Financial Data Analysis

Qing Lou, Qianqian Huang, Luyao Gao, Chunlin Liu

Contents

1 Introduction

Real estate economics is gradually becoming an irreplaceable facet of the social topic in the metropolis. The captivate of the preemption of real estate is a double-edged sword that not only polarizes the rich and the poor but also contains the incentives of the middle-class people of having house possession. One preponderant element of this trading is the house price, while the much latent compounds lying behind it could be the mortgage rate of this particular kind of housing loan. We are considering the mortgage rate in that the central government executes the management control system of the monetary market through various policy tool, the mortgage rate occupies an important position. Common sense makes it possible that the influence of the mortgage rate on the housing price is undeniable, but we still curious about the range of this effect.

In our analysis of the relationship between the housing price index and mortgage rate, the Autoregressive (AR) model fits the data by the order 12. Although this recognizable 12 provides no evidence to quarrel with, the Autoregressive Integrated Moving Average (ARIMA) Model illustrates the connection by more explanatory variables, especially the one from the error term in AR model. Also, the General Autoregressive conditional heteroskedasticity (GARCH) model is used, but it does not yield the significance contribution on data fitting.

The analysis of the dataset of housing price index and mortgage rate provides a good estimation process to their relationship. They are negatively related, and we have given the linear model as the most efficient and simplest model.

2 Literature Review

In recent years, the housing price and the lending rate have gained many people's attention. Researchers study them and want to find the relationship between this two index. The brief definition of housing price index(HPI) is the price changes of residential housing. Lending rate, also called the mortgage rate, is the rate of interest charged on a mortgage. Mortgage rates are determined by the lender in most cases. Since the 1970s of last century, the relationship between housing price and mortgage rate has drawn many real estate researchers' attention in aboard. In a paper called MORTGAGE RATE BUYDOWNS: IMPLICATIONS FOR HOUSING PRICE INDEXES (VB Agarwal RA Phillips,Social Science Quarterly,1984 ,65 (3) :868-875) using FHA-VA transaction to show that the prices of buydown homes tend to be elevated as compared to other properties, and the housing price index are likely to be distorted by the influence of mortgage rate. In paper Do lower mortgage rates mean higher housing prices? (JAMES M. MCGIBANY* and FARROKH NOURZAD, Applied Economics,2004,36 (4) :305-313) analyzes the housing price index and mortgage rate in both long run and short run, compared with each other finds that in long run, they have a negative relationship, but in short run there is no influence from mortgage rate to housing price index. The previous research gives us a brief idea that we want to analyze the long run HPI and mortgage rate with real estate. Relationship between Hong Kong house prices and mortgage flows (RYC Tse, Journal of Property Finance,1996 ,7 (4) :54-63) indicates the price for housing units in Hong Kong

is distorted by mortgage constraint, any changes house prices will have a feedback on mortgage lending, and thus tend to iron out the housing demand. The data in Visually shows the mortgage rate and housing price in Canada. Home price in Toronto rose 5.4% in 2013 and the mortgage rate fallen to historical lows $< 3\%$. This data told us the negative relationship between HPI and Mortgage rate in Canada. However, in the US, one of the world biggest real estate trade market, the relationship between this two index may have a more remarkable result. So we decide to analyze HPI and mortgage rate in the US during recent 30 years using ARMA, ARIMA and GRACH model.

3 Data

3.1 Data source

The data we choose are US National Housing Price Index with a base of 100 in 1975 and 30-Year Conventional MortgageRate(DISCONTINUED). We download the two data from the website "FRED" which is free. Both of the two data are monthly recorded and not seasonally adjusted. We use monthly-based data mainly because two reason: the first one is that we want to make our sample big enough to assure the accuracy of our estimated results; the other one is that we want to find the monthly characteristic of our data by ARIMA model and that is also why we choose the data which is not seasonal adjusted. The time span of our sample data is during 1975.Jan. and 2016.Sep. which is the only period we could find that contains both the two types of data.

```
library(readxl)
data <- read_excel("data.xlsx")
head(data)
     v1 v2 v3    v4   v5
1 1975  1  1 25.25 9.43
2 1975  2  1 25.29 9.11
3 1975  3  1 25.36 8.90
4 1975  4  1 25.40 8.82
5 1975  5  1 25.48 8.91
6 1975  6  1 25.46 8.89
```

