Bayesian Leaning: Bayesian Non Linear Regression

Borja Ruiz Amantegui 100357358. Francesca Sallicati 100341618

March 21t 2018

Contents

Introduction

The data chosen to develop this project on Bayesian nonlinear regression come from Kaggle.com and consist in 65 world indicators; to this purpose we have decided to select some variable among the entire set. The Human Development Indicator is the Y chosen as the dependent variables and some other inputs have been selected consequently:

1. *AB*: Adolescent birth rate per a thousand women between 15 and 19 years of age.
2. *Elec*: Electrification Rate percentage population
3. *GDP*: Gross Domestic Product
4. *ForInv*: Foreign direct investment net inflows.percentageof GDP
5. *IM*: Infant Mortality Rate per thousand people
6. *IntUsers*: Internet users percentage of population
7. *Lifexp*: Life Expectancy at birth
8. *MobSubs*: Mobile phone subscription per thousand people

These are summary statistics for our subset of variables:

```r
summary(data)
```

```
##       HDI               AB               CO2                Elec
##  Min.   :0.3483   Min.   :  0.617   Min.   : 0.02191   Min.   :  5.10
##  1st Qu.:0.5738   1st Qu.: 15.177   1st Qu.: 0.62705   1st Qu.: 58.60
##  Median :0.7241   Median : 40.967   Median : 2.37289   Median : 97.70
##  Mean   :0.6924   Mean   : 49.324   Mean   : 4.66650   Mean   : 77.92
##  3rd Qu.:0.8166   3rd Qu.: 71.516   3rd Qu.: 6.39518   3rd Qu.:100.00
##  Max.   :0.9439   Max.   :204.789   Max.   :43.89304   Max.   :100.00
##      ForInv             GDP                 IM             IntUsers
##  Min.   :-19.378   Min.   :   584.4   Min.   :  1.60   Min.   : 0.99
##  1st Qu.:  1.329   1st Qu.:  3347.1   1st Qu.:  7.00   1st Qu.:17.57
##  Median :  2.986   Median : 11019.7   Median : 14.60   Median :43.40
##  Mean   :  4.514   Mean   : 17157.8   Mean   : 25.27   Mean   :44.11
##  3rd Qu.:  4.846   3rd Qu.: 23127.6   3rd Qu.: 39.65   3rd Qu.:68.19
##  Max.   : 50.017   Max.   :127562.2   Max.   :107.20   Max.   :98.16
##      Lifexp         MobSubs
##  Min.   :49.00   Min.   :  6.39
##  1st Qu.:65.47   1st Qu.: 74.32
##  Median :73.20   Median :106.08
##  Mean   :71.10   Mean   :105.81
##  3rd Qu.:76.80   3rd Qu.:131.35
##  Max.   :84.00   Max.   :239.30
```

Bayesian Linear Model

We start our regression problem by analyzing the full model:

```
##
##  Iterations = 3001:12991
##  Thinning interval  = 10
##  Sample size  = 1000
##
##  DIC: -671.5309
##
##  R-structure:  ~units
##
##       post.mean l-95% CI u-95% CI eff.samp
## units  0.001563 0.001271 0.001874     1000
##
##  Location effects: HDI ~ AB + CO2 + Elec + ForInv + GDP + IM + IntUsers + Lifexp + MobSubs
##
##              post.mean   l-95% CI   u-95% CI eff.samp  pMCMC
## (Intercept)  2.833e-01  1.356e-01  4.561e-01   1000.0  0.004 **
## AB          -2.399e-04 -4.371e-04  2.030e-05   1000.0  0.042 *
## CO2         -9.486e-04 -2.680e-03  5.715e-04   1000.0  0.254
## Elec         1.057e-03  7.063e-04  1.438e-03   1396.2 <0.001 ***
## ForInv      -4.141e-04 -1.309e-03  3.741e-04    826.5  0.326
## GDP          1.222e-06  5.596e-07  1.987e-06   1000.0 <0.001 ***
## IM          -8.807e-04 -1.596e-03 -1.388e-04   1000.0  0.016 *
## IntUsers     1.777e-03  1.315e-03  2.213e-03   1000.0 <0.001 ***
## Lifexp       3.419e-03  1.050e-03  5.224e-03   1000.0 <0.001 ***
## MobSubs      2.339e-04  4.403e-05  4.268e-04   1000.0  0.020 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this model fit we can learn the the variables related to carbon dioxide emissions and foreign direct investments are not significative to we decided to discard them and fit the linear model again.

