Coronavirus in the US

ERROR(CONFIRMED + DEATHS) = ?

GRAD 2

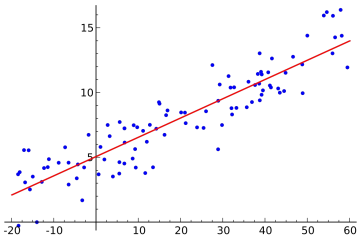
Team Members:

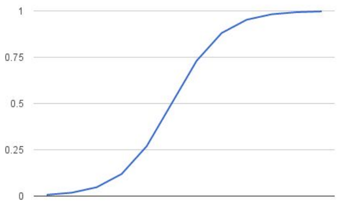
- Ali Al-Ghaithi
- Bikram Maharjan
- Dongqi Lai
- Mohammad H Hasan

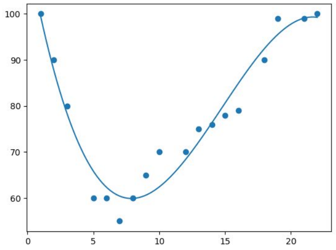
Model Considerations

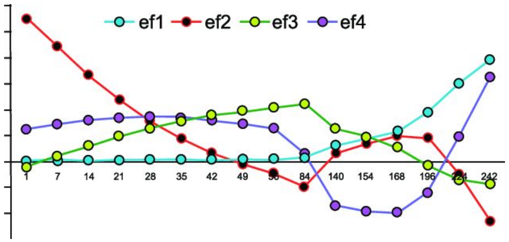
- Time Series Models
  - **ARIMA - winner!**

- Regression Models:
  - Linear Regression
  - Logistic Regression
  - Random Forest - regression
  - Polynomial Regression
  - Exponential Regression









Design Matrix

Model	Other notes	Error
Regression	- **Ntree** = 400 - **Density** of the county (predictors) +1	**Confirmed Cases:** **Error** (Random Forest): 5.4 Million **Error** (Linear Regression): 4.2 Million
Time Series		**Total Error** (Confirmed + Deaths): 304k

Design Matrix

If Time Series:

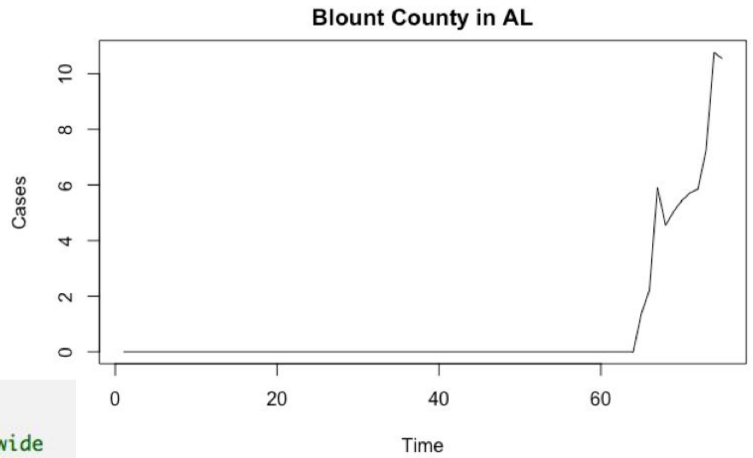
- Delete the last 7 days
- For 3144 counties

```
covid_confirmed_usafacts1    #target <- c("Statewide Unallocated", "Washington")
covid_confirmed_usafacts2 <- covid_confirmed_usafacts1 %>% filter(County.Name != "Statewide
Unallocated" & County.Name != "Washington")
```

- For the 51 Unallocated States

```
target <- c("Statewide Unallocated", "Washington")
covid_confirmed_usafacts2<- covid_confirmed_usafacts1 %>% filter(County.Name %in%target)
```

- Converted from Wide to Long
- Cleaning the date column



2020-03-20	0
2020-03-21	0
2020-03-22	0
2020-03-23	0
2020-03-24	0
2020-03-25	1
2020-03-26	2
2020-03-27	5
2020-03-28	5
2020-03-29	5
2020-03-30	5
2020-03-31	5
2020-04-01	5
2020-04-02	6
2020-04-03	9
2020-04-04	10
2020-04-05	10

What is ARIMA Models?

