State specific time series analysis

Eric Olle

April 23, 2020

This is the basic markdown document for the analysis of the timeseries data.

This is a rough draft and has not been checked for spelling or grammar!

This is a working document that is part of an onging R markdown.

The orginal r markdown will be posted to:

https://github.com/Eric43/SIP-order

This document is licensed under GPLv3 or equivelnet (ie CC-BY-SA)

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

Loading the data from file

Data use if from the nytimes github account.

```
us_counties <- read_csv("https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv
```

```
## Parsed with column specification:
## cols(
##   date = col_date(format = ""),
##   county = col_character(),
##   state = col_character(),
##   fips = col_character(),
##   cases = col_double(),
##   deaths = col_double()
## )
```

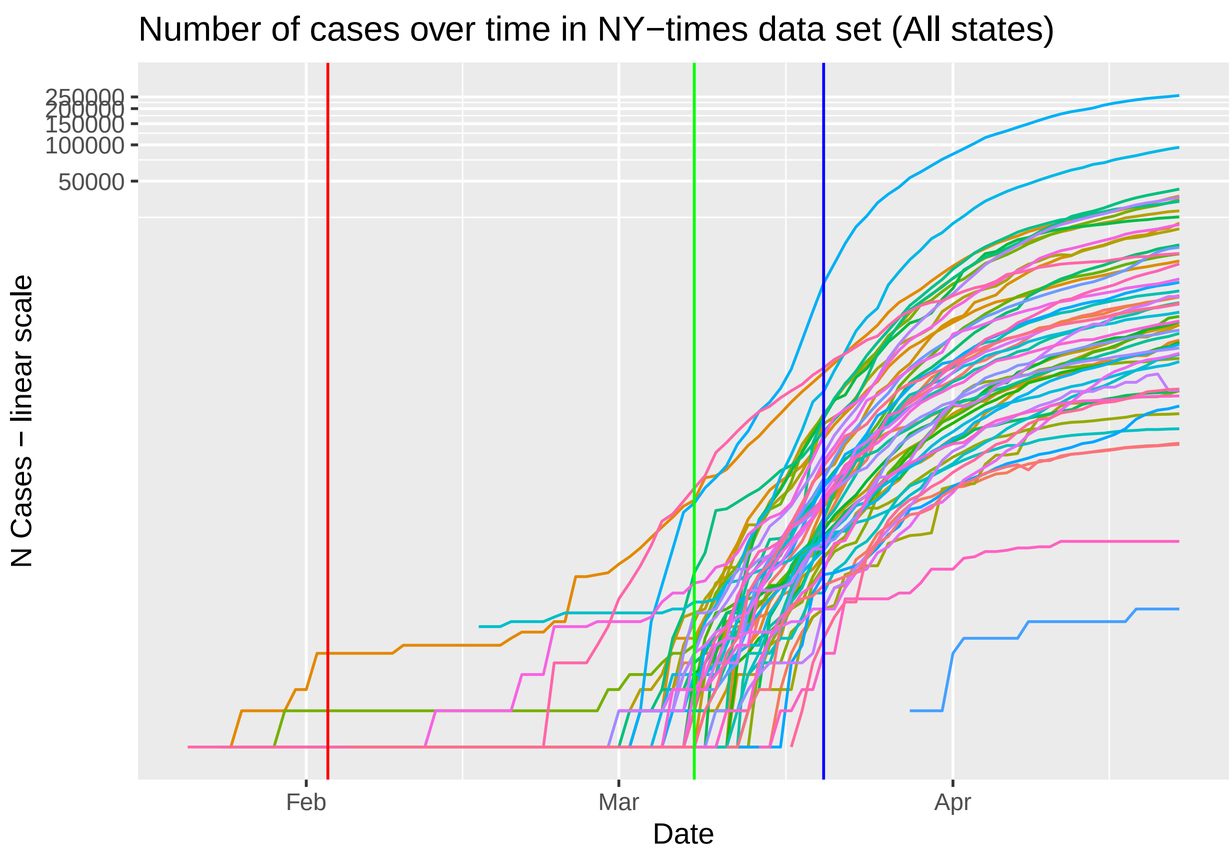
Setting the constants that apply to the individual state

```
# Which specific state?

state2select <- "Illinois"

# What is the state specific lockdown?
# (follow the correct date format)

lockdown_st <- as.Date("2020-03-20")

# Known lockdown data

## Peoples Republic of China
lockdownprc <- as.Date("2020-02-03")
## Italy
lockdownitl <- as.Date("2020-03-08")
## USA - New York
lockdownnyc <- as.Date("2020-03-20")
```

Part 1. Time series analysis

In this section a basic time series plot of the full US along with the selected state will be done.

