Maximum Likelihood Estimation

G.J. LI

1 Likelihood Function

Suppose we can observe a sample $y = (y_1, y_2, \ldots, y_N)$ where N denotes the number of observations. We could model these data by assuming they are from a distribution with underlying parameter vector $\theta_{p\times 1}$ and the **likelihood function** can be written as $p(y|\theta)$. For example, consider the linear regression model below,

$$y_i = x_i'\beta + \epsilon_i, \tag{1}$$

where $x_i = (1, x_{i1}, x_{i2}, \ldots, x_{iK})'$ and $\beta = (\beta_0, \beta_1, \ldots, \beta_K)'$. If we are willing to assume $\epsilon_i \sim i.i.d.N(0, \sigma^2)$, the parameters for this model are β and σ^2 and we could have $y_i|\beta, \sigma^2 \sim i.i.d.N(x_i'\beta, \sigma^2)$[1] with density function,

$$p(y_i|\beta, \sigma^2) = (2\pi)^{-\frac{1}{2}}(\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(y_i - x_i'\beta)^2}{2\sigma^2}\right). \tag{2}$$

By defining $X = (x_1, x_2, \ldots, x_N)'$, $y = (y_1, y_2, \ldots, y_N)'$ and $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_N)'$, we could stack up all the observations and obtain $y \sim N(X\beta, \sigma^2 I_N)$. In other words, we could have the likelihood function for the linear regression

[1] Strictly speaking, the density is also conditional on x_i.

model as

$$\begin{aligned} p(y|\beta,\sigma^2) &= (2\pi)^{-\frac{N}{2}}(\sigma^2)^{-\frac{N}{2}}\exp\left[-\frac{(y-X\beta)'(y-X\beta)}{2\sigma^2}\right], \\ &= (2\pi)^{-\frac{N}{2}}(\sigma^2)^{-\frac{N}{2}}\exp\left[-\frac{\sum_{i=1}^{N}(y_i-x_i'\beta)^2}{2\sigma^2}\right]. \end{aligned} \tag{3}$$

Note that $p(y|\beta,\sigma^2)=\prod_{i=1}^{N}p(y_i|\beta,\sigma^2)$.

2 Parameter Estimation

Given the likelihood function (econometric model), $p(y|\theta)$, firstly we may want to find out what the unknown parameters, θ, are. To estimate θ, we can maximize the LOG likelihood function given the data:[2]

$$\max_{\hat{\theta}} \quad \ln p(y|\hat{\theta}). \tag{4}$$

If all the observations in y are independent, we have $\ln p(y|\theta)=\sum_{i=1}^{N}\ln p(y_i|\theta)$. To solve the maximization problem, we can use the first order condition,

$$\frac{\partial \ln p(y|\hat{\theta})}{\partial\theta}=0, \tag{5}$$

where $\frac{\partial \ln p(y|\theta)}{\partial\theta}$ is called the **score vector** with dimension the same as the number of unknowns in θ. The solution for (5) is the **maximum likelihood estimator** (MLE) for θ, which is a function of the data, denoted as $\hat{\theta}_{MLE}(y)$.[3] For the linear regression model with i.i.d. normal errors, the

[2]Note that in order to estimate θ, usually we should have $N \geq p$.

[3]For some situations, $\hat{\theta}_{MLE}$ may not have analytical or closed form, i.e. we can not write out the function explicitly in terms of y. We can only obtain approximate numerical

maximum likelihood estimators for β and σ^2 are

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'y, \\ \widehat{\sigma^2} &= \frac{SSR}{N} = \frac{(y - X\hat{\beta}_{MLE})'(y - X\hat{\beta}_{MLE})}{N}, \\ &= \frac{y'[I - X(X'X)^{-1}X']y}{N}.\end{aligned} \tag{6}$$

baised

We can see that the MLE estimator of β happens to be the same as the OLS estimator, while the denominator used to calculate $\widehat{\sigma^2}$ is N instead of $N - K - 1$, which implies $\widehat{\sigma^2}$ is a biased estimator. When ϵ_i is not i.i.d. $N(0, \sigma^2)$, $\hat{\beta}$ is said to be a **quasi-maximum likelihood estimator** (QMLE).

3 Properties of MLE

When we derive the properties of MLE, we assume the following **regularity conditions** hold.

1. $\ln p(y_i|\theta)$ is three times differentiable with respect to θ and the derivatives are continuous and finite for all θ and almost all y_i in their support for $i = 1, 2, \ldots, N$.

