P8106 DS2 Midterm Report: Predictive Modeling of Covid-19 Recovery Time

Tianyuan Deng, Qianying Wu, Ze Li

1. Introduction

The spread of COVID-19 across the globe has prompted an urgent need for research into the disease's impacts and recovery patterns. Our study using a dataset of 3000 participants from three cohort studies aims to shed light on the recovery time from COVID-19. By analyzing a variety of factors, including personal characteristics prior to the pandemic and detailed health and infection data, we seek to develop a predictive model for recovery times. This model could significantly contribute to patient management strategies, enhance public health responses, and inform future pandemic preparedness. The dataset includes variables such as gender, race, smoking status, body mass index (BMI), pre-existing health conditions, and COVID-19 infection details, offering a comprehensive overview for analysis.

2. Exploratory Data Analysis

Our initial steps involved preparing the dataset for analysis. We converted numerical codes to categorical labels for variables such as gender, race, smoking status, hypertension, diabetes, vaccination status, severity of the COVID-19 infection, and study group. This process facilitated a more intuitive analysis and visualization. A thorough examination for missing values ensured the integrity of our dataset, enabling accurate and reliable insights.

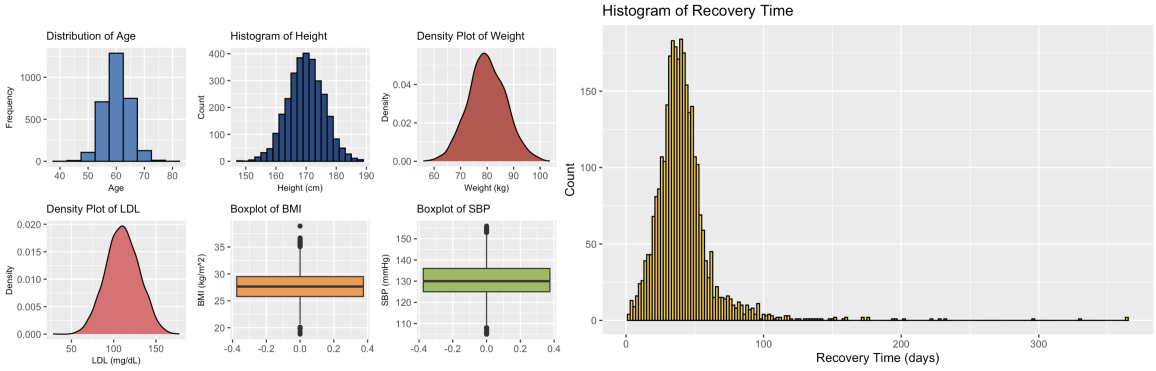
a. Univariate Analysis

We first used univariate analysis to explore individual variables:

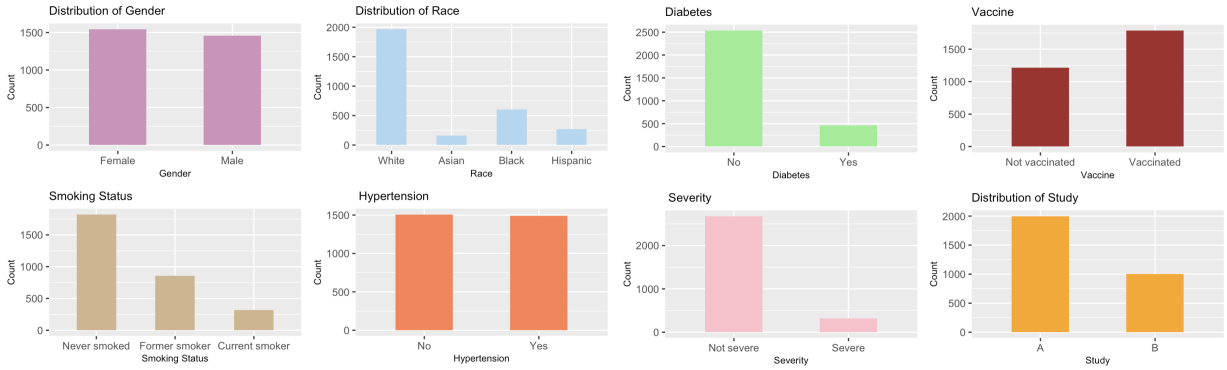
- Age, height, weight, and recovery time were analyzed using histograms, revealing the distribution patterns among participants. The histograms for height and recovery time, in particular, allowed us to understand the physical characteristics of our cohort and the general trend in recovery times.

