Data Science and Data Mining

Spring 2024

Bootstrap Regression for Investigating Macroeconomics Factors Affecting USA Home Prices

Benedict Kongyir
Oklahoma State University, kongyirbenkk@gmail.com

Emil Agbemade
University of Central Florida, emil.agbemade@ucf.esu

Bootstrap Regression for Investigating Macroeconomics Factors Affecting USA Home Prices

By:

Benedict Kongyir, Oklahoma State University

(bkongyi@okstate.edu)

Emil Agbemade, University of Central Florida

(emil.agbemade@ucf.edu)

April 2, 2024

Abstract

This study investigates the impact of macroeconomic indicators on US home prices, underscoring the importance of understanding these dynamics due to their significant socio-economic consequences. Utilizing a dataset from Kaggle, originally collected by FRED, the research examines variables like the Consumer Price Index, Population, Unemployment, GDP, Stock Prices, Income, and Mortgage Rate to discern their effect on housing market fluctuations. The analysis identifies multicollinearity among predictors, necessitating a shift from traditional multiple linear regression to a more robust bootstrap regression method due to violations of parametric assumptions.

Key findings reveal that Real Disposable Income is a significant predictor of home prices, although the presence of multicollinearity complicates the model-building process. The bootstrap regression approach, favored for its resilience to assumption violations, confirms the influence of selected macroeconomic factors on home prices. The study concludes that bootstrap regression provides a reliable alternative to parametric methods in cases of assumption non-compliance and highlights the critical role of addressing multicollinearity in regression analysis. This research offers valuable insights for stakeholders involved in the housing market, emphasizing the need for careful econometric modeling in economic policy and investment decisions.

Keywords: Macroeconomic Indicators,US Home Prices,Multicollinearity,Bootstrap Regression, Real Disposable Income

Introduction

The housing market plays a key role in the overall economic sphere. In the United States, the dynamics of the housing sector are particularly crucial, influencing and being influenced by myriad of macroeconomic factors. Interest in the housing sector has increased in recent years due to the clear impact of the evolution of house prices on the crisis that is currently plaguing most economies [6]. This study embarks on a comprehensive investigation into the intricate relationships between macroeconomic indicators and the fluctuating nature of US Home Prices. The significance of understanding the macroeconomic determinants of home prices cannot be overstated. Fluctuations in home prices can have a cascading effect on consumer spending, borrowing patterns, and broader economic stability. Understanding the macroeconomic factors that influence the real estate market is essential for policyholders, economists, and market participants seeking to invest in real estate. The housing market is a very important topic for both the public and private sector[7]. There are socio-economic consequences of changes in house prices. Changes in house prices has influence on rent, insurance, salaries of workers, etc. It automatically impacts the cost of living and development. Knowledge of macroeconomic factors affecting house prices is invaluable information for various stakeholders such as real estate investors, business enterprises, legislators, etc. This study seeks to shed light on the interactions between macroeconomic variables and US home prices. By delving into key indicators such as Consumer Price Index, Population, Unemployment, GDP, Stock Prices, Income, and Mortgage Rate, we aim to identify which of these factors have a significant influence on US home prices and in what way. We also aim to model home prices using an appropriate statistical model.

Research Questions

As stated earlier, the goal of this work is to investigate the various macroeconomic factors influencing house prices in the United States. This study attempts to answer the following specific questions:

- How do fluctuations in macroeconomic variables influence changes in US Home prices?
- Is there a relationship between the various macroeconomic factors and US house pricing?
- Is there an appropriate statistical model to predict future house prices using macroeconomic factors?
- Do demographic factors, such as population growth influence US Housing market?

Data Description

The data used in this study was obtained from Kaggle `https://www.kaggle.com/datasets/faryarmemon/usa-housing-market-factors`, a free online database, but this data was originally collected by FRED `https://fred.stlouisfed.org/`. Accordingly, this dataset will be updated from time to time to include micro and macroeconomic factors. This data can be used for several statistical analyses such as regression analysis to answer several important questions as well as for predictions. There are no missing values in the dataset. The following macroeconomic factors were investigated on their potential effects on changes in house prices: Population, Unemployment Rate, Consumer Price Index, Stock Price Index, Mortgage Rate, real GDP, and Real Disposable income. All variables are continuous. We excluded the date column from our analysis.

