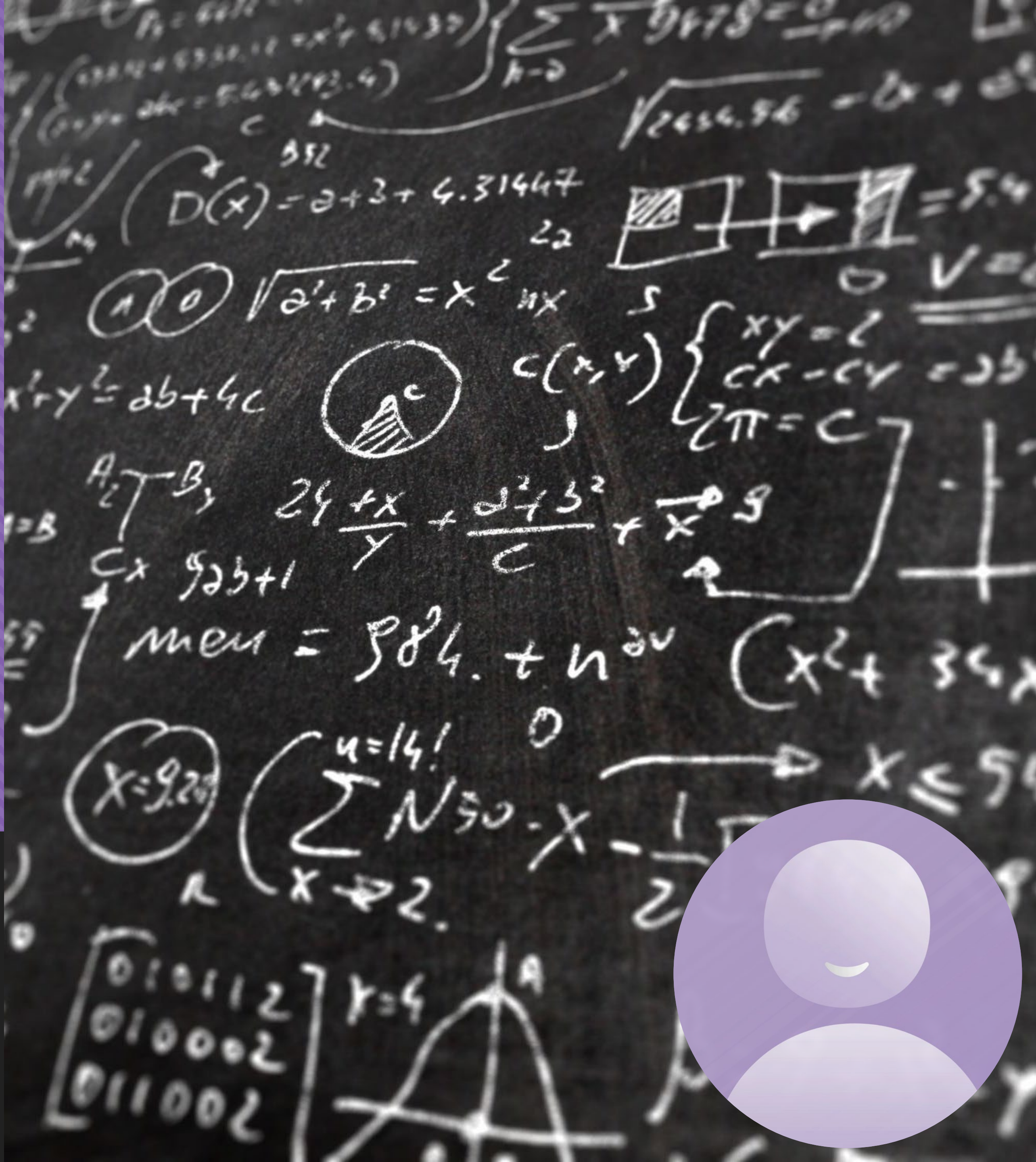
Linear Regression: A Simple Machine Learning Model

Kathleen Ashbaker

HMS 520 Autumn 2023

UNIVERSITY OF WASHINGTON



Introduction

- Hello everyone, I'm Kathleen Ashbaker, and this is my final project for HMS 520, Autumn 2023. Today, I'll be presenting my project titled "Linear Regression: A Simple Machine Learning Model."

