Predicting Hospital Utilization - A Generalized Linear Model Approach

Shinjon Ghosh

Department of Mathematics, Illinois State University

Abstract:

Hospital Utilization is an important aspect of healthcare management. It directly reflects the demand for healthcare services and the overall functioning of the healthcare system. This study proposed the application of Generalized Linear Models (GLMs) to predict and identify factors contributing to hospital visits. Developing a robust model that can anticipate healthcare strategies. The analysis involves building GLMs with a Poisson log link function to predict count outcomes, such as the number of doctor visits. By examining a range of features, the aim to uncover patterns and relationships that improve patient care. The results highlight significant predictors (age, gender, income, private, illness, reduced, freepoor) of hospital visits and offer actionable insights for developing targeted healthcare strategies. Moreover, the model results suggest that the proposed GLMs fitted well and are adequate according to the Doctor Visits dataset. These findings can optimize resource allocation, manage healthcare costs, and contribute to the development of targeted interventions for reducing unnecessary hospital utilization.

Keywords: generalized linear models, machine learning model, poisson glms, negative binomial glms, random forest models, log link function, interaction variable, higher order variable, likelihood ratio test, goodness of fit test, split dataset.

Introduction:

Hospital utilization is a key measure of healthcare system performance, reflecting the frequency and extent to which individuals access hospital services. High utilization rates may indicate an increased burden on healthcare facilities, leading to overcrowding, longer wait times, and rising healthcare expenditures. Conversely, low utilization rates may signal barriers to healthcare access, such as financial constraints, lack of insurance, or geographic disparities. Several factors influence hospital utilization, including demographic characteristics, socioeconomic status, health conditions, and healthcare accessibility. Age, income level, insurance coverage, and chronic illnesses are among the most important predictors of hospital visits. Identifying these factors can help optimize healthcare planning and develop strategies to improve efficiency and accessibility while minimizing unnecessary hospital admissions. According to a commonwealth report, the U.S. allocates a smaller proportion of its total healthcare spending to primary care compared to other high-income countries. In 2021, only 4.7% of healthcare expenditures were directed toward primary care, whereas other high-income nations averaged 14%. This disparity underscores potential challenges in the accessibility and emphasis on primary care services within the U.S. healthcare system. By comparing model performance and identifying significant predictors, the findings will provide insights into the key determinants of hospital utilization. Furthermore, this research will offer policy recommendations to enhance healthcare accessibility and optimize hospital resource management, ultimately contributing to a more efficient and equitable healthcare system.

Background & Literature Review:

The OMOP CDM (Observational Medical Outcomes Partnership Common Data Model) in predicting hospital LoS (length of stay) for planned admissions demonstrates promising predictive capabilities for varying durations and highlights the advantage of standardized data for achieving reproducible results. This approach could serve as a model for enhancing operational efficiency and patient care coordination across healthcare settings. The study also identified key factors influencing LoS, such as surgical operations, patient demographics, and admission characteristics, which can inform strategies for optimizing healthcare resource utilization [1]. Both ML and statistical modeling showed good predictive performance for LOS. Predictive performance was often reported using the Area Under the Receiver Operating Curve (AUROC). The median AUROC values indicated fair-to-good discriminative ability. However, calibration metrics were reported infrequently [3]. The study concludes that RQRs (Randomized Quantile Residuals) offer advantages over traditional Pearson and deviance residuals for diagnosing count regression models, including zero-inflated models. RQRs are approximately standard normally distributed under a correctly specified model, and the SW (Shapiro-Wilk) normality test of RQRs serves as a well-calibrated overall GOF test. The graphical analysis of RQRs against covariates can also provide insights into the nature of model misspecification. While the randomization in RQRs introduces some fluctuation, analyzing multiple realizations can mitigate this [6]. The study concludes that the newly introduced CNBLD (conflation of negative binomial and logarithmic distributions), SCNBLD (shifted conflation of negative binomial and logarithmic distributions, and CNBSLD (conflation of negative binomial shift logarithmic distributions) provide flexible alternatives for modeling count data, particularly when dealing with over-dispersion and an excess of low counts, including zeros (for the modified versions). Their superior performance on real datasets suggests their potential as proficient substitutes for existing models in various fields [7]. The Lasso Regression model provides a feasible approach to predict high acute care utilization in a safety net hospital system by leveraging EHR data and incorporating critical social determinants of health. The model's focus on the top 1% of predicted high utilizers with a strong PPV (positive predictive value) enhances its clinical utility for targeted interventions [2]. The impacts of the CJR (care for joint replacement) program differed by hospital ownership type. Notably, government-owned hospitals did not show a reduction in inpatient length of stay as observed in nonprofit and for-profit hospitals. The authors suggest that the unique financial circumstances and potential challenges faced by government-owned hospitals, such as bureaucratic processes, resource limitations, and soft budget constraints, may have hindered their ability to adapt to the CJR model. The study highlights the need to consider the specific circumstances of government-owned hospitals when implementing bundled payment models [4].

Objectives:

- Fit and compare GLMs & Machine Learning models to predict hospital visits.
- Identification of significant predictors and their impact on hospital utilization.
- Interpret model results to offer policy recommendations for improving healthcare accessibility.

