# Data Science 2 Midterm

Jasmin Martinez (JRM2319), Ixtaccihuatl Obregon, Elliot Kim

03/24/2025

## Exploratory Analysis

For datasets **dat1** and **dat2** initial exploration of the data structure, descriptive statistics of continuous variables, correlation analysis, and various visualization techniques were conducted.

Looking at **dat1**, the distribution of **log_antibody** is approximately normal, as seen in Figure @ref(fig:d1-log-antibody-hist) and Figure @ref(fig:d1-log-antibody-qq). High and low outliers can be observed in the Figure @ref(fig:d1-log-antibody-box). Most covariates have weak or low correlation with each other. There is a high positive correlation between **bmi** and **weight** ($\rho = 0.72$) and a moderate negative correlation between **bmi** and **height** ($\rho = -0.50$). There is a mild negative correlation between **log_antibody** and **bmi** ($\rho = -0.23$), **weight** ($\rho = -0.17$), and **age** ($\rho = -0.15$). There is a mild positive correlation between **log_antibody** and **height** ($\rho = 0.10$). The correlation between **log_antibody** and **SBP** ($\rho = -0.06$), **LDL** ($\rho = -0.04$), and **time** ($\rho = -0.01$) are near zero, indicating no linear relationship. Linear relationships can be seen in Figure @ref(fig:d1-log-antibody-lin).

Exploring **dat2**, the distribution of **log_antibody** is also approximately normal, as shown Figure @ref(fig:d2-log-antibody-hist) and Figure @ref(fig:d2-log-antibody-qq). High and low outliers are again visible in the Figure @ref(fig:d2-log-antibody-box). Most covariates exhibit weak or low correlation with each other. There is a high positive correlation between **bmi** and **weight** ($\rho = 0.72$) and a moderate negative correlation between **bmi** and **height** ($\rho = -0.53$). There is a mild negative correlation between **log_antibody** and **bmi** ($\rho = -0.16$), **weight** ($\rho = -0.11$), and **age** ($\rho = -0.08$). A mild positive correlation exists between **log_antibody** and **height** ($\rho = 0.084$). The correlations between **log_antibody** and **SBP** ($\rho = -0.01$), **LDL** ($\rho = -0.00$), and **time** ($\rho = -0.25$) are near zero or weak, again suggesting no strong linear relationship. Linear relationships can be seen in Figure @ref(fig:d2-log-antibody-lin).

The exploration and evaluation of **dat1** is used to help build a prediction model specific ally using GAM to understand how demographic and clinical factors influence antibody responses and how antibody levels change over time following vaccination. Considering the researcher collects a new and independent dataset, **dat2**, we discover the correlation between **time** and **log_antibody** is stronger than in **dat1** and similarly has weak linear relationships with most covariates. **Dat2** allows us to evaluate the robustness and generalizability of our prediction model.

## Model Trainging

The **dat1** dataset is the designated training data and **dat2** is the testing dataset. Using the **train()** function, a multiple linear regression , ridge, lasso, generalized additive model, and a multivariate adaptive regression splines model were fitted to determine the optimal model for the data. All models used a 10-fold cross-validation that was determined by using **method = cv** and **number = 10** in the **trainControl()** function.

**Multiple Linear Regression Model** A multiple linear regression model is simple and easy to use when working with linearly dependent data but works under the assumption of homoscedasticity and no multicollinearity. Fitting this model required **method = lm** in the **train()** function, resulting in $R^2 = 0.1470498$ and $R^2_{adj} = 0.03320516$ indicating a poor fit.

**Ridge and Lasso Regression**

Ridge and lasso models are useful for improving prediction accuracy when dealing with multiple predictors. These models help prevent overfitting through the inclusion of a penalty term that regularizes coefficient estimates. A key component to these models is the lambda tuning parameter, which controls the strength of the penality applied to the coefficients.

The models were fit using the Caret train() function and specifying the method = "glmnet", which allows the adjustment of the alpha parameter in which alpha = 1 for a lasso regression (L1 regularization) and alpha = 0 for a ridge regression (L2 regularization). The tuning parameter (lambda) for each model was selected by a 10-fold cross-validation. To explore a wide range of lambda values, a sequence of 100 exponentially spaced values between -5 and 6 was specified in the tuneGrid argument of Caret's train() function. These models assume that the relationship between the predictors and the outcome is linear.

While the lasso regression model conducts variable selection by shrinking coefficients to 0, the ridge regression model can only shrink coefficients towards 0. Therefore, the lasso regression model is useful for handling multiple predictors while the ridge regression model helps reduce the size of the coefficients for all predictors and is useful for handling predictors that are highly correlated, thereby preventing multicollinearity.

**Generalized Additive Model (GAM)**

A Generalized Additive Model (GAM) was created to model log-transformed antibody level. Predictors included age, gender, race, smoking, height, weight, BMI, diabetes, hypertension, SBP, LDL, and time since vaccination. To prepare the data for modeling, the model matrix, x, was created using the outcome and all predictors listed above and extracted the response variable, y. Given that race and smoking were categorical variables with multiple categories, some re-coding was necessary to use these variables in the model building. Therefore, race and smoking were converted into factor variables with "White" being the reference group for race and "Never" being the reference group for smoking.

The GAM models were then fitted. The first model, gam.m1, is a standard linear model with no smoothing terms. The second model, gam.m2, uses smoothing on age, bmi, SBP, LDL, and time since vaccination. There was reason to believe these variables were non-linear, therefore the smoothing allowed the variables to be used in the model. Finally, model 3, gam.m3, includes a tensor product to model the interaction between height and weight.

The three GAM models were then compared using an ANOVA test that provides the f-test; this was used to determine which model provides the best fit. Cross-validation for GAM tuning was then conducted using a 10-fold cross-validation to tune to GAM model. The best hyperparameters and the final fitted model were then retrieved. The same model training procedure was used with the new, independent dataset, dat2.

