STAT 6950 Final Project: Modeling the relationship between housing price and Economic factors

Chenze Li and Pin-Hsun Mao

1 Introduction

House price is an essential economic index for people and local government. Its fluctuation reflects the up and down of the local economy and produces potential and general influence on people's daily life. Therefore, it is important to investigate and analyze the trend of house prices and then react to changes it may bring.

In this project, we obtained a data set from Harvard Dataverse including house prices in the United States and other related variables on a monthly scale. The data set consists of 14 house-price-related variables with 372 monthly observations from January the first 1990 to October the first 2020. In this data set, house prices in the United States are represented as HPI. HPI, proposed by three economists in 1980, measures house prices in the United States on a monthly basis.

From this analysis, we hope to grasp the trend of house prices and based on what we learn, we can make reasonable predictions on house prices.

2 Exploratory Data Analysis

The data set contains three parts. The first part contains 14 variables and 372 observations from 2000 to 2020 on a monthly scale. The second and third parts contain observations on a quarterly and yearly scale respectively. To obtain a model with accurate predictions, we only consider the first data set and the explanations of 14 variables are included in the following:

- **UNEMP** Unemployment Rate

- **CONST** Value of Private Residential Construction Put in Place excluding rental, vacant, and seasonal residential improvements - Seasonally Adjusted Annual Rate
- **Months of Supply** Months' Supply at Current Sales Rate
- **Mortgage Rate** 30-Year Fixed Rate Mortgage Average
- **Permits-Number** Total number of housing units of permits, in thousands. New Privately Owned Housing Units Authorized by Building Permits in Permit-Issuing Places
- **Permits-Valuation** Valuation of the number of permits, in millions of dollars. New Privately Owned Housing Units Authorized by Building Permits in Permit-Issuing Places
- **Housing Starts** Number of housing units (total) in thousands, seasonally adjusted
- **Disposable Income** Real Disposable Personal Income, Billions of Chained 2012 Dollars, Seasonally Adjusted Annual Rate
- **Consumption** Personal Consumption Expenditures, Billions of Dollars, Seasonally Adjusted Annual Rate
- **Savings** Personal Saving, Billions of Dollars, Seasonally Adjusted Annual Rate
- **Homes For Sale** Houses For Sale, Number of housing units in thousands
- **Homes Sold** Houses Sold during period, Seasonally Adjusted
- **HPI** SP/Case-Shiller U.S. National Home Price Index (Index Jan 2000 = 100), Seasonally Adjusted

We first investigate the response variable HPI. From Figure 3, we can observe that the median of HPI is around 160, and the variance is not very large. From Figure 2, we can observe that there are two linear trends, specifically, before 2006 and after 2014. There is a nonlinear trend between 2006 and 2014. We then use Pearson correlation coefficient and using heatmap to do the visualization to understand if there is any linearly correlated relationships between the response and the predictors and if there is any potential collinearity among the predictors. Figure 1 shows that Disposable Income and Consumption are positively highly correlated with HPI. For the predictors, Housing Starts and Permits Number

are positively highly correlated with Homes Sold. Permits valuation is positively highly correlated with construction spending. And mortgage rate is negatively highly correlated with consumption and Disposable income. We further investigate the highly correlated variables using scatter plots. Figure 8 and Figure 9 show that the nonlinear relationships. Figure 10 and Figure 11 show that there are potential collinearity between Permits Number and Housing Starts, and between Permits Valuation and CONST. As our data is time series data, we use monthly plot to inestigate whether there is seasonal trend over the years. Figure 12 show that the pattern of each month over the years seems to be similar, which indicates that there is no seasonal trend within our data set.

