Final Report

Topic: Analyzing Homeownership in the United States

Team 04: Qianqian Zhou, Liangqu Chen, Baiyan Ren

Summary

In this study, we analyzed homeownership rate in the United States via visualization of time series data. We further studied several external factors which might contribute to the homeownership rate change. Among these factors, mean sales price, housing permit, and initial claim were found to granger-cause the homeownership rate, though they don't have co-integration. Most importantly, we built multiple time series models to forecast the homeownership rate. To normalize the data, we applied log transformation. We started with benchmark models (e.g., Mean, drift, Naïve, splines regression model) then increased the model complexity (e.g., ARIMA, ARIMAX, VAR). Among these models, the performance of VAR model on the log-transformed data was better than other models. However, ARIMAX model on the original data is the only model capturing the spike in the forecast. Such a model would help stakeholders make decisions responding to market fluctuations. Thus, we would choose ARIMAX and VAR model to make the forecast.

Introduction

For most of us, our largest investment is purchasing a house, often with financial tools like mortgage loans. As of 2020, spending on residential fixed investment was about 885 billion dollars contributing 4.2% of GDP (Gross domestic product), along with 2.8 trillion dollars spent on housing services, which accounted for 13.3% of GDP [1]. Thus, the housing market plays an essential role in the supply and demand of the broader economy. As an indirect impact of the housing market on the macro economy, homeowners are more willing to spend more when housing prices are high, which drives consumer spending as known as wealth effects. This project aims to build a model to forecast the homeownership rate, which is one of the key metrics to study for the housing market.

The homeownership rate in the United States is defined as the percentage of homes that are owned by their occupants [2].

$$Homeownership\ Rate\ (\%) = \left[\frac{Owner\ occupied\ housing\ units}{Total\ Occupied\ housing\ units}\right] * 100$$

It is affected by multiple external factors, including economic status, interest rates, real income, and unemployment [3, 4]. Recently, with the increase in the median existing-home sales price and mortgage interest rates spiking to 6.7%, the barrier to homeownership has become higher [5]. We aim to build a time series model with a group of selected factors that are proven to be statistically significant to predict the homeownership rate. In addition to the dataset of interest rate, median home price, real GDP, and homeownership rate as provided for the project, we also consider adding additional factors (e.g., Nonfarm payrolls, CPI (consumer purchase index), unemployment rate, et al.) to predict the homeownership rate.

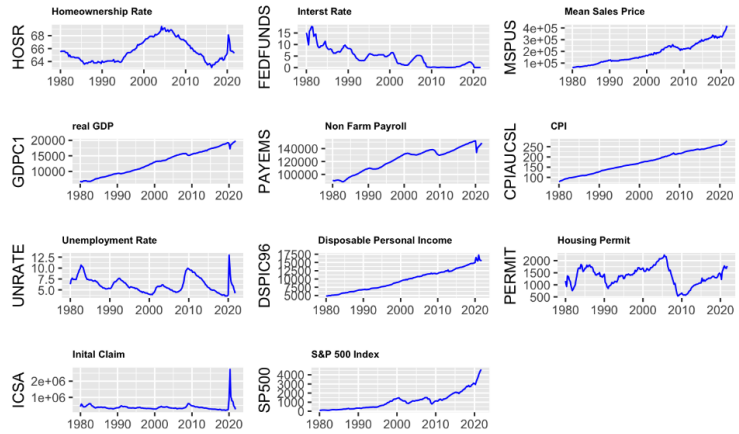


Figure 1. Time series plot of homeownership rate data and other external variables

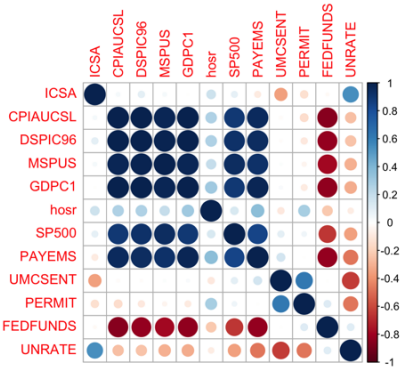


Figure 2. Correlation among variables

Analysis

Exploratory Data Analysis

The homeownership rate has a non-linear trend from 1980 to 2021 (Fig. 1). Similar trend could be observed in the housing permit data. Other variables on an uptrend (e.g., mean sales price, real GDP, nonfarm payrolls) reflect the booming/inflating economy in the last 40 years and are worth exploring.

