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 96000 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 and geographic factors in Brazil that may be associated with TB incidence in each microregion. These factors include indigenous population, illiteracy levels, urbanization rate, dwelling density, poverty levels, sanitation levels, unemployment rates, and timeliness of TB case reporting. The dataset also 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.

Simple exploration of our covariates and their potential relationships with the rate of tuberculosis in each microregion of Brazil was carried out before attempting any type of formal statistical analysis or regression on the data. This was done before attempting to draw any conclusions from the data. The correlations that existed between each of our co-variables and the total number of TB cases were analysed with the help of pairplots. Several of the findings were unexpected, such as the observation that a lower degree of sanitation did not appear to be associated with a higher rate of tuberculosis cases. The same was true in regard to the levels of poverty. Nevertheless, the issue with attempting to infer statistical associations in such a straightforward manner is that we are unable to take into consideration the possibility of changes occurring in other variables for each of the data points. This is the primary reason why we need to use a formal model to investigate the impact of our covariates on the incidence of tuberculosis in relation to the total population.

INSERT PAIR PLOT -1 AND SUMMARY-1 TABLE HERE.

While developing statistical models, we are always forced to choose between two competing priorities: interpretability and flexibility. In most cases, we make an effort to fit the data to a linear model whenever that is at all possible. Linear models are simple, and it is straightforward to draw conclusions and interpretations from them. However, the correlations of interest are frequently far too complicated (and non-linear) to be correctly represented by this method, as is the case with the data that we have regarding tuberculosis. On the opposite end of the spectrum is the use of machine learning models such as neural networks or boosted trees. These techniques provide very accurate predictions of modelled relationships; nevertheless, they call for a substantial quantity of data and, more crucially, they are notoriously challenging to interpret. Generalized additive models, often known as GAMs, provide a reasonable middle ground when compared to these other choices, which is why these models were chosen to serve as this analysis' preferred framework.

INSERT BRAZIL MAP HERE.

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, we 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 :

$$Log(\lambda_i) = offset(log(Population_i)) + f_1(Indigenous_i) +$$

$$f_2(Illiteracy_i) + f_3(Urbanisation_i) + f_4(Density_i) + f_5(Poverty_i) +$$

$$f_6(PoorSanitation_i) + f_7(Unemployment_i) + f_8(Timeliness_i) +$$

$$f_9(Year_i) + f_{10}(Year_i, lon_i) + f_{11}(lat_i, Year_i),$$

$$u_i = E[TB_i] \;\; and \;\; TB_i \sim Pois(\lambda_i)$$

Confirming that our model was accurate was the next step in the process of developing our model. Unfortunately, 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 9,33 (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.

#TODO ADD NEGATIVE BINOMIAL LATEX

Let the model be :

$$\log(\mu_i) = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + ... + f_k(x_{ik}) + \log(\phi) + \log(y_i + r_i)$$

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 and edf was not too similar to k'.

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

Model Fitting and Results:

The most difficult aspect of using this model was figuring out how to account for the interplay that exists between time and place. It was required to add space and time as a product smooth, despite the fact that the majority of the other covariates were modelled using univariate smooth functions. After doing some research on the topic, we came to the conclusion that the interaction term should be incorporated into the model as a tensor product smooth utilising the "te" term in the mgcv package. Because space and time are measured on such distinct scales, this particular sort of interaction works best in situations in which the two important variables are on different scales.

The negative binomial gam model output shows the statistical significance of various predictors on the incidence of tuberculosis (TB) in a given population. The model assumes a negative binomial distribution with a log link function, indicating that the response variable, TB, has a count distribution and that the logarithm of the expected counts is modeled as a linear combination of the predictors.

The intercept term in the model is statistically significant ($p<0.001$), indicating that there is a significant difference in the TB incidence rate even when all other predictor variables are set to zero. The negative coefficient on the intercept suggests that there is a baseline level of protection against TB in the population.

Among the predictors, all except s(Year) are statistically significant at the 0.05 level. Indigenous status, Illiteracy, Poverty, and Timeliness have a statistically significant positive association with the incidence of TB. Urbanisation, Density, Poor Sanitation, and Unemployment are also significantly associated with TB but have a negative effect on incidence.

The s(Year) term has a statistically significant association with TB incidence but its effect is not linear. The term is modeled using a spline with a small number of degrees of freedom (k=3). The p-value for the term is <2e-16, indicating a strong association with TB incidence.

The interaction terms, te(lon,Year) and te(lat,Year), which model the effect of longitude and latitude on TB incidence with respect to year, respectively, are also statistically significant. Both terms have positive effects on TB incidence, indicating that the risk of TB increases with increasing longitude or latitude, depending on the year.

Overall, the model has an adjusted R-squared value of 0.895, indicating that the predictors in the model explain 53.4% of the variability in TB incidence. The scale estimate is 1, suggesting that the model is well-calibrated. The negative binomial distribution is appropriate for modeling the count response variable, and the log link function is appropriate for modeling the logarithm of the expected counts.

Model Evaluation and Discussion:

First, lets discuss the spatial findings of our analysis. As we can see from the yearly graph, figure 4, the analysis shows that there is a gradual decrease of TB cases per capita from 2012 to 2013 at the national level. In contract from 2013 to 2014 the data displays very little decrease of TB risk on a national level.

Secondly,

Thirdly, merging the 2 previous analysis into the spatio-temporal

TODO reference the temporal graph here

Conclusion:

In summary, the gam model identifies several significant predictors of TB incidence, including Indigenous status, Illiteracy, Poverty, and Timeliness, as well as Urbanisation, Density, Poor Sanitation, and Unemployment. The interaction terms with latitude and longitude also play a significant role in predicting TB incidence. The model can be used to identify populations at risk for TB and to develop interventions to reduce TB incidence in these populations.

In conclusion, we discuss some of the restrictions imposed in this study. The relatively little time frame under consideration presents the most evident limitation of the study. If determining any temporal structure of systematic risk was the objective, then additional time ought

to elapse prior to the data being examined and modelled in order to allow for the passage of time. Because the incidence and rate of tuberculosis are probably affected by a wide variety of other factors that are not accounted for in our dataset, additional covariates could also be added to the study. Additional limitations of our study are associated with the application of a generalised additive model (GAM), including the risk that our model overfits our data, as well as its high computing cost and complexity.

Figures

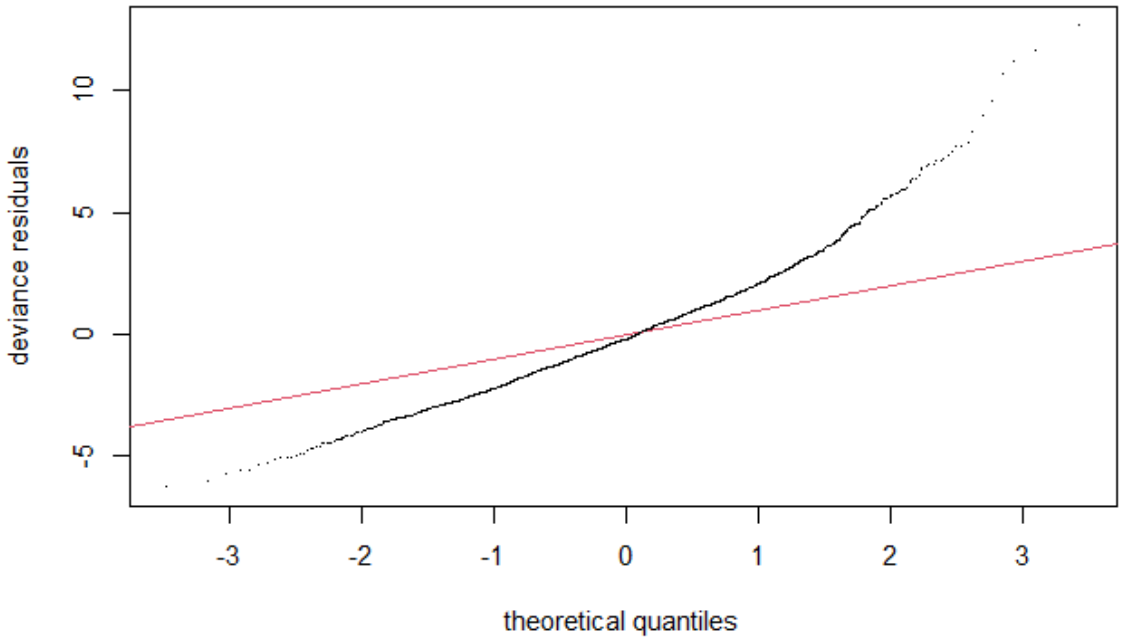
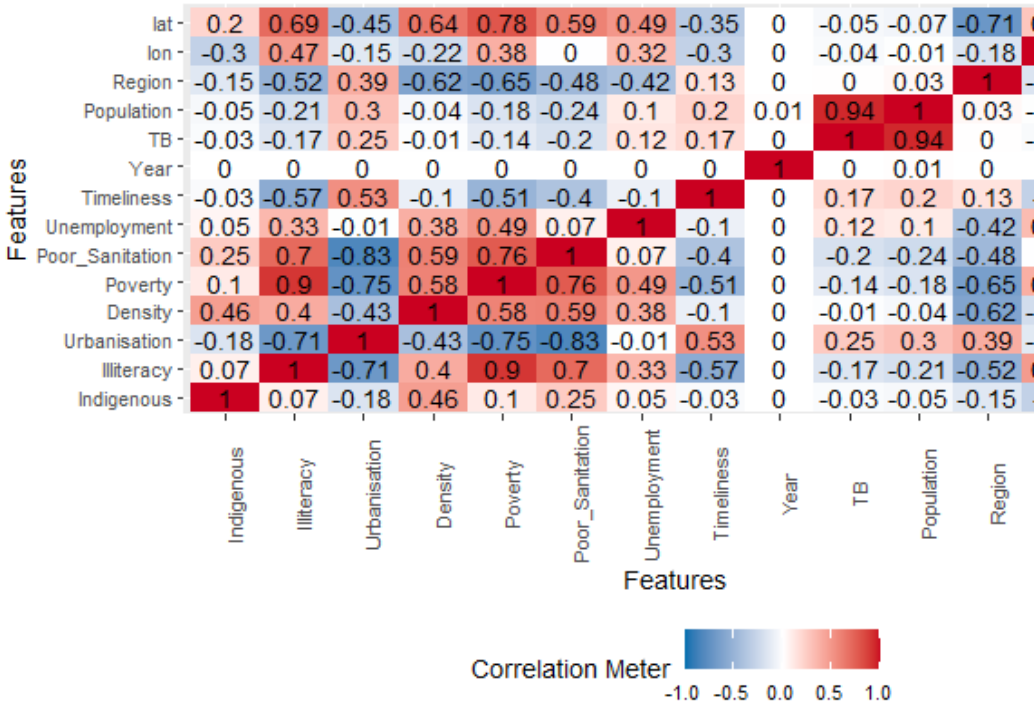


Figure 1: QQ Plot



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

Code Appendix

time series graph

```
TBdata$Risk = (TBdata$TB / TBdata$Population) * 100

## 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")
```

