Università degli studi di Trieste

Master Degree in

Data Science and Scientific Computing

Covid-19 Case Study

Statistical Analysis of Intensive Care in Veneto in Autumn and Winter 2020/2021

Final Project - STATISTICAL METHODS FOR DATA SCIENCE

Group A: Babaei Elham [SM3500466], De Santis Flavia [SM3500482], Doz Romina[SM3500441], Fodor Imola[SM3500474]

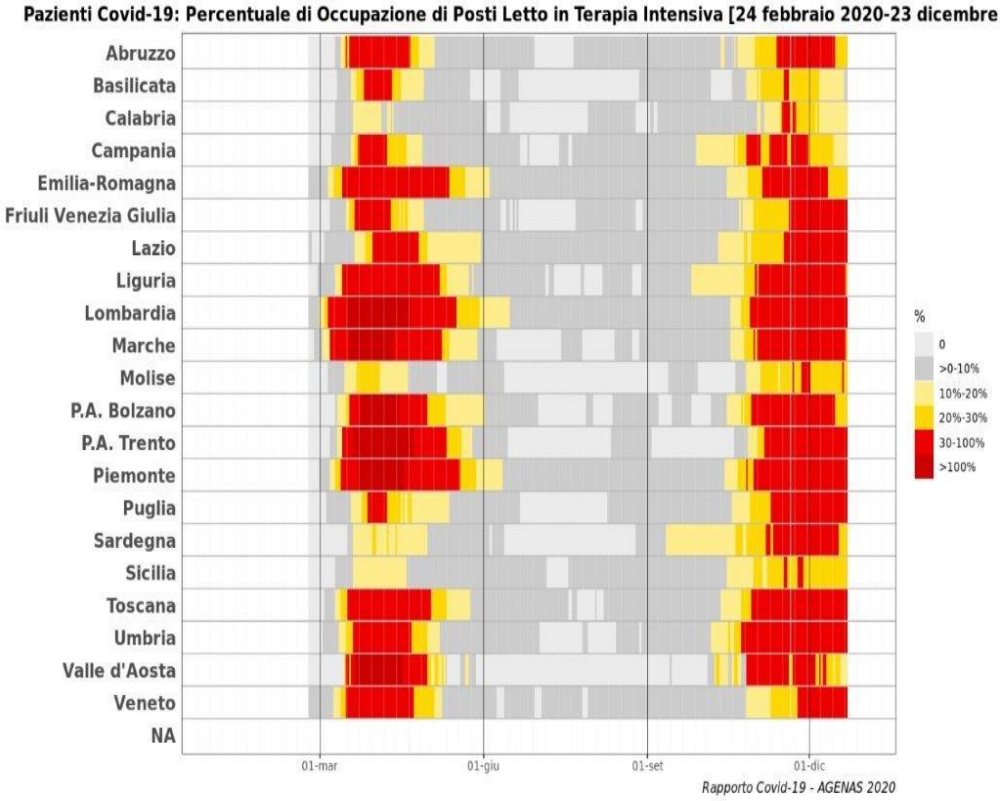
Overview

- Introduction - project motivation
- Explanatory analysis
  - About the dataset
  - Selecting covariates
- Building model
  - Adding new covariates
- Prediction
- An extra approach

Introduction

Why studying intensive care?

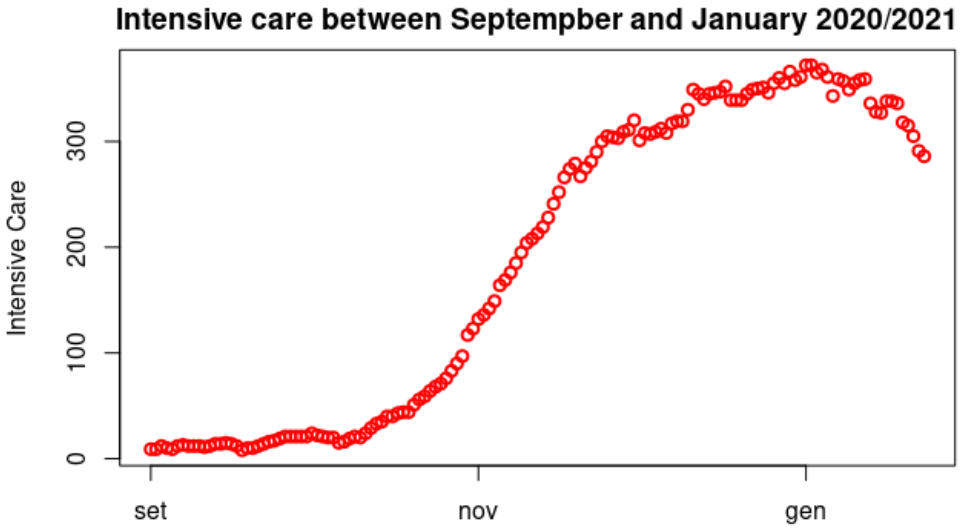
- **Limited ICU beds** to deal with covid-19 patients and those with other pathologies
- Increasing ICU capacity requires more equipment (in particular ventilators) and pharmaceuticals, which might be in **short supply**
- Increasing ICU bed numbers without increasing staff could result in increased mortality. However, doctors and nurses are not easy to find.



Introduction

Why a statistical analysis?

- Derive low-term predictions to get an idea of how to plan/manage the hospital staff/resources in the following weeks
- Understand which are the most relevant factors that determine the increasing of ICU patients
- Suggest possible improvements in the management of the pandemic



Explanatory analysis

The dataset

- The dataset was obtained by the official website of **Protezione Civile starting from 01/09/2020 to 23/01/2021** and considering only region **Veneto**
- Data regarding the place of the survey (latitude, longitude, exc…) have been removed
- Data regarding **variables no longer populated** or that were only collected from a certain date onwards (as ingressi_terapia_intensiva which was collected from 03/12/2020 and so 69% are missing values in the dataset) have been removed, while notes were considered in evaluation of the dataset but not during the modeling procedure

terapia_intensiva	Intensive Care	Intensive_care
ricoverati_con_sintomi	Hospitalised patients with symptoms	Hos_symp
data	Date of notification	Date
totale_ospedalizzati	Total hospitalised patients	Total_Hos
isolamento_domiciliare	Home confinement	Home_con
totale_positivi	Total amount of current positive cases	Total_pos
variazione_totale_positivi	Variation of current positive cases	Variation_pos
nuovi_positivi	Variation of current cases	Variation_cases
dimessi_guariti	Recovered	Recovered
deceduti	Death	Death
totale_casi	Total amount of cases	Total_cases
tamponi	Tests performed	Test
casi_testati	Total number of people tested	People

Explanatory analysis

The dataset

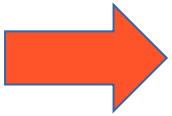
- Blue variables are cumulative and therefore have been **replaced** by the corresponding daily changes.
- Red variables have been **removed** because they are strongly correlated to other variables:

Total_hosp = Hos_symp + Intensive_care

Total_pos = Total_cases – Recovered – Death

Variation_pos = Variation_cases - Recovered_today - Death_today

Variable
Intensive_care
Hos_symp
Date
Total_hosp
Home_con
Total_pos
Variation_pos
Variation_cases
Recovered
Death
Total_cases
Test
People
Zone_color
Lag_zone_color
Season

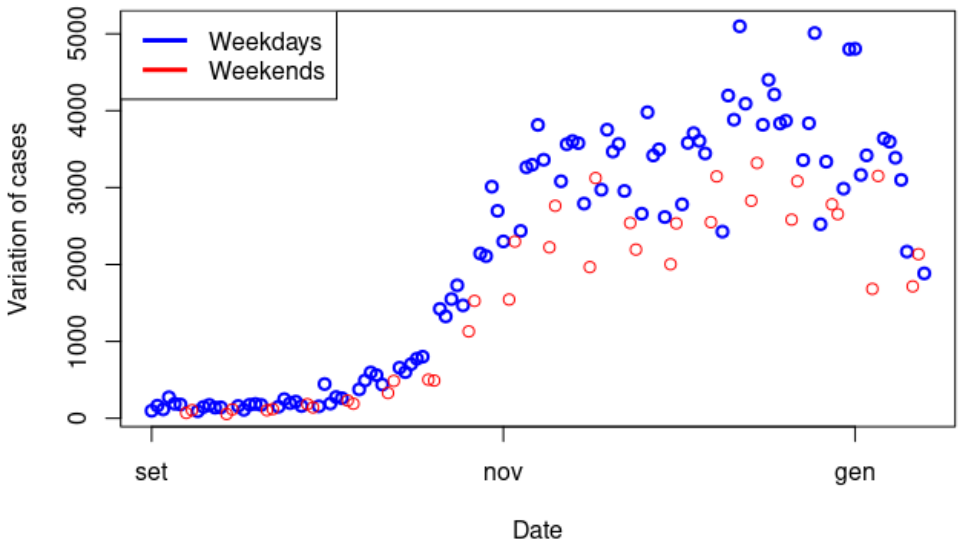


Variable
Intensive_care
Hos_symp
Date
Home_con
Variation_cases
Recovered_today
Death_today
Test_today
People_today
Zone_color
Lag_zone_color
Season

Explanatory analysis

Quality of data

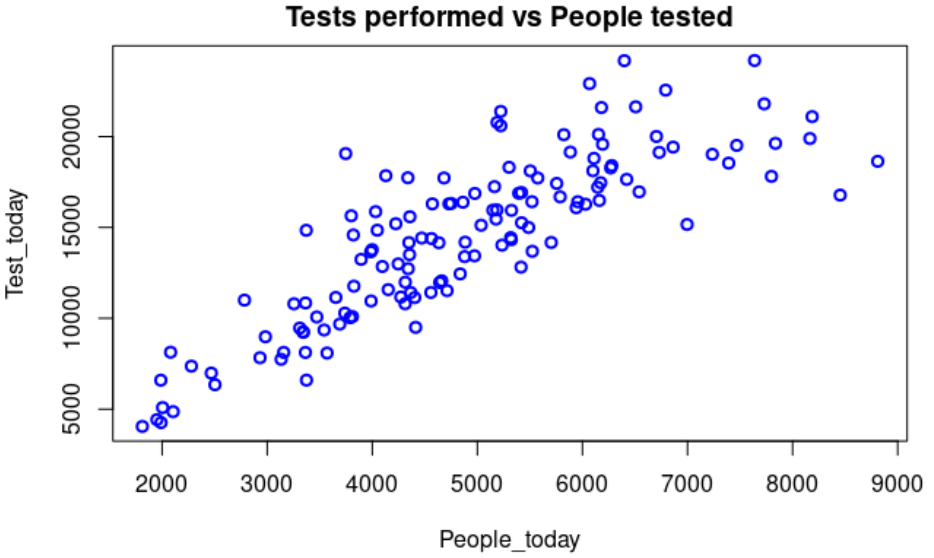
- Despite the fact that data was obtained by the official national source, the reliability depends on the procedures adopted to collect data. In this case, due to relatively **frequent algorithm changes** and new or deleted variables, data-gathering process does not guarantee the most accurate predictions possible
- The dependent variable, intensive care, is not always the effective measured value because there are many **temporal misalignments of the information flow**, as reported in the notes of the dataset.
- On the weekends or holidays, data collection slows down and is retrieved on subsequent weekdays..

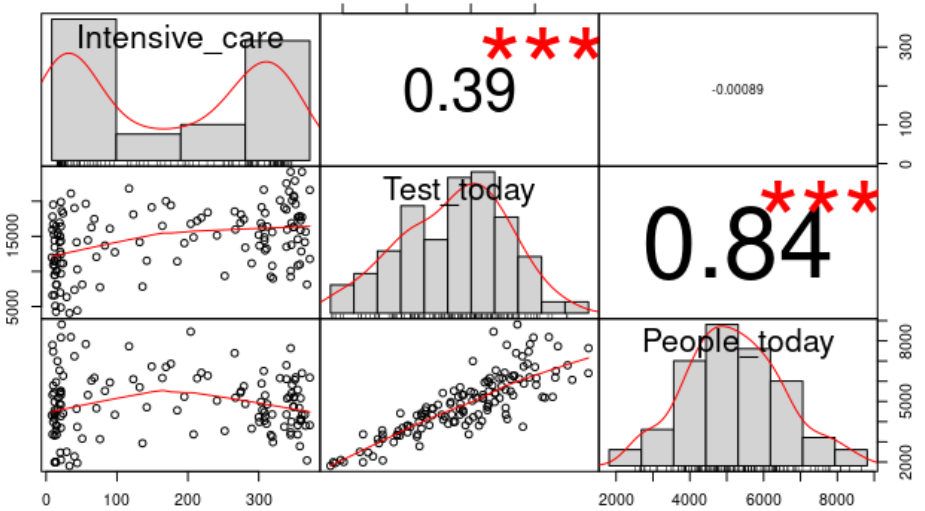


Explanatory analysis

Selecting covariates

- Before starting to create a statistical model, it is convenient to analyze the variables and their relationship with the independent variable or with other covariates.
- There aren't any missing values in the chosen dataset.
- There is a strong correlation between the people tested and the number of tests performed, so only one of them can be used in the model. The number of tests is chosen, being more correlated to the response variable.

