MBA Forecasting Course
Week One

1 Overview

This lecture makes the connection between the introductory statistics courses and time series models. We will review the bivariate conditional distributions with normal errors. The main point is that the needed information is the mean and variance-covariance. You only need the mean and the covariance structure to determine the conditional expectation. This is your forecast.

Next we introduce some basic time series models. For most of these simple models the forecasts can be determine based on the student's knowledge from prior courses.

The issue of model selection is presented. Given all the possible variables, how do we select the best model?

Next is the introduction to the lag operator. Students should be able to write the AR model with lag operator notation. Note that an AR(1) can be written as an MA(∞) with only one parameter.

We will learn about stationarity. This is needed to give some structure to the time series model.

Finally we introduce the Wold Representation Theorem and think of the ARMA(p, q) model as an approximation to the Wold representation for a weakly stationary time series.

2 A review of conditioning

Consider y and x normal random variables

$$\begin{aligned} y &\sim N\left(\mu_y, \sigma_y^2\right) \\ x &\sim N\left(\mu_x, \sigma_x^2\right) \qquad \text{with} \qquad \text{cov}(y, x) = \sigma_{yx}. \end{aligned}$$

If we know x, how do we get the conditional distribution of y given x? How do we get the distribution of y given the value of x?

We will often transform to remove the mean.

$$\begin{aligned} Y &= y - \mu_y \\ X &= x - \mu_x. \end{aligned}$$

This implies

$$\begin{aligned} Y &\sim N\left(0, \sigma_y^2\right) \\ X &\sim N\left(0, \sigma_x^2\right) \qquad \text{with} \qquad \text{cov}(Y, X) = \sigma_{yx}. \end{aligned}$$

Consider the new random variable

$$Y - \frac{\sigma_{yx}}{\sigma_x^2} X.$$

(Note the similarity to $\hat{\beta} = \sum_{t=1}^{T} x_t y_t / \sum_{t=1}^{T} x_t^2$).

The resulting random variables are

$$\begin{aligned} Y - \frac{\sigma_{yx}}{\sigma_x^2} X &\sim N\left(0, \sigma_y^2 - \frac{\sigma_{yx}^2}{\sigma_x^2}\right) \\ X &\sim N\left(0, \sigma_x^2\right) \quad \text{with} \quad \operatorname{cov}\left(Y - \frac{\sigma_{yx}}{\sigma_x^2} X, X\right) = 0. \end{aligned}$$

This is the conditional distribution of Y given the value for X. It shows how to adjust the distribution of Y if we have additional information X.

Points to note about this derivation

1. Used all the information in X because we have zero covariance after conditioning.
2. The conditional variance did not increase. Unless the series were uncorrelated, the conditional variance decreased.
3. To determine the appropriate linear combination of the conditioning information we needed the covariances.
4. A similar structure holds if X is a vector of variables.
5. Suppose X consisted of lagged values of Y

   $$X = Y_{t-k}$$

   for $k = 1, 2, \ldots, p$. This is the situation we will first address in time series modeling. What we want is to determine the distribution of Y_{t+h} given past information $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$.
6. To determine the conditional distribution we will need to determine the covariance between Y_t and $Y_{t-1}, \ldots, Y_{t-p}$.

3 Some basic models

1. White noise

   $$y_t \sim iid(0, \sigma^2).$$

   Not very interesting but is the basic building block for most time series models.

Given the observed sample $\{ y_1, \ \ldots, \ y_T \} \equiv I_T$ (information at time period T), how will you estimate

$$y_{T+h}?$$

The estimated value is 0.

This is the expected value of y_{T+h} given information at time period T. For this model we have

$$E\left[y_{T+h} \,|I_T\right] = 0.$$

The model implies $y_{T+h} \sim (0, \sigma^2)$. We can use the sample to estimate σ^2.

If the variable looks normally distributed, the 90% confidence interval will be

$$[-1.65\sigma, 1.65\sigma].$$

We may need to estimate σ^2 with something like

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T} y_t^2$$

or

$$s^2 = \frac{1}{T-1}\sum_{t=1}^{T} \left(y_t - \overline{y}\right)^2.$$

You might also consider white noise with mean.

$$y_t \sim iid\left(\mu, \sigma^2\right).$$

2. Linear time trend

$$y_t = \beta t + \epsilon_t$$

where $\epsilon_t \sim iid(0, \sigma^2)$ (white noise).

