Bayesian and Frequentist Approaches to

Healthcare Cost Prediction: A Comparative Study

**Yejin Hwang**

Texas A&M University-Corpus Christi

6300 Ocean Dr, Corpus Christi, TX 78412

[yhwang@islander.tamucc.edu](mailto:yhwang@islander.tamucc.edu)

May 5, 2025

# Abstract

Accurately predicting individual medical costs is critical for policy-making and healthcare budgeting. This study explores the use of Bayesian Linear Regression to model healthcare charges and contrasts it with traditional frequentist linear regression. Using the Medical Cost Personal Dataset, we implement a full Bayesian workflow via the `rethinking` package in R, emphasizing prior specification, uncertainty quantification, and model comparison using WAIC. Results reveal that while both approaches yield similar central estimates, Bayesian methods offer clearer uncertainty interpretation and robustness through prior incorporation.

# **1. Introduction**

Healthcare cost prediction plays a pivotal role in insurance pricing, resource allocation, and financial planning. Traditional linear regression models, although widely used, provide limited information about uncertainty. Bayesian methods address this gap by delivering full posterior distributions, enabling nuanced decision-making under uncertainty.  
  
Problem Statement: Can Bayesian Linear Regression improve prediction accuracy or interpretability compared to frequentist regression in the context of healthcare costs?

# **2. Dataset and Features**

This study utilizes the "Medical Cost Personal Dataset" containing 1338 observations with variables such as age, BMI, number of children, sex, smoking status, region, and charges. Data preprocessing includes standardization and one-hot encoding of categorical variables. The dependent variable is `charges`, a continuous measure of medical expenses.

A table with numbers and symbols

AI-generated content may be incorrect.

**Exploratory Data Analysis(EDA)**

To ensure appropriate model design for Bayesian analysis, we conducted a series of preprocessing and exploratory visualization steps. This phase focused on understanding distributional patterns, removing low-value predictors, and identifying key interaction effects.

### **1. Distribution of Charges**

The target variable, charges, is heavily right-skewed. As shown in **Figure 1**, the majority of observations fall below $20,000, but a small number of high-cost outliers exist above $40,000. This highlights the need for models that account for heteroscedasticity.

A comparison of a graph

AI-generated content may be incorrect.

### **2. Feature Filtering: Sex and Children**

We evaluated the predictive strength of demographic features.

* **Sex** showed minimal impact on charges (see **Figure 2**, left).
* **Children** had a slight negative association with charges for large families (4–5 children) as shown in **Figure 2**, right.

### **3. Bivariate Effects: Smoker and Region**

In **Figure 3**, we observe that smokers consistently have higher charges across the board. In contrast, regional differences appear minor and likely do not warrant inclusion in the final model.

A comparison of different colored lines

AI-generated content may be incorrect.

### **4. Interaction Effects**

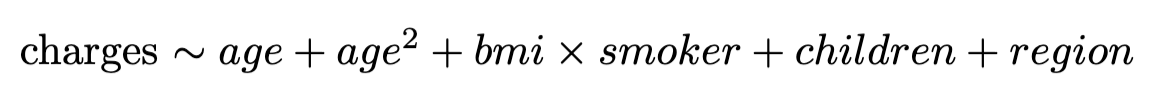
We explored interaction terms to capture possible nonlinear dependencies.  
As shown in **Figure 4**, the **BMI × Smoker** interaction demonstrated the strongest effect and was included in the final Bayesian model. The Age × BMI and Age × Smoker effects were weaker and excluded.

A group of graphs showing different types of smoking

AI-generated content may be incorrect.

# **3. Methods**

This project compares two approaches to healthcare cost modeling: traditional linear regression (`lm()` in R) and Bayesian linear regression using `quap()` and `ulam()`.  
  
**3.1 Frequentist Model**  
- Implemented via the `lm()` function in R.  
- Standard linear model assumptions: linearity, independence, homoscedasticity.



A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer code

AI-generated content may be incorrect.  
  
**3.2 Bayesian Model Specification**

**Variable processing**

**A computer code with many colored text

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.**

To improve numerical stability and model convergence, all continuous predictors (e.g., age, bmi) were standardized (mean = 0, SD = 1). Categorical variables such as sex, smoker, and region were converted to binary or index-based encodings. The response variable charges was scaled by 10,000 to facilitate HMC sampling. These transformations ensured appropriate input for Bayesian inference using the rethinking package.

