Case Study 2

Analyzing and Predicting Emergency Room Visits: An Exploration of Health Insurance Claim Data

JongHee Lee, Sejun Song

In healthcare insurance, understanding the implications of claim data, particularly regarding emergency room (ER) visits, is important for informed policy-making and resource allocation. This report explores the analysis of a dataset from a health insurance company, comprising 788 subscribers with claims resulting from ischemic heart disease during 1998-1999.

Data Collection and Variables

The dataset contains diverse variables related to patient demographics, medical interventions, and costs, outlined as follows:

1. ID: Unique identifier
2. TotalCost: Total cost of claims (USD)
3. Age: Age of the subscriber (Years)
4. Gender: Gender (1 if male; 0 if female)
5. Interventions: Total interventions or procedures
6. Drug: Number of prescribed drugs
7. ERVisits: Number of ER visits
8. Heart_Disease: Number of complications during heart disease treatment
9. Other_Diseases: Number of other diseases (“comorbidities”)
10. Duration: Duration of treatment condition (Days)

Objective

The analysis aims to discover patterns and relationships between ER visits and the aforementioned variables, thereby providing meaningful insights for enhanced healthcare interventions and insurance policy adjustments.

Data Preprocessing

In the initial phase of our data transformation, we concentrated on the dataset with Outlier Management. Outliers can significantly influence analyses and predictive modeling, as they disproportionately affect the estimation of model parameters. Through visual inspection of the 'ERVisits' predictor variable using tools such as scatterplots, we identified specific entries that appeared to be potential outliers, particularly those representing an unusually high number of ER visits.

However, given the potential issues associated with removing outliers from the predictor variable, we decided against their removal. Instead, we settled for Min-Max Scaling as a method to transform the data, adjusting the range of the 'ERVisits' variable to [0, 1], thus mitigating the impact of these extreme values. This approach ensures that the outliers remain part of the analysis, preserving the integrity of the data, while also maintaining the robustness of subsequent analyses and modeling, resulting in a more accurate and reliable outcome. To capture the unique structure and characteristics of the data, we decided to apply scaling after the EDA process.

Moreover, ensuring accurate interpretation of the variables' nature during the modeling process is important. The variables 'Gender', while numerically coded, is inherently categorical. To safeguard against potential numerical interpretation during modeling, it is converted into factor variables. This transformation guarantees that the variable is recognized and treated as categorical predictors when fitting appropriate models, maintaining the analytical integrity.

The data transformation phase ensured that the dataset was not only outliers were managed but also that variables were appropriately scaled and categorized. This solid foundation is prominent for the following Exploratory Data Analysis and modeling stages. Such preprocessing is essential to confirm that the subsequent results are valid, providing a sound basis upon which informed healthcare and policy decisions can be constructed.

Metheology

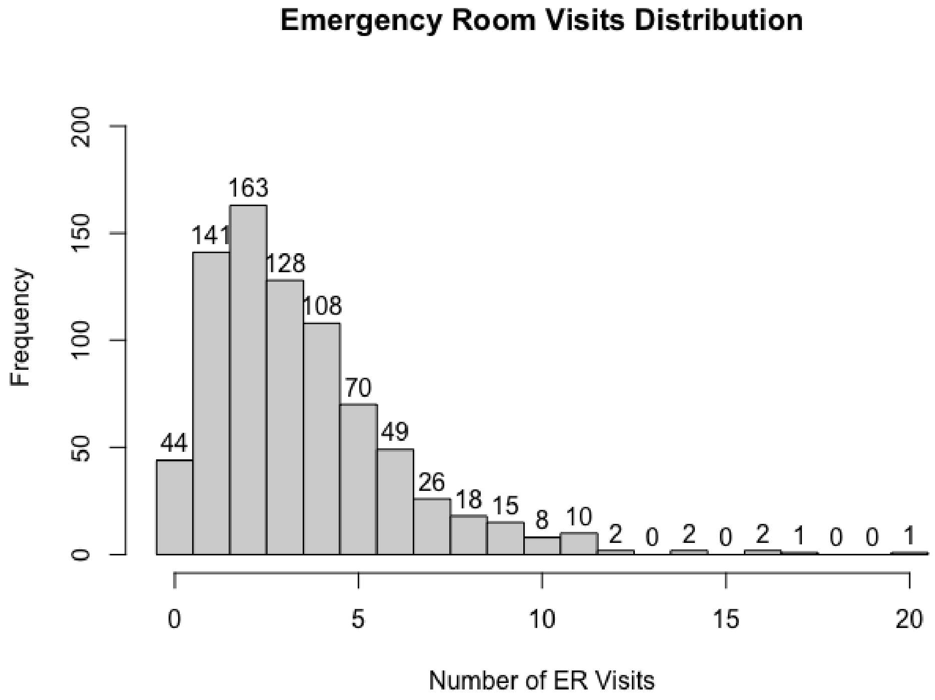
In our analysis, the Poisson regression model serves as a primary tool for characterizing count data, especially when lower values are more frequent, making it suitable for events such as the number of arrests or goals scored in tournaments. A fundamental assumption of this model is the equivalence of the mean and variance of the outcome variable.

However, when variance exceeds the mean, the Negative Binomial (NB) model comes into use. Similar to the Poisson, the NB model characterizes count data but adjusts for overdispersion by introducing an additional term. This model's distinctive feature is the parameter ψ, representing the failures that halt counting. As ψ increases, the NB distribution gravitates toward the Poisson distribution.

On the foundation of our understanding, when encountering overdispersion in Poisson distribution, the Negative Binomial (NB) regression serves as an alternative. To select a better model, we performed stepwise regression for poisson and NB, then conducted a likelihood ratio test (LRT). Through this approach, we aim to better identify variables related to ER Visits and to construct a well-fitted analysis model tailored to the distribution of the data.

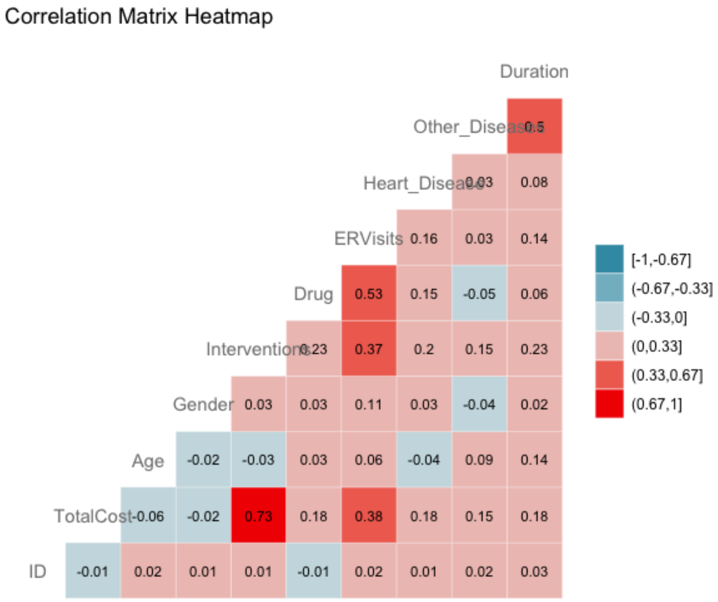
Exploratory Data Analysis

Figure 1. Histogram - Distribution of ERvisits



The histogram presents the distribution of Emergency Room (ER) visits among subscribers. Majority of subscribers have 1, 2, or 3 ER visits, implying a skewed distribution that is characteristic of a Poisson distribution. A notable segment made between 5 to 10 visits, suggesting possible recurrent health challenges. Only a small portion recorded an unusually high number of visits, which might be outliers or specific cases requiring more detailed scrutiny. However, descriptive statistics reveal that the mean is 3.43 and the variance is 6.96. This indicates the presence of overdispersion, and the assumptions required for a Poisson model are not met.