If $\beta \left(\begin{array}{c} < \\ > \end{array} \right) 0$ the series is random variation around a linear $\left(\begin{array}{c} \text{decreasing} \\ \text{increasing} \end{array} \right)$ function.

Given the observed sample $\{ y_1, \ \ldots, \ y_T \}$, how will you estimate

$$y_{T+h}?$$

The estimated value is $y_T + \beta h$.

This is the expected value of y_{T+h} given information at time period T. For this model we have

$$\begin{aligned} y_{T+h} &= \beta(T+h) + \epsilon_{T+h} \\ &= \beta T + \beta h + \epsilon_{T+h}. \end{aligned}$$

In addition

$$y_T = \beta T + \epsilon_T$$

which implies

$$\beta T = y_T - \epsilon_T.$$

Substituting in this gives

$$\begin{aligned} y_{T+h} &= y_T - \epsilon_T + \beta h + \epsilon_{T+h} \\ &= y_T + \beta h - \epsilon_T + \epsilon_{T+h}. \end{aligned}$$

So that $y_{T+h} \sim (y_T + \beta h, 2\sigma^2)$. The expected value given information at time period T is

$$E[y_{T+h} | I_T] = y_T + \beta h.$$

If the variable is normally distributed, the 90% confidence interval will be

$$[y_T + \beta h - 1.65\sqrt{2}\sigma, y_T + \beta h + 1.65\sqrt{2}\sigma].$$

We may need to estimate β. We could perform a regression of y_t on t

$$\hat{\beta} = \frac{\sum_{t=1}^{T} t y_t}{\sum_{t=1}^{T} t^2}.$$

We may also need to estimate σ^2 with something like

$$s^2 = \frac{1}{T-1} \sum_{t=1}^{T} \left(y_t - \hat{\beta} t \right)^2.$$

You can also consider higher order polynomial of time

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \ldots + \beta_r t^r + \epsilon_t.$$

3. Random Walk

$$y_t = y_{t-1} + \epsilon_t$$

where $\epsilon_t \sim iid(0, \sigma^2)$ (white noise).

Given the observed sample $\{ y_1, \ldots, y_T \}$, how will you estimate

$$y_{T+h}?$$

The estimated value is y_T.

This is the expected value of y_{T+h} given information at time period T. For this model we have

$$\begin{aligned} y_{T+h} &= y_{T+h-1} + \epsilon_{T+h} \\ &= y_{T+h-2} + \epsilon_{T+h-1} + \epsilon_{T+h} \\ &\vdots \\ &= y_T + \sum_{j=1}^{h} \epsilon_{T+j} \end{aligned}$$

so that $y_{T+h} \,|I_T \sim (y_T, h\sigma^2)$. The expected value given information at time period T is

$$E\left[y_{T+h} \,|I_T\right] = y_T.$$

If the variable is normally distributed, the 90% confidence interval will be

$$[y_T - 1.65\sigma\sqrt{h}, y_T + 1.65\sigma\sqrt{h}].$$

We may need to estimate σ^2 with something like

$$s^2 = \frac{1}{T-1}\sum_{t=2}^{T}\left(y_t - y_{t-1}\right)^2.$$

You can also consider a random walk with drift

$$y_t = \mu + y_{t-1} + \epsilon_t.$$

4. AR(1) – First order autoregression model.

$$y_t = \rho y_{t-1} + \epsilon_t \qquad\qquad |\rho| < 1$$

where $\epsilon_t \sim iid(0, \sigma^2)$ (white noise).

If $\rho = 1$ this would be the random walk model. For the AR(1) model the observed series is a fixed percent of the previous value with an added random shock.