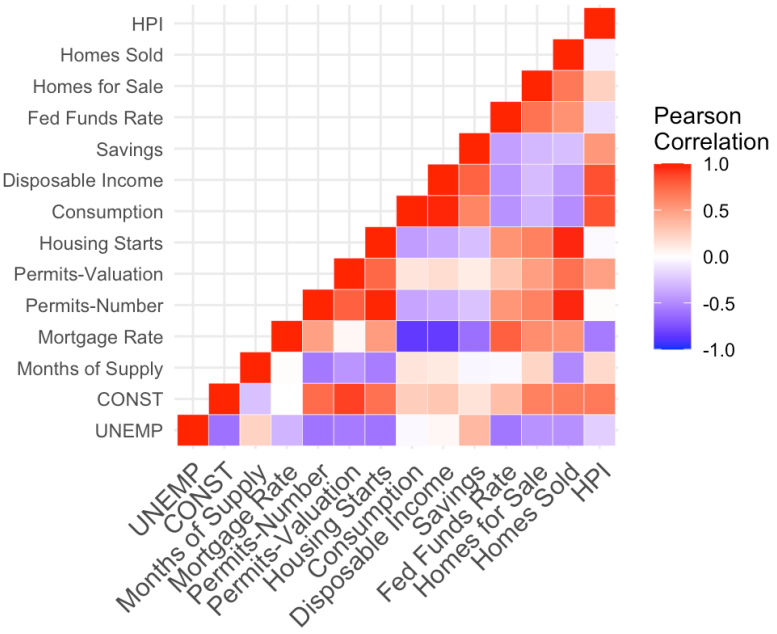


Figure 1

3 Method and Model Building

The response variable HPI is a time series variable. We need to consider a linear model with auto-correlative errors. The data set is divided into a training set and a test set. The training set contains the first 200 observations and the test set contains the left observations. At first, a linear model is considered and during the process of fitting, the transformation

of the response variable is considered which is $HPI_t^{\frac{3}{2}}$. And to reduce the complexity of the model, we use backward stepwise procedure with Bayesian information criterion(BIC) as criteria to select significant predictors. BIC is defined as BIC $= -2 \times l + \log(N) * k$ with l as the log-likelihood of the model, N as the size of training set and k as the number of parameters. The resulting model is stated as:

$$\begin{aligned}\text{HPI}_t^{3/2} = &-1385.58 + 30.92 \times \text{UNEMP}_t + 0.00165 \times \text{CONST}_t - 0.1142 \times \text{Housing.Start}_t \\ &+ 0.21126 \times \text{Consumption}_t - 0.1027 \times \text{Savings}_t + 1.843 \times \text{Homes.for.sale}_t + \epsilon_t\end{aligned}$$

where $\{\epsilon\}_t \sim N(0, \Sigma)$. And from the residual plot, it shows some patterns which implies the error term ϵ_t may follow auto-correlation. Before using ACF and PACF plots, we used Augmented Dickey-Fuller Test to ensure the residuals are stationary. From the ACF and PACF plots, it is better to fit the model with autoregression errors. The $AR(p)$ for residuals are:

$$\epsilon_t = \sum_i \rho_i \epsilon_{t-i} + \omega_t$$

where $\{\omega_t\}_t \overset{\text{i.i.d.}}{\sim} N\left(0, \sigma^2\right)$. And we use Akaike information criterion(AIC) to select optimal p in the $AR(p)$ models and get $p = 2$. AIC $= 2k - 2\ln(\hat{L})$ with $\hat{L}$ as the maximum of likelihood function of the model and k as the number of estimated parameters. Therefore, the $AR(2)$ model for residuals is:

$$\epsilon_t = 0.7187 \times \epsilon_{t-1} + 0.1015 \times \epsilon_{t-2} + \omega_t$$

The linear model with AR(2) errors grasps the whole trend of house prices. However, from Figure 2, we observe two separate periods with increasing trends. And we are interested in investigating the reasons for the increase. Therefore, we built two respective models with data from two separate periods.

We use all covariates to fit a linear regression. From Figure 13 and Figure 14, we observe that there is autocorrelation, and the lag 1 has the largest value of partial autocorrelation function value, which indicates that using the first order autocorrelation structure might be appropriate.

