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.

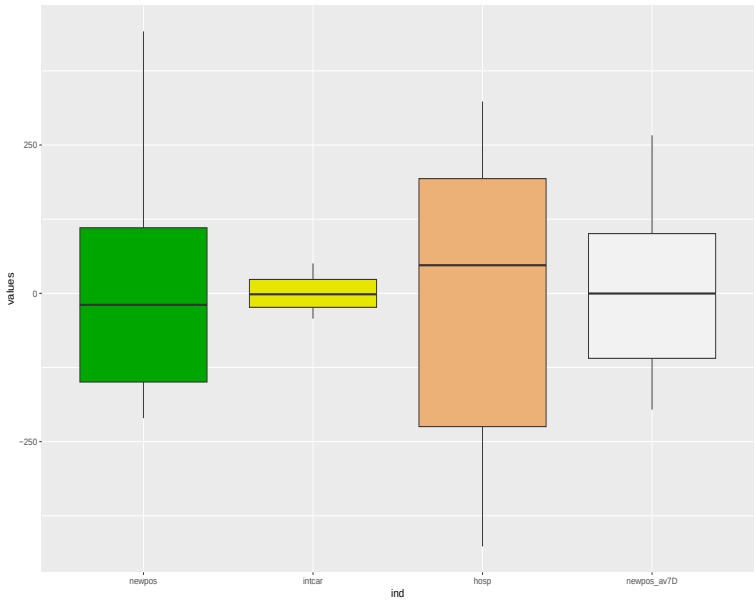

Figure 3: Boxplot of the mean-centered features

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

Figure 5 shows the correlation matrix for each region.

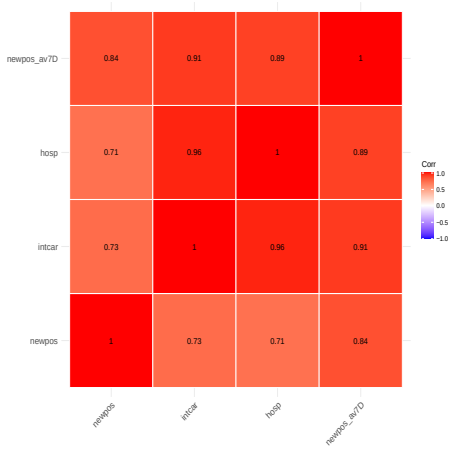

Figure 4: Correlation matrix of the features

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.

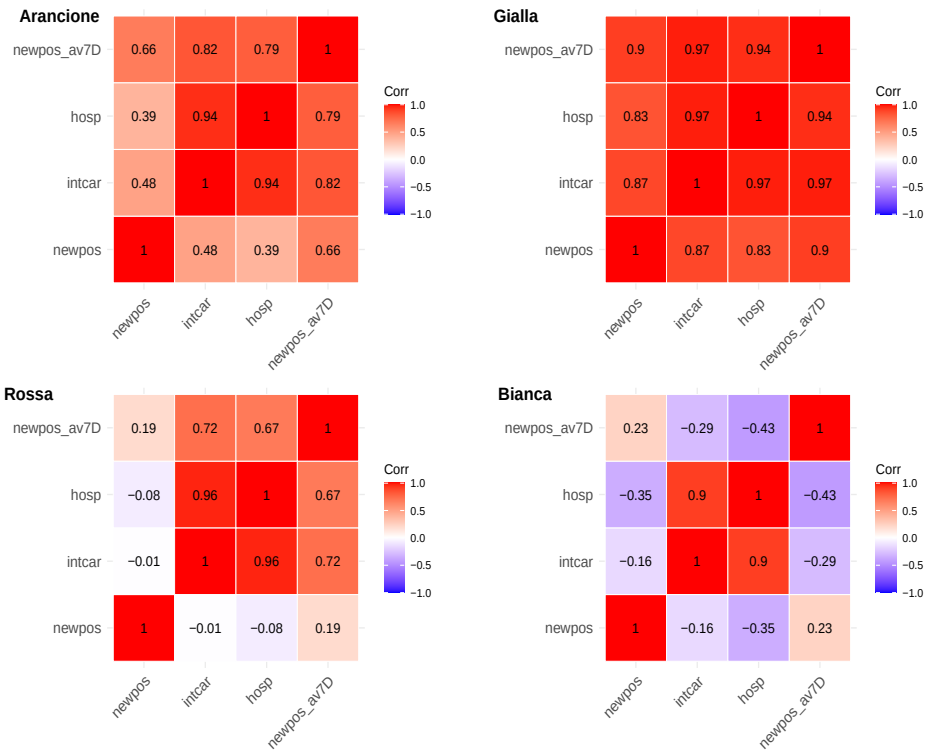

Figure 5: Correlation matrix of the features per region

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

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

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, which was transformed into 3 binary covariates via one-hot encoding. The model is

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

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]

Figure 6 shows the posterior distributions of the coefficients.

