Lecture 5
Maximum Likelihood Method and Discrete Choice Models

CHUNG-MING KUAN

Department of Finance & CRETA

National Taiwan University

March 15, 2022

Lecture Outline

1. Maximum Likelihood Method
   - Maximum Likelihood Estimation
   - Asymptotic Properties of the Maximum Likelihood Estimator
   - Example: MLE under Conditional Normality
2. Discrete Choice Models
   - Probit and Logit Models
   - Estimation of Probit and Logit Models
   - Marginal Effects and Model Performance
   - Asymptotic Properties
   - Multinomial Logit Model

From OLS to Maximum Likelihood

- Drawbacks of linear regressions: Linear regressions are unable to accommodate certain characteristics of the dependent variable, such as binary and truncated responses, and admit specifications for only the conditional mean of the dependent variable. That is, linear regressions ignore data characteristics and other conditional moments.
- Maximum Likelihood method
  - A likelihood function characterizes the random behavior of the dependent variable and permits specifications for distribution parameters (or moments). Hence, a likelihood function provides a more complete model for the dependent variable.
  - Maximizing the likelihood function with respect to unknown parameters yields the maximum likelihood estimators (MLEs).

Why Maximum Likelihood?

Suppose we want to learn the probability of getting a head from tossing a coin. Let $A = \{\text{Getting a head 8 times out of 10 tosses.}\}$. What would be the value of p that is most likely supported by this event? Given the probability function: $\mathbb{P}(A|p) = C_8^{10} p^8 (1-p)^2$, we have:

$$\begin{aligned}
\mathbb{P}(A|p = 0.5) &= C_8^{10}(0.5)^{10} \approx 0.0439, \\
\mathbb{P}(A|p = 0.7) &= C_8^{10}(0.7)^8(0.3)^2 \approx 0.2335, \\
\mathbb{P}(A|p = 0.8) &= C_8^{10}(0.8)^8(0.2)^2 \approx 0.3020, \\
\mathbb{P}(A|p = 0.9) &= C_8^{10}(0.9)^8(0.1)^2 \approx 0.1937.
\end{aligned}$$

We may call $C_8^{10} p^8 (1-p)^2$ the likelihood function of p given the event A. It thus makes sense to maximize this likelihood function with respect to p; see Amemiya (1994) for related discussion.

Maximum Likelihood Estimator

For the random sample $(y_i, \boldsymbol{x}_i')$, $i = 1, \ldots, n$, let $\ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta})$ denote the conditional likelihood function of y_i given $\boldsymbol{x}_i$, with $\boldsymbol{\theta}$ the parameter vector. Maximizing the joint likelihood function, $\prod_{i=1}^n \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta})$, with respect to $\boldsymbol{\theta}$ is equivalent to maximizing its log transformation:

$$L_n(\boldsymbol{\theta}) = \ln\left(\prod_{i=1}^n \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta})\right) = \sum_{i=1}^n \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta}).$$

Solving the following first order condition for $\boldsymbol{\theta}$:

$$\nabla L_n(\boldsymbol{\theta}) = \sum_{i=1}^n \nabla \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta}) = \sum_{i=1}^n \begin{pmatrix} \frac{\partial \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta})}{\partial \theta_1} \\ \frac{\partial \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta})}{\partial \theta_2} \\ \vdots \\ \frac{\partial \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta})}{\partial \theta_k} \end{pmatrix} = \mathbf{0},$$

we obtain the MLE of $\boldsymbol{\theta}$, denoted as $\tilde{\boldsymbol{\theta}}$.

Consistency

Intuition for establishing MLE consistency: Under a LLN, $n^{-1}\nabla L_n(\boldsymbol{\theta})$ is close to $\mathbb{E}[\nabla \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta})]$ on the parameter space when n becomes large. As the MLE $\tilde{\boldsymbol{\theta}}$ solves $\nabla L_n(\boldsymbol{\theta}) = \mathbf{0}$, it is reasonable to expect, in the limit, $\tilde{\boldsymbol{\theta}}$ is close to the parameter of interest $\boldsymbol{\theta}_o$, the solution to

$$\mathbb{E}[\nabla \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta})] = \mathbf{0}.$$

This suggests that under suitable regularity conditions that ensure a proper LLN, $\tilde{\boldsymbol{\theta}} \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_o$.

Remark: Note that the LLN for a sequence of functions of $\boldsymbol{\theta}$ is different from the LLN learned earlier. Indeed, to ensure that $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_o$ are close in the limit, we require an LLN to hold uniformly in $\boldsymbol{\theta}$, known as a uniform LLN (ULLN).

Asymptotic Normality

By the mean-value expansion of $\nabla L_n(\tilde{\boldsymbol{\theta}})$ about $\boldsymbol{\theta}_o$,

$$\frac{1}{n}\nabla L_n(\tilde{\boldsymbol{\theta}}) = \frac{1}{n}\nabla L_n(\boldsymbol{\theta}_o) + \frac{1}{n}\nabla^2 L_n(\boldsymbol{\theta}^\dagger)(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_o),$$

where $\boldsymbol{\theta}^\dagger$ is between $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_o$, and $\nabla^2 L_n(\boldsymbol{\theta})$ denotes the Hessian matrix, the matrix of the second-order derivatives of $L_n(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$:

$$\nabla^2 L_n(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial \theta_1^2} & \frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_3} & \cdots & \frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial \theta_2^2} & \frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_3} & \cdots & \frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_2} & \frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_3} & \cdots & \frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial \theta_k^2} \end{pmatrix},$$

with the (i,j) th element: $\sum_{i=1}^n \partial^2 \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta}) / \partial \theta_i \partial \theta_j$.

Given $\nabla L_n(\tilde{\boldsymbol{\theta}}) = \mathbf{0}$ (why?), we have from the mean-value expansion that

$$\begin{aligned}\frac{1}{n}\nabla^2 L_n(\boldsymbol{\theta}^\dagger)\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) &= -\frac{1}{\sqrt{n}}\nabla L_n(\boldsymbol{\theta}_o) \\ &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nabla \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta}).\end{aligned}$$

When $(y_i, \boldsymbol{x}_i')'$ are i.i.d., a proper weak ULLN ensures

$$\begin{aligned}\frac{1}{n}\nabla^2 L_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\nabla^2 \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta}) \\ &\xrightarrow{\mathbb{P}} \mathbb{E}[\nabla^2 \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta})],\end{aligned}$$

uniformly in $\boldsymbol{\theta}$. Define $\boldsymbol{H}(\boldsymbol{\theta}) = \mathbb{E}[\nabla^2 \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta})]$, the expected value of the Hessian matrix.

