Midterm

Sining Leng

2025-03-24

Load data

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

dat1 =
  dat1 |>
  select(-id)

dat2 =
  dat2 |>
  select(-id)
```

Explorey Data Analysis and Visualization

Summary statistics

dat1

```
summarize_cat = function(data, var_name) {

  var_label = deparse(substitute(var_name))

  out_df =
    data |>
    count({{ var_name }}) |>
    mutate(
      Percent = round(100 * n / sum(n), 1),
      Variable = var_label,
      Level = as.character({{ var_name }})
    ) |>
    rename(N = n) |>
    select(Variable, Level, N, Percent)

  return(out_df)
}

gender <- summarize_cat(dat1, gender)
race <- summarize_cat(dat1, race)
smoking <- summarize_cat(dat1, smoking)
```

```r
diabetes <- summarize_cat(dat1, diabetes)
hypertension <- summarize_cat(dat1, hypertension)
bind_rows(gender, race, smoking, diabetes, hypertension) %>%
  knitr::kable()
```

Variable	Level	N	Percent
gender	0	2573	51.5
gender	1	2427	48.5
race	1	3221	64.4
race	2	278	5.6
race	3	1036	20.7
race	4	465	9.3
smoking	0	3010	60.2
smoking	1	1504	30.1
smoking	2	486	9.7
diabetes	0	4228	84.6
diabetes	1	772	15.4
hypertension	0	2702	54.0
hypertension	1	2298	46.0

```r
summarize_cont = function(data, var_name) {

  var_label = deparse(substitute(var_name))

  out_df =
    data |>
    summarize(
      Variable = var_label,
      Median = round(median({{ var_name }}, na.rm = TRUE), 1),
      Q1 = round(quantile({{ var_name }}, 0.25, na.rm = TRUE), 1),
      Q3 = round(quantile({{ var_name }}, 0.75, na.rm = TRUE), 1)
    ) |>
    mutate(
      IQR = paste0("[", Q1, ", ", Q3, "]")
    ) |>
    select(Variable, Median, IQR)

  return(out_df)
}

age = summarize_cont(dat1, age)
bmi = summarize_cont(dat1, bmi)
height = summarize_cont(dat1, height)
weight = summarize_cont(dat1, weight)
SBP = summarize_cont(dat1, SBP)
LDL = summarize_cont(dat1, LDL)
time = summarize_cont(dat1, time)
log_anti = summarize_cont(dat1, log_antibody)
bind_rows(age, bmi, height, weight, SBP, LDL, time, log_anti) %>%
  knitr::kable()
```

Variable	Median	IQR
age	60.0	[57, 63]
bmi	27.6	[25.8, 29.5]
height	170.1	[166.1, 174.2]
weight	80.1	[75.4, 84.9]
SBP	130.0	[124, 135]
LDL	110.0	[96, 124]
time	106.0	[76, 138]
log_antibody	10.1	[9.7, 10.5]

dat2

