Mercedes Joseph

D208

Professor Middleton

Logistic Regression Modeling Part 2

July 9, 2023

**A1. Research Question:**

What medical conditions and other factors contribute to the readmission of patients?

**A2. Goal for Analysis**

My goal for analysis is to determine what medical conditions and other factors influence the readmission of patients. I will determine this by using a logistic regression model that will help me analyze what factors influence patient readmission. I have a list of independent variables, including medical conditions like HighBlood, Stroke, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Anxiety, Allergic\_rhinitis, Reflux\_esophagitis, Asthma, as well as other factors such as Income, Total charge, and Additional charges.

With logistic regression, I can determine the relationship between these variables and the likelihood of patient readmission. The model will give me coefficients that show the impact of each variable on the probability of readmission. By looking at these coefficients, I can understand which medical conditions and other factors are most important in predicting readmission of patients. This information will help me prioritize and focus on the factors that have the biggest influence on patient outcomes. Additionally, logistic regression provides a probability estimate for each patient's readmission, allowing me to identify those which patients are at a higher risk. Knowing the factors that contribute to readmission, will help the hospital take proactive steps to prevent it, such as providing targeted interventions and follow-up care, ultimately improving patient care outcomes.

**B1. Four Assumptions**

I chose to review four assumptions linearity, multicollinearity, homoscedasticity, and normality of residuals for my logistic regression model. In logistic regression, I need to be able to make certain assumptions about how the predictor variables relate to the outcome variable.

The first assumption in question is linearity, which means that the way the predictors affect the likelihood of patient readmission is a straight line. It assumes that if a predictor changes by a certain amount, the outcome will change by a proportional amount. This assumption helps me understand how each predictor influences the chances of readmission.

Multicollinearity is the second assumption which indicates that the predictors I use in the logistic regression model should not be strongly correlated with each other. When predictors are highly correlated, it becomes challenging to separate and understand their individual effects on readmission. Therefore, it is essential for me to make sure that my predictors are independent and not strongly related to one another.

The third assumption is homoscedasticity I make is about the variability in the differences between the observed and predicted values, known as residuals. This means that the spread of the residuals should be similar, no matter what the predictor values are. When homoscedasticity is violated, “the “spread” of the points across predicted values are not the same” (statsnotebook,2020). This assumption is crucial because it enables me to make dependable predictions across a wide range of predictor values.

The normality of residuals is the final assumption which means I assume that the errors or differences between the observed and predicted values in the logistic regression model follow a bell-shaped curve or a normal distribution. This is crucial because it implies that the residuals should be symmetrical around zero. If the residuals deviate significantly from this pattern, it can impact the accuracy of my predictions and the reliability of the model's results.

To check the accuracy and reliability of my logistic regression analysis, it is important for me to assess whether these assumptions hold true. If any of these assumptions are not met in the logistic regression model, I may need to take additional steps to enhance the accuracy of my analysis. This can include transforming variables or considering alternative logistic regression models that better align with the violated assumptions. By addressing these issues specific to logistic regression, such as linearity, multicollinearity, homoscedasticity, and the normality of residuals, I can improve the quality and robustness of my logistic regression analysis results.

**B2. Programming Language and Benefits**

R is a programming language that I chose to use for this project, which offers built in tools and packages to simplify the process of analyzing data using logistic regression. With R, I can easily clean and prepare my data by organizing and modifying it, handling missing values, and creating new variables. This ensures that my data is in the right format for analysis, saving me time and effort.

R's libraries provide specific functions and tools designed for building and evaluating logistic regression models. These libraries assist in estimating the relationships between variables, checking the model's fit, and identifying any potential issues. By using R for logistic regression, I can interpret the results of my analysis and present them visually, that gives valuable insights into the relationships between variables and effectively communicating my findings.

In my logistic regression analysis, I made use of various packages. The `tidyverse` package provided helpful tools for visualizing and manipulating data, including the `ggplot2` package for creating visualizations and the `dplyr` package for data cleaning and organization. The `stats` package, a core package in R, offered functions for performing logistic regression analysis and obtaining important information about the model's accuracy. I utilized the `vcd` package to handle categorical variables, allowing me to understand their relationships with the predicted outcome. Additionally, I loaded the `gmodels` package, which provided additional tools for summarizing and diagnosing the logistic regression model.

