P8106 Midterm Project Report

Group16:
Sitian Zhou, Qiduo Zhang, Shuchen Dong

Contents

1 Exploratory Analysis and Data Visualization

The original dataset contains COVID-19-related information for 3000 participants from two cohort studies. The function **initial_split()**, combined with a random seed 2527 is used to partition the data into two parts: training data (80%) and testing data (20%) before any model training. The subsequent exploratory analyses are carried out utilizing the training dataset.

The dataset contains 14 predictors: six of them are numeric variables, and eight of them are categorical variables. The recovery time is the response which is also a continuous variable. The summary of the training data is shown in **Table 1** using the **tbl_summary()** function in the **gtsummary** package, containing variable names and the relevant summary statistics.

Continuous variables: Figure 1 and **Figure 2** show scatter plots and histograms for all continuous variables, smoothing lines are used to visualize the relationship between the response and each single variable. The density of the response which will get its maximum when recovery time gets to 38.03 is shown in **Figure 3**, showing the response is uni-modal and slightly right-skewed.

Categorical variables: Figures 4 and **5** show boxplots and ridgeline plots for all categorical variables which visualize the range of response for each level.

As shown in **Figures 1** and **4**, there are some potential outliers in each predictor, especially in the large recovery time area. Particularly, in **Figure 2** we can find that most continuous variables follow a normal distribution except for age. What's more, **Figure 5** suggests that except for the study variable, recovery time has similar distribution patterns within different factor levels in other categorical variables. All these visualizations are plotted using the **ggplot2** package and are good at helping understand the distribution of all variables.

The correlation plot generated using the **corrplot( )** function is presented in **Figure 6**. We eliminated categorical variables here as only continuous variables are meaningful to examine multicollinearity. The variable bmi, as expected, is mediumly correlated with variables height and weight, and age is slightly correlated with sbp. However, no highly correlated variable pairs were detected, leading us to conclude that multicollinearity is not a concern within this dataset.

2 Model Training

In this analysis, we prepare the dataset and proceed to train a variety of regression models, using a systematic approach facilitated by the **caret** package. Here is a concise overview of the model training process:

a. **Training Control Setup:** A training control object is defined with the **trainControl()** function, with a repeated 10-fold cross-validation method. This approach is used in the following model training procedures and ensures a thorough evaluation of model performance.

b. **Model Training:** First, set the random seed to 2527, and then use the **train()** function to fit models. For instance, training the Lasso model involves establishing a tuning grid for the lambda parameter to identify the optimal value that minimizes overfitting and enhances model performance. This step requires passing the model formula, training data, and training control object. RMSE is utilized as a summary metric to evaluate model performance.

c. **Cross-Validation and Tuning:** Cross-validation results are visualized to help select the best tuning parameters. This visualization is key to understanding how different levels affect the model.
d. **Variable Importance:** The importance of predictors in the final model is examined using the **vip()** function, offering insight into which variables have the most significant impact on the outcome.

2.1 Least Squares

The Linear Model is a statistical method used to predict the relationship between one or more independent variables (explanatory variables) and a dependent variable. It assumes linear relationships between the dependent and independent variables, error term with mean 0, no multicollinearity, independence of observations, homoscedasticity, and absence of significant outliers.

In our case, the model fit by least squares approach is trained using all predictor variables and is implemented using the **train** function with the **lm** method, without tuning parameters. The final model coefficients indicate the impact of each predictor on the dependent variable.

2.2 Lasso

Lasso (Least Absolute Shrinkage and Selection Operator) is a technique in regression analysis that performs variable selection and regularization to enhance the prediction accuracy and interpretability of the statistical model. Lasso does this by introducing an adjustment parameter (lambda) that penalizes the absolute sum of the coefficients, thereby achieving variable selection and shrinkage. Besides, Lasso assumes a linear relationship between predictors and response, independence among observations, and does not require residuals to be normally distributed, but outliers can affect the model due to the L1 penalty.

In this case, the Lasso model is trained using the **train** function from the **caret** package in R, with **glmnet** method and a specific tuning grid to select the optimal adjustment parameter (lambda). The tuning grid is set using **expand.grid** for alpha = 1 and lambda values ranging from exp(0) to exp(-9) over 100 points. This range for lambda is chosen to explore a wide array of regularization strengths. The model's performance is evaluated using cross-validation, and the selection of the best tuning parameter for lambda is 0.0116256, ensuring optimal adjustment.

2.3 Ridge

Ridge regression is another technique used for dealing with collinearity, improving prediction accuracy, and estimating coefficients. Unlike Lasso, Ridge Regression penalizes the square of the coefficients, aiming to shrink the coefficients but not necessarily reducing them to zero. Besides, Ridge regression assumes a linear relationship between predictors and response, independence among observations, is more tolerant to multicollinearity due to its L2 penalty, and does not make specific assumptions about the distribution of residuals.

Similar to the lasso model, the Ridge model is trained using the **glmnet** method and a tuning grid to select the best adjustment parameter (lambda). Differing from Lasso, the tuning grid is set using expand.grid for alpha = 0 and lambda values ranging from exp(1) to exp(-5) over 100 points. Through cross-validation, the model's performance is evaluated, achieving the selection of the best tuning parameter for lambda at 0.5622787, ensuring optimal adjustment.

2.4 Elastic Net

Elastic Net combines features of Lasso and Ridge by penalizing both the L1 norm and the L2 norm of the coefficients. It is particularly useful when several predictor variables are highly correlated. Besides, Elastic Net combines the assumptions of Lasso and Ridge, including linearity and independence, and balances multicollinearity handling with outlier sensitivity.

