Notes on Binary Outcome Model

Max Leung

June 8, 2024

The theories underlying binary outcome model were well developed in the second half of the last century. I tried to synthesize those in this note. Many of these are covered in modern PhD level econometrics textbook e.g., Hansen (2022) (disclaimer I don't have a PhD degree).

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, which becomes Iteratively Reweighted Least Squares (IRLS) if $F(.)$ is Logistic function. $ln[L_N(\boldsymbol{\beta})]$ is globally concave because its expected Hessian matrix is negative definite. First order condition becomes sufficient condition and there is unique maximizer. Moreover, 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 i.e. $\widehat{\boldsymbol{\beta}}_{mle} \to_p \boldsymbol{\beta}_0$ 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 (e.g., p.121 of Amemiya, 1985), 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 $\mathbb{E}(y_i|\boldsymbol{x}_i) = 1 \cdot Pr(y_i = 1|\boldsymbol{x}_i) + 0 \cdot Pr(y_i = 0|\boldsymbol{x}_i) = Pr(y_i = 1|\boldsymbol{x}_i) = F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)$
$Var(y_i|\boldsymbol{x}_i) = \mathbb{E}(y_i^2|\boldsymbol{x}_i) - \mathbb{E}(y_i|\boldsymbol{x}_i)^2 = 1^2 \cdot Pr(y_i = 1|\boldsymbol{x}_i) + (1 \cdot Pr(y_i = 1|\boldsymbol{x}_i))^2 = Pr(y_i = 1|\boldsymbol{x}_i) + Pr(y_i = 1|\boldsymbol{x}_i)^2 = Pr(y_i = 1|\boldsymbol{x}_i)(1 - Pr(y_i = 1|\boldsymbol{x}_i))$

For binary outcome model, even not all the regularity conditions for Information Matrix equality hold, the equality still holds algebraically given $Pr(y_i = 1|\boldsymbol{x}_i) = F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)$. i.e., $\boldsymbol{A} = -\boldsymbol{B}$. It can be shown

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

Theorem 4.1.3 of Amemiya (1985 p.111) provides sufficient conditions for the asymptotic normality of extremum estimator. Replacing $Q_N(.)$ in the theorem with $lnL_N(.)$ reduces to MLE (In this general setting, Information Matrix may not hold for general model), $\sqrt{N}(\widehat{\boldsymbol{\beta}}_{mle} - \boldsymbol{\beta}_0) \to_d N(\boldsymbol{0}, \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1})$. For binary outcome model, the asymptotic distribution is $N(\boldsymbol{0}, -\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{A}^{-1}) = N(\boldsymbol{0}, -\boldsymbol{A}^{-1}) = N(\boldsymbol{0}, \boldsymbol{I}(\boldsymbol{\beta_0})^{-1})$. 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.

However, $\boldsymbol{A} = -\boldsymbol{B}$ may not hold if $Pr(y_i = 1|\boldsymbol{x}_i) \neq F(\boldsymbol{x}_i'\boldsymbol{\beta}_0)$ i.e., the CEF is incorrectly specified. Wrong specification of CEF will lead to inconsistency of $\widehat{\boldsymbol{\beta}}_{mle}$ as discussed previously. Thus, if CEF is wrongly specified, using sandwitch standard error can provide a correct estimate of asymptotic variance, but $\widehat{\boldsymbol{\beta}}_{mle}$ is still inconsistent.

1.2.7 Quasi Maximum Likelihood Estimation

If the density function in the likelihood function is misspecified, the resulting MLE or QMLE converges in probability to $\boldsymbol{\theta}^* = argmin_{\boldsymbol{\theta}} KLIC(f, g; \boldsymbol{\theta}) = argmax_{\boldsymbol{\theta}} \mathbb{E}_g[lnL(\boldsymbol{\theta})]$. Such $\boldsymbol{\theta}^*$ may or may not be equal to the true population parameter $\boldsymbol{\theta}_0$ (White, 1982). In the case of binary outcome model, y_i must be Bernoulli distributed. However, its parameter $p_i = F(\boldsymbol{x}_i'\boldsymbol{\beta})$ can be misspecified and thus the resulting MLE may not converge to the parameter of interest $\boldsymbol{\beta}_0$.

Ruud (1983) shows that if $\mathbb{E}_g[x_j|z]$ is a linear function of $z = \alpha + \boldsymbol{x}'\boldsymbol{\beta}$ for $\forall j \in \{2, \cdots, p\}$ (e.g., if $\boldsymbol{x}$ follows multivariate normal), we have $\boldsymbol{\beta}^* = scalar \cdot \boldsymbol{\beta}_0$. Therefore, the QMLE converges in probability to a scaled true population parameter.

