Bayesian Regression Analysis for Predicting Health Insurance Charges Using Demographic and Lifestyle Factors

Harshamala Bandara (S/19/323)
DSC 4043 : Bayesian Statistics
Department of Statistics and Computer Science
Faculty of Science
University of Peradeniya
2025

Contents

1 Introduction 3

1.1 Background . . . 3

1.2 Research Question . . . 3

1.3 Hypotheses . . . 3

1.4 Importance and Objectives . . . 4

2 Methodology 5

2.1 Description of Dataset . . . 5

2.2 Exploratory Data Analysis (EDA) . . . 6

2.2.1 Structure of the dataset . . . 6

2.2.2 Distributions of numeric variables . . . 7

2.2.3 Scatterplots of numeric predictors vs Charges . . . 8

2.2.4 Boxplots for categorical predictors . . . 9

2.2.5 Correlations . . . 10

2.3 Bayesian Methods and Models . . . 10

2.4 Model Specification . . . 11

2.5 Prior Specification . . . 12

2.6 Model Checking . . . 13

3 Results and Discussion 13

3.1 Posterior Summaries . . . 13

3.2 Model Comparison . . . 15

3.3 Interpretation of findings and practical implications . . . 15

4 Conclusion and Recommendations 16

4.1 Key Findings . . . 16

4.2 Comparison with literature . . . 16

4.3 Limitations and suggestions 16

5 Appendix **19**

1 Introduction

1.1 Background

Health insurance charges depend on various individual traits, lifestyle choices, and social factors.Insurance companies and policymakers determine insurance charges by carefully analyzing a range of factors that indicate an individual's potential health risks, such as age, lifestyle habits, and regional differences in healthcare costs.Using statistical models and historical data, they estimate the likelihood of future medical expenses and set premiums accordingly. This analytical approach ensures that premiums are fair, data-driven, and reflective of risk, rather than arbitrary.By understanding these risk factors, insurers can price policies accurately, manage financial risk, and maintain sustainability.

1.2 Research Question

"How can Bayesian regression be used to predict health insurance charges using demographic and lifestyle factors, and how does it improve uncertainty quantification compared to traditional methods?" This study employs Bayesian regression to model health insurance charges, incorporating variables such as age, gender, BMI, and smoking status. By combining prior knowledge with observed data, the Bayesian framework not only estimates the relationships between these factors and insurance costs but also quantifies the uncertainty in these predictions through probabilistic distributions.

1.3 Hypotheses

To investigate how demographic and lifestyle factors influence health insurance charges, this study formulates the following testable hypotheses:

- H1 (Age effect): Older individuals tend to have higher health insurance charges compared to younger individuals.

- H2 (Gender effect): There is a significant difference in health insurance charges between males and females.

- H3 (Smoking effect):Individuals who smoke incur higher health insurance charges than non-smokers.

- H4 (Uncertainty quantification): Bayesian regression provides narrower credible intervals for predictions compared to traditional regression, offering more reliable estimates of health insurance charges.

1.4 Importance and Objectives

This study aims to predict health insurance costs using demographic variables such as age, gender, and BMI, along with smoking habits. Unlike traditional regression approaches that provide point estimates, Bayesian regression offers a probabilistic framework, allowing the incorporation of prior knowledge and the estimation of uncertainty in predictions. By applying Bayesian regression, this research seeks not only to model the relationships between these factors and insurance charges but also to generate predictive distributions that reflect the range of possible outcomes. The findings of this study can support insurers in risk assessment, help policyholders understand factors affecting their premiums, and provide policymakers with insights into the population-level determinants of health expenditure.

