

Unit 7 Live Session

Time Series Analysis Lecture 2: Regression with Time Series, An Intro to Exploratory Time Series Data Analysis, and Time Series Smoothing



Figure 1: South Hall

Class Announcements

- HW 7 is this week
- Teams for Lab-2 have been created.

Roadmap

Rearview Mirror

- Notion of stationarity, ergodicity, and dependency
- Basic Time Series models

Today

- Time series decomposition
- Time series smoothing methods
- Time series decomposition and forecasting using OLS

Looking Ahead

- Autoregressive Moving Average Models (ARMA)
- Autoregressive Integrated Moving Average (ARIMA)

Start-up Code

```
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=80),tidy=TRUE)

# Load required libraries
## Load a set of packages including: broom, cli, crayon, dbplyr , dplyr, dtplyr,forcats,
#googledrive, googlesheets4, ggplot2, haven, hms, httr, jsonlite, lubridate , magrittr,
#modelr, pillar, purrrr, readr, readxl, reprex, rlang, rstudioapi, rvest, stringr, tibble,
#tidyverse
## To laod All data sets in the book "Forecasting: principles and practice"
#by Rob J Hyndman and George Athanasopoulos
library(fpp3)
library(fpp2)
## to create a infrastructure for tidy temporal data
library(tsibble)
### to work with date
library(lubridate)
## to use gg_season
library(feasts)
## Forecasting Models for Tidy Time Series
library(fable)
## To assemble multiple plots
library(gridExtra)
## for simulations
library(simts)
## To use TeX() to write expression in the title of plots
library(latex2exp)
```

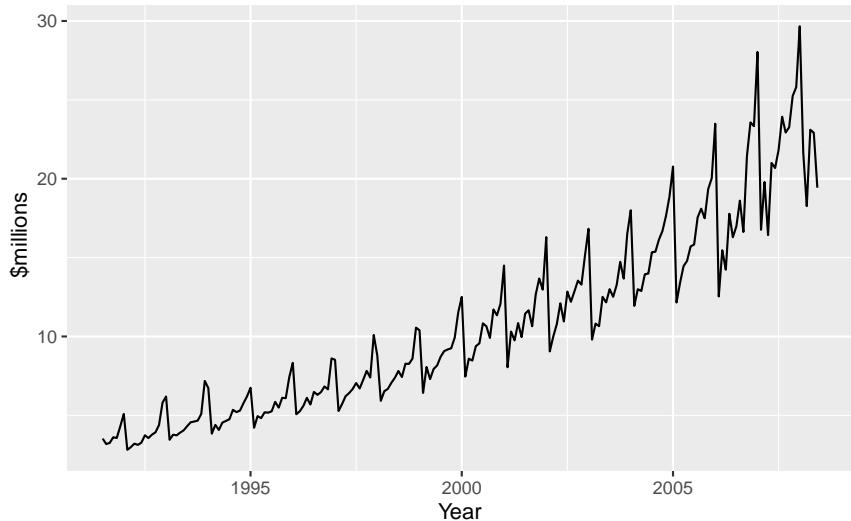
Time series decomposition

Many time series include some or all of the following components.

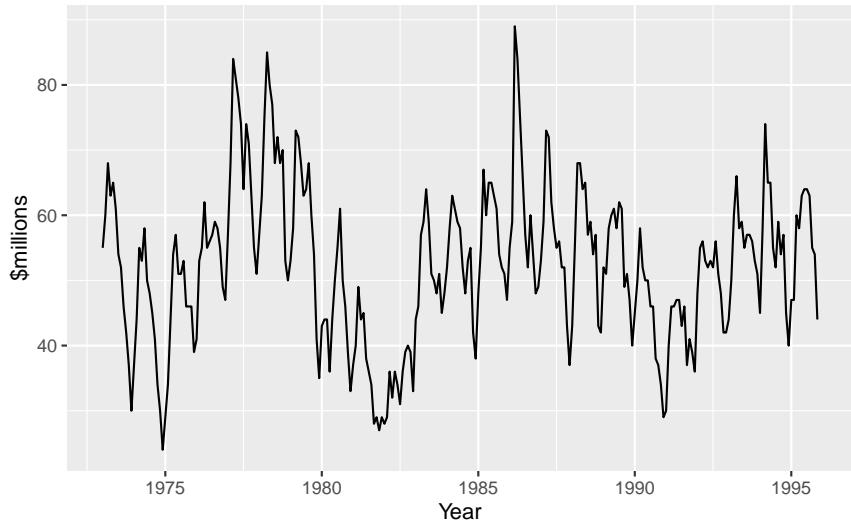
- **Trend:** It is a long-term upward or downward movement of the data. The trend could be linear or nonlinear, and sometimes the trend might change over time.
- **Seasonal:** It is a regular movement in the data based on the season (e.g., every month/quarter/year). Seasonality is always of a fixed and known period.
- **Cyclical:** It is an oscillatory movement of a time series. Its length is not fixed, and it is usually more than one year. In practice, the trend component is assumed to also include the cyclical component.
- **Irregular component** - a stationary process, such as white noise that is random.

a)- Given the above definitions, what are the possible components of the time series in the following graphs?

