DATA SCIENCE
Journal of Computing
and Applied Informatics

Contents lists available online at TALENTA Publisher

DATA SCIENCE: JOURNAL OF COMPUTING AND APPLIED INFORMATICS (JoCAI)

Journal homepage: https://jocai.usu.ac.id

Bayesian Regression for Predicting Price: Empirical Evidence in American Real Estate

Harry Patria

[a] *School of Computing, Newcastle University, England, United Kingdom*

ARTICLE INFO

Article history:
Received 31 October 2022
Revised 4 December 2022
Accepted 26 January 2023
Published online 31 January 2023

Keywords:
Bayesian linear regression
Bayesian inference
Model
Housing Price

Email:
harry.patria@sbm-itb.ac.id

Corresponding Author:
Harry Patria

ABSTRACT

The two foremost aims of classical regression are to assess the structure and magnitude of the relationship between variables. Despite the aforementioned benefits, unlike classical regression, which only offers a point estimate and a confidence interval, Bayesian regression offers the whole spectrum of inferential solutions. The results of this study demonstrate the Bayesian approach's suitability for regression tasks and its advantage in accounting for additional a priori data, which often strengthens studies. Using data from Boston Housing from UCI ML Repository, this study proves that the prior distributions have the benefit of producing analytical, closed-form conclusions, which eliminates the need to use numerical techniques like Markov Chain Monte Carlo (MCMC). Second, software implementations are offered together with formulas for the posterior outcomes that are supplied, clarified, and shown. The assumptions supporting the suggested approach are evaluated in the third step using Bayesian tools. Prior elicitation, posterior calculation, and robustness to prior uncertainty and model sufficiency are the three processes that are essential to Bayesian inference. *The findings confirm that* Bayesian models provide a full posterior distribution over model parameters, which can be used to quantify uncertainty and make more informed predictions and can also incorporate regularization techniques to reduce overfitting and improve generalization performance.

IEEE style in citing this article:
H. Patria, " Bayesian Regression for Predicting Price: Empirical Evidence in American Real Estate," *Data Science: Journal of Computing and Applied Informatics (JoCAI)*, Vol. 7, No. 1, pp. 15-23, 2023.

1. Introduction

The goal of science is to comprehend phenomena and systems to predict and control their advancement. Models, which are advanced mathematical or algorithmically representations in the language of quantitative sciences, play a vital role in the scientific process of knowledge elaboration found in a wide range of industrial applications such as energy [1,2,3], finance [4,5,6,7,8].

Renting and purchasing houses are becoming more popular as cities become more crowded. As a result, predicting and modeling house prices is a crucial subject in a way for determining a more robust method of calculating house prices [9],[10]. Furthermore, this is going to help both sellers and buyers find the best price for their homes [11], strengthening the economy during a recent global recession. This result, which accurately reflects market price become a hot topic in the real estate sector.

This study aims to shed some light on Bayesian linear regression used to model house prices. One advantage of using a Bayesian model for predicting housing prices is that it allows for the incorporation of prior knowledge or belief about model parameters. This can be especially useful in situations where there is limited data available or the data is noisy. Secondly, Bayesian models also allow for the incorporation of regularization techniques, such

e-ISSN: 2580-829X| p- ISSN: 2580-6769

as shrinkage priors, which can help to reduce overfitting and improve the generalization performance of the model. In addition, Bayesian models can naturally handle missing data, by using appropriate probability models to represent the missing data process.

Another potential advantage of using a Bayesian model for predicting housing prices is that it provides a full posterior distribution over model parameters, which can be used to quantify uncertainty and make more informed predictions. This can be especially useful in a decision-making context, as it allows for the explicit consideration of model uncertainty in decision-making processes. Overall, Bayesian models can be a powerful tool for predicting housing prices and may offer some advantages over traditional models such as multiple linear regression. However, it's important to carefully consider the specific needs and characteristics of a given prediction problem and choose the most appropriate modeling approach accordingly.

The advice is organized around the steps of a Bayesian approach. Section 1 outlines the background and purpose. Section 2 describes the literature review while section 3 describes the elicitation of the dataset and method used in this study such as a prior distribution and the estimated posterior distribution. The distributions of the regression parameters, the regression function, and the model, as well as their estimates and uncertainties, are given and explained in Section 4. Lastly, section 5 summarizes the findings and discussion in a way to generalize the findings and application.

