P8106 Data ScienceII Midterm

Yueran Zhang (yz4188)

2023-03-26

Introduction

Background

The COVID-19 pandemic had a significant impact on global health. It is The Covid-19 pandemic had a significant impact on global health. It is one of many mysteries about long Covid: What is the recovery timeline? Are some people more likely than others to experience long coronavirus infections? Long Covid can be a debilitating illness that affects multiple organ systems, with over 200 identified symptoms(Davis et al., 2023). The understanding of important risk factors allows people to correlate long Covid with various determinants such as pre-existing characteristics, medical records, genetics, and lifestyle. According to CDC, some groups of people may be affected more by Post-Covid Conditions, such as people who have experienced more severe COVID-19 illness, especially those who hospitalized, or people who did not get a Covid-19 vaccine(CDC, 2021). Therefore, predicting recovery time from Covid-19 illness and identifying important risk factors for long recovery times are crucial for recovery from Covid-19.

The study was designed to combine three existing cohort studies that have been tracking participants for several years. The study collects recovery information through questionnaires and medical records, and leverages existing data on personal characteristics prior to the pandemic.

Data Descprtion

The dataset is "recovery.RData" that consists of 10000 participants. We generate a random sample of 2,000 participants and create reproducible results by using "set.seed(4188- as my uni numbers)function". ID variables do not convey any useful information and is dropped.The dataset now contains 2000 observations and 15 variables, including pre-existing characteristics(eg.age, gender,BMI, height, weight), medical records(eg.diabetes history,hypertension and vaccination condition), and lifestyle(eg.smoking status). 1 of the 15 variables include recovery time, which is the target variable. The predictors are as following:

- Gender (gender) 1 = Male, 0 = Female
- Race/ethnicity (race) 1 = White, 2 = Asian, 3 = Black, 4 = Hispanic
- Smoking (smoking) Smoking status; 0 = Never smoked, 1 = Former smoker, 2 = Current smoker
- Height (height) Height (in centimeters)
- Weight (weight) Weight (in kilograms)
- BMI (bmi) Body Mass Index; BMI = weight (in kilograms) / height (in meters) squared
- Hypertension (hypertension) 0 = No, 1 = Yes
- Diabetes (diabetes) 0 = No, 1 = Yes
- Systolic blood pressure (SBP) Systolic blood pressure (in mm/Hg)
- LDL cholesterol (LDL) LDL (low-density lipoprotein) cholesterol (in mg/dL)
- Vaccination status at the time of infection (vaccine) 0 = Not vaccinated, 1 = Vaccinated
- Severity of COVID-19 infection (severity) 0 = Not severe, 1= Severe

- Study (study) The study (A/B/C) that the participant belongs to
- Time to recovery (tt_recovery_time) Time from COVID-19 infection to recovery in days

Cleaning the Data

Though there seems to be many numeric/integer variables, not all of them are true numerical variables. Some are displayed as numbers but are really factors("gender", "hypertension", "diabetes", "vaccine", "severity", "study"). For example, for "gender" variable, we use 'number 1' to represent male, so the 'number 1' has no mathematics meaning, only for labeling categories. These variables will be converted from int to factor. Now we have 2,000 observations' data with 8 categorical(factor) variables, 7 numerical variables. After checking, there is no null value or missing data in our dataset.

For training and testing purpose, we randomly divided the dataset of 2,000 particicants into two subsets: training set (70%) and the testing set (30%). The exact same training and testing set was used for the training of all models to ensure the reproducibility of the process.

Table 1: Data summary

Name	Piped data
Number of rows	2000
Number of columns	15
Column type frequency:	
factor	8
numeric	7
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
gender	0	1	FALSE	2	0: 1027, 1: 973
race	0	1	FALSE	4	1: 1285, 3: 416, 4: 202, 2: 97
smoking	0	1	FALSE	3	0: 1207, 1: 601, 2: 192
hypertension	0	1	FALSE	2	0: 1009, 1: 991
diabetes	0	1	FALSE	2	0: 1709, 1: 291
vaccine	0	1	FALSE	2	1: 1194, 0: 806
severity	0	1	FALSE	2	0: 1824, 1: 176
study	0	1	FALSE	3	B: 1204, A: 400, C: 396

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
recovery_time	0	1	43.05	28.33	3.0	28.0	39.0	50.0	365.0	
age	0	1	60.09	4.48	45.0	57.0	60.0	63.0	77.0	
height	0	1	169.99	5.91	149.7	166.2	170.0	173.9	189.1	
weight	0	1	80.09	7.32	57.5	75.3	80.0	85.3	103.7	
bmi	0	1	27.77	2.77	19.7	25.8	27.7	29.6	39.4	
SBP	0	1	130.32	7.98	103.0	125.0	130.0	136.0	157.0	
LDL	0	1	110.08	19.99	33.0	97.0	110.0	124.0	173.0	