Similarly, the Elastic Net model is trained using the **glmnet** method and a tuning grid to select the best adjustment parameter (lambda). In this case, alpha uses a sequence of 21 values from 0 to 1 and lambda values range from exp(0) to exp(-9) over 100 points. The model's performance is evaluated utilizing cross-validation, achieving the selection of the best tuning parameter for alpha at 0.15 and lambda at 0.003905168, ensuring optimal adjustment.

2.5 K-Nearest Neighbour (KNN)

KNN is a basic method for classification and regression. For regression problems, KNN finds the K closest training samples to the test point and predicts the output by averaging the dependent variables of these samples. Besides, KNN could operate without explicit model assumptions, determining outcomes based on the proximity of neighbors and assuming similar instances yield similar outcomes.

In this instance, the model training involves experimenting with various K values and determining the optimal K value via cross-validation. By employing the **knn** method, a tuning parameter grid for k is established, ranging from 1 to 20. Ultimately, the optimal tuning parameter for k is identified as 10.

2.6 Principal Components Regression (PCR)

PCR is an unsupervised approach that utilizes principal components analysis (PCA) to reduce dimension. It uses the principal components as predictors in a linear model, so the model assumptions follow that of a linear model. Furthermore, to optimize the performance of the model, the predictors also need to be centered and scaled. Besides, PCR assumes a linear relationship, which assumes multicollinearity, and results in uncorrelated PCA components.

The PCR model is trained with the method specified to **pcr**. The grid of tuning parameters is set to 1:17, where 17 is the number of predictors in the data, with all predictors being centered and scaled. The best-tuned model contains 17 components, which is the number of total predictors in the original data.

2.7 Partial Least Square (PLS)

Similar to the PCR model, PLS is also employed for dimension reduction. However, it adopts a supervised approach, taking into account the response variable. Like PCR, PLS also adheres to the assumptions of a linear model. Besides, is assumes linearity and focuses on predicting the response, handling multicollinearity by extracting directions most useful for prediction.

In our analysis, the PLS model is trained using the method specified as **pls** with the tuning grid set to 1:17. The predictors are centered and scaled utilizing the "preProcess" argument. The best-tuned model contains 10 components.

2.8 Generalized Additive Model (GAM)

The GAM model provides the flexibility to incorporate nonlinearities across various variables while maintaining the additive structure inherent in linear models. While it assumes the independent observations and the additive structure among the smooth functions, it does not assume the linear relationship between predictors and response. Besides, GAM utilizes the sum of smooth functions, assumes no predetermined form for the relationship between predictors and response, and considers observations independent.

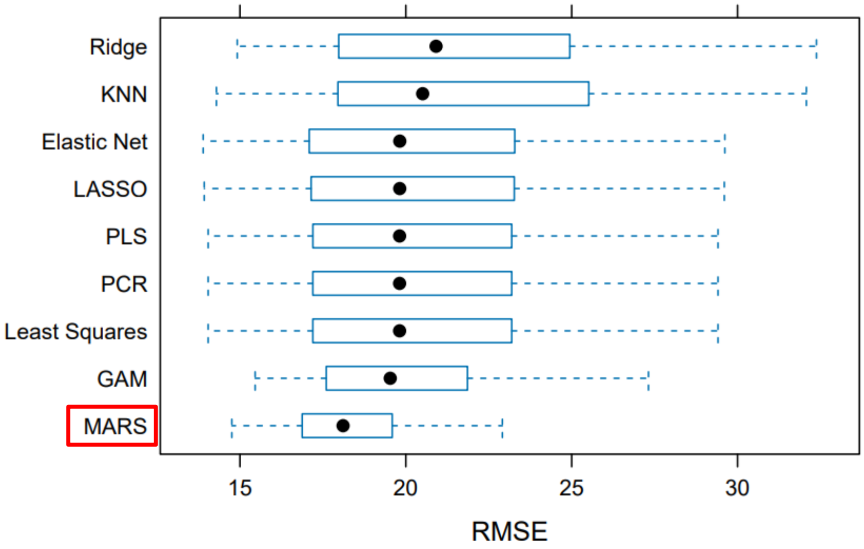
The method **gam** is used when training the GAM model. We retain the default tuning method, **GCV.Cp**, and allow variable selection by setting the selection parameter to True or False. The best-tuned model contains all predictors and sbp, bmi, height, and weight have a degree of freedom greater than 1, suggesting these variables have a non-linear relationship with the recovery time.

2.9 Multivariate Adaptive Regression Splines (MARS)

MARS is another model that allows non-linear relationships between predictors and the response variable. It employs the hinge function to construct a piecewise linear model, assuming that the relationship between predictors and response remains linear within distinct regions of the predictor space. In particular, MARS does not assume linearity, using piecewise linear fits for modeling non-linear relationships and interactions while assuming observation independence.

When training the MARS model, the method **earth** is utilized, with tuning parameter grids set to degrees ranging from 1 to 3 and nprune from 3 to 20. The best-tuned model has a degree of 2 and nprune of 16.

After training all the models, the function **resamples()** is used to assess model performance, and the summary plots are shown below. The MARS model has the smallest RMSE among all the models and was chosen as the final model.



