AUTOREGRESSION AND MOVING AVERAGE MODELS

LECTURE OUTLINE

Last time we studied the following topics:

- Correlation and Autocorrelation
- Feature Engineering
- Feature Selection

In this lecture we will:

- Autoregressive Models
- Moving Average Models

Autoregressive Models vs Linear Regression

Definition: This time series technique assumes that future observations at next time stamp are related to the observations at prior time stamps through a linear relationship

$$y_t = b_0 + b_1 y_{t-1} + \epsilon_t$$

That looks a lot like linear regression, so what's new ?

→ In linear regression, data is supposed i.i.d

→ The models we will study are related to linear regression but account for the correlations that arise between data points in the same time series

Why not linear regression?

- In linear regression, data is supposed i.i.d
  → not the case for time series data: points near in time tend to be strongly correlated with one another.

Does this rule out linear regression?

Linear regression can be applied to time series data provided the following conditions hold:

1. **Assumptions with respect to the behavior of the time series:**
   - The time series has a linear response to its predictors.
   - No input variable is constant over time or perfectly correlated with another input variable.

2. **Assumptions with respect to the error:**
   - For each point in time, the expected value of the error, given all explanatory variables for all time periods (forward and backward), is 0.
   - The error at any given time period is uncorrelated with the inputs at any time period in the past or future → a plot of the autocorrelation function of the errorswill not indicate any pattern.
   - Variance of the error is independent of time.

If these assumptions hold, then ordinary least squares regression is an unbiased estimator of the coefficients given the inputs, even for time series data.

Don’t force linear regression

Some of the consequences of applying linear regression when your data doesn’t meet the required assumptions are:

- Your coefficients will not minimize the error of your model.
- Your p-values for determining whether your coefficients are nonzero will be incorrect because they rely on assumptions that are not met. This means your assessments of coefficient significance could be wrong.

⚠ Linear regressions can be helpful in offering simplicity and transparency when appropriate, but an incorrect model certainly isn’t transparent!

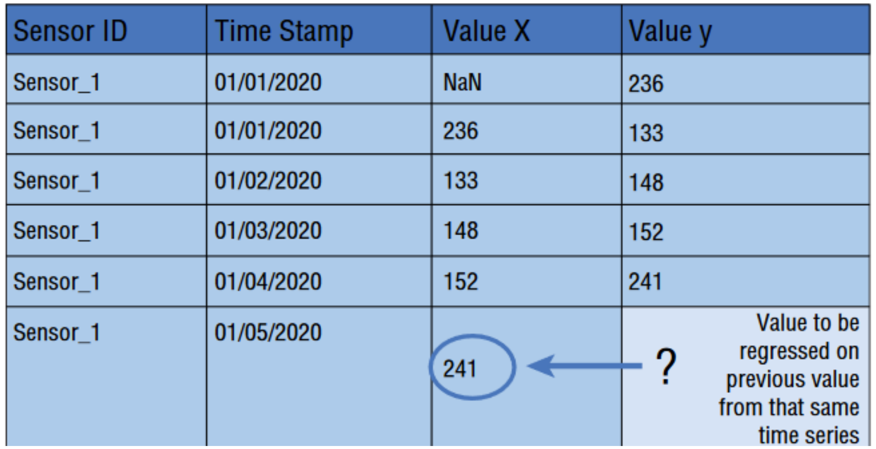
Autoregression

Definition: time series forecasting approach that depends only on the previous outputs of a time series.
→ assumes that future observations at next time stamp are related to the observations at prior time stamps through a linear relationship.

$$y_t = b_0 + b_1 y_{t-1} + \epsilon_t$$

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

- the output value in the previous time stamp becomes the input value to predict the next time stamp value,
- the errors follow theusual assumptions about errors in a simple linear regression model
- the number of preceding input values in the time series that are used to predict next time stamp value is called order (noted p)
- More generally, an nth-order autoregression is a multiple linear regression in which the value of the series at any time t is a linear function of the previous values in that same time series.



Sensor ID	Time Stamp	Value X	Value y
Sensor_1	01/01/2020	NaN	236
Sensor_1	01/01/2020	236	133
Sensor_1	01/02/2020	133	148
Sensor_1	01/03/2020	148	152
Sensor_1	01/04/2020	152	241
Sensor_1	01/05/2020	241	?

Value to be regressed on previous value from that same time series

Stationarity conditions for AR(1): Mean

$$y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t$$

We assume the process is stationary and then work “backward” to see what that implies about the coefficients.

1. the expected value of the process must be the same at all times.

$$\mathbb{E}[y_t] = \mu = \mathbb{E}[y_{t-1}]$$

$$\mathbb{E}[y_t] = \mathbb{E}[\phi_0 + \phi_1 \cdot y_{t-1} + \underbrace{\epsilon_t}_{0}] = \mu$$

$$\Downarrow$$

$$\phi_0 + \phi_1 \cdot \mu = \mu$$

So, if the process is stationary we should have: $\mu = \dfrac{\phi_0}{1 - \phi_1}$

Stationarity conditions for AR(1): Variance

We can do the same to find conditions on the variance to ensure stationarity.

We have: $\phi_0 = \mu(1 - \phi_1)$

$$y_t = \phi_0 + \phi_1 \times y_{t-1} + e_t$$
$$y_t = (\mu - \mu \times \phi_1) + \phi_1 \times y_{t-1} + e_t$$
$$y_t - \mu = \phi_1(y_{t-1} - \mu) + e_t$$

Let's try to compute the variance:

$$Var(y_t - \mu) = Var(\phi_1(y_{t-1} - \mu) + e_t)$$
$$Var(y_t) = Var(\phi_1(y_{t-1} - \mu) + Var(e_t) + 2Cov(\phi_1(y_{t-1} - \mu), e_t)$$
$$Var(y_t) = \phi_1^2 Var(y_{t-1}) + Var(e_t) + 2\phi_1 \cdot Cov(y_{t-1} - \mu, e_t)$$

We will now show that $Cov(y_{t-1} - \mu, e_t) = 0$

Stationarity conditions for AR(1): Variance (1/2)