Exploratory Data Analysis

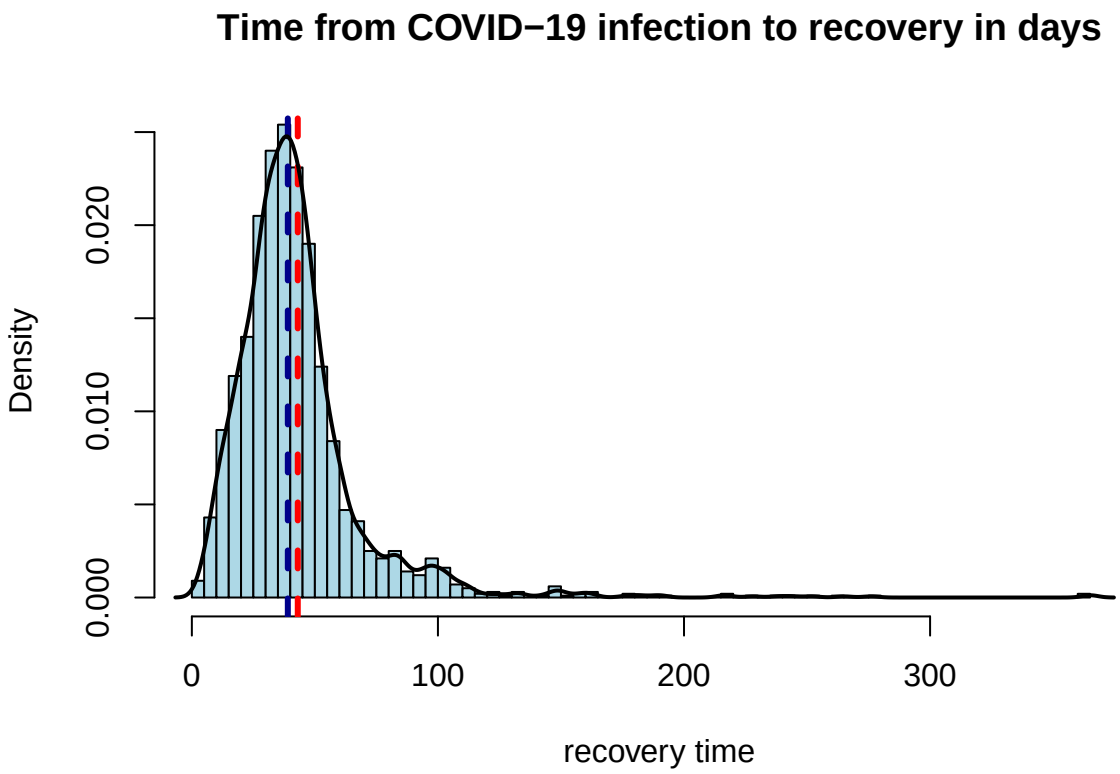
Looking at the Target Feature

```
## Mean recovery time:  43

##
## Median recovery time:  39

##
## Max recovery time:  365

##
## Min recovery time:  3
```



The average mean recovery time from Covid is 43 days; median recovery time is 39 days; the longest recovery process would costs 365 days(1 year); and the minimum days of recovery is only 3 days.The histogram of the recovery time is a little bit of right skewed. The mean is higher than the median. There are also a good number of outliers.

