EDMOND GERAUD

CAPSTONE PROJECT: COMPARISON OF FREQUENTIST AND BAYESIAN APPROACHES ON THE IRIS DATA SET

In the realm of statistics and data analysis, two fundamental frameworks for modeling relationships between variables are Bayesian and frequentist statistics. To illustrate their differences, let's consider a linear regression analysis of the Iris dataset, specifically examining the relationship between Petal Length (independent variable) and Petal Width (dependent variable).

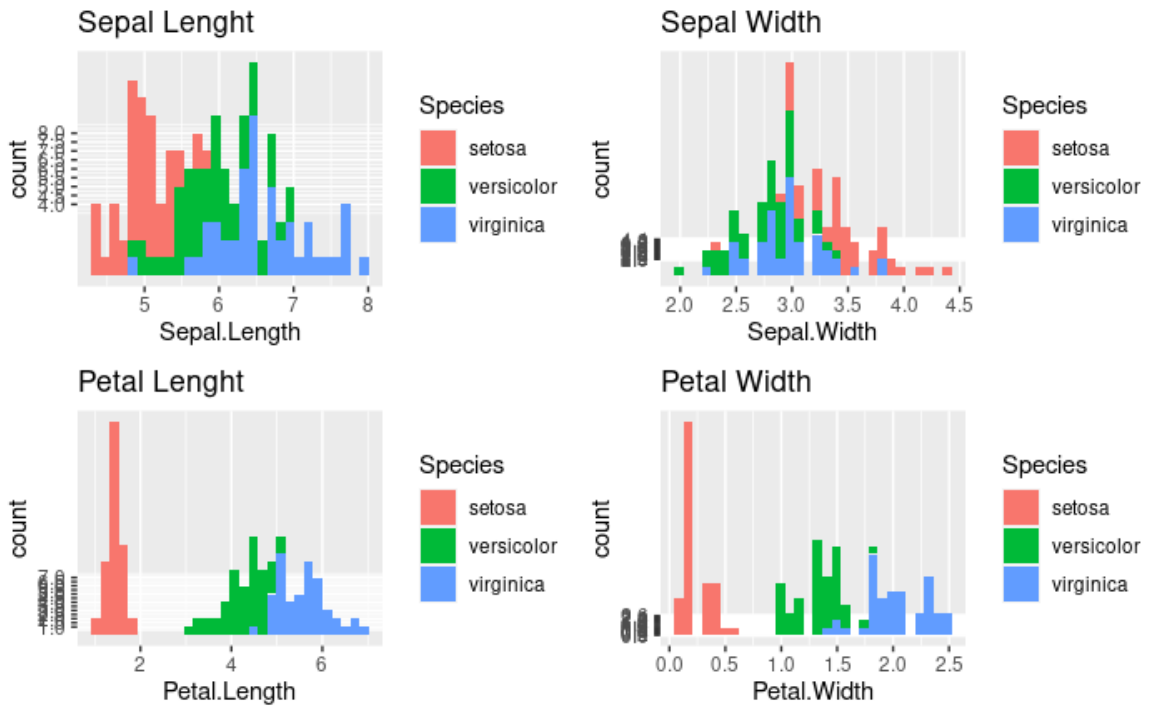
The results show no differences in the frequentist and bayesian framework. This is the main idea of the **capstone project** that I present here. A simple linear regression was made for the frequentist approach and a bayesian one. The model assumptions accomplishes for both methods and similar results were obtained. It is easier to perform a simple linear regression, however, sometimes the assumptions are not made, which is not the case. A bayessian approach is better and more understandable. On the other hand, it is complicated to perform with **rjags.** Though **jags** was used, it was compared with another library called **bayestestR, which is easier to implement.**

