Advanced Econometrics

Discrete Choice Models

- Binary Choice Models
  - Linear Probability Model
  - Index Models for Binary Response
  - Latent variable threshold (LVT) model
  - Random utility models
  - Simple specification tests
  - Goodness of Fit
- Multinomial choice models
  - Multinomial Logit
  - Probabilistic Choice Models

In many applications, the variate of interest is binary, i.e., takes only the values 0 and 1.

Examples:

- Labour force participations.

$$Y = \begin{cases} 1 & \text{if employed} \\ 0 & \text{otherwise} \end{cases} .$$

  We would like to study how labour force participation depends on the characteristics of the individuals.

- House ownership

$$Y = \begin{cases} 1 & \text{if a person owns her house} \\ 0 & \text{otherwise} \end{cases} .$$

  We would like to study how house ownership depends on the characteristics of the individuals.

- Denote $X = (X_1, ..., X_k)'$.
- The objective of a regression model is to estimate $E(Y|X)$.

- $E(Y|X) = \mathcal{P}(Y = 1|X)$, when Y is a binary variable.
- In the *linear probability model* we assume that

$$\mathcal{P}(Y = 1|X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k.$$

- So, the interpretation of β_j is the change in the probability of success when x_j changes:

$$\frac{\partial \mathcal{P}(Y = 1|X)}{\partial X_j} = \beta_j, j = 1, ..., k$$

- The predicted Y is the predicted probability of success.
- The linear probability model is estimated using OLS, that is regressing Y on $X_1, ..., X_k$.

- Potential problem that the fitted values can be outside $[0, 1]$.
- Even without predictions outside of $[0, 1]$, we may estimate effects that imply a change in x changes the probability by more than $+1$ or -1.
- This model will violate assumption of homoskedasticity, so will affect inference. Notice that

$$\begin{aligned} Var(Y|X) &= \mathcal{P}(Y = 1|X)(1 - \mathcal{P}(Y = 1|X)) \\ &= (\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k) \times \\ &\quad (1 - \beta_0 - \beta_1 X_1 - \ldots - \beta_k X_k). \end{aligned}$$

- Therefore we should use the Eicker-Huber-White robust standard errors to make inference.

- An alternative is to assume that $E[Y|X] = \mathcal{P}(Y=1|X) = G(X'\beta_0)$, where the function $G(.)$ is known $0 < G(.) < 1$ thus

$$Y = \begin{cases} 1 & \text{with probability } G(X'\beta_0) \\ 0 & \text{with probability } 1 - G(X'\beta_0) \end{cases}$$

- In most applications, $G(.)$ is a cumulative distribution function.
- The framework is similar to the case of the Bernoulli random variable (conditional on the regressors). The Log-Likelihood function is given by

$$\log\{L(\beta)\} = \sum_{i=1}^{n} Y_i \log(G(X_i'\beta)) + \sum_{i=1}^{n} (1 - Y_i) \log(1 - G(X_i'\beta)).$$

Differentiating with respect to β we have that the MLE estimator $\hat{\beta}_{ML}$ solves

$$
\begin{aligned}
\frac{\partial \log\{\mathcal{L}(\hat{\beta}_{ML})\}}{\partial \beta} &= 0 \\
\sum_{i=1}^{n} \left\{ \frac{Y_i - G(X_i'\hat{\beta}_{ML})}{G(X_i'\hat{\beta}_{ML})\left(1 - G(X_i'\hat{\beta}_{ML})\right)} g(X_i'\hat{\beta}_{ML}) X_i \right\} &= 0
\end{aligned}
$$

where $g(z) = \partial G(z)/\partial z$.

- Define the **generalized residuals** as

$$\hat{\varepsilon}_i^G = \frac{Y_i - G(X_i'\hat{\beta}_{ML})}{G(X_i'\hat{\beta}_{ML})\left[1 - G(X_i'\hat{\beta}_{ML})\right]} g(X_i'\hat{\beta}_{ML})$$

- Likelihood equations are then given by:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^G X_i = 0.$$

This condition requires $\hat{\varepsilon}_i^G$ and X_i are uncorrelated.

Remarks:

- This is a system of non-linear equations hence we have to resort to numerical methods to solve it. There is no closed form solution for this estimator.
- Consistency and asymptotic normality follow from the general results described for the Maximum Likelihood estimator under some regularity conditions.
- Essentially the main requirement for consistency is that $E[Y|X] = \mathcal{P}(Y = 1|X) = G(X'\beta_0)$.

