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 robustness/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). We ultimately selected MARS as the final model, after fine-tuning it using cross-validation. The following sections provide a detailed account of each step in the model training process, from pre-processing to final model selection.

Data Pre-processing

Before modeling, the data underwent the following pre-processing steps:

- Handling Missing Values: We ensured that there were no missing values in the training data. Any missing data would be imputed or removed as appropriate.
- Feature Engineering: Continuous variables were used as they were, while categorical variables were converted to factor types (such as race, gender, smoking).
- Log Transformation: The response variable, log antibody, was log-transformed to normalize its distribution and reduce skewness.
- Also transformed the data

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:

The residuals showed a reasonable fit with no large deviations. Key predictors such as age, gender, bmi, and smoking status showed statistically significant effects, while others, such as race and diabetes, did not appear significant based on p-values.

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.

The training outcome included the estimated coefficients for each feature in the dataset, with significance levels for each predictor.

Overall, the Root Mean Squared Error (RMSE) was 0.544.

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.

In terms of model selection, the best lambda (penalty term) was selected based on cross-validation, which helps to balance model fit and complexity. The model with the lowest cross-validation error was used for evaluation.

The Root Mean Squared Error (RMSE) for LASSO was calculated as 0.544, similar to MLR, indicating comparable predictive performance.

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. This resulted in a complex model of about 13 terms and multiple interactions selected. While it provided a reasonable fit, we sought to prune the model to avoid overfitting.

We fine-tuned the MARS model using cross-validation to determine the optimal number of terms and the degree of interactions. The best parameters were selected as follows:

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

We then fit the final MARS model with those tuned parameters. The RMSE for the tuned MARS model was 0.528, which was slightly better than the MLR and LASSO models. This suggests that the MARS model, after tuning, offers improved predictive performance while avoiding over-fitting.

Based on the RMSE and the results of model selection, MARS was chosen as the final model due to its superior performance and ability to capture complex non-linear relationships between the predictors and the antibody levels. The final MARS model with 10 terms and degree 1 interactions was retained for further evaluation.