3.2 Data description

We download the data from FRED by Excel form. One is about US National Housing Price Index and the other one is 30-Year Conventional MortgageRate(DISCONTINUED). We combine the excel into one in order to make the process of importing data more easily.Therefore, there are 501 rows and 5 columns contained in the final data excel. The 501 rows represent the 501 different months we contained in our sample, from 1975.Jan to 2016.Sep. The first column represents YEAR, the second is MONTH and the third is DATE. The fourth column is the data of US National Housing Price Index Value and the last one is of 30-Year Conventional MortgageRate(DISCONTINUED), which we use v4 and v5 to stand for the two columns in our code. We did some basic analysis for the data as followed:

```
hpi=data$v4 % US National Housing Price Index Value
mort=data$V5 % 30-Year Conventional MortgageRate
summary(hpi)
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   25.25   55.72   81.66   97.76  144.00  184.60
summary(mort)
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.350   5.970   7.910   8.294  10.080  18.450
```

4 Model Specification

4.1 Autoregressive model

An autoregressive (AR) model describes the time-varying processes of some specific variables, and it focuses especially on the relationship among the localized data, always the data before time t. Intuitively speaking, given a time series data $\{x_i\}$, for each x_t, we believe that x_{t-1}, x_{t-2}, x_{t-3} ... have the predictive power that can somewhat explain main portion of x_t as an linear combination. Also, AR model contains a white noise series or a stochastic term that can be imperfectly predicted. And with the time variant assumption, x_t has a statistically significant lag-n autocorrelation, the lagged value x_{t-1} might be the most powerful explanatory variable. In the model specification, R program helps us with the order that has the significant influence on our AR regression.

```
> m1=ar(dhpi,method="mle")
> m1$order
[1] 12
```

And AR(12) is used in our model:

$$(1 - \phi_1 B + \phi_2 B^2 + \cdots + \phi_{12} B^{12}) x_t = a_t. \tag{1}$$

4.2 Autoregressive Integrated Moving Average model

An autoregressive integrated moving average (ARIMA) model provide a parsimonious description of the stochastic process. The AR part is regressed on its own lagged, the (moving average) MA part is a linear combination of lagged error term, the I part illustrates the difference of the value and their previous value. One leading transmutation of ARIMA cut down the burdensome variables used in AR model and MA models, but it still adequate to simulate the data with dynamic structure in our analysis. With the spurn of high order model, some insignificant terms can be removed, the number of parameters used is sharply reduced, and the structure of the estimating model becomes parsimony. Also, as the model comprise three frames, these features should fit the data as well as possible, and we are expecting that the ARIMA model would afford a better interpretation for the interconnection of housing price index.

```
> tsdhpi=ts(dhpi)
> auto.arima(tsdhpi)
Series: tsdhpi
ARIMA(5,0,4) with non-zero mean
sigma^2 estimated as 0.06933:  log likelihood=-39.09
AIC=100.18   AICc=100.72   BIC=146.54
> m4=arima(dhpi,order = c(5,1,4))
> m4
Call:
arima(x = dhpi, order = c(5, 1, 4))
sigma^2 estimated as 0.04959:  log likelihood = 39.46,
aic = -58.93
```

This time we are using the ARIMA(5,1,4) because the aic is smaller:

$$\begin{aligned}(1-\phi_1 B+\phi_2 B^2-\phi_3 B^3-\phi_4 B^4+\phi_5 B^5)x_t \\ = a_t-\theta_1 a_{t-1}+\theta_2 a_{t-2}-\theta_3 a_{t-3}+\theta_4 a_{t-4}.\end{aligned} \tag{2}$$

