Notes on Binary Outcome Model

Max Leung

February 20, 2024

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

$$\begin{aligned} \frac{\partial Pr(y_i = 1|\boldsymbol{x}_i)}{\partial \boldsymbol{x}_i} &= \frac{\partial F(\boldsymbol{x}_i'\boldsymbol{\beta})}{\partial \boldsymbol{x}_i} \\ &= \frac{\partial F(\boldsymbol{x}_i'\boldsymbol{\beta})}{\partial \boldsymbol{x}_i'\boldsymbol{\beta}} \frac{\partial \boldsymbol{x}_i'\boldsymbol{\beta}}{\partial \boldsymbol{x}_i} \\ &= F'(\boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{\beta} \end{aligned}$$

If $F(.)$ is cdf, $F'(.) > 0$. So, $sign(\boldsymbol{\beta})$ decide the sign

Average Marginal Effect is

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

or

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

If $\overline{\boldsymbol{x}}$ includes $\overline{ln(z)}$, it can be replaced by $ln(\overline{z})$. If $\boldsymbol{x}_i$ includes z_i and z_i^2, AME_i for z_i is $F'(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})(\widehat{\beta}_z + 2\widehat{\beta}_z z_i)$. Therefore, AME is $F'(\overline{\boldsymbol{x}}'\widehat{\boldsymbol{\beta}})(\widehat{\beta}_z + 2\widehat{\beta}_z \overline{z})$ where $\overline{z^2}$ in $\overline{\boldsymbol{x}}$ can be replaced by $\overline{z}^2$. If $\boldsymbol{x}_i$ includes an indicator z_i, AME is $F(\overline{\boldsymbol{x}}'_{-z}\widehat{\boldsymbol{\beta}}_{-z} + 1 \cdot \widehat{\beta}_z) - F(\overline{\boldsymbol{x}}'_{-z}\widehat{\boldsymbol{\beta}}_{-z})$.

1.2 Maximum Likelihood Estimation

1.2.1 Probability Mass Function

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

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

1.2.2 Log Likelihood Function

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

1.2.3 Gradient Vector

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

1.2.4 First Order Condition

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

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

There is no closed form solution for $\widehat{\boldsymbol{\beta}}_{mle}$. We usually solve it with Newton method. $ln[L_N(\boldsymbol{\beta})]$ is globally concave because its Hessian matrix is negative definite. As a result, the convergence of the algorithm is fast.

1.2.5 Consistency of Maximum Likelihood Estimator

Correct specification of the likelihood function is a necessary condition for consistent MLE $\widehat{\boldsymbol{\beta}}_{mle} \to_p \boldsymbol{\beta}_0$ as $N \to \infty$. As y_i is binary, it must be Bernoulli distributed. However, $p_i := Pr(y_i = 1|\boldsymbol{x}_i) = F(\boldsymbol{x}_i'\boldsymbol{\beta})$ can be incorrectly specified.

1.2.6 Asymptotic Distribution of Maximum Likelihood Estimator

Under regularity conditions, Information Matrix equality holds and MLE is asymptotically normally distributed. Its asymptotic variance is the inverse of Fisher's Information matrix.

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

$$\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 not all the regularity conditions for Information Matrix equality hold, the equality still holds algebraically. 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 $\sqrt{N}(\widehat{\boldsymbol{\beta}}_{mle} - \boldsymbol{\beta}_0) \to_d N(\boldsymbol{0}, \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}) = N(\boldsymbol{0}, -\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{A}^{-1}) = N(\boldsymbol{0}, -\boldsymbol{A}^{-1}) = N(\boldsymbol{0}, \boldsymbol{I}(\boldsymbol{\beta}_0)^{-1})$. Thus, the asymptotic variance is still the conventional one i.e., the inverse of the Information Matrix. Therefore, we do not have to use the sandwich standard error for binary outcome model.