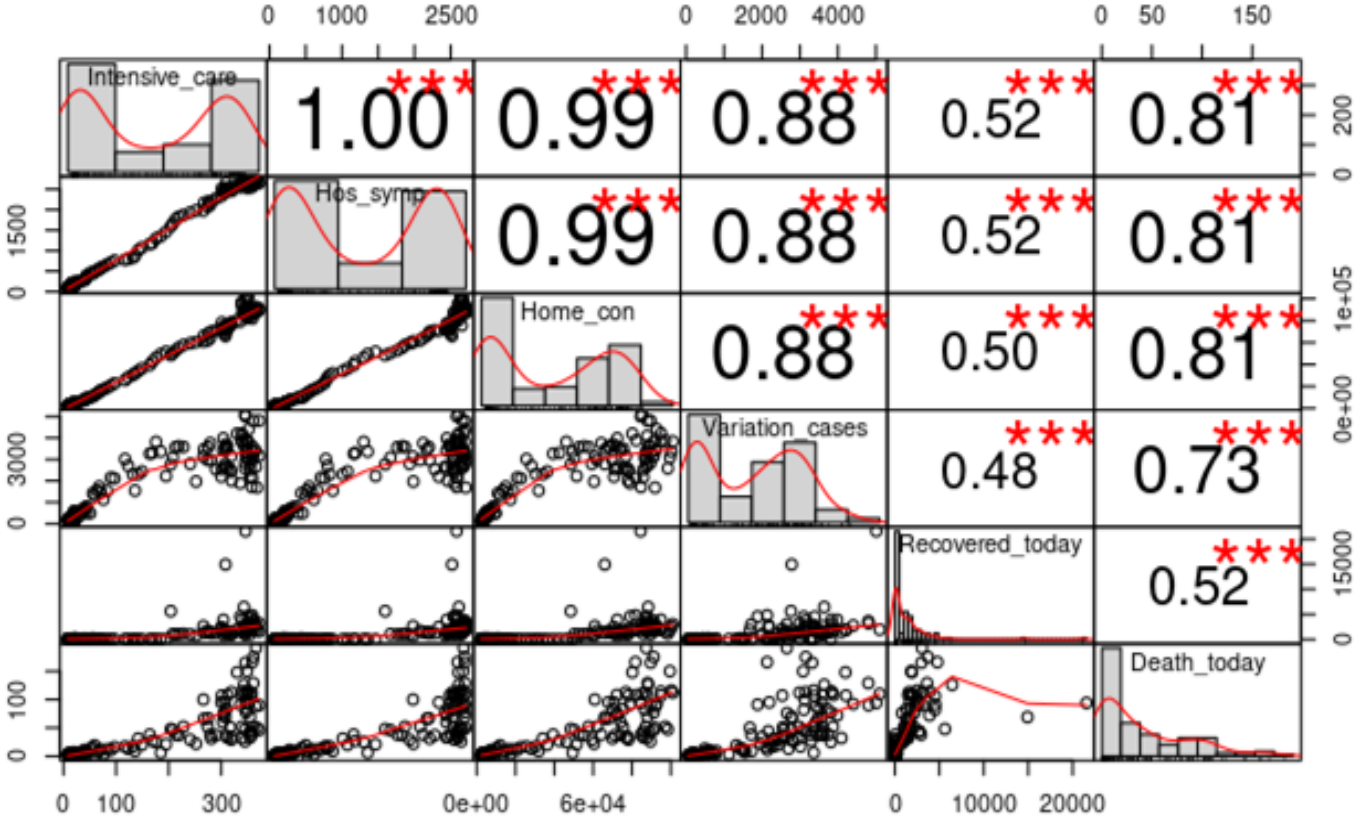




Explanatory analysis

Selecting covariates

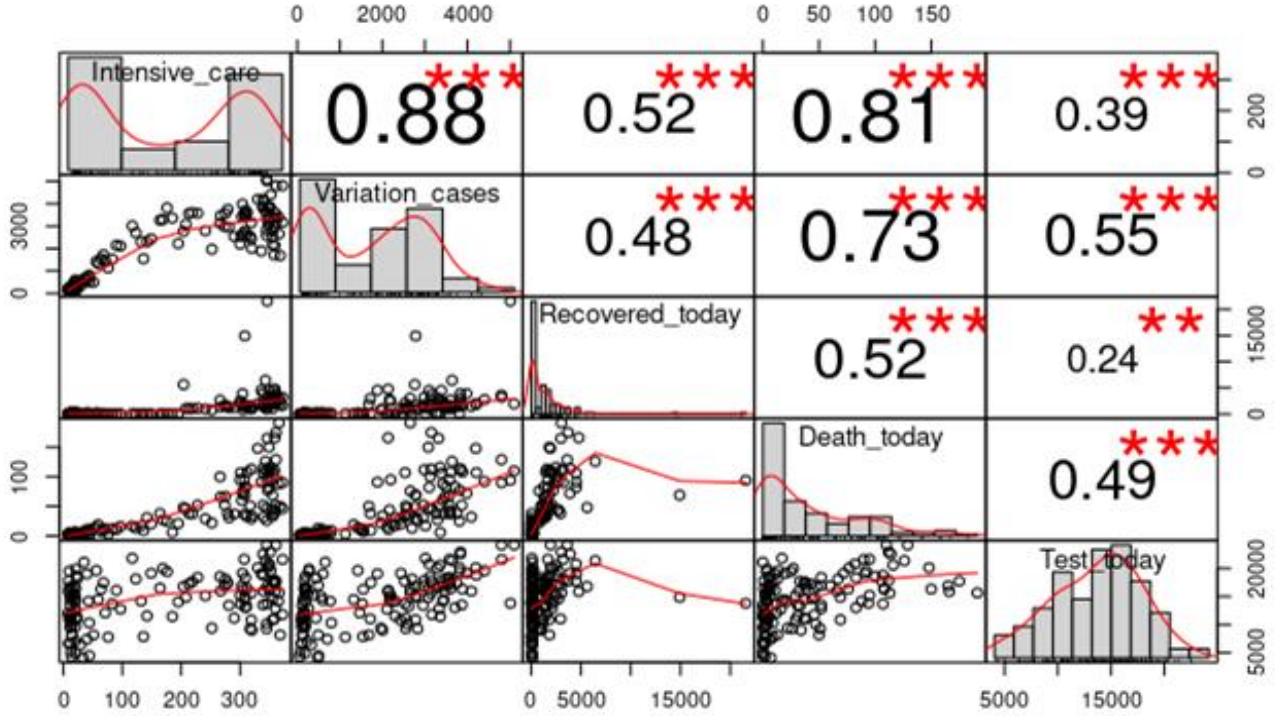
- The other possible predictors are all correlated to the variable intensive care. However, home confinement and hospitalized with symptoms are highly correlated with other covariates and so have been removed.



Explanatory analysis

Selecting covariates

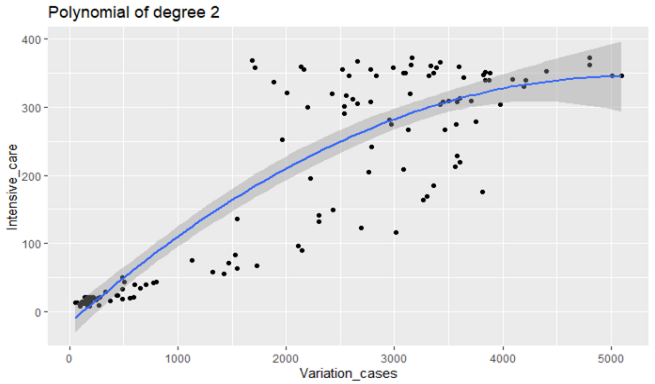
- These are possible predictors that will be considered to obtain a model for the variable intensive care.
- Since there is still a degree of correlation between them (but not so strong like other cases) we will check if any of these variables can be included in the model or not during building the model.

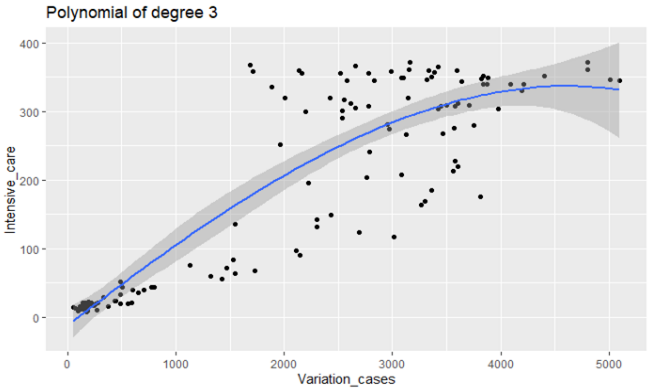


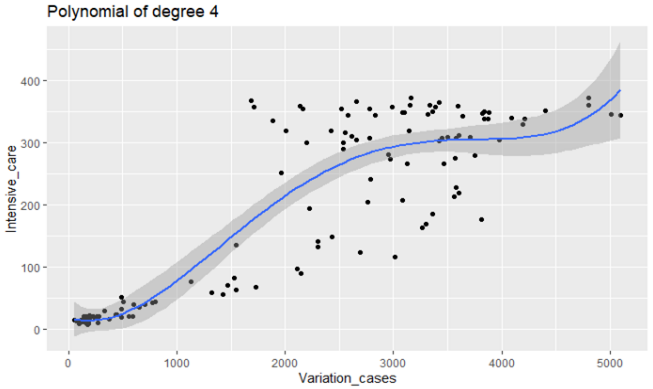
Explanatory analysis

Selecting covariates

- It is possible to approximate the relationship between intensive care and variation of cases via a polynomial of degree two, three or four. As none of them can completely catch the pattern of data due to variation in the variance of data points, the simplest one will be used (degree two) to avoid overfitting.



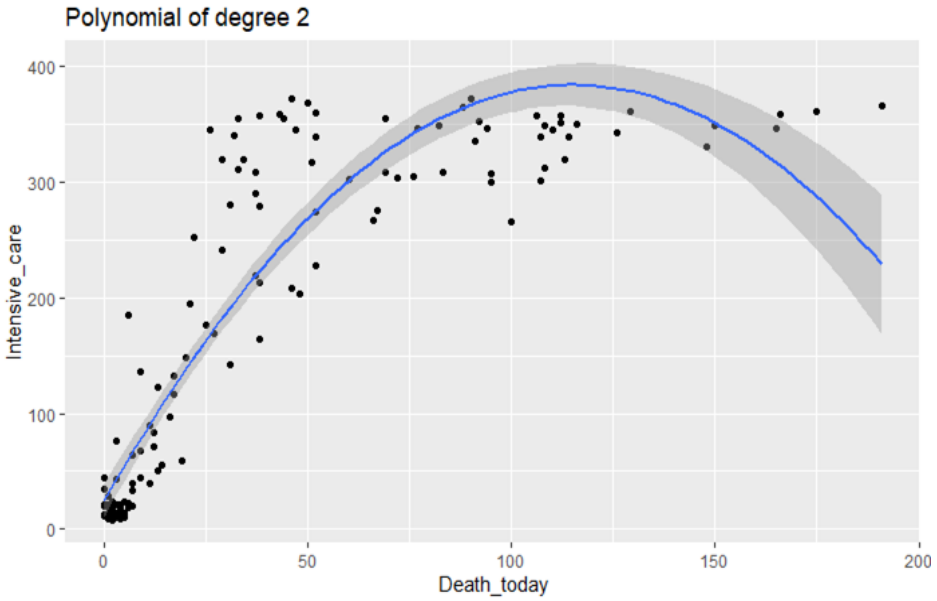


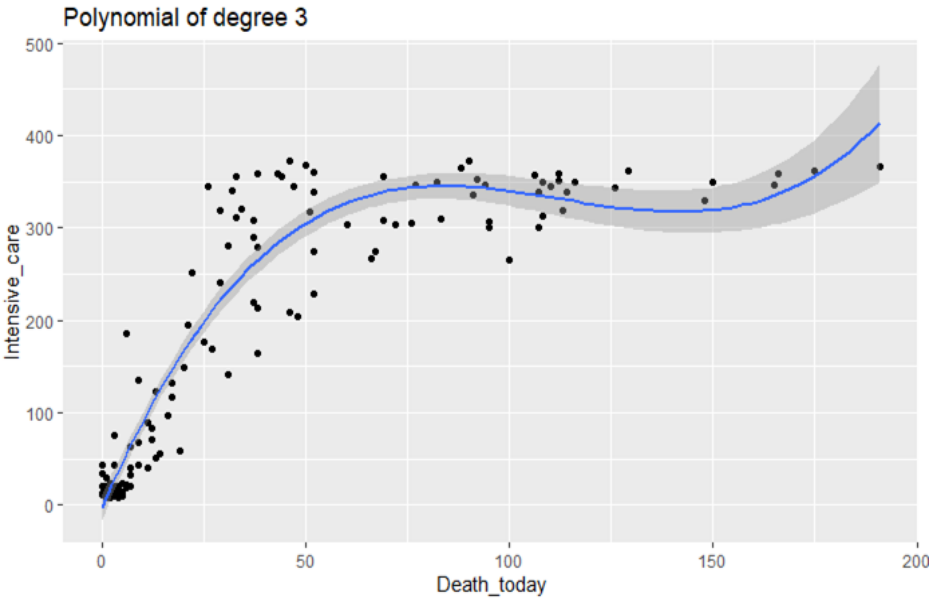


Explanatory analysis

Selecting covariates

- The relationship between intensive care and number of people death can be approximated by a polynomial of degree two or three. The latter is chosen during the building of models because it can capture the curvature of data where the number of deaths is large .

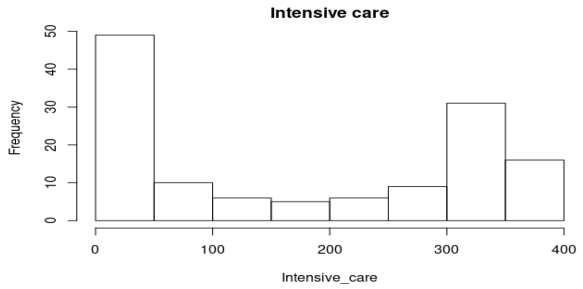




Building model

Model specification

- The aim is to find a model that describes a response variable (intensive care) using multiple predictors.
- The response variable is not normally distributed; it is discrete and non negative. So a simple linear regression model can't be used.
- At first, only the predictors of the original dataset ( ) will be used, later other covariates ( ) will be added.



Response variable
Intensive_care

Possible Predictors
Date
Variation_cases
Recovered_today
Death_today
Test_today
Zone_color
Lag_zone_color
Season

Building model

Generalized Linear Model (GLM)

- It is an extension of linear models, characterized by the following features:
  - linear predictor: $\lambda_i = \sum_{j=1}^{p} x_{ij} \beta_j$
  - link function: $g(E(y_i)) = \lambda_i$
  - the response variable belongs to exponential dispersion family
- The response variable of this problem (intensive care) is a count data and it is assumed to follow a poisson probability distribution in which observations are independent

Link function Linear predictor

$$\ln \lambda_i = b_0 + b_1 x_i$$

$$y_i \sim \text{Poisson}(\lambda_i)$$

Probability distribution

Building model

Criteria

- The strategy used to select the predictors is the stepwise selection, considering the following measures:

  AIC: Akaike Information Criteria
  BIC: Bayesian Information Criteria
  F test: with the hypothesis that data follows the simpler of two nested models
  VIF: Variance Inflation Factor

- **Occam's razor criteria**: the simplest explanation is usually the right one

- The link function is chosen to be the **Canonical** link for Poisson regression which is log (default of Poisson family in R).

Building model

Flow

To build the model:

1. First of all a baseline model is created, by using some of the more important and influential variables which are :

- *Date*
- *Variation_cases*
- *Test_today*

2. We check whether the new variable *Death_today* can be added.
3. We check whether the new variable *Recovered_today* can be added.
4. We build the Poisson model and analyze it.
5. Other possible models are taken into consideration as well:
    - **GLM**
        - **Poisson**
        - **Quasi Poisson**
        - **Negative Binomial**
    - **GAM**
    - **Random Forest**

Building model

Baseline Poisson Model

glm0 <- glm(Intensive_care ~ Date+poly(Variation_cases,2)+Test_today, family = poisson, data=d.train)

AIC	BIC	Residual_Deviance	P_value
1607.79	1622.32	737.67	1.18e-85

- The P_value of F test is small so our model works better than the null model in which only intercept is included.
  P_value <- pchisq(glm0$deviance,glm0$df.residual, lower.tail = F)

- VIF is not greater than 10 for any variable.

Building model

Adding the variable Death_today

model.glm <- glm(Intensive_care ~ Date+poly(Variation_cases,2)+Test_today+ poly(Death_today,3), family = poisson, data=d.train)

AIC	BIC	Residual_Deviance
1469.58	1492.83	593.47

- The above table shows that AIC and BIC are smaller compared to baseline model.
- There is no VIF value greater than 10 for all the variables.
- So we add *Death_today* to the model.

Building model

Adding the variable Recovered_today

model.glm <- glm(Intensive_care ~ Date+poly(Variation_cases,2)+Test_today+ poly(Death_today,3)+Recovered_today, family = poisson, data=d.train)

AIC	BIC	Residual_Deviance
1469.66	1495.81	591.54

- The above table shows there is no improvement in AIC and BIC. Thus we do not add *Recovered_today*.

- **Note**: we also tried other options to check the existence of variables in the model; e.g. we added *Recovered_today* before adding *Death_today,* or we considered transformed version of variables such as log(.) etc, but we couldn't find any considerable form that works better.
  Moreover, no possible interaction between variables could be added to improve the model.

