Advanced Econometrics

Professor: Julie Shi

Timekeeper: Rui Zhou

Fall 2023

Contents

7 Maximum Likelihood Estimation

7.1 The Likelihood Function

The probability density function (p.d.f. in short), for a random variable y, conditioned on a set of parameters θ, is denoted $f(y|\theta)$.

The jointly density, or likelihood function is

$$f(y_1, \cdots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta) = L(\theta|y)$$

It is usually simpler to work with the log of the likelihood function:

$$\ln L(\theta|y) = \sum_{i=1}^{n} \ln f(y_i|\theta)$$

Note that we write the joint density as a function of the data conditioned on the parameters whereas when we form the likelihood function, we will write this function in reverse, as a function of the parameters, conditioned on the data. Though the two functions are the same, it is to be emphasized that the likelihood function is written in this fashion to highlight our interest in the parameters and the information about them that is contained in the observed data. However, it is understood that the likelihood function is not meant to represent a probability density for the parameters.

MLE estimation aims to choose θ to maximize log-likelihood function. The necessary condition is

$$\frac{\partial \ln L(\theta|y, X)}{\partial \theta} = 0$$

which is called the likelihood equation.

Definition (Identification) The parameter vector θ is identified (or estimable) if for any other parameter vector $\theta^* \neq \theta$, for some data y, $L(\theta^*|y) \neq L(\theta|y)$.

7.2 Properties of MLE

Commonly-used densities:

- Normal

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\cdot\left(\frac{y-\mu}{\sigma}\right)^2}$$

- Bernoulli

$$f(y) = p^y (1-p)^{1-y}$$

- Exponential

$$f(y) = \lambda e^{-\lambda y}$$

- Poisson

$$f(y) = \frac{e^{-\lambda}\lambda^y}{y!}$$

Example (Normal Distribution Case) In sampling from a normal distribution with mean μ and variance σ^2, the MLE estimator are

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}_n$$
$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y}_n)^2$$

Example (OLS Revisited)

By assumption,

$$\begin{aligned} f(\varepsilon_i) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right) \\ \Longrightarrow \ln f(\varepsilon_i) &= -\frac{\ln 2\pi}{2} - \frac{\ln \sigma^2}{2} - \frac{\varepsilon_i^2}{2\sigma^2} \\ \Longrightarrow \ln L(\beta, \sigma^2) &= -\frac{n\ln 2\pi}{2} - \frac{n\ln \sigma^2}{2} - \frac{(y - X\beta)'(y - X\beta)}{2\sigma^2} \end{aligned}$$

As we can see from the expression of $\ln L(\beta, \sigma^2)$, the determinant part is the same as minimizing least squares, $(y - X\beta)'(y - X\beta)$. By resorting to the first-order condition, we can get the same result:

$$\hat{\beta} = (X'X)^{-1}X'y$$
$$\hat{\sigma}^2 = \frac{\left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right)}{n}$$

Notice that the $\hat{\sigma}^2$ here is biased but consistent.

Under the regularity, the maximum likelihood estimator (MLE) has the following asymptotic properties:

- Consistency: plim $\widehat{\theta} = \theta_0$.
- Asymptotic normality: $\widehat{\theta} \overset{\text{a}}{\sim} N\left[\theta_0, \{I(\theta_0)\}^{-1}\right]$, where $I(\theta_0) = -\mathrm{E}_0\left[\dfrac{\partial^2 \ln L}{\partial\theta_0 \partial\theta_0'}\right]$.
- Asymptotic efficiency: $\widehat{\theta}$ is asymptotically efficient and achieves Cramer-Rao lower bound for consistent estimators.

An estimator is asymptotically efficient if it is consistent, asymptotically normally distributed, and has an asymptotic covariance matrix that is not larger than the asymptotic covariance matrix of any other consistent, asymptotically normally distributed estimator.