2. Literature Review

The literature makes an effort to extract practical knowledge from historical real estate market data. To find models that are helpful to home buyers and sellers, machine learning techniques are widely used [9,10,11]. Phan [9] investigated historical Australian real estate transactions, showing the significant price disparity between homes in Melbourne's most costly and least expensive suburbs. Additionally, tests show that the Stepwise and Support Vector Machine combo, which is based on mean squared error assessment, is a competitive strategy.

In China, Yu et al. [10] applied multiple prediction models based on deep learning to determine the current real estate data. The results provide a more precise prediction of the housing price or its changing trend in the future. This will allow us to analyze the effects of various factors on housing prices. In light of machine learning, Vineeth et al. [11] scrutinized the house price using machine learning algorithms, ranging from simple linear regression (SLR), Multiple linear regression (MLR), and Neural Networks (NN).

Bayesian regression methods are extremely powerful because they provide us with an entire distribution over the regression parameters rather than just point estimates [12,13,14]. This can be understood as learning not just one model, but an entire family of models and assigning different weights to them based on their likelihood of being correct. Because this weight distribution is affected by the observed data, Bayesian methods can provide us with an uncertainty quantification of our predictions that represents what the model was able to learn from the data [15,16,17,18,19]. The uncertainty measure could be, for example, the standard deviation of all model predictions, which point estimators do not provide by default. Not only transforming AI into more understandable but also changing a paradigm about probability and uncertainty to augment the analytical modeling better.

Bayesian linear regression is a useful tool for forecasting problems, and there are several reasons why it might be preferred over other machine learning techniques such as simple or multiple linear regression or neural networks. Bayesian approaches allow for the incorporation of prior knowledge about model parameters, and can also handle regularization and missing data. In addition, Bayesian linear regression provides a full posterior distribution over model parameters, which can be used to quantify uncertainty and make more informed predictions [15,16,17]. Overall, Bayesian linear regression is a flexible and powerful tool that may be well-suited for certain types of data and modeling scenarios.

3. Methodology

3.1. Data Collection

This study used Boston Housing provided by the UCI repository. Each of the 506 entries provides aggregate information about 14 characteristics of residences from different Boston suburbs, and the data was collected in 1978. The detail of descriptive statistics of the variables can be found in Table 1.

Table 1. Descriptive statistics of the dataset

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
nbr.val	506.00	506.00	506.00	506.00	506.00	506.00	506.00	506.00	506.00	506.00	506.00	506.00	506.00	506.00
nbr.null	0.00	372.00	0.00	471.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
nbr.na	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
min	0.01	0.00	0.46	0.00	0.39	3.56	2.90	1.13	1.00	187.00	12.60	0.32	1.73	5.00
max	88.98	100.00	27.74	1.00	0.87	8.78	100.00	12.13	24.00	711.00	22.00	396.90	37.97	50.00
range	88.97	100.00	27.28	1.00	0.49	5.22	97.10	11.00	23.00	524.00	9.40	396.60	36.24	45.00
sum	1828	5750	5635	35	281	3180	34700	1920	4832	206600	9338	180500	6402	11400
median	0.26	0.00	9.69	0.00	0.54	6.21	77.50	3.21	5.00	330.00	19.05	391.40	11.36	21.20
mean	3.61	11.36	11.14	0.07	0.55	6.29	68.57	3.80	9.55	408.20	18.46	356.70	12.65	22.53
SE.mean	0.38	1.04	0.31	0.01	0.01	0.03	1.25	0.09	0.39	7.49	0.10	4.06	0.32	0.41
CI.mean.0.95	0.75	2.04	0.60	0.02	0.01	0.06	2.46	0.18	0.76	14.72	0.19	7.97	0.62	0.80
var	73.99	543.94	47.06	0.06	0.01	0.49	792.40	4.43	75.82	28400.00	4.69	8335.00	50.99	84.59
std.dev	8.60	23.32	6.86	0.25	0.12	0.70	28.15	2.11	8.71	168.50	2.17	91.29	7.14	9.20
coef.var	2.38	2.05	0.62	3.67	0.21	0.11	0.41	0.55	0.91	0.41	0.12	0.26	0.56	0.41