Building model

Poisson Model

model.glm <- glm(Intensive_care ~ Date+poly(Variation_cases,2)+Test_today+ poly(Death_today,3), family = poisson, data=d.train)

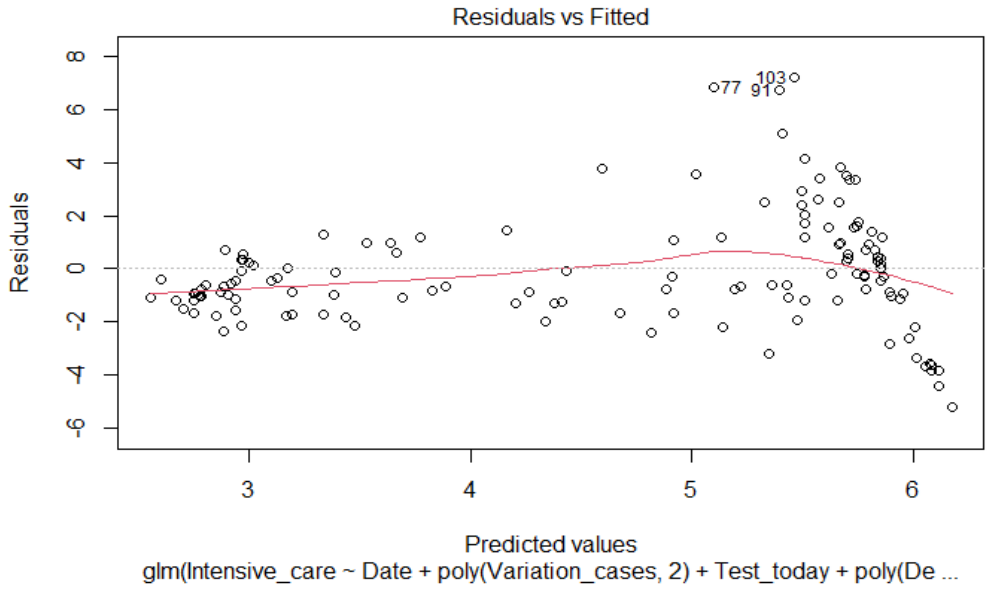
AIC	BIC	Residual_Deviance	P_value
1469.58	1492.83	593.47	2.29e-61

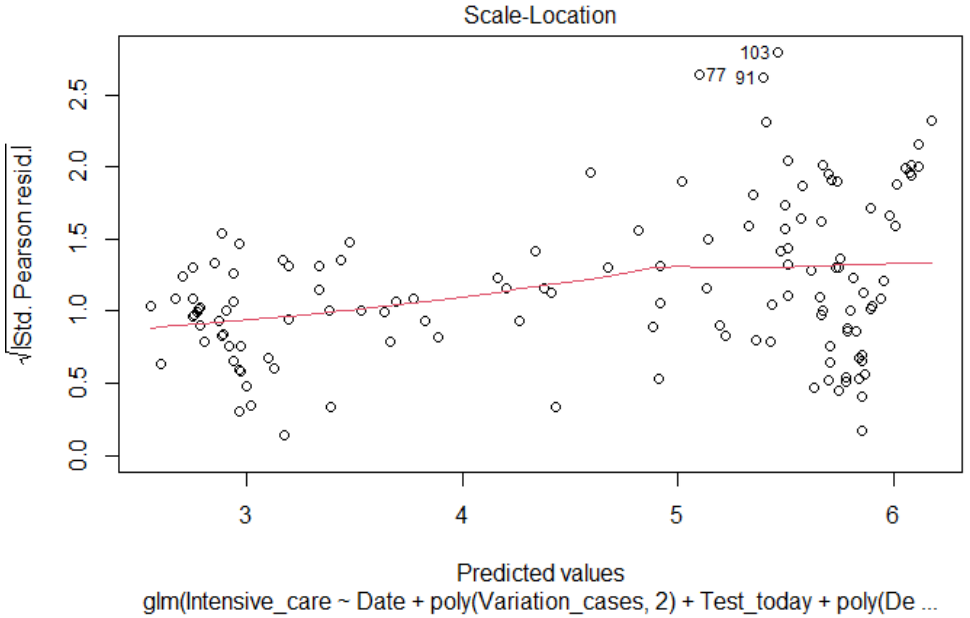
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.4489  -1.2260  -0.4725   0.9442   6.7203

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -1.955e+02  6.927e+00 -28.216  < 2e-16 ***
Date                        1.079e-02  3.726e-04  28.958  < 2e-16 ***
poly(Variation_cases, 2)1   7.618e+00  2.424e-01  31.434  < 2e-16 ***
poly(Variation_cases, 2)2  -2.592e+00  1.059e-01 -24.474  < 2e-16 ***
Test_today                 -2.243e-05  2.244e-06  -9.997  < 2e-16 ***
poly(Death_today, 3)1       2.164e+00  1.906e-01  11.350  < 2e-16 ***
poly(Death_today, 3)2      -1.422e+00  1.249e-01 -11.388  < 2e-16 ***
poly(Death_today, 3)3       7.235e-01  9.472e-02   7.638  2.2e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 18827.45  on 134  degrees of freedom
Residual deviance:   593.47  on 127  degrees of freedom
AIC: 1469.6

Number of Fisher Scoring iterations: 4
```

Predictors
Date
Var_cases
Death_today
Test_today

Building model

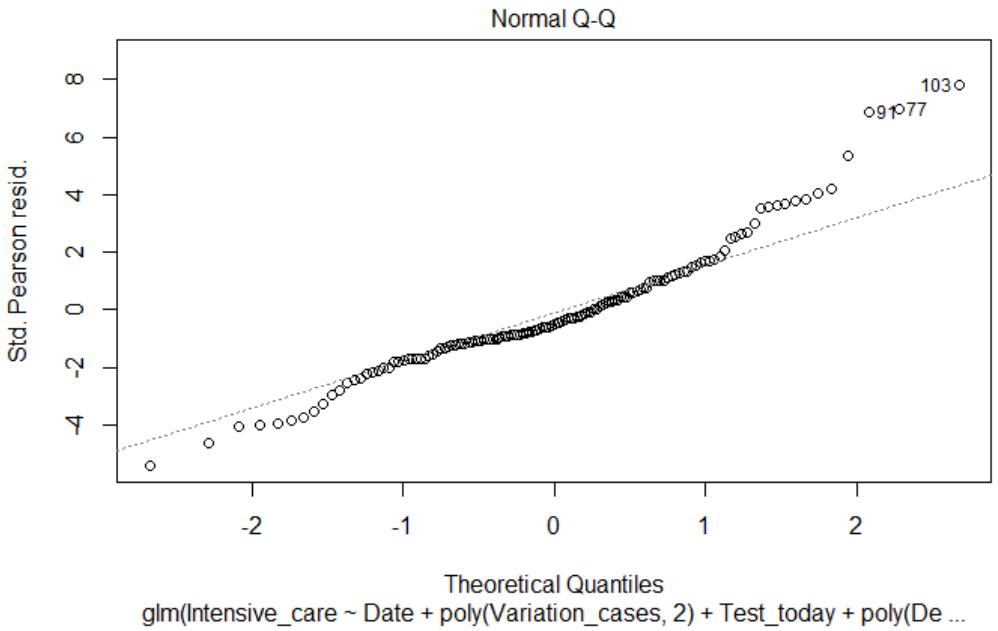
Poisson Model Plots

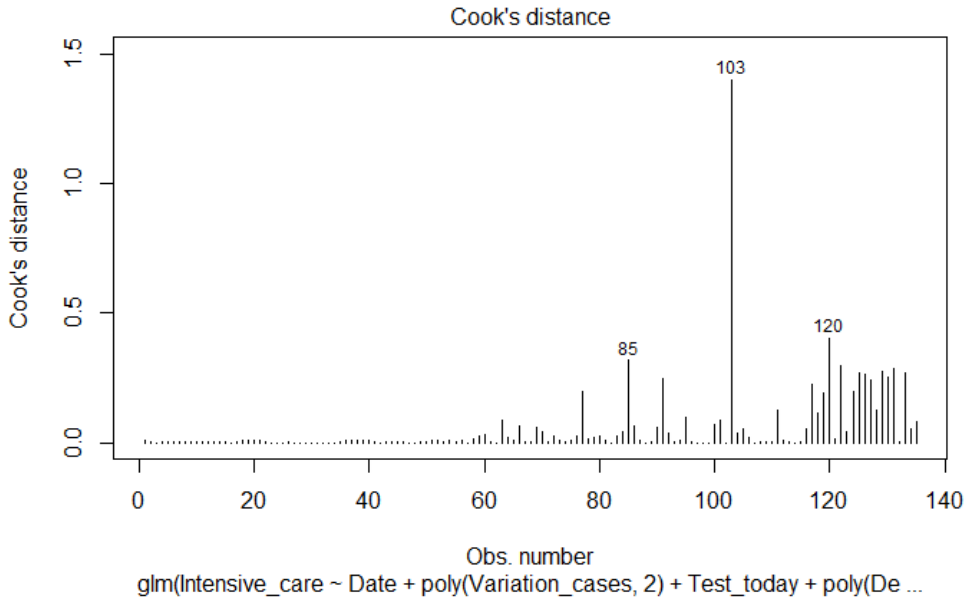




Building model

Poisson Model Plots

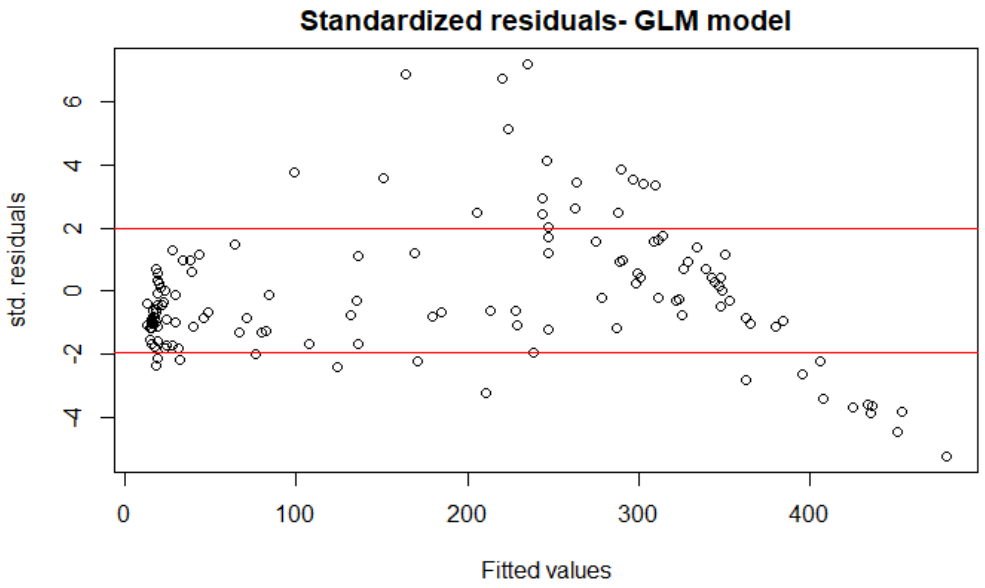


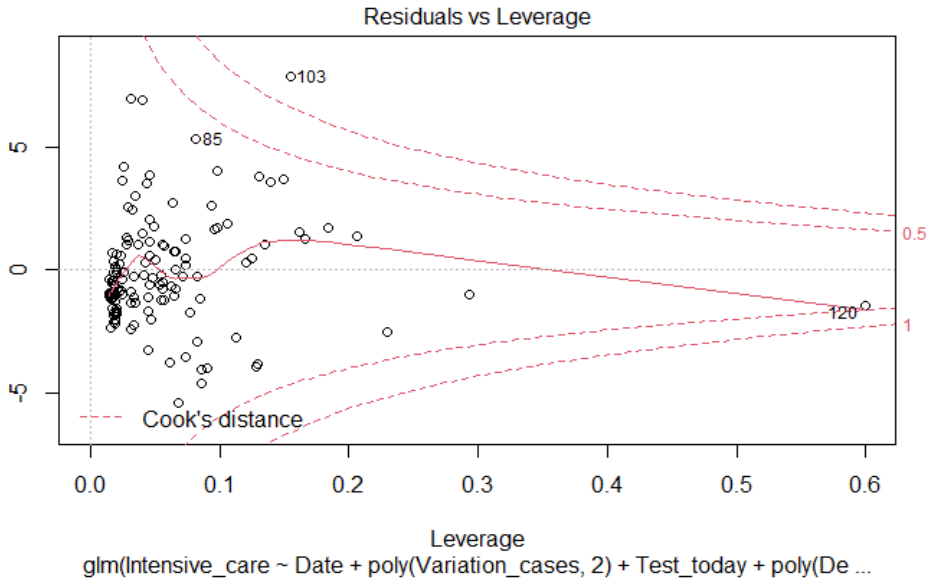


Building model

Poisson Model Plots

- The standardized residual plot shows than many residuals are located outside the interval [-1.96,1.96] of standard normal distribution. It means that there is evidence of overdispersion and we try other models to cope with it such as **Quasi Poisson** and **Negative Binomial**.





- Besides, the Residual vs Leverage plot reveals that the value of Cook's distance metric is larger than 1 for an observation and there is also an observation with large Leverage. It means these observations are probable to be outliers but since the model is over-dispersed, we check them in the next models.

Building model

Quasi-Poisson Model

model.glm.quasi <- glm(Intensive_care ~ Date+poly(Variation_cases,2)+ poly(Death_today,3)+Test_today, family = quasipoisson, data=d.train)

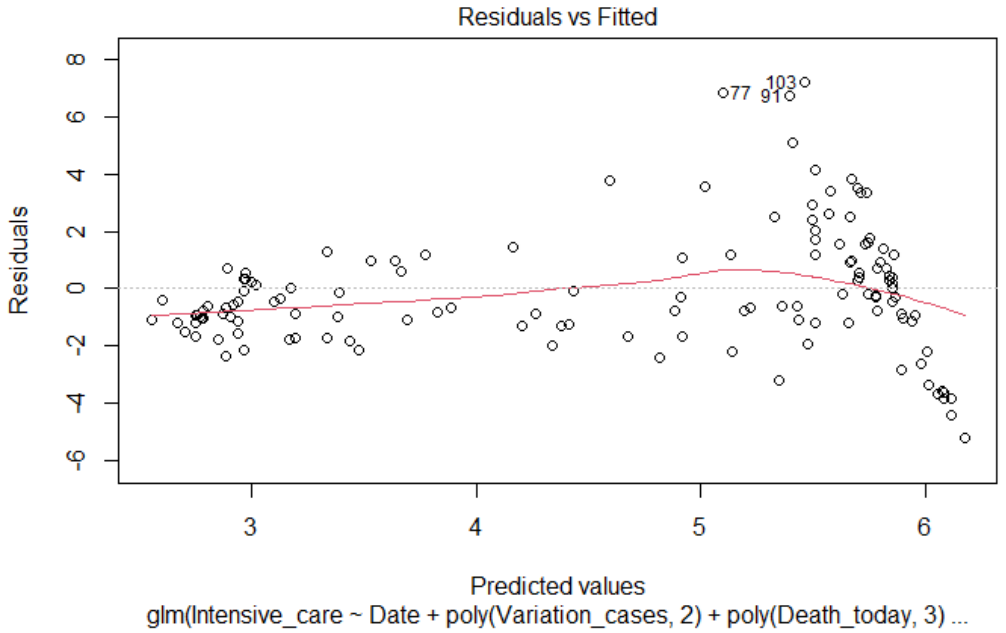
```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.955e+02  1.511e+01 -12.936  < 2e-16 ***
Date                     1.079e-02  8.126e-04  13.276  < 2e-16 ***
poly(Variation_cases, 2)1  7.618e+00  5.286e-01  14.411  < 2e-16 ***
poly(Variation_cases, 2)2 -2.592e+00  2.310e-01 -11.220  < 2e-16 ***
poly(Death_today, 3)1      2.164e+00  4.158e-01   5.204 7.63e-07 ***
poly(Death_today, 3)2     -1.422e+00  2.723e-01  -5.221 7.06e-07 ***
poly(Death_today, 3)3      7.235e-01  2.066e-01   3.502 0.000638 ***
Test_today               -2.243e-05  4.894e-06  -4.583 1.08e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 4.757619)

