Notes on Binary Outcome Model

Max Leung

January 16, 2022

1 General Binary Outcome Model

$$y_i = \begin{cases} 1 & \text{with probability } p_i \\ 0 & \text{with probability } 1 - p_i \end{cases}$$

Model p_i as

$$p_i := Pr(y_i = 1|\boldsymbol{x}_i) = F(\boldsymbol{x}_i'\boldsymbol{\beta}) = F_i$$

1.1 Marginal Effect

$$\begin{aligned} \frac{\partial Pr(y_i = 1|\boldsymbol{x}_i)}{\partial \boldsymbol{x}_i} &= \frac{\partial F(\boldsymbol{x}_i'\boldsymbol{\beta})}{\partial \boldsymbol{x}_i} \\ &= \frac{\partial F(\boldsymbol{x}_i'\boldsymbol{\beta})}{\partial \boldsymbol{x}_i'\boldsymbol{\beta}} \frac{\partial \boldsymbol{x}_i'\boldsymbol{\beta}}{\partial \boldsymbol{x}_i} \\ &= F'(\boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{\beta} && \text{If } F(.) \text{ is cdf, } F'(.) > 0. \text{ So, } sign(\boldsymbol{\beta}) \text{ decide the sign} \end{aligned}$$

Average Marginal Effect

$$AME := N^{-1} \sum_{i=1}^{N} F'(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}}$$

or

$$AME := F'(N^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}} = F'(\bar{\boldsymbol{x}}'\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}}$$

1.2 ML Estimation

As y_i is binary, it must be Bernoulli distributed. The probability mass function (pmf) of such random variable is

$$\begin{aligned} f(y_i|\boldsymbol{x}_i) &= Pr(y_i = 1|\boldsymbol{x}_i)^{y_i} Pr(y_i = 0|\boldsymbol{x}_i)^{1-y_i} \\ &= Pr(y_i = 1|\boldsymbol{x}_i)^{y_i} (1 - Pr(y_i = 1|\boldsymbol{x}_i))^{1-y_i} \\ &= p_i^{y_i} (1 - p_i)^{1-y_i} \\ &= F(\boldsymbol{x}_i'\boldsymbol{\beta})^{y_i} (1 - F(\boldsymbol{x}_i'\boldsymbol{\beta}))^{1-y_i} \end{aligned}$$

Log Likelihood Function is

$$\begin{aligned} ln[L_N(\boldsymbol{\beta})] &= ln[\Pi_{i=1}^{N} f(y_i|\boldsymbol{x}_i)] && \text{assume independence} \\ &= ln[\Pi_{i=1}^{N} F(\boldsymbol{x}_i'\boldsymbol{\beta})^{y_i} (1 - F(\boldsymbol{x}_i'\boldsymbol{\beta}))^{1-y_i}] \\ &= \sum_{i=1}^{N} ln[F(\boldsymbol{x}_i'\boldsymbol{\beta})^{y_i} (1 - F(\boldsymbol{x}_i'\boldsymbol{\beta}))^{1-y_i}] \\ &= \sum_{i=1}^{N} \{y_i ln[F(\boldsymbol{x}_i'\boldsymbol{\beta})] + (1 - y_i) ln[1 - F(\boldsymbol{x}_i'\boldsymbol{\beta})]\} \end{aligned}$$

Gradient Vector

$$
\begin{aligned}
\frac{\partial ln[L_N(\boldsymbol{\beta})]}{\partial \boldsymbol{\beta}} &= \frac{\partial \sum_{i=1}^{N}\{y_i ln[F(\boldsymbol{x}_i'\boldsymbol{\beta})] + (1-y_i)ln[1-F(\boldsymbol{x}_i'\boldsymbol{\beta})]\}}{\partial \boldsymbol{\beta}} \\
&= \sum_{i=1}^{N} \frac{\partial\{y_i ln[F(\boldsymbol{x}_i'\boldsymbol{\beta})] + (1-y_i)ln[1-F(\boldsymbol{x}_i'\boldsymbol{\beta})]\}}{\partial \boldsymbol{\beta}} \\
&= \sum_{i=1}^{N} \{\frac{\partial y_i ln[F(\boldsymbol{x}_i'\boldsymbol{\beta})]}{\partial \boldsymbol{\beta}} + \frac{\partial (1-y_i)ln[1-F(\boldsymbol{x}_i'\boldsymbol{\beta})]}{\partial \boldsymbol{\beta}}\} \\
&= \sum_{i=1}^{N} \{y_i \frac{1}{F(\boldsymbol{x}_i'\boldsymbol{\beta})} F'(\boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i + (1-y_i)\frac{1}{1-F(\boldsymbol{x}_i'\boldsymbol{\beta})}(-1)F'(\boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i\} \\
&= \sum_{i=1}^{N} \{\frac{y_i}{F(\boldsymbol{x}_i'\boldsymbol{\beta})} F'(\boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i - \frac{1-y_i}{1-F(\boldsymbol{x}_i'\boldsymbol{\beta})}F'(\boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i\} \\
&= \sum_{i=1}^{N} \{\frac{y_i}{F_i} F_i'\boldsymbol{x}_i - \frac{1-y_i}{1-F_i}F_i'\boldsymbol{x}_i\} \\
&= \sum_{i=1}^{N} \frac{y_i F_i'\boldsymbol{x}_i(1-F_i) - (1-y_i)F_i'\boldsymbol{x}_i F_i}{F_i(1-F_i)} \\
&= \sum_{i=1}^{N} \frac{(y_i F_i'\boldsymbol{x}_i - y_i F_i'\boldsymbol{x}_i F_i) - (F_i'\boldsymbol{x}_i F_i - y_i F_i'\boldsymbol{x}_i F_i)}{F_i(1-F_i)} \\
&= \sum_{i=1}^{N} \frac{y_i F_i'\boldsymbol{x}_i - F_i'\boldsymbol{x}_i F_i}{F_i(1-F_i)} \\
&= \sum_{i=1}^{N} \frac{y_i - F_i}{F_i(1-F_i)} F_i'\boldsymbol{x}_i
\end{aligned}
$$