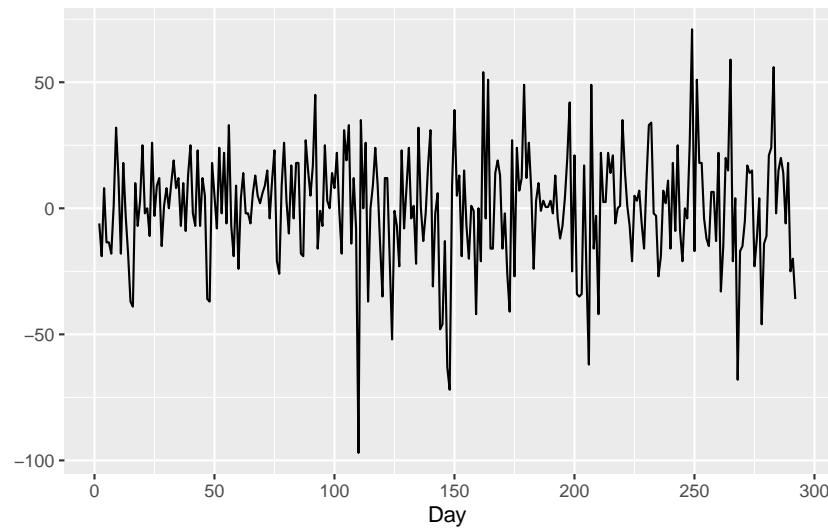
Monthly Antidiabetic drug sales



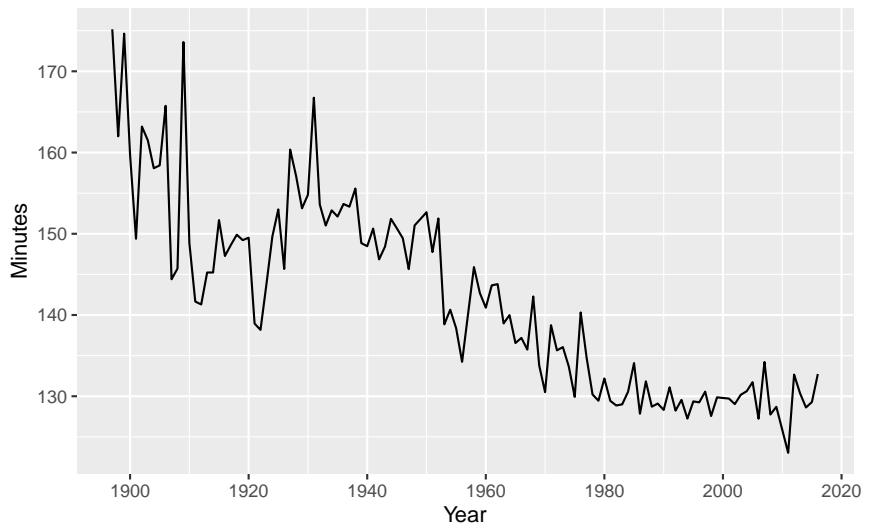
Monthly housing sales



Daily change in Dow Jones index



Winning time for the Boston Marathon



Time series decomposition: Mathematical representation

- To remove the deterministic components, we can decompose our time series into separate stationary and deterministic components.
- There are two common mathematical forms for decomposition:

1- Additive:

$$Y_t = T_t + S_t + E_t$$

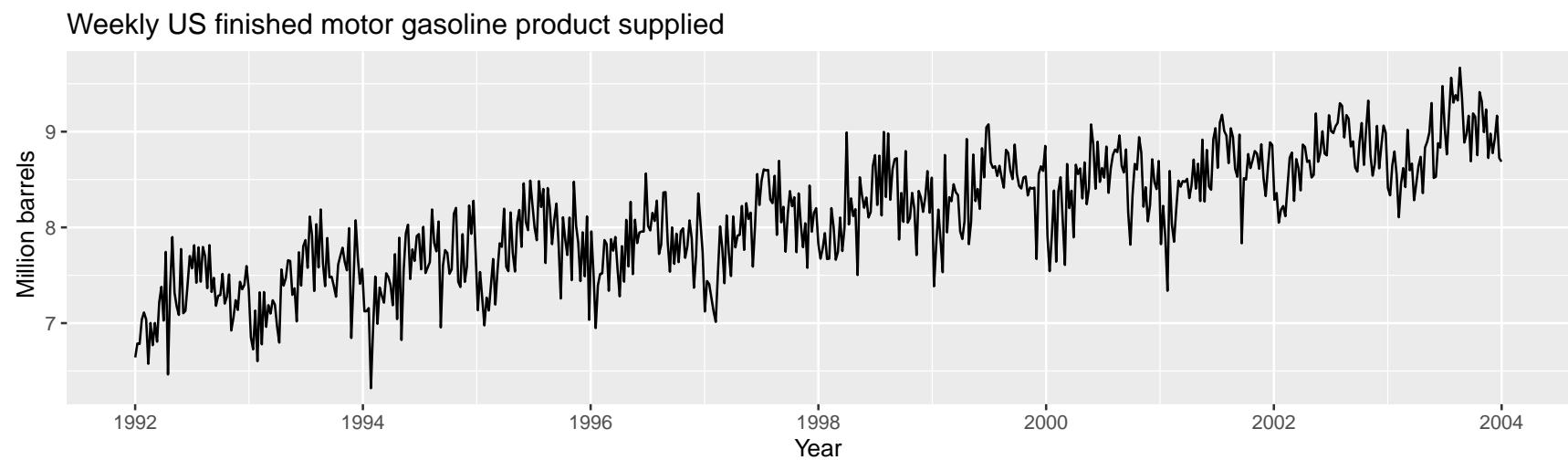
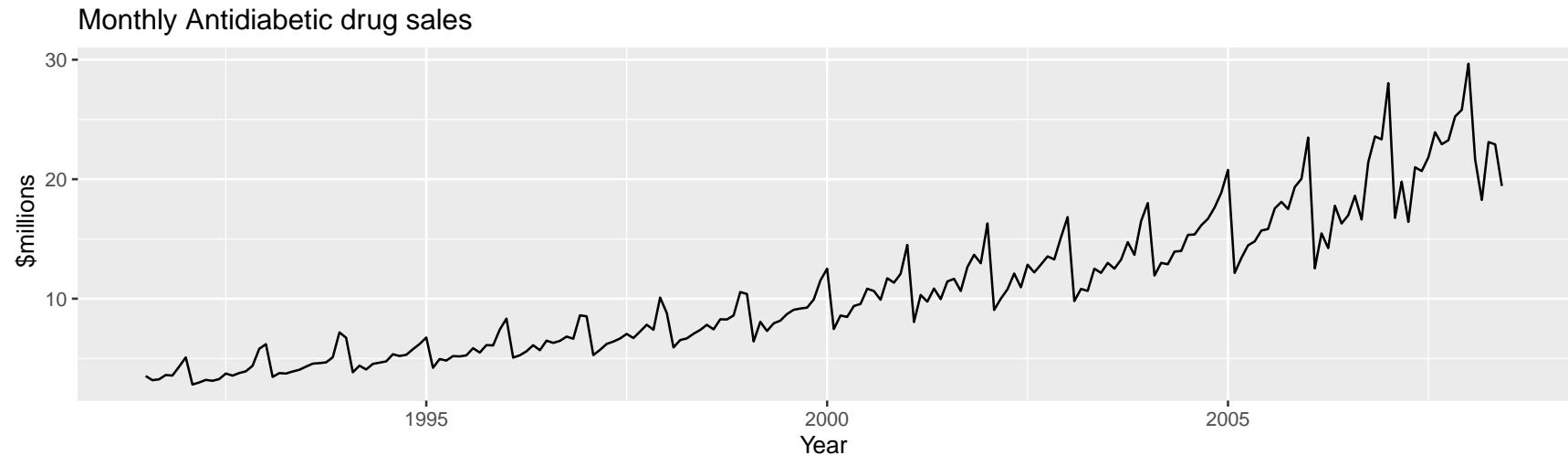
2- Multiplicative:

$$Y_t = T_t \cdot S_t \cdot E_t$$

- We call Y_t a trend stationary time series, if after removing the deterministic part, what remains E_t is a stationary series

a)- Which type of decomposition might be appropriate for the following time series? Why?

b)- Is there any transformation that makes multiplicative decomposition simpler?



