Healthcare Costs Prediction Using Bayesian Linear Regression

A Comparison with Frequentist Models

Yejin Hwang

01

Introduction

Abstract/Overview

- This project compares **Bayesian and Frequentist linear regression models** for predicting healthcare costs.
- The analysis uses the **Medical Cost Personal Dataset**, which includes demographic and medical variables.
- The **Bayesian model** incorporates **priors**, **nonlinear terms**, and **interactions**, producing **credible intervals** for interpretation.
- **Goal**: To evaluate each model's **predictive performance** and ability to **handle uncertainty**.

Introduction & Motivation

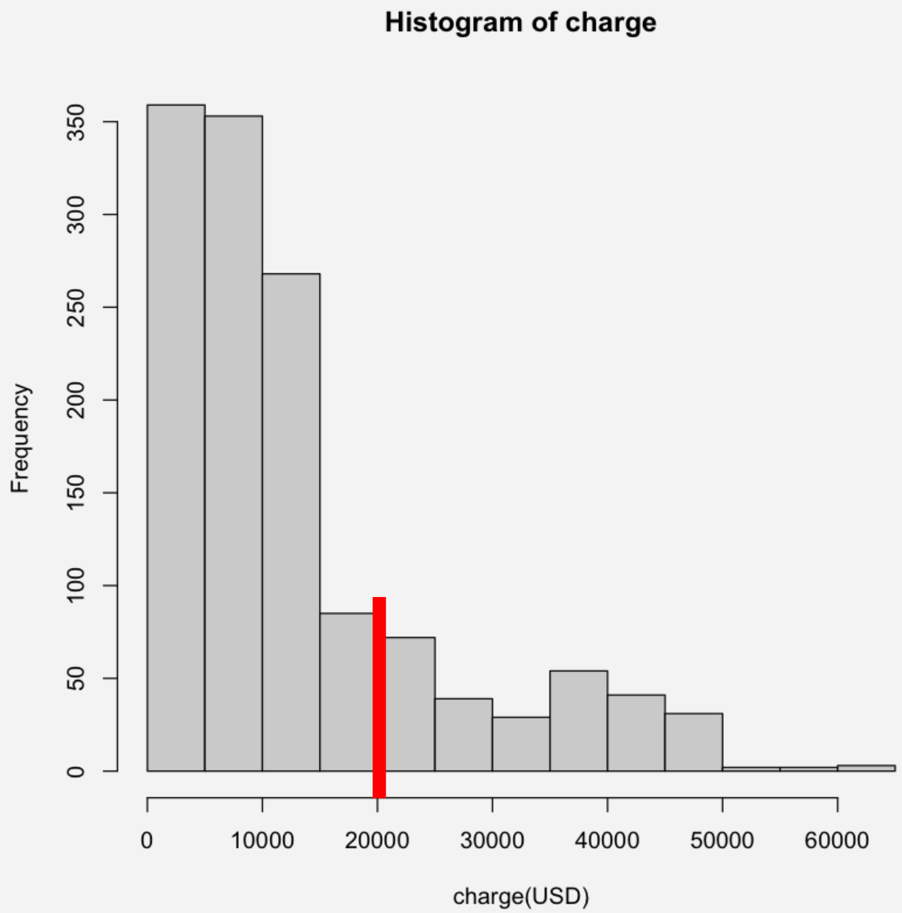
- **Accurate healthcare cost prediction** supports budgeting, insurance planning, and policy decisions.
- Traditional (frequentist) models provide point estimates but **lack uncertainty quantification**.
- **Bayesian methods** offer full posterior inference, flexible modeling, and clearer interpretation.
- **Research Question**: *Can Bayesian regression provide comparable prediction performance while offering better uncertainty estimates and interpretability?*

