P8106 Midterm Project

Zekai Jin (zj2357)
04/04/2023

Introduction

COVID-19, caused by SARS-CoV-2 infection, can lead to a diverse range of clinical manifestations. After infection, most people will have mild to moderate symptoms, and will recover in 2 weeks. However, some people still suffer from prolonged illness, Thus, studying its relationship with other variables will be helpful for improving the recovery rate.

Data Description and exploratory Analysis

The dataset consists of 2000 observations of 14 predictors, 5 of which are binary, 3 are categorical, and 6 are continuous. The outcome is recovery time (in days), which is treated as a continuous variable. A summary of the variables of the dataset is shown in Table 1.

For continuous predictors, a pairwise scatter plot is performed to study their marginal relationship with the outcome (Fig 1). According to the plot, all predictors excluding BMI seem to have no relationship to the outcome, yielding a straight line for all values. BMI relates to the outcome through a "U" like function such that observations with moderate BMI have less recovery time.

For categorical and binary variables, a density plot was performed (Fig.2). Similarly, no obvious difference is seen except for study B. The distribution of recovery time in study B seems to have lower mean and higher variance compared to A and C.

A correlation plot is performed for all predictors including the dummy form of categorical variables (Fig.3). According to the plot, height, weight and BMI are correlated with each other, hypertension is positively correlated to SBP and study B is negatively correlated to study C. Most other predictors are not correlated to others.

Model Training

To address different possible underlying structures of the data, a variety of models were trained, including Linear Model (LM), Elastic Net (Enet), Generalized Linear Model (GLM), Partial Least Squares (PLS), Generalized Additive Model (GAM) and Multiple Adaptive Regression Splines (MARS). Before training, the data was randomly split into training set (n=1600) and testing set (n=400). The training set was used for model training, parameter tuning and model comparison by 10-fold cross validation (CV). The test set was used for estimating the final prediction accuracy. The calculation was done using caret (6.0.93) in R (4.2.2).

Linear Model LM suggests a linear Relationship between response and predictors:

$$\vec{y} = X\vec{\beta} + \vec{\epsilon}$$

Here, $\vec{y}$ is the outcome vector, X is the data matrix and $\vec{\beta}$ is the coefficient vector. $\vec{\epsilon}$ is the error term such that $\epsilon_i \sim iid\ N(0, \sigma^2)$. The coefficient is obtained by minimizing the loss:

$$\left|\left|\vec{y} - X\,\vec{\beta}\right|\right|_2$$

Diagnosis plot suggests that the residual is not normal, which violates the model assumption (Sup.1). Thus, LM is not a good fit of the data.

Elastic Net Enet introduces L1 and L2 penalty to LM, which restricts the number and size of the coefficients by minimizing the following loss function:

$$\left|\left|\vec{y} - X\,\vec{\beta}\right|\right|_2^2 + \lambda\left[(1-\alpha)\left||\beta|\right|_2^2 + \alpha\left||\beta|\right|_1\right]$$

The parameter λ controls the strength of penalty and α controls the balance between L1 and L2 penalty. Both parameters are tuned by 10-fold CV (Fig). The result suggests that no penalty ($\lambda = 0$) yields the best prediction error, which leads to a linear model (Sup.2).

Generalized Linear Model Because the response variable is time for recovery, one may argue that it may not follow normal distribution as LM suggests. Instead, it should follow gamma distribution since recovery time is always positive and right skewed. This yields the following GLM:

$$\log(\vec{u}) = X\,\vec{\beta}$$

Where y_i follows a Gamma distribution with expectation u_i. The coefficient is obtained by maximum likelihood calculated by software.

Partial Least Squares PLS aims to reduce the input dimension by finding the best linear combination of inputs that maximize

$$Var\left(X\,\vec{\phi}\right)Corr^2\left(\vec{y}, X\,\vec{\phi}\right)$$

$$w.r.t. \left|\left|\vec{\phi}\right|\right|_2 = 1, \vec{\phi}^T Corr(X)\vec{\phi} = 0$$

The number of components is selected by the same 10-fold CV as before. The result suggests that 9 components as appropriate (Sup.3).