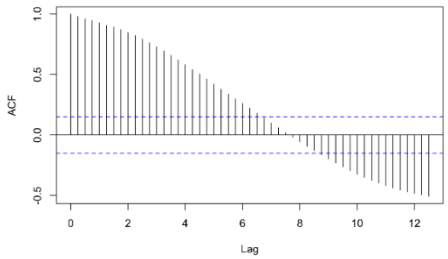


Figure 3. ACF plot of homeownership rate data

Among the external variables, there are some highly correlated pairs (Fig. 2). For example, real GDP and nonfarm payrolls are positively correlated and interest rate is negatively correlated with CPI. Subsequently, non-farm payroll has a positive impact on the homeownership rate and the interest rate does the opposite. These pairs of variables with correlation will be filtered before modeling to avoid collinearity.

The autocorrelation is decreasing slowly with increasing lags, indicating the existence of a trend in the homeownership rate data (Fig. 3). In addition, no seasonality is observed.

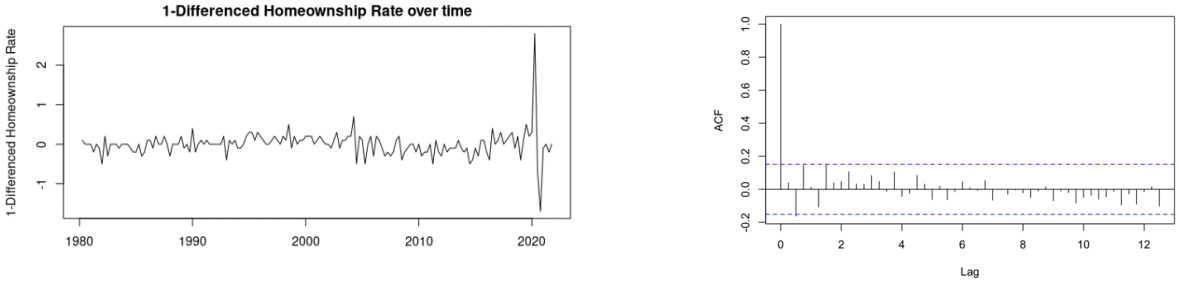


Figure 4. Time series plot and ACF plot of 1-step differenced homeownership rate data

To study the property of homeownership data after removing the trend, we performed a 1-step difference and then visualized the time series and ACF (autocorrelation function) of the differenced data (Fig. 4). The data exhibits a constant mean and stable variance from 1980 to 2020. When the market was hit by COVID-19 in early 2020, the homeownership rate fluctuated and showed increased variability. All the ACF after lag 0 are within a significant band, suggesting that the 1-step differenced data is weakly stationary.

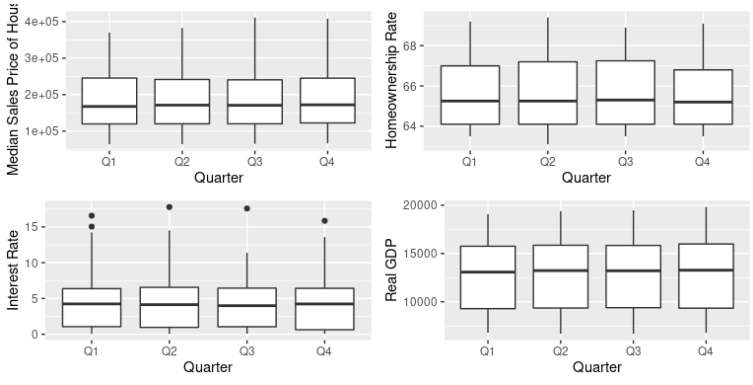


Figure 5. Box plot of homeownership rate, mean sales price of house, interest rate, and real GDP.

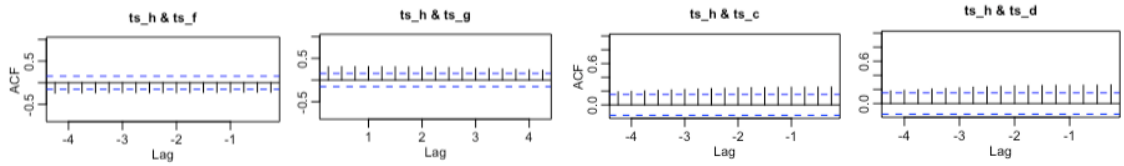
Additionally, we created box plots for the homeownership rate and three external variables (mean sales price of house, interest rate, and real GDP) by quarter (Fig. 5). The plots showed that the values of each variable among quarters are similar, indicating that there is no quarter seasonality.

