Homework 2

Boyu Jiang

9/11/2021

Problem 2

Part A

My learning objectives in this class:

- Getting familiar with R programming, typesetting, and version control;
- Mastering various data visualization tools;
- Learning data handling skills in R.

Part B

1. Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\ exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right],\ \sigma > 0 \tag{1}$$

2. Exponential Distribution

$$f(x) = \frac{1}{\beta}exp(-\frac{x}{\beta}),\ x \geq 0,\ \beta > 0 \tag{2}$$

3. Cauchy Distribution

$$f(x) = \frac{1}{\pi\sigma}\frac{1}{1+(\frac{x-\theta}{\sigma})^2},\ \sigma > 0 \tag{3}$$

Problem 3

Steps in performing Reproducible Research:

1. Before analysis: data storage and organization.

- Storing raw data in multiple locations using multiple media;
- Storing final data in a portable and non-proprietary format;
- Formatting final data appropriately for analysis. (Challenge: I found it hard to process raw data when there are various formatting problems.)

2. During analysis: best coding practices.

- Making code clean, readable, and appropriately formatted;
- Commenting code thoroughly;
- Inviting at least one collaborator to review data and code;
- Documenting all software versions and computing environments.

3. After analysis: finalizing results and sharing.

- Giving explicit instructions on locating data, metadata, and code in the manuscript; (Challenge: although my manuscript is full of figure, table, and formula, readers or reviewers think it is indistinct.)
- Sharing data, metadata, and code at a permanent site.

Problem 4

```
library(data.table)
covid_raw <- fread("https://opendata.ecdc.europa.eu/covid19/casedistribution/csv")
us <- covid_raw[covid_raw$countriesAndTerritories == 'United_States_of_America',]
us_filtered <- us[us$month %in% c(6:7),]
us_filtered$index <- rev(1:dim(us_filtered)[1])
fit<-lm(`Cumulative_number_for_14_days_of_COVID-19_cases_per_100000`~index,
        data=us_filtered)
```

Part A

1.

```
library(knitr)
kable(summary(us_filtered[, c('cases','deaths','popData2019',
                      'Cumulative_number_for_14_days_of_COVID-19_cases_per_100000')]),
      format = 'simple',align='l',col.names=c('cases','deaths','popData2019',
                                              'Cumul.COVID19.Cases for 14 days per 100k'))
```

cases	deaths	popData2019	Cumul.COVID19.Cases for 14 days per 100k
Min. :18665	Min. : 242.0	Min. :329064917	Min. : 89.76
1st Qu.:25540	1st Qu.: 500.0	1st Qu.:329064917	1st Qu.: 92.43
Median :45221	Median : 767.0	Median :329064917	Median :150.94
Mean :44666	Mean : 791.6	Mean :329064917	Mean :170.16
3rd Qu.:61796	3rd Qu.: 982.0	3rd Qu.:329064917	3rd Qu.:247.01
Max. :78427	Max. :2437.0	Max. :329064917	Max. :282.72

```
library(plyr)
kable(count(us_filtered, vars=c('month','year','countriesAndTerritories','geoId',
                        'countryterritoryCode','continentExp')), align='l')
```

month	year	countriesAndTerritories	geoId	countryterritoryCode	continentExp	freq
6	2020	United_States_of_America	US	USA	America	30
7	2020	United_States_of_America	US	USA	America	31

Since the time period was limited from June to July, there are 61 time points in us_filtered data. From these two tables, no missing value exists.

