P8106 Midterm Project

Leyang Rui, Jinghan Zhao

2025-03-28

Load Data

```r
load("data/dat1.RData")
train_data = dat1 |>
  janitor::clean_names() |>
  mutate(
    gender = as.factor(gender),
    diabetes = as.factor(diabetes),
    hypertension = as.factor(hypertension),
    race = fct_recode(race,
                  White = "1",
                  Asian = "2",
                  Black = "3",
                  Hispanic = "4"),
    gender = fct_recode(gender,
                        Male = "1",
                        Female = "0"),
    smoking = fct_recode(smoking,
                         "Never smoked" = "0",
                         "Former smoker" = "1",
                         "Current smoker" = "2"))

load("data/dat2.RData")
test_data = dat2 |>
  janitor::clean_names() |>
  mutate(
    gender = as.factor(gender),
    diabetes = as.factor(diabetes),
    hypertension = as.factor(hypertension),
    race = fct_recode(race,
                  White = "1",
                  Asian = "2",
                  Black = "3",
                  Hispanic = "4"),
    gender = fct_recode(gender,
                        Male = "1",
                        Female = "0"),
    smoking = fct_recode(smoking,
                         "Never smoked" = "0",
                         "Former smoker" = "1",
                         "Current smoker" = "2")
  )
```

Modify Data

```
train_data1 =
  train_data %>%
  select(-id, -height, -weight, -hypertension)

x_train = model.matrix(log_antibody ~ ., train_data1)[, -1]
colnames(x_train) = make.names(colnames(x_train), unique = TRUE)
y_train = train_data1[, "log_antibody"]

test_data =
  test_data %>%
  select(-id, -height, -weight, -hypertension)

x_test = model.matrix(log_antibody ~ ., test_data)[, -1]
colnames(x_test) = make.names(colnames(x_test), unique = TRUE)
y_test = test_data[, "log_antibody"]

ctrl1 = trainControl(method = "cv", number = 10)
```

Descriptive Analysis

Numeric Variables

```
train_data |>
  pivot_longer(
    cols = c(age, height, weight, bmi),
    names_to = "variable",
    values_to = "value"
  ) |>
  ggplot(aes(x = value, y = log_antibody, color = variable)) +
  geom_point(alpha = 0.5, size = 0.6) +
  facet_wrap(variable ~ .,  scales = "free") +
    labs(title = "Distribution of the Demographic Continuous Variables",
         x = "Variables",
         y = "Log_Antibody")
```

```
train_data |>
  pivot_longer(
    cols = c(sbp, ldl, time),
    names_to = "variable",
    values_to = "value"
  ) |>
  ggplot(aes(x = value, y = log_antibody, color = variable)) +
  geom_point(alpha = 0.5, size = 0.6) +
```

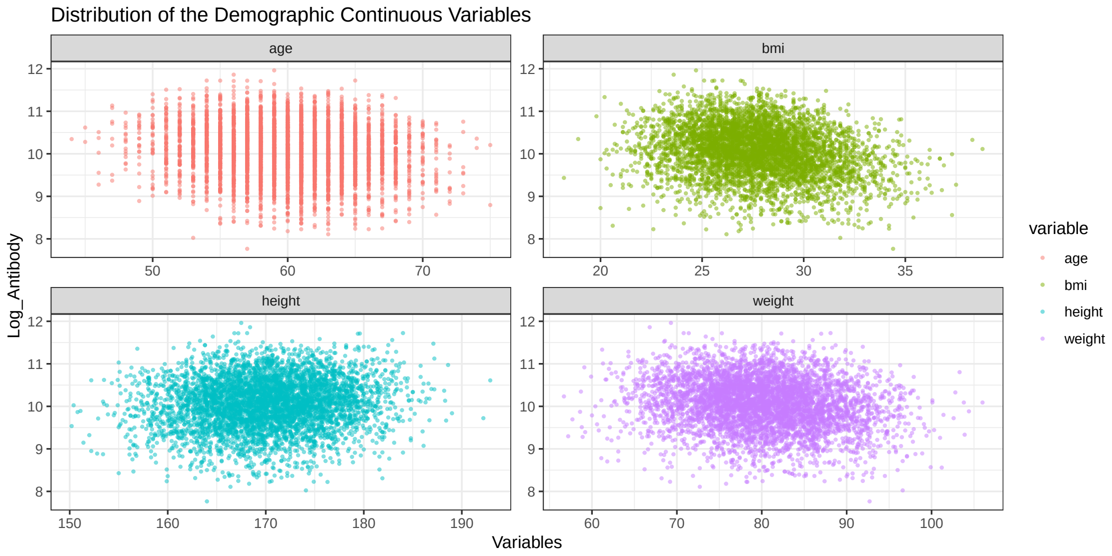

Figure 1: Distribution of the Demographic Continuous Variables

```r
  facet_wrap(variable ~ .,  scales = "free") +
    labs(title = "Distribution of the Clinical Continuous Variables",
         x = "Variables",
         y = "Log_Antibody")
```

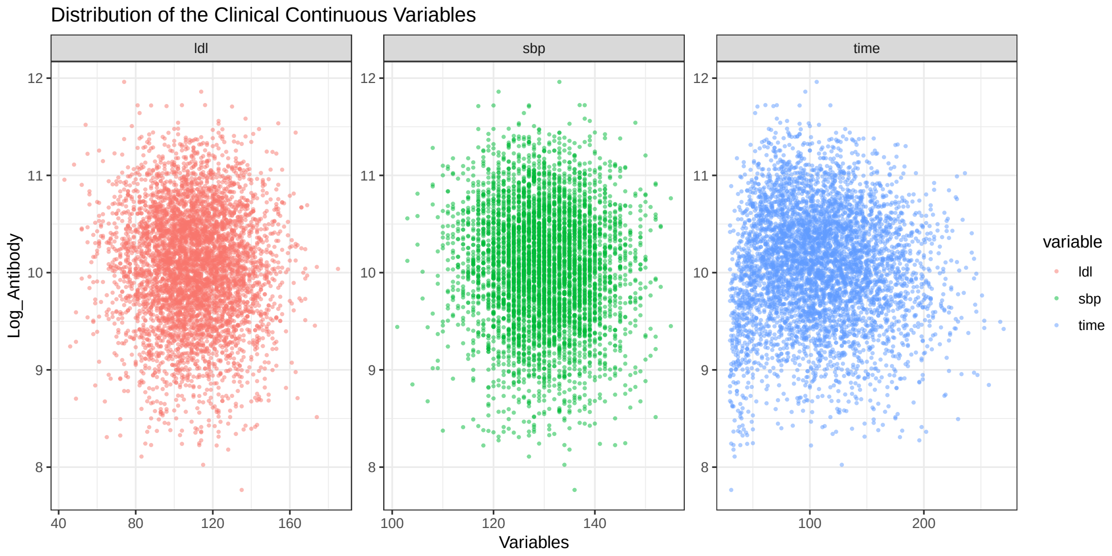

Figure 2: Distribution of the Clinical Continuous Variables

```r
train_data %>%
  pivot_longer(
    cols = c(age, height, weight, bmi, sbp, ldl, time, log_antibody),
    names_to = "variable_name",
    values_to = "value"
  ) %>%
  group_by(variable_name) %>%
  summarize(
    mean = mean(value),
    median = median(value),
    min = min(value),
    first_quantile = quantile(value, probs = 0.25),
    third_quantile = quantile(value, probs = 0.75),
    max = max(value)
  ) %>%
  ungroup() %>%
  arrange(desc(variable_name == "log_antibody"), variable_name) %>%
  knitr::kable(digits = 3, caption = "Descriptive Statistics for the Continuous Variables")
```