Results

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

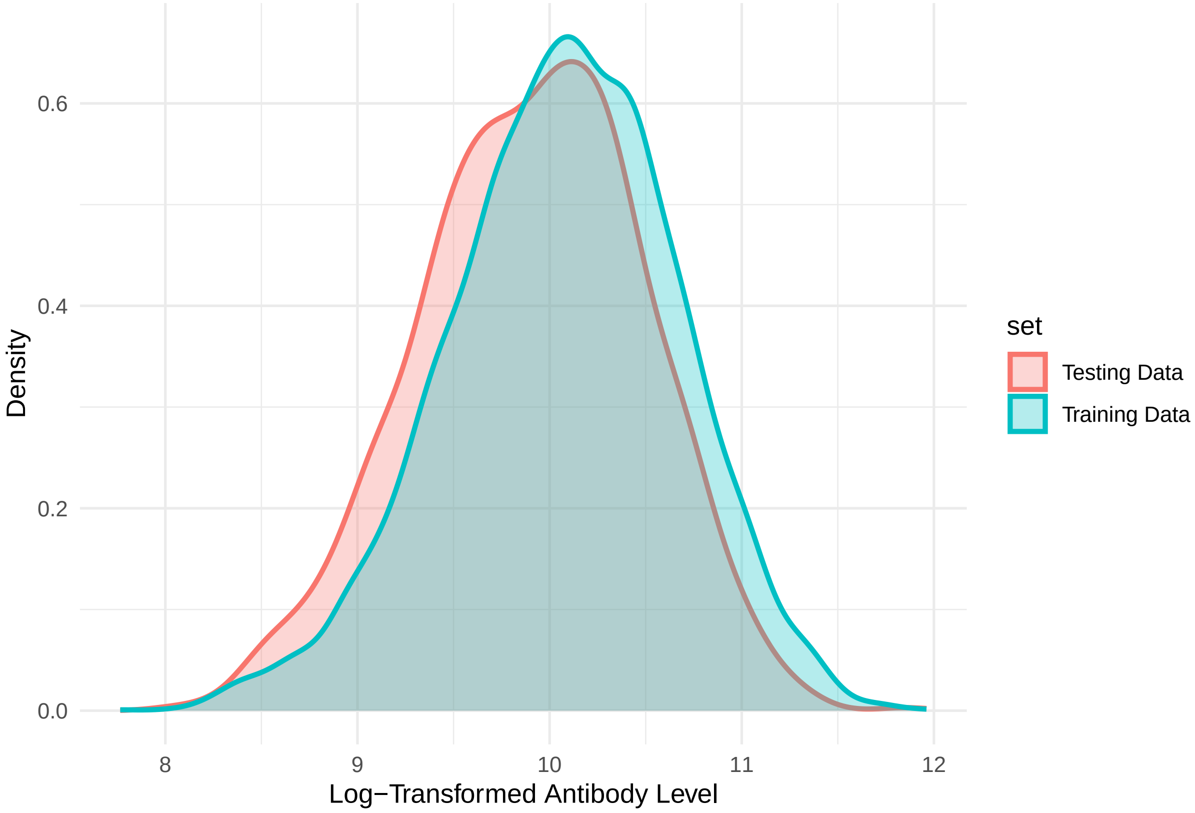


Figure 2: Distribution of Days Since Vaccination, by Data Set

