P8106 Midterm Report

Arthur Starodynov, Ekaterina Hofrenning, Lauren Lazaro

2024-03-26

Introduction

The COVID-19 illness was first identified in late 2019, and quickly spread into a worldwide yearslong pandemic. A study was designed to combine three cohort studies following participants for multiple years. The collection of medical history data and personal characteristics enables the study of risk factors for extended recovery time. In this paper, we will investigate the medical and demographic risk factors for extended recovery times through training several predictive models.

Exploratory analysis and data visualization

Table 1 reports the demographic and medical characteristics of the cohort. The cohort is composed of 3000 participants, 2000 from study A and 100 from study B. The patients have a mean age of 60.2, mean recovery time of 42 days, and 60% vaccination rate. Additionally, there is an approximately equal balance of the genders, the patients are primarily white at 66%, and half of the cohort has hypertension. Overall, it appears that this cohort is slightly older and on the unhealthier side. We set the seed to 2024 for reproducibility of our models.

Table 1: **Table 1. Patient Characteristics**

Characteristic	**N = 3,000**
Age	60.2 (4.5)
Gender	
Female	1,544 (51%)
Male	1,456 (49%)
Race	
Asian	158 (5.3%)
Black	604 (20%)
Hispanic	271 (9.0%)
White	1,967 (66%)
Smoking	
Former Smoker	319 (27%)
Never Smoker	859 (73%)
Unknown	1,822
Height	169.9 (6.0)
Weight	80 (7)
BMI	27.76 (2.79)
Hypertension	1,492 (50%)
Diabetes	463 (15%)
SBP	130 (8)
LDL	110 (20)
Vaccine	1,788 (60%)

Characteristic	N = 3,000
Severity	321 (11%)
Study	
A	2,000 (67%)
B	1,000 (33%)
Recovery_time	42 (23)

Next, we visualized the individual relationships between COVID-19 recovery time and the available predictors through scatter plots for the continuous variables and violin plots for the categorical variables. Figure1 reports the scatter plots and Figure 2 reports the violin plots. The relationships between the continuous variables and recovery time appear to be mostly linear, with some possible non-linearity between BMI and COVID-19 recovery time. The violin plots show that the unhealthier patients tend to see longer recovery times.

