Estimating AR/MA models

Susan Thomas

September 17, 2009

Goals

- The likelihood estimation of AR/MA models
  - AR(1)
  - MA(1)
- Inference
- Model specification for a given dataset

Why MLE?

- Traditional linear statistics is one methodology of estimating models standing on a set of assumptions that are rigidly defined.
  This yields a relative fixed set of models which can be estimated.
- One such assumption is the independence of the error term.
- Maximum Likelihood Estimation (MLE) appears a more complicated way of coming to the same answer, when looking for simple moment estimators (e.g. sample mean) or classical least squares.
- However, MLE permits us to go beyond simple problems. It offers a more generic way to deal with models of stochastic time series processes.

The likelihood approach

- For any model: $y = f(x; \theta)$, MLE involves:
  - setting up the joint likelihood of observing the data
  - finding the θ that maximises the likelihood of the data
- In non time-series problems, assume independence of $y_1, y_2, \ldots, y_N$

$$L = f(y_1, y_2, \ldots, y_N|\theta) = f(y_1|\theta).f(y_2|\theta).\ldots.f(y_N|\theta)$$

- In time series-problems, there is dependence in $x_1, x_2, \ldots, x_T$

$$\begin{aligned} L &= f(y_1, y_2, \ldots, y_N|\phi) \\ &= f(y_1|\phi).f(y_2|y_1, \phi).f(y_3|y_2, y_1, \phi).\ldots.f(y_N|y_{N-1}, \ldots, y_1, \phi) \end{aligned}$$

Here we need to use the *joint probability* of *conditional probabilities*.

MLE setup for AR(1) estimation

The model

- The AR(1) process is

$$Y_t = c + \phi Y_{t-1} + \epsilon_t$$

  where $\epsilon_t \sim \text{i.i.d.} N(0, \sigma^2)$
- We know
  - $E(Y_t) = \mu = c/(1 - \phi)$ and
  - $E(Y_t - \mu)^2 = \sigma^2/(1 - \phi^2)$
- Now we need to setup the Likelihood of the data set:

$$Y_1, Y_2, \ldots, Y_T$$

Likelihood for AR(1)

- Probability of the 1st observation is:

$$
\begin{aligned}
f(y_1;\theta) &= f(y_1; c, \phi, \sigma^2) \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma^2/(1-\phi^2)}} \exp\left(\frac{-\left\{y_1 - (c/(1-\phi))\right\}^2}{2\sigma^2/(1-\phi^2)} \right)
\end{aligned}
$$

The second observation

$$Y_2 = c + \phi Y_1 + \epsilon_2$$

- Conditioning on Y_1, i.e. treating Y_1 as a constant y_1,

$$Y_2 | (Y_1 = y_1) \sim N(c + \phi y_1, \sigma^2)$$

- Conditional mean of $Y_2 = c + \phi y_1$
- Conditional variance of $Y_2 = E(Y_2 - E(Y_2))^2 = E(\epsilon_2)^2 = \sigma^2$.
- Conditional density of Y_2 is:

$$f(Y_2 | Y_1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(Y_2 - c - \phi y_1)^2}{2\sigma^2}\right] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-\epsilon_2^2}{2\sigma^2}\right]$$

Putting it together

- The joint of 1 and 2 is the product of these two elements:

$$f_{Y_1,Y_2}(y_1, y_2; \theta) = f_{Y_1}(y_1; \theta) f_{Y_2|Y_1}(y_2|y_1; \theta)$$

- The conditional for observation 3 is

$$f(Y_3|Y_2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_3 - c - \phi y_2)^2}{2\sigma^2}\right]$$

- In this fashion we can setup all the conditionals, and multiply them together to get the joint.

