14.385 Nonlinear Econometrics: Second part

Instructor: Alberto Abadie, abadie@mit.edu

Class meetings: Mon. and Wed. 1:05 PM - 2:25 PM

Office hours: By appointment

Contents:

- Discrete choice models
- Nonparametric methods
- Treatment effects estimators

Discrete Choice Models I

MIT
Department of Economics
14.385

Alberto Abadie

Binary Choice (Probit and Logit)

Consider an individual deciding whether to enter the labor force or to stay inactive. The utility levels under both alternatives are represented as latent variables which depend on observed and unobserved characteristics:

$$\begin{aligned} u_L &= X'\theta_L + v_L \quad \text{if in the labor force } (Y = 1), \\ u_I &= X'\theta_I + v_I \quad \text{if inactive } (Y = 0). \end{aligned}$$

This individual decides to enter the labor force if $u_L - u_I \geq 0$.

$$\begin{aligned} u_L - u_I &= (X'\theta_L + v_L) - (X'\theta_I + v_I) \\ &= X'(\theta_L - \theta_I) + (v_L - v_I) \end{aligned}$$

Therefore, $u_L - u_I \geq 0$, is equivalent to $v \leq X'\theta_0$, where $v = v_I - v_L$, and $\theta_0 = \theta_L - \theta_I$. If v_L and v_I are independent of X:

$$\begin{aligned} \Pr(Y = 1|X = x) &= \Pr(v \leq X'\theta_0|X = x) = \Pr(v \leq x'\theta_0|X = x) \\ &= \Pr(v \leq x'\theta_0) = F_v(x'\theta_0). \end{aligned}$$

Binary Choice (Probit and Logit)

If v_L and v_I have a joint normal distribution, then $v \sim N(0, \sigma_0^2)$, and $v/\sigma_0 \sim N(0,1)$. Then

$$\begin{aligned}\Pr(Y = 1|X = x) &= \Pr(v \leq x'\theta_0) = \Pr(v/\sigma_0 \leq x'\theta_0/\sigma_0) \\ &= \Phi(x'\beta_0),\end{aligned}$$

where $\beta_0 = \theta_0/\sigma_0$ and

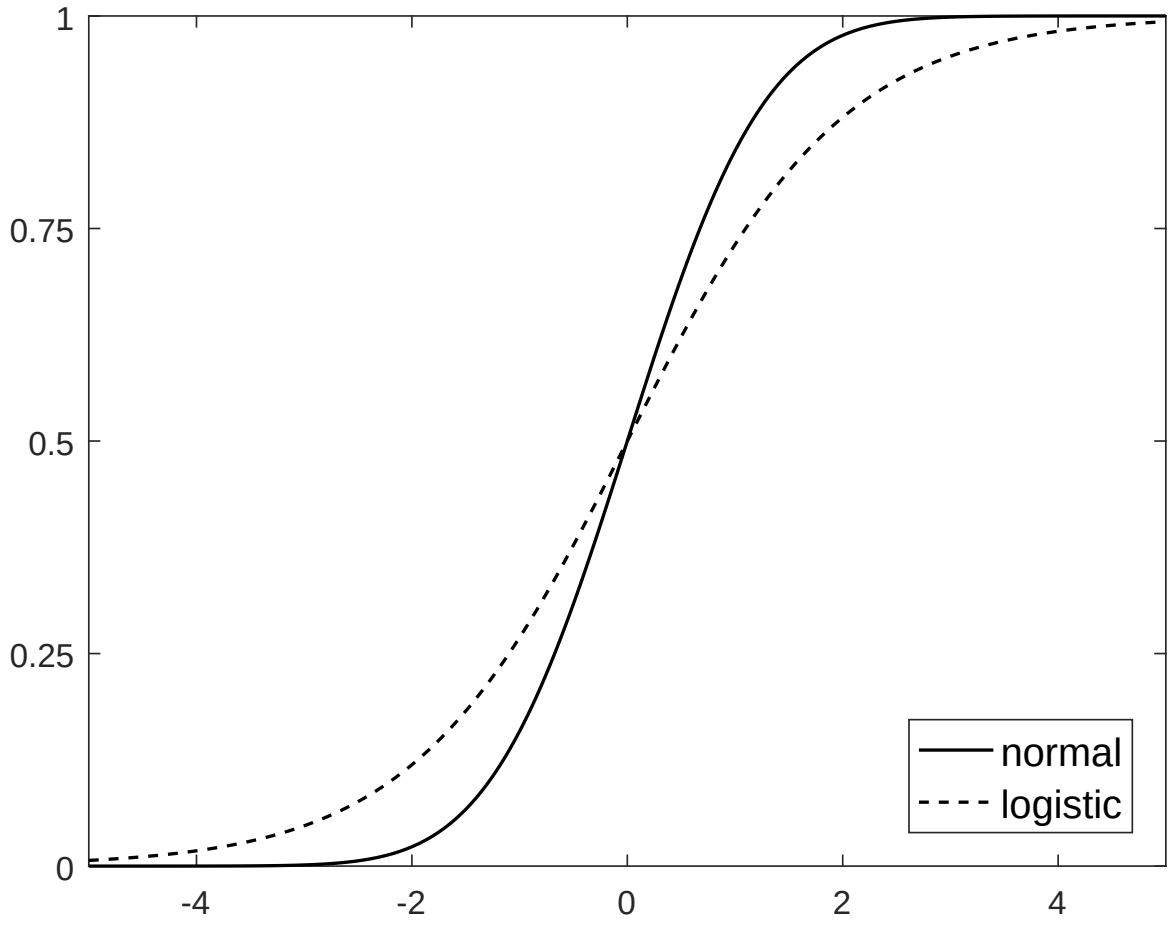
$$\Phi(r) = \int_{-\infty}^{r} \phi(t)\, dt, \qquad \text{where} \quad \phi(t) = \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-t^2/2}.$$

If instead v_L and v_I have independent type I extreme value distributions, then it can be shown that

$$\Pr(Y = 1|X = x) = \Lambda(x'\beta_0) = \frac{\exp(x'\beta_0)}{1 + \exp(x'\beta_0)}.$$

(That is, $\Lambda(\cdot)$ is the cdf of a logistic distribution.)

Binary Choice (Probit and Logit)



Normal and logistic CDFs

Binary Choice (Probit and Logit)

Let $F(x'\beta_0) = \Pr(Y = 1|X = x)$, so $F = \Phi$ for Probit and $F = \Lambda$ for Logit.

For an i.i.d. sample $\{(Y_1, X_1), \ldots, (Y_N, X_N)\}$, the MLE estimator $\widehat{\beta}$ maximizes the log-likelihood

$$\sum_{i=1}^{N} \left(Y_i \ln F(X_i'\beta) + (1 - Y_i) \ln(1 - F(X_i'\beta)) \right).$$

Therefore,

$$\sqrt{N}(\widehat{\beta} - \beta_0) \xrightarrow{d} N(0, J^{-1}),$$

where

$$J = E\left[\frac{f^2(X'\beta_0)}{F(X'\beta_0)(1 - F(X'\beta_0))} XX' \right],$$

and f is the derivative of F.

Computation of Extremum Estimators

Consider extremum estimators:

$$\widehat{\beta} = \text{argmax}_{\beta \in B} \widehat{Q}_N(\beta).$$

In particular, for Probit:

$$\widehat{Q}_N(\beta) = \sum_{i=1}^{N} \Big(Y_i \ln \Phi(X_i'\beta) + (1 - Y_i) \ln(1 - \Phi(X_i'\beta)) \Big).$$

The FOC are:

$$\frac{\partial \widehat{Q}_N(\widehat{\beta})}{\partial \beta} = \sum_{i=1}^{N} X_i \frac{Y_i - \Phi(X'\widehat{\beta})}{\Phi(X'\widehat{\beta})(1 - \Phi(X'\widehat{\beta}))} \phi(X'\widehat{\beta}) = 0,$$

where ϕ is the derivative of Φ. There is no closed-form solution for this problem so the estimators have to be calculated by numerically maximizing the objective function with respect to β.

