Project: Bayesian Linear Regression

Borja Ruiz Amantegui 100357358, Francesca Sallicati 100341618

March 7th 2018

Contents

1 Introduction

In order to work to realize this project we chose a kaggle dataset with 65 world indicators and simplified it by selecting one dependent variable and four other independent variables. The data was gathered manually for most of it at World Bank, Unicef and so on. Some data were not there so K-nnwas used to create some values and have a full dataset that can be used by data science community.

$$y = Genderinequality$$

$$x = AdoBirth100, Fsuicide100, Partnerviolence, fPoliticianPerc$$

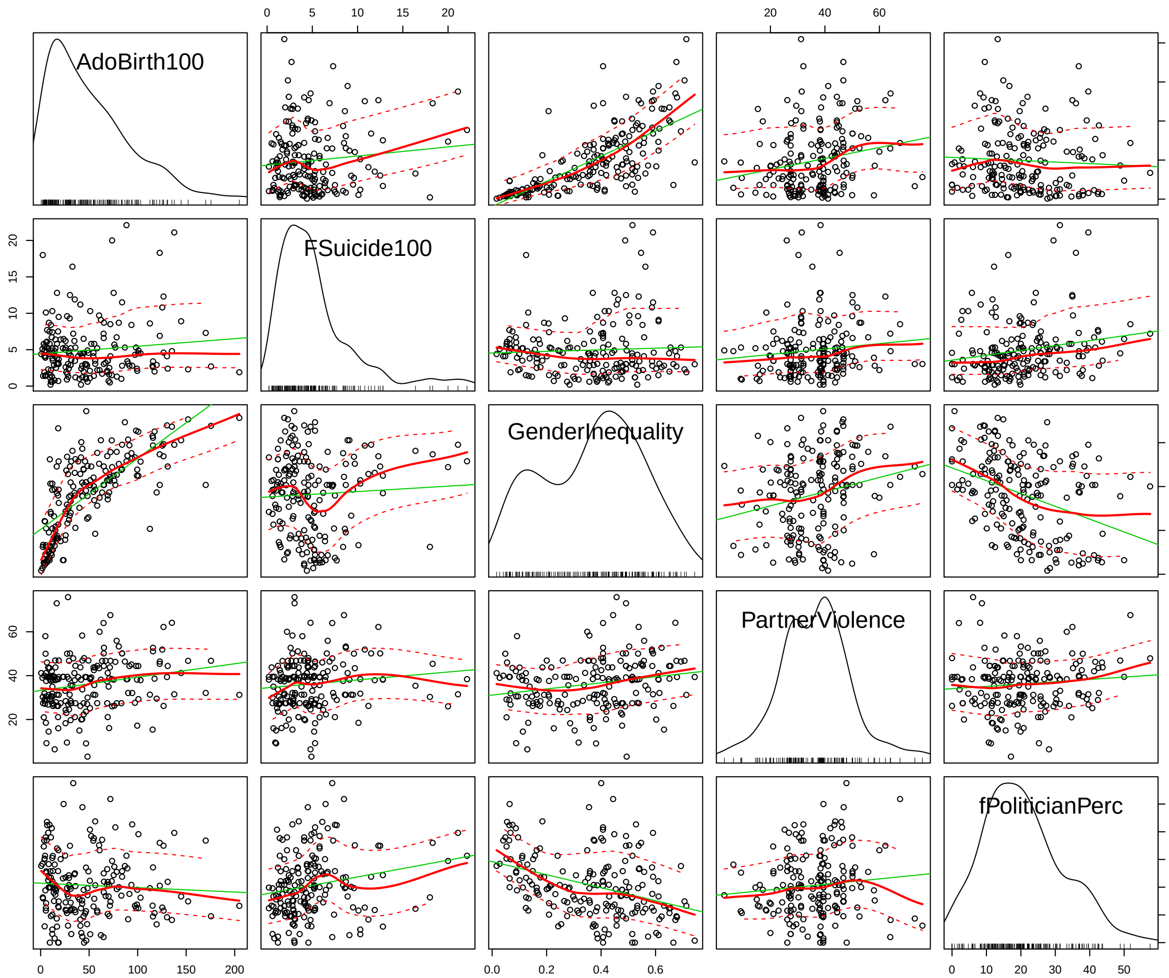
2 Variable description

1. AdoBirth100: birth rate per hundreds of women between 15 and 19 years of age.
2. Gender inequality: index for gender inequality.
3. Fsuicide100: female suicide rate measured in hundreds.
4. Partner violence: Intimate or nonintimate partner violence ever experienced
5. fPoliticianPerc: percentage of female representation in the parliament

```
summary(Kaggle)
```

```
##   AdoBirth100       FSuicide100     GenderInequality  PartnerViolence
##  Min.   :  0.617   Min.   : 0.200   Min.   :0.01642   Min.   : 3.07  
##  1st Qu.: 15.177   1st Qu.: 2.400   1st Qu.:0.20843   1st Qu.:27.43  
##  Median : 40.967   Median : 4.100   Median :0.38669   Median :38.17  
##  Mean   : 49.324   Mean   : 4.947   Mean   :0.36448   Mean   :36.23  
##  3rd Qu.: 71.516   3rd Qu.: 6.125   3rd Qu.:0.50354   3rd Qu.:43.77  
##  Max.   :204.789   Max.   :22.100   Max.   :0.74396   Max.   :75.80  
##  fPoliticianPerc
##  Min.   : 0.00  
##  1st Qu.:12.35  
##  Median :19.63  
##  Mean   :20.66  
##  3rd Qu.:27.20  
##  Max.   :57.55  
```

```
scatterplotMatrix(Kaggle)
```



From the graph we can see that there seems to be a strong linear relationship among Adolescent Birth Rate and Gender Inequality and between Gender Inequality and Female Politician percentage.

