# P8106 Midterm - Report

Kate Colvin (KAC2301), Jeong Yun (Lizy) Choi (JC6452), and Flora Pang (FP2513)

## Introduction

In this project, our team explored the dataset collected from a study on evaluating antibody responses to a newly authorized vaccine. The primary outcome of interest is the log-transformed antibody level measured via dried blood spots. The dataset includes a range of demographic and clinical predictors such as age, gender, race/ethnicity, smoking status, BMI, chronic conditions, and time since vaccination.

Our goal is to develop a predictive model that characterizes how these factors influence antibody responses and asses how well this model generalizes to a new independent dataset collected at a later time point. By doing so, we hope to identify key predictors of antibody levels and evaluate the generalizability of our model across different dataset.

## Exploratory Analysis

Our full combined dataset includes 6,000 patients and contains demographic and health information, time since the patient received the vaccine, and log-transformed antibody level. There are two different subsets of data: data on 5,000 patients was initially collected for model training, and data on 1,000 additional patients was independently collected several months later for model testing and evaluation.

Patients in both datasets have similar demographic and health characteristics (Table 1), but patients from the second collected dataset have a greater time since receiving the vaccine (an additional few months), and therefore slightly lower observed log-transformed antibody levels (Figures 1 and 2). Because we are more likely to observe lower antibody levels from patients in the testing dataset, it's possible that this difference may impact the prediction performance of our models, which are trained using the initial dataset. After plotting the most correlated quantitative variables versus log-transformed antibody level, we can see that the fitted line for the testing data is always flatter than the line of the training data, indicating a weaker relationship between these variables and the response (Figures 7-9).

Across gender and smoking status, there were very slight differences in the observed antibody levels. Women had slightly greater antibody responses than men overall (Figure 3 and Table 2), while current smokers had slightly lower antibody responses than former and never-smokers (Figure 5 and Table 4). There were no observed differences in antibody responses across race (Figure 4 and Table 3). The quantitative variables that were most correlated with log-transformed antibody level were BMI, weight, and age. It's important to note that several predictors are also correlated with each other, such as BMI and weight, BMI and height, and SBP and age (Figure 6), which will impact variable selection.

## Model Training

In this analysis, we trained three different models: Multiple Linear Regression (MLR), LASSO Regression, and Multivariate Adaptive Regression Splines (MARS). The following sections provides each step in the model training process, from pre-processing to final model selection.

### Data Pre-processing

- We ensured that there were no missing values in the training data, missing data were imputed or removed.
- Continuous variables were untouched, while categorical variables were converted to factor types (such as race, gender, smoking).
- The response variable, log antibody, was log-transformed to normalize its distribution and reduce skewness.

### Multiple Linear Regression (MLR) Model

We started by fitting a MLR model with all available predictors in the dataset and the model was fit using ordinary least squares regression (OLS).

The model was trained using the lm() function and the training process involved fitting the model to the data, estimating the regression coefficients for each predictor, and computing the residuals. The code below was used:

```
mlr_model <- lm(log_antibody ~ ., data = dat1)
```

The coefficients were estimated through OLS regression, and the residuals were checked for normality. The model was trained on the entire training dataset, and no regularization was applied.(Table 5)

### LASSO Model

To address potential multi-collinearity and perform feature selection, we used LASSO Regression, applying L1 regularization to shrink the coefficients of less important features to zero. The LASSO model was trained using the glmnet package.

Since LASSO is sensitive to differences in scale among predictor variables, numerical predictors were standardized before training to ensure fair comparison across variables. The preProcess() function from the caret package was used. The same transformations were applied to the validation and test datasets to maintain consistency.

The training procedure involved:

1. Creating a matrix of predictor variables (x) and a vector of the response variable (y)
2. Using cross-validation to select the best lambda (regularization parameter) based on the model's performance. The model with the lowest cross-validation error was used for evaluation.

```
lasso_model <- cv.glmnet(x_train, y_train, alpha = 1)
best_lambda <- lasso_model$lambda.min
lasso_final <- glmnet(x_train, y_train, alpha = 1, lambda = best_lambda)
```

### Multivariate Adaptive Regression Splines (MARS) Model

Non-linear regression MARS model automatically selects the best interactions and non-linear transformations of predictors. We first trained the MARS model without tuning, the used cross-validation to finetune it and determine the optimal number of terms and the degree of interactions (Table 6).

```
mars_model <- earth(log_antibody ~ ., data = dat1)
mars_tune <- train(log_antibody ~ ., data = dat1, method = "earth",
                   trControl = train_control, tuneGrid = tune_grid)
mars_model_tune <- train(log_antibody ~ .,
                         data = dat1,
                         method = "earth",
                         trControl = train_control,
                         tuneGrid = data.frame(nprune = 10, degree = 1))
```

The best parameters were selected as follows:

- nprune = 10: The final model had 10 terms (lowest Generalized Cross Validation score)
- degree = 1: The degree of interaction was set to 1, which considers only pairwise interactions between features.

## Results

To evaluate predictors of log antibody response, we trained three models using the original dataset (dat1): Multiple Linear Regression (MLR), LASSO Regression, and Multivariate Adaptive Regression Splines (MARS). Each model was then tested on an independent dataset (dat2) to assess generalizability. Model performance was evaluated using Root Mean Squared Error (RMSE), which captures prediction accuracy, and Adjusted R-squared (Adj. $R^2$), which measures the proportion of variability explained while accounting for model complexity.

Among the three models, MARS model (Table 6) showed the best overall performance, with the lowest RMSE (0.533) and the highest adjusted $R^2$ (0.169). Both MLR and LASSO had identical RMSE values (0.568), but MLR explained more variance (Adj. $R^2 = 0.149$) than LASSO (Adj. $R^2 = 0.048$), suggesting that LASSO may have underfit slightly by shrinking less informative predictors. These results are summarized in Table 7.