Data transformation

As several external variables are increasing exponentially due to economic booming, such as CPI, GDP and S&P 500, log transformation is conducted for each variable. Thus, there are two datasets we use for training and testing models, one original and another one from log transformation. We also split the datasets into training sets and testing sets. The training sets exclude the last 15 points, and the points are used for testing models.

Cross correlation and contemporaneous correlation

GDP, CPI, disposable income, non-farm payroll, housing permit, sp500, consumer sentiment, unemployment rate exhibits potential leading relationship to homeownership rate, as their cross correlation ACF are outside of the significant band. None of variables shows contemporaneous correlation to the homeownership rate



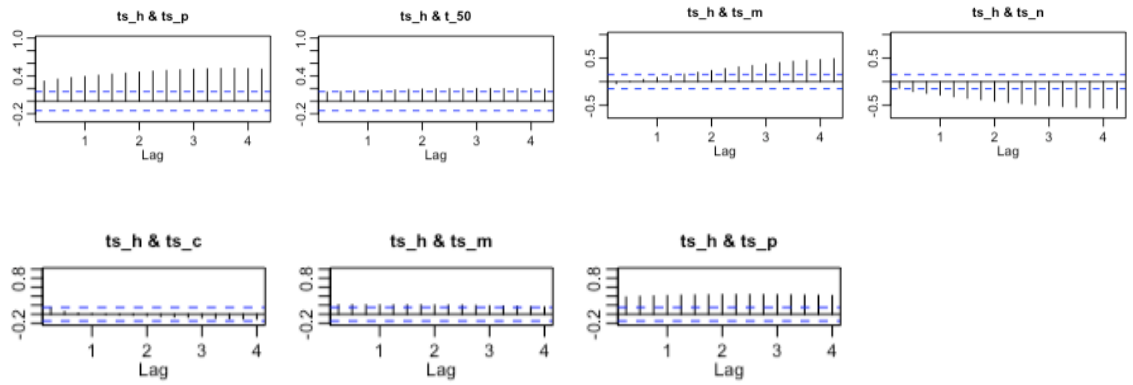


Figure 6. Cross correlation and contemporaneous correlation

Granger test

To identify the external factors which could be useful in forecasting homeownership rate, we employed Granger causality test. For the original data, initial jobless claim, housing permit, and median housing price showed granger causality on homeownership rate based on the p values. For the log transformed data, only initial jobless claim had granger causality. Therefore, in the following analysis, we chose these factors to assist forecasting homeownership rate.

	p value	
granger cause	original	log transformation
CPI	0.77	0.26
Disposable Income	0.25	0.23
Interest rate	0.52	0.52
GDP	0.37	0.06
Initial Jobless Claim	2.90E-05	7.40E-10
Median Housing price	0.013	0.26
Non Farm Payroll	0.29	0.16
Housing permit	0.0061	0.62
S&P 500	0.43	0.67
Consumer Sentiment	0.4	0.62
Unemployment Rate	0.33	0.13

Table 1. Granger causality test on original and log-transformed data

Co-integration

In addition to granger causality test, we also evaluated the co-movement or cointegration of the factors that we believe may influence the homeownership rate. However, both Engle Granger Augmented Dickey Fuller test and hypothesis test procedure showed that none of these factors, which were identified to granger cause the homeownership rate, have co-integration with the homeownership rate.

Time Series Models

Based on the properties of input data that it shows non-linear trend and no seasonality, we employed several models to fit the homeownership rate data. Below is the forecast result for the last 15 time points. From the combined forecast plot, we think the ARIMAX model using the original data performed the best due to the fact that this model can capture the trend and pattern of observed data,

having the peak value at the same time.