2 Methodology

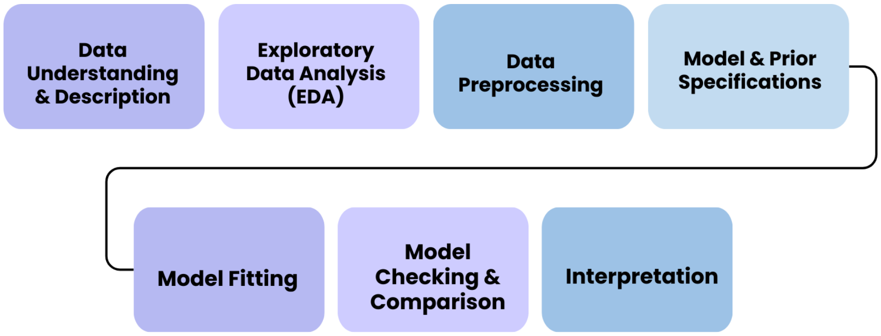

Data Understanding & Description
Exploratory Data Analysis (EDA)
Data Preprocessing
Model & Prior Specifications
Model Fitting
Model Checking & Comparison
Interpretation

Figure 1: Project workflow

2.1 Description of Dataset

This dataset contains details about the healthcare insurance costs for individuals. It consists of several factors that may impact these charges, including age, gender, body mass index (BMI), number of children, smoking habits, region, and the related insurance premiums.There are 1407 observations and 7 variables in the data set. 3 of them are categorical variables and 4 of them are numerical variables. The dataset description is as follows:

Table 1: Description of Variables in the Dataset

Variable	Type	Description
Age	Integer	The age of the individual in years
Gender	Factor	The gender of the individual (male/female)
Body_Mass_IndexḂMI	Continuous	A measure of body fat calculated from weight and height. in kg/m^2
Num_of_children	Integer	The number of children covered by the insurance. (0-5)
Smoking_Status	Factor	Whether the individual is a smoker (yes/no)
Region	Factor	The geographical region of the individual (e.g., southwest, southeast, northwest, northeast)
Insurance_charges	Continuous	Annual medical insurance charges for each individual in USD

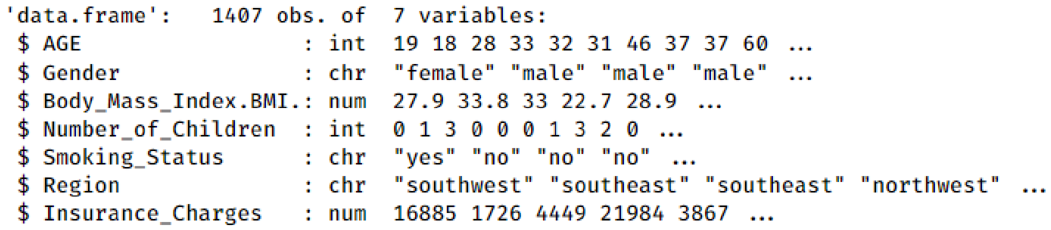
This dataset is suitable for exploring the relationships between the above factors and insurance charges, as well as for building predictive models to estimate insurance charges based on various attributes. It could be used to analyse how factors like age, gender, BMI, smoking status, and region impact insurance costs.

2.2 Exploratory Data Analysis (EDA)

Before fitting a statistical model, it is essential to understand the dataset's structure, distribution, and relationships. Exploratory Data Analysis (EDA) helps to understand the data visually and numerically. It enables us to identify patterns, recognise outliers, and understand how various factors relate to our primary variable.

In this study, an EDA is performed to summarise the demographics and lifestyles of the individuals, check how insurance charges are distributed and explore the connections between different predictors. The findings from this EDA will be used in preparing the data, choosing the right model, and understanding the results from the Bayesian regression analysis.