Time series decomposition: Trend and seasonal components detection

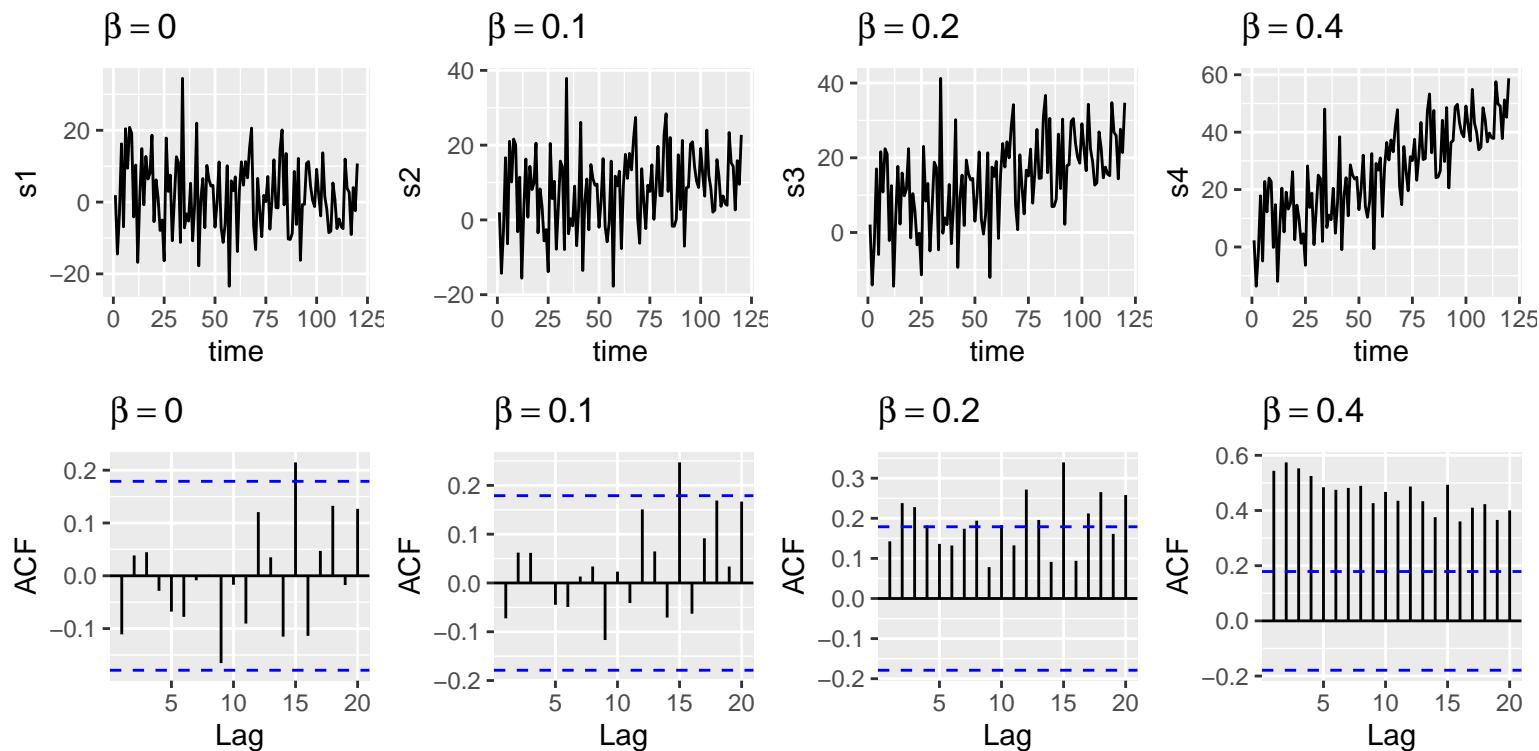
- The primary use of the ACF is to detect autocorrelation in the time series after we have removed the deterministic movement for the time series. ACF is also a helpful tool to determine if a time series has a trend or seasonal movement.
- The following plots display the simulations of X_t for different values of β :

$$X_t = \beta t + W_t$$

- β is the slope over time and measures the change in Y_t over time.

- $\{W\}_{t=-\infty}^{\infty}$ is a white noise process with $E(W_t) = 0$ and $E(W_t^2) = \sigma^2$.

a)- How do the ACF and plot change when the trend becomes more pronounced?

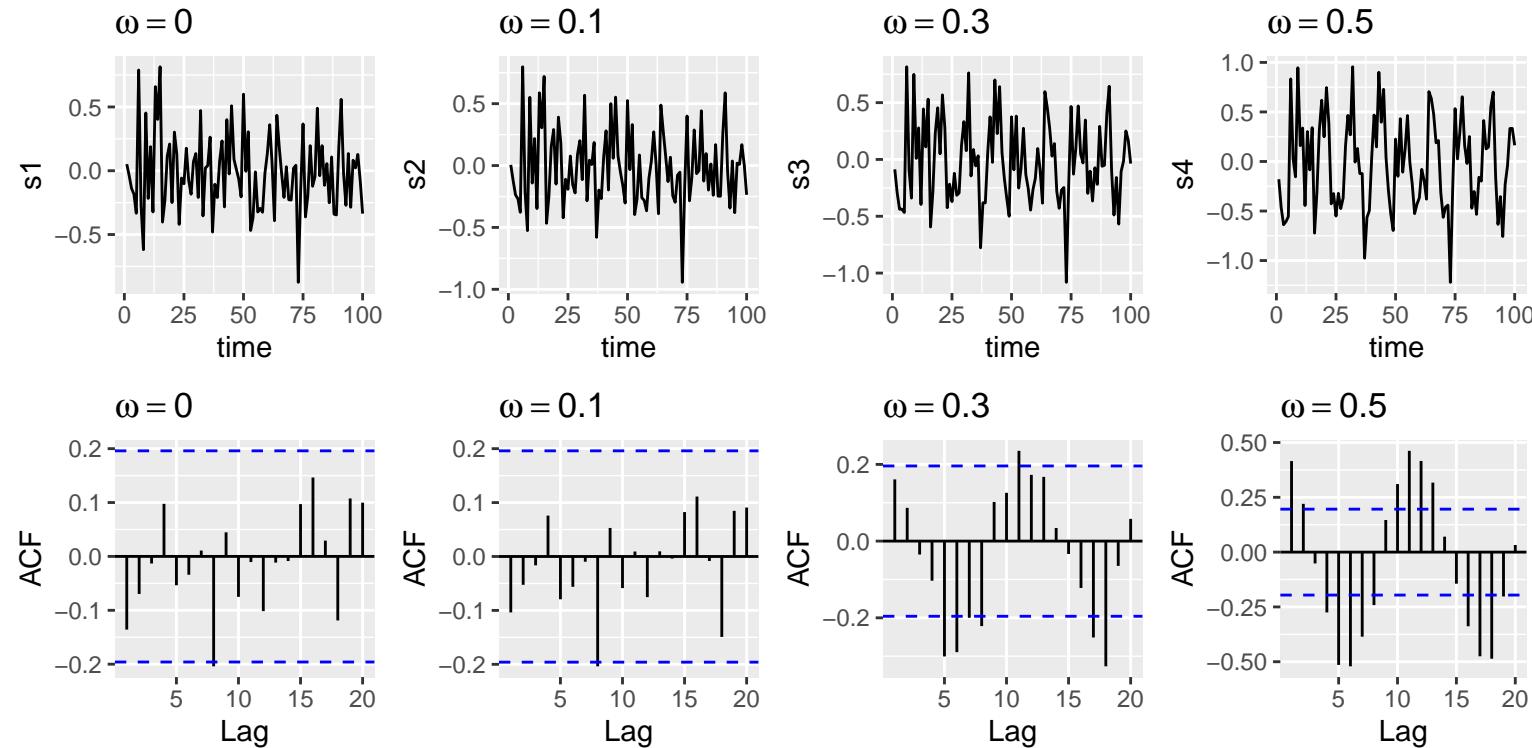


- The following time series process is from simulations of Y_t for different values of ω :

$$Y_t = \omega S_t + W_t$$

- ω measures the strength of the seasonal movement in Y_t over time.

b)- How do the ACF plots change with more pronounced seasonality?

