Lecture 5
Maximum Likelihood Method and Discrete Choice Models

CHUNG-MING KUAN

Department of Finance & CRETA

National Taiwan University

April 14, 2020

Lecture Outline

1. Maximum Likelihood Method
   - Maximum Likelihood Estimation
   - Asymptotic Properties
2. Discrete Choice Models
   - Binary Choice Models
   - 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 (ML) 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 ML 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:

$$\begin{aligned}
\mathbb{P}(A|p = 0.5) &= C_8^{10}(0.5)^{10} \approx 0.01, \\
\mathbb{P}(A|p = 0.7) &= C_8^{10}(0.7)^8(0.3)^2 \approx 0.233, \\
\mathbb{P}(A|p = 0.8) &= C_8^{10}(0.8)^8(0.2)^2 \approx 0.302, \\
\mathbb{P}(A|p = 0.9) &= C_8^{10}(0.9)^8(0.1)^2 \approx 0.194.
\end{aligned}$$

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

Let $\ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta})$ denote the i^{th} likelihood function based on the random sample $(y_i, \boldsymbol{x}_i')'$, where $\boldsymbol{\theta}$ is the parameter vector. The joint likelihood function is $\prod_{i=1}^n \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta})$. Maximizing this likelihood function 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 $\boldsymbol{\theta}$:

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

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

Asymptotic Properties

Consistency: Under a suitable LLN, we expect $n^{-1}\nabla L_n(\boldsymbol{\theta})$ to be close to $\mathbb{E}[\nabla \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta})]$ on the parameter space when n becomes large. In the limit, the solution to $\nabla L_n(\boldsymbol{\theta}) = \mathbf{0}$, which is $\tilde{\boldsymbol{\theta}}$, ought to be close to the solution to $\mathbb{E}[\nabla \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta})] = \mathbf{0}$, $\boldsymbol{\theta}_o$, which is the parameter of interest. This suggests that $\tilde{\boldsymbol{\theta}} \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_o$ under suitable regularity conditions.

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 $\boldsymbol{\theta}$ (see next slide). We will use this expression to derive the asymptotic distribution of $n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)$.

Specifically, the Hessian matrix is

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

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

$$\boldsymbol{H}(\boldsymbol{\theta}) := \frac{1}{n} \mathbb{E}[\nabla^2 L_n(\boldsymbol{\theta})] = \mathbb{E}[\nabla^2 \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta})].$$

Also define the covariance matrix of $n^{-1/2} \nabla L_n(\boldsymbol{\theta})$ as:

$$\boldsymbol{B}(\boldsymbol{\theta}) := \operatorname{var}(n^{-1/2} \nabla L_n(\boldsymbol{\theta})) = \operatorname{var}(\nabla \ln \ell(y_i, \boldsymbol{x}_i'; \boldsymbol{\theta}));$$

it is known as the information matrix when evaluated at $\boldsymbol{\theta}_o$ (i.e. $\boldsymbol{B}(\boldsymbol{\theta}_o)$).

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

$$\frac{1}{n}\nabla^2 L_n(\theta^\dagger)(\tilde{\theta} - \theta_o) = -\frac{1}{n}\nabla L_n(\theta_o).$$

By a suitable LLN and CLT,

$$\sqrt{n}(\tilde{\theta} - \theta_o) = -\Big[\underbrace{\frac{1}{n}\nabla^2 L_n(\theta^\dagger)}_{\xrightarrow{\mathrm{P}}\, \boldsymbol{H}(\theta_o)}\Big]^{-1} \boldsymbol{B}(\theta_o)^{1/2}\Big[\underbrace{\boldsymbol{B}(\theta_o)^{-1/2}\frac{1}{\sqrt{n}}\nabla L_n(\theta_o)}_{\xrightarrow{D}\, \mathcal{N}(\mathbf{0}, \boldsymbol{I})}\Big].$$

where $\boldsymbol{H}(\theta_o)$ is the $\boldsymbol{H}(\theta)$ matrix evaluated at θ_o. It follows that

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

where $\stackrel{d}{=}$ denotes equality in distribution.