FOC

$$
\begin{aligned}
\sum_{i=1}^{N} \frac{y_i - F(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})}{F(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})(1-F(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}))} F'(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})\boldsymbol{x}_i &= \boldsymbol{0} \\
\sum_{i=1}^{N} \underbrace{\{\frac{F'(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})}{F(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})(1-F(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}))}\}}_{\widehat{w}_i}[y_i - F(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})]\boldsymbol{x}_i &= \boldsymbol{0}
\end{aligned}
$$

There is no closed form solution. Thus, we solve it by using Gradient Descent Method or Newton Method. As $ln[L_N(\boldsymbol{\beta})]$ is globally concave for some specifications of F, the convergence is fast.

1.3 Consistency of MLE

If the likelihood function is correctly specified (with other conditions), ML estimator is consistent. y_i must be Bernoulli distributed here. If $p_i := Pr(y_i = 1|\boldsymbol{x}_i) = F(\boldsymbol{x}_i'\boldsymbol{\beta})$ is also correctly specified i.e., $F(.)$ is correctly specified, then $\widehat{\boldsymbol{\beta}} \rightarrow_p \boldsymbol{\beta}$ as $N \rightarrow \infty$.

1.4 Asymptotic Distribution of MLE

Under some regularity conditions, Information Matrix Inequality holds. And under other conditions, ML estimator is asymptotically normally distributed.

