Repo Link: https://github.com/qzhang21/yalehack2021

Multiple Regression for Predicting COVID-19 Death Proportion

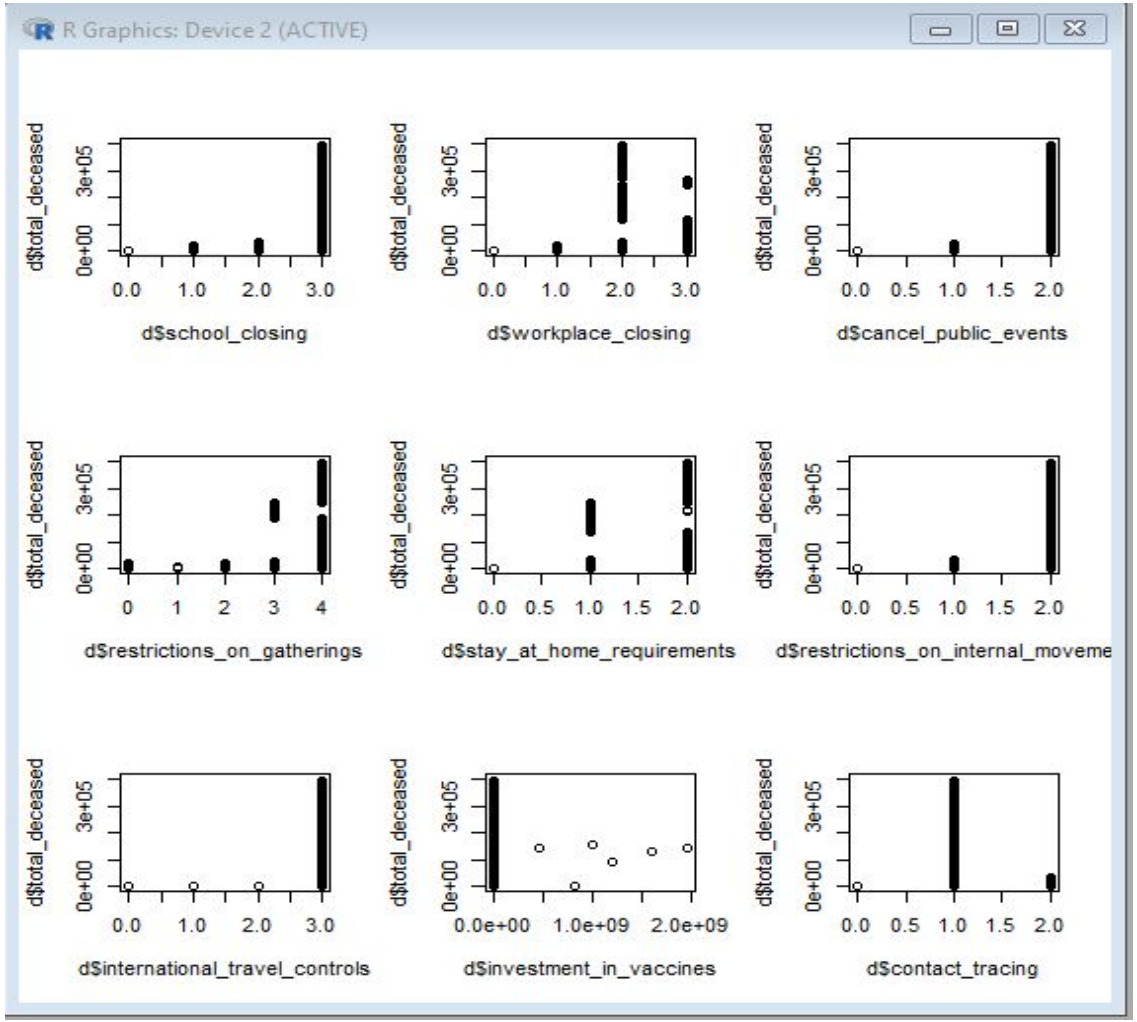
Marie Zhang and Alina Zheng

COVID-19 and its Impacts

The novel coronavirus (COVID-19) emerged into the mainstream about a year ago. Since then, it has wrought unimaginable destruction, resulting in more than two million deaths worldwide at the time of this writing. The disease has also brought to light the precarious economic, government, and healthcare infrastructure of countries that were thought to be unshakeable, most notably the United States. Our goal for this project is to better understand any variables that may have impacted the spread of the disease.

