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 in the notebook include:

- date parsing (`vdate`, `discharged`),
- categorical encoding (`gender`, `rcount`, `facid`),
- target type correction (`lengthofstay`),
- checks for missing values and duplicates,
- filtering of physiologically implausible values in selected lab and vital variables.

These steps improve data consistency and statistical validity before inference and model fitting.

4. Exploratory Data Analysis

EDA findings described in the notebook indicate:

- LOS is heavily right-skewed,
- most encounters have relatively short LOS,
- a smaller subset of encounters has prolonged LOS,
- facility-level distributions appear to differ.

Visuals produced include LOS distribution, gender distribution, and facility distribution.

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.

Notebook interpretation indicates no meaningful gender difference in LOS.

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.

Notebook interpretation indicates facility-level LOS differences are statistically significant.

6. Predictive Modeling

6.1 Baseline Model: Multiple Linear Regression (MLR)

The baseline MLR uses demographic, facility, comorbidity, laboratory, and vital-sign predictors.

Model assessment includes:

- coefficient estimates and confidence intervals,
- R^2,
- AIC and BIC,

- residual diagnostics,
- and holdout test performance via mean squared prediction error (MSPE).

6.2 Model Selection

Three nested linear models are compared:

- full model,
- clinically focused model,
- minimal model.

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

6.3 Improved Model: Gamma GLM (Log Link)

Given LOS is strictly positive and right-skewed, the notebook fits a Gamma GLM with log link and compares it to linear alternatives.

Evaluation includes:

- AIC and BIC,
- pseudo-$R^2 = 1 - \frac{\text{residual deviance}}{\text{null deviance}}$,
- test-set MSPE,
- and residual diagnostics.

Notebook conclusions state Gamma GLM improves fit and predictive stability relative to MLR.

7. Key Findings

- Clinical complexity variables are more informative than basic demographics for LOS.
- Facility differences are meaningful, suggesting operational variability across sites.
- Classical MLR assumptions are strained for this outcome distribution.
- Gamma GLM is statistically more appropriate for LOS-type outcomes.

8. Practical Implications

This workflow can support healthcare operations and analytics use cases such as:

- capacity planning,
- throughput monitoring,
- and prioritization of interventions to reduce prolonged ED stays.

9. Limitations

- Dataset is synthetic and may not fully represent real-world EHR complexity.

- Results may not generalize directly to specific institutions.
- Additional operational variables (arrival time, occupancy, staffing) could improve explanatory power.

10. Recommended Next Steps

1. Add operational features (time-of-day, census, staffing proxies).
2. Evaluate interaction effects and non-linear relationships.
3. Compare additional positive-outcome model families (for example, log-normal or Tweedie).
4. Add reproducible model performance tables exported as artifacts.

11. Reproducibility

Primary analysis source:

- notebooks/ed_length_of_stay_analysis.ipynb

Python script conversion:

- notebooks/ed_length_of_stay_analysis.py

12. References

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