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 $P(y_i)$ 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.