Bayesian Regression with COVID-19 spread data

Michele Guerrini, Davide Mozzi, Carlos Santillán

September 5, 2023

Abstract

We present a Bayesian regression model for predicting the number of patients in a hospital and in its ICU with COVID-19, given the number of patients in the same hospital and in the 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 covariates:

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 covariates and the correlations between them for each region.

Figure 1: Pairs plot between the covariates 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 covariates and the number of patients in the ICU 7 days from now

Again, for some of the covariates, 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 covariate 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 covariates.

Since some of the covariates vary greatly w.r.t. to others, see **intcar** and **hosp**, it could be a convenient to normalize them.

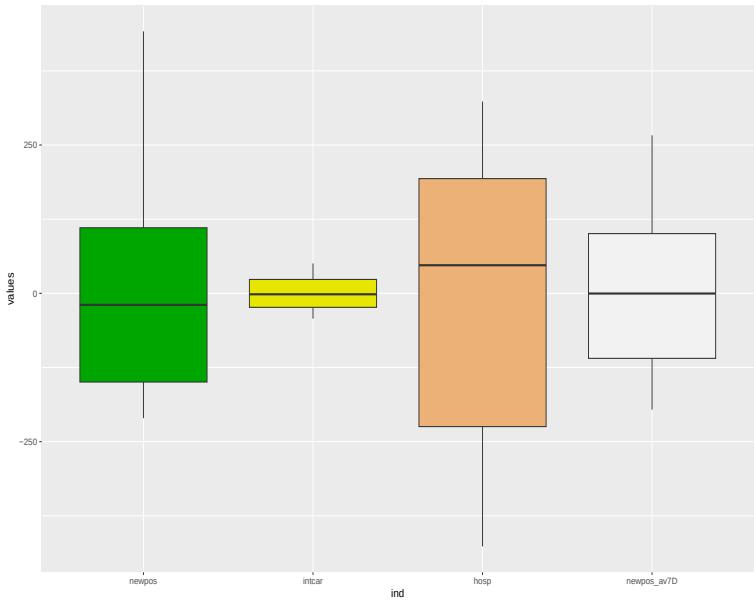

Figure 3: Boxplot of the mean-centered covariates

Lastly, we take a look at the correlation matrices of the covariates. Figure 4 shows the full correlation matrix, no distinction made between regions.

Figure 5 shows the correlation matrix for each region.

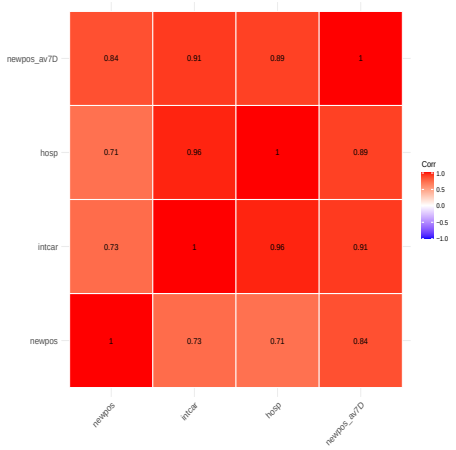

Figure 4: Correlation matrix of the covariates

Before proceeding, we shuffle the dataset and normalize the covariates, but not the target variables.