Note that this implies that

$$E[\frac{\partial \log\{L(\beta_0)\}}{\partial \beta}] = 0.$$

Proof:

$$
\begin{aligned}
E[\frac{\partial \log\{L(\beta_0)\}}{\partial \beta}] &= \sum_{i=1}^{n} E[\frac{Y_i - G(X_i'\beta)}{G(X_i'\beta)\left(1 - G(X_i'\beta)\right)} g(X_i'\beta) X_i] \\
&= \underbrace{\sum_{i=1}^{n} E(E[\frac{Y_i - G(X_i'\beta)}{G(X_i'\beta)\left(1 - G(X_i'\beta)\right)} g(X_i'\beta) X_i | X_i])}_{\text{by the law of iterated expectations}} \\
&= \sum_{i=1}^{n} E(\frac{E[Y_i|X_i] - G(X_i'\beta)}{G(X_i'\beta)\left(1 - G(X_i'\beta)\right)} g(X_i'\beta) X_i) \\
&= 0 \text{ as } E[Y_i|X_i] = G(X_i'\beta_0)
\end{aligned}
$$

- Note that if $E[Y_i|X_i] \neq G(X_i'\beta_0) \Rightarrow E[\frac{\partial \log\{\mathcal{L}(\beta_0)\}}{\partial \beta}] \neq 0$, which can be shown to imply inconsistency of MLE.

- In correctly specified models $\hat{\beta}$ is consistent and asymptotically normally distributed with variance-covariance matrix $[\mathcal{I}(\beta_0)]^{-1}$, that is

$$\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{D} \mathcal{N}(0, [\mathcal{I}(\beta_0)]^{-1})$$

  where

$$\mathcal{I}(\beta_0) = E\{\frac{g(X'\beta_0)^2 XX'}{G(X'\beta_0)[1-G(X'\beta_0)]}\}$$

- An estimator for $\mathcal{I}(\beta_0)$ is

$$\mathcal{I}_n(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\{\frac{g(X_i'\hat{\beta}_{ML})^2 X_i X_i'}{G(X_i'\hat{\beta}_{ML})[1-G(X_i'\hat{\beta}_{ML})]}\}$$

- Inference is done using the Wald, likelihood ratio and Lagrange multiplier statistics.

- The most popular forms of $G(X'\beta_0)$ that are considered in the literature
  - The Logit Model:
$$G(X'\beta_0) = \frac{\exp(X'\beta_0)}{1+\exp(X'\beta_0)}.$$
  - The Probit Model:
$$G(X'\beta_0) = \Phi(X'\beta_0),$$
    where
$$\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-u^2/2}du.$$
    is the Standard Normal Distribution Function.
- Both the probit and logit models are nonlinear and require maximum likelihood estimation.
- No real reason to prefer one over the other

Other possible models:

- $G(X'\beta_0) = \exp(-\exp(X'\beta_0))$ [the log-Weibull distribution]
- $G(X'\beta_0) = 1 - \exp(-\exp(X'\beta_0))$ [the Gompertz distribution, known as the Complementary log-log model]
- $G(X'\beta_0) = \Phi(X'\beta_0)^{\tau}$, $\tau > 0$
- $G(X'\beta_0) = 1 - \left(1 + \omega \exp(X'\beta_0)\right)^{-\frac{1}{\omega}}$, $\omega > 0$. Note that for $\omega = 1$ we have the logit model and $\lim_{\omega \to 0} G(X'\beta_0) = 1 - \exp(-\exp(X'\beta_0))$.

Remark on the Logit Model

- In statistics a common interpretation of the coefficients is in terms of marginal effects on the odds ratio rather than on the probability.

$$\begin{aligned} \mathcal{P}(Y &= 1|x) = \frac{\exp(X'\beta_0)}{1+\exp(X'\beta_0)} \\ &\Rightarrow \frac{\mathcal{P}(Y=1|x)}{1-\mathcal{P}(Y=1|x)} = \exp(X'\beta_0) \\ &\Rightarrow \log(\frac{\mathcal{P}(Y=1|x)}{1-\mathcal{P}(Y=1|x)}) = X'\beta_0 \end{aligned}$$

- $\mathcal{P}(Y=1|X)/(1-\mathcal{P}(Y=1|X))$ measures the probability that $Y=1$ relative to the probability that $Y=0$ and is called the odds ratio or relative risk.
- Example, consider a pharmaceutical drug study where $Y=1$ denotes survival and $Y=0$ denotes death. An odds ratio of 2 means that the odds of survival are twice those of death.
- For the logit model the log-odds ratio is linear in the regressors.

Latent variable threshold (LVT) model

- A possible motivation for the specification $E[Y|X] = \mathcal{P}(Y = 1|X) = G(X'\beta_0)$ can be given by considering the *latent variable threshold model*
- Define a latent random variable:

$$Y^* = X'\beta_0 + \varepsilon,$$

  where Y^* is unobserved $\Rightarrow$ **latent variable.**
- Assume: ε independent of X, $E[\varepsilon] = 0$ and $var(\varepsilon) = \sigma^2$ and distribution function $F(.)$
- **Observation rule:**

$$Y = \begin{cases} 1 & \text{if } Y^* > \lambda \\ 0 & \text{if } Y^* \leq \lambda \end{cases} .$$

  That is, the option is chosen if $Y^* > \lambda$, where λ is a threshold
- **Interpretation:** Y^* propensity of an individual towards option, or net benefit from choosing option.

