Group 34 course work

Table of contents

Introduction:

Tuberculosis (TB) is a bacterial disease caused by the bacterium Mycobacterium tuberculosis. It primarily affects the lungs, but can also affect other parts of the body, such as the kidneys, spine, and brain.

TB is a major public health problem worldwide, affecting millions of people each year. According to the World Health Organization (WHO), TB is one of the top 10 causes of death worldwide, and in 2020 alone, there were an estimated of approximately 10 million cases of TB reported globally. Brazil is one of the countries with a high burden of TB with an estimate of 96000, and it is considered a priority country for TB control by the WHO. *Global Tuberculosis Report 2020* (2020).

The purpose of this report is to provide a comprehensive analysis of TB in Brazil from the 2012 to 2014 data, including an overview of the trends and risk factors associated with TB. Additionally, this report aims to highlight the most important factor for improving TB control and prevention efforts in Brazil.

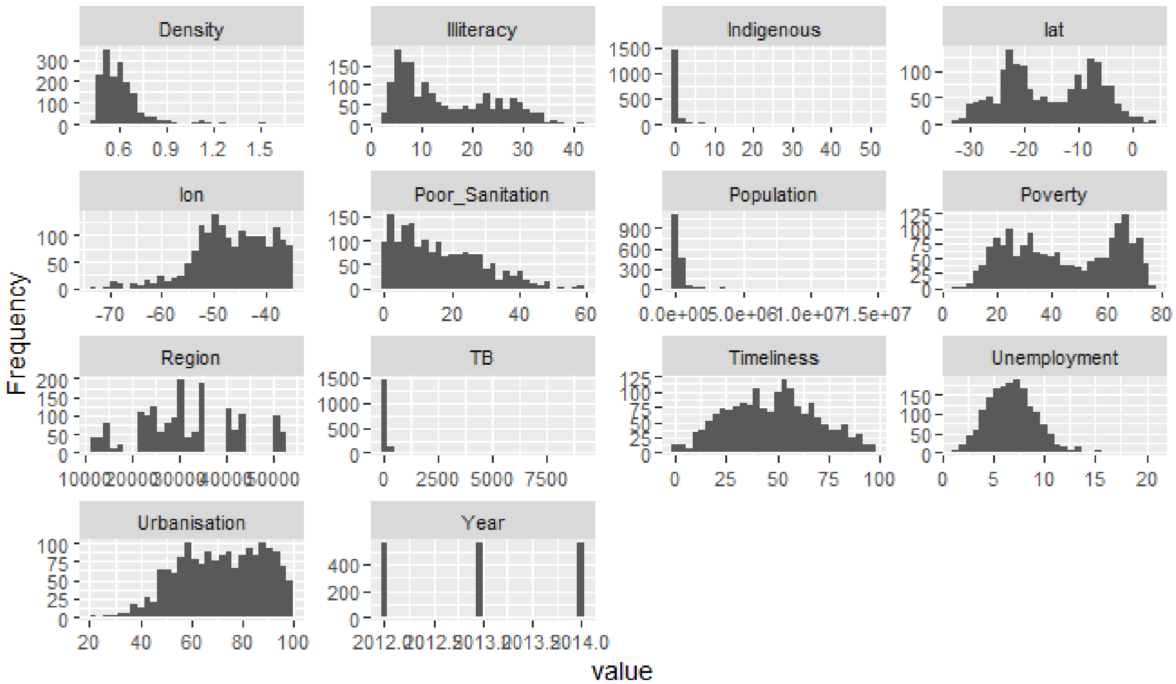
The case study we will consider in this report is the Tuberculosis dataframe where Brazil is divided into 557 administrative microregions and the available data comprises of counts of TB cases in each microregion from 2012 to 2014.

Exploratory Data Analysis:

An exploratory analysis of this dataset can reveal important insights of the potential risk factors for TB in Brazil and help guide public health interventions.

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.

We will start by analysing the distributions of each of the variables to identify any patterns



Firstly, the dwelling density seems to follow a normal distribution that is skewed to the right and a mean of approximately 0,6.

Secondly illiteracy is very heavily skewed to the right but it still displays a normal bell curve around the 5% illiteracy level.

Poor sanitation is extremely skewed to the right.

Unemployment seems to follow a normal distribution with little to no skewness and a mean of approximately 6%.

The other values do not seem to follow any easily identifiable trend nor provide us any meaningful insights with their distribution alone.

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.

Model Selection:

We are utilising generalized additive models (GAMs) in the case study of TB in Brazil as it can model complex and non-linear relationships between TB incidence and risk factors, control for

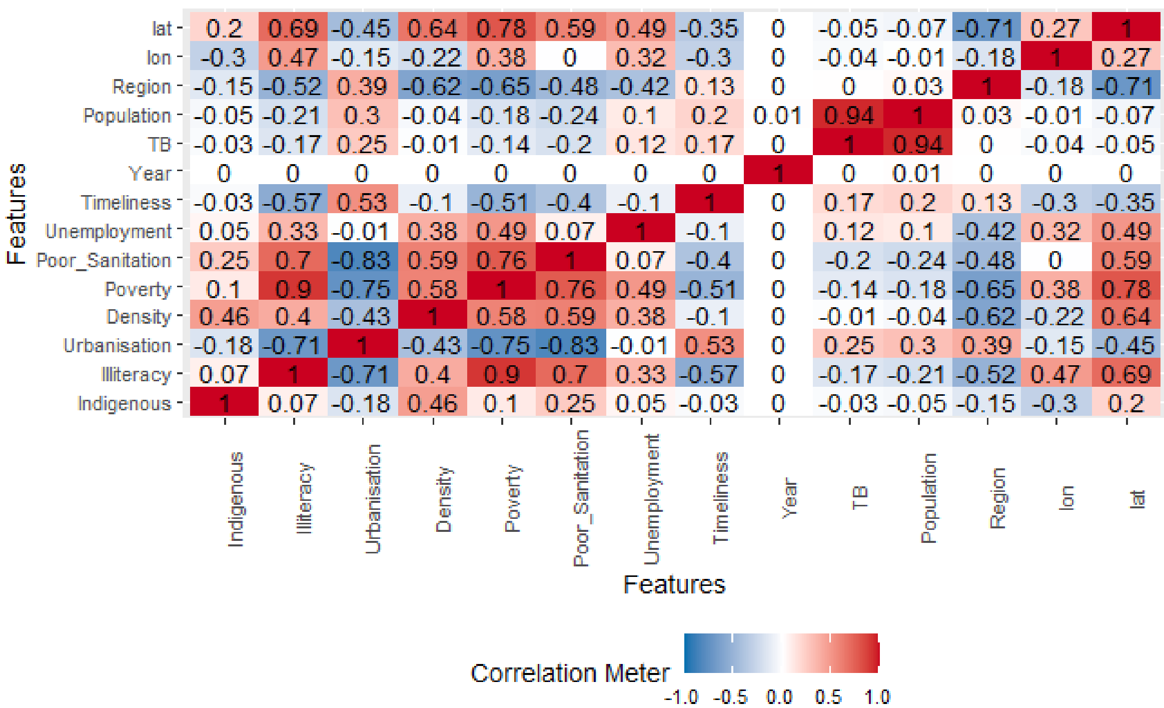


Figure 1: Correlation Matrix

relevant covariates, identify important predictors of TB incidence, and predict TB incidence in different regions of Brazil.

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

Yi Pois(λ) Yi indep.

Log(λi) = $log(Populationi) + fIndigenous(Indigenousi) + fIlliteracy(Illiteracyi) + fUrbanisation(Urbanisationi) + fDensity(Densityi) + fPoverty(Povertyi) + fPoor_Sanitation(Poor_Sanitation$
$fUnemployment(Unemploymenti) + fTimeliness(Timelinessi) + fla(lati) + flon(loni) + fYear(Yeari) + flon, lat, Year(loni, lati, Yeari)$

Model Fitting:

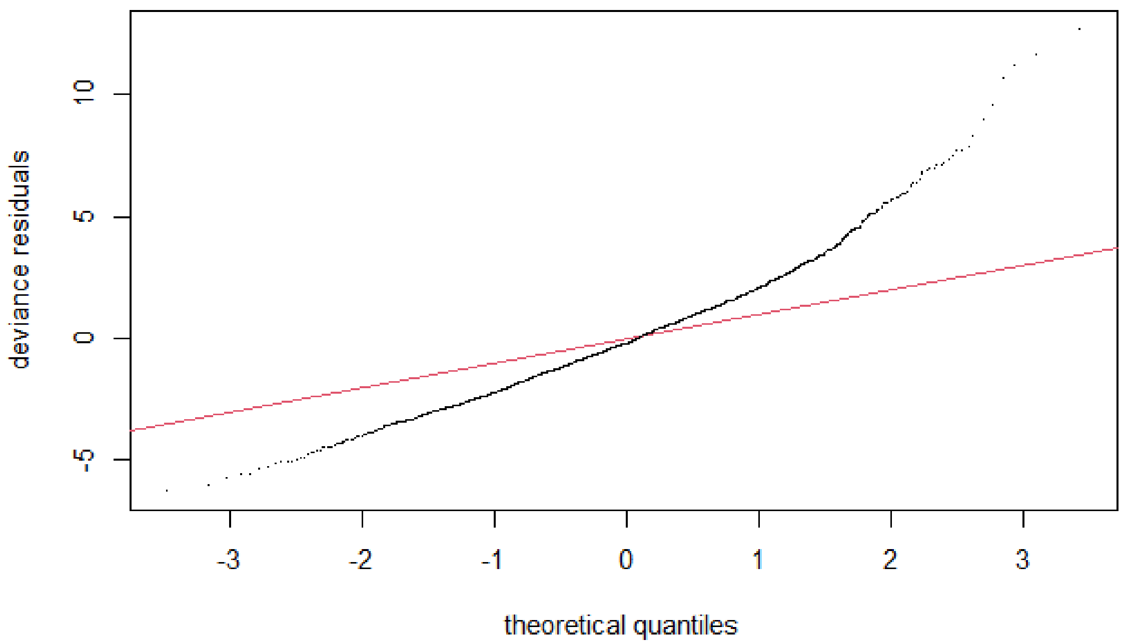


Figure 2: QQ Plot

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.

However the Poisson model seems to not have a good enough fitting results and as such we will extend the model into a negative binomial GAM

Model Evaluation:

TODO make this mathematical istead of the code

Calulating the dispersion parameter: sum(residuals(poisson_model, type = “pearson”)^2) / df.residual(poisson_model) = 6.67687

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 unacounted patern in the model.

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 thre 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.