Dataset

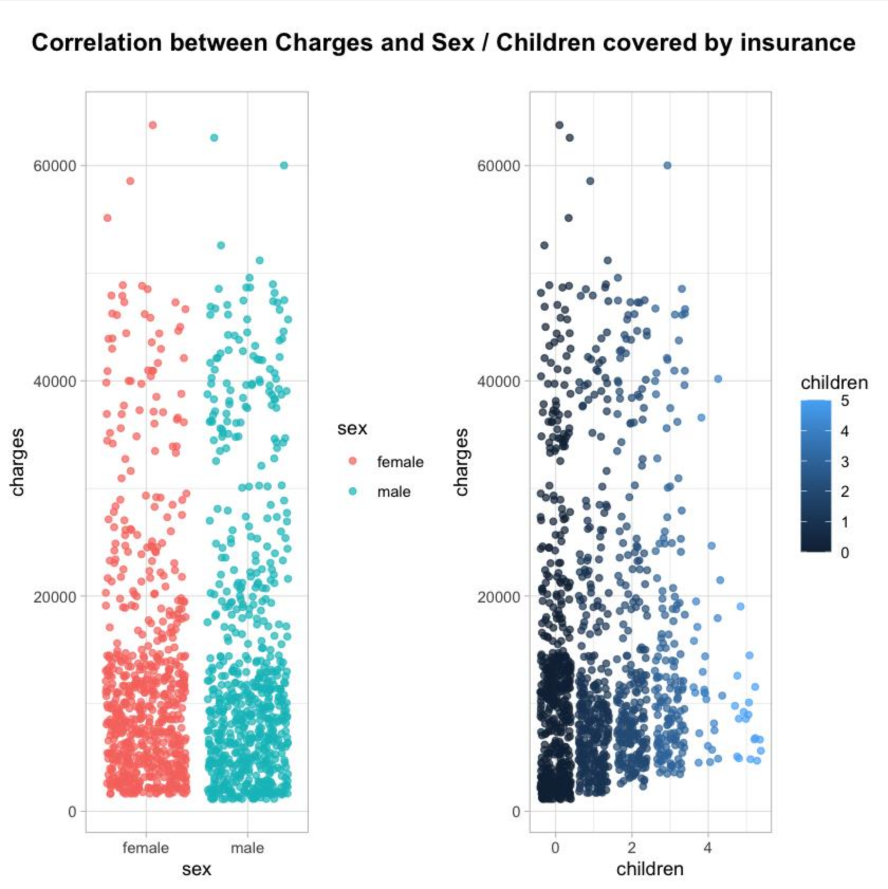
- **Dataset**: Medical Cost Personal Dataset (Kaggle)
- **Sample size**: 1338 individuals
- **Variables(7)**: age, sex, bmi, children, smoker, region, charges
- target variable : charges (Individual medical costs billed by health insurance)
- no missing values

18-64	f/m	15-53	0-5	yes/no	4(US)	Next slide!
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

Exploratory Data Analysis(EDA)



Most charges in this group are below $20,000, and high-cost outliers are rare.



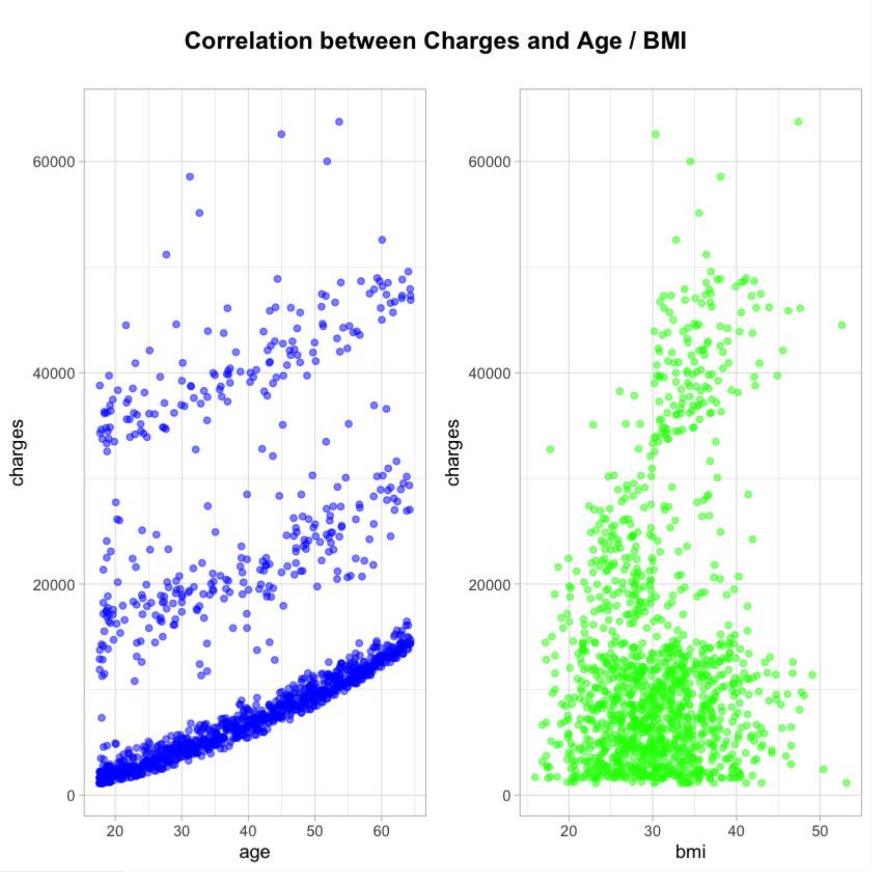
[Charges and Sex]

