



ECON90024 – FORECASTING IN ECONOMICS & BUSINESS

LECTURE 7: ESTIMATING & FORECASTING ARMA MODELS

TODAY'S LECTURE

- Estimating autoregressive models using OLS
- Maximum Likelihood Estimation of ARMA Models (derivations not examinable)
- Computing point and interval forecasts of ARMA processes

ESTIMATING AR(p) MODELS USING OLS

- Let's go back to our simple regression model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad i = 1, 2, \dots, n$$

- If the following ***Gauss-Markov*** assumptions hold:

1. The population model is linear in its parameters and there are no omitted variables.

2. $E[u_i|X] = E[u_i|X_1, X_2, \dots, X_n] = 0$ (Zero Conditional Mean)

3. $Var(u_i|X) = Var(u_i|X_1, X_2, \dots, X_n) = \sigma^2$ (Homoscedastic Errors)

- Then the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are not only consistent estimators of the parameters β_0 and β_1 but they are the ***Best Linear Unbiased Estimators (BLUE)***.

ESTIMATING AR(p) MODELS USING OLS

- Recall from the first lecture that the OLS estimators and corresponding estimates for a simple linear regression are given by:

Parameter	Estimator	Estimate
β_0	$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$	$b_0 = \bar{y} - b_1 \bar{x}$
β_1	$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$	$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$

- Moreover, it can be shown that

$$\hat{\beta}_0 \rightarrow_d N(\beta_0, \sigma_{\hat{\beta}_0}^2)$$

$$\hat{\beta}_1 \rightarrow_d N(\beta_1, \sigma_{\hat{\beta}_1}^2)$$

These limiting results give us the sampling distributions of our estimators and thus allow us to perform inference!

ESTIMATING AR(p) MODELS USING OLS

- We say that an estimator $\hat{\beta}$ is **unbiased** if its sampling distribution is centered around the true population parameter β

$$E[\hat{\beta}] = \beta$$

- We say that an estimator $\hat{\beta}$ is **consistent** if it converges in probability to the true population parameter β

$$\hat{\beta} \rightarrow_p \beta$$

- That is to say that as the sample size $n \rightarrow \infty$, the probability that the estimator $\hat{\beta}$ (remember, it is a random variable) takes value β approaches 1.
- Another way to think about this property is that the sampling distribution of $\hat{\beta}$ will collapse around the true parameter value β as the sample size n becomes infinitely large.

ESTIMATING AR(p) MODELS USING OLS

- Therefore if the Gauss-Markov assumptions are satisfied, the OLS estimators will possess extremely attractive properties for the purposes of inference.
- They will be the best estimators in the sense that they will have the lowest variance of any estimator possible.
- Generalizing these results to a multiple regression context is straightforward. We only need to add an additional condition that there exist ***no multicollinearity*** (i.e. no explanatory variable can be written as a linear function of another).

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_k X_{k,i} + u_i$$

ESTIMATING AR(p) MODELS USING OLS

- Looking at an AR(p) model, we can see that it shares many similarities to the linear regression model:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t$$

$$\varepsilon_t \sim iid(0, \sigma^2)$$

$$E[Y_t] = 0$$

$$E[Y_t | \Omega_{t-1}] = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p}$$

$$var(Y_t | \Omega_{t-1}) = E[\varepsilon_t^2 | \Omega_{t-1}] = \sigma^2$$

- Therefore, intuition tells us that OLS should be an appropriate and desirable way to compute estimates of the autoregressive parameters $\phi_1, \phi_2, \dots, \phi_p$.

ESTIMATING AR(p) MODELS USING OLS

- As it turns out, the OLS estimators for the autoregressive coefficients will be **consistent**!
- However, the estimators will be **biased**. To see how, let's illustrate using the case of an AR(1),

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

$$\varepsilon_t \sim iid(0, \sigma^2)$$

- Given a sample size of T , the OLS estimator of the autoregressive parameter ϕ will be given by,

$$\hat{\phi} = \frac{\sum_{t=2}^T (Y_{t-1} - \bar{Y}_{t-1})(Y_t - \bar{Y}_t)}{\sum_{t=2}^T (Y_{t-1} - \bar{Y}_{t-1})^2} = \frac{\sum_{t=2}^T (Y_{t-1})(Y_t)}{\sum_{t=2}^T (Y_{t-1})^2}$$

