Midterm Project Code

Tianshu Liu

Contents

```
library(tidyverse)
library(summarytools)
library(corrplot)
library(caret)
library(vip)
```

```
# import data
load("./recovery.RData")

set.seed(3196)
lts.dat <- dat[sample(1:10000, 2000),]
set.seed(2575)
lincole.dat <- dat[sample(1:10000, 2000),]
set.seed(5509)
amy.dat <- dat[sample(1:10000, 2000),]

dat1 <- lts.dat %>%
  merge(lincole.dat, all = TRUE) %>%
  na.omit() %>%
  select(-id) %>%
  mutate(
    gender = as.factor(gender),
    race = as.factor(race),
    smoking = as.factor(smoking),
    hypertension = as.factor(hypertension),
    diabetes = as.factor(diabetes),
    vaccine = as.factor(vaccine),
    severity = as.factor(severity),
    study = as.factor(study))

dat2 <- lts.dat %>%
  merge(amy.dat, all = TRUE) %>%
  na.omit() %>%
  select(-id) %>%
  mutate(
    gender = as.factor(gender),
    race = as.factor(race),
    smoking = as.factor(smoking),
    hypertension = as.factor(hypertension),
    diabetes = as.factor(diabetes),
    vaccine = as.factor(vaccine),
    severity = as.factor(severity),
    study = as.factor(study))

dat3 <- lincole.dat %>%
  merge(amy.dat, all = TRUE) %>%
  na.omit() %>%
  select(-id) %>%
  mutate(
    gender = as.factor(gender),
    race = as.factor(race),
    smoking = as.factor(smoking),
    hypertension = as.factor(hypertension),
    diabetes = as.factor(diabetes),
    vaccine = as.factor(vaccine),
    severity = as.factor(severity),
    study = as.factor(study))

summary(dat1)
```

```
##       age            gender   race     smoking       height          weight
##   Min.   :45.00   0:1842   1:2372   0:2223   Min.   :151.2   Min.   : 56.70
##   1st Qu.:57.00   1:1781   2: 172   1:1034   1st Qu.:166.2   1st Qu.: 75.40
##   Median :60.00            3: 716   2: 366   Median :170.2   Median : 80.20
##   Mean   :60.06            4: 363            Mean   :170.2   Mean   : 80.13
##   3rd Qu.:63.00                              3rd Qu.:174.2   3rd Qu.: 84.80
##   Max.   :77.00                              Max.   :188.6   Max.   :103.40
##       bmi            hypertension diabetes      SBP             LDL        vaccine
##   Min.   :19.70   0:1891       0:3065     Min.   :102.0   Min.   : 28.0   0:1469
##   1st Qu.:25.80   1:1732       1: 558     1st Qu.:125.0   1st Qu.: 97.0   1:2154
##   Median :27.60                           Median :130.0   Median :110.0
##   Mean   :27.73                           Mean   :130.2   Mean   :110.5
##   3rd Qu.:29.40                           3rd Qu.:136.0   3rd Qu.:124.0
##   Max.   :39.80                           Max.   :158.0   Max.   :174.0
##   severity study    recovery_time
##   0:3289   A: 728   Min.   :  3.00
##   1: 334   B:2171   1st Qu.: 28.00
##            C: 724   Median : 38.00
##                     Mean   : 42.87
##                     3rd Qu.: 49.00
##                     Max.   :365.00
```

```
summary(dat2)
```

```
##       age            gender   race     smoking       height          weight
##   Min.   :45.00   0:1876   1:2350   0:2220   Min.   :151.2   Min.   : 55.90
##   1st Qu.:57.00   1:1719   2: 173   1:1033   1st Qu.:166.1   1st Qu.: 75.50
##   Median :60.00            3: 703   2: 342   Median :170.2   Median : 80.20
##   Mean   :60.19            4: 369            Mean   :170.1   Mean   : 80.18
##   3rd Qu.:63.00                              3rd Qu.:174.1   3rd Qu.: 84.90
##   Max.   :77.00                              Max.   :190.6   Max.   :104.20
##       bmi            hypertension diabetes      SBP             LDL        vaccine
##   Min.   :19.90   0:1871       0:3056     Min.   :101.0   Min.   : 28.0   0:1427
##   1st Qu.:25.90   1:1724       1: 539     1st Qu.:125.0   1st Qu.: 97.0   1:2168
##   Median :27.70                           Median :130.0   Median :111.0
##   Mean   :27.77                           Mean   :130.2   Mean   :110.6
##   3rd Qu.:29.50                           3rd Qu.:136.0   3rd Qu.:125.0
##   Max.   :39.80                           Max.   :158.0   Max.   :178.0
##   severity study    recovery_time
##   0:3248   A: 739   Min.   :  2.00
##   1: 347   B:2160   1st Qu.: 28.00
##            C: 696   Median : 38.00
##                     Mean   : 42.43
##                     3rd Qu.: 49.00
##                     Max.   :365.00
```

```
summary(dat3)
```

```
##       age            gender   race     smoking       height          weight
##  Min.   :45.00   0:1847   1:2337   0:2246   Min.   :151.2   Min.   : 55.90
##  1st Qu.:57.00   1:1775   2: 206   1:1021   1st Qu.:166.0   1st Qu.: 75.10
##  Median :60.00            3: 709   2: 355   Median :170.1   Median : 80.00
##  Mean   :60.19            4: 370            Mean   :170.1   Mean   : 79.94
##  3rd Qu.:63.00                              3rd Qu.:174.1   3rd Qu.: 84.70
##  Max.   :77.00                              Max.   :190.6   Max.   :104.20
##       bmi       hypertension diabetes      SBP             LDL          vaccine
##  Min.   :19.7   0:1894       0:3070   Min.   :101.0   Min.   : 28.0   0:1435
##  1st Qu.:25.8   1:1728       1: 552   1st Qu.:125.0   1st Qu.: 97.0   1:2187
##  Median :27.6                         Median :130.0   Median :111.0
##  Mean   :27.7                         Mean   :130.2   Mean   :110.3
##  3rd Qu.:29.4                         3rd Qu.:135.8   3rd Qu.:124.0
##  Max.   :39.8                         Max.   :157.0   Max.   :178.0
##  severity study    recovery_time
##  0:3268   A: 719   Min.   :  2.00
##  1: 354   B:2173   1st Qu.: 28.00
##           C: 730   Median : 39.00
##                    Mean   : 42.75
##                    3rd Qu.: 49.75
##                    Max.   :365.00
```

```
dat <- dat1
```

1 Data partition

```
# data partition
dat.matrix <- model.matrix(recovery_time ~ ., dat)[ ,-1]

set.seed(2023)
trainRows <- createDataPartition(y = dat$recovery_time, p = 0.8, list = FALSE)

train_dat <- dat[trainRows,]

train_x <- dat.matrix[trainRows,]
train_y <- dat$recovery_time[trainRows]

test_x <- dat.matrix[-trainRows,]
test_y <- dat$recovery_time[-trainRows]
```

2 Exploratory analysis and data visualization

```
# data summary
st_options(plain.ascii = FALSE,
           style = "rmarkdown",
           dfSummary.silent = TRUE,
           footnote = NA,
           subtitle.emphasis = FALSE)
dfSummary(train_dat)
```

2.0.1 Data Frame Summary

train_dat
Dimensions: 2900 x 15
Duplicates: 0

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	age [numeric]	Mean (sd) : 60.1 (4.5) min < med < max: 45 < 60 < 77 IQR (CV) : 6 (0.1)	33 distinct values	: : . : : : : : . . : : : : : : .	2900 (100.0%)	0 (0.0%)
2	gender [factor]	1. 0 2. 1	1468 (50.6%) 1432 (49.4%)	IIIIIIIIII IIIIIIIII	2900 (100.0%)	0 (0.0%)
3	race [factor]	1. 1 2. 2 3. 3 4. 4	1909 (65.8%) 132 (4.6%) 568 (19.6%) 291 (10.0%)	IIIIIIIIIIIII III II	2900 (100.0%)	0 (0.0%)
4	smoking [factor]	1. 0 2. 1 3. 2	1763 (60.8%) 845 (29.1%) 292 (10.1%)	IIIIIIIIIIII IIIII II	2900 (100.0%)	0 (0.0%)
5	height [numeric]	Mean (sd) : 170.2 (6) min < med < max: 151.2 < 170.1 < 188.6 IQR (CV) : 8 (0)	312 distinct values	: : : : . : : . : : : : . : : : : .	2900 (100.0%)	0 (0.0%)
6	weight [numeric]	Mean (sd) : 80.2 (7) min < med < max: 57.1 < 80.3 < 103.4 IQR (CV) : 9.5 (0.1)	361 distinct values	. : . : : : : : : . : : : : . . : : : : : : .	2900 (100.0%)	0 (0.0%)
7	bmi [numeric]	Mean (sd) : 27.8 (2.7) min < med < max: 19.7 < 27.7 < 39.8 IQR (CV) : 3.6 (0.1)	160 distinct values	: . : : : : : . : : : : : : : : : : .	2900 (100.0%)	0 (0.0%)
8	hypertension [factor]	1. 0 2. 1	1514 (52.2%) 1386 (47.8%)	IIIIIIIIII IIIIIIIII	2900 (100.0%)	0 (0.0%)
9	diabetes [factor]	1. 0 2. 1	2446 (84.3%) 454 (15.7%)	IIIIIIIIIIIIIIII III	2900 (100.0%)	0 (0.0%)
10	SBP [numeric]	Mean (sd) : 130.2 (8.1) min < med < max: 104 < 130 < 158 IQR (CV) : 11 (0.1)	54 distinct values	: : . : : : . . : : : : . : : : : : : :	2900 (100.0%)	0 (0.0%)
11	LDL [numeric]	Mean (sd) : 110.3 (19.9) min < med < max: 32 < 110 < 174 IQR (CV) : 27 (0.2)	116 distinct values	. : : : : : : : . : : : : : . : : : : : : .	2900 (100.0%)	0 (0.0%)
12	vaccine [factor]	1. 0 2. 1	1192 (41.1%) 1708 (58.9%)	IIIIIIII IIIIIIIIIII	2900 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
13	severity	1. 0	2619 (90.3%)	IIIIIIIIIIIIIIIIII	2900	0
	[factor]	2. 1	281 (9.7%)	I	(100.0%)	(0.0%)
14	study	1. A	580 (20.0%)	IIII	2900	0
	[factor]	2. B	1750 (60.3%)	IIIIIIIIIIII	(100.0%)	(0.0%)
		3. C	570 (19.7%)	III		
15	recovery_time	Mean (sd) : 43 (30.5)	144 distinct	:	2900	0
	[numeric]	min < med < max:	values	: :	(100.0%)	(0.0%)
		3 < 38 < 365		: :		
		IQR (CV) : 21 (0.7)		: :		
				: : .		

```
skimr::skim_without_charts(train_dat)
```

Table 2: Data summary

Name	train_dat
Number of rows	2900
Number of columns	15
Column type frequency:	
factor	8
numeric	7
Group variables	None