- Density plots for weight and LDL cholesterol offered insights into the spread and central tendency of these measurements, highlighting potential risk factors associated with recovery.

- Box plots for BMI and systolic blood pressure (SBP) illustrated the variability within our cohort, identifying outliers and the range of these important health indicators.

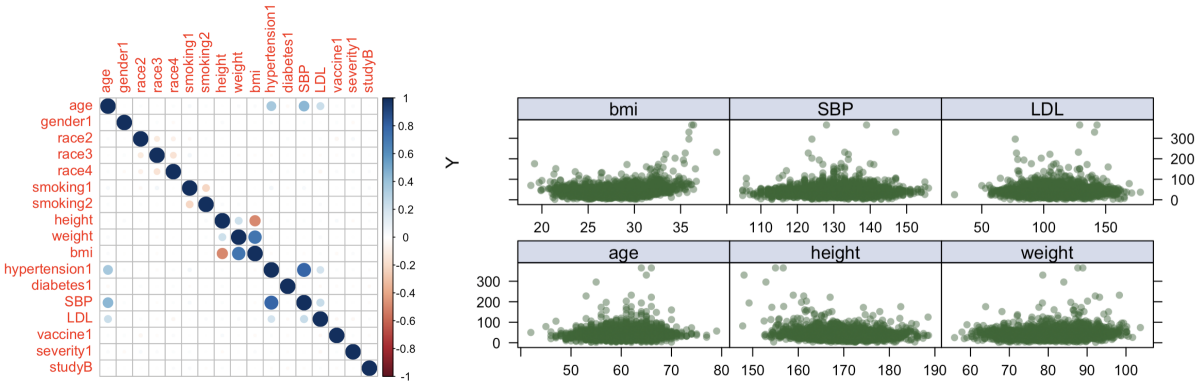


We also used bar plots to provide a visual summary of categorical variables, including gender, race, smoking status, hypertension, diabetes, vaccination status, severity of COVID-19 infection, and study group participation.



B. Bivariate Visualization

Our bivariate analysis focused on exploring relationships between recovery time and other variables. Using scatter plots and correlation matrices, we uncovered intriguing patterns and associations. The correlation analysis, in particular, revealed significant relationships, offering the correlation relationship between different covariates variables. For example, SBP and Hypertension seem to be associated with each other.



3. Model Training

To develop a robust predictive model for COVID-19 recovery time, we employed a multi-model approach. We chose to compare the performance of eight different regression techniques: Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, PLS, PCR, MARS, and GAM. Each of these models helps us to understand the relationship between the predictors and the recovery time. We used Root Mean Square Error (RMSE) to measure the performance of each regression model and compare the performance between different models.

a. Data Preparation

Prior to modeling, we ensured that categorical variables were correctly encoded as factors. To validate our models, we divided the dataset using an 80-20 split for training and testing, respectively, ensuring that we have a representative sample for model validation. The `initial_split`

function was used for this purpose, with a seed set of 7890 to ensure reproducibility across all team members' analyses.

For each model, we used 10-fold cross-validation to evaluate model performance and tune the hyperparameters. The `trainControl` function from the `caret` package was used to set this up, and the `train` function was used to perform the actual model fitting.

1. **Linear Regression (LM)**

Linear Regression assumes a linear relationship between the predictors and the response variable. It's the simplest model that provides a baseline for performance. We used it to estimate the initial relationships between survival time and other covariates without any regularization. The mean RMSE for the linear model was 20.64531.

Since our dataset has multiple predictors, there is a risk of overfitting. To address this, we employed regularization techniques: Ridge, Lasso, and Elastic Net. These methods incorporate penalties into the loss function, which helps reduce overfitting and perform feature selection.

2. **Ridge Regression**

We used the glmnet method in the Caret package to do the ridge regression. We tuned the models by experimenting with various values of lambda (the regularization penalty) from exponential value of 0 to 6 with length 100, and setting alpha = 0. The parameter of best tune is when lambda = 1. The mean RMSE from the Ridge regression model was 21.88024.

