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

May 6, 2025

ABSTRACT

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

Keywords Bayesian Linear regression · Healthcare Cost Forecasting

1 Introduction

The prediction of healthcare costs plays an essential 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 uses 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.

3 Exploratory Data Analysis (EDA)

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

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

Figure 1: Dataset

1. Distribution of charges

The target variable, charges, is heavily skewed to the right. Most observations fall below $20,000, but a small number of high-cost outliers exist above $40,000. This suggests the presence of heteroskedasticity, indicating that models which allow for non-constant variance, such as Bayesian regression, may be more appropriate for accurate estimation.

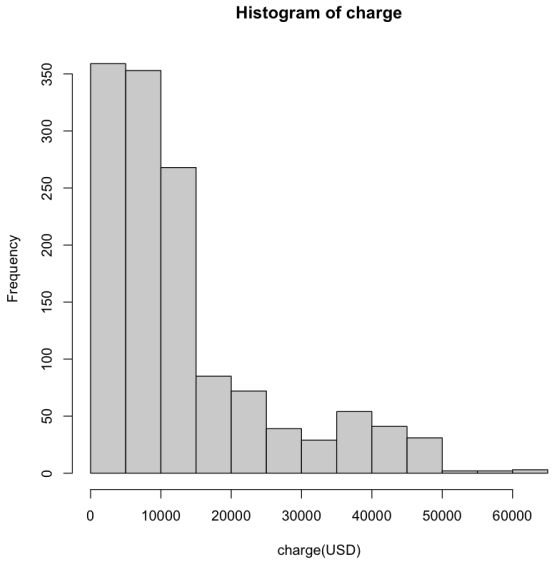

Figure 2: Distribution of medical charges. Most values are under $20,000, with some outliers.

2. Demographic Relationships

Charges and Sex

There is no significant difference in the distribution of charges between males and females. This suggests that sex may not be a strong standalone predictor and was excluded from the final model.

Charges and Number of Children

Charges for insurance with 4-5 children covered tend to decrease slightly. Overall, the number of children alone shows limited predictive power.

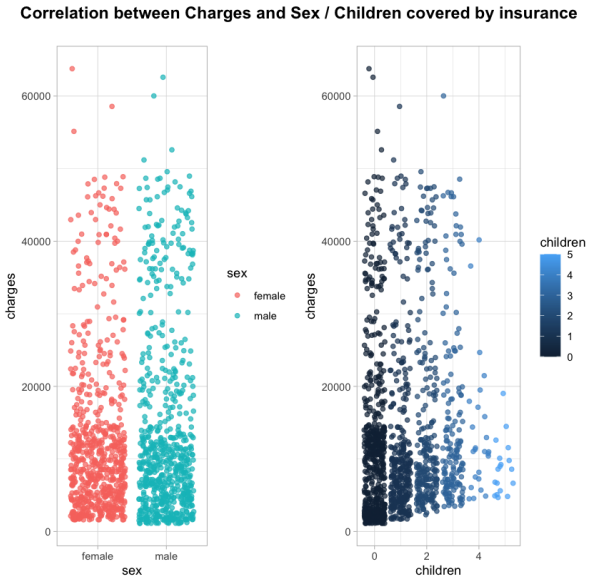

Figure 3: Charges by sex and number of children. Minimal difference by sex; slight decrease for 4–5 children.

3. Age and BMI Correlation

Charges and Age

There is a clear positive correlation between age and medical charges. Older individuals tend to incur higher medical costs. Layered data patterns suggest possible interaction effects with other variables.

Charges and BMI

Unlike age, BMI shows a weaker and more scattered correlation. There is a slight upward trend for higher BMI values (above 35), but the effect is not consistent.

4. Lifestyle and Interaction Effects

Smoker Status and Regional Effect

Smoker status has a strong positive effect on medical charges. The regional effect, however, is negligible.

Interaction: BMI $\times$ Smoker

This interaction effect is much stronger than others. For smokers, charges increase sharply with higher BMI, while for non-smokers the trend is flat. This interaction was included in the final model.

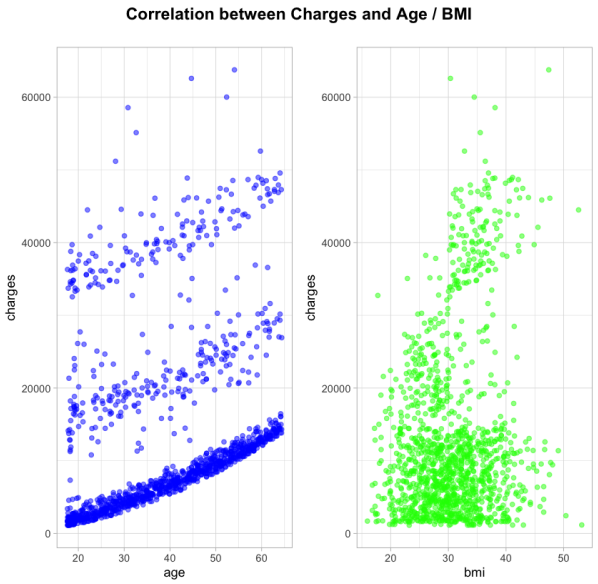

Figure 4: Correlation between age and BMI with charges. Age shows stronger positive association.

