Final Report

Chen Liang (cl4469), Xinyi Shang (xs2529), Yuki Joyama (yj2803)

Exploratory Analysis and Data Visualization

The information of COVID-19 recovery time and other variables (id, gender, race, smoking history, height, weight, body mass index (BMI), history of hypertension and diabetes, systolic blood pressure (SBP), LDL cholesterol (LDL), vaccination status at the time of infection) is collected from two existing cohort studies. Baseline characteristics are presented in Table 1, showing that almost all characteristics are similar between the two study groups, except for COVID-19 recovery time.

Table 1: Baseline Characteristics

Characteristic	**A**, N = 2,000[1]	**B**, N = 1,000[1]
Age	60.2 / 60.0 (4.5)	60.2 / 60.0 (4.4)
Gender		
Female	1,036 (52%)	508 (51%)
Male	964 (48%)	492 (49%)
Race		
Asian	108 (5.4%)	50 (5.0%)
Black	408 (20%)	196 (20%)
Hispanic	172 (8.6%)	99 (9.9%)
White	1,312 (66%)	655 (66%)
Smoking		
Current smoker	218 (11%)	101 (10%)
Former smoker	557 (28%)	302 (30%)
Never smoked	1,225 (61%)	597 (60%)
Height	169.9 / 169.9 (5.9)	170.0 / 170.0 (6.0)
Weight	79.9 / 79.6 (7.1)	80.0 / 80.0 (7.2)
BMI	27.8 / 27.7 (2.8)	27.8 / 27.6 (2.8)
Hypertension		
Hypertension	1,002 (50%)	490 (49%)
No hypertension	998 (50%)	510 (51%)
Diabetes		
Diabetes	322 (16%)	141 (14%)
No diabetes	1,678 (84%)	859 (86%)
SBP	130.6 / 131.0 (8.0)	130.3 / 130.0 (7.9)
LDL	110.3 / 110.0 (19.8)	110.7 / 110.0 (19.8)
Vaccine		
Not vaccinated	797 (40%)	415 (42%)
Vaccinated	1,203 (60%)	585 (59%)
Severity		
Not severe	1,785 (89%)	894 (89%)
Severe	215 (11%)	106 (11%)
Recovery time	40.4 / 40.0 (11.2)	45.7 / 37.0 (36.6)

[1]Mean / Median (SD); n (%)

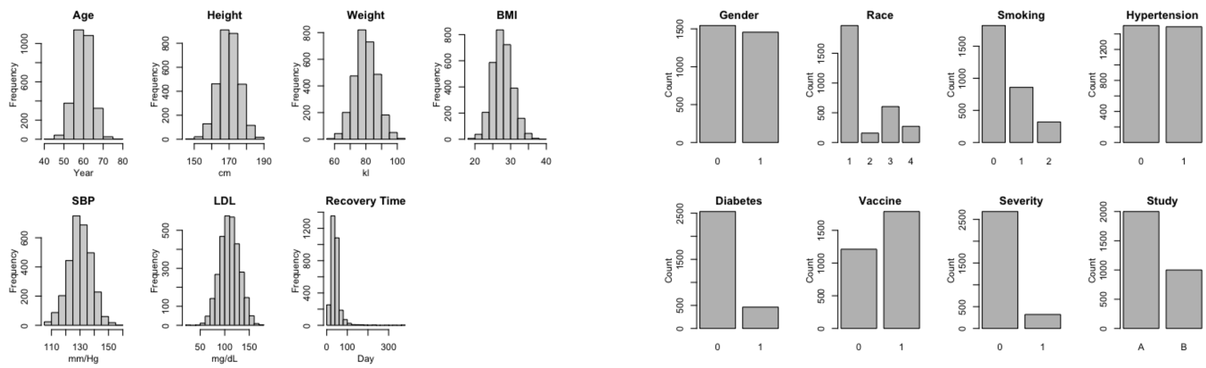

Figure 1: (a) Histogram of Continuous Variables (b) Bar Plot of Categorical Variables

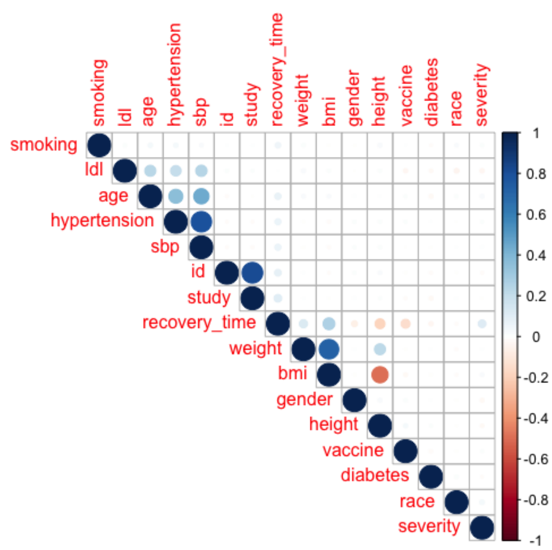

Figure 2: Correlation Plot of Key Variables

Model Training

Model selection

After cleanning and preprocessing the dataset, we partitioned it into an 80-20 training-test split. Then, We employed a variety of regression models to predict the time to recovery from COVID-19. The models include LM, Ridge, PLS, PCR, Lasso, Elastic Net, MARS, and GAM. Each model was trained using the train function from the caret package on the training data with 10-fold cross-validation.

