A statistical analysis of Insurance pricing

Mohamed Sabbah

Introduction

Modern healthcare systems rely heavily on medical insurance as they provide financial protection and access to medical services. Exploring the factors affecting insurance charges can help companies and individuals to make better informed decisions. A statistical analysis is carried out in this project to identify the key factors that affect insurance charges, as well as make predictions regarding said charges.

The Dataset

Medical Cost Personal Dataset consists of 1338 records aimed at examining factors that affect medical insurance costs. It includes the following variables:

- **Age**: Age of the primary beneficiary.
- **Sex**: Gender of the insurance policyholder (female or male).
- **BMI**: Body Mass Index, an objective measure of body weight relative to height (kg/m^2), with an ideal range of 18.5 to 24.9.
- **Children**: Number of children or dependents covered by the health insurance.
- **Smoker**: Indicates whether the individual is a smoker.
- **Region**: The residential area of the beneficiary in the U.S. (northeast, southeast, southwest, northwest).
- **Charges**: Medical costs billed to the individual by the health insurance.

EDA

We start with an explanatory data analysis for the dataset, we first display the dataset as shown below.

Table 1: First 5 rows of the dataset

age	sex	bmi	children	smoker	region	charges
19	female	27.900	0	yes	southwest	16884.924
18	male	33.770	1	no	southeast	1725.552
28	male	33.000	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.471
32	male	28.880	0	no	northwest	3866.855
31	female	25.740	0	no	southeast	3756.622

The next step is getting the summaries for the different columns, these summaries include the mean, median, minimum, maximum, and quantiles of said columns.

Table 2: Dataset statistical summary

age	sex	bmi	children	smoker	region	charges
Min. :18.00	Length:1338	Min. :15.96	0:574	Length:1338	Length:1338	Min. : 1122
1st Qu.:27.00	Class :character	1st Qu.:26.30	1:324	Class :character	Class :character	1st Qu.: 4740
Median :39.00	Mode :character	Median :30.40	2:240	Mode :character	Mode :character	Median : 9382
Mean :39.21	NA	Mean :30.66	3:157	NA	NA	Mean :13270
3rd Qu.:51.00	NA	3rd Qu.:34.69	4: 25	NA	NA	3rd Qu.:16640
Max. :64.00	NA	Max. :53.13	5: 18	NA	NA	Max. :63770

The summary shows important details about the variables, including age (mean: 39.21), bmi (mean: 30.66), and children (mean: 1.095). Charges, which ranges from 1122 to 63770 with a mean of 13270, serves as the target variable.The rest of the variables have NA values because they are categorical variables. The next step is to check for missing values in the dataset, as you can see there are 0 missing values.

Table 3: Count of missing values for each variable

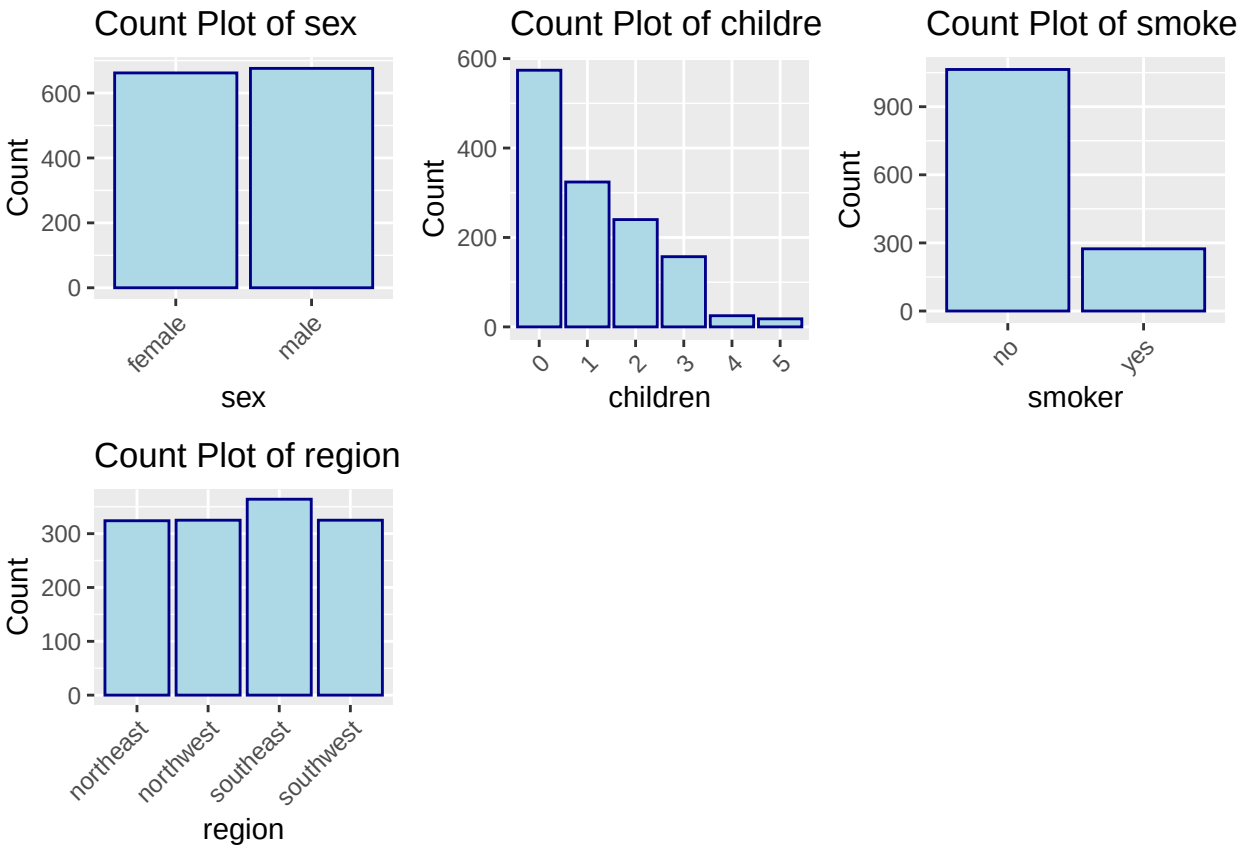
	x
age	0
sex	0
bmi	0
children	0
smoker	0
region	0
charges	0

Univariate analysis and visualizations

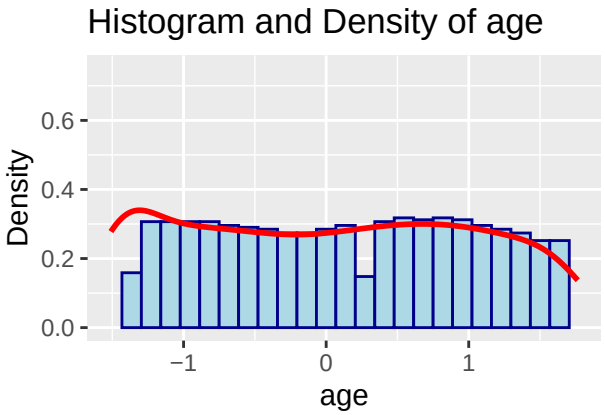
Before we carry out any visualizations, we'll start by centering and scaling the data. Below are some data points after that process is done:

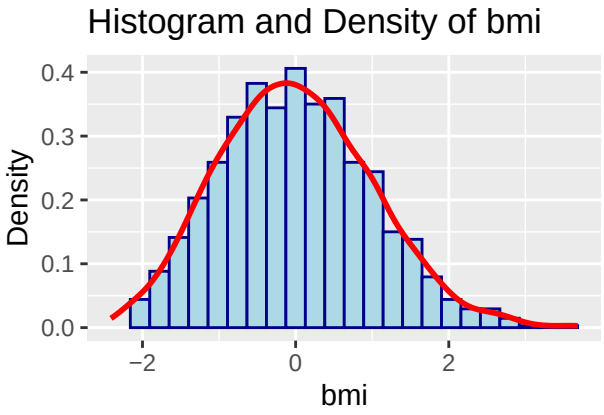
Table 4: First 5 rows of the dataset after scaling and centering it

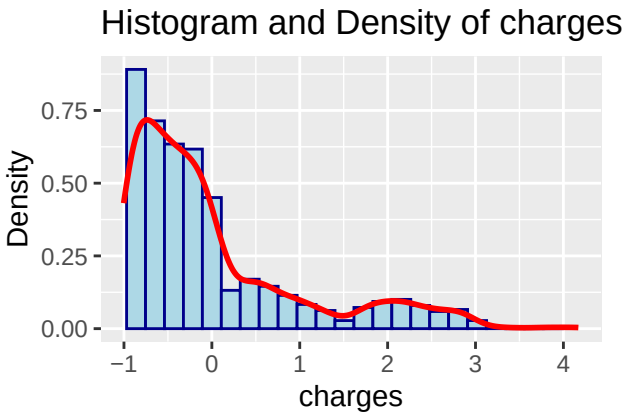
age	sex	bmi	children	smoker	region	charges
-1.4382265	female	-0.4531506	0	yes	southwest	0.2984722
-1.5094011	male	0.5094306	1	no	southeast	-0.9533327
-0.7976553	male	0.3831636	3	no	southeast	-0.7284023
-0.4417824	male	-1.3050431	0	no	northwest	0.7195739
-0.5129570	male	-0.2924471	0	no	northwest	-0.7765118
-0.5841316	female	-0.8073542	0	no	southeast	-0.7856145



For the categorical variables, we use a bar plot as shown above. For sex, we see that it's almost 50% males and 50% females. With children we see that the count decreases as the number of children increases. As for smoking, we see that the number of people who don't smoke is much higher than those who do smoke. Finally, when it comes to the region, we see that almost all regions are equal, with only the southeast being more prevalent.







