

Review of Maximum Likelihood Estimation

- $\{x_1, \dots, x_T\}$ i.i.d. Call that the vector, \mathbf{x} .
- $x \sim$ continuously with density $f(x|\theta)$
- θ is unknown parameter vector that determines the distribution
- Joint Density: $f(x_1|\theta)f(x_2|\theta)\dots f(x_T|\theta)$
- This joint density is the *Likelihood Function*, $L(\mathbf{x}|\theta)$.
- Idea: pick $\hat{\theta}$ to maximize the likelihood $L(\mathbf{x}|\theta)$, of observing the particular realizations \mathbf{x} .

Example: OLS with Normal Distribution

- sample \mathbf{y} of size T is normally distributed with mean $\mathbf{X}\beta$ where
- \mathbf{X} is an $T \times K$ matrix of explanatory variables.
- β is an $K \times 1$ vector of parameters.
- The variance-covariance matrix of the errors from the true regression is $\sigma^2 I$, where I is an $N \times N$ identity matrix.
- The likelihood function is:

$$L(y; X\beta, \sigma^2 I) = \frac{1}{(2\pi\sigma^2)^{T/2}} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^T (y_i - X_i\beta)^2 \right]$$

First-order Conditions: $\hat{\beta}$

$$\begin{aligned}\frac{\partial \ln(L)}{\partial \beta} &= -\frac{1}{2\sigma^2} \frac{\partial}{\partial \beta} [y'y - 2y'X\beta + \beta'X'X\beta] \\ &= \frac{1}{\sigma^2} (X'y - X'X\beta)\end{aligned}$$

Setting this to zero and solving for $\hat{\beta}$ yields:

$$\hat{\beta} = (X'X)^{-1}X'y$$

This is the OLS estimator. Also:

First-order Conditions: $\hat{\sigma}^2$

$$\frac{\partial \ln(L)}{\partial \sigma^2} = -\frac{T}{(2\sigma^2)} + \frac{1}{2\sigma^4} [(y - X\beta)'(y - X\beta)]$$

Setting this to zero and solving for $\hat{\sigma}^2$ yields:

$$\hat{\sigma}^2 = \frac{1}{T} [(y - X\hat{\beta})'(y - X\hat{\beta})],$$

which is just the sample average squared residual.

The Information Matrix

- If θ is our parameter vector,
 - $I(\theta)$ is the *information matrix*,
 - which is minus the expectation of the matrix of second partial derivatives of the log-likelihood with respect to the parameters.

The Information Matrix — Continued

The MLE achieves the Cramer–Rao lower bound, which means that the variance of the estimators equals the inverse of the information matrix,

$$I^{-1}(\mu, \sigma^2).$$

Now,

$$\begin{aligned}
 I(\mu, \sigma^2) &= -E \begin{bmatrix} \frac{\partial^2 \ln L}{\partial \beta^2} & \frac{\partial^2 \ln L}{\partial \beta \partial \sigma^2} \\ \frac{\partial^2 \ln L}{\partial \beta \partial \sigma^2} & \frac{\partial^2 \ln L}{\partial (\sigma^2)^2} \end{bmatrix} \\
 &= -E \begin{bmatrix} -\frac{1}{\sigma^2} X' X & \frac{1}{\sigma^4} (X' y - X' X \beta) \\ \frac{1}{\sigma^4} (y' X' - \beta' X' X) & \frac{T}{\sigma^4} - \frac{1}{\sigma^6} (y - X \beta)' (y - X \beta) \end{bmatrix}
 \end{aligned}$$

The Information Matrix — Continued

Taking the negative of the expectation yields:

$$\begin{bmatrix} \frac{1}{\sigma^2} X'X & 0 \\ 0 & \frac{T}{2\sigma^4} \end{bmatrix}$$

The inverse of this is:

$$\begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{T} \end{bmatrix}$$

Another way of writing $I(\mu, \sigma^2)$

For a vector, θ , of parameters, $I(\theta)$, the information matrix, can be written two ways:

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

The second form is more convenient for estimation, because it does not require estimating second derivatives.