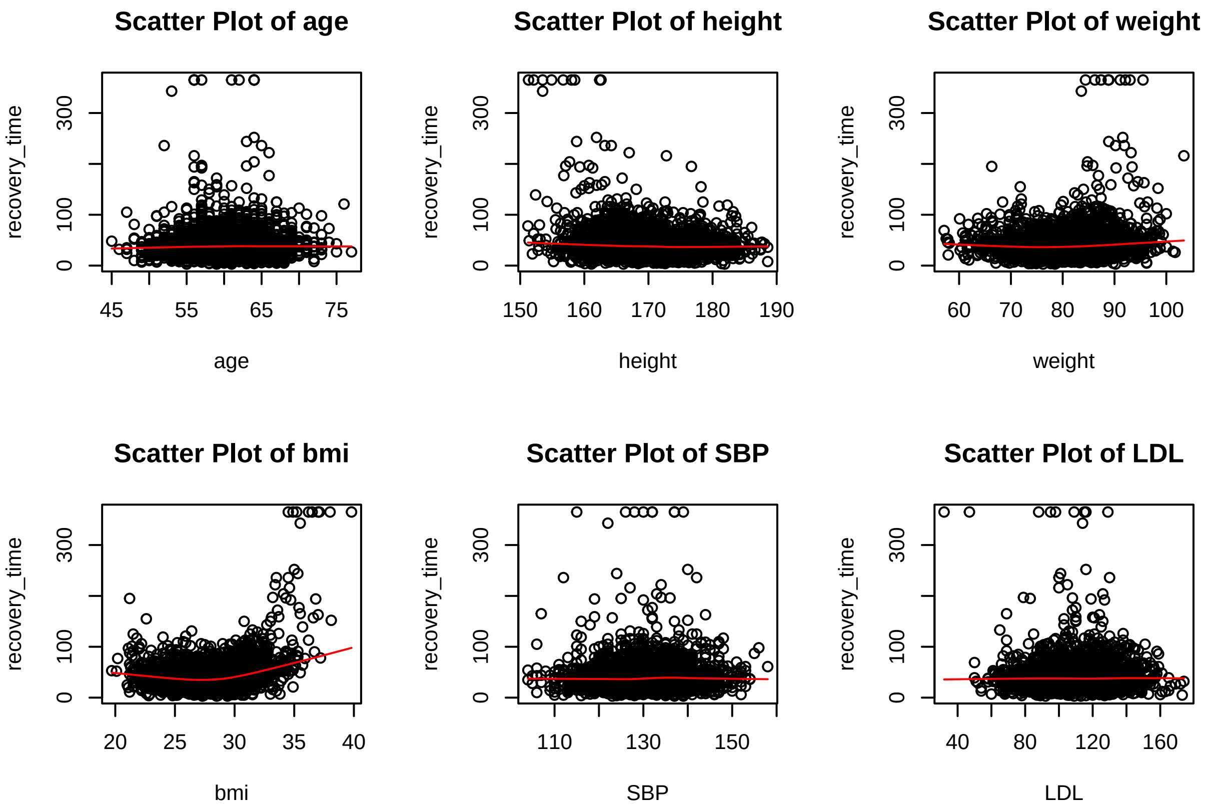
Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
gender	0	1	FALSE	2	0: 1468, 1: 1432
race	0	1	FALSE	4	1: 1909, 3: 568, 4: 291, 2: 132
smoking	0	1	FALSE	3	0: 1763, 1: 845, 2: 292
hypertension	0	1	FALSE	2	0: 1514, 1: 1386
diabetes	0	1	FALSE	2	0: 2446, 1: 454
vaccine	0	1	FALSE	2	1: 1708, 0: 1192
severity	0	1	FALSE	2	0: 2619, 1: 281
study	0	1	FALSE	3	B: 1750, A: 580, C: 570

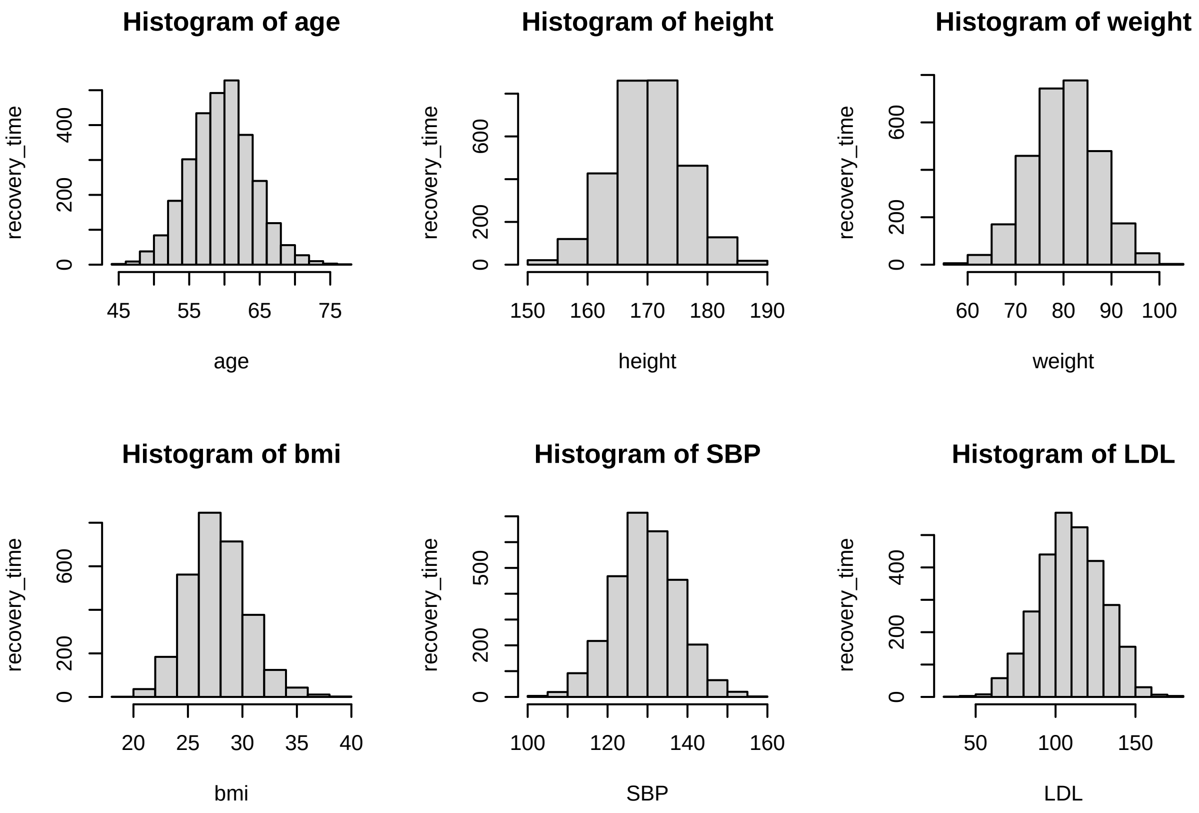
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
age	0	1	60.07	4.51	45.0	57.0	60.00	63.0	77.0
height	0	1	170.17	6.04	151.2	166.1	170.15	174.1	188.6
weight	0	1	80.20	7.00	57.1	75.4	80.30	84.9	103.4
bmi	0	1	27.76	2.73	19.7	25.9	27.70	29.5	39.8
SBP	0	1	130.19	8.08	104.0	125.0	130.00	136.0	158.0
LDL	0	1	110.27	19.87	32.0	97.0	110.00	124.0	174.0
recovery_time	0	1	43.02	30.51	3.0	28.0	38.00	49.0	365.0

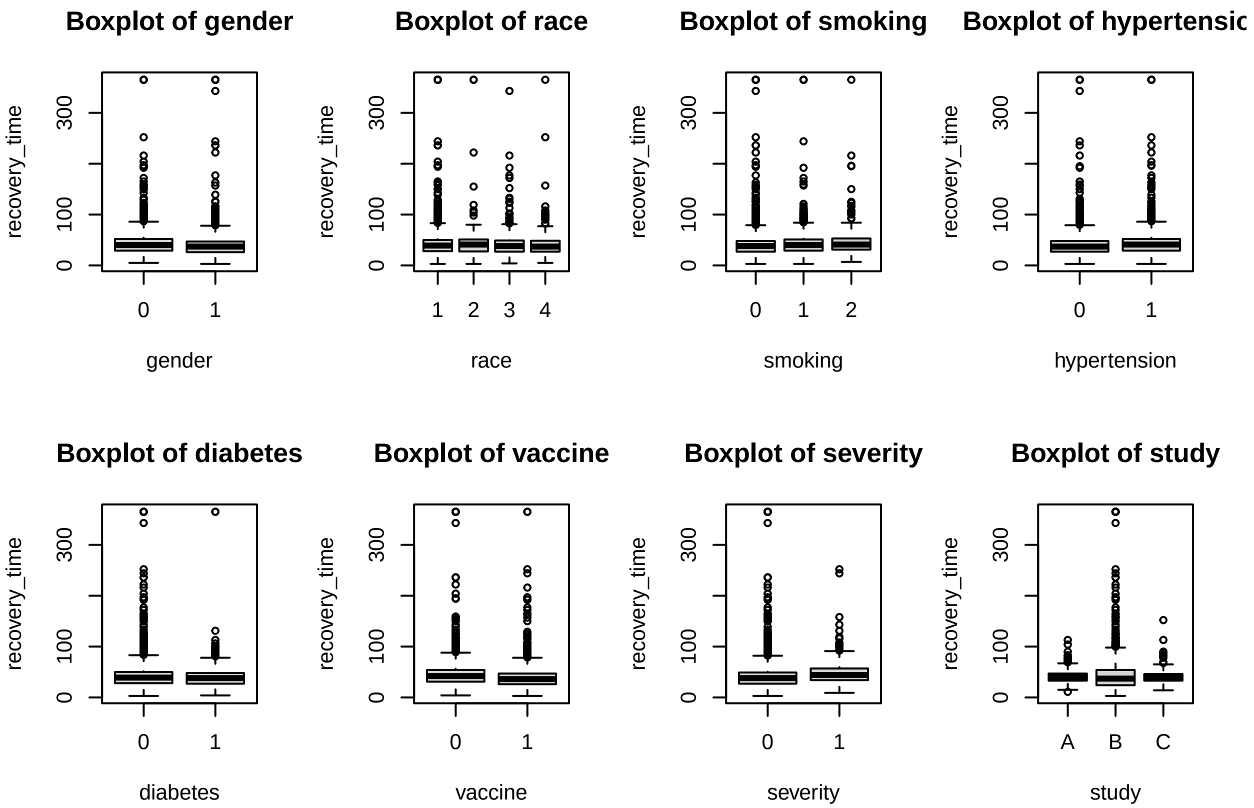
```
####################################################################
## Remember to edit the next chunk if you do any modification here:)
####################################################################

# EDA
# library(GGally)
# ggpairs(dat)

cts_var = c("age", "height", "weight", "bmi", "SBP", "LDL")
fct_var = c("gender", "race", "smoking", "hypertension", "diabetes", "vaccine", "severity", "study")
# featurePlot(x = traindataset1[ ,1:14],
#              y = traindataset1[ ,15],
# plot = "scatter",
# span = .5,
# labels = c("Predictors", "Recovery Time"), type = c("p", "smooth"))

# scatter plot of continuous predictors
par(mfrow=c(2, 3))
for (i in 1:length(cts_var)){
  var = cts_var[i]
  plot(recovery_time~train_dat[,var],
       data = train_dat,
       ylab = "recovery_time",
       xlab = var,
       main = str_c("Scatter Plot of ", var))
  lines(stats::lowess(train_dat[,var], train_dat$recovery_time), col = "red", type = "l")
}
```



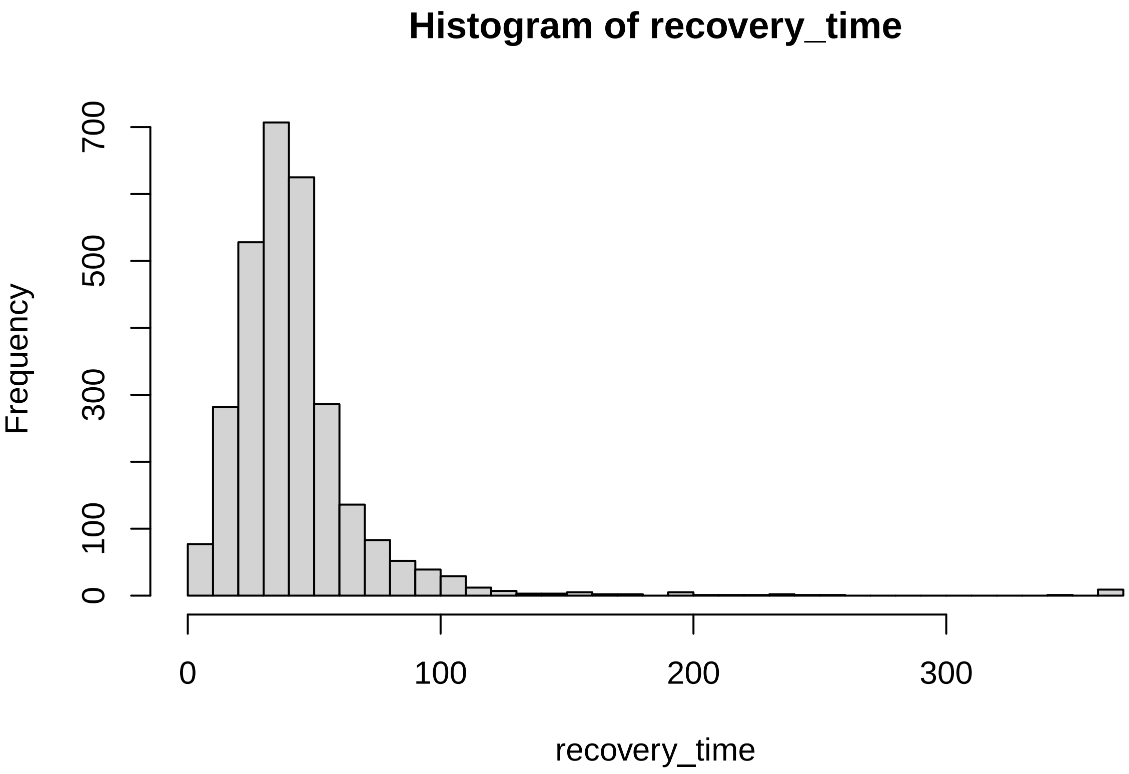
```
for (i in 1:length(cts_var)){
  var = cts_var[i]
  hist(train_dat[,var],
       ylab = "recovery_time",
       xlab = var,
       main = str_c("Histogram of ", var))
}
```