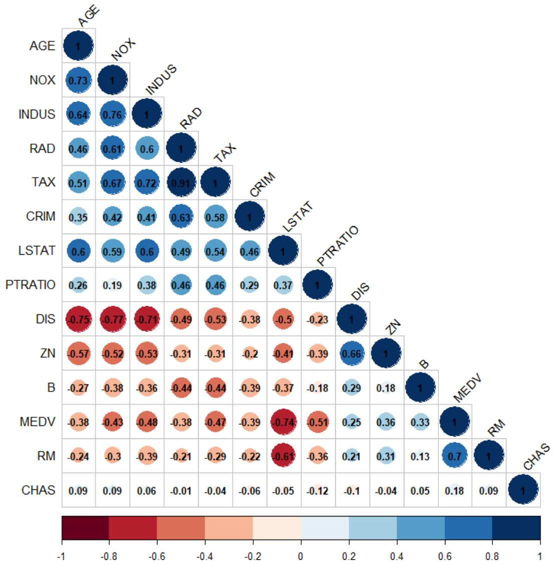


Figure 1. Heatmap and matrix of correlation (Source: Author)

The preliminary analysis using a correlation heatmap describes substantial findings found in Figure 1. First, the average room (RM) and median value of the house (MEDV) are highly correlated, with a score of 0.70. On the other hand, the results found a highly negative correlation between % lower status of the population % lower statusof the population (LSAT) and the aforementioned MEDV, showing a score of -0.61. Since this study merely focuses on house features, therefore, RM will be used in predicting MEDV than LSTAT. In particular, this study scrutinizes the plausible connection between house features and their estimated price by following the Bayesian approach.

3.2. Bayesian Linear Regression

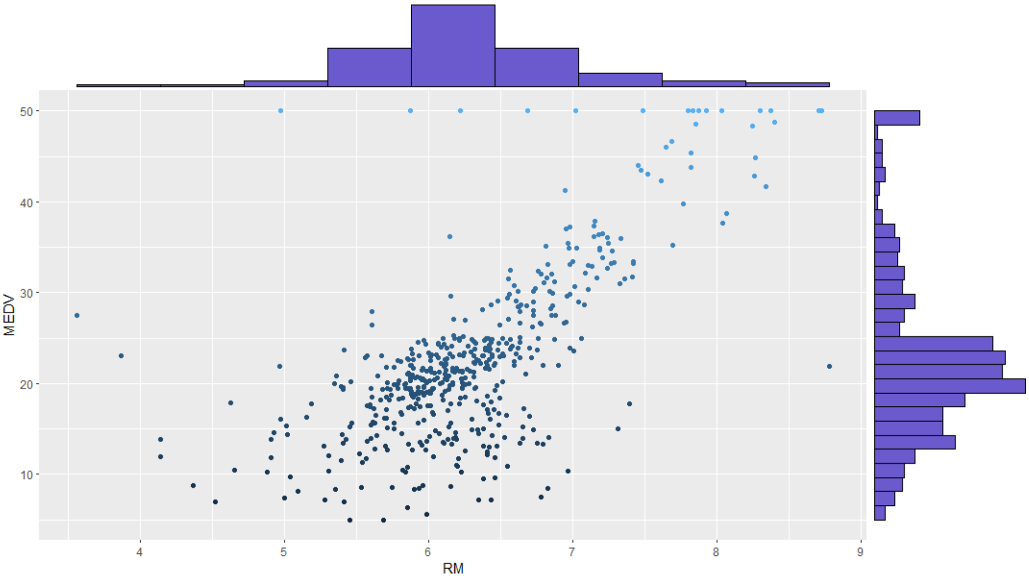
Bayesian linear regression is a statistical approach that is often used to model the relationship between a dependent variable and one or more independent variables. It is based on the idea of using Bayes' theorem to update our belief about the values of the model parameters given the data that we observe. This approach has been widely used in various fields for a variety of purposes, including prediction, estimation, and inference.

