



NYC DATA SCIENCE
ACADEMY

Time Series Analysis

Data Science with R: Machine Learning

Outline

- ❖ **Part 1: The Nature of Time Series Analysis**
- ❖ **Part 2: Decomposition of Time Series Data**
- ❖ **Part 3: ARIMA Models**
- ❖ **Part 4: Assessing Model Fit**
- ❖ **Part 5: Review**

PART 1

The Nature of Time Series Analysis

Cross-Sectional, Longitudinal, & Time Series Data

- ❖ So far our analyses have been limited to **cross-sectional** data; the variables we have been considering have theoretically been measured at a single point in time for each of the observations in our dataset.
- ❖ What if we have data on variables that are repeatedly measured over time? This kind of data is called **longitudinal**, and involves following a phenomenon by measuring its changes through time.
- ❖ Longitudinal data that has been recorded at regularly spaced time intervals for a given span of time comprise a **time series**, and will be the main subject of today's analysis.

The Goal of Time Series Analysis

- ❖ The two main questions we wish to answer when modeling data of a time series nature are:
 - What happened in the **past**?
 - What will happen in the **future**?
- ❖ The analysis of the past events suffices as a **description** of events leading up to the present, whereas the analysis of what will happen offers a **prediction** for what will come after the present.

Applications of Time Series Analysis

- ❖ The prediction of future events, also known as **forecasting**, has vast applications across the social, decision, and classical sciences:
 - Economics: understanding the nature of the stock market.
 - Meteorology: understanding global climate change.
 - Epidemiology: understanding the spread of disease.
- ❖ Before we can forecast, we attempt to break down our time series data into various smaller components that each indicate, in different ways, how a change in the present influences a change in the future.

Why Can't We Use Linear Regression?

- ❖ First of all, linear regression assumes independence among the errors. This might be fine for cross-sectional data, but inherently in a time series model observations that are collected close to each other in time are related.
 - Time series data **violates the independence assumption** of linear regression.
- ❖ In time series data, values of Y_t are **theoretically related** to values of Y_{t-1} , Y_{t-2} , ..., Y_0 simply by the way they were collected over time.
 - Your bank account's balance on a specific day (Y_t) is related to your bank account's balance on the previous day (Y_{t-1}).
- ❖ Regression without accounting for these lags will fail to account for the relationships through time and can lead to **faulty conclusions** about the relationship between our independent and dependent variables.

PART 2

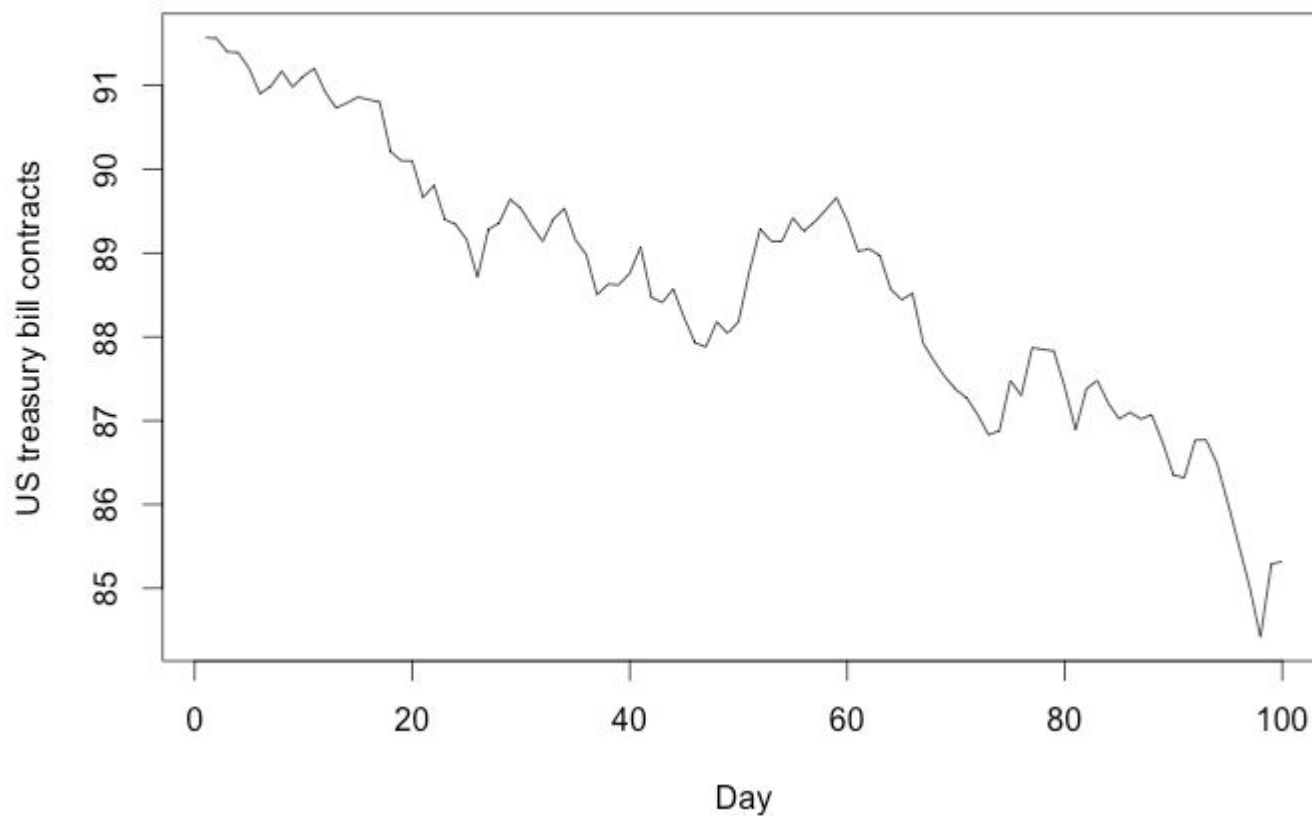
Decomposition of Time Series Data

Basic Components of a Time Series

- ❖ The basic components of a time series boil down to the following components:
 - The **trend** component highlights the long-term nature of the series; it helps describe whether the series is generally increasing or decreasing.
 - The **seasonal** component highlights a repeating effect that is observed over a fixed period of time.
 - The **irregular** or **error** component captures those influences not described by the other effects; it is essentially what is “left over.”
- ❖ You might also hear of a **cyclical** component, which highlights a repeating effect that is observed over **non-fixed** periods of time. This is generally absorbed by the trend and seasonal components, so we won’t focus on it too much.

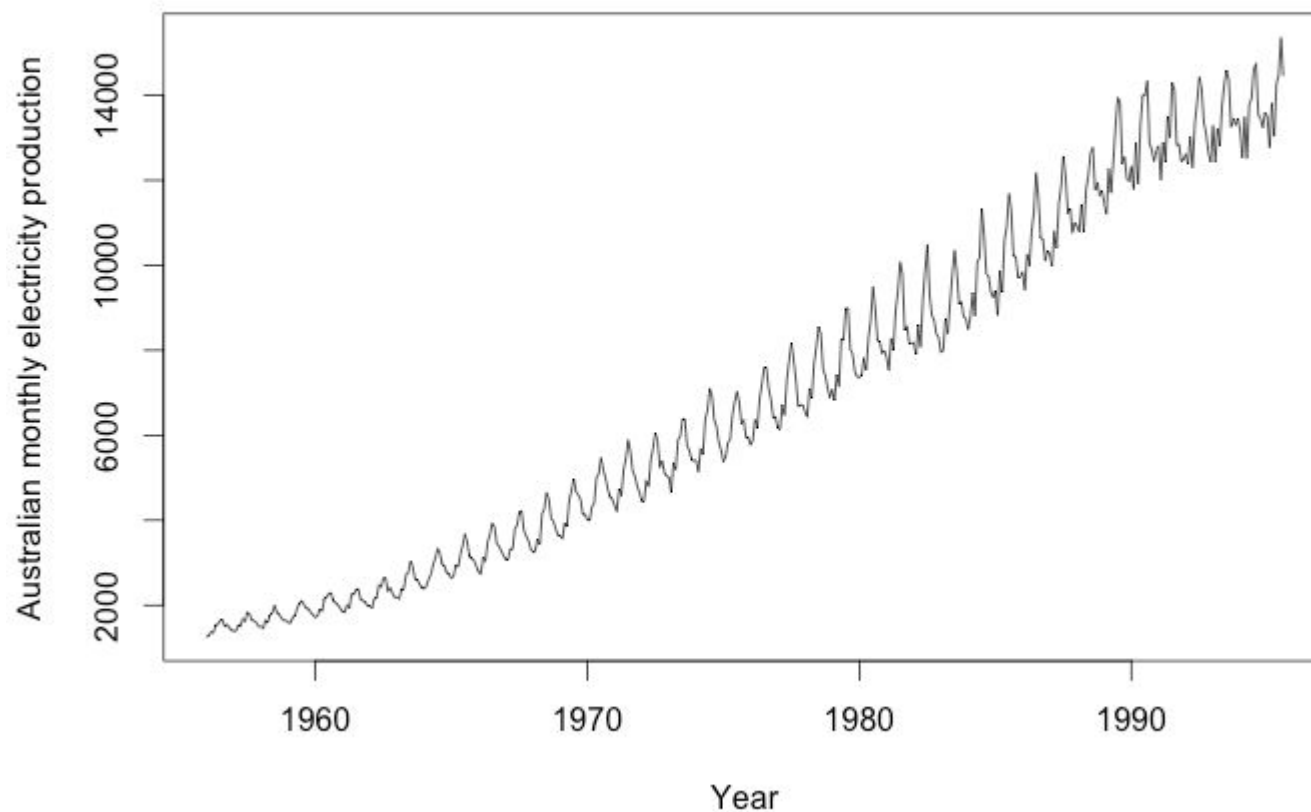
Basic Components of a Time Series

- ❖ What might a **trend** in a time series look like?



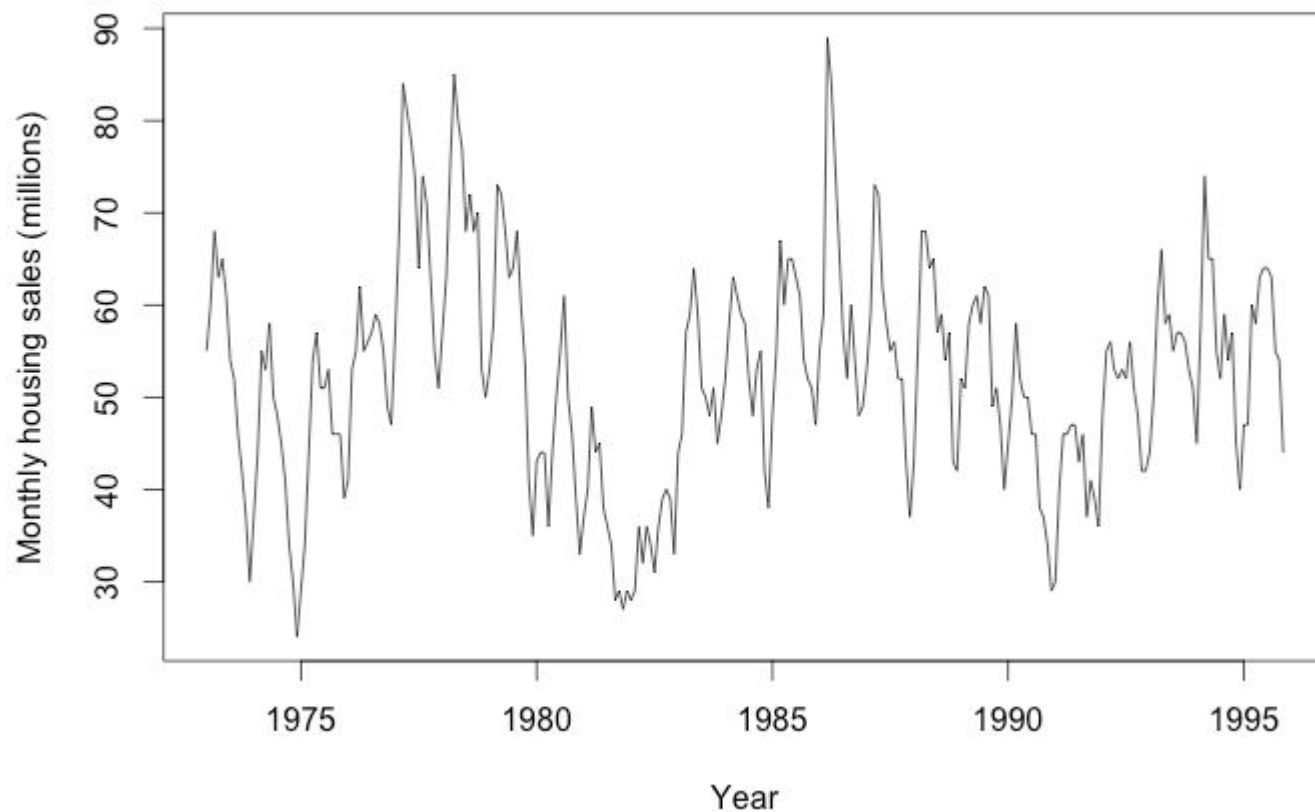
Basic Components of a Time Series

- ❖ What might a **trend** with **seasonality** in a time series look like?