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

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

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 (see Appendix for the details).

Repeat above two steps until convergence of $\widehat{\boldsymbol{\beta}}_{wnls}^{t}$, 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.

The following R code shows that IWNLS estimates of Logit model (F is logistic function) converges to MLE.

```r
binary_model_iwnls <- function (formula_, start_, data_, iter_ = 1000, tol_ = 1e-5) {
  b <- list()
  for (i in seq_len(iter_)) {
    if (i == 1) {
      m <-
        nls(
          as.formula(formula_),
          data = data_,
          start = start_
        )
    } else {
      m <-
        nls(
          as.formula(formula_),
          data = data_,
          start = start_,
          weights = w
        )
    }
    b[[i]] <- summary(m)$coefficients[ ,1]
    if (i > 1 && all(abs(b[[i]] - b[[i - 1]]) < tol_)) {
      return(list(model = m, number_of_iter = i))
    }
    p_hat <- predict(m, type = "response")
    w <- 1 / (p_hat * (1 - p_hat))
  }
  warning("not converge")
  return(list(model = m, number_of_iter = i))
}

logit_model_iwnls <-
  binary_model_iwnls(
    "y ~ 1 / (1 + exp(- b0 - b1 * x1 - b2 * x2))",
    start_ = list(b0 = 0.5, b1 = 0.5, b2 = 0.2),
    data_ = d
  )[["model"]]
summary(logit_model_iwnls)

# check equivalence
# maximum likelihood estimation of logit model with newton method = iterative re-weighted least squares
summary(glm(y ~ x1 + x2, family = binomial(link = "logit"), data = d))
```

1.4 GMM Estimation

1.4.1 Unconditional Moments

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

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

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 $\boldsymbol{S} = \mathbb{E}[\boldsymbol{g}(y_i,\boldsymbol{x}_i;\boldsymbol{\beta}_0)\boldsymbol{g}(y_i,\boldsymbol{x}_i;\boldsymbol{\beta}_0)']$. $\boldsymbol{S}$ 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.

The following R codes shows that GMM estimates of Logit model (F is logistic function) are equal to MLE.

```
library(gmm)

logit_model_gmm <- function (y_, X_, init_par_) {
```

```
  if (!is.numeric(y_)) stop("y_ should be a numeric vector")
  if (!is.matrix(X_)) stop("X_ should be a matrix")
  if (!is.numeric(init_par_)) stop("init_par_ should be a numeric vector")
  if (length(init_par_) != ncol(X_) + 1) stop("wrong init_par_")

  x <- cbind(y_, X_)
  g <- function (theta, x) {
    y <- x[ ,1]
    X <- cbind(1, x[ ,-1])
    b <- theta[1:ncol(X)]
    p <- as.numeric(1 / (1 + exp(- X %*% b)))

    # the first derivative of logistic function is F(1 - F)
    # this leads to simplification i.e., cancel out the weight 1 / F(1 - F)
    (y - p) * X
  }

  # binary model is just-identified, thus G is a square matrix,
  # if G is invertible, and W is chosen to be invertible e.g., I or S^-1,
  # the FOC of GMM reduces to FOC of MM i.e.,
  # G(theta_hat)' W g(theta_hat) = 0 <=> g(theta_hat) = 0
  # therefore, default two-step with optimal weight matrix (W = S^-1 from first step)
  # and one-step with identity matrix (W = I) have the same FOC g(theta_hat) = 0
  # and thus yield the same estimates
  # note g is the sample average of g_i defined in note
  gmm(g = g, x = x, t0 = init_par_, wmatrix = "ident")
}

y <- d[["y"]]
X <- as.matrix(d[c("x1", "x2")])
summary(logit_model_gmm(y, X, c(0.5, 0.5, 0.2)))

# vs mle, same coefficient estimates
summary(glm(y ~ x1 + x2, family = binomial(link = "logit"), data = d))
```

1.5 Special Case: Logit Model

If $F(.) = \Lambda(.)$ i.e., Logistic function (the c.d.f. of standard Logistic random variable),

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

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

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

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

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

Odds Ratio (OR) can be interpreted as

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

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

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

One line of code to run probit model in R.

```
summary(glm(y ~ x1 + x2, family = binomial(link = "probit"), data = d))
```

1.7 Special Case: Robit Model (Liu, 2004)

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.