MARS was particularly well-suited for this dataset because of its ability to flexibly model non-linear relationships and interactions, which are common in immune response data. Variables like age, BMI, and time since vaccination are unlikely to relate to antibody levels in a strictly linear way (Figure 6). While linear models assume constant change across the predictor range, MARS allows for thresholds and curve shapes that better reflect biological processes. This likely contributed to its stronger performance on the test set.

We also assessed model stability using 10-fold cross-validation through the caret package and compared RMSE distributions across models (Figure 10). MARS had the lowest median RMSE, reinforcing its predictive strength, though its variability across folds was slightly wider than MLR. This reflects a common trade-off: MARS offers more flexibility but may be slightly more variable; linear models are simpler and more stable, but risk missing important patterns.

In summary, MARS was selected as the final model due to its ability to handle complex, non-linear patterns in the data and its strong performance on an independent dataset. While it exhibited slightly more variability in resampling, it offered the best balance of accuracy and flexibility. These findings highlight the importance of model selection based on data structure and analysis goals — and support the use of flexible modeling strategies in understanding antibody responses to vaccination.

Table 1: Summary of Patient Testing and Training Data (N=6000)

| **Characteristic** | **Overall** N = 6,000[1] | **Testing Data** N = 1,000[1] | **Training Data** N = 5,000[1] | **p-value**[2] |
|---|---|---|---|---|
| Age | 60.0 (57.0, 63.0) | 60.0 (57.0, 63.0) | 60.0 (57.0, 63.0) | 0.9 |
| Gender | | | | 0.7 |
| Female | 3,082 (51%) | 509 (51%) | 2,573 (51%) | |
| Male | 2,918 (49%) | 491 (49%) | 2,427 (49%) | |
| Race | | | | 0.6 |
| Asian | 333 (5.6%) | 55 (5.5%) | 278 (5.6%) | |
| Black | 1,235 (21%) | 199 (20%) | 1,036 (21%) | |
| Hispanic | 548 (9.1%) | 83 (8.3%) | 465 (9.3%) | |
| White | 3,884 (65%) | 663 (66%) | 3,221 (64%) | |
| Smoking | | | | 0.8 |
| Current | 589 (9.8%) | 103 (10%) | 486 (9.7%) | |
| Former | 1,800 (30%) | 296 (30%) | 1,504 (30%) | |
| Never | 3,611 (60%) | 601 (60%) | 3,010 (60%) | |
| Height (cm) | 170.1 (166.1, 174.2) | 170.2 (166.1, 174.2) | 170.1 (166.1, 174.3) | 0.7 |
| Weight (kg) | 80 (75, 85) | 80 (75, 84) | 80 (75, 85) | 0.8 |
| BMI | 27.60 (25.80, 29.50) | 27.60 (25.80, 29.60) | 27.60 (25.80, 29.50) | 0.9 |
| Diabetes | 929 (15%) | 157 (16%) | 772 (15%) | 0.8 |
| Hypertension | 2,754 (46%) | 456 (46%) | 2,298 (46%) | 0.8 |
| Systolic Blood Pressure (mmHg) | 130 (124, 135) | 130 (124, 135) | 130 (124, 135) | 0.3 |
| LDL Cholesterol (mg/dL) | 110 (96, 124) | 112 (96, 124) | 110 (96, 124) | 0.4 |
| Time Since Vaccinated (days) | 116 (82, 152) | 171 (140, 205) | 106 (76, 138) | <0.001 |
| Log-Transformed Antibody Level | 10.06 (9.65, 10.45) | 9.93 (9.50, 10.32) | 10.09 (9.68, 10.48) | <0.001 |

[1] Median (Q1, Q3); n (%)