2.2.1 Structure of the dataset



```
'data.frame':	1407 obs. of  7 variables:
 $ AGE                 : int  19 18 28 33 32 31 46 37 37 60 ...
 $ Gender              : chr  "female" "male" "male" "male" ...
 $ Body_Mass_Index.BMI.: num  27.9 33.8 33 22.7 28.9 ...
 $ Number_of_Children  : int  0 1 3 0 0 0 1 3 2 0 ...
 $ Smoking_Status      : chr  "yes" "no" "no" "no" ...
 $ Region              : chr  "southwest" "southeast" "southeast" "northwest" ...
 $ Insurance_Charges   : num  16885 1726 4449 21984 3867 ...
```

Figure 2: Dataset Structure

There are 1407 observations and 7 variables in the data set. 3 of them are categorical variables and 4 of them are numerical variables.After getting the summary of this dataset, the following information was obtained,

- Ages range from 18 to 64 years, with a median of 39 and mean of 39.26, indicating a roughly symmetric distribution centered around middle adulthood.
- BMI ranges from 15.96 to 53.13, with a mean of 30.65 and median of 30.3, suggesting that on average, the population is overweight (BMI ¿ 25).
- Most individuals have 0–2 children.
- Charges vary widely indicating a right-skewed distribution.

Frequency tables for categorical variables were obtained and the summary was as follows:

Table 2: Categorical Variables Summary

Table 3: Gender

Female	695
Male	712

Table 4: Smoking Status

No	1123
Yes	284

Table 5: Region

Northeast	336
Northwest	339
Southeast	386
Southwest	346

Next, the distribution of numeric variables was explored, and the plots obtained are included below.

2.2.2 Distributions of numeric variables

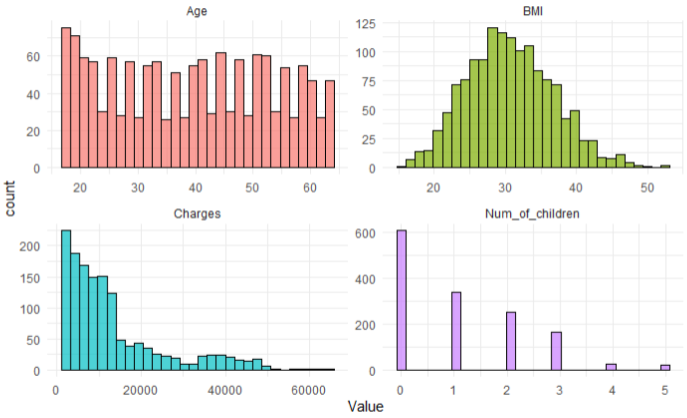


Figure 3: Distributions of numeric variables

The distribution plot for Age illustrates that the Age appears to be fairly uniformly distributed. This suggests a balanced age representation, which is beneficial for generalizability in further analysis.

The plot for BMI (Body Mass Index) displays an approximate bell-shaped distribution centered around the 30 mark. Most individuals in the dataset have a BMI between 25 and 35, indicating that the population is mainly within the overweight category according to standard BMI classifications.

The Charges plot is right-skewed with a long tail towards higher values. Most individuals incur lower charges, as seen by the high frequency at the lower end, while only a few experience very high charges. This pattern is common in healthcare expense data due to the presence of rare, expensive treatments.

The distribution plot for Num_of_children shows that a majority of individuals have zero children, with the frequency decreasing as the number of children increases. Very few individuals have more than three children.

2.2.3 Scatterplots of numeric predictors vs Charges

The following plots show the relationships between numeric predictors and Insurance charges