3 Results

We selected the MARS model as the final model for predicting the recovery time, and the formula is as follows:

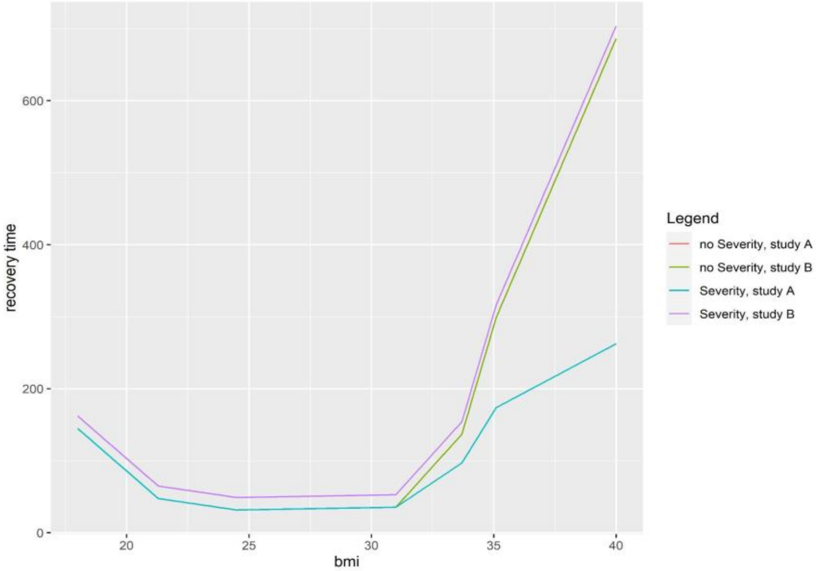
$$
\begin{aligned}
recovery\_time = {} & -229.598 - 31.521h(bmi - 31) + 29.500h(31 - bmi) + 5.310h(bmi - 24.5) \\
& + 31.995h(bmi - 33.7) + 24.536h(bmi - 21.3) + 17.348severity1 * studyB \\
& + 14.776h(bmi - 31) * studyB - 36.7h(bmi - 35.1) - 6.204vaccine1 \\
& - 2.258h(weight - 87.5) * h(bmi - 31) + 2.308h(height - 159.5) * h(bmi - 31) \\
& - 2.943gender1 + 18.022smoking1 * h(bmi - 33.7) + 0.095h(172.6 - height) \\
& * h(bmi - 24.5) + 46.175 * h(bmi - 33.7) * studyB
\end{aligned}
$$

where h( ) denotes the hinge functions. The RMSE in training data and testing data are 16.39721 and 20.18728, respectively. We further measured the variable importance using **varImp()** function. The result is shown in **Figure 7**. The result indicates that bmi and study B are the two most important variables, followed by severity. Their partial dependence plots can be seen in **Figure 8.**

Controlling other factors unchanged, we can get their equations. For example, for participants who are not from study B and have no severity, their recovery time can be predicted using (C is a constant):

$$
\textbf{recovery\_time} = \begin{cases} 304.7446 - 6.379 * bmi + C, & bmi \geq 35.1 \\ -983.3903 + 30.32 * bmi + C, & 33.7 \leq bmi < 35.1 \\ 94.8412 - 1.675 * bmi + C, & 31 \leq bmi < 33.7 \\ 32.1902 + 0.346 * bmi + C, & 24.5 \leq bmi < 31 \\ 162.2852 - 4.964 * bmi + C, & 21.3 \leq bmi < 24.5 \\ 684.902 - 29.5 * bmi + C, & bmi < 21.3 \end{cases}
$$

Now, we assume a male participant who does not smoke, receiving the vaccine, has a 75kg weight and 170cm height. In four different situations, we can observe the relationship between recovery time and BMI in the following figure:



The plot related to participants with no severity from study A overlaps with the plot for those who have severe symptoms from study A because the variable severity only has an effect on recovery time for participants from study B. Based on the figure, we can conclude that for all the participants we assumed before, those with a BMI between 24.7 and 31 need the shortest time to recover from COVID-19. Participants from study B with a BMI greater than 33.7 encounter a substantial increase in recovery time with each unit increase in BMI. However, for those from study A with a BMI greater than 33.7, the recovery time increases more slowly. Furthermore, patients from the same study with more severe symptoms experience longer recovery time compared to those with less severe symptoms.

4 Conclusion & Discussion

Throughout the analysis, the predictor "study" was included as a predictor. From the section "Exploratory analysis and data visualization", we observed the data from study B has a larger variance and more extreme values than that from study A, suggesting potential heterogeneity between data from the two studies. In this case, pooling the data from the two studies directly could introduce noise into the analysis and reduce the model's accuracy. However, using "study" as an additional predictor could reduce the interpretability of the model. The generalizability of the model is limited to the specific population, setting, or conditions studied in each study, which constrains the interpretability of the model.

