Midterm Project

A researcher conducted a study to evaluate antibody responses to a newly authorized vaccine. Understanding antibody responses can help inform vaccine effectiveness and identify populations with weaker immune responses. The antibody level measured via dry blood spot (log-transformed) is the primary outcome of interest. The researcher also collected data on potential predictors to understand factors influencing the antibody response. Here is a description of the dataset (**dat1.RData**):

| **Variable Name (Column Name)** | **Description** |
| --- | --- |
| ID (id) | Participant ID |
| Age (age) | Age (in years) |
| Gender (gender) | 1 = Male, 0 = Female |
| Race/ethnicity (race) | 1 = White, 2 = Asian, 3 = Black, 4 = Hispanic |
| Smoking (smoking) | 0 = Never smoked, 1 = Former smoker, 2 = Current smoker |
| Height (height) | Height (in centimeters) |
| Weight (weight) | Weight (in kilograms) |
| BMI (bmi) | Body Mass Index; BMI = weight (kg) / [height (m)]² |
| Diabetes (diabetes) | 0 = No, 1 = Yes |
| Hypertension (hypertension) | 0 = No, 1 = Yes |
| Systolic blood pressure (SBP) | Systolic blood pressure (mmHg) |
| LDL cholesterol (LDL) | LDL cholesterol (mg/dL) |
| Time since vaccination (time) | Time since vaccination (in days) |
| Log-transformed antibody level (log\_antibody) | Response variable |

Using this dataset, the researcher planned to build a prediction model of antibody levels, aiming to understand how demographic and clinical factors influence antibody responses and how antibody levels change over time following vaccination. Please help the researcher build this model.

A few months after the initial data collection, the researcher collected a new, independent dataset (**dat2.RData**) with the same structure as **dat1.RData**. Evaluate the robustness and generalizability of your prediction model using the new dataset. Does the model generalize well to this new dataset? If the model does not generalize well, discuss possible reasons.

[dat1.RData](https://courseworks2.columbia.edu/courses/217398/files/23611022?wrap=1)[Download dat1.RData](https://courseworks2.columbia.edu/courses/217398/files/23611022/download?download_frd=1)

[dat2.RData](https://courseworks2.columbia.edu/courses/217398/files/23611028?wrap=1)[Download dat2.RData](https://courseworks2.columbia.edu/courses/217398/files/23611028/download?download_frd=1)

Please complete a report that helps the researcher to answer the above questions of interest. **Your report should at least include the following sections:**

**Exploratory analysis**

Use summary statistics and appropriate visualization techniques to explore the data.

**Model training**

Provide a detailed description of the model training procedure and how you obtained the final model. **Your description should include sufficient detail so another statistician can reproduce the same model.** Points will be deducted if details provided are insufficient to reproduce your results.

**Results**

Report the final model. Interpret the model and assess its performance. Summarize your findings from the analysis for a general audience.

**Your submission should include two components:**

1. **Analysis Report (PDF or Word Document):**
   * Your report should succinctly summarize your analysis and findings within a 3-page limit, **excluding** figures and tables.
   * Do not include R code in the report.
   * Avoid inserting raw outputs from R functions; instead, demonstrate your ability to interpret and distill useful information from output tables.
2. **R code and outputs (rmd + knitted file):**
   * Your submissions should contain **all** the code for conducting your analysis, as well as for the creation of any tables/figures/numbers presented in your report.

**Due to submission restrictions, one team member will submit all the files for the entire team. Please coordinate within your team to decide who will be responsible for this task.**