Plotting the full state data set.



Selecting the state specific data from the NY Times dataset

From the unique state names call (above) set the state name by copy/paste or typing in with the quotation marks.

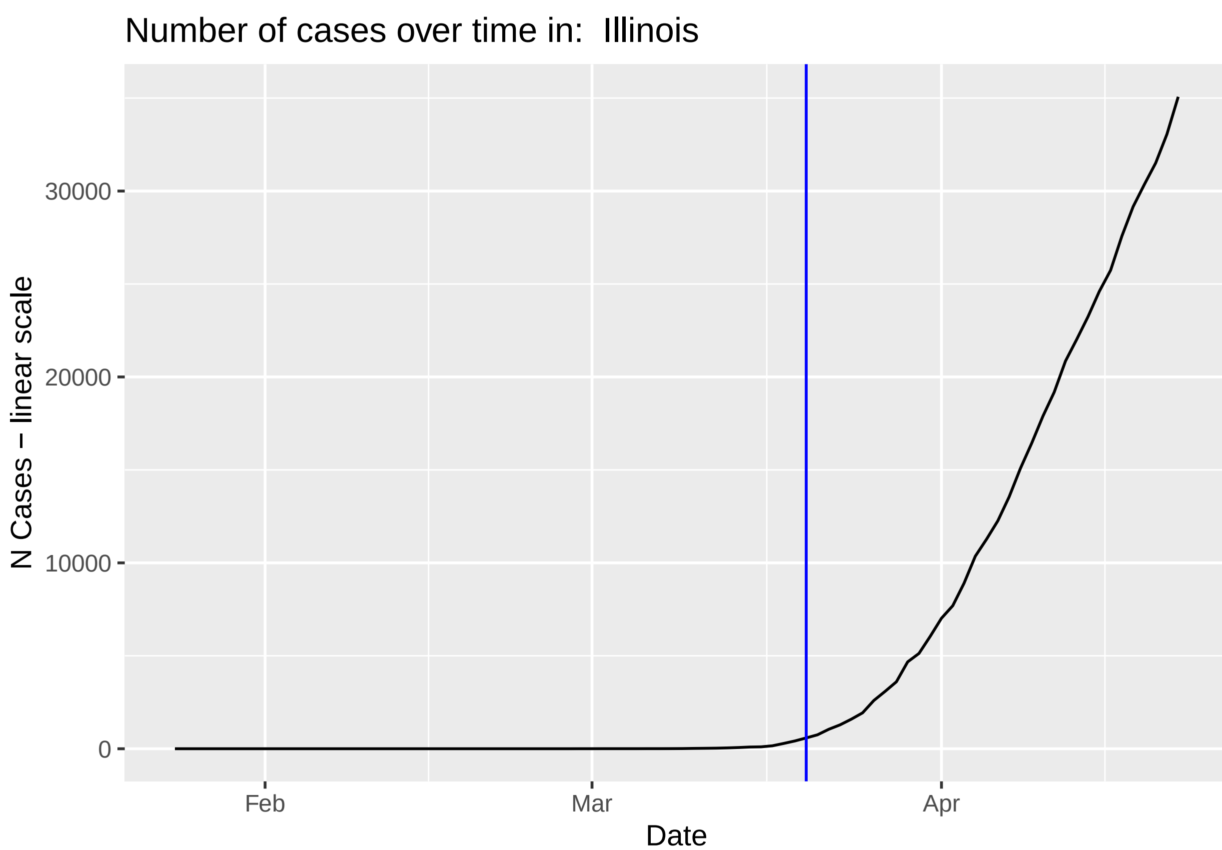
Selecting the state data from state2select var

The st data set was selected in case future geo-spatial on the rate of cases by county is needed.

