POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Bayesian learning and Monte Carlo Simulations: Final Project

Author: **Felipe Azank dos Santos, Gabriel Speranza Pastorello, Giacomo Savazzi**

Student IDs: 10919711, 10946365, 10675184
Advisor: Prof. Federico Bassetti
Co-advisor: Alessandro Carminati
Academic Year: 2022-23

Contents

1 | Introduction

1.1. Introduction

In this project, we aim to perform a regression analysis on the Boston Housing Dataset to predict the median value of owner-occupied homes (MEDV). The dataset provides valuable information collected by the U.S. Census Service regarding housing in the area of Boston, Massachusetts.

The Boston Housing Dataset consists of 14 variables, each representing different characteristics of various neighborhoods in Boston. These features include the per capita crime rate (CRIM), the proportion of residential land zoned for large lots (ZN), the full-value property-tax rate (TAX), and more.

The target variable, MEDV, represents the median value of owner-occupied homes in thousands of dollars. Our goal is to build a regression model that can accurately predict the MEDV based on the available features.

Understanding the relationship between the predictor variables and the target variable is crucial for making informed decisions related to real estate investments and urban planning. This dataset provides an excellent opportunity to explore Bayesian regression techniques and evaluate their effectiveness in predicting housing prices.

In the following sections, we will delve into the model specification, prior selection, posterior analysis, and model comparison. We will also conduct a prediction exercise to evaluate the performance of our model on unseen data. Through this analysis, we aim to uncover meaningful relationships and draw conclusions that can contribute to the understanding of the Boston housing market as well as practicing what was shown during the lectures.

1.2. Description of the data

According to the data source [2], it seems like the values for the MEDV variable have been set to a maximum of 50. That can be observed by the existence of multiple instances of the exact value of 50 in the dataset. Because of that, these values are considered outliers and will not help in prediction, so they were removed, decreasing the total number of samples from 506 to 490.

Some of the main information (minimum and maximum value, values for first and third quantile, median and mean) about each of the 14 variables (including target MEDV) can be found in Fig. 1.1.

Variable	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE
Min.	0.00632	0	0.74	0	0.3850	3561	2.90
Max.	88.97620	100.00	27.74	1	0.8710	8780	100.00
1st Qu.	0.08205	0	5.19	0	0.4490	5881	44.55
3rd Qu.	3.64742	12.50	18.10	0	0.6240	6578	93.88
Median	0.24751	0	9.69	0	0.5380	6185	76.80
Mean	3.64324	11.11	11.11	0.05918	0.5543	6245	68.28
Variable	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
Min.	1137	1000	187.0	12.60	0.32	1980	5.00
Max.	12127	24000	711.0	22.00	396.90	37970	48.80
1st Qu.	2111	4000	280.2	17.40	375.91	7348	16.70
3rd Qu.	5215	24000	666.0	20.20	396.32	17117	24.68
Median	3276	5000	330.0	19.10	391.77	11675	20.90
Mean	3835	9514	408.0	18.52	355.86	12924	21.64

Figure 1.1: Summary of the dataset.

Doing an initial exploratory data analysis, it is possible to observe the individual correlation of the variables towards the target MEDV in Fig. 1.2. A strong red coloring means a powerful negative linear relationship (Pearson coefficient close to –1), while a strong blue means a powerful positive linear relationship (Pearson coefficient close to +1). Some of the most eminent features in this regard are LSTAT and INDUS with a negative correlation, and RM and DIS with a positive correlation. These features can be analyzed individually in Fig. 1.3.

Figure 1.2: Correlation plot between the features and MEDV.

The higher the percentage of lower status of the population (LSTAT) the smaller the value of MEDV, which makes sense. The interesting aspect is the notable behavior of exponential decay. Another variable to also have a significant tendency is the average number of rooms per dwelling (RM), this time having a linear behavior. Also makes sense, with a higher number of rooms meaning a higher value.

The behavior for the weighted distances to five Boston employment centers (DIS) is not as clear as the previous, however it is worth noting that an increase in its value tends to generate an increase in MEDV in general. This is an interesting aspect, since it would be prudent to assume that most people would like to live closer to their work, however these places may have higher pollution and overall lower quality of life. Another feature with intriguing behavior is the proportion of non-retail business acres per town (INDUS), as higher INDUS values tend to be present in lower values of MEDV, which is understandable in the real world for similar reasons as DIS.

