Bayesian Proposal of Final Project

Comparing Bayesian and Frequentist Models in Healthcare Cost Prediction

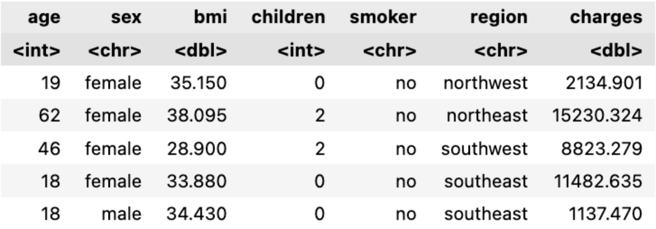
Yejin Hwang

1. Topic of Interest:

My final project will explore **healthcare cost prediction using Bayesian Linear Regression**, with a focus on comparing its predictive performance to that of traditional frequentist models. This topic allows me to apply the full Bayesian workflow introduced in *Statistical Rethinking* and critically evaluate how Bayesian modeling can improve real-world prediction tasks.

2. Data Source:

The dataset used in this project is the **Medical Cost Personal Dataset**, made publicly available on Kaggle. It originates from the book *Machine Learning with R* by Brett Lantz and has been slightly cleaned and reformatted by the uploader to match the book's original format. It contains demographic and medical information (e.g., age, sex, BMI, number of children, region, smoking status) along with the target variable: individual medical charges.



age	sex	bmi	children	smoker	region	charges
<int>	<chr>	<dbl>	<int>	<chr>	<chr>	<dbl>
19	female	35.150	0	no	northwest	2134.901
62	female	38.095	2	no	northeast	15230.324
46	female	28.900	2	no	southwest	8823.279
18	female	33.880	0	no	southeast	11482.635
18	male	34.430	0	no	southeast	1137.470

3. Introduction (Motivation):

In healthcare analytics, accurately predicting patient costs is essential for budgeting, insurance, and policy decisions. Classical linear regression provides point estimates, but it does not quantify uncertainty or nonlinear relationships well.

By applying Bayesian Linear Regression with the *rethinking* package, I will build a richer model that incorporates prior beliefs, estimates posterior distributions via **quadratic approximation using quap()**, and generates predictions through **posterior simulation**, allowing uncertainty to be directly visualized and quantified.

In addition to building a Bayesian model, I aim to **compare its predictive ability with that of traditional frequentist models**, such as classical linear regression, in order to evaluate whether Bayesian methods offer tangible improvements in predictive performance.

4–5. Goals and Project Purpose:

The purpose of this project is to develop and evaluate a Bayesian Linear Regression model to better capture meaningful patterns in healthcare cost data, including nonlinear relationships and interaction effects. While traditional frequentist models typically provide point estimates without uncertainty, Bayesian approaches allow for the incorporation of prior knowledge and generate full posterior distributions, enabling more robust inference and generalization. I will compare the predictive performance of frequentist and Bayesian models using standard metrics such as RMSE, MAE, and R^2, along with Bayesian-specific criteria like WAIC. The Bayesian model will be implemented using the rethinking package, applying quadratic approximation via quap() for fast and interpretable estimation, and exploring MCMC-based inference through ulam() where more flexible posterior sampling is beneficial. Predictions will be generated through posterior simulation using sim(), allowing for uncertainty quantification and model evaluation.