ISA 444: Business Forecasting

03 - Basic Tools and Goodness of Fit

Fadel M. Megahed

Associate Professor
Department of Information Systems and Analytics
Farmer School of Business
Miami University
Email: fmegahed@miamioh.edu
Office Hours: Click here to schedule an appointment

Fall 2020

Outline

1. **Preface**
2. Plots for Time-Series and Cross Sectional Data
3. Summarizing the Data, Correlation & Transformations
4. Recap

Quick Refresher based on Last Class

Main Learning Outcomes

☑ Describe course objectives & structure.

☑ Describe what do we mean by forecasting and explain the **PIVASE** framework.

☑ Explain the differences between cross sectional, time series, and panel datasets.

☑ Identify and describe the basic components of a time series including trends, seasonal components, and cycles.

Non-Graded Activity

For the U.S.'s Advance Retail Sales Dataset, please read the data into R and answer the four questions on the Non-Graded Assignment on Canvas. You have 8 minutes to answer all four questions.

Note that this non-graded activity has NO impact on your grade. Time limit is set so that we do not take a significant portion of class to answer these questions.

Learning Objectives for Today's Class

Main Learning Outcomes

- **Interpret seasonal plots.**
- **Use numerical summaries to describe a time series.**
- **Apply differencing to a time series.**

Outline

1 Preface

2 **Plots for Time-Series and Cross Sectional Data**

3 Summarizing the Data, Correlation & Transformations

4 Recap

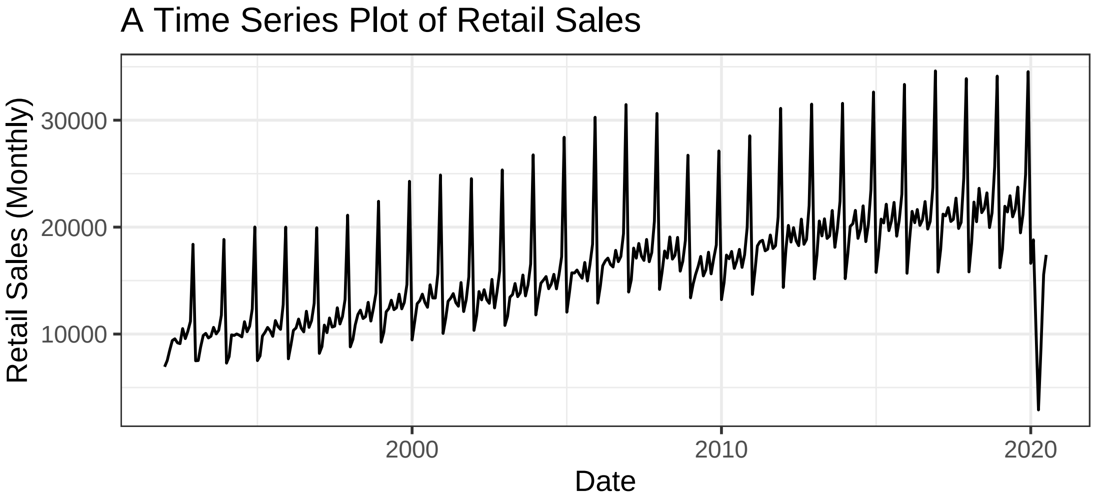
Outline

1. Preface
2. **Plots for Time-Series and Cross Sectional Data**
   - Plots for Time-Series Data
   - Plots for Cross Sectional Data
3. Summarizing the Data, Correlation & Transformations
4. Recap

Recall: Time-Series Plot [1]