2. The expectations of the first, second and third order derivatives of $\ln p(y_i|\theta)$ with respect to θ are finite.

3. Either the boundary points of y_i do not depend on θ or $p(y_i|\theta)$ evaluated at the boundary points is 0.

solutions by computer.

Theorem 3.1. *Given the regularity conditions are satisfied, the following are true:*

$$E\left(\frac{\partial \ln p(y_i|\theta)}{\partial \theta}\right) = 0, \tag{7}$$

$$E\left(-\frac{\partial^2 \ln p(y_i|\theta)}{\partial \theta \partial \theta'}\right) = E\left(\frac{\partial \ln p(y_i|\theta)}{\partial \theta}\frac{\partial \ln p(y_i|\theta)}{\partial \theta'}\right) = Var\left(\frac{\partial \ln p(y_i|\theta)}{\partial \theta}\right). \tag{8}$$

Proof. Since $p(y_i|\theta)$ is a proper density function, we have $\int p(y_i|\theta)\, dy_i = 1$. Given the third regularity condition, by Leibniz Theorem, if we differentiate $\int p(y_i|\theta)\, dy_i$ with respect to θ, we can interchange the operations of integration and differentiation[4], i.e.

$$\frac{\partial \int p(y_i|\theta)\, dy_i}{\partial \theta} = \int \frac{\partial p(y_i|\theta)}{\partial \theta}\, dy_i \tag{9}$$

Therefore, we have $\int \frac{\partial p(y_i|\theta)}{\partial \theta}\, dy_i = \int \frac{\partial \ln p(y_i|\theta)}{\partial \theta} p(y_i|\theta)\, dy_i = E\left(\frac{\partial \ln p(y_i|\theta)}{\partial \theta}\right) = 0$. Differentiating $\int \frac{\partial \ln p(y_i|\theta)}{\partial \theta} p(y_i|\theta)\, dy_i$ again with respect to θ yields

$$\int \left[\frac{\partial^2 \ln p(y_i|\theta)}{\partial \theta \partial \theta'} p(y_i|\theta) + \frac{\partial \ln p(y_i|\theta)}{\partial \theta}\frac{\partial p(y_i|\theta)}{\partial \theta'}\right] dy_i = 0, \tag{10}$$

or

$$-\int \frac{\partial^2 \ln p(y_i|\theta)}{\partial \theta \partial \theta'} p(y_i|\theta)\, dy_i = \int \frac{\partial \ln p(y_i|\theta)}{\partial \theta}\frac{\partial p(y_i|\theta)}{\partial \theta'}\, dy_i = \int \frac{\partial \ln p(y_i|\theta)}{\partial \theta}\frac{\partial \ln p(y_i|\theta)}{\partial \theta'} p(y_i|\theta)\, dy_i. \tag{11}$$

Hence $E\left(-\frac{\partial^2 \ln p(y_i|\theta)}{\partial \theta \partial \theta'}\right) = E\left(\frac{\partial \ln p(y_i|\theta)}{\partial \theta}\frac{\partial \ln p(y_i|\theta)}{\partial \theta'}\right)$. □

[4]When y_i is a scalar, Leibniz Theorem states that $\frac{\partial \int_{a(\theta)}^{b(\theta)} p(y_i|\theta)\, dy_i}{\partial \theta} = \int_{a(\theta)}^{b(\theta)} \frac{\partial p(y_i|\theta)}{\partial \theta}\, dy_i + p(b(\theta)|\theta)\frac{\partial b(\theta)}{\partial \theta} - p(a(\theta)|\theta)\frac{\partial a(\theta)}{\partial \theta}$.

If y_i is independent for $i = 1, 2, \ldots, N$, MLE is essentially the solution for the sample analog of the moment condition in (7). Moreover, with data independence (8) implies

$$E\left(-\frac{\partial^2 \sum_{i=1}^N \ln p(y_i|\theta)}{\partial\theta\partial\theta'}\right) = E\left(\frac{\partial \sum_{i=1}^N \ln p(y_i|\theta)}{\partial\theta}\frac{\partial \sum_{i=1}^N \ln p(y_i|\theta)}{\partial\theta'}\right), \tag{12}$$

or $E\left(-\frac{\partial^2 \ln p(y|\theta)}{\partial\theta\partial\theta'}\right) = E\left(\frac{\partial \ln p(y|\theta)}{\partial\theta}\frac{\partial \ln p(y|\theta)}{\partial\theta'}\right)$ (though this identity does not require data dependence), which is called **information identity**. The resulting matrix is called the **information matrix**.