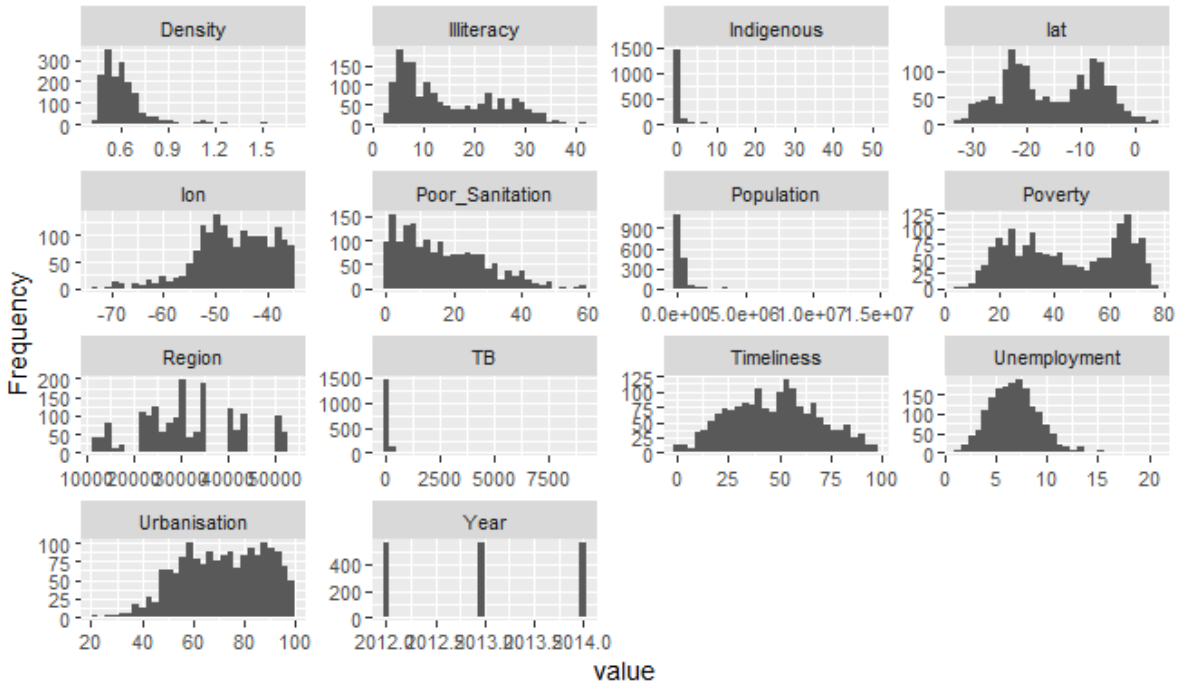


Figure 2: Initial variable analysis

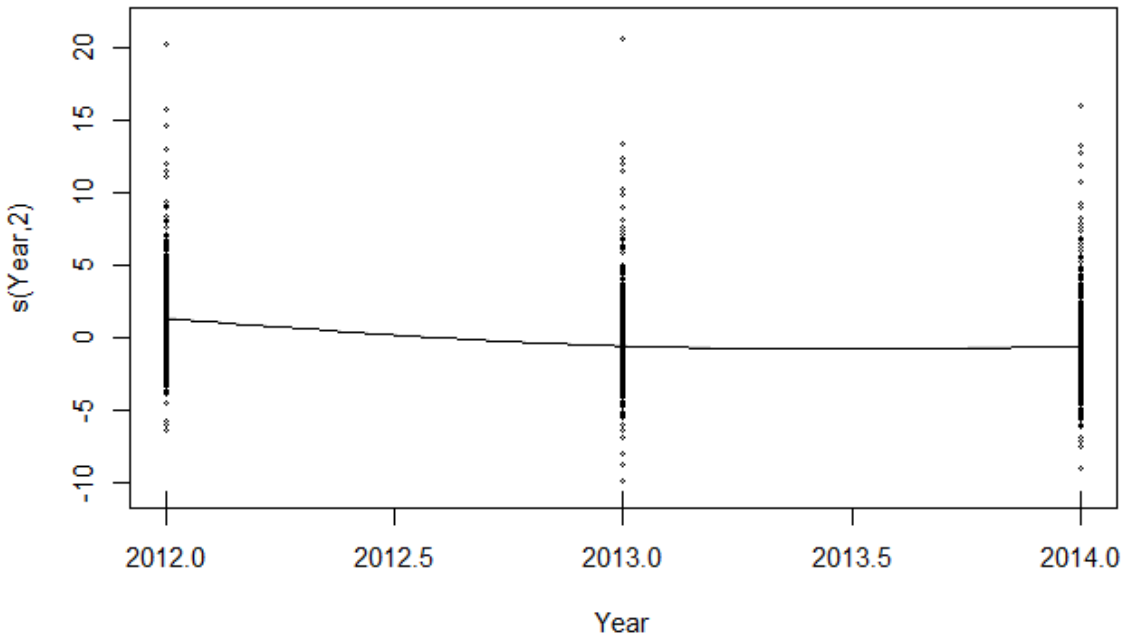


Figure 3: Temporal Analysis of risk of TB per Year

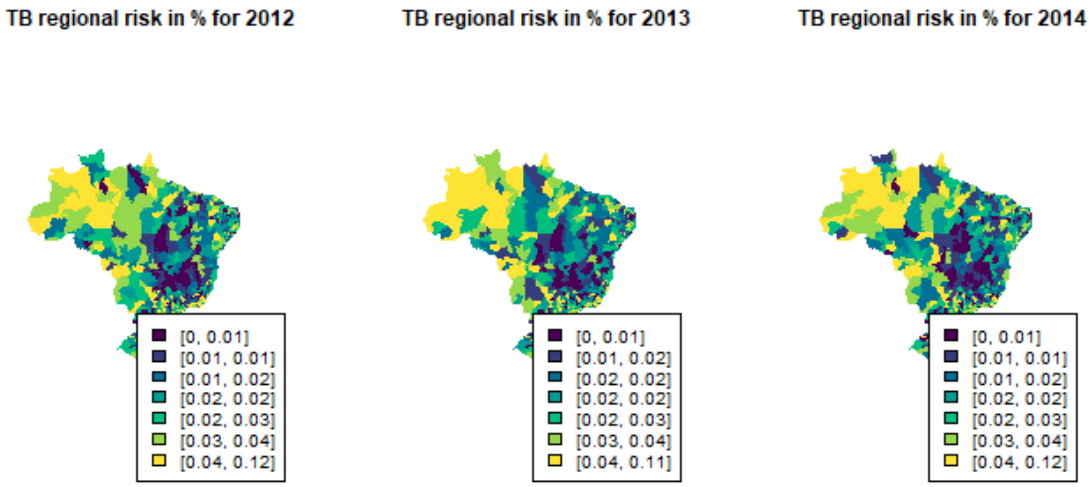
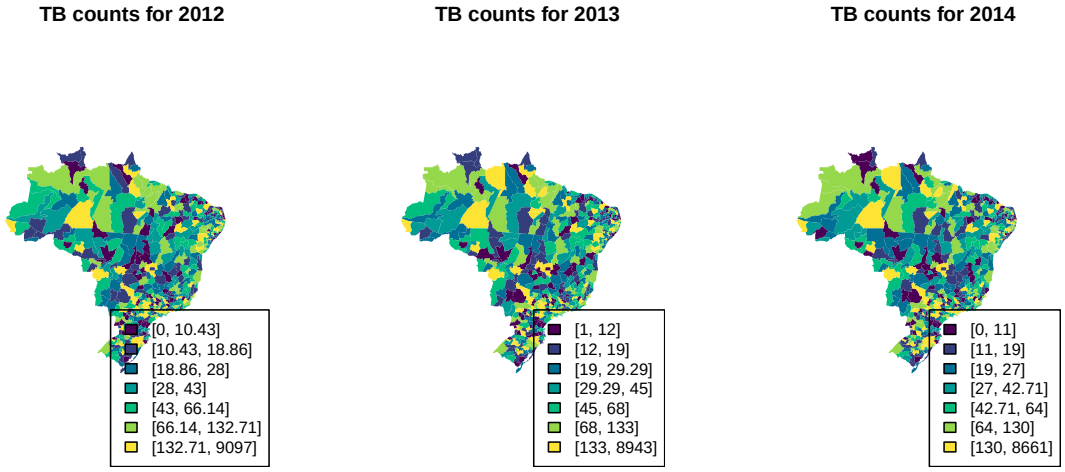
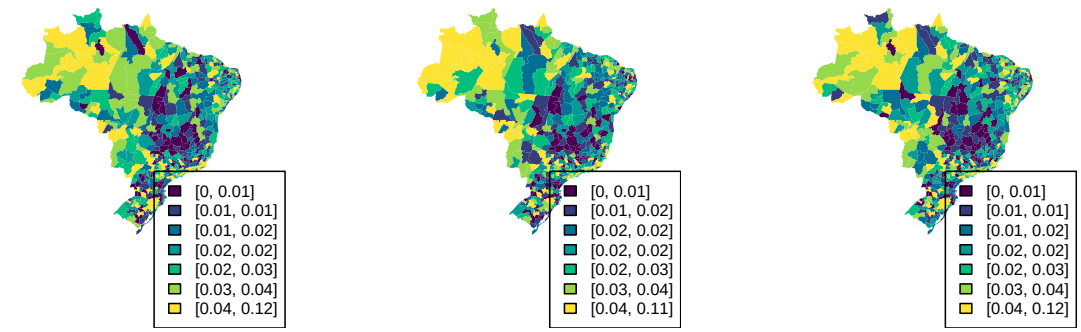


Figure 4: Temporal-Spatial Analysis of risk of TB per Year per region




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



This last one if for the temporal-spacial analysis ## TODO Get this graph for space-time

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
% Thu Mar 23 18:43:39 2023
\begin{table}[ht]
\centering
\begin{tabular}{rlllllllllllllll}
  \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}
```

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 Model 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

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

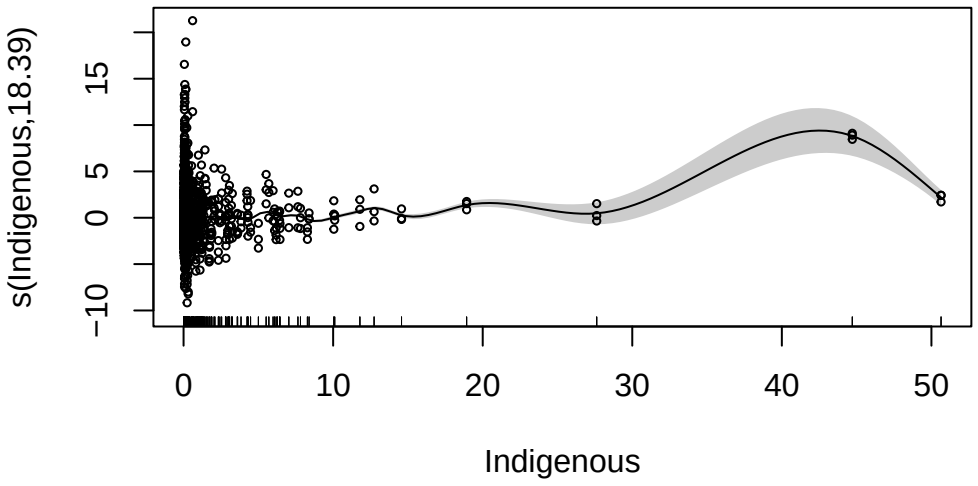
```
Family: poisson
Link function: log