- Where the second equality is obtained using the fact that $\bar{Y}_{t-1} = \bar{Y}_t = E[Y_t] = 0$

ESTIMATING AR(p) MODELS USING OLS

- Using the fact that $Y_t = \phi Y_{t-1} + \varepsilon_t$ we can write,

$$\hat{\phi} = \frac{\sum_{t=2}^T (Y_{t-1})(\phi Y_{t-1} + \varepsilon_t)}{\sum_{t=2}^T (Y_{t-1})^2} = \frac{\phi \sum_{t=2}^T (Y_{t-1})^2}{\sum_{t=2}^T (Y_{t-1})^2} + \frac{\sum_{t=2}^T (Y_{t-1})(\varepsilon_t)}{\sum_{t=2}^T (Y_{t-1})^2}$$

- Simplification yields the following expression

$$\hat{\phi} = \phi + \frac{\sum_{t=2}^T (Y_{t-1})(\varepsilon_t)}{\sum_{t=2}^T (Y_{t-1})^2}$$

- When we apply the expectations operator, we obtain,

$$E[\hat{\phi}] = \phi + E \left[\frac{\sum_{t=2}^T (Y_{t-1})(\varepsilon_t)}{\sum_{t=2}^T (Y_{t-1})^2} \right] = \phi + E \left[\sum_{t=2}^T \left(\frac{Y_{t-1}}{\sum_{t=2}^T (Y_{t-1})^2} \right) \varepsilon_t \right]$$

ESTIMATING AR(p) MODELS USING OLS

- Let's look at the last term more closely:

$$E \left[\sum_{t=2}^T \left(\frac{Y_{t-1}}{\sum_{t=2}^T (Y_{t-1})^2} \right) \varepsilon_t \right] = \sum_{t=2}^T E \left[\left(\frac{Y_{t-1}}{\sum_{t=2}^T (Y_{t-1})^2} \right) \varepsilon_t \right]$$

- The first term in the outer sum (when $t = 2$) is

$$E \left[\frac{Y_1 \varepsilon_2}{\sum_{t=2}^T (Y_{t-1})^2} \right] = E \left[\frac{Y_1 \varepsilon_2}{Y_1^2 + Y_2^2 + \dots + Y_{T-1}^2} \right]$$

- Looking at our model,

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

- We can see clearly that ε_2 is independent of Y_{t-1} . However, ε_2 is not independent of Y_2, Y_3, \dots, Y_T .

ESTIMATING AR(p) MODELS USING OLS

- Specifically, if $\phi > 0$, then a positive realization of ε_2 will contribute positively to Y_2, Y_3, \dots, Y_T . Therefore if we define,

$$Z_1 = \frac{Y_1}{Y_1^2 + Y_2^2 + \dots + Y_{T-1}^2}$$

- We can see that ε_2 will covary negatively with Z_1

$$\text{Cov}(Z_1, \varepsilon_2) = E[Z_1 \varepsilon_2] - E[Z_1]E[\varepsilon_2] = E[Z_1 \varepsilon_2] < 0$$

- We can see that this is going to be true for all t . Therefore when $\phi > 0$, it will be the case that

$$\sum_{t=2}^T E \left[\left(\frac{Y_{t-1}}{\sum_{t=2}^T (Y_{t-1})^2} \right) \varepsilon_t \right] < 0$$

- And so, we have the result that for a finite T , the OLS estimator for the autoregressive coefficient is biased. That is,

$$E[\hat{\phi}] \neq \phi$$

ESTIMATING AR(p) MODELS USING OLS

- Determining the direction of bias will be more complicated for higher order autoregressive models since the manner in which an innovation propagates through time will depend on the combination of autoregressive coefficients,

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t$$

- This result tells us that estimating an autoregressive model that is of a high order with a small number of observations will lead to severely biased and thus potentially misleading estimates! Another reason why we always strive for the most parsimonious model possible.
- The bias will disappear as the sample size becomes arbitrarily large, $T \rightarrow \infty$

MOTIVATING MAXIMUM LIKELIHOOD ESTIMATION (MLE)

- We have seen in the previous few slides that the method of ordinary least squares can be used to estimate AR models.
- However, we run into a problem with OLS when we try to estimate an MA model:

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

- In a standard OLS regression framework, realizations of the errors/innovations ε_t are obtained as a by-product of the estimation (i.e. they are the residuals!).
- We cannot estimate an MA model using OLS as there is no way to specify the errors/innovations as exogenously observed explanatory variables.
- If we cannot estimate MA models using OLS, then we cannot estimate ARMA models either!