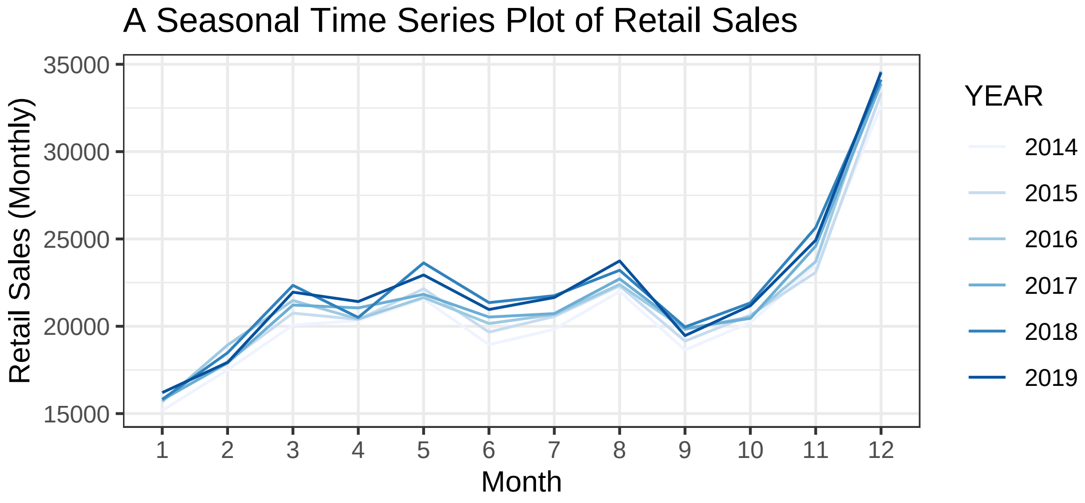
```r
if(require(pacman)==FALSE) install.packages("pacman")
pacman::p_load(tidyverse, lubridate)
retailSales = read.csv("https://fred.stlouisfed.org/graph/fredgraph.csv?bgcol
retailSales$DATE = ymd(retailSales$DATE)

retailSales %>% ggplot(aes(x = DATE, y = RSCCASN)) +
  geom_line() +
  labs(x = "Date", y = "Retail Sales (Monthly)",
       title = "A Time Series Plot of Retail Sales",
       caption = "Data from FRED") +
  theme_bw()
```

Recall: Time-Series Plot [2]



Data from FRED

Constructing a Seasonal Plot – Will be Coded Live in Class

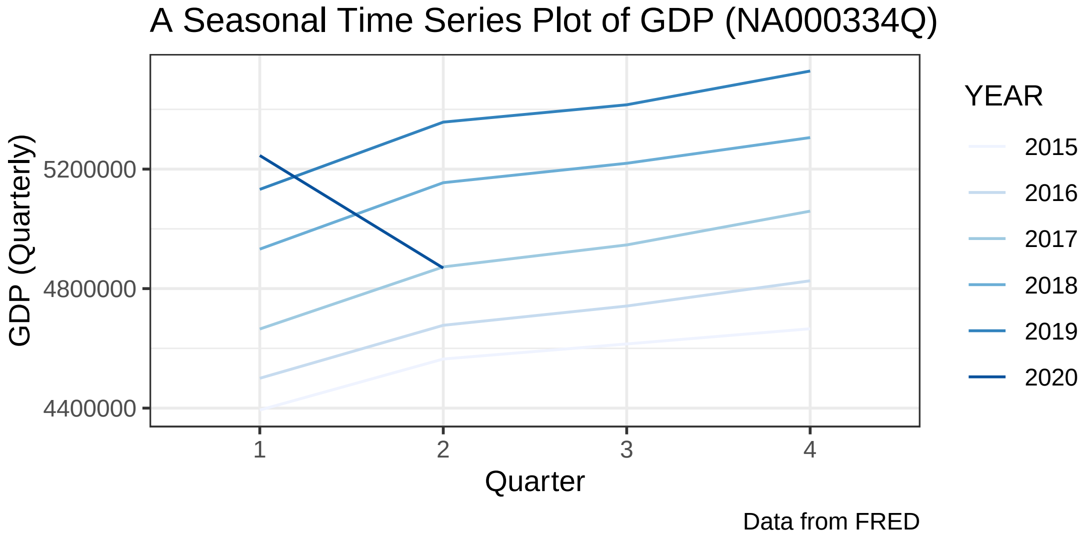


Data from FRED

An Evaluation of your Ability to Extend the Analysis [1]

- In 2-3 minutes, I would like to reproduce and extend the analysis from the previous slide to the Not Seasonally Adjusted GDP data from FRED (click here to access data).
- Your output should be similar to what is shown in the following slide.
- This will **NOT** be graded, but it will force you to ask me questions if you are not able to get a graph similar to the one in the next slide ☺.

