ISA 444: Business Forecasting

04 - Basic Tools and Goodness of Fit (Cont.)

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. **Summarizing the Data, Correlation & Transformations**

3. **Recap**

Quick Refresher based on Last Class

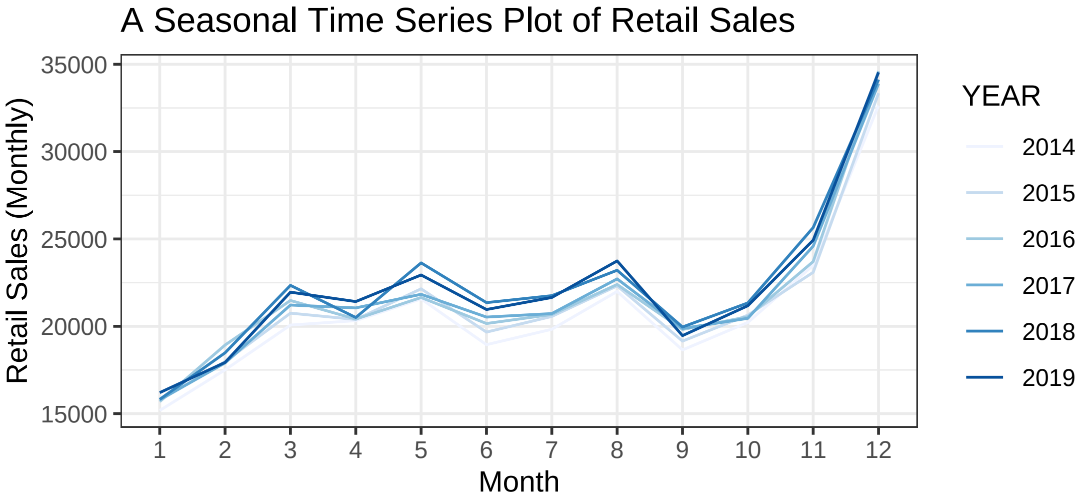
Main Learning Outcomes

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

Non-Graded Mentimeter Poll

Given that we have spent the majority of last class on plotting, please go to www.menti.com, and answer the sliding scale questions, to assess your understanding of the material so far.

Recap: Seasonal Plots

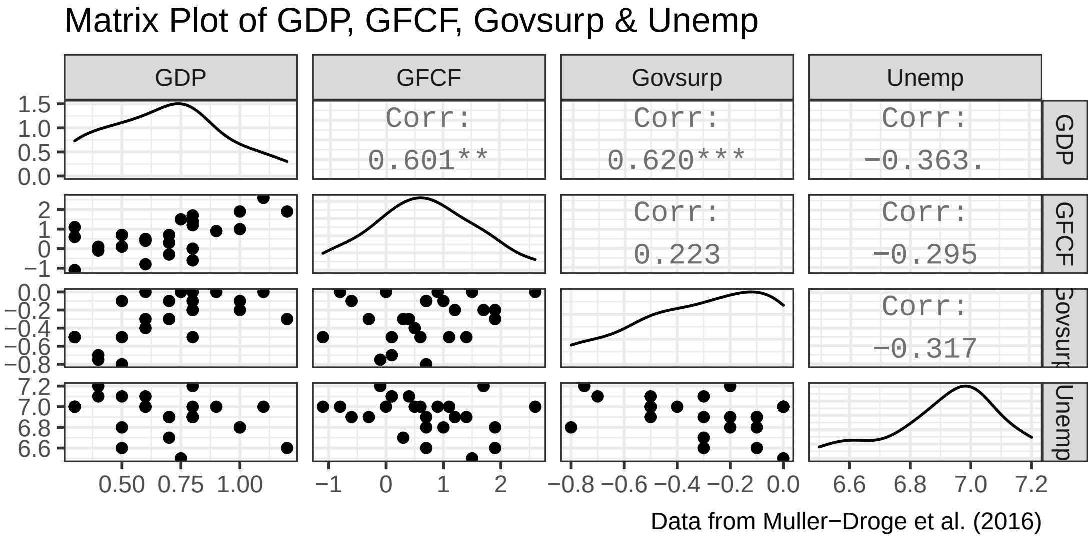


Data from FRED

Recap: Scatterplot Matrix / Generalized Pairs Plots [1]

```
pacman::p_load(tidyverse, GGally, readxl)
read_excel("Data/German_forecasts.xlsx") %>%
  ggpairs(columns = c('GDP', 'GFCF', 'Govsurp', 'Unemp')) +
  labs(title = "Matrix Plot of GDP, GFCF, Govsurp & Unemp",
       caption = "Data from Muller-Droge et al. (2016)") + theme_bw()
```

