House Rent Analysis with Linear Regression Model—— A Case Study of Six Cities in India

Zonglin Cai[1, *, †], Yiqing Zhao[2, †]

[1] United World College of South East Asia, Singapore

[2] University of California, Davis, California, USA

* Corresponding Author Email: cai122427@gapps.uwcsea.edu.sg

[†]These authors contributed equally.

Abstract. House rent in India has been rising significantly ever since the pandemic hit the country. In situations like this, it is important for consumers to have the concept of a reasonable rent price, as otherwise, they may suffer from landlords raising rents deliberately. To resolve this issue, a prediction model for rent prices is necessary. This study analyses rent data from six cities (Kolkata, Mumbai, Bangalore, Delhi, Chennai and Hyderabad) in India with multiple variables, including size, furnishing status, and the number of bathrooms, bedrooms, halls, and kitchens and creates a prediction model based on the data. The main analytical methods used are linear regression and logarithmic transformation. This study also includes a general factor analysis based on the data. The results suggest that this model is reasonably accurate for reference uses, but needs further improvements if it is to be used commercially.

Keywords: House rent; India; Linear regression model; Factor analysis.

1. Introduction

1.1. Background

Information asymmetry always exists in rental markets, and it has been a problem, especially for tenants as they can get abused without even knowing. This has been a problem in India particularly. In fact, the Rent Control Act established by the Indian government was set out to solve this problem primarily. However, legislation cannot solve the problem entirely because individuals might still be subjected to smaller information asymmetries. Moreover, some landlords might be struggling to rent out their house, as they lack the information of market price and set their rent too high. For these reasons, it is important to have such a model predicting rent prices.

1.2. Related research

So far, there have been plenty of studies on rent prediction using various methods. In 2003, Dokmeci et al. analyzed rent data in Istanbul, Turkey using hedonic regression [1]. Their study suggests that external factors like size, number of rooms, bathrooms, balconies, and type of housing unit do have an impact on the rent price. In more recent years, more sophisticated algorithms are used. Ming et al. analyzed the rent price in Chengdu, China with XGBoost [2]. Similarly, Li modeled the rent price in China using the Light Gradient Boosting Model, which is a more efficient variation of the XGBoost model [3]. Ma et al. expanded on their study by additionally using Random Forest Regressor and Extra Tree Regressor [4].

Traditional linear regression models are also frequently used in recent studies. In 2019, Kumar investigated a case study in Ames, Iowa, the United States with linear regression and ridge regression [5]. In 2020, Tomal modeled the rent prices in Cracow, Poland using OLS regression [6]. He especially used SAR and GWR-SAR models to determine the spatial heterogeneity of the parameters and autocorrelation of housing rent. The results showed that the price of rent had a close relationship with houses' structure, location and neighborhood.

Moreover, Andrew Coleman and Grant M. Scobie created a simple model of housing rental and ownership with policy simulations and took New Zealand as an example in 2009 [7]. Their studies

indicated that the demand for rental property is much more elastic than the housing demand in total because of the substitution theory in economics. Therefore, the factor of owner occupancy rate cannot determine the measurement of the housing market. Another article about the factors that influenced the price of rents in Beijing, China indicated that the two most important factors were house area and location [8]. Within the same city, the price of rent was still very different, mainly affected by whether the house is located in the commercial center of the city. Christina stated that bathroom and bedroom numbers, whether pets were allowed in the house, square footage, parking area, whether there were washers or dryers, location and lawn were the eight factors that affect rent prices [9]. Finally, Bogdan researched the renting landscape in thirty countries around the world in 2018 [10]. He found out that the number of renters was growing in most of the thirty countries and the prices also kept increasing.

1.3. Objective

Despite a number of studies surrounding the subject of rent forecast and general analysis of factors that influence house rent, there have been few studies on India's rental market, which was affected heavily by the COVID-19 pandemic. Therefore, our study tries to provide a model for rent prediction in India with linear regression and logarithmic transformation. In addition, we want to have a general analysis of factors that have a big impact on housing rents in India.

