Group 34 course work

Table of contents

Introduction:

Tuberculosis (TB) is a bacterial disease that primarily affects the lungs but can also impact other parts of the body. It is a significant public health problem, with approximately 10 million cases reported globally in 2020. Brazil is a high-burden country for TB, with an estimated 96,000 cases. The purpose of this report is to determine whether various socio-economic variables impact the rate of TB per unit population in Brazil between 2012 and 2014. We will analyze data from 537 micro-regions in Brazil, including the latitude and longitude of each region and the year in which the data was collected. *Global Tuberculosis Report 2020* (2020).

The socioeconomic variables that were recorded for each micro-region in our data set are as follows: the level of illiteracy, urbanisation, poverty, unemployment, and sanitation; the proportion of indigenous population; the dwelling density; and finally, a proxy indicator of the amount of resources in the form of the average amount of time between diagnosing a TB case and reporting it to the health system. In addition, the latitude and longitude of the respective 537 micro-regions, as well as the year in which the data was obtained, are supplied for each of these values. Because of this, we will be better able to explain the geographical, temporal, and spatio-temporal structure of any systematic risk that is not described by the covariates.

Exploratory Data Analysis:

The TBdata dataframe contains information on various socio-demographic previously described and includes information on the number of TB cases and population size for each microregion, as well as unique ID numbers to distinguish between the different regions.

We conducted a preliminary exploration of the relationships between our covariates and the rate of tuberculosis in each microregion of Brazil using pairplots. We found some unexpected results, such as a lower degree of sanitation and poverty not being associated with a higher rate of tuberculosis cases. However, this simple approach doesn't account for changes in other variables for each data point, so we need to use a formal model to investigate the impact of our covariates.

refer to: bellow ## TODO INSERT Reference to PAIR PLOT -1 AND SUMMARY-1 TABLE HERE.

We chose generalized additive models (GAMs) as our preferred framework, which strikes a balance between interpretability and flexibility compared to linear models and machine learning models like neural networks or boosted trees.

TODO reference INSERT BRAZIL MAP HERE.

aka insert the better version of ggpairs

Key trends observed include:

-Normal distribution with right skewness and mean of 0.6 for dwelling density. -Right-skewed distribution with a normal bell curve around 5% for illiteracy. -Extreme right skewness for poor sanitation. -Normal distribution with little to no skewness and mean of 6% for unemployment. -Population is the variable most correlated with TB, while illiteracy, poor sanitation, and poverty have negative correlations.

Model Selection:

At first, the idea was to combine a log link and a Poisson generalised additive model. This was because we were working with count data in the form of TB cases broken down by each microregion. Nonetheless, it was essential to standardise this count because the populations in each region were distinct from one another. After doing some study on the topic, I discovered that using the log of the population in the model is the most effective way to carry out a population standardised regression. While we were trying to fit our model in R, it was necessary for us to include an offset for the log of the population. This provides us with the tuberculosis rate per capita:

Let the model be :

$$Log(\lambda_i) = offset(log(Population_i)) + f_1(Indigenous_i) + f_2(Illiteracy_i) + f_3(Urbanisation_i) + f_4(Density_i) + f_5(I$$

Confirming that our model was accurate was the next step in the process of developing our model. Unfortuitously, a QQ-plot revealed that the quantiles in our data did not closely resemble what they would look like if they followed a theoretical poisson distribution. This was the conclusion that could be drawn from the plot.

After calculating a Pearson estimate of our dispersion parameter to determine whether or not this lack of fit was due to over-dispersion, the results were unequivocal: the parameter was almost 7 (when it should have been 1), which indicated that our data was indeed over dispersed for a Poisson model.

An alternative model to Poisson is a Negative Binomial model. The Negative Binomial (NB) model is likewise suitable for use with count data; however, it differs from the Poisson regression in that it contains an additional parameter that can alter the variance independently from the mean.

mean. Because of this, it is more flexible than the Poisson model, and as a result, it can fit data with a greater degree of fluctuation. A NB model was fitted to the data making use of

the exact same specification as was used before. We can count ourselves fortunate that the QQ-plot5 for the revised model showed a good fit (the majority of the dots fell on the y=x line), and the estimate for the dispersion parameter was 1.133. The AIC was also reduced for our newly developed model. We arrived at the following model after making adjustments to the choice of rank for each of our smooth terms (covariates), until we were certain that they were not too low (edf not too near to k').

INSERT BOTH QQ PLOT HERE, FIGURE -2 at the last of the code.

Model Fitting:

TODO QQ-plot reference

Our QQ-plot suggest that the quantiles in our data our not similar to the line as it deviates from the line in nearly all the values, showing a very flawed fit. As this suggests that our current model doesn't fit the data correctly and required an extension to our model as the Poisson GAM is not accounted for enough deviance as seen in the residuals.

Since the model is not accounting for enough of the variance we will check if there is a significant difference between the variance and the mean. In this analyses we will use the Pearson estimate for the dispersion parameter, this method allow us to estimate the amount of extra variability, or over-dispersion in count data and therefore analyse if the Poisson distribution assumption of equal mean and variance holds.

Model Evaluation:

TODO dispersion parameter

As we can see from the dispersion parameter should be 1 for the assumption of equal mean and variance to hold true, so it seems that there is substantial over-dispersion in the Poisson GAM. This violates one of the Poisson assumptions that the mean and variance are equal therefore we will have to extend the model from e GAM Poisson to a Negative Binomial GAM

As we can see from the residual versus predictor plot, the values seem to be randomly scattered with no clear trend but with some distance from the zero line. As such we can determine that this scatter is due to random errors and not a unaccounted pastern in the model.

##negative binomial QQ-plot reference

The QQ-plot looks much better for the Negative Binomial model. The majority of points lie either on top of very near the y=x line, except for a few towards the extremes. This indicates our assumption about the true distribution of the data is a lot more safe than it was before.

TODO add three residual sum here

sum(residuals(nb_model, type = "pearson")^2) / df.residual(nb_model)

The dispersion parameter is very close to 1, unlike for the Poisson model, meaning that the model that can account for most of the over-dispersion in the data. As such a dispersion parameter value close to 1 can be interpreted as the model is a good fit for the data due to the model adequately capture the variability of the the response variable.

Results and Interpretation:

Conclusion:

Figures

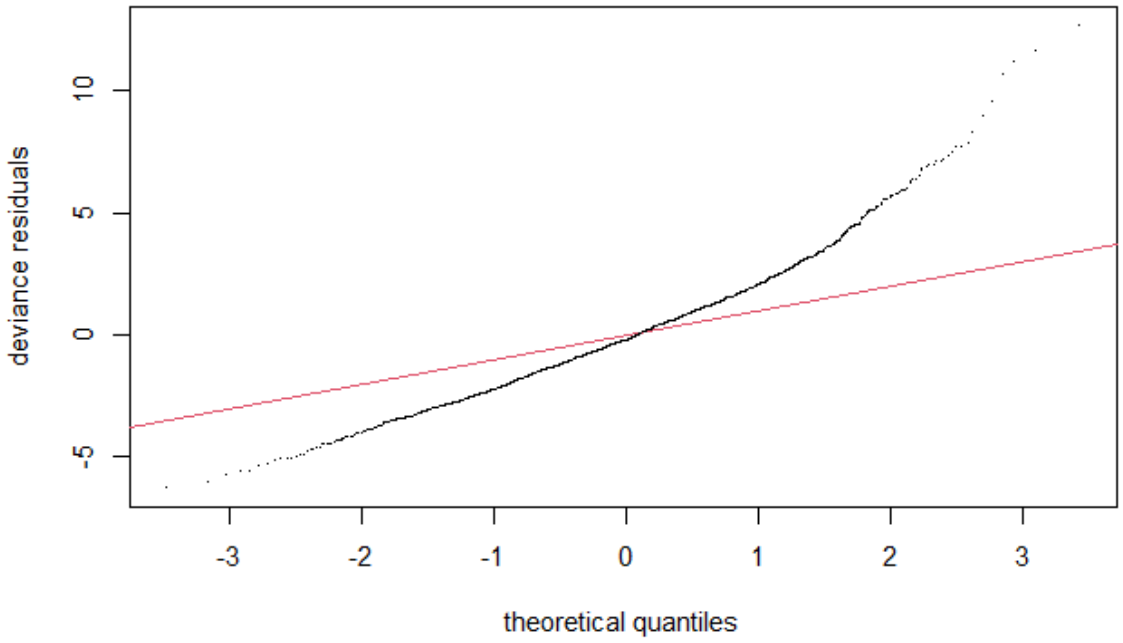
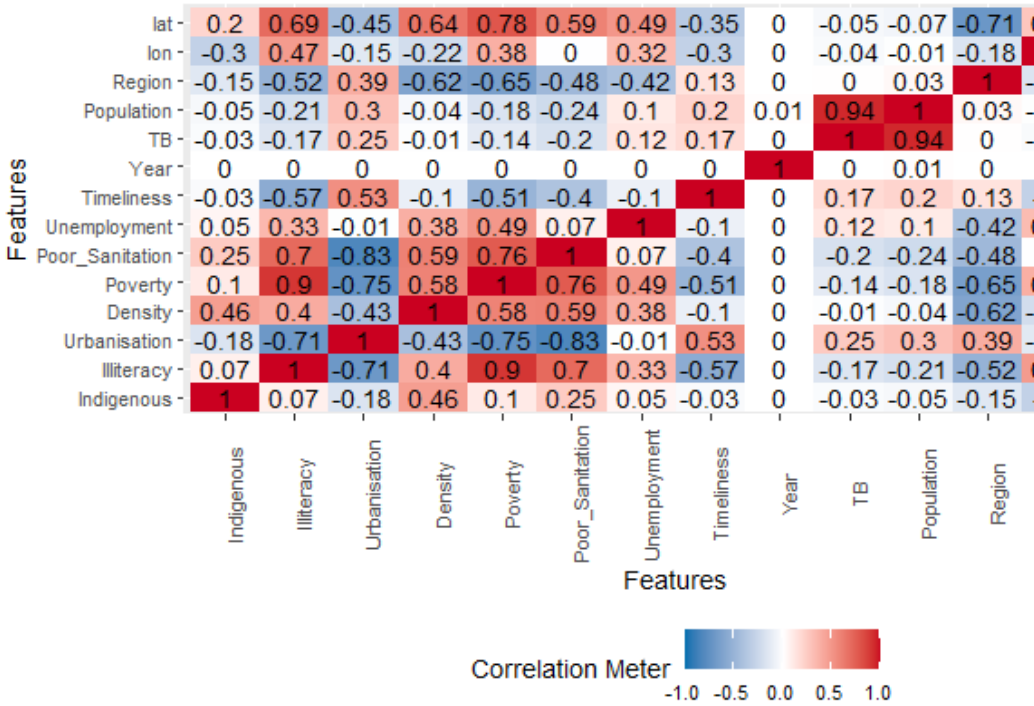


Figure 1: QQ Plot



{figure label="fig:heatMap"}
{caption: This is the heatmap that we will remove and get the better ggpairs} {/figure}

Code Appendix

```
## PLotting map of cases
par(mfrow = c(1,3))
plot.map(TBdata$TB[TBdata$Year==2012],n.levels=7,main="TB counts for 2012")
plot.map(TBdata$TB[TBdata$Year==2013],n.levels=7,main="TB counts for 2013")
plot.map(TBdata$TB[TBdata$Year==2014],n.levels=7,main="TB counts for 2014")
```