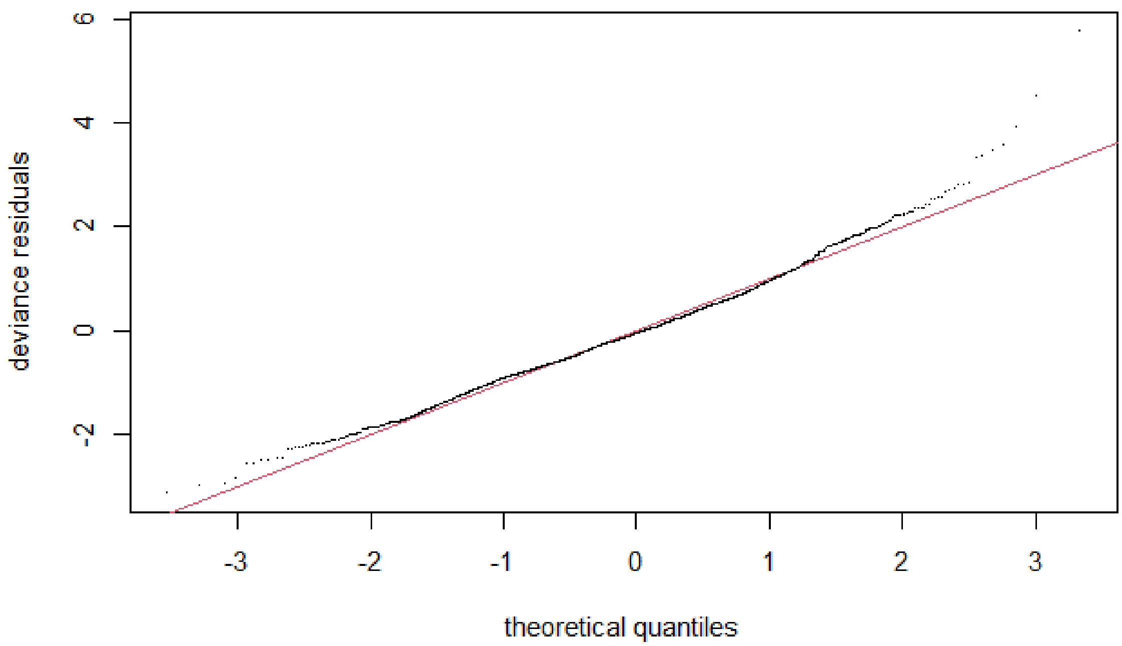


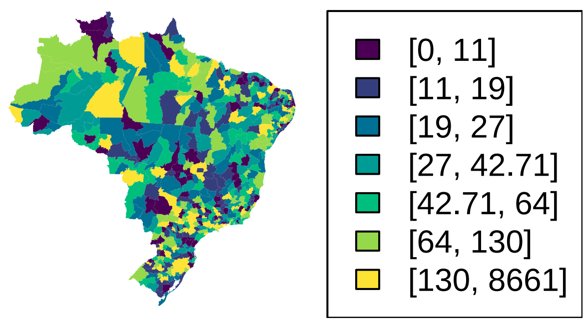
Figure 3: QQ - Negative Binomial

Results and Interpretation:

Apendix

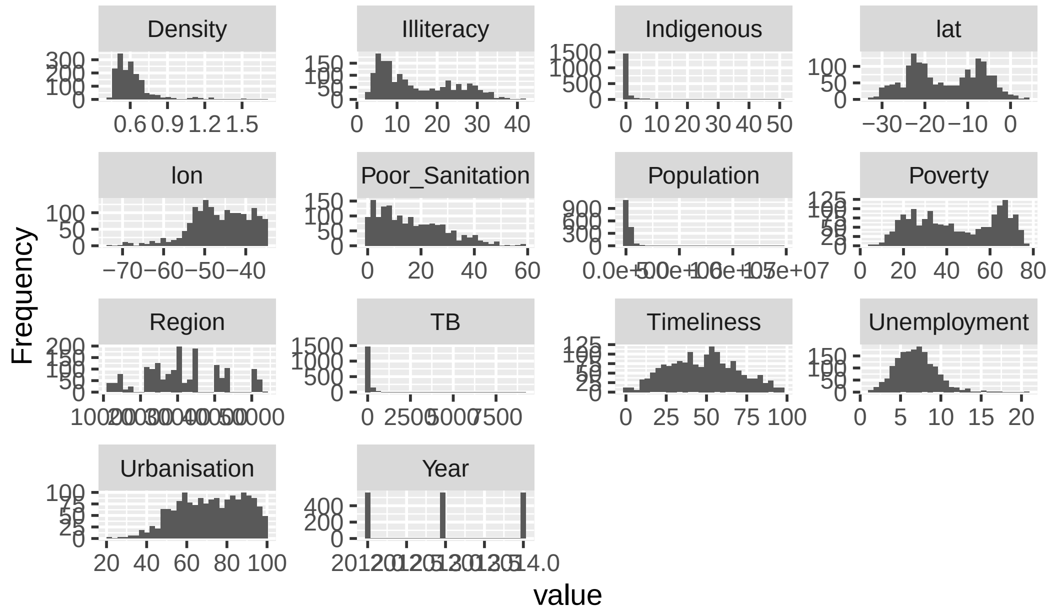
```r
## PLotting map of cases
plot.map(TBdata$TB[TBdata$Year == 2014], n.levels = 7, main = "TB counts for 2014")
```

TB counts for 2014



Exploratory analyses

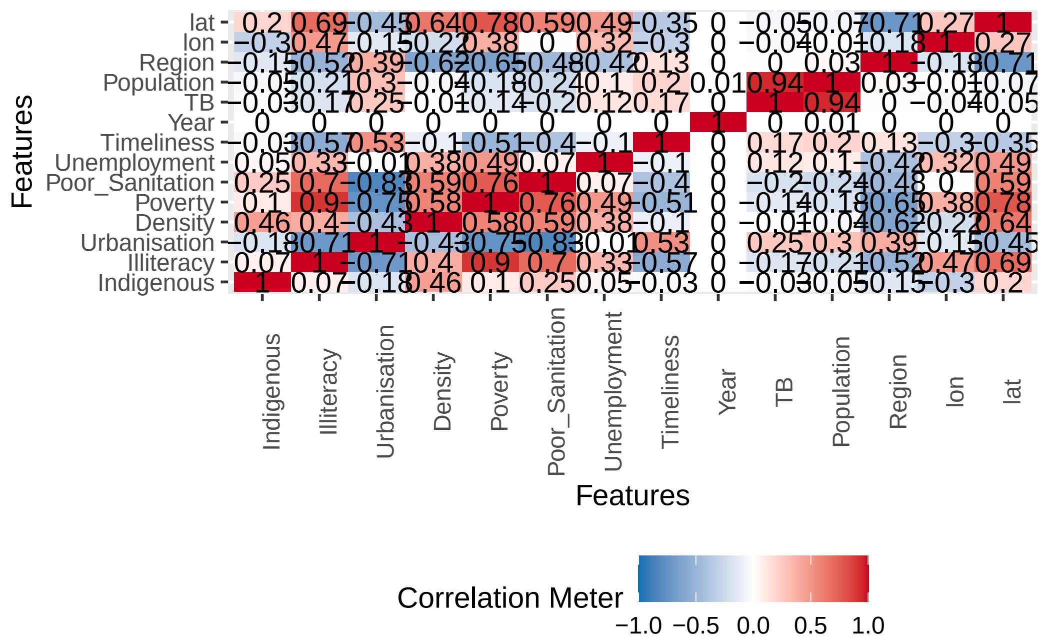
```r
plot_histogram(TBdata)
```



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

Now investigating the correlation matrix of the numerical variables

```
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

