Bayesian Regression with R Squared on the Housing dataset

Tram Ngo

August 2022

1 Summary

The purpose of this project is to discuss the pros and cons of using R-squared as an evaluation metrics for Bayesian regression tasks. Specifically, the paper is divided into 3 big sections: the exploration of the housing dataset, the study of a Bayesian regression model, and the evaluation of R-squared as a Bayesian regression model's evaluation metrics.

$$R^2 = \frac{SumSquaresofResiduals(SSR)}{SumSquaresTotal(SST)} = \frac{\sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

This discussion is inspired by Andrew Gelman's paper "R-squared for Bayesian regression models" that discusses how R-squared is not the best measures for model accuracy when it comes to a Bayesian regression task. In particular, R-squared does not "reflect the posterior uncertainty in the coefficients" and in the case of strong prior information about the coefficients and not much training data, SSR might be greater than SST resulting in $R^2 > 1$. To illustrate this point, we are given the housing dataset which is comprised of 20 explanatory variables including: the number of bedrooms, number of bathrooms, areas of the house, lattitude, and longitude, etc. Our goal is to learn from around 21,000 houses how these explanatory variables can predict the prices of a new, never-seen-before house. This problem can be formulated as a linear regression model:

$$y = \beta^T x + \epsilon = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \epsilon$$

where X and Y are explanatory vector and target vector respectively. The ϵ value is the error value, whose distribution follows N(0,σ^2).

Through the project and with the illustration of the housing dataset, what I found is that for limited data and some prior information about the parameters of interest, the R-squared metric is actually not representative of how powerful a model's predictive ability is.

2 How Bayesian Regression Works

It is worth learning how Bayesian regression works in comparison to the Ordinary Least Squares regression or Frequentism. Let us look into the similarities and differences through a medical example of a patient with abdominal discomfort going to the doctor's office. Both Frequentist and Bayesian doctors will have some existing medical knowledge base that they will use to examine the patient with. Let's call this knowledge base a "model". Now when the abovementioned patient sees the doctor, the Frequentist doctor will start learning about the patient's symptoms, feed them into the "model", and then makes a diagnosis. A Bayesian doctor, on the other hand, will have had some historical medical data about the patient that he will combine with the "model" he has to arrive at an diagnosis. After a Bayesian doctor makes his diagnosis, he will add this new information to his "model" and to the patient's historical data. Loosely speaking, the frequentists believe that the existing "model" is fixed and data points are variable, and Bayesianists think that the existing "model" varies around the fixed data.

In a more technical term, Bayesian regression is developed based on the concept of Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Briefly, the conditional probability of an event A happening given that event B happens is proportionate to the product of the conditional probability of an event B happening given that event A happens (likelihood) and the probability of event A happens on its own (prior information). The term P(B) - the probability of event B happening on its own - is what we call a constant term that exists in the formula so the final probability P(A—B) conforms to the definition of a probability term. Rewriting this in terms of likelihood and prior information to obtain our estimation of the parameters β, we get the following formula:

$$P(\hat{\beta}|data) \propto P(data|\beta)P(\beta)$$

where $P(data|\beta)$ represents the likelihood of observing the data given the parameters whose distribution we have beliefs about. $P(\beta)$ represents the prior distribution of β that we know prior to the data observation. Prior knowledge is subjective, which could lead to different posteriors about the parameters of interest. There are generally 2 types of priors: informative and non-informative or naive priors. The former states a subjective belief about the parameters while the latter does not contain any prior beliefs and could convert a Bayesian model into an Ordinary Least Squares model that does not need prior belief.

3 Exploration Data Analysis

There are about 21,000 houses or records, each with 20 features including the price - our target variable. Our task is to predict the price of new, unseen houses based on the records given and their characteristics. Houses have features such as number of bedrooms, number of bathrooms, the area of the house, the area of the lot, and so on. It is worth examining the dataset by learning about the distribution of each of the 20 features (Figure 1) with histograms. For example, the price of most houses lies under 1 million dollars; similarly, the majority of houses have an area of under 3000 square feet.

Even though the number of explanatory variables for our task is not too big to cause much noise in our model's predictive ability, it is still worth it to understand the correlation among the features with scatter-plots (Figure 2) and the correlation between the target variable (price) and the remaining features (Table 1).

We can see that the most correlated variables (correlation score greater than 0.5) with our "price" variable are "sqft_living", "grade", and "sqft_above" while the least correlated ones are "zipcode", "long" (longtitude), "condition", and "yr_built". In that sense, it is appropriate to choose the most correlated 5 features, whose correlation score is greater than 0.5, to use as predictive variables $x_i(s)$: "sqft_living", "grade", "sqft_above", "sqft_living15", "bathrooms". Detailed exploratory work is provided in the Code Appendix.