4.3 General Autoregressive conditional heteroskedasticity model

First, if an autoregressive moving average model (ARMA) model is assumed for the error variance, the model is a generalized autoregressive conditional heteroscedasticity (GARCH) model. Second, if we look at the basic autoregressive conditional heteroskedasticity (ARCH) model, the variance of the current term is a function of the previous time series' error term. The assumption of the time series data is serially uncorrelated but they are dependent, this time variant data is split into a stochastic piece which is a strong white noise and a time-dependent standard deviation. And in the GARCH (p, q) model, p indicates the order of σ^2, q indicates the order of the stochastic series. One thing needs to mention is that the model is highly relevant in volatility modeling, which means the GARCH model is the best method to test the volatility, for now. The tail distribution of GARCH is heavier than normal, but it still provides a simple parametric function that can be used to describe the volatility evolution.

5 Empirical Results

In this section, we apply the proposed estimation method to the housing price index and mortgage rate to demonstrate its fitness. In the empirical, we target at investigating the following question. (1) How effective is the proposed method? (2) How consilient does the representing formulation describe the data relationship?

First, let's take a look at the inter-correlation within the housing price index. The left picture of figure 1 provides a strong evidence for the trending of this time variant data, we can mainly predicate that this is a unit root process, and it is also stationary. The right picture of figure 1 is the ACF after we have made one difference on the data, it is so clearly that the differentiated data has seasonal adjusting and about 12 months as a cycle. That makes sense to us because this 12 months' cycle reflects indirectly the varying of one year's preference for the housing.

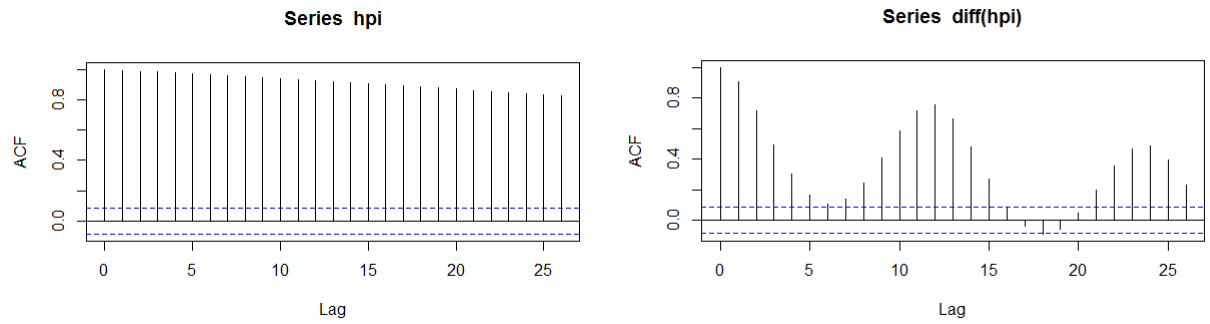

Figure 1: Autocorrelation Function of housing price index and data with differentiate once, from 1975. Jan to 2016. Sep.

Then we have the sense that we can try the AR(12) model with differentiating for once. From the code, we see 11 of the 12 coefficient is significant but does this provides enough evidence to say that this model is a good fit for our data? We use the 'tsdiag' command in the RStudio (see the results in the graph), and it gives much information so that we can have the AR(12) model as:

$$(1 - 1.234B + 0.283B^2 + \cdots + 0.233B^{12})x_t = a_t. \tag{3}$$

```
> m2=arima(hpi,order=c(12,1,0))
> m2
Call:
arima(x = hpi, order = c(12,1,0))
Coefficients:
          ar1      ar2      ar3     ar4      ar5      ar6     ar7
       1.2343  -0.2827  -0.3300  0.3316  -0.0762  -0.1727  0.1563
s.e.   0.0434   0.0687   0.0696  0.0710   0.0725   0.0725  0.0724
          ar8      ar9     ar10    ar11     ar12
       0.0020  -0.0972   0.1159  0.3202  -0.2325
s.e.   0.0729   0.0713   0.0698  0.0689   0.0436
sigma^2 estimated as 0.04963:  log likelihood = 38.27,  aic = -50.54
```

We also tried the MA(12) model, but unfortunately, this MA(12) model has $aic = 110.58$ and $loglikelihood = -42.29$. We do not think that it is a good model for predicting this housing price index dataset, and we do not give the representation of it here.

