Midterm Project

Yuxin Zeng

2020/11/28

Abstract

COVID-19 is an infectious disease caused by SARS-CoV-2, which broke out on a large scale and spread rapidly in countries around the world. I collected the daily number of confirmed cases of COVID-19 in Chicago since March this year, hoping to study whether the number of cases is related to temperature and weekends/holidays. Due to the overdispersion of the data, I use a negative binomial regression model. After excluding insignificant variables, it was found that temperature and holidays have very little influence on the number of confirmed cases. At the same time, the model validation results prove that the model cannot fit the data well. Since the result of the hierarchical model is not ideal neither, I put it in the appendix. I also fit an ARIMA model, which is better than the negative binomial model.

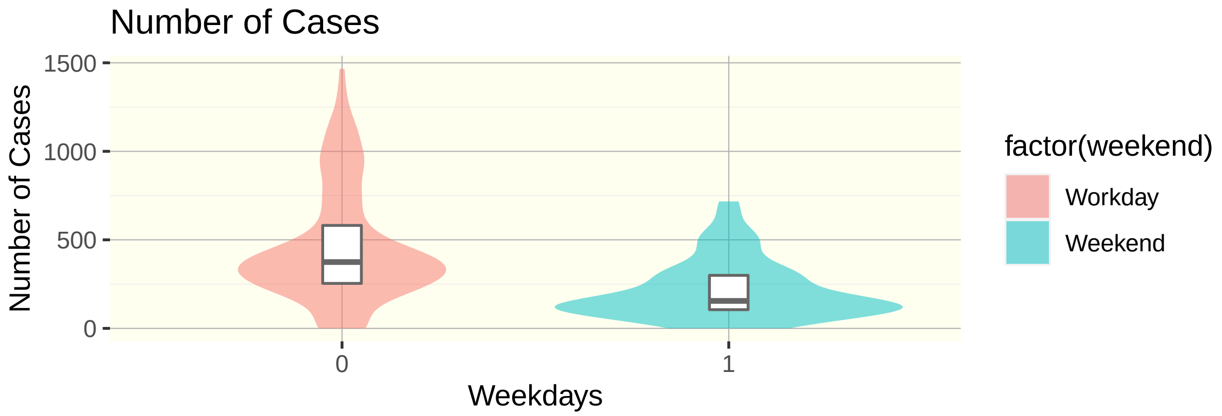
Data

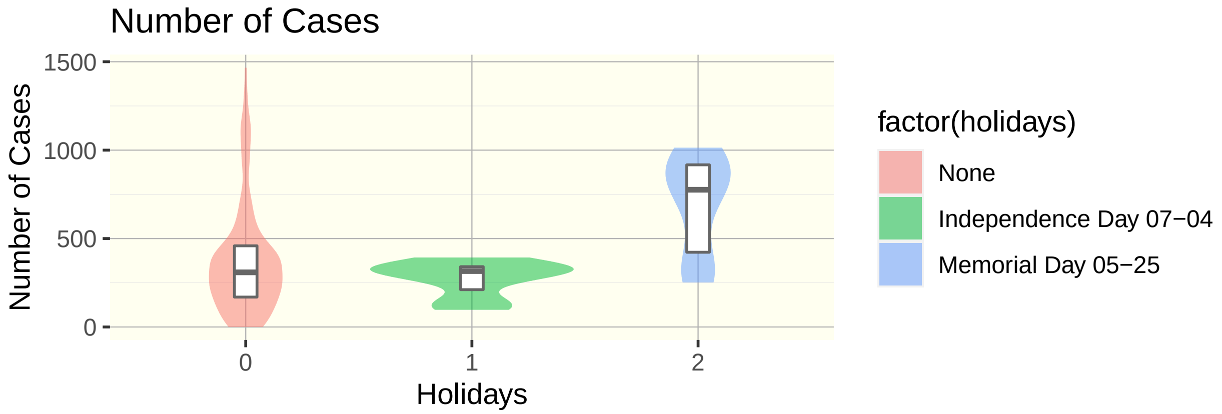
First read the COVID-19 cases data. Only Chicago residents are included based on the home ZIP Code, as provided by the medical provider, or the address, as provided by the Cook County Medical Examiner. Confirmed cases are counted on the date the test specimen was collected.

Since high temperature may inhibit some virus activity. I am curious whether temperature affects the number of confirmed cases per day. I downloaded and organized the temperature data from NDBC website. The air temperature (ATMP) is measured every hour, and I take the temperature at noon as the temperature of that day. The unit is Fahrenheit (F).

After merging the COVID-19 data set and the Temperature data set, I tried to add weekend and holidays variable to indicate the gathering of people. The "weekend" variable is 0 for weekdays, 1 for weekends. Federal holidays that between March and August are Memorial Day (2020-05-25) and Independence Day (2020-07-04), and both have a 3-day off (includes weekend). Taking the asymptomatic of COVID-19 into account, assume there is a two-week incubation period for each holiday. That is, the "holidays" variable is 2 for dates within two weeks after Memorial Day, 1 for dates within two weeks after Independence Day, 0 for otherwise. (Since Independence Day is in summer vacation, for student groups, there is no big difference on that day compared with other days.)