The asymptotic covariance matrix $\boldsymbol{H}(\boldsymbol{\theta}_o)^{-1}\boldsymbol{B}(\boldsymbol{\theta}_o)\boldsymbol{H}(\boldsymbol{\theta}_o)^{-1}$ can be estimated by $\tilde{\boldsymbol{H}}^{-1}\tilde{\boldsymbol{B}}\tilde{\boldsymbol{H}}^{-1}$, 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

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

We may conduct test statistics based on this covariance matrix estimator. For the hypothesis $\theta_j = c$, the Wald statistic compares θ_j and c:

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

where EW-se$(\tilde{\theta}_j)$, the square root of the j^{th} diagonal term of $\tilde{\boldsymbol{H}}^{-1}\tilde{\boldsymbol{B}}\tilde{\boldsymbol{H}}^{-1}$, is the robust (Eicker-White) standard error. 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:

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

an example will be given later. In this case, the asymptotic covariance matrix simplifies to

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

so that

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

When the likelihood function is misspecified, the information matrix equality fails, 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 $-\tilde{\boldsymbol{H}}^{-1}$. For the multiple hypotheses $\boldsymbol{R\theta} = \boldsymbol{c}$, where $\boldsymbol{R}$ is $q \times k$, the Wald statistic is

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

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

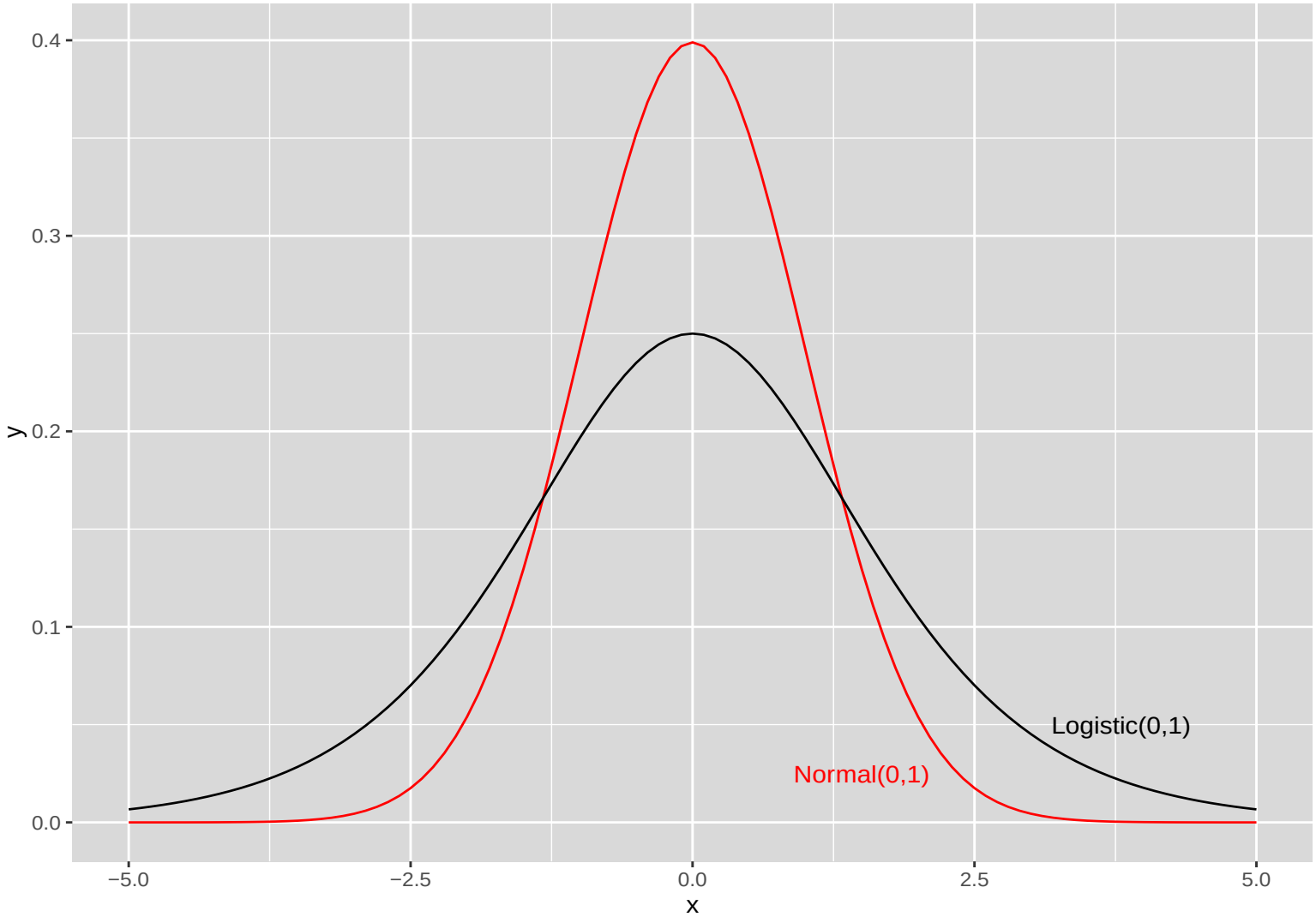

Figure: The logistic distribution vs. standard normal distribution.