1.3 Iteratively Weighted Non-linear Least Squares

1.3.1 Loss Function

$$\sum_{i=1}^{N} w_i(y_i - F(\boldsymbol{x}_i'\boldsymbol{\beta}))^2$$

1.3.2 First Order Condition

$$\begin{aligned}
\frac{\partial \sum_{i=1}^{N} w_i(y_i - F(\boldsymbol{x}_i'\boldsymbol{\beta}))^2}{\partial \boldsymbol{\beta}}|_{\widehat{\boldsymbol{\beta}}_{wnls}} &= \boldsymbol{0} \\
\sum_{i=1}^{N} w_i \frac{\partial (y_i - F(\boldsymbol{x}_i'\boldsymbol{\beta}))^2}{\partial \boldsymbol{\beta}}|_{\widehat{\boldsymbol{\beta}}_{wnls}} &= \boldsymbol{0} \\
-2\sum_{i=1}^{N} w_i(y_i - F(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{wnls})) \frac{\partial F(\boldsymbol{x}_i'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\widehat{\boldsymbol{\beta}}_{wnls}} &= \boldsymbol{0} \\
\sum_{i=1}^{N} w_i(y_i - F(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{wnls}))F'(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{wnls})\boldsymbol{x}_i &= \boldsymbol{0} \\
\sum_{i=1}^{N} w_i F'(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{wnls})[y_i - F(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{wnls})]\boldsymbol{x}_i &= \boldsymbol{0}
\end{aligned}$$

If w_i is the inverse of conditional variance of y_i i.e., $\frac{1}{F(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})[1-F(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})]}$, WNLS and ML estimation have the same FOC, except $\widehat{\boldsymbol{\beta}}$ in WNLS's weight is from NLS or last step while that in MLE is still $\widehat{\boldsymbol{\beta}}_{mle}$.

1.3.3 Algorithm

For $t = 1 \cdots$

Calculate the weight w_i^t with $\widehat{\boldsymbol{\beta}}_{wnls}^{t-1}$ from last step or $\widehat{\boldsymbol{\beta}}_{nls}^{0}$ from NLS if $t = 1$.

Get $\widehat{\boldsymbol{\beta}}_{wnls}^{t} = arg\ min_{\boldsymbol{\beta}} \sum_{i=1}^{N} w_i^t(y_i - F(\boldsymbol{x}_i'\boldsymbol{\beta}))^2 = arg\ min_{\boldsymbol{\beta}} \sum_{i=1}^{N} (\sqrt{w_i^t}y_i - \sqrt{w_i^t}F(\boldsymbol{x}_i'\boldsymbol{\beta}))^2$ with appropriate algorithm for NLS e.g., Gauss-Newton method.

Repeat above two steps until convergence of $\widehat{\boldsymbol{\beta}}^t_{wnls}$, which will be equal to $\widehat{\boldsymbol{\beta}}_{mle}$.

ML and NLS estimator are special cases of M estimator, which is a special case of extremum estimator (Amemiya, 1985). Therefore, under regularity conditions of extremum estimator, NLS estimator is consistent and asymptotic normal. However, it is inefficient. WNLS with an inverse conditional variance weight is efficient. Thus, it is not necessary to apply IWNLS.

1.4 GMM Estimation

1.4.1 Unconditional Moments

Set $w_i = \frac{1}{F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)(1-F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))}$,

$$\mathbb{E}[\frac{\partial w_i(y_i - F(\boldsymbol{x}_i'\boldsymbol{\beta}))^2}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}_0}] = \mathbf{0}$$
$$\mathbb{E}[\underbrace{w_i F'(\boldsymbol{x}_i'\boldsymbol{\beta}_0)(y_i - F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))\boldsymbol{x}_i}_{\boldsymbol{g}(y_i,\boldsymbol{x}_i;\boldsymbol{\beta}_0)}] = \mathbf{0}$$

It is just-identified because the number of moments i.e., $dim(\boldsymbol{x}_i) = dim(\boldsymbol{\beta}_0)$ i.e., the number of parameters. GMM reduced to MM estimation.

1.4.2 Asymptotic Distribution

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{gmm} - \boldsymbol{\beta}_0) \to_d N(\mathbf{0}, (\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G})^{-1}\boldsymbol{G}'\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}\boldsymbol{G}(\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G})^{-1})$$

where $\boldsymbol{G} = \mathbb{E}[\frac{\partial \boldsymbol{g}(y_i,\boldsymbol{x}_i;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}_0}]$ and $\boldsymbol{S} = \mathbb{E}[\boldsymbol{g}(y_i,\boldsymbol{x}_i;\boldsymbol{\beta}_0)\boldsymbol{g}(y_i,\boldsymbol{x}_i;\beta_0)']$. $\boldsymbol{S}$ has this simple form because y_i is assumed to be independent and must be Bernoulli distributed i.e., i.i.d.