2.

```
library(stargazer)
```

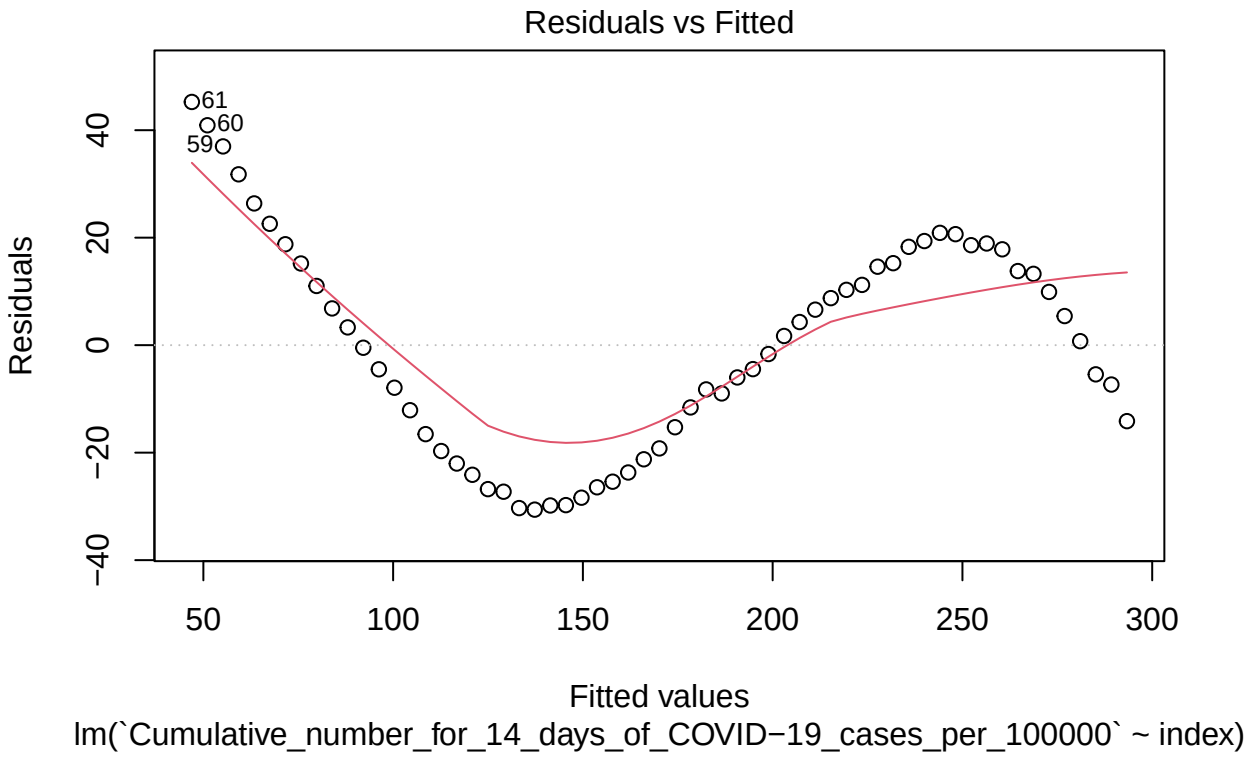
```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
stargazer(fit, title= "Fit Results", align=TRUE, type = 'text')
```