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
```

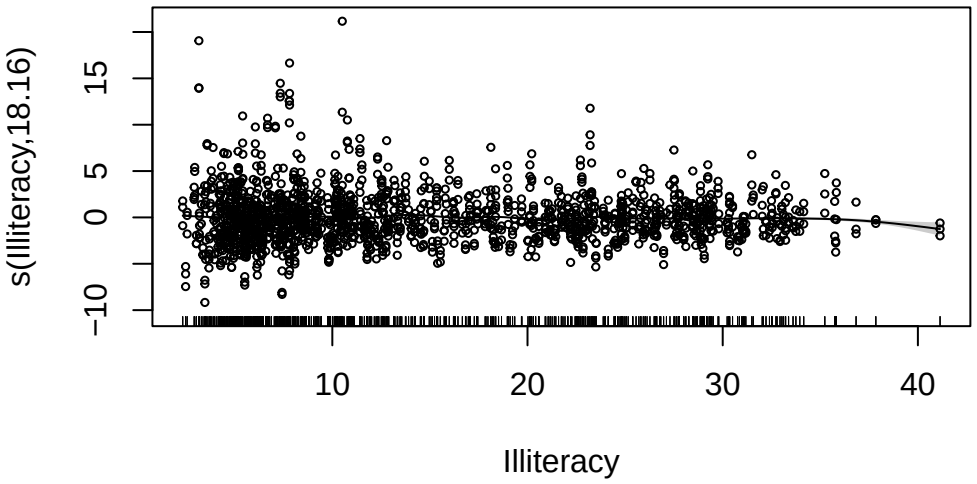
```r
gam.check(poisson_model)
```

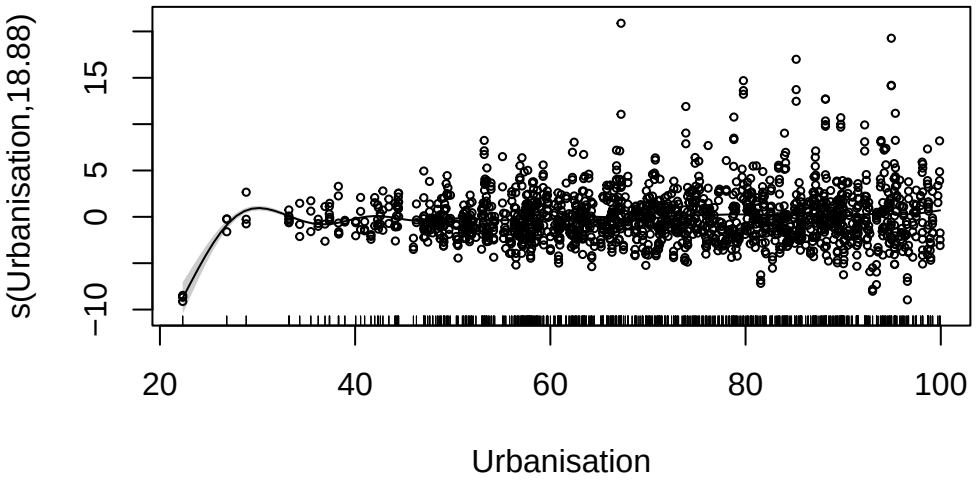
```
#Akaike Information Criterion:
poisson_model$aic
```

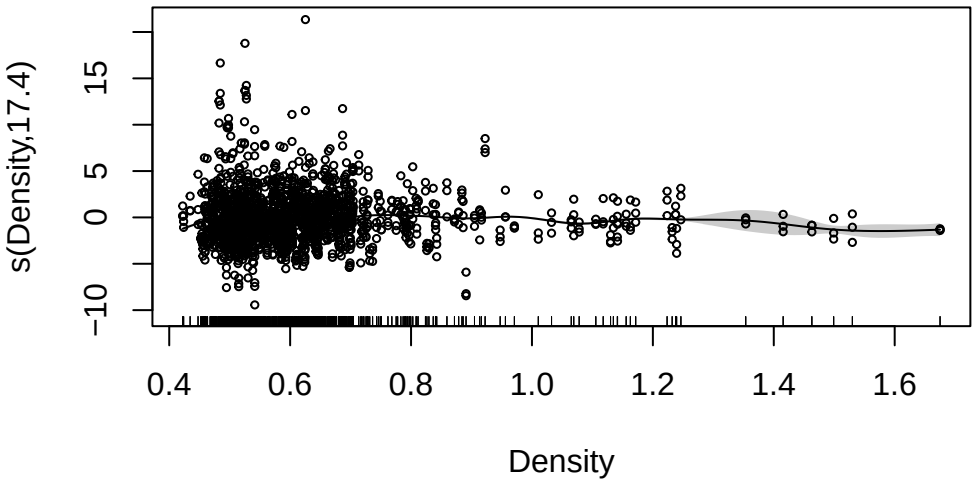
```
[1] 22271.66
```

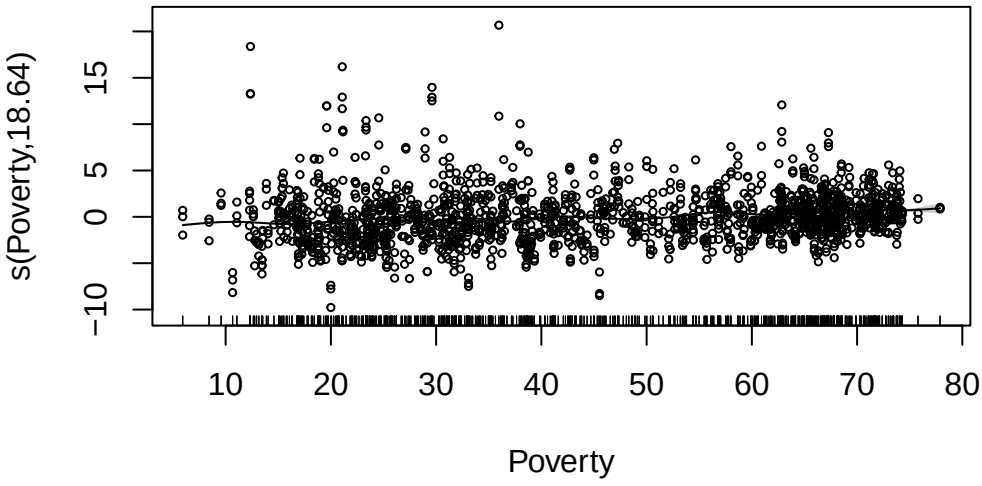
```
plot(poisson_model, shade=T, rug = TRUE, residuals = TRUE,
pch = 1, cex = 0.5)
```

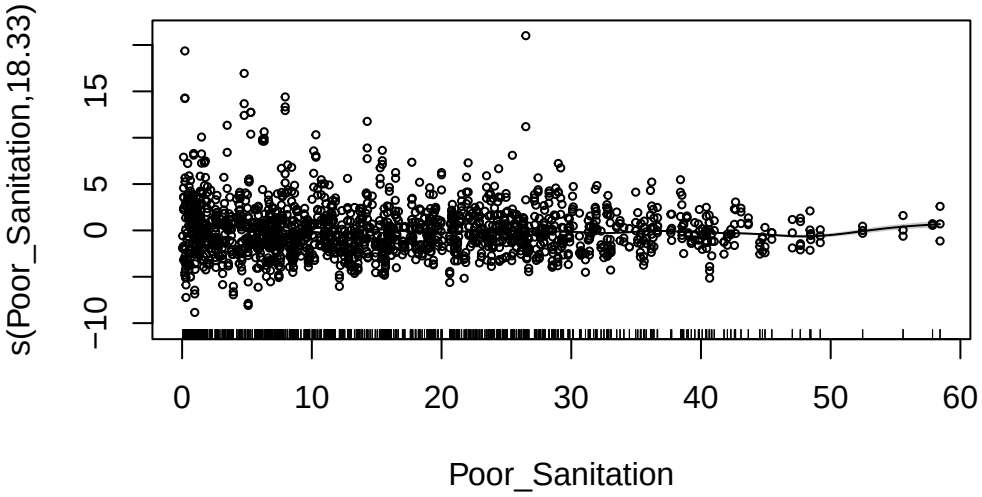


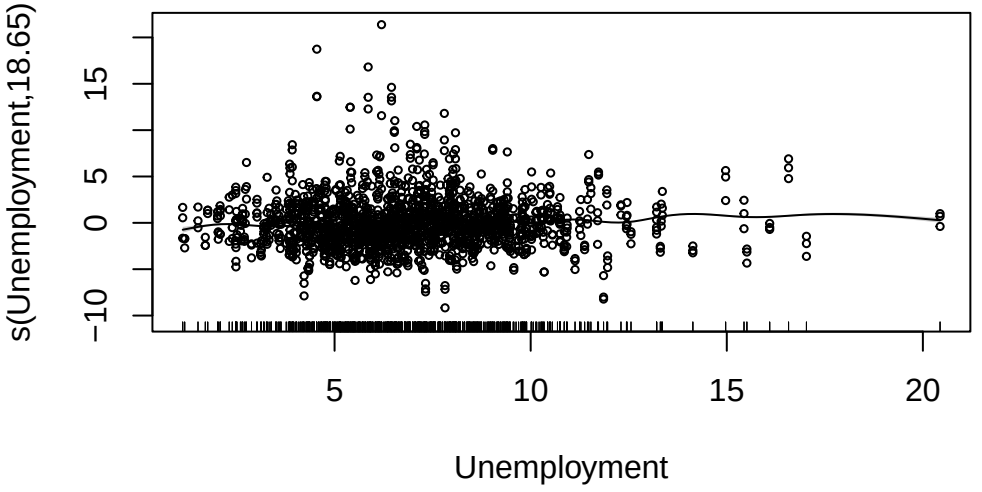

s(Illiteracy,18.16)
−10
0
5
15
10
20
30
40
Illiteracy

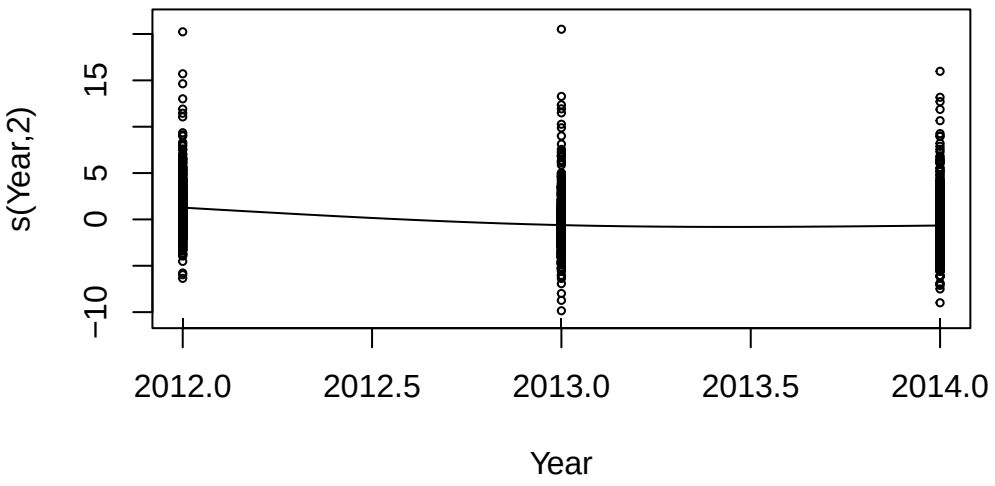

s(Urbanisation,18.88)
−10
0
5
15
20
40
60
80
100
Urbanisation

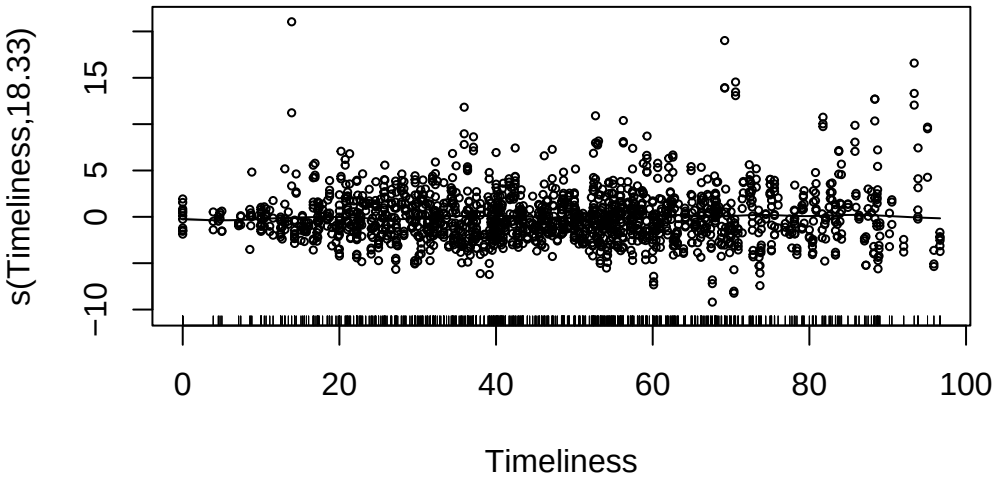

s(Density,17.4)
−10
0
5
15
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Density

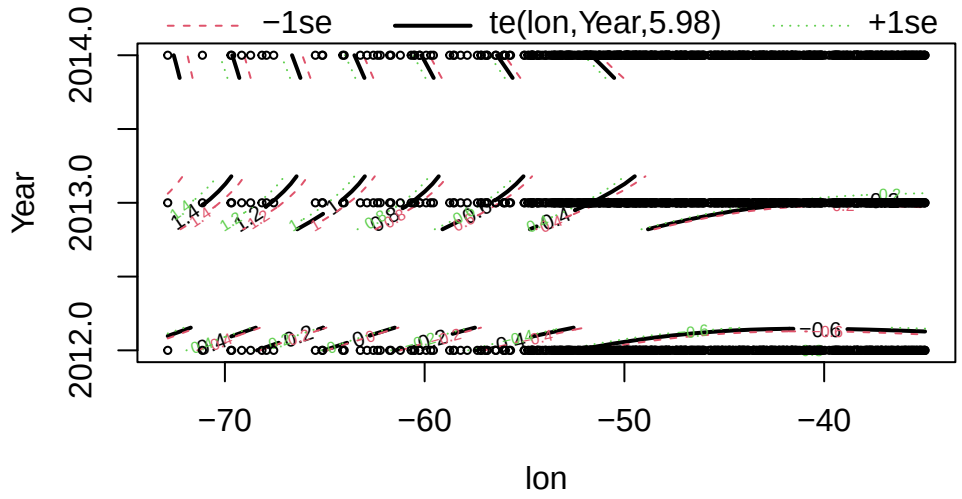

s(Poverty,18.64)
−10
0
5
15
10
20
30
40
50
60
70
80
Poverty

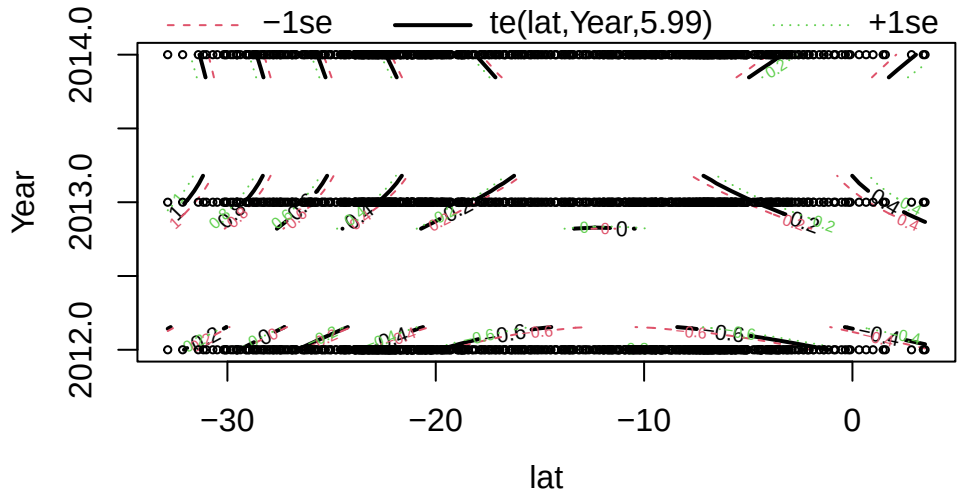

s(Poor_Sanitation,18.33)
−10
0
5
15
0
10
20
30
40
50
60
Poor_Sanitation


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
```

