Economic Statistics Issues under Big Data ——Taking China as an Example

Yining Yang*

Henan Experimental High school, Zhengzhou, Henan province, 450002
**Corresponding author: Yiningyang02@163.com*

Abstract

China's economy has increased dramatically over the past decades. Along with the development of economy, the house prices also increased sharply. Thus, it is necessary to investigate factors contributing to the high house prices. Existing literature demonstrates that GDP, salary, total retail sales of consumer goods and population strongly correlated with increasing house prices. However, there is a FU study that investigated the relationship between house prices and those factors in China. We contacted the study to fill the gap.

This study used a quantitative method to analyze the relationships. Specifically, descriptive statistics and regression models Boo request other house prices. The data was selected from 10 large cities in China from 2011 to 2019. This paper finds that there is a strong relationship between house prices and GDP, house prices and total retail sales of consumer goods and house prices and average yearly salary. However, contrasting with past research, the result shows a negative relationship between house prices and the total number of populations. The limitation of the study includes small sample size end problems of multicollinearity. Further research can be conducted using other factors such as fertility rate as the explanatory variable.

Keywords-*house prices, population, multicollinearity, regression, population structure*

1. INTRODUCTION

With the rapid development of economic, Chinese house price has increased hugely since past few decades. There are many factors that are closely associated with house prices, thus it is significant to explore factors promoting the house prices in China and investigate potential relationships between those factors and house prices.

As cities become more urbanized, a greater number of immigrants will move into large cities, resulting in continued growth in housing demand. Housing is a collateralized commodity; therefore, investors will invest in it when the growth in housing demand is rationally expected, which may lead to the fact that housing prices may rise faster than GDP (and income) per capita (Lu et al., 2014).

The first part of the essay will review the literature regarding house prices and relevant factors. The second part of the essay will use descriptive statistics and inferential statistics to explore the relationship between house prices and relevant factors. The last part of the essay will provide a conclusion and limitation of this analysis.

2. LITERATURE REVIEW

Chinese economy has been developing rapidly in recent years, which can be demonstrated by the booming Gross Domestic Product (hereinafter, GDP). Since 2014, China has become the second largest economy in the world measured by nominal GDP ("World Economic Outbreak Database", 2014). Along with the development of the economy, China's house prices have also experienced a significant increase in the past few decades. Many researchers have explored factors that potentially lead to the growth of house prices.

2.1 Determinants of House Prices

Shen and Liu (2004) analyzed the relationship between house prices and market fundamentals between 1986 and 2002. The authors found that Chinese house prices between 1986-2002 can be largely explained by market fundamentals, which include unemployment rate, total populations changes in construction costs, changes in consumer price index (henceforth CPI). Mankiw and Weil (1989) came up with the concept of the "baby boom" in America. They built a model of housing requirements based on age and believed that the growth of these babies

S. A. R. Khan et al. (Eds.): ICMSSE 2022, AHCS 12, pp. 40–52, 2023.
https://doi.org/10.2991/978-94-6463-056-5_8

in the “baby boom “caused a dramatic increase in house prices. Reen & Razali (2016) used all auto aggressive conditional Heteroscedasticity models to investigate determinants of health prices in Malaysia between 2005 and 2013. The results showed that major determinants of house prices contain GDP, inflation rate and house stock. Current literature demonstrated that leading factors influencing house prices including market fundamentals and population structure. Market fundamental factors include GDP, inflation rate, CPI and construction costs. Population structure includes aging of the population and population distribution. Thus, it is important to fill the gap by exploring the relationship between house prices and population structure.

2.2 Determinants of Population Structure

This section will examine factors that exert influences on population structures, including population mobility, population distribution, ageing of population and number of total populations.

The inter-regional income gap affects population mobility. Due to the existence of an income gap, people will move independently between different regions to maximize their own utility (Lee,1996). Also, research conducted by Zhang et al. (2021) examined urban and the rural population structure in China using Jiangsu province from a climate perspective. The results show that within the Poyang Lake area, the influence of climate factors on population mobility is greater than the economic factor. Also, some scholars generally believe that the natural environment, climatic conditions, sports, culture and educational facilities have an important impact on population mobility (Chen, 2015; Cao, 2018).

Further, one of the major factors influencing the urban population structure is the fertility rate. Because of the one-child policy in the late 1970s, the fertility rate in China continually decreased, which contributes to the decreasing of population, especially the working-age urban population (Chen, 2016; Mu, 2017; Song, 2016; Zhong, 2016), which caused a huge demographic transition in China. According to statistics, the fertility rate in China had decreased sharply from 1960 to 2010 (World Bank, 2016).

With regard to the age structure of population, Ye et al. (2019) demonstrated that the growth of economy has a positive effect on health expenditure, which may lead to an increase in the aging population.