From Figure 2, we observe that there are two linear trends by time. Therefore, we first

use the data with the period before 2006 and the data with the period after 2014 to build two linear regressions. Therefore, we fit two linear regression models with the first order autocorrelation structure, the first one is using the early period data, which is the data before January, 2016, and the second one is using the later period data, which is the data after January, 2014.

For the early period model, Figure 15 show that the response should be transformed by taking log. We then use the transformed response to refit the generalized linear model with the first order autocorrelation structure and delete the non-significant covariates until the model includes all significant covariates. Figure 16 shows that there is no linear trend of the residuals, and the estimated correlation coefficient of this model is around 0.49, which is not very large. This indicates that the autocorrelation is not very strong in the early period dataset. We then decided to use linear regression to fit the early period data. The finalized early period model is as the following.

$$\begin{aligned} log(\text{HPI}_\text{t}) =& 3.656 + (1.961e^{-7}) \times \text{CONST}_\text{t} - (1.025\text{e}^{-2}) \times \text{Mortgage}_\text{t} \\ &+ (2.785e^{-4}) \times \text{Consumption}_\text{t} - (9.747\text{e}^{-5}) \times \text{Disposable.Income}_\text{t} \\ &+ (1.429e^{-4}) \times \text{Savings}_\text{t} + \epsilon_\text{t}, \text{t} = 1, ..., 72, \end{aligned}$$

where $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$

For the later period, we first use all covariates to fit a generalized linear model with the first-order autocorrelation structure. Similar approach as we fit the early period data, we delete the non-significant covariates until all covariates are significant in our model. The following is our finalized model for later period data.

$$\begin{aligned} \text{HPI}_\text{t} &= 170.02458 + 0.00004 \times \text{CONST}_\text{t} + 0.01367 \times \text{Consumption}_\text{t} \\ &- 0.01402 \times \text{Disposable.Income}_\text{t} + 0.0137 \times \text{Savings}_\text{t} + \epsilon_\text{t}, \text{t} = 1, ..., 82 \end{aligned}$$

and $\epsilon \sim N(0, \sigma^2 R)$, where R is the first order autocorrelation structure matrix as the following.

$$R = \begin{pmatrix} 1 & \hat{\rho} & \cdots & \hat{\rho}^{n-1} \\ \hat{\rho} & 1 & \cdots & \hat{\rho}^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}^{n-1} & \hat{\rho}^{n-2} & \cdots & 1 \end{pmatrix} \text{ with } \hat{\rho} = 0.99$$

4 Results and Discussion

Based on the analysis we have made, we can use the test set to evaluate our linear model with auto-regression errors. From Figure 19.(a) and Figure 19.(b), the assumptions of normality and equal variance are satisfied after box cox transformation. And from Figure 19.(f), predictions of expectation of HPI can be seen as the prediction of the trend for the house prices. And the predictions with auto-regressive errors are defined as:

$$\widehat{HPI}_t = \mathrm{E}[\widehat{HPI}_t] + \hat{e}_t$$

where $\mathrm{E}[\widehat{HPI}_t]$ is the prediction of expectation of HPI at time t and $\hat{e}_t$ means the predicted error ϵ_t with AR(2) model. The model collapses in some months in 2020 but fits well in other periods, implying there may be other factors that are not considered in the model and can affect the house prices.

Early period linear regression model shows that construction spending, mortgage, consumption, disposable income, and personal savings are associated with the log scale housing price from January, 2000 to December, 2006. Specifically, as construction spending, consumption, or personal savings, the log scale housing price will also increase. As mortgage, or disposable income increase, the log scale housing price will decrease. Later period generalized linear model with the first order autocorrelation structure shows that the housing price from January, 2014 to October, 2020 is associated with constructuion spending, consumption, disposable income, and personal savings. To be more specific, as construction spending, consumption, or personal savings increases, the housing price will increases. As disposable income increases, the housing price will decrease.