```
gender <- summarize_cat(dat2, gender)
race <- summarize_cat(dat2, race)
smoking <- summarize_cat(dat2, smoking)
diabetes <- summarize_cat(dat2, diabetes)
hypertension <- summarize_cat(dat2, hypertension)
bind_rows(gender, race, smoking, diabetes, hypertension) %>%
  knitr::kable()
```

Variable	Level	N	Percent
gender	0	509	50.9
gender	1	491	49.1
race	1	663	66.3
race	2	55	5.5
race	3	199	19.9
race	4	83	8.3
smoking	0	601	60.1
smoking	1	296	29.6
smoking	2	103	10.3
diabetes	0	843	84.3
diabetes	1	157	15.7
hypertension	0	544	54.4
hypertension	1	456	45.6

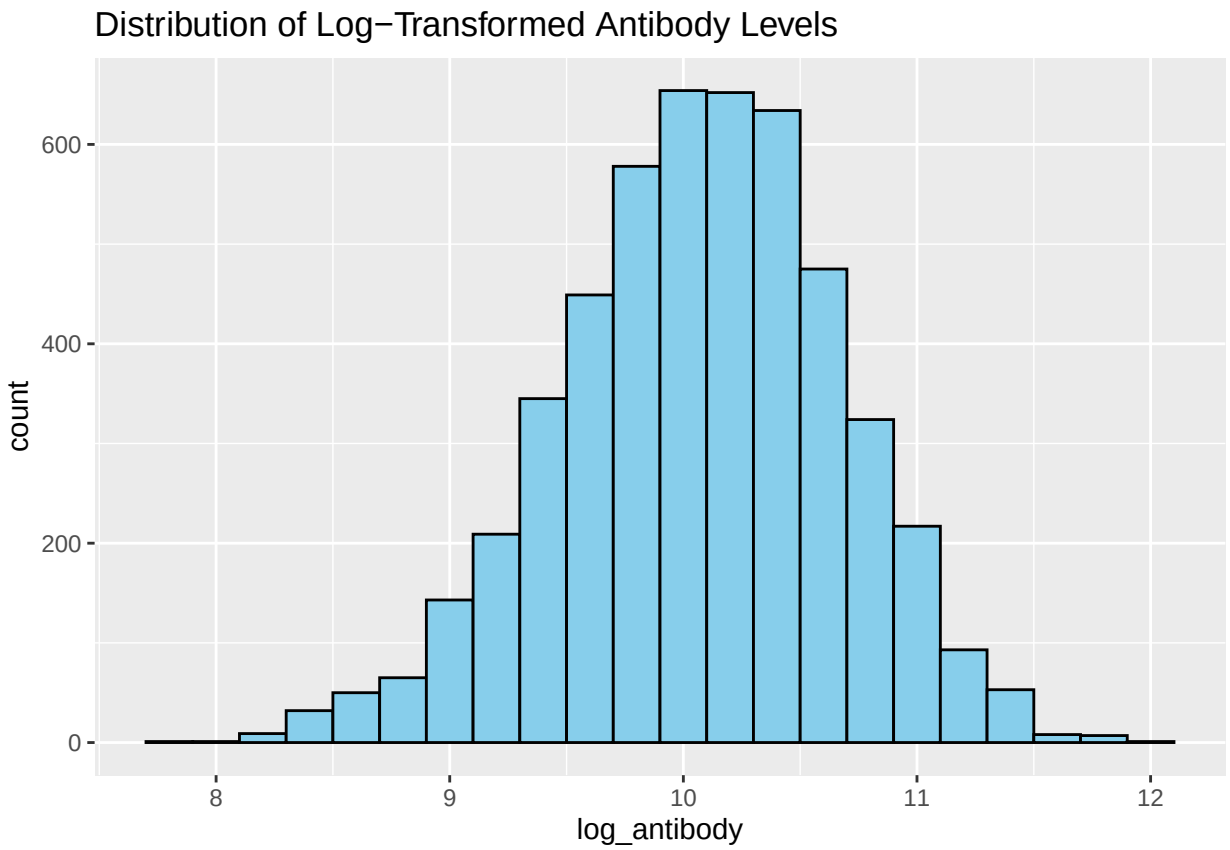
```
age = summarize_cont(dat2, age)
bmi = summarize_cont(dat2, bmi)
height = summarize_cont(dat2, height)
weight = summarize_cont(dat2, weight)
SBP = summarize_cont(dat2, SBP)
LDL = summarize_cont(dat2, LDL)
time = summarize_cont(dat2, time)
log_anti = summarize_cont(dat2, log_antibody)
bind_rows(age, bmi, height, weight, SBP, LDL, time, log_anti) %>%
  knitr::kable()
```

Variable	Median	IQR
age	60.0	[57, 63]
bmi	27.6	[25.8, 29.6]

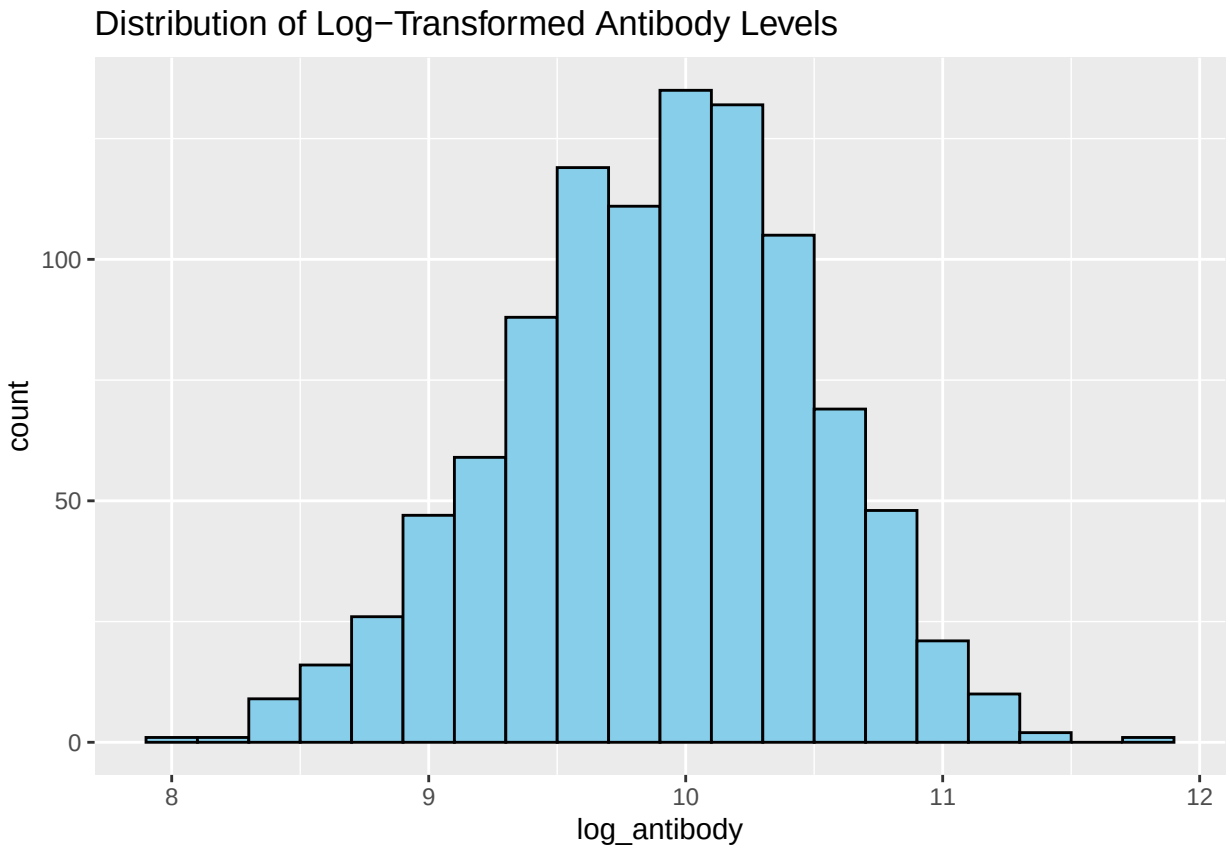
Variable	Median	IQR
height	170.2	[166.1, 174.2]
weight	80.2	[75.3, 84.4]
SBP	130.0	[124, 135]
LDL	112.0	[96, 124]
time	171.0	[140, 205]
log_antibody	9.9	[9.5, 10.3]

Distribution of antibody levels