Final Log likelihood

- The objective function would be to maximise L or minimise $\log L$:

$$logL = -\frac{T-1}{2} 2\pi\sigma^2 - \sum_{t=2}^{T} \frac{\epsilon_t^2}{\sigma^2} - \frac{\pi\sigma^2}{(1-\phi^2)} - \frac{(y_1 - \frac{c}{1-\phi})^2}{\frac{2\sigma^2}{(1-\phi^2)}}$$

Exact vs. Conditional likelihood

- The above strategy yields the “exact MLE”:
  This is because L includes the probability of the first observation, y_1.
- Suppose we just ignore observation 1.
- Then all other observations have an identical and familiar form – it’s just an sum of squared errors, SSE.
  This becomes equivalent to running OLS on the dataset, with Y_t as the LHS and the lagged values Y_{t-1} as the RHS in the equation.
- When the probability of the first observation in an AR(1) model is not included, the MLE is called the “conditional MLE”.

Conditional likelihood for AR(1)

- It is the same as earlier, except for the $f(Y_1|\theta)$ term.

$$logL = -(T-1)\pi\sigma^2 - \sum_{t=2}^{T} \frac{\epsilon_t^2}{\sigma^2}$$

- When T is very large, the exact and the conditional MLE estimates have the same distribution.
  This is true when the series is stationary.
- When the series is non-stationary, $|\phi| > 1$, the conditional MLE gives consistent estimates.
  But the exact MLE does not.
- Thus, for most AR estimations, OLS is used to estimate the parameters of the model.

MLE setup for MA(1) estimation

The MA(1) model

- The model –

$$
\begin{aligned}
Y_t &= \mu + \epsilon_t + \theta \epsilon_{t-1} \\
\epsilon_t &\sim iidN(0, \sigma^2)
\end{aligned}
$$

- In this case, the exact likelihood is harder. So we estimate using a conditional MLE.

Conditional MLE for MA(1)

- Suppose we knew that $\epsilon_0 = 0$ exactly. Then

$$(Y_1 | \epsilon_0 = 0) \sim N(\mu, \sigma^2)$$

- Once Y_1 is observed, we know

$$\epsilon_1 = Y_1 - \mu_1$$

  exactly.

- Then:

$$f_{Y_2|Y_1,\epsilon_0=0}(y_2|y_1, \epsilon_0 = 0; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_2 - \mu - \theta\epsilon_1)^2}{2\sigma^2}\right]$$

Conditional likelihood of MA(1)

- In this fashion, we can go forward, iterating on $\epsilon_t = y_t - \mu - \theta\epsilon_{t-1}$.
- This gives us

$$
\begin{aligned}
\mathcal{L}(\theta) &= \log f_{Y_T, Y_{T-1}, \ldots, Y_1|\epsilon_0=0}(y_T, y_{T-1}, \ldots, y_1|\epsilon_0 = 0; \theta) \\
&= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T}\frac{\epsilon_t^2}{2\sigma^2}
\end{aligned}
$$

- L is different here from the AR(1) process: we need to calculate the L by an iterative process.
  Here, OLS cannot be applied to estimate an MA(1) model.

Summarising MLE for ARMA models

- There are two likelihood functions that can be used for the maximisation of the MLE:
  - Exact MLE: where the probabilities of the first p observations of an AR(p) model or the first q observations of an MA(q) model are explicitly included.
  - Conditional MLE: These are assumed to be known with certainty and are included as inputs in the estimation.
- An AR process can be estimated using OLS under the conditional MLE setup.
- All MA processes have to be estimated using MLE.

Inference

Inference for MLE parameters

- Inference of the estimated model parameters is based on the observed Fischer information.

$$\mathrm{var}(\hat{\theta}_{mle}) = \frac{1}{T} I^{-1}$$

- I is the informaiton matrix and can be estimated either as:
  1. The second derivative estimate:

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

  2. The first derivative estimate:

  $$\hat{I} = -T^{-1} \sum_{t=1}^{T} [\frac{\partial \log L(\theta)}{\partial\theta'} \frac{\partial \log L(\theta)}{\partial\theta'}']$$

  Both estimated at $\theta = \hat{\theta}$
- If T is large enough, then a standard t–test can be performed using $\hat{\theta}_{mle}$ and $\mathrm{var}(\hat{\theta}_{mle})$.

