# P8106 Midterm Final Report

Ila Kanneboyina, Shayne Estill, Naomi Simon-Kumar (ns3782)

03/24/2025

## Contents



# 1. Introduction

## 1.1. Background and Study Objective

This study aims to build a prediction model for antibody responses to a newly authorized vaccine, as measured by log-transformed antibody levels from dried blood spot samples. Our objective is to develop an accurate model to improve understanding of vaccine effectiveness across different population segments and to support the monitoring of immune protection over time.

### 1.2. Data Source and Description

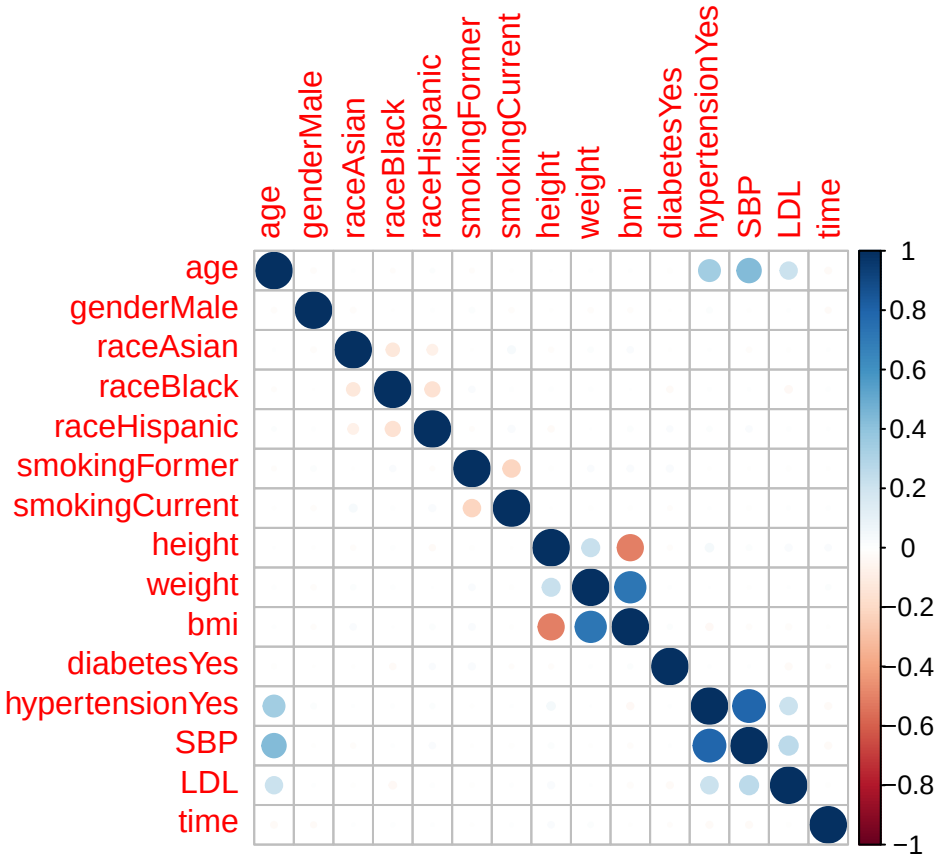
The primary dataset (**dat1.RData**) contains demographic and clinical information from participants in a vaccine response study. Variables include age, gender, race/ethnicity, body mass index (BMI), blood pressure, cholesterol levels, diabetes, hypertension, and time since vaccination (in days). The outcome of interest is the log-transformed antibody level, measured using dried blood spot samples. The dataset consists of records from 5000 participants with 13 variables, excluding the unique identifier variable. A second independent dataset (**dat2.RData**) with identical structure was collected several months later to assess model generalizability. The second dataset consists of records from 1000 participants.

## 2. Exploratory Analysis

### 2.1. Dataset Overview

### 2.2. Summary Statistics

### 2.3. Correlation Analysis



### 2.4. Exploratory Plots

# 3. Model Training

## 3.1. Data Preparation

## 3.2. Cross-validation approach

For model training and evaluation, we implemented 10-fold cross-validation on the training dataset (80% of the original data) using the trainControl function This partitioned the training data into 10 equal subsets, where each model was trained on 9 folds and validated on the remaining fold, rotating through all folds.