3. **Lasso Regression**

We chose to do Lasso regression because it can set some coefficients to zero, effectively performing feature selection. We used the glmnet method in the Caret package to do the Lasso regression. We tuned the models by experimenting with various values of lambda from exponential value of -3 to 5 with length 100, and setting alpha = 1. The parameter of best tune is when lambda = 0.0101. The mean RMSE from the Lasso regression model was 20.63304.

4. **Elastic Net**

We used Elastic Net as a combination of Ridge and Lasso, which balances between feature selection and coefficient shrinkage. We used the glmnet method in the Caret package to do the Elastic Net. We tuned the models by experimenting with various values of lambda from exponential value of -3 to 5 with length 100, and setting alpha ranging from 0 to 1 with the length of 21. The parameter of best tune is when alpha = 0.8 and lambda = 0.00674. The mean RMSE from the Elastic Net model was 20.63289.

5. **Partial Least Squares (PLS)**

We have a large number of predictor variables, so partial least squares can handle multicollinearity between variables well since it creates orthogonal components that summarize the predictors. It also reduces the dimensionality of the data, which simplifies the model. We used the pls method in the Caret package. The optimal number of components is 10 based on the

cross-validation results. This point represents the best trade-off between model complexity (number of components) and predictive accuracy (RMSE).

6. PCR

We utilized Principal Components Regression (PCR) to address multicollinearity in our dataset with numerous predictors. By implementing the 'pcr' function within the Caret package and conducting a cross-validation approach, we tested a range of 1 to 18 components and determined that the optimal number of principal components was 17. This choice minimized the cross-validation root mean square error (RMSE), signifying a harmonious trade-off between model simplicity and predictive accuracy.

7. Multivariate Adaptive Regression Splines (MARS)

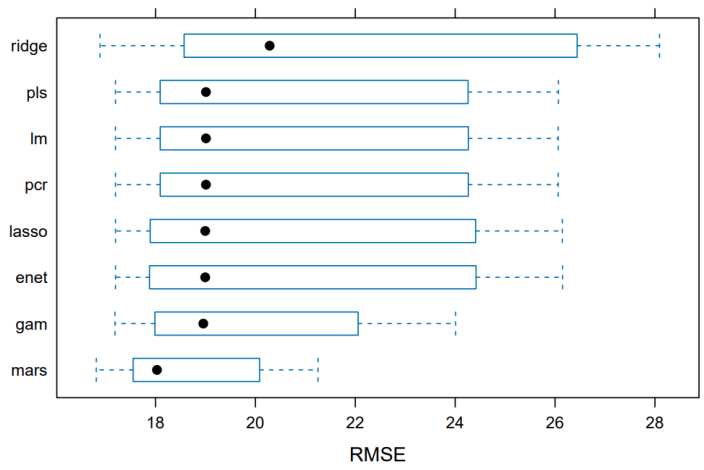
We assume the relationship between predictors and response is non-linear. Without making strong parametric assumptions, it provides a balance between interpretability and model complexity. We used the earth method in the Caret package. We set parameters for a model in three-way interactions and the model training process should try retaining anywhere from 2 to 25 of these functions after the pruning process. The best model uses 20 basis functions (splines) and has the interactions of degree 2.

8. GAM

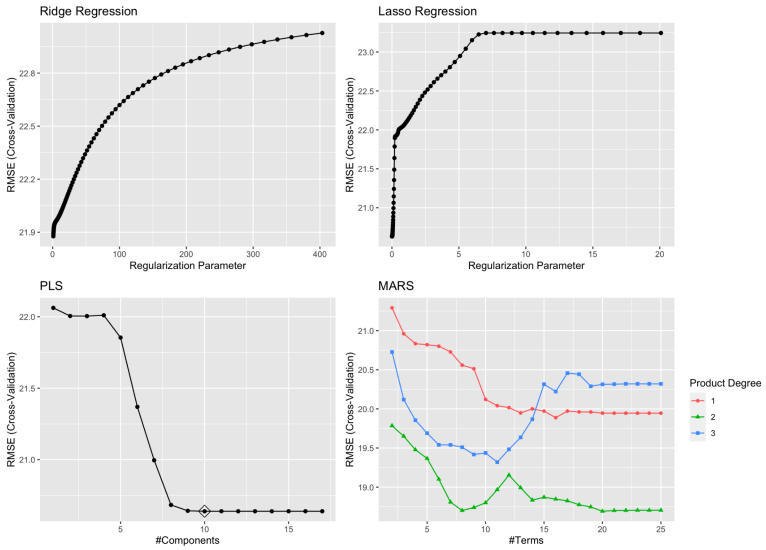
Employing a Generalized Additive Model (GAM) allowed us to capture complex non-linear relationships between predictors and the response variable. By using the 'gam' function within the Caret package, we selected the best model through cross-validation, optimizing the GCV.Cp method. GAM's flexibility in allowing both linear and non-linear terms gave us a nuanced understanding of the data, evident from the model's GCV score of 368.0526. This adaptability is one of GAM's key strengths, as it can uncover underlying patterns in the data that standard linear models may miss, potentially leading to more precise predictions.

