GLM Project: Predicting an Individual's Number of Hospital Visits

Authors: Gabriel Vasquez, Weitong Tie, Rachel Tsai, and Justin Henriquez

STAT 171: General Statistical Models

Dr. Esra Kurum

UC Riverside

Table of Contents

Introduction

The National Health and Nutrition Examination Survey (NHANES), conducted by the Centers for Disease Control and Prevention (CDC), assesses the health and nutritional status of U.S. citizens by collecting medical and demographic data. We analyzed a subset of this survey to develop a generalized linear model which can most accurately predict the number of times an individual has been hospitalized. Our goal was to identify the most significant features in the dataset for inclusion in the model while filtering out less relevant variables.

Data Description

The data set consists of 300 participants. There are no missing values. Each participant is identified by the following eight key features:

- **Hospitalization** (Response Variable) - an integer value representing the number of times an individual has been hospitalized.
- **Age** - an integer value representing an individual's age (years).
- **Gender** - a nominal feature to label an individual as male or female.
- **Smoker** - a nominal feature to label an individual as a smoker or non-smoker.
- **Physically Active** - a nominal feature to label an individual as someone who engages in regular physical activity or not.
- **Diabetes Statu**s - a nominal feature to label an individual as someone who is diabetic or non-diabetic.
- **Cholesterol** - a real number representing an individual's cholesterol level (mg/dL).
- **Body Mass Index (BMI)** - a real number representing an individual's body fat based on height and weight.

Figure 1 Hospitalizations

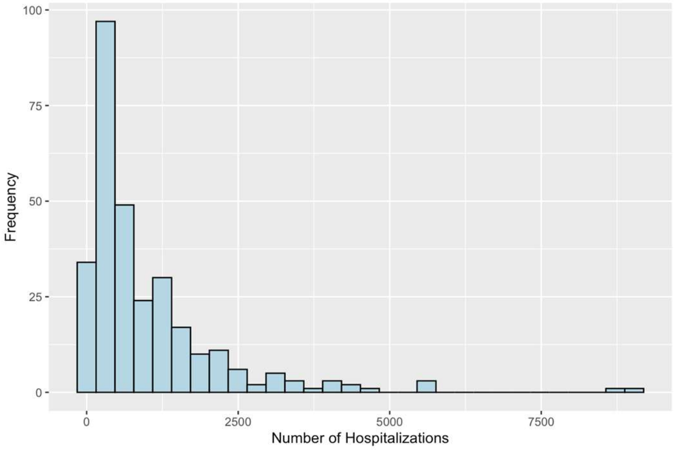


Figure 2: Correlation Matrix

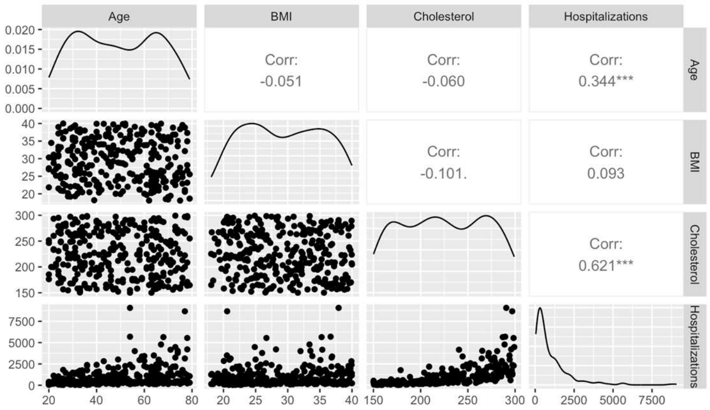


Figure 1 shows the distribution of the response variable (Hospitalizations) is very skewed right. The majority of the subjects' hospitalization visits are between 246 and 1,138. There also appears to be extreme outliers hanging to the far right. We consider these outliers during the residual analysis.

