POLITECNICO
MILANO 1863

# Bayesian Learning and Monte Carlo Simulation Project: Boston Housing Dataset


## Politecnico di Milano

**Authors**: Felipe Azank dos Santos, Gabriel Speranza Pastorello, Giacomo Savazzi
**Advisor**: Prof. Federico Bassetti
**Co-advisors**: Alessandro Carminati

# Introduction

- **Goal:** perform a regression analysis on the Boston Housing Dataset to predict the median value of owner-occupied homes (**MEDV**)
- **Features**: 13 features + target included in the dataset:

| Feature | Description |
|---|---|
| CRIM | per capita crime rate by town |
| ZN | proportion of residential land zoned for lots over 25,000 sq.ft. |
| INDUS | proportion of non-retail business acres per town |
| CHAS | Charles River dummy variable (1 or 0) |
| NOX | nitric oxides concentration (parts per 10 million) |
| RM | average number of rooms per dwelling |
| AGE | proportion of owner-occupied units built prior to 1940 |

| Feature | Description |
|---|---|
| DIS | weighted distances to five Boston employment centres |
| RAD | index of accessibility to radial highways |
| TAX | full-value property tax rate per $10,000 |
| PTRATIO | pupil-teacher ratio by town |
| B | $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town |
| LSTAT | % of lower status of the population |
| **MEDV** | **Median value of owner-occupied homes (in $1000's)** |

# Description of the data

- **MEDV** value has been set to a **maximum of 50** [1], with **multiple instances** of this exact value in the dataset
  - Considered **outliers** and will not help in prediction, so they were **removed**
  - Total number of samples decreased from **506** to **490**

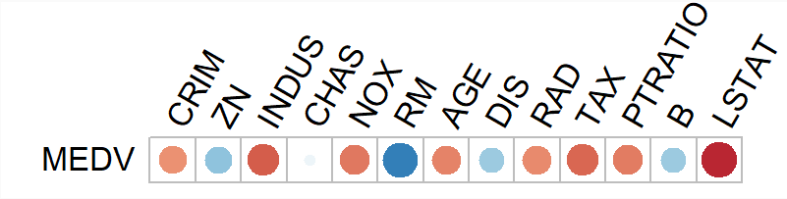


**Fig. 1**: Linear correlation plot between the features and MEDV

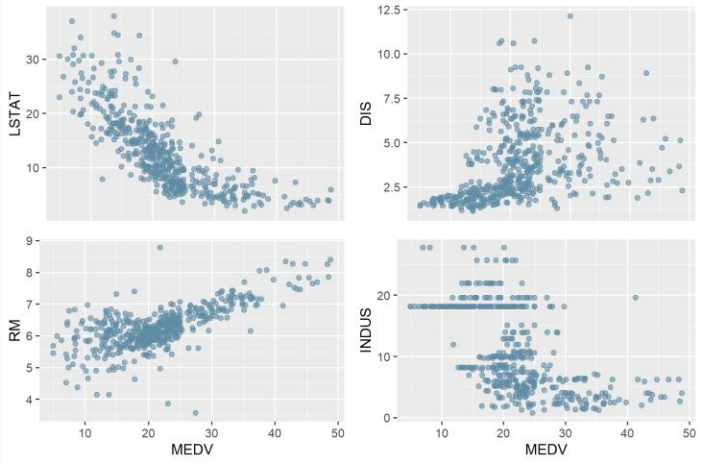


**Fig. 2:** Scatter plots of LSTAT (upper left), DIS (upper right), RM (lower left) and TAX (lower right) in relation to target variable MEDV.

[1] https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset

# Model specification

- **Standard Bayesian Regression**
  - g-prior
  - Jeffreys-Zellner-Siow (JZS) prior
- **Bayesian Information Criterion (BIC)**
- **Just Another Gibbs Sampler (JAGS) Spike and Slab**
  - Median Probability (MP)
  - Highest Probability Density (HPD)

# Standard Bayesian Regression: g-prior

- Assigns a **normal** distribution **prior** to the regression coefficients: $\beta \sim N_k\left(0, \alpha\sigma^2\left(X^tX\right)^{-1}\right)$

- **α = g value**
  - **Small g** = **strong** belief in the prior
  - **Big g** = **not considerable** belief in the prior
  - g values tried: 100, 50, 1
  - Without any major changes, α is then set to 1

