Modeling Health Insurance Expenses in the United States with Hierarchical Bayesian Models

Matt Ludwig

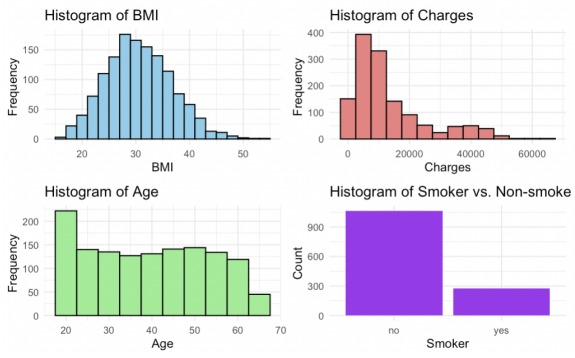
1 Introduction

1.1 Overview

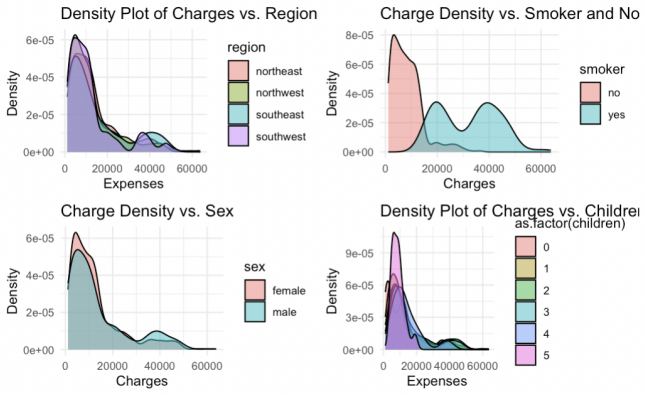
Health insurance or medical insurance is defined as a type of insurance which covers the whole or part of the risk of a person incurring medical expenses. Properly estimating the health insurance expenses is an important task of insurance companies so that they can make accurate predictions about future costs and earnings. In this project, I will construct several Bayesian models to model the cost of health insurance in the United States. In particular, I will first construct a Gamma hierarchical Bayesian model to model the cost of health insurance in each region in the United States based solely on the expense data in each region. The purpose of this model is to allow the information about the cost of health insurance in each of the four geographic regions to inform estimates of the cost of health insurance in the other regions without pooling the data from all of the regions. In effect, this model will optimize the way in which the information between the four geographic regions is shared while still modeling the distributions specific to each region. Following this, I will construct and compare a Bayesian multiple linear regression model and a Bayesian Gamma regression model to model the expenses in each region based on the seven features of the data. This project and its models were inspired by a paper in the Casualty Actuarial Society E-Forum called "Ratemaking for a New Territory: Enhancing GLM Pricing Model with a Bayesian Analysis" by Jing Zhang and Tatjana Miljkovic.[2]

1.2 Description of the Dataset

The Insurance Premium Prediction dataset contains 1338 observations of the seven features.[1] Four of the features are numeric: age, BMI (body mass index), children and exspenses while three of features are nominal: sex, smoker and region. The first plots below are histograms of the frequency of insurees vs. BMI, total expenses or charges (these terms are used interchangeably in this project), age and whether they are smokers or not. The other plots are density plots of the number of insurees vs. their health insurance expenses. Here, several of the covariates are separated into different levels. These plots illustrate how the expenses of the insurees is variable. For the purposes of this project, it is interesting to observe how in the density plot of charges vs. region, the densities in the four regions all appear similar but are slightly different. In the following sections of this report, I will discuss how the Gamma hierarchical models that I implemented captures this similarity and difference.



(a)



(b)

Figure 1

2 Gamma Hierarchical Model for Modeling Health Insurance Exspenses

2.0.1 Gamma Hierarchical Model

Since the severity of insurance claims often follows a Gamma distribution, I first constructed a Gamma hierarchical model to model the distribution of health insurance expenses. This model assumes that the expenses in each region follow a Gamma distribution with parameters α_i and β, that is $\Gamma(\alpha_i, \beta)$ where $\beta \sim \Gamma(1,1)$ is a common parameter. That is, the sampling model for the data in each region is $\Gamma(\alpha_i, beta)$. Further, it assumes that the α_i parameters follow a $\Gamma(\alpha', \beta')$ distribution where the hyperparameters follow $\alpha' \sim \Gamma(1,1)$ and $\beta' \sim \Gamma(1,1)$. In effect, this model will use these prior distributions to obtain posterior estimates for the α_i and common β in each region as well as the α' and β' hyperparameters which control the heterogeneity of the parameters for each of the regions. Therefore, these posterior estimates will essentially provide a description of the Gamma distribution in each of the four regions.