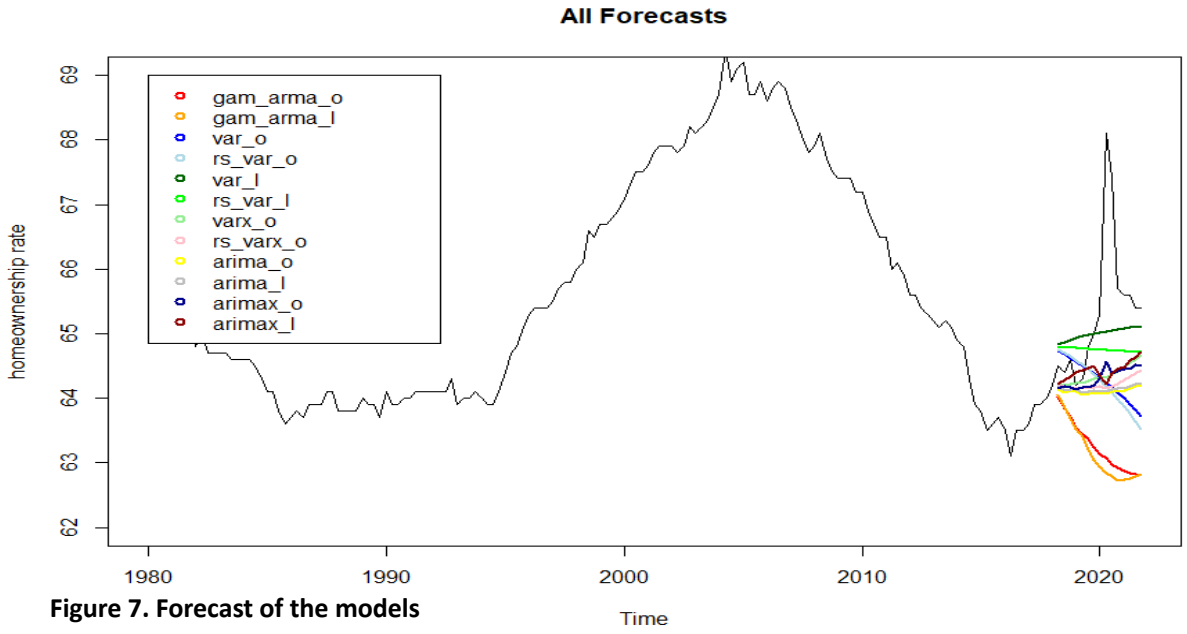


Figure 7. Forecast of the models

When we check PM, MAPE results for every model, it came up with the different model, VAR model using log data, which has the highest predictive power out of these models.

	MAPE	PM
gam_arma_org	0.0320218	5.467065
gam_arma_log	0.03345661	5.973434
m_var_org	0.01844707	2.376715
m_var_log	0.01067006	1.024489
m_rs_var_org	0.01947566	2.582199
m_rs_var_log	0.01288571	1.341864
m_varx_org	0.01500262	1.794987
m_rs_varx_org	0.01700485	2.111995
m_arima_ori	0.9812669	2.360032
m_arima_log	0.9817433	2.290033
m_arimax_ori	0.9844024	1.747695
m_arimax_log	0.9861899	1.786668

Table 2. MAPE & PM result of Models

So based on the plot and accuracy of the forecasts, we choose to provide 2 models: ARIMAX with original data & Unrestricted VAR with log transformed data for forecasting homeownership rate.

Next, we will walk through more details of these models individually.

Benchmark

First, we built benchmark models using simple algorithms (e.g., Mean, drift, Naïve, splines regression model) to provide a baseline comparison. These benchmark models used the homeownership rate as the input data. As we could see from the below 2 plots, no matter whether using the original data or log data, the Spline Model performs best, it captures the peak value and the valley value. Furthermore, we combined the Spline Model with ARMA for more exploration.