Generalized Additive Model GAM smooths the relationship between predictors and outcome, while holding the additive structure:

$$g(\vec{u}) = \beta_0 + \sum_{i=1}^{p} f_i(\vec{x_i})$$

The corresponding smooth functions f_i and coefficients are automatically selected by algorithm. In GAM, feature selection improved the RMSE by 0.05 compared to the full model (Sup.4).

Multiple Adaptive Regression Splines MARS is another model for non-linearity by creating a piecewise linear model with interaction term included. The cut points and slopes are selected by software. Two parameters are tuned: the maximum degree of interaction term and the number of terms. The result suggests that a degree 3 model with 6 terms performed the best (Sup.5).

Results

After building 6 different models addressing different problems, model comparison is done using the same 10-fold CV dataset. The result is shown in Fig.4 and Table 2. In general, the differences between models are not very significant. However, models considering non-linearity performed slightly better than others, yielding less RMSE and less variance of RMSE across different datasets. Thus, the MARS is chosen as our final model for this problem.

In order to estimate the performance of the final model, compare the predicted value and true value for the test set. The resulting RMSE is 21.83, meaning that on average, the predicted recovery time will differ the true value by 22 days.

The final model can be summarized as follows:

$$\begin{aligned} u_i = 6.16 + 7.0 \times h(BMI_i - 25.9) + 5.14 \times h(31.1 - BMI_i) \\ + 21.48 \times h(BMI_i - 31.1) \times studyB_i \\ - 5.43 \times h(BMI_i - 31.1) \times h(weight_i - 86.9.1) \times studyB_i \\ + 1.36 \times h(BMI_i - 31.1) \times h(SBP_i - 123) \times studyB_i \end{aligned}$$

Where $E[y_i] = u_i$ and $h(x) = x_+$. The visualization of marginal relationship between some predictors and the outcome is shown in Fig.5.

Conclusion

Based on the result of the MARS model, the recovery time is related to only 4 of the 14 predictors and some of their interaction terms. Among them, the most important predictor is BMI, which is present in all terms. The partial dependence plot showed the suggested relationship between BMI and the outcome: observations with moderate BMI tend to recover faster. This is similar to our findings in EDA (Fig.1).

Study to be B is also important for our prediction. For people with higher BMI (>31.1), if they were in study B, their predicted recovery time will be significantly longer than those in other studies. This is not caused by Study B having more high-BMI observations (Sup.). Thus, there should be some other difference in study design or target population in study B that caused this effect.

SBP and weight alone may have no influence on the outcome, but their interaction with BMI and study B is still beneficial for prediction. Particularly, for observations in study B with high BMI, high systolic blood pressure further prolonged the recovery time.

In conclusion, our dataset of 2000 observations and 14 predictors suggest a non-linear and interactive relationship between BMI, Study, SBP and weight. Moderate BMI around 26, lower blood pressure are both beneficial for recovery. Data in study B has different behavior, which should be further investigated.

