Group Project Report

2023-03-19

The Model

As the target variable is the count of TB cases, the basic form the model which is designed to explain the ratio of TB cases per capita has the form

$$TB_{i,t} \sim Pois(\eta_{i,t})$$

$$\log(\eta_{i,t}) = \log(Population_{i,t}) + \sum_{j=1}^{8} f_j(x_{i,t,j})$$

where $x_{i,t,j}$ for $j \in 1, ..., 8$ is the value one of the socio-economic variables mentioned in the task-description for year t. Having this model the model coefficients are explaining the relation between the explaining variables and the ratio of TB cases per capita:

Looking at the distribution of the residuals of the model one sees that the Poisson model, which has a fixed dispersion parameter is clearly overdispersed. 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 η_i as described above.

Given the ground model we investigate whether all given socio-economic variables are needed to explain the ratio of TB cases per capita or whether there is a less complex model. There for we try to drop the variables with the highest p-values for the hypothesis test $\beta_j = 0$ and perform an LRT to see whether the reduced model is as good as the more complex model. Leaving one variable out is repeated until the reduced model is significantly worse than more complex model. Dropping the Illiteracy variable does not make the model significantly worse. Next, the Poverty variable which has the 2nd lowest p-value for $\beta_j = 0$ in the initial model is dropped aditionally, but then the null hypothesis of the LRT that this model is as good as the model which only leaves out Illiteracy can be rejected at 5%-level. So in the following we use a model with all of the socio-economic variables except for Illiteracy.

Further extensions to the model can be reached by including 1) time, 2) space, or 3) both.

The temporal model changes the expression for η_i as follows:

$$\log(\eta_i) = \log(Population_{i,t}) + \beta_0 + \sum_{t=2012}^{2014} \sum_{j=1}^{7} \beta_{t,j} f_{t,j}(x_{i,t,j})$$

where $x_{i,t,j}$ is the value of the variable index by j for year t. For this model the AIC does not drop compared to the model which does not consider time.

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 that there are more cases at locations with a certain longitude for any latitude, or the other way round.

$$\log(\eta_i) = \log(Population_{i,t}) + \beta_0 + \sum_{j=1}^{7} \beta_j f_j(x_{i,t,j}) + \beta_8 f_8(lon_{i,t}, lat_{i,t})$$

Furthermore, there 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, so the spatial model is - given that it is simpler - the model which explains the ratio of TB cases per capita best.

$$\log(\eta_i) = \log(Population_{i,t}) + \beta_0 + \sum_{t=2012}^{2014}\left(\sum_{j=1}^{7}\beta_{t,j}f_{t,j}(x_{i,t,j}) + \beta_s f_s(lon_{i,t}, lat_{i,t})\right)$$

Let us know have a look at 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 as it should. 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 stil seems too low for those values given that there are some predicted values in that high segment of response values (absolut 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.

Code

```
library (mgcv)
```

```
## Loading required package: nlme

## This is mgcv 1.8-42. For overview type 'help("mgcv-package")'.
```