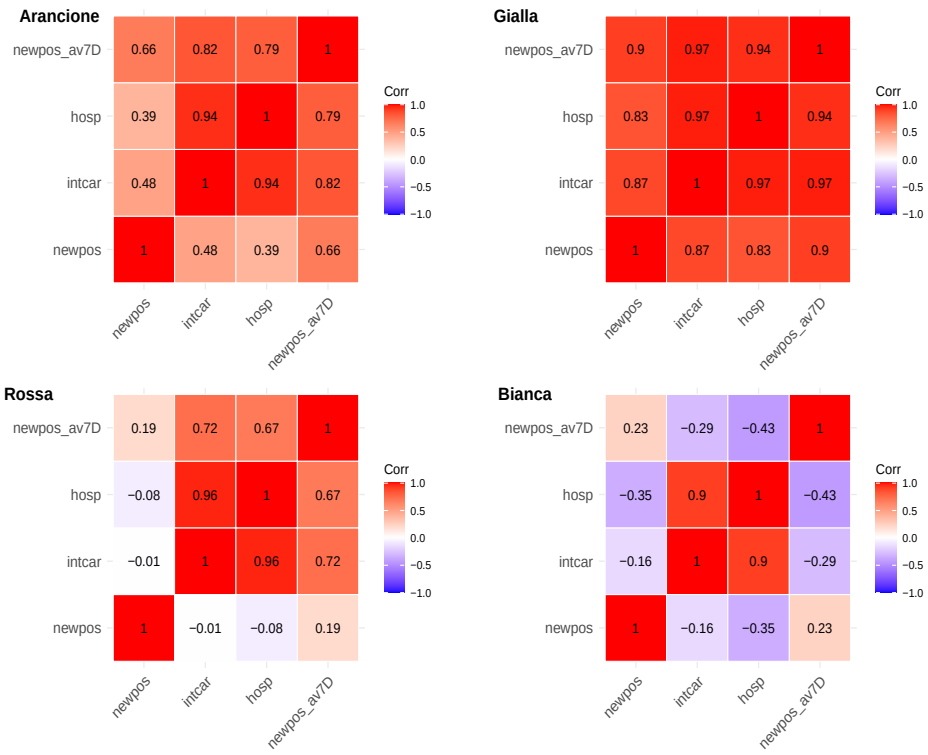

Figure 5: Correlation matrix of the covariates per color of the region

2 Model specification

2.1 Likelihood and prior

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 use default one provided by BAS, an approximation of the Zellner-Siow prior, with the hyperparameter α set to 1. This prior is defined as:

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

We try to predict each target separately and start with **hospH8**, we consider a model with all the covariates. In particular, **color** is considered a categorical variable, which was transformed into 3 binary covariates via one-hot encoding. The model is

$$\begin{aligned} \textbf{hospH8} = \beta_0 + \beta_1\textbf{gialla} + \beta_2\textbf{arancione} + \beta_3\textbf{rossa} + \beta_4\textbf{newpos} \\ + \beta_5\textbf{intcar} + \beta_6\textbf{hosp} + \beta_7\textbf{newpos\_av7D} + \varepsilon \end{aligned}$$

Since we are only using one model which always includes these covariates the probability of them belonging to the model is always 1. Table 1 shows the posterior means and standard deviations of the coefficients[1]

We perform inference on the test set, for which BAS uses the posterior predictive distribution. We recall that in general the predictive posterior distribution is computed as

$$\int \pi(y_{\text{new}}|\sigma^2, \beta, X_{\text{new}})\pi(\sigma^2, \beta|y, X)\, d\sigma^2 d\beta$$

To evaluate the results we use k-fold cross-validation, computing and averaging the mean square error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i is the actual number of people hospitalized after 7 days, and $\hat{y}_i$ is our prediction. We obtain a MSE of 938.4436.

[1]Do not confuse them with the actual covariates, these refer to the coefficents β_i of i-th covariate

For the target variable **intcarH8** we consider the same model

$$\begin{aligned}\textbf{intcarH8} = \beta_0 + \beta_1\textbf{gialla} + \beta_2\textbf{arancione} + \beta_3\textbf{rossa} + \beta_4\textbf{newpos} \\ + \beta_5\textbf{intcar} + \beta_6\textbf{hosp} + \beta_7\textbf{newpos\_av7D} + \varepsilon\end{aligned}$$

The MSE is 23.4831. We show a more in-depth posterior analysis with the final model. Please note that the covariates in the test set are normalized with the training set's mean and standard deviation.

Zellner's g-prior

We change prior to Zellner's g-prior, of the form

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

For the hyperparameter we chose a value of $\alpha = 100$ and obtain an MSE of 943.1065 for the target **hospH8** and of 23.55807 for **intcarH8**.

2.2 Model selection

To perform model selection we use the **Bayesian information criterion (BIC)** defined as

$$BIC = -2\ln\left[\mathcal{L}\left(\mathbf{y} \mid \hat{\beta},\, \hat{\sigma}^2,\, M\right)\right] + (p+1)\ln(n)$$

This approach uses a non-informative prior to train a large number of models trying different subsets of covariates and then chooses the best one based on the above criterion, which is the one with lowest BIC.

In Table 1 we show the posterior probabilities that each covariate is contained in the model for the prediction of **hospH8**. In Table 2 we see the same for **intcarH8**.

As we can see we certainly want to include the covariates **rossa**, **hosp**, and **newpos_av7D** in our model for **hospH8** and **arancione**, **intcar**, and **newpos_av7D** for **intcarH8**.

In the case of **hospH8** the top 5 models have the following posterior probabilities Table 3 and Table 4 show the posterior probabilities of the top 5 models for each target variable.