[2] Wilcoxon rank sum test; Pearson's Chi-squared test

Figure 1: Distribution of Log−Transformed Antibody Level, by Data Set

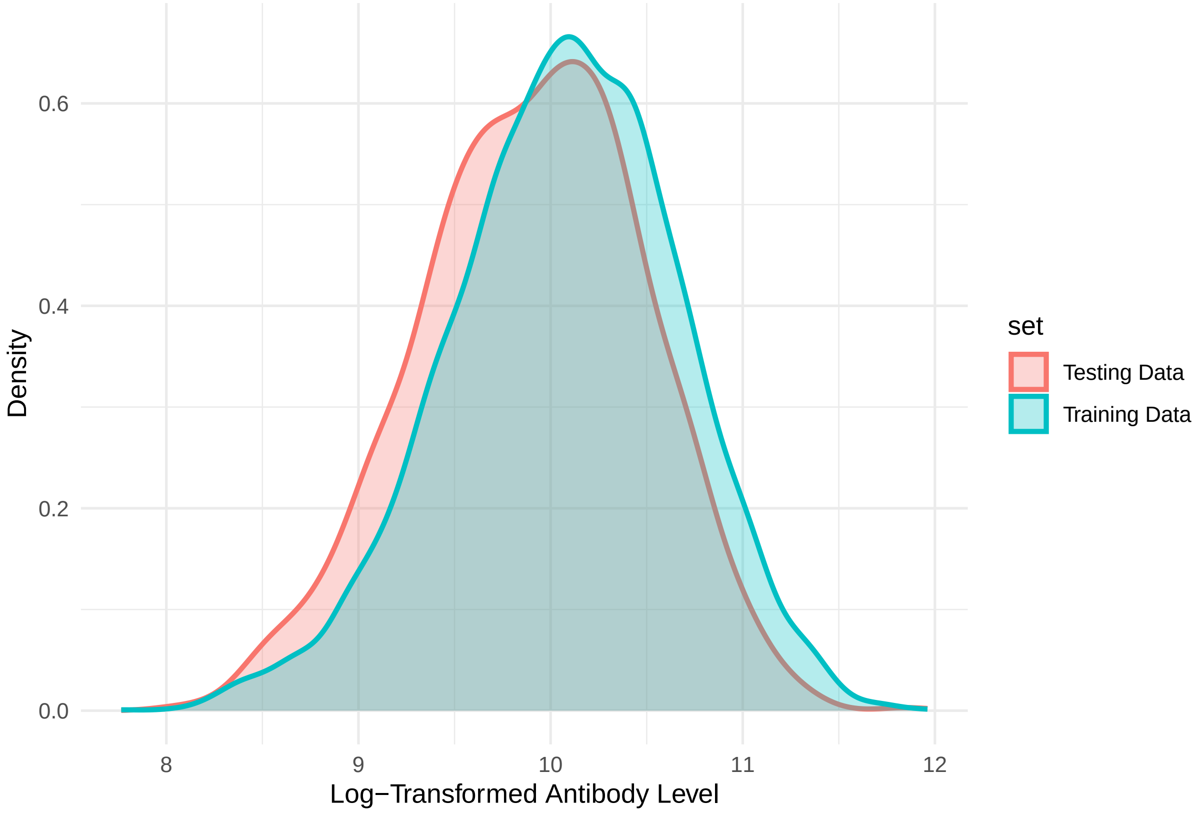


Figure 2: Distribution of Days Since Vaccination, by Data Set

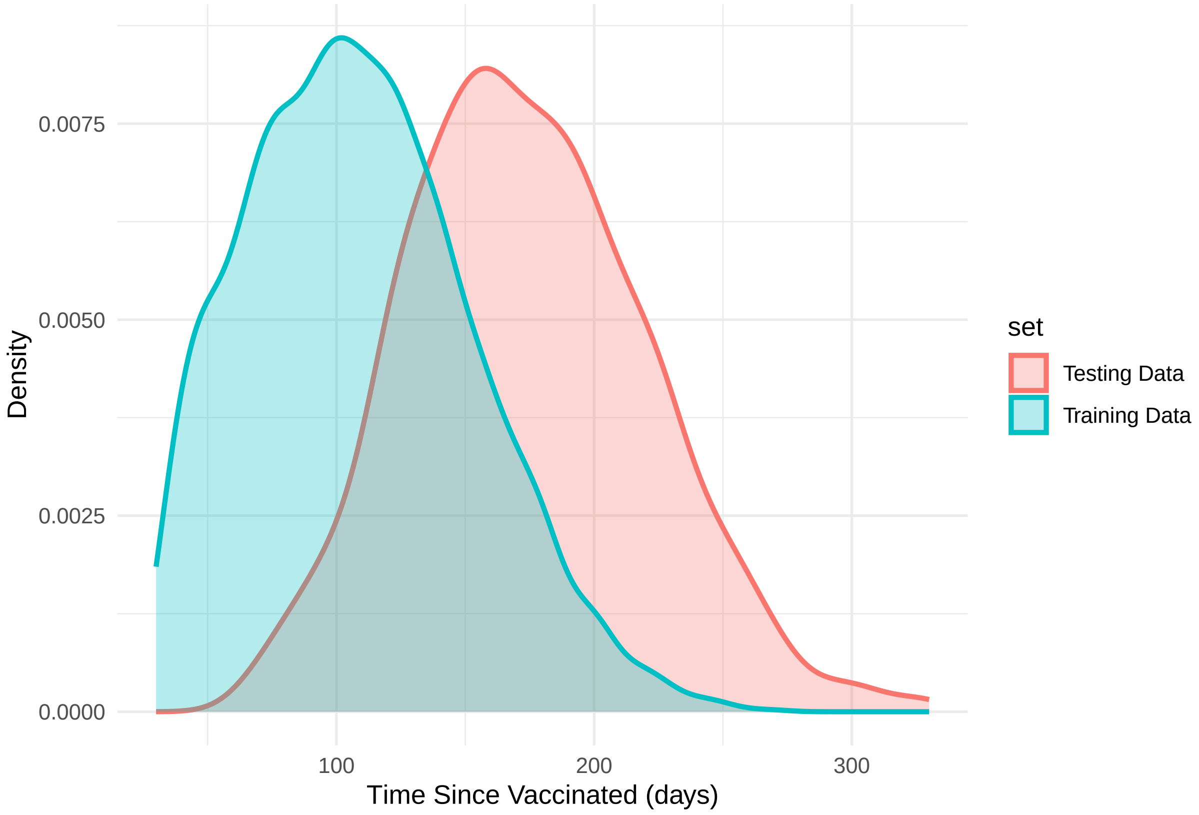


Figure 3: Distribution of Log−Transformed Antibody Level, by Gender

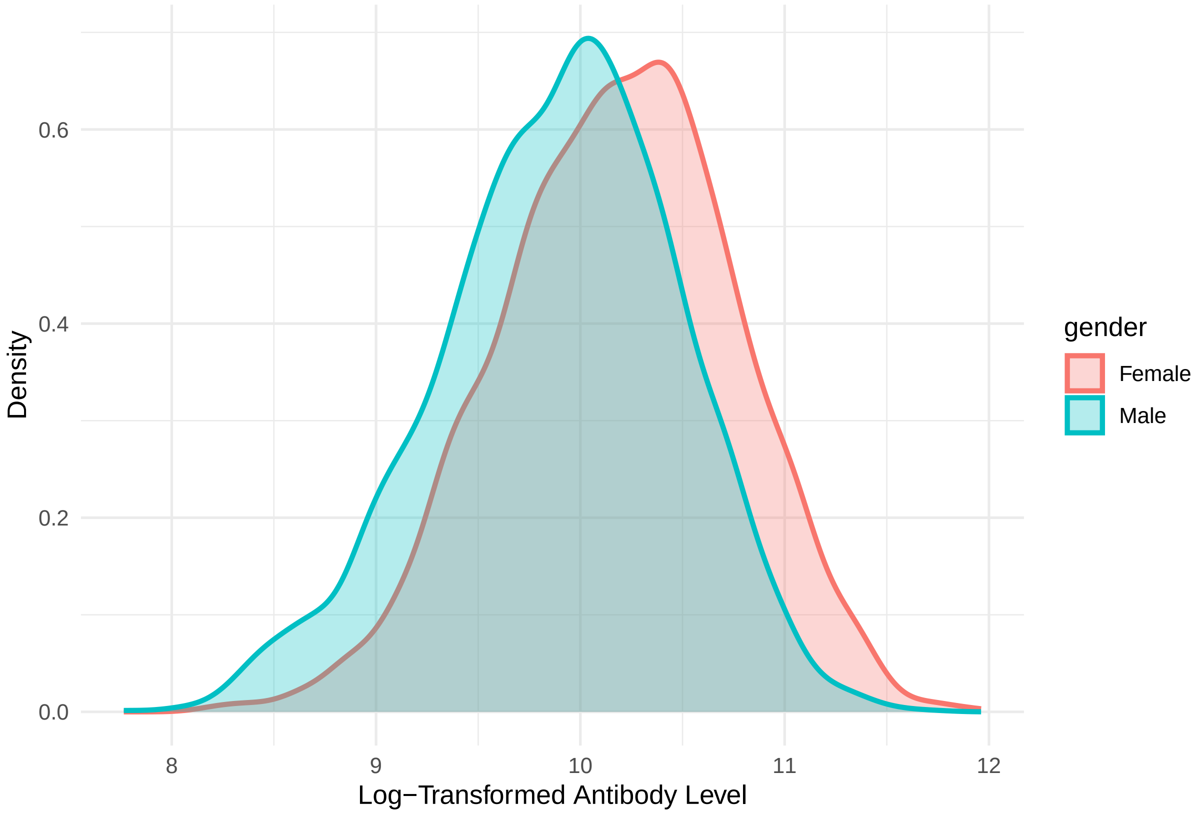


