P8106: Data Science II
Yimin Chen
UNI: yc4195

Midterm Report

Introduction:

This dataset is gathered by three cohort studies and aims to get a better understanding about variables that indicate recovery time from COVID-19 illness. Outcome/response variable is time to recovery and predictors are some personal characteristics related to patient in these studies. A random sample of 2000 over 10,000 patients in total was draw for my data analysis.

Also, I set the random seed number to 4195 which is the last four digit of my UNI for reproducibility. The data is preprocessed by mutating the variable ‘study’ from a character variable to a numeric variable and converting some variables into factor variables. For better model building, I exclude the variable id from all predictors and then drew a numeric analysis table to get an overview of tidied data (**Table 1**).

Exploratory analysis and data visualization:

Our COVID-19 dataset was randomly divided into two parts: training set (70%) and test dataset (30%). Eight models were fitted using the training data and the root mean squared error (RMSE) was calculated for each model based on the test data. The RMSE is a measure of the distance between the predicted values and the actual values, where lower values indicate better model performance. I measured the root mean squared error to quantify the extent to which the predicted response value for a given observation is close to the true response value for that observation. Also, 10-fold cross validation repeating 5 times was employed for all models fitted.

- The correlation plot based on train dataset indicates that there is no apparent correlation between most predictors, while there is multicollinearity between the following: bmi and height&weight; SBP and having hypertension; study group 2 and 3(**Figure 2**).
- Also, by using the featurePlot() function lattice plots were generated to display the correlation between COVID-19 recovery time and each possible predictors (**Figure 3**). From the lattice plot, we notice:
    - People who vaccinated tend to have shorter recovery time than people who not.
    - People who show severe COVID-19 infection tend to have longer recovery time than people who are not severe.
    - People in study B tend to have shorter recovery time than people in study A and C.
    - People with hypertension tend to have longer recovery time than people who not.
    - People with diabetes tend to have longer recovery time than people who not.
    - Male tends to have longer recovery time than female.
    - White tend to have the shortest recovery time among all races.

- People who never smoking tend to have shorter recovery time compared to former and current smoker.
- There is no apparent correlation between remaining continuous numeric variables like LDL, weight, bmi, SBP, age, height.

Model training:

For model training, a **Linear Regression Model** was initially employed to predict the recovery time. However, predictors may exhibit multicollinearity with one another. For example, bmi has a relationship with height and weight. Therefore, it's inappropriate to use linear regression model under such scenario.

Hence, alternative techniques, such as Ridge regression, Lasso, Elastic net, Principal components regression (PCR), and Partial least squares (PLS), were also employed to address the possibility of collinearity among predictors. In contrast to the linear model that used all predictors, these five models above were fine-tuned with a selected range of tuning parameters, and the optimal one was determined using the function like 'bestTune' by 10 fold cross-validation.

Ridge regression and **Lasso** are worth mentioning since they use 2 different penalties /techniques that shrinks the coefficient estimates towards zero, which reduces variance. Alpha was held at a value of 0 as the ridge penalty and was held at a value of 1 as the lasso penalty. **Elastic net model** then was also performed, which is more effective to deal with groups of highly correlated predictors. I use same package "glmnet" but set different tuneGrid for these three models. For ridge, tuneGrid created a grid with alpha equals to 0 and lambda that are exponentially spaced between e^-10 and e^6 over a range of 200 values **(Figure 4)** ; for Lasso, tuneGrid created a grid with alpha equals to 1 and lambda that are exponentially spaced between e^-5 and e^-2 over a range of 100 values **(Figure 5) ;** for Elastic net, grid with alpha ranges from 0 to 1 and lambda that are exponentially spaced between e^-2 and e^2 over a range of 50 values was used **(Figure 6)**.

For **PCR** and **PLS**, we made different assumptions. We set the number of principal components (ncomp) to 19 for PCR **(Figure 7)** and 15 for PLS **(Figure 8)**. We centered and scaled the data for both models to ensure that each predictor variable was on the same scale and had a mean of zero.