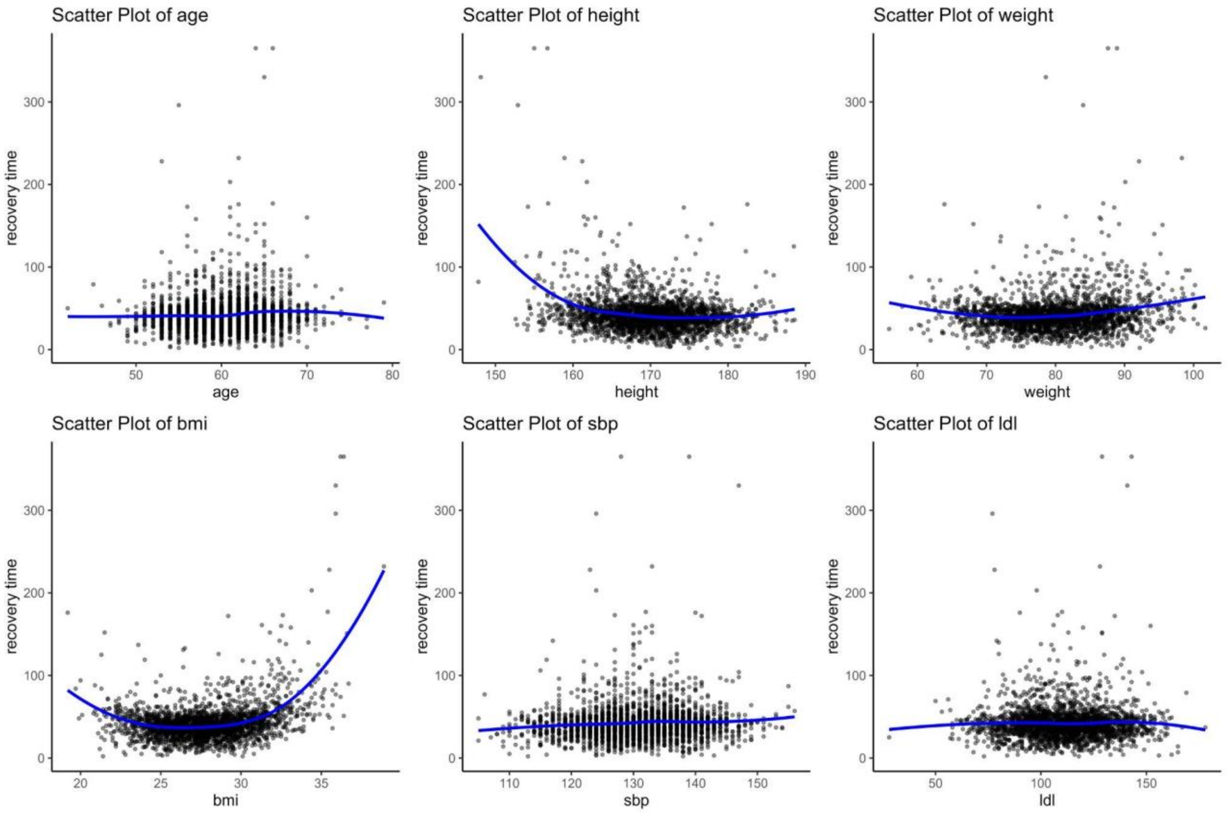
Our analysis indicates that for individuals with a BMI above 33.7, especially those in study group B, recovery time from COVID-19 significantly increases, as evidenced by the coefficient of 46.174615 for the interaction between h(bmi-33.7) and study B. This finding underscores the impact of high BMI on prolonging recovery within this subgroup, highlighting the importance of targeted BMI management to accelerate recovery efforts for those with elevated BMI levels in study group B.

In summary, it identifies specific factors significantly impacting COVID-19 recovery time: higher BMI, especially in study group B, slows recovery; males, smokers, and taller individuals face longer recovery periods; and interestingly, vaccination appears linked to slower recovery, possibly reflecting more complex health backgrounds. These insights emphasize BMI control and lifestyle modifications, like quitting smoking, to support faster recovery, particularly for those in high-risk categories.

5 Appendix

Characteristic			
	N = 3,000		
age	60.0 (57.0, 63.0)	**diabetes**	
gender	1,456 (49%)	*0*	2,537 (85%)
race		*1*	463 (15%)
1	1,967 (66%)	**SBP**	130 (125, 136)
2	158 (5.3%)	**LDL**	110 (97, 124)
3	604 (20%)	**vaccine**	
4	271 (9.0%)	*0*	1,212 (40%)
smoking		*1*	1,788 (60%)
0	1,822 (61%)	**severity**	
1	859 (29%)	*0*	2,679 (89%)
2	319 (11%)	*1*	321 (11%)
height	169.9 (166.0, 173.9)	**study**	
weight	80 (75, 85)	*A*	2,000 (67%)
bmi	27.65 (25.80, 29.50)	*B*	1,000 (33%)
hypertension		**recovery_time**	39 (31, 49)
0	1,508 (50%)		
1	1,492 (50%)		

