Maximum Likelihood Estimation of Normal Linear Regression Model

Estimation

We consider the ML estimator of β for the linear model $y = X\beta + u$ where $u|X \sim N\left(0, \sigma^2 I_n\right)$

Maximum likelihood principle: what values of β and σ^2 make the observed sample most probable.

Define the likelihood function $L\left(\beta, \sigma^2\right) = f\left(\beta, \sigma | X, y\right)$ as the joint pdf of the sample.

Given $u|X \sim N\left(0, \sigma^2 I_n\right)$, then using the fact that $u = y - X\beta$ we have that

$$f\left(y|X; \beta, \sigma\right) = f\left(u\left(y\right)|X; \beta, \sigma\right)\left|u'\left(y\right)\right| = f\left(u\left(y\right)|X; \beta, \sigma\right),$$

since $u'\left(y\right) = 1$. Hence, $y|X \sim N\left(X\beta, \sigma^2 I_n\right)$.

For $y \sim MVNormal\left(\mu, \Sigma\right)$:

$$f\left(y\right) = \left(2\pi\right)^{-n/2} \left|\Sigma\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(y-\mu\right)' \Sigma^{-1} \left(y-\mu\right)\right\},$$

in our case $\Sigma = \sigma^2 I_n$. So, $\left|\Sigma\right|^{-1/2} = \left(\sigma^2\right)^{-n/2}$ and $\mu = X\beta$, so the likelihood function is

$$L\left(\beta, \sigma^2 | y, X\right) = \\ \left(2\pi\right)^{-n/2} \left(\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(y - X\beta\right)' \left(y - X\beta\right)\right\}.$$

We define the log-likelihood function as

$$\ell\left(\beta, \sigma^2\right) \equiv \ln L = \\ -\frac{n}{2}\ln\left(2\pi\right) - \frac{n}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\left(y - X\beta\right)' \left(y - X\beta\right).$$

Note: One advantage of working with $\ln L$ is that since $L = \prod_i f(y_i) \implies \ln L = \sum_i \ln f(y_i)$, then $\frac{\ln L}{n}$ is a sample average and thus can use LLN and CLT.

FOCs:

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{2\sigma^2} 2X' (y - X\beta) = 0 \implies \hat{\beta}_{ML} = (X'X)^{-1} X'y$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (y - X\beta)' (y - X\beta) = 0$$

$$\implies \hat{\sigma}^2_{ML} = \frac{\hat{u}'\hat{u}}{n}, \text{ where } \hat{u} = y - X\hat{\beta}_{ML}.$$

Notice that

$$\hat{\sigma}^2_{ML} = \frac{n-k}{n} s^2 \neq s^2,$$

which implies that $\hat{\sigma}^2_{ML}$ is biased.

Remark: Under the normality assumption $u|X \sim N(0, \sigma^2 I_n)$, we have that $\hat{\beta}_{ML}$ equals b the LS estimator, but $\hat{\sigma}^2_{ML} = \frac{n-k}{n} s^2$ is a biased estimator of σ^2.
In general, LS estimator and MLE are different.

Remark: s^2 a somewhat method of moments estimator. It does not derive from the maximisation of an objective function. We simply adjust by the DoF to get an unbiased estimator.

Remark: Estimation imposes the gradient equality $\partial \ln L / \partial \theta = 0$. This is called the *likelihood equation*. MLE is then a root of the likelihood equation.

Score function

The score function is the vector of first partial derivatives of the log-likelihood function, $\partial \ln L/\partial\theta$; in our case $\theta = (\beta, \sigma^2)'$. That is, the score is just the gradient vector.

The expected value of the score function is 0, that is

$$E\left[\frac{\partial \ln L}{\partial\theta}\right] = 0.$$

To verify this in the normal linear model assume that $E\left[u|X\right] = 0$. Consider

$$\begin{aligned} E\left[\frac{\partial \ln L}{\partial\beta}\right] &= E\left[E\left[\frac{\partial \ln L}{\partial\beta}|X\right]\right] \\ &= E\left[E\left[\left(\frac{1}{2\sigma^2}2X'u\right)|X\right]\right] \\ &= E\left[\frac{1}{\sigma^2}X'E\left[u|X\right]\right] = 0 \end{aligned}$$

and

$$\begin{aligned}
E\left[\frac{\partial \ln L}{\partial \sigma^2}\right] &= E\left[E\left[\frac{\partial \ln L}{\partial \beta}|X\right]\right] \\
&= E\left[E\left[\left(-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}u_i^2\right)|X\right]\right] \\
&= E\left[\left(-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}E\left[u_i^2|X\right]\right)\right] \\
&= -\frac{n}{2\sigma^2}+\frac{n\sigma^2}{2\sigma^4}=0
\end{aligned}$$