2.3 Relationship Between House Prices and Population Structure

Akban and Aydede found that population mobility is significantly related to housing prices and transnational population mobility can most promote the growth of housing prices. According to research conducted by Gabriel et al. (1992), population mobility and migration can be significantly influenced by the differentiation of house prices, using a logistic model. Further, house prices can offset additional incentives for migration such as labour opportunities. Saiz (2003) studied the relationship between population mobility and housing price, he found that housing prices would increase with the increase of the floating population and the growth rate would be larger. Mcquinn and O'Reilly (2008) found that the higher the income level of residents, the higher the housing price.

Houses and children are complements because raising children needs space, Therefore, the rise in house prices increases the cost of childrearing, as the result， the substitution effect would reduce families' demand for children (Dettling & Kearney, 2014; Lindo, 2010). Liu et al. (2020) also studied the effects of house prices on behaviors and intentions of childbearing among women. This means that the relationship between fatality rate and house prices is different among different groups of people. This finding was also supported by other studies. According to research conducted by Dettling and Kearney (2012), a short-term house price increase was associated with a reduced fertility rate among people who do not own houses. Nevertheless, among homeowners, the relationship between house prices and the fertility rate is positively correlated.

3. METHODOLOGY

This section will include research design, sampling method, data collection and descriptive data analysis. The object of this research is to examine whether house prices are related to GDP, population, salary, total retail sales of consumer goods and relations with house prices. Specifically, this essay used a quantitative method (Johnson & Onwuebuzie, 2004) to investigate the research question.

3.1 Research Design

This essay will tackle the research problem by analyzing descriptive statistics using measures of central tendency and data variability. Also, the essay will use multiple regression to explore whether the factors and house prices have linear associations. According to Licht (1999), the complexity of social phenomenon needs adequate tools to investigate behind his explanations. Multiple regression is frequently used to analyze data when there are multiple explanatory variables (Leech et al., 2003). Multiple linear regression allows scholars to investigate combined and independent effects of factors on the dependent variable. In this case, the explain Terry variables are factors potentially influencing house prices including GDP, total number population, average salary and total retail sales of consumer goods. The dependent variable or the outcome variable is the house price.

3.2 Sampling Methods and Data Source

This essay uses a non-random sampling method, purpose sampling, which is frequently used by scholars to select data relying on their judgments. Purposive sampling provides a way of select data in an efficient and manageable manner (Ames et al., 2019). In our essay, data from 2011 to 2019 including GDP, the average salary, the total population and total retail sales of consumer goods were selected from 10 metropolitans in China.

For the selection of data, we chose the average residential commercial house price as the house price because this paper studies the relationship between residential house price and other factors, and therefore the commercial house price should be excluded.

3.3 Hypothesis Development

Descriptive statistics and regression models were constructed to analyze the data. The descriptive statistics were obtained for every year from 2011 to 2019.

Both multiple regression and simple linear regression were regressed on the dependent variable. Also, correlations between explanatory variables were analyzed to examine the existence of the multicollinear problem. The data were analyzed using the tool "Data Analysis" in Excel. Five hypotheses of this study is as following:

H1: there is a significant association between total number of population and house prices

H2: there is a significant association between GDP and city-level house prices;

H3: there is a significant association between average salary and city-level house prices;

H4: there is a significant association between average total retail sale of consumer goods and city-level house prices;

H5: there is a significant association between GDP, total number of populations, total retail sales of consumer goods, and average salary and city-level house prices.

4. EMPIRICAL RESULTS

4.1 Descriptive Statistics

From Table 1, we could see the population in China from 2011 to 2019 has increased only by estimated 12%. Nevertheless, standard deviation of population has decreased from 2011 to 2019. This means that 10 years ago, the distribution of people was more centred in certain cities. In recent years, the distribution of population is more averagely distributed among cities.

TABLE I. MEAN AND STANDARD DEVIATION OF POPULATION OF 10 CITIES FROM 2011-2019

M	1125.80	1134.39	1145.44	1158.52	1168.15	1199.81	1216.54	1238.18	1263.91
Std.	803.90	805.39	806.46	807.56	802.93	805.58	796.73	794.83	786.16

TABLE II. MEAN AND STANDARD DEVIATION OF GDP OF 10 CITIES FROM 2011-2019

M	10503.97	11701.24	12984.94	14207.72	15327.16	16881.23	18806.15	20382.66	20964.10
Std.	4874.97	5101.00	5613.71	6080.47	6514.25	7205.69	7931.65	8791.18	10322.57

From the descriptive table of GDP of 10 cities in China from 2011 to 2019 (Table 2) we can see that GDP half increased hugely. Interestingly, the standard deviation of GDP also increased, this means that the differences in economy between cities have increased in the last 10 years.

TABLE III. MEAN AND STANDARD DEVIATION OF SALARY OF 10 CITIES FROM 2011-2019

M	55713.80	62158.90	70072.80	75560.90	82341.40	90683.10	99743.20	110115.30	122343.90
Std.	12539.43	13096.98	14459.59	15378.14	16693.42	17999.63	19881.74	21833.29	25619.50

From the descriptive statistic table on salary (Table 3), we can see that the average yearly salary has increased from estimated 50,000 Yuan to 20,000 Yuan until 2019. Also, the standard deviation of average yearly salary doubled from 2011 to 2019, this means that the difference of individual wealth also has increased from the past ten years.