Table 1. Summary statistics table


Scatter Plot of age

Scatter Plot of height

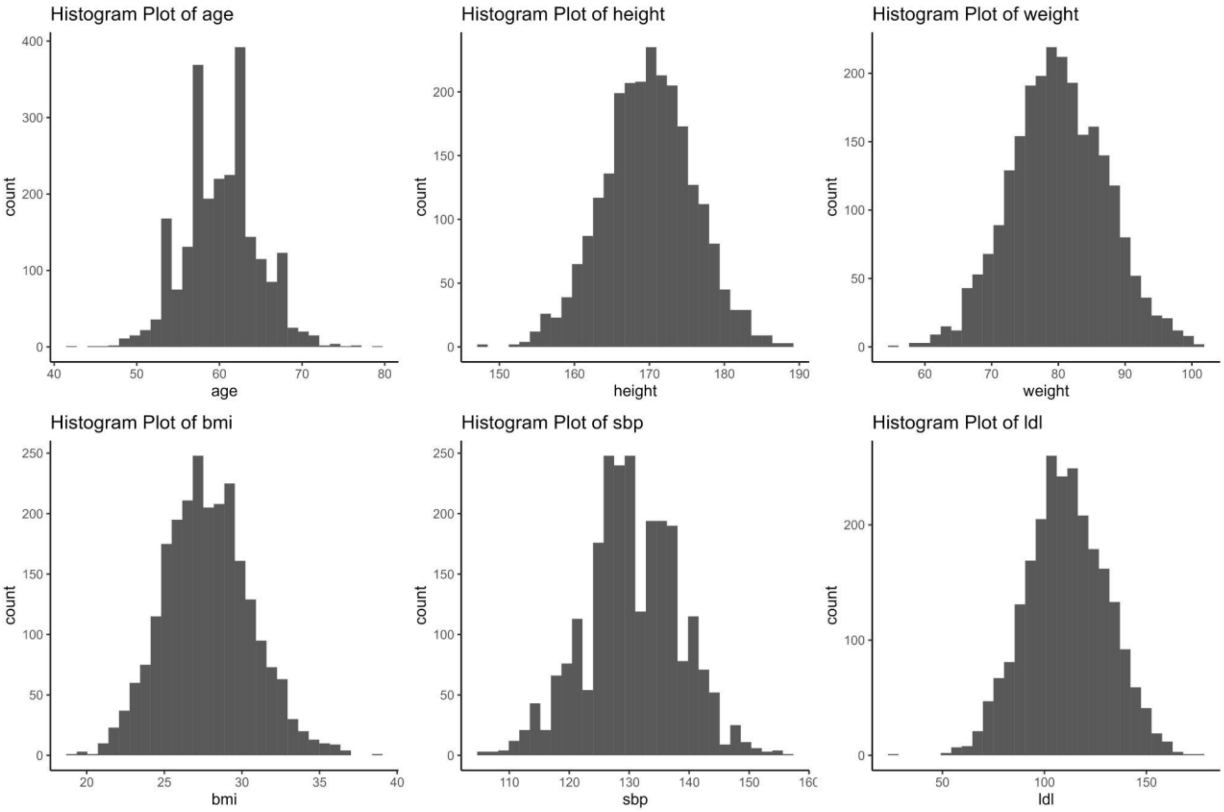
Scatter Plot of weight

Scatter Plot of bmi

Scatter Plot of sbp

Scatter Plot of ldl

Figure 1. Scatter plots for continuous variables against recovery time