After the AR and MA model, we think the combination of these two will also make sense, so we tried the ARIMA model. Before we do the simulation, we transform the data into a time series one by using the 'ts' command in RStudio. The 'auto.arima' command gives us the automatic simulation of these housing price index data on the ARIMA model, and it provides the ARIMA(5,0,4) for the data which has been differentiating for once. That is equivalent to the result that we would have an ARIMA(5,1,4) for the original data, the data without differentiating. And the model is:

$$\begin{aligned}(1 + 0.220B + 0.05B^2 - 0.726B^3 - 0.004B^4 + 0.163B^5)x_t \\ = a_t + 1.728a_{t-1} + 2.145a_{t-2} + 1.184a_{t-3} + 0.429a_{t-4}.\end{aligned} \tag{4}$$

```
> m4=arima(hpi,order = c(5,1,4))
> m4
Call: arima(x = hpi, order = c(5, 1, 4))
Coefficients:
          ar1      ar2     ar3     ar4      ar5     ma1     ma2     ma3
      -0.2191  -0.0476  0.7262  0.0043  -0.1634  1.7278  2.1454  1.1841
s.e.   0.1095   0.0829  0.0678  0.0886   0.0760  0.1066  0.1708  0.1561
          ma4
       0.4289
s.e.   0.0874
sigma^2 estimated as 0.0693:  log likelihood = -44.05,  aic = 108.09
> Box.test(m4$residuals)
        Box-Pierce test
data:  m4$residuals  X-squared = 0.34351, df = 1, p-value = 0.5578
> pp=1-pchisq(0.34,12)
> pp
[1] 1
```

We also have a basic analysis of the mortgage rate, see Figure 2. The ACF shows a strong serial correlation, and the PACF only significant on 1,2,3. And we can conclude that the mortgage rate is almost a stationary data.

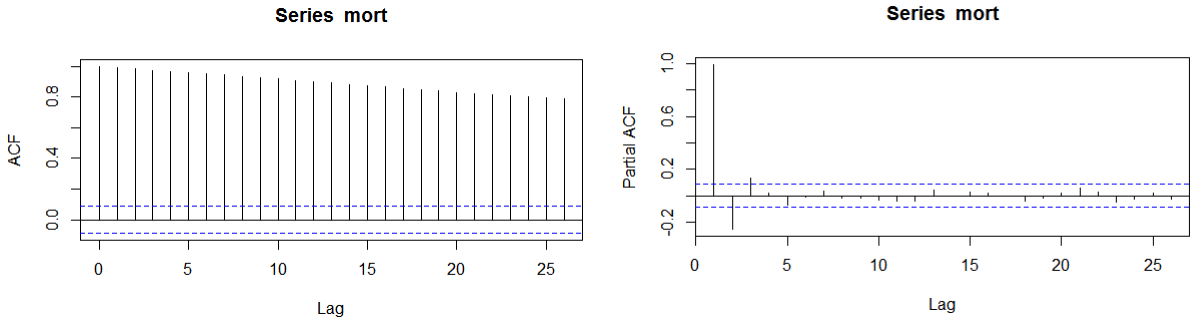

Figure 2: ACF and PACF of mortgage rate fitted from 1975. Jan to 2016. Sep.

The GARCH model is also used, we think it's a good estimation:

$$a_t = \sigma_t \varepsilon_t. \qquad \sigma_t^2 = 0.126 + 1.168 a_{t-1}^2 - 0.356 a_{t-2}^2. \tag{5}$$

```
m7=archTest(dhpi,2)
> m7
Call:
lm(formula = atsq ~ x)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12633    0.03761   3.359 0.000843 ***
x1           1.16821    0.04200  27.815  < 2e-16 ***
x2          -0.35628    0.04200  -8.483 2.55e-16 ***
Residual standard error: 0.7709 on 495 degrees of freedom
Multiple R-squared:  0.7746,    Adjusted R-squared:  0.7737
F-statistic: 850.7 on 2 and 495 DF,  p-value: < 2.2e-16
```

Here comes the most focused part of our analysis on the relationship between the housing price index and mortgage rate. By using the 'plot' command in RStudio, we have an intuitively negative relationship, see the following picture.