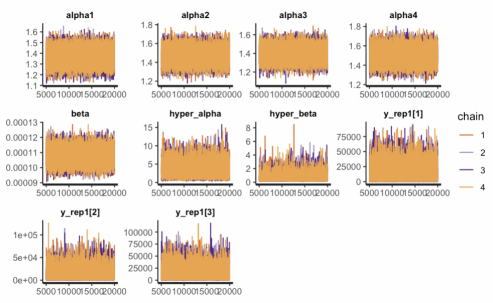
2.1 Results for Gamma Hierarchical Model

After compiling this model in Stan, it produced the following results:

```
Inference for Stan model: anon_model.
4 chains, each with iter=20000; warmup=5000; thin=1;
post-warmup draws per chain=15000, total post-warmup draws=60000.
```

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff
alpha1	1.36	0.00	0.07	1.24	1.32	1.36	1.41	1.50	35544
alpha2	1.47	0.00	0.07	1.33	1.42	1.47	1.51	1.61	33339
alpha3	1.39	0.00	0.07	1.26	1.35	1.39	1.44	1.53	35896
alpha4	1.50	0.00	0.07	1.36	1.45	1.50	1.54	1.64	34656
beta	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	26849
hyper_alpha	3.25	0.01	1.62	1.00	2.08	2.95	4.10	7.19	27823
hyper_beta	0.63	0.00	0.44	0.19	0.35	0.51	0.77	1.79	23911

(a)



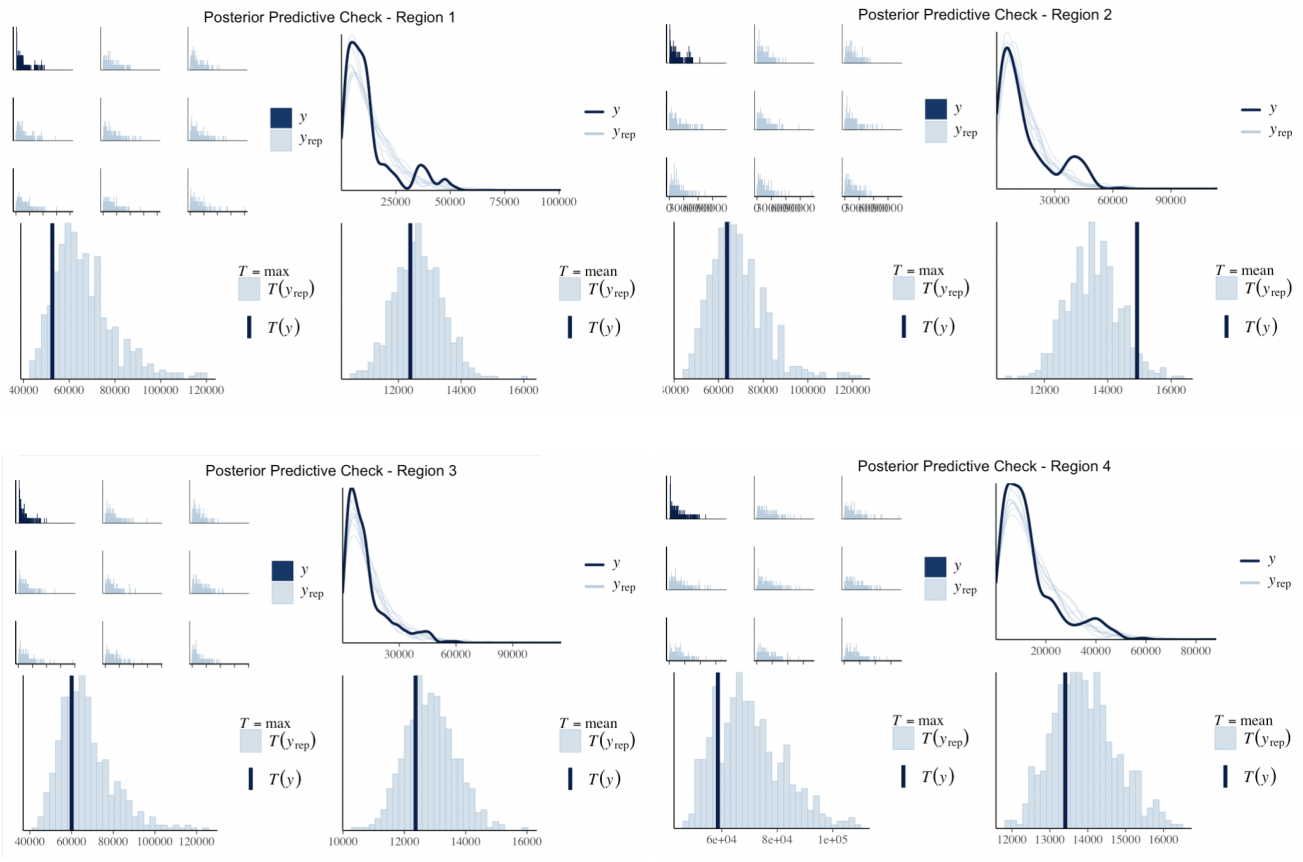
(b)

Figure 2: Posterior Estimates and Trace Plots for the Parameters and Hyperparameters in Each of the Four Geographic Regions

The posterior means, standard errors, confidence intervals and effective sample size of the parameters shown in Figure 2 a.) suggest that shape parameters for the Gamma distributions in the four regions are $1.36, 1.47, 1.39$ and 1.50 which appear to be reasonable estimates. The trace plots for these parameters in Figure 2 b.) also suggest that the Markov chain is fully exploring the state space and that the estimates are also reasonable. The posterior estimate of the common scale parameter beta being zero however does not appear reasonable. Extracting the posterior samples and manually calculating the mean revealed that the mean of the posterior samples of beta is: 0.0001153254. which very small but not zero and is perhaps an adequate value for this data. To gauge the performance of the model, I also calculated the mean squared error and mean absolute error between the observed data and posterior samples in each region and found: MAE values of 11385.72, 13306.46, 11252.18 and 11721.82, and MSE values of 251244361, 321617456, 242353800 and 255951521 respectively.