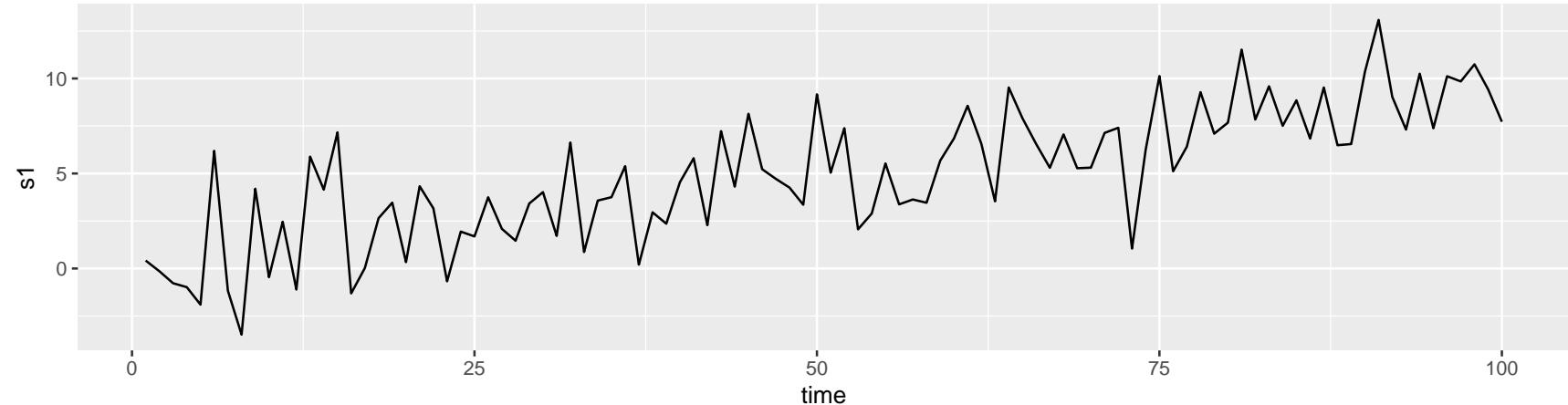


- The following plot shows a simulation of Y_t with both trend and seasonal components:

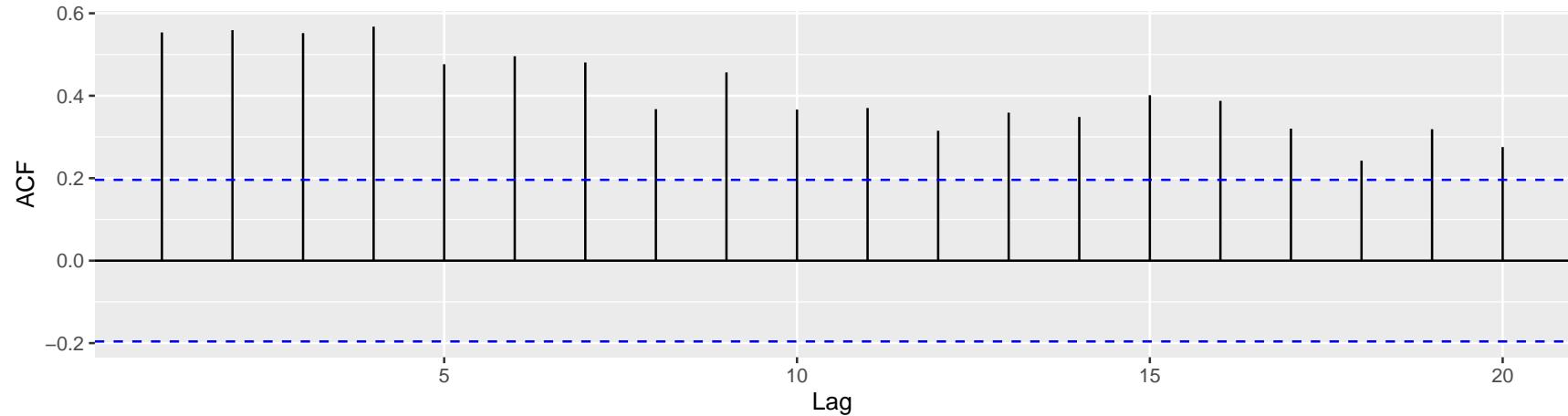
$$Y_t = T_t + S_t + W_t$$

c)- What pattern does the ACF plot exhibit?

$$y_t = 0.1 * \text{time} + 0.1 * \text{season} + w_n$$



$$y_t = 0.1 * \text{time} + 0.1 * \text{season} + w_n$$



Time series decomposition: Estimation

There are two approaches for estimating the deterministic movement in a time series:

- 1- Smoothing procedures
- 2- Linear regression

- What are the advantages and disadvantages of these two methods?

Time series decomposition using smoothing procedures

- Estimation of the deterministic component using a smoothing technique includes the following steps:

1- Estimate the trend using a smoothing procedure such as a moving average

2- De-trending the time series

- By subtracting the trend estimates from the time series for an additive decomposition
- By dividing the time series by the estimated trend values for a multiplicative decomposition

3- Estimating the seasonal factors from the de-trended series

- By calculating the mean (or median) values of the de-trended series for each specific period

4- Normalize the seasonal effects

- For an additive model, seasonal effects are adjusted so that the average of seasonal components is 0
- For a multiplicative model, the seasonal effects are adjusted so that they average 1

5- Calculate the irregular component

- For an additive model $\widehat{E}_t = Y_t - \widehat{T}_t - \widehat{S}_t$
- For a multiplicative model $\widehat{E}_t = \frac{Y_t}{\widehat{T}_t \cdot \widehat{S}_t}$

6- Analyze the residual component for stationarity: Decomposition aims to produce a stationary residual

- The decomposition using smoothing techniques is usually quite successful at describing the time series in question. However, if they do not create an analytic expression for trend and seasonal parts like a moving average, they cannot easily be used for forecasting.
- One exception is Exponential smoothing which could be used either to produce smoothed data for presentation or to make forecasts.