As for the continuous variables, we used a histogram and fitted a distribution on top of it. From what we can see, there might be a chance that bmi follows a normal distribution. We can check that using the Shapiro test.

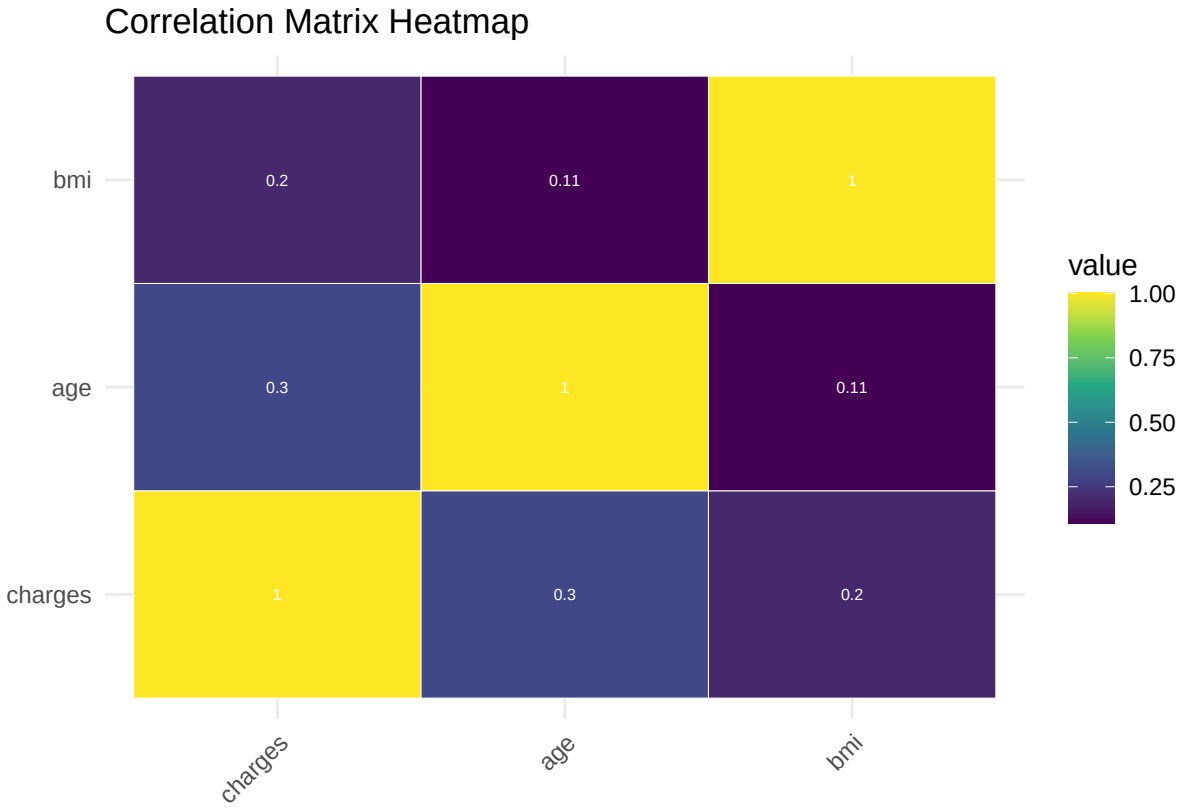
```
##
##  Shapiro-Wilk normality test
##
## data:  data$bmi
## W = 0.99389, p-value = 2.605e-05
```

With a very small value, we reject H_0 (that BMI follows a normal distribution). As for the other 2 variables, we can tell from the density plots that they do not follow a normal distribution either.

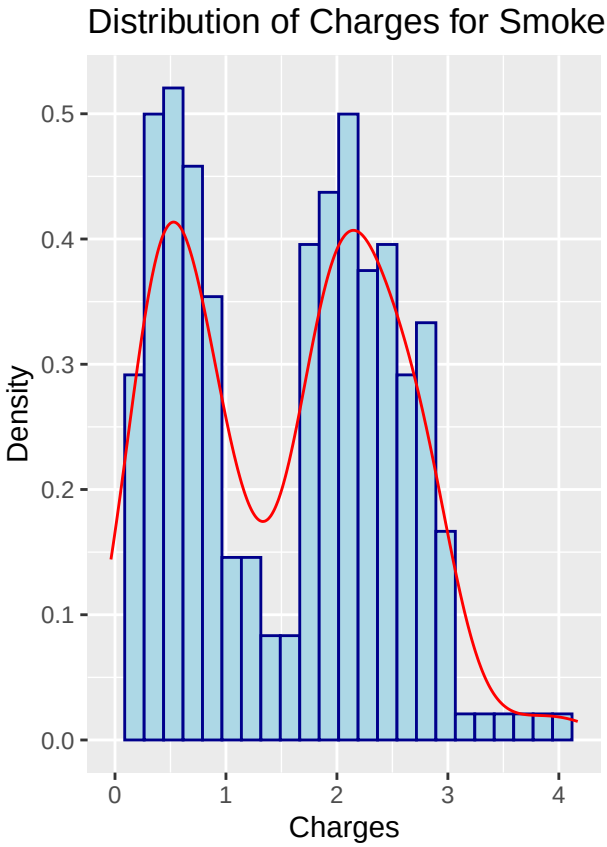
Bivariate analysis and visualizations

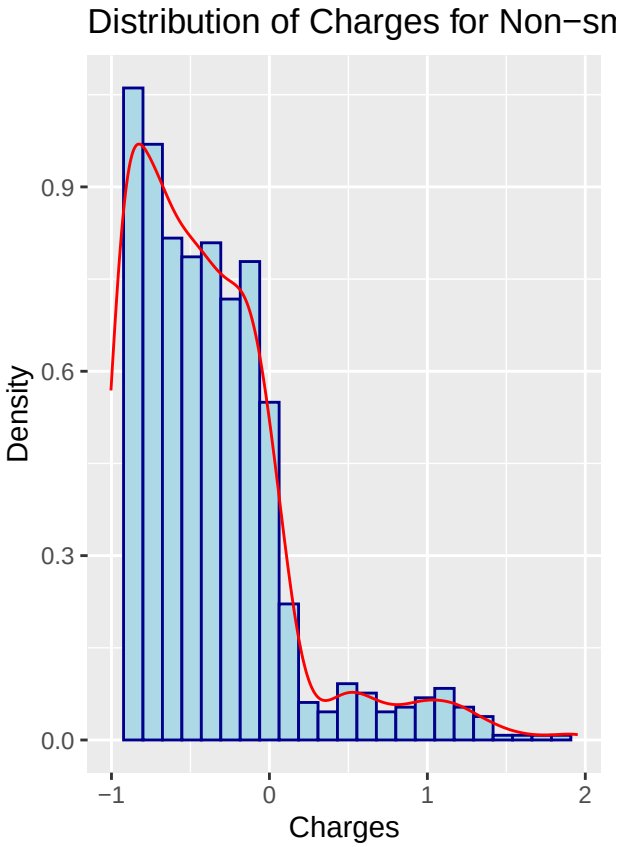
Correlation Matrix Heatmap

We first start with a correlation matrix heatmap to check the variables with the strongest correlation.



As we can see, the variables with the highest correlation with the target variable are the smoker, age, and bmi variables. With smoker having by far the strongest relationship. Let's First check the distributions of smoker vs non smokers when it comes to charges.





