

# Data Science 2 Midterm

Jasmin Martinez (JRM2319), Ixtaccihuatl Obregon, Elliot Kim

03/24/2025

## Contents

1 Exploratory Analysis	1
2 Model Trainging	2
3 Final Model Results	2

## 1 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 1 and Figure 2. High and low outliers can be observed in the Figure 3. 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 4.

Exploring **dat2**, the distribution of **log\_antibody** is also approximately normal, as shown Figure 5 and Figure 6. High and low outliers are again visible in the Figure 7. 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 8.

The exploration and evaluation of **dat1** is used to help build a prediction model specifically 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.

## 2 Model Trainging

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`), ( $pval=<0.001$ ).

## 3 Final Model Results

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 11 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

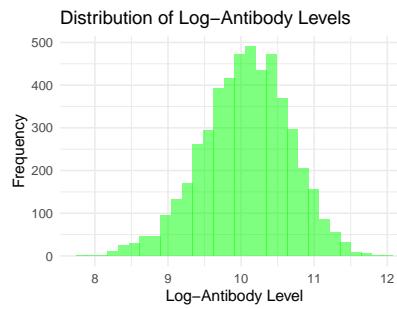
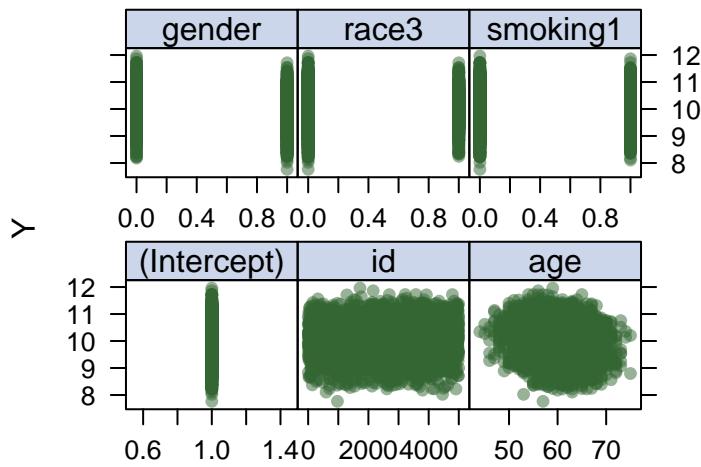