2.2 Posterior Predictive Checks for Gamma Hierarchical Model

In order to further check the validity of this Gamma hierarchical model, I performed posterior predictive checks and plotted histograms of samples from the posterior predictive distribution as well as their dense overlays with the Bayesplot library in R. In addition, for each of the replicated posterior datasets and the original data, I also computed the mean and max values and plotted the results in histograms. These plots are shown below:



(a) (b)

Figure 3: Plots for Posterior Predictive Check in Each of the Four Regions

In each of the four regions, the values of the histograms of the original data set vs. samples from the posterior predictive distribution appear relatively similar and the dense overlays of these plots also appear mostly similar. Furthermore, in the histogram of the max and mean statistics for these data sets the value for the original dataset also appears close to the center of the distribution suggesting that the posterior predictive datasets capture the mean and max of the original data set well.

3 Bayesian Linear Regression Model

In addition, I also implemented a Bayesian linear regression model. Unlike the hierarchical Gamma model, I did not separate the data by region and instead treated the region as a factor and used it as a covariate in the linear regression model. For this model, I placed normal prior distributions, $\mathcal{N}(0, 10)$, over the coefficients of each of the regression coefficients and used a Cauchy distribution, $Cauchy(0, 5)$, for the distribution of errors term. The sampling model for this model was as follows:

$$\begin{aligned} y_i &= \beta_0 + \beta_{\text{age}} \cdot \text{Age}_i + \beta_{\text{sex}} \cdot \text{Sex}_i + \beta_{\text{bmi}} \cdot \text{BMI}_i \\ &\quad + \beta_{\text{children}} \cdot \text{Children}_i + \beta_{\text{smoker}} \cdot \text{Smoker}_i + \beta_{\text{region}} \cdot \text{Region}_i + \epsilon_i \\ \epsilon_i &\sim \text{Normal}(0, \sigma) \end{aligned}$$

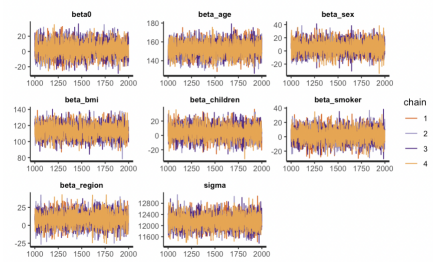
3.1 Results for Bayesian Linear Regression Model

The posterior means of parameters shown in figure below suggest that for every unit increase in age, health care costs increase by \$152.34 dollars and for every unit increase in BMI health care costs increase by \$113.37 dollars. These posterior estimates also suggest that health care costs increases for women, the number of children one has, whether one smokes and regions in which one lives in. However, the models suggests that the effect on the cost

of health care of these other covariates are not as influential as age and BMI. The trace plots for the coefficients estimates are also shown below and suggest that the Markov chain is fully exploring the state space. To gauge the performance of the model, I also calculated the mean squared error and mean absolute error between the observed data and posterior samples in each region and found: MAE value of 13613.5 and an MSE value of 313977652.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%
beta0	3.42	0.12	9.98	-16.22	-3.31	3.33	10.29	22.98
beta_age	152.34	0.10	7.42	137.72	147.36	152.29	157.40	166.91
beta_sex	5.36	0.12	10.03	-14.07	-1.63	5.48	12.23	24.92
beta_bmi	113.37	0.11	8.50	97.20	107.49	113.39	119.36	130.34
beta_children	4.65	0.11	9.71	-14.30	-1.91	4.43	11.18	23.88
beta_smoker	2.49	0.12	9.86	-16.98	-3.86	2.51	8.96	21.41
beta_region	8.58	0.11	9.90	-10.65	2.08	8.68	15.26	27.86
sigma	12146.43	3.24	252.69	11664.76	11975.69	12137.34	12316.52	12662.28

(a)



(b)

Figure 4: Posterior Estimate and Trace Plots for the Parameters in the Bayesian Linear Regression Model

3.2 Posterior Predictive Checks for Bayesian Linear Regression Model

In order to further check the validity of this Bayesian linear regression model, I performed posterior predictive checks and plotted histograms of samples from the posterior predictive distribution as well as their dense overlays with the Bayesplot library. In addition, for each of the replicated posterior datasets and the original data, I also computed the mean and max values and plotted the results in histograms. These plots are shown below:

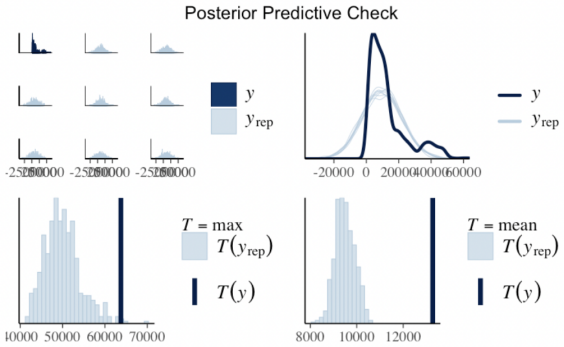


Figure 5: Plots for Posterior Predictive Check for Bayesian Linear Regression Model

These posterior predictive plots illustrate how the model does not fit the data well. The top plots of the distribution of the data illustrate how distribution of the original dataset is more peaked than the distributions of the samples from the posterior predictive datasets. Furthermore, the bottom plots show how the mean and max statistics or the original data is different from those calculated on the posterior predictive datasets. These plots suggest that using a Bayesian linear regression model that assumes that the data is normally distributed is inadequate. Therefore, I then decided to implement a Bayesian Gamma regression model for the data.

