House Price Modeling in Semarang and Surabaya City using Component Regression with Frequentist and Bayesian Approach

Dwi Rantini[1,*], Arip Ramadan[2], Alhassan Sesay[3], Mochammad Fahd Ali Hillaby[1]

[1]Data Science Technology Study Program, Faculty of Advanced Technology and Multidiscipline, Universitas Airlangga, Indonesia
[2]Information System Study Program, Departement of Industrial and System Engineering, Telkom University Surabaya Campus, Indonesia
[3]Faculty of Transformative Education, the United Methodist University, Sierra Leone

dwi.rantini@ftmm.unair.ac.id

Abstract. House prices are getting more and more expensive. Entrepreneurs compete to collect assets. Therefore, this research aims to determine the right house price based on several considerations for that price. Considerations can come from the location of the house or the environment. To find out the relationship, we can use regression analysis. In a regression analysis, predictor variables often found that are interconnected. This results in multicollinearity between these variables. But often these predictor variables should theoretically be significant, and not worth removing. To overcome this, a regression analysis was compared by including all predictor variables, by including predictor variables that were highly correlated with response variables, by including the best subset selection variable, and then the principal component regression. Estimates for the models already mentioned are approached through the frequentist and Bayesian estimation methods. Based on the analysis, it was found that Bayesian analysis is better when compared to frequentist estimates. Then the principal component of the regression turns out to be better used if there is multicollinearity between variables. Thus, it means that we assume all the predictor variables are significant in the model.

Keywords: Bayesian, Frequentist, House Price, Entrepreneurship, Principal Component Regression.

1 Introduction

As everyone knows, the land area on this earth has neither increased nor decreased. According to economics, home is one of the primary needs [1]. Because the land area on earth is fixed, then the price of a house that is certainly built on the land is increasingly expensive [2]. Moreover, land on this earth is also

T. Amrillah et al. (eds.), *Proceedings of the International Conference on Advanced Technology and Multidiscipline (ICATAM 2024)*, Advances in Engineering Research 245,
https://doi.org/10.2991/978-94-6463-566-9_22

to build infrastructure [3]. This situation makes everyone have to think about how to get a house at a price in accordance with economic conditions [4]. Many home offers with various prices. The price of each house sold depends on several factors such as land area, building area, and other facilities. Not only that, but this price is also influenced by the distance of the house to public services such as markets, hospitals, airports, and others. Thus, many factors affect the selling price of a house. And of course, this will have a big impact on immigrants [5]. Who previously did not come from a metropolitan city.

Research related to home prices has increased in recent years [6]. Ordinary least square (OLS) regression is often used to model the selling price of a house against the factors that influence it [7]. Previous research using quantile regression showed that buyers who want high housing prices do indeed want to get luxury amenities [7]. This proves that the facility will indeed be used as a reference for setting house prices. However, in Indonesia, the selling price of the house certainly varies. With the same facilities can make a different price for each different region [8].

Research on housing prices has been done a lot, but most of these studies are more inclined to forecasting. Among them is forecasting conducted data on selling prices of homes in Atlanta, Chicago, Dallas, and San Francisco [9]. Estimates are made for forecasting the selling price of a house using Bayesian analysis. In addition, research on home selling prices using Bayesian analysis and principal component regression has also been conducted [10], but this research focuses on forecasting. This research was conducted to obtain information about what factors are most significant in influencing the selling price of a house. With so many offers for home sales everywhere, the buyer must have a reason why they should buy the house when compared to other houses. Previous information about the selling price of a house can be used as a Bayesian analysis material as a prior, thus comparing the estimation results with a frequentist and Bayesian approach regarding the factors that must be considered when a buyer wants to buy a house.

2 Materials and Methods

In this section, we will explain data sources, determine model variables, and check the collinearity between these variables. Then , some literature reviews will be provided about the methods that will be used in the modeling analysis of this research.

2.1 House Price Secondary Data

House price data is obtained from buying and selling websites in Indonesia. The data taken consists of house prices, land area, building area, number of bathrooms, number of bedrooms, number of floors, and distance of the house to public facilities such as hospitals, malls, and airports. Then the data is taken for the Semarang and Surabaya city. Data on house prices were taken from 50 houses randomly each for the two cities.

2.2 Determining Variables

There are several variables used for home price analysis in the city of Semarang. The response variable is the price of the house, while the predictor variable consists of land area, building area, number of bathrooms, number of bedrooms, number of floors, and distance to public facilities, namely hospital dr. Kariadi, Achmad Yani Airport, and Paragon City Mall. Why are these facilities used as a reference for the distance calculated in the analysis, this hospital and these malls are the largest in Semarang and this airport is the only one in Semarang. As a relevant comparison material, then in Surabaya also taken the same variable as Semarang. As an appropriate comparison, the distance is taken to dr. Soetomo's Hospital, Juanda Airport, and Tunjungan Plaza Mall for variables in Surabaya city.

The data structure of house prices can be seen in Table 1. Then the explanation for each variable is as follows: Y is the house price (in million Rupiah), X_1 is the land area (in m^2), X_2 is the building area (in m^2), X_3 is the number of bathrooms, X_4 is the number of bedrooms, X_5 is the number of floors, X_6 is the distance (in KM) to the hospital dr. Kariadi (for Semarang city) and distance to the hospital dr. Soetomo (for Surabaya city), X_7 is the distance to the Achmad Yani airport (for

Semarang City) and the distance to the Juanda airport (for Surabaya City), X_8 is the distance to the Paragon City Mall (for Semarang City) and distance to the Tunjungan Plaza Mall (for Surabaya City).