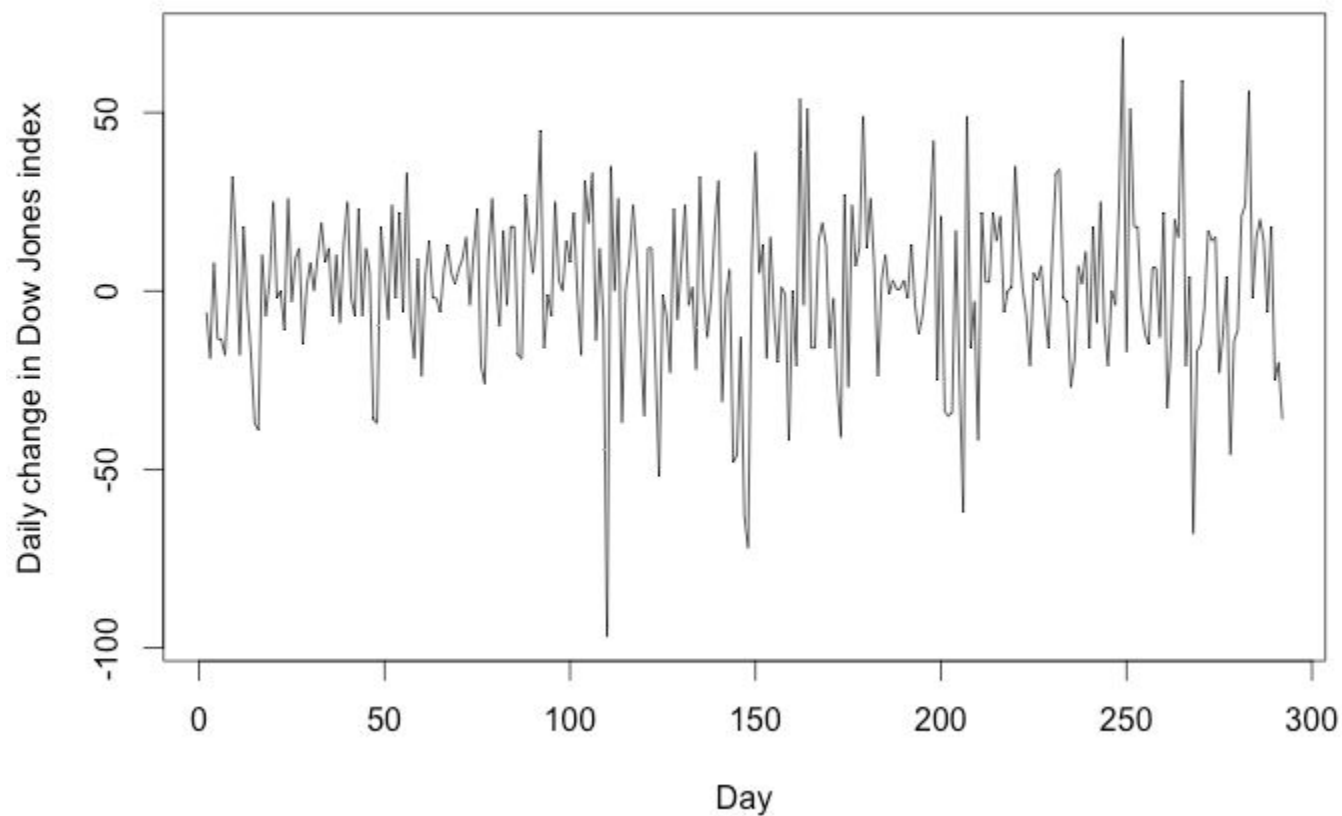
Basic Components of a Time Series

- ❖ What might a **seasonal** and a **cyclical** effect in a time series look like?



Basic Components of a Time Series

- ❖ What might an **irregular** time series look like?



Description: What Happened in the Past?

- ❖ Just as with any other analyses we have seen thus far, we should always begin our process by doing some exploratory data analysis; the EDA will suffice as the **description** component of a time series analysis.
- ❖ For time series analysis, basic numerical and graphical EDA takes the form of:
 - Smoothing
 - Seasonal Decomposition

Smoothing for General Trends

- ❖ Time series often have a bountiful error component which makes it difficult to discern general patterns in the data. What can we do to view the **general trends**?
- ❖ One of the simplest forms of describing the overall pattern of a time series is **smoothing**, which can help dampen the fluctuations we observe in the irregular component and help highlight the more **global** aspects of the series.
- ❖ To graphically depict the general trends, we will consider the idea of **centered moving averages**.

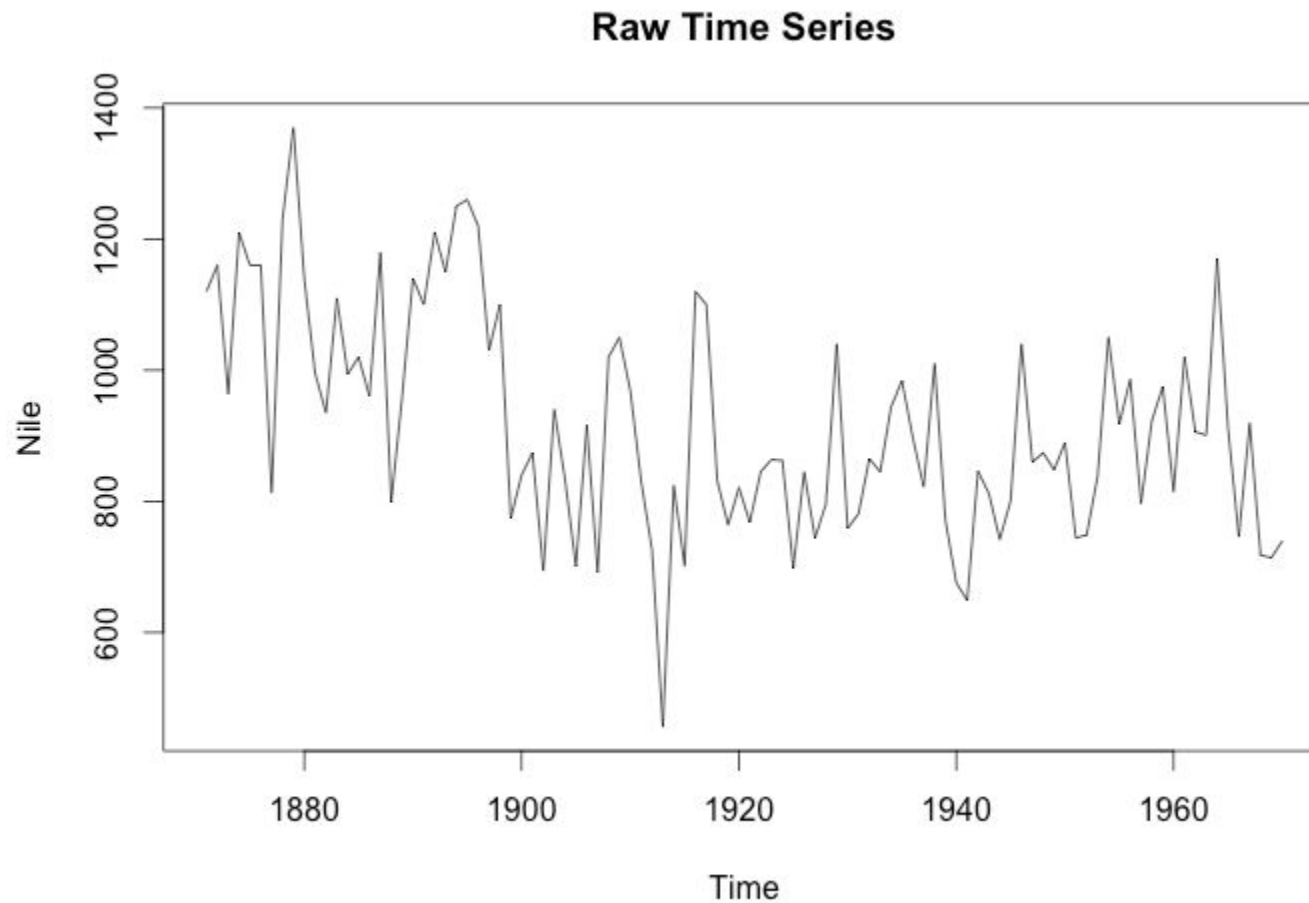
Centered Moving Averages

- ❖ In the construction of a **centered moving average**, each data point is replaced with the mean of that observation and a certain number of observations both before and after.

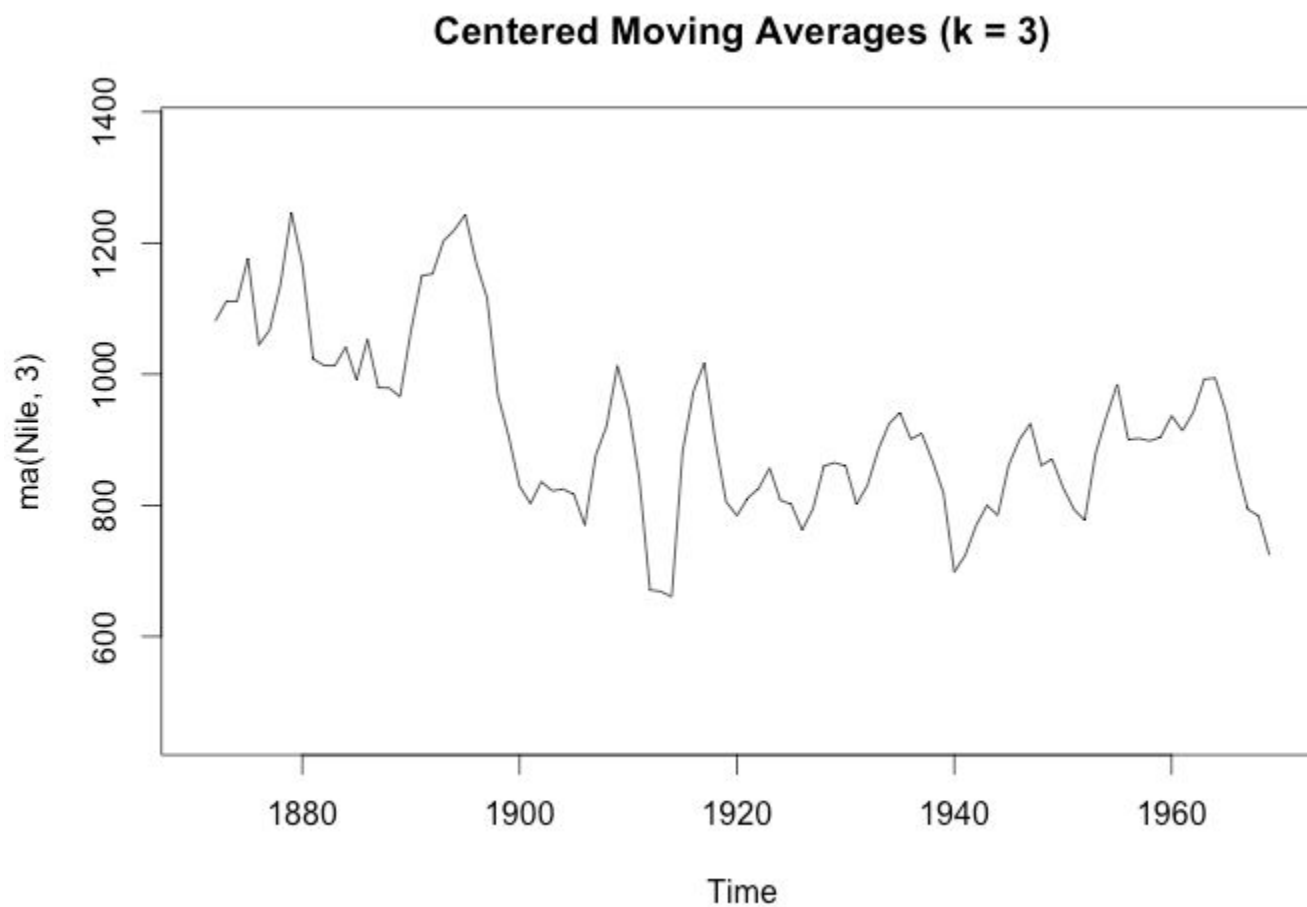
$$S_t = \frac{Y_{t-q} + \dots + Y_t + \dots + Y_{t+q}}{2q+1}$$

- ❖ Here, S_t is the **smoothed value at time t** after taking into account both q terms before and q terms after.
 - What are some problems with this method?
- ❖ Simply by data limitations, when using this smoothing method we “lose” q **observations** at each end of the series because we cannot estimate them.