The below R code shows the ML estimation of Robit model.

```
robit_mle <- function (
    y_,
    X_,
    init_par_,
    df_ = NULL,
    intercept_ = TRUE,
    method_ = "BFGS"
) {

  if (!is.numeric(y_)) stop("y_ should be a numeric vector")
```

```r
  if (!is.matrix(X_)) stop("X_ should be a matrix")
  if (intercept_) X_ <- cbind(1, X_)
  if (!is.numeric(init_par_)) stop("init_par_ should be a numeric vector")
  if (is.null(df_) && (length(init_par_) != ncol(X_) + 1)) stop("wrong init_par_")
  if (!is.null(df_) && (length(init_par_) != ncol(X_))) stop("wrong init_par_")

  negative_lnL <- function (par, y, X) {
    beta <- par[1:ncol(X)]

    if (is.null(df_)) {
      df <- par[ncol(X) + 1]
    } else if (!is.na(df_) && is.numeric(df_) && df_ > 0) {
      df <- df_
    } else {
      stop("df_ should be positve real number")
    }

    t_cdf <- pt(X %*% beta, df = df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
    -sum(y * log(t_cdf) + (1 - y) * log(1 - t_cdf))
  }

  optim(
    par = init_par_,
    fn = negative_lnL,
    y = y_, X = X_,
    method = method_,
    hessian = TRUE
  )
}

y <- d[["y"]]
X <- as.matrix(d[c("x1", "x2")])
robit_mle(y, X, c(0.5, 0.5, 0.2, 4))
```

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.

```
library(lmtest)
library(sandwich)

lpm <- lm(y ~ ., data = d)

# HC0 = White's estimator
# it is robust because of heteroscedasticity consistent
# however, it is biased when there is homoscedasticity
# HC2 adjustment is unbiased under homoscedasticity but biased under heteroscedasticity
coeftest(lpm, vcov = vcovHC(lpm, type = "HC0"))
```

Some scholars e.g., Angrist and Pischke (2009) 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.

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

Imbens and Rubin (2015) provide a theorem (7.1i) saying the indicator coefficient of a linear model identifies ATE even with covariates under random assignment and random sampling. Proof

$$\begin{aligned}
Q(\alpha, \beta, \boldsymbol{\theta}) &:= \mathbb{E}[(Y_i - \alpha - \beta D_i - \boldsymbol{x}_i'\boldsymbol{\theta})^2] \\
&= \mathbb{E}[(Y_i - \alpha - \mathbb{E}[\boldsymbol{x}_i]'\boldsymbol{\theta} - \beta D_i - \boldsymbol{x}_i'\boldsymbol{\theta} + \mathbb{E}[\boldsymbol{x}_i]'\boldsymbol{\theta})^2] \\
&= \mathbb{E}[(Y_i - (\alpha + \mathbb{E}[\boldsymbol{x}_i]'\boldsymbol{\theta}) - \beta D_i - (\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i])'\boldsymbol{\theta})^2] \\
&= \mathbb{E}[(Y_i - \widetilde{\alpha} - \beta D_i - (\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i])'\boldsymbol{\theta})^2] \\
&= \mathbb{E}[(Y_i - \widetilde{\alpha} - \beta D_i)^2] - \mathbb{E}[((\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i])'\boldsymbol{\theta})^2] - 2\mathbb{E}[(Y_i - \widetilde{\alpha} - \beta D_i)((\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i])'\boldsymbol{\theta})] \\
&= \mathbb{E}[(Y_i - \widetilde{\alpha} - \beta D_i)^2] - \mathbb{E}[((\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i])'\boldsymbol{\theta})^2] - 2\mathbb{E}[Y_i(\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i])'\boldsymbol{\theta}] && \text{see below}
\end{aligned}$$

Thus, $\frac{\partial Q}{\partial \beta}|_{\beta^*} = \frac{\partial \mathbb{E}[(Y_i - \widetilde{\alpha} - \beta D_i)^2]}{\partial \beta}|_{\beta^*} = 0$ reduces to the case without covariates. Therefore, as shown before, $\beta^* = ATE$.

$$\begin{aligned}
&\mathbb{E}[\widetilde{\alpha}((\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i])'\boldsymbol{\theta})] = \widetilde{\alpha}(\mathbb{E}[\boldsymbol{x}_i] - \mathbb{E}[\boldsymbol{x}_i])'\boldsymbol{\theta} = 0 \\
&\mathbb{E}[\beta D_i((\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i])'\boldsymbol{\theta})] = \beta\mathbb{E}[D_i(\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i])']\boldsymbol{\theta} = \beta(\mathbb{E}[D_i\boldsymbol{x}_i'] - \mathbb{E}[D_i\mathbb{E}[\boldsymbol{x}_i]'])\boldsymbol{\theta} = \beta(\mathbb{E}[D_i\boldsymbol{x}_i'] - \mathbb{E}[D_i]\mathbb{E}[\boldsymbol{x}_i]')\boldsymbol{\theta} = 0
\end{aligned}$$

The last equality is due to the independence of D_i and $\boldsymbol{x}_i \implies \mathbb{E}[D_i\boldsymbol{x}_i'] = \mathbb{E}[D_i]\mathbb{E}[\boldsymbol{x}_i]'$ given random assignment in large sample. It may not hold in small sample.