What's more, to account for non-linear relationships between predictors and the outcome, we developed a **Generalized Additive Model (GAM)** and a **Multivariate Adaptive Regression Splines Model (MARS)**. For the GAM, we used the smoothing spline method to build each block for every continuous predictor with mentioning "method = gam" in code. For categorical variables like "race," we treated them as dummy variable by setting 'race1' which represents White as reference group and so on. This model automatically accounts for non-linear relationships that standard linear regression ignores. As for the MARS in which method was set as "earth", we explored its effectiveness in modeling non-linear relationships. The tuning parameter of the fitted MARS model is automatically selected by the "caret" training method as degree of features =1 and nprune = 11 to minimize prediction error **(Figure 10)**. In addition, the MARS model gives a rank of importance for predictors in predicting the response. For instance, in our model the most important one is bmi.

Results:

After conducting the EDA and model training, we used a resamples function to pull all the results together and then drew a summarized boxplot to check which one is the best model. GAM was picked as the best model because it had both the smallest mean and median RMSE, a value of 20.66361 and 21.80269 respectively (**Figure 11**).

For the GAM model, all predictors were included and "s()" function was applied for continuous variables which meant these variables had been smoothed using splines. The optimal tuning parameter for this model was recorded in "gam.fit$bestTune". What's more, performing feature selection which refers to the process of deciding which predictors are most crucial for forecasting the outcome showed smaller RMSE than no feature selection (**Figure 9**). In final model, category 'race1' (White) , 'smoking0' (Never smoke) , 'hypertension0' (No hypertension), 'diabetes0' (No diabetes), 'vaccine0' (Not vaccinated), 'severity1' (Not severe), 'study1' (study A) had been treated as reference groups. Also, parametric coefficients were used for explaining these categorical variables at 5% level of significance. For instance, Male tends have a 6.03913 day shorter COVID-19 recovery time than female.

After adjusting for the number of predictor variables, the adjusted R-sq adjusted term 0.383 represents the 38.3% of the variance in the outcome variable is explained by this GAM model. The GCV score, measures the model's total goodness of fit when takes the model's complexity into consideration, is 441.95 in my GAM model. Also, from the final model we can see the predictors bmi and weight matter more than other continuous variables.

Conclusion:

The aim of this project is to develop a prediction model for the duration of COVID-19 patient recovery, with the ultimate goal of comprehending the variables that affect recovery time and identifying significant risk factors for prolonged recovery periods.

Through the implementation of these models, we have gained valuable insights into key predictors that could be used to inform future self-prevention strategies and healthcare services. Out of all the models utilized, the GAM model provided the most accurate forecast for total COVID-19 recovery time. Analysis of the GAM results indicated that body mass index (BMI) and weight are the two significant predictors. Given this, maintaining a healthy lifestyle is crucial during pandemics, as managing weight and maintaining a healthy BMI may potentially reduce recovery time.