Estimation

The Likelihood Ratio Test

- Let θ be a vector of parameters to be estimated.
- Let H_0 be a set of restrictions on these parameters.
- These restrictions could be linear or non-linear.
- Let $\hat{\theta}_U$ be the MLE of θ estimated without regard to constraints (the unrestricted model).
- Let $\hat{\theta}_R$ be the constrained MLE.

The Likelihood Ratio Test Statistic

If $\hat{L}_U(\hat{\theta}_U)$ and $\hat{L}_R(\hat{\theta}_R)$ are the likelihood functions evaluated at these two estimates, the likelihood ratio is given by

$$\lambda = \frac{\hat{L}_R(\cdot)}{\hat{L}_U(\cdot)}.$$

Then, $-2\ln(\lambda) = -2(\ln(\hat{L}_R) - \ln(\hat{L}_U)) \sim \chi^2$ with degrees of freedom equal to the number of restrictions imposed.

Another look at the LR Test

Concentrated Log-Likelihood Many problems can be formulated in terms of partitioning a parameter vector, θ into $\{\theta_1, \theta_2\}$ such that the solution to the optimization problem, $\hat{\theta}_2$ can be written as a function of $\hat{\theta}_1$, e.g.

$$\hat{\theta}_2 = t(\hat{\theta}_1).$$

Then, we can concentrate the log-likelihood function by writing

$$F^*(\theta_1, \theta_2) = F(\theta_1, t(\theta_1)) \equiv F_c(\theta_1)$$

Why do this?

The unrestricted solution to

$$\text{Max}_{\theta_1} F_c(\theta_1)$$

then provides the full solution to the optimization problem, as t is known.

We now use this technique to find estimates for the classical linear regression model.

Example

The log-likelihood function for our CLM with normal disturbances is given by

$$\ln(L) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\sigma^2) - \frac{1}{(2\sigma^2)} [(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)]$$

The solution to the likelihood equation for $\hat{\sigma}^2$ implies that however we estimate β , the estimator will be

$$\hat{\sigma}^2 = \frac{1}{T} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

Concentrating the Likelihood Function

Inserting this back into the log-likelihood yields:

$$\ln(L_c) = -\frac{T}{2} \left[1 + \ln(2\pi) + \ln \left(\frac{1}{T} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \right) \right]$$

Because $(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$ is just the sum of squared residuals from the regression

$$\mathbf{e}'\mathbf{e}$$

we can rewrite $\ln(L_c)$ as

$$\ln(L_c) = -\frac{T}{2} \left[1 + \ln(2\pi) + \ln \left(\frac{1}{T} (\mathbf{e}'\mathbf{e}) \right) \right]$$

The LR Test - redux

For the restricted model, we obtain the restricted concentrated log-likelihood $\ln(L_{cR})$:

$$\ln(L_{cR}) = -\frac{T}{2} \left[1 + \ln(2\pi) + \ln \left(\frac{1}{T} (\mathbf{e}'_{\mathbf{R}} \mathbf{e}_{\mathbf{R}}) \right) \right]$$

So, plugging in these concentrated log-likelihoods into our definition of the LR test, we obtain

$$LR = T \ln \left[\frac{\mathbf{e}'_{\mathbf{R}} \mathbf{e}_{\mathbf{R}}}{\mathbf{e}' \mathbf{e}} \right]$$

or T times the log of the ratio of the restricted SSR and the unrestricted SSR, a nice intuition.

Example — OLS with Normal Errors, ϵ_t

- True regression model:

$$y_t = \alpha + \beta x_t + \epsilon_t$$

- ϵ_t i. i. d. normal.
- sample size T .
- restriction: $\alpha = 1$.