Model 2 (gam.m2) was chosen to be the final model based on the improvements it made upon Model 1 (gam.m1), (pval=<0.001).

**Multivariate Adaptive Regression Splines Model**

The Multivariate Adaptive Regression Splines (MARS) model is a flexible non-linear model technique used to evaluate the relationship between an outcome and multiple covariates. MARS divides the data into segments and fits a piecewise linear regressions for each segment. MARS is used to model **log_antibody** levels based on clinical and demographic predictors. The model is tuned on two hyperparameters: **nprune** for the most number of terms and **degree** for the most amount of degree interactions between predictors.

**Re-sampling and final model selection**

We used the **resample()** function to conduct a robust comparison of all the models listed above, with the performance metric being the Root Mean Squared Error (RMSE). The models included in this analysis were Multiple Linear Regression, Lasso Regression, Ridge Regression, Multivariate Adaptive Regression Splines

(MARS), and the Generalized Additive Model (GAM). As illustrated in the results below, the GAM model demonstrated the lowest RMSE, with a value of 0.531, indicating its superior predictive accuracy. This was closely followed by the MARS model, which had an RMSE of 0.530, highlighting its strong performance as well. Consistent with expectations, the simplest model, Multiple Linear Regression, exhibited the highest RMSE of 0.552, reflecting its limitations in capturing complex relationships within the data compared to more advanced models.

# Final Model Results

The final model to predict **log_antibody** levels was the Generalized Additive (GAM) Model. The model equation is given by:

$$\log_{antibody} = \beta_0 + s_1(\text{age}) + \beta_1 \cdot \text{gender} + \beta_2 \cdot \text{race} + \beta_3 \cdot \text{smoking} + \beta_4 \cdot \text{height} + \beta_5 \cdot \text{weight} + s_2(\text{bmi}) + \beta_6 \cdot \text{diabetes} + \beta_7 \cdot \text{hypertension} + s_3($$

The prediction model's robustness and generalizability is mostly acceptable for **dat2**. Prediction accuracy was determined by the mean squared error (MSE) at 0.325 showing a low average on unseen data. The prediction model shows model stability with generalized cross-validation (GCV) score of 0.279. Looking at @ref(fig:dx-plots) we evaluate Predicted vs Actual log_antibody Levels, Residuals vs Predicted, and the Distribution of Residuals. Predictions are mostly aligned with the observed values especially in the 9.5-10.5 predicted range. There is mild under-prediction below value 9, which indicates that the model may tend to underestimate lower antibody levels but perform well in the 9.5-10.5 predicted range. The predicted vs residual plot shows residuals mostly centered and near 0. Right-skewness can be observed in the predicted vs residual plots, which could suggest some non-linearity. The distribution of residuals looks approximately normal and does not have extreme outliers or multi-modality.

Table 1: Clean Summary Statistics for Numeric Variables

| Variable | Min | Q1 | Median | Mean | Q3 | Max |
|---|---|---|---|---|---|---|
| id | 1.000000 | 1250.750000 | 2500.50000 | 2500.50000 | 3750.25000 | 5000.00000 |
| age | 44.000000 | 57.000000 | 60.00000 | 59.96840 | 63.00000 | 75.00000 |
| gender | 0.000000 | 0.000000 | 0.00000 | 0.48540 | 1.00000 | 1.00000 |
| height | 150.200000 | 166.100000 | 170.10000 | 170.12634 | 174.22500 | 192.90000 |
| weight | 56.700000 | 75.400000 | 80.10000 | 80.10908 | 84.90000 | 106.00000 |
| bmi | 18.200000 | 25.800000 | 27.60000 | 27.74040 | 29.50000 | 38.80000 |
| diabetes | 0.000000 | 0.000000 | 0.00000 | 0.15440 | 0.00000 | 1.00000 |
| hypertension | 0.000000 | 0.000000 | 0.00000 | 0.45960 | 1.00000 | 1.00000 |
| SBP | 101.000000 | 124.000000 | 130.00000 | 129.90040 | 135.00000 | 155.00000 |
| LDL | 43.000000 | 96.000000 | 110.00000 | 109.90860 | 124.00000 | 185.00000 |
| time | 30.000000 | 76.000000 | 106.00000 | 108.86260 | 138.00000 | 270.00000 |
| log | 7.765405 | 9.681635 | 10.08908 | 10.06434 | 10.47758 | 11.96137 |