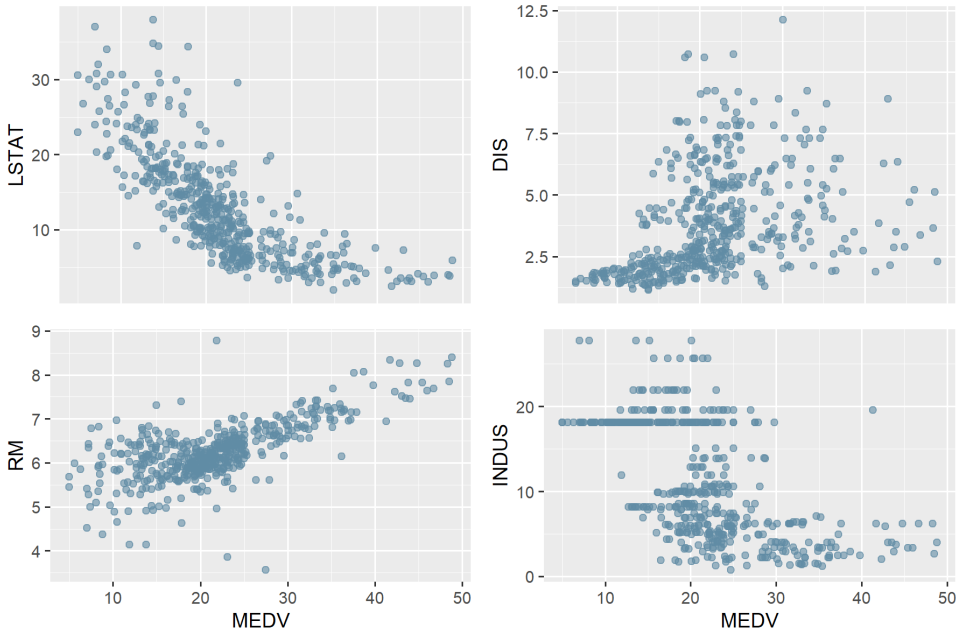

Figure 1.3: Scatter plots of LSTAT (upper left), DIS (upper right), RM (lower left) and TAX (lower right) in relation to target variable MEDV.

Taking a look at the histogram in Fig. 1.4 for the target MEDV, we can observe a certain normal behavior in the data, with some associated noise. This will serve as motivation to choose this distribution in the following steps of the project.

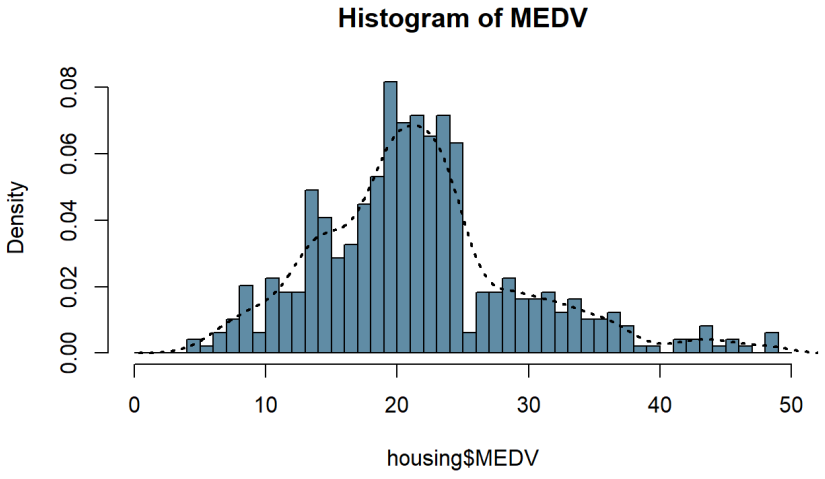

Figure 1.4: Histogram for target MEDV

2 | Model Specification

2.1. Standard Bayesian Regression

In the first of the three main approaches made, the model used in the analysis is a standard Bayesian linear regression model. The relationship between the response variable MEDV and the predictor variables is assumed to be linear with Gaussian noise. Given that, two different priors are considered in the analysis: the g-prior and the Jeffreys-Zellner-Siow (JZS) prior.

Zellner's informative g-prior is a popular choice in Bayesian model selection. It assigns a normal distribution prior to the regression coefficients, and, in the Bayesian Adaptive Sampling (BAS) package in R, it is centered at zero with a variance proportional to the inverse of the model size.

$$\beta \sim N_k \left(0, \alpha\sigma^2 \left(X^t X\right)^{-1}\right) \tag{2.1}$$

Here, β represents the regression coefficients, and α is the hyperparameter equivalent to the g value (described in detail below).