Table 1. Data Structure of House Price

No.	Y	X_1	X_2	...	X_7	X_8
1	450	81	52	...	20.4	10.8
2	650	150	70	⋯	8.2	1.2
3	245	60	36	...	24.7	15.2
4	300	72	72	...	23.2	13.6
5	530	75	45	...	20	10.4
6	396	100	42	...	21.4	11.8
⋮	⋮	⋮	⋮	...	⋮	⋮
47	1100	260	200	...	56.7	50.4
48	700	220	200	...	1155	1152
49	1300	296	200	...	55.5	49.2
50	475	72	36	...	56.1	49.8

2.3 Principal Component Regression (PCR)

In statistics, principal component regression (PCR) is a regression analysis technique that is based on principal component analysis (PCA) [11]. Typically, it considers regressing the outcome (also known as the response or the dependent variable) on a set of covariates (also known as predictors, explanatory variables, or independent variables) based on a standard linear regression model, but uses PCA for estimating the unknown regression coefficients in the model [12]. In PCR, instead of regressing the dependent variable on the explanatory variables directly, the principal components of the explanatory variables are used as regressors [13]. One typically uses only a subset of all the principal components for regression, making PCR a kind of regularized procedure. Often the principal components with higher variances (the ones based on eigenvectors corresponding to the higher eigenvalues of the sample variance-covariance matrix of the explanatory variables) are selected as regressors. However, to predict the outcome, the principal components with low variances may also be important, in some cases even more important [14].

One major use of PCR lies in overcoming the multicollinearity problem which arises when two or more of the explanatory variables are close to being collinear [15]. PCR can aptly deal with such situations by excluding some of the low-variance principal components in the regression step. In addition, by usually regressing on only a subset of all the principal components, PCR can result in dimension reduction through substantially lowering the effective number of parameters characterizing the underlying model. This can be particularly useful in settings with high-dimensional covariates. Also, through the appropriate selection of the principal components to be used for regression, PCR can lead to efficient prediction of the outcome based on the assumed model.

For a more complete explanation, mathematical formulas can be explained through Marx and Smith's research [16]. Let $\mathbf{Y}_{n\times 1} = \left(y_1, \cdots, y_n\right)^T$ denote the vector of the response variable and $\mathbf{X}_{n\times p} = \left(\mathbf{x}_1, \cdots, \mathbf{x}_n\right)^T$ denote the corresponding data matrix of predictor variables, where n and p denote the size of the number of observations and the number of predictor variables, respectively, with $n \geq p$. The primary goal is to obtain an efficient estimator $\hat{\boldsymbol{\beta}}$ for the parameter $\boldsymbol{\beta}$, based on the data. One frequently used approach for this is ordinary least squares regression which, assuming $\mathbf{X}$ is full column rank, gives the unbiased estimator. PCR starts by performing a PCA on the centered data matrix $\mathbf{X}$. For this, let $\mathbf{X} = U\Delta V^T$ denote the singular value decomposition of $\mathbf{X}$, where $\Delta_{p\times p} = \operatorname{diag}\left(\delta_1, \cdots, \delta_p\right)$ with $\delta_1 \geq \cdots \geq \delta_p$ denoting the non-negative singular values of $\mathbf{X}$, while the columns of $U_{p\times p} = \left(\mathbf{u}_1, \cdots, \mathbf{u}_p\right)$ and $\mathrm{V}_{p\times p} = \left(\mathbf{v}_1, \cdots, \mathbf{v}_p\right)$ are both orthonormal sets of vectors denoting the left and right singular vectors of $\mathbf{X}$, respectively. Let $\mathbf{W}_k = \mathbf{X}V_k = \left[\mathbf{X}\mathbf{v}_1, \cdots, \mathbf{X}\mathbf{v}_k\right]$ denote the $n \times k$ matrix having the first k principal components as its column.

$\mathbf{W}$ may be viewed as the data matrix obtained by using the transformed covariates $\mathbf{x}_i^k = V_k^T \mathbf{x}_i \in \mathbb{R}^k$ instead of using the original covariates $\mathbf{x}_i \in \mathbb{R}^p \ \forall\ 1 \le i \le n$. Let $\hat{\gamma}_k = \left(W_k^T W_k\right)^{-1} W_k^T \mathbf{Y} \in \mathbb{R}^k$ denote the vector of estimated regression coefficients obtained by ordinary least squares regression of the response vector $\mathbf{Y}$ on the data matrix W_k. Then, for any $k \in \{1, \cdots, p\}$, the final PCR estimator of $\boldsymbol{\beta}$ based on using the first k principal components is given by $\hat{\boldsymbol{\beta}}_k = V_k \hat{\gamma}_k \in \mathbb{R}^p$.

2.4 Bayesian Modeling

Bayesian inference refers to statistical inference where uncertainty in inferences is quantified using probability. In classical frequentist inference, model parameters and hypotheses are considered to be fixed. Probabilities are not assigned to parameters or hypotheses in frequentist inference. For example, it would not make sense in frequentist inference to directly assign a probability to an event that can only happen once, such as the result of the next flip of a fair coin. However, it would make sense to state that the proportion of heads approaches one-half as the number of coin flips increases [17].

