressors vary or not across alternatives is of practical relevance. The reason is that we can achieve very simple expressions for the less general case (alternative-invariant regressors) and it is very common in practice not to have any alternative-varying regressor in the data set.

When regressors are alternative-invariant (i.e. $\boldsymbol{x}_j = \boldsymbol{x}$ for $j = 1, ..., m$), we typically estimate a single-index model in which the argument of $F(\cdot)$ is $\boldsymbol{x}'\boldsymbol{\beta}_j$. As noted below, a normalization is needed for identification, such as $\boldsymbol{\beta}_1 = 0$. Therefore, the interpretation of the corresponding coefficients is always with respect to the normalized or *base* category. The logit model in which all regressors are alternative-invariant is called the *Multinomial Logit*.

If regressors are alternative-varying, we typically assume that $\boldsymbol{\beta}$ is invariant across alternatives, and, hence, the index is $\boldsymbol{x}_j'\boldsymbol{\beta}$; in this case, we may also interpret the results in terms of a base category by specifying all regressors in deviations with respect to the base category. The logit model with alternative-varying regressors is known as *Conditional Logit*.

Although convenient in practice, this distinction is irrelevant from a theoretical point of view, as the alternative-varying model nests the alternative-invariant one. In particular, we can define a $km \times 1$ vector with zeros everywhere except the jth block, which is $\boldsymbol{x}$, i.e. $\boldsymbol{x}_j = [\mathbf{0}' \; ... \; \mathbf{0}' \; \boldsymbol{x} \; \mathbf{0}' \; ... \; \mathbf{0}']'$, and a $km \times 1$ vector in which each block includes the k regressors of $\boldsymbol{\beta}_j$, i.e. $\boldsymbol{\beta} = [\mathbf{0}' \; \boldsymbol{\beta}_2' \; ... \; \boldsymbol{\beta}_m']'$, and then we can work with $\boldsymbol{x}_j'\boldsymbol{\beta}$ as in the alternative-varying case. This allows us to work with models in which we have regressors of both types.

The Multinomial Logit (MNL) The multinomial logit model is the application of the general model in which we specify:

$$F(\boldsymbol{x}'\boldsymbol{\beta}_j) = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}_j}}{\sum_{l=1}^{m} e^{\boldsymbol{x}'\boldsymbol{\beta}_l}}, \quad j = 1, ..., m. \tag{37}$$

The probabilities are well behaved in the sense that they lie between 0 and 1, and the sum of them across (mutually exclusive) alternatives equals one. Regressors are alternative-invariant.

Each set of first order conditions for the minimization problem reduces to:

$$\frac{\partial \mathcal{L}_{\mathrm{N}}}{\partial \boldsymbol{\beta}_h} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{j=1}^{m} \frac{y_{ij}}{p_{ij}} p_{ij}(\delta_{jh} - p_{ih})\boldsymbol{x}_i = \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{j=1}^{m} y_{ij}(\delta_{jh} - p_{ih})\boldsymbol{x}_i = \sum_{\mathrm{i}=1}^{\mathrm{N}} (y_{ih} - p_{ih})\boldsymbol{x}_i = \mathbf{0}, \tag{38}$$

where $\delta_{jh} = \mathbb{1}\{j = h\}$, and $\mathbb{1}\{A\}$ is an indicator function that takes the value of 1 if A is satisfied and 0 otherwise. We make use of the fact that $\sum_{j=1}^{m} y_{ij}\delta_{jh} = y_{ih}$ and $\sum_{j=1}^{m} y_{ij}p_{ih} = p_{ih}\sum_{j=1}^{m} y_{ij} = p_{ih}$. This expression is equivalent to the one we obtained for the binomial logit except that the probabilities p_{ih} are computed using (37) instead of (12).

Given that $\sum_{j=1}^{m} p_j = 1$, we have to do a normalization to ensure identification. We can easily see that $p_1 = 1 - \sum_{j=2}^{m} p_j = 1 - \sum_{j=2}^{m} \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}_j}}{\sum_{l=1}^{m} e^{\boldsymbol{x}'\boldsymbol{\beta}_l}} = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}_1}}{\sum_{l=1}^{m} e^{\boldsymbol{x}'\boldsymbol{\beta}_l}}$. This leaves us with only $m-1$ instead of m independent vectors of first order conditions like (38). We typically set $\boldsymbol{\beta}_1 = \mathbf{0}$. This is relevant because results for alternatives $j = 2, ..., m$ need to be interpreted in comparison to the *base* category.