**Screenshot of all packages and libraries:**

**A screenshot of a computer program

Description automatically generated**

**B3. Justification of using Regression**

Logistic regression is the right technique for me to analyze my research question, **"What medical conditions and other factors contribute to the readmission of patients?".** This is because logistic regression allows me to view the relationship between predictor variables and my categorical variable “ReAdmis”. With logistic regression, I can estimate the likelihood or probability of readmission based on the given medical conditions and other factors. This helps me understand which predictors are associated with a higher or lower chance of readmission.

Using a logistic regression model is important for answering my research question because it gives me a statistical framework to analyze how different factors influence the likelihood of readmission. It helps me identify the specific medical conditions and other factors that have a significant impact on the risk of readmission. By using logistic regression, I can determine which medical conditions and other factors contribute to readmission, and this information can guide strategies to improve patient outcomes and reduce readmission rates.

**C1. Data Cleaning**

For my D208 assignment, I worked with a CSV file that had already undergone data cleaning and preparation from my D206 assignment. During the cleaning process, I faced several challenges and set specific goals to ensure the data was of high quality and ready for analysis. One challenge was deciding how to handle categorical data related to medical conditions. I carefully examined each variable and determined whether it made sense to convert the categories into numeric values. This decision was based on the nature of each variable and its relevance to the analysis.

Another task involved checking for duplicate columns to avoid redundant information. Fortunately, I did not find any duplicate columns, which ensured the accuracy of the data. To improve the clarity of the dataset, I renamed the columns labeled as "Item 1-8" based on the provided scenario. This allowed for easier understanding and interpretation of each item's meaning. Filling in missing values was another crucial step. I identified columns that had missing information and selected appropriate methods for filling the gaps. For numerical data, I used the median, mode for categorical data, and mean for continuous variables. This step ensured that the dataset was complete and ready for analysis.

To gain a better understanding of the data, I created various visualizations such as boxplots, scatterplots, and histograms. These visual models provided insights into the distribution of the data and the relationships between variables. During the data cleaning process, I also encountered outliers. However, upon further analysis, I determined that these outliers accounted for a very small percentage of the dataset and were primarily related to survey-based questions about patients' experiences. As they were not data errors, I concluded that removing these outliers was unnecessary. Overall, by making informed decisions, addressing challenges, and following a systematic data cleaning approach, I ensured the dataset's quality and prepared it for further analysis. I have attached an R script named **"file\_done.R"** that contains all the code used in the data cleaning process.

**C2. Summary statistics**

**Code for summary statistics of all variables:**

**A picture containing text, font, screenshot, white

Description automatically generated**

**Screenshot of summary statistics:**

**A screenshot of a computer

Description automatically generated**

**Summary of each variable:**

**In terms of the medical conditions and other factors in the dataset, here are the summary statistics:**

**HighBlood**

**The summary statistics suggest that around 40.9% of the individuals in the dataset are identified as having high blood pressure, as represented by the "High Blood" variable**.

* Min.: 0.000
* 1st Qu.: 0.000
* Median: 0.000
* Mean: 0.409
* 3rd Qu.: 1.000
* Max.: 1.000

**Stroke**

**Approximately 19.9% of the individuals in the dataset have encountered a stroke. This information is derived from the "Stroke" variable, which signifies whether a person has had a stroke or not.**

* Min.: 0.0000
* 1st Qu.: 0.0000
* Median: 0.0000
* Mean: 0.1993
* 3rd Qu.: 0.0000
* Max.: 1.0000

**Overweight**

**It can be observed that approximately 73.8% of the individuals in the dataset are categorized as overweight, as indicated by the "Overweight" variable.**

* Min.: 0.0000
* 1st Qu.: 0.0000
* Median: 1.0000
* Mean: 0.7377
* 3rd Qu.: 1.0000
* Max.: 1.0000

**Arthritis**

**Arthritis is a condition characterized by inflammation and stiffness in the joints. The average is 35.74% of the patients in the dataset has arthritis.**

