HW2_Anbin

Anbin Rhee

9/14/2021

Problem 1

I read the article and checked out the tutorials.

Problem 2

Part A

Below are things I hope to get from this Stat Progr Packages class.

- Learning how to use GitHub and getting used to it
- Being better at visualizing data using R
- Being proficient in reproducible research
- Learning how to use variety of R packages (Good programming practice)

Prat B

Below three distributions are Normal, Gamma, Exponential distributions, respectively.

Normal Distribution

$$f(X = x|\mu, \sigma^2) \quad = \quad \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}};\ -\infty < x < \infty;\ -\infty < \mu < \infty, \sigma > 0 \tag{1}$$

Gamma Distribution

$$f(X = x|\alpha, \beta) \quad = \quad \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}};\ 0 \leq x < \infty;\ \alpha, \beta > 0 \tag{2}$$

Exponential Distribution

$$f(X = x|\beta) \quad = \quad \frac{1}{\beta} e^{-\frac{x}{\beta}};\ 0 \leq x < \infty;\ \beta > 0 \tag{3}$$

Problem 3

1. Record the way it was produced for every result. It might be challenging for me when the project is huge and complicated.
2. As manual procedures are hard to reproduce as well as inefficient and error-prone, it is recommended to avoid manual data manipulation steps. In order to avoid such steps, it is important to enhance my programming skills.
3. Record exact versions of every external programs used in the research. If there are lots of external programs in the project, it should be thoroughly examined.

4. Use a version control system for all custom scripts. It is important to know how to run such systems.
5. In standardized format, record every intermediate results. It could be challenging when there are lots of results in the research.
6. For visual consistency between figures, store raw data behind plots. If one data is connected to multiple plots, it is important to store it for all of connected plots.
7. To validate and understand the research result, it is good to inspect the summaries and generating hierarchical analysis output is useful for doing it. It is critical to know each steps of the hierachical anlysis.
8. Give public access to all scripts, runs, and results in the research. It is crucial to make public to easily access to the research.

Problem 4

Part A

```
summary(us_filtered)
```

```
##    dateRep               day            month            year
##  Length:61          Min.   : 1.00   Min.   :6.000   Min.   :2020
##  Class :character   1st Qu.: 8.00   1st Qu.:6.000   1st Qu.:2020
##  Mode  :character   Median :16.00   Median :7.000   Median :2020
##                     Mean   :15.75   Mean   :6.508   Mean   :2020
##                     3rd Qu.:23.00   3rd Qu.:7.000   3rd Qu.:2020
##                     Max.   :31.00   Max.   :7.000   Max.   :2020
##      cases           deaths       countriesAndTerritories    geoId
##  Min.   :18665   Min.   : 242.0   Length:61               Length:61
##  1st Qu.:25540   1st Qu.: 500.0   Class :character        Class :character
##  Median :45221   Median : 767.0   Mode  :character        Mode  :character
##  Mean   :44666   Mean   : 791.6
##  3rd Qu.:61796   3rd Qu.: 982.0
##  Max.   :78427   Max.   :2437.0
##  countryterritoryCode  popData2019        continentExp
##  Length:61            Min.   :329064917   Length:61
##  Class :character     1st Qu.:329064917   Class :character
##  Mode  :character     Median :329064917   Mode  :character
##                       Mean   :329064917
##                       3rd Qu.:329064917
##                       Max.   :329064917
##  Cumulative_number_for_14_days_of_COVID-19_cases_per_100000     index
##  Min.   : 89.76                                              Min.   : 1
##  1st Qu.: 92.43                                              1st Qu.:16
##  Median :150.94                                              Median :31
##  Mean   :170.16                                              Mean   :31
##  3rd Qu.:247.01                                              3rd Qu.:46
##  Max.   :282.72                                              Max.   :61
```

```
sum(is.na(us_filtered))
```

```
## [1] 0
```

We have limited ourselves to 61 time points. As the sum of NA is zero, there is no missing value.

