POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

# Analysis of Covid Dataset through Bayesian Regression

Author(s): **Stefano Baroni**

**Alessia Menozzi**

Academic Year: 2022-2023

# Contents

# 1 | Introduction

The aim of our project is to make predictions for the number of patients expected in hospital and ICU seven days ahead. To accomplish this we'll use a Bayesian regression approach and build linear regression models.

## 1.1. Dataset Description

Out dataset has 205 observations and 10 columns.

The columns are :

- **X** : a unique number associated to each observation.
- **color** : a character indicating the color of the zone, it can be "Arancione", " Bianca", "Gialla", "Rossa". We will need to change this variable and turn it into categorical to use it for our models
- **newpos** : a number indicating the detected COVID positive subjects
- **intcar** : number of patients in intensive care
- **hosp** : number of patients in hospital
- **newpos_av7D** : average number of detected COIVD positive subjects over the previous 7 days
- **day** : current day.
- **hospH8** : number of patients in hospital (7 day head)
- **intcarH8** : number of patients in intensive care (7 day head)
- **dayH8** : (day +7)

## 1.2. Data Processing

We removed the column **X** since it is not needed for our analysis.

Regarding the variable **color** we performed *one-hot encoding* by splitting it into 4 different columns, one for each color, and assign a binary value for each of them. After this process of binarization, we removed the column from the dataset since it was redundant.

Finally the dataset has been normalized, in order to provide the same scales for all the covariates, and shuffled, to create representative folds for the cross validation.

## 1.3. Data Analysis

Before starting with our bayesian analysis we investigated the distributions and relationships between our covariates.
We performed the analysis on: **intcar**, **hosp**, **newpos_av7D** and **newpos**.

### BoxPlot

From the boxplots we can see, for each covariates, the interquantile range ("box") formed by the 25% percentile, the median and the 75% percentile. The whiskers instead represent the remaining distribution.

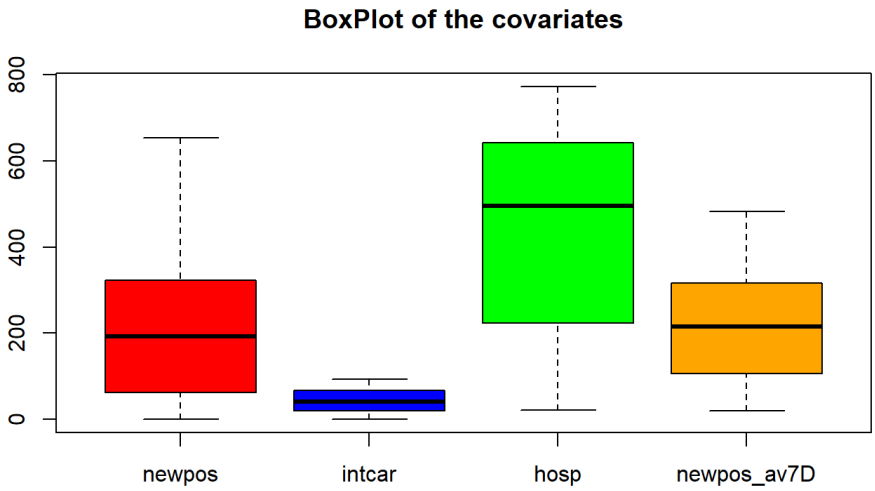


Figure 1.1: BoxPlots of the covariates

We can clearly see how some covariates varie more than others. Due to this fact we decided to normalize the data in order not to let the scale of covariates like *hosp* influence other covariates that varies less and dominate the learning process. With normalization all the variables are on a consistent scale, ready to be better interpreted and compared.

**Scatterplot Matrix**

Then we performed an analysis regarding the data correlations based on the attribute *color* seeing if any value has a stronger influence on the target variables.
As we could logically think, the color *Bianca* is the one having a different behaviour (since we can remember that during the pandemic the disease spread was more stable in that period.)
We can now take a look at the underlying *pairplots* that represent the relationships between all the variables, including the targets *hospH8* and *intcarH8*. We are only showing the scatter plot filtered by the attribute *color = 'Bianca'*. (The others an can be found in chapter A.1)

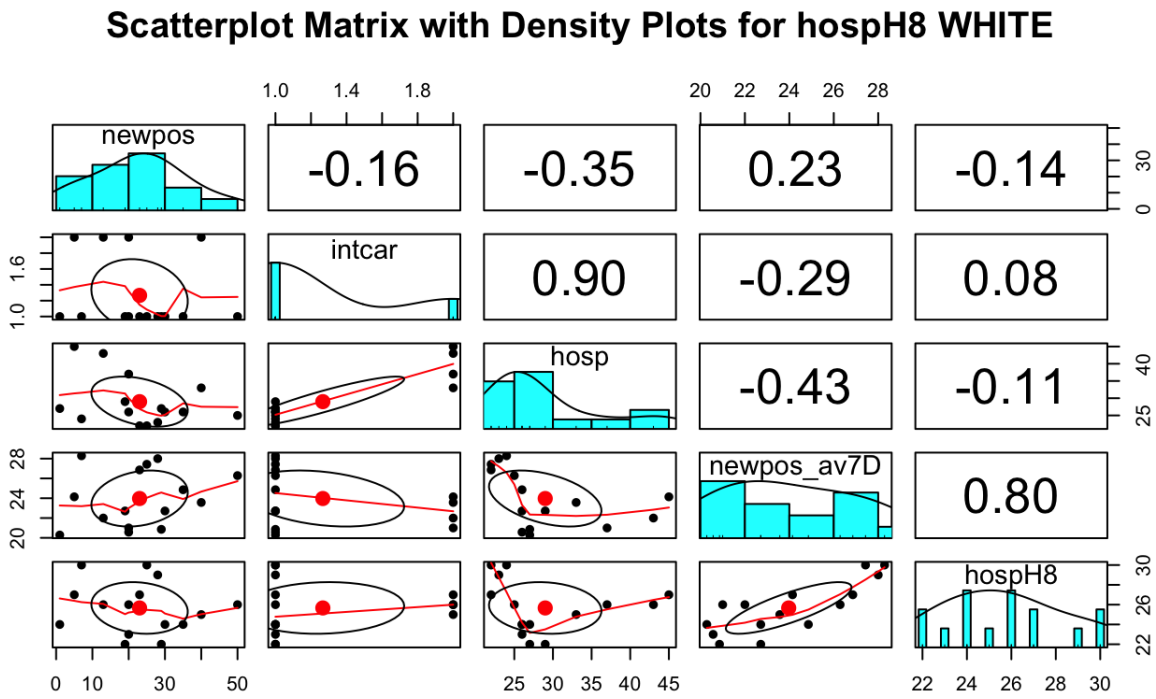


Figure 1.2: Scatterplot matrix with observations for hospH8 when color='Bianca'

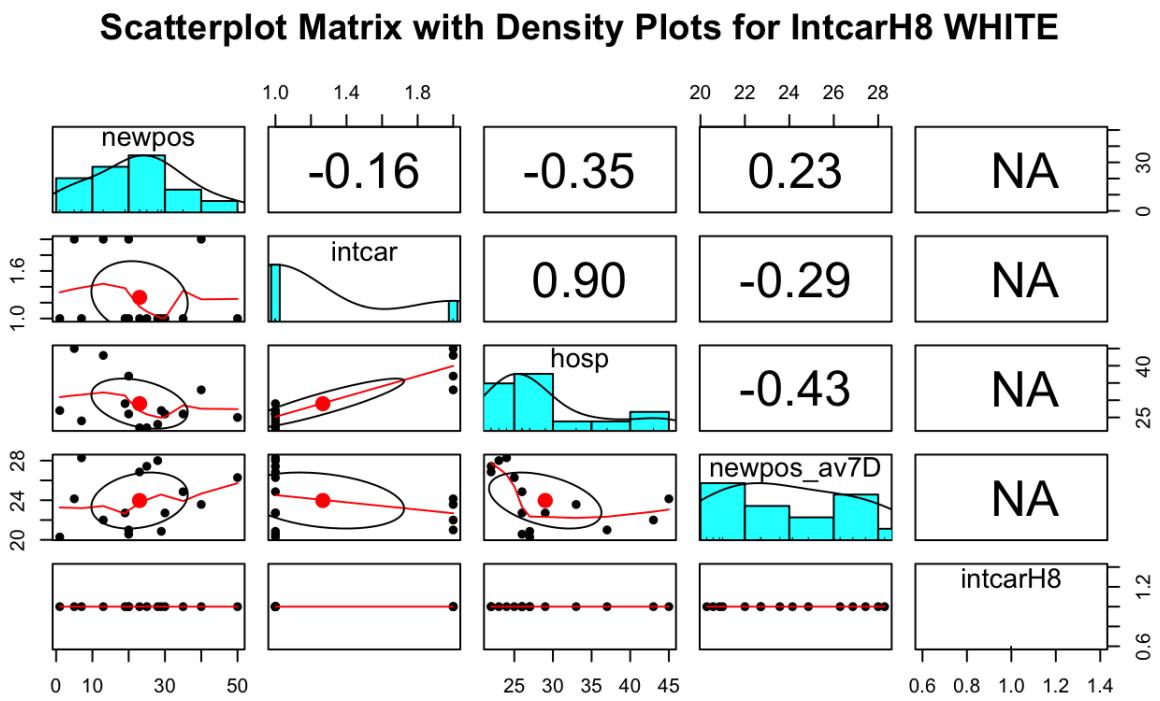


Figure 1.3: Scatterplot matrix with observations for intcarH8 when color='Bianca'

Looking at the last line of the matrix we see that it shows a constant line, this means that the variable *IntcarH8* has constant value for all data points. While in Figure 1.2 we see that *HospH8* doesn't have a linear relationship with the other covariates. In a regression analysis, a variable with little or no variation may not be very useful because it doesn't contribute to explaining the variation in the target variable. This is the reason why we decided not to perform predictions and not to include in our model the observations with attribute $Bianca = 1$ but to keep only the other colors since they are more significantly related to our target variables (see chapter A.1).