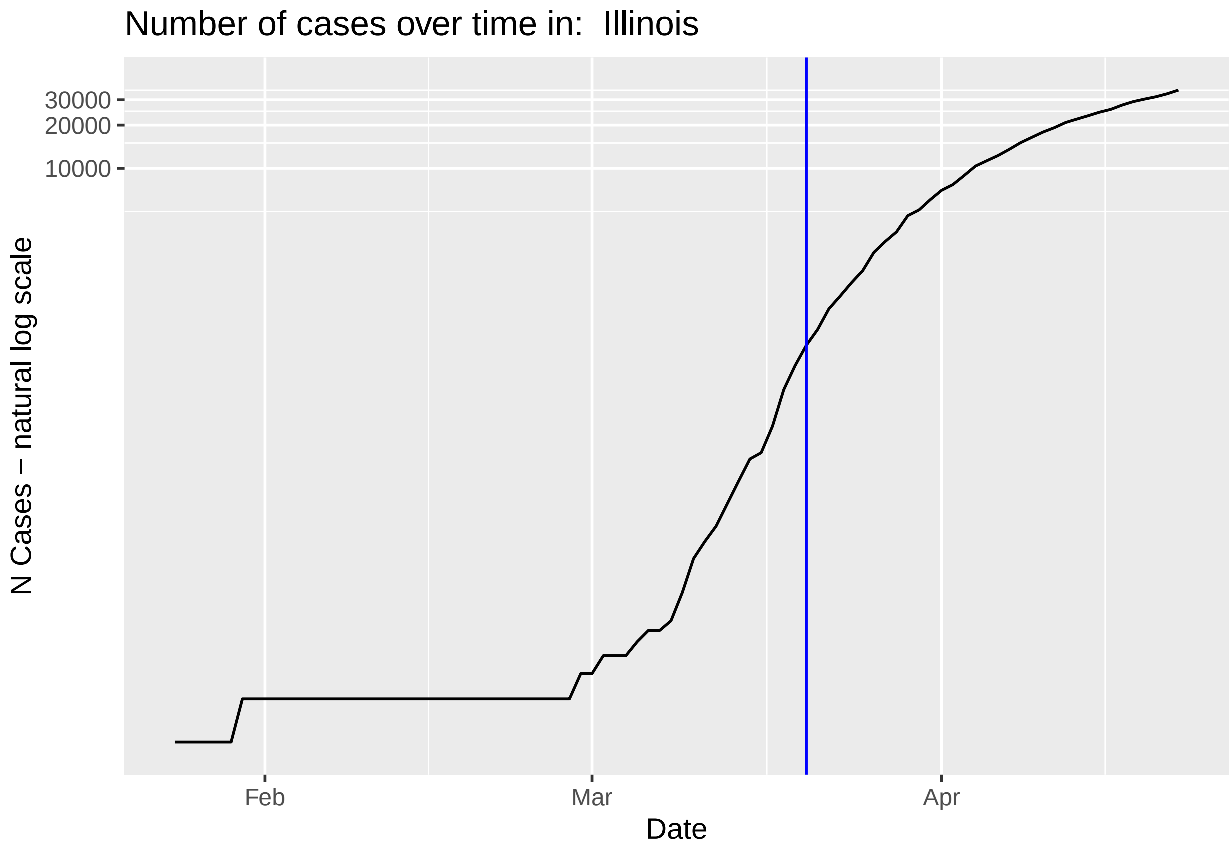
```
st <- us_counties %>%
  select(c(date, state, county, fips, cases, deaths)) %>%
  filter(state == state2select)

st_cases <- st %>% select(c(date, cases)) %>%
    group_by(date) %>%
    tally(cases)
```

Plotting the state specific data

Once the data is selected and grouped by state cases a general total cases per day model can be developed. Depending on the state this should not select the NA's and have different start times. Alternative methods are possible by converting to a wide format to maintain early NA data.