We will now show that: $Cov(y_{t-1} - \mu, e_t) = 0$

We first recall that: $y_t - \mu = \phi_1(y_{t-1} - \mu) + e_t$

Given that this time series is stationary: $y_{t-1} - \mu = \phi_1(y_{t-2} - \mu) + e_{t-1}$

$$y_t - \mu = \phi_1(\phi_1(y_{t-2} - \mu) + e_{t-1}) + e_t$$

Rearranging:

$$y_t - \mu = e_t + \phi_1(e_{t-1} + \phi_1(e_{t-2} + \phi_1(y_{t-3} - \mu)))$$

$$y_t - \mu = e_t + \phi \times e_{t-1} + \phi^2 \times e_{t-2} + \phi^3 \times e_{t-3} + \dots$$

Therefore:

$$y_t - \mu = \sum_{i=1}^{\infty} \phi_1^i \times e_{t-i}$$

y_t minus the process mean is a linear function of the error terms

Stationarity conditions for AR(1): Variance (2/2)

$$y_t - \mu = \sum_{i=1}^{\infty} \phi_1^i \times e_{t-i} \qquad Var(y_t) = \phi_1^2 Var(y_{t-1}) + Var(e_t) + 2\phi_1 \cdot Cov(y_{t-1} - \mu, e_t)$$

We recall that $Cov(X;Y) = \mathbb{E}(XY) – E(X)E(Y)$

$$Cov(y_{t-1} - \mu, e_t) = \mathbb{E}[(y_{t-1} - \mu)e_t] - \mathbb{E}(y_{t-1} - \mu)\mathbb{E}(e_t)$$

$$\mathbb{E}[(y_{t-1} - \mu)e_t] = \mathbb{E}[\sum_{i=1}^{\infty} \phi_1^i \times e_{t-i} \times e_t]$$

Given that the values of e_t at different t values are independent, we have: $\mathbb{E}[(y_{t-1} - \mu)e_t] = 0$

Therefore: $Cov(y_{t-1} - \mu, e_t) = 0$

Hence: $Var(y_t) = \phi_1^2 Var(y_{t-1}) + Var(e_t)$

We assumed that the process was stationary so: $Var(y_t) = Var(y_{t-1})$

Finally: $\boxed{Var(y_t) = \frac{Var(e_t)}{1 - \phi_1^2}}$

- ϕ_1^2must be less than 1
- for a stationary process we must have $-1 < \phi < 1$

→ necessary and sufficient condition for weak stationarity.

Weak stationarity

Definition: Weak stationarity requires only that the mean and variance of a process be time invariant.

Reminder: Strong stationarity requires that the distribution of the random variables output by a process remain the same over time
→ for example, it demands that the statistical distribution of y_1, y_2, y_3 be the same as $y_{101}, y_{102}, y_{103}$ for any measure of that distribution rather than the first and second moments

Also known as second-order stationarity, this is a less restrictive form of stationarity. A time series is weakly stationary if:

1. **Constant mean:** The expected value of the series is the same at all times:

$$\mathbb{E}[X_t] = \mu \quad \text{for all } t$$

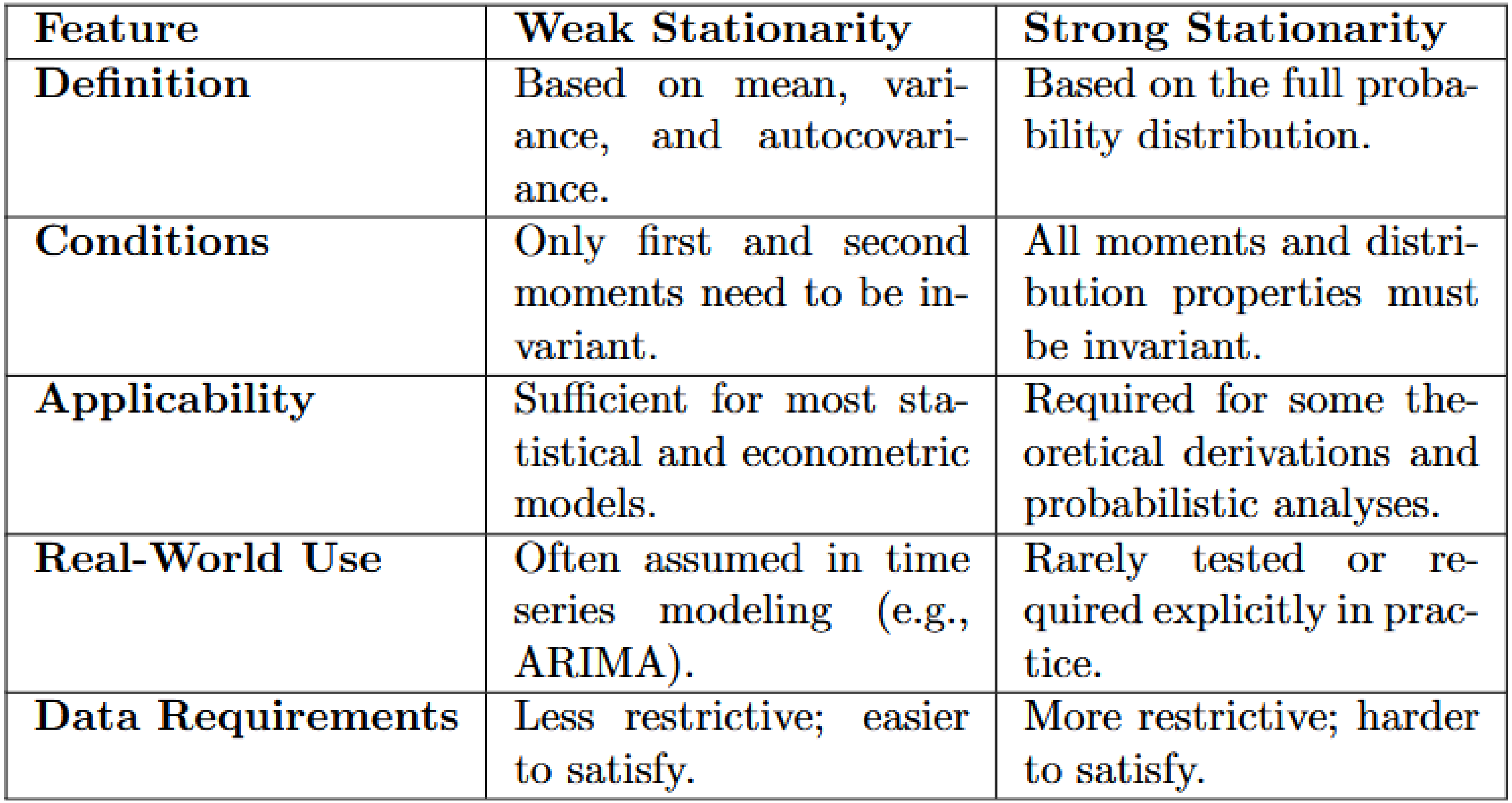
2. **Constant Variance:** The variance is the same at all times