Latent variable threshold (LVT) model

- Probability of choosing the option:

$$
\begin{aligned}
\mathcal{P}[Y=1|X] &= \mathcal{P}[Y^* > \lambda|X] \\
&= \mathcal{P}\left[X'\beta_0 + \varepsilon > \lambda|X\right] \\
&= \mathcal{P}\left[\varepsilon > -X'\beta_0 + \lambda|X\right] \\
&= 1 - \mathcal{P}\left[\varepsilon \leq -X'\beta_0 + \lambda|X\right] \\
&= 1 - F\left(-X'\beta_0 + \lambda\right). \\
&= G(X'\beta)
\end{aligned}
$$

with $G(z) = 1 - F(-z + \lambda)$.

- Note that:
$$\mathcal{P}[Y = 1|X] = 1 - F\left(-X'\beta_0 + \lambda\right)$$
- If $X_1 = 1$, i.e. there is an intercept in the model, it is not possible to identify separately the intercept and $\lambda \Rightarrow$ solution: set $\lambda = 0$.
- **Remark:** If $\lambda = 0$ and ε has a symmetric distribution around zero (as in the Probit or Logit) $G(z) = F(z)$ as in this case $1 - F(-z) = F(z)$

Latent variable threshold (LVT) model

Second identification problem:

- Divide Y^* by $a > 0$

$$\frac{Y^*}{a} = X'\beta_0^* + \frac{\varepsilon}{a}$$

where $\beta_0^* = \beta_0 \,/a$

- Note that the definition of the observable variable Y doesn't change. That is

$$\begin{aligned} Y &= \begin{cases} 1 & \text{if } Y^* > 0 \\ 0 & \text{if } Y^* \leq 0 \end{cases} \\ &= \begin{cases} 1 & \text{if } \frac{Y^*}{a} > 0 \\ 0 & \text{if } \frac{Y^*}{a} \leq 0 \end{cases} \end{aligned}$$

where $\beta_0^* = \beta_0 \,/a$.

- This implies that we cannot identify the variance of ε.
- For given β_0^*, value of β_0 depends on a.
- β_0 identified up to a scale factor.
- **Solution:** normalise distribution of ε- Fix σ^2 at a given number $\Rightarrow$ Assume σ^2 is **known**.

- **Example:**
  - Suppose $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.
  - $P[Y = 1|X] = \Phi(X' \frac{\beta_0}{\sigma}) = \Phi(X' \beta_0^*)$.
  - In the case of Probit model we fix $\sigma^2 = 1$ thus $\varepsilon \sim \mathcal{N}(0, 1)$ and:

$$\mathcal{P}[Y = 1|X] = \Phi(X' \beta_0).$$

Random utility models

- Suppose that an individual has to choose between alternatives a and b, with utilities U^a and U^b.
- The researcher does not observe the utilities, but observes some characteristics of the observation, and writes

$$\begin{aligned} U^a &= X'\beta_a + u_a, \\ U^b &= X'\beta_b + u_b. \end{aligned}$$

  - The researcher observes the chosen alternative, say a, which is indicated by $Y = 1$.
  - Then, we know that

$$\begin{aligned} \mathcal{P}(Y = 1|X) &= \mathcal{P}\left(U^a > U^b|X\right) = \Pr\left(X'\beta_a + u_a > X'\beta_b + u_b|X\right) \\ &= \mathcal{P}\left(u_a - u_b > X'\left(\beta_b - \beta_a\right)|X\right) \\ &= \mathcal{P}\left(\varepsilon > -X'\beta_0|X\right) = 1 - F\left(-X'\beta_0\right). \end{aligned}$$

  where $\varepsilon = u_a - u_b$ and $\beta_0 = \beta_a - \beta_b$

- **Whatever the interpretation**, we have to make inference about $\mathcal{P}(Y = 1|x)$.