Newton-Raphson

If $\widehat{Q}_N(\theta)$ was a quadratic function, we would have that for any $\bar{\theta} \in \Theta$:

$$\widehat{Q}_N(\theta) = \widehat{Q}_N(\bar{\theta}) + (\theta - \bar{\theta})' \frac{\partial \widehat{Q}_N(\bar{\theta})}{\partial \theta} + \frac{1}{2} (\theta - \bar{\theta})' \frac{\partial^2 \widehat{Q}_N(\bar{\theta})}{\partial \theta \partial \theta'} (\theta - \bar{\theta}).$$

The FOC to maximize this function are:

$$0 = \frac{\partial \widehat{Q}_N(\widehat{\theta})}{\partial \theta} = \frac{\partial \widehat{Q}_N(\bar{\theta})}{\partial \theta} + \frac{\partial^2 \widehat{Q}_N(\bar{\theta})}{\partial \theta \partial \theta'} (\widehat{\theta} - \bar{\theta}),$$

so

$$\widehat{\theta} = \bar{\theta} - \left(\frac{\partial^2 \widehat{Q}_N(\bar{\theta})}{\partial \theta \partial \theta'} \right)^{-1} \frac{\partial \widehat{Q}_N(\bar{\theta})}{\partial \theta}.$$

Newton-Raphson

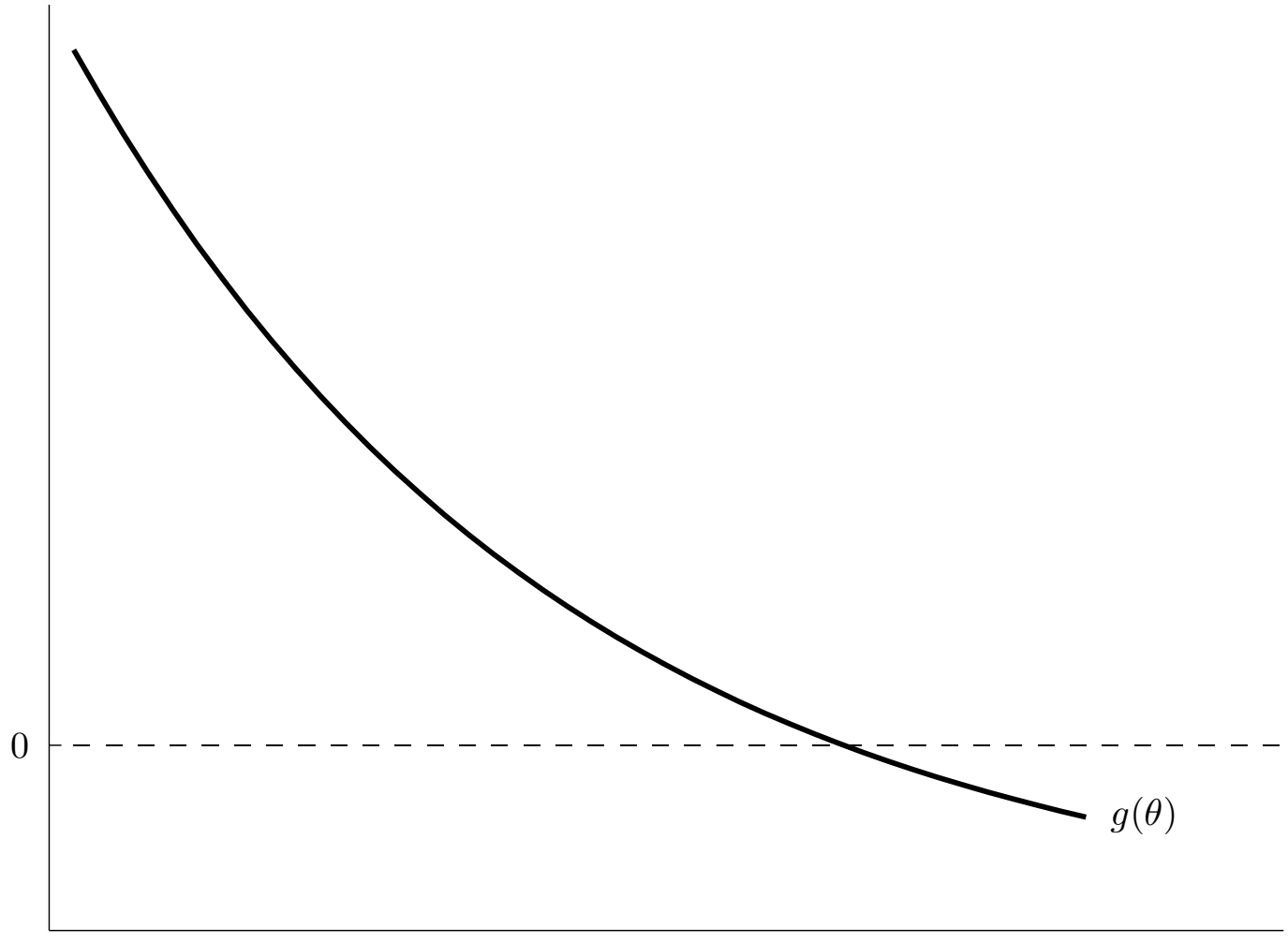
The Newton-Raphson algorithm solves quadratic approximations to the objective function, $\widehat{Q}_N(\theta)$. Let

$$g(\theta) = \frac{\partial \widehat{Q}_N(\theta)}{\partial \theta} \quad \text{and} \quad H(\theta) = \frac{\partial^2 \widehat{Q}_N(\theta)}{\partial \theta \partial \theta'}$$

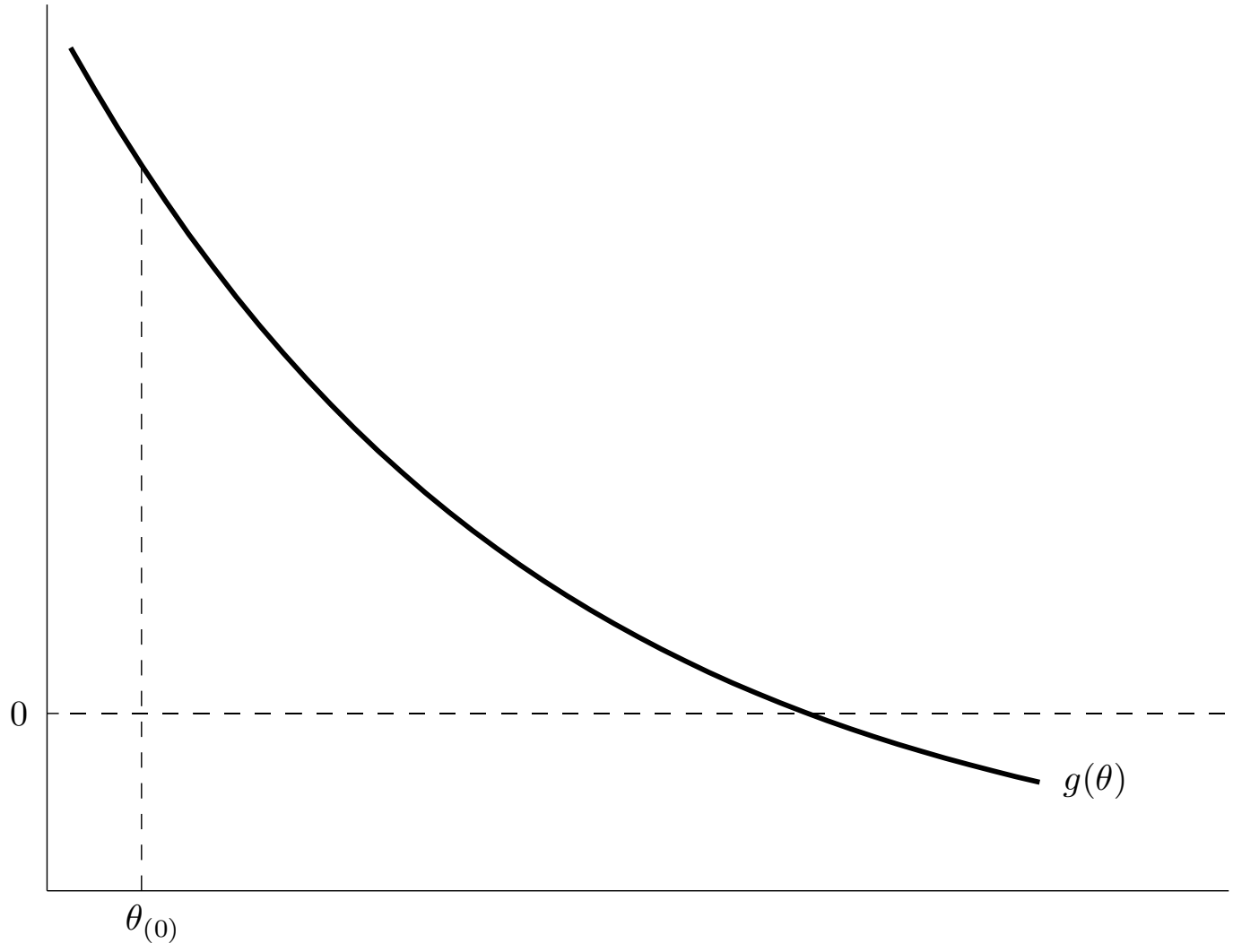
be the **gradient** and the **Hessian**. The Newton-Raphson algorithm proceeds as follows:

1. Assign an initial value to $\widehat{\theta}_{(0)}$ and define the degree of tolerance, dg (e.g., $dg = 0.000001$).
2. Iterate $\widehat{\theta}_{(k+1)} = \widehat{\theta}_{(k)} - H(\widehat{\theta}_{(k)})^{-1} g(\widehat{\theta}_{(k)})$.
3. Stop when:
   - $\|\widehat{\theta}_{(k+1)} - \widehat{\theta}_{(k)}\| < dg$ or
   - $\|g(\widehat{\theta}_{(k)})\| < dg$ or
   - $|\widehat{Q}_N(\theta_{(k+1)}) - \widehat{Q}_N(\theta_{(k)})| < dg$ or, better yet, combine
   - $\|\widehat{\theta}_{(k+1)} - \widehat{\theta}_{(k)}\| + \|g(\widehat{\theta}_{(k)})\| + |\widehat{Q}_N(\theta_{(k+1)}) - \widehat{Q}_N(\theta_{(k)})| < dg$.

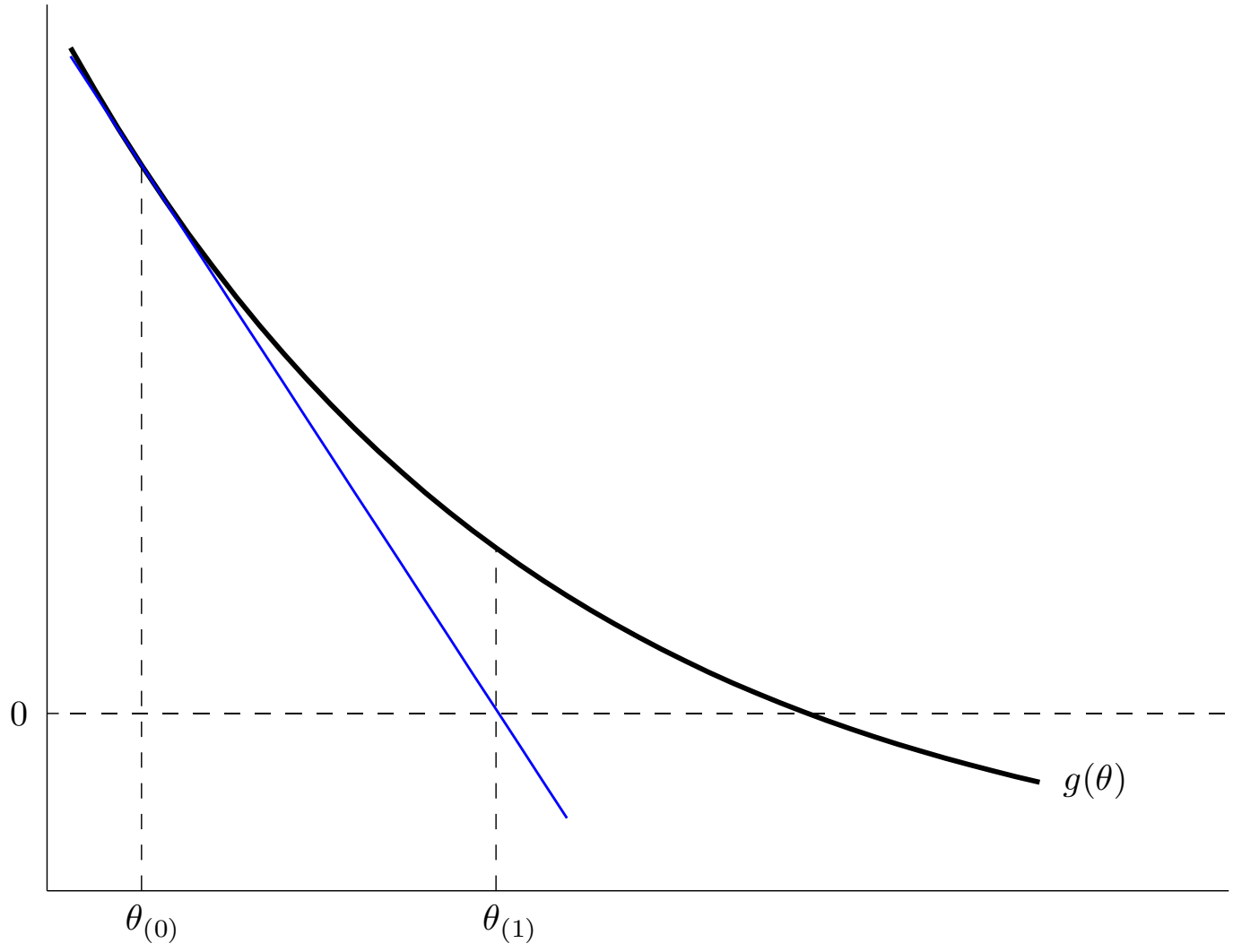
Newton-Raphson



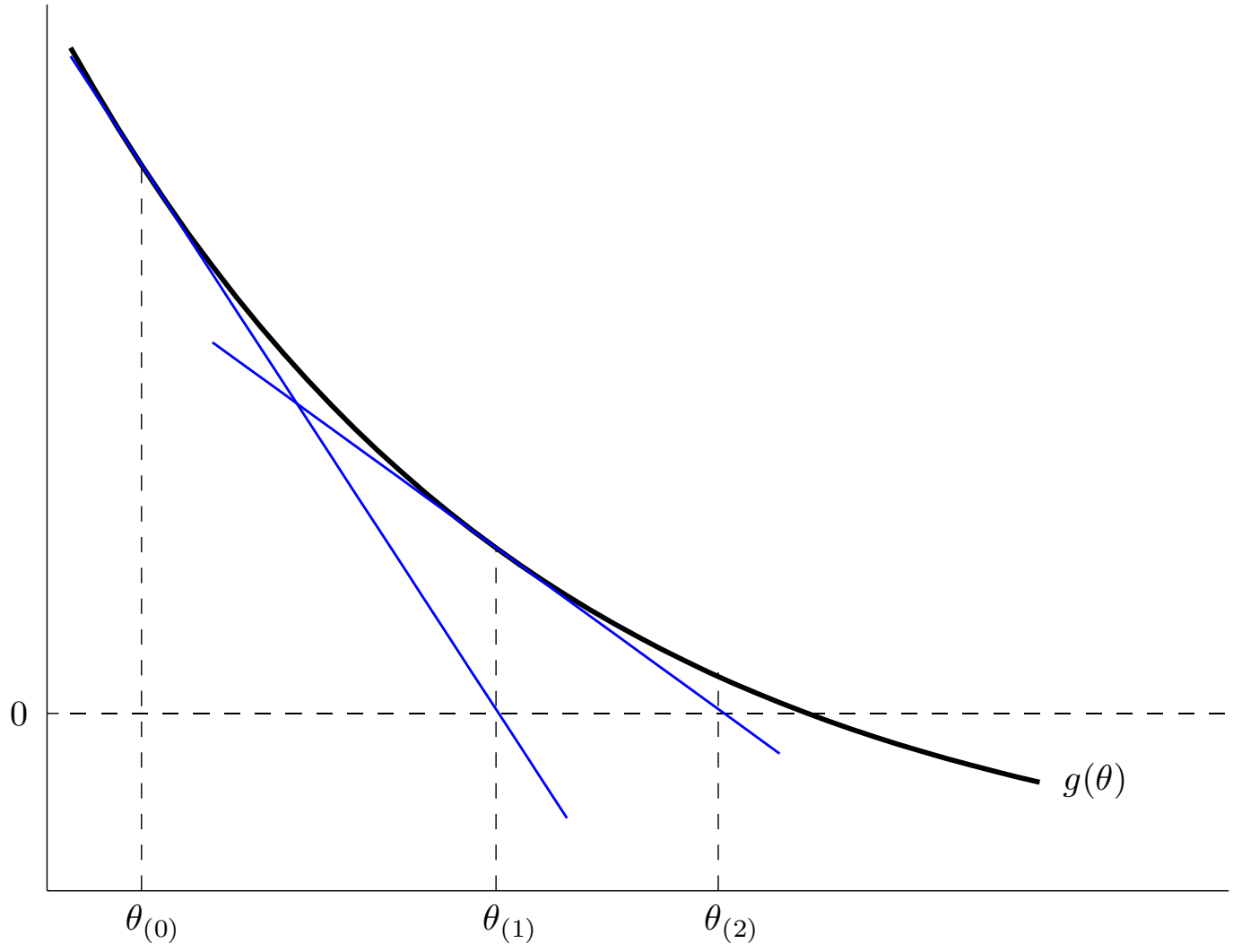
Newton-Raphson



Newton-Raphson



Newton-Raphson



Newton-Raphson

Note that to apply the Newton-Raphson algorithm, we need the objective function to be twice differentiable, and we also need to calculate first and second derivatives.