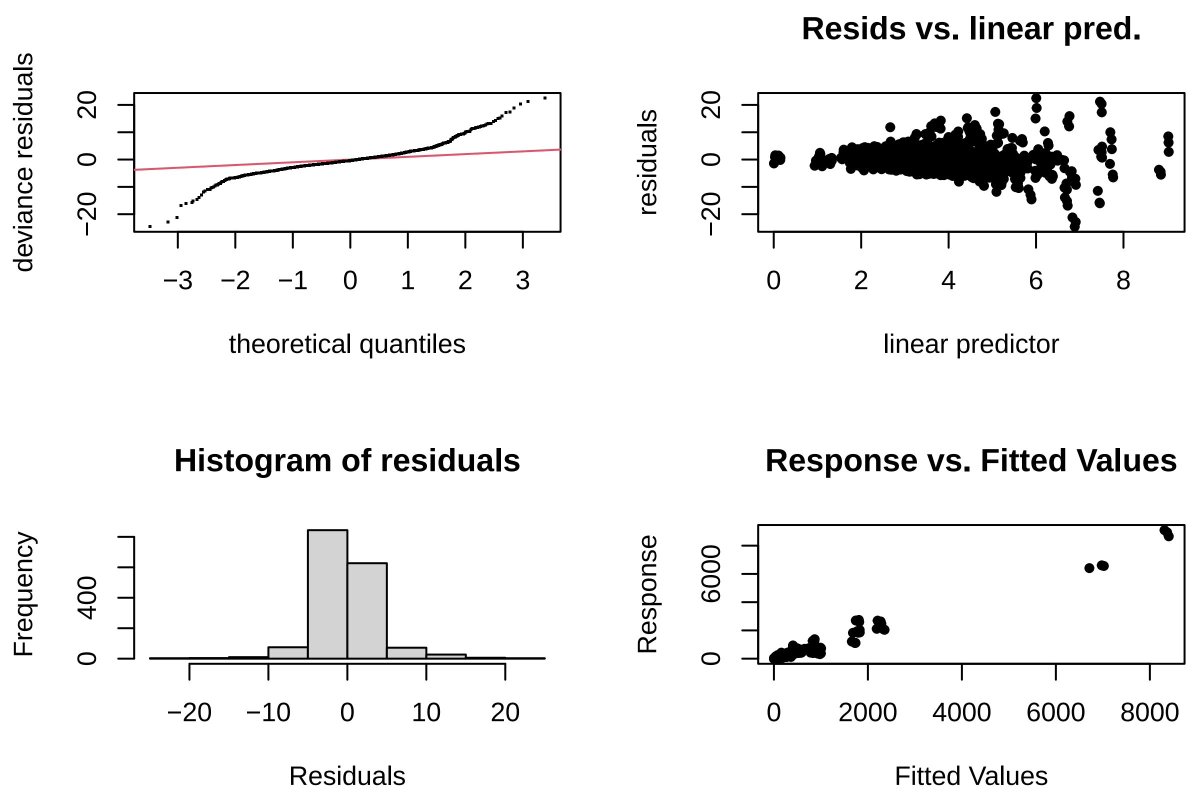
```
par(mfrow = c(2,2))
#fit poisson model with socio-economic variables
model_poisson <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous) +  s(Illiteracy) +  s(Urbani
summary(model_poisson)
```

```
##
## 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
```

```
model_poisson$aic
```

```
## [1] 34047.36
```

```
par(mfrow = c(2,2),pch = 20)
gam.check(model_poisson)
```



```
##
## Method: UBRE   Optimizer: outer newton
## full convergence after 13 iterations.
## Gradient range [2.754533e-08,1.177865e-05]
## (score 13.89908 & scale 1).
## Hessian positive definite, eigenvalue range [6.78691e-06,0.0006391723].
## Model rank =  73 / 73
##
## 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)     9.00 8.96    0.39  <2e-16 ***
## s(Illiteracy)     9.00 8.99    0.41  <2e-16 ***
## s(Urbanisation)   9.00 8.90    0.41  <2e-16 ***
## s(Density)         9.00 8.98    0.39  <2e-16 ***
## s(Poverty)         9.00 8.96    0.39  <2e-16 ***
## s(Poor_Sanitation) 9.00 8.98    0.40  <2e-16 ***
## s(Unemployment)    9.00 8.99    0.39  <2e-16 ***
## s(Timeliness)      9.00 8.35    0.43  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#fit negative binomial model with  socioeconomic
model_nb <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous) +  s(Illiteracy) + s(Urbanisatio
summary(model_nb)
```

```