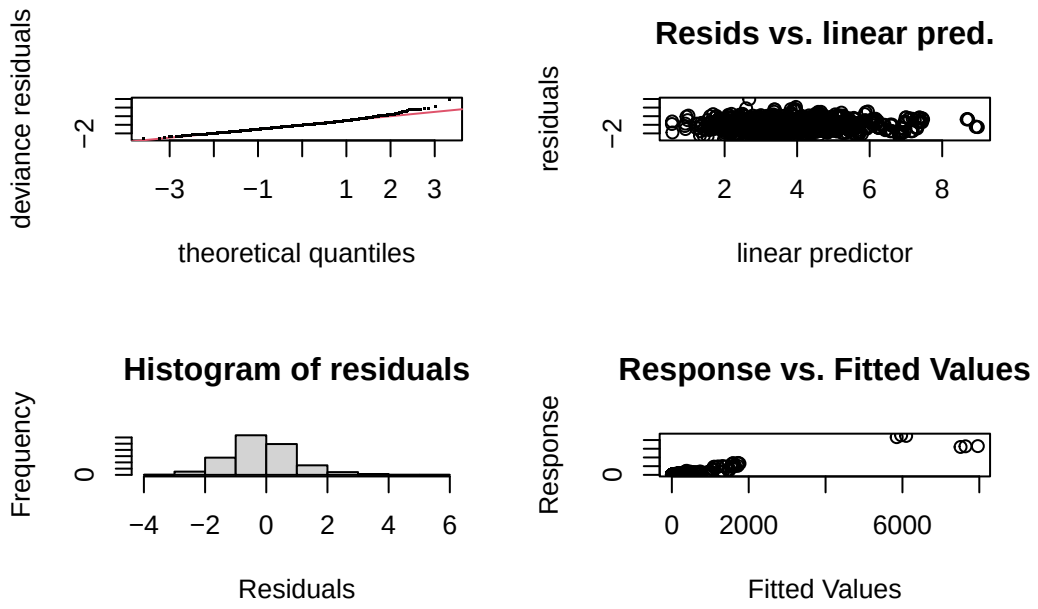
```
#Akaike Information Criterion
nb_model$aic
```

```
[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)
```

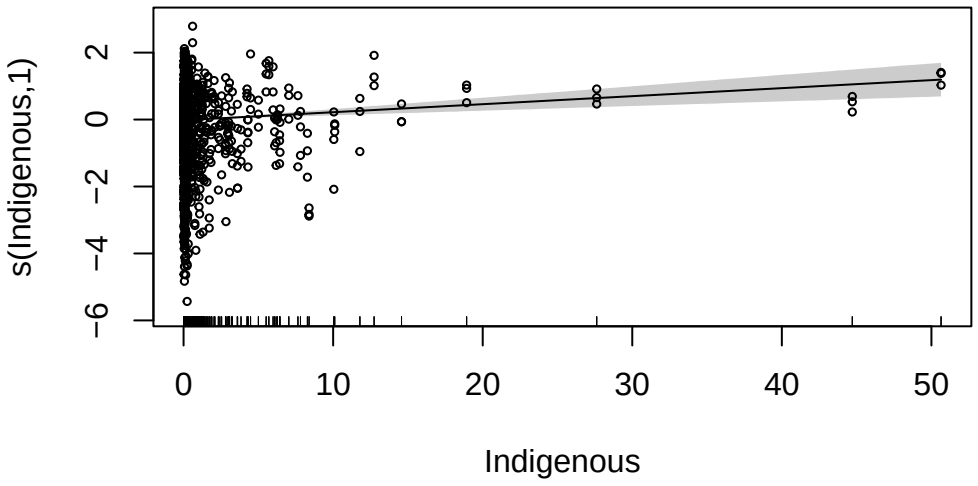


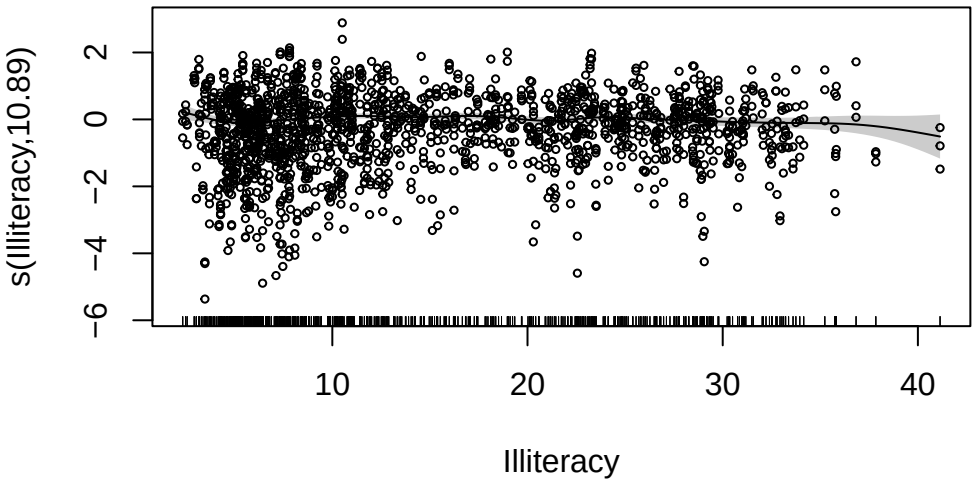

```
Method: REML   Optimizer: outer newton
full convergence after 12 iterations.
Gradient range [-0.0002588633,5.095501e-05]
(score 7203.166 & scale 1).
Hessian positive definite, eigenvalue range [0.0002588044,2.739525].
Model rank =  167 / 167

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    0.19
te(lat,Year)        6.00  5.66    0.96    0.16
---
```

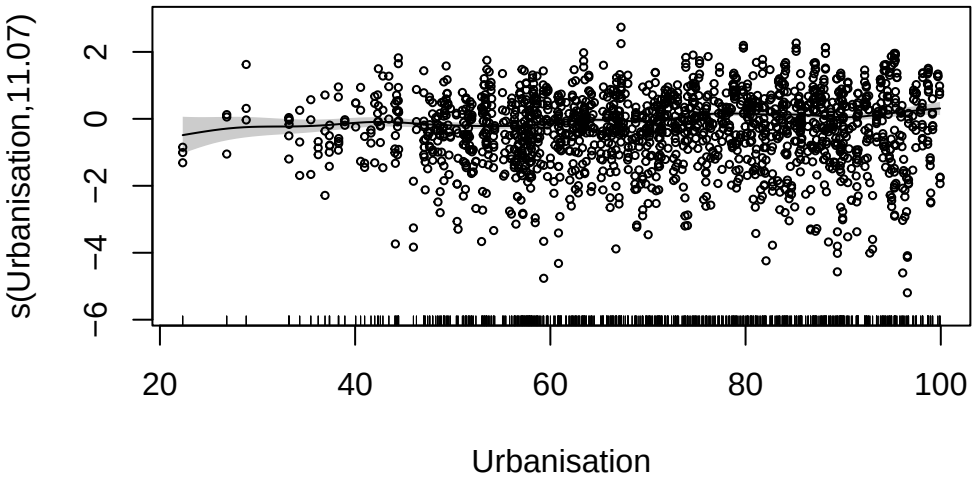
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

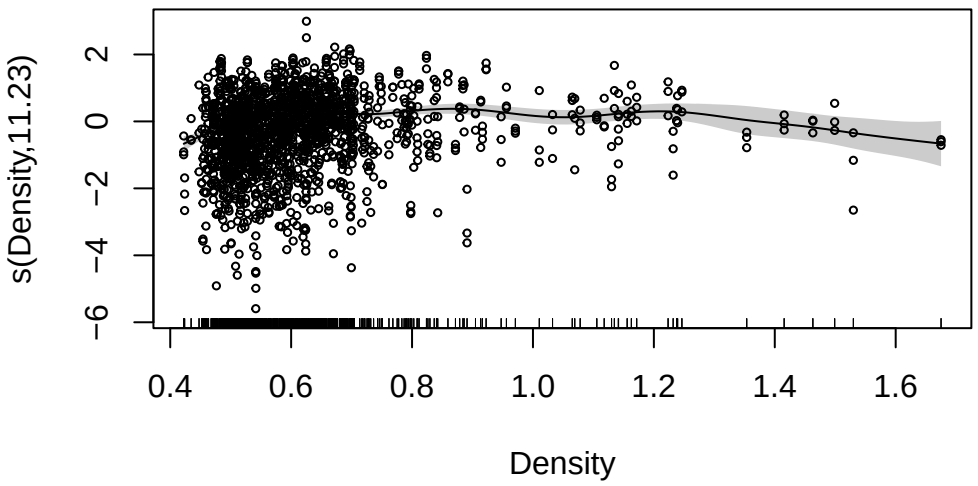
As we can see from the residual versus predictor plot, the values seem to be randomly scattered with no clear trend but with some distance from the zero line. As such we can determine that this scatter is due to random errors and not a uncounted pattern in the model.

