Mercedes Joseph

D208

Professor Middleton

Linear Regression Modeling Part 1

June 27, 2023

Part 1: Research Question

A.  Describe the purpose of this data analysis by doing the following:

**A1. Research Question:**

***How do medical conditions and other factors influence the number of initial days required for treatment?***

**A2. Define the goals of the data analysis.**

**The goal of my research assignment is to determine which variables influence the number of initial days required for patient treatment**. I have chosen independent variables that consist of medical conditions and additional factors, including HighBlood, Stroke, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Anxiety, Allergic\_rhinitis, Reflux\_esophagitis, Asthma, Income, Total charge, and additional charges. By utilizing a linear regression model, I aim to identify the variables that have an impact on the number of initial days patients require for their treatment. This research will ultimately assist the hospital in understanding the underlying causes of initial visits and contribute to the development of a more efficient healthcare system.

Part II: Method Justification

B. Describe multiple linear regression methods by doing the following:

B1.

**I chose to summarize the four assumptions Linearity, Multicollinearity, Homoscedasticity, and Normality of Residuals.**

Linearity this assumption states that the relationship between the predictor variables and the outcome variable is linear. It assumes that the change in the outcome variable is directly proportional to the change in the predictor variables.

Multicollinearity occurs when there is a high correlation between two or more predictor variables in the regression model. This assumption assumes that the predictor variables are not highly correlated, as it can lead to unstable estimates and difficulties in interpreting the individual effects of the variables.

Homoscedasticity assumes that the variability of the residuals is constant across all levels of the predictor variables. Basically, meaning that the spread of the residuals is similar throughout the range of the predictor variables.

Normality of Residuals states that the residuals in the linear regression model follow a normal distribution. It means that the errors or residuals should be symmetrically distributed around zero and not significantly deviate from a bell-shaped curve. Departure from normality can affect the accuracy of statistical inference and the reliability of the model.

***It is important for analysts to assess and satisfy these assumptions to ensure the validity and reliability of the linear regression analysis. Violations of these assumptions may require additional analysis or transformations of the variables to meet the assumptions or alternative regression models to be considered.***

B2.

Programming language R provides user-friendly tools and packages that make it easier for me to clean and prepare my data. I can efficiently organize and modify the data, handle missing values, and create new variables. This ensures that my data is in the right format for analysis, saving me time and effort.

R also provides libraries that simplify the process of building and evaluating multiple linear regression models. These libraries provide functions and tools specifically designed for estimating the relationships between variables, checking model fit, and identifying any potential issues. With R, I can easily interpret the results of my regression analysis and present them visually. Utilizing R for multiple linear regression empowers me to conduct reliable and informative analyses. It allows me to gain insights into the relationships between variables and aids in interpreting and communicating the findings effectively.

**The packages used:**

**A screen shot of a computer program

Description automatically generated with low confidence**

I used the `tidyverse` package because it has useful tools for visualizing and manipulating data, which are important in linear regression modeling. I specifically used the `ggplot2` package for creating visualizations and the `dplyr` package for cleaning and organizing the data. The `stats` package is a core package in R that provides functions for performing linear regression analysis. It helps me fit the regression model and obtain important information about the model's accuracy. I also used the `vcd` package to handle categorical variables in the regression model. This package helps me understand the relationships between these variables and the outcome I'm trying to predict. Lastly, I loaded the `gmodels` package, which gives me additional tools for summarizing and diagnosing the regression model.

B3.

In my research, I have found that multiple linear regression is a valuable method for investigating the factors contributing to the number of initial days required for patient treatment. By utilizing this approach, I can simultaneously examine multiple variables, such as various medical conditions and additional factors, to determine their individual impact on the number of initial days, while controlling for the effects of other factors. The regression coefficients obtained from the analysis provide measures of the relationships between these variables and the number of initial days. This enables me to gain insights into the direction and significance of each variable's effect.