The correlation matrix in figure 2 shows that the variables Age and Cholesterol are the two having the strongest and most significant relationships with hospitalizations in this sample. Their correlations with Hospitalization are much higher than BMI's correlation. Aging can increase a person's hospital visits due to declining organ function and increased disease risk. In our case, hospitalization rates may be influenced by the body's reduced ability to process cholesterol with age, often resulting in unhealthy cholesterol levels. Therefore, we include variables Age and Cholesterol as key predictors in our model. This also shows us that there appears to be no multicollinearity problems with the main effects as of now.

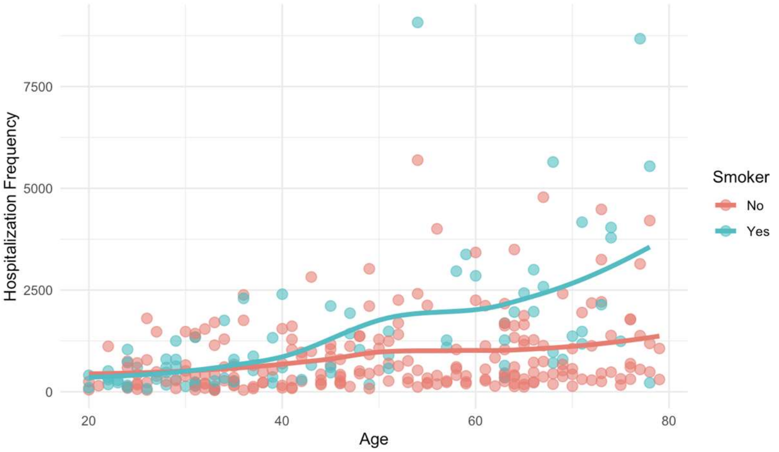


Figure 3: Age vs Hospitalizations vs Smoker

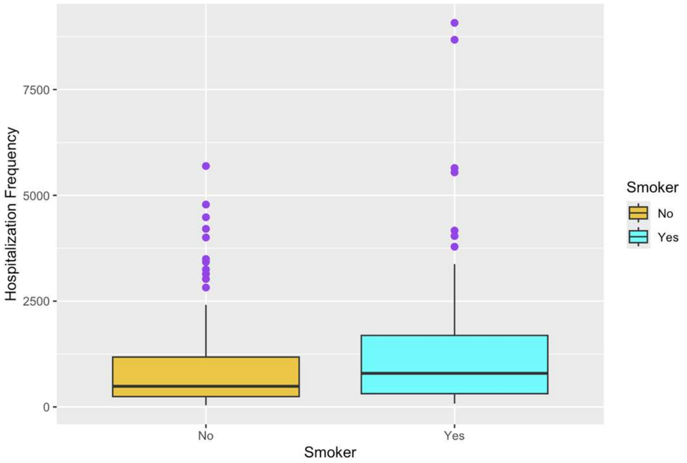


Figure 4: Hospitalizations vs Smoker

The *Hospitalizations vs. Smoker* box plot (figure 4) illustrates a relationship between an individual's smoking status and their number of hospital visits. Given the severe health risks of smoking, this trend is expected. Although only 27% of the subjects are smokers, they have more hospital visits on average than the 73% who do not smoke. Figure 3 further suggests that hospitalization rates increase with age, and that smoking status may interact with age. Among the elderly (age ≥ 65), frequent hospital visits are more common among smokers. A similar observation is witnessed for individuals with diabetes in figures 5 and 6. Therefore, we include variables Smoker and Diabetes to our model.

Furthermore, we investigated directly into possible indicators to see if our assumption would be met. The comparison of the numerical predictor variables compared to their corresponding variance alerted us of a possible case of overdispersion for our data. As for the linearity check, we also compared the $\log(\mu_i)$ to each of the numerical predictors, which consists of Age, Cholesterol, and BMI. Based on the little curvature as seen on the loess curve, we can assume that the linearity assumption has been met. Since the loess curve for each of the variables remained similar in shape and were close to each other, we would infer that minimal interaction effects would be significant.