##
## Family: Negative Binomial(6.146)
## 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) -8.42871    0.01094  -770.5   <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)      1.489  1.795  20.396 2.92e-05 ***
## s(Illiteracy)      1.008  1.017   0.246  0.63129
## s(Urbanisation)    6.634  7.773  24.089  0.00148 **
## s(Density)         4.579  5.672 132.693  < 2e-16 ***
## s(Poverty)         5.733  6.911  17.934  0.01516 *
## s(Poor_Sanitation) 6.123  7.297  73.103  < 2e-16 ***
## s(Unemployment)    5.798  7.000  62.050  < 2e-16 ***
## s(Timeliness)      4.101  5.097  64.474  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.861   Deviance explained = 43.9%
## -REML = 7237.2  Scale est. = 1         n = 1671
```

```
model_nb$aic
```

```
## [1] 14391.19
```

```
#drop Illiteracy
model_nb_2 <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous) +  s(Urbanisation) + s(Density)


#LRT
anova.gam(model_nb_2, model_nb, test = 'LRT')
```

```
## Analysis of Deviance Table
##
```

```
## Model 1: TB ~ offset(log(Population)) + s(Indigenous) + s(Urbanisation) +
##     s(Density) + s(Poverty) + s(Poor_Sanitation) + s(Unemployment) +
##     s(Timeliness)
## Model 2: TB ~ offset(log(Population)) + s(Indigenous) + s(Illiteracy) +
##     s(Urbanisation) + s(Density) + s(Poverty) + s(Poor_Sanitation) +
##     s(Unemployment) + s(Timeliness)
##   Resid. Df Resid. Dev      Df Deviance Pr(>Chi)
## 1    1621.3      14319
## 2    1620.3      14318 0.99721  0.34543   0.5556
```