Figure 4: Distribution of Log−Transformed Antibody Level, by Race

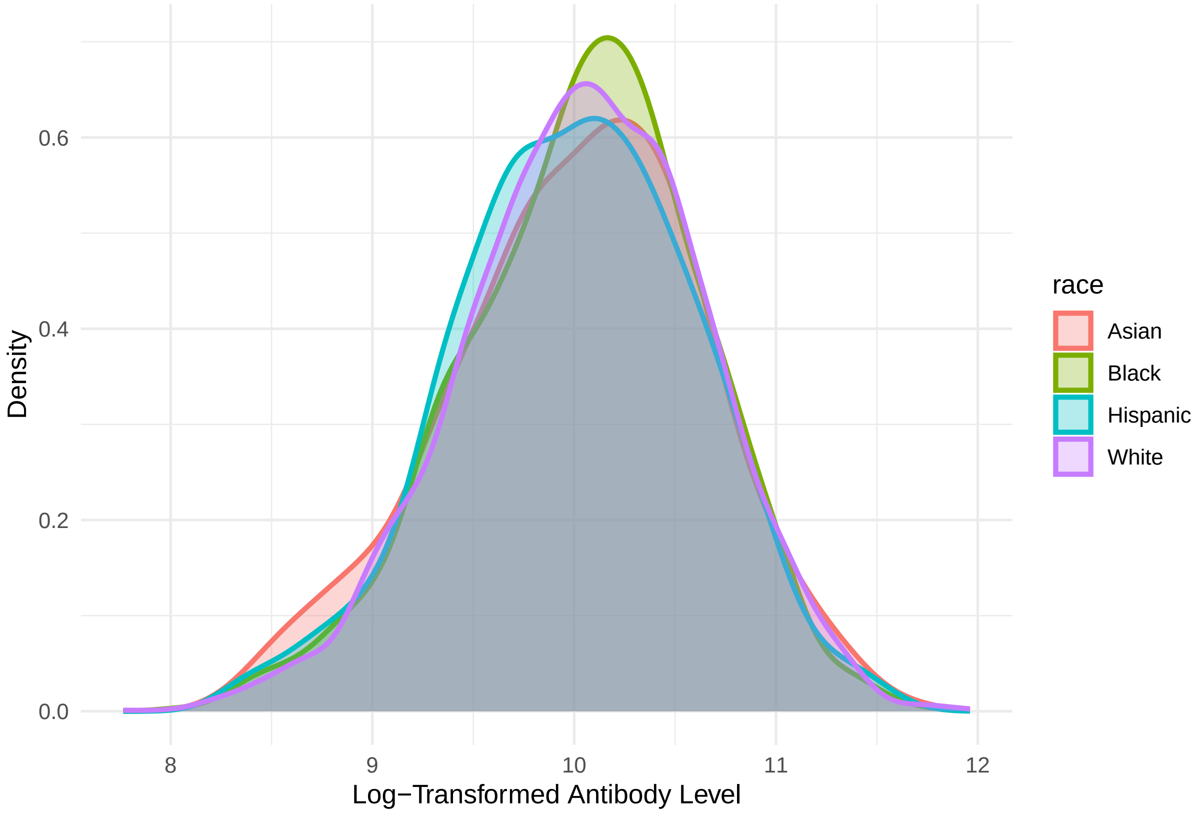


Figure 5: Distribution of Log−Transformed Antibody Level, by Smoking

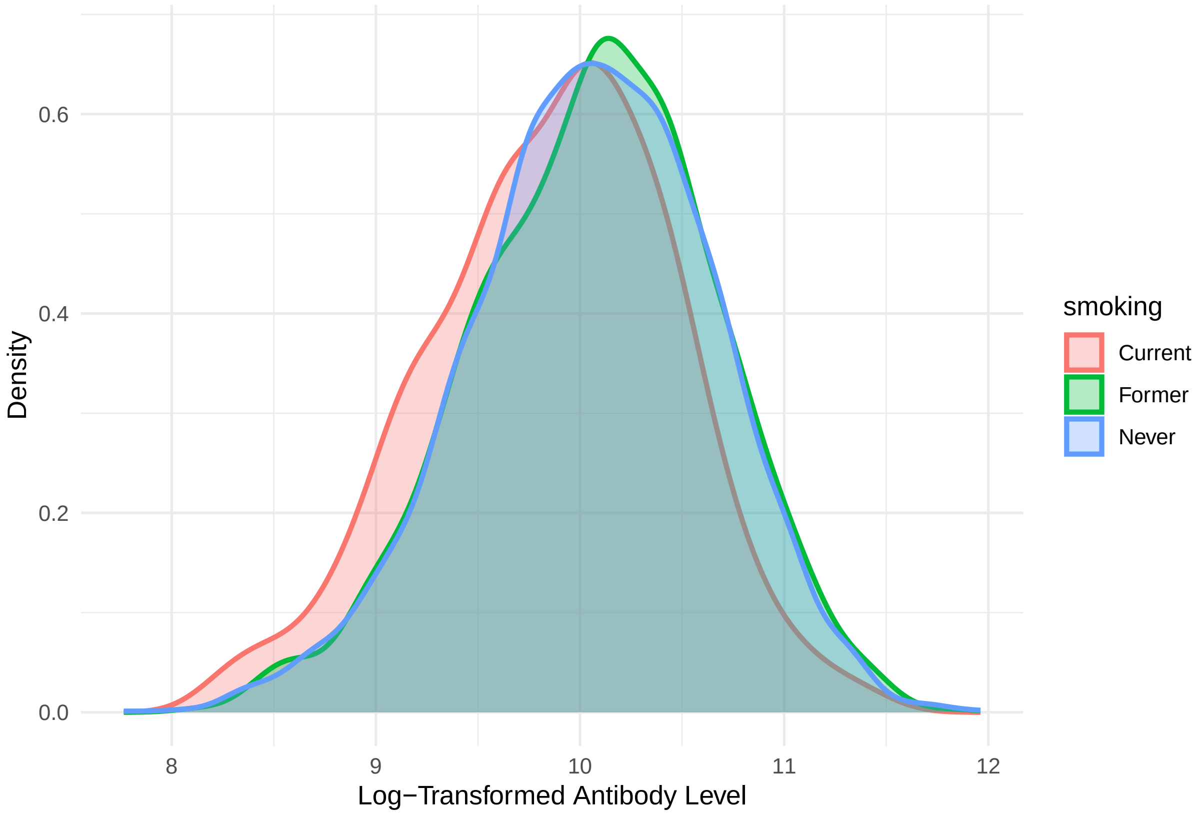


Table 2: Log-Transformed Antibody Level, by Gender

| **Characteristic** | **Female** N = 3,082[1] | **Male** N = 2,918[1] | **p-value**[2] |
|---|---|---|---|
| log_antibody | 10.20 (9.79, 10.58) | 9.93 (9.51, 10.30) | <0.001 |

[1] Median (Q1, Q3)
[2] Wilcoxon rank sum test

Table 3: Log-Transformed Antibody Level, by Race

| **Characteristic** | **Asian** N = 333[1] | **Black** N = 1,235[1] | **Hispanic** N = 548[1] | **White** N = 3,884[1] | **p-value**[2] |
|---|---|---|---|---|---|
| log_antibody | 10.06 (9.62, 10.44) | 10.08 (9.65, 10.44) | 10.03 (9.61, 10.42) | 10.06 (9.65, 10.46) | 0.4 |

