STAT 253 Final Project: Analysis of COVID-19

Max Dodge

Data Context

New York Times Data:

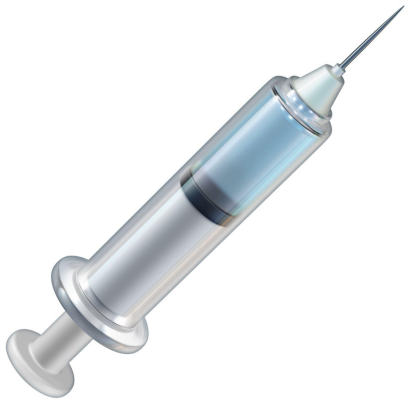
- County level case and death data for each day from December 1st 2020 to April 4th 2021.
- Mask usage by county. Survey conducted July 2020.

CDC Data:

- State level data for weekly vaccine allotment for each of the big three vaccines. (Pfizer, Moderna, Jenssen).

US Census Bureau Data:

- 2019 estimated population by county.



Variables

Location/Date:

- date, county, state, fips (county-state identifier)

COVID:

- cases, deaths, cases_per10k (population adjusted), severe_outbreak

Demographics:

- pop_county, pop_state, prop_statepop (relative population)

Mask Usage:

- NEVER, RARELY, FREQUENTLY, ALWAYS

Vaccine Data:

- janssen_doses, pfizer_dose_1, pfizer_dose_2, moderna_dose_1, moderna_dose_2, total_doses, estimated_doses (population adjusted)

Research Questions

Regression Task:

- Do areas with high mask usage and vaccine distribution correspond to lower levels of covid cases?

Classification Task:

- Which counties are high risk, >50% likelihood of a severe covid outbreak of >2,000 cases per 10,000?

HW3 Investigations - Methods (linear)

OLS

- First run used almost all the variables, almost certainly overfit. MAE = 199.86

LASSO

- Identified date, NEVER, ALWAYS, SOMETIMES, pop_county as the most important variables. MAE almost identical.

Forward Stepwise

- Identified date, ALWAYS, RARELY, FREQUENTLY as the most important variables. MAE almost the same again.

HW3 Investigations - Results (linear)

- Residual plots showed clear nonlinearity. Without penalization OLS will be overfit.
- LASSO and Forward stepwise had much simpler models with very similar error metrics.
- Non-linearity needs to be addressed.

HW3 Investigations - Nonlinear

- Same LASSO and Stepwise as before, just using natural cubic splines with three degrees of freedom.

LASSO

- Different variables deemed important in the nonlinear model. (date, NEVER, RARELY, ALWAYS, FREQUENTLY, SOMETIMES, pop_county, estimated_doses). MAE drops.

Stepwise

- Different variables deemed important in the nonlinear model. (date, ALWAYS). MAE drops.

HW3 Investigations - Summary

Clear nonlinearity.

Mask usage seems to be important.

Vaccine distribution not as important as expected.

“Best” model is the LASSO made from splines.

Summary Statistics

Linear LASSO

	alpha <dbl>	lambda <dbl>	RMSE <dbl>	Rsquared <dbl>	MAE <dbl>	RMSESD <dbl>	RsquaredSD <dbl>	MAESD <dbl>
1	1	0	267.5049	0.3021016	199.866	1.68471	0.005026059	0.9463669

```
15 x 1 sparse Matrix of class "dgCMatrix"
                             1
(Intercept)     -6.726769e+04
date             3.652825e+00
NEVER            4.800653e+02
ALWAYS          -4.185640e+02
SOMETIMES        5.616530e+02
RARELY           .
FREQUENTLY       .
janssen_doses    .
pfizer_dose_1    .
pfizer_dose_2    .
moderna_dose_1   .
moderna_dose_2   .
total_doses      .
pop_county       3.683885e-05
estimated_doses  .
```

Linear Stepwise

	nvmax <int>	RMSE <dbl>	Rsquared <dbl>	MAE <dbl>	RMSESD <dbl>	RsquaredSD <dbl>	MAESD <dbl>
1	10	267.9129	0.299959	200.3371	1.70537	0.004239484	0.5701235

```
  (Intercept)          date        ALWAYS        RARELY    FREQUENTLY
-69010.563822      3.775993   -964.325357   -534.458787   -594.191041
```

Summary Statistics

LASSO with Splines

```
43 x 1 sparse Matrix of class "dgCMatrix"
                                      1
(Intercept)                561.9656913
ns(date, 3)1               338.1709954
ns(date, 3)2               659.6580277
ns(date, 3)3               289.9358191
ns(NEVER, 3)1              152.0726435
ns(NEVER, 3)2               43.7043213
ns(NEVER, 3)3                .
ns(RARELY, 3)1               .
ns(RARELY, 3)2              14.3109165
ns(RARELY, 3)3               .
ns(ALWAYS, 3)1           -118.2638710
ns(ALWAYS, 3)2            -53.4088592
ns(ALWAYS, 3)3           -178.0984231
ns(FREQUENTLY, 3)1         10.5171180
ns(FREQUENTLY, 3)2          .
ns(FREQUENTLY, 3)3        -20.6338843
ns(pfizer_dose_1, 3)1       .
ns(pfizer_dose_1, 3)2       .
ns(pfizer_dose_1, 3)3       .
ns(pfizer_dose_2, 3)1       .
ns(pfizer_dose_2, 3)2       .
ns(pfizer_dose_2, 3)3       .
ns(janssen_doses, 3)1       .
ns(janssen_doses, 3)2       .
ns(janssen_doses, 3)3       .
ns(SOMETIMES, 3)1         157.2126077
ns(SOMETIMES, 3)2         106.1637009
ns(SOMETIMES, 3)3           0.2094672
ns(moderna_dose_1, 3)1      .
ns(moderna_dose_1, 3)2      .
ns(moderna_dose_1, 3)3      .
ns(moderna_dose_2, 3)1      .
ns(moderna_dose_2, 3)2      .
ns(moderna_dose_2, 3)3      .
ns(pop_county, 3)1          7.8789069
ns(pop_county, 3)2          .
ns(pop_county, 3)3          .
ns(total_doses, 3)1         .
ns(total_doses, 3)2         .
ns(total_doses, 3)3         .
ns(estimated_doses, 3)1  -288.7775369
ns(estimated_doses, 3)2     .
ns(estimated_doses, 3)3     .
```

