Autoregressive Moving Average (ARMA) Models

February 4, 2019

The material in this set of notes is based on S&S Chapter 3. We're finally going to define our first time series model! ☺ The first time series model we will define is the **autoregressive (AR)** model. We will then consider a different simple time series model, the **moving average (MA)** model. Putting both models together to create one more general model will give us the **autoregressive moving average** (ARMA) model.

The Autoregressive (AR) Model

The first kind of time series model we'll consider is an **autoregressive (AR)** model. This is one of the most intuitive models we'll consider. The basic idea is that we will model the response at time t x_t as a linear function of its p previous values and some independent random noise, e.g.

$$x_t = 0.5x_{t-1} + w_t, \tag{1}$$

where x_t is stationary and $w_t \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_w^2\right)$. This kind of model is especially well suited to forecasting, as

$$\mathbb{E}\left[x_{t+1}|x_t\right] = 0.5x_{t-1}. \tag{2}$$

We explicitly define an **autoregressive model of order** p, abbreviated as $\mathbf{AR}(p)$ as:

$$(x_t - \mu_x) = \phi_1 (x_{t-1} - \mu_x) + \phi_2 (x_{t-2} - \mu_x) + \cdots + \phi_p (x_{t-p} - \mu_x) + w_t, \tag{3}$$

where x_t is stationary with mean $\mathbb{E}[x_t] = \mu_x$ and $w_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2)$. For convenience:

- We'll often assume $\mu_x = 0$, so

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t. \tag{4}$$

- We'll introduce the **autoregressive operator** notation:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p, \tag{5}$$

  where $B^p x_t = x_{t-p}$ is the **backshift operator**. This allows us to rewrite (3) and (4) more concisely as $\phi(B)(x_t - \mu_x) = w_t$ and

$$\phi(B)(x_t) = w_t, \tag{6}$$

  respectively.

An $\mathbf{AR}(p)$ model looks like a linear regression model, but the covariates are also random variables. We'll start building an understanding of the $\mathbf{AR}(p)$ model by starting with the simpler special case where $p = 1$.

The $\mathbf{AR}(1)$ model with $\mu_x = 0$ is a special case of (3)

$$x_t = \phi_1 x_{t-1} + w_t. \tag{7}$$

A natural thing to do is to try to rewrite x_t as a function of ϕ_1 and the previous values of the random errors. Then (7) will look more like a classical regression problem, as it will no longer have random variables as as covariates. Furthermore, if we could rewrite x_t as a function of ϕ_1 and the random errors $\boldsymbol{w}$, then x_t would be a **linear process**.

A **linear process** x_t is defined to be a linear combination of white noise variates w_t and is given by

$$x_t = \mu_x + \sum_{j=-\infty}^{\infty} \psi_j w_{t-j},$$

where the coefficients satisfy $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, w_t are independent and identically distributed with mean 0 and variance σ_w^2, and $\mu_x = \mathbb{E}[x_t]$. Importantly, it can be shown that the autocovariance function of a linear process is

$$\gamma_x(h) = \sigma_w^2 \sum_{j=-\infty}^{\infty} \psi_{j+h}\psi_j, \tag{8}$$

for $h \geq 0$, recalling that $\gamma_x(h) = \gamma_x(-h)$. This means that once we know the linear process representation of any time series process, we can easily compute its autocovariance (and autocorrelation) functions. We will often use the **infinite moving average operator** shorthand $1 + \psi_1 B + \psi_2 B^2 + \dots \psi_j B^j + \dots = \psi(B)$.

We can start rewriting x_t as follows:

$$\begin{aligned} x_t &= \phi_1^2 x_{t-1} + \phi_1 w_{t-1} + w_t \\ &= \phi_1^3 x_{t-2} + \phi_1^2 w_{t-2} + \phi_1 w_{t-1} + w_t \\ &= \underbrace{\phi_1^k x_{t-k}}_{(*)} + \sum_{j=0}^{k-1} \phi_1^j w_{t-j}. \end{aligned}$$

