Colombia COVID-19 - Central region

A project for the *Statistical Methods for Data Science* course

Angela Carraro, Giullia Monteiro Milano Oliveira, Gaia Saveri

29th July 2020

UniTS - DSSC

Aim of the project

We decided to do central Colombia

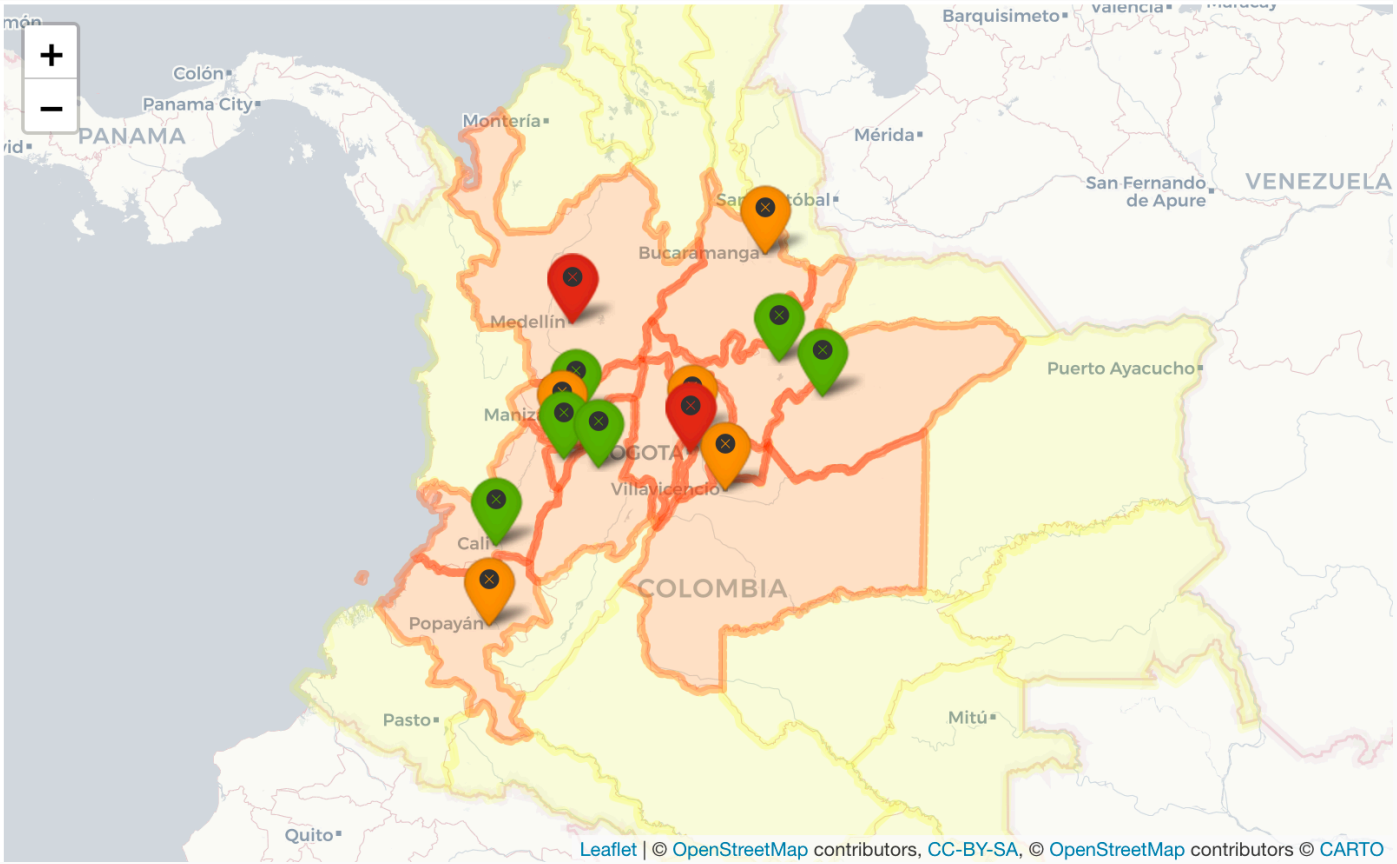
$\Downarrow$

there is the capital Bogotà and big cities like Medellín (Antioquia).

We decided to consider as **central Colombia** the following departments/districts: Bogotà DC, Boyacá, Tolima, Cundinamarca, Meta, Quindío, Valle del Cauca, Risaralda, Celdas, Boyacá, Antioquia, Santander, Casanare.

Map

Here we can see our selected departments. The color of the pins is related with the number of cases: if they are less than 500 the color is green, if they are less than 5000 the color is orange, otherwise it is red.



The dataset

Description of variables

The variables of our dataset are:

- *ID de caso*: ID of the confirmed case.
- *Fecha de diagnóstico*: Date in which the disease was diagnosed.
- *Ciudad de ubicación*: City where the case was diagnosed.
- *Departamento o Distrito*: Department or district where the city belongs to.
- *Atención*: Situation of the patient: recovered, at home, at the hospital, at the ICU or deceased.
- *Edad*: Age of the confirmed case.
- *Sexo*: Sex of the confirmed case.
- *Tipo*: How the person got infected: in Colombia, abroad or unknown.
- *País de procedencia*: Country of origin if the person got infected abroad.

Preprocessing

We expanded the dataset → kaggle link: “covid19co_official.csv”

We had to clean the dataset:

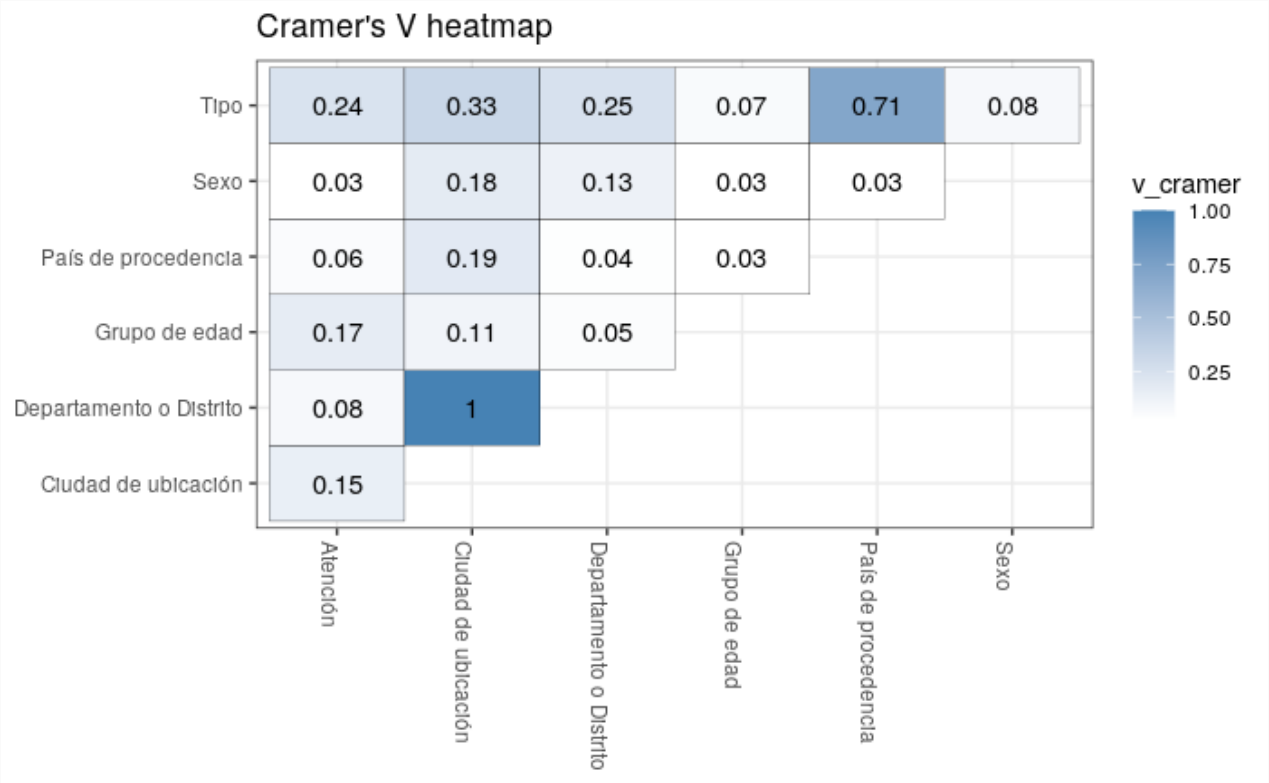
- We transformed the *Fecha de diagnóstico* variable into a *Date* type variable,
- we fixed the variable *Id de caso* (we removed some departments = lines),
- we created a variable *Grupo de edad*,
- we cleaned the column *País de procedencia* (replaced cities with the country) and created the variable *Continente de procedencia* (the first is too fragmented).

Plus we created dummy variables for all the categorical variables.

The dataset

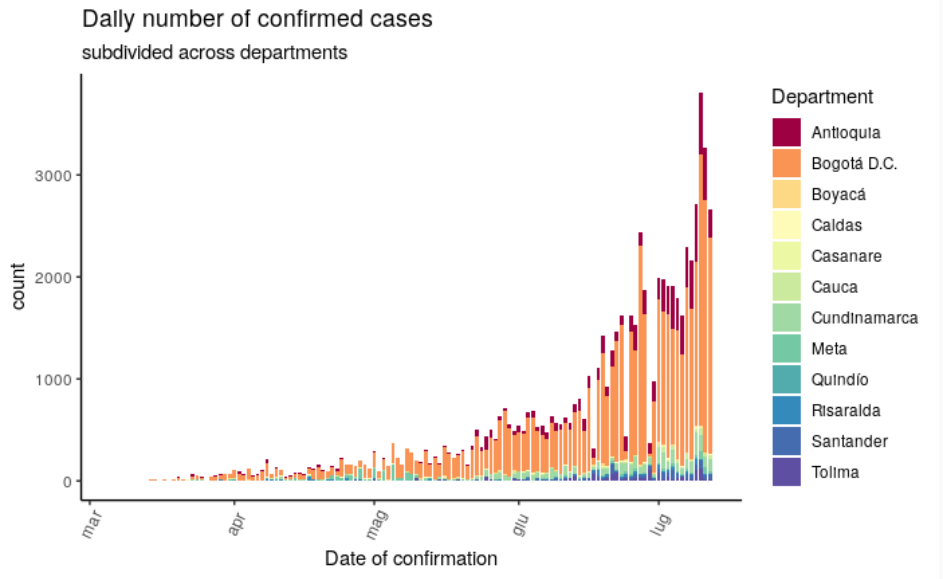
The data has 126 observations and we split it so to leave out the last seven points for prediction, because in this scenario it is sensible to predict only a week as the situation changes really fast.

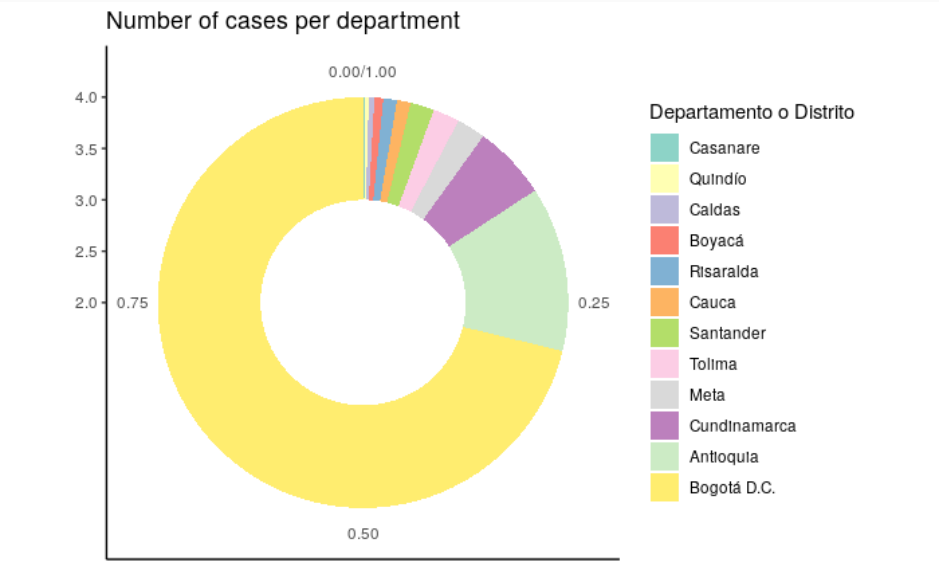
We used Cramer's V to compute the correlation between our categorical variables.

