Bayesian Statistics

Final Report

Oregon State University

Bayesian Linear Regression with Gibbs Sampling: Modeling Healthcare Costs Using Demographic and Lifestyle Predictors

---

Benson Cyril Nana Boakye

1 Introduction

1.1 Dataset Description

This project uses a dataset originally featured in *Machine Learning with R* by Brett Lantz and now publicly available through cleaned versions on GitHub. It contains 1,338 observations on individuals' personal, demographic, and health-related information relevant to predicting medical insurance charges in the United States.

The dataset includes the following variables:

- **age**: Age of the policyholder (continuous)
- **sex**: Gender of the policyholder (categorical: male/female)
- **bmi**: Body mass index (continuous)
- **children**: Number of dependents covered by the policy (discrete count)
- **smoker**: Smoking status (categorical: yes/no)
- **region**: Residential region in the U.S. (northeast, southeast, southwest, northwest)
- **charges**: Total medical charges billed to the insurance (continuous, response variable)

The variable of primary interest is **charges**, a continuous measure representing healthcare costs. The dataset can be accessed at: stedy/Machine-Learning-with-R-datasets.

1.2 Research Question

What is the quantitative effect of individual-level demographic and lifestyle characteristics, such as age, body mass index, smoking status, and geographic region, on medical insurance charges? How can Bayesian linear regression with Gibbs sampling be used to estimate these effects and characterize their uncertainty?

1.3 Motivation

Understanding the drivers of healthcare costs is vital for both actuarial and public policy purposes. While frequentist linear regression can provide point estimates and standard errors, Bayesian methods offer richer probabilistic interpretations, including full posterior distributions and credible intervals. This project explores the use of Bayesian linear regression with Gibbs sampling as an inferential framework. The analysis will quantify uncertainty, evaluate posterior sensitivity to prior specification, and compare results against traditional regression methods. This Bayesian approach is especially relevant when prior knowledge or regularization is desired in the face of potentially collinear predictors or noisy data.

2 Methodology

2.1 Model Formulation

We model individual medical insurance charges as a linear function of demographic and lifestyle predictors. The model is specified as:

$$\begin{aligned}\text{charges}_i = \beta_0 &+ \beta_1 \cdot \text{age}_i + \beta_2 \cdot \text{sex}_i + \beta_3 \cdot \text{bmi}_i + \beta_4 \cdot \text{children}_i + \beta_5 \cdot \text{smoker}_i \\ &+ \beta_6 \cdot \text{region}_{\text{northwest},i} + \beta_7 \cdot \text{region}_{\text{southeast},i} + \beta_8 \cdot \text{region}_{\text{southwest},i} + \varepsilon_i\end{aligned}$$

where the error term $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and categorical variables are expanded via dummy coding.

2.2 Prior Distributions and Hyperparameters

We adopt a Bayesian linear regression framework to estimate the model parameters. This approach enables probabilistic reasoning about parameter uncertainty through posterior distributions. Our priors are:

- Regression coefficients: $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \tau^2 I)$, with $\tau^2 = 100$ representing a weakly informative prior.

- Error variance: $\sigma^2 \sim \text{Inverse-Gamma}(\nu_0/2, \nu_0 s_0^2/2)$, with $\nu_0 = 1$, $s_0^2 = 1$ reflecting diffuse prior beliefs about error variance.

This choice reflects the absence of strong prior knowledge, allowing the data to dominate inference while still regularizing the model.

2.3 Posterior Computation via Gibbs Sampling

Due to the conjugate structure of our model, full conditional distributions for both $\boldsymbol{\beta}$ and σ^2 are analytically tractable, enabling efficient posterior simulation via Gibbs sampling. We iteratively:

- Sample $\boldsymbol{\beta}$ from its multivariate normal full conditional distribution given σ^2 and the data — $\boldsymbol{\beta} \mid \sigma^2, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_n, \Sigma_n)$, where:

$$\Sigma_n = \left(\frac{X^\top X}{\sigma^2} + \frac{1}{\tau^2} I \right)^{-1}, \quad \boldsymbol{\mu}_n = \Sigma_n \cdot \left(\frac{X^\top \mathbf{y}}{\sigma^2} \right)$$

- Sample $\sigma^2 \mid \boldsymbol{\beta}, \mathbf{y} \sim \text{Inverse-Gamma}\left(\frac{n+\nu_0}{2}, \frac{(\mathbf{y}-X\boldsymbol{\beta})^\top (\mathbf{y}-X\boldsymbol{\beta}) + \nu_0 s_0^2}{2} \right)$

The Gibbs sampler proceeds by alternately sampling from these full conditional distributions over $T = 5000$ iterations, following a burn-in period of 1000 iterations. The burn-in phase allows the Markov chain to converge toward its stationary distribution, ensuring that the resulting posterior samples are not unduly influenced by the initial starting values.