2. Data

Our data is collected from Kaggle. In total, there are 4746 data entries of various house types in India. The data is collected from 6 cities in India, which are Kolkata, Mumbai, Bangalore, Delhi, Chennai and Hyderabad. All data is posted on https://www.magicbricks.com/ and is up to date. Most of them are posted in April, May, June and July of 2022. Table 1 shows the names of the variables and their corresponding meaning. For Furnishing Status, we used dummy variables for convenience during data processing.

Table 1. Names and descriptions of each variable

Variable name	Description
BHK	Number of bedrooms, hall and kitchen
Rent	Rent of the houses/apartments/flats in Indian rupees
Size	Size of the houses/apartments/flats in square feet
Floor	Houses/apartments/flats situated on which floor (Ground floor is 0)
City	City where the houses/apartments/flats are located.
Furnishing Status	Furnishing status of the houses/apartments/flats (2 means furnished, 1 is semi-furnished, and 0 is unfurnished)
Bathroom	Number of bathrooms

3. Methodology

3.1. Linear regression

Multiple linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). It can be expressed by the following equation (1)

$$Y = b \bullet X + e \tag{1}$$

where Y is the dependent variable, X is the independent variable, b is an unknown parameter and e is the error term.

3.2. Linear regression with logarithmic transformation

Logarithmic transformation in linear regression is usually used when the relationship between dependent and independent variables is not linear. It can be expressed by the following equation (2), where Y is the dependent variable, X is the independent variable, a and b are unknown parameters and e is the error term.

$$\log Y = a + b \bullet X + e \tag{2}$$

The benefit of this method is that the linear relationship can still be preserved while a non-linear relationship is actually being handled. This transformation will increase the precision of the model compared to the original linear regression. Moreover, this transformation can turn highly skewed data into a more normal distribution, which makes analysis more convenient. This can be shown in Fig. 1 and Fig. 2, which are the distributions of house rent in Bangalore with and without the transformation, respectively.

More specifically, the transformation we are using is a log-linear model, which is to take the log of the dependent variable while remaining independent variables unchanged, as we found out that it is the best way to deal with dummy variables (compared to linear-log and log-log models).