**Model Fitting**

**3.3 Bayesian Model with Hierarchical Structure**

To improve generalization and account for regional variability, we extended the Bayesian linear model to include group-level (hierarchical) intercepts by region. The model is specified as:

A math equations on a white background

AI-generated content may be incorrect.

This hierarchical specification allows **partial pooling** of region-specific intercepts, improving stability in parameter estimation and model generalizability. The nonlinear term age² captures the increasing marginal effect of age on charges, while the interaction BMI × smoker accounts for compounding health risks among smokers with higher BMI.

The model was implemented using **Hamiltonian Monte Carlo (HMC)** via the ulam() function from the rethinking package. This enabled efficient posterior sampling and robust inference of both global and group-level effects.

**Prior Distributions**

A white sheet with black text

AI-generated content may be incorrect.  
To regularize parameter estimation and promote generalization, weakly informative priors were used. All regression coefficients were assigned Normal distributions centered at 0.

Each region was assigned its own intercept, drawn from a common distribution.  
This hierarchical prior structure enables partial pooling, improving stability and generalization.

Finally, a Student‑t was used for the residual standard deviation to enhance robustness against outliers.

**Posterior Summary and Convergence Diagnostics.**  
Table X displays the posterior summaries for all model parameters, including mean estimates, standard deviations, 89% credible intervals, and convergence metrics. The parameters bSb\_SbS​ (smoker) and bIb\_IbI​ (BMI × smoker interaction) showed the strongest effects, with posterior means of 2.38 and 0.88 respectively, indicating that smoking substantially increases expected medical charges, especially in individuals with higher BMI.

Convergence diagnostics were satisfactory for all parameters. The Gelman-Rubin statistic (R^\hat{R}R^) was approximately 1.00 for all parameters, and the effective sample size (ESS bulk) exceeded 1000, suggesting well-mixed chains and reliable posterior inference. These diagnostics confirm the validity of the MCMC results and the stability of parameter estimates.

# **4. Results**

Frequentist Model:  
- RMSE: 6062.0  
- R-squared: 0.75  
  
Bayesian Model (quap):  
- Posterior means aligned closely with `lm()` estimates  
- WAIC: 12789.2  
- Effective sample size (`n\_eff`) > 1000 for all parameters  
- Rhat ≈ 1 for all parameters  
  
Bayesian Model (ulam):  
- 4 chains, 2000 iterations, warmup = 500  
- Rhat = 1.00 for all parameters  
- Runtime: ~35 seconds  
  
Visualization: Posterior plots, trace plots, and predictive distributions clearly show parameter uncertainty and model fit.

# **6. Discussion**

Bayesian regression offers interpretive advantages in quantifying uncertainty and inspecting parameter distributions. Unlike frequentist models, it allows simulation-based predictions and richer diagnostics (e.g., WAIC). While predictive accuracy was similar, Bayesian methods provided more nuanced understanding, particularly useful when decision-making under uncertainty is required.

# **7. Conclusion / Future Work**

**7.1 Conclusion**  
  
Bayesian modeling successfully predicted healthcare charges with comparable accuracy to frequentist regression, while enhancing interpretability. This supports the use of Bayesian methods in policy applications and future healthcare analytics.  
  
**7.2 Future Work**  
1. Expand the model with interaction terms and nonlinear effects using splines.  
2. Explore hierarchical models to account for group-level differences (e.g., by region).  
3. Apply model comparison using LOO-CV and posterior predictive checks.  
4. Extend to other healthcare datasets to test generalizability.

# **9. References**

- Lantz, B. (2019). \*Machine Learning with R\*. Packt Publishing.  
- McElreath, R. (2020). \*Statistical Rethinking\* (2nd ed.). CRC Press.  
- Kaggle dataset: https://www.kaggle.com/datasets/mirichoi0218/insurance

# 10. Appendix

- R code used for Bayesian and frequentist models (quap, ulam, lm)  
- Additional figures: trace plots, posterior densities, WAIC table  
- Data preprocessing steps and summary sta