Appendix

Table 1: Data Summary

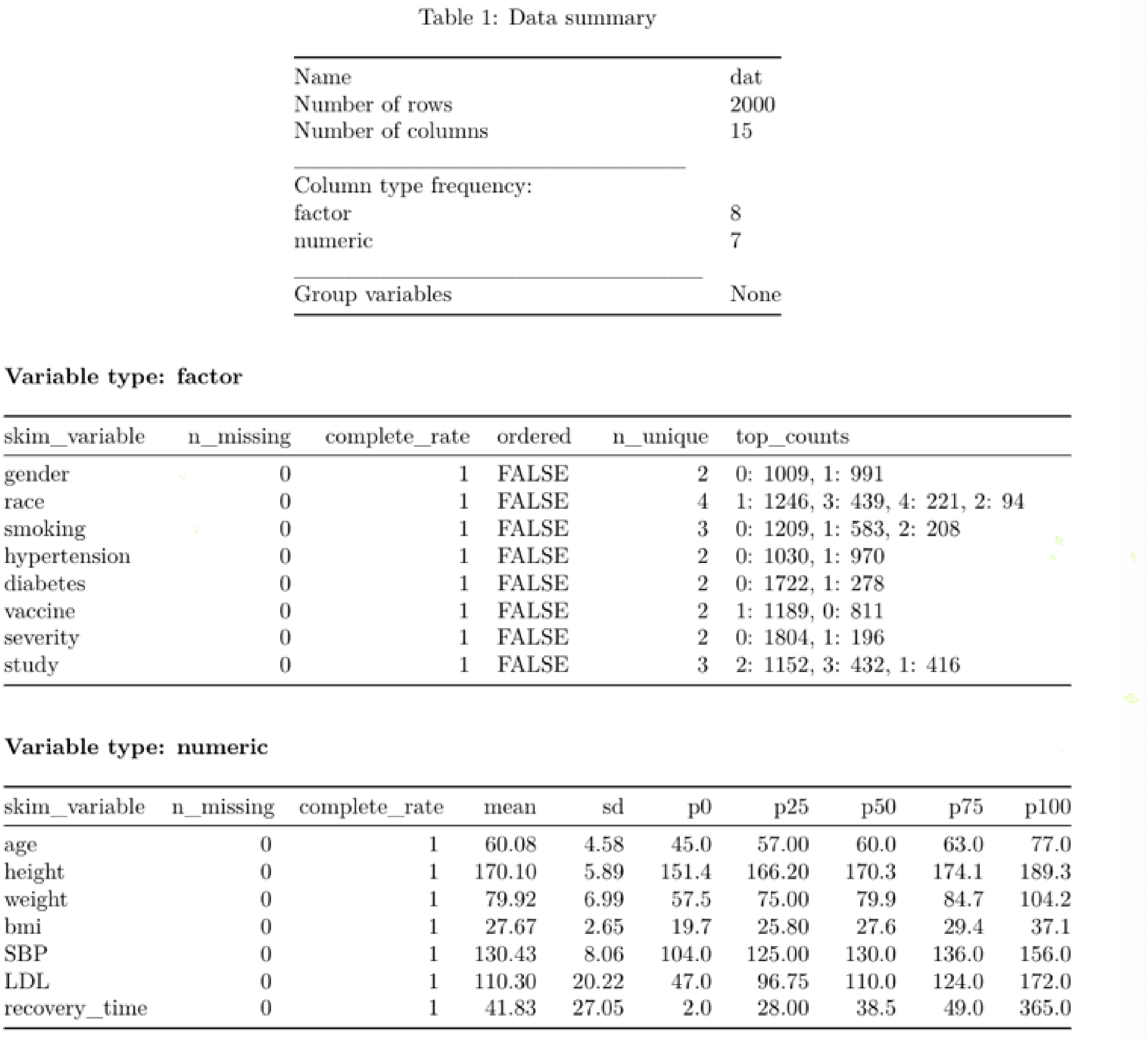


Table 1: Data summary

Name	dat
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: 1009, 1: 991
race	0	1	FALSE	4	1: 1246, 3: 439, 4: 221, 2: 94
smoking	0	1	FALSE	3	0: 1209, 1: 583, 2: 208
hypertension	0	1	FALSE	2	0: 1030, 1: 970
diabetes	0	1	FALSE	2	0: 1722, 1: 278
vaccine	0	1	FALSE	2	1: 1189, 0: 811
severity	0	1	FALSE	2	0: 1804, 1: 196
study	0	1	FALSE	3	2: 1152, 3: 432, 1: 416

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
age	0	1	60.08	4.58	45.0	57.00	60.0	63.0	77.0
height	0	1	170.10	5.89	151.4	166.20	170.3	174.1	189.3
weight	0	1	79.92	6.99	57.5	75.00	79.9	84.7	104.2
bmi	0	1	27.67	2.65	19.7	25.80	27.6	29.4	37.1
SBP	0	1	130.43	8.06	104.0	125.00	130.0	136.0	156.0
LDL	0	1	110.30	20.22	47.0	96.75	110.0	124.0	172.0
recovery_time	0	1	41.83	27.05	2.0	28.00	38.5	49.0	365.0