As we can see from the distributions above, insurance charges for non-smokers is concentrated around the lower values and heavily skewed as oposed to the smokers who have a more spread out distribution. This makes sense as insurance companies will definitely charge smokers more than non-smokers due to the health risks associated with smoking. Let's check the skewness:

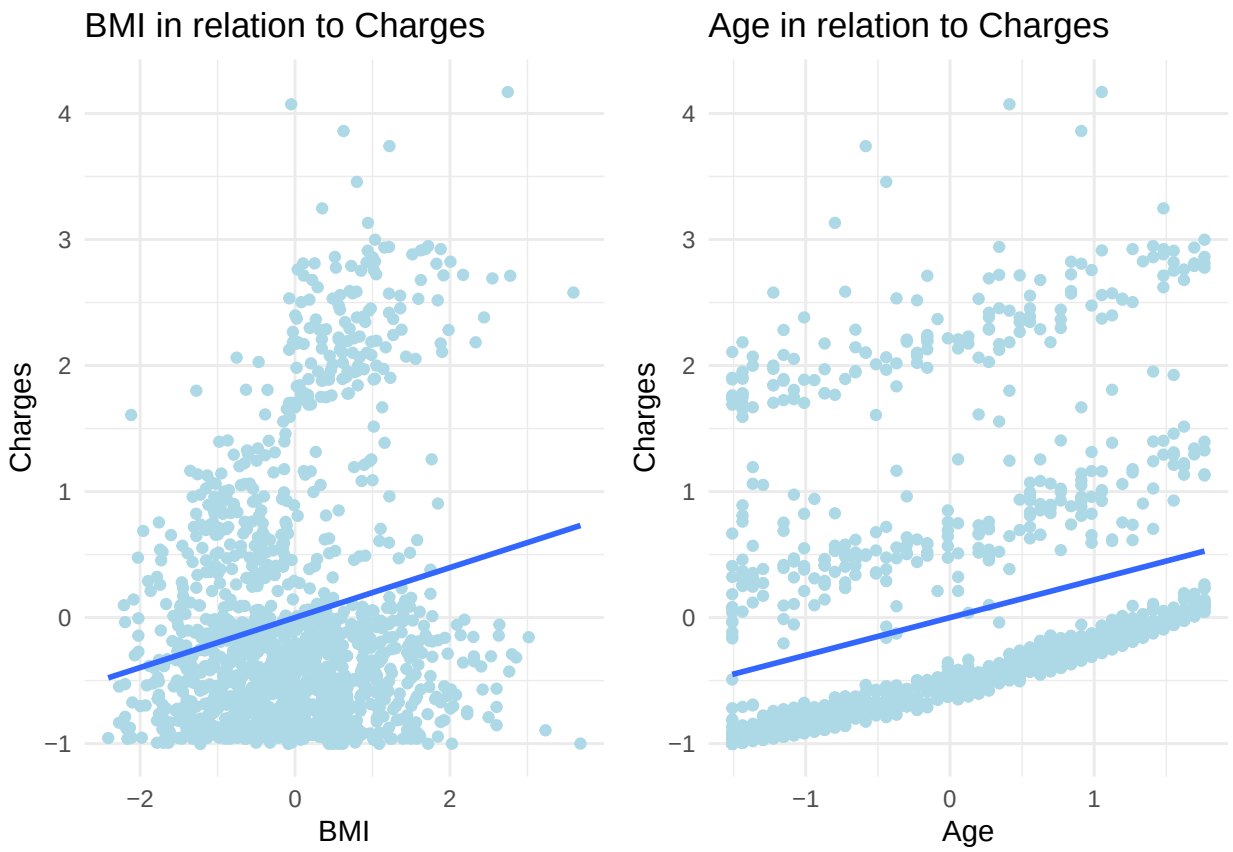
```
## [1] "Smokers' skewness: 0.126816852263372"
```

```
## [1] "Non-smokers' skewness: 1.53378594889156"
```

This supports our claim as the Non-smokers' charges is right skewed, indeed indicating a higher concentration of values for the lower charges.

Now let's check the relationship between bmi and charges, as well as age and charges. We will use a scatter plot with a best fit line to explore the relationships.

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



As we can see, there is a positive correlation between both BMI and age with charges. This makes sense as for the BMI, the larger it is the more likely a person is obese, which is well documented to be the cause of multiple health issues. And as for the Age, as people age they start encountering more health problems.

The models

The frequentist approach

We will first start with a frequentist approach to solve this problem. After conducting an EDA, we can see that a linear model might be appropriate due to the relationships between the variables. A linear model is a mathematical equation that models the relationship between a dependent variable and one or more independent variables by assuming the relationship is linear. The model for multiple predictors can be expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Where: $X_1, X_2, \ldots, X_p$ are the independent variables. $\beta_1, \beta_2, \ldots, \beta_p$ are the corresponding coefficients for each variable. The goal of fitting a linear model is to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$ using ordinary least squares (OLS), which minimizes the sum of squared residuals (the differences between observed and predicted values). The error term ϵ represents random variability and is assumed to follow a normal distribution with zero mean and constant variance.This is a frequentist approach because it estimates parameters from the observed data, without incorporating prior beliefs or distributions about the parameters. This assumes that the data comes from a fixed, true distribution, and the parameters are fixed but unknown quantities estimated from the sample. Inference, such as confidence intervals and hypothesis testing, is based on the sampling distribution of the estimates.

The linear model

```
linear_model <- lm(charges~.,data)
print(summary(linear_model))
```

```
##
## Call:
## lm(formula = charges ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93665 -0.23177 -0.08399  0.11627  2.45677
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.36298    0.06950 -34.002  < 2e-16 ***
## age          0.29850    0.01379  21.647  < 2e-16 ***
## sex         -0.01083    0.02748  -0.394 0.693681
## bmi          0.16747    0.01396  11.997  < 2e-16 ***
## children     0.03958    0.01137   3.483 0.000513 ***
## smoker       1.96700    0.03401  57.839  < 2e-16 ***
## region      -0.02920    0.01255  -2.328 0.020077 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5004 on 1331 degrees of freedom
## Multiple R-squared:  0.7507, Adjusted R-squared:  0.7496
## F-statistic: 668.1 on 6 and 1331 DF,  p-value: < 2.2e-16
```

The model's p_value is < 2.2e-16, which indicates that the model as a whole is statistically significant. This means that there is strong evidence against the null hypothesis, which typically states that there is no effect or no association between the predictor variable and the outcome variable. For our model, we shall discard any variables that have a p_value >0.05, this is the case for the sex variable as it has a very high p_value (0.693681). Hence, we create a second model without it.

```
linear_model <- lm(charges~age+bmi+children+smoker+region,data)
print(summary(linear_model))
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker + region,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94170 -0.23162 -0.08192  0.11563  2.45205
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.37794    0.05819 -40.868  < 2e-16 ***
## age          0.29864    0.01378  21.670  < 2e-16 ***
## bmi          0.16720    0.01394  11.995  < 2e-16 ***
## children     0.03951    0.01136   3.478 0.000522 ***
## smoker       1.96599    0.03390  57.992  < 2e-16 ***
## region      -0.02919    0.01254  -2.327 0.020104 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5002 on 1332 degrees of freedom
## Multiple R-squared:  0.7507, Adjusted R-squared:  0.7498
## F-statistic: 802.2 on 5 and 1332 DF,  p-value: < 2.2e-16
```

The new model has an R-squared of 0.75, which quantifies the proportion of variance in the dependent variables that can be explained by the independent variables in the model. Although it is not bad, it can definitely be improved upon.

The next step is to train the model. We start by splitting the data into training and testing data sets. The split is a 75% training and 25% testing data.

```
## [1] "train length: 1003"
```

```
## [1] "test length: 335"
```

Then we fit the model using the training dataset.

```
linear_model_freq <- lm(charges~age+bmi+children+smoker+region,train)
print(summary(linear_model_freq))
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker + region,
##     data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9174 -0.2415 -0.0904  0.1076  2.4665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.30925    0.06875 -33.589   <2e-16 ***
## age          0.29645    0.01661  17.848   <2e-16 ***
## bmi          0.16965    0.01670  10.160   <2e-16 ***
## children     0.03314    0.01356   2.443   0.0147 *
## smoker       1.94557    0.04010  48.520   <2e-16 ***
## region      -0.04083    0.01500  -2.722   0.0066 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5162 on 997 degrees of freedom
## Multiple R-squared:  0.7351, Adjusted R-squared:  0.7338
## F-statistic: 553.4 on 5 and 997 DF,  p-value: < 2.2e-16
```

After the model was trained successfully, we now make predictions using the testing dataset and calculate the RMSE, which is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where:

- y_i: The observed values.

- $\hat{y}_i$: The predicted values.
- n: The number of observations.

But first,we check the parameter credible intervals.

```
##                   2.5 %     97.5 %
## (Intercept) -2.44415623 -2.1743356
## age          0.26386021  0.3290480
## bmi          0.13688556  0.2024208
## children     0.00652281  0.0597568
## smoker       1.86688076  2.0242534
## region      -0.07026359 -0.0113951
```

Now we use the test dataset to make predictions and calculate the RMSE.

```
test$y_hat <- predict(linear_model_freq, newdata = test)
rmse_value <- rmse(test$charges, test$y_hat)
paste("RMSE:", rmse_value)
```

```
## [1] "RMSE: 0.450634676462848"
```

Considering that our scaled range is (-1,4), this RMSE is moderate. This coupled with the R-squared we obtained shows that there is room for improvement.