INTRODUCTION

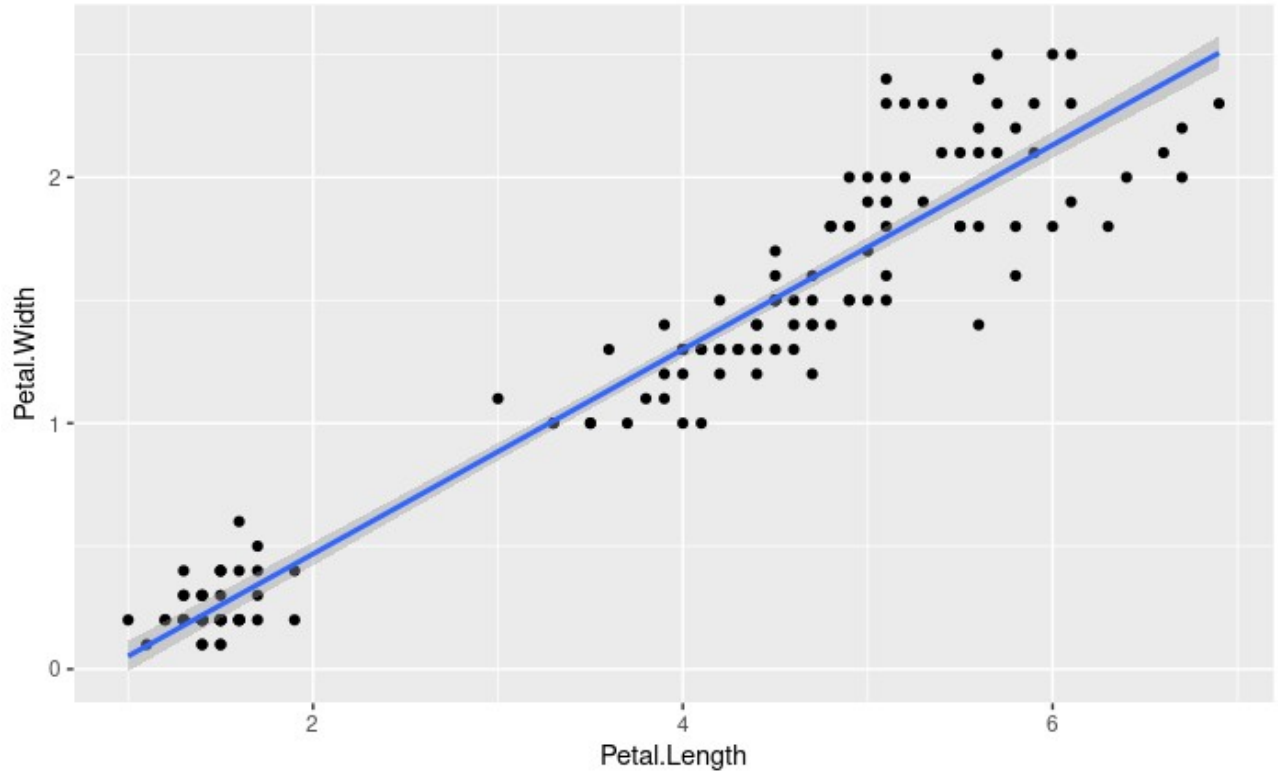
The iris data set was used. It is a well known dataset for data scientists. My itention is to perform a bayesian linear regression between the petal length and width. The data were collected by Anderson and Edgar in 1935. There was no need to clean the data, because it is a toy dataset. The main idea is to consolidate the knowledge obtained in the course.

DATA

The data consists of 150 observations and 5 variables. Four of them are numerical and Species is categorical. All numerical variables appear to be normally distributed as shown in the histograms.



The Petal width is positively correlated with the Petal Length as shown in the next figure.



There is a cluster on the bottom left which is from a particular species. However, here we are not interested in which species are. We want to know the relationship in general of the length and width of the plants.

THE MODEL

The model used was a hierarchical bayesian linear regression. The response variable follos a normal distribution with parameters mu and tau. Were the mean is given by the linear model. The coeficients of the model are set as normal priors and the variance is taken as an inverse gamma distribution.

```
model{

 for(i in 1:n){
  y[i] ~ dnorm(mu[i], tau.sq)
 }

 mu <- X%*%beta

 for(i in 1:p){
  beta[i] ~ dnorm(0, 0.000001)
 }
```

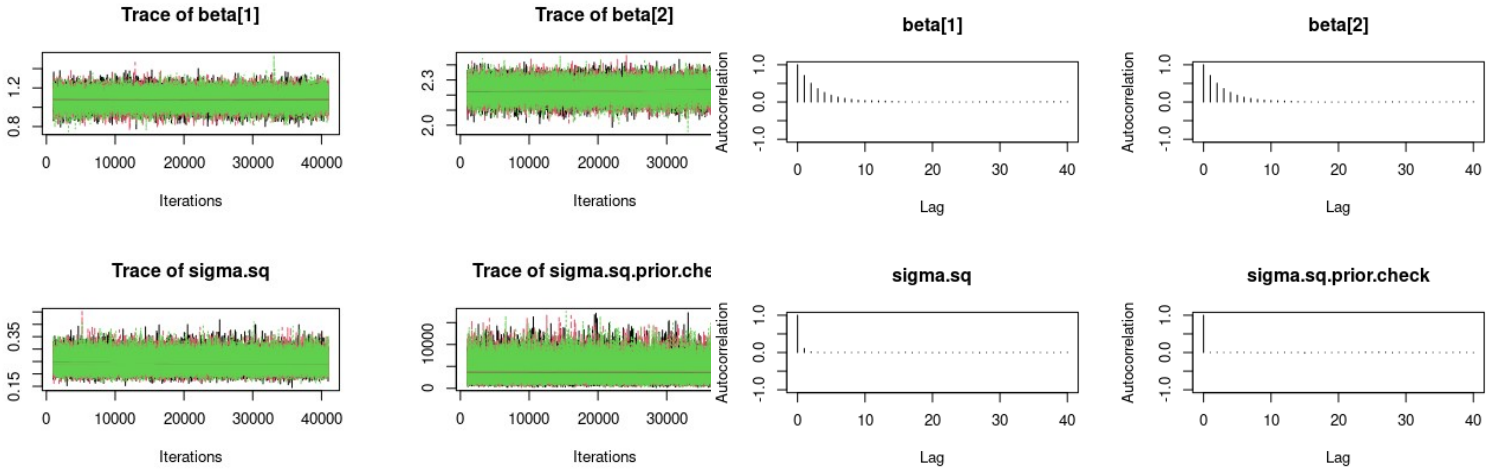
```
  tau.sq <- 1/sigma.sq
  sigma.sq ~ dgamma(4, 0.001)

  sigma.sq.prior.check ~ dgamma(4, 0.001)
 }
```