EDA





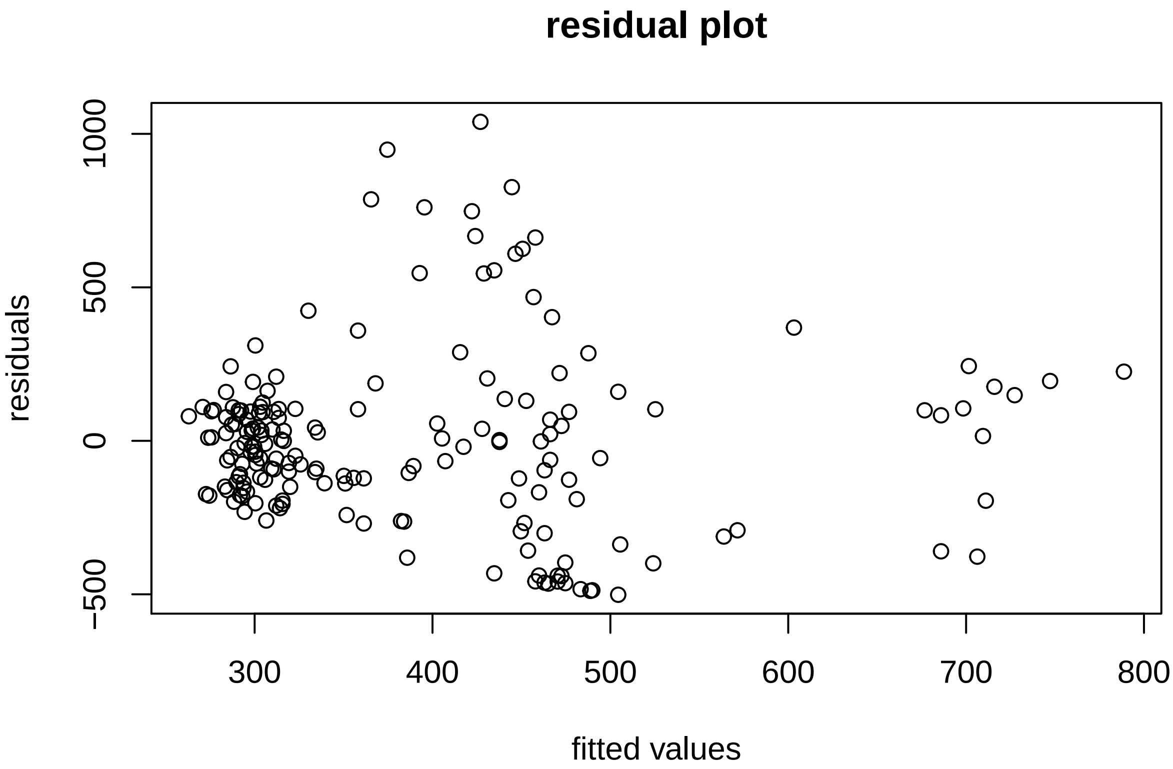
Notice the violin plot comparing the weekdays and weekends, the decrease in cases on weekends may probably due to the fact that hospital's inspectors do not work on weekends. So I made an adjustment on "weekend" variable: "weekend_delay" is 1 for Monday, and 0 for otherwise.

Negative Binomial Model

The dependent variable (number of confirmed cases each day) is a count variable, and the independent variables contain a continuous variable (Air temperature at noon each day) and two categorical variables (weekends and holidays).
The result of over-dispersion test showed that the data was over-dispersion, so I chose negative binomial model rather than Poisson model. The proportion of zero value in the dependent variable is not large, so zero-inflated negative binomial regression is not needed.

I used glm.nb function from MASS package to fit the negative binomial model. From the summary, air temperature and Memorial Day pass the significance test, indicating that they do affect the number of COVID-19 cases, though the coefficient is not very big. Exclude insignificant variables and fit a new model. The formula is Cases_Total~ATMP+holiday.

```
## (Intercept)        ATMP     holiday
## 496.5227890   0.9777186   1.7780014
```



```
## Waiting for profiling to be done...
```