Part B

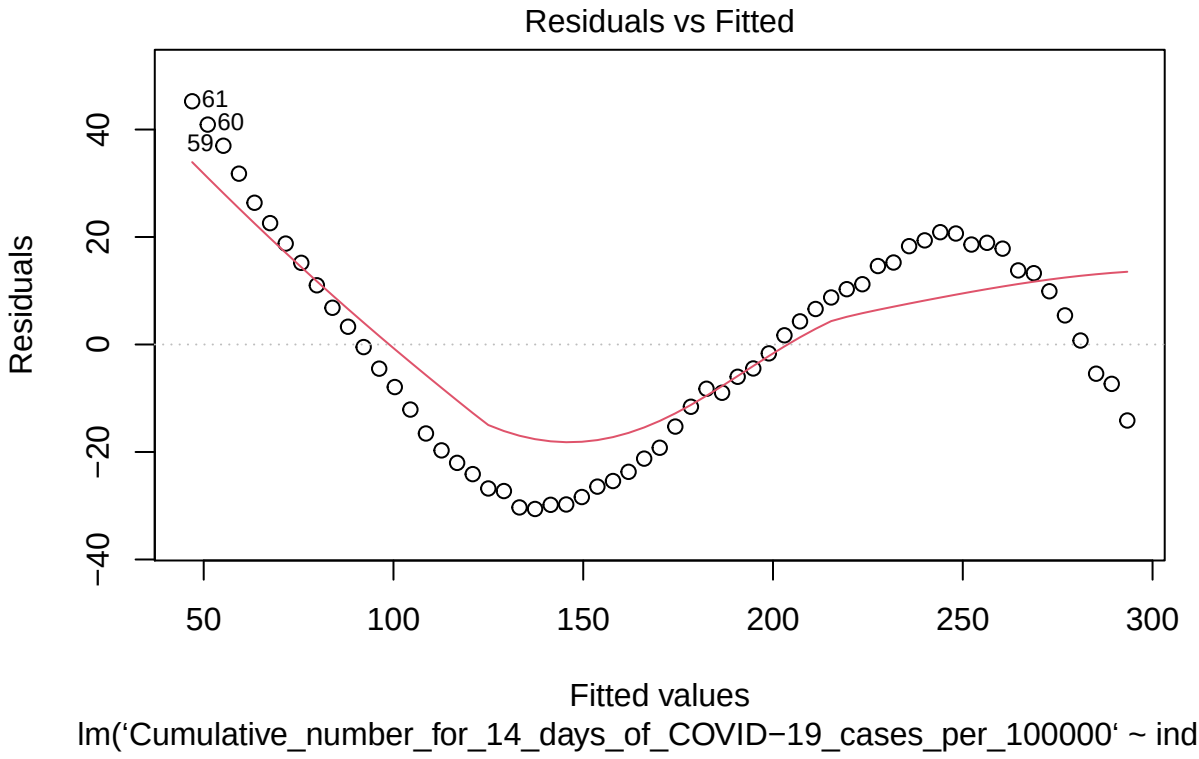
I plotted residuals vs fitted plot, normal Q-Q plot, scale-location plot, residuals vs leverage plot.

```
##
## 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
```