Theorem 3.2. *Apart from the regularity conditions, suppose*

1. $\{y_i\}$ *is an independent sequence generated by* $p(y_i|\theta)$*;*
2. $Var\left(\frac{\partial \ln p(y_i|\theta)}{\partial\theta}\right) = E\left(\frac{\partial \ln p(y_i|\theta)}{\partial\theta}\frac{\partial \ln p(y_i|\theta)}{\partial\theta'}\right) = \mathcal{I}(\theta)$ *evaluated at the true value of* θ *is positive definite.*

Then $\frac{1}{\sqrt{N}}\frac{\sum_{i=1}^N \partial \ln p(y_i|\theta)}{\partial\theta} \overset{A}{\sim} N\left(0, \mathcal{I}(\theta)\right)$ *and* $\sqrt{N}(\hat{\theta} - \theta) \sim N\left(0, \mathcal{I}^{-1}(\theta)\right)$.

Proof. To show that $\frac{1}{\sqrt{N}}\frac{\sum_{i=1}^N \partial \ln p(y_i|\theta)}{\partial\theta}$ is normally distributed asymptotically, the proof is similar to the one in Theorem 8.1 of Lecture 3.

To prove $\hat{\theta}$ is asymptotically normal, by mean value theorem, we can expand $\frac{1}{\sqrt{N}}\frac{\partial \ln p(y|\theta)}{\partial\theta}$ around $\hat{\theta}$: $\frac{1}{\sqrt{N}}\frac{\partial \ln p(y|\theta)}{\partial\theta} = \frac{1}{\sqrt{N}}\frac{\partial \ln p(y|\hat{\theta})}{\partial\theta} + \frac{1}{\sqrt{N}}\frac{\partial^2 \ln p(y|\bar{\theta})}{\partial\theta\partial\theta'}(\theta - \hat{\theta})$, where $\frac{\partial \ln p(y|\hat{\theta})}{\partial\theta} = 0$ and $\bar{\theta}$ is a point between $\hat{\theta}$ and θ. Hence $\sqrt{N}(\hat{\theta} - \theta) = \left(-\frac{1}{N}\frac{\partial^2 \ln p(y|\bar{\theta})}{\partial\theta\partial\theta'}\right)^{-1}\frac{1}{\sqrt{N}}\frac{\partial \ln p(y|\theta)}{\partial\theta}$ and we have $\hat{\theta} \overset{a.s.}{\to} \theta$ and $\bar{\theta} \overset{a.s.}{\to} \theta$. We can see that $\sqrt{N}(\hat{\theta} - \theta)$ is asymptotically equivalent to $\left[E\left(-\frac{\partial^2 \ln p(y_i|\theta)}{\partial\theta\partial\theta'}\right)\right]^{-1}\frac{1}{\sqrt{N}}\frac{\partial \ln p(y|\theta)}{\partial\theta}$. Applying (8) yields the result. □

For the linear regression model with i.i.d. normal errors, we can have

$$\sqrt{N}\begin{pmatrix}\hat{\beta}_{MLE}-\beta\\ \widehat{\sigma^2}_{MLE}-\sigma^2\end{pmatrix}\overset{A}{\sim}N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\sigma^2(plim\frac{X'X}{N})^{-1} & 0\\ 0 & 2\sigma^4\end{bmatrix}\right). \tag{13}$$

Finally, other important properties of MLE are:

1. $\hat{\theta}_{MLE}$ is asymptotically efficient and its asymptotic variance is called the **Cramér-Rao lower bound**;

2. The MLE is invariant under reparameterization. That is, if $\gamma = c(\theta)$, where $c(\cdot)$ is a continuously differentiable one-one function, then $\hat{\gamma}_{MLE} = c(\hat{\theta}_{MLE})$.

4 Hypothesis Testing

general form

Let us consider testing the restrictions: $H_0 : s(\theta) = 0$, where $s : \mathbb{R}^p \to \mathbb{R}^q$ is a continuously differentiable function of θ with $q \leq p$.