## Correlation Matrix

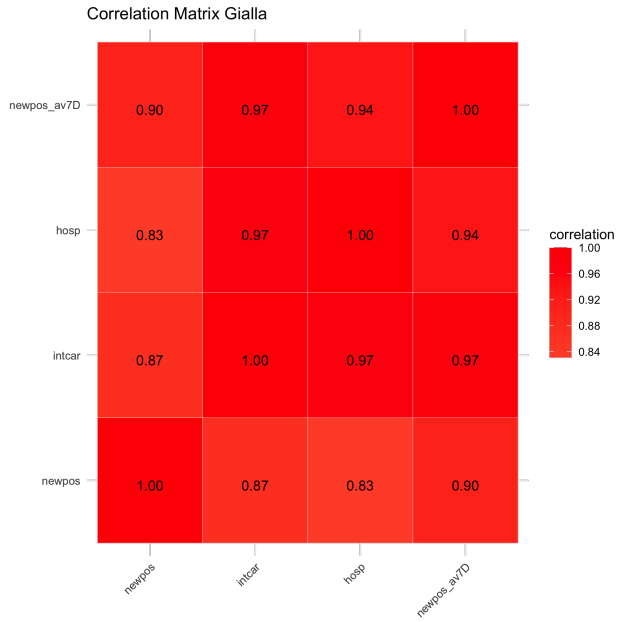


Figure 1.4: Correlation Matrix 'Gialla'

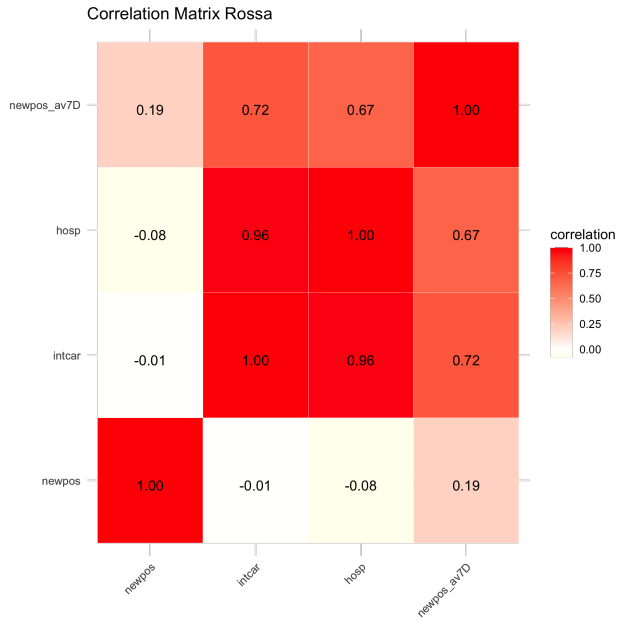


Figure 1.5: Correlation Matrix 'Rossa'

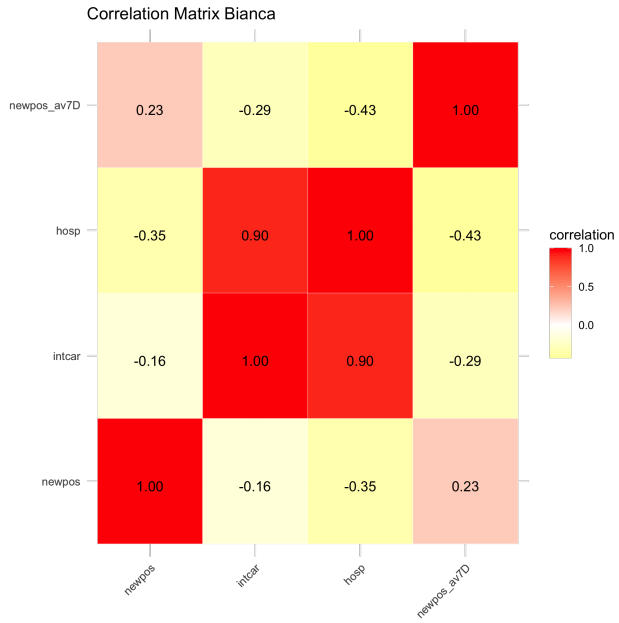


Figure 1.6: Correlation Matrix 'Bianca'

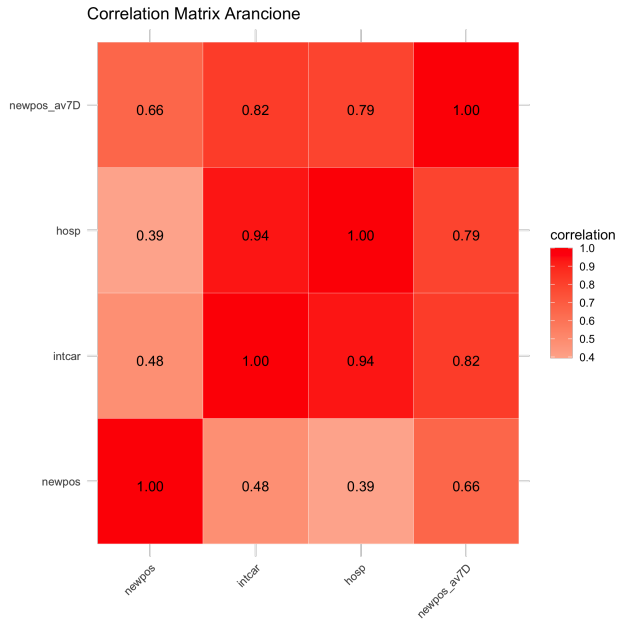


Figure 1.7: Correlation Matrix 'Arancione'

These correlation matrices help us identifying strong and weak correlations between the variables *newpos*, *hosp*, *intcar*, *newpos_av7D*. When building predictive models, understanding correlations helps in selecting relevant features, since some variables may be redundant or highly correlated with each other and can affect the performance of our regression models. The correlation coefficient ranges from -1 to 1. A value of 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation.

We see that for *Gialla* there are variables that have coefficient 0.97 (intcar-hosp and intcar-newpos_av7d), and 0.94 (hosp-newpos_av7d). This indicates a very strong positive relationship and can tell us that they could provide similar information. For *Rossa* there are low correlations meaning that all the coviariates here would bring to my model important information. In Figure 1.6 we see that *Bianca* shows a different trend with negative correlations between some variables. This ensures we are making a good choice when removing it from the models since its behaviour is too different from the other colors.

# 2 | Model Decision

In this chapter we will test some models characterized by different priors.

## 2.1. Prior and likelihood Definition

In order to proceed with a Bayesian analysis we first need to the define the prior distributions of our covariates and the likelihood function. With this two elements we are able to compute the posterior distribution of our covariates and then perform predictive analysis. For what concerns the likelihood we use a Gaussian prior of order n:

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

For the prior instead:

$$(\beta, \sigma^2) \sim \pi(\beta, \sigma^2)$$

In the next sections we will discuss different priors that we have tested showing how each of them behaves, in order to chose the optimal one. To evaluate the different priors we considered two different models for the two different targets:

$$\begin{aligned} \boldsymbol{hospH8} =& \beta_0 + \beta_1 \boldsymbol{gialla} + \beta_2 \boldsymbol{arancione} + \beta_3 \boldsymbol{rossa} + \beta_4 \boldsymbol{newpos} \\ &+ \beta_5 \boldsymbol{intcar} + \beta_6 \boldsymbol{hosp} + \beta_7 \boldsymbol{newpos\_av7D} + \epsilon \end{aligned} \tag{2.1}$$

$$\begin{aligned} \boldsymbol{intcarH8} =& \beta_0 + \beta_1 \boldsymbol{gialla} + \beta_2 \boldsymbol{arancione} + \beta_3 \boldsymbol{rossa} + \beta_4 \boldsymbol{newpos} \\ &+ \beta_5 \boldsymbol{intcar} + \beta_6 \boldsymbol{hosp} + \beta_7 \boldsymbol{newpos\_av7D} + \epsilon \end{aligned} \tag{2.2}$$

Where $\epsilon$ is a pure error term used to introduce noise and with distribution defined as:

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

We can see that the models contain all the covariates. They are only used to test different priors to see which one performs better. To evaluate each model, we perform a prediction and then print the **mean squared erro**r computed as the weighted sum of the squared difference

between the prediction and the true value:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.3}$$

In order to make the analysis accurate and to avoid overfitting we performed the tests using the k-fold cross validation method. This method consists in dividing the dataset in k different folds and use one of them each time as the validation set. Once all the k-MSEs are computed, they are averaged and then the final MSE is shown. To properly apply this method we had to shuffle the data because similar data were placed one after the other inside the dataset. In this way the k randomly generated folds, used at each iteration as the test set, can be considered to be well representative of our dataset. Although it is also important to notice that the process of shuffling the data is random, so a specific seed has been used to allow reproducibility and those results may vary a little bit depending on how the data get shuffled.

For all the models explained below we used as modelprior *Bernoulli(1)*, which only generates one model that includes all the variables.

## 2.2. Zellner g-prior

The Zellner g-prior is an informative prior that tries to balance, through the hyperparameter alpha, the information retrieved from both a base model (an horizontal line for alpha = 0) and the data. As alpha grows the prior becomes less informative and retrieves more information from the data. The prior is defined as:

$$\beta = (\beta_1, \ldots, \beta_k)|\sigma^2 \sim \mathcal{N}_k(0, \alpha\sigma^2(\tilde{X}_t\tilde{X})^{-1})$$

$$\sigma^2|X \sim \pi(\sigma^2) = \sigma^{-2}$$

We can then create our two models and evaluate the results:

| Alpha | MSE |
|---|---|
| 0.01 | 56381.7 |
| 0.1 | 47666.3 |
| 1 | 15016.7 |
| 10 | 1337.12 |
| 100 | 878.44 |
| 200 | 874.54 |

Table 2.1: MSE for g-prior HospH8

| Alpha | MSE |
|---|---|
| 0.01 | 761.06 |
| 0.1 | 645.1 |
| 1 | 210.54 |
| 10 | 28.17 |
| 100 | 21.97 |
| 200 | 21.92 |

Table 2.2: MSE for g-prior IntcarH8

It is clear that with higher alphas we obtain a lower MSE meaning that our distribution was not really informative and an application of a non-informative prior could be a better choice.

## 2.3. Zellner-Siow prior

The Zellner-Siow prior differs from the previous for the fact that now alpha is no longer considered a hyperparameter but instead it's a parameter with it's own distribution. Considering that alpha is now influenced by the data and it's not fixed introduces more flexibility but also an additional level of uncertainty into the model.

$$\beta = (\beta_1, \ldots, \beta_k)|\sigma^2, \alpha, X \sim \mathcal{N}_k(0, \alpha\sigma^2(X_t X)^{-1})$$

$$\sigma^2|X, \alpha \sim \pi(\sigma^2) = \sigma^{-2}$$

$$\frac{1}{\alpha} \sim \pi_0 = \text{Gamma}\left(\frac{1}{2}, \frac{n}{2}\right)$$

## 2.4. Non-informative prior

To simulate a non-informative prior we specified in the model attribute *prior* = BIC.
BIC can be considered a non-informative prior because it doesn't introduce any prior knowledge about the distribution of the model parameters (all parameter values have the same distribution a priori).

## 2.5. Prior Comparison

Finally we can analyze the MSE obtained using the three different priors:

| prior | MSE |
|---|---|
| g-prior alpha= 200 | 874.54 |
| Zellner-Siow | 873.38 |
| BIC | 873.39 |

Table 2.3: MSE for different priors HospH8

| prior | MSE |
|---|---|
| g-prior alpha= 200 | 21.916 |
| Zellner-Siow | 21.896 |
| BIC | 21.895 |

Table 2.4: MSE for different priors IntcarH8

We notice that for both the target variables the model that returns the lowest error is the one created with *modelprior='BIC'*.

## 2.6. Model Selection

In this section we'll discuss the model selection performed using the Bayesian information criterion, or BIC, the most popular criteria in the Bayesian perspective. It works that, when specifying a BIC prior in a model, we are penalizing the model's complexity. This means that models with more parameters will be penalized compared to those with fewer parameters unless the addition of parameters significantly improves the model's fit to the data.

