

COVID Check4

Chen, Xiangnan & Cui, Augustine

2020/11/5

Introduction

COVID-19 is an infectious disease caused by the most recently discovered coronavirus. The disease spreads primarily from person to person through small droplets from the nose or mouth when people with COVID-19 cough, sneeze or speak [1]. The current coronavirus disease COVID-19 pandemic is hitting the globe unprecedentedly. Lives have been taken, and economic activities have been stagnated. Thus, it is crucial for people to understand better this global pandemic's current situation and future development to be better prepared to solve this global crisis. In this report, we are using COVID-19 data gathered in the United States to check if there is a linear relationship between date and change of number of cases. We implement linear regression and construct a null hypothesis to check if we can claim that there is a linear relationship between change of number of cases and the CDC report date.

Background

Among all the studies that focus on COVID-19, we find three sources interesting. The first is CDC's weekly report that gives an overview of the COVID-19 related hospitalizations and deaths. The data from Utah analyzes the relationship between COVID-19 cases, hospitalization, and testing with the deprivation level in different areas within Utah. The third research focuses on the Age distribution of the COVID-19 pandemic.

The CDC report focuses on the relationship between COVID-19 related hospitalizations and mortality. It also includes categorical variables like age and race. The overall cumulative COVID-19 hospitalization rate is 174.8 per 100,000, with the highest rates in people aged 65 years and older (472.3 per 100,000) and 50–64 years (261.5 per 100,000).[2] The second report analyzes COVID-19 Data from the Utah Department of Health. It examines the relationship between cases and hospitalization, the relationship between testing and the level of deprivation. The report shows that the infection number of Utah's high-deprivation areas is three times higher than the lower-deprivation areas, so does hospitalization and testing rates[3]. The third research focuses on the relationship between COVID-19 and ages. The source researches the relationship of incidence between each age group. It also calculates the weekly median age of people with COVID-19. The research shows that the distribution centers at age group from 20-39 years, which means that younger adults are more likely to contribute to community transmission of COVID-19[4].

Our approach to COVID-19 is to analyze the relationship between the change of number of cases and the CDC report date. Different from the previous studies, we are using the change of number of cases. We believe that using the change of number can offer better interpretability when constructing a model with date (Time series data).

Data

```
cov <- read.csv("COVID-19_Case_Surveillance_Public_Use_Data.csv")
```

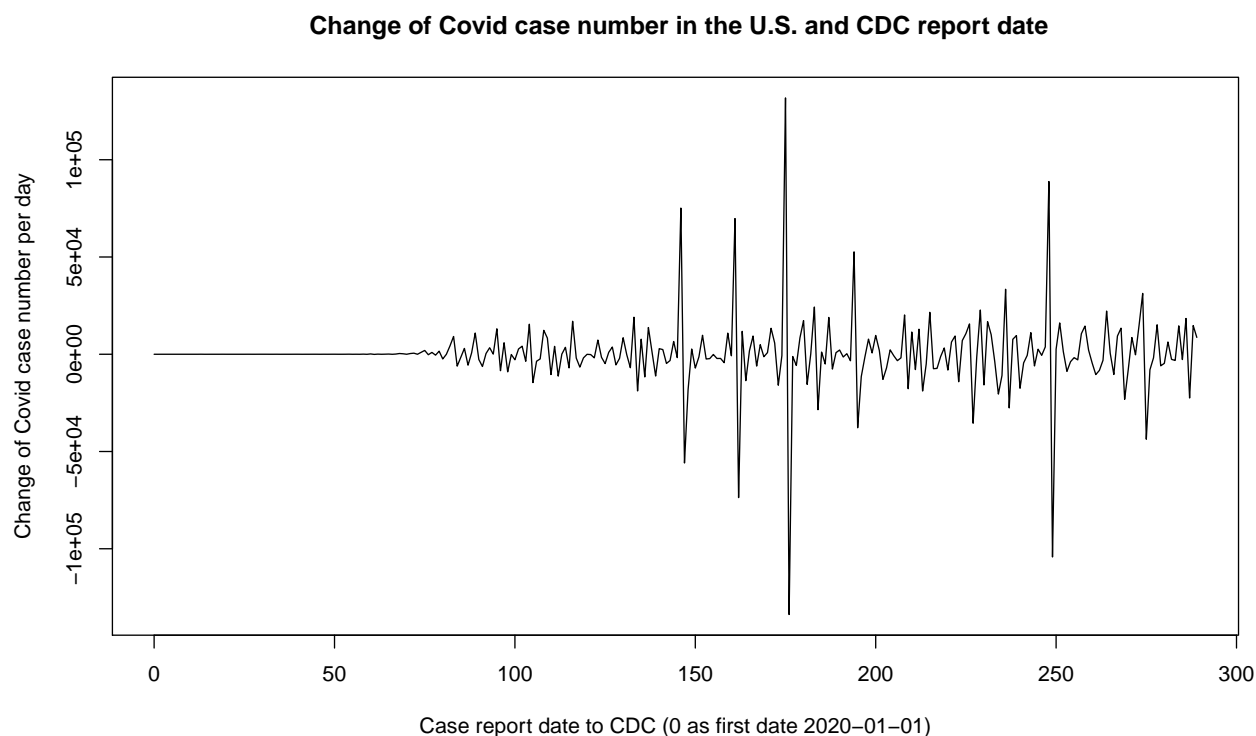
There are three quantitative variables 'cdc_report_dt', 'pos_spec_dt' and 'onset_dt'. There are eight categorical variables 'current_status', 'sex', 'age_group', 'Race.and.ethnicity.combined', 'hosp_yn', 'icu_yn', 'death_yn' and 'medcond_yn'. Description of each variable is in the Appendix.