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

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

We use the mean square error MSE

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

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

Coefficient	Posterior mean	Posterior standard deviation
Intercept	-0.004298	0.012938
gialla	-0.092490	0.058380
arancione	-0.166263	0.070069
rossa	-0.134449	0.072101
newpos	0.147965	0.019453
intcar	0.057368	0.053409
hosp	0.841135	0.057545

Table 1: Posterior means and standard deviations of the first model

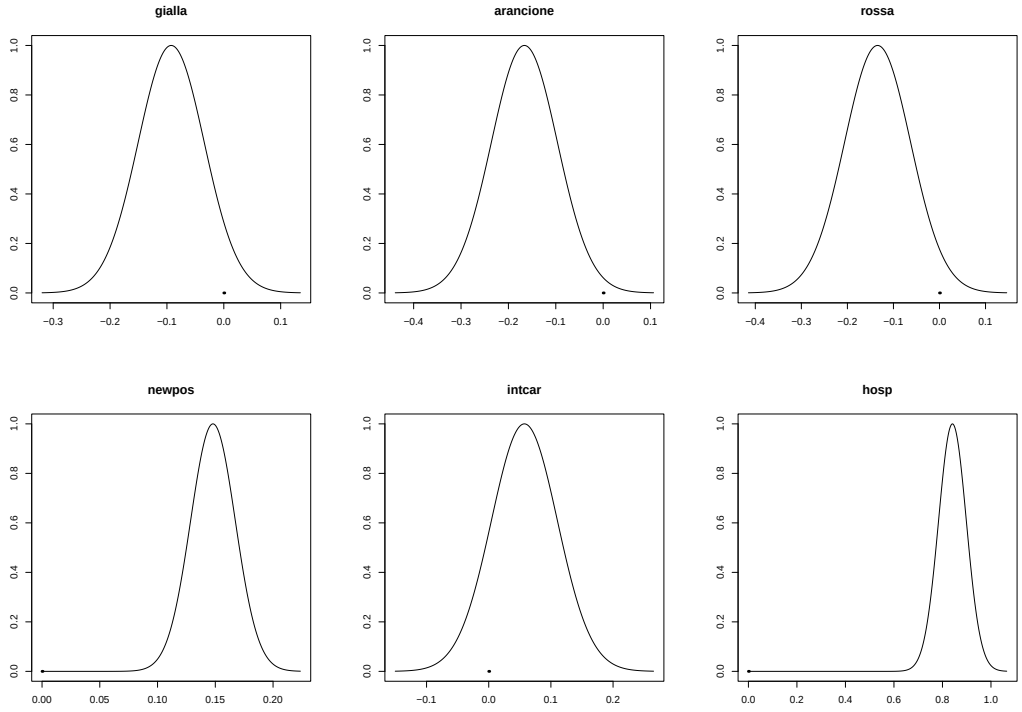

Figure 6: Posterior distributions of each coefficient

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 0.02659245.

For the target variable **intcarH8** we consider an almost identical model

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

The MSE is 0.07319181. 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.

Another model we tried was one that considered the **color** feature as an ordinal variable, note that we used the same prior. The covariate starts from 1 (*Bianca*) and ends in 4 (*Rossa*). For **intcarH8** the model becomes

$$\begin{aligned}\textbf{intcarH8} = \beta_0 + \beta_1\textbf{color} + \beta_2\textbf{newpos} + \beta_3\textbf{intcar} \\ + \beta_4\textbf{hosp} + \varepsilon\end{aligned}$$

which yields a MSE of 0.07816081, slightly worse than the error for the previous model. The same goes for **hospH8**, this is consistent across multiple shufflings and retrainings of the model, so we decide to drop this variant of our model.

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

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

However, we start by setting α to infinity, so the data completely determines our model. We also add the covariate **newpos_av7D**, and also perform model selection and set a uniform prior on all possible models that use a subset of these covariates.

We obtain the following posterior probabilities for each covariate

Coefficient	Posterior
Intercept	1.0000000
gialla	0.1777365
arancione	0.1728493
rossa	0.9542922
newpos	0.1653874
intcar	0.3617246
hosp	1.0000000
newpos_av7D	1.0000000

Table 2: Probabilities of each covariate in the second model

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

Using the same basis for our model, we obtain the following probabilities for the target variable **intcarH8**

Coefficient	Posterior
Intercept	1.0000000
gialla	0.1136039
arancione	0.9989177
rossa	0.0934188
newpos	0.1316067
intcar	1.0000000
hosp	0.1067579
newpos_av7D	1.0000000

Table 3: Probabilities of each covariate in the second model for **intcarH8**

We observe that the most "useful" covariates instead are **arancione**, **intcar**, and **newpos_av7D**.