$$\mathrm{Var}(X_t) = \sigma^2 \quad \text{for all } t$$

3. **Autocovariance Depends Only on Lag:** The autocovariance between X_t and X_{t+h} depends only on the time lag h (not on t):

$$\mathrm{Cov}(X_t, X_{t+h}) = \gamma(h) \quad \text{for all } t \text{ and } h$$

Weak stationarity VS Strong stationarity



Feature	Weak Stationarity	Strong Stationarity
Definition	Based on mean, variance, and autocovariance.	Based on the full probability distribution.
Conditions	Only first and second moments need to be invariant.	All moments and distribution properties must be invariant.
Applicability	Sufficient for most statistical and econometric models.	Required for some theoretical derivations and probabilistic analyses.
Real-World Use	Often assumed in time series modeling (e.g., ARIMA).	Rarely tested or required explicitly in practice.
Data Requirements	Less restrictive; easier to satisfy.	More restrictive; harder to satisfy.

⚠ Strong stationarity does not always imply weak stationarity

Solving AR(1) using linear regression

The AR(1) process can be reformulated as a simple linear regression problem:

$$y_t = \phi_0 + \phi_1 \cdot y_{t-1} + \epsilon_t$$

Fit a linear regression model to estimate ϕ_0 and ϕ_1.

Wait, Linear regression isn’t supposed to be a bad idea ?

Why use linear regression?

- **Approximation:** if the time series is stationary and the AR(1) structure dominates, linear regression can still provide decent parameter estimates.
- **Ease of Implementation:** Linear regression is straightforward to implement compared to Maximum Likelihood Estimation (MLE) or other complex methods.
- **Initialization for Advanced Models:** The estimates from linear regression can serve as initial guesses for iterative fitting methods like MLE or Bayesian estimation

Limitations and alternatives of linear regression

Limitations:

- **Bias in Parameters:** Because the residuals are autocorrelated, the Ordinary Least Squares (OLS) assumptions are violated, leading to biased estimates for the AR coefficient.
- **Underestimated Uncertainty:** The standard errors and confidence intervals for the parameters will likely be too narrow since the residual correlation isn't accounted for

Alternative:

- **Maximum Likelihood Estimation (MLE)**

When to use Linear regression:

- When you need a **quick** approximation of the AR(1) coefficient for exploratory purposes
- When the autocorrelation is **low** and the time series is **stationary**
- For **educational purposes** or **proof of concept** implementations

Maximum Likelihood Estimation

Suppose we have a statistical model denoted by M.
Example: If the law is a normal distribution, for example, there are two parameters, the mean and the variance.
→ Without loss of generation, let's consider a unique parameter noted θ.

Goal: Trying to explain the observed data using the model → find the parameters that best explain the data
Requirement: Find a score that measures how well a certain value of the parameter corresponds to the observed data → Likelihood

$$\mathcal{L} = P(obs|\theta_0, \mathcal{M})$$

Consider a series of i.i.d data x_i:

$$\mathcal{L} = \Pi_{i=1}^{N} P(x_i|\theta_0, \mathcal{M})$$

$$\theta^* = \arg\max_{\theta} P(obs|\theta, \mathcal{M})$$

$$\theta^* = \arg\max_{\theta} \sum_{i}^{m} \log P(x_i|\theta, \mathcal{M})$$

Maximum Likelihood Estimation Example

Let’s try to train an ML model to distinguish between an image of a dog or a muffin. Each image data is denoted x_i, the true label y_i, and the prediction $\hat{y}_i$

1- First we choose the model, in general we choose a gaussian process:

$$p(y|x) = \mathcal{N}(y; \hat{y}(x), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\hat{y}(x) - y)^2}{2\sigma^2}\right)$$

2- Given fixed θ, we estimate:

$$\sum_{i=1}^{m} \log p(y^{(i)}|x^{(i)}; \theta) = -m \log \sigma - \frac{m}{2} \log(2\pi) - \sum_{i=1}^{m} \frac{||\hat{y}^{(i)} - y^{(i)}||^2}{2\sigma^2}$$

3- Find θ that maximises log-likelihood or minimises negative log-likelihood

Fitting AR(1) with MLE

Fitting an AR(1) process using Maximum Likelihood Estimation (MLE) involves estimating the model parameters $(\phi_0, \phi_1, \sigma^2)$ such that the likelihood of observing the given data is maximized under the AR(1) assumptions

$$y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t$$

$$\mathcal{L}(\phi_0, \phi_1, \sigma^2 | y) = \prod_{t=2}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t - \phi_0 - \phi_1 y_{t-1})^2}{2\sigma^2}\right)$$

$$\log \mathcal{L} = -\frac{n-1}{2}\log(2\pi) - \frac{n-1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=2}^{n}(y_t - \phi_0 - \phi_1 y_{t-1})^2$$

Find the parameters $(\phi_0, \phi_1, \sigma^2)$ that maximise L

AR(p) models stationarity conditions

Recall that we define the AR(p) model as follows:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

1. **Mean stability:** The mean of the process must be constant over time. Ensures that the series has a finite and constant mean that does not diverge over time

$$\mu = \frac{\phi_0}{1 - \sum_{i=1}^{p} \phi_i} \quad \text{with} \quad (1 - \sum_{i=1}^{p} \phi_i) \neq 0$$

2. **Variance stability:** The variance of the process must remain constant over time

$$\text{Var}(y_t) = \frac{\sigma^2}{1 - \sum_{i=1}^{p} \phi_i^2}$$

With σ^2 the variance of ϵ_t

3. **Autocovariance stability**: The autocovariance $Cov(y_t, y_{t-k})$ must depend only on the lag k, not on t. Ensures that correlations between observations are a function of lag only, not the specific time points

→ This happens if the characteristic equation satisfies the **stationarity condition**

Characteristic Polynomial

The characteristic polynomial associated with the AR(p) process is: $1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0$

For the AR(p) process to be weakly stationary, the following condition must hold:

All roots of this polynomial must lie outside the unit circle in the complex plane.