* Min.: 0.0000
* 1st Qu.: 0.0000
* Median: 0.0000
* Mean: 0.3574
* 3rd Qu.: 1.0000
* Max.: 1.0000

**Diabetes**

**Diabetes is a chronic condition that affects the body's ability to regulate blood sugar levels. The average is around 27.38% of the patients in the dataset has diabetes.**

* Min.: 0.0000
* 1st Qu.: 0.0000
* Median: 0.0000
* Mean: 0.2738
* 3rd Qu.: 1.0000
* Max.: 1.0000

**Hyperlipidemia**

**According to the summary statistics, approximately 33.7% of the individuals in the dataset exhibit hyperlipidemia, which is characterized by high cholesterol levels.**

* Min.: 0.0000
* 1st Qu.: 0.0000
* Median: 0.0000
* Mean: 0.3372
* 3rd Qu.: 1.0000
* Max.: 1.0000

**BackPain**

**Back pain can range from mild discomfort to severe pain in the back region. The average value of 0.4114 indicates that around 41.14% of the patients in the dataset experience back pain.**

* Min.: 0.0000
* 1st Qu.: 0.0000
* Median: 0.0000
* Mean: 0.4114
* 3rd Qu.: 1.0000
* Max.: 1.0000

**Anxiety**

**According to the summary statistics, around 29.06% of the people in the dataset experience anxiety, as represented by the "Anxiety" variable.**

* Min.: 0.0000
* 1st Qu.: 0.0000
* Median: 0.0000
* Mean: 0.2906
* 3rd Qu.: 1.0000
* Max.: 1.0000

**Allergic\_rhinitis**

**The average value of 0.3941 suggests that around 39.41% of the patients have allergic rhinitis.**

* Min.: 0.0000
* 1st Qu.: 0.0000
* Median: 0.0000
* Mean: 0.3941
* 3rd Qu.: 1.0000
* Max.: 1.0000

**Reflux\_esophagitis**

**The average value of 0.4135 suggests that approximately 41.35% of the patients in the dataset have reflux esophagitis.**

* Min.: 0.0000
* 1st Qu.: 0.0000
* Median: 0.0000
* Mean: 0.4135
* 3rd Qu.: 1.0000
* Max.: 1.0000

**Asthma**

**The average value of 0.2893 indicates that around 28.93% of the patients in the dataset have asthma.**

* Min.: 0.0000
* 1st Qu.: 0.0000
* Median: 0.0000
* Mean: 0.2893
* 3rd Qu.: 1.0000
* Max.: 1.0000

**Income**

**The summary statistics reveal the distribution of income among individuals in the dataset, represented by the "Income" variable. The average income is approximately $38,872.4.**

* Min.: 154.1
* 1st Qu.: 23956.2
* Median: 33942.3
* Mean: 38872.4
* 3rd Qu.: 46466.8
* Max.: 207249.1

**TotalCharge**

**The summary statistics provide information about the distribution of total charges for medical treatment in the dataset, as indicated by the "TotalCharge" variable. The average total charge is approximately $5,892.**

* Min.: 1257
* 1st Qu.: 3253
* Median: 5852
* Mean: 5892
* 3rd Qu.: 7615
* Max.: 21524

**Additional\_charges**

**The summary statistics provide insights into the distribution of additional charges for medical treatment in the dataset, represented by the "Additional Charge" variable. The average additional charge is approximately $12,935.**

* Min.: 3126
* 1st Qu.: 7986
* Median: 11574
* Mean: 12935
* 3rd Qu.: 15626
* Max.: 30566

**ReAdmis**

**This variable indicates whether a patient was readmitted or not. The values are binary, with 0 representing no readmission and 1 representing readmission. The mean value is 0.3669, suggesting that approximately 36.69% of the patients were readmitted.**

* Min.: 0.0000
* 1st Qu.: 0.0000
* Median: 0.0000
* Mean: 0.3669
* 3rd Qu.: 1.0000
* Max.: 1.0000

These summary statistics provide a quick overview of the variables in the dataset giving insights into the prevalence of medical conditions and other factors.