```
summary(glmmModel2)
```

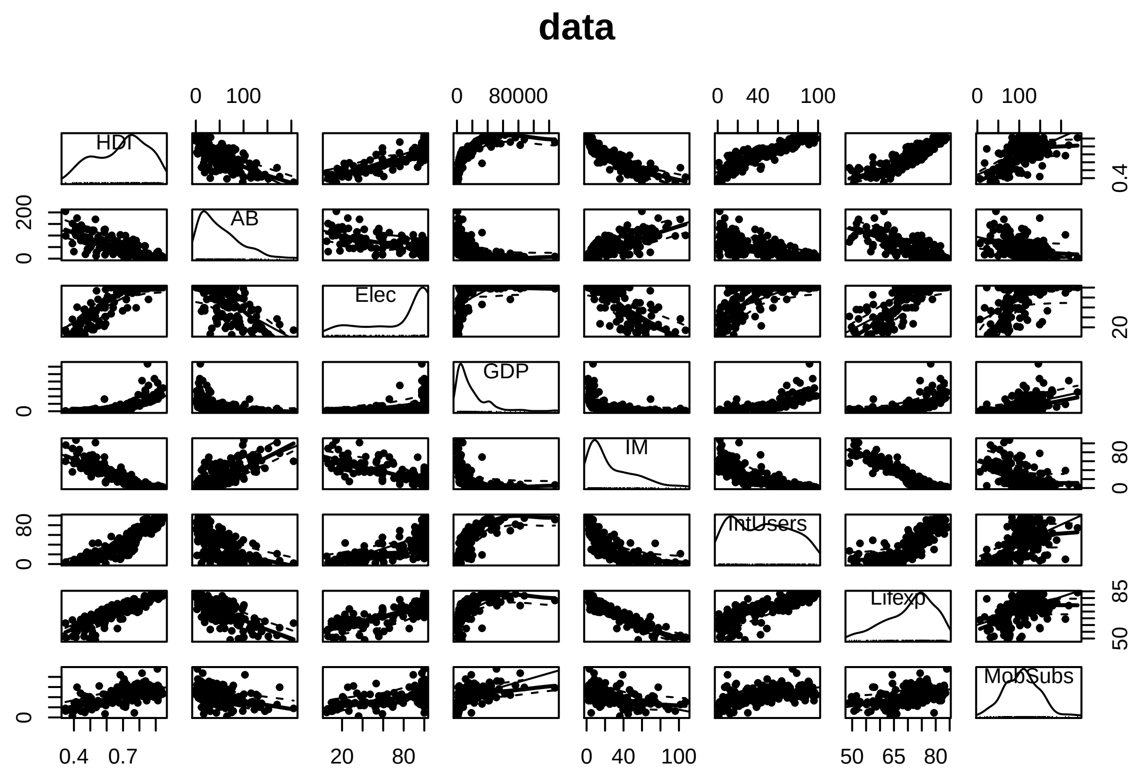
```
##
##  Iterations = 3001:12991
##  Thinning interval  = 10
##  Sample size  = 1000
##
##  DIC: -673.5309
##
##  R-structure:  ~units
##
##       post.mean l-95% CI u-95% CI eff.samp
## units   0.00157 0.001272 0.001894     1000
##
##  Location effects: HDI ~ AB + Elec + GDP + IM + IntUsers + Lifexp + MobSubs
##
##              post.mean   l-95% CI   u-95% CI eff.samp  pMCMC
## (Intercept)  2.565e-01  1.070e-01  4.209e-01     1000 <0.001 ***
## AB          -2.366e-04 -4.743e-04 -2.116e-05     1000  0.034 *
## Elec         1.051e-03  7.235e-04  1.421e-03     1000 <0.001 ***
## GDP          9.017e-07  3.938e-07  1.376e-06     1000 <0.001 ***
## IM          -7.940e-04 -1.484e-03 -9.740e-05     1000  0.032 *
## IntUsers     1.827e-03  1.397e-03  2.240e-03     1138 <0.001 ***
## Lifexp       3.757e-03  1.785e-03  5.887e-03     1000 <0.001 ***
## MobSubs      2.136e-04  4.550e-07  3.886e-04     1000  0.028 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Once checked that all the others features remains significative within the new selection, the aim of the following sections will be trying to improve this model into a non linear one.

