EC402: MEI. Spring Term: The Analysis of Time Series*

Lecture Notes #3: Hypothesis Tests within the Maximum Likelihood Framework

There are three main frequentist[1] approaches to inference within the Maximum Likelihood Framework, the Wald test, the Likelihood ratio test and the Lagrange Multiplier or LM test.

1. Wald Tests

Wald tests are based on the Maximum Likelihood Estimates of the unrestricted model. Suppose we have a model with k unknown parameters ψ that can be expressed and estimated in terms of a log likelihood $\log L(\psi)$. Then the ML estimates of ψ will have limiting distribution

$$\sqrt{T}(\underbrace{\hat{\psi} - \psi}_{k\times 1}) \to N\left(0, \underbrace{IA(\psi)^{-1\prime}}_{k\times k}\right)$$

where $IA(\psi) = \lim \frac{1}{T} I(\psi)$, $I(\psi) = -E\left(\frac{\partial^2 \log L(\psi)}{\partial\psi\partial\psi'}\right)$ (under regularity conditions, this is the *information matrix*).

If we want to consider a linear hypothesis $H_0 : R\psi = q$ against the alternative H_A :

[1] Bayesian inference will not be introduced at this stage.

$R\psi \neq q$, where R has $r < k$ linearly independent rows (r restrictions), then under H_0

$$\sqrt{T}R(\hat{\psi} - \psi) = \sqrt{T}(R\hat{\psi} - q) \xrightarrow{d} N\left(0, \underbrace{RIA(\psi)^{-1}R'}_{r \times r}\right)$$

and[2]

$$\sqrt{T}(R\hat{\psi} - q)' \left[RIA(\psi)^{-1}R'\right]^{-1} \sqrt{T}(R\hat{\psi} - q) \xrightarrow{d} \chi^2(r).$$

As usual we do not observe $IA(\psi)$, but if we can find a consistent estimator, the distribution remains unchanged. Possible estimators are:

1. The empirical information matrix based

$$\frac{1}{T}I(\hat{\psi})$$

2. Empirical Hessian based

$$-\frac{1}{T} \bullet \frac{\partial^2 \log L(\hat{\psi})}{\partial\psi\partial\psi'}.$$

Assuming the first is available, then

$$\begin{aligned} W &= \sqrt{T}(R\hat{\psi} - q)' \left[R\left(\frac{1}{T}I(\hat{\psi})\right)^{-1} R'\right]^{-1} \sqrt{T}(R\hat{\psi} - q) \\ &= (R\hat{\psi} - q)' \left[RI(\hat{\psi})^{-1}R'\right]^{-1} (R\hat{\psi} - q) \xrightarrow{d} \chi^2(r). \end{aligned}$$

So we can use the $\chi^2_\alpha(r)$ tabulated values to test at the α significance level (reject the null if the test statistic is bigger that the tabulated value).

1.1. Nonlinear Constraints

The Wald approach can be extended to nonlinear constraints. The approach is straightforward: linearize the constraints about the null using a first order Taylor expansion and then apply the linear theory above.

So consider $H_0 : R(\psi) = 0$, a set of r linear or nonlinear constraints. R is a column r-vector.

[2]Recall, if the n-dimensional vector $x \sim N(0, A) \Rightarrow x'A^{-1}x \sim \chi^2(n)$.

Let $\frac{\partial R}{\partial \psi} = \left[\frac{\partial R}{\partial \psi_1}, \frac{\partial R}{\partial \psi_2}, ..., \frac{\partial R}{\partial \psi_k}\right]$ be a $r \times k$ matrix, where k is the number of parameters in ψ. Then the statistic

$$W = R(\hat{\psi})' \left[\left(\frac{\partial R(\hat{\psi})}{\partial \psi} \right) I(\hat{\psi})^{-1} \left(\frac{\partial R(\hat{\psi})}{\partial \psi} \right)' \right]^{-1} R(\hat{\psi})$$

is asymptotically $\chi^2(r)$ under H_0.