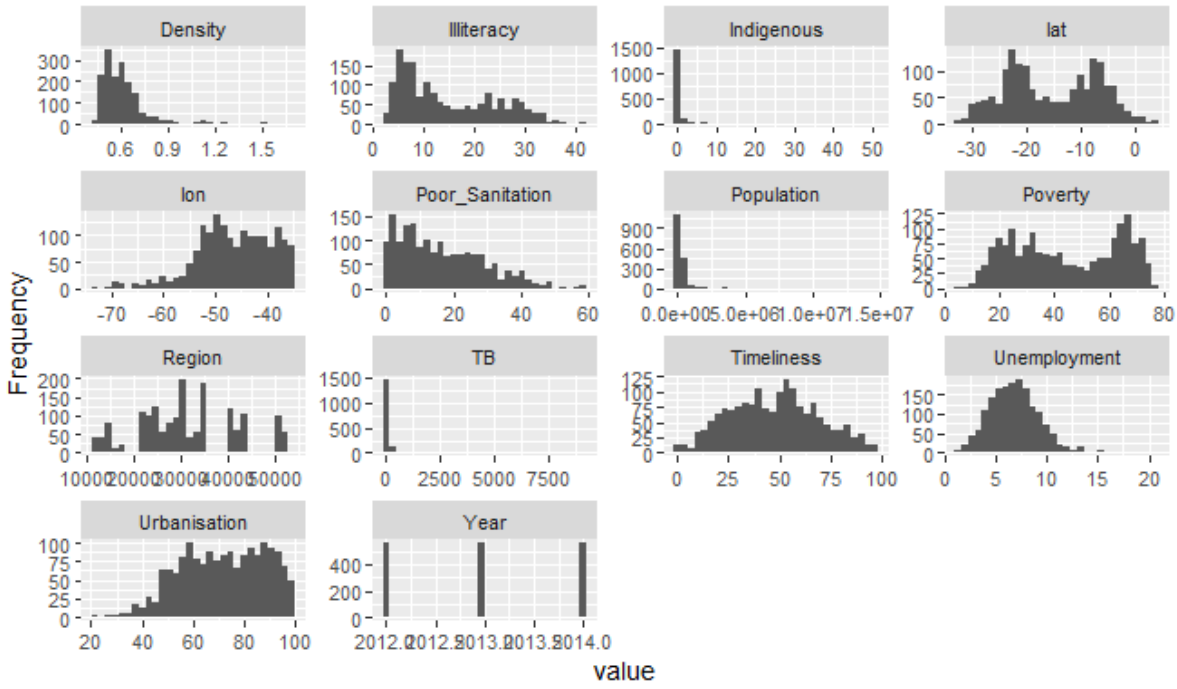
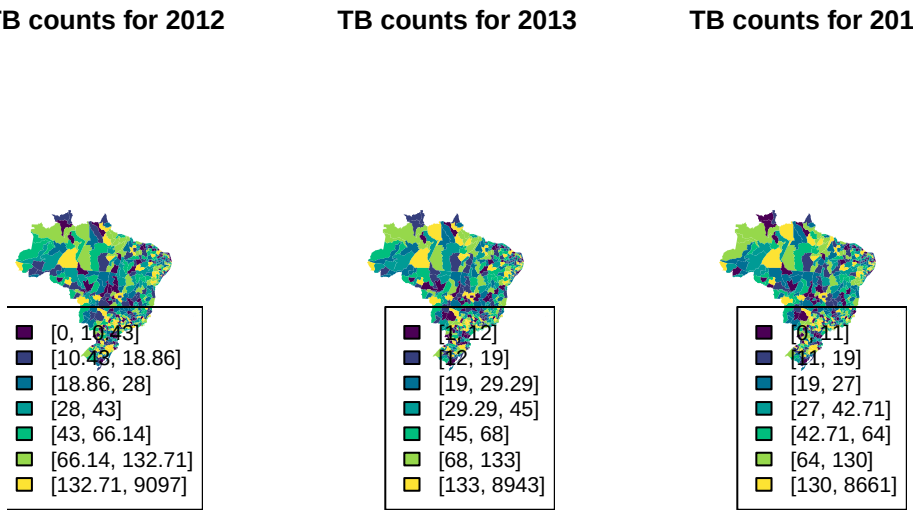


Figure 2: Initial variable analysis




```
par(mfrow = c(1,1))
```

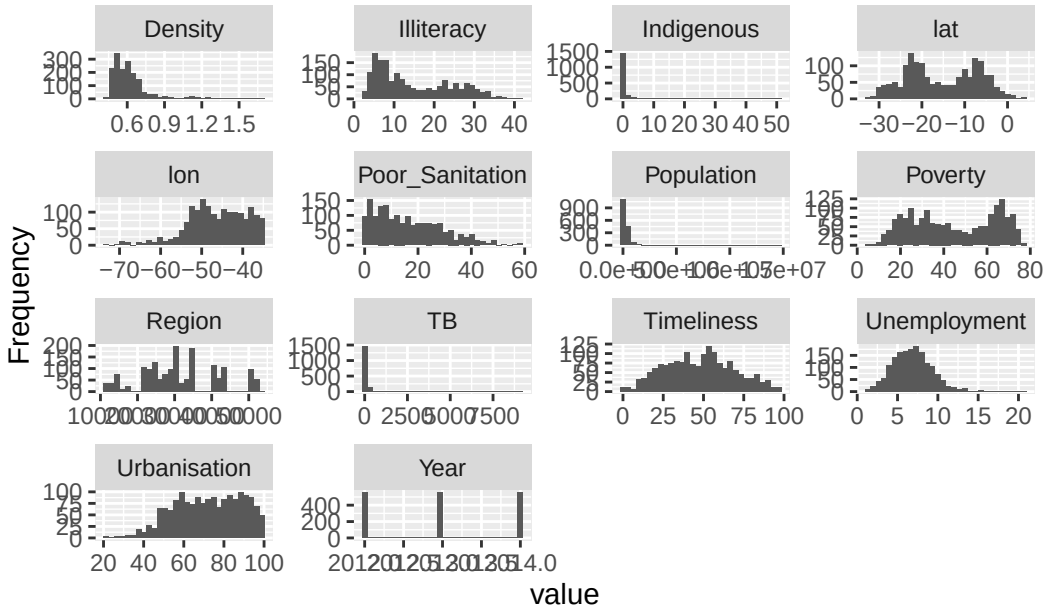
Exploratory analyses

```
# INSERT THIS AS SUMMARY TABLE AS TABLE-1
summary_table <- summary(TBdata)
# Convert the summary table to a LaTeX table using the xtable() function
latex_table <- xtable(summary_table)

# Print the LaTeX table to the console
print(latex_table)
```

```
% latex table generated in R 4.2.2 by xtable 1.8-4 package
% Wed Mar 22 21:20:41 2023
\begin{table}[ht]
\centering
\begin{tabular}{rlllllllllllll}
  \hline
 &   Indigenous &   Illiteracy &  Urbanisation &    Density &    Poverty & Poor\_Sanitation &
  \hline
```

```
X & Min.   : 0.01034   & Min.   : 2.336   & Min.   :22.34   & Min.   :0.4223   & Min.   : 5.9
  X.1 & 1st Qu.: 0.06366   & 1st Qu.: 6.683   & 1st Qu.:58.45   & 1st Qu.:0.5166   & 1st Qu.
  X.2 & Median : 0.10577   & Median :11.516   & Median :72.66   & Median :0.5840   & Median
  X.3 & Mean   : 0.84307   & Mean   :14.802   & Mean   :71.96   & Mean   :0.6212   & Mean
  X.4 & 3rd Qu.: 0.23973   & 3rd Qu.:22.844   & 3rd Qu.:86.16   & 3rd Qu.:0.6585   & 3rd Qu.
  X.5 & Max.   :50.64623   & Max.   :41.137   & Max.   :99.93   & Max.   :1.6751   & Max.
   \hline
\end{tabular}
\end{table}
```

```
# AIM TO SKIP IT AS WE HAVE ALREADY USED CORRPLOT
plot_histogram(TBdata)
```



//TODO talk about the histograms and relevant distributions we can observe

Now investigating the correlation matrix of the numerical variables

```
# NOT REQUIRED AS WE HAVE ALREADY USED CORRPLOT
# plot_correlation(TBdata)
```

As we can see from the matrix, the variable that is the most correlated from TB is the population, with illiteracy, poor sanitation and poverty having a negative correlation with

TB.