Multiple linear regression helps me identify any potential confounding variables and detect possible interactions among the studied factors. By understanding the complex relationships between medical conditions, additional factors, and the number of initial days, I can make informed decisions and develop targeted interventions to optimize patient treatment and reduce the length of initial hospital stays. This knowledge contributes to the development of a more efficient healthcare system by improving resources and patient care planning.

Part III: Data Preparation

C. Summarize the data preparation process for multiple linear regression analysis by doing the following:

C1.

***Data cleaning goals:***

***The goals of data cleaning for my D208 assignment were to:***

1. Convert categorical data: Decide whether to change the categorical information about medical conditions into numbers. I looked at each variable and determined what made sense for the analysis.
2. Check for duplicate columns: Make sure there weren't any columns that had the same information. This would avoid repeating data and keep everything accurate.
3. Rename columns: Give better names to columns labeled as "Item 1-8" based on the provided scenario. This would make it easier to understand what each item represents.
4. Fill in missing values: Identify which columns had missing information and choose the best way to fill in those gaps. For some types of information, I used the middle value, while for others, I used the most common option or the average value.
5. Visualize the data: Create different types of graphs to better understand how the information is spread out and how different variables relate to each other.
6. Handle outliers: Look for unusual values in the data and decide if they were mistakes or if they had a valid reason for being there. Based on that, I decided whether to keep or remove them.

**By setting these goals, I made sure the data was of quality ready for analysis.**

***Steps for data cleaning:***

For my D208 assignment, I chose to utilize a CSV file that had already undergone data cleaning and preparation from my D206 assignment. Throughout the cleaning process, I encountered a few challenges. One major issue was deciding whether to convert the categorical data of the medical condition variables into numeric format. To address this, I carefully examined the data and decided based on the nature of each variable. Another task involved checking for possible duplicate columns, but fortunately, I did not find any duplicates. I also renamed Item 1- 8 the following based on the information from the scenario:

**- Item1: Timely admission**

**- Item2: Timely treatment**

**- Item3: Timely visits**

**- Item4: Reliability**

**- Item5: Options**

**- Item6: Hours of treatment**

**- Item7: Courteous staff**

**- Item8: Evidence of active listening from doctor**

I then identified the columns that contained missing values and made appropriate choices for filling them. For normal data, I used the median to fill missing values, while for categorical data, I used the mode, and for continuous variables, I used the mean. This helped ensure that the dataset was complete and ready for analysis.

To gain a better understanding of the data, I created various visual models, including boxplots, scatterplots, and histograms. These visualizations provided insights into the distribution and relationships between variables. During this process, I also examined outliers in the data. However, upon further analysis, I determined that the outliers accounted for only 0.18% of the dataset and were primarily related to survey-type questions based on patients' experiences, rather than data errors. As a result, I concluded that removing these outliers was not necessary.

Overall, I successfully addressed the challenges encountered during data cleaning and preparation for statistical modeling. By making informed decisions regarding data conversion, filling missing values, and handling outliers, I ensured the dataset's quality and suitability for further analysis.

***I have attached a Rscript that holds all the code for the data cleaning process.***

***File name: “file done.R”.***

C2. Describe the dependent variable and all independent variables using summary statistics that are required to answer the research question, including a screenshot of the summary statistics output for each of these variables.

***Code to Perform Summary Statistics:***



***Summary statistics for all Variables:***

A screenshot of a computer screen

Description automatically generated with low confidence

Results from Summary Statistics:

***Independent Variables:***

***HighBlood:***

- Minimum value: 0.000

- 1st quartile: 0.000

- Median: 0.000

- Mean: 0.409

- 3rd quartile: 1.000

- Maximum value: 1.000

High Blood’s variable represents whether a person has high blood pressure or not. The summary statistics indicate that about 40.9% of the population in the dataset has high blood pressure.