We conduct model diagnostics for the early period and the later period model. Figure 17 shows that the residuals seem to be centered around 0 without any specific patterns, and the spreads of the residuals seem to have equal variance. Therefore, the equal variance assumption seems to be valid, and it also shows that our model is correctly specified. The

normal Q-Q plot shows that the normality assumption of linear regression seems to be valid. Figure 18 shows that the residuals vs. fitted values are linear, and the estimated correlation coefficient is around 0.99, which indicates that the correlation is very high. The normal QQ plot shows that it is a little bite deviates from the normality assumption. Therefore, for the later period model, it may be helpful to use other correlation structures or transform some predictors.

Comparing the linear model with AR(2) errors and two models of two separate periods, we find Consumption, Savings, and CONST are included in all these three models which imply these three variables are the most important in predicting house prices.

References

[1] https://dataverse.harvard.edu/dataverse/HomePrice

[2] https://python-bloggers.com/2021/01/predicting-home-price-trends-based-on-economic-factors-with-python/

[3] https://socialsciences.mcmaster.ca/jfox/Books/Companion/appendices/Appendix-Timeseries-Regression.pdf

[4] Shumway, R. H., Stoffer, D. S., & Stoffer, D. S. (2000). Time series analysis and its applications (Vol. 3). New York: Springer.

[5] Brockwell, P. J., & Davis, R. A. (Eds.). (2002). Introduction to time series and forecasting. New York, NY: Springer New York.

5 Appendix

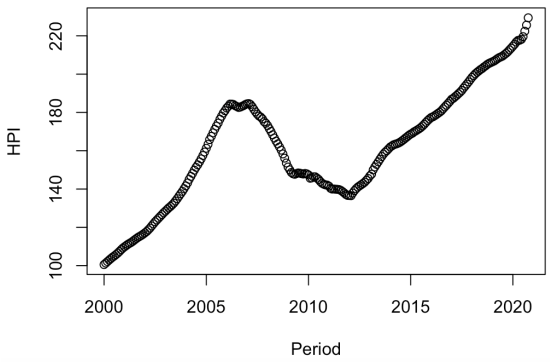


Figure 2

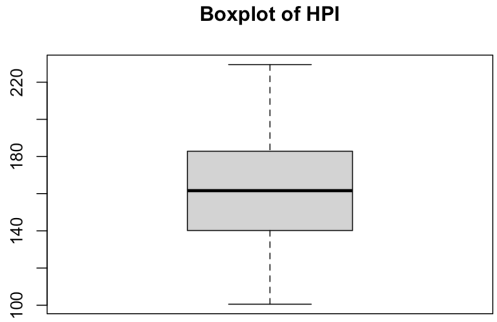


Figure 3

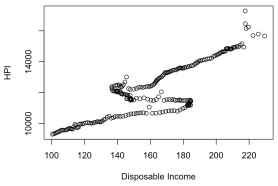


Figure 4

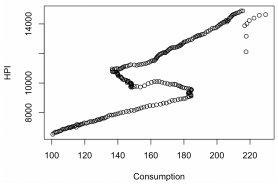


Figure 5

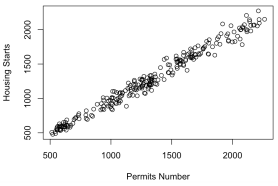


Figure 6

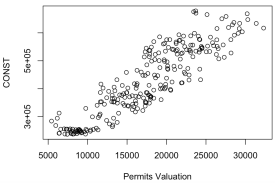


Figure 7

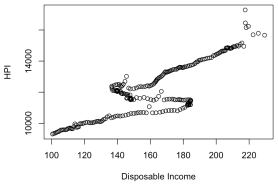


Figure 8

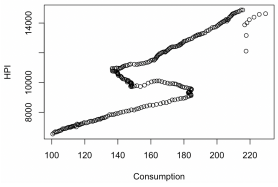


Figure 9

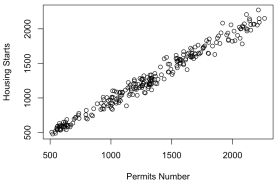


Figure 10

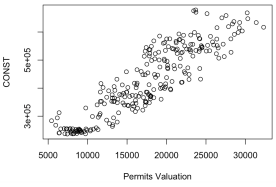


Figure 11

Exploratory Data Analysis: The boxplot and scatter plot of the response variable are shown. Also, the scatter plots contain highly-correlated variables or variables highly correlated with the response variable.