Example — continued

$$\ln L(\theta_U) = -\frac{T}{2} \ln(2\pi) - \frac{T}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=1}^T (y_t - \alpha - \beta x_t)^2$$

The first-order conditions for the estimates $\hat{\alpha}$ and $\hat{\beta}$ simply reduce to the OLS normal equations:

$$\hat{\beta} : \sum_{t=1}^T (y_t - \hat{\alpha} - \hat{\beta} x_t) x_t = 0$$

$$\hat{\alpha} : \sum_{t=1}^T (y_t - \hat{\alpha} - \hat{\beta} x_t) = 0$$

Example — continued

Solving:

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

Substituting into the FOC for $\hat{\beta}$ yields:

$$\hat{\beta} = \frac{\sum_{t=1}^T ((x_t - \bar{x})(y_t - \bar{y}))}{\sum_{t=1}^T (x_t - \bar{x})^2}$$

Example — continued

Solving for $\hat{\sigma}^2$ is as before:

$$\hat{\sigma}^2 = \frac{1}{T} \left[\sum_{t=1}^T (y_t - \hat{\alpha} - \hat{\beta}x_t) \right]^2$$

Example — continued

The restricted model is exactly the same, except that $\hat{\alpha}$ is constrained to be one, so that the normal equation:

$$\hat{\beta}_R : \sum_{t=1}^T (y_t - \hat{\alpha}_R - \hat{\beta}_R x_t) x_t = 0$$

reduces to

$$\hat{\beta}_R : \sum_{t=1}^T (y_t - 1 - \hat{\beta}_R x_t) x_t = 0$$

and

$$\hat{\beta}_R = \frac{\sum_{t=1}^T (x_t y_t - x_t)}{\sum_{t=1}^T x_t^2}$$

One can then plug in to obtain $\hat{\sigma}_R^2$ and form the likelihood ratio, which is distributed $\chi^2(1)$.

The Wald Test

- Problem with LR: Need both restricted and unrestricted model estimates.
- One or other could be hard to compute.
- The Wald test is an alternative that requires estimating the unrestricted model only.
- Suppose $\mathbf{y} \sim N_T(\mathbf{X}\beta, \Sigma)$ Then,

$$(\mathbf{y} - \mathbf{X}\beta)' \Sigma^{-1} (\mathbf{y} - \mathbf{X}\beta) \sim \chi_T^2$$

The Wald Test — continued

- Under the null hypothesis that $E(\mathbf{y}) = \mathbf{X}\beta$, the quadratic form has the χ^2 distribution. If the hypothesis is false, the quadratic form will be larger, on average, than if it were true.
- In particular, it will be a non-central χ^2 with the same d.f., which looks like a central χ^2 , but lies to the right.
- This is the basis for our test.

The Restricted Model

- Now, step back from the normal and let $\hat{\theta}$ be the parameter estimates from the unrestricted model.
- Let restrictions be given by

$$H_0 : f(\theta) = 0.$$

- If the restrictions are valid, then $\hat{\theta}$ should satisfy them.
- If not, $f(\hat{\theta})$ should be farther from zero than would be explained by sampling error alone.

Formalism

The Wald Statistic is

$$W = [f(\hat{\theta})'(Var[f(\hat{\theta})])^{-1}[f(\hat{\theta})]$$

Under H_0 in large samples, $W \sim \chi^2$ with d.f. equal to the number of restrictions. See Greene ch. 9 for details.

Last, to use the Wald test, we need to compute the variance term:

$$Var[f(\hat{\theta})] = G(\hat{\theta})Var[\hat{\theta}]G(\hat{\theta})'$$

$$G(\hat{\theta}) = \left[\frac{\partial f(\hat{\theta})}{\partial \hat{\theta}'} \right]$$

Restrictions on slope coefficients

If the restrictions are on slope coefficients of a linear regression, then

$$\text{Var}[\hat{\theta}] = \text{Var}[\hat{\beta}] = s^2(\mathbf{X}'\mathbf{X})^{-1}$$

where

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{T - K} = \frac{T}{T - K} \hat{\sigma}^2.$$

K is the number of regressors.

- Then, we can write the Wald Statistic:

$$W = \mathbf{f}(\hat{\beta})' (\mathbf{G}(\hat{\beta}) [s^2 \mathbf{X}'\mathbf{X}]^{-1} \mathbf{G}(\hat{\beta})')^{-1} \mathbf{f}(\hat{\beta}) \rightarrow \chi^2[J]$$

where J is the number of restrictions.