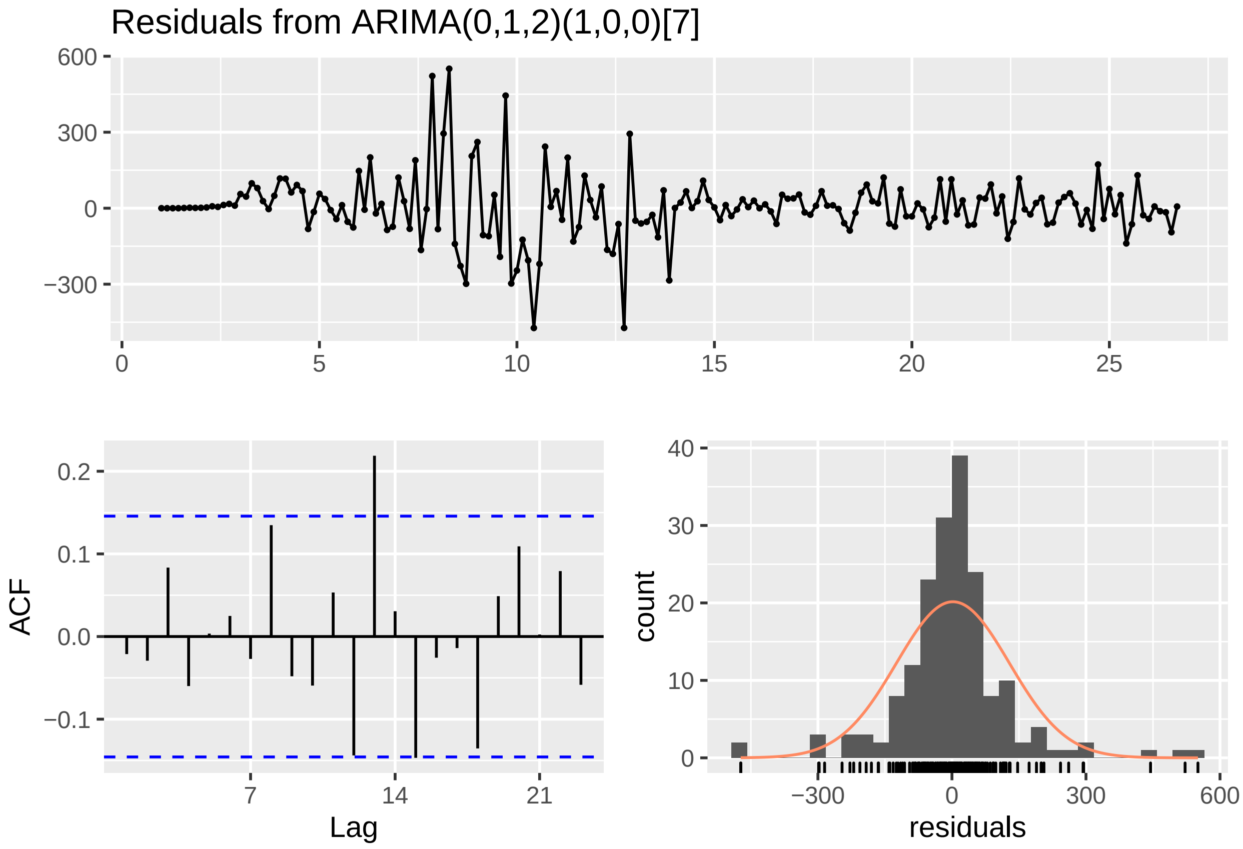
```
##                   2.5 %       97.5 %
## (Intercept)  5.94507739  6.487343886
## ATMP        -0.03855464 -0.006688426
## holiday      0.12467471  1.087818608
```

```
## $results
## [1] "Goodness-of-fit test for Poisson assumption"
##
## $chisq
## [1] 209.6142
##
## $df
## [1] 178
##
## $p.value
## [1] 0.05257753
```

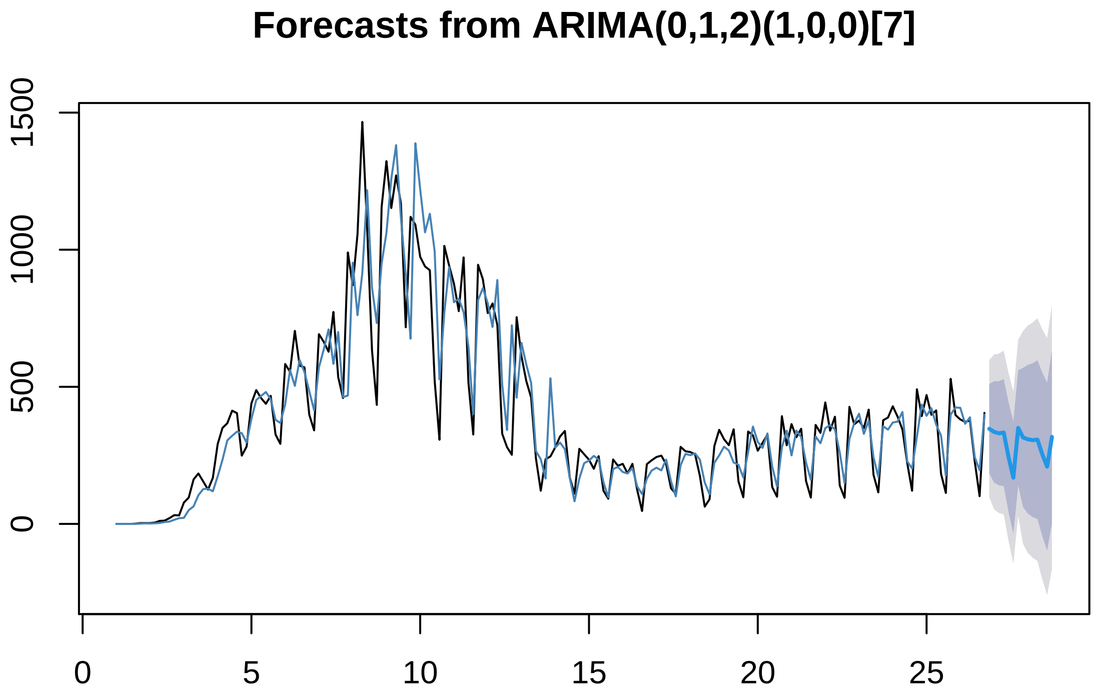
The summary shows as the temperature increases, the number of cases decreases slightly. And the number of cases has increased after holidays. The residuals are mostly clustered on the left in the residual plot. The 97.5% confidence interval of the coefficients is obtained via confint function. After checking the goodness of fit (poisgof), the p value is around 0.05, so it's not a good fit.

