Emergency Department Length of Stay: Statistical Modeling Report

Author: Annelise Thorn
Project: Hospital Length of Stay Analysis
Report Date: February 20, 2026

Executive Summary

This report analyzes emergency department (ED) length of stay (LOS) using a synthetic healthcare dataset with 100,000 encounters and 28 variables. The objective is to identify key factors associated with LOS and compare predictive modeling approaches.

The analysis combines exploratory data analysis, inferential statistics (two-sample t-test and one-way ANOVA), multiple linear regression (MLR), and a Gamma generalized linear model (GLM) with log link.

The central modeling conclusion is that a Gamma GLM is better aligned with LOS data characteristics (positive, right-skewed, heteroscedastic) than standard linear regression.

1. Problem Statement

Emergency departments must balance throughput and quality of care under crowding constraints. LOS is a key operational outcome linked to patient flow and system efficiency.

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

2. Data Source and Scope

- **Dataset:** Hospital Length of Stay Dataset (Microsoft) from Kaggle
- **Rows:** 100,000
- **Columns:** 28
- **Target variable:** `lengthofstay`

The dataset includes demographics, comorbidity flags, lab values, vital signs, facility identifiers, and discharge information.

3. Data Preparation

Preprocessing steps include:

- Date parsing (`vdate`, `discharged`)
- Categorical encoding (`gender`, `rcount`, `facid`)
- Target type correction (`lengthofstay`)

- Missing and duplicate checks
- Filtering physiologically implausible values in selected labs and vitals

4. Exploratory Data Analysis

EDA findings indicate:

- LOS is heavily right-skewed
- Most encounters have shorter LOS with a long-tail subset of prolonged stays
- Facility-level distributions differ

5. Inferential Statistics

5.1 Two-Sample t-Test (Gender)

- **Null hypothesis (H_0):** mean LOS is equal for male and female patients
- **Alternative hypothesis (H_1):** mean LOS differs by gender

Interpretation in the notebook indicates no practically meaningful gender difference.

5.2 One-Way ANOVA (Facility)

- **Null hypothesis (H_0):** all facilities have the same mean LOS
- **Alternative hypothesis (H_1):** at least one facility mean differs

Interpretation indicates statistically significant between-facility differences.

6. Predictive Modeling

6.1 Baseline MLR

The baseline MLR uses demographic, comorbidity, lab, vital, and facility predictors.

Evaluation includes:

- Coefficients and confidence intervals
- R^2
- AIC and BIC
- Residual diagnostics
- Holdout MSPE

6.2 Model Selection

Three nested models were compared:

- Full
- Clinically focused

- Minimal

Selection criteria: AIC, BIC, and test-set MSPE.

6.3 Improved Model: Gamma GLM (Log Link)

Because LOS is positive and right-skewed, the analysis fits a Gamma GLM with log link.

Evaluation includes:

- AIC/BIC
- Pseudo-$R^2 = 1 - \frac{\text{residual deviance}}{\text{null deviance}}$
- Test-set MSPE
- Residual diagnostics

The notebook's conclusion indicates improved fit and predictive stability over baseline MLR.

7. Key Findings

- Clinical complexity variables are stronger LOS drivers than basic demographics
- Facility differences are meaningful and may reflect operational variability
- Gamma GLM is better suited than MLR for LOS-type outcomes

8. Limitations and Next Steps

- Data is synthetic and may not fully represent production EHR complexity
- Additional operational signals (arrival hour, occupancy, staffing) may improve performance
- Future extensions: interaction terms, non-linear effects, and additional outcome distributions

9. Reproducibility

- Notebook: ../notebooks/ed_length_of_stay_analysis.ipynb
- R script: ../notebooks/ed_length_of_stay_analysis.R

10. References

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