

HW2_arflowers

Anna Flowers

9/8/2021

Homework 2

Problem 2

Part A

I am excited for this class because although I have been working with R for about 3 years now, I have never taken a class specifically designed to teach it. The knowledge I have of R has come from classes that used it to accompany the material (so any teaching revolved around that class), or from internet sources that I used to teach myself. I am especially excited to learn more about LaTeX, because although I have used it in the past I have relied on the support of others to create a coherent document. I also want to experience more of the combination of R and LaTeX, because in the past I have really only used LaTeX through Overleaf in documents entirely in LaTeX. I also want to learn more about Github, because although I have used it in the past I have exclusively relied on repositories used by other people and am not used to creating my own material yet.

1. Learn more efficient ways to accomplish tasks I may have been over complicating.
2. Become more confident in my skills using LaTeX.
3. Become more comfortable using Github with repositories that I have created myself.

Part B

Gamma Density Function:

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

Chi squared Density Function:

$$f(x|p) = \frac{1}{\Gamma(p/2) 2^{p/2}} x^{(p/2)-1} e^{-x/2}; 0 \leq x < \infty; p = 1, 2, \dots \quad (2)$$

Lognormal Density Function:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \frac{e^{-(\log x - \mu)^2 / (2\sigma^2)}}{x}; 0 \leq x < \infty; -\infty < \mu < \infty \quad (3)$$

Problem 3

1. Keep track of how all results were produced
 - Challenges: These results are often produced through a lot of trial and error (especially when coding is involved), so it can be difficult to separate the steps that were important and those that were unnecessary.
2. Avoid manually manipulating data
 - Challenges: Some machines may not be able to read specific data types, but opening the file with a different type may change the data without your knowledge.
3. Archive (or keep track of) the exact versions of all programs used
 - Challenges: Even if you do archive the program, some updates will automatically delete older versions on the computer, possibly rendering past research useless.
4. Version control all scripts
 - Challenges: It can be difficult to know when to store versions of code and when not to. Too many versions of code can still make the correct version difficult to find, and too few means you are less likely to have the exact code that you want.
5. Record all intermediate results and standardize if possible
 - Challenges: Intermediate steps might not be in a data form that is easy to save, so it might not be possible to keep track of all intermediate steps.
6. Note seeds used for analyses that include randomness
 - Challenges: Using seeds for randomness may not be appropriate in the context of the experiment.
7. Store raw data used to make plots
 - Challenges: Large data sets may require more storage space than what is available.
8. Keep and inspect all layers of detail of the data
 - Challenges: Amount of data to inspect can grow quickly if there are a lot of layers of data.
9. Connect statements to the results that inspired them
 - Challenges: Research in a specialized field can be difficult to explain to the general public.
10. Provide public access to all data used, programs written, and results discovered
 - Challenges: Some data is not publicly available and perhaps must be purchased, so it cannot be included with the paper.

Problem 4

```
#install.packages('data.table')
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

```
library(knitr)
kable(summary(us_filtered))
```

Part 1

	dateReplay	month	year	cases	deaths	countries	Area	Territories	systems	pop	Day	2019	CFR	Cumulative_number	index	for_14_days_of
														19_cases_per_100000		
Length:	Min.	Min.	Min.	Min.	Min.	Length:61	Length:61	Length:61	Length:61	Min.	Length:61	Min.	Length:61	Min. : 89.76	Min.	
	:	:	:	:	:					:	:	:	:		:	
	1.00	:6.000	:2020	:18665:	:242.0					:329064917					:1	
Class	1st	1st	1st	1st	1st	Class	Class	Class	1st	Class	1st	Qu.: 92.43	1st			
:char-	Qu.: 8.00	Qu.:6.000	:2020	:25640:	:500.0	:char-	:char-	:char-	Qu.:329064917				Qu.:16			
Mode	Median	Median	Median	Median	Median	Mode	Mode	Mode	Median	Mode	Median	:150.94	Median			
:char-	:16.00	:7.000	:2020	:45221:	:767.0	:char-	:char-	:char-	:329064917				:31			
NA	Mean	Mean	Mean	Mean	Mean	NA	NA	NA	Mean	NA	Mean	:170.16	Mean			
	:15.75	:6.508	:2020	:44666:	:791.6				:329064917				:31			
NA	3rd	3rd	3rd	3rd	3rd	NA	NA	NA	3rd	NA	3rd	Qu.:247.01	3rd			
	Qu.:23000	:7.000	:2020	:61796:	:982.0				Qu.:329064917				Qu.:46			
NA	Max.	Max.	Max.	Max.	Max.	NA	NA	NA	Max.	NA	Max.	:282.72	Max.			
	:31.00	:7.000	:2020	:78427:	:2437.0				:329064917				:61			

This data is limited to 61 time points from June 2020 to July 2020. There are no missing points, since there are 30 days in June and 31 in July, so that gives a total of 61 days to survey.