Notations

For each observations, we have log-density $\ln f(y_i|\theta)$. Denote

$$g_i = \frac{\partial \ln f(y_i|\theta)}{\partial\theta}$$
$$H_i = \frac{\partial^2 \ln f(y_i|\theta)}{\partial\theta\partial\theta'}$$

Then we have

$$g = \frac{\partial \ln L(\theta|y)}{\partial\theta} = \sum_{i=1}^{n} \frac{\partial \ln f(y_i|\theta)}{\partial\theta} = \sum_{i=1}^{n} g_i$$
$$H = \frac{\partial^2 \ln L(\theta|y)}{\partial\theta\partial\theta'} = \sum_{i=1}^{n} \frac{\partial^2 \ln f(y_i|\theta)}{\partial\theta\partial\theta'} = \sum_{i=1}^{n} H_i$$

Remarks: H is called *Hessian Matrix*.

7.3 Likelihood Equation

Likelihood equation

$$\mathrm{E}_0\left[\frac{\partial \ln L(\theta_0)}{\partial\theta_0}\right] = \mathrm{E}_0[g_0] = 0$$

Proof

$$\begin{aligned}
\mathrm{E}_0\left[g_{i0}\right] &= \mathrm{E}_0\left[\frac{\partial \ln f\left(y_i|\theta_0\right)}{\partial \theta_0}\right] \\
&= \int \frac{\partial \ln f\left(y_i|\theta_0\right)}{\partial \theta_0} \cdot f\left(y_i|\theta_0\right) \mathrm{d}y_i \\
&= \int \frac{1}{f\left(y_i|\theta_0\right)} \cdot \frac{\partial f\left(y_i|\theta_0\right)}{\partial \theta_0} \cdot f\left(y_i|\theta_0\right) \mathrm{d}y_i \\
&= \int \frac{\partial f\left(y_i|\theta_0\right)}{\partial \theta_0} \mathrm{d}y_i \\
&= \frac{\partial}{\partial \theta_0} \int f\left(y_i|\theta_0\right) \mathrm{d}y_i \\
&= \frac{\partial}{\partial \theta_0}\left(1\right) = 0
\end{aligned}$$

By definition, we have

$$g_0 = \sum_{i=1}^{n} g_{i0} = \sum_{i=1}^{n} \frac{\partial \ln f\left(y_i|\theta_0\right)}{\partial \theta_0} = \frac{\partial \ln L\left(\theta_0|y\right)}{\partial \theta_0}$$

Since we have proved that $\mathrm{E}_0\left[g_{i0}\right] = 0$, it follows that

$$\mathrm{E}_0\left[\frac{\partial \ln L\left(\theta_0\right)}{\partial \theta_0}\right] = \mathrm{E}_0\left[g_0\right] = \mathrm{E}_0\left[\sum_{i=1}^{n} g_{i0}\right] = \sum_{i=1}^{n} \mathrm{E}_0\left[g_{i0}\right] = 0$$

7.4 Information Matrix Equality

Information Matrix Equality

$$\begin{aligned}
\mathrm{Var}\left[\frac{\partial \ln L\left(\theta_0|y\right)}{\partial \theta_0}\right] &= \mathrm{E}_0\left[\left(\frac{\partial \ln L\left(\theta_0|y\right)}{\partial \theta_0}\right)\left(\frac{\partial \ln L\left(\theta_0|y\right)}{\partial \theta_0'}\right)\right] \\
&= -\mathrm{E}_0\left[\frac{\partial^2 \ln L\left(\theta_0|y\right)}{\partial \theta_0 \partial \theta_0'}\right]
\end{aligned}$$

Or equivalently, in our earlier notations, $\mathrm{Var}\left[g_0\right] = -\mathrm{E}_0\left[H_0\right]$.

