Applied Bayes : Price of house per unit area Evaluation in Taiwan

Yitbarek Demesie

Contents

Abstract or Executive Summary

The goal of the project is to build a Bayesian linear regression predictive model that is able to predict price of real estate valuations properties in Taiwan. The unit of measurement for price of real estate valuations per unit area is given by 0000 New Taiwan Dollar/Ping where 1 Ping = $3.3m^2$. The projects aims to predict valuation of a house per unit area after accounting for age of the house(in years), distance to the nearest MRT station (in meter) and number of convenience stores in the living circle on foot (integer). Specifically, we are interested in if increasing any of the explanatory variables leads to the appreciation of the value of the property(both statistically and practically). To measure this, we will be performing one-sided hypothesis testing. Such a model is ideal for investors to determine appropriate valuations for investors interested in buying new properties in the Sindian District, New Taipei City. Additionally, the ability to predict price of real estate valuations per unit area allows renters to make informed decision on whether to hold or sell the property , rent or own a property. This project used 10-fold cross validation technique to measure the predictive accuracy of valuation of properties using weakly informative priors. We used mean absolute error(MAE) to measure model predictive accuracy. Using 10-fold CV estimates of posterior predictive accuracy model we found that the model mean absolute error is 4.90852. This means each of our prediction was with in 4.9 thousands of Taiwan dollar per unit area. Hence, it tells how far off we are in-terms of predicting valuation of a house per unit area from the true valuation of a house per unit area. This mae of 4.90852is quite well compared to mae of 10.7 by predicting average price for all observations. The paper discusses in detail how the results were obtained, prior assumptions and posterior predictive accuracy and hypothesis testing for each of estimates in increasing evaluation of properties price per unit area in Taiwan.

Section 1: Research Question and EDA

The data set of our interest has 414 number of rows or observations and 5 columns. The number of rows corresponds to the market historical data set of real estate valuation collected from Sindian District, New Taipei City, Taiwan. The transactions are recorded during 2012 and 2013. Each observation or row is a unique transaction on a house during that period of time. Additionally, the data set is complete meanining it hasno missing data.

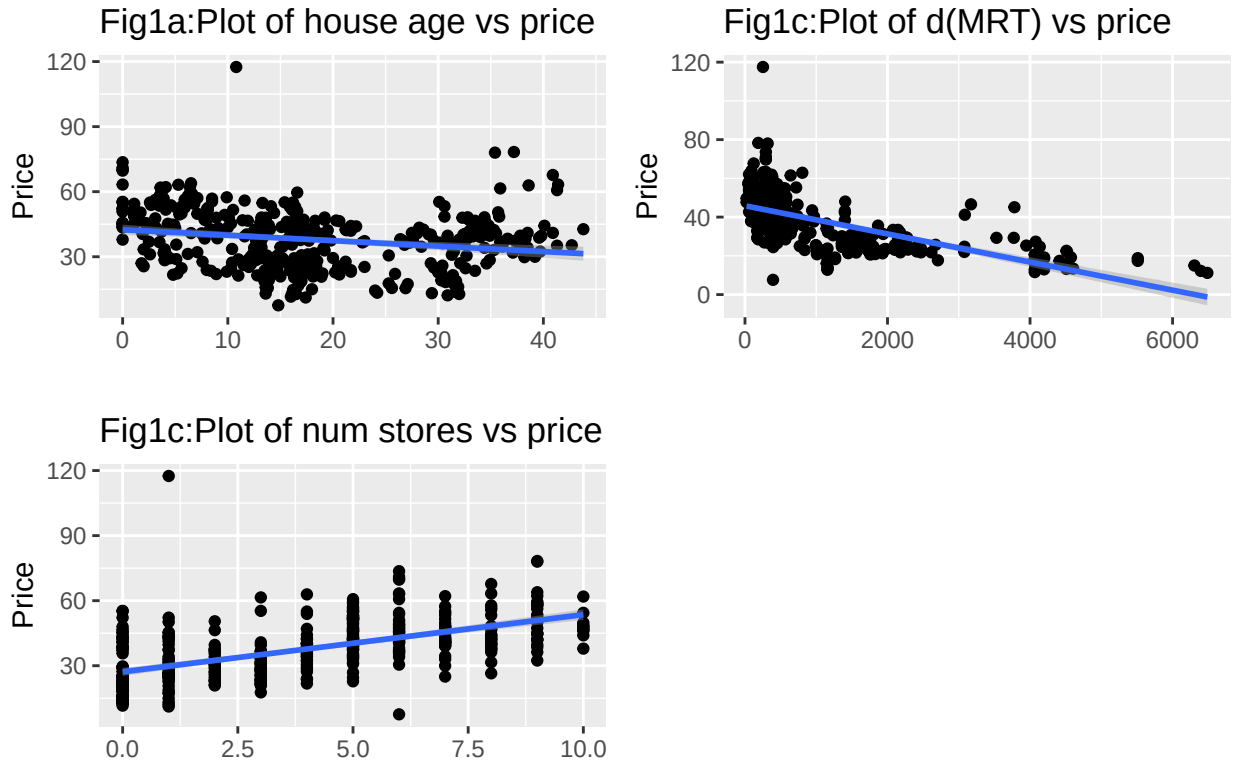
For the research question, we are interested in predicting the house price of unit area given the house age, the number of convenience stores in the living circle on foot, and the distance to the nearest MRT station. Our independent variable is house price of unit area. The house price of a unit area is recorded as per 10000 New Taiwan Dollar/Ping. Ping Ping is a local unit and 1 Ping = $3.3m^2$. On the other hand, the dependent variables house age, distance to the nearest MRT station and number of convenience stores in living circle on foot area measured by units of years, meter, and count(integer) respectively.