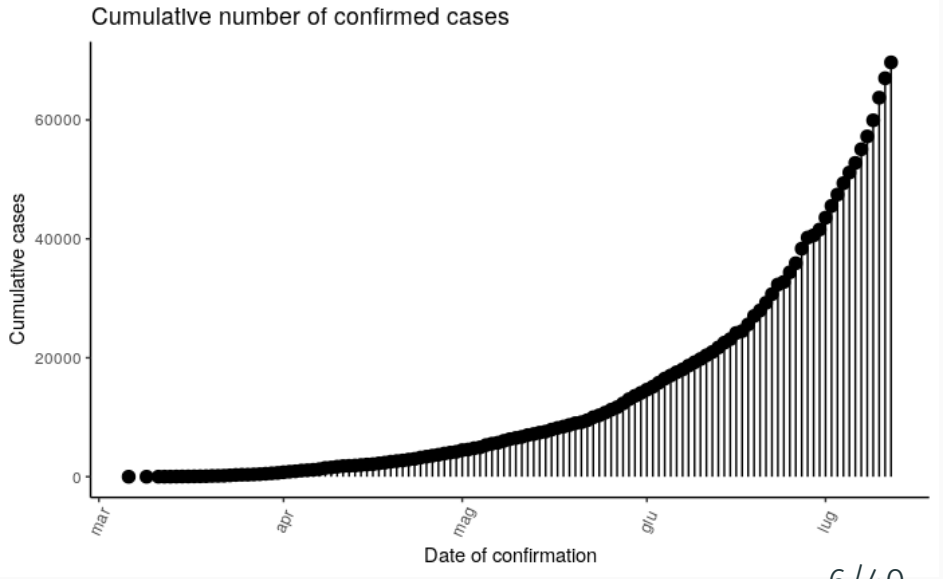


Exploring the dataset

The plots on the right represent the daily and the cumulative incidence of the disease across all the departments we are taking into account.
On the left, summing the total number of cases during the period analyzed shows that the most affected department is the capital Bogotà.





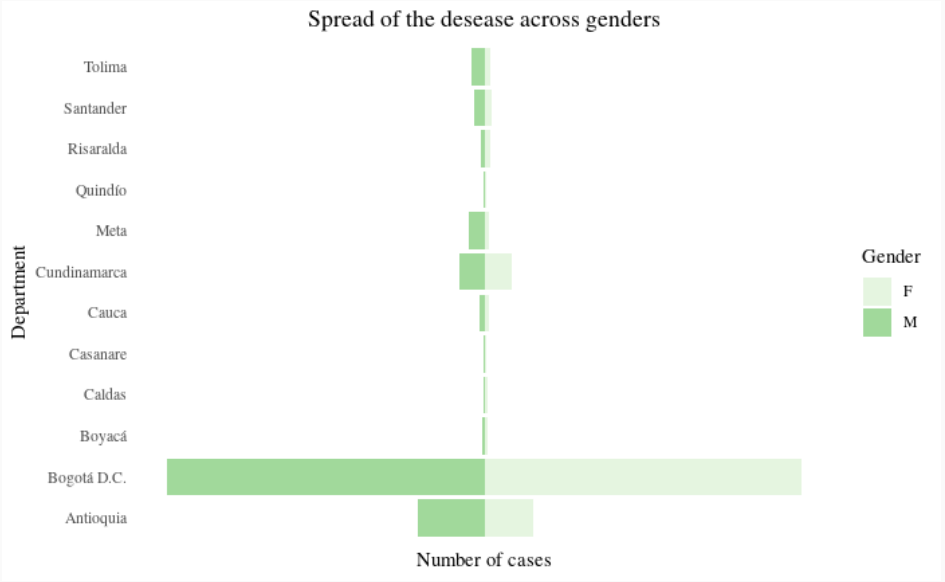


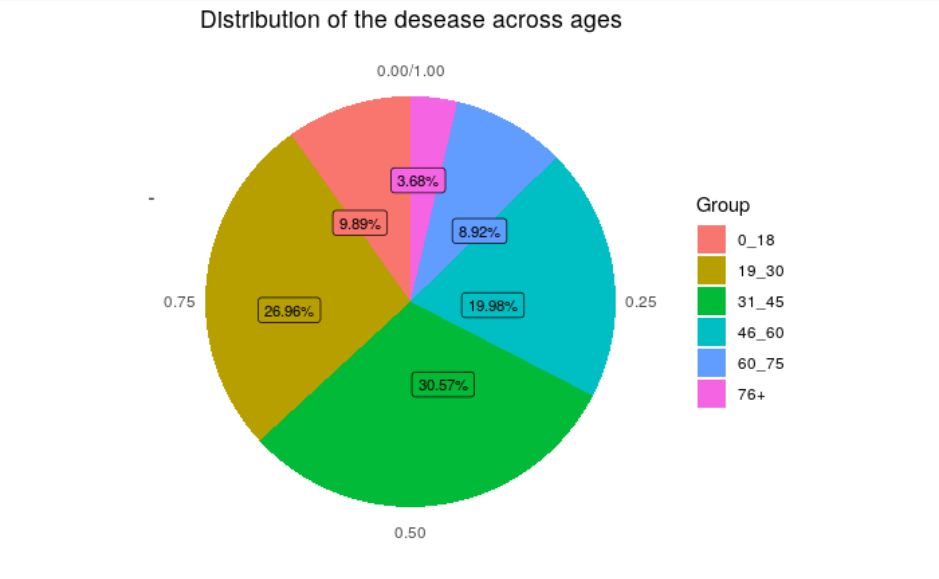
Exploring the variables

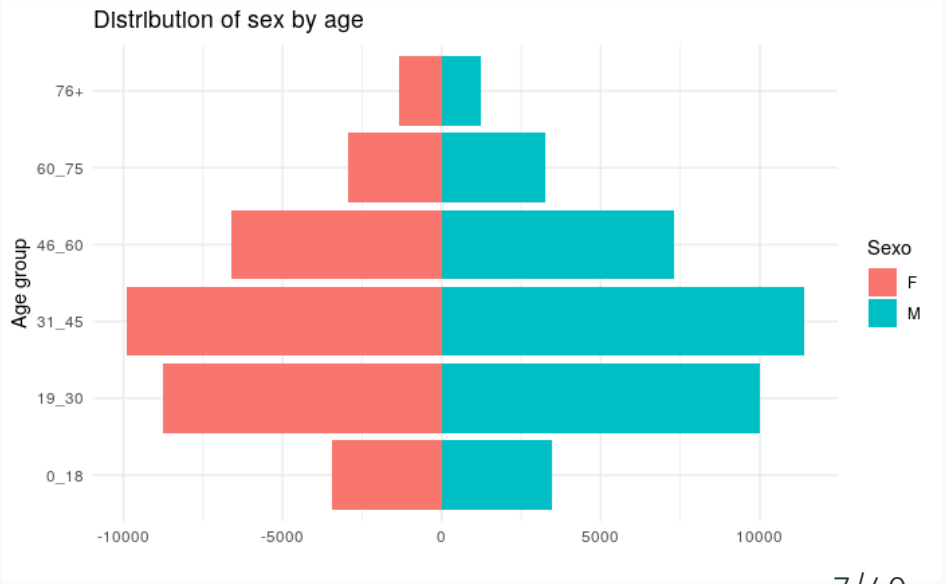
The disease (number of cases) is more or less equally distributed across genders.

People from 31 to 45 years old are the most affected by the disease and people over 76 years old are the least affected.

There isn't much difference between the sexes among the different group of ages.







The frequentist approach

GLM Methods

The Generalized Linear Models (GLM) extends the linear framework allowing for the following transformation involving a twice differentiable link function g such that:

$$g(E[y_i]) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}.$$

We decided to use the GLM methods, with Poisson, Quasi-Poisson and Negative Binomial families, because they are the most used model in case of counting.

The Poisson model

The Poisson model consists in the following:

$$y_i \sim \mathcal{P}oisson(\lambda),$$

with

$$\ln(\lambda) = \ln(E[y_i]) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i.$$

The GLM Poisson model is easily programmed in R with the function *`glm(formula, family = poisson)`*.

Poisson models summary

Poisson models for cumulative cases	AIC	RMSE
Cumulative cases ~ Elapsed time	21915.03	1760.51
Cumulative cases ~ Elapsed time + Elapsed time^2	12062.73	6940.37
Cumulative cases ~ Elapsed time + Sex	20887.67	3892.81
Cumulative cases ~ Elapsed time + Age group	20650.03	3249.32
Cumulative cases ~ Elapsed time + Department	19275.39	3328.87
Cumulative cases ~ Elapsed time + Age group + Department	18085.27	8243.7
Cumulative cases ~ Elapsed time + Elapsed time^2 + Age group + Department	8781.85	8441.28

Poisson models for daily cases	AIC	RMSE
New cases/day ~ Elapsed time	7291.77	753.97
New cases/day ~ Elapsed time + Elapsed time^2	7291.99	739.36
New cases/day ~ Elapsed time + Sex	4079.28	1810.42
New cases/day ~ Elapsed time + Age group	3845.15	1948.8
New cases/day ~ Elapsed time + Department	2989.47	1447.55
New cases/day ~ Elapsed time + Age group + Department	2860.39	1225.26
New cases/day ~ Elapsed time + Elapsed time^2 + Age group + Department	2225.47	3072.66

ANOVA

Anova table for the Poisson models for the Cumulative cases as response.

```
anova(poisson1, poisson4, poisson6, poisson7, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: `Cumulative cases` ~ `Elapsed time`
## Model 2: `Cumulative cases` ~ `Elapsed time` + `Grupo de edad_19_30` +
##     `Grupo de edad_31_45` + `Grupo de edad_46_60` + `Grupo de edad_60_75` +
##     `Grupo de edad_76+`
## Model 3: `Cumulative cases` ~ `Elapsed time` + `Grupo de edad_19_30` +
##     `Grupo de edad_31_45` + `Grupo de edad_46_60` + `Grupo de edad_60_75` +
##     `Grupo de edad_76+` + `Departamento o Distrito_Bogotá D.C.` +
##     `Departamento o Distrito_Boyacá` + `Departamento o Distrito_Caldas` +
##     `Departamento o Distrito_Casanare` + `Departamento o Distrito_Cauca` +
##     `Departamento o Distrito_Cundinamarca` + `Departamento o Distrito_Meta` +
##     `Departamento o Distrito_Quindío` + `Departamento o Distrito_Risaralda` +
##     `Departamento o Distrito_Santander` + `Departamento o Distrito_Tolima`
## Model 4: `Cumulative cases` ~ `Elapsed time` + I(`Elapsed time`^2) + `Grupo de edad_19_30` +
##     `Grupo de edad_31_45` + `Grupo de edad_46_60` + `Grupo de edad_60_75` +
##     `Grupo de edad_76+` + `Departamento o Distrito_Bogotá D.C.` +
##     `Departamento o Distrito_Boyacá` + `Departamento o Distrito_Caldas` +
##     `Departamento o Distrito_Casanare` + `Departamento o Distrito_Cauca` +
##     `Departamento o Distrito_Cundinamarca` + `Departamento o Distrito_Meta` +
##     `Departamento o Distrito_Quindío` + `Departamento o Distrito_Risaralda` +
##     `Departamento o Distrito_Santander` + `Departamento o Distrito_Tolima`
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       118    20716.4
## 2       113    19441.4  5   1275.0 < 2.2e-16 ***
## 3       102    16854.6 11   2586.8 < 2.2e-16 ***
## 4       101     7549.2  1   9305.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

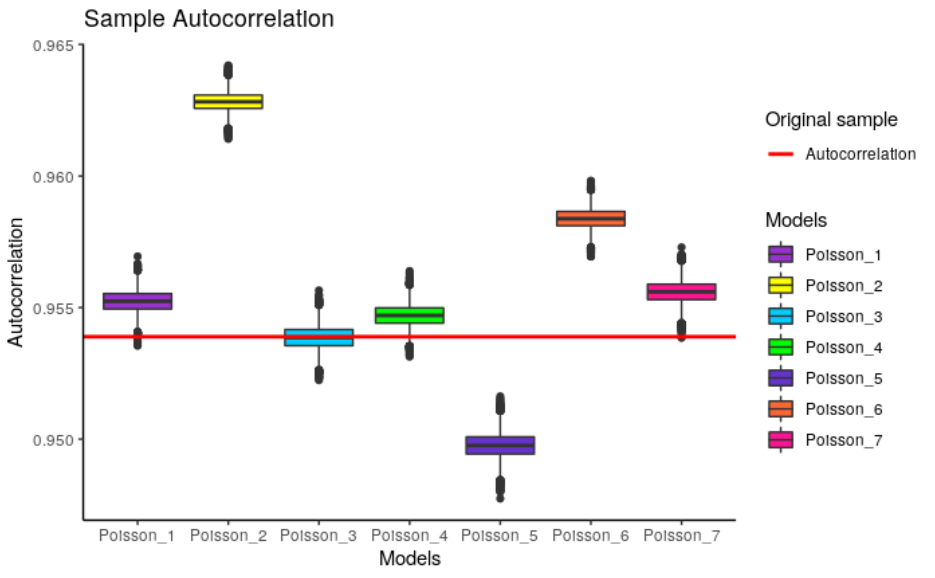
Anova table for the Poisson models for the New cases/day as response.