Proof We only show a proof of this equality in the scalar case. First we have

$$\begin{aligned}\frac{\partial^2 \ln f\left(y_i|\theta\right)}{\partial\theta^2} &= \frac{\partial}{\partial\theta}\left(\frac{\partial \ln f\left(y_i|\theta\right)}{\partial\theta}\right)\\ &= \frac{\partial}{\partial\theta}\left(\frac{1}{f_i}\cdot\frac{\partial f_i}{\partial\theta}\right)\\ &= \frac{1}{f_i}\left(\frac{\partial^2 f_i}{\partial\theta^2}\right) - \left(\frac{1}{f_i}\right)^2\left(\frac{\partial f_i}{\partial\theta}\right)^2\\ &= \frac{1}{f_i}\left(\frac{\partial^2 f_i}{\partial\theta^2}\right) - \left(\frac{\partial \ln f_i}{\partial\theta}\right)^2\end{aligned}$$

Taking expectations on both sides, we have

$$\mathrm{E}\left[\frac{\partial^2 \ln f_i}{\partial\theta^2}\right] = \mathrm{E}\left[\frac{1}{f_i}\left(\frac{\partial^2 f_i}{\partial\theta^2}\right)\right] - \mathrm{E}\left[\left(\frac{\partial \ln f_i}{\partial\theta}\right)^2\right]$$

Compare such result with the equality we hope to prove, it suffices to show that

$$\begin{aligned}\mathrm{E}\left[\frac{1}{f_i}\left(\frac{\partial^2 f_i}{\partial\theta^2}\right)\right] &= \int \frac{1}{f_i}\left(\frac{\partial^2 f_i}{\partial\theta^2}\right)\cdot f_i \mathrm{d}y_i\\ &= \int \frac{\partial^2 f_i}{\partial\theta^2}\mathrm{d}y_i\\ &= \frac{\partial^2}{\partial\theta^2}\int f_i \mathrm{d}y_i\\ &= \frac{\partial^2}{\partial\theta^2}\left(1\right) = 0\end{aligned}$$

So we immediately have

$$\mathrm{E}\left[\frac{\partial^2 \ln f_i}{\partial\theta^2}\right] = -\mathrm{E}\left[\left(\frac{\partial \ln f_i}{\partial\theta}\right)^2\right]$$

Since by definition,

$$\begin{aligned}\frac{\partial \ln L\left(\theta|y\right)}{\partial\theta} &= \sum_{i=1}^{n}\frac{\partial \ln f\left(y_i|\theta\right)}{\partial\theta}\\ \frac{\partial^2 \ln L\left(\theta|y\right)}{\partial\theta^2} &= \sum_{i=1}^{n}\frac{\partial^2 \ln f\left(y_i|\theta\right)}{\partial\theta^2}\end{aligned}$$

So we can conclude that

$$\mathrm{E}\left[\frac{\partial^2 \ln L\left(\theta|y\right)}{\partial\theta^2}\right] = -\mathrm{E}\left[\left(\frac{\partial \ln L\left(\theta|y\right)}{\partial\theta}\right)^2\right]$$

7.5 Consistency of MLE

Consistency of MLE Let θ_0 be the true value of the parameter, $\hat{\theta}$ is the MLE, and θ be any other estimator in the set Θ. Then MLE $\hat{\theta}$ is consistent.

$$\text{plim}\ \hat{\theta} = \theta_0$$

Proof

Jensen's Inequality If $g(x)$ is a concave function, then $\mathrm{E}[g(x)] \leq g(\mathrm{E}[x])$. The inequality is held strict when $g(x)$ is strictly concave.

According the Jensen's inequality, we have

$$\mathrm{E}_0\left[\ln \frac{L(\theta|y)}{L(\theta_0|y)}\right] < \ln \mathrm{E}_0\left[\frac{L(\theta|y)}{L(\theta_0|y)}\right]$$

For the left-hand side,

$$\begin{aligned}
\mathrm{E}_0\left[\frac{L(\theta|y)}{L(\theta_0|y)}\right] &= \int \frac{L(\theta|y)}{L(\theta_0|y)} \cdot L(\theta_0|y)\,\mathrm{d}y \\
&= \int L(\theta|y)\,\mathrm{d}y \\
&= 1
\end{aligned}$$

By the inequality above, it follows that

$$\begin{aligned}
1 = \mathrm{E}_0\left[\ln \frac{L(\theta|y)}{L(\theta_0|y)}\right] &< \ln \mathrm{E}_0\left[\frac{L(\theta|y)}{L(\theta_0|y)}\right] \\
\Longrightarrow \mathrm{E}_0\left[\frac{L(\theta|y)}{L(\theta_0|y)}\right] &< 0 \\
\Longleftrightarrow \mathrm{E}_0\left[\ln L(\theta|y)\right] &< \mathrm{E}_0\left[\ln L(\theta_0|y)\right]
\end{aligned}$$

The intuition is that, the likelihood function gets the maximum when $\theta = \theta_0$.

