Midterm Project

Ravi Brenner, Cameron Chesbrough, Wayne Monical

2025-03-25

Library Load

```
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```
library(tidymodels)
```

```
## -- Attaching packages -------------------------------------- tidymodels 1.3.0 --

## v broom        1.0.7     v rsample      1.2.1
## v dials        1.4.0     v tibble       3.2.1
## v dplyr        1.1.4     v tidyr        1.3.1
## v infer        1.0.7     v tune         1.3.0
## v modeldata    1.4.0     v workflows    1.2.0
## v parsnip      1.3.1     v workflowsets 1.1.0
## v purrr        1.0.4     v yardstick    1.3.2
## v recipes      1.1.1

## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x purrr::discard()         masks scales::discard()
## x dplyr::filter()          masks stats::filter()
## x dplyr::lag()             masks stats::lag()
## x purrr::lift()            masks caret::lift()
## x yardstick::precision()   masks caret::precision()
## x yardstick::recall()      masks caret::recall()
## x yardstick::sensitivity() masks caret::sensitivity()
## x yardstick::specificity() masks caret::specificity()
## x recipes::step()          masks stats::step()
```

```
library(vtable)
```

```
## Loading required package: kableExtra

##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```r
library(patchwork)
library(splines)
library(mgcv)
```

```
## Loading required package: nlme

##
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse

## This is mgcv 1.9-1. For overview type 'help("mgcv-package")'.
```

```r
library(pdp)
```

```
##
## Attaching package: 'pdp'

## The following object is masked from 'package:purrr':
##
##     partial
```

```r
library(earth)
```

```
## Loading required package: Formula

## Loading required package: plotmo

## Loading required package: plotrix

##
## Attaching package: 'plotrix'

## The following object is masked from 'package:scales':
##
##     rescale
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v lubridate 1.9.4     v stringr   1.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x readr::col_factor()      masks scales::col_factor()
## x nlme::collapse()         masks dplyr::collapse()
## x purrr::discard()         masks scales::discard()
## x dplyr::filter()          masks stats::filter()
## x stringr::fixed()         masks recipes::fixed()
## x kableExtra::group_rows() masks dplyr::group_rows()
## x dplyr::lag()             masks stats::lag()
## x purrr::lift()            masks caret::lift()
## x pdp::partial()           masks purrr::partial()
## x readr::spec()            masks yardstick::spec()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(ggplot2)
library(glmnet)
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
## Loaded glmnet 4.1-8
```

Loading data

```
load("dat1.RData")
load("dat2.RData")
```

Exploratory Data Analysis

Go from labels to variable names

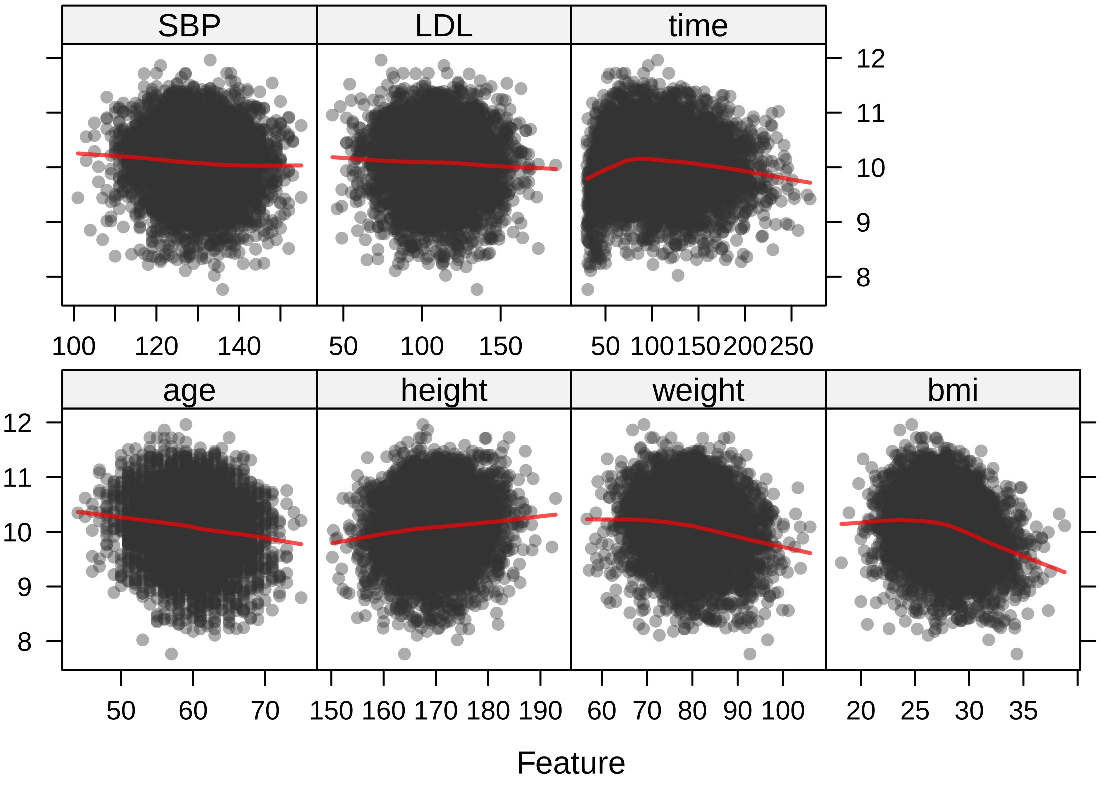
```
dat1 <- dat1 |>
  mutate(gender = factor(gender, levels = c(0,1),
                         labels = c("Female","Male")),
         race = factor(race, levels = c(1,2,3,4),
                       labels = c("White","Asian","Black","Hispanic")),
         smoking = factor(smoking, levels = c(0,1,2),
                          labels = c("Never smoked","Former smoker","Current smoker")),
         diabetes = factor(diabetes, levels = c(0,1),
                           labels = c("No","Yes")),
         hypertension = factor(hypertension, levels = c(0,1),
                               labels = c("No","Yes")),
         ) |>
  dplyr::select(-id)

dat2 <- dat2 |>
  mutate(gender = factor(gender, levels = c(0,1),
                         labels = c("Female","Male")),
         race = factor(race, levels = c(1,2,3,4),
                       labels = c("White","Asian","Black","Hispanic")),
         smoking = factor(smoking, levels = c(0,1,2),
                          labels = c("Never smoked","Former smoker","Current smoker")),
         diabetes = factor(diabetes, levels = c(0,1),
                           labels = c("No","Yes")),
         hypertension = factor(hypertension, levels = c(0,1),
                               labels = c("No","Yes")),
         ) |>
  dplyr::select(-id)
```