Figure 2: Correlation Plot

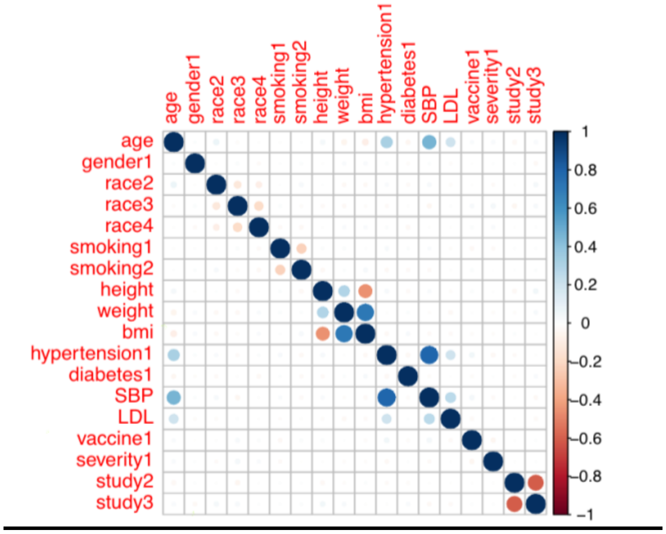


Figure 3: Feature plot which visualizes each predictor's association with the outcome