C3. Univariate graphs

A pie chart of asthma

Description automatically generated with medium confidence

A picture containing text, diagram, screenshot, circle

Description automatically generated

A picture containing text, diagram, graphics, screenshot

Description automatically generated

A picture containing text, diagram, screenshot, font

Description automatically generated

A picture containing text, diagram, screenshot, font

Description automatically generated

A picture containing text, diagram, font, screenshot

Description automatically generated

A picture containing text, screenshot, font, diagram

Description automatically generated

A pie chart of back pain

Description automatically generated with medium confidence

A pie chart of anxiety

Description automatically generated with medium confidence

A picture containing text, diagram, screenshot, circle

Description automatically generated

A picture containing text, diagram, screenshot, circle

Description automatically generated

A graph with blue lines

Description automatically generated

A graph with blue bars

Description automatically generated

A graph of blue bars

Description automatically generated

A blue squares with a black line

Description automatically generated

Bivariate graph:

A screenshot of a graph

Description automatically generated with low confidence

A screenshot of a graph

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated with low confidence

A screenshot of a graph

Description automatically generated with low confidence

A screenshot of a computer screen

Description automatically generated with low confidence

A screenshot of a computer screen

Description automatically generated with low confidence

A screenshot of a graph

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated with low confidence

A screenshot of a graph

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated with low confidence

Bivariate Graphs

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

C4. **Data Transformation (Data Wrangling)**

***I performed six data wrangling activities to make sure all the data is up to quality standards.***

1. **Checking for duplicate columns**: The dataset was checked for any repeated columns, and a duplicate column called "Unnamed" was identified. To avoid redundancy, the duplicate column was removed, and the "Case Order" column was renamed as "index" and set as the index column.

2. **Handling missing values:**

The dataset was examined for missing values in each column. Different approaches were used based on the data type. For categorical columns like **"overweight”, “anxiety",** and **"soft drink"**, the missing values were filled with the mode or the most common value. For numerical columns like **"income"** and "**children"** the missing values were filled with the median or the middle value. Similarly, the mean or the average was used to fill missing values in columns like **"age"** and **"initial day."**

3. **Standardizing data formatting**:

The ZIP codes in the dataset were standardized by adding leading zeros to ensure a consistent format. This helped to maintain uniformity in the representation of ZIP codes.

4. **Converting categorical data:**

The representation of categorical data was standardized by converting **"yes"** values to 1 and **"no"** values to 0. This binary representation ensured consistency in the dataset.

5. **Identifying outliers:**

To detect potential outliers in the longitude and latitude values, a map visualization was created using the ggplot package. This visualization helped to visually examine the distribution of these variables and identify any values that significantly deviated from the rest.

6. **Handling outliers:**

The lapply() function was used to identify columns that contained outlier values. The names of these columns were printed for further analysis. Boxplots were then created for each of these columns to visually represent the range and distribution of the data and determine if any additional steps were needed to handle the outliers. By following these data wrangling steps, the dataset was cleaned and prepared for further analysis. The handling of missing values and outliers improved the overall quality and reliability of the data, ensuring its suitability for obtaining meaningful information.

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

C5. **CSV File**

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

D1. **Initial Model**

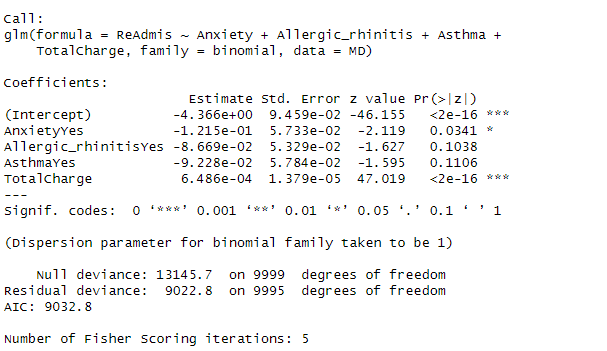
A screenshot of a computer

Description automatically generated

D2. **Model Reduction Method and Justification**

I will improve my logistic regression model using backward stepwise elimination, a method where I remove non-significant variables based on their p-values. By eliminating variables with p-values exceeding the standard of 0.05, I aim to create a simple and more meaningful model (Middleton, 2020). This process continues until all remaining variables are statistically significant. However, I will be to considering limitations such as multicollinearity and the potential removal of contextually or theoretically important variables. In all, I aim to create a model that accurately captures relationships in the data and enhances understanding.

