Washington DC Housing Market Analysis

Nate Lu

Objective

Goals of This Study

- Housing characteristics
- Macroeconomic factors
- Predict DC condo prices

Key Terminology

Feature Selection

Time Series Regression

In-Time and Out-of-Time Modeling

Data Source: Redfin.com

Background

Understanding the Data

Housing Characteristics

Data Source: Redfin.com

Background

Understanding the Data

Macroeconomic factors

Data Source: fred.stlouisfed.org/

Model

Variable Transformation

- **Logarithmic Transform**
- **First Difference Transform**

Model

Feature Selection

- **Forward/Backward Stepwise Regression**
- **K-Best**

Model

Time Series Regression

$$y_t = \beta_0 + \beta_1 x_{t,1} + \beta_2 x_{t,2} + \cdots \beta_k x_{t,k} + u_t, \quad t = 1, 2, \ldots, T \text{ with error } u_t$$

- **Strong Markov Assumption**

- **Weak Markov Assumption**

Model

$$y_t = \beta_0 + \beta_1 x_{t,1} + \beta_2 x_{t,2} + \cdots \beta_k x_{t,k} + u_t, \quad t = 1, 2, \ldots, T \text{ with error } u_t$$

STRONG Gauss Markov Assumption:

- The explanatory variables $x_{.,j}$ are strictly exogenous with respect to the disturbance term. Hence, $E(U_t|\boldsymbol{X}) = 0, \forall\, t = 1, 2, \ldots, T$ where $\boldsymbol{X}$ is the matrix including all K regressors and all T time periods.
- No regressor is constant or can be expressed as a linear function of other regressors. That is, there exists no sets $A = \{a_0, a_1, \ldots, a_k\}$ where $a_j \neq 0,\ for\ some\ j\ \ni a_0 + a_1 x_{t,1} + a_2 x_{t,2} + \cdots a_k x_{t,k} = 0, \forall\, t = 1, 2, \ldots, T$. This implies $\boldsymbol{X}$ has full rank.
- Homoskedasticity: $\text{var}(U_t|\boldsymbol{X}) = \sigma^2, \forall\, t = 1, 2, \ldots, T$.
- No serial correlation: Conditional on X, the disturbance terms are uncorrelated. $\text{Cov}(U_t, U_{t-s}|\boldsymbol{X}) = 0, s = 1, 2, \ldots, T-1$.
- Normality: The disturbance terms are normally distributed. $U_t \sim N(0, \sigma^2)$

Model

$$y_t = \beta_0 + \beta_1 x_{t,1} + \beta_2 x_{t,2} + \cdots \beta_k x_{t,k} + u_t, \quad t = 1, 2, \dots, T \text{ with error } u_t$$

WEAK Gauss Markov Assumption:

- The variables of the model are stationary, ergodic and the explanatory variables $x_{\cdot,j}$ are exogenous with respect to the disturbance term. Hence, $E(U_t | x_{t,1}, x_{t,2}, \dots, x_{t,k}) = 0, \forall\, t = 1, 2, \dots, T.$