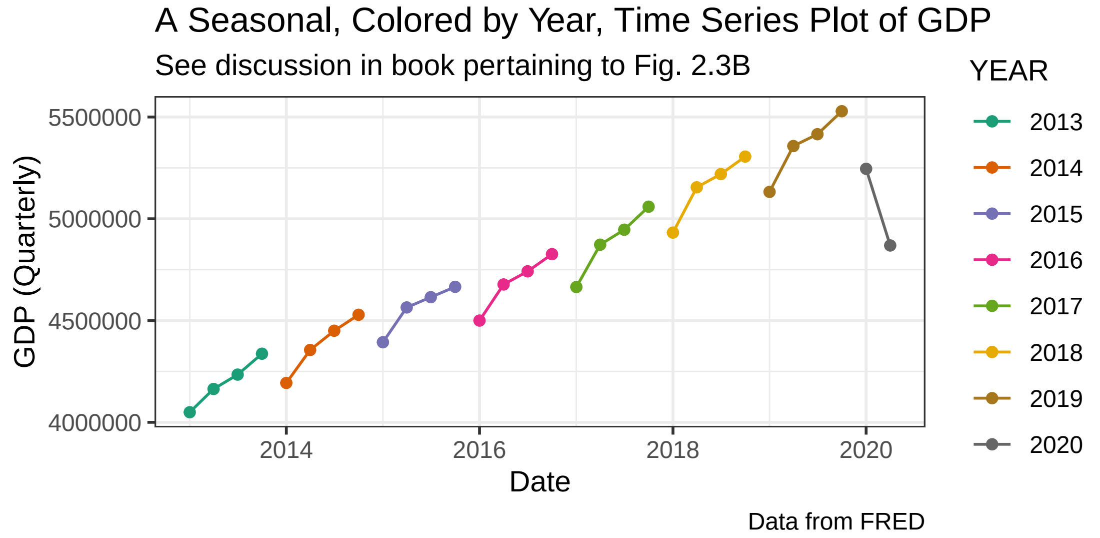
An Evaluation of your Ability to Extend the Analysis [2]



A Time Series Plot of GDP, color coded by Year [1]

```r
gdp %>% filter(YEAR >= 2013) %>%
    mutate(YEAR = as.factor(YEAR)) %>%
    ggplot(aes(x = DATE, y = NA000334Q, color = YEAR)) +
    geom_line() +
    labs(x = "Date", y = "GDP (Quarterly)",
         title = "A Seasonal, Colored by Year, Time Series Plot of GDP",
         subtitle = "See discussion in book pertaining to Fig. 2.3B",
         caption = "Data from FRED") + theme_bw() + scale_x_date() +
    scale_color_brewer(type = "qual", palette = "Dark2") + geom_point()
```

A Time Series Plot of GDP, color coded by Year [2]



Outline

1. Preface

2. **Plots for Time-Series and Cross Sectional Data**
   - Plots for Time-Series Data
   - Plots for Cross Sectional Data

