Stat 220 Lab 5/6

1 Introduction

Recent reports in the health sector have highlighted a potential discrepancy in hospital infection risks across various regions in the U.S. In order to maintain high standards of patient care and ensure equitable healthcare outcomes, it is vital to investigate these claims. In this lab exercise, we will delve into a dataset provided by the U.S. Health Department that represents different hospitals across four major regions: northeast, north-central, south, and west. Our primary objective is to determine the regional effect on the risk of infection in hospitals.

Data Acquisition and Initial Exploration

1. Download the dataset from the following link: `https://richardson.byu.edu/220/infection.csv`
2. Load the dataset into python and perform a preliminary exploration of the data

The variables in the data set are listed here.

Variable	Description
InfctRsk	Infection risk in the hospital
Age	Average age of patients
Stay	Average length of patient's stay (in days)
MedSchool	Presence of a medical school at the hospital ('Yes' for presence, 'No' for absence)
Region	Region in which the hospital is located (north-east, north-central, south, west)
Beds	Number of beds in the hospital

Table 1: Data dictionary for the modified Hospital Infection Risk dataset.

Regression Analysis

1. Given the categorical nature of the 'Region' variable, transform it into dummy variables. Remember that for k categories, you'll need $k-1$ dummy variables.
2. Fit a linear regression model with 'InfctRsk' as the target variable and 'Age', 'Stay', 'MedSchool', 'Beds', and 'Region' as the predictor variables.
3. Examine the regression coefficients. Based on these coefficients, comment on the effect of the different regions on infection risk.

Bayesian Regression

1. Use the 'bambi' package in python to fit a Bayesian regression model using the same variables as in the linear regression.
2. Examine the posterior distributions of the coefficients associated with the dummy variables for 'Region'.
3. Compute the probability that each regression coefficients associated with 'Region is positive (greater than 0) or negative (less than 0).

Discussion

1. Why was it pivotal to incorporate variables like 'Age', 'Stay', 'MedSchool', and 'Beds' in the model, even though the focal interest was 'Region'?
2. Differentiate between the Bayesian approach and the traditional regression approach. Enumerate the advantages and potential limitations of deploying Bayesian regression in this context.

Deliverable

There are two deliverables for this lab:

1. A Python file or notebook that contains the code to produce all the requested results. Also include responses to the discussion questions above.
2. A non-technical report for the U.S. Health department. This is a short and focused report designed to justify your methods and disclose the results specifically with regards to the ultimate question of what effect does region have on infection risk. In this report include the following:
   - No code or formulas
   - Any figures or tables that help with visualizing the important relationship
   - A brief description of the methods you are using and why they will help to answer the main question
   - What your models determined with regards to the question of what effect does region have on infection risk

   This report could be just one page and should not be longer than two pages.