The information matrix $-\mathbb{E}\left[\partial^2\mathcal{L}/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'\right]$ is defined by blocks, each corresponding to $-\mathbb{E}\left[\partial^2\mathcal{L}/\partial\boldsymbol{\beta}_h\partial\boldsymbol{\beta}_l'\right]$. Thus, differentiating (38) with respect to $\boldsymbol{\beta}_l'$, we obtain:

$$-\mathbb{E}\left[\partial^2\mathcal{L}/\partial\boldsymbol{\beta}_h\partial\boldsymbol{\beta}_l'\right] = \mathbb{E}\left[p_h(\delta_{hl} - p_l)\boldsymbol{x}\boldsymbol{x}'\right] = \begin{cases} \mathbb{E}[p_h(1 - p_l)\boldsymbol{x}\boldsymbol{x}'] & \text{if } h = l, \\ \mathbb{E}[-p_h p_l \boldsymbol{x}\boldsymbol{x}'] & \text{if } h \neq l, \end{cases} \tag{39}$$

for $h = 1, ..., m$ and $l = 1, ..., m$, and the asymptotic variance of the estimator is given by $-\mathbb{E}\left[\partial^2\mathcal{L}/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'\right]^{-1}$.

The marginal effects in this model are the effect of changing a regressor by one unit on the probabilities of choosing each alternative:

$$\frac{\partial p_j}{\partial x_k} = p_j \left(\beta_{jk} - \sum_{h=1}^{m} p_h \beta_{hk} \right) \equiv p_j(\beta_{jk} - \bar{\beta}_{\boldsymbol{p}k}). \tag{40}$$

The term $\bar{\beta}_{\boldsymbol{p}k}$ is a probability weighted average of alternative-specific β_{jk}s, using the choice probabilities $\boldsymbol{p}$ as weights. From this expression we can see that the sign of a parameter estimate does not necessarily correspond to the sign of the effect of an increase in the regressor on the probability of choosing this alternative, and, in that sense, it does not make much sense to test whether a coefficient is different from zero or not. More subtly, the sign of the individual marginal effects can differ across individuals, as the weighted average $\bar{\beta}_{\boldsymbol{p}}$ uses individual-specific choice probabilities as weights.

The Conditional Logit (CL) The conditional logit model is the application of the general model in which we specify:

$$F_j(\boldsymbol{x}'\boldsymbol{\beta}) = \frac{e^{\boldsymbol{x}_j'\boldsymbol{\beta}}}{\sum_{l=1}^{m} e^{\boldsymbol{x}_l'\boldsymbol{\beta}}}, \quad j = 1, ..., m. \tag{41}$$

As in the MNL, the probabilities are well behaved in the sense that they lie between 0 and 1, and sum to one. As in the MNL, given $\sum_{j=1}^{m} p_j = 1$, we can rewrite these probabilities in terms of a *base category* (although in this case we do not have to do it explicitly for identification, one of the categories is redundant). For instance, we can set $j = 1$ as the base category, in which case we can rewrite the model with the variables in deviations from $\boldsymbol{x}_1$, i.e. $\tilde{\boldsymbol{x}}_j \equiv (\boldsymbol{x}_j - \boldsymbol{x}_1)$ (and hence set $\tilde{\boldsymbol{x}}_1 = 0$) so that there is no redundant category. This gives the correct interpretation of the estimates, which is with respect to the base category.

The first order conditions in this case are given by:

$$\frac{\partial \mathcal{L}_{\mathrm{N}}}{\partial \boldsymbol{\beta}} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{j=1}^{m} \frac{y_{ij}}{p_{ij}} p_{ij}(\boldsymbol{x}_{ij} - \sum_{h=1}^{m} p_{ih}\boldsymbol{x}_{ih}) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{j=1}^{m} y_{ij}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_{\boldsymbol{p}_i}) = \boldsymbol{0}, \tag{42}$$

where $\bar{\boldsymbol{x}}_{\boldsymbol{p}_i} \equiv \sum_{h=1}^{m} p_{ih}\boldsymbol{x}_{ih}$.

The expression for the asymptotic variance can be obtained by using the information matrix equality:

$$\begin{aligned} \Omega_0 &= \mathbb{E}\left[\sum_{l=1}^{m}\sum_{j=1}^{m} y_j y_l (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_l - \bar{\boldsymbol{x}})'\right]^{-1} \\ &= \mathbb{E}\left[\sum_{j=1}^{m} y_j^2 (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})'\right]^{-1} = \mathbb{E}\left[\sum_{j=1}^{m} p_j (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})'\right]^{-1}, \end{aligned} \tag{43}$$

where we use the law of iterated expectations, the 0/1 nature of the variable y, so that $\mathbb{E}[y_j^2|\boldsymbol{x}_j] = \mathbb{E}[y_j|\boldsymbol{x}_j] = p_j$, and that $y_j y_l = 0 \; \forall j \neq l$.