Definition of Variables:

- **House˙Price˙Index**: House price changes according to the index base period set.
- **Stock˙Price˙Index**: Stock price changes according to the index base period set.
- **Consumer˙Price˙Index**: The Consumer Price Index measures the overall change in consumer prices based on a representative basket of goods and services over time.
- **Population**: Population of USA (unit: thousands).
- **Unemployment˙Rate**: Unemployment rate of USA (unit: percentage).
- **Real˙GDP**: Gross Domestic Product(GDP)with adjusted inflation
- **Mortgage˙Rate**: Interest charged on mortgages (unit: percentage).
- **Real˙Disposable˙Income (Real Disposable Personal Income)**: Money left from salary after all the taxes are paid

Methods

There are several candidate models that we can consider for this problem. One of these is the well-known multiple linear regression model. However, parametric models like ordinary linear regression rely on several assumptions including the normality of errors. In this study, we built the linear regression model on the data and performed the residual analysis to check for the validity of the model assumptions. As you will see below in the residual diagnostics section, constant variance and normality assumptions are violated. There is also evidence of a presence of outliers in the data from the histogram. Thus, we chose to use the nonparametric counterpart, bootstrap regression which is known to be robust to these assumptions. This study uses a nonparametric bootstrap regression method to model US house prices on some macroeconomic variables. Nonparametric bootstrap regression because the ordinary least squares regression model did not suit the data. The residual analysis from the least squares regression indicates that the errors are not normal. This renders the ordinary least squares regression results unreliable. That is the standard error estimates as well as the confidence intervals obtained are unreliable and cannot be used for any meaningful inferences.

Exploratory Data Analysis

Assessing Relationship Between House Price and the Predictors.

Before building the model, we first of all would like to visually check if at all, there exists some form of relationship between our response and the predictors. This will also help us identify possible predictor variables that are highly correlated which may lead to multicollinearity problems in our final model. We do this using a correlation matrix plot. The correlation matrix plot below indicates a very strong positive correlation between House price and Real disposable income, Real GDP, Consumer price, and Stock price. There is also a strong negative correlation between House price and Mortgage, and a moderate negative correlation between House price, Unemployment rate, and Population. The plot below also indicates that some of the predictor variables are highly correlated and so there is certainly multicollinearity in our data. Multicollinearity is a statistical phenomenon that occurs when two or more independent variables in a regression model are highly correlated with each other. In other words, multicollinearity indicates a strong linear relationship among the predictor variables. This can create challenges in the regression analysis because it becomes difficult to determine the individual effects of each independent variable on the dependent variable accurately [3]

Ordinary Least Squares Regression Model

We build the Ordinary Least Squares Regression Model and check to see if the model assumptions are valid. Variable selection is important in multiple regression [1]. We use forward stepwise regression as a way to include only significant predictors in the model. The