Poisson definition

As the data is count data we will first fit a Poisson module since this distribution is a good fit for the nature of the data

```
poisson_model <- gam(TB ~ offset(log(Population)) + s(Indigenous, k = 20) + s(Illiteracy ,
summary(poisson_model)
```

```
Family: poisson
Link function: log

Formula:
TB ~ offset(log(Population)) + s(Indigenous, k = 20) + s(Illiteracy,
    k = 20) + s(Urbanisation, k = 20) + s(Density, k = 20) +
    s(Poverty, k = 20) + s(Poor_Sanitation, k = 20) + s(Unemployment,
    k = 20) + s(Year, k = 3) + s(Timeliness, k = 20) + te(lon,
    Year, k = 3) + te(lat, Year, k = 3)

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.467263   0.004435   -1909   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                      edf Ref.df Chi.sq p-value
s(Indigenous)      18.393  18.92 1051.3  <2e-16 ***
s(Illiteracy)      18.157  18.89  341.5  <2e-16 ***
s(Urbanisation)    18.879  18.99 1502.8  <2e-16 ***
s(Density)         17.396  18.05 1763.1  <2e-16 ***
s(Poverty)         18.636  18.95 2027.8  <2e-16 ***
s(Poor_Sanitation) 18.327  18.91 1251.0  <2e-16 ***
s(Unemployment)    18.648  18.98 2622.8  <2e-16 ***
s(Year)             1.999   2.00 1797.0  <2e-16 ***
s(Timeliness)      18.328  18.91  927.0  <2e-16 ***
te(lon,Year)        5.980   6.00 1440.1  <2e-16 ***
te(lat,Year)        5.986   6.00 1301.2  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

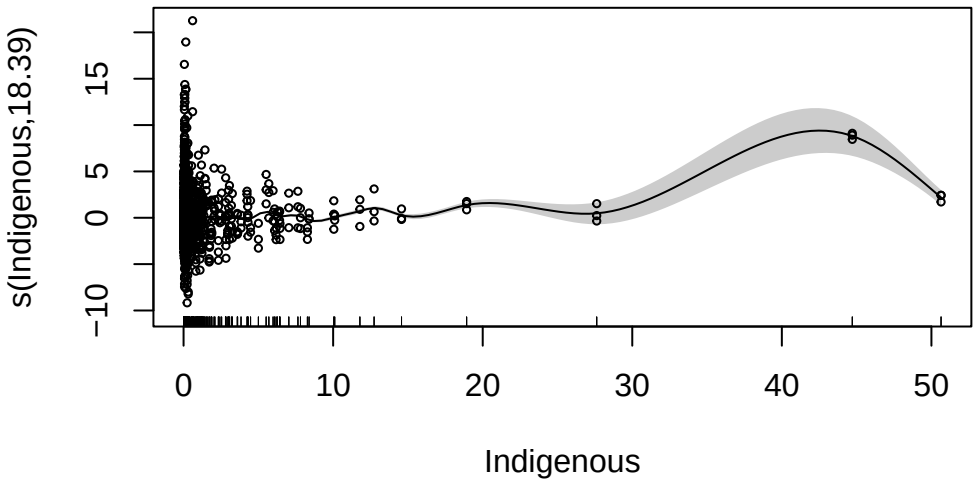
```
R-sq.(adj) =  0.995   Deviance explained = 82.9%
-REML =  11564  Scale est. = 1         n = 1671
```

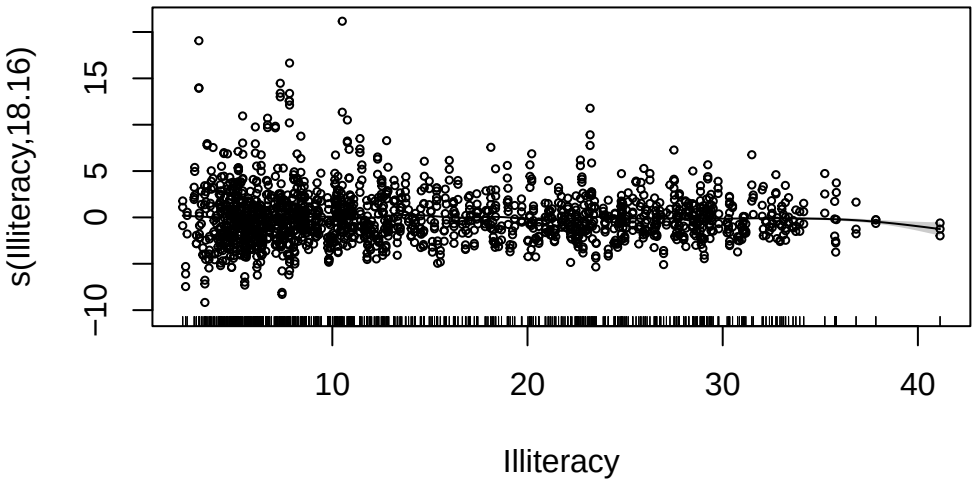
```
gam.check(poisson_model)
```

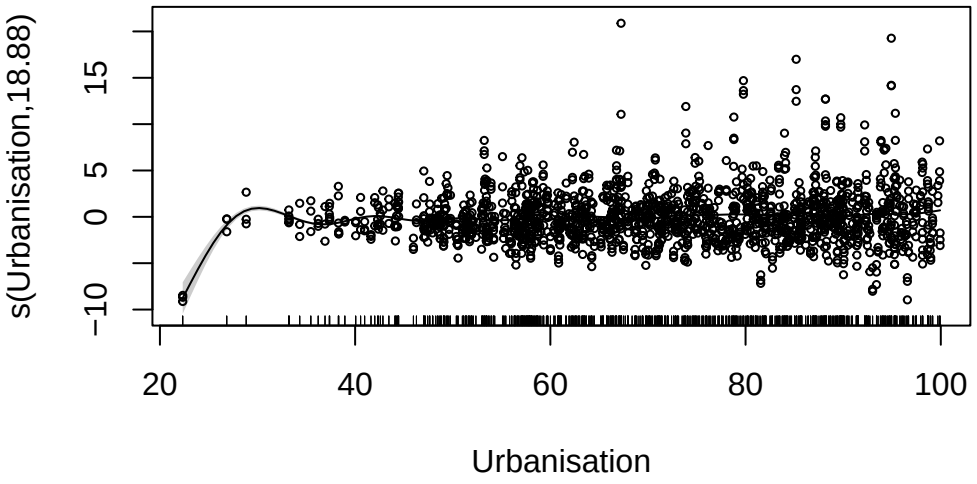
```
#Akaike Information Criterion:
poisson_model$aic
```

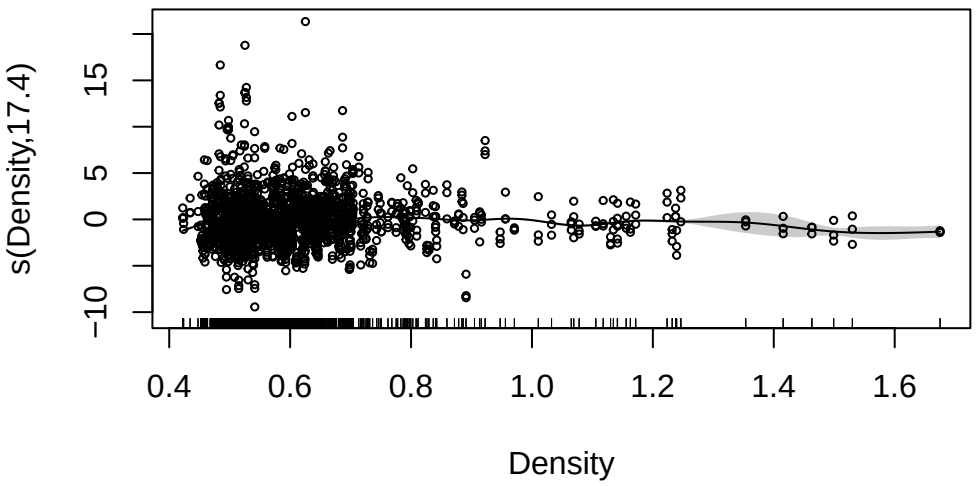
```
[1] 22271.66
```

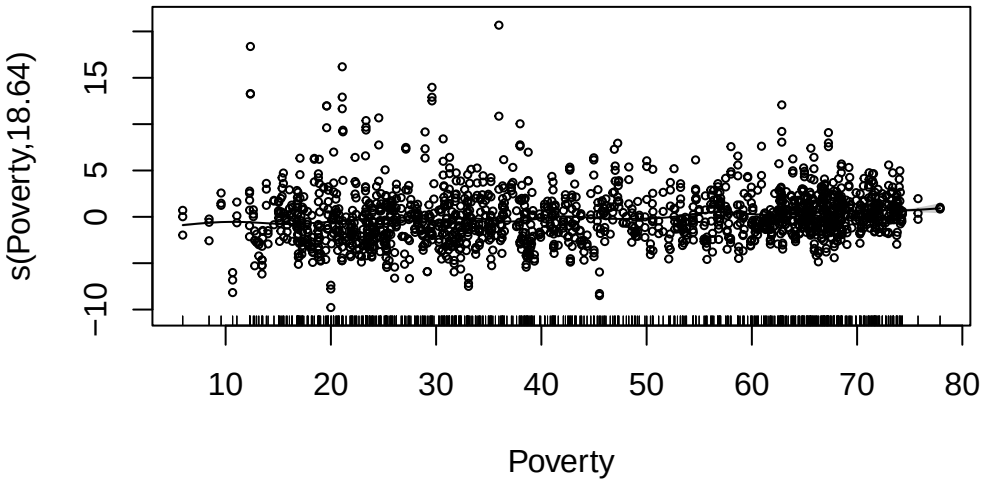
```
plot(poisson_model, shade=T, rug = TRUE, residuals = TRUE,
pch = 1, cex = 0.5)
```

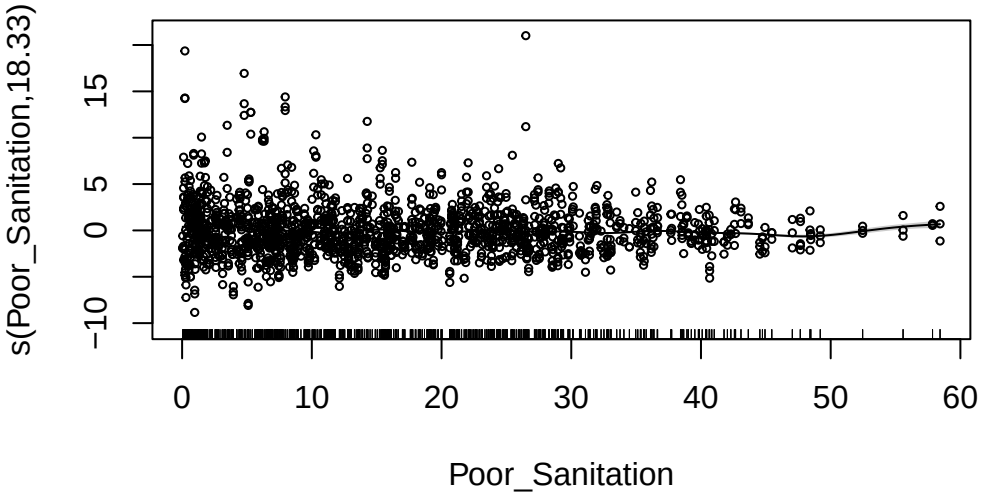


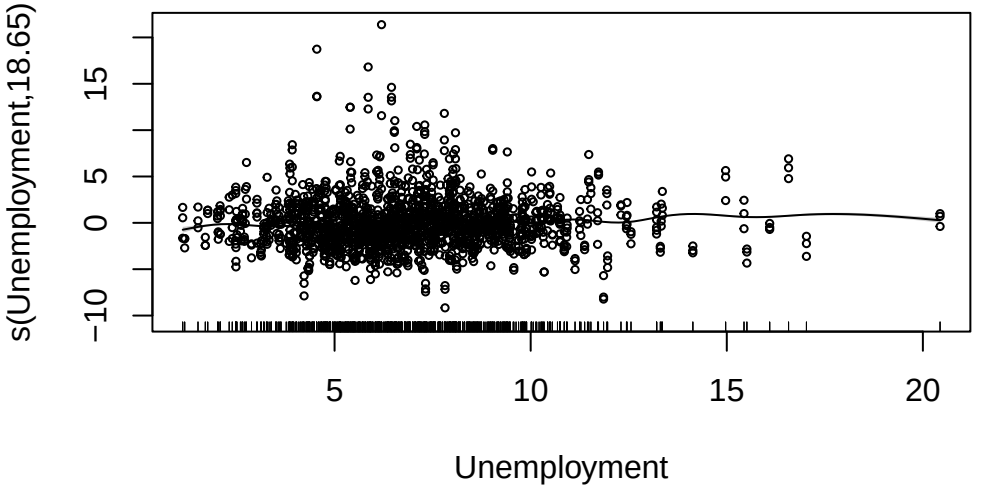

s(Illiteracy,18.16)
−10
0
5
15
10
20
30
40
Illiteracy

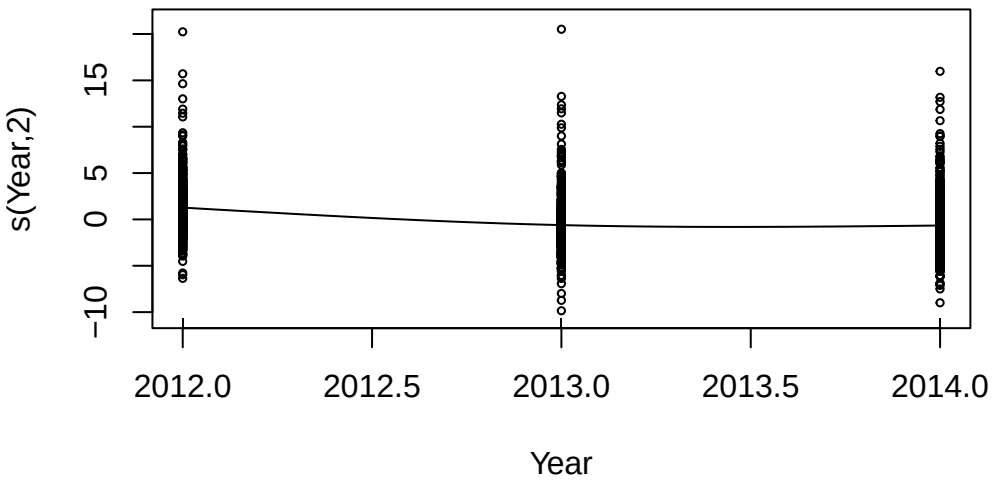

s(Urbanisation,18.88)
−10
0
5
15
20
40
60
80
100
Urbanisation

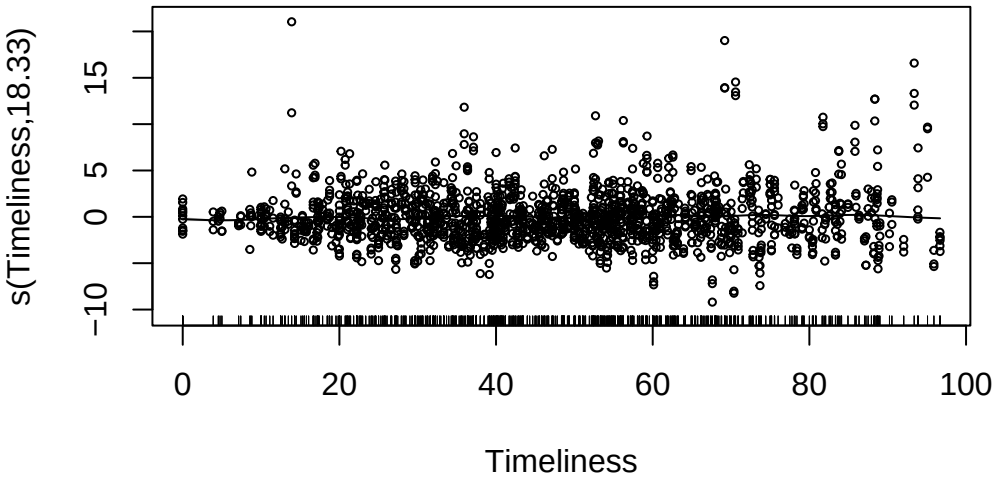

s(Density,17.4)
−10
0
5
15
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Density

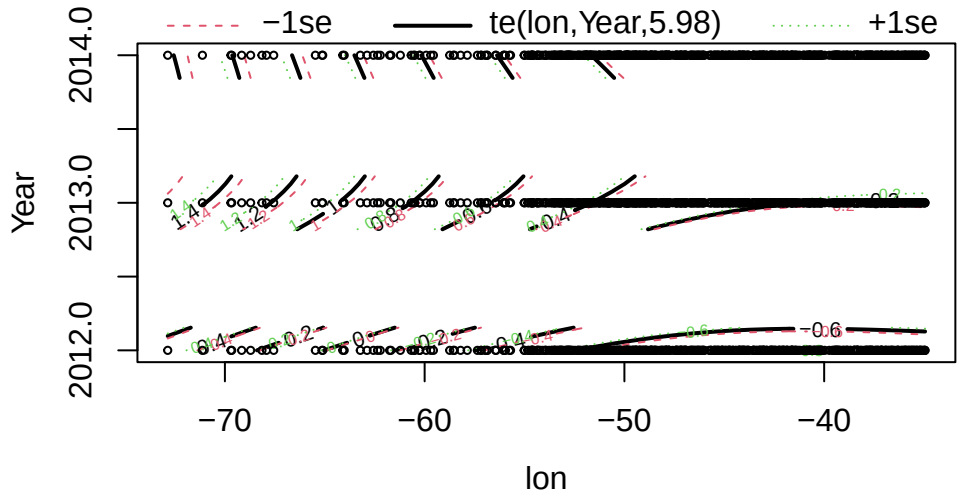

s(Poverty,18.64)
−10
0
5
15
10
20
30
40
50
60
70
80
Poverty

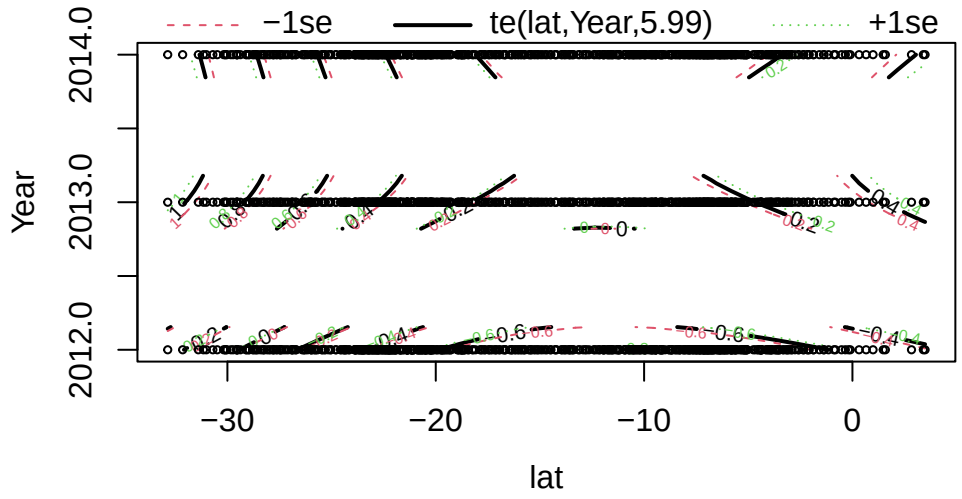

s(Poor_Sanitation,18.33)
−10
0
5
15
0
10
20
30
40
50
60
Poor_Sanitation


s(Unemployment,18.65)
−10
0
5
15
5
10
15
20
Unemployment


s(Year,2)
−10
0
5
15
2012.0
2012.5
2013.0
2013.5
2014.0
Year


s(Timeliness,18.33)
−10
0
5
15
0
20
40
60
80
100
Timeliness





Our QQ-plot suggest that the poisson model fit deviates from the theoretical quantiles in nearly all the values, showing a very flawed fit. As this suggests that our current model doesn't fit the

data correctly and required an extension to our model as the Poisson GAM is not accounted for enough deviance as seen in the residuals.

Since the model is not accounting for enough of the variance we will check if there is a significant difference between the variance and the mean. In this analyses we will use the Pearson estimate for the dispersion parameter, this method allow us to estimate the amount of extra variability, or over-dispersion in count data and therefore analyse if the Poisson distribution assumption of equal mean and variance holds.

