Bayesian Regression with COVID-19 spread data

Michele Guerrini, Davide Mozzi, Carlos Santillán

September 4, 2023

Abstract

We present a Bayesian regression model for predicting the number of patients in a hospital and the ICU with COVID -19, given the number of patients in the same hospital and ICU, the number of new positive subjects and the region color of the previous week.

Contents

1 Problem Description and Dataset

1.1 Problem description

We are given the problem of predicting the number of patients in a hospital and the ICU 7 days from the current date. The prediction is based on the current date's patients both in the hospital and the ICU, new positive subjects, the average number of new positive subjects over the previous 7 days, and the color of the region. We tackle the problem via Bayesian regression

1.2 Dataset description

The dataset contains 205 entries. Each entry corresponds to a date with the following features

1. **newpos:** Number of newly detected COVID-19 positive subjects. An integer value.
2. **intcar:** Number of COVID patients in the ICU. An integer value.
3. **hosp:** Number of COVID patients at the hospital. An integer value.
4. **newpos-av7D:** Average number of newly detected COVID-19 positive subjects over the previous 7 days. A float value.
5. **color:** Color of the region over the previous 7 days. String that can contain the following values: *Bianca*, *Gialla*, *Arancione*, *Rossa*.
6. **day:** Current date. String in R date format.
7. **hospH8:** A target variable, the number of COVID patients at the hospital 7 days from now.
8. **intcarH8:** A target variable, the number of COVID patients in the ICU 7 days from now.
9. **dayH8:** The current date plus 7 days.

We will extend this by adding a categorical variable **season**, which will have values: *winter*, *spring*, *summer*, *fall*.

1.3 Data exploration

We start by looking at a pairs plot of the **newpos**, **intcar**, **hosp**, **newpos-avg7D**, and **hospH8**. The pairs plot contains a kernel density estimation of the distribution along the diagonal, the estimations are distinguished by color of the region. We also see the scatter plots for each pair of features, and the correlations between them for each region.

