MTHM506 Statistical Data Modelling Group Project

Analysis of 2012–2014 Brazil Tuberculosis Data

Group 2

March 2023

1 Introduction

1.1 Problem Statement

Analysis of tuberculosis (TB) data originating from Brazil using Generalized Additive Models (GAMs). Brazil is divided into 557 administrative microregions, and the available data comprises counts of TB cases in each microregion for each of the years from 2012 to 2014.

1.2 Exploratory Analysis of Data and Problem

The TB data from Brazil includes 1,671 entries or samples with 14 columns of numeric data types that specify the characteristics of each sample. The columns are self-explanatory because they are called Indigenous, Illiteracy, Urbanisation, Density, Poverty, Poor Sanitation, Unemployment, Timeliness, Year, TB, Population, Region, lon, and lat. TB stands for tuberculosis, whereas lon and lat stand for longitude and latitude. The dataset has no missing values in the technical sense, but it contains some abnormalities, which increases the amount of pre-processing needed. The region is stored as a continuous variable despite being a factor variable. Nonetheless, changing it depends on the task at hand. Moreover, the collection includes coordinates that describe the precise geospatial locations of the micro-regions listed in the column. The next part gives a detailed exploration of the data.

1.3 Data Exploration

An in-depth analysis of the datasets reveals that the mean and median values for the indigenous population are low, but the maximum value is 50, which suggests that there are individual areas where the indigenous population is concentrated and that these areas may be areas of potential poverty and poor sanitation and should be areas where there are more cases of tuberculosis. The mean and median illiteracy rates are only 14 and 11, respectively. However, the maximum value of 41 suggests that there are specific backward areas with significant populations lacking access to education, which also suggests that the area seems poor and has poor sanitation. There are still some places that are less urbanised, where there may be more occurrences of tuberculosis, but the mean, median, and minimum values for urbanisation are 70 and 22, respectively, suggesting that most areas are highly urbanised. Based on the population density data, most locations can fit one person in a room, but the highest value of 1.6 highlights the existence of some places with very high population densities, which sharply raise the rate of TB transmission. The distribution of the poverty data suggests that each district has different poverty levels, with only a limited number of districts where poverty is not a significant issue. Although the general results on inadequate sanitation are low, the maximum score of 58 indicates that some districts have poor sanitation and substantial disease risk. Although average unemployment rates are low, a maximum value of 20 indicates that some isolated regions may experience severe economic hardship, protracted social unrest, and potentially significant morbidity rates. With a minimum value of 0, notification timeliness data has a fairly wide range.

Timeliness, Unemployment, and Urbanization approximately follow a Normal Distribution with few extremes, whilst the remainder is multimodal normal distributions. The above suggests that employing semi-parametric or non-parametric models to demonstrate the relationship between the target and predictors would be advantageous. The target variable in this study is a risk, defined as 'TB/Population',' whereas the remaining variables are possible predictors. As demonstrated in the table below, most features in the dataset exhibit some connection. As some features are correlated, basic regression cannot be used because it would yield false results; rather, models that account for the connection can be used. It is vital to note that some characteristics are anticipated to have positive correlations (tuberculosis versus population, density versus poverty) and vice versa. Specifically, population density, poverty, health conditions, unemployment, and notification timeliness are likely high due to the high population density, the low economic share per capita, the high poverty rate, and the high jobless rate. See Figure 2 for a correlogram of the 8 socio-economic covariates

2 Model

We want to model the count of cases TB_i by actually modelling ρ_i using

$$TB_i \sim Pois(\lambda_i = z_i\rho_i) \; TB_i \text{ indep.}$$
$$log(\lambda_i) = log(z_i) + log(\rho_i)$$

where TB_i is the count of TB cases. We are using the canonical link function - log. z_i is the total population, which is taken as an offset. Model $log(\rho_i)$ as

$$log(\rho_i) = \sum_{j=1}^{8} f_j(x_{i,j})$$
$$f(x_i) = \sum_{k=1}^{q} \beta_k b_k(x_i)$$

where $x_{i,j}$ is the jth covariate (out of 8 socio-economic covariates) for the ith instance/datum in the dataset, $f(\cdot)$ is a smooth function of said covariate and $b_k(\cdot)$ is a basis function with k knots. Hence, the model boils down to

$$TB_i \sim Pois(\lambda_i = z_i\rho_i) \; TB_i \text{ indep.}$$
$$log(\lambda_i) = log(z_i) + \sum_{j=1}^{8}\sum_{k=1}^{q} \beta_{j,k} b_{j,k}(x_{i,j})$$