```
library(stargazer)
```

Part 2

```
##
## 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)
```

Part B

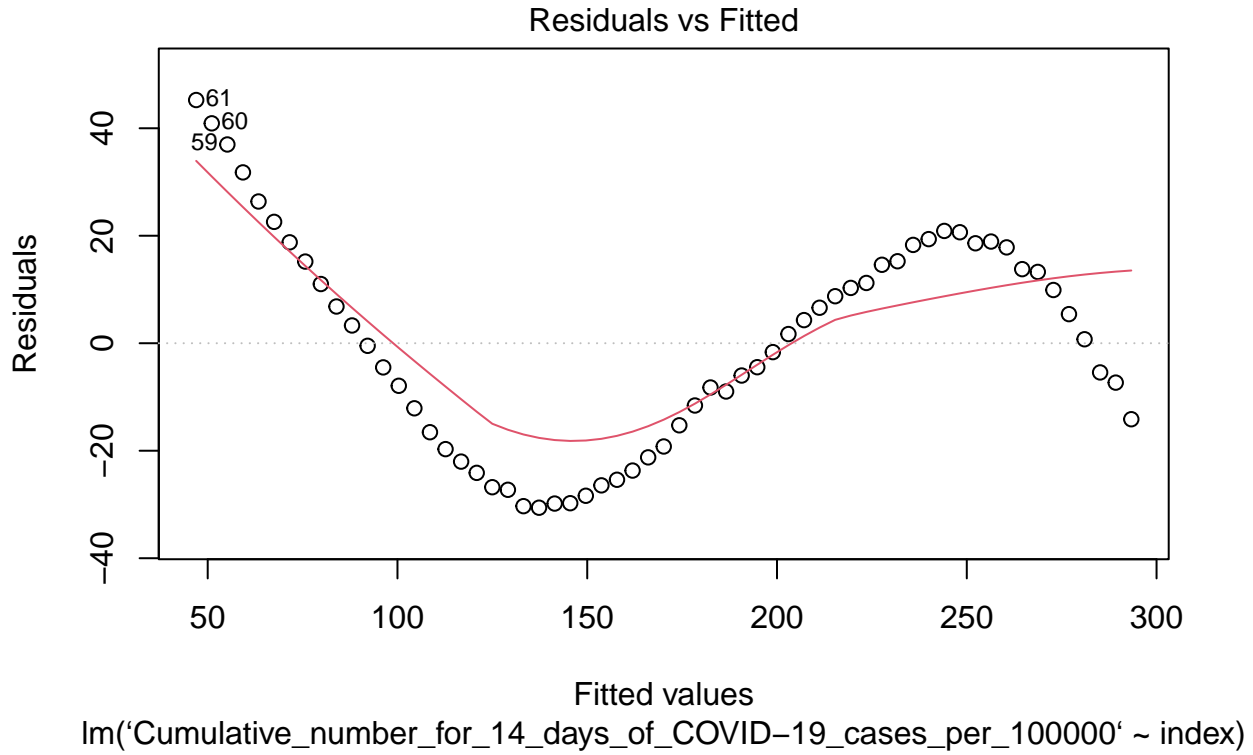
```
#install.packages("broom")
fit.diags <- broom::augment(fit)
plot(fit,c(1:3,5))
```