Figure 1. Scatterplots

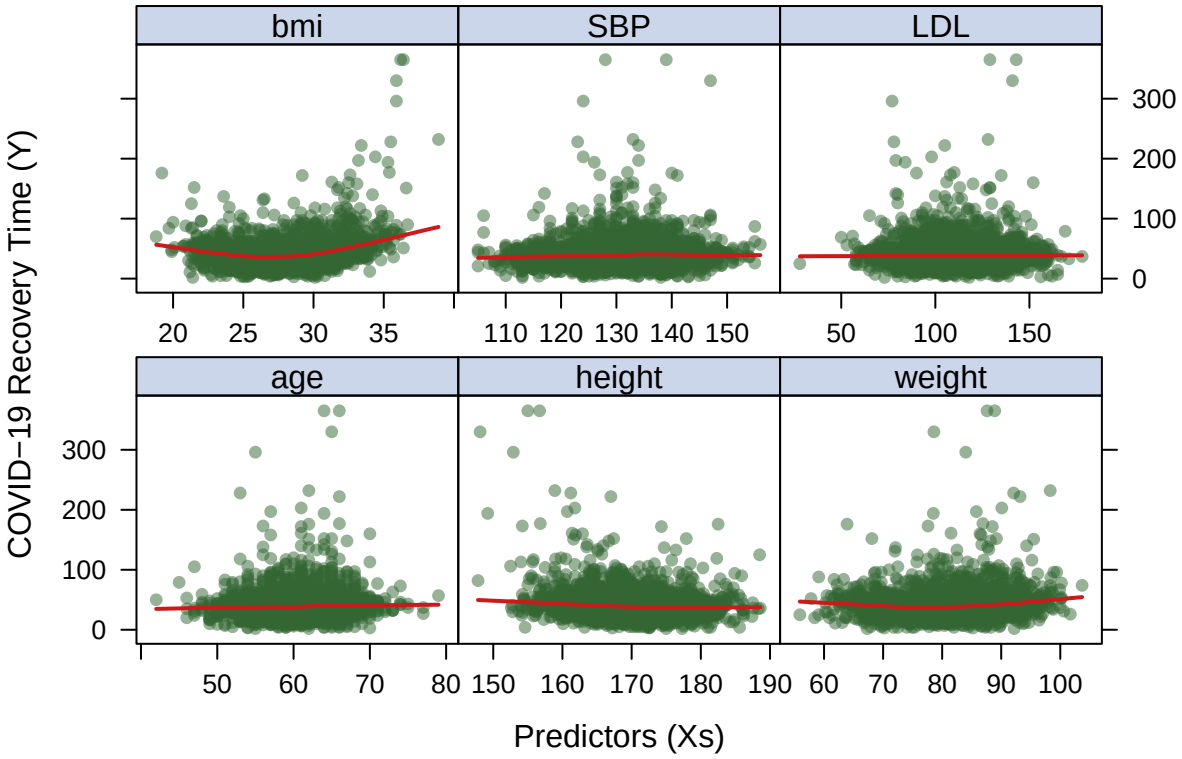
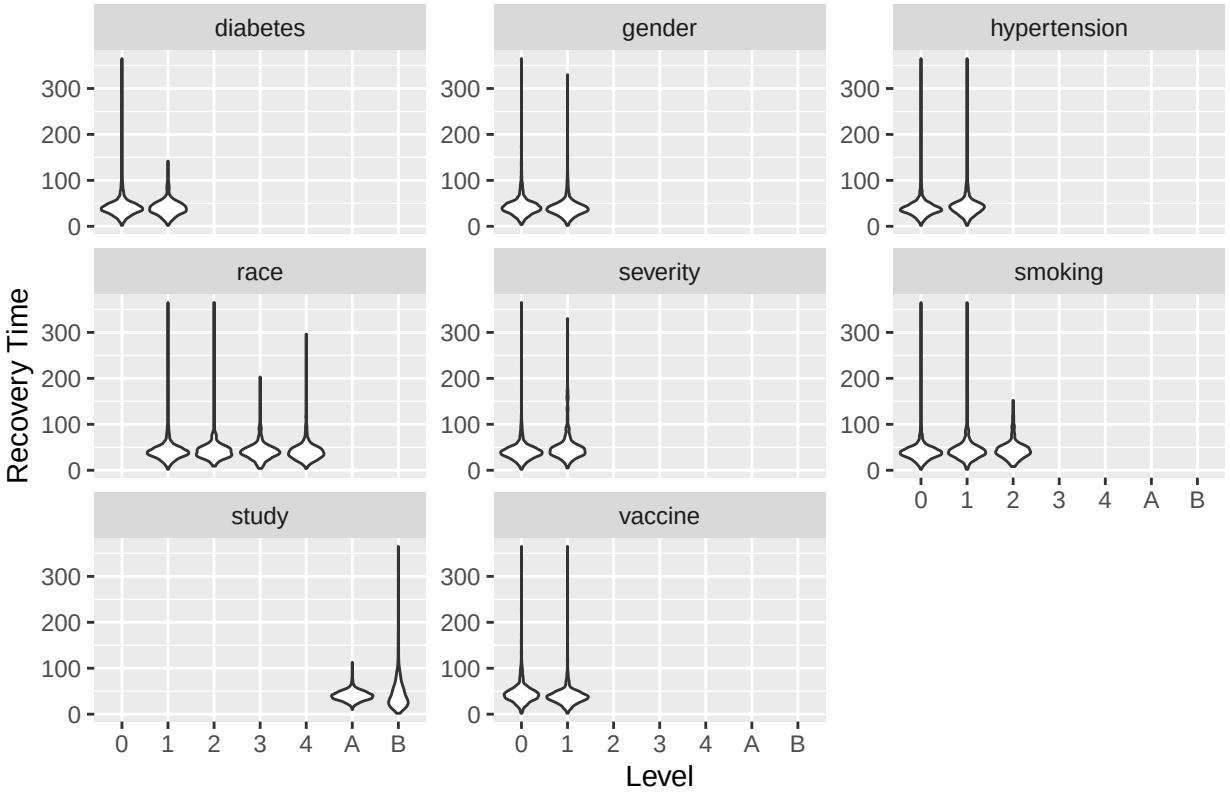