$$
\begin{aligned}
\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &\rightarrow_d N(\mathbf{0}, [\boldsymbol{I}(\boldsymbol{\beta}_0)]^{-1}) \\
&= N(\mathbf{0}, [\sum_{i=1}^{N} \frac{1}{F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)(1-F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))} F'(\boldsymbol{x}_i'\boldsymbol{\beta}_0)^2 \boldsymbol{x}_i\boldsymbol{x}_i']^{-1}) \\
&= N(\mathbf{0}, [\sum_{i=1}^{N} \frac{1}{Var(y_i|\boldsymbol{x}_i)} F'(\boldsymbol{x}_i'\boldsymbol{\beta}_0)^2 \boldsymbol{x}_i\boldsymbol{x}_i']^{-1})
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{I}(\boldsymbol{\beta}_0) &:= -\mathbb{E}[\frac{\partial^2 lnL_N(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}|_{\boldsymbol{\beta}_0}|\boldsymbol{x}_i] = \mathbb{E}[\frac{\partial lnL_N(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} \cdot \frac{\partial lnL_N(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}'}|_{\boldsymbol{\beta}_0}|\boldsymbol{x}_i] \\
&= -\mathbb{E}[\frac{\partial \sum_{i=1}^{N}\{\frac{y_i}{F_i}F_i'\boldsymbol{x}_i - \frac{1-y_i}{1-F_i}F_i'\boldsymbol{x}_i\}}{\partial\boldsymbol{\beta}'}|_{\boldsymbol{\beta}_0}|\boldsymbol{x}_i] \\
&= -\sum_{i=1}^{N}\mathbb{E}[\frac{\partial(\frac{y_i}{F_i} - \frac{1-y_i}{1-F_i})F_i'\boldsymbol{x}_i}{\partial\boldsymbol{\beta}'}|_{\boldsymbol{\beta}_0}|\boldsymbol{x}_i] \\
&= -\sum_{i=1}^{N}\mathbb{E}[\frac{\partial(\frac{y_i}{F_i} - \frac{1-y_i}{1-F_i})}{\partial\boldsymbol{\beta}'}F_i'\boldsymbol{x}_i|_{\boldsymbol{\beta}_0} + (\frac{y_i}{F_i} - \frac{1-y_i}{1-F_i})\frac{\partial F_i'\boldsymbol{x}_i}{\partial\boldsymbol{\beta}'}|_{\boldsymbol{\beta}_0}|\boldsymbol{x}_i] \\
&= -\sum_{i=1}^{N}\mathbb{E}[(\frac{\partial\frac{y_i}{F_i}}{\partial\boldsymbol{\beta}'} - \frac{\partial\frac{1-y_i}{1-F_i}}{\partial\boldsymbol{\beta}'})F_i'\boldsymbol{x}_i + (\frac{y_i}{F_i} - \frac{1-y_i}{1-F_i})\frac{\partial F_i'\boldsymbol{x}_i}{\partial\boldsymbol{\beta}'}|\boldsymbol{x}_i]|_{\boldsymbol{\beta}_0} \\
&= -\sum_{i=1}^{N}\mathbb{E}[(-y_iF_i^{-2}F_i'\boldsymbol{x}_i' - (1-y_i)(1-F_i)^{-2}F_i'\boldsymbol{x}_i')F_i'\boldsymbol{x}_i + (\frac{y_i}{F_i} - \frac{1-y_i}{1-F_i})F_i^{''}\boldsymbol{x}_i\boldsymbol{x}_i'|\boldsymbol{x}_i]|_{\boldsymbol{\beta}_0} \\
&= \sum_{i=1}^{N}\{\mathbb{E}[(y_iF_i^{-2}F_i'\boldsymbol{x}_i' + (1-y_i)(1-F_i)^{-2}F_i'\boldsymbol{x}_i')F_i'\boldsymbol{x}_i|\boldsymbol{x}_i]|_{\boldsymbol{\beta}_0} + \mathbb{E}[(\frac{y_i}{F_i} - \frac{1-y_i}{1-F_i})F_i^{''}\boldsymbol{x}_i\boldsymbol{x}_i'|\boldsymbol{x}_i]|_{\boldsymbol{\beta}_0}\} \\
&= \sum_{i=1}^{N}\{\mathbb{E}[(y_iF_i^{-2} + (1-y_i)(1-F_i)^{-2})F_i'^2\boldsymbol{x}_i\boldsymbol{x}_i'|\boldsymbol{x}_i]|_{\boldsymbol{\beta}_0} + \mathbb{E}[(\frac{y_i}{F_i} - \frac{1-y_i}{1-F_i})|\boldsymbol{x}_i]F_i^{''}\boldsymbol{x}_i\boldsymbol{x}_i'|_{\boldsymbol{\beta}_0}\} \\
&= \sum_{i=1}^{N}\{(\mathbb{E}[y_i|\boldsymbol{x}_i]F_i^{-2} + (1-\mathbb{E}[y_i|\boldsymbol{x}_i])(1-F_i)^{-2})F_i'^2\boldsymbol{x}_i\boldsymbol{x}_i'|_{\boldsymbol{\beta}_0} + (\frac{\mathbb{E}[y_i|\boldsymbol{x}_i]}{F_i} - \frac{1-\mathbb{E}[y_i|\boldsymbol{x}_i]}{1-F_i})F_i^{''}\boldsymbol{x}_i\boldsymbol{x}_i'|_{\boldsymbol{\beta}_0}\} \\
&= \sum_{i=1}^{N}\{(F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)^{-2} + (1-F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))(1-F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))^{-2})F'(\boldsymbol{x}_i'\boldsymbol{\beta}_0)^2\boldsymbol{x}_i\boldsymbol{x}_i' + \\
&\quad (\frac{F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)}{F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)} - \frac{1-F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)}{1-F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)})F^{''}(\boldsymbol{x}_i'\boldsymbol{\beta}_0)\boldsymbol{x}_i\boldsymbol{x}_i'\} \\
&= \sum_{i=1}^{N}(\frac{1}{F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)} + \frac{1}{1-F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)})F'(\boldsymbol{x}_i'\boldsymbol{\beta}_0)^2\boldsymbol{x}_i\boldsymbol{x}_i' \\
&= \sum_{i=1}^{N}\frac{(1-F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)) + F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)}{F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)(1-F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))}F'(\boldsymbol{x}_i'\boldsymbol{\beta}_0)^2\boldsymbol{x}_i\boldsymbol{x}_i' \\
&= \sum_{i=1}^{N}\frac{1}{F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)(1-F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))}F'(\boldsymbol{x}_i'\boldsymbol{\beta}_0)^2\boldsymbol{x}_i\boldsymbol{x}_i'
\end{aligned}
$$

As $\mathbb{E}(y_i|\boldsymbol{x}_i) = 1 \cdot Pr(y_i = 1|\boldsymbol{x}_i) + 0 \cdot Pr(y_i = 0|\boldsymbol{x}_i) = Pr(y_i = 1|\boldsymbol{x}_i) = F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)$
$Var(y_i|\boldsymbol{x}_i) = \mathbb{E}[(y_i - \mathbb{E}(y_i|\boldsymbol{x}_i))^2|\boldsymbol{x}_i] = (1-\mathbb{E}(y_i|\boldsymbol{x}_i))^2Pr(y_i = 1|\boldsymbol{x}_i) + (0-\mathbb{E}(y_i|\boldsymbol{x}_i))^2Pr(y_i = 0|\boldsymbol{x}_i) = (1-Pr(y_i = 1|\boldsymbol{x}_i))^2Pr(y_i = 1|\boldsymbol{x}_i) + Pr(y_i = 1|\boldsymbol{x}_i)^2(1-Pr(y_i = 1|\boldsymbol{x}_i)) = (1-Pr(y_i = 1|\boldsymbol{x}_i))Pr(y_i = 1|\boldsymbol{x}_i)(1-Pr(y_i = 1|\boldsymbol{x}_i) + Pr(y_i = 1|\boldsymbol{x}_i)) = Pr(y_i = 1|\boldsymbol{x}_i)(1-Pr(y_i = 1|\boldsymbol{x}_i)) = F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)(1-F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))$