THE METHOD OF MAXIMUM LIKELIHOOD (MLE)

- Maximum likelihood estimation starts from the specification of the joint distribution function of the data. Let's suppose that we have a sequence of n *i. i. d.* Normal random variables with mean μ and variance σ^2 . Each X_i will have the density function

$$f(X_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(X_i - \mu)^2}{2\sigma^2} \right\}$$

- Therefore, if we observe a realization x_i , the value of the density function is given by

$$f(X_i = x_i; \mu, \sigma^2) = f(x_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\}$$

THE METHOD OF MAXIMUM LIKELIHOOD (MLE)

- The ***joint density function*** of the sequence will then be given by,

$$f(X_1, X_2, \dots, X_n; \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(X_i - \mu)^2}{2\sigma^2} \right\}$$

- If μ and σ^2 as *known*, it is clear that we would be able to evaluate the density function for a set of observations

$$f(x_1, x_2, \dots, x_n; \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\}$$

THE METHOD OF MAXIMUM LIKELIHOOD (MLE)

- The fundamental problem that we face as statisticians however is that the parameters are ***unknown quantities***. So let's define

$$L(\mu, \sigma^2; x_1, x_2, \dots, x_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\}$$

- This is known as the ***likelihood function***. This function represents the value of the joint density function for a set of observations $\{x_1, x_2, \dots, x_n\}$ as a function of μ and σ^2 .
- Using the likelihood function, we can derive estimators $\hat{\mu}$ and $\hat{\sigma}^2$ for the unknown parameters μ and σ^2 by asking the following question:

“Given the set of observations $\{x_1, x_2, \dots, x_n\}$, what are the values of $\hat{\mu}$ and $\hat{\sigma}^2$ that maximize the likelihood function?”

THE METHOD OF MAXIMUM LIKELIHOOD

(MLE)

- As with OLS, we will compute the maximum likelihood estimates using the tools of high school calculus. But first we recognize that working directly with the likelihood function can be cumbersome due to the fact that it is a n -product.

$$L(\mu, \sigma^2; x_1, x_2, \dots, x_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\} = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right\}$$

- For analytical convenience, we will transform the likelihood function using natural logs to obtain the *log likelihood*,

$$\log L(\mu, \sigma^2; x_1, x_2, \dots, x_n) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

- Taking logs is a **monotonic transformation**, the values of $\hat{\mu}$ and $\hat{\sigma}^2$ that maximize the likelihood function will also maximize the log likelihood function.

THE METHOD OF MAXIMUM LIKELIHOOD (MLE)

- To compute the maximum likelihood estimates, we take the first derivative of the log likelihood with respect to each parameter and set it equal to zero

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^n (x_i - \hat{\mu}) = 0$$

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i=1}^n (x_i - \hat{\mu})^2 = 0$$

- Solving for $\hat{\mu}$ and $\hat{\sigma}^2$, we obtain,

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 = s^2$$

THE METHOD OF MAXIMUM LIKELIHOOD (MLE)

- To estimate a simple linear regression via maximum likelihood, we begin with the population model, where for $i = 1, \dots, n$

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- To use maximum likelihood, we need to impose a distributional assumption on the errors u_i . Let's suppose that

$$u_i \sim iid N(0, \sigma^2)$$

- Note that by imposing this assumption we will immediately satisfy the Gauss-Markov assumptions:

$$E[u_i|X] = E[u_i|X_1, X_2, \dots, X_n] = 0$$

$$Var(u_i|X) = Var(u_i|X_1, X_2, \dots, X_n) = \sigma^2$$

THE METHOD OF MAXIMUM LIKELIHOOD (MLE)

- Therefore, the log likelihood for the simple regression model is given by,

$$\log L(\mu, \sigma^2; u_1, u_2, \dots, u_n) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (u_i)^2$$

