STAT5014 Hwk 2

Samuel Soon

9/10/2021

Problem 2

a.

Since I have a little prior experience with R, my goals for this class will be to strengthen my knowledge of basic R concepts, and hopefully get used to operations such as data tidying, scraping, etc.

Goals

- Improve on creating graphs using ggplot or other software
- Learn how to write my own functions for estimating parameters
- Learn how to use latex outside of mathematical expressions

b.

$$\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(\frac{-(x-\mu)}{\sqrt{2}\sigma})^2} \tag{1}$$

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{2}$$

$$f_X(x) = 0.5, a \leq x \leq a+2 \tag{3}$$

Problem 3

- Step 1: Results of interest should have their related procedures, parameters, and so on recorded down for reproduceability.

  In projects requiring collaboration between multiple people, it could be difficult to create a centralized archive of steps taken.

- Step 2: Manual computation is often inefficient and prone to human error. Automating processes outside of specific situations allows for more consistently effective analysis.

  Some processes are so esoteric that automating may be deemed inefficient compared to the time it takes to manually calculate a result.

- Step 3: Because of the various dependencies certain versions of programs may require, saving current versions will help ensure that results can be easily reproduced, and won't be lost due to version updates of related programs.

  Saving virtual machines for every version of an experiment might be expensive in terms of computer storage?

- Step 4: Keeping track of custom scrips is necessary for ensuring that certain results can be 100

  Human error may make keeping track of every change in code difficult.

- Step 5: Intermediate results provide a good view of the consequences of idiosyncrasies within procedures. Recording intermediate results allows processes to be examined in more detail, which gives a greater understanding of the final result.

  People can be lazy.

- Step 6: Since analysis often uses random variables to draw conclusions from, and seeds are used to approximate randomness, archiving the seed used in an experiment allows results of the experiment to be reproduced in the future.

  Saving the seed may not be useful to other parties if they do not also know the algorithm for random number generation you used.

- Step 7: In analysis, data often needs to be readjusted for tidying purposes, or data transformation. In these cases, it is helpful to have raw data saved, as said data can easily be transformed. If the data is not stored, the analysis will have to be done again, wasting time.

  Again, human error can cause raw data to be lost.

- Step 8: When presenting findings, it is helpful for readers if you provide ways to further learn about the details of your research. Within an article, links or references to more detailed articles should be provided, so that the audience can easily examine your findings in greater detail.

  You may not necessarily have access to resources that help your audience with understanding your research.

- Step 9: When drawing conclusions, it is best if you include references to results early on, so that audiences can examine results themselves, and track down the sources of your conclusions.

  Audiences may not be familiar with your presented material. You may need to take steps to ease readers in to what your research deals with.

- Step 10: All non-confidential contents of your research should be available to the public, so that they can reproduce and verify your results.

  In certain cases, results may not be able to be reproduced without confidential information being given.

4.

a.

```
## [1] "Summary Table:"
```

dateRep	day	month	year	cases	deaths	countriesAndTerr
Length:61	Min. : 1.00	Min. :6.000	Min. :2020	Min. :18665	Min. : 242.0	Length:61
Class :character	1st Qu.: 8.00	1st Qu.:6.000	1st Qu.:2020	1st Qu.:25540	1st Qu.: 500.0	Class :character
Mode :character	Median :16.00	Median :7.000	Median :2020	Median :45221	Median : 767.0	Mode :character
NA	Mean :15.75	Mean :6.508	Mean :2020	Mean :44666	Mean : 791.6	NA
NA	3rd Qu.:23.00	3rd Qu.:7.000	3rd Qu.:2020	3rd Qu.:61796	3rd Qu.: 982.0	NA
NA	Max. :31.00	Max. :7.000	Max. :2020	Max. :78427	Max. :2437.0	NA