Histogram Plot of age

Histogram Plot of height

Histogram Plot of weight

Histogram Plot of bmi

Histogram Plot of sbp

Histogram Plot of ldl

Figure 2. Histograms for continuous variables

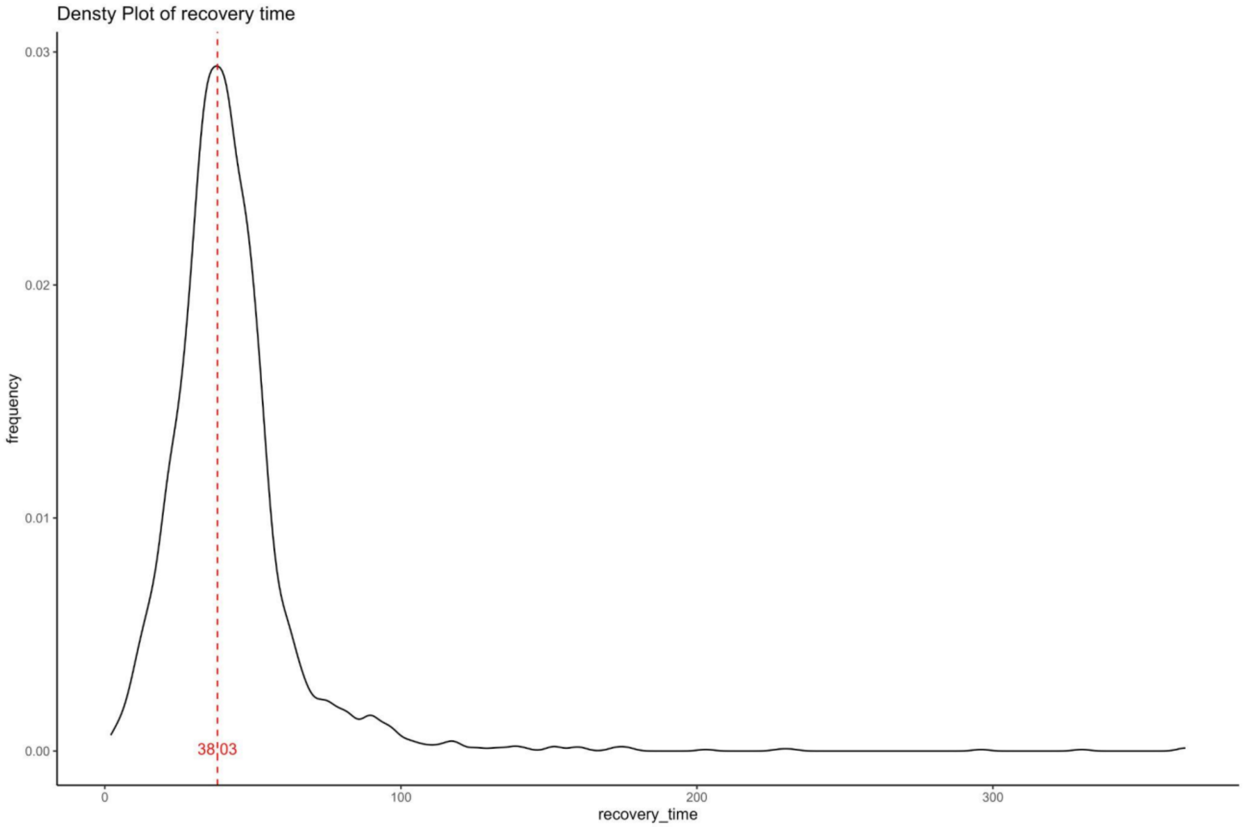