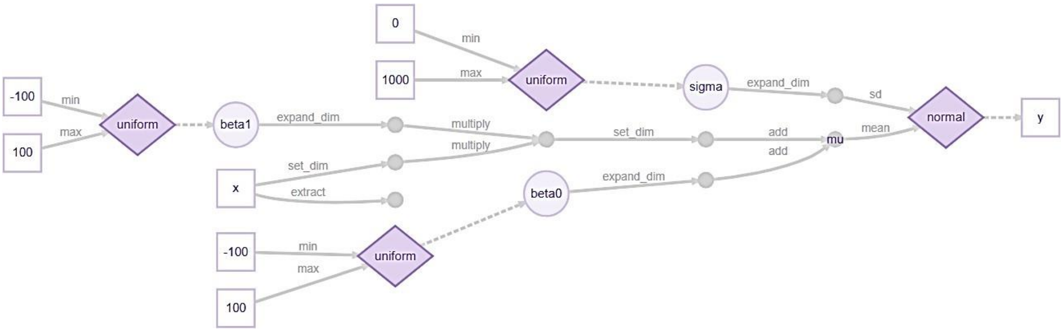
One notable application of Bayesian linear regression is in the field of finance, where it has been used to model a variety of financial data including stock prices, exchange rates, and interest rates [20]. Bayesian linear regression has also been used in the field of engineering, where it has been applied to problems such as surface roughness prediction [21] and short-term travel speed.

This section outlines linear regression using probability distributions rather than point estimates in light of a Bayesian perspective. The response, y, is assumed to be drawn from a probability distribution rather than being estimated as a single value. The Bayesian Linear Regression model with a response sampled from a normal distribution is illustrated by Equation 1 below:

$$y \sim \mathrm{N}(\beta^{\mathrm{T}}\mathrm{X}, \sigma^{2}\mathrm{I}) \qquad (1)$$

A notable symbol of y is defined as an output of a normal (Gaussian) distribution with a mean and variance. For linear regression, the mean is calculated by multiplying the weight matrix by the predictor matrix. The variance is equal to the standard deviation squared (multiplied by the Identity matrix because this is a multi-dimensional formulation of the model).





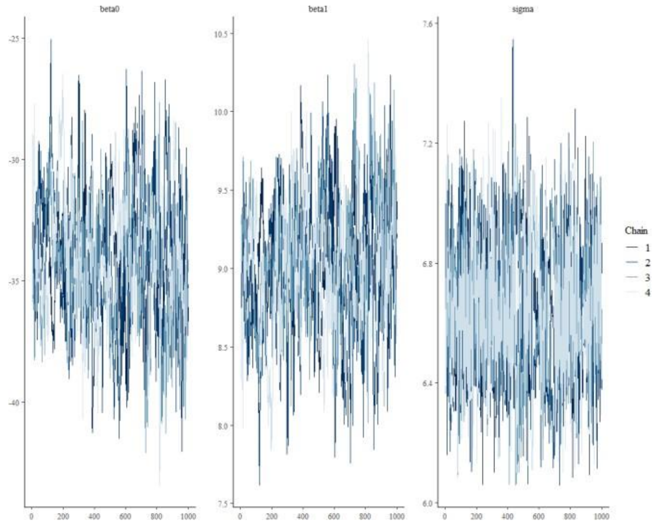


Figure 2. Distribution and Correlation, prior and posterior probability distribution

The goal of Bayesian Linear Regression is not to find a single "best" value for the model parameters, but rather to determine the posterior distribution for the model parameters. Not only is the response generated from a probability distribution, but again the model parameters are also assumed to be generated from a distribution. The posterior probability of the model parameters is completely reliant on the training inputs and outputs:

$$P(\beta|y,X) = \frac{P(y|\beta,X)\ x\ P(\beta|X)}{P(y|X)} \qquad (2)$$

3.3. Posterior Probability Distribution

Given the inputs and outputs, P(|y, X) represents the posterior probability distribution of the model parameters. This is equal to the prior probability of the parameters multiplied by the likelihood of the data, P(y|, X), and divided by a normalization constant. This is a straightforward formulation of the Bayes Theorem, which serves as the cornerstone of Bayesian inference. As opposed to OLS, our model's parameters have a posterior distribution that is proportional to the likelihood of the data times the prior probability of the parameters.