Given the observed sample $\left\{\ y_1,\ \ \ldots,\ \ y_T\ \right\}$, how will you estimate

$$y_{T+h}?$$

For this model

$$\begin{aligned} y_{T+h} &= \rho y_{T+h-1} + \epsilon_{T+h} \\ &= \rho^2 y_{T+h-2} + \rho\epsilon_{T+h-1} + \epsilon_{T+h} \\ &\vdots \\ &= \rho^h y_T + \sum_{j=1}^{h} \rho^{h-j}\epsilon_{T+j} \end{aligned}$$

so that $y_{T+h}\,|I_T \sim \left(\rho^h y_T, \sigma^2\left(\frac{1-\rho^{2h}}{1-\rho^2}\right)\right)$. So the expected value given information at time period T is

$$E\left[y_{T+h}\,|I_T\right] = \rho^h y_T.$$

If the variable is normally distributed, the 90% confidence interval will be

$$\left[\rho^h y_T - 1.65\sigma\left(\frac{1-\rho^{2h}}{1-\rho^2}\right)^{1/2}, \rho^h y_T + 1.65\sigma\left(\frac{1-\rho^{2h}}{1-\rho^2}\right)^{1/2}\right].$$

We may need to estimate ρ. We could perform a regression of y_t on y_{t-1}

$$\hat{\rho} = \frac{\sum_{t=2}^{T} y_t y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2}.$$

We may also need to estimate σ^2 with something like

$$s^2 = \frac{1}{T-2}\sum_{t=2}^{T}\left(y_t - \hat{\rho} y_{t-1}\right)^2.$$

You can also consider the AR(p) model

$$\begin{aligned} y_t &= \rho_0 + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \ldots + \rho_p y_{t-p} + \epsilon_t \\ &= \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \epsilon_t. \end{aligned}$$

3.1 (optional) An Aside on the AR(1) forecast variance

$$\begin{aligned} y_{T+h} &= \rho^h y_T + \rho^{h-1}\epsilon_{T+1} + \rho^{h-2}\epsilon_{T-2} + \ldots + \rho\epsilon_{T+h-1} + \epsilon_{T+h} \\ &= \rho^h y_T + \sum_{s=0}^{h-1}\rho^s \epsilon_{T+h-s} \\ &= \rho^h y_T + \sum_{s=0}^{h-1}\rho^{h-1-s}\epsilon_{T+1+s} \end{aligned}$$

The ϵ_t are iid so the variance is just the sum of the individual variances

$$\begin{aligned} \mathrm{Var}\left(\sum_{s=0}^{h-1}\rho^s\epsilon_{T+h-s}\right) &= \sum_{s=0}^{h-1}\mathrm{Var}\left(\rho^s\epsilon_{T+h-s}\right) && (1) \\ &= \sum_{s=0}^{h-1}\rho^{2s}\sigma^2 && (2) \\ &= \sigma^2\sum_{s=0}^{h-1}\rho^{2s} && (3) \\ &= \sigma^2\frac{1-\rho^{2h}}{1-\rho^2} && (4) \end{aligned}$$

where we are summing a finite number of terms in a geometric series.

$$\begin{aligned} A &= 1 + \rho^2 + \rho^4 + \rho^6 + \ldots + \rho^{2(h-1)} && (5) \\ \Rightarrow \rho^2 A &= \rho^2 + \rho^4 + \rho^6 + \rho^8 + \ldots + \rho^{2h} && (6) \\ \Rightarrow A - \rho^2 A &= 1 - \rho^{2h} && (7) \\ \Rightarrow A &= \frac{1 - \rho^{2h}}{1 - \rho^2} && (8) \end{aligned}$$

Note that as $h \to \infty$, the variance goes to

$$\sigma^2 \frac{1}{1 - \rho^2}.$$

5. MA(1) – First order moving average model.

$$y_t = \epsilon_t + \theta \epsilon_{t-1} \qquad |\theta| < 1$$

where $\epsilon_t \sim iid(0, \sigma^2)$ (white noise).

The observed series is a weighted average of the previous shock and a new shock.

Given the observed sample $\left\{ \ y_1, \ \ldots, \ y_T \ \right\}$, how will you estimate

$$y_{T+h}?$$

Results similar to those derived above are possible. However, the algebra gets complicated so we will let the computer provide the answers numerically.

You can also consider the MA(q) model

$$y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q}.$$

4 Model Selection

Almost any model can be estimated for a given data set. How do we select between the different models?

The model represents y_t in terms of lagged values and other variables.