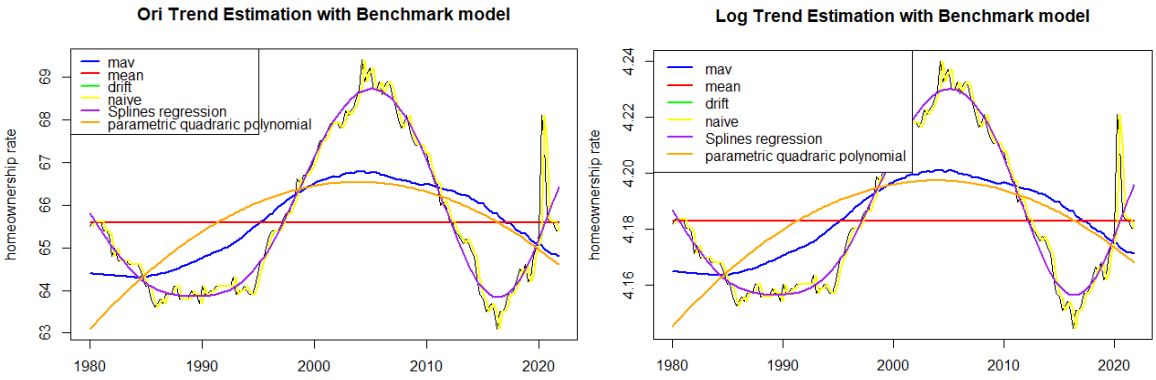


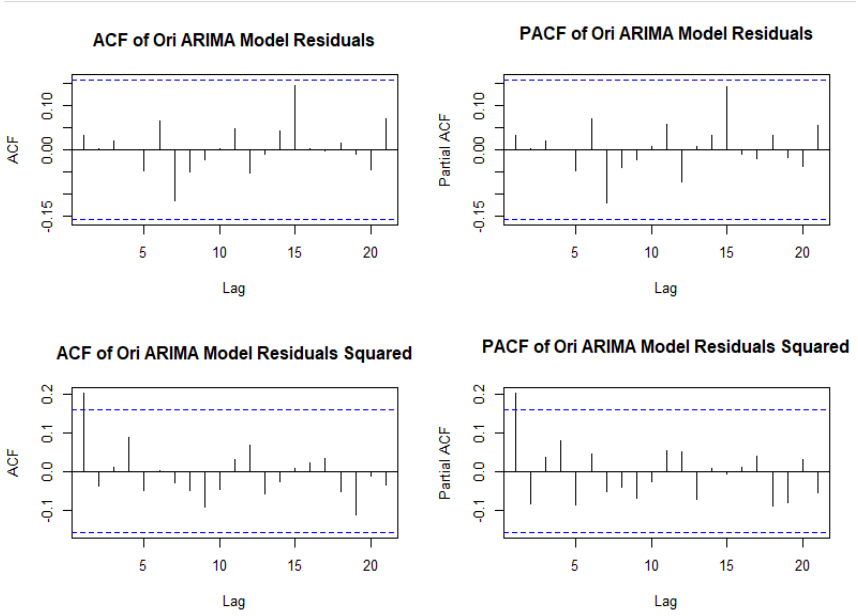
Figure 8. Combined benchmark models on original data and log-transformed data

Spline + ARMA

Splines is chosen for fitting the trend due to its great fit to the data and an ARMA model is used for fitting the residuals. We applied the model combination to both original and log transformed data. The residuals of both models resemble white noise and are normally distributed.

ARIMA

We further applied ARIMA (4, 0, 3) model to fit the trend of homeownership rate. We used original data and log transformed data. Both models captured the time series change of homeownership rate well. Importantly, the squared residuals of both models didn't show serial correlation, they are both white noises. Thus, ARIMA model combined with GARCH did not fit here.



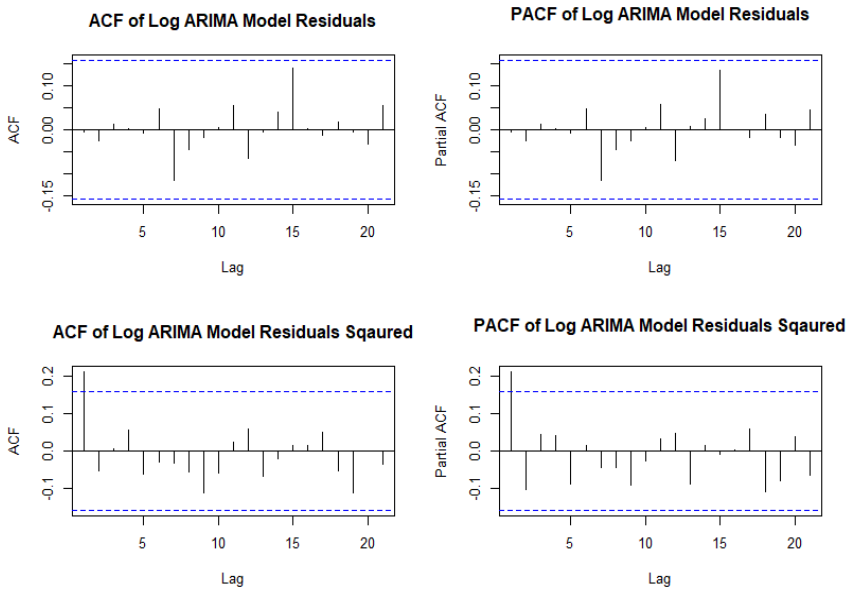


Figure 9. ACF and PACF plot of residuals and squared residuals of ARIMA model on original and log-transformed data