We can see that we can almost lagged values of $\boldsymbol{x}$ out of the right hand side. Fortunately, when $|\phi_1| < 1$, then

$$\lim_{k\to\infty} \mathbb{E}\left[\left(x_t - \sum_{j=0}^{k-1} \phi_1^j w_{t-j}\right)^2\right] = \lim_{k\to\infty} \phi^{2k} \mathbb{E}\left[x_{t-k}^2\right] = 0,$$

because $\mathbb{E}\left[x_{t-k}^2\right]$ is constant as long as x_t is stationary is assumed. This means that when $|\phi_1| < 1$, then we can write elements of the response x_t as a linear function the previous values of the random errors:

$$x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j}. \tag{9}$$

(9) is the **linear process** representation of an $\mathbf{AR}(1)$ model. It follows that the autocovariance function

$$\begin{aligned}\gamma_x(h) &= \sigma_w^2 \sum_{j=0}^{\infty} \phi^{j+h} \phi^j \\ &= \sigma_w^2 \phi^h \sum_{j=0}^{\infty} \phi^{2j} \\ &= \sigma_w^2 \phi^h \left(\frac{1}{1-\phi^2} \right).\end{aligned} \tag{10}$$

and the autocorrelation function is

$$\rho_x(h) = \phi^h. \tag{11}$$

Note that this is not the only way to compute the values of autocovariance function. We could compute them directly from (4),

$$\begin{aligned}\gamma_x(h) =& \mathbb{E}\left[x_{t-h} x_t\right] \qquad (12) \\ =& \mathbb{E}\left[x_{t-h}\left(\phi_1 x_{t-1} + w_t\right)\right] \\ =& \phi_1 \mathbb{E}\left[x_{t-1-(h-1)} x_{t-1}\right] + \mathbb{E}\left[x_{t-h} w_t\right] \\ =& \phi_1 \gamma_x(h-1).\end{aligned}$$

This gives us a recursive relation that we can use to compute the autocovariance function

$\gamma_x(h)$, starting from $\gamma_x(0)$. We can compute $\gamma_x(0)$ using substitution:

$$\begin{aligned}
\gamma_x(0) =& \mathbb{E}\left[x_t^2\right] && (13)\\
=& \mathbb{E}\left[x_t^2\right]\\
=& \mathbb{E}\left[(\phi_1 x_{t-1} + w_t)^2\right]\\
=& \mathbb{E}\left[\phi_1^2 x_{t-1}^2 + 2\phi_1 w_t x_{t-1} + w_t^2\right]\\
=& \phi_1^2 \mathbb{E}\left[x_{t-1}^2\right] + \sigma_w^2\\
=& \sigma_w^2 \sum_{j=0}^{\infty} \phi_1^{2j} && \text{(follows from continued substitution)}\\
=& \frac{\sigma_w^2}{1-\phi_1^2}, && \text{if } |\phi_1| < 1,\ \gamma_x(0) = \infty \text{ otherwise!}
\end{aligned}$$

Whether or not $|\phi_1| < 1$ is closely related to whether or not $\boldsymbol{x}$ is a stationary time series. If $|\phi_1| < 1$, then it is easy to see that the $\mathbf{AR}(1)$ time series $\boldsymbol{x}$ is stationary because the mean of each x_t is zero and the autocovariance function $\gamma_x(h) = \sigma_w^2 \phi_h \left(\frac{1}{1-\phi^2}\right)$ depends only on the lag, h. What happens when $\phi_1 > 1$?

When $\phi_1 < 1$, we can revisit (7) and note that $x_{t+1} = \phi_1 x_t + w_{t+1}$. Rearranging gives

$$x_t = \left(\frac{1}{\phi_1}\right) x_{t+1} - \left(\frac{1}{\phi_1}\right) w_{t+1}.$$

If $\phi > 1$, then $\left(\frac{1}{\phi_1}\right) < 1$ and we can use the same approach we used previously to write

$$x_t = -\sum_{j=1}^{\infty} \left(\frac{1}{\phi_1}\right)^j w_{t+j}.$$

The problem, however, is that this requires that x_t is a function of *future* values, which may not be known at time t. We call such a time series **non-causal**. Using a **non-causal** model will rarely make sense in practice, and furthermore makes forecasting impossible - in the future, whenever we talk about $\mathbf{AR}(p)$ models we restrict our attention to **causal** models.

Understanding when a $\mathbf{AR}(p)$ model is causal is more difficult than understanding when an $\mathbf{AR}(1)$ model is causal. We figured out when an $\mathbf{AR}(1)$ model is causal by finding the coefficients $\psi_{-\infty}, \ldots, \psi_{\infty}$ of its linear process representation as a function of the AR coefficient ϕ_1, and showing that all of the coefficients $\psi_{-\infty}, \ldots, \psi_{-1}$ for future errors are exactly equal to zero.