If $\widehat{Q}_N(\theta)$ is not globally concave, the Newton-Raphson algorithm may behave erratically because $-H^{-1}(\widehat{\theta}_{(k)})$ becomes negative definite so the algorithm

$$\widehat{\theta}_{(k+1)} = \widehat{\theta}_{(k)} - H(\widehat{\theta}_{(k)})^{-1} g(\widehat{\theta}_{(k)})$$

moves in a direction "opposite" to the one indicated by the gradient.

Alternative gradient-based methods are based on

$$\widehat{\theta}_{(k+1)} = \widehat{\theta}_{(k)} + A(\widehat{\theta}_{(k)})\, g(\widehat{\theta}_{(k)})$$

where $A(\widehat{\theta}_{(k)})$, which replaces $-H^{-1}(\widehat{\theta}_{(k)})$, is positive definite.

BHHH

Berndt, Hall, Hall, and Hausman (1974) proposed an alternative to Newton-Raphson for cases when computation of second derivatives is complicated.

We will study the MLE case. For an MLE estimator:

$$\widehat{Q}_N(\theta) = \sum_{i=1}^{N} \ln f_i(\theta),$$

where f_i is the likelihood of observation i. The gradient is:

$$g(\theta) = \sum_{i=1}^{N} \frac{\partial \ln f_i(\theta)}{\partial \theta}.$$

The BHHH algorithm proceeds exactly like the Newton-Raphson algorithm but replaces the Hessian matrix with:

$$H_{\text{BHHH}}(\theta) = -\sum_{i=1}^{N} \frac{\partial \ln f_i(\theta)}{\partial \theta} \frac{\partial \ln f_i(\theta)}{\partial \theta'}.$$

BHHH

$H_{\text{BHHH}}(\theta)$ is easier to calculate than the Hessian because it depends on first derivatives only. In addition, $-H_{\text{BHHH}}^{-1}(\theta)$ is always positive definite.

The BHHH algorithm can also be adapted to other (non-MLE) M-estimators. In that case:

$$\widehat{Q}_N(\theta) = \sum_{i=1}^{N} m_i(\theta),$$

$$g(\theta) = \sum_{i=1}^{N} \frac{\partial m_i(\theta)}{\partial \theta},$$

and

$$H_{\text{BHHH}}(\theta) = -\sum_{i=1}^{N} \frac{\partial m_i(\theta)}{\partial \theta} \frac{\partial m_i(\theta)}{\partial \theta'}.$$

Other Methods

There exist non-gradient based methods that can be used when computing derivatives is costly or when the objective function is not differentiable. These are typically slower than gradient-based methods but do not require computing derivatives.

One of the most popular among the non-gradient-based methods is the **Nelder-Mead** algorithm (which is the default optimizer in Matlab).

Optimization is complicated in the presence of local optima. In that case, we should:

- restart our algorithms from different starting values, or
- use randomized optimization algorithms, like **simulated annealing** or **genetic algorithms**.

Binary Choice (Probit and Logit)

Partial derivatives, also called **marginal effects**, are:

$$\frac{\partial}{\partial x}\Pr(Y=1|X=x) = f(x'\beta_0)\beta_0.$$

In particular, the partial derivative with respect to the j-th covariate (with coefficient β_0^j) is

$$f(x'\beta_0)\beta_0^j.$$

Because f is positive, partial derivatives have the same sign as the coefficients in β_0. Notice that ratios between partial derivatives for different covariates equal ratios of coefficients:

$$\frac{f(x'\beta_0)\beta_0^j}{f(x'\beta_0)\beta_0^k} = \frac{\beta_0^j}{\beta_0^k}.$$

We can estimate marginal effects evaluated at particular values, x, of the regressors with

$$f(x'\widehat{\beta})\widehat{\beta}.$$

Standard errors can be calculated using the delta method or the bootstrap.

Binary Choice (Probit and Logit)

When the regressors of interest are discrete or categorical, it often makes more sense to directly calculate the change in the probability that $Y = 1$ associated with a change in the value of one of the regressors.

Suppose that the dependent variable, Y, refers to labor market participation of married women and that the regressor of interest, X^j (with coefficient β_0^j) is the number of children.

Suppose also that the model includes an additional set of regressors X^{-j} with coefficients β_0^{-j}.

For a given value $X^{-j} = x^{-j}$, the change in the probability of participation associated with the number of children changing from x_1^j to x_2^j is

$$F(x_2^j \beta_0^j + x^{-j\prime} \beta_0^{-j}) - F(x_1^j \beta_0^j + x^{-j\prime} \beta_0^{-j}).$$

Binary Choice (Probit and Logit)

Notice that when $x_2^j - x_1^j$ is positive, because F is an increasing function, it follows that the change in probability

$$F(x_2^j \beta_0^j + x^{-j\prime} \beta_0^{-j}) - F(x_1^j \beta_0^j + x^{-j\prime} \beta_0^{-j}) \tag{1}$$

has the same sign as β_0^j.

The probability change in equation (1) can be estimated by

$$F(x_2^j \widehat{\beta}^j + x^{-j\prime} \widehat{\beta}^{-j}) - F(x_1^j \widehat{\beta}^j + x^{-j\prime} \widehat{\beta}^{-j})$$

and standard errors can be calculated using the delta method or the bootstrap.

Binary Choice (Probit and Logit)

Sometimes instead of evaluating the partial derivatives at some arbitrary value x, we may be interested in the average value of the partial derivatives in the population. The vector of average partial derivatives is:

$$E\left[\frac{\partial}{\partial X}\Pr(Y=1|X)\right] = E\left[f(X'\beta_0)\right]\beta_0,$$

which can be estimated using:

$$\left(\frac{1}{N}\sum_{i=1}^{N} f(X_i'\widehat{\beta})\right)\widehat{\beta}.$$

Standard errors for average partial derivatives can be derived using techniques for two-step estimators or the bootstrap.

Binary Choice (Probit and Logit)

Similarly, we may want to evaluate the population average of a probability change associated with a change in the value of regressor X^j from x_1^j to x_2^j:

$$E\Big[F(x_2^j\beta_0^j + X^{-j\prime}\beta_0^{-j}) - F(x_1^j\beta_0^j + X^{-j\prime}\beta_0^{-j})\Big].$$

This can be estimated using the sample analog

$$\frac{1}{N}\sum_{i=1}^{N}\left(F(x_2^j\widehat{\beta}^j + X_i^{-j\prime}\widehat{\beta}^{-j}) - F(x_1^j\widehat{\beta}^j + X_i^{-j\prime}\widehat{\beta}^{-j})\right),$$

and standard errors can be derived using techniques for two-step estimators or the bootstrap.

Other estimands are possible, of course, as demonstrated later (see the HMDA example below).

Binary Choice (Probit and Logit)

The logit coefficients have also an direct interpretation in terms of odds ratios. Notice that for the logit model:

$$\frac{\Pr(Y_i = 1|X_i)}{\Pr(Y_i = 0|X_i)} = e^{X_i'\beta_0}.$$

The expression $\Pr(Y_i = 1|X_i)/\Pr(Y_i = 0|X_i)$ is called the **odds**.

After a one-unit increase in X^j, the odds become $e^{X_i'\beta_0+\beta_0^j}$. Therefore, the **odds ratio** associated with a one-unit increase in X^j is:

$$\frac{e^{X_i'\beta_0+\beta_0^j}}{e^{X_i'\beta_0}} = e^{\beta_0^j}.$$

That is, a one-unit increase in X^j is associated with an increase in the odds of $100 \times (e^{\beta_0^j} - 1)$ percent. A equivalent interpretation is given by the fact that β_0^j is change in the log of the odds ratio associated with a one unit increase in X^j.