```
ggplot(dat1, aes(x = log_antibody)) +
  geom_histogram(binwidth = 0.2, fill = 'skyblue', color = 'black') +
  labs(title = "Distribution of Log-Transformed Antibody Levels")
```



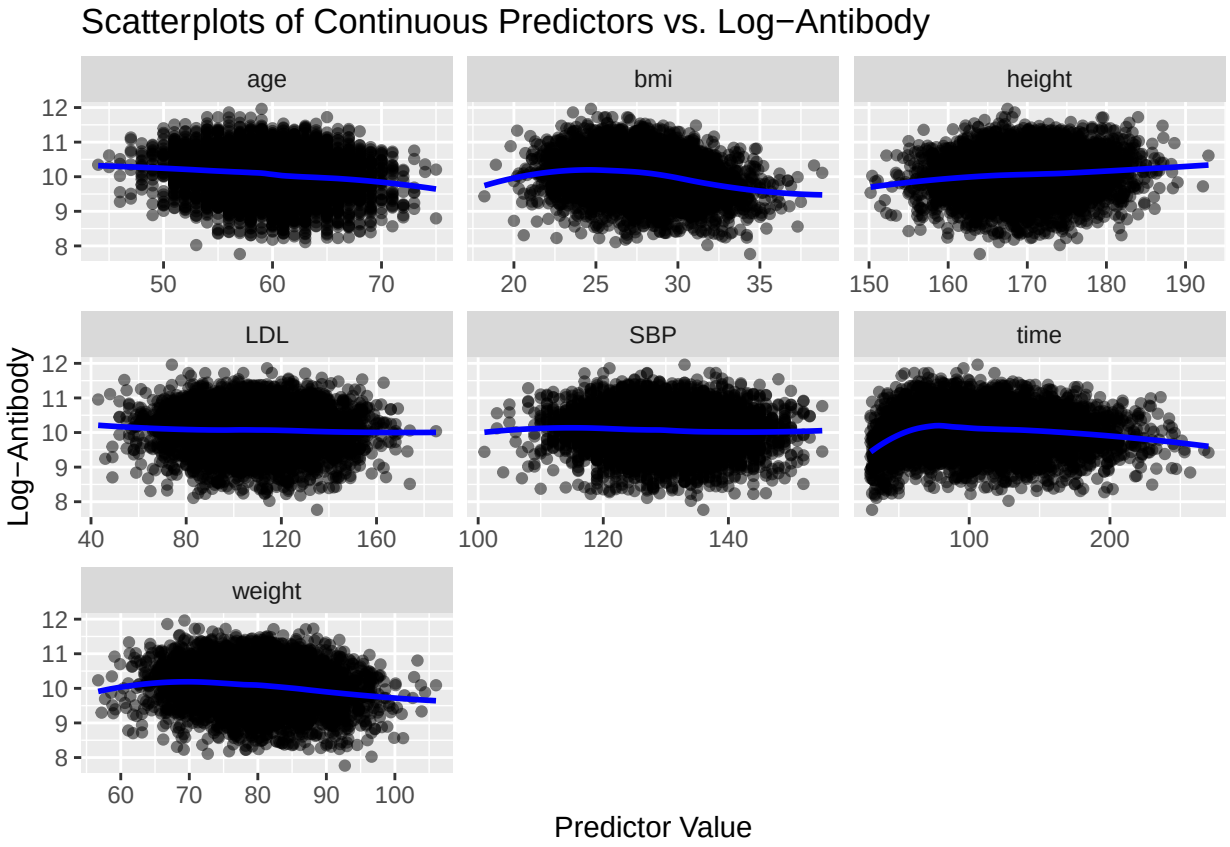

```
ggplot(dat2, aes(x = log_antibody)) +
  geom_histogram(binwidth = 0.2, fill = 'skyblue', color = 'black') +
  labs(title = "Distribution of Log-Transformed Antibody Levels")
```



Scatterplots of Continuous Predictors vs. Log-Antibody

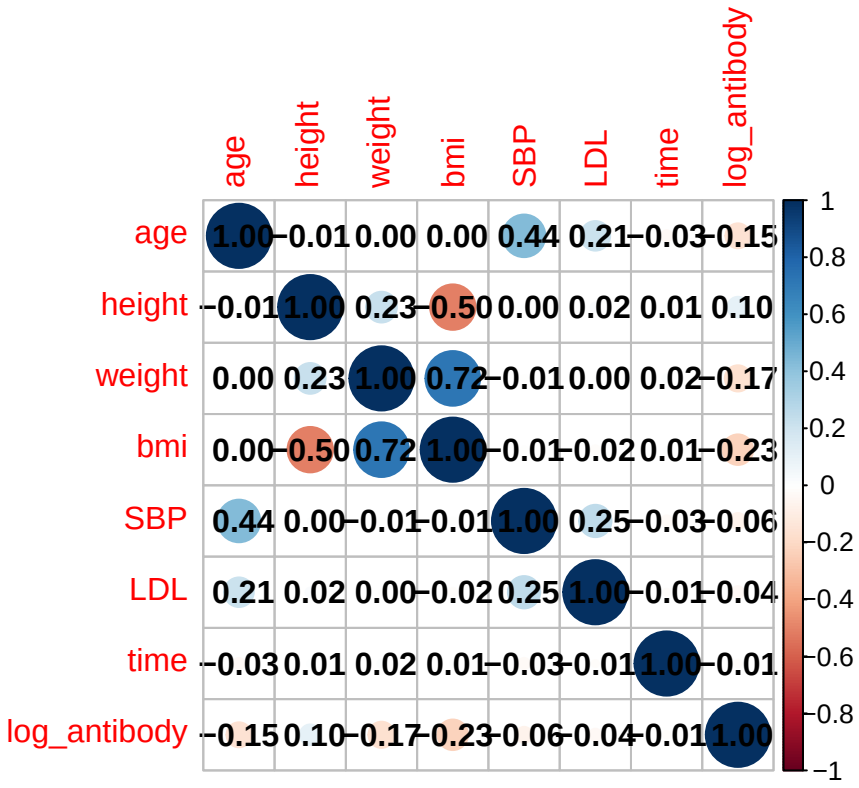
```
dat1 %>%
  select(age, height, weight, bmi, SBP, LDL, time, log_antibody) %>%
  pivot_longer(
    cols = c(age,  height, weight, bmi, SBP, LDL, time),
    names_to = "predictor",
    values_to = "value"
  ) %>%
  ggplot(aes(x = value, y = log_antibody)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "loess", se = FALSE, color = "blue") +
  facet_wrap(~ predictor, scales = "free_x") +
  labs(
    x = "Predictor Value",
    y = "Log-Antibody",
    title = "Scatterplots of Continuous Predictors vs. Log-Antibody"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Correlation Matrix of Continuous Variables