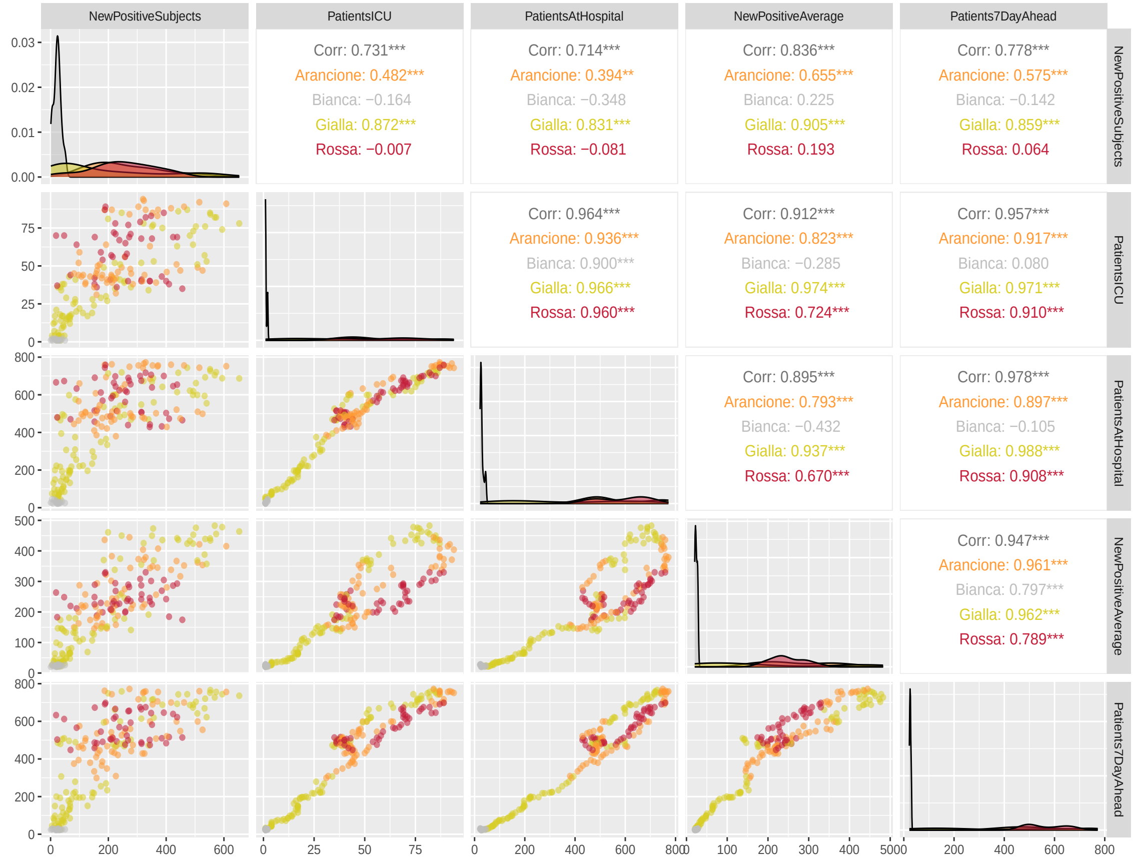

Figure 1: Pairs plot between the features and the number of patients at the hospital 7 days from now

We can immediately see an almost linear relationship between the number of patients in the ICU and those at the hospital, suggesting that we may remove one of them. The density plots are as we would expect, the white regions (lower COVID-19 presence) have all the mass concentrated around low values of the features, while the other regions are more spread out and have a center of mass further to the right (higher values). We also see that the relationship between the number of patients at the hospital, and that of 7 days in the future could be linear.

We now look at the correlations with the number of patients in the ICU

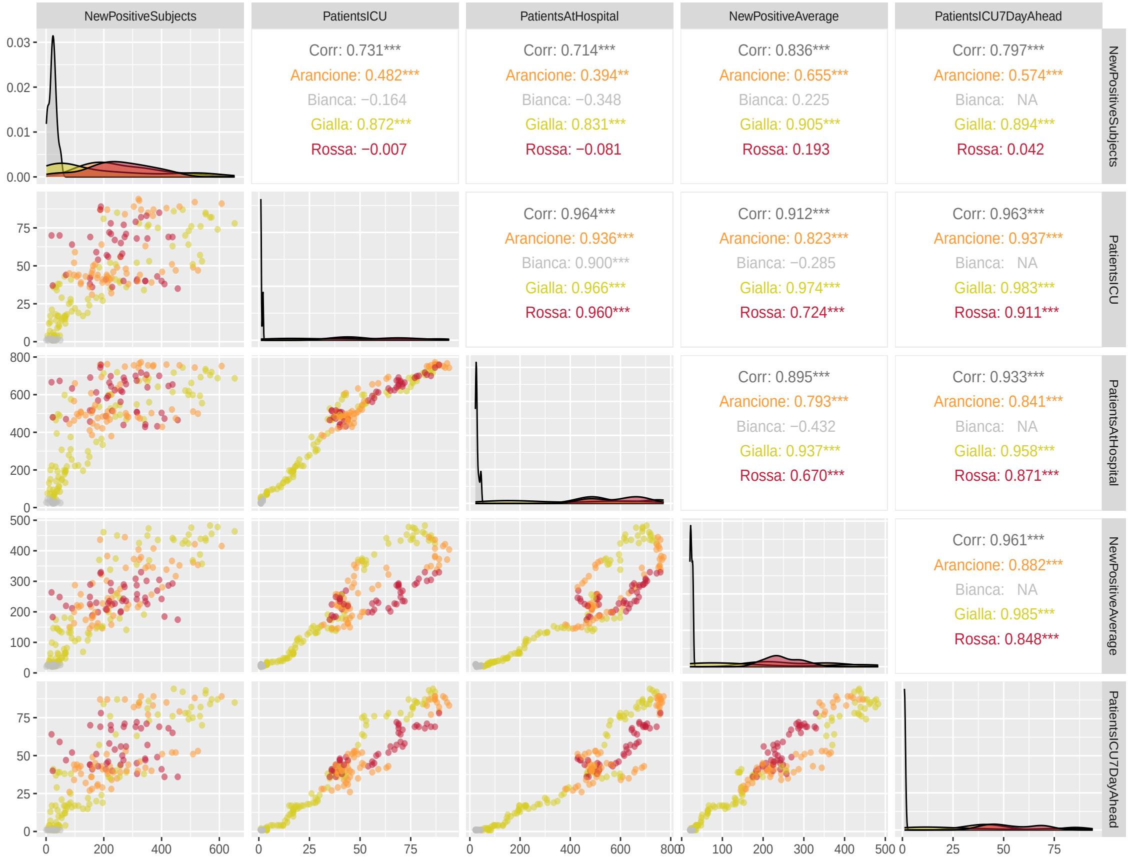

Figure 2: Pairs plot between the features and the number of patients in the ICU 7 days from now

Again, for some of the features the relationship appears to be linear w.r.t. the target variable. Note that where the correlation is marked NA the value is zero. We used the Pearson correlation coefficient $r_{X,Y}$, defined as

$$r_{X,Y} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}$$

Now we qualitatively assess how much each feature varies via boxplots (Figure 3). A boxplot allows us to see the spread of our data, the boxes themselves contain the interquantile range, the whiskers show the range of the rest of the data. We also mean center the features.