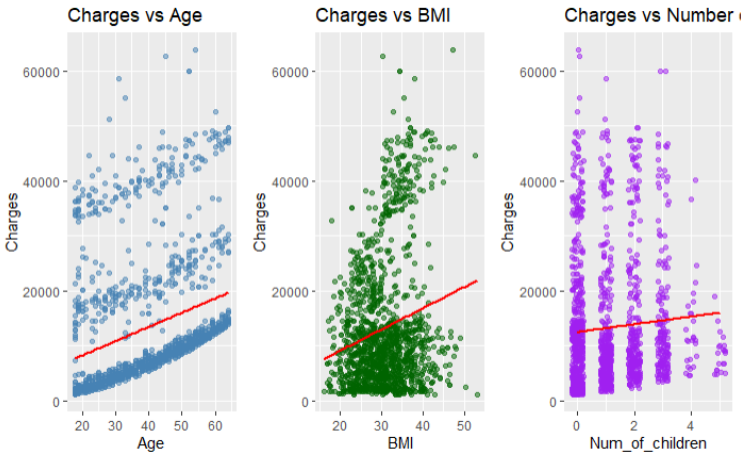


Figure 4: Scatterplots of numeric predictors vs Charges

Charges vs. Age plot shows a positive upward trend: as age increases, charges tend to rise.However, the spread widens with age suggesting that some elderly individuals have moderate costs, while smokers/obese have extreme charges.

In Charges vs BMI plot, higher BMI ranges (¿30, obesity),the clusters of very high charges (can be smokers) can be observed.BMI alone is not strong, but combined with smoking status, it amplifies charges.

Charges vs Number of Children plot shows a Weak linear trend.It seems that the Charges don't significantly rise with more children.

2.2.4 Boxplots for categorical predictors

The following box plots show the relationships between categorical predictors and Insurance charges

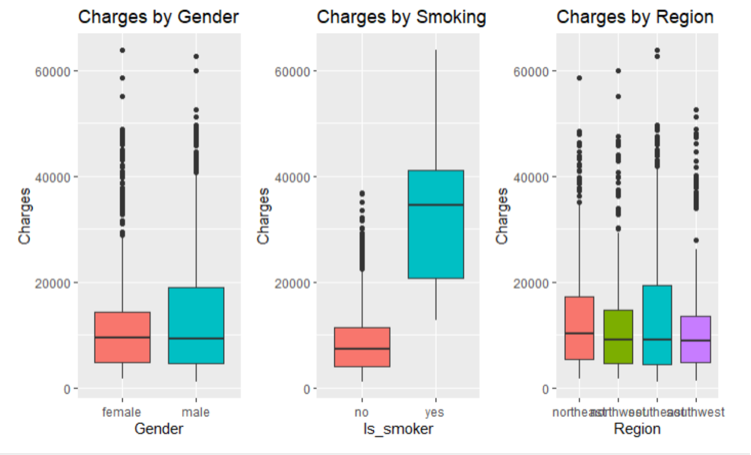


Figure 5: Boxplots for categorical predictors

The boxplot of charges by smoking shows that smokers generally have significantly higher insurance charges than non-smokers, with greater variability and more extreme high values, suggesting the impact of smoking as a significant risk factor in insurance cost calculations.

Charges by gender shows no significant difference in median charges implying that the gender doesn't strongly influence medical costs in this dataset.

Charges by region boxplots across Northeast, Northwest, Southeast, Southwest also look fairly similar, suggesting that the region is not a strong driver of charges.

2.2.5 Correlations

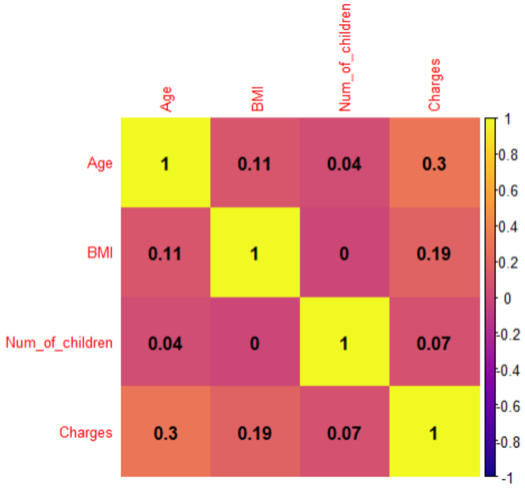


Figure 6: Correlation heatmap

The charges variable shows a moderate positive correlation with age variable. As age increases, charges tend to increase moderately. Older individuals usually have higher medical expenses.There is a weak positive correlation between charges and BMI. This makes sense since overweight/obesity may lead to higher medical costs, but the effect is not strong.