4 Bayesian Gamma Regression Model

The Bayesian Gamma regression model is similar to the Bayesian normal linear regression with the exception that it assumes the response (expenses in this project) follows a Gamma distribution. It also utilizes a log link function to relate the linear predictors to the response variable. In this model, I used normal priors, $\mathcal{N}(0, 10)$, for the intercept and regression coefficients and Gamma priors, $\Gamma(1, .1)$, for the shape and rate parameters. The likelihood or sampling model is as follows:

$$\mu = \exp(\beta_0 + \beta_{\text{age}} \cdot \text{Age} + \beta_{\text{sex}} \cdot \text{Sex} + \beta_{\text{bmi}} \cdot \text{BMI} + \beta_{\text{children}} \cdot \text{Children} + \beta_{\text{smoker}} \cdot \text{Smoker} + \beta_{\text{region}} \cdot \text{Region})$$

$$\text{for } i = 1 \text{ to } N: \quad \text{Charges}[i] \sim \text{gamma}(\text{shape}, 1/\mu[i]) \quad \text{where Scale} = 1/\mu$$

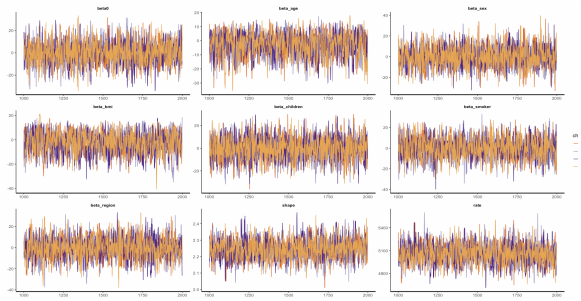
4.1 Results for Gamma Linear Regression Model

The posterior means of parameters and trace plots for the coefficients estimates are also shown below.

4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff
beta0	-0.17	0.21	10.19	-19.71	-7.11	-0.25	6.77	20.13	2385
beta_age	-2.89	0.25	8.31	-20.81	-8.19	-2.17	3.11	11.08	1069
beta_sex	0.17	0.20	9.93	-19.29	-6.58	0.17	6.83	19.24	2363
beta_bmi	-2.46	0.24	8.72	-20.70	-8.34	-1.88	3.72	12.53	1346
beta_children	-0.45	0.21	9.95	-20.30	-7.04	-0.43	6.14	18.83	2211
beta_smoker	0.11	0.20	9.98	-20.32	-6.59	0.02	6.98	19.45	2414
beta_region	-0.31	0.21	10.06	-20.98	-6.89	-0.17	6.40	18.88	2293
shape	2.25	0.00	0.06	2.13	2.20	2.25	2.29	2.36	1555
rate	5059.45	3.32	129.88	4811.39	4971.77	5059.58	5145.02	5316.41	1530

(a)



(b)

Figure 6: Posterior Estimate and Trace Plots for the Parameters in the Bayesian Linear Regression Model

4.2 Posterior Predictive Checks for the Gamma Regression Model

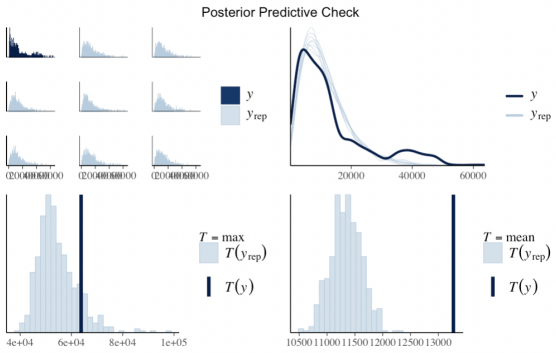


Figure 7: Plots for Posterior Predictive Check for Bayesian Linear Regression Model

These posterior predictive plots illustrate how the model fits the data better than it fits the normal linear regression model. The top plots comparing the distribution of the original dataset and the posterior predictive datasets illustrate how the model is capturing the shape of the data better than the normal model does. The bottom two plots of the max and mean statistics for these data sets the value for the original dataset illustrate how the model does not do a great job of modeling this data as there is still some discrepancy between the mean and max statistics of the original data and those of the posterior predictive datasets.

5 Conclusion

In this project, I implemented three Bayesian models to model the cost of health insurance in the United States. I first implemented a Bayesian hierarchical Gamma model to model health insurance expenses in the four geographic regions in the United States. The hierarchical model gave the following mean posterior values for the shape parameter of the four regions: 1.36, 1.47, 1.39 and 1.5. It also gave a common beta parameter of 0.0001153254. Furthermore, posterior predictive checks suggested that that the model fit the data well. Following this, I implemented a Bayesian linear regression models to predict the cost of health insurance in the United States based on seven covariates. While the coefficients in this model appeared reasonable, posterior predictive checks indicated that the model did not fit the data entirely well as the original data does not appear to be normally distributed. Therefore,

I implemented a Bayesian Gamma regression model which appeared to fit the original data better. Overall, the biggest takeaways from this project were not to underestimate the difficulty of fitting generalized linear models in Stan and the importance of picking appropriate prior distributions and initial values when performing MCMC sampling so as not to encounter overflow errors. Further investigations could refine the current models as well as implement others. In particular, it would be interesting to implement a hierarchical Gamma regression model on this data which would most likely perform better than the current implementation of the Gamma regression model. In addition to this, future work could construct other hierarchical models which capture different groupings of the data. In particular, given the apparent importance of BMI and Age in the predicting the cost of health insurance, it would be interesting to construct a hierarchical model that could capture age and BMI groupings possibly model the data more accurately.

