Hypothesis Test within the Maximum Likelihood Framework

Christian Julliard

Department of Economics and FMG
London School of Economics

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 test.

[1] Bayesian inference will not be presented.

Key assumptions

We have already seen that even if observations are dependent, the results derived for the MLE in the iid setting carry over for *ergodic* processes, and we'll be assuming that:

1. the MLE of a vector of parameters ψ is consistent
2. And that

$$\sqrt{T}(\hat{\psi} - \psi) \xrightarrow{D} N\left(0, \left(\frac{1}{T} I(\psi)\right)^{-1}\right)$$

where $I(\psi)$ is the information matrix.

Outline

The Wald tests

Idea: use the MLE of the *unrestricted* model.

- Suppose we have a model with k unknown parameters ψ that delivers the log likelihood $logL(\psi)$.
- We know that

$$\sqrt{T}(\hat{\psi}-\psi)\xrightarrow{d}N\left(0,IA(\psi)^{-1}\right)$$

where $IA(\psi)=\frac{1}{T}I(\psi)$, $I(\psi)=-E\left(\frac{\partial^2\log L(\psi)}{\partial\psi\partial\psi'}\right)$.

- Suppose we want to test a linear hypothesis $H_0:R\psi=q$ vs. $H_A:R\psi\neq q$, where R has $r<k$ linearly independent rows (r restrictions)

⇒ then under H_0

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

Wald Test LR test Lagrange Multiplier Comparison Durbin-Watson

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

- This implies that

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

But: we do not observe $IA(\psi)$. If we can find a consistent estimator, the distribution remains unchanged

- Possible estimators are: the empirical information matrix based, $\frac{1}{T}I(\hat{\psi})$, and the empirical hessian based, $\left[-\frac{1}{T}\times\frac{\partial^2\log L(\hat{\psi})}{\partial\psi\partial\psi'}\right]^{-1}$.
- Assuming the first is available, then

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

Wald Test LR test Lagrange Multiplier Comparison Durbin-Watson

Linear restrictions for the standard linear regression

- We showed last time that for the standard linear regression model

$$y_i=x_i'\beta+\varepsilon_i\sim N\left(0,\sigma^2\right);\ i=1,...,n;\ E\left[x_i\varepsilon_s\right]=0\ \forall s,i \qquad (1)$$

with unknown parameters β, σ^2, we have that

$$\sqrt{T}\left(\hat{\beta}'-\beta\right)\xrightarrow{d}N\left(0,\left(\frac{1}{T}\frac{\sum_{t=1}x_tx_t'}{\hat{\sigma}^2}\right)^{-1}\right)$$

- Suppose $\beta=[\beta_1,\beta_2]'$ and we want to test the null:

$$\beta_1-\beta_2=q$$

or equivalently in matrix form

$$\underbrace{[1,-1]}_{R}\underbrace{\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}}_{\beta'}=q$$

Wald Test LR test Lagrange Multiplier Comparison Durbin-Watson

Wald test for the standard linear regression

Noting that under the null

$$\sqrt{T}R\left(\hat{\beta}'-\beta\right)=\sqrt{T}\left(R\hat{\beta}'-q\right)\xrightarrow{D}N\left(0,R\left(\frac{1}{T}\frac{\sum_{t=1}x_tx_t'}{\hat{\sigma}^2}\right)^{-1}R'\right)$$

We can construct the Wald test as

$$W=\sqrt{T}\left(R\hat{\beta}'-q\right)\left[R\left(\frac{1}{T}\frac{\sum_{t=1}x_tx_t'}{\hat{\sigma}^2}\right)^{-1}R'\right]^{-1}\sqrt{T}\left(R\hat{\beta}'-q\right)$$

$$=\left(\hat{\beta}_1-\hat{\beta}_2-q\right)^2\left[R\left(\frac{\sum_{t=1}x_tx_t'}{\hat{\sigma}^2}\right)^{-1}R'\right]^{-1}\xrightarrow{d}\chi^2(1)$$

Wald test for nonlinear constraints

- Consider $H_0 : R(\psi) = 0$, a set of r linear or nonlinear constraints. (R is a column r-vector).
- 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 well defined $r \times k$ matrix (k is the number of parameters in ψ).
- Then, under H_0 the statistic

$$W = R(\hat{\psi})' \left[\left(\frac{\partial R(\hat{\psi})}{\partial \psi}\right) I(\hat{\psi})^{-1} \left(\frac{\partial R(\hat{\psi})}{\partial \psi}\right)' \right]^{-1} R(\hat{\psi}) \to \chi^2(r)$$

- Intuition: Delta Method/Taylor Expansion.