TABLE IV. MEAN AND STANDARD DEVIATION OF TOTAL RETAIL SALE OF CONSUMER GOODS OF 10 CITIES FROM 2011-2019.

M	4065.8	4616.2	5189.0	5737.9	6643.8	7298.3	7930.5	8277.3	9234.2
Std.	1856.3	2016.6	2163.8	2230.4	3137.4	3347.7	3597.1	3887.5	3932.3

From this descriptive statistic table of total retail sales of consumer goods (Table 4), we can see that total retail sales of consumer goods have increased from estimated for 1000 to 9000 until 2019. This means that the purchasing power of citizens in China has usually increased from the past ten years. Also, the standard deviation of total retail sale consumer goods increased dramatically.

TABLE V. MEAN AND STANDARD DEVIATION OF HOUSE PRICES IN 10 CITIES FROM2011-2019

M	11419.55	11636.06	13005.60	13214.30	15566.40	19251.20	21825.80	24689.81	26013.40
Std.	5268.26	5068.34	6175.84	6465.58	9312.31	12794.06	13783.04	15472.65	15562.51

From the mean and standard deviation table of house prices (Table 5), we can see that the house prices in China are around 10,000 per meteR-squared 10 years ago and the price doubled until 2019. This means that the house prices increased around 10% on average per year. Additionally, the standard deviation of house prices tripled from the last 10 years.

TABLE VI. CORRELATION MATRIX

	House price	Population	GDP	T R	Salary
House Price	1				
Population	-0.36907	1			
GDP	0.56290508	0.22704105	1		
T R	0.41478768	0.33240306	0.93652754	1	
Salary	0.69823209	-0.1013411	0.78844075	0.77866615	1

From the correlation matrix, we can see that there is a moderate negative association between house price and population. The correlation between salary and house price is the strongest which is almost 0.7. The second strongest correlation between house price and the factors is GDP, which is around 0.56. There is also a moderate to strong correlation between total retail sales of consumer goods and house prices.

With regard to multicollinearity, we can see that there is a strong correlation between GDP and total retail sale of consumer goods there is also a strong correlation between GDP and salary, salary and total retail sales of consumer goods. Thus, to avoid multicollinearity, some factors should be removed from the multi regression model.

4.2 Regression Analysis

This sectional analyzes linear relationships between house prices and four factors, total number of populations, GDP, total retail sales of consumer goods and average salary.

Model 1: House prices and population

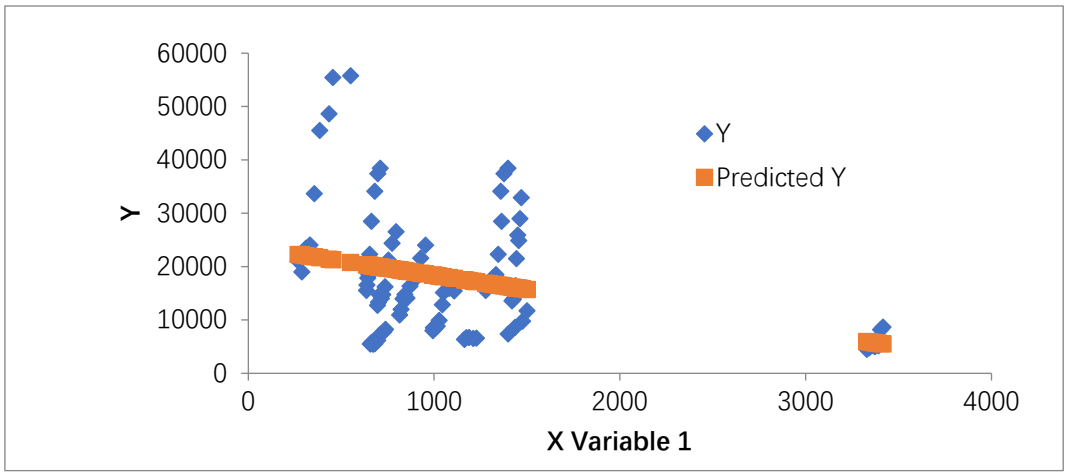

Figure 1. Relationship between house prices and population

TABLE VII. REGRESSION STATISTICS OF HOUSE PRICES AND POPULATION

Regression Statistics	
Multiple R	0.36906997
R-squared	0.13621264
Adjusted R-squared	0.12639687
Standard Error	10895.289
Observations	90

TABLE VIII. ANOVA TABLE FOR SIMPLE REGRESSION (HOUSE PRICES AND POPULATION)

	Regression	Residual	Total
df	1.00	88.00	89.00
SS	1647292610.32	10446244413.49	12093537023.80
MS	1647292610.32	118707322.88	
F	13.88		
Significance F	0.00		

	Coefficients	Standard Error	t Stat	P-value
Intercept	23712.20	2046.45	11.59	0.00
Population	-5.33	1.43	-3.73	0.00

From above figure and tables, we can see that there is a negative linear relationship between house prices and total number of populations. Specifically, one unit increase of population results in estimated 5 unit of house price reduction. The R-squared is around 14%, which means that 14% of their ability in house prices can be explained by number of population. Although the regression model is significant, the relationship between house prices and population is not strong based on the R-squared.