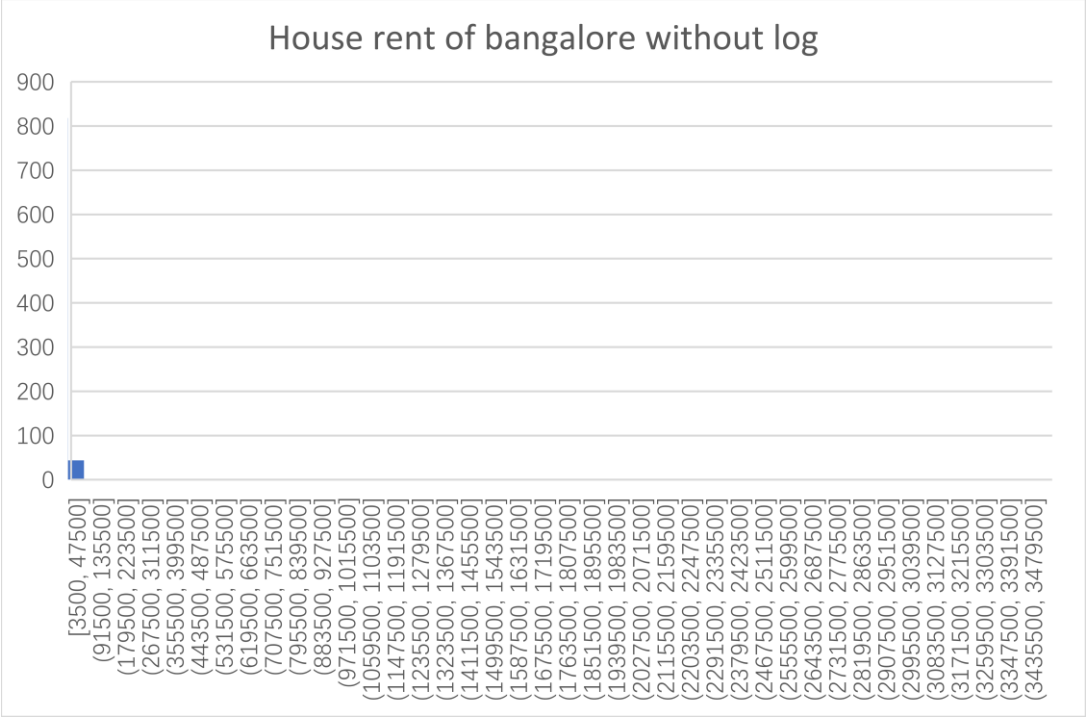

Fig. 1 Distribution of house rent in Bangalore without logarithmic transformation (Photo credit: Original)

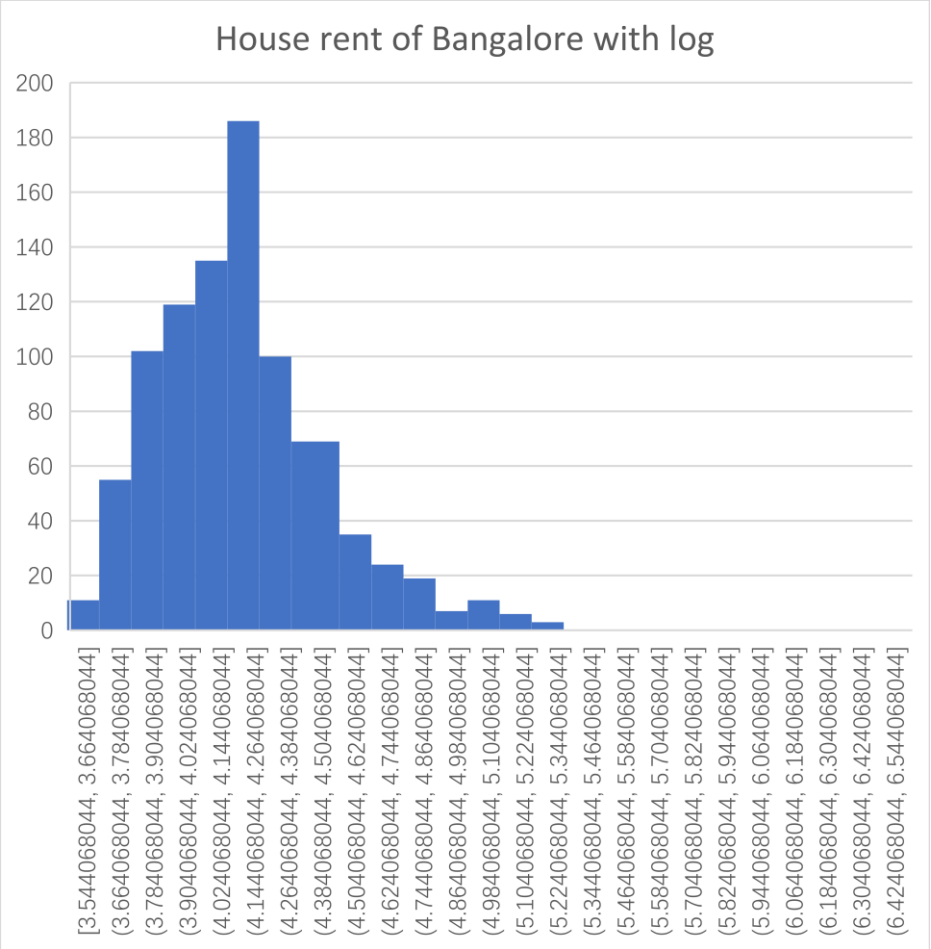

Fig. 2 Distribution of house rent in Bangalore with logarithmic transformation (Photo credit: Original)

4. Results

Our results can be divided into three parts: general analysis on single factors, linear regression results on different cities, and results of regression with logarithmic transformation. By isolating individual factors and performing linear regressions between each of them and the rent price, we want to see what factors are more relevant to rent prices, and then we provide a simplistic model using linear regression only. At last, we use logarithmic transformation to improve the accuracy of our model.