According to the correlation plot, here is no multicollinearity problem among predictors (all correlations ¡ 0.7).

2.3 Bayesian Methods and Models

In contrast to classical regression which provides single-point estimates and confidence intervals, Bayesian regression yields full posterior distributions for model parameters, allowing a richer understanding of uncertainty. The response variable, health insurance charges, displayed a highly right-skewed distribution showing a small proportion experiencing extremely high charges. Since such skewness violates the assumption of normally distributed residuals in regression modeling, two Bayesian regression models were constructed:

- Model 1: Raw Charges as Response
  Here, the untransformed charges were used directly. This allows the model to capture the true scale of the data, but requires priors and likelihoods that are robust to heavy-tailed distributions.

- Model 2: Log-Transformed Charges as Response
  In this approach, a natural logarithm transformation was applied to the charges. The transformation reduces right skewness, producing a distribution that more closely resembles normality. This helps stabilizing variance and improving model fit.

2.4 Model Specification

The general Multiple Linear Regression models are:

- Model 1: Raw Charges as Response

$$\text{Log}(\text{Charges}_i) \sim \mathcal{N}(\mu_i, \sigma^2)$$

$$\begin{aligned} Charges_i = \beta_0 + \beta_1\,\text{Age\_s}_i + \beta_2\,\text{BMI\_s}_i + \beta_3\,\text{Num\_of\_Children}_i \\ + \beta_4\,\text{Gender}_i + \beta_5\,\text{Is\_smoker}_i + \beta_6\,\text{Region}_i \end{aligned} \tag{1}$$

- Model 2: Log-Transformed Charges as Response

$$\text{Log}(\text{Charges}_i) \sim \mathcal{N}(\mu_i, \sigma^2)$$

$$Log(Charges_i) = \beta_0 + \beta_1 \, \text{Age\_s}_i + \beta_2 \, \text{BMI\_s}_i + \beta_3 \, \text{Num\_of\_Children}_i \\ + \beta_4 \, \text{Gender}_i + \beta_5 \, \text{Is\_smoker}_i + \beta_6 \, \text{Region}_i \quad (2)$$

2.5 Prior Specification

- Model 1: For a robustness check, a model with the raw Charges without transformation also fitted using a Gaussian likelihood. Since the response variable spans a wide range ($\tilde{1}$,100 to 63,700 USD) and exhibits significant variability, a broader, weakly informative priors were applied to account for the large scale of the data. The intercept was assigned a prior of $\mathcal{N}(0, 5000)$ and the slopes for predictors were given a prior of $\mathcal{N}(0, 1000)$ reflecting the possibility of larger coefficient magnitudes due to the high variance in the response. The residual standard deviation was assigned a Student-t prior with 3 degrees of freedom and scale 10,000, allowing for heavy tails and outliers while remaining weakly informative. These priors provide flexibility for the model to capture the wide range and extreme values of raw Charges, while still incorporating reasonable regularization to stabilize estimation.

- Model 2: For the main analysis, we modeled the log-transformed Charges using a Gaussian likelihood, which is appropriate because the logarithmic transformation reduces the right-skewness of the response and makes it approximately normally distributed. The model includes predictors such as standardized Age and BMI, the number of children, Gender, Smoker status, and Region. We applied weakly informative priors for the coefficients, assigning $\mathcal{N}(0, 1)$ to slopes and $\mathcal{N}(0, 5)$ to the intercept, reflecting the assumption that most effects are small but allowing the data to guide the estimates. The residual standard deviation was given a Studentt prior with 3 degrees of freedom, which is flexible enough to accommodate moderate deviations from normality.