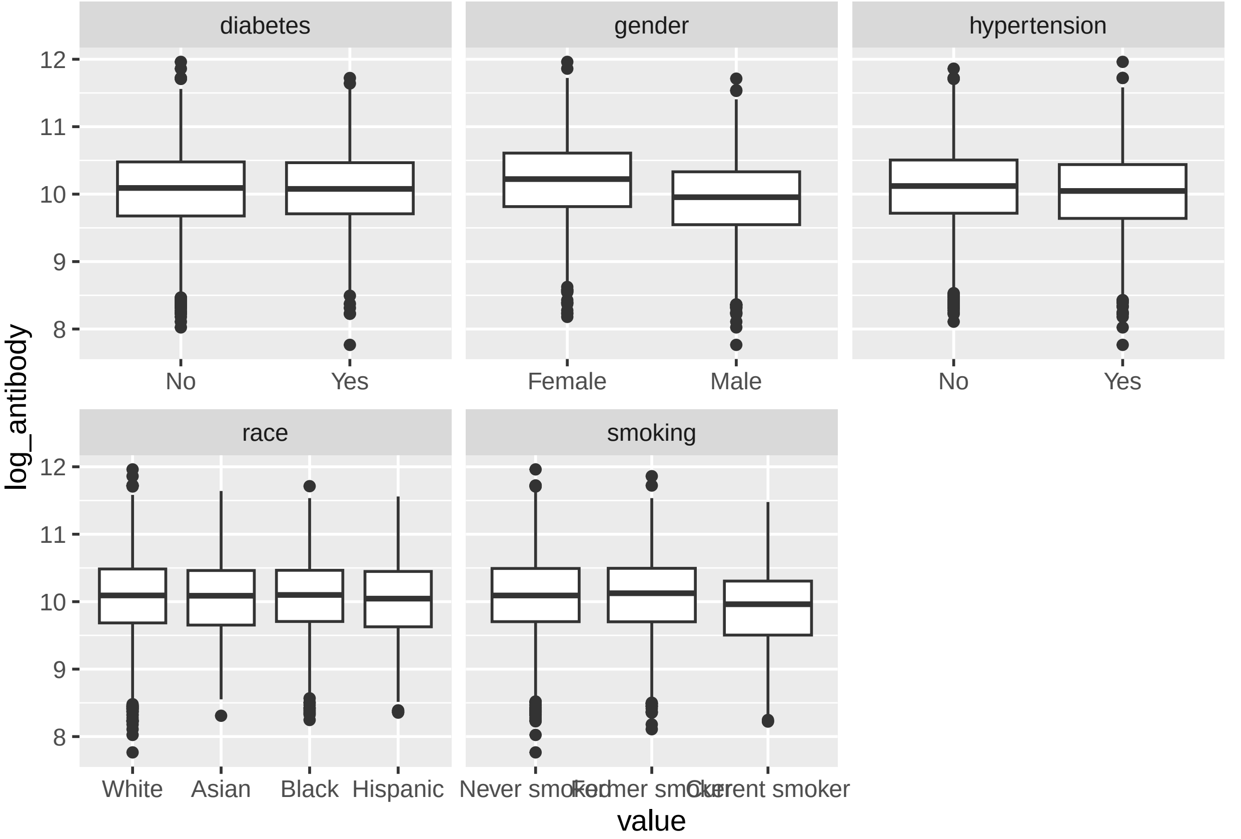
Use featureplot from caret to plot the training data

```
theme1 <- trellis.par.get()
theme1$plot.symbol$col = rgb(.2, .2, .2, .4)
theme1$plot.symbol$pch = 16
theme1$plot.line$col = rgb(1, 0, 0, .7)
theme1$plot.line$lwd <- 2
trellis.par.set(theme1)
```

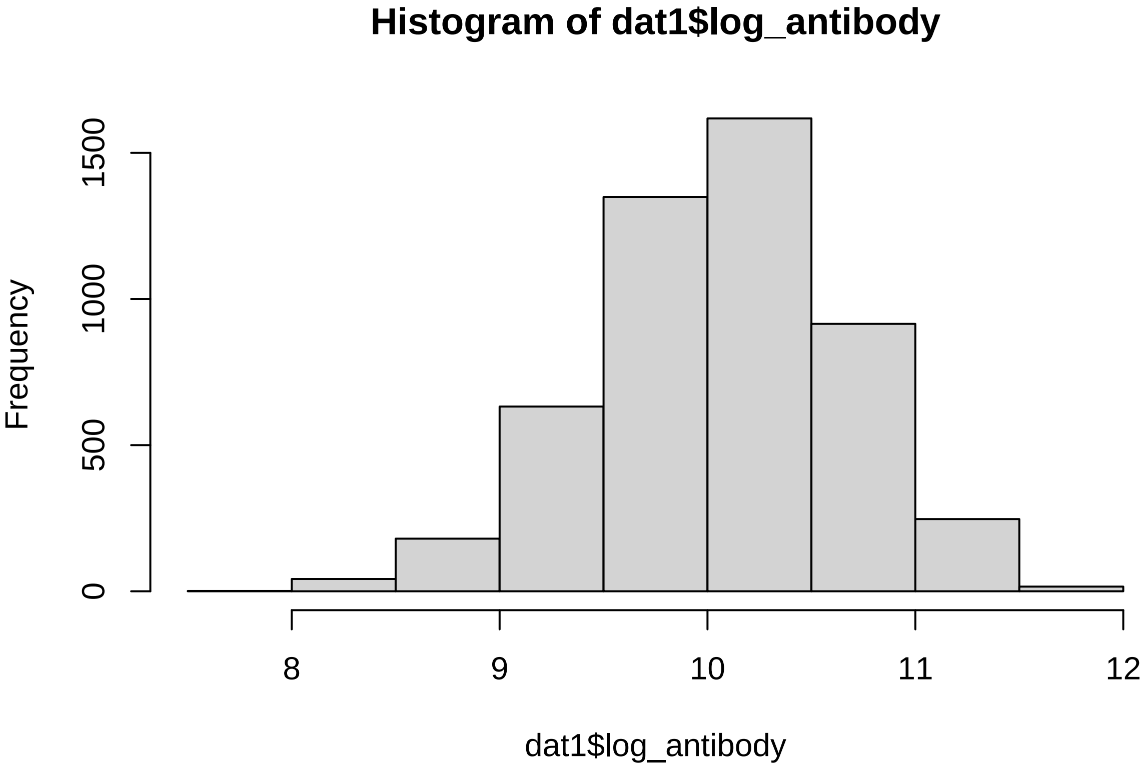
```
featurePlot(x = dat1 |> dplyr::select(-log_antibody,
                                      -where(is.factor)),
            y = dat1$log_antibody,
            type = c("p","smooth"))
```