In a g-prior approach, the value g is used to add subjective information to the regression process, reflecting the amount of information in the prior relative to the data presented. In other words, if a small g is chosen, this means there's a strong belief that the prior is correct, while if g is big, that means there is not a considerable belief that the prior is the correct one (like choosing a non-informative prior). Given that we are using a Zellner's g-prior, the value of g will be masked under the value alpha (alpha = g).

After trying the values of 100 and 1 and not noticing any major changes in the coefficient's values and confidence intervals, this serves as indication that the chosen prior distribution is coherent. Because of that, the value of α is set to 1.

The posterior value obtained by using the g-prior as a prior can be expressed as a Normal distribution:

$$\beta|\sigma^2, y, X \sim N_{k+1} \left(\frac{\alpha}{\alpha+1}\hat{\beta}, \frac{\sigma^2\alpha}{\alpha+1}\left(X^t X\right)^{-1}\right) \tag{2.2}$$

The JZS prior (Jeffreys prior on sigma and the Zellner-Siow prior on the coefficients) is another commonly used prior in Bayesian model selection because it offers a reasonable

compromise between regularization and flexibility in the model. It is a mixture of Zellner prior and does not introduce additional information but rather reflects a conservative belief about the coefficients' distribution.

This allows for the modeling of small deviations from the null hypothesis, acknowledging the possibility of subtle effects or variations in the data. As described by A.F. Jarosz and J. Wiley [1], cases with small samples may be better served by employing the JZS method. In this analysis, the JZS prior is employed with default settings in the BAS package.

In this first approach, Bayesian Adaptive Sampling (BAS) algorithm for model or variable selection was not used, which would lead to the sampling without replacement from the models generated. However, when compared with more sophisticated approaches that will be seen below, we can see that the lack of this process impacted the overall performance of the model, seen in section 3.

2.2. Regression with Bayesian Information Criterion (BIC)

In the second approach considered, we applied the Bayesian Informative Criterion process, with backward elimination, to select from a range of linear models produced, the best ones considering both information lost due to simplification, as well as complexity of the regression.

The Bayesian Information Criterion (BIC) can be better described as a criterion used to select the best model among different candidates given by Eq. 2.3:

$$BIC = kln(n) - 2ln(\hat{L}) \tag{2.3}$$

Where p is the number of predictors (without the intercept), n is the number of samples and $\hat{L}$ is the maximum value of the likelihood function for the model. It balances model fit and complexity, as it penalizes models that do not fit the data well, as indicated by lower log-likelihood values, and penalizes models with a larger number of parameters, discouraging overfitting (high variance) and favoring simpler models. The specific set of values for the hyperparameters in the BIC prior is determined automatically based on the data and model structure and a uniform prior is assigned for simplicity and execution.

This estimator was chosen instead of the standard Akaike Information Criterion (AIC) because the BIC approach provides a more incisive penalty for complexity as well as being independent from the prior adopted.

By default, the model selection process using the BIC criterion was the Backward Selection, a process in which starts with the regression considering all n available features (13), extract the BIC value and compares it with the best model with n-1 features (12 in our case). Then, if the BIC estimation of the simpler models is better, the process is repeated until it cannot be improved more.

2.3. Just Another Gibbs Sampler (JAGS) Spike and Slab

In the third approach adopted, we used Just Another Gibbs Sampler (JAGS), which is specifically designed for the implementation of Markov Chain Monte Carlo (MCMC) algorithms. The main idea behind JAGS is to allow users to define their statistical models and specify the relationships between variables, priors and likelihoods. JAGS then utilizes MCMC methods, specifically Gibbs sampling, to estimate the posterior distribution of the model parameters.

The Gibbs Sampling method is a famous sampling procedure that can be used to calculate marginal probabilities distributions (such as the posterior). It is used especially in cases where the full conditional distributions are easier to calculate than the joint distribution.