As $\boldsymbol{\theta}^\dagger$ is between $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_o$, the consistency of $\tilde{\boldsymbol{\theta}}$ implies that $\boldsymbol{\theta}^\dagger \approx \boldsymbol{\theta}_o$ in the limit. Following the weak ULLN, $n^{-1}\nabla^2 L_n(\boldsymbol{\theta}^\dagger) \xrightarrow{\mathbb{P}} \boldsymbol{H}(\boldsymbol{\theta}_o)$, so that

$$\begin{aligned}\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) &= -\left(\frac{1}{n}\nabla^2 L_n(\boldsymbol{\theta}^\dagger)\right)^{-1}\left(\frac{1}{\sqrt{n}}\nabla L_n(\boldsymbol{\theta}_o)\right) \\ &\approx -\boldsymbol{H}(\boldsymbol{\theta}_o)^{-1}\left(\frac{1}{\sqrt{n}}\nabla L_n(\boldsymbol{\theta}_o)\right).\end{aligned}$$

To determine the asymptotic distribution of the right-hand side, define:

$$\boldsymbol{B}(\boldsymbol{\theta}) := \operatorname{var}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nabla \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta})\right) = \operatorname{var}(\nabla \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta}));$$

$\boldsymbol{B}(\boldsymbol{\theta}_o)$ is known as the information matrix. By a suitable CLT,

$$\boldsymbol{B}(\boldsymbol{\theta}_o)^{-1/2}\frac{1}{\sqrt{n}}\nabla L_n(\boldsymbol{\theta}_o) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{I}).$$

Combining the results above, we have

$$\begin{aligned}\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) &\approx -\boldsymbol{H}(\theta_o)^{-1}\frac{1}{\sqrt{n}}\nabla L_n(\boldsymbol{\theta}_o) \\ &= -\boldsymbol{H}(\theta_o)^{-1}\boldsymbol{B}(\boldsymbol{\theta}_o)^{1/2}\left[\boldsymbol{B}(\boldsymbol{\theta}_o)^{-1/2}\frac{1}{\sqrt{n}}\nabla L_n(\boldsymbol{\theta}_o)\right] \\ &\xrightarrow{D} -\boldsymbol{H}(\theta_o)^{-1}\boldsymbol{B}(\boldsymbol{\theta}_o)^{1/2}\mathcal{N}(\mathbf{0}, \boldsymbol{I}).\end{aligned}$$

This shows that

$$\begin{aligned}\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) &\xrightarrow{D} -\boldsymbol{H}(\boldsymbol{\theta}_o)^{-1}\,\boldsymbol{B}(\boldsymbol{\theta}_o)^{1/2}\mathcal{N}(\mathbf{0}, \boldsymbol{I}) \\ &\stackrel{d}{=} \mathcal{N}\Big(\mathbf{0},\ \boldsymbol{H}(\boldsymbol{\theta}_o)^{-1}\boldsymbol{B}(\boldsymbol{\theta}_o)\boldsymbol{H}(\boldsymbol{\theta}_o)^{-1}\Big),\end{aligned}$$

where $\stackrel{d}{=}$ denotes equality in distribution. We use $\boldsymbol{D}_o$ to denote the asymptotic covariance matrix $\boldsymbol{H}(\boldsymbol{\theta}_o)^{-1}\boldsymbol{B}(\boldsymbol{\theta}_o)\boldsymbol{H}(\boldsymbol{\theta}_o)^{-1}$.

As before, $\boldsymbol{D}_o$ can be consistently estimated by $\tilde{\boldsymbol{D}} = \tilde{\boldsymbol{H}}^{-1}\tilde{\boldsymbol{B}}\tilde{\boldsymbol{H}}^{-1}$, where $\tilde{\boldsymbol{H}}$ and $\tilde{\boldsymbol{B}}$ are, respectively, the sample counterparts of $\boldsymbol{H}(\boldsymbol{\theta}_o)$ and $\boldsymbol{B}(\boldsymbol{\theta}_o)$, both evaluated at the MLE $\tilde{\boldsymbol{\theta}}$. For example, the $(i,j)^{\text{th}}$ element of $\tilde{\boldsymbol{H}}$ is

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}.$$

The estimator $\tilde{\boldsymbol{D}}$ is also referred to as the Eicker-White covariance matrix estimator. For the hypothesis $\theta_j = c$, the Wald statistic is:

$$\frac{\tilde{\theta}_j - c}{\text{EW-se}(\tilde{\theta}_j)} \xrightarrow{D} \mathcal{N}(0,1),$$

where EW-se$(\tilde{\theta}_j)$ is the square root of the j^{th} diagonal term of $\tilde{\boldsymbol{D}}/n$, an Eicker-White standard error. The Wald tests of joint hypotheses can be computed similarly.

Information Matrix Equality

When the likelihood function is specified correctly, we have the well-known information matrix equality:

$$\boldsymbol{H}(\boldsymbol{\theta}_o) + \boldsymbol{B}(\boldsymbol{\theta}_o) = \mathbf{0}, \quad \text{or} \quad \boldsymbol{B}(\boldsymbol{\theta}_o) = -\boldsymbol{H}(\boldsymbol{\theta}_o).$$

In this case, the asymptotic covariance matrix simplifies to

$$\boldsymbol{H}(\boldsymbol{\theta}_o)^{-1}\boldsymbol{B}(\boldsymbol{\theta}_o)\boldsymbol{H}(\boldsymbol{\theta}_o)^{-1} = -\boldsymbol{H}(\boldsymbol{\theta}_o)^{-1} = \boldsymbol{B}(\boldsymbol{\theta}_o)^{-1},$$

where $\boldsymbol{B}(\boldsymbol{\theta}_o)^{-1}$ is the Cramér-Rao lower bound of the variance of unbiased estimators. Thus,

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, -\boldsymbol{H}(\boldsymbol{\theta}_o)^{-1}\right) \overset{d}{=} \mathcal{N}\left(\mathbf{0}, \boldsymbol{B}(\boldsymbol{\theta}_o)^{-1}\right).$$

On the other hand, the information matrix equality fails under misspecified likelihood function, and the asymptotic covariance matrix remains the "sandwich" form.