Time series decomposition using linear regression

- Linear regression can be used to estimate trend or seasonality terms in the following steps:

1- Estimate the trend or seasonal movement or both with the following specifications:

- A Linear trend: $Y_t = \beta_0 + \text{beta}_1 \cdot t + W_t$
- A quadratic trend: $Y_t = \beta_0 + \text{beta}_1 \cdot t + \beta_2 \cdot t^2 + W_t$
- A seasonal movement: $Y_t = \beta_0 + \sum_{i=1}^{s-1} \beta_i \cdot S_{it} + W_t$
- A linear trend with seasonal movement: $Y_t = \beta_0 + \beta_1 \cdot t + \sum_{i=2}^{s-1} \beta_i \cdot S_{it} + W_t$
- **We can include higher polynomial trends but it is not recommended for forecasting due to the risk of overfitting.**

2- Analyze the residual component:

- If it is white noise, we can use the estimated model for description and prediction.
- If it is not white noise but stationary, we can use a model to fit the stationary residuals, such as ARMA models.

3- If we have a few competing trend specifications, the best one can be chosen by AIC, BIC, RMSE (Root mean square error), or similar criteria.

4- Finally, forecasting can be achieved by forecasting the residuals and combining them with the forecasts of the trend and seasonal components.

Time series decomposition using both smoothing techniques and linear regression

- We can also combine smoothing techniques with linear regression in the following steps:

1- Estimate the trend, \hat{T}_t^1 via a smoothing method

2- Estimate and normalize the seasonal factors, \hat{S}_t , from the de-trended series

3- Deseasonalize the original data by removing the seasonal component $\hat{Y}_t = Y_t - \hat{S}_t$

4- Reestimate the trend, \hat{T}_t^2 from the deseasonalized data using a (polynomial) regression

5- Analyse the residuals $E_t = Y_t - \hat{S}_t - \hat{T}_t^2$ to verify if they are stationary and specify their model (if needed)

6- Forecast the series Y_{T+h} . Remember that $\hat{S}_t = \widehat{S_{t+d}}$ means that we can always forecast the seasonal component.

Differencing to de-trend and deseasonalize a time series

- Instead of using the smoothing technique or OLS regression to remove the deterministic movement in a time series, we can use differencing:

1- To remove a linear trend:

$$\nabla y_t = y_t - y_{t-1}$$

2- To remove a polynomial trend of degree k :

$$\nabla^k y_t = \nabla^{k-1}(y_t - y_{t-1}) = \nabla^{k-1}y_t - \nabla^{k-1}y_{t-1}$$

3- To remove a seasonal movement where $S_{t-d} = S_t = S_{t+d}$:

$$\nabla_d y_t = y_t - y_{t-d}$$

4- To remove both trend and seasonal movement, we need to apply both a non-seasonal first difference and a seasonal difference

- Usually, the differences of order 1 or 2 are enough for removing a trend, and for seasonality, differences of order one are sufficient
- By differencing the data, our sample size is reduced.
- The interpretation also changes since we are now working with differences ∇y_t , rather than levels of Y_t

a)- Use the appropriate type of difference to completely remove the deterministic movement in Y_t .

$$Y_t = \beta_0 + \beta_1 \cdot t + S_t + W_t$$

- Where $S_t = S_{t+d}$

Case Study: Airline passenger bookings in the U.S.

Introduction

An airline company usually needs to predict future demand before ordering new aircraft. The data science team obtained the latest U.S. carrier, foreign carrier, and individual airport passenger and flight data from the Bureau of Transportation Statistics (BTS). Their goal is to:

- **Describe how to model demand over time and predict future demand before ordering new aircraft**

Data Description and wrangling

The data set “Passengers_2019” contains a monthly time series of the number of domestic and international passenger bookings for the period 2010–2019.

```
passenger <- read_csv("./data/Passengers_2019.csv")
df.ts <- ts(passenger[,3], start = c(2010,1), end=c(2019,12), frequency = 12)

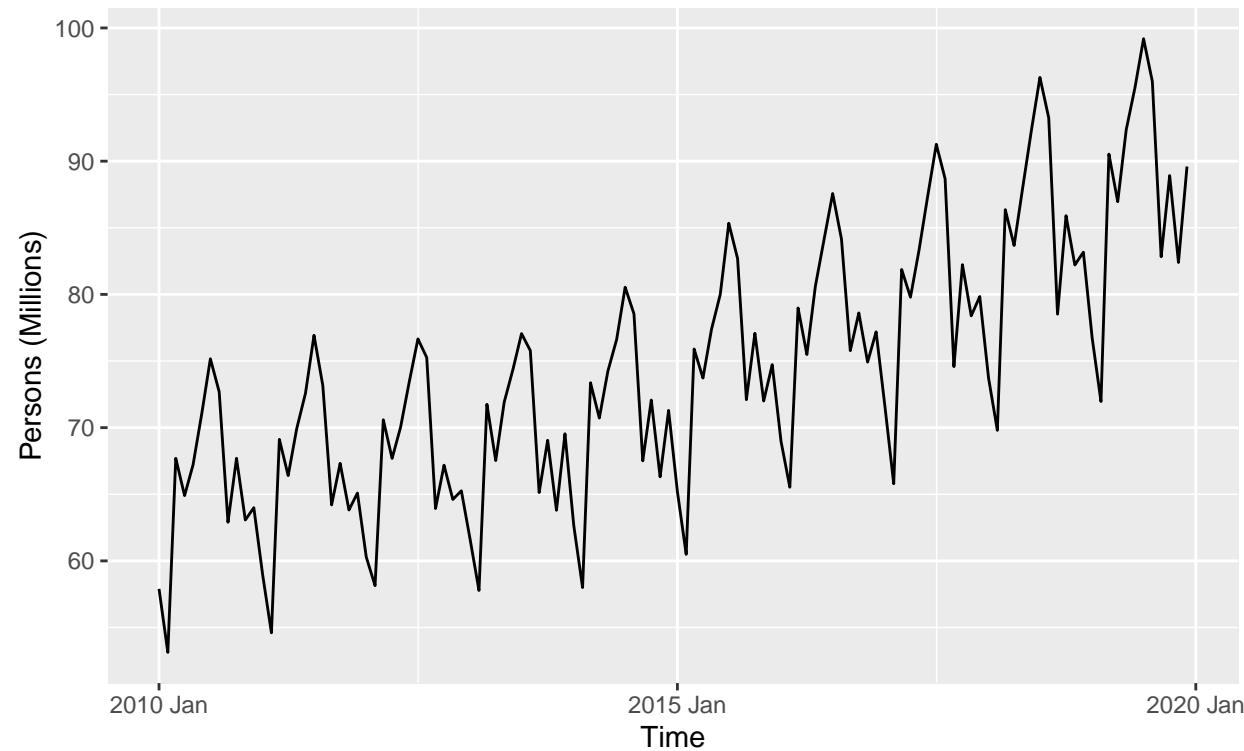
## convert dataframe to tsibble abject
df <- df.ts %>%
  as_tsibble(df.ts) %>%
  mutate(Total = round(value/1000000,2),
        log_total = log(Total)) %>%
  dplyr::select(index, Total, log_total)
```

Descriptive Statistics

- Produce a time plot of the data and describe the patterns in the graph. Identify any unusual or unexpected fluctuations in the time series.
- Is there any indication of non-stationarity?