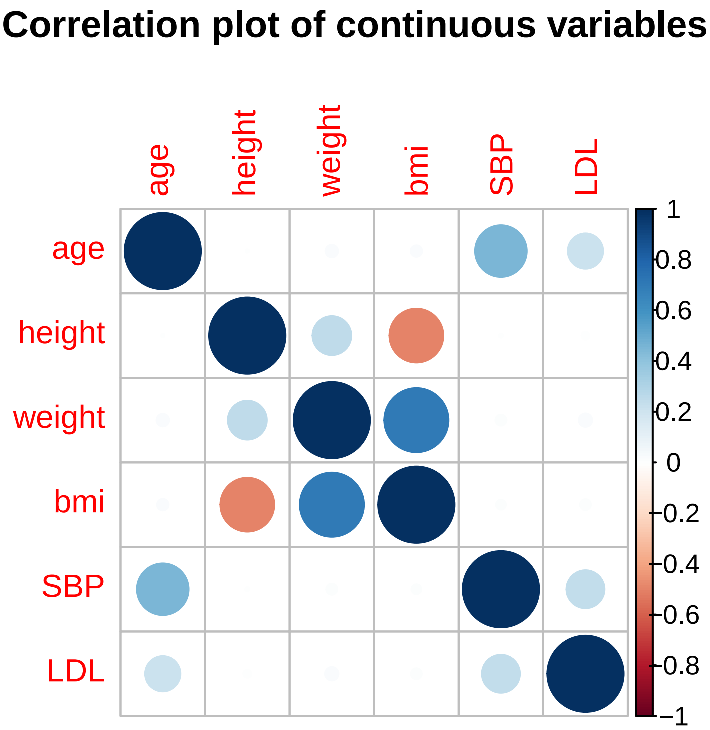
```
# boxplot of categorical predictors
par(mfrow=c(2, 4))
for (i in 1:length(fct_var)){
  var = fct_var[i]
  plot(recovery_time~train_dat[,var],
       data = train_dat,
       ylab = "recovery_time",
       xlab = var,
       main = str_c("Boxplot of ", var))
}
```



```r
# histogram of response
par(mfrow=c(1, 1))
hist(train_dat$recovery_time,
     breaks = 50,
     main = "Histogram of recovery_time",
     xlab = "recovery_time")
```



```
# correlation
par(mfrow=c(1, 1))
corrplot(cor(train_dat[,cts_var]), method = "circle", type = "full",
         title = "Correlation plot of continuous variables",
         mar = c(2, 2, 4, 2))
```



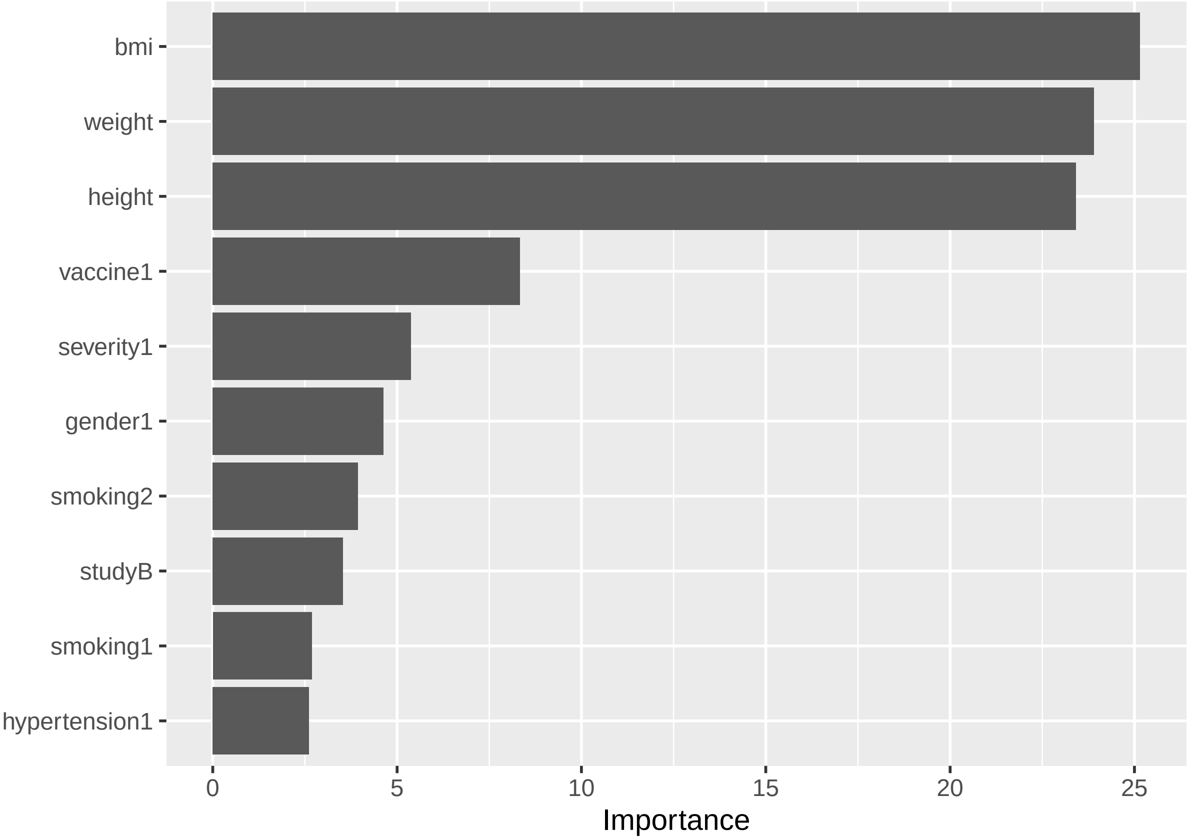
3 Model Training

3.1 Linear Model

```
ctrl1 <- trainControl(method = "cv", number = 10)
set.seed(3196)

lm.fit <- train(train_x, train_y,
                method = "lm",
                trControl = ctrl1)

coef(lm.fit$finalModel)
```

```
##   (Intercept)           age       gender1         race2         race3
## -3.190120e+03  1.163953e-01 -4.443893e+00  2.189010e+00 -6.599719e-01
##         race4      smoking1      smoking2        height        weight
## -1.156806e+00  2.905693e+00  6.427376e+00  1.866280e+01 -2.014323e+01
##           bmi hypertension1     diabetes1           SBP           LDL
##  6.056969e+01  4.165589e+00 -1.152370e+00 -7.863399e-02 -4.215262e-02
##      vaccine1     severity1        studyB        studyC
## -8.133542e+00  8.747096e+00  4.368587e+00 -6.869681e-01
```

```
vip(lm.fit$finalModel)
```



3.2 LASSO

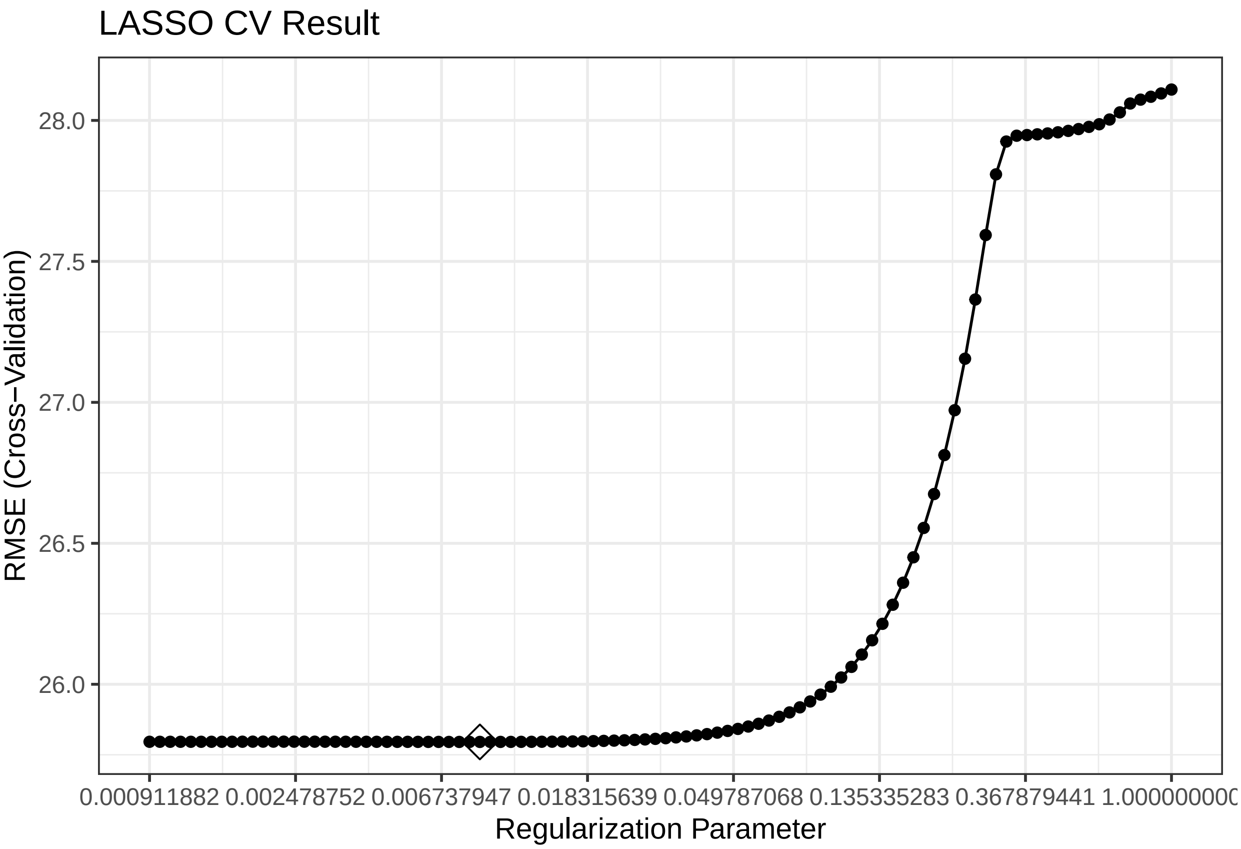
```
set.seed(3196)
lasso.fit <- train(train_x, train_y,
                   method = "glmnet",
                   tuneGrid = expand.grid(
                     alpha = 1,
                     lambda = exp(seq(0, -7, length=100))),
                   trControl = ctrl1)

lasso.fit$bestTune
```

```
##    alpha      lambda
## 33     1 0.008761626
```

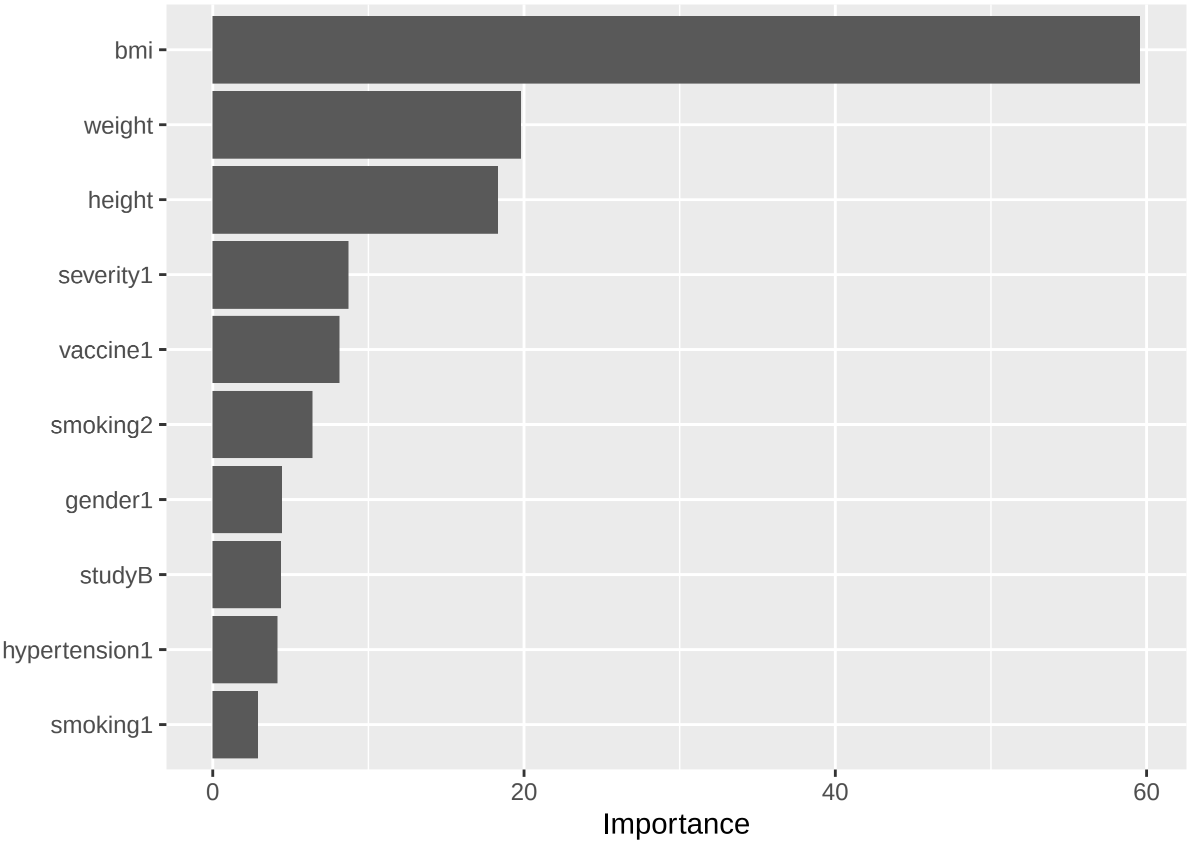
```
coef(lasso.fit$finalModel, s = lasso.fit$bestTune$lambda)
```