For example, suppose we have

$$H_0 : \psi_1 \psi_2 \psi_3 = 1$$

$$\psi_3 = 4\psi_4 - 2$$

where $k = 4$. So

$$\begin{aligned} R(\psi) &= \begin{bmatrix} \psi_1\psi_2\psi_3 - 1 \\ \psi_3 - 4\psi_4 + 2 \end{bmatrix} \\ \frac{\partial R}{\partial \psi} &= \begin{bmatrix} \psi_2\psi_3 & \psi_1\psi_3 & \psi_1\psi_2 & 0 \\ 0 & 0 & 1 & -4 \end{bmatrix}. \end{aligned}$$

So the Wald statistic is

$$\begin{aligned} &\left(\hat{\psi}_1\hat{\psi}_2\hat{\psi}_3 - 1, \hat{\psi}_3 - 4\hat{\psi}_4 + 2\right) \left[\begin{pmatrix} \hat{\psi}_2\hat{\psi}_3 & \hat{\psi}_1\hat{\psi}_3 & \hat{\psi}_1\hat{\psi}_2 & 0 \\ 0 & 0 & 1 & -4 \end{pmatrix} I(\hat{\psi})^{-1} \right. \\ &\left. \begin{pmatrix} \hat{\psi}_2\hat{\psi}_3 & 0 \\ \hat{\psi}_1\hat{\psi}_3 & 0 \\ \hat{\psi}_1\hat{\psi}_2 & 1 \\ 0 & -4 \end{pmatrix} \right]^{-1} \begin{pmatrix} \hat{\psi}_1\hat{\psi}_2\hat{\psi}_3 - 1 \\ \hat{\psi}_3 - 4\hat{\psi}_4 + 2 \end{pmatrix}. \end{aligned}$$

2. The Likelihood Ratio Test

Again suppose that we have a model with unknown parameters that can be expressed in terms of a likelihood function $L(\psi)$. Suppose we also have a set of r restrictions, either linear

$$R\psi = q$$

or nonlinear

$$R(\psi) = 0.$$

The procedure is to estimate the unrestricted model to obtain ML estimates, $\hat{\psi}$, and hence maximized likelihood $L(\hat{\psi})$. Then estimate the model under the restrictions to obtain restricted estimates, $\hat{\psi}_0$, and the likelihood $L(\hat{\psi}_0)$. Then compare the two.

It can be shown that under the null

$$LR = -2\log\left\{\frac{L(\hat{\psi}_0)}{L(\hat{\psi})}\right\} = 2\left\{\log L(\hat{\psi}) - \log L(\hat{\psi}_0)\right\}$$

is asymptotically $\chi^2(r)$. Obviously LR is proportional to $\log\left\{\frac{L(\hat{\psi}_0)}{L(\hat{\psi})}\right\}$, hence LR is greater than 0. If the data conforms with the null you expect $L(\hat{\psi})$ to be close to $L(\hat{\psi}_0)$ and for LR to be close to 0. If the data does not conform you expect $L(\hat{\psi}) >> L(\hat{\psi}_0)$ and $LR >> 0$. Hence the test is to reject H_0 at the α level if $LR > \chi^2_\alpha(r)$.

3. Lagrange Multiplier Tests

The LM test is based on the restricted estimates. Again suppose that we have a model with unknown parameters that can be expressed in terms of likelihood $L(\psi)$. Suppose we have r restrictions $R(\psi) = 0$.

Let $\hat{\psi}_0$ denote the ML estimator of ψ in the restricted model. If the restrictions are valid $\hat{\psi}_0$ will be close to $\hat{\psi}$ and the partial derivatives in the vector $\frac{\partial \log L(\hat{\psi}_0)}{\partial \psi}$ will also be close to zero.[3] It can be shown that under the null, the quadratic form

$$LM = \frac{1}{T}\frac{\partial \log L(\hat{\psi}_0)}{\partial \psi'} IA(\psi_0)^{-1} \frac{\partial \log L(\hat{\psi}_0)}{\partial \psi} \xrightarrow{d} \chi^2(r).$$

As usual, this is often not an operational statistic as we do not know $IA(\psi_0)$ and this must be replaced by a consistent estimate. Assuming that $\frac{1}{T}I(\hat{\psi})$ or a consistent alternative is available, then

$$\frac{\partial \log L(\hat{\psi}_0)}{\partial \psi'} I(\hat{\psi}_0)^{-1} \frac{\partial \log L(\hat{\psi}_0)}{\partial \psi} \xrightarrow{d} \chi^2(r) \tag{3.1}$$

and is referred to as a Lagrange Multiplier statistic.

[3] Note that

$$\frac{\partial \log L(\hat{\psi})}{\partial \psi} = 0$$

by first order optimality condition.