Figure 1: Histogram of log\_antibody levels for dat1



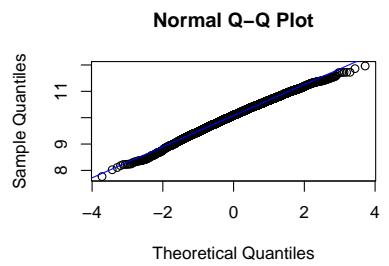


Figure 2: Q-Q plot of log\_antibody levels for dat1

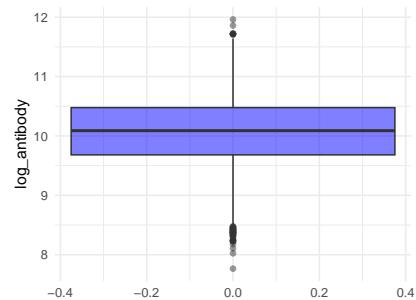


Figure 3: Boxplot of log\_antibody levels for dat1

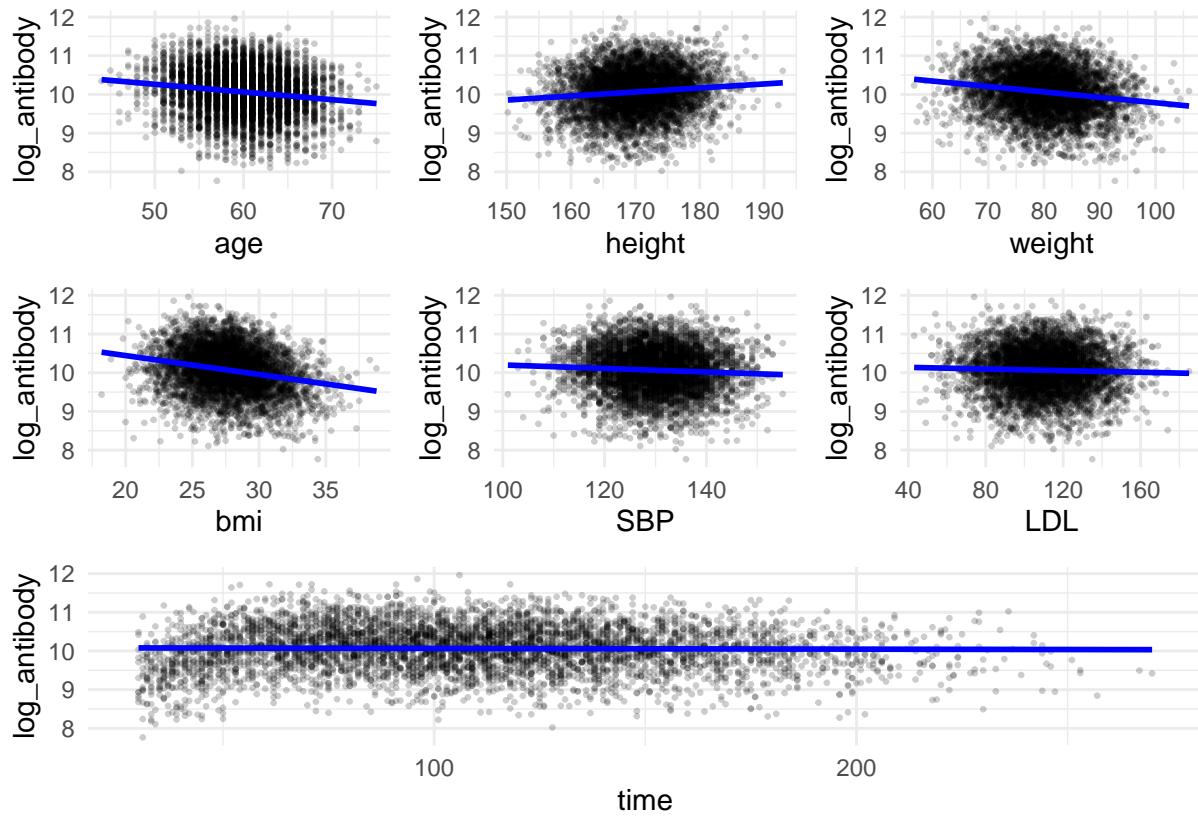


Figure 4: Linearity between log\_antibody levels and covariates for dat1

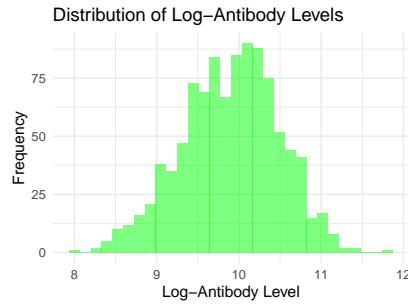