Table 1: Table 1:Summary Statistics for Data

	Mean	SD	Min	Max
The house age (unit: year)	17.71	11.39	0.00	43.80
The distance to the nearest MRT station (unit: meter)	1083.89	1262.11	23.38	6488.02
The number of convenience stores in the living circle on foot (integer)	4.09	2.95	0.00	10.00
House price of unit area (10000 New Taiwan Dollar/Ping)	37.98	13.61	7.60	117.50

As we can see from the above Table, in our data set the house age ranges from 0 to 43.8 with mean 17.71 and standard deviation of 11.39. In our data set the distance to the nearest MRT station ranges from 23.38 to 6488.02 with mean 1083.89 and standard deviation of 1262.11. The number of convenience stores in the living circle on foot ranges from 0 to 10 with mean 4.09 and standard deviation of 2.95. Moreover,house price of unit area (10000 New Taiwan Dollar/Ping) ranges from 7.6 to 117.5 with mean 37.98 and standard deviation of 13.61.

Figure 1: Graph to Check for Linear Assumptions



As we can see from Figure 1, we can confirm that the normal regression assumptions are met. This is because each of the scatter plots of our dependent variables(house age in years, distance to the nearest MRT station in meters, number of convenience plots in living circle) vs house price of a unit area are all linear. Hence, this

confirms the assumption that the structure of the relation between our dependent and independent variables is linear to perform an Ordinary Least Squares(OLS) regression. Additionally, structure of the data shows independence among each observation. That is, accounting for all our predictors(dependent variables) say X, the observed housing price price unit area (Y_i) on case i is independent of any other case j. While, we have yet to check on the structure of variability assumption(observed values of Y will vary normally around their average and standard deviation) we can proceed to specifying our priors and building our bayesian regression model.