Model 2: houses prices and GDP

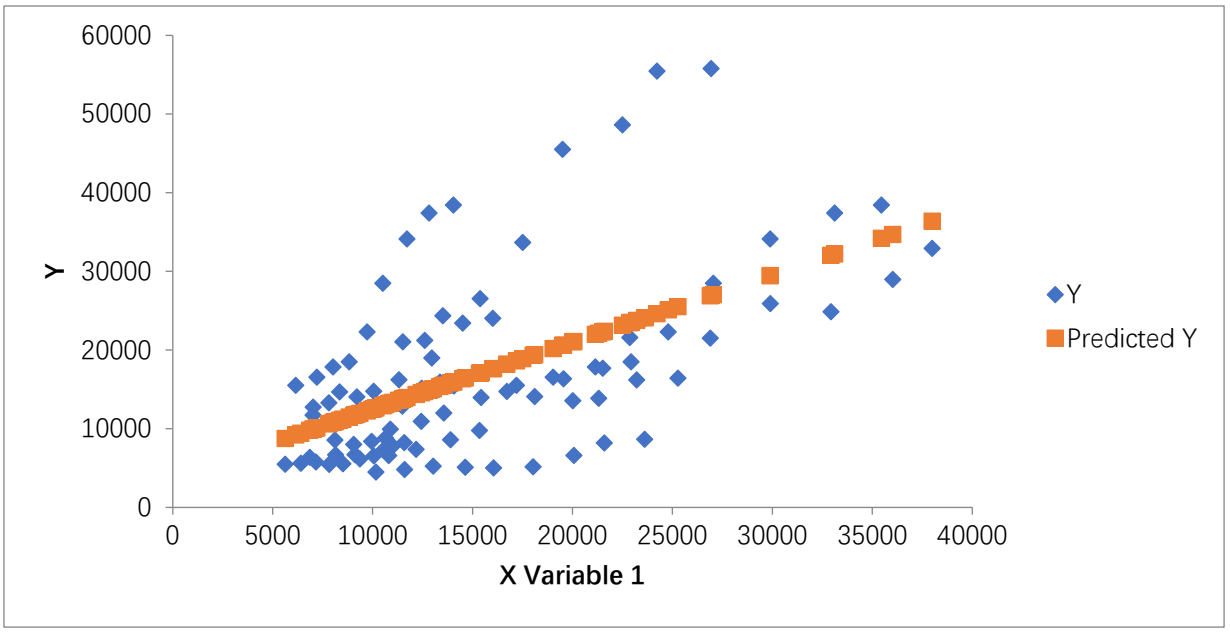

Figure 2. House prices and GDP

TABLE IX. REGRESSION STATISTICS FOR RELATIONSHIP BETWEEN HOUSE PRICES AND GDP

Regression Statistics	
Multiple R	0.5629
R-squared	0.3169
Adjusted R-squared	0.3091
Standard Error	9689.2355
Observations	90.0000

TABLE X. ANOVA TABLE FOR LINEAR RELATIONSHIP BETWEEN HOUSE PRICES AND GDP

	Regression	Residual	Total
df	1.00	88.00	89.00
SS	3831983932.22	8261553091.58	12093537023.80
MS	3831983932.22	93881285.13	
F	40.82		
Significance F	0.00		

	Coefficients	Standard Error	t Stat	P-value
Intercept	3969.036	2337.565	1.698	0.093
GDP	0.853	0.133	6.389	0.000

The above figure and data show that there is a clear positive linear relationship between house prices and GDP. Specifically, why do you need increase of GDP without in estimated 0.85 increases house prices. The P value of the regression model is extremely small; thus, we can reject the null hypothesis at 95% confidence level. The R-squared is around 30%, meaning that 30% their ability of house prices can be explained by GDP.