The variables Physical, BMI, and Gender showed little to no meaningful relationship with Hospitalizations, based on graphical analysis. BMI had a low correlation of -0.051, making it a weak predictor and unsuitable for inclusion in the baseline model. Physical activity did not show a clear relationship with hospitalizations in its box plot. However, the Wald test suggested it was statistically significant. The Physical variable appeared to be not ideal for the baseline model but may be reconsidered in later testing. Gender showed no strong association with hospitalization rates, both graphically and through the Wald test, making it too weak to be considered an ideal candidate.

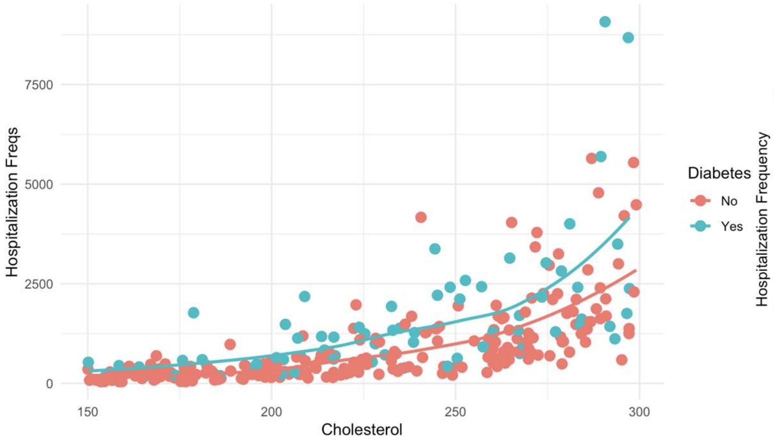


Figure 5: Cholesterol vs Hospitalizations vs Diabetes

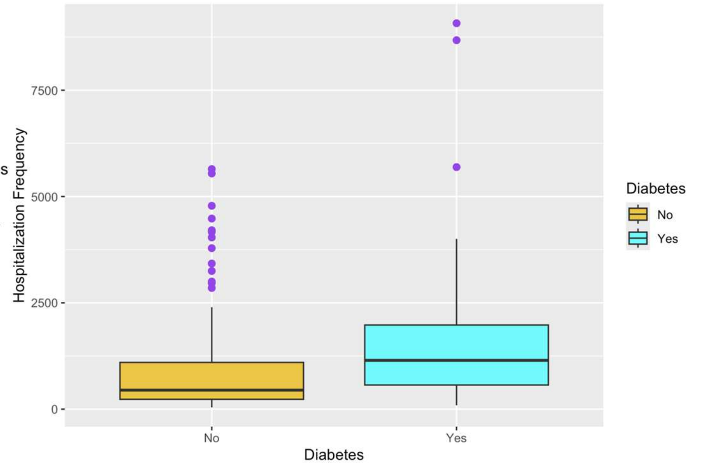


Figure 6: Hospitalizations vs Diabetes

We chose the features **Age, Cholesterol, Smoker,** and **Diabetes** as key predictors in our baseline model due to their clear graphical relationship with hospitalization rates. When we fit a Negative Binomial model using only these four features, the results confirmed their significance in predicting the number of hospital visits. These variables are deeply tied to human health, making their inclusion both logical and evidence-backed.

On the other hand, while factors like Physical Activity and BMI might intuitively seem just as relevant, our data visualization did not strongly support their predictive power. As a result, we excluded them from our baseline model. However, this exclusion isn't absolute—the Negative Binomial model reveals the significance of these omitted features or their interactions, prompting further refinements.

Methodology

We first decided to fit a generalized linear model that best predicts the relationship between Hospitalizations with our other predictor variables, we believe that the appropriate statistical procedure to use would be a regression model. The outcome variable in this study, Hospitalizations, represents a count variable. Therefore, we assumed that the appropriate probability distribution for this type of outcome would be the Poisson distribution with its corresponding Poisson Regression Model.