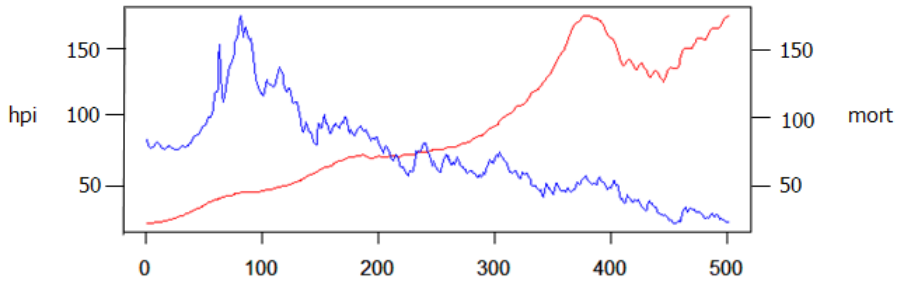

Figure 3: Relationship of housing price index and the mortgage rate from 1975. Jan to 2016. Sep.

And we are also wondering if we could just do it using a basic linear model. We think it provides bad estimation because the t-test of both the intercept and coefficient is significant at 99% confidence interval and the adjusted R-squared equals 0.603. And also because we see the steep change both in the mortgage rate in about 1980 and the housing price index in about 2008. But we here still give the function of this linear regression:

$$\begin{array}{rcc} hpi = & 193.19 & -11.48\ mort. \\ & 3.738 & 0.418 \end{array} \tag{6}$$

6 Discussion of Linear Model

In our paper, the basic topic is the relationship between housing price index and the mortgage rate, so we will have a deeper discussion in this section.

6.1 Reasons for this Bad Estimation

As we have mentioned before, we do not think that the linear model is a good estimation to relate housing price index with the mortgage rate. Because we see the steep change both in the mortgage rate in about 1980s and the housing price index in about 2008, it is reasonable to think about the expectations. Would it be the residual that made that bad estimation? Or would here be some special cases in this time period?

First, consider the residuals' part. The ACF of our linear residuals shows a stationary trend in this relationship between housing price index and mortgage rate. The PACF of our linear residuals in the right graph in Figure 4 shows significance in the first two lags. We use this information to further discuss the residuals in our linear regression model.

The residuals form an AR(2) model with housing price index on the left side of the equation, while mortgage rate on the right. The equation is:

$$(1 - 1.473B + 0.601B^2)hpi = 0.075mort. \tag{7}$$

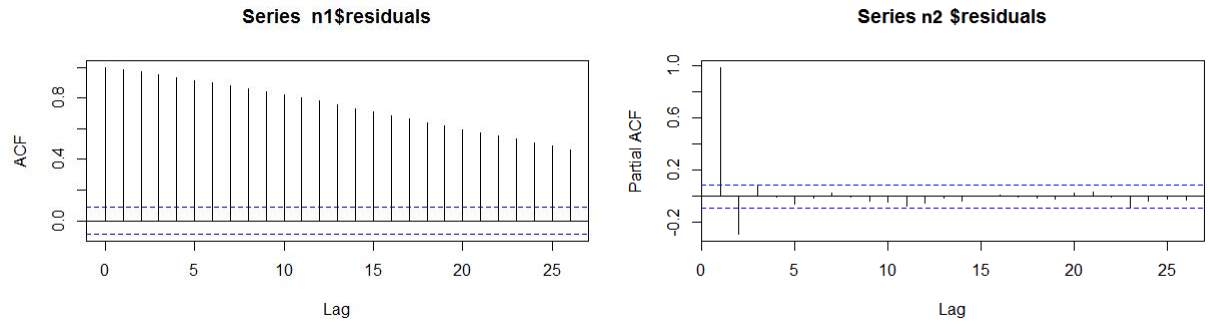

Figure 4: ACF and PACF of n1 residuals fitted from 1975. Jan to 2016. Sep.

```
> m5=arima(hpi,order=c(2,0,0),include.mean=F,xreg=mort)
> m5
Call:
arima(x = hpi, order = c(2, 0, 0), xreg = mort, include.mean = F)
Coefficients:
          ar1      ar2    mort
       1.4729  -0.6005  0.0748
s.e.   0.0356   0.0356  0.0340
sigma^2 estimated as 0.07448:  log likelihood = -61.54,  aic = 131.08
```

But it's not a good explanation, the aic is big and the t-test is significant.