Then I used glmer function to fit a multilevel negative binomial model. I group the data by month, and the formula is Cases_Total~ATMP+holiday+(1|(month(Date))). In this sense, the influence of air temperature on the dependent variable goes down a lot because the temperature increases month by month and grouping data by months will weaken the effect of temperature.

ARIMA



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,2)(1,0,0)[7]
## Q* = 21.299, df = 11, p-value = 0.03041
##
## Model df: 3.   Total lags used: 14
```



I fitted a ARIMA(0,1,2)(1,0,0)[7] model. The result of Ljung-Box test shows that the sequence is correlated, not a white noise sequence. The AIC of ARIMA model is smaller than that of negative binomial model.

Discussion

When counting the number of COVID-19 cases each day, ARIMA model is better than negative binomial regression model, but it is better to use the infectious disease dynamics models like SIR, SEIR, etc.,which can reflect the epidemic law from the aspect of disease transmission mechanism. The poor fit of the negative binomial regression model is reasonable. Traditional statistical regression models are not suitable for infectious diseases. Compared with the SIR model, they are static, lack geospatiality, and cannot perfectly measure the impact of human activities on the number of cases. The impact of air temperature and holidays on the number of COVID-19 cases is extremely slight, especially at the beginning of the outbreak, the number is more determined by human activities. If I could collect data on whether people in Chicago wear masks and how often they go out from March to August as independent variables, the model might fit better.

Biography

[1] healthdata.gov

[2] National Data Buoy Center

[3]Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

[4]Andrew Gelman and Yu-Sung Su (2020). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.11-2. https://CRAN.R-project.org/package=arm

[5]Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2020). *forecast: Forecasting functions for time series and linear models*. R package version 8.13, <URL:https://pkg.robjhyndman.com/forecast/>.

Appendix

```
#Data Preparation
dt=read.csv("COVID-19_Daily_Cases__Deaths__and_Hospitalizations.csv")
dt$Date=as.Date(dt$Date,"%m/%d/%Y")
dt=dt[month(dt$Date)<=08,]

Mar=data.frame(read.table(file="mar.txt"))
Apr=data.frame(read.table(file="apr.txt"))
May=data.frame(read.table(file="may.txt"))
Jun=data.frame(read.table(file="jun.txt"))
Jul=data.frame(read.table(file="jul.txt"))
Aug=data.frame(read.table(file="aug.txt"))
temp=data.frame(rbind(Mar,Apr,May,Jun,Jul,Aug))
colnames(temp)=c("YY","MM","DD","hh","mm","WDIR","WSPD","GST","WVHT","DPD","APD","MWD","PRES","ATMP","W
temp=temp[temp$hh=="12"&temp$mm=="0",]
temp$Date=make_datetime(temp$YY,temp$MM,temp$DD)
temp=temp[,c("ATMP","Date")]

dt=left_join(temp,dt,by="Date")
which(is.na(dt))
```

```
## integer(0)
```

```
dt=dt[,c(1:5,27:31)]

#Add variable
Sys.setlocale("LC_TIME", "English")
```

```
## [1] "English_United States.1252"
```

```
dt$weekday=weekdays(as.Date(dt$Date))
dt$weekend=ifelse(dt$weekday=="Saturday" | dt$weekday=="Sunday",1,0)
dt$weekend_delay=ifelse(dt$weekday=="Monday",1,0)

M=difftime(as.Date("2020-05-25"),as.Date(dt$Date),units="days")
I=difftime(as.Date("2020-07-24"),as.Date(dt$Date),units="days")
dt$holidays=ifelse(0<=I & I<=14,1,ifelse(0<=M & M<=14,2,0))
```

```
#Negative Binomial Model
#Over Dispersion
qcc.overdispersion.test(dt$Cases...Total,type="poisson")
```

```
##
## Overdispersion test Obs.Var/Theor.Var Statistic p-value
##        poisson data          238.1927  42874.69       0
```

```
#Zero Inflation
table(dt$Cases...Total=="0")
```

```
##
## FALSE  TRUE
##   177     4
```

```r
#Model
fit1=glm.nb(formula=Cases...Total~ATMP+factor(weekend_delay)+factor(holidays),data=dt)
summary(fit1)
```