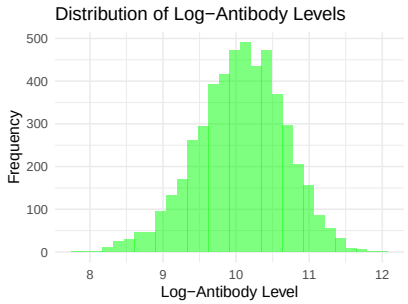



Figure 1: Histogram of log_antibody levels for dat1

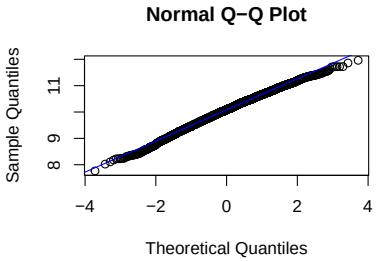



Figure 2: Q-Q plot of log_antibody levels for dat1

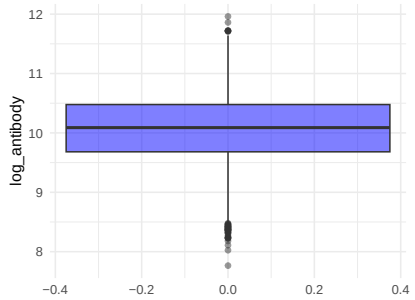



Figure 3: Boxplot of log_antibody levels for dat1

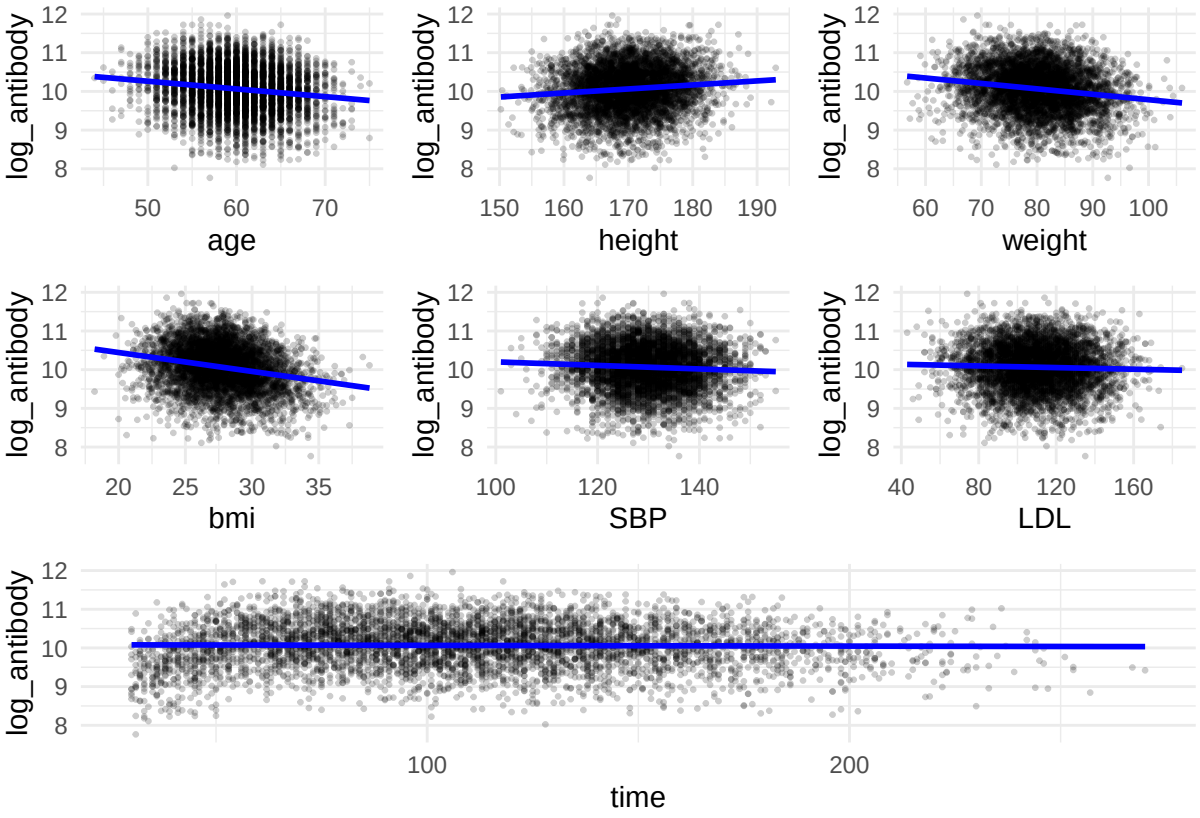



Figure 4: Linearity between log_antibody levels and covariates for dat1