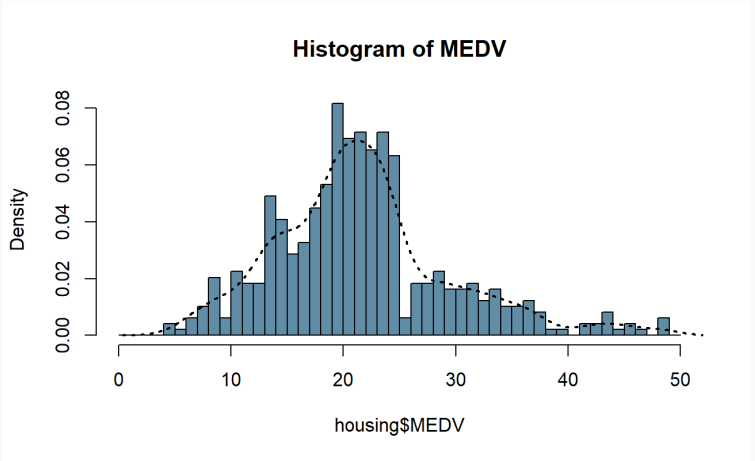


**Fig. 3:** Histogram for target MEDV

# Standard Bayesian Regression: g-prior

- 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 |

**Fig. 4:** g-prior posterior distribution of coefficients

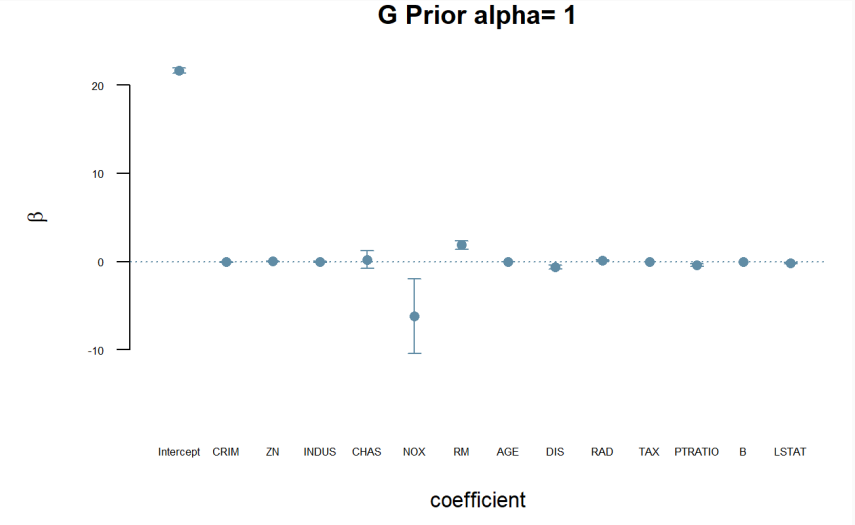


**Fig. 5:** Parameters mean values and confidence intervals with g-prior

# Standard Bayesian Regression: JZS prior

- Jeffreys prior on **σ** and the Zellner-Siow prior on the **coefficients**

- It is a **mixture of the g-prior** and does not introduce additional information but rather reflects a conservative belief about the coefficients' distribution.

| 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 |

**Fig. 6:** JZS posterior distribution of coefficients

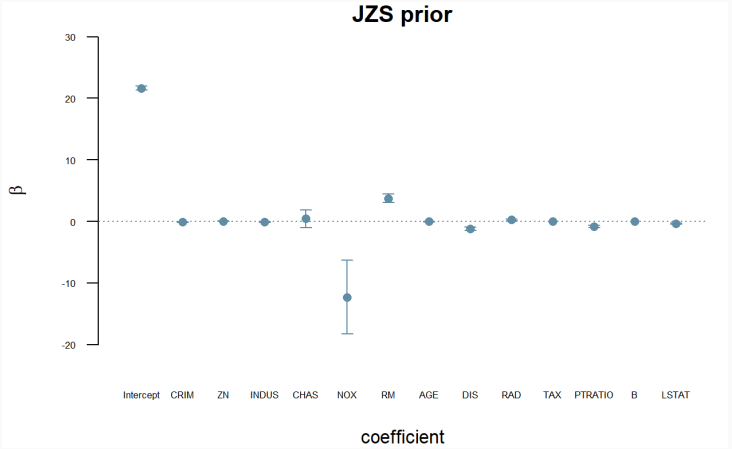


**Fig. 7:** Parameters mean values and confidence intervals with JZS prior

# Bayesian Information Criterion (BIC)

**The Bayesian Information Criterion is defined as follow:**

$$BIC = Kln(n) - 2\ln(\mathcal{L})$$

**Where:**

- $K$ is the number of features
- $n$ is the number of datapoints
- $\mathcal{L}$ is the maximized value of the likelihood function for the model