Nonlinear Bayesian Regression

Before decidicing which method best fits our needs, we realised a scatterplot of our target and feature variables.

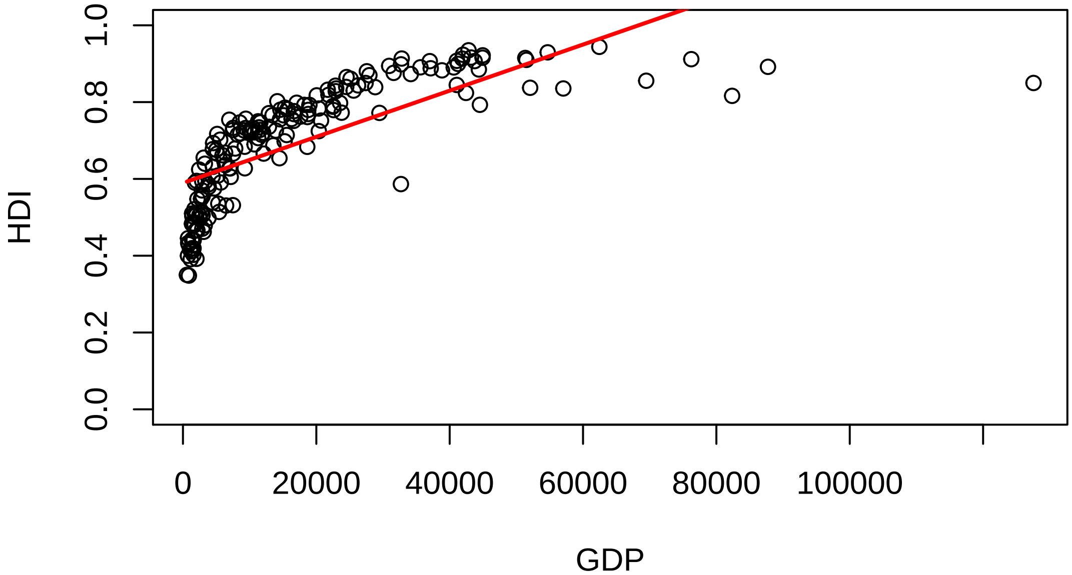
Polynomial Bayesian Regression

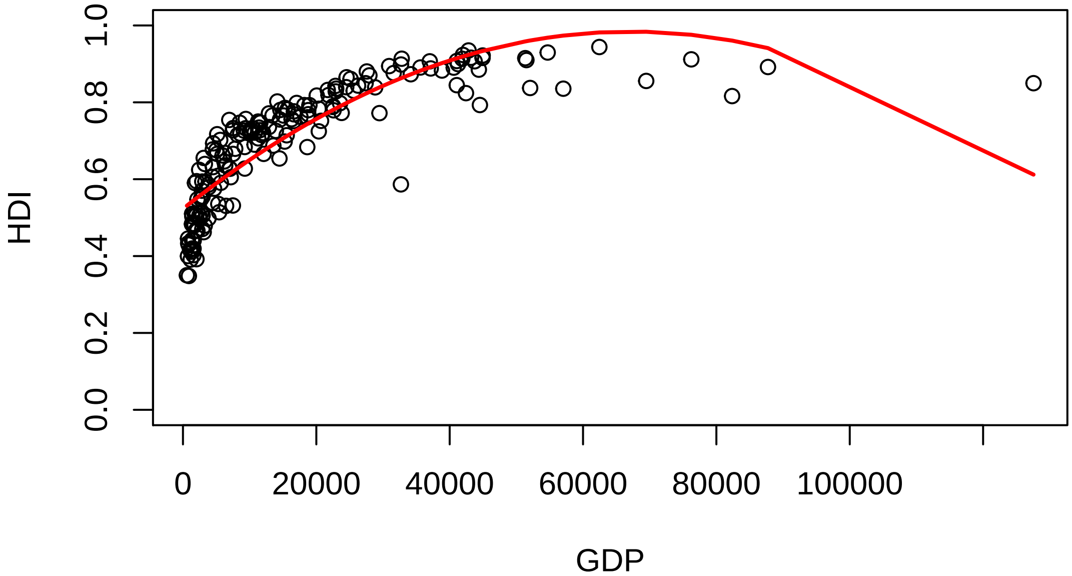


The plot suggest that a polynomial transformation maybe worth for the Gross Domestic Product may be worth, therefore we analysis its single contribution to the model in order to understand in a quadratic degree would produce a better fit.

```
## [1] -307.4475 -433.4232
```

The DIC for the goodness of fit is decreasing for the quadratic model, so we decide to use a degree equal to 2.





The DIC statistics is decreasing for the quadratic model, moreover the plot clearly shows that the degree to is more useful to fit the data.

Let's finally check if the goodness of fit increases by introducing a second polynomial degree for GDP in our previous model.

```r
summary(glmmModel3)
```