Section 2: The Bayesian Model and Posterior convergence

Once we have confirmed that we can use linear regression, the next step would be to specify our prior parameters. In the data model, we have four data variables, 1 response variable Y and 3 explanatory variables in X. As a result, we have seven unknown regression parameters that encode the relationship between price of house per unit with our dependent variables. Since, there was limited reliable informative prior information on the prior distributions, I have resorted to weakly informative priors. Using weak informative prior assumptions, the Bayesian linear regression is given by:

data: $Y_i|\beta_0^C, \beta_1^C, \beta_2^C, \beta_3^C \overset{ind}{\sim} N(\mu_i, \sigma^2)$ \

with $\mu_i = \beta_0 + \beta_1^C * age_i + \beta_2^C * dMRT_i + \beta_3^C * NCStore_i$

The priors:

$\beta_0 \sim N(m_0, s_0^2) = N(37.98, 13.60649^2)$

$\beta_1 \sim N(m_1, s_1^2) = N(0, 0.2986^2)$

$\beta_2 \sim N(m_1, s_1^2) = N(0, 0.0.0027^2)$

$\beta_3 \sim N(m_3, s_3^2) = N(0, 1.1548^2)$

$\sigma \sim Exp(l) = Exp(0.073)$

Once we specified the structure of our data using the weakly informative priors, our next step is to simulate MCMC posterior post-burn in draws. For this paper, the burn-in sample was 1000 with 20000 iterations for each of the four chains. As a result, we have 19000 post-burn in samples for posterior inference from each chain for analysis or inference.

After tossing out the first 1000 iterations of Markov chain values from the burn-in phase, the stan_glm() simulation produces four parallel chains of length 20000 for each model parameter: $\beta_0^{(1)}, \beta_0^{(2)}, \ldots, \beta_0^{(20000)}$, $\beta_1^{(1)}, \beta_1^{(2)}, \ldots, \beta_1^{(20000)}$, $\beta_2^{(1)}, \beta_2^{(2)}, \ldots, \beta_2^{(20000)}$, $\beta_3^{(1)}, \beta_3^{(2)}, \ldots, \beta_3^{(20000)}$, and $\sigma^{(1)}, \sigma^{(2)}, \ldots, \sigma^{(20000)}$ .These are stored as (Intercept), age , dMRT, N_CStores, and sigma respectively. The results are summarized in Table 2 below:

Table 2: Table 2: Effective Sample Size Ratio and Rhat

	(Intercept)	age	dMRT	N_CStores	sigma
Effective Sample Size	1.0258289	1.1281974	0.7961842	0.7691974	1.105816
Rhat	0.9999604	0.9999763	0.9999810	0.9999916	1.000011

As we can see from Table 2, using diagnostics indicate that these chains are independent sample, stable, mixing quickly, and trustworthy. This follows from the observation that the effective sample size ratios are above 1 for house of age, and sigma, indicating the independence of out posterior post-burn in samples.However, he effective sample size of distance to the nearest MRT and number of stores in living circle are close to 0.8. This means that for every 10 post-burn in samples, we get 8 independent post-burn in samples. Hence, we would consider this sufficient as the number of independent post-burn in samples is very close to the post-burn samples. Additionally, $\hat{R}$ values are close to 1.This confirms that the posterior post-burn in draws

for all four of our chains have converged and we can trust the results.Hence, we can confirm that by effective size ration our MCMC samples are efficient and converging by their respective $\hat{R}$ values.

An additional check we postponed was to check if our prior assumptions were reasonable. In order to do that we can perform a a posterior predictive check (PPC). A PPC means that we use our MCMC samples to simulate our response variable Y from the posterior predictive distribution.