**It is used for model selection in bayesian linear regression due to it's relatioship with the Bayes Factor:**

$$BF = \frac{p(M1|D)p(M1)}{p(M2|D)p(M2)}$$

$$p(M|D) = e^{-\frac{BIC}{2}} p(M)$$

# Bayesian Information Criterion (BIC)

The BIC in the BAS package consider the following non-informative prior over the regression coefficients:

$$\pi(\beta|\sigma^2) \sim 1$$
$$\pi(\sigma^2) \sim \frac{1}{\sigma^2}$$

Then we define a uniform prior over the possible model and performs backward elimination keeping the model with the smallest BIC value, that is the best BIC model. In the model selection process, we obtain the following inclusion probabilities over the features:

| Features | CRIM | ZN | INDUS | CHAS | NOX | RM | AGE |
|---|---|---|---|---|---|---|---|
| $p(\beta_i \neq 0 \mid Y)$ | 0.994 | 0.931 | 0.061 | 0.050 | 1.000 | 1.000 | 0.358 |
| **Features** | **DIS** | **RAD** | **TAX** | **PTRATIO** | **B** | **LSTAT** | |
| $p(\beta_i \neq 0 \mid Y)$ | 1.000 | 1.000 | 1.000 | 1.000 | 0.974 | 1.000 | |

# Bayesian Information Criterion (BIC)

Then we consider the **best BIC model** and obtain the following value for the regression coefficients:

| 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 |

**Fig. 8:** BIC posterior distribution of coefficients

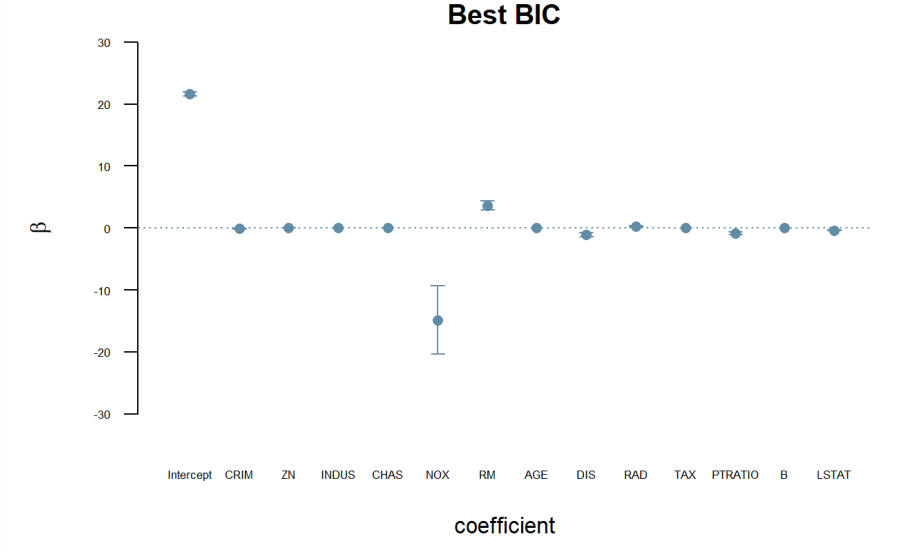


**Fig. 9:** Parameters mean values and confidence intervals with BIC

# Just Another Gibbs Sampler (JAGS) Spike and Slab

**The Spike and Slab defines and index $\gamma$ over the model defined as an array of k element such that for $\beta_i = 0$ we have that $\gamma_i = 0$ and for $\beta_i \neq 0$ we have that $\gamma_i = 1$. This is done using the following prior for the regression coefficients:**

$$\beta_i|\gamma_i \sim (1-\gamma_i)\delta_{\{0\}} + \gamma_i \mathcal{N}(0, \sigma^2_{\beta_i})$$
$$\gamma_i|\theta_i \sim \mathcal{B}e(\theta_i)$$
$$\theta_i \sim \pi(\theta_i)$$

**Over the following model:**

$$Y|X \sim Norm(\mu, \sigma^2)$$
$$\mu = \beta 0 + Xi * \beta$$
$$\tau = \frac{1}{\sigma^2} \sim Gamma(0.001, 0.001)$$
$$(\beta 0, \beta) \sim \pi(\beta 0, \beta)$$

**Then we select the model using two different criteria.**

# JAGS Spike and Slab: Median Probability Model

In the **Median Probability** model, we consider the posterior of the inclusion variable and select the features that have **probability higher than 0.5** to be included.