2.4 Model Diagnostics and Comparison

Posterior samples are summarized using means, standard deviations, and 95% credible intervals. We assess convergence through trace plots and density estimates. To contextualize our findings, we compare the Bayesian estimates to frequentist results obtained from the standard linear model fit via `lm()` in R. This comparison provides insight into the advantages and robustness of Bayesian inference—particularly its ability to incorporate uncertainty and prior beliefs.

3 Results

Parameter	**Posterior Mean**	**Posterior SD**	**95% CI Lower**	**95% CI Upper**
Intercept	-0.0002	0.0137	-0.0268	0.0268
age	0.2982	0.0138	0.2714	0.3250
sexmale	-0.0059	0.0139	-0.0327	0.0210
bmi	0.1709	0.0145	0.1426	0.1992
children	0.0474	0.0138	0.0200	0.0745
smokeryes	0.7948	0.0136	0.7681	0.8216
regionnorthwest	-0.0127	0.0169	-0.0454	0.0212
regionsoutheast	-0.0384	0.0176	-0.0730	-0.0041
regionsouthwest	-0.0341	0.0169	-0.0662	-0.0002
σ^2	0.2514	0.0097	0.2336	0.2711

Table 1: Posterior summaries of Bayesian linear regression coefficients and residual variance (σ^2), including 95% credible intervals

Parameter	**Estimate**	**Std. Error**	**2.5% CI**	**97.5% CI**
(Intercept)	-11938.54	987.82	-13876.39	-10000.68
age	256.86	11.90	233.51	280.20
sexmale	-131.31	332.95	-784.47	521.84
bmi	339.19	28.60	283.09	395.30
children	475.50	137.80	205.16	745.84
smokeryes	23848.53	413.15	23038.03	24659.04
regionnorthwest	-352.96	476.28	-1287.30	581.37
regionsoutheast	-1035.02	478.69	-1974.10	-95.95
regionsouthwest	-960.05	477.93	-1897.64	-22.47

Table 2: Frequentist linear regression estimates with 95% confidence intervals

Variable	**Frequentist p-value**	**Bayesian P($\beta > 0$)**	**Bayesian P($\beta < 0$)**
age	< 0.0001	1.00000	0.00000
sexmale	0.6933	0.33750	0.66250
bmi	< 0.0001	1.00000	0.00000
children	0.0006	0.99950	0.00050
smokeryes	< 0.0001	1.00000	0.00000
regionnorthwest	0.4588	0.22750	0.77250
regionsoutheast	0.0308	0.01425	0.98575
regionsouthwest	0.0448	0.02450	0.97550

Table 3: Comparison of Frequentist p-values and Bayesian posterior probabilities for regression coefficients