Now consider the sample analogy,

$$\frac{1}{n}\left[\ln f(y|\theta) - \ln f(y|\theta_0)\right] = \frac{1}{n}\sum_{i=1}^{n} \ln f(y_i|\theta) - \frac{1}{n}\sum_{i=1}^{n} \ln f(y_i|\theta_0)$$

As sample mean is a consistent estimator of the expectation, we have

$$\begin{aligned}
\text{plim}\ \frac{1}{n}\sum_{i=1}^{n} \ln f(y_i|\theta) &= \mathrm{E}_0\left[\frac{1}{n}\ln L(\theta|y)\right] \\
\text{plim}\ \frac{1}{n}\sum_{i=1}^{n} \ln f(y_i|\theta_0) &= \mathrm{E}_0\left[\frac{1}{n}\ln L(\theta_0|y)\right]
\end{aligned}$$

It follows that

$$\text{plim } \frac{1}{n}\sum_{i=1}^{n}\ln f\left(y_i|\theta\right) - \text{plim } \frac{1}{n}\sum_{i=1}^{n}\ln f\left(y_i|\theta_0\right) < 0$$

According to the definition of $\hat{\theta}$ that maximizes the likelihood function, for any finite sample, we have

$$\frac{1}{n}\sum_{i=1}^{n}\ln f\left(y_i|\hat{\theta}\right) \geq \frac{1}{n}\sum_{i=1}^{n}\ln f\left(y_i|\theta_0\right)$$

Hence, it must be that

$$\text{plim } \frac{1}{n}\sum_{i=1}^{n}\ln f\left(y_i|\hat{\theta}\right) \geq \text{plim } \frac{1}{n}\sum_{i=1}^{n}\ln f\left(y_i|\theta_0\right)$$

The two inequalities of plim $\frac{1}{n}\sum_{i=1}^{n}\ln f\left(y_i|\hat{\theta}\right)$ and plim $\frac{1}{n}\sum_{i=1}^{n}\ln f\left(y_i|\theta_0\right)$ can hold if and only if

$$\text{plim } \frac{1}{n}\sum_{i=1}^{n}\ln f\left(y_i|\theta\right) = \text{plim } \frac{1}{n}\sum_{i=1}^{n}\ln f\left(y_i|\theta_0\right)$$

which implies that plim $\hat{\theta} = \theta_0$.

7.6 Asymptotic Normality

Asymptotic Normality The MLE $\hat{\theta}$ has an asymptotic normal distribution,

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{\text{d}} N\left[0, \left\{-\text{E}_0\left[\frac{1}{n}H\left(\theta_0\right)\right]\right\}^{-1}\right]$$

So we have

$$\hat{\theta} \overset{\text{a}}{\sim} N\left[\theta_0, \left\{I\left(\theta_0\right)\right\}^{-1}\right], \text{ where } I\left(\theta_0\right) = -\text{E}_0\left[H\left(\theta_0\right)\right]$$

Proof

At the maximum likelihood estimator, the gradient of the log likelihood equals zero (which is by definition), so $g\left(\hat{\theta}\right) = 0$. We expand this set of equations in a Taylor series around the true parameters θ_0. We will use the mean value theorem to truncate the Taylor series for each element of $g\left(\hat{\theta}\right)$ at the second order (which is Lagrange theorem),

$$g\left(\hat{\theta}\right) = g\left(\theta_0\right) + H\left(\bar{\theta}\right)\cdot\left(\hat{\theta} - \theta_0\right) = 0$$

The K rows of the Hessian are each evaluated at a point $\bar{\theta}_k$ that is between $\hat{\theta}$ and θ_0. Specifically.