Finally, the marginal effects are given by:

$$\frac{\partial p_j}{\partial x_{hk}} = p_j(\delta_{jh} - p_h)\beta_k = \begin{cases} p_j(1-p_j)\beta_k & \text{if } j = h, \\ -p_j p_h \beta_k & \text{if } j \neq h. \end{cases} \tag{44}$$

Therefore, in this case, the sign of the *own* marginal effects is the same as the sign of β_k, while the sign of the *cross* effects is opposite to the sign of the coefficient.

D. Latent Variable Representation

The latent variable representation in the multinomial context is based on the Additive Random Utility Model (ARUM). As in the binary outcome case, consider the utility of choosing alternative j as the sum of a deterministic or observable component and a random component:

$$U_j = V_j + \varepsilon_j, \quad j = 1, ..., m. \tag{45}$$

We define the deterministic component as a single-index $V_j \equiv \boldsymbol{x}'\boldsymbol{\beta}_j$ or $V_j \equiv \boldsymbol{x}_j'\boldsymbol{\beta}$. Individuals choose the alternative that gives them the highest utility. Therefore, the probability that they choose alternative j is given by:

$$\begin{aligned} \Pr[y = j|\boldsymbol{x}] &= \Pr[U_j \geq U_h \; \forall h \neq j|\boldsymbol{x}] \\ &= \Pr[\varepsilon_h - \varepsilon_j \leq -(V_h - V_j) \; \forall h \neq j|\boldsymbol{x}] \equiv \Pr[\tilde{\varepsilon}_{hj} \leq -\tilde{V}_{hj} \; \forall h \neq j|\boldsymbol{x}], \end{aligned} \tag{46}$$

where we define $\tilde{z}_{hj} = z_h - z_j$.

Different multinomial models can be generated by different assumptions on the joint distribution of the error terms. The probabilities are computed with the corresponding cdfs. For instance, in the three-choice model, the probability of choosing alternative 1 is given by:

$$\Pr[y = 1|\boldsymbol{x}] = \Pr[\tilde{\varepsilon}_{21} \leq -\tilde{V}_{21}, \tilde{\varepsilon}_{31} \leq -\tilde{V}_{31}|\boldsymbol{x}] = \int_{-\infty}^{-\tilde{V}_{31}} \int_{-\infty}^{-\tilde{V}_{21}} f(\tilde{\varepsilon}_{21}, \tilde{\varepsilon}_{31}) d\tilde{\varepsilon}_{21} d\tilde{\varepsilon}_{31}. \tag{47}$$

Computing $m - 1$ dimensional integrals is computationally demanding, and the computational burden increases exponentially as we increase the number of available choices. This complexity favors logit models as opposed to probit models when the number of alternatives is large (even when the normality is a more natural assumption for the errors than the extreme value) because they often provide analytical solutions to the integrals.

The MNL and the CL are obtained by assuming that errors $\varepsilon_1, ..., \varepsilon_m$ are independently and identically distributed as a Type I extreme value, as in the binomial case. The independence assumption implies that the joint distribution of error differences can be factorized and, hence, the multiple dimensional integral reduces to the product of one-dimensional integrals.

As discussed below, we may allow the random components to be correlated. In this case, some covariance restrictions are necessary, as the model is only identified up to $m-1$ error-difference pairs. Additionally —for the same reason as in the the binomial case—, one variance needs to be imposed, as U_j is only determined up to scale.

E. Relaxing the Independence of Irrelevant Alternatives Assumption.

The assumption of uncorrelated $\varepsilon_1, ..., \varepsilon_m$ from the MNL and CL models is known as *independence of irrelevant alternatives* (IIA). The implication of the IIA is that the problem is reduced to a comparison between any pair of alternatives.

This assumption is often seen as too restrictive. A well known extreme example is known as the red bus-blue bus problem. Consider that we are analyzing the choice of the transportation mean used to commute: car, red bus or blue bus. The only difference between red and blue bus is the color. The MNL and CL models assume that the conditional probability of commute by car given commute by car or red bus (Pr[*car*|*car or red bus*]) is independent of whether there is a blue bus option or not. In other words, if the alternatives were only car or red bus, we could estimate a (binary) logit on this:

$$\Pr[c|c{\cup}rb] = \frac{\Pr[c]}{\Pr[c \cup rb]} = \frac{\frac{e^{\boldsymbol{x}'\boldsymbol{\beta}_c}}{\sum_{j=1}^{m} e^{\boldsymbol{x}'\boldsymbol{\beta}_j}}}{\frac{e^{\boldsymbol{x}'\boldsymbol{\beta}_c}}{\sum_{j=1}^{m} e^{\boldsymbol{x}'\boldsymbol{\beta}_j}} + \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}_{rb}}}{\sum_{j=1}^{m} e^{\boldsymbol{x}'\boldsymbol{\beta}_j}}} = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}_c}}{e^{\boldsymbol{x}'\boldsymbol{\beta}_c} + e^{\boldsymbol{x}'\boldsymbol{\beta}_{rb}}} = \frac{e^{\boldsymbol{x}'(\boldsymbol{\beta}_c - \boldsymbol{\beta}_{rb})}}{1 + e^{\boldsymbol{x}'(\boldsymbol{\beta}_c - \boldsymbol{\beta}_{rb})}}. \tag{48}$$