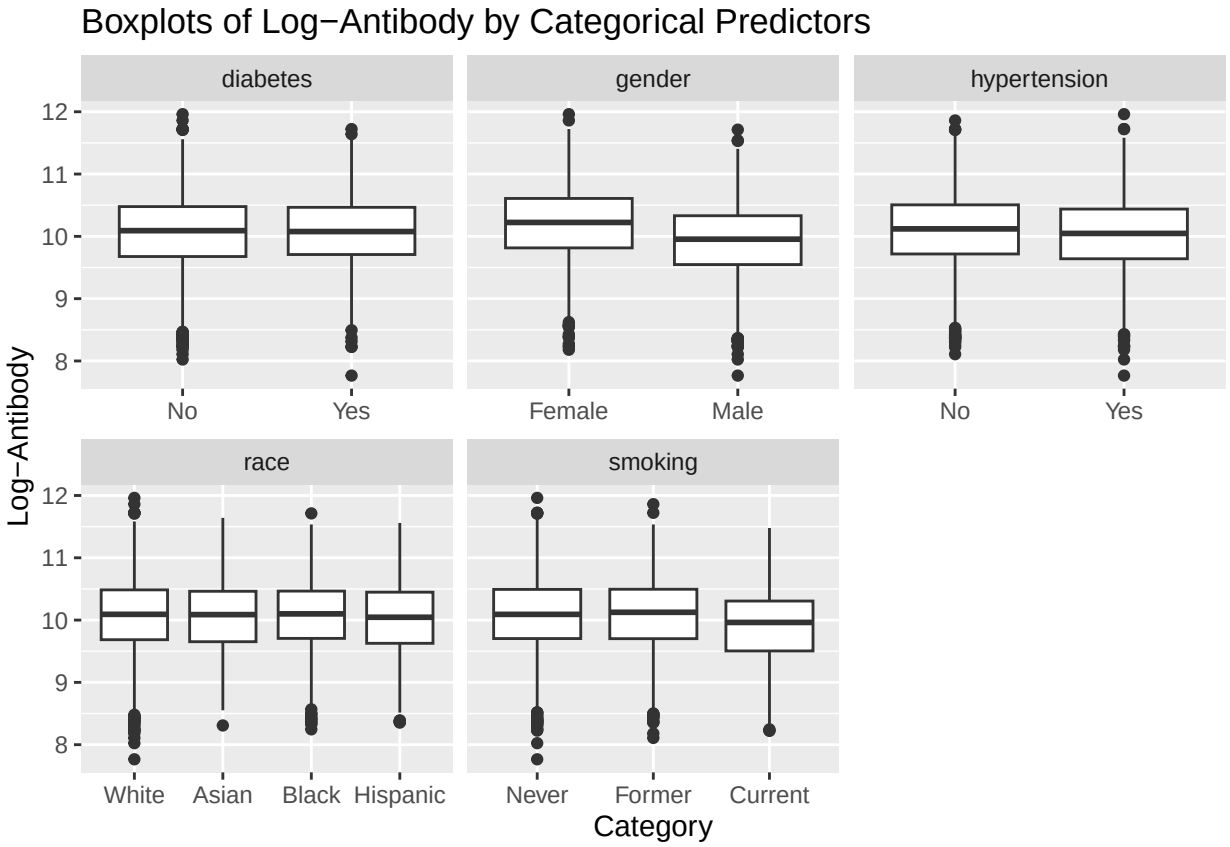
```
dat1 %>%
  select(age, height, weight, bmi, SBP, LDL, time, log_antibody) %>%
  cor(use = "complete.obs") %>%
  corrplot(type = "full",
           title = "Correlation Matrix of Continuous Variables",
           addCoef.col = "black",
           mar = c(0,0,2,0))
```



Boxplots of log_antibody by categorical variables

```
dat1 %>%
  select(log_antibody, gender, race, smoking, diabetes, hypertension) %>%
  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", "Former", "Current")),
    diabetes = factor(diabetes,
                      levels = c(0, 1),
                      labels = c("No", "Yes")),
    hypertension = factor(hypertension,
                          levels = c(0, 1),
                          labels = c("No", "Yes"))
  ) %>%
  pivot_longer(
    cols = c(gender, race, smoking, diabetes, hypertension),
    names_to = "predictor",
```

```
    values_to = "category"
  ) %>%
  ggplot(aes(x = category, y = log_antibody)) +
  geom_boxplot() +
  facet_wrap(~ predictor, scales = "free_x") +
  labs(
    x = "Category",
    y = "Log-Antibody",
    title = "Boxplots of Log-Antibody by Categorical Predictors"
  )
```