- This is an expression in terms of the regression errors which are not directly observed. To express the log likelihood in terms of the observable data, we use the fact that

$$u_i = Y_i - \beta_0 - \beta_1 X_i$$

- To write

$$\log L(\mu, \sigma^2; x_1, y_1, \dots, x_n, y_n) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

THE METHOD OF MAXIMUM LIKELIHOOD (MLE)

- Again, to compute maximum likelihood estimates of the slope and intercept parameters, we derive the following first order conditions,

$$\frac{\partial \log L}{\partial \beta_0} = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial \log L}{\partial \hat{\beta}_1} = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial \log L}{\partial \hat{\sigma}^2} = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 0$$

- Notice that the first order conditions for the maximum likelihood estimates will yield the same solutions as OLS! Therefore, under the assumption of normally distributed regression errors, OLS and MLE will yield the same estimates.

THE METHOD OF MAXIMUM LIKELIHOOD (MLE) FOR AN AR(1)

- Now let's consider a first order autoregressive process,

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

$$\varepsilon_t \sim i.i.d. N(0, \sigma^2)$$

- For notational compactness, let's collect the set of population parameters to be estimated as $\boldsymbol{\theta} = \{c, \phi, \sigma^2\}$.
- As a first step, let's consider the **unconditional** probability density function of Y_1 , the first observation in the sample. We can easily show (using backward substitution or lag operators) that when $|\phi| < 1$,

$$var(Y_1) = \phi^2 var(Y_0) + \sigma^2$$

If $|\phi| < 1$, then $var(Y_1) = var(Y_0)$ so that

$$var(Y_1)(1 - \phi^2) = \sigma^2$$

$$E[Y_1] = \mu = \frac{c}{1 - \phi}$$

$$E[(Y_1 - \mu)^2] = \frac{\sigma^2}{1 - \phi^2}$$

$$E[Y_1] = c + \phi E[Y_0]$$

If $|\phi| < 1$, then $E[Y_1] = E[Y_0]$ so that

$$E[Y_1](1 - \phi) = c$$

THE METHOD OF MAXIMUM LIKELIHOOD (MLE) FOR AN AR(1)

- Since all the ε_t 's are normal, it must be the case that Y_1 is normal as the sum of normal random variables is itself a normal random variable. Thus the density for the first observation takes the form,

$$f(y_1; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi} \sqrt{\frac{\sigma^2}{(1-\phi^2)}}} \exp \left\{ \frac{-\left(y_1 - \frac{c}{1-\phi}\right)^2}{\frac{2\sigma^2}{(1-\phi^2)}} \right\}$$

- Next, consider the distribution of the second observation Y_2 **conditional** on observing $Y_1 = y_1$. From the AR(1) specification we know that,

$$Y_2 = c + \phi Y_1 + \varepsilon_2$$

- Therefore,

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

THE METHOD OF MAXIMUM LIKELIHOOD (MLE) FOR AN AR(1)

- That is,

$$f(y_2|Y_1 = y_1; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(y_2 - c - \phi y_1)^2}{2\sigma^2}\right\}$$

- From introductory stats, we know that by rearranging Bayes rule, the joint density can be obtained as the product of the marginal and the conditional densities,

$$f(X, Y) = f(X|Y)f(Y)$$

- Therefore,

$$f(y_1, y_2; \boldsymbol{\theta}) = f(y_2|y_1; \boldsymbol{\theta})f(y_1; \boldsymbol{\theta})$$

- Using the same argument for the third observation y_3 , we have that

$$f(y_1, y_2, y_3; \boldsymbol{\theta}) = f(y_3|y_1, y_2; \boldsymbol{\theta})f(y_1, y_2; \boldsymbol{\theta}) = f(y_3|y_1, y_2; \boldsymbol{\theta})f(y_2|y_1; \boldsymbol{\theta})f(y_1; \boldsymbol{\theta})$$

THE METHOD OF MAXIMUM LIKELIHOOD (MLE) FOR AN AR(1)

- Therefore the likelihood for the complete sample can be calculated as,

$$L(\boldsymbol{\theta}; y_1, \dots, y_T) = f(y_1; \boldsymbol{\theta}) \times f(y_2, \dots, y_T | y_1; \boldsymbol{\theta}) = f(y_1; \boldsymbol{\theta}) \times \prod_{t=2}^T f(y_t | y_{t-1}; \boldsymbol{\theta})$$

