

Estimating Linear Time Series Models 1

The Autoregression:

$$x_t = \alpha + \lambda_1 x_{t-1} + \cdots + \lambda_p x_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim iid(0, \sigma^2)$$

Given a sample x_t , $t = 1, \dots, T$, consider the estimation of $\alpha, \lambda_1, \dots, \lambda_p, \sigma^2$. The natural approach is ordinary least squares.

For simplicity, consider the AR(1), $x_t = \alpha + \lambda_1 x_{t-1} + \varepsilon_t$. The least squares estimators are

$$\hat{\lambda}_1 = \frac{\sum_{t=2}^T (x_t - \bar{x})(x_{t-1} - \bar{x}_{-1})}{\sum_{t=2}^T (x_{t-1} - \bar{x}_{-1})^2} \quad \hat{\alpha} = \bar{x} - \hat{\lambda}_1 \bar{x}_{-1}.$$

NOTE: regression in time series is different! The usual classical regression model (CRM) assumptions don't hold.

- $\hat{\lambda}_1$ is *biased*: $E(\hat{\lambda}_1) - \lambda_1 = O(T^{-1})$ – but not zero!
- The usual confidence intervals and t tests are approximately valid, but the justification is quite different from the CRM case.

We can however show that $\hat{\lambda}_1$ is a consistent, asymptotically normal (CAN) estimator of λ_1 .

Proposition 1: If $|\lambda_1| < 1$ then

i) (consistency)

$$\hat{\lambda}_1 \xrightarrow{\text{pr}} \lambda_1$$

ii) (asymptotic normality)

$$\sqrt{T}(\hat{\lambda}_1 - \lambda_1) \xrightarrow{d} N(0, 1 - \lambda_1^2)$$

.

Consistency

Consider the error-of-estimate:

$$\hat{\lambda}_1 = \lambda_1 + \frac{\sum_{t=2}^T (x_{t-1} - \bar{x}_{-1}) \varepsilon_t}{\sum_{t=2}^T (x_{t-1} - \bar{x}_{-1})^2} = \lambda_1 + \frac{T^{-1} \sum_{t=2}^T x_{t-1} \varepsilon_t - \bar{x}_{-1} \bar{\varepsilon}}{T^{-1} \sum_{t=2}^T x_{t-1}^2 - \bar{x}_{-1}^2}$$

By Slutsky's Theorem,

$$\text{plim } \hat{\lambda}_1 = \lambda_1 + \frac{\text{plim } T^{-1} \sum_{t=2}^T x_{t-1} \varepsilon_t - \text{plim } \bar{x}_{-1} \text{plim } \bar{\varepsilon}}{\text{plim } T^{-1} \sum_{t=2}^T x_{t-1}^2 - (\text{plim } \bar{x}_{-1})^2}.$$

Consider the following facts:

- Finite lag transformations of stationary ergodic series are stationary and ergodic.
- Hence, the following sequences are all stationary and ergodic, on the assumptions:
 $\{x_t\}$, $\{\varepsilon_t\}$, $\{x_t^2\}$, $\{x_{t-1}\varepsilon_t\}$.
- By the LIE,

$$E(x_{t-1}\varepsilon_t) = E(E(x_{t-1}\varepsilon_t | \mathcal{F}_{t-1})) = E(x_{t-1}E(\varepsilon_t | \mathcal{F}_{t-1})) = E(0) = 0$$

It follows from the above facts, that

$$1. \quad \bar{x}_{-1} \xrightarrow{\text{pr}} \mu_x = \frac{\alpha}{1 - \lambda_1}$$

$$2. \quad \bar{\varepsilon} \xrightarrow{\text{pr}} 0$$

$$3. \quad T^{-1} \sum_{t=2}^T x_{t-1}^2 \xrightarrow{\text{pr}} \sigma_x^2 + \mu_x^2 = \frac{\sigma^2}{1 - \lambda_1^2} + \frac{\alpha^2}{(1 - \lambda_1)^2}$$

$$4. \quad T^{-1} \sum_{t=2}^T x_{t-1} \varepsilon_t \xrightarrow{\text{pr}} 0$$

Hence, by Slutsky's theorem,

$$\text{plim } \hat{\lambda}_1 = \lambda_1 + \frac{0}{\sigma_x^2} = \lambda_1.$$

Please note: we have to show *two* conditions, both necessary for consistency:

- Numerator of $\text{plim } \hat{\lambda}_1 - \lambda_1$ is equal to zero,
- Denominator of $\text{plim } \hat{\lambda}_1 - \lambda_1$ is non-zero.