Other Features Compare with recovery time

Numerical Features

Figure 1. Numercial Varibables − Scatter Plots

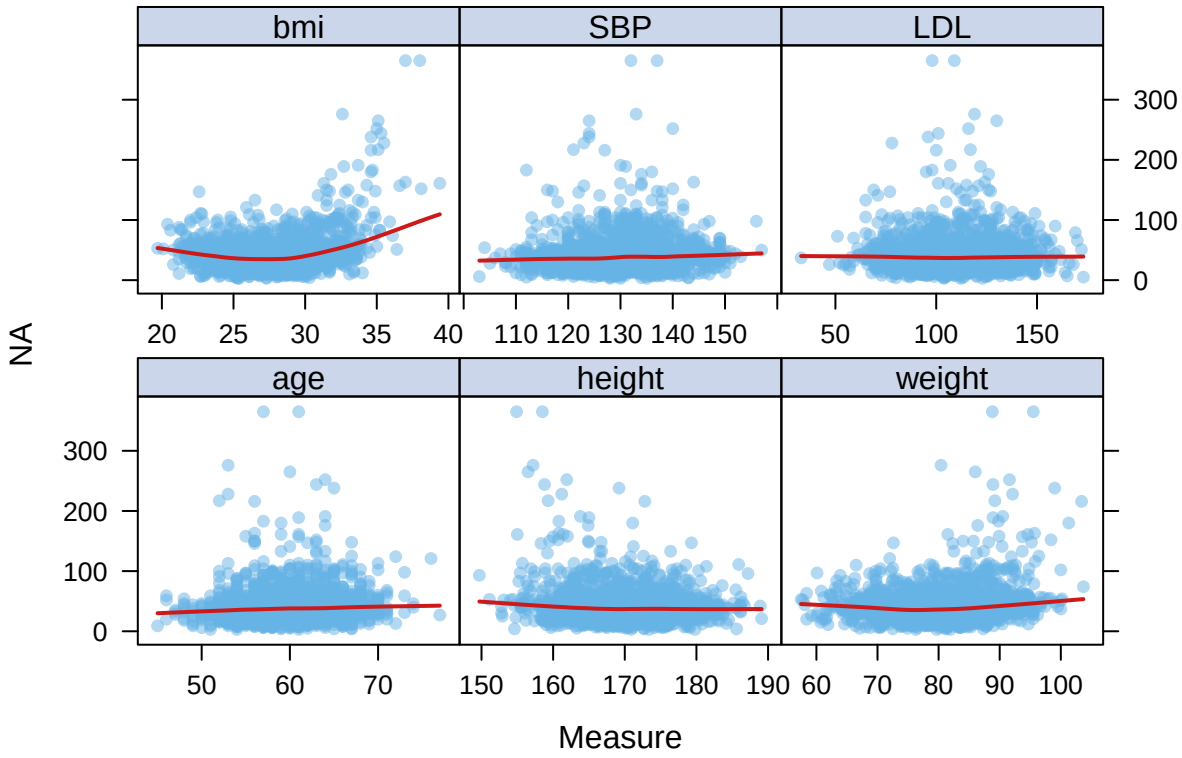
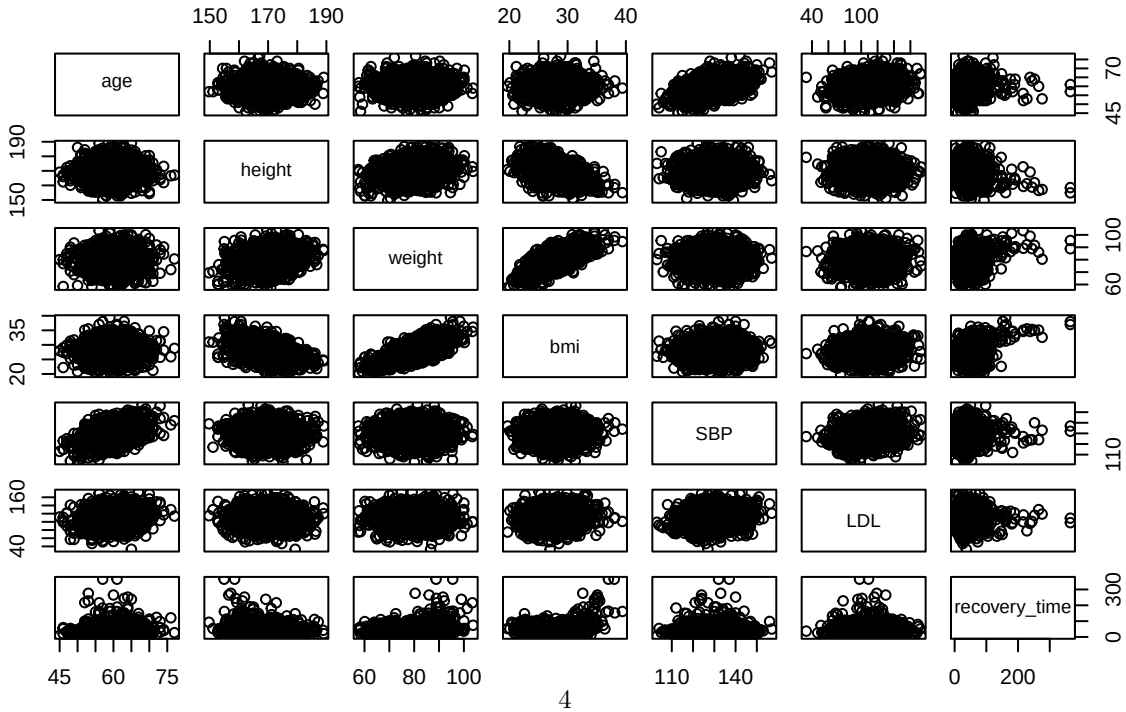