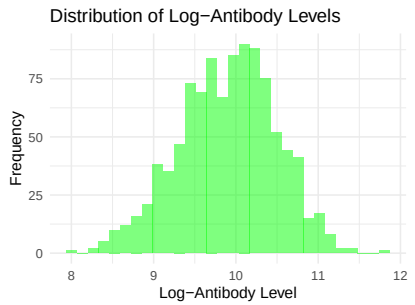



Figure 5: Histogram of log_antibody levels for dat2

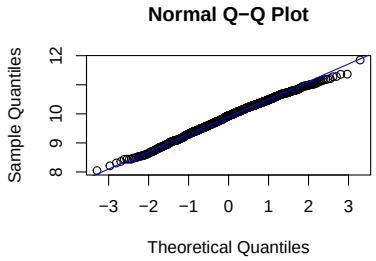



Figure 6: Q-Q plot of log_antibody levels for dat2

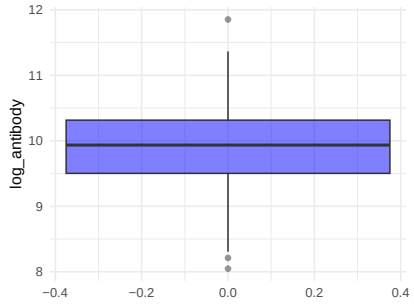



Figure 7: Boxplot of log_antibody levels for dat2

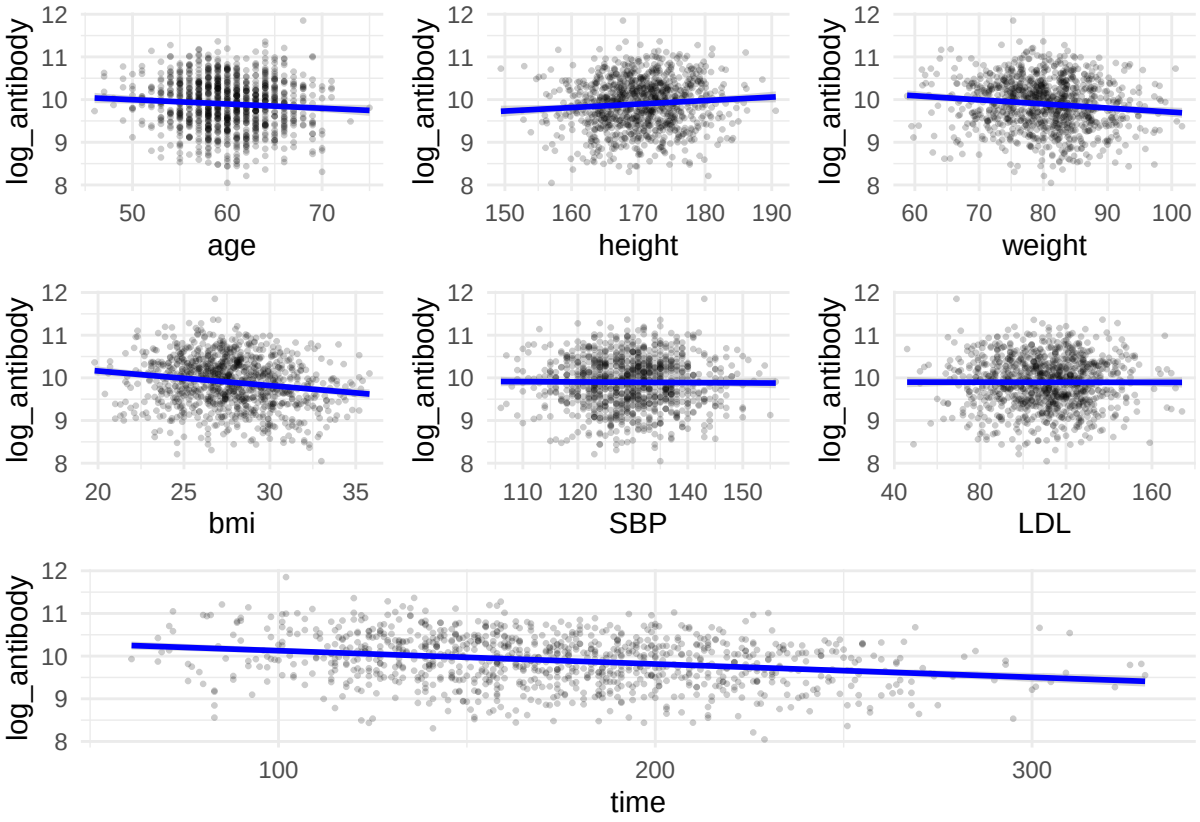
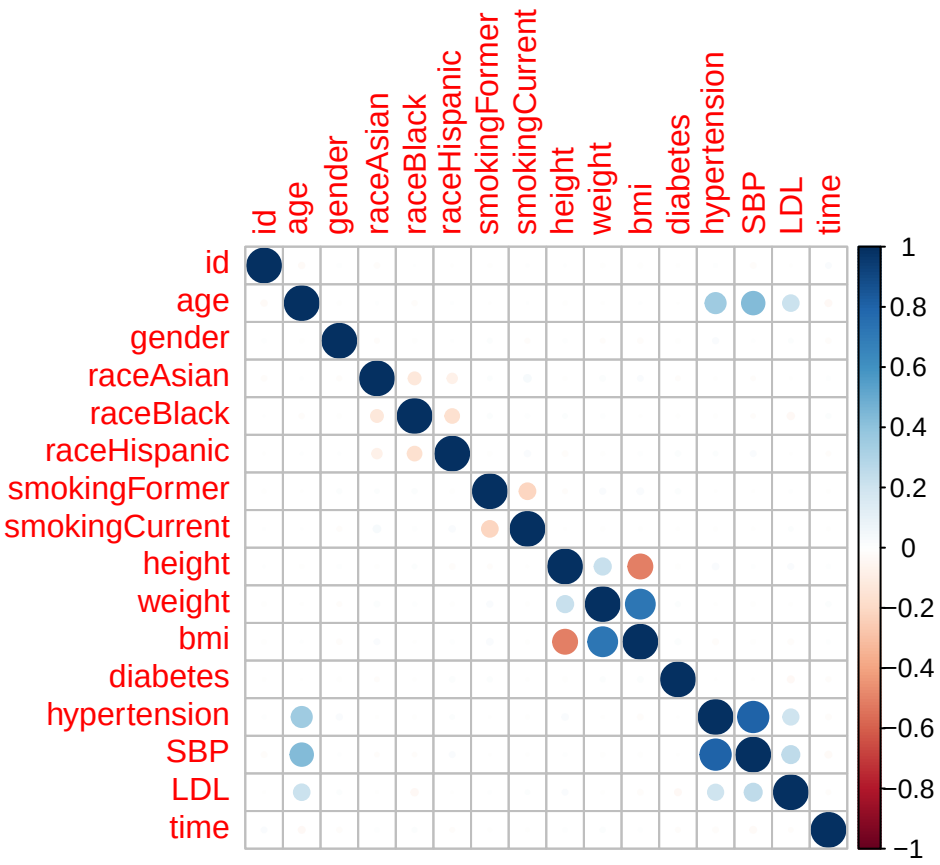
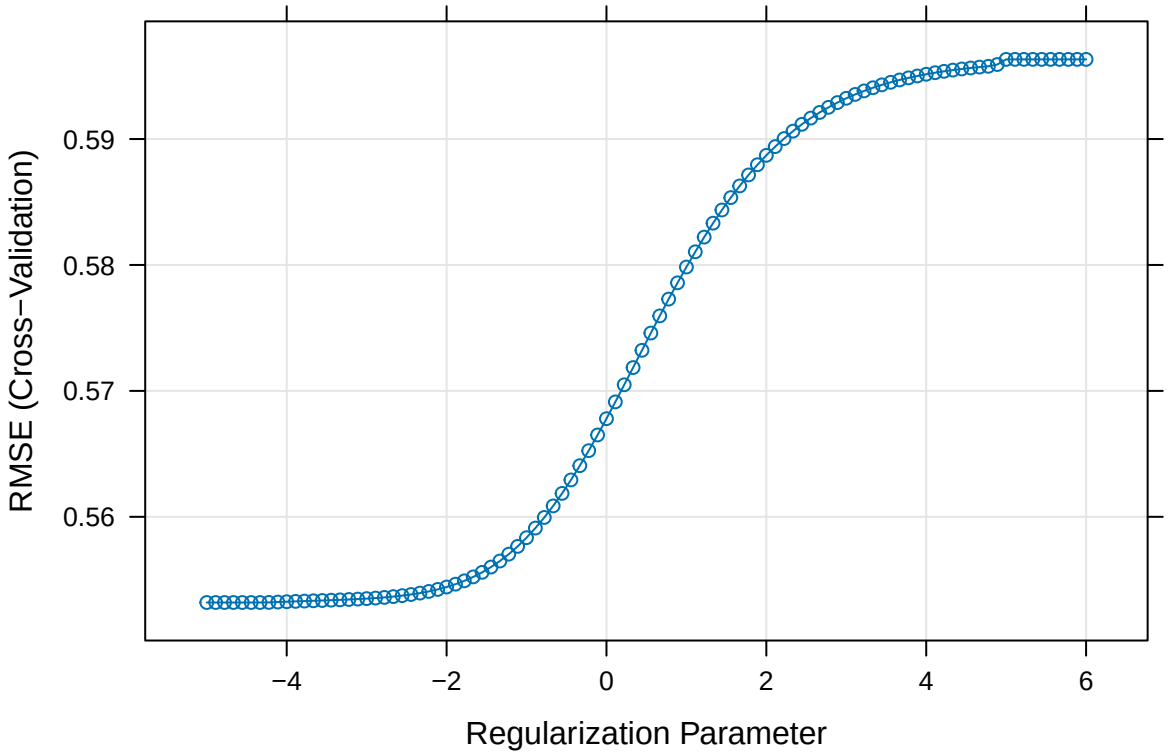