When the information matrix equality holds, the Wald statistic for the hypothesis $\theta_j = c$ is

$$\frac{\tilde{\theta}_j - c}{\text{se}(\tilde{\theta}_j)} \xrightarrow{D} \mathcal{N}(0,1),$$

where $\text{se}(\tilde{\theta}_j)$ is the square root of the j^{th} diagonal term of $-\tilde{\boldsymbol{H}}^{-1}/n$, i.e.,

$$-\left[\sum_{i=1}^{n} \frac{\partial^2 \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta})}{\partial\theta_j \partial\theta_j}\bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}\right]^{-1}.$$

For the multiple hypotheses $\boldsymbol{R\theta} = \boldsymbol{c}$, where $\boldsymbol{R}$ is $q \times k$, the Wald statistic has a χ^2 distribution in the limit:

$$-n(\boldsymbol{R}\tilde{\boldsymbol{\theta}} - \boldsymbol{c})'[\boldsymbol{R}\tilde{\boldsymbol{H}}^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\tilde{\boldsymbol{\theta}} - \boldsymbol{c}) \xrightarrow{D} \chi^2(q).$$

Example: MLE under Conditional Normality

Assume conditional normality for y_i: $y_i|\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2)$, such that the specification of the likelihood function is

$$\ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta}) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}.$$

The log of joint likelihood function is:

$$L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n} \frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}.$$

Straightforward calculation shows

$$\nabla \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta}) = \begin{bmatrix} \frac{\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})}{\sigma^2} \\ -\frac{1}{2\sigma^2} + \frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2(\sigma^2)^2} \end{bmatrix}.$$

Hence,

$$\nabla L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \nabla \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta}) = \begin{bmatrix} \frac{\sum_{i=1}^{n} \boldsymbol{x}_i (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})}{\sigma^2} \\ -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2(\sigma^2)^2} \end{bmatrix}.$$

Solving $\nabla L_n(\boldsymbol{\theta}) = \mathbf{0}$ is equivalent to solving

$$\begin{bmatrix} \sum_{i=1}^{n} \boldsymbol{x}_i (y_i - \boldsymbol{x}_i'\boldsymbol{\beta}) \\ -n + \frac{\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{\sigma^2} \end{bmatrix} = \mathbf{0},$$

from which we obtain the MLEs for $\boldsymbol{\beta}$ and σ^2:

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i \right),$$

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})^2 = \frac{1}{n} \sum_{i=1}^{n} \tilde{u}_i^2,$$

where $\tilde{u}_i$ are the MLE (OLS) residuals. Note that $\tilde{\boldsymbol{\beta}}$ is the OLS estimator.

When the specification is correct for $y_i|\boldsymbol{x}_i$, there exists $\boldsymbol{\theta}_o = (\boldsymbol{\beta}_o' \ \sigma_o^2)'$ such that $y_i|\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{x}_i'\boldsymbol{\beta}_o, \sigma_o^2)$. It follows from the law of iterated expectation:

$$\mathbb{E}[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})]|_{\boldsymbol{\beta}=\boldsymbol{\beta}_o} = \mathbb{E}[\boldsymbol{x}_i(\mathbb{E}(y_i|\boldsymbol{x}_i) - \boldsymbol{x}_i'\boldsymbol{\beta}_o)] = \mathbf{0},$$

because the conditional mean of y_i is $\boldsymbol{x}_i'\boldsymbol{\beta}_o$. Similarly,

$$\mathbb{E}[(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2]|_{\boldsymbol{\beta}=\boldsymbol{\beta}_o} = \mathbb{E}\big\{\mathbb{E}[(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_o)^2]|\boldsymbol{x}_i\big\} = \sigma_o^2,$$

because the conditional variance of y_i is σ_o^2. Thus, $\boldsymbol{\theta}_o = (\boldsymbol{\beta}_o', \sigma_o^2)$ solves

$$\mathbb{E}[\nabla \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta})] = \mathbb{E}\begin{bmatrix} \frac{\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})}{\sigma^2} \\ -\frac{1}{2\sigma^2} + \frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2(\sigma^2)^2} \end{bmatrix} = \mathbf{0}.$$

That is, the parameter of interest is $\boldsymbol{\theta}_o$, for which $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}', \tilde{\sigma}^2)$ is consistent.

Moreover,

$$\nabla^2 \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta}) = \begin{bmatrix} -\frac{\boldsymbol{x}_i \boldsymbol{x}_i'}{\sigma^2} & -\frac{\boldsymbol{x}_i (y_i - \boldsymbol{x}_i' \boldsymbol{\beta})}{(\sigma^2)^2} \\ -\frac{(y_i - \boldsymbol{x}_i' \boldsymbol{\beta}) \boldsymbol{x}_i'}{(\sigma^2)^2} & \frac{1}{2(\sigma^2)^2} - \frac{(y_i - \boldsymbol{x}_i' \boldsymbol{\beta})^2}{(\sigma^2)^3} \end{bmatrix},$$

and its expected value is

$$\boldsymbol{H}(\boldsymbol{\theta}) = \begin{bmatrix} -\frac{\mathbb{E}(\boldsymbol{x}_i \boldsymbol{x}_i')}{\sigma^2} & -\frac{\mathbb{E}[\boldsymbol{x}_i (y_i - \boldsymbol{x}_i' \boldsymbol{\beta})]}{(\sigma^2)^2} \\ -\frac{\mathbb{E}[(y_i - \boldsymbol{x}_i' \boldsymbol{\beta}) \boldsymbol{x}_i']}{(\sigma^2)^2} & \frac{1}{2(\sigma^2)^2} - \frac{\mathbb{E}(y_i - \boldsymbol{x}_i' \boldsymbol{\beta})^2}{(\sigma^2)^3} \end{bmatrix}.$$

When $\boldsymbol{H}(\boldsymbol{\theta})$ is evaluated at $\boldsymbol{\theta}_o$,

$$\boldsymbol{H}(\boldsymbol{\theta}_o) = \begin{bmatrix} -\frac{\mathbb{E}(\boldsymbol{x}_i \boldsymbol{x}_i')}{\sigma_o^2} & \mathbf{0} \\ \mathbf{0}' & -\frac{1}{2\sigma_o^4} \end{bmatrix}.$$

Recall that $\boldsymbol{B}(\boldsymbol{\theta}) := \mathrm{var}(\nabla \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta}))$. Evaluating $\boldsymbol{B}(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_o$ yields

$$
\begin{aligned}
&\mathbb{E}[(\nabla \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta}))(\nabla \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta}))']|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \\
&= \mathbb{E}\left(\begin{bmatrix} \frac{\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})}{\sigma^2} \\ -\frac{1}{2\sigma^2} + \frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2(\sigma^2)^2} \end{bmatrix} \begin{bmatrix} \frac{\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})}{\sigma^2} \\ -\frac{1}{2\sigma^2} + \frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2(\sigma^2)^2} \end{bmatrix}' \right)\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \\
&= \begin{bmatrix} \frac{\mathbb{E}[(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_o)^2 \boldsymbol{x}_i \boldsymbol{x}_i']}{\sigma_o^4} & \mathbf{0} \\ \mathbf{0}' & \frac{1}{4\sigma_o^4} + \frac{\mathbb{E}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_o)^4}{4\sigma_o^8} - \frac{\mathbb{E}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_o)^2}{2\sigma_o^6} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\mathbb{E}(\boldsymbol{x}_i \boldsymbol{x}_i')}{\sigma_o^2} & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2\sigma_o^4} \end{bmatrix},
\end{aligned}
$$

where $\mathbb{E}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_o)^4 = \mathbb{E}[\mathbb{E}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_o)^4 | \boldsymbol{x}_i)] = 3\sigma_o^4$, by the property of the fourth moment of conditional normality. This proves the information matrix equality: $\boldsymbol{H}(\boldsymbol{\theta}_o) + \boldsymbol{B}(\boldsymbol{\theta}_o) = \mathbf{0}$.