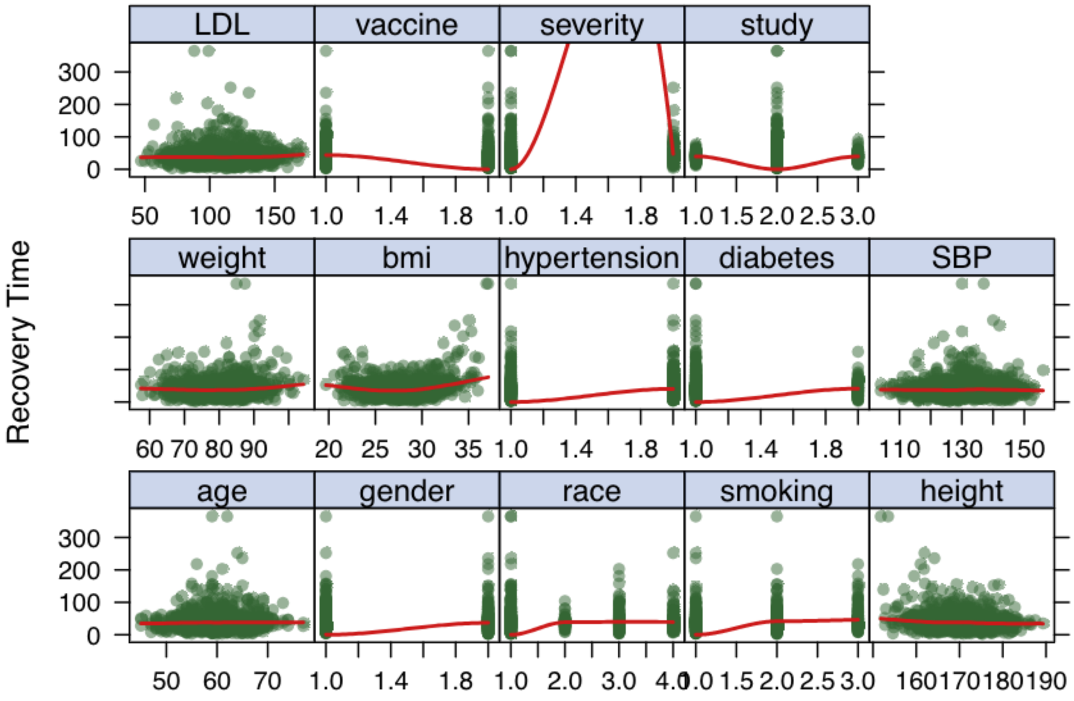


Figure 4: RMSE Differences in Different Lambda in Ridge Regression

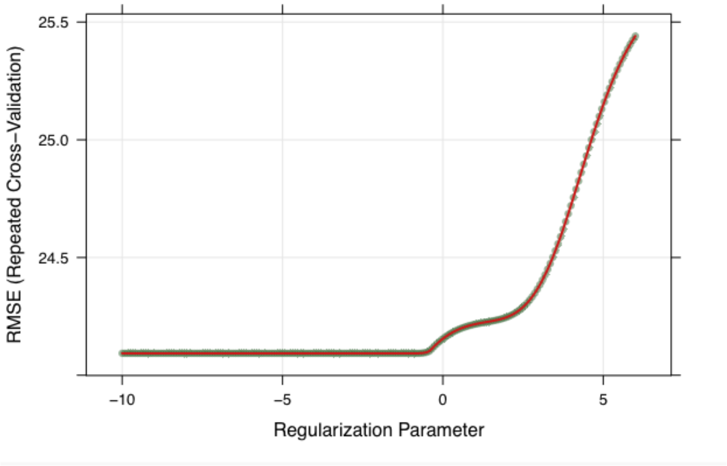


Figure 5: RMSE Differences in Different Lambda in Lasso

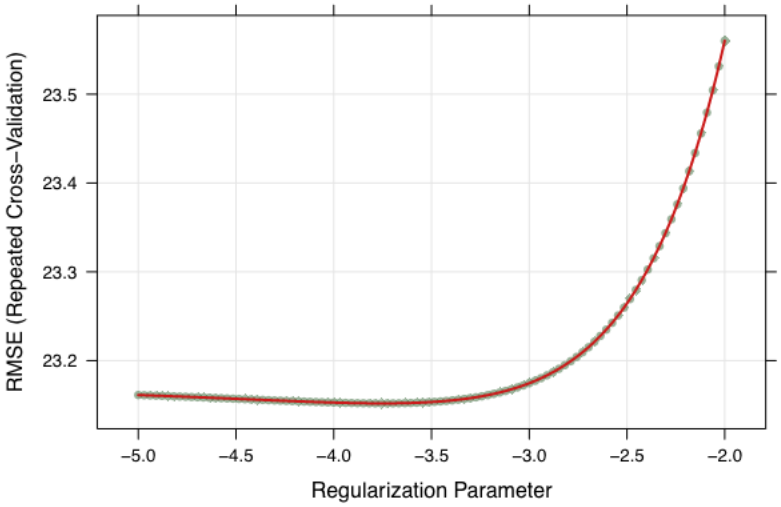


Figure 6: RMSE Differences in Different Lambda in Elastic Net