Figure 8: Linearity between log_antibody levels and covariates for dat2




```
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.
```




```
##   alpha     lambda
## 7     0 0.01312373
```

```
## 17 x 1 sparse Matrix of class "dgCMatrix"
##                              s1
## (Intercept)     1.274791e+01
## id             -1.894551e-06
## age            -1.977970e-02
## gender         -2.880556e-01
## raceAsian      -4.051537e-03
## raceBlack      -6.683819e-03
## raceHispanic   -4.179393e-02
## smokingFormer   2.418667e-02
## smokingCurrent -1.847083e-01
## height         -2.046729e-04
## weight         -9.107320e-04
## bmi            -4.735061e-02
## diabetes        1.128129e-02
## hypertension   -1.636703e-02
## SBP             1.066327e-03
## LDL            -1.608202e-04
## time           -2.795454e-04
```

```
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.
```

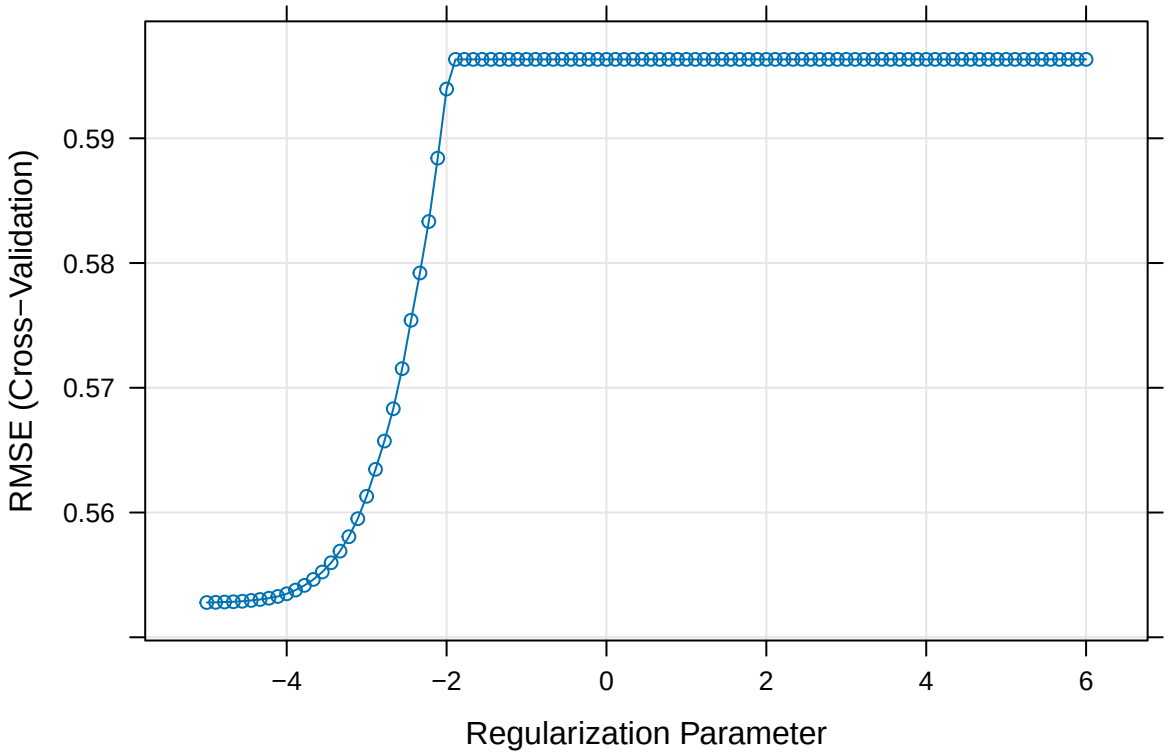



Table 2: Optimal Tuning Parameters for Ridge and Lasso Regression

| Model | Alpha | Lambda |
|---|---|---|
| Ridge | 0 | 0.0131237 |
| Lasso | 1 | 0.0075298 |

Table 3: ANOVA Comparison of GAM Models (Dat1)

| Model | Resid. Df | Resid. Dev | Df | Deviance | F | Pr(>F) |
|---|---|---|---|---|---|---|
| gam.m1 | 4984.000 | 1509.442 | NA | NA | NA | NA |
| gam.m2 | 4971.177 | 1380.053 | 12.82350 | 129.389611 | 36.37749 | 0.0000000 |
| gam.m3 | 4968.540 | 1378.818 | 2.63652 | 1.234568 | 1.68820 | 0.1738572 |