Because plim $\left(\hat{\theta}-\theta_0\right)=0$, plim $\left(\hat{\theta}-\bar{\theta}\right)=0$ as well. The second derivatives are continuous functions . Therefore, if the limiting distribution exists, then

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{\mathrm{d}} -\left[H\left(\theta_0\right)\right]^{-1}\left[\sqrt{n}g\left(\theta_0\right)\right]$$

By dividing $H\left(\theta_0\right)$ and $g\left(\theta_0\right)$ by n, we obtain

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{\mathrm{d}} -\left[\frac{1}{n}H\left(\theta_0\right)\right]^{-1}\left[\sqrt{n}\bar{g}\left(\theta_0\right)\right]$$

We may apply the Lindeberg-Levy CLT to $\sqrt{n}\bar{g}\left(\theta_0\right)$, because it is $\sqrt{n}$ times the mean of a random sample.

Example (Information Matrix for a Normal Distribution) For a normal distribution with mean μ and variance σ^2,

$$\ln L\left(\mu,\sigma^2\right)=-\frac{1}{2}\sum_{i=1}^{n}\left[\ln 2\pi+\ln\sigma^2+\frac{\left(y_i-\mu\right)^2}{\sigma^2}\right]$$

$$\Longrightarrow\begin{cases}\frac{\partial^2\ln L}{\partial\mu^2}=-\frac{n}{\sigma^2}\\ \frac{\partial^2\ln L}{\partial\left(\sigma^2\right)^2}=\frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\sum_{i=1}^{n}\left(y_i-\mu\right)^2\\ \frac{\partial^2\ln L}{\partial\mu\partial\left(\sigma^2\right)}=-\frac{1}{\sigma^4}\sum_{i=1}^{n}\left(y_i-\mu\right)\end{cases}$$

For the asymptotic variance of the maximum likelihood estimator, we need the expectations of these derivatives.

The first is non-stochastic, and the third has expectation of 0, since $\mathrm{E}\left[y_i\right]=\mu$. That leaves the second, which we can verify has expectation of $-\frac{n}{2\sigma^4}$, because each of the n terms $\left(y_i-\mu\right)^2$ has expected value σ^2.

Collecting these in the information matrix, reversing the sign, and inverting the matrix gives the asymptotic covariance matrix for the maximum likelihood estimators,

$$\left\{-\mathrm{E}_0\left[H_0\right]\right\}^{-1}=\begin{bmatrix}\frac{\sigma^2}{n} & 0\\ 0 & \frac{2\sigma^4}{n}\end{bmatrix}$$

7.7 Discrete Choice Model

Let U_a and U_b represent an individual's utility of two choices:

$$U_a=w'\beta_1+z_a'\gamma_a+\varepsilon_a$$
$$U_b=w'\beta_b+z_b'\gamma_b+\varepsilon_b$$

If we denote by $Y = 1$ the consumer's choice of alternative a, we infer from $Y = 1$ that $U_a > U_b$.

$$\begin{aligned}\Pr[Y = 1|w, z_a, z_b] &= \Pr[U_a > U_b] \\ &= \Pr[w'(\beta_a - \beta_b) + z_a'\gamma_a - z_b'\gamma_b + (\varepsilon_a - \varepsilon_b)|w, z_a, z_b] \\ &= \Pr[X\beta + \varepsilon > 0|X]\end{aligned}$$

We model the net benefit of a choice as an variable y^* such taht

$$y^* = x'\beta + \varepsilon$$

where ε has mean zero and has either a standardized logistic or normal distribution.

