11. **Optimization (最適化):**

MLE of θ results in the following maximization problem:

$$\max_{\theta} \quad \log L(\theta; x).$$

We often have the case where the solution of θ is not derived in closed form.

$\Longrightarrow$ Optimization procedure

$$0 = \frac{\partial \log L(\theta; x)}{\partial \theta} = \frac{\partial \log L(\theta^*; x)}{\partial \theta} + \frac{\partial^2 \log L(\theta^*; x)}{\partial \theta \partial \theta'}(\theta - \theta^*).$$

Solving the above equation with respect to θ, we obtain the following:

$$\theta = \theta^* - \left(\frac{\partial^2 \log L(\theta^*; x)}{\partial \theta \partial \theta'} \right)^{-1} \frac{\partial \log L(\theta^*; x)}{\partial \theta}.$$

Replace the variables as follows:

$$\theta \longrightarrow \theta^{(i+1)}, \qquad \theta^* \longrightarrow \theta^{(i)}.$$

Then, we have:

$$\theta^{(i+1)} = \theta^{(i)} - \left(\frac{\partial^2 \log L(\theta^{(i)}; x)}{\partial\theta\partial\theta'}\right)^{-1} \frac{\partial \log L(\theta^{(i)}; x)}{\partial\theta}.$$

$\Longrightarrow$ **Newton-Raphson method (ニュートン・ラプソン法)**

Replacing $\dfrac{\partial^2 \log L(\theta^{(i)}; x)}{\partial\theta\partial\theta'}$ by $\mathrm{E}\left(\dfrac{\partial^2 \log L(\theta^{(i)}; x)}{\partial\theta\partial\theta'}\right)$, we obtain the following optimization algorithm:

$$\begin{aligned}
\theta^{(i+1)} &= \theta^{(i)} - \left(\mathrm{E}\left(\frac{\partial^2 \log L(\theta^{(i)}; x)}{\partial\theta\partial\theta'}\right)\right)^{-1} \frac{\partial \log L(\theta^{(i)}; x)}{\partial\theta} \\
&= \theta^{(i)} + \left(I(\theta^{(i)})\right)^{-1} \frac{\partial \log L(\theta^{(i)}; x)}{\partial\theta}
\end{aligned}$$

$\Longrightarrow$ **Method of Scoring (スコア法)**

2 Qualitative Dependent Variable (質的従属変数)

1. **Discrete Choice Model (離散選択モデル)**

2. **Limited Dependent Variable Model (制限従属変数モデル)**

3. **Count Data Model (計数データモデル)**

Usually, the regression model is given by:

$$y_i = X_i\beta + u_i, \qquad u_i \sim N(0, \sigma^2), \qquad i = 1, 2, \cdots, n,$$

where y_i is a continuous type of random variable within the interval from $-\infty$ to ∞.

When y_i is discrete or truncated, what happens?

2.1 Discrete Choice Model (離散選択モデル)

2.1.1 Binary Choice Model (二値選択モデル)

Example 1: Consider the regression model:

$$y_i^* = X_i\beta + u_i, \qquad u_i \sim (0, \sigma^2), \qquad i = 1, 2, \cdots, n,$$

where y_i^* is unobserved, but y_i is observed as 0 or 1, i.e.,

$$y_i = \begin{cases} 1, & \text{if } y_i^* > 0, \\ 0, & \text{if } y_i^* \le 0. \end{cases}$$

Consider the probability that y_i takes 1, i.e.,

$$\begin{aligned} P(y_i = 1) &= P(y_i^* > 0) = P(u_i > -X_i\beta) = P(u_i^* > -X_i\beta^*) = 1 - P(u_i^* \le -X_i\beta^*) \\ &= 1 - F(-X_i\beta^*) = F(X_i\beta^*), \quad \text{(if the dist. of } u_i^* \text{ is symmetric.)}, \end{aligned}$$

where $u_i^* = \dfrac{u_i}{\sigma}$, and $\beta^* = \dfrac{\beta}{\sigma}$ are defined.

(*) β^* can be estimated, but β and σ^2 cannot be estimated separately (i.e., β and σ^2 are not identified).

The distribution function of u_i^* is given by $F(x) = \displaystyle\int_{-\infty}^{x} f(z)\mathrm{d}z$.

If u_i^* is standard normal, i.e., $u_i^* \sim N(0, 1)$, we call **probit model**.

$$F(x) = \int_{-\infty}^{x} (2\pi)^{-1/2} \exp(-\frac{1}{2}z^2)\mathrm{d}z, \qquad f(x) = (2\pi)^{-1/2} \exp(-\frac{1}{2}x^2).$$

If u_i^* is logistic, we call **logit model**.

$$F(x) = \frac{1}{1 + \exp(-x)}, \qquad f(x) = \frac{\exp(-x)}{(1 + \exp(-x))^2}.$$

We can consider the other distribution function for u_i^*.