Figure 2a: Posterior Predictive Check(PPC) for 100 simulations

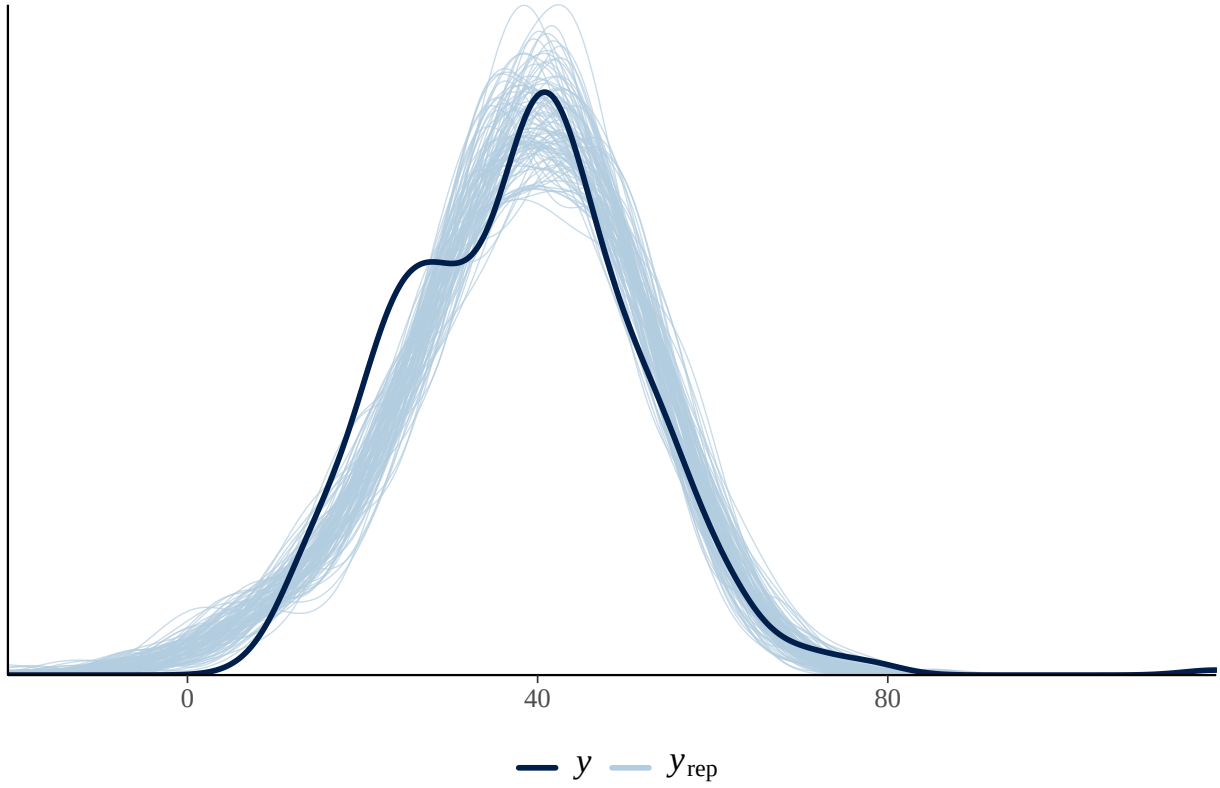
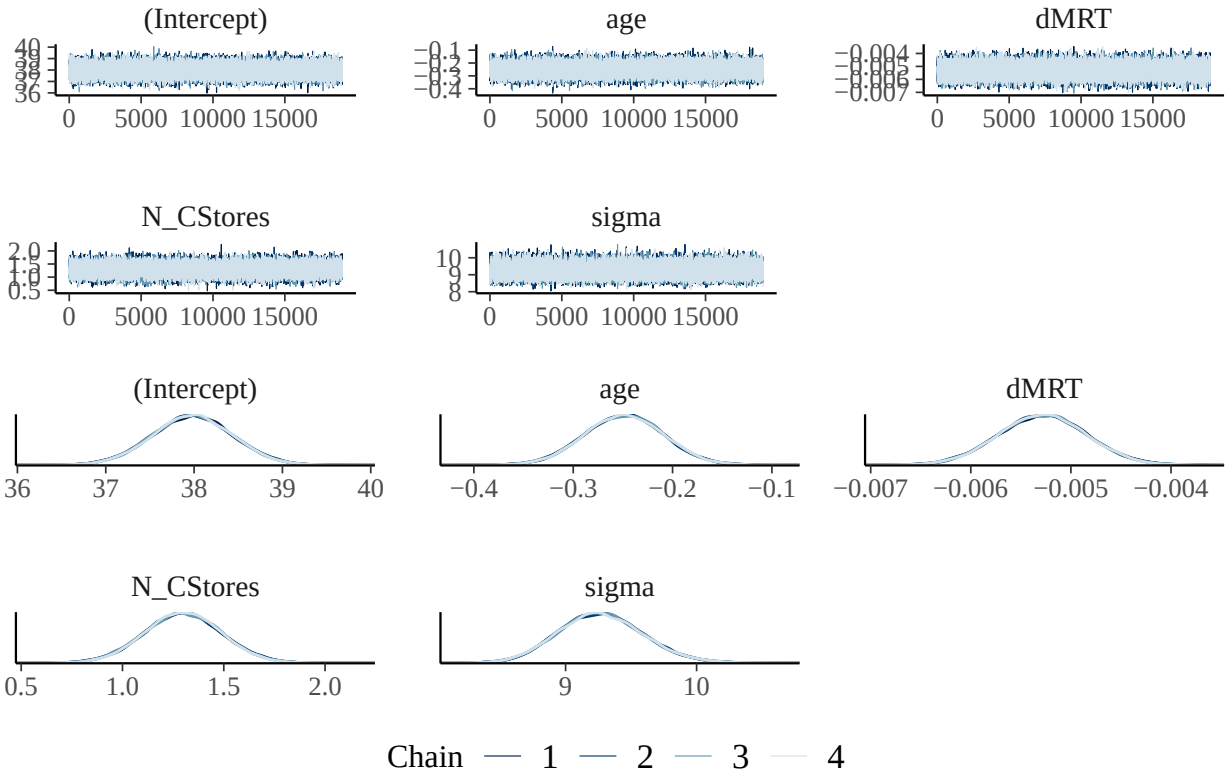