Figure 2. Correlation Matrix Heatmap



The heatmap displays the correlation matrix detailing the relationships between Emergency Room (ER) visits and various potential predictor variables. Strong positive correlations are represented in deep red, while strong negative correlations appear in deep blue. The variable with the highest correlation to 'ER Visits' is 'Drug' with a coefficient of 0.53, indicating a moderate positive correlation. Following this, 'TotalCost' and 'Interventions' had coefficients of 0.38 and 0.37, respectively. However, the strong correlation observed between 'TotalCost' and 'Interventions' suggests potential multicollinearity, which could compromise the reliability of coefficient estimates and complicate the modeling process. Variables such as 'Gender' and 'Duration' demonstrate minimal correlation values, underscoring their limited linear association with ER Visits.

Figure 3a. Scatterplot of ER Visits in Relation to TotalCost

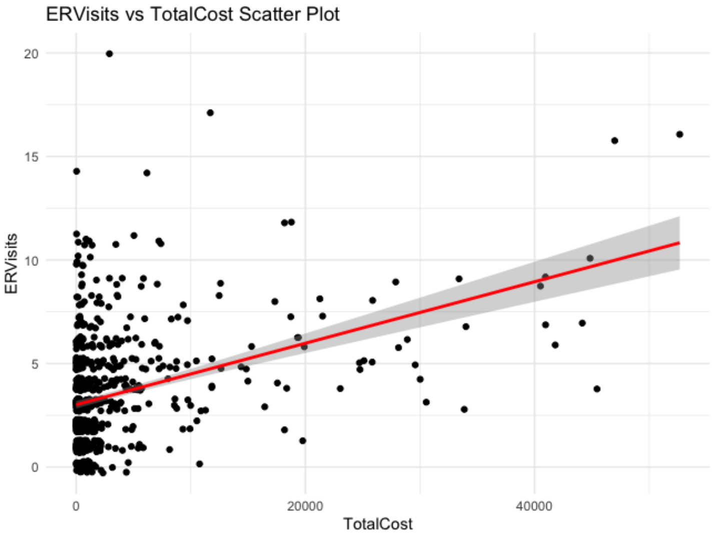


Figure 3b. Scatterplot of ER Visits in Relation to Drug Usage

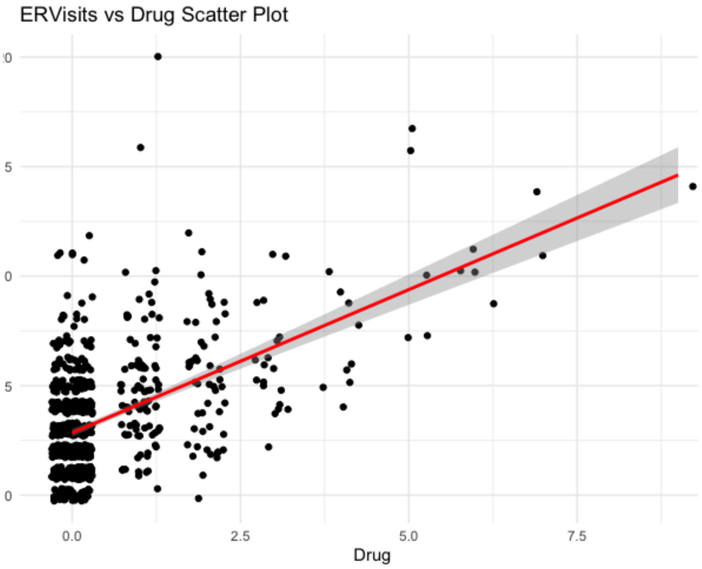


Figure 3a and 3b present scatterplots visualizing the association between ER Visits and each predictor. The plot features a red linear regression line showing an upward trend, suggesting a potential association where higher predictors are observed alongside an increase in ER Visits. The shaded area around the regression line represents a confidence interval, providing insight into the potential variability of this observed relationship. However, it is crucial to emphasize that this is an observational study. Therefore, we cannot infer causation from this correlation. The spread of the data points also indicates that additional variables may be influencing this relationship, underscoring the need for further investigation to fully understand the factors.

Figure 4. Boxplot of ER Visits Based on Complications During Heart Disease Treatment

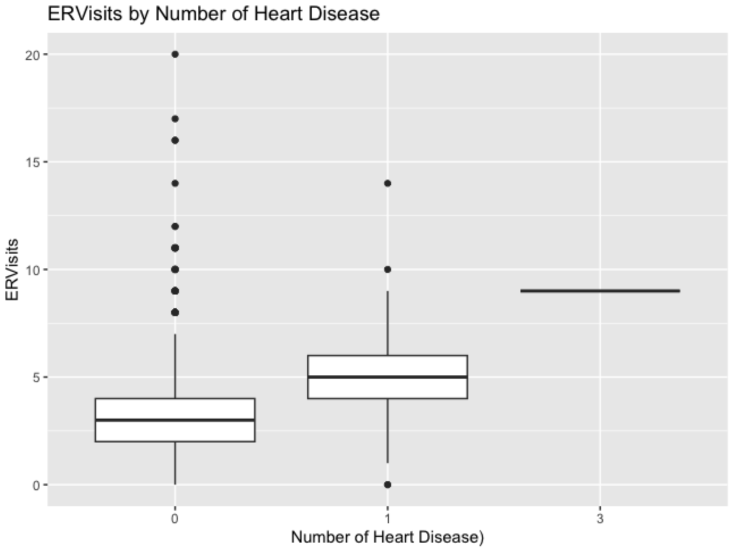
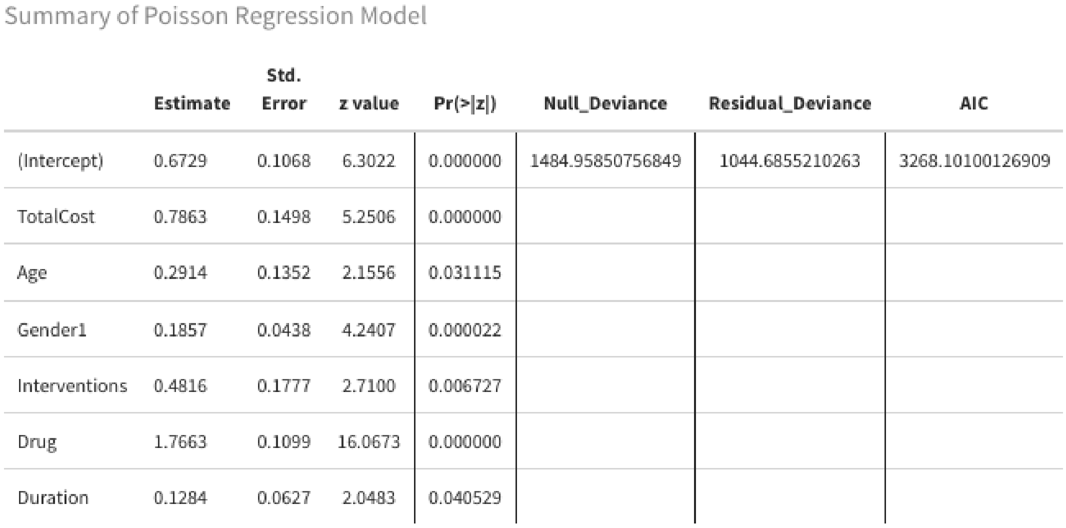