To ensure that all the assumptions were valid, we ran the dispersion test to check for equidispersion in our full model following Poisson. After running the test, we ended up rejecting the null hypothesis and concluded that the true dispersion was greater than 1, otherwise known as overdispersion. After comparing the Negative Binomial model to the Poisson model, we found out that the Negative Binomial model has a much lower AIC than the Poisson model. This suggests that the Negative Binomial model is a better fit. The mean number of hospitalizations was also confirmed to be less than its variance, concluding that we should model our study after the Negative Binomial distribution instead of Poisson. Ultimately, we employed a Generalized Linear Model (GLM) framework with a Negative Binomial distribution to model the relationship between the number of hospitalizations and the predictor variables. The final model is structured as follows:

Final Model Equation

$$\log(\mu_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9$$

- x_1 - Age
- x_2 - Cholesterol
- x_3 - Smoker
- x_4 - Diabetes
- x_5 - Physical
- x_6 - BMI
- x_7 - Gender
- x_8 - Cholesterol × Diabetes (x_2x_4)
- x_9 - BMI × Gender (x_6x_7)

Definitions

- Response Variable: μ_i = the expected number of hospitalizations for individual i.
- Link Function: $\log(\mu_i)$ ensures that the predictions remain non-negative, aligning with the count nature of the response variable.
- Predictors:
  - Main Effects:
    - Continuous: Age, Cholesterol, BMI
    - Binary: Smoker (Yes/No), Diabetes (Yes/No), Physical Activity (Yes/No), Gender (Male/Female).
  - Interaction Terms:
    - Cholesterol × Diabetes
    - BMI × Gender.

- Dispersion Parameter (θ): Models excess variability beyond the Poisson assumption, validated by $\theta = 32.78$ in our final model.
- Model Interpretation: For each one-unit increase in a given predictor, assuming all other predictors remain unchanged, the expected number of hospitalizations increases by a factor equal to the exponential of that predictor's coefficient.

Model selection

Baseline Model (Model 1)

Predictors: Age, Cholesterol, Smoking Status, Diabetes Status

We began with a Negative Binomial regression mode, informed by the exploratory data analysis, which includes predictors with strong theoretical and graphical support. Age and cholesterol exhibited the highest correlations with hospitalization counts in exploratory analysis, while smoking and diabetes showed clear graphical associations. All predictors were statistically significant ($p<0.001$), with age and cholesterol demonstrating the largest effect sizes.

Expanded Model (Model 2)

Predictors: Age, Cholesterol, Smoker, Diabetes, Physical Activity, BMI, Gender

To refine the baseline, three additional variables — Physical Activity, BMI, and Gender were incorporated. While exploratory visualizations suggested weak associations for these variables, their inclusion was motivated by potential clinical relevance. The significant improvements they provided:

- Physical Activity: Reduced hospitalizations by 25% ($e^{-0.29} = 0.75$, $p < 0.001$), a counterintuitive finding given weak graphical trend.
- BMI: Each unit increase in BMI raised hospitalizations by 4.8% ($e^{0.047} = 1.048$, $p < 0.001$), aligning with literature on obesity-related health risks.
- Gender: While the main effect for gender was non-significant ($p = 0.108$), it was retained for subsequent interaction testing.

Table 1: Model Comparison (Model 1 vs. Model 2)

	Log-Likelihood	AIC	BIC
Model 1	-4,080.6	4,092.6	4,114.8
Model 2	-3,661.9	3,679.9	3,713.2

A likelihood ratio test confirmed the superiority of Model 2 ($p<0.001$), with a significant reduction in deviance (residual deviance: 302.6 vs. 306.6). Mode 1 and 2's likelihood ratio statistic of 4,080.6 - 3,661.9 = 418.7 says that Model 2 improves more than Model 1. The lower AIC and BIC values further supported Model 2, suggesting that the added predictors (Physical Activity, BMI, Gender) meaningfully enhanced explanatory power.