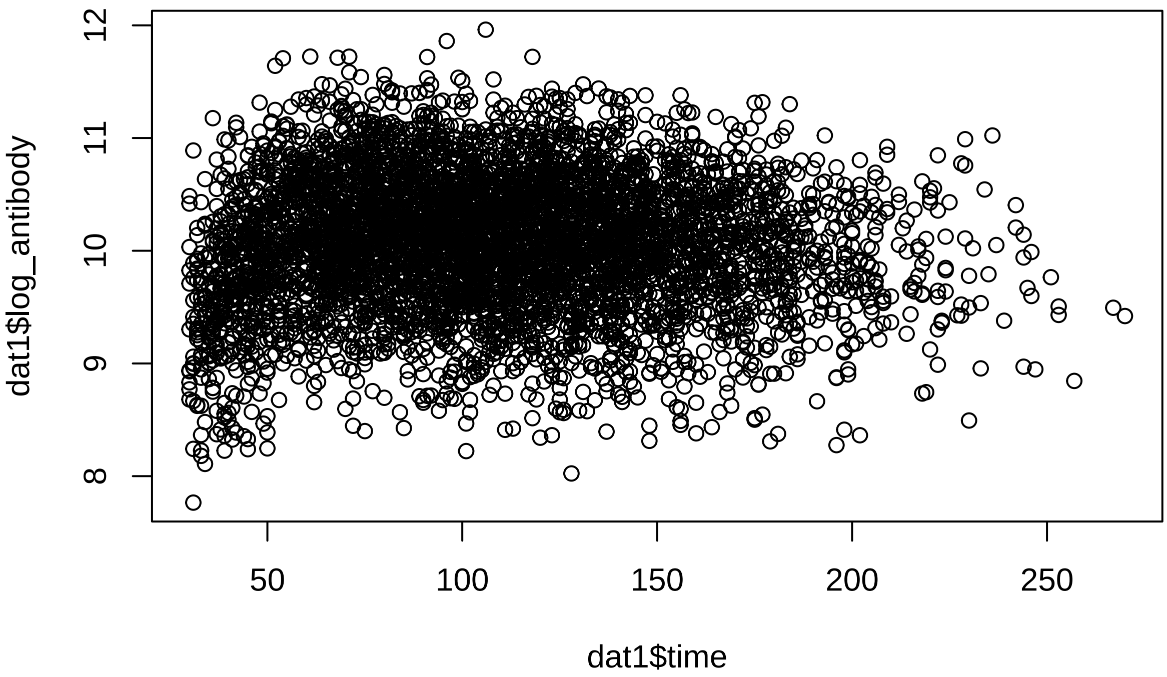
```
dat1 |>
  dplyr::select(log_antibody,gender,race, smoking,diabetes, hypertension) |>
  pivot_longer(cols = 2:6,
               names_to = "variable",
               values_to = "value") |>
  ggplot(aes(x = value, y = log_antibody)) +
  geom_boxplot() +
  facet_wrap(.~variable,scales = "free_x")
```



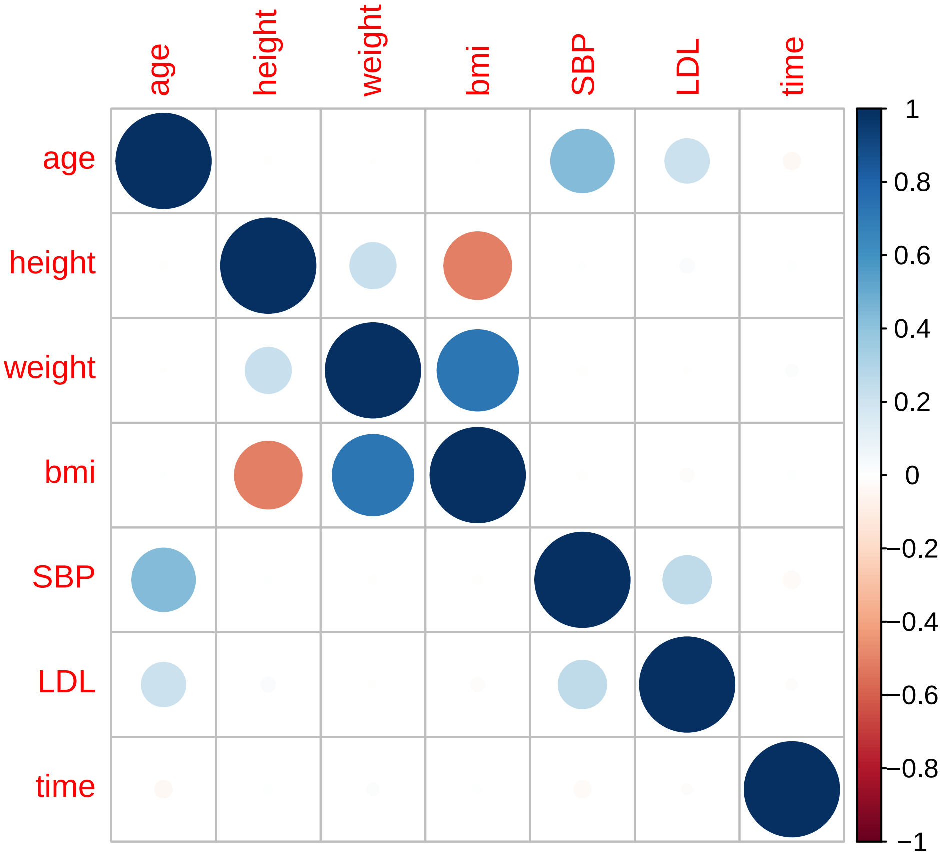
```r
antibody_hist = hist(dat1$log_antibody)
```



```
antibody_scatter = plot(x = dat1$time, y = dat1$log_antibody)
```



```
summ_table = sumtable(dat1, out = 'return')
continuous = dat1[c(1,5:7,10:12)]
correlations = cor(continuous)
corr_plot = corrplot(correlations)
```