Monthly airlines passenger in the U.S. for 2010–2019

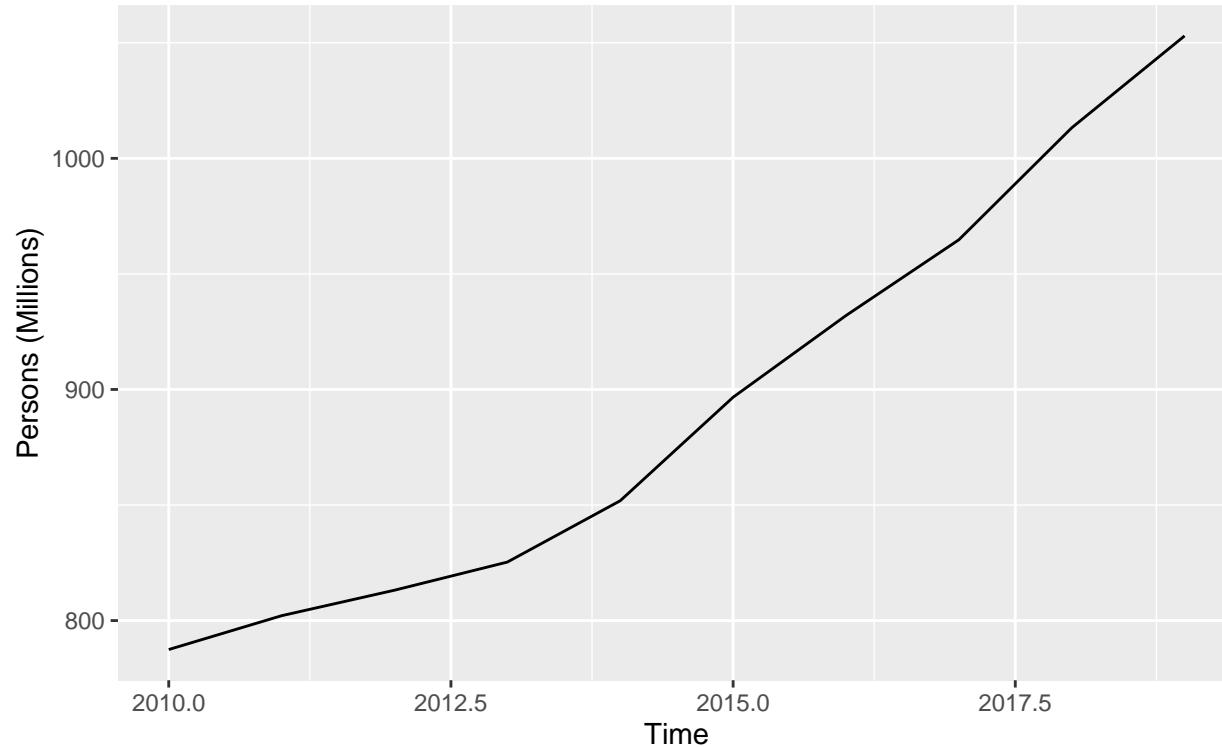
Domestic and international



- To have a clearer view of the trend, we can remove the seasonal movement by aggregating the data to the annual level.

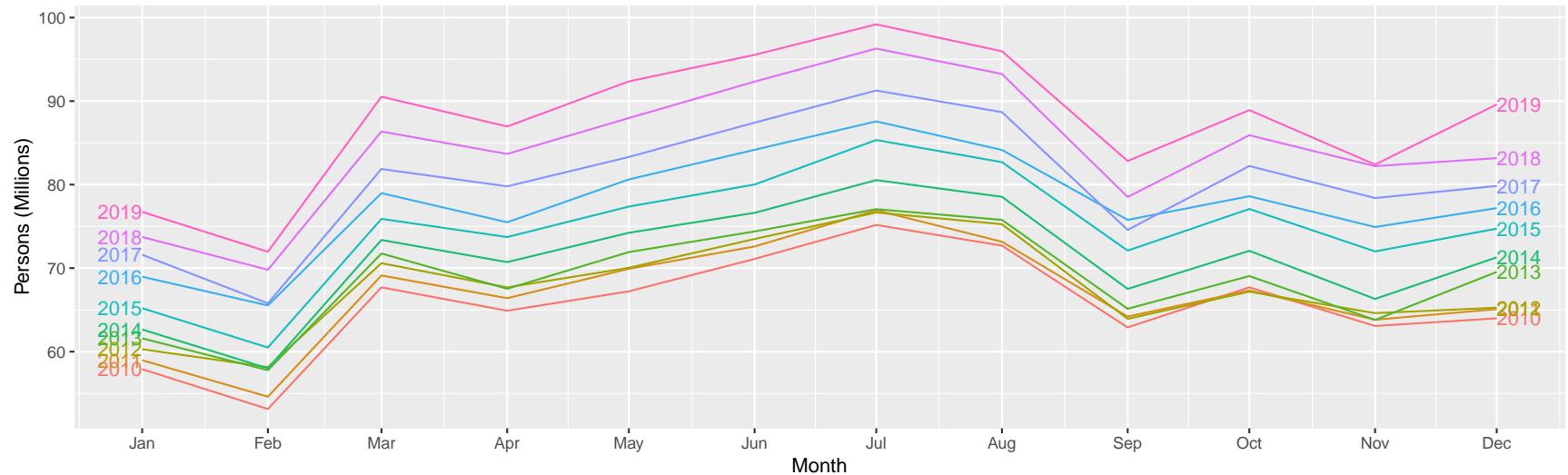
Annual airlines passenger in the U.S. for 2010–2019