Model 3: house prices and total retails sales of consumer goods

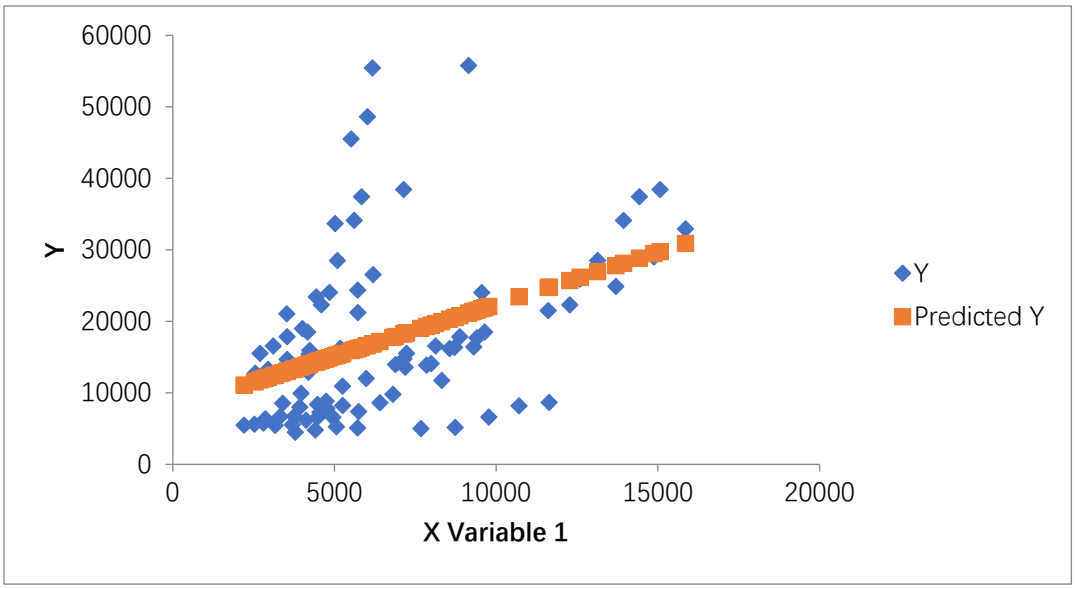

Figure 3. House prices and total retail sales of consumer goods (TR stands for total retail sales of consumer goods

TABLE XI. REGRESSION STATISTICS FOR LINEAR RELATIONSHIP BETWEEN HOUSE PRICES AND TOTAL RETAIL SALES OF CONSUMER GOODS

Regression Statistics	
Multiple R	0.4148
R-squared	0.1720
Adjusted R-squared	0.1626
Standard Error	10666.8871
Observations	90.0000

Though both figure and statistic show that there is a strong positive linear relationship between house prices and total retail sales of consumer goods. Specifically, since the P value is less than 0.05, we can reject the null hypothesis at 95% confidence level. The R-squared is around 17%, which means that 70% of their ability in house prices can be explained by total retail sales of consumer goo

TABLE XII. ANOVA TABLE FOR LINEAR RELATIONSHIP BETWEEN HOUSE PRICES AND TOTAL RETAIL SALES OF CONSUMER GOODS

	Regression	Residual	Total
df	1.00	88.00	89.00
SS	2080678723.55	10012858300.25	12093537023.80
MS	2080678723.55	113782480.68	
F	18.29		
Significance F	0.00		

	Coefficients	Standard Error	t Stat	P-value
Intercept	7882.8490	2493.9908	3.1607	0.0022
TR	1.4523	0.3396	4.2763	0.0000

Model 4: house prices and average salary

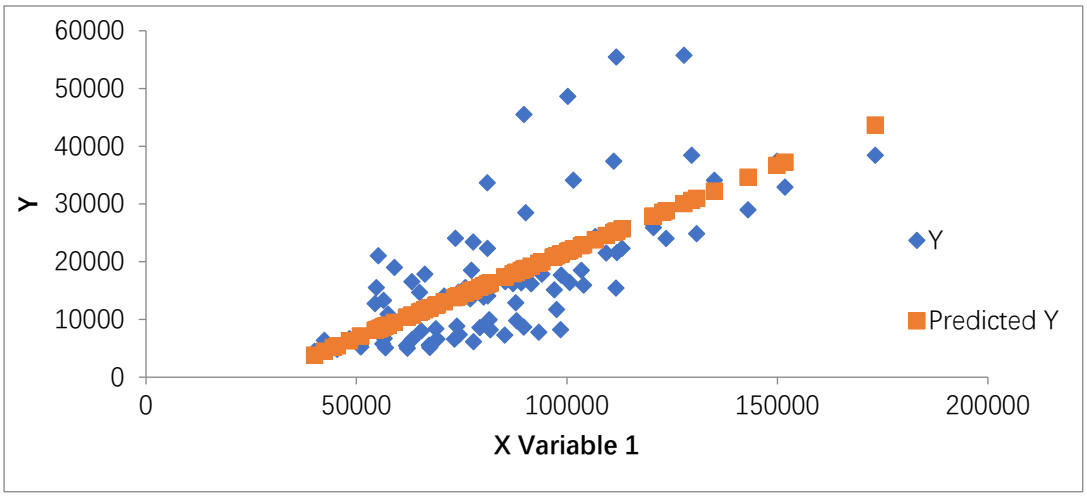

Figure 4. House prices and average salary

TABLE XIII. REGRESSION STATISTICS FOR SIMPLE REGRESSION (HOUSE PRICES AND SALARY)

Regression Statistics	
Multiple R	0.6982
R-squared	0.4875
Adjusted R-squared	0.4817
Standard Error	8392.0949
Observations	90.0000

TABLE XIV. ANOVA TABLE FOR SIMPLE REGRESSION (HOUSE PRICES AND SALARY)

	Regression	Residual	Total
df	1.00	88.00	89.00
SS	5895938483.79	6197598540.02	12093537023.80
MS	5895938483.79	70427256.14	
F	83.72		
Significance F	0.00		

	Coefficients	Standard Error	t Stat	P-value
Intercept	-8152.6614	2929.7450	-2.7827	0.0066
Salary	0.2992	0.0327	9.1497	0.0000

The above figure and data demonstrate that there is a strong positive linear relationship between house prices and average yearly salary. Since the P value is less than 0.05, we can reject the null hypothesis at 95% confidence level. The R-squared is around 49 percent, which means that almost 50% of for ability house prices can be explained by total retail sales of consumer goods.