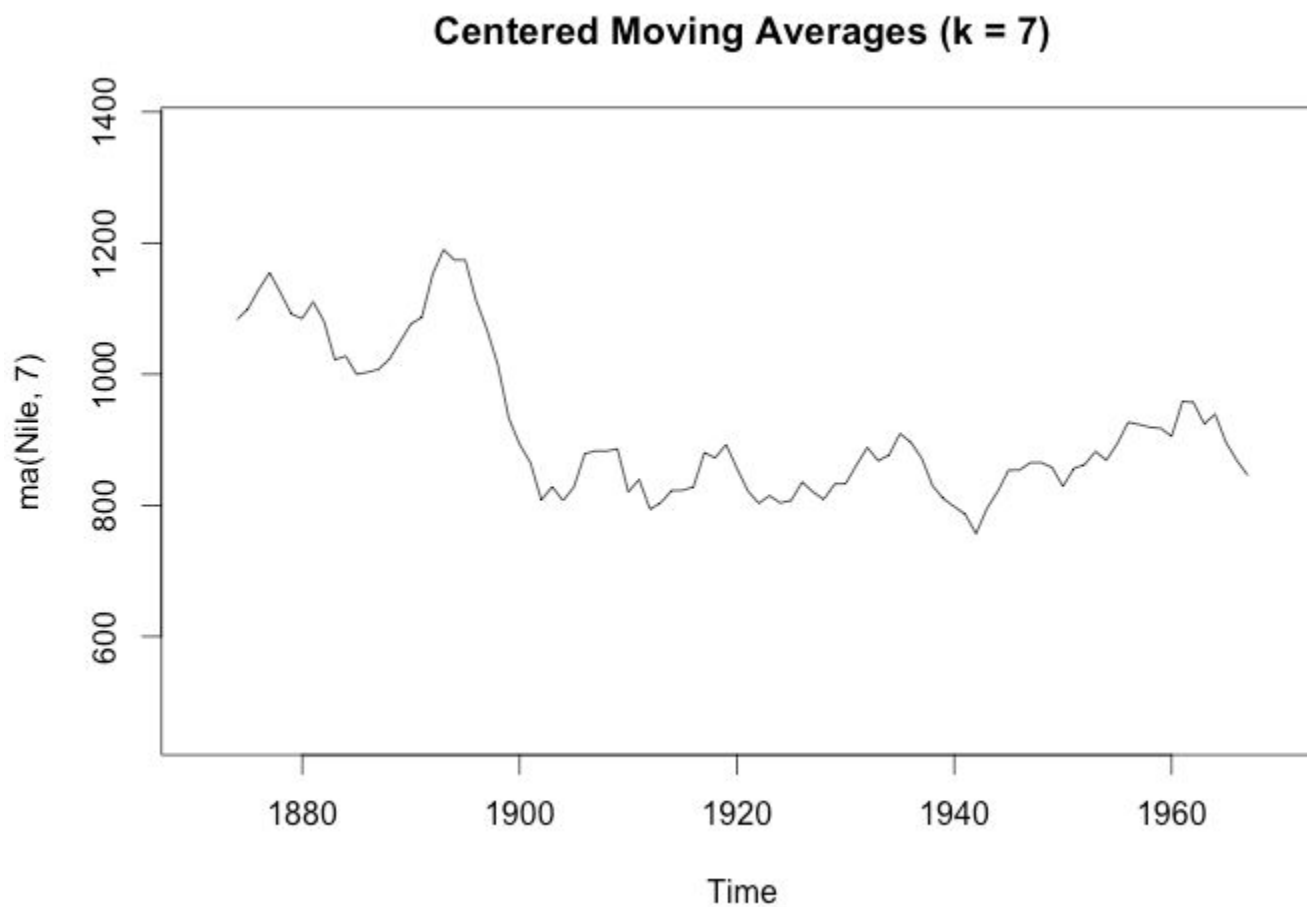
Centered Moving Averages



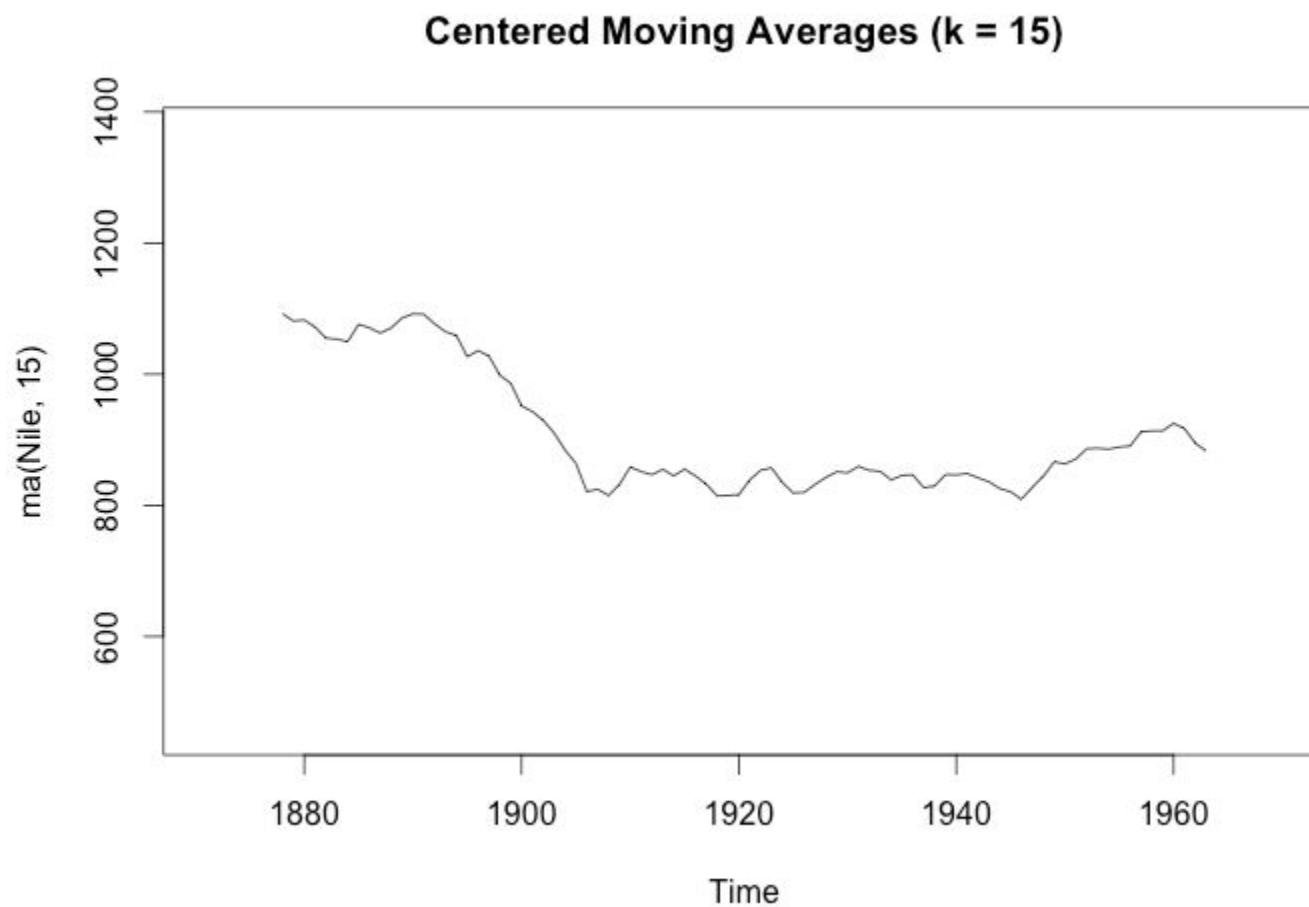
Centered Moving Averages



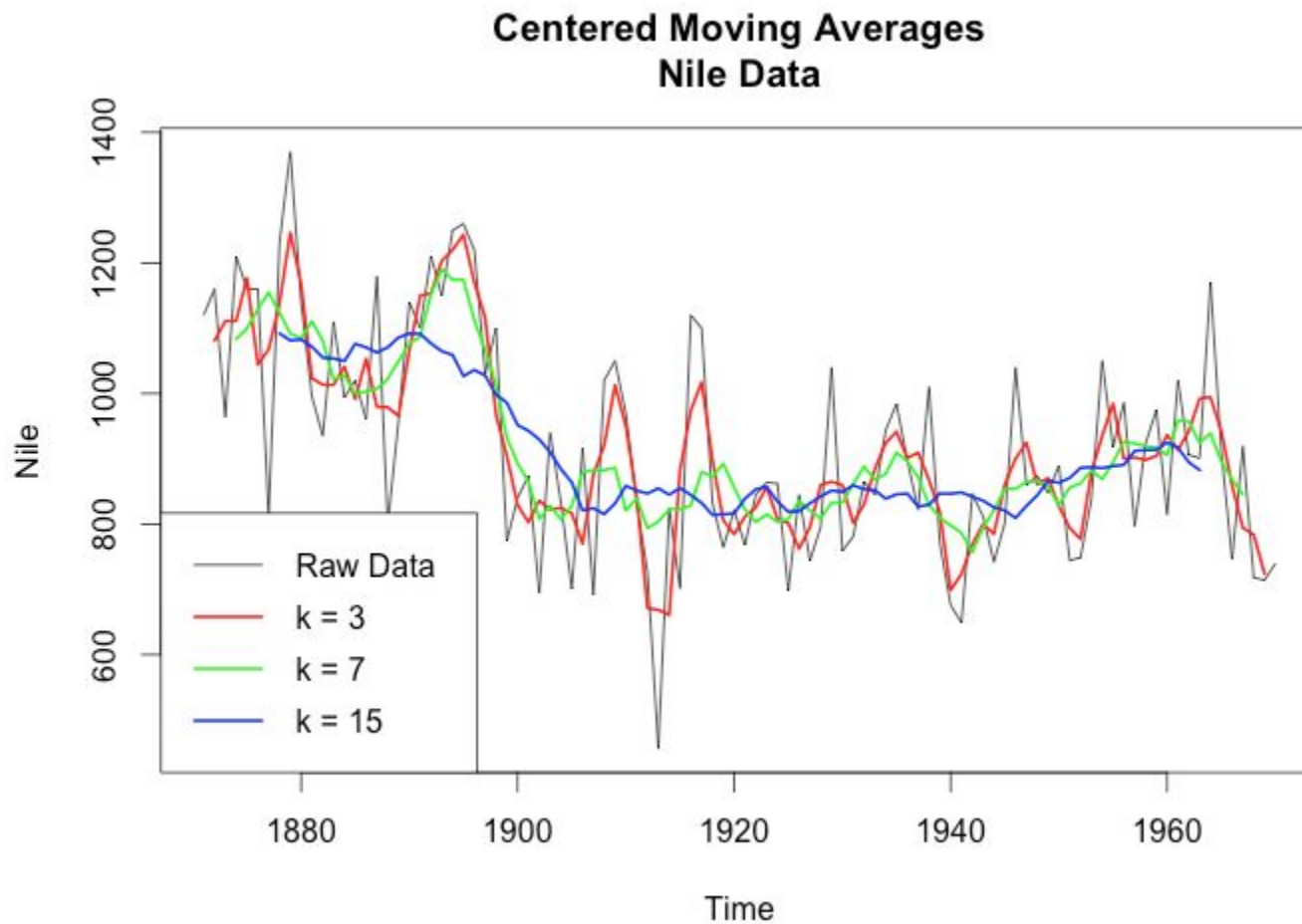
Centered Moving Averages



Centered Moving Averages



Centered Moving Averages



Seasonal Decomposition

- ❖ When time series data displays some type of **periodicity**, this is an indication that seasonality exists within the series.
 - Why might a centered moving average no longer suffice for description?
- ❖ In **seasonal decomposition**, we aim to break down the series into the following model, which can be either additive or multiplicative:

$$Y_t = Trend_t + Seasonal_t + Irregular_t$$

$$Y_t = Trend_t \times Seasonal_t \times Irregular_t$$

- ❖ An observation at time t is the sum (or product) of the contributions of the trend, seasonal, and irregular components existent at time t .

Seasonal Decomposition

- ❖ The choice between additive or multiplicative decompositions is simple:
 - Use the **additive model** when the magnitude of the seasonal fluctuations or the variations surrounding the general trend **does not vary** over time.
 - Use the **multiplicative model** when the magnitude of the seasonal fluctuations or the variations surrounding the general trend appears to be **changing in a proportional manner** over time.
- ❖ **NB:** Multiplicative models can be transformed into additive models by simply applying a **log transformation**; the results can then be back-transformed onto the original scale by exponentiation:

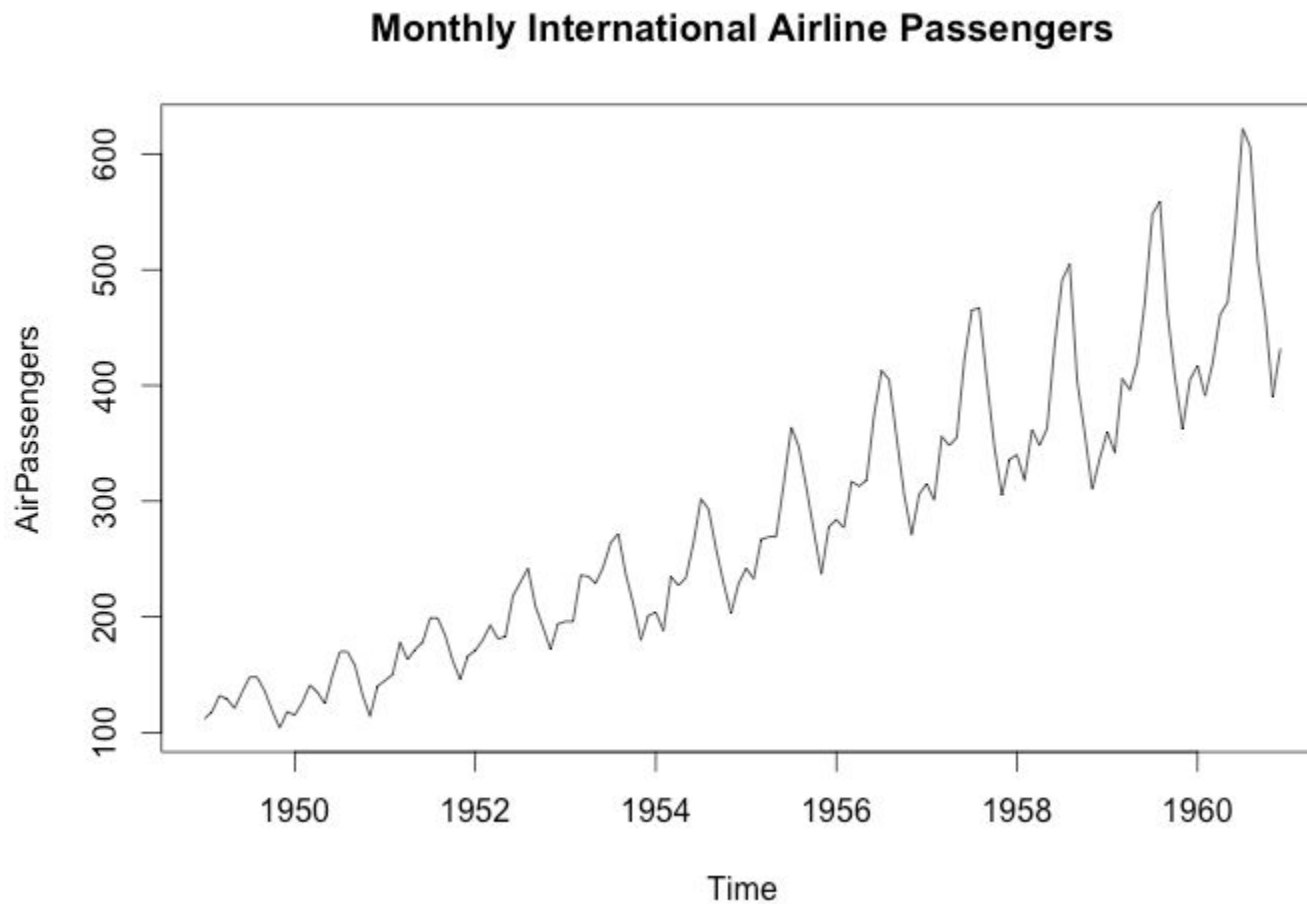
$$\ln(Y_t) = \ln(Trend_t \times Seasonal_t \times Irregular_t)$$