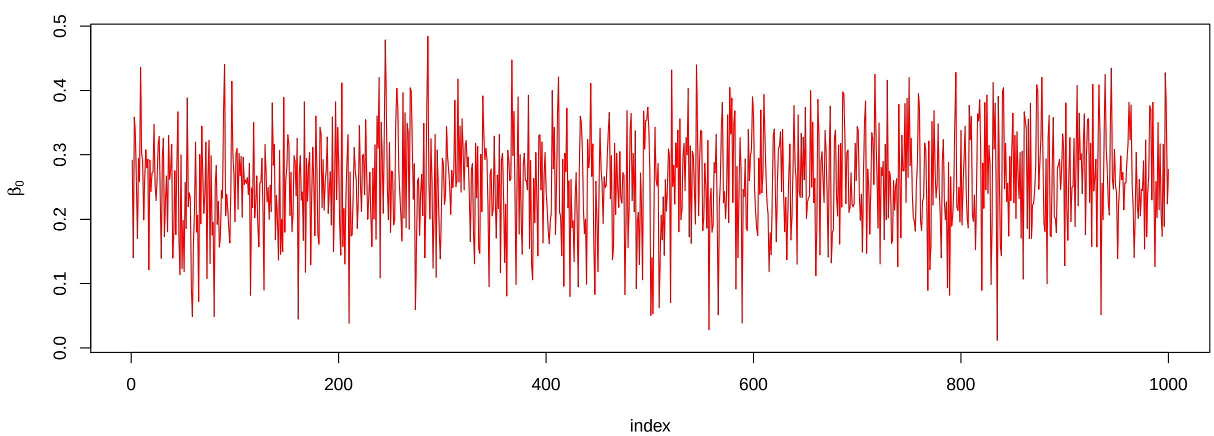
```
##
##  Iterations = 3001:12991
##  Thinning interval  = 10
##  Sample size  = 1000
##
##  DIC: -722.3928
##
##  R-structure:  ~units
##
##       post.mean  l-95% CI u-95% CI eff.samp
## units  0.001195 0.0009608 0.001426     1000
##
##  Location effects: HDI ~ AB + Elec + poly(GDP, 2) + IM + IntUsers + Lifexp + MobSubs
##
##                  post.mean   l-95% CI   u-95% CI eff.samp  pMCMC
## (Intercept)    4.050e-01  2.652e-01  5.401e-01   1000.0 <0.001 ***
## AB            -1.867e-04 -3.681e-04  4.776e-06   1000.0  0.046 *
## Elec           1.037e-03  7.204e-04  