There are 5760066 observations and each observation contains 11 variables.

```
reportDate <- unique(cov$cdc_report_dt)
case_date <- count(cov, cov$cdc_report_dt)
case_date$date <- as.Date(case_date$cov$cdc_report_dt)
case_date$date <- case_date$date - as.Date("2020-01-01")
case_date$cum_case <- cumsum(case_date$n)

plot(case_date$date, c(0, diff(case_date$n)), xlab = "Case report date to CDC (0 as first date 2020-01-01)",
      ylab = "Change of Covid case number per day", main = "Change of Covid case number in the U.S. and CDC report date",
      type = "l")
```

Data Preprocessing



There is a positive nonlinear relationship between the CDC report date and change of Covid case number. We can see that among the first 80 dates—there are very few cases reported to CDC. After the 100th day, the curves starts to fade out and becomes like a fan shape. There is not an obvious trend showing an increase in slope.

We are not using date of first positive specimen collection ('pos_spec_dt') and date of symptom onset ('onset_dt'), because we specifically choose to use the CDC report date as our predictor variable. We are currently not using any categorical variables within our model because we want first to see the relationship between the case number and the CDC report date.

We fitted a linear model, and the summary:

```
changeMod <- lm(c(0, diff(case_date$n)) ~ case_date$date)
## prettify(summary(changeMod)) %>% kable()
summary(changeMod)
```

```
##
## Call:
## lm(formula = c(0, diff(case_date$n)) ~ case_date$date)
##
## Residuals:
```

##	Min	1Q	Median	3Q	Max
##	-134063	-3444	-57	2837	131535

```
##
## Coefficients:
```

##		Estimate	Std. Error	t value	Pr(> t)
##	(Intercept)	7.875	2270.511	0.003	0.997
##	case_date\$date	1.339	13.502	0.099	0.921

```
##
## Residual standard error: 18880 on 284 degrees of freedom
## Multiple R-squared: 3.464e-05, Adjusted R-squared: -0.003486
## F-statistic: 0.009839 on 1 and 284 DF, p-value: 0.9211
```

Due to Appendix II, we implement wild bootstrap to conduct a null hypothesis test to see if there is a linear relationship between change of number of cases and the CDC report date. We set the significance level to $\alpha = 0.05$.

$\alpha = 0.05$ $H_0: B_1 = 0$ $H_1: B_1 \neq 0$

```
wildChangeMod <- wild.boot(c(0, diff(case_date$n)) ~ case_date$date, B = 10000,
  seed = 8888)
quantile(wildChangeMod$bootEstParam[, 2], probs = c(0.025, 0.975))
```

```
##      2.5%      97.5%
## -19.20775  21.84842
```



95% confidence interval for estimate of the population slope is (-19.20775, 21.84842). Since 0 is within the confidence interval, it is a plausible value. We cannot reject null hypothesis and do not have evidence to support the claim that there is a linear relationship between change of number of cases and the CDC report date.

Besides, we also want to check if the population mean of changes of number of cases differs from 0. We set the significance level to $\alpha = 0.05$.

Conclutions

Due to the null hypothesis tests we conduct, we do not have evidence to support the claim that there is a linear relationship between change of number of cases and the CDC report date. It means that there is no evidence that the number of cases accelerates everyday. This result offers us a peace of mind that there is no evidence that the speed of infections accelerates. We hope this research can also help people that are frustrated about this global pandemic and give them a optimistic opinion about the future development. However, since we are using a time series data and we do not completely offset the time dependence variable. The null hypothesis may be conducted when errors are not independent. For further study, a specific time series analysis method should be implemented to offset the dependence and conduct the null hypothesis.