Interaction Testing (Model 3)

Predictors: Age, Cholesterol, Smoker, Diabetes, Physical, BMI, Gender, [All Two-Way Interactions]

To explore potential synergies between predictors, Model 3 was constructed. This included 21 interaction terms between all pairs of predictors.

Model 3 key findings:

1. Significant Interactions:
   - Cholesterol × SmokerYes ($p = 0.0181$): The effect of cholesterol on hospitalizations is different for smokers vs. non-smokers.
   - Cholesterol × DiabetesYes ($p = 0.0472$): The impact of cholesterol on hospitalizations depends on diabetes status.
   - BM × GenderMale ($p = 0.0017$): The effect of BMI on hospitalizations differs for males vs. females.

2. Marginally Significant Interactions:
   - PhysicalYes($p = 0.4147$): Significant in our initial model.
   - Cholesterol × BMI ($p = 0.0879$): Cholesterol's effect on hospitalizations may depend on BMI.
   - PhysicalYes × BMI ($p = 0.0926$): The relationship between physical activity and BMI may affect hospitalizations.

We believe these interactions may have an impact for prediction due to their marginal significance. We are mindful that they are not strongly significant. Still, we consider keeping these variables for further investigation as studies suggest their interactions have a significant effect on human health.

Reduced Interaction Model (Model 4)

Predictors: Age, Cholesterol, Smoker, Diabetes, Physical, BMI, Gender, Cholesterol × Smoker, Cholesterol × Diabetes, Cholesterol × BMI, Physical × BMI, BMI × Gender

To address the complexity and overfitting risks of Model 3, a reduced interaction model (Model 4) was constructed. This model retained only the most significant and theoretically justified interactions: Cholesterol × Smoker, Cholesterol × Diabetes, Cholesterol × BMI, Physical × BMI, BMI × Gender. It was interesting to see that Cholesterol× BMI has better performance, therefore we kept for further exploration.

Model 4 key findings:

1.Significant Interactions:

- Cholesterol × DiabetesYes (p = 0.0207)
- BMI × GenderMale (p = 0.0076)

2.Marginally Significant Interactions:

- Cholesterol × Smoker (p=0.054)
- Cholesterol × BMI (p= 0.0659)

3.Model Fit:

- AIC = 3,666.2 (lower than Model 3)
- BIC = 3,718.1 (lower than Model 3)

Further Refinement (Model 5)

Predictors: Age, Cholesterol, Smoker, Diabetes, Physical, BMI, Gender, Cholesterol × Smoker, Cholesterol × Diabetes, Cholesterol × BMI, BMI × Gender

To further simplify the model, the non-significant interaction Physical × BMI was removed from Model 4 because it was the most insignificant interaction, resulting in Model 5. This step aims to reduce complexity of our model while maintaining predictive power and interpretability.

Model 5 key Findings:

1.Significant Interactions:

- Cholesterol × Diabetes (p=0.020)
- BMI × Gender (p=0.005)

2.Marginally Significant Interactions:

- Cholesterol × Smoker (p=0.057)
- Cholesterol × BMI (p=0.087)

3.Model Fit:

- AIC = 3,666.1 (slightly lower than Model 4).
- BIC = 3,714.3 (slightly lower than Model 4).

Final Refinement (Model 6)

Predictors: Age, Cholesterol, Smoker, Diabetes, Physical, BMI, Gender, Cholesterol × Diabetes, BMI × Gender

To further simplify the model, the interaction terms Cholesterol × Smoker and Cholesterol × BMI was removed from Model 5 because of their insignificance, resulting in Model 6. This step aimed to retain only the most significant and theoretically justified interactions while minimizing complexity.