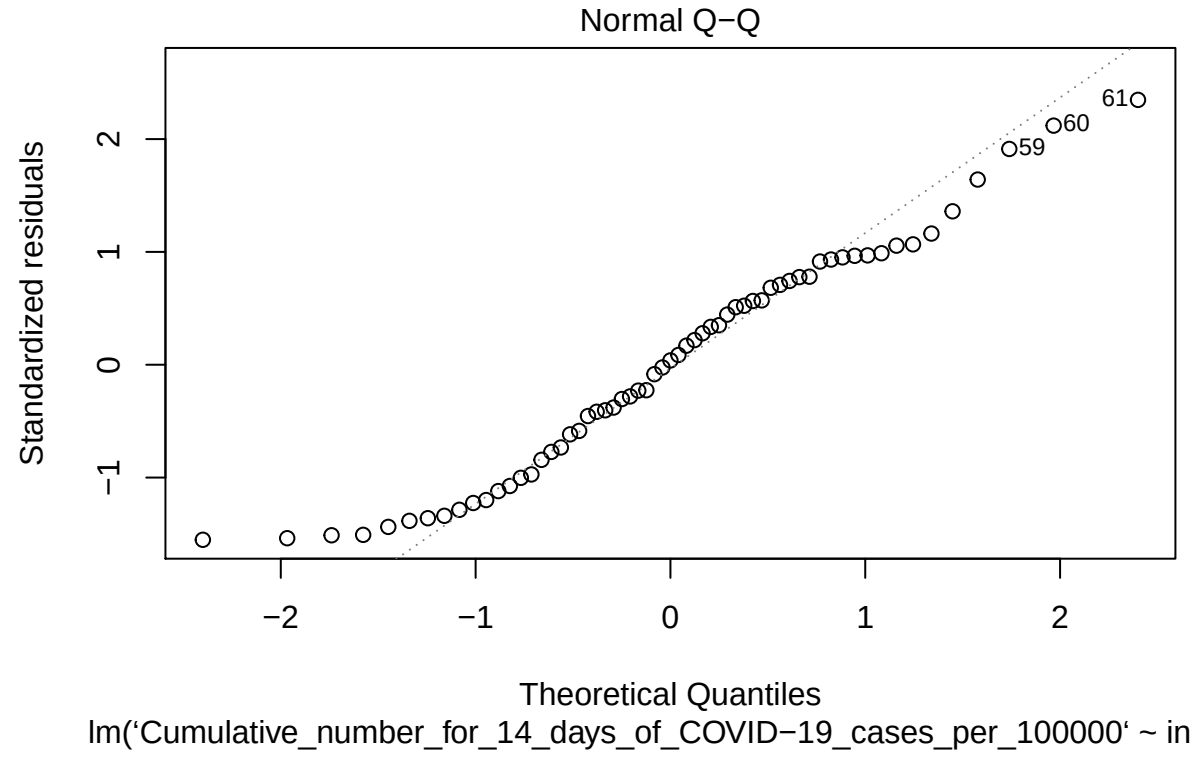
```
summary.lm(fit)
```

```
##
## Call:
## lm(formula = `Cumulative_number_for_14_days_of_COVID-19_cases_per_100000` ~
##     index, data = us_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.602 -16.555   0.738  15.196  45.251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.8532     5.1649   8.297 1.72e-11 ***
## index         4.1065     0.1449  28.345  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.92 on 59 degrees of freedom
## Multiple R-squared:  0.9316, Adjusted R-squared:  0.9304
## F-statistic: 803.5 on 1 and 59 DF,  p-value: < 2.2e-16
```