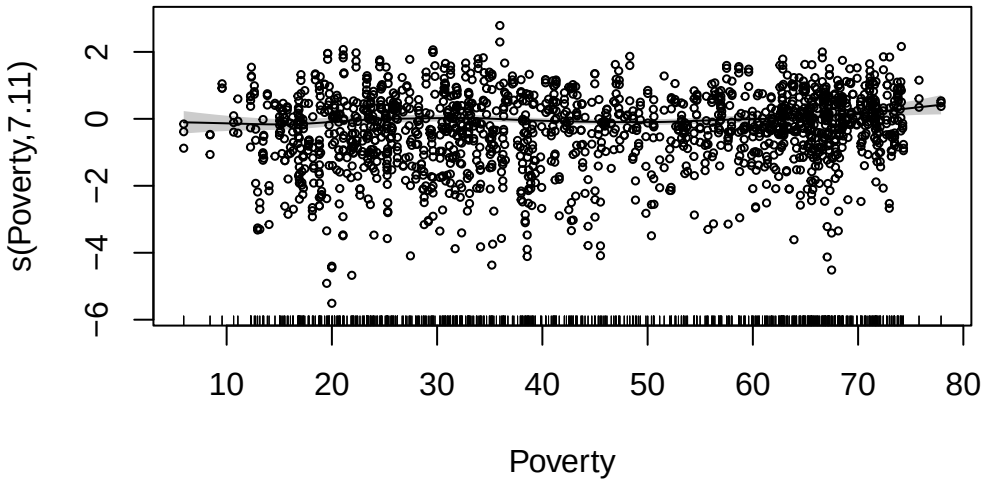
```
plot(nb_model, shade=T, rug = TRUE, residuals = TRUE,
pch = 1, scheme =1,  cex = 0.5)
```

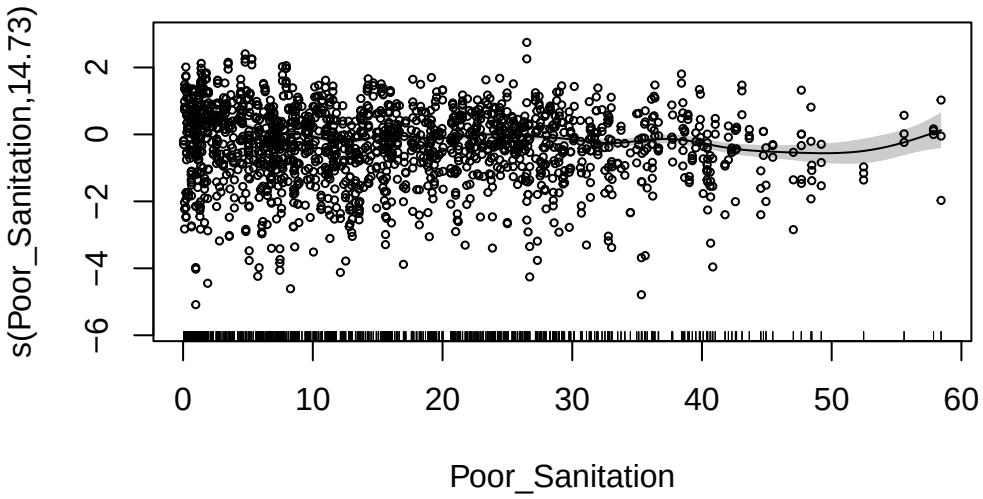


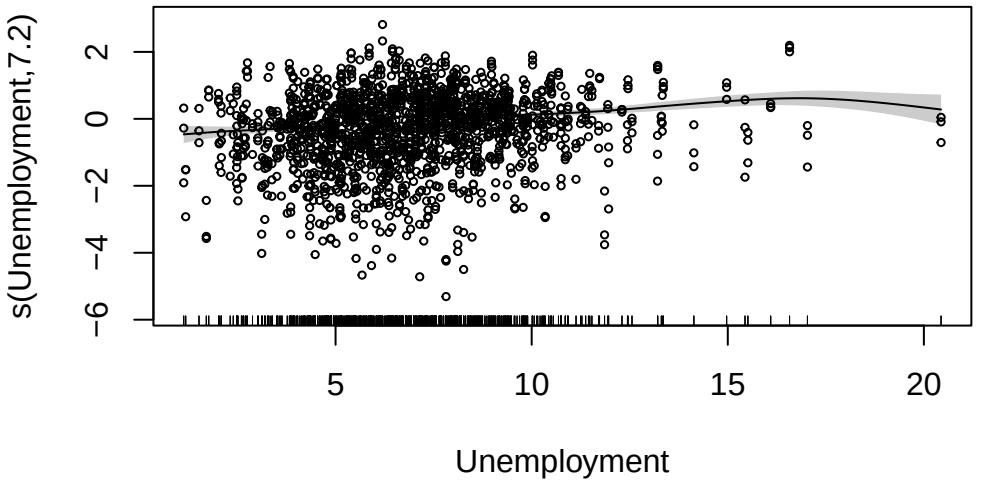

s(Illiteracy,10.89)
−6
−4
−2
0
2
10
20
30
40
Illiteracy

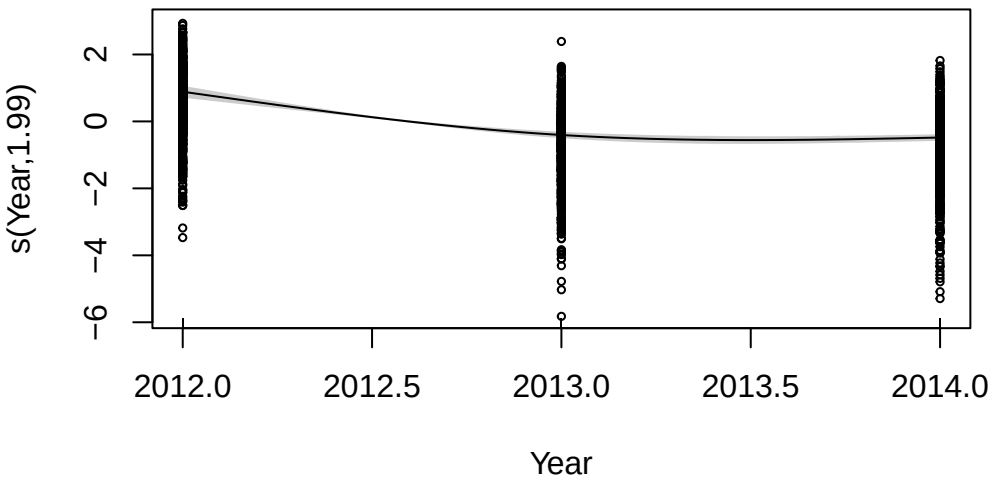

s(Urbanisation,11.07)
−6
−4
−2
0
2
20
40
60
80
100
Urbanisation

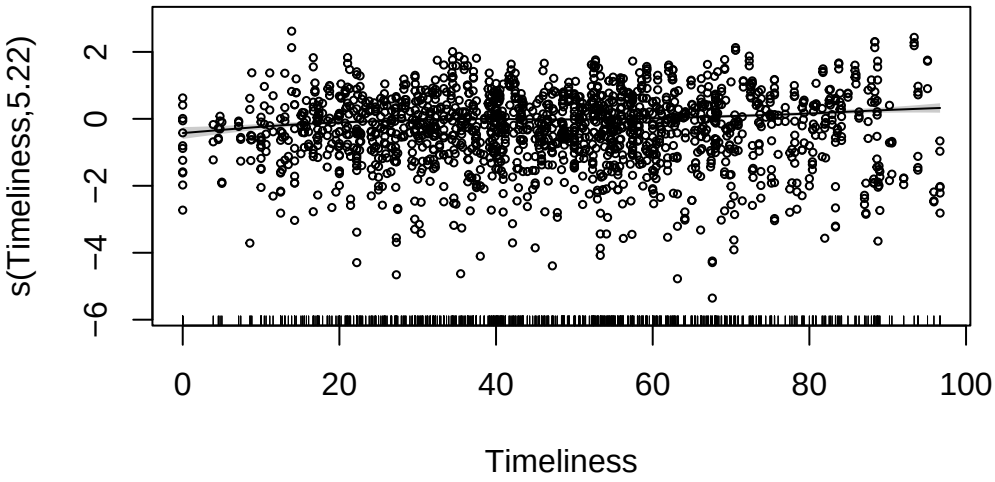

s(Density,11.23)
−6
−4
−2
0
2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Density

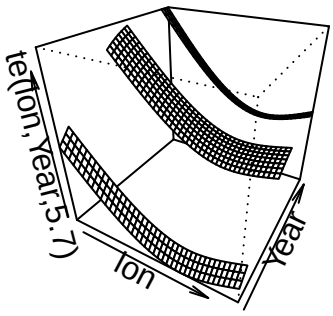

s(Poverty,7.11)
−6
−4
−2
0
2
10
20
30
40
50
60
70
80
Poverty

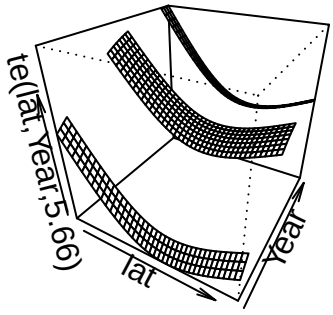

s(Poor_Sanitation,14.73)
−6
−4
−2
0
2
0
10
20
30
40
50
60
Poor_Sanitation


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


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


s(Timeliness,5.22)
2
0
−2
−4
−6
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

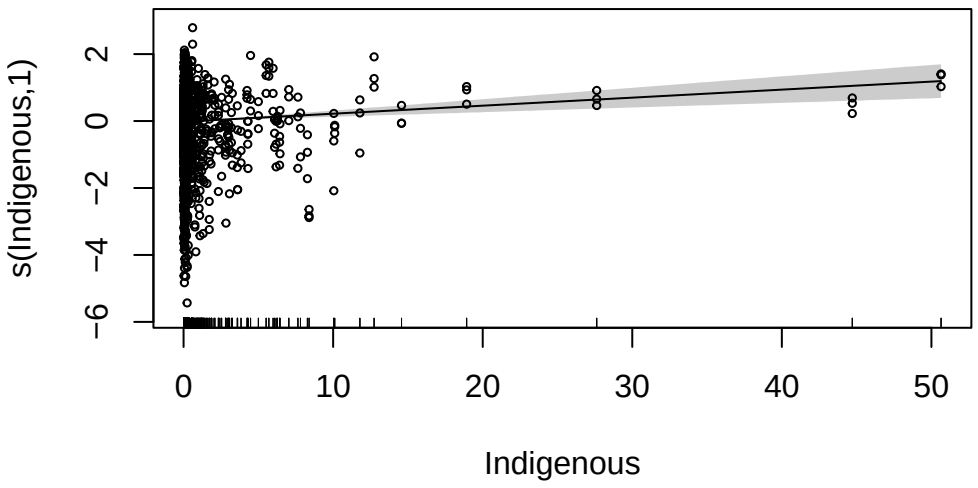
our assumption about the true distribution of the data is a lot more safe than it was before.