Figure 2b: Trace Plots and Density Plots for Parallel Chain(4)



As we can see from Figure 2a, after performing PPC for 100 simulations, we can conclude that the the assumptions we made while building the model match the actual data. That is our simulations of posterior prediction conditional on our MCMC draws of all parameters aligns with the weakly informative prior we specified for the data. This is to be expected as our prior assumptions are obtained from the data as a result of lack of prior information. We can reach at the sample conclusion from trace and density plots in Figure 2b as well.

Section 3: Posterior Inference

So far we have assess convergence and the necessary MCMC checks to check if prior assumptions were reasonable. Since we have confirmed convergence and sensibility of our prior specification, we can continue to make posterior inference.Hence, in this section, we will Construct and interpret 95% HPD intervals (or 95% posterior credible intervals) for all model parameters.

Table 3: Table 3: Model Parameters and 95% HPD Intervals

Term	Estimate	Standard Error	Lower Credible Interval	Upper Credible Interval
(Intercept)	37.9831334	0.4540359	37.0862322	38.8745242
age	-0.2486497	0.0401342	-0.3273101	-0.1702280
dMRT	-0.0052800	0.0004427	-0.0061525	-0.0044112
N_CStores	1.2984805	0.1903949	0.9252949	1.6741753
sigma	9.2644889	0.3229443	8.6669170	9.9338895
mean_PPD	37.9844727	0.6478912	36.7109329	39.2516210

Table 3 summarizes the model estimates and we can conclude that on an average house with average age,

average distance to MRT and average number of stores in the near circle, house prices are typically around 38, though this average could be somewhere between 37.092 and 38.908. Additionally, For every 1 year increase in house of age, house price per unit area typically decreases by 0.249 thousands of Taiwan dollar per ping, all else equal. However, this this average reduction in the price of the house per unit area could be as low as 0.3292 or as high as 0.1688 thousands of Taiwan dollar per ping. Additionally, for every 1 meter increase in the distance to the nearest MRT station, house price per unit area typically decreases by 0.00528 thousands of Taiwan dollar per ping, all else equal.However, this this average reduction in the price of the house per unit area could be as low as 0.006166 or as high as 0.004394 thousands of Taiwan dollar per ping. Moreover, for every 1 additional increase in the number of convenience stores in the living circle on foot, house price per unit area typically increases by 1.3 thousands of Taiwan dollar per ping, all else equal.However, this this average increase in the price of the house per unit area could be as low as 0.92 or as high as 1.68 thousands of Taiwan dollar per ping.