However we do not observe y^* directly; instead, our observation is

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

Then we have

$$\begin{aligned}\Pr[y = 1|x] &= \Pr[y^* > 0|x] \\ &= \Pr[x'\beta + \varepsilon > 0|x] \\ &= \Pr[\varepsilon > -x'\beta|x] \\ &= \Pr[\varepsilon < x'\beta|x] \\ &= F(x'\beta)\end{aligned}$$

Remarks:

- The assumptions of known variance is an innocent normalization.
  - Suppose the variance of ε is scaled by an unrestricted parameter σ^2. The latent regression will be

$$\begin{aligned} y^* &= x'\beta + \sigma\varepsilon \\ \Longrightarrow \left(\frac{y^*}{\sigma}\right) &= x'\left(\frac{\beta}{\sigma}\right) + \varepsilon \end{aligned}$$

  - The transformed model is the same model with the same data. The observed data will be unchanged; y is still 0 or 1, only depending on the sign of y^* instead of its scale.
  - This also means that there is no information about σ in the sample data so σ cannot be estimated.
- The assumption of zero cutoff is another innocent normalization.

– Let a be the supposed nonzero cutoff and α be the unknown constant term in the model, then

$$\begin{aligned}\Pr[y^* > a|x] &= \Pr[\alpha + x'\beta + \varepsilon > a|x] \\ &= \Pr[(\alpha - a) + x'\beta + \varepsilon > 0|x]\end{aligned}$$

– Since α is unknown, the difference $(\alpha - a)$ remains an unknown parameter. If the model contains a constant term, it is unchanged by the choice of the cutoff.

7.7.1 Model Setup

The binary outcome suggests a regression model

$$F(x'\beta) = x'\beta$$

with $\mathrm{E}[y|x] = \{0 \cdot [1 - F(x'\beta)]\} + [1 \cdot F(x'\beta)] = F(x'\beta)$.

This implies the linear probability model:

$$\begin{aligned}y &= \mathrm{E}[y|x] + (y - \mathrm{E}[y|x]) \\ &= x'\beta + \varepsilon\end{aligned}$$

- Shortcoming:
  - We cannot constrain $x'\beta$ to the $[0, 1]$ interval.
  - Heteroskedasticity: $\mathrm{Var}[\varepsilon_i|X] = (x_i'\beta)(1 - x_i'\beta)$, as $x'\beta + \varepsilon$ must equal 0 or 1, and ε equals either $x'\beta$ or $1 - x'\beta$, with probability $1 - F$ and F, respectively.
- Advantages:
  - Simplicity: The coefficient is easy to be interpreted.
  - Robustness: The assumptions of normality or logisticality are fragile while linearity is distribution free.

Generally, we want to construct a model which produces predictions consistent with the underlying theory

$$\Pr[y^* > 0|x] = \Pr[\varepsilon < x'\beta|x] = F(x'\beta)$$

and we expect that

$$\begin{aligned}\lim_{x'\beta \to +\infty} \Pr[Y = 1|x] &= \lim_{x'\beta \to +\infty} F(x'\beta) = 1 \\ \lim_{x'\beta \to -\infty} \Pr[Y = 1|x] &= \lim_{x'\beta \to -\infty} F(x'\beta) = 0\end{aligned}$$

The normal distribution has been commonly used, denoted as the probit model

$$\Pr\left[Y=1|x\right]=\int_{-\infty}^{x'\beta}\phi\left(t\right)\mathrm{d}t=\Phi\left(x'\beta\right)$$

Another commonly used model is the logit model, assuming logistic distribution:

$$\begin{aligned}\Pr\left[Y=1|x\right]&=\frac{1}{1+e^{-x'\beta}}=\Lambda\left(x'\beta\right)\\&=\left(\int_{-\infty}^{x'\beta}\Lambda\left(x'\beta\right)\left(1-\Lambda\left(x'\beta\right)\right)\mathrm{d}\left(x'\beta\right)\right)\end{aligned}$$

The probability model is

$$\mathrm{E}\left[y|x\right]=F\left(x'\beta\right)$$

The parameters of the model are not necessarily the marginal effects

$$\frac{\partial\mathrm{E}\left[y|x\right]}{\partial x}=\frac{\mathrm{d}F\left(x'\beta\right)}{\mathrm{d}\left(x'\beta\right)}\times\frac{\mathrm{d}\left(x'\beta\right)}{\mathrm{d}x}=f\left(x'\beta\right)\times\beta$$