1.352e-03    783.9 <0.001 ***
## poly(GDP, 2)1  5.190e-01  3.913e-01  6.566e-01   1000.0 <0.001 ***
## poly(GDP, 2)2 -3.431e-01 -4.328e-01 -2.544e-01   1012.1 <0.001 ***
## IM            -1.118e-03 -1.674e-03 -4.742e-04   1000.0  0.002 **
## IntUsers       1.133e-03  7.516e-04  1.566e-03   1000.0 <0.001 ***
## Lifexp         2.583e-03  9.154e-04  4.340e-03    908.0  0.004 **
## MobSubs        9.850e-05 -8.027e-05  2.630e-04   1000.0  0.246
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

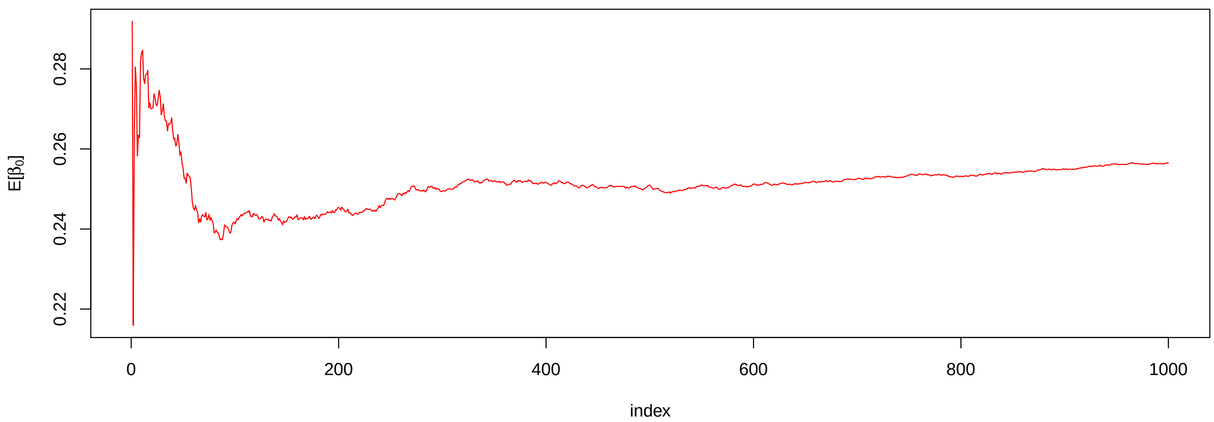
The model is improving as expect with all the significative coefficients and a lower DIC value.