Methods & Materials:

Data Source & Overview

The dataset used in this study is the DoctorVisits dataset from the AER (Applied Econometrics with R) package. It contains individual-level data on the number of doctor visits, along with demographic, economic, and health-related variables. The dataset contains 5190 observations and 12 features. Among 12 variables private [Indicator of private health insurance (factor: "yes" = 2 or "no" = 1)], freepoor [Indicator for eligibility for free government healthcare (factor: "yes" = 2 or "no" = 1)], freerepat [individual have free government health insurance due to old age, disability or veteran status (factor: “yes” = 2 or “no” = 1)], nchronic [chronic condition not limiting activity (factor: “yes” = 2 or “no” = 1)], lchronic [Chronic condition limiting activity (factor: “yes” = 2 or “no” = 1)], gender (Male & Female) are categorical features and age, income (Annual income in tens of thousands of dollars), illness (Number of illnesses in past 2 weeks), reduced (Number of days of reduced activity in past 2 weeks due to illness or injury), health (General health questionnaire score using Goldberg's method) are numerical variables. Moreover, there is a count variable visits, which indicates the number of doctor visits within the past two weeks. Each row corresponds to a specific user. This dataset allows for the analysis of how demographic, economic, and health-related factors influence hospital utilization. It is a useful resource for statistical modeling, hypothesis testing, and developing strategies for policymakers, healthcare providers, and insurance companies to optimize resource allocation, improve patient care, and manage healthcare costs.

Data Preparation

By applying Generalized Linear Models (GLMs), this study aims to uncover significant predictors of hospital visits and provide insights into improving healthcare accessibility and resource management. In the dataset, there are no missing values. After that, I use the factor function for factorizing Gender, Private, Freepoor, Freerepat, nchronic, and lchronic variables. Later, I define visits (number of doctor visits within past two weeks) as a response variable. On the other hand, I set categorical and numerical variable as predictor or explanatory variables according to dataset. Moreover, I split the dataset by 80-20 rules for model evaluation and measured performance metrics. The training set (80% = 4152 observations) is used to build and learn the patterns in the data. On the other hand, The testing set (20% = 1038 observations) is used to evaluate the model's performance on unseen data.

Result & Discussion:

Exploratory Data Analysis

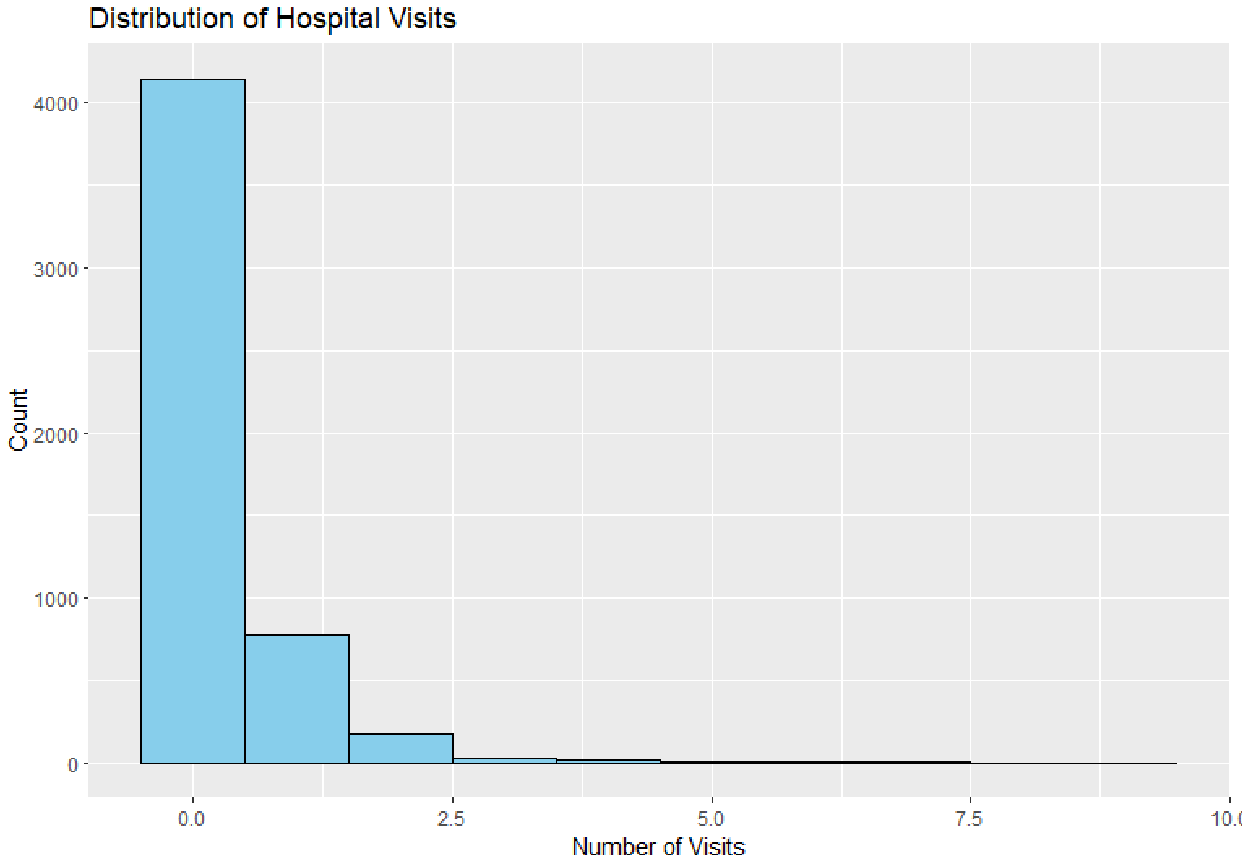

Fig 1: Distribution of Doctor Visits

The plot shows the distribution of doctor visits, specifically the number of visits per individual. From the histogram analysis, we can say that the distribution is heavily right-skewed, indicating that most individuals have zero or very few hospital visits. The highest bar corresponds to individuals who did not visit the hospital at all, suggesting that a large proportion of the population did not utilize hospital services during the study period. Moreover, A small number of individuals had 1 or more doctor visits, and very few had more than 3 visits. In addition, beyond 5 visits, the counts are almost negligible. According to Fig 1, we can say that this type of data is typical for count models, such as Poisson regression or negative binomial regression, which can handle skewed count data effectively.

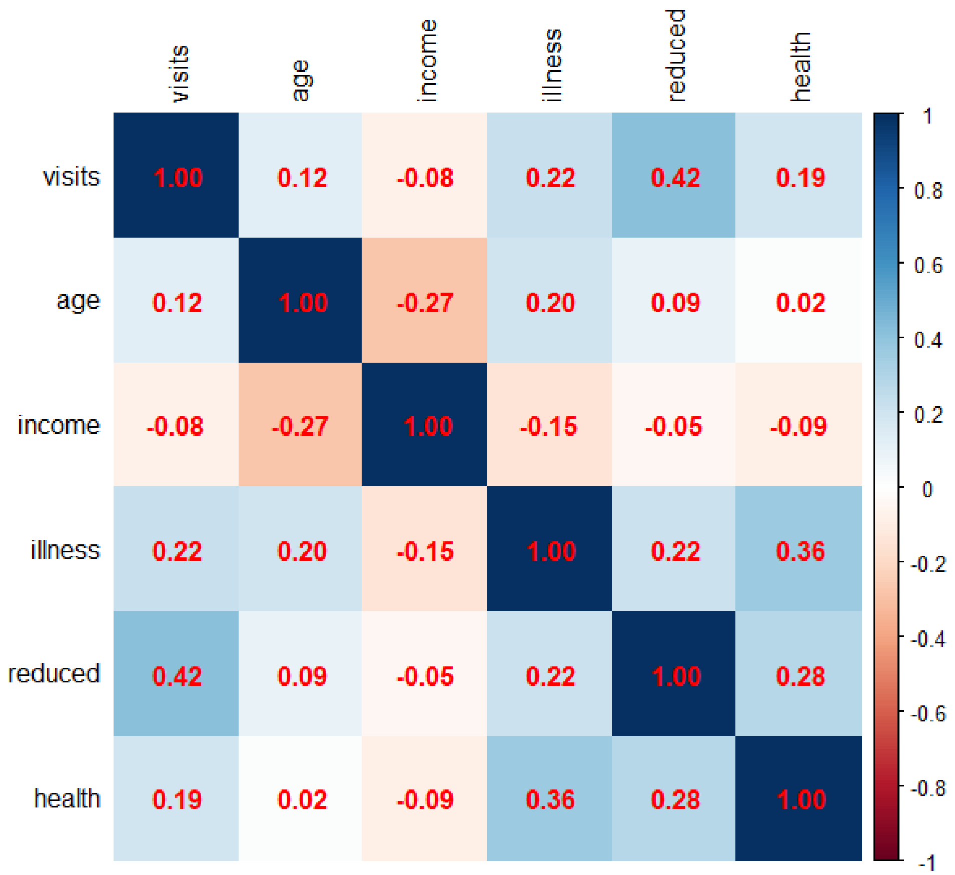

Fig 2: Correlation between Numeric Variables

This correlation plot visualizes the relationships between variables: visits, age, income, illness, reduced, and health. The values range from -1 (strong negative correlation) to +1 (strong positive correlation), with color intensity representing the strength of the correlations. A moderate positive correlation of illness (0.22) indicates that individuals with more illnesses tend to have more hospital visits. Moreover, a highly moderate positive correlation of reduced (0.42) suggests that reduced activity in the past 2 weeks due to illness or injury is highly associated with increased hospital utilization. On the other hand, a weak positive correlation of age (0.12) and health (0.19) illustrates that poorer health and older individuals are slightly more likely to visit hospitals. In addition, A weak negative correlation of income (-0.08) suggests that higher income is slightly associated with fewer doctor visits. According to the Fig 2 we can say that, reduced, health, and illness show moderate correlations with visits, making them important predictors for hospital utilization management.