$$\begin{aligned}
\boldsymbol{G} &= \mathbb{E}[\frac{\partial \boldsymbol{g}(y_i,\boldsymbol{x}_i;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}_0}] \\
&= \mathbb{E}[\frac{\partial w_i F'(\boldsymbol{x}_i'\boldsymbol{\beta})(y_i - F(\boldsymbol{x}_i'\boldsymbol{\beta}))\boldsymbol{x}_i}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}_0}] \\
&= \mathbb{E}[\mathbb{E}[\frac{\partial \frac{F'(\boldsymbol{x}_i'\boldsymbol{\beta})}{F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)(1-F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))}(y_i - F(\boldsymbol{x}_i'\boldsymbol{\beta}))\boldsymbol{x}_i}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}_0}|\boldsymbol{x}_i]] \\
&= \mathbb{E}[-\frac{(F'(\boldsymbol{x}_i'\boldsymbol{\beta}_0))^2}{F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)(1-F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))}\boldsymbol{x}_i\boldsymbol{x}_i']
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{S} &= \mathbb{E}[\boldsymbol{g}(y_i,\boldsymbol{x}_i;\boldsymbol{\beta}_0)\boldsymbol{g}(y_i,\boldsymbol{x}_i;\boldsymbol{\beta}_0)'] \\
&= \mathbb{E}[w_i F'(\boldsymbol{x}_i'\boldsymbol{\beta}_0)(y_i - F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))\boldsymbol{x}_i(w_i F'(\boldsymbol{x}_i'\boldsymbol{\beta}_0)(y_i - F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))\boldsymbol{x}_i)'] \\
&= \mathbb{E}[(w_i F'(\boldsymbol{x}_i'\boldsymbol{\beta}_0))^2\mathbb{E}[(y_i - F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))^2|\boldsymbol{x}_i]\boldsymbol{x}_i\boldsymbol{x}_i'] \\
&= \mathbb{E}[(w_i F'(\boldsymbol{x}_i'\boldsymbol{\beta}_0))^2 Var(y_i|\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i'] \\
&= \mathbb{E}[(\frac{F'(\boldsymbol{x}_i'\boldsymbol{\beta}_0)}{F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)(1-F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))})^2 F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)(1-F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))\boldsymbol{x}_i\boldsymbol{x}_i'] \\
&= \mathbb{E}[\frac{(F'(\boldsymbol{x}_i'\boldsymbol{\beta}_0))^2}{F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)(1-F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))}\boldsymbol{x}_i\boldsymbol{x}_i'] = -\boldsymbol{G}
\end{aligned}$$

Because of just-identification, GMM reduces to MM estimation, $\boldsymbol{G}$ is a square matrix and thus invertible.

$$\begin{aligned}
\sqrt{N}(\widehat{\boldsymbol{\beta}}_{mm} - \boldsymbol{\beta}_0) &\to_d N(\mathbf{0}, (\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G})^{-1}\boldsymbol{G}'\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}\boldsymbol{G}(\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G})^{-1}) \\
&= N(\mathbf{0}, \boldsymbol{G}^{-1}\boldsymbol{W}^{-1}\boldsymbol{G}'^{-1}\boldsymbol{G}'\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}\boldsymbol{G}\boldsymbol{G}^{-1}\boldsymbol{W}^{-1}\boldsymbol{G}'^{-1}) \\
&= N(\mathbf{0}, \boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}'^{-1}) \\
&= N(\mathbf{0}, [\boldsymbol{I}(\boldsymbol{\beta}_0)]^{-1})
\end{aligned}$$

Because

$$\boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}'^{-1} = -\boldsymbol{G}'^{-1} = \{\mathbb{E}[\frac{(F'(\boldsymbol{x}_i'\boldsymbol{\beta}_0))^2}{F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)(1-F(\boldsymbol{x}_i'\boldsymbol{\beta}_0))}\boldsymbol{x}_i\boldsymbol{x}_i']\}^{-1} = [\boldsymbol{I}(\boldsymbol{\beta}_0)]^{-1}$$

Thus, GMM/MM and ML estimator has the same asymptotic distribution.

1.5 Special Case: Logit Model

If $F(.) = \Lambda(.)$ i.e., Logistic function (the c.d.f. of standard 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 p.d.f. of standard Logistic random variable is symmetric at 0 (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 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, an unit increase in x_k means the odds $\frac{p_j}{1-p_j}$ is multiplied by the Odds Ratio (OR) $exp(\beta_k)$.