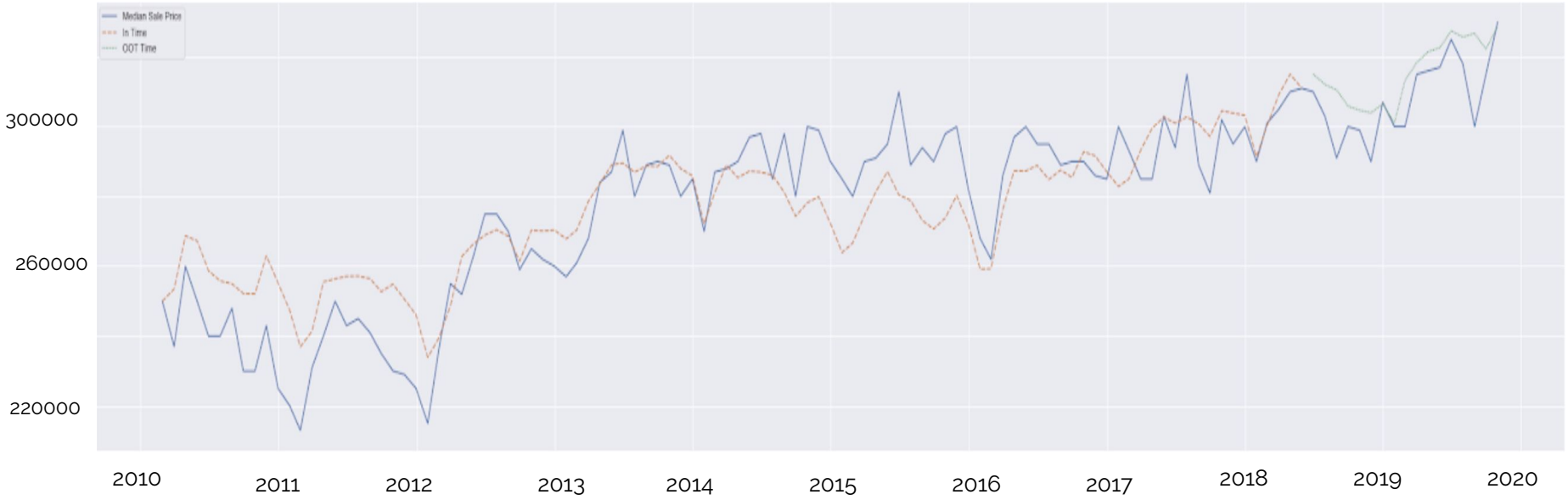
- No regressor is constant or can be expressed as a linear function of other regressors. That is, there exists no sets $A = \{a_0, a_1, \dots, a_k\}$ where $a_j \neq 0, \ for\ some\ j \ni a_0 + a_1 x_{t,1} + a_2 x_{t,2} + \cdots a_k x_{t,k} = 0, \forall\, t = 1, 2, \dots, T.$

- Homoskedasticity: $\text{var}(U_t | x_{t,1}, x_{t,2}, \dots, x_{t,k}) = \sigma^2, \forall\, t = 1, 2, \dots, T.$

- No serial correlation: Conditional on X, the disturbance terms are uncorrelated. $\text{Cov}(U_t, U_{t-s} | x_{t,1}, x_{t,2}, \dots, x_{t,k}) = 0, s = 1, 2, \dots, T-1.$

Results

Statistical Analysis



Results

Residual Analysis

	MAE	MAPE
In Time	$10591.33	4.02 %
OOT Time	$7882.99	2.61%

*
MAE: Mean Absolute Error
MAPE: Mean Absolute Percentage Error

What's Next

- Model Use -

We can use Washington DC housing market characteristics and macroeconomic factors to predict condo prices.

- Next Step -

We will enhance the predictive model by using Machine Learning techniques such as Random Forest, Support Vector Machines, and Natural Language Processing.

Thank You!

GitHub: https://github.com/iuniorhsiung/mod4_project_DC_housing_price

Appendix

Appendix

- Data Source: Redfin.com
- Data Source: FRED(Federal Reserve Economic Data)
- Results: Time Series Regression
- Results: Normality Check
- Results: Multicollinearity Check
- Model Approach: Regression Flow Chart

Data Source

Data Source: https://www.redfin.com/blog/data-center/

Redfin:

- Redfin is a real estate brokerage that has direct access to data from multiple listing services, as well as insight from real estate agents across the country.
- Redfin provides housing market data for all metropolitan areas, cities, neighborhoods, and zip codes across the nation.
- Redfin's housing market data includes data for prices (median sale price, percentage of homes sold above list price, percentage of homes that had price drop, etc.), inventory (number of homes on market, new listings, months of supply, etc.), and sales (number of homes sold, median days on market, etc.).
- Redfin's housing market data can be filtered by area, property type, month-over-month change, year-over-year change, and the time period.