Average derivatives (Stoker, 1986) provide the average marginal effect without specifying $F(.)$. Define $\mathbb{E}[y|\boldsymbol{x}] = m(\boldsymbol{x})$,

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

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

LPM can also be justified on its analogy with linear discriminant function (not linear discriminant analysis LDA). Set $\widehat{y}_i = 1(\widehat{Pr}(y_i = 1|\boldsymbol{x}_i) > threshold)$, where $\widehat{Pr}(y_i = 1|\boldsymbol{x}_i) = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{lpm}$. If the threshold is set at $(\bar{\boldsymbol{x}}_1'\widehat{\boldsymbol{\beta}}_{lpm} + \bar{\boldsymbol{x}}_2'\widehat{\boldsymbol{\beta}}_{lpm})/2$, LPM provides the same prediction with linear discriminant function. Since linear discriminant function does not require any distributional or functional assumption, even the true CEF is non-linear i.e., our specification is wrong, the prediction from LPM is still justified. Linear discriminant function is $\boldsymbol{x}_i'\widehat{\boldsymbol{\lambda}}$ where $\widehat{\boldsymbol{\lambda}} = arg\ max_{\boldsymbol{\lambda}} \frac{\boldsymbol{\lambda}'(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)^2}{\boldsymbol{\lambda}'\boldsymbol{S}\boldsymbol{\lambda}}$. This gives $\widehat{\boldsymbol{\lambda}} = \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$. Given new observation $\boldsymbol{x}_0$, it will be classify as group 1 if $\widehat{\boldsymbol{\lambda}}'\boldsymbol{x}_0 > \frac{1}{2}(\widehat{\boldsymbol{\lambda}}'\bar{\boldsymbol{x}}_1 + \widehat{\boldsymbol{\lambda}}'\bar{\boldsymbol{x}}_2)$. it can be shown $\widehat{\boldsymbol{\beta}}_{lpm} = \widehat{\boldsymbol{\lambda}}\frac{SSErr_{lpm}}{N_1+N_2-2}$ (see Maddala, 1983 for proof). Thus,

$$\begin{aligned} \widehat{y}_i &= 1(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{lpm} > (\bar{\boldsymbol{x}}_1'\widehat{\boldsymbol{\beta}}_{lpm} + \bar{\boldsymbol{x}}_2'\widehat{\boldsymbol{\beta}}_{lpm})/2) \\ &= 1(\boldsymbol{x}_i'\widehat{\boldsymbol{\lambda}}\frac{SSErr_{lpm}}{N_1+N_2-2}) > \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\widehat{\boldsymbol{\lambda}}\frac{SSErr_{lpm}}{N_1+N_2-2}) \\ &= 1(\boldsymbol{x}_i'\widehat{\boldsymbol{\lambda}} > \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\widehat{\boldsymbol{\lambda}}) \end{aligned}$$

1.9 Model Evaluation

1.9.1 Accuracy

Accuracy $= N^{-1}\sum_{i=1}^{N} 1(1\{\widehat{Pr}(y_i = 1|\boldsymbol{x}_i) \geq threshold\} = y_i)$. The threshold is usually set at 0.5. If the data used to compute the accuracy are the data used to estimate the model, it is called in-sample accuracy. If new data are used, it is called out-of-sample accuracy.

```
logit_model <- glm(y ~ x1 + x2, family = binomial(link = "logit"), data = d)
p_hat <- predict(logit_model, type = "response")

mean(as.numeric(as.numeric(p_hat >= 0.5) == y), na.rm = TRUE)
```

1.9.2 Confusion Matrix and ROC

It is actually not confusing/ed. It partitions the sample into True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). Precision $= \frac{TP}{Predict\ is\ positive} = \frac{TP}{TP+FP}$, Recall $= \frac{TP}{Actual\ is\ positive} = \frac{TP}{TP+FN}$. In general, Precision and Recall is negatively related.