Our Process

- Extensive data cleaning!
  - Only kept U.S. data because different healthcare systems will introduce many confounding variables
- Plotting bivariate data to find correlations between two variables (with particular interest in seeing how **total deceased** was affected by contact tracing, school closings, travel restrictions, etc.)
- Multiple linear regression
  - Eliminate multicollinearity with VIF
  - Compare the full model to AIC and BIC from stepwise elimination
  - Run model diagnostics

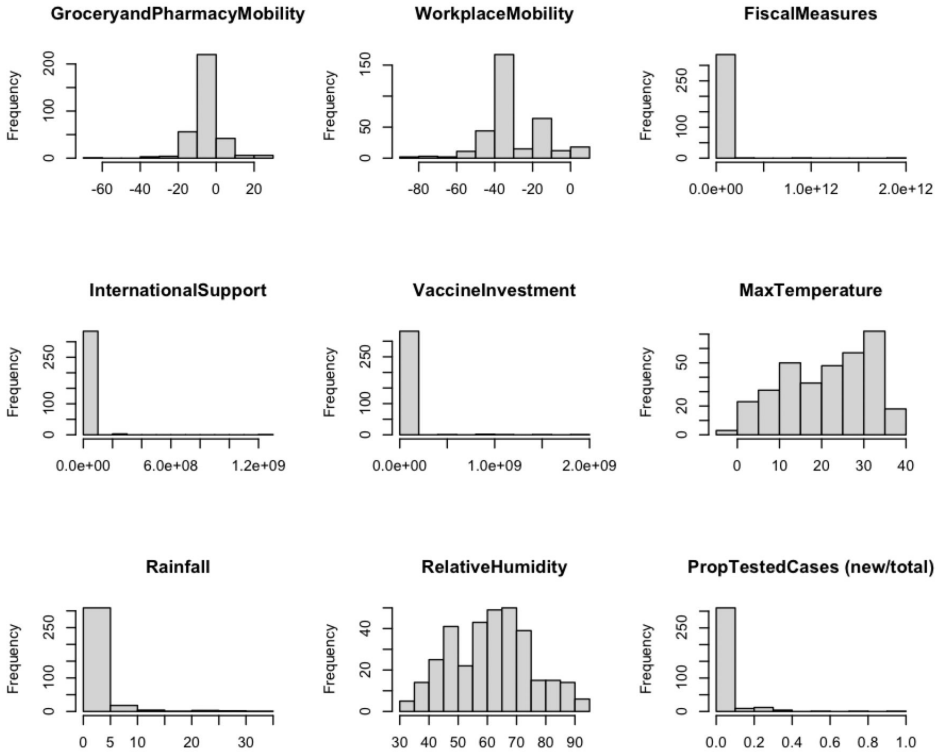

R Graphics: Device 2 (ACTIVE)
d$total_deceased
0e+00
3e+05
0.0
1.0
2.0
3.0
d$school_closing
d$workplace_closing
0.5
1.5
d$cancel_public_events
0
1
2
3
4
d$restrictions_on_gatherings
d$stay_at_home_requirements
d$restrictions_on_internal_moveme
d$international_travel_controls
0.0e+00
1.0e+09
2.0e+09
d$investment_in_vaccines
d$contact_tracing

VIF Procedure for Multicollinearity

- Utilized stepwise elimination procedures using *vifstep()* from the usdm library in R
  - Tried multiple thresholds, went with th=2 (highest R^2 between variables could not exceed 0.5)
- Ultimately eliminated categorical variables (leveled ones such as restrictions) due to resulting rank deficiencies. There was high imbalance across categories (most areas adopted policies of similar levels across the board).