Table 4: ANOVA Comparison of GAM Models (Dat2)

| Model | Resid. Df | Resid. Dev | Df | Deviance | F | Pr(>F) |
|---|---|---|---|---|---|---|
| gam.m1 | 984.0000 | 275.4546 | NA | NA | NA | NA |
| gam.m2 | 977.7820 | 268.5783 | 6.217997 | 6.8762407 | 4.0341782 | 0.0004416 |
| gam.m3 | 976.6323 | 268.2745 | 1.149679 | 0.3038207 | 0.9640406 | 0.3384056 |

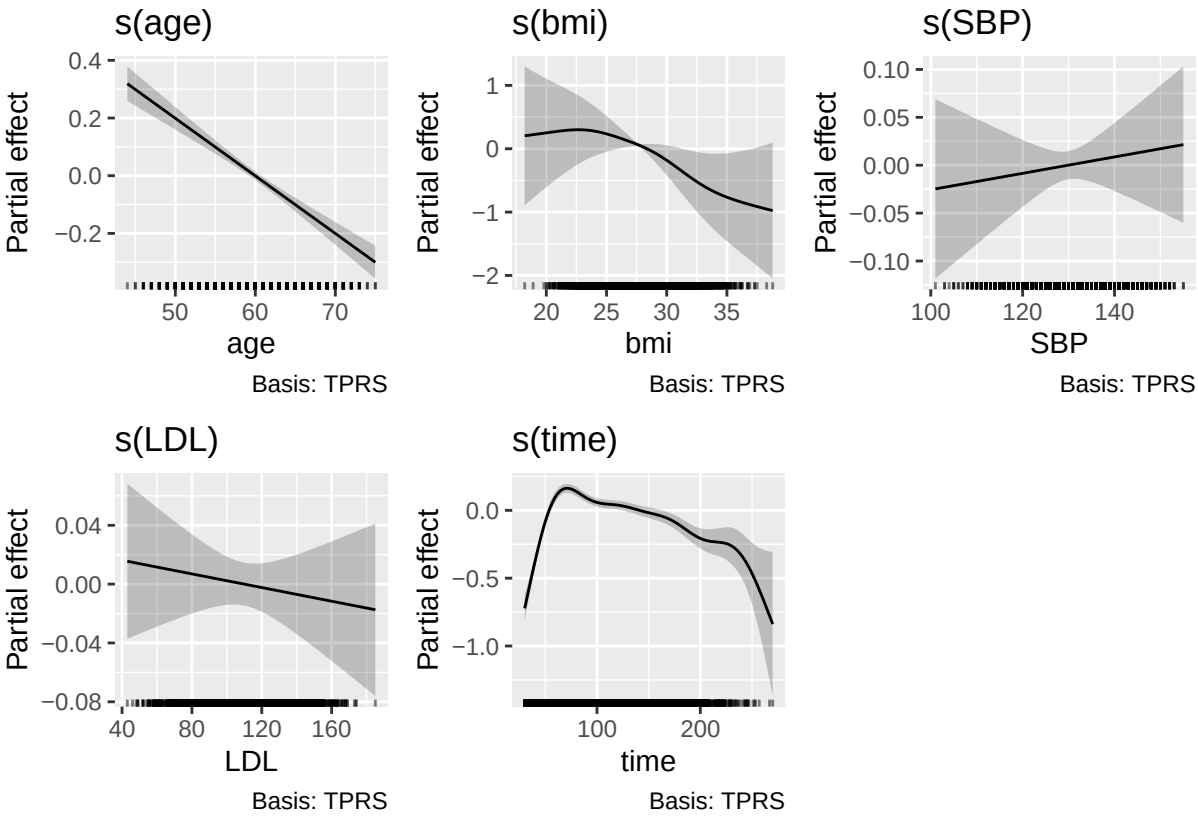
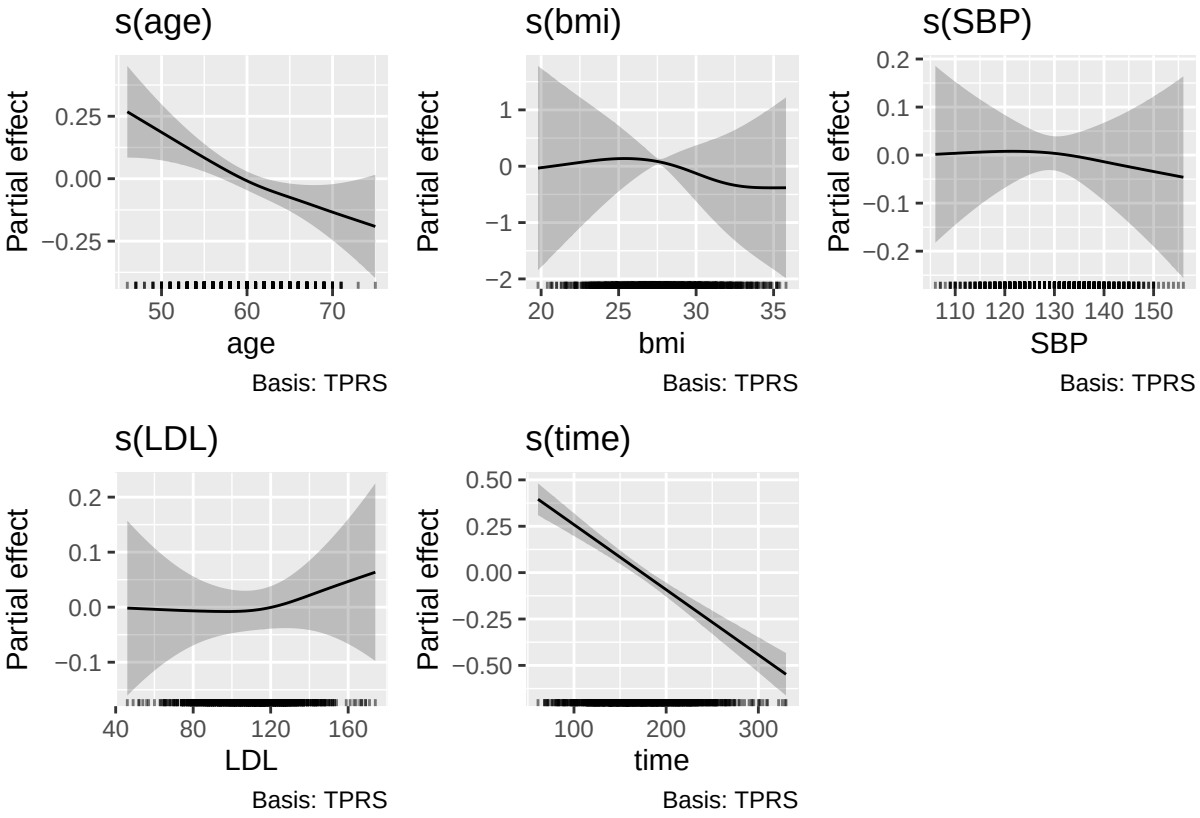