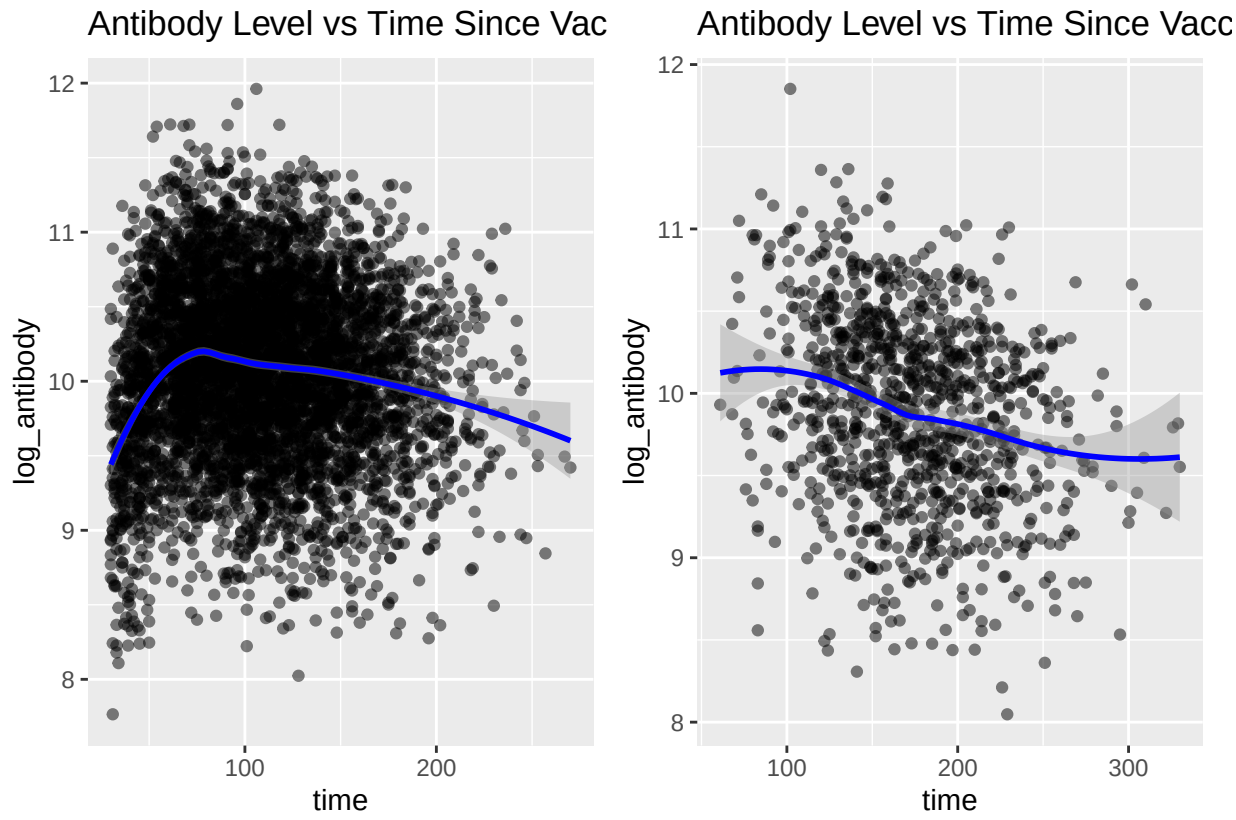


Antibody level and time

```
p1 <- ggplot(dat1, aes(x = time, y = log_antibody)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "loess", color = "blue") +
  labs(title = "Antibody Level vs Time Since Vaccination (dat1)")

p2 <- ggplot(dat2, aes(x = time, y = log_antibody)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "loess", color = "blue") +
  labs(title = "Antibody Level vs Time Since Vaccination (dat2)")
```

```
p1+p2
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



Models Building

cross-validation

```
set.seed(123)
ctrl = trainControl(method = "cv", number = 10)
```

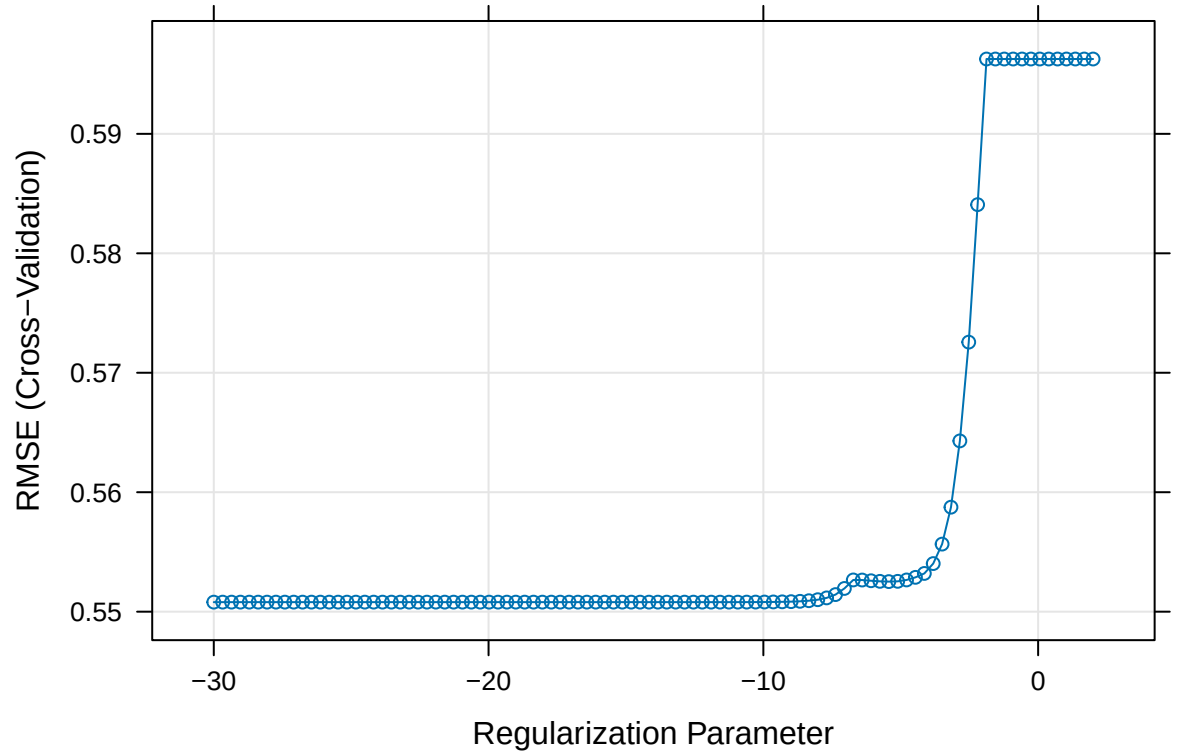
LASSO