Theorem 4.1. ***Wald Test****: Let the conditions of Theorem 3.2 hold and the $q \times (K+1)$ gradient matrix, $\frac{\partial s(\theta)}{\partial \theta'}$ has rank q. Then under $H_0 : s(\theta) = 0$,*

1. $\sqrt{N}s(\hat{\theta}) \overset{A}{\sim} N\left(0, \frac{\partial s(\theta)}{\partial \theta'}\mathcal{I}^{-1}(\theta)\frac{\partial s(\theta)}{\partial \theta'}'\right)$;

unrestricted

2. *The Wald statistic is calculated as* $Ns(\hat{\theta})'\left[\frac{\partial s(\hat{\theta})}{\partial \theta'}\hat{\mathcal{I}}^{-1}(\hat{\theta})\frac{\partial s(\hat{\theta})}{\partial \theta'}'\right]^{-1}s(\hat{\theta}) \overset{A}{\sim} \chi^2_q$, *where* $\hat{\mathcal{I}}(\hat{\theta})$ *is a symmetrical positive definite matrix computed from the unconstrained model such that* $\hat{\mathcal{I}}(\hat{\theta}) \xrightarrow{p} \mathcal{I}(\theta)$.

Proof. Note that $s(\cdot)$ is a vector function. We can apply mean value theorem to each of its element around θ to get $s_i(\hat{\theta}) = s_i(\theta) + \frac{\partial s_i(\bar{\theta}^{(i)})}{\partial \theta'}(\hat{\theta} - \theta)$ with

$s_i(\theta) = 0$ under the restriction for $i = 1, 2, \ldots, q$. Since $\hat{\theta} \xrightarrow{p} \theta$, we have $\bar{\theta}^{(i)} \xrightarrow{p} \theta$. Hence $s_i(\hat{\theta})$ is asymptotically equivalent to $\frac{\partial s_i(\theta)}{\partial \theta'}(\hat{\theta} - \theta)$ and $s(\hat{\theta})$ is asymptotically equivalent to $\frac{\partial s(\theta)}{\partial \theta'}(\hat{\theta} - \theta)$. Given the result in Theorem 3.2, we can have $\sqrt{N} s(\hat{\theta}) \overset{A}{\sim} N\left(0, \frac{\partial s(\theta)}{\partial \theta'} \mathcal{I}^{-1}(\theta) \frac{\partial s(\theta)}{\partial \theta'}'\right)$. □

Under the restriction, the maximization problem now is to find $\tilde{\theta}$ and $\tilde{\lambda}$ to maximize the following Lagrangian,

$$\mathcal{L}(\tilde{\theta}, \tilde{\lambda}) = \ln p(y|\tilde{\theta}) + \tilde{\lambda}' s(\tilde{\theta}). \tag{14}$$

The first order conditions are

$$\begin{aligned} \frac{\partial \mathcal{L}(\tilde{\theta}, \tilde{\lambda})}{\partial \theta} &= \frac{\partial \ln p(y|\tilde{\theta})}{\partial \theta} + \frac{\partial s(\tilde{\theta})}{\partial \theta'}' \tilde{\lambda} = 0, & (15) \\ \frac{\partial \mathcal{L}(\tilde{\theta}, \tilde{\lambda})}{\partial \lambda} &= s(\tilde{\theta}) = 0. & (16) \end{aligned}$$

Taking mean value expansions of $\frac{\partial \ln p(y|\tilde{\theta})}{\partial \theta}$ and $s(\tilde{\theta})$ around $\hat{\theta}$ in (15) and (16) respectively gives

$$\begin{aligned} \frac{\partial \mathcal{L}(\tilde{\theta}, \tilde{\lambda})}{\partial \theta} &= \frac{\partial \ln p(y|\hat{\theta})}{\partial \theta} + \frac{\partial^2 \ln p(y|\bar{\theta})}{\partial \theta \partial \theta'}(\tilde{\theta} - \hat{\theta}) + \frac{\partial s(\tilde{\theta})}{\partial \theta'}' \tilde{\lambda} = 0, \\ &= \frac{\partial^2 \ln p(y|\bar{\theta})}{\partial \theta \partial \theta'}(\tilde{\theta} - \hat{\theta}) + \frac{\partial s(\tilde{\theta})}{\partial \theta'}' \tilde{\lambda}, & (17) \\ \frac{\partial \mathcal{L}(\tilde{\theta}, \tilde{\lambda})}{\partial \lambda} &= s(\hat{\theta}) + \frac{\partial s(\bar{\theta})}{\partial \theta'}(\tilde{\theta} - \hat{\theta}) = 0. & (18) \end{aligned}$$

