Final project

Statistical Methods for Data Science

Mariani, Marturini, Tinto

Università degli Studi di Trieste

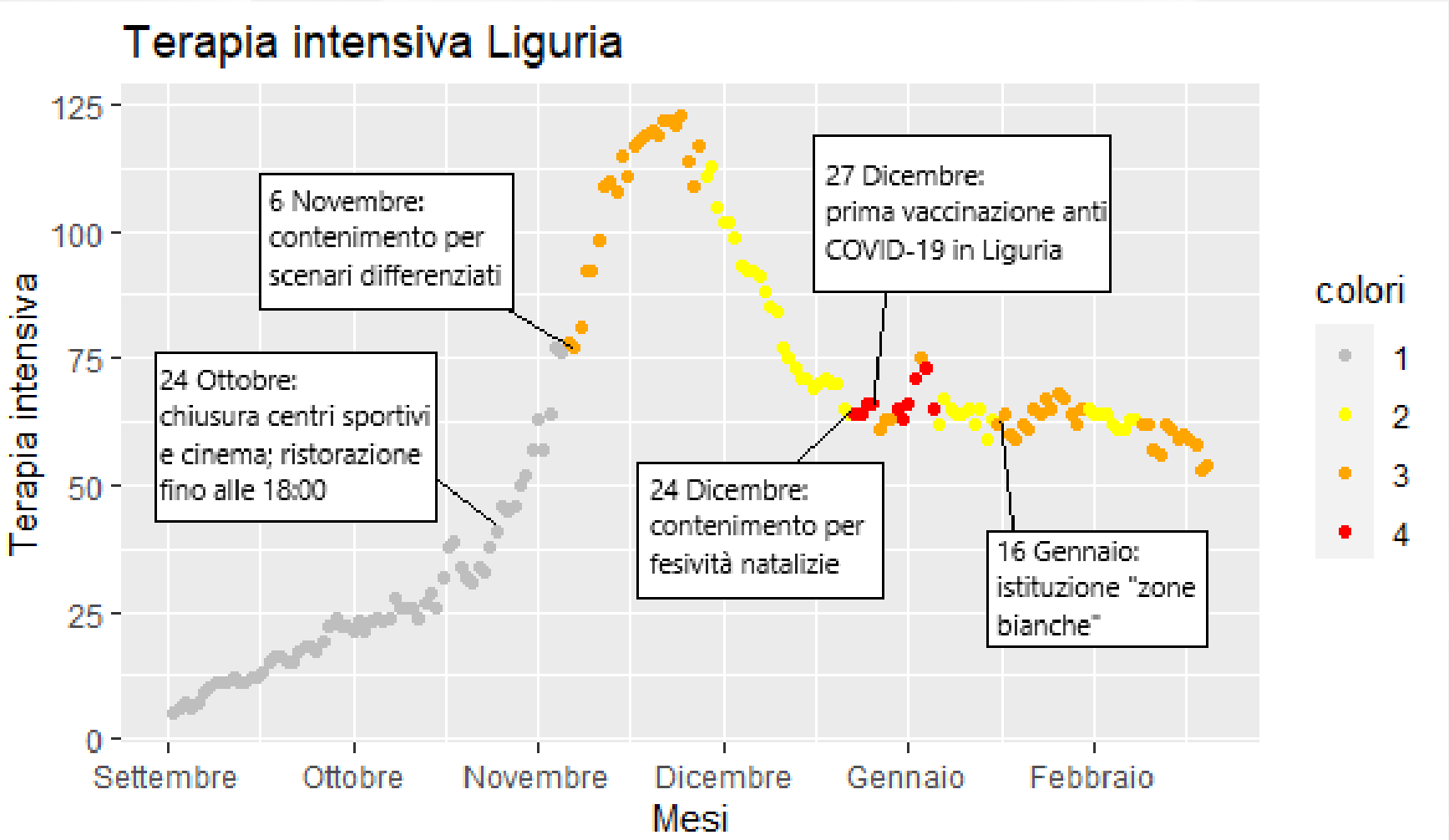
Introduction

Goal

Statistical modeling of the intensive care units (ICU) due to COVID-19 in the region Liguria, from September 2020 to February 2021

- description of the dataset
- investigation of explanatory variables
- evaluation and comparison of alternative models
- predictions and possible improvements

The dataset overview



The dataset quality

variables	type
data	Date
stato	factor
codice_regione	factor
denominazione_regione	factor
lat	numeric
long	numeric
ricoverati_con_sintomi	integer
terapia_intensiva	integer
totale_ospedalizzati	integer
isolamento_domiciliare	integer
totale_positivi	integer
variazione_totale_positivi	integer
nuovi_positivi	integer
dimessi_guariti	integer
deceduti	integer
casi_da_sospetto_diagnostico	integer
casi_da_screening	integer
totale_casi	integer
tamponi	integer
casi_testati	integer
note	character
ingressi_terapia_intensiva	integer
note_test	character
note_casi	character
totale_positivi_test_molecolare	integer
totale_positivi_test_antigenico_rapido	integer
tamponi_test_molecolare	integer
tamponi_test_antigenico_rapido	integer
codice_nuts_1	character
codice_nuts_2	character

source: Protezione civile (https://github.com/pcm-dpc/COVID-19)

Discarding variables:

- not applicable
- not significant
- multicollinearity

The dataset missing values

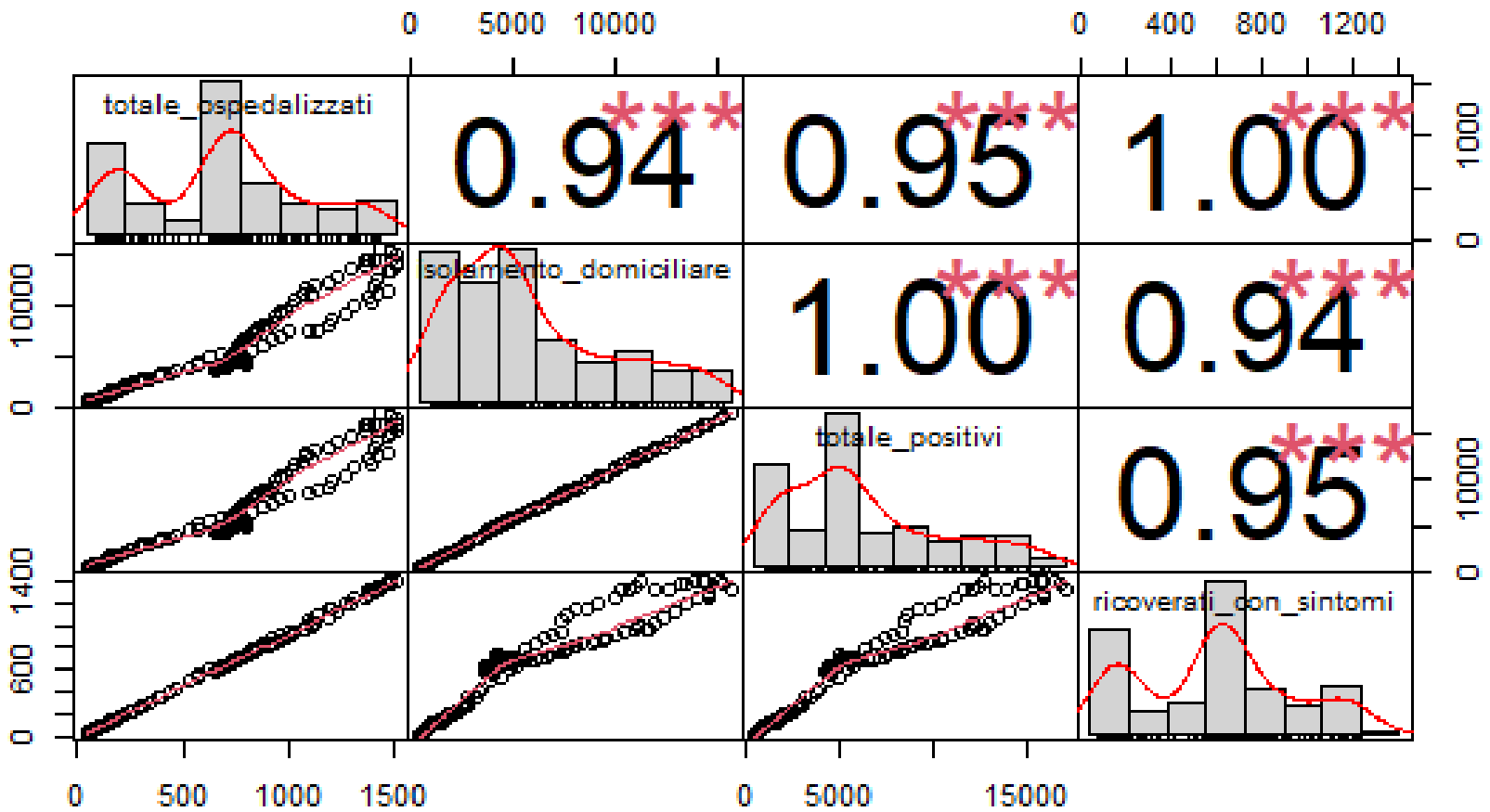
Missing values

We discarded the following independent variables due to an high number of missing values

Variables	missing (n)	missing (%)
totale positivi test molecolare	136	84
totale positivi test antigenico rapido	136	84
tamponi test molecolare	136	84
tampni test antigenico rapido	136	84
ingressi terapia intensiva	93	57.4
casi da sospetto diagnostico	68	42
casi da screening	68	42

The dataset multicollinearity

An example of multicollinearity between predictors



The dataset predictors

Considered predictors

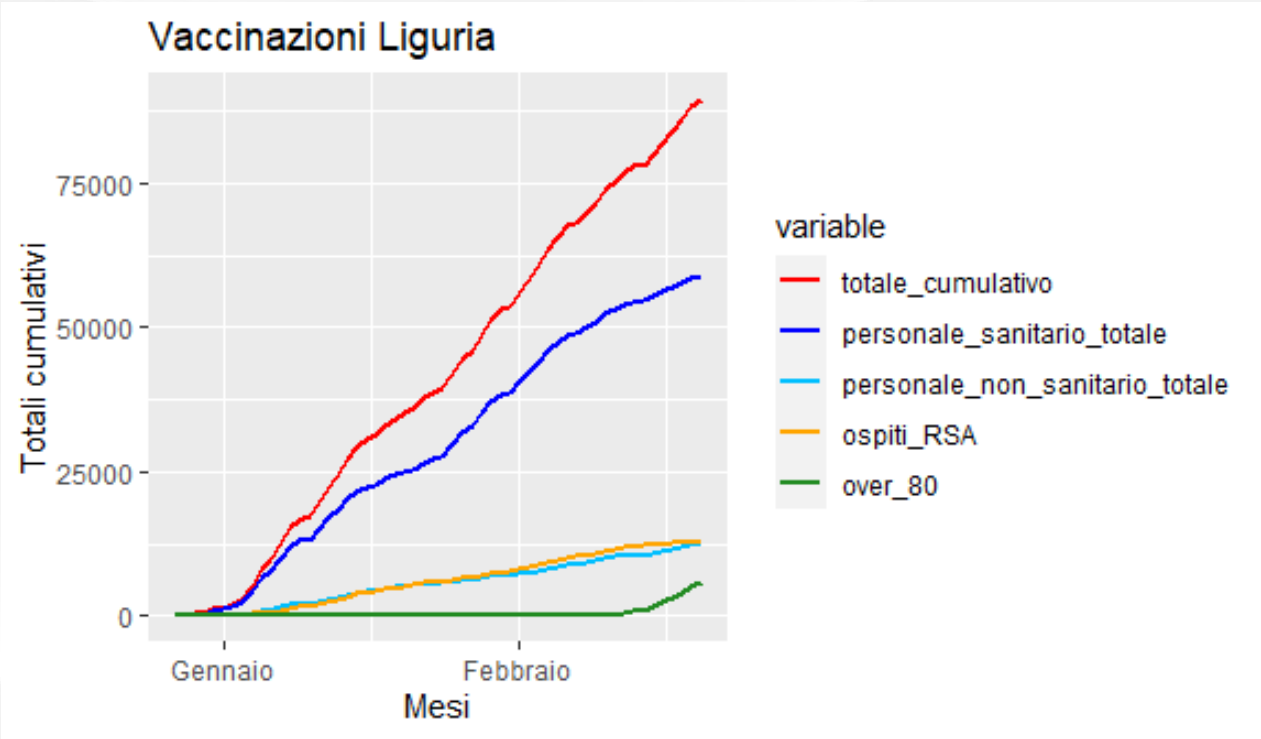
- totale_positivi
- variazione_totale_positivi
- ricoverati_con_sintomi
- deceduti_giornaliero
- dimessi_guariti_giornaliero

Added predictors

- colori
- vaccinazioni

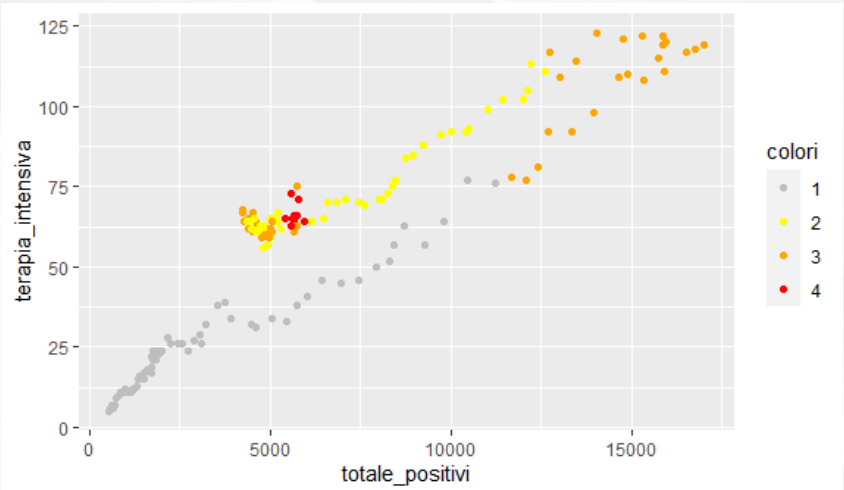
Vaccinations

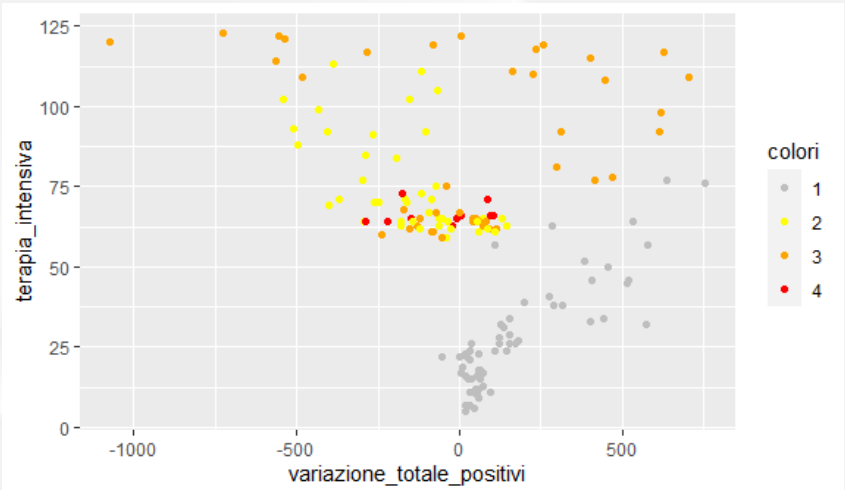
index	integer
area	string
data_somministrazione	datetime
totale	integer
sesso_maschile	integer
sesso_femminile	integer
categoria_operatori_sanitari_sociosanitari	integer
categoria_personale_non_sanitario	integer
categoria_ospiti_rsa	integer
categoria_over80	integer
prima_dose	integer
seconda_dose	integer
codice_NUTS1	string
codice_NUTS2	string
codice_regione_ISTAT	integer
nome_regione	string