```
set.seed(123)
model_lasso <- train(
  log_antibody ~ .,
  data = dat1,
  method = "glmnet",
  trControl = ctrl,
```

```
  tuneGrid = expand.grid(alpha = 1,
                         lambda = exp(seq(2, -30, length = 100))

))
```

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

```
plot(model_lasso, xTrans = log)
```




```
model_lasso$bestTune
```

```
##    alpha       lambda
## 60     1 1.792538e-05
```

GAM

```
set.seed(123)
model_gam <- train(
  log_antibody ~ .,
  data = dat1,
  method = "gam",
```

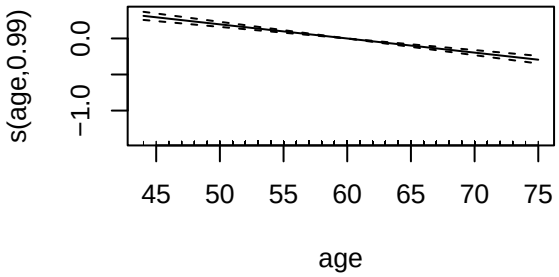
```
  trControl = ctrl
)

model_gam$finalModel
```

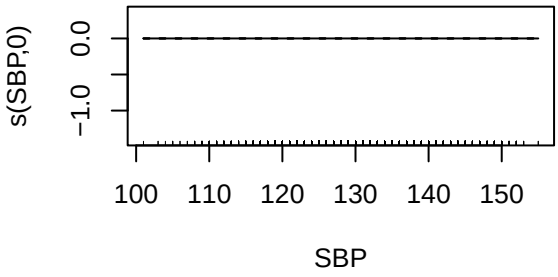
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## .outcome ~ gender + race2 + race3 + race4 + smoking1 + smoking2 +
##     diabetes + hypertension + s(age) + s(SBP) + s(LDL) + s(bmi) +
##     s(time) + s(height) + s(weight)
##
## Estimated degrees of freedom:
## 0.991 0.000 0.000 4.179 7.892 1.234 0.000
##  total = 23.3
##
## GCV score: 0.2786734
```

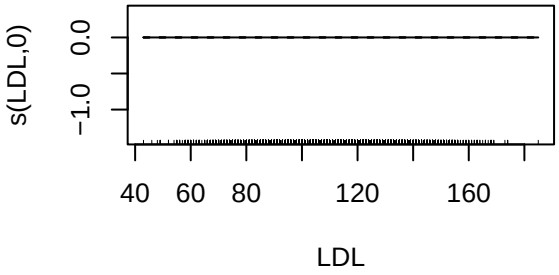
```
model_gam$bestTune
```

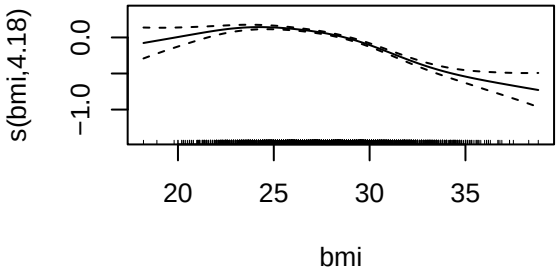
```
##   select method
## 2   TRUE GCV.Cp
```

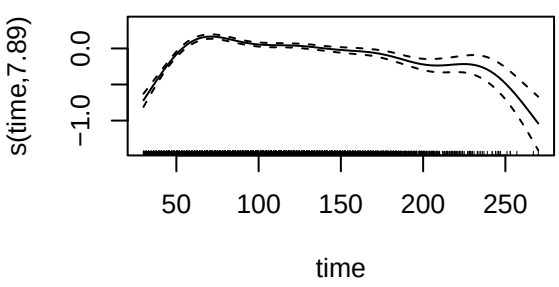
```
par(mfrow = c(2,2))
plot(model_gam$finalModel)
```