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

When creating the model with BIC as prior, the modelprior used is the *uniform()* distribution which assigns equal probabilities to all models.

To compare the models returned, we look at the logmarg information. We use this information to retrieve the model with the largest log of marginal likelihood, which corresponds to the model with the smallest BIC. In this way we find the most influent variables which are the ones included in the best model.

These are:

**- For HospH8**

$$\boldsymbol{hospH8} = \beta_0 + \beta_1 \boldsymbol{rossa} + \beta_2 \boldsymbol{intcar} + \beta_3 \boldsymbol{hosp} + \beta_4 \boldsymbol{newpos\_av7D} + \epsilon \tag{2.4}$$

For this target variable, we found that the first model has PostProbs = 0.415, the second 0.346 and the others lower than 0.044.
The following figures give some insights about the model obtained:

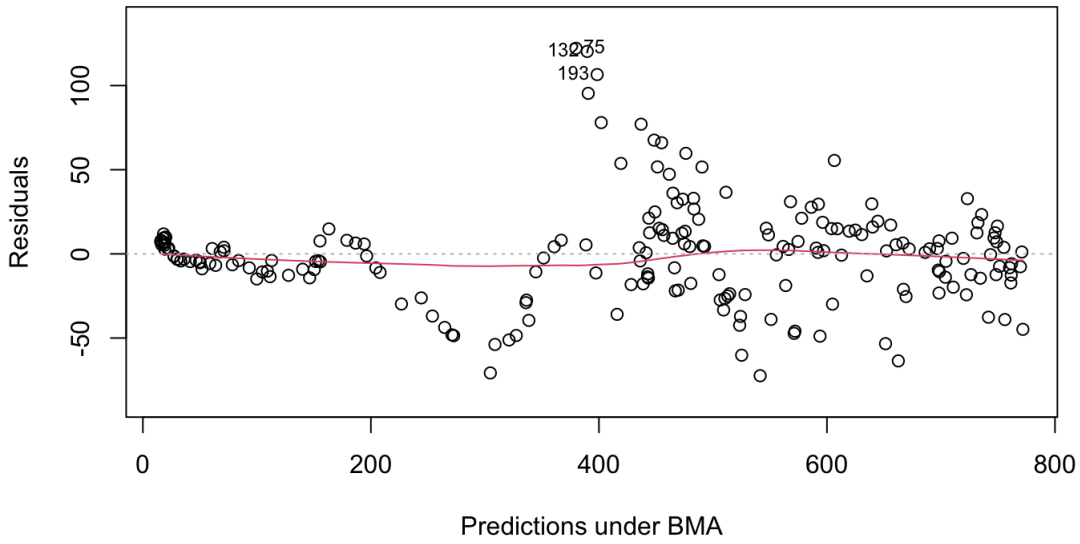


Figure 2.1: Residuals Plot for HospH8

This is a plot of residuals and fitted values under BMA. Ideally, we would want to see a random scatter of points around the horizontal line at zero, showing no systematic pattern in the residuals, indicating that the linear relationship between the dependent variable and the predictors is appropriate. In our plot the majority of the points is concentrated around the red line (residuals = 0), meaning that the predictions are generally accurate. On the other hand we see that some outliers are indicated (75,132,193), these point could potentially have a significant impact on the model fit.

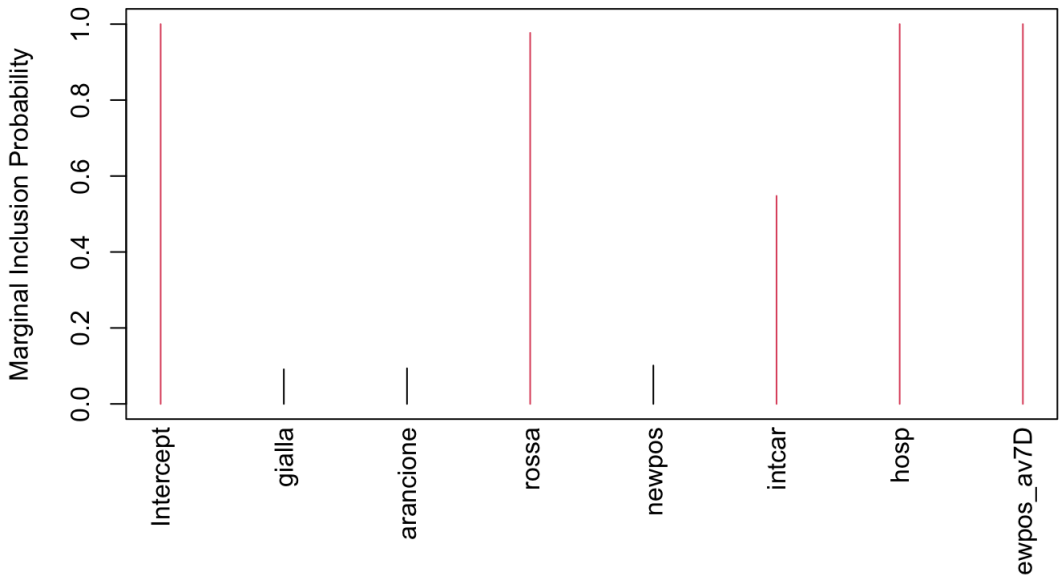


Figure 2.2: Marginal Inclusion probability for HospH8

Figure 2.2 shows the posterior probability of each covariate to be contained in the model. We see the coviarates are the same as the ones in (2.4)

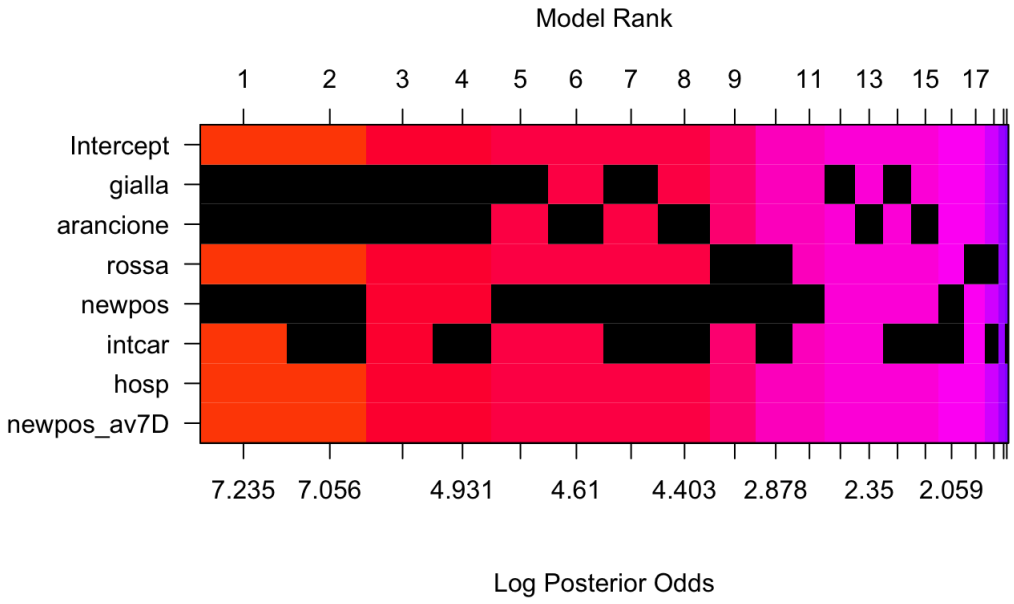


Figure 2.3: Model Ranking Matrix for hospH8

Figure 2.3 shows the covariates included in every model and it's interesting to notice how *hosp* and *newpos_ av7D* are always included in the models generated. We were expecting this since these two attributes were presented in the summary with post p(B != 0) = 1

**- For IntcarH8**

$$\boldsymbol{intcarH8} = \beta_0 + \beta_1 \boldsymbol{arancione} + \beta_2 \boldsymbol{intcar} + \beta_3 \boldsymbol{newpos\_av7D} + \epsilon \tag{2.5}$$

For this target variable, we found that the first model has PostProbs = 0.691 and the others lower than 0.116.

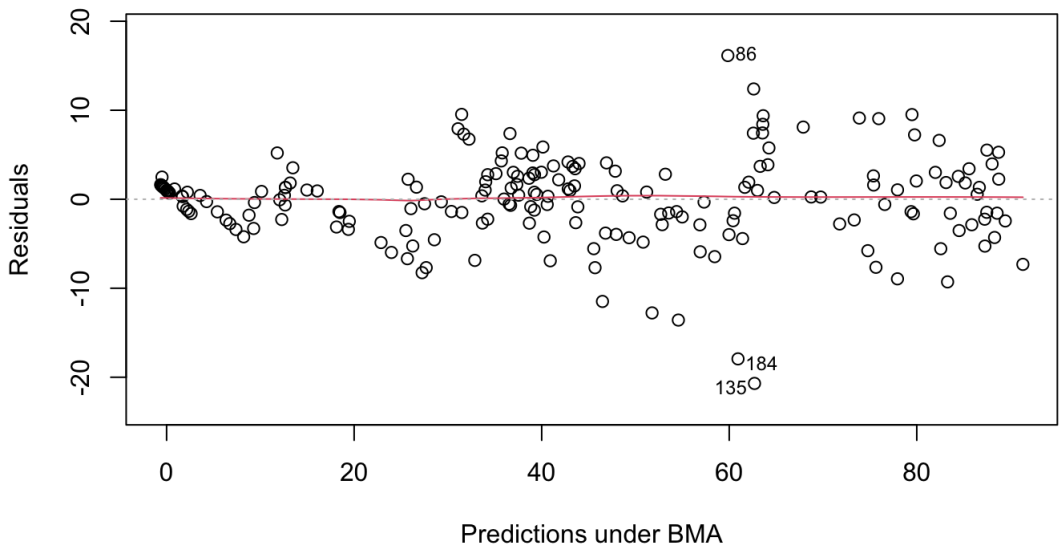


Figure 2.4: Residuals Plot for IntcarH8

Comparing this residual plot with the previous one, we see that the points are even more concentrated around the red line. This could tell us that this model is predicting with a slightly better accuracy. Still there are some points (86,184,135) marked as outliers. (Check A.2 for outlier analysis.)