Figure 4 illustrates a boxplot comparing ER visits to the number of complications during heart disease treatment. For subscribers without complications, the median ER visits is about 5, though there are several potential outliers, with visits ranging up to 20. For those with one complication, the median lies close to 7, but outliers also exist in this category, with a few cases having 10 to 15 visits. When subscribers encounter three complications, ER visits consistently hover around 10. The plot suggests a correlation between increased complications and ER visits. However, it's worth noting the absence of data for subscribers with two complications and the mere presence of one data point for those with three complications. Out of the 788 data points, only one falls outside the range of zero or one complication. This hints at the presence of an outlier in the dataset.

Model Testing Based on Stepwise Regression

Figure 5a. Poisson Regression Model

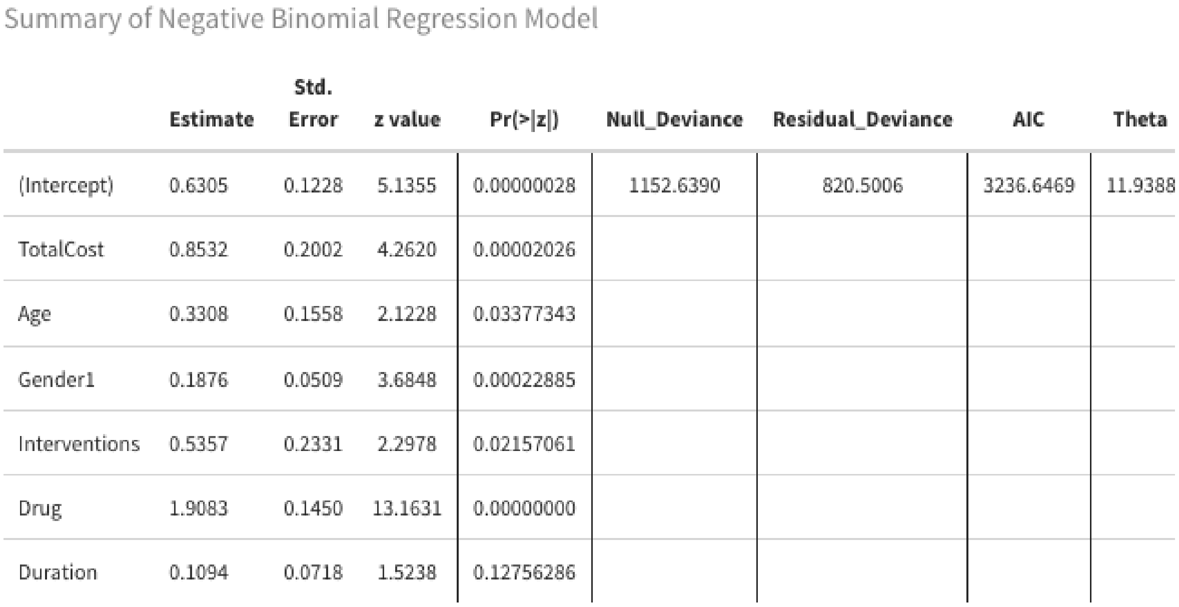


Summary of Poisson Regression Model

	Estimate	Std. Error	z value	Pr(>|z|)	Null_Deviance	Residual_Deviance	AIC
(Intercept)	0.6729	0.1068	6.3022	0.000000	1484.95850756849	1044.6855210263	3268.10100126909
TotalCost	0.7863	0.1498	5.2506	0.000000			
Age	0.2914	0.1352	2.1556	0.031115			
Gender1	0.1857	0.0438	4.2407	0.000022			
Interventions	0.4816	0.1777	2.7100	0.006727			
Drug	1.7663	0.1099	16.0673	0.000000			
Duration	0.1284	0.0627	2.0483	0.040529			

In the analysis of the model, it was observed that all variables, with the exception of 'Duration', demonstrated p-values less than 0.05. This indicates that these variables are statistically significant predictors within the framework of the model, highlighting their influential role in the outcome. Furthermore, the model's Akaike Information Criterion (AIC) value stands at 3268.10.

Figure 5b. Negative Binomial Regression Model



Summary of Negative Binomial Regression Model

	Estimate	Std. Error	z value	Pr(>|z|)	Null_Deviance	Residual_Deviance	AIC	Theta
(Intercept)	0.6305	0.1228	5.1355	0.00000028	1152.6390	820.5006	3236.6469	11.9388
TotalCost	0.8532	0.2002	4.2620	0.00002026				
Age	0.3308	0.1558	2.1228	0.03377343				
Gender1	0.1876	0.0509	3.6848	0.00022885				
Interventions	0.5357	0.2331	2.2978	0.02157061				
Drug	1.9083	0.1450	13.1631	0.00000000				
Duration	0.1094	0.0718	1.5238	0.12756286				

The Negative Binomial Regression Model indicates statistically significant predictors, with all variables, except for 'Duration', having p-values under 0.05. The model's Akaike Information Criterion (AIC) is 3236.65, suggesting a balance between model complexity and fit.