6 Appendix : R Code

```
---
title: "Hierarchical Linear Regression Project"
output: html_document
---

# Bayesian Final Project: Using Bayesian Models to Model Health Insurance Costs in America

Goal:

Develop a Bayesian hierarchical linear regression model to predict insurance costs

Covariates:

Age, Sex, BMI, Children, Smoker, Region \~ Charges

Groups:

Region \~ Northeast, Northwest, Southeast, Southwest

# Load and Explore Data

‘‘‘{r}

#install.packages("gridExtra")
#install.packages("ggplot2")

library(gridExtra)
library(ggplot2)

insurance_data <- read.csv("/Users/matthewludwig/Desktop/insurance.csv")
insurance_data$sex<- factor(insurance_data$sex)
insurance_data$smoker<- factor(insurance_data$smoker)
insurance_data$region<- factor(insurance_data$region)



density_plot_children <- ggplot(insurance_data, aes(x = expenses, fill = as.factor(children))) +
  geom_density(alpha = 0.5) +
  labs(title = "Density Plot of Charges vs. Children",
```

```
       x = "Expenses",
       y = "Density") +
  theme_minimal()


density_plot_region <- ggplot(insurance_data, aes(x = expenses, fill = region)) +
  geom_density(alpha = 0.5) +
  labs(title = "Density Plot of Charges vs. Region",
       x = "Expenses",
       y = "Density") +
  theme_minimal()

density_plot_smoker <- ggplot(insurance_data, aes(x = expenses, fill = smoker)) +
  geom_density(alpha = 0.5) +
  labs(title = "Charge Density vs. Smoker and Non-smoker",
       x = "Charges",
       y = "Density") +
  theme_minimal()


density_plot_sex <- ggplot(insurance_data, aes(x = expenses, fill = sex)) +
  geom_density(alpha = 0.5) +
  labs(title = "Charge Density vs. Sex",
       x = "Charges",
       y = "Density") + theme_minimal()

histogram_bmi <- ggplot(insurance_data, aes(x = bmi)) +
  geom_histogram(binwidth = 2, fill = "skyblue", color = "black") +
  labs(title = "Histogram of BMI",
       x = "BMI",
       y = "Frequency") +
  theme_minimal()

histogram_expenses <- ggplot(insurance_data, aes(x = expenses)) +
  geom_histogram(binwidth = 5000, fill = "lightcoral", color = "black") +
  labs(title = "Histogram of Charges",
       x = "Charges",
       y = "Frequency") +
  theme_minimal()

histogram_age <- ggplot(insurance_data, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "lightgreen", color = "black") +
  labs(title = "Histogram of Age",
       x = "Age",
       y = "Frequency") +
  theme_minimal()

histogram_smoker <- ggplot(insurance_data, aes(x = smoker)) +
  geom_bar(fill = "purple") +
  labs(title = "Histogram of Smoker vs. Non-smoker",
       x = "Smoker",
       y = "Count") +
  theme_minimal()
```

```
grid.arrange(density_plot_region, density_plot_smoker, density_plot_sex,density_plot_children, ncol =

grid.arrange(histogram_bmi, histogram_expenses, histogram_age, histogram_smoker,
             ncol = 2)




histogram_region_1 <- ggplot(subset(insurance_data, region == "northeast"), aes(x = expenses)) +
  geom_histogram(binwidth = 5000, fill = "lightblue", color = "black") +
  labs(title = "Histogram of Charges in Region 1",
       x = "Charges",
       y = "Frequency") +
  theme_minimal()

histogram_region_2 <- ggplot(subset(insurance_data, region == "northwest"), aes(x = expenses)) +
  geom_histogram(binwidth = 5000, fill = "lightgreen", color = "black") +
  labs(title = "Histogram of Charges in Region 2",
       x = "Charges",
       y = "Frequency") +
  theme_minimal()

histogram_region_3 <- ggplot(subset(insurance_data, region == "southeast"), aes(x = expenses)) +
  geom_histogram(binwidth = 5000, fill = "lightcoral", color = "black") +
  labs(title = "Histogram of Charges in Region 3",
       x = "Charges",
       y = "Frequency") +
  theme_minimal()

histogram_region_4 <- ggplot(subset(insurance_data, region == "southwest"), aes(x = expenses)) +
  geom_histogram(binwidth = 5000, fill = "lightyellow", color = "black") +
  labs(title = "Histogram of Charges in Region 4",
       x = "Charges",
       y = "Frequency") +
  theme_minimal()

grid.arrange(histogram_region_1, histogram_region_2, histogram_region_3, histogram_region_4, ncol = 2




'''