***Stroke:***

- Minimum value: 0.000

- 1st quartile: 0.000

- Median: 0.000

- Mean: 0.1993

- 3rd quartile: 0.000

- Maximum value: 1.000

Stroke’s variable indicates whether a person has had a stroke or not. The summary statistics show that approximately 19.9% of the individuals in the dataset have experienced a stroke.

***Overweight:***

- Minimum value: 0.0000

- 1st quartile: 0.0000

- Median: 1.0000

- Mean: 0.7377

- 3rd quartile: 1.0000

- Maximum value: 1.0000

The "Overweight" variable represents whether a person is overweight or not. The summary statistics reveal that around 73.8% of the population in the dataset is classified as overweight.

***Arthritis:***

- Minimum value: 0.0000

- 1st quartile: 0.0000

- Median: 0.0000

- Mean: 0.3574

- 3rd quartile: 1.0000

- Maximum value: 1.0000

Arthritis’ variable indicates whether a person has arthritis or not. The summary statistics suggest that about 35.7% of the individuals in the dataset have arthritis.

***Diabetes:***

- Minimum value: 0.0000

- 1st quartile: 0.0000

- Median: 0.0000

- Mean: 0.2738

- 3rd quartile: 1.0000

- Maximum value: 1.0000

The "Diabetes" variable represents whether a person has diabetes or not. The summary statistics show that approximately 27.4% of the population in the dataset has diabetes.

***Hyperlipidemia:***

- Minimum value: 0.0000

- 1st quartile: 0.0000

- Median: 0.0000

- Mean: 0.3372

- 3rd quartile: 1.0000

- Maximum value: 1.0000

Hyperlipidemia’s variable indicates whether a person has hyperlipidemia (high cholesterol) or not. The summary statistics suggest that about 33.7% of the individuals in the dataset have hyperlipidemia.

***BackPain:***

- Minimum value: 0.0000

- 1st quartile: 0.0000

- Median: 0.0000

- Mean: 0.4114

- 3rd quartile: 1.0000

- Maximum value: 1.0000

The "BackPain" variable represents whether a person experiences back pain or not. The summary statistics indicate that approximately 41.1% of the population in the dataset has back pain.

***Anxiety:***

- Minimum value: 0.0000

- 1st quartile: 0.0000

- Median: 0.0000

- Mean: 0.2906

- 3rd quartile: 1.0000

- Maximum value: 1.0000

Anxiety’s variable indicates whether a person experiences anxiety or not. The summary statistics suggest that about 29.1% of the individuals in the dataset have anxiety.

***Allergic\_rhinitis:***

- Minimum value: 0.0000

- 1st quartile: 0.0000

- Median: 0.0000

- Mean: 0.3941

- 3rd quartile: 1.0000

- Maximum value: 1.0000

The "Allergic\_rhinitis" variable represents whether a person has allergic rhinitis (hay fever) or not. The summary statistics indicate that approximately 39.4% of the population in the dataset has allergic rhinitis.

***Reflux\_esophagitis:***

- Minimum value: 0.0000

- 1st quartile: 0.0000

- Median: 0.0000

- Mean: 0.4135

- 3rd quartile: 1.0000

- Maximum value: 1.0000

Reflux Esophagitis’s variable indicates whether a person has reflux esophagitis (acid reflux) or not. The summary statistics suggest that about 41.4% of the individuals in the dataset have reflux esophagitis.

***Asthma:***

- Minimum value: 0.0000

- 1st quartile: 0.0000

- Median: 0.0000

- Mean: 0.2893

- 3rd quartile: 1.0000

- Maximum value: 1.0000

The "Asthma" variable represents whether a person has asthma or not. The summary statistics indicate that approximately 28.9% of the population in the dataset has asthma.

***Income:***

- Minimum value: 154.1

- 1st quartile: 23956.2

- Median: 33942.3

- Mean: 38872.4

- 3rd quartile: 46466.8

- Maximum value: 207249.1

Income’s variable represents the income of individuals. The summary statistics show the distribution of income in the dataset, with the mean income being approximately $38,872.4.