```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(Timeliness, k = 20) + s(lat, k = 30) + s(lon,
    k = 30) + s(Year, k = 3) + ti(lon, lat, Year, k = 3)
```

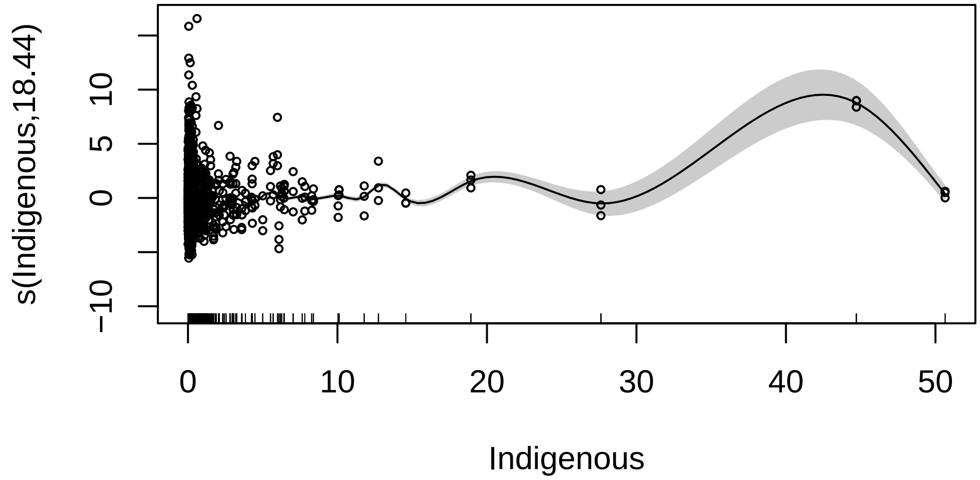
```
Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.480233   0.004557   -1861   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                       edf Ref.df  Chi.sq p-value
s(Indigenous)       18.443 18.925  649.18 < 2e-16 ***
s(Illiteracy)       18.262 18.907  439.58 < 2e-16 ***
s(Urbanisation)     18.888 18.991  741.26 < 2e-16 ***
s(Density)          17.130 17.913  525.38 < 2e-16 ***
s(Poverty)          18.502 18.917 1176.23 < 2e-16 ***
s(Poor_Sanitation) 17.517 18.656  858.89 < 2e-16 ***
s(Unemployment)     18.475 18.944 1450.40 < 2e-16 ***
s(Timeliness)       18.159 18.868  911.93 < 2e-16 ***
s(lat)              28.310 28.939 2406.85 < 2e-16 ***
s(lon)              28.447 28.951 2601.82 < 2e-16 ***
s(Year)              1.847  1.962   46.82 < 2e-16 ***
ti(lon,lat,Year)     3.913  5.017   16.61 0.00535 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.997   Deviance explained = 87.9%
-REML = 9825.5  Scale est. = 1         n = 1671
```

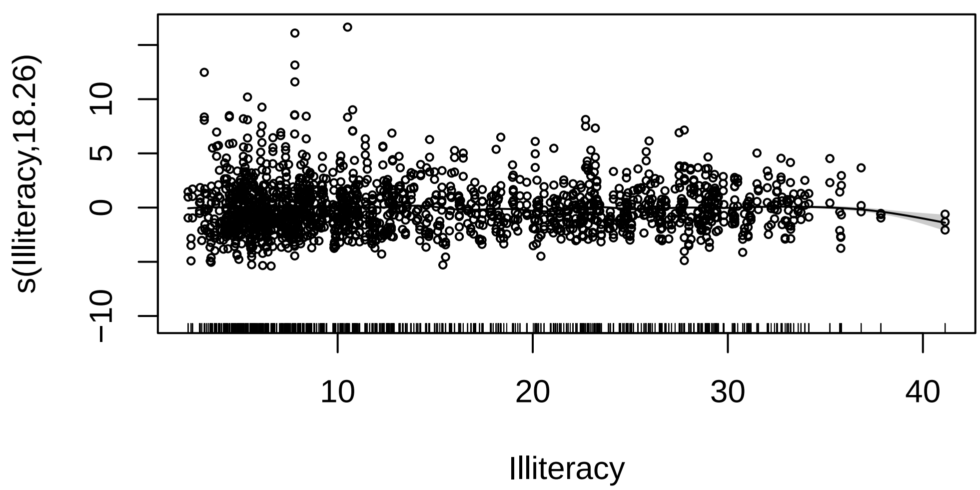
```r
gam.check(poisson_model)
```

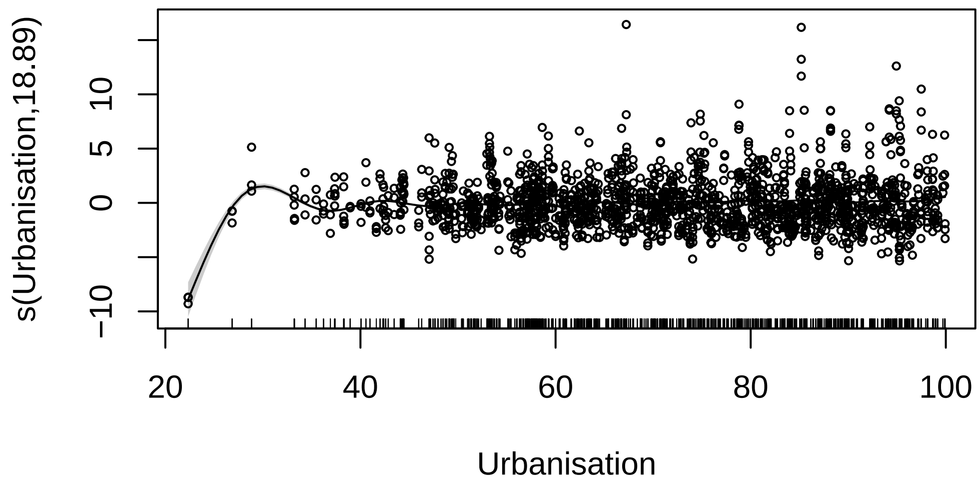
```r
#Akaike Information Criterion:
poisson_model$aic
```

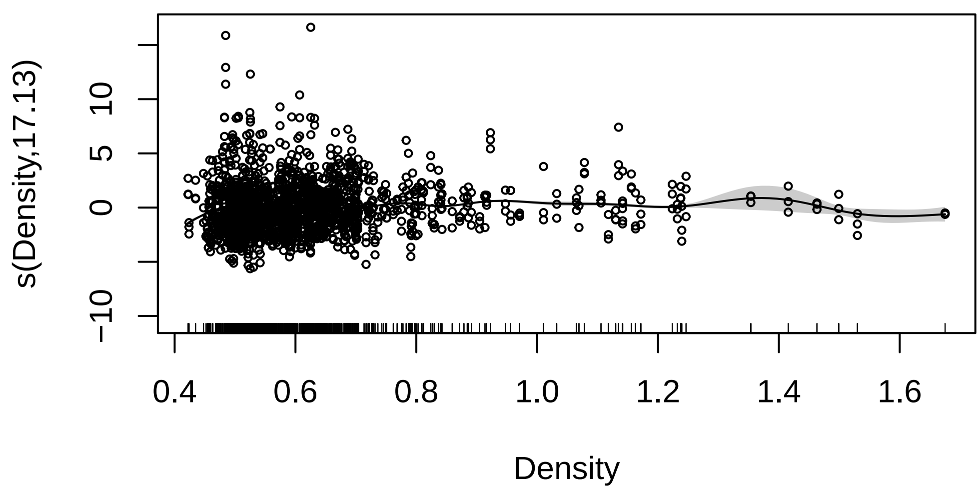
```
[1] 18585.52
```

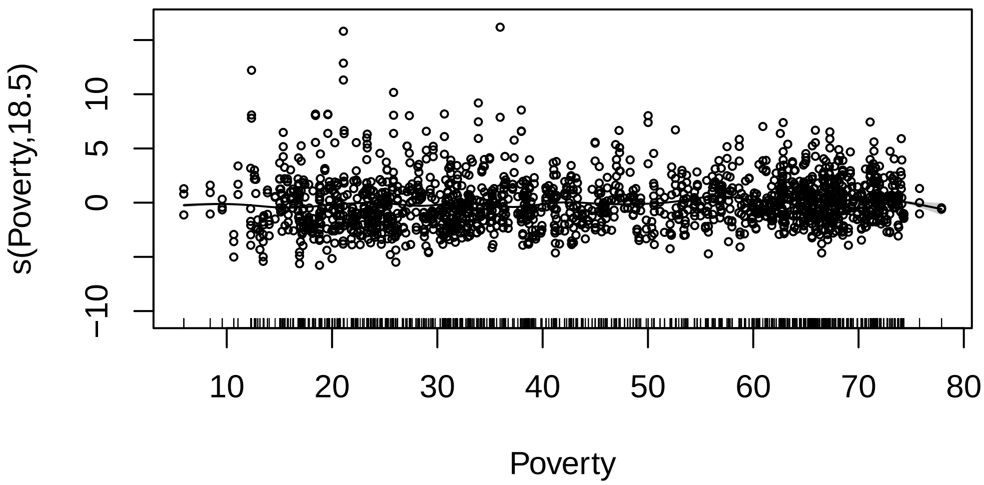
```r
plot(poisson_model, shade=T, rug = TRUE, residuals = TRUE,
pch = 1, cex = 0.5)
```

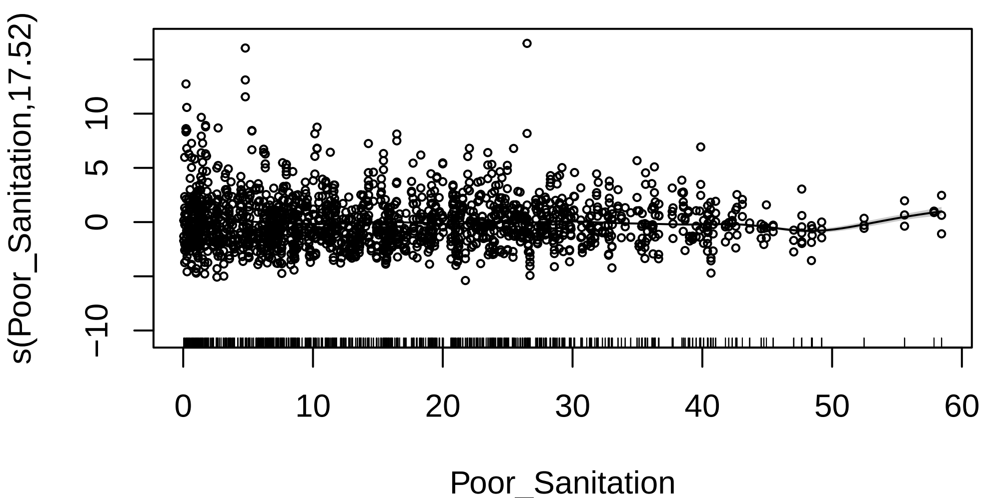

s(Indigenous,18.44)
−10
0
5
10
0
10
20
30
40
50
Indigenous

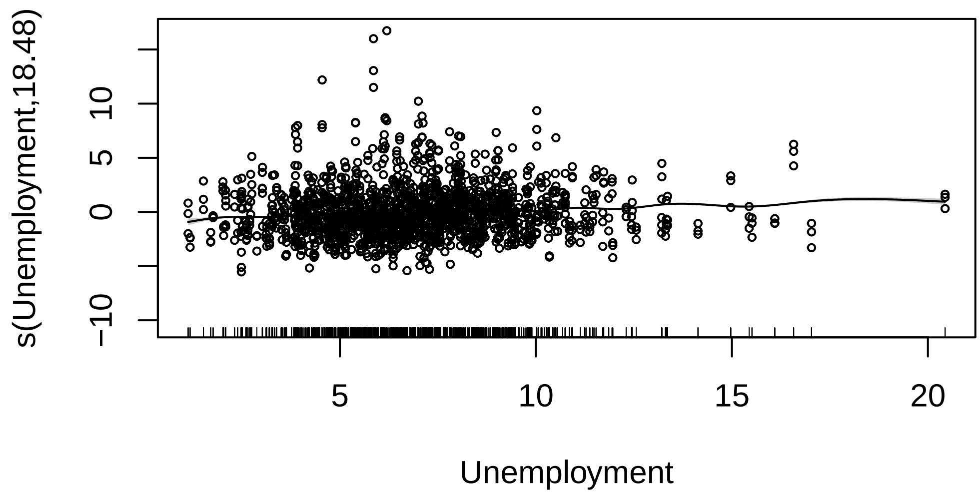

s(Illiteracy,18.26)
−10
0
5
10
10
20
30
40
Illiteracy


s(Urbanisation,18.89)
−10
0
5
10
20
40
60
80
100
Urbanisation

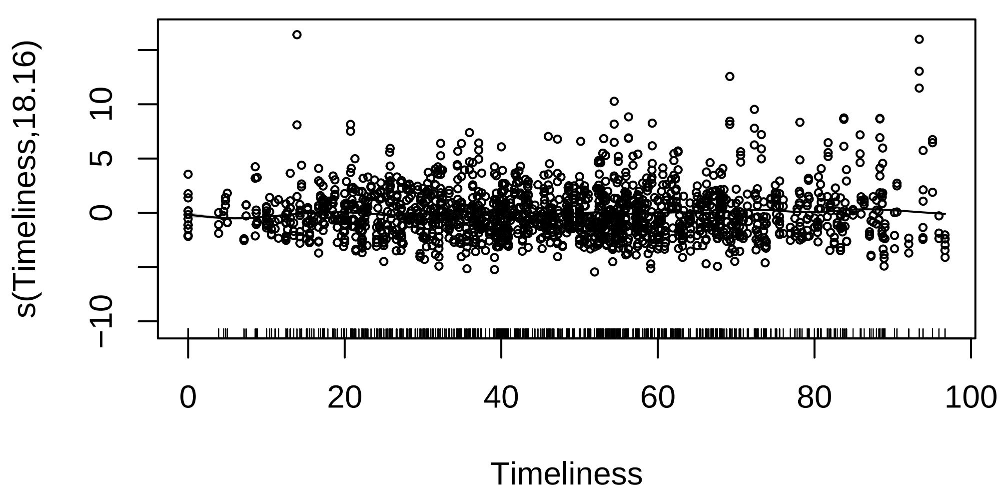

s(Density,17.13)
−10
0
5
10
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Density


s(Poverty,18.5)
−10
0
5
10
10
20
30
40
50
60
70
80
Poverty

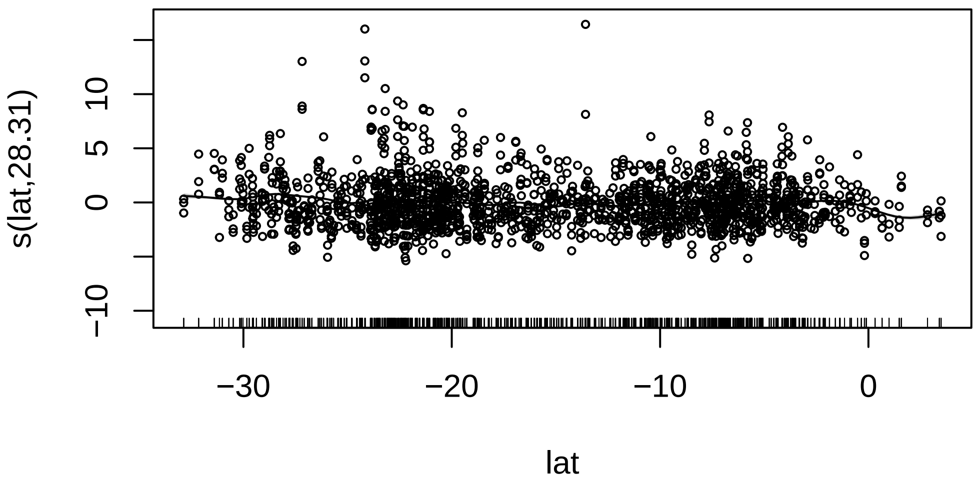

s(Poor_Sanitation,17.52)
−10
0
5
10
0
10
20
30
40
50
60
Poor_Sanitation

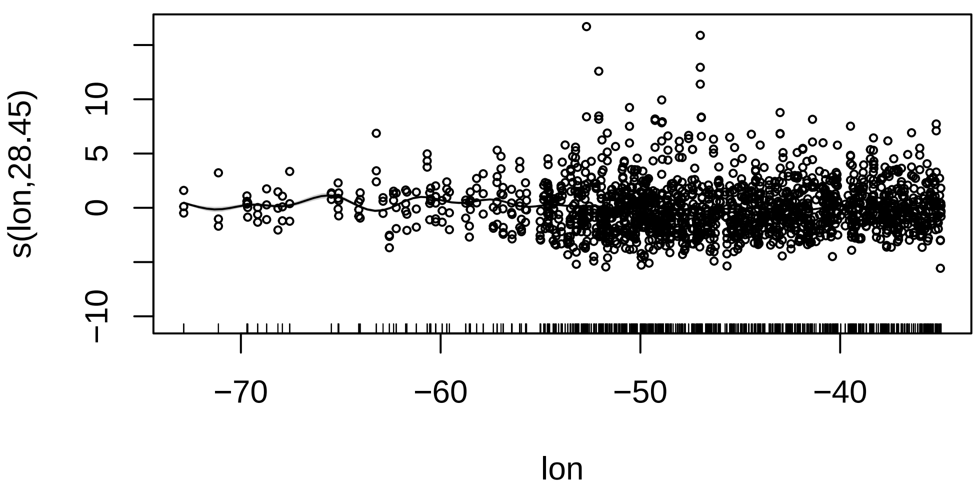


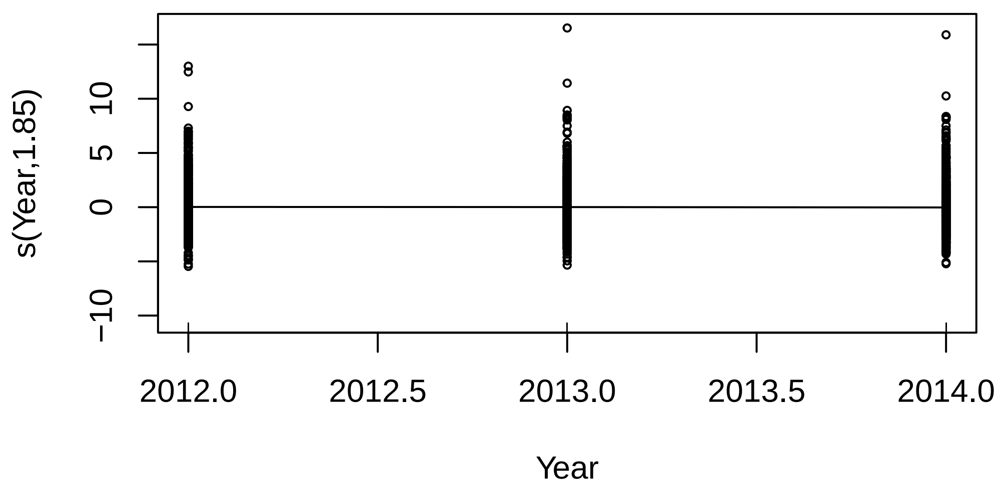
s(Unemployment,18.48)
−10
0
5
10
5
10
15
20
Unemployment

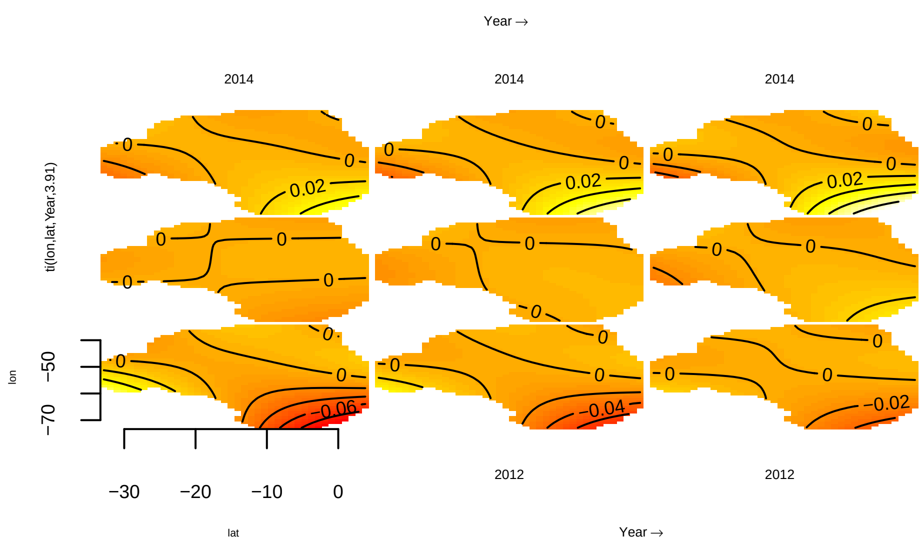


s(Timeliness,18.16)
−10
0
5
10
0
20
40
60
80
100
Timeliness

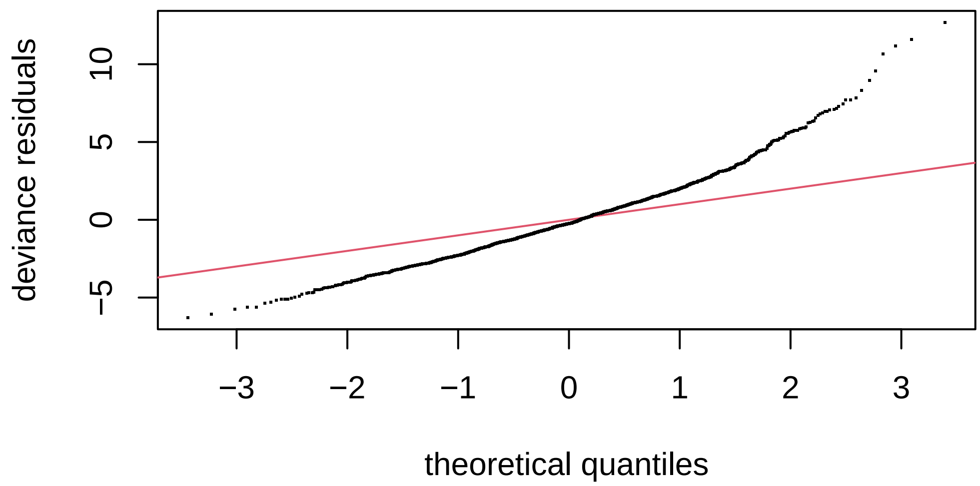

s(lat,28.31)
−10
0
5
10
−30
−20
−10
0
lat


s(lon,28.45)
−10
0
5
10
−70
−60
−50
−40
lon


s(Year,1.85)
−10
0
5
10
2012.0
2012.5
2013.0
2013.5
2014.0
Year


Year →
2014
2014
2014
ti(lon,lat,Year,3.91)
0.02
0.02
0.02
−0.06
−0.04
−0.02
lon
−50
−70
−30
−20
−10
0
lat
2012
2012
Year →