VAR

As suggested by the granger causality test, initial jobless claim, median sale price of house and housing permit were helpful to predict homeownership rate of next period for original dataset. Thus, these variables were used to build VAR models on the original dataset. For log transformed data, only initial jobless claim was significant in the granger causality tests. To build VAR on the log transformed data, only initial jobless claim was picked. In addition, we also applied restrict function in R to obtain restricted VAR model.

VAR model trained on the log transformed data exhibited the best predicting accuracy, even though there are few actualized homeownership rate points outside of the upper bound. Order of 2 was used as selected by BIC. The only variable that was significant was homeownership rate at lag 1. The residuals were normally distributed and as we can tell from ACF plots, they showed weak stationarity.

P-value of the ARCH test was 0.24, thus we cannot reject the null hypothesis of constant volatility. And p-value of JB test was 0.126, which suggests that the residuals were normally distributed. The portmanteau test had a p-value < 0.05 thus, we rejected the null hypothesis of uncorrelated errors, meaning the residuals were serially correlated.

Var models can capture the impact from other variables/factors at different lags, which as a result, provides the best forecast accuracy out of all the models. In the analysis of economy study, we would like to take advantage of this because the macroeconomics works like a giant cycle and each part of the economy has impacts on each other. Thus, VAR model is the best choice for this reason. In this case, the initial jobless claim as a leading predictor help to improve its forecast accuracy, which implies that when people start losing their jobs, they are less likely to own a house in near future.

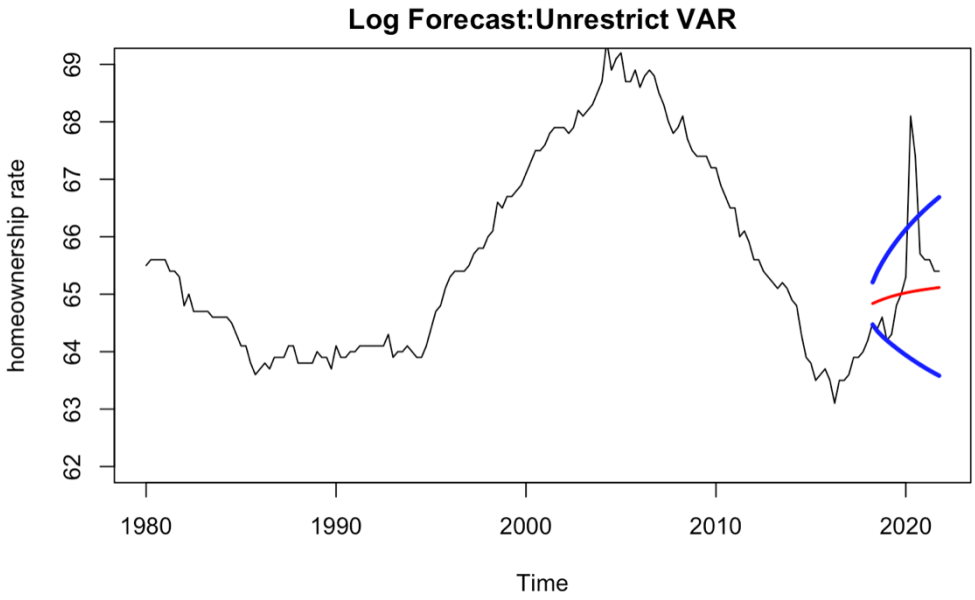


Figure 10. unrestricted VAR prediction trained on log transformed data with lower and upper bounds