Figure 6 and Figure 7 show a heatmap-like plot of the posterior probabilities and the inclusion of covariates of each model.

Since, for **hospH8**, none of the models have a posterior probability above 50%, we do not choose the best model, Model 1, but we use Bayesian Model Averaging

Coefficient	Posterior
Intercept	1.0000000
gialla	0.1736033
arancione	0.1845924
rossa	0.9420397
newpos	0.1676675
intcar	0.4934792
hosp	1.0000000
newpos_av7D	1.0000000

Table 1: Probabilities of each covariate for **hospH8**

Coefficient	Posterior
Intercept	1.0000000
gialla	0.09773376
arancione	0.99781126
rossa	0.11141621
newpos	0.14062014
intcar	1.0000000
hosp	0.08919494
newpos_av7D	1.0000000

Table 2: Probabilities of each covariate for **intcarH8**

Models	Model 1	Model 2	Model 3	Model 4	Model 5
Posterior probability	0.3664	0.2734	0.0496	0.0478	0.0385

Table 3: Posterior probabilities of the best 5 models for **hospH8**

Models	Model 1	Model 2	Model 3	Model 4	Model 5
Posterior probability	0.6669	0.0877	0.0586	0.0514	0.0504

Table 4: Posterior probabilities of the best 5 models for **intcarH8**

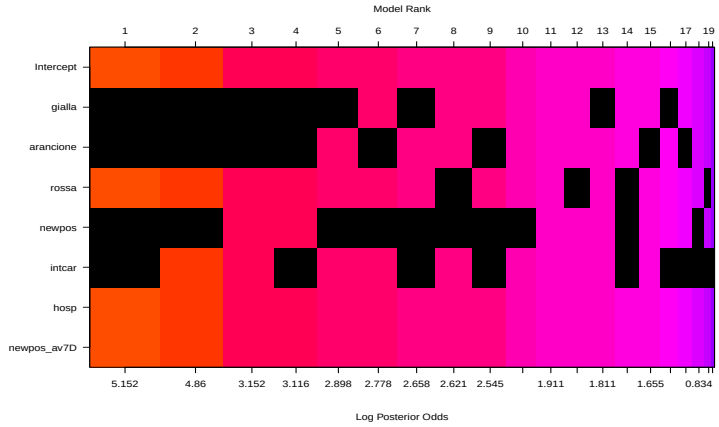

Figure 6: Model ranking and posterior probabilities

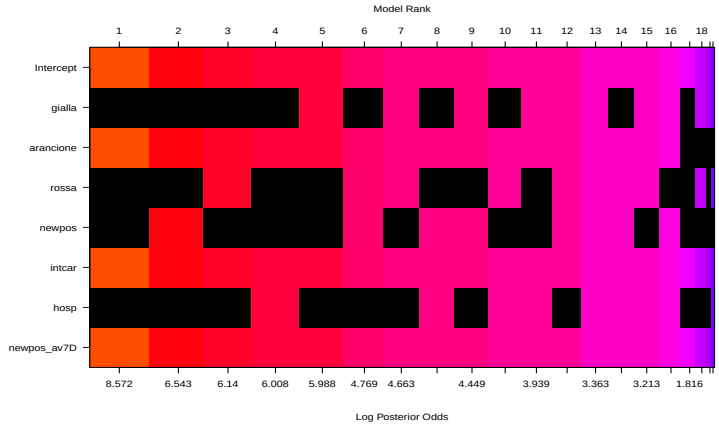

Figure 7: Model ranking for **intcarH8**

(BMA), the posterior probability of a quantity of interest Δ, computed via BMA, is

$$\pi(\Delta|\mathbf{y}) = \sum_{j=1}^{2^p} \pi(\Delta|M_j, \mathbf{y})\pi(M_j|\mathbf{y})$$

where p is the number of covariates.

We can therefore compute our predictions as a weighted average of each model's prediction, where the weights are the model's posterior probabilities

$$\hat{Y} = \sum_{j=1}^{2^p} \hat{Y}_j^* p(M_j|X)$$

The MSE for this aggregate model is 929.9491, sligthly better than the first model with all the covariates (while using only the Highest Probability Model the MSE is 950.7221).

For **intcarH8** the best model has a posterior probability of about 67%, so we could try to use only that one to perform inference. So we choose the model

$$\textbf{intcarH8} = \beta_0 + \beta_1 \textbf{arancione} + \beta_2 \textbf{intcar} + \beta_3 \textbf{newpos\_av7D} + \varepsilon$$

which yields a MSE of 23.00852, slightly better than the one using all the covariates (while using BMA the MSE is 23.11804, which is very similar).