## 3.3. Models

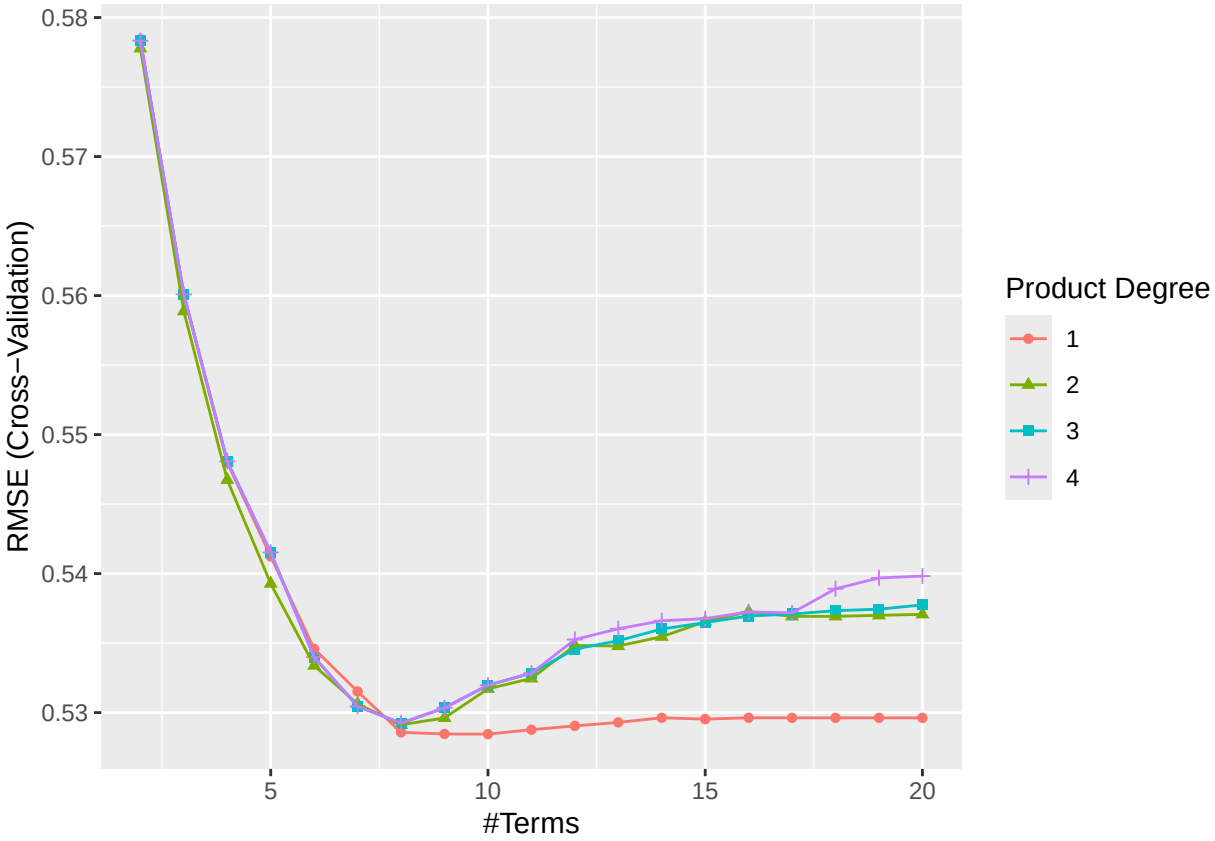
#### 3.3.1. Linear Regression

#### 3.3.2. Elastic Net

#### 3.3.3. MARS

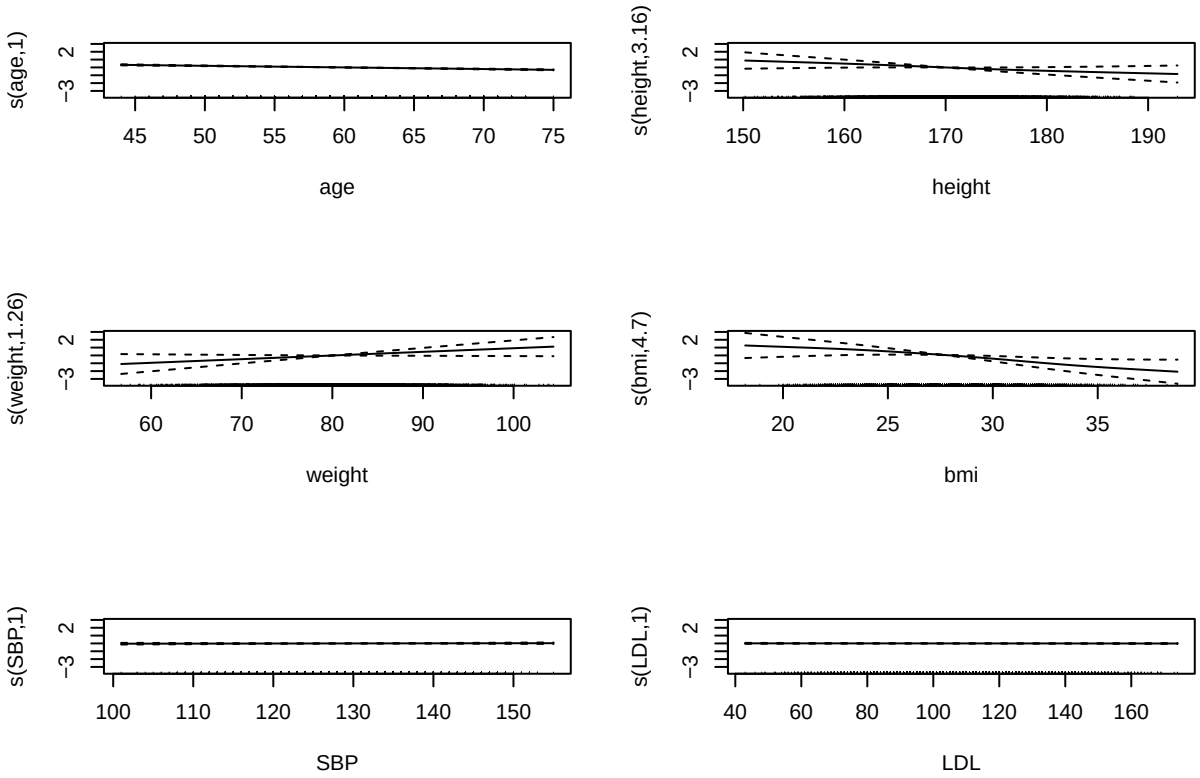
MARS was selected as a candidate model as its ability to automatically detect important predictors and their interactions made it suitable for our dataset's combination of demographic and clinical variables.