3 Frequentist approach

We decided to predict Gender Inequality and we started our analysis by the regressing over the full model:

```
freq.reg1 <- lm(GenderInequality ~ ., data = Kaggle)
summary(freq.reg1)
```

```
##
## Call:
## lm(formula = GenderInequality ~ ., data = Kaggle)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35381 -0.08215 -0.00923  0.07180  0.30084
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)      0.2632773  0.0293509   8.970 3.48e-16 ***
## AdoBirth100      0.0030376  0.0002043  14.869  < 2e-16 ***
## FSuicide100      0.0017172  0.0021634   0.794   0.4284
## PartnerViolence  0.0015767  0.0006953   2.268   0.0245 *
## fPoliticianPerc -0.0055302  0.0007188  -7.694 8.56e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1106 on 183 degrees of freedom
## Multiple R-squared:  0.6428, Adjusted R-squared:  0.635
## F-statistic: 82.35 on 4 and 183 DF,  p-value: < 2.2e-16
```

From which we can see that the variable Female Suicide Rate is not significant, so we discard it from the model.

```
freq.reg2 <- lm(GenderInequality ~ .-FSuicide100, data = Kaggle)
summary(freq.reg2)
```

```
##
## Call:
## lm(formula = GenderInequality ~ . - FSuicide100, data = Kaggle)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34832 -0.08003 -0.01375  0.07691  0.30086
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.2671008  0.0289238   9.235  < 2e-16 ***
## AdoBirth100      0.0030548  0.0002029  15.053  < 2e-16 ***
## PartnerViolence  0.0016170  0.0006927   2.334   0.0207 *
## fPoliticianPerc -0.0054159  0.0007035  -7.698 8.17e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1105 on 184 degrees of freedom
## Multiple R-squared:  0.6416, Adjusted R-squared:  0.6358
## F-statistic: 109.8 on 3 and 184 DF,  p-value: < 2.2e-16
```

It is a good choice since the adjusted R^2 is slightly increasing without the useless predictor.