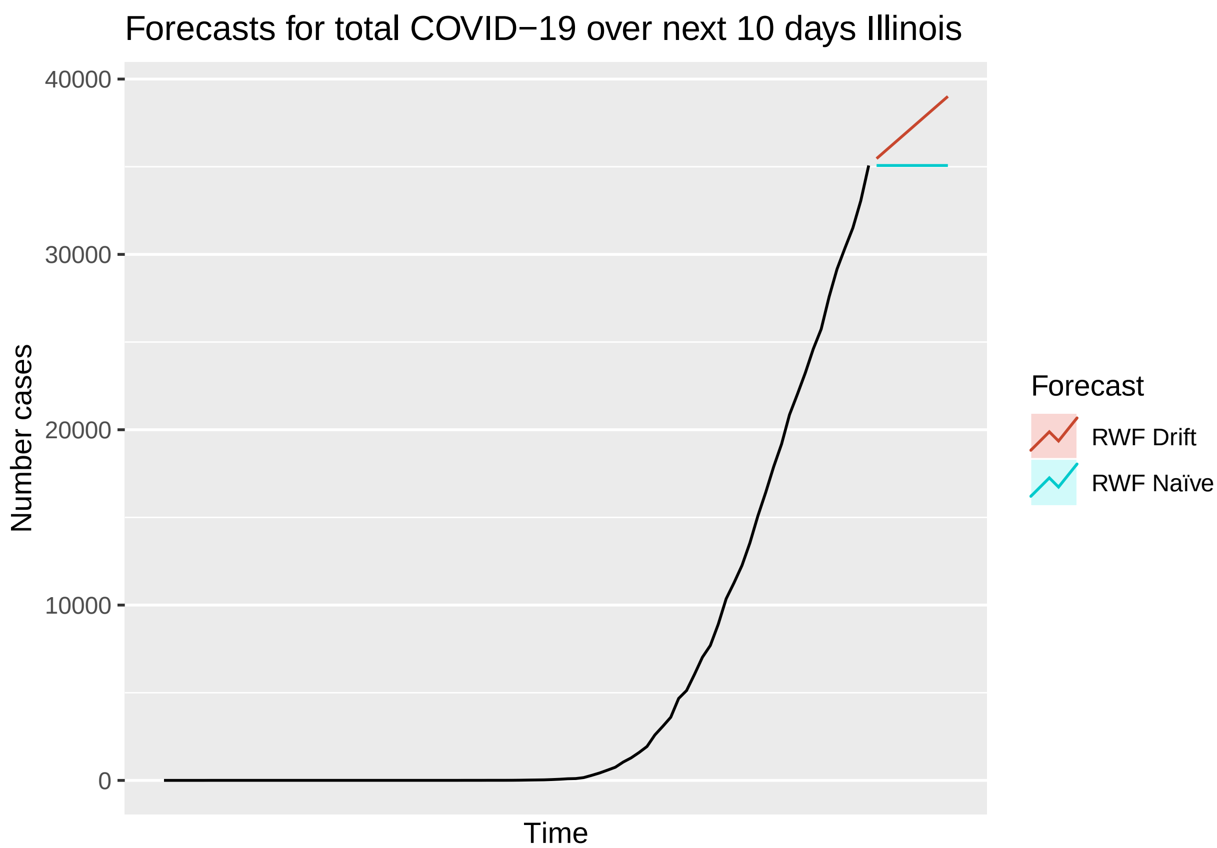


Depending on total case numbers(i.e. greater than 1000-5000) a log scale maybe easier to show trends and see recent trends. If under 5000 cases total this may over-represent trends (up or down) that are only part of the standard variance of the data.



Basic forecast model for the next 10 days.

To do this you will need to convert the date data into a time-series using the Forecast and lubridate packages in R.