ROC plots Recall $\frac{TP}{TP+FN}$ (y-axis) versus Fall-out $\frac{FP}{TN+FP}$ (x-axis). Recall and Fall-out must be positively related. Model increases Recall with a smaller increase in Fall-out is superior.

```
library(pROC)

logit_model <- glm(y ~ x1 + x2, family = binomial(link = "logit"), data = d)
p_hat <- predict(logit_model, type = "response")

table(
  actual = y,
  prediction = as.numeric(p_hat >= 0.5)
)

roc(y ~ x1 + x2, data = d, na.rm = TRUE, plot = TRUE)
```

1.9.3 Akaike Information Criterion (AIC)

Claeskens and Hjort (2008) provide a robust Aic formula for generalized linear model.

$$\text{AIC} = 2 \cdot ln[L(\widehat{\boldsymbol{\beta}}_{mle})] - 2 \cdot \alpha$$

where $\alpha = \sum_{i=1}^{N} \mathbb{E}(y_i - \mathbb{E}[y_i|\boldsymbol{x}_i])\widehat{\theta}_i/\widehat{\eta}$. Since $\theta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ and $\eta = 1$ for Logit model, $\alpha = \sum_{i=1}^{N} \mathbb{E}(y_i - \overbrace{\mathbb{E}[y_i|\boldsymbol{x}_i]}^{Pr(y_i=1|\boldsymbol{x}_i)})\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{mle}$. α can be estimated by bootstrap method.

Estimate the model and get $\widehat{Pr(y_i = 1|\boldsymbol{x}_i)}$ and $y_i^* = 1(\widehat{Pr(y_i = 1|\boldsymbol{x}_i)} \geq 0.5)$ for $\forall i$.
For $b = 1$ to B
Draw N random samples with replacement from original dataset. Define index set $\mathbb{I}_b$ where $|\mathbb{I}_b| = N$.
Use bootstrapped dataset $\{y_j^*, \boldsymbol{x}_j\}_{j\in\mathbb{I}_b}$ to estimate a Logit model and get $\widehat{\boldsymbol{\beta}}_{mle}^*$.
Calculate $\widehat{\alpha}_b = \sum_j (y_j^* - \widehat{Pr(y_j = 1|\boldsymbol{x}_j)})\boldsymbol{x}_j'\widehat{\boldsymbol{\beta}}_{mle}^*$.

Thus, we have B number of $\widehat{\alpha}_b$ for a model. AIC of this model is $2 \cdot ln[L(\widehat{\boldsymbol{\beta}}_{mle})] - 2 \cdot \sum_{b=1}^{B} \widehat{\alpha}_b/B$. The original AIC formula sets α equals to the number of parameter, which is easier to compute. However, it can be less accurate if the true model is not the specified model.

The following R code computes the above algorithm.

```
logit_model <- glm(y ~ x1 + x2, family = binomial(link = "logit"), data = d)

alpha_hat <- list()
p_hat <- predict(logit_model, type = "response")
y_star <- as.numeric(p_hat >= 0.5)
B <- 5000
for (b in seq_len(B)) {
  boot_index <- sample(nrow(d), replace = TRUE)
  boot_y_star <- y_star[boot_index]
  boot_p_hat <- p_hat[boot_index]
  boot_d <- cbind(boot_y_star, d[boot_index,c("x1", "x2")])
  boot_logit_model <-
    glm(
      y ~ x1 + x2,
      family = binomial(link = "logit"),
      data = boot_d
    )
  alpha_hat[[b]] <-
    t(boot_y_star - boot_p_hat) %*%
    as.matrix(cbind(1, boot_d[c("x1", "x2")])) %*%
    as.matrix(boot_logit_model$coefficients)
}

alpha_hat_bar <- mean(as.numeric(alpha_hat), na.rm = TRUE)
as.numeric(boot_aic <- 2 * logLik(logit_model) - 2 * alpha_hat_bar)
```

1.9.4 McFadden (1974) Pseudo-R^2

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

With binary dependent variable, $lnL_0 \leq lnL_{fit} \leq 0$ means $0 \leq R_{mcf}^2 \leq 1$.

```
logit_model <- glm(y ~ x1 + x2, family = binomial(link = "logit"), data = d)
mc_fadden_r2 <- 1 - as.numeric(logLik(logit_model) / logLik(update(logit_model, formula = y ~ 1)))
```

1.9.5 Cox & Snell (1970) Pseudo-R^2

This is from Statistics.

$$R_{cs}^2 = 1 - exp(-\frac{2}{N}(lnL_{fit} - lnL_0)) = 1 - (\frac{L_0}{L_{fit}})^{2/N}$$

With binary dependent variable, $lnL_0 \leq lnL_{fit} \leq 0$ means $0 \leq R^2_{cs} \leq 1 - exp(\overbrace{\frac{2}{N} lnL_0}^{\leq 0}) < 1$. Thus, it is less than one (lnL_0 cannot be negative infinity). For normal linear regression with homoscedasticity and known variance (not necessarily binary dependent variable), we have

$$\begin{aligned} R^2_{cs} &= 1 - exp[-\frac{2}{N}(-\frac{1}{2\sigma^2})(\sum_{i=1}^{N}(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{mle})^2 - \sum_{i=1}^{N}(y_i - \bar{y})^2)] \\ &= 1 - exp[\frac{1}{N\sigma^2}(SSError - SSTotal)] = 1 - exp[\frac{1}{N\sigma^2}(-SSReg)] \\ &\approx \frac{1}{N\sigma^2} SSReg \approx \frac{1}{N \cdot \sum_{i=1}^{N}(y_i - \bar{y})^2/N} SSReg = \frac{SSReg}{SST} = R^2 \end{aligned}$$