An alternative model with polynomial features

One of the limitations of our linear model is that it assumes a linear relationship between the independent variables and the dependent variable. A solution to combat this is by adding polynomial features to our model. This extends the model to capture non-linear relationships between the predictors and the dependent variable. But first, we need to define polynomial features. Polynomial features are new features added to the model that are created by raising the existing features to a degree or by creating interaction terms between them. For example, if we have 2 features x_1 and x_2, creating polynomial features up to the second degree will leave us with $(c, x_1, x_2, x_1^2, x_2^2, x_1 * x_2)$. In our case, we'll use 2 degrees polynomial features.

```
data_pol <-subset(data,select=-c(sex,charges))

data_pol <- as.data.frame(model.matrix(~ .^2 -1, data = data_pol))
data_pol$age_squared <-data_pol$age^2
data_pol$bmi_squared <-data_pol$bmi^2
data_pol$children_squared <-data_pol$children^2
data_pol$smoker_squared <-data_pol$smoker^2
data_pol$region_squared <-data_pol$region^2




data_pol$charges<-data$charges
colnames(data_pol) <- gsub(":", "_", colnames(data_pol))
```

```
linear_model_poly_2 <-  lm(charges ~ ., data = data_pol)

print(summary(linear_model_poly_2))
```

```
##
## Call:
## lm(formula = charges ~ ., data = data_pol)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92040 -0.14877 -0.10570 -0.03852  2.50562
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.527218   0.133274 -18.963  < 2e-16 ***
## age               0.261827   0.045331   5.776 9.54e-09 ***
## bmi              -0.658825   0.047191 -13.961  < 2e-16 ***
## children          0.178662   0.051739   3.453 0.000572 ***
## smoker            2.000178   0.083967  23.821  < 2e-16 ***
## region           -0.089835   0.066207  -1.357 0.175050
## age_bmi           0.008531   0.011127   0.767 0.443404
## age_children     -0.001948   0.009768  -0.199 0.841935
## age_smoker       -0.005247   0.027255  -0.193 0.847367
## age_region        0.019107   0.010122   1.888 0.059287 .
## bmi_children      0.001032   0.009321   0.111 0.911866
## bmi_smoker        0.720801   0.026916  26.780  < 2e-16 ***
## bmi_region       -0.018807   0.010745  -1.750 0.080289 .
## children_smoker  -0.032773   0.023320  -1.405 0.160150
## children_region  -0.012664   0.008108  -1.562 0.118543
## smoker_region     0.014478   0.025355   0.571 0.568101
## age_squared       0.064332   0.013416   4.795 1.81e-06 ***
## bmi_squared      -0.024542   0.008240  -2.978 0.002951 **
## children_squared -0.009556   0.006760  -1.414 0.157682
## smoker_squared           NA         NA      NA       NA
## region_squared    0.012246   0.010978   1.115 0.264855
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3953 on 1318 degrees of freedom
## Multiple R-squared:  0.8459, Adjusted R-squared:  0.8437
## F-statistic: 380.9 on 19 and 1318 DF,  p-value: < 2.2e-16
```

Already, we see a massive improvement in the R-squared as it is now 0.85 rising from 0.75. Now, let's remove the statistically insignificant variables.

```
linear_model_poly_2 <-  lm(charges ~age+bmi+children+smoker+bmi_smoker+age_squared +bmi_squared, data =

print(summary(linear_model_poly_2))
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker + bmi_smoker +
##     age_squared + bmi_squared, data = data_pol)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94968 -0.14533 -0.11159 -0.04934  2.50760
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.520036   0.044052 -57.206  < 2e-16 ***
## age          0.301393   0.011008  27.380  < 2e-16 ***
## bmi         -0.713772   0.033958 -21.019  < 2e-16 ***
```

```
## children     0.055235   0.009461   5.838 6.62e-09 ***
## smoker       1.965897   0.026957  72.926  < 2e-16 ***
## bmi_smoker   0.722473   0.026303  27.467  < 2e-16 ***
## age_squared  0.061445   0.013152   4.672 3.29e-06 ***
## bmi_squared -0.026668   0.007985  -3.340 0.000861 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3976 on 1330 degrees of freedom
## Multiple R-squared:  0.8427, Adjusted R-squared:  0.8419
## F-statistic:  1018 on 7 and 1330 DF,  p-value: < 2.2e-16
```

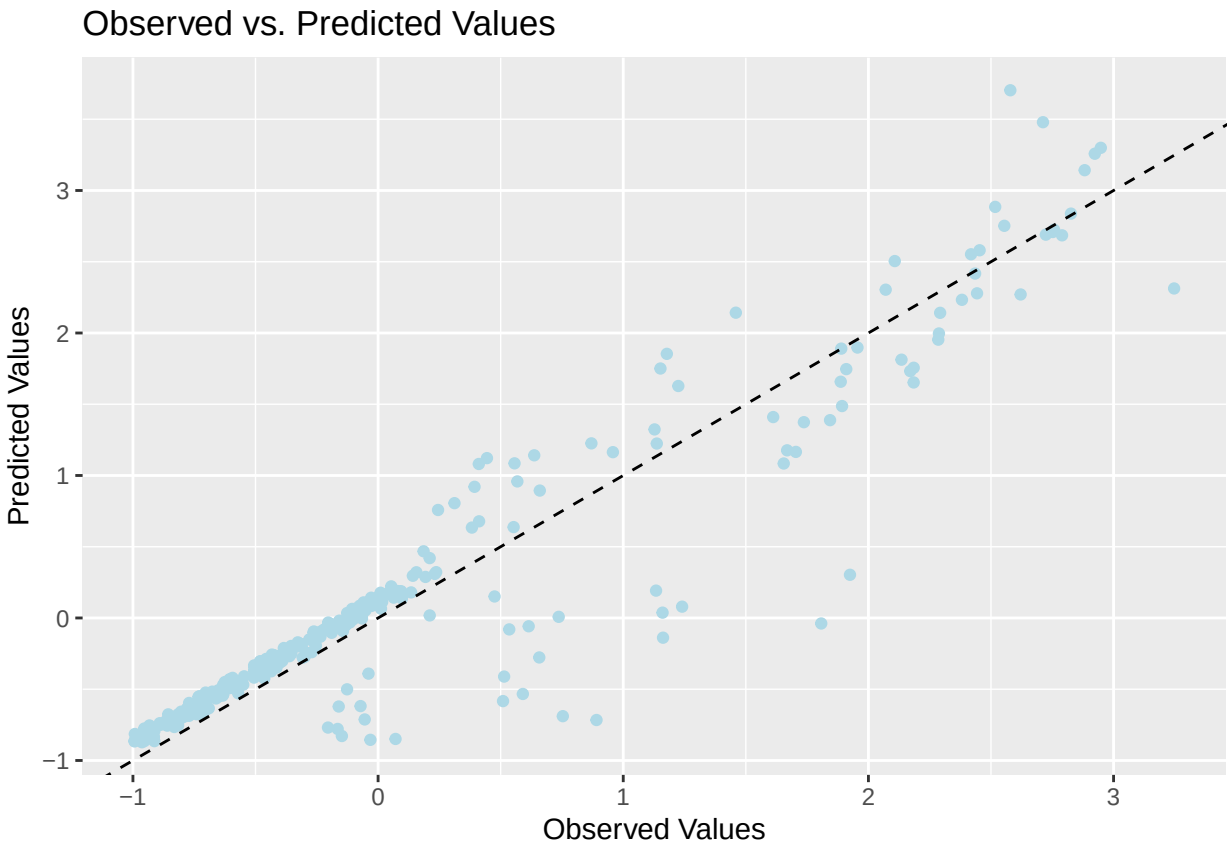
Now let's do the train-test split

Checking the RMSE

```
test$y_hat_pol <- predict(linear_model_poly_2, newdata = test_pol)
rmse_value <- rmse(test$charges, test$y_hat_pol)
paste("RMSE:", rmse_value)
```

```
## [1] "RMSE: 0.344693932381953"
```

This shows an obvious improvement in both the R-squared and the RMSE. Finally, let's visualize our results.



The Bayesian Approach

In this model, we use a Bayesian linear regression approach to estimate the relationship between a continuous response variable and several predictor variables. The model assumes that the response variable, Y, is normally distributed, with its mean, μ, being a linear combination of the predictor variables, and the error is

modeled by a precision parameter, τ (the inverse of variance). The general form of the model is as follows:

$$Y_i \sim N(\mu_i, \tau)$$

$$\mu_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi}$$

We adopt a Bayesian framework, which involves specifying prior distributions for all the unknown parameters in the model. In this case, we use weakly informative priors, which are designed to provide some initial regularization but are broad enough not to dominate the posterior in the presence of sufficient data.For the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$, we assume independent normal priors with a mean of 0 and a small precision (large variance):

$$\beta_j \sim N(0, 0.1) \quad \text{for } j = 0, 1, \ldots, p$$

These weakly informative priors reflect our initial belief that the parameters are likely to be close to 0, but they allow for large deviations based on the data. This helps to regularize the model without imposing strong assumptions.For the precision parameter τ, we assign a Gamma prior:

$$\tau \sim \text{Gamma}(0.1, 0.1)$$

The Gamma distribution is commonly used as a prior for precision in Bayesian models, and the parameters 0.1 and 0.1 make it a weakly informative prior, allowing the data to significantly influence the posterior distribution.The likelihood function specifies how the observed data are generated given the model parameters. For each observation i, the response variable Y_i is assumed to follow a normal distribution centered around μ_i with precision τ:

$$Y_i \sim N(\mu_i, \tau)$$

The mean μ_i is modeled as a linear combination of the predictors $X_{1i}, X_{2i}, \ldots, X_{pi}$ and the corresponding coefficients $\beta_1, \beta_2, \ldots, \beta_p$.In the Bayesian framework, we combine the prior distributions with the likelihood to obtain the posterior distributions of the model parameters. This is achieved using Markov Chain Monte Carlo (MCMC) methods, such as Gibbs sampling in JAGS, which allows us to generate samples from the posterior distribution.The posterior distributions provide not just point estimates for the parameters but a full distribution, allowing us to quantify uncertainty and make probabilistic statements about the parameters.This Bayesian model incorporates weakly informative priors to regularize the parameter estimates while allowing the data to drive the inference. The Bayesian approach provides a probabilistic interpretation of the model parameters, giving us posterior distributions that reflect the uncertainty in our estimates.