We have shown that, when the model $y_i|\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2)$ is correctly specified,

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \xrightarrow{D} \mathcal{N}(\mathbf{0}, -\boldsymbol{H}(\boldsymbol{\theta}_o)^{-1}).$$

In particular,

$$\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) \xrightarrow{D} \mathcal{N}\big(\mathbf{0}, \sigma_o^2[\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1}\big),$$

where the asymptotic covariance matrix agrees with that of the OLS estimator under conditional homoskedasticity, and

$$\sqrt{n}(\tilde{\sigma}^2 - \sigma_o^2) \xrightarrow{D} \mathcal{N}\big(\mathbf{0}, 2\sigma_o^4\big).$$

On the other hand, when data are conditionally heteroskedastic, e.g., $h(\boldsymbol{x}_t'\boldsymbol{\alpha})$, the specification σ^2 in $y_i|\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2)$ is incorrect, and the derivation of the information matrix equality above would be invalid.

Likelihoods of Binary Variable

In practice, we may be interested in learning the individual characteristics that affect ownership of a car or attendance of an event. In such cases, the dependent variable is binary such that $y_i = 1$ when the individual i owns a car or attends an event, and $y_i = 0$ otherwise. Writing

$$y_i = \begin{cases} 1, & \text{with conditional probability } \mathbb{P}(y_i = 1|\boldsymbol{x}_i), \\ 0, & \text{with conditional probability } 1 - \mathbb{P}(y_i = 1|\boldsymbol{x}_i), \end{cases}$$

where $\boldsymbol{x}_i$ are the variables that may affect the decision y_i. The likelihood function of y_i given $\boldsymbol{x}_i$ is then

$$\mathbb{P}(y_i = 1|\boldsymbol{x}_i)^{y_i}[1 - \mathbb{P}(y_i = 1|\boldsymbol{x}_i)]^{1-y_i}.$$

It is common to specify a distribution $F(\boldsymbol{x}_i;\boldsymbol{\theta})$ for the unknown probability $\mathbb{P}(y_i = 1|\boldsymbol{x}_i)$, where $\boldsymbol{\theta}$ is $k \times 1$. The specified likelihood function for individual i is

$$\ell(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}) = F(\boldsymbol{x}_i;\boldsymbol{\theta})^{y_i}[1 - F(\boldsymbol{x}_i;\boldsymbol{\theta})]^{1-y_i}.$$

We can compute the MLE $\tilde{\boldsymbol{\theta}}$ by maximizing

$$\begin{aligned} L_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \ln \ell(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \big[y_i \ln F(\boldsymbol{x}_i;\boldsymbol{\theta}) + (1 - y_i) \ln(1 - F(\boldsymbol{x}_i;\boldsymbol{\theta}))\big]. \end{aligned}$$

The resulting MLE depends on the specification $F(\boldsymbol{x}_i;\boldsymbol{\theta})$.

Probit and Logit Models

The probit and logit models are two different specifications of $F(\boldsymbol{x}_i;\boldsymbol{\theta})$.

- Probit model:

$$F(\boldsymbol{x}_i;\boldsymbol{\theta}) = \Phi(\boldsymbol{x}_i'\boldsymbol{\theta}) = \int_{-\infty}^{\boldsymbol{x}_i'\boldsymbol{\theta}} \frac{1}{\sqrt{2\pi}} e^{-v^2/2}\,\mathrm{d}v,$$

  where Φ denotes the standard normal distribution function.

- Logit model:

$$F(\boldsymbol{x}_i;\boldsymbol{\theta}) = G(\boldsymbol{x}_i'\boldsymbol{\theta}) = \frac{1}{1+\exp(-\boldsymbol{x}_i'\boldsymbol{\theta})} = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\theta})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\theta})},$$

  where G is the logistic distribution function with mean zero and variance $\pi^2/3$. Note that the logistic distribution has thicker tails than the standard normal distribution.