1.5.3 First Order Condition

$$\sum_{i=1}^{N} \frac{y_i - \Lambda(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})}{\Lambda(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})(1-\Lambda(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}))}\Lambda'(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})\boldsymbol{x}_i = \boldsymbol{0}$$

$$\sum_{i=1}^{N} \frac{y_i - \Lambda(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})}{\Lambda(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})(1-\Lambda(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}))}\Lambda(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})(1-\Lambda(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})\boldsymbol{x}_i = \boldsymbol{0}$$

$$\sum_{i=1}^{N}(y_i - \Lambda(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}))\boldsymbol{x}_i = \boldsymbol{0}$$

$$\sum_{i=1}^{N}(y_i - \mathbb{E}(y_i|\boldsymbol{x}_i))\boldsymbol{x}_i = \boldsymbol{0}$$

This is similar to the first order condition of OLS estimation of linear model. Moreover, if intercept is included in $\boldsymbol{x}_i$.

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

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

Interesting result, $\bar{y}$ is the percentage of one in the sample, which is equal to the average predicted probability of Logit Model.

1.6 Special Case: Probit Model

If $F(.) = \Phi(.)$ i.e., the c.d.f. of standard Normal random variable,

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

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

1.6.1 Marginal Effect

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

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

1.6.2 First Order Condition

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

1.7 Special Case: Robit Model

if $F(.) = F_{t,\nu}(.)$ i.e., the c.d.f. of standard student's t random variable,

$$p_i := Pr(y_i = 1|\boldsymbol{x}_i) = F_{t,\nu}(\boldsymbol{x}_i'\boldsymbol{\beta}) = \int_{-\infty}^{\boldsymbol{x}_i'\boldsymbol{\beta}} t(z)dz$$

$$F_{t,\nu}^{-1}(p_i) = \boldsymbol{x}_i'\boldsymbol{\beta} \qquad F_{t,\nu}^{-1}(.) \text{ is called Robit function, no closed form}$$

As standard student's t random variable converges to standard normal random variable when the degree of freedom $\nu \to \infty$, Robit model is a generalization of Probit model. The extra flexibility provided by ν seems to handle outliers well. Except some special cases, $F_{t,\nu}(.)$ does not have a closed form. The additional parameter ν can be estimated by ML method with $\boldsymbol{\beta}$ simultaneously, or a grid of ν can be pre-specified.

1.8 Special Case: Linear Probability Model (LPM)

If $F(.)$ is an identity function,

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

However, it is not likely that $F(.)$ is an identity function because the resulting predicted probability can be larger than 1 or smaller than 0. MLE's first order condition is

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

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. Heteroskedasticity can be solved by using GLS estimation i.e., WLS estimation with an inverse conditional variance weight. First, use $\widehat{\boldsymbol{\beta}}_{ols}$ to get the weight $[\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{ols}(1-\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{ols})]^{-1}$ for $\forall i$. Second, get

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{wls} &= argmin_{\boldsymbol{\beta}} \sum_{i=1}^{N}\{\frac{1}{\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{ols}(1-\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{ols})}\}[y_i - \boldsymbol{x}_i'\boldsymbol{\beta}]^2 \\
&= argmin_{\boldsymbol{\beta}} \sum_{i=1}^{N}\{\frac{1}{\sqrt{\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{ols}(1-\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{ols})}}\}^2[y_i - \boldsymbol{x}_i'\boldsymbol{\beta}]^2 \\
&= argmin_{\boldsymbol{\beta}} \sum_{i=1}^{N}[\frac{y_i}{\sqrt{\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{ols}(1-\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{ols})}} - \frac{\boldsymbol{x}_i'}{\sqrt{\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{ols}(1-\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{ols})}}\boldsymbol{\beta}]^2 \\
&= argmin_{\boldsymbol{\beta}} \sum_{i=1}^{N}[\tilde{y}_i - \tilde{\boldsymbol{x}}_i'\boldsymbol{\beta}]^2
\end{aligned}$$

Therefore, $\widehat{\boldsymbol{\beta}}_{wls} = (\sum_{i=1}^{N}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i')^{-1}\sum_{i=1}^{N}\tilde{\boldsymbol{x}}_i\tilde{y}_i$. Its first order condition is

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

If $\widehat{\boldsymbol{\beta}}_{wls}$ is then used in estimating the weight and loop the above process until convergence, the resulting IWLS estimate is equal to ML estimate. It is a special case of IWNLS = ML estimation derived in the previous section for general model. $\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \to 0$ or $\to 1$, $\widehat{w}_i$ is large and lead to numerical instability during optimization.