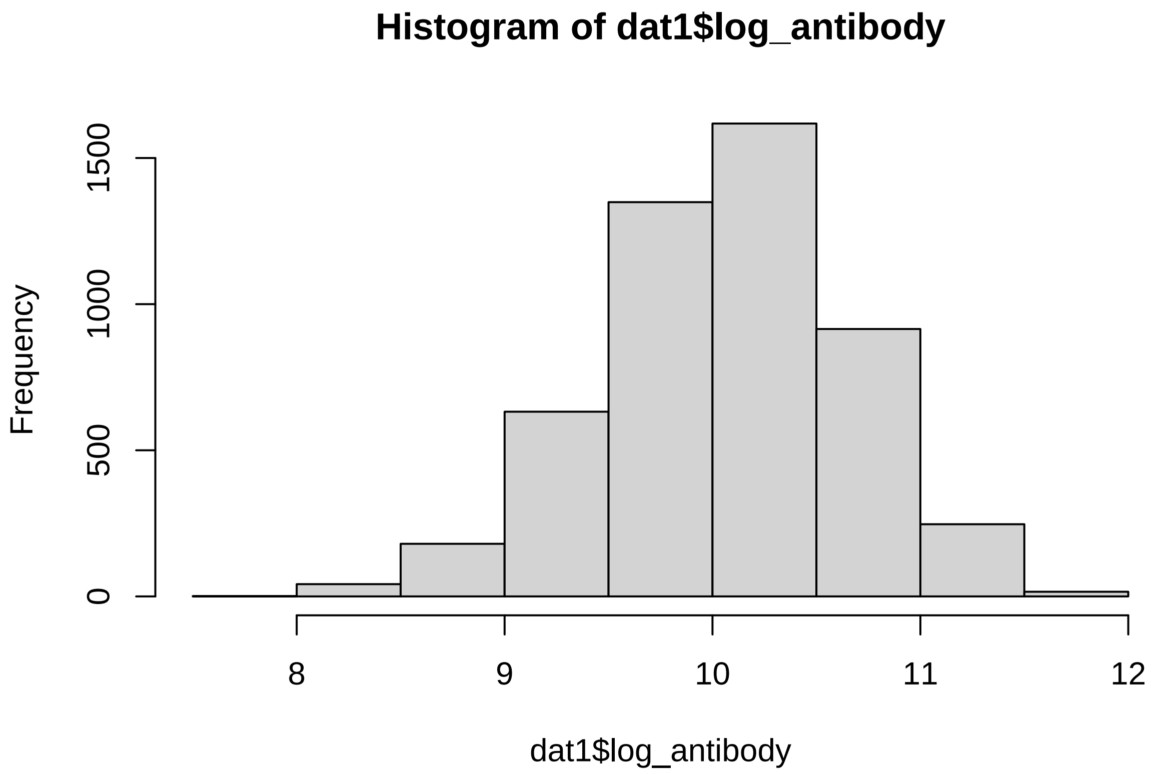
```
report_table = sumtable(dat1, out = 'kable')
report_table
```

Looking at the second dataset

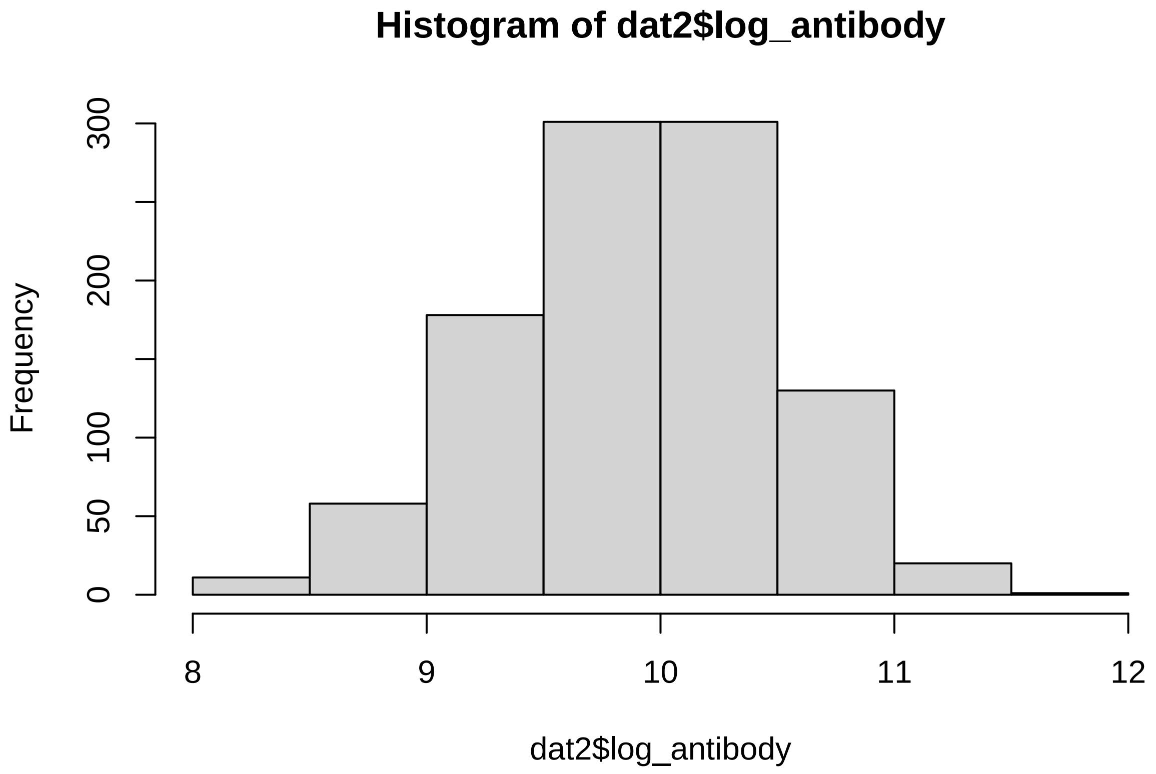
```
antibody_hist = hist(dat1$log_antibody)
```