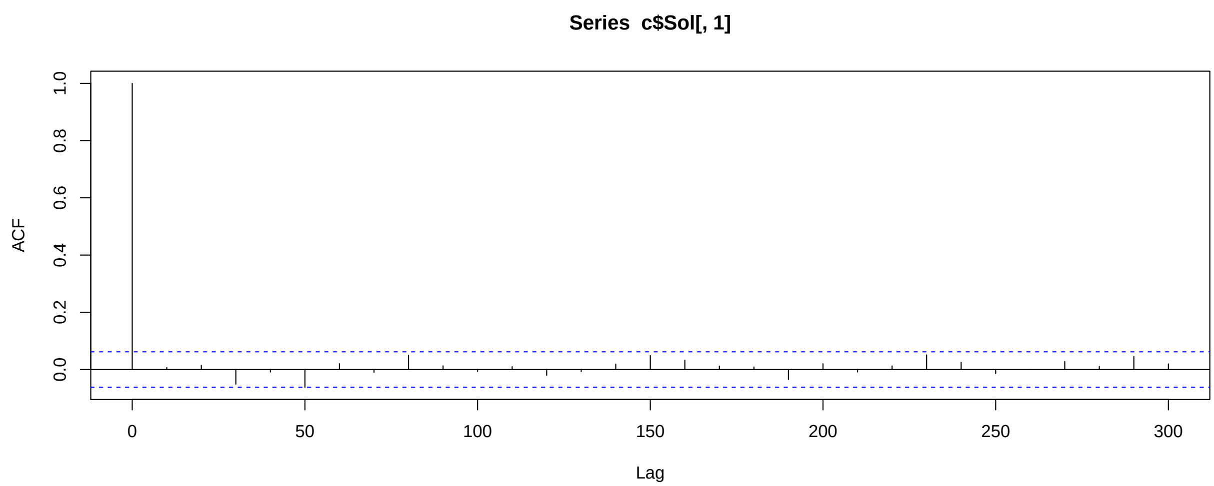
Convergence

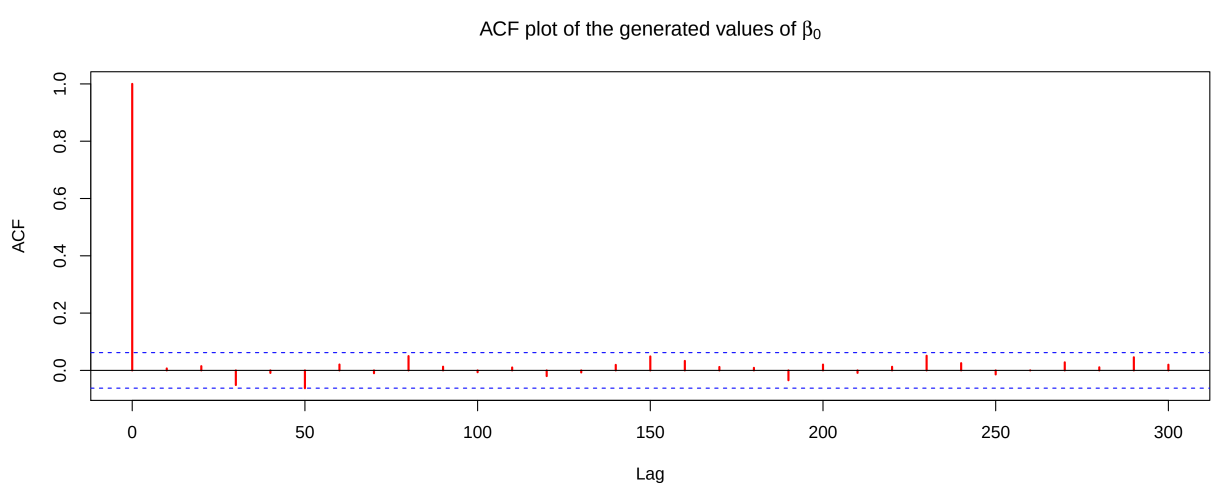
From the graphs:

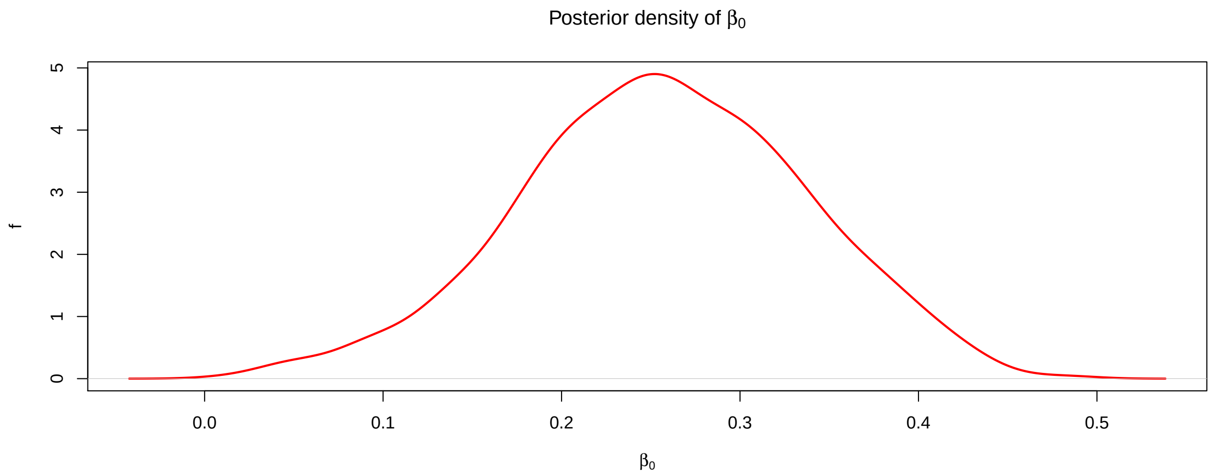
1. The traceplot represents a stationary dataset.
2. Our mean stabilizies.
3. Autocorrelation decreases drastically in 1 step.
4. We have a normal distribution.




E[β0]
0.22
0.24
0.26
0.28
0
200
400
600
800
1000
index


Series c$Sol[, 1]
ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
Lag


ACF plot of the generated values of β0
ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
Lag



Therefore we can not reject the convergence hypothesis.