If $F(.)$ is truly an identity function, OLS estimator is still unbiased and consistent. The default heteroskedasticity can be tackled by using robust standard error e.g., Eicker-Huber-White standard error. However, it is still inefficient.

Some scholars e.g., Angrist advocate using LPM e.g., if your target is to estimate Average Treatment Effect (ATE) and you have only one indicator regressor, the OLS slope estimate is estimated ATE under independent assumption, which can be justified by random assignment. This result holds even y is not binary. $\beta = \frac{Cov(Y,D)}{Var(D)} = \mathbb{E}(Y|D=1) - \mathbb{E}(Y|D=0)$, second equality uses double expectation to derive.

$$\begin{aligned}
\mathbb{E}(Y|D=1) - \mathbb{E}(Y|D=0) &= \mathbb{E}(Y_1|D=1) - \mathbb{E}(Y_0|D=0) \\
&= \mathbb{E}(Y_1) - \mathbb{E}(Y_0) \qquad Y_1, Y_0 \perp D \text{ under random assignment} \\
&= \mathbb{E}(Y_1 - Y_0) = ATE
\end{aligned}$$

Stoker's average derivatives provide the average marginal effect without specifying $F(.)$. Define $\mathbb{E}[y|\boldsymbol{x}] = m(\boldsymbol{x})$,

$$\begin{aligned} AME := \mathbb{E}[\frac{\partial m(\boldsymbol{x})}{\partial \boldsymbol{x}}] &= -\mathbb{E}[m(\boldsymbol{x})\frac{\partial ln(f(\boldsymbol{x}))}{\partial \boldsymbol{x}}] \qquad \text{by generalized Information matrix equality} \\ &= -\mathbb{E}[\mathbb{E}[y|\boldsymbol{x}]\frac{f'(\boldsymbol{x})}{f(\boldsymbol{x})}] \\ &= -\mathbb{E}[\mathbb{E}[y\frac{f'(\boldsymbol{x})}{f(\boldsymbol{x})}|\boldsymbol{x}]] \\ &= -\mathbb{E}[y\frac{f'(\boldsymbol{x})}{f(\boldsymbol{x})}] \end{aligned}$$

Thus,

$$\widehat{AME} = -N^{-1}\sum_{i=1}^{N} y_i \frac{\widehat{f}'(\boldsymbol{x}_i)}{\widehat{f}(\boldsymbol{x}_i)}$$

If $\boldsymbol{x}$ follow multivariate normal,

$$\begin{aligned} AME &= -\mathbb{E}[y \cdot -(\boldsymbol{x}-\boldsymbol{\mu}_x)\boldsymbol{\Sigma}_x^{-1}] \\ &= \mathbb{E}[y(\boldsymbol{x}-\boldsymbol{\mu}_x)\{\mathbb{E}[(\boldsymbol{x}-\boldsymbol{\mu}_x)(\boldsymbol{x}-\boldsymbol{\mu}_x)']\}^{-1}] \\ &= \{\mathbb{E}[(\boldsymbol{x}-\boldsymbol{\mu}_x)(\boldsymbol{x}-\boldsymbol{\mu}_x)']\}^{-1}\mathbb{E}[(\boldsymbol{x}-\boldsymbol{\mu}_x)y] \end{aligned}$$

Thus, it is an OLS estimator of regressing y on $\boldsymbol{x}-\boldsymbol{\mu}_x$. This result holds no matter y is binary or not. If your target is AME with multivariate normal $\boldsymbol{x}$, you can simply regress y on $\boldsymbol{x}-\boldsymbol{\mu}_x$, the estimated coefficients is estimated AME. Even $\boldsymbol{x}$ is not multivariate normal, AME can be non-parametrically estimated using $-N^{-1}\sum_{i=1}^{N} y_i \frac{\widehat{f}'(\boldsymbol{x}_i)}{\widehat{f}(\boldsymbol{x}_i)}$.

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.

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

It can be motivated by Latent Variable Models and Generalized Linear Model (GLM) discussed below.

2 Latent Variable Models

2.1 Index Function Model

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

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 standard Logistic distribution, the model is Logit model. If u_i follows standard Normal distribution, it is Probit model. If u_i follows standard studnet's t distribution, it is Robit 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 standard 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 (log Weibull distribution). It can be shown $\varepsilon_0 - \varepsilon_1$ follows standard Logistic distribution i.e., $F_{\varepsilon_0 - \varepsilon_1}(.) = \Lambda(.)$.