```
##
## Call:
## glm.nb(formula = Cases...Total ~ ATMP + factor(weekend_delay) +
##     factor(holidays), data = dt, init.theta = 1.247933766, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8867  -0.6326  -0.0252   0.3008   1.8136
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              6.17460    0.13536  45.618  < 2e-16 ***
## ATMP                    -0.02149    0.00800  -2.686  0.00723 **
## factor(weekend_delay)1  0.17450    0.18723   0.932  0.35134
## factor(holidays)1      -0.11437    0.25320  -0.452  0.65147
## factor(holidays)2       0.56377    0.24360   2.314  0.02065 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.2479) family taken to be 1)
##
##     Null deviance: 227.12  on 180  degrees of freedom
## Residual deviance: 209.57  on 176  degrees of freedom
## AIC: 2516
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.248
##           Std. Err.:  0.124
##
##  2 x log-likelihood:  -2504.050
```

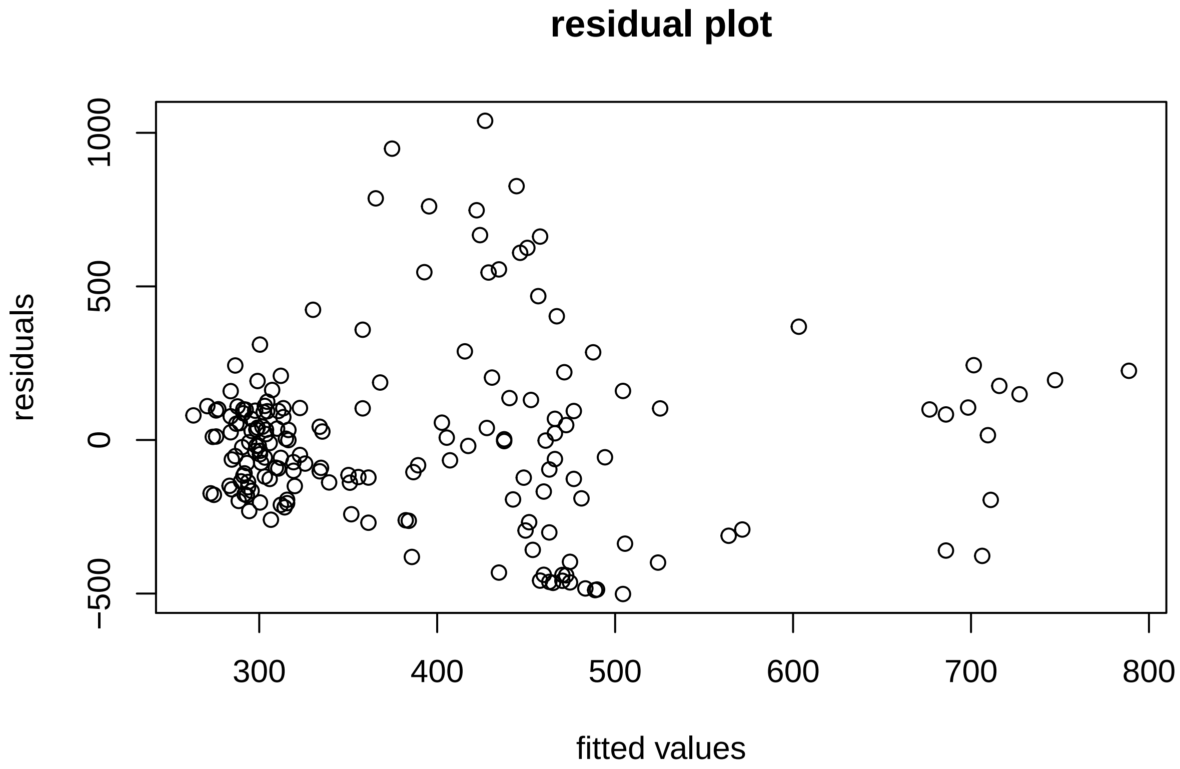
```r
dt$holiday=ifelse(0<=M & M<=14,1,0)
fit2=glm.nb(formula=Cases...Total~ATMP+holiday,data=dt)
summary(fit2)
```

```
##
## Call:
## glm.nb(formula = Cases...Total ~ ATMP + holiday, data = dt, init.theta = 1.241275342,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8524  -0.6360  -0.0363   0.3083   1.7733
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.207629   0.132519  46.843  < 2e-16 ***
## ATMP        -0.022533   0.007695  -2.928  0.00341 **
## holiday      0.575490   0.243544   2.363  0.01813 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.2413) family taken to be 1)
##
##     Null deviance: 225.98  on 180  degrees of freedom
## Residual deviance: 209.61  on 178  degrees of freedom
## AIC: 2513.1
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.241
##           Std. Err.:  0.123
##
##  2 x log-likelihood:  -2505.145
```

```
#Validation
exp(coef(fit2))
```

```
## (Intercept)        ATMP     holiday
## 496.5227890   0.9777186   1.7780014
```

```
plot(fitted(fit2),resid(fit2,type="response"),xlab="",ylab="")
title(main="residual plot",xlab="fitted values",ylab="residuals")
```



```
confint(fit2)
```

```
## Waiting for profiling to be done...