In this analysis, the likelihood of the model is specified as a Gaussian regression model,the response variable is assumed to follow a Gaussian distribution. The intercept depends directly on a linear combination of the feature where the regression coefficients are assigned a Spike-and-Slab prior. The precision of the Gaussian distribution is assigned a Gamma distribution while the hyper-parameters for the coefficient priors are set to 1 and 0, respectively, resulting in a diffuse prior centered at zero.

$$\begin{aligned} &Y|X \sim \text{Norm}(\mu, \sigma^2) \\ &\mu = \beta_0 + X_i\beta \\ &\tau = \frac{1}{\sigma^2} \sim Gamma(0.001, 0.001) \\ &(\beta_0, \beta) \sim \pi(\beta_0, \beta) \end{aligned} \tag{2.4}$$

The Spike-and-Slab prior assign for each feature an index γ_i such that$\gamma_i = 0$ if $\beta_i = 0$ and $\gamma_i = 1$ if $\beta_i \neq 0$ with:

$$\gamma_j \overset{\text{ind}}{\sim} \mathcal{B}e\left(\theta_j\right) \quad j = 1, \ldots, p$$

where θ_j is the probability that β_j is big enough to be included in the model. The prior can then be described as

$$\beta_j \mid \gamma_j \overset{\text{ind}}{\sim} (1 - \gamma_j)\, \delta_{\{0\}} + \gamma_j\, \mathcal{N}\left(0, \sigma^2_{\beta_j}\right) \tag{2.5}$$

$$\gamma_j \mid \theta_j \overset{\text{ind}}{\sim} \mathcal{B}e\left(\theta_j\right) \tag{2.6}$$

$$\theta_j \overset{\text{iid}}{\sim} \pi\left(\theta_j\right) \tag{2.7}$$

Where $\delta_{\{0\}}$ is a Dirac measure with mass in zero and $\mathcal{N}\left(0, \sigma^2_{\beta_j}\right)$ is a diffuse distribution. As previously said in our case $\sigma^2_{\beta_j}$ was chosen to be 1.

The posterior can be analytically calculated, however, its computation is expensive enough to substitute this process for an approximate inference, such as the Gibbs Sampler.

3 | Posterior analysis

3.1. Standard Bayesian Regression

Starting from the g-prior approach, with the defined posterior distribution seen in Eq. 2.1, the following values for the coefficients are obtained:

G Prior	Posterior Mean	Posterior SD	p(B != 0)
Intercept	21,6359	0,1698	1
CRIM	-0,0534	0,0185	1
ZN	0,0177	0,0080	1
INDUS	-0,0219	0,0352	1
CHAS	0,2261	0,5244	1
NOX	-6,1991	2,1615	1
RM	1,8797	0,2527	1
AGE	-0,0118	0,0075	1
DIS	-0,6055	0,1136	1
RAD	0,1257	0,0375	1
TAX	-0,0069	0,0021	1
PTRATIO	-0,4191	0,0745	1
B	0,0039	0,0015	1
LSTAT	-0,1751	0,0301	1

Figure 3.1: g-prior posterior distribution of coefficients

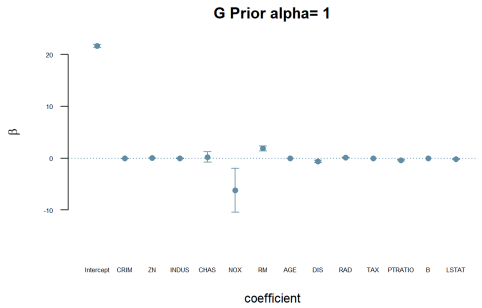

Figure 3.2: Parameters mean values and confidence intervals with g-prior

With the JZS prior, we can notice quite similar results, which is a initial indication that the prior distribution may not be as much significant.

JZS Prior	Posterior Mean	Posterior SD	p(B != 0)
Intercept	21,6359	0,1698	1
CRIM	-0,1060	0,0260	1
ZN	0,0351	0,0112	1
INDUS	-0,0436	0,0496	1
CHAS	0,4493	0,7393	1
NOX	-12,3178	3,0469	1
RM	3,7351	0,3562	1
AGE	-0,0235	0,0106	1
DIS	-1,2031	0,1601	1
RAD	0,2497	0,0529	1
TAX	-0,0137	0,0030	1
PTRATIO	-0,8328	0,1050	1
B	0,0078	0,0021	1
LSTAT	-0,3478	0,0424	1