Table 1: Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
age	5000	60	4.5	44	57	63	75
gender	5000						
... Female	2573	51%					
... Male	2427	49%					
race	5000						
... White	3221	64%					
... Asian	278	6%					
... Black	1036	21%					
... Hispanic	465	9%					
smoking	5000						
... Never smoked	3010	60%					
... Former smoker	1504	30%					
... Current smoker	486	10%					
height	5000	170	5.9	150	166	174	193
weight	5000	80	7.1	57	75	85	106
bmi	5000	28	2.8	18	26	30	39
diabetes	5000						
... No	4228	85%					
... Yes	772	15%					
hypertension	5000						
... No	2702	54%					
... Yes	2298	46%					
SBP	5000	130	8	101	124	135	155
LDL	5000	110	20	43	96	124	185
time	5000	109	43	30	76	138	270
log_antibody	5000	10	0.6	7.8	9.7	10	12



```
antibody_hist_data2 = hist(dat2$log_antibody)
```



```
summ_table_data2 = sumtable(dat2, out = 'return')
```

Model Training

Creating the design Matrix

```r
load('dat1.RData')

design_matrix =
  dat1 |>
  mutate(
    race_asian = as.numeric(race == 2),
    race_black = as.numeric(race == 3),
    race_hispanic = as.numeric(race == 4),
    smoking_former = as.numeric(smoking == 1),
    smoking_current = as.numeric(smoking ==2)
  ) %>%
  select(
    age, gender, race_asian, race_black, race_hispanic,
    smoking_former, smoking_current, height, weight,
    bmi, diabetes, hypertension, SBP, LDL, time,
    log_antibody
  )
```

Specify X and Y for model training

```r
y = design_matrix$log_antibody
x = select(design_matrix, -log_antibody) %>%
  as.matrix()
```

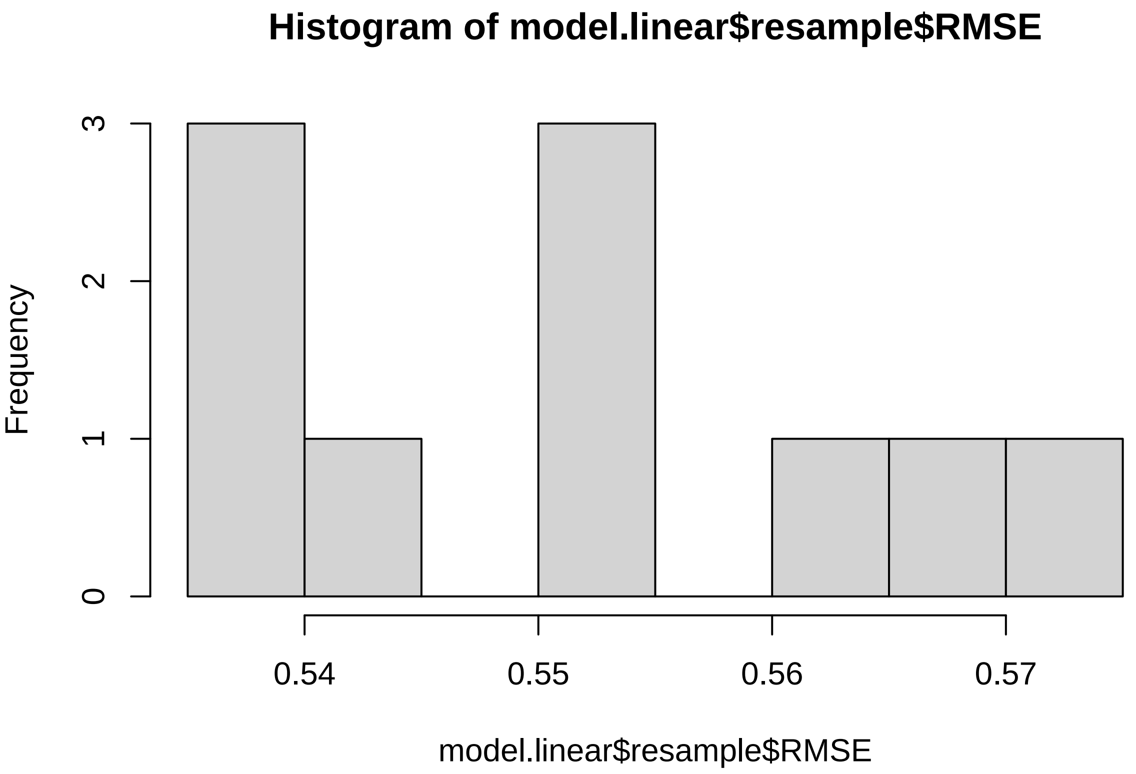
Specify CV Procedure

```r
ctrl <- trainControl(method = "cv", number = 10)
```

Linear Regression

```r
set.seed(1)
model.linear =
  train(x = x,
        y = y,
        method = "lm",
        metric = "RMSE",
        trControl = ctrl)

hist(model.linear$resample$RMSE)
```



```
coef(model.linear$finalModel)
```