Table 1: Descriptive Statistics for the Continuous Variables

variable_name	mean	median	min	first_quantile	third_quantile	max
log_antibody	10.064	10.089	7.765	9.682	10.478	11.961
age	59.968	60.000	44.000	57.000	63.000	75.000
bmi	27.740	27.600	18.200	25.800	29.500	38.800
height	170.126	170.100	150.200	166.100	174.225	192.900
ldl	109.909	110.000	43.000	96.000	124.000	185.000
sbp	129.900	130.000	101.000	124.000	135.000	155.000
time	108.863	106.000	30.000	76.000	138.000	270.000
weight	80.109	80.100	56.700	75.400	84.900	106.000

Categorical Variables

```
train_data |>
  pivot_longer(
    cols = c(gender, race, smoking, diabetes, hypertension),
    names_to = "variable",
    values_to = "value"
  ) |>
  mutate(
    variable = factor(variable, levels = c("gender", "race", "smoking", "diabetes", "hypertension"))
  ) |>
  ggplot(aes(x = value, y = log_antibody, fill = variable)) +
  geom_boxplot(alpha = 0.5) +
  facet_wrap(variable ~ .,  scales = "free") +
    labs(title = "Distribution of the Categorical Variables",
         x = "Variables",
         y = "Log_Antibody") +
  theme(axis.text.x = element_text(angle = 30, vjust = 1, hjust = 1))
```

Correlation Plot

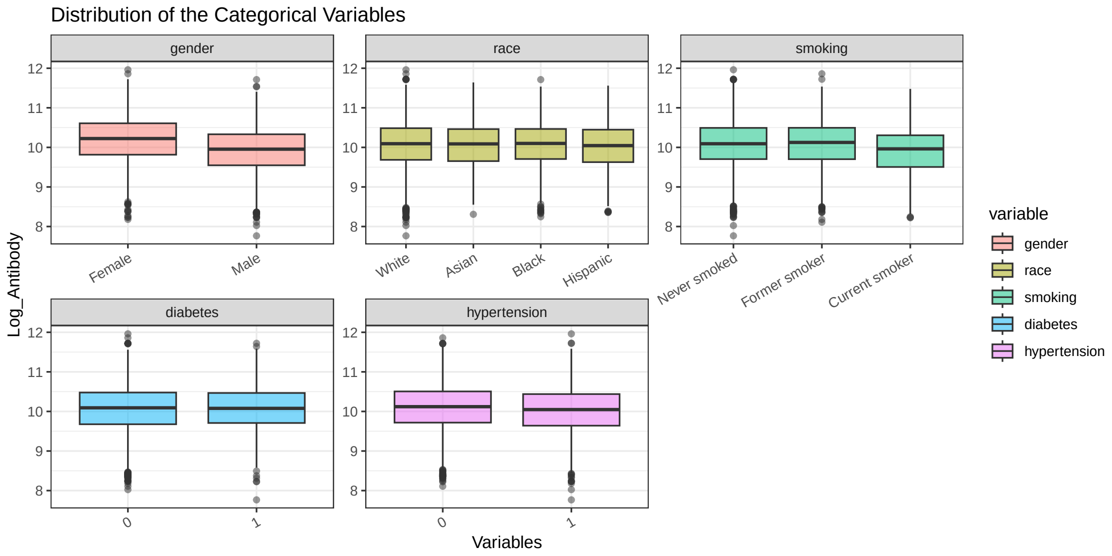

Figure 3: Distribution of the Categorical Variables

```r
x_corr = model.matrix(log_antibody ~ ., train_data[, -1])[, -1]
corrplot(cor(x_corr), method = "circle", type = "full")
```

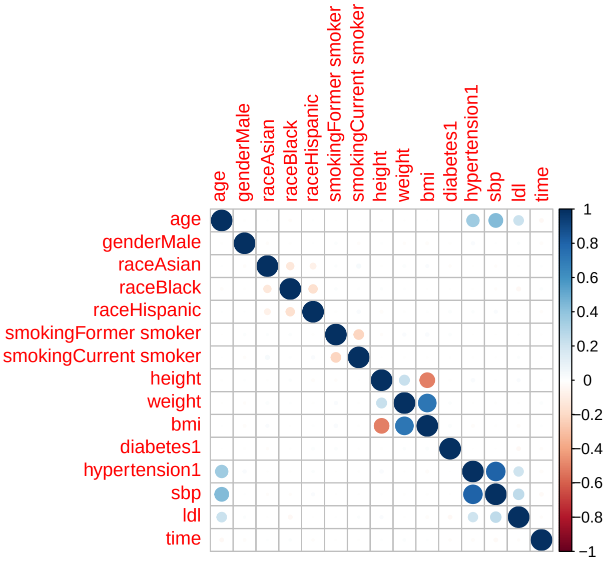

Figure 4: Correlation Plot

Regression

Elastic Net

```r
set.seed(37)
enet_fit = train(log_antibody ~ .,
                 data = train_data1,
                 method = "glmnet",
                 tuneGrid = expand.grid(alpha = seq(0, 1, length = 21),
                                        lambda = exp(seq(-2, -8, length = 100))),
                 trControl = ctrl1)
enet_fit$bestTune
```

```
##      alpha      lambda
## 2044     1 0.004544037
```

```r
coef(enet_fit$finalModel, enet_fit$bestTune$lambda)
```

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                                  s1
## (Intercept)            1.272351e+01
## age                   -1.914947e-02
## genderMale            -2.859704e-01
## raceAsian              .           
## raceBlack              .           
## raceHispanic          -2.469963e-02
## smokingFormer smoker   1.592525e-02
## smokingCurrent smoker -1.770936e-01
## bmi                   -4.820177e-02
## diabetes1              4.263422e-05
## sbp                    .           
## ldl                    .           
## time                  -1.850422e-04
```

```r
mycol = rainbow(25)
mypar = list(superpose.symbol = list(col = mycol),
superpose.line = list(col = mycol))

plot(enet_fit, par.settings = mypar, xTrans = log)
```

PCR

```r
set.seed(37)
pcr_fit = train(x_train, y_train,
                method = "pcr",
                tuneGrid = data.frame(ncomp = 1:12),
                trControl = ctrl1,
                preProcess = c("center", "scale"))
summary(pcr_fit)
```

```
## Data:    X dimension: 5000 12
##  Y dimension: 5000 1
```

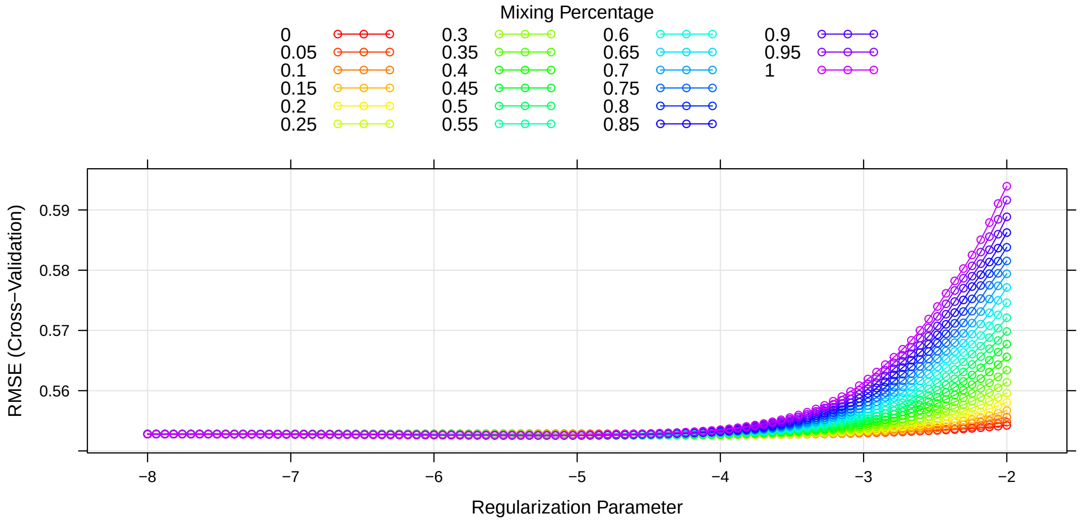

Figure 5: Effect of Tuning Parameters on Train Error (Elastic Net)