3.1. The LM test in Nonlinear Least Squares

This result can be specialised for nonlinear least squares problems. Thus we have

$$y_t = g(x_t;\beta) + \varepsilon_t, \qquad \varepsilon_t \text{ iid } N(0,\sigma^2)$$

$$x_t \text{ independent of } \varepsilon_t, \qquad t = 1, ..., T.$$

Then the unrestricted log likelihood has the form

$$\log L(\beta,\sigma^2) = -\frac{T}{2}\log 2\pi - \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\varepsilon_t(\beta)^2,$$

where

$$\varepsilon_t(\beta) = y_t - g(x_t;\beta).$$

Assume that the r restrictions involve only β not σ^2, so they have the form $R(\beta) = 0$. Then

$$\begin{aligned} \frac{\partial \log L(\beta,\sigma^2)}{\partial\beta} &= \frac{1}{\sigma^2}\sum_t z_t\varepsilon_t, \\ z_t &= -\frac{\partial\varepsilon_t}{\partial\beta}. \end{aligned} \tag{3.2}$$

As before,

$$\frac{1}{T}I(\psi) = -E\left[\frac{1}{T}\frac{\partial^2 \log L(\psi)}{\partial\psi\partial\psi'}\right]$$

but as σ^2 is not in the restriction the information matrix is block diagonal. Consider only the sub matrix associated with β. Since x_t is independent of ε_t,

$$I_{\beta\beta}(\psi) = -E\left[\frac{\partial^2 \log L}{\partial\beta\partial\beta'}\right] = \frac{1}{\sigma^2}E\sum_t z_t z_t'. \tag{3.3}$$

Substituting (3.2) and (3.3) into (3.1), evaluating at $(\hat\beta_0, \hat\sigma_0^2)$, where $\hat\sigma_0^2 = \frac{1}{T}\sum\varepsilon_t^2(\hat\beta_0)$, and replacing $E\sum z_t z_t'$ by $\sum z_t z_t'$,

$$\begin{aligned} LM &= \frac{1}{(\hat\sigma_0^2)^2}\left(\sum z_t\varepsilon_t\right)'\left[\frac{1}{\hat\sigma_0^2}\sum z_t z_t'\right]^{-1}\left(\sum z_t\varepsilon_t\right) \\ &= \frac{1}{\hat\sigma_0^2}\left(\sum z_t\varepsilon_t\right)'\left[\sum z_t z_t'\right]^{-1}\left(\sum z_t\varepsilon_t\right). \end{aligned}$$

By inspection LM is related to the regression of ε_t on z_t (i.e. $\varepsilon_t = z_t'\gamma + u_t$, $\hat{\gamma} = (\Sigma z_t z_t')^{-1}\Sigma z_t \varepsilon_t$). Define fitted values for such a regression as

$$\eta_t = z_t'\hat{\gamma} = z_t' \left[\sum z_t z_t'\right]^{-1} \left(\sum z_t \varepsilon_t\right).$$

Now consider the R^2 from this regression

$$\begin{aligned} T \times R^2 &= \frac{\eta'\eta}{\frac{1}{T}\varepsilon'\varepsilon} \\ &= \frac{\left(\sum z_t\varepsilon_t\right)' \left[\sum z_t z_t'\right]^{-1} \left[\sum z_t z_t'\right] \left[\sum z_t z_t'\right]^{-1} \left(\sum z_t \varepsilon_t\right)}{\hat{\sigma}_0^2} \\ &= LM. \end{aligned}$$

Hence a valid LM statistic can always be obtained by regressing $\varepsilon_t(\hat{\psi}_0)$ on $z_t(\hat{\psi}_0)$ and calculating $LM^* = TR^2$. Then reject H_0 at the α level if $LM^* > \chi^2_\alpha(r)$.

Example: Testing for AR(1) error

The model:

$$\begin{aligned} y_t &= x_t'\beta + u_t \\ u_t &= \phi u_{t-1} + \varepsilon_t, \qquad |\phi| < 1, \qquad \varepsilon_t \text{ iid } (0, \sigma^2). \end{aligned}$$

x_t contemporaneously independent of ε_t.