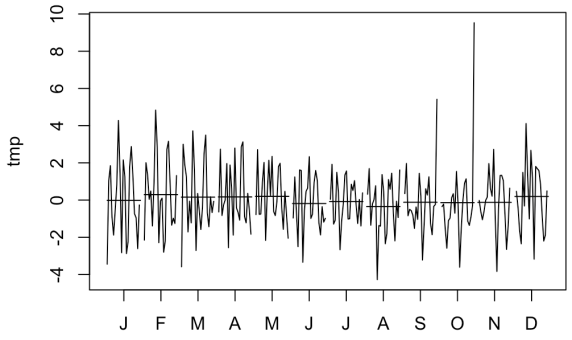


Figure 12

Month plot of the residuals of the linear regression model using all time periods.

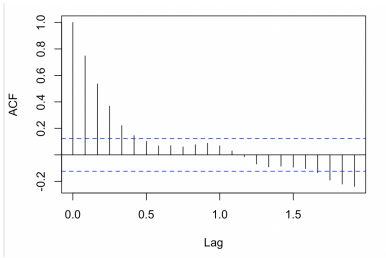


Figure 13

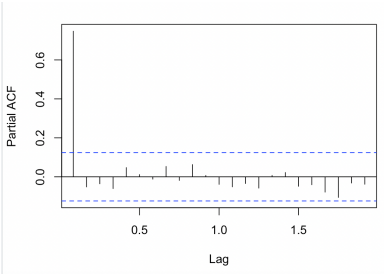


Figure 14

Model with all time period: The plots show the residuals of ACF and PACF for the linear regression model with all time period.

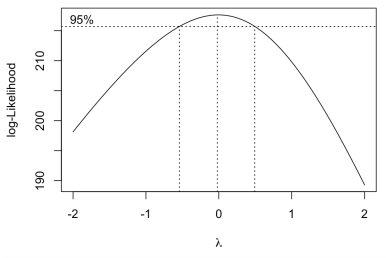


Figure 15

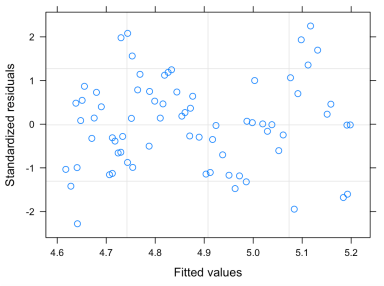
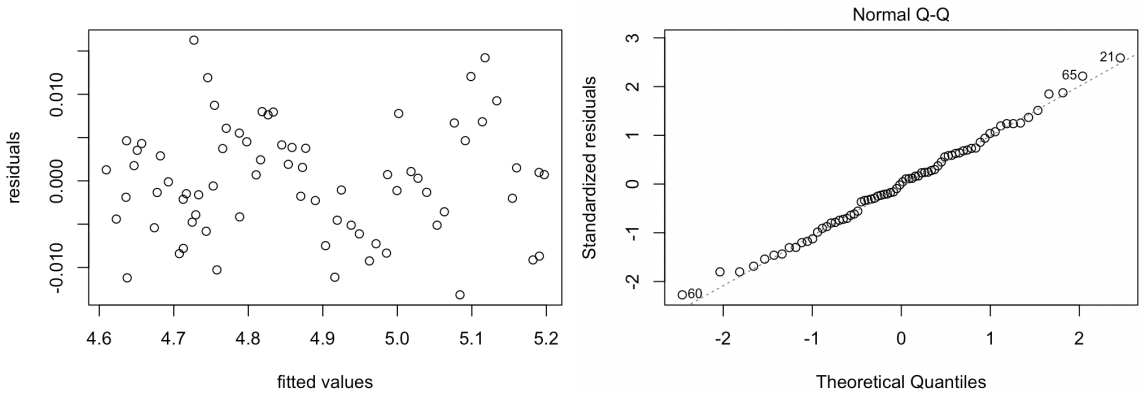