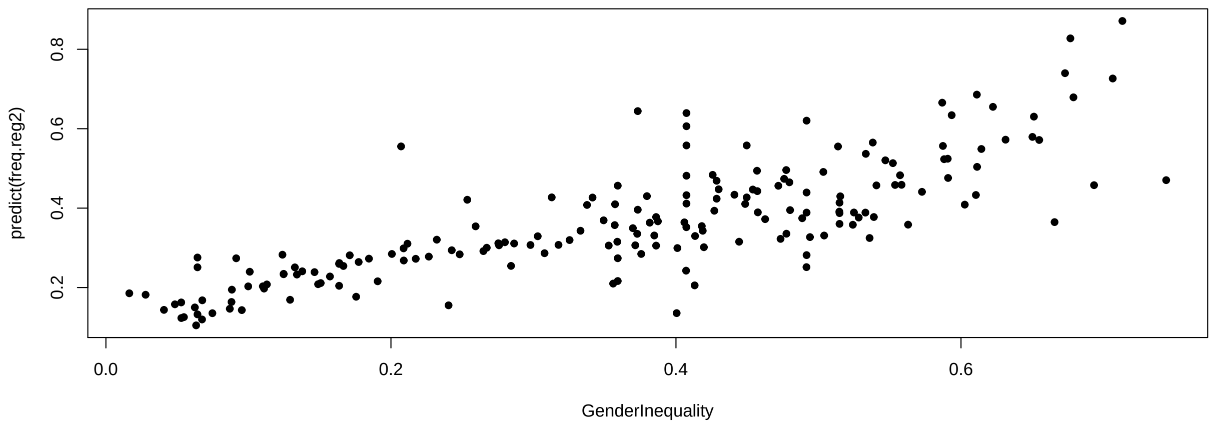
Let's check the confidence intervals for the coefficients:

```
confint(freq.reg2,level=.95)
```

```
##                         2.5 %       97.5 %
## (Intercept)      0.2100357481  0.324165787
## AdoBirth100      0.0026544507  0.003455198
## PartnerViolence  0.0002503047  0.002983742
## fPoliticianPerc -0.0068038768 -0.004027926
```

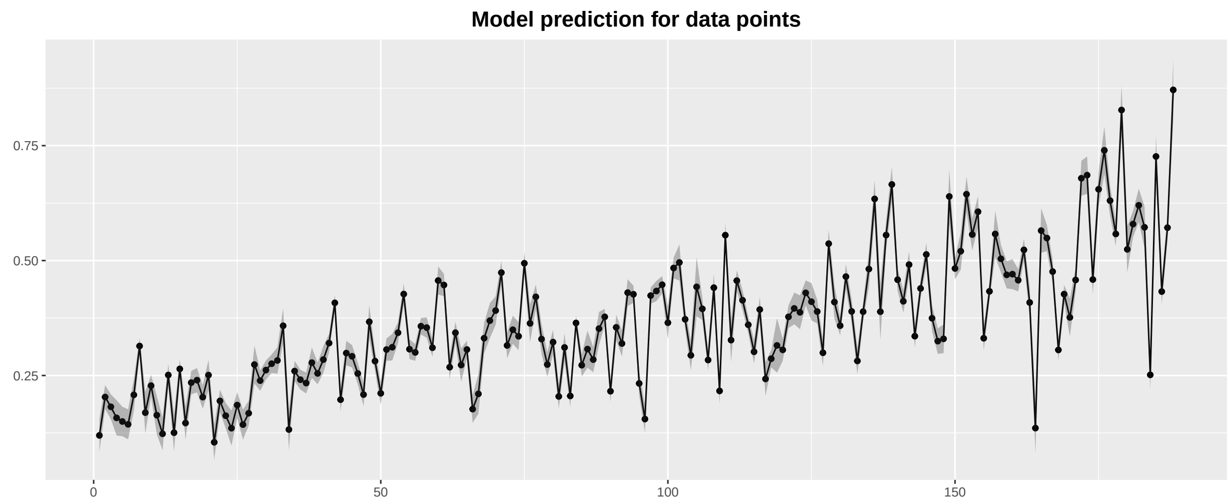
Since none of them contains 0 the estimates of the β_i coefficients are significatively different from 0.

```
plot(GenderInequality, predict(freq.reg2),pch=16)
```