- In general we care about the effect of X on $E(Y|X) = \mathcal{P}(Y = 1|X)$, that is, we care about $\partial \mathcal{P}(Y = 1|X)/\partial X_j$, $j = 2, ..., k$
- For the linear case, this is easily computed as the coefficient on X_j
- For the nonlinear probit and logit models, it's more complicated: $\partial \mathcal{P}(Y = 1|X)/\partial X_j = g(X'\beta_0)\beta_{0j}$, where $g(z)$ is $\partial G(z)/\partial z$ and β_{0j} is the element j of β_0.
- Clear that it's incorrect to just compare the coefficients across different models
- Can compare sign and significance (based on a standard t test) of coefficients, though
- To compare the magnitude of effects, need to calculate the derivatives, say at the means of the regressors

Simple specification tests

As pointed out above if $G(.)$ is misspecified, then $\hat{\beta}_{ML}$ is inconsistent. Some simple specification tests are available:

- A RESET-type test can be performed by testing $H_0 : \delta_1 = \delta_2 = 0$ in the model

$$E[Y_i|X_i] = G(X_i'\beta_0 + \delta_1(X_i'\hat{\beta}_{ML})^2 + \delta_2(X_i'\hat{\beta}_{ML})^3), i = 1, ..., n$$

  * This is actually a normality test in the probit.

- The model can be tested against more general parametric specifications, which include additional shape parameters. Examples:
  - Consider $G(X'\beta_0) = \Phi(X'\beta_0)^\tau$, and use the score statistic to test $H_0 : \tau = 1$ (Probit)
  - Consider $G(X'\beta_0) = 1 - \left(1 + \omega \exp(X'\beta_0)\right)^{-\frac{1}{\omega}}$, $\omega > 0$. and use the score statistic to test $H_0 : \omega = 1$ (Logit).

Note that heteroskedasticity in the LVT model leads to misspecification of the conditional mean of Y: Define a latent random variable:

$$Y^* = X'\beta_0 + k \times h(Z'\gamma_0)\varepsilon,$$

where Y^* is unobserved. Assume ε independent of X, $E[\varepsilon] = 0$ and $var(\varepsilon) = 1$ and distribution function $F(.)$, Z are a vector function of X of size d and h any function with $h > 0$, $h(0) = 1$, $h'(0) \neq 0$

- $k = 1$ for probit; $k = \sqrt{\pi^2/3}$ for logit.
- Observation rule:

$$Y = \begin{cases} 1 & \text{if } Y^* > 0 \\ 0 & \text{if } Y^* \leq 0 \end{cases} .$$

- In this case

$$\begin{aligned}
\mathcal{P}[Y=1|X] &= \mathcal{P}[Y^* > 0|X] \\
&= \mathcal{P}\left[X'\beta_0 + kh(Z'\gamma_0)\varepsilon > 0|X\right] \\
&= \mathcal{P}\left[\varepsilon > -\frac{X'\beta_0}{kh(Z'\gamma_0)}|X\right] \\
&= 1-\mathcal{P}\left[\varepsilon \le -\frac{X'\beta_0}{kh(Z'\gamma_0)}|X\right] \\
&= 1-F\left(-\frac{X'\beta_0}{kh(Z'\gamma_0)}\right). \\
&= G(\frac{X'\beta_0}{kh(Z'\gamma_0)}) \neq G(X'\beta_0)
\end{aligned}$$

- To test the hypothesis $H_0 : \gamma_0 = 0$ (homoskedasticity), we can construct a LM test based on the so called generalized residuals

- LM test statistic can be calculated as

$$\xi_{LM} = \iota' S(S'S)^{-1} S' \iota \sim \chi^2(d)$$

where ith row of S equal to

$$S_i = (\hat{\varepsilon}_i^G X_i', \hat{\varepsilon}_i^G (X_i' \hat{\beta}_{ML}) Z_i')$$

where $\hat{\varepsilon}_i^G$ are the Generalised residuals.

- This is asymptotically equivalent to testing $H_0 : \gamma_0 = 0$ in the model

$$E[Y_i | X_i] = G(X_i' \beta_0 + (X_i' \hat{\beta}_{ML}) Z_i' \gamma_0)), i = 1, ..., n.$$

- Unlike the Linear Probability Model, where we can compute an R^2 to judge goodness of fit, we need new measures of goodness of fit
- One possibility is a pseudo R^2 based on the log likelihood and defined as $1 - \log(\mathcal{L}_{ur})/\log(\mathcal{L}_r)$. Where $\log(\mathcal{L}_r)$ corresponds to the log-likelihood computed only with the intercept.
- Can also look at the percent correctly predicted – if predict a probability $> .5$ then that matches $Y = 1$ and vice versa.

- Two ways to extend the binary response: *unordered* and *ordered outcomes*. In both cases, it is convenient to label the possible outcomes on Y as $\{0, 1, ..., J\}$, so Y takes on $J+1$ different values.
- In the *unordered* (or nominal) case, the labeling of outcomes is totally arbitrary. For example, if Y is mode of transportation to work, we might use the follow labels: 0 is by car without pooling, 1 is car pooling, 2 is bus, and 3 is rapid transit (metro). Nothing changes if we switch the labels.
- Another example of an unordered outcome is different kinds of health insurance.