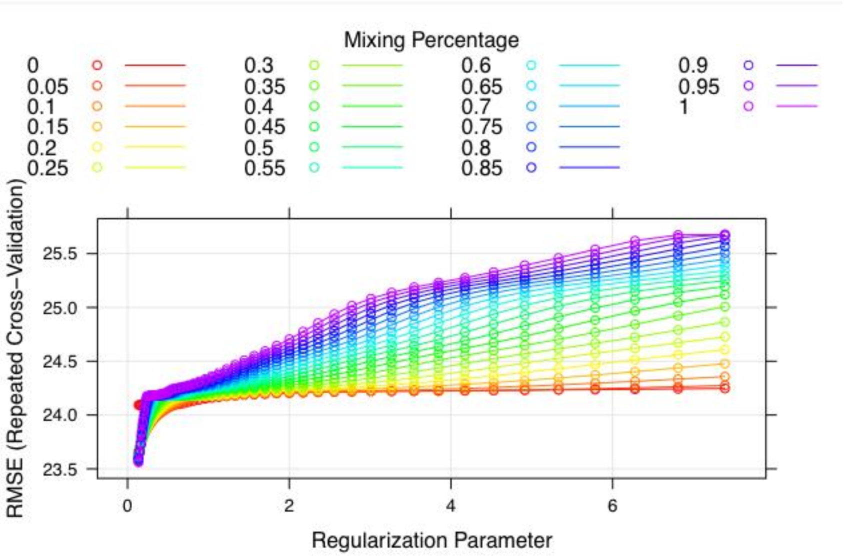


Figure 7: RMSE Differences in Different Number of Components in PCR

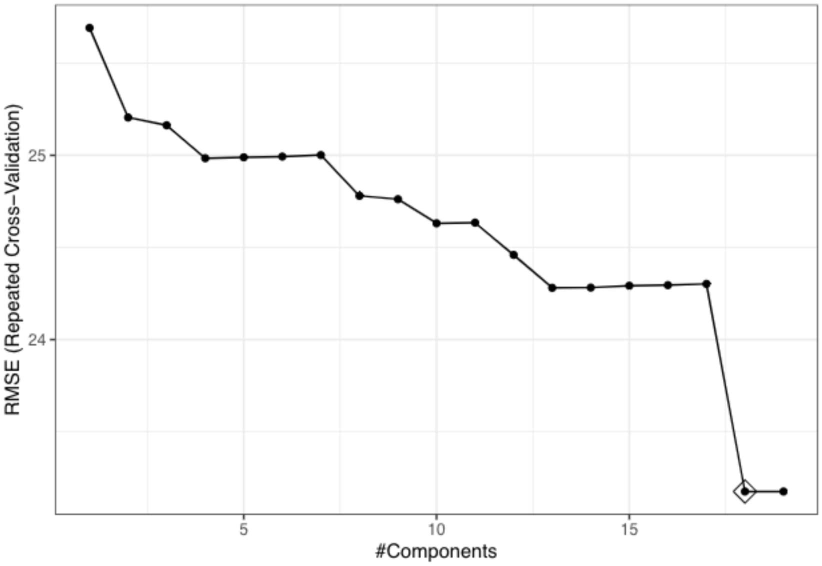


Figure 8: RMSE Differences in Different Number of Components in PLS

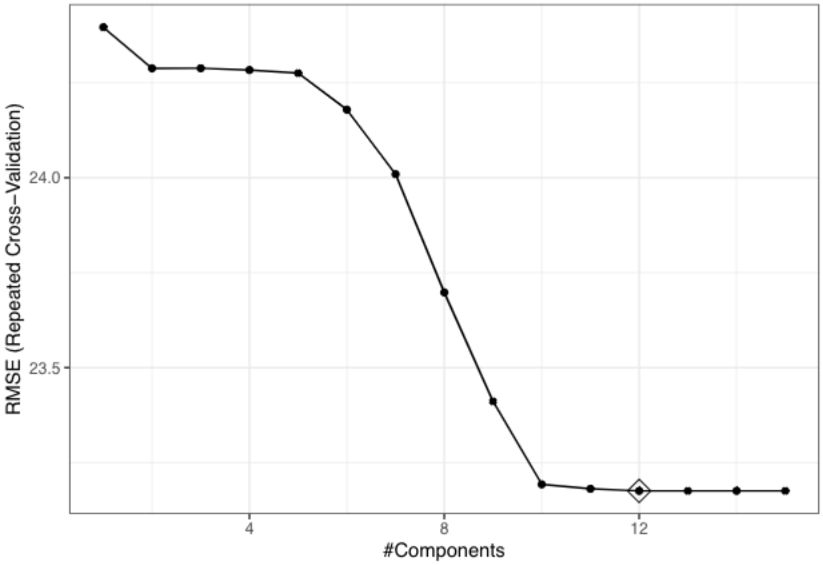


Figure 9: RMSE Differences in Feature Selection and No Feature Selection in GAM