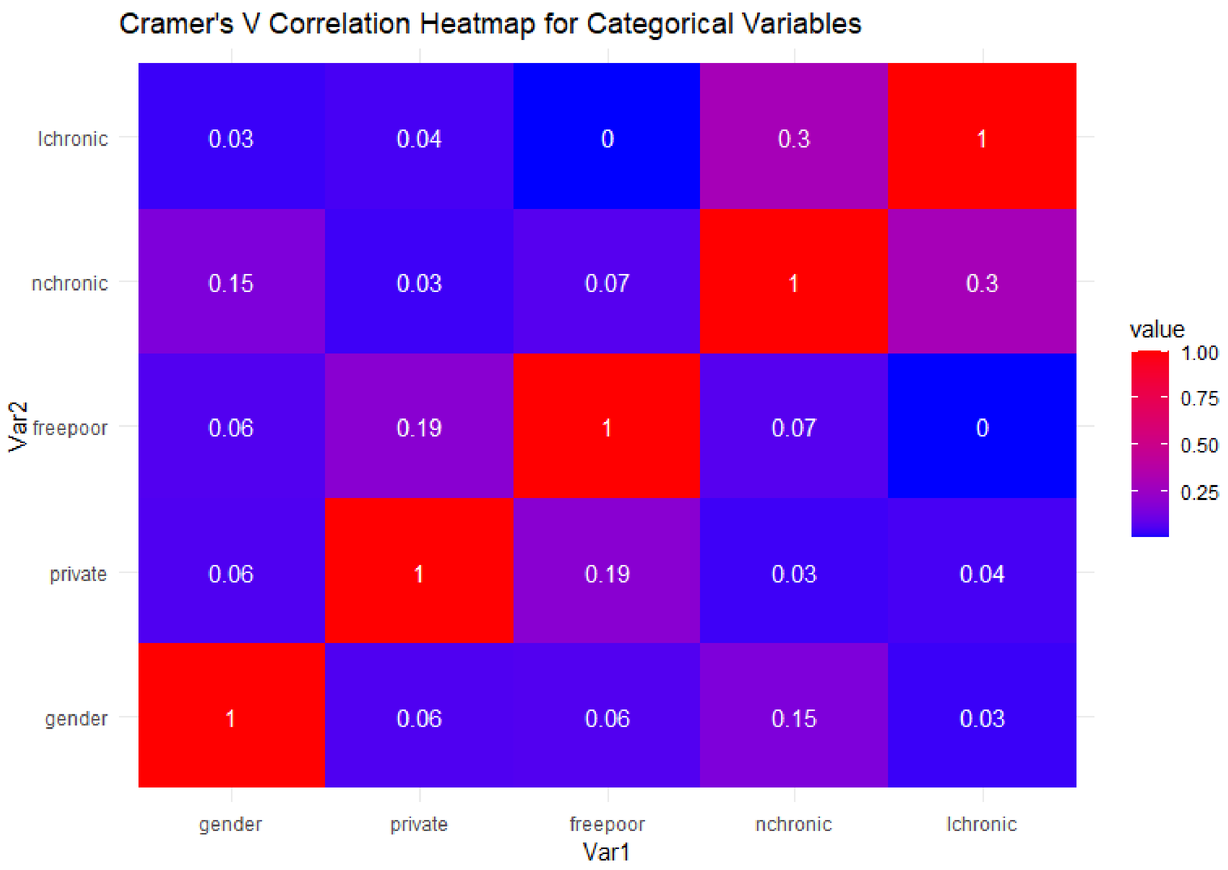

Fig 3: Correlation Heatmap for Categorical Variables

This heatmap visualizes the Cramér's V correlation coefficients between pairs of categorical variables: gender, private, freepoor, nchronic, and lchronic. Cramér's V measures the strength of association between two categorical variables, ranging from 0 (no association) to 1 (perfect association). The color gradient represents the magnitude of the correlation, with red indicating strong correlations and blue indicating weak correlations. According to the plot, we can say that there is a moderate relationship between freepoor with private and lchronic with nchronic variables. On contrary, gender has weak correlation with other variables. From Fig 3, we can say that the lack of strong associations across most variable pairs highlights that these categorical features are relatively independent, which can be advantageous for modeling purposes as it reduces multicollinearity concerns.

Generalized Linear Models

From processed Doctor Visits dataset, I use visits as a response variable which is count data (number of doctor visits in past 2 weeks). That’s why I set and run several poisson generalized linear models according to different predictor variables with log link functions.

Poisson Model : $\log(E(yi)) = \log(\mu i) = \beta 0 + \beta 1Xi1 + \cdots + \beta pXip$

where, $y_i = Response\ variable$

$\beta_0 = Intercept$

$\beta_1 \ldots.\beta_p = Slopes$

$X_{i1} \ldots..X_{ip} = Predictor\ variables$

In Table 1, Model 1 is a full model with all predictor variables (age, gender, health, reduced, private, freepoor, freerepat, lchronic, nchronic, income, and illness). On the other hand, Model 2 is a reduced model with limited predictor variables (gender, illness, reduced, health, freepoor, lchronic). Moreover, Models 4, 5, & 6 are forward regression, backward elimination, stepwise regression models, respectively. Furthermore, Model 3 is based on interaction predictor variables (age*gender, income*private, illness, reduced, health, freepoor, lchronic, nchronic). In addition, Model 7 is based on higher order with interaction predictor variables ($age^2, income^2 * private^3$, $nchronic^2$, gender, illness, reduced, health, freepoor, lchronic).