***TotalCharge:***

- Minimum value: 1257

- 1st quartile: 3253

- Median: 5852

- Mean: 5892

- 3rd quartile: 7615

- Maximum value: 21524

The "TotalCharge" variable represents the total charge for medical treatment. The summary statistics show the distribution of total charges in the dataset, with the mean total charge being approximately $5,892.

***Additional\_charges:***

- Minimum value: 3126

- 1st quartile: 7986

- Median: 11574

- Mean: 12935

- 3rd quartile: 15626

- Maximum value: 30566

Additional Charge’s variable represents additional charges for medical treatment. The summary statistics show the distribution of additional charges in the dataset, with the mean additional charge being approximately $12,935.

***Dependent Variable:***

***Initial\_days:***

- Minimum value: 1.002

- 1st quartile: 8.929

- Median: 34.432

- Mean: 34.432

- 3rd quartile: 59.460

- Maximum value: 71.981

The "Initial\_days" variable represents the number of initial days of hospitalization. The summary statistics show the distribution of initial hospitalization days in the dataset, with the mean duration being approximately 34.432 days.

**Brief Overview of Summary Statistics:**

Summary statistics provide a way to summarize important information about a dataset without having to look at every single data point. They give us a quick overview of the data and help us understand its main characteristics. Some common summary statistics include the mean (average), which tells us the typical value of the data, the median, which represents the middle value, and the mode, which is the most frequent value. These measures help us understand the central tendency of the data.

Other summary statistics include the standard deviation, which tells us how spread out the data points are from the mean, and the minimum and maximum values, which give us an idea of the range of the data. Percentiles are also useful summary statistics. They divide the data into equal parts, such as the 25th percentile, which represents the value below which 25% of the data falls. By looking at summary statistics, we can quickly get an idea of the overall pattern and variability in the data. This information allows us to make comparisons, identify patterns, and gain initial insights before delving into more detailed analysis. I

C3.

Univariate

A picture containing diagram, screenshot, text, line

Description automatically generated

#Histogram for Initial\_ days

A picture containing diagram, text, screenshot, line

Description automatically generated

#Histogram for Income

A picture containing text, diagram, screenshot, design

Description automatically generated

#Histogram Total Charge

A picture containing diagram, screenshot, text, plot

Description automatically generated

#Histogram Additional Charge

A picture containing graphics, diagram, font, graphic design

Description automatically generated

#Pie Chart for Asthma

A pie chart of high blood

Description automatically generated with medium confidence

Pie chart for HighBlood

A red circle with a blue triangle

Description automatically generated with medium confidence

Pie chart for Stroke

A pie chart of overweight

Description automatically generated with medium confidence

Pie chart Overweight

A red and blue pie chart

Description automatically generated with medium confidence

Pie Chart Arthritis

A red and blue pie chart

Description automatically generated with medium confidence

Pie chart Diabetes

A red and blue pie chart

Description automatically generated with medium confidence

Pie chart for Hyperlipidemia

A red and blue pie chart

Description automatically generated with medium confidence

Pie chart Back pain

A picture containing text, diagram, graphics, screenshot

Description automatically generated

Pie chart for Anxiety

A red and blue pie chart

Description automatically generated with medium confidence

Pie chart Allergic Rhinitis

A pie chart of reflux

Description automatically generated with medium confidence

Pie chart Reflux esophagitis

***Bivariate graphs for continuous variables:***

***A screenshot of a computer

Description automatically generated with low confidence***

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing screenshot, diagram, text, line

Description automatically generated

A picture containing screenshot, rectangle, diagram, line

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer screen

Description automatically generated with low confidence

***Bivariate Graphs for continuous variables:***

A screen shot of a graph

Description automatically generated with low confidence

A screenshot of a graph

Description automatically generated with low confidence

A picture containing text, screenshot, black and white

Description automatically generated

C4.