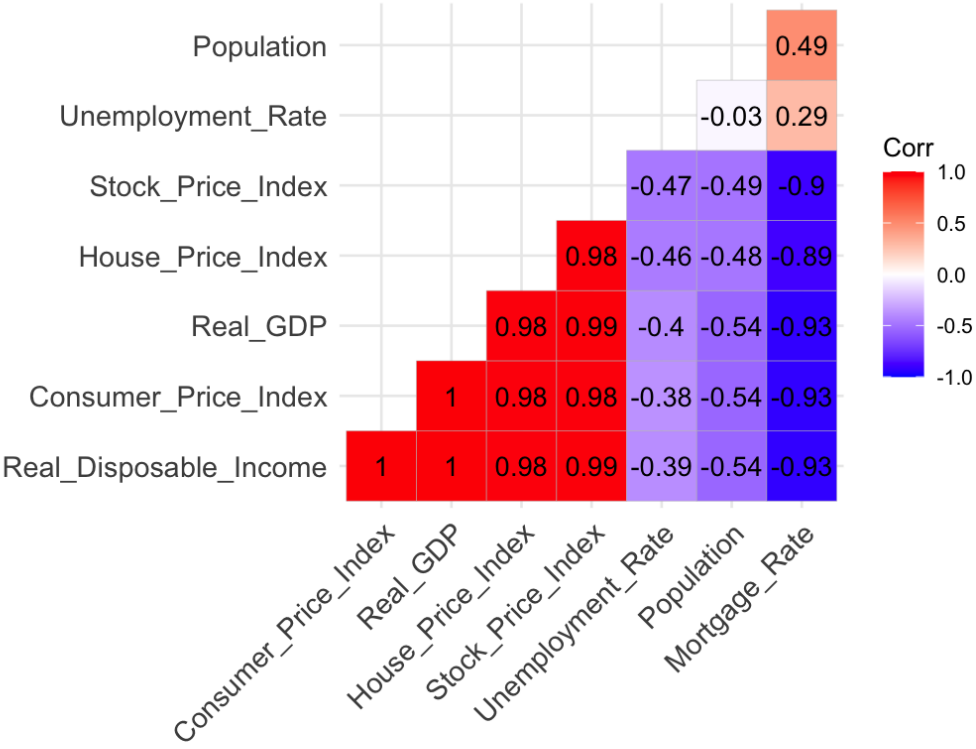


Figure 1: Correlation Matrix Plot of Variables

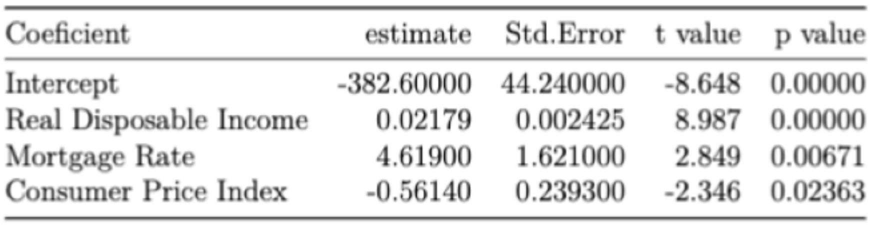
stepwise selection procedure selects Real Disposable Income, Mortgage, and Consumer Price Index as the only significant predictors with multiple R-squared of 0.979. However these three predictors are correlated as shown in the correlation matrix plot above and so to fix this issue we shall drop the Mortgage and Consumer Price Index from our final model. This reduces the multiple linear regression to a simple linear regression problem. The final simple linear regression model is given below with an R-squared of 0.9733 similar to that of multiple R-squared. This further emphasizes the effects of the multicollinearity in the data. Only one of the predictors explains approximately the same amount of variation in the response as all three did.

$$\hat{y} = -235.2 + 0.01484x_1$$

where $\hat{y}$: House Price and x_1 : Real Disposable Income.

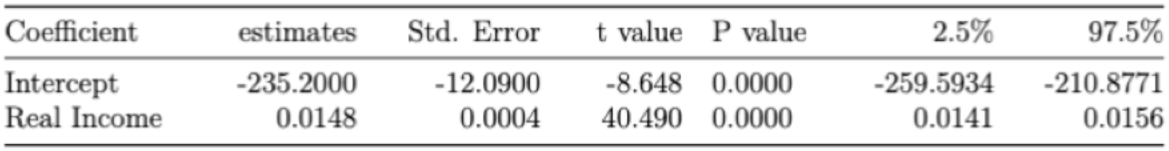
However, as stated earlier, we cannot trust the test statistics as well as the confidence inter-

Table 1: Stepwise Regression Model ANOVA



Coeficient	estimate	Std.Error	t value	p value
Intercept	-382.60000	44.240000	-8.648	0.00000
Real Disposable Income	0.02179	0.002425	8.987	0.00000
Mortgage Rate	4.61900	1.621000	2.849	0.00671
Consumer Price Index	-0.56140	0.239300	-2.346	0.02363

Table 2: Final Simple Linear Regression Model ANOVA



Coefficient	estimates	Std. Error	t value	P value	2.5%	97.5%
Intercept	-235.2000	-12.0900	-8.648	0.0000	-259.5934	-210.8771
Real Income	0.0148	0.0004	40.490	0.0000	0.0141	0.0156

vals and so no meaningful inferences can be made with this model if the model assumptions are not valid. We will check the model assumptions in the next section using the residuals.

Ordinary Least Squares Regression Model Assumptions Diagnostics:

The model diagnostics analysis indicates that the constant variance assumption is not satisfied as depicted in the residual plot below. The residual plot below suggests the variance is increasing with time as depicted by the fun-shape scatter plot of the residuals. The normality assumption is also violated as shown by the histogram and the quantile-quantile plot. This conclusion is also supported by the Shapiro-Wilk normality test results with p-value = 0.008381. Normality assumption is very important for linear regression analysis [2]. These violations warrant a search for an alternative robust model. Bootstrap regression is our best choice in this situation.