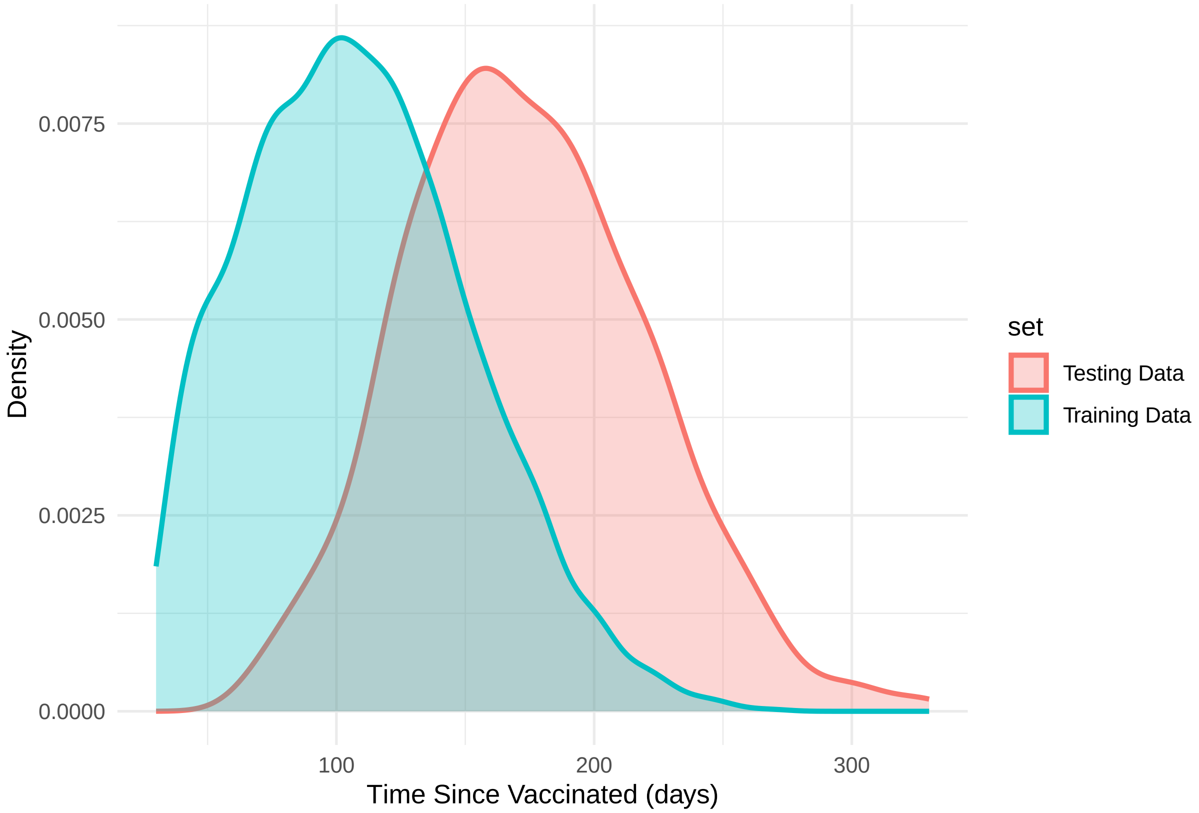


Figure 3: Distribution of Log−Transformed Antibody Level, by Gender