In data wrangling, I first checked if there were any repeated columns in the dataset. I found a duplicate column called "Unnamed" that was identical to the "Case Order" column. To avoid redundancy, I removed the duplicate column and made the "Case Order" column the index by renaming it as "index". Next, I focused on identifying missing values in the dataset. I counted the number of null values in each column to determine which columns had missing data.

To handle these missing values, I employed different approaches based on the type of data. For categorical columns like "overweight," "anxiety," and "soft drink," I filled in the missing values with the mode, which is the most common value in the column. For numerical columns like "income" and "children," I used the median, which is the middle value, to fill in the missing values. Similarly, for columns like "age" and "initial day," I used the mean, which is the average, to fill in the missing values.

To ensure consistent formatting, I added leading zeros to all ZIP codes in the dataset. This ensured that all ZIP codes had a uniform format with zeros at the beginning. Additionally, I standardized the representation of categorical data by converting "yes" values to 1 and "no" values to 0, creating a binary representation. To identify potential outliers in the longitude ("Lng") and latitude ("Lat") values, I utilized a visualization tool called ggplot to create a map visualization. This allowed me to visually examine the distribution of these variables and detect any values that deviated significantly from the rest.

To further investigate outliers, I used the lapply() function to check for columns that contained outlier values. I printed the names of these columns for further analysis. To visualize the outliers, I created boxplots for each of these columns. Boxplots provided a clear representation of the range and distribution of the data, enabling me to determine if any further actions were required to handle the outliers.

By performing these data wrangling steps, I successfully cleaned and prepared the dataset, ensuring its reliability and readiness for further analysis. Addressing missing values and outliers enhanced the overall quality of the data, setting the foundation for obtaining meaningful insights from the dataset.

I have a Rfile attached named, “**file done.R**”. to display data wrangling processes.

C5. Provide the prepared data set as a CSV file.

I have attached the CSV file named “**MD.csv**”.

Part IV: Model Comparison and Analysis

D. Compare an initial and a reduced linear regression model by doing the following:

D1.

***Linear Regression model:***

A screenshot of a computer

Description automatically generated with medium confidence

D2.

I'm planning to use a method called backward stepwise elimination to improve my linear regression model. This method involves removing variables that are not statistically significant one at a time. Starting with a model called BSE, I will remove the “least significant feature at each iteration”, based on its p-value, which has been shown to enhance the overall performance of the model (Middleton, 2022). If a variable's p-value exceeds a certain threshold, such as 0.05, I will remove it from the model. By eliminating non-significant variables, I aim to create a simpler and more meaningful model. After removing variables, I will assess how well the final model fits the data and interpret the effects of the remaining variables. It's important to consider other factors and be aware of the limitations of this method. The goal is to have a better model that effectively explains the relationship between the variables.

D3.

***The screenshots below are of each model:***

**The starting model:**

**A screenshot of a computer

Description automatically generated with medium confidence**

**Remove Variable Diabetes**

**A screenshot of a computer

Description automatically generated with medium confidence**

**Remove Variable Income**

**A screenshot of a computer program

Description automatically generated with medium confidence**

**Remove Variable BackPain**

**A screenshot of a computer

Description automatically generated with medium confidence**

**Removed variable Allergic\_rhintis**

**A screenshot of a computer

Description automatically generated with medium confidence**

**Removed variable Overweight**

**A screenshot of a computer

Description automatically generated with medium confidence**

**Removed Variable Arthritis**

**A screenshot of a computer

Description automatically generated with medium confidence**

**Removed Variable Reflux\_esophagis**

**A screenshot of a computer

Description automatically generated with medium confidence**

**Remove variable Stroke**

**A screenshot of a computer program

Description automatically generated with low confidence**

**Removed variable Additional\_charges**

**A screenshot of a computer

Description automatically generated with medium confidence**

**Removed variable Hyperlipidemia**

**A screenshot of a computer

Description automatically generated with medium confidence**

**Removed variable Anxiety**

**A screenshot of a computer

Description automatically generated with medium confidence**

**The reduced linear regression Model:**

**A screenshot of a computer code

Description automatically generated with low confidence**

E. Analyze the data set using your reduced linear regression model by doing the following:

E1.

