P8106 Midterm Project: Predicting COVID-19 Recovery Time

Guadalupe Antonio Lopez, Gustavo Garcia-Franceschini, Derek Lamb

UNI's: GA2612, GEG2145, DRL2168

Introduction

This analysis combines three cohort studies regarding recovery time from COVID-19 illness. We have the individual's gender and race, along with other medical information. Among these, stand out their vaccination status and the study (A or B) they were a part of. With this information, we aim to fit a model that can both help us predict recovery time, and help us understand variables strongly associated with increased risk for long COVID-19 recovery times.

EDA

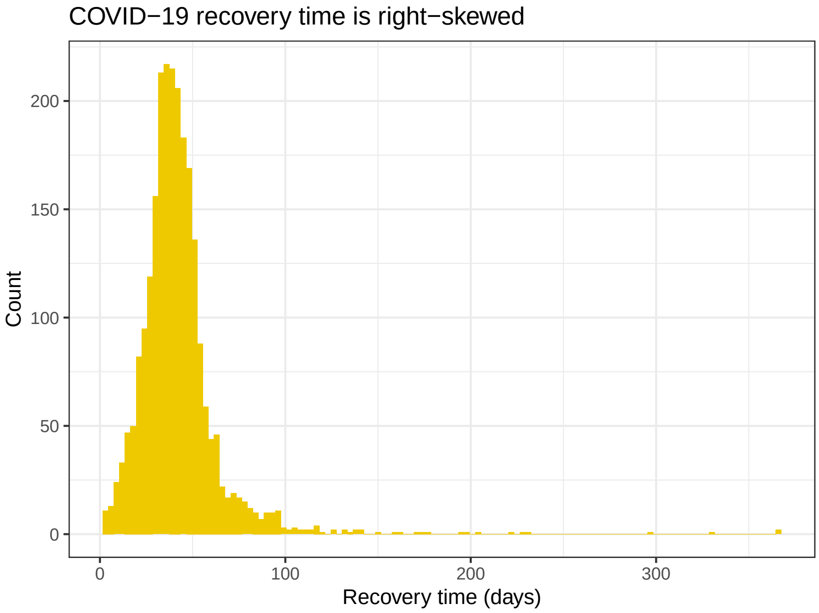
Table 1: Summary statistics

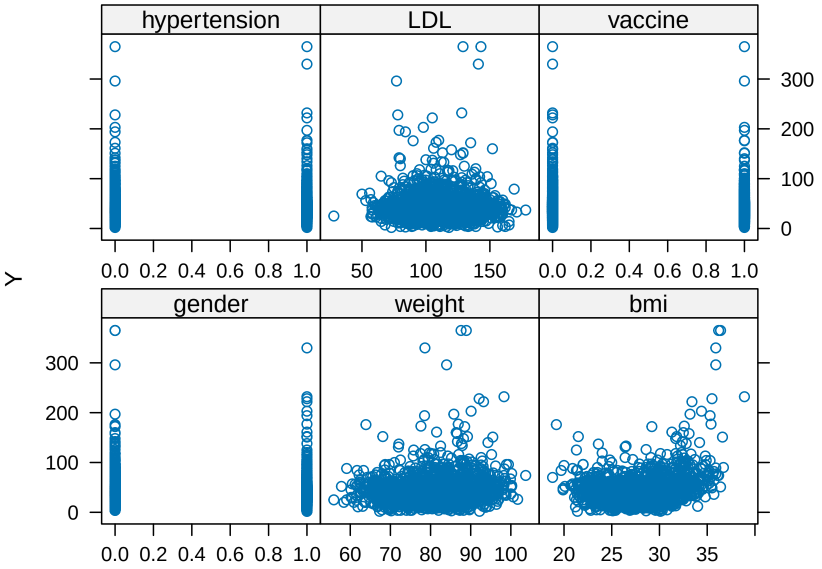
Variable	**A**, N = 2,000	**B**, N = 1,000
age	60.206 (4.531)	60.187 (4.380)
gender	964 (48%)	492 (49%)
race		
1	1,312 (66%)	655 (66%)
2	108 (5.4%)	50 (5.0%)
3	408 (20%)	196 (20%)
4	172 (8.6%)	99 (9.9%)
smoking		
0	1,225 (61%)	597 (60%)
1	557 (28%)	302 (30%)
2	218 (11%)	101 (10%)
height	169.858 (5.946)	169.993 (6.014)

Variable	**A**, N = 2,000	**B**, N = 1,000
weight	79.930 (7.128)	80.020 (7.156)
bmi	27.764 (2.765)	27.759 (2.848)
hypertension	1,002 (50%)	490 (49%)
diabetes	322 (16%)	141 (14%)
SBP	130.553 (8.023)	130.306 (7.877)
LDL	110.307 (19.756)	110.748 (19.765)
vaccine	1,203 (60%)	585 (59%)
severity	215 (11%)	106 (11%)
recovery_time	40.423 (11.175)	45.659 (36.622)

The table probably goes first.