Figure 3.3: ZSL posterior distribution of coefficients

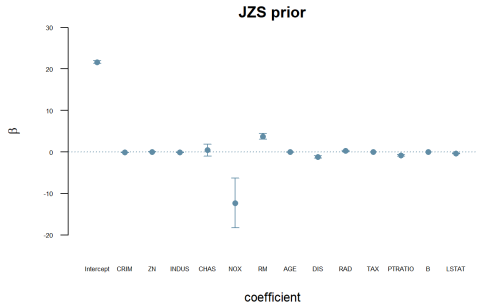

Figure 3.4: Parameters mean values and confidence intervals with ZSL

3.2. Bayesian Information Criterion (BIC)

In the second approach, we can see in Fig. 3.5 that 3 of the features were not included given the BIC criteria, excluding variables INDUS, CHAS and AGE, leaving the model with 10 features. The choice for the CHAS feature is understandable, mainly because it is a dummy variable and is the one with the least linear correlation, as seen in Fig. 1.2.

However, the INDUS variable is an interesting choice, as a stronger linear relationship can be noticed in both Fig. 1.2 and 1.3. One possibility is that, above a certain value, the INDUS feature appears to have rounded values, differently from lower values where it shows a more continuous behavior, as noticed in Fig 1.3. This points to a procedure in data collection that can result in loss of information. By comparing the values with the previous models, it's possible to notice that the INDUS and AGE features that were dropped indeed showed to have a small coefficient value in the previous approach.

Best BIC	Posterior Mean	Posterior SD	p(B != 0)
Intercept	21,6359	0,1703	1
CRIM	-0,1066	0,0261	1
ZN	0,0388	0,0112	1
INDUS	0	0	0
CHAS	0	0	0
NOX	-14,8492	2,8125	1
RM	3,6297	0,3483	1
AGE	0	0	0
DIS	-1,0840	0,1508	1
RAD	0,2741	0,0511	1
TAX	-0,0151	0,0027	1
PTRATIO	-0,8755	0,1038	1
B	0,0077	0,0021	1
LSTAT	-0,3881	0,0394	1

Figure 3.5: BIC posterior distribution of coefficients

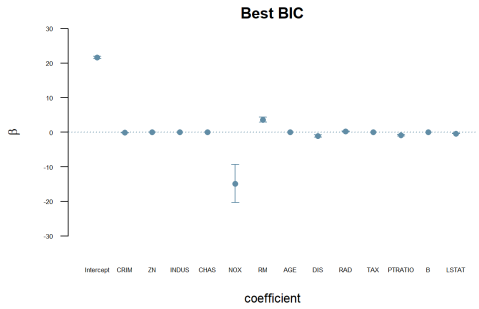

Figure 3.6: Parameters mean values and confidence intervals with BIC

3.3. Just Another Gibbs Sampler (JAGS) Spike and Slab

The approach using JAGS with the Spike and Slab prior estimates both the posterior over the regression coefficients and the posterior of the inclusion probability of a certain feature in the model.

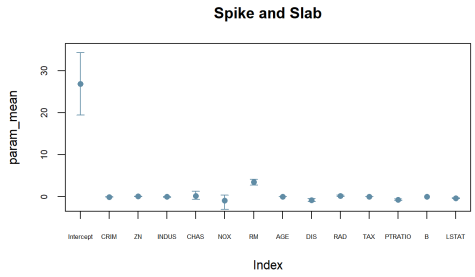

Figure 3.7: Parameters mean values and confidence intervals of Gibbs approach with spike & slab prior

The posterior of the inclusion probability was then used to perform model selection using two different approaches. The first approach selects a model called Median Probability Model by excluding all the features that have mean inclusion probability less than 0.5.

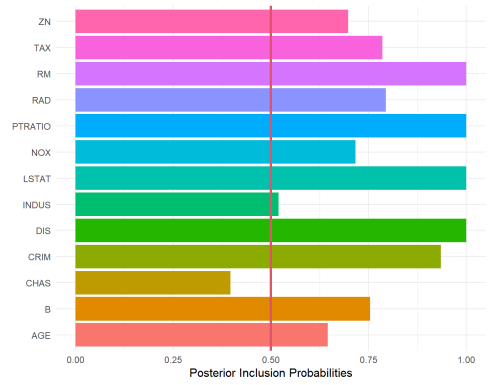

Figure 3.8: Histogram of the g values used to determine the selected features

In this case the only feature excluded was CHAS feature, this is consistent with the feature excluded by BIC and also with the correlation analysis since CHAS is the feature with less linear correlation.

The second approach selects a model called Highest Posterior Density Model that is the model that was visited most frequently in the MCMC generated by JAGS.