Let $\hat{y}_t$ denote the value implied (estimated value or fitted value) by the model and denote the residuals

$$\hat{e}_t = y_t - \hat{y}_t.$$

One measure of the model's "fit" is the mean squared error (MSE)

$$MSE = \frac{\sum_{t=1}^{T} \hat{e}_t^2}{T}.$$

The mean of the squared errors implied by the model. Sometimes the square root of this measure is reported. The root mean squared error ($RMSE$) is simply

$$RMSE = \sqrt{MSE}.$$

You have seen this before in your linear regression class. You selected the model's parameters to minimize the residual sum of squares (RSS)

$$RSS = T \cdot MSE.$$

You judged the fit of your regression by its R^2

$$R^2 = 1 - \frac{\sum_{t=1}^{T} \hat{e}_t^2}{\sum_{t=1}^{T} (y_t - \bar{y})^2} = 1 - \frac{MSE}{\frac{1}{T}\sum_{t=1}^{T} (y_t - \bar{y})^2}$$

where $\bar{y}$ is the sample mean. So selecting between models based on the highest R^2 is equal to selecting between models based on the lowest MSE.

Recall this is generally a bad idea.

As you add variables your R^2 never falls. You will get a better "fit" to your data but your forecasts may actually be worse. This means that you have *fit your data but not your model.*

When this occurs it is called in-sample overfitting. It is characterized by high R^2, low MSE and poor forecasts.

The problem is that the reduction in MSE is not sufficient to justify the estimated parameter (losing a degree of freedom).

To adjust for this we can use a better measure by doing a degrees of freedom correction and measure the model's "fit" with

$$s^2 = \frac{\sum_{t=1}^{T} \hat{e}_t^2}{T - k}$$

where k is the number of parameters estimated.

In a regression model with normally distributed errors, s^2 is an unbiased estimate of the regression disturbance variance.

In linear regression the related measure is the **adjusted** R^2.

$$\overline{R^2} = 1 - \frac{\frac{1}{T-k}\sum_{t=1}^{T} \hat{e}_t^2}{\frac{1}{T-1}\sum_{t=1}^{T} (y_t - \bar{y})^2} = 1 - \frac{s^2}{\frac{1}{T-1}\sum_{t=1}^{T} (y_t - \bar{y})^2}.$$

Selecting a model based on the highest $\overline{R^2}$ is the same as using the lowest s^2.

Note that another way to think of this is as a penalty weight on the MSE.

$$s^2 = \left(\frac{T}{T-k}\right) MSE.$$

Estimating a new parameter will likely lower the MSE but will increase $\left(\frac{T}{T-k}\right)$. Hence s^2 may increase or decrease.

There are two widely used model selection criteria.

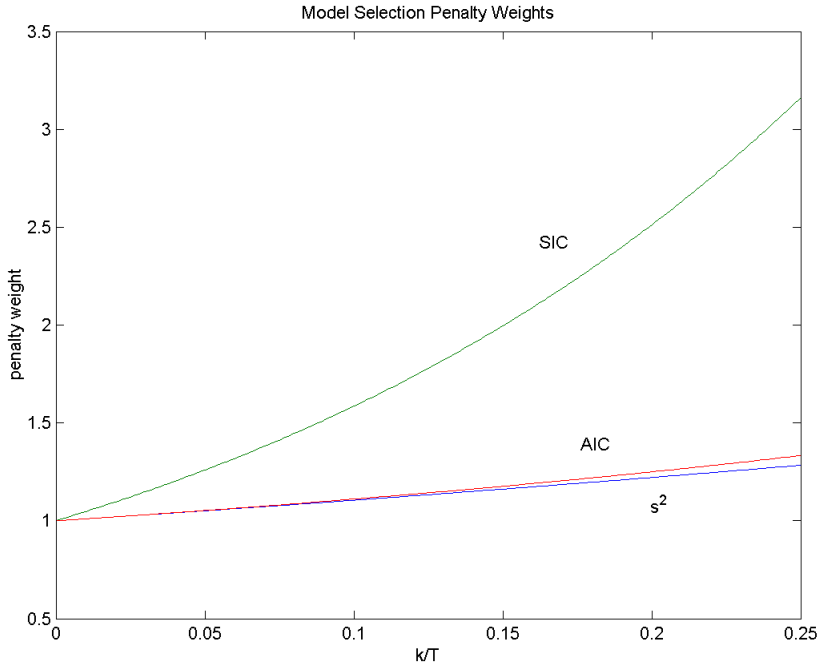
1. The Akaike information criteria (AIC)

$$AIC = \exp\left(\frac{2k}{T}\right)\frac{\sum_{t=1}^{T}\hat{e}_t^2}{T} = \exp\left(\frac{2k}{T}\right)MSE.$$

2. The Schwarz information criteria (SIC) or Bayesian information criteria (BIC). Note that SAS reports this as the SBC.

$$SIC = T^{k/T}\frac{\sum_{t=1}^{T}\hat{e}_t^2}{T} = \left(T^{k/T}\right)MSE.$$

Just like s^2, these can be thought of as penalty weights for the MSE to account for additional parameters being estimated.