The Initial model

The initial model we will be using is a linear model without any transformations done on the features, this is shown below.

```
jags_data<- list(
  N = nrow(train),
  charges = train$charges,
  age = train$age,
  sex = train$sex,
  bmi = train$bmi,
  children = train$children,
  smoker = train$smoker,
  region = train$region
)

jags_code<- "
model {
#priors
beta0 ~ dnorm(0,0.1)
beta_age ~ dnorm(0,0.1)
```

```
beta_sex ~ dnorm(0,0.1)
beta_bmi ~ dnorm(0,0.1)
beta_children ~ dnorm(0,0.1)
beta_smoker ~ dnorm(0,0.1)
beta_region ~ dnorm(0,0.1)
tau ~ dgamma(0.1, 0.1)

#likelihood
for (i in 1:N) {
    mu[i] <- beta0 +beta_age * age[i] + beta_sex * sex[i]+beta_bmi * bmi[i] + beta_children * children[i

    charges[i] ~ dnorm(mu[i], tau)
  }







}
"
jags_model <- jags(data = jags_data, inits = NULL,
                   parameters.to.save = c("beta0", "beta_age", "beta_sex","beta_bmi","beta_children","be
                   model.file = textConnection(jags_code),
                   n.chains = 3, n.iter = 12000, n.burnin = 2000, n.thin = 10)
```

```
## module glm loaded
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 1003
##    Unobserved stochastic nodes: 8
##    Total graph size: 8577
##
## Initializing model
```

```
jags_model
```

```
## Inference for Bugs model at "4", fit using jags,
##  3 chains, each with 12000 iterations (first 2000 discarded), n.thin = 10
##  n.sims = 3000 iterations saved
##                mu.vect sd.vect     2.5%      25%      50%      75%    97.5%
## beta0           -2.330   0.083   -2.493   -2.385   -2.330   -2.273   -2.172
## beta_age         0.297   0.017    0.263    0.285    0.297    0.308    0.329
## beta_bmi         0.170   0.017    0.137    0.158    0.169    0.181    0.202
## beta_children    0.033   0.014    0.006    0.024    0.033    0.042    0.060
## beta_region     -0.041   0.015   -0.070   -0.051   -0.041   -0.032   -0.013
## beta_sex         0.016   0.033   -0.049   -0.006    0.016    0.038    0.081
## beta_smoker      1.945   0.040    1.867    1.917    1.944    1.971    2.023
## tau              3.749   0.168    3.437    3.635    3.745    3.861    4.084
## deviance      1521.772   4.127 1515.941 1518.828 1521.131 1523.863 1531.146
##                Rhat n.eff
```

```
## beta0         1.001  3000
## beta_age      1.001  2700
## beta_bmi      1.001  3000
## beta_children 1.001  3000
## beta_region   1.002  1300
## beta_sex      1.001  2400
## beta_smoker   1.001  3000
## tau           1.001  3000
## deviance      1.001  3000
##
## For each parameter, n.eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
##
## DIC info (using the rule, pD = var(deviance)/2)
## pD = 8.5 and DIC = 1530.3
## DIC is an estimate of expected predictive error (lower deviance is better).
```

If a parameter's credible interval includes 0, it suggests that the effect of the parameter could plausibly be zero, meaning there may be no significant association between that predictor and the outcome. The parameter that fits this condition is beta_sex, hence, we shall remove it.

```
jags_data<- list(
  N = nrow(train),
  charges = train$charges,
  age = train$age,
  bmi = train$bmi,
  children = train$children,
  smoker = train$smoker,
  region = train$region
)

jags_code<- "
model {
#priors
beta0 ~ dnorm(0,0.1)
beta_age ~ dnorm(0,0.1)
beta_bmi ~ dnorm(0,0.1)
beta_children ~ dnorm(0,0.1)
beta_smoker ~ dnorm(0,0.1)
beta_region ~ dnorm(0,0.1)
tau ~ dgamma(0.1, 0.1)

#likelihood
for (i in 1:N) {
    mu[i] <- beta0 +beta_age * age[i] +beta_bmi * bmi[i] + beta_children * children[i] + beta_smoker * s


    charges[i] ~ dnorm(mu[i], tau)
  }
```

```
}
"
jags_model <- jags(data = jags_data, inits = NULL,
                   parameters.to.save = c("beta0", "beta_age","beta_bmi","beta_children","beta_smoker",
                   model.file = textConnection(jags_code),
                   n.chains = 3, n.iter = 12000, n.burnin = 2000, n.thin = 10)
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 1003
##    Unobserved stochastic nodes: 7
##    Total graph size: 7570
##
## Initializing model
```

```
jags_model
```

```
## Inference for Bugs model at "5", fit using jags,
##  3 chains, each with 12000 iterations (first 2000 discarded), n.thin = 10
##  n.sims = 3000 iterations saved
##                mu.vect sd.vect     2.5%      25%      50%      75%    97.5%
## beta0           -2.307   0.069   -2.439   -2.354   -2.309   -2.259   -2.171
## beta_age         0.297   0.016    0.264    0.286    0.296    0.307    0.328
## beta_bmi         0.169   0.017    0.137    0.158    0.169    0.181    0.202
## beta_children    0.033   0.013    0.007    0.023    0.033    0.042    0.060
## beta_region     -0.041   0.015   -0.071   -0.051   -0.041   -0.031   -0.011
## beta_smoker      1.945   0.040    1.868    1.918    1.945    1.972    2.020
## tau              3.747   0.161    3.437    3.640    3.745    3.856    4.060
## deviance      1520.802   3.766 1515.591 1518.071 1520.210 1522.731 1529.788
##                Rhat n.eff
## beta0         1.002  1400
## beta_age      1.002  1600
## beta_bmi      1.001  2500
## beta_children 1.001  3000
## beta_region   1.001  3000
## beta_smoker   1.001  2300
## tau           1.001  3000
## deviance      1.001  3000
##
## For each parameter, n.eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
##
## DIC info (using the rule, pD = var(deviance)/2)
## pD = 7.1 and DIC = 1527.9
## DIC is an estimate of expected predictive error (lower deviance is better).
```

In order to compare between models, we need to discuss the different metrics that we will use.

Deviance and DIC

The next step is to look at the deviance and the DIC(Deviance Information Criterion) of the models, but first we'll explain exactly what these are. Deviance is a measure of the goodness of fit of a statistical model.

it is calculated as:

$$\text{Deviance} = -2 \times (\log\text{-likelihood of the fitted model})$$

with the lower the deviance, the better the fit of the model. The DIC on the other hand incorporates both the deviance of a model as well as its complexity, and it is defined as:

$$DIC(m) = 2D(\bar{\theta}_m, m) + 2p_m$$

where:

- p_m: Can be interpreted as the number of effective parameters for model m given by $p_m = \overline{D(\theta_m, m)} - D(\bar{\theta}_m, m)$
- $\underline{D(\theta_m, m)}$:The deviance.
- $\overline{D(\theta_m, m)}$: The posterior mean of the deviance.
- $\bar{\theta}_m$: The posterior mean of the parameters involved in model m.

$\hat{R}$

Rhat, also known as the potential scale reduction factor, is a diagnostic statistic used in Bayesian analysis to assess the convergence of Markov Chain Monte Carlo (MCMC) simulations. It is calculated by:

$$\hat{R} = \sqrt{\frac{V_{\text{between}} + (N+1) \cdot V_{\text{within}}}{N}}$$

where:

- $V_{between}$: The variance of the means of the chains.
- V_{within}: The average variance within each chain.
- N: The number of iterations in each chain.

An $\hat{R}$ value of 1 indicates that the chains have converged and are sampling from the same distribution, $\hat{R} < 1.05$ Generally indicates good convergence, and $1.05 < \hat{R} < 1.1$ suggests potential convergence issues.

n.eff.

n.eff. or effective sample size, is a measure used to assess the number of independent samples drawn from a posterior distribution after accounting for autocorrelation in MCMC simulations. A higher n.eff.. indicates more independent information in the samples, suggesting that the MCMC has mixed well and the posterior distribution is well approximated. A good n.eff. also needs to be close to the number of samples.

Now since we outlined the metrics we will be looking at, let's compare between both our models. We can see that the model without beta_sex has a lower (in other words, lower), a better n.eff. when it comes to most parameters, and the same $\hat{R}$. Hence, it is the better option.