3 Posterior analysis

3.1 Posterior distributions

We start with the aggregate model chosen for **hospH8**. The posterior distributions of the coefficients are

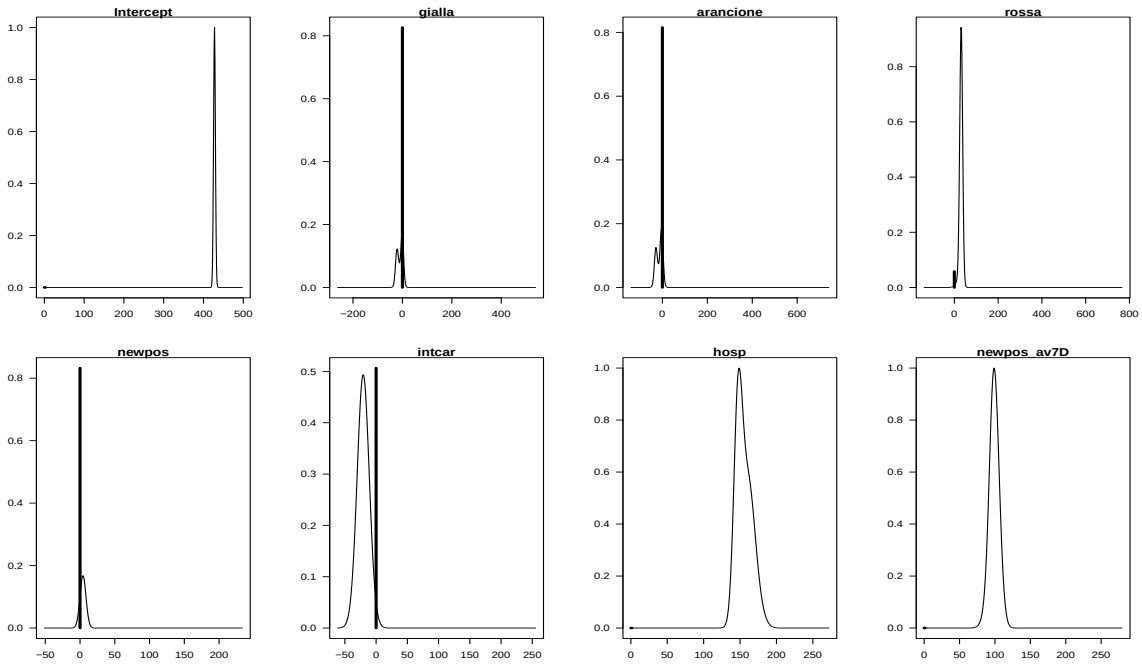

Figure 8: Posterior distributions of coefficients for best **hospH8** model

Judging by where they are centered we could say that the covariate **hosp** has more weight when performing inference w.r.t. **newpos_av7D** and **rossa**.

In Figure 9 we show the credible intervals of the coefficients.

We now consider the best model for predicting **intcarH8**. Figure 10 shows the posterior distribution of the coefficients. It appears that both **intcar** and **newpos_av7D** have a higher weight w.r.t. **arancione**. Figure 11 shows the credible intervals.