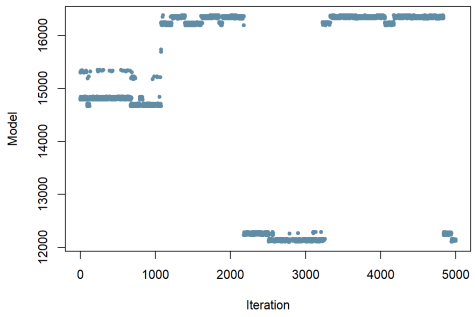

Figure 3.9: Visited Models JAGS approach

In this case the selected model depended partly on the number of samples used in the JAGS. With the standard fifty thousand samples we would obtain a model with eleven features that would exclude the CHAS and INDUS features.

Instead by using a higher number of samples, around a hundred thousand samples, the selected model alternated between the previous one and a model with two less features, that also excluded the AGE and B features. In both cases the obtained model would exclude CHAS and INDUS just like BIC and the second would also exclude AGE, like BIC, but also B. This is in part surprising since, in the BIC approach, B had a coefficient close to zero but a high inclusion probability.

4 | Prediction analysis

4.1. Performance metrics chosen

To compare the model performance, as well as to understand how each approach differs from one another, we started off the process by making the standard 70/30 train test split, generating a test set with 147 random samples.

The Regression Metrics used to compare and evaluate the models were:

* Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{4.1}$$

* Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \mid \hat{y}_i - y_i \mid \tag{4.2}$$

* Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \mid \frac{\hat{y}_i - y_i}{y_i} \mid \tag{4.3}$$

The first 2 metrics were chosen considering that they have a straightforward, easy to understand interpretation because they are presented in the same metric as the target value itself. RMSE calculates the root mean square errors between actual values and predictions. It is widely used because it is a metric that punishes large errors (when squaring), but it is in the same unit as the variable of interest (when rooting). That is, the lower its value, the better.

On the other hand, the Mean Absolute Error (MAE), does not penalize outliers in the same way and, therefore, when compared with the RMSE, can be used to determine how's the performance of the model is with regards to generating outliers. Finally, the MAPE metric is used to interpret percentage how far off from the target the predictions are generated, eliminating the comparison problems with magnitude of the values.

4.2. Standard Bayesian Regression

Analysing the results for prediction using g-prior and JZS prior in Fig. 4.1 and 4.2, it is possible to observe very similar results. The RMSE, MAE and MAPE were 3.72, 2.82 and 13,99% respectively for the model with g-prior, and 3.73, 2.82 and 13.99% for the model with JZS prior.

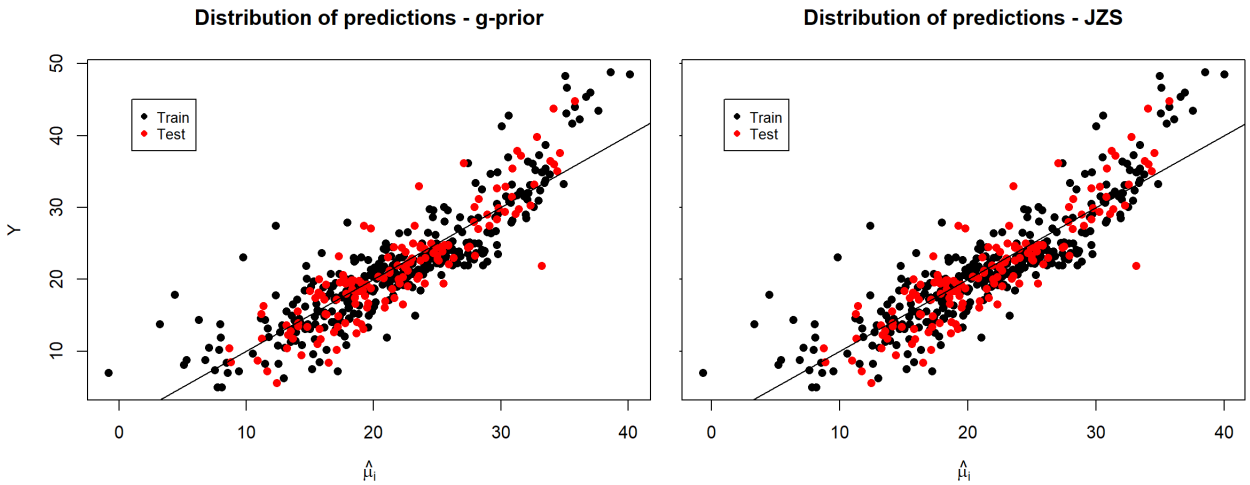

Figure 4.1: Distribution of predictions - first approach

This similarity in results provides evidence that the choice of prior may not be critical for making predictions in this specific context. It suggests that the model's performance is robust and not heavily influenced by the specific prior choice.