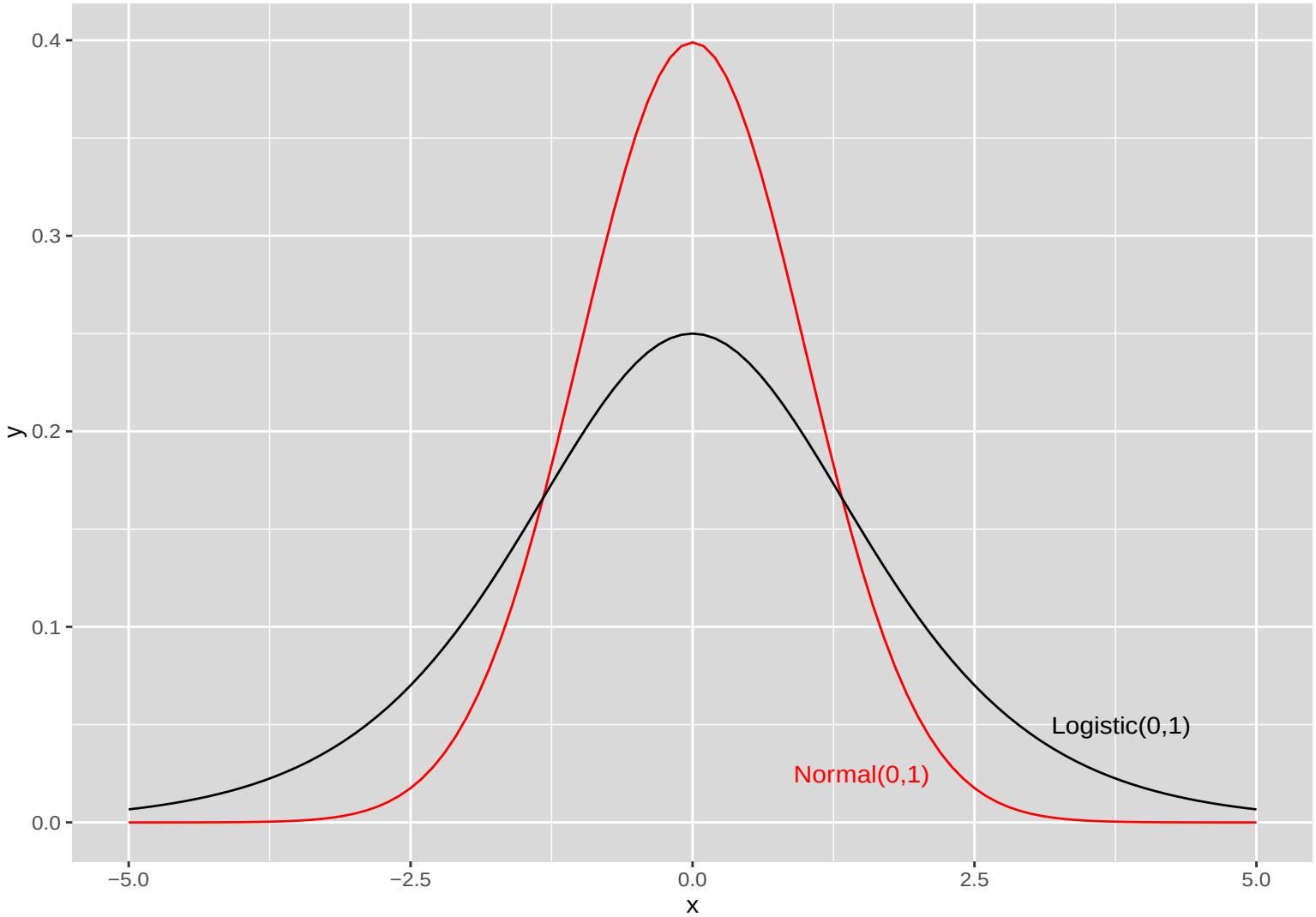

Figure: The logistic distribution vs. standard normal distribution.

For the probit model, $\mathrm{d}\Phi(u)/\,\mathrm{d}u = \phi(u)$, where ϕ is the standard normal density function, and hence

$$\begin{aligned}\nabla L_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n}\left[y_i\frac{\phi(\boldsymbol{x}_i'\boldsymbol{\theta})}{\Phi(\boldsymbol{x}_i'\boldsymbol{\theta})} - (1-y_i)\frac{\phi(\boldsymbol{x}_i'\boldsymbol{\theta})}{1-\Phi(\boldsymbol{x}_i'\boldsymbol{\theta})}\right]\boldsymbol{x}_i \\ &= \sum_{i=1}^{n}\frac{y_i - \Phi(\boldsymbol{x}_i'\boldsymbol{\theta})}{\Phi(\boldsymbol{x}_i'\boldsymbol{\theta})[1-\Phi(\boldsymbol{x}_i'\boldsymbol{\theta})]}\phi(\boldsymbol{x}_i'\boldsymbol{\theta})\boldsymbol{x}_i.\end{aligned}$$

Setting the right-hand side to zero yields a system of k nonlinear functions:

$$\sum_{i=1}^{n}\frac{y_i - \Phi(\boldsymbol{x}_i'\boldsymbol{\theta})}{\Phi(\boldsymbol{x}_i'\boldsymbol{\theta})[1-\Phi(\boldsymbol{x}_i'\boldsymbol{\theta})]}\phi(\boldsymbol{x}_i'\boldsymbol{\theta})\boldsymbol{x}_i = \mathbf{0}.$$

Unlike the MLE under conditional normality, the MLE $\tilde{\boldsymbol{\theta}}$ cannot be solved analytically. Yet, $\tilde{\boldsymbol{\theta}}$ can be computed using numerical methods. Plugging $\tilde{\boldsymbol{\theta}}$ into $\Phi(\boldsymbol{x}_i'\boldsymbol{\theta})$ we obtain the predicted probabilities $\Phi(\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}})$.

For the logit model, $\mathrm{d}\,G(u)/\,\mathrm{d}\,u = G'(u) = G(u)[1 - G(u)]$, so that

$$\begin{aligned}
\nabla L_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \left[y_i \frac{G'(\boldsymbol{x}_i'\boldsymbol{\theta})}{G(\boldsymbol{x}_i'\boldsymbol{\theta})} - (1 - y_i) \frac{G'(\boldsymbol{x}_i'\boldsymbol{\theta})}{1 - G(\boldsymbol{x}_i'\boldsymbol{\theta})} \right] \boldsymbol{x}_i \\
&= \sum_{i=1}^{n} \left\{ y_i [1 - G(\boldsymbol{x}_i'\boldsymbol{\theta})] - (1 - y_i) G(\boldsymbol{x}_i'\boldsymbol{\theta}) \right\} \boldsymbol{x}_i \\
&= \sum_{i=1}^{n} [y_i - G(\boldsymbol{x}_i'\boldsymbol{\theta})] \boldsymbol{x}_i .
\end{aligned}$$

The MLE $\tilde{\boldsymbol{\theta}}$ is obtained by numerically solving the nonlinear system:

$$\sum_{i=1}^{n} [y_i - G(\boldsymbol{x}_i'\boldsymbol{\theta})] \boldsymbol{x}_i = \mathbf{0}.$$

Plugging $\tilde{\boldsymbol{\theta}}$ into $G(\boldsymbol{x}_i'\boldsymbol{\theta})$ we obtain the predicted probabilities $G(\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}})$.

Remark: A Binary y_i has conditional mean $\mathbb{E}(y_i|\boldsymbol{x}_i) = \mathbb{P}(y_i = 1|\boldsymbol{x}_i)$ and conditional variance:

$$\text{var}(y_i|\boldsymbol{x}_i) = \mathbb{P}(y_i = 1|\boldsymbol{x}_i)[1 - \mathbb{P}(y_i = 1|\boldsymbol{x}_i)].$$

That is, y_i are conditionally heteroskedastic. Given $F(\boldsymbol{x}_i; \boldsymbol{\theta})$ is a specification for $\mathbb{P}(y_i = 1|\boldsymbol{x}_i) = \mathbb{E}(y_i = 1|\boldsymbol{x}_i)$, we may write

$$y_i = F(\boldsymbol{x}_i; \boldsymbol{\theta}) + u_i,$$

and estimate $\boldsymbol{\theta}$ by the nonlinear LS (NLS) method, i.e., minimizing $\sum_{i=1}^n [y_i - F(\boldsymbol{x}_i; \boldsymbol{\theta})]^2$ with respect to $\boldsymbol{\theta}$. Note that the NLS estimator is not the same as the MLE discussed earlier (check), and it is not efficient because the objective function does not take into account conditional heteroskedasticity and binary feature of y.

