

Using The Regression Method to Discover the Relationship Between Real Estate and Stock Market

Zixuan Li *

Sydney University Business School, University of Sydney, Sydney, Australia

* Corresponding Author Email: zili9137@uni.sydney.edu.au

Abstract. Recently, with the economic fluctuation, the relationship between house price and the stock market index has been paying much attention. However, this relationship has not been fully clarified yet through historical research. In this paper, several regression models including linear regression model and non-linear regression model are constructed to test for both single and dual relationship between the house price and market index. The assumption and method used in building the model will also be discussed. From testing, there is a positive linear effect on house price from the market index (approximately 80% of house price variation has been explained by the market index). On the opposite side, house price affects 93% of the market index in a positive non-linear pattern. Besides that, GDP can explain around 90% of house price variation in a positive linear relationship. This paper aims to help investors to have a better understanding of the relationship between these two and to help them build a better investment portfolio.

Keywords: Real estate; Stock market index; Regression; Machine learning; Finance.

1. Introduction

Both the real estate market and stock market play an important role in a country's economy, the deviation of the economy is largely related to these two parts. Real estate is related to the whole population in making the choice of renting/buying a house in daily life. Additionally, houses are gradually considered as a speculating financial instrument for investors in the financial market in recent decades. Housing prices also closely relate to the standard of living of citizens. People who live in a house, that may have a high house price, tend to have a high living standard with high income and wealth. Stock as an important financial instrument is also highly related to the economy, the fluctuation can impact the whole country and even the world pandemically, for example Financial Crisis in 2008. Stock can also determine the destiny of a company, the performance of the company can determine the employment, salary and subsidies for employees, and in such ways, it correlates to people's daily lives.

In recent days, researchers and scholars have discovered that there is a relationship between the house price and the market index because both the market index and the house can be considered financial instruments and are all related to the economy to a large extent [1]. From a government perspective, the government can directly use regulation in both real estate and stock markets to take control of the economy efficiently. There are some articles discussing the relationship between house price and market index, especially the unidirectional relationship. However, the unidirectional and dual directional relationships have not been identified yet.

This paper aims to figure out this double direction relationship. According to the data visualization, two variables will be explanatory analysed based on a graph. Several simple linear regression models will be constructed to analyse the simple relationship between two variables, in addition, several polynomial non-linear regressions will also be constructed to transform the variable and analyse the non-linear relationship. GDP and inflation effects will also be discussed in this paper. The paper aims to help the investors have a better understanding of the relationship between house price and market index, and further help them to construct an ideal investment portfolio.

2. Method

2.1. Data resource

The data used to represent house price is the median house price in the USA in each quarter from FRED [2]. As the house price has a high variance and may be very different according to several factors, the mean of the house price makes less sense than the median of the house price. To match with the quarterly sequences, “Total Share Price of All Shares in United States” from FRED [3] is chosen for market index data, rather than S&P 500, as S&P 500 data is daily basis which cannot be reliable in quarterly sequence. GDP data is the “Gross Domestic Product” from FRED [4] which also has a quarterly frequency. CPI data is the “Consumer Price Index Total All Items for the United States” from FRED [5].

For the following regression analysis, both linear and non-linear regression models are based on the assumption of normality distribution. However due to the limited information about the popularity and the inability to gather all market index and house price from the population, this normality assumption exists in doubt, the hold of the assumption is unknown in this case. Although the normality assumption is violated, with a large sample size, this violation has little impact on the result [6].

2.2. Model predetermination and hypothesis testing

1) Linear relationship

Assuming that there is a linear relationship between the market index and the house price, the relationship between house price with GDP and CPI is also assumed to be linear relationships. Therefore, the basic true function and relationship can be assumed to be:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon \quad (1)$$

2) Predicted function

In this case, variables can either be different transformations or different factors.

The predicted function is expected to be:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_n X_n \quad (2)$$

3) Hypothesis testing

With holding CLT, T-test and F-test are mainly used for the following regressions. T-test is mainly for simple linear regression, F-test with joint-testing is mainly used for multiple linear regression and polynomial regression.