Since the some features vary greatly w.r.t. to others, see **intcar** and **hosp**, it could be a conventient to normalize the features. In Figure 4 we see the boxplots for the normalized data.

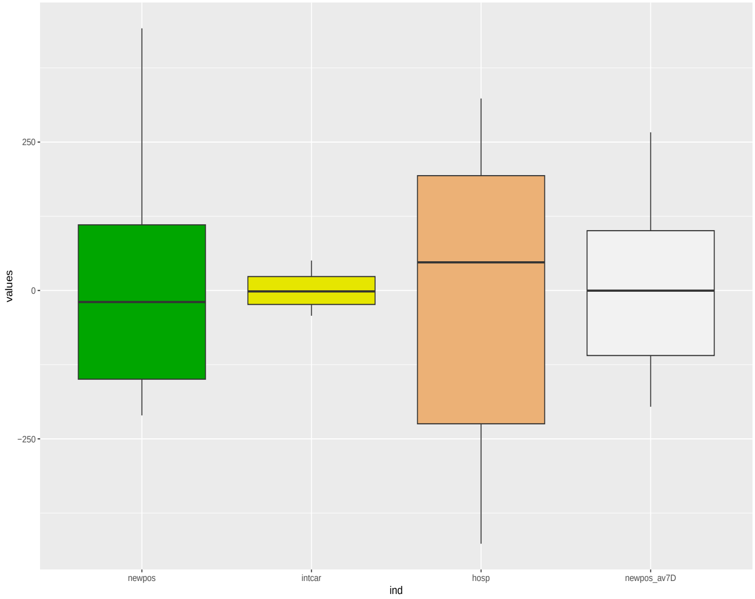

Figure 3: Boxplot of the mean-centered features

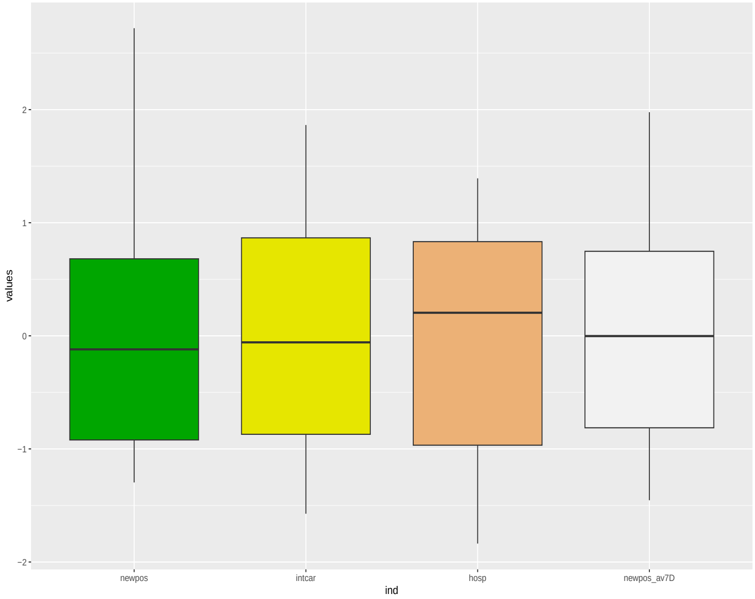

Figure 4: Boxplot of the mean-centered features

Lastly, we take at the correlation matrices of the features. Figure 5 shows the full correlation matrix, no distinction made between regions.

Figure 6 shows the correlation matrix for each region.