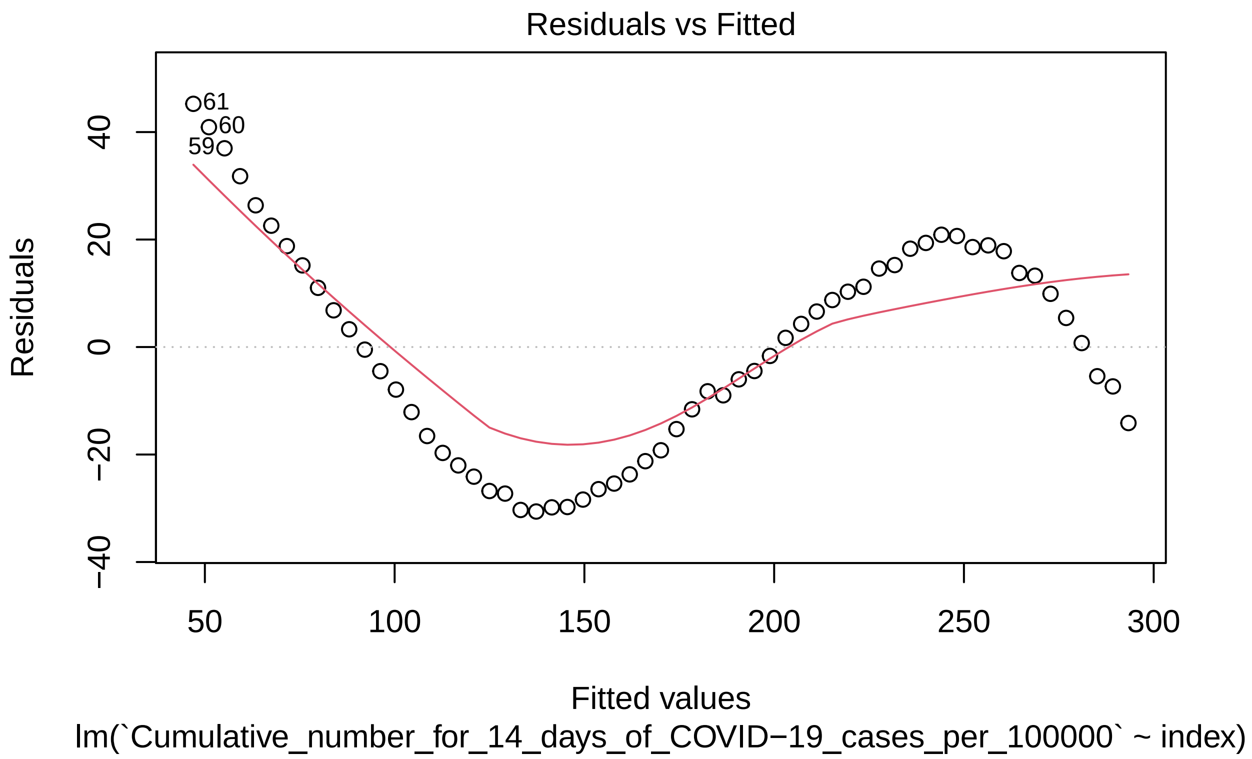
```
## [1] "There are 3 time points: Day, Month, Year."
```