```
anova(poisson1bis, poisson2bis, poisson7bis, test="Chisq")
```

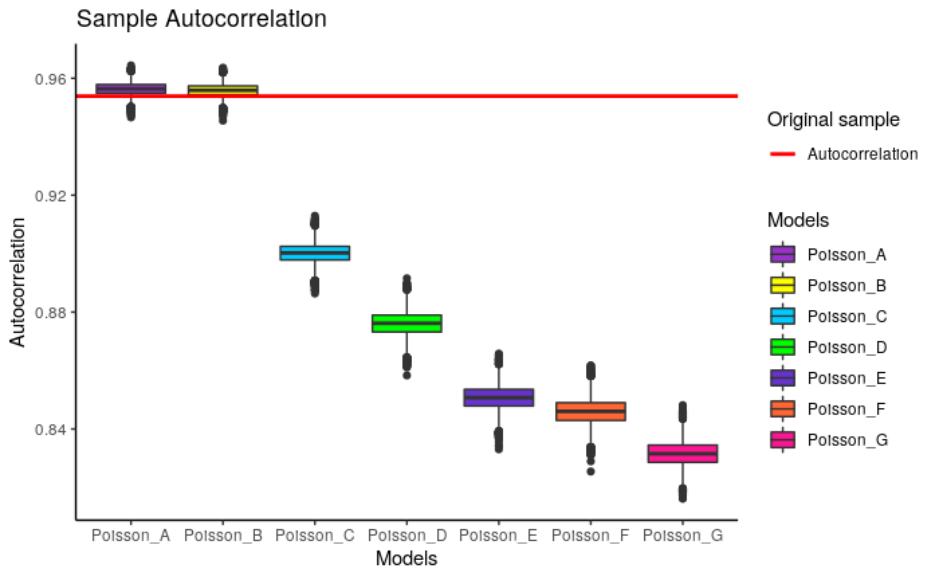
```
## Analysis of Deviance Table
##
## Model 1: `New cases/day` ~ `Elapsed time`
## Model 2: `New cases/day` ~ `Elapsed time` + I(`Elapsed time`^2)
## Model 3: `New cases/day` ~ `Elapsed time` + I(`Elapsed time`^2) + `Grupo de edad_19_30` +
##     `Grupo de edad_31_45` + `Grupo de edad_46_60` + `Grupo de edad_60_75` +
##     `Grupo de edad_76+` + `Departamento o Distrito_Bogotá D.C.` +
##     `Departamento o Distrito_Boyacá` + `Departamento o Distrito_Caldas` +
##     `Departamento o Distrito_Casanare` + `Departamento o Distrito_Cauca` +
##     `Departamento o Distrito_Cundinamarca` + `Departamento o Distrito_Meta` +
##     `Departamento o Distrito_Quindío` + `Departamento o Distrito_Risaralda` +
##     `Departamento o Distrito_Santander` + `Departamento o Distrito_Tolima`
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       118     6442.5
## 2       117     6440.7  1      1.8   0.1817
## 3       101     1342.2 16   5098.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Autocorrelation

Autocorrelation for the Poisson models for the Cumulative cases as response.



Autocorrelation for the Poisson models for the New cases/day as response.



Best Poisson model 1

Best Poisson model for the cumulative number of cases, considering the prediction interval, the RMSE, the AIC and the Occam's razor principle.

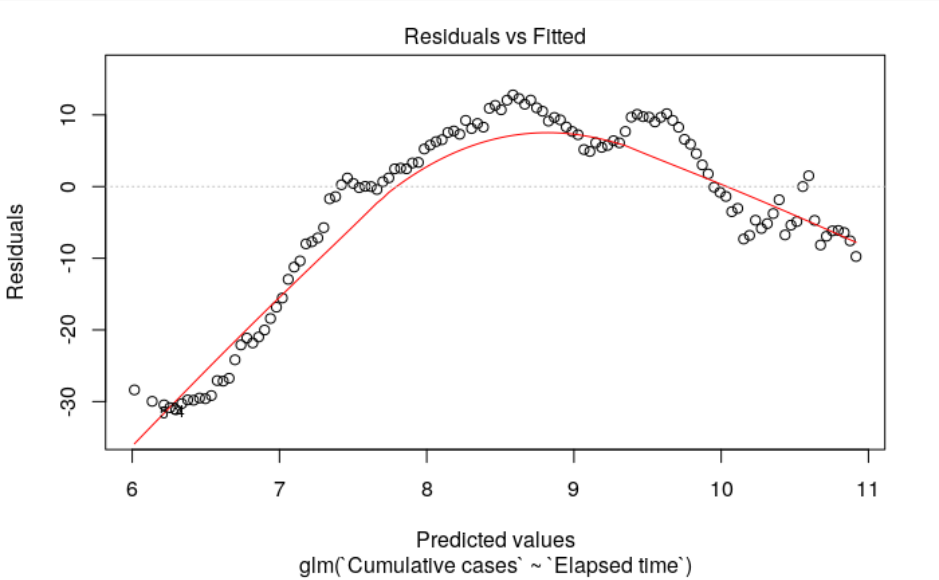
The model

```
poisson1 <- glm(`Cumulative cases` ~ `Elapsed time`,
                data=data1[1:120, ], family=poisson)
```

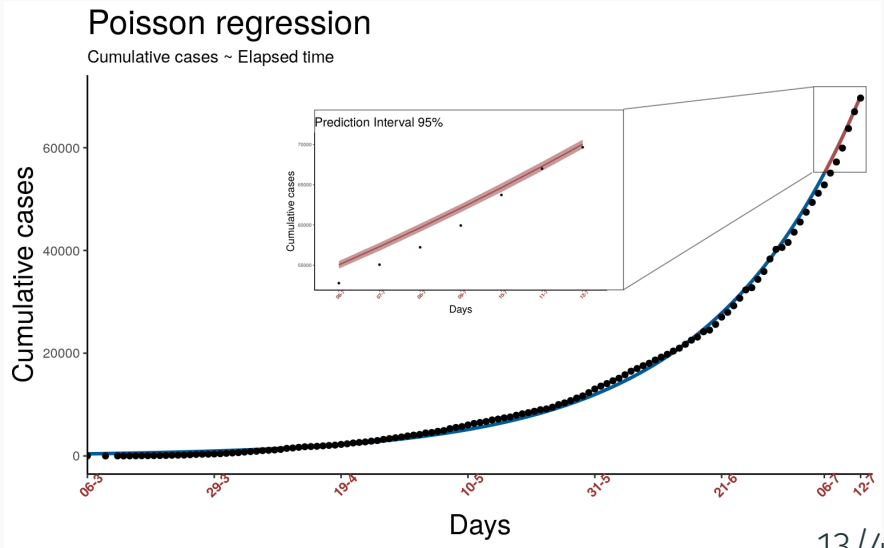
Performance indexes

```
## [1] "Estimated overdispersion 126.060786858437"
## [1] "RMSE: 1760.50922940986"
## [1] "AIC: 21915.0291925917"
## [1] "Null deviance:  1740689.09"  "Residual deviance: 20716.39"
```

Residuals plot



Prediction interval



Best Poisson model 2

Best Poisson model for the daily number of cases, considering the prediction interval, the RMSE, the AIC and the Occam's razor principle.

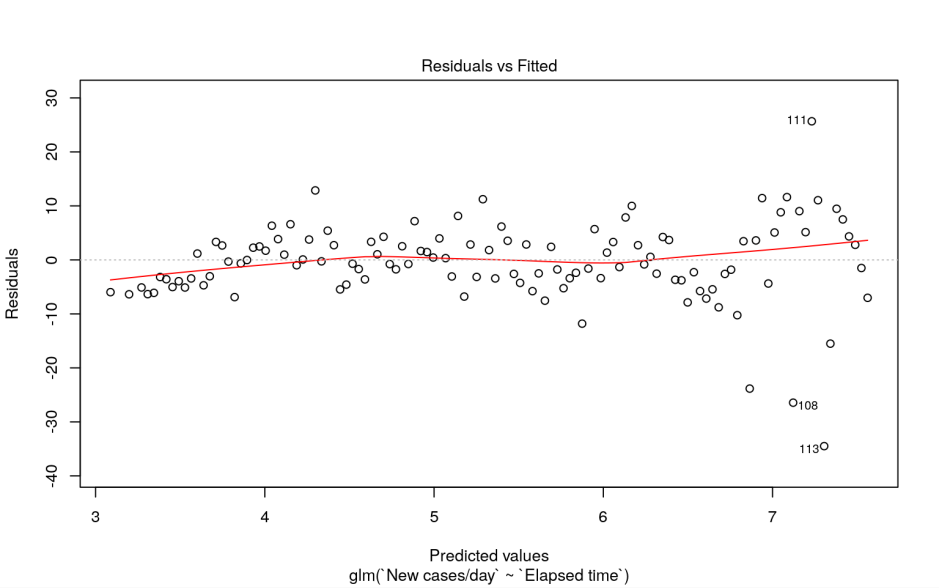
The model

```
poisson1bis <- glm(`New cases/day` ~ `Elapsed time`,
                   data=data1[1:120, ], family=poisson)
```

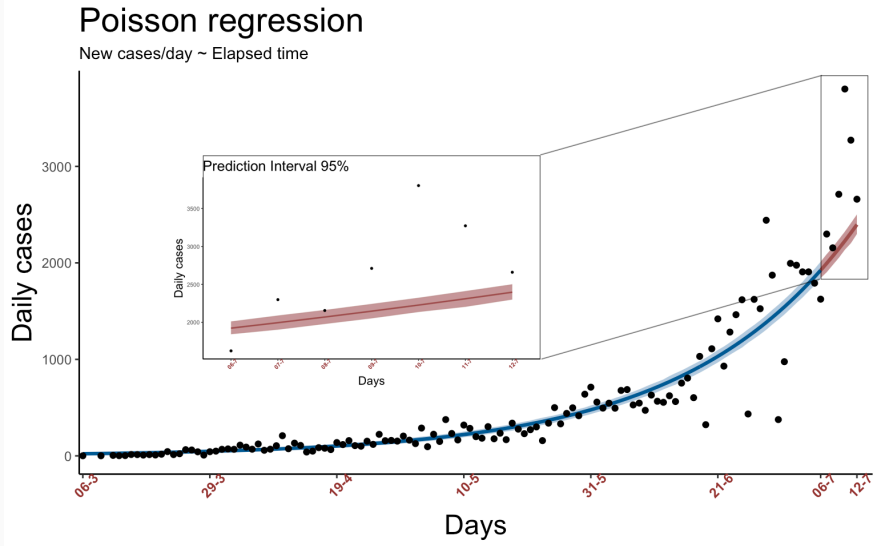
Performance indexes

```
## [1] "Estimated overdispersion 50.354334425333"
## [1] "RMSE: 753.966239397641"
## [1] "AIC: 7291.77457395168"
## [1] "Null deviance:  64369.08"   "Residual deviance: 6442.53"
```

Residuals plot



Prediction interval



The Quasi-Poisson model

Poisson models have a main drawback: the mean and the variance both equal λ. However, in many cases this is a major restriction, since data show a non negligible extent of overdispersion: data exhibit a greater variance than that assumed by the model. The idea is then to specify a Quasi-Poisson model, with:

$$\begin{cases} E[y_i] = \lambda \\ \text{var}(y_i) = \phi E[y_i] \end{cases}$$

where ϕ is the overdispersion parameter.