Figure 5: Histogram of log\_antibody levels for dat2

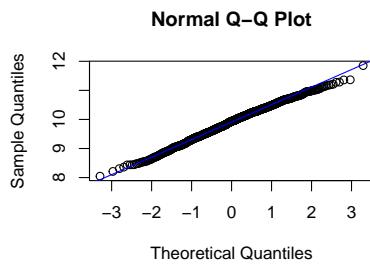


Figure 6: Q-Q plot of log\_antibody levels for dat2

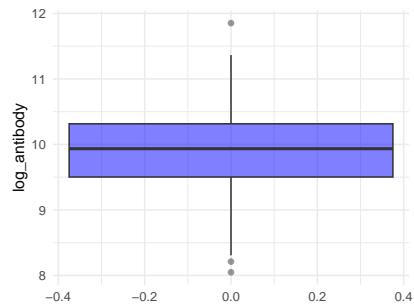


Figure 7: Boxplot of log\_antibody levels for dat2

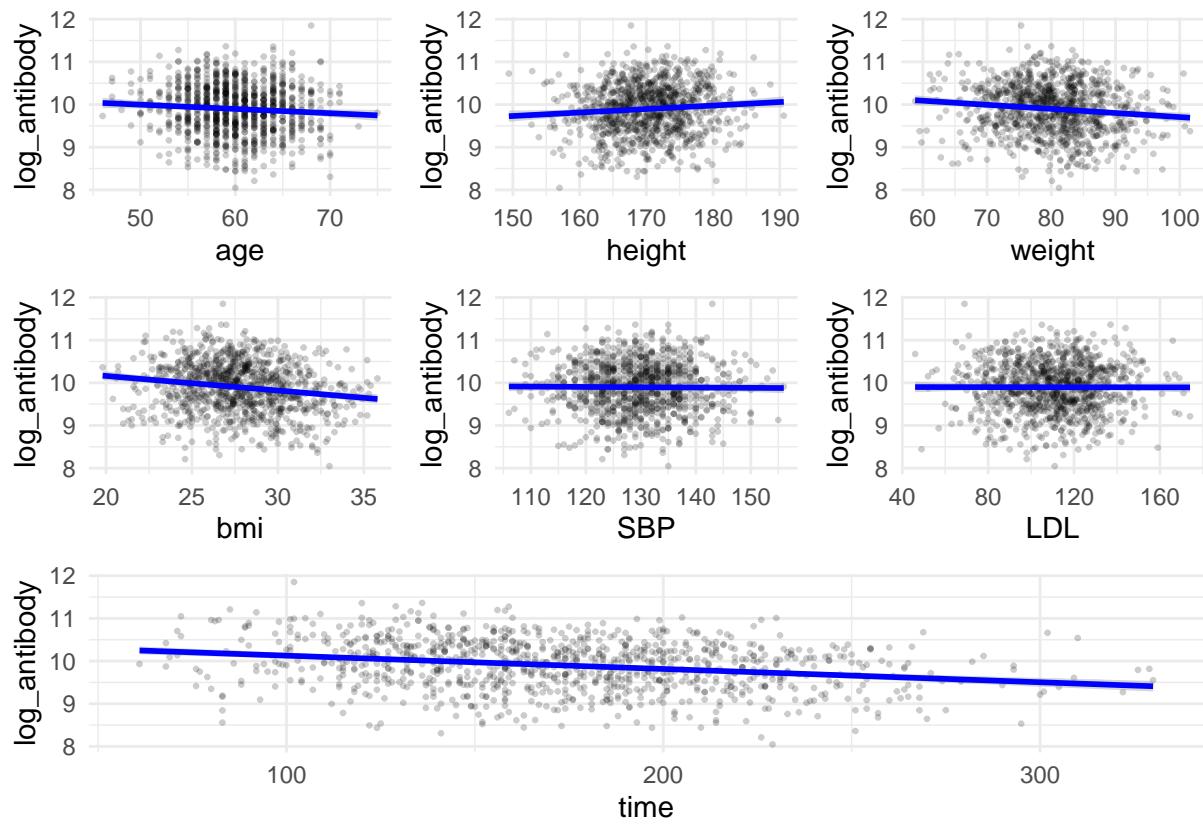
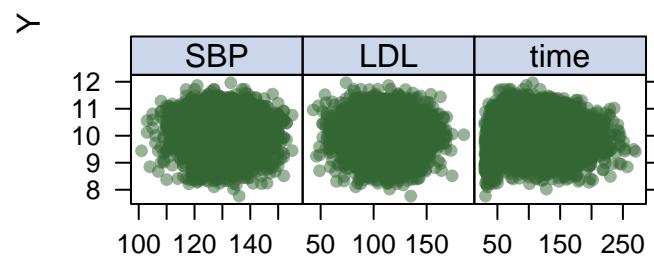
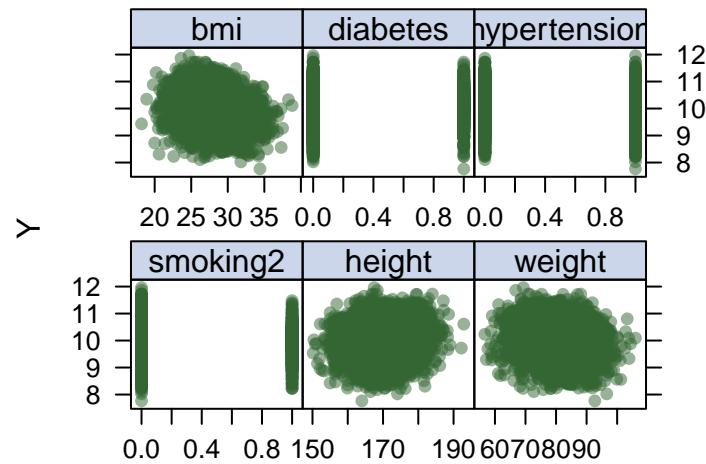