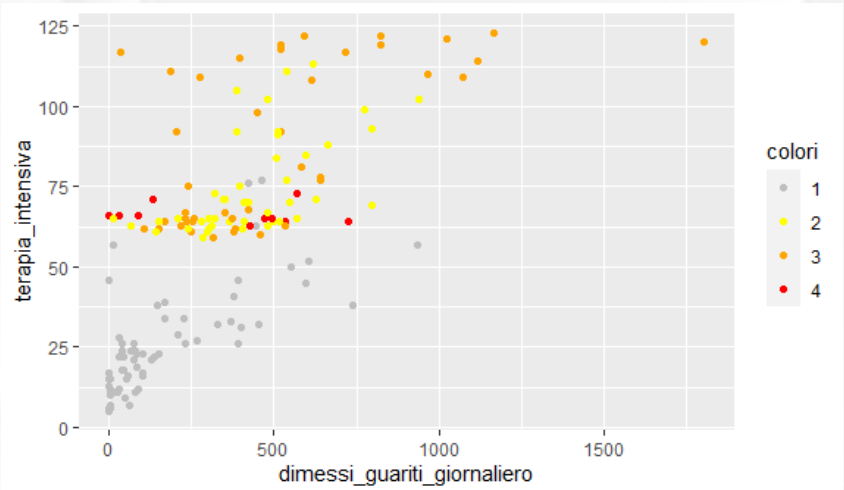


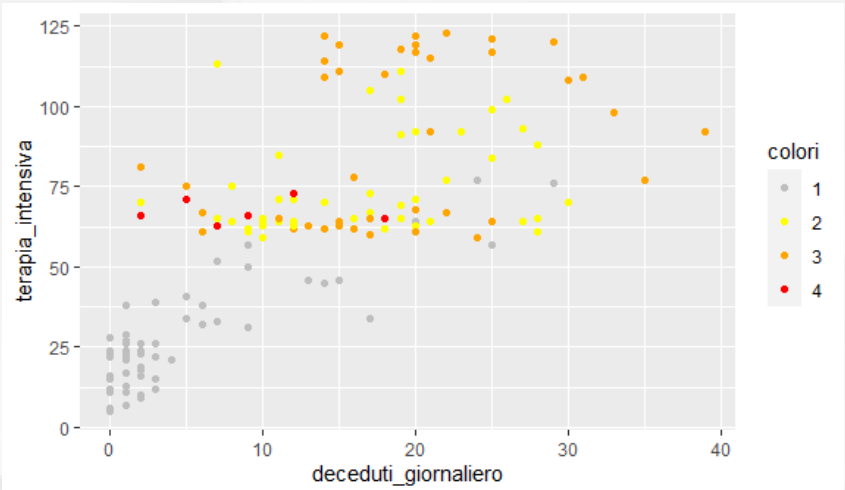
source: https://www.governo.it/it/cscovid19/report-vaccini/

Terapia intensiva vs predictors









Linear model

Linear model

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{ip-1}$$

Assumptions

- Normality
- Homoscedasticity
- Independence

Linear model

```
lm.model1 <- lm(covid.liguria, formula = terapia_intensiva ~ totale_positivi + colori)
lm.model2 <- lm(covid.liguria, formula = terapia_intensiva ~ totale_positivi + colori + variazione_totale_positivi)
lm.model3 <- lm(covid.liguria, formula = terapia_intensiva ~ totale_positivi + colori + variazione_totale_positivi + dimessi_guariti_giornaliero)
lm.model4 <- lm(covid.liguria, formula = terapia_intensiva ~ totale_positivi + colori + variazione_totale_positivi + deceduti_giornaliero)
lm.model5 <- lm(covid.liguria, formula = terapia_intensiva ~ totale_positivi + colori + variazione_totale_positivi + totale_vaccini)
```

- dimessi_guariti_giornaliero and deceduti_giornaliero not significant

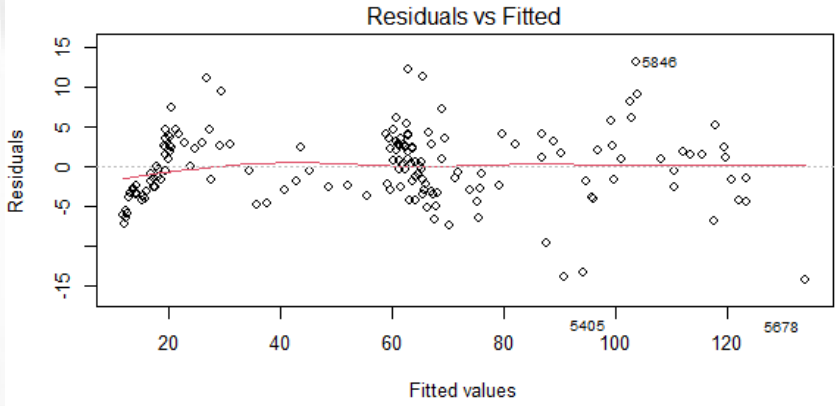
- anova(lm.model1, lm.model2, lm.model5)

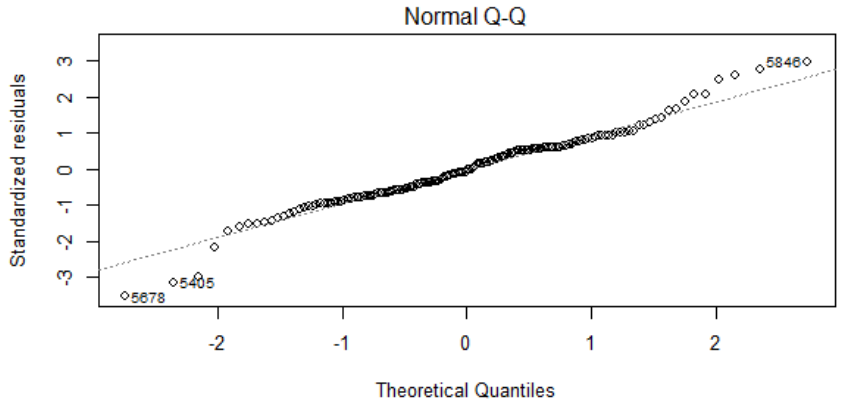
```
Analysis of Variance Table

Model 1: terapia_intensiva ~ totale_positivi + colori
Model 2: terapia_intensiva ~ totale_positivi + colori + variazione_totale_positivi
Model 3: terapia_intensiva ~ totale_positivi + colori + variazione_totale_positivi +
    totale_vaccini
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    157 5190.7
2    156 4622.4  1    568.21 27.879 4.299e-07 ***
3    155 3159.2  1   1463.28 71.794 1.735e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

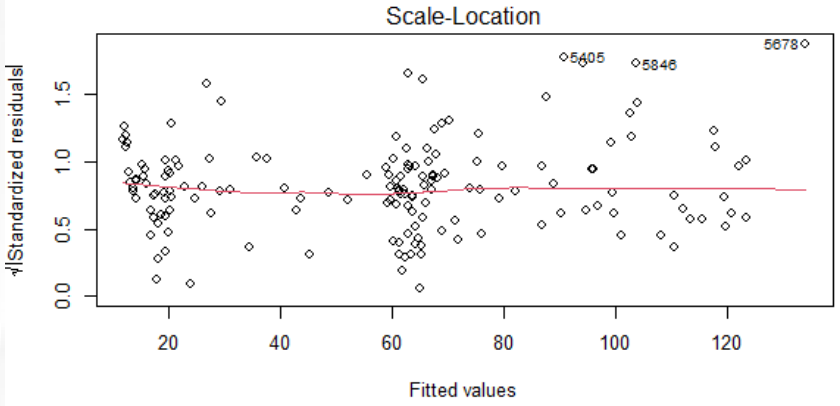
Linear model summary

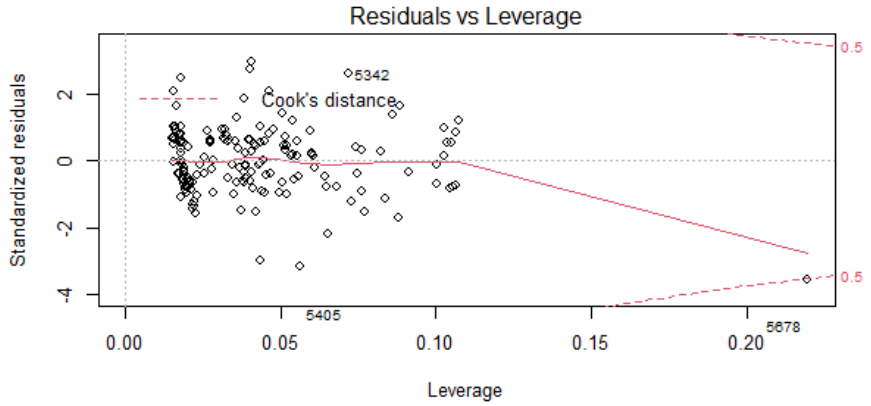
```
Call:
lm(formula = terapia_intensiva ~ totale_positivi + colori + variazione_totale_positivi +
    totale_vaccini, data = raw.data)

Residuals:
     Min       1Q   Median       3Q      Max
-14.0838  -2.8790  -0.1842   2.7266  13.2252