SCOPE OF PROJECT

- The aim of this project is to provide a comprehensive guide to simple linear regression using R.
- This technique is a fundamental aspect of machine learning and this guide is suitable for learners and researchers at all levels, whether you're a novice or an advanced user.

LOAD LIBRARIES

- Let's begin with Step 0, where we load essential R libraries like 'readr', 'tidyverse', 'mice', 'dplyr', and 'ggplot2'. These libraries are crucial for data manipulation and visualization.

UPLOAD DATA AND STORE IN DATA FRAME

- Moving to Step 1, we upload our data. Here, we're using a CSV file named "Patient_List_aaliyah_washington_md". It's important to note that we're using 'show_col_types = FALSE' to simplify our data frame's readability.
- This masks the object type in the column name, i.e. ‘double’

DATA EXAMINATION AND CLEANING

- In Step 2, we focus on examining and cleaning the data. Using 'dplyr' and the 'mutate_all()' function, we replace 'None' values with NA. We then perform a summary to get an overview of our data. This step is crucial for ensuring the quality and accuracy of our analysis.

DATA CHECKING AND IMPUTATION

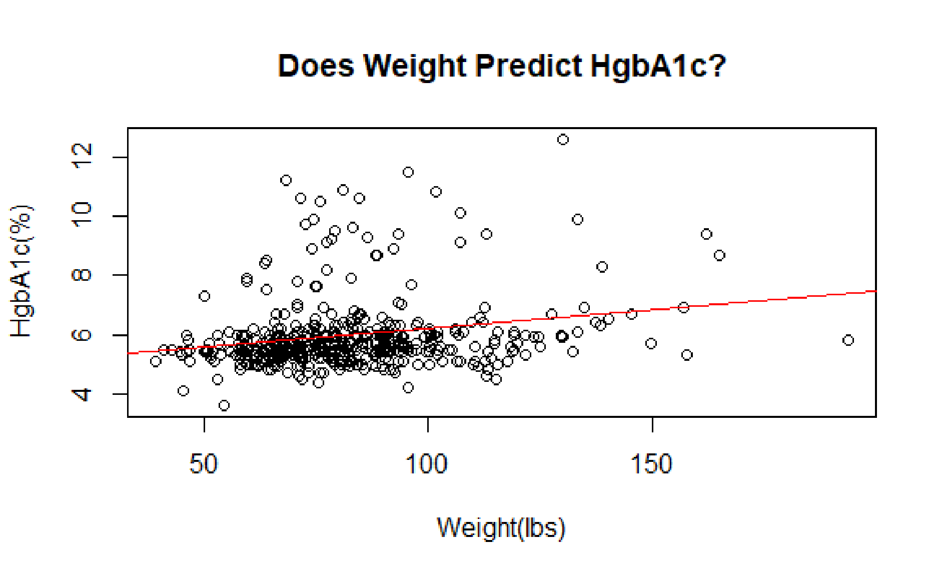
- Next, we explore optional methods to check specific column values and count 'NA' values, using functions like 'is.numeric' and 'impute_data'. This helps us handle missing data effectively, a common challenge in data analysis.
- ‘mice’ is the essential library for this