	alpha <dbl>	lambda <dbl>	RMSE <dbl>	Rsquared <dbl>	MAE <dbl>	RMSESD <dbl>	RsquaredSD <dbl>	MAESD <dbl>
1	1	0	263.73	0.3216572	195.858	1.888258	0.004812359	0.8371789

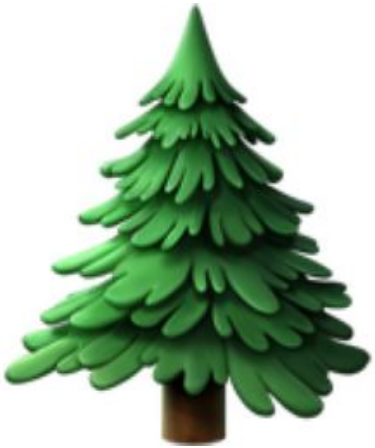
Summary Statistics

Stepwise with Splines

(Intercept)	ns(date, 3)1	ns(date, 3)2	ns(date, 3)3	ns(ALWAYS, 3)3
409.9869	354.3039	712.1190	310.0315	-388.2847

	nvmax <int>	RMSE <dbl>	Rsquared <dbl>	MAE <dbl>	RMSESD <dbl>	RsquaredSD <dbl>	MAESD <dbl>
1	10	264.8274	0.3160049	196.9682	1.673656	0.004735475	0.7865038

Classification Analysis - Methods



Logistic Regression

- Cutoff was 50% or higher, used a LASSO logistic regression model to predict high/low risk counties. 10 fold CV.

Trees

- Used a random forest model with 50 trees to predict counties where outbreaks were present.

Classification Analysis - Results

Logistic Regression:

- Variable Importance echos the variable importance from the cubic splines LASSO. Pretty High accuracy but also high and similar NIR.

Random Forest:

- Mask usage again comes up as important. Vaccines not so much.

Classification Analysis - Summary

Logistic Regression

alpha <dbl>	lambda <dbl>	AUC <dbl>	Sens <dbl>	Spec <dbl>	Accuracy <dbl>	AUCSD <dbl>	SensSD <dbl>	SpecSD <dbl>	AccuracySD <dbl>
1	0	0.7660099	0.9420646	0.2192787	0.770469	0.002483562	0.001337234	0.004927102	0.001758362

	severe_outbreak <chr>	n <int>
1	Not_Severe	298643
2	Severe	92973

```
[1] 0.7625914
```

Random Forests

```
Call:
 randomForest(x = x, y = y, ntree = 50, mtry = param$mtry)
               Type of random forest: classification
                     Number of trees: 50
No. of variables tried at each split: 5

        OOB estimate of  error rate: 0.41%
Confusion matrix:
           Not_Severe Severe class.error
Not_Severe     298120    523 0.001751255
Severe           1092  91881 0.011745345
```

Variable Importance

Logistic Regression

```
15 x 1 sparse Matrix of class "dgCMatrix"
                         1
(Intercept)     -4.361385e+02
date             2.329714e-02
NEVER            3.717505e+00
ALWAYS          -1.803531e+00
SOMETIMES        4.199349e+00
RARELY           1.147487e+00
FREQUENTLY       .
janssen_doses    .
pfizer_dose_1    .
pfizer_dose_2    .
moderna_dose_1   1.598996e-07
moderna_dose_2   4.900417e-20
total_doses      .
pop_county       1.826308e-07
estimated_doses  .
```

Random Forest

	predictor <chr>	MeanDecreaseGini <dbl>
ALWAYS	ALWAYS	21657.21041
pop_county	pop_county	21128.64861
date	date	20792.37165
SOMETIMES	SOMETIMES	19174.20377
NEVER	NEVER	18861.87501
RARELY	RARELY	18362.74993
FREQUENTLY	FREQUENTLY	17606.16673
estimated_doses	estimated_doses	322.50068
total_doses	total_doses	206.89655
moderna_dose_1	moderna_dose_1	206.25595

Conclusion

Cubic Splines LASSO is the best regression model, Random Forest is the preferred classification model. Reason it is likely so accurate is because cases are cumulative, makes it easier to split by date and get a very accurate prediction.

Mask usage seems to be very important in predicting how many cases a particular county will see.

Vaccine data seems less strong as a predictor.

More variables are likely needed. Such as population density and political leanings would be interesting.

Changing cases from cumulative to new cases would likely help make analyses more understandable.