Figure2. Correlations between numercial variables

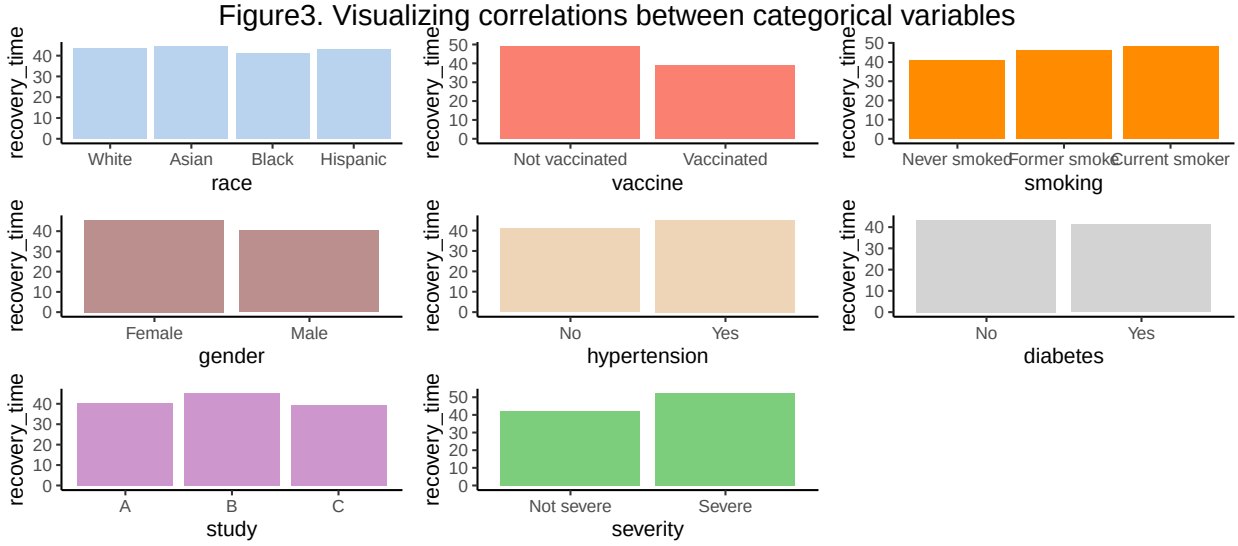


As observed the numerical variables plots, we can assume that,

- **BMI** - Participants BMI tends to curvilinear(U-shaped) correlated with the recovery time. If participants starts BMI with 20-30, then increasing the BMI value would decrease the recovery time(Minimum recovery time is BMI=19.7). However once participants pass the BMI 30 range and start getting to longer recovery time(Maximum recovery time is BMI=39.4), increasing the BMI value would increase the recovery time. The average shorter recovery time falls into BMI 25-30.
- **SBP** - Participants' Systolic blood pressure value seems slightly positive statistically associated with recovery time. There is a slightly upward sloping line, that's to say, people with lower SBP measurement value seems to have shorter recovery time than people with high SBP value.
- **LDL** - Participants' LDL cholesterol value seems no significant statistically correlated with recovery time. We observe a nearly horizon linear line, and we cannot say that people with minimum LDL value has huge difference recovery time with people had the maximum LDL value.
- **Age** - Participants' age seems slightly positive statistically associated with recovery time. From the plot, we see a slightly upward sloping line, so older people may need more time to recover from Covid compared younger people.
- **Height** - Participants' height seems slightly negative statistically associated with recovery time. It shows that a slightly downward sloping line, indicating that much shorter people may need more recovery time, but taller people need less time to recovery from Covid.
- **Weight** - Participants weight is slightly curvilinear(U-shaped) correlated with the recovery time. If people within 60-80 kgs, then increasing the weight would have less the recovery time. However if participants pass 80 kgs range and start needing longer recovery time, increasing in weight value would increase the recovery time after people.
- Finally, most variables are a little bit correlated with one another. As we can tell that **age and BMI** value are significantly negative correlated. **weight and BMI** seems positive correlated, that is we observed as above that they seems both have an 'u-shaped' correlated with recovery time.

Categorical Features

Figure3. Visualizing correlations between categorical variables



From the above categorical graphs, we can assume that,

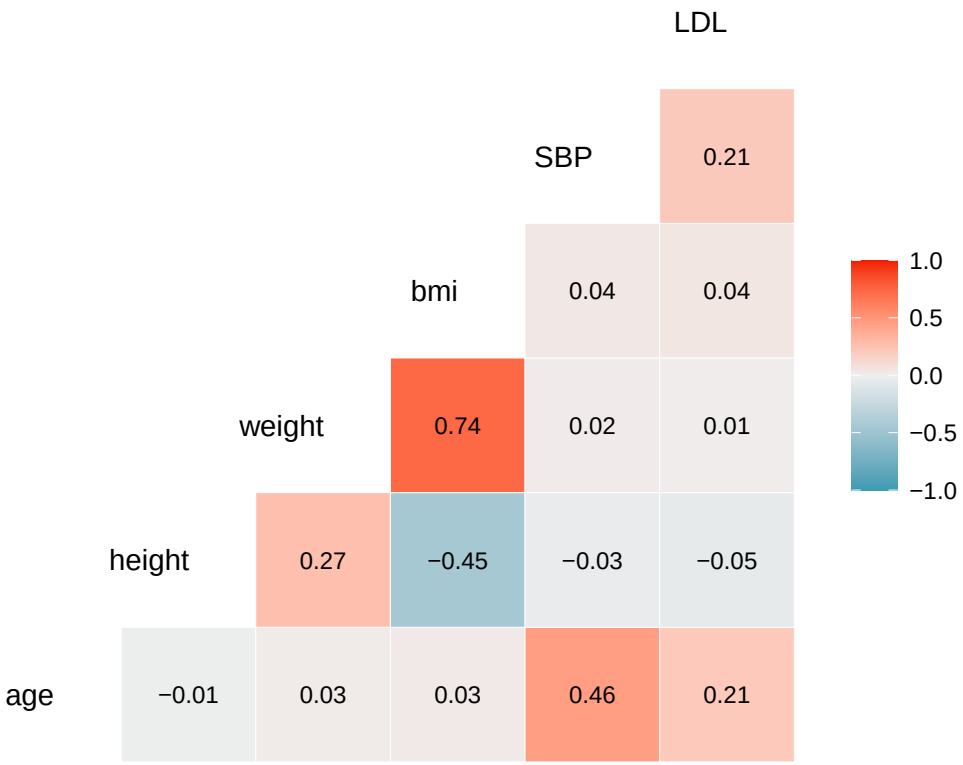
- **Race** - There is no significant difference of recovery time among various race.The average of recovery days is more than 40 days. However, people self-identified as Asian tend to have the longest average recovery time and self-identified as American African tend to have the average shortest recovery time.
- **Vaccine** - From the graph, we can assume that vaccine is a related factors with recovery time. Since people in the not vaccined group(around 50 days of recovery time) have longer recovery time than the people in vaccined group (around 40 days of recovery time).
- **Smoke** - It seems that smoking status is correlated with recovery time. There is difference of recovery time among among disparate smoking status.If people never smoke, they tend to less recovery time(around 40 days). People with smoking history or still are smokers are more likely to need more time to recovery from COVID-19.
- **Gender** - It seems that female may need longer time to recover after COVID-19 infection. We can assume that gender tends to influence recovery time.
- **hypertension** - There is a positive association between participants hypertension status and recovery time, as people with hypertension need more time to recover after COVID-19 infection.
- **diabetes** - From the diabetes variable, it is suprisingly that people without diabetes would be need more time to recovery while people with diabetes seem to less recovery time after Covid infection.
- **Study** - People from Study Group B would have a longer recovery time compared participants from Group A or Group C.
- **Severity** - From the above information, we can assume that severity condition is a related factors with recovery time. People in severe condition would be likely more time to recover from Covid than people not in severe condition.

