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)
```

Linear Regression

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

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

```
## 
## Call:
## lm(formula = log_antibody ~ age + gender + race + smoking + height + 
##     weight + bmi + diabetes + hypertension + SBP + LDL + time, 
##     data = training)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.14743 -0.35065  0.03211  0.37738  1.53018 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  28.2069787  2.6457948  10.661  < 2e-16 ***
## age          -0.0196829  0.0021607  -9.110  < 2e-16 ***
## gender       -0.2797813  0.0173438 -16.132  < 2e-16 ***
## race2        -0.0139482  0.0386090  -0.361   0.7179    
## race3        -0.0080486  0.0218346  -0.369   0.7124    
## race4        -0.0463573  0.0301577  -1.537   0.1243    
## smoking1      0.0219875  0.0193608   1.136   0.2562    
## smoking2     -0.1815792  0.0297480  -6.104 1.13e-09 ***
## height       -0.0919586  0.0154999  -5.933 3.23e-09 ***
## weight        0.0953372  0.0164227   5.805 6.93e-09 ***
## bmi          -0.3264716  0.0471923  -6.918 5.32e-12 ***
## diabetes      0.0030653  0.0243426   0.126   0.8998    
## hypertension -0.0287531  0.0290736  -0.989   0.3227    
## SBP           0.0024700  0.0019002   1.300   0.1937    
## LDL          -0.0001017  0.0004518  -0.225   0.8219    
## time         -0.0003804  0.0001988  -1.914   0.0557 .  
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 0.5471 on 3984 degrees of freedom
## Multiple R-squared:  0.147,  Adjusted R-squared:  0.1438 
## F-statistic: 45.78 on 15 and 3984 DF,  p-value: < 2.2e-16
```

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

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

```
## [1] 0.5639064
```

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

```
## [1] 0.1641537
```

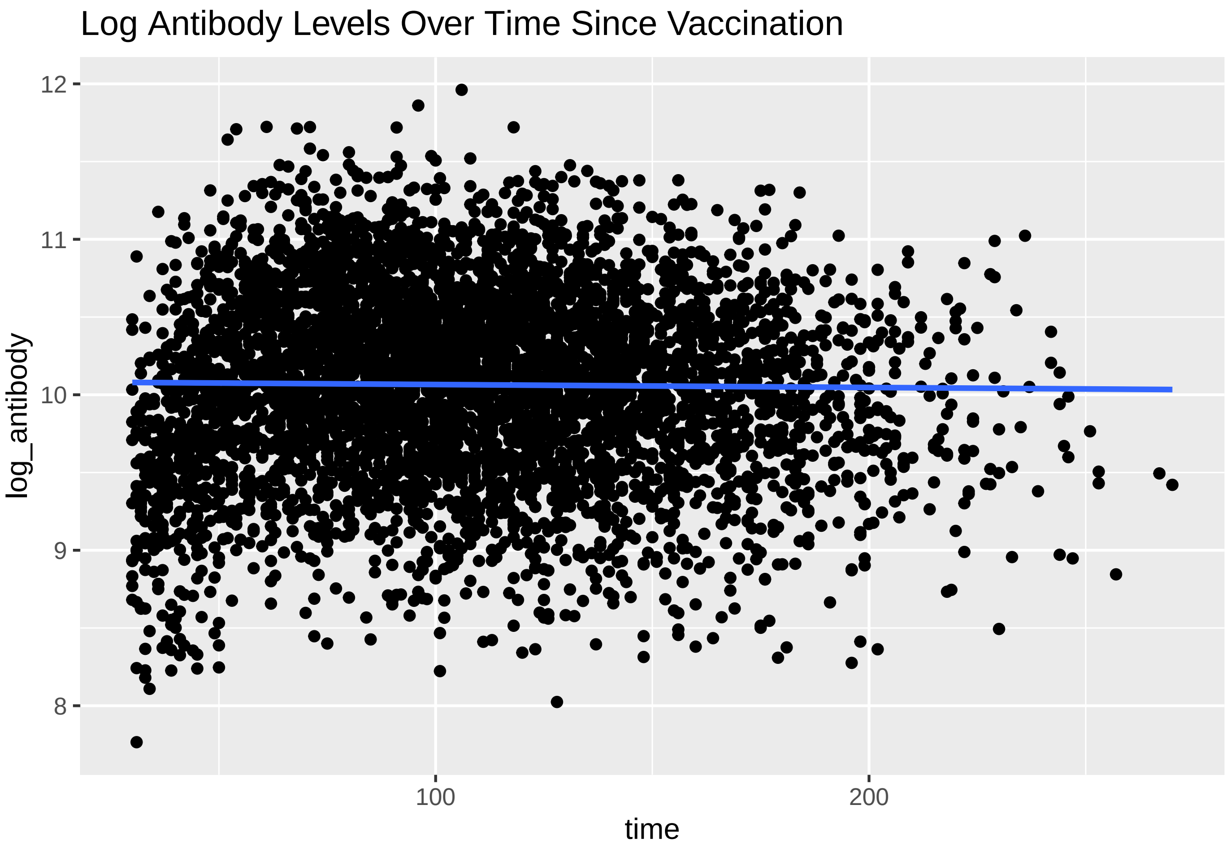
```
generalization = predict(model, newdata = initial_test)

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

```
## [1] 0.5662817
```

```
# 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
```

```
## [1] 0.06952672
```

```
ggplot(initial_training, aes(x = time, y = log_antibody)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Log Antibody Levels Over Time Since Vaccination")
```


Log Antibody Levels Over Time Since Vaccination
log_antibody
time