It can also be shown by direct integration,

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

If $V_1 - V_0 = \boldsymbol{x}'\boldsymbol{\beta}$, $Pr(y=1|\boldsymbol{x}) = \Lambda(\boldsymbol{x}'\boldsymbol{\beta})$. 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 Lagrange Multiplier Test, that is LM Test

that is easy to go through, no reason to take a break.

$$\boldsymbol{S} = \begin{pmatrix} s_1(\widehat{\boldsymbol{\beta}})' \\ \vdots \\ s_N(\widehat{\boldsymbol{\beta}})' \end{pmatrix} \qquad \text{where } s_i(\widehat{\boldsymbol{\beta}}) = \frac{\partial lnl_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\widehat{\boldsymbol{\beta}}}$$

$$\alpha_{LM} = N\frac{\mathbf{1}'\boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'\mathbf{1}}{\mathbf{1}'\mathbf{1}} = N \cdot R^2 \to_d \chi^2$$

$\boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'$ is the Hat matrix. Thus, $\boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'\mathbf{1}$ is an OLS predicted value of Outer-Product-of-the-Gradient (OPG) regression in which $\mathbf{1}$ is dependent variable and $\boldsymbol{S}$ is regressors. The Sum of Squares Regression i.e., Explained Sum of Squares is

$$\begin{aligned}(\boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'\mathbf{1})'\boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'\mathbf{1} &= \mathbf{1}'\boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'\mathbf{1} \\ &= \mathbf{1}'\boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'\mathbf{1}\end{aligned}$$

$\mathbf{1}'\mathbf{1}$ is Sum of Squares Total. Thus, α_{LM} is N times R^2.

3.1 Normality Test

If the density function of error term u_i in latent variable model is within the Pearson family, its c.d.f. is

$$Pr(u_i \leq t) = \Phi(t + \gamma_1 t^2 + \gamma_2 t^3)$$

where γ_1 controls the third moment i.e., skewness while γ_2 controls the fourth moment i.e., excess kurtosis. Clearly, u_i is standard normally distributed if both γ_1 and γ_2 are zero. This is the zero hypothesis.

Extended probit model becomes

$$Pr(y_i = 1|\boldsymbol{x}_i) = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta} + \gamma_1(\boldsymbol{x}_i'\boldsymbol{\beta})^2 + \gamma_2(\boldsymbol{x}_i'\boldsymbol{\beta})^3)$$

$$s_i(\begin{pmatrix} \widehat{\boldsymbol{\beta}}_{mle} \\ \widehat{\gamma}_{1,mle} \\ \widehat{\gamma}_{2,mle} \end{pmatrix}) = \begin{pmatrix} \frac{\partial lnl_i(\boldsymbol{\beta},\gamma_1,\gamma_2)}{\partial \boldsymbol{\beta}}|_{\widehat{\boldsymbol{\beta}}_{mle}} \\ \frac{\partial lnl_i(\boldsymbol{\beta},\gamma_1,\gamma_2)}{\partial \gamma_1}|_{\widehat{\gamma}_{1,mle}} \\ \frac{\partial lnl_i(\boldsymbol{\beta},\gamma_1,\gamma_2)}{\partial \gamma_2}|_{\widehat{\gamma}_{2,mle}} \end{pmatrix} = \begin{pmatrix} \widehat{\varepsilon}_i^G \boldsymbol{x}_i \\ \widehat{\varepsilon}_i^G(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{mle})^2 \\ \widehat{\varepsilon}_i^G(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{mle})^3 \end{pmatrix}$$

where $\widehat{\varepsilon}_i^G := \frac{y_i - \Phi(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{mle})}{\Phi(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{mle})(1-\Phi(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{mle}))}\Phi'(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{mle})$ is the generalized residual. We know that matrix $\boldsymbol{S}$ has variables $\widehat{\varepsilon}_i^G\boldsymbol{x}_i'$, $\widehat{\varepsilon}_i^G(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{mle})^2$ and $\widehat{\varepsilon}_i^G(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{mle})^3$. We can then regress 1 on these variables. The R^2 of such model times N is α_{LM}, which is asymptotically Chi-squared distributed with two degree of freedom under zero hypothesis. If α_{LM} is large enough, we reject zero hypothesis i.e., reject normality of u_i and $F(.)$ should not be $\Phi(.)$.