When evaluating my linear regression models, I used the R-squared to measure how well the models fit the data. R-squared shows the “percentage variation in dependent that is explained by independent variables” (Yadav, 2019). In the initial model, the R-squared value was 0.3651232, which means that about 36.51% of the variation in the outcome variable could be explained by the predictor variables. As I simplified the model to make it more understandable, I developed a reduced model that had a slightly lower R-squared value of 0.364594. This means that the reduced model explained a tiny bit less of the variation in the outcome variable compared to the initial model.

While the initial model showed a slightly better fit, I also considered other factors like simplicity and the significance of the variables. The reduced model was chosen because it was simpler and still captured most of the important information while removing unnecessary or less impactful variables. So, despite the reduced model having a slightly lower R-squared value, it provided a good balance between simplicity and accuracy in explaining the relationship between the predictors and the outcome variable.

E2.

***Residual plot:***

A picture containing diagram, text, map

Description automatically generated

**Model Residual Standard Error:**

The Residual Standard Error (RSE) is a measure that tells me how much, on average, my predictions in the linear regression model differ from the actual values. In my case, the RSE value of 19.81965 means that, on average, my predictions are off by about 19.82 units. A lower RSE is better because it means my predictions are closer to the actual values. A higher RSE means my predictions are more spread out and less accurate. By looking at the RSE, I can get an idea of how well my model fits the data and how reliable my predictions are.

E3.

I have attached the Rscript used to implement the linear regression models. Named “**D208.R”.**

(To ensure that formatting of code is executable for the evaluator.)

Part V: Data Summary and Implications

F.

F1.

**The Regression Equation:**

Y = 9.044174509 - 1.207225679(HighBlood) - 0.985255214(Asthma) + 0.004441403(TotalCharge)

or

Initial\_days = 9.044174509 - 1.207225679(HighBlood) - 0.985255214(Asthma) + 0.004441403(TotalCharge)

**Interpretation of Coefficients:**

The coefficient for "HighBlood1" is -1.2072. This suggests that, holding all other variables constant, having high blood pressure (HighBlood1 = 1) is associated with a decrease of approximately 1.2072 in my "Initial\_days" compared to not having high blood pressure (HighBlood1 = 0).

The coefficient for "Asthma1" is -0.9853. This indicates that, holding all other variables constant, having asthma (Asthma1 = 1) is associated with a decrease of approximately 0.9853 in my "Initial\_days" compared to not having asthma (Asthma1 = 0).

The coefficient for "TotalCharge" is 0.0044. This means that for each unit increase in the total charge, my "Initial\_days" is expected to increase by approximately 0.0044, assuming all other variables are held constant.

**Statistical and Practical significance of the reduced model:**

In my reduced model, I found that certain variables have both practical and statistical significance in predicting the number of initial days. The variables "HighBlood1" and "Asthma1" indicate that having high blood pressure and asthma, respectively, have a notable impact on the number of initial days. This information is important for understanding and managing these health conditions. Additionally, the variable "TotalCharge" has a significant effect on the number of initial days. This means that the total charge incurred also plays a role in predicting the length of the initial days. This finding has implications for patients and healthcare providers, as it suggests that the financial aspect can influence the duration of the initial period.