Figure 3. Density plot for the response variable

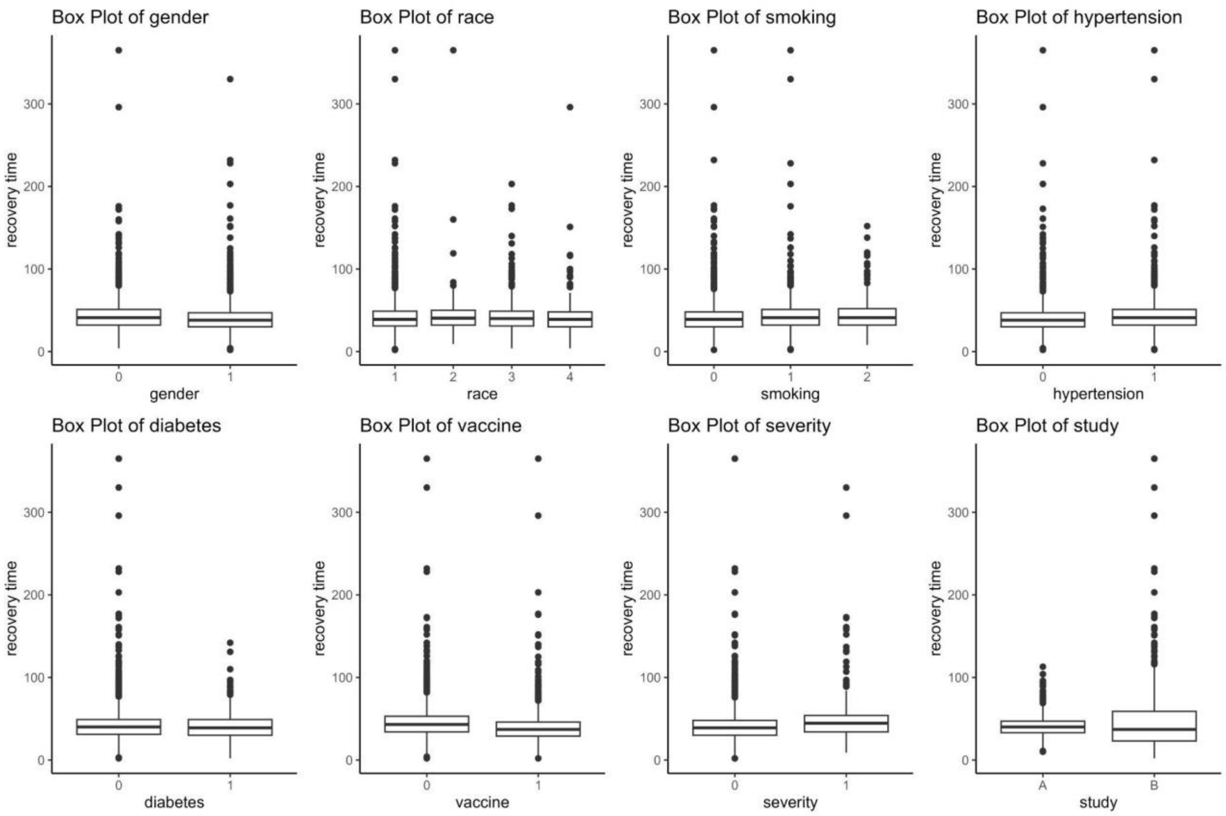


Figure 4. Box plots of recovery time by each categorical variable

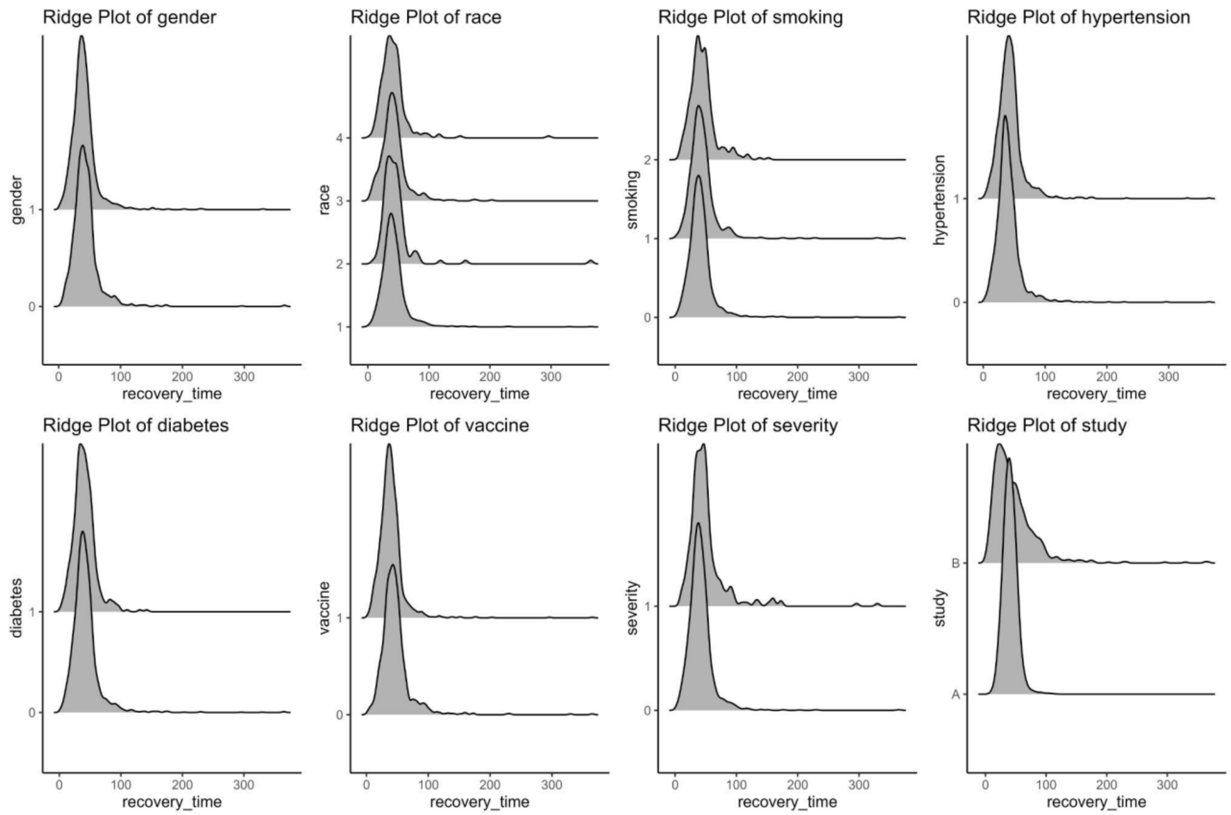


Figure 5. Ridge plots of recovery time by each categorical variable