4.1. General analysis of single factors

In this study, we first used regression models to analyze the five factors and compared their impact on housing rent. The following table includes the R values and standard errors of each of the five regressions.

Table 2. R, R^2 and standard error values of 5 variables

	BHK	Size	Floor	Furnishing	Bathroom
Multiple R	0.369717574	0.413550758	0.215498864	0.146251109	0.441215229
R^2	0.136691085	0.17102423	0.04643976	0.021389387	0.194670878
Adjusted R^2	0.136509106	0.170849488	0.046238757	0.021183103	0.194501121
Standard Error	72579.76777	71121.90486	76279.27004	77274.71679	70100.18488
Observation	4746	4746	4746	4746	4746

Among the 5 factors, the biggest value of R^2 is from the bathroom number (0.19), which means that it has the largest impact on housing rent. The overall order of R^2 is bathroom (0.19) > size (0.17) > BHK (0.14) > floor level (0.05) > furnishing status (0.02). The greater R2 is, the larger of an impact each variable has on rent price. Interestingly, our results show that whether a house has been furnished does not influence the house rent by much.

4.2. Effects of different cities

When it comes to another very important factor --- location, we also analyzed R and R^2 values in different cities. In the first part, we compare all the data of the same factor in different cities with their corresponding house prices to get R and R^2, while in the second part, we aggregated the data for all the different factors in the same city with their housing rents and compared the R^2 between different cities. The number of observations in six cities are 972, 524, 886, 868, 605 and 891 respectively, and the R2 values are 0.75, 0.30, 0.07, 0.58, 0.50 and 0.36, as shown in Table 3.

Table 3. R, R^2, standard error values and number of observations of the six cities

Regression	Mumbai	Kolkata	Bangalore	Hyderabad	Delhi	Chennai
Multiple R	0.866266182	0.54864719	0.267578933	0.761630229	0.703898743	0.599804339
R^2	0.750417098	0.301013739	0.071598485	0.580080605	0.49547344	0.359765246
Adjusted R^2	0.749125261	0.294266767	0.066323477	0.577644878	0.491262033	0.3561481
Standard Error	51352.16277	9356.378713	116006.5973	17180.58125	31056.74366	26535.4225
Observation	972	524	886	868	605	891

From Fig. 3, it is very obvious that Mumbai has the largest R^2 value (0.75) followed by Hyderabad (0.58) while Bangalore has the smallest (0.07). More specifically, it indicates that 75% of the variation of all dependent variables in Mumbai can be explained by the regression model, however, only 7% of the variation in Bangalore has a close correlation.

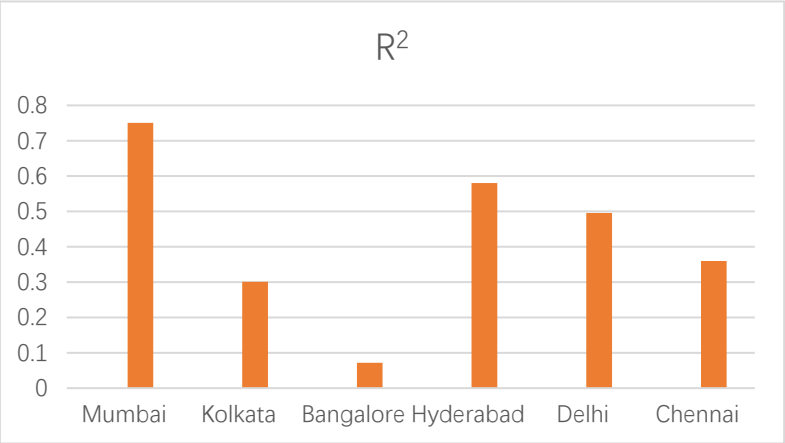

Fig. 3 Comparison of R^2 values between the six cities (Photo credit: Original)

4.3. Logarithmic transformation

After performing another round of regression with logarithmic transformation, we have some additional results.