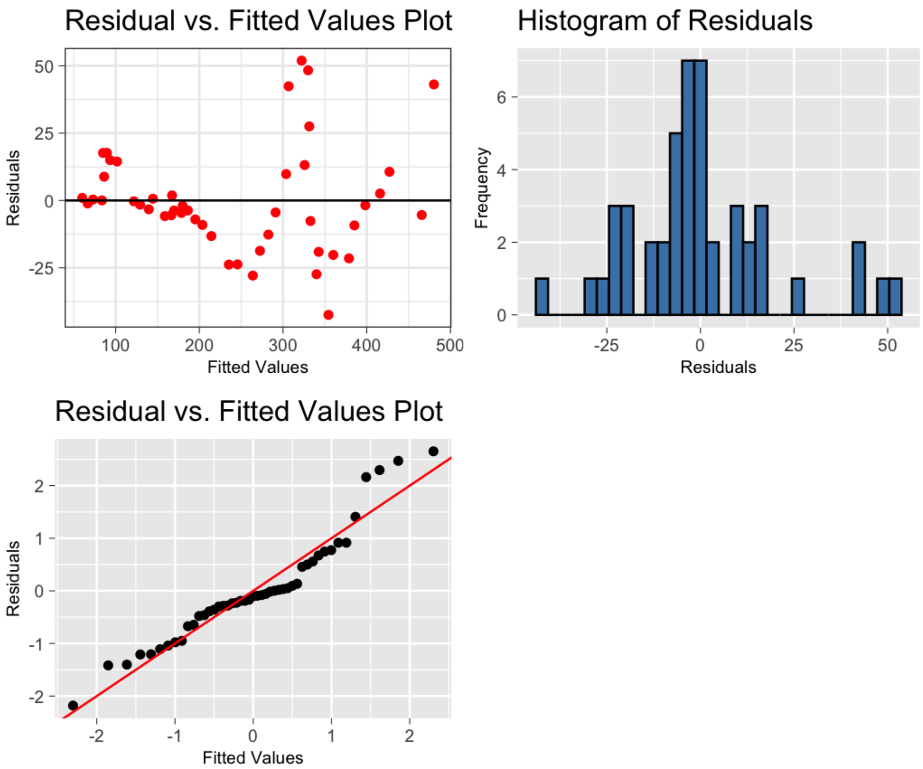


Figure 2: Residual Analysis Plots

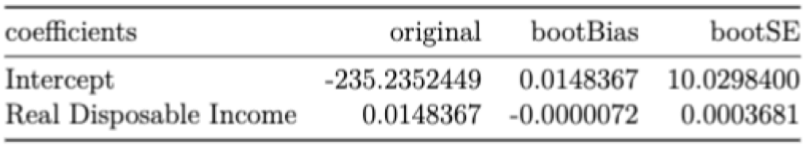
Bootstrap Regression model

Bootstrapping is a general approach to statistical inference based on building a sampling distribution for a statistic by resampling from the data at hand [5]. It is known to be robust to model violations and thus gives more accurate and reliable results than parametric methods when parametric assumptions are not satisfied. nonparametric bootstrap allows us to estimate the sampling distribution of a statistic empirically without making assumptions about the form of the population, and without deriving the sampling distribution explicitly [5].

Since the linear regression model assumptions are violated, we resort to building our robust nonparametric bootstrap regression. We first use the resampling in R and use the boot function in R to confirm our results. The results of the two procedures are pretty similar. We report the results of the boot function.

Bootstrap Estimates

Table 3: Bootstrap Estimates

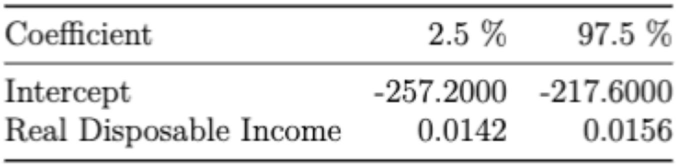


coefficients	original	bootBias	bootSE
Intercept	-235.2352449	0.0148367	10.0298400
Real Disposable Income	0.0148367	-0.0000072	0.0003681

Bootstrap Confidence Intervals

There are several approaches to constructing bootstrap confidence intervals but the bias-corrected(BCA) intervals are preferred[5]. In this work, We report only the *bca* confidence interval as it is known to be less biased and more accurate. From the bootstrap confidence interval table below, we are 95% confident that the average consumer price index is between 0.0142 and 0.0156

Table 4: Bootstrap Confidence Interval



Coefficient	2.5 %	97.5 %
Intercept	-257.2000	-217.6000
Real Disposable Income	0.0142	0.0156

Distributions of bootstrap parameter estimates

Let's take a look at the density plots of the bootstrap parameter estimates. We can see from the density plots that each of the bootstrap estimates is approximately Gaussian. This is the sampling distribution of the parameter estimates.