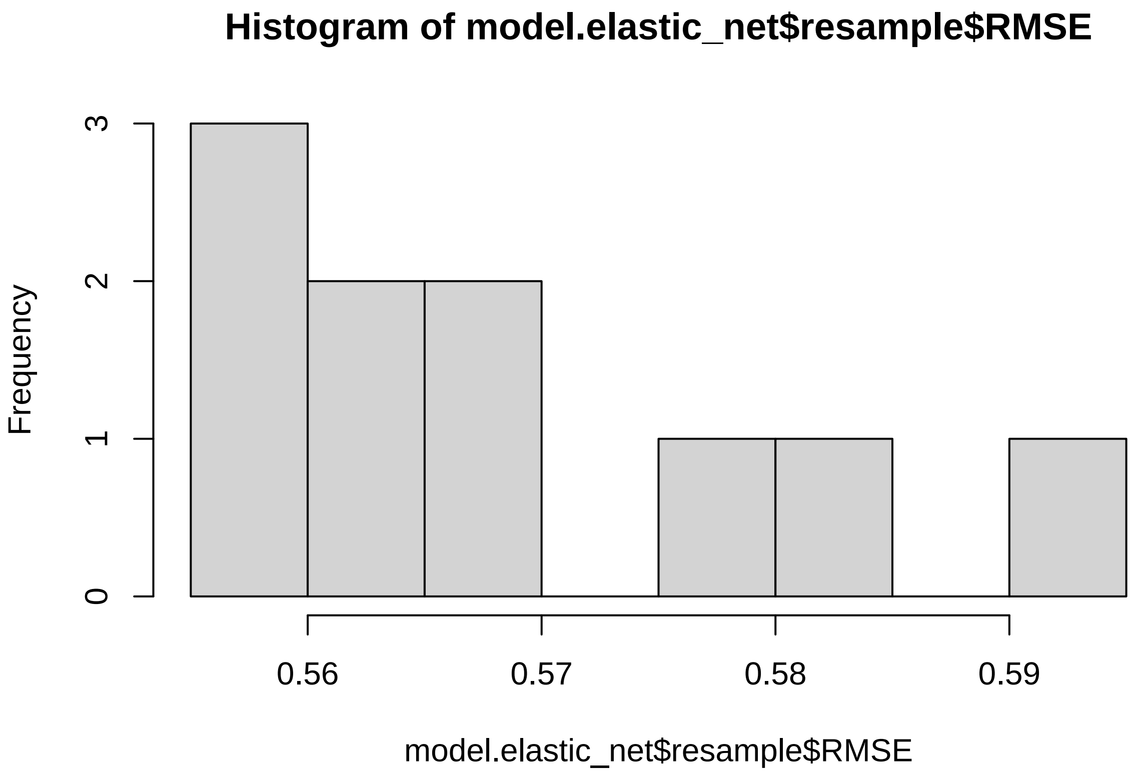
```
##     (Intercept)             age          gender      race_asian      race_black 
##   26.6751961468   -0.0205978746   -0.2974929370   -0.0060422043   -0.0075294859 
##   race_hispanic  smoking_former smoking_current          height          weight 
##   -0.0417570580    0.0219906714   -0.1934834467   -0.0821380676    0.0859034194 
##             bmi        diabetes    hypertension             SBP             LDL 
##   -0.2977934503    0.0112794933   -0.0179106155    0.0015181119   -0.0001645307 
##            time 
##   -0.0003010641
```

Elastic Net

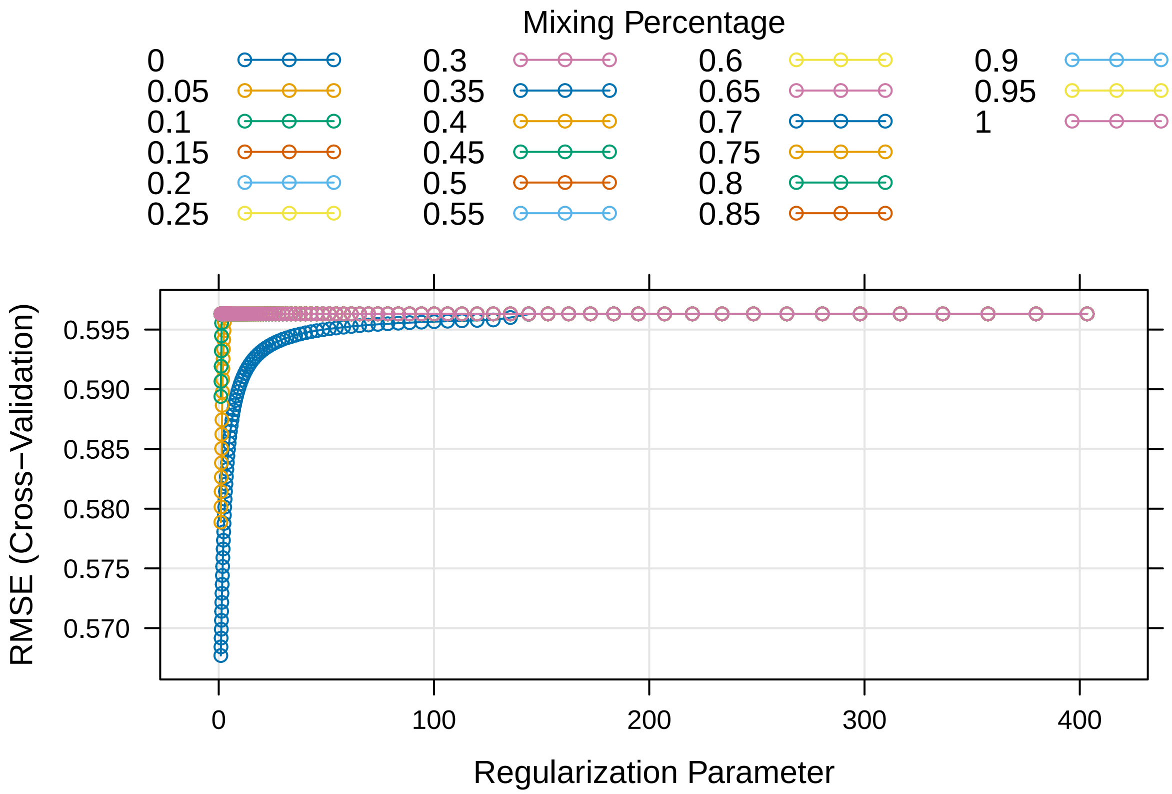
```
set.seed(1)
model.elastic_net =
  train(x = x,
        y = y,
        method = "glmnet",
        metric = "RMSE",
        trControl = ctrl,
        tuneGrid = expand.grid(.alpha = seq(0,1, length = 21),
                               .lambda = exp(seq(6,0, length = 100))))
```

```
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.
```

```
hist(model.elastic_net$resample$RMSE)
```