→ This just means the magnitude of each root z_i must satisfy $|z_i| > 1$)

Special conditions:

- **AR(1):** The model is weakly stationary if $|\phi_1| < 1$
- **AR(2):** The roots of $(1 - \phi_1 z\ - \phi_2 z^2 = 0)$ must lie outside the unit circle. In practice, this implies constraints on ϕ_1 and ϕ_2
  1. $|\phi_2| < 1$
  2. $|\phi_1 + \phi_2| < 1$
  3. $|\phi_1 - \phi_2| < 1$

Time for Exercise 1

Dickey-Fuller test (1/2)

To make a series stationary, use differencing. But how many times do you have to difference your series? Once, twice? more? After differencing, how to make sure that the result is stationary?
→ **Dickey-Fuller test**

- Statistical test used to determine whether a given time series is stationary or has a unit root: A time series with a unit root is non-stationary. It exhibits a persistent, random walk-like behavior, where shocks to the system have a permanent effect.
- A simple autoregressive process of order 1, AR(1), is represented as: $Y_t = \rho Y_{t-1} + \epsilon_t$

  → If $\rho = 1$, the process has a unit root: implies non-stationarity, as the variance of y_t grows over time, and shocks have a permanent effect
- The test evaluates the null hypothesis H_0 that the time series has a unit root ($\rho = 1$), against the alternative hypothesis H_1 that it is stationary ($\rho > 1$)

Dickey-Fuller test (2/2)

Hypotheses:

Null hypothesis $\boldsymbol{H_0}$: the time series has a unit root $(\rho = 1)$ → non-stationary

Alternative Hypothesis H_1: The series has no unit root → stationary

The test transforms the AR(1) process by subtracting y_{t-1} from both sides:

$$Y_t - Y_{t-1} = \rho Y_{t-1} + \epsilon_t - Y_{t-1} = (\rho - 1)Y_{t-1} + \epsilon_t$$

Let $\gamma = \rho - 1$, so

$$\Delta Y_t = \gamma Y_{t-1} + \epsilon_t$$

Now, testing for a unit root is equivalent to testing $\gamma = 0$:

- If $\gamma = 0$: The series has a unit root (non-stationary)
- If $\gamma \neq 0$: the series is stationary

The test evaluates whether the estimated γ is significantly different from zero using a t-statistic.

Augmented Dickey-Fuller test

Limit of DF test: The Dickey-Fuller test assumes that the error term ϵ_t is white noise. If ϵ_t exhibits autocorrelation, the results may be invalid.

→ The Augmented Dickey-Fuller (ADF) test addresses this by including lagged differences of y_t in the model:

$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \sum_{i=1}^{p} \phi_i \Delta Y_{t-i} + \epsilon_t$$

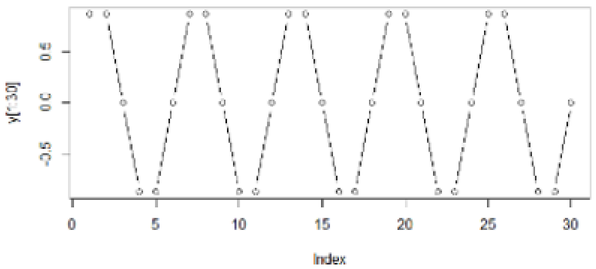
→ The ADF test is widely used because it is robust to higher-order autocorrelation

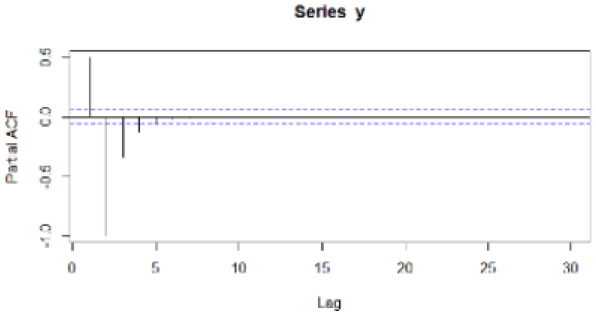
How to choose the number of lags

Because of serial dependence, it is crucial to take care of autocorrelation. The stronger the correlation between the output and a specific lagged variable, the more weight that autoregression can put on that specific variable. So that variable is considered to have a strong predictive power.

Reminder about PACF: The partial autocorrelation of a time series for a given lag is the partial correlation of the time series with itself at that lag given all the information between the two points in time removed.

→ Use the PACF !!





ADF test to choose the number of lags

You can also use the ADF to choose the number of lags.

$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \sum_{i=1}^{p} \phi_i \Delta Y_{t-i} + \epsilon_t$$

2 methods:

- add lags until the ϕ_i are unsignificant.

- use the residuals. You can add lag until you have no serial correlation in the residuals ϵ_t.

  → Durbin Watson or Breusch-Godfrey Test

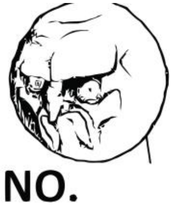
  → Plot the ACF of residuals at different lags. Look for significant spikes beyond the confidence bounds, indicating autocorrelation at specific lags

Evaluation Metrics

Akaike Information Criterion (AIC)

Using the PACF it is not always easy to decide what is a good lag value, so you might want to try a bunch of them.

→ Should you choose the one with the highest performance ?



You want a model that achieves a high performance (low error/ high likelihood) but also with the lowest complexity:

- The less complex is a model the more it is general.
- The more complex it is, the more chances it has to overfit on your data.