##                   2.5 %       97.5 %
## (Intercept)  5.94507739  6.487343886
## ATMP        -0.03855464 -0.006688426
## holiday      0.12467471  1.087818608
```

```
poisgof(fit2)
```

```
## $results
## [1] "Goodness-of-fit test for Poisson assumption"
##
## $chisq
## [1] 209.6142
##
## $df
## [1] 178
##
## $p.value
## [1] 0.05257753
```

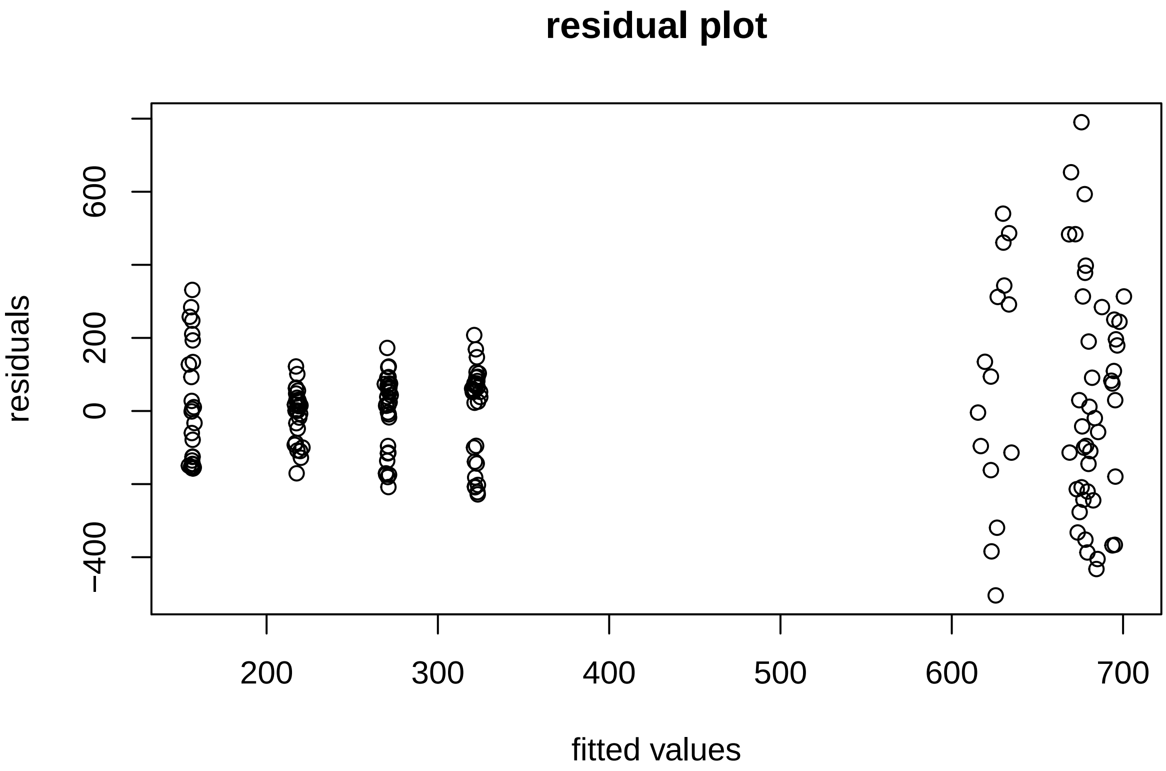
```
#Multilevel negative binomial
fit3=glmer(formula=Cases...Total~ATMP+holiday+(1|(month(dt$Date))),data=dt,family=negative.binomial(the
summary(fit3)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: Negative Binomial(1.241)  ( log )
## Formula: Cases...Total ~ ATMP + holiday + (1 | (month(dt$Date)))
##    Data: dt
##
##      AIC      BIC   logLik deviance df.resid
##   2476.4   2492.4  -1233.2   2466.4      176
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.10963 -0.47440  0.06682  0.30600  2.34721
##
## Random effects:
##  Groups           Name        Variance Std.Dev.
##  (month(dt$Date)) (Intercept) 0.3048   0.5521
## Number of obs: 181, groups:  (month(dt$Date)), 6
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.817577   0.334887  17.372   <2e-16 ***
## ATMP        -0.001556   0.016606  -0.094    0.925
## holiday      0.102756   0.316006   0.325    0.745
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr) ATMP
## ATMP    -0.708
## holiday -0.057 -0.024
```

```
anova(fit3)
```

```
## Analysis of Variance Table
##         npar   Sum Sq  Mean Sq F value
## ATMP       1 0.009846 0.009846  0.0098
## holiday    1 0.104706 0.104706  0.1047
```