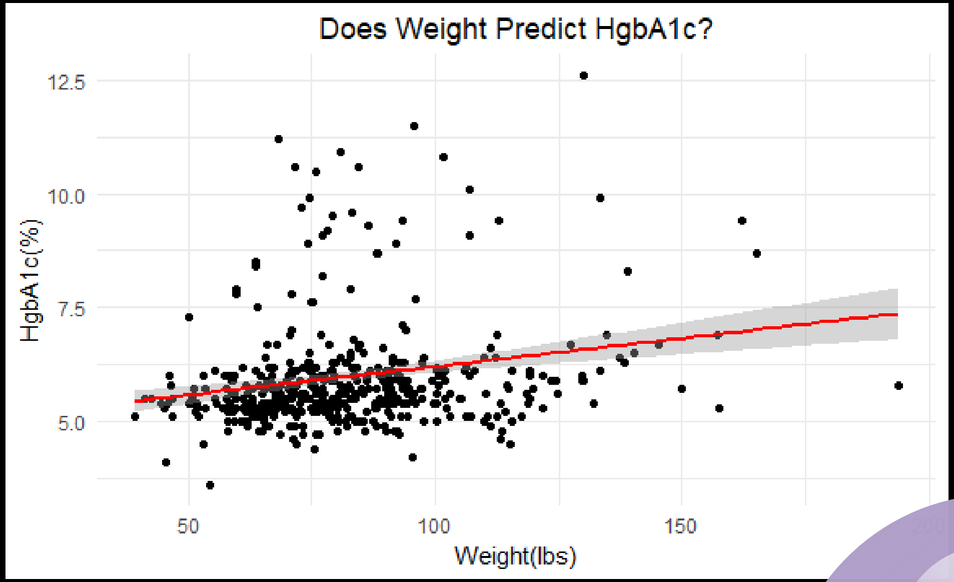
DATA VISUALIZATION

- In Step 3, we dive into data visualization. Using 'ggplot2' and base R plotting functions, we create scatter plots to visually examine the relationship between variables like weight and HgbA1c levels.

- We also compare visualizations(only clean data shown) and summary outputs between the raw and cleaned data, which highlights the importance of data cleaning in revealing accurate trends and patterns.

COMPARATIVE VISUALIZATION OF WEIGHT AND HGBA1C USING BASE AND GGPLOT2





LINEAR REGRESSION ANALYSIS

- Finally, we conduct a simple linear regression analysis, first with the original dataset and then with the cleaned data. We use the 'lm' function in R for this purpose and then summarize the output to interpret our model's findings.

LINEAR REGRESSION ANALYSIS

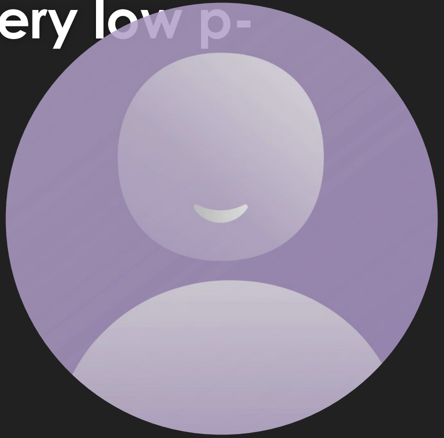
- "In this part of our analysis, we summarized the results of our linear regression model, where we tried to understand the relationship between HgbA1c levels and weight using our cleaned data set.

- The summary provides us with a few key insights. First, it shows the range of residuals, which are the differences between the observed values and the values predicted by our model. In our case, these residuals vary from about -2 to +6, indicating how far off some predictions might be.

LINEAR REGRESSION ANALYSIS

- Then, we look at the coefficients, which tell us how much HgbA1c levels change with a unit change in weight. The coefficient for weight is positive, suggesting a slight increase in HgbA1c levels with an increase in weight.

- The 't value' and 'Pr(>|t|)' are statistical measures indicating how significant these coefficients are. In our model, both the intercept and the weight have very low p-values, indicating that these findings are statistically significant.



LINEAR REGRESSION ANALYSIS

- Lastly, the model gives us the 'R-squared' value, around 0.04, which tells us how well our model explains the variability in HgbA1c levels. While this number is not very high, it's common in real-world data, where many factors can influence the outcome.

- In summary, this model gives us a basic understanding of the relationship between weight and HgbA1c levels, with statistical significance, though it explains a relatively small portion of the variation in HgbA1c levels."

CONCLUSION, REPO LINK, AND ACKNOWLEDGEMENTS

- **In conclusion, this project not only demonstrates the application of simple linear regression in R but also underscores the significance of data preparation and visualization in machine learning.**
- **Kathleen’s Repo Link here!!!!!!! :**
  - **https://github.com/QueenKatherys/HMS-Autumn-2023-Final-Project-**
- THANK YOU FOR WATCHING !!!!!!!