Asymptotic Normality

$$\sqrt{T}(\hat{\lambda}_1 - \lambda_1) = \frac{T^{-1/2} \sum_{t=2}^T x_{t-1} \varepsilon_t - \bar{x}_{-1} T^{-1/2} \sum_{t=2}^T \varepsilon_t}{T^{-1} \sum_{t=2}^T x_{t-1}^2 - \bar{x}_{-1}^2}$$

First note that ε_t and x_{t-1} are independent, and

- $E|x_{t-1}\varepsilon_t| = E|x_{t-1}|E|\varepsilon_t| < \infty$
- $E(x_{t-1}\varepsilon_t|\mathcal{F}_{t-1}) = x_{t-1}E(\varepsilon_t|\mathcal{F}_{t-1}) = 0$ a.s.

Therefore, $(x_{t-1}\varepsilon_t, \varepsilon_t)'$, is a vector martingale difference.

It can be shown

$$\frac{1}{\sqrt{T}} \begin{bmatrix} \sum_{t=2}^T x_{t-1} \varepsilon_t \\ \sum_{t=2}^T \varepsilon_t \end{bmatrix} \xrightarrow{d} N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} (\sigma_x^2 + \mu_x^2)\sigma^2 & \mu_x\sigma^2 \\ \mu_x\sigma^2 & \sigma^2 \end{bmatrix} \right)$$

The second step is to apply Cramér's Theorem for vectors:

If $Z_T \xrightarrow{d} N(0, \Omega)$ and $A_T \xrightarrow{pr} A$ then $A_T Z_T \xrightarrow{d} N(0, A\Omega A')$.

Consider the case

$$A_T = \frac{1}{T^{-1} \sum_{t=2}^T x_{t-1}^2 - \bar{x}_{-1}^2} \begin{bmatrix} 1 & -\bar{x}_{-1} \end{bmatrix} \xrightarrow{\text{pr}} \frac{1}{\sigma_x^2} \begin{bmatrix} 1 & -\mu_x \end{bmatrix}.$$

Note that

$$\begin{aligned} A\Omega A' &= \frac{1}{\sigma_x^4} \begin{bmatrix} 1 & -\mu_x \end{bmatrix} \begin{bmatrix} (\sigma_x^2 + \mu_x^2)\sigma^2 & \mu_x\sigma^2 \\ \mu_x\sigma^2 & \sigma^2 \end{bmatrix} \begin{bmatrix} 1 \\ -\mu_x \end{bmatrix} \\ &= \frac{\sigma^2\sigma_x^2}{\sigma_x^4} = \frac{\sigma^2}{\sigma_x^2} = 1 - \lambda_1^2. \end{aligned}$$

We have therefore shown, as required, that

$$\sqrt{T}(\hat{\lambda}_1 - \lambda_1) \xrightarrow{d} N(0, 1 - \lambda_1^2)$$

Remarks:

- These results can be used to generate approximate confidence intervals, perform significance tests, etc.
- Our assumption of i.i.d. shocks can be weakened to a m.d. assumption - slightly complicates the argument (see ET Ch 6).
- The usual OLS variance formula is $\hat{\sigma}^2/T\hat{\sigma}_x^2$ - a consistent estimate of $(1 - \lambda_1^2)/T$, the approx. variance of $\hat{\lambda}_1$.
- Of course, corresponding results exist for $\hat{\alpha}$.
- The extension from AR(1) to AR(p) is exactly the same in principle. The complications are merely in the matrix algebra. The estimator has the form

$$\hat{\lambda} = (X'_-X_-)^{-1}X'_-x$$

where $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_p)'$,

$$x = (x_{p+1}, \dots, x_T)', X_- = [x_{-1}, x_{-2}, \dots, x_{-p}]$$

Estimating Linear Time Series Models 2

The Moving Average Model

Consider the MA(1)

$$x_t = \alpha + \varepsilon_t + \theta_1 \varepsilon_{t-1}, \quad \varepsilon_t \sim iid(0, \sigma^2).$$

The least squares method can also be applied to this model. However, the model is *nonlinear*. Therefore there is a *numerical* problem to solve.

Write

$$\varepsilon_t = x_t - \alpha - \theta_1 \varepsilon_{t-1} \tag{*}$$

The residuals can be calculated *recursively* from formula (*), for $t = 1, 2, \dots, T$.

Thus,

$$\begin{aligned} \varepsilon_t &= x_t - \alpha - \theta_1(x_{t-1} - \alpha - \theta_1(x_{t-2} - \alpha - \theta_1(\dots))) \\ &= \sum_{j=0}^{t-1} (-\theta_1)^j (x_{t-j} - \alpha) + (-\theta_1)^t \varepsilon_0. \end{aligned}$$

As an approximation, set $\varepsilon_0 = 0$.