Second, consider the information part. All of us think this part makes more sense to the reason that the steep change in both datasets. The financial crisis in about 1980 and 2008 obviously make the change in the housing price index and the mortgage rate.

But why the financial crisis in about 1980 have such huge impact on the mortgage rate, but almost no change in housing price? It has some issue rooted in history, like the baby boomers. Baby boomers in the United States, those born in the post-war generation over the 19 year period between 1946 and 1965, has become to the age of marriage. And that has the significant increase in the demand of the house because we know that most of them moved out from their parents' house after marriage. We do not know the most exactly why the housing price does not change that much, but we can guess that it is for some reality reasons.

And if we look at the financial crisis in 2008, it provides enough information for the suddenly climb. In 2003, the federal fund rate is about 1.01%, and after considering the inflation, it was actually a negative interest. This breaks the balance of the prudent savings and speculative risk, and it stimulates the opportunist in housing markets. That drives the housing price up in a very short period. We see the final results of this property bubble, it's the 2008 Financial Crisis. And that makes our regression model more reasonable for it does not fit the data that much.

6.2 Trend of the Difference

In this section, we talk about the variance in the dataset. To be specific, we are analysis the relationship between the $hpi_n - phi_{n-1}$and $mort_n - nort_{n-1}$. We are curious that the difference in the dataset will following which trend? Will they have the positive relationship? The code provides the answer:

$$\begin{array}{rcc} dhpi = & 0.322 & -0.275\ dmort \\ & 0.037 & 0.126 \end{array} \tag{8}$$

```
> y=dhpi[2:500]; x=dmort[1:499];
> n2=lm(y~x)
> n2
lm(formula = y ~ x)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.32190    0.03661   8.794   <2e-16 ***
x            0.27533    0.12558   2.193   0.0288 *
```

Now we can conclude that the t-test of the coefficient is significant at 95% confidence interval, and the intercept is significant on 99% confidence interval. As we just did to the linear regression for the non-differentiated term, we take a look at the linear residuals. The ACF provides the evidence that there exists a seasonal trend in the residuals, that may explain the model.

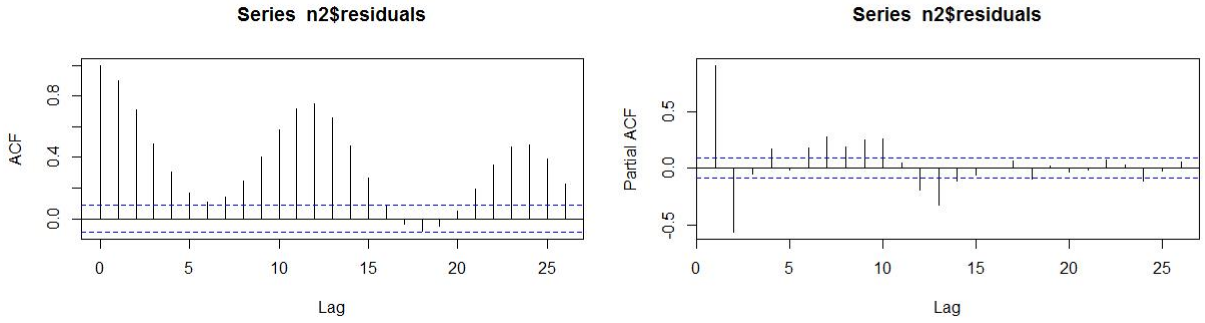

Figure 5: ACF and PACF of n2 residuals fitted from 1975. Jan to 2016. Sep.

6.3 New regression for data without exception

7 Conclusion