Let's visualize the two kinds of prediction intervals, that is to say inference the conditional means of Y given $X = x, E[Y|X = x]$:

```
p<-predict(freq.reg2,interval="confidence")
p<-as.data.frame(p)
ggplot(p, aes(c(1:188),p$fit))+geom_point()+geom_line()+geom_ribbon(data=p,aes(ymin=lwr,ymax=upr),alpha
```

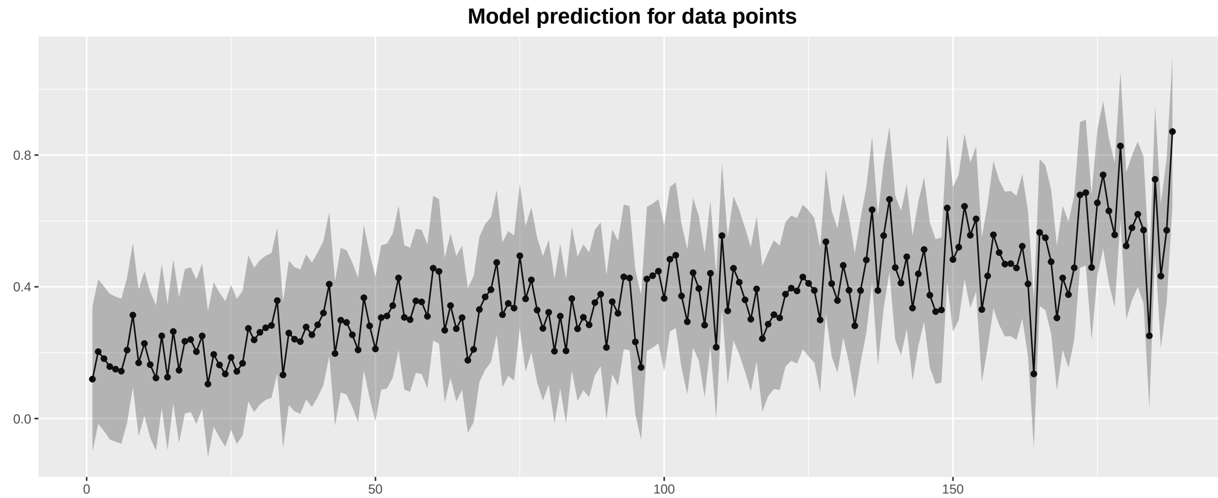


and the conditional response $Y|X = x$:

```
p<-predict(freq.reg2,interval="prediction")
```

```
## Warning in predict.lm(freq.reg2, interval = "prediction"): predictions on current data refer to _futu
```

```
p<-as.data.frame(p)
ggplot(p, aes(c(1:188),p$fit))+geom_point()+geom_line()+geom_ribbon(data=p,aes(ymin=lwr,ymax=upr),alpha
```