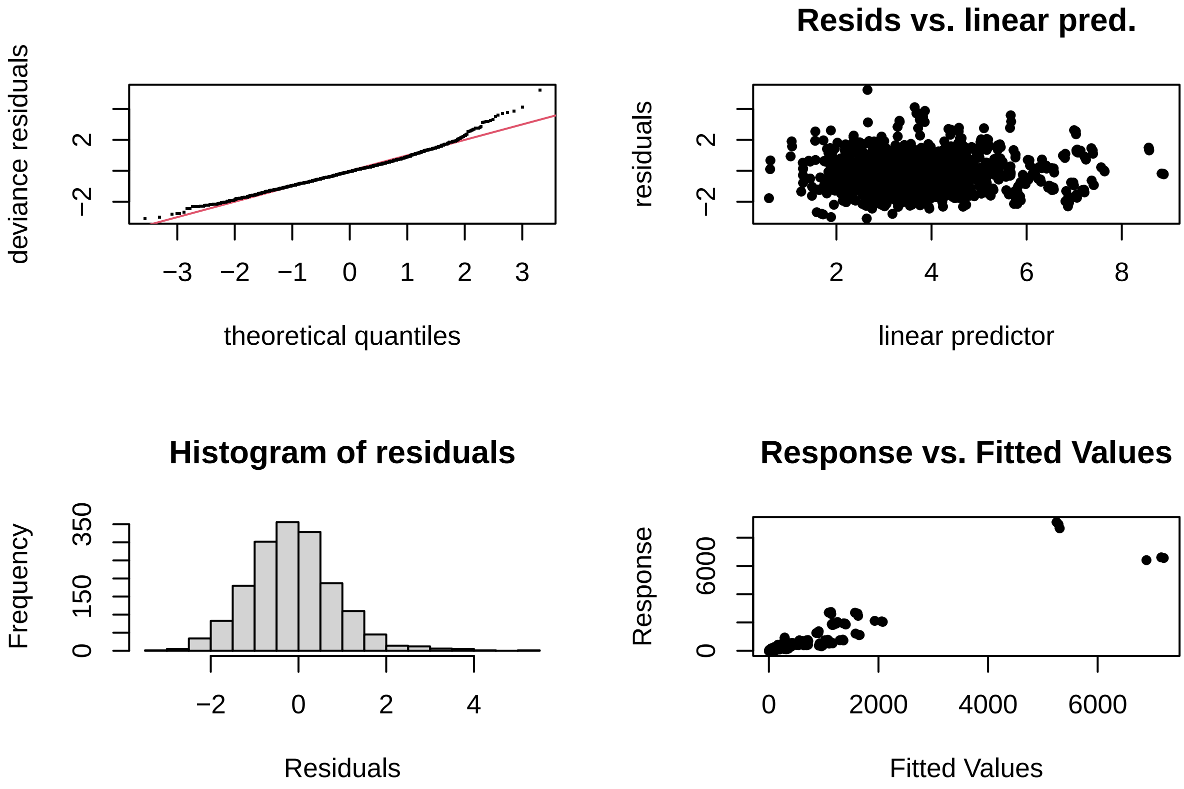
```
#Null hypothesis not rejected -> drop poverty
model_nb_3 <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous) +  s(Urbanisation) + s(Density
#LRT
anova.gam(model_nb_3, model_nb_2, test = 'LRT')
```

```
## Analysis of Deviance Table
##
## Model 1: TB ~ offset(log(Population)) + s(Indigenous) + s(Urbanisation) +
##     s(Density) + s(Poor_Sanitation) + s(Unemployment) + s(Timeliness)
## Model 2: TB ~ offset(log(Population)) + s(Indigenous) + s(Urbanisation) +
##     s(Density) + s(Poverty) + s(Poor_Sanitation) + s(Unemployment) +
##     s(Timeliness)
##   Resid. Df Resid. Dev     Df Deviance Pr(>Chi)
## 1    1630.5      14345
## 2    1621.3      14319 9.1169   26.444 0.001861 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model_nb_final <- model_nb_2
summary(model_nb_final)
```

```
##
## Family: Negative Binomial(6.146)
## Link function: log
##
## Formula:
## TB ~ offset(log(Population)) + s(Indigenous) + s(Urbanisation) +
##     s(Density) + s(Poverty) + s(Poor_Sanitation) + s(Unemployment) +
##     s(Timeliness)
##
## Parametric coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.42863    0.01094  -770.6   <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)      1.518  1.833  21.13 2.08e-05 ***
## s(Urbanisation)    6.610  7.752  23.73  0.00167 **
## s(Density)         4.578  5.667 147.64  < 2e-16 ***
## s(Poverty)         5.771  6.945  21.36  0.00394 **
## s(Poor_Sanitation) 6.119  7.293  76.07  < 2e-16 ***
## s(Unemployment)    5.776  6.977  64.21  < 2e-16 ***
## s(Timeliness)      4.106  5.103  66.42  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.86   Deviance explained = 43.9%
## -REML = 7234.9  Scale est. = 1         n = 1671
```