Recap: Scatterplot Matrix / Generalized Pairs Plots [2]



Learning Objectives for Today's Class

Main Learning Outcomes

- **Use numerical summaries to describe a time series.**
- **Apply transformations to a time series.**

Outline

1. Preface
2. **Summarizing the Data, Correlation & Transformations**
3. Recap

Outline

1. Preface
2. **Summarizing the Data, Correlation & Transformations**
   - Summarizing the Data
   - Correlation
   - Transformations
3. Recap

Measures of Average

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

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:

```
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 **Summarizing the Data, Correlation & Transformations**

- Summarizing the Data
- Correlation
- Transformations

3 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:[1]*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.

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

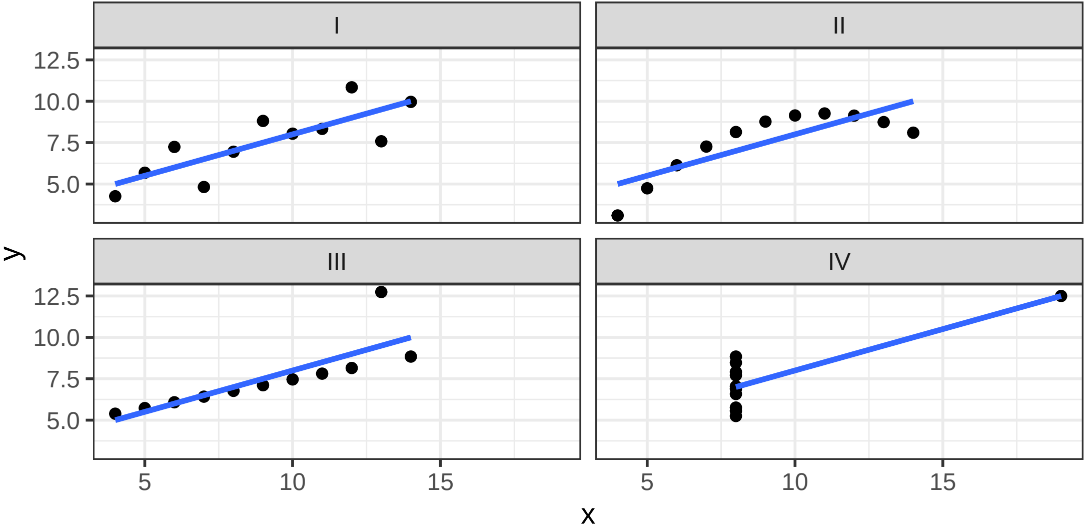
A Synthetic Example: The Anscombe Dataset [2]

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]

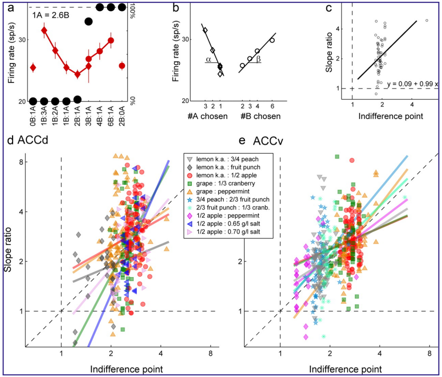
set	x.mean	x.sd	y.mean	y.sd	corr
I	9.00	3.32	7.50	2.03	0.82
II	9.00	3.32	7.50	2.03	0.82
III	9.00	3.32	7.50	2.03	0.82
IV	9.00	3.32	7.50	2.03	0.82

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[2].



[2] 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. **Summarizing the Data, Correlation & Transformations**
   - Summarizing the Data
   - Correlation
   - Transformations
3. 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	2020-08-17	27844.91		
^DJI	2020-08-18	27778.07	27844.91	-66.84
^DJI	2020-08-19	27692.88	27778.07	-85.19
^DJI	2020-08-20	27739.73	27692.88	46.85
^DJI	2020-08-21	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))
```