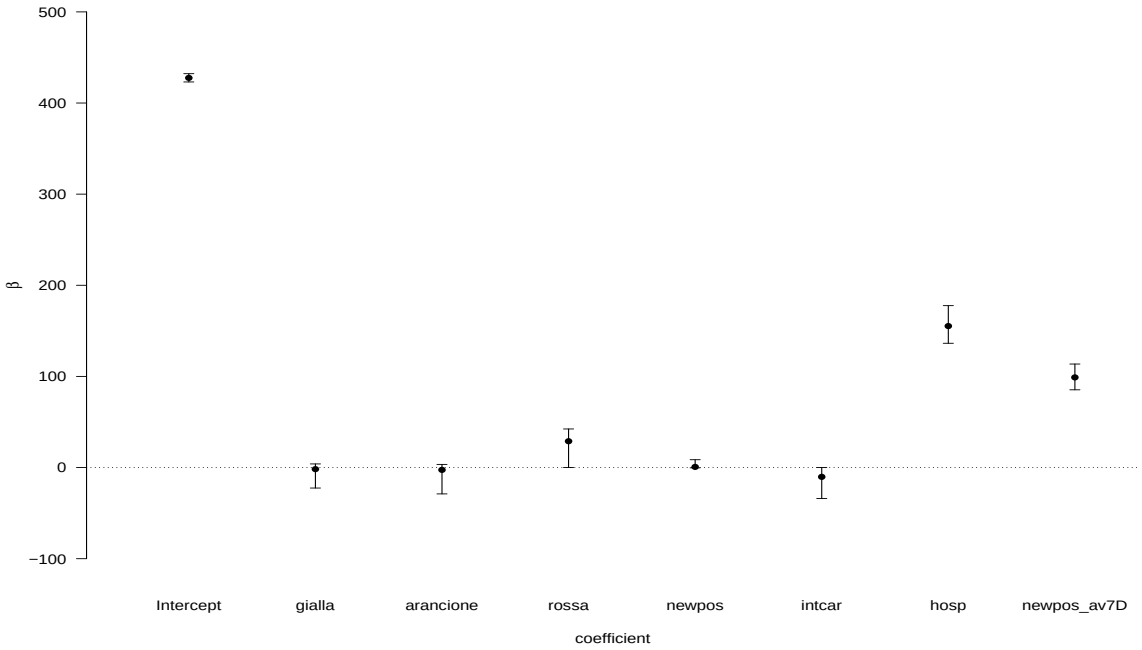

Figure 9: 95% credible intervals for the coefficients in the **hospH8** model

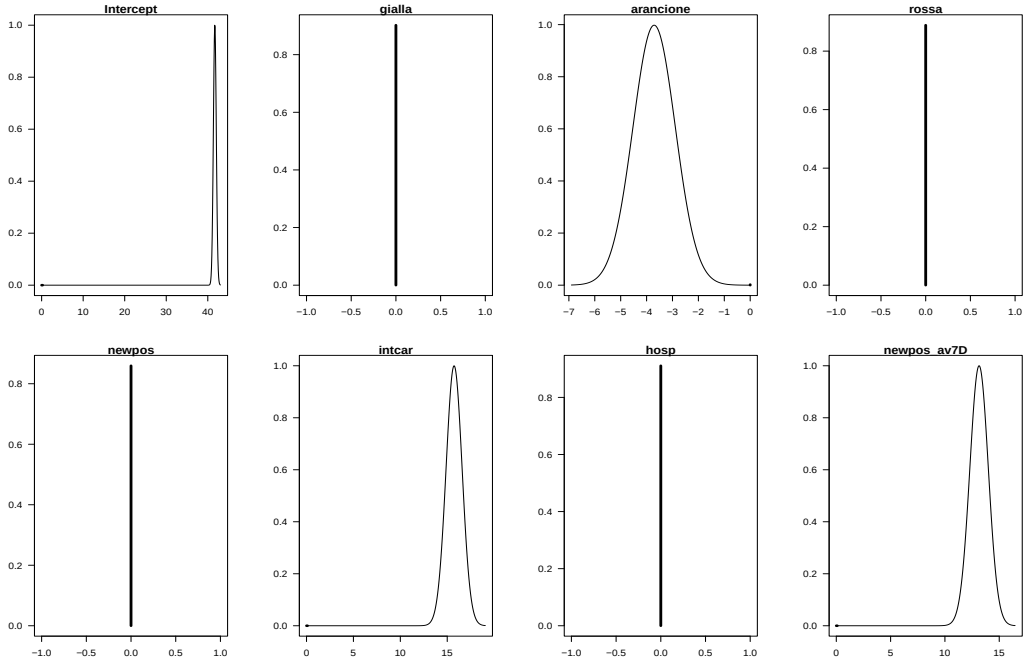

Figure 10: Posterior distributions of coefficients for best **intcarH8** model

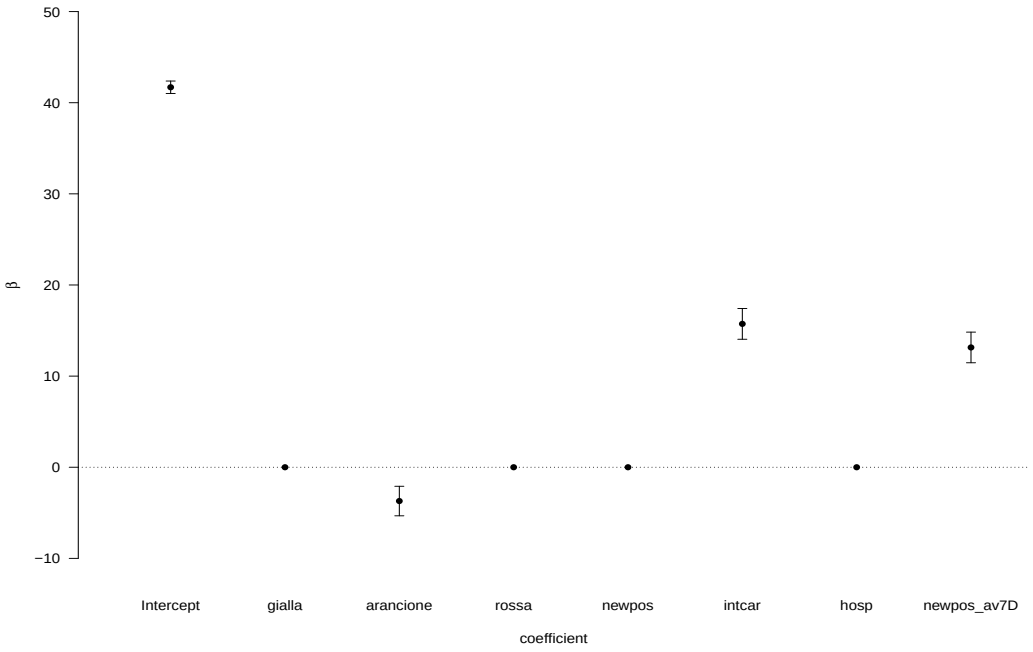

Figure 11: 95% credible intervals for the coefficients in the **intcarH8** model

3.2 Sensitivity analysis

Although we chose to be guided completely by the data, we could still use a g-prior and tune the hyperparameter α by choosing the MSE as the criterion. We use the same models as before, so BMA with **hospH8** and the best model with 3 covariates for **intcarH8**.

In both cases the MSE is high for low values of α. However, it decreases by increasing the value of α until it converges. We settle on $\alpha = 100$, as used in Subsection 2.1.

Since there is no substantial difference in performance we kept using the non-informative prior with `BIC` in Subsections 2.3 and 3.1.

4 Conclusion

In conclusion we see that, given the evaluation based on MSE, the best models for predicting the target variables are the ones obtained via model selection using BIC.

An interesting thing to notice about these models is how different the covariates' inclusion probabilities are (see Table 1 and 2). In fact we see that for predicting the target **hospH8** the model cares whether the region is a red zone or not while for **intcarH8** the attention is shifted on the orange zone. This seems counterintuitive but if we take into account the fact the primary goal of the restrictions during the pandemic, here described by the color of the region, was to save hospitals from being overcrowded, it actually makes sense.

Appendix A Additional analysis

A.1 Frequentist Linear Regression

In a frequentist setting we consider the following model

$$y = \boldsymbol{\beta}^T X + \varepsilon$$

where

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