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               8.835e+00  7.224e-01  12.230  < 2e-16 ***
totale_positivi           6.222e-03  1.559e-04  39.918  < 2e-16 ***
colori2                   1.394e+01  1.677e+00   8.311 4.47e-14 ***
colori3                   1.205e+01  1.861e+00   6.473 1.20e-09 ***
colori4                   1.772e+01  1.809e+00   9.797  < 2e-16 ***
variazione_totale_positivi -1.298e-02  1.596e-03  -8.136 1.23e-13 ***
totale_vaccini            1.931e-04  2.279e-05   8.473 1.74e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.515 on 155 degrees of freedom
Multiple R-squared:  0.9815,    Adjusted R-squared:  0.9808
F-statistic:  1372 on 6 and 155 DF,  p-value: < 2.2e-16
```

Linear regression diagnostic plots

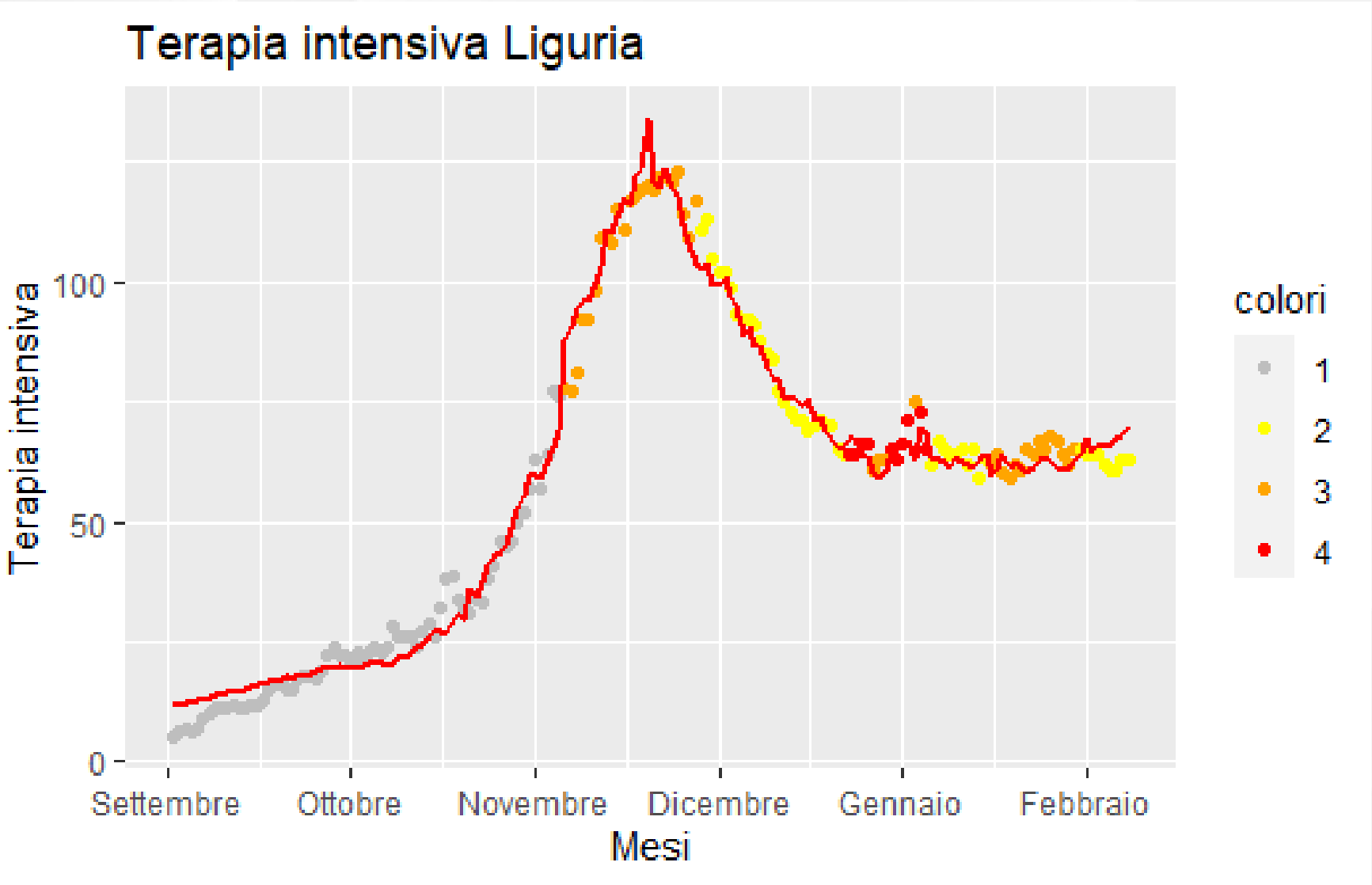




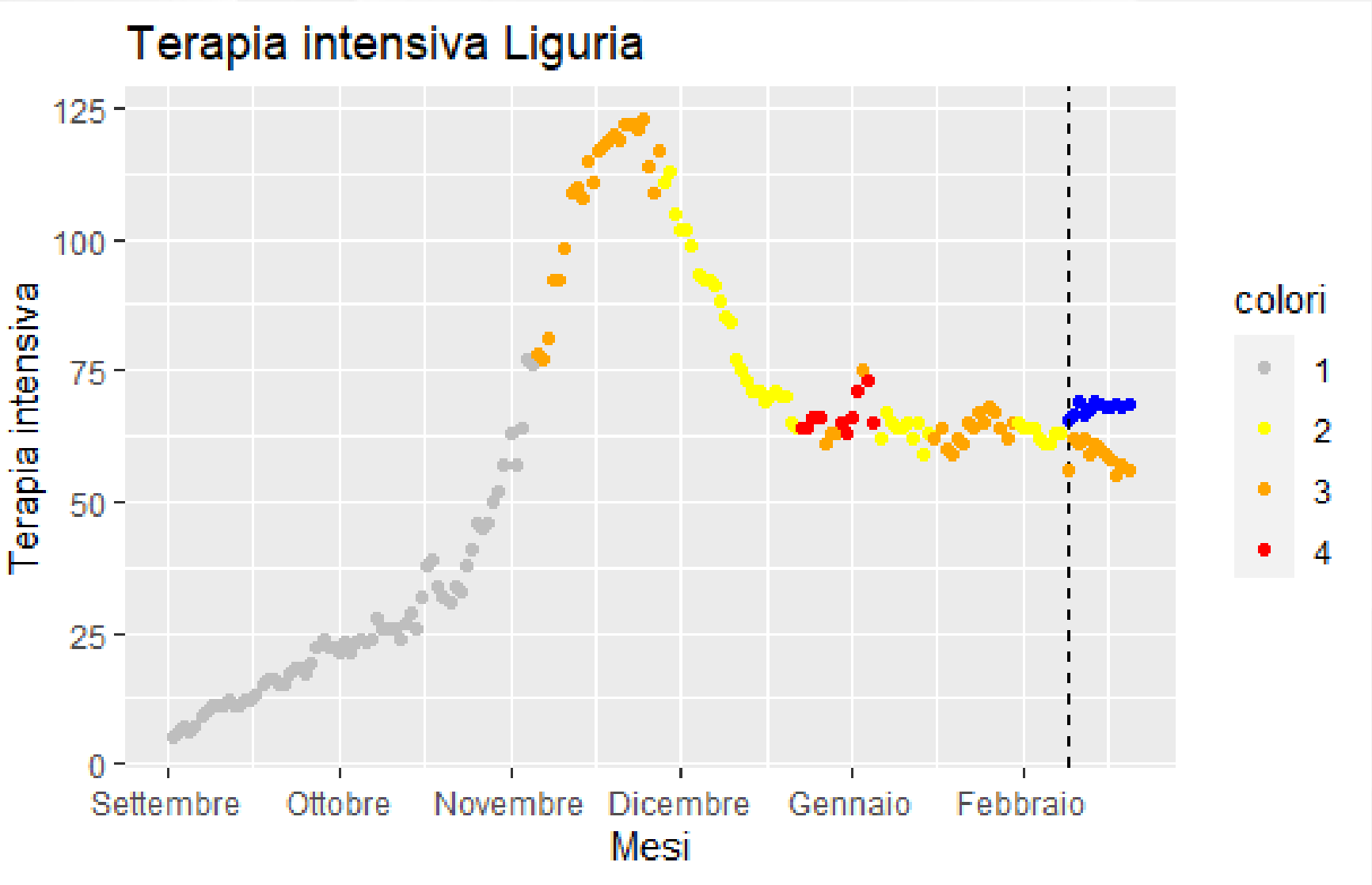




Linear model model fit



Linear model predictions



Linear model problems

Problems

- Correlation between observations: time dependency
- Predictions of negative values
- Not enough vaccines data

Poisson regression

One approach can be using a Poisson regression:

$$Y \sim \text{Poi}(\lambda).$$

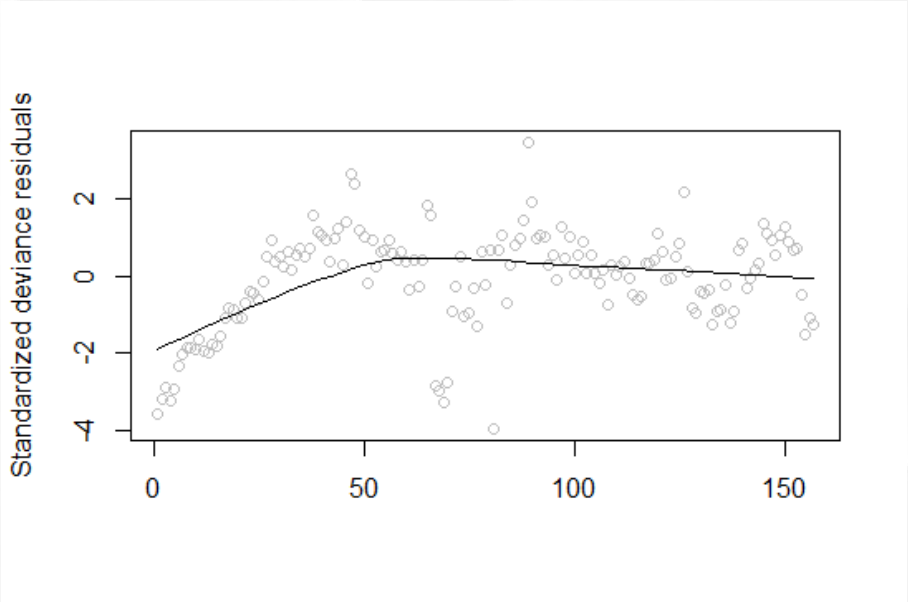
The default link function for Poisson regression is the logarithm:

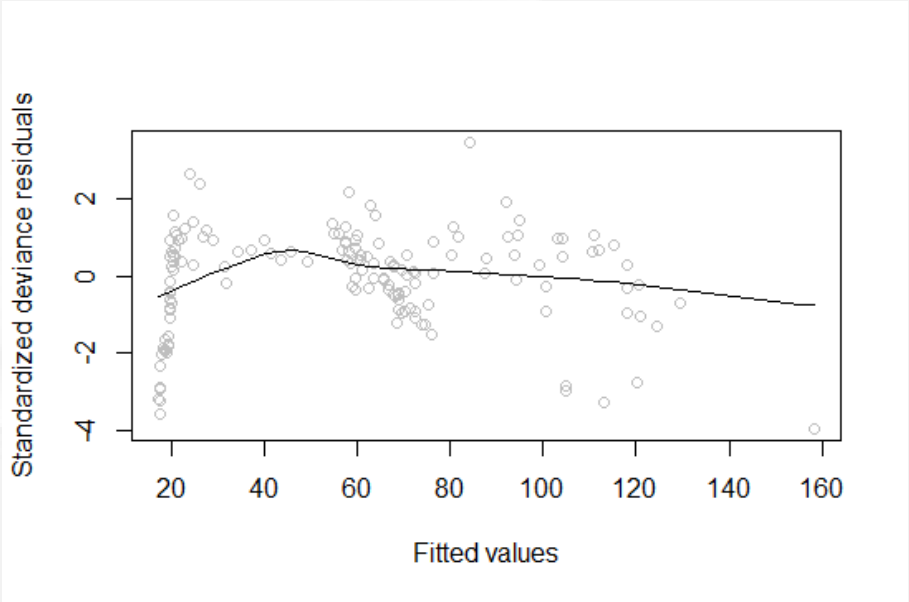
$$\log\big(\text{E}[Y]\big) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4.$$

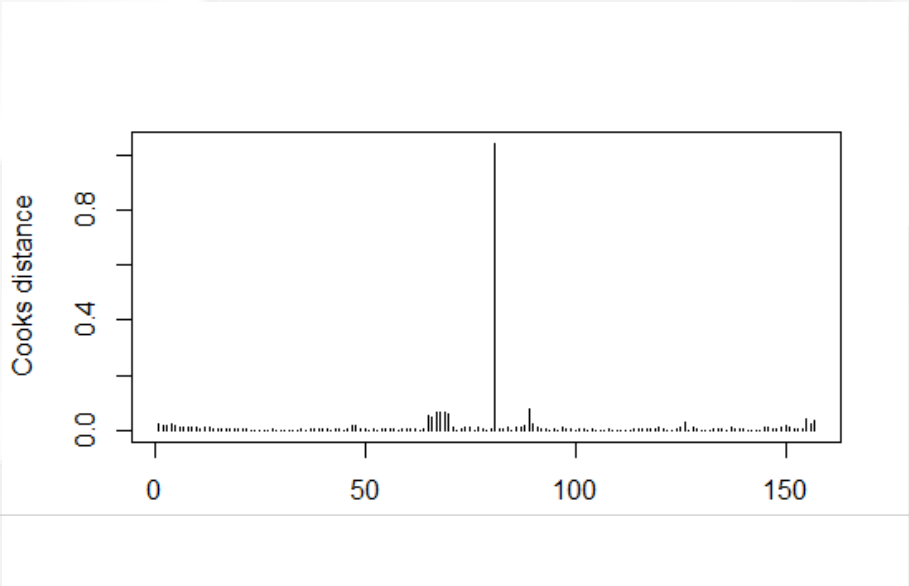
Note: in this model, we will not consider the variables DIMESSI_GUARITI_GIORNALIERO and DECEDUTI_GIORNALIERO (not significant), nor the variable TOTALE_POSITIVI.