    Null deviance: 18827.45  on 134  degrees of freedom
Residual deviance:   593.47  on 127  degrees of freedom
AIC: NA
```

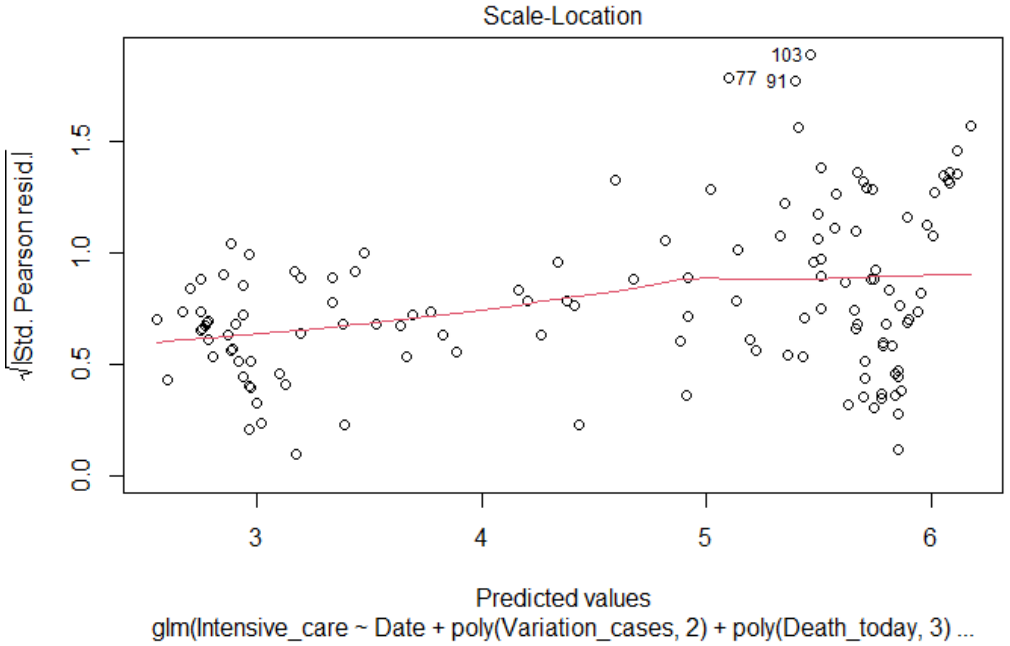
Predictors
Date
Var_cases
Death_today
Test_today

- The coefficients are the same as poisson.
- Dispersion parameter is 4.76 > 1 .

Building model

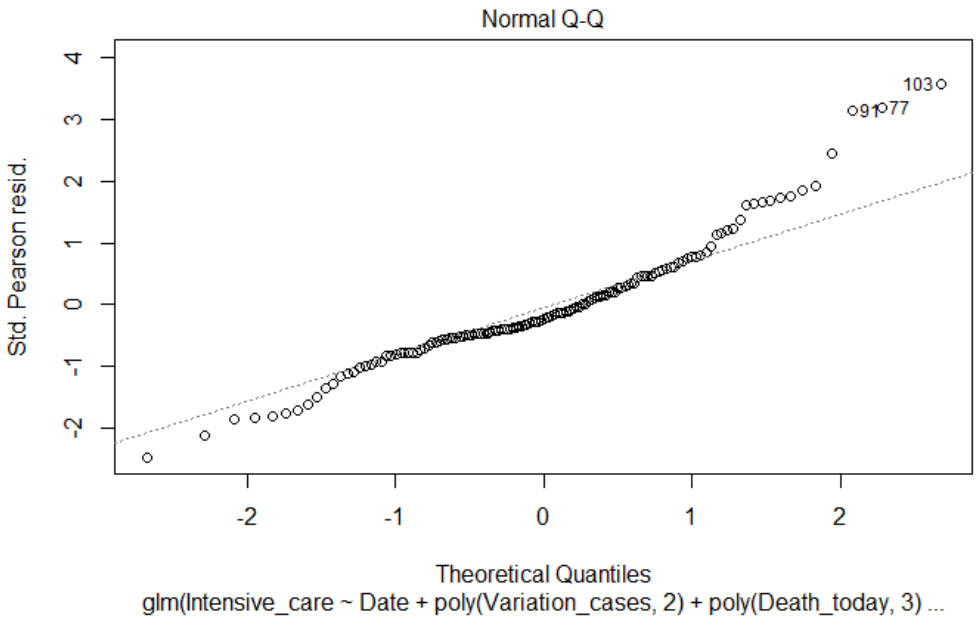
Quasi-Poisson Model Plots

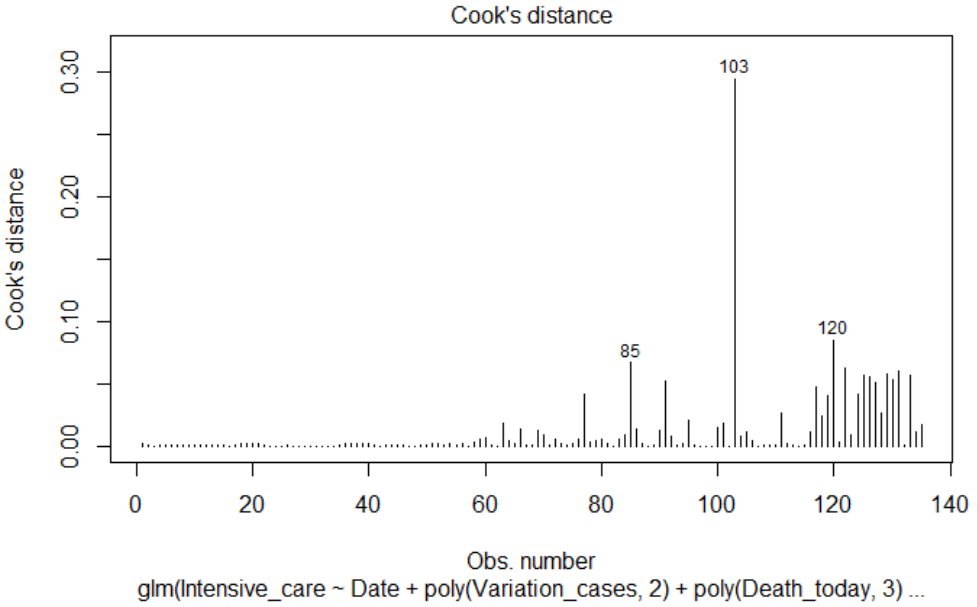




Building model

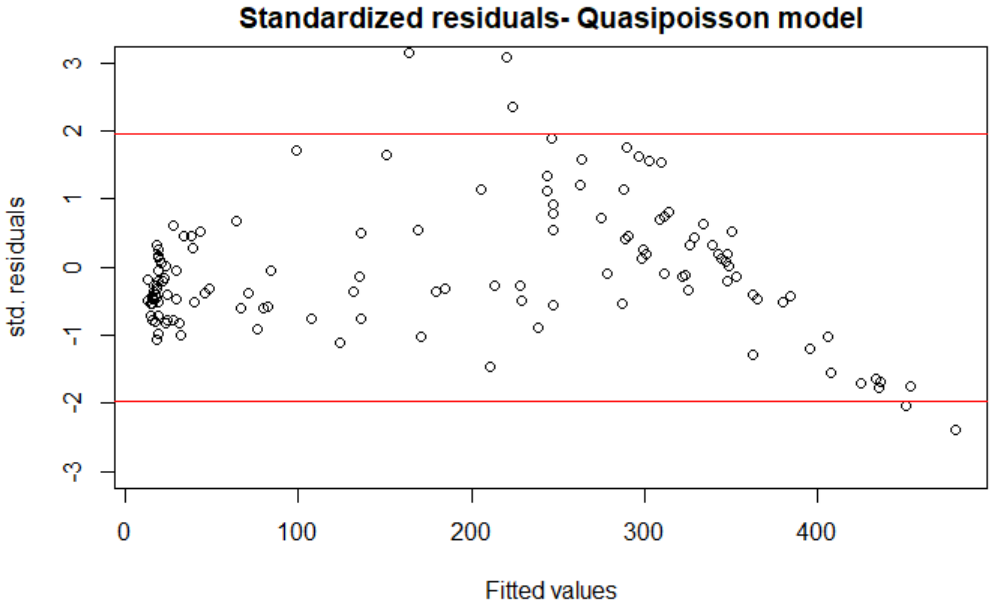
Quasi-Poisson Model Plots

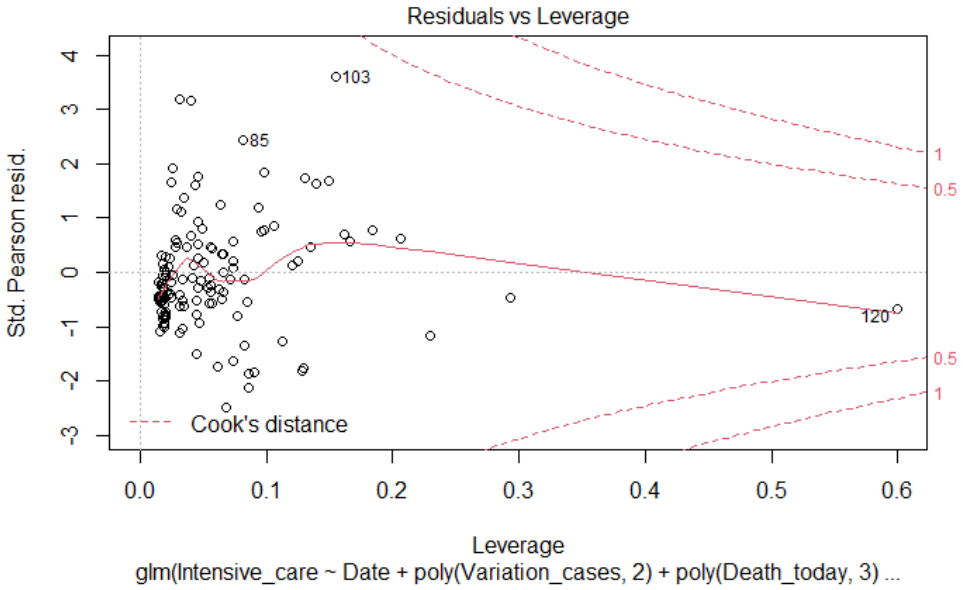




Building model

Quasi-Poisson Model Plots





- The standardized residual plot shows that now many of the residuals are between [-1.96,1.96].
- There is no strong outlier according to Residual vs Leverage plot anymore and also other plots. By looking at the dataset it turns out datapoint 120 which still has a larger leverage corresponds to an observation with maximum number of death from 1th Sep to 13th Jan.

Building model

Negative Binomial Model (NB)

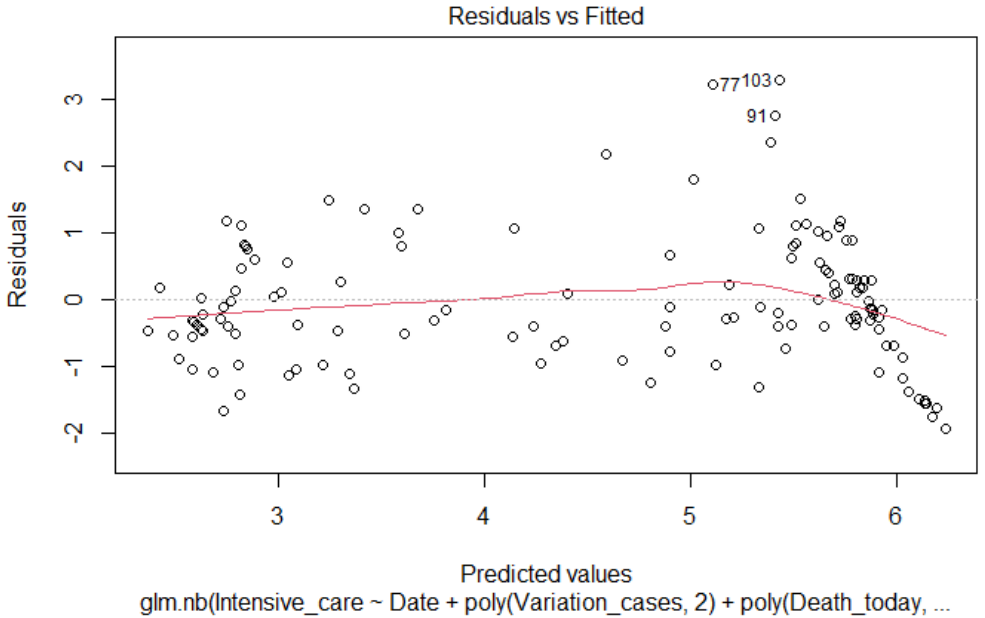
model.glm.nb <- glm.nb(Intensive_care ~ Date+poly(Variation_cases,2)+ poly(Death_today,3)+Test_today, data=d.train)

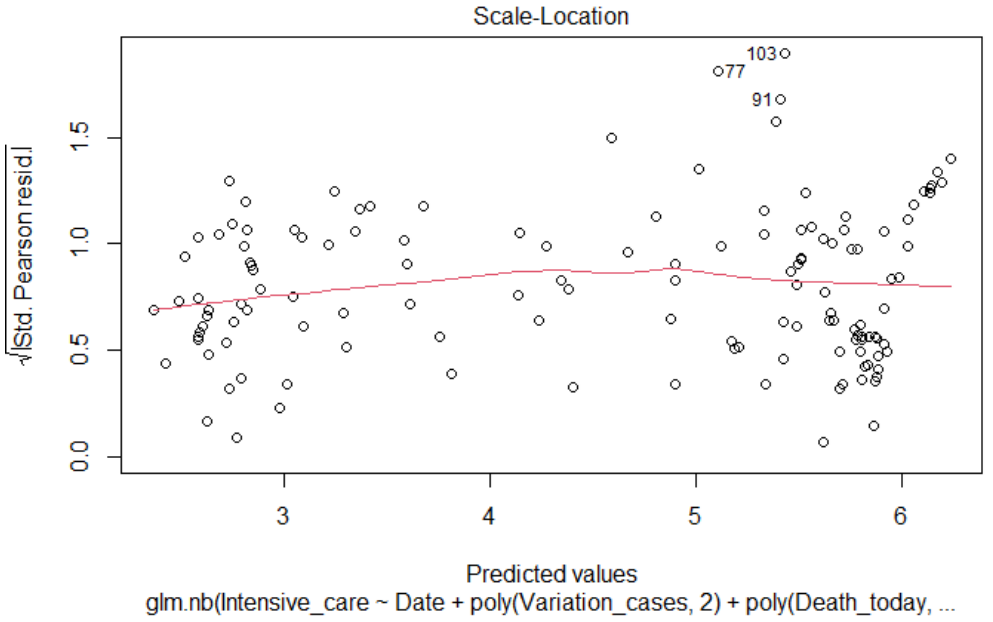
```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.238e+02  1.767e+01 -12.664  < 2e-16 ***
Date                      1.231e-02  9.504e-04  12.954  < 2e-16 ***
poly(Variation_cases, 2)1 8.216e+00  4.895e-01  16.784  < 2e-16 ***
poly(Variation_cases, 2)2 -2.844e+00  2.260e-01 -12.586  < 2e-16 ***
poly(Death_today, 3)1      1.804e+00  4.335e-01   4.162 3.15e-05 ***
poly(Death_today, 3)2     -1.286e+00  2.846e-01  -4.518 6.23e-06 ***
poly(Death_today, 3)3      5.810e-01  2.304e-01   2.522   0.0117 *
Test_today                -2.499e-05  4.849e-06  -5.153 2.56e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(50.4787) family taken to be 1)