Data Imbalance

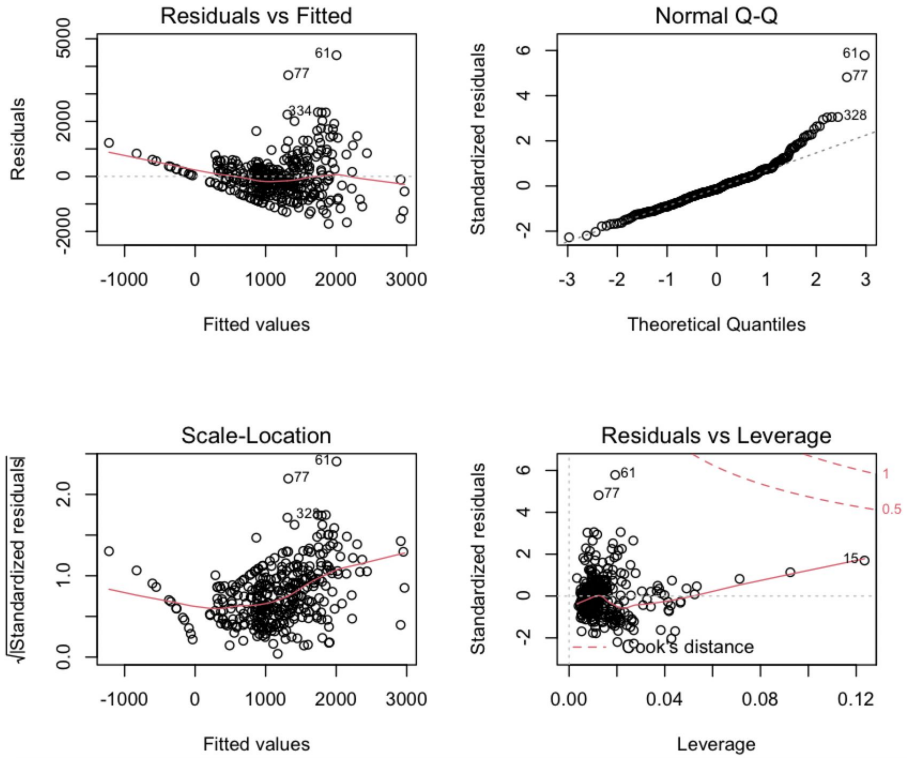
- Lots of highly skewed data! :(



Solution: transform!

Outlier Identification after Transformation

- We identified outliers after transformation since transformation may have taken care of some outliers
- Here is a sample plot series from the BIC model (the others look very similar)



Additive Model Comparison

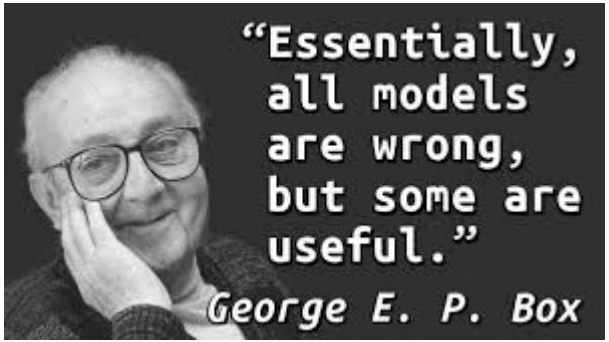
- Obtained full, AIC, and BIC models
- Most additive elements were significant in the AIC and BIC models
- Ended up choosing the most parsimonious model, since additional complexity always better explains the data.

```
> summary(model_bic)

Call:
lm(formula = y.var ~ mobility_workplaces + maximum_temperature +
    relative_humidity + prop_tested, data = x.matrix)

Residuals:
     Min       1Q   Median       3Q      Max
-1637.77  -465.25   -79.78   324.20  2405.04

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -49.898    322.319  -0.155  0.87707
mobility_workplaces -21.255      2.575  -8.253 3.83e-15 ***
maximum_temperature -38.812      3.958  -9.806  < 2e-16 ***
relative_humidity    -7.880      2.955  -2.667  0.00803 **
prop_tested        -917.973    124.567  -7.369 1.41e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 689.8 on 327 degrees of freedom
Multiple R-squared:  0.3983,  Adjusted R-squared:  0.3909
F-statistic: 54.11 on 4 and 327 DF,  p-value: < 2.2e-16
```

Additive Model Equation+Next Steps

- Equation: new_deceased-hat = -49.898 - 21.255*mobility_workplaces - 38.812*maximum_temperature - 7.880*relative_humidity - 917.973*prop_tested
- Next Steps:
  - Explore interactions between variables to refine the model
  - Cross-validate on another dataset
  - Consider adding other variables such as demographics
- Note:
  - All models are wrong, but some are useful. Our model may not always be correct, but hopefully it gives some insight on factors to explore in the spread of the virus.

Thanks!



Original dataset citation:

```
@article{Wahltinez2020,

  author = "O. Wahltinez and others",

  year = 2020,

  title = "COVID-19 Open-Data: curating a fine-grained, global-scale data repository for SARS-CoV-2",

  note = "Work in progress",

  url = {https://goo.gle/covid-19-open-data},

  }
```