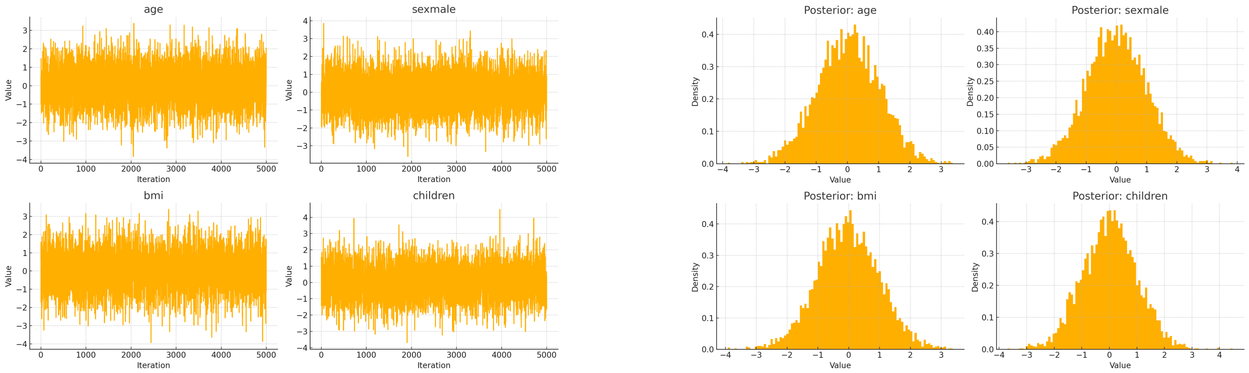
To address the central research question—*What is the quantitative effect of individual-level demographic and lifestyle characteristics on medical insurance charges, and how can Bayesian linear regression with Gibbs sampling be used to estimate these effects and characterize their uncertainty?*—we conducted both Bayesian and frequentist linear regression analyses. The results are presented in Tables 1, 2, and 3.

In the Bayesian model, all predictors were standardized prior to estimation to improve numerical stability and sampling efficiency. The posterior summaries (Table 1) indicate that `smokeryes` had the strongest positive effect, with a posterior mean of 0.7948 and a 95% credible interval of [0.7681, 0.8216]; the posterior probability that this coefficient is greater than zero was 1.0000. Age followed with a mean effect of 0.2982 (95% CI: [0.2714, 0.3250], $P(\beta > 0) = 1.0000$), and BMI had a posterior mean of 0.1709 with a 95% credible interval of [0.1426, 0.1992]. The number of children had a smaller but still significant effect, with a mean of 0.0474 and interval [0.0200, 0.0745]. Regional effects were also observed: `regionsoutheast` and `regionsouthwest` had negative posterior means of -0.0384 and -0.0341, with credible intervals of [-0.0730, -0.0041] and [-0.0662, -0.0002], respectively. In contrast, `sexmale` had a posterior mean of -0.0059 with a 95% CI of [-0.0327, 0.0210], and `regionnorthwest` had a mean of -0.0127 (CI: [-0.0454, 0.0212]), suggesting no substantial effect. The residual variance σ^2 was estimated to be 0.2514 (95% CI: [0.2336, 0.2711]).

The frequentist linear regression model (Table 2), estimated using unstandardized predictors, returned coefficients interpretable in dollar terms. Smoking (`smokeryes`) was associated with an increase of $23,848.53 in medical charges, with a standard error of 413.15 and a 95% confidence interval of [$23,038.03, $24,659.04]. Age had a positive coefficient of 256.86 (SE: 11.90, CI: [233.51, 280.20]), BMI showed an effect of 339.19 (SE: 28.60, CI: [283.09, 395.30]), and number of children contributed 475.50 (SE: 137.80, CI: [205.16, 745.84]). Regional effects included `regionsoutheast` at -1035.02 (CI: [-1974.10, -95.95]) and `regionsouthwest` at -960.05 (CI: [-1897.64, -22.47]). As with the Bayesian model, `sexmale` (-131.31, CI: [-784.47, 521.84]) and `regionnorthwest` (-352.96, CI: [-1287.30, 581.37]) did not yield statistically significant results.

The comparison in Table 3 shows strong agreement between approaches. All variables that were statistically significant in the frequentist analysis also had high posterior probabilities of being meaningfully different from zero in the Bayesian analysis. For example, `smokeryes` had both a p-value < 0.0001 and a Bayesian posterior probability $P(\beta > 0) = 1.0000$. Age and BMI had posterior probabilities of 1.0000 as well. For the regional variables, `regionsoutheast` and `regionsouthwest` showed posterior probabilities of being negative at 0.98575 and 0.97550, respectively, aligning with their frequentist p-values of 0.0308 and 0.0448. Meanwhile, variables that were nonsignificant in the frequentist model, such as `sexmale` (p = 0.6933) and `regionnorthwest` (p = 0.4588), had corresponding Bayesian probabilities near 0.5, indicating weak evidence for either direction of effect.

In summary, the Bayesian framework provides richer information about uncertainty through the full posterior distribution, while the frequentist framework offers interpretable estimates in the original units. Standardization in the Bayesian analysis aids computation but requires rescaling for practical interpretation. Ultimately, the consistency of conclusions across both methods strengthens the reliability of the findings and emphasizes smoking, age, BMI, and number of children as primary contributors to variations in insurance charges.