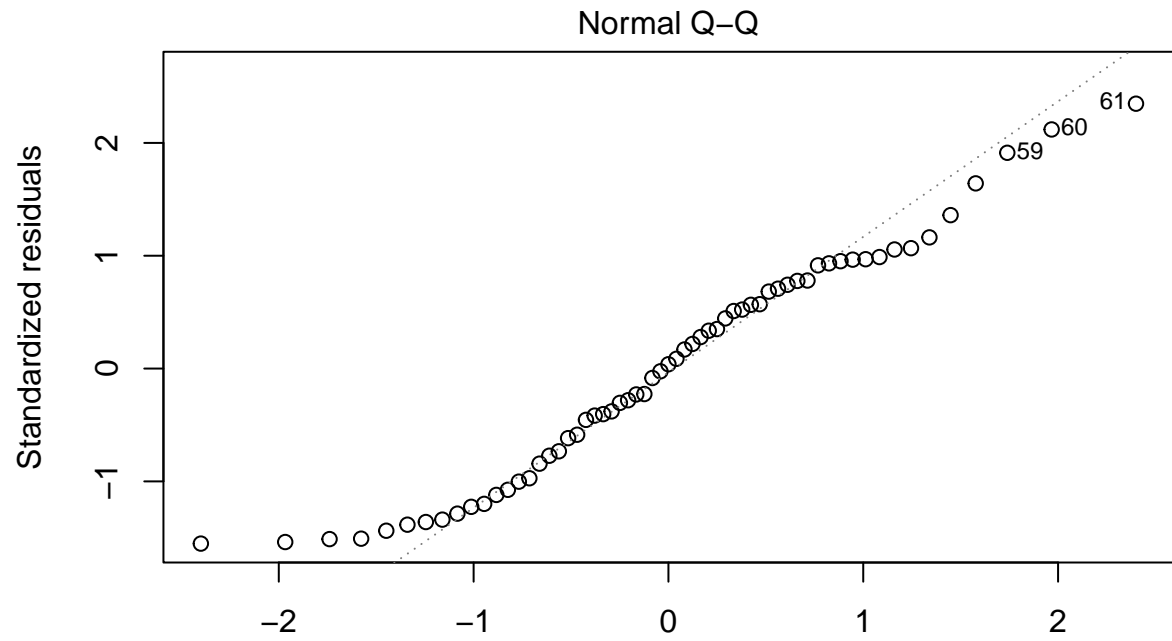
Table 2:

<i>Dependent variable:</i>	
‘Cumulative_number_for_14_days_of_COVID-19_cases_per_100000’	
index	4.107*** (0.145)
Constant	42.853*** (5.165)
Observations	61
R ²	0.932
Adjusted R ²	0.930
Residual Std. Error	19.922 (df = 59)