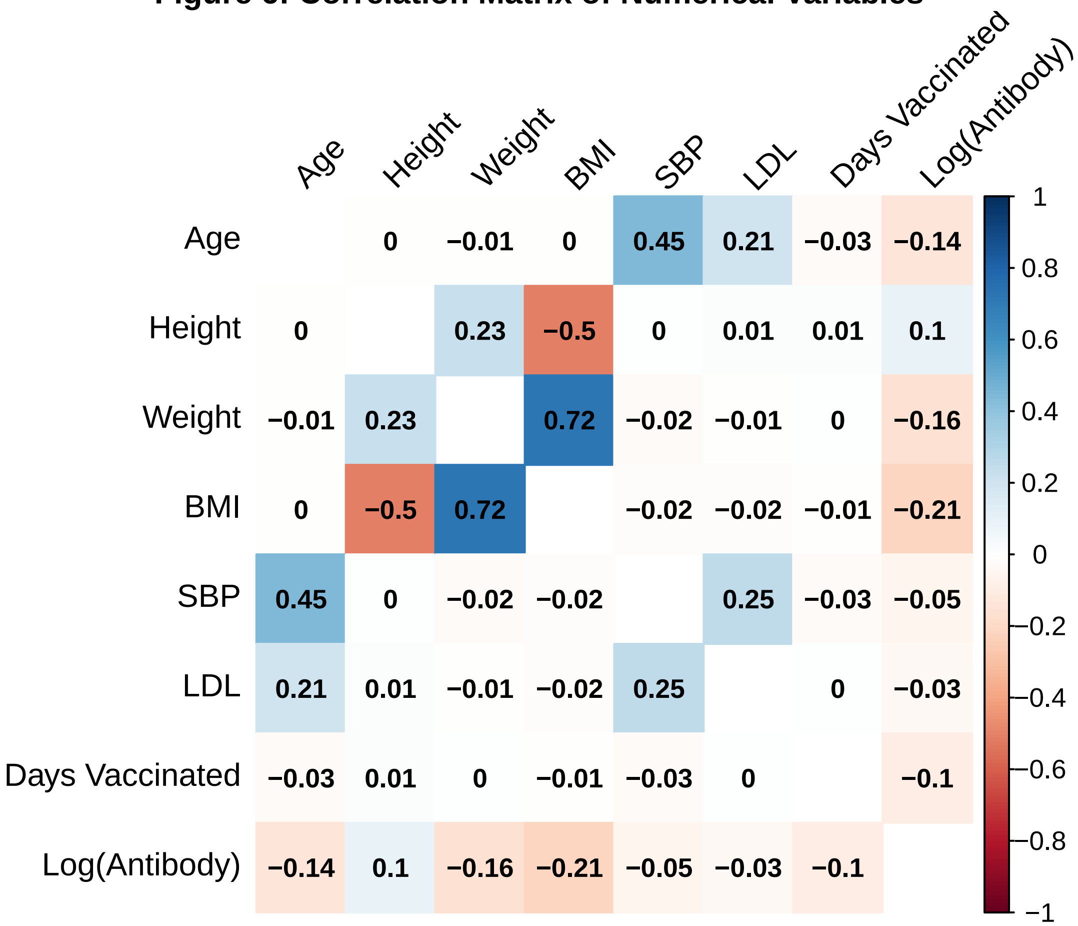
[1] Median (Q1, Q3)
[2] Kruskal-Wallis rank sum test

Table 4: Log-Transformed Antibody Level, by Smoking Status

| **Characteristic** | **Current** N = 589[1] | **Former** N = 1,800[1] | **Never** N = 3,611[1] | **p-value**[2] |
|---|---|---|---|---|
| log_antibody | 9.91 (9.46, 10.28) | 10.10 (9.66, 10.48) | 10.07 (9.68, 10.46) | <0.001 |

[1] Median (Q1, Q3)
[2] Kruskal-Wallis rank sum test

**Figure 6: Correlation Matrix of Numerical Variables**



| | Age | Height | Weight | BMI | SBP | LDL | Days Vaccinated | Log(Antibody) |
|---|---|---|---|---|---|---|---|---|
| Age | | 0 | −0.01 | 0 | 0.45 | 0.21 | −0.03 | −0.14 |
| Height | 0 | | 0.23 | −0.5 | 0 | 0.01 | 0.01 | 0.1 |
| Weight | −0.01 | 0.23 | | 0.72 | −0.02 | −0.01 | 0 | −0.16 |
| BMI | 0 | −0.5 | 0.72 | | −0.02 | −0.02 | −0.01 | −0.21 |
| SBP | 0.45 | 0 | −0.02 | −0.02 | | 0.25 | −0.03 | −0.05 |
| LDL | 0.21 | 0.01 | −0.01 | −0.02 | 0.25 | | 0 | −0.03 |
| Days Vaccinated | −0.03 | 0.01 | 0 | −0.01 | −0.03 | 0 | | −0.1 |
| Log(Antibody) | −0.14 | 0.1 | −0.16 | −0.21 | −0.05 | −0.03 | −0.1 | |