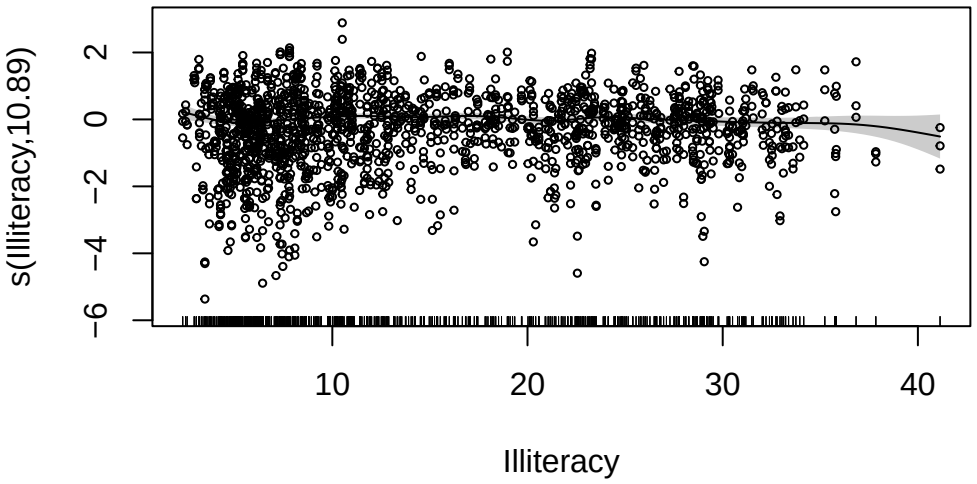
```
#Calculating Pearson estimate for dispersion parameter using Pearson residuals:
sum(residuals(nb_model, type = "pearson")^2) / df.residual(nb_model)
```

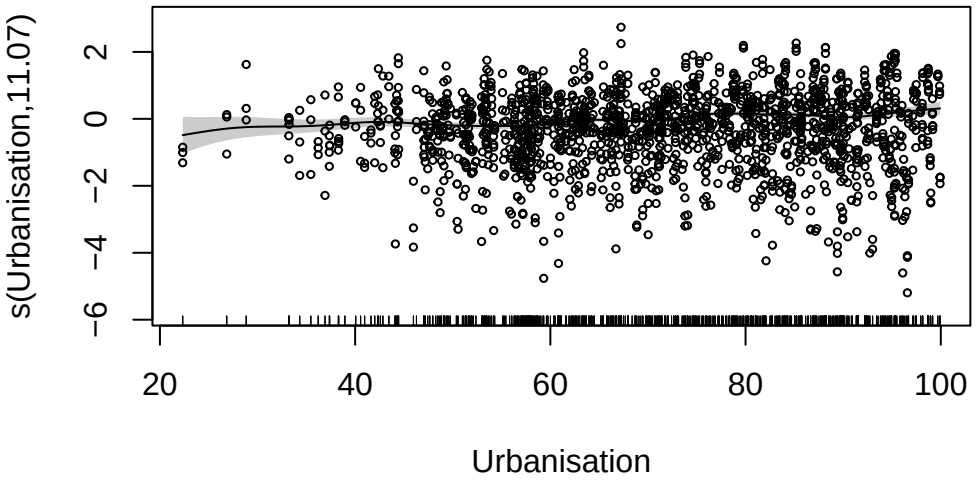
```
[1] 1.338156
```

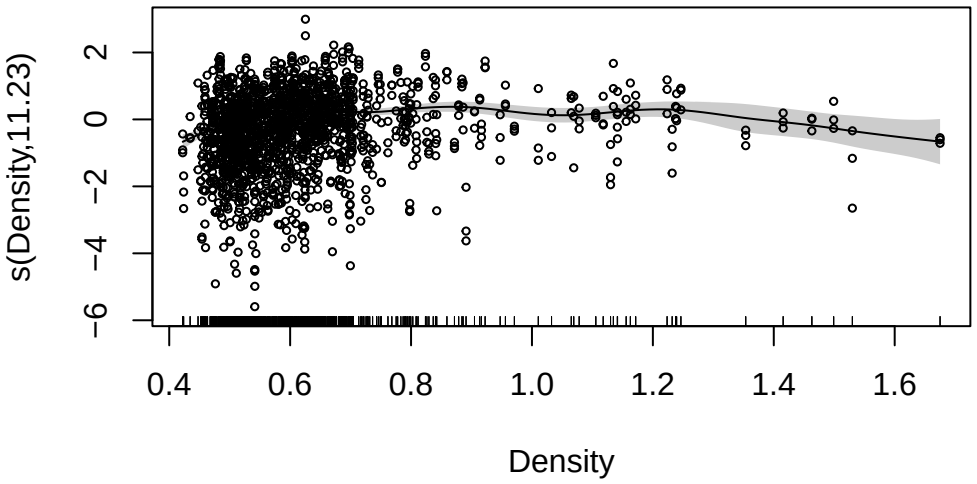
The dispersion parameter is very close to 1, unlike for the Poisson model, meaning that the model that can account for most of the over-dispersion in the data. As such a dispersion parameter value close to 1 can be interpreted as the model is a good fit for the data due to the model adequately capture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

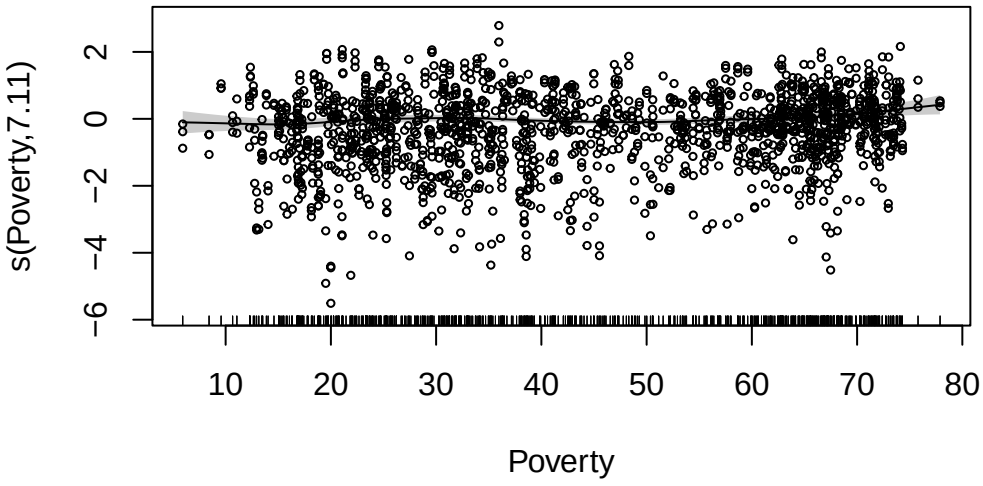
```
plot(nb_model, shade=T, rug = TRUE, residuals = TRUE,scheme=1,
pch = 1, cex = 0.5)
```

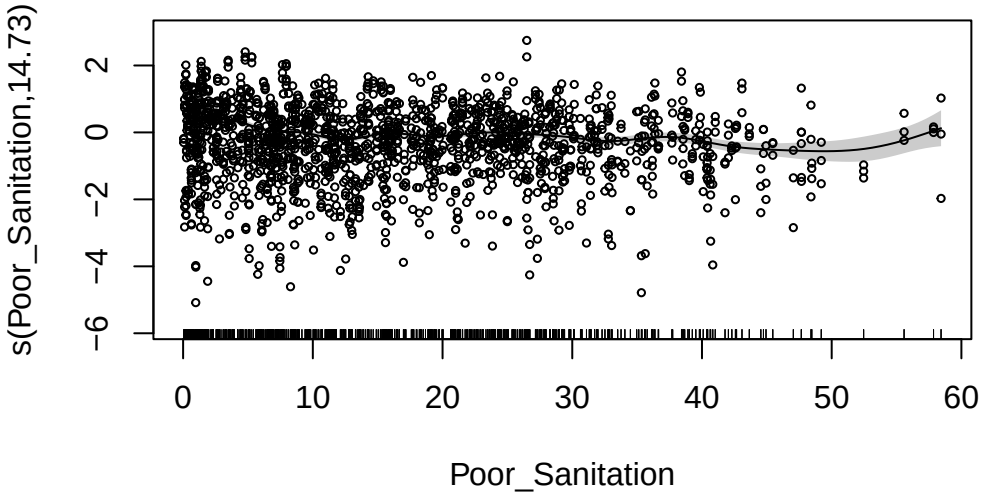


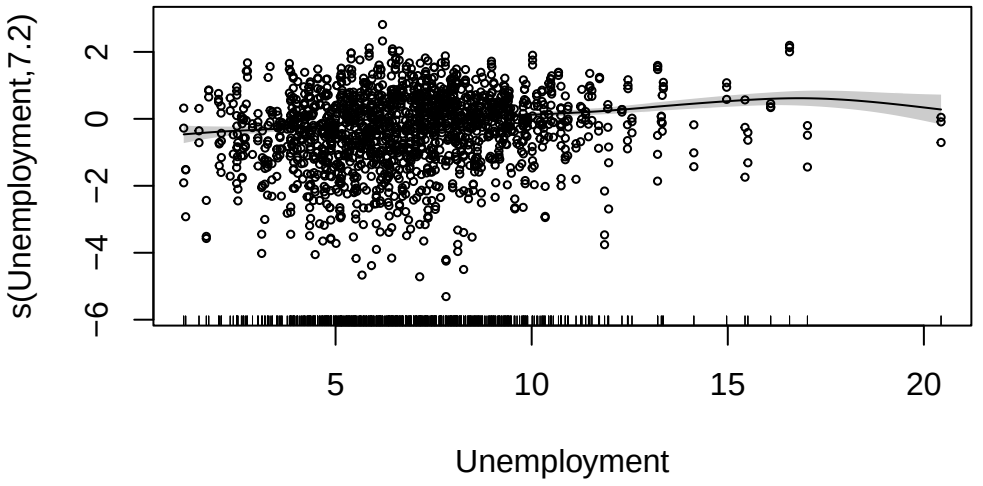

s(Illiteracy,10.89)
−6
−4
−2
0
2
10
20
30
40
Illiteracy

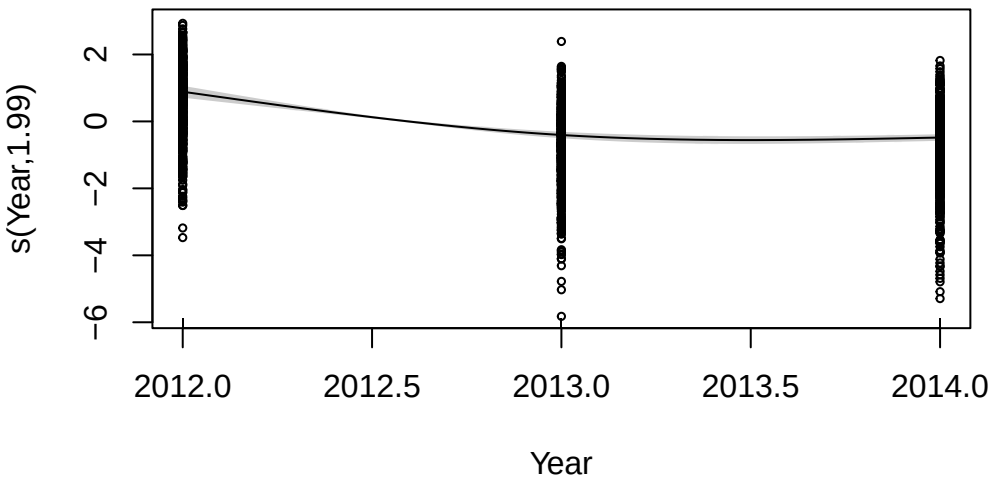

s(Urbanisation,11.07)
−6
−4
−2
0
2
20
40
60
80
100
Urbanisation

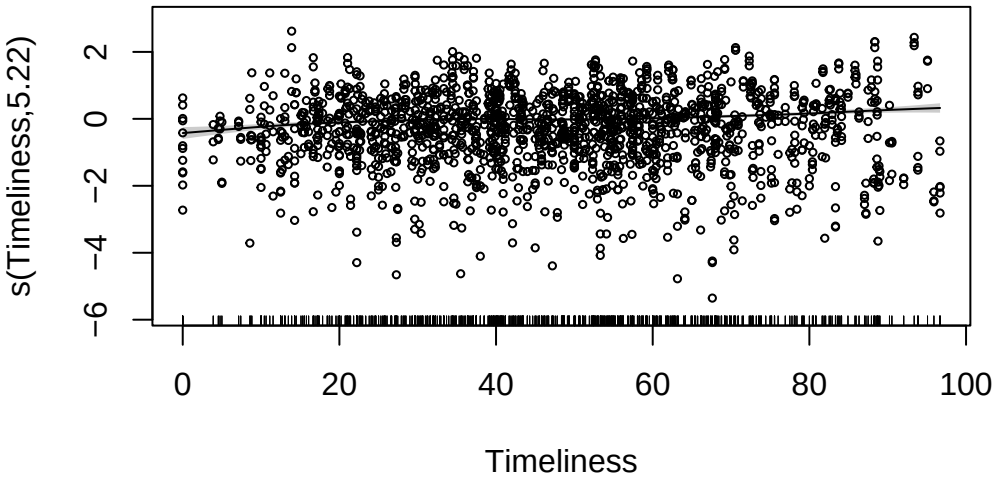

s(Density,11.23)
2
0
−2
−4
−6
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Density