However, in practice, one would expect that introducing the blue bus would have a larger effect on the red bus commuting than on car commuting. This should increase Pr[*car*|*car or red bus*], as the introduction of the blue bus "steals" more observations to the red bus than to the car alternative. In this section we discuss some of the most popular models that break the IIA assumption.

The Nested Logit (NL) This is the most analytically tractable of the generalizations of the multinomial model. It is ideal when there is a clear nesting structure, although this is not the case of all applications. The nested logit model breaks the decision tree into limbs and branches; IIA is assumed only between limbs, a fixed correlation within each limb is estimated. Here we discuss the model with two levels, but it could be generalized to introduce further levels.

The decision tree is as follows. First, the individual chooses a limb j among the available J limbs (e.g. college or work). Then, she chooses one of the H_j branches available within limb j (e.g. if she chooses to work, which occupation; if she chooses to go to college, which college). Hence, the joint probability of being in limb j and choosing branch h is the product of the probability of being in limb j times the conditional probability of choosing branch h given that limb j have been chosen.

We consider a single-index with regressors that vary across limbs and branches and others that only vary across limbs:

$$V_{jh} \equiv \boldsymbol{z}_j'\boldsymbol{\alpha} + \boldsymbol{x}_{jh}'\boldsymbol{\beta}_j, \quad h = 1, ..., H_j,\ j = 1, ..., J, \tag{49}$$

where $\boldsymbol{z}_j'$ only vary across limbs and $\boldsymbol{x}_{jh}'$ vary across limbs and branches. This variation is important for identification. It can also be adapted to alternative invariant regressors by considering $V_{jh} \equiv \boldsymbol{z}'\boldsymbol{\alpha}_j + \boldsymbol{x}'\boldsymbol{\beta}_{jh}$.

In the nested logit model, the probability of choosing alternative jh is given by:

$$p_{jh} = p_j \times p_{h|j} = \frac{\exp\left(\boldsymbol{z}_j'\boldsymbol{\alpha} + \rho_j IV_j\right)}{\sum_{l=1}^{J} \exp\left(\boldsymbol{z}_l'\boldsymbol{\alpha} + \rho_l IV_l\right)} \times \frac{\exp\left(\boldsymbol{x}_{jh}'\boldsymbol{\beta}_j/\rho_j\right)}{\sum_{r=1}^{H_j} \exp\left(\boldsymbol{x}_{jr}'\boldsymbol{\beta}_j/\rho_j\right)}, \tag{50}$$

where:

$$IV_j = \ln\left(\sum_{r=1}^{H_j} \exp\left(\boldsymbol{x}_{jr}'\boldsymbol{\beta}_j/\rho_j\right)\right). \tag{51}$$

This expression can be derived from the latent variable representation by assuming that $\varepsilon_1, ..., \varepsilon_m$ are distributed according to a particular type of generalized extreme value distribution (see, for instance, Cameron and Trivedi, Chapter 15.6). Parameters ρ_j are known as *scale parameters* (as opposed to the *regression parameters* $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_j$) because they scale the regression parameters in the previous expression. They are associated with the correlation between random components of the branch j; in particular, $\rho_j = \sqrt{1 - \text{Corr}(\varepsilon_{jh}, \varepsilon_{jl})}$. The term IV_j is known as the *inclusive value*.

The log-likelihood function is rewritten as:

$$\mathcal{L}_{\text{N}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \rho_1, ..., \rho_j) = \sum_{\text{i}=1}^{\text{N}} \sum_{j=1}^{J} y_{ij} \ln p_{ij} + \sum_{\text{i}=1}^{\text{N}} \sum_{j=1}^{J} \sum_{h=1}^{H_j} y_{ijh} \ln p_{ih|j}. \tag{52}$$

The ML estimator obtained from maximizing this log-likelihood function is known as *Full Information ML* (or FIML).