2.6 Model Checking

After fitting the Bayesian regression models, several model checking steps were conducted to ensure reliability and good fit. First, the convergence diagnostics were examined using traceplots, Rhat values, and effective sample sizes (ESS) to confirm that the Markov chains mixed well and that the posterior estimates are stable. Next, the overall model fits were evaluated with Bayesian R squared values which measures the proportion of variance in the response explained by the predictors. The summary of model checking is as follows,

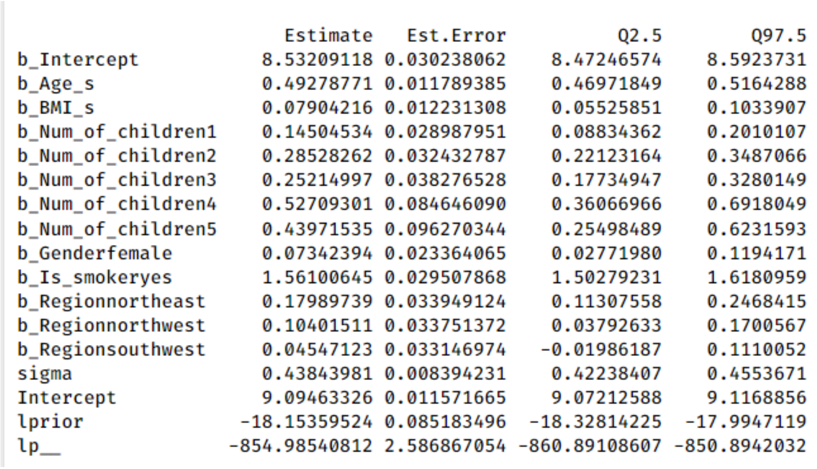
Table 6: Bayesian MLR Model Checking Summary

Metric / Check	**Log-Charges Model**	**Raw-Charges Model**	**Notes**
Rhat	1.00	1.00	Values close to 1 indicate convergence
Bulk ESS	18137	18556	Large ESS imply reliable posterior estimates
Tail ESS	8674	9376	Checks stability of tail estimates
Bayesian R^2	0.773333	0.6868734	Higher value implies a better fit
LOOIC	1687.6	28583.2	Lower implies a better predictive performance

3 Results and Discussion

3.1 Posterior Summaries

For the Log(Charges) model:

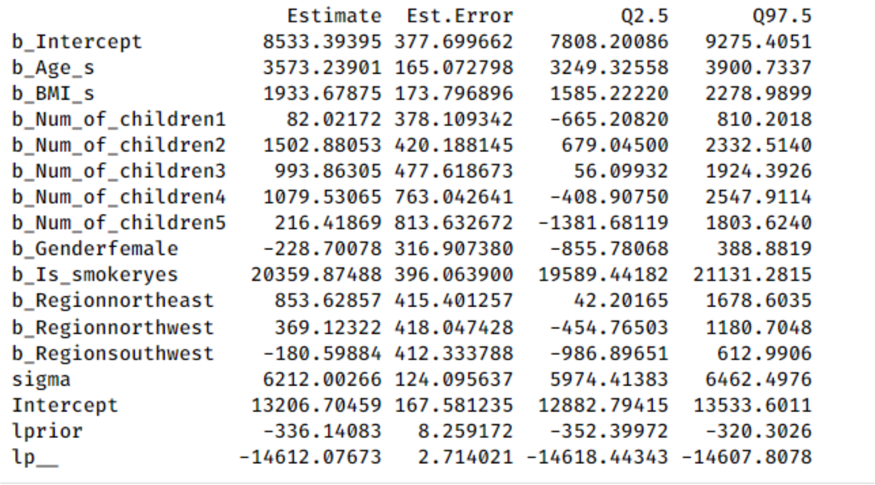


```
                       Estimate   Est.Error         Q2.5        Q97.5
b_Intercept          8.53209118 0.030238062   8.47246574    8.5923731
b_Age_s              0.49278771 0.011789385   0.46971849    0.5164288
b_BMI_s              0.07904216 0.012231308   0.05525851    0.1033907
b_Num_of_children1   0.14504534 0.028987951   0.08834362    0.2010107
b_Num_of_children2   0.28528262 0.032432787   0.22123164    0.3487066
b_Num_of_children3   0.25214997 0.038276528   0.17734947    0.3280149
b_Num_of_children4   0.52709301 0.084646090   0.36066966    0.6918049
b_Num_of_children5   0.43971535 0.096270344   0.25498489    0.6231593
b_Genderfemale       0.07342394 0.023364065   0.02771980    0.1194171
b_Is_smokeryes       1.56100645 0.029507868   1.50279231    1.6180959
b_Regionnortheast    0.17989739 0.033949124   0.11307558    0.2468415
b_Regionnorthwest    0.10401511 0.033751372   0.03792633    0.1700567
b_Regionsouthwest    0.04547123 0.033146974  -0.01986187    0.1110052
sigma                0.43843981 0.008394231   0.42238407    0.4553671
Intercept            9.09463326 0.011571665   9.07212588    9.1168856
lprior             -18.15359524 0.085183496 -18.32814225  -17.9947119
lp__              -854.98540812 2.586867054 -860.89108607 -850.8942032
```