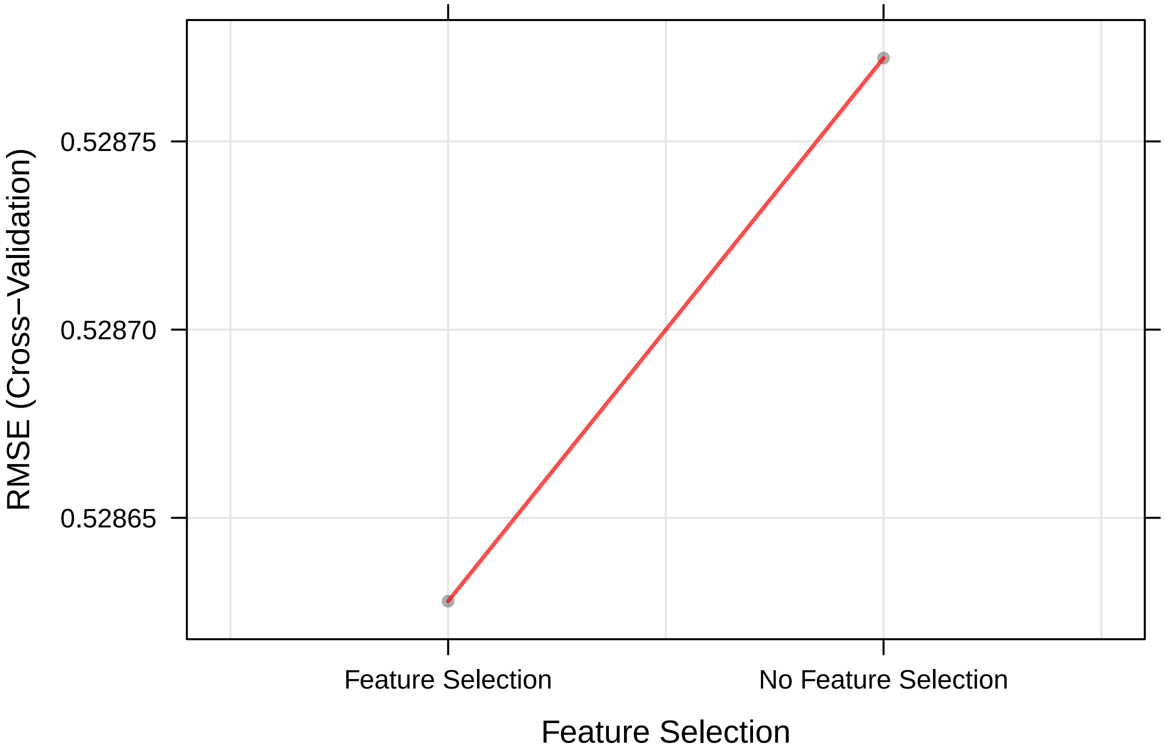
```r
plot(model.elastic_net)
```



GAM

```
set.seed(1)
model.gam =
  train(x = x,
        y = y,
        method = "gam",
        metric = "RMSE",
        trControl = ctrl)
```

```
plot(model.gam)
```

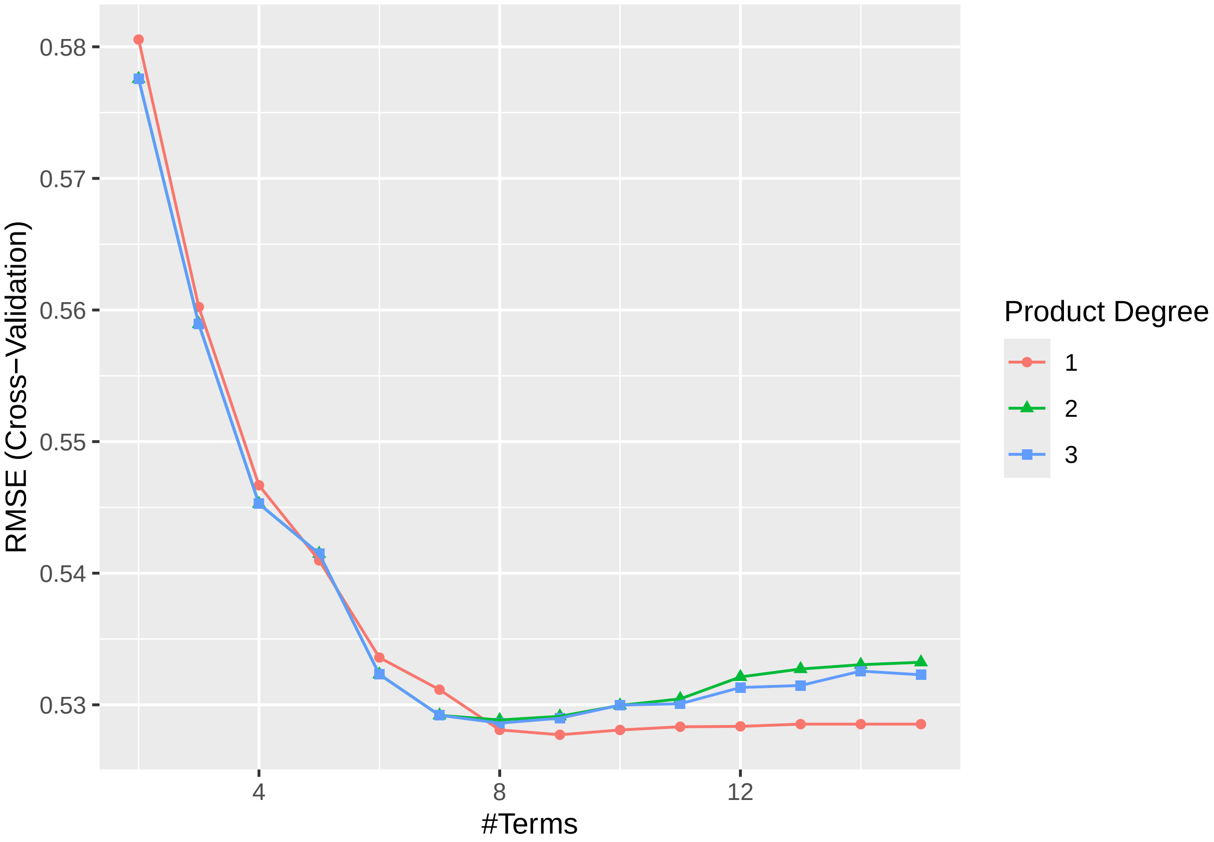


MARS

```
mars_grid =
  expand.grid(degree = 1:3,
              nprune = 2:15)

set.seed(1)
model.mars =
  train(x, y,
        method = "earth",
        tuneGrid = mars_grid,
        trControl = ctrl)

ggplot(model.mars) +
  labs('MARS Model Evaluation')
```