4. Results

Upon examining the boxplot, the Multivariate Adaptive Regression Splines (MARS) model is distinguished by its notably lower median RMSE, as well as a relatively tight interquartile range. The compactness of the MARS boxplot signifies consistent performance across different folds of cross-validation, highlighting the model's robustness and reliability. Moreover, the absence of extended whiskers or outliers in the MARS plot further reinforces the stability of the model.



The MARS plot illustrates how the number of terms and the complexity of interactions within the model influence the predictive accuracy, as measured by RMSE. Degree 1 interactions show a steady decrease in RMSE with more terms, stabilizing quickly, suggesting that main effects alone offer limited predictive improvements. Degree 2 interactions provide a lower RMSE, indicating that incorporating pairwise interactions captures more complex relationships effectively. However, degree 3 interactions, after an initial drop, lead to an increase in RMSE with additional terms, implying overfitting. The optimal MARS model, therefore, seems to include a moderate number of terms with second-degree interactions, balancing the trade-off between complexity and accuracy.



The MARS approach has effectively identified non-linear relationships and pivotal interactions, which are critical in predicting the recovery time from COVID-19. The coefficients from the MARS model reveal a complex interplay of variables impacting recovery time from COVID-19. Hinge functions like **h(bmi-31)** suggest that BMI values above 31 are associated with a different recovery trajectory than values below this point. Interactions such as **h(bmi-31) * studyB** imply that the effect of BMI on recovery is modified by the study group. Significant coefficients for terms involving vaccine and gender indicate their influence on recovery time. The negative

coefficient for **race4 * h(bmi-34)** suggests that being in race group 4 and having a BMI over 34 is associated with longer recovery times. Interaction terms like **h(bmi-22) * hypertension** indicates that the impact of BMI below 22 on recovery time changes in the presence of hypertension.

The MARS model exhibits a solid performance in predicting the recovery time from COVID-19, with a test RMSE of 16.52678, which suggests that the model's predictions are, on average, within approximately 16.5 days of the actual recovery times. Additionally, the model achieves a reasonable Mean Absolute Error (MAE) and possesses a relatively high R-squared value among the models evaluated, indicating that it captures a significant portion of the variability in the recovery times.

5. Conclusion

Our analysis began by considering a suite of linear models, including LASSO, Ridge, and Elastic Net, to predict the recovery time based on 13 explanatory variables. The identifier variable id was appropriately excluded from the analysis as it provides no predictive power. Despite the inclusion of regularization techniques to address potential overfitting, these linear models demonstrated a lack of fit, indicating that the assumptions of linearity may not hold for the data.

Subsequently, we shifted our focus to models capable of capturing non-linear relationships. We employed Partial Least Squares (PLS), Principal Components Regression (PCR), Multivariate Adaptive Regression Splines (MARS), and Generalized Additive Models (GAM) for this purpose. Through a resampling procedure, we evaluated the models' performance based on the Root Mean Squared Error (RMSE) metric. The MARS model emerged as the superior model, exhibiting the lowest RMSE and the most consistent results across resamples, as evidenced by the narrowest interquartile range. Then, we perform this model in the test data. The test error is 273.1345, which is reasonable within the context of our dataset.

Therefore, based on our analysis, the MARS model is recommended for predicting recovery time given its demonstrated ability to handle non-linearity and produce reliable predictions.

In our predictive model, we included the variable 'study', representing the cohort study to which participants belonged. Its inclusion was statistically justified, as it proved to be a significant predictor in the models, likely capturing variations in recovery times due to methodological differences between studies A and B. By integrating 'study', our model accounts for these underlying effects, enhancing accuracy and interpretability, and ensuring our findings remain relevant across different cohort characteristics.