Hypothesis for T-test are as follows:

$$H_0: \beta = 0 \quad (3)$$

$$H_1: \beta \neq 0 \quad (4)$$

Hypothesis for F-testing are as follows:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_n = 0 \quad (5)$$

$$H_1: \text{at least one beta is not equal to 0} \quad (6)$$

2.3. Model Construction

In the beginning, two simple linear models were constructed to directly test for the bi-directional relationship between the two variables (Model 1 and Model 2). According to the scatter plot, it shows a slight non-linear relationship between two variables, therefore several transformations are considered to construct the model. Model 3 is to detect the effect of house price on the market index with log transformed house price. Model 4 is to detect the effect of market index on house price with log transformed market index. Squared transformations are then considered in the model. In Model

5, the house price has been squared. In Model 6, the market index is squared. As has been discussed in the previous section, the house price has increased sharply from the 50th percentile position in the time horizon. For the market index, there are changes that happen to be around the 25th percentile and 75th percentile. Therefore, the splines are considered, and the knots are set in 25th, 50th, and 75th percentile position. Model 7 is to test the efficiency of the splines of house price on market index, and Model 9 is to test the efficiency of the splines of market index on house price. In Model 8, both log transformation and splines are combined. From Model 10-13, both GDP and CPI are tested for the relationship with house price individually and combined with market index.

The formula built for each model is shown in the Table 1.

Table. 1 Formula for Each Model

Model	Formula
Model 1	$HP = \beta_0 + \beta_1 MI + \varepsilon$
Model 2	$MI = \beta_0 + \beta_1 HP + \varepsilon$
Model 3	$MI = \beta_0 + \beta_1 HP + \beta_2 \log(HP) + \varepsilon$
Model 4	$HP = \beta_0 + \beta_1 MI + \beta_2 \log(MI) + \varepsilon$
Model 5	$MI = \beta_0 + \beta_1 HP + \beta_2 HP^2 + \varepsilon$
Model 6	$HP = \beta_0 + \beta_1 MI + \beta_2 MI^2 + \varepsilon$
Model 7	$MI = \beta_0 + \beta_1 HP + \beta_2 S_{25th} HP + \beta_3 S_{50th} HP + \beta_4 S_{75th} HP + \varepsilon$
Model 8	$MI = \beta_0 + \beta_1 HP + \beta_2 S_{25th} HP + \beta_3 S_{50th} HP + \beta_4 S_{75th} HP + \beta_5 \log(HP) + \varepsilon$
Model 9	$HP = \beta_0 + \beta_1 MI + \beta_2 S_{25th} MI + \beta_3 S_{50th} MI + \beta_4 S_{75th} MI + \varepsilon$
Model 10	$HP = \beta_0 + \beta_1 MI + \beta_2 GDP + \varepsilon$
Model 11	$HP = \beta_0 + \beta_1 MI + \beta_2 CPI + \varepsilon$
Model 12	$HP = \beta_0 + \beta_1 CPI + \varepsilon$
Model 13	$HP = \beta_0 + \beta_1 GDP + \varepsilon$

Note: HP stands for house price and MI stands for market index

2.4. Model evaluation

To evaluate the model, there are several considerations used. First is the adjusted R^2 , in this article adjusted R^2 is been chosen instead of R^2 , this is because most of the model that is built has multiple factors, and the explanatory proportion always increase as the number of predictors increases, however this increase will generate bias and have little effect on explaining y [7], therefore use the adjusted R^2 , which have already adjusted for the number of predictors, the formula for adjusted R^2 is as follows:

$$R^2 = 1 - \frac{(1-R)(N-1)}{N-P-1} \quad (7)$$

Where P is the number of predictors and N is the sample size. For adjusted R^2 , the higher the better, high adjusted R^2 means that the model has explained the relationship a lot.

Another consideration of model evaluation is the error, which is the standard error of residual (SER). SER is the standard deviation of the error; this measures the variation of error between the

predicted value $\hat{Y}$ and the true value Y . In this case, the lower the SER, the better. Low SER means that the error that the model made is less and the prediction of the model is more accurate.

The final consideration is the number of predictors, the number of predictors is related to the degree of freedom, when the number of predictors increases, the model will become too complex, then the overfitting problem may exist. According to the Principle of Parsimony [8], for similar efficiency, the model is simpler the better. Therefore, degree of freedom (df) may be an evaluation factor here.