We start by fitting the model selected in the Bayesian setting for **intcarH8**:

$$\textbf{intcarH8} = \beta_0 + \beta_1 \textbf{arancione} + \beta_2 \textbf{intcar} + \beta_3 \textbf{newpos\_av7D} + \varepsilon$$

We fit the model with the training set and obtain an adjusted R^2 of 0.9462. We recall that R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(\overline{y} - \hat{y}_i)}$$

where $\overline{y}$ is the target mean in the training set, and $\hat{y}_i$ is our prediction of the i-th sample. Adjusted R^2 simply accounts for the degrees of freedom (number of covariates). Figure 12 shows the diagnostics plots for the model. As we can see the residuals are clearly not Gaussian, and as a matter of fact a Shapiro test for normality returns a p-value close to zero. The residuals v. fitted plot seems to be random enough, and it appears the model has captured the variability of the data well enough. The MSE on the test set is 36.36943.

Lastly, we test the hypothesis $H_0 : \ \beta_i = 0 \ \ v. \ \ H_1 : \ \beta_i \neq 0$ for each coefficient. We reject each hypothesis at any significance level since all p-values are equal to zero, meaning that all the covariates contribute significantly to the model.

For predicting **hospH8** we choose the best model found in the Bayesian case, to emulate the aggregate model we would have to resort to ensemble methods, which we avoid for simplicity.

$$\textbf{hospH8} = \beta_0 + \beta_1 \textbf{rossa} + \beta_2 \textbf{hosp} + \beta_3 \textbf{newpos\_av7D} + \varepsilon$$

We obtain an adjusted R^2 of 0.9836. Figure 13 shows the diagnostics plots for the model. Again, the residuals do not pass a normality test, and we reject the null hypothesis for every covariate. We also do not see a complete random pattern in the residuals v. fitted plot, meaning there could be some complexity in the data that our model did not capture. The MSE on the test set is 752.7621.