- The log likelihood is therefore given by

$$\log L(\boldsymbol{\theta}; y_1, \dots, y_T) = \log(f(y_1; \boldsymbol{\theta})) + \sum_{t=2}^T \log(f(y_t | y_{t-1}; \boldsymbol{\theta}))$$

- We have previously computed

$$f(y_1; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi} \sqrt{\frac{\sigma^2}{(1-\phi^2)}}} \exp \left\{ \frac{-(y_1 - \frac{c}{1-\phi})^2}{\frac{2\sigma^2}{(1-\phi^2)}} \right\}$$

$$f(y_t | y_{t-1}; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ \frac{-(y_t - c - \phi y_{t-1})^2}{2\sigma^2} \right\}$$

THE METHOD OF MAXIMUM LIKELIHOOD (MLE) FOR AN AR(1)

- Then,

$$\log(f(y_1; \boldsymbol{\theta})) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log\left(\frac{\sigma^2}{1-\phi^2}\right) - \frac{\left(y_1 - \frac{c}{1-\phi}\right)^2}{\frac{2\sigma^2}{(1-\phi^2)}}$$

$$\log(f(y_t|y_{t-1}; \boldsymbol{\theta})) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{-(y_t - c - \phi y_{t-1})^2}{2\sigma^2}$$

- Putting it all together,

$$\log L(\boldsymbol{\theta}; y_1, \dots, y_T) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log\left(\frac{\sigma^2}{1-\phi^2}\right) - \frac{\left(y_1 - \frac{c}{1-\phi}\right)^2}{\frac{2\sigma^2}{(1-\phi^2)}} - \frac{T-1}{2} \log(2\pi) - \frac{T-1}{2} \log(\sigma^2) - \sum_{t=2}^T \frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2}$$

- Differentiating this with respect to $\boldsymbol{\theta}$ produces a set of nonlinear equations for which there is no simple closed form solution for $\boldsymbol{\theta}$ in terms of the observations y_1, y_2, \dots, y_T . To compute estimates from this exact likelihood function, we will need a numerical solver.

THE METHOD OF MAXIMUM LIKELIHOOD (MLE) FOR AN AR(1)

- However, if we condition on the first observation y_1 , that is, treat it as given, we can obtain the following ***conditional log likelihood function***,

$$\log L(\boldsymbol{\theta}; y_2, \dots, y_T | y_1) = -\frac{T-1}{2} \log(2\pi) - \frac{T-1}{2} \log(\sigma^2) - \sum_{t=2}^T \frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2}$$

- This conditional log likelihood function will produce the same first order conditions as OLS!
- If the sample size T is large, then the first observation makes a negligible contribution to the total likelihood.
- The exact *MLE* and conditional *MLE* turn out to have the same large sample properties. Therefore, in practice, almost all software packages will perform conditional MLE when estimating ARMA models.

THE METHOD OF MAXIMUM LIKELIHOOD FOR AN MA(1)

- Now let's consider an MA(1) process,

$$Y_t = \mu + \varepsilon_t + \theta \varepsilon_{t-1}$$

$$\varepsilon_t \sim i.i.d. N(0, \sigma^2)$$

- Again, for notational compactness, let's collect the set of population parameters to be estimated as $\boldsymbol{\theta} = \{\mu, \theta, \sigma^2\}$.
- We saw that in the previous slide that the approach to estimating parameters by MLE involves the construction of the **likelihood function** as the product of conditional densities. We are going to pursue a similar approach with regards to the MA(1) model.
- Thus if we were to condition on ε_{t-1} , then according to the above specification, the conditional density of y_t is given by

$$f(y_t | \varepsilon_{t-1}; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ \frac{-(y_t - \mu - \theta \varepsilon_{t-1})^2}{2\sigma^2} \right\}$$

THE METHOD OF MAXIMUM LIKELIHOOD FOR AN MA(1)

- Suppose that we knew with certainty that $\varepsilon_0 = 0$, then

$$f(y_1|\varepsilon_0 = 0; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(y_1 - \mu)^2}{2\sigma^2}\right\}$$