4 Bayesian Model Training With Truncated Data

To illustrate the main point of this project, we need to mock the situation where weak data and strong prior information are present. We are training a Bayesian regression model with truncated data from the housing dataset. Specifically, we only use the first 10 houses as data points into our Bayesian model, with β parameters following a N($\mu = 0$, $\sigma = 10$). Feeding these into the regression model gives us posterior means for our parameters of interest: $\beta(s)$ as in Figure 3: Mean Posterior Parameters. It can be interpreted from the chart that the 94 percent Credible intervals contain the true parameters with 94 percent chance. For example, the parameter β_1 for the "sqft_living" feature has its posterior mean of -0.47, with the credible interval being (-0.74, -0.19). We can intuitively say that this interval contains the true mean of β_1.

The mean R-squared for our toy project of 10 data points after sampling 1000 predictions is -5.834, which is a negative number and which means that the explained variance from the model exceeds the total variance of the data.

5 Model Evaluation using R-squared

A brief definition of R-squared (or coefficient of determination) in regression analysis can be that it is the "measure of goodness-of-fit" for many types of regression tasks according to A.Colin Cameron. In particular, it measures the explanatory power of the regression model at hand by evaluating the ratio of the variance explained by the model over the total variance in the data. R-squared, by definition, ranges from 0 to 100 percent, or from 0 to 1. There are two limitations when it comes to using R-squared to evaluate a Bayesian regression task, as outlined by Gelman:

1. It can be outside of its defined bound, which means that it can get beyond the 100 percent or 1 value. The case usually happens when we have weak or very limited data yet strong prior information about the predicted parameters. This is caused by the fact that in such a case, "it is possible for the fitted variance" (SSR) to be greater than the total variance (SST), which ultimately makes the ratio greater than 1 or 100 percent. While training the truncated housing dataset where we feed smaller subsets of data into our Bayesian model, we did notice R-squared to be greater than 1.

$$R^2 = \frac{SumSquaresofResiduals(SSR)}{SumSquaresTotal(SST)} = \frac{\sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

2. The irrelevance of an increase or decrease in the value of R-squared when it comes to explanatory power. R-squared when used as a standard Linear regression model is a point estimate. Therefore, it is common to interpret R-squared as it reflects better explanatory power of the model when it increases and worse when it decreases. However, in a Bayesian model, posterior parameters' probability or uncertainty is desired, so R-squared fluctuation in value might not necessarily accurately reflect our end results. As we see in Figure 4, an increase in R-squared is irrelevant and cannot be used to reflect the true changes in the model's predictive ability.

6 R-squared Alternatives

According to Gelman, R-squared can be specified into an alternative R-squared that reflects the posterior parameter uncertainty as well as keeps the bound of R-squared below 1:

$$\texttt{Alternative R}^2 = \frac{\texttt{explained var}}{\texttt{explained var} + \texttt{residual var}} = \frac{var_{fit}}{var_{fit} + var_{residual}}$$

Another alternative metric to evaluate a Bayesian model in the case of limited data is to use Leave-one-out cross validation, which was outlined in the paper "Practical Bayesian model evaluation using leave-one-out cross-validation

(LOOCV) and WAIC" by Andrew Gelman and Aki Vehtari. It is worth noting that the use of cross validation in machine learning is common and easy to employ. LOOCV is an extreme case of cross validation when one data point is left out the validation point while the rest of the dataset is trained. This method can be applied to a Bayesian regression model for its low bias and great adaptation to small datasets because of such extremity in the amount used to train versus the amount used to validate.

Information criteria such as AIC (Akaike information criterion), DIC (Deviance information criterion), and WAIC (Watanabe–Akaike information criterion) can also be used to compare likelihoods, which can ultimately help in the comparison of models. Moreover, in the highly insightful paper "Evaluating Bayesian Models with Posterior Dispersion Indices", Kucukelbir also discusses the concept of "Posterior Dispersion Indices" to combine posterior dispersion along with predictive accuracy of a model. In this method, Kucukelbir believes that it is useful to learn in-depth how different data points contribute to the dispersion of the probability of posterior parameters.

7 Conclusion

In summary, even though R-squared is a quick, easy to interpret, and ubiquitous metric to evaluate a regression model, it is not always useful and relevant. In fact, for regression models that use the Bayesian technique to include uncertainty of parameter likelihood, R-squared, being a point estimate, poses many limitations outlined in this paper.

As the continuation of this project, I personally think an interesting subject to study further is the approach that uses experiments and prior information simultaneously. This approach incorporates the advantages and limitations of both Bayesian and Frequentist methods. The interplay of the two methods is well explained and suggested in the paper "The Interplay of Bayesian and Frequentist Analysis" by Bayarri and Berger.