This implies

$$y_t = \phi y_{t-1} + x_t'\beta - \phi x_{t-1}'\beta + \varepsilon_t.$$

We want to test

$$\begin{aligned} H_0 &: \phi = 0 \text{ against} \\ H_A &: \phi \neq 0. \end{aligned}$$

Define

$$\begin{aligned} \varepsilon_t(\beta, \phi) &= y_t - \phi y_{t-1} - (x_t - \phi x_{t-1})'\beta \\ z_t(\beta, \phi) &= -\begin{bmatrix} \frac{\partial \varepsilon_t}{\partial \beta} \\ \frac{\partial \varepsilon_t}{\partial \phi} \end{bmatrix} = \begin{bmatrix} x_t - \phi x_{t-1} \\ y_{t-1} - x_{t-1}'\beta \end{bmatrix}. \end{aligned}$$

Estimate the model *under the null*. ML is least squares on $y_t = x_t'\beta + u_t$, and calculate the least squares residuals $\hat{u}_t$. Then evaluating under the null

$$\begin{aligned} \varepsilon_t(\hat{\beta}, 0) &= y_t - x_t'\hat{\beta} = \hat{u}_t \\ z_t(\hat{\beta}, 0) &= \begin{bmatrix} x_t \\ \hat{u}_{t-1} \end{bmatrix}. \end{aligned}$$

Calculate the LM statistic by regressing $\hat{u}_t$ on $(x_t, \hat{u}_{t-1})$, obtain TR^2, reject if $TR^2 > \chi^2_\alpha(1)$. Note this is valid even if x contains the lagged dependent variable, since it requires only contemporaneous independence.

4. Comparison between the Wald, LR and LM tests

The first point is that all three are asymptotically equivalent. The issue in finite samples is difficult. However it is possible to give some tentative conclusions that have emerged from several studies.

The basic conclusion is that in general the LR test is the best, in the sense that its finite sample behaviour most closely approximates its expected large sample properties. The Wald test is second best and the LM procedure worst. However, there are exceptions.

5. Durbin Watson Test

You may be a little surprised that we have spent so much time discussing tests of serial correlation without mentioning the Durbin Watson test.

The Durbin Watson test is the only test for which we have small sample properties. Unfortunately the circumstances in which it is valid are so restricted that it is almost always inappropriate.

The model:

$$\begin{aligned} y_t &= x_t'\beta + u_t \\ u_t &= \phi u_{t-1} + \varepsilon_t, \qquad \varepsilon_t \text{ iid } N(0, \sigma^2). \end{aligned}$$

We want to test

$$\begin{aligned} H_0 &: \phi = 0 \text{ against} \\ H_A &: \phi > 0. \end{aligned}$$

Under the null, estimate the model by least squares and calculate the test statistic

$$d = \frac{\sum_{t=2}^{T} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T} \hat{u}_t^2} = \frac{\sum_{2}^{T} \hat{u}_t^2}{\sum_{1}^{T} \hat{u}_t^2} + \frac{\sum_{2}^{T} \hat{u}_{t-1}^2}{\sum_{1}^{T} \hat{u}_t^2} - 2\frac{\sum_{2}^{T} \hat{u}_t \hat{u}_{t-1}}{\sum_{1}^{T} \hat{u}_t^2}.$$

Note that $d \approx 2(1 - r_1)$, where r_1 is the simple correlation between $\hat{u}_t$ and $\hat{u}_{t-1}$. d lies in the interval [0,4] with $r_1 = 0$ corresponding to $d = 2$ and $r_1 = 1$ corresponding to $d = 0$.

Unfortunately the exact distribution of d depends on X and as a result it cannot be tabulated concisely. However, for a specific X it is possible to obtain the exact distribution of d numerically and in particular to calculate the critical point d_α such that under the null $\Pr ob(d < d_\alpha) = \alpha$. A test based on this is called an exact Durbin Watson test.

Although the exact distribution of d depends on X, it is subject to an upper and lower bound. These bounds d_U and d_L have been tabulated. They depend on both T the sample size and K the number of regressors. To be valid, the regression **must** contain a constant term. We are testing against positive serial correlation. Hence we reject if d is too small.

If $d < d_L$ reject, if $d > d_U$ fail to reject. If $d_L < d < d_U$ inconclusive. In practice if d is small in the sense of being close to its critical points, it is sensible to re-estimate allowing for serial correlation.

Though the Durbin Watson statistic is often reported, the circumstances in which it is valid are very restricted. In particular it requires not only a constant term on the right hand side but it also requires that all right hand side variables are processed independent of the errors. If the x_t are only contemporaneously independent, the Durbin Watson test is invalid. Under these circumstances the test tends to fail to reject in the presence of serial correlation.