Figure 8: Linearity between log\_antibody levels and covariates for dat2

Table 2: 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



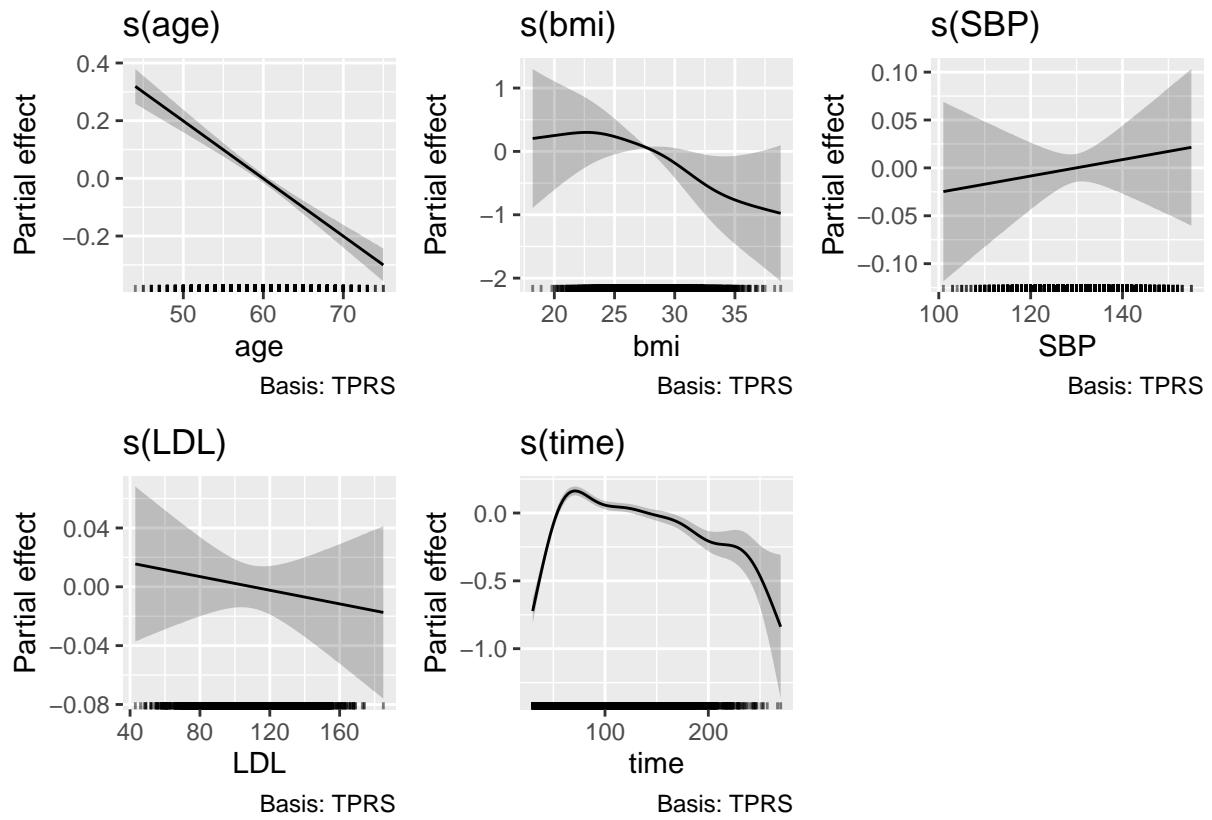


Figure 9: Smooth terms for GAM model (gam.m2)

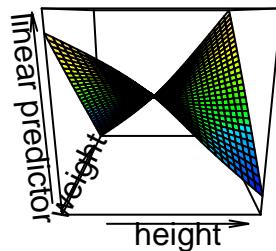


Figure 10: Visualization of Model 3 (dat1)