A nice characteristic of the previous expression is that the probabilities p_j and $p_{h|j}$ are of conditional logit form. For the times (or cases) in which the computation of the FIML estimator was (or is) challenging, a *Limited Information ML* (LIML) or sequential estimator that takes advantage of this feature was proposed. The method consists of two stages. In the first stage $\boldsymbol{\beta}_j/\rho_j$ are estimated from conditional logits for the conditional probabilities in each of the limbs. The second stage consists of a conditional logit across limbs with the predicted inclusive value $\widehat{IV}_{jh}$ as an added regressor; this second stage delivers $\hat{\boldsymbol{\alpha}}$ and $\hat{\rho}_j$, and $\hat{\boldsymbol{\beta}}_j$ is recovered combining $\widehat{\boldsymbol{\beta}_j/\rho_j}$ and $\hat{\rho}_j$. This method is less efficient but still consistent; standard errors need to be corrected for the fact that estimates of the inclusive value instead of real values are used in the second stage. This method is also very useful to produce starting values to the FIML estimation, as the log-likelihood in equation (52) is not globally concave.

Random Parameters Logit (RPL) The random parameters logit (RPL) model specifies the utility of individual i of choosing alternative j to be:

$$U_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}_i + \varepsilon_{ij}, \quad \boldsymbol{\beta}_i \sim \mathcal{N}(\boldsymbol{\beta}, \Sigma_{\boldsymbol{\beta}}), \tag{53}$$

where ε_{ij} are i.i.d. Type I extreme value as for the CL model. The difference with CL is that it permits parameters $\boldsymbol{\beta}$ to be random. Although the normality assumption is the most common, there are other alternatives when the support of the parameters is not $[-\infty, \infty]$.

To see how this model generates correlation between unobservables across alternatives, we can rewrite the model as follows:

$$U_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \nu_{ij}; \qquad \nu_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{u}_i + \varepsilon_{ij}, \quad \boldsymbol{u}_i \sim \mathcal{N}(\boldsymbol{0}, \Sigma_{\boldsymbol{\beta}}). \tag{54}$$

Then, the covariance between unobservables is $\mathrm{Cov}(\nu_{ij}, \nu_{ih}) = \boldsymbol{x}'_{ij}\Sigma_{\boldsymbol{\beta}}\boldsymbol{x}_{ih}$ for $j \neq h$. In most of the applications, $\Sigma_{\boldsymbol{\beta}}$ is specified to be diagonal and, additionally, some of the diagonal values are set to zero (i.e., some parameters are assumed to be deterministic).

Given the extreme value assumption, the probability for individual i of choosing j is expected to be:

$$p_{ij} = \int \frac{e^{\boldsymbol{x}'_{ij}\boldsymbol{\beta}_i}}{\sum_{l=1}^{m} e^{\boldsymbol{x}'_{il}\boldsymbol{\beta}_i}} \phi(\boldsymbol{\beta}_i; \boldsymbol{\beta}, \Sigma_{\boldsymbol{\beta}}) d\boldsymbol{\beta}_i, \tag{55}$$

where the integral is k-dimensional, and $\phi(\boldsymbol{\beta}_i; \boldsymbol{\beta}, \Sigma_{\boldsymbol{\beta}})$ denotes the k-variate normal density with mean $\boldsymbol{\beta}$ and variance $\Sigma_{\boldsymbol{\beta}}$.[3] The log-likelihood has the general form

[3] In practice, the dimension of the integral is given by the parameters with non-zero variance,

seen in Section II.B, and the parameters of interest are $\boldsymbol{\beta}$ and $\Sigma_{\boldsymbol{\beta}}$. However, there is no closed form solution to the integral, and simulation methods are needed. A common and simple simulation method uses Monte-Carlo integration:

$$\widehat{\mathcal{L}_{\mathrm{N}}}(\boldsymbol{\beta}, \Sigma_{\boldsymbol{\beta}}) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{j=1}^{m} y_{ij} \ln \left[\frac{1}{S} \sum_{s=1}^{S} \frac{e^{\boldsymbol{x}'_{ij} \boldsymbol{\beta}_i^{(s)}}}{\sum_{l=1}^{m} e^{\boldsymbol{x}'_{il} \boldsymbol{\beta}_i^{(s)}}} \right], \tag{56}$$

where $\boldsymbol{\beta}_i^{(s)}$ for $s = 1, ..., S$ are random draws from the density $\phi(\boldsymbol{\beta}_i; \boldsymbol{\beta}, \Sigma_{\boldsymbol{\beta}})$. Given that $\boldsymbol{\beta}$ and $\Sigma_{\boldsymbol{\beta}}$ are unknown, this estimation is an iterative problem.