```r
#Lets check our model residuals:
qq.gam(poisson_model)
```



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.

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

```
[1] 6.67687
```

```r
#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

Negatice 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(Timeliness, k = 20) + s(lat, k = 30) + s(lon,
    k = 30) + s(Year, k = 3) + ti(lon, lat, Year, k = 3)

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.451092   0.009451  -894.2   <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)       2.875  3.451  18.887 0.000643 ***
s(Illiteracy)       4.772  6.034  11.091 0.079313 .
s(Urbanisation)     6.599  8.228  33.224 8.79e-05 ***
s(Density)          6.942  8.637 103.826  < 2e-16 ***
s(Poverty)          2.028  2.608   1.074 0.652302
s(Poor_Sanitation)  8.696 10.680  70.808  < 2e-16 ***
s(Unemployment)     5.532  6.935  92.111  < 2e-16 ***
s(Timeliness)       4.764  5.967  79.993  < 2e-16 ***
s(lat)             21.577 25.088 190.344  < 2e-16 ***
s(lon)             19.462 23.059 239.215  < 2e-16 ***
s(Year)             1.400  1.636   1.888 0.207855
```

```
ti(lon,lat,Year)     1.363  1.592   0.819 0.681850
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.883   Deviance explained = 58.1%
-REML = 7113.6  Scale est. = 1         n = 1671
```

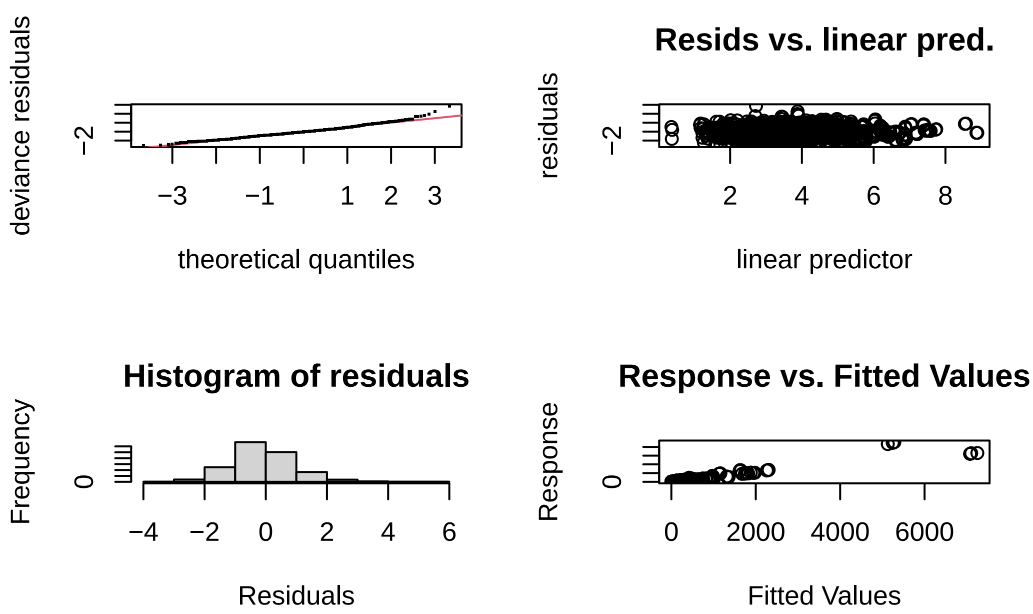
```
#Akaike Information Criterion
nb_model$aic
```