From a statistical perspective, the significance of these variables is determined by their p-values. In my analysis, the p-values associated with "HighBlood1," "Asthma1," and "TotalCharge" are all below the commonly used threshold of 0.05. This indicates that the relationships observed are unlikely to have occurred by chance and are considered statistically significant. Overall, these results provide valuable insights into the factors that affect the length of the initial days. Understanding the impact of high blood pressure, asthma, and total charges can help healthcare professionals, policymakers, and individuals make informed decisions and develop strategies to manage and improve outcomes in healthcare settings.

**Limitations:**

As someone without expert guidance, there are limitations to consider in this analysis. One limitation is the risk of including irrelevant variables or overlooking important ones. I might miss key factors like age or gender that could have a significant impact on the outcome.

Another limitation is the potential for misinterpreting the statistical results and their practical implications. Without expertise, I might misunderstand the significance of variables like "HighBlood1" (high blood pressure) or "Asthma1" (asthma) in relation to "Initial\_days". An expert could provide insights into the practical implications of these variables and help me understand their true effects.

The absence of an expert makes it difficult to assess data quality and potential biases. I might not recognize missing data or outliers that could affect the reliability of the analysis. Variables like "TotalCharge" might be influenced by other factors that I am not aware of, and an expert could help identify and address these issues. An expert would help in selecting and validating the regression model. They could guide me in choosing the most appropriate model for the research question, considering variables like "HighBlood1", "Asthma1", and "TotalCharge". They could also evaluate the model's performance and determine how well it predicts the value of "Initial\_days".

In summary, the absence of expert guidance limits my ability to select relevant variables, interpret results accurately, ensure data quality, and validate the chosen model. Collaborating with experts would enhance the reliability and validity of the analysis, providing clearer insights into the relationship between variables and their practical significance in predicting "Initial\_days".

F2.

Based on the overall analysis of the reduced linear regression model variables "Initial\_days," "Asthma," "TotalCharge," and "HighBlood," it is recommended to collaborate with medical experts and specialists to further investigate the impact of asthma on the length of initial hospital stays. Also conducting a detailed analysis of the relationship between total charges and initial days can help identify opportunities to reduce costs. It is also important to engage with hypertension specialists to explore the influence of high blood pressure on the length of initial hospital stays. Based on the findings, targeted interventions can be implemented to address these factors. Making sure to continuously monitor and assess outcomes will help improve patient care and overall quality of care. This data-driven approach will support decision-making and lead to improvements in patient management and hospital operations.

**References**

*Data exploration through summary statistics - data exploration*. Coursera. (n.d.). https://www.coursera.org/lecture/big-data-machine-learning/data-exploration-through-summary-statistics-ZLttd?utm\_medium=sem&utm\_source=bg&utm\_campaign=B2C\_NAMER\_\_coursera\_FTCOF\_\_branded-search-country-US-country-CA&campaignid=278172071&adgroupid=1209463253056290&device=c&keyword=https%3A%2F%2Fwww.coursera.org%2F&matchtype=b&network=o&devicemodel=&adposition=&creativeid=&hide\_mobile\_promo&msclkid=f8b74d1acc1b1ee1546629805d31f8f9&utm\_term=https%3A%2F%2Fwww.coursera.org%2F&utm\_content=01+-+Branded+Search+-+US+-+dynamic+ad+group

Neal. (2022, March 31). *Residual standard error of a regression in python*. Tech Help Notes. https://techhelpnotes.com/residual-standard-error-of-a-regression-in-python/

*Using and interpreting indicator (dummy) variables¶*. Using and Interpreting Indicator (Dummy) Variables - Unifying Data Science. (n.d.). <https://www.unifyingdatascience.org/html/interpreting_indicator_vars.html>

Yadav, J. (2019, August 15). *Statistics: How should I interpret results of ols?*. Medium. https://jyotiyadav99111.medium.com/statistics-how-should-i-interpret-results-of-ols-3bde1ebeec01

Zach. (2021, May 19). *How to interpret an odds ratio less than 1*. Statology. https://www.statology.org/interpret-odds-ratio-less-than-1/