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                          s1
## (Intercept)   -3.064684e+03
## age            1.121503e-01
## gender1       -4.430103e+00
## race2          2.178570e+00
## race3         -6.668347e-01
## race4         -1.121978e+00
## smoking1       2.881801e+00
## smoking2       6.347112e+00
## height         1.791867e+01
## weight        -1.935584e+01
## bmi            5.831981e+01
## hypertension1  4.085224e+00
## diabetes1     -1.158180e+00
## SBP           -7.271934e-02
## LDL           -4.182399e-02
## vaccine1      -8.155397e+00
## severity1      8.695698e+00
## studyB         4.364254e+00
## studyC        -6.597810e-01
```

```
ggplot(lasso.fit, highlight = TRUE) +
  labs(title="LASSO CV Result") +
  scale_x_continuous(trans='log',n.breaks = 10) +
  theme_bw()
```



```
ggsave("./figure/lasso_cv.jpeg", dpi = 500)

vip(lasso.fit$finalModel)
```



3.3 Ridge

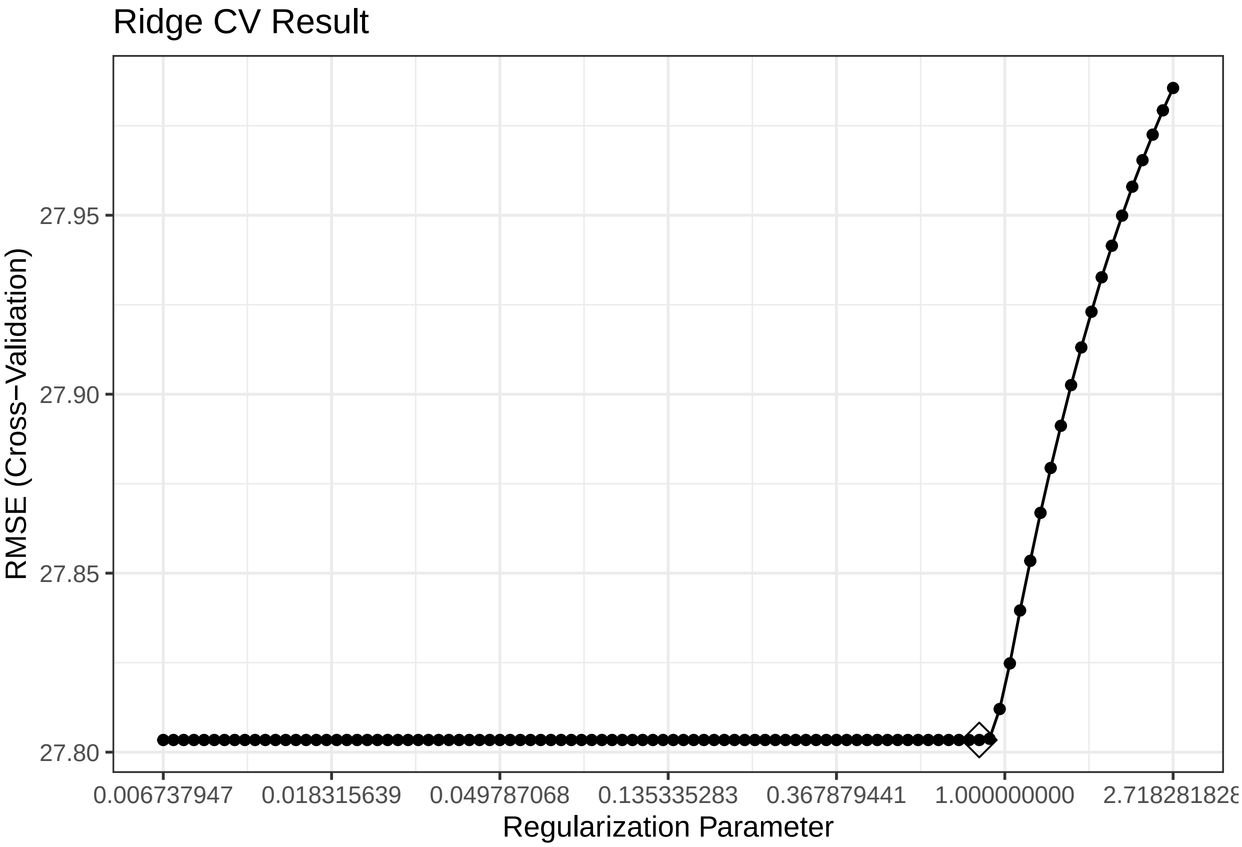
```
set.seed(3196)
ridge.fit <- train(train_x, train_y,
                   method = "glmnet",
                   tuneGrid = expand.grid(alpha = 0,
                                          lambda = exp(seq(1, -5, length=100))),
                   trControl = ctrl1)

ridge.fit$bestTune
```

```
##    alpha    lambda
## 81     0 0.8594049
```

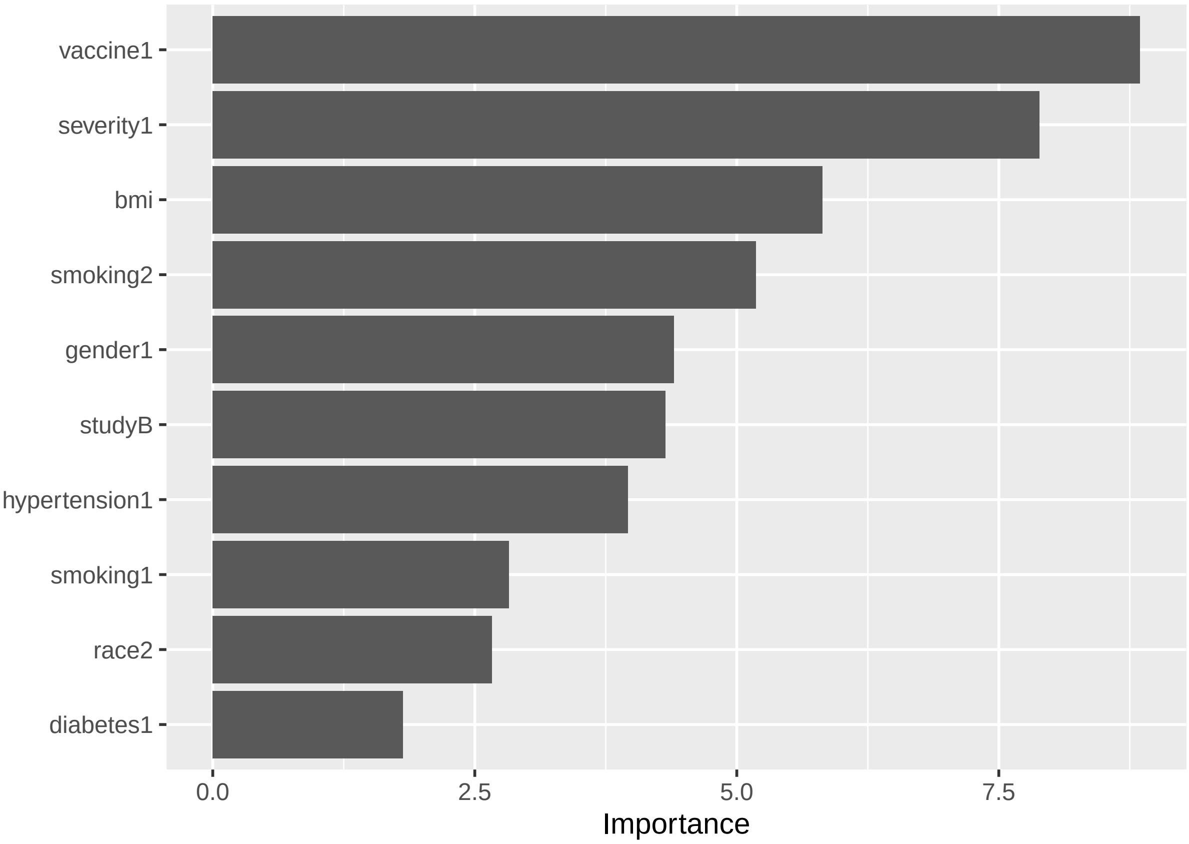
```
coef(ridge.fit$finalModel, s = ridge.fit$bestTune$lambda)
```

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                        s1
## (Intercept)   -131.33806374
## age              0.09731228
## gender1         -4.40320528
## race2            2.66527141
## race3           -1.32710400
## race4           -1.12570977
## smoking1         2.82624366
## smoking2         5.18400128
## height           0.60404463
## weight          -1.01341715
## bmi              5.81922510
## hypertension1    3.96367066
## diabetes1       -1.81677375
## SBP             -0.06303616
## LDL             -0.04440780
## vaccine1        -8.84608080
## severity1        7.88676978
## studyB           4.32156225
## studyC          -0.51357417
```

```
ggplot(ridge.fit,highlight = TRUE) +
  scale_x_continuous(trans='log', n.breaks = 6) +
  labs(title="Ridge CV Result") +
  theme_bw()
```



```
ggsave("./figure/ridge_cv.jpeg", dpi = 500)

vip(ridge.fit$finalModel)
```



3.4 Elastic Net

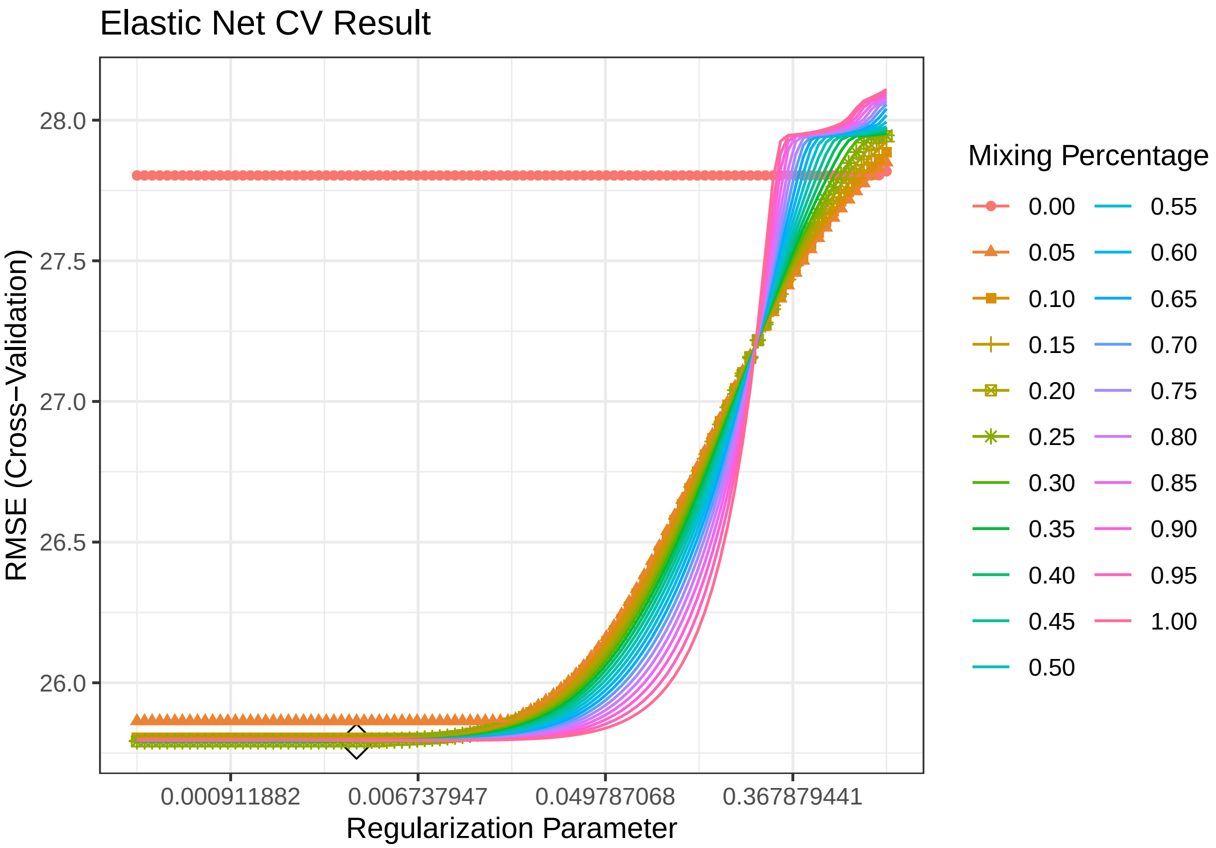
```
set.seed(3196)
enet.fit <- train(train_x, train_y,
                  method = "glmnet",
                  tuneGrid = expand.grid(
                    alpha = seq(0, 1, length = 21),
                    lambda = exp(seq(0, -8, length = 100))),
                  trControl = ctrl1)

enet.fit$bestTune
```

```
##     alpha      lambda
## 530  0.25 0.003494498
```