```
##
## Call:
## glm(formula = terapia_intensiva ~ ricoverati_con_sintomi + variazione_totale_positivi +
##     vaccini + colori, family = poisson, data = train.lig)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5093  -0.7591   0.2569   0.7042   3.3772
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 2.812e+00  3.304e-02  85.106  < 2e-16 ***
## ricoverati_con_sintomi      1.263e-03  4.989e-05  25.309  < 2e-16 ***
## variazione_totale_positivi -2.872e-04  3.742e-05  -7.675 1.65e-14 ***
## vaccini                     3.700e-06  8.354e-07   4.430 9.44e-06 ***
## colori2                     4.531e-01  4.318e-02  10.494  < 2e-16 ***
## colori3                     3.163e-01  4.849e-02   6.523 6.87e-11 ***
## colori4                     4.893e-01  4.908e-02   9.969  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3234.85  on 156  degrees of freedom
## Residual deviance:  238.05  on 150  degrees of freedom
## AIC: 1142.6
##
## Number of Fisher Scoring iterations: 4
```

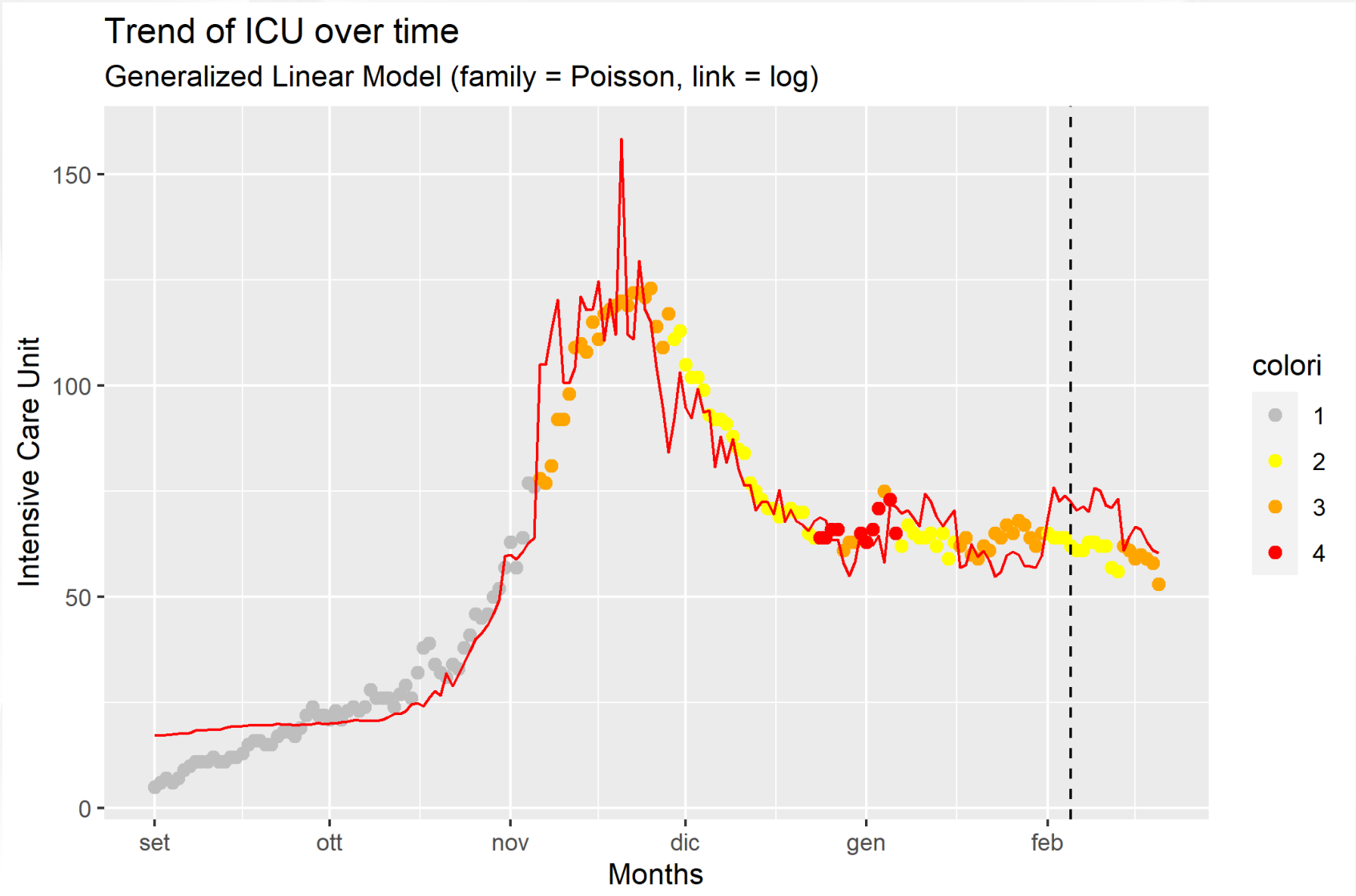
Poisson regression Residuals



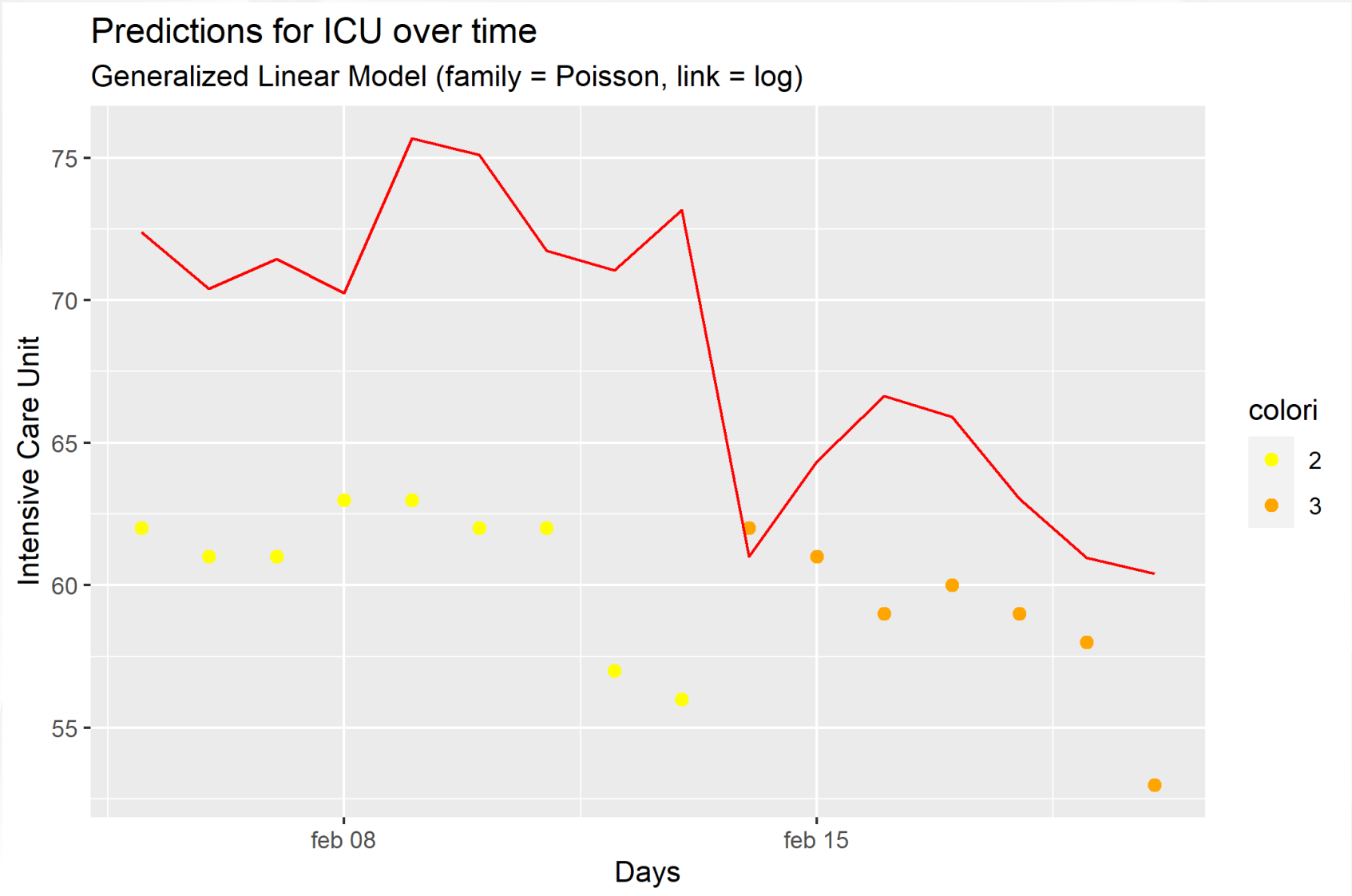




Poisson regression model fit



Poisson regression predictions



Poisson regression Identity link function

We tried and changed the link function from the logarithm to the identity[1], which allows the difference between covariate patterns to be quantified using the difference instead of the ratio:

$$\mathrm{E}[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4.$$

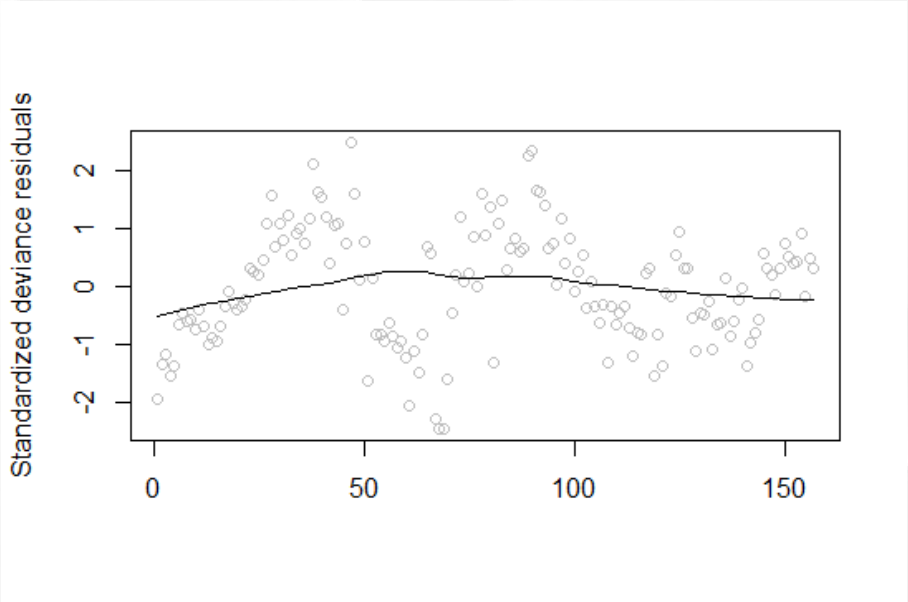
In terms of performances we see an improvement:

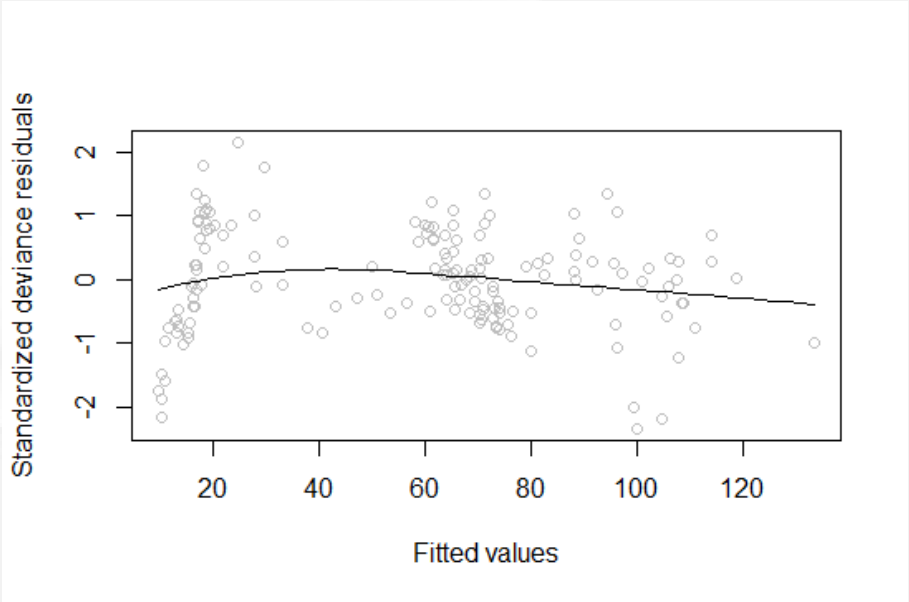
Link function	df	AIC
Logarithm	7	1142.600
Identity	7	1052.751

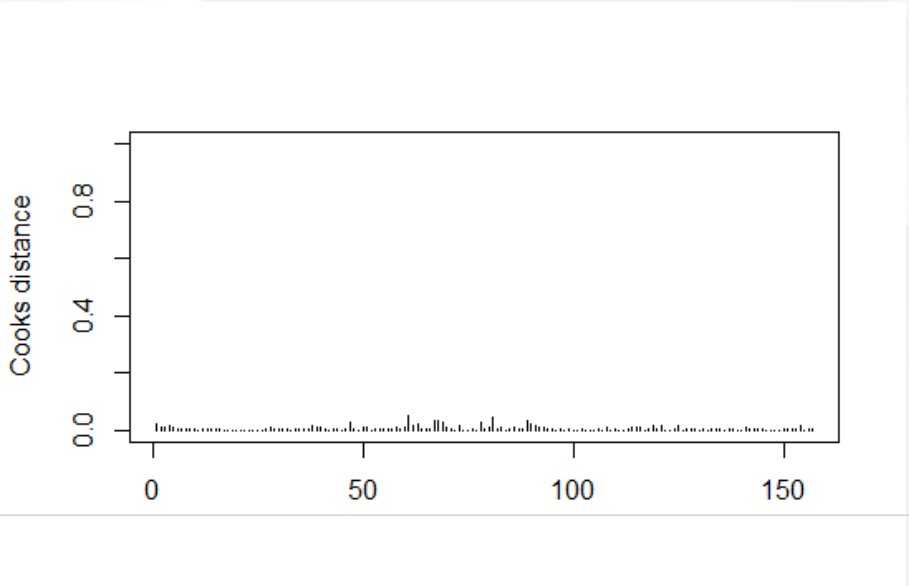
[1] Usually done in epidemiology; less stable than the logarithm.