4.1 Which to use?

AIC versus SIC (you never use s^2).

AIC selects a more highly parameterized model.

SIC selects a model with fewer parameters.

For forecasts SIC generally gives more accurate confidence intervals. But it can sometimes overstate the level of uncertainty, i.e. the confidence intervals will be too large.

For model fitting to understand the dynamics of a system the AIC is better.

You should generally report both, but for forecasting you should use the SIC (SBC in SAS).

4.2 SAS and AIC and SBC

SAS reports the natural log of the measures that we previously defined. Because the natural log is a monotonically increasing function, the best model will still be selected. However, the values of the measure will be different and can sometimes be negative.

In SAS

$$AIC = \frac{2k}{T} + \ln(MSE)$$

and

$$SBC = \frac{k}{T}\ln(T) + \ln(MSE).$$

You still select the model with the lowest value (i.e., closest to negative infinity).

5 The Lag Operator

Time series are random variables that are indexed by time

$$\ldots, y_{-3}, y_{-2}, y_{-1}, y_0, y_1, y_2, y_3, \ldots.$$

We often want to map from one series to another, for example consider the linear combination

$$y_t = \alpha x_t + \beta z_t + \epsilon_t.$$

This can be more complicated due to the time dimension. We will need to consider linear combinations across the time dimension, for example

$$y_t = \alpha x_{t-2} + \beta z_{t-3} + \epsilon_t.$$

We need some notation to keep track of these linear combinations across time.

The lag operator allows this. The lag operator shifts a series back one time period

$$Ly_t = y_{t-1}.$$

If we need a two period lag we can write it as

$$L^2 y_t = LLy_t = y_{t-2}.$$

In general, to lag h periods we write

$$L^h y_t = y_{t-h}.$$

L^{-1} does the inverse operation.

$$L^{-1} y_t = y_{t+1}.$$

Some examples

1. AR(1)

$$\begin{aligned} y_t &= \rho y_{t-1} + \epsilon_t \\ y_t &= \rho L y_t + \epsilon_t \\ y_t - \rho L y_t &= \epsilon_t \\ (1 - \rho L) y_t &= \epsilon_t \end{aligned}$$

2. AR(p)

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \epsilon_t \\ y_t &= \phi_1 L y_t + \phi_2 L^2 y_t + \ldots + \phi_p L^p y_t + \epsilon_t \\ y_t - \phi_1 L y_t - \phi_2 L^2 y_t - \ldots - \phi_p L^p y_t &= \epsilon_t \\ (1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p) y_t &= \epsilon_t \\ \Phi(L) y_t &= \epsilon_t \end{aligned}$$

where $\Phi(L)$ is a p^{th} ordered polynomial in the lag operator.

3. MA(q)

$$\begin{aligned} y_t &= \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q} \\ y_t &= \epsilon_t + \theta_1 L \epsilon_t + \theta_2 L^2 \epsilon_t + \ldots + \theta_q L^q \epsilon_t \\ y_t &= (1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q) \epsilon_t \\ y_t &= \Theta(L) \epsilon_t \end{aligned}$$

where $\Theta(L)$ is a q^{th} ordered polynomial in the lag operator.

We can use the lag operator to simplify some of our calculations that move between the time dimensions.

6 An AR(1) as an MA(∞)

First we can get the result by repeatedly substituting the model into itself.

$$\begin{aligned}
y_t = \rho y_{t-1} + \epsilon_t \qquad & |\rho| < 1 \\
y_{t-1} = \rho y_{t-2} + \epsilon_{t-1} \qquad & y_t = \rho^2 y_{t-2} + \rho\epsilon_{t-1} + \epsilon_t \\
y_{t-2} = \rho y_{t-3} + \epsilon_{t-2} \qquad & y_t = \rho^3 y_{t-3} + \rho^2\epsilon_{t-2} + \rho\epsilon_{t-1} + \epsilon_t \\
\vdots \qquad & \vdots \\
& y_t = \rho^\infty y_{t-\infty} + \sum_{s=0}^{\infty} \rho^s \epsilon_{t-s} \\
\Rightarrow \qquad & y_t = \sum_{s=0}^{\infty} \rho^s \epsilon_{t-s}.
\end{aligned}$$