Spline

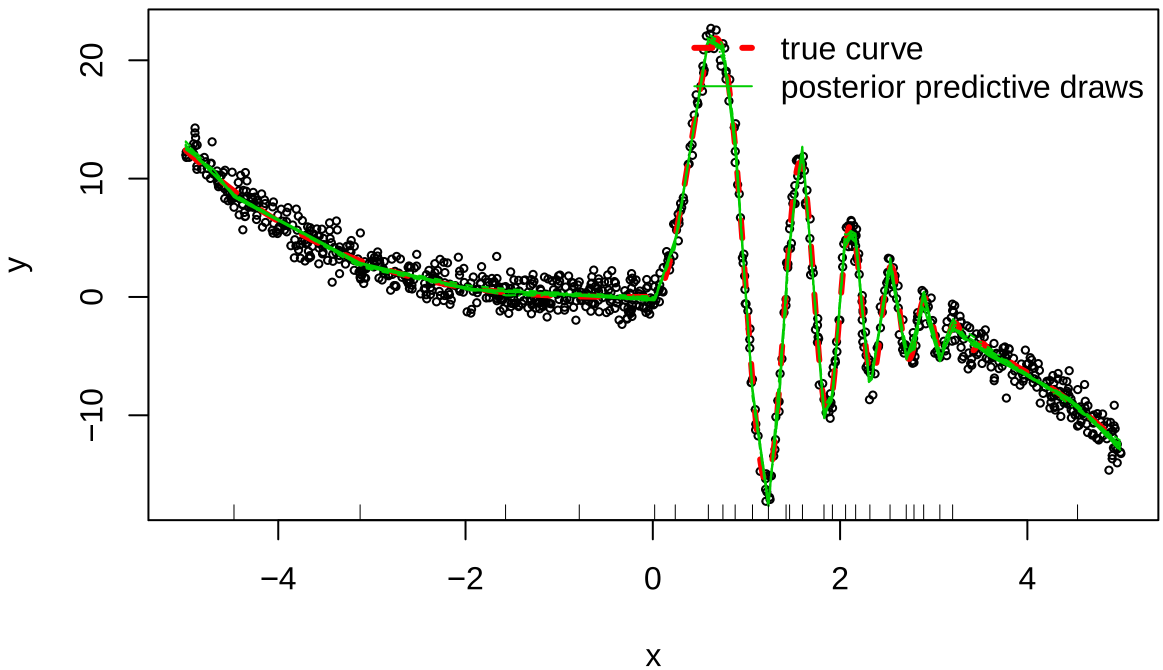
Example

Now we will address the *Mobile phone subscription* variable.

Splines are more easily able to capture different trends in the data, since a polynomial of different orders are fitted the subranges of the variable, so that with lower maximum polynomial degree in each section the model is usually more accurate than higher order degree polynomial regression. To illustrates this fact we generated some simulated data from a Gaussian Distribution in order to show how smooth this method can capture different behaviours in the range on inputs.