Looking at the distribution of the residuals of the model, we can see that the data is clearly far too overdispersed to be modelled by a Poisson, which has a fixed dispersion parameter. Even with 60 knots per smooth term the model doesn't seem to have enough flexibility which may be another indicator that a Poisson model is unsuitable for the data. The residuals vs fitted plot fans out, indicating that the model does not have enough flexibility to fit well (the edf is also close to the maximum degrees of freedom, and increasing the number of knots doesn't resolve the problem). We propose the conventional alternative to the Poisson - the Negative Binomial model. Doing so, leads to a drop in the AIC. So the model distribution is changed to Negative Binomial with the same parameterisation except for the feature that the count of TB cases is now Negative Binomial distributed with mean as described above. See Table 1 appendix for a showcase of different model configurations and their associated AIC.

$$TB_i \sim NB(\lambda_i, \sigma_i^2) \; TB_i \text{ indep.}$$
$$\lambda_i = z_i\rho_i; \; \sigma_i^2 = \lambda_i + \frac{\lambda_i^2}{\phi}$$
$$log(\lambda_i) = log(z_i) + \sum_{j=1}^{8}\sum_{k=1}^{q} \beta_{j,k} b_{j,k}(x_{i,j})$$

where ϕ is a dispersion parameter, later estimated by the `gam` function in R.

When having a look at the relationship between the squared residuals and the fitted values one sees that the relation is not exactly quadratic, but rather close to 0, which would reflect the relation between model variance and the expected value in a Gaussian Distribution Model (additional evidence is provided by the Residuals vs. Fitted plot). However, fitting a Gaussian model leads to very skewed residuals, indicating that the data is apparently not Gaussian. So the model distribution is changed to Negative Binomial with the same parameterization except for the feature that the count of TB cases is now Negative Binomial distributed with mean λ_i as described above.

Given this base model, we investigate whether all given socio-economic variables are needed to explain the response or whether there exists a model with fewer parameters. The p-value for the smooth term of Illiteracy points towards it not being statistically significant. Poverty, although not statistically insignificant, has the second-largest p-value. These terms are sequentially dropped and the resulting model checked against the original model via a Likelihood Ratio Test (conducted using the `anova` function in R). We find that leaving out Illiteracy does not alter the model at a 5%-level of significance, whereas taking out both Poverty and Illiteracy does. So, in the following, we use a model with all of the socio-economic variables except Illiteracy. Note that this converts our linear predictor to

$$log(\lambda_i) = log(z_i) + \sum_{j=1}^{7} \sum_{k=1}^{q} \beta_{j,k} b_{j,k}(x_{i,j})$$

This leaves us with a model with AIC = 14,391.19 and 43.9% of deviance explained. Running `gam.check()` lets us analyse the residual plots (see Figure) and examine the basis functions for the model. The QQ plot tells us that the model fails to predict well on the upper and lower ends of the response variable. Increasing the knots to 20 per covariate leads to marginal improvement with 44.9% deviance explained. More efficient extensions can be to add 1) spatial, 2) temporal and 3) spatio-temporal covariates.

First, we will try adding spatial terms. The spatial model adds a smoothed term which is function of the longitude and the latitude. A bivariate function is used because it makes sense to assume that there are more cases at certain locations (defined by the interaction between latitude and longitude) than others, rather than there being more cases at locations with a certain longitude for any latitude, or the other way round. Hence, our linear predictor is now

$$log(\lambda_i) = log(z_i) + \sum_{j=1}^{7} \sum_{k=1}^{q} \beta_{j,k} b_{j,k}(x_{i,j}) + \sum_{k=1}^{q} \beta_k b_k(lon_i, lat_i)$$

Using this model with the regular `s` smoother function from the `mgcv` package leads to a model that can explain 56.4% of the deviance and has a slightly lower AIC of 14,013.13. The QQ plot still points to the upper and lower tails being incorrectly predicted. At the cost of significantly more computation, using a tensor product smooth `te` on the bivariate spatial term with 20 knots allows us to make a decent improvement on this. See Appendix for different numbers of knots that were tested. This gets us to 69.9% deviance explained. The QQ plot looks considerably better with only a few problematic instances at the top and bottom quantiles.