Fig.1 Scatter plot of response variable vs. continuous predictors

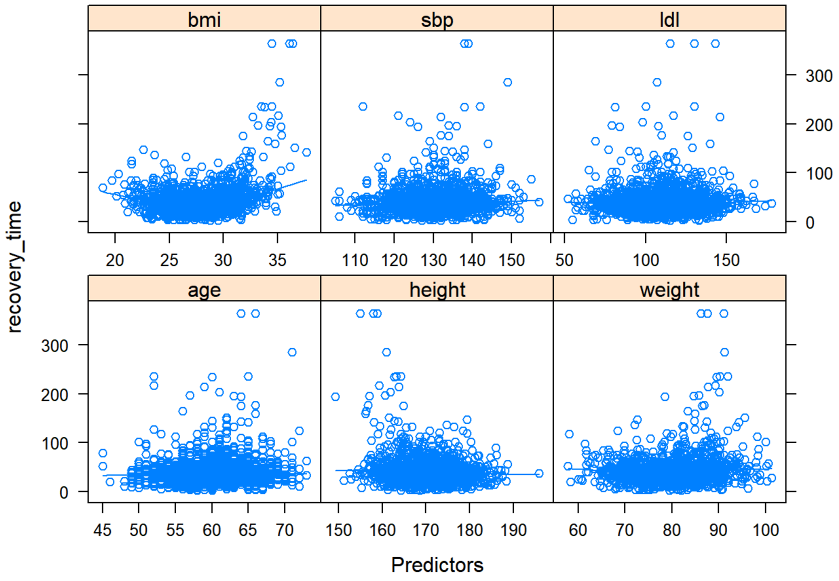


Fig.2 Density plot of response variable among different categorical variables

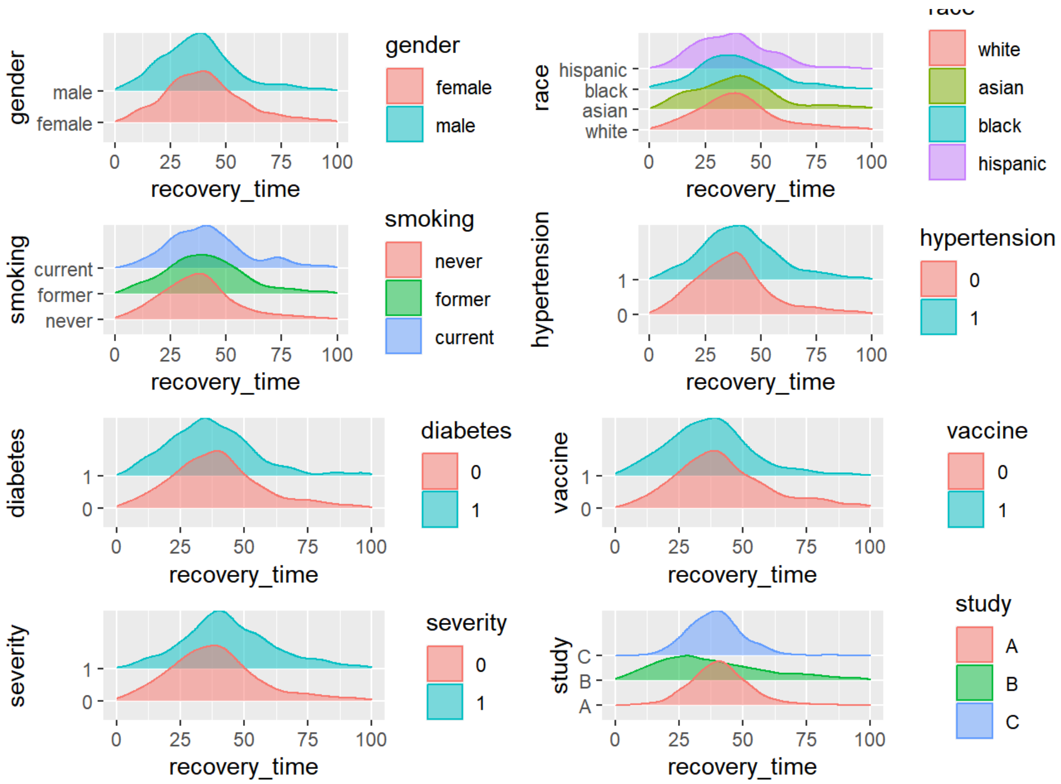


Fig.3 Correlation plot of predictors

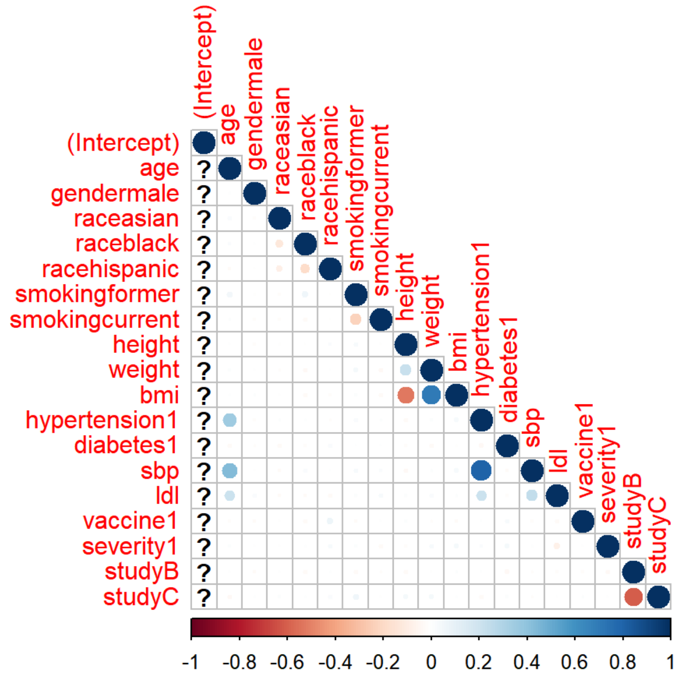


Fig.4 Comparison of RMSE for different models

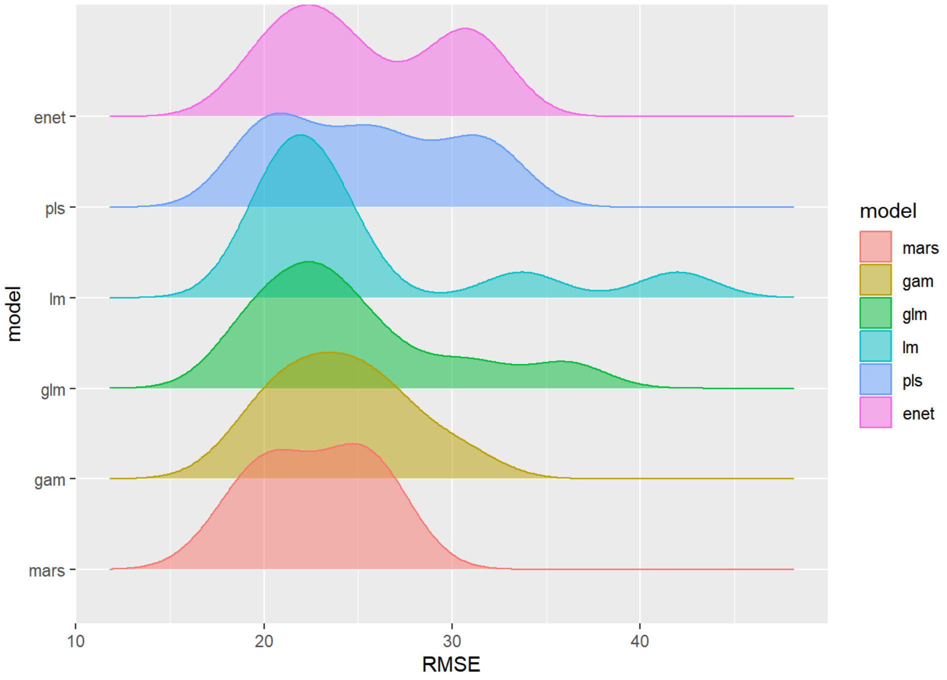