```
gam.check(model_nb_final)
```



```
##
## Method: REML   Optimizer: outer newton
## full convergence after 6 iterations.
## Gradient range [-5.66156e-06,7.265887e-06]
## (score 7234.878 & scale 1).
## Hessian positive definite, eigenvalue range [0.1154615,588.9389].
## Model rank =  64 / 64
##
## 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)       9.00 1.52    0.49  <2e-16 ***
## s(Urbanisation)     9.00 6.61    0.50  <2e-16 ***
## s(Density)          9.00 4.58    0.50  <2e-16 ***
## s(Poverty)          9.00 5.77    0.49  <2e-16 ***
## s(Poor_Sanitation) 9.00 6.12    0.50  <2e-16 ***
## s(Unemployment)     9.00 5.78    0.50  <2e-16 ***
## s(Timeliness)       9.00 4.11    0.56  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#temporal model
model_nb_time <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous, by = Year) + s(Urbanisation
summary(model_nb)
```

```
##
## Family: Negative Binomial(6.146)
## 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) -8.42871    0.01094  -770.5   <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)      1.489  1.795  20.396 2.92e-05 ***
## s(Illiteracy)      1.008  1.017   0.246  0.63129
## s(Urbanisation)    6.634  7.773  24.089  0.00148 **
## s(Density)         4.579  5.672 132.693  < 2e-16 ***
## s(Poverty)         5.733  6.911  17.934  0.01516 *
## s(Poor_Sanitation) 6.123  7.297  73.103  < 2e-16 ***
## s(Unemployment)    5.798  7.000  62.050  < 2e-16 ***
## s(Timeliness)      4.101  5.097  64.474  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.861   Deviance explained = 43.9%
## -REML = 7237.2  Scale est. = 1         n = 1671
```

```
model_nb$aic
```

```
## [1] 14391.19
```