So far, we have explained what the estimates tell us. The next step would be to perform 1-sided Bayesian hypothesis test. We are interested in testing if each of our parameters $\beta_1^C, \beta_2^C, \beta_3^C$ are greater than 0. The reason for choosing if each of the estimate is greater than 0 to is to allow owners to see which of the estimates lead to statistcally significant appreciation of the properties in price per unit area. The table below summarizes the posterior probability that each of our estimates are greater than 0.

Table 4: Table 4: Hypothesis Test(1-sided) and are each Estimates > 0?

	Probability(out of 100)
The house age (unit: year)	0
The distance to the nearest MRT station (unit: meter)	0
The number of convenience stores in the living circle on foot (integer)	100

As we can see from Table 4, the only way to increase price of house per unit area(in thousands of Taiwan dollar per ping) is by increasing the number of convenience stores in the living circle on foot. There is 100% posterior probability that the house per unit area of a house increases by increasing the the number of convenience stores in the living circle.The additional increase in age of a house and its distance to the nearest MRT station don't lead to an appreciation on the value of a house in Taiwan per unit area.

Now we understand that increasing the number of convenience stores in the living circle on foot appreciates the value of a house. We are now interested in the practically significant relationships between the response variable and the explanatory variables. By assuming that inflation in Taiwan is close to 3%, which is standard rate for most middle-income countries. Since we are interested in the increase in the price of house per unit area and the average house per unit area is 37.98 thousands Taiwan dollar per ping. Therefore, one would consider an increase in in the price of house per unit area 0.03(37.98)= 1.1394 to be practically significant.

Table 5: Table 5: Practical Hypothesis Test(1-sided) and are each Estimates > 1.1394?

	Probability(out of 100)
The house age (unit: year)	0.00000
The distance to the nearest MRT station (unit: meter)	0.00000
The number of convenience stores in the living circle on foot (integer)	80.02632

As we can see from Table 5 above, there is 80% posterior probability that $\beta_3^c > 1.1394$. This is practically significant result as significant majority of our β_3^c estimates are bigger than one would consider an increase in in the price of house per unit area which is 1.1394 to be practically significant. As a result, the best strategy for any new home buyer is to find properties where there is investment plans to increase the number of convenience stores in the living circle on foot.

Section 4: Prediction

So far, we have discussed about our posterior inference. In this section, we will use our model to make a prediction for one new observation and make a prediction interval for that observation. Additionally, we will assess the accuracy of our prediction using 10-fold cross validation(CV).

Assume that we have a new observation with the following the house age(in years), The distance to the nearest MRT station(in meters) and The number of convenience stores in the living circle on foot is given by 10, 1200 and 6 respectively.

After scaling each of our continuous explanatory variables and plugging them back in to our model, we can conclude that there is 95% posterior predictive probability that the house price of unit area (10000 New Taiwan Dollar/Ping) is between 40 and 44. Since we made up the observation above we won't be able to assess predictive accuracy. However, we can assess predictive accuracy of our model on our whole training data set and using 10-fold cross validation.