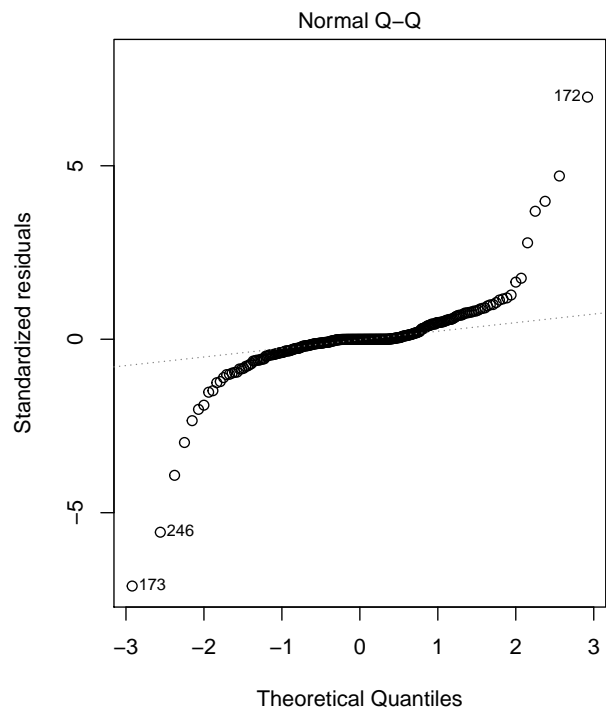
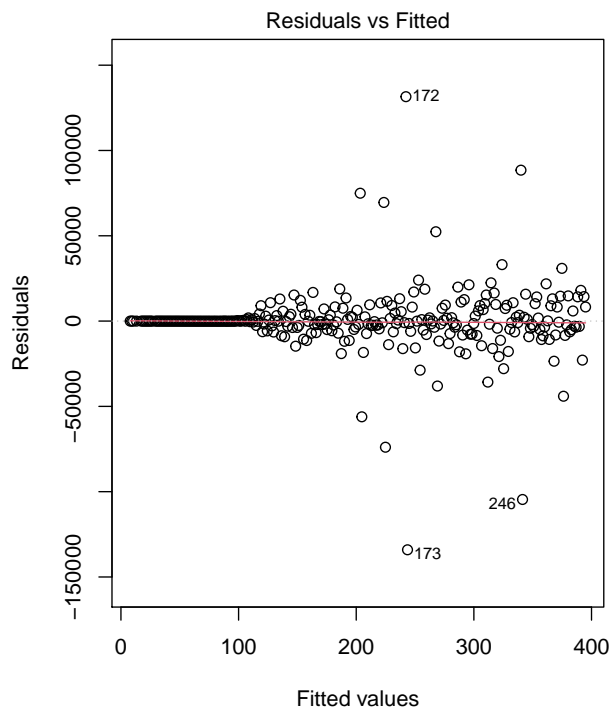
References

- [1] Q&A on coronaviruses (COVID-19) <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses>
- [2] “COVIDView: A Weekly Surveillance Summary of U.S. COVID-19 Activity,” Centers for Disease Control and Prevention. [Online]. Available: <https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/index.html>. [Accessed: 28-Sep-2020].
- [3] Lewis NM, Friedrichs M, Wagstaff S, et al. Disparities in COVID-19 Incidence, Hospitalizations, and Testing, by Area-Level Deprivation — Utah, March 3–July 9, 2020. MMWR Morb Mortal Wkly Rep 2020;69:1369–1373. DOI: <http://dx.doi.org/10.15585/mmwr.mm6938a4>
- [4] Boehmer TK, DeVies J, Caruso E, et al. Changing Age Distribution of the COVID-19 Pandemic — United States, May–August 2020. MMWR Morb Mortal Wkly Rep. ePub: 23 September 2020. DOI: <http://dx.doi.org/10.15585/mmwr.mm6939e1>
- [5] Centers for Disease Control and Prevention. COVID-19 Case Surveillance Public Use Data, September 30, 2020 <https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf>

