MTHM506 Group Project Report

Group 2

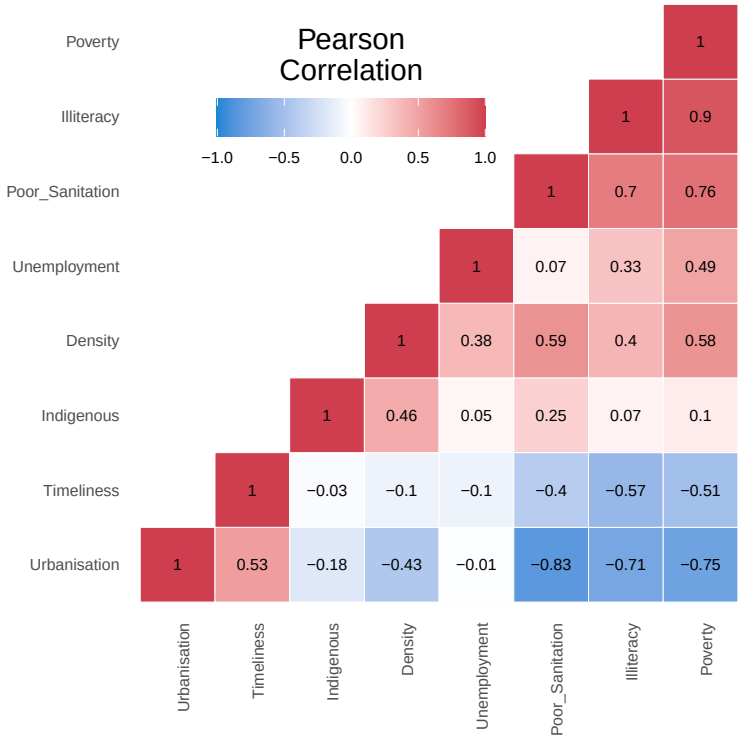
2023-03-20

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

Appendix

```
# Import Libraries
library(mgcv) # required for GAM
library(tidyverse)
library(ggplot2) # required for plotting
library(dplyr) # required for filtering dataset
library(fields) # required for maps
library(maps) # required for maps
library(reshape2) # only required for melt in corr plot
library(car) # only required for VIF

# Load Data
load('datasets_project.RData')

# Investigate correlation
### Resource -> http://www.sthda.com/english/wiki/ggplot2-quick-
### correlation-matrix-heatmap-r-software-and-data-visualization
cormat <- cor(TBdata[,c(1,2,3,4,5,6,7,8)])
# Reorder
reorder_cormat <- function(cormat){
  # Use correlation between variables as distance
  dd <- as.dist((1-cormat)/2)
  hc <- hclust(dd)
  cormat <-cormat[hc$order, hc$order]
}

# Reorder the correlation matrix
cormat <- reorder_cormat(cormat)
# Get lower triangular matrix
cormat[lower.tri(cormat)] <- NA

melted_cormat <- melt(cormat , na.rm = TRUE)
melted_cormat$value = round(melted_cormat$value, 2)

# Create a ggheatmap
ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value))+
```

```
 geom_tile(color = "white")+
 scale_fill_gradient2(low = "#1a85d6", high = "#cf3e4f", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
    name="Pearson\nCorrelation") +
 theme_minimal()+ # minimal theme
 theme(axis.text.x = element_text(angle = 90, vjust = 1,
    size = 12, hjust = 1))+
 coord_fixed()

# Add correlation coefficients
ggheatmap +
geom_text(aes(Var2, Var1, label = value), color = "black", size = 2) +
theme(
  axis.text.x = element_text(size = 6),
  axis.text.y = element_text(size = 6),
  axis.title.x = element_blank(),
  axis.title.y = element_blank(),
  panel.grid.major = element_blank(),
  panel.border = element_blank(),
  panel.background = element_blank(),
  axis.ticks = element_blank(),
  legend.justification = c(1, 0),
  legend.position = c(0.6, 0.7),
  legend.direction = "horizontal",
  legend.text = element_text(size = 6)
  ) +
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                title.position = "top", title.hjust = 0.5))
```



```
#### Illiteracy is highly correlated with Poverty
#### Carry out a Variance Inflation Test
model_all <- lm(TB ~ . , data = select(TBdata, 'Indigenous' , 'Illiteracy' ,
'Urbanisation' , 'Density' , 'Poverty', 'Unemployment' , 'Timeliness' , 'Year' ,
'TB' , 'Population'))  # with all the independent variables

vif(model_all) # Several variables are highly correlated
```

```
##   Indigenous   Illiteracy Urbanisation      Density      Poverty Unemployment
##     1.451133     6.623719     4.153478     2.499462    13.815146     2.360352
##   Timeliness         Year   Population
##     1.642305     1.000056     1.117419
```

```
model_no_illiteracy <- lm(TB ~ . , data = select(TBdata, 'Indigenous',
'Urbanisation' , 'Density' , 'Poverty', 'Unemployment' , 'Timeliness' , 'Year' ,
'TB' , 'Population'))  # with all the independent variables

vif(model_no_illiteracy) # Poverty and Unemployment still seem highly correlated
```

```
##   Indigenous Urbanisation      Density      Poverty Unemployment   Timeliness
##     1.417551     4.101467     2.262345     5.978840     2.233231     1.588746
##         Year   Population
##     1.000056     1.116952
```