```
[1] 14015.77
```

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)
```

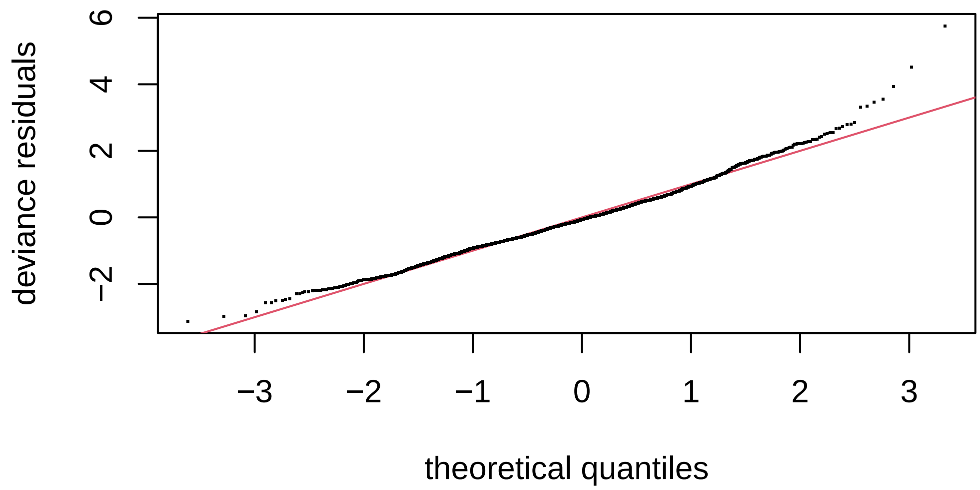



```
Method: REML   Optimizer: outer newton
```

```
full convergence after 7 iterations.
Gradient range [-0.0008538501,2.782458e-05]
(score 7113.619 & scale 1).
Hessian positive definite, eigenvalue range [0.0001209016,3.084703].
Model rank =  221 / 221 

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  2.87    0.54  <2e-16 ***
s(Illiteracy)      19.00  4.77    0.53  <2e-16 ***
s(Urbanisation)    19.00  6.60    0.54  <2e-16 ***
s(Density)         19.00  6.94    0.54  <2e-16 ***
s(Poverty)         19.00  2.03    0.54  <2e-16 ***
s(Poor_Sanitation) 19.00  8.70    0.53  <2e-16 ***
s(Unemployment)    19.00  5.53    0.53  <2e-16 ***
s(Timeliness)      19.00  4.76    0.60  <2e-16 ***
s(lat)             29.00 21.58    0.54  <2e-16 ***
s(lon)             29.00 19.46    0.56  <2e-16 ***
s(Year)             2.00  1.40    0.79  <2e-16 ***
ti(lon,lat,Year)    8.00  1.36    0.94    0.01 ** 
---
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 unacounted patern in the model.