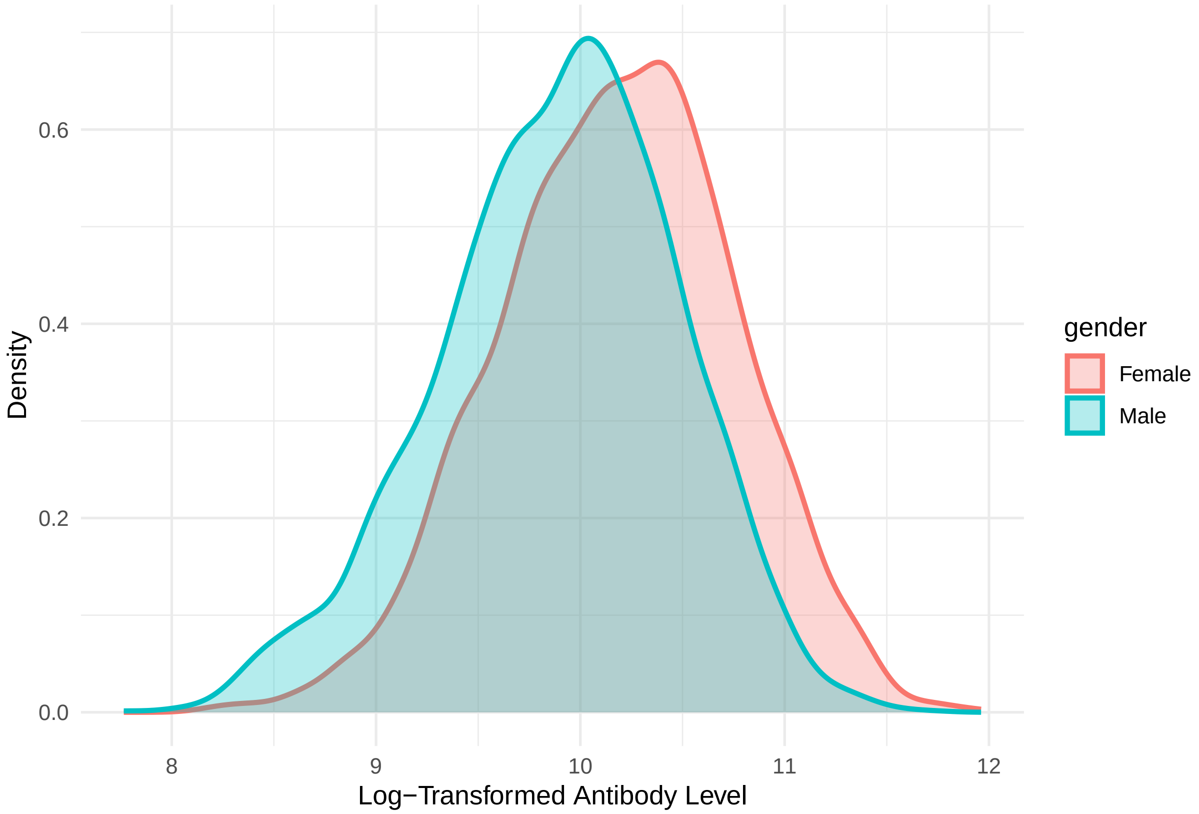


Figure 4: Distribution of Log−Transformed Antibody Level, by Race

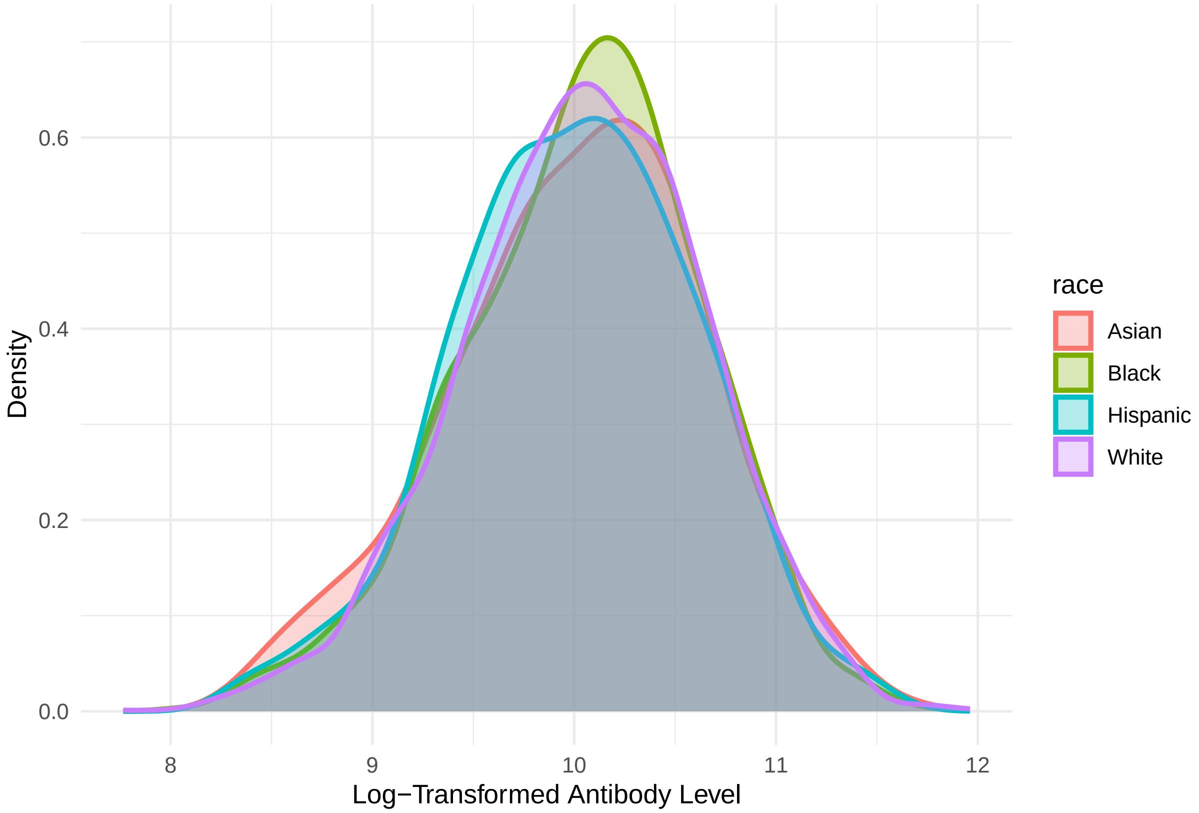