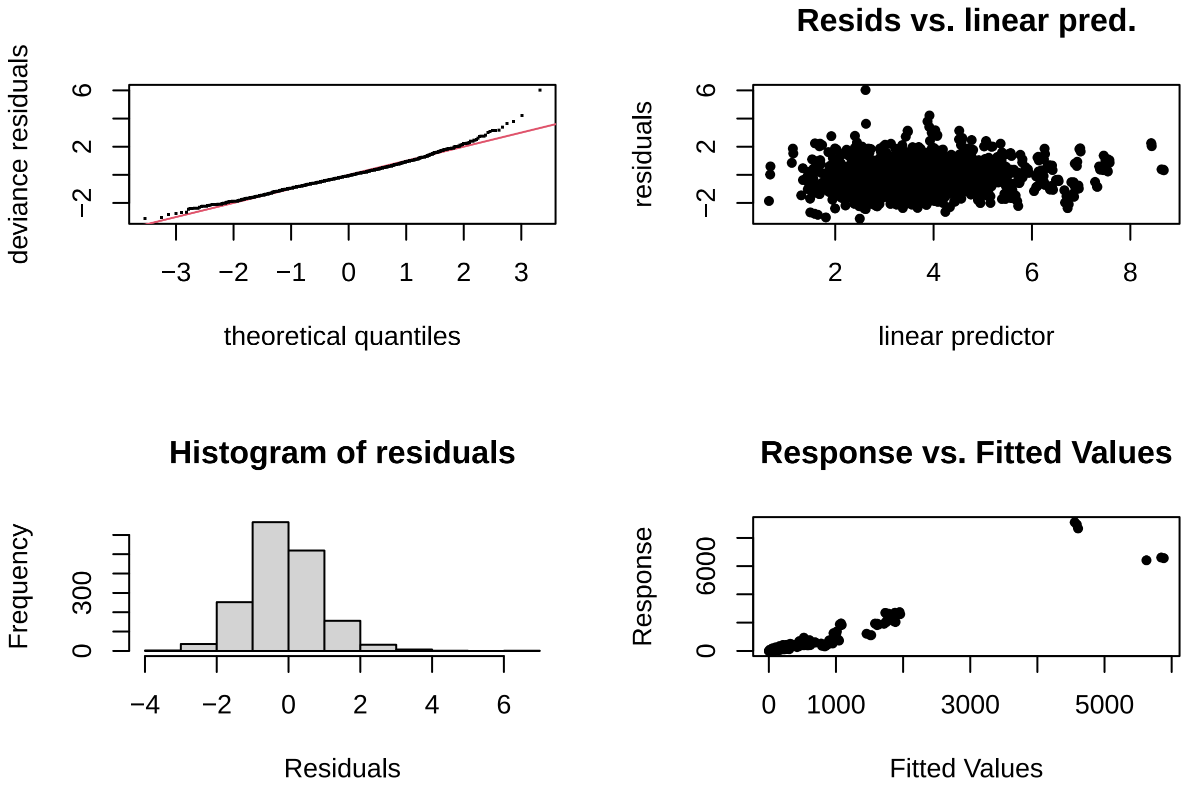
```
#spatial model
model_nb_space <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous) +  s(Illiteracy) + s(Urbani
summary(model_nb_space)
```

```
##
## Family: Negative Binomial(8.318)
## 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) + s(lon, lat)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.44740    0.00972  -869.1   <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)        1.002  1.004  11.968 0.000548 ***
## s(Illiteracy)        3.235  4.101   7.349 0.125235
## s(Urbanisation)      4.943  6.095  34.874 3.33e-06 ***
## s(Density)           4.222  5.286  45.898  < 2e-16 ***
## s(Poverty)           1.954  2.494   7.667 0.027494 *
## s(Poor_Sanitation)  6.009  7.173  62.150  < 2e-16 ***
## s(Unemployment)      4.388  5.496  77.071  < 2e-16 ***
## s(Timeliness)        4.083  5.079  73.290  < 2e-16 ***
## s(lon,lat)          26.060 28.422 483.248  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.847   Deviance explained = 56.6%
## -REML = 7069.9  Scale est. = 1         n = 1671
```

```
model_nb_space$aic
```

```
## [1] 14008.78
```

```
gam.check(model_nb_space)
```



```
##
```

```
## Method: REML   Optimizer: outer newton
## full convergence after 5 iterations.
## Gradient range [-0.0006694501,0.00015351]
## (score 7069.921 & scale 1).
## Hessian positive definite, eigenvalue range [0.0006676393,510.0227].
## Model rank =  102 / 102
##
## 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)        9.00  1.00    0.54  <2e-16 ***
## s(Illiteracy)        9.00  3.24    0.53  <2e-16 ***
## s(Urbanisation)      9.00  4.94    0.54  <2e-16 ***
## s(Density)           9.00  4.22    0.53  <2e-16 ***
## s(Poverty)           9.00  1.95    0.54  <2e-16 ***
## s(Poor_Sanitation)  9.00  6.01    0.53  <2e-16 ***
## s(Unemployment)      9.00  4.39    0.54  <2e-16 ***
## s(Timeliness)        9.00  4.08    0.61  <2e-16 ***
## s(lon,lat)          29.00 26.06    0.50  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova.gam(model_nb_space, model_nb_final)
```

```
## Analysis of Deviance Table
##
## Model 1: TB ~ offset(log(Population)) + s(Indigenous) + s(Illiteracy) +
##     s(Urbanisation) + s(Density) + s(Poverty) + s(Poor_Sanitation) +
##     s(Unemployment) + s(Timeliness) + s(lon, lat)
## Model 2: TB ~ offset(log(Population)) + s(Indigenous) + s(Urbanisation) +
##     s(Density) + s(Poverty) + s(Poor_Sanitation) + s(Unemployment) +
##     s(Timeliness)
##   Resid. Df Resid. Dev      Df Deviance
## 1    1596.9      13895
## 2    1621.3      14319 -24.469  -423.61
```

```
#spatio-temporal model
model_nb_time_and_space <- gam(formula = TB ~ offset(log(Population)) + s(Urbanisation, by = Year) + s(
summary(model_nb_time_and_space)
```