Binary Choice (Probit and Logit)

There is some confusion around the notion of the odds ratio, as in the social sciences the odds, $\Pr(Y_i = 1|X_i)/\Pr(Y_i = 0|X_i)$, are often referred to as the "odds ratio".

In addition, $\Pr(Y_i = 1|X_i)/\Pr(Y_i = 0|X_i)$ is sometimes referred to as "relative risk". However, the correct definition of **relative risk** is the ratio in the probabilities of the same event for two population groups:

$$\frac{\Pr(Y = 1|X = x_1)}{\Pr(Y = 1|X = x_0)}.$$

For example, it has been estimated that the probability of developing lung cancer during the lifetime is around 0.16 for smokers and 0.01 for non-smokers. Therefore, the relative risk is 16: smokers are 16 times more likely to develop lung cancer than non-smokers.

The odds ratio, however, is:

$$\frac{.16/(1 - .16)}{.01/(1 - .01)} = 18.86.$$

Example: The Boston HMDA Data

In order to detect potential discriminatory practices of mortgage credit lenders against minority applicants, the U.S. Home Mortgage Disclosure Act (HMDA, pronounced "hum-duh") of 1975 requires lenders to routinely disclose information on mortgage applications, including the race and ethnicity of the applicants.

The information collected under the HMDA does not include, however, data on the credit histories of the applicants, and other loan and applicant characteristics that are considered to be important factors in determining the approval or denial of mortgage loans.

To overcome these deficiencies in the HMDA data, the Federal Reserve Bank of Boston collected an additional set of 38 variables included in mortgage applications for a sample of applications in the Boston metropolitan area in 1990.

Example: The Boston HMDA Data

The Boston HMDA data set includes all mortgage applications by minority applicants in the Boston metropolitan area in 1990, as well as a random sample of mortgage applications by non-minority applicants in the same year and geographical area. We will treat them as separate random samples from some super-populations of interest.

The outcome variable, Y, takes value one if the mortgage application was denied, and zero if the mortgage application was approved, and X is a vector of applicant and loan characteristics:

- race,
- housing expense to income ratio,
- total debt payments to income ratio,
- consumer credit history,
- mortgage credit history,
- regional unemployment rate in the applicant's industry,
- loan amount to appraised value ratio.

(see Munnell et al., *AER* 1996, for precise variable definitions).

Example: The Boston HMDA Data

We restrict our sample to single-family residences and male applicants who are white non-Hispanic or black non-Hispanic, not self-employed, who were approved for private mortgage insurance, and who do not have a public record of default or bankruptcy at the time of the application.

This leaves us with a sample of 148 black applicants and 1336 white applicants, for a total of 1484 applicants.

In this sample, the mortgage application denial rates are 6.36 percent for non-minority applicants and 20.27 percent for minority applicants, with a difference of 13.91 percentage points.

We will investigate how much of this difference can be attributed to differences in observable characteristics (other than race) between black and whites.

Example: The Boston HMDA Data

Let

$$X^j = \begin{cases} 1 & \text{for minority applicants,} \\ 0 & \text{for non-minority applicants.} \end{cases}$$

We aim to estimate:

$$\tau_0 = E\left[F(\beta_0^j + X^{-j\prime}\beta_0^{-j}) - F(X^{-j\prime}\beta_0^{-j})\middle| X^j = 1\right].$$

τ_0 represents the difference between the average rejection rate for minority applicants and the average rejection rate for non-minority applicants after setting the distribution of X^{-j} to be the same as for minority applicants.

Let $N_1 = \sum_{i=1}^N X_i^j$ be the number of minority applicants in our sample. Our estimator of τ_0 is:

$$\widehat{\tau} = \frac{1}{N_1}\sum_{i=1}^N X_i^j \left(F(\widehat{\beta}^j + X_i^{-j\prime}\widehat{\beta}^{-j}) - F(X_i^{-j\prime}\widehat{\beta}^{-j})\right).$$

Example: The Boston HMDA Data

Calculating standard errors for $\widehat{\tau}$ is complicated because:

- sampling is not completely i.i.d., because minority and non-minority applicants are sampled separately,
- $\widehat{\tau}$ averages only over minority applicants.

To overcome this difficulties we use the bootstrap, adjusting the bootstrap sampling scheme to mimic the sampling scheme used to generate the original sample.

That is, in each bootstrap iteration we resample 148 black applicants and 1336 white applicants with replacement.

Example: The Boston HMDA Data

The following table reports estimates the Logit coefficient on the minority dummy variable, $\widehat{\beta}^j$, and the difference between the probability of mortgage denial for minority applicant and non-minority applicants after adjusting for differences in applicant and loan characteristics, along with standard errors.

	$\widehat{\beta}^j$	$\widehat{\tau}$
estimate	0.8829	0.0990
s.e.	0.2632	0.0333

The Logit coefficient indicates that, after taking into account other variables, mortgage applications from minority applicants are still rejected with higher probability that mortgage applications from non-minority applicants.

$\widehat{\tau}$ indicates that the difference in denial rates between minority applicants and a comparable group of non-minority applicants is 9.9 percentage points.

Ordered Probit

Sometimes the dependent variable of interest is categorical and ordered, but takes on more than two values:

- Number of cars in a household: $0, 1, 2, ...$
- Highest educational degree attained: $0 =$ high school dropout, $1 =$ high school, $2 =$ college, etc.
- Vote in an election: $0 =$ left-wing party, $1 =$ center-left, $2 =$ center-right, etc.
- Capital controls on FDI: $0 =$ no restrictions, $1 =$ moderate restrictions (restrictions only in "strategic" industries), $2 =$ severe restrictions (e.g. restrictions to repatriation of capitals).

Ordered Probit

Suppose that the variable Y^* represents the latent preference for the number of cars in a household (0, 1, or 2).

Assume that the number of cars in a household is given by:

$$Y = \begin{cases} 0 & \text{if } Y^* \leq c_1, \\ 1 & \text{if } c_1 < Y^* \leq c_2, \\ 2 & \text{if } c_2 < Y^*, \end{cases}$$

where $Y = 2$ codes "two or more cars".

Suppose also that:

$$Y^* = X'\beta + u, \qquad \text{where} \quad u \sim N(0, 1).$$

Ordered Probit

Then,

$$\begin{aligned}
\Pr(Y = 0 \mid X) &= \Pr(Y^* \leq c_1 | X) = \Pr(X'\beta + u \leq c_1 | X) \\
&= \Pr(u \leq c_1 - X'\beta | X) = \Phi(c_1 - X'\beta),
\end{aligned}$$

$$\begin{aligned}
\Pr(Y = 1 \mid X) &= \Pr(c_1 < Y^* \leq c_2 | X) = \Pr(c_1 < X'\beta + u \leq c_2 | X) \\
&= \Pr(c_1 - X'\beta < u \leq c_2 - X'\beta | X) \\
&= \Phi(c_2 - X'\beta) - \Phi(c_1 - X'\beta),
\end{aligned}$$

$$\begin{aligned}
\Pr(Y = 2 \mid X) &= \Pr(c_2 < Y^* | X) = \Pr(c_2 < X'\beta + u | X) \\
&= \Pr(c_2 - X'\beta < u | X) = 1 - \Phi(c_2 - X'\beta).
\end{aligned}$$

Ordered Probit

Now, suppose that we have a sample $(Y_1, X_1), \cdots, (Y_N, X_N)$. The likelihood of the sample is:

$$\left(\prod_{Y_i=0} \Phi(c_1 - X_i'\beta) \right) \left(\prod_{Y_i=1} \left(\Phi(c_2 - X_i'\beta) - \Phi(c_1 - X_i'\beta) \right) \right) \times \left(\prod_{Y_i=2} \left(1 - \Phi(c_2 - X_i'\beta) \right) \right).$$

The ordered probit estimator maximizes this likelihood.

An ordered logit model can be derived similarly.

Ordered Probit

For the case of $m+1$ categories, the likelihood of the sample becomes:

$$\left(\prod_{Y_i=0}\Phi(c_1-X_i'\beta)\right)\left(\prod_{Y_i=1}\Big(\Phi(c_2-X_i'\beta)-\Phi(c_1-X_i'\beta)\Big)\right)$$
$$\cdots\times\left(\prod_{Y_i=m-1}\Big(\Phi(c_m-X_i'\beta)-\Phi(c_{m-1}-X_i'\beta)\Big)\right)$$
$$\times\left(\prod_{Y_i=m}\Big(1-\Phi(c_m-X_i'\beta)\Big)\right).$$

Ordered Probit

An appealing feature of the ordered probit (as opposed to using linear regression) is that it produces predicted probabilities in between zero and one and that sum to one.

$$\begin{aligned}
\widehat{\Pr}(Y = 0 \,|\, X = x) &= \Phi(\widehat{c}_1 - x'\widehat{\beta}), \\
\widehat{\Pr}(Y = 1 \,|\, X = x) &= \Phi(\widehat{c}_2 - x'\widehat{\beta}) - \Phi(\widehat{c}_1 - x'\widehat{\beta}), \\
\widehat{\Pr}(Y = 2 \,|\, X = x) &= 1 - \Phi(\widehat{c}_2 - x'\widehat{\beta}).
\end{aligned}$$

Notice that, by construction, these predicted probabilities are in between zero and one and sum to one.