ARIMA Models

The Idea

Capture autocorrelation in the series by modeling it directly

Uses

Forecasting

Advantages: Strong underlying theory, Flexible

Key concepts: order, differencing

Autoregressive Model: *AR(p)*

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \varepsilon_t$$

Autoregressive Moving Average Model: *ARMA(p,q)*

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \ldots + \beta_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

The future is "similar" to the past (in a probabilistic sense)

Autoregressive Integrated Moving Average Model:
ARIMA(p,d,q)

1. First apply differencing (order d)
2. Then fit ARMA(p,q):

Lag-1 differencing

Seasonal differencing

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \ldots + \beta_p Y_{t-p}$$
$$+ \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

Seasonal ARIMA(p,d,q)(P,D,Q)

Lag-1 differencing:
$y_t - y_{t-1}$

Used for removing **trend**

Order d = {0,1,2,…}

Seasonal (lag-M) differencing:
$y_t - y_{t-M}$

Used for removing **seasonality**

Order D = {0=none, 1=once}

Major Assumption: *Stationarity*

No trend/seasonality
Constant level, variance & autocorrelations

Model Development

- Using Auto.Arima models
- Each county is a time series

Create a function that will 3144 counties

1. Filter data using countyFIPS
2. Use xts function to convert to Time Series
3. Apply Auto.Arima function
4. Forecast 7 days ahead

```
for ( i in 1:3144) {
  countyFIPS_IN =  countyFIPS_vc[i]
  new_data_Abbeville <- new_data %>% filter(countyFIPS == countyFIPS_IN)

library(xts)
new_data_Abbeville_ts <- xts(new_data_Abbeville$count, order.by=new_data_Abbeville$dates)
new_data_Abbeville_ts
library(forecast)

new_data_Abbeville_ts_model <- auto.arima(new_data_Abbeville_ts)
futurVal <- forecast(new_data_Abbeville_ts_model,h=7)
pred_out[i] = round(as.data.frame(futurVal)[7,1])
}
```

for 51 States

1. Filter data using stateFIPS
2. Use xts function to convert Time Series
3. Apply Auto.Arima function
4. Forecast 7 days ahead

Model Validation - (Find the test/validation error):

$$
\begin{aligned}
\text{total error} &= \text{case error} + \text{death error} \\
&= \sum_{i=1}^{n} w_i (c_i - \hat{c}_i)^2 + \sum_{i=1}^{n} w_i (d_i - \hat{d}_i)^2
\end{aligned}
$$

Total error= 3144 counties cases error + 51 unallocated states cases error + 3144 counties deaths error+51 unallocated states deaths error

Traditional cross validation cannot be applied for time-series problems (data is temporally dependent)

Error_for_deaths_unallocated: 11.26609

Error_for_deaths_Counties: 9491.097

Deaths

9502.363

Error_for_cases_unallocated: 9212.122

Error_for_cases_counties: 285629.4

Cases

294841.6

Total_Error:

304343.9

Dr. Zoe Total Error = 283725.16

Model Comparison (Progress over time):

<u>Time Series models</u>:	<u>Regression Models</u>:
Good: - We Lose a lot of information - Help us understand important days during the pandemic	**Good**: - More information - Better Validation
Bad: - Need to convert data from Wide -> Long - We do not have many days to disguise - More complex CV methods. For eg: Nested CV.	**Bad**: - High error margin - Not all regression model works - Logistic [family = “bernoulli”] does not run - Hard to find predictors for county level

Time Series model worked much better than Regression

We are GRAD 2 Team