Emergency Department Length of Stay: Statistical Modeling Report

Author: Annelise Thorn
Project: Hospital Length of Stay Analysis

Executive Summary

This report evaluates emergency department (ED) length of stay (LOS) drivers using a synthetic dataset of 100,000 encounters. The workflow combines exploratory analysis, inferential statistics, and predictive modeling to identify meaningful factors associated with LOS and compare model families.

The primary modeling conclusion is that a Gamma GLM with log link is better aligned with LOS characteristics (positive, right-skewed, heteroscedastic) than a standard multiple linear regression model.

1. Objective

Research question: What factors drive emergency department LOS, and how well can LOS be predicted?

2. Data

- Source: Kaggle (Hospital Length of Stay Dataset – Microsoft)
- Size: 100,000 rows, 28 variables
- Target: `lengthofstay`

Variables include demographics, comorbidities, labs, vitals, readmission count, and facility ID.

3. Methodology

3.1 Data Preparation

- Type conversion for date and factor variables
- Missingness and duplicate checks
- Filtering physiologically implausible values for selected labs and vitals

3.2 Exploratory Analysis

- LOS distribution assessment
- Group comparisons across gender and facility
- Summary diagnostics for candidate predictors

3.3 Inference

- Two-sample t-test for LOS by gender

- One-way ANOVA for LOS across facilities

3.4 Predictive Modeling

- Baseline multiple linear regression (MLR)
- Model selection using AIC/BIC and holdout MSPE
- Improved Gamma GLM with log link
- Diagnostic review of residual behavior

4. Key Findings

- LOS is strongly right-skewed with a long tail.
- Facility-level differences are statistically meaningful.
- Clinical complexity variables contribute more to LOS variation than basic demographics.
- Gamma GLM provides stronger distributional fit and predictive stability for LOS-like outcomes.

5. Limitations

- Synthetic data may not reflect all real-world operational complexity.
- Additional operational variables (arrival hour, staffing, occupancy) may improve explanatory power.

6. Reproducibility

- Analysis script: ../notebooks/ed_length_of_stay_analysis.R
- Environment details: session_info.txt
- Full reproducibility instructions: ../REPRODUCIBILITY.md

7. References

- https://www.kaggle.com/datasets/aayushchou/hospital-length-of-stay-dataset-microsoft?resource=download
- https://www.sciencedirect.com/science/article/pii/S1755599X20301026