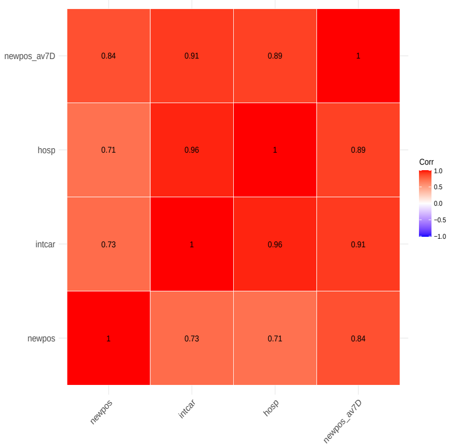

Figure 5: Correlation matrix of the features

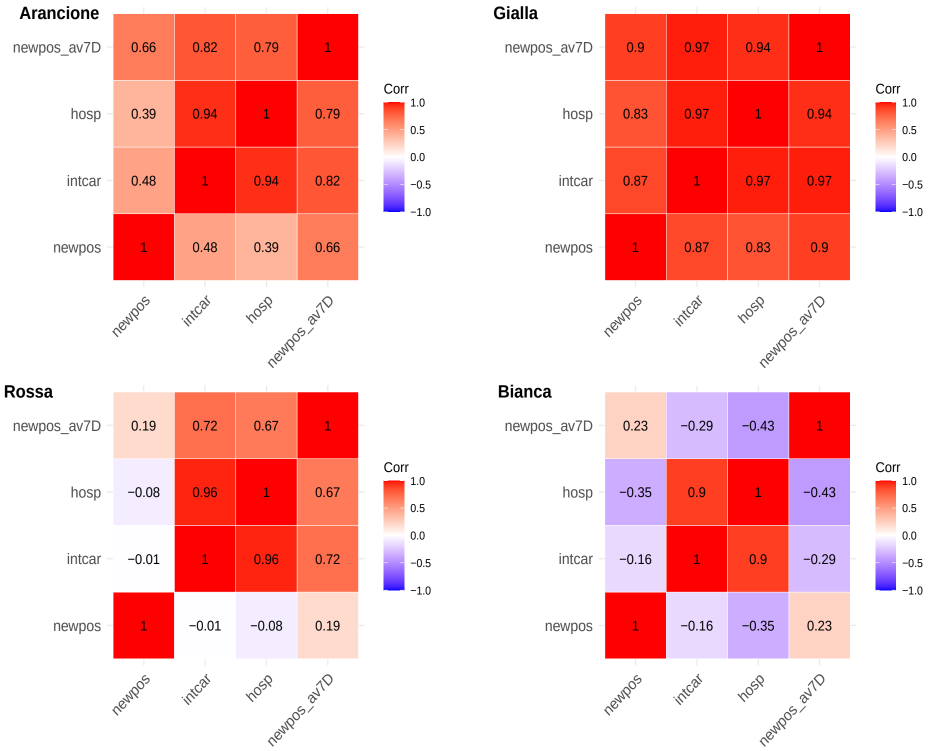

Figure 6: Correlation matrix of the features per region

Before proceeding, we shuffle and split the dataset in training and test sets with a 87/13 ratio. We also normalize the covariates in the training set, but not the target variables.

1.4 Model specification

In the Bayesian setting, we use a Gaussian likelihood

$$y|\beta, \sigma^2, X \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$$

For the prior we initially used default one provided by BAS. BAS uses an approximation of the Zellner-Siow prior

$$\begin{aligned}
\beta = (\beta_1, \dots, \beta_k)|\sigma^2, \alpha, X &\sim \mathcal{N}_k(0, \alpha\sigma^2(X^T X)^{-1}) \\
\sigma^2|X, \alpha &\sim \pi(\sigma^2) = \sigma^{-2} \\
1/\alpha &\sim \pi_0 = \Gamma(1/2, n/2)
\end{aligned}$$

BAS sets the hyperparameter α to 1.

We try to predict each target separately and start with **hostH8**, we consider a model with all the covariates. In particular, **color** is considered a categorical variable.

$$\begin{aligned}
\textbf{hospH8} = \beta_0 + \beta_1\textbf{Gialla} + \beta_2\textbf{Arancione} + \beta_3\textbf{Rossa} + \beta_4\textbf{newpos} \\
+ \beta_5\textbf{intcar} + \beta_6\textbf{hosp} + \beta_7\textbf{newpos-av7D} + \varepsilon
\end{aligned}$$

We transform the **color** covariate into an ordinal feature starting from 1 (*Bianca*) and ending in 4 (*Rossa*).

We must choose a prior, we initially set it to Zellner's G-prior

$$\begin{aligned} \beta = (\beta_1, \ldots, \beta_k) | \sigma^2 &\sim \mathcal{N}_k(0, \alpha\sigma^2(X^T X)^{-1}) \\ (\beta_0, \sigma^2) &\sim \pi(\beta_0, \sigma^2) = \sigma^{-2} \end{aligned}$$