Linear Probability Model

One may consider estimating the linear regression for the binary variable y:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\theta} + u_i.$$

This approach, however, completely ignores the binary feature of y, and the linear specification is a poor approximation to the conditional mean function $F(\boldsymbol{x}, \boldsymbol{\theta})$. In particular, its fitted values are not bounded between zero and one, and a fitted value greater than one or less than zero cannot be interpreted as a probability. This also explains why the probit and logit models are preferred in practice.

Marginal Effects

Given Φ and G are nonlinear functions, it should be emphasized that the estimated coefficients are not the marginal effects of the regressors on y. This is very different from the results of linear regressions. Indeed, for the probit model, the marginal effect of the j^{th} regressor is:

$$\frac{\partial \Phi(\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}})}{\partial x_{ij}} = \phi(\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}})\tilde{\theta}_j, \quad j = 1, \ldots, k;$$

for the logit model,

$$\frac{\partial G(\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}})}{\partial x_{ij}} = G(\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}})[1 - G(\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}})]\tilde{\theta}_j, \quad j = 1, \ldots, k.$$

Thus, the marginal effects are the products of $\tilde{\theta}_j$ and a scaling factor that changes with $\boldsymbol{x}_i$.

To circumvent the problem of changing factors, one may evaluate the marginal effects at the sample average $\bar{\boldsymbol{x}}$:

$$\phi(\bar{\boldsymbol{x}}'\tilde{\boldsymbol{\theta}})\tilde{\theta}_j, \quad \text{or} \quad G(\bar{\boldsymbol{x}}'\tilde{\boldsymbol{\theta}})[1 - G(\bar{\boldsymbol{x}}'\tilde{\boldsymbol{\theta}})]\tilde{\theta}_j.$$

This is known as the marginal effect at the average.

Computing the marginal effect at the average makes sense when regressors are continuous variables. If one of the regressors is binary, say, a gender dummy variable, its sample average is the sample proportion of a gender. It would be hard to interpret the resulting marginal effect (for example, what does it mean by the marginal effect at gender = 0.42?). When a regressor is not continuous, one may compute the average marginal effect instead:

$$\frac{1}{n}\sum_{i=1}^{n} \phi(\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}})\tilde{\theta}_j, \quad \text{or} \quad \frac{1}{n}\sum_{i=1}^{n} G(\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}})[1 - G(\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}})]\tilde{\theta}_j.$$

When there are continuous and binary regressors, let $\bar{\boldsymbol{x}}(1)$ denote the vector that contains the sample averages of all continuous regressors and 1 for the binary regressor. Similarly, $\bar{\boldsymbol{x}}(0)$ is the same as $\bar{\boldsymbol{x}}(1)$, except that 1 is replaced by 0 for the binary regressor. The marginal effect of this binary regressor may be computed as:

$$\Phi(\bar{\boldsymbol{x}}(1)'\tilde{\boldsymbol{\theta}}) - \Phi(\bar{\boldsymbol{x}}(0)'\tilde{\boldsymbol{\theta}}), \quad \text{or} \quad G(\bar{\boldsymbol{x}}(1)'\tilde{\boldsymbol{\theta}}) - G(\bar{\boldsymbol{x}}(0)'\tilde{\boldsymbol{\theta}}).$$

Remark: Observe that the odds ratio is $G(\boldsymbol{x}_i'\boldsymbol{\theta})/[1 - G(\boldsymbol{x}_i'\boldsymbol{\theta})] = \exp(\boldsymbol{x}_i'\boldsymbol{\theta})$, i.e., the probability of $y = 1$ relative to the probability of $y = 0$. Thus,

$$\log\left(\frac{G(\boldsymbol{x}_i'\boldsymbol{\theta})}{1 - G(\boldsymbol{x}_i'\boldsymbol{\theta})}\right) = \boldsymbol{x}_i'\boldsymbol{\theta}.$$

so that the coefficient θ_j can be understood as the marginal effect of x_j on the log odds ratio.

Model Performance

In practice, we use the predicted probabilities $G(\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}})$ or $\Phi(\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}})$ to determine the model prediction. A model would predict the outcome 1 if the predicted probability is greater than a threshold value; otherwise, the prediction would be zero.

- A natural candidate for the threshold is 0.5. An outcome is considered more probable if the predicted probability is greater than 0.5. Yet, this threshold makes sense when the sample proportions of $y_i = 1$ and $y_i = 0$ are approximately equal.
- A better threshold is $\bar{y}$, the sample proportion of $y_i = 1$. Since the predicted probabilities are approximations to $\mathbb{E}(y_i|\boldsymbol{x}_i)$, an outcome is considered more probable when the conditional mean is greater than the sample average (unconditional mean).

For each i, there are 4 possible outcome pairs:

$$(y_i, \text{prediction}_i) = (1,1),\ (1,0),\ (0,1),\ (0,0),$$

where $(1,1)$ and $(0,0)$ are correct predictions, and $(0,1)$ and $(1,0)$ are incorrect predictions. The model performance is based on the percentage correctly predicted, i.e., the proportion of the pairs $(1,1)$ and $(0,0)$ out of n observations:

$$\frac{1}{n}[\text{Number of } (1,1) + \text{Number of } (0,0)].$$

We may also compute the percentage correctly predicted for each group: the proportions of $(1,1)$ (or $(0,0)$) out of the number of $y_i = 1$ (or $y_i = 0$):

$$\frac{\text{Number of } (1,1)}{\text{Number of } y_i = 1}, \quad \text{or} \quad \frac{\text{Number of } (0,0)}{\text{Number of } y_i = 0}.$$

Asymptotic Properties

Under suitable regularity conditions, the MLE of the logit model is $\tilde{\boldsymbol{\theta}} \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_o$, and

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \xrightarrow{D} \mathcal{N}\big(\mathbf{0},\, \boldsymbol{H}(\boldsymbol{\theta}_o)^{-1}\boldsymbol{B}(\boldsymbol{\theta}_o)\boldsymbol{H}(\boldsymbol{\theta}_o)^{-1}\big),$$

where $\boldsymbol{H}(\boldsymbol{\theta}_o) = \mathbb{E}[\nabla^2 \ln \ell(\boldsymbol{\theta}_o)]$, so that

$$\begin{aligned}\boldsymbol{H}(\boldsymbol{\theta}_o) &= \mathbb{E}\big[\nabla\{[y_i - G(\boldsymbol{x}_i'\boldsymbol{\theta}_o)]\boldsymbol{x}_i\}\big] \\ &= -\mathbb{E}\big[G(\boldsymbol{x}_i'\boldsymbol{\theta}_o)(1 - G(\boldsymbol{x}_i'\boldsymbol{\theta}_o))\boldsymbol{x}_i\boldsymbol{x}_i'\big],\end{aligned}$$

and $\boldsymbol{B}(\boldsymbol{\theta}_o) = \mathrm{var}(\nabla \ln \ell(\boldsymbol{\theta}_o))$, so that

$$\boldsymbol{B}(\boldsymbol{\theta}_o) = \mathrm{var}\big([y_i - G(\boldsymbol{x}_i'\boldsymbol{\theta}_o)]\boldsymbol{x}_i\big) = \mathbb{E}\big[(y_i - G(\boldsymbol{x}_i'\boldsymbol{\theta}_o))^2\boldsymbol{x}_i\boldsymbol{x}_i'\big].$$