Premultiplying (17) by $\frac{\partial s(\bar{\theta})}{\partial \theta'}\left(\frac{\partial^2 \ln p(y|\bar{\theta})}{\partial \theta \partial \theta'}\right)^{-1}$, we have $\tilde{\lambda} = \left[\frac{\partial s(\bar{\theta})}{\partial \theta'}\left(\frac{\partial^2 \ln p(y|\bar{\theta})}{\partial \theta \partial \theta'}\right)^{-1} \frac{\partial s(\tilde{\theta})}{\partial \theta'}'\right]^{-1} \frac{\partial s(\bar{\theta})}{\partial \theta'}(\hat{\theta} - \tilde{\theta}) = \left[\frac{\partial s(\bar{\theta})}{\partial \theta'}\left(\frac{\partial^2 \ln p(y|\bar{\theta})}{\partial \theta \partial \theta'}\right)^{-1} \frac{\partial s(\tilde{\theta})}{\partial \theta'}'\right]^{-1} s(\hat{\theta})$, which leads to the next theorem.

Theorem 4.2. ***Lagrange Multiplier Test****: Let the conditions of Theorem 3.2 hold. Then under* $H_0 : s(\theta) = 0$*,*

1. $\frac{1}{\sqrt{N}}\tilde{\lambda} \overset{A}{\sim} N\left(0, (\frac{\partial s(\theta)}{\partial \theta'}\mathcal{I}^{-1}(\theta)\frac{\partial s(\theta)}{\partial \theta'}')^{-1}\right)$;
2. *The Lagrange multiplier (LM) test statistic is defined as*

$$\frac{1}{N}\tilde{\lambda}'\frac{\partial s(\tilde{\theta})}{\partial \theta'}\hat{\mathcal{I}}^{-1}(\tilde{\theta})\frac{\partial s(\tilde{\theta})}{\partial \theta'}'\tilde{\lambda} \overset{A}{\sim} \chi^2_q,$$

where $\hat{\mathcal{I}}(\tilde{\theta})$ *is a symmetrical positive definite matrix computed from the constrained model such that* $\hat{\mathcal{I}}(\tilde{\theta}) \xrightarrow{p} \mathcal{I}(\theta)$*. From (15), we have* $\frac{\partial s(\tilde{\theta})}{\partial \theta'}'\tilde{\lambda} = -\frac{\partial \ln p(y|\tilde{\theta})}{\partial \theta}$*. Therefore, the LM statistic can also be calculated as* $\frac{1}{N}\frac{\partial \ln p(y|\tilde{\theta})}{\partial \theta}'\hat{\mathcal{I}}^{-1}(\tilde{\theta})\frac{\partial \ln p(y|\tilde{\theta})}{\partial \theta}$.

Proof. Since $\frac{1}{\sqrt{N}}\tilde{\lambda} = \left[\frac{\partial s(\bar{\theta})}{\partial \theta'}\left(\frac{1}{N}\frac{\partial^2 \ln p(y|\bar{\theta})}{\partial \theta \partial \theta'}\right)^{-1}\frac{\partial s(\tilde{\theta})}{\partial \theta'}'\right]^{-1}\sqrt{N}s(\hat{\theta})$ and both $\tilde{\theta}$ and $\bar{\theta}$ converges to θ almost surely under the null, $\frac{1}{\sqrt{N}}\tilde{\lambda}$ is asymptotically equivalent to $\left[\frac{\partial s(\theta)}{\partial \theta'}\mathcal{I}^{-1}(\theta)\frac{\partial s(\tilde{\theta})}{\partial \theta'}'\right]^{-1}\sqrt{N}s(\hat{\theta})$. Using Theorem 4.1 proves the asymptotic normality of $\frac{1}{\sqrt{N}}\tilde{\lambda}$. Applying Corollary 9.2 in Lecture 3 yields the LM statistic. □

Once we have the solution of $\tilde{\lambda}$ in terms of $\tilde{\theta}$, we can substitute into (17) to obtain

$$\tilde{\lambda} = \left[\frac{\partial s(\bar{\theta})}{\partial \theta'}\left(\frac{\partial^2 \ln p(y|\bar{\theta})}{\partial \theta \partial \theta'}\right)^{-1}\frac{\partial s(\tilde{\theta})}{\partial \theta'}'\right]^{-1}s(\hat{\theta}), \tag{19}$$

$$\tilde{\theta} = \hat{\theta} - \left(\frac{\partial^2 \ln p(y|\bar{\theta})}{\partial \theta \partial \theta'}\right)^{-1}\frac{\partial s(\tilde{\theta})}{\partial \theta'}'\left[\frac{\partial s(\bar{\theta})}{\partial \theta'}\left(\frac{\partial^2 \ln p(y|\bar{\theta})}{\partial \theta \partial \theta'}\right)^{-1}\frac{\partial s(\tilde{\theta})}{\partial \theta'}'\right]^{-1}s(\hat{\theta}). \tag{20}$$