Quasi-Poisson models summary

Quasi Poisson models for cumulative cases	RMSE
Cumulative cases ~ Elapsed time	1760.51
Cumulative cases ~ Elapsed time + Elapsed time^2	6940.37
Cumulative cases ~ Elapsed time + Sex	3892.81
Cumulative cases ~ Elapsed time + Age group	3249.32
Cumulative cases ~ Elapsed time + Department	3328.87
Cumulative cases ~ Elapsed time + Department + Age group	8243.7
Cumulative cases ~ Elapsed time + Elapsed time^2 + Department + Age group	8441.28

Quasi Poisson models for daily cases	RMSE
New cases/day ~ Elapsed time	753.97
New cases/day ~ Elapsed time + Elapsed time^2	739.36
New cases/day ~ Elapsed time + Sex	1810.42
New cases/day ~ Elapsed time + Age group	1948.8
New cases/day ~ Elapsed time + Department	1447.55
New cases/day ~ Elapsed time + Department + Age group	1225.26
New cases/day ~ Elapsed time + Elapsed time^2 + Department + Age group	3072.66

The Negative Binomial model

The Negative Binomial model consists in the following:

$$y_i \sim \mathcal{N}egative\ \mathcal{B}inomial(\lambda, \phi),$$

$$\ln(\lambda) = \ln(E[y_i]) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i.$$

where the parameter ϕ is called *precision* and it is such that:

$$E[y_i] = \lambda, \qquad \text{Var}[y_i] = \lambda + \frac{\lambda^2}{\phi}$$

The GLM Negative Binomial model is easily programmed in R with the function `glm.nb(formula)`.

Negative Binomial models summary

Negative Binomial models for cumulative cases	AIC	RMSE
Cumulative cases ~ Elapsed time	21911.18	1765.66
Cumulative cases ~ Elapsed time + Elapsed time^2	12059.27	6946.45
Cumulative cases ~ Elapsed time + Sex	20882.94	3898.5
Cumulative cases ~ Elapsed time + Age group	20645.27	3253.45
Cumulative cases ~ Elapsed time + Department	19270.25	3335.37
Cumulative cases ~ Elapsed time + Age group + Department	18080.49	8253.9
Cumulative cases ~ Elapsed time + Elapsed time^2 + Age group + Department	8780.25	8445.61

Negative Binomial models for daily cases	AIC	RMSE
New cases/day ~ Elapsed time	1445.58	630.34
New cases/day ~ Elapsed time + Elapsed time^2	1435.03	1196.18
New cases/day ~ Elapsed time + Sex	1441.88	1182.57
New cases/day ~ Elapsed time + Age group	1444.38	1943.63
New cases/day ~ Elapsed time + Department	1431.44	2185.73
New cases/day ~ Elapsed time + Age group + Department	1435.14	1379.57
New cases/day ~ Elapsed time + Elapsed time^2 + Age group + Department	1376.64	9304.27

ANOVA

Anova table for the NB models for the Cumulative cases as response.

```
anova(nb1, nb4, nb6, nb7, test="Chisq")
```

```
## Likelihood ratio tests of Negative Binomial Models
##
## Response: Cumulative cases
##
## 1
## 2
## 3                      `Elapsed time` + `Grupo de edad_19_30` + `Grupo de edad_31_45` + `Grupo de
## 4 `Elapsed time` + I(`Elapsed time`^2) + `Grupo de edad_19_30` + `Grupo de edad_31_45` + `Grupo de
##       theta Resid. df    2 x log-lik.   Test    df LR stat. Pr(Chi)
## 1 11252779       118      -21905.177
## 2 12919544       113      -20629.270 1 vs 2     5 1275.907       0
## 3 13728065       102      -18042.489 2 vs 3    11 2586.780       0
## 4 13798920       101       -8740.254 3 vs 4     1 9302.235       0
```

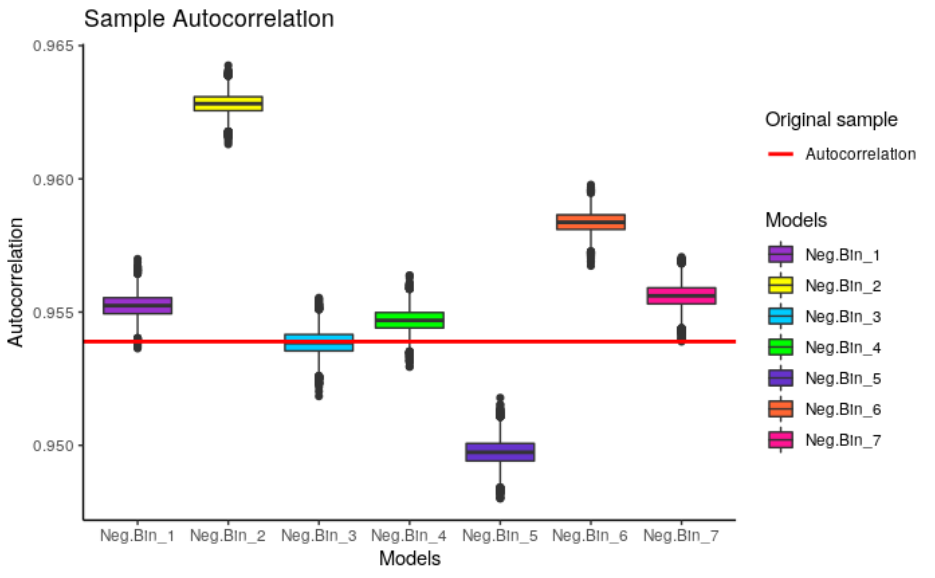
Anova table for the NB models for the New cases/day as response.

```
anova(nb1bis, nb4bis, nb6bis, nb7bis, test="Chisq")
```

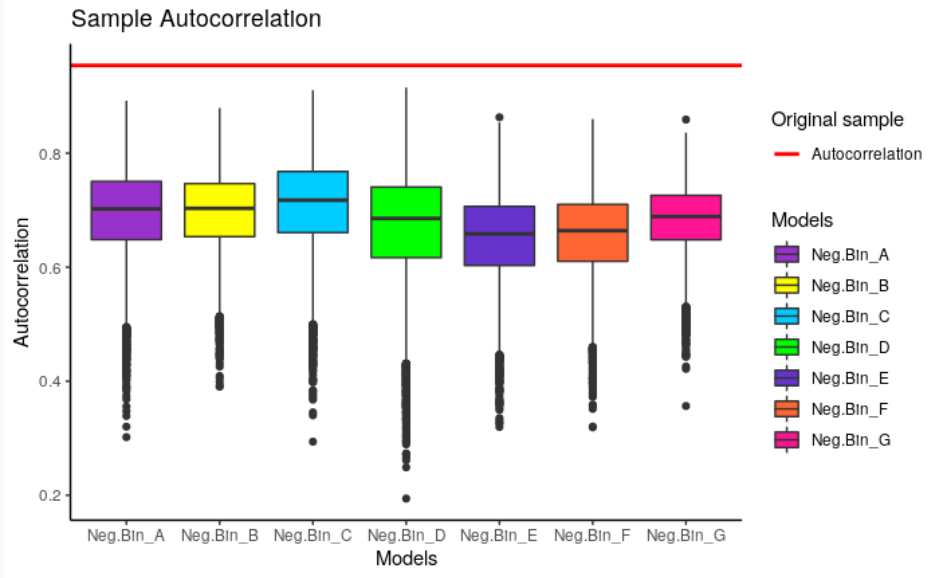
```
## Likelihood ratio tests of Negative Binomial Models
##
## Response: New cases/day
##
## 1
## 2
## 3                      `Elapsed time` + `Grupo de edad_19_30` + `Grupo de edad_31_45` + `Grupo de
## 4 `Elapsed time` + I(`Elapsed time`^2) + `Grupo de edad_19_30` + `Grupo de edad_31_45` + `Grupo de
##       theta Resid. df    2 x log-lik.   Test    df LR stat.      Pr(Chi)
## 1  4.571965       118       -1439.581
## 2  5.156688       113       -1428.377 1 vs 2     5 11.20369 4.748776e-02
## 3  7.241381       102       -1397.144 2 vs 3    11 31.23285 1.011534e-03
## 4 12.990325       101       -1336.637 3 vs 4     1 60.50701 7.327472e-15
```

Autocorrelation

Autocorrelation for the NB models for the Cumulative cases as response.



Autocorrelation for the NB models for the New cases/day as response.



Best NB model 1

Best NB model for the cumulative number of cases, considering the prediction interval, the RMSE, the AIC and the Occam's razor principle.

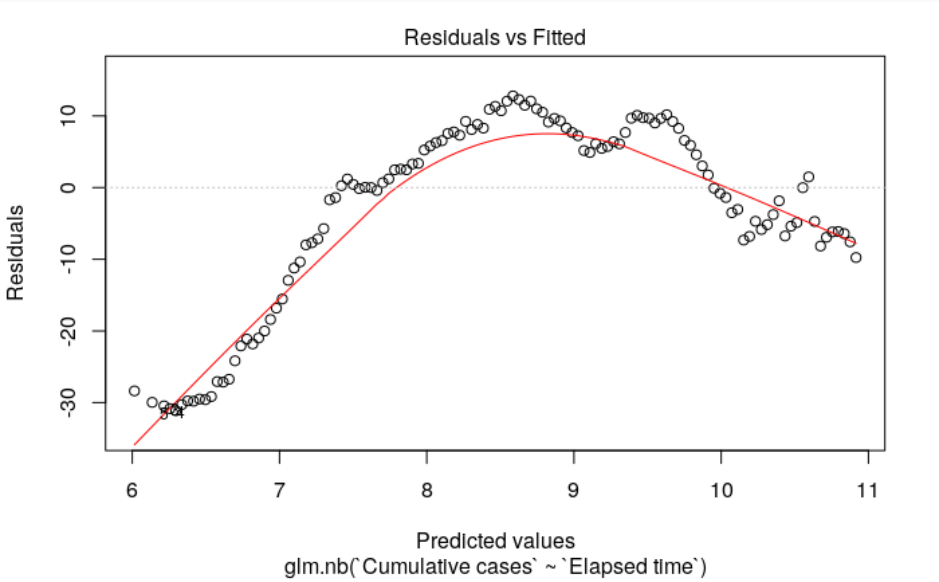
The model

```
nb1 <- glm.nb(`Cumulative cases` ~ `Elapsed time`,
              data=data1[1:120, ])
```

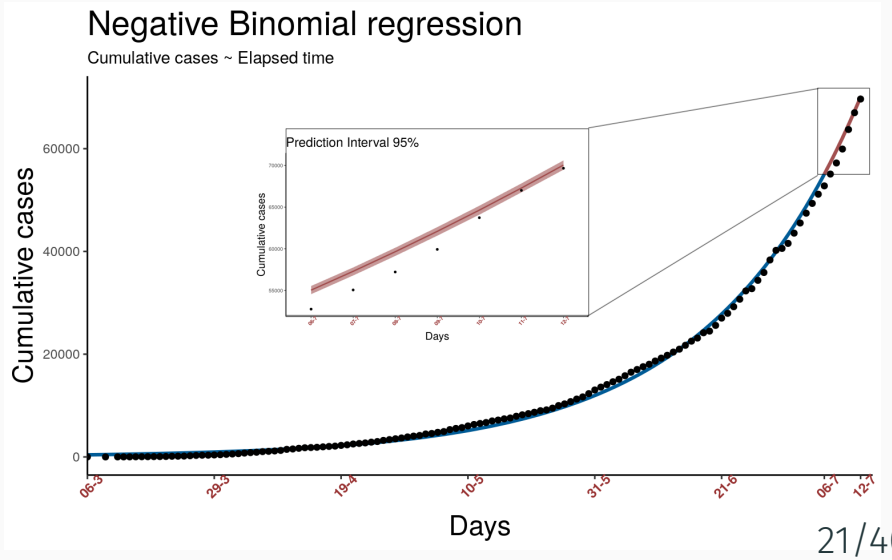
Performance indexes

```
## [1] "Estimated overdispersion 177.301055416919"
## [1] "RMSE: 1765.66109000166"
## [1] "AIC: 21911.1770461798"
## [1] "Null deviance:  1738722.15"  "Residual deviance: 20710.41"
```

Residuals plot



Prediction interval



Best NB model 2

Best NB model for the daily number of cases, considering the prediction interval, the RMSE, the AIC and the Occam's razor principle.