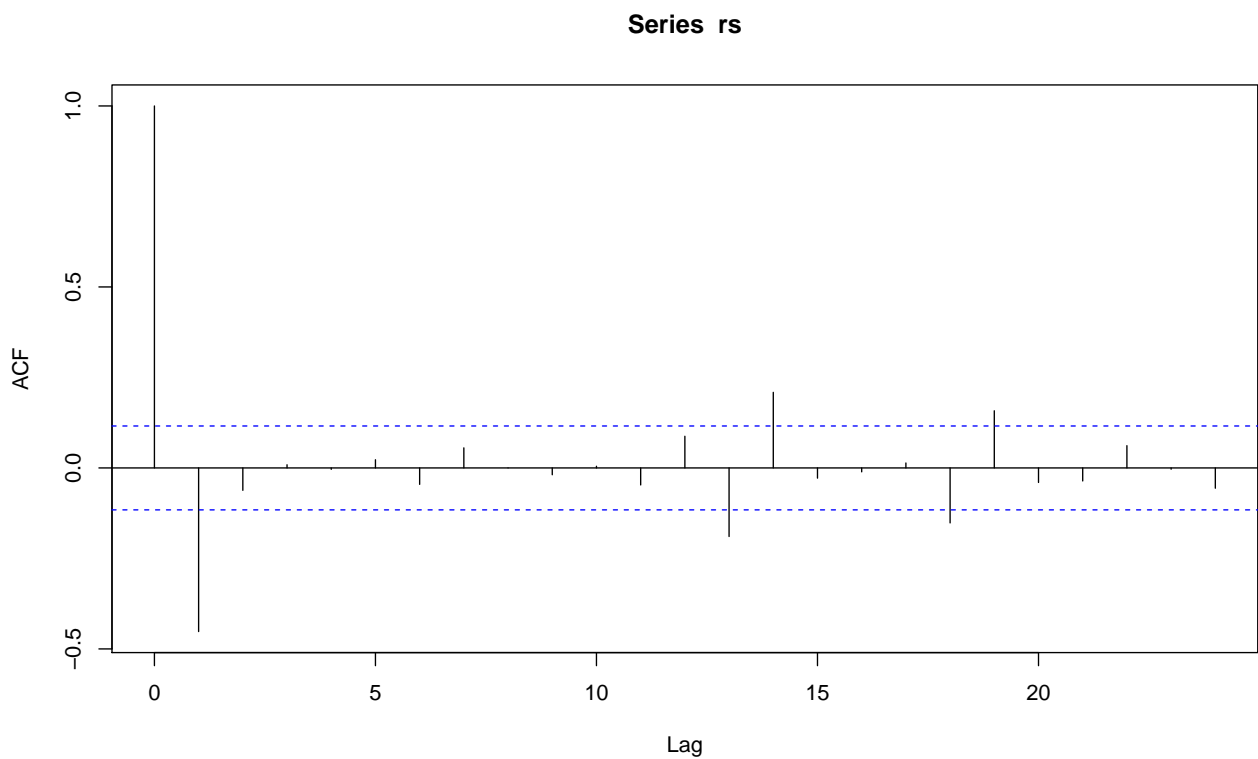
Appendix I

Column.names	Description	Type
cdc_report_dt	Initial case report date to CDC	Quantative: Date & Time
pos_spec_dt	Date of first positive specimen collection	Quantative: Date & Time
onset_dt	Symptom onset date, if symptomatic	Quantative: Date & Time
current_status	Case Status: Laboratory-confirmed case; Probable case	Categorical: Text
sex	Sex: Male; Female	Categorical: Text
age_group	Age Group: 0 - 9 Years; 10 - 19 Years; 20 - 39 Years; 40 - 49 Years; 50 - 59 Years; 60 - 69 Years; 70 - 79 Years; 80 + Years	Categorical: Text
Race and ethnicity (combined)	Race and ethnicity (combined): Hispanic/Latino; American Indian / Alaska Native, Non-Hispanic; Asian, Non-Hispanic; Black, Non-Hispanic; Native Hawaiian / Other Pacific Islander, Non-Hispanic; White, Non-Hispanic; Multiple/Other, Non-Hispanic	Categorical: Text
hosp_yn	Hospitalization status: Yes/No	Categorical: Text
icu_yn	ICU admission status: Yes/No	Categorical: Text
death_yn	Death status: Yes/No	Categorical: Text
medcond_yn	Presence of underlying comorbidity or disease: Yes/No	Categorical: Text