Fig.5 Partial dependence plot of the final MARS model

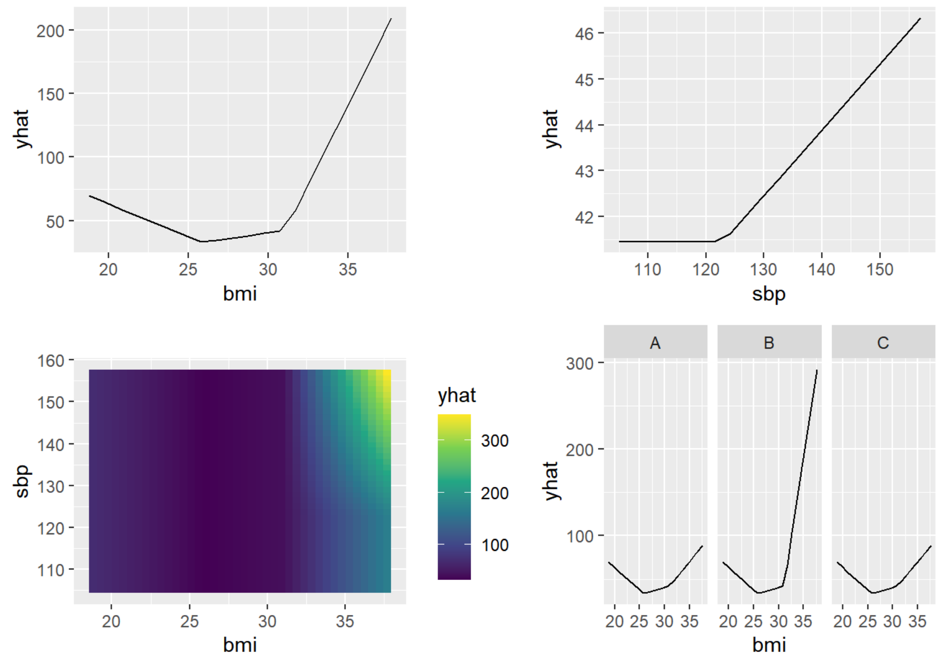


Table 1 Description of the dataset

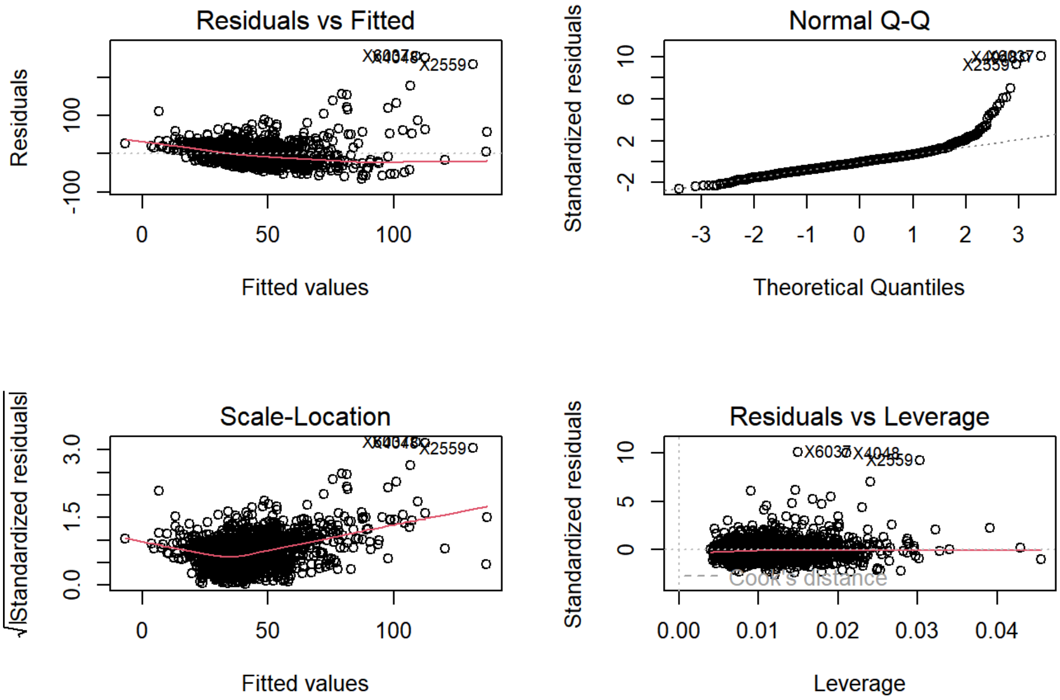
Variable	Type	Mean (prob)	Description
age	continuous	60.14	unit in years
gender	binary	0.478	male encoded as 1
race	nominal	/	white, Asian, black or hispanic
smoking	ordinary	/	no/former/current smokers
height	continuous	170.2	unit in centimeters
weight	continuous	80.00	unit in kilograms
BMI	continuous	27.67	height/weight^2
hypertension	binary	0.476	yes encoded as 1
diabetes	binary	0.152	yes encoded as 1
SBP	continuous	130.2	Systolic blood pressure in mmHg
LDL	continuous	110.4	unit in mg/L
vaccine	binary	0.600	vaccinated encoded as 1
severity	binary	0.109	severe encoded as 1
study	nominal	/	A, B or C
recovery_time	continuous	42.78	unit in days