When $G(\boldsymbol{x}_i'\boldsymbol{\theta})$ is correctly specified for $\mathbb{E}(y_i|\boldsymbol{x}_i)$,

$$\begin{aligned}\mathbb{E}\big[(y_i - G(\boldsymbol{x}_i'\boldsymbol{\theta}_o))^2\boldsymbol{x}_i\boldsymbol{x}_i'\big] &= \mathbb{E}\big[\mathbb{E}\big((y_i - G(\boldsymbol{x}_i'\boldsymbol{\theta}_o))^2|\boldsymbol{x}_i\big)\boldsymbol{x}_i\boldsymbol{x}_i'\big] \\ &= \mathbb{E}\big[G(\boldsymbol{x}_i'\boldsymbol{\theta}_o)(1 - G(\boldsymbol{x}_i'\boldsymbol{\theta}_o))\boldsymbol{x}_i\boldsymbol{x}_i'\big],\end{aligned}$$

because $\mathbb{E}\big((y_i - G(\boldsymbol{x}_i'\boldsymbol{\theta}_o))^2|\boldsymbol{x}_i\big)$ is the conditional variance of y_i. This shows that the information matrix equality holds: $\boldsymbol{B}(\boldsymbol{\theta}_o) + \boldsymbol{H}(\boldsymbol{\theta}_o) = \boldsymbol{0}$, and

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \xrightarrow{D} \mathcal{N}\big(\boldsymbol{0}, -\boldsymbol{H}(\boldsymbol{\theta}_o)^{-1}\big).$$

The covariance matrix $-\boldsymbol{H}(\boldsymbol{\theta}_o)^{-1}$ can be consistently estimated by its sample counterpart:

$$-\tilde{\boldsymbol{H}}^{-1} = \left(\frac{1}{n}\sum_{i=1}^{n} G(\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}})[1 - G(\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}})]\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}.$$

The result for the probit model can be derived similarly (homework).

Similar to linear regressions, the hypothesis $\theta_j = c$ can be tested using the Wald statistic:

$$(\tilde{\theta}_j - c)/\text{se}(\tilde{\theta}_j),$$

where $\text{se}(\tilde{\theta}_j)$ is the square root of the j^{th} diagonal term of $-\tilde{\boldsymbol{H}}^{-1}/n$. The Wald statistic for multiple hypotheses can be computed similarly. Note that some researchers use the Eicker-White standard errors based on the "sandwich" estimator $\tilde{\boldsymbol{H}}^{-1}\tilde{\boldsymbol{B}}\tilde{\boldsymbol{H}}^{-1}$, with

$$\tilde{\boldsymbol{B}} = \frac{1}{n}\sum_{i=1}^{n}[y_i - G(\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}})]^2\boldsymbol{x}_i\boldsymbol{x}_i'.$$

Example: Labor Force Participation

Wooldridge (2016, p. 570): Labor force participation by a married woman during 1975. The estimated logit and probit models are, respectively,

$$\begin{aligned} &0.425 - 0.021\,\text{nwifeinc} + 0.221\,\text{educ} + 0.206\,\text{exper} - 0.0032\,\text{exper}^2 \\ &\quad - 0.088\,\text{age} - 1.443\,\text{kidslt6} + 0.06\,\text{kidsge6}, \\ &0.270 - 0.012\,\text{nwifeinc} + 0.131\,\text{educ} + 0.123\,\text{exper} - 0.0019\,\text{exper}^2 \\ &\quad - 0.053\,\text{age} - 0.868\,\text{kidslt6} + 0.036\,\text{kidsge6}, \end{aligned}$$

where "nwifeinc" denotes husband's income and "kidslt6" denotes the number of kids less than 6-year old. Note that the estimated coefficients are not the marginal response to the change of regressors.

Likelihood of Multinomial Variable

In practice, an individual (firm) may face more than 2 choices, e.g., employment status and commuting mode. Suppose there are $J+1$ mutually exclusive choices that do not have a natural ordering. Let that

$$y_i = \begin{cases} 0, & \text{with probability } \mathbb{P}(y_i = 0|\boldsymbol{x}_i), \\ 1, & \text{with probability } \mathbb{P}(y_i = 1|\boldsymbol{x}_i), \\ \vdots & \\ J, & \text{with probability } \mathbb{P}(y_i = J|\boldsymbol{x}_i). \end{cases}$$

Define the new binary variable $d_{i,j}$, $j = 0, 1, \ldots, J$, as

$$d_{i,j} = \begin{cases} 1, & \text{if } y_i = j, \\ 0, & \text{otherwise;} \end{cases}$$

note that $\sum_{j=0}^{J} d_{i,j} = 1$.

For individual i, the conditional density of $d_{i,0}, \ldots, d_{i,J}$ is

$$g(d_{i,0}, \ldots, d_{i,J} | \boldsymbol{x}_i) = \prod_{j=0}^{J} \mathbb{P}(y_i = j | \boldsymbol{x}_i)^{d_{i,j}}.$$

We may specify a distribution function $F_j(\boldsymbol{x}_i; \boldsymbol{\theta})$ for $\mathbb{P}(y_i = j | \boldsymbol{x}_i)$ and obtain the specified log-likelihood function:

$$L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=0}^{J} d_{i,j} \ln F_j(\boldsymbol{x}_i; \boldsymbol{\theta}).$$

The first order condition is

$$\nabla L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=0}^{J} d_{i,j} \frac{1}{F_j(\boldsymbol{x}_i; \boldsymbol{\theta})} \left[\nabla F_j(\boldsymbol{x}_i; \boldsymbol{\theta})\right] = \mathbf{0},$$

from which we can solve for the MLE $\tilde{\boldsymbol{\theta}}$.