Information matrix

The information matrix is minus the expectation of the Hessian matrix, evaluated at the true parameters.

$$I(\theta) \equiv -E\left[\frac{\partial^2 \ln L}{\partial\theta\partial\theta'}\right]$$

In our linear normal model, we have $\theta = (\beta, \sigma^2)'$.

$$\begin{aligned}
\frac{\partial^2 \ln L}{\partial\beta\partial\beta'} &= \frac{\partial}{\partial\beta'}\left(\frac{\partial \ln L}{\partial\beta}\right) \\
&= \frac{\partial}{\partial\beta'}\left(\frac{1}{2\sigma^2}2X'(y - X\beta)\right) \\
&= -\frac{X'X}{\sigma^2} \\
\implies E\left[-\frac{\partial^2 \ln L}{\partial\beta\partial\beta'}|X\right] &= \frac{X'X}{\sigma^2}.
\end{aligned}$$

Also,

$$\begin{aligned}
\frac{\partial^2 \ln L}{\partial \beta \partial \sigma^2} &= \frac{\partial}{\partial \sigma^2}\left(\frac{\partial \ln L}{\partial \beta}\right) \\
&= \frac{\partial}{\partial \sigma^2}\left(\frac{1}{\sigma^2} X' u\right) \\
&= -\frac{X' u}{\sigma^4} \\
\implies E\left[-\frac{\partial^2 \ln L}{\partial \beta \partial \sigma^2} | X\right] &= \frac{E\left[X' u | X\right]}{\sigma^4} = \frac{0}{\sigma^4} = 0.
\end{aligned}$$

Finally,

$$\begin{aligned}
\frac{\partial^2 \ln L}{\partial\left(\sigma^2\right)^2} &= \frac{\partial}{\partial\sigma^2}\left(-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}u'u\right)\\
&= \frac{n}{2\sigma^4}-2\frac{1}{2\sigma^6}\varepsilon'\varepsilon\\
\Longrightarrow\ E\left[-\frac{\partial^2 \ln L}{\partial\left(\sigma^2\right)^2}|X\right] &= -E\left[\frac{n}{2\sigma^4}-2\frac{1}{2\sigma^6}\varepsilon'\varepsilon|X\right]\\
&= -\frac{n}{2\sigma^4}+2\frac{n\sigma^2}{2\sigma^6}\\
&= \frac{n}{2\sigma^4}.
\end{aligned}$$

The information matrix is

$$I\left(\beta,\sigma|X\right)=\left[\begin{array}{cc}\frac{X'X}{\sigma^{2}} & 0_{k\times1}\\ 0_{1\times k} & \frac{n}{2\sigma^{4}}\end{array}\right].$$

Variance of score function

The variance of the score is:

$$\begin{aligned} Var\left[\frac{\partial \ln L}{\partial\theta}\right] &= E\left[\frac{\partial \ln L}{\partial\theta}\left(\frac{\partial \ln L}{\partial\theta}\right)'\right] - \underbrace{E\left[\frac{\partial \ln L}{\partial\theta}\right]}_{=0}\underbrace{E\left[\left(\frac{\partial \ln L}{\partial\theta}\right)'\right]}_{=0} \\ &= E\left[\frac{\partial \ln L}{\partial\theta}\left(\frac{\partial \ln L}{\partial\theta}\right)'\right], \end{aligned}$$

Note that the information matrix also equals the variance of the score and in fact we have the result called the *information matrix equality*

$$Var\left[\frac{\partial \ln L}{\partial\theta}\right] = -E\left[\frac{\partial^2 \ln L}{\partial\theta\partial\theta'}\right] = I\left(\theta\right).$$

Remark: The result that

$$Var\left[\frac{\partial \ln L}{\partial\theta}\right] = -E\left[\frac{\partial^2 \ln L}{\partial\theta\partial\theta'}\right]$$

depends on the assumption that we have specified the *true density*. We could test whether we have specified the true density by measuring the scaled distance of the two.

CR-Lower Bound (CRLB)

Theorem[CR Lower bound]: If $E\left[\hat{\theta}\right] = \theta$ then $Var\left[\hat{\theta}\right] \geq I(\theta)^{-1}$. $I(\theta)^{-1}$ is called the Cramèr-Rao Lower Bound.