```
#Calculating Pearson estimate for dispersion parameter using Pearson residuals:
sum(residuals(poisson_model, type = "pearson")^2) / df.residual(poisson_model)
```

```
[1] 9.353406
```

```
#The dispersion parameter should be 1, so it seems that there is substantial over-dispersi
```

As we can see from the dispersion parameter should be 1 for the assumption of equal mean and variance to hold true, so it seems that there is substantial over-dispersion in the Poisson GAM. This violates one of the Poisson assumptions that the mean and variance are equal therefore we will have to extend the model from e GAM Poisson to a Negative Binomial GAM

Negative binomial

```
#fitting a negative-binomial model to our TB data:
nb_model <- gam(TB ~ offset(log(Population)) + s(Indigenous, k = 20) + s(Illiteracy , k =

summary(nb_model)
```

```
Family: Negative Binomial(9)
Link function: log

Formula:
TB ~ offset(log(Population)) + s(Indigenous, k = 20) + s(Illiteracy,
    k = 20) + s(Urbanisation, k = 20) + s(Density, k = 20) +
    s(Poverty, k = 20) + s(Poor_Sanitation, k = 20) + s(Unemployment,
    k = 20) + s(Year, k = 3) + s(Timeliness, k = 20) + te(lon,
    Year, k = 3) + te(lat, Year, k = 3)

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept) -8.442775   0.009432  -895.1   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                     edf Ref.df Chi.sq  p-value
s(Indigenous)      1.001  1.002  22.58 1.90e-06 ***
s(Illiteracy)     10.893 13.279  37.81 0.000278 ***
s(Urbanisation)   11.070 13.410  58.68  < 2e-16 ***
s(Density)        11.231 13.495 155.64  < 2e-16 ***
s(Poverty)         7.109  8.897  36.63 2.36e-05 ***
s(Poor_Sanitation) 14.726 16.913 134.41  < 2e-16 ***
s(Unemployment)    7.203  8.927  99.22  < 2e-16 ***
s(Year)            1.986  1.998 102.59  < 2e-16 ***
s(Timeliness)      5.216  6.533  71.00  < 2e-16 ***
te(lon,Year)       5.700  5.961 101.69  < 2e-16 ***
te(lat,Year)       5.664  5.953  83.69  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.895   Deviance explained = 53.4%
-REML = 7203.2  Scale est. = 1         n = 1671
```

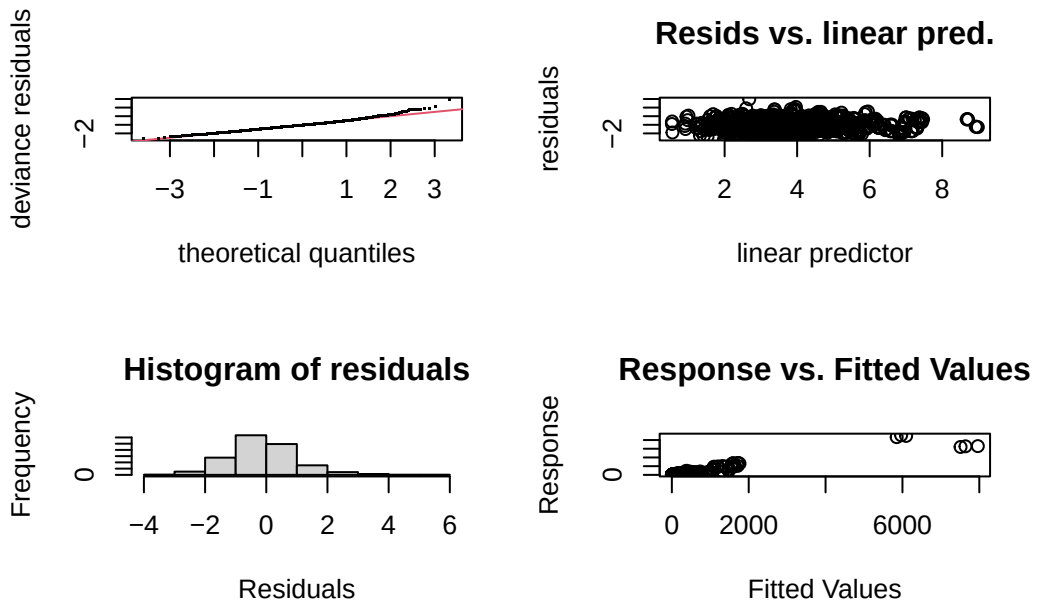
```
#Akaike Information Criterion
nb_model$aic
```

```
[1] 14199.83
```

As we can see from this Akaike Information Criterion(AIC) the Negative Binomial has a significantly lower value than the previous 18585,52 from the GAM Poisson, meaning this is already a better fitting model than the previous one.

Now we will check the residuals to check for any anomalies on our model prediction

```
gam.check(nb_model)
```

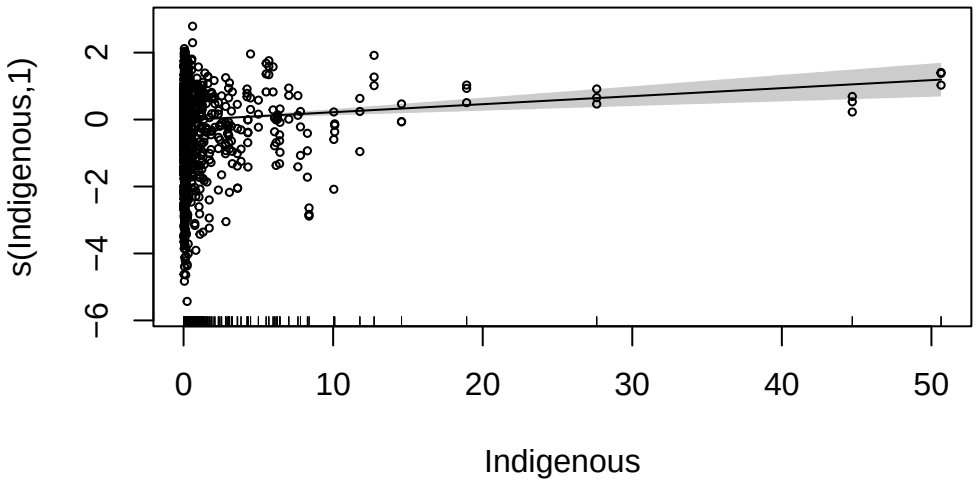


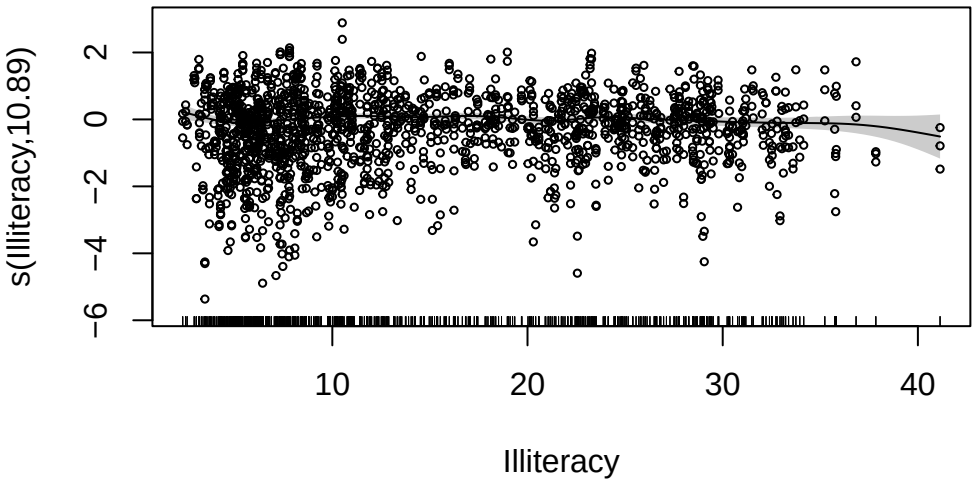

```
Method: REML   Optimizer: outer newton
full convergence after 12 iterations.
Gradient range [-0.0002588633,5.095501e-05]
(score 7203.166 & scale 1).
Hessian positive definite, eigenvalue range [0.0002588044,2.739525].
Model rank =  167 / 167