```
## Fit method: svdpc
## Number of components considered: 12
## TRAINING: % variance explained
##           1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X          13.522   23.760   33.489   42.534    51.17   59.563    67.70
## .outcome    1.296    1.493    1.512    1.512     1.54    2.032    13.44
##           8 comps  9 comps  10 comps  11 comps  12 comps
## X           75.76    82.63     89.32     95.36     100.0
## .outcome    13.48    13.68     13.78     13.84      14.5
```

```
ggplot(pcr_fit, highlight = TRUE)
```

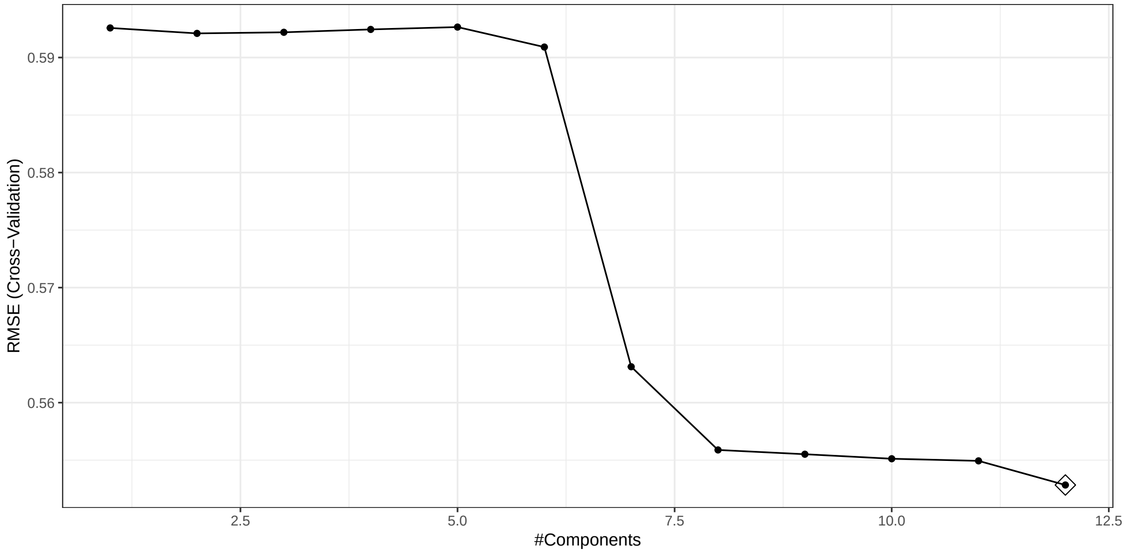

Figure 6: Component Selection (PCR)

PLS

```r
set.seed(37)
pls_fit = train(x_train, y_train,
                method = "pls",
                tuneGrid = data.frame(ncomp = 1:12),
                trControl = ctrl1,
                preProcess = c("center", "scale"))
summary(pls_fit)
```

```
## Data:    X dimension: 5000 12
##  Y dimension: 5000 1
## Fit method: oscorespls
## Number of components considered: 3
## TRAINING: % variance explained
##           1 comps  2 comps  3 comps
## X           9.295    21.21    26.88
## .outcome   13.885    14.45    14.50
```

```r
ggplot(pls_fit, highlight = TRUE)
```

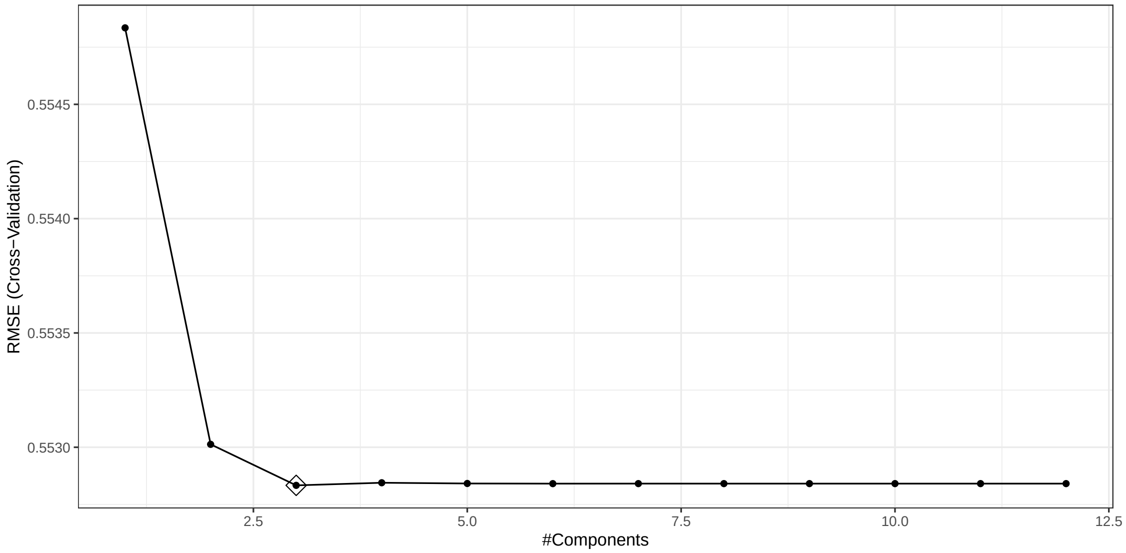

Figure 7: Component Selection (PLS)

GAM

```r
set.seed(37)

gam.fit = train(x_train, y_train,
                 method = "gam",
                 trControl = ctrl1)
```

```
gam.fit$bestTune
```

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

```
gam.fit$finalModel
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## .outcome ~ genderMale + raceAsian + raceBlack + raceHispanic +
##     smokingFormer.smoker + smokingCurrent.smoker + diabetes1 +
##     s(age) + s(sbp) + s(ldl) + s(bmi) + s(time)
##
## Estimated degrees of freedom:
## 0.992 0.000 0.000 4.179 7.915  total = 21.09
##
## GCV score: 0.2786375
```

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

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

MARS