Remark: ML uses more information than LS since we know the the pdf. LS is an curve-fitting approach, of semi-parametric nature.

$\hat{\beta}_{ML} = b \Longrightarrow b$ **is Best Unbiased Estimator**
The CRLB is

$$I(\beta, \sigma | X)^{-1} = \left[\begin{array}{cc} \sigma^2 (X'X)^{-1} & 0_{k\times 1} \\ 0_{1\times k} & \frac{2\sigma^4}{n} \end{array} \right].$$

Remark: Therefore, $\hat{\beta}_{ML} = b$ attains the CR lower bound. That is, under the linear normal model the LS estimator b is the best among the class of unbiased estimators. This is stronger result than GM result according to which b is the best among the class of unbiased and *linear* estimators. In our case, we obtain the stronger result for the LS estimator b under the error *normality* assumption.

Remark: There does not exist an unbiased estimator of σ^2 that attains the CR lower bound. To see that s^2 does not attain the bound suppose $u|X \sim N\left(0, \sigma^2 I_n\right)$. Then,

$$\frac{\hat{u}'\hat{u}}{\sigma^2} \sim \chi^2_{n-k} \implies \frac{(n-k)\,s^2}{\sigma^2} \sim \chi^2_{n-k}.$$

Then,

$$E\left[\frac{(n-k)\,s^2}{\sigma^2}\right] = n-k \implies E\left[s^2\right] = \sigma^2$$
$$Var\left[\frac{(n-k)\,s^2}{\sigma^2}\right] = 2\,(n-k) \implies Var\left[s^2\right] = \frac{2\sigma^4}{n-k} > \frac{2\sigma^4}{n}.$$

And,

$$Var\left[\hat{\sigma}^2_{ML}\right] = \frac{2\,(n-k)}{n^2}\sigma^4.$$

Properties of ML estimator

Definition: An estimator $\hat{\theta}$ is said to be *asymptotically efficient* if it is *CAN* (consistent, asymptotically normal) and there is no other estimator in this class with smaller variance/covariance matrix. That is, $Var\left[\tilde{\theta}\right] - Var\left[\hat{\theta}\right]$ is positive semi-definite for any $\tilde{\theta}$ in the CAN class.

Proposition: Under regularity conditions (on $f(y|X;\theta)$), the MLE estimator $\hat{\theta}$ has the following asymptotic properties:

M1: Consistency: $p\lim\left(\hat{\theta}\right) = \theta_0$

M2: Asymptotic normality: $\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, \left(\frac{I(\theta_0)}{n}\right)^{-1}\right)$, or in practice $\hat{\theta} \overset{a}{\sim} N\left(\theta_0, I(\theta_0)^{-1}\right)$, where $I(\theta_0) = E\left[-\partial^2 \ln L / \partial\theta\partial\theta'\right]$.

M3: Asymptotic efficiency: $\hat{\theta}$ achieves the CRLB and it's asymptotically efficient.

M4: Invariance: MLE of $\gamma_0 = c(\theta_0)$ is $\hat{\gamma} = c\left(\hat{\theta}\right)$.

The underlying assumption above is that $f(\cdot)$ is the true density.

Hypothesis tests

Consider the maximum likelihood estimation of a parameter θ and a test of the hypothesis $H_0 : c(\theta) = 0$. The logic of the 3 different test bases are illustrated in the next graph. But this testing bases extent to other estimation procedures, such as the LS.

Figure: Three Bases for Hypothesis Tests

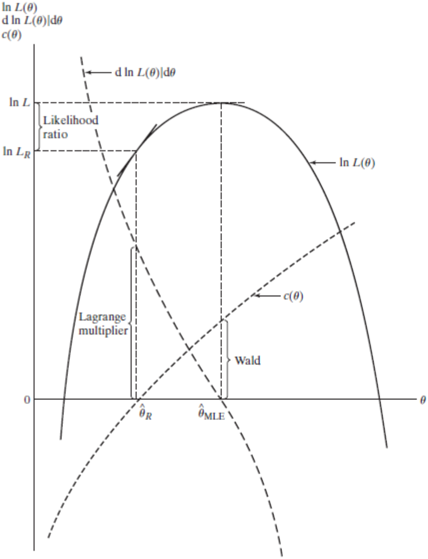

FIGURE 14.2 Three Bases for Hypothesis Tests.

1) LR test: If $c\left(\theta\right) = 0$, the imposing it should not lead to a large reduction in the log-likelihood function. So, we test the difference $\ln L_U - \ln L_R$.

This requires that we estimate both models. The F-test with SSR_U and SSR_R is the LR form of the F-test.