Figure 7: Posterior summary of Log model

According to the posterior summary of log-transformed model, Smoking status and age are the strongest predictors. BMI, children, and region also matter, but gender has only a small effect. The model is well-calibrated, with narrow credible intervals indicating high certainty.

For the Raw Charges model:



```
                      Estimate   Est.Error        Q2.5       Q97.5
b_Intercept         8533.39395 377.699662  7808.20086   9275.4051
b_Age_s             3573.23901 165.072798  3249.32558   3900.7337
b_BMI_s             1933.67875 173.796896  1585.22220   2278.9899
b_Num_of_children1    82.02172 378.109342  -665.20820    810.2018
b_Num_of_children2  1502.88053 420.188145   679.04500   2332.5140
b_Num_of_children3   993.86305 477.618673    56.09932   1924.3926
b_Num_of_children4  1079.53065 763.042641  -408.90750   2547.9114
b_Num_of_children5   216.41869 813.632672 -1381.68119   1803.6240
b_Genderfemale      -228.70078 316.907380  -855.78068    388.8819
b_Is_smokeryes     20359.87488 396.063900 19589.44182  21131.2815
b_Regionnortheast    853.62857 415.401257    42.20165   1678.6035
b_Regionnorthwest    369.12322 418.047428  -454.76503   1180.7048
b_Regionsouthwest   -180.59884 412.333788  -986.89651    612.9906
sigma               6212.00266 124.095637  5974.41383   6462.4976
Intercept          13206.70459 167.581235 12882.79415  13533.6011
lprior              -336.14083   8.259172  -352.39972   -320.3026
lp__              -14612.07673   2.714021 -14618.44343 -14607.8078
```

Figure 8: Posterior summary of raw charges model

The posterior estimates indicate that smoking, age, and BMI are the most influential factors driving higher insurance charges, with smoking showing the strongest effect. Having two or three children also increases costs to some extent, while gender and most regional differences exhibit little to no meaningful impact.

3.2 Model Comparison

Table 7: Comparison of Bayesian Regression Models

Criteria	**Model 1 (Log Charges)**	**Model 2 (Raw Charges)**
Likelihood	Gaussian	Gaussian
Prior specification	Normal(0, 1) slopes; Normal(0, 5) intercept	Normal(0, 1000) slopes; Normal(0, 5000) intercept
LOOIC	1687.6	28583.2
Bayesian R^2	0.773333	0.6868734
$\hat{R}$ (convergence)	1.00	1.00
95% Credible Intervals	Include 0 for some predictors	Include 0 for most predictors
Overall fit	Better under Gaussian and weakly informative priors. Narrow CI	Less precise. wide CI

3.3 Interpretation of findings and practical implications

The Bayesian regression analysis shows that key factors such as age, BMI, and smoking status strongly influence health insurance charges, while variables like gender and number of children have weaker or uncertain effects. Between the two models, the Gaussian likelihood provided more stable estimates and credible intervals, making it more suitable for this dataset. The posterior summaries and Bayesian R squared values indicate that the model explains a substantial amount of variation in charges, and the convergence diagnostics (R hat) confirm that the results are reliable.