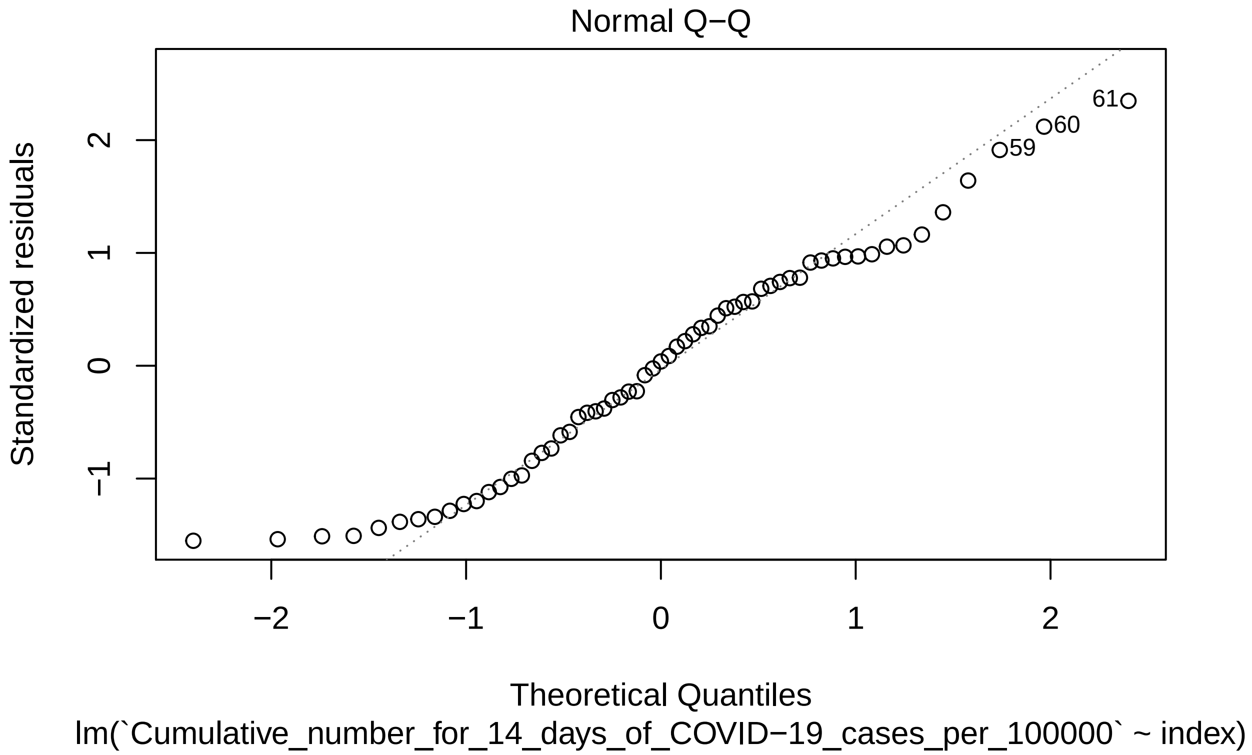
```
## [1] "Number of Missing Values in Each Column:"
##                                                    dateRep
##                                                          0
##                                                        day
##                                                          0
##                                                      month
##                                                          0
##                                                       year
##                                                          0
##                                                      cases
##                                                          0
##                                                     deaths
##                                                          0
##                                    countriesAndTerritories
##                                                          0
##                                                      geoId
##                                                          0
##                                       countryterritoryCode
##                                                          0
##                                                popData2019
##                                                          0
##                                               continentExp
##                                                          0
## Cumulative_number_for_14_days_of_COVID-19_cases_per_100000
##                                                          0
##                                                      index
##                                                          0
```