2) Wald test: If the restriction is valid, then $c\left(\hat{\theta}_{MLE}\right)$ should be close to zero because MLE is consistent. Therefore, the test is based on $c\left(\hat{\theta}_{MLE}\right)$ and we reject the hypothesis if this value is significantly different from zero.

This requires that we estimate only the unrestricted model and evaluate the restriction at the unrestricted estimator value. For example, if $c\left(\theta\right) = R\beta - r = 0$, then as we have seen above we can express the F-test in terms of $c\left(\hat{\theta}\right) = Rb - r = 0$, which is the Wald representation of the F-test.

3) Lagrange Multiplier (LM) or Rao's Score test: If the restriction is valid, then the restricted estimator should be near the point that maximises the log-likelihood function. Therefore, the slope of the log-likelihood function $\partial \ln L / \partial \theta$ should be near zero at the restricted estimator. The test is based on the slope of the log-likelihood at the point where the function is maximised subject to the restriction.

These three tests are asymptotically equivalent under the null hypothesis, but they can behave differently in a small sample.

- All LR, LM (Score), Wald test statistics asymptotically follow

$$\chi^2(J), \quad J = \text{number of restrictions under } \; H_0$$

Suppose $H_0 : c(\theta) = q$.

Theorem:**[LR test]** $\lambda = L_R / L_U$. We have $-2 \ln \lambda = 2 (\ln L_U - \ln L_R) \xrightarrow[H_0]{d} \chi^2_J$.

Theorem:**[Wald test]** Wald statistic

$$w = \left[c\left(\hat{\theta}\right) - q\right]' \left(AVar\left[c\left(\hat{\theta}\right) - q\right]\right)^{-1} \left[c\left(\hat{\theta}\right) - q\right] \xrightarrow[H_0]{d} \chi^2_J,$$

where

$$AVar\left[c\left(\hat{\theta}\right) - q\right] = C \;\, AVar\left[\hat{\theta}\right] \;\, C',$$

$$C = \frac{\partial c(\theta)}{\partial \theta}$$

or in practice

$$A\hat{V}ar\left[c\left(\hat{\theta}\right) - q\right] = \hat{C} \;\, AVar\left[\hat{\theta}\right] \;\, \hat{C}',$$

$$\hat{C} = \frac{\partial c(\theta)}{\partial \theta}|_{\theta = \hat{\theta}}.$$

For the LM test, write the Lagrangian

$$\ln L^* (\theta) = \ln L (\theta) + \lambda' (c (\theta) - q)$$

FOC (define $C = \frac{\partial c(\theta)}{\partial \theta}$):

$$\begin{aligned} \frac{\partial \ln L^* (\theta)}{\partial \theta} &= \frac{\partial \ln L (\theta)}{\partial \theta} + C' \lambda = 0 \\ \frac{\partial \ln L^* (\theta)}{\partial \lambda} &= c (\theta) - q = 0, \end{aligned}$$

If the restrictions are valid, then imposing them will not lead to a significant difference in the maximised value of the likelihood function. In the FOC, this means that the second term in the derivative vector should be small. In particular, λ will be small. We could test this, i.e. $H_0 : \lambda = 0$, which leads to the ML test.

An equivalent and simpler formulation is the following. At the restricted maximum

$$\frac{\partial \ln L\left(\hat{\theta}_R\right)}{\partial \theta} = -\hat{C}'\hat{\lambda} \equiv \hat{g}_R.$$

If the restrictions are valid, within the range of sampling variability, then $\hat{g}_R = 0$. That is, the derivatives of the log-likelihood evaluated at the restricted parameter will be approximately zero.[1] The vector of first derivatives is the score function. The variance of the score function is the information matrix, which we use to find the asymptotic covariance matrix of the MLE.

Theorem:**[LM test]** LM test

$$LM = \left[\frac{\partial \ln L\left(\hat{\theta}_R\right)}{\partial \hat{\theta}_R}\right]' \left(I\left(\hat{\theta}_R\right)\right)^{-1} \left[\frac{\partial \ln L\left(\hat{\theta}_R\right)}{\partial \hat{\theta}_R}\right] \xrightarrow[H_0]{d} \chi^2_J.$$

[1] Below we do this for the linear regression model, where the score function is $g = X'\varepsilon$.

F-test as an LR test

The LR test of the null hypothesis compares L_U, the maximised likelihood of the unrestricted model, and L_R, the maximised likelihood of the restricted model.

If the LR $\lambda \equiv L_U/L_R$ is too large then the null might be wrong.