For binary outcome model, even the regularity conditions for Information matrix Inequality does not hold, the Information matrix equality still holds. i.e. $\boldsymbol{A} = -\boldsymbol{B}$. It can be seen:

$$
\begin{aligned}
\boldsymbol{B} &= \mathbb{E}[\frac{\partial lnL_N(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \cdot \frac{\partial lnL_N(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}|_{\boldsymbol{\beta}_0}|\boldsymbol{x}_i] \\
&= \mathbb{E}(\sum_{i=1}^{N} \frac{y_i - F_i}{F_i(1-F_i)} F_i' \boldsymbol{x}_i \cdot \sum_{i=1}^{N} \frac{y_i - F_i}{F_i(1-F_i)} F_i' \boldsymbol{x}_i'|_{\boldsymbol{\beta}_0}|\boldsymbol{x}_i) \\
&= \mathbb{E}(\sum_{i=1}^{N} \frac{y_i - F_i}{F_i(1-F_i)} \frac{y_i - F_i}{F_i(1-F_i)} F_i'^2 \boldsymbol{x}_i \boldsymbol{x}_i'|_{\boldsymbol{\beta}_0}|\boldsymbol{x}_i) \\
&= \sum_{i=1}^{N} \frac{\mathbb{E}[(y_i - F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))^2|\boldsymbol{x}_i]}{F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)(1-F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))} \frac{1}{F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)(1-F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))} F'(\boldsymbol{x}_i'\boldsymbol{\beta}_0)^2 \boldsymbol{x}_i \boldsymbol{x}_i' \\
&= \sum_{i=1}^{N} \frac{1}{F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)(1-F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))} F'(\boldsymbol{x}_i'\boldsymbol{\beta}_0)^2 \boldsymbol{x}_i \boldsymbol{x}_i' \\
&= -\mathbb{E}[\frac{\partial^2 lnL_N(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}|_{\boldsymbol{\beta}_0}|\boldsymbol{x}_i] \\
&:= \boldsymbol{I}(\boldsymbol{\beta}_0) = -\boldsymbol{A}
\end{aligned}
$$

Without satisfying all the regularity conditions, the asymptotic distribution of MLE is:
$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \rightarrow_d N(\boldsymbol{0}, \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}) = N(\boldsymbol{0}, -\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{A}^{-1}) = N(\boldsymbol{0}, -\boldsymbol{A}^{-1}) = N(\boldsymbol{0}, \boldsymbol{I}(\boldsymbol{\beta_0})^{-1})$
Still the conventional one. So, we do not need to use the sandwich standard error for binary outcome model.

1.5 Special Case: Logit Model

If $F(.) = \Lambda(.)$ which is Logistic function (the c.d.f. of Logistic random variable),

$$p_i := Pr(y_i = 1|\boldsymbol{x}_i) = \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta}) = \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}}} = \frac{1}{1 + e^{-\boldsymbol{x}_i'\boldsymbol{\beta}}}$$

$$ln\frac{p_i}{1-p_i} = \Lambda^{-1}(p_i) = \boldsymbol{x}_i'\boldsymbol{\beta} \qquad \Lambda^{-1}(.) \text{ is called Logit function}$$

1.5.1 Marginal Effect

$$
\begin{aligned}
\Lambda'(z) &= \frac{d(1+e^{-z})^{-1}}{dz} \\
&= -(1+e^{-z})^{-2} e^{-z}(-1) \\
&= (1+e^{-z})^{-1}(1+e^{-z})^{-1} e^{-z} \\
&= \Lambda(z)\frac{e^{-z}}{1+e^{-z}} \\
&= \Lambda(z)\frac{1}{1+e^{z}} \\
&= \Lambda(z)\frac{1+e^{z}-e^{z}}{1+e^{z}} \\
&= \Lambda(z)(1-\Lambda(z))
\end{aligned}
$$

$$\Lambda'(\boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{\beta} = \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})(1-\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta}))\boldsymbol{\beta} \leq 0.25\boldsymbol{\beta}$$

As

$$
\begin{aligned}
\frac{dz(1-z)}{dz}|_{z^*} &= 1 - 2z^* = 0 \\
z^* &= 1/2 \\
z^*(1-z^*) &= 1/2 \cdot 1/2 = 1/4 = 0.25
\end{aligned}
$$

$\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})(1-\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})) = 0.25$ when $\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta}) = 1/2$ which happens when $\boldsymbol{x}_i'\boldsymbol{\beta} = 0$ as Logistic p.d.f. is symmetric at 0 (so Logistic c.d.f. = 0.5 at 0).