The sum of squares

$$S(\theta_1, \alpha) = \sum_{t=1}^T \varepsilon_t^2$$

must be minimised using a numerical algorithm.

Note that

$$\begin{aligned}\frac{\partial S}{\partial \theta_1} &= 2 \sum_{t=1}^T \varepsilon_t \frac{\partial \varepsilon_t}{\partial \theta_1} \\ \frac{\partial S}{\partial \alpha} &= 2 \sum_{t=1}^T \varepsilon_t \frac{\partial \varepsilon_t}{\partial \alpha}\end{aligned}$$

where the derivatives can be computed recursively, starting from approximate initial conditions

$$\frac{\partial \varepsilon_0}{\partial \theta_1} = \frac{\partial \varepsilon_0}{\partial \alpha} = 0.$$

Thus,

$$\begin{aligned}\frac{\partial \varepsilon_t}{\partial \theta_1} &= -\varepsilon_{t-1} - \theta_1 \frac{\partial \varepsilon_{t-1}}{\partial \theta_1} \\ &= -\varepsilon_{t-1} + \theta_1 \varepsilon_{t-2} + \theta_1^2 \frac{\partial \varepsilon_{t-2}}{\partial \theta_1} \\ &= \dots \approx -\sum_{j=1}^{t-1} (-\theta_1)^j \varepsilon_{t-j}\end{aligned}$$

$$\begin{aligned}\frac{\partial \varepsilon_t}{\partial \alpha} &= -1 - \theta_1 \frac{\partial \varepsilon_{t-1}}{\partial \alpha} \\ &= -1 + \theta_1 - \theta_1^2 \frac{\partial \varepsilon_{t-1}}{\partial \alpha} = \dots \approx -\frac{1}{1 + \theta_1}.\end{aligned}$$

Gauss-Newton Algorithm

To compute LS estimates, we must solve the pair of nonlinear equations,

$$\frac{\partial S}{\partial \theta_1} = 0, \quad \frac{\partial S}{\partial \alpha} = 0.$$

Note: the derivatives, $\partial \varepsilon_t / \partial \theta_1$ and $\partial \varepsilon_t / \partial \alpha$, must be orthogonal to (uncorrelated with) ε_t at the solution point. This suggests a regression-based search algorithm.

Let $d_t = (\partial \varepsilon_t / \partial \theta_1, \partial \varepsilon_t / \partial \alpha)'$, and let d_t^r and ε_t^r denote the evaluation of d_t and ε_t at trial points (θ_1^r, α^r) for $r = 0, 1, 2, \dots$

The updating rule is

$$\begin{bmatrix} \theta_1^{r+1} \\ \alpha^{r+1} \end{bmatrix} = \begin{bmatrix} \theta_1^r \\ \alpha^r \end{bmatrix} + \mu^r \left(\sum_t d_t^r d_t^{r'} \right)^{-1} \sum_t d_t^r \varepsilon_t^r.$$

- The (optional) factor μ^r is a scalar, chosen to minimize S in a line search. Setting $\mu^r = 1$ will also work, but may require more steps to converge.
- $\sum_t \hat{d}_t \hat{\varepsilon}_t = 0$ by definition. Hence the updating vector approaches 0 as $(\theta_1^r, \alpha^r)'$ approaches the least squares estimate $(\hat{\theta}_1, \hat{\alpha})'$. This guarantees the algorithm ‘stops’ at the right place.

Whether the algorithm finds the true (global) minimum depends on the choice of starting values, $(\theta_1^0, \alpha^0)'$. Ideally, try two or more starting points and compare.

Asymptotic Properties

- It can be proved that the nonlinear least squares (NLS) estimator is CAN.
- Because no closed formula exists for the estimator, the proof of these results requires advanced methods.
- However, the fact that $\{d_t \varepsilon_t\}$ (evaluated at the true parameters) is a vector martingale difference process is a key ingredient of the proof.
- Also note that the asymptotic covariance matrix of $(\hat{\theta}_1, \hat{\alpha})'$ is consistently estimated by

$$\hat{V} = \hat{\sigma}^2 \left(\sum_t \hat{d} \hat{d}' \right)^{-1}.$$

Generalizations

- Extension to MA(q) is just a matter of elaborating the same method.
- To extend to ARMA(p, q), combine the two least squares procedures.