```
set.seed(37)

mars_grid = expand.grid(degree = 1:3,
                        nprune = 2:12)

mars.fit = train(x_train, y_train,
                 method = "earth",
                 tuneGrid = mars_grid,
                 trControl = ctrl1)

ggplot(mars.fit)
```

```
mars.fit$bestTune
```

```
##   nprune degree
## 8      9      1
```

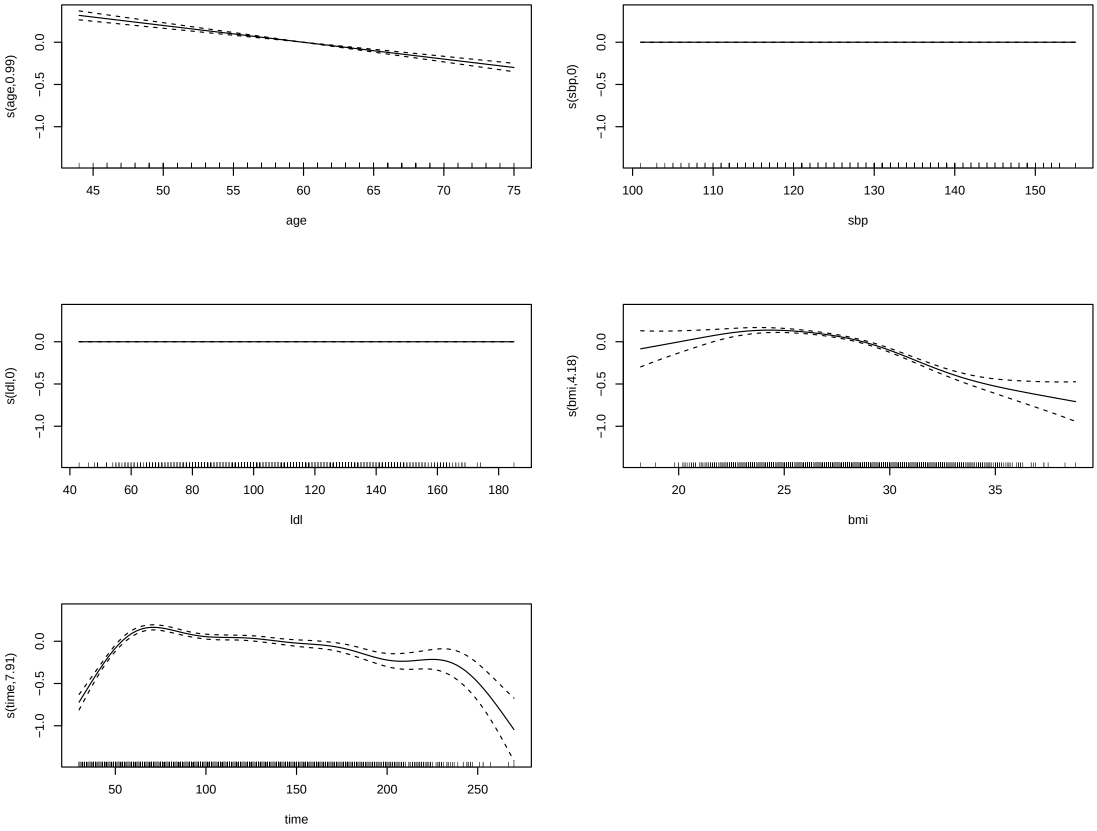

Figure 8: Degree of Predictors (GAM)

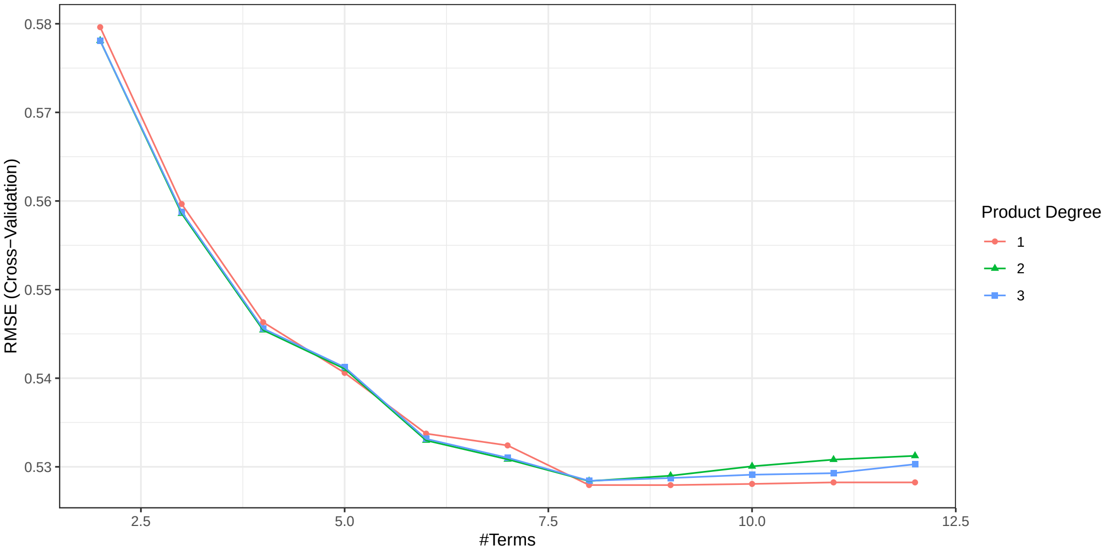

Figure 9: Term and Degree Selection (MARS)

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

```
##           (Intercept)           h(27.8-bmi)           h(time-57)
##          10.847446930          -0.061997354         -0.002254182
##            h(57-time)            genderMale           h(age-59)
##          -0.033529326          -0.296290451         -0.022957648
##             h(59-age) smokingCurrent.smoker          h(bmi-23.7)
##           0.016138468          -0.205126851         -0.084380175
```

Regression Trees

```
set.seed(37)

tree_full = rpart(formula = log_antibody ~ ., data = train_data1,
                  control = rpart.control(cp = 0))

cpTable = tree_full$cptable
plotcp(tree_full)
```

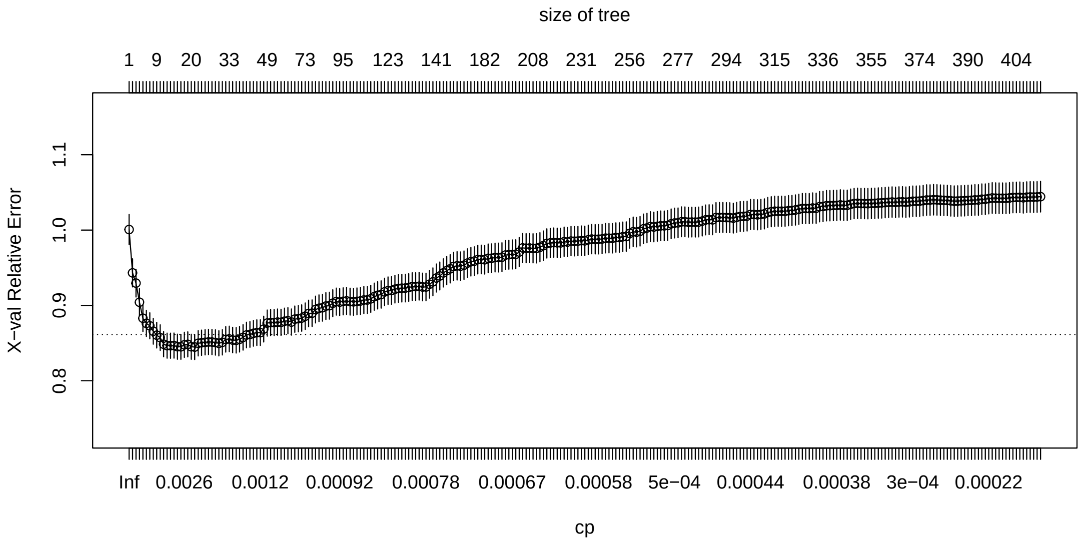

Figure 10: Tuning Parameter Selection (Regression Tree)

```
## Find the cp that yields the minimum cross-validation error
minErr = which.min(cpTable[,4])
tree_final = rpart::prune(tree_full, cp = cpTable[minErr, 1])
rpart.plot(tree_final)
```