Domestic and international

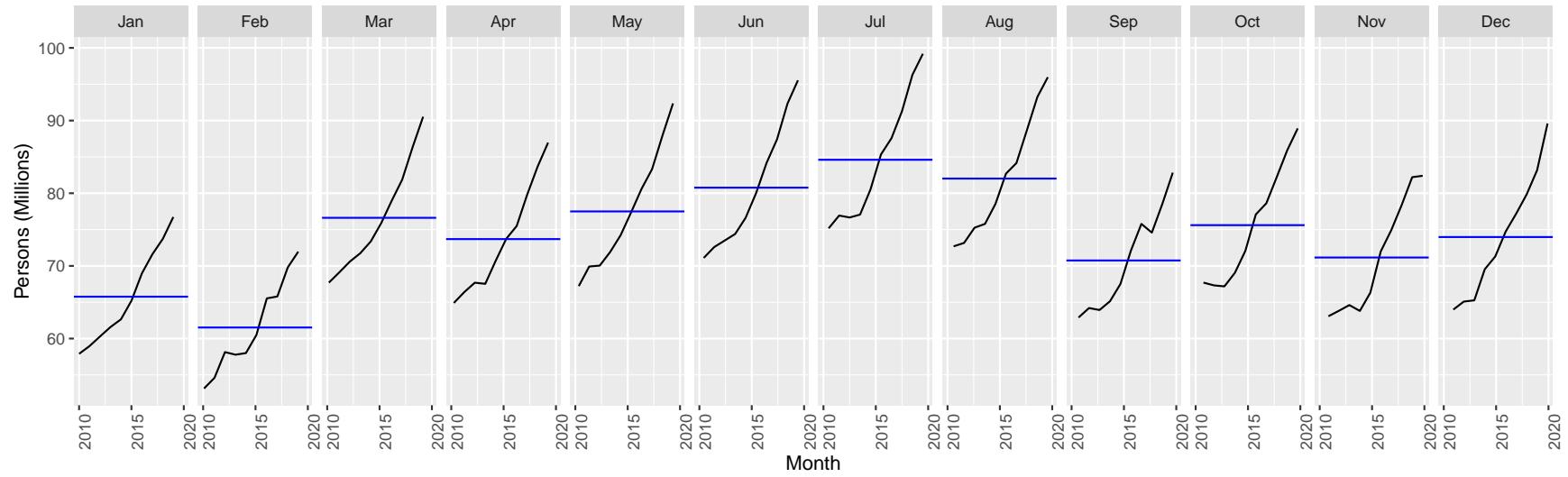


- We can view seasonality by using different types of plots.

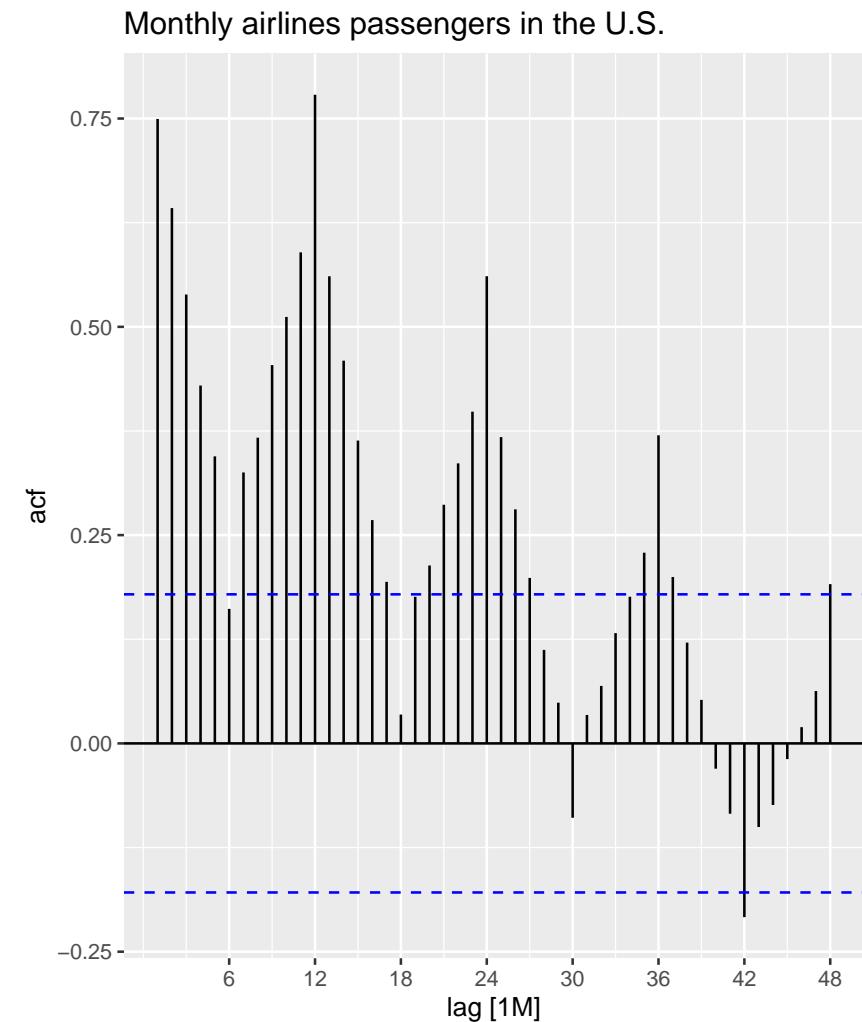
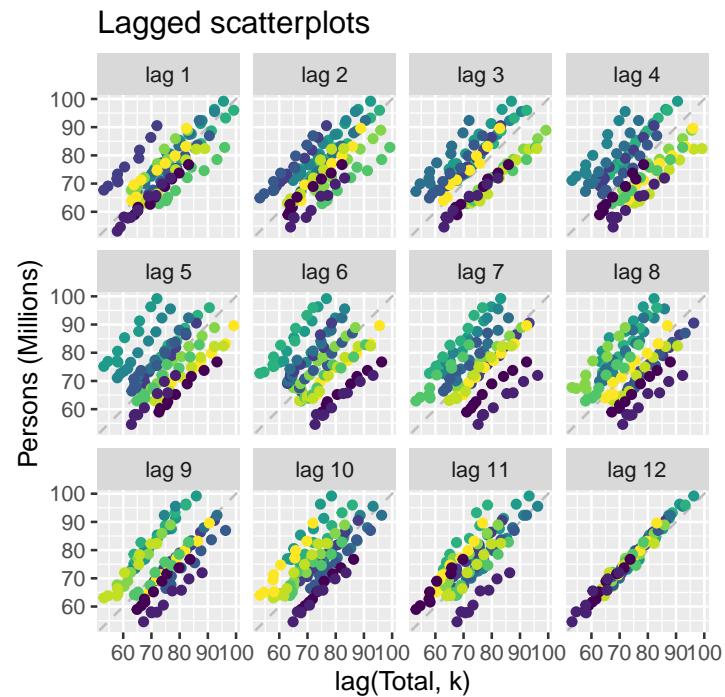
Seasonal plot: Airlines passenger in the U.S. for 2010–2019



Seasonal plot: Airlines passenger in the U.S. for 2010–2019



- One way to check for autocorrelation is to examine the ACF and lag plots. They are also helpful in determining if a time series has a trend and seasonality.
- What do you see in these two plots?



Model Development

Time series decomposition

- a)- Use both additive and multiplicative decomposition to remove trend and seasonal movement from the air passenger data set.
- b)- Check the random component. Is it stationary?