- Moreover, given observations of y_1 and y_2 the values of ε_1 and ε_2 can be calculated from

$$\begin{aligned}\varepsilon_1 &= y_1 - \mu \\ \varepsilon_2 &= y_2 - \mu - \theta\varepsilon_1\end{aligned}$$

- In fact, given $\varepsilon_0 = 0$, the full sequence $\{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_T\}$ can be calculated from the set of observations $\{y_1, y_2, \dots, y_T\}$ by iterating on

$$\varepsilon_t = y_t - \mu - \theta\varepsilon_{t-1}$$

THE METHOD OF MAXIMUM LIKELIHOOD FOR AN MA(1)

- The conditional density of the t -th observation can then be calculated as:

$$f(y_t|y_1, y_2, \dots, y_{t-1}, \varepsilon_0 = 0; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(y_t - \mu)^2}{2\sigma^2}\right\} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-\varepsilon_t^2}{2\sigma^2}\right\}$$

- It follows that the sample likelihood would then be the product of these individual conditional densities,

$$L(\boldsymbol{\theta}; y_1, y_2, \dots, y_T | \varepsilon_0 = 0) = \prod_{t=1}^T \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-\varepsilon_t^2}{2\sigma^2}\right\}$$

- So that the conditional log likelihood is,

$$\log L(\boldsymbol{\theta}; y_1, y_2, \dots, y_T | \varepsilon_0 = 0) = -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log(\sigma^2) - \sum_{t=1}^T \frac{\varepsilon_t^2}{2\sigma^2}$$

- Thus, For a particular numerical value of $\boldsymbol{\theta}$, we can calculate the sequence of ε_t implied by the data and evaluate the above log likelihood. The value of $\boldsymbol{\theta}$ that maximizes the conditional log likelihood will be found by numerical optimization.

Setting $\varepsilon_0 = 0$ is not a very strong assumption since the impact of innovations in an MA model are very short lived!

THE METHOD OF MAXIMUM LIKELIHOOD FOR AN ARMA(p, q)

- Now let's consider an ARMA(p, q)

$$Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

$$\varepsilon_t \sim iid N(0, \sigma^2)$$

- The set of parameters to be estimated are, $\boldsymbol{\theta} = \{c, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \sigma^2\}$
- From our analysis of the AR and MA models, we can see that the approach to estimating an ARMA(p, q) via maximum likelihood will involve conditioning on a set of p initial values of y_t as well as a set of q initial values of ε_t

THE METHOD OF MAXIMUM LIKELIHOOD FOR AN ARMA(p, q)

- Specifically, taking initial values $\mathbf{y}_0 = \{y_0, y_{-1}, y_{-2}, \dots, y_{-p+1}\}$ and $\boldsymbol{\varepsilon}_0 = \{\varepsilon_0, \varepsilon_{-1}, \varepsilon_{-2}, \dots, \varepsilon_{-q+1}\}$ as given, the sequence $\{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_T\}$ can be calculated from $\{y_1, y_2, \dots, y_T\}$ by iterating on

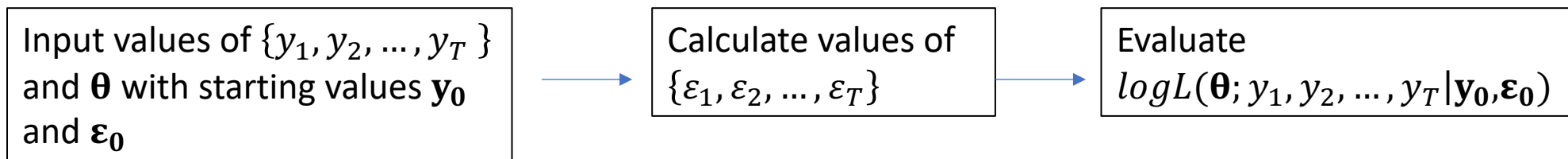
$$\varepsilon_t = y_t - c - \phi_1 y_{t-1} - \dots - \phi_p y_{t-p} - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q}$$

- Doing so for $t = 1, 2, \dots, T$, we can build the conditional log likelihood as

$$\log L(\boldsymbol{\theta}; y_1, y_2, \dots, y_T | \mathbf{y}_0, \boldsymbol{\varepsilon}_0) = -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log(\sigma^2) - \sum_{t=1}^T \frac{\varepsilon_t^2}{2\sigma^2}$$