4 Bayesian approach

With the Bayesian approach we start again by studing the full model:

```r
summary(bayes.reg1)
```

```
##
##  Iterations = 3001:12991
##  Thinning interval  = 10
##  Sample size  = 1000
##
##  DIC: -287.4239
##
##  R-structure:  ~units
##
##       post.mean l-95% CI u-95% CI eff.samp
## units   0.01242 0.009811  0.01497     1057
##
##  Location effects: GenderInequality ~ AdoBirth100 + FSuicide100 + PartnerViolence + fPoliticianPerc
##
##                  post.mean  l-95% CI  u-95% CI eff.samp  pMCMC
## (Intercept)       0.262816  0.205553  0.323773     1000 <0.001 ***
## AdoBirth100       0.003028  0.002607  0.003421     1000 <0.001 ***
## FSuicide100       0.001885 -0.002650  0.006097     1000  0.378
## PartnerViolence   0.001576  0.000127  0.002910     1000  0.036 *
## fPoliticianPerc  -0.005527 -0.007009 -0.004138     1000 <0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Once again the Female Suicide Rate is not statistically significant so let's delete it:

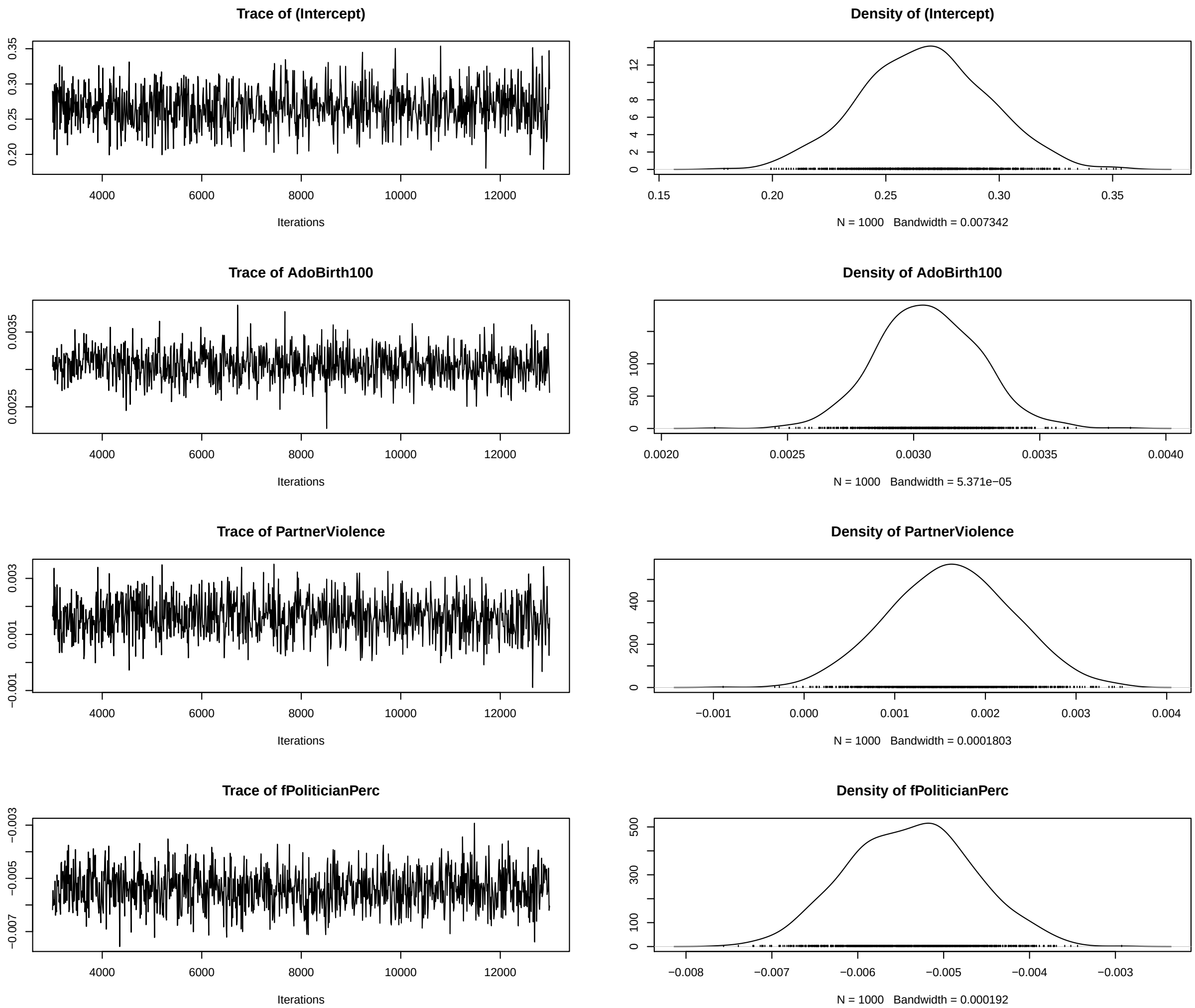
```r
summary(bayes.reg2)
```