Figure 2. Violin Plots



Model training

We trained several different models in order to predict the time to recovery from COVID-19. We included all demographic and medical variables, including study because there might be important differences across study to account for. All models were trained on a training data set (70%), with 10-fold cross-validation. First, we used a simple linear model to predict recovery time. Linear models use least squares estimation and the assumptions of this model include: independent observations, homoscedasticity, and linearity. The common assumption of normally distributed errors is not needed. Next, we conducted a LASSO model, which adds a penalty term onto a normal linear model to shrink coefficients. We tuned this model using the "best" method, choosing the lowest test error, across a large grid of penalty values. Following this, we conducted a partial least squares (PLS) model. PLS models are a supervised dimension reduction procedure where the response variable is used to create new features that approximate the old features in addition to being related to the response. Next, we conducted an elastic net model, which combines the regularization process of a ridge regression penalty and the feature selection of a LASSO model penalty. We trained this model across a large grid of alpha and lambda values. Following this, we conducted a Multivariate Adaptive Regression Splines (MARS) model to consider non-linearity. This is an extension of linear models but makes no assumption about the relationship between the predictors and outcome, essentially creating a piece-wise linear model. We trained this model across grids of pruning parameter and degree parameter. Finally, we conducted a Generalized Additive Model (GAM) model which uses smooth functions to help model complex relationships between predictors and outcome. This can be described as a penalized generalized linear model. In order to determine the best prediction model, we tested the models on the testing data set and chose the model with the lowest cross-validated test error; the chosen error statistic was Root Mean Squared Error (RMSE).

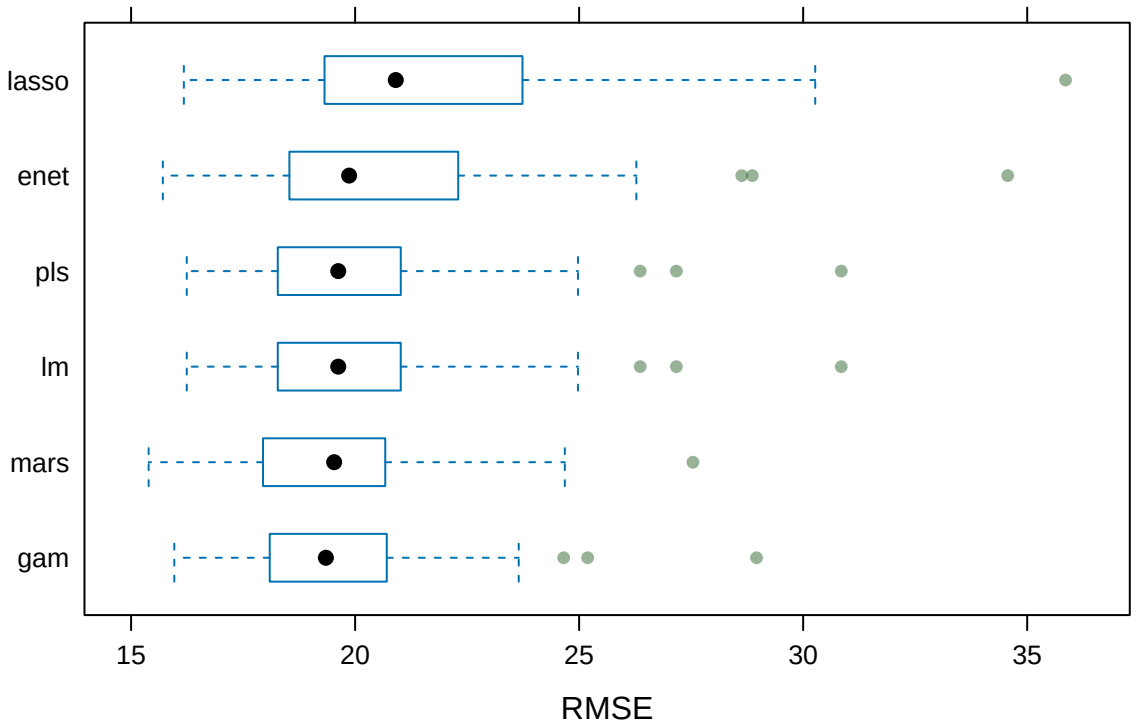
Table 2: Median Resampled RMSE's

lm~RMSE	lasso~RMSE	enet~RMSE	pls~RMSE	gam~RMSE	mars~RMSE
19.62327	20.90705	19.86654	19.62328	19.34727	19.53205

Results

In order to assess which model performed the best, comparing the RMSE was the best method. A model with a low RMSE indicates the best performing model(low prediction error), hence the opposite with a high RMSE.