Multinomial Logit Model

A common specification for conditional probability is:

$$F_j(\boldsymbol{x}_i;\boldsymbol{\theta}) = G_{i,j} = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\theta}_j)}{\sum_{k=0}^{J}\exp(\boldsymbol{x}_i'\boldsymbol{\theta}_k)}, \quad j = 0,\ldots,J,$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_0' \ \boldsymbol{\theta}_1' \ \ldots \ \boldsymbol{\theta}_J')'$. Note that individual characteristics $\boldsymbol{x}_i$ in this model do not vary with choices. This specification has a problem that

$$\begin{aligned}
&\frac{\exp[\boldsymbol{x}_i'(\boldsymbol{\theta}_j+\boldsymbol{\gamma})]}{\exp[\boldsymbol{x}_i'(\boldsymbol{\theta}_j+\boldsymbol{\gamma})] + \sum_{k\neq j}\exp(\boldsymbol{x}_i'\boldsymbol{\theta}_k)} \\
&\qquad = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\theta}_j)}{\exp(\boldsymbol{x}_i'\boldsymbol{\theta}_j) + \sum_{k\neq j}\exp[\boldsymbol{x}_i'(\boldsymbol{\theta}_k-\boldsymbol{\gamma})]}, \quad j = 0,1,\ldots,J;
\end{aligned}$$

that is, the parameters are not identified because they may take any value.

To identify the parameters properly, set $\boldsymbol{\theta}_0 = \mathbf{0}$ so that

$$F_0(\boldsymbol{x}_i;\boldsymbol{\theta}) = G_{i,0} = \frac{1}{1+\sum_{k=1}^{J}\exp(\boldsymbol{x}_i'\boldsymbol{\theta}_k)},$$

$$F_j(\boldsymbol{x}_i;\boldsymbol{\theta}) = G_{i,j} = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\theta}_j)}{1+\sum_{k=1}^{J}\exp(\boldsymbol{x}_i'\boldsymbol{\theta}_k)}, \quad j = 1,\ldots,J,$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1' \ \boldsymbol{\theta}_2' \ \ldots \ \boldsymbol{\theta}_J')'$, and $G_{i,0}$ is the "base" choice. This leads to the multinomial logit model, with the log-likelihood function:

$$\begin{aligned} L_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n}\sum_{j=0}^{J} d_{i,j}\log G_{i,j} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J} d_{i,j}\boldsymbol{x}_i'\boldsymbol{\theta}_j - \sum_{j=0}^{J} d_{i,j}\left[\sum_{i=1}^{n}\log\left(1+\sum_{k=1}^{J}\exp(\boldsymbol{x}_i'\boldsymbol{\theta}_k)\right)\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J} d_{i,j}\boldsymbol{x}_i'\boldsymbol{\theta}_j - \sum_{i=1}^{n}\log\left(1+\sum_{k=1}^{J}\exp(\boldsymbol{x}_i'\boldsymbol{\theta}_k)\right). \end{aligned}$$

It is easy to see that

$$\begin{aligned}\nabla_{\boldsymbol{\theta}_j} L_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n} d_{i,j}\boldsymbol{x}_i - \sum_{i=1}^{n} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\theta}_j)}{1+\sum_{k=1}^{J}\exp(\boldsymbol{x}_i'\boldsymbol{\theta}_k)}\boldsymbol{x}_i \\ &= \sum_{i=1}^{n}(d_{i,j} - G_{i,j})\boldsymbol{x}_i, \quad j = 1,\ldots,J.\end{aligned}$$

Setting these J functions to zero we can solve for the MLE $\tilde{\boldsymbol{\theta}}$ using numerical methods. The predicted probabilities for the j^{th} choice are $\widehat{G}_{i,j}$, $G_{i,j}$ evaluated at $\tilde{\boldsymbol{\theta}}$. Under suitable conditions, $\tilde{\boldsymbol{\theta}} \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_o$. We can also show that the information matrix equality holds when $G_{i,j}$ are correctly specified (we omit the proof). It follows that

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \xrightarrow{D} \mathcal{N}\big(\mathbf{0}, -\boldsymbol{H}(\boldsymbol{\theta}_o)^{-1}\big).$$

Noting that

$$\nabla_{\boldsymbol{\theta}_j} G_{i,k} = -G_{i,k} G_{i,j} \boldsymbol{x}_i, \quad k \neq j,$$
$$\nabla_{\boldsymbol{\theta}_j} G_{i,j} = G_{i,j}\left[1 - G_{i,j}\right] \boldsymbol{x}_i,$$

the Hessian matrix contains the following diagonal blocks:

$$\nabla_{\boldsymbol{\theta}_j \boldsymbol{\theta}_j'} L_n(\boldsymbol{\theta}) = -\sum_{i=1}^{n} G_{i,j}(1 - G_{i,j}) \boldsymbol{x}_i \boldsymbol{x}_i', \quad j = 1, \ldots, J,$$

and the following off-diagonal blocks:

$$\nabla_{\boldsymbol{\theta}_j \boldsymbol{\theta}_k'} L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} (G_{i,j} G_{i,k}) \boldsymbol{x}_i \boldsymbol{x}_i', \quad k \neq j,\ j, k = 1, \ldots, J.$$

Using the sample averages of these blocks we obtain a consistent estimator for $\boldsymbol{H}(\boldsymbol{\theta}_o)$. As before, the Wald statistics can be computed using the standard errors in this estimator.

Marginal Effects

The marginal effect of the change of $\boldsymbol{x}_i$ on $\widehat{G}_{i,0}$ and $\widehat{G}_{i,j}$ are, respectively,

$$\nabla_{\boldsymbol{x}_i} \widehat{G}_{i,0} = -\widehat{G}_{i,0} \sum_{k=1}^{J} \widehat{G}_{i,k} \tilde{\boldsymbol{\theta}}_k,$$

$$\nabla_{\boldsymbol{x}_i} \widehat{G}_{i,j} = \widehat{G}_{i,j} \left(\tilde{\boldsymbol{\theta}}_j - \sum_{k=1}^{J} \widehat{G}_{i,k} \tilde{\boldsymbol{\theta}}_k \right), \quad j = 1, \ldots, J.$$

Clearly, all regressors and all coefficients enter $\nabla_{\boldsymbol{x}_i} \widehat{G}_{i,0}$ and $\nabla_{\boldsymbol{x}_i} \widehat{G}_{i,j}$, so that the marginal effects change with $\boldsymbol{x}_i$. To obtain a constant marginal effect, we may evaluate $\widehat{G}_{i,0}$ and $\widehat{G}_{i,j}$ above at $\bar{\boldsymbol{x}}$ to obtain the "marginal effect at the average". We may also compute the "average marginal effect" as:

$$-\frac{1}{n} \sum_{i=1}^{n} \widehat{G}_{i,0} \left(\sum_{k=1}^{J} \widehat{G}_{i,k} \tilde{\boldsymbol{\theta}}_k \right), \quad \frac{1}{n} \sum_{i=1}^{n} \widehat{G}_{i,j} \left(\tilde{\boldsymbol{\theta}}_j - \sum_{k=1}^{J} \widehat{G}_{i,k} \tilde{\boldsymbol{\theta}}_k \right).$$