It applies Taylor first order approximation centered at zero i.e., $1 - e^z \approx -z$. We also substitute σ^2 with $\widehat{\sigma}^2 = \sum_{i=1}^{N}(y_i - \bar{y})^2/N$

1.9.6 Cragg & Uhler (1970) Pseudo-R^2

Since Cox & Snell's R^2 is smaller than one, it is easy to adjust this by dividing its maximum value.

$$\begin{aligned} R^2_{cu} &= R^2_{cs}/(1 - exp(\frac{2}{N} lnL_0)) \\ &= \frac{1 - (\frac{L_0}{L_{fit}})^{2/N}}{1 - L_0^{2/N}} \end{aligned}$$

Many authors refer this as Nagelkerke (1991) Pseudo-R^2. However, it can at least date back to Cragg and Uhler (1970) published on The Canadian Journal of Economics. Maddala (1983) cited this correctly. Wiki also said Pseudo-R^2 on Cragg and Uhler (1970) and Nagelkerke (1991) are the same.

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

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

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,

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

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

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

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 $\Phi(.)$.

Implementation in R.

```
probit_model <- glm(y ~ x1 + x2, family = binomial(link = "probit"), data = d)

xb_hat <- predict(probit_model, type = "link")
p_hat <- predict(probit_model, type = "response")
e_g <- (d[["y"]] - p_hat) * dnorm(xb_hat) / (p_hat * (1 - p_hat))

opg_d <-
  data.frame(
    y = 1,
    e_g_x = e_g * as.matrix(cbind(1, d[c("x1", "x2")])),
    e_g_xb_square = e_g * xb_hat^2,
    e_g_xb_cube = e_g * xb_hat^3
  )

alpha_lm <- nrow(opg_d) * summary(lm(y ~ . -1, data = opg_d))$r.squared
p_value <- pchisq(alpha_lm, df = 2, lower.tail = FALSE, ncp = 0, log.p = FALSE)
if (p_value < 0.01) print("reject gamma_1 = gamma_2 = 0 <=> reject u_i is standard normal")
```

4 Generalized Linear Model (GLM)

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 $\theta = \boldsymbol{x}'\boldsymbol{\beta}$, we have $\mathbb{E}(y) = b'(\boldsymbol{x}'\boldsymbol{\beta})$. 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 $\mathbb{E}(y) = logistic(\boldsymbol{x}'\boldsymbol{\beta})$ by setting $n = 1$ and $\theta = \boldsymbol{x}'\boldsymbol{\beta}$. 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

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

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

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

First order Taylor expansion of $F^{-1}(\bar{p}_t)$ centered at p_t,

$$F^{-1}(\bar{p}_t) \approx F^{-1}(p_t) + \frac{dF^{-1}(\bar{p}_t)}{d\bar{p}_t}|_{p_t}(\bar{p}_t - p_t) = F^{-1}(p_t) + \frac{1}{f(p_t)}(\bar{p}_t - p_t) \iff v_t = F^{-1}(\bar{p}_t) - F^{-1}(p_t) = \frac{\bar{p}_t - p_t}{f(p_t)}$$
$$Var(v_t) \approx Var(\frac{\bar{p}_t - p_t}{f(p_t)}) = \frac{1}{(f(p_t))^2}Var(\bar{p}_t) = \frac{1}{(f(p_t))^2}\frac{p_t(1-p_t)}{N_t} = \frac{p_t(1-p_t)}{(f(p_t))^2 N_t}$$

Since $Var(\bar{p}_t) = N_t^{-2}Var(\sum_{i=1}^{N_t} y_{it}) = N_t^{-2}\cdot N_t p_t(1-p_t) = \frac{p_t(1-p_t)}{N_t}$.