Ordered Probit

Example: Education in the CPS.

Since 1992, the Current Population Survey (CPS) reports education as "highest degree attained" (high school, college, more than college).

March 2006 CPS extract with the following variables:

`educ`:	=0 if high school dropout,
	=1 high school graduate,
	=2 if college graduate,
	=3 if higher degree,
`hispanic`:	=1 if Hispanic, =0 otherwise,
`age`:	age in years.

The individuals in our sample are heads of households, white, male, aged 35-60, and working full time for pay (not self-employed).

Ordered Probit

We want to estimate:

$$\Pr(\texttt{educ} = n \,|\, \texttt{hispanic}, \texttt{age}),$$

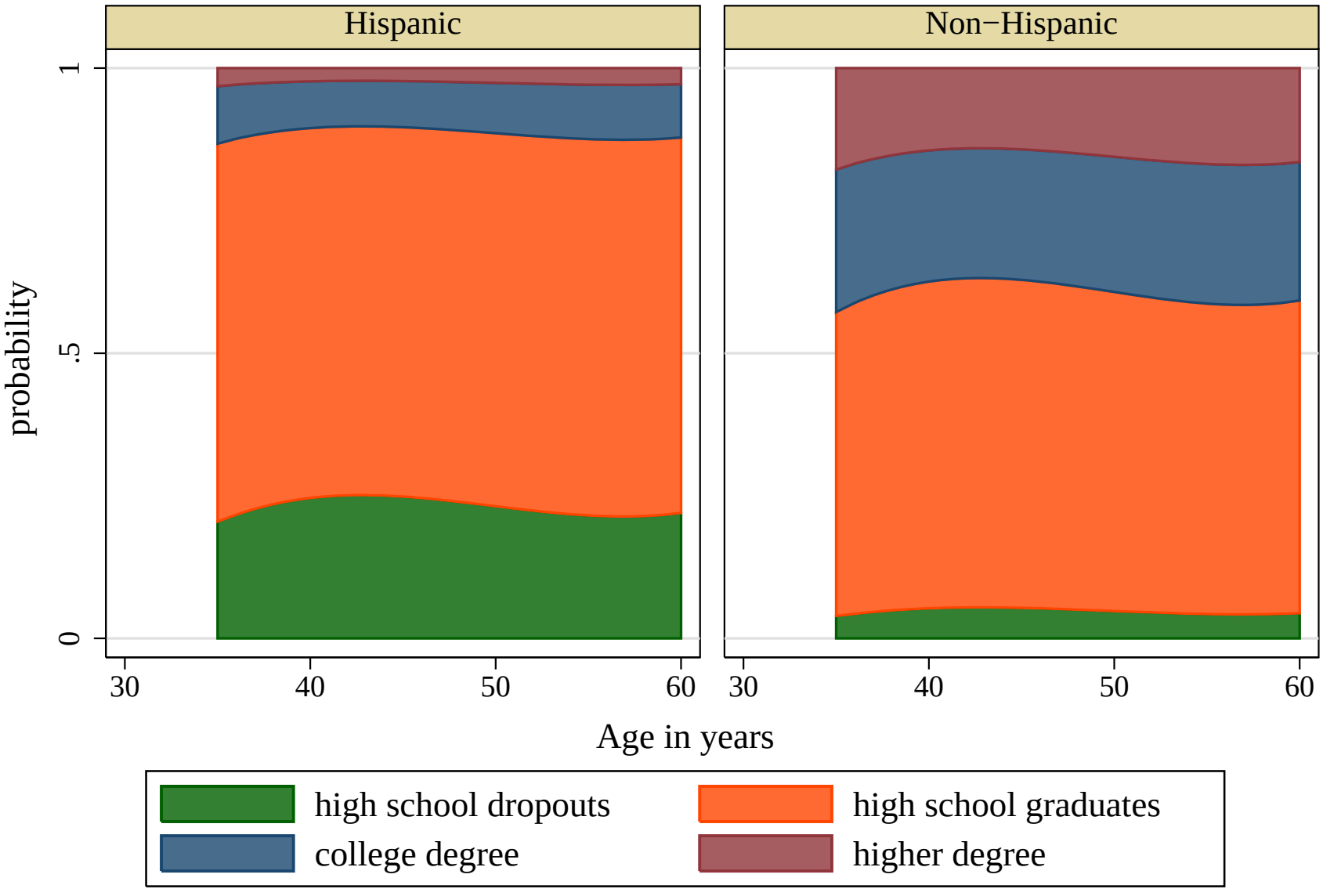
for $n = 0, 1, 2, 3$.

Ordered Probit Estimates

Dependent variable: educ

	$\widehat{\beta}^j$	s.e.$(\widehat{\beta}^j)$
hispanic	-.99704	(.03003)
age	-.50216	(.21095)
age^2	.01024	(.00454)
age^3	-.00007	(.00003)
$\widehat{c}_1$	-9.5476	(3.2280)
$\widehat{c}_2$	-7.6707	(3.2278)
$\widehat{c}_3$	-6.9042	(3.2277)

Ordered Probit



Highest degree attained as a function of ethnicity and age

Multinomial Choice Models

In many problems of interest choices are made from a set of unordered alternatives. Examples:

- Commuting mode: car, train, walk, ...
- Mobile phone network.

We will assume that for a random sample of N individuals, we observe the choice

$$Y_i \in \{0, 1, \ldots, m\}.$$

Let $Y_{ij} = 1$ if $Y_i = j$, and $Y_{ij} = 0$ otherwise, for $i = 1, \ldots, N$ and $j = 0, \ldots, m$.

We observe also characteristics of the individual,

$$X_i \quad \text{for} \quad i = 1, \ldots, N \quad \text{(e.g., income)},$$

and we may also observe covariates with values that change depending on the alternative,

$$X_{ij} \quad \text{for} \quad i = 1, \ldots, N \quad \text{and} \quad j = 0, \ldots, m \quad \text{(e.g., prices)}.$$

Multinomial Logit Model

Suppose first that the covariates of interest only vary across individuals, X_i. The **multinomial logit model** is then a straightforward extension of the basic logit model.

Let $p_{ij} = \Pr(Y_i = j|X_i)$. The multinomial logit model postulates:

$$p_{ij} = \frac{e^{X_i'\beta_j}}{\sum_{k=0}^{m} e^{X_i'\beta_k}}.$$

Equal translation of all vectors β_k to $\beta_k + \alpha$ (where α is any conformable vector) leaves all choice probabilities, p_{ij}, unchanged. Therefore, we typically normalize $\beta_0 = 0$, (and we say that $j = 0$ is the "base case"):

$$p_{ij} = \begin{cases} \frac{1}{1+\sum_{k=1}^{m} e^{X_i'\beta_k}} & \text{if } j = 0, \\ \frac{e^{X_i'\beta_j}}{1+\sum_{k=1}^{m} e^{X_i'\beta_k}} & \text{if } j > 1, \end{cases}$$

and $p_{ij}/p_{i0} = e^{X_i'\beta_j}$.

Multinomial Logit Model

The log-likelihood is:

$$\sum_{i=1}^{N}\sum_{j=0}^{m} Y_{ij} \ln p_{ij} = \sum_{i=1}^{N}\sum_{j=0}^{m} Y_{ij} \ln \left(\frac{e^{X_i' \beta_j}}{\sum_{j=0}^{m} e^{X_i' \beta_j}} \right).$$

We maximize this log-likelihood with respect to $\beta_1, \ldots, \beta_m$ (remember that we set $\beta_0 = 0$).