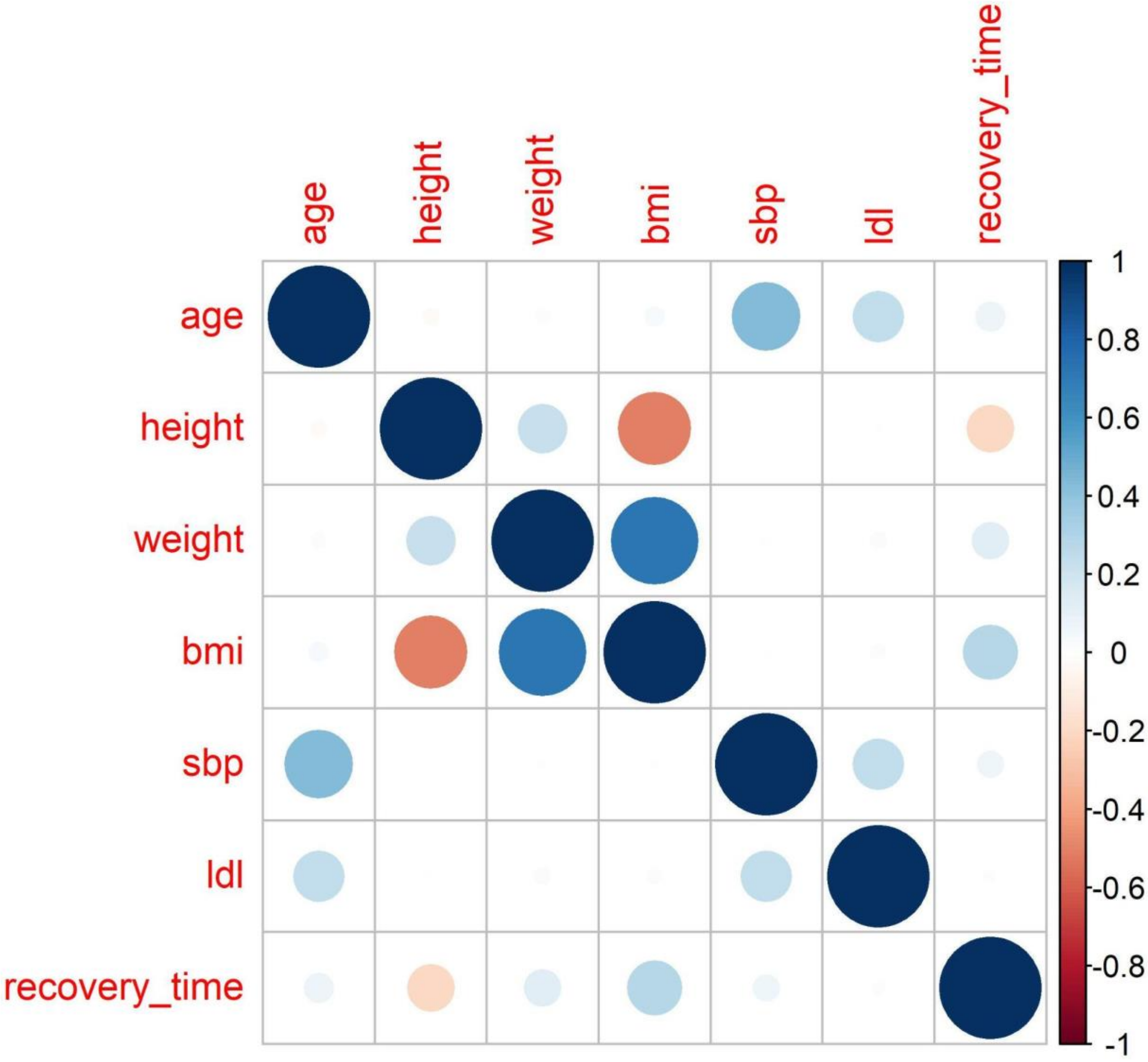


Figure 6. Correlation among continuous variables

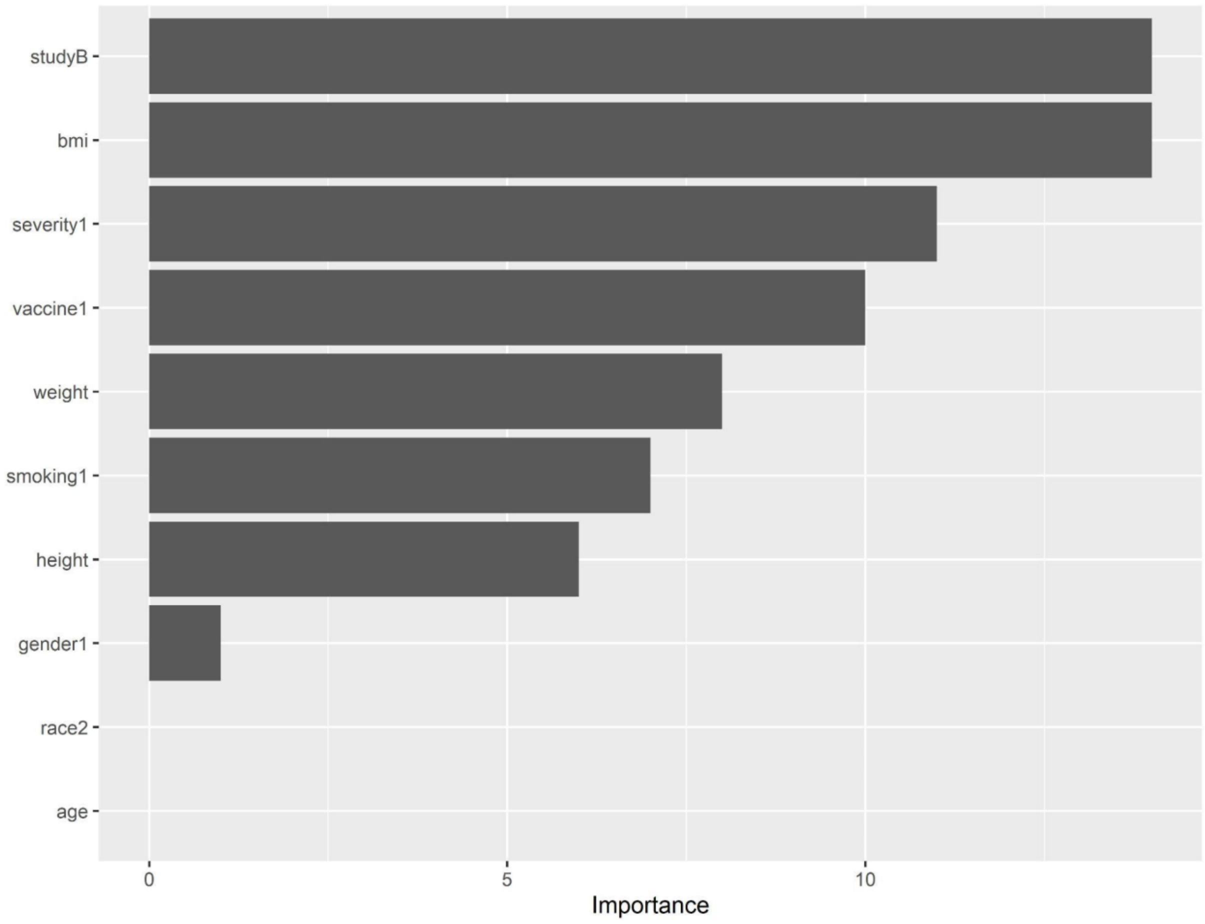


Figure 7. Variable importance for the MARS model

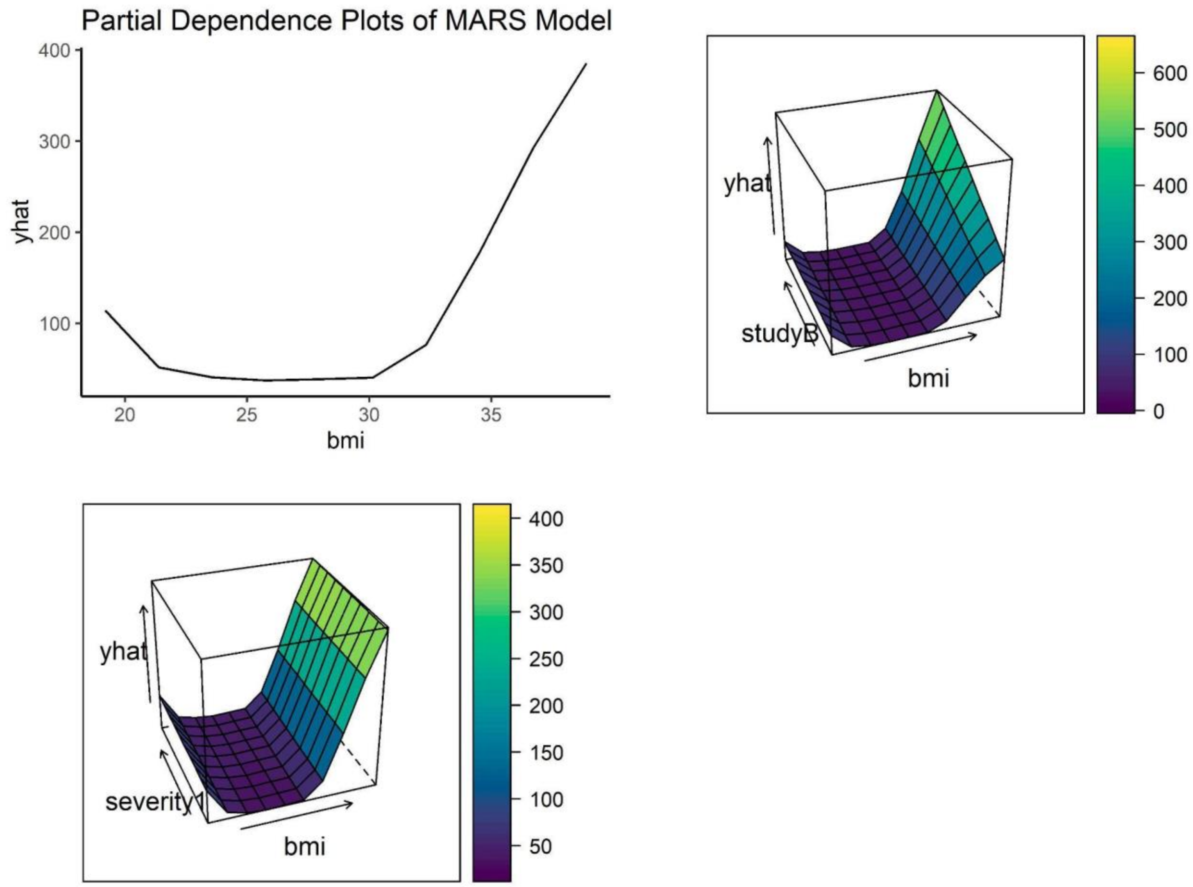


Figure 8. Partial dependence plots