```
summary(tree_final)
```

```
## Call:
## rpart(formula = log_antibody ~ ., data = train_data1, control = rpart.control(cp = 0))
```

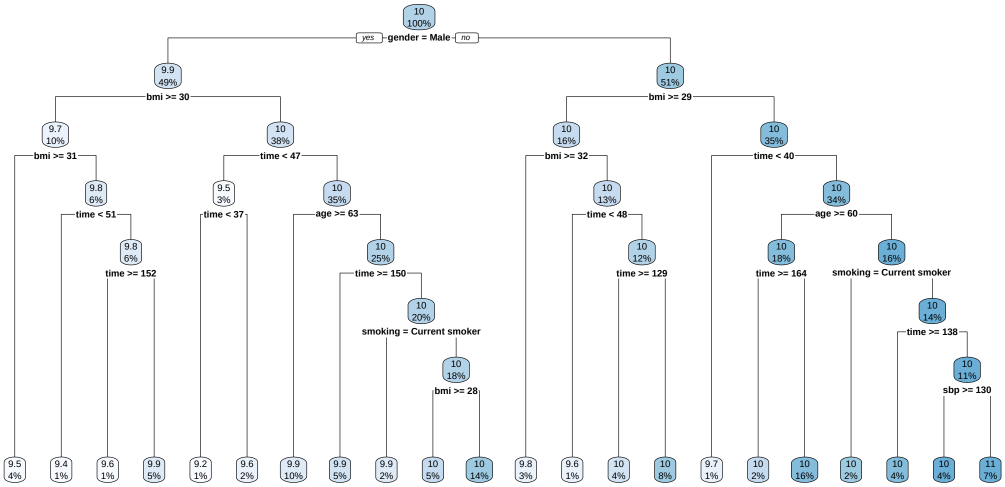

Figure 11: Final Regression Tree Model