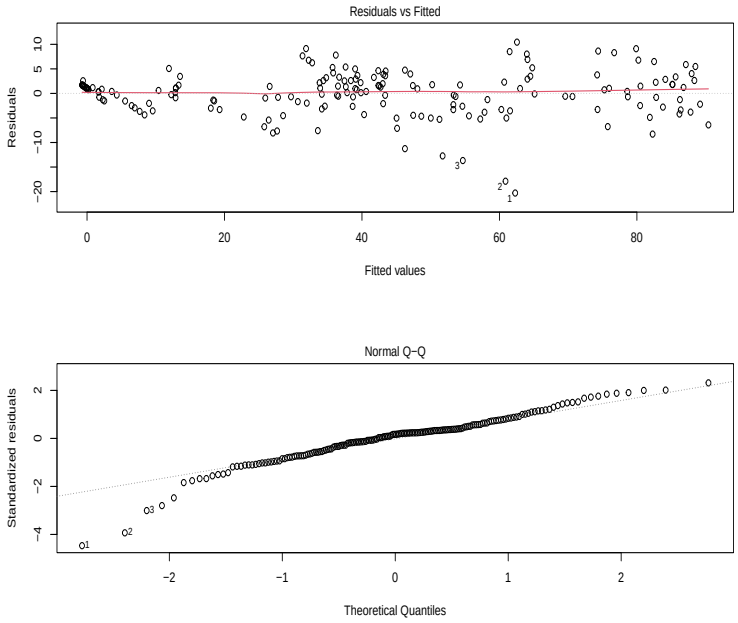

Figure 12: Diagnosis plots of the frequentist model for **intcarH8**

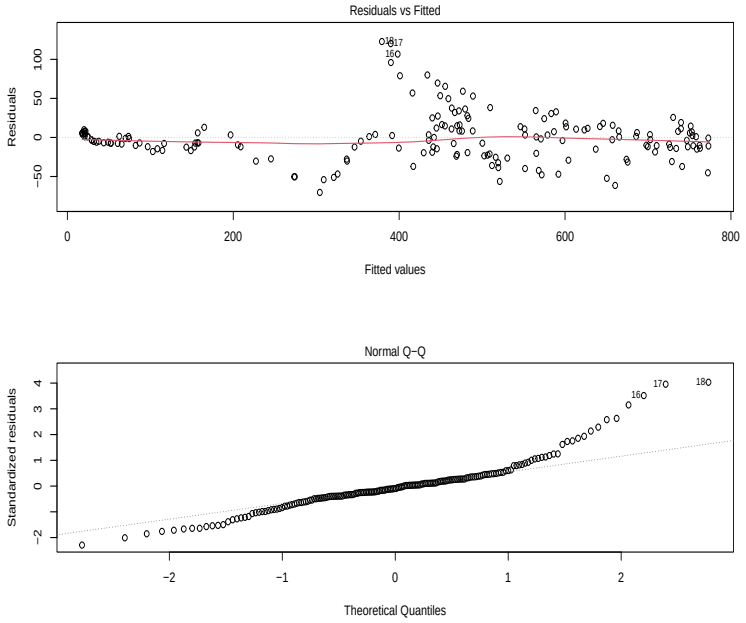

Figure 13: Diagnosis plots of the frequentist model for **hospH8**

A.2 Different representations of some covariates

Color as a numerical variable

This model considers the **color** covariate as an ordinal variable, note that this model uses the non-informative prior with BIC as the model selection criterion. The new covariate starts from 1 (*Bianca*) and ends in 4 (*Rossa*).

For **intcarH8** the model becomes:

$$\textbf{intcarH8} = \beta_0 + \beta_1 \textbf{color} + \beta_2 \textbf{newpos} + \beta_3 \textbf{intcar} + \beta_4 \textbf{hosp} + \beta_5 \textbf{newpos\_av7D} + \varepsilon$$

which yields a MSE of 25.85215, slightly worse than the error of the other models. The same goes for **hospH8**, which yields a MSE of 1019.115, so we decide to drop this variant of our model.

Season

Here we transform the variable **day** into **season** and include it in the model.