Bayesian Analysis in general, many have reviewed, how the steps or principles in the Bayesian analysis. To be able to learn more about Bayesian analysis, it can be explained in a book [18] which is already complete about Bayesian analysis. In general, Bayesian analysis formulas can be written in Equation (1)

$$p\left(\boldsymbol{\theta} \mid \mathbf{x}\right) \propto p\left(\boldsymbol{\theta}\right) l\left(\mathbf{x} \mid \boldsymbol{\theta}\right). \tag{1}$$

where $p\left(\boldsymbol{\theta} \mid \mathbf{x}\right)$ is the posterior distribution, $p\left(\boldsymbol{\theta}\right)$ is the prior distribution and $l\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$ is the likelihood of data research.

This research will emphasize the use of Pseudo priors which will be used as information to obtain posterior distribution. Pseudo-prior is a term that was popularized by [19] which is defined as the determination of prior distributions that

use the prior distribution parameter values from the results of modeling or processing by the frequentist approach.

2.5 Methodology

The following are the steps taken to do modeling with regression analysis

1. Detect the presence of multicollinearity by showing the correlation value between variables and also the VIF value of each variable.
2. If there is multicollinearity, then it is likely to be done with principal component regression.
3. Before going to the principal component regression stage, a regression will be carried out by entering all the predictor variables.
4. Regressing predictor variables that have a high correlation value to the response variable.
5. Regressing the best variable for the best subset selection
6. With principal component analysis, principal component regression will be performed.
7. Comparing R^2 values, the largest R^2 values that will be considered later, will then be compared with the estimation results using principal component regression.

For the series of steps above, we will get the estimated regression coefficient values for each model. This regression coefficient value will later be used as a Pseudo before being estimated by Bayesian analysis.

2.6 Dealing with Multicollinearity

To detect the relationship between predictor variables, correlation can be used. The higher the correlation value, the more predictor variables are interconnected, or the presence of multicollinearity. The correlation between predictor variables can be seen in Table 2 for Semarang City and Table 3 for Surabaya city. From Table 2, it can be seen that there are four high correlation values, while in Table 3 there are two high correlation values (in the red box).

Table 2. Variable Predictor Correlation Matrix for Home Prices in the Semarang City

	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	0.345							
X_2	0.774	0.54						
X_3	0.816	0.38	0.85					
X_4	0.748	0.52	0.85	0.84				
X_5	0.379	0.05	0.41	0.42	0.418			
X_6	-0.001	0.06	0.24	0.11	-0.06	0.29		
X_7	-0.002	0.06	0.24	0.11	-0.05	0.29	1	
X_8	0.000	0.06	0.25	0.11	-0.05	0.29	1	1

Table 3. Variable Predictor Correlation Matrix for Home Prices in the Surabaya city

	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	0.924							
X_2	0.929	0.887						
X_3	0.28	0.337	0.344					
X_4	0.324	0.481	0.404	0.702				
X_5	0.446	0.329	0.555	0.423	0.22			
X_6	-0.113	-0.09	-0.153	0.355	0.359	-0.09		
X_7	-0.13	-0.211	-0.224	-0.186	-0.452	0.002	-0.281	
X_8	-0.184	-0.172	-0.21	0.263	0.29	-0.071	0.912	-0.244

Among the many indices, the variance inflation factor (VIF) is widely used to determine collinearity. Although any fixed numbers for VIFs as extremes have not yet been introduced, but it is obvious that the higher the indices are, the more serious the collinearity problem exists. If any of the VIFs exceeds 5 or 10, it is an indication that the associated regression coefficients will poorly be estimated [20].

Table 4. VIF values for Predictor Variables for Semarang and Surabaya city

	VIF	
Variables	Semarang	Surabaya
X_1	1.66	6.69
X_2	8.28	8.11
X_3	4.99	2.68
X_4	7.76	3.04
X_5	1.70	2.21
X_6	19316.16	7.01
X_7	3817.02	1.45
X_8	21781.24	6.55

Based on

Table **4**, there is a very extreme VIF value, namely the relationship between X_6 , X_7 and X_8 in Semarang, besides that other VIF values are also quite high.

3 Main Results

Before going to the analysis step, we will first look at the pattern of house price data in Semarang and Surabaya. To help see the pattern of data, a histogram is displayed for each house price in Semarang and Surabaya cities in Figure 1 and Figure 2. Data on house prices in Semarang and Surabaya in Figure 1 and Figure 2 are accompanied by Normal distribution fit according to the Kolmogorov-Smirnov goodness of fit test.

After a histogram of the house price is displayed, descriptive statistics for the predictor variables are displayed in Table 5 and Table 6. Based on Table 5 and Table 6 it can be seen that there are significant differences in variance on the predictor variables. In addition to this, also for reasons of unit differences in each of the predictor variables, then for the principal component analysis of regression the correlation unit will be used, not covariance.