The linear process representation is especially useful for an $\mathbf{AR}(p)$ time series when $p > 1$, because computing the autocovariance function $\gamma_x(h)$ directly as we did in (12) and (13) gets much more cumbersome when $p > 1$. We can see this in the $\mathbf{AR}(2)$ case, where we have

$$x_t = \phi_2 x_{t-2} + \phi_1 x_{t-1} + w_t. \tag{14}$$

We can get a recursive relation for the autocovariance function $\gamma_x(h)$ starting from $\gamma_x(0)$ and $\gamma_x(1)$ as follows:

$$\begin{aligned}
\gamma_x(h) =& \mathbb{E}\left[x_{t-h} x_t\right] \\
=& \mathbb{E}\left[x_{t-h}\left(\phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t\right)\right] \\
=& \phi_1 \mathbb{E}\left[x_{t-1-(h-1)} x_{t-1}\right] + \phi_2 \mathbb{E}\left[x_{t-2-(h-2)} x_{t-2}\right] + \mathbb{E}\left[x_{t-h} w_t\right] \\
=& \phi_1 \gamma_x(h-1) + \phi_2 \gamma_x(h-2).
\end{aligned}$$

We can try to compute $\gamma_x(0)$ and $\gamma_x(1)$ using substitution:

$$\begin{aligned}
\gamma_x(0) =& \mathbb{E}\left[x_t^2\right] \\
=& \mathbb{E}\left[\left(\phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t\right)^2\right] \\
=& \mathbb{E}\left[\phi_1^2 x_{t-1}^2 + \phi_2^2 x_{t-2}^2 + 2\phi_1\phi_2 x_{t-1} x_{t-2} + 2\phi_1 x_{t-1} w_t + 2\phi_2 x_{t-2} w_t + w_t^2\right] \\
=& \mathbb{E}\left[\phi_1^2 x_{t-1}^2 + \phi_2^2 x_{t-2}^2 + 2\phi_1\phi_2 x_{t-1} x_{t-2}\right] + \sigma_w^2.
\end{aligned}$$

However, this gets *very* complicated, even though we only have two lags!

Unfortunately, it's much harder to find the linear process representation of an $\mathbf{AR}(p)$ time series by simple substitution as we did with an $\mathbf{AR}(1)$ time series. Substituting according to (14)

$$
\begin{aligned}
x_t =& \phi_1\phi_2 x_{t-3} + \left(\phi_2 + \phi_1^2\right) x_{t-2} + \phi_1 w_{t-1} + w_t \\
=& \left(\phi_2 + \phi_1^2\right)\phi_2 x_{t-4} + \phi_1\left(2\phi_2 + \phi_1^2\right) x_{t-3} + \left(\phi_2 + \phi_1^2\right) w_{t-2} + \phi_1 w_{t-1} + w_t \\
=& \left(\phi_2 + \phi_1^2\right)\phi_2 x_{t-4} + \phi_1\left(2\phi_2 + \phi_1^2\right)\left(\phi_2 x_{t-5} + \phi_1 x_{t-4} + w_{t-3}\right) + \left(\phi_2 + \phi_1^2\right) w_{t-2} + \phi_1 w_{t-1} + w_t \\
=& \phi_1\phi_2\left(2\phi_2 + \phi_1^2\right) x_{t-5} + \left(\phi_2^2 + \phi_1^2\phi_2 + 2\phi_1\phi_2^2 + \phi_1^3\phi_2\right) x_{t-4} + \\
& \phi_1\left(2\phi_2 + \phi_1^2\right) w_{t-3} + \left(\phi_2 + \phi_1^2\right) w_{t-2} + \phi_1 w_{t-1} + w_t \dots
\end{aligned}
$$

Again, this is *not* working out nicely!

Instead, we can find the values of $\psi_1, \dots, \psi_\infty$ that satisfy $\phi(B)\psi(B)w_t = w_t$, which follows from substituting $x_t = \sum_{j=-\infty}^{\infty}\psi_j w_{t-j}$ into (19). This is equivalent to finding the inverse function $\phi^{-1}(B)$ that satisfies $\phi(B)\phi^{-1}(B)w_t = w_t$.

We can see how this method for finding the values of $\psi_1, \dots, \psi_\infty$ works by returning to the $\mathbf{AR}(1)$ case. The values $\psi_1, \dots, \psi_\infty$ that satisfy $\phi(B)\psi(B)w_t = w_t$ solve:

$$
1 + (\psi_1 - \phi_1) B + (\psi_2 - \psi_1\phi_1) B^2 + \dots + \psi_j B^j + \dots = 1, \tag{15}
$$

where (15) follows from expanding $\phi(B)$ and $\psi(B)$. This allows us to recover the linear process representation of the $\mathbf{AR}(1)$ process in a different way, as (15) holds if all of the coefficients for B^j with $j > 0$ are equal to zero, i.e. $\psi_k - \psi_{k-1}\phi_1 = 0$ for $k > 1$.