F Statistic	803.464*** (df = 1; 59)

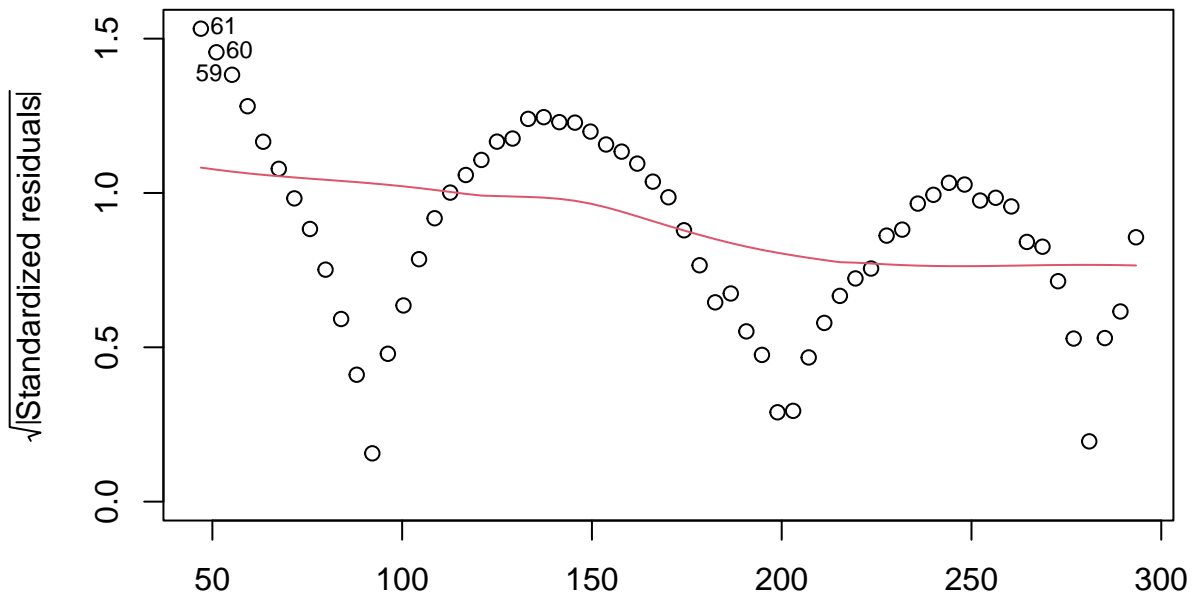
Note:

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

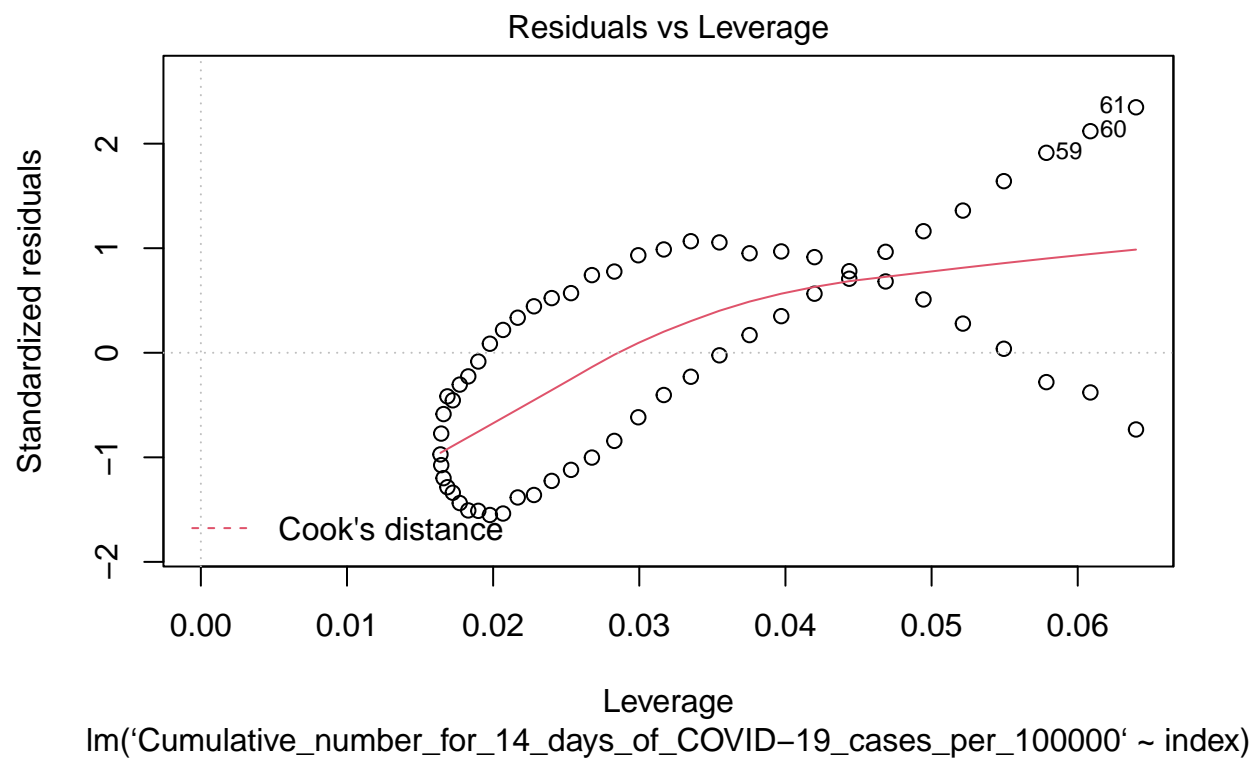




Im('Cumulative_number_for_14_days_of_COVID-19_cases_per_100000' ~ index)
