Franklin County PA Covid

Matthew Angle

Current Data

The data within this model is limited. There exists no easy package in R for PA Coronavirus cases by county. I've entered in this data manually.

```
knitr::opts_chunk$set(error = TRUE)
#load libs
library("tidyverse")
library("ggplot2")
library("httr")
library("rvest")
```

##Scraping PA Tables Making DF's

```
franklinCountyCorona <- data.frame("day" = c(seq(1,16)), "dates" = seq(as.Date("2020-03-20"), by = "days

franklinCountyCorona
```

```
##    day      dates cases
## 1    1 2020-03-20     1
## 2    2 2020-03-21     1
## 3    3 2020-03-22     1
## 4    4 2020-03-23     1
## 5    5 2020-03-24     3
## 6    6 2020-03-25     5
## 7    7 2020-03-26     5
## 8    8 2020-03-27     5
## 9    9 2020-03-28     7
## 10  10 2020-03-29    11
## 11  11 2020-03-30    12
## 12  12 2020-03-31    19
## 13  13 2020-04-01    21
## 14  14 2020-04-02    23
## 15  15 2020-04-03    26
## 16  16 2020-04-04    30
```

```
url <- 'https://www.health.pa.gov/topics/disease/coronavirus/Pages/Archives.aspx'
ws <- GET(url)

tbls <- html_nodes(content(ws), "table")
print((html_table(tbls[[4]])))
```

```
##                X1              X2     X3
## 1          County Number of Cases Deaths
## 2           Adams              21
## 3       Allegheny             552      3
```

```
## 4       Armstrong             12
## 5          Beaver             69      6
## 6         Bedford              4
## 7           Berks            235      2
## 8           Blair              5
## 9         Bradford             10
## 10          Bucks            488     11
## 11         Butler             84      2
## 12        Cambria              6
## 13        Cameron              1
## 14         Carbon             46      1
## 15         Centre             39
## 16        Chester            250      2
## 17        Clarion              4
## 18     Clearfield              7
## 19        Clinton              1
## 20       Columbia             20
## 21       Crawford              5
## 22     Cumberland             54      2
## 23        Dauphin             99      1
## 24       Delaware            616     13
## 25           Erie             19
## 26        Fayette             23      1
## 27         Forest              2
## 28       Franklin             27
## 29         Greene             12
## 30     Huntingdon              4
## 31        Indiana              9
## 32        Juniata              7
## 33     Lackawanna            146      6
## 34      Lancaster            291      5
## 35       Lawrence             22      2
## 36        Lebanon             87
## 37         Lehigh            804      7
## 38        Luzerne            648      5
## 39       Lycoming             10
## 40         McKean              1
## 41         Mercer             14
## 42        Mifflin              4
## 43         Monroe            484     11
## 44     Montgomery            982     17
## 45        Montour             19
## 46    Northampton            588     11
## 47 Northumberland              9
## 48          Perry              5      1
## 49   Philadelphia           2610     24
## 50           Pike             97      1
## 51         Potter              3
## 52     Schuylkill             77
## 53         Snyder              6      1
## 54       Somerset              3
## 55       Sullivan              1
## 56    Susquehanna              5
## 57          Tioga              3
```

```
## 58          Union               5
## 59        Venango               3
## 60         Warren               1
## 61     Washington              46
## 62          Wayne              28
## 63   Westmoreland             135
## 64        Wyoming               5
## 65           York             144      1
```

##Fit

The fit model

```
fit <- lm(formula = log(cases) ~ day , data = franklinCountyCorona)


summary(fit)
```

```
##
## Call:
## lm(formula = log(cases) ~ day, data = franklinCountyCorona)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62387 -0.13153  0.04508  0.21350  0.47008
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.40709    0.15030  -2.709    0.017 *
## day          0.25774    0.01554  16.582 1.34e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2866 on 14 degrees of freedom
## Multiple R-squared:  0.9515, Adjusted R-squared:  0.9481
## F-statistic:   275 on 1 and 14 DF,  p-value: 1.342e-10
```

##Using the model

Using the model to generate data for an additional amount of time. Placed in final model

```
newDay <-data.frame("day" = c(seq(1, 30)))
nextTwentyDays <-predict(fit, newDay)
nextTwentyDays <- as.data.frame(nextTwentyDays)
tmp <- seq(as.Date("2020-03-20"), by = "days", length.out = 30)
names(nextTwentyDays)[1] <- "cases"
#has a null value assume model starts at 1
#nextTwentyDays[1,1] <- 1
nextTwentyDays <- mutate(nextTwentyDays,
                         "day"   = c(seq(1, 30)),
                         "cases" = ceiling(exp(nextTwentyDays$cases)),
                         "dates" = tmp)
finalModel<- merge(nextTwentyDays, franklinCountyCorona, by = "dates", all = TRUE)

(finalModel)
```

```
##         dates cases.x day.x day.y cases.y
```

```
## 1  2020-03-20       1     1     1       1
## 2  2020-03-21       2     2     2       1
## 3  2020-03-22       2     3     3       1
## 4  2020-03-23       2     4     4       1
## 5  2020-03-24       3     5     5       3
## 6  2020-03-25       4     6     6       5
## 7  2020-03-26       5     7     7       5
## 8  2020-03-27       6     8     8       5
## 9  2020-03-28       7     9     9       7
## 10 2020-03-29       9    10    10      11
## 11 2020-03-30      12    11    11      12
## 12 2020-03-31      15    12    12      19
## 13 2020-04-01      19    13    13      21
## 14 2020-04-02      25    14    14      23
## 15 2020-04-03      32    15    15      26
## 16 2020-04-04      42    16    16      30
## 17 2020-04-05      54    17    NA      NA
## 18 2020-04-06      69    18    NA      NA
## 19 2020-04-07      90    19    NA      NA
## 20 2020-04-08     116    20    NA      NA
## 21 2020-04-09     150    21    NA      NA
## 22 2020-04-10     194    22    NA      NA
## 23 2020-04-11     250    23    NA      NA
## 24 2020-04-12     324    24    NA      NA
## 25 2020-04-13     419    25    NA      NA
## 26 2020-04-14     542    26    NA      NA
## 27 2020-04-15     701    27    NA      NA
## 28 2020-04-16     907    28    NA      NA
## 29 2020-04-17    1174    29    NA      NA
## 30 2020-04-18    1519    30    NA      NA
```