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

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

6.1 Special Case: $F(.)$ is an identity function

$$\mathbb{E}(v_t) = \mathbb{E}(\bar{p}_t - p_t) = N_t^{-1}\mathbb{E}(\sum_{i=1}^{N_t} y_{it}) - p_t = N_t^{-1}\cdot N_t p_t - p_t = 0$$
$$Var(v_t) = \frac{p_t(1-p_t)}{1^2 N_t} = \frac{p_t(1-p_t)}{N_t}$$

$$\widehat{\boldsymbol{\beta}}_{chi} = arg\ min_{\boldsymbol{\beta}} \sum_{t=1}^{T} \frac{N_t}{\bar{p}_t(1-\bar{p}_t)}(\bar{p}_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2 = arg\ min_{\boldsymbol{\beta}} \sum_{t=1}^{T} [\sqrt{\frac{N_t}{\bar{p}_t(1-\bar{p}_t)}}\bar{p}_t - (\sqrt{\frac{N_t}{\bar{p}_t(1-\bar{p}_t)}}\boldsymbol{x}_t)'\boldsymbol{\beta}]^2$$

It is called minimum chi-square method.

6.2 Special Case: $F(.)$ is a logistic function

$$Var(v_t) \approx \frac{p_t(1-p_t)}{[p_t(1-p_t)]^2 N_t} = \frac{1}{N_t p_t(1-p_t)}$$

$$\widehat{\boldsymbol{\beta}}_{chi} = arg\ min_{\boldsymbol{\beta}} \sum_{t=1}^{T} N_t\bar{p}_t(1-\bar{p}_t)(logit(\bar{p}_t) - \boldsymbol{x}_t'\boldsymbol{\beta})^2 = arg\ min_{\boldsymbol{\beta}} \sum_{t=1}^{T} [\sqrt{N_t\bar{p}_t(1-\bar{p}_t)}logti(\bar{p}_t) - (\sqrt{N_t\bar{p}_t(1-\bar{p}_t)}\boldsymbol{x}_t)'\boldsymbol{\beta}]^2$$

It is called minimum logit chi-square method. It can be shown

$$\sqrt{T}(\widehat{\boldsymbol{\beta}}_{chi} - \boldsymbol{\beta}_0) \to_d N(\boldsymbol{0}, [\sum_{t=1}^{T} \frac{N_t exp(\boldsymbol{x}_t'\boldsymbol{\beta}_0)}{(1+exp(\boldsymbol{x}_t'\boldsymbol{\beta}_0))^2}\boldsymbol{x}_t\boldsymbol{x}_t']^{-1})$$

6.3 Special Case: $F(.)$ is standard normal cdf

$$Var(v_t) \approx \frac{p_t(1-p_t)}{(\phi(p_t))^2 N_t}$$

$$\widehat{\boldsymbol{\beta}}_{chi} = arg\ min_{\boldsymbol{\beta}} \sum_{t=1}^{T} \frac{(\phi(\bar{p}_t))^2 N_t}{\bar{p}_t(1-\bar{p}_t)}(\Phi^{-1}(\bar{p}_t) - \boldsymbol{x}_t'\boldsymbol{\beta})^2 = arg\ min_{\boldsymbol{\beta}} \sum_{t=1}^{T} [\sqrt{\frac{(\phi(\bar{p}_t))^2 N_t}{\bar{p}_t(1-\bar{p}_t)}}\Phi^{-1}(\bar{p}_t) - (\sqrt{\frac{(\phi(\bar{p}_t))^2 N_t}{\bar{p}_t(1-\bar{p}_t)}}\boldsymbol{x}_t)'\boldsymbol{\beta}]^2$$

7 Semi-parametric Estimation

7.1 Maximum Score Estimation (Manski, 1975)

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

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

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 Semi-parametric MLE (Klein & Spady, 1993)

7.2.1 Algorithm

For $t = 1 \cdots$

Use $\boldsymbol{\beta}^{(t)}$, estimate $F^{(t)}$ by local constant (Nadaraya–Watson) estimator,

$$F^{(t)}(\boldsymbol{x}'\boldsymbol{\beta}^{(t)}) = \frac{\sum_{j=1}^N y_j K((\boldsymbol{x}_j - \boldsymbol{x})'\boldsymbol{\beta}^{(t)}/h)}{\sum_{j=1}^N K((\boldsymbol{x}_j - \boldsymbol{x})'\boldsymbol{\beta}^{(t)}/h)}$$

Given $F^{(t)}$ and $\boldsymbol{\beta}^{(t)}$, compute $\boldsymbol{\beta}^{(t+1)}$ by gradient descent method i.e., $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \alpha \frac{\partial Q_N^{(t)}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}^{(t)}}$ where

$$\frac{\partial Q_N^{(t)}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}^{(t)}} = \sum_{i=1}^N \frac{y_i - F^{(t)}(\boldsymbol{x}_i'\boldsymbol{\beta}^{(t)})}{F^{(t)}(\boldsymbol{x}_i'\boldsymbol{\beta}^{(t)})(1 - F^{(t)}(\boldsymbol{x}_i'\boldsymbol{\beta}^{(t)}))} F^{(t)'}(\boldsymbol{x}_i'\boldsymbol{\beta}^{(t)})\boldsymbol{x}_i$$

Loop until convergence of $\boldsymbol{\beta}^{(t)}$.