Multinomial Probit (MNP) An obvious way to introduce correlation of unobservables across alternatives is to specify them to be distributed as a multivariate normal, $\varepsilon \sim \mathcal{N}(\mathbf{0}, \Sigma)$. Different models arise from different assumptions on Σ. Even though off-diagonal elements can be set to non-zero values, some restrictions need to be placed for identification. From the ARUM we know that choices are determined by differences in utilities; for this reason, some elements of Σ (at least all the corresponding elements of the base category plus one variance) are not identified. The most common normalization is to set one of the variances —and hence all covariances involving this alternative— equal to zero, and to fix an additional parameter (e.g. , in the bivariate case, $\sigma_{11} = \sigma_{12} = 0$ and $\sigma_{22} = 1$, which leads to $\varepsilon_2 - \varepsilon_1 \sim \mathcal{N}(0, 1)$, the binary probit model). If regressors are alternative-invariant, additional restrictions may be needed to avoid obtaining very imprecise estimates.

The conditional choice probabilities are given by the $(m-1)$-dimensional integral over the normal distribution. For instance, if $m = 3$:

$$\Pr[y = 1|\boldsymbol{x}] = \int_{-\infty}^{-\tilde{V}_{31}} \int_{-\infty}^{-\tilde{V}_{21}} \phi(\tilde{\varepsilon}_{21}, \tilde{\varepsilon}_{31}; \mathbf{0}, \Sigma) d\tilde{\varepsilon}_{21} d\tilde{\varepsilon}_{31}. \tag{57}$$

The absence of closed form for this integral requires the use of simulation methods such as Monte Carlo integration to evaluate the log-likelihood, as in the case of the RPL:

$$\widehat{\mathcal{L}_{\mathrm{N}}}(\boldsymbol{\beta}, \Sigma) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{j=1}^{m} y_{ij} \ln \widehat{p}_{ij}, \tag{58}$$

where $\widehat{p}_{ij}$ are obtained using Monte-Carlo simulation.

F. Ordered Outcomes

All the previous analysis in this section is developed for unordered outcomes. We have seen a latent variable interpretation of all this by means of the ARUM.

which may be a subset of K.

The analysis of ordered multinomial outcomes corresponds to the index function latent variable interpretation that we have seen for binomial models. So consider the index function model for the latent variable y^*:

$$y^* = \boldsymbol{x}'\boldsymbol{\beta} + u, \quad u|\boldsymbol{x} \sim F(\cdot). \tag{59}$$

The variable that we observe is y, which is given by:

$$y = j \text{ if } \alpha_{j-1} < y^* \leq \alpha_j. \tag{60}$$

Therefore, the probability of choosing alternative j is given by:

$$\begin{aligned}\Pr[y = j|\boldsymbol{x}] = \Pr[\alpha_{j-1} < y^* \leq \alpha_j|\boldsymbol{x}] &= \Pr[\alpha_{j-1} - \boldsymbol{x}'\boldsymbol{\beta} < u \leq \alpha_j - \boldsymbol{x}'\boldsymbol{\beta}|\boldsymbol{x}] \\ &= F(\alpha_j - \boldsymbol{x}'\boldsymbol{\beta}) - F(\alpha_{j-1} - \boldsymbol{x}'\boldsymbol{\beta}). \end{aligned} \tag{61}$$

The parameters are then estimated by maximizing the general log-likelihood function given by these probabilities.

III. Endogenous Variables

There are several approaches to deal with endogeneity of regressors in the context of discrete choice. The most common way to proceed is to assume normality and proceed with a probit model. However, either if the number of endogenous variables is large or if the outcome variable is a large-dimensional multinomial outcome, this may be unfeasible or may impose too much structure. An alternative is GMM. We discuss both methods.

For simplicity, we proceed throughout this section with a binary outcome and only one endogenous regressor. Results may be generalized both to multinomial outcomes and to several endogenous regressors.

A. Probit with Continuous Endogenous Regressor

Consider the model:

$$y_1 = \mathbb{1}\{\boldsymbol{x}'\boldsymbol{\alpha} + \beta y_2 + \varepsilon \geq 0\}, \tag{62}$$

$$y_2 = \boldsymbol{z}'\boldsymbol{\gamma} + \nu, \tag{63}$$

where $\boldsymbol{z}$ strictly contains $\boldsymbol{x}$, and:

$$\begin{pmatrix}\varepsilon \\ \nu\end{pmatrix} \Bigg| \boldsymbol{z} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}1 & \rho\sigma \\ \rho\sigma & \sigma^2\end{bmatrix}\right). \tag{64}$$

Note that endogeneity is introduced by $\rho \neq 0$. To see it, note that the conditional distribution of ε given $\boldsymbol{z}$ and ν is given by:

$$\varepsilon|\boldsymbol{z}, \nu \sim \mathcal{N}\left(\frac{\rho}{\sigma}\nu, 1-\rho^2\right). \tag{65}$$