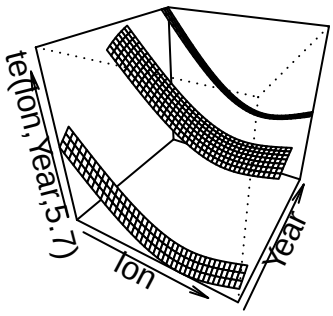

s(Poverty,7.11)
2
0
−2
−4
−6
10
20
30
40
50
60
70
80
Poverty

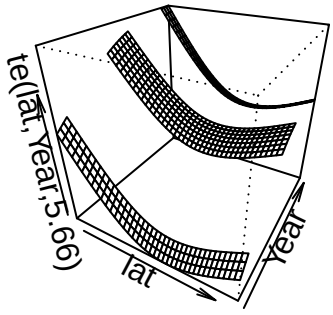

s(Poor_Sanitation,14.73)
−6
−4
−2
0
2
0
10
20
30
40
50
60
Poor_Sanitation


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

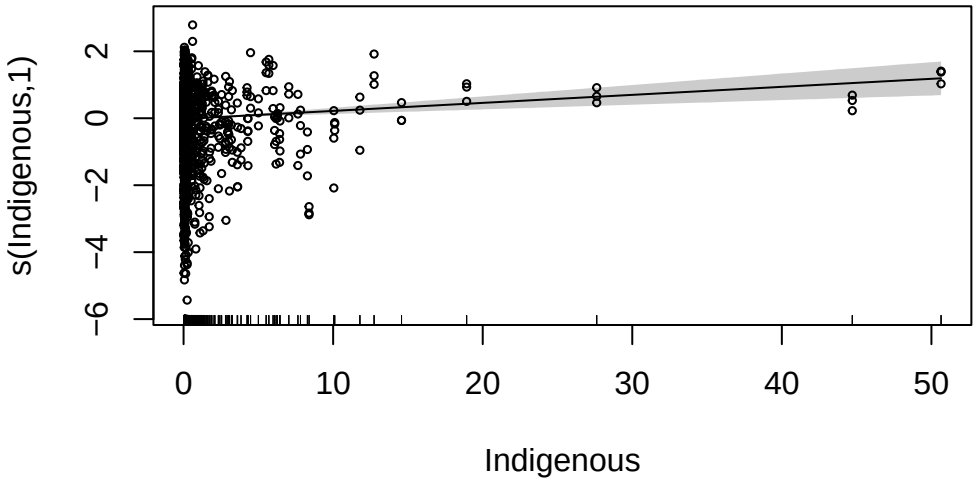

s(Year,1.99)
−6
−4
−2
0
2
2012.0
2012.5
2013.0
2013.5
2014.0
Year

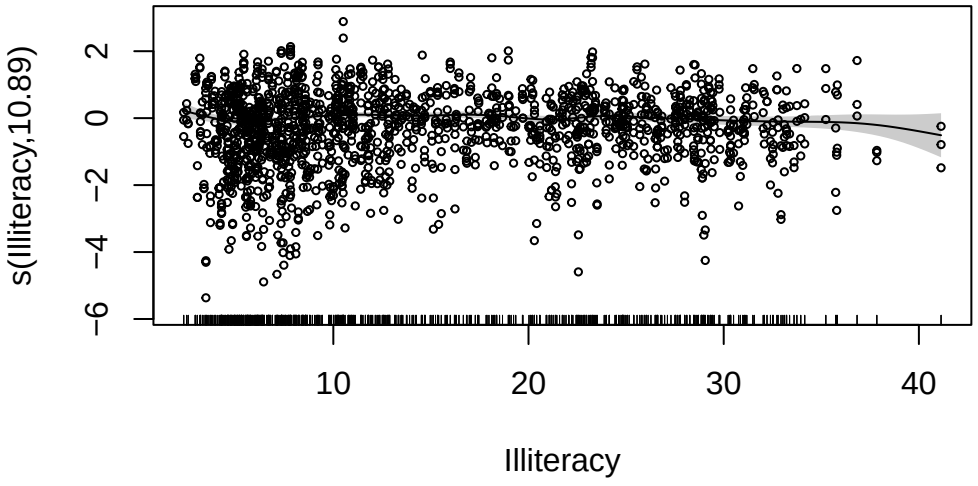

s(Timeliness,5.22)
−6
−4
−2
0
2
0
20
40
60
80
100
Timeliness

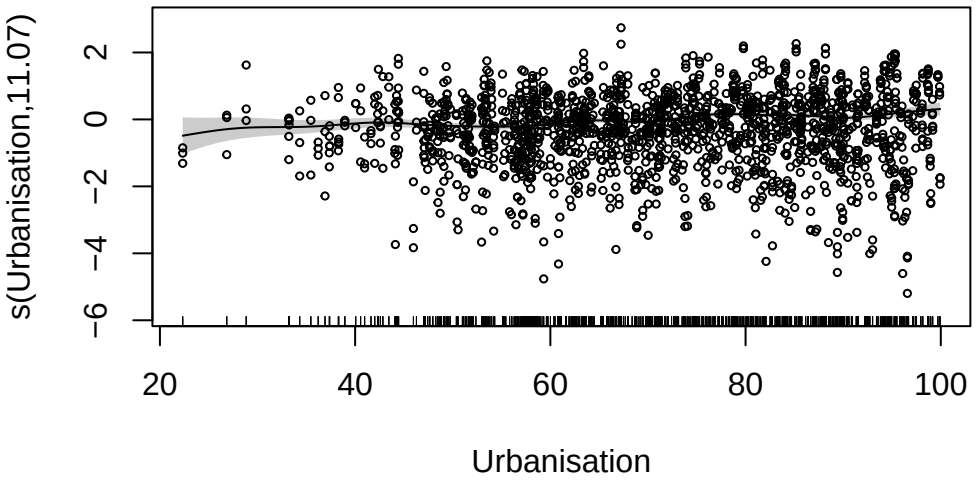

te(lon,Year,5.7)
lon
Year

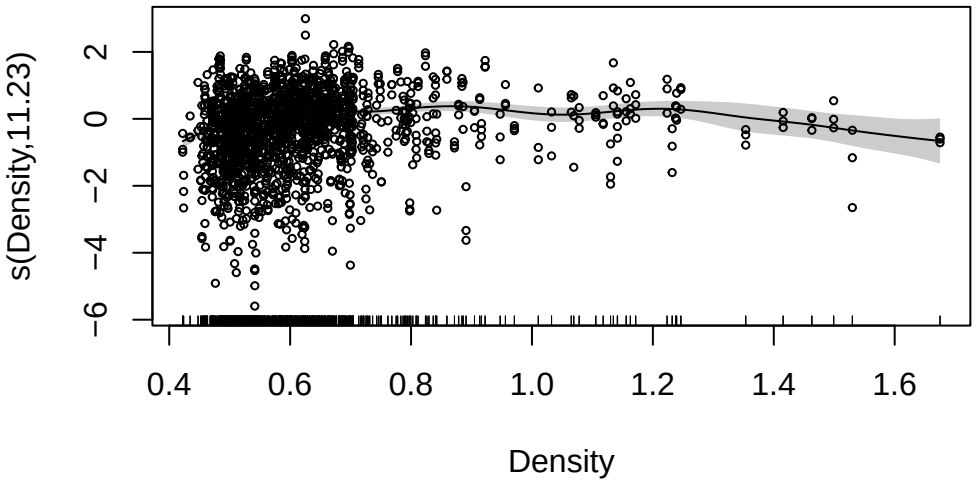

te(lat,Year,5.66)
lat
Year

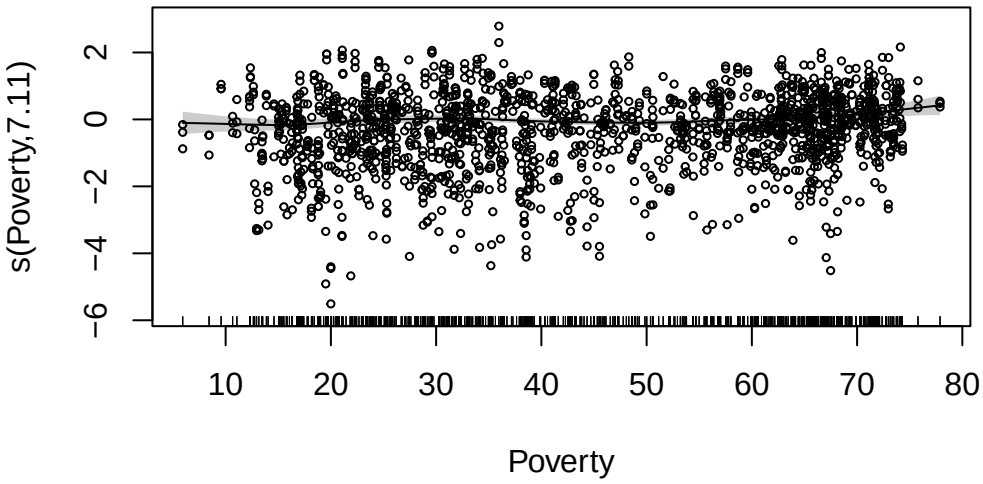
```
plot(nb_model, shade = T, rug = TRUE, residuals = TRUE, scheme = 2, pch = 1,
    cex = 0.5)
```