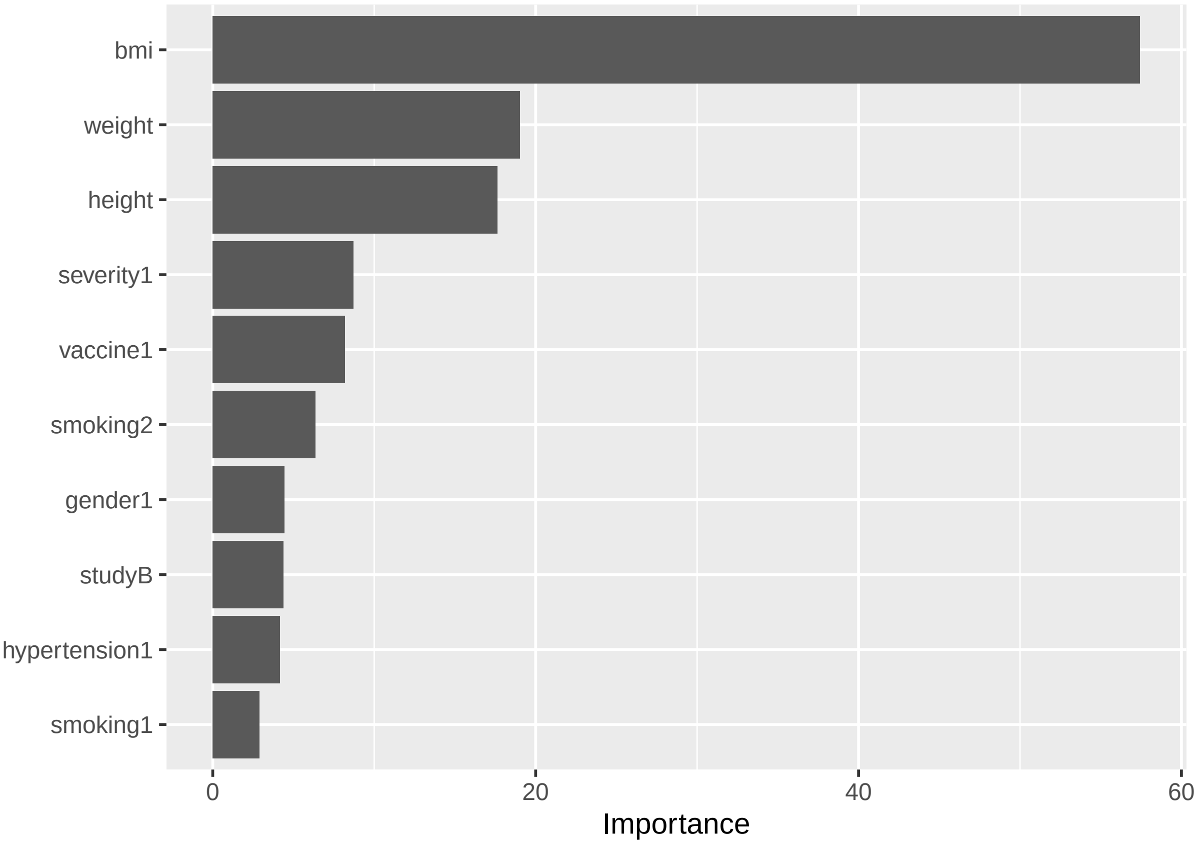
```
coef(enet.fit$finalModel, enet.fit$bestTune$lambda)
```

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                          s1
## (Intercept)   -3.014181e+03
## age            1.152745e-01
## gender1       -4.446710e+00
## race2          2.215140e+00
## race3         -6.976962e-01
## race4         -1.153433e+00
## smoking1       2.905529e+00
## smoking2       6.363631e+00
## height         1.762383e+01
## weight        -1.904323e+01
## bmi            5.742555e+01
## hypertension1  4.166158e+00
## diabetes1     -1.190108e+00
## SBP           -7.830526e-02
## LDL           -4.231365e-02
## vaccine1      -8.188023e+00
## severity1      8.709786e+00
## studyB         4.376008e+00
## studyC        -6.676946e-01
```

```
ggplot(enet.fit, highlight = TRUE) +
  scale_x_continuous(trans='log', n.breaks = 6) +
  labs(title ="Elastic Net CV Result") +
  theme_bw()
```

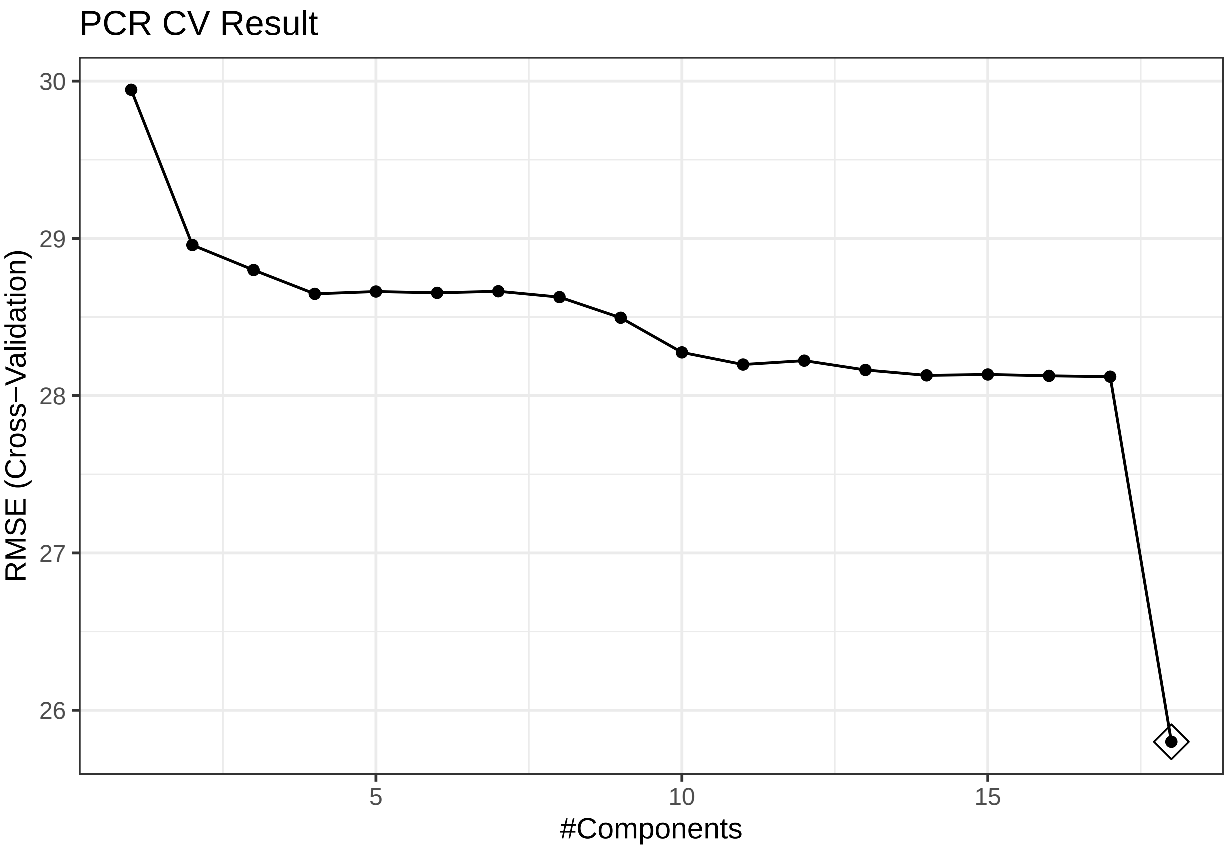


```
ggsave("./figure/enet_cv.jpeg", dpi = 500)

vip(enet.fit$finalModel)
```



3.5 Principal components regression (PCR)

```
set.seed(3196)
pcr.fit <- train(train_x,
                 train_y,
                 method = "pcr",
                 tuneGrid  = data.frame(ncomp = 1:ncol(train_x)),
                 trControl = ctrl1,
                 preProcess = c("center", "scale"))
ggplot(pcr.fit, highlight = TRUE) +
  labs(title  ="PCR CV Result") +
  theme_bw()
```



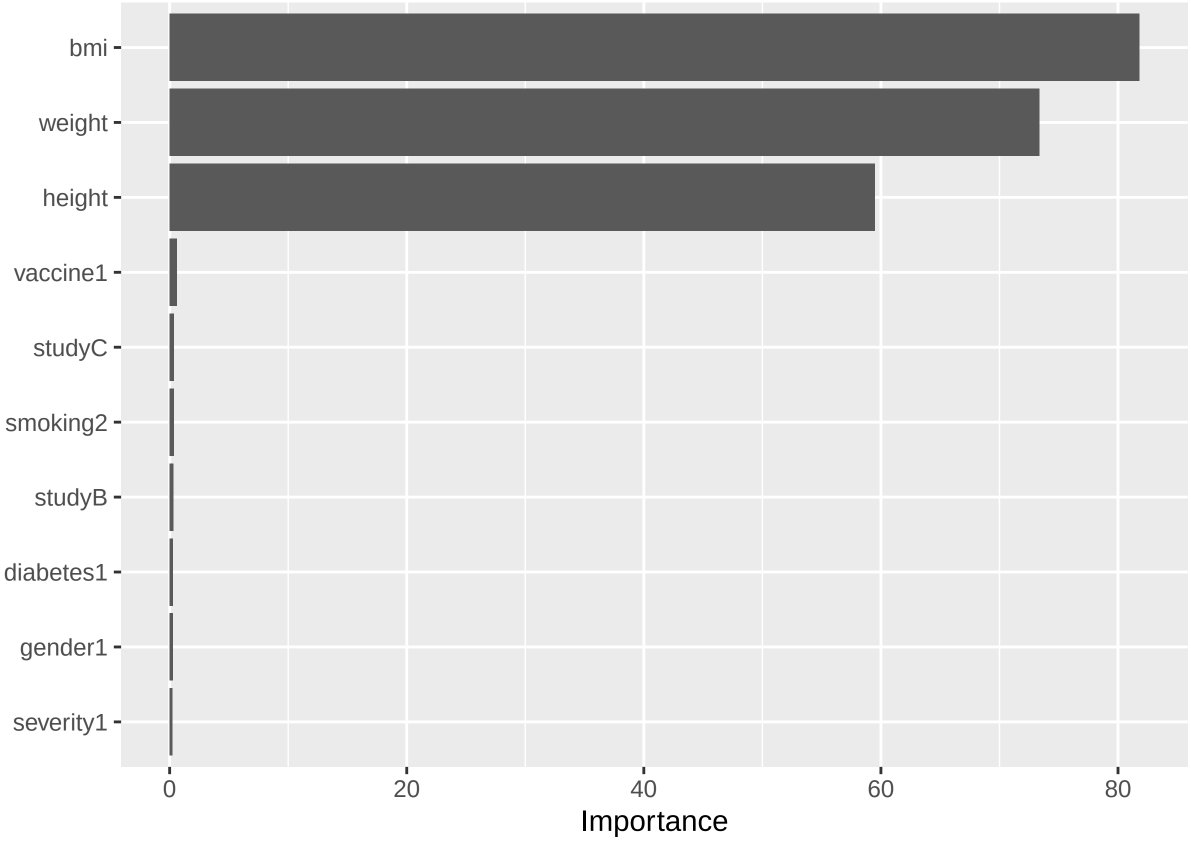
```
ggsave("./figure/pcr_cv.jpeg", dpi = 500)