$$\ln(Y_t) = \ln(Trend_t) + \ln(Seasonal_t) + \ln(Irregular_t)$$

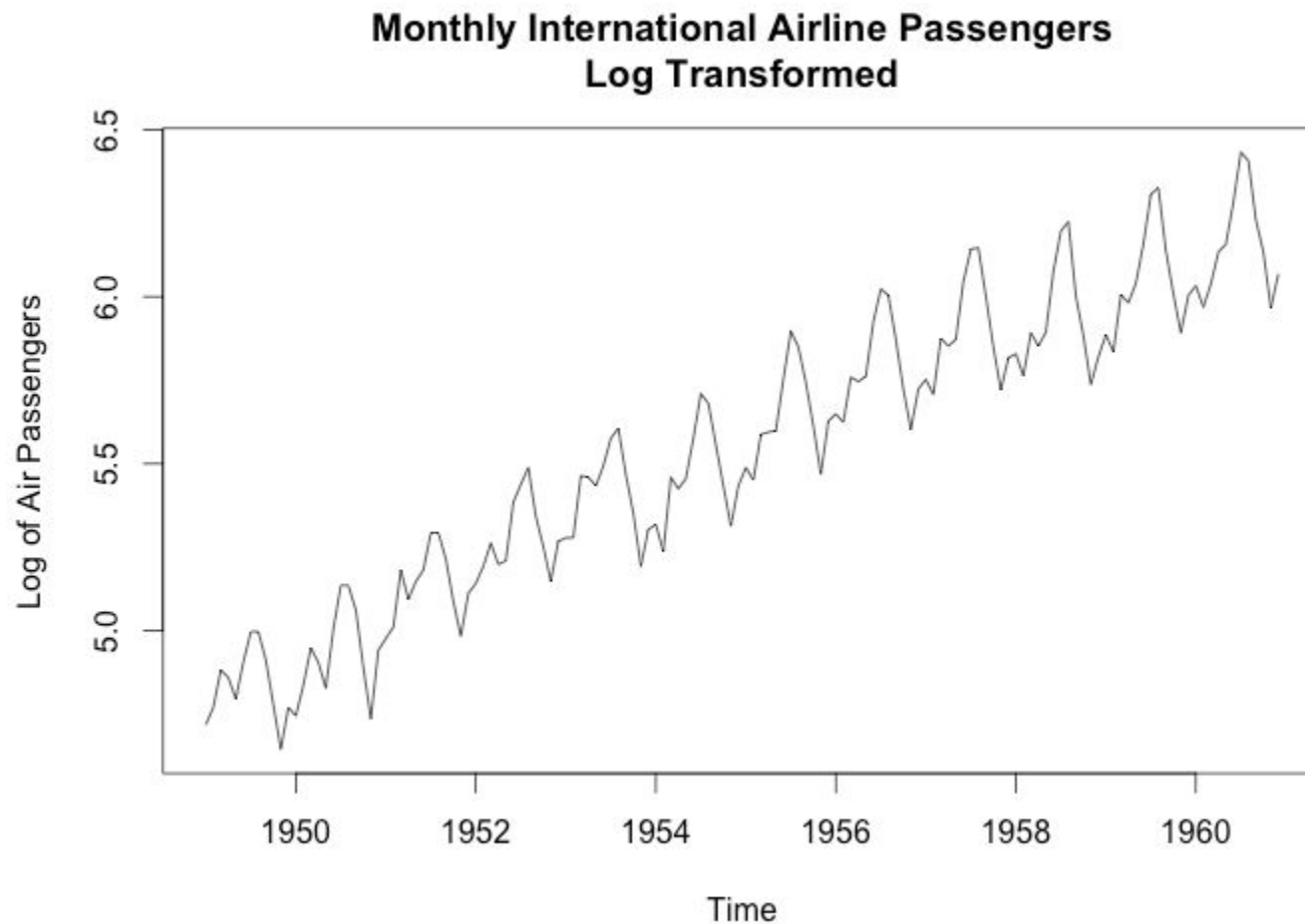
Seasonal Decomposition

- ❖ The most popular method for performing seasonal decomposition was developed by [Cleveland et al. \(1990\)](#) and is called “Seasonal and Trend Decomposition using LOESS,” or **STL** for short.
- ❖ The method is composed of a series of filtering procedures that repeatedly use the **LOESS (locally estimated smoothing)** procedure.
 - The derivation of this method is outside the scope of our discussion, but provides a procedure that is both **versatile** and **robust**.
- ❖ Let’s see the power of the STL decomposition with an example...

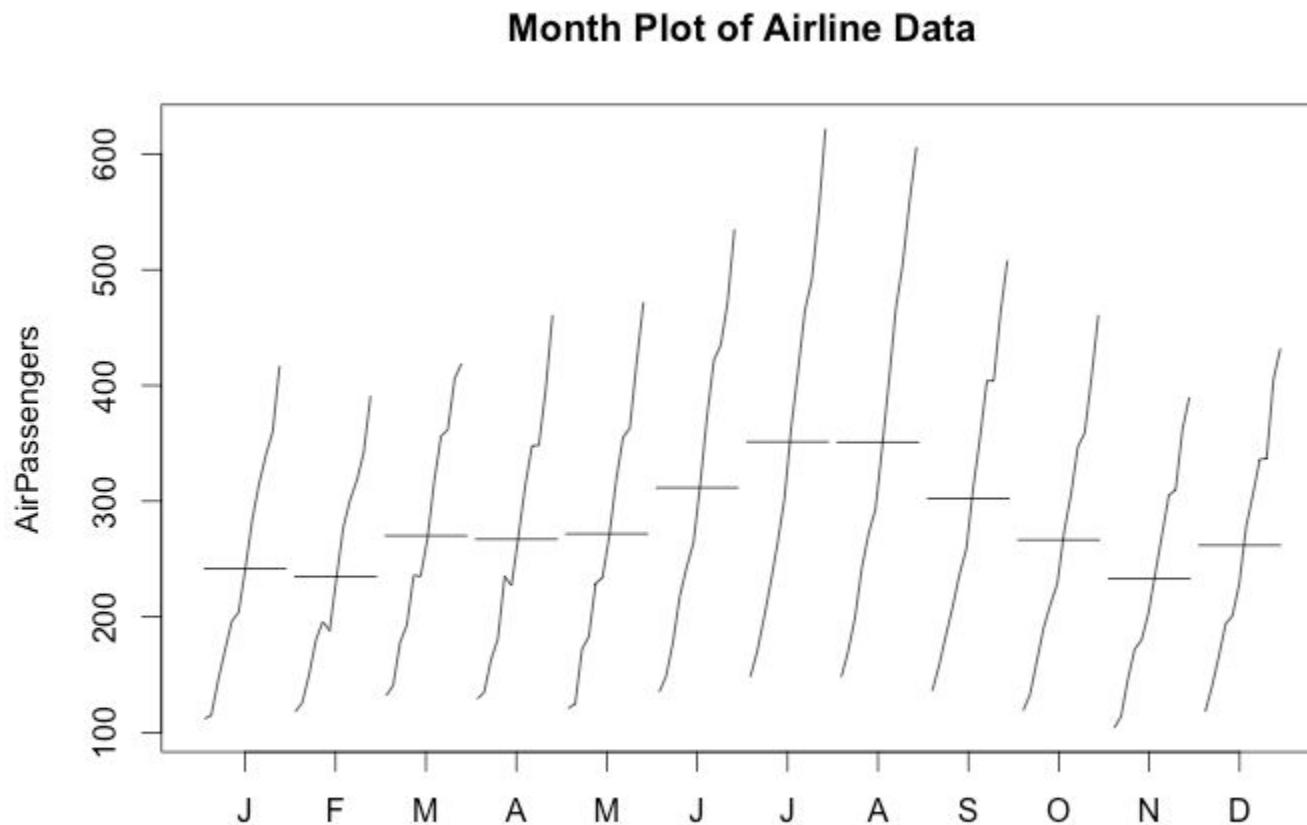
Seasonal Decomposition



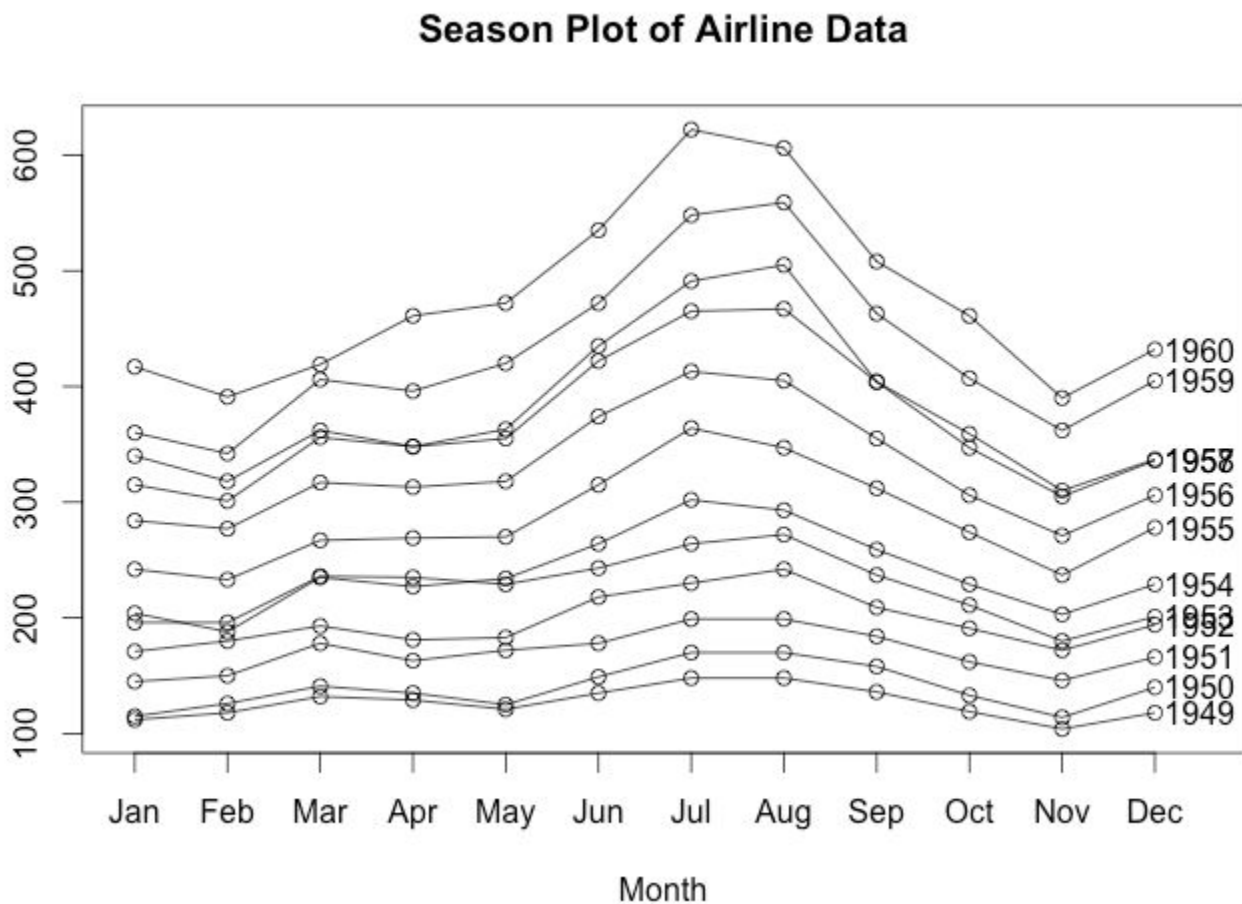
Seasonal Decomposition



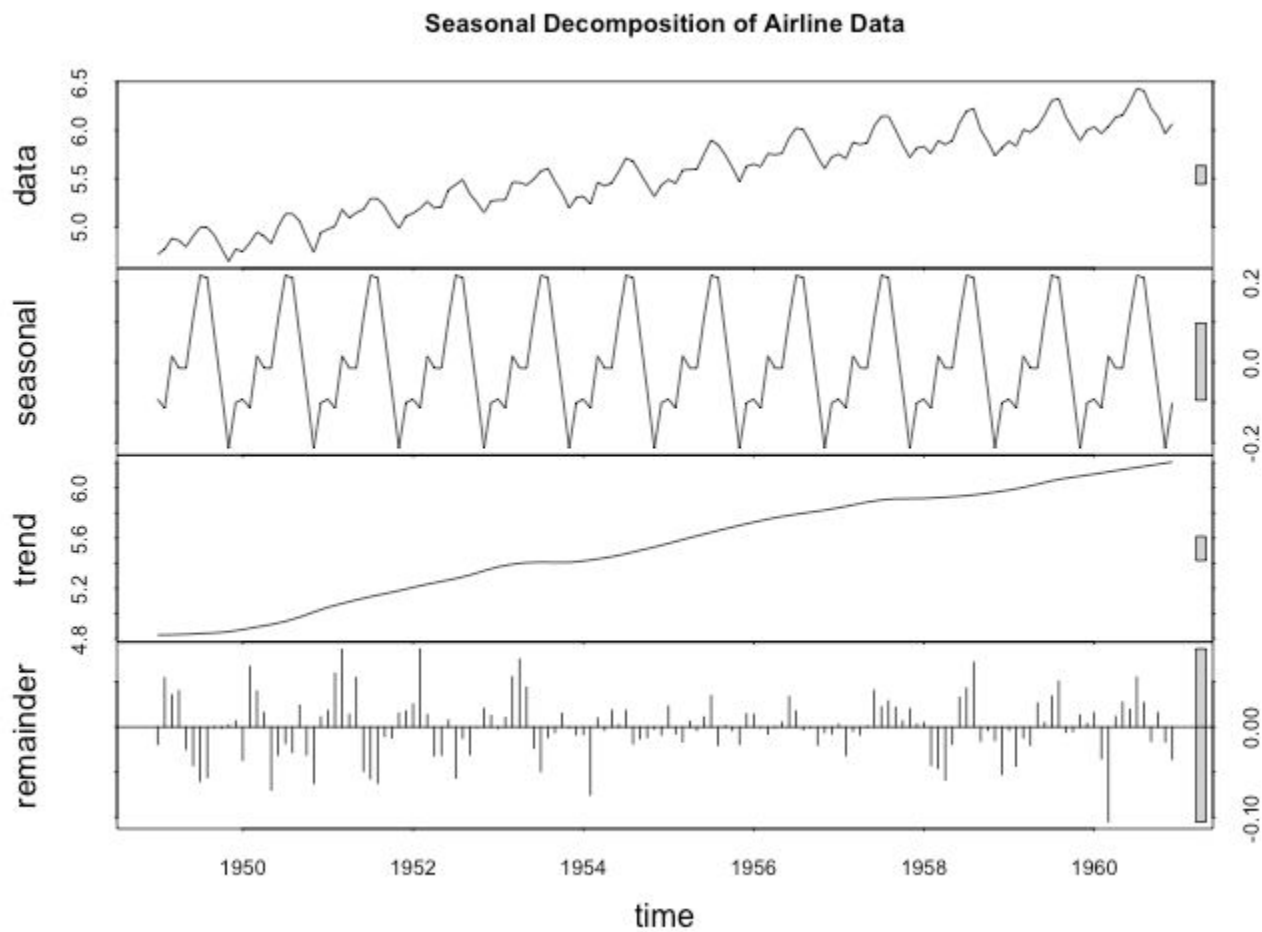
Seasonal Decomposition



Seasonal Decomposition



Seasonal Decomposition



The Idea of White Noise

- ❖ A time series that seems to depict irregularities (as observed within the most recent plot) can also be referred to as **white noise**.
- ❖ White noise describes the assumption that each element in the time series is a random draw from a population with a mean of zero and a constant variance (normally distributed); another term for a time series that is just white noise is a **stationary** series.
 - Thus, time series with trends or seasonality are **not stationary** because the properties of Y_t depend on the time at which they are observed.
- ❖ Before we begin modeling and forecasting, we ideally would want a time series that is stationary in nature; how can we do this?

PART 3

ARIMA Models

ARIMA Models

- ❖ ARIMA stands for **Auto-Regressive Integrated Moving Average**; ARIMA models provide a complicated method for forecasting particularly non-seasonal time series by combining the ideas of multiple methodologies.
 - ARIMA models are also referred to as **Box-Jenkins models**; they were developed by George Box and Gwilym Jenkins.
- ❖ The components of an ARIMA model are:
 - **AR**: The auto-regressive component for lags on the stationary series.
 - **I**: The integrated component for a series that needs to be differenced to become stationary.
 - **MA**: The moving average component for lags of the forecast errors.