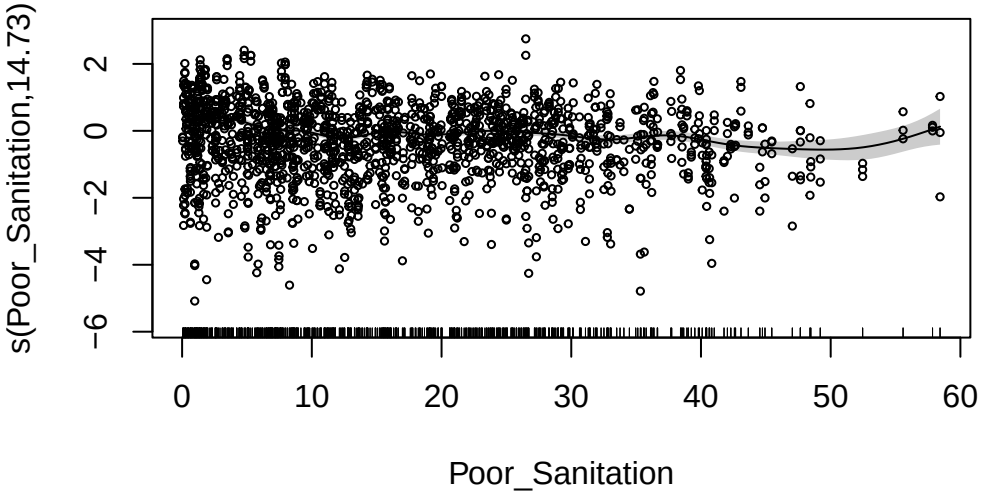


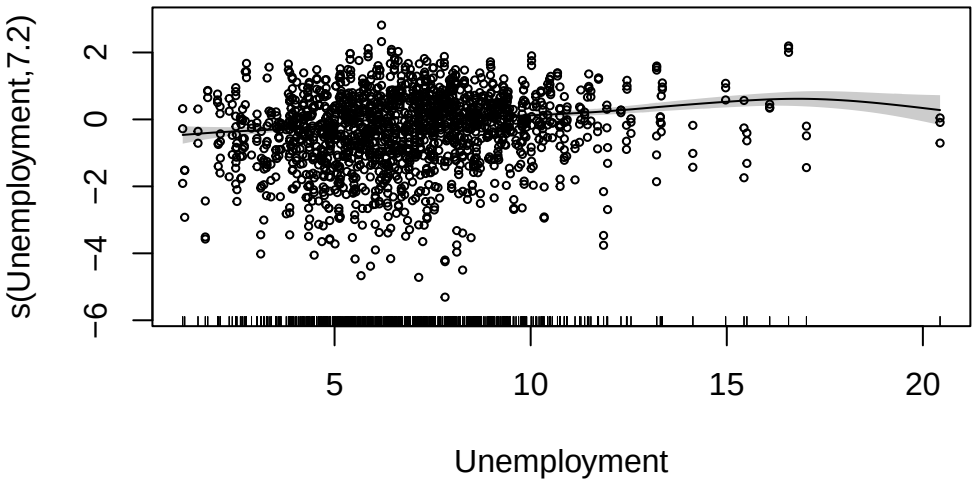


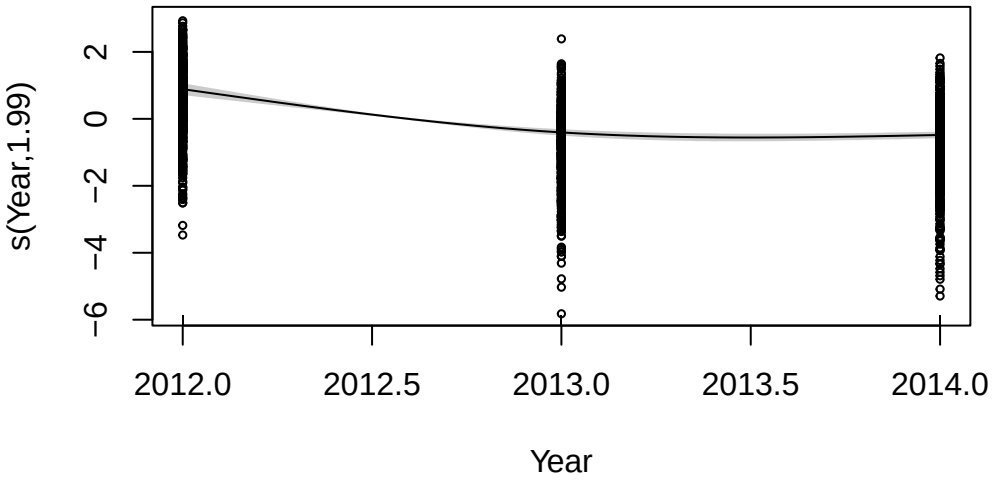

s(Urbanisation,11.07)
2
0
−2
−4
−6
20
40
60
80
100
Urbanisation

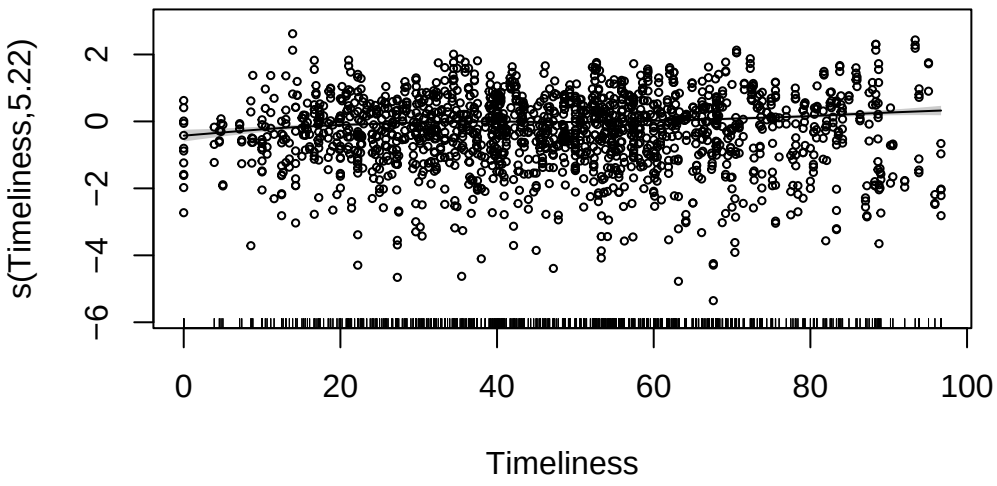

s(Density,11.23)
2
0
−2
−4
−6
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Density

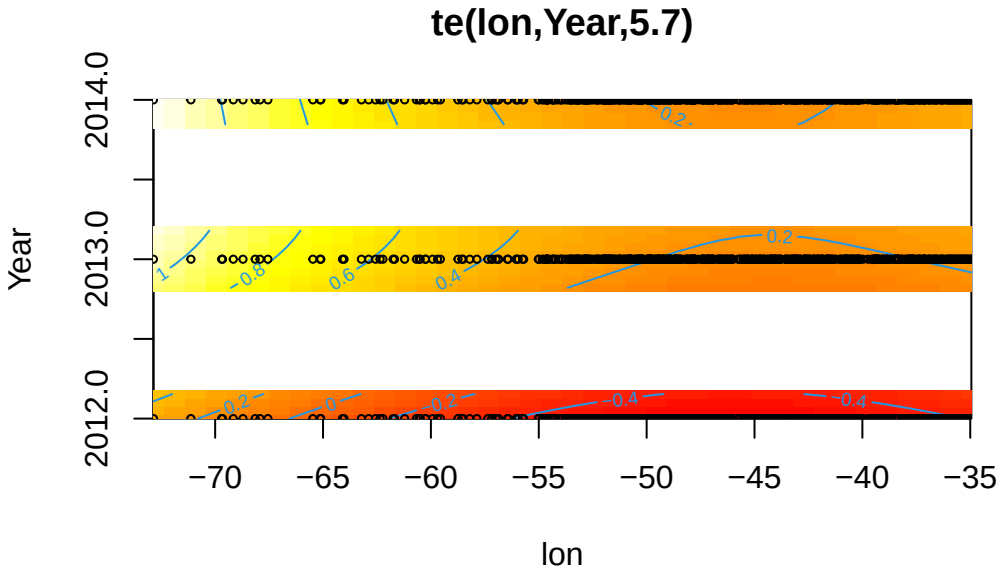

s(Poverty,7.11)
−6
−4
−2
0
2
10
20
30
40
50
60
70
80
Poverty

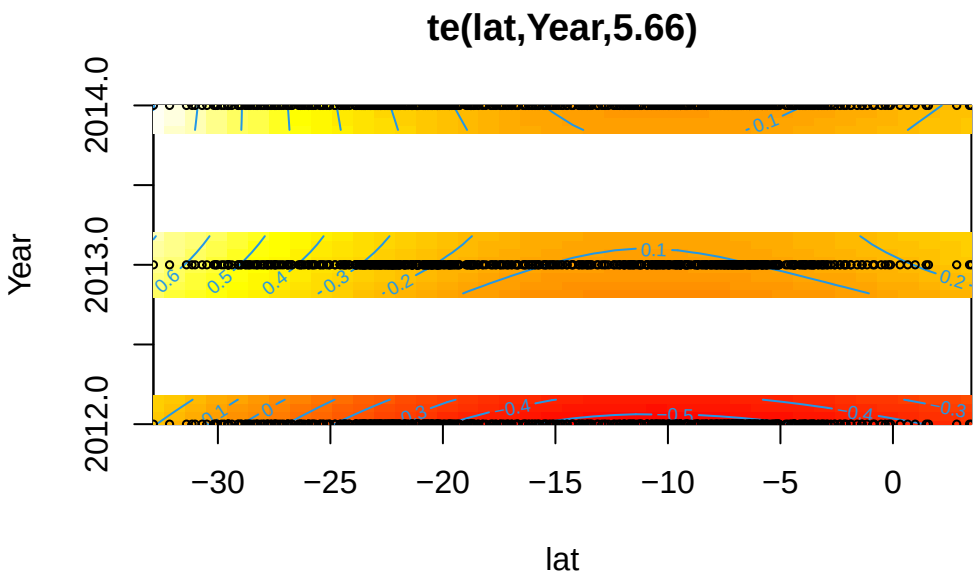

s(Poor_Sanitation,14.73)
−6
−4
−2
0
2
0
10
20
30
40
50
60
Poor_Sanitation


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

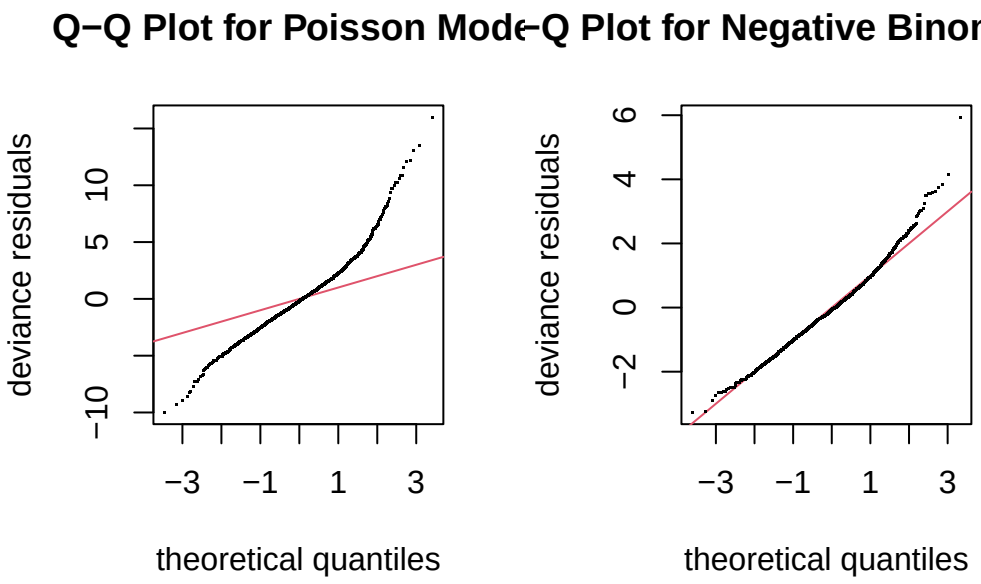

s(Year,1.99)
2
0
−2
−4
−6
2012.0
2012.5
2013.0
2013.5
2014.0
Year


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")
```



##spatial graph does not render automaticall have to be added manually

```
par(mfrow = c(3, 1))
vis.gam(nb_model, view = c("lon", "lat"), cond = list(Year = 2012), plot.type = "contour")
vis.gam(nb_model, view = c("lon", "lat"), cond = list(Year = 2013), plot.type = "contour")
vis.gam(nb_model, view = c("lon", "lat"), cond = list(Year = 2014), plot.type = "contour")
par(mfrow = c(1, 1))
```

Bibliography

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