```
##
## Call:
## glm(formula = terapia_intensiva ~ ricoverati_con_sintomi + variazione_totale_positivi +
##     vaccini + colori, family = poisson(link = "identity"), data = train.lig)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.42279  -0.70342  -0.03817   0.65250   2.46038
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                8.460e+00  7.190e-01  11.767  < 2e-16 ***
## ricoverati_con_sintomi     6.497e-02  2.928e-03  22.189  < 2e-16 ***
## variazione_totale_positivi -2.312e-02  3.646e-03  -6.342 2.27e-10 ***
## vaccini                    -4.040e-05  5.463e-05  -0.739 0.459635
## colori2                     1.412e+01  3.160e+00   4.468 7.89e-06 ***
## colori3                     1.661e+01  3.284e+00   5.057 4.26e-07 ***
## colori4                     1.222e+01  3.227e+00   3.786 0.000153 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3234.8  on 156  degrees of freedom
## Residual deviance:  148.2  on 150  degrees of freedom
## AIC: 1052.8
##
## Number of Fisher Scoring iterations: 5
```

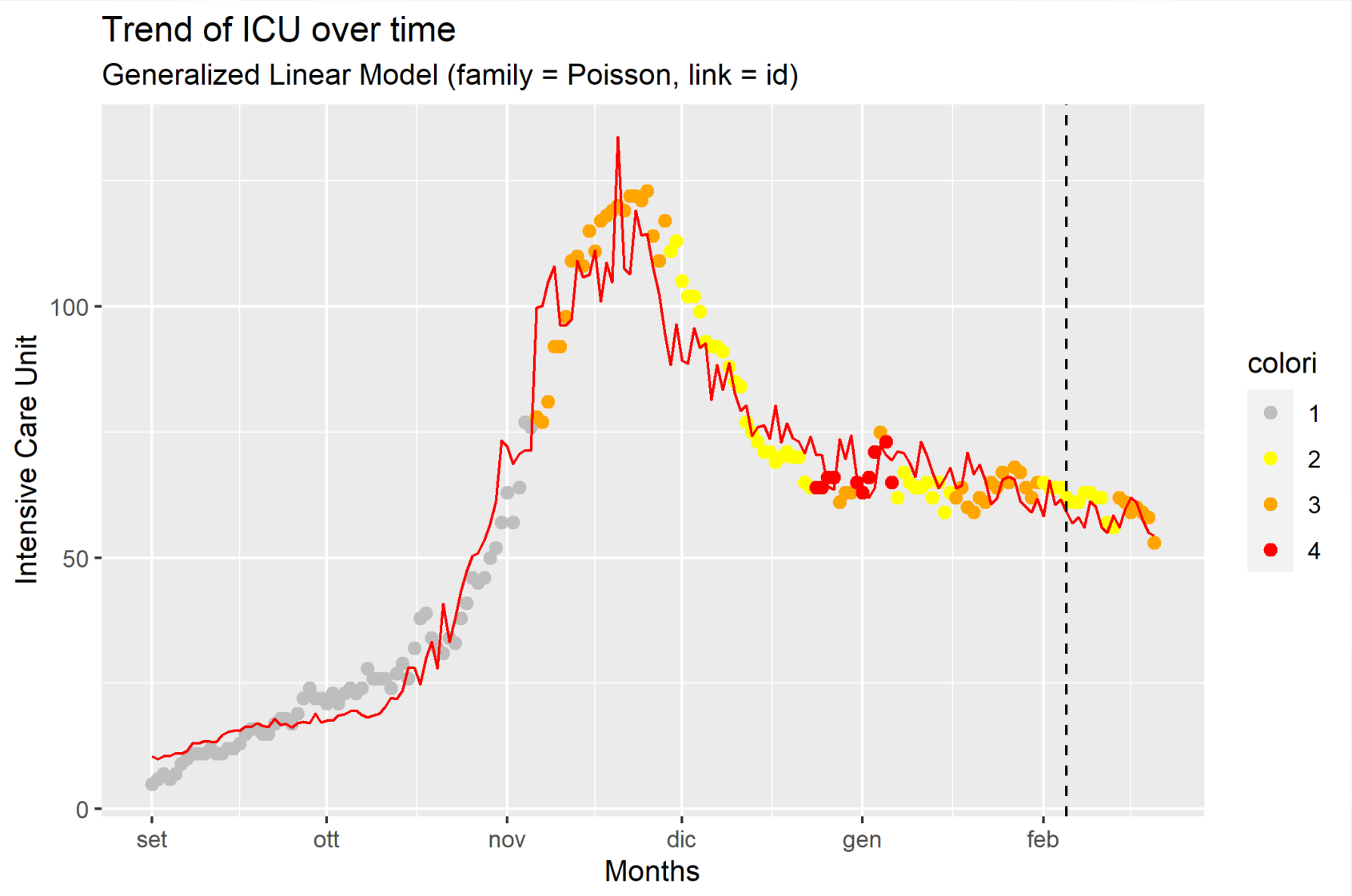
Poisson regression Residuals



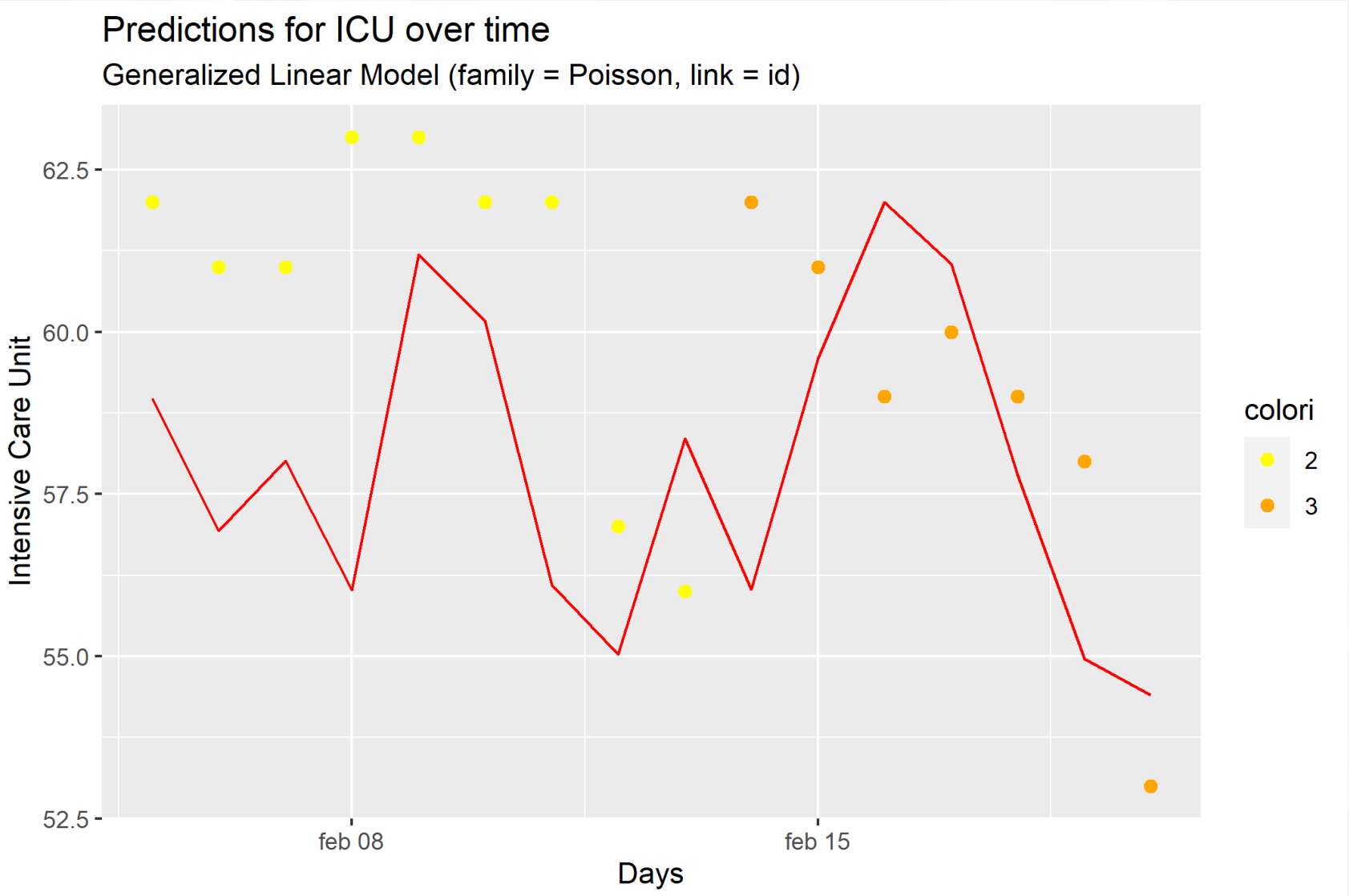




Poisson regression model fit



Poisson regression predictions



Negative binomial

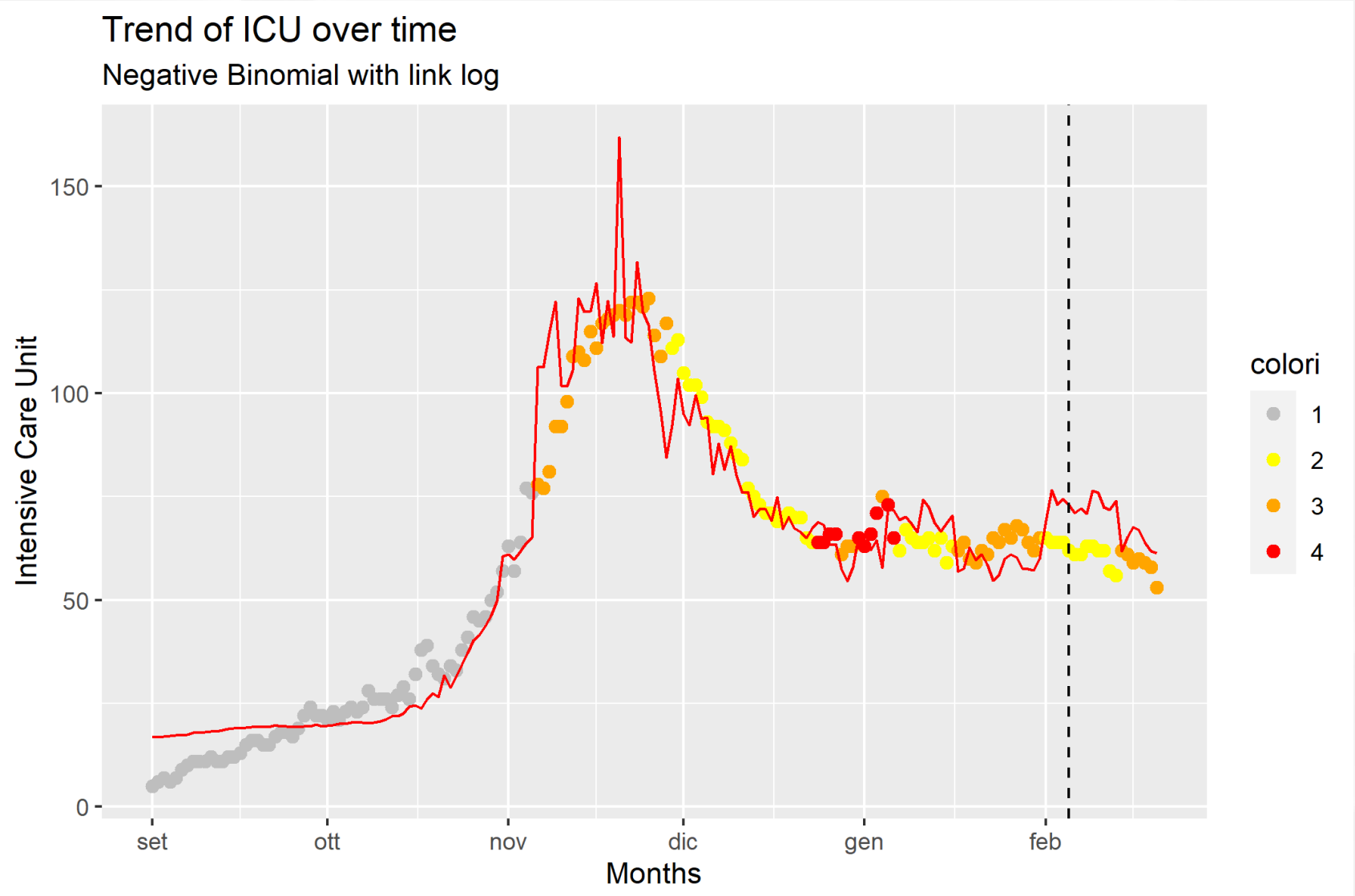
Another problem of the Poisson regression is that we did not consider the overdispersion of the data.

To overcome the issue, we tried a Negative Binomial with the logarithm link function, but in terms of performances there is no sign of improvement.

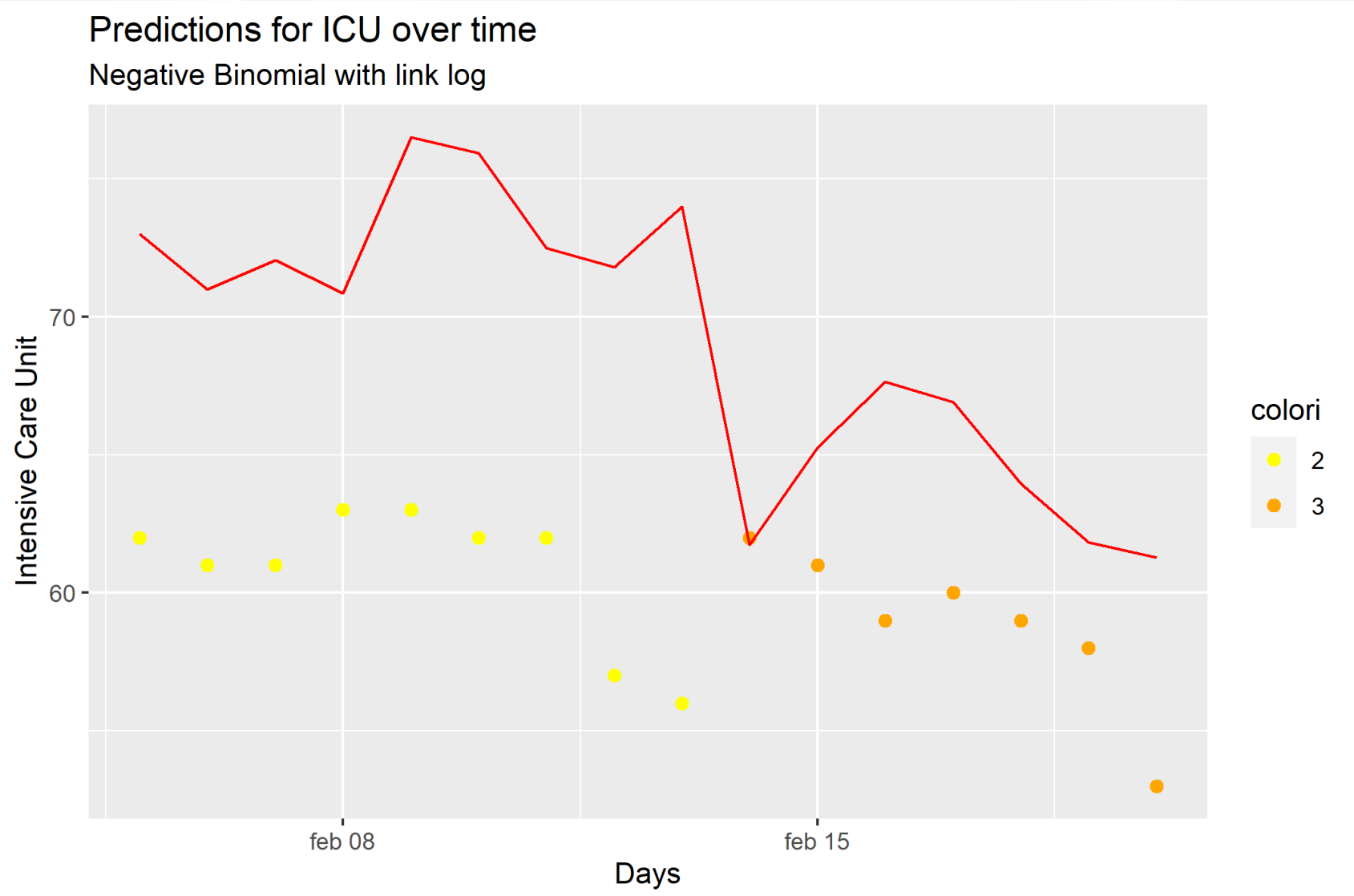
Model	df	AIC
Poisson	7	1142.600
Neg. Binomial	8	1137.361