```
##
##  Iterations = 3001:12991
##  Thinning interval  = 10
##  Sample size  = 1000
```

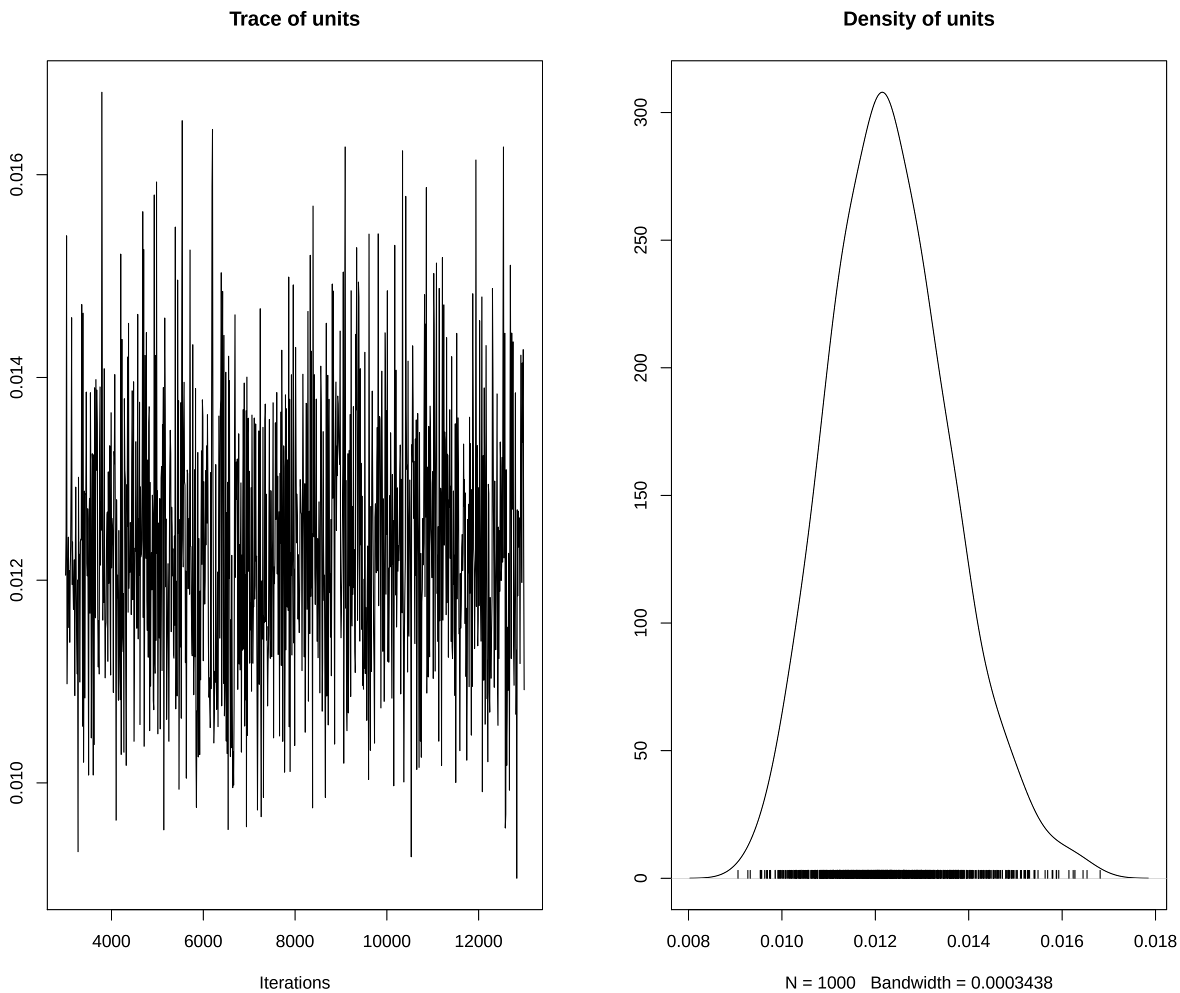
```
##
##  DIC: -288.932
##
##  R-structure:  ~units
##
##       post.mean l-95% CI u-95% CI eff.samp
## units   0.01234 0.009928  0.01492     1000
##
##  Location effects: GenderInequality ~ AdoBirth100 + PartnerViolence + fPoliticianPerc
##
##                  post.mean   l-95% CI   u-95% CI eff.samp  pMCMC
## (Intercept)      0.2669580  0.2125940  0.3201701     1000 <0.001 ***
## AdoBirth100      0.0030557  0.0026566  0.0034409     1000 <0.001 ***
## PartnerViolence  0.0016186  0.0003022  0.0029078     1000  0.014 *
## fPoliticianPerc -0.0053861 -0.0066854 -0.0039363     1000 <0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Without that variables Partner Violence is more significant within the new subset.

```r
plot(bayes.reg2)
```