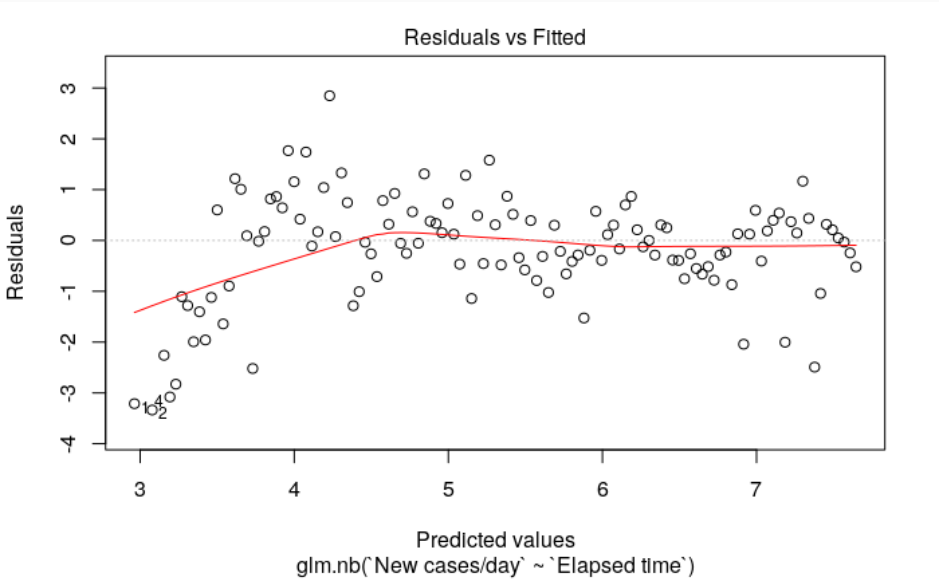
The model

```
nb1bis <- glm.nb(`New cases/day` ~ `Elapsed time`,
                 data=data1[1:120, ])
```

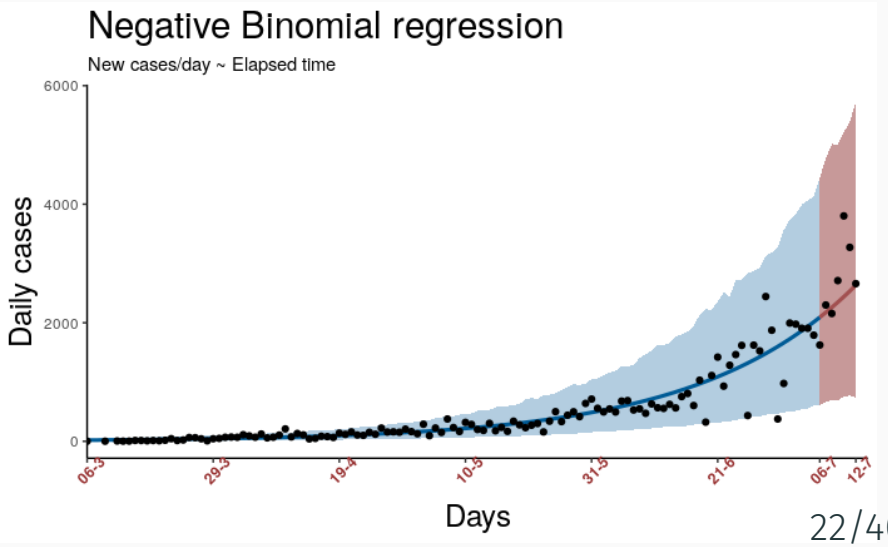
Performance indexes

```
## [1] "Estimated overdispersion 1.18509200751555"
## [1] "RMSE: 630.336144980397"
## [1] "AIC: 1445.58057844612"
## [1] "Null deviance:  916.37"    "Residual deviance: 136.75"
```

Residuals plot



Prediction interval



GAM Methods

A Generalized Additive Model (GAM) is a generalized linear model (GLM) in which the linear predictor is given by a user specified sum of smooth functions of the covariates plus a conventional parametric component of the linear predictor. Given $\mu_i = E[Y_i]$, we used the model:

$$\log(\mu_i) = \beta_0 + \sum_{j=1}^{p} s_j(x_{ij}) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the dependent variable $Y_i \sim \mathcal{P}oisson(\mu_i)$ and s_j are smooth functions of covariates x_j.

GAM models summary

GAM models for cumulative cases	AIC	RMSE
Cumulative cases ~ s(Elapsed time)	1587.12	1282.66
Cumulative cases ~ s(Elapsed time) + s(Elapsed time) ^2)	1517.88	6047.01
Cumulative cases ~ s(Elapsed time) + s(Sex)	1525.13	2301.65
Cumulative cases ~ s(Elapsed time) + Age group	1543.48	911.82
Cumulative cases ~ s(Elapsed time) + Department	1517.1	704.14
Cumulative cases ~ s(Elapsed time) + Age group + Department	1486.71	1038.67
Cumulative cases ~ s(Elapsed time) + s(Elapsed time^2) + Age group + Department	1337.17	4599.76

GAM models for daily cases	AIC	RMSE
New cases/day ~ s(Elapsed time)	6361.03	1121.42
New cases/day ~ s(Elapsed time) + s(Elapsed time^2)	6341.57	756.74
New cases/day ~ s(Elapsed time) + s(Sex)	1239.36	1301.35
New cases/day ~ s(Elapsed time) + s(Sex) + Age group	1094.53	412.27
New cases/day ~ s(Elapsed time) + s(Sex) + Department	1095.72	398.09
New cases/day ~ s(Elapsed time) + s(Sex) + Age group + Department	1078.03	868.4
New cases/day ~ s(Elapsed time) + s(Elapsed time^2) + s(Sex) + Age group + Department	1077.19	1237.69

ANOVA

Anova table for the GAM models for the Cumulative cases as response.

```
anova(gam1, gam5, gam6, gam7, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Cumulative_cases ~ s(Elapsed_time)
## Model 2: Cumulative_cases ~ s(Elapsed_time) + Departamento_o_Distrito_Bogotá_D.C. +
##     Departamento_o_Distrito_Boyacá + Departamento_o_Distrito_Caldas +
##     Departamento_o_Distrito_Casanare + Departamento_o_Distrito_Cauca +
##     Departamento_o_Distrito_Cundinamarca + Departamento_o_Distrito_Meta +
##     Departamento_o_Distrito_Quindío + Departamento_o_Distrito_Risaralda +
##     Departamento_o_Distrito_Santander + Departamento_o_Distrito_Tolima
## Model 3: Cumulative_cases ~ s(Elapsed_time) + Grupo_de_edad_19_30 + Grupo_de_edad_31_45 +
##     Grupo_de_edad_46_60 + Grupo_de_edad_60_75 + Grupo_de_edad_76 +
##     Departamento_o_Distrito_Bogotá_D.C. + Departamento_o_Distrito_Boyacá +
##     Departamento_o_Distrito_Caldas + Departamento_o_Distrito_Casanare +
##     Departamento_o_Distrito_Cauca + Departamento_o_Distrito_Cundinamarca +
##     Departamento_o_Distrito_Meta + Departamento_o_Distrito_Quindío +
##     Departamento_o_Distrito_Risaralda + Departamento_o_Distrito_Santander +
##     Departamento_o_Distrito_Tolima
## Model 4: Cumulative_cases ~ s(Elapsed_time) + s(I(Elapsed_time^2)) + Grupo_de_edad_19_30 +
##     Grupo_de_edad_31_45 + Grupo_de_edad_46_60 + Grupo_de_edad_60_75 +
##     Grupo_de_edad_76 + Departamento_o_Distrito_Bogotá_D.C. +
##     Departamento_o_Distrito_Boyacá + Departamento_o_Distrito_Caldas +
##     Departamento_o_Distrito_Casanare + Departamento_o_Distrito_Cauca +
##     Departamento_o_Distrito_Cundinamarca + Departamento_o_Distrito_Meta +
##     Departamento_o_Distrito_Quindío + Departamento_o_Distrito_Risaralda +
##     Departamento_o_Distrito_Santander + Departamento_o_Distrito_Tolima
##   Resid. Df Resid. Dev      Df Deviance  Pr(>Chi)
## 1   110.001     372.54
## 2    99.001     280.52 11.0002   92.022 6.690e-15 ***
## 3    94.001     240.12  5.0001   40.402 1.239e-07 ***
## 4    86.005      74.68  7.9960  165.435 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova table for the GAM models for the New cases/day as response.

```
anova(gam1bis,gam4bis, gam6bis, gam7bis, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: New_cases_day ~ s(Elapsed_time)
## Model 2: New_cases_day ~ s(Elapsed_time) + s(Sexo_M) + Grupo_de_edad_19_30 +
##     Grupo_de_edad_31_45 + Grupo_de_edad_46_60 + Grupo_de_edad_60_75 +
##     Grupo_de_edad_76
## Model 3: New_cases_day ~ s(Elapsed_time) + s(Sexo_M) + Grupo_de_edad_19_30 +
##     Grupo_de_edad_31_45 + Grupo_de_edad_46_60 + Grupo_de_edad_60_75 +
##     Grupo_de_edad_76 + Departamento_o_Distrito_Bogotá_D.C. +
##     Departamento_o_Distrito_Boyacá + Departamento_o_Distrito_Caldas +
##     Departamento_o_Distrito_Casanare + Departamento_o_Distrito_Cauca +
##     Departamento_o_Distrito_Cundinamarca + Departamento_o_Distrito_Meta +
##     Departamento_o_Distrito_Quindío + Departamento_o_Distrito_Risaralda +
##     Departamento_o_Distrito_Santander + Departamento_o_Distrito_Tolima
## Model 4: New_cases_day ~ s(Elapsed_time) + s(I(Elapsed_time^2)) + s(Sexo_M) +
##     Grupo_de_edad_19_30 + Grupo_de_edad_31_45 + Grupo_de_edad_46_60 +
##     Grupo_de_edad_60_75 + Grupo_de_edad_76 + Departamento_o_Distrito_Bogotá_D.C. +
##     Departamento_o_Distrito_Boyacá + Departamento_o_Distrito_Caldas +
##     Departamento_o_Distrito_Casanare + Departamento_o_Distrito_Cauca +
##     Departamento_o_Distrito_Cundinamarca + Departamento_o_Distrito_Meta +
##     Departamento_o_Distrito_Quindío + Departamento_o_Distrito_Risaralda +
##     Departamento_o_Distrito_Santander + Departamento_o_Distrito_Tolima
##   Resid. Df Resid. Dev      Df Deviance  Pr(>Chi)
## 1   110.004     5496.0
## 2    96.007      201.6 13.9967   5294.4 < 2.2e-16 ***
## 3    85.007      163.1 11.0004     38.5 6.459e-05 ***
## 4    82.878      158.6  2.1287      4.5    0.1148
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Best GAM model 1

Best GAM model for the cumulative number of cases, considering the prediction interval, the RMSE, the AIC and the Occam's razor principle.

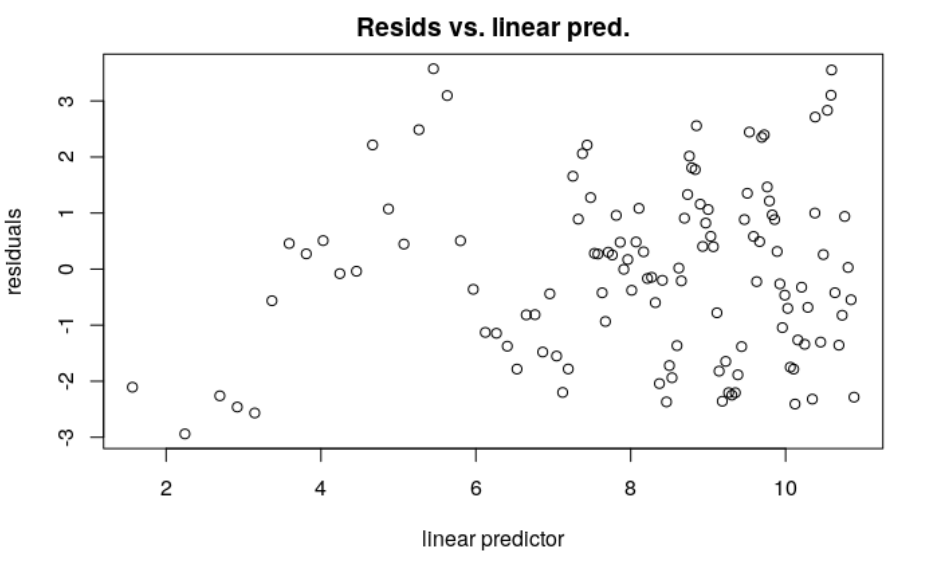
The model

```
gam5 <- gam(Cumulative_cases ~ s(Elapsed_time) + Departamento_o_Distrito_Bogotá_D.C. +
              Departamento_o_Distrito_Boyacá + Departamento_o_Distrito_Caldas +
              Departamento_o_Distrito_Casanare + Departamento_o_Distrito_Cauca +
              Departamento_o_Distrito_Cundinamarca + Departamento_o_Distrito_Meta +
              Departamento_o_Distrito_Quindío + Departamento_o_Distrito_Risaralda +
              Departamento_o_Distrito_Santander + Departamento_o_Distrito_Tolima,
            family = poisson(), data = df[1:120,])
```