```
##   n= 5000
##
##             CP nsplit rel error    xerror       xstd
## 1  0.057918182      0 1.0000000 1.0007146 0.02004169
## 2  0.027293800      1 0.9420818 0.9432020 0.01870086
## 3  0.025172140      2 0.9147880 0.9295966 0.01846651
## 4  0.020759383      3 0.8896159 0.9042136 0.01812179
## 5  0.010560350      4 0.8688565 0.8829836 0.01767741
## 6  0.007928454      5 0.8582961 0.8761783 0.01736189
## 7  0.007483148      6 0.8503677 0.8729469 0.01732773
## 8  0.006830408      7 0.8428845 0.8657230 0.01716499
## 9  0.006412912      8 0.8360541 0.8604904 0.01704889
## 10 0.006407846      9 0.8296412 0.8572604 0.01697315
## 11 0.004852623     10 0.8232334 0.8485364 0.01678947
## 12 0.003284421     11 0.8183808 0.8465873 0.01664462
## 13 0.003033803     12 0.8150963 0.8465219 0.01667932
## 14 0.002914817     13 0.8120625 0.8466921 0.01669623
## 15 0.002790737     14 0.8091477 0.8452142 0.01665712
## 16 0.002730117     15 0.8063570 0.8450512 0.01661721
## 17 0.002560197     16 0.8036269 0.8476923 0.01669101
## 18 0.002327749     17 0.8010667 0.8483089 0.01668839
## 19 0.002223936     19 0.7964112 0.8451645 0.01661475
## 20 0.002164228     21 0.7919633 0.8447269 0.01660173
##
## Variable importance
##     bmi    time  gender     age smoking     sbp     ldl
##      31      28      27       7       3       3       1
##
## Node number 1: 5000 observations,    complexity param=0.05791818
##   mean=10.06434, MSE=0.355722
##   left son=2 (2427 obs) right son=3 (2573 obs)
##   Primary splits:
##       gender  splits as  RL,        improve=0.057918180, (0 missing)
##       bmi     < 29.65 to the right, improve=0.049512600, (0 missing)
##       time    < 46.5  to the left,  improve=0.041303690, (0 missing)
##       age     < 60.5  to the right, improve=0.018144830, (0 missing)
##       smoking splits as  RRL,       improve=0.008813525, (0 missing)
##   Surrogate splits:
##       sbp  < 119.5 to the left,  agree=0.518, adj=0.007, (0 split)
##       time < 37.5  to the left,  agree=0.518, adj=0.006, (0 split)
##       ldl  < 140.5 to the right, agree=0.517, adj=0.005, (0 split)
##       age  < 66.5  to the right, agree=0.516, adj=0.003, (0 split)
##       bmi  < 20.75 to the left,  agree=0.515, adj=0.002, (0 split)
##
## Node number 2: 2427 observations,    complexity param=0.02517214
##   mean=9.91655, MSE=0.3395789
##   left son=4 (524 obs) right son=5 (1903 obs)
##   Primary splits:
##       bmi     < 29.85 to the right, improve=0.054323820, (0 missing)
##       time    < 46.5  to the left,  improve=0.049666810, (0 missing)
##       age     < 60.5  to the right, improve=0.017649280, (0 missing)
##       smoking splits as  RRL,       improve=0.009762305, (0 missing)
##       ldl     < 141.5 to the right, improve=0.004210017, (0 missing)
##
## Node number 3: 2573 observations,    complexity param=0.0272938
##   mean=10.20375, MSE=0.3309125
##   left son=6 (820 obs) right son=7 (1753 obs)
##   Primary splits:
##       bmi     < 28.95 to the right, improve=0.057015340, (0 missing)
##       time    < 46.5  to the left,  improve=0.036007110, (0 missing)
##       age     < 59.5  to the right, improve=0.023746020, (0 missing)
##       smoking splits as  RRL,       improve=0.010252650, (0 missing)
##       sbp     < 128.5 to the right, improve=0.006713466, (0 missing)
##   Surrogate splits:
##       ldl < 164.5 to the right, agree=0.683, adj=0.005, (0 split)
##       sbp < 108.5 to the left,  agree=0.682, adj=0.002, (0 split)
##
## Node number 4: 524 observations,    complexity param=0.004852623
##   mean=9.657717, MSE=0.3166048
##   left son=8 (207 obs) right son=9 (317 obs)
##   Primary splits:
##       bmi     < 31.45 to the right, improve=0.05202457, (0 missing)
##       time    < 50.5  to the left,  improve=0.04364456, (0 missing)
##       smoking splits as  RRL,       improve=0.01948327, (0 missing)
##       sbp     < 120.5 to the right, improve=0.01676732, (0 missing)
##       age     < 59.5  to the right, improve=0.01274723, (0 missing)
##   Surrogate splits:
##       ldl  < 152   to the right, agree=0.615, adj=0.024, (0 split)
##       time < 31.5  to the left,  agree=0.609, adj=0.010, (0 split)
##       sbp  < 109   to the left,  agree=0.607, adj=0.005, (0 split)
##
## Node number 5: 1903 observations,    complexity param=0.02075938
##   mean=9.987821, MSE=0.3223782
##   left son=10 (137 obs) right son=11 (1766 obs)
##   Primary splits:
##       time    < 46.5  to the left,  improve=0.060185330, (0 missing)
##       age     < 62.5  to the right, improve=0.023530880, (0 missing)
##       ldl     < 141.5 to the right, improve=0.008688514, (0 missing)
##       bmi     < 28.35 to the right, improve=0.008031965, (0 missing)
##       smoking splits as  RRL,       improve=0.007860098, (0 missing)
##
## Node number 6: 820 observations,    complexity param=0.006830408
##   mean=10.00291, MSE=0.3091452
##   left son=12 (162 obs) right son=13 (658 obs)
##   Primary splits:
##       bmi  < 32.25 to the right, improve=0.047923770, (0 missing)
##       time < 47.5  to the left,  improve=0.047697360, (0 missing)
##       age  < 60.5  to the right, improve=0.021396790, (0 missing)
##       ldl  < 73.5  to the right, improve=0.011921340, (0 missing)
##       sbp  < 137.5 to the right, improve=0.008408437, (0 missing)
##
## Node number 7: 1753 observations,    complexity param=0.01056035
##   mean=10.29769, MSE=0.313402
##   left son=14 (53 obs) right son=15 (1700 obs)
##   Primary splits:
##       time    < 39.5  to the left,  improve=0.034188120, (0 missing)
##       age     < 59.5  to the right, improve=0.028638110, (0 missing)
##       smoking splits as  RRL,       improve=0.017123260, (0 missing)
##       sbp     < 128.5 to the right, improve=0.009433160, (0 missing)
##       bmi     < 26.75 to the right, improve=0.005238462, (0 missing)
##
## Node number 8: 207 observations
##   mean=9.498897, MSE=0.2980599
##
## Node number 9: 317 observations,    complexity param=0.002223936
##   mean=9.761427, MSE=0.3014877
##   left son=18 (27 obs) right son=19 (290 obs)
##   Primary splits:
##       time    < 51    to the left,  improve=0.03944451, (0 missing)
##       smoking splits as  RRL,       improve=0.03307206, (0 missing)
##       age     < 58.5  to the right, improve=0.02350311, (0 missing)
##       sbp     < 121.5 to the right, improve=0.02005357, (0 missing)
##       ldl     < 80.5  to the right, improve=0.01272572, (0 missing)
##
## Node number 10: 137 observations,    complexity param=0.003033803
##   mean=9.487714, MSE=0.3536406
##   left son=20 (46 obs) right son=21 (91 obs)
##   Primary splits:
##       time < 36.5  to the left,  improve=0.11137440, (0 missing)
##       sbp  < 118.5 to the right, improve=0.05219794, (0 missing)
##       bmi  < 26.65 to the right, improve=0.04629031, (0 missing)
##       age  < 65.5  to the right, improve=0.04068284, (0 missing)
##       ldl  < 88.5  to the left,  improve=0.02321797, (0 missing)
##
## Node number 11: 1766 observations,    complexity param=0.007483148
##   mean=10.02662, MSE=0.2990454
##   left son=22 (508 obs) right son=23 (1258 obs)
##   Primary splits:
##       age     < 62.5  to the right, improve=0.025202130, (0 missing)
##       time    < 97.5  to the right, improve=0.015378130, (0 missing)
##       ldl     < 142.5 to the right, improve=0.012513270, (0 missing)
##       smoking splits as  RRL,       improve=0.009228690, (0 missing)
##       bmi     < 28.35 to the right, improve=0.008363651, (0 missing)
##   Surrogate splits:
##       sbp < 140.5 to the right, agree=0.737, adj=0.085, (0 split)
##       ldl < 159.5 to the right, agree=0.715, adj=0.008, (0 split)
##
## Node number 12: 162 observations
##   mean=9.757604, MSE=0.2946382
##
## Node number 13: 658 observations,    complexity param=0.006407846
##   mean=10.06331, MSE=0.2942538
##   left son=26 (51 obs) right son=27 (607 obs)
##   Primary splits:
##       time < 47.5  to the left,  improve=0.058863320, (0 missing)
##       age  < 61.5  to the right, improve=0.018077840, (0 missing)
##       bmi  < 31.35 to the right, improve=0.011334980, (0 missing)
##       ldl  < 73.5  to the right, improve=0.009520452, (0 missing)
##       sbp  < 137.5 to the right, improve=0.009235878, (0 missing)
##
## Node number 14: 53 observations
##   mean=9.71145, MSE=0.4536552
##
## Node number 15: 1700 observations,    complexity param=0.007928454
##   mean=10.31597, MSE=0.2979808
##   left son=30 (914 obs) right son=31 (786 obs)
##   Primary splits:
##       age     < 59.5  to the right, improve=0.027837610, (0 missing)
##       smoking splits as  RRL,       improve=0.016679170, (0 missing)
##       time    < 159.5 to the right, improve=0.014116890, (0 missing)
##       sbp     < 128.5 to the right, improve=0.009671418, (0 missing)
##       bmi     < 26.75 to the right, improve=0.004398652, (0 missing)
##   Surrogate splits:
##       sbp  < 126.5 to the right, agree=0.645, adj=0.232, (0 split)
##       ldl  < 91.5  to the right, agree=0.569, adj=0.069, (0 split)
##       time < 176.5 to the left,  agree=0.545, adj=0.017, (0 split)
##       bmi  < 21.35 to the right, agree=0.540, adj=0.005, (0 split)
##
## Node number 18: 27 observations
##   mean=9.404035, MSE=0.2771699
##
## Node number 19: 290 observations,    complexity param=0.002223936
##   mean=9.794701, MSE=0.2907525
##   left son=38 (62 obs) right son=39 (228 obs)
##   Primary splits:
##       time    < 151.5 to the right, improve=0.04911456, (0 missing)
##       smoking splits as  RRL,       improve=0.04214043, (0 missing)
##       sbp     < 121.5 to the right, improve=0.02250072, (0 missing)
##       age     < 58.5  to the right, improve=0.02190715, (0 missing)
##       ldl     < 80.5  to the right, improve=0.01055996, (0 missing)
##
## Node number 20: 46 observations
##   mean=9.208578, MSE=0.2594872
##
## Node number 21: 91 observations
##   mean=9.628815, MSE=0.3419384
##
## Node number 22: 508 observations
##   mean=9.890003, MSE=0.3020677
##
## Node number 23: 1258 observations,    complexity param=0.006412912
##   mean=10.08178, MSE=0.287245
##   left son=46 (250 obs) right son=47 (1008 obs)
##   Primary splits:
##       time    < 149.5 to the right, improve=0.031564790, (0 missing)
##       age     < 54.5  to the right, improve=0.011762700, (0 missing)
##       ldl     < 126.5 to the right, improve=0.011752420, (0 missing)
##       smoking splits as  RRL,       improve=0.010707140, (0 missing)
##       sbp     < 109.5 to the left,  improve=0.007957896, (0 missing)
##   Surrogate splits:
##       ldl < 162.5 to the right, agree=0.802, adj=0.004, (0 split)
##
## Node number 26: 51 observations
##   mean=9.609269, MSE=0.2289111
##
## Node number 27: 607 observations,    complexity param=0.002730117
##   mean=10.10146, MSE=0.2809679
##   left son=54 (225 obs) right son=55 (382 obs)
##   Primary splits:
##       time < 128.5 to the right, improve=0.02847191, (0 missing)
##       age  < 61.5  to the right, improve=0.02293396, (0 missing)
##       bmi  < 31.35 to the right, improve=0.01993356, (0 missing)
##       ldl  < 111.5 to the right, improve=0.01092987, (0 missing)
##       sbp  < 137.5 to the right, improve=0.01007648, (0 missing)
##   Surrogate splits:
##       sbp < 114.5 to the left,  agree=0.636, adj=0.018, (0 split)
##       age < 52.5  to the left,  agree=0.631, adj=0.004, (0 split)
##       bmi < 32.15 to the right, agree=0.631, adj=0.004, (0 split)
##
## Node number 30: 914 observations,    complexity param=0.002914817
##   mean=10.23151, MSE=0.2886223
##   left son=60 (101 obs) right son=61 (813 obs)
##   Primary splits:
##       time    < 163.5 to the right, improve=0.019652420, (0 missing)
##       smoking splits as  RRL,       improve=0.012542120, (0 missing)
##       bmi     < 24.85 to the right, improve=0.010077130, (0 missing)
##       age     < 70.5  to the right, improve=0.004980359, (0 missing)
##       ldl     < 79.5  to the right, improve=0.004331232, (0 missing)
##
## Node number 31: 786 observations,    complexity param=0.002790737
##   mean=10.41418, MSE=0.2909223
##   left son=62 (78 obs) right son=63 (708 obs)
##   Primary splits:
##       smoking splits as  RRL,       improve=0.021707010, (0 missing)
##       time    < 184.5 to the right, improve=0.020284770, (0 missing)
##       sbp     < 129.5 to the right, improve=0.013958710, (0 missing)
##       race    splits as  LRLL,      improve=0.008200332, (0 missing)
##       age     < 54.5  to the right, improve=0.007226599, (0 missing)
##   Surrogate splits:
```