- In other cases the order matters. For example, each person applying for a mortgage is given a credit rating in the set $\{0, 1, 2, 3, 4, 5, 6\}$. The fact that a credit rating of 5 is better than 4, and that 1 is better than 0, is important.
- In this chapter we will discuss the estimation of *unordered response models* and leave the discussion of ordered response models for the next chapter.

Multinomial Logit

- Start with the case where Y is an *unordered outcome* taking on values in $\{0, 1, ..., J\}$. Assume we have conditioning variables, $\mathbf{X}$, that change with the unit (i.e. observation) but not with the alternative.
- For example, in modeling type of health insurance, we include observable characteristics of the individual but not characteristics of the different kinds of health plans. For occupational choice, $\mathbf{X}$ can include years of schooling, age, gender, and so on – but not characteristics of the occupations.

Multinomial Logit

- In this setting, we are interested in the *response probabilities*,

$$p_j(\mathbf{X}) = \mathcal{P}(Y = j|\mathbf{X}), j = 0, ..., J.$$

- Because one and only one choice is possible,

$$p_0(\mathbf{X}) + p_1(\mathbf{X}) + \ldots + p_J(\mathbf{X}) = 1 \text{ for all } \mathbf{X}$$

- We are interested in how changing elements of **X** affects the response probabilities.

- In the basic *multinomial logit (MNL) model*, the response probabilities are

$$\mathcal{P}(Y = j|\mathbf{X}) = \frac{\exp(\mathbf{X}'\boldsymbol{\beta}_j)}{\left[1+\sum_{h=1}^{J}\exp(\mathbf{X}'\boldsymbol{\beta}_h)\right]}, \; j = 1, ..., J$$
$$\mathcal{P}(Y = 0|\mathbf{X}) = \frac{1}{\left[1+\sum_{h=1}^{J}\exp(\mathbf{X}'\boldsymbol{\beta}_h)\right]}$$

where in almost all applications $X_1 \equiv 1$ (the first element of $\mathbf{X}$).

Multinomial Logit

- Unless $J = 1$ (*binary response logit*), the partial effects on the $p_j(\cdot)$ are complicated. For a continuous X_k (k^{th} element of **X)**,

$$\frac{\partial p_j(\mathbf{X})}{\partial X_k} = p_j(\mathbf{X}) \left\{ \beta_{jk} - \frac{\left[\sum_{h=1}^{J} \beta_{hk} \exp(\mathbf{X}' \boldsymbol{\beta}_h)\right]}{\left[1 + \sum_{h=1}^{J} \exp(\mathbf{X}' \boldsymbol{\beta}_h)\right]} \right\},$$

  where β_{hk} is the k^{th} element of $\boldsymbol{\beta}_h$. $\partial p_j(\mathbf{X})/\partial X_k$ might not have the same sign as β_{jk}.

- Easier to interpret is the response on $p_j(\mathbf{X})$ relative to $p_0(\mathbf{X})$:

$$\begin{aligned} r_j(\mathbf{X}) &\equiv \frac{p_j(\mathbf{X})}{p_0(\mathbf{X})} = \exp(\mathbf{X}' \boldsymbol{\beta}_j) \\ \frac{\partial r_j(\mathbf{X})}{\partial X_k} &= \beta_{jk} \exp(\mathbf{X}' \boldsymbol{\beta}_j) \end{aligned}$$

Multinomial Logit

- The *log odds* of response j relative to response 0 is

$$logodds_j(\mathbf{X}) \equiv \log\left[\frac{p_j(\mathbf{X})}{p_0(\mathbf{X})}\right] = \mathbf{X}'\boldsymbol{\beta}_j,$$

and so β_{jk} measures the partial effect of x_k on the log odds of j relative to outcome 0:

$$\frac{\partial logodds_j(\mathbf{X})}{\partial X_k} = \beta_{jk}.$$

- A key feature of the *MNL* model is that if we condition on the even that Y can take on any of two outcomes, the resulting model for choosing between the outcomes is a binary response logit.
- Formally, suppose we condition on the event that $Y \in \{j, h\}$:

$$\begin{aligned}
\mathcal{P}(Y &= j|Y = j \text{ or } Y = h) = p_j(\mathbf{X}, \boldsymbol{\beta}) / [p_j(\mathbf{X}, \boldsymbol{\beta}) + p_h(\mathbf{X}, \boldsymbol{\beta})] \\
&= \frac{\exp(\mathbf{X}'\boldsymbol{\beta}_j)}{[\exp(\mathbf{X}'\boldsymbol{\beta}_j) + \exp(\mathbf{X}'\boldsymbol{\beta}_h)]} = \frac{\exp[\mathbf{X}'(\boldsymbol{\beta}_j - \boldsymbol{\beta}_h)]}{\{[\exp[\mathbf{X}'(\boldsymbol{\beta}_j - \boldsymbol{\beta}_h)] + 1\}} \\
&= \Lambda[\mathbf{X}'(\boldsymbol{\beta}_j - \boldsymbol{\beta}_h)]
\end{aligned}$$

where $\Lambda[\mathbf{a}] = \exp(a) / [1 + \exp(a)]$.