Using this expression, we can factorize the conditional likelihood function as $f(y_1|\boldsymbol{z}, y_2)f(y_2|\boldsymbol{z})$. Therefore, the log-likelihood of a random sample of N independent observations conditional on Z is:

$$\mathcal{L}_{\mathrm{N}}(\boldsymbol{\alpha}, \beta, \rho, \sigma, \boldsymbol{\gamma}) \propto \sum_{\mathrm{i}=1}^{\mathrm{N}} \left\{ y_{1i} \ln \Phi\left(a\right) + (1-y_{1i}) \ln\left[1-\Phi\left(a\right)\right] - \ln \sigma - \frac{(y_{2i}-\boldsymbol{z}_i'\boldsymbol{\gamma})^2}{2\sigma^2} \right\}, \tag{66}$$

where $a = \frac{\boldsymbol{x}_i'\boldsymbol{\alpha}+\beta y_{2i}+\frac{\rho}{\sigma}(y_{2i}-\boldsymbol{z}_i'\boldsymbol{\gamma})}{\sqrt{1-\rho^2}}$.

The parameter vector $\theta \equiv (\boldsymbol{\alpha}', \beta, \rho, \sigma, \boldsymbol{\gamma}')'$ can be estimated by ML on the previous expression. If this is costly or the log-likelihood is not very well behaved, we may proceed in two steps, with a LIML estimation in the same spirit of Exercise 1 of Chapter 1. The estimates obtained by the two-step estimation are less efficient, but still consistent. Standard errors of the second stage should be corrected to account for the fact that we are using estimates of ν instead of observed values.

B. Probit with Binary Endogenous Regressor

In this case, the model is:

$$y_1 = \mathbb{1}\{\boldsymbol{x}'\boldsymbol{\alpha} + \beta y_2 + \varepsilon \geq 0\}, \tag{67}$$

$$y_2 = \mathbb{1}\{\boldsymbol{z}'\boldsymbol{\gamma} + \nu \geq 0\}, \tag{68}$$

where $\boldsymbol{z}$ strictly contains $\boldsymbol{x}$, and:

$$\begin{pmatrix}\varepsilon \\ \nu\end{pmatrix} \Bigg| \boldsymbol{z} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}1 & \rho \\ \rho & 1\end{bmatrix}\right). \tag{69}$$

This is, in fact, a bivariate binomial Probit. There is no two-step procedure to estimate the parameter vector, and, hence, the estimation should be performed by maximizing the conditional log-likelihood, which is given by:

$$\begin{aligned}\mathcal{L}_{\mathrm{N}}(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}, \rho) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \{ & y_{1i}y_{2i} \ln P_{11i} + (1-y_{1i})y_{2i} \ln P_{01i} + \qquad (70) \\ & + y_{1i}(1-y_{2i}) \ln P_{10i} + (1-y_{1i})(1-y_{2i}) \ln P_{00i} \},\end{aligned}$$

where:

$$
\begin{aligned}
P_{00} &\equiv \Pr[y_1 = 0, y_2 = 0|\boldsymbol{z}] = \Phi_2(-\boldsymbol{x}'\boldsymbol{\alpha}, -\boldsymbol{z}'\boldsymbol{\gamma}; \rho), && (71)\\
P_{10} &\equiv \Pr[y_1 = 1, y_2 = 0|\boldsymbol{z}] = \Pr[\varepsilon \geq -\boldsymbol{x}'\boldsymbol{\alpha}|y_2 = 0, \boldsymbol{z}] \Pr[y_2 = 0|\boldsymbol{z}] \\
&\qquad = (1 - \Pr[\varepsilon \leq -\boldsymbol{x}'\boldsymbol{\alpha}|y_2 = 0, \boldsymbol{z}]) \Pr[y_2 = 0|\boldsymbol{z}] \\
&\qquad = \Phi(-\boldsymbol{z}'\boldsymbol{\gamma}) - P_{00}, && (72)\\
P_{01} &\equiv \Pr[y_1 = 0, y_2 = 1|\boldsymbol{z}] = \Pr[y_2 = 1|\varepsilon \leq -\boldsymbol{x}'\boldsymbol{\alpha} - \beta, \boldsymbol{z}] \Pr[\varepsilon \leq -\boldsymbol{x}'\boldsymbol{\alpha} - \beta|\boldsymbol{z}] \\
&\qquad = (1 - \Pr[y_2 = 0|\varepsilon \leq -\boldsymbol{x}'\boldsymbol{\alpha} - \beta, \boldsymbol{z}]) \Pr[\varepsilon \leq -\boldsymbol{x}'\boldsymbol{\alpha} - \beta|\boldsymbol{z}] \\
&\qquad = \Phi(-\boldsymbol{x}'\boldsymbol{\alpha} - \beta) - \Phi_2(-\boldsymbol{x}'\boldsymbol{\alpha} - \beta, -\boldsymbol{z}'\boldsymbol{\gamma}; \rho), && (73)\\
P_{11} &\equiv \Pr[y_1 = 1, y_2 = 1|\boldsymbol{z}] = 1 - P_{00} - P_{10} - P_{01}, && (74)
\end{aligned}
$$

and $\Phi_2(.,.;\rho)$ indicates the cdf of a bivariate standard normal distribution with correlation parameter ρ. Simulation-based methods can be used to obtain the different probabilities.