Figure 3. Model Comparison Plot Using RMSE




```
## Call: earth(x=matrix[2102,18], y=c(31,47,40,34,3...), keepxy=TRUE, degree=2,
##             nprune=2)
##
##                     coefficients
## (Intercept)            40.14471
## studyB * h(bmi-31)     30.05710
##
## Selected 2 of 26 terms, and 2 of 18 predictors (nprune=2)
## Termination condition: Reached nk 37
## Importance: studyB, bmi, gender-unused, race1-unused, race2-unused, ...
## Number of terms at each degree of interaction: 1 0 1
## GCV 383.1087    RSS 802615    GRSq 0.3266962    RSq 0.3282976

##
## Family: gaussian
## Link function: identity
##
```

```
## Formula:
## .outcome ~ gender + race3 + race4 + smoking1 + smoking2 + hypertension +
##     diabetes + vaccine + severity + studyB + s(age) + s(SBP) +
##     s(LDL) + s(bmi) + s(height) + s(weight)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   44.1392     1.1752  37.560  < 2e-16 ***
## gender        -3.5459     0.8381  -4.231 2.43e-05 ***
## race3         -0.6689     1.0599  -0.631  0.52803
## race4         -1.0546     1.4686  -0.718  0.47276
## smoking1       2.6229     0.9504   2.760  0.00583 **
## smoking2       3.6774     1.3880   2.649  0.00813 **
## hypertension   1.9348     1.4111   1.371  0.17048
## diabetes      -2.0844     1.1518  -1.810  0.07050 .
## vaccine       -7.0273     0.8545  -8.224 3.43e-16 ***
## severity       9.3538     1.3413   6.974 4.13e-12 ***
## studyB         4.6053     0.8908   5.170 2.57e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df      F  p-value
## s(age)    1.000  1.000  1.677 0.195457
## s(SBP)    1.478  1.813  0.786 0.543422
## s(LDL)    1.219  1.408  3.834 0.027315 *
## s(bmi)    7.199  8.167 57.524  < 2e-16 ***
## s(height) 6.908  7.956  9.740  < 2e-16 ***
## s(weight) 1.670  2.194  6.925 0.000453 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.358   Deviance explained = 36.7%
## GCV = 370.41  Scale est. = 365.04    n = 2102
```

According to Figure 3, the linear model was the worst performing followed by PLS, Lasso, elastic net, GAM, and finally, the MARS model, which had the lowest median and mean RMSE(showing best performance). However, the final model for predicting time to recovery from COVID-19 was our GAM model. Although technically the MARS model "performed the best" when looking at the formula used, it should be noted that the MARS model only used 2 predictors, which would not be an accurate depiction of the recovery time with COVID-19. Therefore, when compared to the GAM model, which had a slightly higher mean and median RMSE, had to be selected as the final model for comprehensivly predicting time to recovery from COVID-19 within the study.

The final GAM model has recovery_time as the outcome with "White" (race = 0) as a reference category, with similar usage for "Never Smoker" refrencing smoking, and "Study A" referencing study predictor. When looking at the formula we notice s() around some of the variable names which will indicate that a smoothing function was applied on those variables. In addition any term with an * shows a statistically significant term at 5% level of significance. Taking the GAM model into consideration we see that all the predictors show a 36.7% of the deviance in COVID 19 recovery time. The RMSE (training error) of the GAM model was about 16.0, showing that on average the model's prediction on training data will deviate from the actual data around 16.0 units, meanwhile while using the test data we see that the RMSE was around 18.5 shows that on data that the model has never seen before is just a little bit worse than on the testing set.

Conclusions

The final GAM model showed several factors that were statistically significant in predicting the recovery time from COVID-19. It is seen that on average, having a history of former or current smoking, having hypertension, and experiencing severe COVID-19 infections were shown to have a longer predicted recovery time. In addition, with the inclusion of the study predictor, it is seen that being in study B was also associated with having a longer recovery time in comparison to those in study A. It is seen that being male and being vaccinated was shown to have a shorter recovery time. The model did not show any significant associations with diabetes or race predictors. Finally, BMI, height,weight, and age were also significantly associated with predicting the recovery time from COVID-19. This model and data can give us an insight into whhat predictors should be looked at closer when looking at COVID-19 recovery time.