From calculus, the formula for the marginal effects is:

$$\frac{\partial p_{ij}}{\partial X_i} = p_{ij} \left(\beta_j - \sum_{k=0}^{m} p_{ik} \beta_k \right).$$

Notice that $\beta_j = 0$ does not imply $\partial p_{ij} / \partial X_i = 0$. The average marginal effect is:

$$\frac{1}{N} \sum_{i=1}^{N} p_{ij} \left(\beta_j - \sum_{k=0}^{m} p_{ik} \beta_k \right).$$

Conditional Logit Model

The **conditional logit model** considers the case where regressors vary by alternative (and possibly by individual), X_{ij}. This model postulates:

$$p_{ij} = \frac{e^{X_{ij}'\beta}}{\sum_{k=0}^{m} e^{X_{ik}'\beta}}.$$

The model cannot contain variables that do not vary by alternative (like a constant, or individual income). Imagine that we try to introduce such a variable, Z_i, with coefficient α. The probability of choice j for individual i is now:

$$p_{ij} = \frac{e^{X_{ij}'\beta + Z_i\alpha}}{\sum_{k=0}^{m} e^{X_{ik}'\beta + Z_i\alpha}} = \frac{e^{Z_i\alpha} e^{X_{ij}'\beta}}{e^{Z_i\alpha} \sum_{k=0}^{m} e^{X_{ik}'\beta}} = \frac{e^{X_{ij}'\beta}}{\sum_{k=0}^{m} e^{X_{ik}'\beta}}.$$

So, α is not identified.

Conditional Logit Model

The marginal effects for the conditional logit model are:

$$\partial p_{ij}/\partial X_{ij} = p_{ij}(1 - p_{ij})\beta, \quad \partial p_{ij}/\partial X_{ik} = -p_{ij}p_{ik}\beta.$$

McFadden (1974) showed that:

- The conditional logit model can be derived as the solution of a utility maximizing agent with utility from choice j given by

  $$U_{ij} = X'_{ij}\beta + u_{ij}$$

  where u_{ij} are independent and have a type I extreme value distribution.
- By choosing X_{ij} and β appropriately, it can be shown that the multinomial logit model is a particular case of the conditional logit model. To simplify, suppose that there are three alternatives $j = 0, 1, 2$, then we can make:

  $$X_{i0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad X_{i1} = \begin{pmatrix} X_i \\ 0 \end{pmatrix} \quad X_{i2} = \begin{pmatrix} 0 \\ X_i \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix},$$

  obtaining the multinomial logit model.

Conditional Logit Model

An implication of the fact the conditional logit model nests the multinomial logit model is that we can specify conditional logit models with individual specific variables (as long as they have coefficients that vary by alternative):

$$p_{ij} = \frac{e^{X_{ij}'\beta + Z_i'\alpha_j}}{\sum_{k=0}^{m} e^{X_{ik}'\beta + Z_i'\alpha_k}}.$$

Again, a normalization on the coefficients is needed to attain identification. Usually, $\alpha_0 = 0$ is adopted.

Choice of Heating System (K. Train and Y. Croissant)

Data from 900 California households on the choice of heating system.

Possible choices are:

- gas central
- gas room
- electric central
- electric room
- heat pump

We will consider two variables, X_{ij}, that vary by alternative and household:

- installation cost ($\texttt{ic}_{ij}$)
- annual operating cost ($\texttt{oc}_{ij}$)

The model will also include an intercept, α_j, for each alternative. The base case, $j = 0$, is heat pump, and we normalize $\alpha_0 = 0$.

Choice of Heating System (K. Train and Y. Croissant)

We estimate the model:

$$p_{ij} = \frac{e^{\beta^{\mathtt{ic}}\mathtt{ic}_{ij}+\beta^{\mathtt{oc}}\mathtt{oc}_{ij}+\alpha_j}}{\sum_{k=0}^{4} e^{\beta^{\mathtt{ic}}\mathtt{ic}_{ik}+\beta^{\mathtt{oc}}\mathtt{oc}_{ik}+\alpha_k}}.$$

Conditional Logit Estimates

	coeff.		
	symbol	value	s.e.
Intercepts:			
electric central	$\widehat{\alpha}_1$	1.6588	(0.4484)
electric room	$\widehat{\alpha}_2$	1.8534	(0.3620)
gas central	$\widehat{\alpha}_3$	1.7110	(0.2267)
gas room	$\widehat{\alpha}_4$	0.3083	(0.2066)
installation cost	$\widehat{\beta}^{\mathtt{ic}}$	-0.0015	(0.0006)
operating cost	$\widehat{\beta}^{\mathtt{oc}}$	-0.0070	(0.0016)

Choice of Heating System (K. Train and Y. Croissant)

Suppose now that the California Energy Commission (CES) plans to offer 10 percent rebates on the installation costs of heat pumps.

We can use the model to predict the increase in the choice probability of heat pumps in California induced by the CES rebate:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{e^{\widehat{\beta}^{\mathtt{ic}}(0.9\,\mathtt{ic}_{i0})+\widehat{\beta}^{\mathtt{oc}}\mathtt{oc}_{i0}}}{e^{\widehat{\beta}^{\mathtt{ic}}(0.9\,\mathtt{ic}_{i0})+\widehat{\beta}^{\mathtt{oc}}\mathtt{oc}_{i0}}+\sum_{k=1}^{4}e^{\widehat{\beta}^{\mathtt{ic}}\mathtt{ic}_{ik}+\widehat{\beta}^{\mathtt{oc}}\mathtt{oc}_{ik}+\widehat{\alpha}_k}}$$

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{e^{\widehat{\beta}^{\mathtt{ic}}\mathtt{ic}_{i0}+\widehat{\beta}^{\mathtt{oc}}\mathtt{oc}_{i0}}}{e^{\widehat{\beta}^{\mathtt{ic}}\mathtt{ic}_{i0}+\widehat{\beta}^{\mathtt{oc}}\mathtt{oc}_{i0}}+\sum_{k=1}^{4}e^{\widehat{\beta}^{\mathtt{ic}}\mathtt{ic}_{ik}+\widehat{\beta}^{\mathtt{oc}}\mathtt{oc}_{ik}+\widehat{\alpha}_k}}$$

$$= 0.0645 - 0.0555$$

$$= 0.0090.$$

Notice that we could have calculated the choice probability of heat pump without the rebate (0.0555) directly in the data, without using the model. If the model includes alternative-specific intercepts both calculations coincide.

Independence of Irrelevant Alternatives (IIA)

The conditional logit model implies that the conditional probability of $Y_i = j$ given $Y_i = j$ or k is:

$$\frac{p_{ij}}{p_{ij} + p_{ik}} = \frac{1}{1 + e^{-(X_{ij} - X_{ik})'\beta}}.$$

That is, this conditional probability depends only on $X_{ij} - X_{ik}$ and not on the characteristics of other alternatives. Moreover, it does not change even if other alternatives become available. This is called **independence of irrelevant alternatives** (IIA).

Another way to express IIA is that ratios of probability choices:

$$\frac{p_{ij}}{p_{ik}} = e^{(X_{ij} - X_{ik})'\beta}$$

depend only on $X_{ij} - X_{ik}$.

Independence of irrelevant alternatives is a potentially restrictive feature of the conditional logit model, as illustrated by McFadden's **red bus/blue bus example**.

Red Bus/Blue Bus

Suppose commuters choose between two modes of transportation: **car** and **red bus**. Initially the choice probability for each mode is $1/2$:

$$p_{\texttt{car}} = p_{\texttt{red bus}} = 1/2,$$

and $p_{\texttt{car}}/p_{\texttt{red bus}} = 1$.

Suppose now that a new mode of transportation is introduced: the **blue bus**. If consumers do not value the color of the bus (the color of the bus is not in X_{ij}) then blue buses and red buses have the same covariates, so $p_{\texttt{red bus}}/p_{\texttt{blue bus}} = 1$. By the independence of irrelevant alternatives property, it is still true that $p_{\texttt{car}}/p_{\texttt{red bus}} = 1$. Therefore, we obtain:

$$p_{\texttt{car}} = p_{\texttt{red bus}} = p_{\texttt{blue bus}} = 1/3.$$

Just by introducing the blue bus, the choice probability of car went from $1/2$ to $1/3$!

However, because blue and red buses are close substitutes, we would expect that $p_{\texttt{car}}$ stays unchanged.

Red Bus/Blue Bus

The conditional logit model fails to take into account that blue and red buses are close substitutes, essentially the same alternative.

"This example suggests that the application of the model should be limited to situations where the alternatives can plausibly be assumed to be distinct ... in the eyes of each decision maker." (McFadden, 1974)

The example shows that the conditional logit model will tend to overestimate the demand for new alternatives that are close substitutes of existing ones.