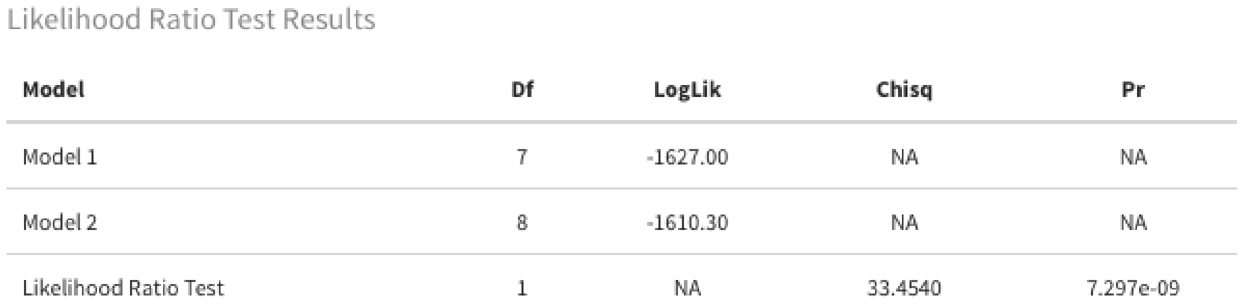
However, it is important to note that due to the stepwise selection procedure implemented during the analysis, the p-values of the selected variables may be biased downwards, exhibiting a tendency to be smaller. This potential bias arises from the iterative nature of stepwise procedures, which repeatedly tests variables for inclusion or exclusion, potentially leading to an overestimation of the significance of the selected variables. Therefore, caution should be exercised when interpreting these p-values.

Comparing the two models using their AIC values, the Negative Binomial Regression Model has a slightly lower AIC value, suggesting that it might be a better fit to the data compared to the Poisson Regression Model.

However, the final decision on which model to use should consider the theoretical assumptions of the models and the nature of the data. Additionally, looking at other diagnostic metrics and plots can provide a more comprehensive understanding of the model fit.

Likelihood Ratio Test

Figure 6. Result of LRT



Likelihood Ratio Test Results

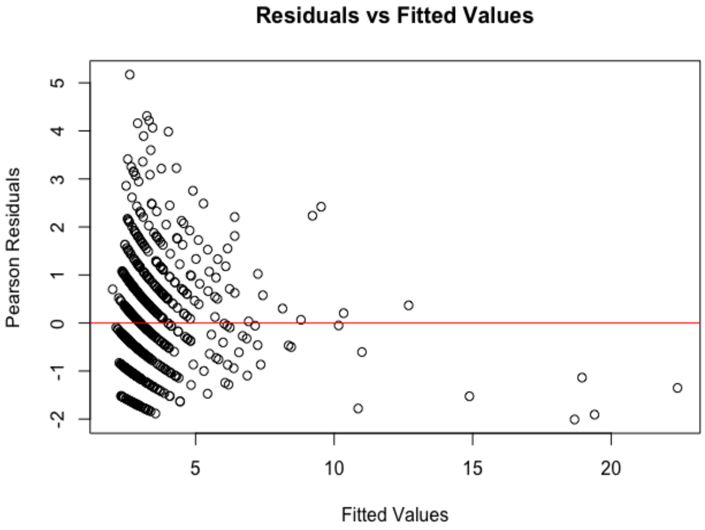
Model	Df	LogLik	Chisq	Pr
Model 1	7	-1627.00	NA	NA
Model 2	8	-1610.30	NA	NA
Likelihood Ratio Test	1	NA	33.4540	7.297e-09

In the provided Likelihood Ratio Test (LRT) comparing two regression models, Model 1 and Model 2, the results suggest a distinction in their performance. Model 2, with a log likelihood of -1610.30, exhibits a better fit to the data than Model 1, which has a log likelihood of -1627.0. The Chisq value of 33.45 and the extremely low p-value, indicate that this difference in fit is statistically significant. When investigating these models as a comparison between a Poisson regression (Model 1) and a negative binomial regression (Model 2), the results suggest that the negative binomial model offers a superior fit to the data. This could hint at the presence of overdispersion in the data set, where the observed variance exceeds the mean. Given this, the negative binomial regression, which accounts for overdispersion, seems more appropriate for this data than the Poisson regression.

Assumptions for Negative Binomial Regression

When deploying a Negative Binomial Regression, it is essential to ensure that certain assumptions are met. Below we delineate each assumption and validate them with respect to our dataset:

Figure 7. Residual vs Fitted Values plot



Linearity:

For the negative binomial regression to function effectively, the relationship between the linear predictor and the logarithm of the mean of the dependent variable must be linear. One of the primary indicators to determine the fit of the model is the distribution of residuals. Ideally, the residuals should be randomly distributed around 0. However, a clear pattern is observed in our dataset (figure 7), this suggests that the link function may not be appropriately modeling the relationship between the linear predictor and the dependent variable, indicating a violation of the model assumptions.

Independence:

Each observation should be independent of others. Given that each data point represents an individual entity, the assumption of independence holds valid in our case.

Multicollinearity:

This arises when there's a high correlation between independent variables within a model. The Variance Inflation Factor (VIF) is a widely used metric to evaluate this. In our data, all variables have a VIF less than 5 (figure 2), suggesting that multicollinearity is not an issue.

Non-Negativity:

The non-negativity assumption states that both the dependent variable and the predicted values should never be negative. To ensure this assumption, we employed Min-Max scaling to standardize the relevant variables.

Overdispersion:
A hallmark of the negative binomial regression is its capability to operate well in scenarios where the variance of the dependent variable is greater than its mean, indicating overdispersion. However, it might struggle with its opposite, underdispersion. Examining our data, the variance is significantly larger at 6.96 compared to the mean of 3.43, clearly indicating overdispersion in the dataset.

In conclusion, while most assumptions are met for our Negative Binomial Regression, the linearity assumption causes a challenge, and appropriate alternatives should be considered.

Negative Binomial Regression Model Interpretation

The result of the negative binomial regression model (Figure 5b) provides insights into factors affecting the number of emergency room visits by patients with ischemic heart disease. Specifically, the Total Cost of Treatment indicates a relationship between treatment expenses and ER visits. For every unit increase in the total cost of treatment, the expected log count of ER visits rises by approximately 0.81. This could suggest that patients with higher treatment costs may frequent the ER more, possibly due to dealing with more severe or complicated health challenges. Additionally, the model highlights a gender distinction. Being male, represented by the value "1" in the Gender variable, is linked to an increase in the expected log count of ER visits by 0.16 compared to females. This suggests that males may have certain gender-specific health concerns or behaviors leading to more ER visits than females. By understanding these factors, we can better predict ER visit trends and work towards more efficient and personalized healthcare solutions.

Conclusion:

In the context of subscribers with ischemic heart disease, several factors like the total cost of treatment, age, gender, interventions, and drugs significantly influence the frequency of ER visits. Health providers and insurance companies can utilize these insights to strategize patient care, allocate resources, and design intervention strategies to specific patient groups. Understanding these relationships also provide effective communication strategies, ensuring that patients are well-informed about potential risks.

