Statistics 5350/7110
Forecasting

Lecture 13
Estimating ARMA Models

Professor Robert Stine

Preliminaries

- Questions?

- Assignments
  - Assignment #3 due Tuesday
  - We will review all techniques today (R stuff)

- Quick review
  - Partial correlation and the PACF
  - Use of ACF and PACF for identifying ARMA models: Cut off versus geometric decay
  - Role of model selection criteria (AIC, BIC)

Lecture_12.Rmd

Today's Topics

Text, §4.3

- Context
  - Presume we know the identifying order of the ARMA model (p, q), normally distributed
  - Start with initial estimates, typically moment estimators
  - Refine those initial estimates using maximum likelihood (ML)

- Methods of estimation
  - Moment estimation as an initial, easy-to-do starting value
  - Maximum likelihood to gain more efficiency (iterative, nonlinear procedure)

The next lecture continues these topics with more examples

- Estimating AR processes
  - Resembles a least squares regression
  - Example of maximum likelihood estimates for an AR(1) process (page 89)
- Estimating MA processes
  - Fitting a regression but you don't see values of the predictors
  - Textbook example 4.26, p85 has details for MA(1)

Moment Estimation

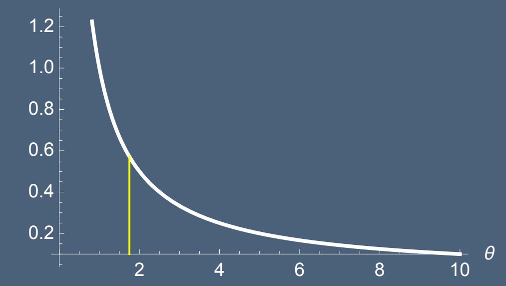
- Method of moments
  - Parameter estimate comes from solving equation(s) based on expected values
- Example
  - Data $Y_1, \ldots, Y_n$ is sample from Uniform[0, θ]
  - Want to find an estimator for unknown parameter θ
- Moment estimator
  - Expected value is $E(Y_i) = \theta/2$
  - Solve for a parameter estimate by substituting first sample moment for $E(Y_i)$

$$E(Y_i) = \theta/2 \quad \Rightarrow \quad \overline{Y} = \tilde{\theta}/2 \quad \Rightarrow \quad \tilde{\theta} = 2\,\overline{Y}$$

- Discussion
  - Simple to apply in many problems, though may have to solve system of equations
  - Central limit theorem implies “nice” properties for the moment estimator
  - Modern generalization known as “estimating equations”

Maximum Likelihood Estimation

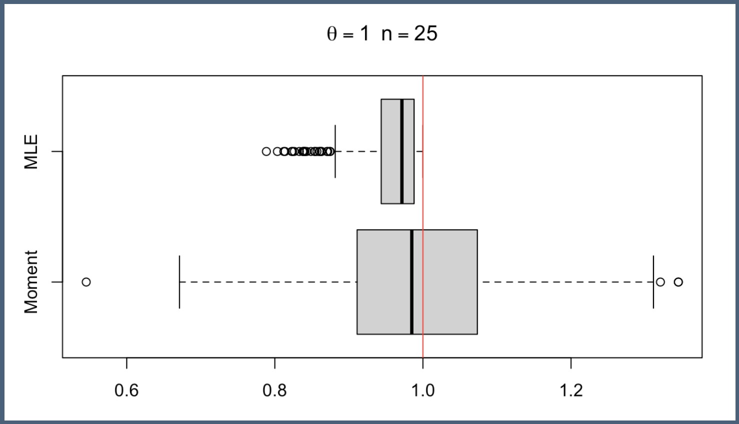
- Maximum likelihood
  - Requires a probability model rather than an expectation
- Example
  - Data $Y_1, \ldots, Y_n$ is sample from Uniform[0, θ]
  - Want to find an estimator for unknown parameter θ
- MLE
  - Likelihood $$P(Y_1, \ldots, Y_n) = (1/\theta)^n, \quad 0 \le Y_i \le \theta$$
    which we can express as $$P(Y_1, \ldots, Y_n) = (1/\theta)^n \, I_{\max(Y_i) \le \theta}$$
  - Maximum likelihood estimator (MLE) is $\hat{\theta} = \max_i Y_i$