Model 5: multiple regression model

TABLE XV. MULTIPLE REGRESSION STATISTICS

Regression Statistics	
Multiple R	0.8269
R-squared	0.6837
Adjusted R-squared	0.6688
Standard Error	6708.5214
Observations	90.0000

TABLE XVI. ANOVA TABLE FOR MULTIPLE LINEAR RELATIONSHIP

	Regression	Residual	Total
df	4.00	85.00	89.00
SS	8268174929.15	3825362094.66	12093537023.80
MS	2067043732.29	45004259.94	
F	45.93		
Significance F	0.00		

	Coefficients	Standard Error	t Stat	P-value
Intercept	-3152.52	3242.05	-0.97	0.33
Population	-3.47	1.19	-2.92	0.00
GDP	1.47	0.28	5.34	0.00
TR	-3.06	0.72	-4.26	0.00
Salary	0.25	0.05	4.74	0.00

The above data demonstrate that all of the factors including population, GDP, total retail sales of consumer goods and average yearly salary contributed to the multiple regression models, since the P value is less than 0.05. The adjusted R-squared is around 67%, meaning that 60% their ability house prices can be explained by the four factors.

4.3 Residual Plot Tests

For any regression model, there are two parts, the fitted value and the error term, the latter is called residual and is defined as the distance between fitted value and true value of data. It is significant to assume that the residuals are randomly distributed, since errors normally do not happen with a regular pattern (Kim, 2019). In this section, we will examine whether the residuals of the regression models are randomly distributed by plotting residuals against explanatory variables.

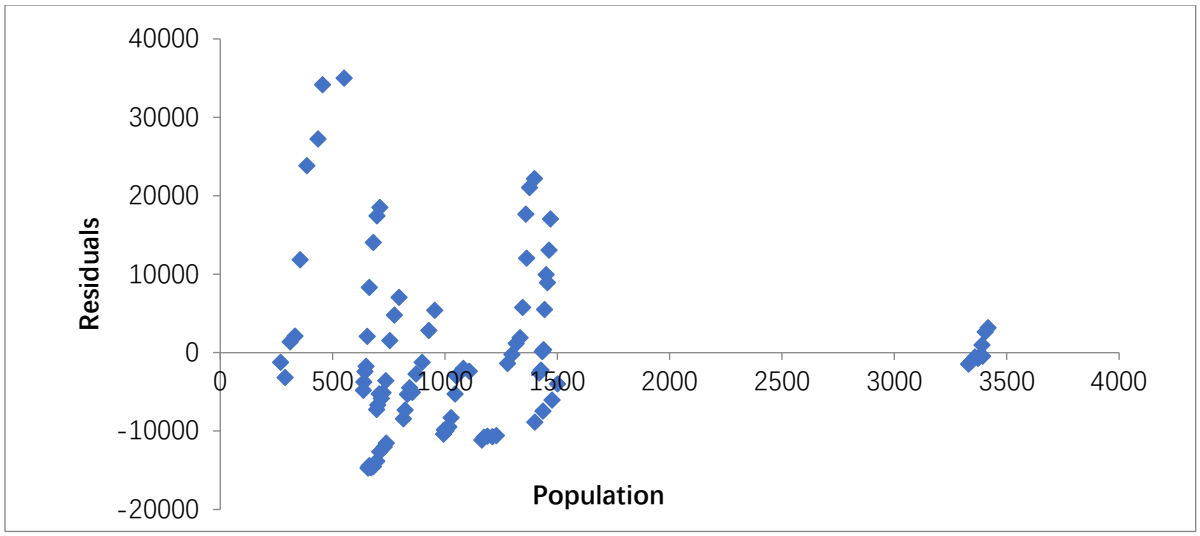

Figure 5. Residual Plot of predicted house prices in simple linear model with explanatory variable total number of populations

It can be seen that the residuals of model 1 are not randomly distributed, the error terms are centred on the left side of the independent variable. It may suggest that there is some correlation between the independent variable.