→ There is a trade off between model performance and complexity.

The AIC metric offers a way to settle the trade-off.

$$AIC = -2\log(\hat{L}) + 2k$$

Where:

- $\hat{L}$ is the maximum value of the likelihood function for the model
- k: The number of estimated parameters in the model (including any intercept or variance terms)

A lower AIC value indicates a better balance between model fit and complexity.
AIC does not provide an absolute measure of model quality; it only allows comparison between models.

Likelihood $-2\log(\hat{L})$:

- Measures how well the model explains the observed data.
- Higher likelihood means the model fits the data better.
- The term $-2\log(\hat{L})$ is negative because a larger likelihood corresponds to a smaller $-2\log(\hat{L})$, favoring better fits.

Penalty Term (2k):

- Increases with the number of parameters k in the model.
- Discourages overfitting by penalizing overly complex models with too many parameters.

Bayesian Information Criterion (BIC)

$$\mathrm{BIC} = -2\ln(\hat{L}) + k\ln(n)$$

Where:

- $\widehat{L}$ is the maximum value of the likelihood function for the model
- k: The number of estimated parameters in the model (including any intercept or variance terms)
- n: the number of data points

- Lower BIC values indicate better models: A model with a smaller BIC is preferred as it balances a good fit to the data with simplicity.
- Model Selection: When comparing models, the one with the smallest BIC is typically chosen.

Key Differences Between AIC and BIC:

- **Penalty**: BIC penalizes model complexity more heavily than the AIC especially as the sample size increases.
- BIC tends to select simpler models in large datasets due to its stronger penalty for complexity.
- AIC is more lenient and often used when predictive accuracy is the primary concern

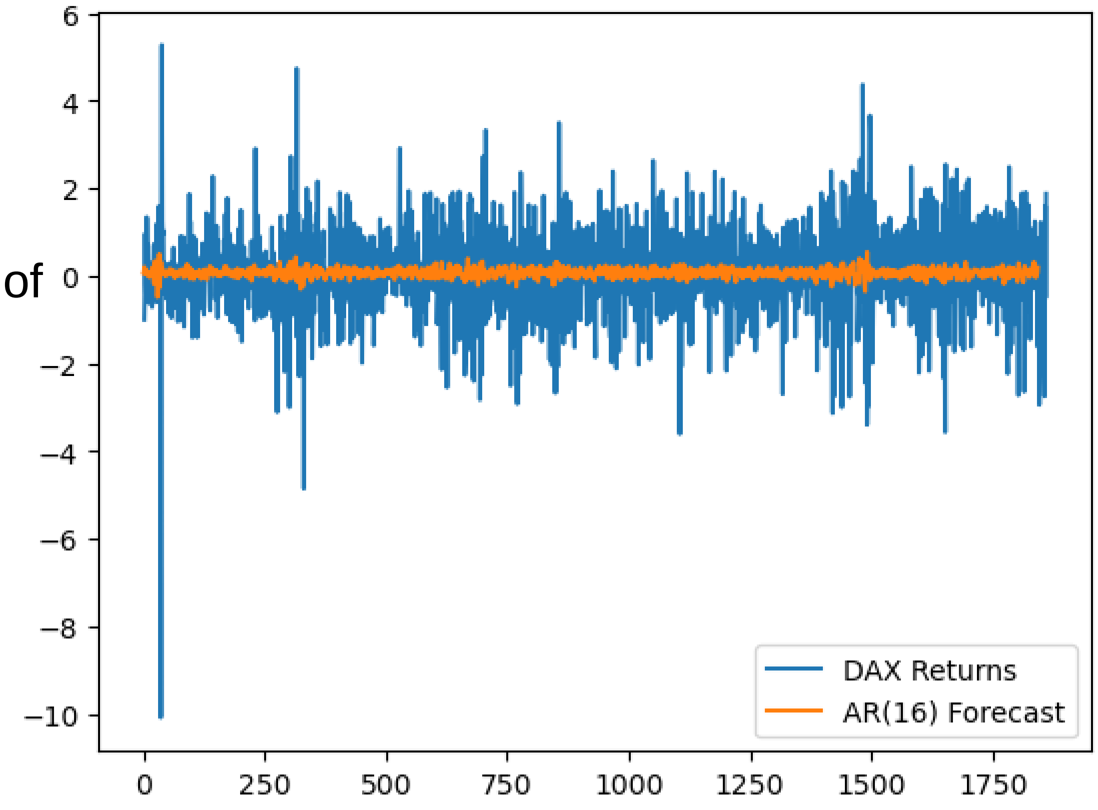
Time for Exercise 2

Forecasting with AR(p)

Regression towards the mean

The main difference between the forecast and the data is that the forecast is less variable than the data.

It may predict the direction of the future correctly, but not the scale of the change from one time period to another.



Forecasts are means of the predicted distributions and so necessarily will have lower variability than sampled data.

⚠ **Warning:** Predictions suggest a more stable future than will usually be the case. Forecasts suggest a much smoother future than is likely to be the case

N-step forecasts

$$y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t$$

$$y_{t+1} = \phi_0 + \phi_1 y_t + \epsilon_{t+1}$$

$$y_{t+2} = \phi_0 + \phi_1 y_{t+1} + \epsilon_{t+2}$$

If you are at time t and want to predict y_{t+2}, you need to predict y_{t+1} first.

Variance for forecasts made increasingly far into the future from the same underlying model.