The alternative model with polynomial features

As we did in the "frequentist approach" section, we will introduce polynomial features to our model to see if it can improve it.

```
jags_alt_data<- list(
  N = nrow(train_pol),
  charges = train_pol$charges,
  age = train_pol$age,
  bmi = train_pol$bmi,
  children = train_pol$children,
  smoker = train_pol$smoker,
```

```
  region = train_pol$region,
  age_bmi=train_pol$age_bmi,
  age_children=train_pol$age_children,
  age_smoker=train_pol$age_smoker,
  age_region=train_pol$age_region,
  bmi_children = train_pol$bmi_children,
  bmi_smoker = train_pol$bmi_smoker,
  bmi_region = train_pol$bmi_region,
  children_smoker = train_pol$children_smoker,
  children_region = train_pol$children_region,
  smoker_region = train_pol$smoker_region,
  age_squared = train_pol$age_squared,
  bmi_squared = train_pol$bmi_squared,
  children_squared = train_pol$children_squared,
  smoker_squared = train_pol$smoker_squared,
  region_squared = train_pol$region_squared)

jags_alt_code <- "
model {
#priors
beta0 ~ dnorm(0,0.1)
beta_age ~ dnorm(0,0.1)
beta_bmi ~ dnorm(0,0.1)
beta_children ~ dnorm(0,0.1)
beta_smoker ~ dnorm(0,0.1)
beta_region ~ dnorm(0,0.1)
beta_age_bmi ~ dnorm(0,0.1)
beta_age_children ~ dnorm(0,0.1)
beta_age_smoker ~ dnorm(0,0.1)
beta_age_region ~ dnorm(0,0.1)
beta_bmi_children ~ dnorm(0,0.1)
beta_bmi_smoker ~ dnorm(0,0.1)
beta_bmi_region ~ dnorm(0,0.1)
beta_children_smoker ~ dnorm(0,0.1)
beta_children_region ~ dnorm(0,0.1)
beta_smoker_region ~ dnorm(0,0.1)
beta_age_squared ~ dnorm(0,0.1)
beta_bmi_squared ~ dnorm(0,0.1)
beta_children_squared ~ dnorm(0,0.1)
beta_smoker_squared ~ dnorm(0,0.1)
beta_region_squared ~ dnorm(0,0.1)









tau ~ dgamma(0.1, 0.1)

#likelihood
for (i in 1:N) {
    mu[i] <- beta0 +beta_age * age[i] +beta_bmi * bmi[i] + beta_children * children[i] + beta_smoker * s
```

```
    charges[i] ~ dnorm(mu[i], tau)
  }




}
"
jags_alt_model <- jags(data = jags_alt_data, inits = NULL,
                   parameters.to.save = c("beta0", "beta_age","beta_bmi","beta_children","beta_smoker",
"beta_age_smoker" ,
"beta_age_region" ,
"beta_bmi_children",
"beta_bmi_smoker",
"beta_bmi_region" ,
"beta_children_smoker",
"beta_children_region",
"beta_smoker_region" ,
"beta_age_squared" ,
"beta_bmi_squared" ,
"beta_children_squared" ,
"beta_smoker_squared",
"beta_region_squared" ,
"tau"),
                   model.file = textConnection(jags_alt_code),
                   n.chains = 3, n.iter = 12000, n.burnin = 2000, n.thin = 10)
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 1003
##    Unobserved stochastic nodes: 22
##    Total graph size: 26655
##
## Initializing model
```

```
jags_alt_model
```

```
## Inference for Bugs model at "6", fit using jags,
##  3 chains, each with 12000 iterations (first 2000 discarded), n.thin = 10
##  n.sims = 3000 iterations saved
##                         mu.vect sd.vect     2.5%      25%      50%      75%
## beta0                    -0.940   1.714   -4.262   -2.099   -0.918    0.198
## beta_age                  0.286   0.055    0.181    0.249    0.285    0.324
## beta_age_bmi              0.013   0.013   -0.014    0.004    0.013    0.021
## beta_age_children        -0.004   0.012   -0.028   -0.012   -0.004    0.004
## beta_age_region           0.009   0.012   -0.014    0.001    0.009    0.017
## beta_age_smoker          -0.001   0.034   -0.068   -0.024   -0.002    0.021
## beta_age_squared          0.056   0.017    0.023    0.045    0.056    0.068
## beta_bmi                 -0.703   0.059   -0.815   -0.742   -0.703   -0.663
```

```
## beta_bmi_children        0.005   0.012   -0.017   -0.002    0.005    0.013
## beta_bmi_region         -0.013   0.013   -0.039   -0.022   -0.013   -0.004
## beta_bmi_smoker          0.741   0.033    0.677    0.718    0.741    0.764
## beta_bmi_squared        -0.017   0.010   -0.037   -0.024   -0.017   -0.010
## beta_children            0.223   0.065    0.098    0.180    0.223    0.267
## beta_children_region    -0.013   0.010   -0.032   -0.019   -0.013   -0.006
## beta_children_smoker    -0.056   0.029   -0.113   -0.075   -0.055   -0.036
## beta_children_squared   -0.014   0.008   -0.030   -0.019   -0.014   -0.009
## beta_region             -0.105   0.080   -0.256   -0.158   -0.105   -0.051
## beta_region_squared      0.017   0.013   -0.009    0.008    0.017    0.026
## beta_smoker             -0.387   2.568   -5.260   -2.088   -0.427    1.328
## beta_smoker_region       0.000   0.030   -0.061   -0.020    0.000    0.020
## beta_smoker_squared      0.828   0.856   -0.851    0.252    0.839    1.403
## tau                      5.896   0.267    5.396    5.711    5.892    6.072
## deviance              1066.711   7.363 1056.053 1061.741 1065.835 1070.765
##                          97.5%  Rhat n.eff
## beta0                    2.313 1.001  3000
## beta_age                 0.394 1.002  1700
## beta_age_bmi             0.039 1.001  3000
## beta_age_children        0.018 1.001  3000
## beta_age_region          0.032 1.001  2100
## beta_age_smoker          0.065 1.001  3000
## beta_age_squared         0.089 1.002  1900
## beta_bmi                -0.586 1.001  2500
## beta_bmi_children        0.028 1.001  3000
## beta_bmi_region          0.013 1.002  2000
## beta_bmi_smoker          0.803 1.001  3000
## beta_bmi_squared         0.003 1.002  3000
## beta_children            0.348 1.001  2500
## beta_children_region     0.007 1.001  2800
## beta_children_smoker     0.000 1.001  2600
## beta_children_squared    0.001 1.002  1200
## beta_region              0.050 1.002  1800
## beta_region_squared      0.043 1.002  1500
## beta_smoker              4.629 1.001  3000
## beta_smoker_region       0.058 1.003   910
## beta_smoker_squared      2.458 1.001  3000
## tau                      6.437 1.002  1100
## deviance              1081.745 1.001  3000
##
## For each parameter, n.eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
##
## DIC info (using the rule, pD = var(deviance)/2)
## pD = 27.1 and DIC = 1093.8
## DIC is an estimate of expected predictive error (lower deviance is better).
```