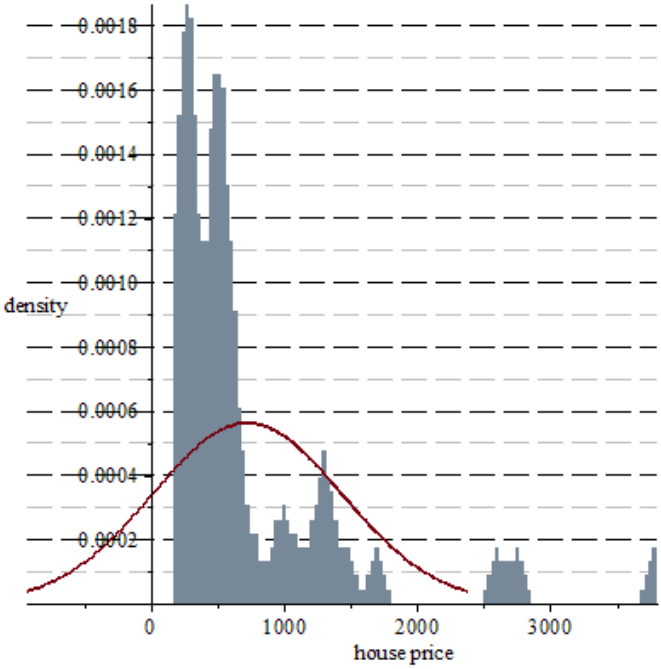

Figure 1. House Price Histogram in Semarang city

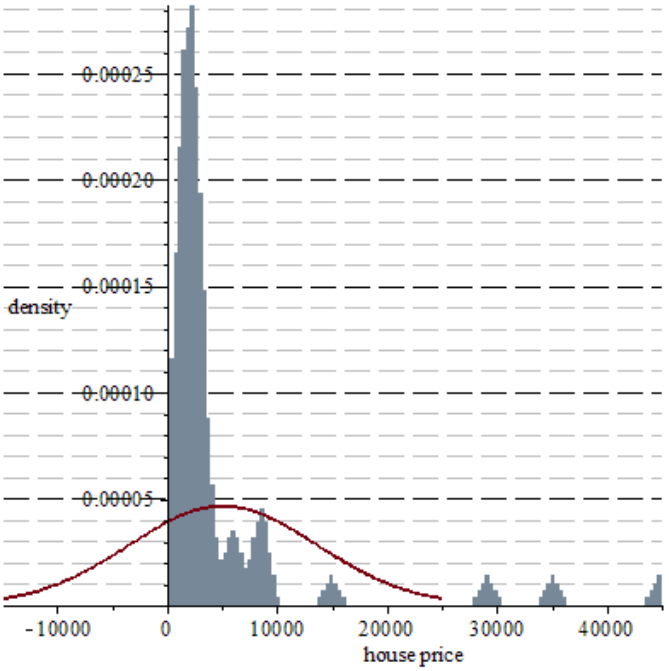

Figure 2. House Price Histogram in Surabaya city

Table 5. Descriptive Statistics for Predictor Variables in Semarang city

Variable	Mean	StDev	Var	Q1	Median	Q3
X_1	168.3	282.5	79826.7	72.0	87.0	150.0
X_2	89.3	77.9	6075.6	39.0	49.0	122.5
X_3	1.500	0.839	0.704	1.000	1.000	2.000
X_4	2.780	1.234	1.522	2.000	2.000	3.000
X_5	1.2000	0.4041	0.1633	1.0000	1.0000	1.0000
X_6	39.2	161.6	26120.1	7.5	11.6	15.1
X_7	46.9	160.8	25842.8	14.9	20.3	24.8
X_8	40.0	161.3	26009.2	8.3	11.8	15.4

Table 6. Descriptive Statistics for Predictor Variables in Surabaya city

Variable	Mean	StDev	Var	Q1	Median	Q3
X_1	225.2	214.5	46017.6	103.0	160.5	240.0
X_2	267.6	279.2	77930.7	120.0	180.0	287.5
X_3	3.100	1.374	1.888	2.000	3.000	4.000

Variable	Mean	StDev	Var	Q1	Median	Q3
X_4	4.140	1.552	2.409	3.000	4.000	5.000
X_5	1.7000	0.5440	0.2959	1.0000	2.0000	2.0000
X_6	6.008	2.871	8.242	3.500	5.400	7.500
X_7	16.644	3.072	9.440	14.100	17.400	19.400
X_8	7.816	2.850	8.120	5.175	7.200	10.000

Based on the steps outlined in the section 2.5,

Table 7 and Table 8 are the results of regression analysis if all predictor variables are included in the model. With a significance level of a = 5%, only X_3 is significant for the model in Semarang city and X_1, X_2 and X_4 for Surabaya city. Then the results of this regression estimate will be made Pseudo priors for further estimation using Bayesian analysis, which is on the right side. For Bayesian analysis, a variable is said to be significant if the estimated results of its parameters at the interval of credible intervals do not contain zero.

Table 7. Estimation Results by Entering All Predictor Variables into the Model in the Semarang city

	Frequentist			Bayesian		
Estimator	Coef	SE Coef	P-Value	Coef	2.5%	9.75%
$\hat{\beta}_0$	-334	278	0.236	-321.6	-592	-48.8
$\hat{\beta}_1$	-0.109	0.269	0.688	-0.1	-0.4	0.2
$\hat{\beta}_2$	3.80	2.18	0.089	3.7	1.6	5.9
$\hat{\beta}_3$.	4	157	0.007	450.7	289.8	609.4
$\hat{\beta}_4$	-52	133	0.697	-48.5	-177.6	82.1
$\hat{\beta}_5$	165	191	0.390	163.3	-28.9	360.1
$\hat{\beta}_6$	-34.5	50.8	0.501	-35.1	-72.9	-3.7
$\hat{\beta}_7$	1.8	22.7	0.939	-0.5	-22.7	20.1
$\hat{\beta}_8$	32.0	54.0	0.557	34.9	-1.2	73.7
R^2			71.77%			71.77%

Table 8. Estimation Results by Entering All Predictor Variables into the Model in the Surabaya city