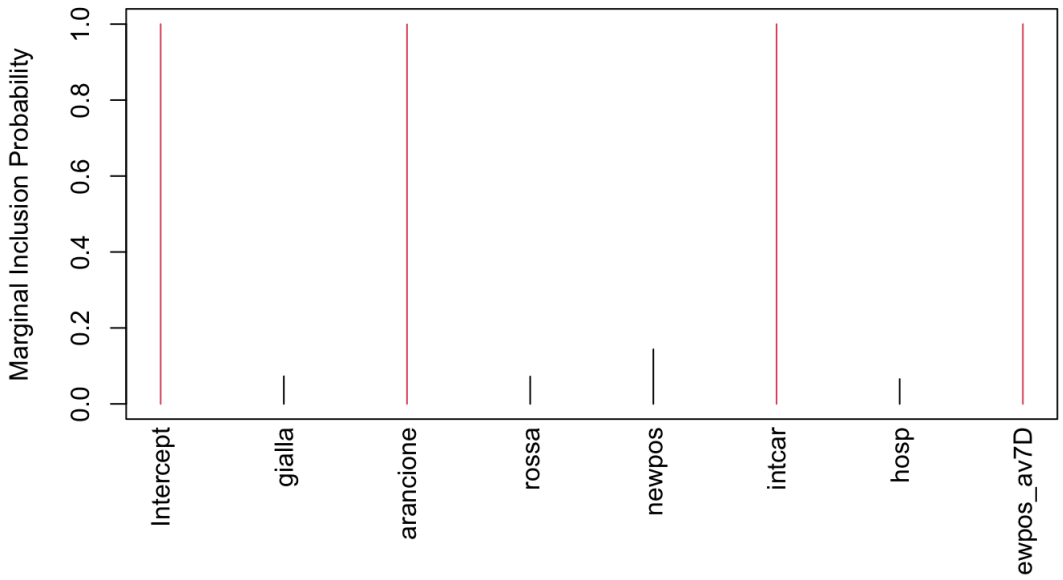


Figure 2.5: Marginal Inclusion probability for IntcarH8

In Figure 2.5 we obtain the same results as we said before when finding bestgamma using BIC. These attributes have probability = 1 to be included in the best model.

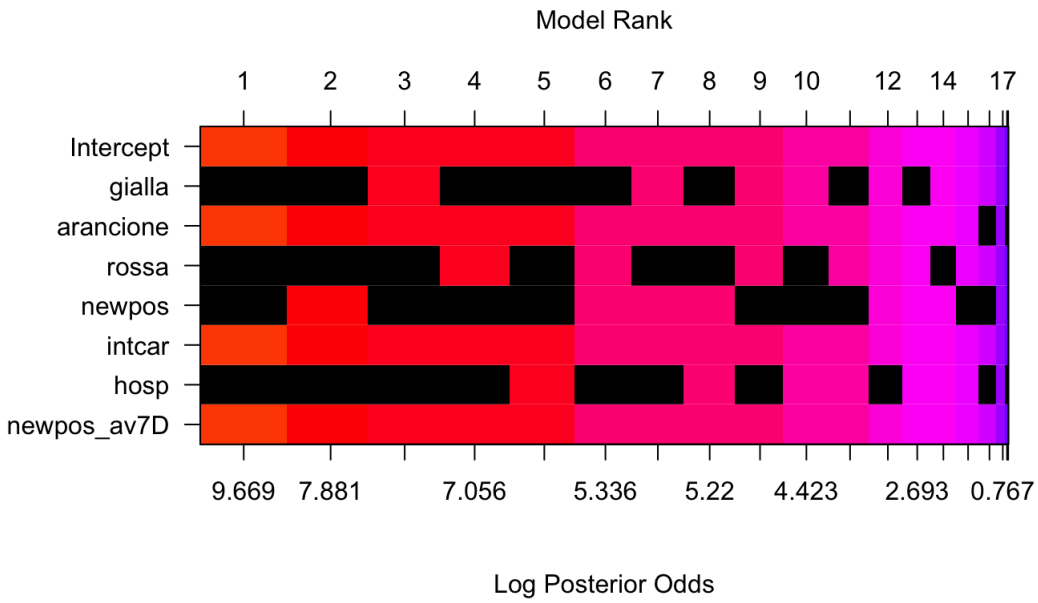


Figure 2.6: Model Ranking Matrix for IntcarH8

We see in figure 2.6 that *newpos_ av7D* and *intcar* are present in all the models, meaning they have post p(B != 0) = 1, while *arancione* is present in almost all the models (post p(B != 0)= 0.99945)

At last, given the obtained results of the two best models, we decided to test the performance of the application of BMA against using the estimator HPM that only usedthe best model.

As said HPM considers only the model with the highest posterior probability while BMA, computes the prediction as weighted sum of all the models, where the weights are the posterior probabilities of each model.

This analysis was conducted in order to understand if the best models could be used instead of considering the weighted sum of all the other models:

| Estimator | MSE |
|---|---|
| BMA | 878.80 |
| HPM | 883.71 |

Table 2.5: MSE HospH8

| Estimator | MSE |
|---|---|
| BMA | 21.94 |
| HPM | 21.86 |

Table 2.6: MSE IntcarH8

It is interesting to notice how in the case of hospH8, BMA results to be the best approach, due to the best model not having a high enough posterior probability.
In the case of intcarH8 the results are comparable and HPM performs slightly better than BMA. Obviously we have to take into account that this results may vary a little bit due to the shuffle function and the creation of folds but we decided to keep the model obtained with BMA for hospH8 and to keep HPM estimator for intcarH8.

# 3 Posterior Analysis of the covariates

In this chapter we will analyze the posterior distributions of the covariates.

## 3.1. HospH8

These plots are obtained from the best model for hospH8 which is the one that uses BIC as prior and the covariates calculated using BMA as estimator.

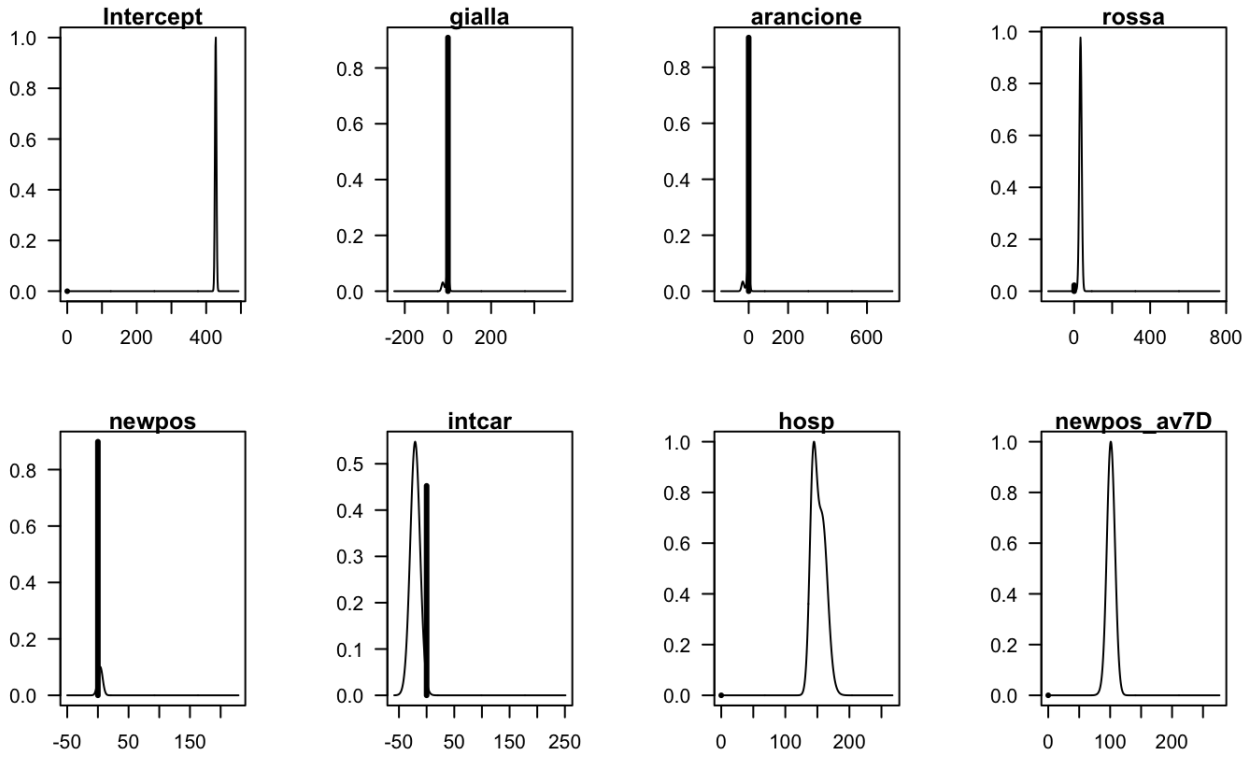


Figure 3.1: Posterior distributions of coefficients for best hospH8 model

To see if one coviarate as an impact on our target variable, the mean should not be centered

in 0. We see, as said before, that the coviariates that have the biggest influence on hospH8 are *hosp*, *newpos_ av7D* (since their mean is very far from 0). Others coviariates that have influence are *rossa* and *intcar* (but way less). This plots are useful since they give us the information about how the different covariates influence the model. For example up to now we have seen that the covariate *hosp* and *newpos_ av7D* were always included in our models. From their distribution we can also infer that the covariate *hosp*, due to its higher mean, will have more influence on the prediction with respect to *newpos_ av7D*

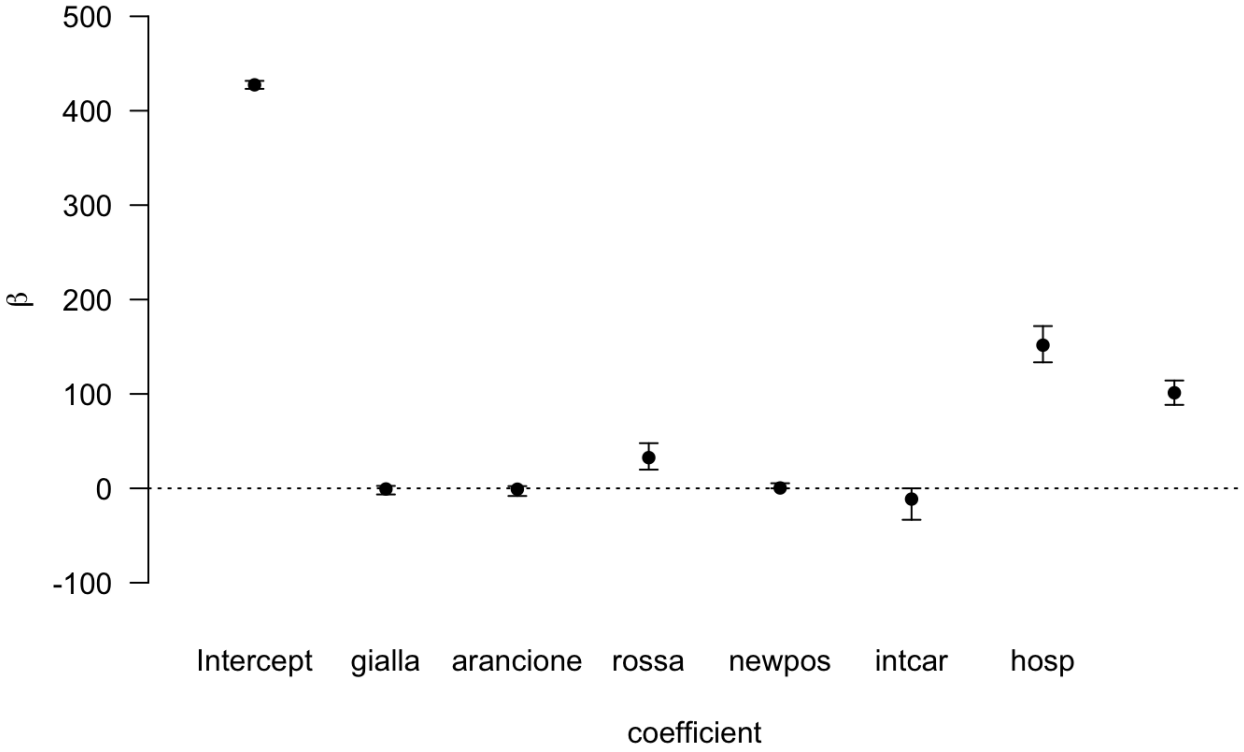


Figure 3.2: 95 % Credible Intervals Coviariates HospH8

Analyzing the figure 3.2 we get the same result as above, knowing that one covariate is more influent when its credible interval doesn't contain 0 and the more far, the better.