Model 6 key Findings:

1.Significant Interactions:

- Cholesterol × Diabetes(p=0.026)
- BMI × Gender (p=0.002)

2.Model Fit

- AIC = 3,668.2 (slightly higher than Model 5)
- BIC = 3,708.9 (slightly lower than Model 5)

Table 2: Model Comparison (Model 5 vs Model 6)

	Log-Likelihood	AIC	BIC
Model 5	-3,646.2	3666.11	3714.26
Model 6	-3,640.1	3668.2	3708.94

The likelihood ratio test yielded a chi-squared statistic of 6.09 (p=0.048), indicating that those two extra interactions in Model 5 provide a marginal improvement in fit. However, BIC, which penalizes complexity more harshly than AIC, is lower for Model 6 (3708.9 vs 3714.3), favoring the simpler specification. In other words, both models are supportable by the evidence, but we prioritize simplicity and interpretability. We also want to avoid overfitting. Since BIC favors Model 6, we therefore select Model 6 as the final model and proceed with it for interpretation.

Multicollinearity Check

To assess multicollinearity in Model 6, we calculated the Variance Inflation Factor (VIF) score for each predictor. Below are the VIF results:

- Age and Physical Activity had the lowest VIF values of 1.08 and 1.04, respectively. This indicated no significant multicollinearity.
- Cholesterol, Smoker, Diabetes, BMI, and Gender had very high VIF values (>10), suggesting multicollinearity, which is expected due to their involvement in interaction terms.
- All interaction terms (Cholesterol:Smoker, Cholesterol:Diabetes, Cholesterol:BMI, and BMI:Gender) also had very high VIF values (>10). This is typical in models with interaction terms, as they are naturally correlated with their main effects.

Centering Predictors

To address the high multicollinearity observed in Model 6, we centered the continuous predictors (Cholesterol and BMI) by subtracting their mean values. This helped to reduce the correlation between main effects and their interaction terms. We calculated the Variance Inflation Factor (VIF) for each predictor after centering the continuous variables (Cholesterol and BMI). All VIF values were below five, indicating low multicollinearity among the predictors. Specifically:

Table 3: Main Effects VIF Scores

Main Effects	Age	Cholesterol	Smoker	Diabetes	Physical	BMI	Gender
VIF	1.04	1.37	1.07	1.06	1.04	1.81	1.03

Table 4: Interaction Term VIF Scores

Interaction Terms	Cholesterol:Diabetes	BMI:Gender
VIF	1.42	1.82

This confirms that centering the continuous predictors successfully addressed multicollinearity concerns, ensuring stable and reliable coefficient estimates in the final model.

Outlier Analysis

In our analysis of the hospitalization dataset, we identified several outliers—individuals with extremely high hospitalization counts (e.g., 8,676 visits for a 77-year-old female smoker with diabetes). Despite their extreme values, we chose to retain these outliers for the following reasons:

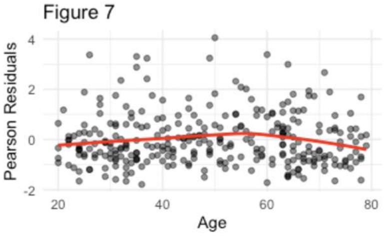
1. Outliers often represent high-risk patients with severe health conditions, such as chronic illnesses. Removing these cases would bias the model toward healthier populations, reducing its ability to predict hospitalization patterns for vulnerable subgroups. These outliers are clinically meaningful and should not be excluded simply because they are extreme.

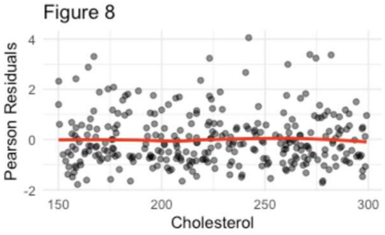
2. The dataset contains 300 observations, which is relatively small for a regression model with multiple predictors. Removing even a few outliers would significantly reduce the sample size, potentially leading to loss of statistical power and unstable parameter estimates. Retaining outliers ensures the model reflects the full heterogeneity of the population.