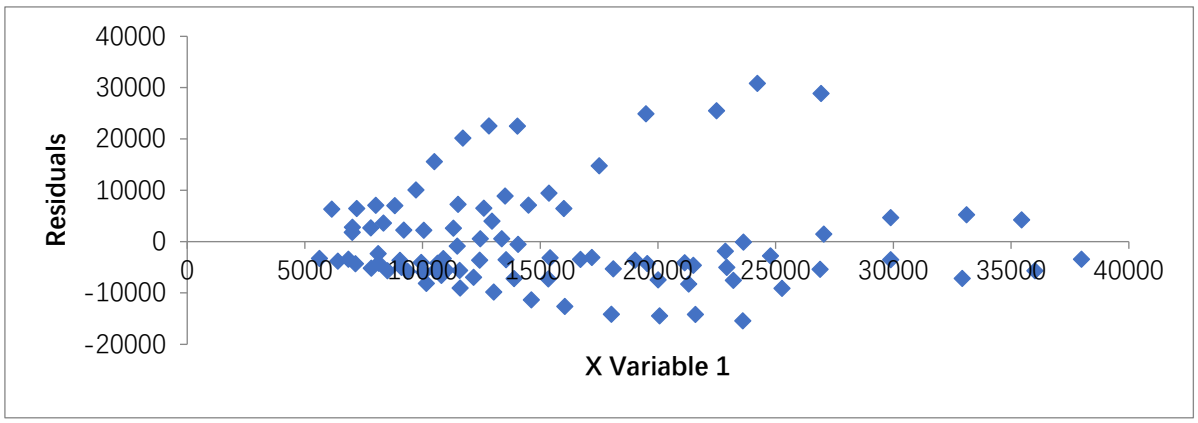

Figure 6. Residual Plot of predicted house prices in simple linear model with explanatory variable total number of GDP

It can be seen that the error terms in model 2 are not randomly distributed either. it may suggest that the independent variable is autocorrelated since this is time series data.

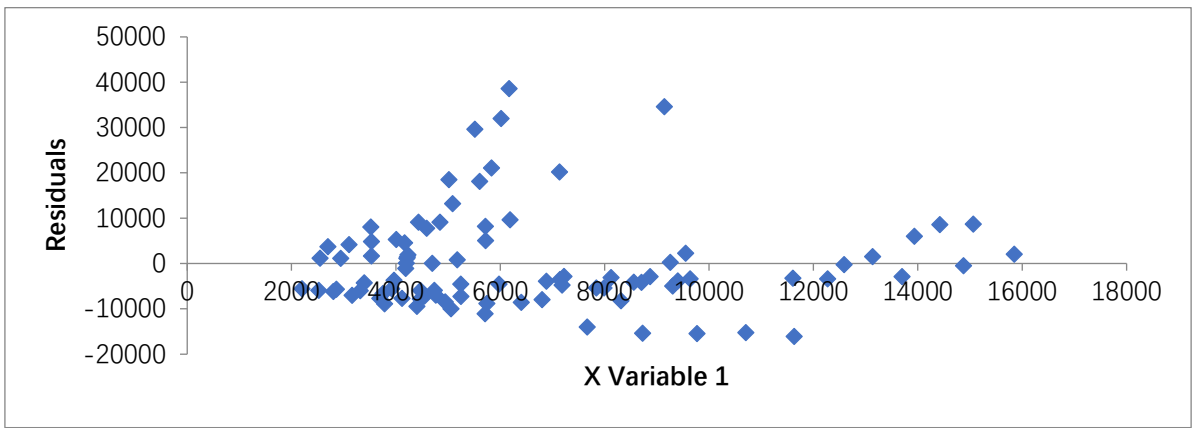

Figure 7. Residual Plot of predicted house prices in simple linear model with explanatory variable total retail sales of consumer goods

It can be observed that the residuals of model 3 are not randomly distributed, it may suggest the same problem as model 1 and model 2.

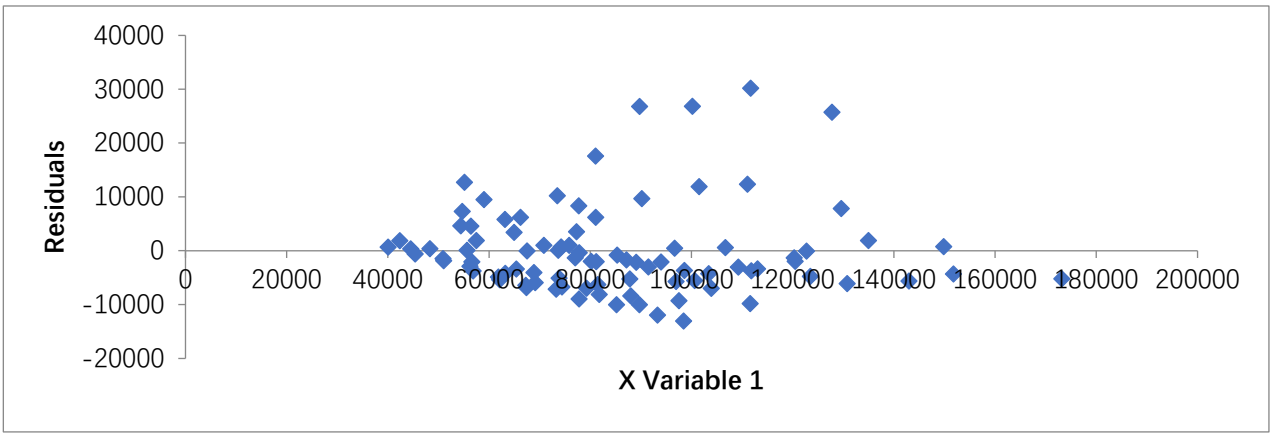

Figure 8. Residual plot of predicted house prices with average salary

The residuals of model 4 are not randomly distributed. Half of the residuals increased within the range from 5000 to 10000 and the other half residuals decreased within the range from 5000 to 10000. Also, half of the residuals decreased within the range from 10000 to 15000 and the other half of the residuals increased within the range from 10000 to 15000. There is a clear pattern of the residuals.