- The previous formula shows that $\mathcal{P}(Y = j | Y = j \text{ or } Y = h)$ has the logit form with parameter vector $\boldsymbol{\beta}_j - \boldsymbol{\beta}_h$.
- If we set $h = 0$ it follows that $\mathcal{P}(Y = j | Y = j \text{ or } Y = 0) = \Lambda(\mathbf{X}'\boldsymbol{\beta}_j)$, which means we can estimate $\boldsymbol{\beta}_j$ by using a binary response logit on the sample of people choosing either 0 or j.
- This simplification is an artifact of the *MNL* functional form.

Multinomial Logit

- Full maximum likelihood estimation of the $\boldsymbol{\beta}_j$ is straightforward. The log likelihood function is:

$$\log L(\boldsymbol{\beta}) = \sum_{i=0}^{n} \sum_{j=0}^{J} 1[Y_i = j] \log[p_j(\mathbf{X}_i, \boldsymbol{\beta})].$$

- Inference is standard. The expected Hessian given $\mathbf{X}_i$ is easy to compute.
- In terms of goodness of fit and prediction, the MNL model often works well. We can choose $\mathbf{X}$ to be flexible functions of underlying explanatory variables.

Probabilistic Choice Models

- Again, let there be $J+1$ choices, but now explicitly view the response (choice) as *maximizing underlying utility*. For a random draw i, the latent utilities are

$$U_{ij} = \mathbf{X}'_{ij}\boldsymbol{\beta} + a_{ij},\ j = 0, ..., J,$$

  where $\mathbf{X}_{ij}$ can vary by unit (i) and choice (j). Notice that $\boldsymbol{\beta}$, in this formulation, does not depend on j. It is almost always true that $\mathbf{X}_{ij}$ includes unity.
- **Example:** $\mathbf{X}_{ij}$ can include the costs of various modes of transportation j for each unit i. Its coefficient measures the effect of cost on utility across any mode of transportation.
- Sometimes a variable will change only by choice and not individual (such as the price of a car if geographic homogeneity is assumed).

Probabilistic Choice Models

- Let $\mathbf{X}_i$ include all nonredundant elements of $(\mathbf{X}_{i0}, \mathbf{X}_{i1}, ..., \mathbf{X}_{iJ})'$. Let $\mathbf{a}_i = (a_{i0}, a_{i1}, ..., a_{iJ})'$ and assume $\mathbf{a}_i$ is independent of $\mathbf{X}_i$ (exogeneity).
- The observed choice $Y_i \in \{0, 1, ..., J\}$ is the one that maximizes utility:

$$Y_i = \underset{j \in \{0,1,...,J\}}{\operatorname{argmax}} \{U_{ij}\};$$

  that is, $Y_i = j$ if choice j yields the highest utility.
- McFadden (1974) showed that if the $\{a_{ij} : j = 0, 1, ..., J\}$ are independent, identically distributed with the *type I extreme value distribution*, that is, with cdf $F(a) = \exp[-\exp(-a)]$, then it can be shown that

$$\mathcal{P}(Y_i = j | \mathbf{X}_i) = \frac{\exp(\mathbf{X}_{ij}'\boldsymbol{\beta})}{\left[1 + \sum_{h=1}^{J} \exp(\mathbf{X}_{ih}'\boldsymbol{\beta})\right]}, \ j = 0, 1, ..., J,$$

  where this expression uses a normalization $\mathbf{X}_{i0} \equiv \mathbf{0}$. (Equivalently, the covariates of choices $j = 1, ..., J$ are measured net of $\mathbf{X}_{i0}$.)

Probabilistic Choice Models

- Often it is useful to write

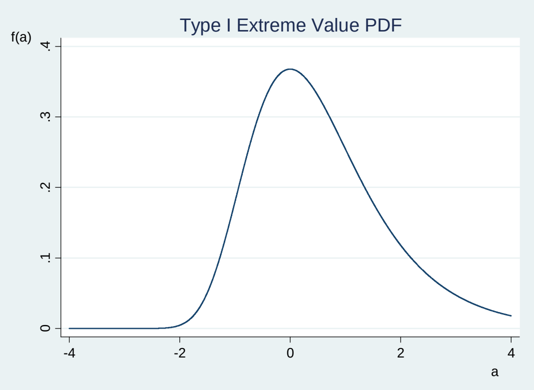
$$\mathcal{P}(Y_i = j|\mathbf{X}_i) = \frac{\exp(\mathbf{X}_{ij}'\boldsymbol{\beta})}{\left[\sum_{h=0}^{J}\exp(\mathbf{X}_{ih}'\boldsymbol{\beta})\right]}, \; j = 0, 1, ..., J,$$

in which case the $\mathbf{X}_{ij}$ are not measured net of $\mathbf{X}_{i0}$.