- Overall, 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 of healthcare costs in this dataset,

→ Therefore, it was **excluded from the final model**.

[Charges and Children]

- Charges for insurance with 4-5 children covered seems to go down.
- In general, the number of children alone still shows limited predictive power.

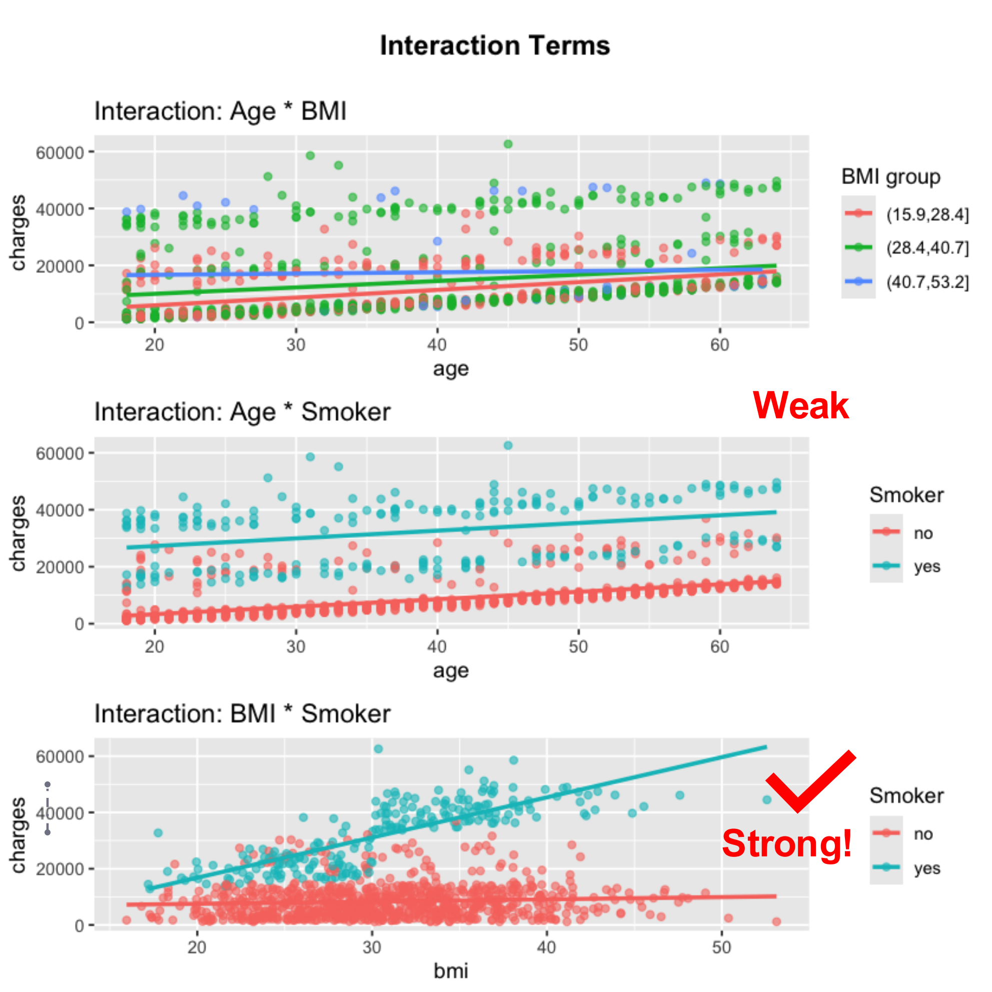


[Charges and Age]