We already see a huge improvement in the DIC (1527 to 1093), but not necessarily the n.eff.. Now let's remove the statistically insignificant parameters. These are the parameters that have 0 in their credible interval.

```
jags_alt_data<- list(
  N = nrow(train_pol),
  charges = train_pol$charges,
  age = train_pol$age,
```

```
  bmi = train_pol$bmi,
  children = train_pol$children,
  region = train_pol$region,
  age_bmi=train_pol$age_bmi,
  age_children=train_pol$age_children,
  age_region=train_pol$age_region,
  bmi_smoker = train_pol$bmi_smoker,
  bmi_region = train_pol$bmi_region,
  children_smoker = train_pol$children_smoker,
  children_region = train_pol$children_region,
  smoker_region = train_pol$smoker_region,
  age_squared = train_pol$age_squared,
  bmi_squared = train_pol$bmi_squared,
  smoker_squared = train_pol$smoker_squared,
  region_squared = train_pol$region_squared)

jags_alt_code <- "
model {
#priors
beta0 ~ dnorm(0,0.1)
beta_age ~ dnorm(0,0.1)
beta_bmi ~ dnorm(0,0.1)
beta_children ~ dnorm(0,0.1)
beta_region ~ dnorm(0,0.1)
beta_age_bmi ~ dnorm(0,0.1)
beta_age_children ~ dnorm(0,0.1)
beta_age_region ~ dnorm(0,0.1)
beta_bmi_smoker ~ dnorm(0,0.1)
beta_bmi_region ~ dnorm(0,0.1)
beta_children_smoker ~ dnorm(0,0.1)
beta_children_region ~ dnorm(0,0.1)
beta_smoker_region ~ dnorm(0,0.1)
beta_age_squared ~ dnorm(0,0.1)
beta_bmi_squared ~ dnorm(0,0.1)
beta_smoker_squared ~ dnorm(0,0.1)
beta_region_squared ~ dnorm(0,0.1)









tau ~ dgamma(0.1, 0.1)

#likelihood
for (i in 1:N) {
    mu[i] <- beta0 +beta_age * age[i] +beta_bmi * bmi[i] + beta_children * children[i] +  beta_region *


    charges[i] ~ dnorm(mu[i], tau)
  }
```

```
}
"
jags_alt_model <- jags(data = jags_alt_data, inits = NULL,
                   parameters.to.save = c("beta0", "beta_age","beta_bmi","beta_children","beta_region",
"beta_age_region" ,
"beta_bmi_smoker",
"beta_bmi_region" ,
"beta_children_smoker",
"beta_children_region",
"beta_smoker_region" ,
"beta_age_squared" ,
"beta_bmi_squared" ,
"beta_smoker_squared",
"beta_region_squared" ,
"tau"),
                   model.file = textConnection(jags_alt_code),
                   n.chains = 3, n.iter = 12000, n.burnin = 2000, n.thin = 10)
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 1003
##    Unobserved stochastic nodes: 18
##    Total graph size: 21747
##
## Initializing model
```

```
jags_alt_model
```

```
## Inference for Bugs model at "4", fit using jags,
##  3 chains, each with 12000 iterations (first 2000 discarded), n.thin = 10
##  n.sims = 3000 iterations saved
##                         mu.vect sd.vect     2.5%      25%      50%      75%
## beta0                    -1.113   0.107   -1.322   -1.184   -1.115   -1.042
## beta_age                  0.284   0.039    0.207    0.257    0.283    0.309
## beta_age_bmi              0.014   0.014   -0.012    0.005    0.014    0.024
## beta_age_children        -0.002   0.012   -0.025   -0.010   -0.002    0.006
## beta_age_region           0.009   0.012   -0.016    0.000    0.009    0.017
## beta_age_squared          0.050   0.017    0.018    0.039    0.050    0.061
## beta_bmi                 -0.696   0.054   -0.806   -0.732   -0.697   -0.659
## beta_bmi_region          -0.013   0.013   -0.038   -0.022   -0.013   -0.004
## beta_bmi_smoker           0.743   0.033    0.679    0.722    0.743    0.766
## beta_bmi_squared         -0.017   0.010   -0.036   -0.023   -0.017   -0.010
## beta_children             0.137   0.041    0.056    0.110    0.137    0.164
## beta_children_region     -0.013   0.010   -0.032   -0.019   -0.013   -0.006
## beta_children_smoker     -0.048   0.028   -0.103   -0.066   -0.048   -0.030
## beta_region              -0.101   0.078   -0.255   -0.154   -0.100   -0.047
## beta_region_squared       0.016   0.013   -0.010    0.008    0.016    0.025
## beta_smoker_region       -0.002   0.030   -0.062   -0.023   -0.002    0.018
```

```
## beta_smoker_squared     0.696   0.033    0.632    0.673    0.697    0.719
## tau                     5.889   0.266    5.374    5.704    5.886    6.061
## deviance             1067.043   6.316 1056.934 1062.759 1066.295 1070.605
##                         97.5%  Rhat n.eff
## beta0                  -0.908 1.003   970
## beta_age                0.362 1.002  1500
## beta_age_bmi            0.042 1.001  2800
## beta_age_children       0.020 1.001  3000
## beta_age_region         0.032 1.002  1300
## beta_age_squared        0.083 1.003  1600
## beta_bmi               -0.590 1.001  3000
## beta_bmi_region         0.013 1.001  3000
## beta_bmi_smoker         0.808 1.001  3000
## beta_bmi_squared        0.002 1.001  3000
## beta_children           0.218 1.003  1300
## beta_children_region    0.006 1.002  1300
## beta_children_smoker    0.007 1.002  1800
## beta_region             0.054 1.002  1300
## beta_region_squared     0.042 1.002  1600
## beta_smoker_region      0.057 1.001  3000
## beta_smoker_squared     0.760 1.002  1400
## tau                     6.428 1.001  3000
## deviance             1080.486 1.001  3000
##
## For each parameter, n.eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
##
## DIC info (using the rule, pD = var(deviance)/2)
## pD = 20.0 and DIC = 1087.0
## DIC is an estimate of expected predictive error (lower deviance is better).
```

Looks like this model is even better as it has better DIC(1087 in comparison to 1093), comparable n.eff. and $\hat{R}$, and tighter intervals in certain parameter.

Predictive Accuracy

```r
# Extract posterior samples for each parameter
set.seed(1245)
beta0_samples <- jags_model$BUGSoutput$sims.list$beta0
beta_age_samples <- jags_model$BUGSoutput$sims.list$beta_age
beta_bmi_samples <- jags_model$BUGSoutput$sims.list$beta_bmi
beta_children_samples <- jags_model$BUGSoutput$sims.list$beta_children
beta_smoker_samples <- jags_model$BUGSoutput$sims.list$beta_smoker
beta_region_samples <- jags_model$BUGSoutput$sims.list$beta_region

# Prepare the test data
test_data <- data.frame(
  age = test$age,
  bmi = test$bmi,
  children = test$children,
  smoker = test$smoker,
  region = test$region
)
```

```r
# Number of posterior samples and number of test observations
n_samples <- length(beta0_samples)
n_test <- nrow(test_data)

# Initialize matrix to store predictions
predictions <- matrix(NA, nrow = n_samples, ncol = n_test)

# Loop through each test observation and compute predictions for each posterior sample
for (j in 1:n_test) {
  mu <- beta0_samples +
        beta_age_samples * test_data$age[j] +
        beta_bmi_samples * test_data$bmi[j] +
        beta_children_samples * test_data$children[j] +
        beta_smoker_samples * test_data$smoker[j] +
        beta_region_samples * test_data$region[j]

  # Store predictions from the normal distribution with posterior tau
  predictions[, j] <- rnorm(n_samples, mu, sqrt(1 / jags_model$BUGSoutput$sims.list$tau))
}

# Calculate the mean prediction for each test observation
predicted_means <- apply(predictions, 2, mean)

# Output predicted means
test$bayes <- predicted_means
rmse_value <- rmse(test$charges, test$bayes)
paste("RMSE:", rmse_value)
```

Initial model

```
## [1] "RMSE: 0.451565037895899"
```

polynomial features model

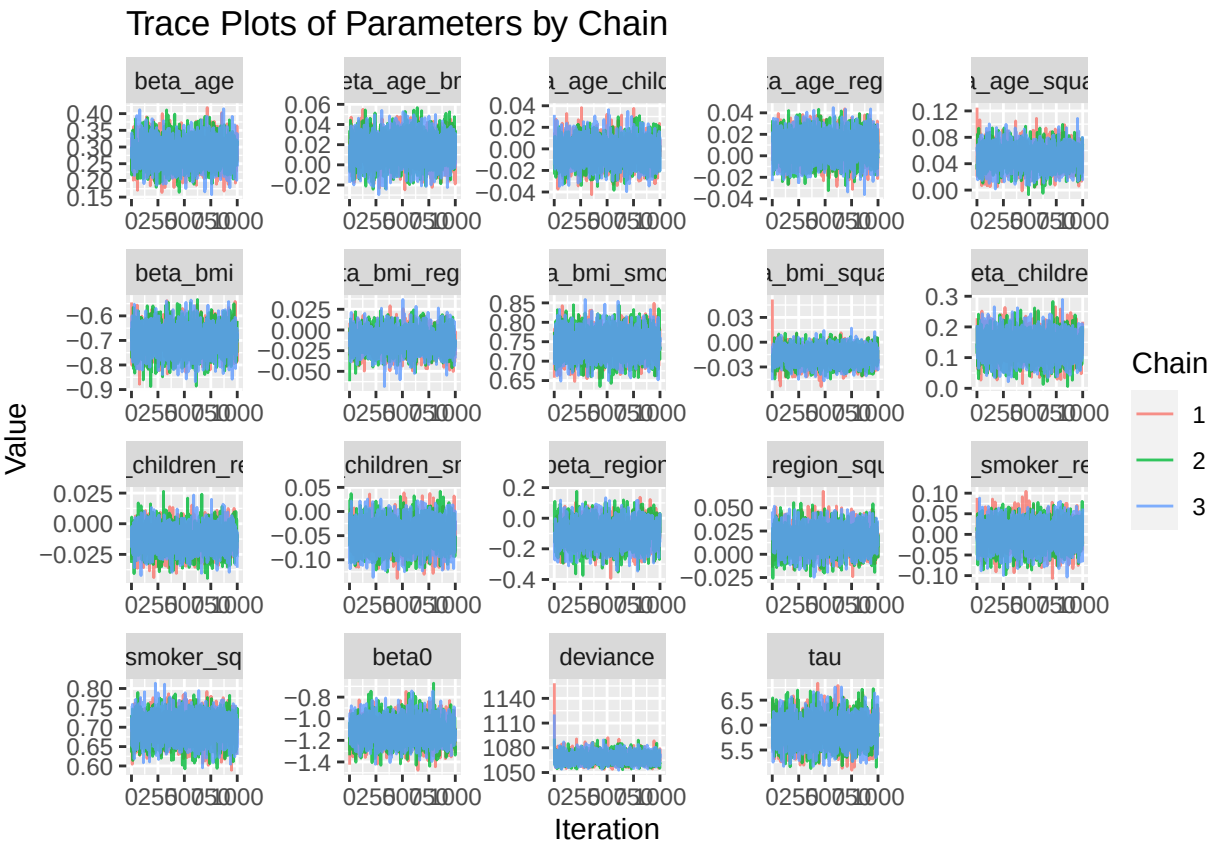
```
## [1] "RMSE: 0.347201479701525"
```

As we can see, the alternative model performs better even when it comes to the predictive accuracy as the RMSE dropped from 0.45 to 0.35.