The Auto-Regressive Component

- ❖ In an **auto-regressive model of order p** , each value in a time series is predicted from a linear combination of the **previous p values**:

$$AR(p) : Y_t = \mu + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + \epsilon_t$$

- ❖ What makes up the auto-regressive component:
 - Y_t is a given value of the series.
 - μ is the mean of the series.
 - β_i are the coefficients of each lag Y_{t-i} .
 - ϵ_t is the irregular component (error of prediction).
- ❖ The AR model is essentially saying that the **value of a variable at a specific time** is related to the **values of the variable at previous times**.

The Moving Average Component

- ❖ In a **moving average model of order q** , each value in a time series is predicted from a linear combination of the **previous q errors**:

$$MA(q) : Y_t = \mu - \theta_1\epsilon_{t-1} - \theta_2\epsilon_{t-2} - \dots - \theta_q\epsilon_{t-q} + \epsilon_t$$

- ❖ What makes up the moving average component:
 - Y_t is a given value of the series.
 - μ is the mean of the series.
 - θ_i are the coefficients of each error ϵ_{t-i} .
 - ϵ_t is the irregular component (error of prediction).
- ❖ The MA model is essentially saying that the **value of a variable at a specific time** is related to the **residuals of prediction at previous times**.

The Integrated Component

- ❖ The **integrated component** refers to a time series that has been differenced d times; a differenced series represents the change between consecutive observations in the original series:

$$I(1) : Y'_t = Y_t - Y_{t-1}$$

- ❖ Note that as we difference multiple times, instead of differencing prior lags, we **difference the previous difference**. What does this mean?

$$\begin{aligned} I(2) : Y''_t &= Y'_t - Y'_{t-1} \\ &= (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \\ &= Y_t - 2Y_{t-1} + Y_{t-2} \end{aligned}$$

Putting it All Together

- ❖ An $ARIMA(p, d, q)$ model is a model that combines the ideas of each of the previously discussed components, in which:
 - The time series has been differenced d times.
 - Representing the **integrated component** $I(d)$.
 - The resulting values are predicted from the previous p actual values.
 - Representing the **auto-regressive component** $AR(p)$.
 - The resulting values are predicted from the previous q error terms.
 - Representing the **moving average component** $MA(q)$.

Fitting ARIMA(p, d, q) Models: The Procedure

- ❖ The general procedure for fitting an ARIMA(p, d, q) model is as follows:
 1. Ensure that the time series is stationary.
 - a. Use the residual values after detrending using linear regression.
 - b. Use the residual values after seasonally decomposing.
 - c. Possibly difference d times using the integrated component.
 2. Identify a reasonable subset of models.
 - a. Determine possible values of p .
 - b. Determine possible values of q .
 3. Fit the models based on the parameter selections.
 - a. Evaluate the model fit.
 4. Make forecasts with the final selected model.

Ensuring Stationarity: The Augmented Dickey-Fuller Test

- ❖ The **Augmented Dickey-Fuller test** helps us determine whether a model is stationary. Essentially, it boils down to an assessment as to whether or not differencing will help in making the series stationary:
 - Null Hypothesis (H_0): The series **is not stationary**.
 - Alternative Hypothesis (H_A): The series **is stationary**.
- ❖ Should we **retain** the null hypothesis:
 - Difference the series (possibly again) and conduct the Augmented Dickey-Fuller test once more.
- ❖ Should we **reject** the null hypothesis:
 - Conclude that the series is stationary and move forward with the analysis; you have now found d .

Determining Possible Values for p & q : ACF & PACF

- ❖ To help determine possible values for p & q , we need to look at the series **autocorrelation** and **partial autocorrelation** functions.
- ❖ **Autocorrelation AC** measures the way observations relate to each other:
 - $AC(k)$ is the correlation between a set of observations Y_t and the observations k time periods earlier Y_{t-k} .
 - The autocorrelation function $ACF(k)$ computes $AC(1), AC(2), \dots, AC(k)$.
- ❖ **Partial autocorrelation PAC** measures the way observations relate to each other after accounting for all other intervening observations:
 - $PAC(k)$ is the correlation $AC(k)$ with the effects of $Y_{t-1}, Y_{t-2}, \dots, Y_{t-k+1}$ removed.
 - The partial autocorrelation function $PACF(k)$ computes $PAC(1), PAC(2), \dots, PAC(k)$.

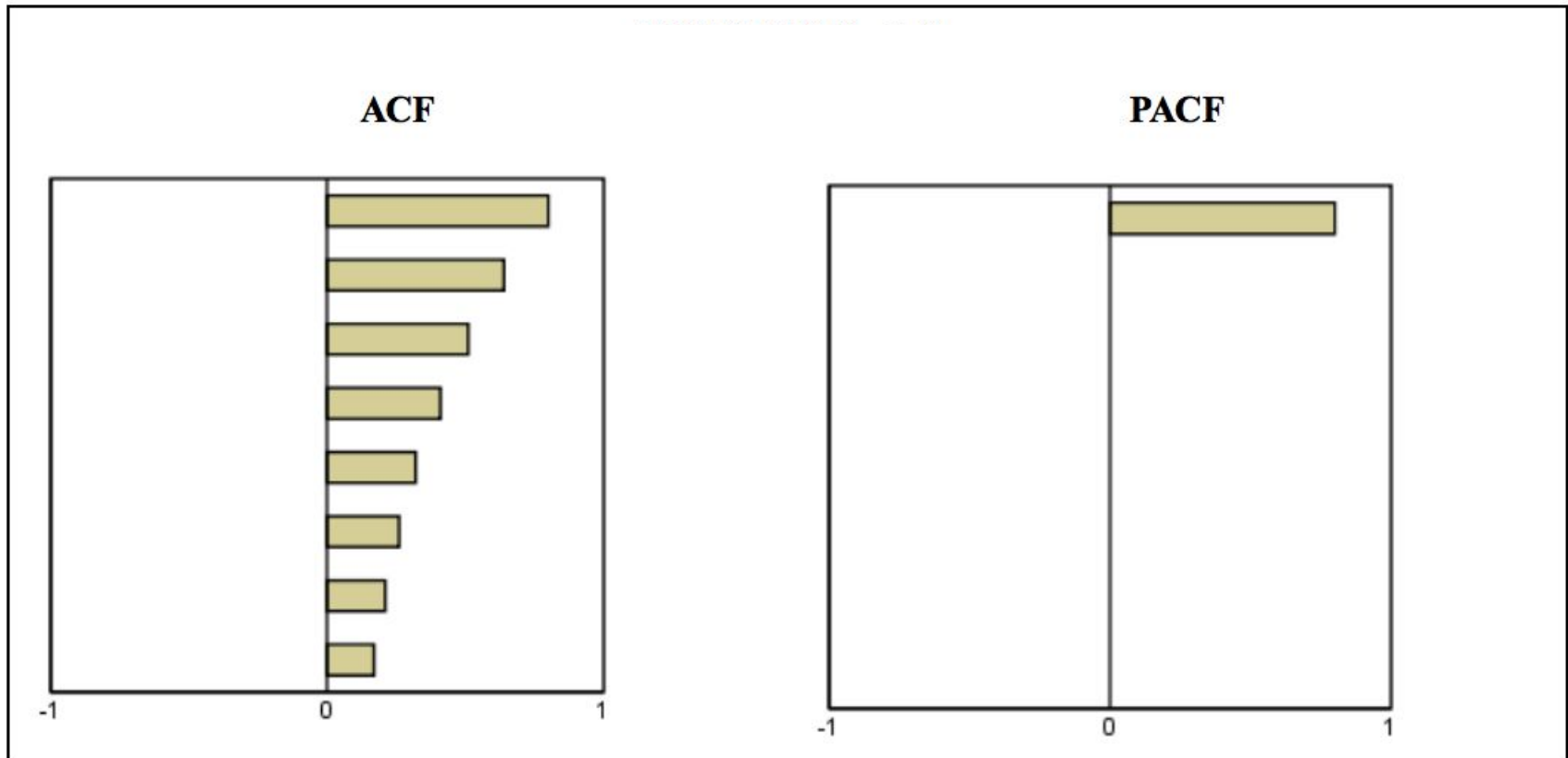
Determining Possible Values for p & q : ACF & PACF

- ❖ A plot of the autocorrelation function ACF displays the correlation of the series with itself at different lags.
- ❖ A plot of the partial autocorrelation function PACF displays the amount of autocorrelation that is not explained by lower order autocorrelations.
- ❖ An inspection of the ACF and PACF plots in tandem will help in determining possible values of p & q .
 - **NB:** This procedure is definitely an art rather than a science; it is a jump-off point to start exploring possible models.

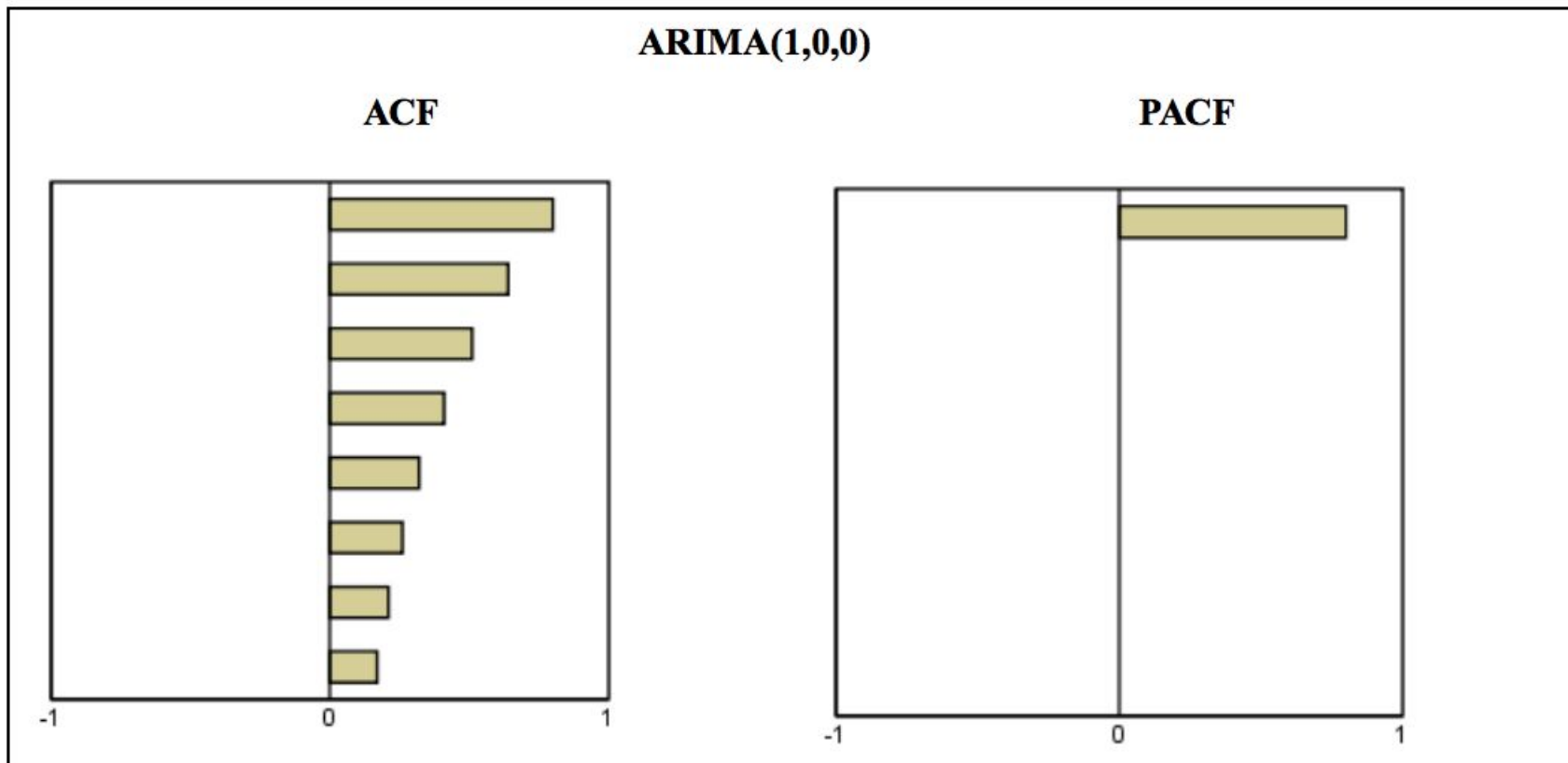
Determining Possible Values for p & q : ACF & PACF

- ❖ Guide to selecting p & q from ACF and PACF plots:
 - AR processes have a quickly decaying ACF with spikes in the first few PACF lags. Choose p as the number of spikes in the PACF.
 - MA processes have a quickly decaying PACF with spikes in the first few ACF lags. Choose q as the number of spikes in the ACF.
 - ARMA processes have a quickly decaying ACF and PACF. Choose p & q in tandem as though the AR and MA processes are independent.
 - In general, do not worry about the sign of the values; we are mostly interested in the magnitude of the correlations.
- ❖ **NB:** If you see an ACF that decays very slowly, this is an indication that you have a **nonstationary series** and should difference the model. Increase the value of d and start the search for p & q over again.