```
##
## Call:
## glm.nb(formula = terapia_intensiva ~ ricoverati_con_sintomi +
##     variazione_totale_positivi + vaccini + colori, data = train.lig,
##     init.theta = 187.4830039, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3269  -0.6774   0.2244   0.6607   2.7218
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 2.789e+00  3.581e-02  77.892  < 2e-16 ***
## ricoverati_con_sintomi      1.300e-03  5.755e-05  22.596  < 2e-16 ***
## variazione_totale_positivi -2.942e-04  4.727e-05  -6.224 4.85e-10 ***
## vaccini                     4.028e-06  9.714e-07   4.147 3.37e-05 ***
## colori2                     4.397e-01  5.095e-02   8.631  < 2e-16 ***
## colori3                     3.047e-01  5.641e-02   5.402 6.60e-08 ***
## colori4                     4.856e-01  5.689e-02   8.535  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(187.483) family taken to be 1)
##
##     Null deviance: 2540.75  on 156  degrees of freedom
## Residual deviance:  189.76  on 150  degrees of freedom
## AIC: 1137.4
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  187.5
##           Std. Err.:  86.1
##
##  2 x log-likelihood:  -1121.361
```

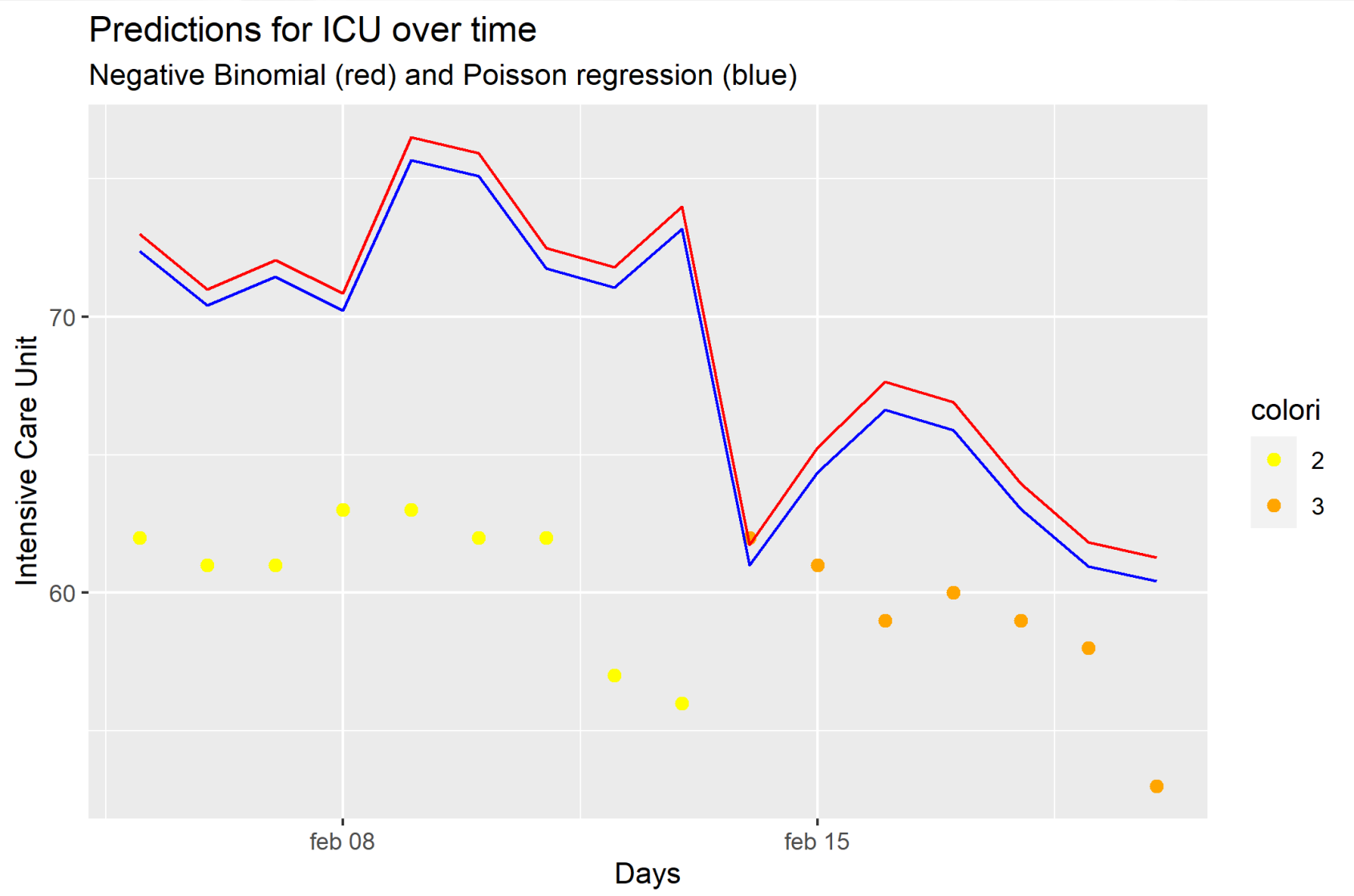
Negative Binomial model fit



Negative Binomial predictions



NBin and Poisson comparison



NBin and Poisson conclusions

The Negative Binomial does not prove to be a better alternative to the Poisson regression model.

Changing the link function reduces the deviance residuals and the overdispersion of the data.

Generalized Additive Model

Due to the non-linearity between the response variable and the covariates we may try to use the Generalized Additive Model (Gam).

We can start with a gam with Family Poisson and link function "identity".

In this model we will not consider the variables:

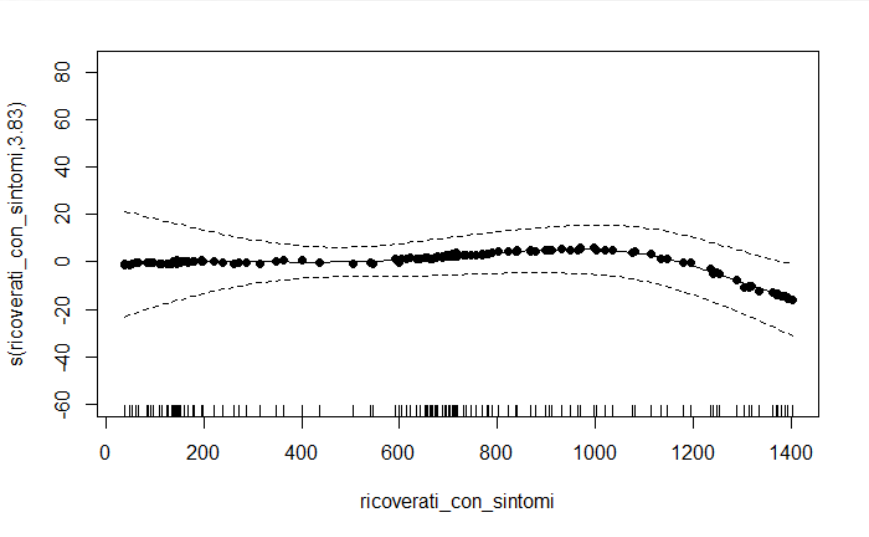
- deceduti_giornaliero
- dimessi_guariti_giornaliero

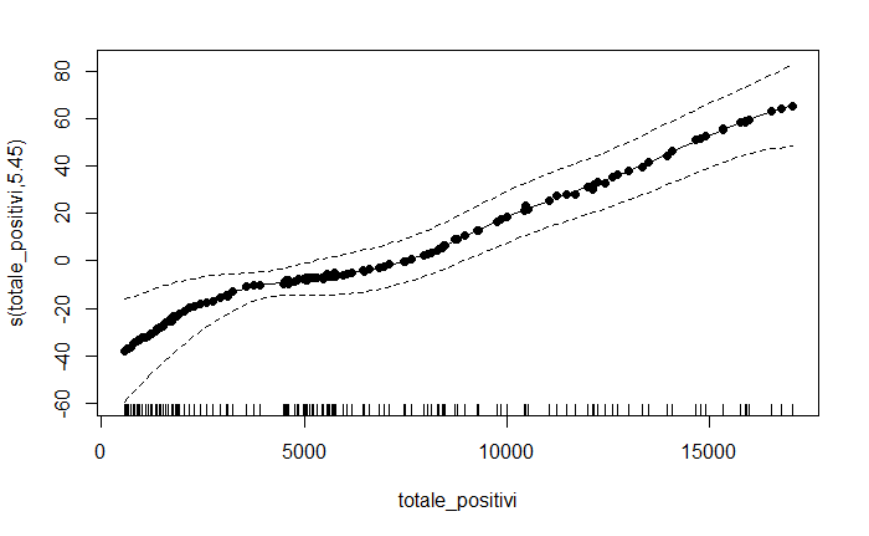
They are not significant.

Generalized Additive Model

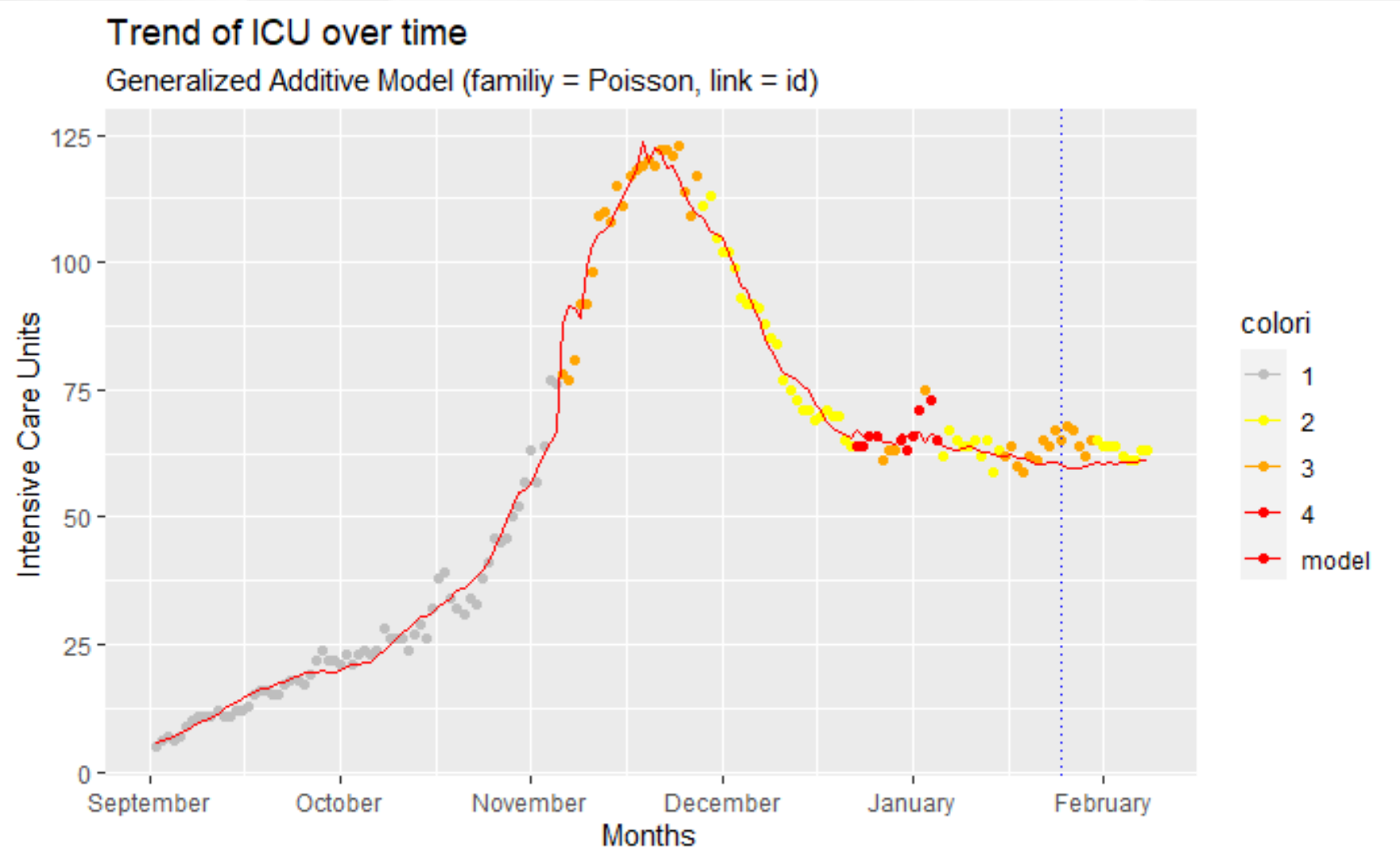
```
Family: poisson
Link function: identity

Formula:
terapia_intensiva ~ s(ricoverati_con_sintomi) + s(totale_positivi) +
    colori