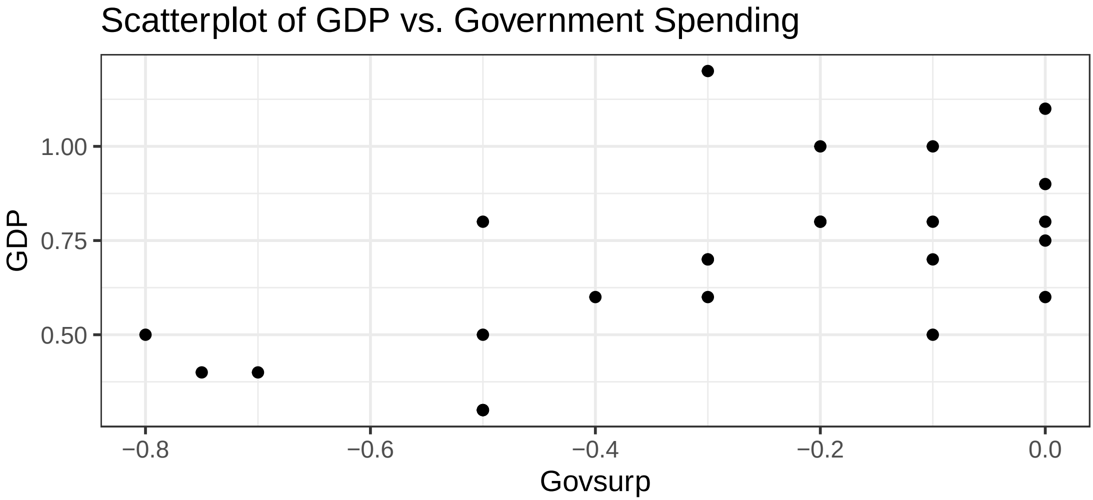
3. Summarizing the Data, Correlation & Transformations

4. Recap

Scatterplots [1]

- **Scatterplots** are frequently used to visualize the correlation between two continuous variables.
- In this Example, we will be using the German_Forecast Data. The file can be downloaded to your working directory, using the `download.file()` from base R.
- Note that the data is an xlsx file, which would require us to use the `read_excel()` from the readxl package.
- We will remake the plot of GDP vs Govsurp (Figure 2.4 in our textbook) using R. As noted in the chapter, the figure was created using Minitab for the book.
- The plot using the `ggplot()` is shown in the next slide. We will recreate it in class.

Scatterplots [2]



Data from Muller−Droge et al. (2016)

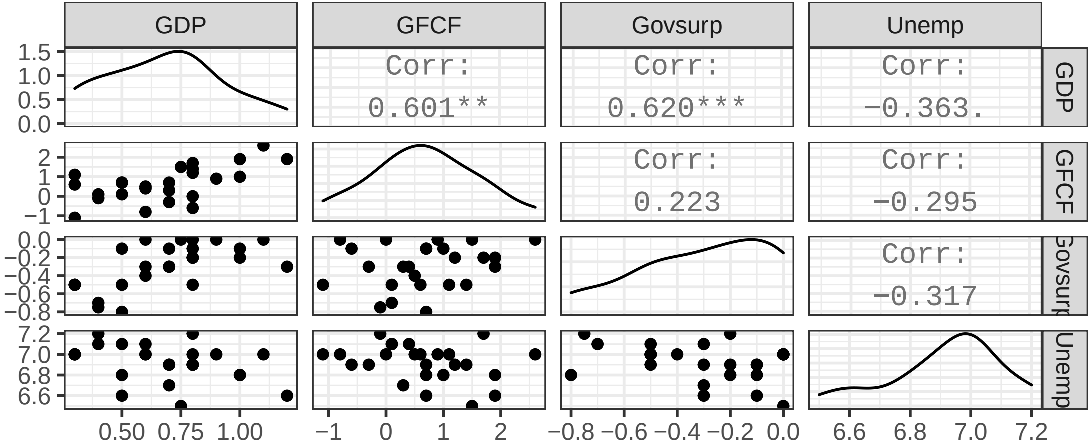
Scatterplot Matrix / Generalized Pairs Plots [1]

- Extending scatterplots for when we have more than two variables.[1]
- Can be easily created in R using the `ggpairs()` from the GGally package.

[1] John W Emerson, Walton A Green, Barret Schloerke, Jason Crowley, Dianne Cook, Heike Hofmann, Hadley Wickham. The Generalized Pairs Plot. Journal of Computational and Graphical Statistics, vol. 22, no. 1, pp. 79–91, 2012. Click here to access paper.