Figure 5: Distribution of Log−Transformed Antibody Level, by Smoking

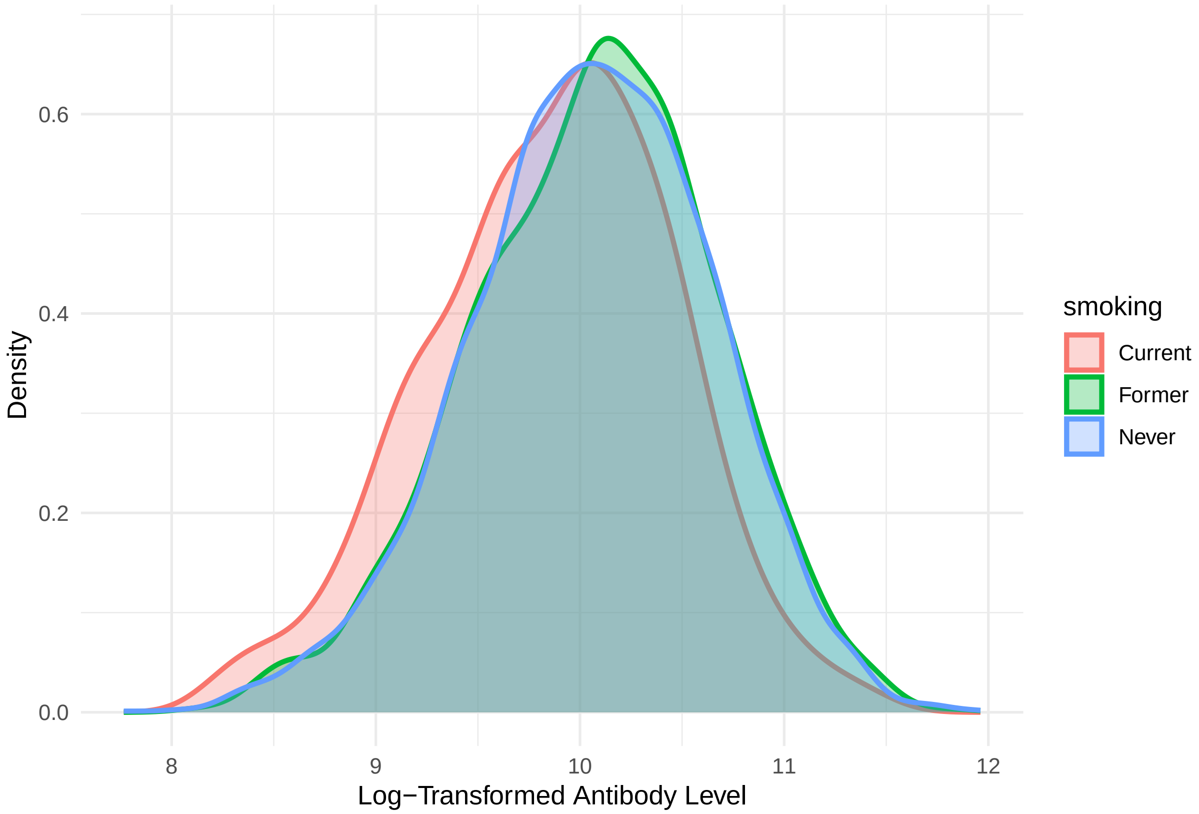


Table 2: Log-Transformed Antibody Level, by Gender