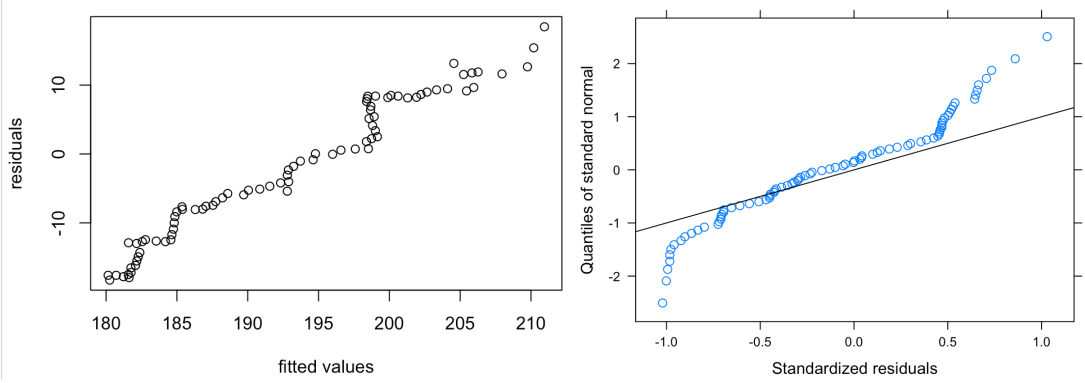
Figure 16

Early period model building plots.



((a)) fig 1 ((b)) fig 2

Figure 17: Diagnostic plots for the early period model



((a)) fig 1 ((b)) fig 2

Figure 18: Diagnostic plots for the later period model

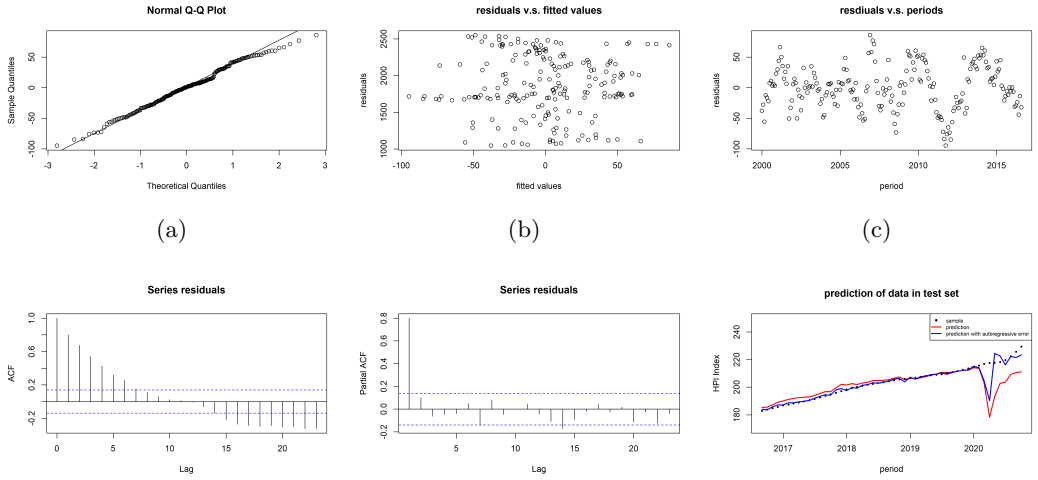


Figure 19: **Model with AR(2) errors**: The plots are the predictions and diagnostics of the model. The first line(from left to right): the Q-Q plot of residuals from the linear model; the residuals v.s. the fitted values; the residuals v.s. time periods; The second line(from left to right): the first two plots are ACF and PACF plots of the residuals. The last plot is the predictions with the test set.