## 3.2. IntcarH8

These plots are obtained from the best model for intcarH8 which is the one that uses BIC as prior and the covariates calculated using HPM instead as estimator.

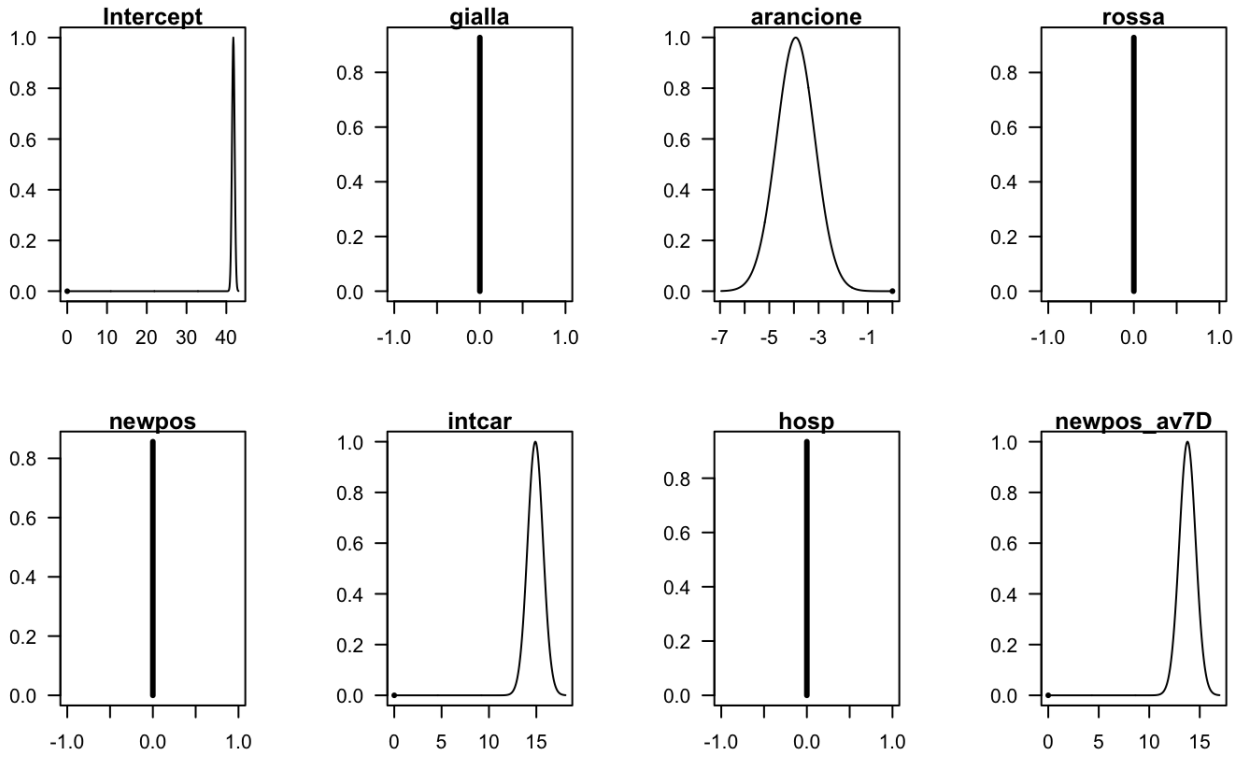


Figure 3.3: Posterior distributions of coefficients for best intcarH8 model

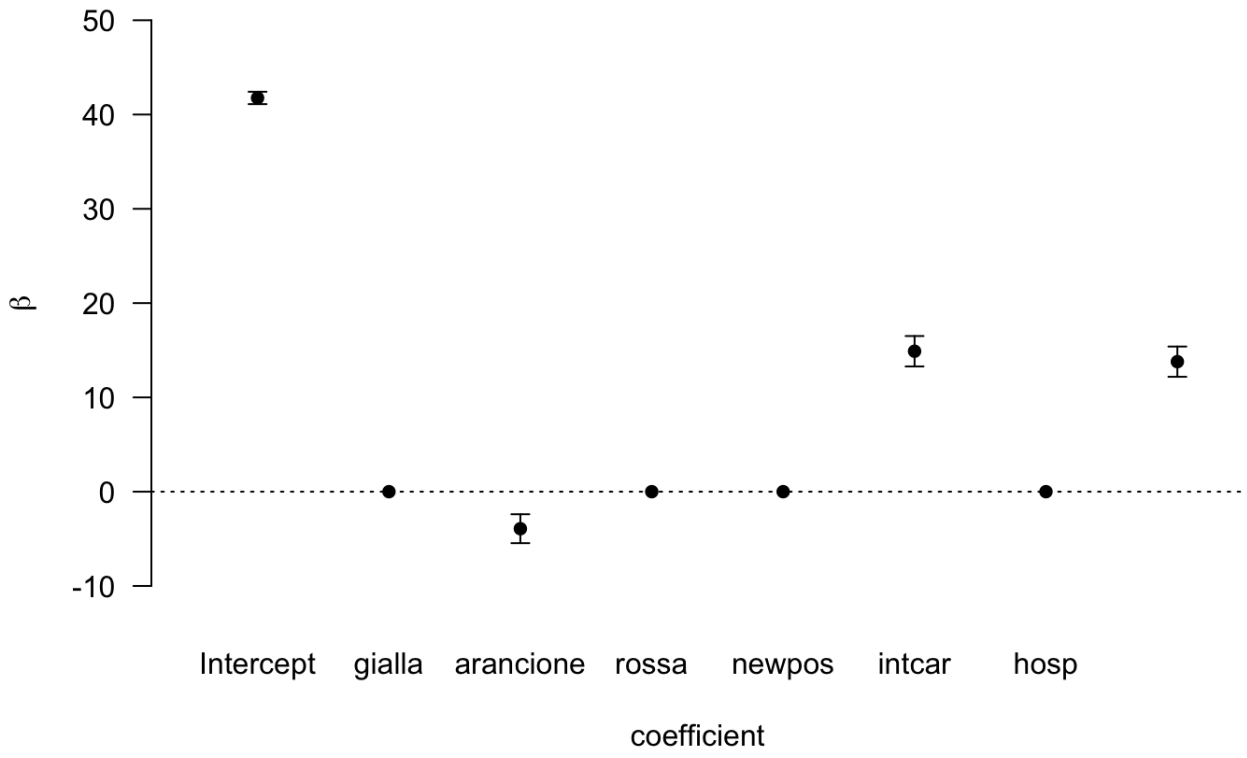


Figure 3.4: 95 % Credible Intervals Coviariates intcarH8

From 3.3 and 3.4 we see that the most influent coviariates are *newpos_ av7D* and *intcar*, while *gialla*, *rossa*, *hosp* and *newpos* have a mean of 0 and credible interval completly in 0. This means that the contribution of these variables in the prediction is null. This is obviously expected since those covariates were not included in the model, hence they cannot influence it.

# 4 | Predictive Analysis

Finally in this chapter we want to show the results obtained by the two models while trying to predict new data.

## 4.1. HospH8

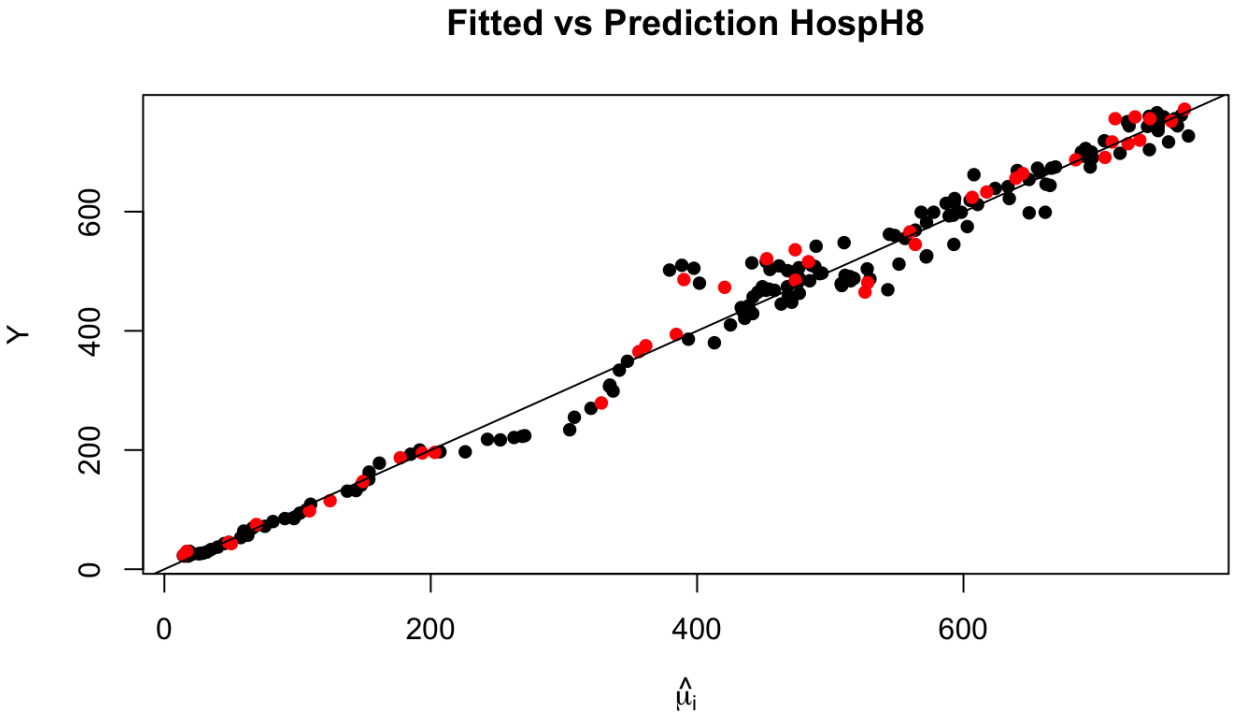


Figure 4.1: Posterior distributions of coefficients for best hospH8 model

This graph shows the difference in behaviour between **fitting** and **prediction**. The red points represent the predicted values which are obtained from new data (different from the training ones) and the black ones are the values predicted by our model on the data used for training. We can see that both points follow the straight line, meaning that the model performs well

both on known data,the training one, and unknown data, the new observations.

