and $\Big(\frac{1}{n}\frac{\partial^2 \log L(\theta; X)}{\partial\theta\partial\theta'}\Big)^{-1}\Big(\frac{1}{\sqrt{n}}\frac{\partial \log L(\theta; X)}{\partial\theta}\Big)$ has the same asymptotic distribution as $\Sigma^{-1}\Big(\frac{1}{\sqrt{n}}\frac{\partial \log L(\theta; X)}{\partial\theta}\Big)$.

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.,

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

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

$$P(y_i = 1) = P(y_i^* > 0) = P(u_i > -X_i\beta) = P(u_i^* > -X_i\beta^*) = 1 - P(u_i^* \leq -X_i\beta^*)$$

$$= 1 - F(-X_i\beta^*) = F(X_i\beta^*), \quad \text{(if the dist. of } u_i^* \text{ is symmetric.)},$$

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:

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

is a negative definite matrix.

For maximization, the method of scoring is given by:

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

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.

Another Interpretation: This maximization problem is equivalent to the nonlinear least squares estimation problem from the following regression model:

$$y_i = F(X_i\beta^*) + u_i,$$

where $u_i = y_i - F_i$ takes $u_i = 1 - F_i$ with probability $P(y_i = 1) = F(X_i\beta^*) = F_i$ and $u_i = -F_i$ with probability $P(y_i = 0) = 1 - F(X_i\beta^*) = 1 - F_i$.

Therefore, the mean and variance of u_i are:

$$\mathrm{E}(u_i) = (1 - F_i)F_i + (-F_i)(1 - F_i) = 0,$$

$$\sigma_i^2 = \mathrm{V}(u_i) = \mathrm{E}(u_i^2) - (\mathrm{E}(u_i))^2 = (1 - F_i)^2 F_i + (-F_i)^2(1 - F_i) = F_i(1 - F_i).$$

The weighted least squares method solves the following minimization problem:

$$\min_{\beta^*} \sum_{i=1}^{n} \frac{(y_i - F(X_i\beta^*))^2}{\sigma_i^2}.$$

The first order condition is:

$$\sum_{i=1}^{n} \frac{X_i' f(X_i\beta^*)(y_i - F(X_i\beta^*))}{\sigma_i^2} = \sum_{i=1}^{n} \frac{X_i' f_i(y_i - F_i)}{F_i(1 - F_i)} = 0,$$

which is equivalent to the first order condition of MLE.

Thus, the binary choice model is interpreted as the nonlinear least squares.

Prediction: $\mathrm{E}(y_i) = 0 \times (1 - F_i) + 1 \times F_i = F_i \equiv F(X_i\beta^*)$.

Example 2: Consider the two utility functions: $U_{1i} = X_i\beta_1 + \epsilon_{1i}$ and $U_{2i} = X_i\beta_2 + \epsilon_{2i}$. A linear utility function is problematic, but we consider the linear function for simplicity of discussion.

We purchase a good when $U_{1i} > U_{2i}$ and do not purchase it when $U_{1i} < U_{2i}$.

We can observe $y_i = 1$ when we purchase the good, i.e., when $U_{1i} > U_{2i}$, and $y_i = 0$ otherwise.

$$\begin{aligned} P(y_i = 1) &= P(U_{1i} > U_{2i}) = P(X_i(\beta_1 - \beta_2) > -\epsilon_{1i} + \epsilon_{2i}) \\ &= P(-X_i\beta^* < \epsilon_i^*) = P(-X_i\beta^{**} < \epsilon_i^{**}) = 1 - F(-X_i\beta^{**}) = F(X_i\beta^{**}) \end{aligned}$$

where $\beta^* = \beta_1 - \beta_2, \quad \epsilon_i^* = \epsilon_{1i} - \epsilon_{2i}, \quad \beta^{**} = \dfrac{\beta^*}{\sigma^*}$ and $\epsilon_i^{**} = \dfrac{\epsilon_i^*}{\sigma^*}$.

We can estimate β^{**}, but we cannot estimate ϵ_i^* and σ^*, separately.

Mean and variance of ϵ_i^{**} are normalized to be zero and one, respectively.

If the distribution of ϵ_i^{**} is symmetric, the last equality holds.

We can estimate β^{**} by MLE as in Example 1.

Example 3: Consider the questionnaire:

$$y_i = \begin{cases} 1, & \text{if the } i\text{th person answers YES,} \\ 0, & \text{if the } i\text{th person answers NO.} \end{cases}$$

Consider estimating the following linear regression model:

$$y_i = X_i\beta + u_i.$$

When $\mathrm{E}(u_i) = 0$, the expectation of y_i is given by:

$$\mathrm{E}(y_i) = X_i\beta.$$

Because of the linear function, $X_i\beta$ takes the value from $-\infty$ to ∞.

However, $\mathrm{E}(y_i)$ indicates the ratio of the people who answer YES out of all the people, because of $\mathrm{E}(y_i) = 1 \times P(y_i = 1) + 0 \times P(y_i = 0) = P(y_i = 1)$.

That is, $\mathrm{E}(y_i)$ has to be between zero and one.

Therefore, it is not appropriate that $\mathrm{E}(y_i)$ is approximated as $X_i\beta$.