Trace of (Intercept)
Iterations
Density of (Intercept)
N = 1000 Bandwidth = 0.007342
Trace of AdoBirth100
Iterations
Density of AdoBirth100
N = 1000 Bandwidth = 5.371e−05
Trace of PartnerViolence
Iterations
Density of PartnerViolence
N = 1000 Bandwidth = 0.0001803
Trace of fPoliticianPerc
Iterations
Density of fPoliticianPerc
N = 1000 Bandwidth = 0.000192



Let's check the empirical mean of the β coefficients:

```
beta=bayes.reg2$Sol
colMeans(beta)
```

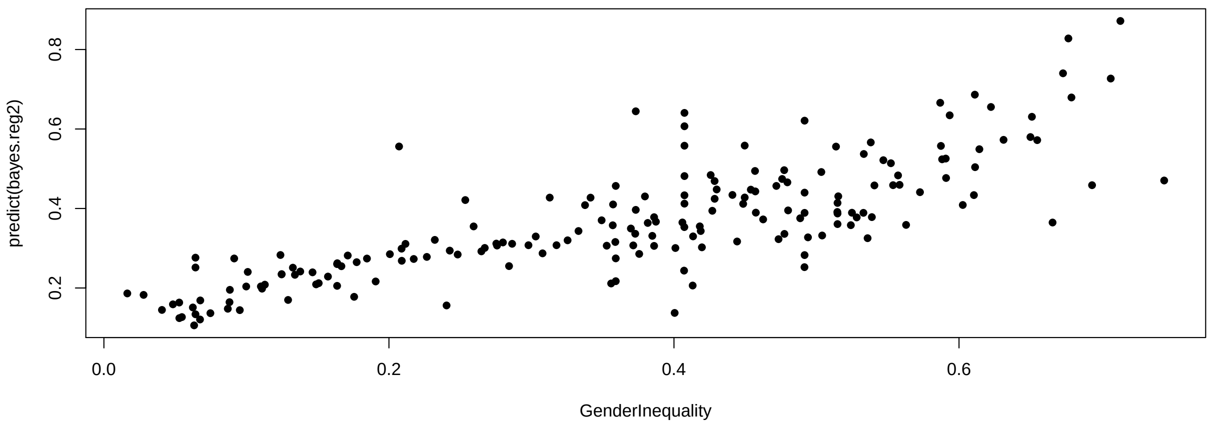
```
##     (Intercept)     AdoBirth100 PartnerViolence fPoliticianPerc
##     0.266957971     0.003055703     0.001618626    -0.005386089
```

and that their 95 credible intervals do no contain 0.

```
HPDinterval(beta,.95)
```