pcr.fit$bestTune
```

```
##    ncomp
## 18    18
```

```
coef(pcr.fit$finalModel)
```

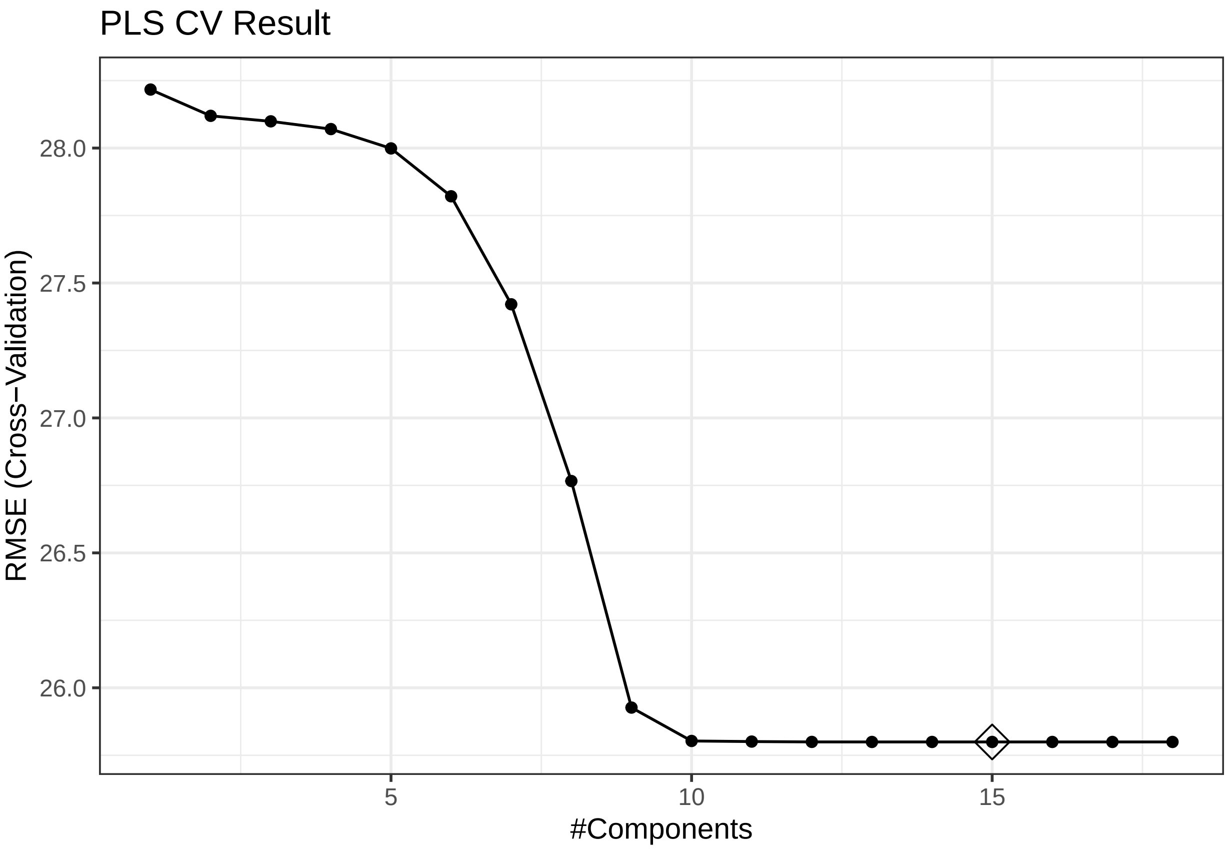
```
## , , 18 comps
##
##                    .outcome
## age               0.5252538
## gender1          -2.2221586
## race2             0.4563464
## race3            -0.2619635
## race4            -0.3476329
## smoking1          1.3205684
## smoking2          1.9344423
## height          112.6936931
## weight         -141.0001175
## bmi             165.1518985
## hypertension1     2.0811234
## diabetes1        -0.4188178
## SBP              -0.6356938
## LDL              -0.8376686
## vaccine1         -4.0025673
## severity1         2.5879846
## studyB            2.1374000
## studyC           -0.2730416
```

```
vip(pcr.fit$finalModel)
```



3.6 Partial Least Squares (PLS)

```
set.seed(3196)
pls.fit <- train(train_x,
                 train_y,
                 method = "pls",
                 tuneGrid = data.frame(ncomp = 1:ncol(train_x)),
                 trControl = ctrl1,
                 preProcess = c("center", "scale"))
ggplot(pls.fit, highlight = TRUE) +
  labs(title  ="PLS CV Result") +
  theme_bw()
```



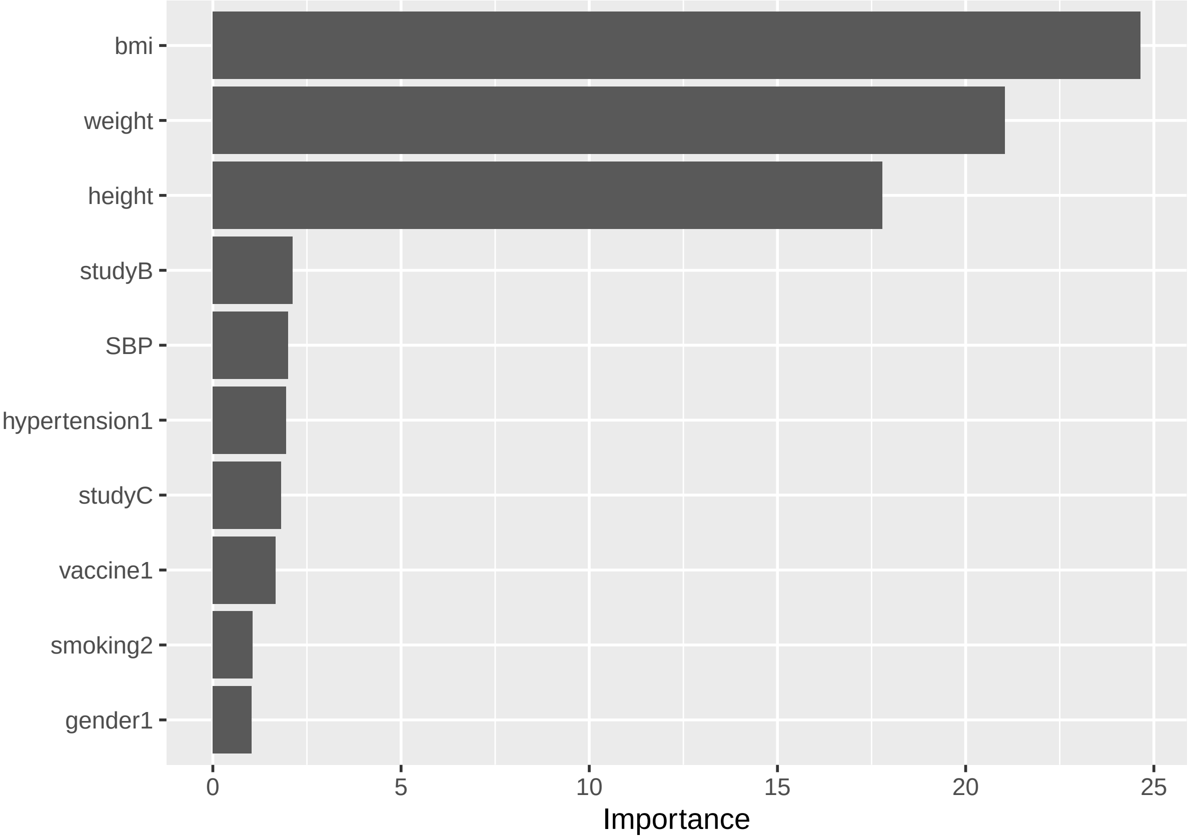
```
ggsave("./figure/pls_cv.jpeg", dpi = 500)

pls.fit$bestTune
```

```
##    ncomp
## 15    15
```

```
coef(pls.fit$finalModel)
```

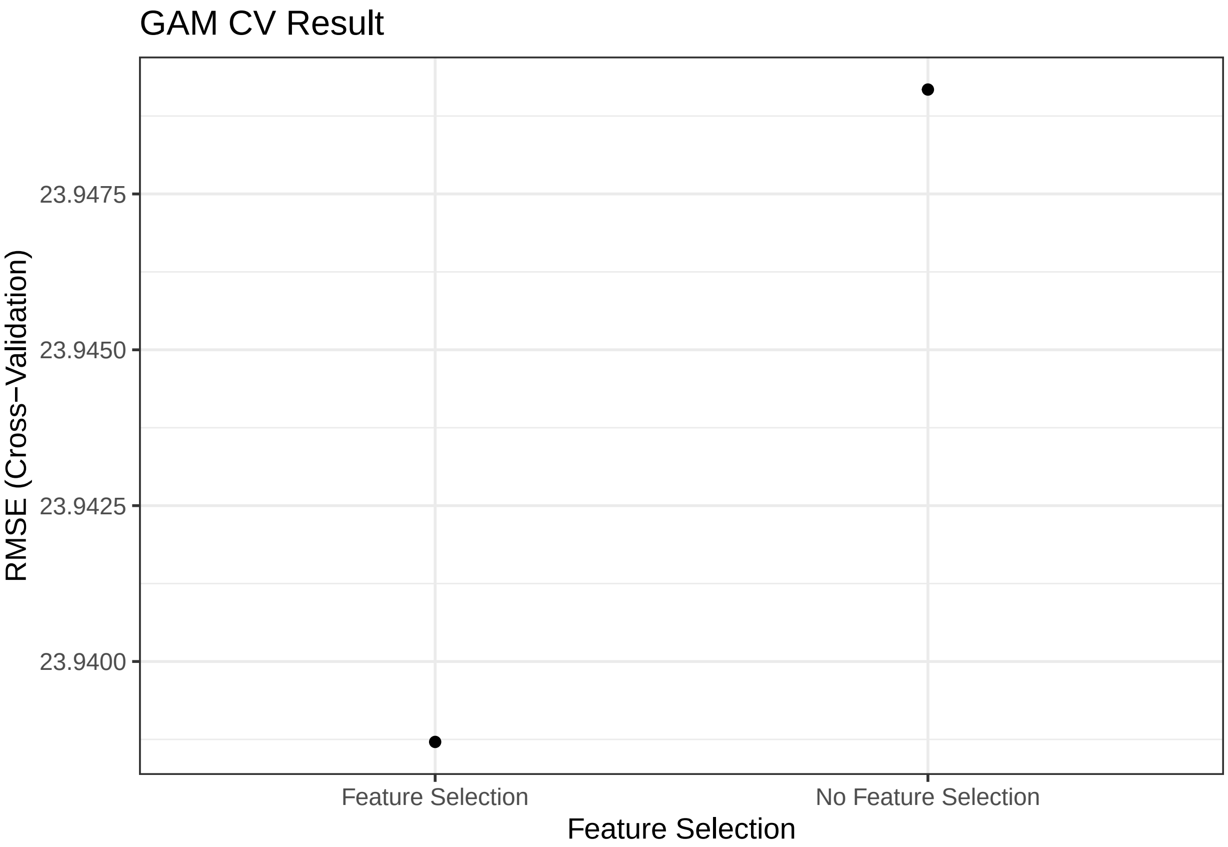
```
## , , 15 comps
##
##                    .outcome
## age               0.5252540
## gender1          -2.2221591
## race2             0.4563454
## race3            -0.2619627
## race4            -0.3476322
## smoking1          1.3205686
## smoking2          1.9344419
## height          112.6936929
## weight         -141.0001175
## bmi             165.1518987
## hypertension1     2.0811232
## diabetes1        -0.4188173
## SBP              -0.6356940
## LDL              -0.8376679
## vaccine1         -4.0025660
## severity1         2.5879873
## studyB           2.1373998
## studyC          -0.2730415
```

```
vip(pls.fit$finalModel)
```



3.7 Generalized additive model (GAM)