3. Result

3.1. Data Visualisation and Explanatory Analysis

3.1.1 House price trend and explanatory analysis

According to the time series Fig.1, the house price is increasing, and the house price is experiencing an upsurge from around 2020 and reaches the highest around 2022, then begins to fall a bit in 2023. The reason for this trend may be because of the government's support after the pandemic to encourage the economy in the United States. This dramatic increase may result from inflation [9]. Therefore, the relationship between inflation and house price is also tested in this paper.

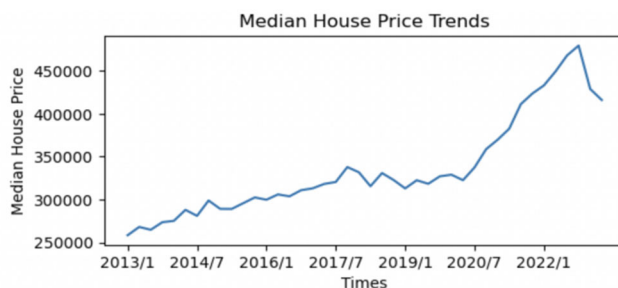


Fig. 1 Time series graph of median house price trends

3.1.2 market index trend and explanatory analysis

For the market index, Fig.2 also indicates an increasing pattern, however unlike house prices which have a dramatic increase in a concentrated period, the market index increases in a more volatility pattern, sometimes it experiences an increase, and sometimes it drops immediately.

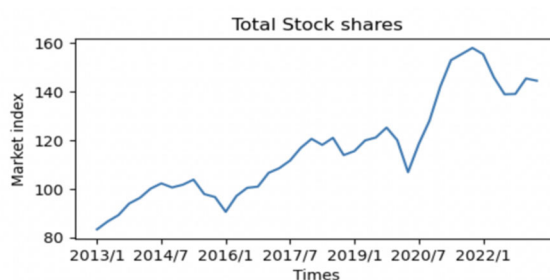


Fig. 2 Time series graph of total stock shares

3.1.3 market index and house price relationship scatter plots

Fig.3 and Fig.4 shows the scatter plot between the house price and the market index, house price is in the x-axis and market index is in the y-axis. From the shows, it can be briefly shown a positive relationship between the market index and house price. And from the right part when the house price is greater than 400,000, it shows a non-linear curve between the two variables. From this, the relationship can be expected to be positive non-linear. However, scatter plots only show the proxy relationship, regressions are needed for further clarification.

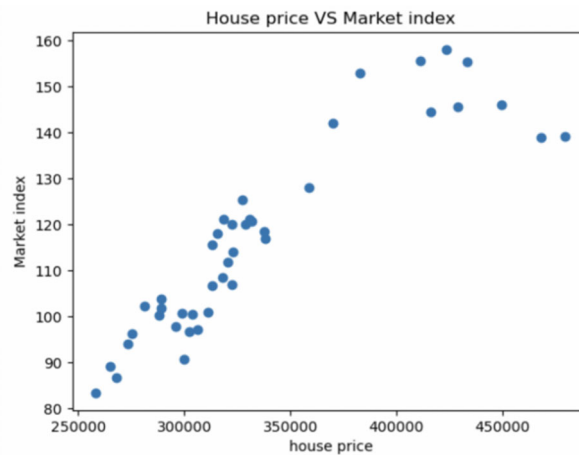


Fig. 3 Scatter plot of house price impact on market index



Fig. 4 Scatter plot of market index effect on house price

3.2. The selection of the best model

Table .2 Evaluation Result for Each Model

	1	2	3	4	5	6	7	8	9	10	11	12	13
R^2	0.80	0.80	0.87	0.80	0.89	0.80	0.89	0.94	0.81	0.91	0.80	0.32	0.91
SER	25529	9	7.5	25644	7.1	25729	7	5.1	25053	17624	25854	47768	17612
DF	1	1	2	2	2	2	4	5	4	2	2	1	1

Table 2 summaries the result of fitness of each model, which is used to measure the performance of the models. Fitness and model measurement can be considered by three components: R^2 , SER, DF. For the SER result in this case, there are two kinds of numbers, some numbers are really large, while some are very small, this is because the unit of response is different, SER has the same unit as response variables. House price is always very high, therefore SER of the model when the house price is the response variable is high.