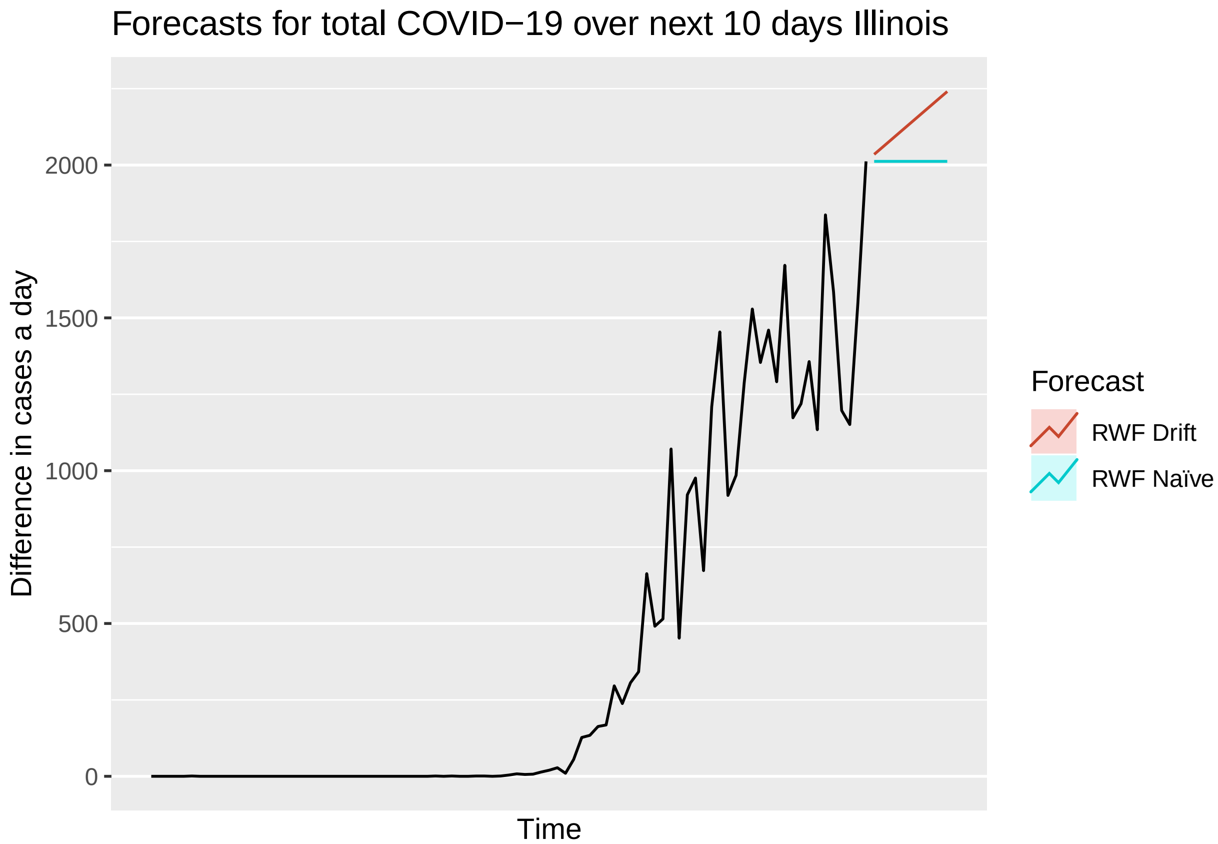
Part 2. Using linear modeling in daily difference to determine the trends

In this part it will be broken down into thre sections. First section will look at the standard model for the 10-days bofere and after a lock-down/stay in place order. Then a standard last 10-14 days compared to previous 10-14 days and both normalized to days 1 through 10 (or 14). Second part is the start of the difference model by looking at the previous 10-14 days and comparing to the time before that. This is appropriate in states that had not transistioned from lag-phase to exponetial growth phase (i.e. West Virginia). This shoud show a basic day-over-day trend. Third part is looking at the effect of the stay in place/lockdown order to determine if it had a measurable effect. NOTE: in states with minimal/no time in exponential growth this may not be an accurate measure and recommend that Part 2, section B be used (ie. comparing two different time frames).

NOTE: This is still being worked on and needs to have the stay in place order model done.