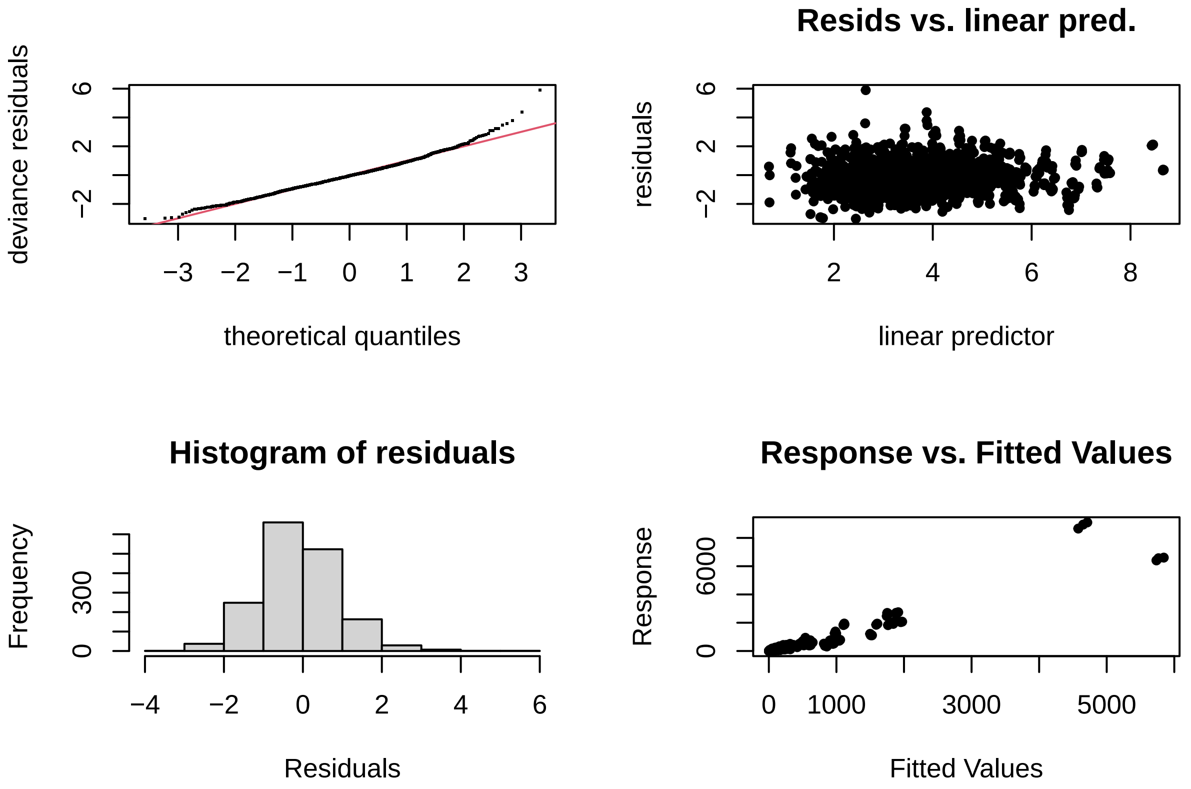
```
##
## Family: Negative Binomial(8.258)
## Link function: log
##
## Formula:
## TB ~ offset(log(Population)) + s(Urbanisation, by = Year) + s(Density,
##     by = Year) + s(Poverty, by = Year) + s(Poor_Sanitation, by = Year) +
##     s(Timeliness, by = Year) + s(Unemployment, by = Year) + s(lon,
##     lat, by = Year)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    31.98      23.70   1.349    0.177
##
```

```
## Approximate significance of smooth terms:
##                              edf Ref.df Chi.sq  p-value    
## s(Urbanisation):Year       4.770  5.901  32.32 2.27e-05 ***
## s(Density):Year            4.886  6.050  51.54  < 2e-16 ***
## s(Poverty):Year            3.198  4.041  14.75  0.00549 ** 
## s(Poor_Sanitation):Year    5.519  6.687  55.25  < 2e-16 ***
## s(Timeliness):Year         4.273  5.303  75.51  < 2e-16 ***
## s(Unemployment):Year       4.391  5.501  85.93  < 2e-16 ***
## s(lon,lat):Year           27.207 29.476 493.44  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Rank: 85/91
## R-sq.(adj) =  0.851   Deviance explained = 56.2%
## -REML = 7129.5  Scale est. = 1         n = 1671
```

```
model_nb_time_and_space$aic
```

```
## [1] 14019.8
```

```
gam.check(model_nb_time_and_space)
```



```
## 
## Method: REML   Optimizer: outer newton
## full convergence after 5 iterations.
## Gradient range [-8.903051e-06,0.0001832814]
## (score 7129.522 & scale 1).
## Hessian positive definite, eigenvalue range [0.3581337,509.4814].
## Model rank =  85 / 91
##
## 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(Urbanisation):Year    10.00  4.77    0.53  <2e-16 ***
## s(Density):Year         10.00  4.89    0.53  <2e-16 ***
## s(Poverty):Year         10.00  3.20    0.53  <2e-16 ***
## s(Poor_Sanitation):Year 10.00  5.52    0.53  <2e-16 ***
## s(Timeliness):Year      10.00  4.27    0.60  <2e-16 ***
## s(Unemployment):Year    10.00  4.39    0.54  <2e-16 ***
## s(lon,lat):Year         30.00 27.21    0.50  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.