The regression results for optimal models are shown below:

Table. 3 Best Model Selection Result

Models	Variables	Coefficient	P-value	[0.025	0.975]
Model 1	Market index	2484.5113	0.00	2098.832	2870.190
Model 8	Price	-0.0070	0.00	-0.009	-0.005
	Log_price	2042.7906	0.00	1374.399	2711.182
	Step_1	0.0013	0.00	0.001	0.002
	Step_2	6.62e-05	0.758	-0.00	0.00
	Step_3	0.0008	0.002	0.000	0.001
Model 12	CPI	4.26e+04	0.00	2.98e+04	5.54e+04
Model 13	GDP	18.9671	0.00	16.526	21.408

1) How house price explains the market index

Model 8 has the lowest SER and highest adjusted R^2 . According to the formula in Table 1, in this model both log transformation and spline are considered. In this model, the main contributor is the log price which has a high coefficient between the log price and the market index. Table 3, clearly shows that the second knot is useless, contributes less to the explanation and its p-value is greater than 0.05(which is the significant level that has been set for this test).

2) How market index explains the house price

To explain the house price, Model 1 is doing best without considering the effect of GDP and CPI. According to the formula in Table 1: the formula only considers the direct effect of house price and the market index. From Table 3, it can be shown that the effectiveness of market index is high, the market index explains 80.4% deviation of house price, and clearly shows the high correlation between the explanatory and response variables. If the market index increases by 1 unit, house prices will increase by 2485 dollars in price.

3) What is the relationship between the house price with GDP and CPI.

From Model 12 which is formulated with only the CPI and the house price, the R^2 of the model is only 31.5%. From Table 3: the coefficient of CPI that explains the house price is nearly 0. This means that when CPI increases by 1 unit, this has little effect on the house price. This is violated against the previous expectation which the increase in house price is because of inflation.

GDP seems to have a high correlation with house price. From Model 13 which is constructed purely with the GDP with the house price, the R^2 is 90.7%, which means that GDP can explain 90.7% of the variation in house prices. This high proportion also reflects the high correlation between the house price and the GDP.

4. Discussion

4.1. Time horizon chosen

House price and market index sample that has been chosen at the beginning is from 1963 - 2023, however, according to the consideration that the financial market in both the real estate market and stock market around the 1960s are not complete and not perfect, and not popular as recent days, the data collection in this incomplete market is not verified and may not be accurate. Also, for the house price, the data in the early days may not be accurate.

Therefore, more recent years of data is considerable to slice from the whole population. According to [10], the normal circle for an economy is around 10 years, therefore data from 2013 to 2023 is slicing from the raw data and being used for analysis. By considering sample size and Central Limit Theorem, a very large sample size with a long horizon is chosen, which is from 2013 to 2023 (nearly 60 years of data).

4.2. discussion on the relationship between house price and market index

Market index can positively affect the house price. From the selected Model 1, the coefficient between the market index and house price is positive. In addition, I can conclude that the market index is linearly affecting the house price, because both SER and R^2 of other non-linear models which are destructured to test for the effectiveness of market index on house price have not improved too much. Although Model 9 has the lowest SER(25053) and highest R^2 (80.1%), the fitness is only improved by very little, for R^2 only improved by 0.1%, due to the principle of parsimony [8], with similar effectiveness, model the simpler the better, Model 9 is definitely more complex than Model 1 with 4 variables, therefore Model 9 is considerably not as good as Model 1. Then can conclude that the market index positively and linearly affects house prices.

House price also positively affects the market index. The selected Model 8 is a non-linear model which contains log and spline transformations, which indicate that house price non-linearly affects the market index, and this non-linear relationship is quite strong ($R^2 = 93.9\%$). This non-linear positive relationship supports the previous expectation according to the discussion on the scatter plot.

Besides the mutually positive relationship between house price and market index, GDP also has a positive relationship with house price, from Model 13, GDP can individually explain house price by a large proportion linearly. From Model 12 when CPI increases by 1 unit, this has little effect on the house price. This is violated against the previous expectation which inflation had a large impact on house price. Thus, GDP is significantly positively linearly affected by house price and CPI has no significant relationship with house price.

4.3. Improvement

The third regression assumption, which has been discussed previously, seems to be violated in this case as the data are correlated in the form of time series. The solution can be to try to use the time series method to test for relationship again. Besides that, OVB bias can also be an issue to improve. OVB bias means that there may exist some other variables that can help to explain the house price and market index (besides GDP and CPI) but have not been added into consideration in this case [11]. The ignored variables may change or affect the relationship (especially change the linear relationship to non-linear relationship), however, this bias may always exist as it is impossible to consider and figure out all the possible variables.