Figure 9: Smooth terms for GAM model (gam.m2)


s(age)
Partial effect
0.25
0.00
−0.25
50
60
70
age
Basis: TPRS
s(bmi)
Partial effect
1
0
−1
−2
20
25
30
35
bmi
Basis: TPRS
s(SBP)
Partial effect
0.2
0.1
0.0
−0.1
−0.2
110 120 130 140 150
SBP
Basis: TPRS
s(LDL)
Partial effect
0.2
0.1
0.0
−0.1
40
80
120
160
LDL
Basis: TPRS
s(time)
Partial effect
0.50
0.25
0.00
−0.25
−0.50
100
200
300
time
Basis: TPRS

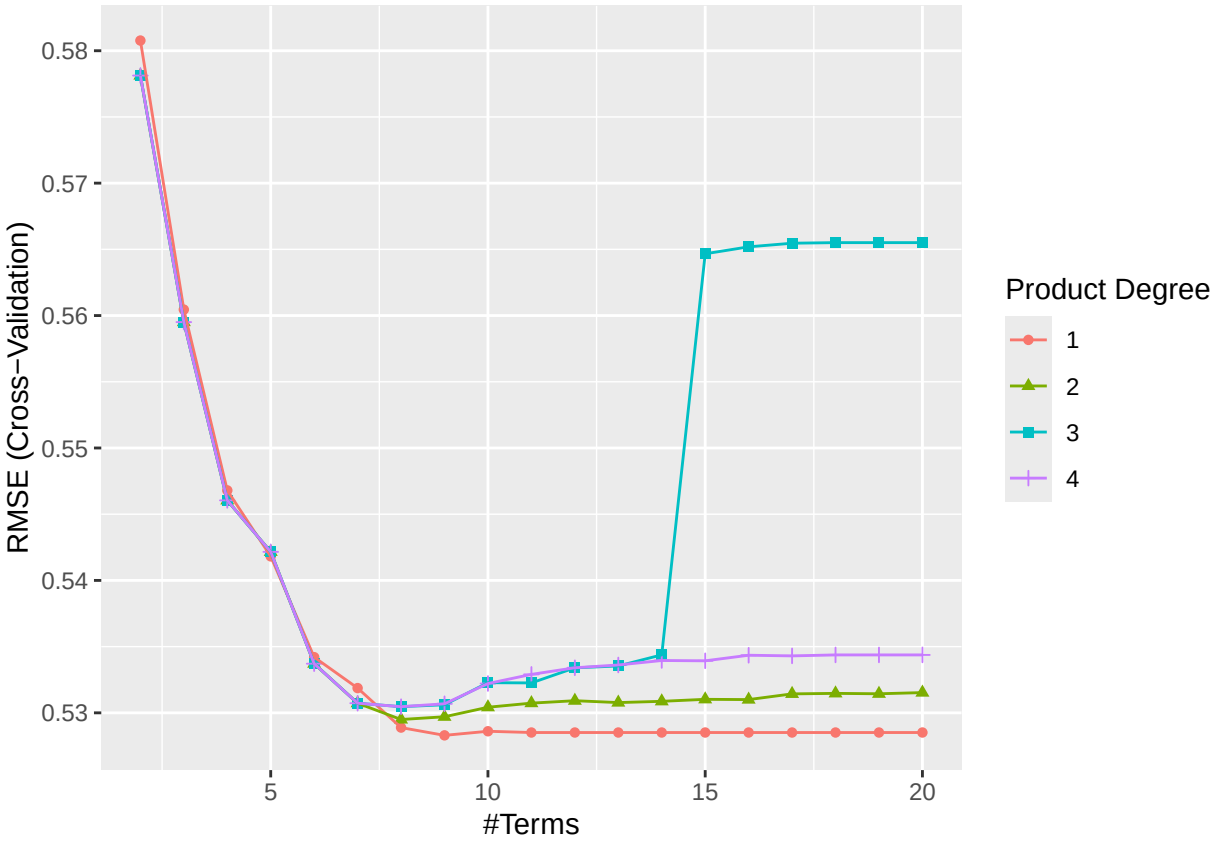



Figure 10: MARS Tuning Grid Selection

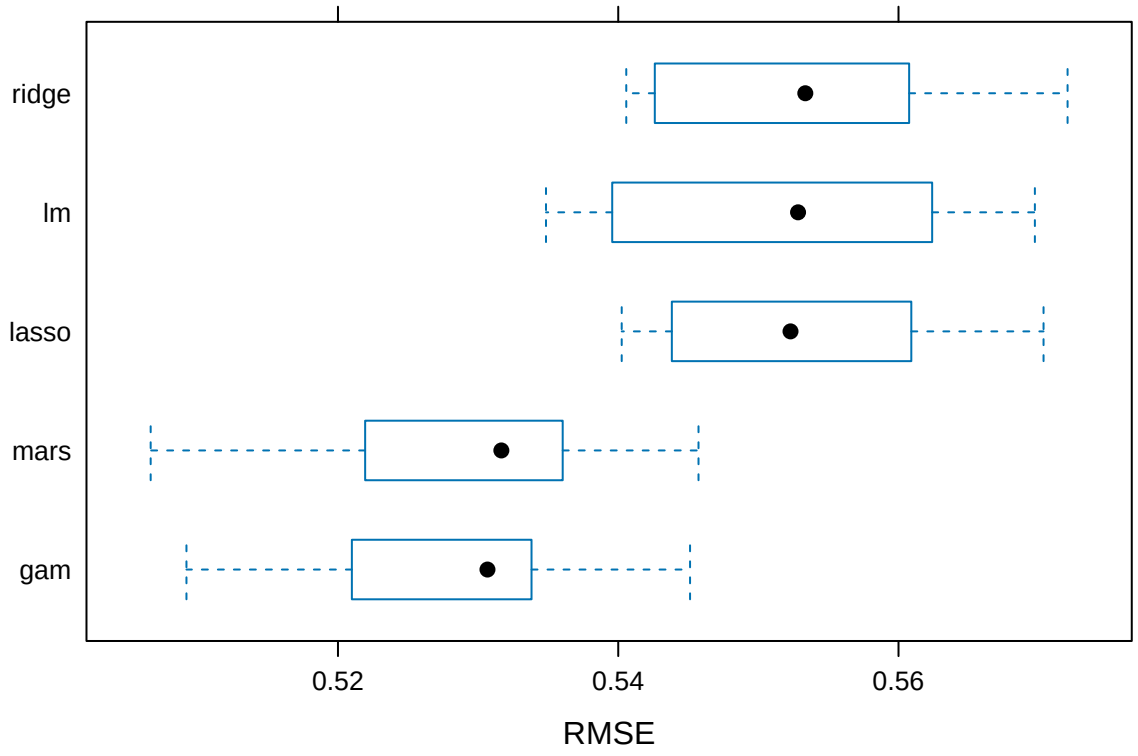



Figure 11: Model Comparison (RMSE)

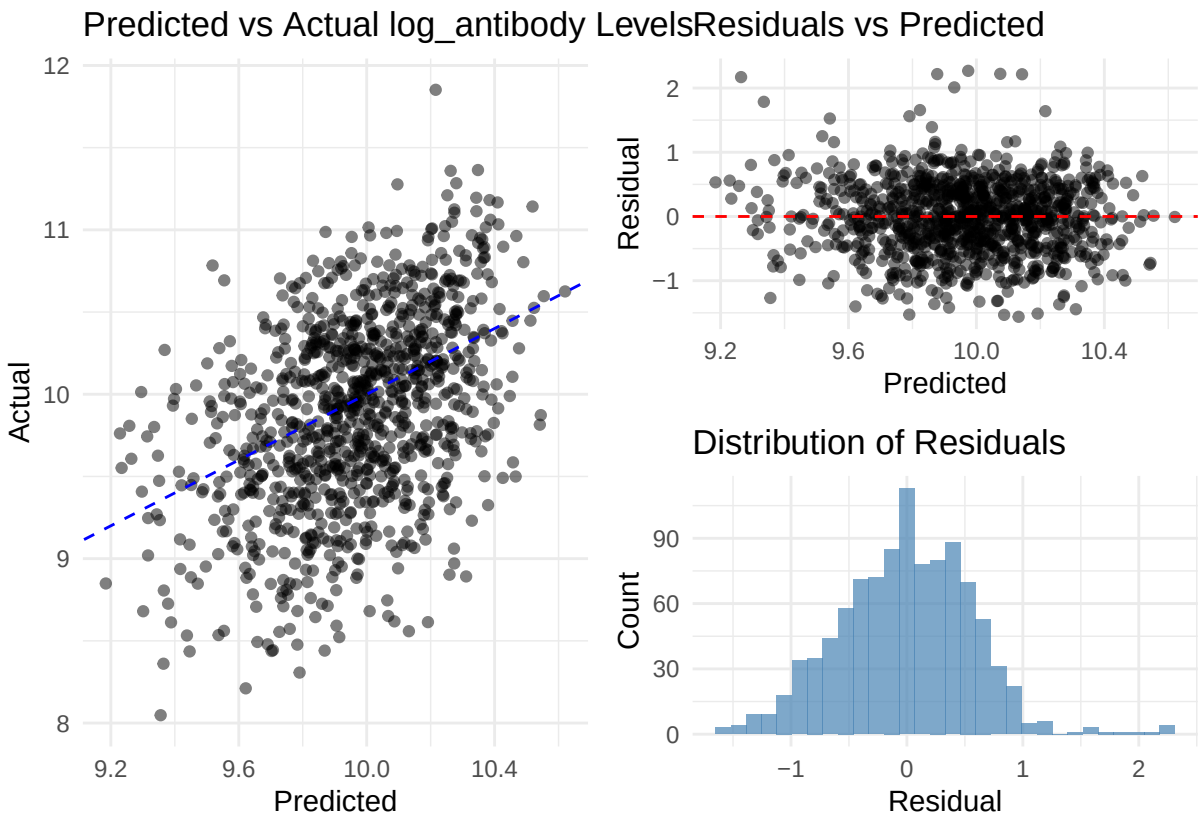



Figure 12: Diagnostic Plots for Prediction Model on dat2