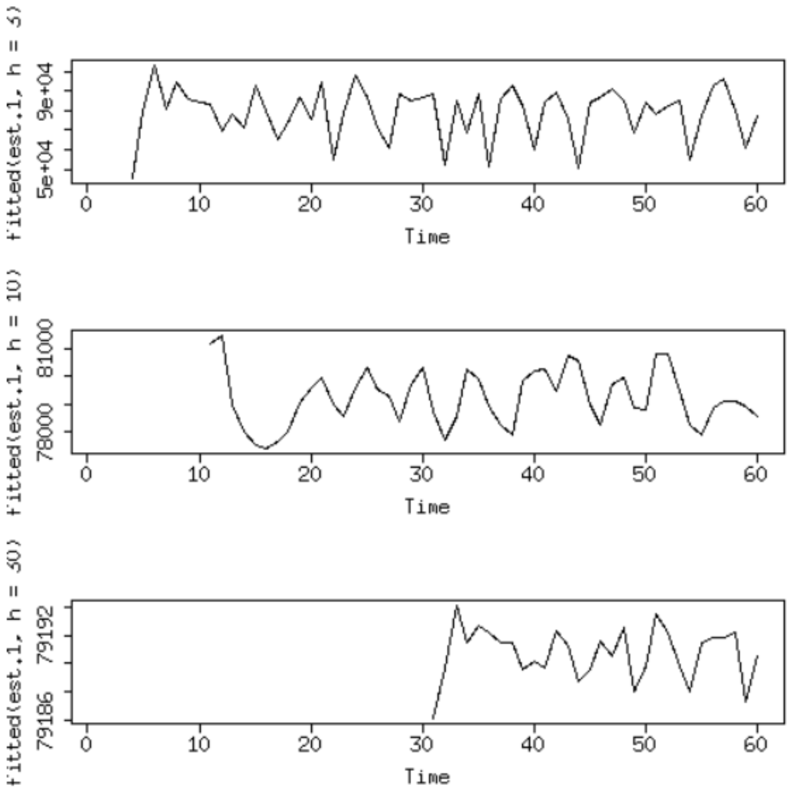
The variance of the prediction decreases with increasing forward horizon.

→ the further forward in time we go, the less the actual data matters because the coefficients for input data look only at a finite previous set of time points.

Forecasts further out in time converge to being the unconditional prediction that is, unconditioned on data.

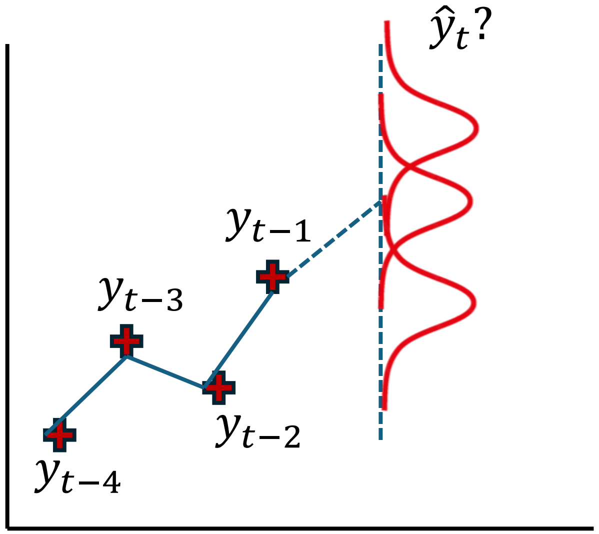
As the time horizon grows:

- The future prediction approaches the mean value of the series
- The variance of both the error term and of the forecast values shrinks to 0



Time for Exercise 3

Some Clarifications on MLE for AR(p)



$$\mathcal{L} = P(obs|\theta_0, \mathcal{M}) \longleftarrow$$ Normal Distribution

y_t

What is this exactly? It is the parameters of the model M, **its mean and variance**

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

$$p(y|x) = \mathcal{N}(y; \hat{y}(x), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\hat{y}(x) - y)^2}{2\sigma^2}\right)$$

$\hat{y}_t$ is the mean of a Normal distribution conditionned by the previous values

→ Given that we use an AR(p), the mean is modeled as $y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-1} + \cdots + \phi_p y_{t-p}$.

→ The mean itself is parametered by the coefficients ϕ_i !

→ Finding the mean is equivalent to finding the parameters ϕ_i

→ Once the AR(p) model is fitted, the coefficients ϕ_i do not change but the y_{t-i} change, so the mean changes. Every time you are estimating a new distribution with a different mean. $\mathbb{E}[y_t|y_{t-1}, \ldots, y_{t-p}] = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p}$

→ **Bonus:** MLE allows to estimate the variance of the distribution → you get uncertainty over your prediction for free

Moving Average Models

MA models principle

AR model: $y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t$

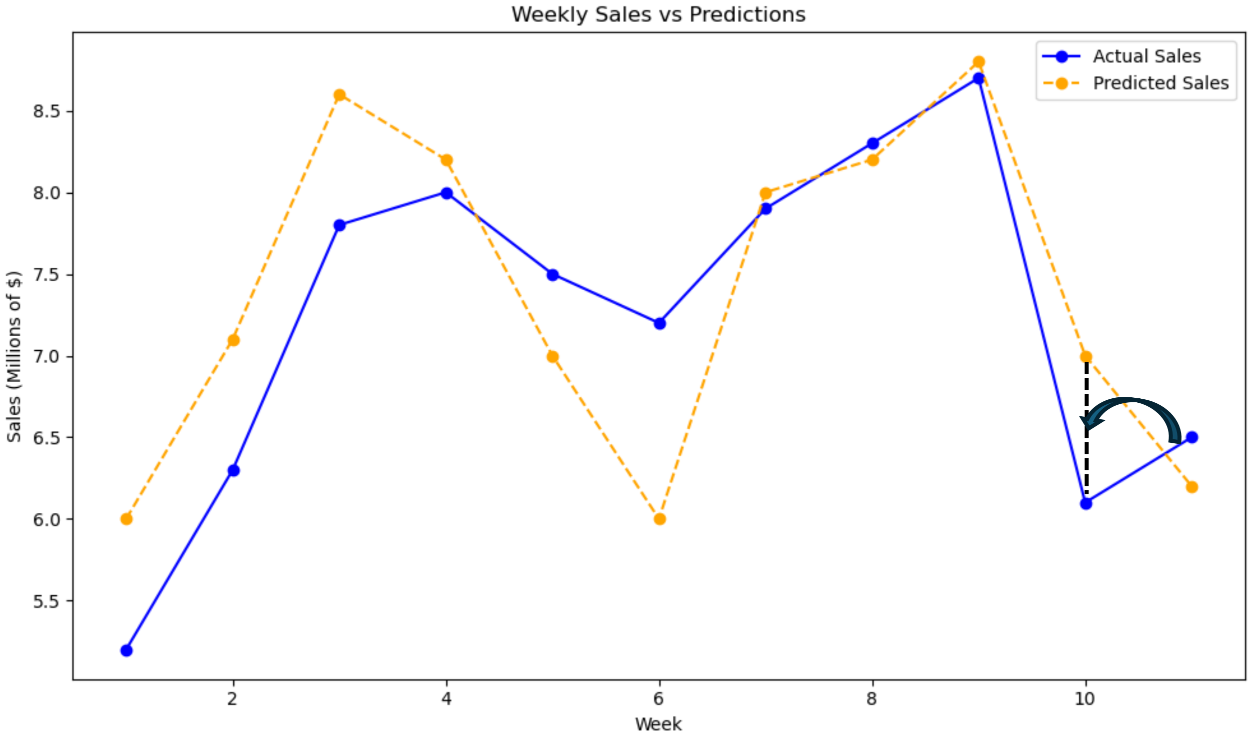
Instead of looking at previous values of Y we are instead going to look at previous errors:

$$y_t = \phi_0 + \phi_1 e_{t-1} + \phi_2 e_{t-2} + \cdots + \phi_p e_{t-p} + \epsilon_t$$

Intuition: The error from yesterday affects the current value. Some unknown predicted shock that shifted you off of where you expected to be is what's actually affecting the current time point.

There is no error for the first prediction, so we use the mean of the series as a starting point.

$$\boxed{\phi_0 = \mu}$$



Example

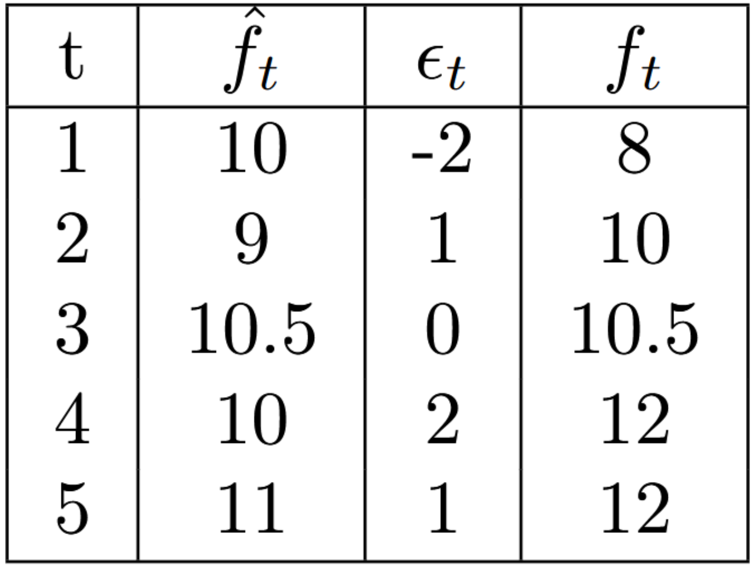
You’re a survivor in a zombie apocalypse, and every week you stock up on supplies. On average, you gather 10 cans of food

$$\phi_0 = 10$$

Scavenging is unpredictable.

$$\hat{f}_t = \mu + \phi \times e_{t-1}$$

So if we say $\phi_1 = 0.5$ and recall that the error e_t is normally distributed with mean 0 and standard deviation 1.



t	$\hat{f}_t$	ϵ_t	f_t
1	10	-2	8
2	9	1	10
3	10.5	0	10.5
4	10	2	12
5	11	1	12

Stationarity of MA

We've seen earlier some conditions that make the AR process stationary. What are the conditions that make an MA process stationary?

⚠ MA model is always stationary regardless of its parameters.

- Stationarity is inherent by design
- MA model does not have a dependence on its own past values but only on past error terms
- Errors are assumed to have a constant mean and variance.

Proof of Stationarity for MA(1)

The MA process of order 1 is defined as:

$$y_t = \mu + \epsilon_t + \theta\epsilon_{t-1}$$

Mean:

$$\mathbb{E}[y_t] = \mathbb{E}[\mu + \epsilon_t + \theta\epsilon_{t-1}] = \mu + \mathbb{E}[\epsilon_t] + \theta\mathbb{E}[\epsilon_{t-1}]$$

Since $\mathbb{E}[\epsilon_t] = 0$

$$\mathbb{E}[y_t] = \mu$$

Variance:

$$\begin{aligned}\mathrm{Var}(y_t) &= \mathrm{Var}(\mu + \epsilon_t + \theta\epsilon_{t-1}) \\ &= \sigma^2 + \theta^2\sigma^2 = (1+\theta^2)\sigma^2\end{aligned}$$

Thus, the variance is constant over time.

Proof of Stationarity for MA(1)

Autocovariance: The autocovariance at lag k is: $\mathrm{Cov}(y_t, y_{t-k}) = \mathbb{E}[(y_t - \mu)(y_{t-k} - \mu)]$

For k = 0 (autocovariance at lag 0): $\mathrm{Cov}(y_t, y_t) = \mathrm{Var}(y_t) = (1 + \theta^2)\sigma^2$

Substituting $y_t = \mu + \epsilon_t + \theta\epsilon_{t-1}$: $y_t - \mu = \epsilon_t + \theta\epsilon_{t-1},$

$$y_{t-k} - \mu = \epsilon_{t-k} + \theta\epsilon_{t-k-1}$$

For k = 1 (autocovariance at lag 1): $\mathrm{Cov}(y_t, y_{t-1}) = \mathbb{E}[(\epsilon_t + \theta\epsilon_{t-1})(\epsilon_{t-1} + \theta\epsilon_{t-2})]$

$$\mathrm{Cov}(y_t, y_{t-1}) = \underbrace{\mathbb{E}[\epsilon_t\epsilon_{t-1}]}_{0} + \theta\underbrace{\mathbb{E}[\epsilon_t\epsilon_{t-2}]}_{0} + \theta\mathbb{E}[\epsilon_{t-1}^2] + \underbrace{\theta^2\mathbb{E}[\epsilon_{t-1}\epsilon_{t-2}]}_{0}$$

$$\mathrm{Cov}(y_t, y_{t-1}) = \theta\sigma^2$$

For k > 1:

$$\mathrm{Cov}(y_t, y_{t-k}) = 0$$

Since ϵ_t and ϵ_{t-2} are uncorrelated for $k \geq 2$

Short memory model

MA models are called **short memory** models:

- the first prediction is the mean of the series → starting point
- with a long enough series, it does not really matter what is the starting point.

$$y_{t-1} = \phi_0 + \theta\epsilon_{t-2} + \epsilon_{t-1}$$

$$y_t = \phi_0 + \theta\epsilon_{t-1} + \epsilon_t$$

$$y_{t+1} = \phi_0 + \theta\epsilon_t + \epsilon_{t+1} \longrightarrow \text{No more } \epsilon_{t-1} \text{ !!}$$

What does stationarity really mean?