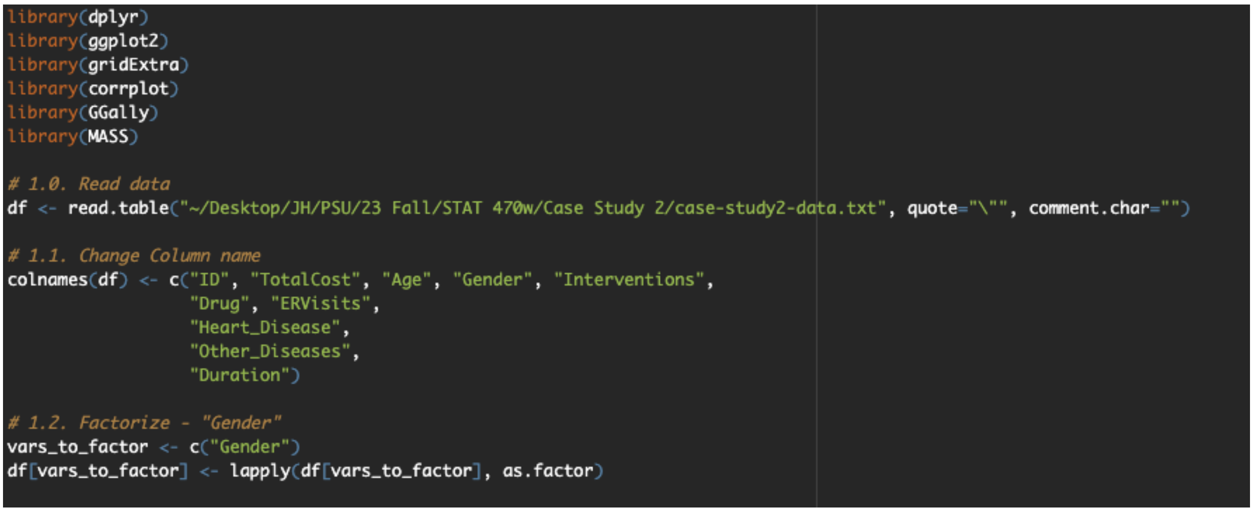
Reference

Malek-Ahmadi, M. (n.d.). *Regression analysis with right-skewed data: Applications for pre ...* Regression Analysis with Right-Skewed Data: Applications for Pre-Clinical Alzheimer's Disease. https://files.alz.washington.edu/presentations/2020/fall/Malek-Ahmadi.pdf

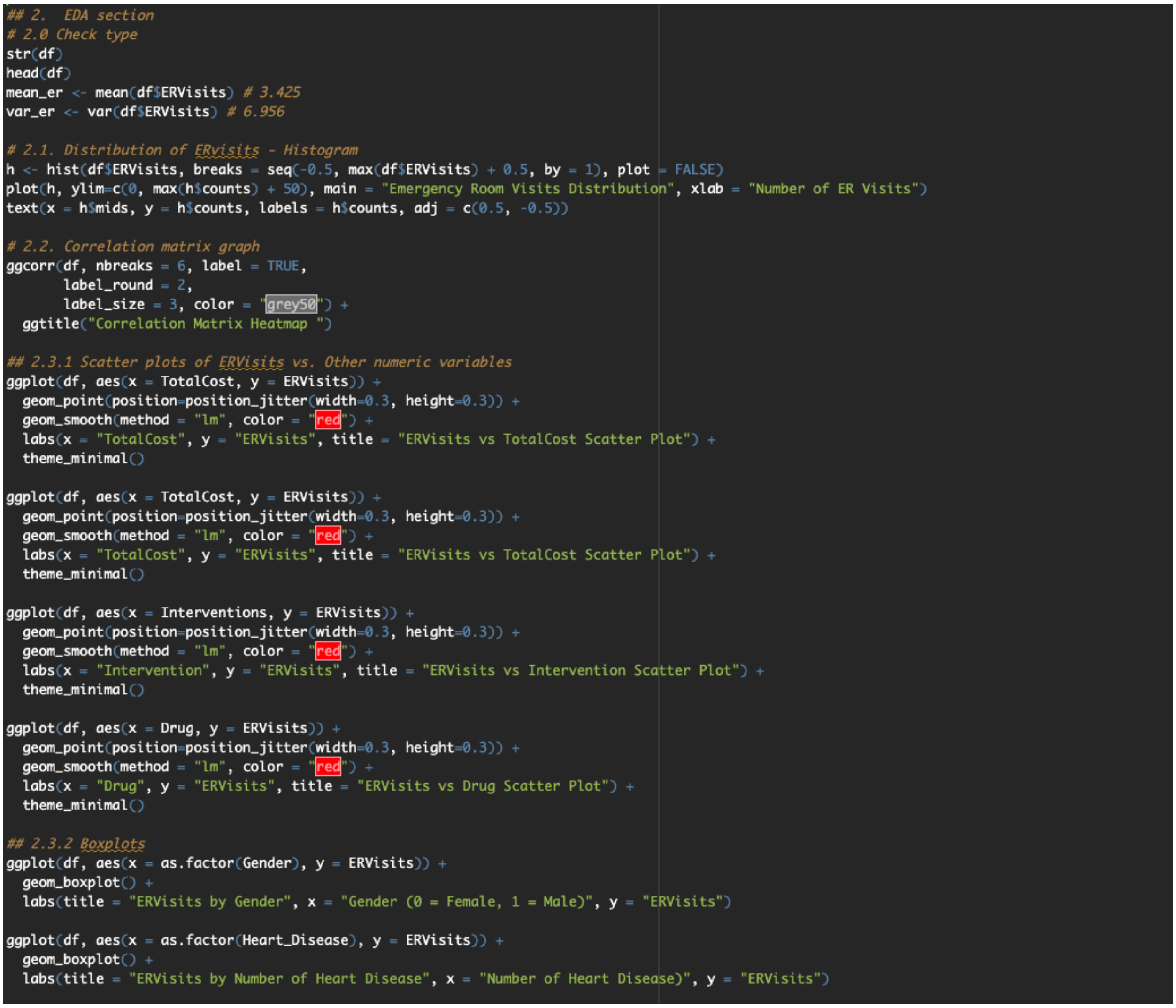
Feature scaling with Scikit-learn - Michael Fuchs python. MFuchs. (2019, August 31). https://michael-fuchs-python.netlify.app/2019/08/31/feature-scaling-with-scikit-learn/

View of the negative binomial regression: The southwest respiratory and critical care chronicles. View of The Negative Binomial regression | The Southwest Respiratory and Critical Care Chronicles. (n.d.). https://pulmonarychronicles.com/index.php/pulmonarychronicles/article/view/200/491