# Predict the Claim Distribution with a Hierarchical Gamma Model

'''{r}

library(rstan)
library(dplyr)
```

```
insurance_data$region <- as.numeric(factor(insurance_data$region, levels = unique(insurance_data$regi

insurance_data$sex <- as.numeric(factor(insurance_data$sex, levels = unique(insurance_data$sex)))

insurance_data$smoker <- as.numeric(factor(insurance_data$smoker, levels = unique(insurance_data$smok

head(insurance_data)

insurance_data$expenses

region_1_expsenses <- subset(insurance_data,region== 1)
region_2_expsenses <- subset(insurance_data,region== 2)
region_3_expsenses <- subset(insurance_data,region== 3)
region_4_expsenses <- subset(insurance_data,region == 4)

min_length <- 324


region_1_expsenses <- region_1_expsenses[1:324, ]
region_2_expsenses <- region_2_expsenses[1:324, ]
region_3_expsenses <- region_3_expsenses[1:324, ]
region_4_expsenses <- region_4_expsenses[1:324, ]
region_1_expsenses$expenses

‘‘‘

‘‘‘{r}

#install.packages("V8")


library(rstan)

stan_model <- "
data {
  int<lower=0> N1;          //observations in region 1
  int<lower=0> N2;
  int<lower=0> N3;
  int<lower=0> N4;

  real<lower=0> y1[N1];       // Observed expenses
  real<lower=0> y2[N2];
  real<lower=0> y3[N3];
  real<lower=0> y4[N4];

}




parameters {
  real<lower=0> alpha1;   //alpha for region 1
  real<lower=0> alpha2;
```

```
  real<lower=0> alpha3;
  real<lower=0> alpha4;

  real<lower=0> beta;    //common Beta parameter


  real<lower=0> hyper_alpha; //hyperparam for alpha
  real<lower=0> hyper_beta;  //hypermarpam for beta
}




model {

  //priors for hyperparameters
  hyper_alpha ~ gamma(2, 1);
  hyper_beta ~ gamma(1, 1);
  beta ~ gamma(1, 1);


  //hierarchical model
  alpha1 ~ gamma(hyper_alpha, 1/hyper_beta);
  alpha2 ~ gamma(hyper_alpha, 1/hyper_beta);
  alpha3 ~ gamma(hyper_alpha, 1/hyper_beta);
  alpha4 ~ gamma(hyper_alpha, 1/hyper_beta);


  //likelihood
 y1 ~ gamma(alpha1, beta);
 y2 ~ gamma(alpha2, beta);
 y3 ~ gamma(alpha3, beta);
 y4 ~ gamma(alpha4, beta);

}


generated quantities {

  real<lower=0> y_rep1[N1];
  real<lower=0> y_rep2[N2];
  real<lower=0> y_rep3[N3];
  real<lower=0> y_rep4[N4];

  for (i in 1:N1) y_rep1[i] = gamma_rng(alpha1, beta);
  for (i in 1:N2) y_rep2[i] = gamma_rng(alpha2, beta);
  for (i in 1:N3) y_rep3[i] = gamma_rng(alpha3, beta);
  for (i in 1:N4) y_rep4[i] = gamma_rng(alpha4, beta);

}

"
```

```
data_list <- list(
  N1 = 324,
  y1 = region_1_expsenses$expenses,
  N2 = 324,
  y2 = region_2_expsenses$expenses,
  N3 = 324,
  y3 = region_3_expsenses$expenses,
  N4 = 324,
  y4 = region_4_expsenses$expenses
)

stan_model <- stan_model(model_code = stan_model)
fit <- sampling(stan_model, data = data_list, chains = 4, iter = 20000, warmup = 5000)
print(fit)
plot(fit)
traceplot(fit)


posterior_samples <- extract(fit)

beta_samples <- posterior_samples$beta
beta_mean <- mean(beta_samples)
print(beta_mean)
# [1] 0.0001153254


‘‘‘


# MSE, R\^2, MAE

‘‘‘{r}

#install.packages("Metrics")
library(Metrics)

posterior_samples <- extract(fit)

y_rep1 <- posterior_samples$y_rep1
y_rep2 <- posterior_samples$y_rep2
y_rep3 <- posterior_samples$y_rep3
y_rep4 <- posterior_samples$y_rep4

observed_data <- list(
  y1 = region_1_expsenses$expenses,
  y2 = region_2_expsenses$expenses,
  y3 = region_3_expsenses$expenses,
  y4 = region_4_expsenses$expenses
)

mse_region_1 <- mean((y_rep1 - observed_data$y1)^2)
mse_region_2 <- mean((y_rep2 - observed_data$y2)^2)
mse_region_3 <- mean((y_rep3 - observed_data$y3)^2)
```

```
mse_region_4 <- mean((y_rep4 - observed_data$y4)^2)

mae_region_1 <- mae(y_rep1, observed_data$y1)
mae_region_2 <- mae(y_rep2, observed_data$y2)
mae_region_3 <- mae(y_rep3, observed_data$y3)
mae_region_4 <- mae(y_rep4, observed_data$y4)

cat("MAE for Region 1:", mae_region_1, "\n")
cat("MAE for Region 2:", mae_region_2, "\n")
cat("MAE for Region 3:", mae_region_3, "\n")
cat("MAE for Region 4:", mae_region_4, "\n")

cat("MSE for Region 1:", mse_region_1, "\n")
cat("MSE for Region 2:", mse_region_2, "\n")
cat("MSE for Region 3:", mse_region_3, "\n")
cat("MSE for Region 4:", mse_region_4, "\n")