In practical terms, the findings highlight that smoking and higher BMI significantly increase insurance costs, which can guide insurers in setting fair premiums and encourage healthier lifestyles among policyholders. The weaker role of gender and children suggests that demographic factors may matter less compared to lifestyle choices. Overall, this Bayesian approach not only quantifies uncertainty around estimates but also provides more robust insights than traditional methods, making it useful for decision-making in healthcare policy and insurance pricing.

4 Conclusion and Recommendations

4.1 Key Findings

This study applied Bayesian regression to predict health insurance charges based on demographic and lifestyle factors. The results show that smoking status and BMI are the strongest predictors of higher charges, followed by age, while gender and number of children have little or uncertain impact. Between the two models tested, the Gaussian likelihood model produced more stable results and better explained variation in the data.

4.2 Comparison with literature

These findings are consistent with existing research, which repeatedly highlights the role of lifestyle-related risks (such as smoking and obesity) in driving healthcare costs Moriarty et al., 2012, Kim, 2025. Similar to prior studies, demographic variables like gender and family size contribute less significantly compared to behavioral and health-related factors. The use of Bayesian methods also aligns with recent literature emphasising improved uncertainty quantification and more flexible inference compared to traditional regression.

4.3 Limitations and suggestions

A key limitation is that the dataset may not fully capture other important drivers of healthcare costs, such as chronic illnesses or physical activity. Additionally, the analysis assumes that relationships are linear, which may oversimplify the complex effects of certain predictors. Although Bayesian methods provide credible intervals and robust estimates, the results still depend on model assumptions and chosen priors. Future studies should expand the set of predictors to include medical history, lifestyle habits etc. Moreover, applying alternative Bayesian models such as hierarchical or mixture models could provide deeper insights into subgroup differences.

References

Kim, Y. (2025). The effects of smoking, alcohol consumption, obesity, and physical inactivity on healthcare costs: A longitudinal cohort study. *BMC Public Health*, *25*, 873. https://doi.org/10.1186/s12889-025-22133-4

Kongyir, B., & Agbemade, E. (2024). Modeling health insurance premium using bayesian hierarchical models. *Data Science and Data Mining*, (25). https://stars.library.ucf.edu/data-science-mining/25

Moriarty, J. P., Branda, M. E., Olsen, K. D., Shah, N. D., Borah, B. J., Wagie, A. E., Egginton, J. S., & Naessens, J. M. (2012). The effects of incremental costs of smoking and obesity on health care costs among adults: A 7-year longitudinal study. *Journal of Occupational and Environmental Medicine*, *54*(3), 286–291. https://doi.org/10.1097/JOM.0b013e318246f1f4

Muth, C., Oravecz, Z., & Gabry, J. (2018). User-friendly bayesian regression modeling: A tutorial with `rstanarm` and `shinystan`. *The Quantitative Methods for Psychology*, *14*(2), 99–119. https://doi.org/10.20982/tqmp.14.2.p099

5 Appendix

The full R Markdown file used for the analyses could not be included due to some compilation issues. The analysis steps performed in this project are summarised below:

- Data preprocessing: Cleaning and transforming the insurance dataset.
- Exploratory Data Analysis: Visualising numeric and categorical variables, including histograms and bar plots.
- Modelling: Bayesian regression modeling.
- Diagnostics and Results: Posterior summaries, plots, and model comparisons.

The R markdown file is attached separately for further reference.