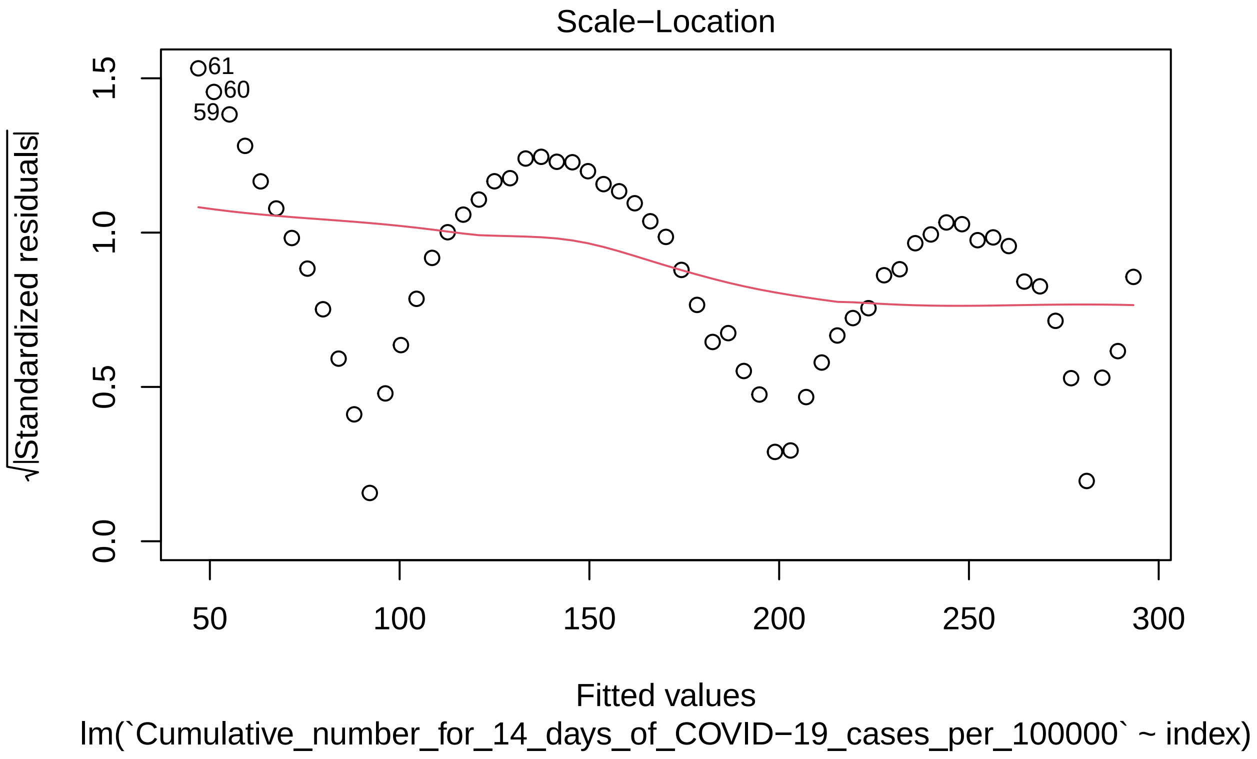
```
## [1] "LM Summary:"
```

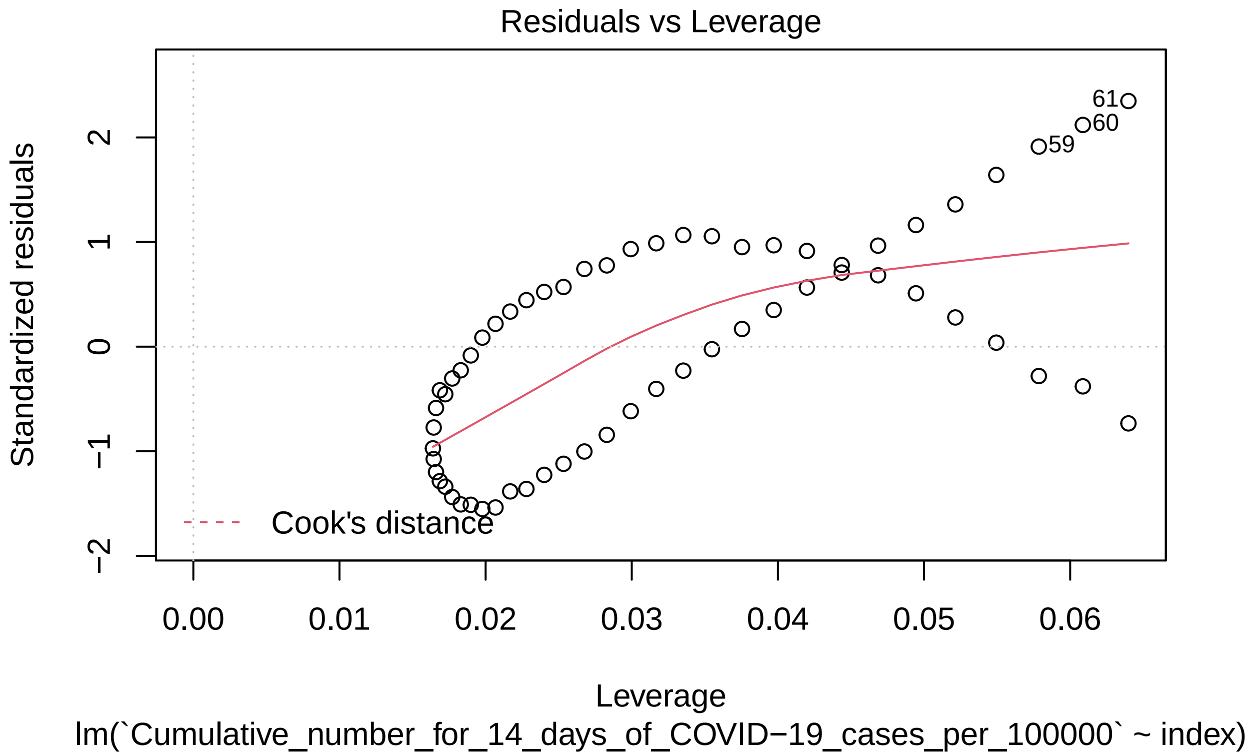
```
## 
## ================================================================================
##                                         Dependent variable:                     
##                     ------------------------------------------------------------
##                     `Cumulative_number_for_14_days_of_COVID-19_cases_per_100000`
## --------------------------------------------------------------------------------
## 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
```