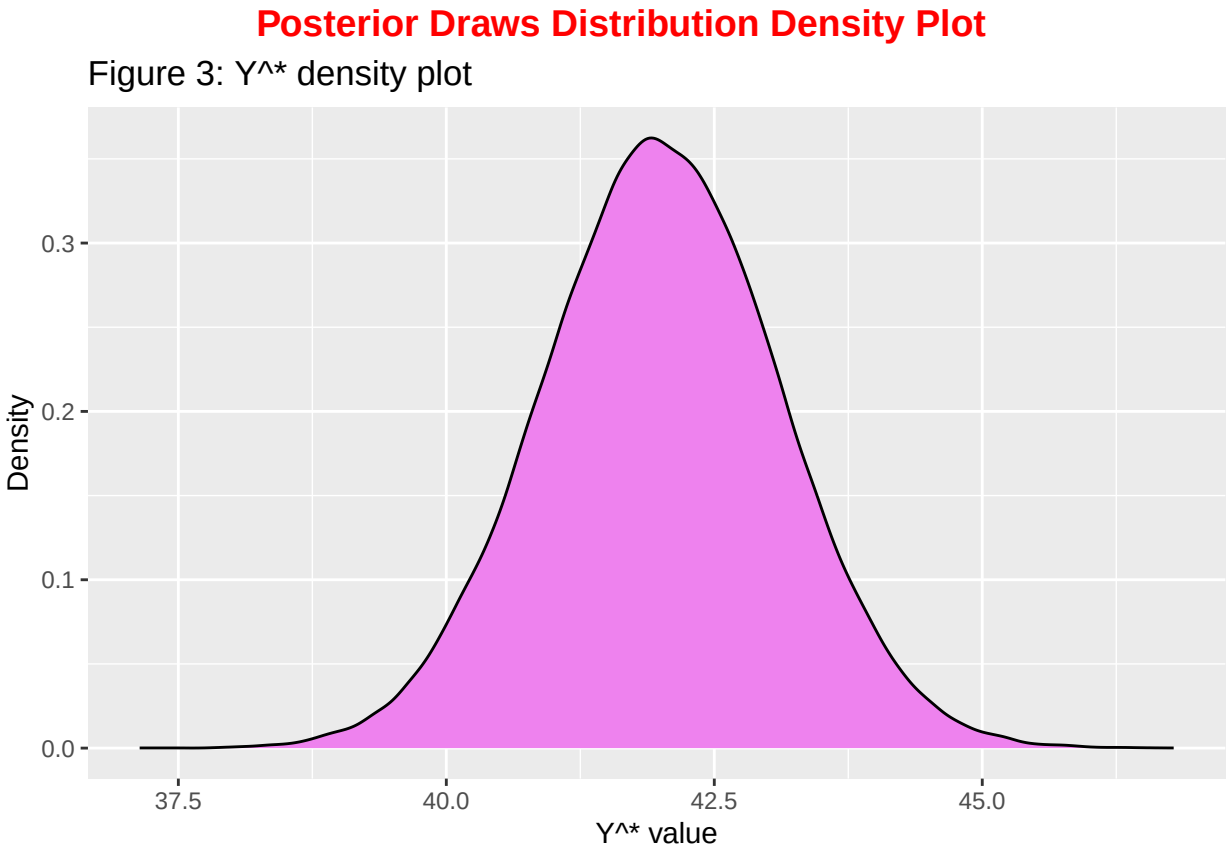


Figure 3: Y^* density plot

To examine the overall model predictive accuracy of our model we will use the predictive summary of our model. Let's examine the posterior predictive summaries for our data:

As we can see from table 6, Among all 414 number of properties in the data set, we see that the observed price per unit area of a house is typically 4.9 or 0.52 standard deviations, from the respective posterior predictive mean. As we can see, our model is doing quite well, as it is able to predict House price of unit area (10000 New Taiwan Dollar/Ping) with in 4.9. Given the magnitude of house price of a unit area, this prediction is doing well. Furthermore, 62% of test observations fall within their respective 50% prediction interval whereas 95% fall within their 95% prediction interval. As we can see, this model is doing quite well both interms of accuracy of the result, and the distributions being within their predictive interval.

Once concern that comes up when evaluating a model on the training data set is over-fitting. To check that our model is not overfitting to the training data, we can perform 10-fold cross validation to confirm if the

results from 10-fold CV align with the models from our overall model on the training set.