Model Checking diagnostics and discussion

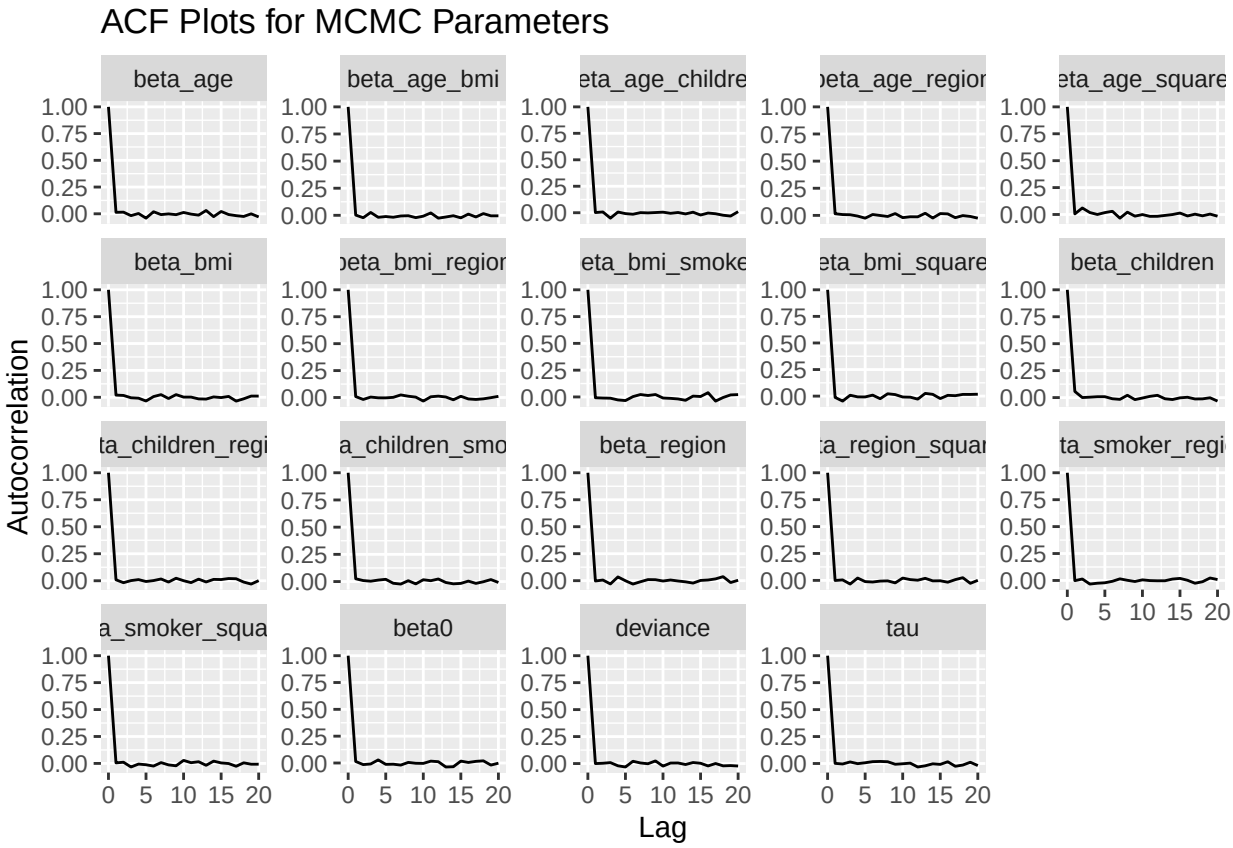
Since we will go with the model with polynomial features, we shall run some diagnostics to evaluate the reliability and validity of the estimates obtained from the JAGS model.

Trace Plots Trace plots are used to visually assess the sampling behavior of each parameter in a Bayesian model. They show how the sampled values of a parameter evolve over the iterations of the MCMC algorithm, helping to check if the chain has converged and is mixing well. What we are looking for is random fluctuation around a stable mean value and having good mixing, which is indicated by frequent transitions across the parameter space, without long periods where the chain gets stuck in one region. We should also be looking for the convergence of all 3 chains, which is indicated by the overlapping of the chains after some iterations.



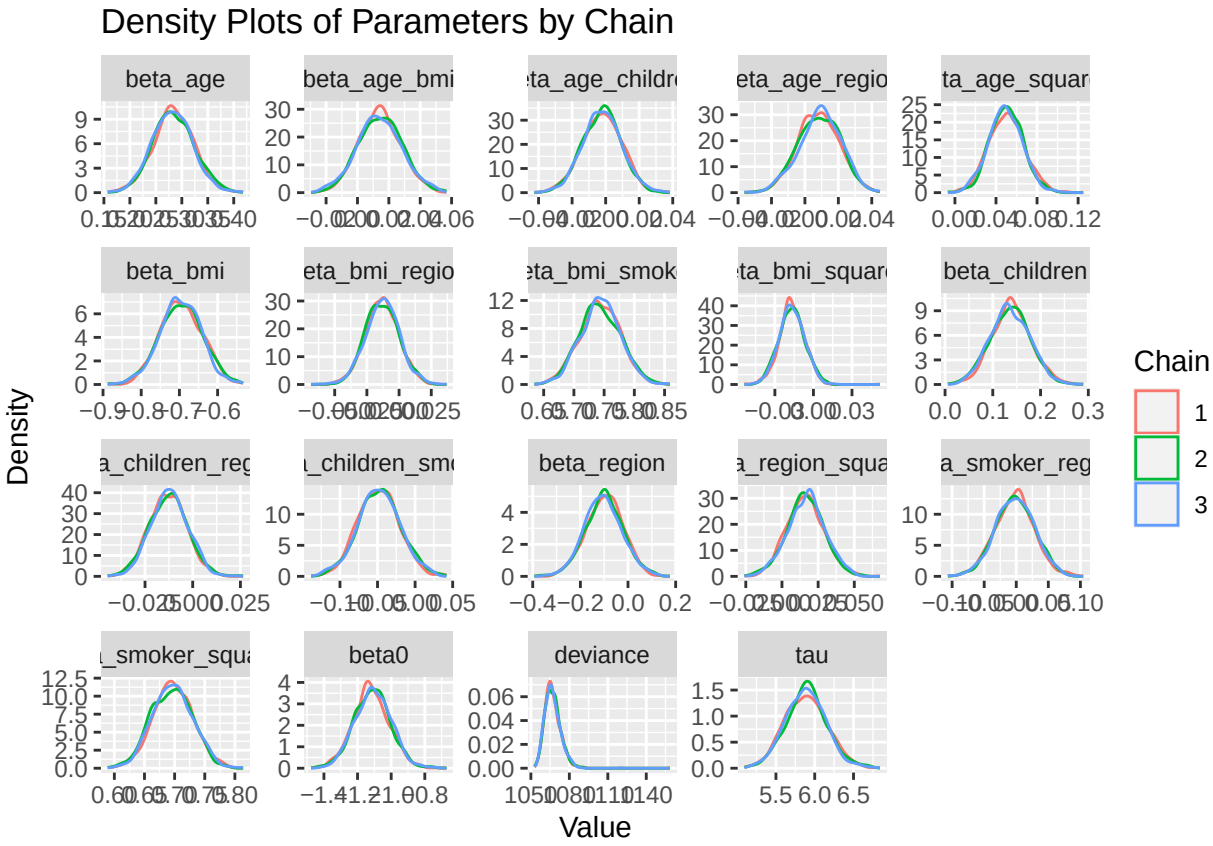
In our trace plots we see that the values fluctuate randomly around a central value, suggesting that the MCMC process has reached a stable posterior distribution. We also see that there is no upward or downward trend. Both of these aspects show convergence. We also see that the chains are mixing and overlapping.

ACF plot An ACF plot is used to analyze how a variable's values are correlated with its own past values at various time lags and it helps assess how well the sampling algorithm is performing and whether the samples in the chain are independent of each other.



As all plots shown above have a drop in autocorrelation to a value close to 0 very quickly, this shows that the chain is moving independently from one iteration to the next, indicating good mixing.

Density plot A density plot shows the estimated probability density function of the posterior distributions of the parameters sampled from the model. What we need to check is the symmetry of the density and if the chains overlap.



As seen by the density plots above, the densities for the parameters are all symmetric. Symmetry implies that the parameter estimates are stable and not biased in one direction. We can also see that the chains are overlapping, this is a strong indicator that our MCMC sampling has reached a stationary distribution.

All diagnostic graphs shows that our model is reliable, is performing well, and that the MCMC sampling has converged effectively. This is also backed up by our good predictions as shown by the low mean residual obtained.

Findings and Conclusion

In conclusion, we find that the variables we have give us a strong indicator of what an individual's charge is going to be. With some of the most significant ones being unsurprisingly the ones that affect one's health (Age, smoking, BMI). We also find that both the Frequentist and Bayesian approaches provide us with good predictive models for this problem,with models that utilize polynomial features performing better than the ones without them as they help us capture the non-linear relationships between the dependent variables. This was evaluated by the RMSE of the prediction in the frequentist model's case and the DIC, $\hat{R}$, n.eff. and the RMSE in the Bayesian model's case.