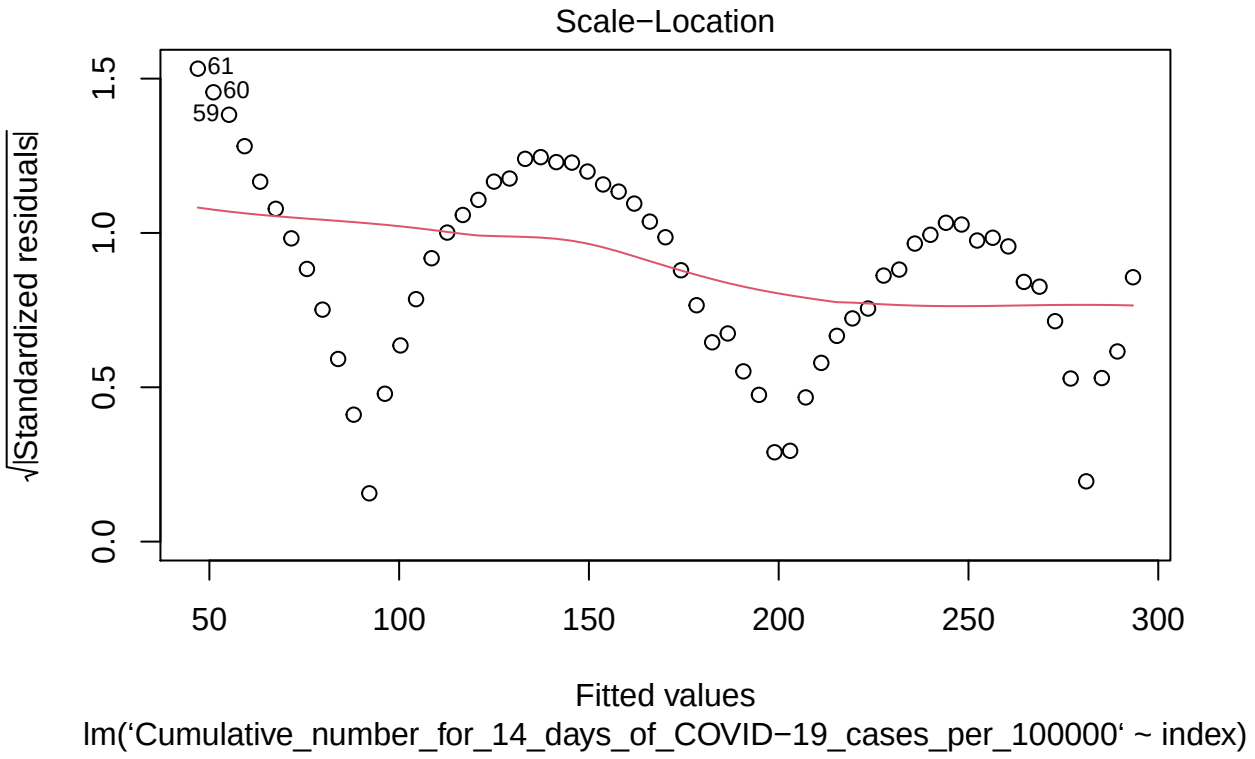
```
fit.diags <- broom::augment(fit)
plot(fit,1)
```