```
##       sbp < 149   to the right, agree=0.903, adj=0.026, (0 split)
##
## Node number 38: 62 observations
##   mean=9.565542, MSE=0.2656172
##
## Node number 39: 228 observations
##   mean=9.857017, MSE=0.2794241
##
## Node number 46: 250 observations
##   mean=9.890584, MSE=0.2961252
##
## Node number 47: 1008 observations,    complexity param=0.003284421
##   mean=10.12921, MSE=0.273727
##   left son=94 (87 obs) right son=95 (921 obs)
##   Primary splits:
##       smoking splits as  RRL,       improve=0.021171970, (0 missing)
##       bmi     < 28.25 to the right, improve=0.016615490, (0 missing)
##       ldl     < 126.5 to the right, improve=0.013732310, (0 missing)
##       age     < 52.5  to the right, improve=0.009102790, (0 missing)
##       sbp     < 109.5 to the left,  improve=0.008120837, (0 missing)
##
## Node number 54: 225 observations
##   mean=9.984915, MSE=0.2376411
##
## Node number 55: 382 observations
##   mean=10.1701, MSE=0.293776
##
## Node number 60: 101 observations
##   mean=10.01783, MSE=0.30527
##
## Node number 61: 813 observations
##   mean=10.25805, MSE=0.2801773
##
## Node number 62: 78 observations
##   mean=10.17476, MSE=0.3236525
##
## Node number 63: 708 observations,    complexity param=0.002327749
##   mean=10.44056, MSE=0.2803057
##   left son=126 (182 obs) right son=127 (526 obs)
##   Primary splits:
##       time < 137.5 to the right, improve=0.019352530, (0 missing)
##       sbp  < 133.5 to the right, improve=0.014196050, (0 missing)
##       age  < 55.5  to the right, improve=0.009180245, (0 missing)
##       race splits as  LRLL,      improve=0.008380573, (0 missing)
##       bmi  < 28.45 to the right, improve=0.007482769, (0 missing)
##   Surrogate splits:
##       ldl < 154.5 to the right, agree=0.744, adj=0.005, (0 split)
##
## Node number 94: 87 observations
##   mean=9.881514, MSE=0.2388818
##
## Node number 95: 921 observations,    complexity param=0.002560197
##   mean=10.1526, MSE=0.2706758
##   left son=190 (236 obs) right son=191 (685 obs)
##   Primary splits:
##       bmi  < 28.25 to the right, improve=0.018266070, (0 missing)
##       ldl  < 127.5 to the right, improve=0.013821740, (0 missing)
##       age  < 55.5  to the right, improve=0.012300610, (0 missing)
##       sbp  < 109.5 to the left,  improve=0.011041330, (0 missing)
##       time < 95.5  to the right, improve=0.007309987, (0 missing)
##   Surrogate splits:
##       ldl < 164   to the right, agree=0.746, adj=0.008, (0 split)
##       age < 46.5  to the left,  agree=0.745, adj=0.004, (0 split)
##
## Node number 126: 182 observations
##   mean=10.31535, MSE=0.2119669
##
## Node number 127: 526 observations,    complexity param=0.002327749
##   mean=10.48388, MSE=0.2966499
##   left son=254 (191 obs) right son=255 (335 obs)
##   Primary splits:
##       sbp  < 129.5 to the right, improve=0.028452590, (0 missing)
##       time < 50.5  to the left,  improve=0.013522720, (0 missing)
##       race splits as  LRLL,      improve=0.012365230, (0 missing)
##       bmi  < 28.35 to the right, improve=0.009747016, (0 missing)
##       age  < 54.5  to the right, improve=0.007912211, (0 missing)
##   Surrogate splits:
##       ldl  < 145.5 to the right, agree=0.646, adj=0.026, (0 split)
##       bmi  < 22.7  to the left,  agree=0.644, adj=0.021, (0 split)
##       time < 42.5  to the left,  agree=0.641, adj=0.010, (0 split)
##
## Node number 190: 236 observations
##   mean=10.03281, MSE=0.2619733
##
## Node number 191: 685 observations
##   mean=10.19387, MSE=0.2670264
##
## Node number 254: 191 observations
##   mean=10.36221, MSE=0.2972116
##
## Node number 255: 335 observations
##   mean=10.55325, MSE=0.2830768
```

Comparison

```r
resamp = resamples(list(elastic_net = enet_fit,
                        pcr = pcr_fit,
                        pls = pls_fit,
                        gam = gam.fit,
                        mars = mars.fit))
summary(resamp)
```

```
##
## Call:
## summary.resamples(object = resamp)
##
```

```
## Models: elastic_net, pcr, pls, gam, mars
## Number of resamples: 10
##
## MAE
##                  Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## elastic_net 0.4294666 0.4333146 0.4414904 0.4404188 0.4479709 0.4502491    0
## pcr         0.4283770 0.4331431 0.4418042 0.4406867 0.4484017 0.4516552    0
## pls         0.4283860 0.4331977 0.4418812 0.4407023 0.4483486 0.4517004    0
## gam         0.4038951 0.4165022 0.4250940 0.4229309 0.4317016 0.4337601    0
## mars        0.4068067 0.4144508 0.4248340 0.4221903 0.4299231 0.4327355    0
##
## RMSE
##                  Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## elastic_net 0.5346950 0.5401593 0.5540536 0.5525646 0.5599590 0.5733181    0
## pcr         0.5348357 0.5398323 0.5544163 0.5528409 0.5604510 0.5739030    0
## pls         0.5347609 0.5398986 0.5544173 0.5528333 0.5603197 0.5740022    0
## gam         0.5101573 0.5227377 0.5286720 0.5286310 0.5327635 0.5511329    0
## mars        0.5115426 0.5227742 0.5278403 0.5279391 0.5326822 0.5440279    0
##
## Rsquared
##                  Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## elastic_net 0.1039358 0.1362602 0.1422293 0.1424292 0.1471805 0.1856091    0
## pcr         0.1036311 0.1369615 0.1396188 0.1414866 0.1456933 0.1846009    0
## pls         0.1037262 0.1368722 0.1397305 0.1415124 0.1458437 0.1846014    0
## gam         0.1629446 0.2041048 0.2183934 0.2154693 0.2374062 0.2522148    0
## mars        0.1561416 0.2106884 0.2225692 0.2174307 0.2382161 0.2471323    0
```

```r
bwplot(resamp, metric = "RMSE")
```

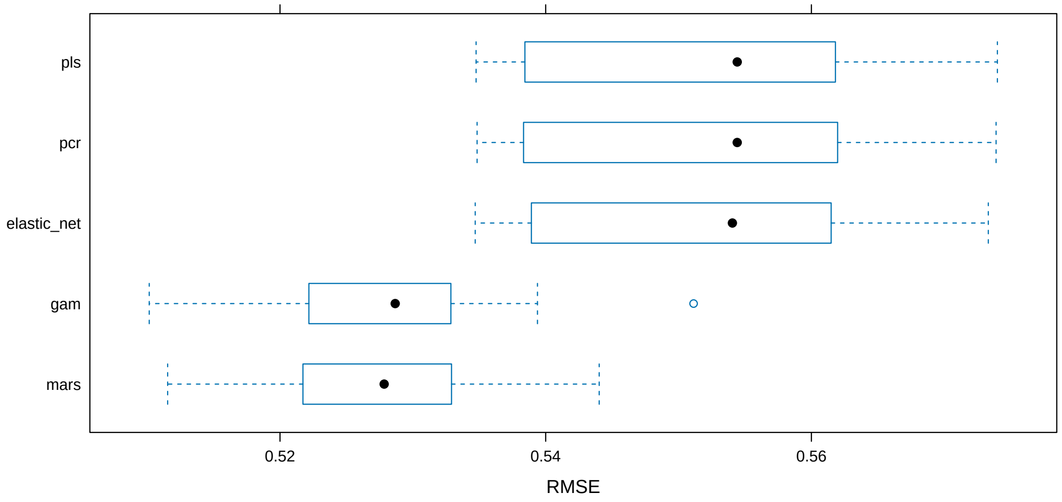

Figure 12: Model Selection

Model Performance