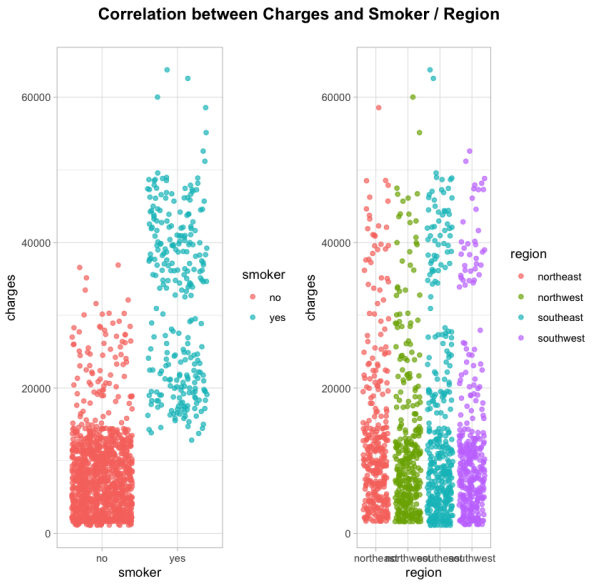

Figure 5: Smoker status strongly increases charges; regional effect is negligible.

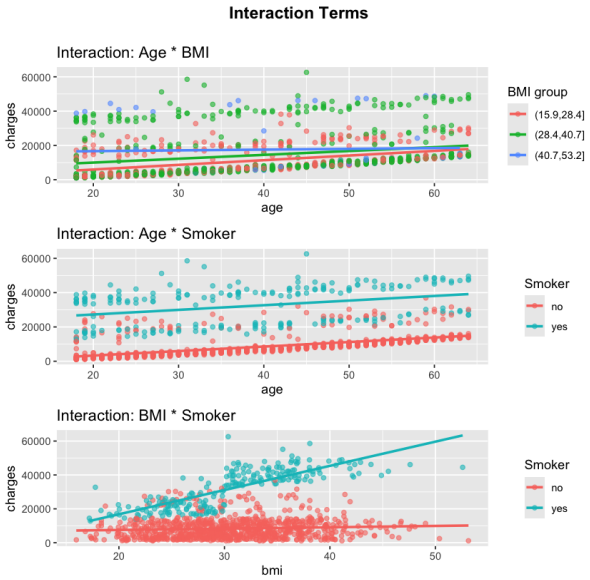

Figure 6: Interaction effects: BMI $\times$ Smoker shows strongest effect and is included in the final model.

4 Methods

This project compares two approaches to healthcare cost modeling: traditional linear regression using `lm()` in R, and Bayesian linear regression using `quap()` and `ulam()` from the `rethinking` package.

4.1 Frequentist Model

The frequentist model was implemented via the `lm()` function. It assumes linearity, independence of errors, and homoscedasticity. The model specification was:

```
charges ~ age + age^2 + bmi *smoker + children + region
```

Model summary:

- Residual standard error: 4834
- Multiple R-squared: 0.8409
- Adjusted R-squared: 0.8397
- F-statistic: 701.1 on 8 and 1061 DF, p-value: $< 2.2 \times 10^{-16}$

Significant predictors included age, BMI, smoker status, and regions. The interaction between BMI and smoker also had a strong effect.

4.2 Bayesian Model with Hierarchical Structure

To account for regional variability, group-level(hierarchical) intercepts are implemented. The model includes:

- Nonlinear effect: age^2 capturing increasing marginal cost.
- Interaction term: BMI $\times$ smoker to account for compounding health risks.

The model was implemented using Hamiltonian Monte Carlo (HMC) via `ulam()` for efficient posterior sampling and robust inference.