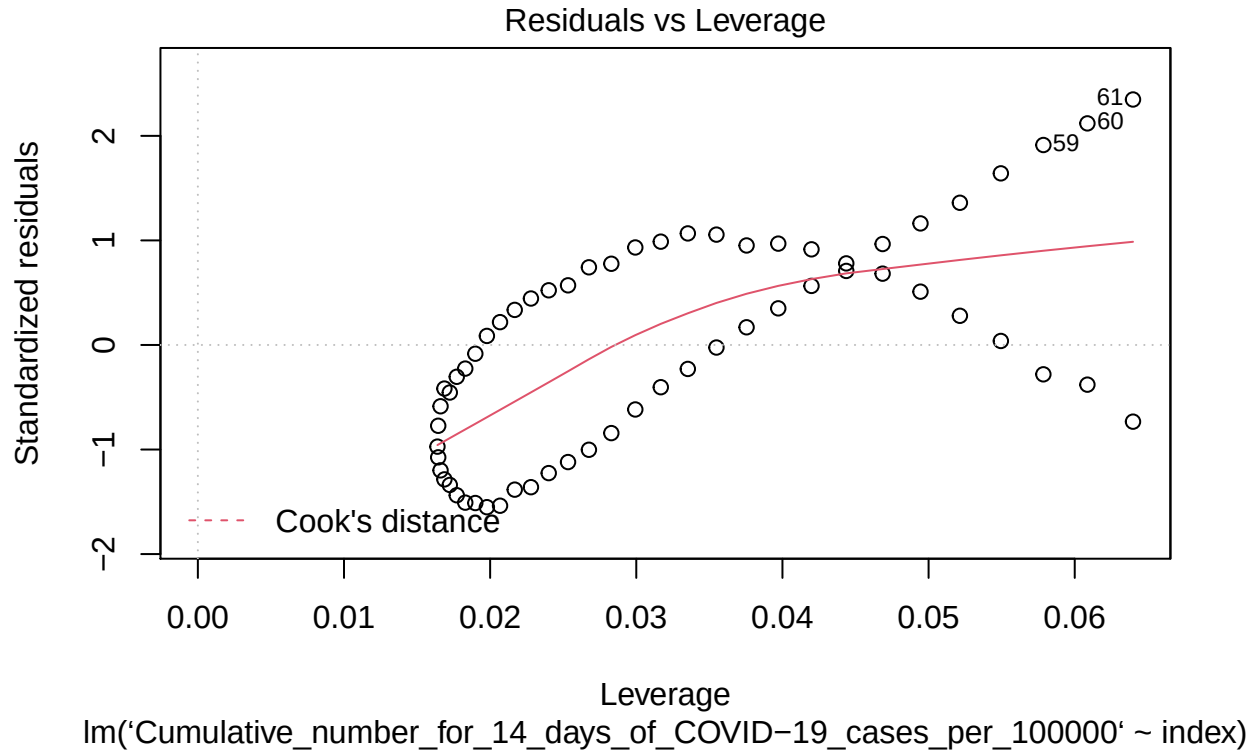

```
plot(fit,2)
```




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



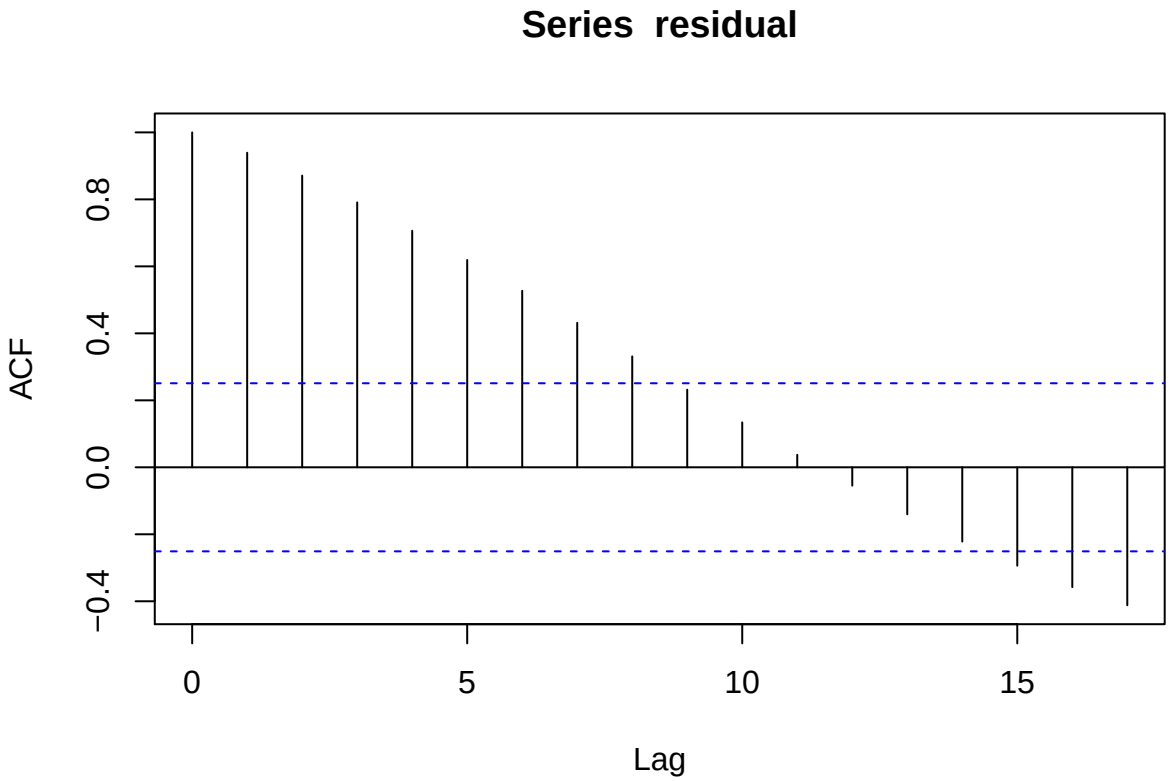

```
plot(fit,5)
```



Part C

I created auto correlation plot of the residuals.