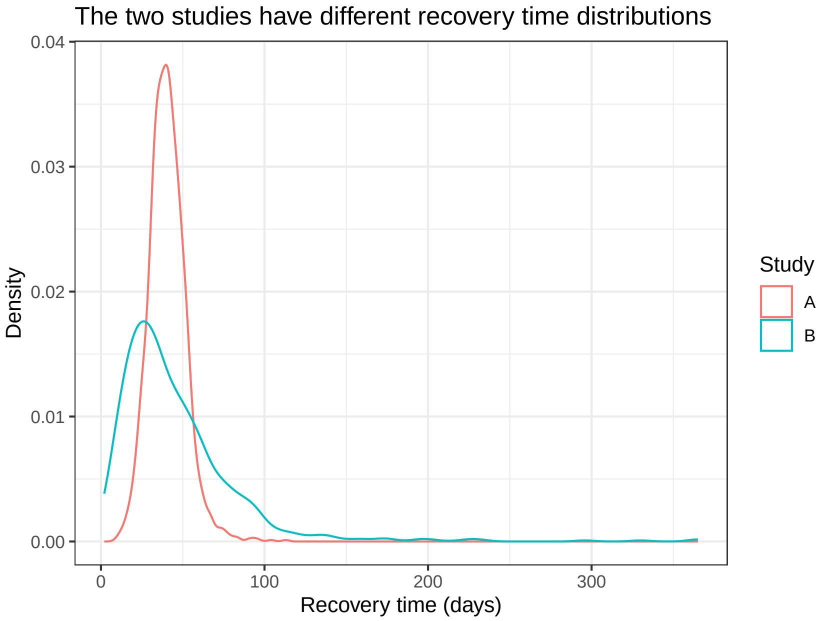
We found that the COVID-19 infection recovery time is heavily right-skewed: most of the individuals recovered at around 6 weeks, but there are individuals that only recovered from the infection after three months (or more) of being infected. This may mean that we'll need flexible models to capture the skewness of the response.





And also the scatterplots with continuous variables

When evaluating the distribution of recovery time, split into study groups, we find that its different for the two distributions. Study A has a later peak, while Study B has a heavier tail, corresponding to more individuals in that study experiencing longer recovery time. This is an early indication that study group might be an important variable when predicting recovery time.



We also examined the pairwise correlations of the variables, and the correlations of the covariates with the recovery time. There were two clusters of strong correlation (height, weight, and BMI; hypertension and SBP), but these covariates were functionally dependent upon each other. There were no other strong correlations between variables, and no one covariate had an exceptional correlation to recovery time.

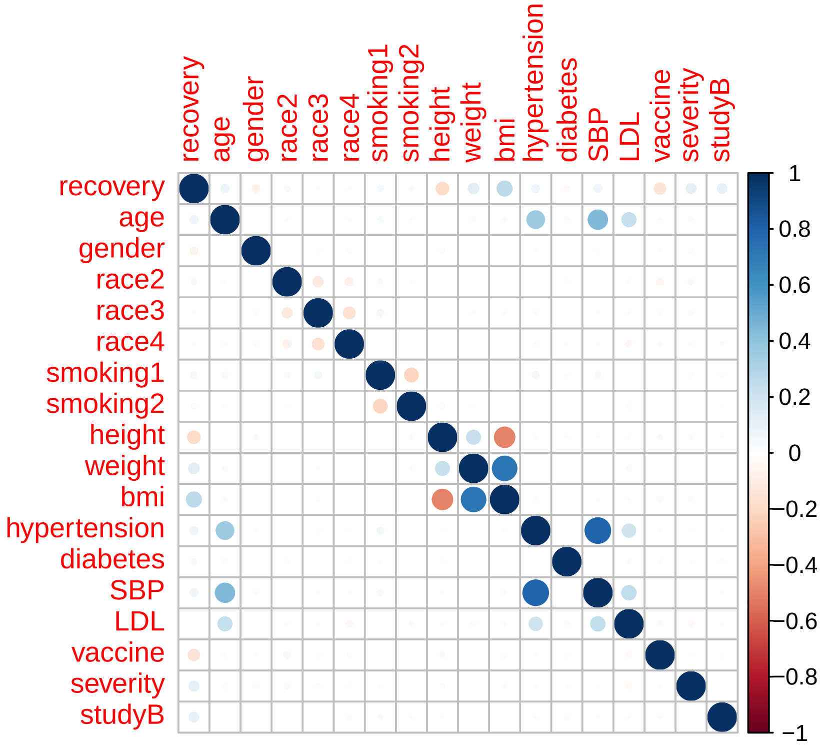


Figure M. Correlogram of study variables.

Model Training

To train the models, we partitioned the data into training and testing sets, with 80% of the data (2400 subjects) being assigned to the training set, and the remaining 20% (600 subjects) being assigned to the test set.