8 Appendix

8.1 Table Appendix

Feature	Correlation	Feature	Correlation
price	1.000	floors	0.253
sqft_living	0.702	waterfront	0.253
grade	0.665	yr_renovated	0.128
sqft_above	0.603	sqrft_lot	0.091
sqft_living15	0.583	sqrft_lot15	0.079
bathrooms	0.527	yr_built	0.049
view	0.392	condition	0.036
sqft_basement	0.321	long (longtitude)	0.023
lat (lattitude)	0.311	zipcode	-0.054
bedrooms	0.308		

Table 1: Correlation table

8.2 Graph Appendix

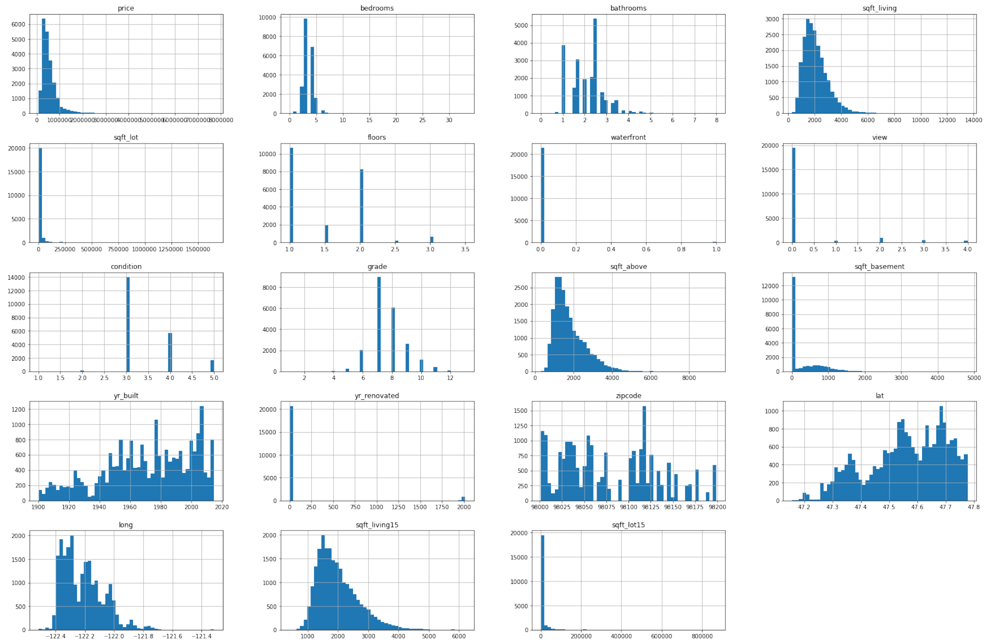

Figure 1: Distributions of each feature

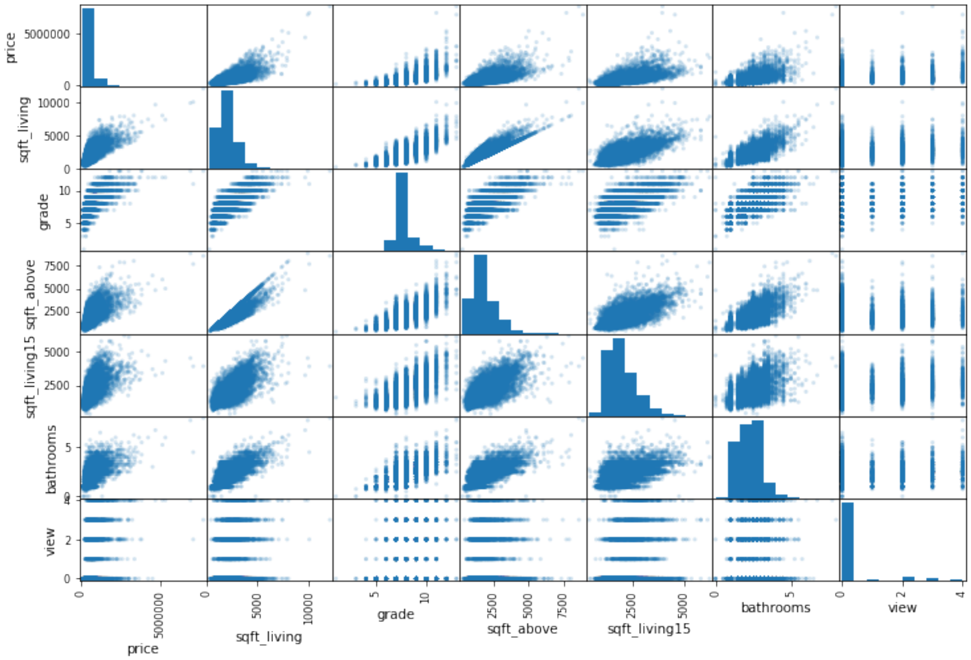

Figure 2: Correlation between price and top 6 most correlated features

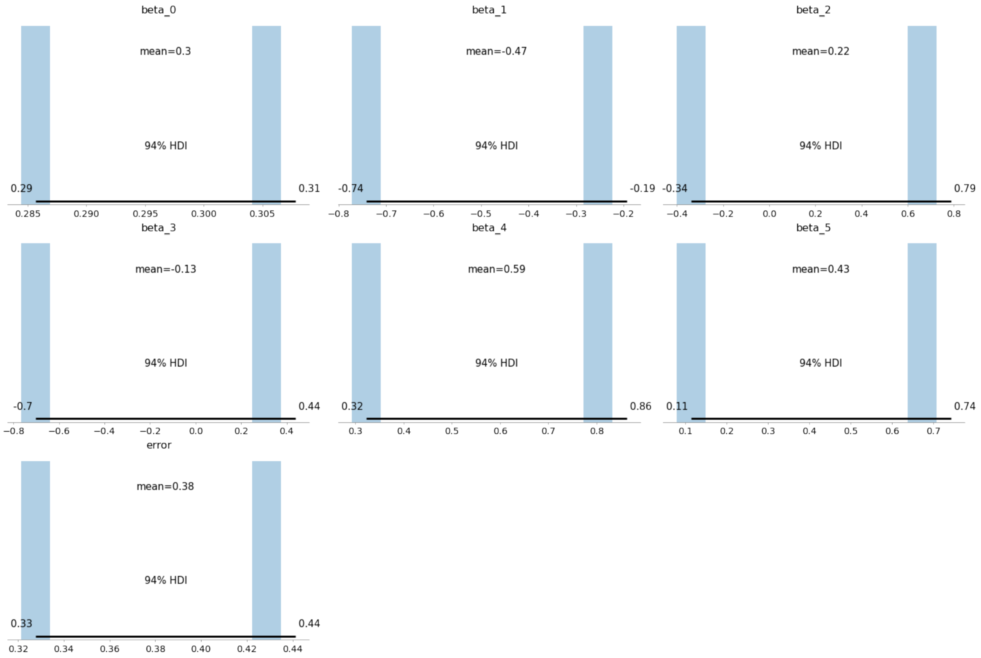

Figure 3: Mean posterior parameters

8.3 Code Appendix

The code that produces the graphs, charts, and tables for this project is hosted at: https://github.com/tramngo1603/Bayesian_Regression_JHU

9. Bibliography

Bayarri, M. J., & Berger, J. O. (n.d.). *The interplay of bayesian and frequentist analysis*. Project Euclid. Retrieved August 22, 2022, from https://projecteuclid.org/journals/statistical-science/volume-19/issue-1/The-Interplay-of-Bayesian-and-Frequentist-Analysis/10.1214/088342304000000116.full?tab=ArticleLinkCited

Bishop, CM & Tipping, ME 2003, Bayesian Regression and Classification. in JAK Suykens, I Horvath, S Basu, C Micchelli & JV (eds), *Advances in Learning Theory: Methods, Models and Applications*. NATO Science Series, III: Computer and Systems Sciences, vol. 190, IOS Press, pp. 267-285.