- In the context of probabilistic choice models, this is usually called the *conditional logit (CL) model* (the name given by McFadden).
- Fairly easy to estimate $\boldsymbol{\beta}$ by MLE, even for lots of choices.

- The type I extreme value distribution is perhaps not natural because it is not symmetric – it has a thicker right tail.
- The density for the type I extreme value distribution is

$$f(a) = \exp(-a)\exp(-\exp(-a))$$



- The MNL model can be shown to be a special case of the CL model.
- Suppose we have an MNL model with covariates $\mathbf{W}_i$ and parameters $\delta_1, \delta_2, ..., \delta_J$. Let d_{jh} be a dummy variable equal to 1 when $j = h$ and zero otherwise. Define $\mathbf{X}_{ij} = (d_{1j}\mathbf{W}_i, d_{2j}\mathbf{W}_i, ..., d_{Jj}\mathbf{W}_i)'$ and $\boldsymbol{\beta} = (\delta_1', \delta_2', ..., \delta_J')'$.
- Therefore for $j = 1, ..., J$ we have $\mathbf{X}_{ij}'\boldsymbol{\beta} = \mathbf{W}_i'\delta_j$.
- Consequently the focus is often on CL model.

Remark: McFadden shared the 2000 Nobel Memorial Prize in Economic Sciences with James Heckman. McFadden's share of the prize was "*for his development of theory and methods for analyzing discrete choice*".

Probabilistic Choice Models

- This model has the *Independence of Irrelevant Alternatives* (IIA) property which means that for any pair (j, l) the odds ratio

$$\frac{\Pr(Y_i = j|\mathbf{X}_i)}{\Pr(Y_i = \ell|\mathbf{X}_i)} = \frac{\exp(\mathbf{X}'_{ij}\beta)}{\exp(\mathbf{X}'_{i\ell}\beta)}$$

  does not depend on the characteristics or availability of any other options.
- This is called the i*ndependence from irrelevant alternatives* (IIA) assumption because it implies that adding another alternative or changing the characteristics of a third alternative does not affect the relative odds between alternatives.
- *IIA* can have unattractive implications for the probabilities when alternatives are similar, and for predicting substitution patterns when new alternatives are introduced or old choices are taken away.

Probabilistic Choice Models

- **Red Bus/Blue Bus example:**
  - Commuters face a decision between **car** and *red* **bus**.
  - Suppose that a commuter chooses between these two options with equal probability,0.5, so that the odds ratio equals 1.
  - Now suppose a third mode, *blue* **bus**, is added. Assuming **bus** commuters do not care about the color of the **bus**, they are expected to choose between **bus** and **car** still with equal probability, so the probability of **car** is still 0.5, while the probability of each of the two **bus** types is 0.25.
  - **IIA** implies that this is not the case: for the odds ratio between **car** and *red* **bus** to be preserved, and the odds of *red* and *blue* **bus** to be equal The new probabilities must be **car** 0.33; *red* **bus** 0.33; *blue* **bus** 0.33.
- Another way to characterize the problem: In

$$U_{ij} = \mathbf{X}'_{ij}\boldsymbol{\beta} + a_{ij},\ j = 0, ..., J,$$

the $a_{ij}, j = 0, 1, ..., J$, are assumed to be independent. This is unrealistic when some choices are similar.

- The *IIA* property is driven partly by the specific form of the type I extreme value distribution, but more importantly by the independence of the a_{ij} across j. (Independence across i is a given with random sampling.)
- There are a number of ways to relax *IIA*. All effectively relax the independence of the errors but in different ways
- We consider here two: the *Multinomial Probit*. and *Nested Logit*.

Multinomial Probit.

- Directly allow correlation among the $\{a_{ij} : j = 0, 1, ..., J\}$.
- Usually done by specifying multivariate normal. That is, assume $\mathbf{a}_i = (a_{i1}, ..., a_{iJ})$ has a multivariate normal distribution (with unit variances) and an unrestricted correlation matrix Σ. Leads to the **multinomial probit** model. (A better name is **conditional probit**, in the spirit of the probabilistic choice framework.)
- Multinomial probit is computationally very difficult for even a handful of alternatives.