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

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

The MLE $\tilde{\boldsymbol{\theta}}$ is obtained by solving

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

This is a system of k nonlinear functions and can be solved using numerical methods. The fitted values of the probit model are the predicted probabilities $\Phi(\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}})$.

For the logit model, note that $\mathrm{d}\,G(u)/\,\mathrm{d}\,u = G'(u) = G(u)[1 - G(u)]$. We thus have

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

The MLE $\tilde{\boldsymbol{\theta}}$ solves the nonlinear system: $\sum_{i=1}^{n} [y_i - G(\boldsymbol{x}_i'\boldsymbol{\theta})]\boldsymbol{x}_i = \mathbf{0}$. The fitted values of the logit model are 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. As $F(\boldsymbol{x}_i; \boldsymbol{\theta})$ is specified for $\mathbb{P}(y_i = 1|\boldsymbol{x}_i)$, we can 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 $\boldsymbol{\theta}$. The NLS estimator is, however, not the same as the MLEs discussed earlier (check), and it is not efficient because the objective function does not take into account conditional heteroskedasticity and binary feature of y.

Marginal Effect

For the probit model, the marginal effect of the j^{th} regressor is:

$$\frac{\partial \Phi(\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}})}{\partial x_{ij}} = \phi(\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}})\tilde{\theta}_j;$$

for the logit model,

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

These marginal effects are not a constant $\tilde{\theta}_j$ but the product of $\tilde{\theta}_j$ and a scaling factor which changes with $\boldsymbol{x}_i$. To circumvent the problem of changing factors, one may evaluate these 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 $= 0.42$?). Instead, one may compute the average marginal effect as:

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

We may also compute the marginal effect more directly as:

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

where $\bar{\boldsymbol{x}}(1)$ includes the sample averages of continuous regressors and 1 for the binary regressor, and $\bar{\boldsymbol{x}}(0)$ is the same as $\bar{\boldsymbol{x}}(1)$, except 1 is replaced by 0 for the binary regressor.

Model Performance

Estimating a logit/probit model leads to the predicted probabilities $G(\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}})$ or $\Phi(\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}})$. We can then make a prediction of 1 when a predicted probability is greater than the threshold value (0.5 or $\bar{y}$, the sample proportion of $y_i = 1$); otherwise, the prediction is zero. The possible outcome pairs $(y_i, \text{prediction}_i)$ are $(1,1)$, $(1,0)$, $(0,1)$, and $(0,0)$, in which $(1,1)$ and $(0,0)$ are correct predictions. Thus, we may determine the model performance using the percentage correctly predicted, i.e., the proportion of the pairs $(1,1)$ and $(0,0)$ out of n observations.

To know how well the model predicts each group, we may compute the percentage correctly predicted for $y_i = 1$ (the proportion of the number of $(1,1)$ out of the number of $y_i = 1$) and that for $y_i = 0$ (the proportion of the number of $(0,0)$ out of the number of $y_i = 0$).

Asymptotic Properties

Under suitable 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

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

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 $\mathbb{P}(y_i|\boldsymbol{x}_i)$,

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

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: $G(\boldsymbol{x}_i'\boldsymbol{\theta}_o)[1 - G(\boldsymbol{x}_i'\boldsymbol{\theta}_o)]$. It follows that the information matrix equality holds: $\boldsymbol{B}(\boldsymbol{\theta}_o) + \boldsymbol{H}(\boldsymbol{\theta}_o) = \boldsymbol{0}$, and hence

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

$\boldsymbol{H}(\boldsymbol{\theta}_o)$ can be consistently estimated by its sample counterpart:

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

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 $-\tilde{\boldsymbol{H}}^{-1}$. 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

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

Linear Probability Model

The linear probability model is the linear regression for binary y:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i,$$

which can be estimated using the OLS method. Yet, this model has the following drawbacks.

- The linear probability model does not take into account the binary feature of y_i: bounded conditional mean (between zero and one) and conditional heteroskedasticity.
- The fitted values of a linear probability model may not be between zero and one and hence cannot be interpreted as probabilities.

Example: Labor Force Participation

Wooldridge (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 QMLE $\tilde{\boldsymbol{\theta}}$.

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

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

that is, the parameters are not identified.

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 sets of functions to zero we can solve for the QMLE $\tilde{\boldsymbol{\theta}}$ 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 Effect

The marginal effect of $\widehat{G}_{i,0}$ and $\widehat{G}_{i,j}$ to the change of $\boldsymbol{x}_i$ 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 compute $\widehat{G}_{i,0}$ 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":

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