C. Moment Estimation

A different approach to handle endogeneity is with a moment-based estimation instead of ML. Assuming that probabilities are correctly specified, we can consider an estimator that is the solution of:

$$\sum_{\text{i=1}}^{\text{N}} \sum_{j=1}^{m} (y_i - p_{ij})\boldsymbol{z}_i = \boldsymbol{0}, \tag{75}$$

where $\boldsymbol{z}$ strictly contains the set of exogenous regressors and is of the same dimension as the vector of regressors.[4] This estimator is consistent, and its efficiency depends on the choice of $\boldsymbol{z}$. It is worth noting that equation (75) coincides with the expression for the MNL estimator if $\boldsymbol{z} = \boldsymbol{x}$.

IV. Binary Models for Panel Data

Consider the binary choice panel data model with individual effects:[5]

$$y_{it} = \mathbb{1}\{\boldsymbol{x}_{it}'\boldsymbol{\beta} + \eta_i + v_{it} > 0\}. \tag{76}$$

This is a non-linear panel data model. In particular, it belongs to the subclass in which errors are not additively separable. This distinction is important because

[4] If we have more instruments than regressors, when our GMM problem is overidentified, and the estimator is the result of minimizing a combination of the $k^z > k^x$ conditions given by the equation (75).

[5] We retake the convention of using subscripts for random variables in the panel data context.

additively separable models allow the construction of moment conditions that mimic the linear ones. This is not the case for non-additively separable models.

The estimation problem can be addressed from a fixed effects or a random effects perspective (in the sense of treating η_i as parameters to be estimated or as a random variable from a given distribution). From a fixed effects perspective, the log-likelihood is:

$$\mathcal{L}_{\mathrm{N}}(\boldsymbol{\beta},\boldsymbol{\eta}) = \sum_{\mathrm{i=1}}^{\mathrm{N}}\sum_{t=1}^{T}\{y_{it}\ln F(\boldsymbol{x}_{it}'\boldsymbol{\beta}+\eta_i)+(1-y_{it})\ln(1-F(\boldsymbol{x}_{it}'\boldsymbol{\beta}+\eta_i))\}. \quad (77)$$

In this case, the vector of fixed effects, $\boldsymbol{\eta}$, is a vector of *nuisance parameters*, this is, parameters that are not of immediate interest, but which must be accounted for in the analysis of those parameters which are of interest. The problem with this approach, however, is that the number of nuisance parameters to estimate becomes very large when N is relatively large compared to T. Often, we get rid of these parameters through the concentrated likelihood, $\mathcal{L}_{\mathrm{N}}(\boldsymbol{\beta},\hat{\boldsymbol{\eta}}(\boldsymbol{\beta}))$, where $\hat{\boldsymbol{\eta}}(\boldsymbol{\beta})$ is the MLE of $\boldsymbol{\eta}$ for a given $\boldsymbol{\beta}$.

Another approach is from the random effects perspective. In this case, we optimize the integrated likelihood, like in the Random Parameters Logit case. Specifically:

$$\mathcal{L}_{\mathrm{N}}(\boldsymbol{\beta}) = \sum_{\mathrm{i=1}}^{\mathrm{N}}\sum_{t=1}^{T}\ln\int f(y_{it}|\boldsymbol{x}_{it};\boldsymbol{\beta},\eta_i)g(\eta_i;\boldsymbol{\gamma})d\eta_i, \quad (78)$$

where $g(\eta_i;\boldsymbol{\gamma})$ can or cannot be the pdf of η_i. In the latter case, $\mathcal{L}_{\mathrm{N}}(\boldsymbol{\beta})$ is a pseudo-likelihood that can still deliver consistent estimates as $N\to\infty$ and $T\to\infty$, but that produces inconsistent estimators as $N\to\infty$ for fixed T. In particular, it produces biases of order $1/T$. This situation is known as the *incidental parameters problem* and it is of particular concern when T is small relative to N. This problem also applies to the fixed effects case, as the concentrated likelihood can indeed be expressed in the form of (78) with a specific functional form for g. The incidental parameters problem is one of the main challenges in modern econometrics.