```
plot(fitted(fit3),resid(fit3,type="response"),xlab="",ylab="")
title(main="residual plot",xlab="fitted values",ylab="residuals")
```

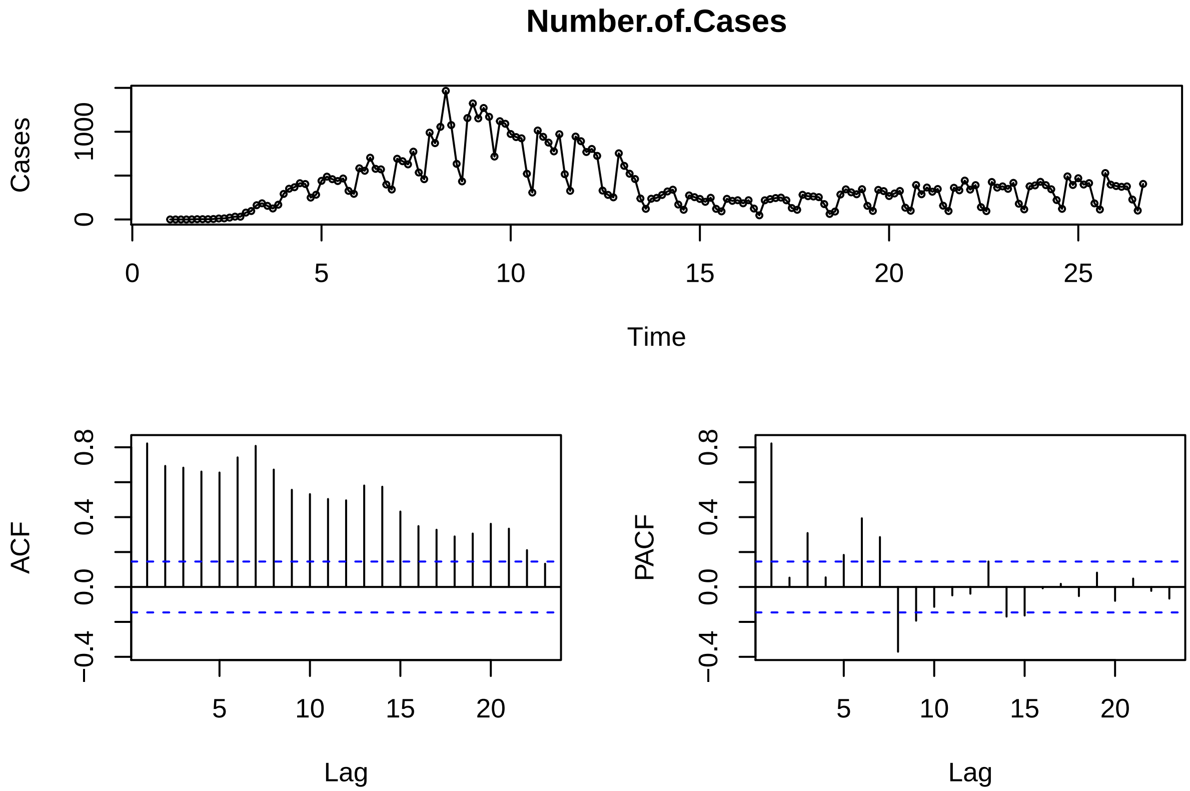