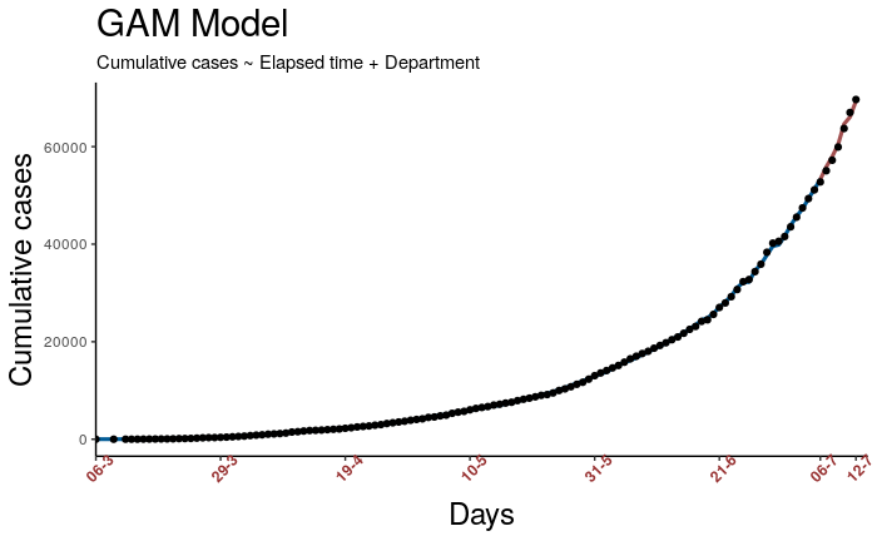
Performance indexes

```
## [1] "RMSE: 704.139095232087"
## [1] "AIC: 1517.10332972498"
```

Residuals plot



Prediction interval



Best GAM model 2

Best GAM model for the daily number of cases, considering the prediction interval, the RMSE, the AIC and the Occam's razor principle.

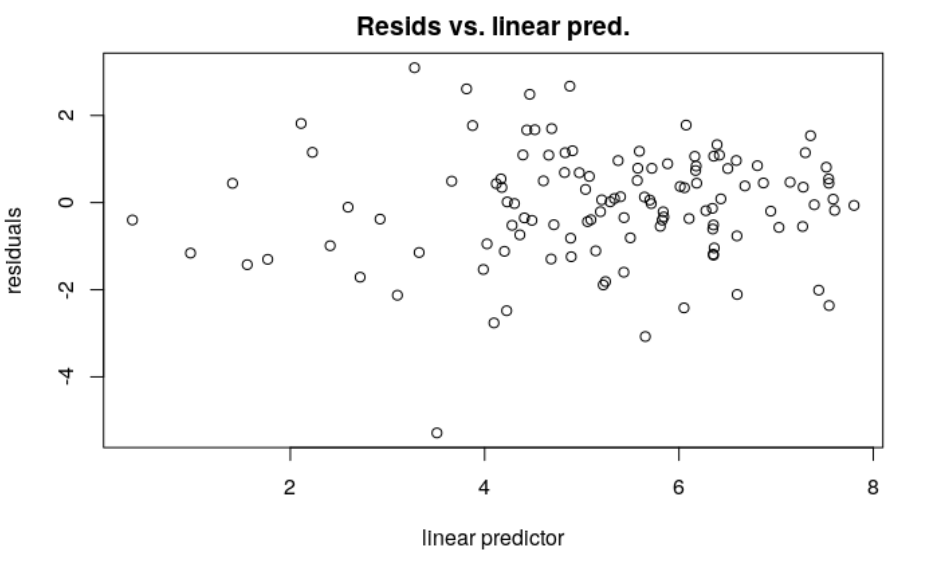
The model

```
gam5bis <- gam(New_cases_day~s(Elapsed_time) + s(Sexo_M) + Departamento_o_Distrito_Bogotá_D.C. +
                 Departamento_o_Distrito_Boyacá + Departamento_o_Distrito_Caldas +
                 Departamento_o_Distrito_Casanare + Departamento_o_Distrito_Cauca +
                 Departamento_o_Distrito_Cundinamarca + Departamento_o_Distrito_Meta +
                 Departamento_o_Distrito_Quindío + Departamento_o_Distrito_Risaralda +
                 Departamento_o_Distrito_Santander + Departamento_o_Distrito_Tolima,
               family = poisson(), data = df[1:120,])
```

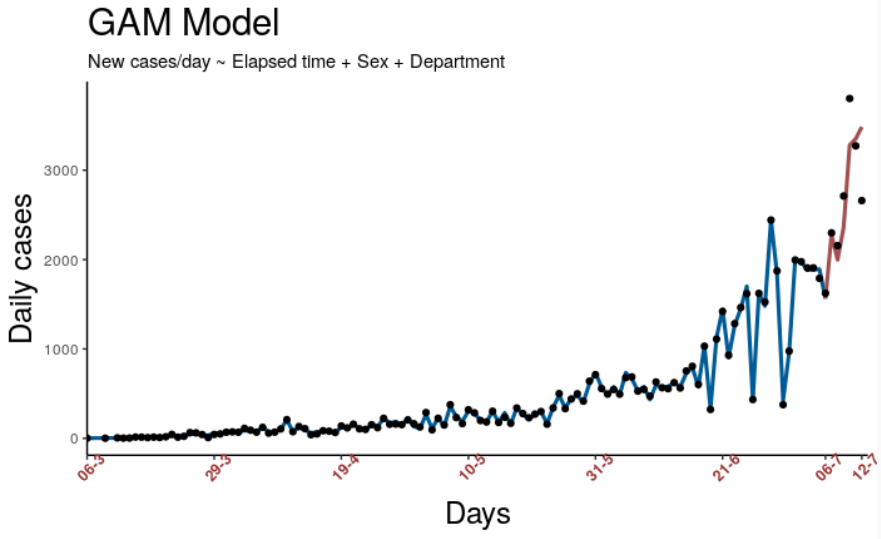
Performance indexes

```
## [1] "RMSE: 398.085987887759"
## [1] "AIC: 1095.71518740438"
```

Residuals plot



Prediction interval



The Bayesian approach

Poisson regression

We fit a simple Poisson regression:

$$\begin{aligned} \ln(\lambda_i) &= \alpha + \beta \cdot \texttt{Elapsed time}_i \\ y_i &\sim \mathcal{P}oisson(\lambda_i) \\ \alpha &\sim \mathcal{N}(0, 1) \\ \beta &\sim \mathcal{N}(0.25, 1) \end{aligned}$$

with $i = 1, \ldots, 1051$, being 1051 the number of rows of our dataset.

For what concerns the *stan* program, the functions we used are:

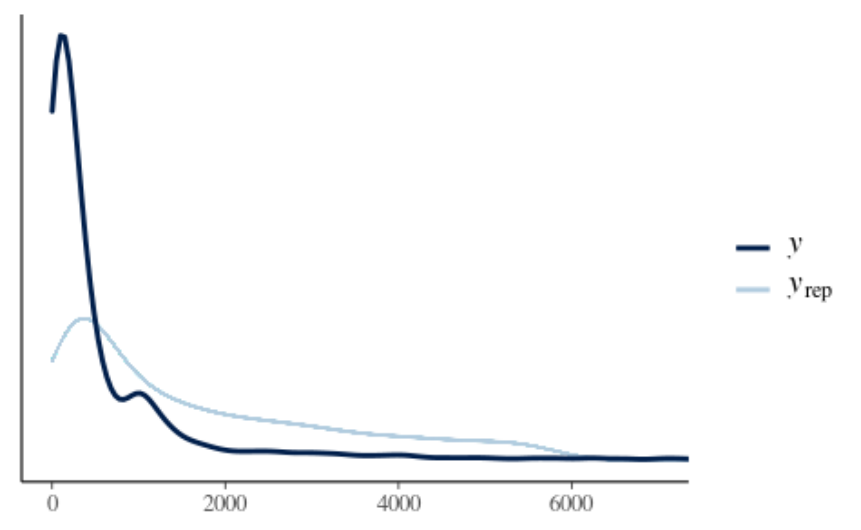
- *poisson_log_rng* to describe the distribution of y_i;
- *poisson_log_lpmf* to specify the likelihood.

Poisson model

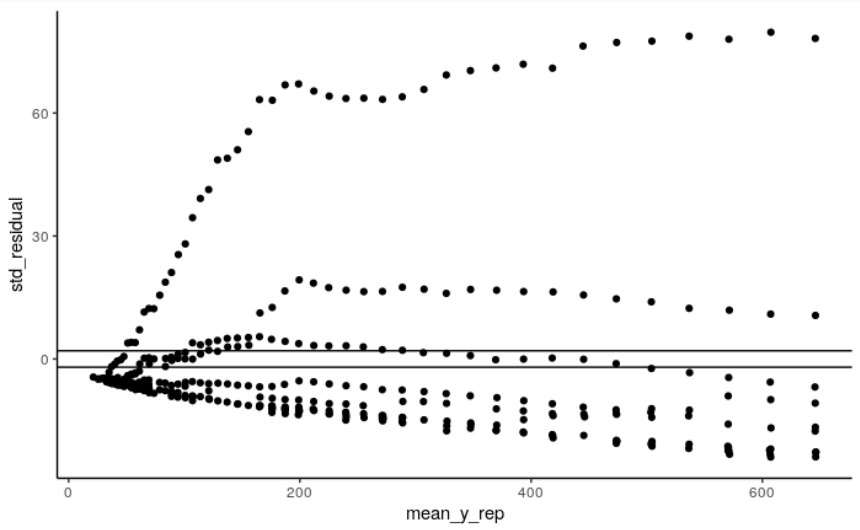
The fit is not satisfactory: the distribution of data is not coherent with replicated data.

The plot of the standardized residuals indicates a large amount of overdispersion.

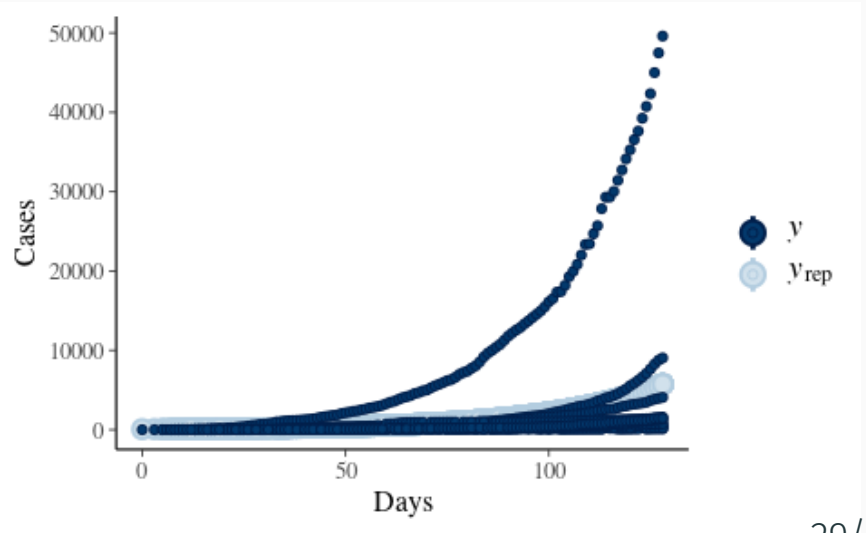
Posterior draws



Residuals Plot



Intervals' plot



Negative Binomial regression

We try to improve the previous model using the following Negative Binomial model:

$$\begin{aligned} \ln(\lambda_i) &= \alpha + \beta \cdot \texttt{Elapsed time}_i \\ y_i &\sim \mathcal{N}egative\ \mathcal{B}inomial(\lambda_i, \phi) \\ \alpha &\sim \mathcal{N}(0, 1) \\ \beta &\sim \mathcal{N}(0.25, 1) \end{aligned}$$

Where the parameter ϕ is called *precision* and it is such that:

$$E[y_i] = \lambda_i, \qquad \text{Var}[y_i] = \lambda_i + \frac{\lambda_i^2}{\phi}$$

with $i = 1, \ldots, 1051$.

