Data Science II Midterm

Megan Panier, Shiying Wu, and Rita Wang

2025-03-25

Libraries

```
library(readxl) # to import excel files
library(tidyverse)
library(ggplot2)
library(tidymodels)
library(glmnet)
library(caret)
library(splines)
library(mgcv)
library(earth)
library(pROC)
library(pdp)
library(vip)
library(AppliedPredictiveModeling)
```

Importing and Organizing Data

```
load("./data/dat1.RData") #importing training data
  # Log-transformed antibody level (log_antibody) --> y
initial_training = dat1 #renaming the original training data name

load("./data/dat2.RData") #importing training data
initial_test = dat2 #renaming the original training data name

set.seed(2222)

# partition data into training and validation data sets
datSplit = initial_split(data = initial_training, prop = 0.8)
training = training(datSplit)
validation = testing(datSplit)
```

```
model = lm(log_antibody ~ age + gender + race + smoking + height + weight + bmi + diabetes +
             hypertension + SBP + LDL + time, data = training)

# View the model summary
summary(model)
```

```r
predictions_train = predict(model, newdata = validation)

# RMSE
rmse_train = sqrt(mean((predictions_train - validation$log_antibody)^2))
rmse_train

# R^2
rsq_train = 1 - sum((predictions_train - validation$log_antibody)^2) /
  sum((mean(training$log_antibody) - validation$log_antibody)^2)
rsq_train
```

```r
generalization = predict(model, newdata = initial_test)

# Calculate RMSE for dat2
rmse_dat2 = sqrt(mean((generalization - initial_test$log_antibody)^2))
rmse_dat2

# Calculate R-squared for dat2
rsq_dat2 = 1 - sum((generalization - initial_test$log_antibody)^2) /
  sum((mean(initial_test$log_antibody) - initial_test$log_antibody)^2)
rsq_dat2
```

Understand how demographic and clinical factors influence antibody responses

Understand how antibody levels change over time following vaccination