# Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Lab 3

*Professor Jeffrey Yau*

*March 18, 2018*

---

## Introduction

### Load packages set some formatting preferences

```
pkg <- c('knitr', 'Hmisc', 'ggcorrplot', 'car',
         'dplyr', 'ggplot2', 'jtools', 'readr')
invisible(lapply(pkg, require, character.only = T))
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

### Question 1: E-Commerce Retail Sales as a Percent of Total Sales

### Load data and quality-check

```
# Load csv file as df
df_q1 <- read.csv("ECOMPCTNSA.csv", header = TRUE, sep = ",")
str(df_q1)
```

```
## 'data.frame':    69 obs. of  2 variables:
##  $ DATE      : Factor w/ 69 levels "1999-10-01","2000-01-01",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ ECOMPCTNSA: num  0.7 0.8 0.8 0.9 1.1 1.1 1 1 1.3 1.3 ...
```

```
# View(df_q1) names(df_q1)
head(df_q1)
```

```
##         DATE ECOMPCTNSA
## 1 1999-10-01        0.7
## 2 2000-01-01        0.8
## 3 2000-04-01        0.8
## 4 2000-07-01        0.9
## 5 2000-10-01        1.1
## 6 2001-01-01        1.1
```

```
describe(df_q1)
```

```
## df_q1
##
##  2  Variables      69  Observations
## ---------------------------------------------------------------------------
## DATE
##        n  missing distinct
##       69        0       69
##
## lowest : 1999-10-01 2000-01-01 2000-04-01 2000-07-01 2000-10-01
## highest: 2015-10-01 2016-01-01 2016-04-01 2016-07-01 2016-10-01
## ---------------------------------------------------------------------------
## ECOMPCTNSA
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       69        0       50        1    3.835    2.524     0.94     1.10
##      .25      .50      .75      .90      .95
##     2.00     3.60     5.30     6.92     7.70
##
## lowest : 0.7 0.8 0.9 1.0 1.1, highest: 7.0 7.5 7.7 8.7 9.5
## ---------------------------------------------------------------------------
```

```
# Create an R time-series object with our quarterly data
# starting with final quarter of 1999
q1_ts <- ts(df_q1$ECOMPCTNSA, frequency = 4, start = c(1999,
    4))
str(q1_ts)
```

```
##  Time-Series [1:69] from 2000 to 2017: 0.7 0.8 0.8 0.9 1.1 1.1 1 1 1.3 1.3 ...
```

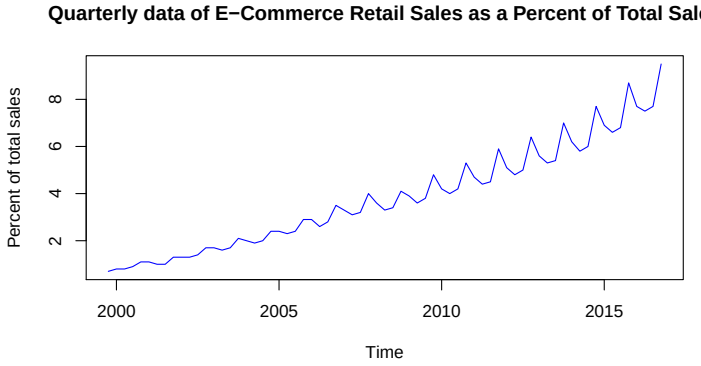
```
head(q1_ts)
```

```
##      Qtr1 Qtr2 Qtr3 Qtr4
## 1999                 0.7
## 2000  0.8  0.8  0.9  1.1
## 2001  1.1
```

We see that there are no missing values and that we are working with quarterly time series data. No anomalies are detected and there is not potential for top or bottom code. On gross visual inspection of the time series, we see that the starting values are all less than 1, with the later values all being greater than 7. This leads us to already suspect we are not dealing with a stationary series. We cannot make a confident comment on seasonality without further EDA for visualization.

### EDA

```
plot.ts(q1_ts, main = "Quarterly data of E-Commerce Retail Sales as a Percent of Total Sales",
    ylab = "Percent of total sales", col = "blue")
```




We are now able to clearly see that our the e-commernce retail sales time series is not stationary and contains a strong seasonal component.

### Order Identification

### Model Creation

### Fit Evaluation

### Question 2: data_2018Spring_MTS.txt

### Load data

```
# df_q2 <- read.csv('correlate-flight_prices.csv', header =
# TRUE, sep=',')
```

### EDA

### Order Identification

### Model Creation

### Fit Evaluation

## Conclusion