We contest this with an extension on the model with only socio-economic covariates, but instead of adding spatial terms, we add the temporal dimension `Year`. The linear predictor becomes

$$log(\lambda_i) = log(z_i) + \sum_{j=1}^{7} f_{2012,j}(x_{i,j}) \times x_{2012} + \sum_{j=1}^{7} f_{2013,j}(x_{i,j}) \times x_{2013} + \sum_{j=1}^{7} f_{2014,j}(x_{i,j}) \times x_{2014}$$

where the new terms $x_{2012}, x_{2013}, x_{2014}$ are indicator variables equating to 1 if `Year` is respectively 2012,2013,2014 and zero otherwise. Exercising some shorthand, it can be expressed as

$$log(\lambda_i) = log(z_i) + \sum_{t=2012}^{2014} \sum_{j=1}^{7} f_{t,j}(x_{i,j}) \times x_t$$

where x_t is now the indicator variable for `Year`. A slightly separate approach can be tested with `Year` as a covariate instead of a grouping variable. In that case, the linear predictor would be

$$log(\lambda_i) = log(z_i) + \sum_{j=1}^{7} f_{t,j}(x_{i,j}) + \sum_{t=2012}^{2014} \beta_t x_t$$

Neither of the temporal formulations show much increase in deviance explained (the one with year as grouping variable actually shows a decrease to 41.5%!). Their QQ plots are also much worse than the spatial model, showing gross deviations on high as well as low quantiles. Finally, we create a spatio-temporal model, including both `Year` as well as `lon,lat`. Its linear predictor is formulated as below

$$log(\lambda_i) = log(z_i) + \sum_{t=2012}^{2014} \left(\sum_{j=1}^{7} \sum_{k=1}^{q} \beta_{t,j,k} b_{t,j,k}(x_{t,i,j}) + \sum_{k=1}^{q} \beta_{t,k} b_{t,k}(lon_{i,t}, lat_{i,t}) \right) \times x_t$$

This is a model which includes the term for the location and estimates a functional relation for each year and each explaining variable. The AIC of this model does not drop compared to the spatial model. Reasons for this may be that 3 factors is perhaps not enough granularity to discern any effect from the temporal dimension. Naive estimates would point towards the spread being higher in winter months as TB spreads through inhaling tiny droplets from coughs or sneezes. Having more granular data at the season or month level could bring to light any temporal patterns, if they exist.

So the spatial model (given that it is simpler) is the model we choose to best explain the ratio of TB cases per capita. To recall, it is formulated as

$$log(\lambda_i) = log(z_i) + \sum_{j=1}^{7}\sum_{k=1}^{q} \beta_{j,k} b_{j,k}(x_{i,j}) + \sum_{k=1}^{q} \beta_k b_k(lon_i, lat_i)$$

Considering the fit of the spatial model, it fits well even though the largest residuals are higher than expected from the model distribution. For districts that have a high number of cases, the predictor does not seem as accurate. But the highest residuals do not arise when the ratio of TB cases per capita is extraordinarily high, but rather when the absolute number of TB cases is high (see residuals vs. response). The variance of the model still seems too low for those values given that there are some predicted values in that high segment of response values (absolute number of TB cases) where the prediction for the response value is lower than the actual value, and some where the prediction of the actual value is higher than the actual value. Using this model, we predict the rate of TB per 100,000 inhabitants. See Figure 1

3 Critical Review

Drawbacks of the Model:

1. Predictions do not cover full range of data, as evinced by deviations in the QQ plot

2. The independence assumption of the `Year` variable can be questioned in two ways, thus providing justification for leaving it out: The data points of a certain district in 2012, 2013 and 2014 may not be fully independent of one another because it is likely that the conditions in that region have not substantially changed. The model seems unnecessarily complex, if we add additional smooth terms for each variable grouped on `Year` - it hardly brings any additional explantory power. Another possible violation of the independece assumption arises from spatial correlation - the fact that regions which are located closely to one another are mutually dependent of one another in terms of the number of TB cases as well as the socio-economic determinants of the spread of infectious diseases. Parametric coefficients of the `Year` factor in `temporal.model` barely differ from each other, at around -8.4 for `Year`=2012, and varying by 0.004 and -0.038 respectively for 2013 and 2014, which respectively correspond to multipliers of 1.004 and 0.963 on the response scale.