Table 2 Comparison of RMSE for different models

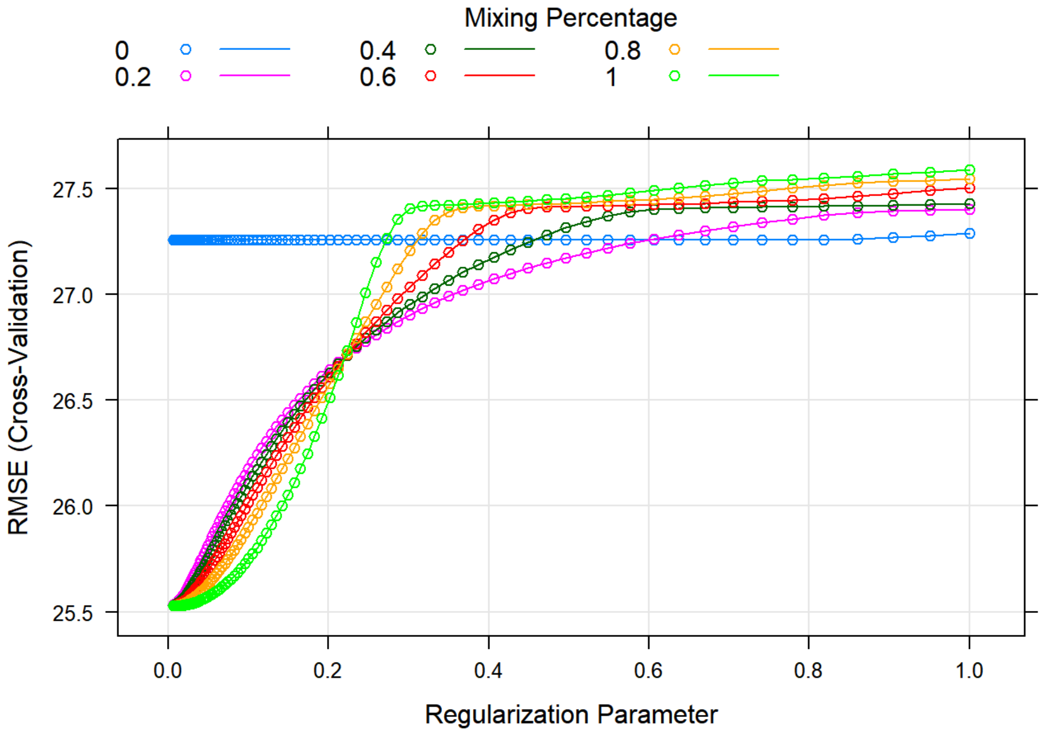
Model	mean (RMSE)	Var (RMSE)
MARS	22.7	3.07
GAM	24.1	3.52
GLM	24.7	5.35
LM	25.2	7.08
PLS	25.4	4.84
Enet	25.5	4.60