```
set.seed(3196)
gam.fit <- train(train_x,
                 train_y,
                 method = "gam",
                 tuneGrid = data.frame(select = c(TRUE, FALSE),
                                       method = "GCV.Cp"),
                 trControl = ctrl1)


ggplot(gam.fit) +
  labs(title = "GAM CV Result") +
  theme_bw()
```



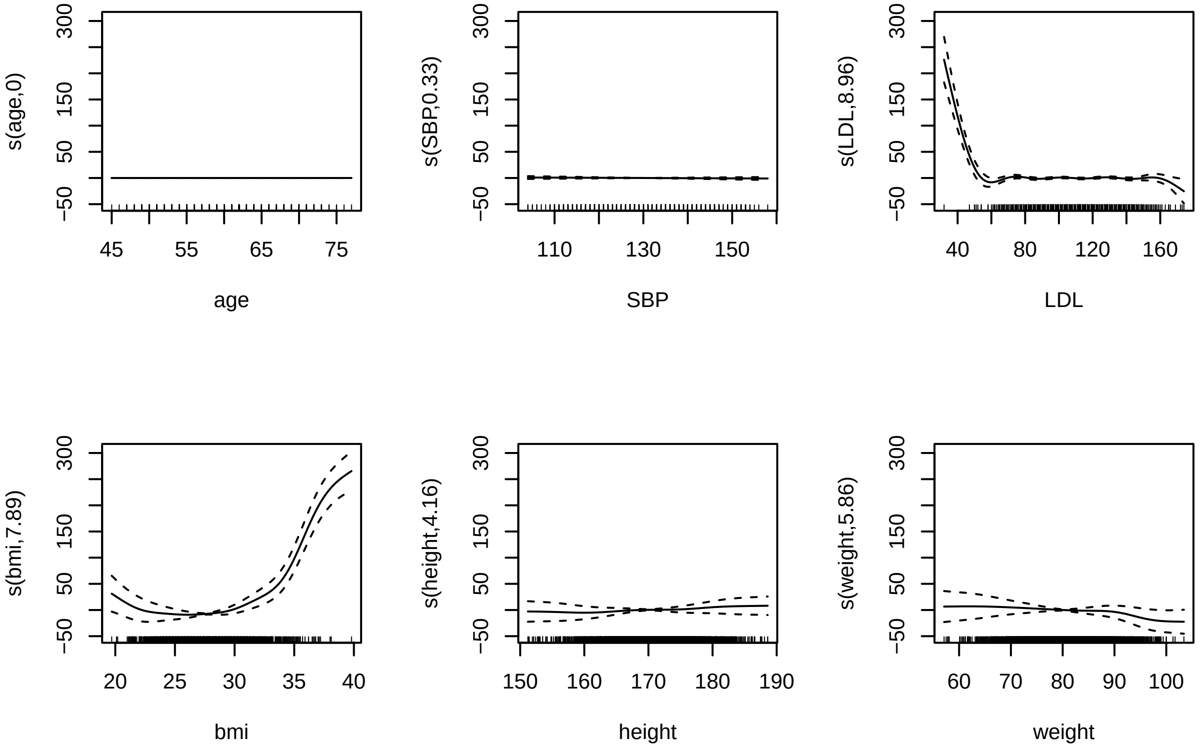
```
ggsave("./figure/gam_cv.jpeg", dpi = 500)

gam.fit$bestTune
```

```
##   select method
## 2   TRUE GCV.Cp
```

```
# coef(gam.fit$finalModel)
gam.fit$finalModel
```

```
##
## 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)
##
## Estimated degrees of freedom:
## 0.000 0.329 8.959 7.893 4.163 5.856  total = 39.2
##
## GCV score: 524.051
```

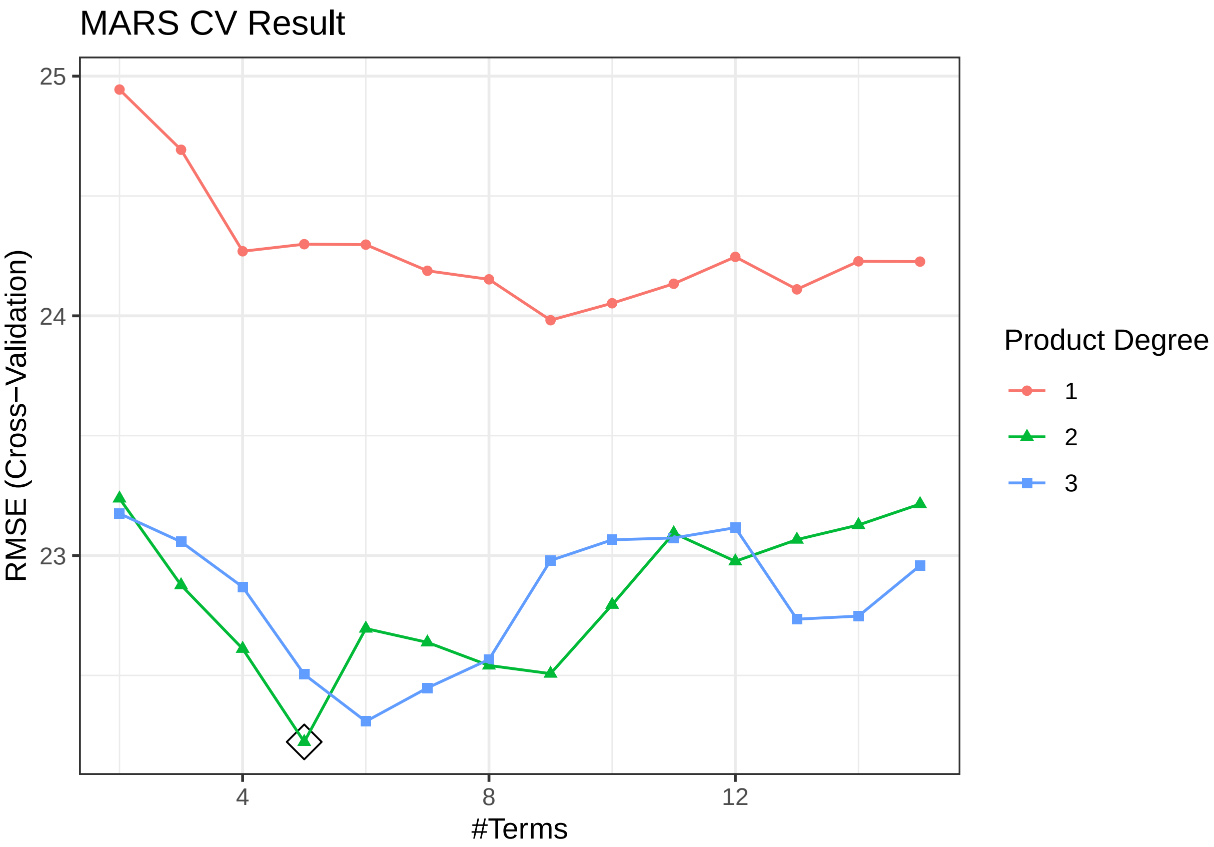
```
par(mfrow=c(2, 3))
plot(gam.fit$finalModel)
```



```
par(mfrow=c(1, 1))
```

3.8 Multivariate Adaptive Regression Splines (MARS)

```
mars_grid <- expand.grid(degree = 1:3,
                         nprune = 2:15)
set.seed(3196)
mars.fit <- train(train_x,
                  train_y,
                  method = "earth",
                  tuneGrid = mars_grid,
                  trControl = ctrl1)

ggplot(mars.fit, highlight = TRUE)+
  labs(title  ="MARS CV Result") +
  theme_bw()
```



```
ggsave("./figure/mars_cv.jpeg", dpi = 500)

mars.fit$bestTune
```

```
##    nprune degree
## 18      5      2
```

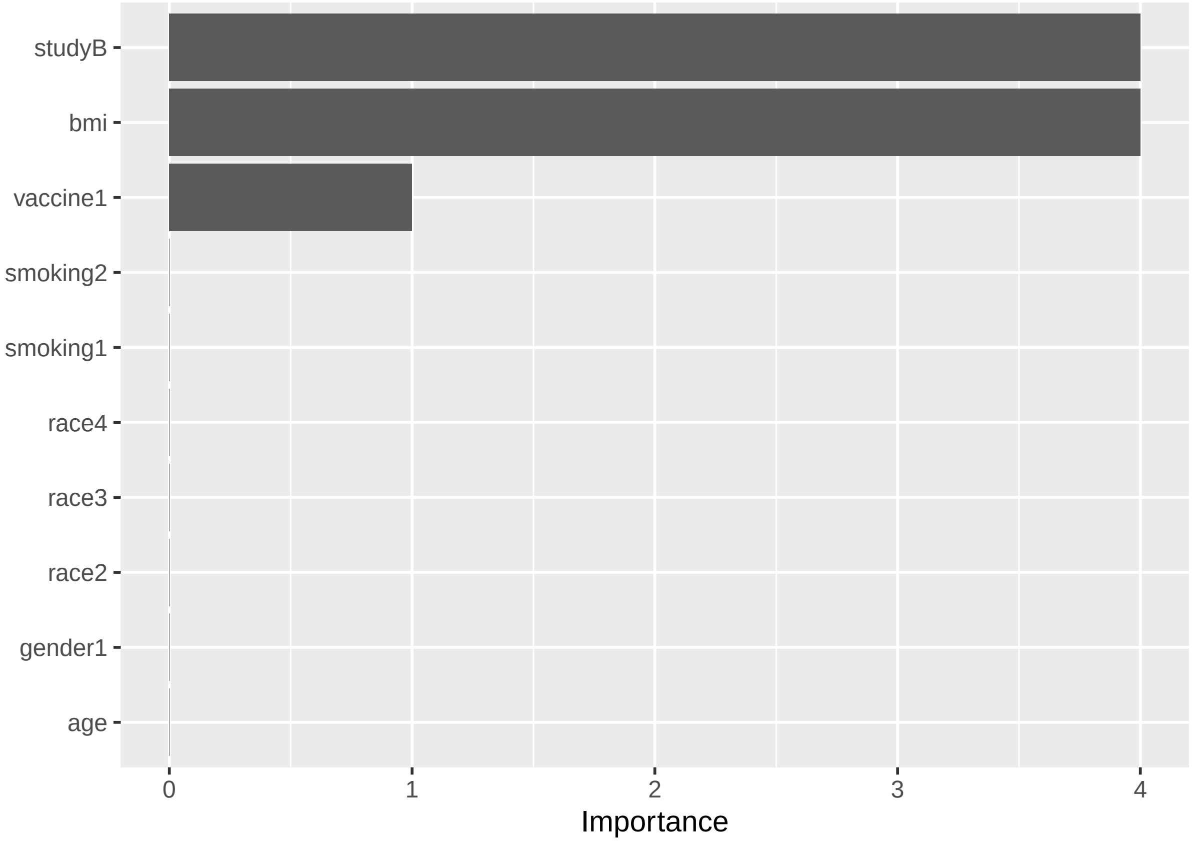
```
coef(mars.fit$finalModel)
```

```
##         (Intercept)          h(31.7-bmi) h(bmi-31.7) * studyB
##           19.366730             3.705371            34.383832
##         h(bmi-26.8)             vaccine1
##            6.695655            -7.788338
```

```
summary(mars.fit$finalModel)
```

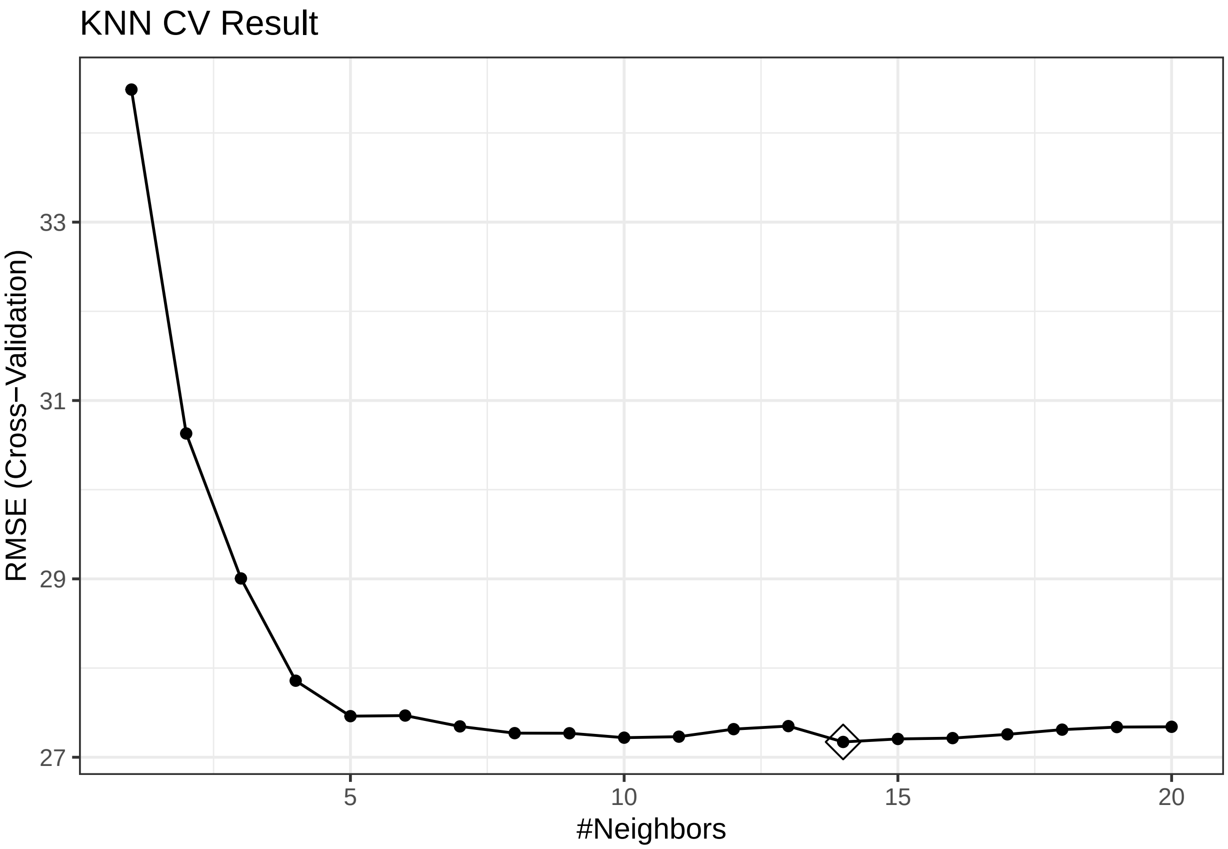
```
## Call: earth(x=matrix[2900,18], y=c(40,34,31,50,3...), keepxy=TRUE, degree=2,
##             nprune=5)
##
##                        coefficients
## (Intercept)               19.366730
## vaccine1                  -7.788338
## h(bmi-26.8)                6.695655
## h(31.7-bmi)                3.705371
## h(bmi-31.7) * studyB      34.383832
##
## Selected 5 of 25 terms, and 3 of 18 predictors (nprune=5)
## Termination condition: Reached nk 37
## Importance: bmi, studyB, vaccine1, age-unused, gender1-unused, ...
## Number of terms at each degree of interaction: 1 3 1
## GCV 491.1694    RSS 1413606    GRSq 0.4723714    RSq 0.4760052
```