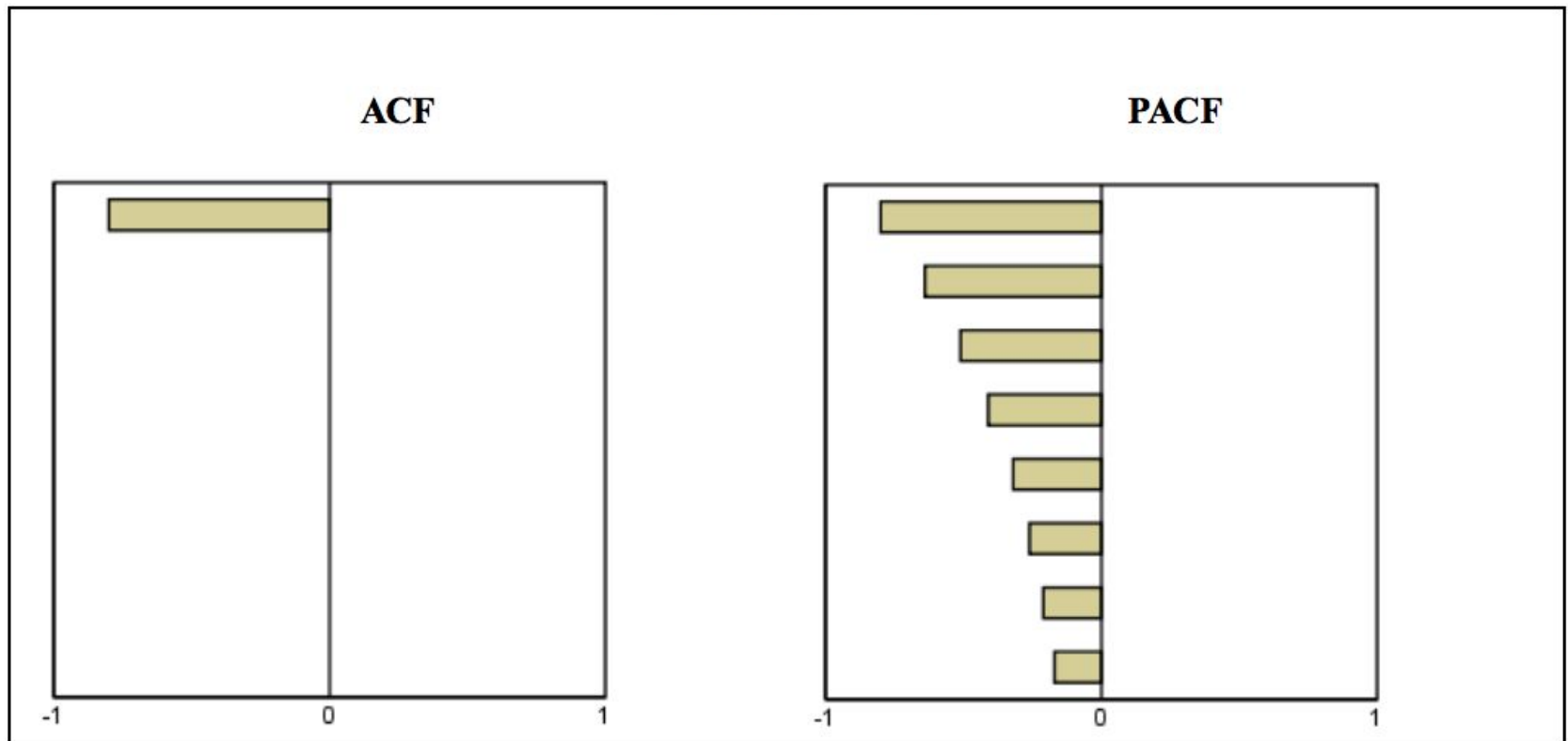
Determining Possible Values for p & q : ACF & PACF



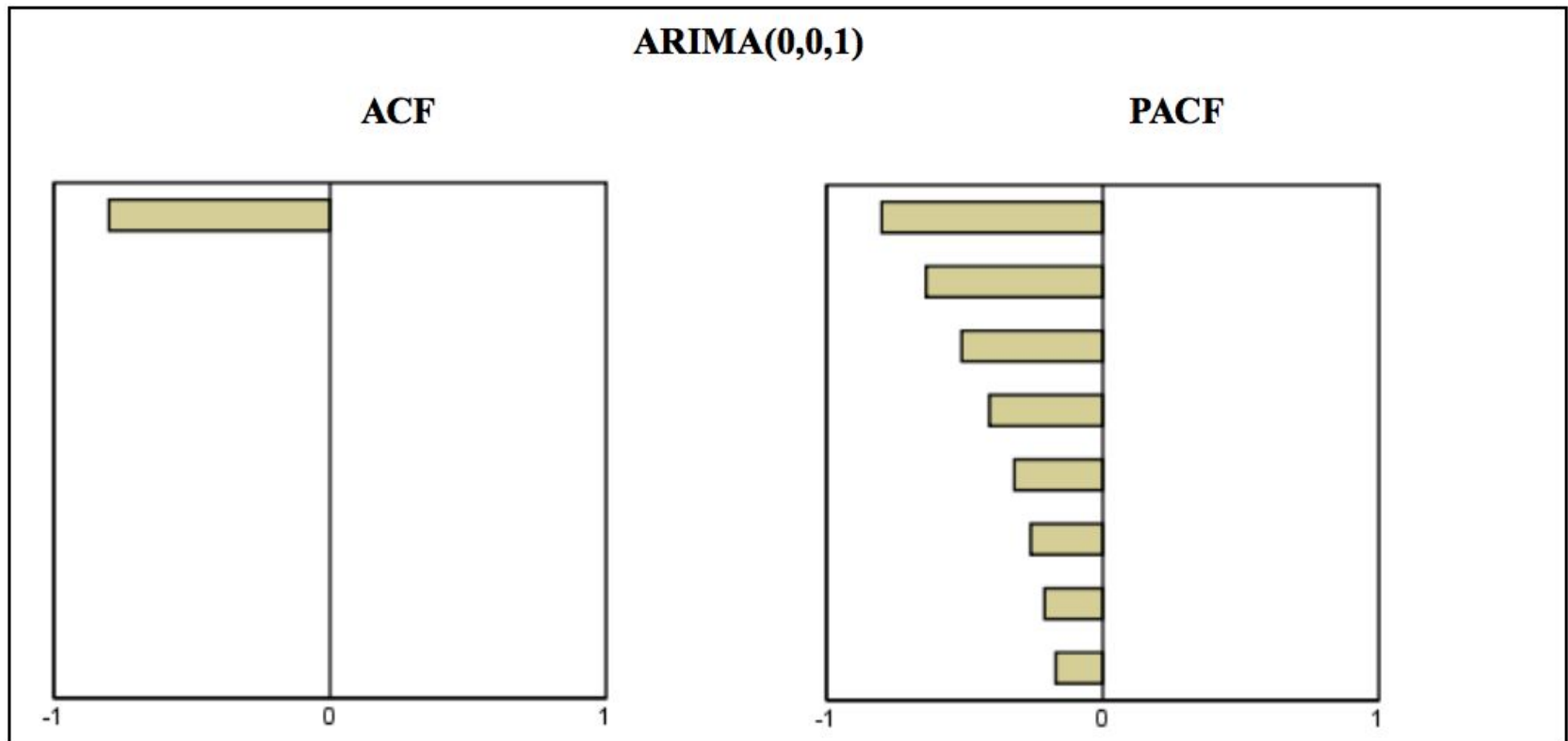
Determining Possible Values for p & q : ACF & PACF



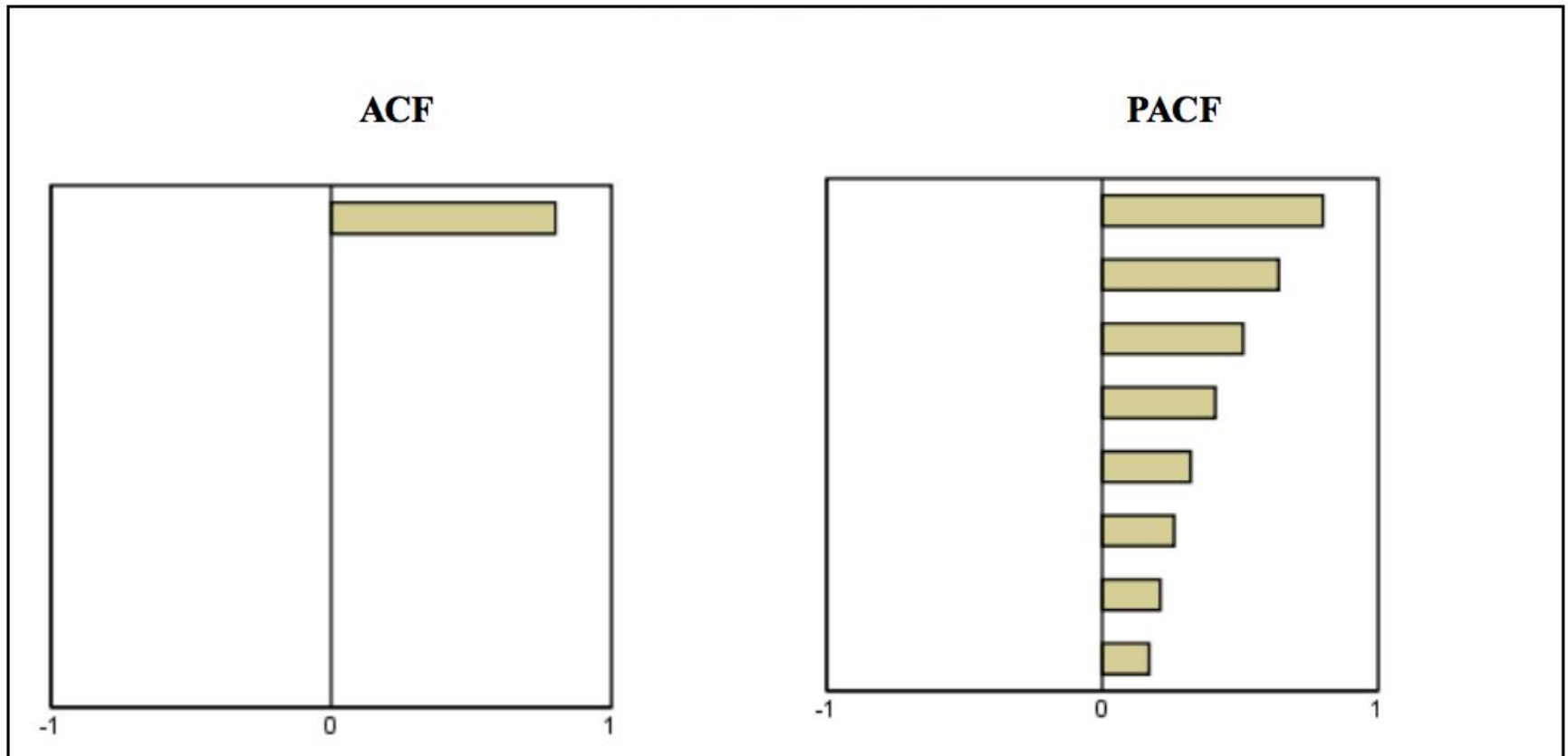
Determining Possible Values for p & q : ACF & PACF



Determining Possible Values for p & q : ACF & PACF



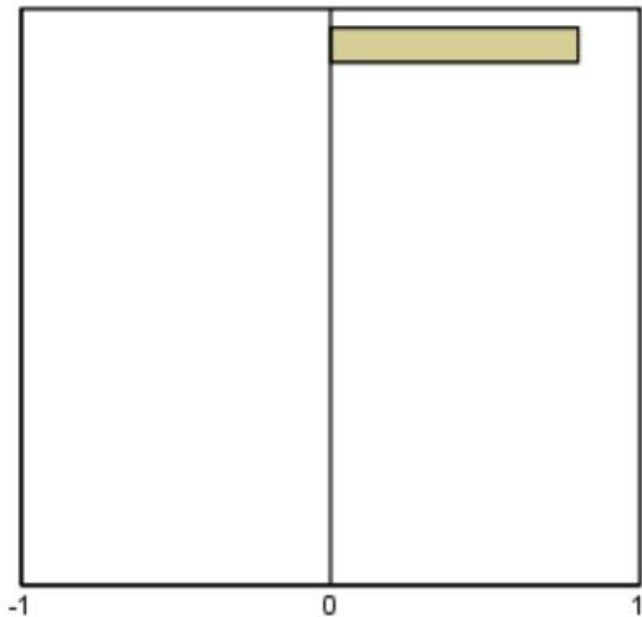
Determining Possible Values for p & q : ACF & PACF



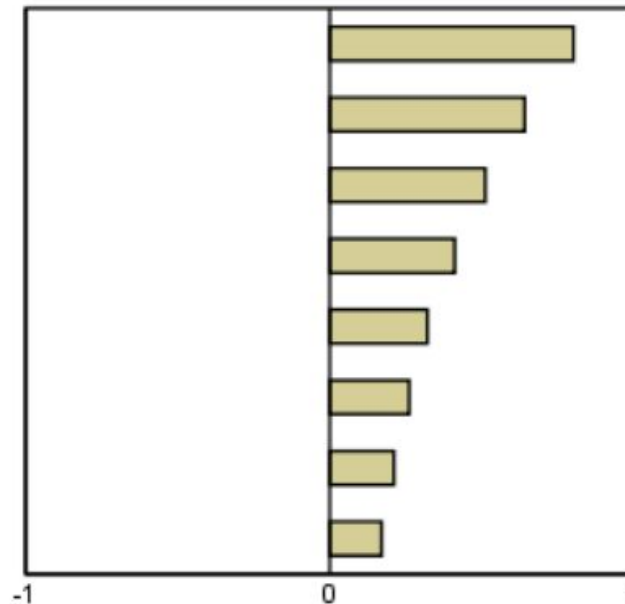
Determining Possible Values for p & q : ACF & PACF

ARIMA(0,0,1)