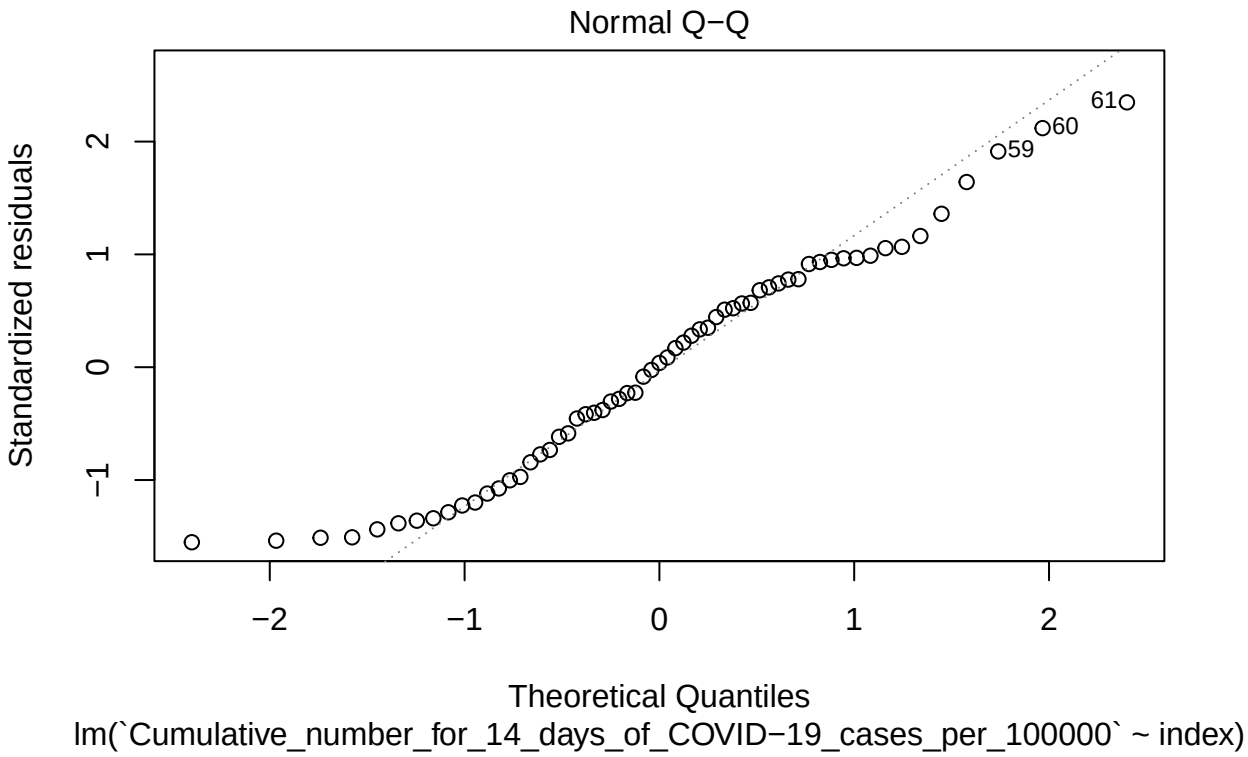
```
##
## Fit Results
## ================================================================================
##                                          Dependent variable:
##                     ---------------------------------------------------------
##                     `Cumulative_number_for_14_days_of_COVID-19_cases_per_100000`
## --------------------------------------------------------------------------------
## index                                          4.107***
##                                                (0.145)
##
## Constant                                      42.853***
##                                                (5.165)
##
## --------------------------------------------------------------------------------
## Observations                                     61
## R2                                              0.932
## Adjusted R2                                     0.930
## Residual Std. Error                       19.922 (df = 59)
## F Statistic                             803.464*** (df = 1; 59)
## ================================================================================
## Note:                                                *p<0.1; **p<0.05; ***p<0.01
```

Part B

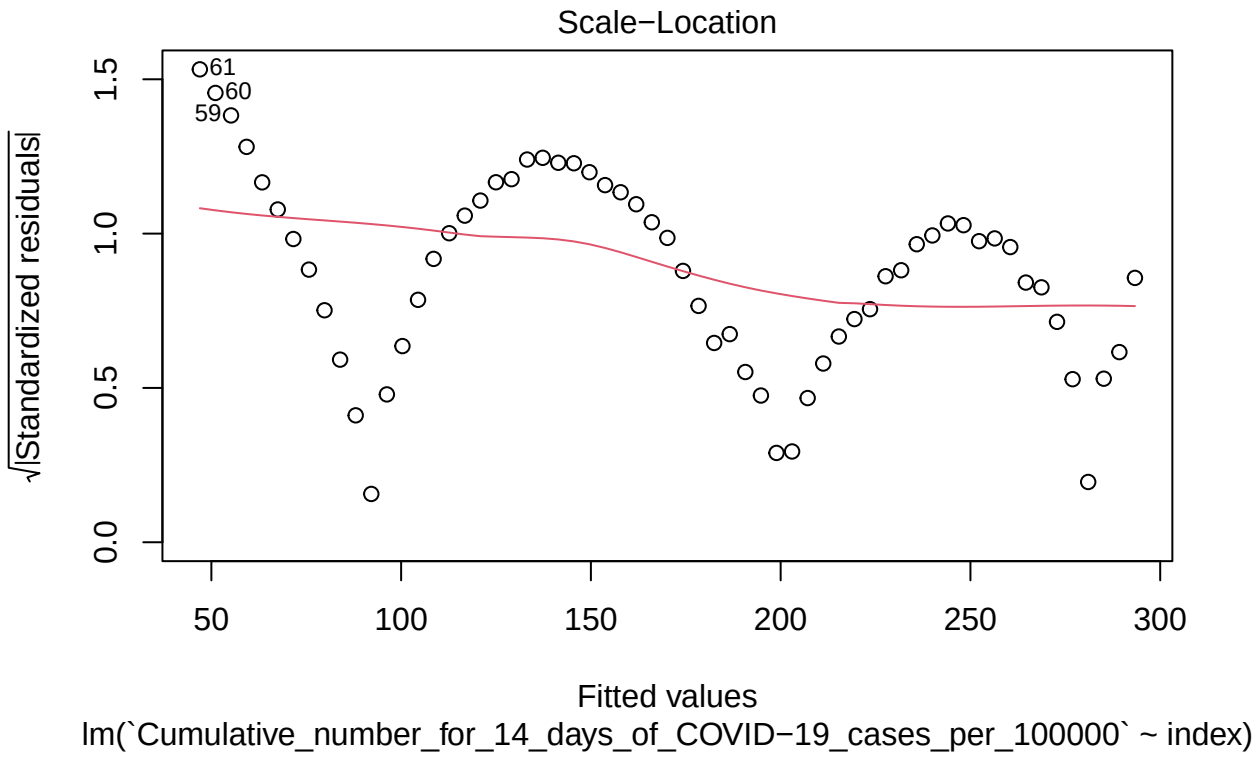
```
library(broom)
fit.diags <- broom::augment(fit)
plot(fit, 1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(`Cumulative_number_for_14_days_of_COVID−19_cases_per_100000` ~ index)
61
60
59
40
20
0
−20
−40
50
100
150
200
250
300

