MODELING HOSPITAL STAY DURATION BASED ON PATIENT DEMOGRAPHICS AND TREATMENT VARIABLES USING SAS REGRESSION

1. Overview

Client:

A regional healthcare system operating multiple inpatient facilities across the U.S.

Objective:

To build a multiple linear regression model in SAS to predict the number of inpatient days based on patient demographics, diagnosis, and treatment variables. The goal was to improve hospital bed management and inform early discharge protocols.

2. Background

Length of stay (LOS) is a critical efficiency metric for hospitals. Excessive LOS increases costs and impacts quality-of-care scores. The client lacked an interpretable predictive model for LOS based on admission data. SAS was chosen due to its capacity for clinical data integration and its use in regulatory health reporting.

3. Data Summary

Dataset:

Admissions data for 9,200 patients admitted across three hospitals in 2022

Variables Used:

Variable	Type	Description
LOS_Days	Continuous	Length of stay (dependent variable)
Age	Continuous	Patient age in years
Gender	Categorical	Male / Female
Diagnosis_Category	Categorical	Cardiac, Respiratory, Ortho, Other
Surgery_Performed	Binary	1 = Yes, 0 = No
ICU_Admission	Binary	1 = ICU included in stay, 0 = No

Comorbidity_Count	Continuous	Number of documented chronic conditions
Admission_Source	Categorical	ER, Referral, Elective

4. Methodology

Software Used:

SAS 9.4

SAS Workflow:

1. **Data Cleaning & Prep:**
    - PROC IMPORT and DATA steps for categorical encoding
    - Imputed missing values using hospital median and mode values
    - Created dummy variables for Diagnosis_Category and Admission_Source
2. **Exploratory Analysis:**
    - PROC MEANS, PROC FREQ, and box plots to understand LOS distribution
    - Detected positive skewness in LOS → applied log transformation
3. **Regression Modeling:**
    - PROC REG with transformed dependent variable log(LOS_Days)
    - Checked multicollinearity using VIF
    - Residual diagnostics performed with PLOTS=DIAGNOSTICS
4. **Model Selection:**
    - Stepwise selection using AIC/BIC
    - Tested interaction: ICU × Surgery_Performed

5. Key Results

Predictor	**Coefficient (β)**	**p-value**	**Interpretation**
Age	+0.015	0.003	Older patients tend to have longer stays
Comorbidity_Count	+0.072	<0.001	More conditions → longer LOS

Surgery_Performed	+0.231	0.018	Surgery increases LOS
ICU_Admission	+0.362	<0.001	ICU use significantly lengthens stay
Admission_Source (ER)	+0.108	0.039	ER admissions lead to longer stays than elective ones

Model Diagnostics:

- $R^2 = 0.68$ (log-transformed model)
- VIF values < 2
- Residuals approximately normal
- Cook's Distance: no influential points identified

6. Visual Outputs (from SAS)

- Residual vs. Fitted plot for homoscedasticity
- Histogram and Q-Q plot for residuals
- Variable importance bar chart
- Scatterplot matrix of key predictors

7. Deliverables

- .sas regression code with log transformation workflow
- Technical report (18 pages) including:
  - Variable coding and rationale
  - Regression results and diagnostics
  - Back-transformed LOS predictions
  - Recommendations for operational use
- Briefing deck (5 slides) for hospital administration:
  - Key findings
  - Predictors of prolonged stay
  - Implications for bed turnover and discharge strategy

8. Application & Outcome

- Used by hospital operations to flag high-LOS patients on admission
- Integrated into daily discharge planning dashboard
- Enabled prioritization of resource-intensive patients during bed shortage periods

9. Strategic Value Delivered

- Provided a **clinically interpretable tool** for stay prediction
- Supported **proactive care coordination and capacity planning**
- Increased hospital readiness for **surge scenarios** using SAS-driven forecasts