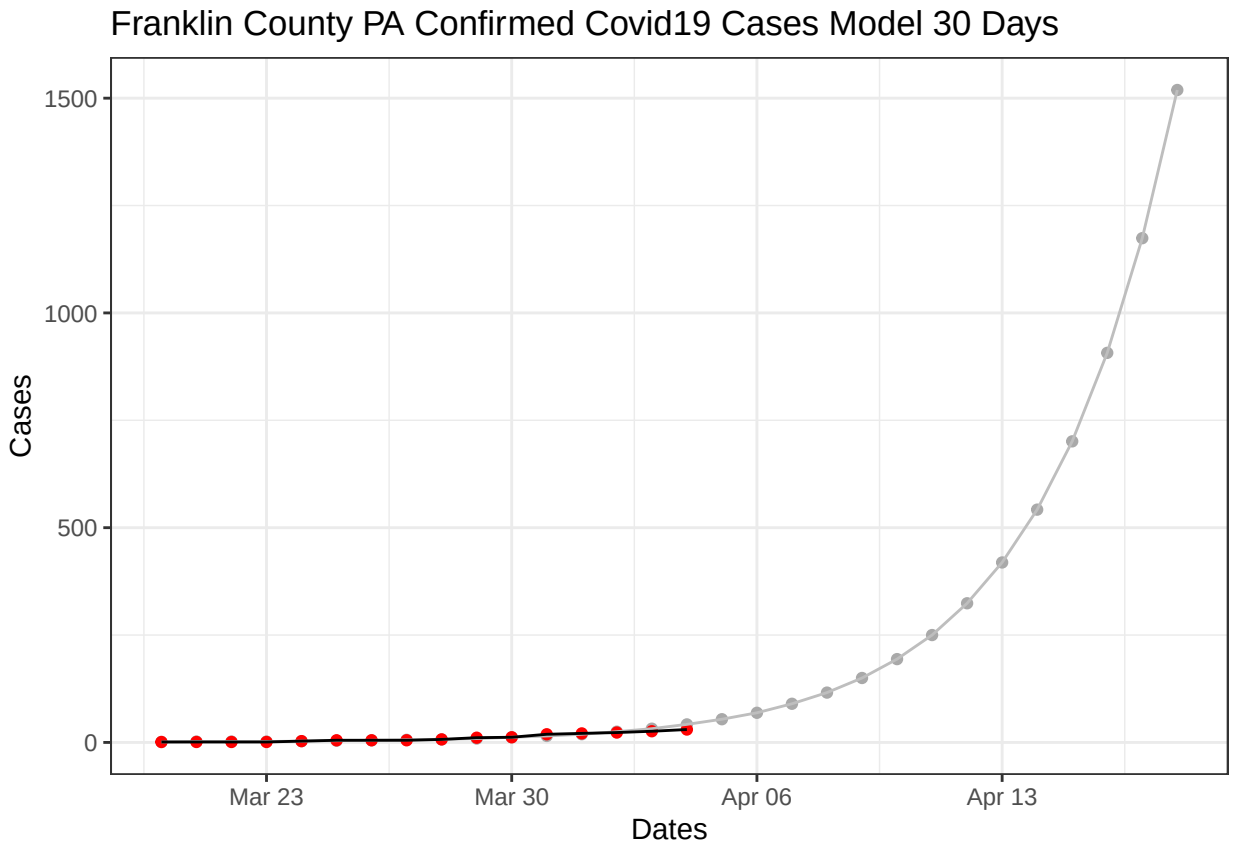
##Plot the data

Used the data from the model to plot

```
ggplot(finalModel, aes(x = dates)) +
  geom_point(aes(y = cases.x), color = "darkgrey") +
  geom_point(aes(y = cases.y), color = "red") +
  geom_path(aes(y = cases.x), color = "grey") +
  geom_path(aes(y = cases.y), color = "black") +
  labs(x = "Dates", y = "Cases") +
  ggtitle("Franklin County PA Confirmed Covid19 Cases Model 30 Days") +
  theme_bw()
```

```
## Warning: Removed 14 rows containing missing values (geom_point).
```

```
## Warning: Removed 14 row(s) containing missing values (geom_path).
```


Franklin County PA Confirmed Covid19 Cases Model 30 Days
1500
1000
500
0
Cases
Mar 23
Mar 30
Apr 06
Apr 13
Dates