```
set.seed(0)
f <- function(x) { -.1 * x^3 + 2 * as.numeric((x < 4) * (x > 0)) * sin(pi * x^2) * (x - 4)^2 }
sigma <- 1
n <- 1000
x <- runif(n, -5, 5)
y <- rnorm(n, f(x), sigma)
```

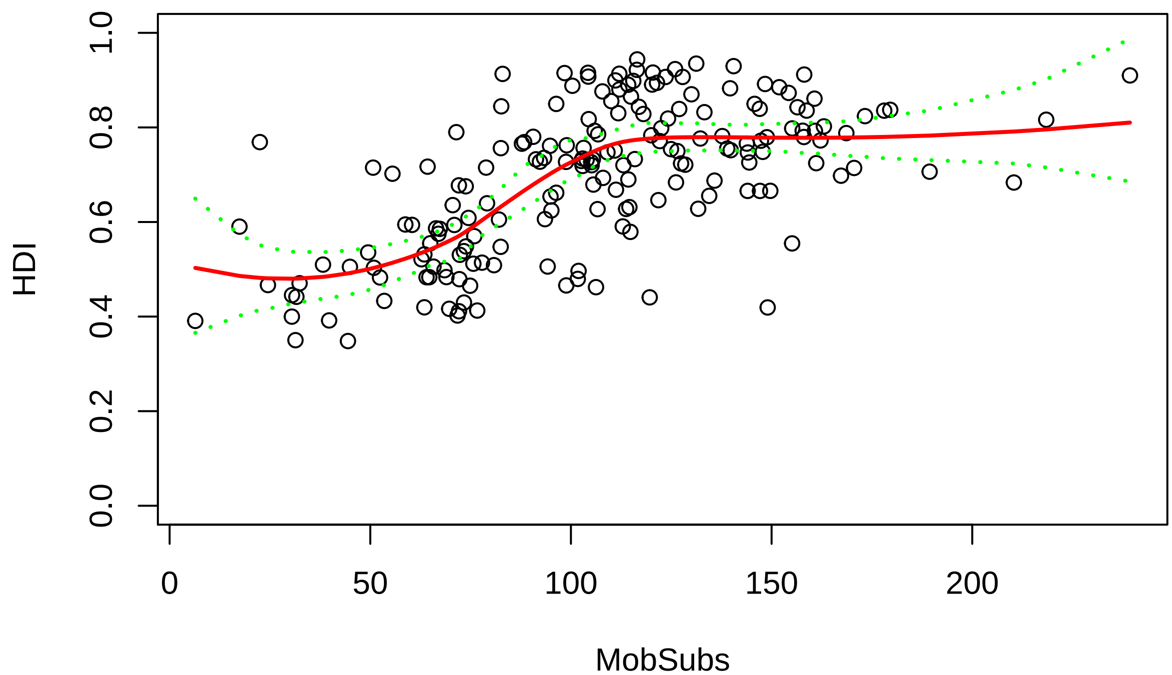
We now plot the artificial data and fit it with spline segmentation:



Since the data is so diverse in the center to the edges, without splines fitting so accurately wouldn't have been possible. In the graph we can see how the fitting has adjusted itself according to the data, by introducing many more knots in the center in order to capture the sinusoidal trend.

Data processing

We now address the knot in *Mobile phone suscription* variable, by imposing its value to 1 according to our scatterplot and by setting the maximum degree allowed for the polynomials to 2.



We can see that the model fits the data better, therefore the goodness of fit of the glmmModel3 will increase by introducing the spline regression to the *Mobile Users* variable (Know = 1, Degree = 2).