D3. **Reduced Model**



E1. **Model Comparison**

A screenshot of a computer

Description automatically generated

When analyzing the results of the logistic regression model, I found that some variables, including **‘HighBlood’, ‘Stroke’, ‘Overweight’, ‘Arthritis’, ‘Diabetes’, ‘Hyperlipidemia’**, **‘BackPain’, ‘Allergic\_rhinitis’, "Reflux\_esophagitis",** and **"Asthma"** does not show a significant impact on the outcome. However, the variable **"Anxiety"** had a statistically significant effect, suggesting that it may influence the outcome **‘ReAdmis’**. With an addition, of other variables like **‘Income’, ‘TotalCharge’,** and **‘Additional\_charges’** showed a strong relationship with the variable **‘ReAdmis’**, indicating that these variables are important to predicting the outcome. This will be important to consider these findings when creating a plan for the outcome and understand how they contribute to our understanding of the outcome.

A screenshot of a computer

Description automatically generated with medium confidence

Based on my logistic regression analysis, I found that the presence of **‘Anxiety’** had a statistically significant impact on the outcome, indicating that it can influence the likelihood of the outcome ‘**ReAdmis**’. Although variables like **‘Allergic\_rhinitis’** and **‘Asthma’** did not show strong evidence of significance, further investigation will be needed to understand their potential relationship. Additionally, **"TotalCharge"** had a strong influence on the outcome, with higher charges associated with increased odds of patient **‘ReAdmis’** happening. The model's fit to the data was reasonably good fit, as indicated by the lower residual deviance (9022.8) compared to the null deviance (13145.7). Overall, this logistic regression analysis provides valuable insights into the predictive factors influence on the dependent variable (**ReAdmis)** of interest.

**Comparison between the two models:**

When comparing the two models I used the evaluation metric AIC or Akaike Information Criterion metric for the comparison between **initial\_model** and **final\_model.** AIC **“**is an estimator of prediction error which measures a statistical model in order to quantify the goodness of fit of the model (Li, 2023). The **initial\_model** had an AIC value of (9048.176), while the **final\_model** had an AIC value of (9032.815). The AIC is a measure how well a model fits the data, considering the model's complexity. A lower AIC value indicates a better fit. In this case, the **final\_model** has a slightly lower AIC value, suggesting that it fits the data a bit better than the **initial\_model**. Although, the difference between the two models is slight the final\_model clearly shows a more accurate representation of which variables are significant.

E2. **Calculation and Analysis**

1. **Confusion Matrix:**

A number on a white background

Description automatically generated

Looking at the results from the matrix, I can see that the model made quite a few correct predictions. For the instances that were **“No”,** the model accurately predicted 5,736 of them as **"No"**, but it mistakenly predicted 595 as **"Yes”.** Also, for the instances that were actually "Yes," the model correctly predicted 3,322 of them as **"Yes"**, but it incorrectly predicted 347 as **"No"**. The results allowed me to evaluate the model's performance and assess its accuracy. While the model has a good number of correct predictions, there may still be room for improvement to reduce the number of false positives and false negatives.

1. **Accuracy Calculation**

****

With an accuracy of 0.9058, my final logistic regression model is correct in its predictions around 90.58% of the time. Meaning the model can accurately classify the outcomes for most cases. Accuracy being important in measuring because it tells us how well the model is performing overall. My accuracy score shows a potential high accuracy of 0.9058 indicating that the model does a good job in predicting readmission outcomes.

**E3. Code for Logistic Regression**

I have attached the R script "**D208\_task2.R**" for your review and execution.