1.5.2 Odds Ratio

$$ln\frac{p_i}{1-p_i} = \Lambda^{-1}(p_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$$

If $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are only different in x_k by 1 unit, then

$$\begin{aligned}
\boldsymbol{x}_i'\boldsymbol{\beta} - \boldsymbol{x}_j'\boldsymbol{\beta} &= ln\frac{p_i}{1-p_i} - ln\frac{p_j}{1-p_j} \\
\boldsymbol{x}_j'\boldsymbol{\beta} + 1\cdot\beta_k - \boldsymbol{x}_j'\boldsymbol{\beta} = \beta_k &= ln\frac{p_i/(1-p_i)}{p_j/(1-p_j)} \\
exp(\beta_k) &= \frac{p_i/(1-p_i)}{p_j/(1-p_j)} := OR
\end{aligned}$$

Odds Ratio (OR) can be interpreted as

$$\begin{aligned}
\frac{p_j}{1-p_j} &= exp(\boldsymbol{x}_j'\boldsymbol{\beta}) \\
exp(\boldsymbol{x}_i'\boldsymbol{\beta}) = exp(\boldsymbol{x}_j'\boldsymbol{\beta} + 1\cdot\beta_k) &= exp(\boldsymbol{x}_j'\boldsymbol{\beta})exp(\beta_k) \\
&= \frac{p_j}{1-p_j}exp(\beta_k)
\end{aligned}$$

So, 1 unit increase in x_k multiplies the odds $\frac{p_j}{1-p_j}$ by $exp(\beta_k)$, which is the Odds Ratio (OR)

1.5.3 FOC

$$\begin{aligned}
\sum_{i=1}^{N}\frac{y_i - \Lambda(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})}{\Lambda(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})(1-\Lambda(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}))}\Lambda'(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})\boldsymbol{x}_i &= \boldsymbol{0} \\
\sum_{i=1}^{N}\frac{y_i - \Lambda(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})}{\Lambda(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})(1-\Lambda(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}))}\Lambda(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})(1-\Lambda(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})\boldsymbol{x}_i &= \boldsymbol{0} \\
\sum_{i=1}^{N}(y_i - \Lambda(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}))\boldsymbol{x}_i &= \boldsymbol{0} \\
\sum_{i=1}^{N}(y_i - \mathbb{E}(y_i|\boldsymbol{x}_i))\boldsymbol{x}_i &= \boldsymbol{0}
\end{aligned}$$

Similar to OLS. Moreover, if intercept included 1 i.e. $\boldsymbol{x}_i$ has 1

$$\begin{aligned}
\sum_{i=1}^{N}(y_i - \Lambda(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}))\cdot 1 &= 0 && \text{"residual" sum to 0} \\
N^{-1}\sum_{i=1}^{N}\Lambda(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}) &= \bar{y}
\end{aligned}$$

Interesting result, $\bar{y}$ is the percentage of 1 in the sample, which is the same as average predicted probability of Logit Model

1.6 Special Case: Probit Model

f $F(.) = \Phi(.)$ which is the c.d.f. of Standard Normal random variable,

$$p_i := Pr(y_i = 1|\boldsymbol{x}_i) = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}) = \int_{-\infty}^{\boldsymbol{x}_i'\boldsymbol{\beta}}\phi(z)dz$$

$$\Phi^{-1}(p_i) = \boldsymbol{x}_i'\boldsymbol{\beta} \qquad \Phi^{-1}(.) \text{ is called Probit function, no closed form}$$

Marginal Effect

$$\begin{aligned}
\Phi'(z) &= \frac{d\int_{-\infty}^{z}\phi(a)da}{dz} \\
&= \phi(z) && \text{by Fundamental Theorem of Calculus} \\
&= \frac{1}{\sqrt{2\pi}}exp(-\frac{1}{2}z^2)
\end{aligned}$$

$$\begin{aligned}
\Phi'(\boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{\beta} &= \frac{1}{\sqrt{2\pi}}exp(-\frac{1}{2}(\boldsymbol{x}_i'\boldsymbol{\beta})^2)\boldsymbol{\beta} \\
&\leq \frac{1}{\sqrt{2\pi}}\cdot 1\cdot\boldsymbol{\beta} && \text{as } 0 < exp(z) \leq 1 \text{ if } z \leq 0 \\
&\approx 0.4\boldsymbol{\beta}
\end{aligned}$$

FOC

$$\sum_{i=1}^{N}\underbrace{\{\frac{\Phi'(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})}{\Phi(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})(1-\Phi(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}))}\}}_{\widehat{w}_i}[y_i - \Phi(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})]\boldsymbol{x}_i = \boldsymbol{0}$$

1.7 Special Case: Linear Probability Model (LPM)

$F(.)$ is the identity function.

$$p_i := Pr(y_i = 1|\boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$$

However, it is not likely that $F(.)$ is identity function as the predicted probability is likely larger than 1 or smaller than 0.

There is default heteroskedasticity problem. As shown before, $Var(y_i|\boldsymbol{x}_i) = p_i(1-p_i) = \boldsymbol{x}_i'\boldsymbol{\beta}(1-\boldsymbol{x}_i'\boldsymbol{\beta})$ which depend on i.

It can be estimated by OLS with robust standard error or GLS or MLE. MLE FOC is:

$$\sum_{i=1}^{N}\underbrace{\{\frac{1}{\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}(1-\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})}\}}_{\widehat{w}_i}[y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}]\boldsymbol{x}_i = \boldsymbol{0}$$

However, if $\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \to 0$ or $\to 1$, $\widehat{w}_i$ is large and lead to numerical unstability.

1.8 The Motivation of the choice of $F(.)$

It can be motivated by Latent Variable Models or Generalized Linear Model (GLM), which will be discussed below.