The F-test of the $H_0 : R\beta = r$ is an LR test since it's a monotone transformation of λ.

$$
\begin{aligned}
L_R &= \left(\frac{2\pi}{n}\right)^{-n/2} \exp\left(-\frac{n}{2}\right) SSR_R^{-n/2} \\
L_U &= \left(\frac{2\pi}{n}\right)^{-n/2} \exp\left(-\frac{n}{2}\right) SSR_U^{-n/2},
\end{aligned}
$$

Hence

$$\lambda \equiv \left(\frac{SSR_U}{SSR_R}\right)^{-n/2} \iff \lambda^{n/2} \equiv \left(\frac{SSR_R}{SSR_U}\right)$$

and

$$\begin{aligned} F &= \frac{\left(SSR_R - SSR_U\right)/J}{SSR_U/\left(n-K\right)} = \frac{\left(\frac{SSR_R}{SSR_U} - 1\right)/J}{1/\left(n-K\right)} \\ &= \frac{\left(\lambda^{n/2} - 1\right)/J}{1/\left(n-K\right)} \end{aligned}$$

so that two tests are the same.

Quasi-Maximum Likelihood

Without the normality assumption $\hat{\beta}_{ML}$ may not be the LS estimator or that the LS estimator achieves the CR lower bound. But the LS estimator b is a quasi-ML-estimator, an estimator that maximises a misspecified likelihood function, which is assumed to be normal.

Generalised Least Squares

Generalised Linear Regression Model
We extend the linear regression model by allowing a more general form for the variance/covariance matrix of the error term. Specifically, we consider the (data generating) model

$$y = X\beta + u, \;\; E\left[uu'|X\right] = \sigma^2 V\left(X\right) \neq \sigma^2 I_n;$$

where $V\left(X\right)$ is a symmetric and positive definite matrix and a function of X, i.e. non-spherical errors. We write V for convenience.

Example:

Under heteroskedasticity with no autocorrelation,

$$V = \begin{bmatrix} h_1 & 0 & \cdots & 0 \\ 0 & h_2 & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ 0 & 0 & \vdots & h_n \end{bmatrix} \neq I_n.$$

Under autocorrelation with homoskedasticity,

$$V = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \vdots & \vdots & & \ddots & \vdots \\ \rho_{n-1} & \cdots & \rho_2 & \rho_1 & 1 \end{bmatrix} \neq I_n.$$

Estimation: GLS Estimator

Consequences for Least Squares estimator: Not BLUE

Under the assumption of strict exogeneity (zero conditional mean), $E\left[\varepsilon|X\right]=0$, the LS estimator $b=\left(X'X\right)^{-1}X'y$ is unbiased and consistent; $E\left[b\right]=E\left[b|X\right]=\beta$.

The general form of the (conditional) variance/covariance matrix of the the LS estimator b is

$$\begin{aligned}
Var\left[b|X\right] &= E\left[\left(b-E\left[b|X\right]\right)\left(b-E\left[b|X\right]\right)'|X\right] \\
&= E\left[\left(b-\beta\right)\left(b-\beta\right)'|X\right] \\
&= E\left[\left(\left(X'X\right)^{-1}X'u\right)\left(\left(X'X\right)^{-1}X'u\right)'|X\right] \\
&= E\left[\left(X'X\right)^{-1}X'uu'X\left(X'X\right)^{-1}|X\right] \\
&= \left(X'X\right)^{-1}X'E\left[uu'|X\right]X\left(X'X\right)^{-1} \\
\Longrightarrow \quad Var\left[b|X\right] &= \sigma^2\left(X'X\right)^{-1}X'VX\left(X'X\right)^{-1}.
\end{aligned}$$

Notice that $V = I_n \implies Var\left[b|X\right] = \sigma^2 \left(X'X\right)^{-1}$. But in general this is not the case. Therefore, using $se\left(b_k\right) = \sqrt{s^2 \left[\left(X'X\right)^{-1}\right]_{kk}}$ is not valid; it's biased and inconsistent. Similarly, using $se\left(b_k\right) = \sqrt{s^2 \left[\left(X'X\right)^{-1}\right]_{kk}}$ implies that the t-statistic does not, under normality, follow t-distribution or does not converge to the normal distribution asymptotically. The same result holds for the F-statistic and Wald-statistic.

LS is Not BLUE:

Since $V \neq I_n$ violates the GM theorem premises; the GM theorem requires $Var\left[u|X\right] = \sigma^2 I_n$. Hence, b it's not MVUE, in the framework of MLE.