Data Source

Data Source: https://fred.stlouisfed.org/

FRED:

- FRED is a website providing Federal Reserve Economic Data across the country.
- FRED provides over 500,000 financial and economic data series from more than 85 public and proprietary sources.
- FRED's economic data includes home price index, unemployment rate and all employment for different activities.
- Data used for model inputs:

  https://github.com/iuniorhsiung/mod4_project_DC_housing_price/blob/master/data/readme.md

Results

Statistical Analysis

Regression output

OLS Regression Results

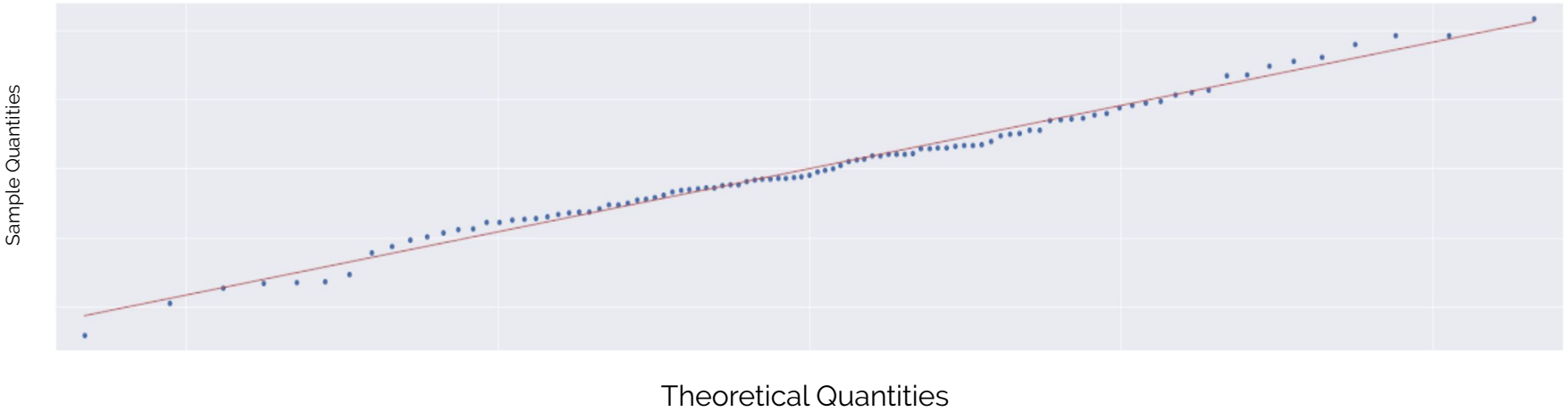
Dep. Variable:	Median Sale Price_1diff	R-squared:	0.337
Model:	OLS	Adj. R-squared:	0.309
Method:	Least Squares	F-statistic:	11.94
Date:	Mon, 17 Feb 2020	Prob (F-statistic):	6.88e-08
Time:	22:24:53	Log-Likelihood:	-1039.7
No. Observations:	99	AIC:	2089.
Df Residuals:	94	BIC:	2102.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	920.4555	1094.821	0.841	0.403	-1253.337	3094.248
Days on Market_1diff	-592.2738	119.719	-4.947	0.000	-829.978	-354.570
US_UR_1diff	1.727e+04	6483.813	2.664	0.009	4400.316	3.01e+04
New Listings MoM_1diff	-66.9649	25.346	-2.642	0.010	-117.289	-16.641
WDXRSA_1diff	1184.0909	1054.197	1.123	0.264	-909.042	3277.223

Omnibus:	0.432	Durbin-Watson:	2.801
Prob(Omnibus):	0.806	Jarque-Bera (JB):	0.119
Skew:	0.038	Prob(JB):	0.942
Kurtosis:	3.152	Cond. No.	265.

Results

Normality Check: Q-Q plot



Results

Multicollinearity

	Days on Market	US UR	New Listing MOM	DC HPI
VIF	1.0576	1.0605	1.0676	1.0504

Model Approach