# Time Series SOLUTIONS

For this lab you need to upload the TS_lab data available in the class R package.

There are **4 sets of questions** to be answered. You can get up to 100 points + bonus questions. Points are indicated next to each question.

Remember to:

- Format your table using stargazer
- Submit the r markdown file with all your codes.
- Bold your answers.

```
# setwd("O:/Users/ffusi/Jesse/Labs/TimeSeries")
data = read.csv("TS_lab.csv")
```

## 1 The policy problem

> **Research question**: Does an increase in bus frequency affect bus ridership in City A?

Several cities aim to support public transportation and increase ridership. There are several ways to encourage residents to utilize public transportation more often; one of those is to modify the transit schedule so as to increase the frequency of buses in peak times and reduce waiting time.

City A has opted for this solution and starting from May 1st has implemented a new bus schedule which increases bus frequency. At the end of the year, the mayor wants to estimate whether the decision was effective and ask you to analyze one year of ridership data to test the following hypothesis:

> **Hypothesis**: An increase in bus frequency has a positive effect on bus ridership in City A (i.e., the new schedule is effective).

## 2 Data

Data on ridership were provided by the local transit agency. The variable "Passengers" represents the number of daily passengers on all buses in the city (in thousands). Data were collected from January 1st to December 31st. The intervention was implemented on May 1st (day 121 is the first day of the new schedule).

| Variable name | Description |
|---|---|
| Passengers | Daily passengers on the buses (in thousands) |

As policy analyst, you propose to utilize a time series model to analyze the data.

**Q1: Prepare the data for the analysis.**

- **Q1a:** Create the three variables you will need to run a time series model: Time, TimeSince, and Treatment. (5 points for each variable)
- **Q1b:** Provide a summary statistics table. (5 points)

**Q1a ANSWER:**

```
# First we need to create a variable that accounts for the time
data$Time = rep(1:365)
```

```r
# Then we create a variable that indicates days before and after treatment (i.e., before and after the 
data$Treatment = ifelse(data$Time > 120, 1, 0)

# Finally we create a variable that measures time passed after the intervention
data$TimeSince = c(rep(0, 120), rep(1:245))
```

**Q1B ANSWER:**

```r
stargazer( data,
           type = "html",
           digits = 2 )
```

Statistic

N

Mean

St. Dev.

Min

Pctl(25)

Pctl(75)

Max

passengers

365

1,403.63

90.26

1,100.00

1,333.79

1,469.41

1,700.00

Time

365

183.00

105.51

1

92

274

365

Treatment

365

0.67

0.47

0

0

1

1

TimeSince

365

82.56

81.94

0

0

154

245

**Q2: We estimate the time series model.**

- **Q2a:** Run the model and provide a result table in stargazer. (5 points + 5 points)
- **Q2b:** What does the intercept represent? (5 points)
- **Q2c:** What does the coefficient of Time represent? (5 points)
- **Q2d:** Which coefficient represents the immediate effect of the policy? (5 points)
- **Q2e:** Which coefficient represents the sustained effect of the policy? (5 points)

**Q2a ANSWER:**

```r
regTS = lm ( passengers ~ Time + Treatment + TimeSince, data = data)  # Our time series model

stargazer( regTS,
           type = "html",
           dep.var.labels = ("Wellbeing"),
           column.labels = ("Model results"),
           covariate.labels = c("Time (b1)", "Treatment (b2)", "Time Since Treatment (b3)"),
           omit.stat = "all",
           digits = 2 )
```

Dependent variable:

Wellbeing

Model results

Time (b1)

0.11

(0.18)

Treatment (b2)

21.74

(15.02)

Time Since Treatment (b3)

0.50***

(0.19)

Constant

1,327.88***

(12.42)

Note:

*p<0.1;* ***p<0.05;*** p<0.01

**Q2b ANSWER:** The intercept (coefficient b0) represents the baseline outcome - i.e., the average ridership before the intervention.

**Q2c ANSWER:** The coefficient of Time (coefficient b1) represents the slope of the line before the intervention - i.e., whether ridership was increasing or decreasing before the intervention.

**Q2d ANSWER:** The coefficient of Treatment (coefficient b2) represents the immediate effect after the intervention - i.e., whether ridership has increased or decreased immediately after the intervention.

**Q2e ANSWER:** The coefficient of TimeSince (coefficient b3) represents the sustained effect - i.e., whether the trend of ridership has changed after the intervetion compared to before the intervention.

**Q3: Now let's look at the results more closely**

- **Q3a:** Has the new schedule increased or decreased the use of public transportation in the short term? Indicate the magnitude of the effect and whether it is statistically significant. (3 * 3 points)
- **Q3b:** Has the new schedule increased or decreased the use of public transportation in the long term? Indicate the magnitude of the effect and whether it is statistically significant. (3 * 3 points)
- **Q3c:** Provide a brief (1-2 statements) possible explanation for these results. (4 points)

**Q3a ANSWER:** Given the results, we cannot reject the null hypothesis that there is no difference in ridership immediately before and after the new schedule is implemented. The coefficient of 21.74 is not significant at the 0.05 level. There is no short term (immediate) effect of the policy.