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                      k'   edf k-index p-value
s(Indigenous)      19.00  1.00    0.52  <2e-16 ***
s(Illiteracy)      19.00 10.89    0.51  <2e-16 ***
s(Urbanisation)    19.00 11.07    0.52  <2e-16 ***
s(Density)         19.00 11.23    0.52  <2e-16 ***
s(Poverty)         19.00  7.11    0.52  <2e-16 ***
s(Poor_Sanitation) 19.00 14.73    0.52  <2e-16 ***
s(Unemployment)    19.00  7.20    0.52  <2e-16 ***
s(Year)             2.00  1.99    0.73  <2e-16 ***
s(Timeliness)      19.00  5.22    0.57  <2e-16 ***
te(lon,Year)        6.00  5.70    0.97    0.12
te(lat,Year)        6.00  5.66    0.96    0.12
---
```

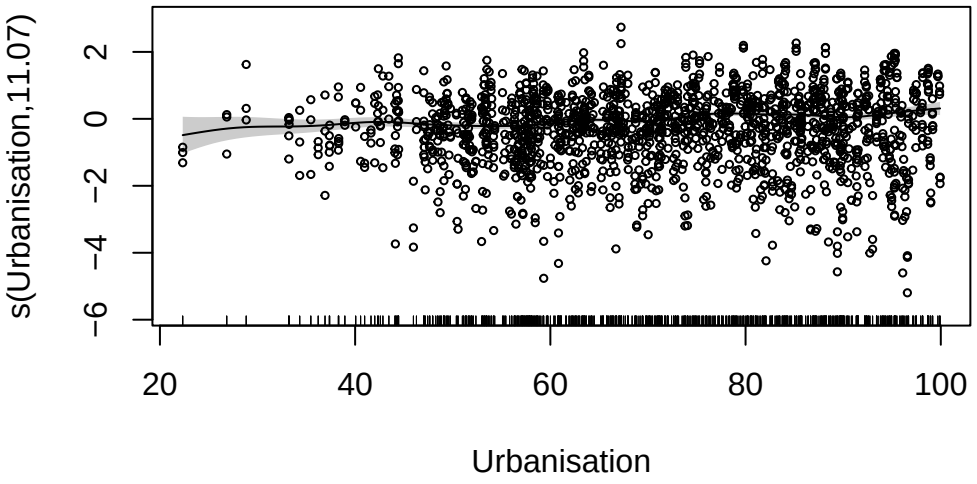
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

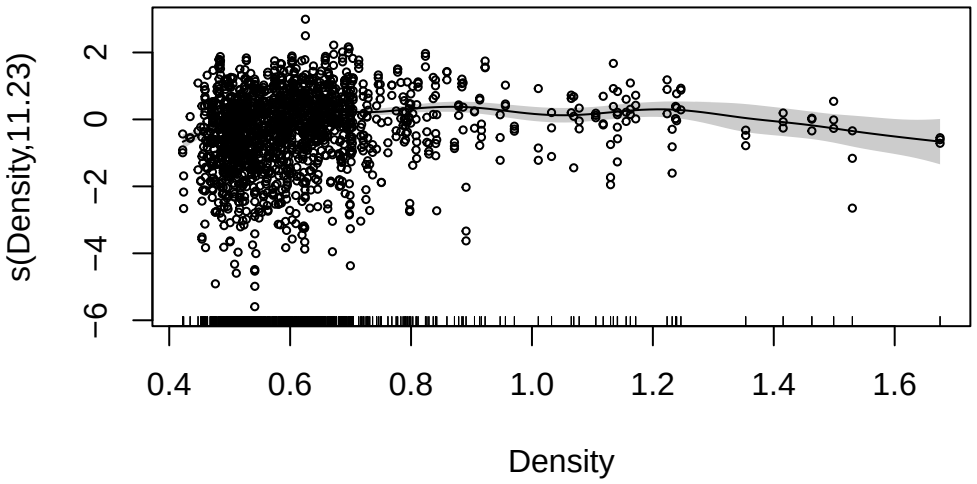
As we can see from the residual versus predictor plot, the values seem to be randomly scattered with no clear trend but with some distance from the zero line. As such we can determine that this scatter is due to random errors and not a uncounted pattern in the model.

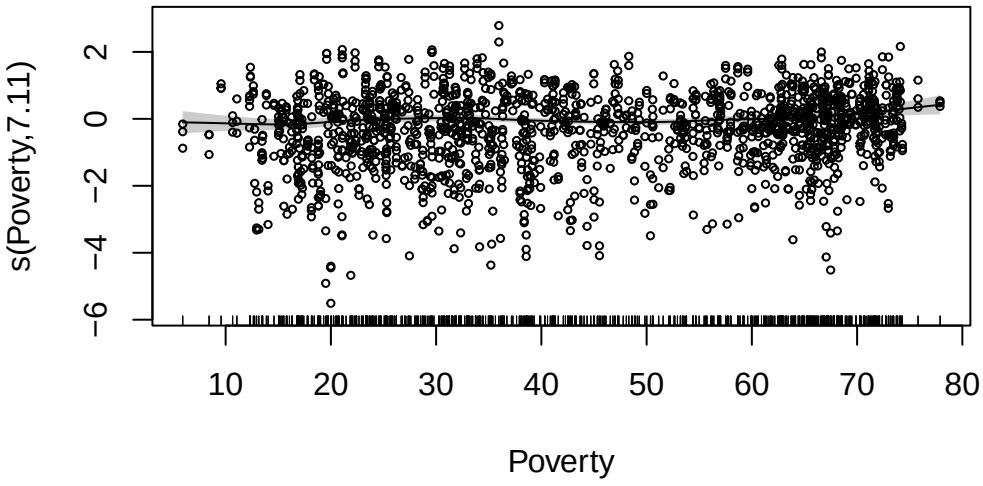
```
plot(nb_model, shade=T, rug = TRUE, residuals = TRUE,
pch = 1, scheme =1,  cex = 0.5)
```

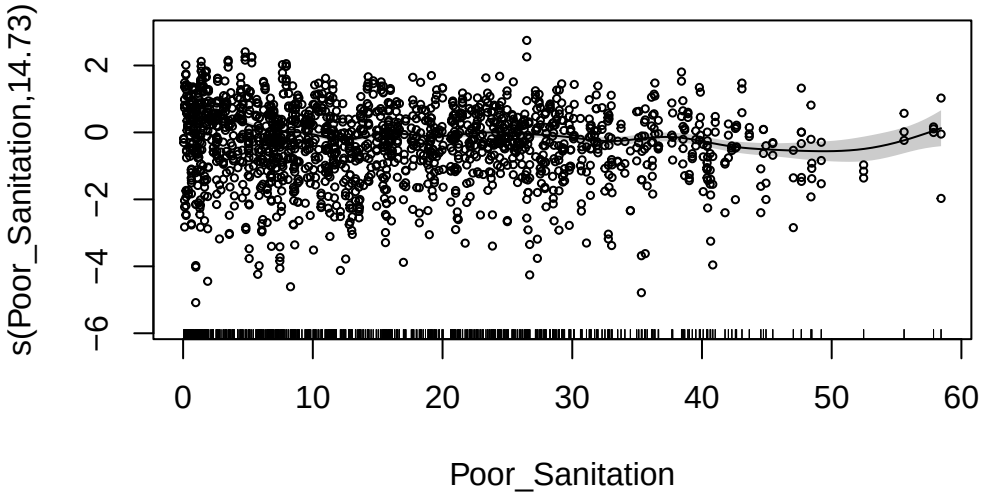


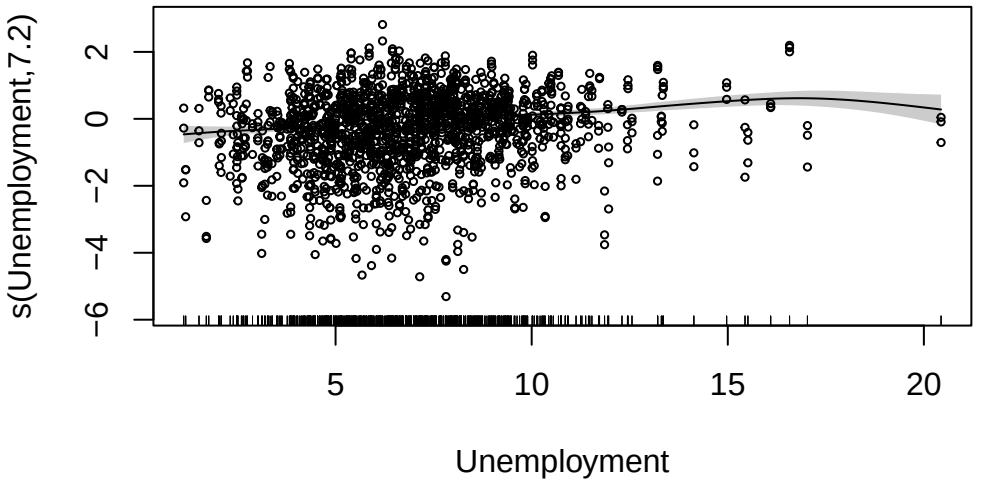

s(Illiteracy,10.89)
−6
−4
−2
0
2
10
20
30
40
Illiteracy

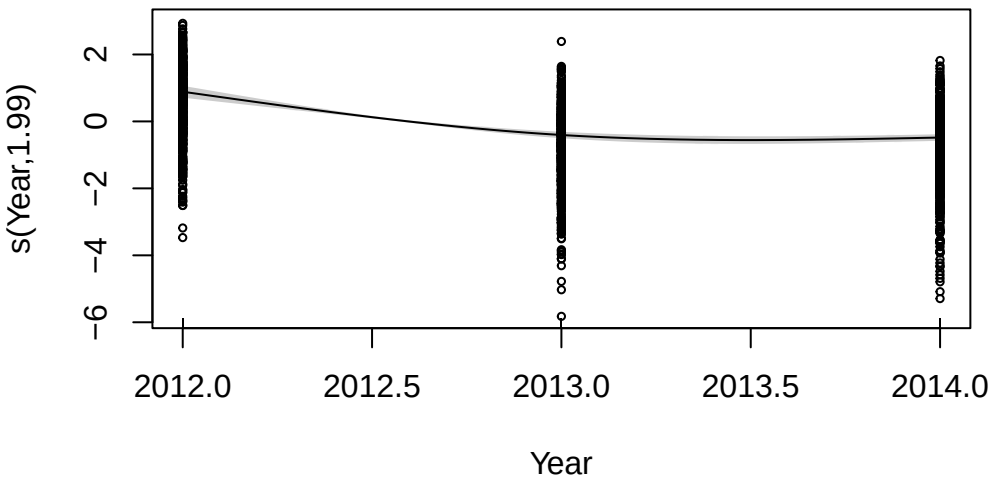

s(Urbanisation,11.07)
−6
−4
−2
0
2
20
40
60
80
100
Urbanisation

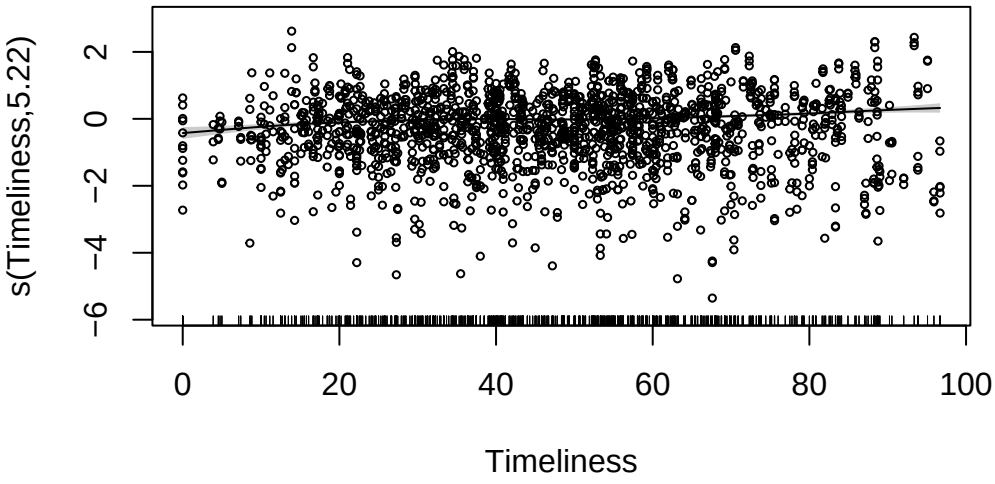

s(Density,11.23)
2
0
−2
−4
−6
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Density

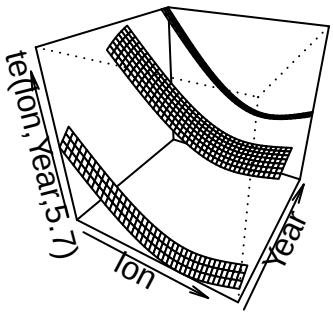

s(Poverty,7.11)
2
0
−2
−4
−6
10
20
30
40
50
60
70
80
Poverty

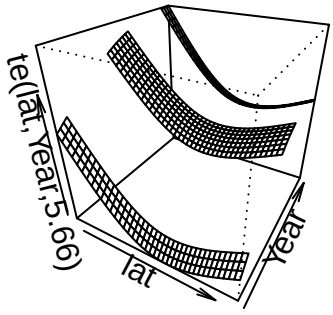