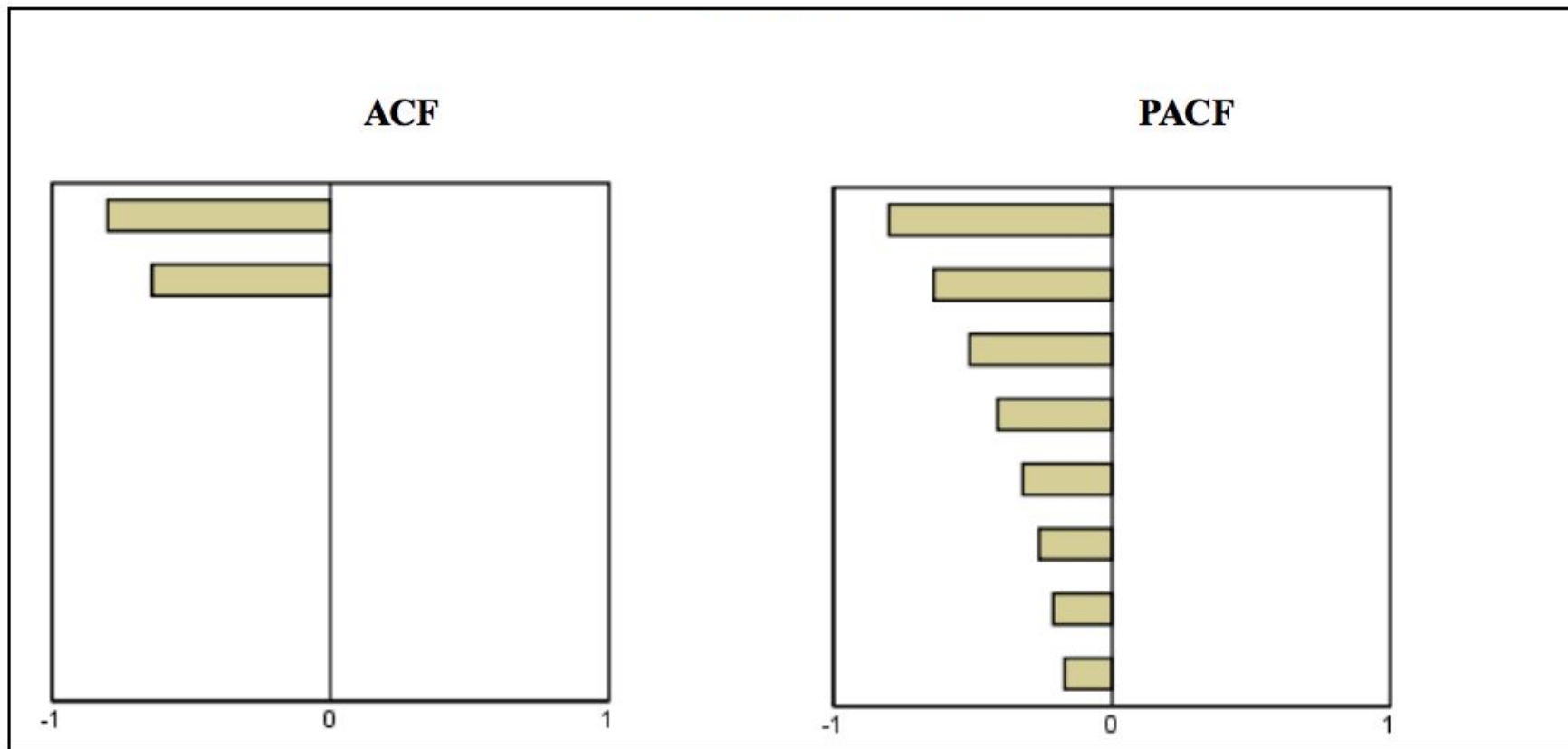
ACF



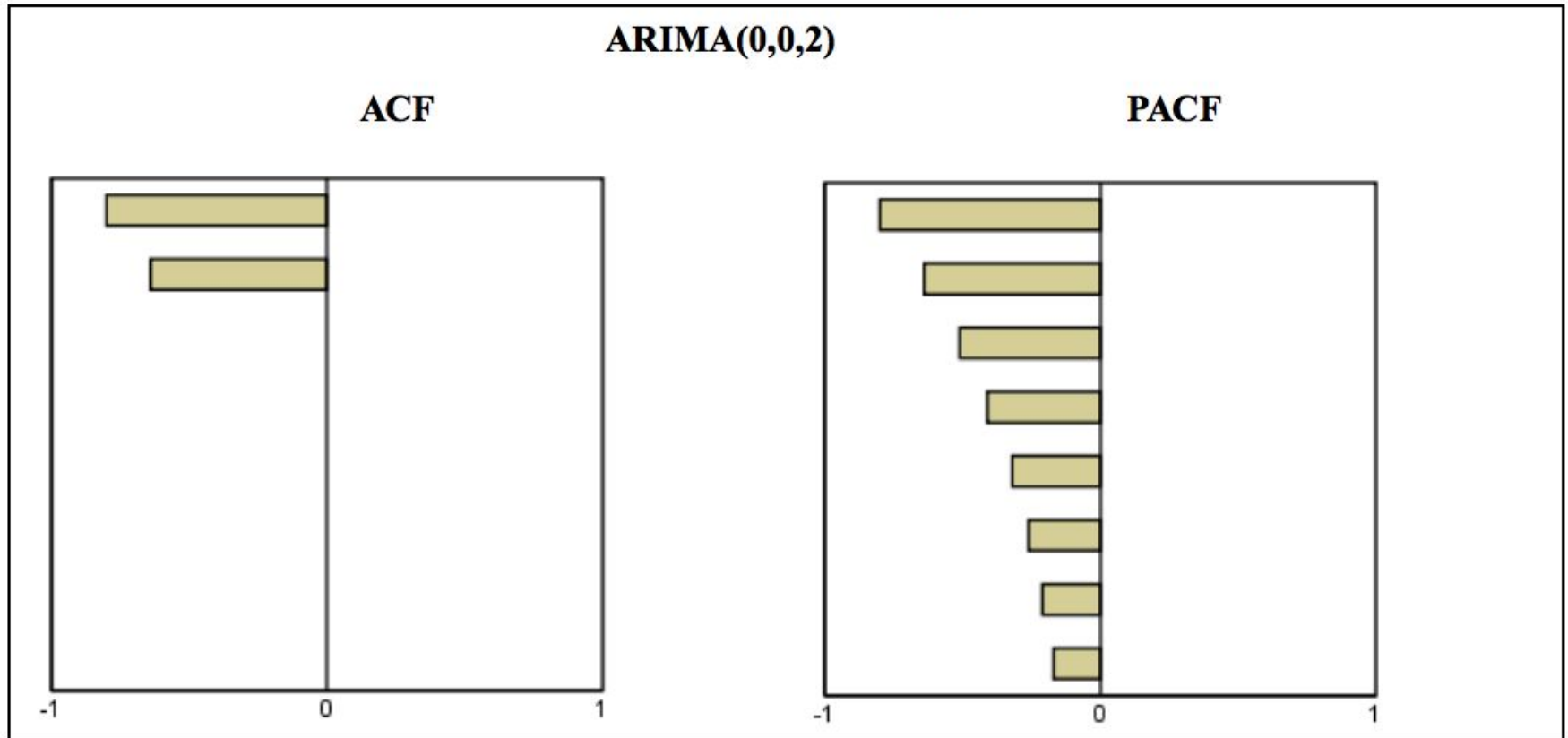
PACF



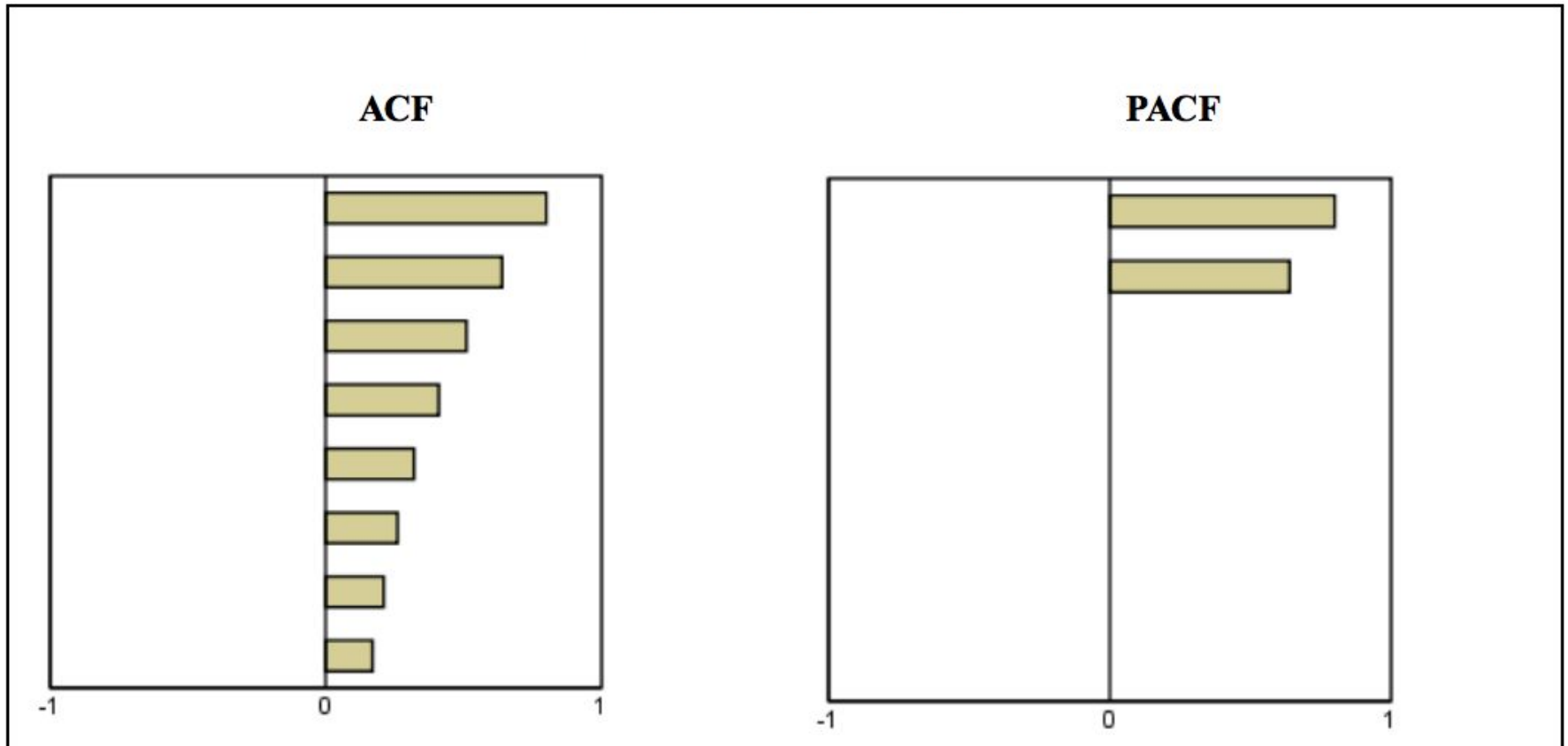
Determining Possible Values for p & q : ACF & PACF



Determining Possible Values for p & q : ACF & PACF



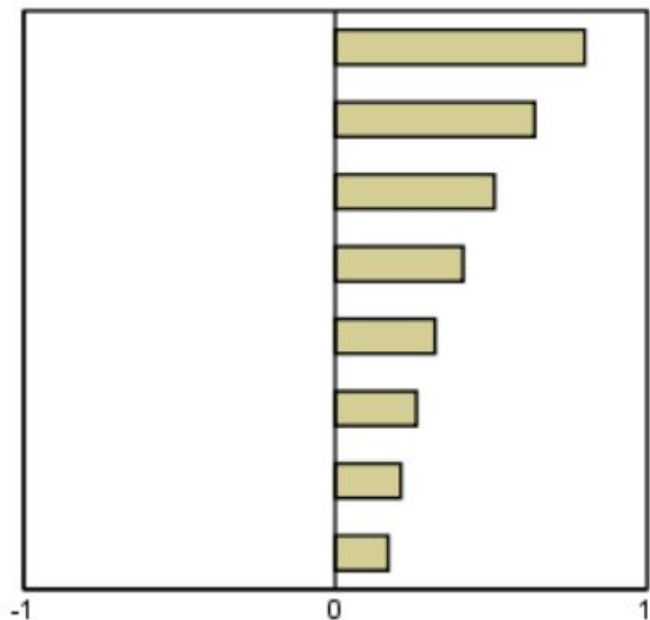
Determining Possible Values for p & q : ACF & PACF



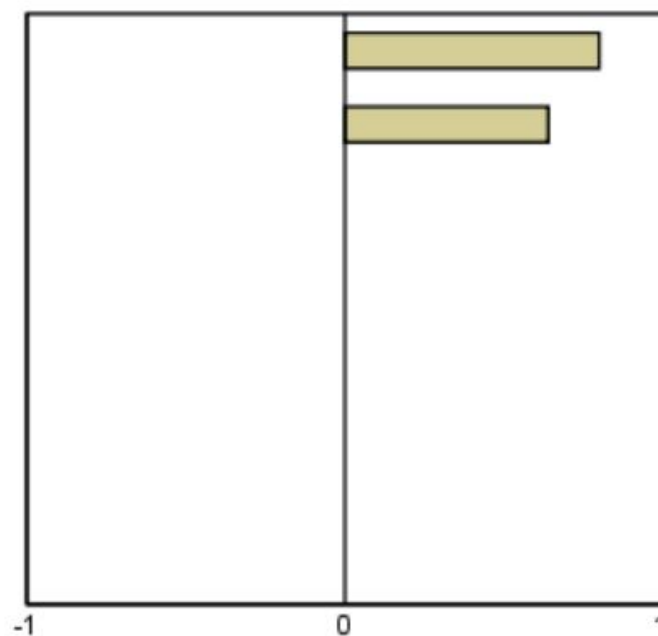
Determining Possible Values for p & q : ACF & PACF

ARIMA(2,0,0)