Linear restrictions

$$H_0 : \mathbf{R}\beta - \mathbf{q} = \mathbf{0}$$

For example, suppose there were three betas, β_1 , β_2 , and β_3 . Let's look at three tests:

1. $\beta_1 = 0$,
2. $\beta_1 = \beta_2$
3. $\beta_1 = 0$ and $\beta_2 = 2$

Each row of \mathbf{R} is a single linear restriction on the coefficient vector.

Writing $R\beta$

- Case 1:

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \quad \mathbf{q} = \begin{bmatrix} 0 \end{bmatrix}$$

- Case 2:

$$\mathbf{R} = \begin{bmatrix} 1 & -1 & 0 \end{bmatrix} \quad \mathbf{q} = \begin{bmatrix} 0 \end{bmatrix}$$

$$\mathbf{R} = [1 \ -1 \ 0] \quad \mathbf{q} = 0$$

- Case 3:

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad \mathbf{q} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

The Wald Statistic

In general, the Wald statistic with J linear restrictions reduces to

$$W = [\mathbf{R}\hat{\beta} - \mathbf{q}]' [\mathbf{R}s^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} [\mathbf{R}\hat{\beta} - \mathbf{q}]$$

with J d.f. We will use these tests extensively in our discussion of chapters 5 and 6 of CLM.

The F Test

A related test for testing the validity of the J restrictions

$$\mathbf{R}\beta - \mathbf{q} = \mathbf{0}$$

Recall that the F test can be written in terms of a comparison of SSR for the restricted and unrestricted models:

$$F(J, T - K) = \frac{(\mathbf{e}'_{\mathbf{R}}\mathbf{e}_{\mathbf{R}} - \mathbf{e}'\mathbf{e})/J}{\mathbf{e}'\mathbf{e}/(T - K)}$$

or

$$F(J, T - K) = \frac{[\mathbf{R}\hat{\beta} - \mathbf{q}]'[\mathbf{R}s^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}[\mathbf{R}\hat{\beta} - \mathbf{q}]}{J}$$

Why do we care?

We care because in a linear model with normally distributed disturbances, under the null the test statistic derived above is *exact*.

- This will be important later because under normality, some of our cross-sectional CAPM tests will be of this form and
- A sufficient condition for the (static) CAPM to be “correct,” is for asset returns to be normally distributed.

The LM Test

A test that involves computing only the *restricted* least-squares estimator.

- If hypothesis is valid, at the restricted estimator, the derivative of the log-likelihood function should be close to zero.
- We will next form the LM test with the J restrictions $\mathbf{f}(\theta) = \mathbf{0}$.

The LM Test — continued

$$\ln(L_{LM}) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\sigma^2) - \frac{1}{(2\sigma^2)} [(y - X\beta)'(y - X\beta)] + \lambda' \mathbf{f}(\beta)$$

This is maximized by choice of $\hat{\beta}$ and $\hat{\sigma}^2$.

First-order Conditions

$$\frac{\partial \ln(L_{LM})}{\partial \beta} = \frac{1}{\sigma^2} \mathbf{X}'(\mathbf{y} - \mathbf{X}\beta + \frac{\partial \mathbf{f}(\beta)'}{\partial \beta'} \lambda) = \mathbf{0},$$

$$\frac{\partial \ln(L_{LM})}{\partial \sigma^2} = -\frac{T}{(2\sigma^2)} + \frac{1}{2\sigma^4} [(y - X\beta)'(y - X\beta)] = 0$$

and

$$\mathbf{f}(\beta) = \mathbf{0}$$

The LM Test — continued

The test then, is whether the Lagrange multipliers equal zero. When the restrictions are linear, the test statistic becomes (see Greene, chapter 7)

$$W = [\mathbf{R}\hat{\beta} - \mathbf{q}]'[\mathbf{R}s_R^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}[\mathbf{R}\hat{\beta} - \mathbf{q}]$$

where J is the number of restrictions.

W, LR, LM, and F

We compare them for J linear restrictions in the linear model with K regressors. It can be shown that

- $W = \frac{T}{T-K} JF$,
- $LR = T \ln \left[1 + \frac{1}{T-K} JF \right]$,
- $LM = \frac{T}{(T-K)[1+(1/(T-K))JF]} JF$,

and that

$$W > LR > LM$$