	Frequentist			Bayesian		
Estimator	Coef	SE Coef	P-Value	Coef	2.5%	9.75%
$\hat{\beta}_0$	-4979	3159	0.123	-4922.3	-7968	-1890.9
$\hat{\beta}_1$	22.63	4.17	0.000	22.6	18.6	26.7
$\hat{\beta}_2$	15.49	3.53	0.000	15.5	12.2	18.9
$\hat{\beta}_3$.	-30	412	0.942	-29.1	-445.2	390.4
$\hat{\beta}_4$	-909	389	0.024	-911.1	-1301	-524.8
$\hat{\beta}_5$	413	945	0.664	405.6	-541.4	1371
$\hat{\beta}_6$	317	319	0.326	316.2	11.2	618.9
$\hat{\beta}_7$	148	136	0.281	146.1	13.8	277.7
$\hat{\beta}_8$	-43	311	0.891	-42.2	-340.7	257.9
R^2			93.32%			93.32%

The next step is to regress the predictor variable that has the highest correlation value to the house price variable. Based on Table 2 and Table 3, we get X_2 , X_3 and X_4 for Semarang City and X_1 and X_2 variables for Surabaya. Like as before, the estimation results from the frequentist approach will be used as Pseudo priors for Bayesian analysis. The estimation results can be seen in Table 9 and Table 10. For Surabaya city, the estimation results using the frequentist and Bayesian approaches have the same results, which are all significant parameters. Unlike Surabaya City, there are differences regarding the results of the estimation for Semarang City.

Table 9. Estimation Results by Entering the Highest Correlation Predictor Variables into the Model in the Semarang City

	Frequentist			Bayesian		
Estimator	Coef	SE Coef	P-Value	Coef	2.5%	9.75%
$\hat{\beta}_0$	-308	171	0.079	-307.4	-485.4	-130.7
$\hat{\beta}_2$	2.10	1.64	0.205	2.1	0.5	3.8
$\hat{\beta}_3$.	449	145	0.003	449.4	298.9	599.4
$\hat{\beta}_4$	62.0	99.0	0.534	61.6	-37.7	160.2
R^2			69.13%			69.13%

Table 10. Estimation Results by Entering the Highest Correlation Predictor Variables into the Model in the Surabaya City

	Frequentist			Bayesian		
Estimator	Coef	SE Coef	P-Value	Coef	2.5%	9.75%
$\hat{\beta}_0$	-3281	542	0.000	-3865	-3483	-2692
$\hat{\beta}_1$	18.76	3.79	0.000	16.7	18	20.8
$\hat{\beta}_2$	15.71	2.91	0.000	13.6	15	17.8
R^2			90.97%			90.97%

The next step is to regress the predictor variable that has the best subset selection. Based on Table 4, the variables X_6 , X_7 and X_8 for Semarang city have very extreme VIF values. Therefore, only one is taken to represent the three, namely X_7 because it has the highest correlation based on Table 2. The selection of the best subset is based on the highest R^2 value, which can be seen in

Table **11** and Table 12.

Table 11. The selection of the Best Subset Model is based on the Largest Determination Coefficient for Semarang City

Number of Variables	R^2	X_1	X_2	X_3	X_4	X_5	X_7
2	67.5		V	V			
2	66.7			V	V		
3	68.9		V	V			V
3	67.1		V	V	V		
4	68.6		V	V		V	V
4	68.5	V	V	V			V
5	68.1	V	V	V		V	V
5	68.1		V	V	V	V	V
6	67.5	V	V	V	V	V	V

Table 12. The selection of the Best Subset Model is based on the Largest Determination Coefficient for Surabaya City

Number of Variables	R^2	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	86.0		V						
1	85.1	V							
2	90.6	V	V						
2	87.1	V				V			
3	92.0	V	V		V				
3	91.2	V	V					V	
4	92.4	V	V		V		V		
4	92.3	V	V		V				V
5	92.5	V	V		V		V	V	
5	92.3	V	V		V			V	V
6	92.4	V	V		V	V	V	V	
6	92.3	V	V	V	V		V	V	
7	92.2	V	V		V	V	V	V	V
7	92.2	V	V	V	V	V	V	V	
8	92.0	V	V	V	V	V	V	V	V

Based on

Table **11** and Table 12, the estimation results for the frequentist and Bayesian approaches can be seen in Table 13 and Table 14.

Table 13. Estimation Results by Entering the Best Subset Predictor Variables into the Model in the Semarang City

	Frequentist			Bayesian		
Estimator	Coef	SE Coef	P-Value	Coef	2.5%	9.75%
$\hat{\beta}_0$	-203	124	0.108	-203.4	-333	-71.6
$\hat{\beta}_2$	3.34	1.45	0.026	3.3	1.9	4.8
$\hat{\beta}_3$.	441	132	0.002	441.1	317.5	564
$\hat{\beta}_7$	-0.652	0.370	0.085	-0.7	-1.1	-0.3
R^2			70.84%			70.84%

Table 14. Estimation Results by Entering the Best Subset Predictor Variables into the Model in

the Surabaya City

	Frequentist			Bayesian		
Estimator	Coef	SE Coef	P-Value	Coef	2.5%	9.75%
$\hat{\beta}_0$	-4789	2818	0.096	-4828	-7445.1	-2167
$\hat{\beta}_1$	21.85	3.55	0.000	21.9	18.3	25.4
$\hat{\beta}_2$	16.45	2.66	0.000	16.5	13.8	19.1
$\hat{\beta}_4$	-899	292	0.004	-897.6	-1200	-597.5
$\hat{\beta}_6$	277	135	0.047	278	133	421.6
$\hat{\beta}_7$	160	125	0.205	161.8	38.8	282.7
R^2			93.28%			93.28%

It has been explained before that predictor variables have a high multicollinearity, therefore the model to be analyzed last is to use principal component regression. The number of principal components taken can be seen in the scree plot in Figure 3 and Figure 4 and also the eigenvalues of more than 1 in Table 15 and Table 16.