1.9 Model Evaluation: Pseudo-R^2

McFadden (1974) suggests:

$$\begin{aligned}
R^2_{Binary} &= 1 - \frac{lnL_{fit}}{lnL_0} \\
&= 1 - \frac{\sum_{i=1}^{N}\{y_i ln\widehat{p}_i + (1-y_i)ln(1-\widehat{p}_i)\}}{\sum_{i=1}^{N}\{y_i ln\bar{y} + (1-y_i)ln(1-\bar{y})\}} \\
&= 1 - \frac{\sum_{i=1}^{N}\{y_i ln\widehat{p}_i + (1-y_i)ln(1-\widehat{p}_i)\}}{(\sum_{i=1}^{N} y_i)ln\bar{y} + (N - \sum_{i=1}^{N} y_i)ln(1-\bar{y})\}} \\
&= 1 - \frac{\sum_{i=1}^{N}\{y_i ln\widehat{p}_i + (1-y_i)ln(1-\widehat{p}_i)\}}{N\bar{y}ln\bar{y} + N(1-\bar{y})ln(1-\bar{y})} \\
&= 1 - \frac{\sum_{i=1}^{N}\{y_i ln\widehat{p}_i + (1-y_i)ln(1-\widehat{p}_i)\}}{N(\bar{y}ln\bar{y} + (1-\bar{y})ln(1-\bar{y}))}
\end{aligned}$$

1.10 Other Model Evaluation Methods

In-sample accuracy, out-of-sample accuracy, cross validation accuracy, confusion matrix, ROC, etc.

2 Latent Variable Models

2.1 Index Function Model

$$y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i \qquad\qquad y_i^* \text{ is unobservable}$$

But we can observe y_i

$$\begin{aligned}
y_i &= 1(y_i^* > 0) \\
\mathbb{E}(y_i|\boldsymbol{x}_i) &= 1 \cdot Pr(y_i = 1|\boldsymbol{x}_i) + 0 \cdot Pr(y_i = 0|\boldsymbol{x}_i) \\
&= Pr(y_i = 1|\boldsymbol{x}_i) \\
&= Pr(y_i^* > 0|\boldsymbol{x}_i) \\
&= Pr(\boldsymbol{x}_i'\boldsymbol{\beta} + u_i > 0|\boldsymbol{x}_i) \\
&= Pr(u_i > -\boldsymbol{x}_i'\boldsymbol{\beta}|\boldsymbol{x}_i) \\
&= Pr(u_i \leq \boldsymbol{x}_i'\boldsymbol{\beta}|\boldsymbol{x}_i) && \text{if } u_i \text{ is symmetric at } 0 \\
&= F_u(\boldsymbol{x}_i'\boldsymbol{\beta})
\end{aligned}$$

If u_i follows Logistic distribution, it is Logit model. If u_i follows standard normal distribution, it is Probit model.

2.2 Identification of parameters

$$y_i = 1 \implies y_i^* > 0 \implies \boldsymbol{x}_i'\boldsymbol{\beta} + u_i > 0$$

However, for any constant $c > 0$,

$$\boldsymbol{x}_i'\boldsymbol{\beta} + u_i > 0 \iff \boldsymbol{x}_i'c\boldsymbol{\beta} + cu_i > 0$$

So, $\boldsymbol{\beta}$ is not identified with $y_i = 1(y_i^* > 0)$. Thus, we restrict $Var(u_i|\boldsymbol{x}_i)$ to identify $\boldsymbol{\beta}$
If u_i follows Logistic distribution, $Var(u_i|\boldsymbol{x}_i) = \pi^2/3$
If u_i follows Standard Normal distribution, $Var(u_i|\boldsymbol{x}_i) = 1$

2.3 Additive Random Utility Model (ARUM)

$y = 0$ means choosing option 0. Utility obtained from this is U_0; $y = 1$ means choosing option 1. Utility obtained from this is U_1.

$$\begin{aligned}
U_0 &= V_0 + \varepsilon_0 && V_0 \text{ is deterministic component of utility} \\
U_1 &= V_1 + \varepsilon_1
\end{aligned}$$

$$\begin{aligned}
y &= 1(U_1 > U_0) \\
\mathbb{E}(y|\boldsymbol{x}) &= 1 \cdot Pr(y = 1|\boldsymbol{x}) + 0 \cdot Pr(y = 0|\boldsymbol{x}) \\
&= Pr(y = 1|\boldsymbol{x}) \\
&= Pr(U_1 > U_0|\boldsymbol{x}) \\
&= Pr(V_1 + \varepsilon_1 > V_0 + \varepsilon_0|\boldsymbol{x}) \\
&= Pr(V_1 - V_0 > \varepsilon_0 - \varepsilon_1|\boldsymbol{x}) \\
&= Pr(\varepsilon_0 - \varepsilon_1 < V_1 - V_0|\boldsymbol{x}) \\
&= F_{\varepsilon_0 - \varepsilon_1}(V_1 - V_0)
\end{aligned}$$

2.3.1 Special Case: Logit Model

If ε_0 and ε_1 are independent and both follows Type 1 Extreme Value Distribution (or log Weibull Distribution). It can be shown $\varepsilon_0 - \varepsilon_1$ follows Logistic distribution. i.e. $F_{\varepsilon_0 - \varepsilon_1}(.) = \Lambda(.)$