Scatterplot Matrix / Generalized Pairs Plots [2]

Matrix Plot of GDP, GFCF, Govsurp & Unemp



Data from Muller−Droge et al. (2016)

Outline

1. Preface
2. Plots for Time-Series and Cross Sectional Data
3. **Summarizing the Data, Correlation & Transformations**
4. Recap

Outline

1. Preface
2. Plots for Time-Series and Cross Sectional Data
3. **Summarizing the Data, Correlation & Transformations**
   - Summarizing the Data
   - Correlation
   - Transformations
4. Recap

Measures of Average

Mean: Given a set of n values Y_1, Y_2, …, Y_n, the arithmetic mean can be computed as:

$$\bar{Y} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n} = \frac{1}{n}\sum_{i=1}^{i=n} Y_i. \tag{1}$$

Order Statistics: Given a set of n values Y_1, Y_2, …, Y_n, we place them in an ascending order to define the order statistics, written as $Y_{(1)}$, $Y_{(2)}$, …, $Y_{(n)}$.

Median:

- If n is odd, $n = 2m + 1$ and the median is $Y_{(m+1)}$.
- If n is even, $n = 2m$ and the median is the average of the two middle numbers, i.e., $\frac{1}{2}[Y_{(m)} + Y_{(m+1)}]$.

Measures of Variation

The **range** denotes the difference between the largest and smallest value in a sample:

$$\text{Range} = Y_{(n)} - Y_{(1)}. \tag{2}$$

The **deviation** is defined as the difference between a given observation Y_i and the mean $\bar{Y}$.

The **mean absolute deviation (MAD)** is the average deviations about the mean, irrespective of their sign:

$$\text{MAD} = \frac{\sum_{i=1}^{i=n} |d_i|}{n}. \tag{3}$$

The **variance** is the average of the squared deviations around the mean:

$$S^2 = \frac{\sum_{i=1}^{i=n} d_i^2}{n-1}. \tag{4}$$

Applications of Measures of Average/Variance: MAD Function [1]

- The `mad()` in R is used for computing the median absolute deviation and **Not** the mean absolute deviation. This can be easily checked using `?mad()` in your R console.
- Thus, we will have to create our custom R function, `MAD()`, which we will define as follows:

```r
MAD = function(x){
  return( mean( abs(x-mean(x)) ) )
  }
```

- Now, let us make sure that this formula works as expected by testing it on the vector `x = c(1, 2, 3)` and comparing it with manually computing the MAD.

Applications of Measures of Average/Variance: $^DJI [1]

- Let us examine the stock prices for the Dow Jones Industrial Average Index from March 23, 2020 up to and including August 21, 2020.
- Let us compute the aforementioned measures, on the adjusted closing price, using the following two approaches: (a) averages **across** all months, and (b) averages **by/within** month. The printout for those two methods are shown in the tables below.

meanACP	medianACP	madACP	varACP	sdACP
25151.08	25605.54	1584.40	3668632.08	1915.37

month	meanACP	medianACP	madACP	varACP	sdACP
March	21275.85	21636.78	951.48	1801923.06	1342.36
April	23293.90	23515.26	762.02	1064521.85	1031.76
May	24271.02	24214.42	533.18	435018.17	659.56
June	26062.27	25948.21	501.86	446276.23	668.04
July	26385.83	26449.11	317.40	146256.40	382.43
August	27585.58	27739.73	321.74	165154.65	406.39

Outline

1. Preface
2. Plots for Time-Series and Cross Sectional Data
3. **Summarizing the Data, Correlation & Transformations**
   - Summarizing the Data
   - Correlation
   - Transformations
4. Recap

The Pearson Correlation Coefficient

- **Correlation:** measures the strength of the **linear relationship** between two quantitative variables.

- It can be computed using the `cor()` from base R. Mathematically speaking, the pearson correlation coefficient, r, can be computed as

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{5}$$

- Do **not** use the Pearson Correlation coefficient if both variables are not quantitative. Instead, refer to the `mixed.cor()` from the psch package to compute the correlations for mixtures of continuous, polytomous, and/or dichotomous variables.

- You should supplement any descriptive summaries with visualizations to ensure that you are able to interpret the computations correctly.