The model is well suited for the question of interest due to the fact that all variables follow a symetrical distributions. The priors are normal for the mean and inverse gamma for the variance. The justification of the priors came from the bayesian normal linear model

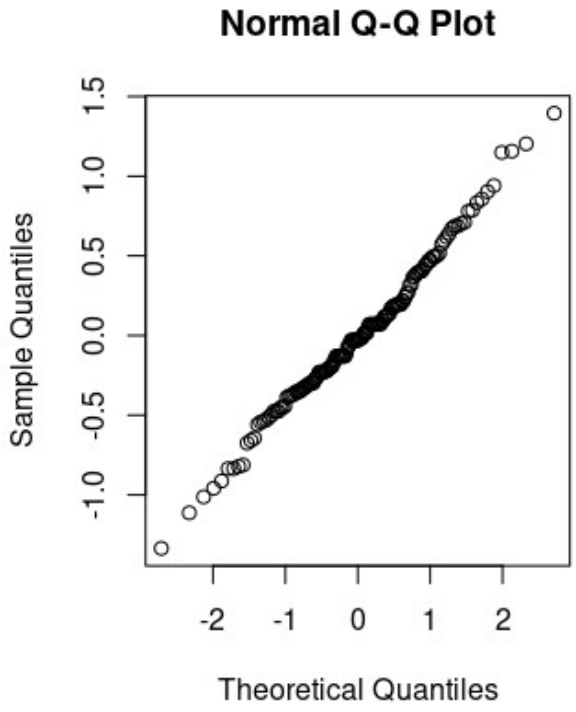
CHECK THE ASSUMPTIONS AND THE MODEL.

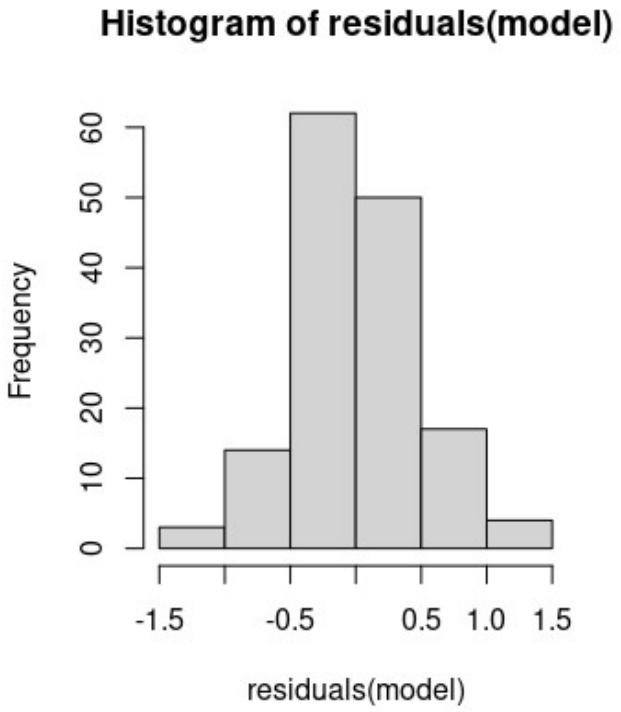
The trace plot look like they converge rapidly. Moreover the autocorrelation plots show no major implications. And



regarding the model, the residuals are normally distributed. This is telling us that the model is well suited for our data.

EDMOND GERAUD





RESULTS

The correlation between the length and width of the petal is the same as the frequentist approach with 96%. The bayes factor is above >100, and the probability of direction is 100%. Indicating that the correlation is statistically significant.

CONCLUSIONS

Because we are playing with a very known dataset, the aim of this project was to obtain the same results in the frequentist and bayes framework. The results are satisfying.