- clear **positive correlation** between age and medical charges. As people get older, their medical expenses tend to increase.
- Furthermore, the layered appearance of data points suggests that age **may interact with other variables**, such as smoking status or BMI, influencing medical charges in a more complex way.

[Charges and BMI]

- Unlike age, the correlation here is weaker and more scattered. While there is a slight upward trend, particularly among individuals with higher BMI (above 35), the association is not consistent.
- This suggests that BMI alone is not a strong predictor of medical charges, and its effect may depend **on interactions with other variables** such as age or smoking status.



[Age * BMI] & [Age * Smoker]
There is a slight upward trend between variables, but the interaction effect remains relatively weak.

→ Given the weak effects, these interactions were excluded from the final model.

[BMI * Smoker]
This interaction is **much stronger**. For smokers, charges increase sharply with higher BMI, while for non-smokers, the pattern remains flat.

→ Thus, only BMI × Smoker was included as an interaction term in the final Bayesian model.

02

Methodology

Frequentist Model

$$\text{charges} \sim age + age^2 + bmi \times smoker + children + region$$

```
Call:
lm(formula = formula_1, data = df_train)

Residuals:
     Min       1Q   Median       3Q      Max
-14279.0  -1881.0  -1309.2   -444.7  30232.7
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1789.022    943.214  -1.897  0.05814 .
age                259.782     10.645  24.404  < 2e-16 ***
bmi                  5.269     28.458   0.185  0.85316
smokeryes       -19955.662   1847.198 -10.803  < 2e-16 ***
children           503.729    123.203   4.089 4.67e-05 ***
regionnorthwest   -672.383    421.652  -1.595  0.11109
regionsoutheast  -1159.931    426.484  -2.720  0.00664 **
regionsouthwest  -1165.246    426.087  -2.735  0.00635 **
bmi:smokeryes     1429.994     58.874  24.289  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4834 on 1061 degrees of freedom
Multiple R-squared:  0.8409,    Adjusted R-squared:  0.8397
F-statistic: 701.1 on 8 and 1061 DF,  p-value: < 2.2e-16
```

- **Nonlinear term**: age^2 to capture acceleration in healthcare costs with age.
- **Key interactions**: BMI × smoker
- **Significant predictors** include age, children, smoker, and bmi × smoker

Bayesian Model

Variable Processing

```r
df <- df %>%
  mutate(
    smoker = ifelse(smoker == "yes", 1, 0),
    sex = as.integer(factor(sex)) - 1,
    region = as.integer(factor(region)),  # 1~4
    charges = charges / 10000,            # scale down
    bmi = (bmi - mean(bmi)) / sd(bmi),
    age = (age - mean(age)) / sd(age)
  )
```

→ converted to binary (1 = "yes", 0 = "no")

→ converted to integer index(1-4)

→ scaled down by 10,000 to improve model stability

→ age and bmi standardized (mean = 0, sd = 1) for better convergence

```r
dat_list <- list(
  charges = df$charges,
  age = df$age,
  age2 = df$age^2,
  bmi = df$bmi,
  smoker = df$smoker,
  region = df$region,
  children = df$children
)
```

Model Fitting

$$
\begin{aligned}
\text{charges} &\sim \text{Normal}(\mu, \sigma) \\
\mu &= a[\text{region}] \\
&\quad + b_A \cdot \text{age} \\
&\quad + b_{A2} \cdot \text{age}^2 \\
&\quad + b_B \cdot \text{bmi} \\
&\quad + b_S \cdot \text{smoker} \\
&\quad + b_I \cdot \text{bmi} \cdot \text{smoker} \\
&\quad + b_C \cdot \text{children} \\
a[\text{region}] &\sim \text{Normal}(a_{\text{bar}}, \sigma_{\text{region}})
\end{aligned}
$$

- Fit using rethinking R package with HMC(Hamiltonian Monte Carlo)
- **Nonlinear term**: age² to capture acceleration in healthcare costs with age.
- **Key interactions**: BMI × smoker
- **Group-level intercepts** for region, using **partial pooling** through hierarchical priors to capture individual and group-level effects.