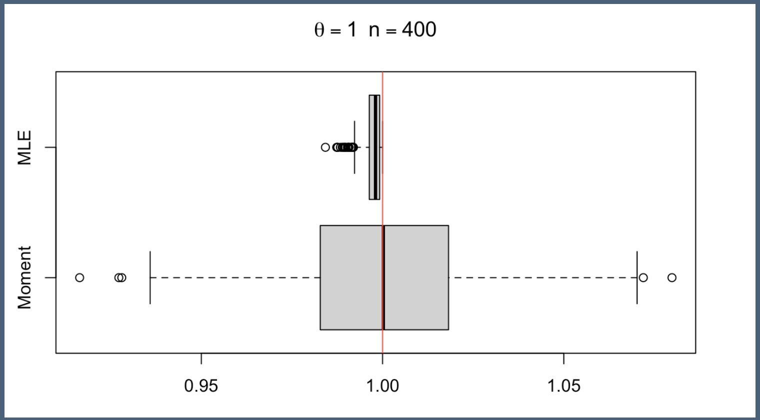


- Discussion
  - Need to know the probability distribution
  - Maximizing the likelihood usually implies calculus, but not always!
  - Theory implies that MLE is typically the best estimator (smallest standard error)

Comparison in Example

- The MLE makes more efficient use of the data
  - Slightly biased but less variable
  - MLE has broad collection of good properties (approximately unbiased, small standard error)
- Example
  - Varying sample sizes when sampling from uniform[0,1] ($\theta = 1$)
  - Moment estimator is unbiased but has much larger variability





Estimating Autoregressions

- Assume time series is AR(p) with p known, normally distributed

Presume mean is known to be zero

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + w_t$$

- Moment estimates from Yule-Walker equations

Definition 4.2.2

  - Multiply by X_t, then take expectation

$$\gamma(0) = \phi_1 \gamma(1) + \cdots + \phi_p \gamma(p) + \sigma_w^2$$

  - Multiply by lags, then take expectation

$$\gamma(1) = \phi_1 \gamma(0) + \cdots + \phi_p \gamma(p-1) \quad \Rightarrow \quad \rho(1) = \phi_1 \; 1 + \cdots + \phi_p \rho(p-1)$$

$$\gamma(2) = \phi_1 \gamma(1) + \cdots + \phi_p \gamma(p-2) \quad \Rightarrow \quad \rho(2) = \phi_1 \rho(1) + \cdots + \phi_p \rho(p-2)$$

  - Obtain system of equations to solve for ϕ in terms of the autocovariances/autocorrelations

$$\begin{aligned}
\gamma(1) &= \phi_1 \gamma(0) + \phi_2 \gamma(1) + \cdots + \phi_p \gamma(p-1) \\
\gamma(2) &= \phi_1 \gamma(1) + \phi_2 \gamma(2) + \cdots + \phi_p \gamma(p-2) \\
\vdots &= \vdots \\
\gamma(p) &= \phi_1 \gamma(p-1) + \phi_2 \gamma(2) + \cdots + \phi_p \gamma(0)
\end{aligned}$$

Yule-Walker Estimates vs LS

- Assume time series is AR(p) with p known, normally distributed

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + w_t$$

Presume mean is known to be zero

- Yule-Walker estimates have large bias
  - Consider an AR(1) model

$$\tilde{\phi}_{1,YW} = \frac{\hat{\gamma}_1}{\hat{\gamma}_0} = \frac{\sum_{t=2}^{n} Y_t Y_{t-1}}{\sum_{t=1}^{n} Y_t^2}$$