    Null deviance: 5497.96  on 134  degrees of freedom
Residual deviance:  122.33  on 127  degrees of freedom
AIC: 1169.1
```

Predictors
Date
Var_cases
Death_today
Test_today

Building model

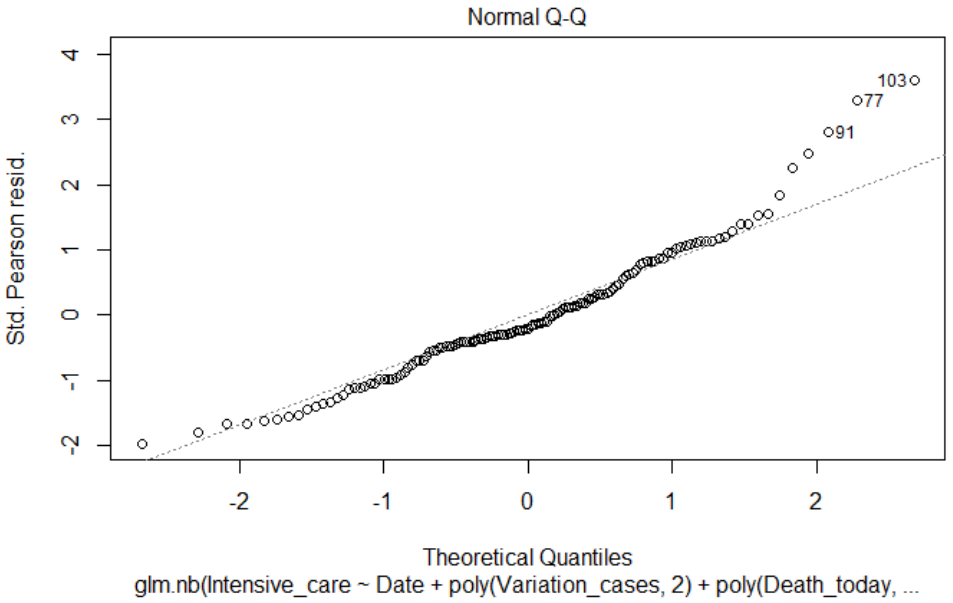
Negative Binomial Model (NB) Plots

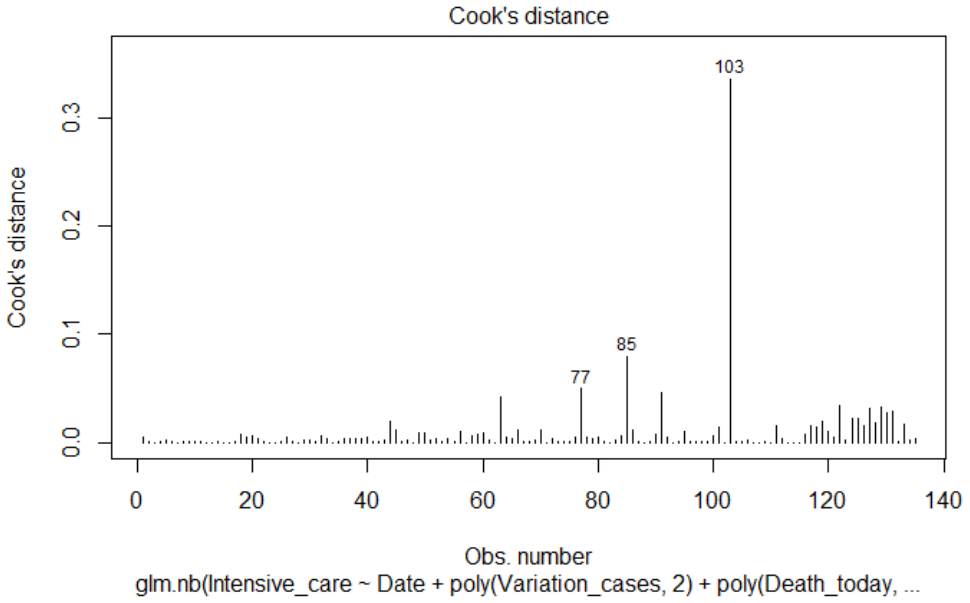




Building model

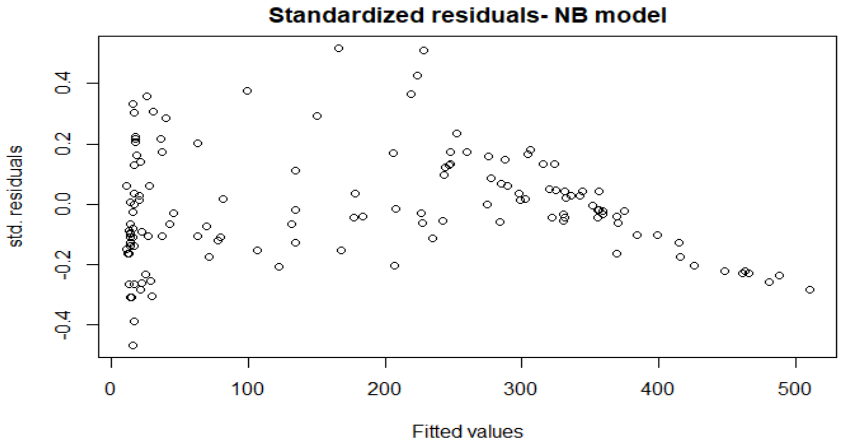
Negative Binomial Model (NB) Plots

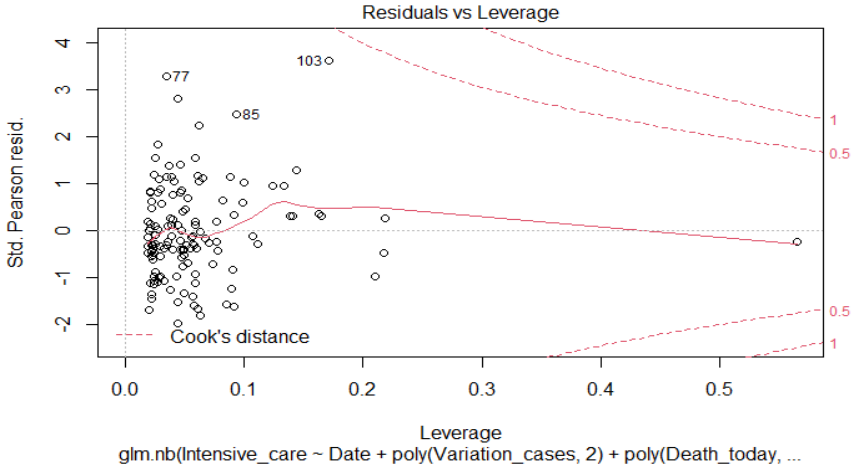


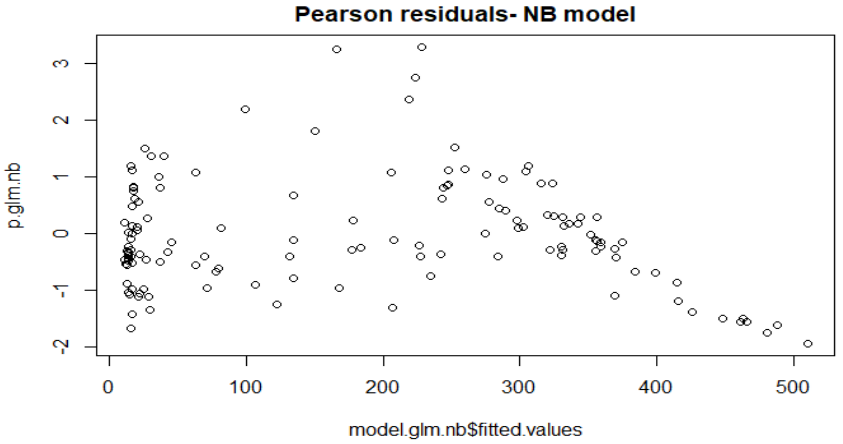


Building model

Negative Binomial Model (NB) Plots







Building model

GAM Model

model.gam <- gam(Intensive_care ~ Date+s(Variation_cases)+ s(Death_today)+Test_today, family = poisson, data=d.train)

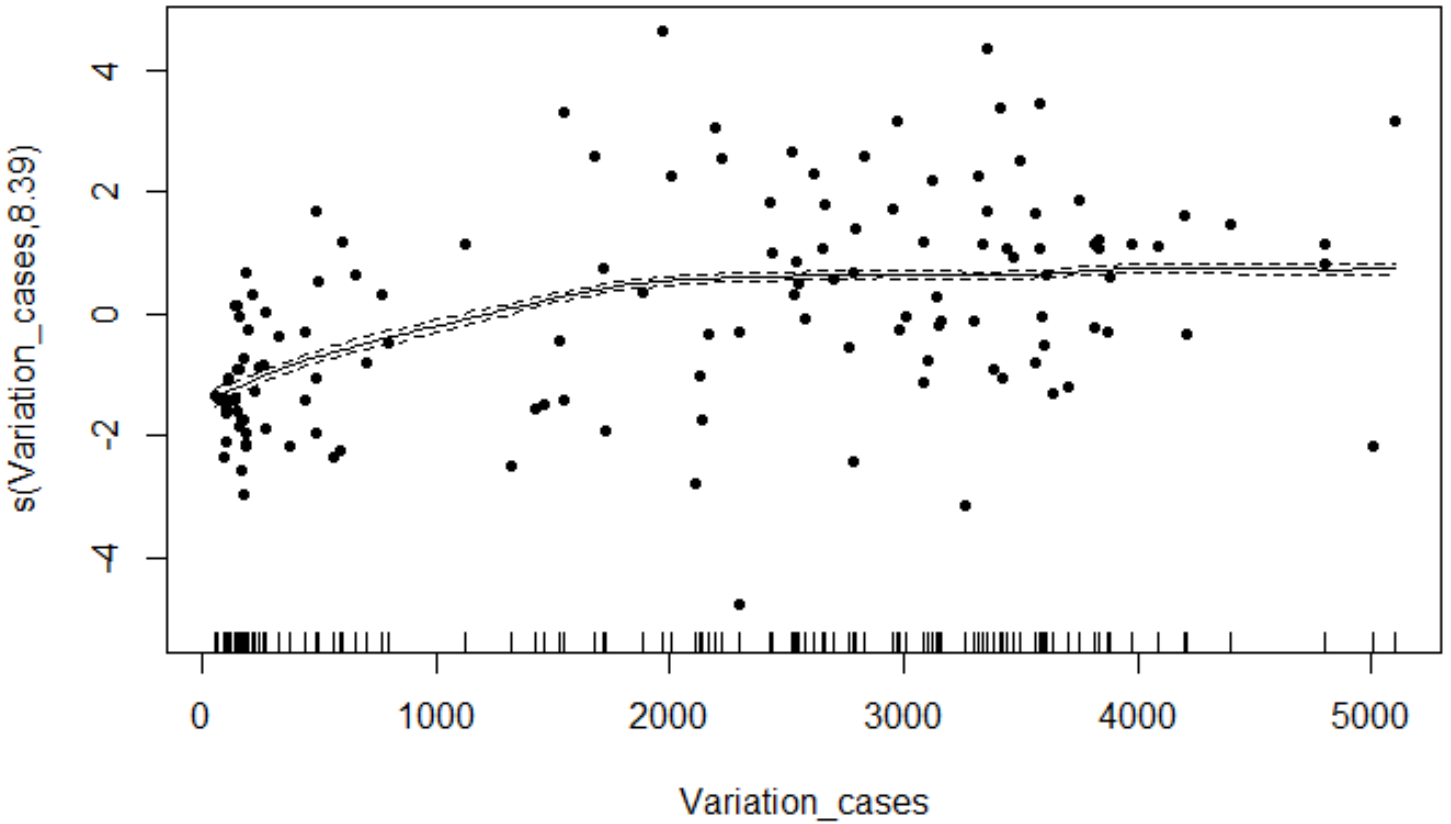
```
Parametric coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.517e+02  7.866e+00 -19.280   <2e-16 ***
Date         8.428e-03  4.232e-04  19.912   <2e-16 ***
Test_today  -2.200e-05  2.352e-06  -9.355   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                     edf Ref.df Chi.sq p-value
s(Variation_cases) 8.395  8.879  535.2  <2e-16 ***
s(Death_today)     8.851  8.992  208.0  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.975   Deviance explained = 98.4%
UBRE = 1.5428  Scale est. = 1         n = 135
s(Variation_cases)     s(Death_today)
        0.10128214         0.03071374
```

Predictors
Date
Var_cases
Death_today
Test_today

- estimated degree of freedom (edf) is larger than 1 for both nonlinear parts. So it is true to consider smooth function for them
- Adjusted R-square is 97% which is good. Also, the deviance in response variable explained by the model is 98%.

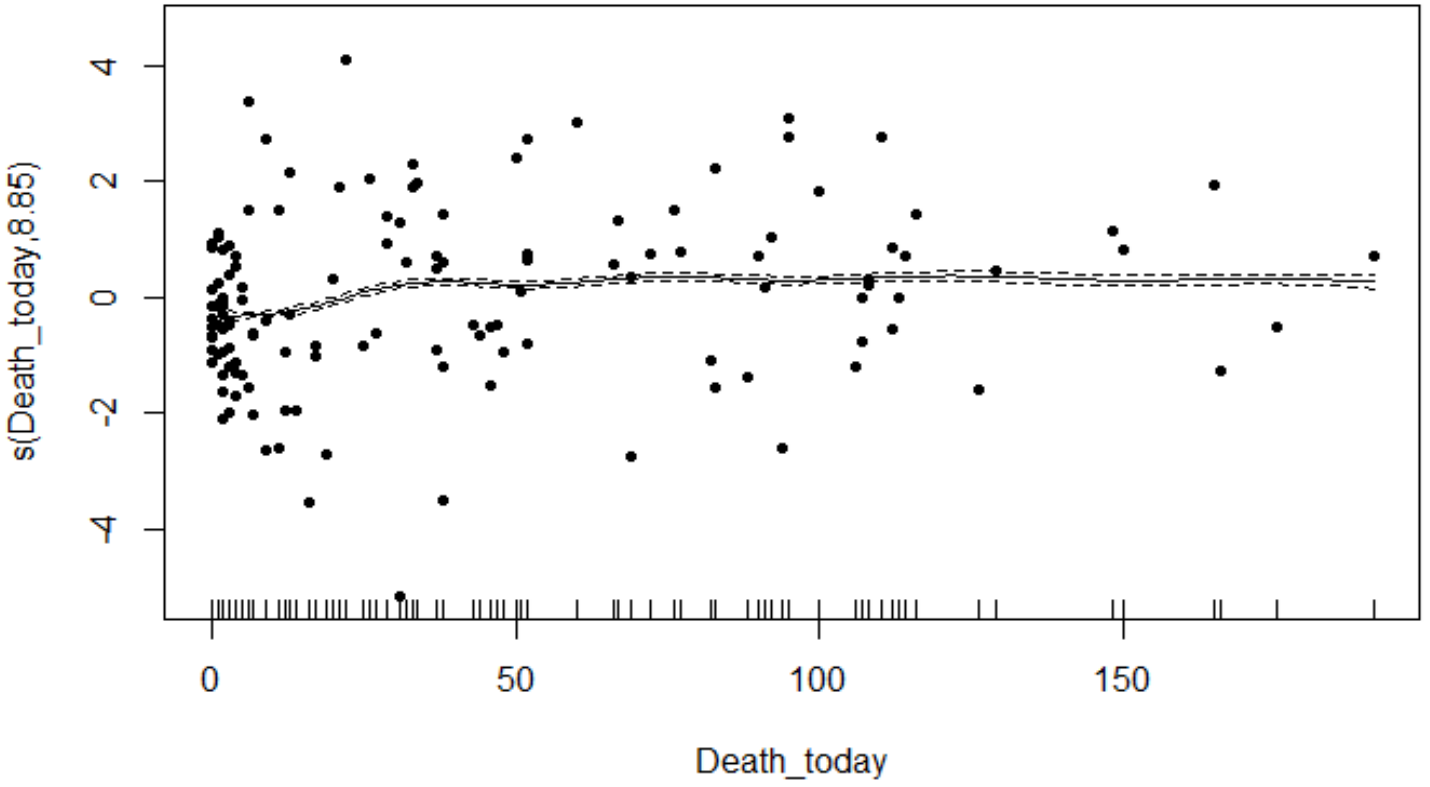
Building model

GAM Model Plots



Building model

GAM Model Plots



Building model

Random Forest model (RF)

- Random Forest is one of the ensemble techniques that use a bag of trees instead of just one single tree to build a model. Hence, the outputs are more reliable.