```
model_no_illiteracy_no_poverty <- lm(TB ~ . , data = select(TBdata, 'Indigenous',
'Urbanisation' , 'Density', 'Unemployment' , 'Timeliness' , 'Year' ,
'TB' , 'Population'))  # with all the independent variables

vif(model_no_illiteracy_no_poverty) # almost no variable is highly correlated
```

```
##   Indigenous Urbanisation      Density Unemployment   Timeliness         Year
##     1.307511     1.946109     1.959290     1.295781     1.484281     1.000056
##   Population
##     1.115992
```

```
## More formal tests are conducted to confirm the dropping of Illiteracy.
## Check to see if Poverty should be dropped as well
prelim.model.1 <- 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(prelim.model.1)
```

```
##
## 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
```

```
### Only the effect of illiteracy cannot be reliably stated to be non-zero
prelim.model.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')
)
# Show summary
summary(prelim.model.2)
```

```
##
## 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
```

```
# Likelihood ratio test
anova(prelim.model.1 , prelim.model.2 , test = 'F') # p-value is over 0.05
```

```
## 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)
## 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    1620.3      14318
## 2    1621.3      14319 -0.99721 -0.34543   0.5556
```

```
# The models are statistically indistinguishable

### Only the effect of illiteracy cannot be reliably stated to be non-zero
prelim.model.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')
)
# Show summary
summary(prelim.model.3)
```

```
##
## Family: Negative Binomial(6.069)
## Link function: log
##
## Formula:
## TB ~ offset(log(Population)) + s(Indigenous) + s(Urbanisation) +
##     s(Density) + s(Poor_Sanitation) + s(Unemployment) + s(Timeliness)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.42781    0.01099  -766.8   <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)       1.868  2.265  32.29 1.33e-06 ***
## s(Urbanisation)     6.934  8.025  36.71 8.70e-06 ***
## s(Density)          4.249  5.281 152.98  < 2e-16 ***
## s(Poor_Sanitation) 6.105  7.281  90.00  < 2e-16 ***
## s(Unemployment)     5.682  6.881  82.74  < 2e-16 ***
## s(Timeliness)       4.137  5.140  78.14  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.856   Deviance explained =   43%
## -REML = 7237.4  Scale est. = 1         n = 1671
```

```
# Likelihood ratio test
anova(prelim.model.2 , prelim.model.3 , test = 'F') # p-value is less than 0.05
```

```
## 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(Urbanisation) +
##     s(Density) + s(Poor_Sanitation) + s(Unemployment) + s(Timeliness)
##   Resid. Df Resid. Dev      Df Deviance Pr(>Chi)
## 1    1621.3      14319
## 2    1630.5      14345 -9.1169  -26.444 0.001861 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# The models are statistically different. Poverty should not be excluded.

### Model chosen (with social covariates) is the negative binomial without
### Illiteracy
summary(prelim.model.2) # Only 44% of the deviance is explained. Adding temporal
```

```
##
## 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
```

```
# and spatial covariates may improve this

### Adding spatial covariates
```