3. The fitted vs. response plot shows a deviance from the 45-degree line for high absoulute values of the number of cases. One reason for this is that the model is designed to predict the ratio of TB cases per capita, not the absolute number. Another reason is that the model variance increases with the mean value.

4 Conclusions

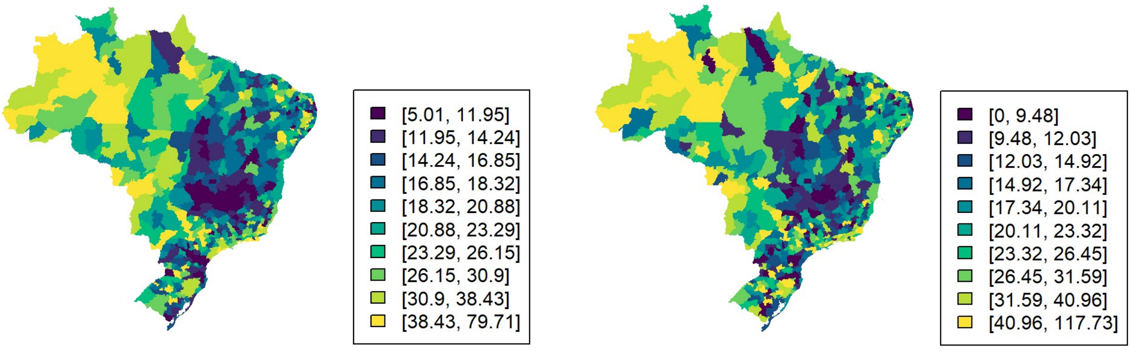
Firstly, based on the Correlogram we can conclude that no single socio-economic covariate has much linear correlation with TB incidence, but illiteracy, urbanisation, poverty, sanitation, unemployment and timeliness of notification are all weakly correlated with TB incidence, and given that there are more strong correlations between these socio-economic covariates, an increase in illiteracy, poverty, unemployment and poor sanitation will simultaneously correlate to decreases in urbanisation and timeliness of notification, all contributing to increases in TB incidence. Poverty is strongly correlated with several socio-economic covariates and is the primary factor that governments need to improve. Sanitation and urbanisation are also more strongly correlated, implying that good urban infrastructure and quality health resources have a greater impact on reducing TB incidence. According to our predicted TB incidence map, Brazil's central region has a lower incidence overall, so we recommend that the health sector invest significant health resources to improve the current situation throughout Brazil's north-west, followed by localised areas in the south and east, although these areas could also rely on assistance from neighbouring regions with fewer cases to improve their situation.