```
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 1

          Mean of squared residuals: 365.7855
                    % Var explained: 98.23
                   %IncMSE IncNodePurity
Date             13476.1531     1022872.9
Variation_cases   7120.0330      788339.4
Death_today       8792.0797      785508.2
Test_today         436.5875      154501.9
```

Predictors
Date
Var_cases
Death_today
Test_today

- The number of trees is by default 500.
- The variation of response variable explained by the model is 98%.
- The most important variable (ignoring the variable date) is *Variation_cases* with the largest amount increasing in node purity.

Building model

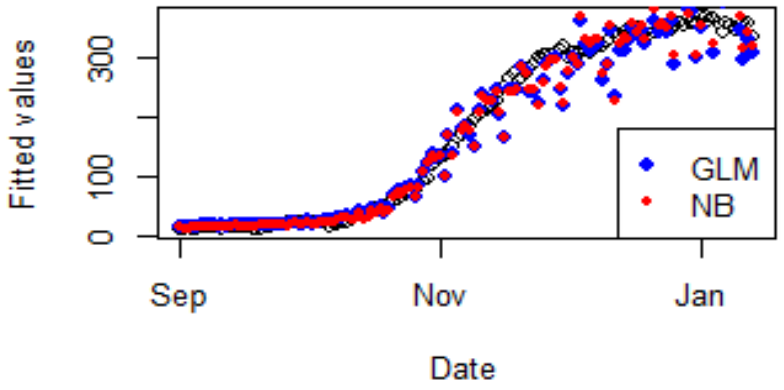
Models comparison

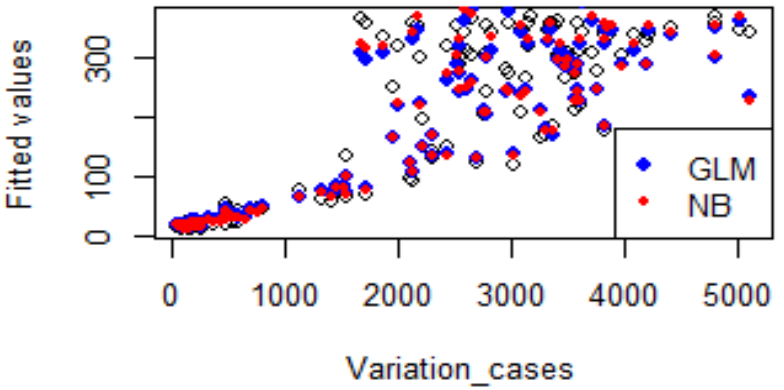
- NB seems to be the best model as it has lower AIC and also BIC.
- AIC and BIC are not available for Quasi-Poisson because it doesn't use the distribution function of the data and therefore no log-likelihood function.

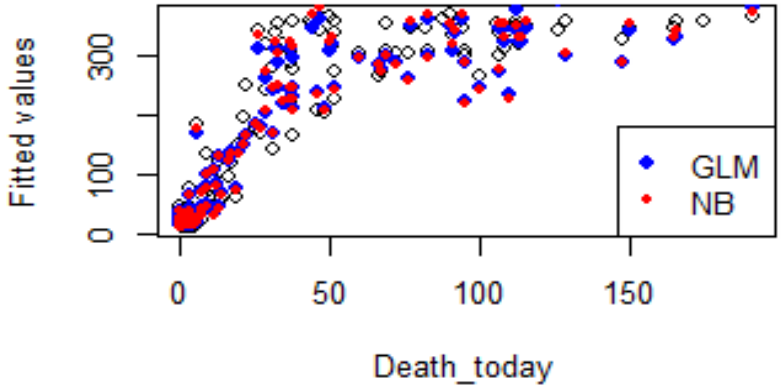
Model	df	AIC	BIC
Poisson	8	1469.58	1492.83
Quasi Poisson	8	NA	NA
Negative Binomial (NB)	9	1169.13	1195.28
GAM	20	1203.40	1262.22

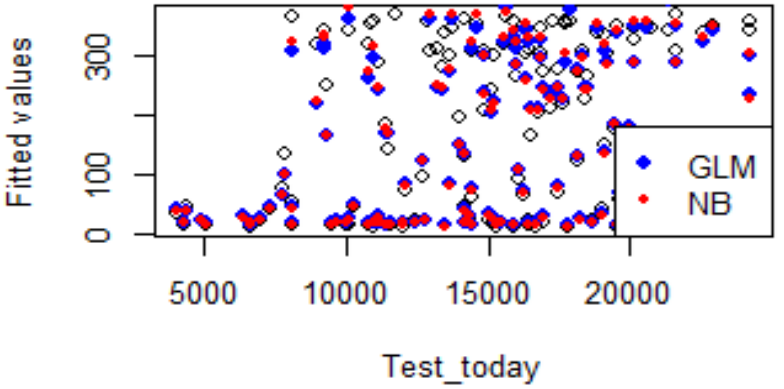
Building model

Fitted values for NB model









Building model

Adding new covariates

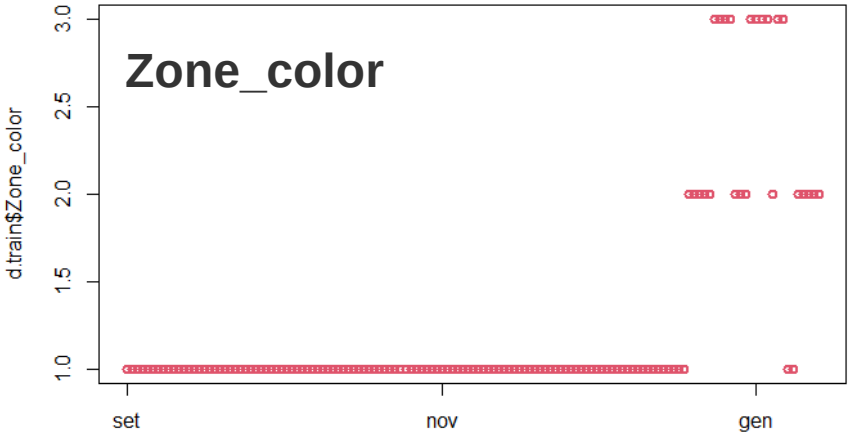
- By building the models, we evaluated the inclusion of some **covariates** and their effect on the response variable *Intensive_care*.
- We decided to include in the models:
  - The regional colors (yellow = 1, orange = 2, red = 3), to which is associated a set of laws that regulate the permitted and forbidden behaviours, defined as **Zone_color**;
  - The 7-days delay on the *Zone_color* variable defined as **Lag_zone_color** that is usually needed to see some effect on the data;
  - The **Season** categorical variable that takes into account the seasonal effect on the spreading of respiratory diseases like Covid-19.

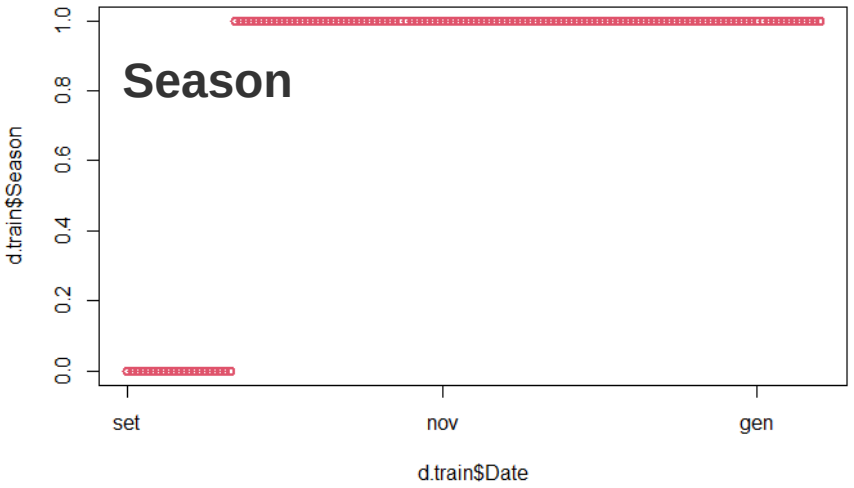
- We decided not to take into account some additional considered covariates like the region-specific laws contained in the regional edict "*Decreto legge*".
  - It seems probably strictly correlated to *Zone_color* covariate that we already introduced; although surely these regional laws have had an impact on the displacement of large groups of people (for example we assisted to a larger migration in the beginning of December and a very reduced one during the holidays, when it was not allowed to exit the Municipality borders except for rare cases).

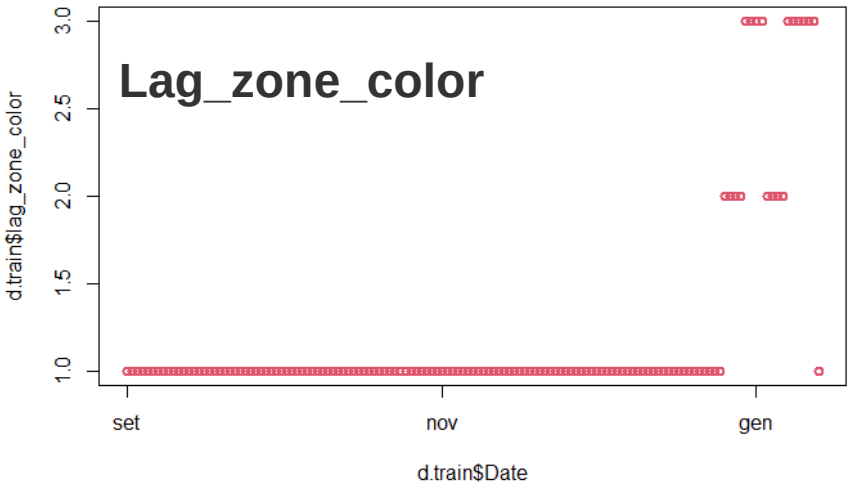
Building model

Adding new covariates

- **Zone_color**
  - 1 for yellow
  - 2 for orange
  - 3 for red
- **Lag_zone_color**: 7 days lag
- **Season**: categorical
  - 0 for Summer
  - 1 for Autumn and Winter







Building model

Adding new covariates

- All the three new variables can be added to our Poisson model as they **decrease AIC and BIC**, also there is no any considerable collinearity between them and old variables based on VIF, and all of them are significant in the model as well.
- No improvement with **interaction** terms.

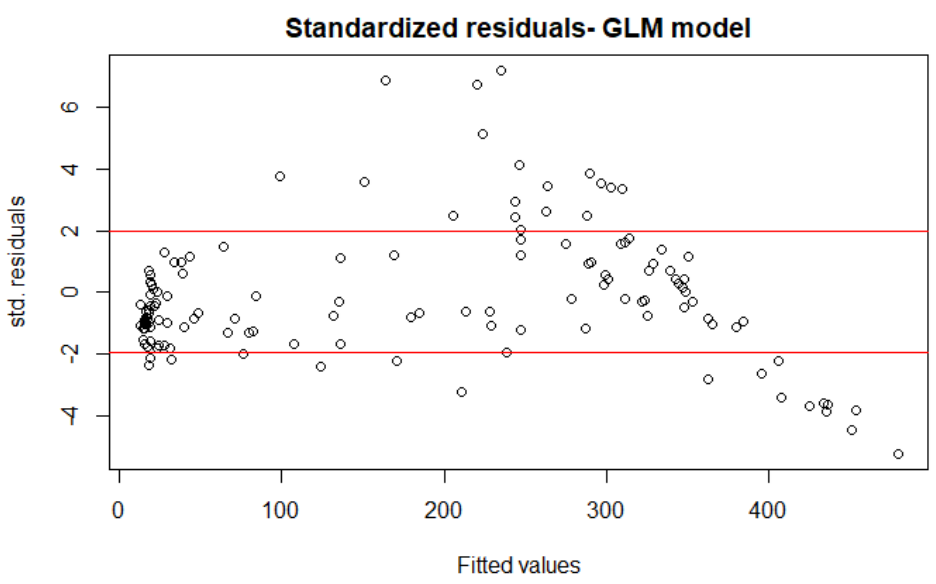
Model	AIC	BIC	RD	VIF
Original GLM Poisson model (from slide 19)	1469.58	1492.83	593.47	
Adding *Zone_color*	1428.49	1454.64	550.38	ok
Adding *Lag_zone_color*	1377.68	1406.73	497.56	ok
Adding *Season*	1351.23	1383.19	469.11	ok

Building model

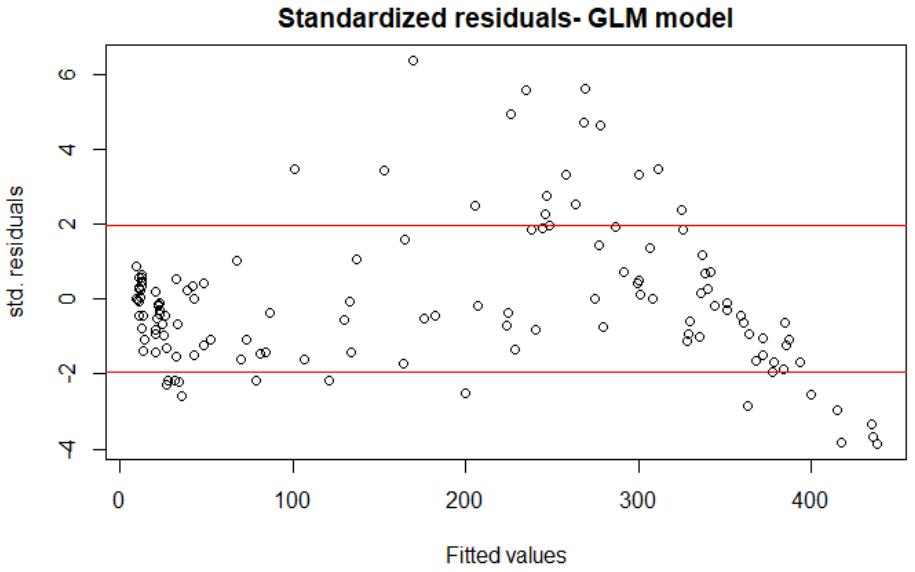
Poisson model- with new covariates

We go through **all the previous steps** to build our models including new variables. So we don't repeat all of them again. We just show that our new Poisson model seems to be overdispersed according to the residual plot.
The point is that comparing the two residual plots before and after adding new covariates, turns out that **we managed to alleviate the tail of residual plot to some extent.**

Before adding new covariates



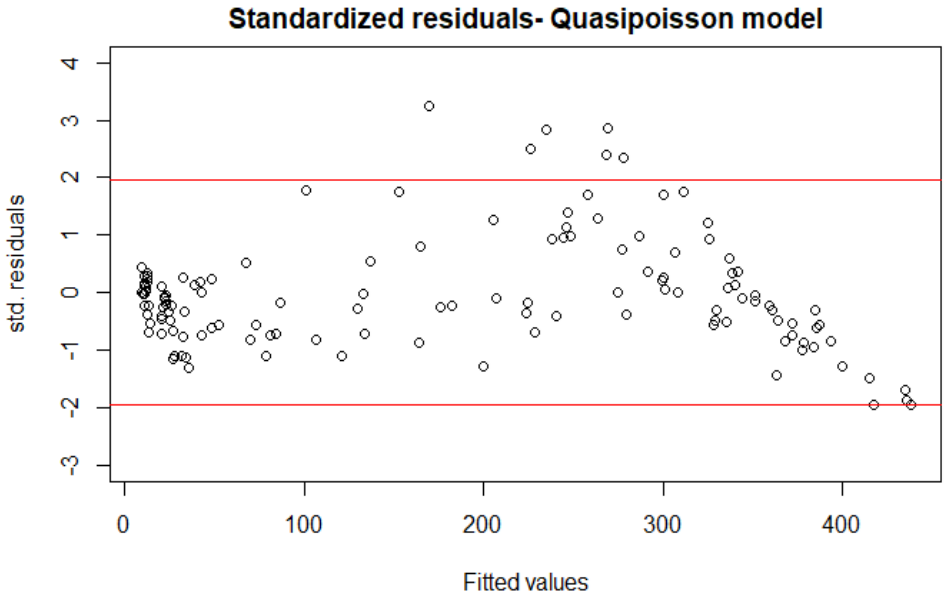
After adding new covariates

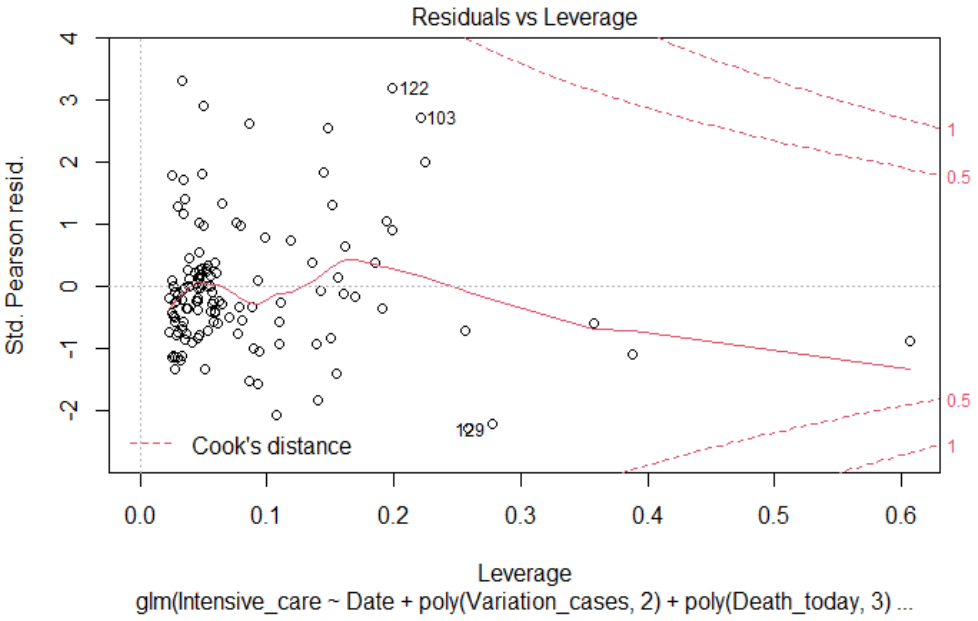


Building model

Quasi-Poisson model- with new covariates

The dispersion parameter is 3.87 >1 .





Model comparison

After adding new covariates

- **GLM**
  - **Poisson**
  - **Quasi Poisson**
  - **Negative Binomial**
- **GAM**
- **Random Forest**

Predictors
Date
Var_cases
Death_today
Test_today
Zone_color
Lag_zone_color
Season