The model is written as:

$$y_i = P(y_i = 1) + u_i,$$

where u_i is a discrete type of random variable, i.e., u_i takes $1 - P(y_i = 1)$ with probability $P(y_i = 1)$ and $-P(y_i = 1)$ with probability $1 - P(y_i = 1) = P(y_i = 0)$.

Consider that $P(y_i = 1)$ is connected with the distribution function $F(X_i\beta)$ as follows:

$$P(y_i = 1) = F(X_i\beta),$$

where $F(\cdot)$ denotes a distribution function such as normal dist., logistic dist., and so on. $\longrightarrow$ probit model or logit model.

The probability function of y_i is:

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

The joint distribution of $y_1, y_2, \cdots, y_n$ is:

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

which corresponds to the likelihood function. $\longrightarrow$ MLE

Example 4: Ordered probit or logit model:

Consider the regression model:

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

where y_i^* is unobserved, but y_i is observed as $1, 2, \cdots, m$, i.e.,

$$y_i = \begin{cases} 1, & \text{if } -\infty < y_i^* \le a_1, \\ 2, & \text{if } a_1 < y_i^* \le a_2, \\ \vdots, & \\ m, & \text{if } a_{m-1} < y_i^* < \infty, \end{cases}$$

where a_1, a_2, $\cdots$, a_{m-1} are assumed to be known.

Consider the probability that y_i takes $1, 2, \cdots, m$, i.e.,

$$
\begin{aligned}
P(y_i = 1) &= P(y_i^* \le a_1) = P(u_i \le a_1 - X_i\beta) \\
&= F(a_1 - X_i\beta), \\
P(y_i = 2) &= P(a_1 < y_i^* \le a_2) = P(a_1 - X_i\beta < u_i \le a_2 - X_i\beta) \\
&= F(a_2 - X_i\beta) - F(a_1 - X_i\beta), \\
P(y_i = 3) &= P(a_2 < y_i^* \le a_3) = P(a_2 - X_i\beta < u_i \le a_3 - X_i\beta) \\
&= F(a_3 - X_i\beta) - F(a_2 - X_i\beta), \\
&\vdots \\
P(y_i = m) &= P(a_{m-1} < y_i^*) = P(a_{m-1} - X_i\beta < u_i) \\
&= 1 - F(a_{m-1} - X_i\beta).
\end{aligned}
$$

Define the following indicator functions:

$$I_{i1} = \begin{cases} 1, & \text{if } y_i = 1, \\ 0, & \text{otherwise.} \end{cases} \qquad I_{i2} = \begin{cases} 1, & \text{if } y_i = 2, \\ 0, & \text{otherwise.} \end{cases} \qquad \cdots \qquad I_{im} = \begin{cases} 1, & \text{if } y_i = m, \\ 0, & \text{otherwise.} \end{cases}$$

More compactly,

$$P(y_i = j) = F(a_j - X_i\beta) - F(a_{j-1} - X_i\beta),$$

for $j = 1, 2, \cdots, m$, where $a_0 = -\infty$ and $a_m = \infty$.

$$I_{ij} = \begin{cases} 1, & \text{if } y_i = j, \\ 0, & \text{otherwise,} \end{cases}$$

for $j = 1, 2, \cdots, m$.

Then, the likelihood function is:

$$L(\beta) = \prod_{i=1}^{n} \Big(F(a_1 - X_i\beta)\Big)^{I_{i1}} \Big(F(a_2 - X_i\beta) - F(a_1 - X_i\beta)\Big)^{I_{i2}} \cdots \Big(1 - F(a_{m-1} - X_i\beta)\Big)^{I_{im}}$$
$$= \prod_{i=1}^{n} \prod_{j=1}^{m} \Big(F(a_j - X_i\beta) - F(a_{j-1} - X_i\beta)\Big)^{I_{ij}},$$

where $a_0 = -\infty$ and $a_m = \infty$. Remember that $F(-\infty) = 0$ and $F(\infty) = 1$.

The log-likelihood function is:

$$\log L(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{m} I_{ij} \log\Big(F(a_j - X_i\beta) - F(a_{j-1} - X_i\beta)\Big).$$

The first derivative of $\log L(\beta)$ with respect to β is:

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{-I_{ij} X_i' \Big(f(a_j - X_i\beta) - f(a_{j-1} - X_i\beta)\Big)}{F(a_j - X_i\beta) - F(a_{j-1} - X_i\beta)} = 0.$$

Usually, normal distribution or logistic distribution is chosen for $F(\cdot)$.

Example 5: Multinomial logit model:

The ith individual has $m+1$ choices, i.e., $j = 0, 1, \cdots, m$.

$$P(y_i = j) = \frac{\exp(X_i\beta_j)}{\sum_{j=0}^{m} \exp(X_i\beta_j)} \equiv P_{ij},$$

for $\beta_0 = 0$. The case of $m = 1$ corresponds to the bivariate logit model (binary choice).

Note that

$$\log \frac{P_{ij}}{P_{i0}} = X_i\beta_j$$

The log-likelihood function is:

$$\log L(\beta_1, \cdots, \beta_m) = \sum_{i=1}^{n} \sum_{j=0}^{m} d_{ij} \ln P_{ij},$$

where $d_{ij} = 1$ when the ith individual chooses jth choice, and $d_{ij} = 0$ otherwise.