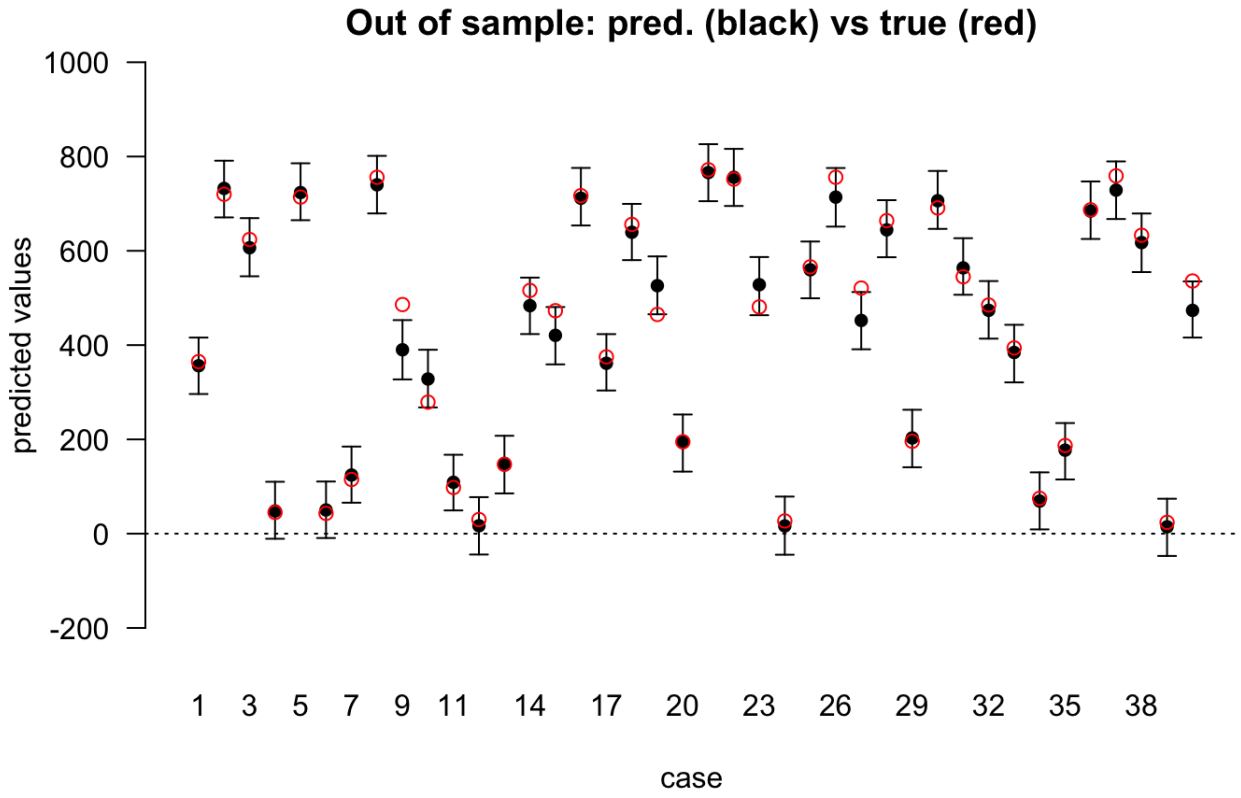


Figure 4.2: 95 % Credible Intervals Coviariates hospH8

Figure 4.2 shows the errors in the predictions with the confidence intervals. There is a big error when the red point is not even included in the credible interval, and this happens only in one of the cases above, there are others where the red point is in the borders. Overall we can notice a good accuracy of predictions.

## 4.2. IntcarH8

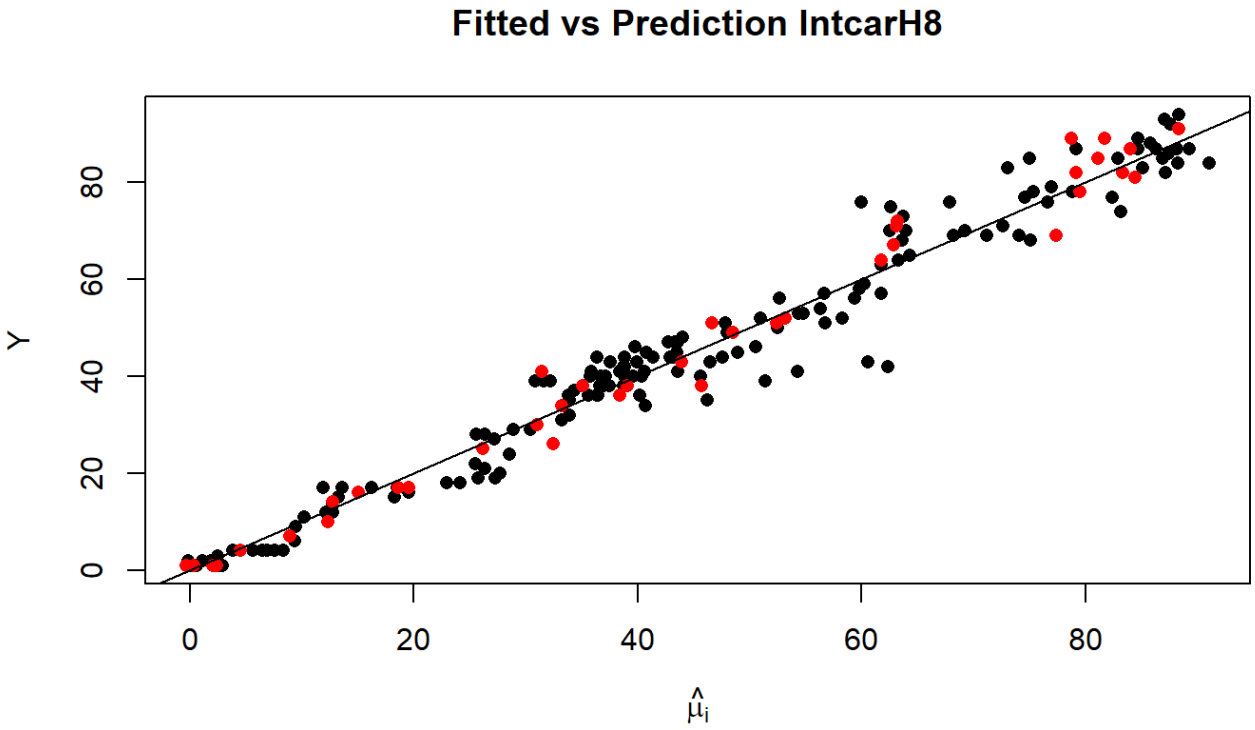


Figure 4.3: Posterior distributions of coefficients for best intcarH8 model

Again for *intcarH8* we can see that the model performs pretty well on both the type of data.

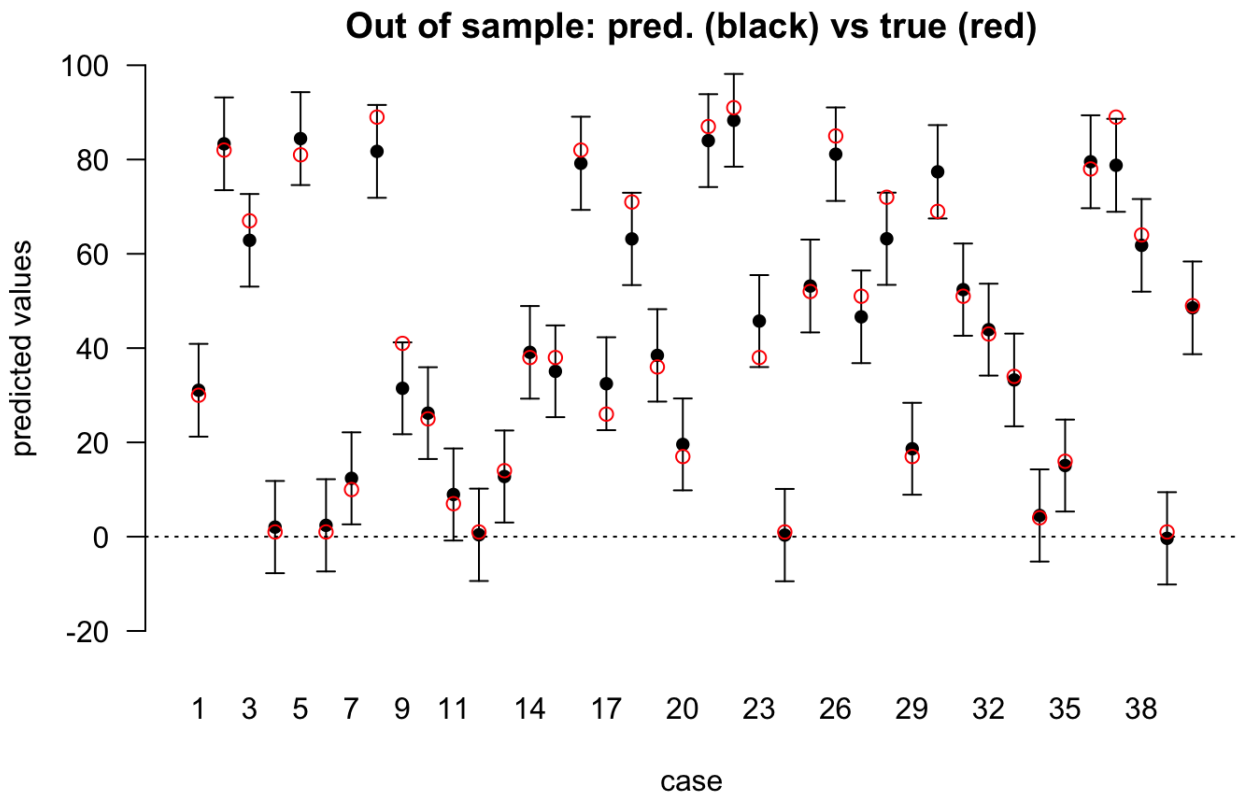


Figure 4.4: Complete scatterplot matrix for hospH8

Figure 4.4 shows that the predictions in most of the cases are accurate since almost all the red points are contained in the intervals.

# 5 | Conclusion

In the end we can conclude that according to our analysis the best prior which gives us the best performance is the non-informative one, even though the results obtained appear to be comparable especially with the Zellner-Siow prior. After the analysis we were also able to find out that the color of Lombardy region, during the Covid pandemic, played a crucial role in predicting our targets.

# A | Appendix

## A.1. Plots

This section contains different scatterplots related to different colors.