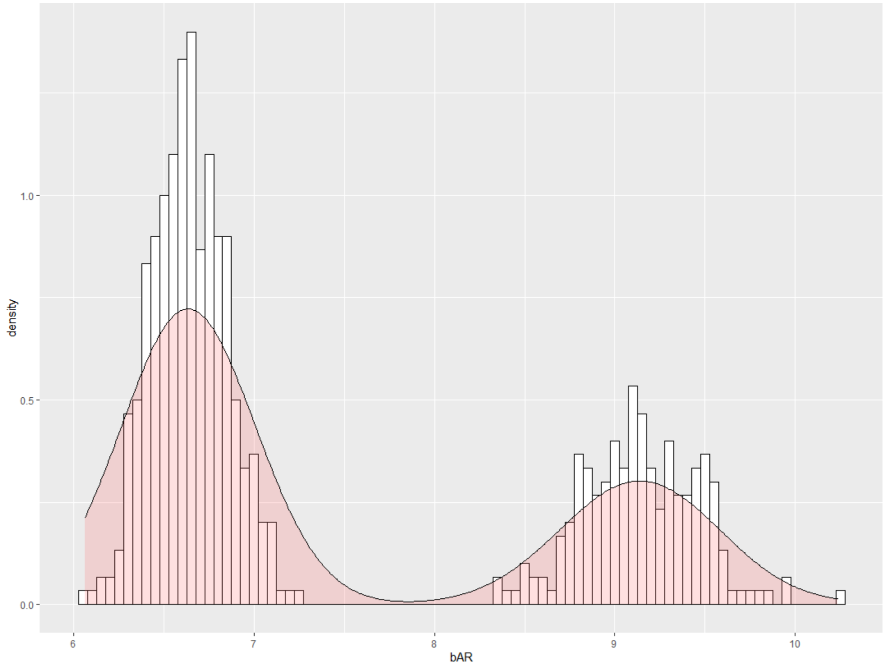


Figure 3. Density and distribution

In terms of priors, we can integrate domain knowledge or an educated estimate as to what the model parameters should be in our model, unlike the frequentist approach, which presupposes that all information about the parameters comes from the data. In this vein, we can utilize non-informative priors for the parameters, like a normal distribution, if we do not already have any estimations.

4. Result and Discussion

The posterior distributions of the model parameters are approximated in Figure 2 and Figure 3. According to the findings, Bayesian models offer a complete posterior distribution of model parameters, enabling quantification of uncertainty and more informed predictions. They also incorporate regularization techniques to prevent overfitting and enhance generalization performance. The model produces the stable outcomes of 1000 MCMC steps, which means that 1000 steps were taken from the posterior distribution by the method.

Table 2. Testing and result validation using 1,000 datapoints

	Mean	SD	Naive SE	Time Ser. SE	2.50%	25%	50%	75%	97.50%
Q_0	-34.09	2.511	0.03971	0.15304	-39.1	-35.76	-34.09	-32.41	-29.03
Q_1	9.01	0.397	0.00628	0.02402	8.23	8.75	9.01	9.28	9.8
σ	6.64	0.217	0.00343	0.00668	6.22	6.49	6.63	6.79	7.07

Instead of just showing an estimate of the linear fit in classical regression, Bayesian allows us to draw a range of lines, each representing a different estimate of the model parameters. Because the model parameters are less uncertain as the number of data points increases, the lines begin to overlap.

Table 3. Testing and result validation using 10,000 datapoints

	Mean	SD	Naive SE	Time Ser. SE	2.50%	25%	50%	75%	97.50%
Q_0	-34.68	2.632	0.01316	0.03408	-39.79	-36.46	-34.7	-32.9	-29.48
Q_1	9.1	0.416	0.00208	0.00544	8.28	8.82	9.11	9.39	9.91
σ	6.63	0.211	0.00105	0.00388	6.23	6.49	6.63	6.77	7.07

To demonstrate the effect of the number of data points in the model, author used two attempts using a different number of data points (1,000 and 10,000), and the resulting fits are shown consecutively in Figures 4 and 5.