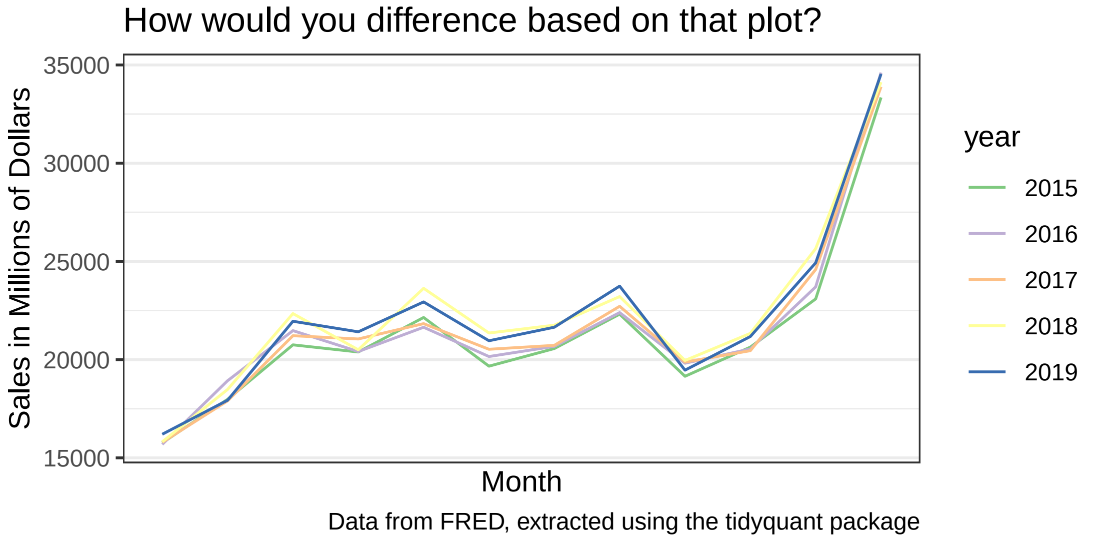
date	symbol	adjusted	DYt
2020-08-17	^DJI	27844.91	
2020-08-18	^DJI	27778.07	-66.84
2020-08-19	^DJI	27692.88	-85.19
2020-08-20	^DJI	27739.73	46.85
2020-08-21	^DJI	27930.33	190.60

Differencing for Seasonal Data [1]

Recall the plot of the Advance Retail Sales: Clothing and Clothing Accessory Stores (RSCCASN) FRED dataset.

```
pacman::p_load(tidyquant)
retailSales = tq_get("RSCCASN", get = "economic.data",
                     from = "2015-01-01", to = "2019-12-31")
retailSales$month = month(retailSales$date)
retailSales$year = year(retailSales$date) %>% factor()
retailSales %>%
  ggplot(aes(x = month, y = price, color = year)) + geom_line() +
  labs(title = "How would you difference based on that plot?",
       x = "Month", y = "Sales in Millions of Dollars",
       caption = "Data from FRED, extracted using the tidyquant package") +
  scale_x_discrete() + scale_color_brewer(type = "qual") + theme_bw()
```

Differencing for Seasonal Data [2]



Differencing with Seasonal Data [3]

The table below is the result of Approach #1 in R

date	price	Yt-m	DYt-m
2015-01-01	15764		
2015-02-01	17980		
2015-03-01	20752		
2015-04-01	20389		
2015-05-01	22145		
2015-06-01	19667		
2015-07-01	20564		
2015-08-01	22314		
2015-09-01	19151		
2015-10-01	20637		
2015-11-01	23090		
2015-12-01	33345		
2016-01-01	15685	15764	-79
2016-02-01	18926	17980	946
2016-03-01	21477	20752	725
2016-04-01	20412	20389	23
2016-05-01	21648	22145	-497
2016-06-01	20157	19667	490

Differencing with Seasonal Data [4]

The table below is the result of Approach #2 in R

date	price	DYt-m
2015-01-01	15764	
2015-02-01	17980	
2015-03-01	20752	
2015-04-01	20389	
2015-05-01	22145	
2015-06-01	19667	
2015-07-01	20564	
2015-08-01	22314	
2015-09-01	19151	
2015-10-01	20637	
2015-11-01	23090	
2015-12-01	33345	
2016-01-01	15685	-79
2016-02-01	18926	946
2016-03-01	21477	725
2016-04-01	20412	23
2016-05-01	21648	-497
2016-06-01	20157	490