4 Generalized Linear Model

GLM is originated from Statistics, and usually not covered in Econometrics textbook.

4.1 Exponential Family of Distribution

4.1.1 Probability Mass Function

$$f(y|\theta,\eta) = a(y)exp[\frac{y\cdot\theta - b(\theta)}{\eta} + c(y,\eta)]$$

4.1.2 Moments

$$\mathbb{E}(y) = b'(\theta)$$
$$Var(y) = b''(\theta)\eta$$

If $\theta = \boldsymbol{x_i'\beta}$, we have $\mathbb{E}(y) = b'(\boldsymbol{x_i'\beta})$. Thus, $b'(.)$ is the canonical mean function and $b'^{-1}(.)$ is the canonical link function.

4.1.3 Special Case: Binomial Distribution

$$\begin{aligned}
f(y|p) &= \binom{n}{p} p^y(1-p)^{n-y} \\
&= exp[ln(\binom{n}{p} p^y(1-p)^{n-y})] \\
&= exp[ln\binom{n}{p} + ln(p^y(1-p)^{n-y})] \\
&= exp[ln\binom{n}{p} + y\cdot ln(p) + (n-y)ln(1-p)] \\
&= exp[ln\binom{n}{p} + y\cdot ln(p) + n\cdot ln(1-p) - y\cdot ln(1-p)] \\
&= exp[ln\binom{n}{p} + y(ln(p) - ln(1-p)) + n\cdot ln(1-p)] \\
&= exp[y(ln\frac{p}{1-p}) - (-n\cdot ln(1-p)) + ln\binom{n}{p}]
\end{aligned}$$

Thus, $\eta = 1$, $a(y) = 1$, $\theta = ln\frac{p}{1-p}$, $b(\theta) = -n\cdot ln(1-p)$, $c(y,\eta) = ln\binom{n}{p}$

$$\theta = ln\frac{p}{1-p} = logit(p) \iff logistic(\theta) = p$$

$$\begin{aligned}
b(\theta) &= -n\cdot ln(1-p) \\
&= -n\cdot ln(1-logistic(\theta)) \\
&= -n\cdot ln(1-\frac{1}{1+e^{-\theta}}) \\
&= -n\cdot ln(\frac{e^{-\theta}}{1+e^{-\theta}}) \\
&= -n\cdot ln(\frac{1}{1+e^{\theta}}) \\
&= n\cdot ln(1+e^{\theta})
\end{aligned}$$

$$b'(\theta) = \frac{d\ n\cdot ln(1+e^{\theta})}{d\theta} = n\frac{1}{1+e^{\theta}}e^{\theta} = n\cdot logistic(\theta)$$

Thus, $\mathbb{E}(y) = b'(\theta) = n\cdot logistic(\theta)$. Bernoulli distribution is a special case of Binomial distribution with $n = 1$. Therefore, binary outcome model, i.e., y is Bernoulli distributed, is $\mathbb{E}(y) = logistic(\boldsymbol{x'\beta})$ by setting $n = 1$ and $\theta = \boldsymbol{x'\beta}$. Logistic function is the canonical mean function for binary outcome model. $F(.)$ is thus selected to be logistic function.

5 Conditional Logistic Regression

Conditional MLE is applied in non-linear panel model in Econometrics.

$$Pr(y_i = 1) = logistic(\gamma + \boldsymbol{x}_i'\boldsymbol{\beta}) = \frac{e^{(\gamma+\boldsymbol{x}_i'\boldsymbol{\beta})}}{1 + e^{(\gamma+\boldsymbol{x}_i'\boldsymbol{\beta})}}$$

5.1 Full Likelihood Function

$$\begin{aligned}
L(\boldsymbol{y}; \boldsymbol{\beta}, \gamma) &= \prod_{i=1}^{N} \frac{e^{(\gamma+\boldsymbol{x}_i'\boldsymbol{\beta})y_i}}{1 + e^{(\gamma+\boldsymbol{x}_i'\boldsymbol{\beta})}} && \text{assume independence of } y_i \\
&= \frac{\prod_{i=1}^{N} e^{(\gamma+\boldsymbol{x}_i'\boldsymbol{\beta})y_i}}{\prod_{i=1}^{N}[1 + e^{(\gamma+\boldsymbol{x}_i'\boldsymbol{\beta})}]} \\
&= \frac{e^{\sum_{i=1}^{N}(\gamma+\boldsymbol{x}_i'\boldsymbol{\beta})y_i}}{\prod_{i=1}^{N}[1 + e^{(\gamma+\boldsymbol{x}_i'\boldsymbol{\beta})}]} \\
&= \frac{e^{\gamma\sum_{i=1}^{N} y_i + \sum_{i=1}^{N}(\boldsymbol{x}_i'\boldsymbol{\beta})y_i}}{\prod_{i=1}^{N}[1 + e^{(\gamma+\boldsymbol{x}_i'\boldsymbol{\beta})}]}
\end{aligned}$$