A Synthetic Example: The Anscombe Dataset [1]

In a seminal paper, Anscombe stated:[2]*Few of us escape being indoctrinated with these notions*

- *numerical calculations are exact, but graphs are rough;*
- *for any particular kind of statistical data there is just one set of calculations constituting a correct statistical analysis;*
- *performing intricate calculations is virtuous, whereas actually looking at the data is cheating.*

He proceeded by stating that *a computer should make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding.*

Now, let us consider his four datasets, each consisting of eleven (x,y) pairs.

[2] Anscombe, Francis J. 1973. “Graphs in Statistical Analysis." *The American Statistician* 27 (1): 17–21. (Click here to access the full paper).

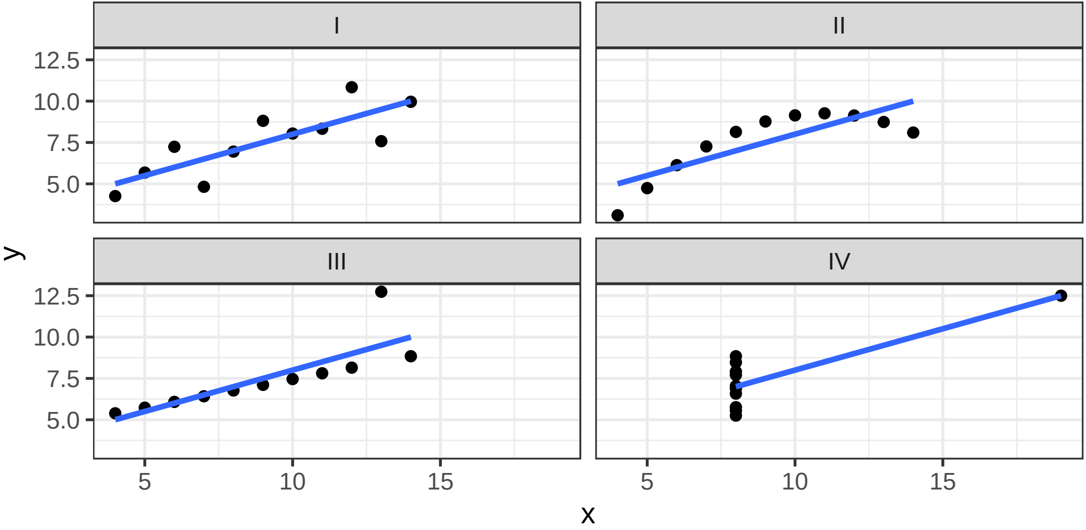
A Synthetic Example: The Anscombe Dataset [2]

x1	x2	x3	x4	y1	y2	y3	y4
10.00	10.00	10.00	8.00	8.04	9.14	7.46	6.58
8.00	8.00	8.00	8.00	6.95	8.14	6.77	5.76
13.00	13.00	13.00	8.00	7.58	8.74	12.74	7.71
9.00	9.00	9.00	8.00	8.81	8.77	7.11	8.84
11.00	11.00	11.00	8.00	8.33	9.26	7.81	8.47
14.00	14.00	14.00	8.00	9.96	8.10	8.84	7.04
6.00	6.00	6.00	8.00	7.24	6.13	6.08	5.25
4.00	4.00	4.00	19.00	4.26	3.10	5.39	12.50
12.00	12.00	12.00	8.00	10.84	9.13	8.15	5.56
7.00	7.00	7.00	8.00	4.82	7.26	6.42	7.91
5.00	5.00	5.00	8.00	5.68	4.74	5.73	6.89

A Synthetic Example: The Anscombe Dataset [3]

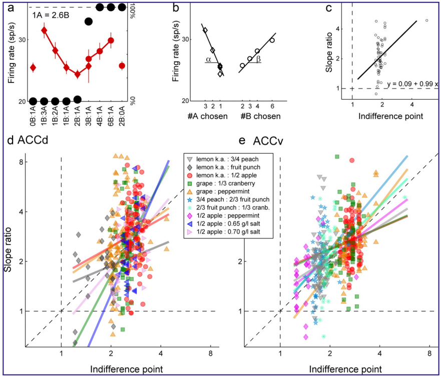
set	x.mean	x.sd	y.mean	y.sd	corr
I	9.00	3.32	7.50	2.03	0.82
II	9.00	3.32	7.50	2.03	0.82
III	9.00	3.32	7.50	2.03	0.82
IV	9.00	3.32	7.50	2.03	0.82

A Synthetic Example: The Anscombe Dataset [4]



Anscombe-Like Mistakes in Research and Practice

In my estimation, Figure 8c represents an example where regression should not have been performed[3].