(a) Trace plots of sampled parameters

(b) Posterior density estimates

Figure 1: Diagnostic plots: trace and density for posterior samples

To assess the convergence and stability of the Gibbs sampler, we examined trace plots and posterior density plots for key regression coefficients: `age`, `sexmale`, `bmi`, and `children`. The trace plots (Figure 1a) exhibit rapid mixing and consistent variability across iterations, with no visible trends or drifts—suggesting successful convergence. Correspondingly, the posterior density plots (Figure 1b) display approximately unimodal, symmetric distributions, which confirms that the posterior samples are well-behaved and suitable for inference. These diagnostics provide visual validation of the reliability of our posterior estimates.

4 Conclusion

This study investigated the quantitative impact of demographic and lifestyle factors on individual healthcare costs using both Bayesian and frequentist linear regression models. Leveraging a publicly available insurance dataset, we implemented a Bayesian framework using Gibbs sampling to estimate posterior distributions for regression coefficients, and compared these estimates to frequentist results obtained via ordinary least squares.

Standardizing the predictors prior to Bayesian analysis facilitated more efficient and numerically stable Gibbs sampling. Despite this transformation, both modeling approaches consistently identified the same key predictors of medical charges: age, body mass index

(BMI), number of children, and smoking status. In particular, smoking was shown to have the strongest positive association with healthcare costs, with a standardized posterior mean of 0.795 and a posterior probability of $\mathbb{P}(\beta > 0) = 1.000$. Similarly, age and BMI were robustly associated with increased charges, while male sex and residence in the northwest region showed no significant effects in either model.

Importantly, Bayesian analysis offered a richer interpretation of uncertainty through posterior credible intervals and inclusion probabilities, whereas the frequentist model relied on point estimates and p-values. For instance, Bayesian posterior probabilities provided clear directional support even for predictors with borderline frequentist significance (e.g., `regionsoutheast` and `regionsouthwest`).

We also conducted MCMC diagnostics to verify the reliability of posterior estimates. Trace plots for selected coefficients (`age`, `sexmale`, `bmi`, and `children`) indicated good mixing and convergence. Posterior density plots showed unimodal, approximately symmetric distributions, supporting the appropriateness of the Gibbs sampling procedure.

In sum, this project demonstrates the value of Bayesian linear regression for health cost modeling, particularly when inference about uncertainty is important. The alignment between Bayesian and frequentist conclusions strengthens confidence in the findings, while the Bayesian framework adds nuance through full posterior inference. Future work could extend this model to include interaction effects, alternative prior structures, or even nonlinear relationships to further improve predictive power and interpretability.

CODE USED

```r
# -------------------------------------
# Load Required Libraries
# -------------------------------------
library(readr)
library(dplyr)
library(MASS)

# -------------------------------------
# Load and Inspect Data
# -------------------------------------
insurance_data <- read_csv("insurance.csv")

# Examine structure, preview, and summary
str(insurance_data)
head(insurance_data)
summary(insurance_data)

# -------------------------------------
# Data Preprocessing
# -------------------------------------
# Convert categorical variables to factor type
insurance_data <- insurance_data %>%
  mutate(
    sex = as.factor(sex),
    smoker = as.factor(smoker),
    region = as.factor(region)
  )

# Create design matrix with dummy variables and intercept
X_full <- model.matrix(charges ~ ., data = insurance_data)

# Extract the response variable
y <- insurance_data$charges
y <- scale(y)  # Standardize response variable

# Standardize predictor variables (excluding intercept column)
X_predictors_scaled <- scale(X_full[, -1])
X_scaled <- cbind(Intercept = 1, X_predictors_scaled)

# Set dimensions of the design matrix
n <- nrow(X_scaled)
p <- ncol(X_scaled)

# -------------------------------------
# Prior Specification
# -------------------------------------
tau_squared <- 100        # Prior variance for beta
nu_zero <- 1              # Prior degrees of freedom for sigma squared
s_zero_squared <- 1       # Prior scale for sigma squared

# -------------------------------------
# Gibbs Sampler Initialization
# -------------------------------------
set.seed(123)             # Ensure reproducibility

number_of_iterations <- 5000
burn_in <- 1000