```
vip(mars.fit$finalModel)
```



3.9 K-Nearest Neighbour (KNN)

```
set.seed(3196)
knn.fit <- train(train_x,
                 train_y,
                 tuneGrid  = data.frame(k = 1:20),
                 method = "knn",
                 trControl = ctrl1)

ggplot(knn.fit, highlight = TRUE) +
  labs(title  ="KNN CV Result") +
  theme_bw()
```



```
ggsave("./figure/knn_cv.jpeg", dpi = 500)

knn.fit$bestTune
```

```
##     k
## 14 14
```

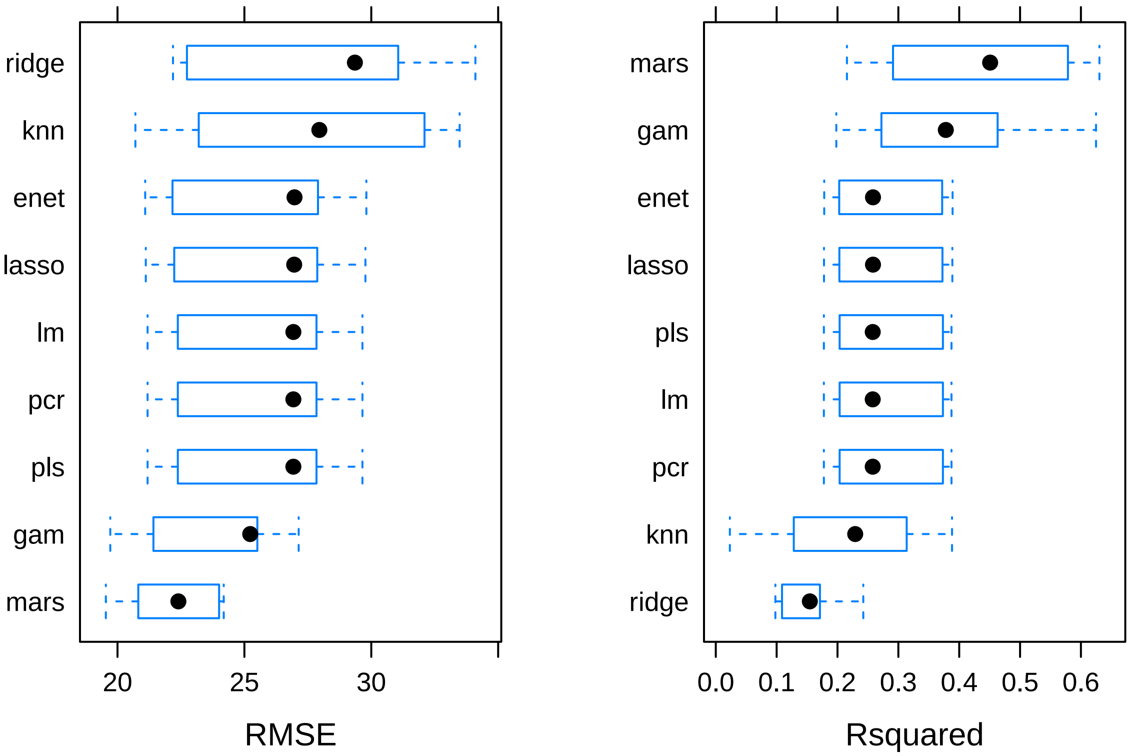
4 Model Selection

```
resamp <- resamples(list(lm = lm.fit,
                         lasso = lasso.fit,
                         ridge = ridge.fit,
                         enet = enet.fit,
                         pcr = pcr.fit,
                         pls = pls.fit,
                         gam = gam.fit,
                         mars = mars.fit,
                         knn = knn.fit))

summary(resamp)
```

```
##
## Call:
## summary.resamples(object = resamp)
##
## Models: lm, lasso, ridge, enet, pcr, pls, gam, mars, knn
## Number of resamples: 10
```

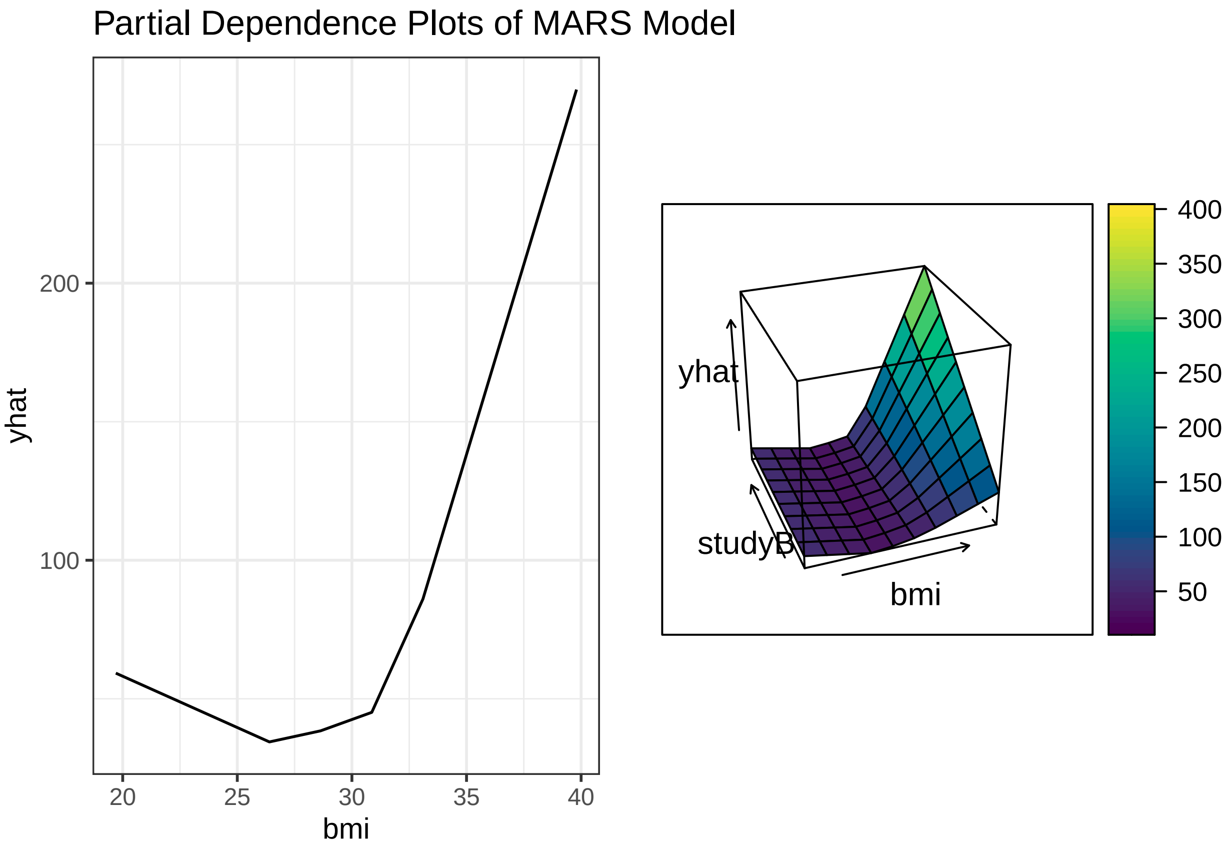
```
##
## MAE
##           Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## lm    15.51396 15.83948 16.88102 16.61921 17.14341 17.96742    0
## lasso 15.44129 15.76841 16.80674 16.55187 17.07095 17.92603    0
## ridge 15.68400 16.02330 16.92856 16.82838 17.25775 18.56533    0
## enet  15.41158 15.73749 16.78372 16.52957 17.04973 17.91601    0
## pcr   15.51396 15.83948 16.88102 16.61921 17.14341 17.96742    0
## pls   15.51396 15.83948 16.88101 16.61921 17.14341 17.96741    0
## gam   14.60978 15.08810 15.43250 15.46444 15.67318 16.85977    0
## mars  13.64202 14.46532 14.77166 14.86003 15.47625 15.91164    0
## knn   14.39020 15.47735 16.02523 15.98084 16.50950 17.33583    0
##
## RMSE
##           Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## lm    21.18486 22.72303 26.93048 25.79936 27.83037 29.65227    0
## lasso 21.11493 22.59901 26.96284 25.79552 27.85980 29.76940    0
## ridge 22.18691 23.07543 29.35410 27.80340 30.82227 34.10345    0
## enet  21.09063 22.54005 26.97513 25.79174 27.88113 29.80609    0
## pcr   21.18486 22.72303 26.93048 25.79936 27.83037 29.65227    0
## pls   21.18486 22.72303 26.93048 25.79936 27.83037 29.65227    0
## gam   19.71529 21.64375 25.22940 23.93871 25.49934 27.13899    0
## mars  19.54048 20.85027 22.39887 22.22258 23.91929 24.18597    0
## knn   20.70529 23.31530 27.95344 27.17151 31.25128 33.48263    0
##
## Rsquared
##             Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lm    0.17762897 0.2039281 0.2578972 0.2769133 0.3671569 0.3873461    0
## lasso 0.17784426 0.2034880 0.2583431 0.2766401 0.3661592 0.3887409    0
## ridge 0.09795596 0.1134920 0.1545560 0.1530016 0.1700872 0.2424090    0
## enet  0.17818275 0.2033071 0.2583457 0.2765137 0.3655471 0.3891442    0
## pcr   0.17762897 0.2039281 0.2578972 0.2769133 0.3671569 0.3873461    0
## pls   0.17762894 0.2039281 0.2578972 0.2769133 0.3671570 0.3873461    0
## gam   0.19802433 0.2803342 0.3780187 0.3882578 0.4533286 0.6248481    0
## mars  0.21553836 0.3047383 0.4508321 0.4330195 0.5589093 0.6304219    0
## knn   0.02314974 0.1357417 0.2291133 0.2137684 0.3009374 0.3882280    0
```

```
# jpeg("./figure/resample.jpeg", width = 8, height=6, units="in", res=500)
p1=bwplot(resamp, metric = "RMSE")
p2=bwplot(resamp, metric = "Rsquared")
grid.arrange(p1, p2 ,ncol=2)
```



```
# dev.off()
```

```
p1<- pdp::partial(mars.fit, pred.var = c("bmi"), grid.resolution = 10) %>% autoplot() +
  theme_bw()+
  labs(title = "Partial Dependence Plots of MARS Model")

p2 <-pdp::partial(mars.fit, pred.var = c("bmi", "studyB"),
                   grid.resolution = 10) %>%
      pdp::plotPartial(levelplot = FALSE, zlab = "yhat", drape = TRUE,
                       screen = list(z = 20, x = -60))

# jpeg("./figure/partial_dependence.jpeg", width = 8, height=6, units="in", res=500)
gridExtra::grid.arrange(p1, p2, ncol = 2)
```



```
# dev.off()

# Important variables
varImp(mars.fit$finalModel)
```

```
##            Overall
## bmi      100.00000
## studyB   100.00000
## vaccine1  17.78457
```

5 Training / Testing Error

```
# training error
mars.train.pred = predict(mars.fit, newdata = train_x)
RMSE(train_y, mars.train.pred)
```

```
## [1] 22.07828
```

```
# testing error
mars.pred = predict(mars.fit, newdata = test_x)
RMSE(test_y, mars.pred)
```

```
## [1] 22.1712
```