```
dispersion_glmer(fit3)
```

```
## [1] 0.9274866
```

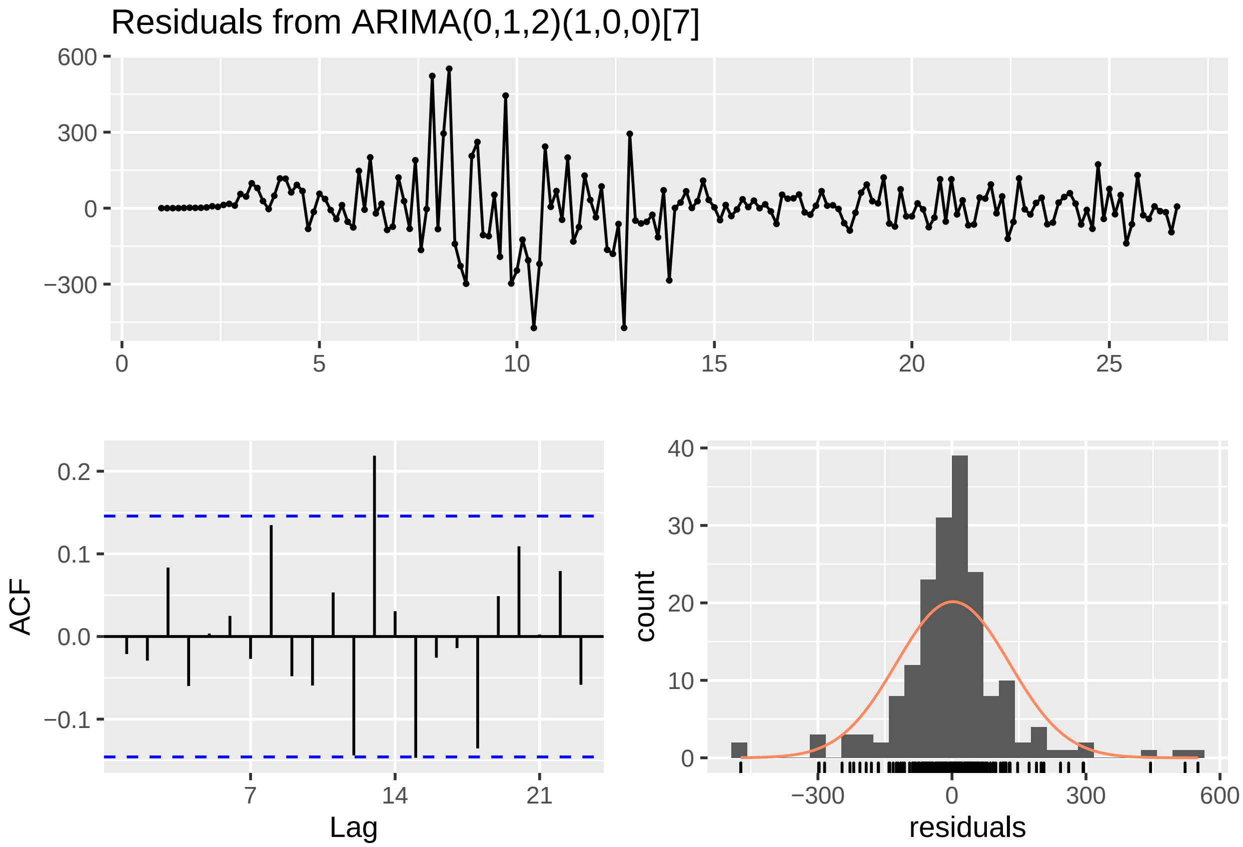
```
#ARIMA
Number.of.Cases=ts(dt$Cases...Total,frequency=7)

tsdisplay(Number.of.Cases,xlab="Time",ylab="Cases")
```



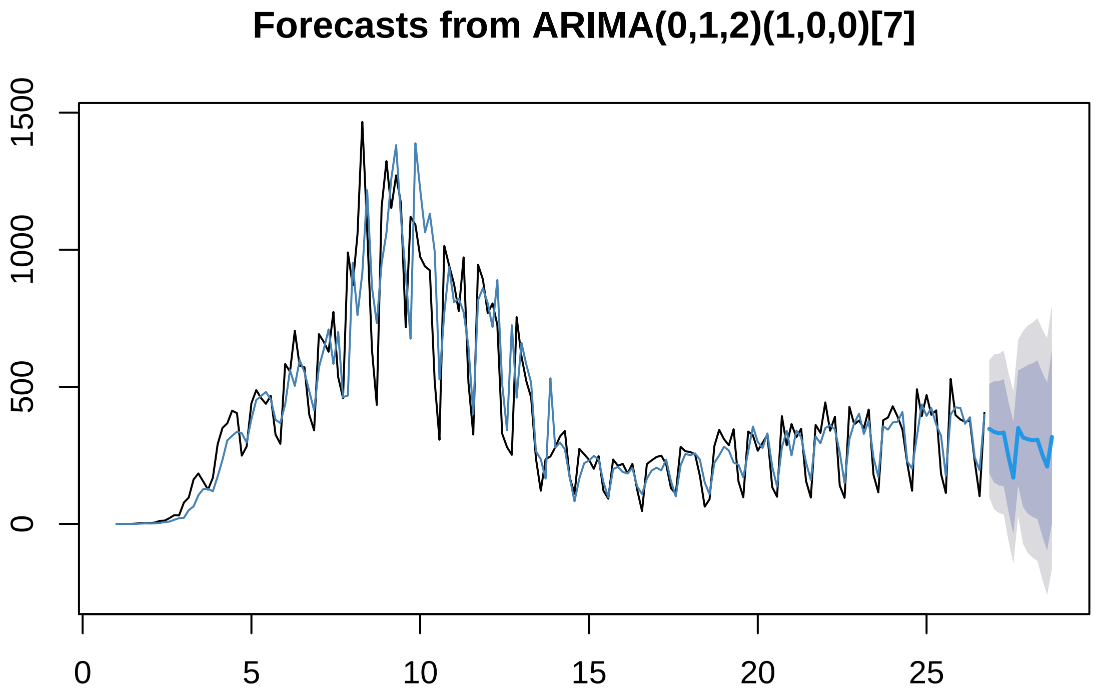
```
fit3=auto.arima(Number.of.Cases)

checkresiduals(fit4)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,2)(1,0,0)[7]
## Q* = 21.299, df = 11, p-value = 0.03041
##
## Model df: 3.   Total lags used: 14
```

```
fore=forecast(fit4)
plot(fore)
lines(fore$fitted,col="steelblue")
```


Forecasts from ARIMA(0,1,2)(1,0,0)[7]
1500
1000
500
0
0
5
10
15
20
25