```r
#checking the model residuals
qq.gam(nb_model)
```



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.

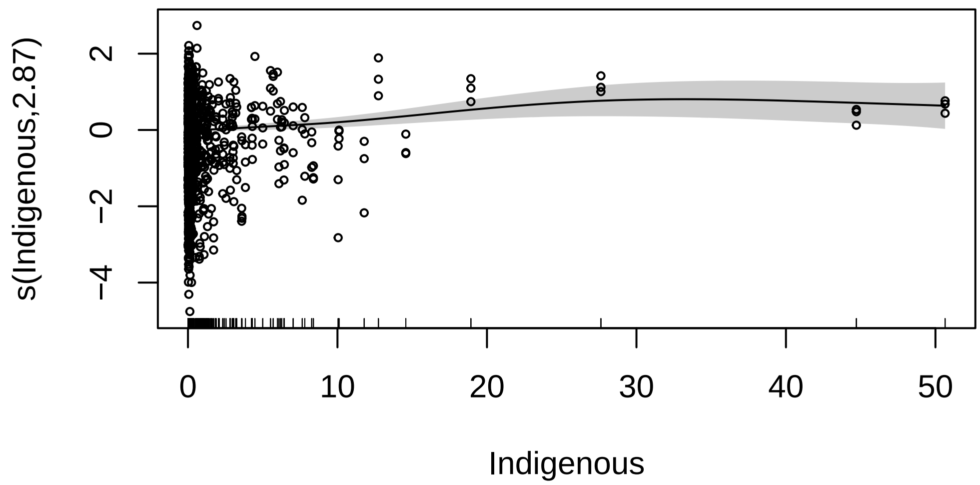
```
#Calculating Pearson estimate for dispersion parameter using Pearson residuals:
sum(residuals(nb_model, type = "pearson")^2) / df.residual(nb_model)
```

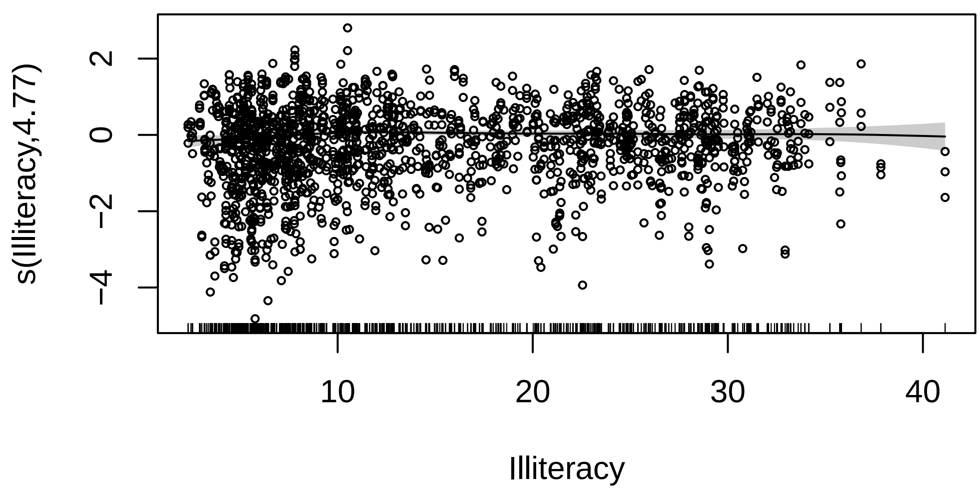
```
[1] 1.190792
```

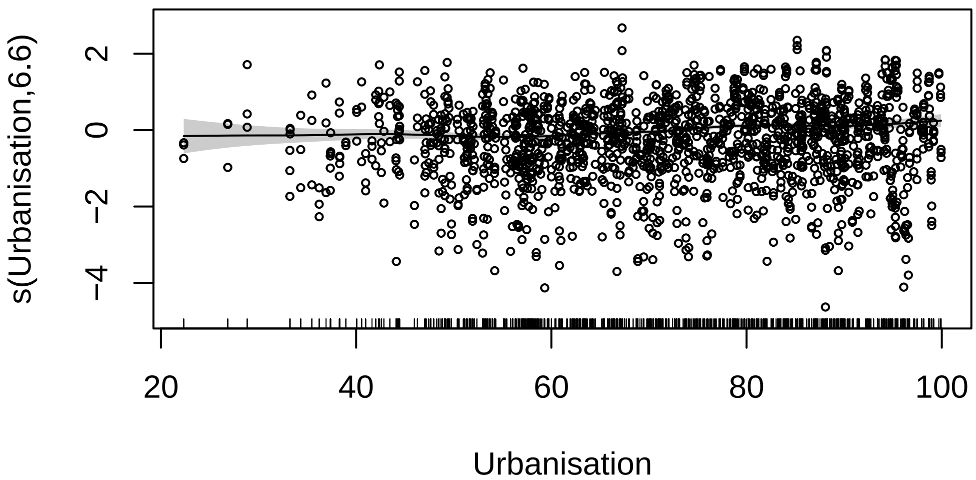
The dispersion parameter is very close to 1, unlike for the Poisson model, meaning that the model that can account for most of the over-dispersion in the data. As such a dispersion parameter value close to 1 can be interpreted as the model is a good fit for the data due to the model adequately capture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

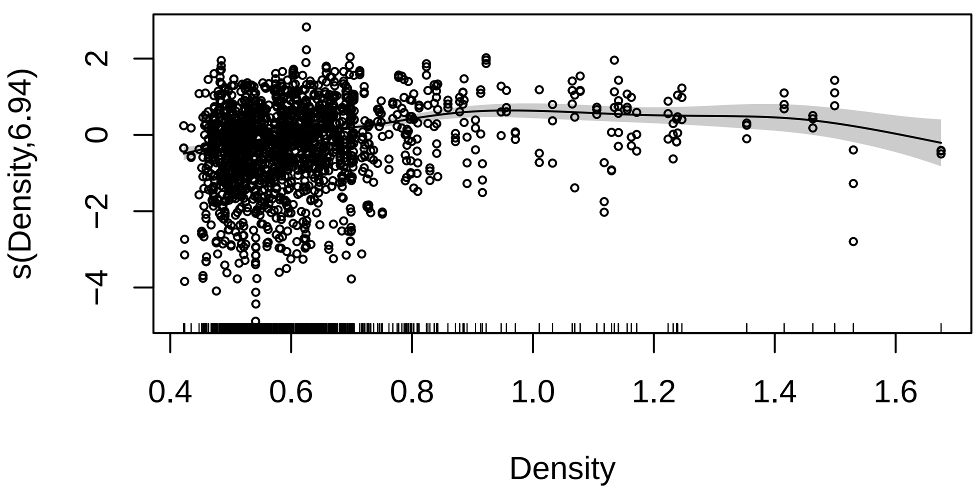
Again tidy up the plots of the Negative Binomial

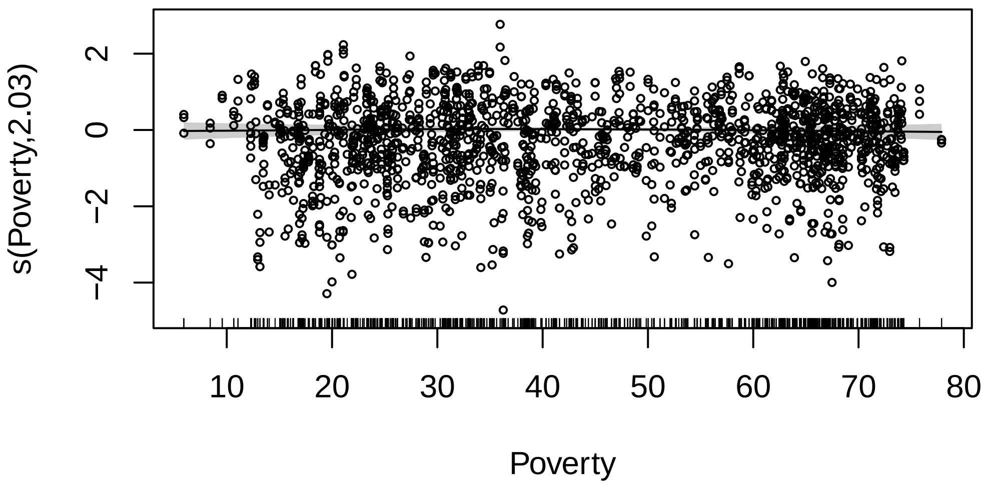
```
plot(nb_model, shade=T, rug = TRUE, residuals = TRUE,scheme=1,
pch = 1, cex = 0.5)
```

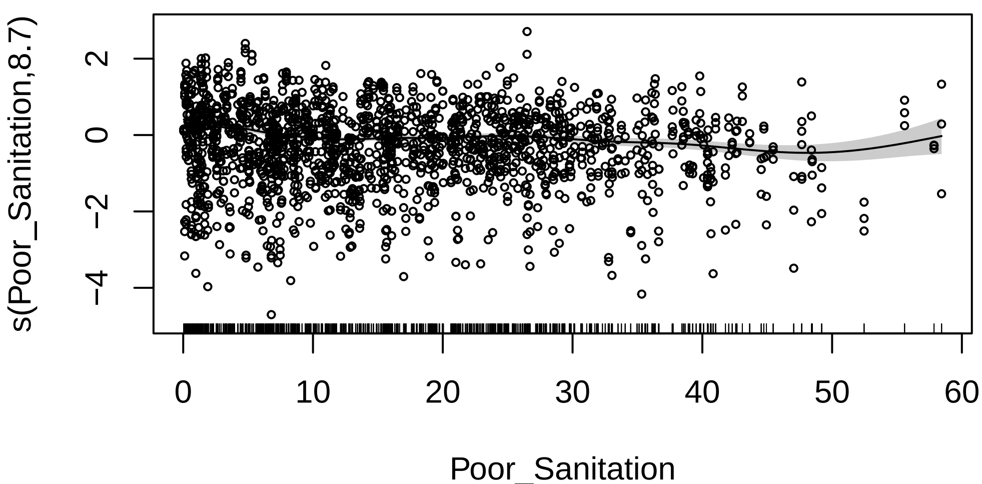

s(Indigenous,2.87)
2
0