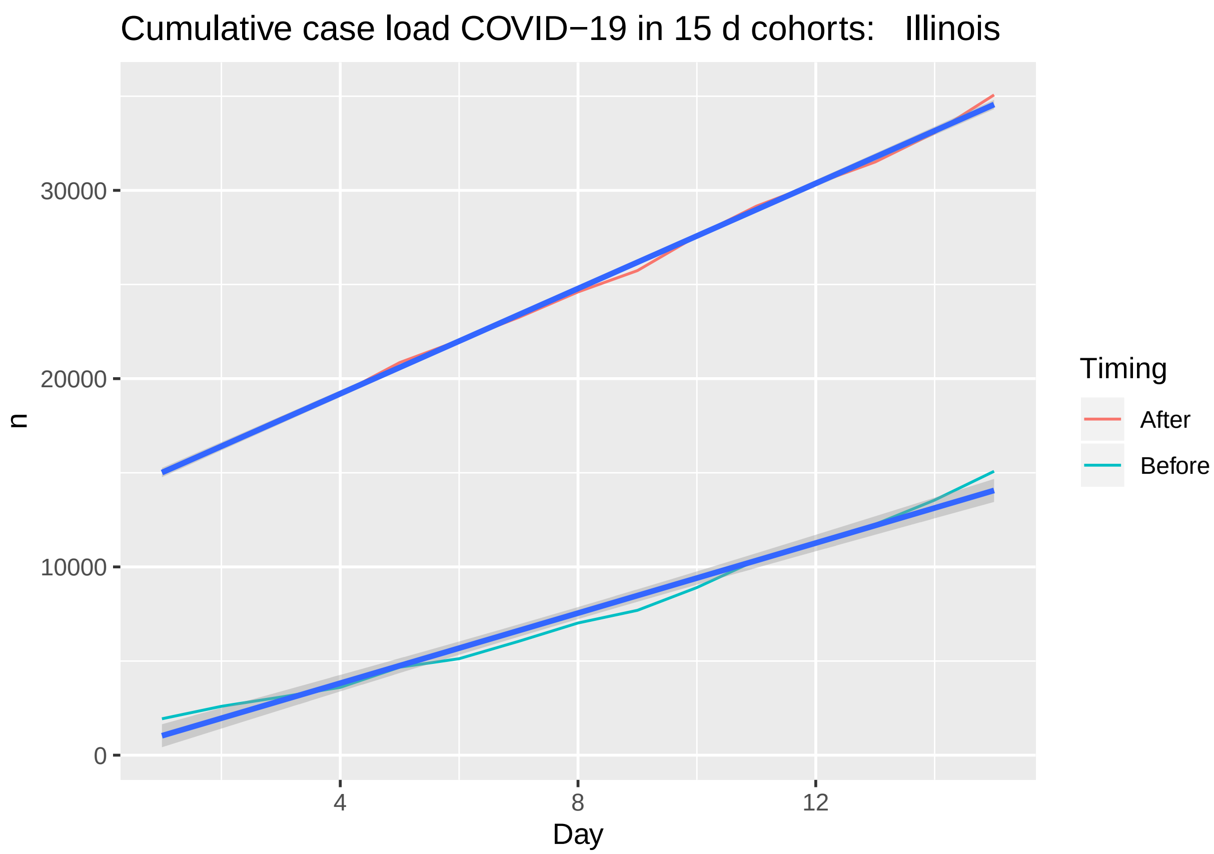
Part 2. Section A

Comparing the last n days versus previous time frame using standard case number and difference modeling.