Model Specification:

$$\text{charges}_i \sim Normal(\mu_i, \sigma)$$
$$\mu_i = a_{region[i]} + b_A \cdot age_i + b_{A2} \cdot age_i^2 + b_B \cdot bmi_i + b_S \cdot smoker_i + b_I \cdot (bmi_i \cdot smoker_i) + b_C \cdot children_i$$
$$a_{region} \sim Normal(\bar{a}, \sigma_{region})$$

Prior Distributions:

$$\text{a}_{region} \sim Normal(a_{bar}, \sigma_{region})$$
$$a_{bar} \sim Normal(0, 1)$$
$$b_A \sim Normal(0, 1)$$
$$b_{A2} \sim Normal(0, 0.5)$$
$$b_B \sim Normal(0, 1)$$
$$b_S \sim Normal(0, 1)$$
$$b_I \sim Normal(0, 2)$$
$$b_C \sim Normal(0, 1)$$
$$\sigma_{region} \sim Exponential(1)$$
$$\sigma \sim Student{-}t(3, 0, 2)$$

3.3 Posterior Summary and Convergence Diagnostics

Posterior estimates included means, standard deviations, and 89% credible intervals. Notably:

- b_S (smoker): posterior mean ≈ 2.38
- b_I (BMI $\times$ smoker): posterior mean ≈ 0.88

Convergence diagnostics:

- Gelman-Rubin $\hat{R} \approx 1.00$ for all parameters
- Effective sample sizes (ESS bulk) > 1000

These confirm well-mixed chains and stable MCMC results.

Posterior Summary and Convergence Diagnostics

	mean	sd	5.5%	94.5%	rhat	ess_bulk
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
a[1]	0.75647597	0.037017649	0.69737312	0.81534236	1.000594	3179.635
a[2]	0.70632372	0.035192934	0.65071844	0.76272944	1.000708	3277.199
a[3]	0.65588925	0.034848647	0.60097880	0.71189982	1.000401	3117.549
a[4]	0.65382788	0.035675805	0.59686141	0.71071261	1.000061	3449.681
a_bar	0.68656155	0.086205270	0.59139100	0.78273907	1.002358	1799.225
bA	0.36646392	0.013438788	0.34521645	0.38798128	1.000753	7265.132
bA2	0.07366381	0.015966326	0.04860666	0.09909054	1.000337	3353.504
bB	0.01035157	0.015337207	-0.01408597	0.03488432	1.000410	5432.374
bS	2.37726936	0.032430441	2.32550615	2.42898165	1.000930	6261.116
bI	0.87481319	0.032263668	0.82395350	0.92712248	1.000618	6290.132
bC	0.06696070	0.011538463	0.04878287	0.08553685	1.000382	3883.143
sigma_region	0.10386663	0.109011462	0.02762506	0.25992625	1.001658	1815.567
sigma	0.48179579	0.009147451	0.46740256	0.49655605	1.000353	7835.471

Figure 7: Posterior Summary

3.4 Posterior Distribution Visualization

Posterior samples of parameters were visualized using pairwise correlation plots.

- Most parameter pairs have weak correlations (near zero), indicating low collinearity.
- Distributions of parameters appear approximately normal.

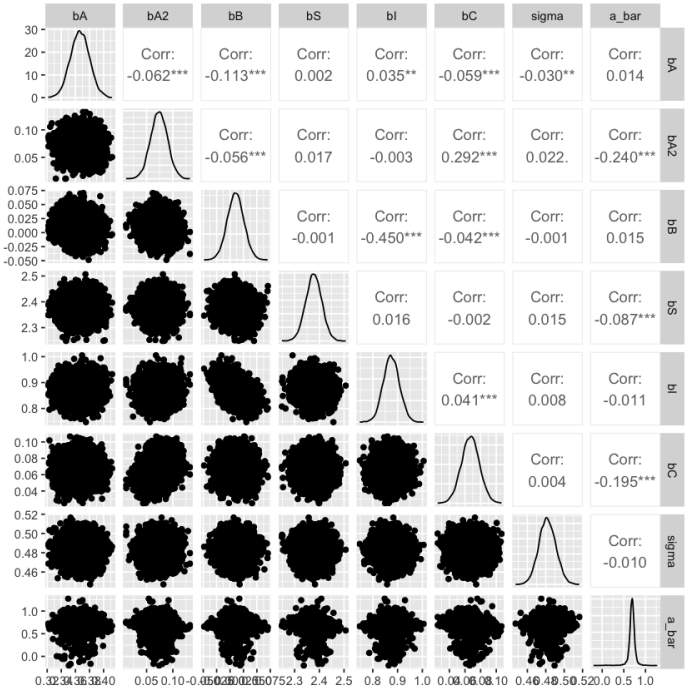

Figure 8: Pairwise posterior distributions and correlation summary for parameters.

5 Result

4.1 Trank Plot Diagnostics

To further assess mixing and convergence, rank plots of parameter draws across chains is inspected.

- Visualizes the Tranks of parameter draws for each chain.
- All chains appear well mixed with no major divergences.

4.2 Traceplot Diagnostics

To evaluate sampling performance, traceplots for each parameter were analyzed.

- All chains mix well, showing no signs of divergence or drift.
- Burn-in phase is shaded gray.
- The results suggest that convergence is satisfactory.