Consider the ARMA(1,1) to illustrate. To perform the NLS iterations, the following recursive formulae are needed.

$$\begin{aligned}\varepsilon_t &= x_t - \alpha - \lambda_1 x_{t-1} - \theta_1 \varepsilon_{t-1} \\ \frac{\partial \varepsilon_t}{\partial \alpha} &= -1 - \theta_1 \frac{\partial \varepsilon_{t-1}}{\partial \alpha} = \frac{-1}{1 + \theta_1} \\ \frac{\partial \varepsilon_t}{\partial \lambda_1} &= -x_{t-1} - \theta_1 \frac{\partial \varepsilon_{t-1}}{\partial \lambda_1} \\ \frac{\partial \varepsilon_t}{\partial \theta_1} &= -\varepsilon_{t-1} - \theta_1 \frac{\partial \varepsilon_{t-1}}{\partial \theta_1}\end{aligned}$$

- Could start the extra recursion by setting

$$\frac{\partial \varepsilon_0}{\partial \lambda_1} = \frac{-\bar{x}}{1 + \theta_1}$$

say.

Method of Maximum Likelihood

Assume ARMA(1,1). Suppose it is known that

$$\varepsilon_t \sim \text{NI}(0, \sigma^2).$$

The probability density function (p.d.f.) is

$$\phi(\varepsilon_t) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left\{-\frac{\varepsilon_t^2}{2\sigma^2}\right\}.$$

The *conditional* p.d.f. of the data is therefore

$$\phi(x_t|\mathcal{F}_{t-1}) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left\{-\frac{(x_t - \alpha - \lambda_1 x_{t-1} - \theta_1 \varepsilon_{t-1})^2}{2\sigma^2}\right\}$$

noting that:

$$E(x_t|\mathcal{F}_{t-1}) = \alpha + \lambda_1 x_{t-1} + \theta_1 \varepsilon_{t-1}$$

$$\text{Var}(x_t|\mathcal{F}_{t-1}) = \sigma^2.$$

Now, extend the well-known conditional probability relation for events A and B ,

$$P(A|B)P(B) = P(A \cap B)$$

the joint p.d.f. of the sample is

$$\begin{aligned}\phi(x_1, x_2, \dots, x_T) &= \phi(x_1)\phi(x_2|x_1)\dots\phi(x_T|x_1, \dots, x_{T-1}) \\ &= \phi(x_1)(2\pi\sigma^2)^{-(T-1)/2} e^{-\frac{1}{2\sigma^2} \sum_{t=2}^T \varepsilon_t^2}\end{aligned}$$

where $\varepsilon_t = x_t - \alpha - \lambda_1 x_{t-1} - \theta_1 \varepsilon_{t-1}$, and $f(x_1)$ is to be determined.

- This formula defines a family of p.d.f.s, depending on parameters $\alpha, \lambda_1, \theta_1, \sigma^2$.
- Define the likelihood function as

$$\mathcal{L}(\alpha, \lambda_1, \theta_1, \sigma^2; x_1, \dots, x_T) = \phi(x_1, \dots, x_T; \alpha, \lambda_1, \theta_1, \sigma^2)$$

In other words, a function of parameters, given observed sample.

Maximum Likelihood - Procedure

1. Choose the parameter values to maximize \mathcal{L} , computed from the observed x_1, \dots, x_T . These are the values "most likely" to have generated the sample, in the sense that the p.d.f. is largest at that point.
2. Maximizing $L = \log \mathcal{L}$ has the same solution, since the logarithm is a monotonic transformation. This is obviously easier, converting products into sums:

$$L = \log \phi(x_1) - \frac{T}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=2}^T \varepsilon_t^2.$$

3. $\log \phi(x_1) = O(1)$ as $T \rightarrow \infty$, while the other terms in L are $O(T)$. Therefore, ignoring the "start-up" term gives an equivalent estimator in large samples.
4. Ignoring $\log \phi(x_1)$, note that parameters $\alpha, \lambda_1, \theta_1$ occur only in last term. This will have the same maximum for any value of σ^2 . Therefore, ML is equivalent in large samples to NLS
5. Finally, maximize L w.r.t. σ^2 , ignoring $\log \phi(x_1)$ and holding other parameters at solution values: solving the first-order conditions

$$\frac{\partial L}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{t=2}^T \hat{\varepsilon}_t^2 = 0$$

yields solution $\sigma^2 = T^{-1} \sum_{t=2}^T \hat{\varepsilon}_t^2$.

Points:

- Generalizing to ARMA(p, q) is done in the obvious way. Note that in this case the start-up term is the joint density of x_1, \dots, x_p .
- ‘Exact’ ML can also be performed, by determining the formula for $\phi(x_1)$, as a function of parameters, and maximizing the full function.