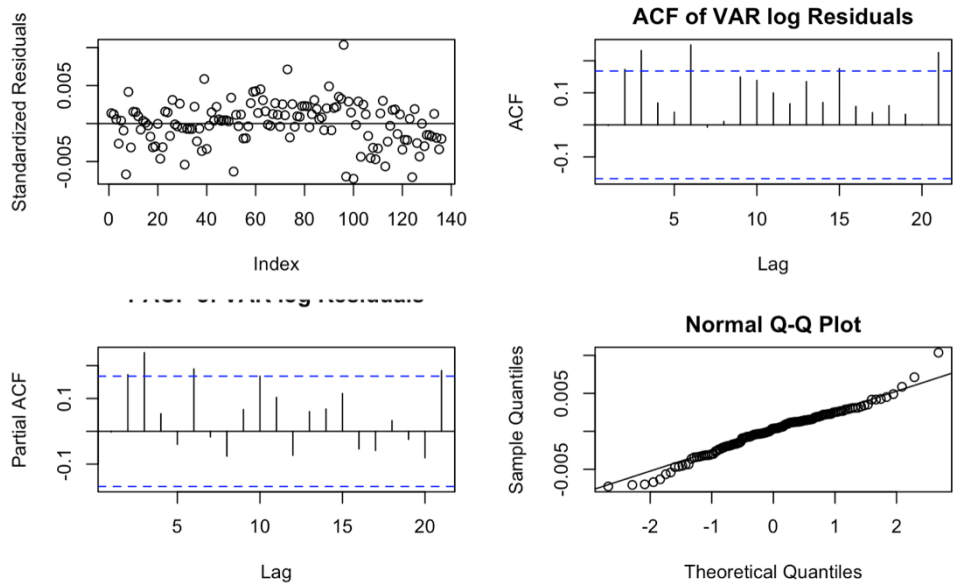


Figure 11. Residual analysis of unrestricted VAR trained on log transformed data

VARX

Like VAR models, we applied VARX models and their restricted counterpart to the original dataset. The only difference was that median sale price of house and housing permit were used as external variables with homeownership rate and initial jobless claim being as internal variables. An order of 2 was picked based on BIC.

ARIMAX

Another model considering the influence of external factors is ARIMAX. For the model using original data, we used original initial jobless claim, housing permit, and median housing price data as exogenous variables to build a ARIMAX (0, 1, 0) model. The p value for both Box-Pierce and Box-Ljung test were less than 0.001, suggesting that the residuals have serial correlation. The autocorrelation

could be observed in the ACF/PACF plot as well. However, the model successfully predicted the homeownership spike in the testing dataset. Though the amplitude is lower than the actual value, it is significant to forecast such fluctuation for stakeholders to make decisions. The limitation of this model is that we didn't use lagged external variables, therefore, we have to use predicted value of external factors for forecasting.

For the model using log transformed data, we used initial jobless claim as the external factor. The performance of ARIMAX (2, 1, 1) was not as good as the one on original data. Similarly, the residuals showed serial correlation.

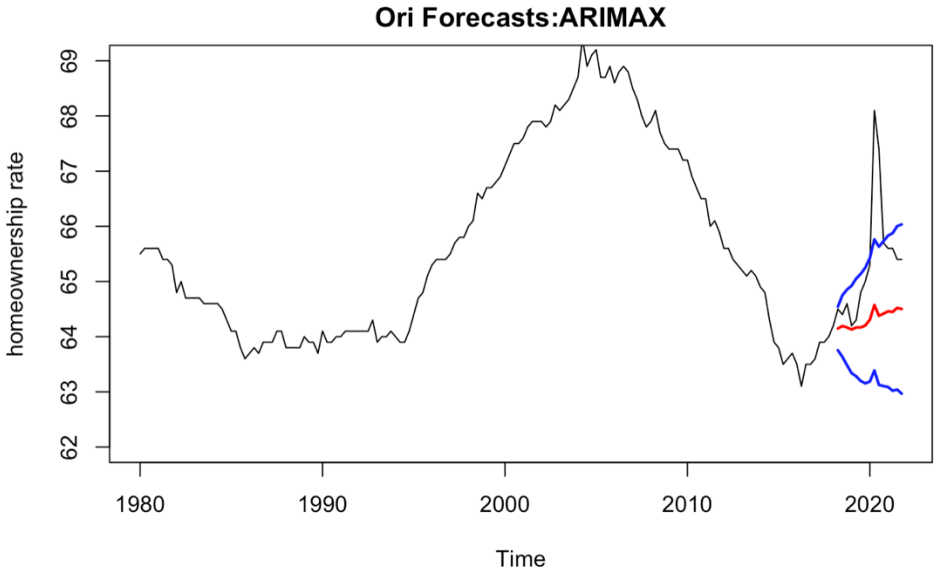


Figure 12. Forecasting and 95% confidence intervals of ARIMAX (0, 1, 0) on testing dataset