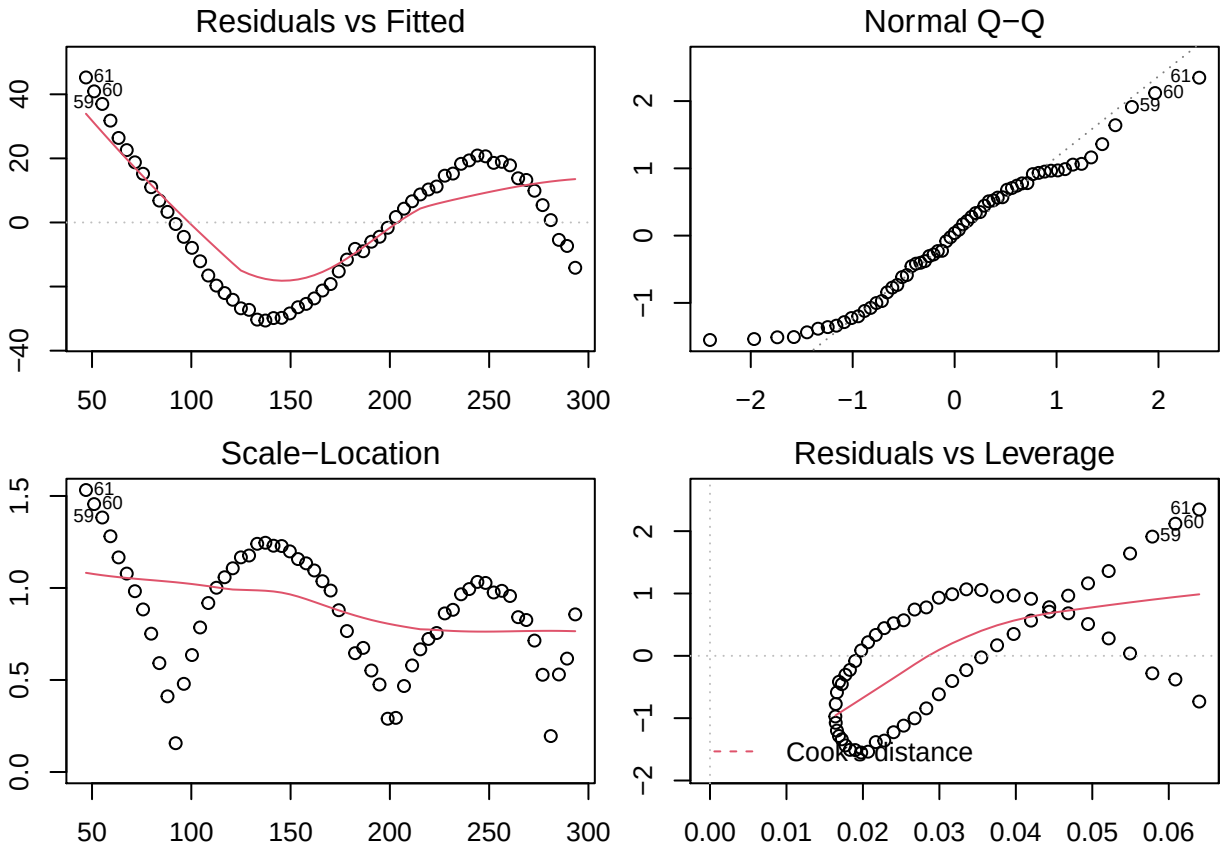
```
residual <- fit.diags$.resid
acf(residual)
```



Problem 5

I combined the four plots from Problem 4 into a single plot with the smallest margin.

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



Problem 6

I knitted this documentto PDF and pushed to GitHub.