Modeling

Methods

Looking back our target outcome is a continuous variable(recovery time from COVID-19 infection), we would first start with regression model,as the most simple and popular technique for predicting a continuous variable.As we only have 14 variables,all these variables in the data will be set to fit the model. From the above EDA, our dataset contains some correlated predictors, where we could perform Principal components regression (PCR) and partial least squares regression((PLS).This technique constructs a set of linear combinations of the inputs for regression. In order to to simplify a large multivariate model is to use penalized regression, we would use Ridge regression,lasso regression and Elastic net model. For some variables, the relationship between the target outcome and the predictor variables is not linear.In these situations, we need to build a non-linear regression, like Generalized additive model (GAM) and Multivariate Adaptive Regression Splines(MARS).

One of the most robust and popular approach for estimating a model performance is k-fold cross-validation, and note that, the best model is the model that has the lowest cross-validation error, RMSE. In our study, we use 15-fold cross-validation repeated 5 times.



It shows that **weight and BMI**, **height and weight**, **height and BMI**, **age and SBP**, **age and LDL** and **LDL and SBP** are correlated. In this situation, there might be an interaction effect between some predictors.

1. Multiple linear regression

In this section, we'll build a multiple regression model to predict recovery time based on the other participants' characteristics variables, such as gender, race, smoking status and so on. Once identified the model, we continue the diagnostic by checking how well the model fits the data.We use the method = 'lm' syntax for linear regression models.

The Residual standard error (RSE) = 24.07, meaning that the observed recovery time deviate from the predicted recovery time by approximately 24.07 units in average. This corresponds to an error rate of 24.07/mean(training.data$recovery_time) = 24.07/42.93581 = 56.06%, which is pretty high. The Adjusted R-square value in the summary output is a correction for the number of 15 variables included in the predictive model. R-Squared(RSq) is 0.2524,the regression model did not explain much of the variability in the outcome.

2. Penalized Regression

The standard linear model performs poorly in this situation.A better alternative is the penalized regression allowing to create a linear regression model that is penalized.This is also known as shrinkage or regularization methods.The consequence of imposing this penalty, is to reduce (i.e. shrink) the coefficient values towards zero. This allows the less contributive variables to have a coefficient close to zero or equal zero.In this section, we will use penalized regression methods, including ridge regression, lasso regression and elastic net regression.

1) Ridge Regression

Ridge regression shrinks the regression coefficients, so that variables, with minor contribution to the outcome, have their coefficients close to zero.The shrinkage of the coefficients is achieved by penalizing the regression model with a penalty term called L2-norm, which is the sum of the squared coefficients.Alpha=0 the ridge penalty. The output shows that the RMSE and R-squared values for the ridge regression model on the training data are 27.414 and 15.57%, respectively. It seems that no improvement from linear model or we say that even worse. Let's move on to the lasso model.

2) Lasso Model

One disadvantage of the ridge regression is that, it will include all the predictors in the final model.Ridge regression shrinks the coefficients towards zero, but it will not set any of them exactly to zero. The lasso regression is an alternative that overcomes this drawback. It shrinks the regression coefficients toward zero by penalizing the regression model with a penalty term called L1-norm, which is the sum of the absolute coefficients. Alpha=1 is the lasso penalty.The summary output shows the model did not tune alpha because I held it at 1 for lasso regression. The optimal tuning values (at the minimum RMSE) were alpha = 1 and lambda = 0.0012. The lasso model finds that the an R-squared of 27.90% with RMSE of 25.37%.There is some improvement in the performance compared with linear model. Next, we we want to explore the Elastic net model.

3) Elastic Net Model

Elastic Net produces a regression model that is penalized with both the L1-norm and L2-norm. The consequence of this is to effectively shrink coefficients (like in ridge regression) and to set some coefficients to zero (as in LASSO).The caret packages tests a range of possible alpha and lambda values, then selects the best values for lambda and alpha, resulting to a final model that is an elastic net model. The optimal tuning values (at the mininum RMSE) were alpha = 0.0 and lambda = 1, so the mix is 100% ridge, 0% lasso. The lElastic Net shows that the an R-squared of 15.43% with RMSE of 27.437%.