- To see this note that

$$\begin{aligned}\mathcal{P}(Y_i &= j|\mathbf{X}_i) = \mathcal{P}\left(U_{ij} > U_{i\ell}; \ell = 0, ..., J; \ell \neq j\right) \\ &= \mathcal{P}(\mathbf{X}_{ij}'\boldsymbol{\beta} + a_{ij} > \mathbf{X}_{i\ell}'\boldsymbol{\beta} + a_{i\ell}; \ell = 0, ..., J; \ell \neq j) \\ &= \mathcal{P}(a_{ij} - a_{i\ell} > \left(\mathbf{X}_{i\ell} - \mathbf{X}_{ij}\right)' \boldsymbol{\beta}; \ell = 0, ..., J; \ell \neq j) \\ &= \mathcal{P}(\varepsilon_{i,j,\ell} > \mathbf{Z}_{i,j\,\ell}'\boldsymbol{\beta}; \ell = 0, ..., J; \ell \neq j)\end{aligned}$$

where $\varepsilon_{i,j,\ell} = a_{ij} - a_{i\ell}$, and $\mathbf{Z}_{i,j\,\ell} = \left(\mathbf{X}_{i\ell} - \mathbf{X}_{ij}\right)$. Write $\boldsymbol{\varepsilon}_{i,j} = \left(\varepsilon_{i,j,1}, \varepsilon_{i,j,2}, ..., \varepsilon_{i,j,j-1}, \varepsilon_{i,j,j+1}, ..., \varepsilon_{i,j,J}\right)'$ and consider the subset of $\mathbb{R}^J : \Gamma_{i,j}(\beta) = \prod_{\ell=0,\ell\neq j}^{J}\left(\mathbf{Z}_{i,j,\ell}'\boldsymbol{\beta}, +\infty\right)$ (Cartesian product).

- Therefore we need to compute the multiple integral:

$$P(\varepsilon_{i,j,\ell} > \mathbf{Z}'_{i,j\,\ell}\boldsymbol{\beta}; \ell = 0, ..., J; \ell \neq j) = \int_{\Gamma_{i,j}(\beta)} f(\varepsilon_{i,j}) d\boldsymbol{\varepsilon}_{i,j},$$

  where $f(\varepsilon_{i,j})$ is the density function of $\boldsymbol{\varepsilon}_{i,j}$.
- We need to resort to numerical integration or simulation methods to compute this integral.
- If we only ever observe a single choice for each unit, it is difficult to estimate the matrix Σ when the choice set is large.
- This can be partly overcome by assuming a special structure of the correlation matrix Σ.

- McFadden (1981) proposed the *Nested Logit Model*.
- Suppose we can group alternatives into S groups of "similar" alternatives. Let there be G_s alternatives in subgroup s, $s = 1, ..., S$. Now specify a nested structure:

$$\mathcal{P}(Y \in G_s|\mathbf{X}) = \frac{\left\{\alpha_s \left[\sum_{j \in G_s} \exp(\rho_s^{-1}\mathbf{X}_j'\boldsymbol{\beta})\right]^{\rho_s}\right\}}{\sum_{r=1}^{S} \alpha_r \left[\sum_{j \in G_r} \exp(\rho_r^{-1}\mathbf{X}_j'\boldsymbol{\beta})\right]^{\rho_r}}$$

$$\mathcal{P}(Y = j|Y \in G_s, \mathbf{X}) = \frac{\exp(\rho_s^{-1}\mathbf{X}_j'\boldsymbol{\beta})}{\left[\sum_{h \in G_s} \exp(\rho_s^{-1}\mathbf{X}_h'\boldsymbol{\beta})\right]}$$

- Notice that $\mathcal{P}(Y = j|\mathbf{X}) = \mathcal{P}(Y = j|Y \in G_s, \mathbf{X})\mathcal{P}(Y \in G_s|\mathbf{X})$
- The second probability is a CL model conditional on being in subgroup s.
- The first probability gives the probability that the outcome is in group s (conditional on $\mathbf{X}$);
- Need a normalization, usually $\alpha_1 = 1$.

- *Important Issue:* How can the nesting structure be chosen? Gets even more complicated with more than one level of nesting.
- Structure leads to a simple two-step estimation method. Let $\boldsymbol{\lambda}_s = \rho_s^{-1}\boldsymbol{\beta}$, $s = 1, ..., S$. These can be easily estimated by applying *conditional logit* within each subgroup s. Let $\widehat{\lambda}_s$ be the estimator of $\boldsymbol{\lambda}_s$.
- Then estimate the α_s and ρ_s by maximizing

$$\sum_{i=1}^{n}\sum_{s=1}^{S} 1[Y_i \in G_s] \log[q_s(\mathbf{X}_i; \widehat{\lambda}_s, \boldsymbol{\alpha}, \boldsymbol{\rho})],$$

where $q_s(\mathbf{X}; \boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\rho})$ is $\mathcal{P}(Y \in G_s|\mathbf{X})$.