Likelihood Function: y_i is the following Bernoulli distribution:

$$f(y_i) = (P(y_i = 1))^{y_i}(P(y_i = 0))^{1-y_i} = (F(X_i\beta^*))^{y_i}(1 - F(X_i\beta^*))^{1-y_i}, \qquad y_i = 0, 1.$$

[Review — Bernoulli Distribution (ベルヌイ分布)]

Suppose that X is a Bernoulli random variable. the distribution of X, denoted by $f(x)$, is:

$$f(x) = p^x(1 - p)^{1-x}, \qquad x = 0, 1.$$

The mean and variance are:

$$\mu = \mathrm{E}(X) = \sum_{x=0}^{1} xf(x) = 0 \times (1 - p) + 1 \times p = p,$$

$$\sigma^2 = \mathrm{V}(X) = \mathrm{E}((X - \mu)^2) = \sum_{x=0}^{1}(x - \mu)^2 f(x) = (0 - p)^2(1 - p) + (1 - p)^2 p = p(1 - p).$$

[End of Review]

The likelihood function is given by:

$$L(\beta^*) = f(y_1, y_2, \cdots, y_n) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} (F(X_i\beta^*))^{y_i}(1 - F(X_i\beta^*))^{1-y_i},$$

The log-likelihood function is:

$$\log L(\beta^*) = \sum_{i=1}^{n} \Big(y_i \log F(X_i\beta^*) + (1 - y_i)\log(1 - F(X_i\beta^*)) \Big),$$

Solving the maximization problem of $\log L(\beta^*)$ with respect to β^*, the first order condition is:

$$\begin{aligned}
\frac{\partial \log L(\beta^*)}{\partial \beta^*} &= \sum_{i=1}^{n} \Big(\frac{y_i X_i' f(X_i\beta^*)}{F(X_i\beta^*)} - \frac{(1 - y_i) X_i' f(X_i\beta^*)}{1 - F(X_i\beta^*)} \Big) \\
&= \sum_{i=1}^{n} \frac{X_i' f(X_i\beta^*)(y_i - F(X_i\beta^*))}{F(X_i\beta^*)(1 - F(X_i\beta^*))} = \sum_{i=1}^{n} \frac{X_i' f_i (y_i - F_i)}{F_i(1 - F_i)} = 0,
\end{aligned}$$

where $f_i \equiv f(X_i\beta^*)$ and $F_i \equiv F(X_i\beta^*)$. Remember that $f(x) \equiv \dfrac{\mathrm{d}F(x)}{\mathrm{d}x}$.

The second order condition is:

$$\frac{\partial^2 \log L(\beta^*)}{\partial\beta^*\partial\beta^{*\prime}} = \sum_{i=1}^n \frac{X_i' \dfrac{\partial f_i}{\partial\beta^*}(y_i - F_i)}{F_i(1-F_i)} + \sum_{i=1}^n \frac{X_i' f_i \dfrac{\partial(f_i - F_i)}{\partial\beta^*}}{F_i(1-F_i)}$$
$$+ \sum_{i=1}^n X_i' f_i (y_i - F_i) \frac{\partial (F_i(1-F_i))^{-1}}{\partial\beta^*}$$
$$= \sum_{i=1}^n \frac{X_i' X_i f_i' (y_i - F_i)}{F_i(1-F_i)} - \sum_{i=1}^n \frac{X_i' X_i f_i^2}{F_i(1-F_i)} + \sum_{i=1}^n X_i' f_i (y_i - F_i) \frac{X_i f_i (1-2F_i)}{(F_i(1-F_i))^2}$$

is a negative definite matrix.

For maximization, the method of scoring is given by:

$$\beta^{*(j+1)} = \beta^{*(j)} + \left(-\mathrm{E}\Big(\frac{\partial^2 \log L(\beta^{*(j)})}{\partial\beta^*\partial\beta^{*\prime}}\Big)\right)^{-1} \frac{\partial \log L(\beta^{*(j)})}{\partial\beta^*}$$
$$= \beta^{*(j)} + \left(\sum_{i=1}^n \frac{X_i' X_i (f_i^{(j)})^2}{F_i^{(j)}(1-F_i^{(j)})}\right)^{-1} \sum_{i=1}^n \frac{X_i' f_i^{(j)} (y_i - F_i^{(j)})}{F_i^{(j)}(1-F_i^{(j)})},$$

where $F_i^{(j)} = F(X_i\beta^{*(j)})$ and $f_i^{(j)} = f(X_i\beta^{*(j)})$. Note that

$$I(\beta^*) = -\mathrm{E}\Big(\frac{\partial^2 \log L(\beta^*)}{\partial\beta^*\partial\beta^{*\prime}}\Big) = \sum_{i=1}^{n} \frac{X_i' X_i f_i^2}{F_i(1-F_i)}.$$

because of $\mathrm{E}(y_i) = F_i$.

It is known that

$$\sqrt{n}(\hat{\beta}^* - \beta^*) \longrightarrow N\left(0,\ \lim_{n\to\infty}\left(-\frac{1}{n}\mathrm{E}\Big(\frac{\partial^2 \log L(\beta^*)}{\partial\beta^*\partial\beta^{*\prime}}\Big)\right)^{-1}\right),$$

where $\hat{\beta}^* \equiv \lim_{j\to\infty} \beta^{*(j)}$ denotes MLE of β^*.

Practically, we use the following normal distribution:

$$\hat{\beta}^* \sim N\Big(\beta^*,\ I(\hat{\beta}^*)^{-1}\Big),$$

where $I(\hat{\beta}^*) = -\mathrm{E}\left(\dfrac{\partial^2 \log L(\hat{\beta}^*)}{\partial\beta^*\partial\beta^{*\prime}}\right) = \displaystyle\sum_{i=1}^{n} \frac{X_i' X_i \hat{f}_i^2}{\hat{F}_i(1-\hat{F}_i)}$, $\hat{f}_i = f(X_i\hat{\beta}^*)$ and $\hat{F}_i = F(X_i\hat{\beta}^*)$.

Thus, the significance test for β^* and the confidence interval for β^* can be constructed.