b.

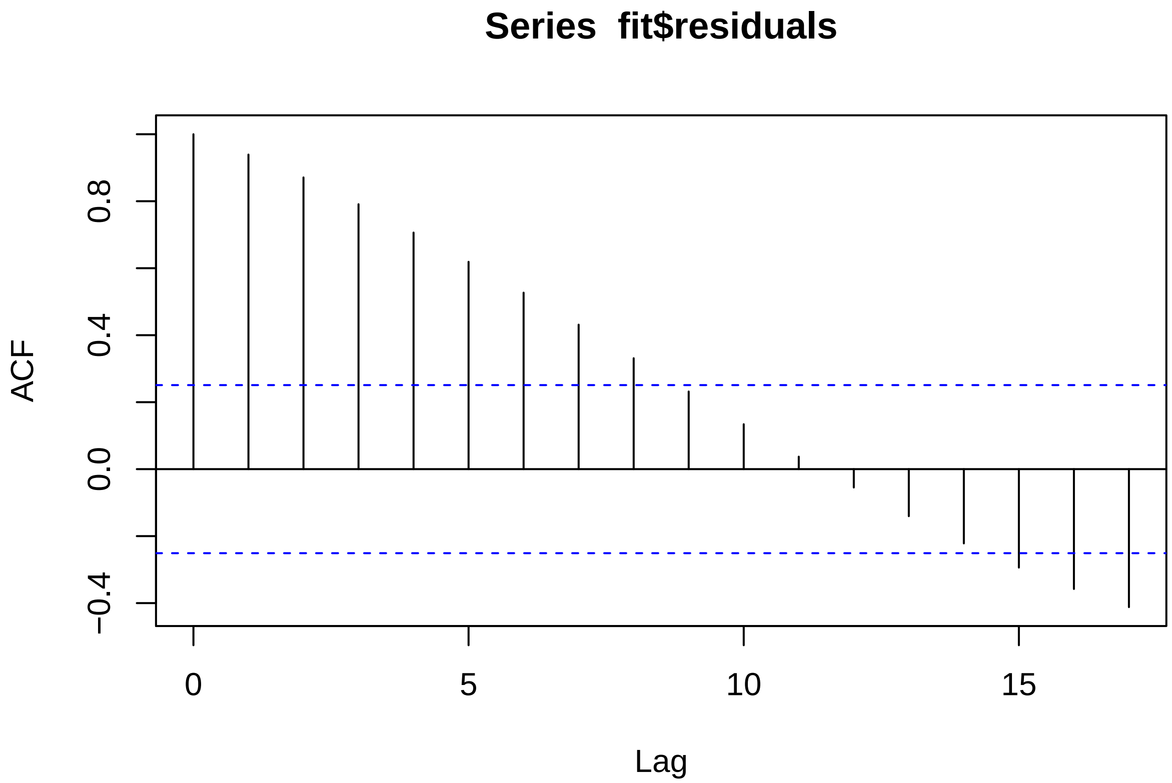

Residuals vs Fitted
Residuals
40
20
0
−20
−40
61
60
59
50
100
150
200
250
300
Fitted values
lm(`Cumulative_number_for_14_days_of_COVID−19_cases_per_100000` ~ index)