Model	df	AIC	BIC
Poisson	11	1351.23	1383.19
Quasi Poisson	11	NA	NA
Negative Binomial (NB)	12	1148.19	1183.05
GAM	20	1162.13	1222.06

AIC and BIC for all the models are smaller compared to the previous ones without adding new variables. It means we also managed to improve our models by adding new variable.

The table shows NB is the best model again according to AIC.

Prediction

Predictive information criteria

- **MSE (mean-square error)**: measures the average squared difference between the estimated values and the actual value. It is always positive, and values closer to zero are better.

$$MSE = \frac{1}{n}\sum_i (\hat{Y}_i - Y_i)^2$$

- **RMSE (root-mean-square error)**: square root of the MSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_i (\hat{Y}_i - Y_i)^2}$$

- **NRMSE (normalized-root-mean-square error)**: normalizing the RMSD facilitates the comparison between datasets or models with different scales.

$$NRMSE = \frac{RMSE}{Y_{max} - Y_{min}} \; or \; NRMSE = \frac{RMSE}{\bar{Y}}$$

Prediction

On test dataset from 14th to 23rd Jan

- Model comparison **without and with new covariates**

Models(without)	MSE	RMSE	NRMSE
Poisson	27883.18	166.98	**0.53**
Quasi Poisson	27883.18	166.98	**0.53**
Negative Binomial	29539.56	171.87	**0.54**
GAM	29260.05	171.06	**0.54**
RF	4055.92	63.69	**0.20**

Models(with)	MSE	RMSE	NRMSE
Poisson	16466.95	128.32	**0.40**
Quasi Poisson	16466.95	128.32	**0.40**
Negative Binomial	20399.62	142.83	**0.45**
GAM	20124.960	141.86	**0.45**
RF	456.98	21.38	**0.07**

Prediction

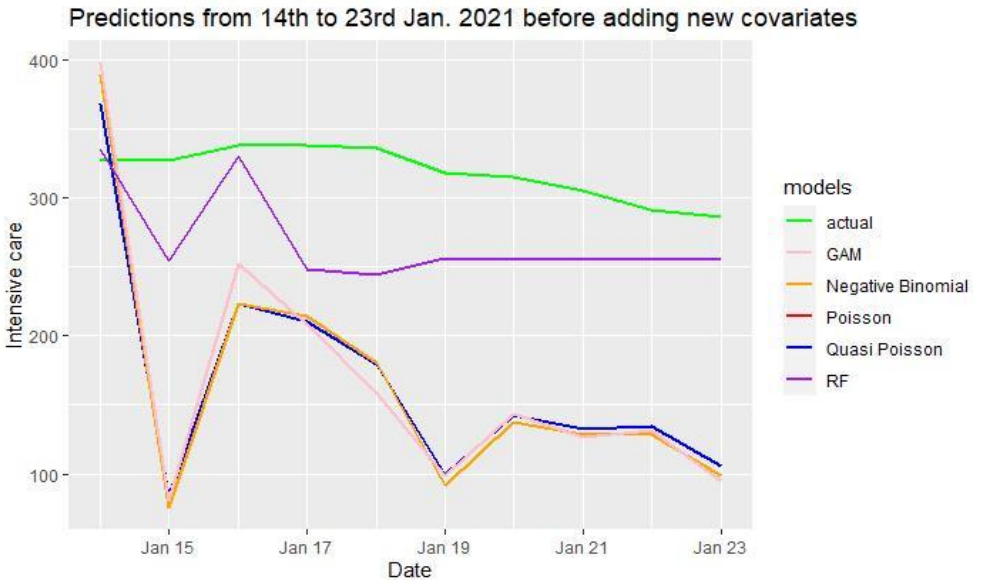
On test dataset

- After adding new variables the error rate decreases in all the models.
- Random forest has the lowest error rate.
- There is no considerable difference between other models; however Poisson is a bit better than NB for this test dataset.

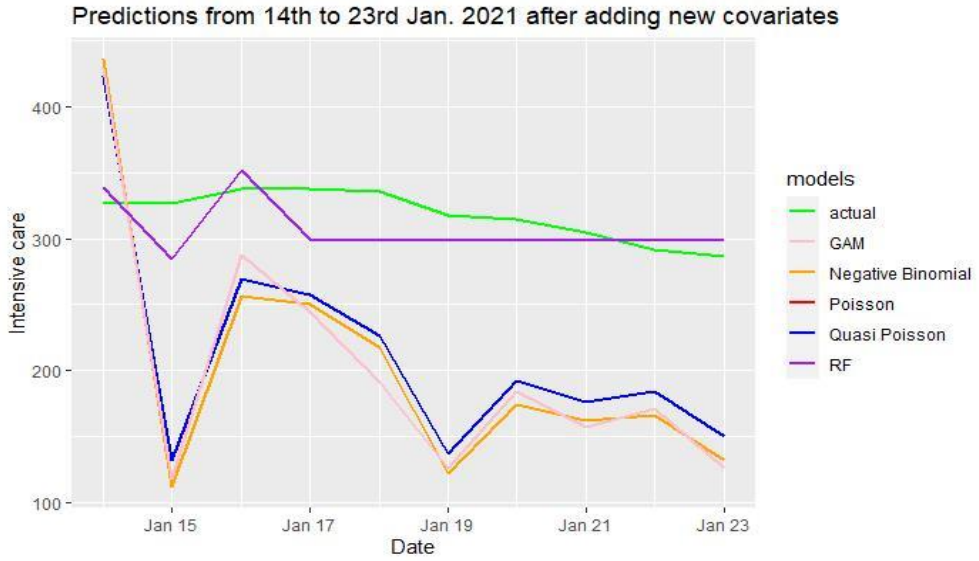
Prediction

On test dataset

- The **first** plot shows the predictions for 10 days of January, from 14th to 23th as the test dataset, by different models <u>without</u> the adding of the covariates.



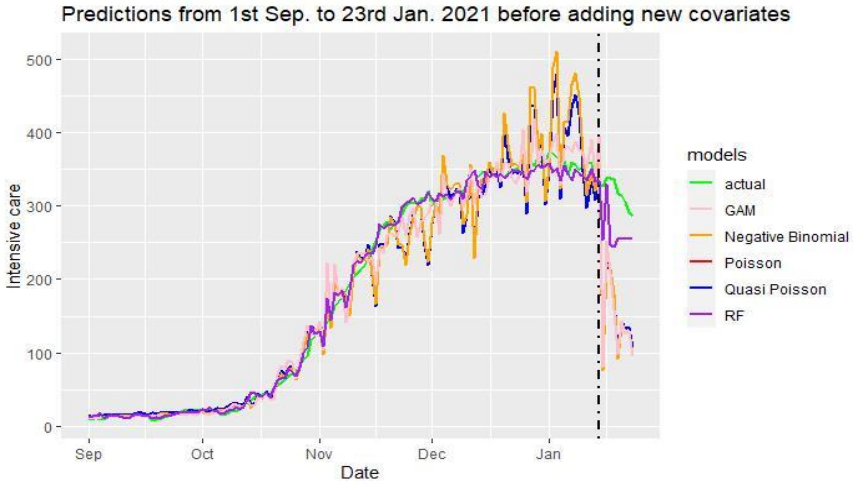
- The **second** plot shows how the models change after adding all the previously introduced **covariates**.

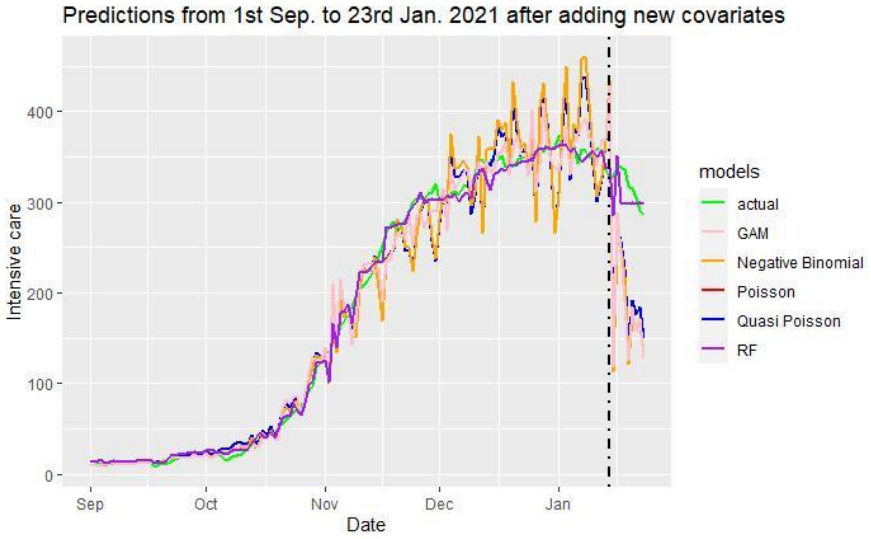


Prediction

The fit on all the dataset

- The following plots show the predictions for all the dataset (after dashed line for test dataset).
- Excessive distortion in predictions from the middle of December afterwards, could be due to significant difference in data compared to the previous 3 months that play a remarkable role in building the model.





Extra approach

Predicting on shifted/historical data

Why this extra approach?

Not being able to say with certainty number of deaths in the span of 14 days, the idea came to use past data to predict future.
The current time (t) and future times (t+1, t+n) are forecast times and past observations (t-1, t-n) are used to make forecasts.

> *Sequence prediction attempts to predict elements of a sequence on the basis of the preceding elements*
>
> — Sequence Learning: From Recognition and Prediction to Sequential Decision Making, 2001.

Extra approach

Shifting the dataset

- We could frame our forecast problem with an input sequence of 7 past observations to forecast 7 future observations and use the data as follows:

date_only <date>	terapia_intensiva <int>	deceduti <int>	lag_deceduti <int>
2020-09-01	9	2122	2107
2020-09-02	9	2123	2116
2020-09-03	12	2123	2117
2020-09-04	10	2126	2119
2020-09-05	9	2130	2120
2020-09-06	12	2130	2120
2020-09-07	13	[illegible]	2120
2020-09-08	12	[illegible]	2122
2020-09-09	12	2133	2123
2020-09-10	12	2135	2123

TRAIN

Previous 1 2 3 4 5 6 ... 14 Next

date_only <date>	terapia_intensiva <int>	lag_deceduti <int>
2021-01-11	358	6813
2021-01-12	359	6988
2021-01-13	336	7114
2021-01-14	328	7157
2021-01-15	327	7263
2021-01-16	338	7345
2021-01-17	338	7389
2021-01-18	336	7427
2021-01-19	318	7593
2021-01-20	315	7684

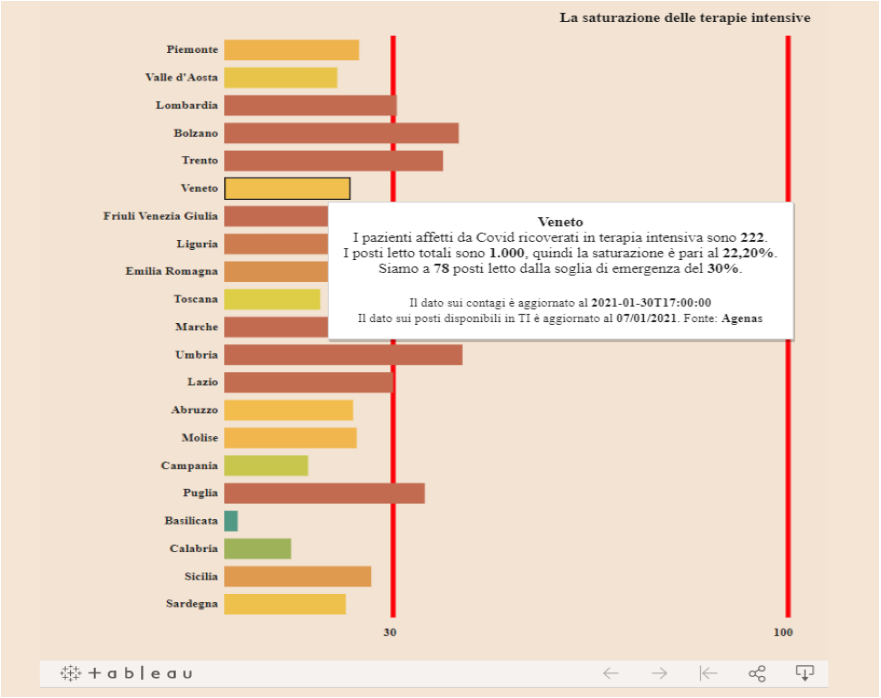
TEST

Previous 1 2 Next

Extra approach

The added covariate *percentage occupancy*

- Added a feature indicating the percentage from the total available ICU beds for the individual days
- Did the hospitals make decisions who gets received/who stays in ICU based on this threshold? Based on this doubt, we state that the **independence assumption is no longer valid**, between the individual response variables.