5.2 Conditional Likelihood Function

Define the set $G_t = \{\boldsymbol{y} \in \mathbb{R}^{dim(\boldsymbol{y})}; \boldsymbol{y}'\boldsymbol{1} = t\}$

$$\begin{aligned}
L(\boldsymbol{y}|\boldsymbol{y}'\boldsymbol{1} = t; \boldsymbol{\beta}, \gamma) &= \frac{Pr(\boldsymbol{y} \cap \boldsymbol{y}'\boldsymbol{1} = t)}{Pr(\boldsymbol{y}'\boldsymbol{1} = t)} && \text{conditional probability} \\
&= \frac{Pr(\boldsymbol{y} \cap \boldsymbol{y}'\boldsymbol{1} = t)}{Pr(\cup_{\boldsymbol{z}\in\mathbb{R}^{dim(\boldsymbol{z})}}\{\boldsymbol{z} \cap \boldsymbol{z}'\boldsymbol{1} = t\})} && \text{total probability} \\
&= \frac{Pr(\boldsymbol{y} \cap \boldsymbol{y}'\boldsymbol{1} = t)}{\sum_{\boldsymbol{z}\in\mathbb{R}^{dim(\boldsymbol{z})}} Pr(\boldsymbol{z} \cap \boldsymbol{z}'\boldsymbol{1} = t)} && \text{no intersection of sets} \\
&= \frac{Pr(\boldsymbol{y} \cap \boldsymbol{y}'\boldsymbol{1} = t)}{\sum_{\boldsymbol{z}\in G_t} Pr(\boldsymbol{z} \cap \boldsymbol{z}'\boldsymbol{1} = t)} \\
&= \frac{\frac{e^{\gamma t + \sum_{i=1}^{N}(\boldsymbol{x}_i'\boldsymbol{\beta})y_i}}{\prod_{i=1}^{N}[1+e^{(\gamma+\boldsymbol{x}_i'\boldsymbol{\beta})}]}}{\sum_{\boldsymbol{z}\in G_t}\frac{e^{\gamma t + \sum_{i=1}^{N}(\boldsymbol{x}_i'\boldsymbol{\beta})z_i}}{\prod_{i=1}^{N}[1+e^{(\gamma+\boldsymbol{x}_i'\boldsymbol{\beta})}]}} \\
&= \frac{e^{\gamma t + \sum_{i=1}^{N}(\boldsymbol{x}_i'\boldsymbol{\beta})y_i}}{\sum_{\boldsymbol{z}\in G_t} e^{\gamma t + \sum_{i=1}^{N}(\boldsymbol{x}_i'\boldsymbol{\beta})z_i}} \\
&= \frac{e^{\gamma t} e^{\sum_{i=1}^{N}(\boldsymbol{x}_i'\boldsymbol{\beta})y_i}}{e^{\gamma t}\sum_{\boldsymbol{z}\in G_t} e^{\sum_{i=1}^{N}(\boldsymbol{x}_i'\boldsymbol{\beta})z_i}} \\
&= \frac{e^{\sum_{i=1}^{N}(\boldsymbol{x}_i'\boldsymbol{\beta})y_i}}{\sum_{\boldsymbol{z}\in G_t} e^{\sum_{i=1}^{N}(\boldsymbol{x}_i'\boldsymbol{\beta})z_i}}
\end{aligned}$$

Parameter γ is gone.

6 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.

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

7 Semi-parametric Estimation

7.1 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 the number of correct observations, 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

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

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.

7.2 Smooth Maximum Score Estimation (Horowitz, 1992)

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

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

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

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 $F^{(2)}$. Repeat until convergence of $\boldsymbol{\beta}$

8 References

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

Davidson, R., & MacKinnon, J. G. (1993). Estimation and Inference in Econometrics.

Ruud, P. A. (2000). An Introduction to Classical Econometric Theory.