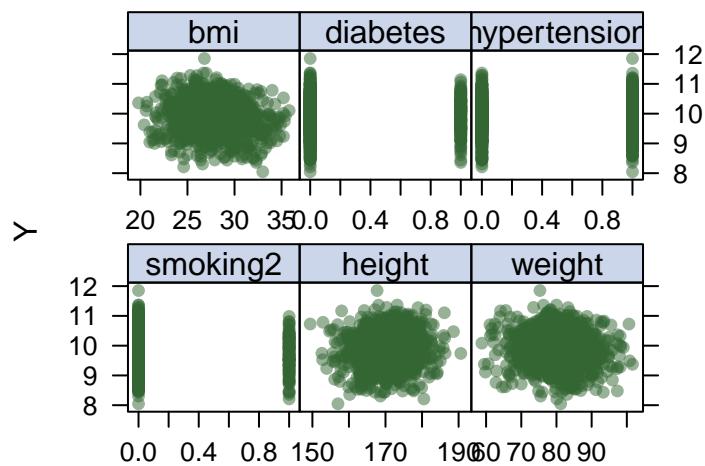
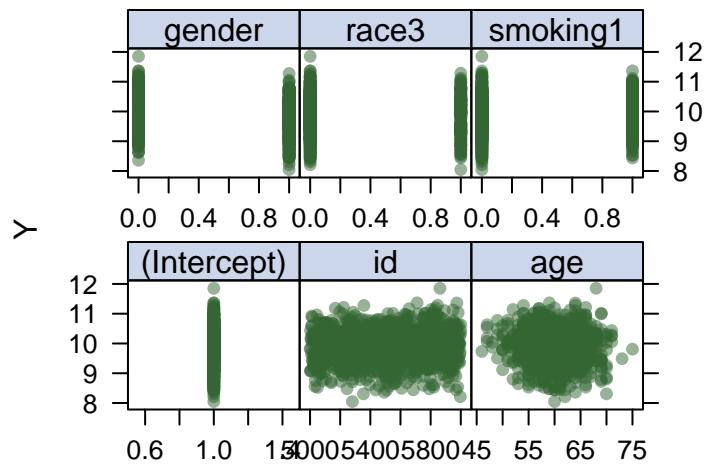
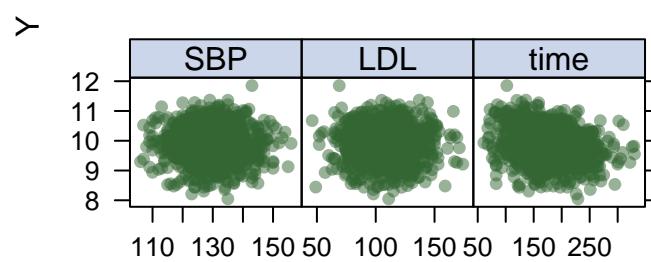
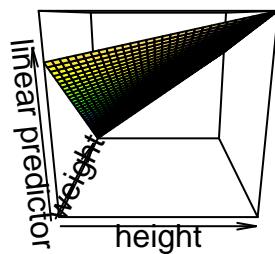
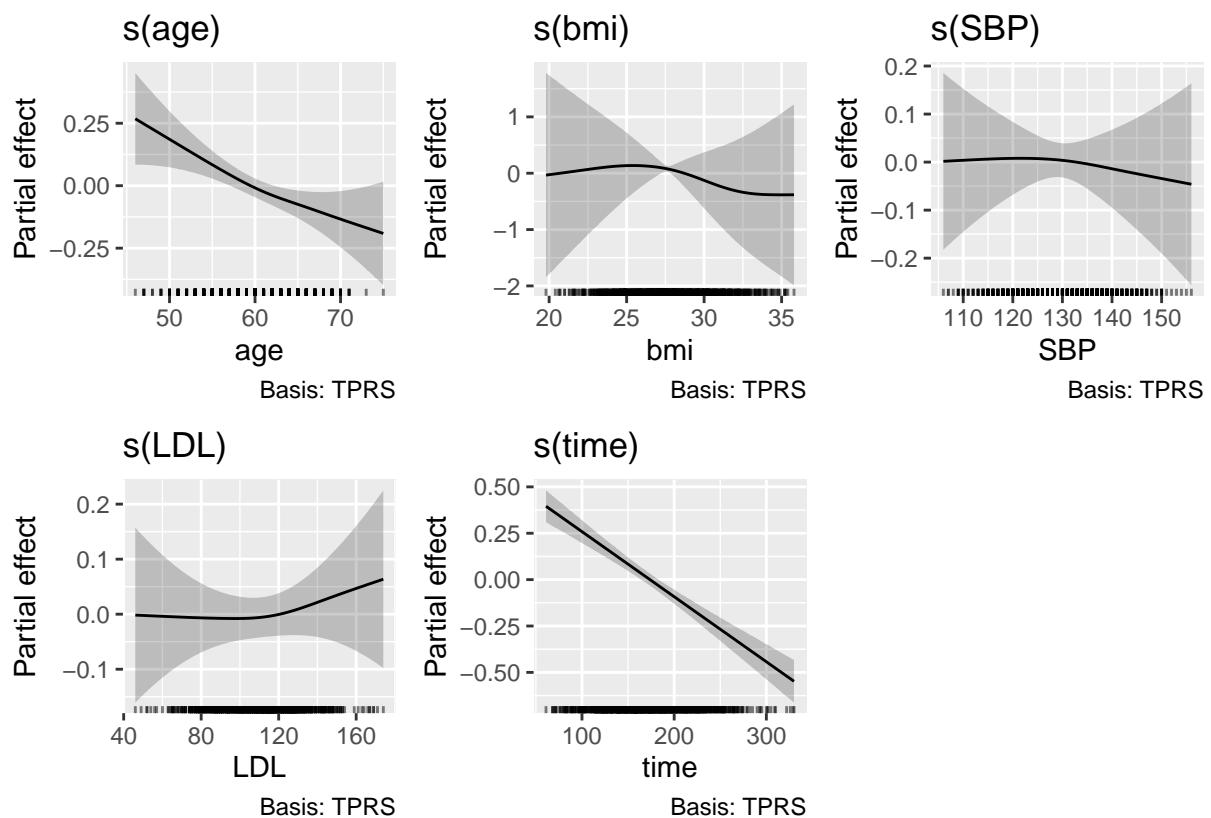