Supplementary Files

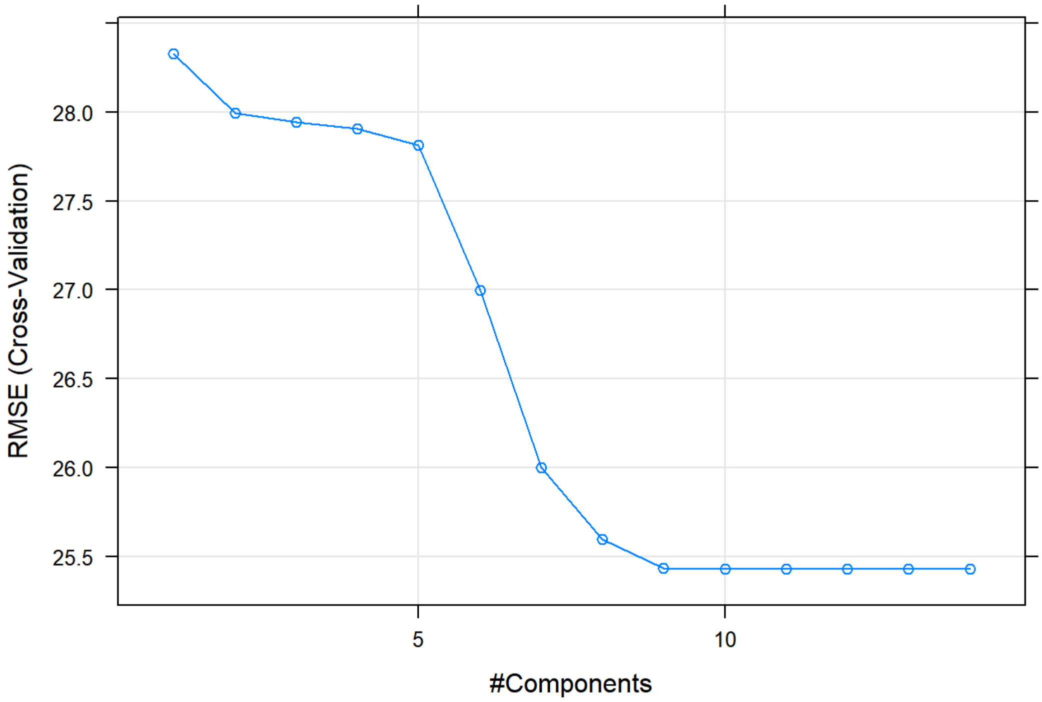
Sup.1 Linear model diagnosis plot



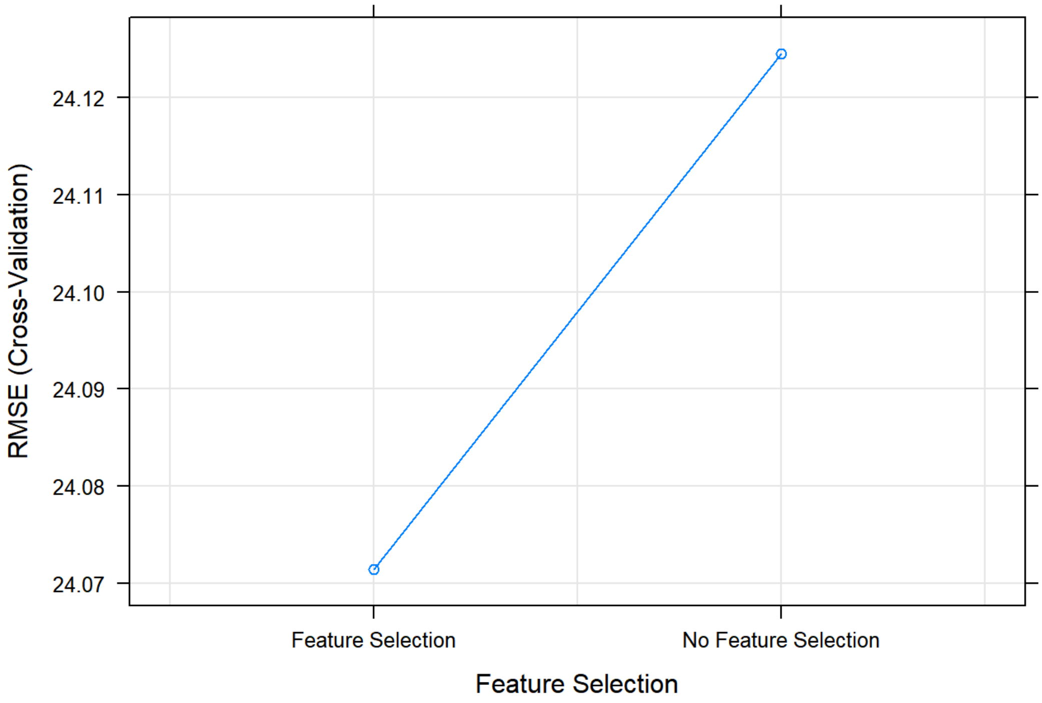
Sup.2 RMSE of Enet tuning grid



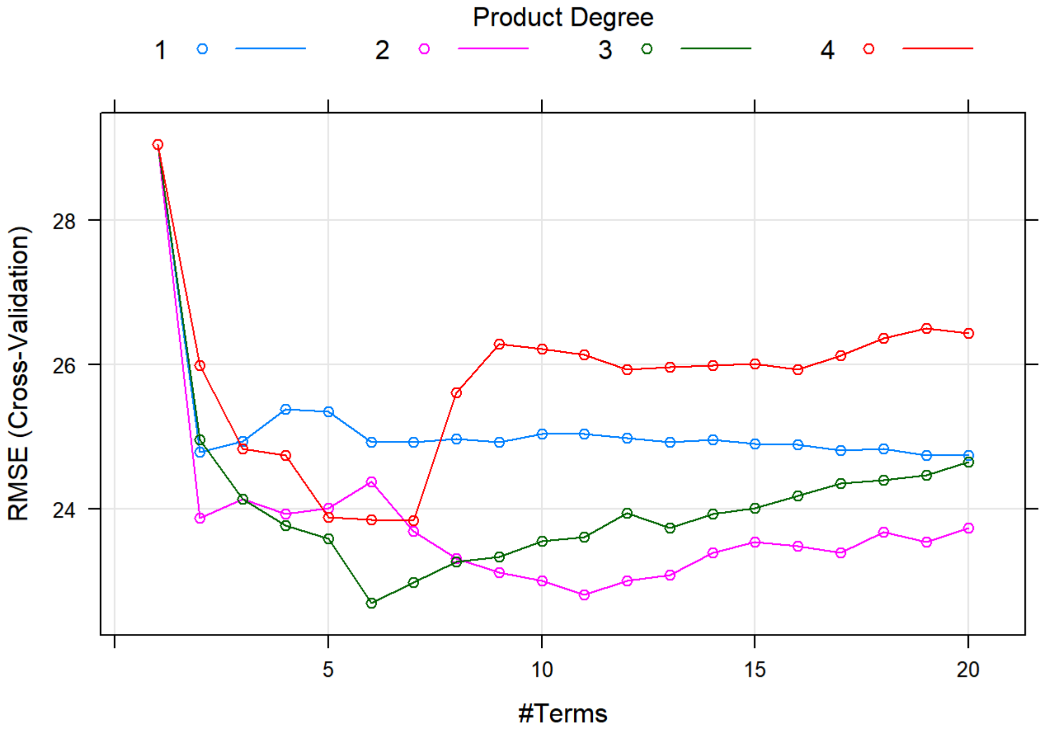
Sup.3 RMSE of PLS for different numbers of components