THE METHOD OF MAXIMUM LIKELIHOOD FOR AN ARMA(p, q)

- Again, a particular numerical value of $\boldsymbol{\theta}$, we can calculate the sequence of ε_t implied by the data and evaluate the above log likelihood. The value of $\boldsymbol{\theta}$ that maximizes the conditional log likelihood will be found by numerical optimization.
- When T is large, the choice of initial y 's and ε 's will have a negligible effect on the log likelihood. Typically, the optimization algorithms in R or other software packages will set the initial ε 's to zero and the initial y 's to their expected value or the actual observed values (Box and Jenkins, 1976).
- There are many different techniques that can be used in numerical maximization (grid search, steepest ascent, etc) but the overall approach is the same:



FORECASTING AN MA PROCESS

- Let's suppose we have a data generating process that is described as an MA(2)

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}$$

$$\varepsilon_t \sim i.i.d.(0, \sigma^2)$$

- Suppose that we are standing at time T and we want to compute a one-step ahead forecast for time $T + 1$. Since the forecast is simply the conditional mean, we can write this as,

$$E[Y_{T+1} | \Omega_T] = \theta_1 \varepsilon_T + \theta_2 \varepsilon_{T-1}$$

- The two-step ahead forecast will therefore be,

$$E[Y_{T+2} | \Omega_T] = \theta_2 \varepsilon_T$$

- The h -step ahead forecast for $h > 2$ will be,

$$E[Y_{T+h} | \Omega_T] = 0$$

FORECASTING AN MA PROCESS

- The forecast errors will therefore be given by,

$$Y_{T+1} - E[Y_{T+1}|\Omega_T] = \varepsilon_{T+1}$$

$$Y_{T+2} - E[Y_{T+2}|\Omega_T] = \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1}$$

$$Y_{T+h} - E[Y_{T+h}|\Omega_T] = \varepsilon_{T+h} + \theta_1 \varepsilon_{T+h-1} + \theta_2 \varepsilon_{T+h-2}$$

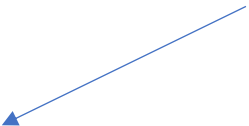
- With the associated forecast error variances,

$$\sigma_1^2 = \sigma^2$$

$$\sigma_2^2 = \sigma^2(1 + \theta_1^2)$$

$$\sigma_h^2 = \sigma^2(1 + \theta_1^2 + \theta_2^2)$$

This is the unconditional
variance of the MA(2)
process!



FORECASTING AN MA PROCESS

- We can see how this generalizes to an MA(q) process,

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

$$E[Y_{T+1} | \Omega_T] = \theta_1 \varepsilon_T + \theta_2 \varepsilon_{T-1} + \cdots + \theta_q \varepsilon_{T-q}$$

$$E[Y_{T+2} | \Omega_T] = \theta_2 \varepsilon_T + \theta_3 \varepsilon_{T-1} + \cdots + \theta_q \varepsilon_{T-q+1}$$

\vdots

$$E[Y_{T+q} | \Omega_T] = \theta_q \varepsilon_T$$

$$E[Y_{T+h} | \Omega_T] = 0 \qquad h > q$$

FORECASTING AN MA PROCESS

- The forecast errors will then be given by,

$$Y_{T+1} - E[Y_{T+1}|\Omega_T] = \varepsilon_{T+1}$$

$$Y_{T+2} - E[Y_{T+2}|\Omega_T] = \varepsilon_{T+1} + \theta_1 \varepsilon_T$$

$$Y_{t+q} - E[Y_{T+q}|\Omega_T] = \varepsilon_{T+q} + \theta_1 \varepsilon_{T+q-1} + \theta_2 \varepsilon_{T+q-2} + \cdots + \theta_{q-1} \varepsilon_{T+1}$$

$$Y_{T+h} - E[Y_{T+h}|\Omega_T] = \varepsilon_{T+h} + \theta_1 \varepsilon_{T+h-1} + \theta_2 \varepsilon_{T+h-2} + \cdots + \theta_q \varepsilon_{T+h-q}$$