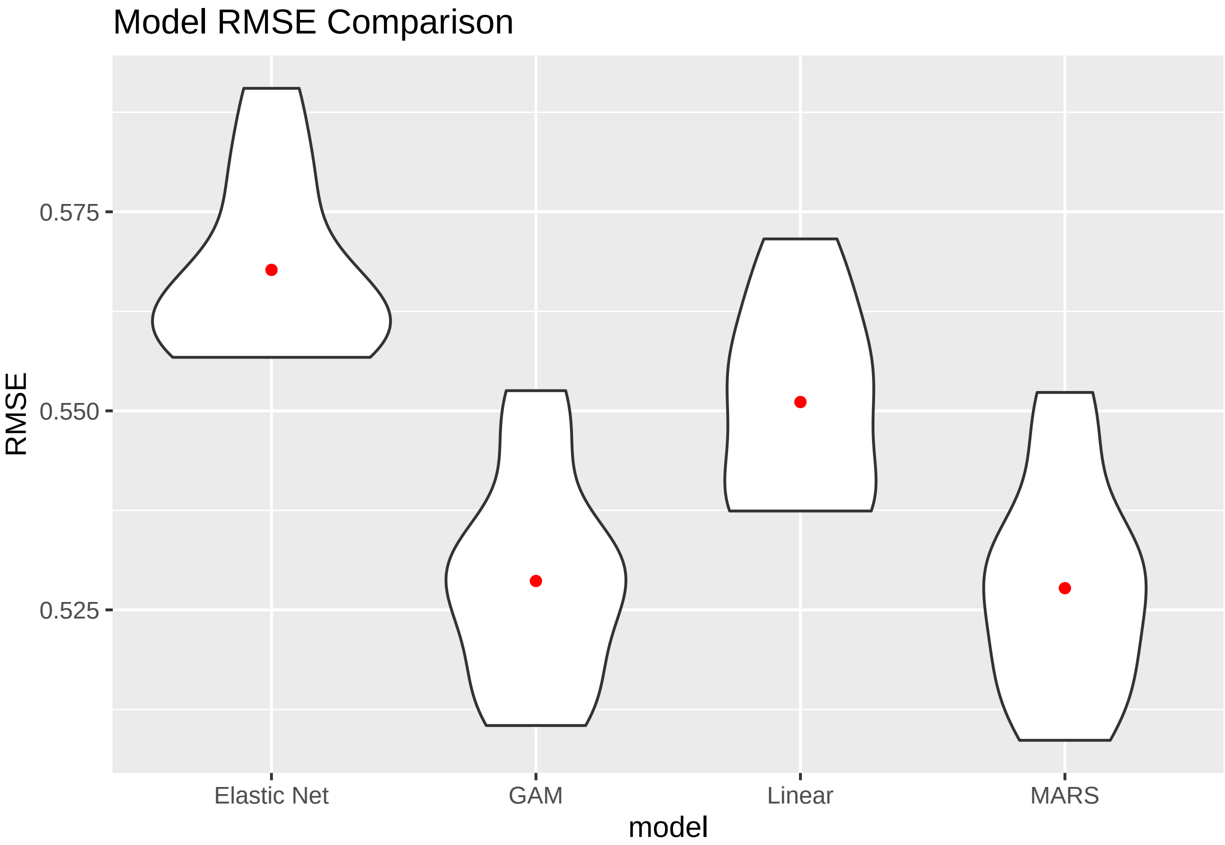
Comparing Cross Validated RMSE

```
model.RMSE=
  rbind(
  data.frame(
    model = 'Linear',
    RMSE = model.linear$resample$RMSE
  ),
  data.frame(
    model = 'Elastic Net',
    RMSE = model.elastic_net$resample$RMSE),
  data.frame(
    model = 'GAM',
    RMSE = model.gam$resample$RMSE
  ),
  data.frame(
    model = 'MARS',
    RMSE = model.mars$resample$RMSE
  )
)
```

```
model.RMSE %>%
  group_by(model) %>%
  summarize(mean(RMSE))
```

```
## # A tibble: 4 x 2
##   model       `mean(RMSE)`
##   <chr>              <dbl>
## 1 Elastic Net        0.568
## 2 GAM                0.529
## 3 Linear             0.551
## 4 MARS               0.528
```

```
model.RMSE |>
  ggplot(aes(x = model, y = RMSE)) +
  geom_violin()+
  stat_summary(
    fun = "mean",
               geom = "point",
               color = "red")+
  labs(title = "Model RMSE Comparison", xlab = "Model")
```



Model Evaluation

Clean new data set

```
load('dat2.RData')

design_matrix2 =
  dat2 |>
  mutate(
    race_asian = as.numeric(race == 2),
    race_black = as.numeric(race == 3),
    race_hispanic = as.numeric(race == 4),
    smoking_former = as.numeric(smoking == 1),
    smoking_current = as.numeric(smoking ==2)
  ) %>%
  select(
    age, gender, race_asian, race_black, race_hispanic,
    smoking_former, smoking_current, height, weight,
    bmi, diabetes, hypertension, SBP, LDL, time,
    log_antibody
  )

x2 = as.matrix(dplyr::select(design_matrix2, -log_antibody))
y2 = design_matrix2$log_antibody
```

```
head(x)
```

```
##   age gender race_asian race_black race_hispanic smoking_former smoking_current
## 1  50      0          0          0             0              0               0
## 2  71      1          0          0             0              0               0
## 3  58      1          0          0             0              1               0
## 4  63      0          0          0             0              0               0
## 5  56      1          0          0             0              0               0
## 6  59      1          0          1             0              0               0
##   height weight  bmi diabetes hypertension SBP LDL time
## 1  176.1   68.3 22.0        0            0 130  82   76
## 2  175.7   69.6 22.6        0            1 149 129   82
## 3  168.7   76.9 27.0        0            0 127 101  168
## 4  167.4   90.0 32.1        0            1 138  93  105
## 5  162.7   83.9 31.7        0            0 123  97  193
## 6  167.8   86.8 30.8        0            1 132 108  143
```

```
head(x2)
```

```
##      age gender race_asian race_black race_hispanic smoking_former
## 5001  58      0          0          0             1              1
## 5002  62      0          0          0             0              1
## 5003  71      0          0          0             1              0
## 5004  59      1          0          0             0              0
## 5005  69      1          0          0             0              0
## 5006  56      0          0          0             0              0
##      smoking_current height weight  bmi diabetes hypertension SBP LDL time
## 5001               0  176.4   86.4 27.7        0            0 130 115  205
## 5002               0  167.5   82.4 29.4        1            0 123 118  229
## 5003               0  179.3   79.2 24.6        1            1 145 149  206
## 5004               0  170.0   81.0 28.0        0            0 123 119  163
## 5005               0  166.5   74.8 27.0        1            1 150 142  240
## 5006               0  167.6   74.8 26.6        0            0 121 112  206
```

Make predictions and get test set RMSE

```
pred = predict(model.gam, x2)

dat2_rmse = sqrt(mean((pred - y2)^2))

dat2_rmse
```

```
## [1] 0.5700836
```