It can also be by the direct integration method

$$
\begin{aligned}
Pr(y=1|\boldsymbol{x}) &= Pr(\varepsilon_0 - \varepsilon_1 < V_1 - V_0|\boldsymbol{x}) \\
&= Pr(\varepsilon_0 < \varepsilon_1 + V_1 - V_0|\boldsymbol{x}) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\varepsilon_1+V_1-V_0} f(\varepsilon_0,\varepsilon_1)\partial\varepsilon_0\partial\varepsilon_1 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\varepsilon_1+V_1-V_0} f_{\varepsilon_0}(\varepsilon_0)f_{\varepsilon_1}(\varepsilon_1)\partial\varepsilon_0\partial\varepsilon_1 && \text{as independence} \\
&= \int_{-\infty}^{\infty} f_{\varepsilon_1}(\varepsilon_1)[\int_{-\infty}^{\varepsilon_1+V_1-V_0} f_{\varepsilon_0}(\varepsilon_0)\partial\varepsilon_0]\partial\varepsilon_1 \\
&= \int_{-\infty}^{\infty} f_{\varepsilon_1}(\varepsilon_1)[\int_{-\infty}^{\varepsilon_1+V_1-V_0} e^{-\varepsilon_0}exp(-e^{-\varepsilon_0})\partial\varepsilon_0]\partial\varepsilon_1 \\
&= \int_{-\infty}^{\infty} f_{\varepsilon_1}(\varepsilon_1)exp(-e^{-\varepsilon_0})|_{-\infty}^{\varepsilon_1+V_1-V_0}\partial\varepsilon_1 \\
&= \int_{-\infty}^{\infty} f_{\varepsilon_1}(\varepsilon_1)[exp(-e^{-(\varepsilon_1+V_1-V_0)}) - exp(-e^{--\infty})]\partial\varepsilon_1 \\
&= \int_{-\infty}^{\infty} f_{\varepsilon_1}(\varepsilon_1)exp(-e^{-(\varepsilon_1+V_1-V_0)})\partial\varepsilon_1 \\
&= \int_{-\infty}^{\infty} e^{-\varepsilon_1}exp(-e^{-\varepsilon_1})exp(-e^{-(\varepsilon_1+V_1-V_0)})\partial\varepsilon_1 \\
&= \int_{-\infty}^{\infty} e^{-\varepsilon_1}exp(-e^{-\varepsilon_1} - e^{-(\varepsilon_1+V_1-V_0)})\partial\varepsilon_1 \\
&= \int_{-\infty}^{\infty} e^{-\varepsilon_1}exp(-e^{-\varepsilon_1} - e^{-\varepsilon_1}e^{-(V_1-V_0)})\partial\varepsilon_1 \\
&= \int_{-\infty}^{\infty} e^{-\varepsilon_1}exp(-e^{-\varepsilon_1}(1+e^{-(V_1-V_0)}))\partial\varepsilon_1 \\
&= 1/(1+e^{-(V_1-V_0)}) \\
&= \Lambda(V_1-V_0)
\end{aligned}
$$

As $\int_{-\infty}^{\infty} ae^{-\varepsilon}exp(-ae^{-\varepsilon})d\varepsilon = 1$
If $V_1 - V_0 = \boldsymbol{x'\beta}$, $Pr(y=1|\boldsymbol{x}) = \Lambda(\boldsymbol{x'\beta})$. It is Logit model.

2.3.2 Special Case: Probit Model

If ε_0 and ε_1 are multivariate (bivariate here) standard normally distributed, any linear combination of ε_0 and ε_1 also follow standard normal. So, $\varepsilon_0 - \varepsilon_1$ follows univariate standard normal. i.e., $F_{\varepsilon_0-\varepsilon_1}(.) = \Phi(.)$

3 Berkson's Minimum Chi-square Estimator

$$p_t := Pr(y_i = 1|\boldsymbol{x}_i = \boldsymbol{x}_t) = F(\boldsymbol{x}_t'\boldsymbol{\beta})$$
$$F^{-1}(p_t) = \boldsymbol{x}_t'\boldsymbol{\beta}$$

p_t can be estimated by $\bar{p}_t = \bar{y}_t = N_t^{-1}\sum_{i=1}^{N_t} y_{it}$

$$F^{-1}(\bar{p}_t) - F^{-1}(\bar{p}_t) + F^{-1}(p_t) = \boldsymbol{x}_t'\boldsymbol{\beta}$$
$$F^{-1}(\bar{p}_t) = \boldsymbol{x}_t'\boldsymbol{\beta} + \underbrace{F^{-1}(\bar{p}_t) - F^{-1}(p_t)}_{v_t}$$

As $Var(v_t|\boldsymbol{x}_t$ depends on t, there is heteroskedasticity. GLS (WLS here) can be used to estimate $\boldsymbol{\beta}$ efficiently.

$$\widehat{\boldsymbol{\beta}} = arg\ min_{\boldsymbol{\beta}} \sum_{t=1}^{T} Var(v_t|\boldsymbol{x}_t)^{-1}(F^{-1}(\bar{p}_t) - \boldsymbol{x}_t'\boldsymbol{\beta})^2$$