FORECASTING AN MA PROCESS

- Therefore, the forecast error variances will grow as the forecast horizon increases,

$$\sigma_1^2 = \sigma^2$$

$$\sigma_2^2 = \sigma^2(1 + \theta_1^2)$$

$$\sigma_q^2 = \sigma^2(1 + \theta_1^2 + \cdots + \theta_{q-1}^2)$$

$$\sigma_h^2 = \sigma^2(1 + \theta_1^2 + \cdots + \theta_{q-1}^2 + \theta_q^2)$$

- If we make the assumption that the innovations ε_t are normally distributed, then the forecast errors will themselves be normal random variables. Thus our 95% prediction interval for the m -step ahead forecast will be defined by

$$E[Y_{T+m} | \Omega_T] \pm 1.96\sigma_m$$

FORECASTING AN AR PROCESS

- Let's now consider an AR(1) model

$$Y_t = \phi_1 Y_{t-1} + \varepsilon_t$$

- Again, let's suppose that we are standing at time T and forecasting into future periods, then we have that,

$$E[Y_{T+1} | \Omega_T] = \phi_1 Y_T$$

$$E[Y_{T+2} | \Omega_T] = \phi_1 (\phi_1 Y_T) = \phi_1^2 Y_T$$

$$E[Y_{T+h} | \Omega_T] = \phi_1^h Y_T$$

FORECASTING AN AR PROCESS

- Then the forecast error for the 1-step ahead forecast will be given by,

$$Y_{T+1} - E[Y_{T+1}|\Omega_T] = \phi_1 Y_T + \varepsilon_{T+1} - \phi_1 Y_T = \varepsilon_{T+1}$$

- The forecast error for the 2-step ahead forecast will be given by

$$Y_{T+2} - E[Y_{T+2}|\Omega_T] = \phi_1 Y_{T+1} + \varepsilon_{T+2} - \phi_1^2 Y_T$$

- Note that since $Y_{T+1} = \phi_1 Y_T + \varepsilon_{T+1}$ we will obtain,

$$Y_{T+2} - E[Y_{T+2}|\Omega_T] = \varepsilon_{T+2} + \phi_1 \varepsilon_{T+1}$$

- Using the same reasoning, we will be able to show that the forecast error for the h -step forecast error will be given by,

$$Y_{T+h} - E[Y_{T+h}|\Omega_T] = \varepsilon_{T+h} + \phi_1 \varepsilon_{T+h-1} + \phi_1^2 \varepsilon_{T+h-2} + \cdots + \phi_1^{h-1} \varepsilon_{T+1}$$

FORECASTING AN AR PROCESS

- The variance of the forecast errors will then be given by:

$$\sigma_1^2 = \sigma^2$$

$$\sigma_2^2 = \sigma^2(1 + \phi_1^2)$$

$$\sigma_h^2 = \sigma^2(1 + \phi_1^2 + \phi_1^4 + \cdots + \phi_1^{2h-2})$$

- Therefore, the variance of the forecast error for an autoregressive process grows with the forecast horizon. That is, for autoregressive processes, forecasts far into the future will be less precise than near horizon forecasts.
- Again, if we make the assumption that the innovations ε_t are normally distributed, then the forecast errors will themselves be normal random variables. Thus our 95% interval forecast for the m -step ahead forecast will be defined by

$$E[Y_{T+m}|\Omega_T] \pm 1.96\sigma_m$$

FORECASTING AN ARMA PROCESS

- Now let's consider an ARMA(p,q)

$$Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

$$\varepsilon_t \sim_{iid} N(0, \sigma^2)$$

- Generating the point and interval forecasts then involves the following steps:
 1. The m -step ahead point forecast is given by $E[Y_{T+m} | \Omega_T]$
 2. The m -step ahead forecast error is given by $Y_{T+m} - E[Y_{T+m} | \Omega_T]$
 3. From the structure of the ARMA model, we know that the forecast error will be a linear function of *i. i. d.* innovations ε_t from which we can compute the variance of the forecast error as a function of $\boldsymbol{\theta} = \{c, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \sigma^2\}$
 4. Then the 95% interval forecast for the m -step ahead forecast will be defined by

$$E[Y_{T+m} | \Omega_T] \pm 1.96 \sigma_m$$

5. We will use the MLE estimates to compute all of these objects.