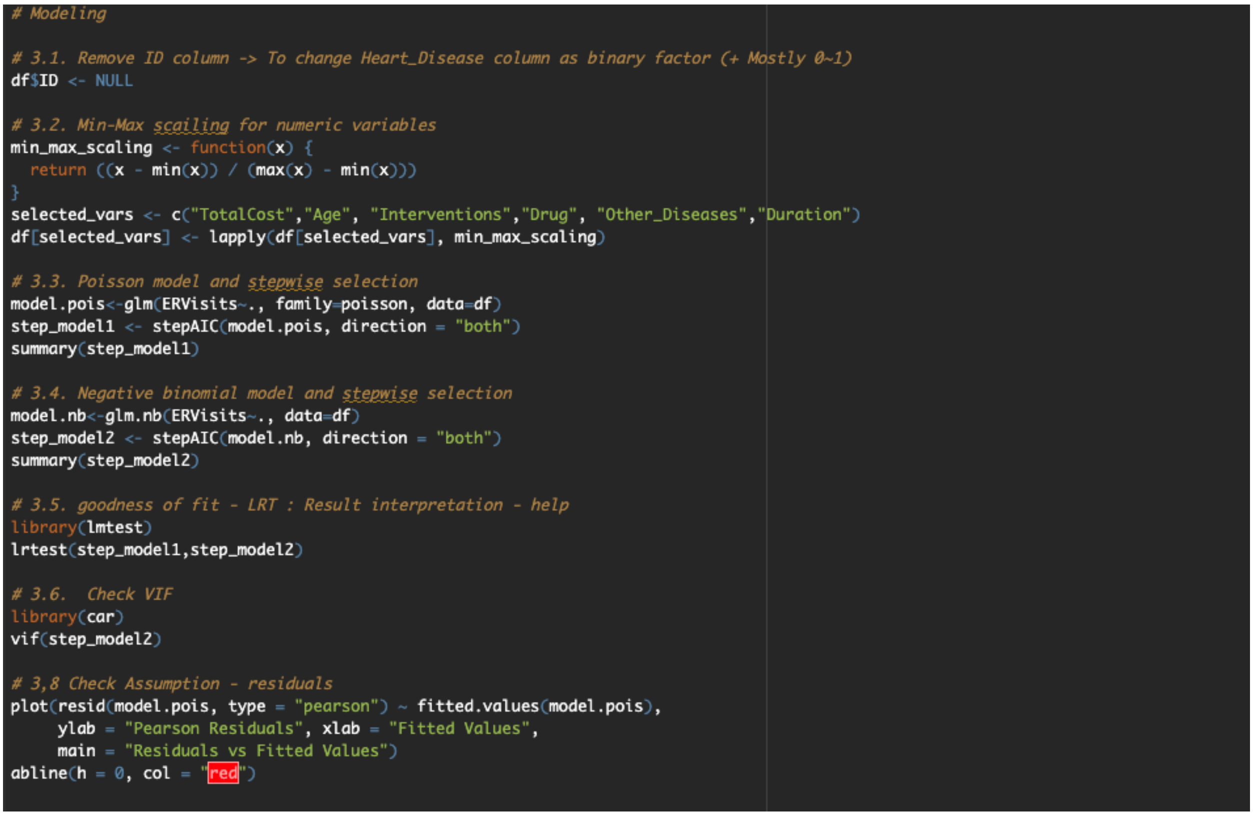
Code Appendix



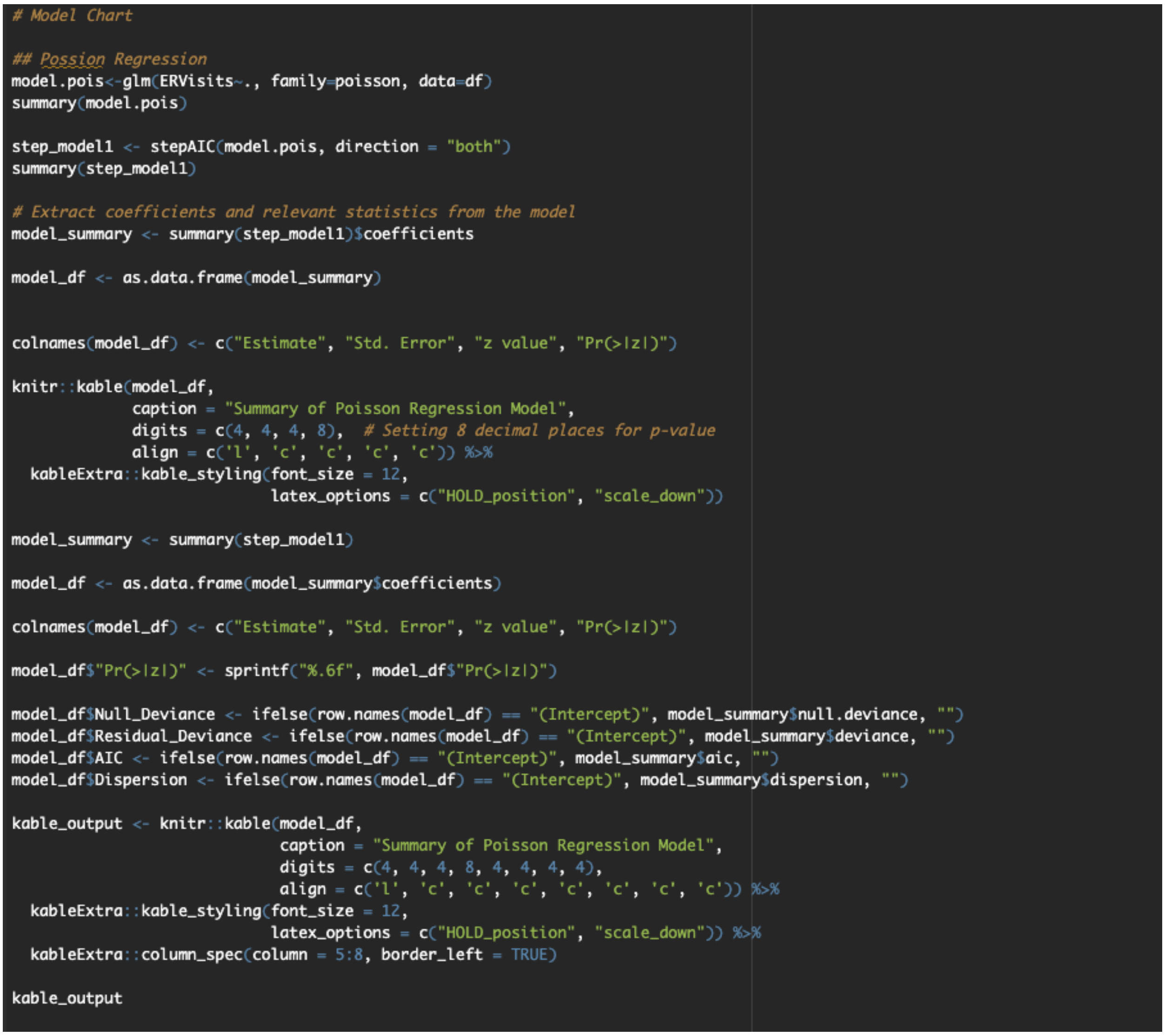
```
library(dplyr)
library(ggplot2)
library(gridExtra)
library(corrplot)
library(GGally)
library(MASS)

# 1.0. Read data
df <- read.table("~/Desktop/JH/PSU/23 Fall/STAT 470w/Case Study 2/case-study2-data.txt", quote="\"", comment.char="")

# 1.1. Change Column name
colnames(df) <- c("ID", "TotalCost", "Age", "Gender", "Interventions",
                  "Drug", "ERVisits",
                  "Heart_Disease",
                  "Other_Diseases",
                  "Duration")

# 1.2. Factorize - "Gender"
vars_to_factor <- c("Gender")
df[vars_to_factor] <- lapply(df[vars_to_factor], as.factor)
```