Example: nonlinear restrictions for the SLR

- Suppose $\beta = [\beta_1, \beta_2]'$ and we want to test the null:

$$\underbrace{\frac{1}{2}\beta_1^2 + \frac{1}{2}\beta_2^2 - q}_{R(\psi)} = 0$$

in the standard linear regression model (1)
- Noticing that

$$\frac{\partial R(\psi)}{\partial \psi} = \left[\frac{\partial R(\psi)}{\partial \beta_1}; \frac{\partial R(\psi)}{\partial \beta_2}\right] = [\beta_1; \beta_2]$$

- Than the Wald statistic becomes

$$\left(\frac{1}{2}\hat{\beta}_1^2 + \frac{1}{2}\hat{\beta}_2^2 - q\right)^2 \left\{ \left[\hat{\beta}_1; \hat{\beta}_2\right] \left(\frac{\sum_{t=1} x_t x_t'}{\hat{\sigma}^2}\right)^{-1} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \right\}^{-1} \xrightarrow{d} \chi^2(1)$$

The Likelihood Ratio Test

- Again suppose the model can be expressed in terms of a likelihood function $L(\psi)$.
- Suppose we also have a set of r restrictions, either linear or nonlinear i.e.

$$R\psi = q \text{ or } R(\psi) = 0.$$

Idea:
1. Estimate the unrestricted model to obtain ML estimates, $\hat{\psi}$ and $L(\hat{\psi})$.
2. Estimate the model under the restrictions to obtain restricted estimates $\hat{\psi}_0$ and $L(\hat{\psi}_0)$.
3. Then compare $L(\hat{\psi})$ and $L(\hat{\psi}_0)$

- It can be shown that under the null

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

- 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 $LR > \chi^2_\alpha(r)$.

Lagrange Multiplier Tests

- Again suppose the model can be expressed in terms of a likelihood function $L(\psi)$ and that we have r restrictions $R(\psi) = 0$.
- If the restrictions are valid $\hat{\psi}_0$ (the MLE of the restricted model) will be close to $\hat{\psi}$ (the MLE of the unrestricted model) and the partial derivatives in the vector $\frac{\partial \log L(\hat{\psi}_0)}{\partial \psi}$ will also be close to zero (note: $\frac{\partial \log L(\hat{\psi})}{\partial \psi} = 0$ by construction)

- 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, we normally 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

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

and is referred to as a Lagrange Multiplier statistic.

LM test for nonlinear least squares

- This result can be specialized 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

$$\begin{aligned} \log L(\beta, \sigma^2) &= -\frac{T}{2} \log 2\pi - \frac{T}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^{T} \varepsilon_t(\beta)^2 \\ \varepsilon_t(\beta) &= y_t - g(x_t; \beta). \end{aligned}$$

- Assume that the r restrictions involve only β (not σ^2): $R(\beta) = 0$.
- Then

$$\begin{aligned} \frac{\partial \log L(\beta, \sigma^2)}{\partial \beta} &= \frac{1}{\sigma^2} \sum_t z_t \varepsilon_t, \\ z_t &= -\frac{\partial \varepsilon_t}{\partial \beta}. \end{aligned} \qquad (3)$$

and 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,

$$I_{\beta\beta}(\psi) = -E\left[\frac{\partial^2 \log L}{\partial \beta \partial \beta'}\right] = \frac{1}{\sigma^2} E \sum_t z_t z_t'. \qquad (4)$$

- Evaluating the LM-statistics at $(\hat{\beta}_0, \hat{\sigma}_0^2)$, where $\hat{\sigma}_0^2 = \frac{1}{T}\sum_t \varepsilon_t^2(\hat{\beta}_0)$, and replacing the expectations with their sample analog

$$LM = \frac{1}{\hat{\sigma}_0^2}\left(\sum_t z_t\varepsilon_t\right)'\left[\sum_t z_t z_t'\right]^{-1}\left(\sum_t z_t\varepsilon_t\right).$$

- 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

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

- Now consider the R^2 from this regression

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

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

Intuition: if $\hat{\beta}_0$ is close to $\hat{\beta}$, the $\varepsilon_t(\hat{\beta}_0)$ shouldn't be forecastable.

- All three tests are asymptotically equivalent.

Warning: these are asymptotic distribution results, so caution should be used in small sample.

- In small sample (but there are exceptions):
  1. In general the LR test is the best, in the sense that its finite sample behavior most closely approximates its expected large sample properties.
  2. The Wald test is second best and the LM procedure worst.

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

$$H_0 : \phi = 0 \text{ against } H_A : \phi > 0.$$

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

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

Note: $d \approx 2(1 - r_1)$, where r_1 is the simple correlation between $\hat{u}_t$ and $\hat{u}_{t-1}$.

$\Rightarrow$ d lies in the interval [0,4].

- Unfortunately the exact distribution of d depends on X

But d is subject to an upper (d_U) and lower bound (d_L) that depend on both the sample size and the number of regressors.

- We are testing against *positive* serial correlation so 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.

Note: to be valid, *i*) the regression must contain a constant, *ii*) all RHS variables are processed independent of the errors