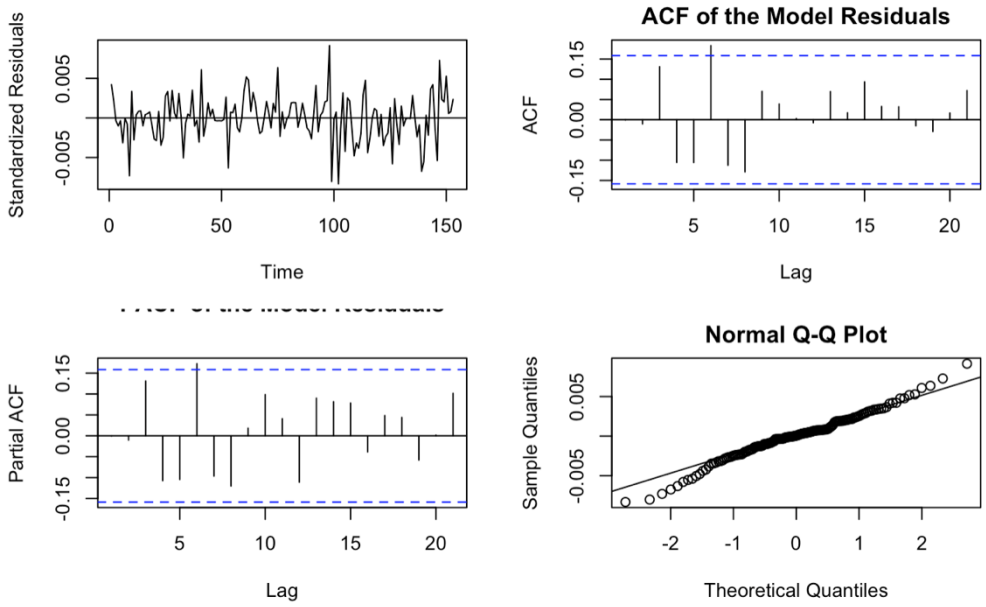


Figure 13. Visualization of residuals of ARIMAX (0, 1, 0)

Conclusion

In our study, the time series analysis helped us understand the properties of homeownership rate in the past 40 years. We built several models to forecast the rate. Among these models, the performance of VAR model on the log-transformed data was better than other models. However, ARIMAX model on the original data is the only model capturing the spike in the forecast. Such a model would help stakeholders make decisions responding to market fluctuations. Thus, we will choose ARIMAX to perform the time series forecasting. In the future, it is possible to ensemble the VAR model and ARIMAX model to optimize the performance of forecasting. Also, as a note for the use of the models, for ARIMAX perform forecasting, prediction of external variables shall be conducted.

Reference

[1] Weinstock, L. R. (2022, May 3). Introduction to U.S. economy: Housing market. Congressional Research Service. Retrieved September 18, 2022, from https://sgp.fas.org/crs/misc/IF11327.pdf

[2] *QUARTERLY RESIDENTIAL VACANCIES AND HOMEOWNERSHIP, SECOND QUARTER 2022*. (n.d.). Retrieved September 17, 2022, from https://www.census.gov/housing/hvs/files/currenthvspress.pdf

[3] pvs_admin. (2021, June 28). *Top 11 important economic factors affecting housing market*. Builders in Calicut | Apartments for Sale in Kozhikode | Flats in Calicut. Retrieved September 17, 2022, from https://pvsbuilders.com/economic-factors-affecting-housing-market/

[4] Haurin , D. R., & Rosenthal , S. S. (2004, December). *The sustainability of homeownership: Factors affecting the duration of of Homeownership and Rental Spells*. U.S. Department of Housing and Urban Development. Retrieved September 18, 2022, from https://www.huduser.gov/Publications/pdf/homeownsustainability.pdf

[5] McCue, D. (2022, August 10). *Across the nation, rising prices and increased interest rates limit access to homeownership*. Across the Nation, Rising Prices and Increased Interest Rates Limit Access to Homeownership | Joint Center for Housing Studies. Retrieved September 17, 2022, from https://www.jchs.harvard.edu/blog/across-nation-rising-prices-and-increased-interest-rates-limit-access-homeownership