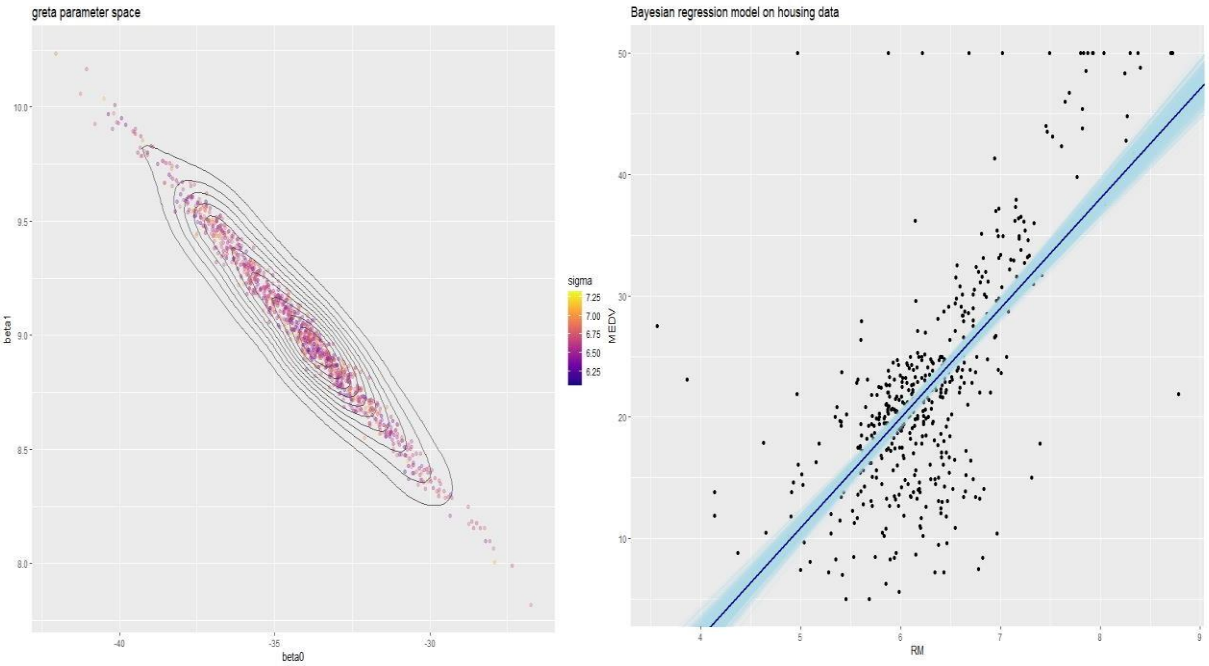


Figure 4. Bayesian Regression of RM and MEDV on the first attempt (1,000 data points)