Table 4 shows the coefficients of the regression for each of the cities. From these coefficients, we can derive a model for each city. For example, the model for Kolkata can be expressed as

$$\log rent = 0.0792783 * \text{BHK} + 0.00021176 * \text{Size} + (-0.0635775) * \text{Floor} + 0.03635651 * \text{Furnishing} + 0.12363036 * \text{Bathrooms} + 3.4873161$$

Table 4. Coefficients of the variables of each city

	Kolkata	Mumbai	Bangalore	Delhi	Chennai	Hyderabad
Intercept	3.4873161	3.7115403	4.13178995	3.6699341	3.65330958	3.69299258
BHK	0.0792783	0.06357035	0.09007733	0.07253449	0.08497004	0.08972869
Size	0.00021176	0.00025701	0.00032202	0.00012619	0.00025124	0.00018509
Floor	-0.0635775	-0.1129652	-0.1409431	-0.0483275	-0.0594339	-0.0574164
Furnishing	0.03635651	0.0400852	0.07481495	0.07248787	0.06042515	0.05895562
Bathrooms	0.12363036	0.07367975	0.03918431	0.1749855	0.04648302	0.03366166

Except for R and R2, we also used p-value to determine the correlation between independent variables and dependent variables.

Table 5 shows the p-values of variables for each of the 6 cities. As we can see, all the p-values are below 0.01, meaning that all observations are statistically significant, and there are strong relationships between independent and dependent variables. Interestingly, the p-value of the variable Bathrooms in Bangalore, Chennai and Hyderabad are significantly larger than the others, all exceeding 0.001, indicating that the relationship between Bathrooms and the rent price might not be so strong. Moreover, the p-value of Floor in Delhi is also greater than 0.001, which means that floor might not be as relevant to the rent price. Lastly, p-values of Size in all six cities are considerably smaller than other variables, showing that among all variables Size has the strongest relationship to rent, which makes sense intuitively as well.

Table 5. P-values of the variables of each city

	P-value					
	Kolkata	Mumbai	Bangalore	Delhi	Chennai	Hyderabad
Intercept	0	0	0	0	0	0
BHK	1.8727E-08	6.8572E-06	8.3828E-09	7.6176E-06	3.7144E-08	2.1879E-13
Size	4.4169E-16	4.6736E-51	3.7473E-40	1.0881E-13	7.6665E-47	1.1809E-42
Floor	6.1915E-05	4.5656E-22	7.4858E-14	0.00432628	2.4392E-06	9.9201E-06
Furnishing	0.00141311	4.2049E-05	5.9122E-18	1.9134E-09	4.9372E-10	2.9086E-11
Bathrooms	3.3772E-12	3.3935E-07	0.00361025	2.5992E-23	0.00216301	0.00764831

Table 6 shows the adjusted R^2 value of rent in the six cities with and without logarithmic transformation. We can clearly see that other than Mumbai, which suffered a minor decrease in R^2 value, all other five models had considerable increases to their adjusted R^2 value, indicating a strong ability of the models to represent the data. Most notably, Bangalore has a significant increase in the accuracy of the model, with the adjusted R^2 value increasing by more than 10 times. This further proves that the log-linear regression model is a plausible model for rent prices in general.