The advantage of this approach is that we do not have to specify $F(.)$. However, the algorithm is slow by its nature. *np* package in R provides a convenient function for performing the above algorithm.

```
library(np)

# the first beta is normalized to one
summary(npindex(y ~ x1 + x2, method = "kleinspady", gradients = TRUE, data = d))
```

7.2.2 Asymptotic Distribution

Given certain regularity conditions,

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{KS} - \boldsymbol{\beta}_0) \to_d N(\mathbf{0}, \{\mathbb{E}[\frac{\partial F(\boldsymbol{x}'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}(\frac{\partial F(\boldsymbol{x}'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}})' \frac{1}{F(\boldsymbol{x}'\boldsymbol{\beta})(1 - F(\boldsymbol{x}'\boldsymbol{\beta}))}]\}^{-1})$$

8 References

Amemiya, T. (1985). Advanced Econometrics.

Angrist, J. D., & Pischke, J.-S. (2009). Mostly Harmless Econometrics: An Empiricist's Companion.

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

Claeskens, G., & Hjort, N. L. (2008). Model Selection and Model Averaging.

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

Imbens, G. W., & Rubin, D. B. (2015). Causal inference for statistics, social, and biomedical sciences: An introduction.

Klein, R. W., & Spady, R. H. (1993). An Efficient Semiparametric Estimator for Binary Response Models. Econometrica, 61(2), 387–421.

Liu, C. (2004). Robit Regression: A Simple Robust Alternative to Logistic and Probit Regression. In Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives (eds W.A. Shewhart, S.S. Wilks, A. Gelman and X.-L. Meng).

Maddala, G. (1983). Limited-Dependent and Qualitative Variables in Econometrics.

Manski, C. F. (1975). Maximum Score Estimation of the Stochastic Utility Model of Choice. Journal of Econometrics. 3(3), 205–228.

Ruud, P. A. (1983). Sufficient Conditions for the Consistency of Maximum Likelihood Estimation Despite Misspecification of Distribution in Multinomial Discrete Choice Models. Econometrica, 51(1), 225–228.

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

Stoker, T. M. (1986). Consistent Estimation of Scaled Coefficients. Econometrica, 54(6), 1461–1481.

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. Econometrica, 50(1), 1–25.

9 Appendix: Gauss-Newton Method

9.1 Objective Function

For non-linear model $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{\theta}_0) + \boldsymbol{\varepsilon}$,

$$f(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))' \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))$$

9.2 Approximate Newton Method

9.2.1 Gradient Vector

$$\begin{aligned}
\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{\partial (\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))' \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \\
&= \frac{\partial (\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))'}{\partial \boldsymbol{\theta}} (\boldsymbol{W} + \boldsymbol{W}') (\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta})) \\
&= -2 (\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'})' \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))
\end{aligned}$$

9.2.2 Hessian Matrix

$$\begin{aligned}
\frac{\partial^2 f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} &= -2 \frac{\partial (\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'})' \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}'} \\
&= -2 [\frac{\partial^2 \boldsymbol{g}(\boldsymbol{\theta})'}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta})) - (\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'})' \boldsymbol{W} \frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}] \\
&\approx 2 (\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'})' \boldsymbol{W} \frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}
\end{aligned}$$

The first term is discarded because $\mathbb{E}(\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta})) \to_p \boldsymbol{0}$. Proof: $\boldsymbol{\theta} \to_p \boldsymbol{\theta}_0$ when $N \to \infty$. By continuous mapping theorem, $\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}) \to_p \boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta_0}) = \boldsymbol{\varepsilon}$. By continuous mapping theorem again, $\mathbb{E}(\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta})) \to_p \mathbb{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$.

9.2.3 Apply Newton Method

$$\begin{aligned}
\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - [\frac{\partial^2 f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}_t}]^{-1} \frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}_t} \\
&\approx \boldsymbol{\theta}_t - [2 (\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}_t})' \boldsymbol{W} \frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}_t}]^{-1} [-2 (\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}_t})' \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))] \\
&= \boldsymbol{\theta}_t + [(\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}_t})' \boldsymbol{W} \frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}_t}]^{-1} [(\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}_t})' \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))]
\end{aligned}$$

It can also be written as

$$= \boldsymbol{\theta}_t - [(\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}_t})' \boldsymbol{W} \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}_t}]^{-1} [(\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}_t})' \boldsymbol{W} \boldsymbol{u}]$$

where $\boldsymbol{u} = \boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta})$.