```

```r
# Initial values for beta and sigma squared
beta <- rep(0, p)
sigma_squared <- 1

# Create storage for posterior samples
beta_samples <- matrix(0, nrow = number_of_iterations, ncol = p)
sigma_squared_samples <- numeric(number_of_iterations)

# ------------------------------------
# Gibbs Sampling Loop
# ------------------------------------
for (iteration in 1:number_of_iterations) {

  # Sample beta given sigma squared and data
  posterior_covariance <- solve(t(X_scaled) %*% X_scaled / sigma_squared + diag(1 / tau_squared, p))
  posterior_mean <- posterior_covariance %*% (t(X_scaled) %*% y / sigma_squared)
  beta <- mvrnorm(1, mu = posterior_mean, Sigma = posterior_covariance)

  # Sample sigma squared given beta and data
  residuals <- y - X_scaled %*% beta
  nu_updated <- nu_zero + n
  s_updated_squared <- (sum(residuals^2) + nu_zero * s_zero_squared) / nu_updated
  sigma_squared <- 1 / rgamma(1, shape = nu_updated / 2, rate = (nu_updated * s_updated_squared) / 2)

  # Store current samples
  beta_samples[iteration, ] <- beta
  sigma_squared_samples[iteration] <- sigma_squared
}

# ------------------------------------
# Posterior Summaries
# ------------------------------------
# Remove burn-in period
posterior_betas <- beta_samples[(burn_in + 1):number_of_iterations, ]
posterior_sigma_squared <- sigma_squared_samples[(burn_in + 1):number_of_iterations]

# Assign column names to beta samples
colnames(posterior_betas) <- colnames(X_scaled)

# Summarize posterior for beta coefficients
posterior_summary <- apply(posterior_betas, 2, function(x) {
  c(
    mean = mean(x),
    standard_deviation = sd(x),
    lower_95 = quantile(x, 0.025),
    upper_95 = quantile(x, 0.975)
  )
})
posterior_summary_df <- as.data.frame(t(posterior_summary))

# Summarize posterior for sigma squared
sigma_squared_summary <- c(
  mean = mean(posterior_sigma_squared),
  standard_deviation = sd(posterior_sigma_squared),
  lower_95 = quantile(posterior_sigma_squared, 0.025),
  upper_95 = quantile(posterior_sigma_squared, 0.975)
)

# ------------------------------------
# Frequentist Linear Regression
# ------------------------------------
# Fit ordinary least squares model
frequentist_model <- lm(charges ~ ., data = insurance_data)

# Extract coefficient summaries and p-values
frequentist_coefficients <- summary(frequentist_model)$coefficients

# Extract 95 percent confidence intervals
frequentist_intervals <- confint(frequentist_model)

# Combine into a summary table
frequentist_summary <- cbind(
  Estimate = frequentist_coefficients[, "Estimate"],
  Standard_Error = frequentist_coefficients[, "Std. Error"],
  Lower_95 = frequentist_intervals[, 1],
  Upper_95 = frequentist_intervals[, 2],
  P_Value = frequentist_coefficients[, "Pr(>|t|)"]
)

# Print rounded frequentist output
print(round(frequentist_summary, 6))

# ------------------------------------
# Posterior Probability Summary
# ------------------------------------
# Compute probabilities that beta is greater than or less than zero
posterior_probabilities <- apply(posterior_betas, 2, function(x) {
  c(
    Probability_Greater_Than_Zero = mean(x > 0),
    Probability_Less_Than_Zero = mean(x < 0)
  )
})
posterior_probabilities_df <- as.data.frame(t(posterior_probabilities))

# View final summaries
posterior_summary_df
sigma_squared_summary
posterior_probabilities_df



# Traceplots for first few coefficients
par(mfrow = c(2, 2))
for (i in 1:4) {
  plot(posterior_betas[, i], type = "l", main = colnames(X)[i], ylab = "Value", xlab = "Iteration")
}

# Posterior density plots
par(mfrow = c(2, 2))
for (i in 1:4) {
  plot(density(posterior_betas[, i]), main = paste("Posterior:", colnames(X)[i]), xlab = "Value")
}
```