5 References

1. Wood, S. N. (2017). Generalized Additive Models: An Introduction with R (2nd ed.). CRC Press.

6 Appendix

6.1 Figures



(a) Predicted

(b) True

Figure 1: Predicted (a) and True (b) rates of TB per 100k inhabitants

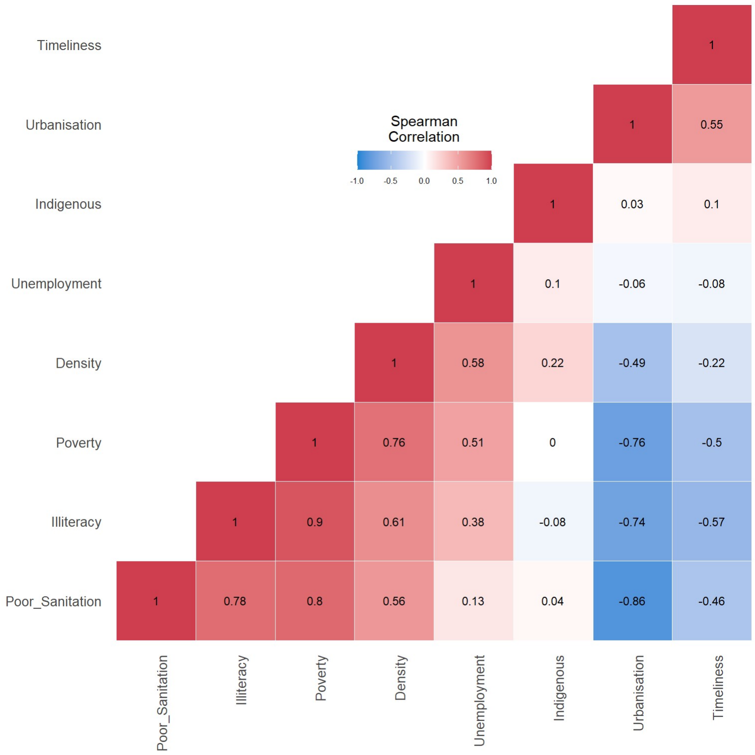


Figure 2: Correlogram shows covariates with highest positive and negative correlations.

6.2 Tables

Model	AIC	Deviance Explained
model_poisson	34047.36	66.9%
Gadgets	13	14

Table 1: An example table.

6.3 Code

```r
# Load Data
load("C:/Users/soura/Documents/COMM511/group_coursework/datasets_project.RData")
# Import Libraries
library(mgcv) # required for GAM

#fit poisson model with socio-economic variables
model_poisson <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous)
+ s(Illiteracy) + s(Urbanisation) + s(Density) + s(Poverty) + s(Poor_Sanitation)
+ s(Unemployment) + s(Timeliness),
data = TBdata,
family = poisson(link = 'log')
)

#add flexibility
model_poisson.more.knots <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous, k = 60)
+ s(Illiteracy, k = 60) +  s(Urbanisation, k = 60) + s(Density, k = 60)
+ s(Poverty, k = 60) + s(Poor_Sanitation, k = 60) + s(Unemployment, k = 60)
+ s(Timeliness, k = 60),
data = TBdata,
family = poisson(link = 'log')
)
# check summary
gam.check(model_poisson.more.knots)
summary(model_poisson.more.knots) # SIGNS OF OVERFIT
plot(model_poisson.more.knots) # SIGNS OF OVERFIT

#fit negative binomial model with  socioeconomic
model_nb <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous)
+  s(Illiteracy) + s(Urbanisation) + s(Density) + s(Poverty) + s(Poor_Sanitation)
+ s(Unemployment) + s(Timeliness),
data = TBdata, family = nb(link = 'log')
)
#Show summary
summary(model_nb)
model_nb$aic

#fit a linear relation between squared residuals and prediction to see whether another model describes
#the variance-fitted values relation better
summary(lm(log(model_nb$residuals^2) ~ log(predict(model_nb, type = 'response'))))

#drop Illiteracy
model_nb_2 <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous) +  s(Urbanisation) + s(Density)
+ s(Poverty) + s(Poor_Sanitation) + s(Unemployment) + s(Timeliness),
data = TBdata,
family = nb(link = 'log')
)

# Likelihood ratio test
anova.gam(model_nb_2, model_nb, test = 'F') # p-value is over 0.05
# The models are statistically indistinguishable

model_nb_3 <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous) +  s(Urbanisation) + s(Density)
+ s(Poor_Sanitation) + s(Unemployment) + s(Timeliness),
data = TBdata,
family = nb(link = 'log')
)
# Likelihood ratio test
anova.gam(model_nb_3, model_nb_2, test = 'F')# p-value is less than 0.05
# The models are statistically different. Poverty should not be excluded.