s(Poor_Sanitation,14.73)
−6
−4
−2
0
2
0
10
20
30
40
50
60
Poor_Sanitation


s(Unemployment,7.2)
−6
−4
−2
0
2
5
10
15
20
Unemployment


s(Year,1.99)
−6
−4
−2
0
2
2012.0
2012.5
2013.0
2013.5
2014.0
Year


s(Timeliness,5.22)
−6
−4
−2
0
2
0
20
40
60
80
100
Timeliness





The QQ-plot looks much better for the Negative Binomial model. The majority of points lie either on top of very near the y=x line, except for a few towards the extremes. This indicates

our assumption about the true distribution of the data is a lot more safe than it was before.

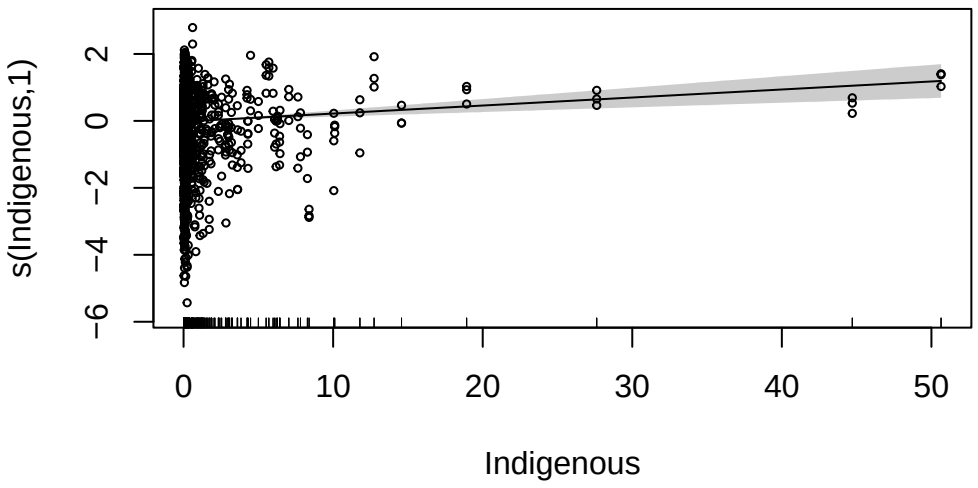
```
#Calculating Pearson estimate for dispersion parameter using Pearson residuals:
sum(residuals(nb_model, type = "pearson")^2) / df.residual(nb_model)
```

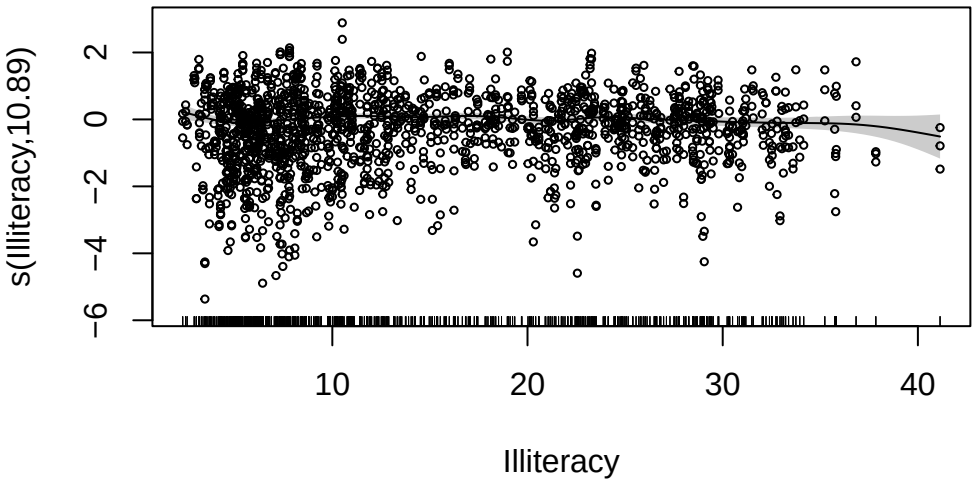
```
[1] 1.338156
```

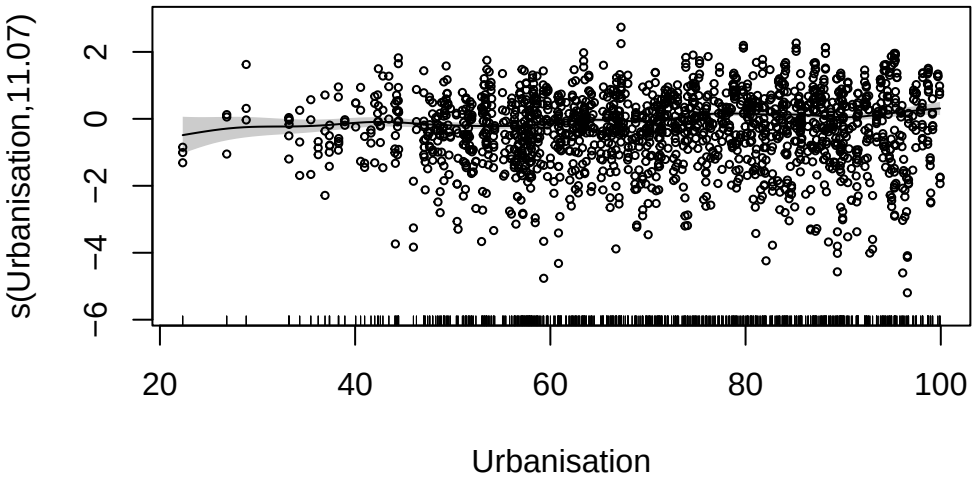
The dispersion parameter is very close to 1, unlike for the Poisson model, meaning that the model that can account for most of the over-dispersion in the data. As such a dispersion parameter value close to 1 can be interpreted as the model is a good fit for the data due to the model adequately capture the variability of the the response variable.

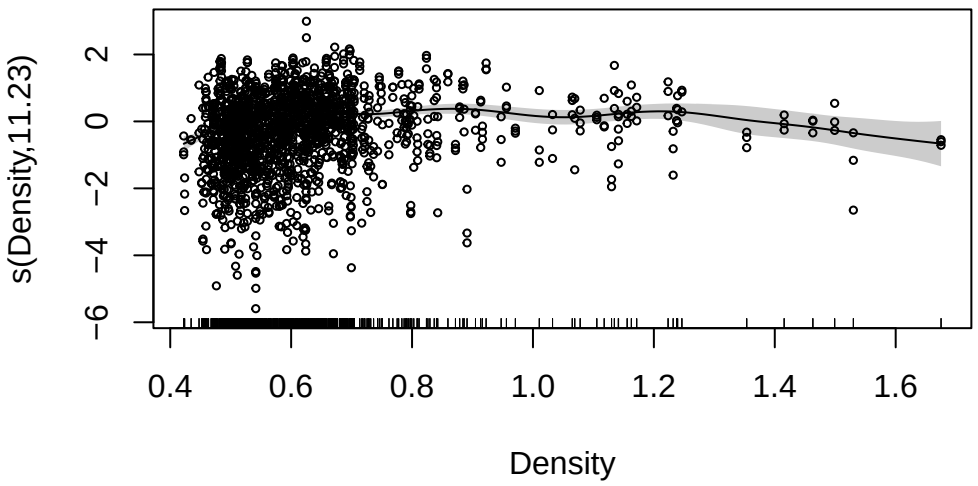
Again tidy up the plots of the Negative Binomial

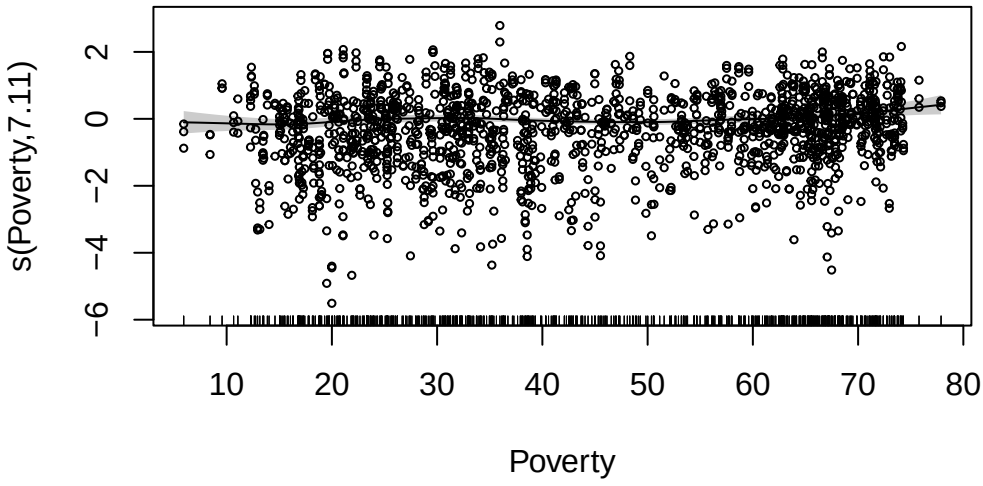
```
plot(nb_model, shade=T, rug = TRUE, residuals = TRUE,scheme=1,
pch = 1, cex = 0.5)
```

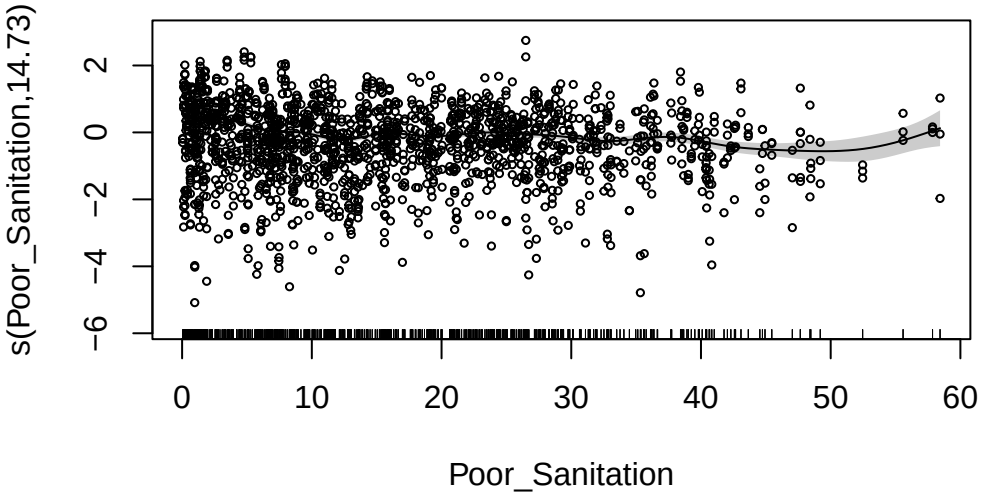


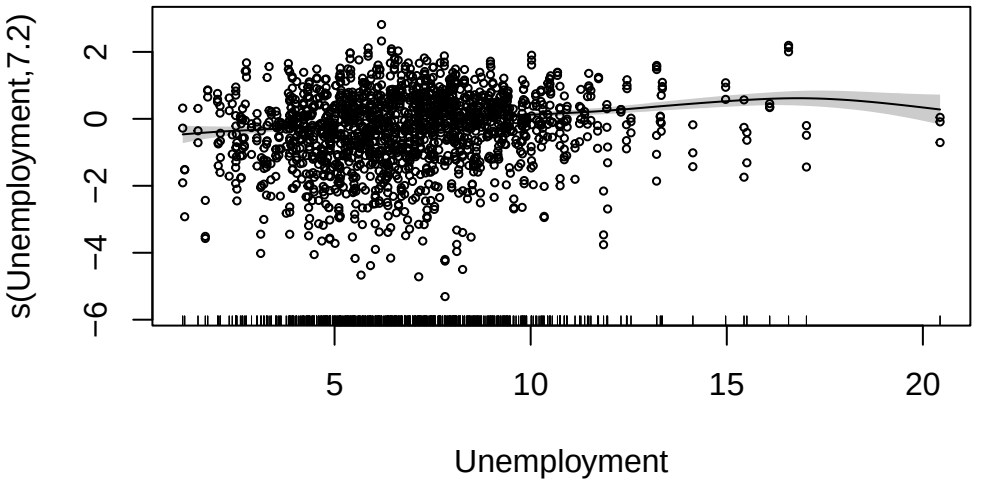

s(Illiteracy,10.89)
−6
−4
−2
0
2
10
20
30
40
Illiteracy

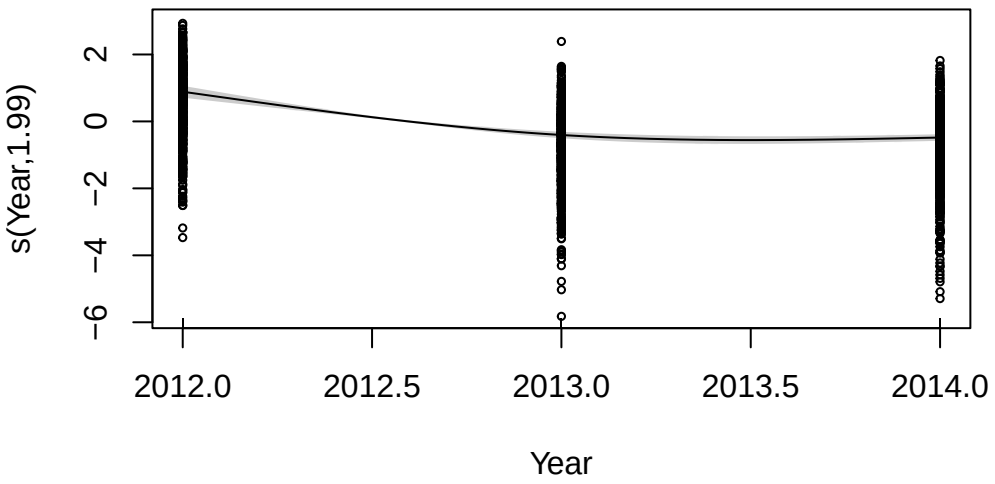

s(Urbanisation,11.07)
−6
−4
−2
0
2
20
40
60
80
100
Urbanisation

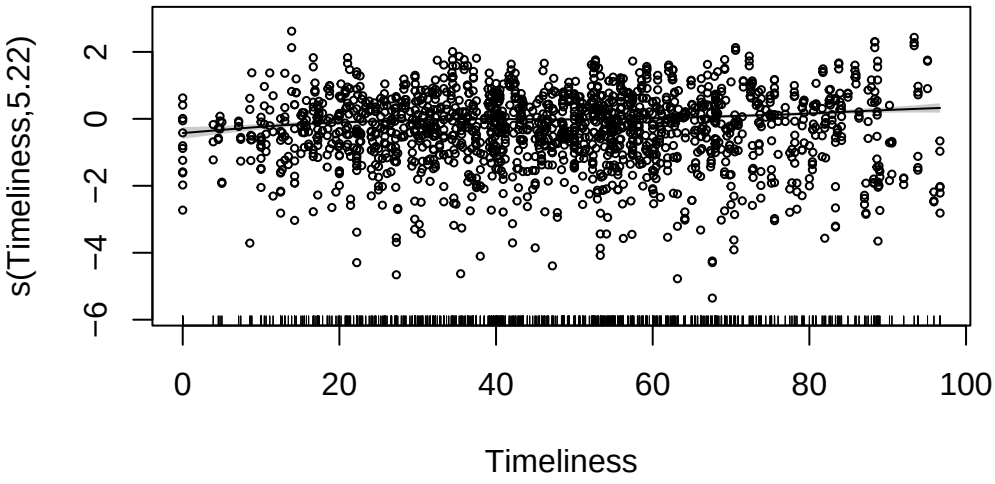

s(Density,11.23)
2
0
−2
−4
−6