```
plot(fit, 2)
```


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(`Cumulative_number_for_14_days_of_COVID−19_cases_per_100000` ~ index)
61
60
59
−2
−1
0
1
2

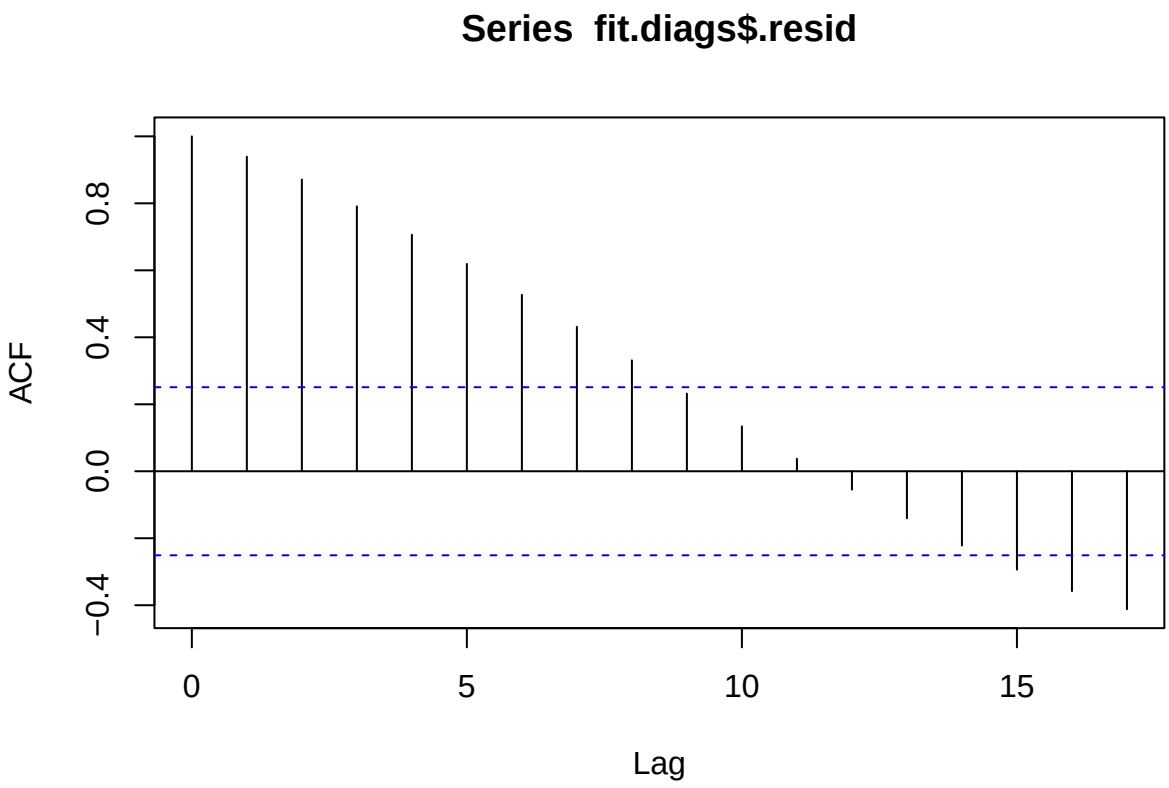
```
plot(fit, 3)
```


Scale−Location
√|Standardized residuals|
61
60
59
0.0
0.5
1.0
1.5
50
100
150
200
250
300
Fitted values
lm(`Cumulative_number_for_14_days_of_COVID−19_cases_per_100000` ~ index)