4 Semi-parametric Estimation

4.1 Maximum Score Estimation (Manski, 1975, 1985)

We predict $y_i = 1$ if $\boldsymbol{x}_i'\boldsymbol{\beta} > 0$. We predict $y_i = 0$ if $\boldsymbol{x}_i'\boldsymbol{\beta} \leq 0$. The Score function, which counts how many observations we predict correctly, is defined as

$$
\begin{aligned}
S_N(\boldsymbol{\beta}) &:= \sum_{i=1}^{N}\{y_i 1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0) + (1-y_i)1(\boldsymbol{x}_i'\boldsymbol{\beta} \leq 0)\} \\
&= \sum_{i=1}^{N}\{y_i 1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0) + (1-y_i)(1 - 1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0))\} \\
&= \sum_{i=1}^{N}\{y_i 1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0) + 1 - y_i - (1-y_i)1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0)\} \\
&= \sum_{i=1}^{N}\{y_i 1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0) + 1 - y_i - 1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0) + y_i 1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0)\} \\
&= \sum_{i=1}^{N}\{(2y_i - 1)1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0) + 1 - y_i\} \\
&= \sum_{i=1}^{N}(2y_i - 1)1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0) + N - \sum_{i=1}^{N} y_i
\end{aligned}
$$

The Maximum Score Estimator is

$$
\widehat{\boldsymbol{\beta}} = arg\ max_{\boldsymbol{\beta}} \sum_{i=1}^{N}(2y_i - 1)1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0) \qquad \text{can not use differentiation}
$$

MSE can be regarded as Least Absolute Deviation (LAD) Estimator. It can be seen

$$
\begin{aligned}
Q_N(\boldsymbol{\beta}) &= N - S_N(\boldsymbol{\beta}) \\
&= \sum_{i=1}^{N}(1 - \{(2y_i - 1)1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0) + 1 - y_i\}) \\
&= \sum_{i=1}^{N}\{y_i - (2y_i - 1)1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0)\} \\
&= \sum_{i=1}^{N}\begin{cases} 1 - 1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0) & \text{if } y_i = 1 \\ 1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0) & \text{if } y_i = 0 \end{cases} \\
&= \sum_{i=1}^{N}\begin{cases} y_i - 1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0) & \text{if } y_i = 1 \\ -(y_i - 1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0)) & \text{if } y_i = 0 \end{cases} \\
&= \sum_{i=1}^{N}\begin{cases} |y_i - 1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0)| & \text{if } y_i = 1 \text{ as } 1 - 1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0) \geq 0 \\ |y_i - 1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0)| & \text{if } y_i = 0 \text{ as } 0 - 1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0) \leq 0 \end{cases} \\
&= \sum_{i=1}^{N} |y_i - 1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0)| \\
&= \sum_{i=1}^{N} |y_i - Median(y_i|\boldsymbol{x}_i)|
\end{aligned}
$$

The last line since

$$\begin{aligned}
Median(y_i|\boldsymbol{x}_i) &= Median(1(y_i^* > 0)|\boldsymbol{x}_i) \\
&= 1(Median(y_i^*|\boldsymbol{x}_i) > 0) \\
&= 1(Median(\boldsymbol{x}_i'\boldsymbol{\beta} + u_i|\boldsymbol{x}_i) > 0) \\
&= 1(Median(\boldsymbol{x}_i'\boldsymbol{\beta}|\boldsymbol{x}_i) + \underbrace{Median(u_i|\boldsymbol{x}_i)}_{0} > 0) && \text{assume } 0 \\
&= 1(\boldsymbol{x}_i'\boldsymbol{\beta} > 0)
\end{aligned}$$

Assume $Median(u_i|\boldsymbol{x}_i) = 0$, $\widehat{\boldsymbol{\beta}}$ is consistent but $N^{1/3}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to non-normal distribution.

4.2 Smooth Maximum Score Estimation (Horowitz, 1992)

$$\widehat{\boldsymbol{\beta}} = arg\ max_{\boldsymbol{\beta}} \sum_{i=1}^{N} (2y_i - 1) K(\boldsymbol{x}_i'\boldsymbol{\beta}/h_N)$$

4.3 Semi-parametric MLE (Klein & Spady, 1993)

$$\sum_{i=1}^{N} \underbrace{\{\frac{\widehat{F}'(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})}{\widehat{F}(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})(1 - \widehat{F}(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}))}}_{\widehat{w}_i} \}[y_i - \widehat{F}(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})]\boldsymbol{x}_i = \boldsymbol{0}$$

Initialize $\boldsymbol{\beta}^{(1)}$, estimate $F^{(1)}$ by kernel estimation.
Given $F^{(1)}$ and $\boldsymbol{\beta}^{(1)}$, estimate $\boldsymbol{\beta}^{(2)}$ by gradient descent method i.e., $\boldsymbol{\beta}^{(2)} = \boldsymbol{\beta}^{(1)} + \boldsymbol{A}_N \frac{\partial Q_N(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}^{(1)}}$
Given $\boldsymbol{\beta}^{(2)}$, estimate $F^{(1)}$. Repeat until convergence of $\boldsymbol{\beta}$

5 References

Cameron, A. C., & Trivedi, P. K. (2005). Microeconometrics: Methods and Applications