```
predicted_values = predict(mars.fit$finalModel, newdata = x_test)

residuals = y_test - predicted_values

plot(predicted_values, residuals,
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Residuals vs Fitted Values (MARS)",
     pch = 20, col = "mediumpurple1")
abline(h = 0, col = "blue", lwd = 2)
```

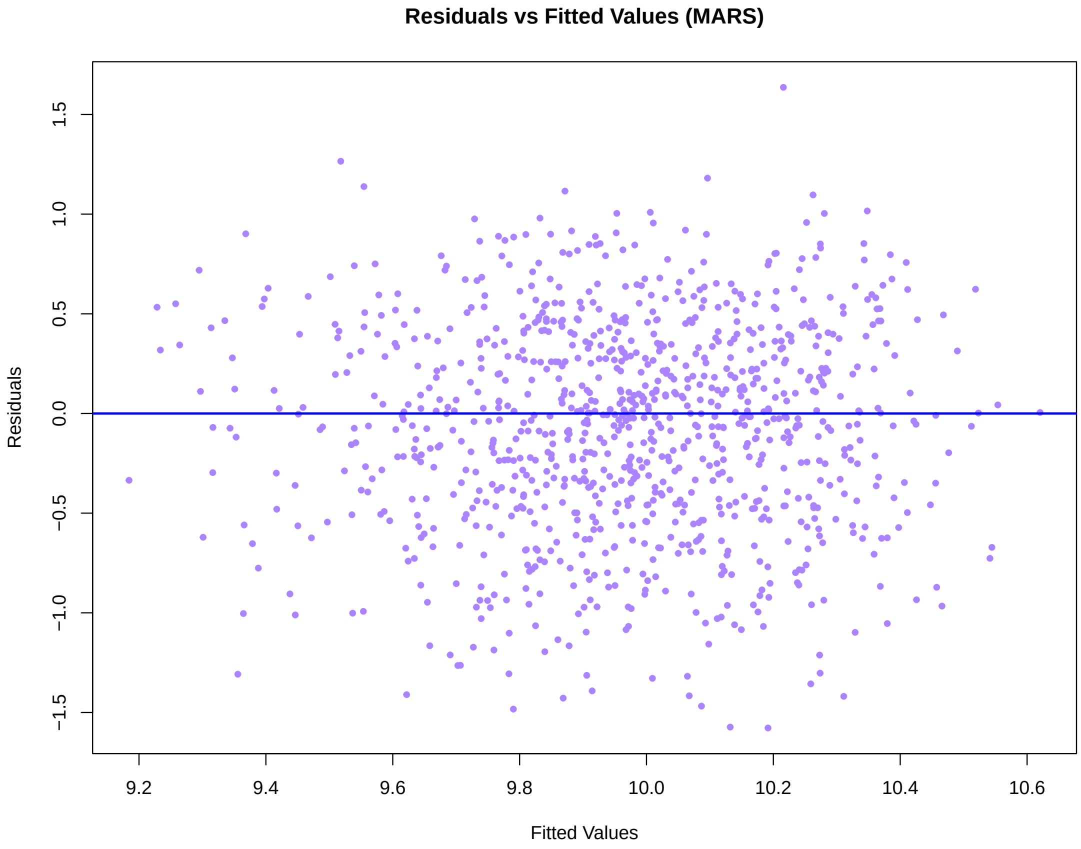

Figure 13: Residuals vs Fitted Values (MARS)

```
plot(y_test, predicted_values,
     xlab = "Actual Values", ylab = "Predicted Values",
     main = "Prediction vs Actual (MARS)",
     pch = 20, col = "orange")
abline(0, 1, col = "mediumseagreen", lwd = 2)
```

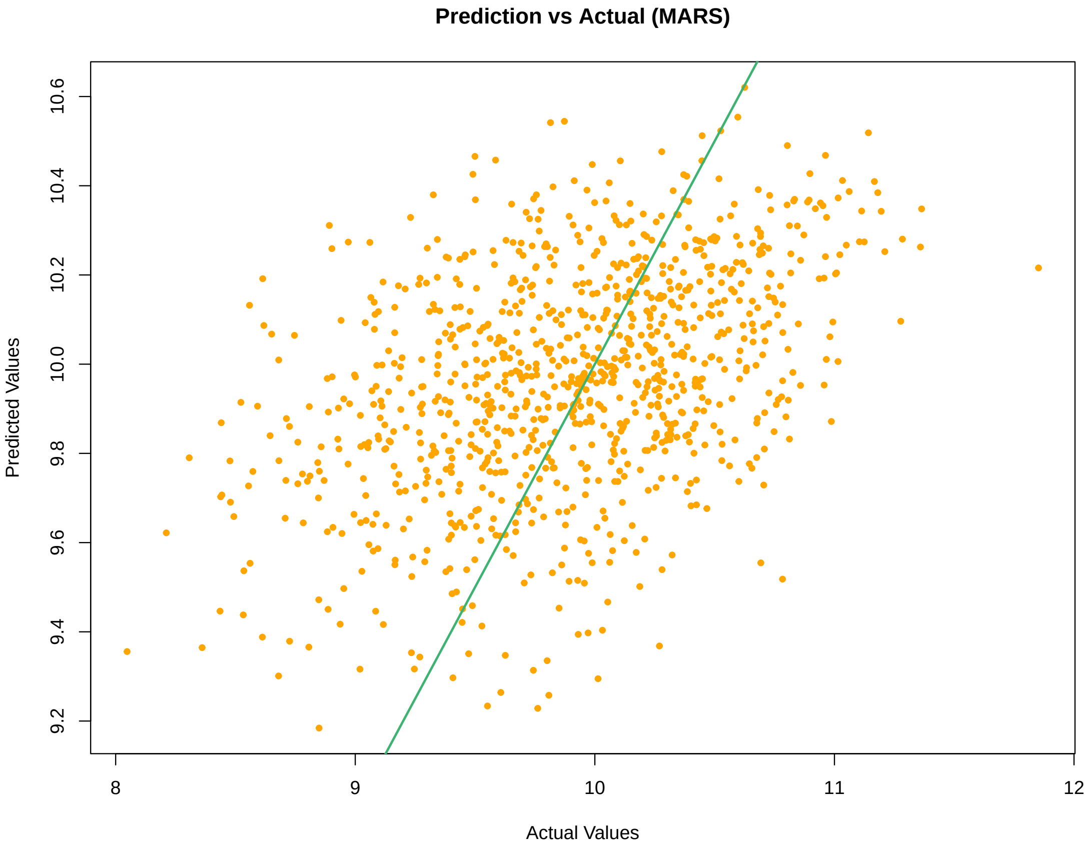

Figure 14: Prediction vs Actual (MARS)

```r
rmse = sqrt(mean((y_test - predicted_values)^2))
rmse
```

```
## [1] 0.5327718
```