```
plot(fit, 5)
```

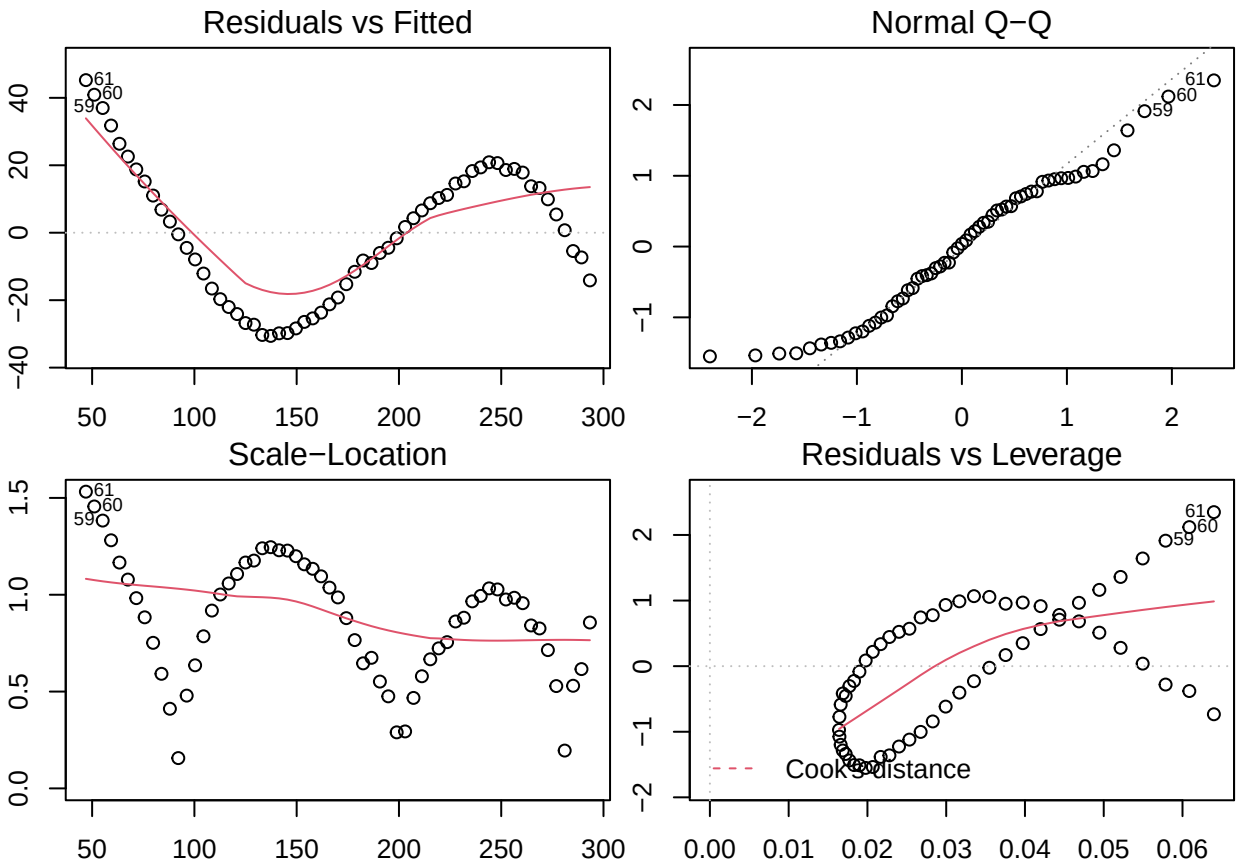
Part C

```
acf(fit.diags$.resid,type = "correlation")
```



Problem 5

```
par(mfrow=c(2,2),mar = c(2,2,1.5,0.5))
plot(fit, 1)
plot(fit, 2)
plot(fit, 3)
plot(fit, 5)
```


Residuals vs Fitted
Normal Q−Q
Scale−Location
Residuals vs Leverage
Cook's distance
61
60
59