For what concerns the `stan` program, the functions we used are:

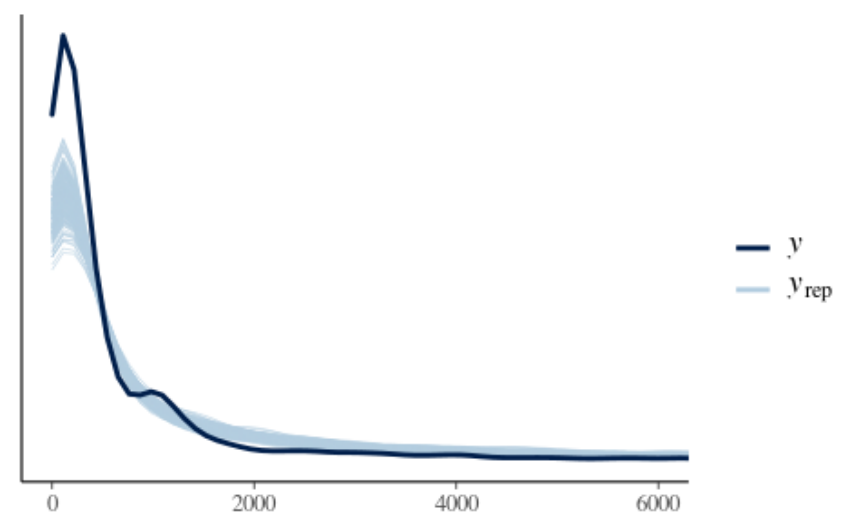
- `neg_binomial_2_log_rng` to specify the distribution of y_i;
- `neg_binomial_2_log_lpmf` for the likelihood.

Negative Binomial model

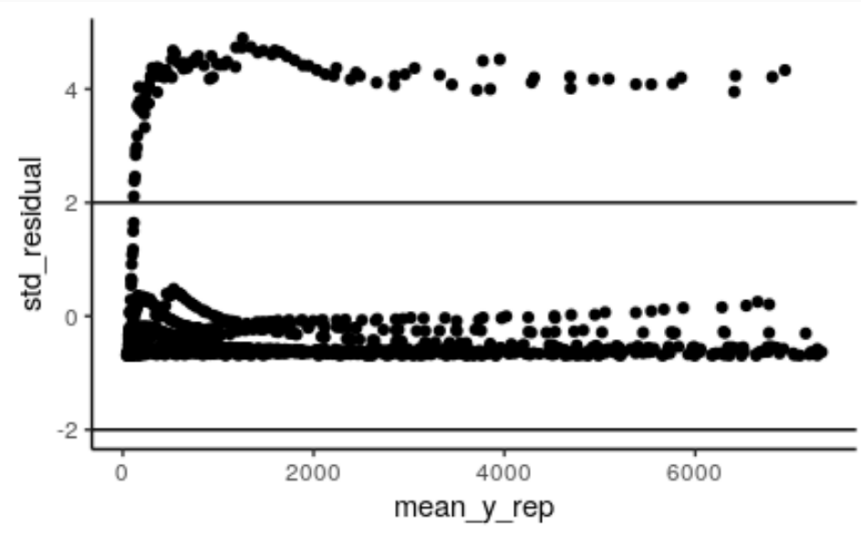
The fit has slightly improved, but overdispersion is still present.

The intervals' plot shows that the model is not able to capture the data.

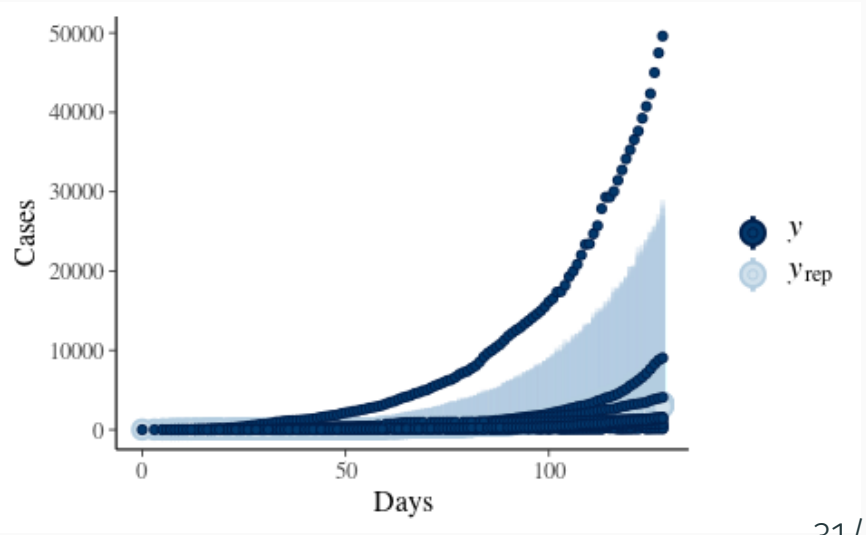
Posterior draws



Residuals Plot



Intervals' plot



Multilevel Negative Binomial regression

We try to fit the following Negative Binomial model, which also includes ***Mean Age*** as covariate:

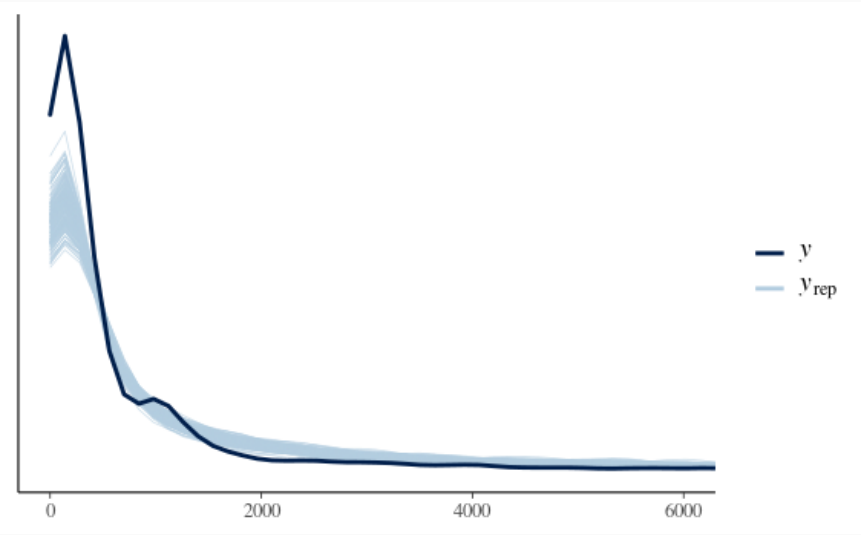
$$
\begin{aligned}
\ln(\lambda_i) &= \alpha + \beta_{time} \cdot \texttt{Elapsed time}_i + \beta_{age} \cdot \texttt{Mean Age} \\
y_i &\sim \mathcal{N}egative\ \mathcal{B}inomial(\lambda_i, \phi) \\
\alpha &\sim \mathcal{N}(0, 1) \\
\beta_{time} &\sim \mathcal{N}(0.5, 1) \\
\beta_{age} &\sim \mathcal{N}(0, 1)
\end{aligned}
$$

The ***stan*** functions we used are the same as before.

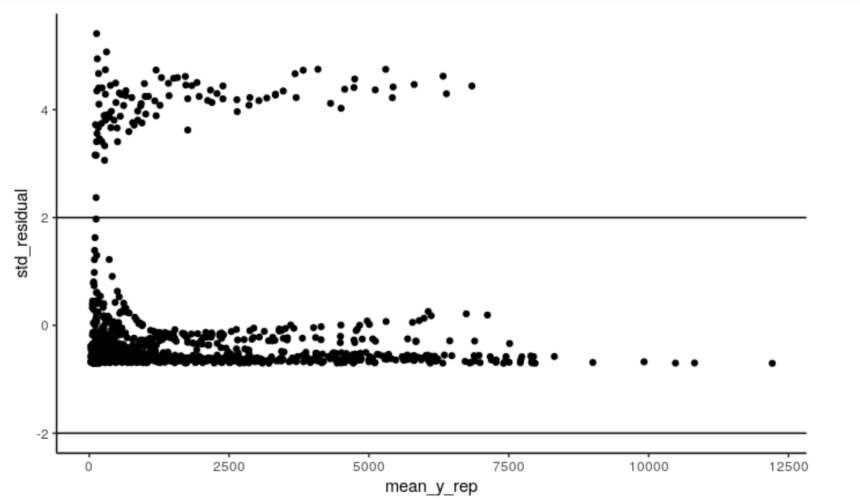
Multilevel Negative Binomial model

The model doesn't seem to improve much adding the *Mean Age* predictor.

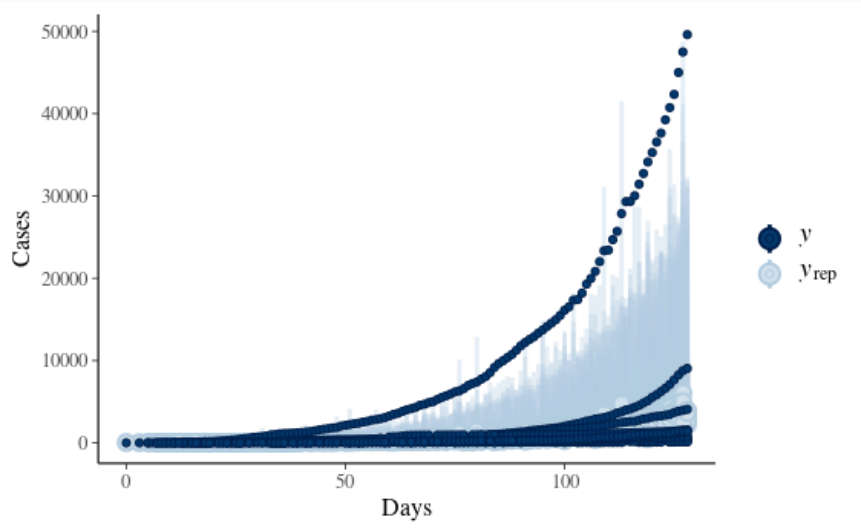
Posterior draws



Residuals Plot



Intervals' plot



Accuracy across departments

A common problem of the previous Bayesian models is that for most departments the predictions are not accurate:

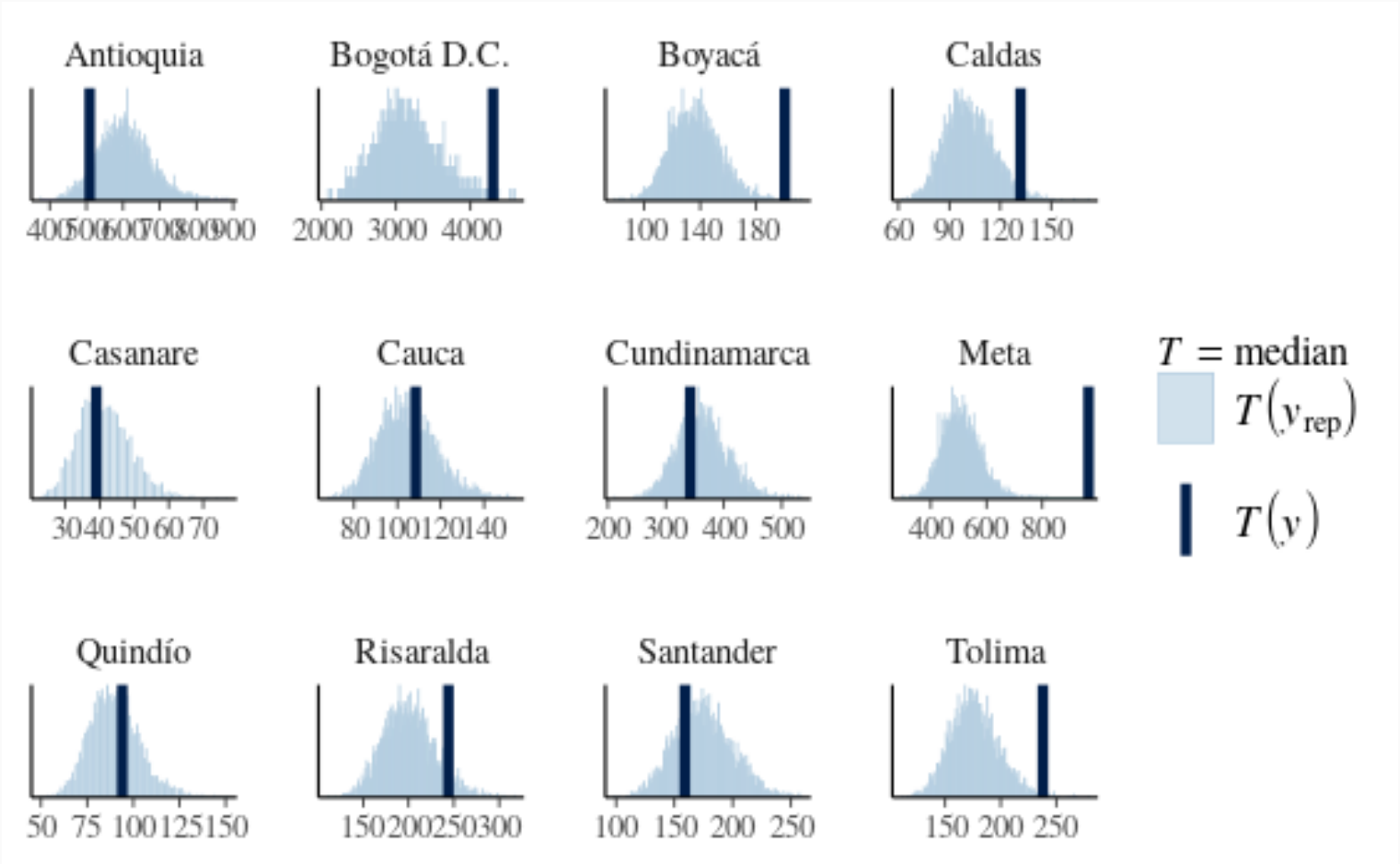


Figure 1: Accuracy for the NB model

Hierarchical model

In order to solve that problem, we fit a model with department-specific intercept term.