3. Dimension reduction

From above EDA, this data set has multiple correlated predictor variables. Here, we used two well known regression methods based on dimension reduction: Principal Component Regression (PCR) and Partial Least Squares (PLS) regression.

1) Principal Component Regression

The principal component regression (PCR) first applies Principal Component Analysis on the data set to summarize the original predictor variables into few new variables also known as principal components (PCs), which are a linear combination of the original data.We simply specify method = "pcr" within train() to perform PCA on all our numeric predictors prior to fitting the model.The PCR model perform 15-fold cross validation repeated 5 times a PCR model tuning the number of principal components to use as predictors from 1-18.

By controlling for multicollinearity with PCR, we can experience significant improvement in our predictive accuracy compared to the previously obtained linear models (reducing the cross-validated RMSE from about 25 to nearly 23).

2) Partial least squares

Similar to PCR, we can easily fit a PLS model by changing the method argument in train(). We found that the RMSE is also dropped as 23.88 when compared with linear model.

3. Beyond linearity

1) Generalized additive model(GAM)

We have detected a non-linear relationship in your data, the advantage of GAM is that they automatically model non-linear relationships so we do not need to manually try out many different transformations on each variable individually. Here, we used train() with method = 'gam'syntax to perform GAM model.GCV is used for smoothness selection in the model; smoothing parameters are chosen to minimise prediction error.

From the output, it shows that the R-sqaured is 0.417, which is a great improvement from the former models.

2) Multivariate adaptive regression splines (MARS)

This model is a non-parametric algorithm that creates a piecewise linear model to capture nonlinearities and interactions effects. We use earth() function performs the MARS algorithm.The caret implementation tunes two parameters: nprune and degree. In our model setting, the nprune is between 2 to 19, that is the maximum number of terms in the model. The degree is set 1-3, set as the maximum degree of interaction.

For the MARS model, the best tuning parameters are nprune is 4 and degree is 2.