Model	Fitted Model in Details	AIC	Deviance	BIC
Model – 1 (Full Model)	glm(formula = visits ~ . , family = poisson(link = “log”), data = train_data)	5324.2	3432.3	5400.164
Model – 2 (Reduced Model)	glm(formula = visits ~ gender + illness + reduced + health + freepoor + lchronic, family = poisson(link = “log”), data = train_data)	5336	3454.1	5380.278
Model – 3 (Interaction Variables Model)	glm(formula = visits ~ age * gender + income * private + illness + reduced + health + freepoor + nchronic + lchronic, family = poisson(link = “log”), data = train_data)	5309.9	3416	5392.191
Model – 4 (Forward Model)	glm(formula = visits ~ reduced + illness + age + health + gender + freepoor, family = poisson(link = “log”), data = train_data)	5325.4	3443.6	5331.3

Model – 5 (Backward Model)	glm(formula = visits ~ gender + income + illness + reduced + health + private + freepoor + freerepat + nchronic + lchronic, family = poisson(link = “log”), data = train_data)	5323.8	3434	5329.8
Model – 6 (Step-wise Model)	glm(formula = visits ~ gender + income + illness + reduced + health + private + freepoor + freerepat + nchronic + lchronic, family = poisson(link = “log”), data = train_data)	5323.8	3434	5329.8
Model – 7 (Higher order with Interaction variables model)	glm(formula = visits ~ age^2 + gender + income^2 * private^3 + illness + reduced + health + freepoor + nchronic^2 + lchronic, family = poisson(link = “log”), data = train_data)	5316.6	3424.7	5392.541

Table 1: Model comparisons by AIC, Deviance & BIC

Now, I check the better-fitted model based on AIC (Akaike Information Criterion), Residual Deviance, and BIC (Bayesian Information Criterion).

$AIC = n \log(SSEp/n) + 2p$

$BIC = n \log(SSEp/n) + p\log(n)$

Where, n = number of observations

$SSEp$ = Residual Sum Squared Error of the model

p = number of estimated parameters

We know that lower AIC, lower BIC, and lower deviance indicate a better-fitted model. From Table 1, we can say that Model 3, which is fitted by interaction predictor variables, illustrates lower deviance (3416) and lower AIC (5309.9) compared with other models. On the contrary, Model 5 (Backward elimination), and model 6 (stepwise selection) produced lower BIC, which is

5329.8. However, these two models do not capture interaction variables. In that case, I select Model 3 (interaction predictor variables) is a better-fitted model.

Likelihood Ratio Tests to Compare Nested Models:

The Likelihood Ratio Test (LRT) compares nested models to assess whether the additional predictors in the more complex model significantly improve model fit. It does this by testing whether the reduction in deviance is statistically significant.

Null Hypothesis (H_0) : The additional parameters in the full model do not significantly improve the model fit. This means the extra predictors in the full model have no effect.

Alternative Hypothesis (H_a) : There are significant effect of additional predictor variables. So, the extra predictors improve the model's explanatory power.

```
Model 1: visits ~ age + gender + income + illness + reduced + health +
    private + freepoor + freerepat + nchronic + lchronic

Model 2: visits ~ gender + illness + reduced + health + freepoor + lchronic

Model 3: visits ~ age * gender + income * private + illness + reduced +
    health + freepoor + nchronic + lchronic

Model 4: visits ~ reduced + illness + age + health + gender + freepoor

Model 5: visits ~ gender + income + illness + reduced + health + private +
    freepoor + freerepat + nchronic + lchronic

Model 6: visits ~ age^2 + gender + income^2 * private^3 + illness + reduced +
    health + freepoor + nchronic^2 + lchronic

  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      4140     3432.3
2      4145     3454.1 -5  -21.771 0.0005788 ***
3      4139     3416.0  6   38.075 1.086e-06 ***
4      4145     3443.6 -6  -27.540 0.0001147 ***
5      4141     3434.0  4    9.586 0.0480015 *
6      4140     3424.7  1    9.273 0.0023255 *
```

According to the decision rule, if the p-value is small (p-value < 0.05), we reject null hypothesis, meaning there are significant effects of additional predictor variables. So, the extra predictors improve the model's explanatory power. Here, all of the additional predictor variables' p-value is less than 0.05. Thus, we can say the additional predictor variables fits significantly better than full model (Model 1), meaning interaction terms improve model performance. Meanwhile, Model 3 (interaction predictor variables) p-value is lowest (1.086e-06) than other models. That means model 3 predictor variables significantly more improved the model performance.

Overdispersion:

Overdispersion occurs in count data models (such as Poisson regression) when the variance exceeds the mean, violating the assumption of equidispersion Var(Y)=E(Y) in the Poisson model. According to the decision rules, if Dispersion > 1, the data exhibit overdispersion.

Dispersion Ratio	1.272289

Table 2: Dispersion Ratio

From Table 2, we can say that Model 3 dispersion ratio is 1.28, and it has an overdispersion issue. Now, I use the Negative Binomial generalized linear model to fix overdispersion because it allows the variance to be greater than the mean by introducing a dispersion parameter.

Negative Binomial GLMs:

Because of the overdispersion issue, I run Negative Binomial GLMs with interaction predictor variables.

Model	AIC	Deviance	BIC	Dispersion Ratio
Negative Binomial	5110.5	2460.5	5199.134	0.9920598

Table 3: Negative Binomial Model Performance

From Table 1 and Table 3, we can say that the Negative Binomial model has lower AIC (5110) and BIC (5199.134) which is a better fit compared to the Poisson models. Moreover, much lower deviance (2460.5) suggests a significantly improved goodness-of-fit. In addition, Dispersion ratio close to 1 (ϕ=0.992) indicates that the model variance is now correctly estimated, resolving the overdispersion issue.