Scale-Location

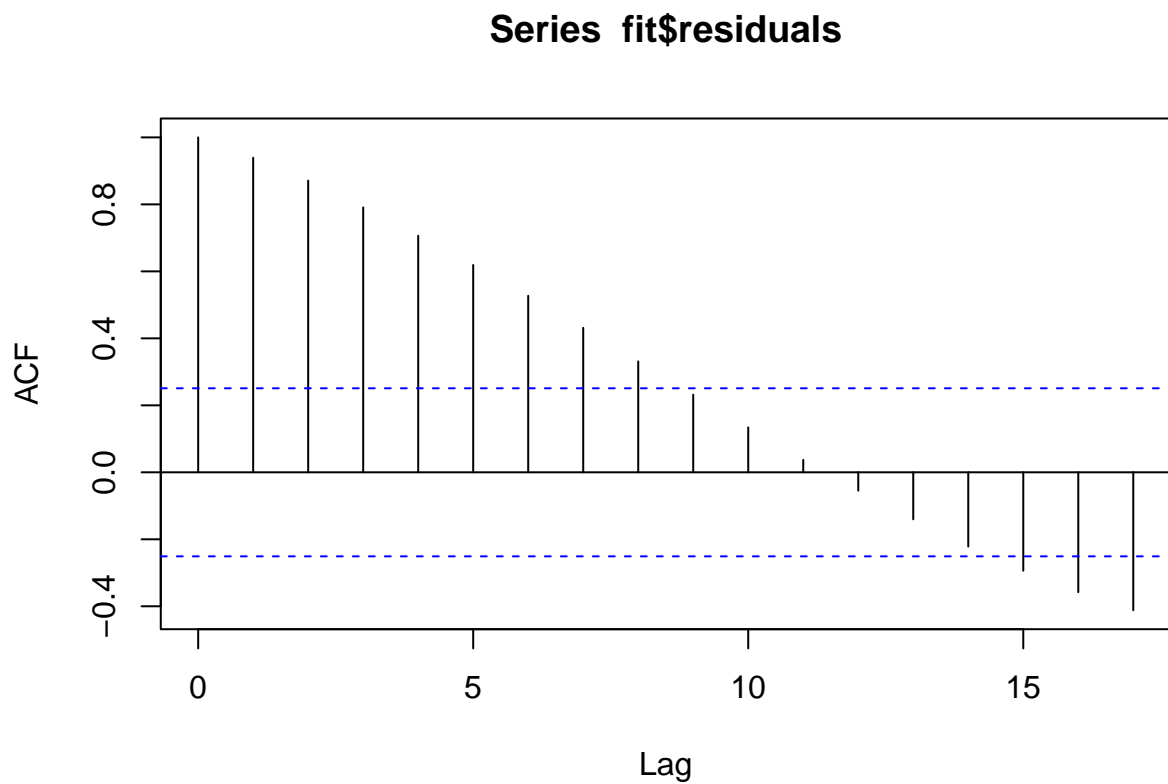


Im('Cumulative_number_for_14_days_of_COVID-19_cases_per_100000' ~ index)



Part C

```
acf(fit$residuals)
```



Problem 5

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