  - The threshold to be around is 30% (of 1016 beds in Veneto).



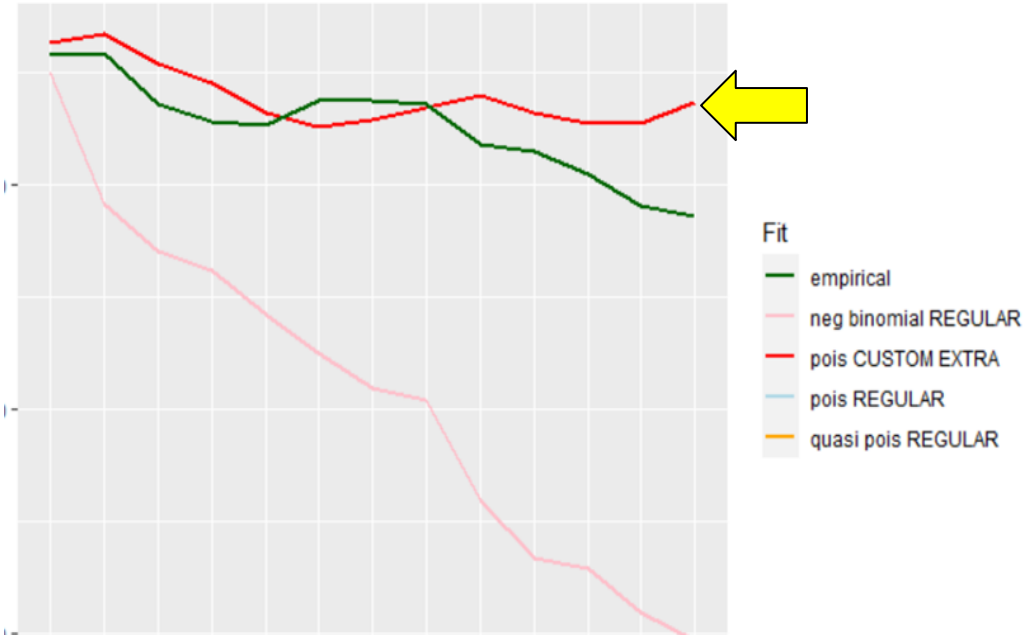
NOTE This metric is one of the predictors for a custom “new” model developed for the shifted approach itself

Extra approach

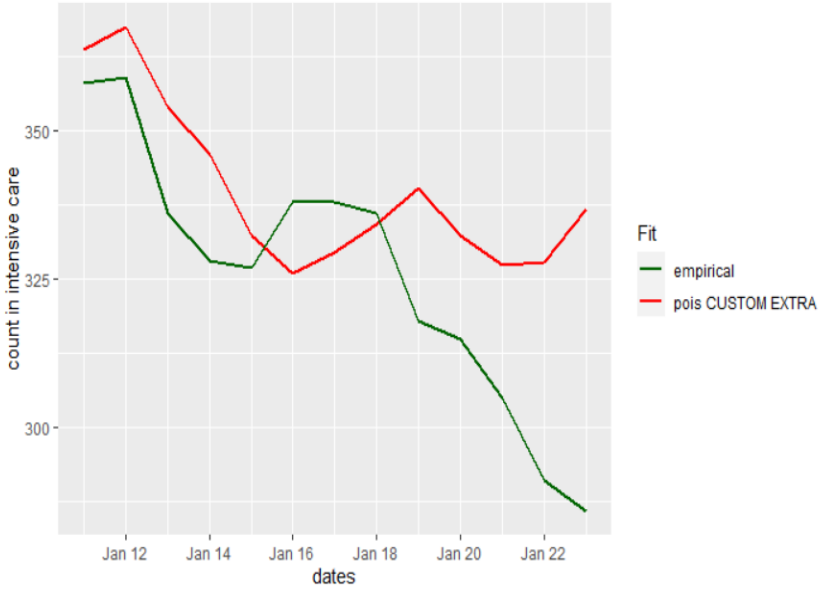
Performance with ultimate chosen model

Well-suited Predictors
Date
Var_cases
Death_today
Test_today
Zone
Lag_zone
Season

- Shifted data + **proved well-suited model**
- Trying the approach with the covariates and models that worked best for the original dataset shows quite good results in AIC scores, see further slides



NOTE: GAM REGULAR MISSING HERE, COMPUTATIONAL DIFFICULTIES

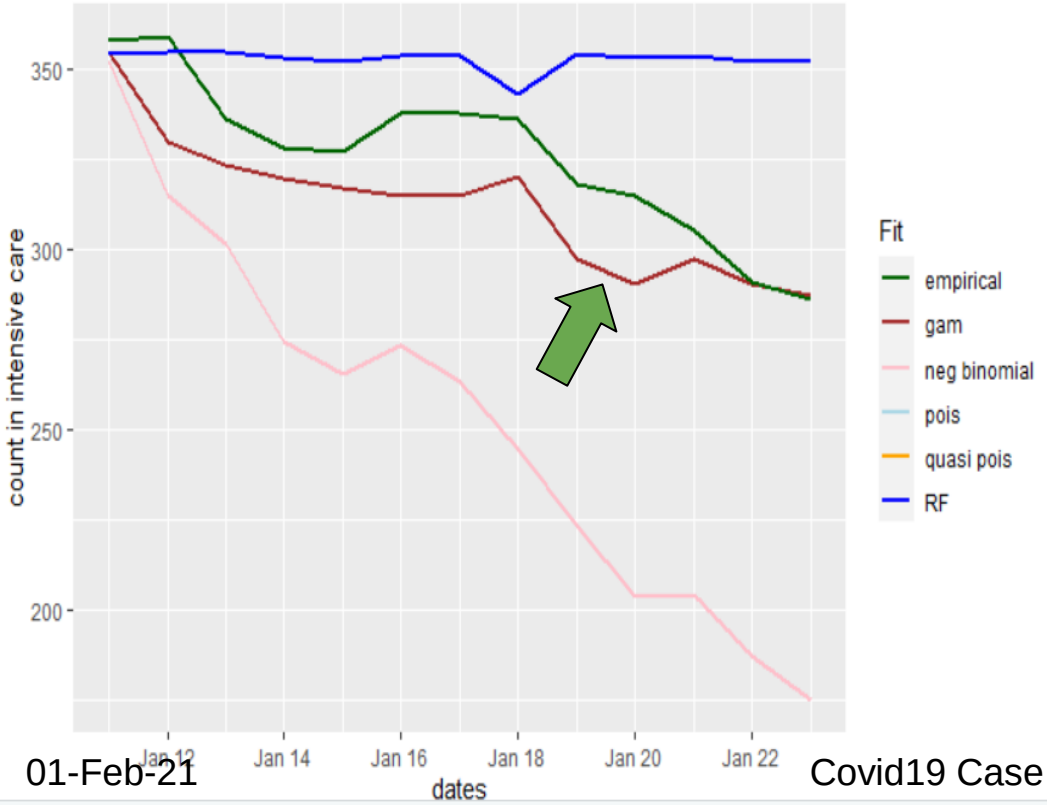


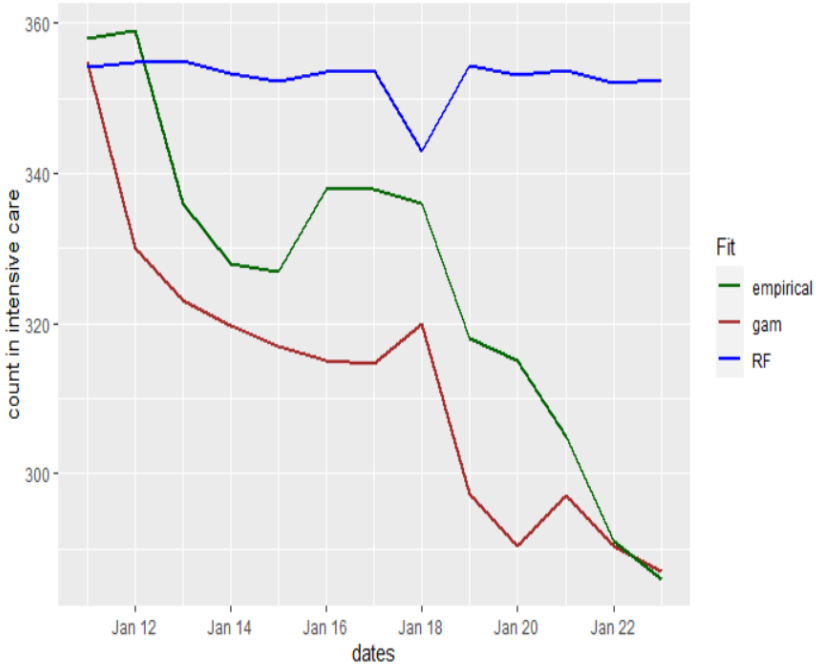
Extra approach

Performance with penultimate chosen model

- Shifted data + **penultimate proved well-suited model**
- Trying the approach with the penultimate set of covariates, indicates a promising solution with a gam model

Possible Predictors
Date
Var_cases
Hosp_symp
Death_today
Zone
Lag_zone



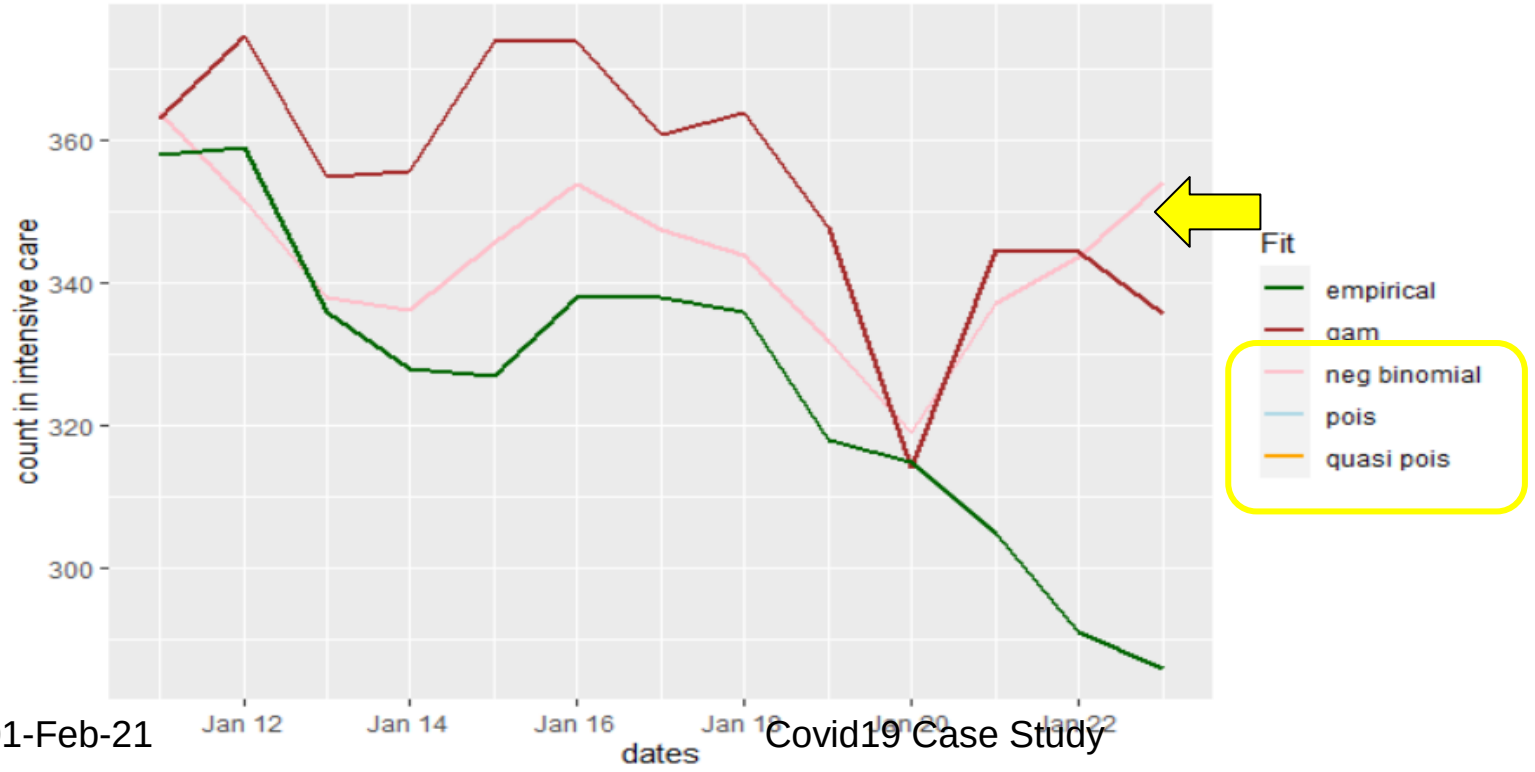


Extra approach

Performance with a new model

Chosen Predictors
Date
Moving avg Death
Moving avg Total_Hosp
Variation Total Hosp
Percentage Occupied

- Shifted by 14 the data + **new model**
- NOTE: Two methodologies for selecting predictors: backward selection & “it worked best”; here showing the version that had the best prediction when **adding least correlated covariates**, observing the **least residual deviance** + the conviction that **the percentage** influences decisions on accepted patients for ICU



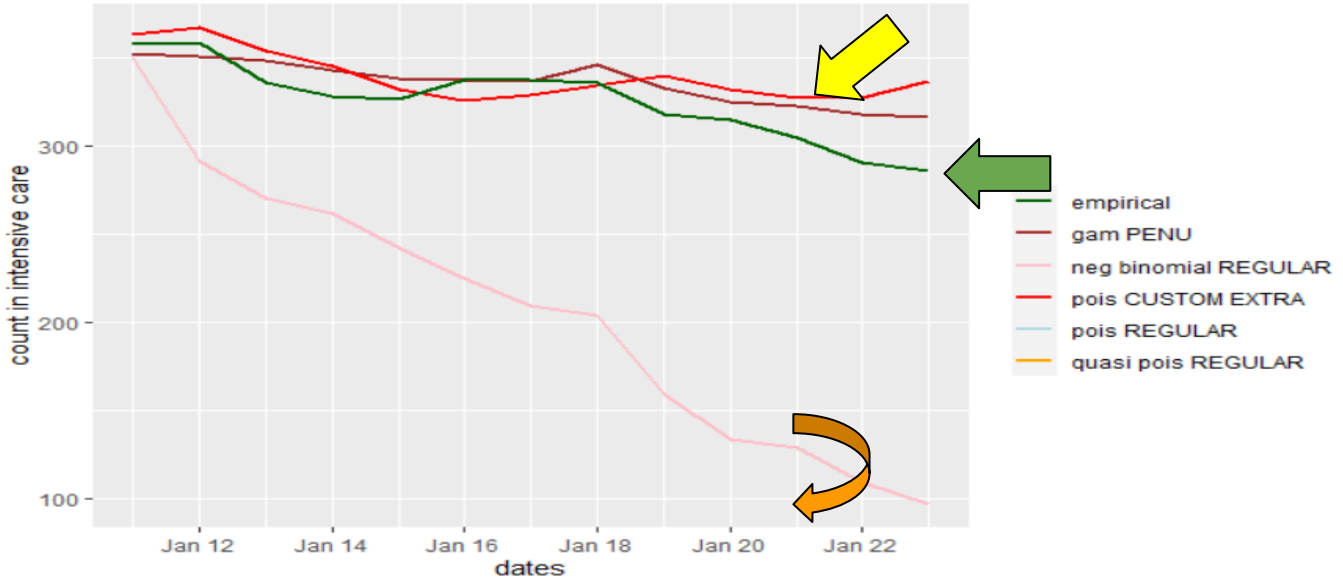
Extra approach

Comparing the fit and predictions graph

- AIC comparison of the models

	df <dbl>	AIC <dbl>
poisson_extra.model	6.00000	875.4858
gam_extra.model	10.00000	911.8659
quasipoisson_r.model	11.00000	*NA*
negbin_r.model	12.00000	854.5101
gam_penu.model	19.21097	798.1246
gam_r.model	20.03568	800.0119

6 rows



Extra approach

Comparing predictions - metrics

- Model comparison when using shifted/historical data from **10th to 23rd January**.

models	MSE	RMSE	NRMSE
Poisson EXTRA	480.4996	21.92030	**0.067**
GAM Penultimate R	234.2662	15.30576	**0.047**

AIC for the used models, together with the metrics MSE, RMSE, NRMSE proved the good fit of the penultimate gam model.
The decreasing trend was caught for January. Predicting using historical data ***gives good predictions***.

Extra approach

Possible next steps

- reduce days to predict
- apply transformations to covariates, to improve their efficiency
- Leverage the analysis from **AR (Autoregressive), ARMA (Autoregressive Moving Average) and ARIMA methods (Autoregressive Integrated Moving Average)** methodologies.

- OR

- We suggest the **Recursive Multi-step Forecast** technique, where the predicted values would be used as input for the successive data points.

- OR

- **Direct Multi-step Forecast Strategy** The direct method involves developing a separate model for each forecast time step.

Possible improvements

To further improve our analysis:

- Use the newly added covariates by the “Protezione Civile”.
- Use 1 month of data as training set.
- Explore better the reasoning behind the peculiar relationship between certain metrics of the dataset.
- Take data from e.g. Oct/Nov when new algorithms for gathering data were introduced (government rules).

References

Course books:

- https://moodle2.units.it/pluginfile.php/340058/mod_resource/content/1/core-statistics.pdf
- https://moodle2.units.it/pluginfile.php/340059/mod_resource/content/1/Data-Analysis-and-Graphics-Using-R-An-Example-Based-Approach-Cambridge-Series-in-Statistical-and-Probabilistic-Mathematics-.pdf

Resources from the internet:

- Rapporto Covid-19
- Intensive care management of coronavirus disease 2019 (COVID-19): challenges and recommendations
- Generalized linear models. Introduction to advanced statistical… | by Yuho Kida
- Terapie intensive, scopri (in tempo reale) quanti posti sono occupati - Info Data
- https://www.ilmessaggero.it/salute/focus/terapia_intensiva_covid_rianimazione_ospedali_precedenza_a_chi_puo_sopravvivere_iss_ricoveri_criteri_news-5600322.html
- Interpreting Generalized Linear Models
- Certain terminology gotten from: 171 Responses to 4 Strategies for Multi-Step Time Series Forecasting

- SMDS Course slides and labs