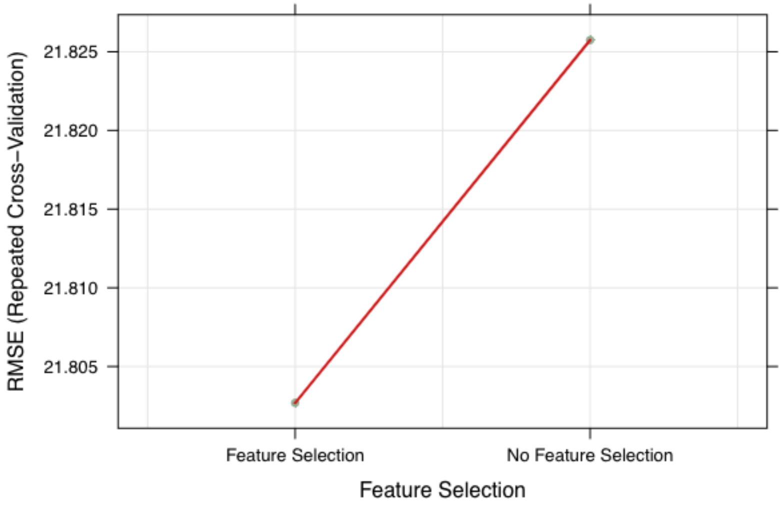


Figure 10: RMSE Differences in Different Number of Terms in MARS

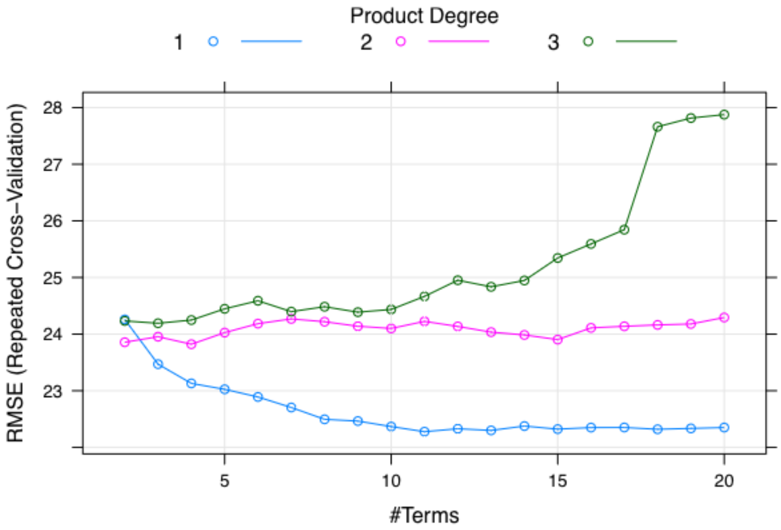
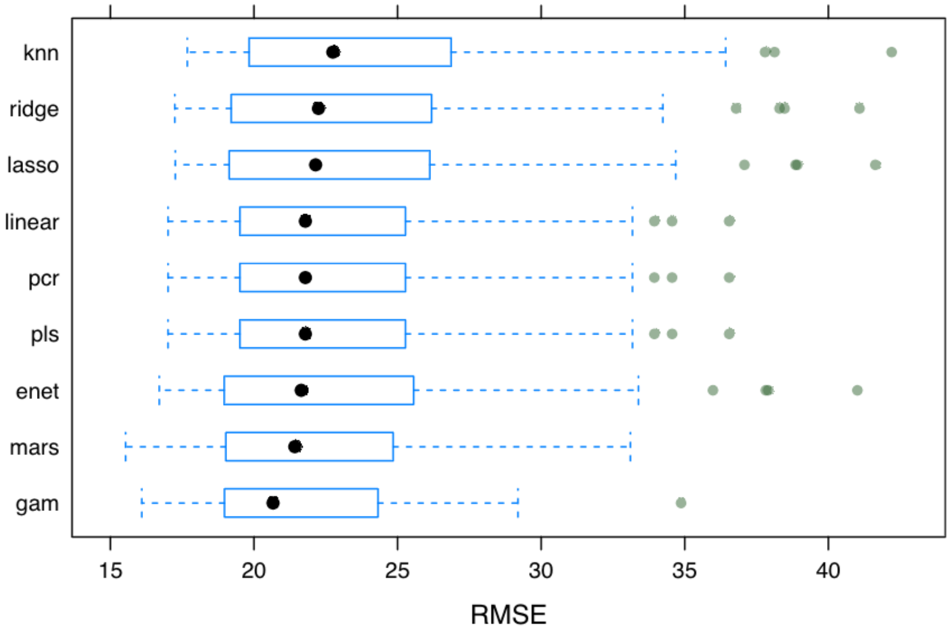


Figure 11: RMSE of Models used for testing


knn
ridge
lasso
linear
pcr
pls
enet
mars
gam
RMSE