5. Summary

In this paper, both explanatory analysis and regression analysis have been used to test for the relationship between house price and market index, in addition to the relationship between house price and GDP/CPI. In explanatory analysis, when analysing individual distributions from time-series plots, house price and market index all show an increased trend with time passed, the dramatic increment seems to happen have the same time. When analyzing the relationship from the scatter plot, it shows that house price seems to have a positive non-linear relationship with market index, in the opposite side, the market index seems to have a positive linear relationship with house price. As the explanatory analysis is only used to provide a basic expectation on the relationship, the detailed relationship is tested from regression to verify these previous expectations.

From regression analysis, several models have been constructed including simple and multiple linear models, as well as the non-linear models. From these models, several models have been selected based on the model selection criteria (R^2 , SER, DF) to best explain the relationship. House prices have a positive non-linear impact on the market index, while the market index has a positive linear impact on house price, therefore can conclude that the two variables both have a positive relationship with each other, the dual direction between these two are positive, but one is linear, and one is non-linear.

Also from the government statement, CPI and GDP seem to have a relationship with house price, from regression analysis, the relationship between GDP and house price has been verified, as GDP

can largely explain the variation in house price (greater than 90%), however, the relationship between house price and CPI has violated the previous expectation as the proportion of explanation by the model is not greater than 50% (only 32%), thus CPI not have a significant impact on the house price.

With quite favourable regression result and quite certain statements, hopefully, the positive relationships between these variables can help investors build the investment portfolio well, and help people get a better understanding of how each economic section are linked together, how a change in one can result in a large sequential result.

References

- [1] Okunev, J., Wilson, P., & Zurbrugg, R. Relationships between Australian real estate and stock market prices - A case of market inefficiency. *Journal of Forecasting*, 2022, 21(3): 181-192.
- [2] U.S. Census Bureau and U.S. Department of Housing and Urban Development, Median Sales Price of Houses Sold for the United States [MSPUS], retrieved from FRED, Federal Reserve Bank of St. Louis; <https://fred.stlouisfed.org/series/MSPUS>, September 6, 2023
- [3] Organization for Economic Co-operation and Development, Total Share Prices for All Shares for the United States [SPASTT01USQ661N], retrieved from FRED, Federal Reserve Bank of St. Louis; <https://fred.stlouisfed.org/series/SPASTT01USQ661N>, September 7, 2023.
- [4] U.S. Bureau of Economic Analysis, Gross Domestic Product [GDP], retrieved from FRED, Federal Reserve Bank of St. Louis; <https://fred.stlouisfed.org/series/GDP>, September 8, 2023.
- [5] Organization for Economic Co-operation and Development, Consumer Price Index: Total All Items for the United States [CPALTT01USQ657N], retrieved from FRED, Federal Reserve Bank of St. Louis; <https://fred.stlouisfed.org/series/CPALTT01USQ657N>, September 8, 2023.
- [6] Schmidt, A. F., & Finan, C. Linear regression and the normality assumption. *Journal of Clinical Epidemiology*, 2018, 98: 146–151.
- [7] Akossou, A. Y. J., & Palm, R. Impact of data structure on the estimators R-square and adjusted R-square in linear regression. *Int. J. Math. Comput*, 2013, 20(3), 84-93.
- [8] Matzke, D., Wagenmakers, E.-J., & Vandekerckhove, J. Model Comparison and the Principle of Parsimony. In *The Oxford Handbook of Computational and Mathematical Psychology*. Oxford University Press, 2015: 300–319
- [9] Aladangady, A., Anenberg, E., & Garcia, D. (2022, November 17). House price growth and inflation during COVID-19. The Fed - House Price Growth and Inflation During COVID-19. <https://www.federalreserve.gov/econres/notes/feds-notes/house-price-growth-and-inflation-during-covid-19-20221117.html>
- [10] Bernard, L., Gevorkyan, A. V., Palley, T. I., & Semmler, W. Time Scales and mechanisms of economic cycles: A review of theories of long waves. *Review of Keynesian Economics*, 2014, 2(1): 87–107.
- [11] Wilms, R., Mäthner, E., Winnen, L., & Lanwehr, R. Omitted variable bias: A threat to estimating causal relationships. *Methods in Psychology*, 2021, 5: 100075.