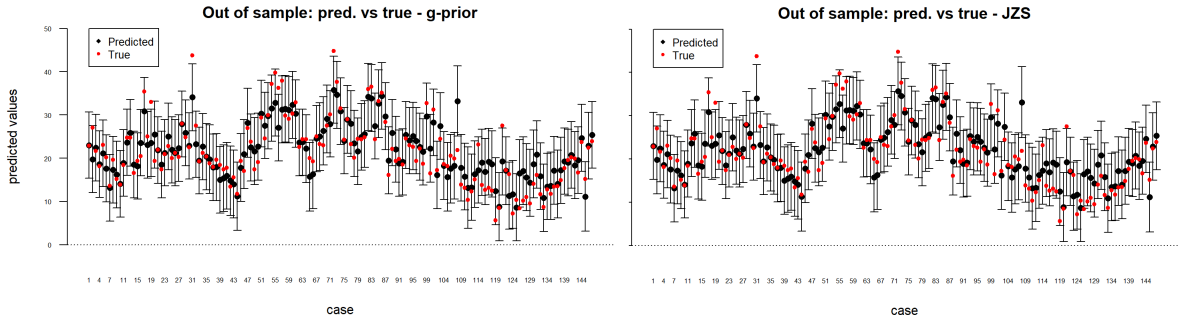

Figure 4.2: Predictions on test samples - first approach

4.3. Bayesian Information Criterion (BIC)

The results obtained with the best BIC model were of 3.62, 2.75 and 13,80% for RMSE, MAE and MAPE, respectively, corresponding to slightly better results compared to the standard models. These results can be observed in Fig. 4.3 and 4.4.

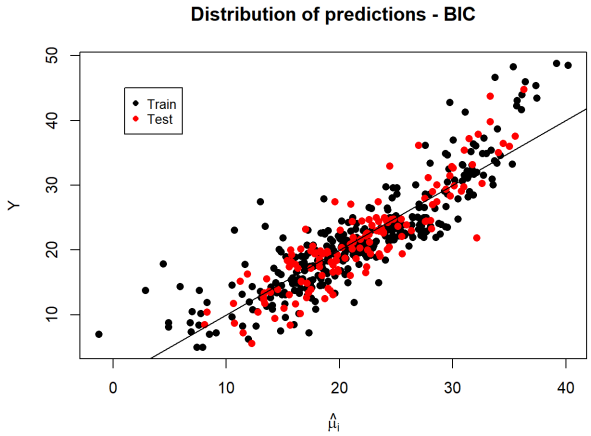

Figure 4.3: Prediction distribution - BIC approach

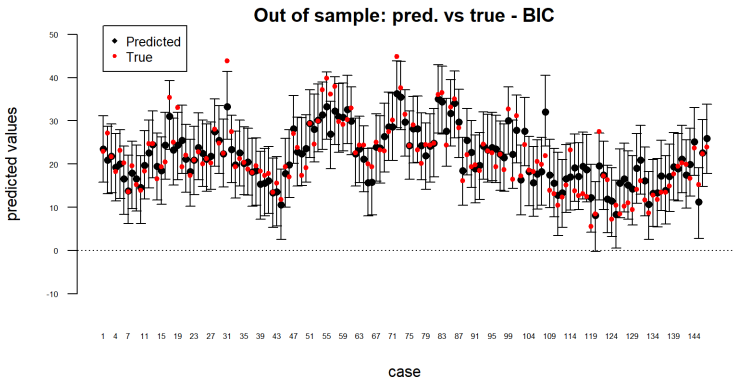

Figure 4.4: Predicted values of test set of best BIC model

4.4. Just Another Gibbs Sampler (JAGS) Spike and Slab

For the prediction analysis we decided to use both the Median Probability Model and the Highest Posterior Density Model, in this case the one with eleven features.

For the Median Probability model, we obtained 3.55, 2.64 and 13.11% values respectively for the RMSE, MAE and MAPE.