Compare to Example 4.23. Sums written to make it easier to compare to LS

    Ratio has 1 more term in the sum in the denominator than in the numerator

  - Approach always yields a stationary model, $|\tilde{\phi}_{1,YW}| \leq 1$

- Least squares estimator
  - Regress X_t on X_{t-1}

$$\tilde{\phi}_1 = \frac{\sum_{t=2}^{n} Y_t Y_{t-1}}{\sum_{t=2}^{n} Y_{t-1}^2}$$

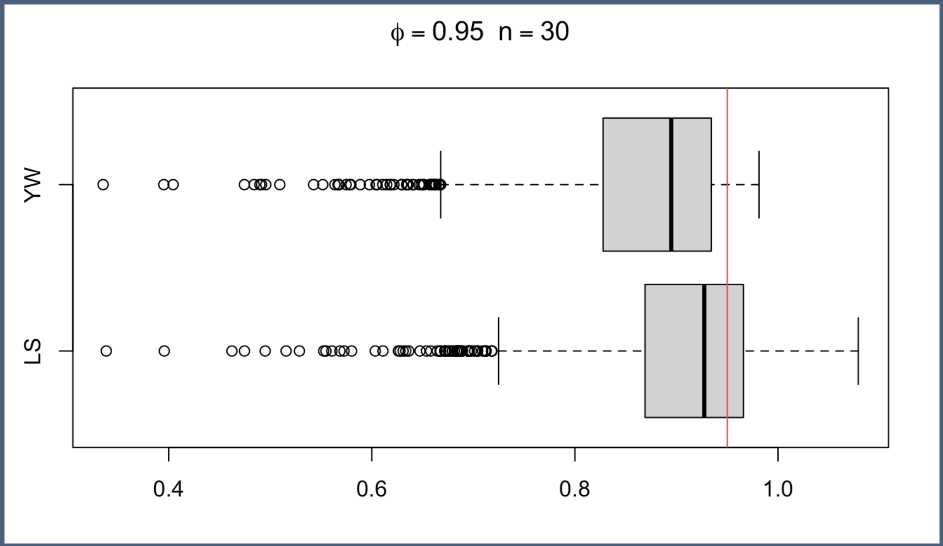
  - Magnitude of estimate could be larger than 1.

Yule-Walker Estimates vs LS

- Assume time series is AR(1), normally distributed

$$X_t = \phi_1 X_{t-1} + w_t$$

  - Mean is known to be 0
- Yule-Walker estimates have larger MSE
  - Yule Walker implies a stationary model but resulting estimator has large bias with similar variance

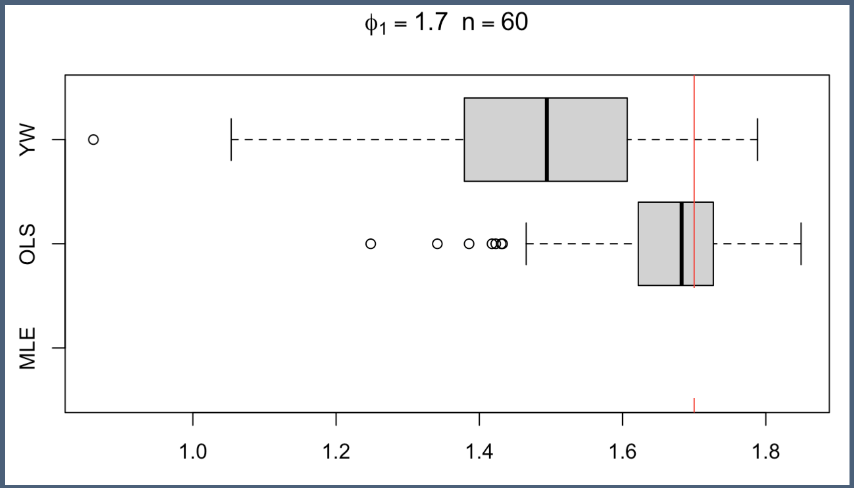


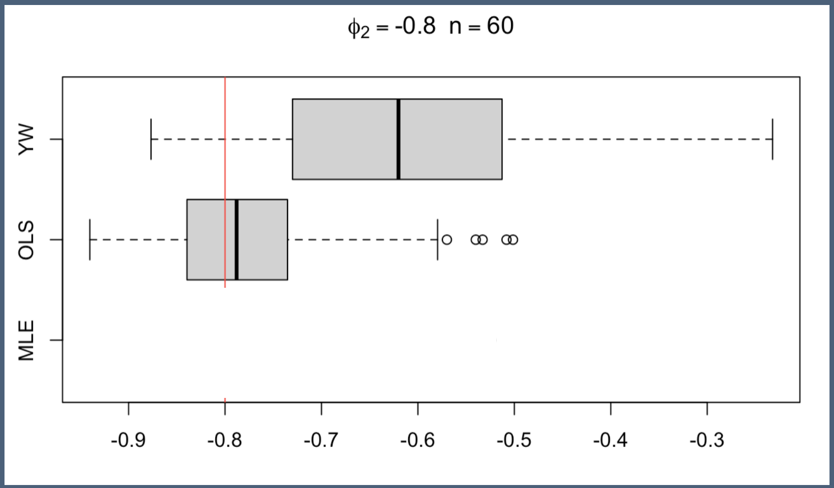
Yule-Walker Estimates vs LS

- Assume time series is AR(1), normally distributed

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + w_t$$

  - R-code estimates allow non-zero mean (center the data using sample mean)
- Yule-Walker estimates have larger MSE
  - Yule Walker implies a stationary model but result is a biased estimator





Maximum Likelihood for AR

- Likelihood function
  - Use of the sequential structure of a time series
  - Factor the joint distribution into a product of conditional distributions

$$P(X_1, \ldots, X_n) = P(X_1)\, P(X_2 \,|\, X_1)\, P(X_3 \,|\, X_1, X_2)\, P(X_4 \,|\, X_1, X_2, X_3) \cdots P(X_n \,|\, X_{n-1}, \ldots, X_1)$$