```

```
### Model chosen (with socio-economic covariates) is the negative binomial without Illiteracy
### as the effect of illiteracy cannot be reliably stated to be non-zero
summary(model_nb_2)# Only 44% of the deviance is explained. Adding temporal
# and spatial covariates may improve this
par(mfrow=c(2,2))
gam.check(model_nb_2)

### Increasing the knots to 20 yields minor improvement - discarded
prelim.model.4 <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous , 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),
data = TBdata ,
family = nb(link = 'log')
)
# Show summary
summary(prelim.model.4)
par(mfrow=c(2,2))
gam.check(prelim.model.4)

#### Introducing temporality as a grouping variable
#Temporal model
par(mfrow = c(2,2), pch = 20)
model_nb_time <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous, by = Year)
+ s(Urbanisation, by = Year) + s(Density, by = Year) + s(Poverty, by = Year) + s(Poor_Sanitation, by = Year)
+ s(Unemployment, by = Year) + s(Timeliness, by = Year),
data = TBdata,
family = nb(link = 'log')
)
# Show summary
summary(model_nb_time)
model_nb_time$aic

#### Temporality as a covariate
TBdata$Year.asFactor <- factor(TBdata$Year)

temporal.model <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous)
+ s(Urbanisation) + s(Density) + s(Poor_Sanitation) + s(Unemployment) + s(Poverty)
+ s(Timeliness) + Year.asFactor,
data = TBdata ,
family = nb(link = 'log')
)
# Check summary
summary(temporal.model) # Temporal alone doesn't add much to explaining the variance

par(mfrow=c(2,2))
gam.check(temporal.model)
par(mfrow = c(1,1))

### Adding spatial covariates
spatial.model <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous)
+ s(Urbanisation) + s(Density) + s(Poor_Sanitation) + s(Unemployment) +s(Poverty)
+ s(Timeliness) + s(lon , lat),
data = TBdata ,
family = nb(link = 'log')
)
# Check summary
summary(spatial.model)

par(mfrow=c(2,2))
gam.check(spatial.model)
par(mfrow = c(1,1))
spatial.model$aic

```

```
### Using separate smoothers
spatial.model.2 <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous)
+ s(Urbanisation) + s(Density) + s(Poor_Sanitation) + s(Unemployment) + s(Poverty)
+ s(Timeliness) + te(lon , lat , k = 20),
data = TBdata ,
family = nb(link = 'log')
)
# Check summary
summary(spatial.model.2)

spatial.model.2$aic
par(mfrow=c(2,2))
gam.check(spatial.model.2 , pch = 20)
par(mfrow = c(1,1))

# Check if this model is significantly different from one with only socio-economic covariates
anova.gam(spatial.model.2, model_nb_2, test = 'LRT')

#### Spatio-temporal model
spatio.temporal.model <- gam(formula = TB ~ offset(log(Population)) + s(Urbanisation, by = Year.asFactor)
+ s(Density, by = Year.asFactor) + s(Poverty, by = Year.asFactor)
+ s(Poor_Sanitation, by = Year.asFactor) + s(Timeliness, by = Year.asFactor)
+ s(Unemployment, by = Year.asFactor) + te(lon,lat, by = Year.asFactor), data = TBdata, family = nb(link = 'log'))
summary(spatio.temporal.model)
spatio.temporal.model$aic
gam.check(spatio.temporal.model)

### Spatio-temporal model (with Year as parametric covariate)
spatio.temporal.model.2 <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous)
+ s(Urbanisation) + s(Density) + s(Poor_Sanitation) + s(Unemployment) + s(Poverty)
+ s(Timeliness) + te(lon , lat , k = 20) + Year.asFactor,
data = TBdata ,
family = nb(link = 'log')
)
# Check summary
summary(spatio.temporal.model.2) # Temporal dimension as a parametric variable explains
# slightly more of the deviance
spatio.temporal.model.2$aic
gam.check(spatio.temporal.model.2 , pch = 20)

```

6.4 Code Outputs

```
# check summary
summary(model_poisson)

Family: poisson
Link function: log

Formula:
TB ~ offset(log(Population)) + s(Indigenous) + s(Illiteracy) +
    s(Urbanisation) + s(Density) + s(Poverty) + s(Poor_Sanitation) +
    s(Unemployment) + s(Timeliness)

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.449827   0.004199   -2012   <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)      8.961  8.999  569.4  <2e-16 ***
s(Illiteracy)      8.989  9.000 2704.0  <2e-16 ***
s(Urbanisation)    8.900  8.996 1490.4  <2e-16 ***
s(Density)         8.985  9.000 1758.4  <2e-16 ***
s(Poverty)         8.956  8.999 1470.2  <2e-16 ***
s(Poor_Sanitation) 8.979  9.000 1327.0  <2e-16 ***
s(Unemployment)    8.993  9.000 2423.5  <2e-16 ***
s(Timeliness)      8.352  8.864  600.7  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.976   Deviance explained = 66.9%
UBRE = 13.899  Scale est. = 1         n = 1671

```