→ **This allows the model to generalize better while capturing important structure in the data.**

Priors

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

- Normal priors used for all coefficients (centered at 0).
- Narrower prior for age² to regularize nonlinear effect.
- Wider prior for bmi × smoker to allow stronger interaction.
- Each region gets its own intercept a[region], drawn from a common distribution

→ It allows partial pooling and better generalization.

- Student-t prior on σ to improve robustness against outliers.

Posterior summary

Effective sample size

	mean	sd	5.5%	94.5%	rhat	ess_bulk
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
a[1]	0.754815804	0.037101148	0.69524852	0.81319317	1.0010766	3405.729
a[2]	0.704751033	0.035125848	0.64867023	0.76120848	1.0015224	3356.370
a[3]	0.655461571	0.035061227	0.59918968	0.71162514	1.0006304	3223.618
a[4]	0.653460423	0.035336899	0.59675505	0.70965157	1.0012693	3228.597
a_bar	0.687475746	0.071921111	0.58915328	0.78082251	1.0013757	2766.301
bA	0.366438888	0.012979543	0.34544085	0.38730460	1.0001156	7316.193
bA2	0.074006091	0.016113620	0.04838000	0.09979718	1.0010022	3386.566
bB	0.009912174	0.015423978	-0.01484373	0.03410852	1.0010531	5486.035
bS	2.377448367	0.032956742	2.32380000	2.42885000	0.9999566	7257.481
bI	0.875087819	0.031414759	0.82490155	0.92494169	1.0012652	6103.980
bC	0.067317408	0.011645894	0.04831901	0.08566258	1.0006048	3932.715
sigma_region	0.099450858	0.095837541	0.02703233	0.24677113	1.0012941	2502.534
sigma	0.481895559	0.009435628	0.46686184	0.49730700	1.0003315	7616.260