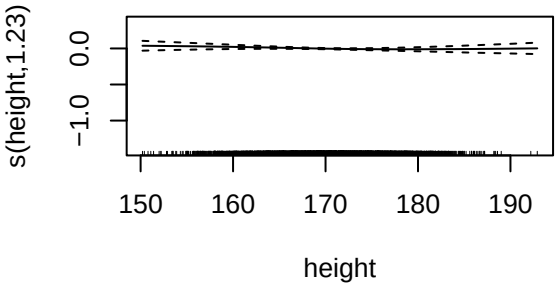

s(age,0.99)
−1.0
0.0
45
50
55
60
65
70
75
age

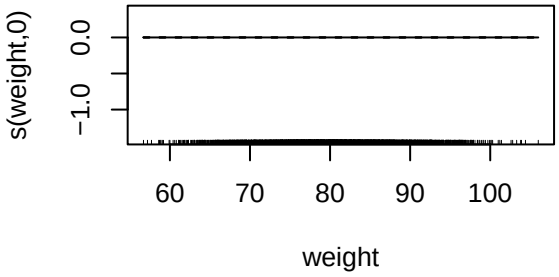

s(SBP,0)
−1.0
0.0
100
110
120
130
140
150
SBP


s(LDL,0)
−1.0
0.0
40
60
80
120
160
LDL


s(bmi,4.18)
−1.0
0.0
20
25
30
35
bmi

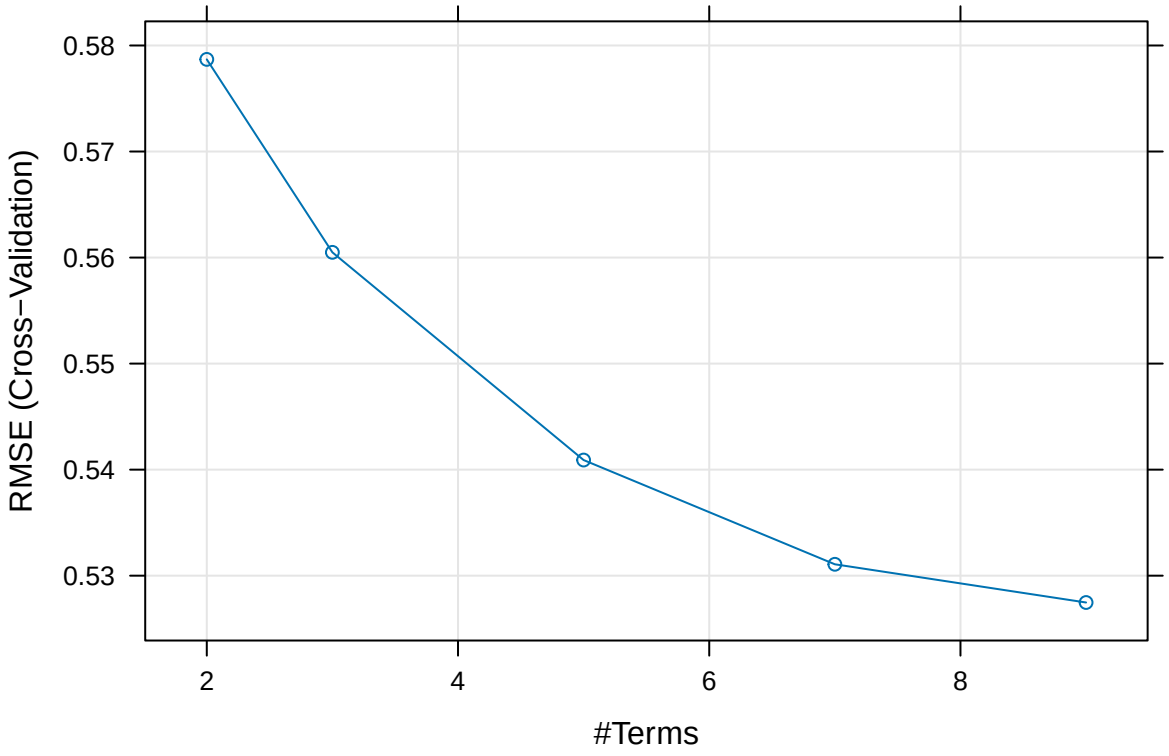






MARS

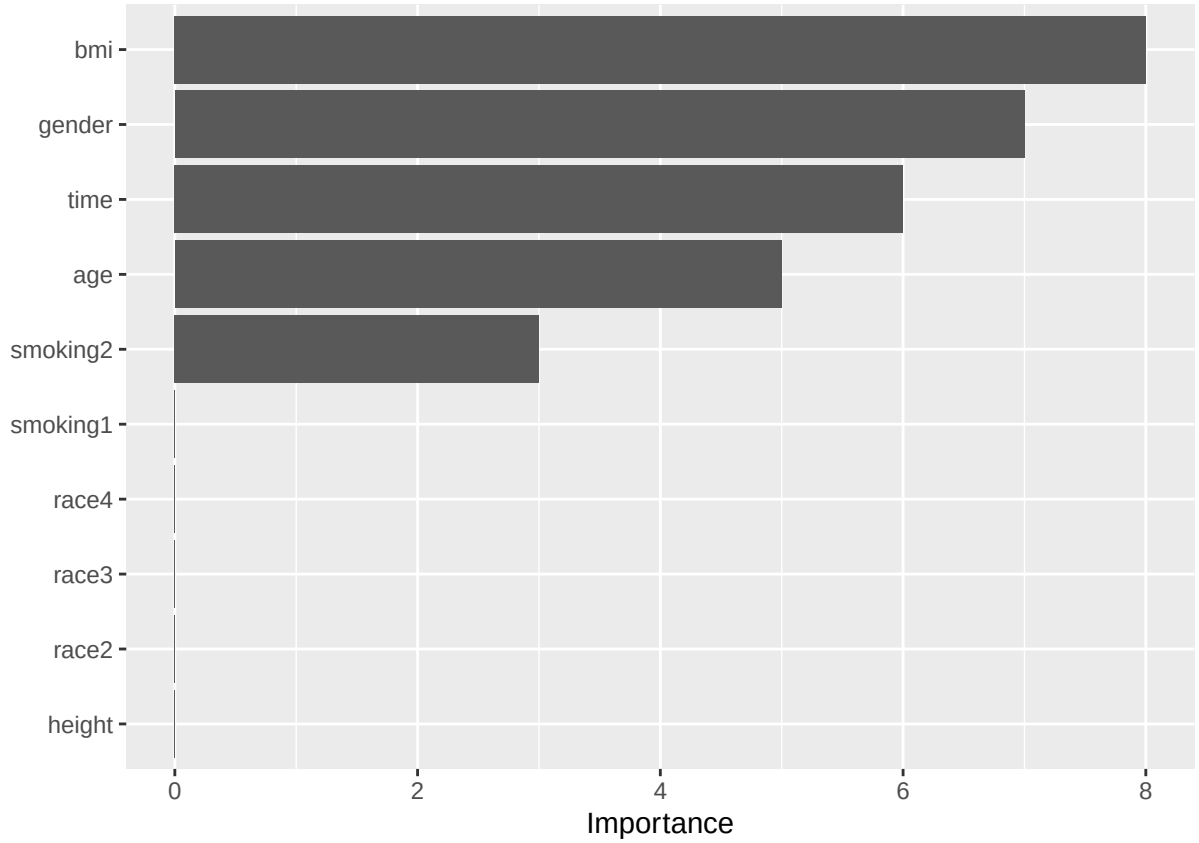
```
set.seed(123)
model_mars <- train(
  log_antibody ~ .,
  data = dat1,
  method = "earth",
  trControl = ctrl,
  tuneLength = 5
)

plot(model_mars)
```