As (23) in Lecture 3, the solution in (20) in general does not have closed form. An asymptotic equivalent estimator can by obtained by replacing θ and $\bar{\theta}$ by $\hat{\theta}$ on the right hand side.

$$\tilde{\theta} = \hat{\theta} - \left(\frac{\partial^2 \ln p(y|\hat{\theta})}{\partial\theta\partial\theta'} \right)^{-1} \frac{\partial s(\hat{\theta})}{\partial\theta'}' \left[\frac{\partial s(\hat{\theta})}{\partial\theta'} \left(\frac{\partial^2 \ln p(y|\hat{\theta})}{\partial\theta\partial\theta'} \right)^{-1} \frac{\partial s(\hat{\theta})}{\partial\theta'}' \right]^{-1} s(\hat{\theta}) \tag{21}$$

We can now obtain

$$-(\tilde{\theta}-\hat{\theta})' \frac{\partial^2 \ln p(y|\hat{\theta})}{\partial\theta\partial\theta'} (\tilde{\theta}-\hat{\theta}) = N s(\hat{\theta})' \left[\frac{\partial s(\hat{\theta})}{\partial\theta'} \left(-\frac{1}{N} \frac{\partial^2 \ln p(y|\hat{\theta})}{\partial\theta\partial\theta'} \right)^{-1} \frac{\partial s(\hat{\theta})}{\partial\theta'}' \right]^{-1} s(\hat{\theta}) \tag{22}$$

The right hand side is the Wald statistic. The result in (22) motivates another test statistic under the likelihood framework.

Theorem 4.3. ***Likelihood Ratio Test****: Let the conditions of Theorem 3.2 hold. Then under* $H_0 : s(\theta) = 0$, $-2 \ln \frac{p(y|\tilde{\theta})}{p(y|\hat{\theta})}$ *will asymptotically follow chi-squared distribution with degrees of freedom* q*, where* $\tilde{\theta}$ *is the estimate from the restricted model while* $\hat{\theta}$ *is the estimate from the unrestricted model.*

Proof. Expand $\ln p(y|\tilde{\theta})$ around $\hat{\theta}$:

$$\begin{aligned} -2 \ln \frac{p(y|\tilde{\theta})}{p(y|\hat{\theta})} &= -2 \left[\ln p(y|\hat{\theta}) + \frac{\partial \ln p(y|\hat{\theta})}{\partial\theta'} (\tilde{\theta} - \hat{\theta}) + \frac{1}{2} (\tilde{\theta} - \hat{\theta})' \frac{\partial^2 \ln p(y|\bar{\theta})}{\partial\theta\partial\theta'} (\tilde{\theta} - \hat{\theta}) - \ln p(y|\hat{\theta}) \right] \\ &= -(\tilde{\theta} - \hat{\theta})' \frac{\partial^2 \ln p(y|\bar{\theta})}{\partial\theta\partial\theta'} (\tilde{\theta} - \hat{\theta}) \end{aligned} \tag{23}$$

Since $\frac{\partial \ln p(y|\hat{\theta})}{\partial\theta'} = 0$, $-2 \ln \frac{p(y|\tilde{\theta})}{p(y|\hat{\theta})}$ will hence be asymptotically equivalent to $-(\tilde{\theta} - \hat{\theta})' \frac{\partial^2 \ln p(y|\hat{\theta})}{\partial\theta\partial\theta'} (\tilde{\theta} - \hat{\theta})$, which is the same as the Wald statistic from

(22). □

For the linear regression model with i.i.d. normal error, under linear restrictions, the three test statistics turn out to have the following forms:

1. $LR = N \ln \frac{SSR_R}{SSR_U}$,
2. $W = N \frac{SSR_R - SSR_U}{SSR_U}$,
3. $LM = N \frac{SSR_R - SSR_U}{SSR_R} = NR_*^2$, where R_*^2 is the r-squared obtained by regressing the estimated residuals from the restricted model on the regressors under the unrestricted model.

special

Note that $W \geq LR \geq LM$ and all the test statistics are asymptotically equivalent and are closely related to the F test statistic: $F = \frac{(SSR_R - SSR_U)/q}{SSR_U/(N-K-1)}$. In fact $q \times F = \frac{N-K-1}{N} W = \frac{(N-K-1)LM}{N-LM}$.