Figure 7: Log−Transformed Antibody Level vs. BMI

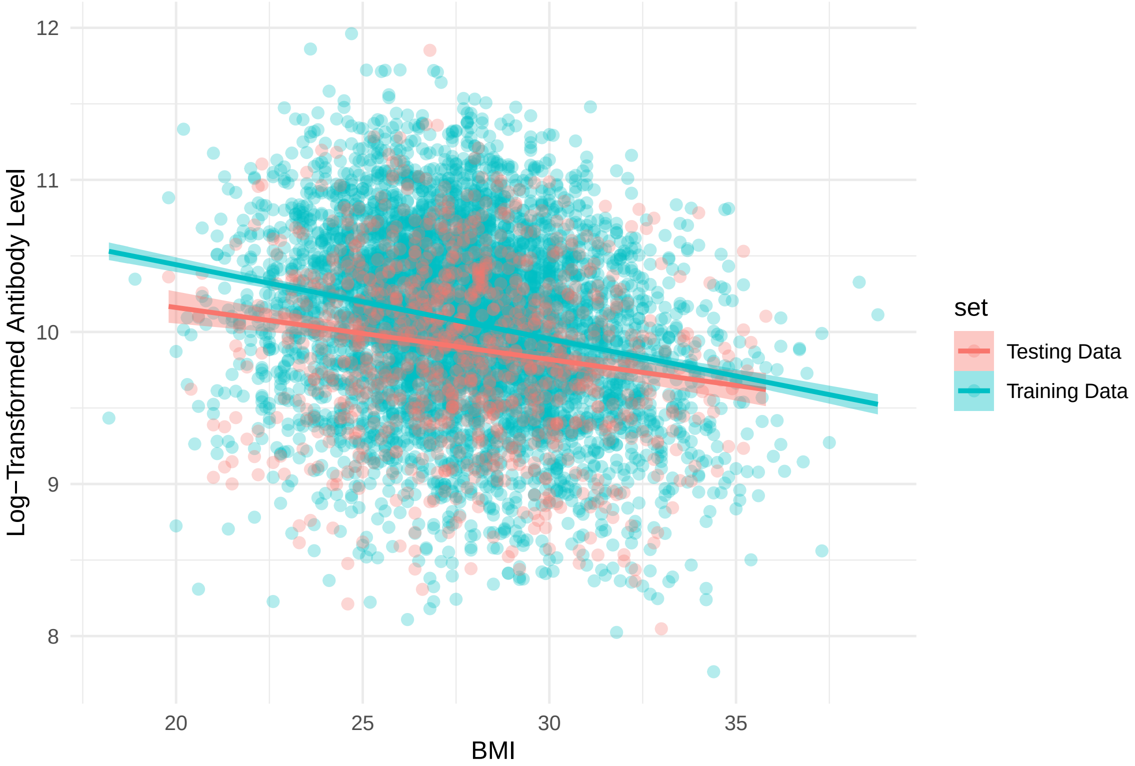
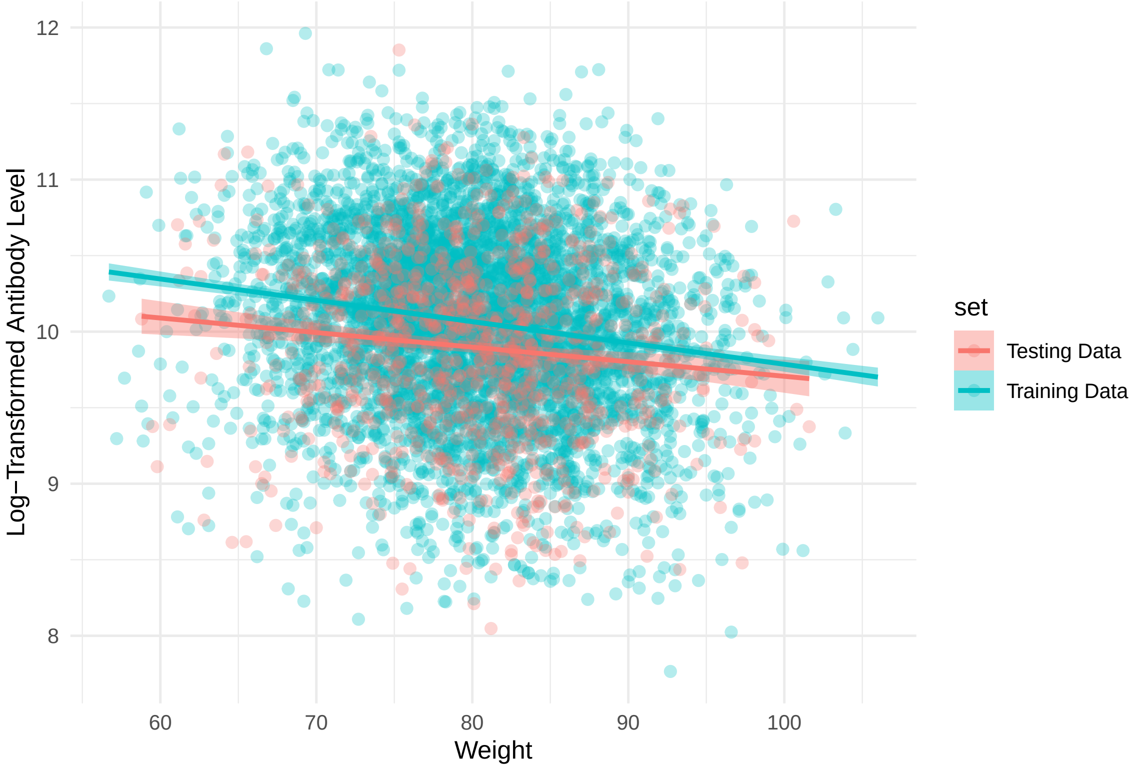