Time series model specification

Specification = formulation + selection

- Formulation is done based on a mixture of prior, theoretical knowledge about the problem and diagnostic, exploratory tests of the data.
- Selection is based on estimation and hypothesis tests.
- The **Box–Jenkins** methodology of forecasting: seperate the identification of the model from the estimation of the model.

Formulating a model

- Examples of prior knowledge driving model formulation:
  - Monthly series of agricultural produce will have a seasonal behaviour for the kharif and rabi crop.
  - Daily data on the call money rates will have a fortnightly pattern because of banks having report their capital requirements to the central bank every fortnight.
  - The time series of prices of a single futures contract will have a steadily decreasing trend as the contract comes close to expiration.
- Diagnostic tests: the Box–Jenkins methodology of *a priori* identification.

The Box–Jenkins identification methodology

- Graphs of the raw data: To pick out the possible existence of a trend, seasonality, etc.
  These are only indicative. The question of how the seasonality or trend affects the time series dyanamics – whether as an additive component of $f(t)$, or as part of the polynomial structure of $g(L)$ – depends upon more rigorous tests.
- ACFs, PACFs: More subjective measures of whether there is a stochastic trend, or a seasonal pattern.
  A plot of the autocorrelation function is also useful to detect the manner of time dependance – whether it is an AR, or MA, or a mixed ARMA process, and how many lags are likely to be required to describe the DGP.

- The statistical significance of each correlation coefficient is tested as a t-test, where the σ of the coefficient is given by Bartlett(1928). Typically, we test against the null of white noise, $\phi = \theta = 0$. Here, the Bartlett's formula approximates to

$$\mathrm{var}(\hat{\rho}_k) = 1/T$$

- Another test is the Portmanteau test of significance of the sum of a set of k autocorrelation coefficients, Q_k.

$$\begin{aligned} Q_k &= T \sum_{i=1}^{k} \hat{\rho}_k^2 \\ &\sim \chi^2(k-p-q) \end{aligned}$$

Problems of underestimating the significance of $\hat{\rho}_k$

- We neither know the true model nor the true model parameters. In this case, our bound is typically an over–estimate of the true σ.
- For example, an AR(1) model will have

$$\text{var}(\hat{\rho}_1) = \phi^2/T$$

If the model is stationary, then $-1 < \phi < 1$, and $\phi^2/T << 1/T$.

- Therefore, we end up underestimating the presence of temporal dependance when using $\text{var}(\hat{\rho}_k) = 1/T$.

Model estimation

- Once the form and the order of the temporal dependance has been approximately identified, the outcome is a set of possible ARMA models that should be estimated.
- Estimation is done using MLE/OLS depending upon whether it has an MA term or not.

Tests for model selection

We use one of the standard MLE tests to do a first brush selection of a model.

- The standard tests are:
  1. Likelihood Ratio (LR)
  2. Wald
  3. Lagrange Multiplier
- Tests that incorporate a penality for over-parameterisation are:
  1. Akaike Information Criteria (AIC): $(2 * \log L)/T + (k * 2)/T$
  2. Schwarz–Bayes Criteria (SBC): $(2 * \log(L))/T + (k * \log T)/T$

  where k is the number of parameters in the model, and T is the number of obsrevations.

Tests for model selection

These tests are superior to simple hypothesis testing for parameters because:

- They give numerical values.
- In some cases, they can be used to compare non–nested models.
- With these tests, one model is being tested against the other, whereas hypothesis testing requires a null of a “true” model.

Box–Jenkin's *a posteriori* identification

The last stage of the modelling process is checking whether the model chosen using the processes listed above is a suitable approximation to the "true" DGP or not.

1. A model must be consistent with the prior/theoretical knowledge and properties of the data.
2. Apply in–sample checks, *residual analysis*:
   Use the model to calculate the residuals, and analyse the properties of the residuals for consistency with prior assumptions/knowledge.
3. Apply out–of–sample checks, *forecast bias*:
   The dataset used for estimation must be a subset of the total dataset.
   Once the model is estimated, it can be used for forecasting future values – the data not used in estimation should be used to check the quality of the forecasts from the model.