5 Information Criterion

Apart from doing hypothesis tests, we can also use information criteria to compare different model specifications under likelihood framework. Any information criterion can be written in the form of $-2 \ln p(y|\hat{\theta}_{MLE}) + c$, where c is some constant and $-2 \ln p(y|\hat{\theta}_{MLE})$ is sometimes called deviance. A model with smaller information criterion value fits the data better. The following are two commonly seen criteria:

1. Akaike information criterion: $AIC = -2 \ln p(y|\hat{\theta}_{MLE}) + 2p$, where p is the dimension of θ;

2. Bayesian information criterion (Schwarz criterion): $BIC = -2 \ln p(y|\hat{\theta}_{MLE}) + p \ln N$.

For the linear regression model, AIC and BIC respectively are

$$AIC = N(\ln \frac{2\pi SSR}{N} + 1) + 2(K+2), \tag{24}$$

$$BIC = N(\ln \frac{2\pi SSR}{N} + 1) + (K+2) \ln N. \tag{25}$$

6 Binary Dependent Variable Models

In economic studies, it is sometimes useful to incorporate qualitative information into the model. Qualitative information, for example, can refer to the gender or race of an individual, the industry of a firm (manufacturing, retail, etc.), and the region where a city is located (south, north, west, etc.). Quite often qualitative information comes in binary form: a person is female or male; a person does or does not own a personal computer; a firm offers a certain kind of employee pension plan or it does not. We can model such binary information using dummy variables, whose values are either 0 or 1. You may probably have seen dummy variables used as explanatory variables in a regression. In this section, we will discuss models related to binary dependent variables.

6.1 Linear Probability Model

For the linear regression model,

$$y_i = x_i'\beta + \epsilon_i, \quad \text{for } i = 1, 2, \ldots, N, \tag{26}$$

where $x_i = (1, x_{1i}, x_{2i}, \ldots, x_{Ki})'$ and $\beta = (\beta_0, \beta_1, \ldots, \beta_K)'$. Here y_i only takes value of either 0 or 1. For example, we want to find out what factors can explain a married woman's decision to stay in the labour force: $y_i = 1$ to stay and $y = 0$ otherwise. Due to the binary nature of y_i, we can have

$$E(y_i|X) = 1 \times P(y_i = 1|X) + 0 \times P(y_i = 0|X) = P(y_i = 1|X). \tag{27}$$

In other words, the probability of $y_i = 1$ is represented by

$$P(y_i = 1|X) = x_i'\beta \tag{28}$$

Therefore β is interpreted as the effect of a change of x_i on the probability for a woman to continue to work. Note also that due to the binary nature of y_i,

$$Var(y_i) = [1 - P(y_i = 1|X)]\, P(y_i = 1|X) = (1 - x_i'\beta)x_i'\beta. \tag{29}$$

Hence we have heteroskedasticity in the model.

Though we can use linear regression model to study binary dependent variables, such application is rarely seen in the literature due to the following limitations.

1. It is possible to have probability point forecast, $x_i'\hat{\beta}$, outside the range of $[0, 1]$, which does not make sense for $P(y_i = 1|X)$. Typically, point forecast works well only for observations close to the sample average of x_i.

2. The probability of $y_i = 1$ is linearly related to the independent variables for all their possible values. For example, we may use number of children a woman has to explain her work decision. Intuitively speaking, the effect of going from zero children to one young child could be bigger than having two children from one on a woman's decision to work.

6.2 Logit and Probit Model

To overcome the limitations of linear probability model, we need more sensible way to relate $P(y_i = 1|X)$ to the linear combination $x_i'\beta$. For a continuous random variable, its cumulative distribution functions (CDF), $G(\cdot)$ is defined as

$$G(x) = \int_{-\infty}^{x} g(t)dt. \tag{30}$$

where $g(\cdot)$ is the probability density function (PDF) and $g(x) = \frac{dG(x)}{dx}$. Note that $G(\cdot)$ always takes on values between 0 and 1. A possible way to relate $P(y_i = 1|X)$ to $x_i'\beta$ is to define

$$P(y_i = 1|X) = G(x_i'\beta). \tag{31}$$

dp/dx=dG(x*beta)*d(x*beta)=dg*beta

Now the point forecast is always inside the range of $[0, 1]$. Popular choices for $G(\cdot)$ include standard normal distribution (mean is 0 and variance is 1) and logistic distribution. For **probit model**, $G(\cdot)$ is the standard normal CDF, while for **logit model**, $G(\cdot)$ is the standard logistic CDF. Sometimes they are referred to as **binary probability response model**. The standard normal density takes the following form