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

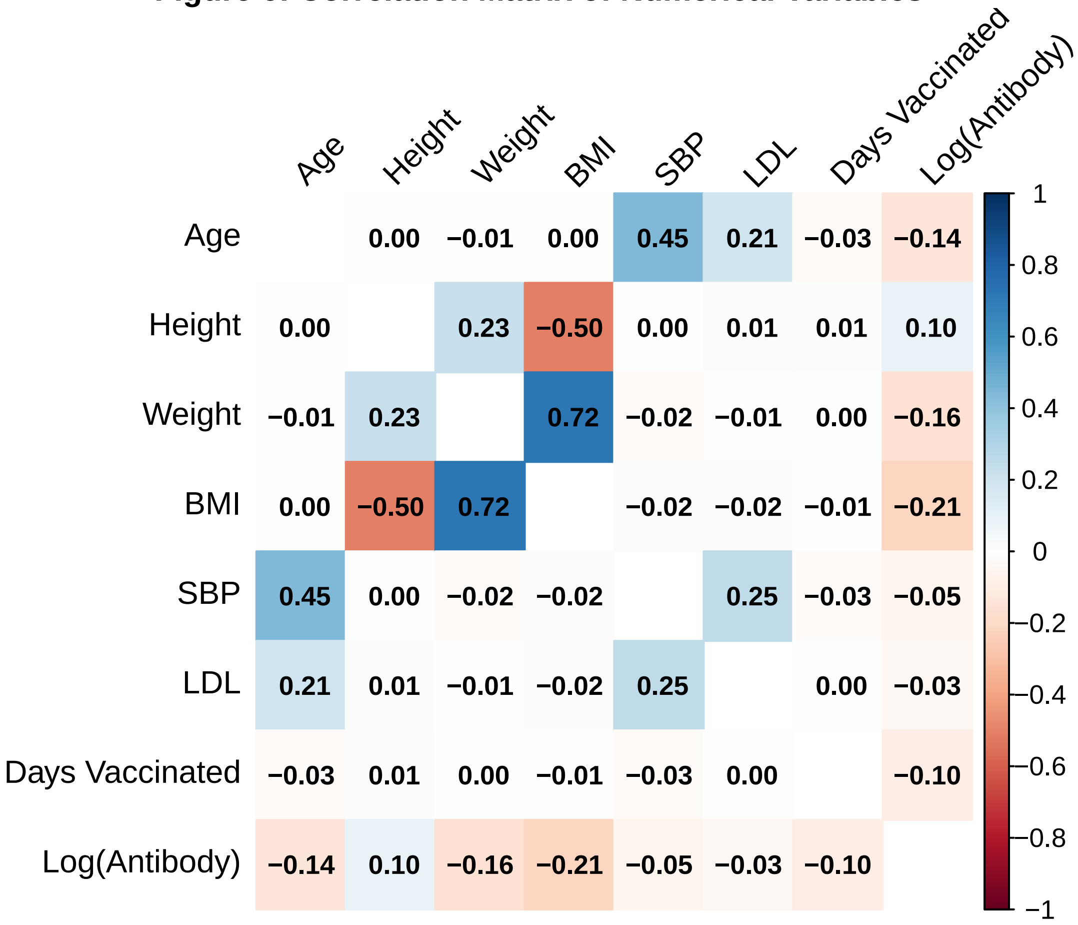


Figure 7: Log−Transformed Antibody Level vs. BMI