Expression appears in theory of large language models

- Likelihood function for an AR(1) with mean μ = 0
  - Only need to condition on one prior value

$$P(X_1, \ldots, X_n) = P(X_1)\, P(X_2 \,|\, X_1)\, P(X_3 \,|\, X_2)\, P(X_4 \,|\, X_3) \cdots P(X_n \,|\, X_{n-1})$$

  - Logs convert product into a sum

$$\log P(X_1, \ldots, X_n) = \log P(X_1) + \sum_{t=2}^{n} \log P\left(X_t \,|\, X_{t-1}\right)$$

  - Summands have a simple form when P denotes a normal distribution

$$\sum_{t=2}^{n} \log P(X_t \,|\, X_{t-1}) = -\frac{n-1}{2} \log(2\pi\, \sigma_w^2) - \frac{\boxed{\sum_{t=2}^{n} (X_t - \phi_1 X_{t-1})^2}}{2\sigma_w^2}$$

Least squares

MLE for AR(1) Model

- Marginal distribution of first term
  - Recall formula for a normal distribution with mean μ and variance σ^2

$$f(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}$$

  - First term of the log likelihood for zero-mean AR(1)

$$\log P(X_1) = -\frac{1}{2}\left(\log(2\pi\sigma_x^2) + \frac{(X_1-\mu)^2}{\sigma_x^2}\right) \text{ where } \sigma_x^2 = \frac{\sigma_w^2}{1-\phi^2} \text{ and } \mu = 0$$

- MLE estimator
  - Take derivative of the log-likelihood
  - Derivative of sum of the last n-1 terms produces

$$\tilde{\phi} = \frac{\sum_{t=2}^{n} X_t X_{t-1}}{\sum_{t=2}^{n} X_{t-1}^2} \quad \text{and} \quad \widetilde{\sigma_w^2} = \frac{\sum_{t=2}^{n} (X_t - \tilde{\phi} X_{t-1})^2}{n-1}$$

  - First term of the likelihood modifies these slightly

Estimating the MA(1) Model

- Process is a first-order, invertible moving average

$$X_t = w_t + \theta_1 w_{t-1}$$

  - Consider a moment estimator
  - First two autocovariances are

$$\gamma(0) = (1 + \theta_1^2)\ \sigma_w^2$$
$$\gamma(1) = \theta_1\ \sigma_w^2$$

  - Ratio of covariances removes variance term allowing solution from quadratic equation

$$\frac{\gamma(1)}{\gamma(0)} = \frac{\theta_1}{1 + \theta_1^2} \quad \Rightarrow \quad \hat{\theta}_1 = \frac{1 \pm \sqrt{1 - 4\,\hat{\rho}(1)^2}}{2\hat{\rho}(1)}$$

- Problem: quadratic equation has two solutions!
  - Take the solution that's an invertible process – if there is one.
  - If $0.5 < \hat{\rho}(1)$, then take value close to 1.0 (max value for an invertible MA(1) process)

Text has an example of a process for which estimated correlation is larger than .5

Complications: Redundant Models

- Easy for ARMA(p,q) model to masquerade as ARMA(p+1,q+1)
  - Problem becomes evident when consider the polynomial representation
- Extraneous terms
  - Backshift polynomial representation of process
  - Following process is stationary and invertible (a well-posed ARMA(p,q) process)

$$\phi(B)\, X_t = \theta(B)\, w_t$$

  - Multiply both sides by $\eta(B) = (1 - 0.5\, B)$

$$\eta(B)\, \phi(B)\, X_t = \eta(B)\, \theta(B)\, w_t$$

$$\widetilde{\phi}\,(B)\, X_t = \widetilde{\theta}(B)\, w_t$$

  - Process is stationary and invertible ARMA(p+1,q+1), but has a "common factor"
- Detection
  - Not evident in the coefficients themselves, so you need to inspect zeros of polynomials

$$X_t = 0.8X_{t-1} + 0.5w_{t-1} + w_t \quad \equiv \quad X_t = 0.3X_{t-1} + 0.4X_{t-2} + w_{t-1} + 0.25w_{t-2} + w_t$$

Text, Example 4.9

Complication for estimation?

What's next?

- Maximum likelihood
  - More examples
  - Sampling properties of the estimates: standard errors and confidence intervals

- Practical question for data analysis
  - These look a lot like regression models, but where are the diagnostic plots?