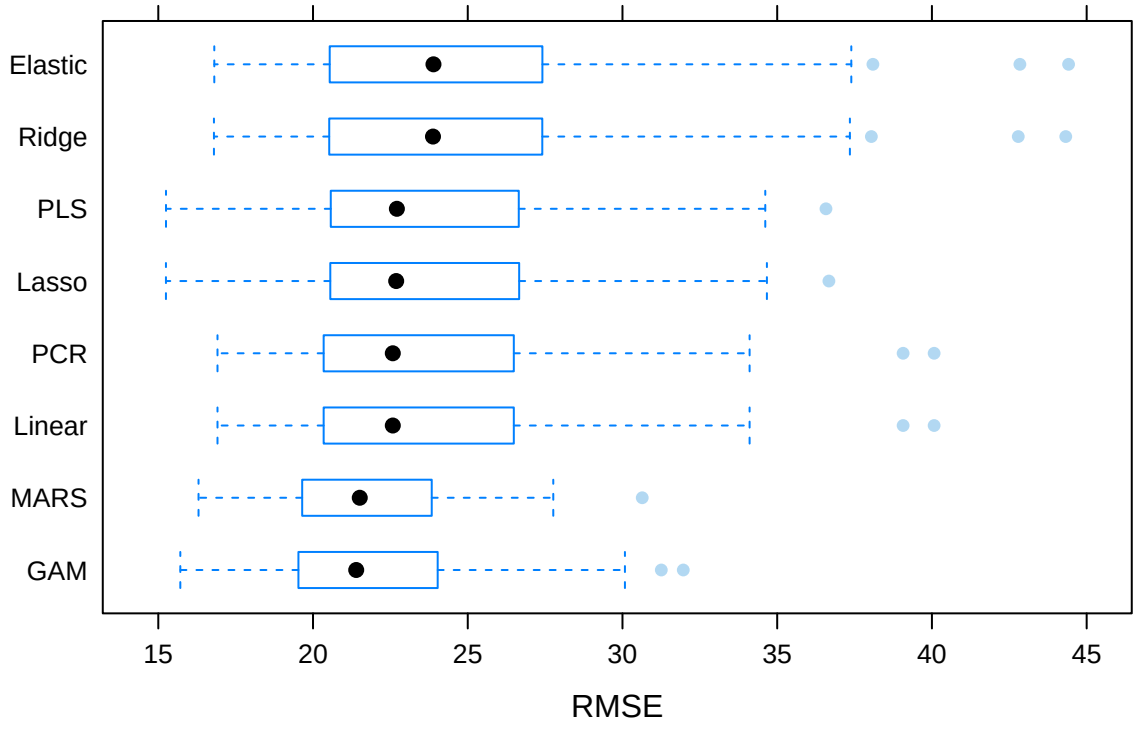
Results

Model comparing

```
##
## Call:
## summary.resamples(object = res)
##
## Models: Linear, Ridge, Lasso, Elastic, PCR, PLS, GAM, MARS
## Number of resamples: 75
##
## MAE
##             Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## Linear  12.40749 14.98670 15.80654 16.03241 17.25612 21.24611    0
## Ridge   12.63789 14.91455 16.18251 16.25742 17.46053 22.07011    0
## Lasso   11.93503 14.72117 15.81299 15.99679 17.04037 21.11045    0
## Elastic 12.63924 14.91500 16.18184 16.26059 17.46417 22.07681    0
## PCR     12.40749 14.98670 15.80654 16.03241 17.25612 21.24611    0
## PLS     11.94196 14.76273 15.82870 16.01731 17.06492 21.12309    0
## GAM     11.92519 14.01003 14.88853 15.09315 16.20321 18.74405    0
## MARS    12.34664 14.15148 14.88598 15.09019 15.89449 18.12048    0
##
## RMSE
##             Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## Linear  16.91310 20.34373 22.57789 23.85554 26.48867 40.07092    0
## Ridge   16.79863 20.52402 23.87653 24.74003 27.40849 44.32468    0
## Lasso   15.24744 20.55835 22.68774 23.87973 26.66135 36.67105    0
## Elastic 16.80970 20.53943 23.89226 24.75558 27.41249 44.41587    0
## PCR     16.91310 20.34373 22.57789 23.85554 26.48867 40.07092    0
## PLS     15.24901 20.57350 22.70986 23.88424 26.64884 36.57629    0
## GAM     15.71197 19.52856 21.39551 21.90043 24.02857 31.96522    0
## MARS    16.30067 19.65013 21.51042 21.77451 23.83634 30.63837    0
```

```
##
## Rsquared
##                 Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## Linear  0.0156711497 0.1687404 0.2242848 0.2376198 0.3077384 0.4498982    0
## Ridge   0.0042248581 0.1205466 0.1702171 0.1733856 0.2146821 0.3774604    0
## Lasso   0.0436044876 0.1505055 0.2174820 0.2319088 0.3258687 0.4790890    0
## Elastic 0.0041620560 0.1198326 0.1691600 0.1723974 0.2140520 0.3756773    0
## PCR     0.0156711497 0.1687404 0.2242848 0.2376198 0.3077384 0.4498982    0
## PLS     0.0437655739 0.1494952 0.2178169 0.2320316 0.3274695 0.4806585    0
## GAM     0.0502564255 0.2369335 0.3324919 0.3593539 0.4715109 0.6837745    0
## MARS    0.0003428931 0.1600101 0.3207824 0.3444724 0.5370344 0.7719202    0
```

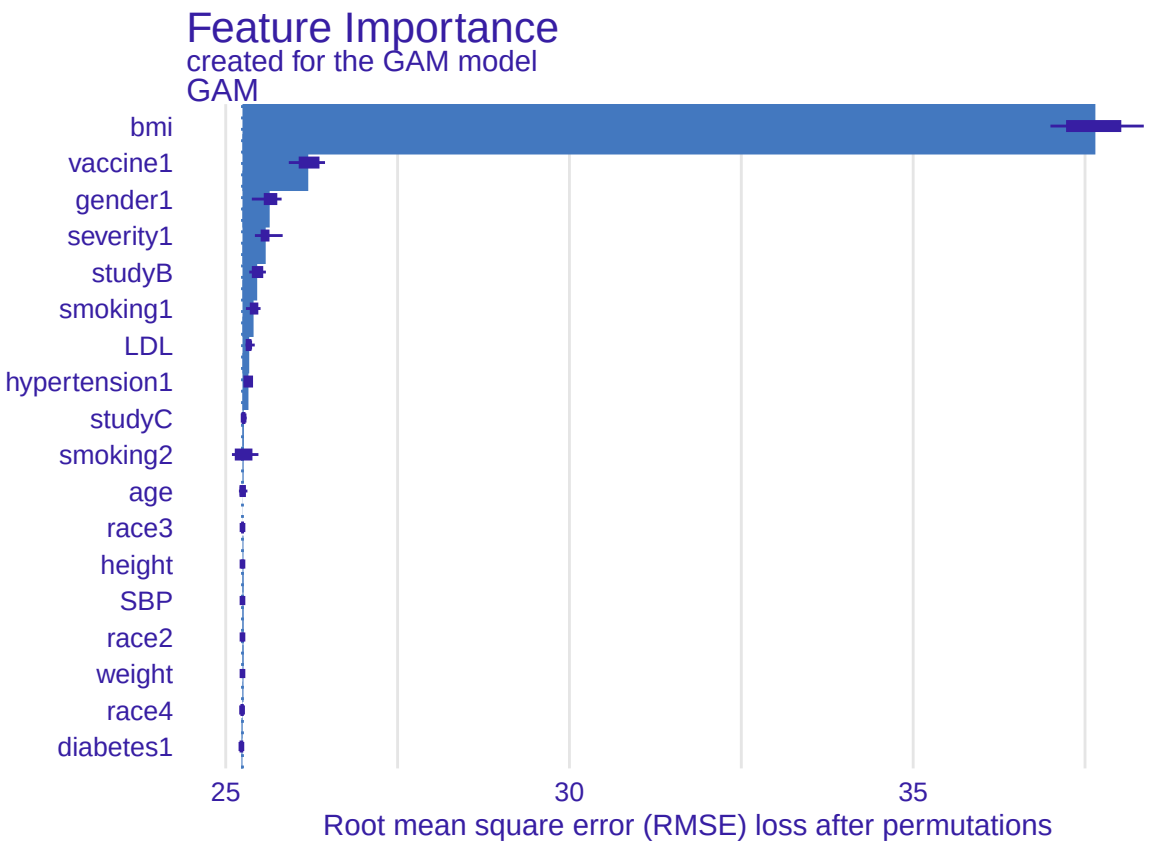
Figure 4. Model Comparing Plot based on RMSE



As our target outcome is a continuous variable - recovery time, we could use RMSE(Root mean squared error), that measure error is to take the difference between the actual and predicted value for a given observation. Our objective is minimize RMSE. Extracting the results for each model, the GAM model with the lowest median training RMSE.In this case, the GAM model performs the "best" (compared with the others).