Rigourous definition:

- Mean and variance are steady
- covariance does not depend on time

Intuition:

the dependence on previous observations declines over time

In MA models, they not only decline but they disappear!

Why is it called moving average?

$$y_t = \phi_0 + \phi_1 e_{t-1} + \phi_2 e_{t-2} + \cdots + \phi_p e_{t-p} + \epsilon_t$$

it represents the current value of a time series as a weighted average of current and past errors in the system.

These errors are considered to "move" across time, and their weighted contributions to the series change dynamically.

$$y_{t-1} = \phi_0 + \theta\epsilon_{t-2} + \epsilon_{t-1}$$

$$y_t = \phi_0 + \theta\epsilon_{t-1} + \epsilon_t$$

$$y_{t+1} = \phi_0 + \theta\epsilon_t + \epsilon_{t+1}$$

AR model: the current value is expressed as a function of its past values,
MA model: averages the impact of past random shocks.
→ good model for capturing short-term correlations caused by residual effects or random fluctuations.

No linear regression for MA!

For AR, we gave some reasons to fit it using linear regression → **Do not do that for MA!** ⚠

MA models depend on the errors from previous time steps, and these residuals are not observable, they must be estimated during the model-fitting process.

$$y_t = \mu + \epsilon_t + \theta\epsilon_{t-1}$$

The residuals depend on the model parameters so they are not known beforehand
→ Estimating ϵ requires prior knowledge of the model, which introduces circular dependency.

Using MLE does not have the same limitations as linear regression:
→ MLE explicitly models the unobserved residuals as part of the parameter estimation process

Why MLE is better suited for MA

Linear regression limitations:

- Linear regression assumes the predictors (independent variables) are observable and fixed. In an MA(1) process, the lagged residual ϵ_{t-1} is unobserved and depends on the model parameters being estimated.

- This circular dependency makes it impossible to fit an MA(1) process directly using linear regression.

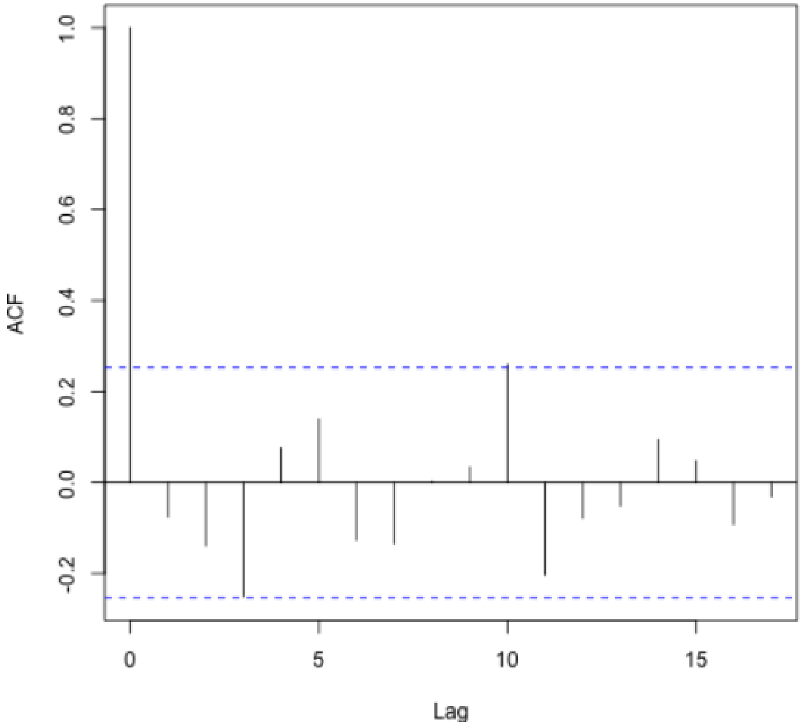
MLE advantages:

- MLE treats the residuals ϵ_t as random variables and estimates the likelihood of the observed data given the parameters μ, θ, σ^2

- It uses the full time series and accounts for the unobserved residuals by optimizing the parameters to maximize the probability of the observed data.

Selecting the number of parameters for MA(q)

Should we use PACF as for AR models?



MA models are called **short memory** models:

- the definition of the MA process ensures a sharp cutoff of the ACF for any value greater than q

We see significant values at lags 3 and 9, so we fit an MA model with these lags.

Forecasting with MA(q)

An MA(1) process is defined as:

$$y_t = \mu + \epsilon_t + \theta\epsilon_{t-1}$$

For one step ahead forecasting (y_{t+1}):

- You can use ϵ_t, which is known at time t, to predict y_{t+1} because it directly influences both y_t and y_{t+1}.

For k step ahead forecasting:

- The error terms beyond the current known values are unknown

  → the prediction collapses to the mean of the process

For an MA(q) process, the general form is:

$$y_t = \phi_0 + \phi_1 e_{t-1} + \phi_2 e_{t-2} + \cdots + \phi_p e_{t-p} + \epsilon_t$$

- You can only provide an informed prediction up to q-steps ahead because the process depends on q lagged noise terms.
- Beyond q-steps, all the error terms needed for prediction are unknown, and the forecast defaults to the mean mu of the process.

⚠

The further your prediction, the weaker it is !

$$y_t = \phi_0 + \phi_1 \cancel{e_{t-1}} + \phi_2 \cancel{e_{t-2}} + \cdots + \phi_p e_{t-p} + \epsilon_t$$