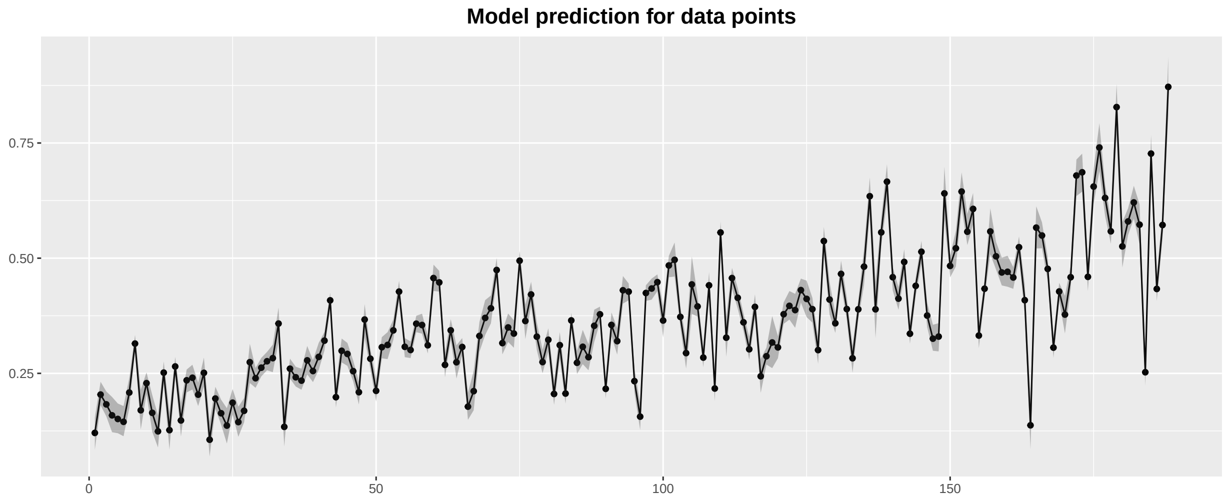
```
##                          lower        upper
## (Intercept)      0.2125940332  0.320170100
## AdoBirth100      0.0026565813  0.003440905
## PartnerViolence  0.0003022287  0.002907773
## fPoliticianPerc -0.0066854279 -0.003936320
## attr(,"Probability")
## [1] 0.95
```

```
plot(GenderInequality, predict(bayes.reg2),pch=16)
```

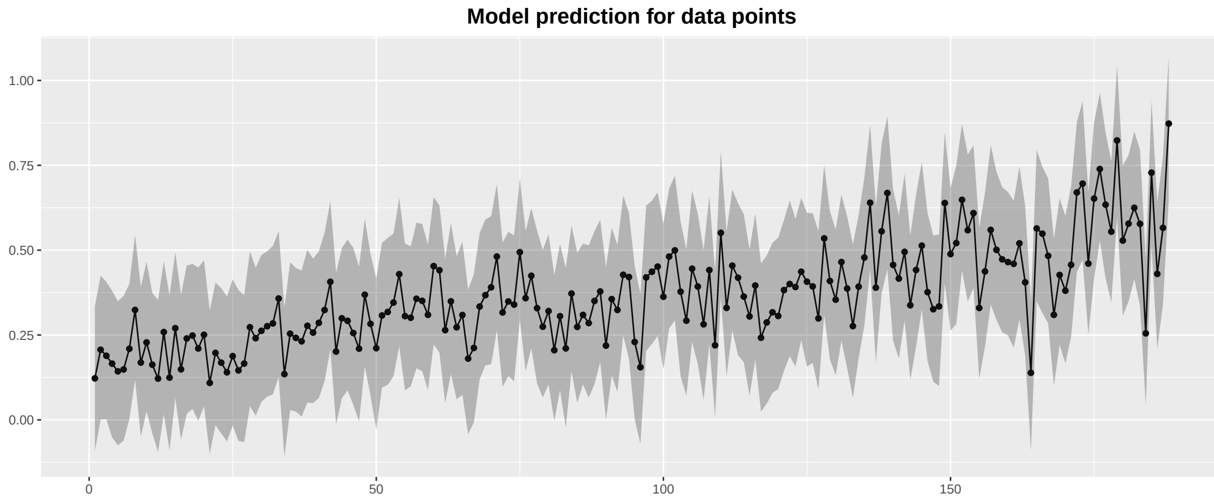


Let's visualize the two kinds of prediction intervals:

```
p<-predict(bayes.reg2,interval="confidence")
p<-as.data.frame(p)
ggplot(p, aes(c(1:188),p$fit))+geom_point()+geom_line()+geom_ribbon(data=p,aes(ymin=lwr,ymax=upr),alpha
```




```
p<-predict(bayes.reg2,interval="prediction")
p<-as.data.frame(p)
ggplot(p, aes(c(1:188),p$fit))+geom_point()+geom_line()+geom_ribbon(data=p,aes(ymin=lwr,ymax=upr),alpha
```



5 Conclusion

In this project we have applied a frequentist and a bayesian approach in order to predict the gender inequality of each country. After evaluating the models we observe that there are no major differences in our predictions. Therefore we determine that the bayesian approach is also valid for prediction.