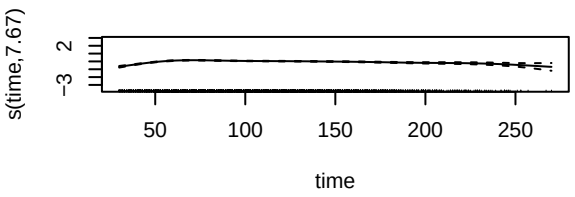
For model tuning, we initially specified a grid with degrees 1 to 3 (maximum number of interactions) and 2 to 15 retained terms, then expanded to degrees 1 to 4 and 2 to 20 retained terms to evaluate whether higher-order interactions might better capture complex relationships between variables. Cross-validation results showed that models with degree = 1 consistently achieved the lowest RMSE. As interaction degree increased, RMSE values became more variable rather than decreasing steadily, suggesting no gain in prediction accuracy. The range of retained terms appeared appropriate as RMSE decreased initially with more terms but then plateaued. The final MARS model (**degree = 1, 10 retained terms**) was able to represent nonlinear relationships through hinge functions at specific threshold values.




### 3.3.4. GAM

GAM was selected because immune responses typically follow smooth, nonlinear patterns that change gradually with predictors like age, BMI, and time since vaccination. GAM's ability to model these relationships as flexible smooth functions while also being interpretable made it an appropriate candidate model to examine how each predictor independently contributes to antibody levels. Our analysis showed that some predictors have estimated degrees of freedom (edf) greater than 1, specifically height (edf=3.163), weight (edf=1.255), BMI (edf=4.700), and time (edf=7.665), indicating they are modeled as nonlinear smooth functions. This is supported by the GAM predictor plots. Weight, while slightly above 1, was not statistically significant. Among the nonlinear terms, height, BMI, and time were statistically significant ($p < 0.05$), indicating meaningful nonlinear relationships with the outcome. Predictors age, SBP, and LDL each have an edf of 1.00, indicating that they are modeled as approximately linear. Of these, only age was statistically significant ($p < 2e\text{-}16$), suggesting a strong linear association with log antibody levels.


s(age,1)
−3 2
45 50 55 60 65 70 75
age
s(height,3.16)
−3 2
150 160 170 180 190
height
s(weight,1.26)
−3 2
60 70 80 90 100
weight
s(bmi,4.7)
−3 2
20 25 30 35
bmi
s(SBP,1)
−3 2
100 110 120 130 140 150
SBP
s(LDL,1)
−3 2
40 60 80 100 120 140 160
LDL




# 4. Results

## 4.1. Model Comparison and Selection

We evaluated six prediction models for antibody levels: multiple linear regression, elastic net, principal component regression (PCR), partial least squares (PLS), multivariate adaptive regression splines (MARS), and generalized additive model (GAM). To objectively compare model performance, we used 10-fold cross-validation with RMSE as the evaluation metric. The cross-validation results showed that the MARS model achieved the lowest mean RMSE (0.5285), followed closely by GAM (0.5304) and elastic net (0.5317). The linear model performed notably worse (RMSE = 0.5409). This supported our initial analysis from EDA that linear relationships may not adequately capture the complex relationships in the data. PCR and PLS models performed moderately well, with RMSE values of 0.5367 and 0.5323, respectively.

The MARS model, with optimal parameters of degree = 1 (no interaction terms) and nprune = 10 (retaining 10 terms), was selected as the final prediction model of antibody levels due to its optimal predictive performance, effectively capturing non-linear relationships between predictors and antibody levels.

## 4.2. Final Model: xxx

### 4.2.1. Evaluation of model performance

Test set performance metrics & interpretation

e.g.

When applied to the held-out test set from the original dataset, the MARS model demonstrated consistent performance with a test RMSE of 0.5270, indicating good generalizability to unseen data.

### 4.3. Model Generalizability

Evaluation of model based on dat2.

## 5. Conclusion

## 6. References

Phillips, N. D. (n.d.). YaRrr! The Pirate's Guide to R. Retrieved March 29, 2025, from https://bookdown.org/ndphillips/YaRrr/arranging-plots-with-parmfrow-and-layout.html