```

```
# Posterior Predictive Check:

```{r}
library(gridExtra)

yrep1 <- as.array(fit, "y_rep1")
yrep2 <- as.array(fit, "y_rep2")
yrep3 <- as.array(fit, "y_rep3")
yrep4 <- as.array(fit, "y_rep4")

combine_plots <- function(hist_plot, dens_plot, stat_plot1, stat_plot2, title) {
  grid.arrange(hist_plot, dens_plot, stat_plot1, stat_plot2,
               ncol = 2, top = title)
}

hist_plot1 <- ppc_hist(region_1_expsenses$expenses, yrep1[1:8, 1,], binwidth = 50)
dens_plot1 <- ppc_dens_overlay(region_1_expsenses$expenses, yrep1[1:8, 1,], binwidth = 50)
stat_plot1_1 <- ppc_stat(region_1_expsenses$expenses, yrep1[1:500, 1,], stat = 'max')
stat_plot1_2 <- ppc_stat(region_1_expsenses$expenses, yrep1[1:500, 1,])
combined_plot1 <- combine_plots(hist_plot1, dens_plot1, stat_plot1_1, stat_plot1_2, "Posterior Predic

hist_plot2 <- ppc_hist(region_2_expsenses$expenses, yrep2[1:8, 1,], binwidth = 50)
dens_plot2 <- ppc_dens_overlay(region_2_expsenses$expenses, yrep2[1:8, 1,], binwidth = 50)
stat_plot2_1 <- ppc_stat(region_2_expsenses$expenses, yrep2[1:500, 1,], stat = 'max')
stat_plot2_2 <- ppc_stat(region_2_expsenses$expenses, yrep2[1:500, 1,])
combined_plot2 <- combine_plots(hist_plot2, dens_plot2, stat_plot2_1, stat_plot2_2, "Posterior Predic

hist_plot3 <- ppc_hist(region_3_expsenses$expenses, yrep3[1:8, 1,], binwidth = 50)
dens_plot3 <- ppc_dens_overlay(region_3_expsenses$expenses, yrep3[1:8, 1,], binwidth = 50)
stat_plot3_1 <- ppc_stat(region_3_expsenses$expenses, yrep3[1:500, 1,], stat = 'max')
stat_plot3_2 <- ppc_stat(region_3_expsenses$expenses, yrep3[1:500, 1,])
combined_plot3 <- combine_plots(hist_plot3, dens_plot3, stat_plot3_1, stat_plot3_2, "Posterior Predic

hist_plot4 <- ppc_hist(region_4_expsenses$expenses, yrep4[1:8, 1,], binwidth = 50)
```
```

````
dens_plot4 <- ppc_dens_overlay(region_4_expsenses$expenses, yrep4[1:8, 1,], binwidth = 50)
stat_plot4_1 <- ppc_stat(region_4_expsenses$expenses, yrep4[1:500, 1,], stat = 'max')
stat_plot4_2 <- ppc_stat(region_4_expsenses$expenses, yrep4[1:500, 1,])
combined_plot4 <- combine_plots(hist_plot4, dens_plot4, stat_plot4_1, stat_plot4_2, "Posterior Predic

grid.arrange(combined_plot1, combined_plot2, combined_plot3, combined_plot4, ncol = 1)



```

#

\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\~\

# Bayesian Linear Regression Model with Stan

```{r}


library(rstan)
library(bayesplot)


stan_code <- '
data {

  int<lower=0> N; //observations
  vector[N] Age;
  vector[N] Sex;
  vector[N] BMI;
  vector[N] Children;
  vector[N] Smoker;
  vector[N] Region;
  vector[N] Charges;

}

parameters {
  real beta0;
  real beta_age;
  real beta_sex;
  real beta_bmi;
  real beta_children;
  real beta_smoker;
  real beta_region;
  real<lower=0> sigma;  //std of the error
}

model {
  vector[N] y_hat;
````

```
  //prior
  beta0 ~ normal(0, 10);
  beta_age ~ normal(0, 10);
  beta_sex ~ normal(0, 10);
  beta_bmi ~ normal(0, 10);
  beta_children ~ normal(0, 10);
  beta_smoker ~ normal(0, 10);
  beta_region ~ normal(0, 10);
  sigma ~ cauchy(0, 5);

  //likelihood
  y_hat = beta0 + beta_age * Age+beta_sex*Sex+beta_bmi*BMI+beta_children*Children+beta_smoker * Smoke
  Charges~normal(y_hat, sigma);
}




generated quantities {
  vector[N] y_pred;

  // Posr pred dist
  for (i in 1:N) {
    y_pred[i] = normal_rng(beta0+beta_age * Age[i]+beta_sex*Sex[i] + beta_bmi * BMI[i] + beta_childre
  }
}

'

stan_model <- stan_model(model_code = stan_code)

stan_data <- list(
  N = nrow(insurance_data),
  Age = insurance_data$age,
  Sex = insurance_data$sex,
  BMI = insurance_data$bmi,
  Children = insurance_data$children,
  Smoker = insurance_data$smoker,
  Region = insurance_data$region,
  Charges = insurance_data$expenses
)

stan_fit <- sampling(stan_model, data = stan_data, chains = 4, iter = 2000, warmup = 1000)
print(stan_fit)

posterior_predictive <- as.array(stan_fit, "y_pred")
#mcmc_trace(stan_fit)
#mcmc_acf(stan_fit)
traceplot(stan_fit, pars = c("beta0", "beta_age", "beta_sex", "beta_bmi", "beta_children", "beta_smok
```

```

# MSE