−2
−4
0
10
20
30
40
50
Indigenous

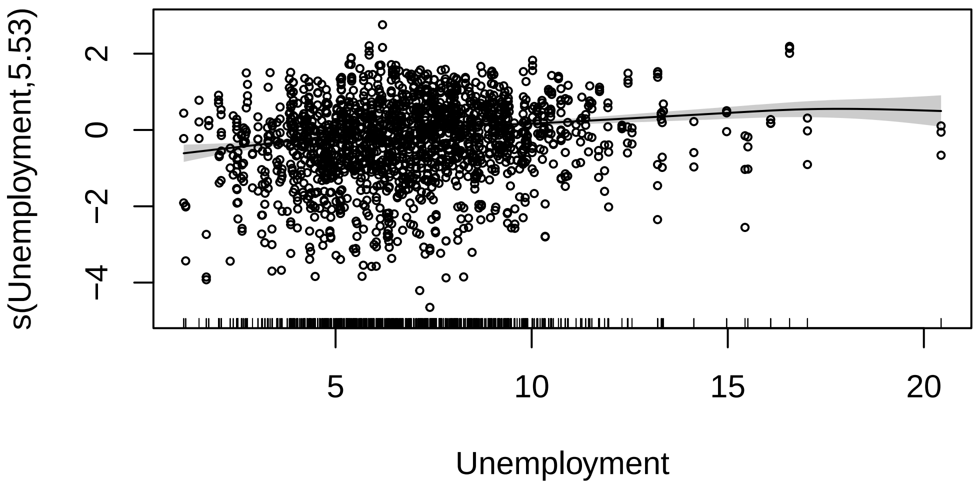

s(Illiteracy,4.77)
2
0
−2
−4
10
20
30
40
Illiteracy


s(Urbanisation,6.6)
2
0
−2
−4
20
40
60
80
100
Urbanisation

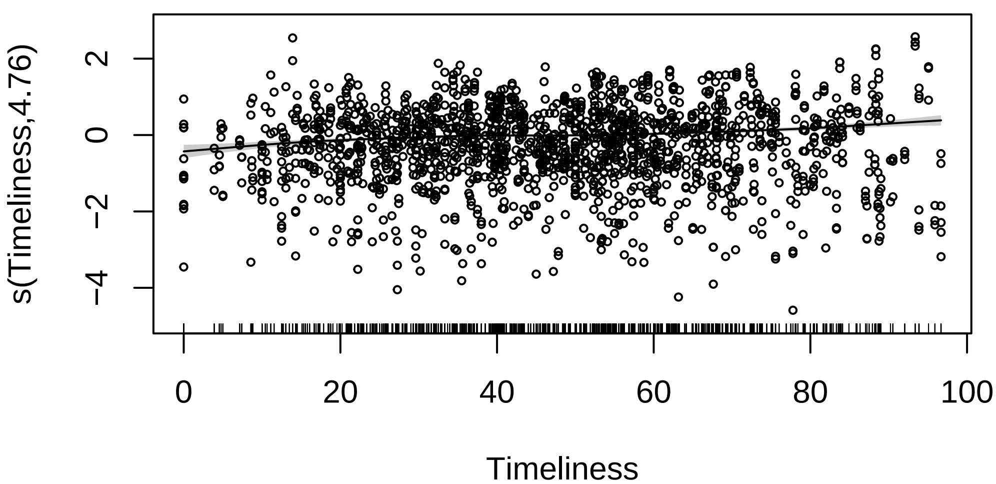

s(Density,6.94)
2
0
−2
−4
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Density


s(Poverty,2.03)
2
0
−2
−4
10
20
30
40
50
60
70
80
Poverty

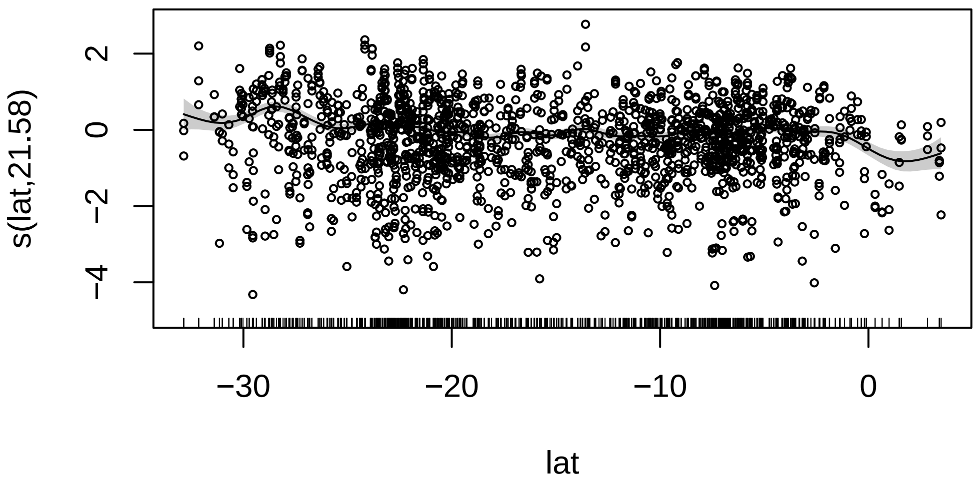

s(Poor_Sanitation,8.7)
2
0
−2
−4
0
10
20
30
40
50
60
Poor_Sanitation



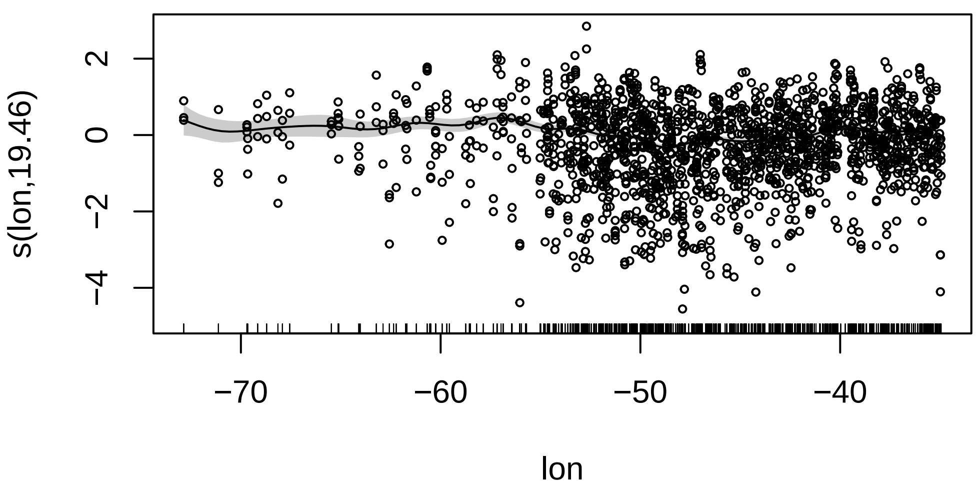
s(Unemployment,5.53)
2
0
−2
−4
5
10
15
20
Unemployment



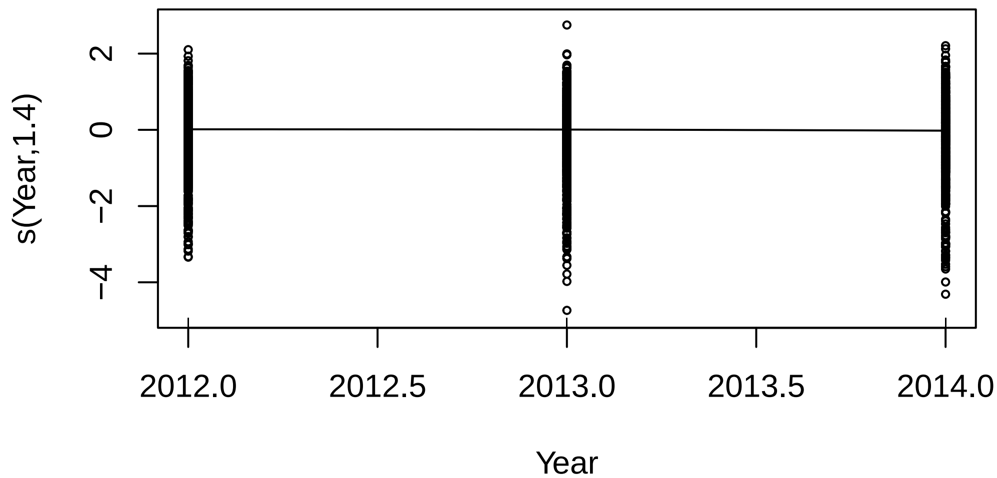
s(Timeliness,4.76)
2
0
−2
−4
0
20
40
60
80
100
Timeliness

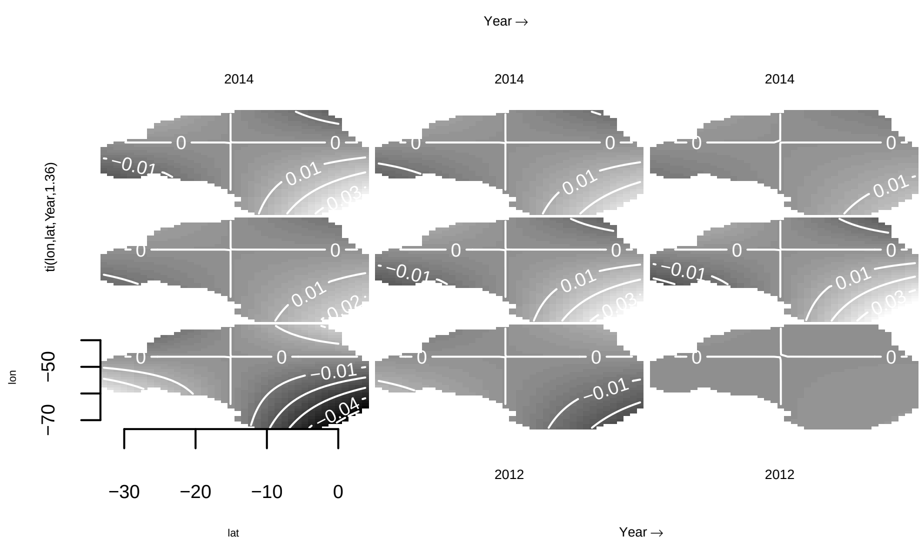


s(lat,21.58)
2
0
−2
−4
−30
−20
−10
0
lat

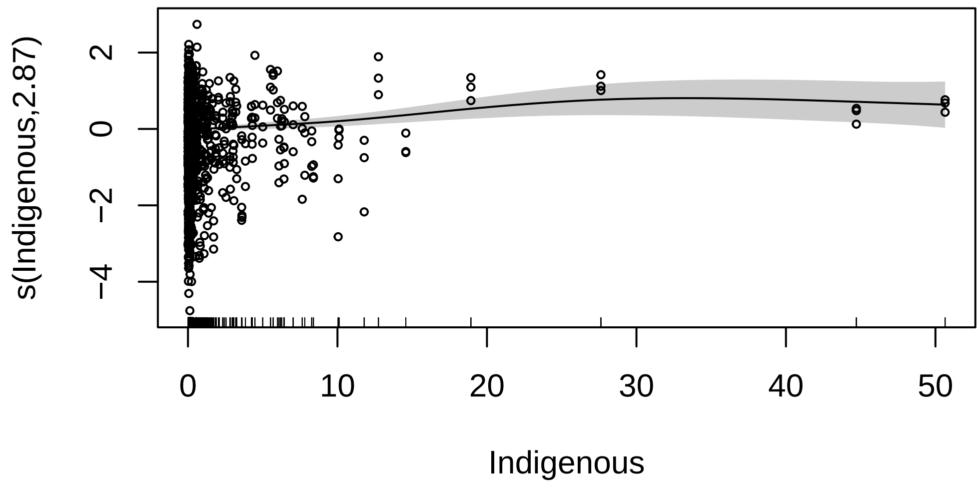


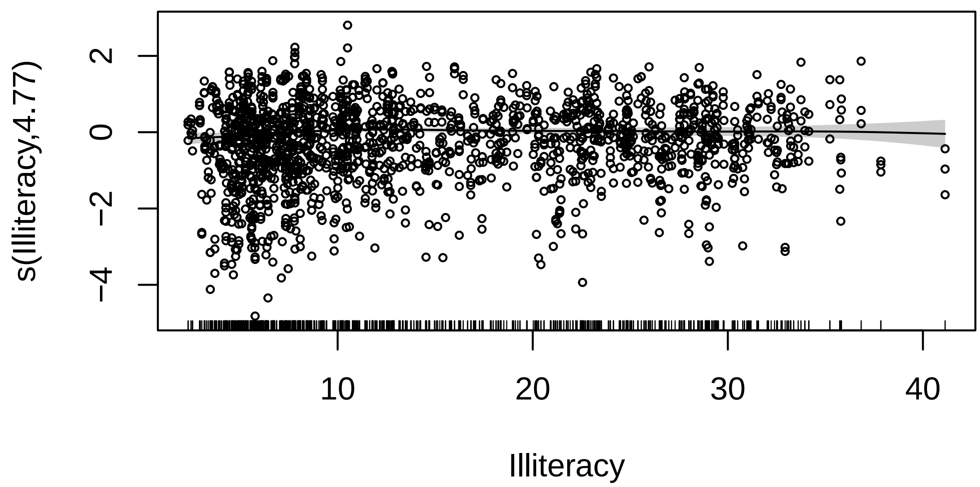
s(lon,19.46)
2
0
−2
−4
−70
−60
−50
−40
lon

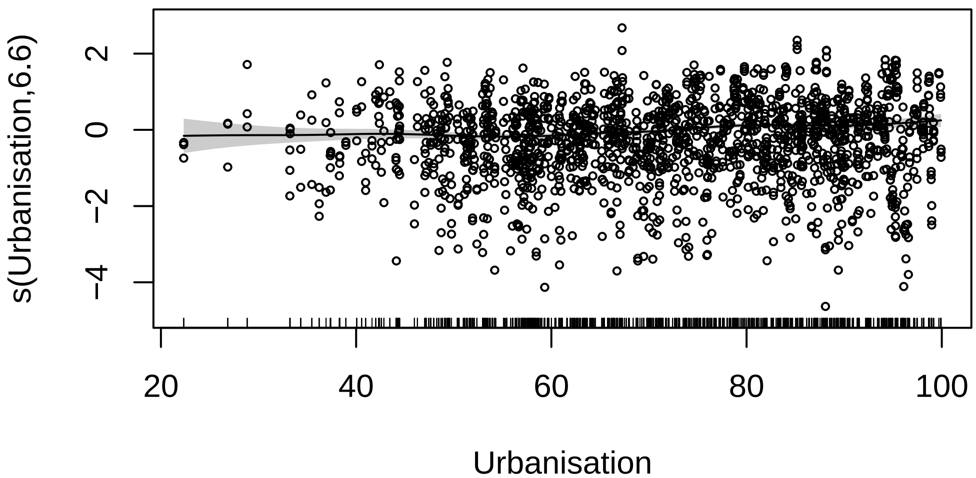

s(Year,1.4)
2
0
−2
−4
2012.0
2012.5
2013.0
2013.5
2014.0
Year


Year →
2014
2014
2014
ti(lon,lat,Year,1.36)
lon
−50