Goodness of Fit test

For the generalized linear model with Negative Binomial, I use the chi-square test and McFadden's Pseudo R-Squared for checking the goodness-of-fit of the model. A low p-value (typically < 0.05) suggests rejection of the null hypothesis, indicating poor fit. A high p-value suggests that the model fits the data well.

p-value	1
McFadden’s Pseudo R-Squared	100%

Table 4 : Goodness of Fit Test

According to Table 4, we can say that the chi-square test p-valus is 1, indicating that response variable (visits) is perfectly explained by the model. Moreover, the McFadden’s Pseudo R-Squared is 100%. It means 100% of the variation in the response y is explained by the model.

Negative Binomial Model Summary:

Coefficients	Estimate	p-value
Age	0.975418	0.000999
GenderFemale	0.616668	0.000293
Income	-0.537322	0.003057
PrivateYes	-0.307022	0.025958
Illness	0.188026	4.87e-13
Reduced	0.140732	< 2e-16
Health	0.046821	0.001558
FreepoorYes	-0.651108	0.005449
nchronicYes	0.132838	0.126492
lchronicYes	0.193847	0.086889
Age : GenderFemale	-0.922826	0.008926
Income : PrivateYes	0.644924	0.003068

Table 5 : Model Summary

From this model summary, we can say that all of the predictor variables are significant without nchronic and lchronic variables according to p-values. For each one-unit increase in age, the expected number of hospital visits increases by a log of 0.9754 unit, when other factors are constant. The p-value of age (0.000999) indicates that it has strongly influenced the number of visits. Moreover, individuals with illness, reduce, and health have an expected increase log of 0.1880, 0.140732, & 0.0468 units in hospital visits with holding other factors, respectively. The small p-value shows that illness, reduce, and health are highly significant factors influencing doctor visits. Furthermore, one unit increase of genderfemale, the expected number of visits increases by a log of 0.6167 unit compared to gendermale, when other factor are constant. In addition, one unit increase of nchronicyes and lchronicyes, the expected number of doctor visits increases by a log of 0.1328 and 0.1938 unit with holding other factors. This p-value of nchronic and lchronic is not significant (greater than 0.05), so nchronic and lchronic is not significant predictor of hospital visits in this model. On the other hand, for each one-unit increase in income, the expected number of hospital visits decreases by a log of 0.5373 units when other factors are constant. Moreover, individuals with private health insurance have log of 0.307 fewer hospital visits than those without private insurance and individual is from a poorer background (compared to others), their expected number of hospital visits decreases by a log of 0.6511 unit, with holding other factors. Later, the interaction terms (age:genderfemale, income:privateyes) have significant effect in this model based on lower p-value.

Model Prediction & Validation:

Both Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are common evaluation metrics for regression models. They measure how well a model's predictions match the actual values. Lower RMSE and lower MAE suggest better prediction performance.

Model	RMSE		MAE	
	Train	Test	Train	Test
Negative Binomial	0.7395115	0.8504019	0.4163303	0.4458986
Random Forest	0.4694575	0.8315878	0.2753708	0.4297475

Table 6: Model Predictions by RMSE & MAE

According to Table 6, the Random Forest test dataset model RMSE (0.832) and MAE (0.429) is slightly lower than the Negative Binomial Model test dataset RMSE (0.850) and MAE (0.446). Moreover, the negative binomial model train dataset RMSE (0.739) and MAE (0.416) is closer to test dataset model RMSE (0.850) and MAE (0.416). Thus, the negative binomial model generalizes well. It performs equally well on both training and unseen (test) data. On the other hand, the random forest model train dataset RMSE (0.469) and MAE (0.275) are significant lower than test dataset model RMSE (0.832) and MAE (0.429). So, the model has an overfitting issue and performs well on training data but poorly on test data.

Limitations:

The DoctorVisits dataset comes from the 1977–1978 Australian Health Survey, which is a limitation of this study. Healthcare policies, insurance coverage, and accessibility have significantly changed since 1978. The doctor visit patterns from that era may not reflect the current healthcare landscape. Inflation, income distribution, and the cost of healthcare have shifted over time. The dataset does not account for modern factors like telemedicine, private health insurance expansion, or government health subsidies. New treatments, diagnostic techniques, and disease prevalence have changed over the decades. Chronic diseases and healthcare utilization patterns are different now.

Conclusion:

This project aims to leverage predictive analytics to forecast doctor visits and provide actionable insights for hospital management strategies. Specifically, constructing an optimum generalized linear model with the help of R language to predict hospital visits. According to the EDA analysis, the distribution is heavily right-skewed. It indicates data is typical for count models, such as Poisson regression or negative binomial regression, which can handle skewed count data effectively. Moreover, I found a better generalized linear model by AIC when the response variable is visits and the predictor variables are age*gender, income*private, illness, reduced, health, freepoor, lchronic, and nchronic. According to the likelihood ratio test, the additional interaction predictor variables model (Model 3) performance improves better than the full model. In addition, there is not any overdispersion issue in the Negative Binomial model. The goodness-of-fit test indicates the response variable is perfectly explained by the model. From the Negative Binomial model summary, we can say that all of the predictor variables (age, genderfemale, privateyes, illness, reduce, health, income, freepooryes, age:genderfemale, and income:privateyes) are significant without nchronic and lchronic variables according to p-values. According to RMSE and MAE, we can say that the negative binomial model generalizes well than the Random Forest model. It performs equally well on both training and unseen (test) data. Overall, these results offer a guideline for predicting doctor visits with influential predictor variables and enhance hospital management strategies.