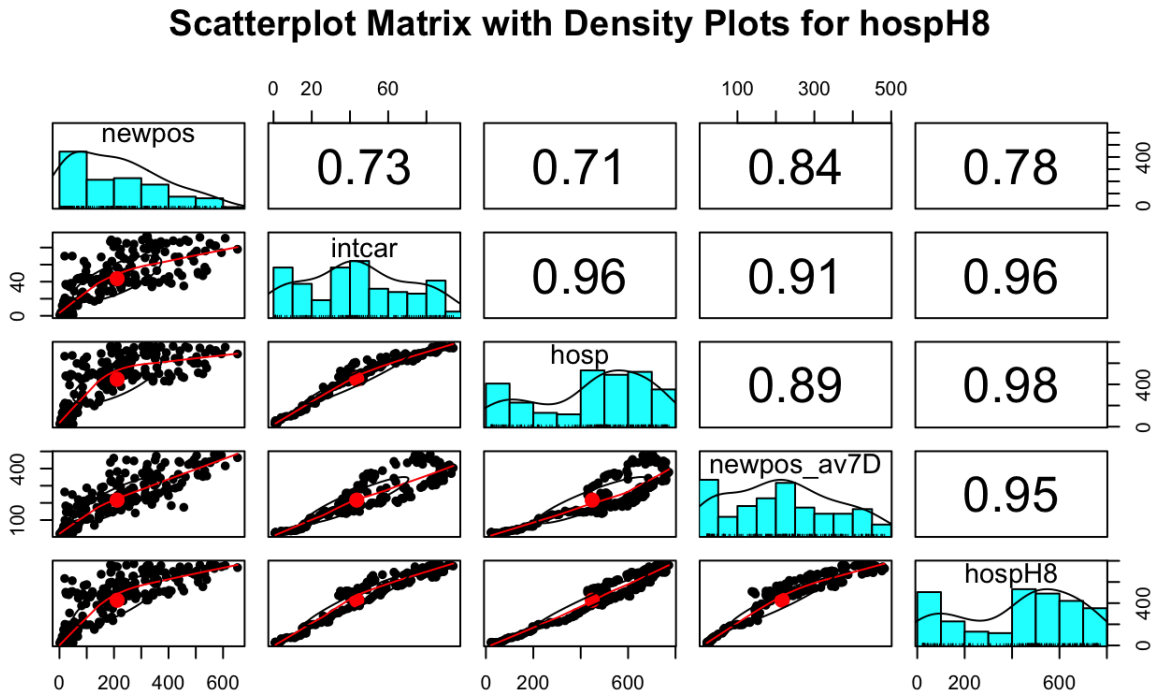


Figure A.1: Complete scatterplot matrix for hospH8

We see from this one that hospH8 has a very strong positive correlation with intcar, hosp and newpos_av7D. Also we see that the covariates are also highly correlated between each other, like for example intcar and hosp have corr=0.96. We also see that when it shows the scatterplot of their correlation is tight suggesting a strong relationship and it resembles a straight line. When the correlation is not that strong, the scatterplot is more sparse like for newpos-intcar.
When two variables are so much related, it suggests that they influence each other and so one

could be left out of the model. Still this is not something we can say for sure, further analysis are needed.

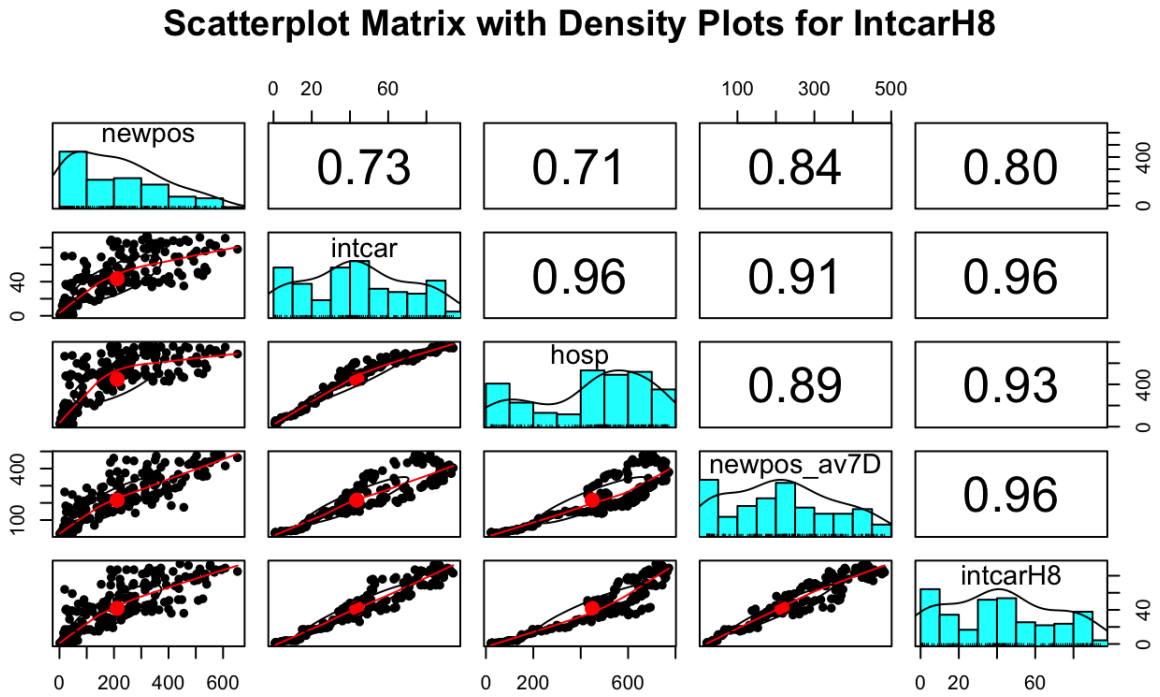


Figure A.2: Complete scatterplot matrix for intcar8

We see that intcarH8 has a very strong relationship with intcar, newpos_av7D (0.96) and hosp (0.93).

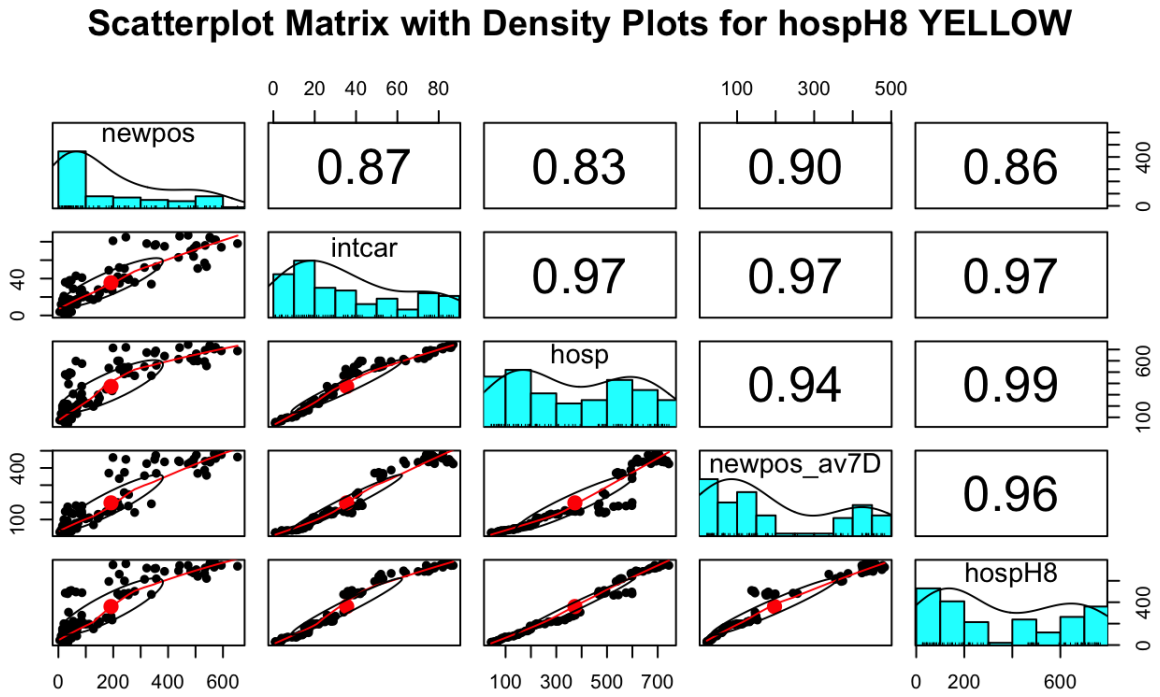


Figure A.3: Scatterplot matrix with observations for hospH8 when color='Gialla'

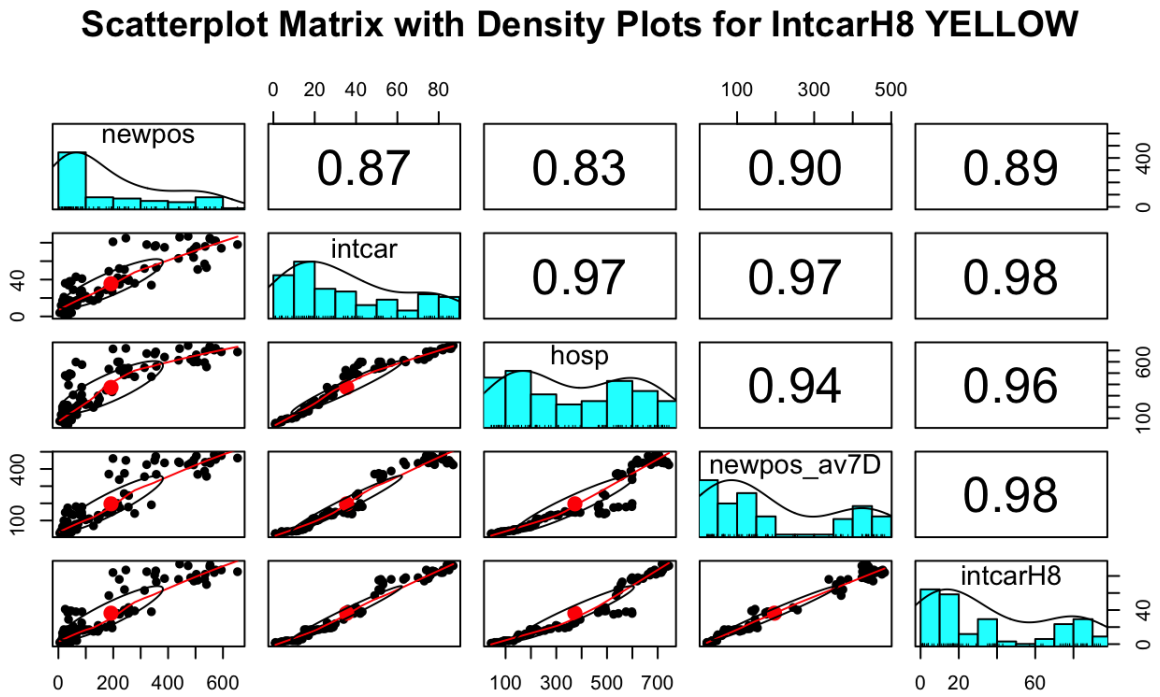


Figure A.4: Scatterplot matrix with observations for intcarH8 when color='Gialla'