The Strongest effects

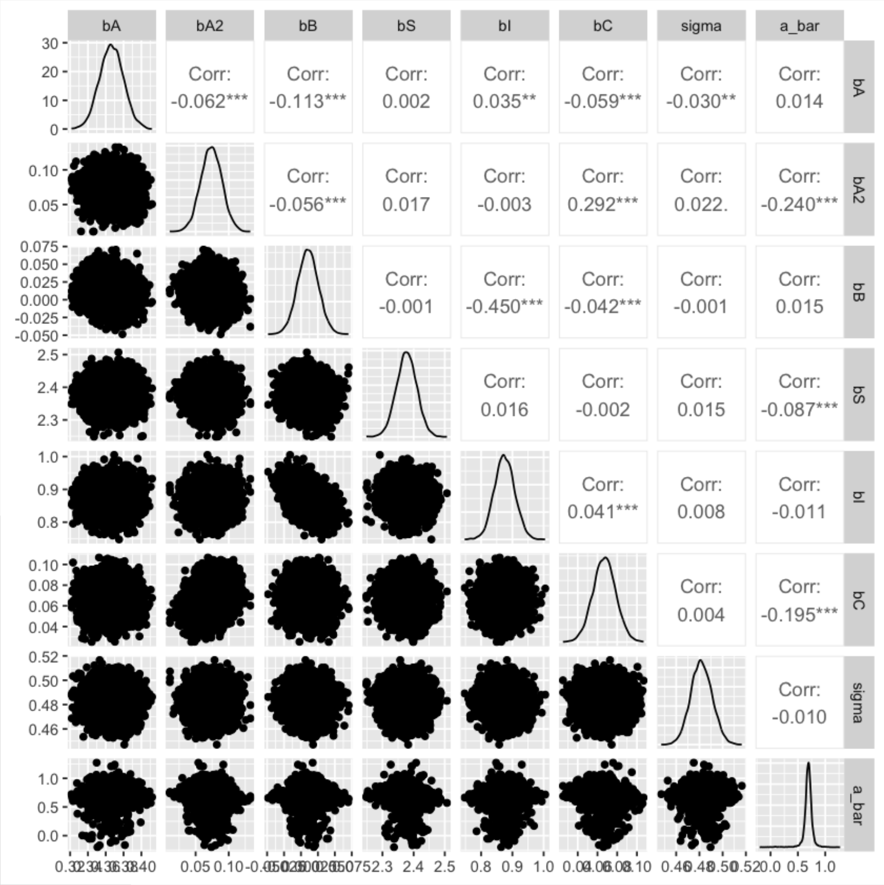
Rhat ≈ 1.00 Ess bulk > 1000 for all

→ **good convergence and effective sampling**

→ Being a smoker increases expected charges

→ Charges increase more steeply with BMI for smokers

Posterior distribution



- Posterior samples of parameters are shown with pairwise correlations.
- Most parameters have weak correlations (near zero), suggesting low collinearity.
- Distributions look approximately normal.

03

RESULT

Result

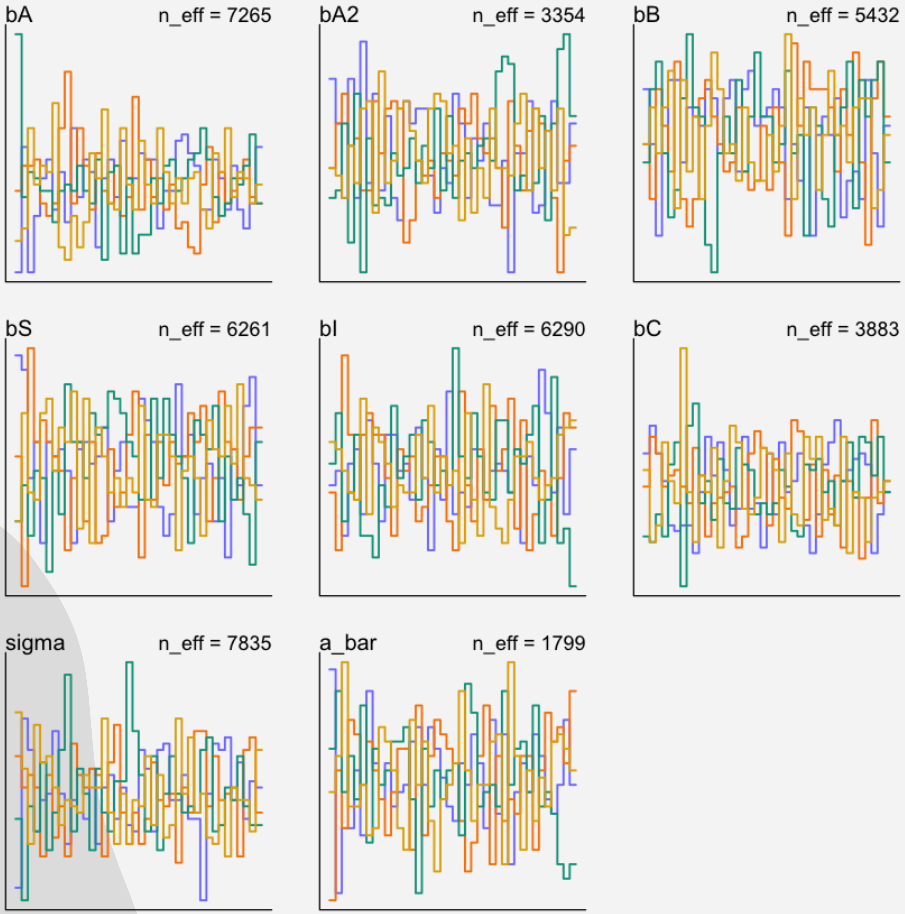
```
Hamiltonian Monte Carlo approximation
8000 samples from 4 chains

Sampling durations (seconds):
  chain_id warmup sampling total
1        1   7.10    11.69 18.79
2        2   6.86     7.36 14.22
3        3   6.65    16.75 23.40
4        4   7.22     7.46 14.68
```

- The model was fit using 8000 samples with 4 chains, each running for 4000 iterations.