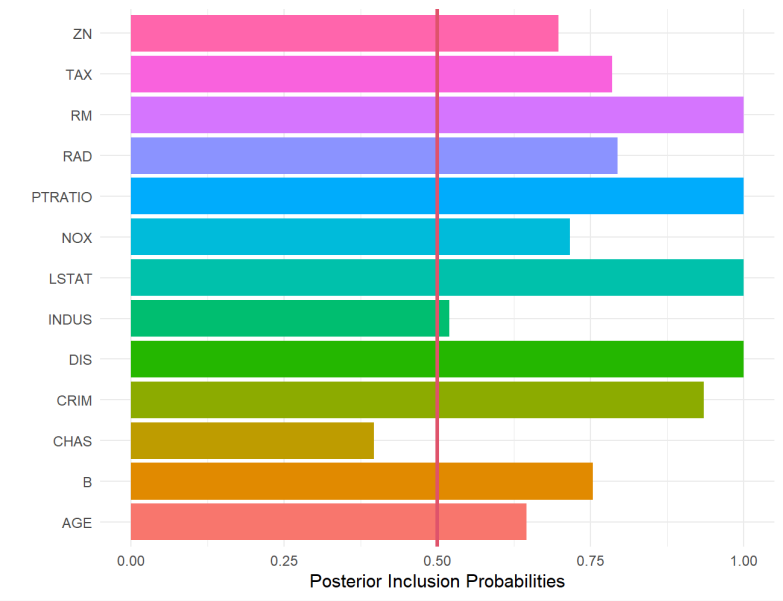


**Fig. 10:** Histogram of the g values used to determine the selected features

In this case **only the CHAS feature is excluded** from the model as all the other features have inclusion probability higher than 0.5

# JAGS Spike and Slab: Highest Posterior Density Model

In the **Highest Posterior Density** model, we select the model that was **visited more times** in the MCMC generated by JAGS.

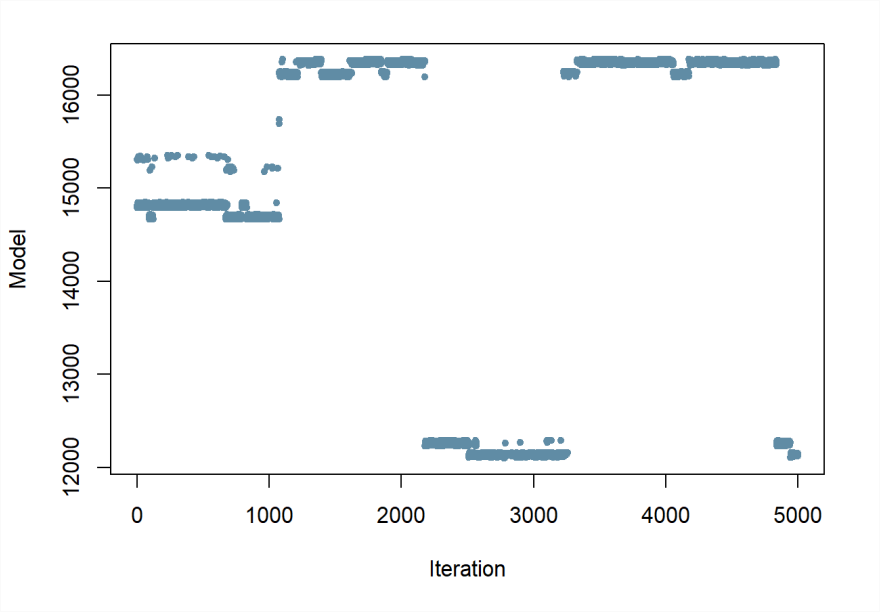


**Fig. 11:** Visited Models JAGS approach

The models are indexed based on the features included, so from the index of the model we can extract what features were excluded.

In this case the features excluded were **INDUS** and **CHAS**.

# Prediction Analysis

- **Train-test split with Shuffling**
  - 70/30 proportion
  - Removing bias
  - 147 test samples
- **Regression Metrics**
  - RMSE
  - MAE
  - MAPE
- **Performance Comparison**

# Regression Metrics

**Root Mean Squared Error (RMSE)**

- Easy interpretability (same unit as the target)
- Large errors penalization

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

**Mean Absolute Error (MAE)**

- Overall performance (equal penalization)

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

**Mean Absolute Percentage Error (MAPE)**

- Most intuitive (eliminates magnitude)
- Ideal for low range target values

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

# Prediction Analysis – Performance Comparison

- Just like the posterior, the predictive distribution of the standard Bayesian Regression showed similar results, proving once again that the model was quite robust and **indifferent to the priors.**
- BIC approach was successful in improving **both performance and complexity**, proving to be a "go to" approach when dealing with simple regression problems.
- Both MCMC model selection techniques (MP & HPD) generated a better performance (this could be a sign of overfitting since the models are more complex).

| 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% |

**Fig. 12:** Comparison of the models' performances

# Conclusions

- ✓ **All feature selection approaches dropped CHAS feature. Proving that, for our analysis, the CHAS column proved to be useless.**
- ✓ **The Bayesian Regression Analysis proved to be relativelly sucessfull, corroborating our choice to use the Normal Linear Regression.**
- ✓ **The project was extremelly useful in reviewing the main points talked in the class and lab sessions.**

# Thank you!


- Felipe A. dos Santos (10919711)
- Gabriel S. Pastorello (10946365)
- Giacomo Savazzi (10675184)