The Use of Count Data Models in the Health Industry

Introduction

In the evolving landscape of healthcare analytics, the ability to accurately model and interpret discrete health outcomes is critical. Count data models offer a robust statistical framework for analyzing outcomes expressed as non-negative integers—such as the number of hospital visits, prescriptions, or disease cases. These models are increasingly vital for understanding healthcare utilization, identifying risk factors, and guiding evidence-based policy decisions.

Applications in Health Research

Count data models are widely used across various facets of the health industry. Below are three examples from the literature that illustrate their versatility:

- **Outpatient Health Services Utilisation:** Abu Bakar et al. (2022) applied various count data models—including Poisson, Negative Binomial, and Zero-Inflated models—to analyze outpatient visit frequency using data from the Malaysian National Health and Morbidity Survey. Their study found that zero-inflated models provided the best fit for data with a high proportion of zero counts.
  `https://doi.org/10.1186/s12874-022-01733-3`
- **Disease Surveillance:** Held et al. (2006) developed a two-component model to analyze counts of infectious diseases. Their approach combined endemic and epidemic components using Poisson and autoregressive structures, providing a flexible framework for modeling disease incidence over time.
  `https://doi.org/10.1093/biostatistics/kxj016`
- **Medication Adherence:** Tibble et al. (2023) used count-based methods to estimate asthma medication adherence from electronic health records. By analyzing over 1.6 million prescriptions, they compared multiple adherence metrics and highlighted the importance of model choice in interpreting treatment patterns.
  `https://doi.org/10.1186/s12874-023-01935-3`

The Project

It is anticipated that this project will investigate a range of count data models, including the Poisson, Negative Binomial, Poisson-Inverse Gaussian, as well as Zero-Inflated and Zero-Deflated variants. A suitable dataset will be selected and analyzed (in R) using these models to evaluate their performance in capturing the underlying data structure. Model fit and suitability will be assessed using a combination of statistical criteria—such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Quantile-Quantile (Q-Q) plots of residuals, and Quantile-Band plots. The relative emphasis on each component of the project will be guided by the specific interests and expertise of the authors.