```{r}
mse <- mean(apply(posterior_predictive, 2, function(pred) (pred - stan_data$Charges)^2))
mae <- mean(apply(posterior_predictive, 2, function(pred) abs(pred - stan_data$Charges)))
cat("MAE:", mae, "\n")
cat("MSE:", mse, "\n")
```

Posterior Predictive Check

```{r}

library(rstan)
library(bayesplot)

yrep1 <- as.array(stan_fit, "y_pred")
Charges <- insurance_data$expenses



library(gridExtra)
yrep1fixed <- yrep1[1:8, 1,]
yrep1full <- yrep1[1:500, 1,]

hist_plot <- ppc_hist(Charges, yrep1fixed, binwidth = 50)
dens_plot <- ppc_dens_overlay(Charges, yrep1fixed, binwidth = 50)
stat_plot1 <- ppc_stat(Charges, yrep1full, stat = 'max')
stat_plot2 <- ppc_stat(Charges, yrep1full)

# make afunction to combine the plots
combine_plots <- function(hist_plot, dens_plot, stat_plot1, stat_plot2) {
  grid.arrange(hist_plot, dens_plot, stat_plot1, stat_plot2,
               ncol = 2, top = "Posterior Predictive Check")
}
combine_plots(hist_plot, dens_plot, stat_plot1, stat_plot2)




```

Bayesian Gamma Regression Model:

```{r}


library(rstan)
library(bayesplot)
```
```

```
stan_code_gamma <- '
data {
  int<lower=0> N;
  vector[N] Age;
  vector[N] Sex;
  vector[N] BMI;
  vector[N] Children;
  vector[N] Smoker;
  vector[N] Region;
  vector[N] Charges;
}

parameters {
  real beta0;
  real beta_age;
  real beta_sex;
  real beta_bmi;
  real beta_children;
  real beta_smoker;
  real beta_region;
  real<lower=0> shape;
  real<lower=0> rate;
}

model {
  vector[N] mu;

  //prior
  beta0 ~ normal(0, 10);
  beta_age ~ normal(0, 10);
  beta_sex ~ normal(0, 10);
  beta_bmi ~ normal(0, 10);
  beta_children ~ normal(0, 10);
  beta_smoker ~ normal(0, 10);
  beta_region ~ normal(0, 10);
  shape~gamma(1, 1);
  rate~gamma(1, 1);

  mu = exp(beta0 + beta_age * Age + beta_sex * Sex + beta_bmi * BMI +
           beta_children * Children + beta_smoker * Smoker + beta_region * Region);

  //likelihood
  for (i in 1:N) {
    Charges[i]~gamma(shape, 1 / mu[i]); // Scale = 1/mu
  }
}

generated quantities {
  vector[N] y_pred;

  //post pred dist
  for (i in 1:N) {
```

```
    y_pred[i] = gamma_rng(shape, 1 / rate);  // Scale = 1/rate
  }
}

'

stan_model_gamma <- stan_model(model_code = stan_code_gamma)

stan_data_gamma <- list(
  N = nrow(insurance_data),
  Age = insurance_data$age,
  Sex = insurance_data$sex,
  BMI = insurance_data$bmi,
  Children = insurance_data$children,
  Smoker = insurance_data$smoker,
  Region = insurance_data$region,
  Charges = insurance_data$expenses
)

stan_fit_gamma <- sampling(stan_model_gamma, data = stan_data_gamma, chains = 4, iter = 2000, warmup
print(stan_fit_gamma)

posterior_predictive_gamma <- as.array(stan_fit_gamma, "y_pred")
traceplot(stan_fit_gamma, pars = c("beta0", "beta_age", "beta_sex", "beta_bmi", "beta_children", "bet
```

```
'''

# Posterior Predictive Check:

'''{r}


library(rstan)
library(bayesplot)


yrep1 <- as.array(stan_fit_gamma, "y_pred")

Charges <- insurance_data$expenses


yrep1fixed <- yrep1[1:8, 1,]
hist_plot <- ppc_hist(Charges, yrep1fixed, binwidth = 50)
dens_plot <- ppc_dens_overlay(Charges, yrep1fixed, binwidth = 50)

yrep1fixed <- yrep1[1:500, 1,]
```

```
stat_plot1 <- ppc_stat(Charges, yrep1fixed, stat = 'max')
stat_plot2 <- ppc_stat(Charges, yrep1fixed)

combine_plots <- function(hist_plot, dens_plot, stat_plot1, stat_plot2) {
  grid.arrange(hist_plot, dens_plot, stat_plot1, stat_plot2,
               ncol = 2, top = "Posterior Predictive Check")
}

combine_plots(hist_plot, dens_plot, stat_plot1, stat_plot2)
```

```

## References


[1] Nursnaaz. *Insurance Premium Prediction*. https://www.kaggle.com/nursnaaz/insurance-premium-prediction. Kaggle Dataset. 2017.

[2] Jing Zhang and Tatjana Miljkovic. "Ratemaking for a New Territory: Enhancing GLM Pricing Model with a Bayesian Analysis". In: *Casualty Actuarial Society E-Forum* Spring 2018-Volume 2 (2018). Available online. URL: https://www.casact.org/sites/default/files/database/forum_18spforumv2_03_zhang_miljkovic.pdf.
```