To predict COVID-19 recovery time, we modeled the data using four approaches – two linear and two non-linear. For the linear approaches, we selected elastic net and partial least squares regression. For the nonlinear approaches, we selected multivariate adaptive regression splines (MARS) and a general additive model (GAM).

All models were fitted using the `train()` function in the `caret` package. Although some inputs varied by model, the common inputs were formula or model matrix and response vector, data, method, tuning parameters grid, and a 10-fold cross validation method.

Note that all models were fit using a seed of 1.

Elastic Net Model

To fit the elastic net model, we used the model formula with `recovery_time` as the response and all other variables in our training data set to be predictors. Given that we fit an elastic net model, the method specified was `glmnet`, with tuning parameter alpha to be sequenced between 0 and 1 (with length 21) and lambda to be exponentially sequenced between -6 and 1 (with length 100). We settled on this lambda region after fitting the model various times with different regions. We started with a large region (-4 to 4), but realized that our preferred lambda value was close to our lower boundary. Thus, we continued to expand our region until we settled on -6 to 1.

After fitting the elastic net model, the final model based on the optimal lambda contained 17 predictors and an intercept – no predictors were shrunk to zero. The values for each predictor represent the estimated effect of each predictor on recovery time. Based on our output, age, height, bmi, and systolic blood pressure had positive coefficients. This suggests that an increase in a given predictor is associated with an increase in recovery time. There were also positive coefficients for categorical variables race (asian), former and current smoking status, hypertension, systolic blood pressure, severity, and study B. These positive coefficients indicate the difference in the outcome compared to their reference level.

Partial Least Squares (PLS) Regression Model

To fit the PLS model, we used the training model matrix, based on our training data, and training response vector. Given that we fit a PLS model, the method specified was `pls`, with number of components ranging between 1 and 15. This range is based on the number of variables in our training model matrix. Additionally, the predictor data was centered and scaled – specified in the `preProcess` input.

After fitting the PLS model, the final model is based on 13 components.

Multivariate Adaptive Regression Splines Model

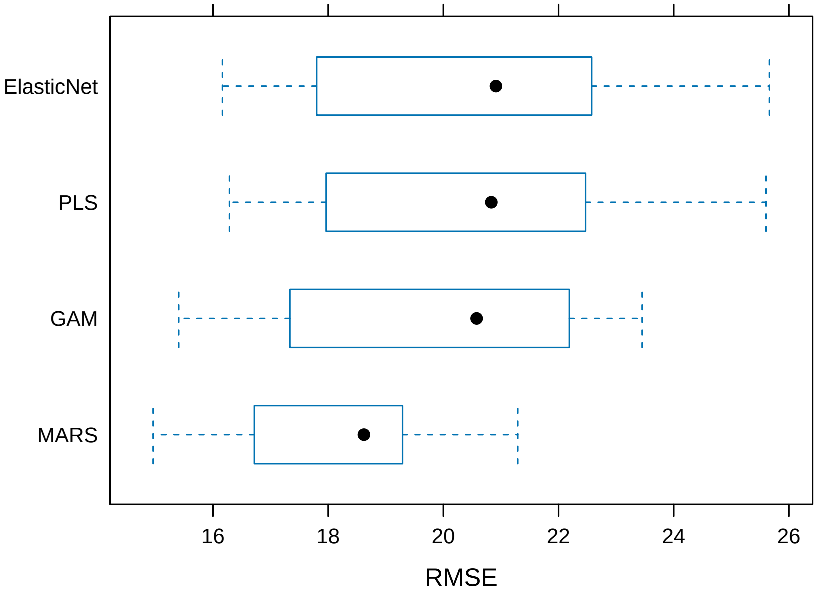
To fit the MARS model, we used the training model matrix and training response vector. Given that we fit a MARS model, the method specified was `earth`, with degrees ranging between 1 and 3 and the maximum number of terms in the pruned model to range between 2 and 14.

General Additive Model

To fit the GAM, we used the training model matrix and training response vector. Given that we fit a GAM, the method specified was `gam` and a 10-fold cross validation.

Results

We then compared the four models that we fit by resampling on the training set. In the figure below, we constructed boxplots to compare the four different models by their resampled RMSE. The two linear models and GAM perform about the same, though GAM was a bit better on average. MARS noticeably outperformed the other models.



Once we decided on MARS as a final model, we calculated the test error using the 20% partition of the initial data set.

The test RMSE for the MARS model is 17.18.

Conclusion

In this project, our goal was to use statistical learning to gain insight into the recovery process of people infected with COVID-19. We fit four models to predict COVID-19 recovery time from a set of 14 covariates, two linear and two nonlinear. Our linear models achieved similar performance in predicting recovery time, but they were outdone by the nonlinear methods, MARS in particular. This improvement in prediction is due to the greater flexibility of the nonlinear methods, but comes at a trade-off of the interpretability of such models. However, as our goal was to develop the best model for predicting COVID-19 recovery time, we are comfortable giving up some of this interpretability, and recommending the MARS model developed above for this task.