So the varying intercept model that we take into account is:

$$\ln(\lambda_{i,d}) = \alpha_d + \beta_{time} \cdot \texttt{Elapsed time}_i + \beta_{age} \cdot \texttt{Mean age}_i$$

$$\alpha_d \sim \mathcal{N}(\mu + \beta_{pop} \cdot \text{People}_d + \beta_{sur} \cdot \text{Surface}_d + \beta_{dens} \cdot \text{Density}_d, \sigma_\alpha)$$

$$y_i \sim \mathcal{N}egative\ \mathcal{B}inomial(\lambda_{i,d}, \phi)$$

The priors used are:

$$\beta_{time} \sim \mathcal{N}(0.5, 1)$$

$$\beta_{age} \sim \mathcal{N}(0, 1)$$

$$\psi \sim \mathcal{N}(0, 1)$$

being $\psi = [\beta_{pop}, \beta_{sur}, \beta_{dens}]$.

New dataset

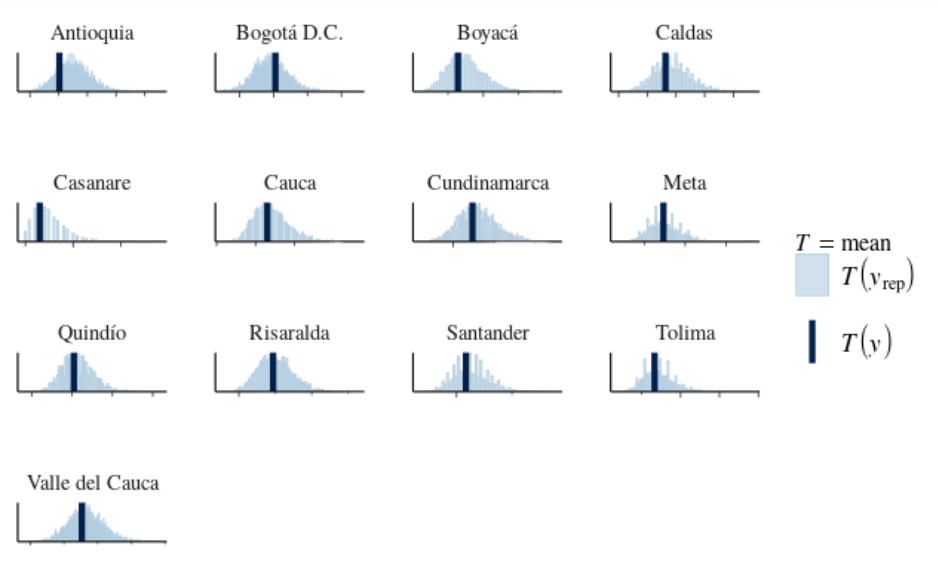
We added the following covariates into the dataset, which are constants for each department:

- `People`: millions of inhabitants for each region;
- `Surface`: km^3, extent of each region;
- `Density`: $\frac{people}{km^2}$, density of the population in each region.

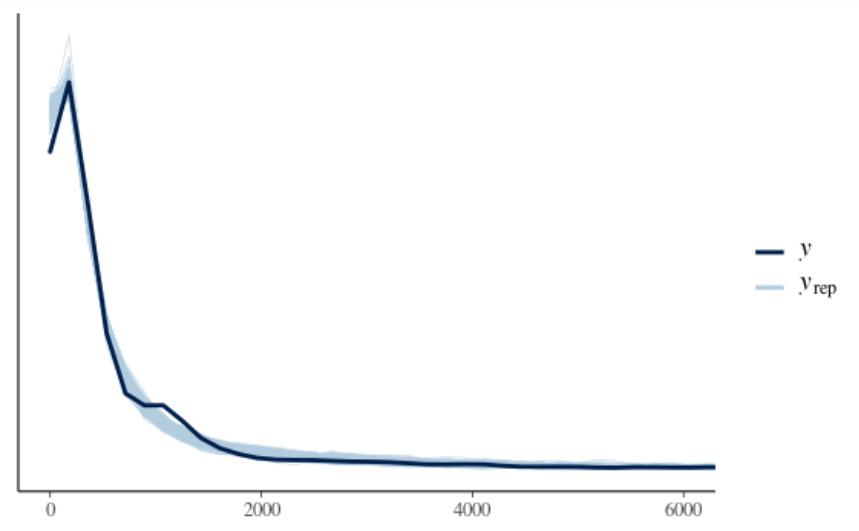
Hierarchical model

The fit has improved: now the distribution of data is coherent with the replicated data.

Departments' accuracy



Posterior draws

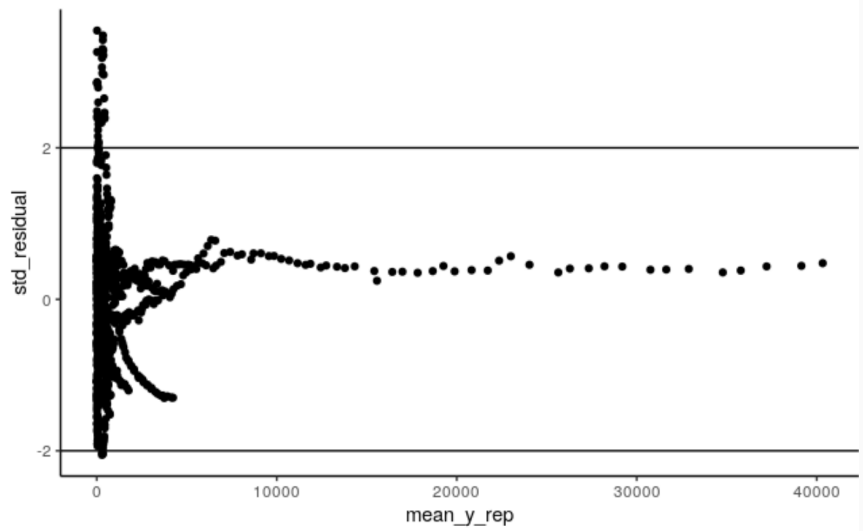


The problem of low accuracy across departments is also solved: now for almost all departments predictions are plausible.

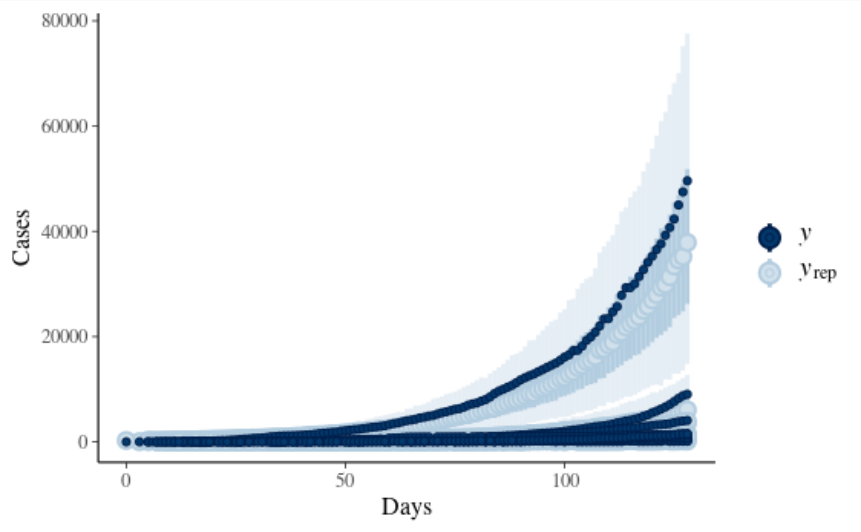
Hierarchical model

From the intervals' plot we can now infer that the model captures the data.

Residuals plot



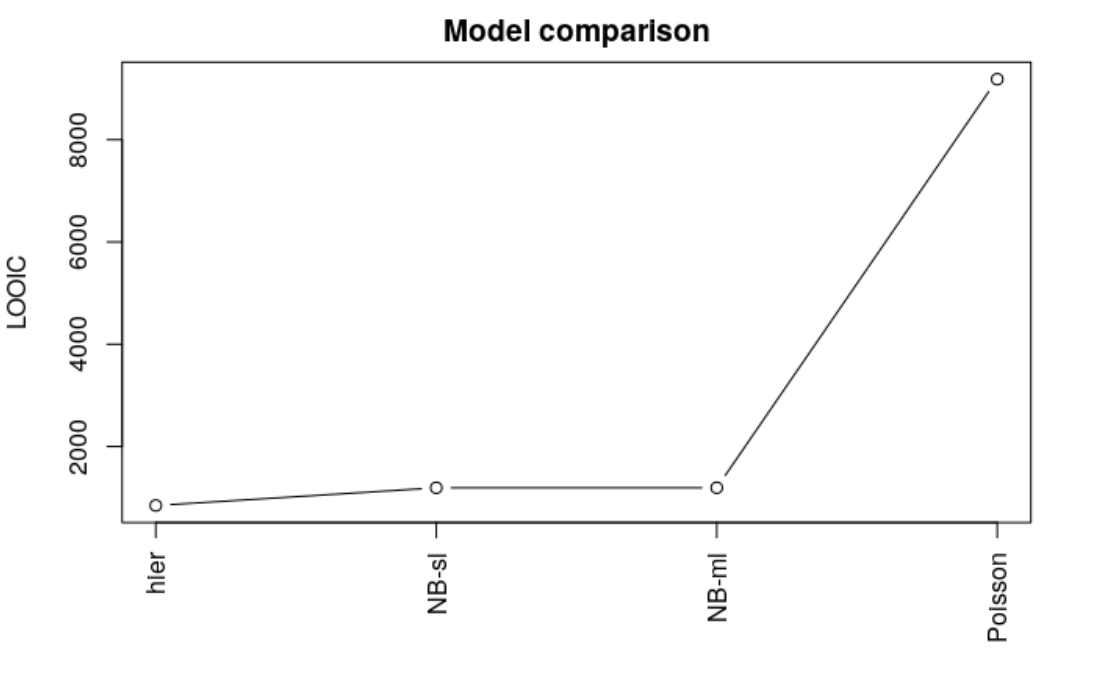
Intervals' plot



Also the residuals are now for the vast majority inside the 95% confidence interval.

LOOIC

The **L**eave **O**ne **O**ut **I**nformation **C**riteria is a method for estimating pointwise out-of-sample prediction accuracy from a fitted Bayesian model using the log-likelihood evaluated at the posterior simulations of the parameter values.



Conclusions

Among all the frequentist model we tried, considering both the reduction on the AIC (among models belonging to the same family) and the reduction on the RMSE, we think that the best are the GAM models.

For the Bayesian models we obtained the best accuracy with the hierarchical model.

Other interesting models we could try with these data are the ARIMA model (which is often used for time series analysis) or the SIR model (which is one of the most popular model for epidemiological data).

Thank you for your attention!

You can find our project in GitHub here: GitHub repository