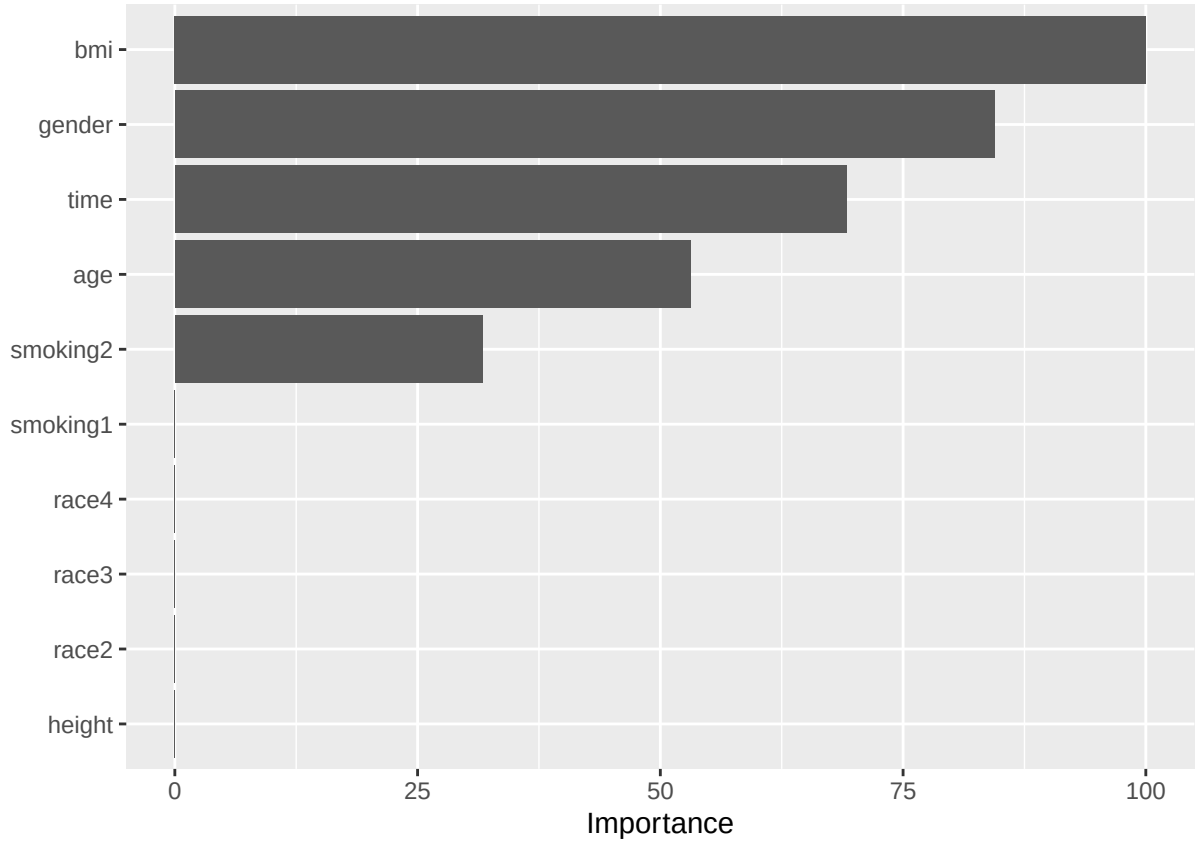

```
coef(model_mars$finalModel)
```

```
##  (Intercept)  h(27.8-bmi)   h(time-57)   h(57-time)       gender    h(age-59)
## 10.847446930 -0.061997354 -0.002254182 -0.033529326 -0.296290451 -0.022957648
##    h(59-age)     smoking2  h(bmi-23.7)
##  0.016138468 -0.205126851 -0.084380175
```

```
vip(model_mars$finalModel, type = "nsubsets")
```




```
vip(model_mars$finalModel, type = "rss")
```



Predictions and Model Evaluation

```
set.seed(123)

pred_lasso <- predict(model_lasso, newdata = dat2)
pred_mars  <- predict(model_mars, newdata = dat2)
pred_gam   <- predict(model_gam, newdata = dat2)

resample = resamples(list(lasso = model_lasso, gam = model_gam, mars = model_mars))
summary(resample)
```

```
##
## Call:
## summary.resamples(object = resample)
##
## Models: lasso, gam, mars
## Number of resamples: 10
##
## MAE
##            Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lasso 0.4262302 0.4320751 0.4340211 0.4390288 0.4427211 0.4724617    0
## gam   0.4078470 0.4112544 0.4168810 0.4225034 0.4281610 0.4635896    0
## mars  0.4078837 0.4098168 0.4163554 0.4220725 0.4285910 0.4641090    0
```

```
##
## RMSE
##            Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lasso 0.5327885 0.5395192 0.5514583 0.5507990 0.5562623 0.5834205    0
## gam   0.5055708 0.5150506 0.5243183 0.5277724 0.5349189 0.5715483    0
## mars  0.5048757 0.5121685 0.5240607 0.5274715 0.5361255 0.5724126    0
##
## Rsquared
##            Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lasso 0.1001182 0.1324550 0.1530212 0.1478575 0.1617642 0.1910705    0
## gam   0.1586929 0.1935757 0.2259290 0.2177415 0.2449893 0.2595686    0
## mars  0.1566346 0.1954579 0.2274358 0.2187657 0.2465799 0.2627660    0
```

MARS model has the lowest RMSE (0.5276)