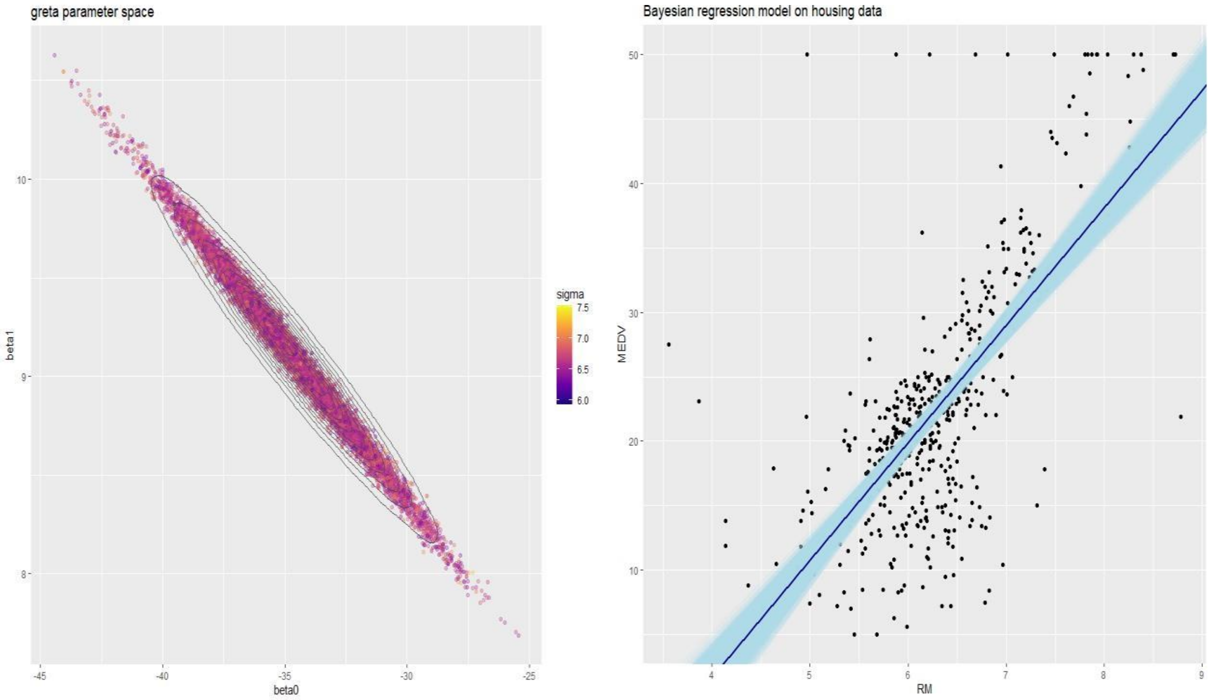


Figure 5. Bayesian Regression of RM and MEDV on the second attempt (10,000 data points)

From the two findings in Figure 4 and Figure 5, it can be concluded that ~~when~~ fewer data points lead to adjustments significantly, indicating more uncertainty in the model. Thus, we also obtain a distribution rather than a single result when using our Bayesian Linear Model to forecast the outcome for a single data point. When it comes to posterior, a distribution of potential model parameters depending on the data and the prior is the outcome of performing Bayesian Linear Regression. With fewer data points, the posterior distribution will be more dispersed, allowing us to measure how doubtful we are about the model.

5. Conclusions

Based on the findings, it can be concluded that Bayesian models are a useful tool for predicting and modeling various types of data. These models provide a full posterior distribution over model parameters, which allows for the quantification of uncertainty and the ability to make more informed predictions. Additionally, Bayesian models can incorporate regularization techniques to reduce overfitting and improve generalization performance. Overall, the use of Bayesian models can be a valuable approach for a wide range of prediction and modeling tasks.

In cases where we have limited data or prior knowledge that we want to incorporate into our model, the Bayesian Linear Regression approach can incorporate prior knowledge while also displaying our uncertainty. The Bayesian framework is reflected in Bayesian Linear Regression: we form an initial estimate and improve it as more data is gathered. The Bayesian approach is a sensible way to perceive the world, and Bayesian Inference can be a valuable counterpoint to frequentism. The goal of data science is to identify the best tool for the job, not to take sides.

The forthcoming study could include a discussion of potential future studies in the conclusion section that further demonstrate the value of the Bayesian approach and how it can be a useful counterpoint to frequentism. This could include outlining specific research questions or directions that could be pursued in future work, and explaining how these studies would help to further illustrate the benefits of using the Bayesian approach in various contexts. Additionally, the authors could consider providing concrete examples of challenges or problems that the Bayesian approach is particularly well-suited to addressing, and explaining how future studies could help to address these challenges and contribute to a better understanding of the usefulness of the Bayesian approach.”.

References

[1] H. Patria and V. Adrison, “Oil Exploration Economics: Empirical Evidence from Indonesian Geological Basins,” Economics and Finance in Indonesia, vol. 61, no. 3, p. 196, Dec. 2015, doi: 10.7454/efi.v61i3.514.

[2] H. Patria, “The Role of Success Rate, Discovery, Appraisal Spending, and Transitioning Region on Exploration Drilling of Oil and Gas in Indonesia in 2004–2015,” Economics and Finance in Indonesia, vol. 67, no. 2, p. 183, Dec. 2021, doi: 10.47291/efi.v67i2.952.

[3] H. Patria, “Predicting the Oil Investment Decision through Data Mining,” Data Science: Journal of Computing and Applied Informatics, vol. 6, no. 1, pp. 1–11, Jan. 2022, doi: 10.32734/jocai.v6.i1-7539.