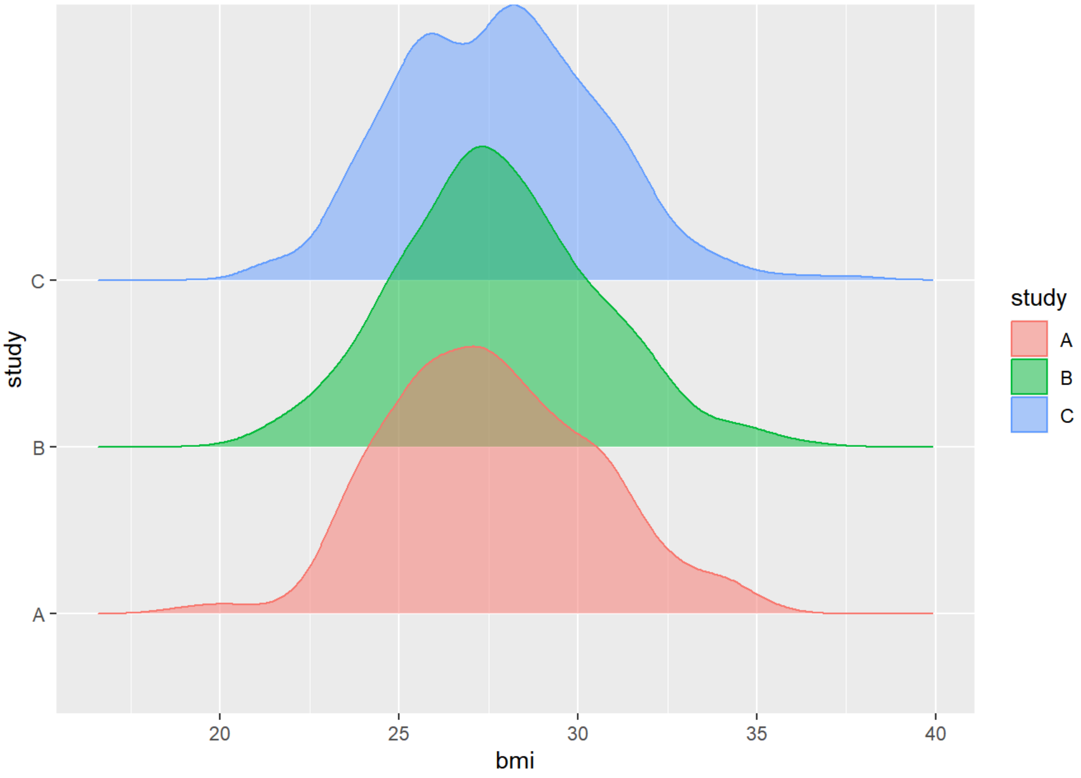
Sup.4 RMSE of GAM with/without feature selection



Sup.5 RMSE of MARS tuning grid



Sup.6 density plot of BMI across different studies


C
B
A
study
bmi
study
A
B
C