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Density

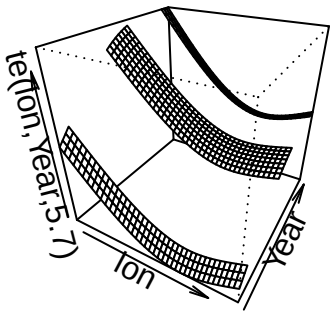

s(Poverty,7.11)
2
0
−2
−4
−6
10
20
30
40
50
60
70
80
Poverty

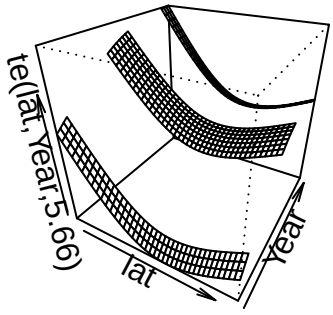

s(Poor_Sanitation,14.73)
−6
−4
−2
0
2
0
10
20
30
40
50
60
Poor_Sanitation


s(Unemployment,7.2)
−6
−4
−2
0
2
5
10
15
20
Unemployment

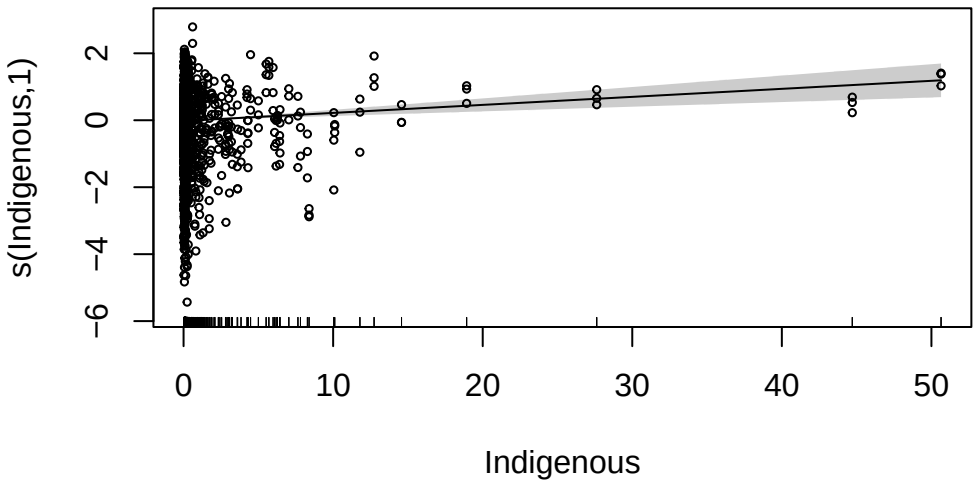

s(Year,1.99)
−6
−4
−2
0
2
2012.0
2012.5
2013.0
2013.5
2014.0
Year

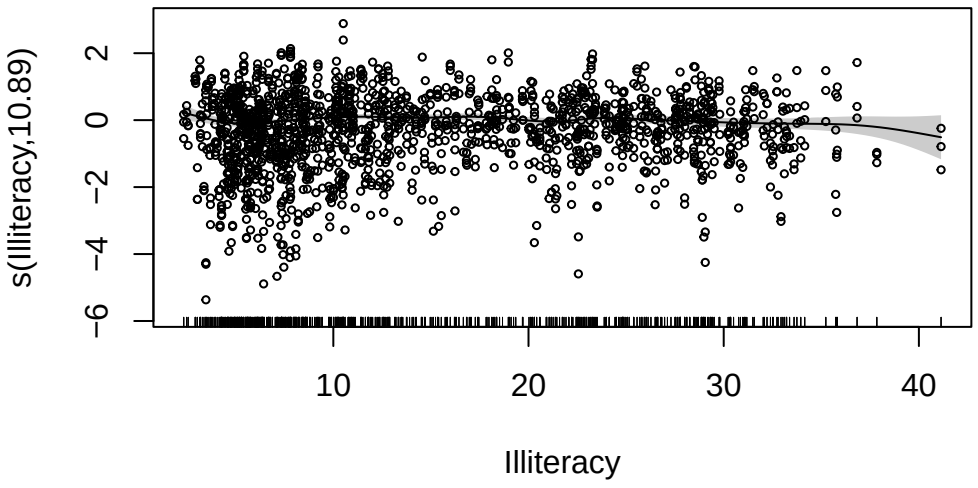

s(Timeliness,5.22)
−6
−4
−2
0
2
0
20
40
60
80
100
Timeliness

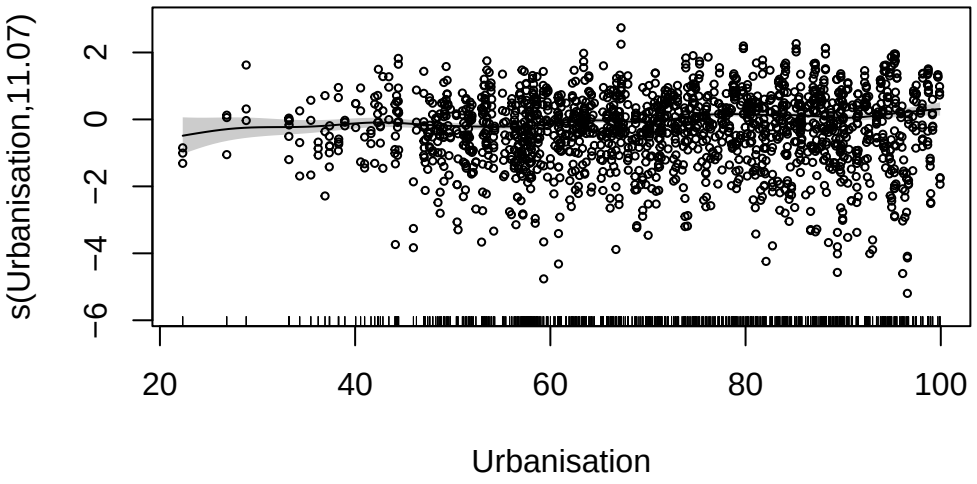

te(lon,Year,5.7)
lon
Year

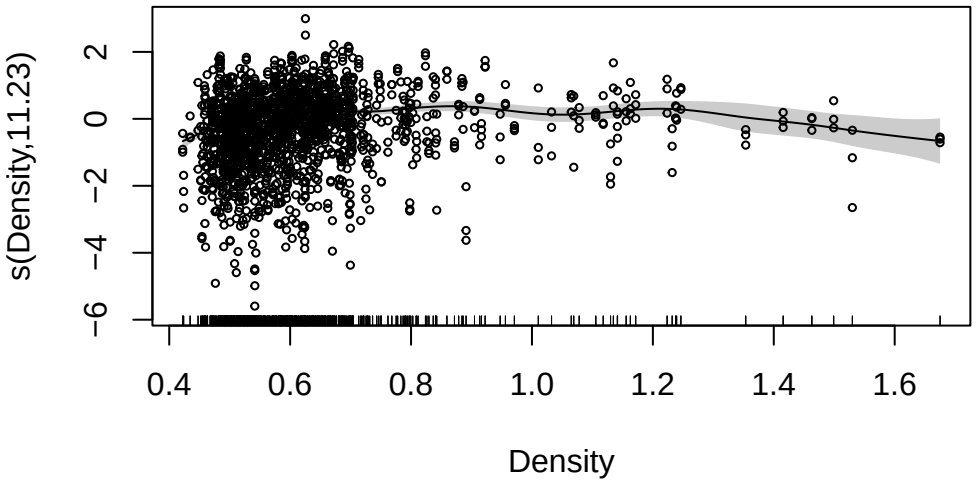

te(lat,Year,5.66)
lat
Year

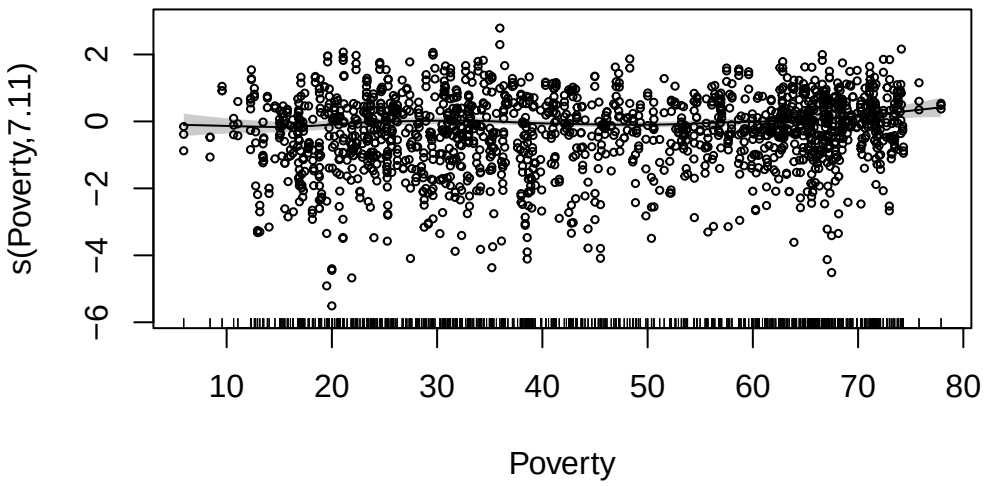
```
plot(nb_model, shade = T, rug = TRUE, residuals = TRUE, scheme = 2, pch = 1,
    cex = 0.5)
```

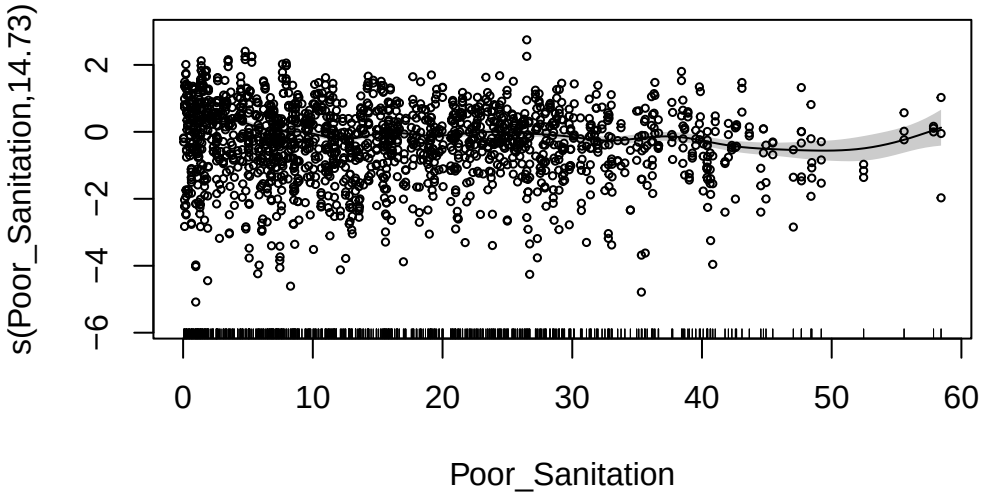


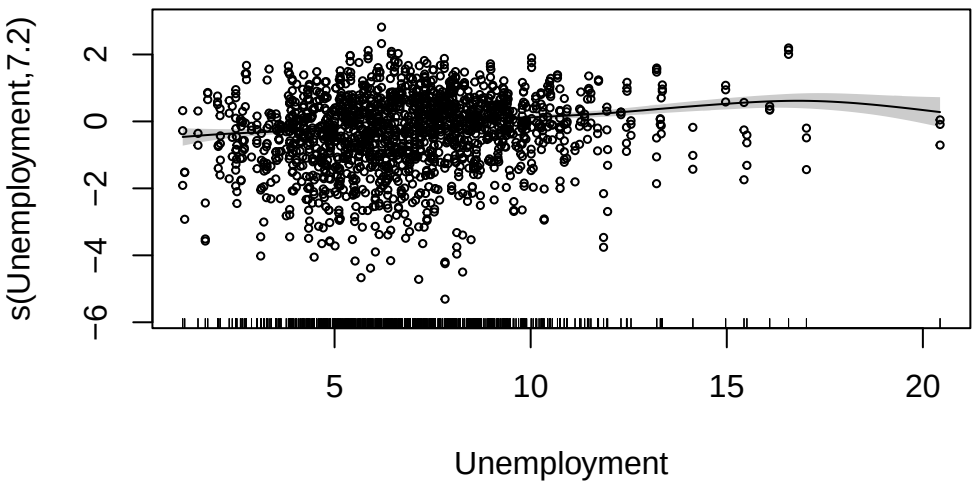


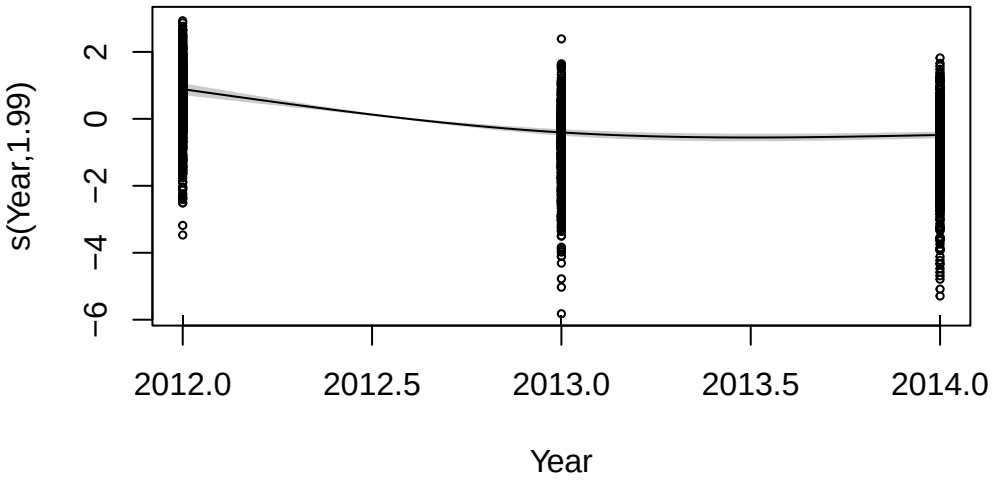

s(Urbanisation,11.07)
−6
−4
−2
0
2
20
40
60
80
100
Urbanisation

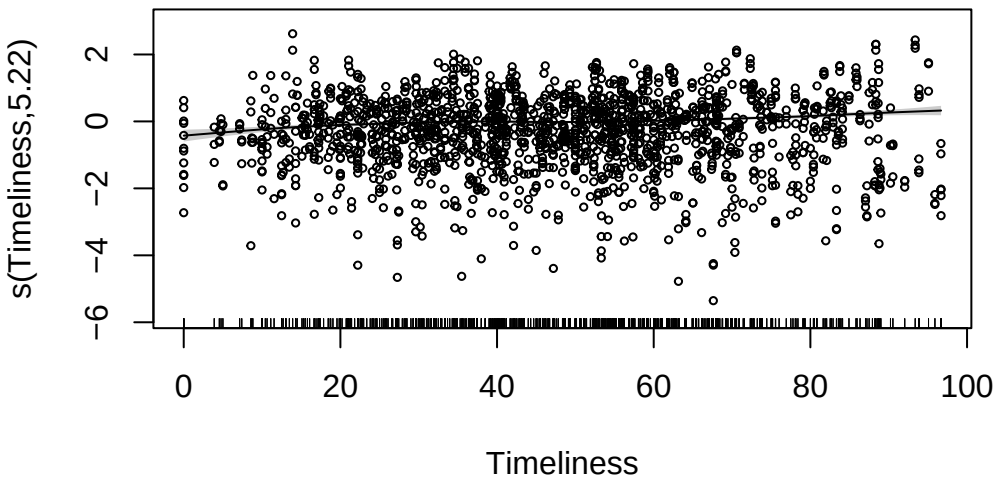

s(Density,11.23)
−6
−4
−2
0
2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Density

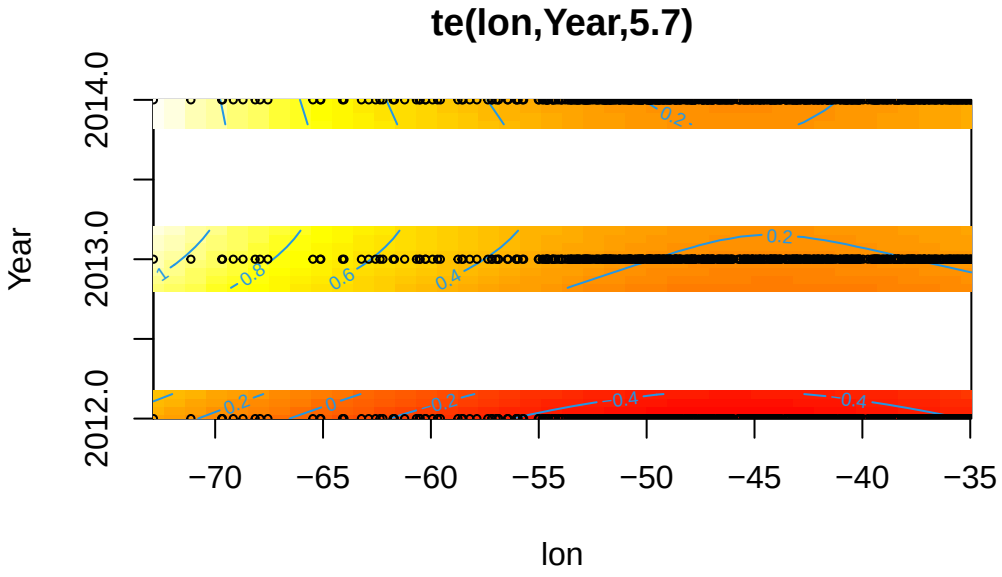

s(Poverty,7.11)
−6
−4
−2
0
2
10
20
30
40
50
60
70
80
Poverty

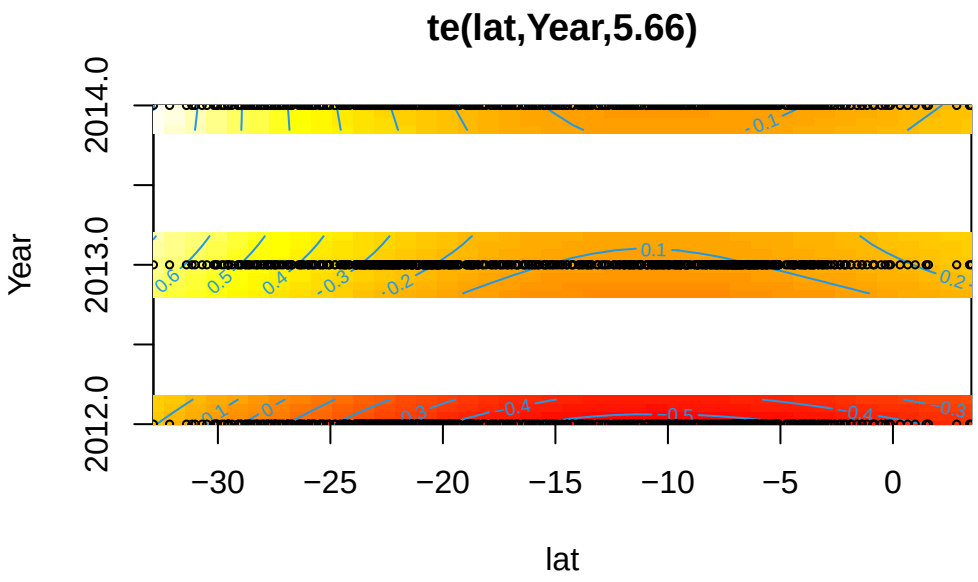

s(Poor_Sanitation,14.73)
−6
−4
−2
0
2
0
10
20
30
40
50
60
Poor_Sanitation


s(Unemployment,7.2)
−6
−4
−2
0
2
5
10
15
20
Unemployment

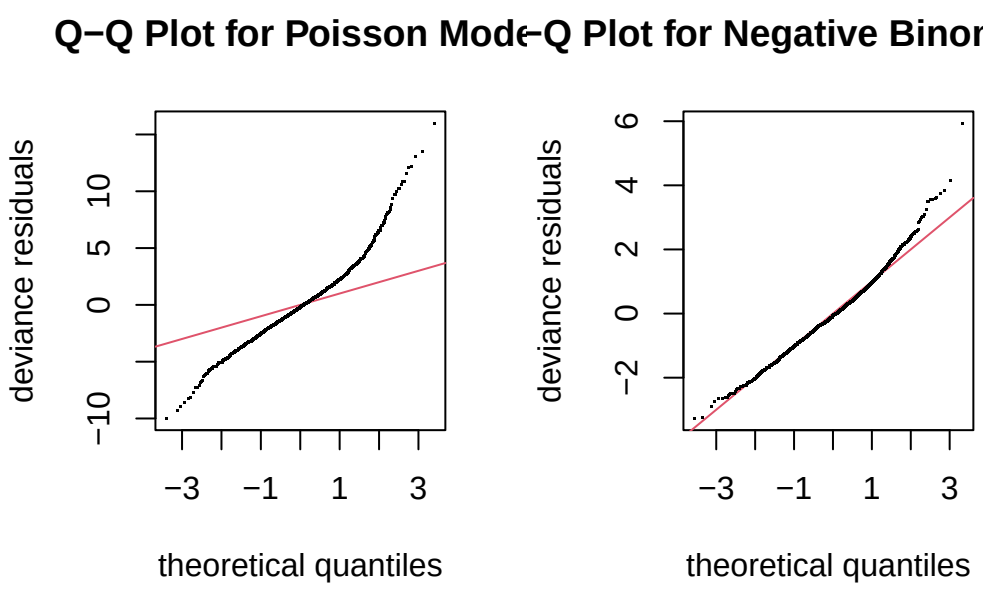

s(Year,1.99)
−6
−4
−2
0
2
2012.0
2012.5
2013.0
2013.5
2014.0
Year


s(Timeliness,5.22)
−6
−4
−2
0
2
0
20
40
60
80
100
Timeliness


te(lon,Year,5.7)
Year
2012.0
2013.0
2014.0
−70
−65
−60
−55
−50
−45
−40
−35
lon

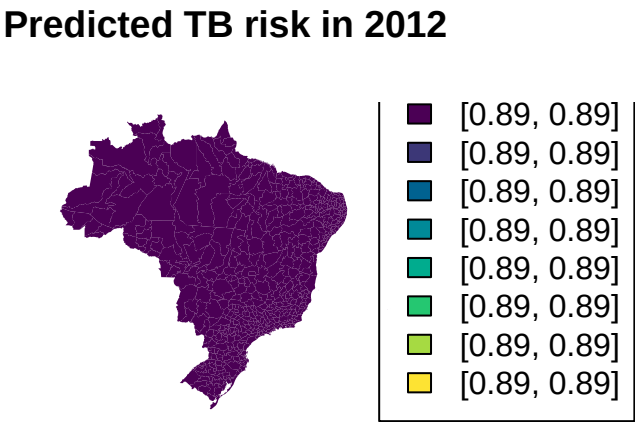