LM assumes linearity, homoscedasticity, no multicollinearity, and normal distribution of residuals. Lasso has the same assumption as LM. Elastic Net is a linear regression model that combines Lasso and Ridge regularization penalties, it assumes that a balance of Lasso and Ridge penalties will produce a better model.Ridge regression sssumes multicollinearity in the data and attempts to mitigate its effects by introducing a penalty term. Principal Component Regression assumes that the principal components capture most of the variance in the predictors and that these components have a linear relationship with the outcome. Partial Least Squares assumes that the new components created from the predictors will have a linear relationship with the outcome. Gam assumes that the data can be better modeled by allowing non-linear relationships between predictors and the response. MARS is a non-parametric regression method that builds flexible models by fitting piecewise linear regressions, it asssumes that the relationships between the predictors and the outcome can be modeled with splines, which are piecewise polynomials joined at knots.

Selection of Tuning Parameters

In predictive modeling, tuning parameters are crucial as they can significantly affect the model's performance. For our final analysis, we employed a meticulous process to identify the optimal tuning parameters for each model, aiming to strike a delicate balance between bias and variance, ultimately to improve prediction accuracy.

To select the best tuning parameters, Initially, we used a wide range and search pattern, we created a grid of potential models with different degrees and numbers of terms to prune, then used 10-fold cross-validation to select the optimal combination. After identifying promising ranges for the selected parameters where show the best cross-validation performance, we then searched parameter patterns within a narrower range and with more density by decreasing the step within each parameter sequence. For example, with the MARS model, we started with a relatively large number of maximum terms in the initial grid search to capture potential model complexity. We then narrowed down the search space for the degree of interaction and number of terms based on cross-validation performance. The best tuning parameters given by the cross-validation is: `nprune = 7`, `degree = 4`.

Model Comparison

After fitting all the models, we used the resamples function to compare their performance based on RMSE. The performance of all models was assessed through 10-fold cross-validation on the training set. Repeated cross-validation was not employed to avoid excessive computational cost. The results of the cross-validation are presented below

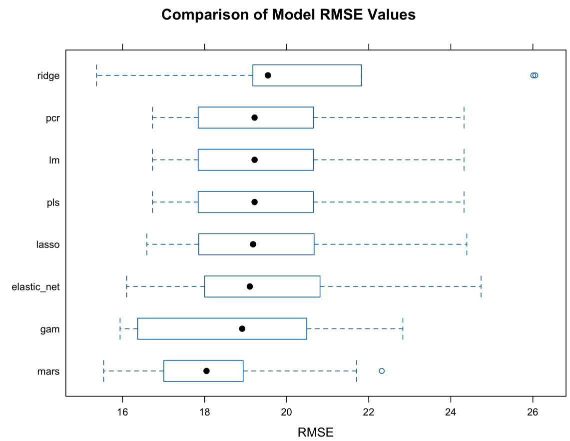

Figure 3: Comparison of Model RMSE Values

This box plot illustrates the distribution of Root Mean Square Error (RMSE) values across different predictive models used to estimate the time to recovery from COVID-19. The MARS model has the lowest median RMSE, suggesting that it is the best performing model in terms of prediction accuracy on the validation sets used during cross-validation. Moreover, there is a clear distinction between the group of models with the lowest RMSE values (MARS, GAM, Elastic Net) and the other models, indicating that incorporating non-linearity and regularization seems beneficial for this dataset.

In summary, based on this plot, MARS offers the best balance between accuracy and consistency for the given dataset. However, it's important to consider other factors such as the complexity of the model, interpretability, and computational efficiency when making a final selection for practical application.

Why choose MARS

Multivariate adaptive regression splines is an effective technique for simplifying models and constructing them. It is a nonparametric, multivariate regression method that can estimate complex nonlinear relations by a series of spline functions of the predictor variables, which makes it flexible in modeling the shape of the recovery time distribution in the COVID-19 dataset.

One of the strengths of MARS is its flexibility. It can handle various types of predictor variables, from continuous to categorical, and is adept at managing a high number of them. Its nonparametric nature is especially beneficial as it operates without presumptions about how the predictor variables are distributed.