2.2 Model selection

In the case of **hospH8** the top 5 models have the following posterior probabilities

Models	Model 1	Model 2	Model 3	Model 4	Model 5
Posterior probability	0.3755	0.2975000	0.05410000	0.0429000	0.04250000

Table 4: Posterior probabilities of the best 5 models for **hospH8**

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

Since none of the models have a posterior probability above 50%, we do not choose the best model, which in this case is model 1, but we use Bayesian model averaging (BMA), the posterior probability of a quantity of interest Δ, computer 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 then compute our predictions as a weighted average of each model's pre-

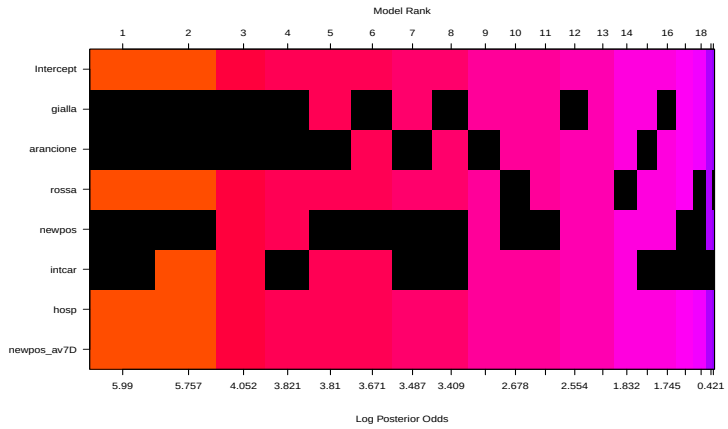

Figure 7: Model ranking and posterior probabilities

diction, 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 0.01251516, sligthly better than the first model with all the covariates.

The ranking for the models predicting **intcarH8** is shown in Figure 8

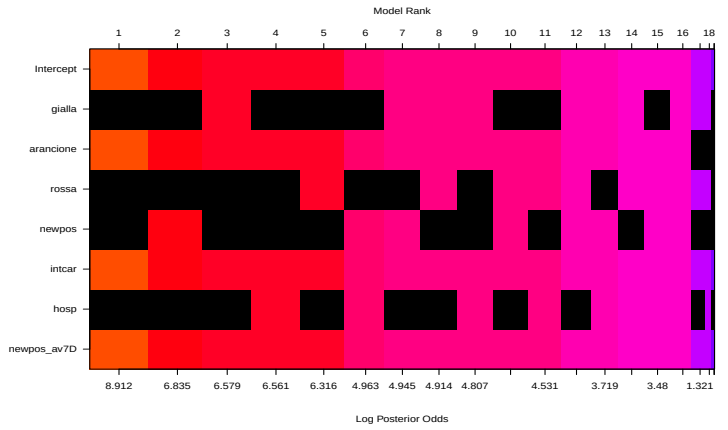

Figure 8: Model ranking for **intcarH8**

For **intcarH8** the best model has a posterior probability of 65%, 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 0.04632954[2], better than the one using all the covariates.

We also added a categorical covariate **season**, as mentioned in the first section. However, it yielded slightly worse results in terms of MSE (for **intcarH8** 0.0495), so we decided to drop it.

[2]Note that we judge the improvement in terms of the MSE over a large number of shufflings and iterations