Appendix II

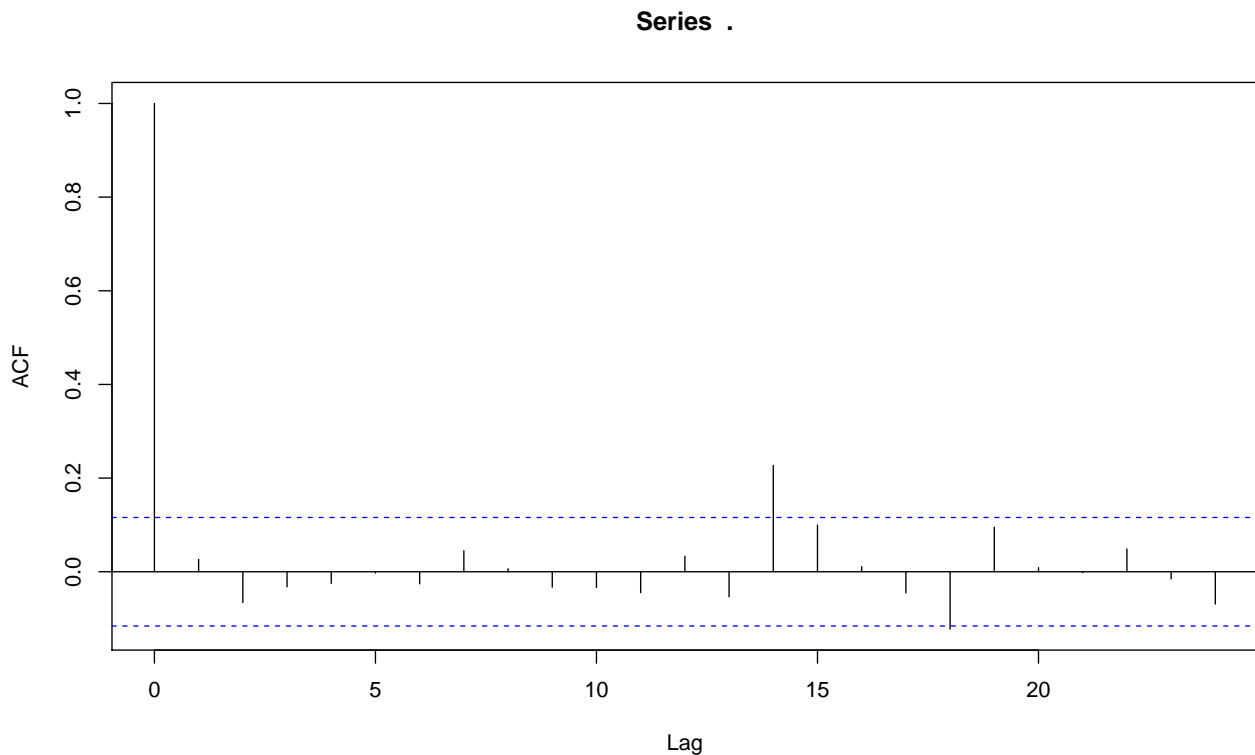


```
rs <- residuals(changeMod)
acf(rs)
```



```
mod_new <- arima(x = c(0, diff(case_date$n)), order = c(0, 0, 1))
print(mod_new)

##
## Call:
## arima(x = c(0, diff(case_date$n)), order = c(0, 0, 1))
##
## Coefficients:
##          ma1  intercept
##       -0.9516  145.8037
## s.e.    0.0195   42.6458
##
## sigma^2 estimated as 191588854:  log likelihood = -3134.13,  aic = 6274.26
residuals(mod_new) %>% acf()
```



```
coeftest(mod_new)

##
## z test of coefficients:
##
##          Estimate Std. Error  z value  Pr(>|z|)
## ma1       -0.951648   0.019471 -48.8741 < 2.2e-16 ***
## intercept 145.803650  42.645821   3.4189 0.0006286 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

auto.arima(c(0, diff(case_date$n)))

## Series: c(0, diff(case_date$n))
## ARIMA(0,0,1) with non-zero mean
```

```
##
## Coefficients:
##          ma1      mean
##      -0.9516  145.8037
## s.e.    0.0195   42.6458
##
## sigma^2 estimated as 192938071:  log likelihood=-3134.13
## AIC=6274.26   AICc=6274.35   BIC=6285.23
```

Errors are independent: Since the change of number of cases is a time series data, there has a time dependence in the errors. We include the date as our predictor variable to offset the time dependence. The assumption that errors are independent is questionable, but we hope that including date as predictor variable make errors independent. According to the ACF diagram, we can observe a significant lag 1 correlation. So the data in the present day has a negative correlation with the data in the previous day. So we cannot claim that the errors are independent.

Errors have mean 0: There does not appear to be a pattern away from the 0 horizontal line. It is reasonable to assume that errors have mean 0.

Errors have constant variance: There does appear a fan shape in the residual vs Fitted plot. It is not reasonable to assume that errors have constant variance.

Errors are from normal distribution: There does appear to be a clear pattern away from the diagonal in the QQ plot. It is not reasonable to assume that errors come from a normal distribution.

Since we can only assume that errors are independent and have mean 0, we implement wild bootstrap to further analysis.