5. CONCLUSION

This paper investigates the relationship between house prices and other factors. Specifically, their result verified the positive relationship between house prices and GDP, house prices and salary and housing prices and total retail sales of consumer goods. However, no positive relationship between house prices and the number of populations was observed.

For a total number of populations, we observed that number of populations has decreased in the past ten years. This is also the case for GDP, total retail sales of consumer goods, average yearly salary and house prices. More importantly, we observed that this distribution of population among cities has increased in the past ten years. Since there is a tendency for people to move from developing areas to develop areas, this possibly means that China's economic center of gravity is no longer concentrated in a few cities but is gradually shifting and dispersing to individual cities. We also observed that the variability of GDP, total retail sales of consumer goods, salary and house prices have doubled or tripled. Although the centre of gravity is no longer central in a few cities, the differences in individual wealth and the economy of cities have increased.

From the regression analysis, we can see that there exist moderate to strong associations between house prices and the four factors we have investigated. The interesting fact is that there is a moderate negative correlation between house prices and population. That behind rational could be that the increased population suggests a more competitive social environment, thus it is more difficult for people to be employed. Thus, the demand for houses also decreased, resulting in decrease in house prices. Overall, we found that all of the five models are significant; however, the variability of house prices can be explained by salary to the largest extent. Then, the second independent variable which accounts for the variability of house prices moderately is GDP. The total number of population and total retails sales of consumer goods have a relatively weak association with house prices, compared to salary and GDP. This result actually aligns with the reality since salary directly reflects the purchasing power of citizens and GDP directly reflects the status of the economy of a country. It can be concluded that when the economy strong, whether at a personal level or national level, the house prices are also strong.

There are some limitations of this research. First, the total sample size is relatively small, and its generalizability needs to be further investigated. The second limitation exists in the regression model, the correlation metrics demonstrate that there is a problem of multicollinearity of our data.

REFERENCES

[1] Akban, A. & Aydede, Y., (2013), Are the educational credentials of immigrant and native-born workers perfect substitutes in Canadian labour markets? a production function analysis. Education economics. 5. London: Routledge, Taylor and Francis..

[2] Ames, H., Glenton, C. & Lewin, S (2019). Purposive sampling in a qualitative evidence synthesis: a worked example from a synthesis on parental perceptions of vaccination communication. BMC Med Res Methodol 19(26).

[3] Chen, Y., Gibb, K., Leishman, C. & Wright, R., (2012). The Impact of Population Ageing on House Prices: A Microsimulation Approach. Scottish Journal of Political Economy. 59.

[4] Dettling, L. & Kearney, M. S., (2014), House prices and birth rates: The impact of the real estate market on the decision to have a baby, Journal of Public Economics, Elsevier, 110(C), 82-100, DOI: 10.1016/j.jpubeco.2013.09.009

[5] Gabriel, S., Shack-Marquez, J.,& Wascher, W. (1992), Regional house-price dispersion and interregional migration, Journal of Housing Economics, 2(3), 235-236,

[6] Glaeser E., Huang, W., Ma, Y. & Shleifer, A., (2017), A Real Estate Boom with Chinese Characteristics, Journal of Economic Perspectives, 3(1), 93-116,

[7] John, R. B., & Onwuegbuzie, A.J. (2004). Mixed methods research: A research paradigm whose time has come. Educational Researcher, 33(7), pp.14-26. doi:10.3102/0013189X033007014

[8] Kim H. Y. (2019). Statistical notes for clinical researchers: simple linear regression 3 - residual analysis. Restorative dentistry & endodontics, 44(1), e11.

[9] LEECH, N., GLINER, J., MORGAN, G., HARMON, R. (2003), Use and Interpretation of Multiple Regression, Clinicians' Guide To Research Methods And Statistics, 42(6), 738-740,

[10] Lu, M., Ou, H., & Chen, B., (2014), Rationality or Froth. An Empirical Study of Urbanization, Migration, and House Prices, National Social Sciences Database, 1, 30.

[11] Mankiw, N. & Weil, D., (1989), The Baby Boom, The Baby Bust and The Housing Market, Regional Science and Urban Economics, 19, 235-258,

[12] Mcquinn, K. & O'Reilly, G., (2008). Assessing the role of income and interest rates in determining house prices. Economic Modelling. 25. 377-390.

[13] Reen, T & Razali, M. (2016). The Dynamics of House Price Volatility In Malaysia.

[14] Journal of Technology Management 3(2).

[15] Shen, Y. & L, H. (2004), Housing Prices and Economic Fundamentals: A Cross City Analysis of China for 1995—2002,

[16] Zhang, L., Lin, X., Leng, L. & Zeng, Y., (2021). Spatial distribution of rural population from a climate perspective: Evidence from Jiangxi Province in China. PloS one.