Casella, George, and Roger L. Berger. *Statistical Inference*. 2nd ed., Brooks/Cole Cengage Learning, 2002.

Cameron, A. C., & Windmeijer, F. A. G. (1998, June 11). *An R-squared measure of goodness of fit for some common nonlinear regression models*. Journal of Econometrics. Retrieved August 22, 2022, from https://www.sciencedirect.com/science/article/pii/S0304407696018180

Gelman, Andrew. “R-Squared for Bayesian Regression Models.” *Taylor & Francis*, https://www.tandfonline.com/doi/full/10.1080/00031305.2018.1549100.

Gelman, A., & Vehtari, A. (n.d.). *Practical bayesian model evaluation using leave-one-out cross* ... Retrieved August 23, 2022, from http://www.stat.columbia.edu/~gelman/research/published/loo_stan.pdf

Hastie, Trevor, et al. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2017.

Knuth, Kevin H. *Bayesian Inference and Maximum Entropy Methods In Science and Engineering : 27th International Workshop On Bayesian Inference and Maximum Entropy Methods In Science and Engineering, Saratoga Springs, New York, 8-13 July 2007*. Melville, NY: American Institute of Physics, 2007.

Kucukelbir, A., Wang, Y., & Blei, D. M. (2017, July 17). *Evaluating bayesian models with posterior dispersion indices*. PMLR. Retrieved August 22, 2022, from http://proceedings.mlr.press/v70/kucukelbir17a.html

Mitchell, T. J. “Bayesian Variable Selection in Linear Regression.” *Taylor & Francis*, https://www.tandfonline.com/doi/abs/10.1080/01621459.1988.10478694.

WAKEFIELD, JON. *Bayesian and Frequentist Regression Methods*. SPRINGER-VERLAG NEW YORK, 2016.