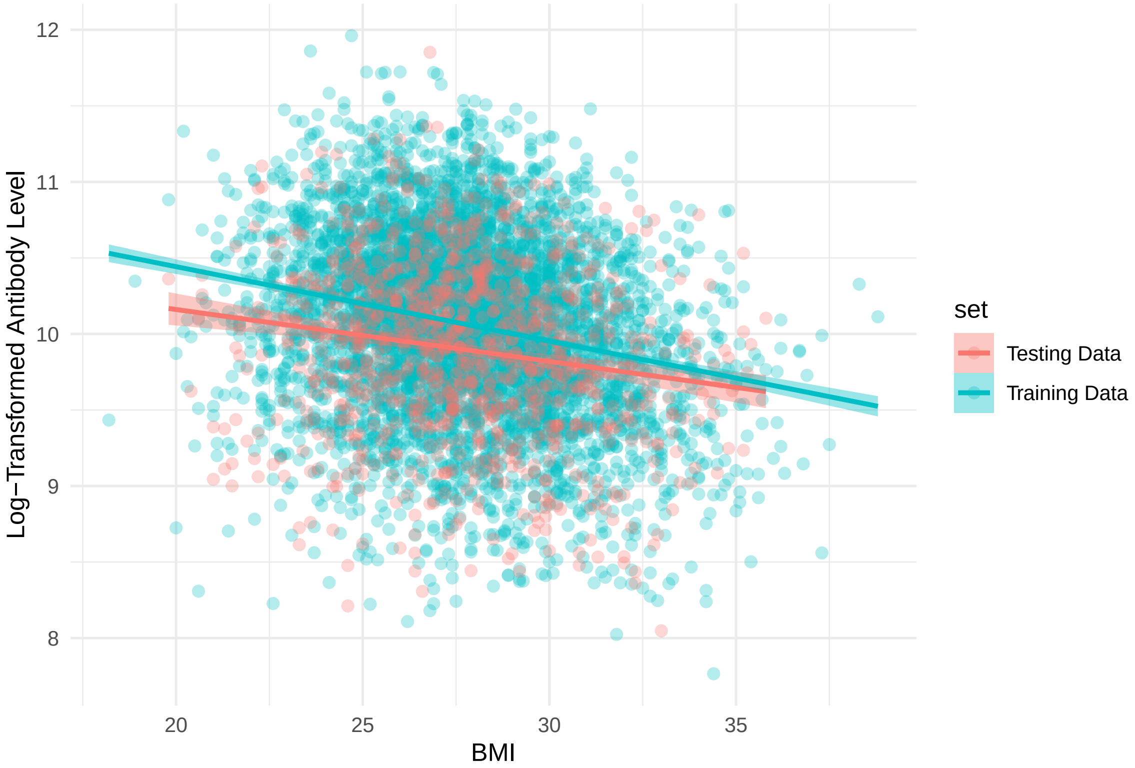


Figure 8: Log−Transformed Antibody Level vs. Weight

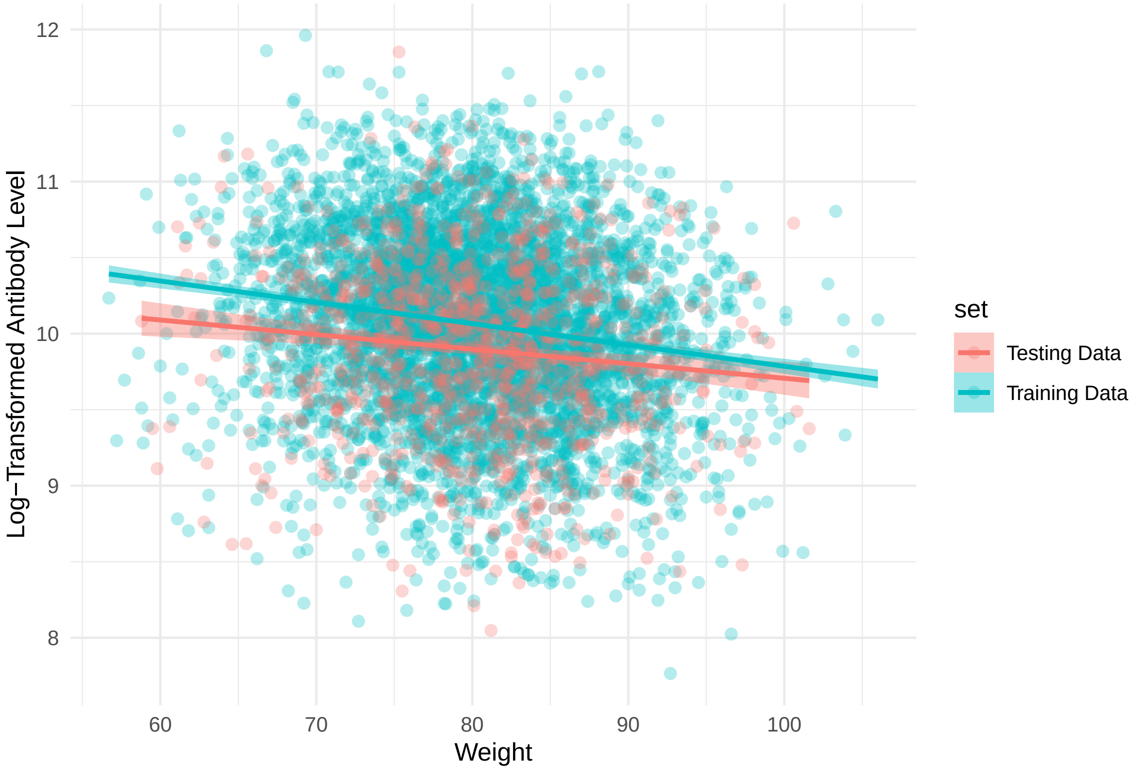