F1. **Regression Equation, Coefficients, etc.**

1. **Regression Equation:**

ln(p/(1-p)) = -4.366 - 0.1215(Anxiety) - 0.08669(Allergic\_rhinitis) - 0.09228(Asthma) + 0.0006486(TotalCharge)

1. **Provide Interpretation of Coefficients:**

In a logistic regression model, the coefficients provide insights into the relationship between the predictors (Independent variables) and the outcome (dependent variables) **(Schober, P., Boer, C., & Schwarte, L. A., 2018).** The intercept, sits at (-4.366), represents the starting point of the log-odds when all predictors are zero. **Anxiety** shows a negative coefficient (-0.1215) suggests that higher levels of anxiety are associated with a decreased likelihood of the outcome. Along with the coefficients for **Allergic\_rhinitis** (-0.08669) and **Asthma** (-0.09228) indicate a slightly lower probability of the outcome for individuals with these conditions. On the other hand, the positive coefficient for **TotalCharge** (0.0006486) suggests that higher total charges are associated with a slight increase in the likelihood of the outcome. These coefficients help us understand the direction and magnitude of the effects of each predictor on the outcome **ReAdmis** in a logistic regression model.

1. **Discussion of significance**

The reduced model in logistic regression includes the variables **Anxiety,** **Allergic\_rhinitis, Asthma**, and **TotalCharge**. The statistical significance of these variables is determined by their p-values. In this case, **Anxiety** has a p-value of (0.0341), indicating that this variable has statistical significance at the conventional threshold of (0.05). This suggests that the presence of anxiety has a significant effect on the likelihood of the outcome. However, the variables **Allergic\_rhinitis** and **Asthma** p-values are (0.1038) and (0.1106), which are slightly above the threshold level of significance (0.5). These variables are not statistically significant, but they may still have some practical significance or potential importance in certain contexts or when a part of a larger sample sizes. **TotalCharge** has a very small p-value (<2e-16), suggesting a strong statistical significance. In conclusion, the reduced model provides glimpses into the independent variables that are statistically significant that cause an impact on **ReAdmis**.

1. **Limitations**

In my data analysis process, it's important to acknowledge that not having an expert along the way can limit my understanding of the variables in the reduced model, including the dependent variable "**ReAdmis"**. Variables like **"Anxiety",** **"Allergic rhinitis”,"Asthma",** and **"TotalCharge"** in the reduced model may have significant relationships with **"ReAdmis"** that an expert can help explain in simple terms. An expert’s expertise can shed light on the underlying factors and mechanisms at play, allowing for a deeper understanding of how these variables affect the likelihood of readmission (**ReAdmis**). Additionally, experts can provide insights on potential confounding factors or alternative explanations that I may have overlooked. Their guidance can help me interpret the coefficients in a more meaningful way and assess the practical significance of the findings in relation to readmission rates. Without an expert, I may miss out on valuable context and explanations that could enhance the overall quality and reliability of my analysis.

F2. **Recommended Course of Action**

Based on the results of the data analysis, it is evident that **"Anxiety"**, **"Allergic Rhinitis"**, **"Asthma"**, and **"TotalCharge"** are the most significant factors in predicting readmission. To address this, I recommend implementing specific actions from a data analytics standpoint. Firstly, healthcare providers should leverage patient data and employ predictive analytics models to identify individuals at a higher risk of readmission due to these factors. This proactive approach will enable the creation of intervention plans and targeted interventions to prevent readmissions. Additionally, conducting further analysis is crucial to understanding the underlying mechanisms and relationships between these factors and readmission rates. Exploring additional variables, performing subgroup analyses, and considering interactions between variables can provide deeper insights into the readmission process. Lastly, it is essential to establish a continuous monitoring and evaluation system that tracks the implementation of medical interventions and their impact on readmission rates. This system will allow for the assessment of the effectiveness of the recommended actions and facilitate data-driven adjustments when needed. By combining data analytics with clinical expertise, healthcare providers can develop more targeted interventions and improve patient outcomes by reducing readmission rates.

**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

LI, J. (n.d.). *How to evaluate time series models using AIC in R -*. ProjectPro. https://www.projectpro.io/recipes/evaluate-time-series-models-aic

Middelton. (2022). Geeting started with D208  Part 2. file:///C:/Users/merce/Downloads/Getting%20Started%20with%20D208%20(Part%20II)COIT%20(6).pdf

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, *126*(5), 1763-1768.