Table 15. Eigenvalues and Eigenvectors of Correlation Matrix for Semarang city

Variables	Component				
	1	2	3	4	5
	Eigenvalues				
	3.6248	2.7343	0.9251	0.4903	0.1309
	Eigenvectors				
X_1	0.242	-0.252	-0.681	-0.606	-0.218
X_2	0.431	-0.289	-0.069	0.219	0.356
X_3	0.377	-0.338	0.124	0.419	-0.735
X_4	0.334	-0.433	0.043	0.094	0.528
X_5	0.306	-0.049	0.706	-0.627	-0.072
X_6	0.369	0.428	-0.073	0.048	0.014
X_7	0.369	0.428	-0.076	0.050	0.024
X_8	0.369	0.427	-0.075	0.049	0.020

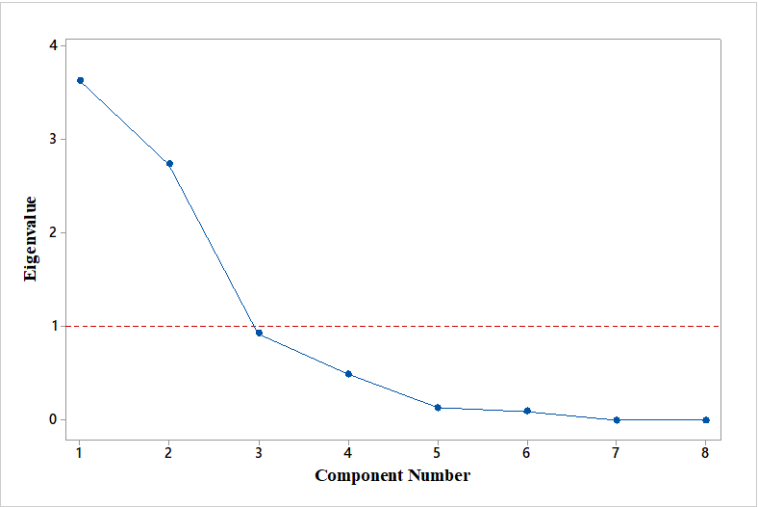

Figure 3. Scree Plot Principal Component for the Semarang city

Table 16. Eigenvalues and Eigenvectors of Correlation Matrix for Surabaya city

Variables	Component				
	1	2	3	4	5
	Eigenvalues				
	3.0780	2.3434	0.9875	0.6336	0.6037
	Eigenvectors				
X_1	0.416	0.311	-0.220	-0.287	0.429
X_2	0.423	0.365	-0.085	-0.372	0.087
X_3	0.438	-0.107	0.363	0.521	0.071
X_4	0.474	-0.129	-0.112	0.463	0.165
X_5	0.304	0.241	0.559	-0.209	-0.613
X_6	0.199	-0.568	0.094	-0.311	0.129
X_7	-0.275	0.172	0.680	-0.051	0.618
X_8	0.416	0.311	-0.220	-0.287	0.429

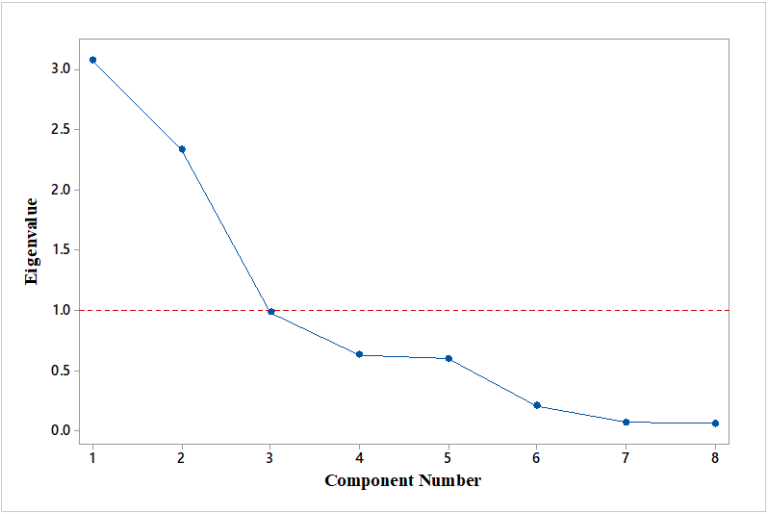

Figure 4. Scree Plot Principal Component for the Surabaya city

Based on eigenvalues and scree plots, the regression was performed for PC_1 and PC_2 for both cities. Estimation results can be seen in Table 17 and

Table 18.