Lalime's/Chez Panisse/Bongo Burger

Imbens provides another example where IIA creates unrealistic substitutions patterns.

Suppose that two upscale restaurants, Lalime's and Chez Panisse, and a fast food restaurant, Bongo Burger, compose a local restaurant market with market shares 0.10, 0.25, and 0.65, respectively.

The IIA property implies that if Lalime's closes (or becomes unaffordably expensive), then Lalime's market share will be divided between Chez Panisse and Bongo Burger in a way proportional to their previous market shares.

This means that $0.65/(0.25+0.65)=72\%$ of the previous Lalime's diners substitute to Bongo Burger!

IIA and the Random Utility Model

As discussed before, the conditional logit model can be derived as the solution of an agent maximizing utility over choices, where utility from choice j is given by

$$U_{ij} = X'_{ij}\beta + u_{ij}$$

and u_{ij} are independent and have a type I extreme value distribution.

It can be shown that the IIA property arise from the combination of two assumptions:

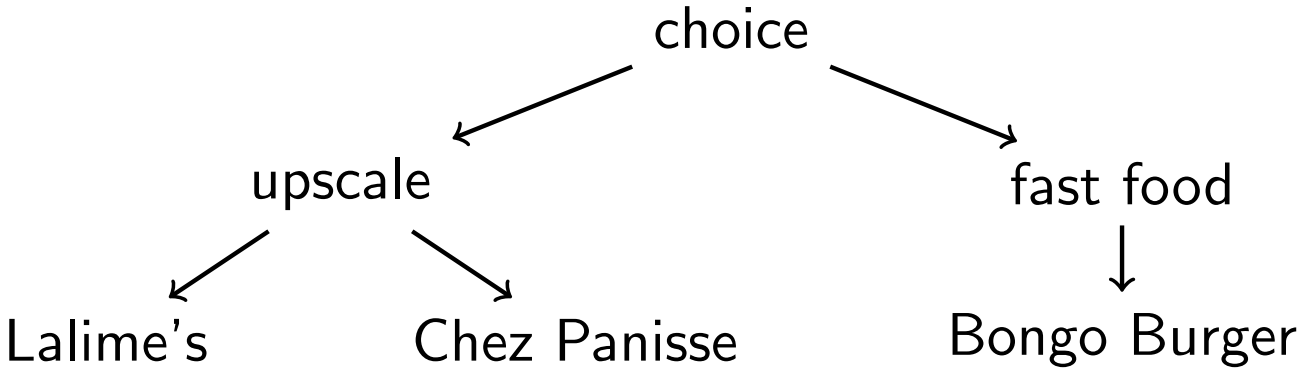
(1) independence between the utilities of different alternatives (given the covariates),
(2) type I extreme value distribution for u_{ij}.

There exist extensions of the conditional logit model that dispose of the IIA property by relaxing either (1) or (2). We will study two models that relax (1):

- nested logit
- random coefficient logit

The Nested Logit Model

A simple way to relax IIA is to nest similar alternatives.



Choices are nested into "upscale" and "fast food", in order to make the following distinction:

- Choices within nests: Among upscale restaurants p_{ij}/p_{ik} may not depend on the attributes of the fast food restaurant (IIA within nests). E.g., if prices at Bongo Burger increase, Lalime's and Chez Panisse gain clients in a way proportional to their previous market shares.
- Choices between nests: The ratio of choice probabilities, p_{ij}/p_{ik}, between Chez Panisse and Bongo Burger may very well depend on the attributes (e.g., prices) at Lalime's (no IIA between nests).

The Nested Logit Model

Suppose that alternatives are grouped into S nests: $B_1, \ldots, B_S$.

Assume that alternative j belongs to nest s. Conditional on the choice being in nest s, the probability of choosing j is:

$$\frac{e^{X_{ij}'\beta/\rho_s}}{\sum_{k \in B_s} e^{X_{ik}'\beta/\rho_s}}.$$

The probability of choosing an alternative that belongs to nest s is:

$$\frac{e^{Z_s'\alpha}\left(\sum_{k \in B_s} e^{X_{ik}'\beta/\rho_s}\right)^{\rho_s}}{\sum_{r=1}^{S} e^{Z_r'\alpha}\left(\sum_{k \in B_r} e^{X_{ik}'\beta/\rho_r}\right)^{\rho_r}}.$$

The variables Z_s are nest-specific characteristics, often nest-specific constants, in which case $e^{Z_s'\alpha} = e^{\alpha_s}$ (usually normalizing $\alpha_1 = 0$).

The Nested Logit Model

These probabilities can be derived as the solution of a model with a utility-maximizing agent that derives utility from choice j equal to:

$$U_{ij} = X'_{ij}\beta + Z'_s\alpha_s + u_{ij},$$

where the joint distribution function of $(u_{i0}, \ldots, u_{im})$ has a generalized extreme value (GEV) distribution:

$$F(u_{i0}, \ldots, u_{im}) = e^{-\sum_{s=1}^{S}\left(\sum_{j\in B_s} e^{-u_{ij}/\rho_s}\right)^{\rho_s}},$$

where $0 < \rho_s \leq 1$. Between nests u_{ij} and u_{ik} are independent. Within a nest the correlation between u_{ij} and u_{ik} is equal to $1 - \rho_s^2$, where s is the index of the nest.

The Nested Logit Model

Multiplying the probability of j given B_s times the probability of B_s, we obtain the probability of choice j:

$$p_{ij} = \frac{e^{X_{ij}'\beta/\rho_s + Z_s'\alpha}\left(\sum_{k\in B_s} e^{X_{ik}'\beta/\rho_s}\right)^{\rho_s - 1}}{\sum_{r=1}^{S} e^{Z_r'\alpha}\left(\sum_{k\in B_r} e^{X_{ik}'\beta/\rho_r}\right)^{\rho_r}}.$$

Using this formula, it is easy to see that:

- If both j and k belong to the same nest, then p_{ij}/p_{ik} depends only on $X_{ij} - X_{ik}$, so IIA holds.
- If j and k belong to different nests, B_s and B_r, and it is not the case that $\rho_s = \rho_r = 1$, then p_{ij}/p_{ik} may depend on the attributes of other alternatives in B_s and B_r, so IIA does not hold.
- If $\rho_1 = \cdots = \rho_S = 1$ we go back to the conditional logit model.

Random Coefficient Logit

Another way to introduce correlation between choices is to postulate a model of random coefficients. Consider:

$$U_{ij} = X_{ij}'\beta_i + u_{ij},$$

where u_{ij} are independent type I extreme value and the β_i are Normal and independent of any other variable in the model:

$$\beta_i \sim N(\beta_0, \Sigma_0),$$

Let $\varepsilon_i = \beta_i - \beta_0$, then

$$U_{ij} = X_{ij}'\beta_0 + v_{ij},$$

where $v_{ij} = X_{ij}'\varepsilon_i + u_{ij}$ are correlated across alternatives.

Therefore, the model does not impose IIA.

Random Coefficient Logit

For any given values of β and Σ we can obtain the choice probabilities, by first integrating over the distribution of u_{ij} and then over the distribution of β_i:

$$p_{ij}(\beta, \Sigma) = \int \frac{e^{X_{ij}'\beta_i}}{\sum_{k=0}^{m} e^{X_{ik}'\beta_i}} \phi(\beta_i|\beta, \Sigma) d\beta_i,$$

where $\phi(\cdot|\beta, \Sigma)$ is the density of a multivariate Normal variable with mean β and variance Σ.

There is no close-form solution to this integral, which is usually computed using Monte-Carlo simulation:

(1) Obtain a large number $\beta_i^{(1)}, \ldots, \beta_i^{(R)}$ of computer generated values of a $N(\beta, \Sigma)$.

(2) Approximate $p_{ij}(\beta, \Sigma)$ as:

$$\widehat{p}_{ij}(\beta, \Sigma) = \frac{1}{R} \sum_{r=1}^{R} \frac{e^{X_{ij}'\beta_i^{(r)}}}{\sum_{k=0}^{m} e^{X_{ik}'\beta_i^{(r)}}}.$$

Random Coefficient Logit

The parameters β_0 and Σ_0 are estimated by maximizing the simulated log-likelihood:

$$\sum_{i=1}^{N} \sum_{j=1}^{m} Y_{ij} \ln \widehat{p}(\beta, \Sigma),$$

where Y_{ij} takes value one if $Y_i = j$, and value zero otherwise.

This is computationally expensive because the choice probabilities have to be simulated for each individual in each iteration of the optimization procedure (as the values for β and Σ change in each iteration of the optimization procedure).