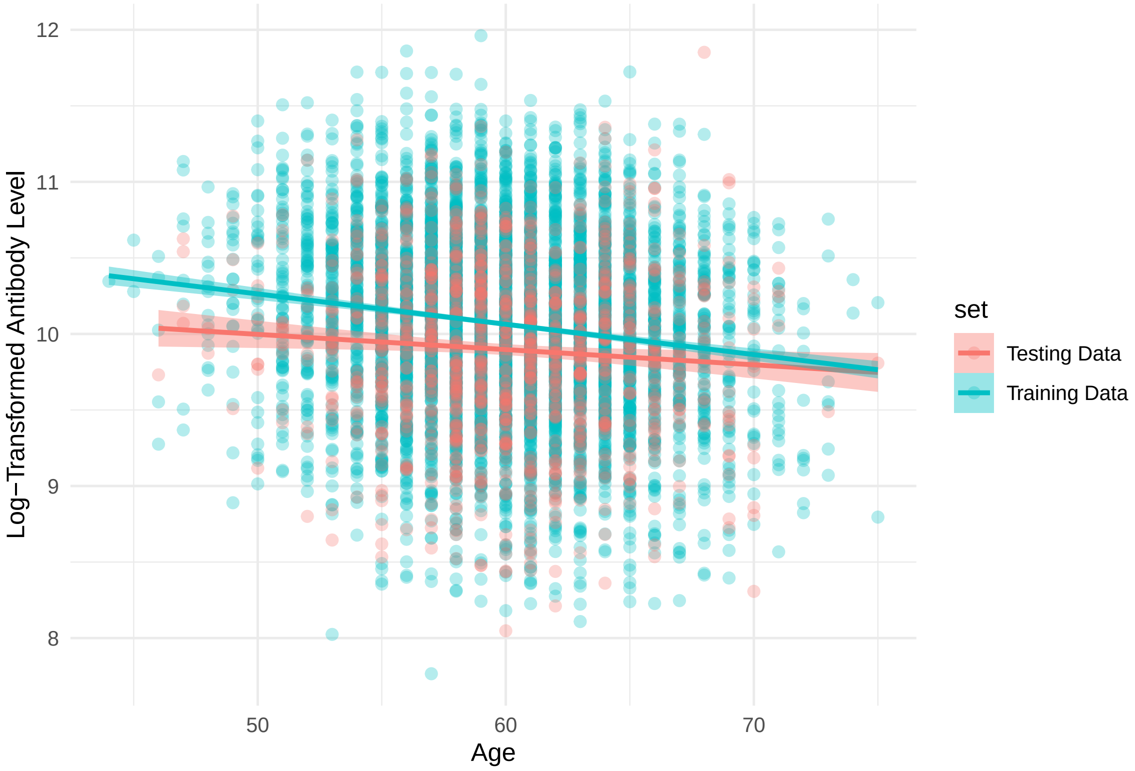


Figure 9: Log−Transformed Antibody Level vs. Age


Log−Transformed Antibody Level
Age
set
Testing Data
Training Data