[3] Cai, Xinying, and Camillo Padoa-Schioppa. 2012. "Neuronal Encoding of Subjective Value in Dorsal and Ventral Anterior Cingulate Cortex." *Journal of Neuroscience* 32(11):3791–3808.

Outline

1. Preface
2. Plots for Time-Series and Cross Sectional Data
3. **Summarizing the Data, Correlation & Transformations**
   - Summarizing the Data
   - Correlation
   - Transformations
4. Recap

First Differences [1]

The change in the time series from one period to the next is known as the (first) difference. It can be computed as follows:

$$DY_t = Y_t - Y_{t-1} \tag{6}$$

```r
dowJonesIndex = tq_get("^DJI", from = "2020-08-17", to = "2020-08-22") %>%
  select(symbol, date, adjusted)
dowJonesIndex$`Yt-1` = lag(dowJonesIndex$adjusted)
dowJonesIndex$DYt = dowJonesIndex$adjusted - dowJonesIndex$`Yt-1`
```

symbol	date	adjusted	Yt-1	DYt
^DJI	18491.00	27844.91		
^DJI	18492.00	27778.07	27844.91	-66.84
^DJI	18493.00	27692.88	27778.07	-85.19
^DJI	18494.00	27739.73	27692.88	46.85
^DJI	18495.00	27930.33	27739.73	190.60

First Differences [2]

Note that the differences can be computed in one step using the function `diff()` from base R as follows.

```
dowJonesIndex = tq_get("^DJI", from = "2020-08-17", to = "2020-08-22") %>%
  select(symbol, date, adjusted)
dowJonesIndex$DYt = c(NA, diff(dowJonesIndex$adjusted))
```

symbol	date	adjusted	DYt
^DJI	18491.00	27844.91	
^DJI	18492.00	27778.07	-66.84
^DJI	18493.00	27692.88	-85.19
^DJI	18494.00	27739.73	46.85
^DJI	18495.00	27930.33	190.60

Outline

1. Preface
2. Plots for Time-Series and Cross Sectional Data
3. Summarizing the Data, Correlation & Transformations
4. **Recap**

Summary of Main Points

Main Learning Outcomes

- **Interpret seasonal plots.**
- **Use numerical summaries to describe a time series.**
- **Apply differencing to a time series.**

Things to Do

- Thoroughly read Chapter 2 of our book. We covered up to the beginning of Section 2.6.1).
- Go through the slides, examples and make sure you have a good understanding of what we have covered.
- Complete the graded assignment for more details.
- If you are interested in additional practice problems, please consider the following problems from your textbook.
  - Exercise 2.4
  - Exercise 2.7

Graded Assignment 03: Evaluating your Retention/Focus

Please go to Canvas (click here) and answer the two questions. **Due August 27, 2020 [11:59 PM, Ohio local time].**

What/Why/Prep? The purpose of this assignment is to evaluate your understanding and retention of the material covered up to the end of Class 03. In order to prepare for this, you should have either actively attended class and/or watched the recording from WebEx. Furthermore, you should have thoroughly read up to the begining of Section 2.6.1 from your textbook.

General Guidelines:

- Individual assignment.
- This is **NOT** a timed assignment (i.e. once you start the assignment you will have 25 minutes to complete 4 questions). If the concepts we covered are well-understood, this should take $\leq$ 10 minutes.
- Proctorio is NOT required for this assignment.
- You will need to have R installed (or accessible through the Remote Desktop)

ISA 444: Business Forecasting

03 - Basic Tools and Goodness of Fit

Fadel M. Megahed

Associate Professor
Department of Information Systems and Analytics
Farmer School of Business
Miami University
Email: fmegahed@miamioh.edu
Office Hours: Click here to schedule an appointment

Fall 2020