Discussion of Results

In the results above we see both the ordinary least squares regression and the bootstrap are candidate models for solving this problem however, the nonparametric bootstrap procedure

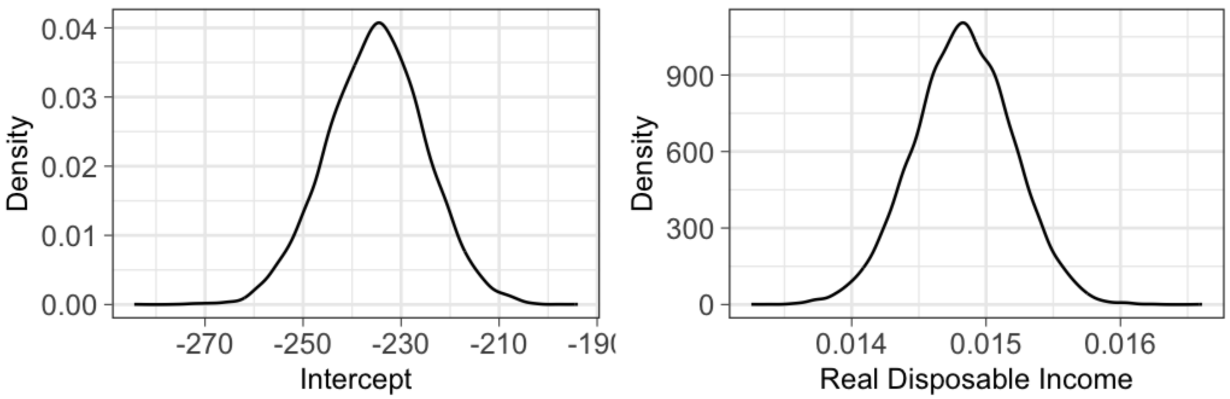


Figure 3: Correlation Matrix Plot of Variables

is more appropriate due to the non-normality observed from the residual analysis. This makes the results of the ordinary least squares regression procedure unreliable. The test statistic as well as the confidence intervals obtained from the ordinary least squares regression procedure can not be trusted and so cannot be used for any meaningful inferences. The confidence intervals for both procedures are similar as reported in the tables above. In theory, the bootstrap confidence intervals are considered valid and reliable. We also see how the multicollinearity can be an issue in multiple regression if not addressed properly. Its presence can cause the regression coefficients to become unstable and difficult to interpret, which can lead to wide confidence intervals and increased variability in the predicted values of the dependent variable. Understanding what causes it and how to detect and fix it can help us to overcome these problems[3]

Conclusion and Recommendations

In this work, we have highlighted the power of the bootstrap approach for estimating regression coefficients. We also demonstrated some of the advantages that the bootstrap method has over the parametric procedure. In general, the bootstrap procedure uses the ordinary least squares model without relying on the assumptions, which is a cool idea. We have also demonstrated the importance of checking multicollinearity in data when building a multiple regression model and the need to address it appropriately to avoid inflated standard errors and p-values, which can lead to incorrect conclusions about their statistical significance. This work does not seek to discredit ordinary least squares regression in general, but to encourage the use of nonparametric bootstrap regression when the need arises, that is if the least squares regression assumptions are violated. [4]

Bibliography

[1] Emil Agbemade. "Variable Selection and Regression Analysis". In: (2023). URL: `https://stars.library.ucf.edu/data-science-mining/11`.

[2] Amir Alipour Yengejeh. "A linear regression model to predict the critical temperature of a superconductor". In: (2023). URL: `https://stars.library.ucf.edu/data-science-mining/12`.

[3] Aniruddha Bhandari. *What is Multicollinearity? Here's Everything You Need to Know.* Analytics Vidhya, Mar. 2020. URL: `https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/`.

[4] S Dakurah et al. "A Model for Pricing Insurance using Options". In: ().

[5] John Fox and Sanford Weisberg. "Bootstrapping regression models". In: *An R and S-PLUS Companion to Applied Regression: A Web Appendix to the Book. Sage, Thousand Oaks, CA. URL http://cran. r-project. org/doc/contrib/Fox-Companion/appendix-bootstrapping. pdf* (2002).

[6] Antonio Montanés and Lorena Olmos. "Convergence in US house prices". In: *Economics Letters* 121.2 (2013), pp. 152–155.

[7] Panos Pashardes and Christos S Savva. "Factors affecting house prices in Cyprus: 1988-2008". In: *Cyprus Economic Policy Review* 3.1 (2009), pp. 3–25.