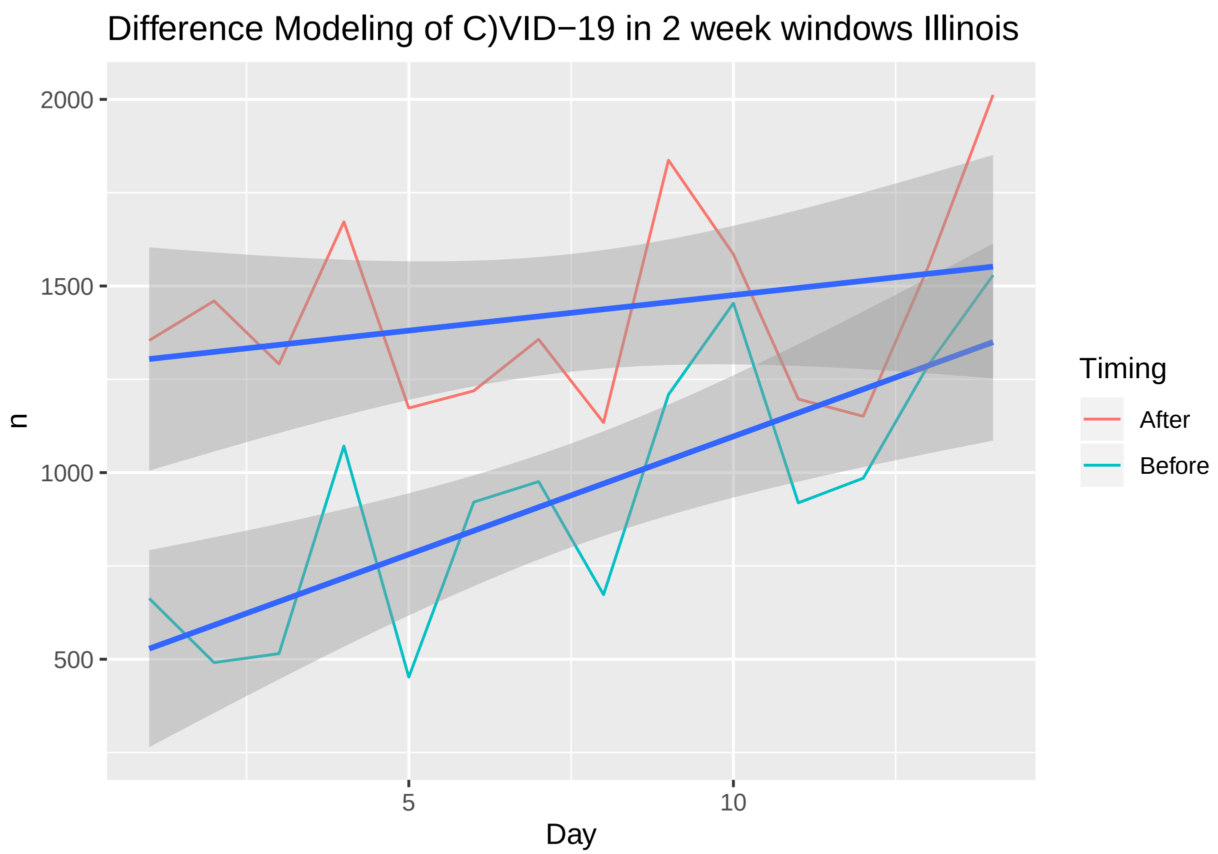


Above shows a basic difference model. If it appears that a form of stasis (i.e. random variation around an estimated mean) was achieved then it maybe possible to use ARIMA modeling for a better determination of actual Difference(cases).

After setting up the two different time frames in a tibble (i.e. 14 days for difference series is 15 days in standard series), the two differnt data sets are plotted.



After the days are set now its time to plot.



Checking the Linear Model of the before and after.

First part is to look at the LM for the before group. Slope of difference in cases per day is located under the Day row, Estimate column.

```
##
## Call:
## lm(formula = n ~ Day, data = diff_tib %>% filter(Timing == "Before"))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -329.0 -213.6   34.6  165.2  357.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   465.01     134.89   3.447  0.00483 **
## Day            63.19      15.84   3.989  0.00180 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 239 on 12 degrees of freedom
## Multiple R-squared:   0.57,  Adjusted R-squared:  0.5342
## F-statistic: 15.91 on 1 and 12 DF,  p-value: 0.001798
```

Second part is to look at the LM for the After group. Slope of difference in cases per day is located under the Day row, Estimate column.

```
##
## Call:
## lm(formula = n ~ Day, data = diff_tib %>% filter(Timing == "After"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -362.73 -200.74  -16.59  129.80  460.20
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1285.31     152.94   8.404 2.26e-06 ***
## Day            19.04      17.96   1.060     0.31
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 270.9 on 12 degrees of freedom
## Multiple R-squared:  0.08558,    Adjusted R-squared:  0.00938
## F-statistic: 1.123 on 1 and 12 DF,  p-value: 0.3101
```

Comparing the slopes of the line can give you an idea of how the last 14 days compare to the previous.

Linear modeling to look at before and after the stay in place order

Note: This is to be completed ASAP but was playing around with different windows for the above section to look at a moving window differenc model.