Residual Analysis

After selecting our final model—which includes all main effects along with the interaction terms Cholesterol × Diabetes and BMI × Gender—we conducted a thorough residual analysis to confirm that all model assumptions were satisfied. As previously established, the response variable follows a negative binomial distribution and exhibits overdispersion. The assumption of independence also holds true, as each of the 300 observations corresponds to a unique individual. This left us with two key assumptions to verify: linearity and overall model fit.

To assess linearity, we plotted the Pearson residuals against each numeric predictor: Age, Cholesterol, and BMI. The plots revealed no suspicious wave-like patterns that would suggest non-linearity or the need for transformation. For model fit, we used the global Wald Test, which produced a p-value close to 0.0—well below our 0.05 significance threshold. This result allows us to confidently reject the null hypothesis and conclude that at least one predictor in the final model is statistically significant to an individual's hospitalization rate.





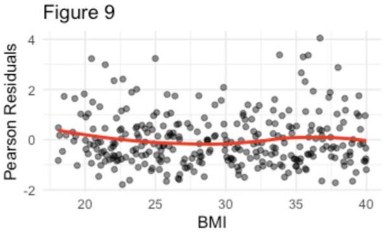


Figure 7: A scatterplot showing the Pearson residuals vs. age with the red loess curve, illustrating the linearity assumption is considered satisfied.

Figure 8: A scatterplot showing the Pearson residuals vs. cholesterol with the red loess curve, illustrating the linearity assumption is considered satisfied.

Figure 9: A scatter plot showing the Pearson residuals versus BMI with the red loess curve, illustrating the linearity assumption is considered satisfied.

Results

Our estimated regression equation using our centered variables of Cholesterol and BMI resulted the final model:

- x_1 - Age
- x_2 - Cholesterol
- x_3 - Smoker
- x_4 - Diabetes
- x_5 - Physical
- x_6 - BMI
- x_7 - Gender
- x_8 - Cholesterol × Diabetes (x_2x_4)
- x_9 - BMI × Gender (x_6x_7)

$$\log(\mu_i) = 4.755 + 0.03(x_1) + 0.02(x_2) + 0.409(x_3) + 0.592(x_4) - 0.275(x_5) + 0.051(x_6) - 0.031(x_7) - 0.001(x_8) - 0.011(x_9)$$

which once converted to our expected response μ_i, we have:

$$\mu_i = e^{(4.755 + 0.03(x1) + 0.02(x2) + 0.409(x3) + 0.592\,(x4) - 0.275(x5) + 0.051(x6) - 0.031(x7) - 0.001(x8) - 0.011(x9))} = e^{(\beta 0)} \times \Pi_J e\left(\beta_J x_{Ji}\right)$$

$$\mu_i = 116.116 \times 1.031(x_1) \times 1.020(x_2) \times 1.502(x_3) \times 1.807(x_4) \times 0.759(x_5) \times 1.053(x_6) \times 0.970(x_7) \times 0.999(x_8) \times 0.989(x_9)$$

The expected count of hospitalizations is now a product of factors.

Interpretations

Table 5: Multiplicative Effects on Expected Hospitalizations

	Age	Cholesterol	Smoker	Diabetes	Physical	BMI	Gender	Cholesterol:Diabetes	BMI:Gender
1 - β_i	-0.031	-0.02	-0.502	-0.807	0.241	-0.53	0.03	0.001	0.011

Based on these estimates, the variables can be interpreted as follows:

- Age: an individual's expected hospitalization rate increases approximately 3.1% every year they get older.
- Cholesterol: an individual's expected hospitalization rate increases approximately 2% for each additional milligrap per deciliter (mg/dL).
- Smoker: a person who smokes has about 50.2% higher expected hospitalizations than non-smokers.
- Diabetes: a person who is diabetic has about 80.7% higher expected hospitalizations than non-diabetics.
- Physical: an individual who is physically active has about 24.1% fewer expected hospitalizations than individuals who are inactive.
- BMI: an individual's expected hospitalizations rate increases approximately 53% for every 1-unit BMI increase.
- Gender: a male is 3% less likely to have higher expected hospitalizations than females.
- Cholesterol x Diabetes: Cholesterol increases hospitalizations in both groups, but somewhat less for diabetics. For diabetics, each 1 mg/dL increase in cholesterol raises expected hospitalizations by about 1.9% (vs 2% for non-diabetics).
- BMI x Gender: Higher BMI increases hospitalizations in both sexes, but somewhat less for males. For males, each 1-unit increase in BMI raises expected hospitalizations by about 4.1% (vs 5.3% for females).
- The intercept (β_0) no realistic interpretation because an individual will require at least some of these features (like gender).

Discussion & Conclusion

Our stated goal was to identify the most significant features to include in a parsimonious model of hospitalization counts while filtering out less relevant variables. Model 6 achieves this. After comparing candidate specifications with LRT/AIC/BIC and addressing multicollinearity by centering continuous predictors, we selected a final Negative Binomial GLM that keeps the features with stable, independent signal and drops weaker or redundant terms. Concretely, the retained predictors are Age, Cholesterol, BMI, Smoking, Diabetes, Physical Activity, plus two interaction modifiers—Cholesterol×Diabetes and BMI×Gender. We deliberately excluded the weaker interactions (Cholesterol×Smoker, Cholesterol×BMI) because they offered only marginal in-sample gains and reduced interpretability.

Interpreting Model 6 on the single-unit scale provides a clear answer to *which features matter and by how much*: expected hospitalizations rise by ~3.1% per +1 year of age, ~2.0% per +1 mg/dL of cholesterol, and ~5.3% per +1 BMI unit for females; smokers average ~50% higher, diabetics ~81% higher, and physically active individuals ~24% lower counts relative to their reference groups. The interactions refine—rather than replace—these conclusions: the cholesterol slope is slightly weaker in diabetics (+1.9% vs +2.0% per mg/dL), and the BMI slope is weaker in males (+4.1% vs +5.3% per unit). The male main effect itself is small and not statistically significant after adjustment. Taken together, these results deliver a minimal, interpretable feature set that captures the core drivers and the two group-specific slope differences most supported by the data.

Because Model 6 balances fit and simplicity, it directly serves our goal: it tells us which variables to keep, which to discard, and how to explain the effects in a way that is actionable and generalizable. The model is diagnostically sound for this dataset (overdispersion handled; residuals without systematic pattern; VIFs < 2 after centering), and its multiplicative interpretation makes it suitable for communication and deployment (e.g., as a small risk calculator). As next steps, we recommend validating out-of-sample performance (cross-validation and calibration plots), probing limited non-linearities for Age/Cholesterol/BMI, incorporating survey design if applicable, and externally validating on an independent cohort. These steps would stress-test the selected feature set and confirm that the Model 6 variables—and no more—are the right ones to carry forward.

Author Contribution

R.T. performed model selection, residual analysis, participated in reporting and wrote the Results and Discussion & Conclusion section

W.T performed model selection, outlier analysis, and centering of predictors to address multicollinearity. Conducted Likelihood Ratio Tests and AIC/BIC comparisons for model evaluation.

G.V. performed data manipulation, EDA, variable selection, residual analysis, multicollinearity checking, model selection, coding, and reporting.

J.H. performed residual analysis, coding, and reporting.

References

Centers for Disease Control and Prevention. (2020, February 27). *Data Briefs, Number 360*. National Center for Health Statistics www.cdc.gov/nchs/products/databriefs/db360.htm.

Li, M., et al. (2022, December 16). Trends in body mass index, overweight and obesity among adults in the USA, the NHANES from 2003 to 2018: A repeat cross-sectional survey. *BMJ Open*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9764609/