Table 17. Estimation Results by Principal Component Regression for Semarang city

	Frequentist			Bayesian		
Estimator	Coef	SE Coef	P-Value	Coef	2.5%	9.75%
constant	726.5	61.2	0.000	726.3	656.8	795.3
PC_1	213.9	32.5	0.000	213.9	177.5	250.7
PC_2.	-242.3	37.4	0.000	-242.5	-284.9	-199.9
R^2			64.48%			64.48%

Table 18. Estimation Results by Principal Component Regression for Surabaya city

	Frequentist			Bayesian		
Estimator	Coef	SE Coef	P-Value	Coef	2.5%	9.75%
constant	5148	645	0.000	5146.4	4414	5873

	Frequentist			Bayesian		
Estimator	Coef	SE Coef	P-Value	Coef	2.5%	9.75%
PC_1	3268	371	0.000	3268.3	2852	3689
PC_2.	2961	426	0.000	2958.9	2477	3444
R^2			72.82%			72.82%

Based on Table 17 and

Table 18, it can be seen that using both frequentist and Bayesian estimates gives the same results, i.e. all significant components.

Based on

Table 7, Table 8, Table 9, Table 10, Table 13, Table 14, Table 17 and

Table 18, it can be concluded that the highest coefficient of determination is obtained if all variables are included in the model. However, not all predictor variables significantly influence the selling price of houses in Semarang or Surabaya. To be able to see how the results of the estimation of the original price of each house sold in Semarang and Surabaya can be seen in Figure 5 and Figure 6.

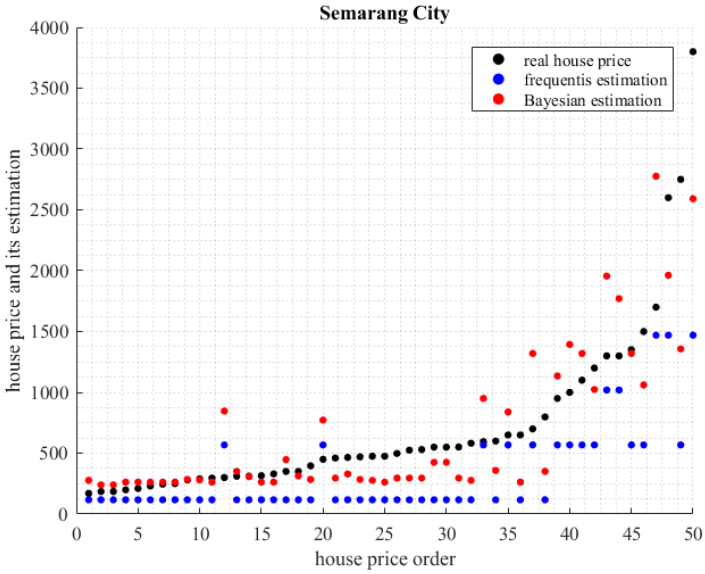

Figure 5. Scatter Plot of Estimated Results by Incorporating All Variable Predictors into the Model through the Frequentist and Bayesian Approaches for Semarang City

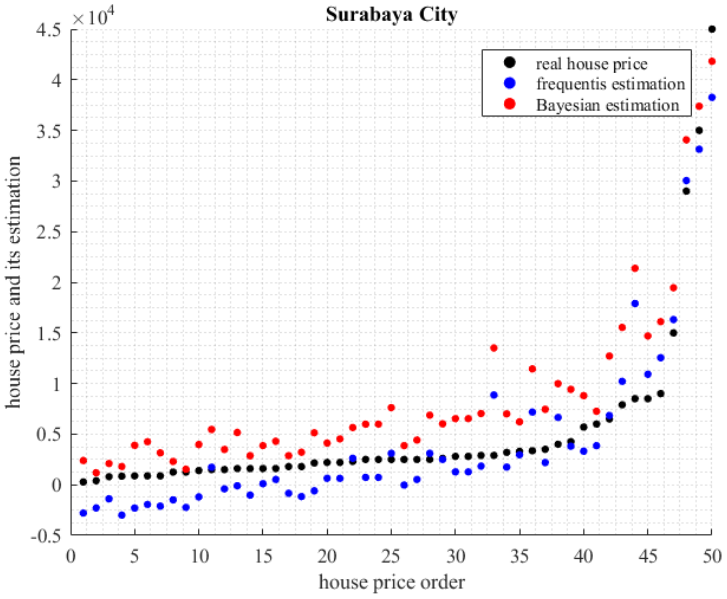

Figure 6. Scatter Plot of Estimated Results by Incorporating All Variable Predictors into the Model through the Frequentist and Bayesian Approaches for Surabaya City

Then, the scatter plot above will be compared with the scatter plot from the estimation results using PCR. The estimation results for the frequentist and Bayesian approaches are separated into two different scatter plots because of the very large scale of house price data, so the dots coincide. Scatter plots for the Semarang and Surabaya cities can be seen in Figure 7 and Figure 8, respectively.

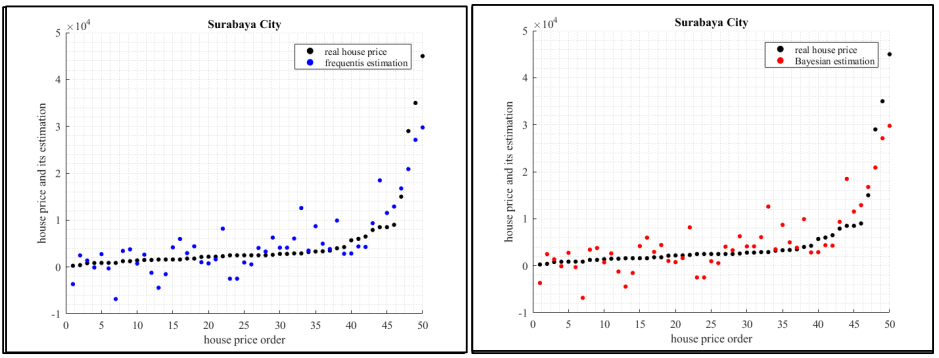

Figure 8. Scatter Plot of Estimated Results by Principal Component Regression through the Frequentist and Bayesian Approaches for Surabaya City