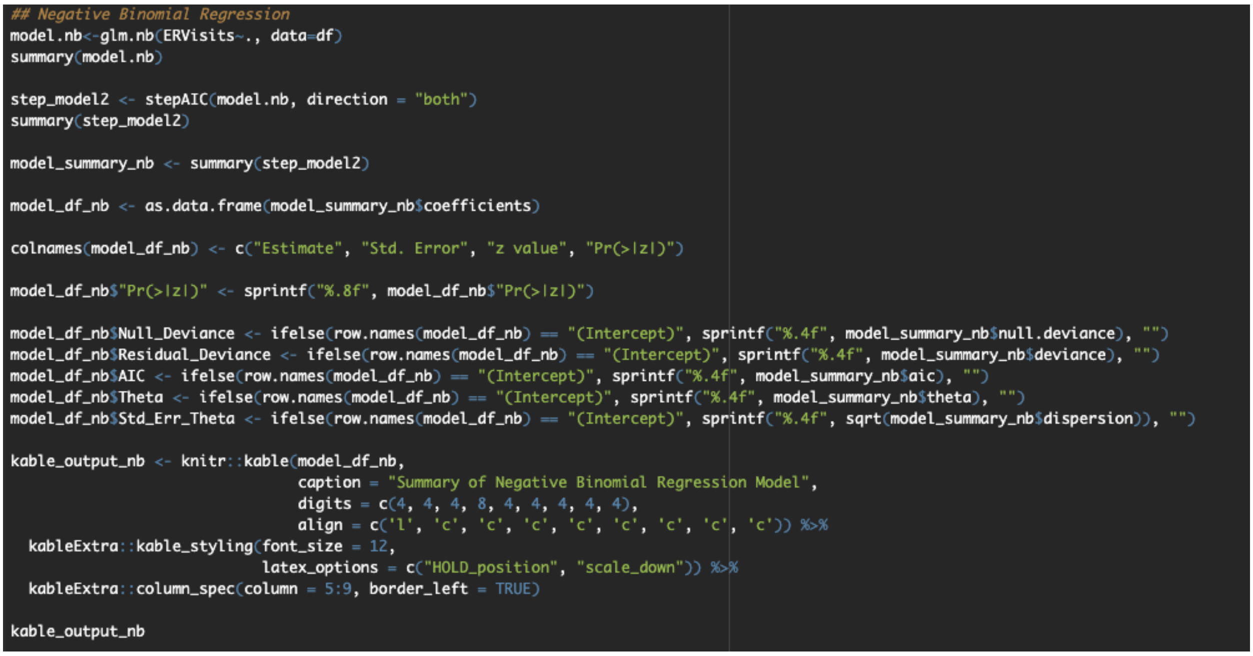
```
## 2.  EDA section
# 2.0 Check type
str(df)
head(df)
mean_er <- mean(df$ERVisits) # 3.425
var_er <- var(df$ERVisits) # 6.956

# 2.1. Distribution of ERvisits - Histogram
h <- hist(df$ERVisits, breaks = seq(-0.5, max(df$ERVisits) + 0.5, by = 1), plot = FALSE)
plot(h, ylim=c(0, max(h$counts) + 50), main = "Emergency Room Visits Distribution", xlab = "Number of ER Visits")
text(x = h$mids, y = h$counts, labels = h$counts, adj = c(0.5, -0.5))

# 2.2. Correlation matrix graph
ggcorr(df, nbreaks = 6, label = TRUE,
       label_round = 2,
       label_size = 3, color = "grey50") +
  ggtitle("Correlation Matrix Heatmap ")

## 2.3.1 Scatter plots of ERVisits vs. Other numeric variables
ggplot(df, aes(x = TotalCost, y = ERVisits)) +
  geom_point(position=position_jitter(width=0.3, height=0.3)) +
  geom_smooth(method = "lm", color = "red") +
  labs(x = "TotalCost", y = "ERVisits", title = "ERVisits vs TotalCost Scatter Plot") +
  theme_minimal()

ggplot(df, aes(x = TotalCost, y = ERVisits)) +
  geom_point(position=position_jitter(width=0.3, height=0.3)) +
  geom_smooth(method = "lm", color = "red") +
  labs(x = "TotalCost", y = "ERVisits", title = "ERVisits vs TotalCost Scatter Plot") +
  theme_minimal()

ggplot(df, aes(x = Interventions, y = ERVisits)) +
  geom_point(position=position_jitter(width=0.3, height=0.3)) +
  geom_smooth(method = "lm", color = "red") +
  labs(x = "Intervention", y = "ERVisits", title = "ERVisits vs Intervention Scatter Plot") +
  theme_minimal()

ggplot(df, aes(x = Drug, y = ERVisits)) +
  geom_point(position=position_jitter(width=0.3, height=0.3)) +
  geom_smooth(method = "lm", color = "red") +
  labs(x = "Drug", y = "ERVisits", title = "ERVisits vs Drug Scatter Plot") +
  theme_minimal()

## 2.3.2 Boxplots
ggplot(df, aes(x = as.factor(Gender), y = ERVisits)) +
  geom_boxplot() +
  labs(title = "ERVisits by Gender", x = "Gender (0 = Female, 1 = Male)", y = "ERVisits")

ggplot(df, aes(x = as.factor(Heart_Disease), y = ERVisits)) +
  geom_boxplot() +
  labs(title = "ERVisits by Number of Heart Disease", x = "Number of Heart Disease)", y = "ERVisits")
```