Growth Rates: The Formulation

In the absence of seasonality, the growth rate for a time series is given by

$$GY_t = 100\frac{Y_t - Y_{t-1}}{Y_{t-1}} \tag{7}$$

In the presence of seasonality (with period $= m$), the growth rate for a time series is given by

$$GY_t = 100\frac{Y_t - Y_{t-m}}{Y_{t-m}} \tag{8}$$

Growth Rates in Practice – a Non-Graded Class Activity

- **Question:** Has anyone in this class bought any cryptocurrencies?
- **Follow up question:** Hypothetically speaking let us say that Fadel has purchased \$10 worth of Stellar (\$XLM). Each \$XLM coin was worth 0.1 on 08-25-2020.
  - Let us say that Fadel was lucky and his investment went up 20% on 08-26-2020, i.e. $GY_t = 20$. Using the formula, compute the value of each coin on 08-26-2020.
  - Due to the volatility of cryptocurrencies, let us assume that Fadel's growth rate for the following day was -20% (i.e., in comparison with the coin's value on 08-26-2020). Compute the value of the coin on 08-27-2020.
  - Provide the answer to both computational questions on Canvas.

The Log Transform [1]

The log transformation can be computed as follows:

$$L_t = \ln{(Y_t)} \tag{9}$$

Note that the `log()` in R takes the natural logarithm as its default base, i.e., would transform a variable/statistic based on the above equation.

The reverse transformation using the exponential function is:

$$e^{L_t} = e^{\ln{(Y_t)}} = Y_t \tag{10}$$

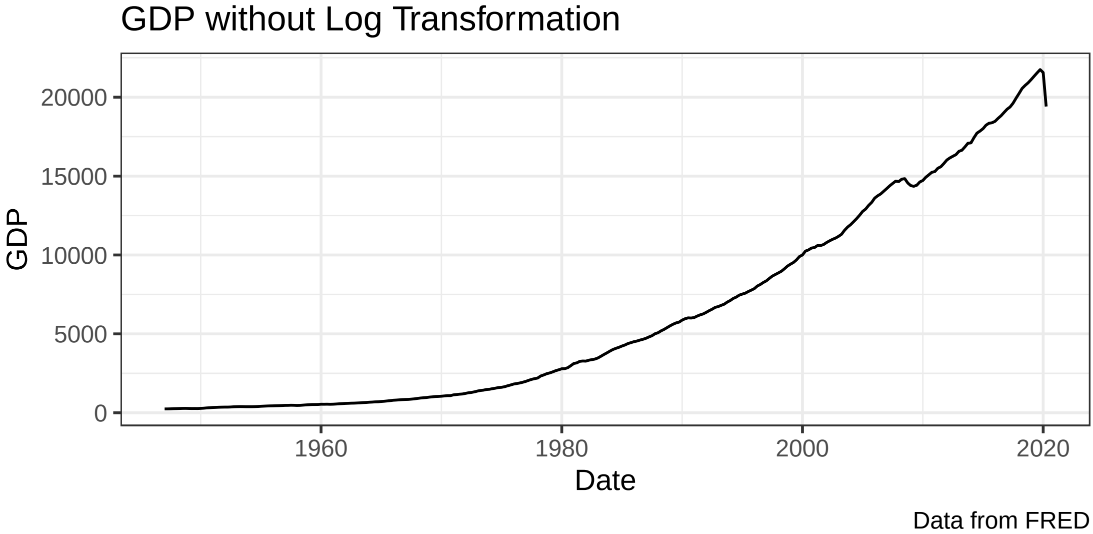
The first difference in logarithms represents the logarithm of the ratio:

$$L_t = \ln{(\frac{Y_t}{Y_{t-1}})} = \ln{(Y_t)} - \ln{(Y_{t-1})} \tag{11}$$

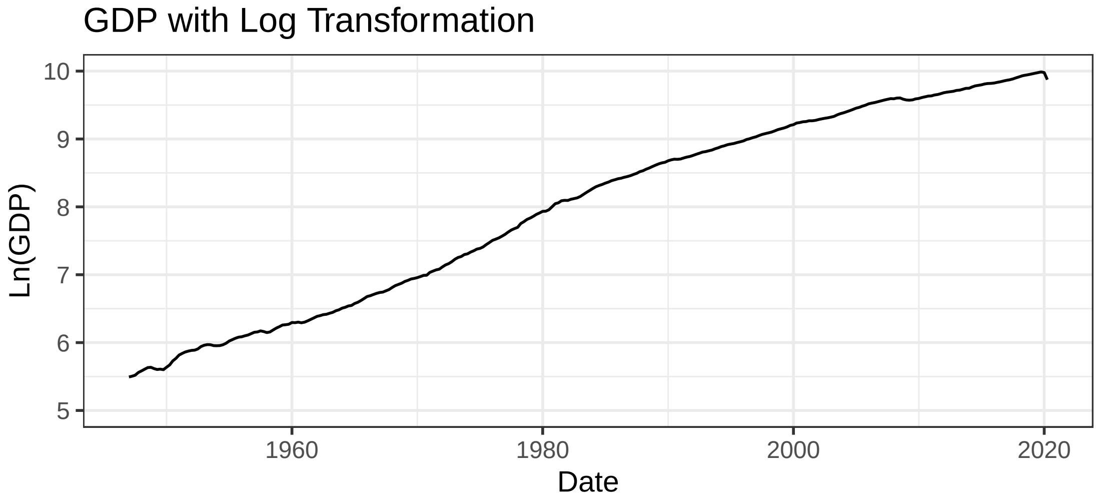
The Log Transform [2]

- The primary purpose of the log transform is to **convert exponential growth into linear growth.**
- The transform often has the **secondary purpose of balancing the variance.**
- Difference in logs and growth rate transformations produce similar results and interpretations.

Plots with and without the Log Transformation [1]

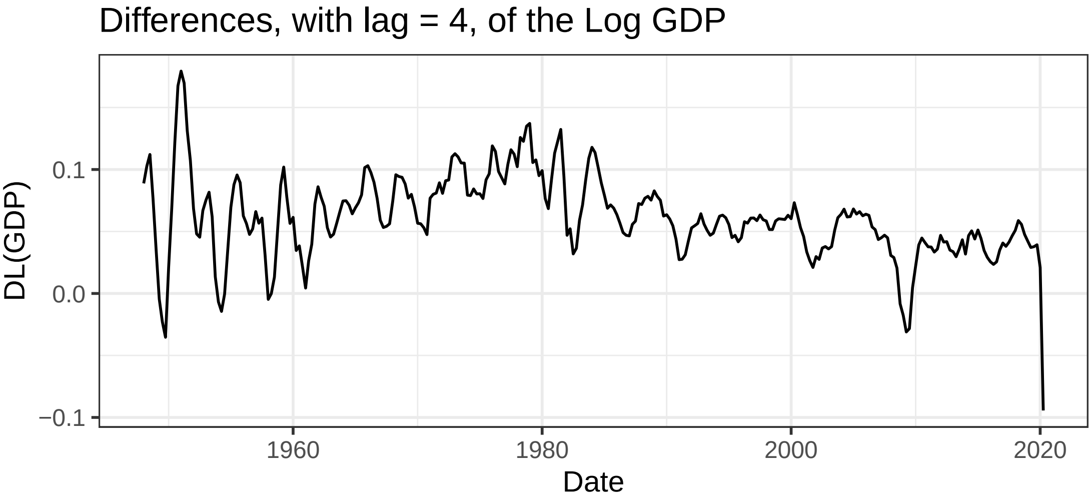


Plots with and without the Log Transformation [2]



Data from FRED

Plots with and without the Log Transformation [3]



Data from FRED

Outline

1 Preface

2 Summarizing the Data, Correlation & Transformations

3 Recap

Summary of Main Points

Main Learning Outcomes

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

Things to Do

- Thoroughly read Chapter 2 of our book. We covered up to the end of Section 2.6, p. 43).
- 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. To access these datasets,please click here.
  - Exercise 2.4
  - Exercise 2.7
  - Exercise 2.10

Graded Assignment 03: Evaluating your Retention/Focus

Please go to Canvas (click here) and answer the questions. **Due August 31, 2020 [2:50 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 04. 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 end of Section 2.6 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 ≤ 10 minutes.
- Proctorio is NOT required for this assignment.
- You will need to have R installed (or accessible through the Remote Desktop)

Graded Assignment 04: Evaluating your Retention/Focus

Please go to Canvas (click here) and answer the questions. **Due August 31, 2020 [2:50 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 04. 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 end of Section 2.6 from your textbook.

General Guidelines:

- Individual assignment.
- This is **NOT** a timed assignment (i.e. once you start the assignment you will have 10-15 minutes to complete the one question).
- Proctorio is NOT required for this assignment.
- You will need to have R installed (or accessible through the Remote Desktop)

ISA 444: Business Forecasting

04 - Basic Tools and Goodness of Fit (Cont.)

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