Feature interpretation

Once we've found the model that maximizes the predictive accuracy, our next goal is to interpret the model structure.Variable importance seeks to identify those variables that are most influential in our model.




```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## .outcome ~ gender1 + race3 + race4 + smoking1 + smoking2 + hypertension1 +
##     diabetes1 + vaccine1 + severity1 + studyB + studyC + s(age) +
##     s(SBP) + s(LDL) + s(bmi) + s(height) + s(weight)
##
## Parametric coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept)   44.84105    1.76300  25.435 < 0.0000000000000002 ***
## gender1       -5.91458    1.13457  -5.213 0.000000214033603973 ***
## race3         -0.04456    1.43711  -0.031              0.97527
## race4          0.74892    1.92912   0.388              0.69791
## smoking1       3.47958    1.26685   2.747              0.00610 **
## smoking2      11.74760    1.97495   5.948 0.000000003429570379 ***
## hypertension1  3.40135    1.21087   2.809              0.00504 **
## diabetes1     -1.10140    1.62493  -0.678              0.49800
## vaccine1      -9.60867    1.15927  -8.289 0.000000000000000269 ***
## severity1      9.15336    2.05305   4.458 0.000008926137572436 ***
## studyB         3.84225    1.46814   2.617              0.00897 **
## studyC        -0.86473    1.80010  -0.480              0.63103
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Approximate significance of smooth terms:
##                      edf Ref.df      F             p-value
## s(age)    0.77450353031      9  0.373              0.0374 *
## s(SBP)    0.00000039197      9  0.000              0.4097
## s(LDL)    2.64399407486      9  0.561              0.1076
## s(bmi)    7.91677303440      9 91.832 <0.0000000000000002 ***
## s(height) 0.00000080036      9  0.000              0.5303
## s(weight) 0.00000007039      9  0.000              0.7246
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.417   Deviance explained = 42.6%
## GCV = 453.64  Scale est. = 446.09    n = 1402
```

- Some of the same features that were considered highly influential in our GAM model, and importance is determined by magnitude of the standardized coefficients.
- **BMI** -As we observed that BMI is a crucial factor associated with recovery time as being the highest importance. With the coefficient of 7.92, we would say that with one unit change in BMI value, the predicted Covid recovery time would average increase 7.92 days while controlling for other variables.
- **Vaccination** -Followed with vaccined ('not vaccined' as reference group) status, is negative associated with the preidcted Covid recovery time, as the parameter coefficient is negative value, which means one unit change in vaccination status(change from Not-vaccined[code:0] to Vaccined[code:1]), the predicted recovery time would average decrease 9.61 days.
- **Gender** - It is shows that gender is negative associated with the recovery time.As setting male as reference group, if we change one unit in gender, for biological gender of male[code:0] to female[code:1], the predicted recovery time would average decrease 5.9 days.
- **severe status** - Participants' severe status is positive associated with the predicted recovery time('Not sereve" as reference). When controlling other variables, people who are in severe status are more likely to need another 9.15 days to recover from Covid compared with individuals who are not in severe status.
- **Smoking** - Smoking status is also postive associated with the predicted recovery time ('Never smoked' as reference group). We would say that if people are former when compared people who never smoke, the predicted Covid recovery time would average increase 3.48 days while controlling for other variables. What's more, if people are current smoker, it would takes average more than 11.75 predicted recovery days when compared with people never smoke.
- **Hypertension** - Hypertension ('no hypertension' as reference group) status, is positive associated with the predicted recovery time, indicating people with hypertension that the predicted recovery time would average increase 3.40 days.
- **study** - People in study B is associated with 3.84 days longer predicted recovery time than people from Study A(Study A as reference group).
- **Age** - Age is positive associated with the predicted Covid recovery time. People with one unit change in age, the recovery time increase 0.774 days.
- For the GAM model, the Deviance explained by the model is 42.6%, and the adjusted R-Square is 41.7%, that showing moderate level of correlation.

Conclusion

Lastly, we use the GAM model with all the predictors as the final model to predict Covid recovery time. Also, we identify the extract top 20 influential variables that significantly associated with our target. It illustrates that BMI is the most influential followed by vaccination status, and gender,which matches our EDA section finding in some parts. The limitation of using GAM as the final model is that the model is restricted to be additive; GAM are additive in nature, which means there are no interaction terms in the Model.

Reference

CDC. "Long COVID or Post-COVID Conditions." Centers for Disease Control and Prevention, 16 Sept. 2021, www.cdc.gov/coronavirus/2019-ncov/long-term-effects/index.html.

Davis, H.E., McCorkell, L., Vogel, J.M. et al. Long COVID: major findings, mechanisms and recommendations. Nat Rev Microbiol 21, 133–146 (2023). https://doi.org/10.1038/s41579-022-00846-2