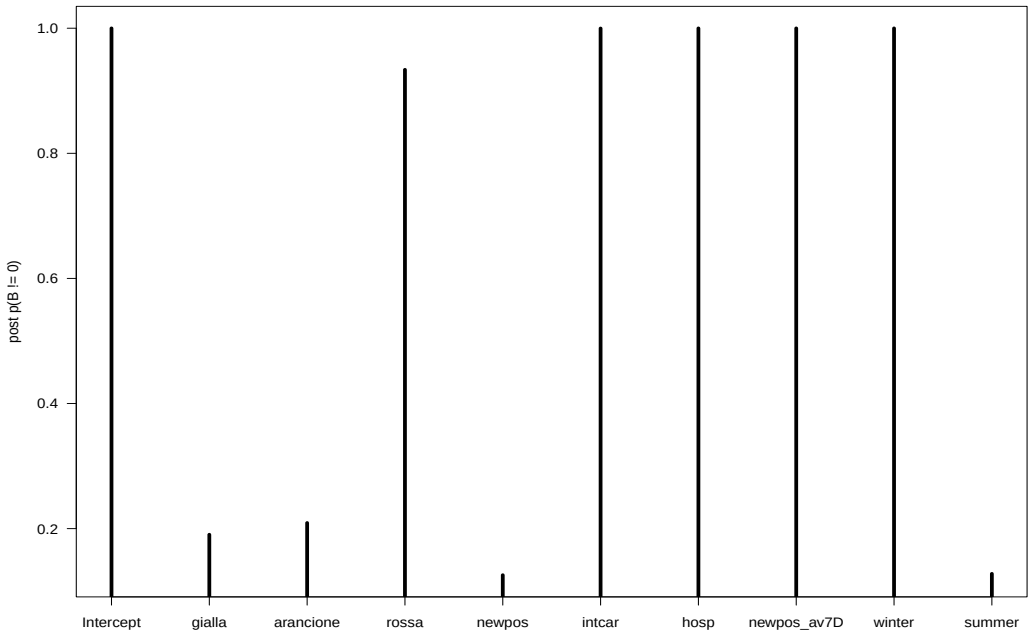

Figure 14: Feature inclusion probabilities for **hospH8**.

As we can see in Figure 14, the **winter** indicator variable for the season (with target **hospH8**) seems to be relevant. Indeed the MSE of this model is 781.8348, which is the best one we achieved.

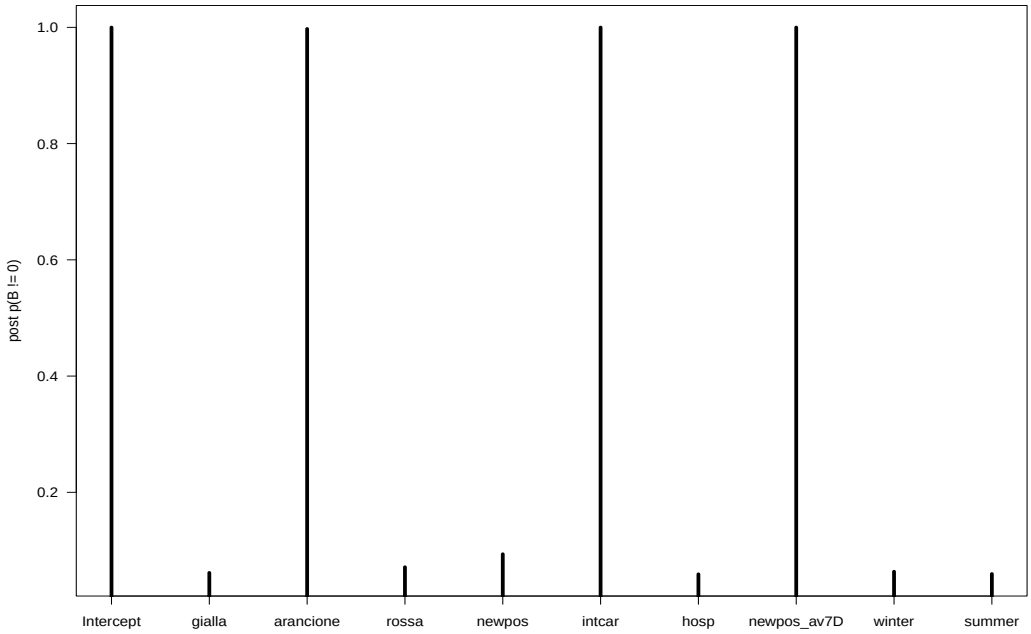

Figure 15: Feature inclusion probabilities for **intcarH8**.

Instead in Figure 15 we can see that for the prediction of **intcarH8** the season doesn't seem relevant. Indeed the MSE results 23.0993, which is barely better than the other models without the season.