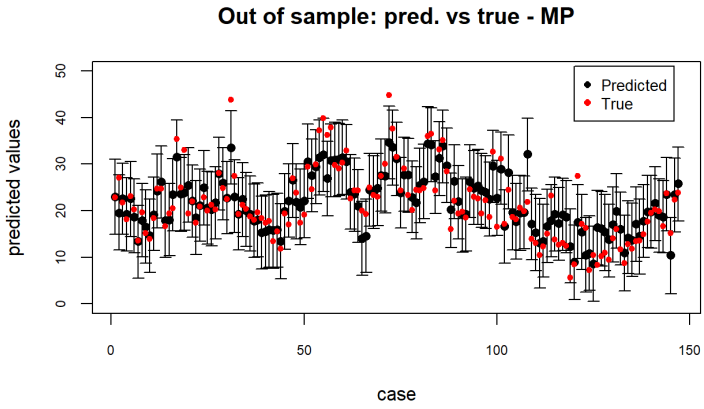

Figure 4.5: Predicted distribution - Median Probability model

For the Highest Posterior Density model, we obtained 3.58, 2.66 and 13.18% values respectively for the RMSE, MAE and MAPE.

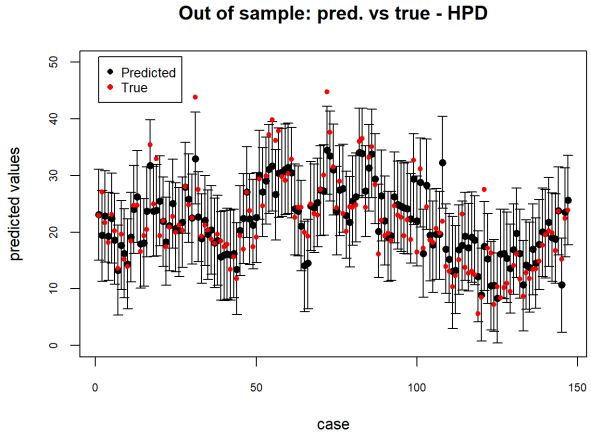

Figure 4.6: Predicted distribution - Highest Probability density

The comparison of all the values, as well as its key takeaways are present in the next section.

5 | Conclusions

Finally, the final prediction results can be found in the following table:

Model	Metric
g-prior standard	RMSE: 3.72
	MAE: 2.82
	MAPE: 13.99%
JZS standard	RMSE: 3.73
	MAE: 2.82
	MAPE: 13.99%
Best BIC	RMSE: 3.62
	MAE: 2.75
	MAPE: 13.80%
JAGS - MP	RMSE: 3.55
	MAE: 2.64
	MAPE: 13.11%
JAGS - HPD	RMSE: 3.58
	MAE: 2.66
	MAPE: 13.18%

Figure 5.1: Comparison of the models' performances

As it's possible to see, the results are in line with the theoretical interpretations of the models and the approaches. The simpler elaborations (Bayesian Regression without model or feature selection), using g-prior or JSZ, have extremely similar performances, proving that the model was quite robust and indifferent to the priors, even with the changes made in the g/alpha value.

Furthermore, it's possible to see and increase in performance in every metric with the utilization of the BIC strategy, in which beyond improving the result, it removed unnecessary complexity to the model (which helps computation), proving to be a good strategy to model selection.

At last, the JAGS approach proved to be successful given that, beyond applying the feature selection using the Median Probability Model, the model selection process using the frequency of the Markov Chain (Highest Posterior Density) led to the fine tuning of the parameters and a significant improvement in the overall performance.

From the the model selected by BIC, Median Probability and Highest Posterior Density we can conclude that the CHAS feature is not strongly related to our target. On the other hand, the INDUS feature is excluded by BIC and HPD, that consider the best model, while it's kept by Median Probability that consider the inclusion probability over all the

models. This could support our hypothesis that while INDUS and MEDV are correlated, there is a loss of data that cause BIC and HPD to exclude it.

In conclusion, we can affirm that the project proved to be helpful with regards to applying the main concepts and theories seen in the lectures as well as giving evidence that the Bayesian Learning approach and modeling is effective in the day-to-day job of a Data Scientist.

Bibliography

[1] A. F. Jarosz and J. Wiley. What are the odds? a practical guide to computing and reporting bayes factors. *Journal of Problem Solving*, 2014.

[2] P. Perera. The boston housing dataset, 2018. URL `https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset`.