−70
−30
−20
−10
0
lat
2012
2012
Year →

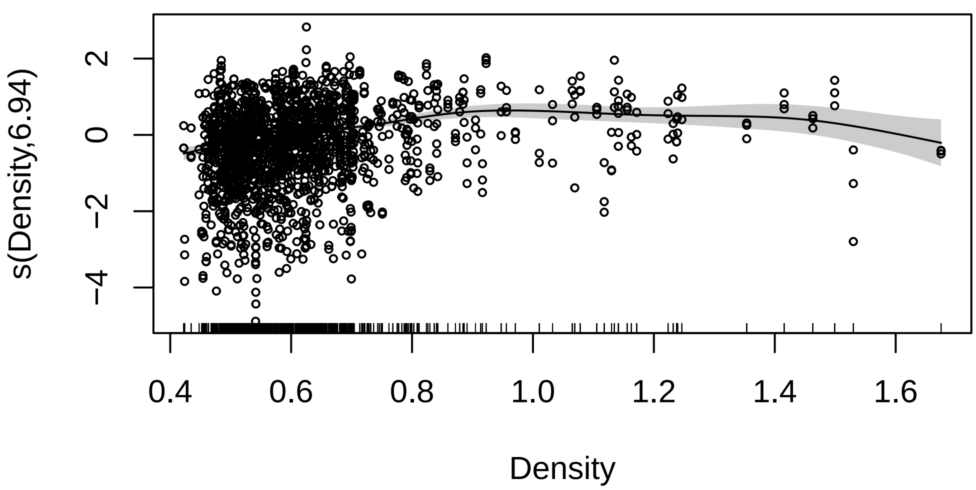
```r
plot(nb_model, shade = T, rug = TRUE, residuals = TRUE, scheme = 2, pch = 1,
    cex = 0.5)
```



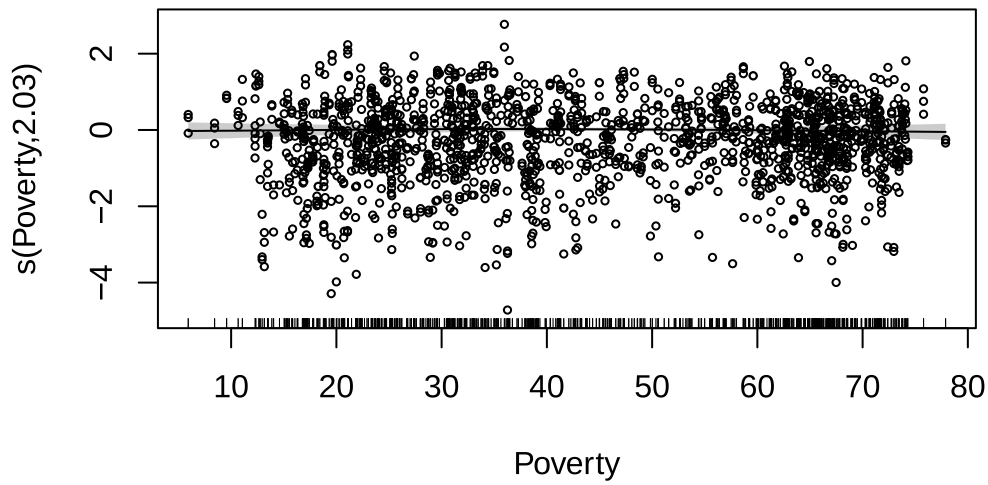


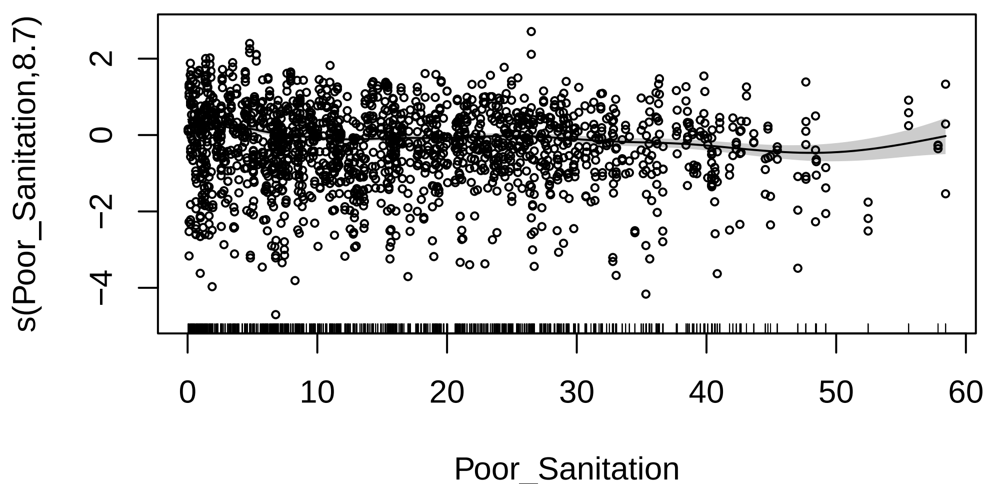


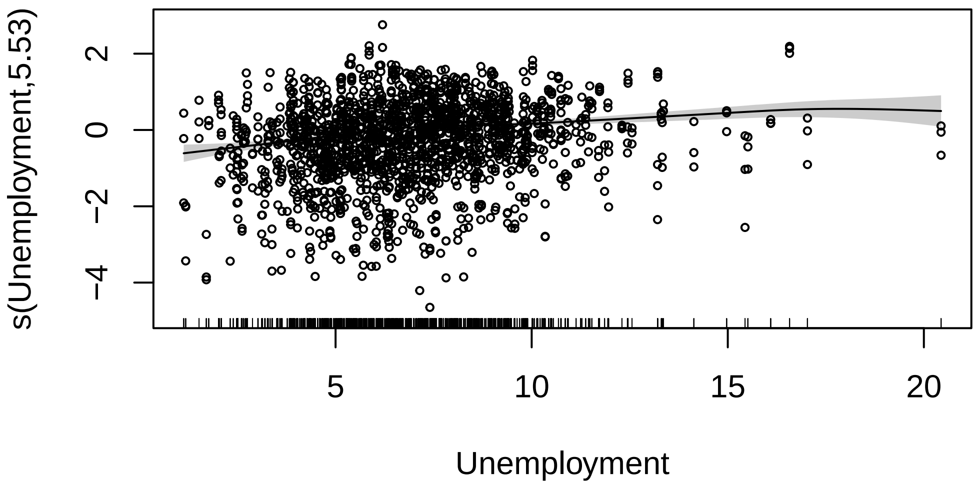
s(Urbanisation,6.6)
2
0
−2
−4
20
40
60
80
100
Urbanisation

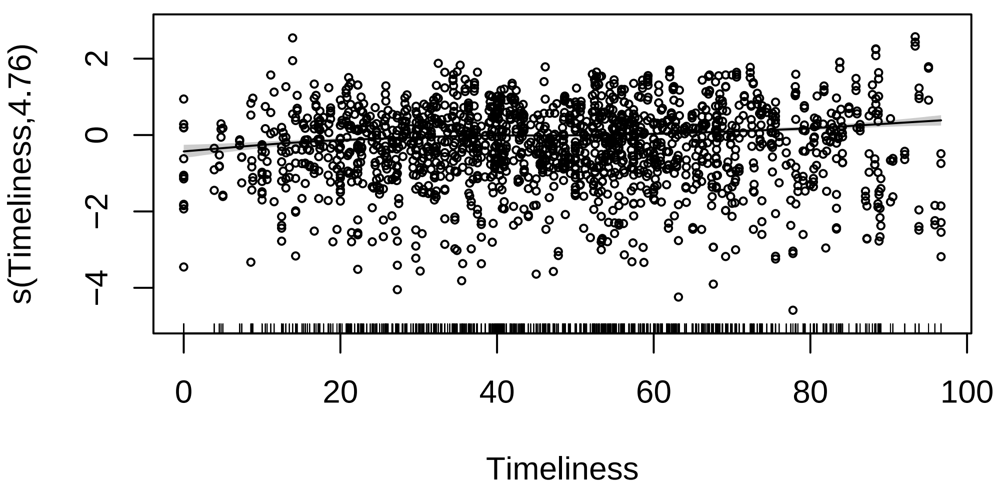


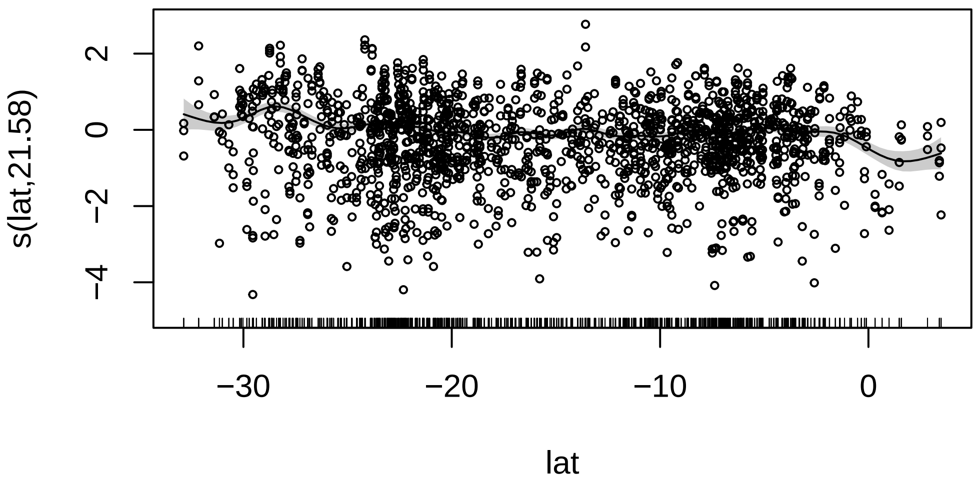
s(Density,6.94)
2
0
−2
−4
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Density


s(Poverty,2.03)
2
0
−2
−4
10
20
30
40
50
60
70
80
Poverty

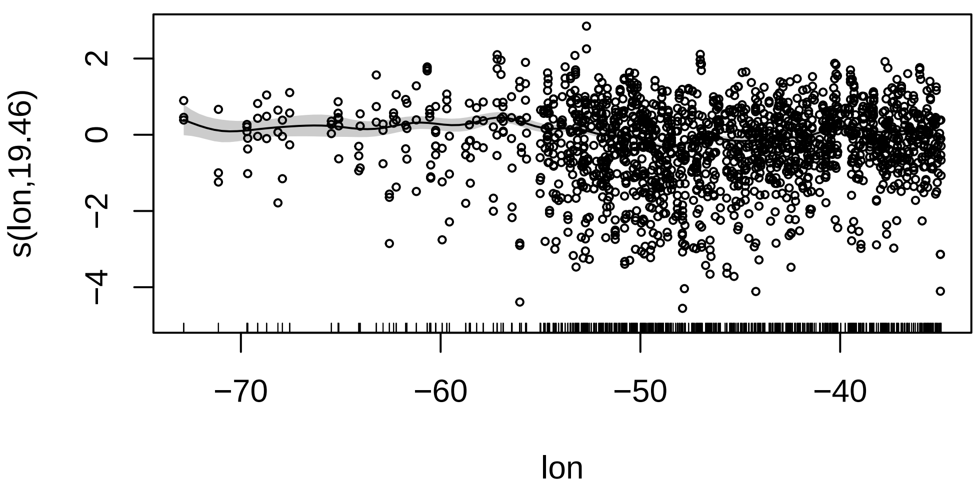

s(Poor_Sanitation,8.7)
2
0
−2
−4
0
10
20
30
40
50
60
Poor_Sanitation


s(Unemployment,5.53)
2
0
−2
−4
5
10
15
20
Unemployment

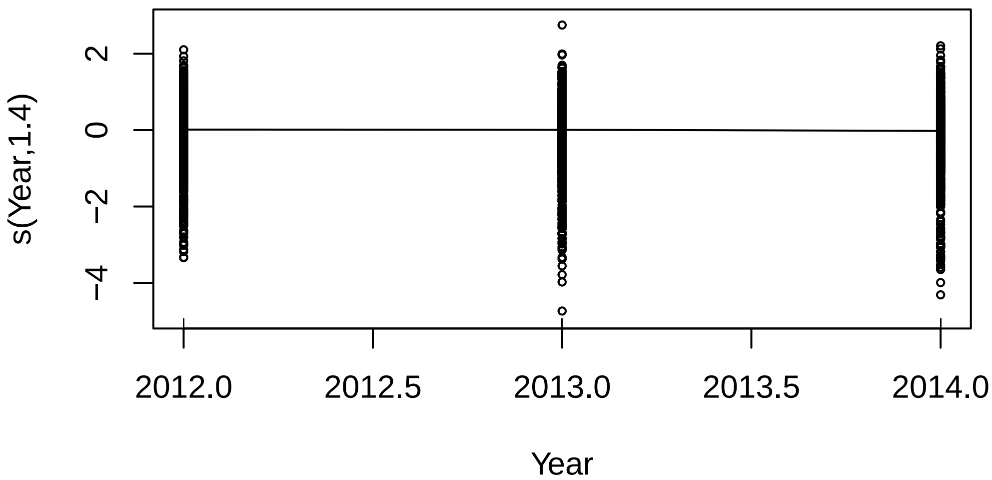

s(Timeliness,4.76)
2
0
−2
−4
0
20
40
60
80
100
Timeliness

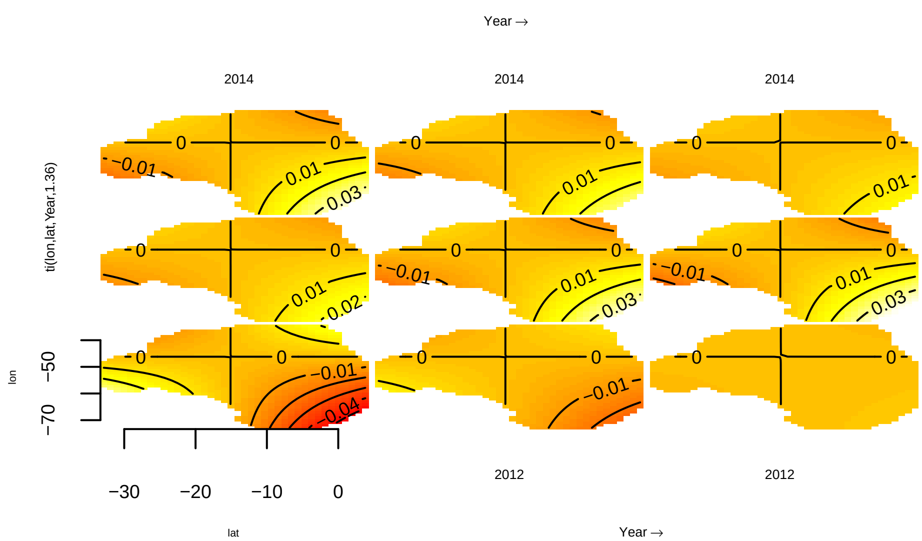


s(lat,21.58)
2
0
−2
−4
−30
−20
−10
0
lat



s(lon,19.46)
2
0
−2
−4
−70
−60
−50
−40
lon





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