Table 6. Comparison of adjusted R^2 of each city with and without logarithmic transformation

City	Adjusted R^2 with log	Adjusted R^2 without log
Kolkata	0.294267	0.53294205
Mumbai	0.749125	0.73906754
Bangalore	0.066323	0.73226164
Delhi	0.491262	0.68253842
Chennai	0.356148	0.66305786
Hyderabad	0.577645	0.67094438

5. Discussion

The results from analyzing the R2 value and p-value are slightly different. The p-values show that all the factors have a strong correlation with housing rent while R2 indicates that two of the factors,

furnishing status, and floor level do not show close relationships with housing rents. This means that the model or the data we analyzed needs to be improved. We suggest that the difference in the two analyses is caused by the influence of extreme values. The standard errors shown in Table 2 are very large, which means that our data is highly discrete and the final results will be highly affected by extreme values. To improve our model results, it is better to cull out outliers and extreme values.

Moreover, while the models are fairly accurate according to the R2 value, there is no doubt that there is still room for improvement. The R2 values are still not high enough for the model to be considered reliable. There might be a few solutions to this. Firstly, removing these outliers can make our model more accurate, as previously mentioned. Using other advanced modeling methods, such as gradient boosting and ridge regression, can help make the model more accurate as well.

Our model also fails to account for different areas within a city. This may also be a potential point of improvement as intuitively location should have a non-trivial impact on house rents.

6. Conclusion

In conclusion, all six factors (number of BHK, size, floor level, number of bathrooms, furnishing status and city) have a certain effect on housing rents. Among them, in the same location, the number of bathrooms and size of the house has the largest impact on the rents while floor level and furnishing status have the smallest influence. Mumbai has the highest housing rent and correlation with different factors, but Bangalore has the lowest one. In other words, if the buyer's budget is limited, they should consider a house with a small number of bathrooms or smaller areas in Bangalore. Choosing a furnished house can help the buyer save some costs to a certain extent as well as they would not need to spend extra money on furnishing.

References

[1] Dökmeci, Vedia, et al. “External Factors, Housing Values, and Rents: Evidence from Survey Data.” Journal of Housing Research, vol. 14, no. 1, 2003, pp. 83–99. JSTOR, http://www.jstor.org/stable/44944775. Accessed 6 Oct. 2022.

[2] Ming, Yue, et al. “Prediction and Analysis of Chengdu Housing Rent Based on XGBoost Algorithm.” Proceedings of the 2020 3rd International Conference on Big Data Technologies, 2020,

[3] Li, Jinze. “Monthly Housing Rent Forecast Based on LightGBM (Light Gradient Boosting) Model.” International Journal of Intelligent Information and Management Science, vol. 7, no. 6, Dec. 2018, pp. 58–65.

[4] Ma, Tao, et al. “A Study on House Rent Prediction Based on Ensemble Learning.” Finance, Chinese Academic Journal, Hans Publishers, 25 Oct. 2019.

[5] Kumar, Adarsh. “House Rent Price Prediction.” International Research Journal of Engineering and Technology, vol. 06, no. 04, Apr. 2019, pp. 3188–3191.

[6] Mateusz Tomal. “Modelling Housing Rents Using Spatial Autoregressive Geographically Weighted Regression: A Case Study in Cracow, Poland.” International Journal of Geo-Information, May. 2020.

[7] Colema, Andrew, et al. "A Simple Model of Housing Rental and Ownership with Policy Simulations," Working Papers 09_08, Motu Economic and Public Policy Research.

[8] Workspace – Heywhale.com, https://www.heywhale.com/mw/project/5d52f2cac143cf002b22727b.

[9] A., Christina. “8 Factors Affecting Rent Prices.” LinkedIn, 5 Feb. 2021, https://www.linkedin.com/pulse/8-factors-affecting-rent-prices-christina-capoche.

[10] Bogdan. “International Study: Renting Landscape in 30 Countries around the World.” RentCafe Rental Blog, 13 May 2021, https://www.rentcafe.com/blog/rental-market/renting-landscape-30-countries-around-world/