P8106 Midterm Project: Predicting COVID-19 Recovery Time

Guadalupe Antonio Lopez, Gustavo Garcia-Franceschini, Derek Lamb

UNI's: GA2612, GEG2145, DRL2168

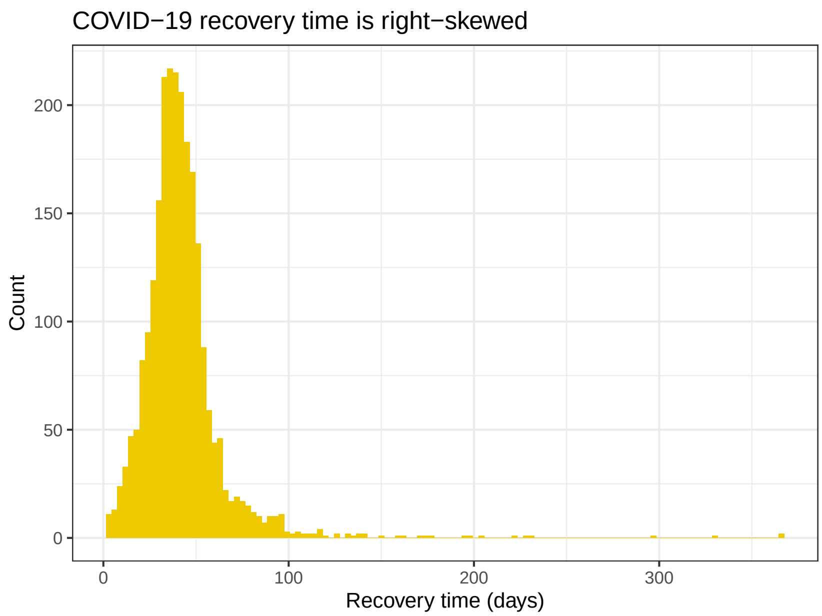
Introduction

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.

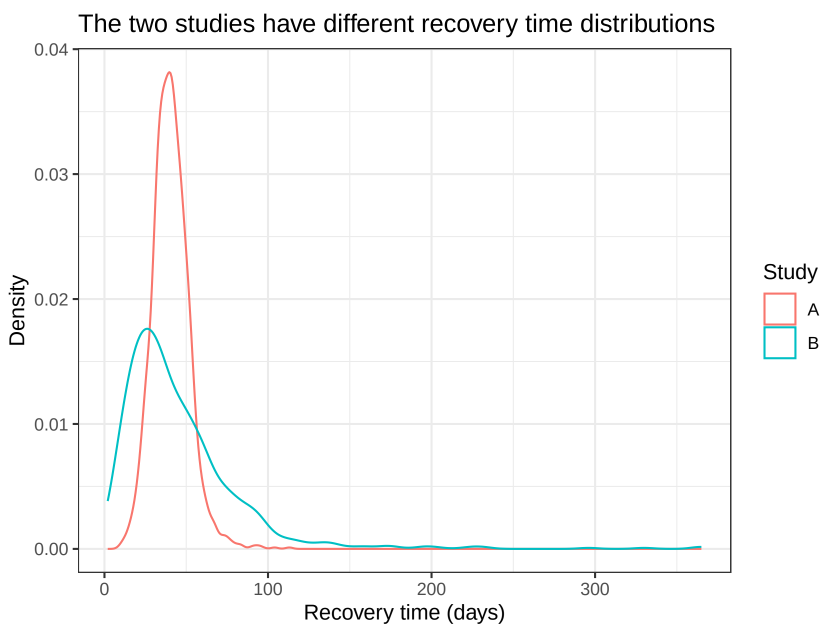
EDA

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.



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.

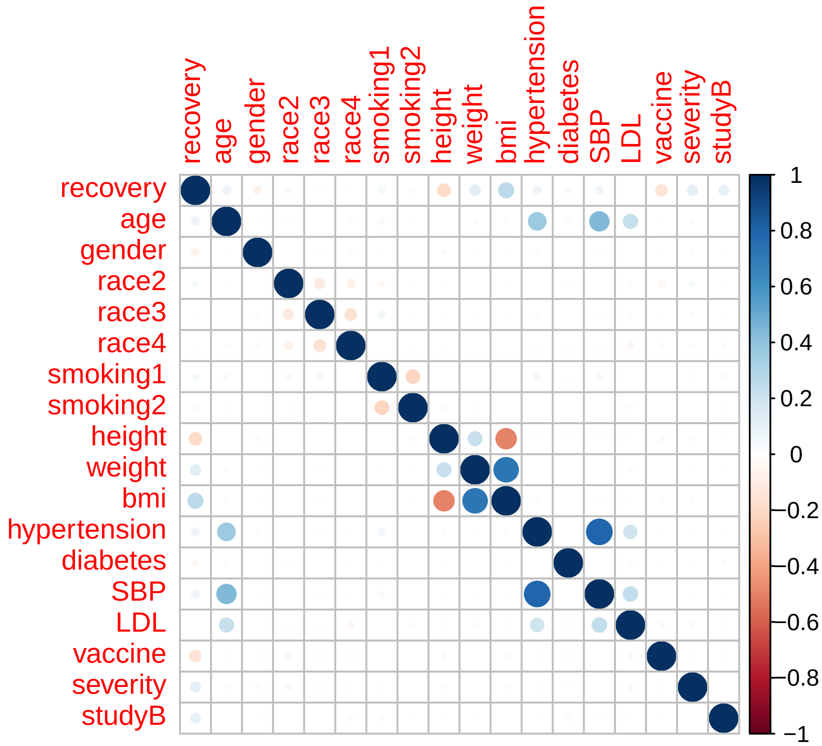


Figure M. Correlogram of study variables.

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 cross validation method.

the `caret` package

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. Lastly, we used a 10-fold cross validation method.

Partial Least Squares Regression Model

Multivariate Adaptive Regression Splines Model

General Additive Model

Results

Remember to talk about `study` as a variable.

Conclusion