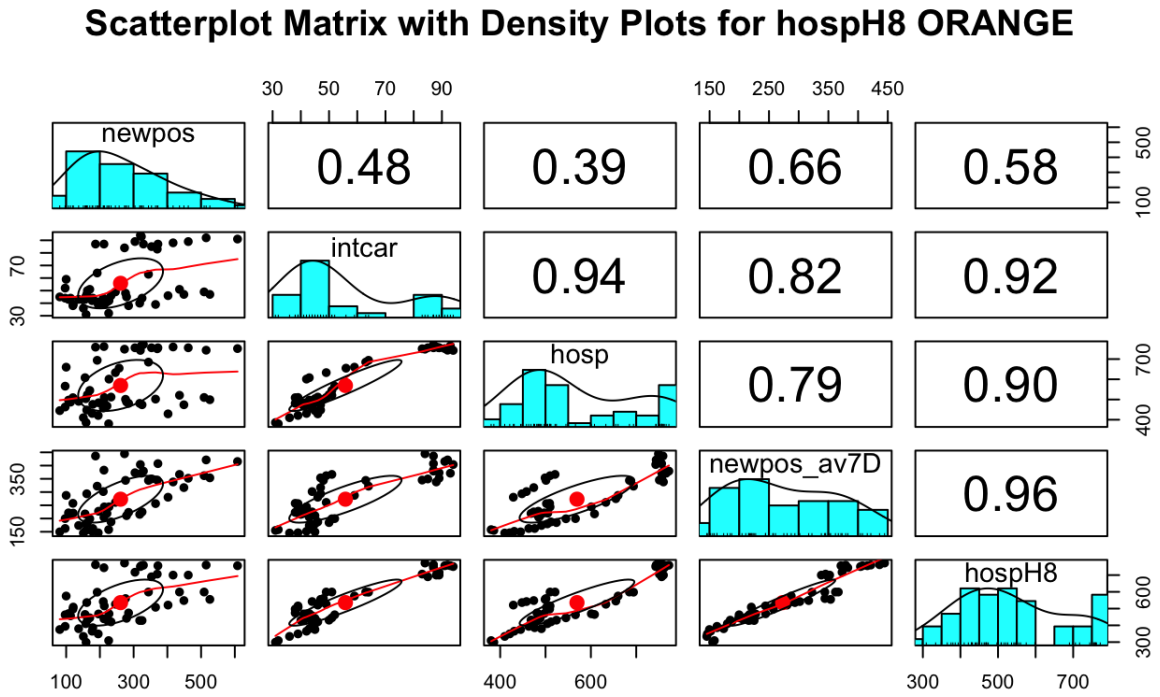


Figure A.5: Scatterplot matrix with observations for hospH8 when color='Arancione'

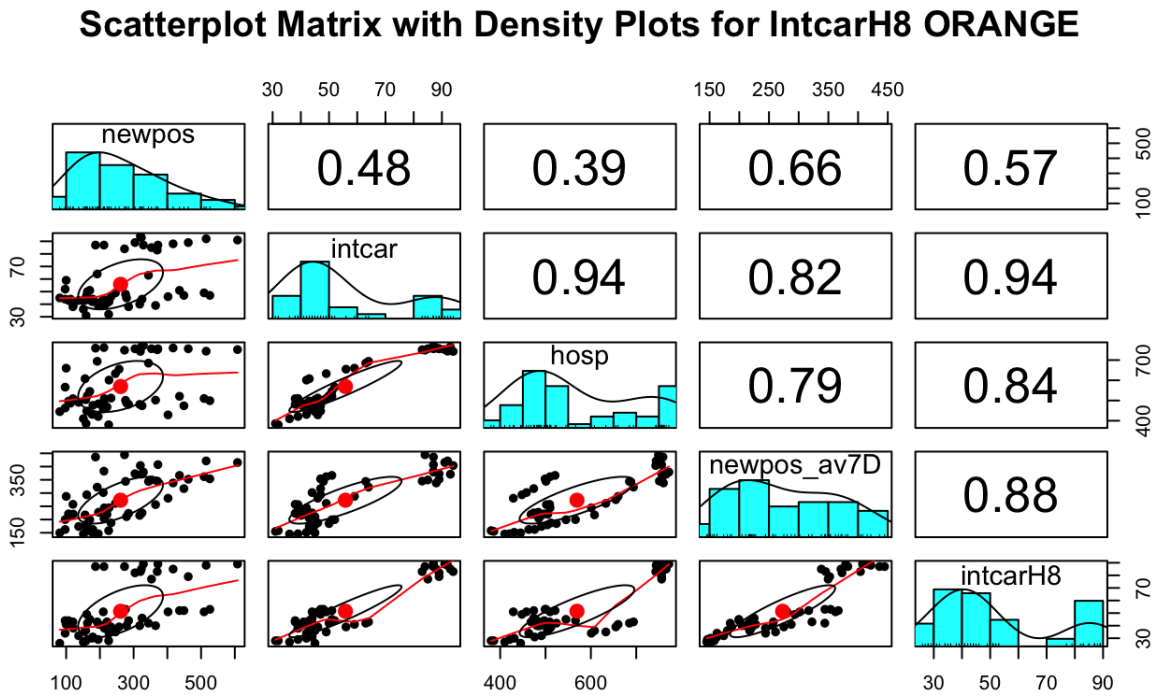


Figure A.6: Scatterplot matrix with observations for intcarH8 when color='Arancione'

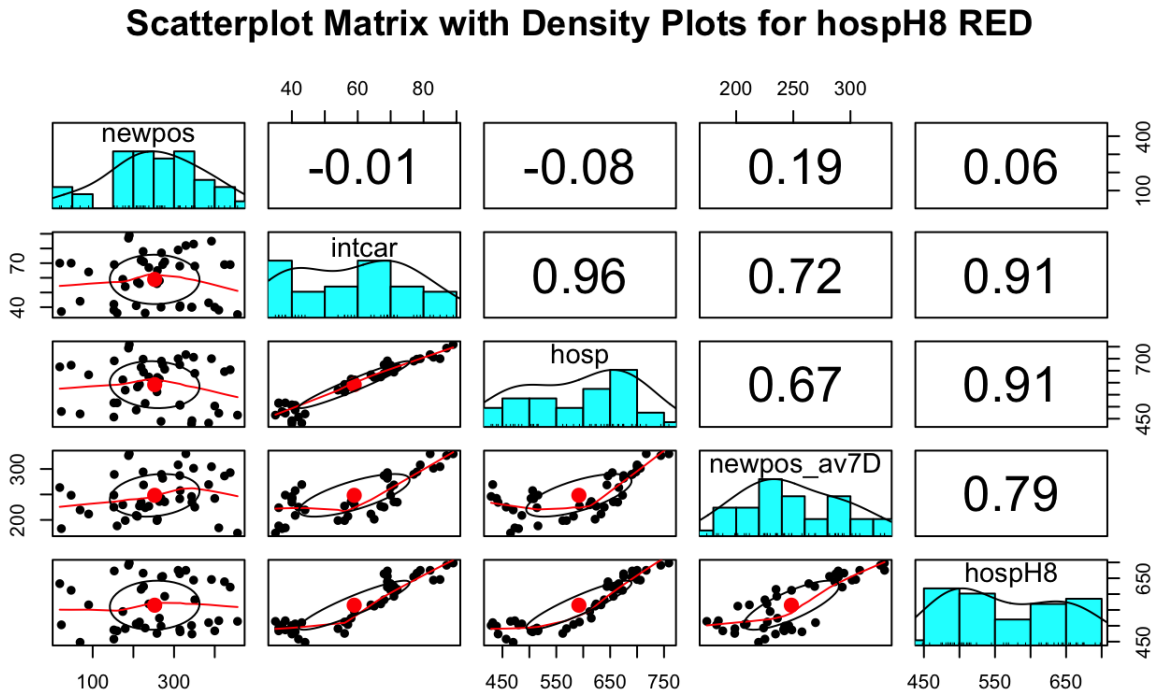


Figure A.7: Scatterplot matrix with observations for hospH8 when color='Rossa'

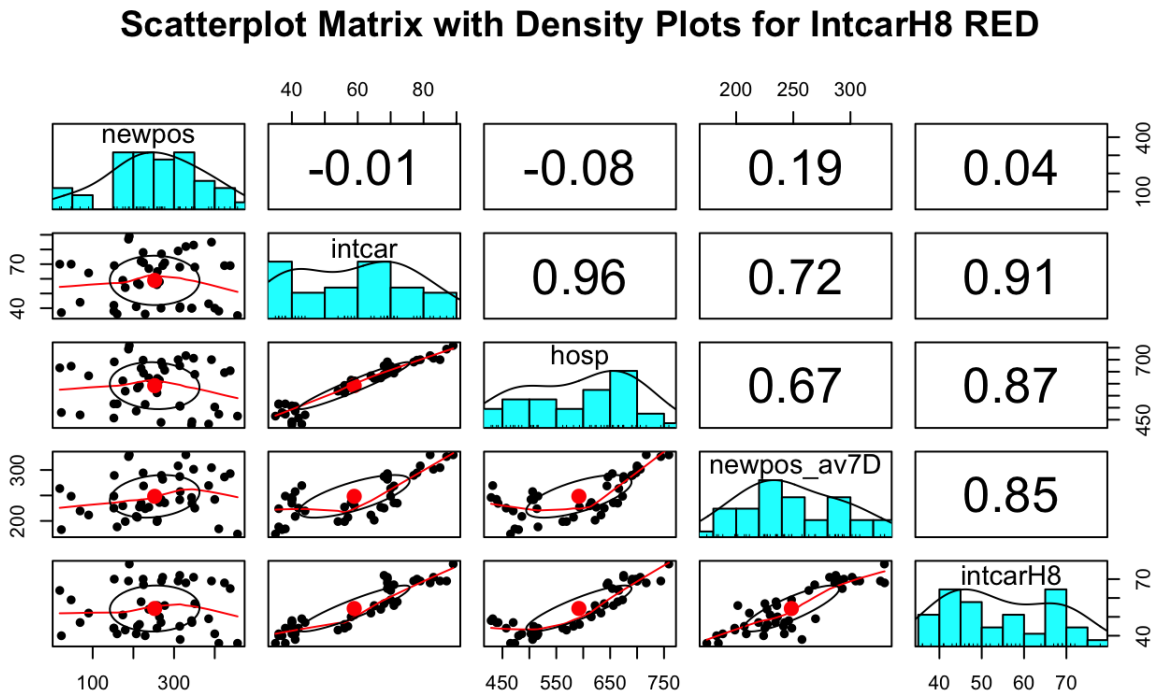


Figure A.8: Scatterplot matrix with observations for intcarH8 when color='Rossa'

## A.2. Outliers Analysis

**- For HospH8**

| Point | Actual HospH8 | 2.5% | 97.5% | Prediction |
|---|---|---|---|---|
| 75 | 502 | 327.8080 | 450.2849 | 389.5364 |
| 184 | 505 | 596.6542 | 716.9518 | 658.8055 |
| 132 | 510 | 339.8858 | 460.0887 | 399.8365 |

Table A.1: Outliers points with predicted credible intervals vs Actual value

**- For IntcarH8**

| Point | Actual IntcarH8 | 2.5% | 97.5% | Prediction |
|---|---|---|---|---|
| 86 | 76 | 50.72509 | 68.85411 | 59.78960 |
| 184 | 43 | 52.72225 | 70.84950 | 61.78588 |
| 135 | 42 | 54.35397 | 72.50540 | 63.42968 |

Table A.2: Outliers points with predicted credible intervals vs Actual value

We notice that for both the target variables the outliers found in the residuals graph are real since the actual value is not included in the credible predicted region. To increase the accuracy of our model we could think of removing these points from out dataset since they could be misleading.