```
par(mfrow = c(1, 2))

qq.gam(poisson_model, main = "Q-Q Plot for Poisson Model")
qq.gam(nb_model, main = "Q-Q Plot for Negative Binomial Model")
```

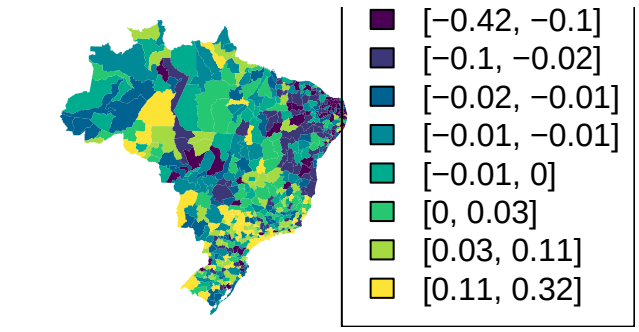



```
preds = predict(nb_model, type = "terms", se.fit = FALSE)
year2012 <- preds[1:557, ]
year2013 <- preds[558:1114, ]
year2014 <- preds[1115:1671, ]
cord.pred_2012 <- as.numeric(year2012[, 8])
cord.pred_2013 <- as.numeric(year2013[, 9])
cord.pred_2014 <- as.numeric(year2014[, 10])
List = list(cord.pred_2012, cord.pred_2013, cord.pred_2014)
cord.preds.mean <- rowMeans(simplify2array(List))
cord.change_1 <- cord.pred_2013 - cord.pred_2012
cord.change_2 <- cord.pred_2014 - cord.pred_2013
```

```
plot.map(cord.pred_2012, n.levels = 8, main = "Predicted TB risk in 2012")
```




```
plot.map(cord.pred_2013, n.levels = 8, main = "Predicted TB risk in 2013")
```



Global Tuberculosis Report 2020. 2020. Genève, Switzerland: World Health Organization.