**Q3b ANSWER:** But in the long term, we observe a positive change in the ridership trend as the coefficient of TimeSince is significant at 0.05 level. The slope of the line changes after the intervention. Each day, the ridership increases of 500 rides. The coefficient is 0.5 but the unit of analysis is expressed in thousands (0.50 * 1000). We can conclude that there is a significant and positive sustained effect of the policy.

**Q3c ANSWER:** It might be that it took some time for the new schedule to attract new passengers. Therefore, we do not observe a short term effect, but in the long run, the highest bus frequency has a positive effect in attracting ridership.

**Q4:** An important aspect of a time series model is the counterfactual.

- **Q4a:** What is the number of passengers 100 days after the intervention? (5 points)
- **Q4b:** What is the counterfactual? Provide both its formula and estimation. (5 + 5 points)
- **Q4c:** What would the counterfactual be after 150 days? Provide an estimation. (5 points)
- **Q4d:** Are the two counterfactuals the same? Why? (5 points)

**Q4a ANSWER:** The number of passengers 100 days after the intervention is 1423.625.

```r
#Calculate the number of passengers 100 days after the intervention
data_100 <- as.data.frame(cbind( Time = 221, Treatment = 1, TimeSince = 100)) #
passengers_100 <- predict(regTS, data_100)

passengers_100
```

```
##        1
## 1423.625
```

**Q4b ANSWER:**

The counterfactual formula is:

$$Y = b_0 + b_1 * 221 + b_2 * 0 + b_3 * 0 + e \quad (1)$$

The counterfactual is 1351.602.

```
#Calculate the counterfactual
data_100_C <- as.data.frame(cbind( Time = 221, Treatment = 0, TimeSince = 0)) # Data if the interventio

passengers_100_C <- predict(regTS, data_100_C)
passengers_100_C
```

```
##        1
## 1351.602
```

**Q4c ANSWER:** The number of passengers after 150 days at t = 271 is 1454.135. The counterfactual is 1356.97.

```
#Calculate the number of passengers 100 days after the intervention
data_150 <- as.data.frame(cbind( Time = 271, Treatment = 1, TimeSince = 150)) #
passengers_150 <- predict(regTS, data_150)

passengers_150
```

```
##        1
## 1454.135
```

```
#Calculate the counterfactual
data_150_C <- as.data.frame(cbind( Time = 271, Treatment = 0, TimeSince = 0)) # Data if the interventio

passengers_150_C <- predict(regTS, data_150_C)
passengers_150_C
```

```
##        1
## 1356.969
```

**Q4d ANSWER:** The counterfactual 150 days after the intervetion is different as the slope of the line constantly increases and we need to account for it.

**BONUS QUESTION: Time series with a control group**

We have learned that there are threat to the validity of time series analysis. In particular, another event might have occurred at the same time of the intervention and caused the immediate and sustained effect that we observe.

A way to address this issue is to use a control group that is not subject to the intervention. This design makes sure that the effect we observe is the result of the policy intervention.

The mayor proposes to utilize city B as a control group. City B is a neighbor city with very similar characteristics to city A. Yet city B has not changed its bus schedule in the past year.

- **BQ1** Upload the data TS_Groups_Lab.csv from the class package. Look at the variables that are included and estimate a new time series model which includes the control group. (5 points)
- **BQ2** Interpret only the new coefficients in the model. Indicate whether they are stastitically significant and what they represent. Can you confirm the results of the previous model? (5 points)

**BQ1 ANSWER**

```
# setwd("O:/Users/ffusi/Jesse/Labs/TimeSeries")
data2 = read.csv("TS_Groups_lab.csv")
```

```r
regTS = lm ( passengers ~ Time + Treatment + TimeSince + Group + Time * Group + Treatment * Group + Tim

stargazer( regTS,
           type = "html",
           dep.var.labels = ("Wellbeing"),
           column.labels = ("Model results"),
           covariate.labels = c("Time (b1)", "Treatment (b2)", "Time Since Treatment (b3)", "Group (b4)
           omit.stat = "all",
           digits = 2 )
```

Dependent variable:

Wellbeing

Model results

Time (b1)

0.11

(0.07)

Treatment (b2)

-5.72

(5.51)

Time Since Treatment (b3)

0.15**

(0.07)

Group (b4)

-30.63***

(6.45)

Time*Group (b5)

0.47***

(0.09)

Treatment*Group (b6)

7.75

(7.80)

TimeSince*Group (b7)

1.16***

(0.10)

Constant

1,171.88***

(4.56)

Note:

*p<0.1;* ***p<0.05;*** p<0.01

**BQ2 ANSWER**

- b4 (Group) is significant and negatively correlated with ridership. This indicates that daily ridership in the treated city (City A) is, on average, lower than the ridership in the control city (City B).
- b5 shows that before the treatment, the treated city's ridership was growing at a faster rate than the ridership in the control city. The difference between the two slopes is positive and significant.
- b6 shows that after the treatment there is no significant difference between the two cities. As before, we observe no immediate effect.
- But b7 is positive and significant, suggesting that after the treatment City A (Treament) grows at an even higher rate than City B (Control group). This confirms our previous findings.