For the probit model,

$$\frac{\partial\mathrm{E}\left[y|x\right]}{\partial x}=\phi\left(x'\beta\right)\times\beta$$

For the logit model,

$$\frac{\partial\mathrm{E}\left[y|x\right]}{\partial x}=\Lambda\left(x'\beta\right)\left[1-\Lambda\left(x'\beta\right)\right]\times\beta$$

The partial effects at the average (PEA)

$$PEA=\hat{\gamma}\left(\bar{x}\right)=f\left(\bar{x}'\hat{\beta}\right)\hat{\beta}$$

The average partial effects (APE):

$$APE=\hat{\bar{\gamma}}=\frac{1}{n}\sum_{i=1}^{n}f\left(x_i'\hat{\beta}\right)\hat{\beta}$$

For the logit model, the odds "in favor of $Y=1$" are

$$\begin{aligned}Odds&=\frac{\Pr\left(Y=1|x\right)}{\Pr\left(Y=0|x\right)}\\&=\exp\left(x'\beta\right)\end{aligned}$$

Consider the effect on the odds of the change of a dummy variable d (whose coefficient is δ),

$$\begin{aligned} OddsRatio &= \frac{Odds\left(x, d=1\right)}{Odds\left(x, d=0\right)} \\ &= \exp\left(\delta\right) \end{aligned}$$

For discrete choice model, the likelihood function is

$$\Pr\left[Y_1 = y_1, Y_2 = y_2, \cdots, Y_n = y_n\right] = \prod_{y_i=0}\left[1 - F\left(x_i'\beta\right)\right] \prod_{y_i=1} F\left(x_i'\beta\right)$$

It can be written as (which is the most important application):

$$\begin{aligned} L\left(\beta|\mathbf{y}\right) &= \prod_{i=1}^{n} F\left(x_i'\beta\right)^{y_i}\left[1 - F\left(x_i'\beta\right)\right]^{1-y_i} \\ \Longrightarrow \ln L &= \sum_{i=1}^{n}\left\{y_i \ln F\left(x_i'\beta\right) + \left(1 - y_i\right)\ln\left[1 - F\left(x_i'\beta\right)\right]\right\} \end{aligned}$$

The likelihood equations is then

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n}\left[\frac{y_i f_i}{F_i} - \left(1 - y_i\right)\cdot\frac{f_i}{1 - F_i}\right]x_i = 0$$

Remarks: Take care when taking derivatives with respect to $\ln F\left(x_i'\beta\right)$: $\dfrac{\partial \ln F\left(x'\beta\right)}{\partial \beta} = \dfrac{1}{F\left(x'\beta\right)}\cdot f\left(x'\beta\right)\cdot x$.

The likelihood equation for a logit model:

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n}\left(y_i - \Lambda_i\right)x_i = 0$$

Likelihood Ratio Index:

$$LRI = 1 - \frac{\ln L}{\ln L_o}$$

7.7.2 Hypothesis Testing

If J is the number of restrictions,

- Likelihood ratio test

$$-2\ln\frac{\hat{L}_R}{\hat{L}_U} \sim \chi^2\left(J\right)$$

- Wald test

$$W = \left[c\left(\hat{\theta}\right) - q\right]' \left\{\text{Asy.Var}\left[c\left(\hat{\theta}\right) - q\right]\right\}^{-1} \left[c\left(\hat{\theta}\right) - q\right] \sim \chi^2\left(J\right)$$

- Lagrange multiplier test

$$LM = \left(\frac{\partial \ln L\left(\hat{\theta}_R\right)}{\partial \hat{\theta}_R}\right)' \left\{I\left(\hat{\theta}_R\right)^{-1}\right\} \left(\frac{\partial \ln L\left(\hat{\theta}_R\right)}{\partial \hat{\theta}_R}\right) \sim \chi^2\left(J\right)$$