Results

Our final MARS model is as follows:

$\hat{y}$ = 22.435 + 3.574 × h(30.3 - bmi) + 9.783 × h(bmi - 30.3) * studyB + -6.264 × vaccine + 2.991 × h(164 - height) * h(bmi - 30.3) * studyB + 4.898 × h(bmi - 25.7) + -2.64 × h(87.6 - weight) * h(bmi - 30.3) * studyB, where $h(.)$ is a hinge function

Table 2: Summary of the MARS model

Equation	Coefficients
(Intercept)	22.435204
vaccine	-6.264022
h(bmi-25.7)	4.898496
h(30.3-bmi)	3.574364
h(bmi-30.3) * studyB	9.782606
h(164-height) * h(bmi-30.3) * studyB	2.990502
h(87.6-weight) * h(bmi-30.3) * studyB	-2.640353

The summary of the final MARS model is shown in Table 2. Vaccinated people have 6.264 shorter recovery time (days) compared to non-vaccinated ones, holding other variables constant. The model shows that BMI has two knots (25.7 and 30.3). This can be expressed as follows:

$$\text{Recovery time} = \begin{cases} 22.435 & \text{for BMI} \leq 25.7 \\ 22.435 + 4.898\ (\text{BMI - }25.7) & \text{for } 25.7 \leq \text{BMI} \leq 30.3 \\ 22.435 + 3.574\ (30.3\text{ - BMI}) & \text{for } 30.3 \leq \text{BMI} \end{cases}$$

All else being equal, if BMI is in the range (25.7, 30.3), the recovery time increases by 4.898 days for every unit increase in BMI; for those with BMI larger than 30.3, the recovery time increases by 3.574 days for every unit increase in BMI. The model also tells us that there are interactions between h(bmi - 30.3) and studyB; h(164 - height), h(bmi - 30.3) and studyB; h(87.6 - weight), h(bmi - 30.3) and studyB. We will discuss this in the later section ("Additional Considerations"). Given the results, we can infer that the followings are the important risk factors for longer recovery time:

- No history of vaccination
- BMI over 25.7
- BMI over 30.3 in Study B
- Height under 164 cm and BMI over 30.3 in Study B

Figure 4 illustrates that study B, BMI, height, weight, and vaccination status have the non-zero importance value in the final model.

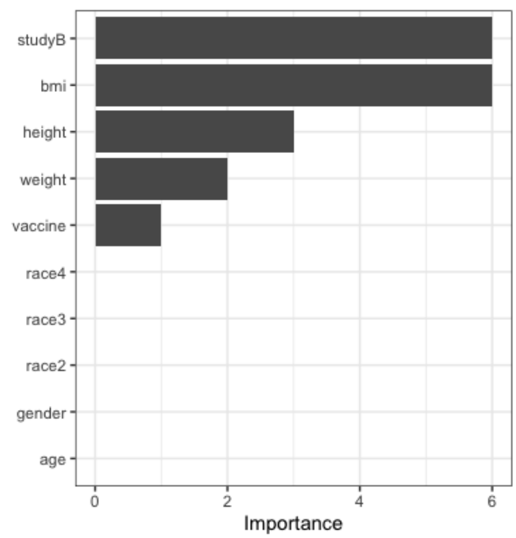

Figure 4: Variance Importance Plot

Conclusions

Our analysis using the Multivariate Adaptive Regression Splines (MARS) model has provided significant insights into factors influencing COVID-19 recovery times. Key findings include the substantial impact of vaccination, which significantly reduces recovery time, highlighting the critical role of immunization in managing COVID-19 outcomes. The model also reveals the nuanced effects of body mass index (BMI) on recovery, with distinct thresholds where recovery time increases, underscoring the importance of metabolic health in the COVID-19 recovery process.

Furthermore, the interaction between BMI, study variables, and demographic factors such as height and weight, suggests a complex relationship affecting recovery time. These interactions emphasize the need for a tailored approach to treatment, considering the multifaceted nature of individual health profiles.

In conclusion, our model analysis underscores the necessity of vaccination and the management of metabolic health in improving COVID-19 recovery times. It also highlights the importance of personalized healthcare strategies that account for the interplay of various factors affecting individual recovery trajectories. Future research should focus on understanding the mechanisms behind these associations, to develop more effective, targeted interventions for COVID-19 recovery.

Additional Considerations

In our work with the Multivariate Adaptive Regression Splines (MARS) model for predicting COVID-19 recovery times, we chose to include "study" as a predictor. This decision was based on recognizing that factors like socioeconomic status, geography, and demographics can greatly affect recovery outcomes. These factors vary from one study to another but weren't directly included in our datasets. Our analysis showed that the "study" variable significantly interacts with other variables, especially BMI, highlighting that the influence of certain predictors on recovery time can change depending on the study context. This finding underlines the importance of considering the "study" variable to accurately capture the diverse experiences of COVID-19 recovery.

To deepen our understanding of these effects, a stratified analysis is suggested as a future step. Such an analysis would allow us to dissect how recovery dynamics change across distinct study conditions, providing a better understanding of the factors influencing recovery times. By segmenting data according to specific study characteristics, we can tailor our model to more precisely predict the COVID-19 recovery time.