[4] C. D. Mariana and H. Patria, “Are Electric Vehicle Stocks in ASEAN Countries Investible during the Covid-19 Pandemic?,” IOP Conf Ser Earth Environ Sci, vol. 997, no. 1, p. 012002, Feb. 2022, doi: 10.1088/1755-1315/997/1/012002.

[5] D. Anggraeni, K. Sugiyanto, M. Irwan, Z. Zam, and H. Patria, “STOCK PRICE MOVEMENT PREDICTION USING SUPERVISED MACHINE LEARNING ALGORITHM: The Konstanz Information Miner (KNIME)”.

[6] F. Zulfikri, D. Tryanda, A. Syarif, and H. Patria, “Predicting Peer to Peer Lending Loan Risk Using Classification Approach,” International Journal of Advanced Science Computing and Engineering, vol. 3, no. 2, pp. 94–100, Oct. 2021, doi: 10.30630/ijasce.3.2.57.

[7] H. Patria, “Predicting Fraudulence Transaction under Data Imbalance using Neural Network (Deep Learning),” Data Science: Journal of Computing and Applied Informatics, vol. 6, no. 2, pp. 67–80, Jul. 2022, doi: 10.32734/jocai.v6.i2-8309.

[8] A. Kurniawan, A. Rifa’i, M. A. Nafis, N. S. Andriaswuri, H. Patria, and D. Purwitasari, “Feature Selection and Sensitivity Analysis of Oversampling in Big and Highly Imbalanced Bank’s Credit Data,” in 2022 10th International Conference on Information and Communication Technology (ICoICT), Aug. 2022, pp. 35–40. doi: 10.1109/ICoICT55009.2022.9914889.

[9] T. D. Phan, “Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia,” in 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), Dec. 2018, pp. 35–42. doi: 10.1109/iCMLDE.2018.00017.

[10] L. Yu, C. Jiao, H. Xin, Y. Wang, and K. Wang, “ Prediction on Housing Price Based on Deep Learning,” World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, vol. 12, pp. 90–99, 2018.

[11] N. Vineeth, M. Ayyappa, and B. Varadharajulu, “House Price Prediction Using Machine Learning Algorithms,” in International Conference on Soft Computing Systems, Apr. 2018, pp. 425–433.

[12] M. West, “Outlier Models and Prior Distributions in Bayesian Linear Regression,” Journal of the Royal Statistical Society. Series B (Methodological), vol. 46, no. 3, pp. 431–439, 1984.

[13] T. J. Mitchell and J. J. Beauchamp, “Bayesian variable selection in linear regression. Journal of the american statistical association,” J Am Stat Assoc, vol. 83, no. 404, pp. 1023–1032, 1988.

[14] D. J. C. MacKay, “Bayesian nonlinear modeling for the prediction competition,” ASHRAE Trans, vol. 100, no. 2, pp. 1053–1062, 1994.

[15] A. E. Raftery, D. Madigan, and J. A. Hoeting, “Bayesian model averaging for linear regression models,” J Am Stat Assoc, vol. 92, no. 437, pp. 179–191, Mar. 1997.

[16] T. P. Minka, “Bayesian linear regression,” Technical Report,MIT. 2000.

[17] S. M. Lynch, Introduction to Applied Bayesian Statistics and Estimation for Social Scientists, vol. 1. New York: Springer, 2001.

[18] P. Pérez, G. de Los Campos, J. Crossa, and D. Gianola, “Genomic-Enabled Prediction Based on Molecular Markers and Pedigree Using the Bayesian Linear Regression Package in R,” Plant Genome, vol. 3, no. 2, 2010.

[19] I. Castillo, J. Schmidt-Hieber, and A. van der Vaart, “ Bayesian linear regression with sparse priors,” The Annals of Statistics, vol. 45, no. 3, pp. 1986–2018, 2015.

[20] Y. Chen and D. B. Dunson, “ Posterior convergence for Bayesian functional linear regression,” Biometrika, vol. 100, no. 2, pp. 371–382, 2013.

[21] J. Chen and X. Xu, "Bayesian linear regression for surface roughness prediction," Measurement, vol. 45, no. 8, pp. 1892–1898, 2012.

[22] A. Gelman, B. Goodrich, J. Garby, and A. Vehtari, "R-squared for Bayesian regression models," Am Stat, May 2019.