Figure 8: Log−Transformed Antibody Level vs. Weight



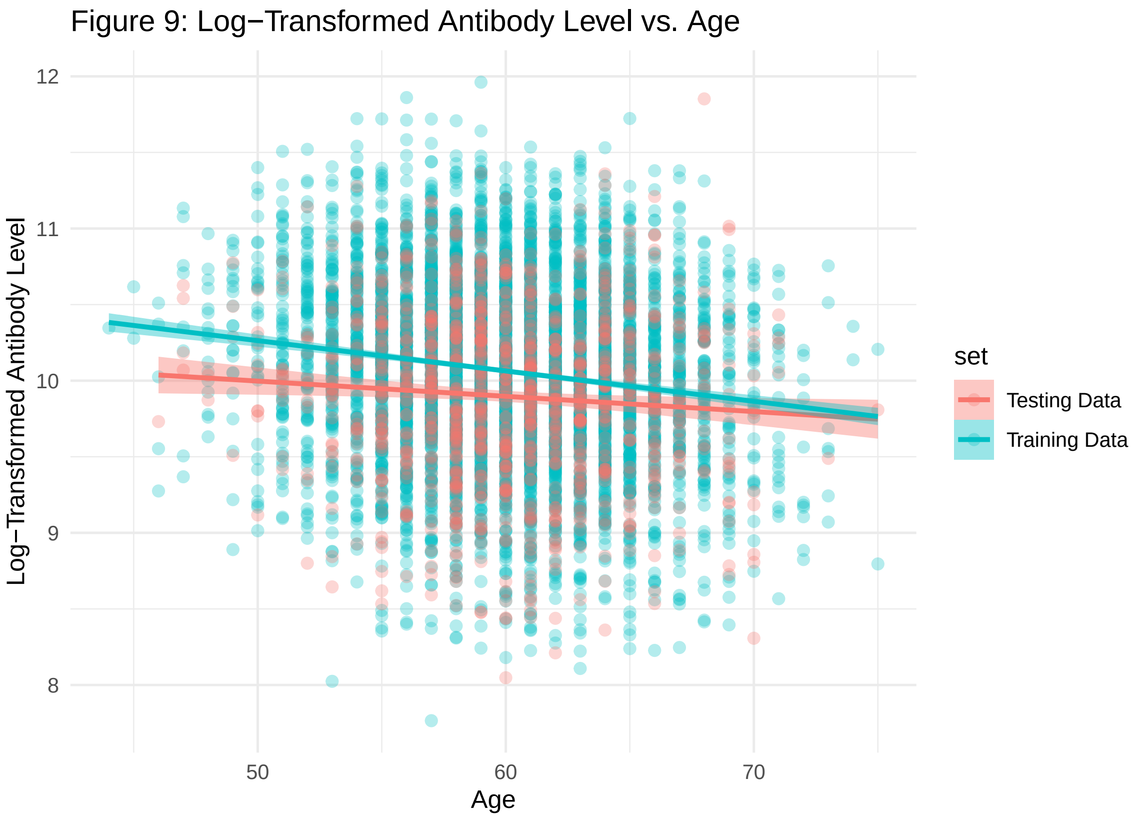


Table 5: Regression Coefficients for MLR Model

| Term | Estimate |
|---|---|
| (Intercept) | 26.7074 |
| id | 0.0000 |
| age | -0.0206 |
| gender | -0.2975 |
| race2 | -0.0064 |
| race3 | -0.0076 |
| race4 | -0.0418 |
| smoking1 | 0.0219 |
| smoking2 | -0.1935 |
| height | -0.0823 |
| weight | 0.0860 |
| bmi | -0.2982 |
| diabetes | 0.0112 |
| hypertension | -0.0175 |
| sbp | 0.0015 |
| ldl | -0.0002 |
| time | -0.0003 |

Table 6: Regression Coefficients for MARS Model

| Term | Estimate |
|---|---|
| (Intercept) | 10.8474 |
| h(27.8-bmi) | -0.0620 |
| h(time-57) | -0.0023 |
| h(57-time) | -0.0335 |
| gender | -0.2963 |
| h(age-59) | -0.0230 |
| h(59-age) | 0.0161 |
| smoking2 | -0.2051 |
| h(bmi-23.7) | -0.0844 |

Figure 10: 10−Fold Cross−Validation RMSE Comparison

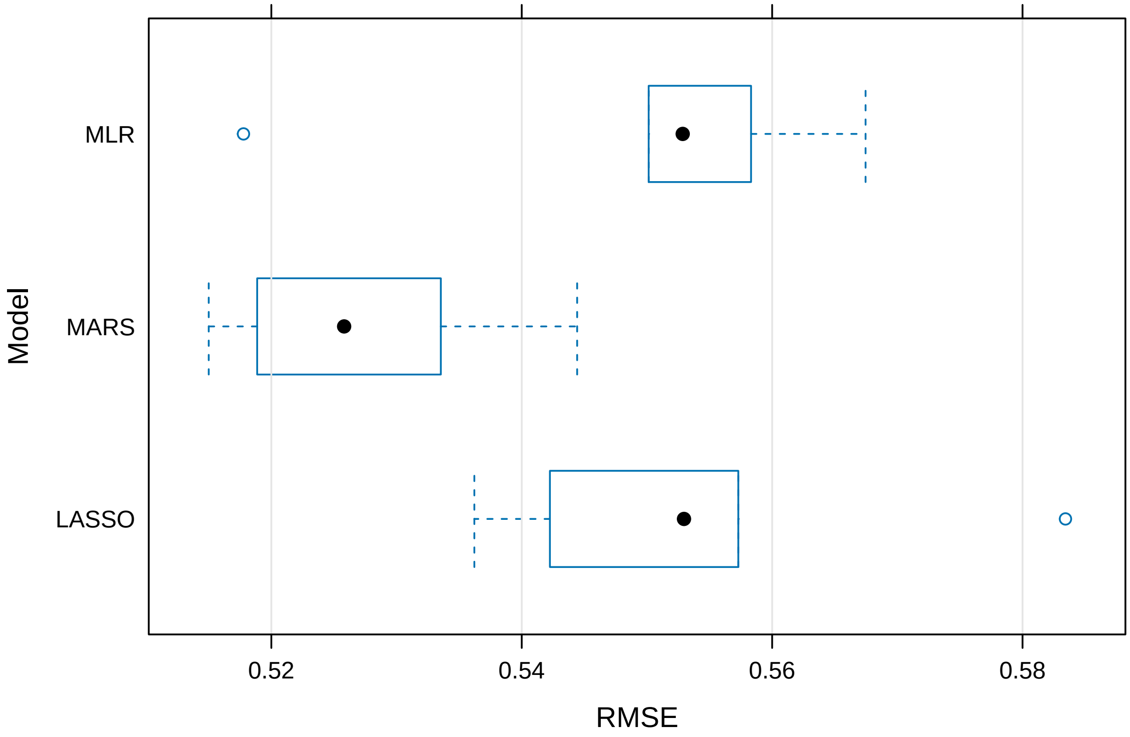


Table 7: Model Performance on Independent Test Set (dat2)

| Model | RMSE | Adjusted R-squared | Notes |
|---|---|---|---|
| Multiple Linear Regression (MLR) | 0.568 | 0.149 | Baseline model; assumes linear relationships |
| LASSO Regression | 0.568 | 0.048 | Performs variable selection via L1 regularization |
| MARS | 0.533 | 0.169 | Best performance; captures non-linear effects |