Modeling Deterministic Trend and Seasonality

- a)- Fit following regression models to the logarithm of the number of airline passengers data with a linear trend, quadratic trend, seasonal dummies variables.

$$\log(AP_t) = \beta_0 + \beta_1 \cdot t + \epsilon_t$$

$$\log(AP_t) = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2 + \epsilon_t$$

$$\log(AP_t) = \beta_0 + \beta_1 \cdot t + \sum_{i=2}^{12} \beta_i Month_i + \epsilon_t$$

$$\log(AP_t) = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2 + \sum_{i=2}^{12} \beta_i Month_i + \epsilon_t$$

b)- Plot the fitted values against time and against the airline passenger time series? What is the best-fitting model?

c)- How do we interpret the values of the coefficients?

```
#fit_linear <- # uncomment and replace with your code  
  
#fit_quadratic <- # uncomment and replace with your code  
  
#fit_linear_season <- # uncomment and replace with your code  
  
#fit_quadratic_season <- # uncomment and replace with your code
```

Regression Diagnostic Results

- The OLS procedure for time series data has good asymptotic properties under the following assumption (from Introductory Econometrics by Jeffrey Wooldridge):

1-Linearity: The stochastic process $\{(x_{1t}, x_{2t}, \dots, x_{kt}, y_t) : t = 1, 2, \dots, n\}$ follows the linear model:

$$y_t = \beta_0 + \beta_1 \cdot x_{1t} + \dots + \beta_k \cdot x_{kt} + u_t$$

- The x_{tj} can be lagged dependent

2-Ergodic stationarity: The stochastic process $\{(x_{1t}, x_{2t}, \dots, x_{kt}, y_t) : t = 1, 2, \dots, n\}$ is stationary and ergodic.

- Random sample assumption for cross-sectional data is replaced with this assumption. Intuitively, This assumption allows observations to be dependent but:
 - Stationarity does require that the any correlation between $(x_{1t}, x_{2t}, \dots, x_{kt}, y_t)$ and $(x_{1t+h}, x_{2t+h}, \dots, x_{kt+h}, y_{t+h})$ is the same across all time periods t and only depends on h .
 - Ergodicity does require that correlation between the $(x_{1t}, x_{2t}, \dots, x_{kt}, y_t)$ and $(x_{1t+h}, x_{2t+h}, \dots, x_{kt+h}, y_{t+h})$ disappears sufficiently quickly as observations get farther apart and $h \rightarrow \infty$

3- No Perfect Collinearity

4- Zero conditional mean or predetermined regressors All explanatory variables (x_{t1}, \dots, x_{tk}) are uncorrelated with the contemporaneous error term u_t :

$$E(x_{tk} \cdot u_t) = 0$$

- For all t and k .
- This assumption puts no restrictions on how ϵ_t is related to the explanatory variables in other time periods.

5- Homoskedasticity: The error u_t are contemporaneously homoskedastic

$$Var(u_t | x_{it}) = \sigma^2$$

6- No serial correlation: For all time period $t \neq s$ and i variables:

$$E(u_t u_s | x_{it} x_{is}) = 0$$

- Under assumptions 1 to 4, the OLS estimators are consistent

- Under these six assumptions, the OLS estimators are asymptotically normally distributed, and the usual OLS standard errors, t statistics, F statistics, and LM statistics are asymptotically valid.

Test for serial correlation Serial correlation means that error terms from different (usually adjacent) periods are correlated.

Serial correlation will not affect the unbiasedness or consistency of OLS estimators, but it does affect their efficiency.

With a positive serial correlation, the OLS estimates of the standard errors will be smaller than the actual standard errors. This will lead to the conclusion that the parameter estimates are more precise than they really are, inflating t statistics. There will be a tendency to reject null hypotheses when they should not be rejected.

- There are different types of serial correlation.

1- First-order AR(1) serial correlation:

$$u_t = \rho u_{t-1} + e_t$$

- $\rho < 1$
- u_t is the measured error term
- e_t is an uncorrelated random variable with mean zero and variance σ_e^2

We can use the Durbin-Watson test for AR(1) serial correlation. The null hypothesis and the Durbin-Watson (DW) statistic are:

- $H_0 : \rho = 0$
- $H_a : \rho \neq 0$

$$DW = \frac{\sum_{t=2}^T (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=2}^T \hat{u}_t^2}$$

2- Higher-Order AR(p) Serial Correlation:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \dots + \rho_q u_{t-q} + e_t$$

- To test for higher-order serial correlation, we can use the Ljung-Box test. The null hypothesis and the Ljung-Box statistic are:

– $H_0 : \rho_1 = \rho_2 = \dots = \rho_q = 0$

$$Q_h = T(T+2) \sum_{j=1}^h \frac{\hat{\rho}_j^2}{T-j}$$

- where h represents the maximum lag.

- Under the null hypothesis $Q_h \sim \chi_h^2$ since asymptotically $\hat{\rho}_j \sim N(0, \frac{1}{T})$ under the null and therefore the statistic is proportional to the sum of squared standard normal random variables.

a)- Plot the residuals against time and the fitted values and try to check the 6 CLM assumptions. Do these plots reveal any problems with the model?

b)- Test for autocorrelation of the residuals

```
# Replace with your code
```

Model Selection

- Recall that **Akaike Information Criterion (AIC)** is defined as:

$$AIC = -2 \times \log L_k + 2 \times k$$

- where $\log L_k$ is the maximized log-likelihood and k is the number of parameters in the model.

- One could normalize AIC by n , the number of observations used to estimate the model, and obtain

$$AIC = \frac{-2\log L_k + 2k}{n} \approx \ln(\hat{\sigma}^2) + \frac{2k}{n} + c$$

- where $\hat{\sigma}^2$ denotes the MLE of σ^2 and c is some constant.

- Also **Bayesian Information Criteria (BIC)** is given by:

$$BIC = \ln(\hat{\sigma}^2) + k \frac{\ln(n)}{n}$$

- Note that BIC imposes a greater penalty for the number of estimated model parameters than AIC. As such, BIC will always give a model whose number of parameters is no greater than that chosen under AIC.

- The information criterion model selection process should **NOT** be used as a substitute for careful examination of the characteristics of the estimated autocorrelation and partial autocorrelation functions; they can be used as supplementary guidelines.
- Critical examination of the residuals for model inadequacies should always be included as a major aspect of the overall model selection process.

a)- Now compare the three estimated models using AIC and BIC

```
# Replace with your code
```

Forecasting

- Recall that the predictions of y can be obtained using:

$$\hat{y}_{t+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{t+1}$$

Assuming that the regression errors are normally distributed, an approximate 95% prediction interval associated with this forecast is given by:

$$\begin{aligned}\hat{y} &\pm 1.96\hat{\sigma}_u SE \\ SE &= \sqrt{1 + \frac{1}{T} + \frac{(x - \bar{x})^2}{(T - 1)s_x^2}}\end{aligned}$$

- Where T is the total number of observations, s_x is the standard deviation of the observed x values, and $\hat{\sigma}_u$ is the standard error of the regression.
- SE includes two sources of variance or uncertainty in our forecasts. First is σ^2 , the variance of the error in the population, and it does not change with the sample size. Second, the sampling error in \hat{y} , which arises because we have estimated the β_0 and β_1 . Because β_0 and β_1 have a variance proportional to $1/T$, for large samples, the sampling error can be very small.

a)- Regardless of your answers to the above questions, use your regression model to predict the monthly airline passengers for 2020, 2021, and 2022. Produce prediction intervals for each of your forecasts.

```
# Replace with your code
```