Table 3: ANOVA Comparison of GAM Models (Dat2)

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





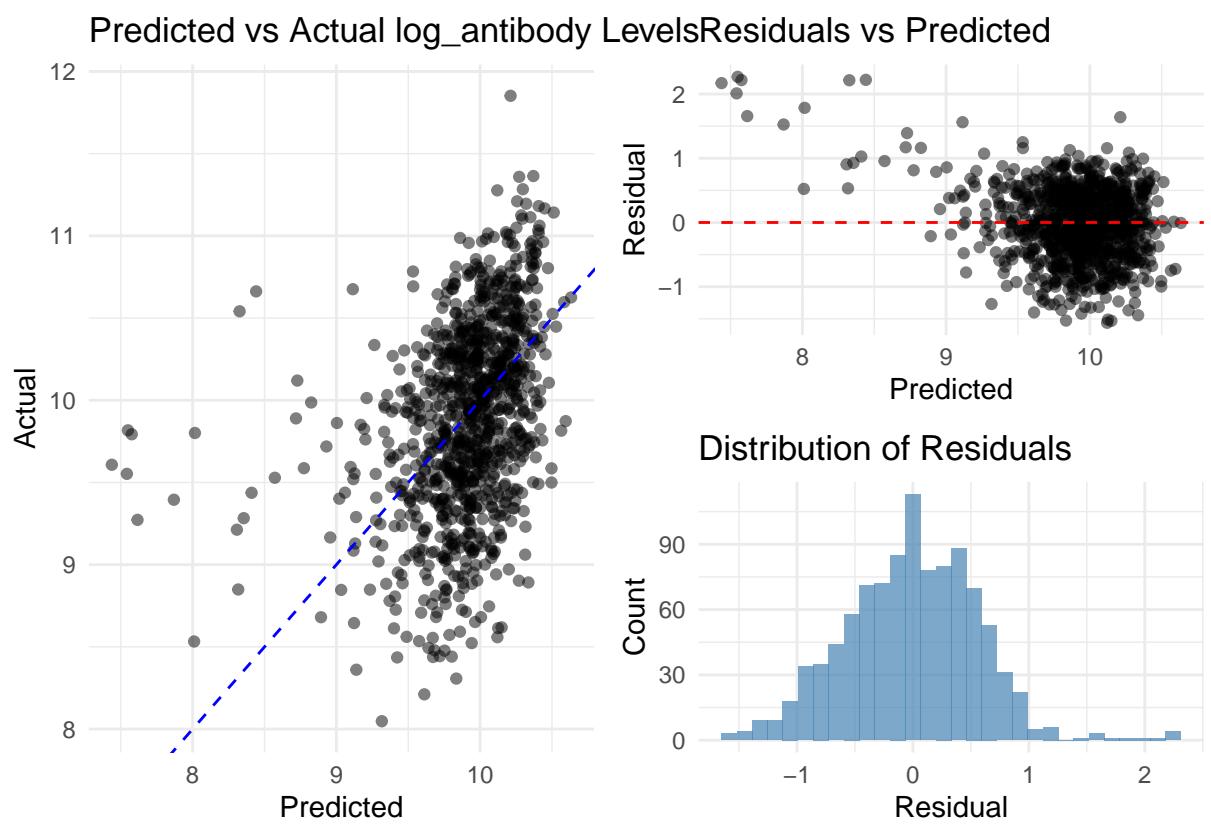


Figure 11: Diagnostic Plots for Prediction Model on dat2