3 Posterior analysis

3.1 Posterior distributions

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

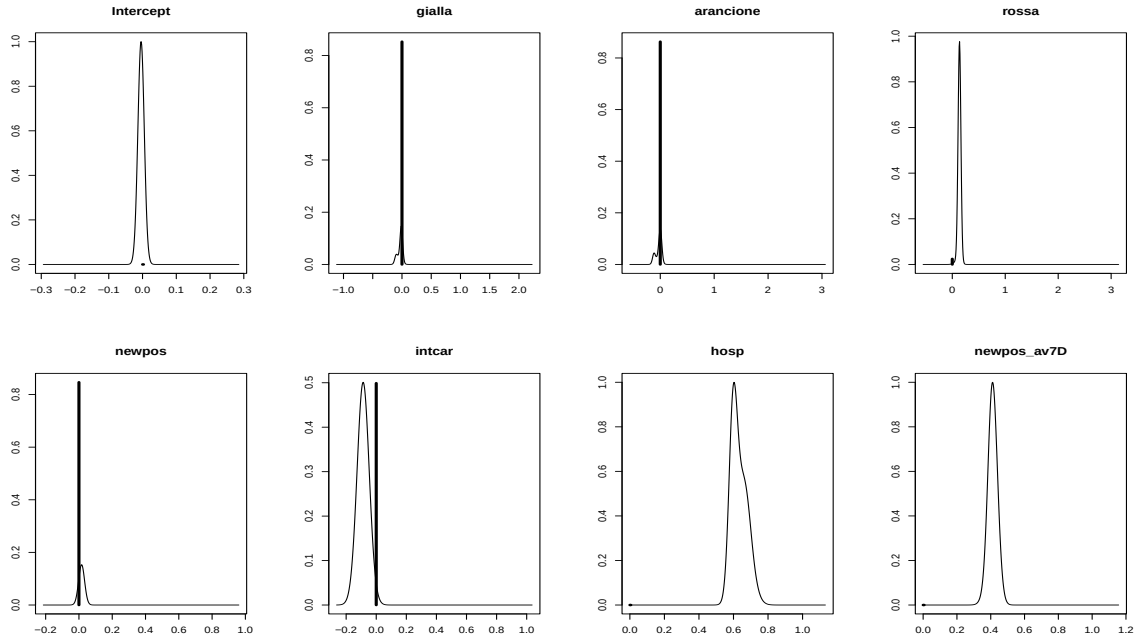

Figure 9: Posterior distributions of coefficients for best **hospH8** model

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**. We also show the posterior means, standard deviations, and 95% credible intervals in Table 5.

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**.

Lastly, Table 6 shows the posterior means, standard deviations, and 95% credible intervals.

Coefficient	2.5%	Post. mean	97.5%	Post. std. dev.
Intercept	-0.022873976	-0.004298	0.01451821	0.009491
gialla	-0.044262972	-0.004576	0.02548325	0.019832
arancione	-0.041202061	-0.003878	0.04235637	0.023008
rossa	0.070003523	0.131494	0.20234863	0.035352
newpos	-0.001460963	0.002445	0.03236543	0.009053
intcar	-0.140785618	-0.043676	0.00000000	0.052065
hosp	0.555651734	0.631584	0.72833902	0.046662
newpos_av7D	0.356364413	0.411558	0.47048409	0.028693

Table 5: Posterior means and standard deviations of the BMA model

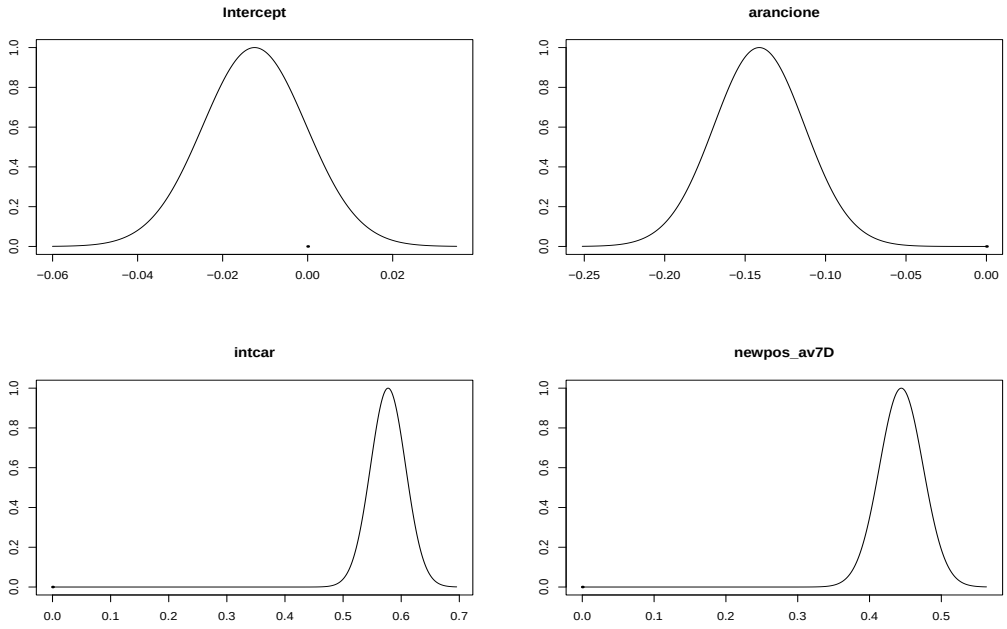

Figure 10: Posterior distributions of coefficients for best **hospH8** model

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 model with 3 covariates for **intcarH8**

Covariate	2.5%	Post. mean	97.5%	Post. std. dev.
Intercept	-0.03659953	-0.01249	0.01162539	0.01222
arancione	-0.19703226	-0.14122	-0.08541724	0.02828
intcar	0.51756701	0.57743	0.63728508	0.03033
newpos_av7D	0.38412313	0.44426	0.50438846	0.03047

Table 6: Posterior means and standard deviations of the BMA model

In both cases the MSE is high for low values of α, around 1 for both targets. However, it decreases until it converges to 0.0127 for **hospH8** and 0.0495 for **intcarH8**. We settle on $\alpha = 100$.

Since there is no substantial difference in performance we kept setting α to infinity, by using the `BIC` prior on BAS.

3.3 Comparison with frequentist linear regression

4 Conclusion