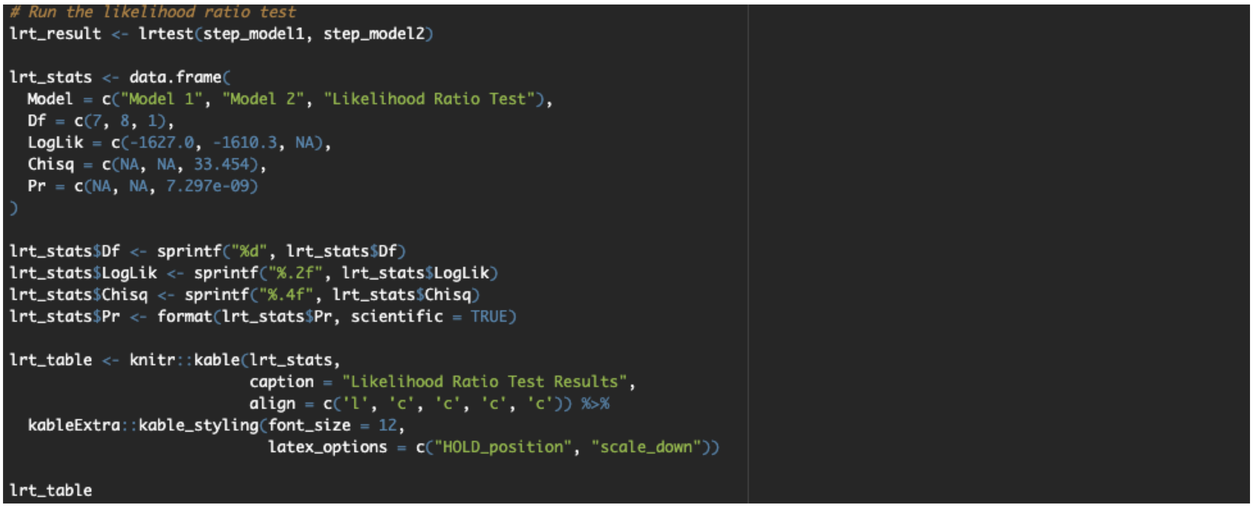
```r
# Modeling

# 3.1. Remove ID column -> To change Heart_Disease column as binary factor (+ Mostly 0~1)
df$ID <- NULL

# 3.2. Min-Max scailing for numeric variables
min_max_scaling <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}
selected_vars <- c("TotalCost","Age", "Interventions","Drug", "Other_Diseases","Duration")
df[selected_vars] <- lapply(df[selected_vars], min_max_scaling)

# 3.3. Poisson model and stepwise selection
model.pois<-glm(ERVisits~., family=poisson, data=df)
step_model1 <- stepAIC(model.pois, direction = "both")
summary(step_model1)

# 3.4. Negative binomial model and stepwise selection
model.nb<-glm.nb(ERVisits~., data=df)
step_model2 <- stepAIC(model.nb, direction = "both")
summary(step_model2)

# 3.5. goodness of fit - LRT : Result interpretation - help
library(lmtest)
lrtest(step_model1,step_model2)

# 3.6.  Check VIF
library(car)
vif(step_model2)

# 3,8 Check Assumption - residuals
plot(resid(model.pois, type = "pearson") ~ fitted.values(model.pois),
     ylab = "Pearson Residuals", xlab = "Fitted Values",
     main = "Residuals vs Fitted Values")
abline(h = 0, col = "red")
```



```
# Model Chart

## Possion Regression
model.pois<-glm(ERVisits~., family=poisson, data=df)
summary(model.pois)

step_model1 <- stepAIC(model.pois, direction = "both")
summary(step_model1)

# Extract coefficients and relevant statistics from the model
model_summary <- summary(step_model1)$coefficients

model_df <- as.data.frame(model_summary)


colnames(model_df) <- c("Estimate", "Std. Error", "z value", "Pr(>|z|)")

knitr::kable(model_df,
             caption = "Summary of Poisson Regression Model",
             digits = c(4, 4, 4, 8),  # Setting 8 decimal places for p-value
             align = c('l', 'c', 'c', 'c', 'c')) %>%
  kableExtra::kable_styling(font_size = 12,
                            latex_options = c("HOLD_position", "scale_down"))

model_summary <- summary(step_model1)

model_df <- as.data.frame(model_summary$coefficients)

colnames(model_df) <- c("Estimate", "Std. Error", "z value", "Pr(>|z|)")

model_df$"Pr(>|z|)" <- sprintf("%.6f", model_df$"Pr(>|z|)")

model_df$Null_Deviance <- ifelse(row.names(model_df) == "(Intercept)", model_summary$null.deviance, "")
model_df$Residual_Deviance <- ifelse(row.names(model_df) == "(Intercept)", model_summary$deviance, "")
model_df$AIC <- ifelse(row.names(model_df) == "(Intercept)", model_summary$aic, "")
model_df$Dispersion <- ifelse(row.names(model_df) == "(Intercept)", model_summary$dispersion, "")

kable_output <- knitr::kable(model_df,
                             caption = "Summary of Poisson Regression Model",
                             digits = c(4, 4, 4, 8, 4, 4, 4, 4),
                             align = c('l', 'c', 'c', 'c', 'c', 'c', 'c', 'c')) %>%
  kableExtra::kable_styling(font_size = 12,
                            latex_options = c("HOLD_position", "scale_down")) %>%
  kableExtra::column_spec(column = 5:8, border_left = TRUE)

kable_output
```



```
## Negative Binomial Regression
model.nb<-glm.nb(ERVisits~., data=df)
summary(model.nb)

step_model2 <- stepAIC(model.nb, direction = "both")
summary(step_model2)

model_summary_nb <- summary(step_model2)

model_df_nb <- as.data.frame(model_summary_nb$coefficients)

colnames(model_df_nb) <- c("Estimate", "Std. Error", "z value", "Pr(>|z|)")

model_df_nb$"Pr(>|z|)" <- sprintf("%.8f", model_df_nb$"Pr(>|z|)")

model_df_nb$Null_Deviance <- ifelse(row.names(model_df_nb) == "(Intercept)", sprintf("%.4f", model_summary_nb$null.deviance), "")
model_df_nb$Residual_Deviance <- ifelse(row.names(model_df_nb) == "(Intercept)", sprintf("%.4f", model_summary_nb$deviance), "")
model_df_nb$AIC <- ifelse(row.names(model_df_nb) == "(Intercept)", sprintf("%.4f", model_summary_nb$aic), "")
model_df_nb$Theta <- ifelse(row.names(model_df_nb) == "(Intercept)", sprintf("%.4f", model_summary_nb$theta), "")
model_df_nb$Std_Err_Theta <- ifelse(row.names(model_df_nb) == "(Intercept)", sprintf("%.4f", sqrt(model_summary_nb$dispersion)), "")

kable_output_nb <- knitr::kable(model_df_nb,
                                caption = "Summary of Negative Binomial Regression Model",
                                digits = c(4, 4, 4, 8, 4, 4, 4, 4, 4),
                                align = c('l', 'c', 'c', 'c', 'c', 'c', 'c', 'c', 'c')) %>%
  kableExtra::kable_styling(font_size = 12,
                            latex_options = c("HOLD_position", "scale_down")) %>%
  kableExtra::column_spec(column = 5:9, border_left = TRUE)

kable_output_nb
```



```r
# Run the likelihood ratio test
lrt_result <- lrtest(step_model1, step_model2)

lrt_stats <- data.frame(
  Model = c("Model 1", "Model 2", "Likelihood Ratio Test"),
  Df = c(7, 8, 1),
  LogLik = c(-1627.0, -1610.3, NA),
  Chisq = c(NA, NA, 33.454),
  Pr = c(NA, NA, 7.297e-09)
)

lrt_stats$Df <- sprintf("%d", lrt_stats$Df)
lrt_stats$LogLik <- sprintf("%.2f", lrt_stats$LogLik)
lrt_stats$Chisq <- sprintf("%.4f", lrt_stats$Chisq)
lrt_stats$Pr <- format(lrt_stats$Pr, scientific = TRUE)

lrt_table <- knitr::kable(lrt_stats,
                          caption = "Likelihood Ratio Test Results",
                          align = c('l', 'c', 'c', 'c', 'c')) %>%
  kableExtra::kable_styling(font_size = 12,
                            latex_options = c("HOLD_position", "scale_down"))

lrt_table
```