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right). \tag{32}$$

However, its CDF, denoted as $\Phi(\cdot)$, does not have closed form. The density function of standard logistic distribution is

$$g_L(x) = \frac{\exp(-x)}{[1 + \exp(-x)]^2}. \tag{33}$$

Its CDF has the following closed form

$$G_L(x) = \frac{1}{1 + \exp(-x)} = \frac{\exp(x)}{1 + \exp(x)}. \tag{34}$$

Note that for standard normal and logistic distribution, $G(\cdot)$ is a nonlinear function with $G(0) = 0.5$ and both PDFs have symmetrical bell shape around 0. The marginal effect of x_i is now $\frac{\partial P(y_i=1|X)}{\partial x_i'} = g(x_i'\beta)\beta$, which will depend on not only β, but also on $g(x_i'\beta)$, which is non-negative. Hence, the sign of the marginal effect will depend on β only. Since both the logistic and standard normal distribution have bell shape densities, the magnitude increase of x_i (i.e. $|x_i|$) will have diminishing marginal effect on the probability response. When investigating the marginal effect, we need to focus on

some interesting values of x_i, such as the sample average or the quantiles. $x_i'\beta = 0$ (the maximum point for the standard normal and logistic density) is also an interesting value. Note that $\phi(0) = \frac{1}{\sqrt{2\pi}} \approx 0.4$ and $g_L(0) = 0.25$. To make the βs from logit and probit roughly comparable to that obtained from linear probability model, we could multiply the β from probit (logit) by 0.4 (0.25).

To estimate β, we have to use maximum likelihood. The likelihood function is

$$\begin{aligned} P(y|X,\beta) &= P(y_1|x_1,\beta)P(y_2|x_2,\beta)\dots P(y_N|x_N,\beta), \\ &= \prod_{i=1}^{N} G(x_i'\beta)^{y_i}\left[1-G(x_i'\beta)\right]^{1-y_i}. \end{aligned} \tag{35}$$

We can then estimate β by taking log of the likelihood function and finding the maximum by computer numerically.

As mentioned earlier, under regularity conditions, the maximum likelihood estimator (MLE) is consistent, asymptotically normal, and asymptotically efficient. The z-statistic $\frac{\hat{\beta}_{MLE}}{std(\hat{\beta}_{MLE})}$ calculated in R to test $H_0 : \hat{\beta}_{MLE} = 0$ will follow standard normal distribution asymptotically. The three tests related to maximum likelihood estimation can be used to test restrictions on β.

follow standard normal distribution

Question: Find the score vector and information matrix for binary probability response model. ($\frac{\partial \ln P(y|\beta)}{\partial \beta} = \sum_{i=1}^{N} \frac{g(x_i'\beta)[y_i - G(x_i'\beta)]}{G(x_i'\beta)[1-G(x_i'\beta)]} x_i$ and $E\left(-\frac{\partial^2 \ln P(y|\beta)}{\partial\beta\partial\beta'}\right) = E\left(\frac{\partial \ln P(y|\beta)}{\partial\beta}\frac{\partial \ln P(y|\beta)}{\partial\beta'}\right) = \sum_{i=1}^{N} \frac{g^2(x_i'\beta)}{G(x_i'\beta)[1-G(x_i'\beta)]} x_i x_i'$)

To measure the fitness of the model, we can use information criterion or Pseudo (McFadden) R-squared measure: $1 - \frac{\ln P(y|X,\hat{\beta}_{MLE})}{\ln P(y|X,\hat{\beta}_0)}$, where the

(normal) R^2=1-(SSR/SST)

SST=(SSR_r within all coef=0)

numerator is the unrestricted log likelihood evaluated at the MLE and denominator is the log likelihood obtained by including just the intercept in the binary response model.

References

[1] W.H. Greene, *Econometric Analysis*, 7th ed., Pearson Education, Chapter 16, 2011.

[2] H. White, *Asymptotic Theory for Econometricians*, Chapter 4.2, Prentice Hall, 2001.

[3] J.M. Wooldridge, *Introductory Econometrics (a modern approach)*, 4th ed., South Western College, 7.5, 7.6, 8.5, 17.1, 17A, 2008.