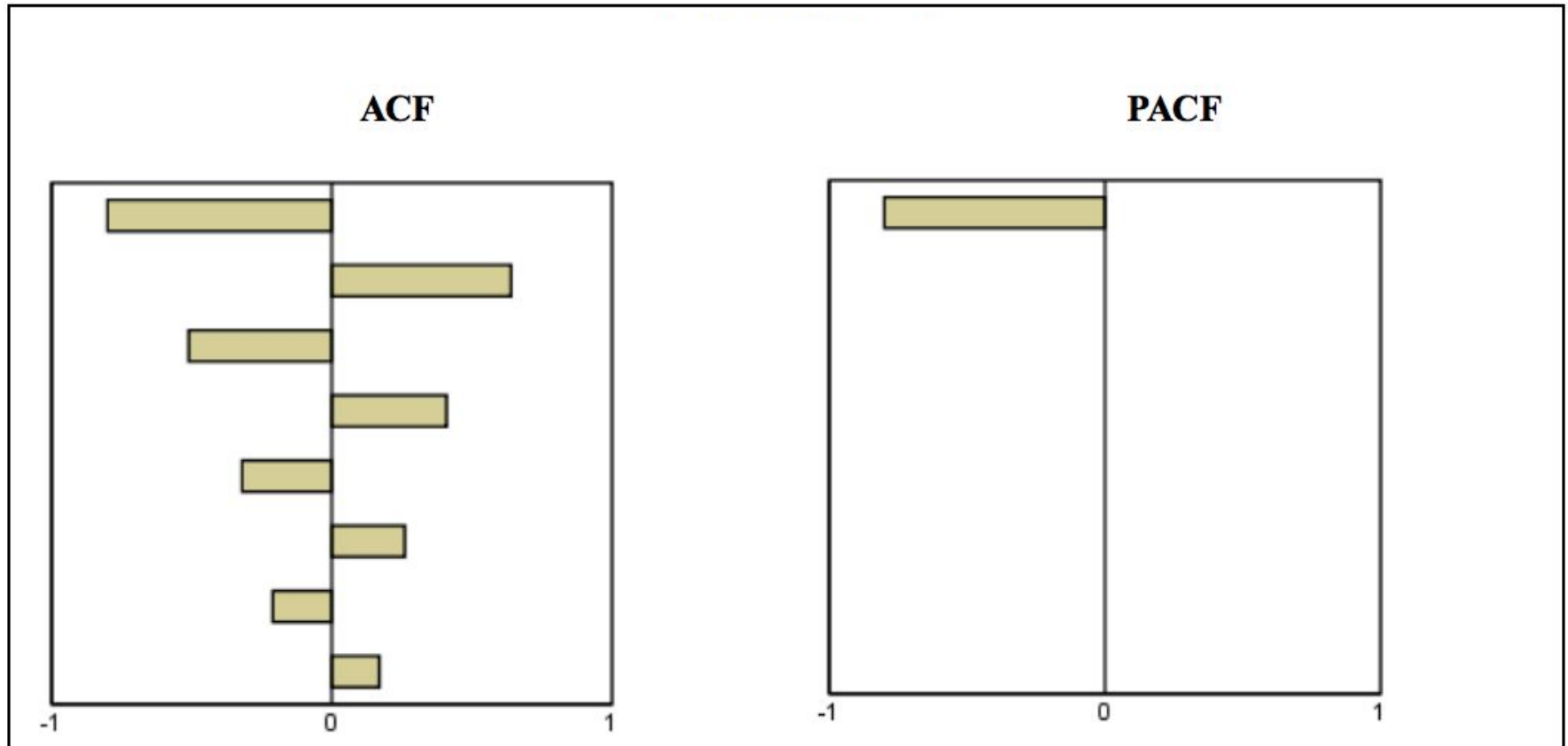
ACF



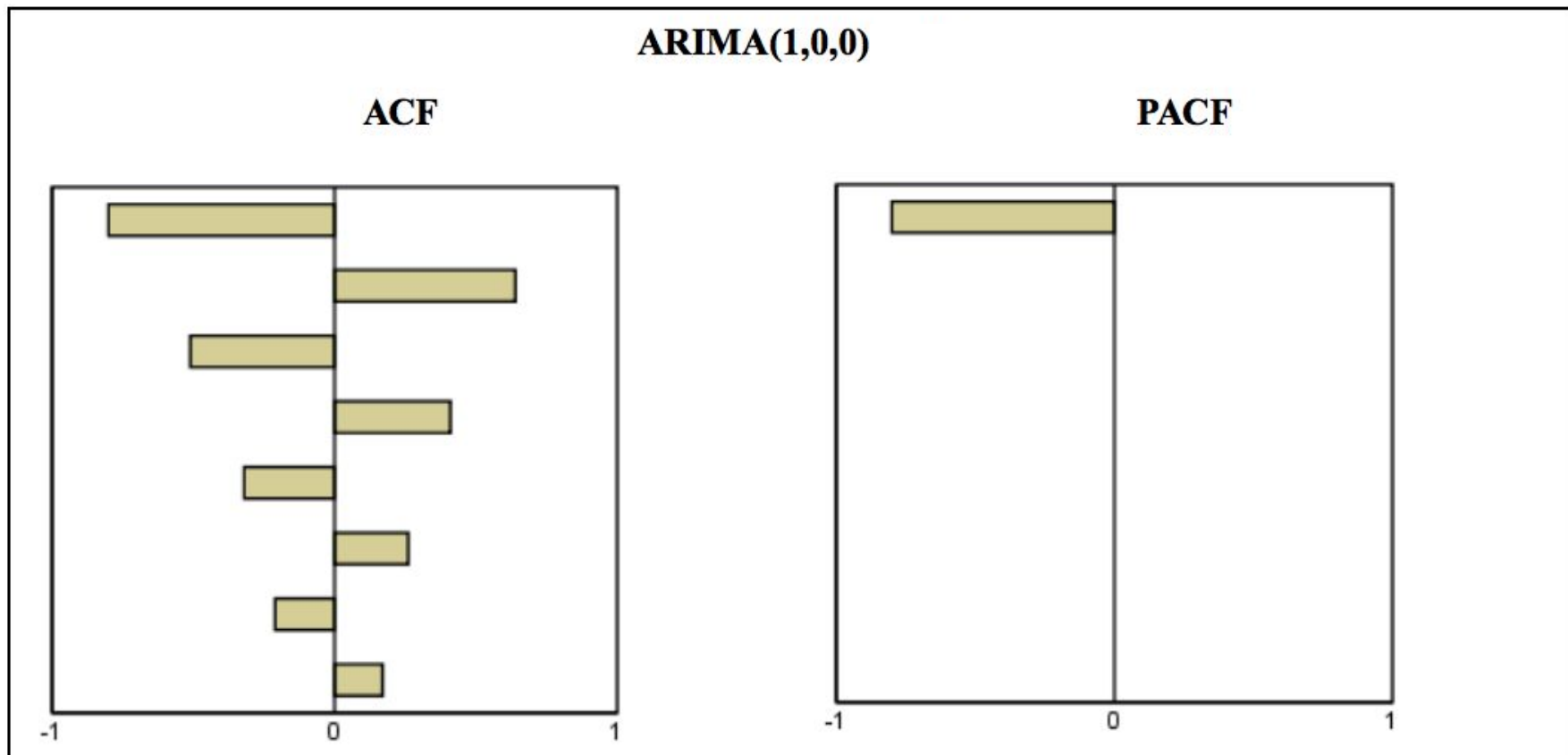
PACF



Determining Possible Values for p & q : ACF & PACF

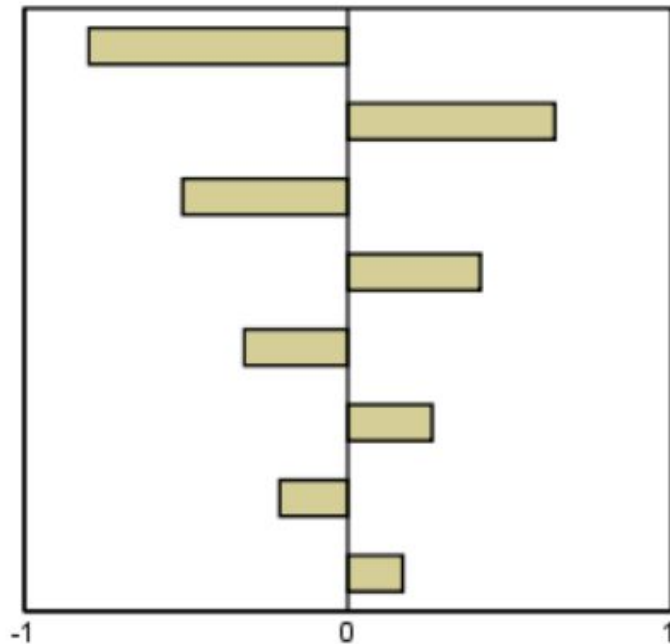


Determining Possible Values for p & q : ACF & PACF

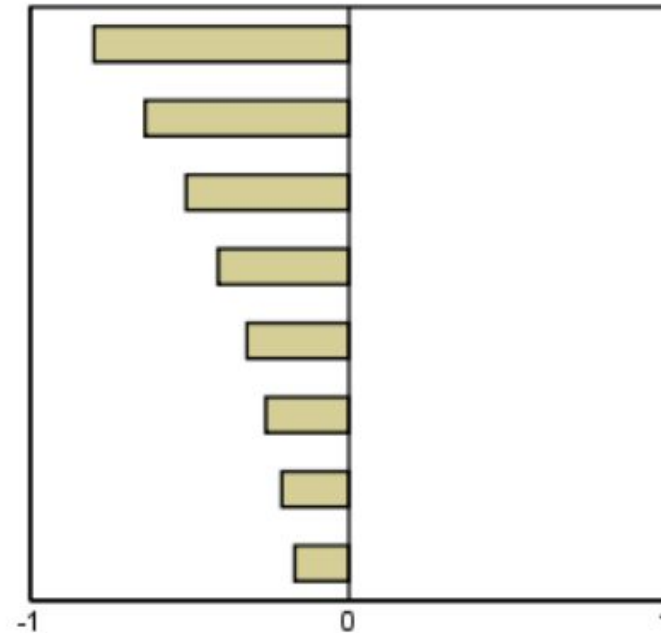


Determining Possible Values for p & q : ACF & PACF

ACF



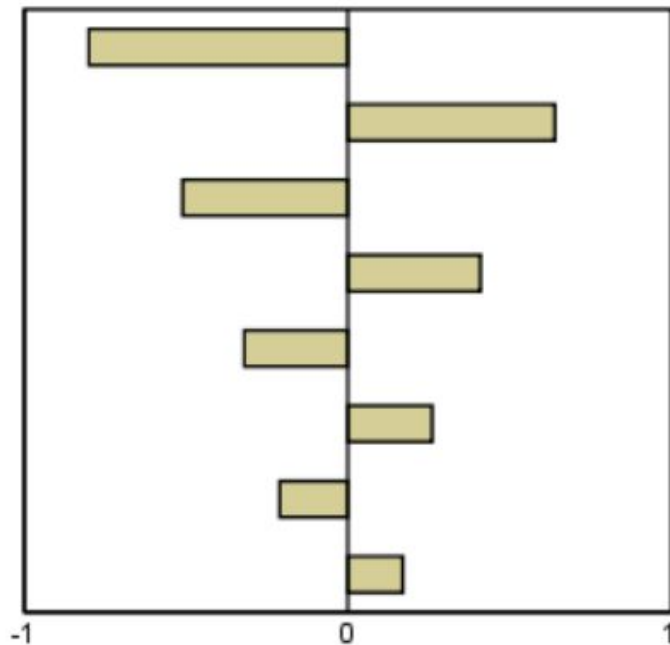
PACF



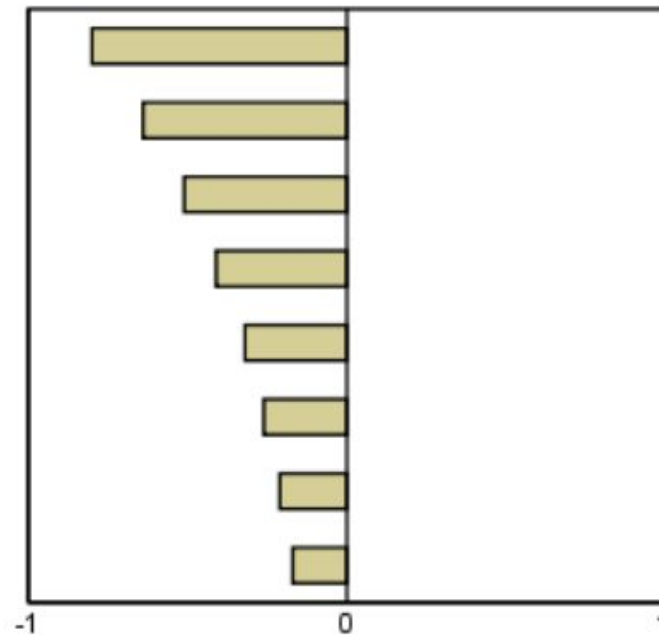
Determining Possible Values for p & q : ACF & PACF

ARIMA(1,0,1)

ACF



PACF



PART 4

Assessing Model Fit

Assessing Model Fit

- ❖ Once we have selected values for p , d , & q , and thus fit an $ARIMA(p, d, q)$ model, we need to [assess the model fit](#).
- ❖ The three main methods for assessing the fit of an $ARIMA(p, d, q)$ model are:
 - Residual analysis
 - The Box-Ljung test
 - Manual overfitting

Assessing Model Fit: Residual Analysis

- ❖ Recall that if the model is appropriate, the **residuals should resemble white noise** and be normally distributed with a mean of 0 and a constant variance.
 - This idea is familiar from linear regression; we can simply check:
 - A scatterplot of the **residuals versus fit** to see if we violate the assumption of **constant variance**.
 - A **QQ plot** to see if we violate the assumption of **normality**.
- ❖ Additionally, we should check the **ACF and PACF of the residuals**; what would we hope to see?
 - The autocorrelations should essentially be zero for every possible lag, indicating that we do not violate the assumption of **independent errors**.

Assessing Model Fit: The Box-Ljung Test

- ❖ The Box-Ljung test takes the residual analysis one step further in assessing whether **all the autocorrelations are zero**; in effect, it tests for whether our series is of white noise.
 - Null Hypothesis (H_0): The autocorrelations are **all 0**.
 - Alternative Hypothesis (H_A): At least one of the autocorrelations is **not 0**.
- ❖ Should we **retain** the null hypothesis:
 - The time series is made up of white noise and we have an indication of a valid model fit.
- ❖ Should we **reject** the null hypothesis:
 - The time series is not made up of white noise and we have an indication of an invalid model fit.

Assessing Model Fit: Manual Overfitting

- ❖ Although we have already decided upon values of p , d , & q , we can continue to “tweak” the model by [attempting to overfit](#).
 - In this case, the process of overfitting is used to ensure that we have not accidentally left any significant terms out.
- ❖ The process of overfitting an ARIMA model:
 1. Fit an extra AR term:
 - a. If the extra AR term is helpful, repeat this step.
 - b. If the extra AR term is not helpful, move forward.
 2. Fit an extra MA term:
 - a. If the extra MA term is helpful, go back to fitting an extra AR term.
 - b. If the extra MA term is not helpful, you have successfully overfit.
- ❖ Compare models based on AIC, BIC, or the p-values for the added terms.

Loose Interpretation of Coefficient Estimates

- ❖ The interpretation for model coefficients in ARIMA is **not the same** as we have seen in the past for generalized linear models.
- ❖ We are not relating Y to a slew of predictor variables X and uncovering some underlying structure -- instead, **the ARIMA model depends on Y itself**.
- ❖ ARIMA models are simply modeling past values of the series plus previous error terms and assuming that they contain information for forecasting purposes; this is often called an **atheoretical model**.
- ❖ Although we don't have the same type of interpretation of coefficient estimates, we can still get a **general understanding** of what the coefficients mean.

Loose Interpretation of Coefficient Estimates

- ❖ An ARIMA model with significant AR terms is a signature of a time series that feels a **restoring force** that tends to pull the series back towards the general mean.
- ❖ Consider a basic AR(1) model; the coefficient for the AR(1) term is an indicator of how fast the series is restored towards its average value:
 - Suppose that the coefficient is **relatively large (near 1)**. This would mean that the series **returns to the mean slowly** because there are long-lasting effects of previously observed values.
 - Suppose that the coefficient is **relatively small (near 0)**. This would mean that the series **returns to the mean quickly** because the effects of previously observed values taper off quickly.

Loose Interpretation of Coefficient Estimates

- ❖ An ARIMA model with significant MA terms is a signature of a time series that can feel **shocks** where the effects are felt in consecutive periods.
- ❖ Consider a few basic MA models:
 - The coefficient for the MA(1) term is an indicator of how much the shock of **the previous time period** is still retained in the current time period.
 - The coefficient for the MA(2) term is an indicator of how much the shock of **two previous time periods** is still retained in the current time period.
- ❖ **NB:** In general, an AR term describes more about the time series than an MA term. Why is this?

PART 5

Review

Review

❖ Part 1: The Nature of Time Series

Data

- Cross-Sectional, Longitudinal, & Time Series Data
- The Goal of Time Series Analysis
- Applications of Time Series Analysis
- Why Can't We Use Linear Regression?

❖ Part 2: Decomposition of Time Series

Data

- Basic Components of a Time Series
- Description: What Happened in the Past?
- Smoothing for General Trends
- Centered Moving Averages
- Seasonal Decomposition
- The Idea of White Noise

❖ Part 3: ARIMA Models

- The AR Component
- The MA Component
- The I Component
- Putting it All Together
- Fitting ARIMA(p, d, q) Models: The Procedure
- Ensuring Stationarity: The Augmented Dicky-Fuller Test
- Determining Possible Values for p & q : ACF & PACF

❖ Part 4: Assessing Model Fit

- Residual Analysis
- The Box-Ljung Test
- Manual Overfitting
- Loose Interpretation of Coefficient Estimates