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   44.357      1.472  30.131  < 2e-16 ***
colori2       23.961      2.863   8.369  < 2e-16 ***
colori3       23.874      3.079   7.754 8.88e-15 ***
colori4       26.144      3.494   7.484 7.23e-14 ***
---
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(ricoverati_con_sintomi) 3.829  4.781  12.57  0.0142 *
s(totale_positivi)        5.451  6.518  87.30  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.988   Deviance explained =   99%
UBRE = -0.59997  Scale est. = 1         n = 147
```

Generalized Additive Model Residuals

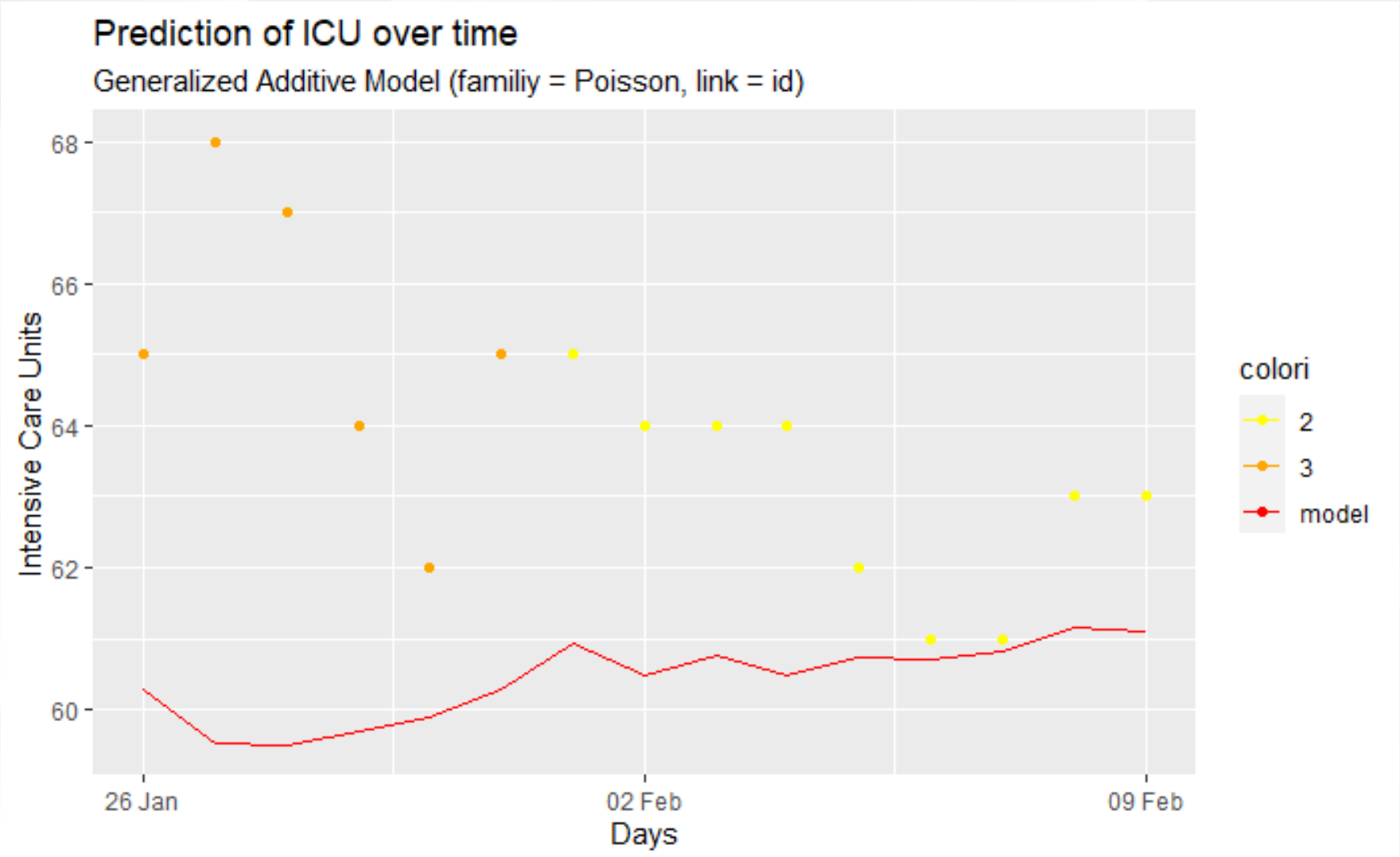




Generalized Additive Model Model fit



Generalized Additive Model Predictions

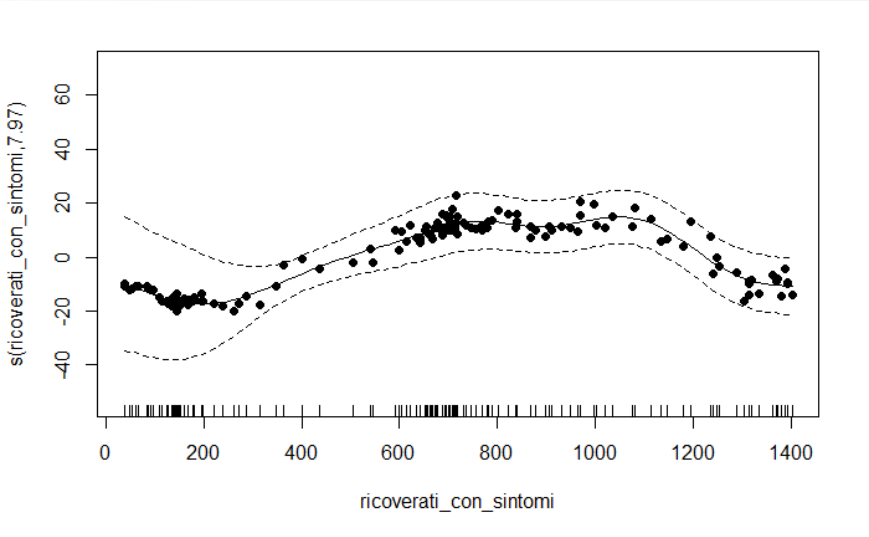


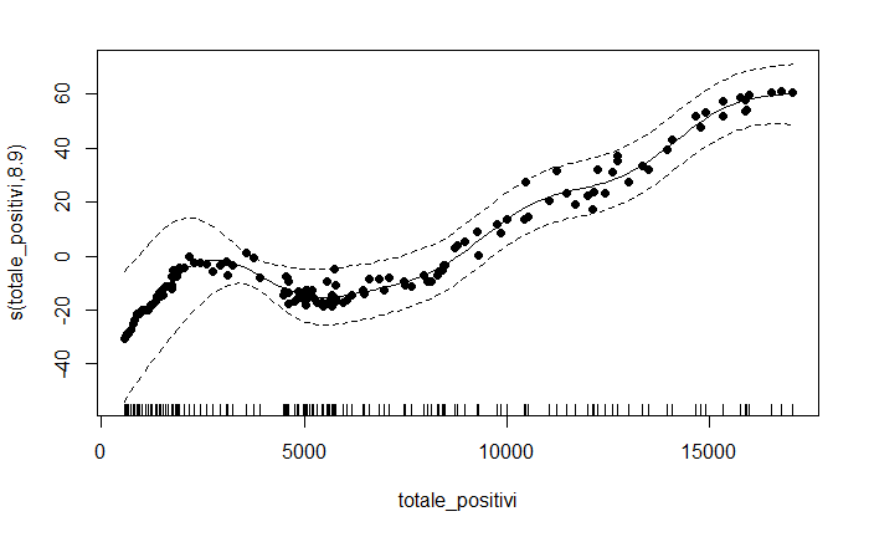
Generalized Additive Model

Now we can try with the Gaussian Family with the same covariates.

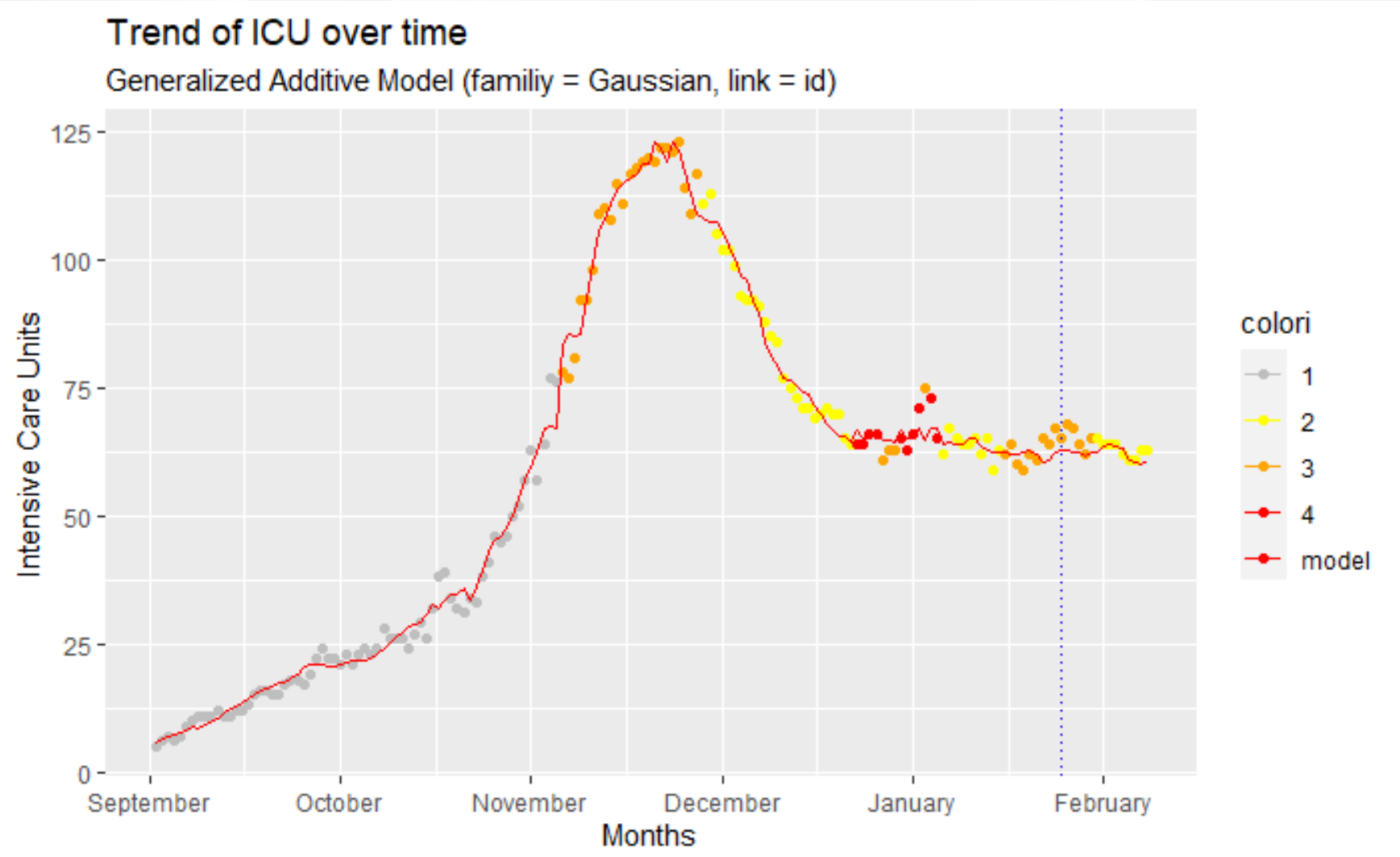
```
Family: gaussian
Link function: identity

Formula:
terapia_intensiva ~ s(ricoverati_con_sintomi) + s(totale_positivi) +
    colori

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   45.641      1.088   41.95   <2e-16 ***
colori2       21.440      2.096   10.23   <2e-16 ***
colori3       21.666      2.020   10.72   <2e-16 ***
colori4       24.642      2.013   12.24   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                           edf Ref.df     F p-value
s(ricoverati_con_sintomi) 7.975  8.675 22.79  <2e-16 ***
s(totale_positivi)        8.902  8.969 57.62  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.991   Deviance explained = 99.2%
GCV = 12.254  Scale est. = 10.514    n = 147
```

Generalized Additive Model Residuals

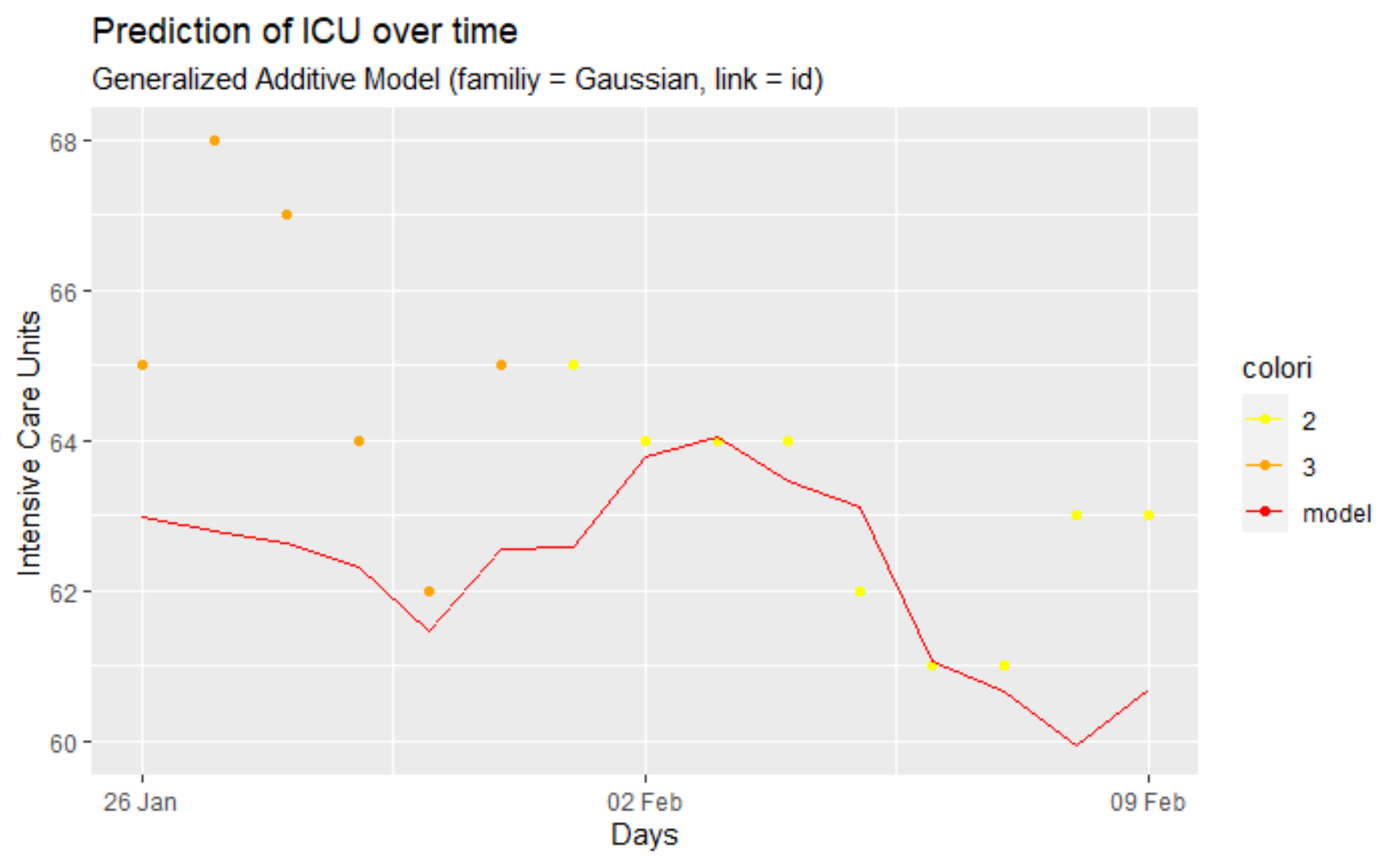




Generalized Additive Model Model fit



Generalized Additive Model Predictions



Generalized Additive Model

In terms of AIC we can see a significant improvement.

Family	df	AIC
Poisson	13.2	889.2437
Gaussian	21.8	784.2528

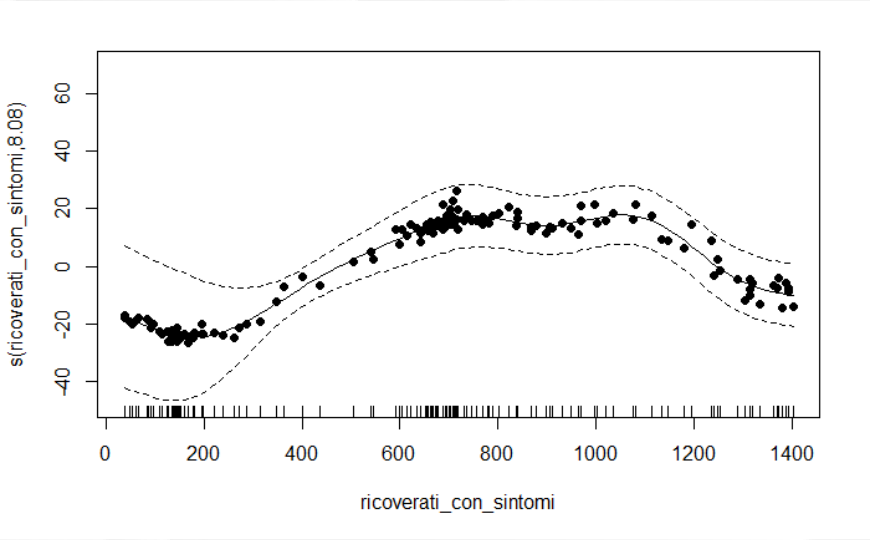
Generalized Additive Model

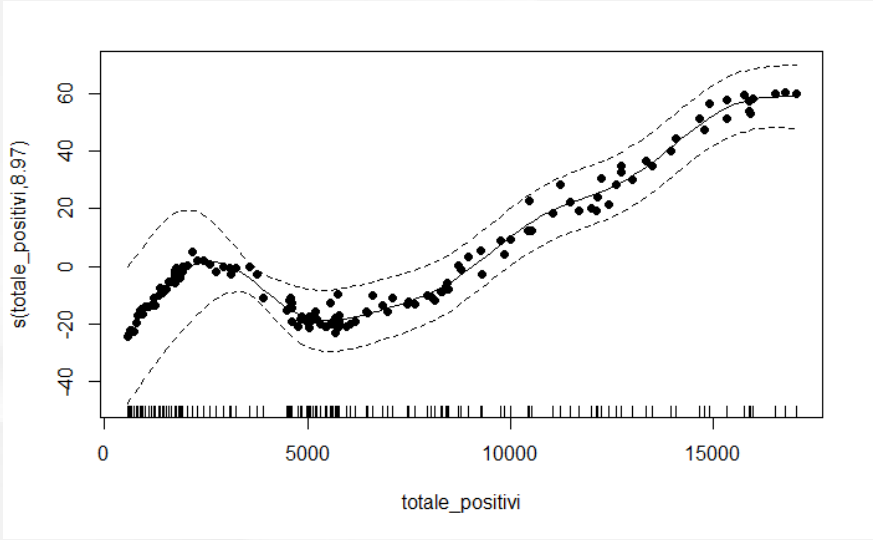
The last model is a Generalize Additive Model with Family Gaussian and link function identity but in this case we added the two covariates that we dropped for the previous two models.

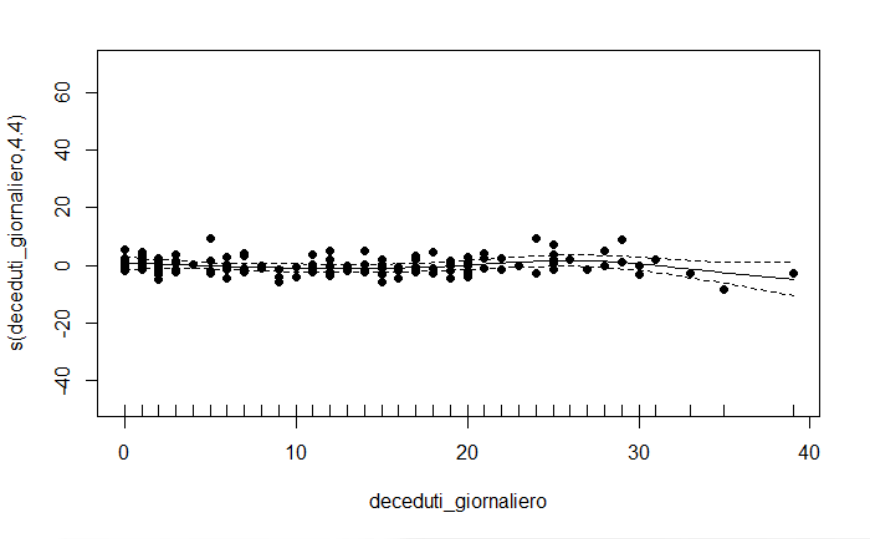
Generalized Additive Model

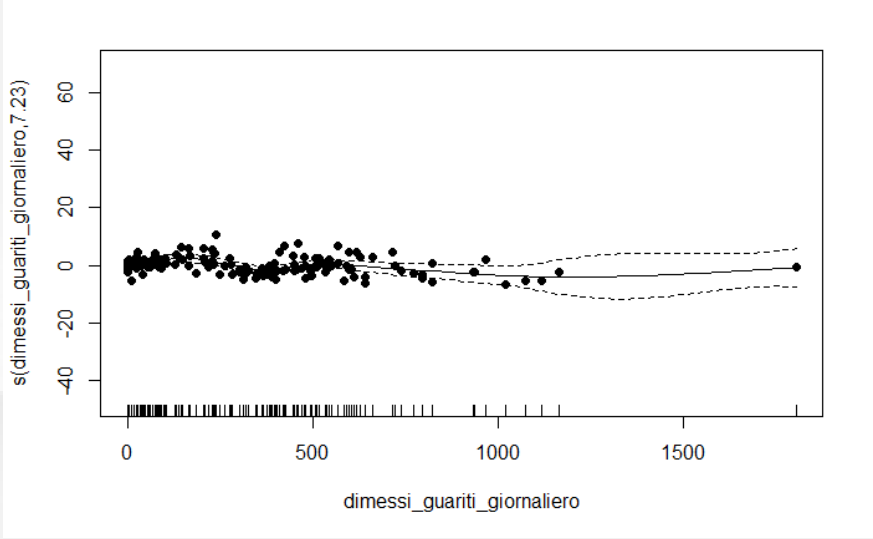
```
Family: gaussian
Link function: identity