Reference:

[1] Lee, H., Kim, S., Moon, H., Lee, H., Kim, K., Jung, S., & Yoo, S. (2024). Hospital length of stay prediction for planned admissions using observational medical outcomes partnership common data model: Retrospective study. Journal of Medical Internet Research, 26, e59260. https://doi.org/10.2196/59260.

[2] Li, Z., Gogia, S., Tatem, K. S., Cooke, C., Singer, J., Chokshi, D. A., & Newton-Dame, R. (2023). Developing a Model to Predict High Health Care Utilization Among Patients in a New York City Safety Net System. Medical care, 61(2), 102–108. https://doi.org/10.1097/MLR.0000000000001807.

[3] Gokhale, S., Taylor, D., Gill, J., Hu, Y., Zeps, N., Lequertier, V., Prado, L., Teede, H., & Enticott, J. (2023). Hospital length of stay prediction tools for all hospital admissions and general medicine populations: systematic review and meta-analysis. Frontiers in medicine, 10, 1192969. https://doi.org/10.3389/fmed.2023.1192969.

[4] Kim, N., & Jacobson, M. (2024). Medicare's comprehensive care for joint replacement model increased public hospitals' inpatient length of stay. BMC health services research, 24(1), 1495. https://doi.org/10.1186/s12913-024-11905-0.

[5] Gupta, A., Hall, M., Masserano, B., Wilson, A., Johnson, K., Chen, C., Challa, L., Katragadda, H., & Mittal, V. (2025). Trends in resource utilization for new-onset psychosis hospitalizations at children's hospitals. Journal of hospital medicine, 10.1002/jhm.13597. Advance online publication. https://doi.org/10.1002/jhm.13597.

[6] Feng, C., Li, L., & Sadeghpour, A. (2020). A comparison of residual diagnosis tools for diagnosing regression models for count data. BMC medical research methodology, 20(1), 175. https://doi.org/10.1186/s12874-020-01055-2.

[7] Alqefari, A. A., Alzaid, A. A., & Qarmalah, N. (2024). On the Conflation of Negative Binomial and Logarithmic Distributions. Axioms, 13(10), 707. https://doi.org/10.3390/axioms13100707.

[8] Cameron, A. C., & Trivedi, P. K. (1986). Econometric models based on count data: Comparisons and applications of some estimators and tests. Journal of Applied Econometrics, 1(1), 29–53.

[9] Cameron, A. C., & Trivedi, P. K. (1998). Regression analysis of count data. Cambridge University Press.

[10] Mullahy, J. (1997). Heterogeneity, excess zeros, and the structure of count data models. Journal of Applied Econometrics, 12(3), 337–350.

Appendix:

R Code:

```
##Install Packages
install.packages("AER")
library(survival)
library(sandwich)
library(AER)

#Load Dataset
data("DoctorVisits")
head(DoctorVisits)
names(DoctorVisits)
str(DoctorVisits)

# Factorize categorical variables
DoctorVisits$freepoor  <- as.numeric(as.factor(DoctorVisits$freepoor))  # "no" -> 1, "yes" -> 2
DoctorVisits$freerepat <- as.numeric(as.factor(DoctorVisits$freerepat)) # "no" -> 1, "yes" -> 2
DoctorVisits$private   <- as.numeric(as.factor(DoctorVisits$private))   # "no" -> 1, "yes" -> 2
DoctorVisits$gender    <- as.numeric(as.factor(DoctorVisits$gender))    # "male" -> 1, "female" -> 2
DoctorVisits$nchronic <- as.numeric(as.factor(DoctorVisits$nchronic))   # "no" -> 1, "yes" -> 2
DoctorVisits$lchronic <- as.numeric(as.factor(DoctorVisits$lchronic))   # "no" -> 1, "yes" -> 2

#Check null value
colSums(is.na(DoctorVisits))

# Load visualization libraries
install.packages("ggplot2")
library(ggplot2)
library(skimr)   # For an overall dataset summary
library(corrplot) # For correlation plots
library(dplyr)

# Histogram of hospital visits
ggplot(DoctorVisits, aes(x = visits)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Hospital Visits", x = "Number of Visits", y = "Count")

# Correlation Plot for numeric variables
numeric_vars <- DoctorVisits %>%
  select_if(is.numeric)
cor_matrix <- cor(numeric_vars, use = "complete.obs")
corrplot(cor_matrix, method = "color", addCoef.col = "black", tl.col = "black")

# Correlation Heatmap for Categorical Variables
library(rcompanion)
library(vcd)

categorical_vars <- c("gender", "private", "freepoor", "nchronic", "lchronic")
cramer_matrix <- matrix(NA, nrow = length(categorical_vars), ncol = length(categorical_vars))
rownames(cramer_matrix) <- categorical_vars
colnames(cramer_matrix) <- categorical_vars

for (i in 1:length(categorical_vars)) {
  for (j in 1:length(categorical_vars)) {
    if (i == j) {
      cramer_matrix[i, j] <- 1  # Perfect correlation with itself
    } else {
      table_temp <- table(DoctorVisits[[categorical_vars[i]]], DoctorVisits[[categorical_vars[j]]])
      cramer_matrix[i, j] <- assocstats(table_temp)$cramer
    }
  }
}

print(cramer_matrix)
library(reshape2)
melted_cramer <- melt(cramer_matrix)

ggplot(melted_cramer, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  geom_text(aes(label = round(value, 2)), color = "white") +
  scale_fill_gradient(low = "blue", high = "red") +
  labs(title = "Cramer's V Correlation Heatmap for Categorical Variables") +
  theme_minimal()

##split the dataset 80% and 20%
install.packages("caret")
library(ggplot2)
library(lattice)
library(caret)

set.seed(123)

n <- nrow(DoctorVisits)
train_indices <- sample(1:n, size = 0.8 * n)
```

```
train_data <- DoctorVisits[train_indices, ]
test_data  <- DoctorVisits[-train_indices, ]

dim(train_data)
dim(test_data)

#Model Evaluation
dv1 <- glm(visits ~ age + gender + income + illness + reduced + health + private + freepoor + freerepat + nchronic + lchronic, data = train_data, family = poisson(link = "log"))
summary(dv1)
BIC1 <- BIC(dv1)
print(BIC1)



dv2 <- glm(visits ~ gender + illness + reduced + health + freepoor + lchronic, data = train_data, family = poisson(link = "log"))
summary (dv2)
BIC2 <- BIC(dv2)
print(BIC2)

dv3 <- glm(visits ~ age * gender + income*private + illness + reduced + health + freepoor +nchronic + lchronic, data = train_data, family = poisson(link = "log"))
summary(dv3)
BIC3 <- BIC(dv3)
print(BIC3)
```

```
library(MASS)
#Forward, Backward, & Stepwise
# Fit the null model (no predictors)
null_model <- glm(visits ~ 1, family = poisson(link = "log"), data = train_data)
dv4 <- step(null_model, scope = list(lower = null_model, upper = dv1),
   direction = "forward")
summary(dv4)

# Calculate BIC
forward_model <- step(null_model, scope = list(lower = null_model, upper = dv1),
                      direction = "forward", k = log(nrow(DoctorVisits)))
summary(forward_model)

#Backward
dv5 <- step(dv1, direction = "backward")
summary(dv5)

#Calculate BIC
backward_model <- step(dv1, direction = "backward", k = log(nrow(DoctorVisits)))
summary(backward_model)

#Step-wise
dv6 <- step(dv1, direction = "both")
summary(dv6)

#Calculate BIC
stepwise_model <- step(dv1, direction = "both", k = log(nrow(DoctorVisits)))
summary(stepwise_model)
```

```
dv7 <- glm(visits ~ age^2 + gender + income^2 * private^3 + illness + reduced + health + freepoor
        +nchronic^2 + lchronic, data = train_data, family = poisson(link = "log"))
summary(dv7)
BIC7 <- BIC(dv7)
print(BIC7)

# LRT
lrt_result <- anova(dv1, dv2, dv3, dv4, dv5, dv6, dv7, test = "Chisq")
print(lrt_result)

#Overdispersion
dp3 <- sum(residuals(dv3, type = "pearson")^2) / dv3$df.residual
dp3

p_value3 <- pchisq(deviance(dv3), df.residual(dv3), lower.tail = FALSE)
print(p_value3)

#Negative Binomial Model
install.packages("MASS")
library(MASS)
dv8 <- glm.nb(visits ~ age + gender + income + illness + reduced + health + private + freepoor +
freerepat + nchronic + lchronic, data = train_data)
summary (dv8)

dv9 <- glm.nb(visits ~ age * gender + income*private + illness + reduced + health + freepoor
          +nchronic + lchronic, data = train_data)
summary(dv9)
```

```r
deviance_nb <- sum(residuals(dv9, type = "pearson")^2)
df_nb <- dv9$df.residual
dispersion_nb <- deviance_nb / df_nb
dispersion_nb

# Goodness of Fit test
p_value <- pchisq(deviance(dv9), df.residual(dv9), lower.tail = FALSE)
print(p_value)

install.packages("pscl")
library(pscl)
mcfadden_r2 <- pR2(dv9)["McFadden"]
print(mcfadden_r2)

# Model Prediction by RMSE and MAE
# Predict on Train data
train_data$predicted_nb <- predict(dv9, newdata = train_data, type = "response")

# Predict on Test data
test_data$predicted_nb <- predict(dv9, newdata = test_data, type = "response")

install.packages("Metrics")
library(Metrics)
rmse(test_data$visits, test_data$predicted_nb)
rmse(train_data$visits, train_data$predicted_nb)

mae(test_data$visits, test_data$predicted_nb)
mae(train_data$visits, train_data$predicted_nb)
```

```
#Random Forest Model
install.packages("randomForest")
library(randomForest)

# Train Random Forest model
dv10 <- randomForest(visits ~ age * gender + income*private + illness + reduced + health +
freepoor +nchronic + lchronic, data = train_data, ntree = 500, mtry = 3)
summary(dv10)

# Predict on test data
test_data$predicted_rf <- predict(dv10, newdata = test_data)

# Predict on Train data
train_data$predicted_rf <- predict(dv10, newdata = train_data)

# Evaluate RMSE
rmse(test_data$visits, test_data$predicted_rf)
rmse(train_data$visits, train_data$predicted_rf)

mae(test_data$visits, test_data$predicted_rf)
mae(train_data$visits, train_data$predicted_rf)
```