Inference in ML

Let φ denote the vector of parameters of the likelihood function. It is known that, under standard regularity conditions,

$$\sqrt{T}(\hat{\varphi} - \varphi) \rightarrow_d N(0, \mathfrak{I}^{-1}) \quad (*)$$

where $\mathfrak{I} = - \lim_{T \rightarrow \infty} \frac{1}{T} E \frac{\partial^2 L}{\partial \varphi \partial \varphi'}$. This is the *information matrix*.

This formula provides approximate standard errors of the estimates, replacing \mathfrak{I} by

$$\hat{\mathfrak{I}} = -\frac{1}{T} \frac{\partial^2 L(\hat{\varphi})}{\partial \varphi \partial \varphi'}$$

- In practice, $\hat{\mathfrak{I}}$ is the average of second derivatives of the terms of L , Should be close to its expected value. Easily calculated, given a formula for L .
- The *Cramér-Rao Theorem* states that $\hat{\mathfrak{I}}^{-1}$ is a lower bound on the asymptotic variance matrix of CAN estimators.
i.e., ML is asymptotically efficient in the class of CAN estimators.

ML and QML

Gaussianity is a strong assumption! If it is incorrect, the method is called *quasi*-ML.

- QML is in many cases a CAN estimator (e.g. equivalent to NLS).
- However, the result (*) must be replaced by

$$\sqrt{T}(\hat{\varphi} - \varphi) \rightarrow_d N(0, \mathfrak{I}^{-1}A\mathfrak{I}^{-1}) \quad (**)$$

where

$$A = \lim_{T \rightarrow \infty} \frac{1}{T} E \left(\begin{array}{cc} \frac{\partial L}{\partial \varphi} & \frac{\partial L}{\partial \varphi'} \end{array} \right).$$

- The property $A = \mathfrak{I}$ holds in general only for true ML estimators.
 - The assumption of normality must be valid. Not true for "quasi-ML".
- Cramér-Rao Theorem depends on $A = \mathfrak{I}$, does not hold for QML.
- There is no reason to think 'exact' QML better than NLS (justification is large-sample).
- There are cases (we will not encounter them) where QML is inconsistent. Caution needed!
- Even if the normality assumption is correct, simulation studies suggest exact ML may perform no better than NLS in small samples.
- All these estimators typically biased in small samples.

Tests

Consider a restriction (null hypothesis) $g(\varphi) = 0$.

Let this be a hypothesis (or set of hypotheses) about the model.

Examples: $g(\varphi) = \varphi_1$; $g(\varphi) = \varphi_2 - \varphi_3$; $g(\varphi) = \varphi_4\varphi_5 - 1$.

1. Wald tests.

Determine the distribution of the random variable $g(\hat{\varphi})$ when H_0 true, and hence derive a test statistic. This statistic is based on variance estimates $\hat{\mathfrak{S}}^{-1}$ or $\hat{V} = \hat{\mathfrak{S}}^{-1} \hat{A} \hat{\mathfrak{S}}^{-1}$. The usual t test is an example.

2. Lagrange Multiplier tests.

Estimate the model *subject* to the restrictions. Base the test on distribution of Lagrange multipliers for the *constrained* maximisation of L (or minimisation of S), w.r.t. φ , subject to $g(\varphi) = 0$.

- These are also called *score tests*. According to the first-order conditions, the Lagrange multipliers are a multiple of the scores (first derivatives) of the log-likelihood.

3. Likelihood Ratio tests

Estimate model *both* subject to restriction *and* unrestricted. Under H_0 ,

$$2(L(\hat{\varphi}) - L(\hat{\varphi})) \sim \chi^2(p)$$

in large samples where $\hat{\varphi}$ is the constrained ML estimate.

Points:

- All of these methods give rise to test statistics which are asymptotically $\chi^2(p)$ when H_0 is true, where p = number of restrictions under test.

- If:

1. the model is linear-in-parameters, (e.g. AR model)

2. the restriction is linear (i.e., $g(\varphi) = R\varphi + c$ for constant R and c).

then the Wald, LM and LR principles all give rise to the same formula, and are equivalent to the F test.

In other cases, they can give different results.

- The Wald test can depend critically on how the restriction is stated. Thus

$$\varphi_4\varphi_5 - 1 = 0$$

and

$$\varphi_4 - 1/\varphi_5 = 0$$

represent the same restriction, but in finite samples the W statistics for the two tests could be *very* different.

- The LR test is *only* valid when the information matrix equality holds! Use with caution in the QML context.