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

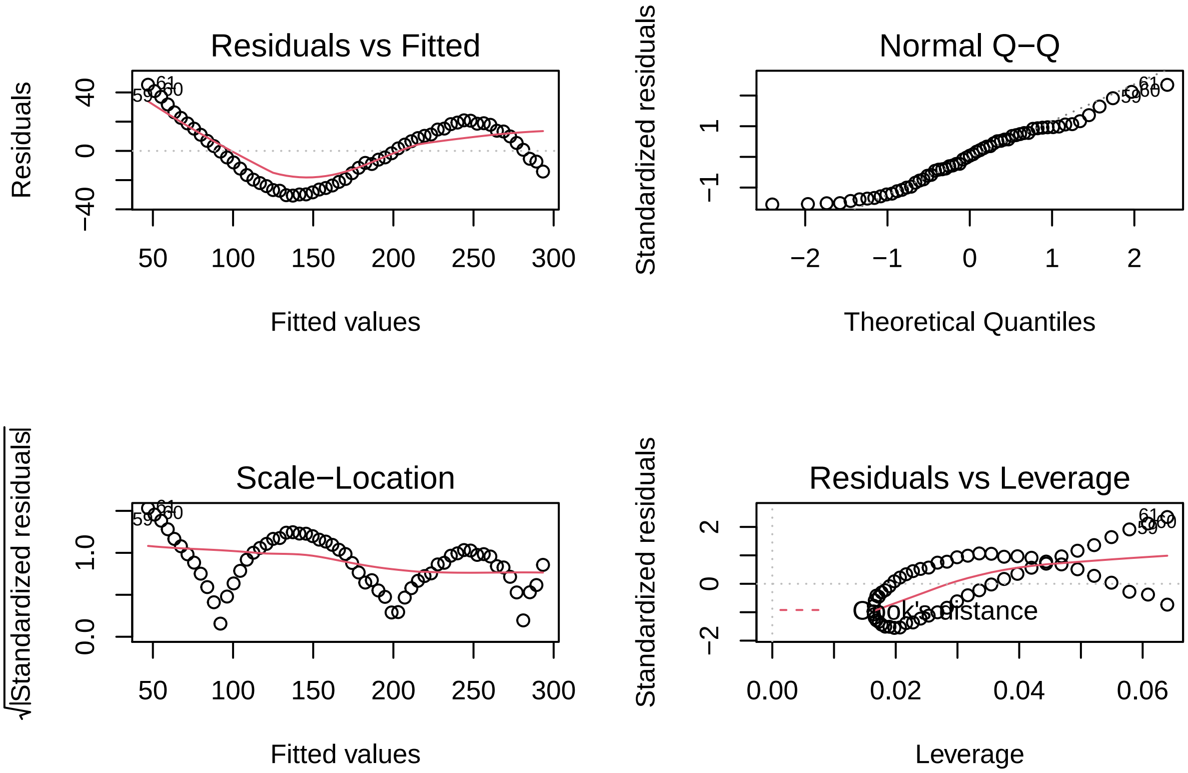

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
61
60
59
50
100
150
200
250
300
Fitted values
lm(`Cumulative_number_for_14_days_of_COVID−19_cases_per_100000` ~ index)


Residuals vs Leverage
Standardized residuals
2
1
0
−1
−2
61
60
59
Cook's distance
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Leverage
lm(`Cumulative_number_for_14_days_of_COVID−19_cases_per_100000` ~ index)

c.



5.


Residuals vs Fitted
Residuals
40
0
−40
50
100
150
200
250
300
Fitted values
59
61
60
Normal Q−Q
Standardized residuals
1
−1
−2
−1
0
1
2
Theoretical Quantiles
61
59
60
Scale−Location
√|Standardized residuals|
1.0
0.0
50
100
150
200
250
300
Fitted values
59
61
60
Residuals vs Leverage
Standardized residuals
2
0
−2
0.00
0.02
0.04
0.06
Leverage
Cook's distance
61
59
60