Now let's try this approach for the $\mathbf{AR}(2)$ case. We have

$$
\begin{aligned}
1 =& \left(1 - \phi_1 B - \phi_2 B^2\right)\left(1 + \psi_1 B + \psi_2 B^2 + \dots + \psi_j B^j + \dots\right) \\
=& 1 + (\psi_1 - \phi_1) B + (\psi_2 - \phi_2 - \phi_1\psi_1) B^2 + (\psi_3 - \phi_1\psi_2 - \phi_2\psi_1) B^3 + \dots + \\
& (\psi_j - \phi_1\psi_{j-1} - \phi_2\psi_{j-2}) B^j + \dots
\end{aligned}
$$

We see that we can compute the values of $\psi_1, \ldots, \psi_\infty$ recursively,

$$\psi_1 = \phi_1$$

$$\psi_2 = \phi_2 + \phi_1^2$$

$$\psi_3 = 2\phi_1 \left(\phi_2 + \phi_1^2\right),$$

and so on.

It's also very tricky to figure out when **AR** (p) model is **causal** for $p > 1$ - we'll come back to this later.

The Moving Average (MA) Model

Instead of assuming that elements of a time series x_t are linear function of previous elements of the time series $x_1, \ldots, x_{t-1}$ and independent, identically distributed noise w_t, we might assume that elements of a time series x_t are a linear function of all of the current and previous noise variates, $w_1, \ldots, w_{t-1}$. The latter gives us the **moving average model of order** q, abbreviated as **MA** (q). The **MA** (q) model is explicitly defined as

$$x_t - \mu_x = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q}, \tag{16}$$

where $\mathbb{E}\left[x_t\right] = \mu_x$ and $w_t \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_w^2\right)$. For convenience:

- We'll often assume $\mu_x = 0$, so

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q}. \tag{17}$$

- We'll introduce the **moving average operator** notation:

$$\theta\left(B\right) = 1 + \phi_1 B + \phi_2 B^2 + \cdots + \phi_p B^p, \tag{18}$$

which allows us to rewrite (16) and (17) more concisely as $x_t - \mu_x = \theta(B) w_t$ and

$$x_t = \theta(B) w_t, \tag{19}$$

respectively.

Again, the $\mathbf{MA}(q)$ model looks like a linear regression model. Importantly, the $\mathbf{MA}(q)$ model is stationary for any values of the parameters $\theta_1, \ldots, \theta_q$.

Like we did with the $\mathbf{AR}(p)$ model, we'll start building an understanding of the $\mathbf{MA}(q)$ by starting with the simpler special case where $q = 1$,

$$x_t = \theta_1 w_{t-1} + w_t. \tag{20}$$

It is easy to see that this $\mathbf{MA}(q)$ time series is mean zero. We can compute the autocovariance function as follows:

$$\begin{aligned}
\gamma_x(h) &= \mathbb{E}[x_t x_{t-h}] \\
&= \mathbb{E}[(\theta_1 w_{t-1} + w_t)(\theta_1 w_{t-h-1} + w_{t-h})] \\
&= \mathbb{E}\left[\theta_1^2 w_{t-1} w_{t-h-1} + \theta_1 w_t w_{t-h-1} + \theta_1 w_{t-1} w_{t-h} + w_t w_{t-h}\right] \\
&= \mathbb{E}\left[\theta_1^2 w_{t-1} w_{t-h-1} + \theta_1 w_{t-1} w_{t-h} + w_t w_{t-h}\right] \\
&= \begin{cases} \sigma_w^2(\theta_1^2 + 1) & h = 0 \\ \theta_1 & h = 1 \\ 0 & h > 1 \end{cases}.
\end{aligned} \tag{21}$$

The corresponding autocorrelation function is

$$\rho_x(h) = \begin{cases} \frac{\theta_1}{\theta_1^2 + 1} & h = 1 \\ 0 & h > 1 \end{cases}. \tag{22}$$

The autocovariance and autocorrelation functions of the $\mathbf{MA}(q)$ model are noteworthy in two ways:

(•) The autocorrelation function $\rho_x(h)$ is bounded, $\rho_x(h) \leq 1/2$ for $h > 1$.

(*) The parameters of the **MA** (q) model do not uniquely determine the autocovariance and autocorrelation function values. θ_1 and σ_w^2 do not uniquely determine the value of the autocovariance function $\gamma_x(h)$, and θ_1 does not determine the value of the autocorrelation function.

It is easiest to understand (*) via an example.