→ This provides robust posterior samples for each parameter.

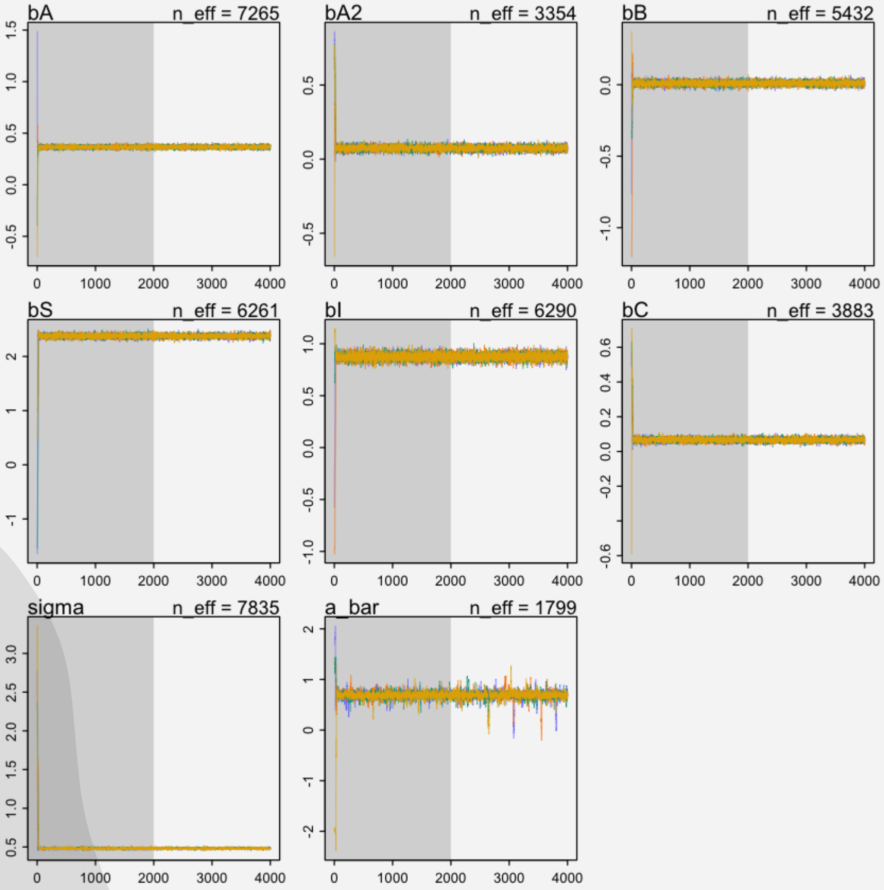
- Each chain converged quickly, with total runtimes between 14–23 seconds.

Trankplot



- Visualizes the ranks of parameter draws across chains.
- All chains are well mixed, no major divergences observed.

Traceplot



- All chains mix well with no signs of divergence or drift.
- Burn-in (warmup) phase is shaded gray.
- The results suggest **convergence is satisfactory** and posteriors are reliable for interpretation.

Frequent vs Bayesian

	Model	RMSE	MAE	R2	WAIC
1	Frequentist Linear Regression	4930.554	2952.016	0.8409286	NA
2	Bayesian Linear Regression (full)	4796.795	2888.051	0.8429861	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 available only for Bayesian model, which aids in model comparison and validation
- Bayesian model provides uncertainty quantification and more interpretable posterior estimates

Conclusion/Limitation

Key Findings

• Bayesian regression offered competitive predictive performance

• Interaction terms and nonlinear effects improved model expressiveness

• Posterior distributions allow uncertainty visualization and deeper interpretation

Limitations

• Small dataset size (n = 1338) may limit generalizability

• The model assumes Gaussian residuals (could be improved with more flexible likelihoods)

• Regional effects are modest. more hierarchical structure (e.g., for smoker subgroups) could be explored

Thank you