Formula:
terapia_intensiva ~ s(ricoverati_con_sintomi) + s(totale_positivi) +
    colori + s(deceduti_giornaliero) + s(dimessi_guariti_giornaliero)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   46.407      1.089  42.613   <2e-16 ***
colori2       20.218      2.078   9.731   <2e-16 ***
colori3       20.079      2.076   9.674   <2e-16 ***
colori4       23.365      1.960  11.918   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                                edf Ref.df      F p-value
s(ricoverati_con_sintomi)     8.077  8.718 18.423  <2e-16 ***
s(totale_positivi)            8.971  8.991 55.040  <2e-16 ***
s(deceduti_giornaliero)       4.396  5.456  1.627  0.1496
s(dimessi_guariti_giornaliero) 7.229  8.219  2.193  0.0268 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.992   Deviance explained = 99.4%
GCV =  11.42  Scale est. = 8.882     n = 147
```

Generalized Additive Model Residuals

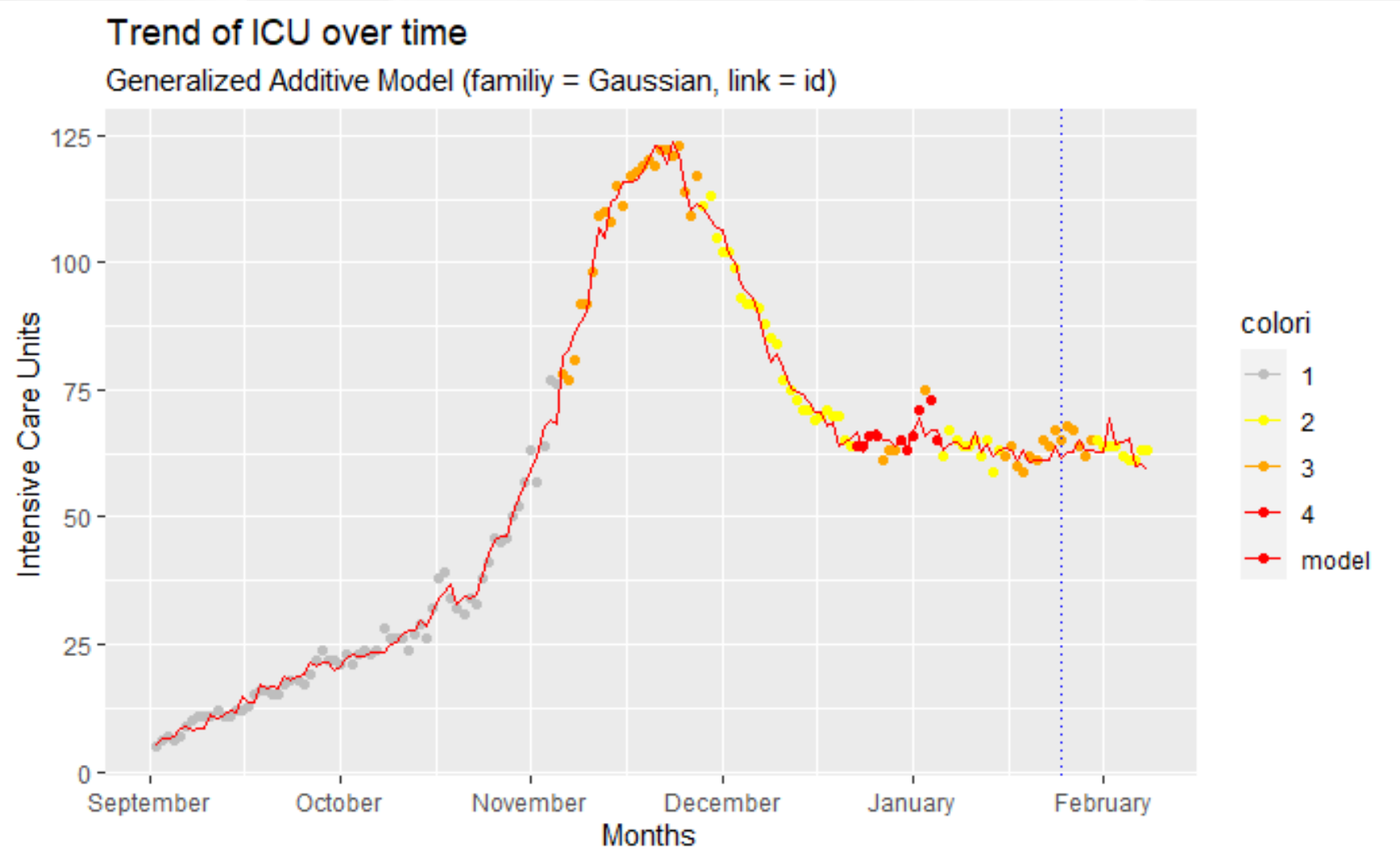




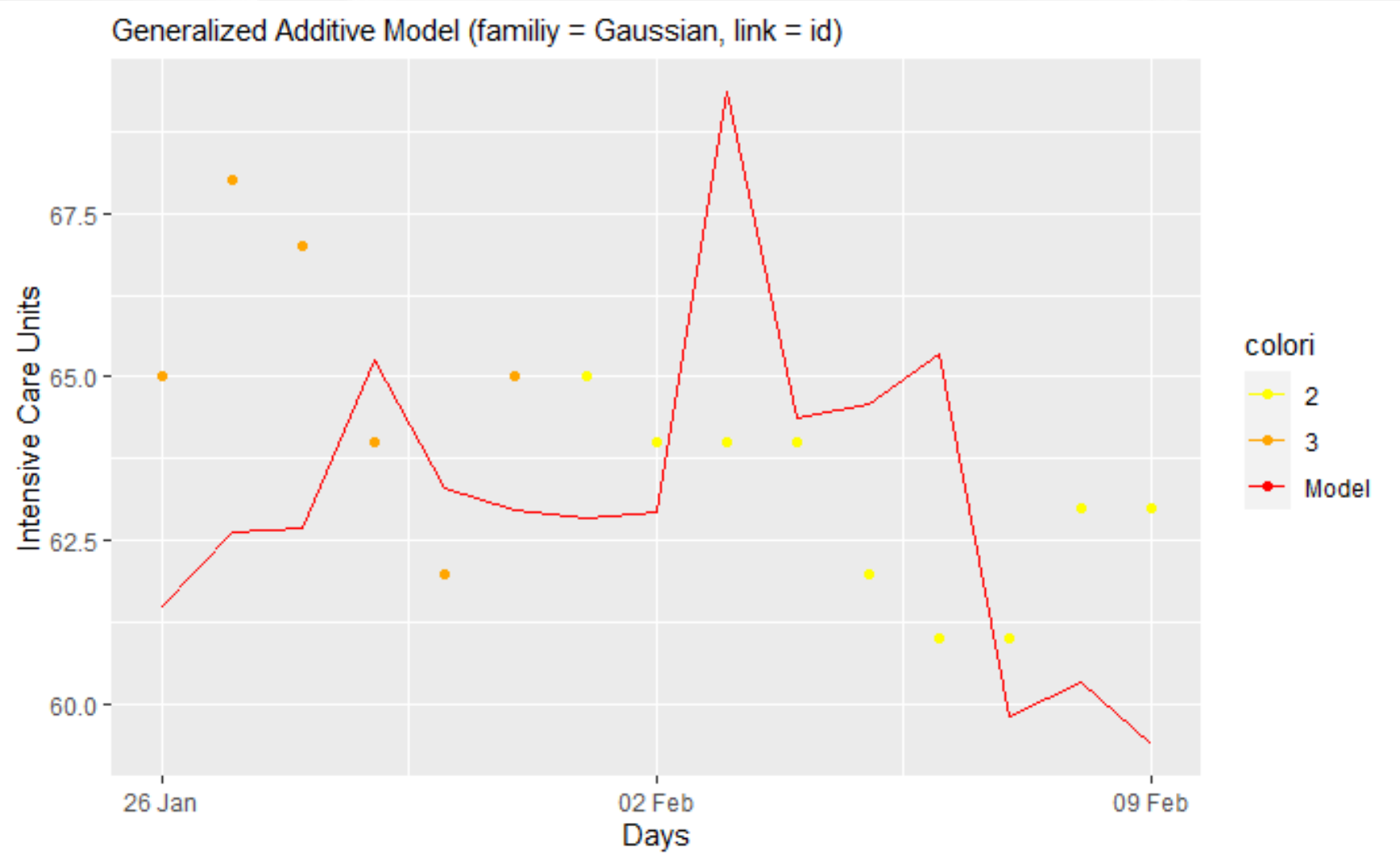




Generalized Additive Model Model fit



Generalized Additive Model Predictions



The AIC in this case is 768.6147 so we have a little improvement.

Conclusions

Model	AIC
Linear model	495
Poisson regression (id)	1052.751
GAM (Gaussian)	768

- Predictions on new data
- Alternative: time series regression
- Other factors: COVID-19 variants