4.3 Frequentist vs Bayesian Comparison

Model	RMSE	MAE	R^2	WAIC
Frequentist Linear Regression	4930.554	2952.016	0.8409	NA
Bayesian Linear Regression	4796.795	2888.051	0.8430	1857.318

- Bayesian model slightly outperformed the frequentist model in both RMSE and MAE.
- R^2 values were nearly identical, indicating similar explanatory power.
- WAIC was available only for the Bayesian model and supports model validation.

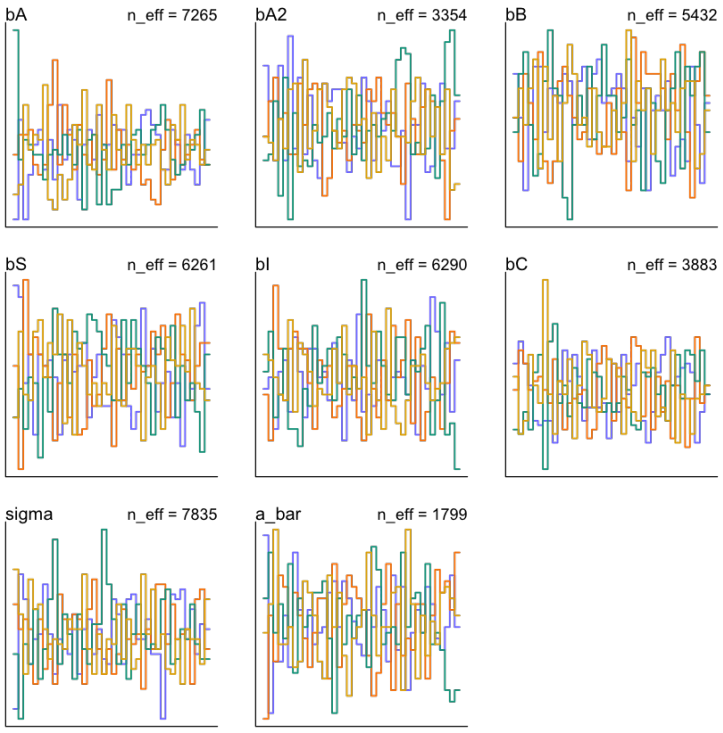

Figure 9: Trank plots showing parameter sampling consistency across chains.

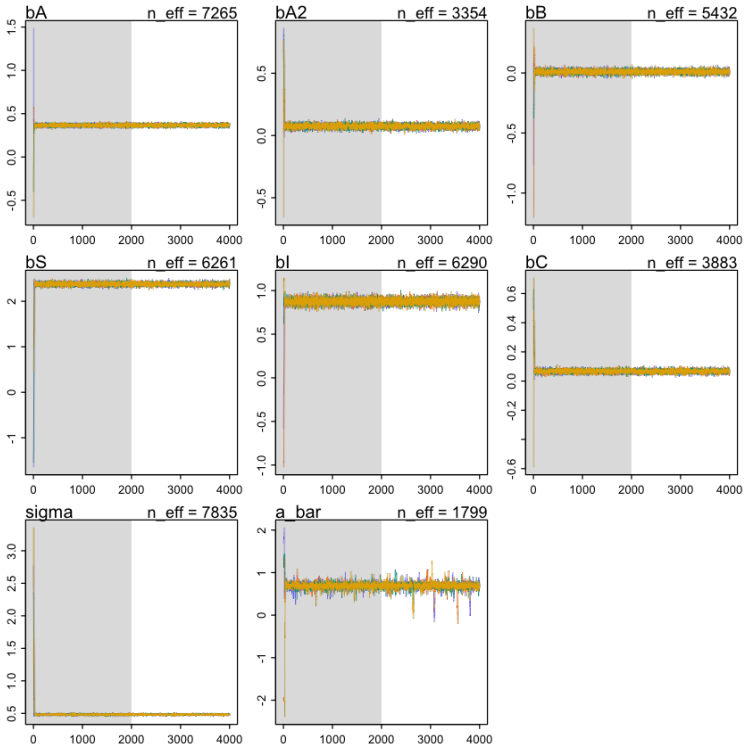

Figure 10: Traceplots of MCMC samples for selected parameters across chains.

- Bayesian approach allows for uncertainty quantification and interpretable posterior estimates.

4.4 Conclusion and Limitations

Key Findings

- Bayesian regression offered competitive predictive performance.
- Interaction terms and nonlinear effects improved expressiveness.
- Posterior distributions allowed deeper interpretation.

Limitations

- Small dataset size ($n = 1338$) may limit generalizability.
- The model assumes Gaussian residuals (can be improved with flexible likelihoods).
- Modest regional effects — future work may explore more complex hierarchical structures.