4 Conclusions

Based on the results and analysis, there are several things that can be learned. The first is different estimation results when using the frequentist and Bayesian approaches. However, estimations using the frequentist and Bayesian approaches have the same results for the principal component regression. This statement is supported by the estimation results in Figure 5, in Figure 5 shows that Bayesian estimates are closer to the original house price in Semarang. The second is to look in more detail, in Figure 6, Figure 7 and Figure 8, the estimation results are at a minus value. This should be avoided because the selling price of a house cannot be negative. The cause of the results of this negative estimate is because the regression is done using the Normal distribution approach whose domains are from infinite negative to positive infinite. The third is estimation using principal component regression is found to be

better than entering all variables, even though entering the coefficient of determination is higher. This statement is supported by the fact that in the regression model that includes all the variables it turns out that not all are significant, whereas in PCR, all variables are considered influential.

Based on the above statement, in general it can be concluded that Bayesian estimates are better than frequentist estimates. Then for data in the presence of multicollinearity, it is better to use PCR. For future research that uses regression analysis, first look at the data from the response variable. When the response variable is always positive, it is better to use a distribution whose domain is always positive as well, such as Log Normal, Weibull, and others.

References

[1] G. Mattioli, C. Roberts, J. K. Steinberger, and A. Brown, "The political economy of car dependence: A systems of provision approach," *Energy Res Soc Sci*, vol. 66, p. 101486, 2020.

[2] M. Lipton, "Why poor people stay poor," in *Rural development*, Routledge, 2023, pp. 66–81.

[3] D. Hongbo and C. Mulley, "Relationship between transport accessibility and land value: Local model approach with geographically weighted regression," *Transp Res Rec*, vol. 1977, no. 1977, pp. 197–205, 2006, doi: 10.3141/1977-25.

[4] J. Vonlanthen, "Interest rates and real estate prices: a panel study," *Swiss J Econ Stat*, vol. 159, no. 1, p. 6, 2023.

[5] F. Helfer, V. Grossmann, and A. Osikominu, "How does immigration affect housing costs in Switzerland?," *Swiss J Econ Stat*, vol. 159, no. 1, p. 5, 2023.

[6] M. Cho, "House price dynamics: A survey of theoretical and empirical issues," *Journal of Housing Research*, vol. 7, pp. 145–172, 1996.

[7] H. Kim, S. W. Park, S. Lee, and X. Xue, "Determinants of house prices in Seoul: A quantile regression approach," *Pacific Rim Property Research Journal*, vol. 21, no. 2, pp. 91–113, 2015, doi: 10.1080/14445921.2015.1058031.

[8] E. L. Glaeser and J. E. Gyourko, "The Impact of Zoning on Housing Affordability," *SSRN Electronic Journal*, vol. 9, no. 2, 2005, doi: 10.2139/ssrn.302388.

[9] C. L. Kuo, "A Bayesian Approach to the Construction and Comparison of Alternative House Price Indices," *Journal of Real Estate Finance and Economics*, vol. 14, no. 1–2, pp. 113–132, 1997, doi: 10.1023/A:1007724103085.

[10] R. Gupta and A. Kabundi, "Forecasting Real US House Prices: Principal Components Versus Bayesian Regressions," *International Business & Economics Research Journal (IBER)*, vol. 9, no. 7, 2010.

[11] M. S. I. Khan, N. Islam, J. Uddin, S. Islam, and M. K. Nasir, "Water quality prediction and classification based on principal component regression and gradient boosting classifier approach," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 8, pp. 4773–4781, 2022.

[12] Z. Li, L. Tian, Q. Jiang, and X. Yan, "Dynamic nonlinear process monitoring based on dynamic correlation variable selection and kernel principal component regression," *J Franklin Inst*, vol. 359, no. 9, pp. 4513–4539, 2022.

[13] E. Costantini, K. M. Lang, K. Sijtsma, and T. Reeskens, "Solving the many-variables problem in MICE with principal component regression," *Behav Res Methods*, vol. 56, no. 3, pp. 1715–1737, 2024.

[14] I. T. Jolliffe, "A Note on the Use of Principal Components in Regression," *Appl Stat*, vol. 31, no. 3, p. 300, 1982, doi: 10.2307/2348005.

[15] M. R. Espejo, *The Oxford Dictionary of Statistical Terms*, vol. 167, no. 2. Oxford University Press on Demand, 2004. doi: 10.1111/j.1467-985x.2004.t01-3-.x.

[16] B. D. Marx and E. P. Smith, "Principal Component Estimation for Generalized Linear Regression," *Biometrika*, vol. 77, no. 1, p. 23, 1990, doi: 10.2307/2336046.

[17] S. Basu, *Bayesian and Frequentist Regression Methods*, vol. 84, no. 1. Springer Science & Business Media, 2016. doi: 10.1111/insr.12171.

[18] M. Andrews and T. Baguley, *Bayesian data analysis*. Chapman and Hall/CRC, 2017. doi: 10.1017/9781316216491.030.

[19] B. P. Carlin and S. Chib, "Bayesian Model Choice Via Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 3, pp. 473–484, 1995, doi: 10.1111/j.2517-6161.1995.tb02042.x.

[20] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*, vol. 821. John Wiley & Sons, 2012.