Table 6: Table 6: Overall Model Performance

mae	mae_scaled	within_50	within_95
4.869275	0.5218091	0.6183575	0.9541063

Table 7: Table 7: Model Performance Using 10-fold CV

fold	mae	mae_scaled	within_50	within_95
1	5.851879	0.6176122	0.5952381	1.0000000
2	3.302832	0.3608082	0.6341463	0.9268293
3	5.040682	0.5356907	0.5853659	0.9512195
4	4.891664	0.5143526	0.6428571	0.9761905
5	5.551251	0.5889161	0.6341463	0.9756098
6	4.627868	0.4967337	0.6097561	0.9024390
7	5.054135	0.5326287	0.6190476	1.0000000
8	4.515604	0.4749012	0.7073171	0.9756098
9	5.603744	0.6613575	0.5365854	0.9268293
10	4.645538	0.4988136	0.5952381	0.9285714

As we can see from table 7 , our model performs very similarly to the whole model fit to the training data set. Table 7 displays, each of the posterior prediction metrics corresponding to each of the folds which were created randomly. Since the splits are performed equally and at random , the training models performance varies. This stems from the nature of the test set and how close it is to the training data set. Nonetheless, our mae for each of the folds was as low as 3.3 and as high as 5.9. All these values in very small proximity to the overall model performance.

From table 8, we can conclude that the predictive accuracy our model is good as well using 10-fold CV estimates. As we can see from table 8, Among all 414number of properties in the data set, we see that the observed price per unit area of a house is typically 4.9 or 0.53 standard deviations, from the respective posterior predictive mean. As we can see, our model using 10-fold CV is doing quite well, as it is able to predict House price of unit area (10000 New Taiwan Dollar/Ping) with in 4.9. Given the magnitude of house price of a unit area, this prediction is doing well. Furthermore, 62% of test observations fall within their respective 50% prediction interval whereas 96% fall within their 95% prediction interval. As we can see, this model is doing quite well both in-terms of accuracy of the result, and the distributions being within their predictive interval.

Section 5 : Results and Conclusion

Table 8: Table 8: Ultimate 10-fold CV estimates of posterior predictive accuracy Model

mae	mae_scaled	within_50	within_95
4.90852	0.5281814	0.6159698	0.9563298

From table 8, we can conclude that the predictive accuracy our model is good as well using 10-fold CV estimates. As we can see from table 8, Among all 414number of properties in the data set, we see that the observed price per unit area of a house is typically 4.9 or 0.53 standard deviations, from the respective

posterior predictive mean. As we can see, our model using 10-fold CV is doing quite well, as it is able to predict House price of unit area (10000 New Taiwan Dollar/Ping) with in 4.9. Given the magnitude of house price of a unit area, this prediction is doing well. Furthermore, 62% of test observations fall within their respective 50% prediction interval whereas 96% fall within their 95% prediction interval. As we can see, this model is doing quite well both in-terms of accuracy of the result, and the distributions being within their predictive interval.

We can also conclude that the best way to appreciate the value of the house per unit area in Taiwan is to increase the number of convenience stores in the living circle on foot. Additionally, for every 1 meter increase in the distance to the nearest MRT station, house price per unit area typically decreases by 0.00528 thousands of Taiwan dollar per ping, all else equal. For every 1 year increase in house of age, house price per unit area typically decreases by 0.249 thousands of Taiwan dollar per ping, all else equal. for every 1 additional increase in the number of convenience stores in the living circle on foot, house price per unit area typically increases by 1.298 thousands of Taiwan dollar per ping, all else equal.

Overall, our Bayesian linear regression model performed well and will be able to provide good insight into valuation of properties in Taiwan given the predictive features provided.Additionally when the Bayesian linear regression model is compared to a baseline model, it performs quite well . A baseline model is a model which predicts price of new rentals as the average price of rentals from the training set. We will find that the baseline model mae is 10.7108147 where as the ultimate 10-fold Bayesian linear regression model estimates mae is 4.9085198. This is a 118% increase in mae which is achieved by using Bayesian linear regression. Therefore,we can conclude our model did quite well.

Data Sources

-Yeh, I. C., & Hsu, T. K. (2018). Building real estate valuation models with comparative approach through case-based reasoning. Applied Soft Computing, 65, 260-271.

- UCI Machine Learning Repository: Real Estate valuation data set data set. (n.d.). Retrieved May 4, 2023, from https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set