Now the same result can be achieved by using lag operators.

$$\begin{aligned}
y_t &= \rho y_{t-1} + \epsilon_t \\
y_t &= \rho L y_t + \epsilon_t \\
y_t - \rho L y_t &= \epsilon_t \\
(1 - \rho L) y_t &= \epsilon_t \\
y_t &= \frac{1}{(1-\rho L)} \epsilon_t \\
&= \sum_{s=0}^{\infty} (\rho L)^s \epsilon_t \\
&= \sum_{s=0}^{\infty} \rho^s L^s \epsilon_t \\
&= \sum_{s=0}^{\infty} \rho^s \epsilon_{t-s}
\end{aligned}$$

An AR(1) series is a linear combination of an iid series

$$y_t = B(L)\epsilon_t = \sum_{s=0}^{\infty} b_s \epsilon_{t-s}$$

where $b_s = \rho^s$.

This MA(∞) model has an infinity number of terms but they all depend on only one parameter, ρ.

7 Stationarity

For forecasting we will need to restrict attention to a set of models that only change in ways that will allow statistical analysis.

If we do not have a series in this set, we transform it to get it into the set. Make our forecasts on the transformed series. Finally, reverse the transformation to get forecasts in terms of the original series.

7.1 Weak stationarity

(covariance stationarity or second order stationarity)

The first two moments of the joint distribution do not change with time shifts.

For the first moment this means $Ey_t = \mu_y$ for all t. In English this means that your series does not have any trend or predictable cycles.

For the second moment this means

$$E\left[(y_{t-k} - \mu_y)(y_t - \mu_y)\right] = \gamma(k).$$

does not depend on t.

The covariance is

$$E\left[(y_{t-k} - \mu_y)(y_t - \mu_y)\right] = \gamma(t, k).$$

In general, the covariance term depends on both t and k. For a series to be weakly stationary the covariance needs to depend only on k and not on t

$$E\left[(y_{t-k} - \mu_y)(y_t - \mu_y)\right] = \gamma(k).$$

This is called the autocovariance function and shows the covariance between terms that are separated by k time periods.

8 Wold Theorem

Let y_t be any zero-mean covariance-stationary (weakly stationary) process. Then it can be written as

$$y_t = B(L)\epsilon_t = \sum_{s=0}^{\infty} b_s \epsilon_{t-s}$$

where $b_0 = 1$, $\sum_{s=0}^{\infty} b_s^2 < \infty$ and $\epsilon_t \sim iid(0, \sigma^2)$.

This is called the Wold decomposition of a time series.

This means that once we remove the mean, any trend and any deterministic cycles, we can restrict attention to linear combinations of an iid process.

1. Must remove any trend or cycle.

2. The linear combinations may have an infinite number of terms.

3. The term $\sum_{s=0}^{\infty} b_s^2$ is just the variance of y_t.

The natural question is how do we deal with an infinite number of coefficients?

The answer is that we use a finite number of parameters to represent the infinite number of coefficients.

Just as it was shown in the AR(1) model.

8.1 ARMA models as the Wold decomposition

We have $y_t = B(L)\epsilon_t$.

Consider parameterizing

$$B(L) = \sum_{s=0}^{\infty} b_s L^s$$

with the ratio of finite ordered polynomial.

$$B(L) = \frac{\Theta(L)}{\Phi(L)} = \frac{1 + \theta_1 L + \ldots + \theta_q L^q}{1 - \phi_1 L - \ldots - \phi_p L^p} = \sum_{s=0}^{\infty} b_s L^s.$$

The infinite number of coefficients b_s for $s = 1, 2, \ldots$ are all written as functions of the finite number of parameters $\phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q$.

This would imply that we could write the model for y_t as

$$\Phi(L) y_t = \Theta(L)\epsilon_t.$$

This is called an ARMA(p, q) model and denoted $y_t \sim$ ARMA(p, q).

This raises some obvious questions.

- How do we remove the mean, trend and any cycles?
- How do we select p and q?
- How do we estimate the parameters $\phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q$?

We start to consider these questions next week.

Let's estimate some basic ARMA models.