# Housing prices as a benchmark for us economy

**By: Alan Zaragoza, Eric Reingardt, and Jordan Small**

# Creating an Econometric Model

STEP 1

# Equation

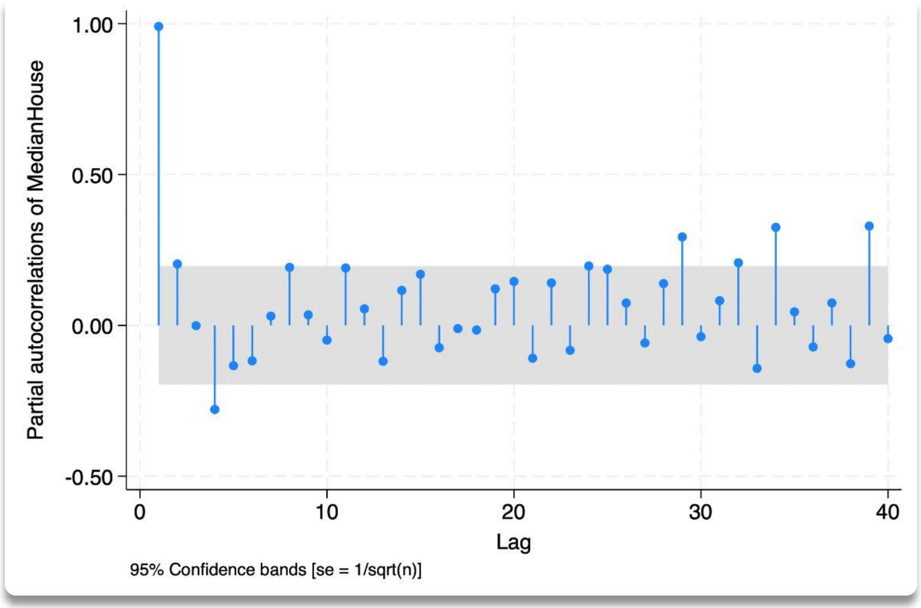
Land Price=$\beta 1+\beta 2$(GDP Growth)$+\beta 3$(Interest Rate)$+\beta 4$(Income)$+\beta 5$(Inflation)$+\beta 6$ (Unemployment)$+\varepsilon$

There is **one endogenous variable**, **Land Price**, and **five exogenous variables: GDP**, **Interest Rate**, **Real Income**, **Inflation**, and **Unemployment**. The model also includes **one error term** ($\varepsilon$).

An **endogenous variable** is a variable whose value is **determined by the model itself**.
In other words, it's a variable that is **explained by other variables** in the system.

An **exogenous variable** is a variable whose value is **determined outside the model** and is **not explained** by other variables in the system.
It is assumed to **influence the outcome**, but **not be influenced by it**.

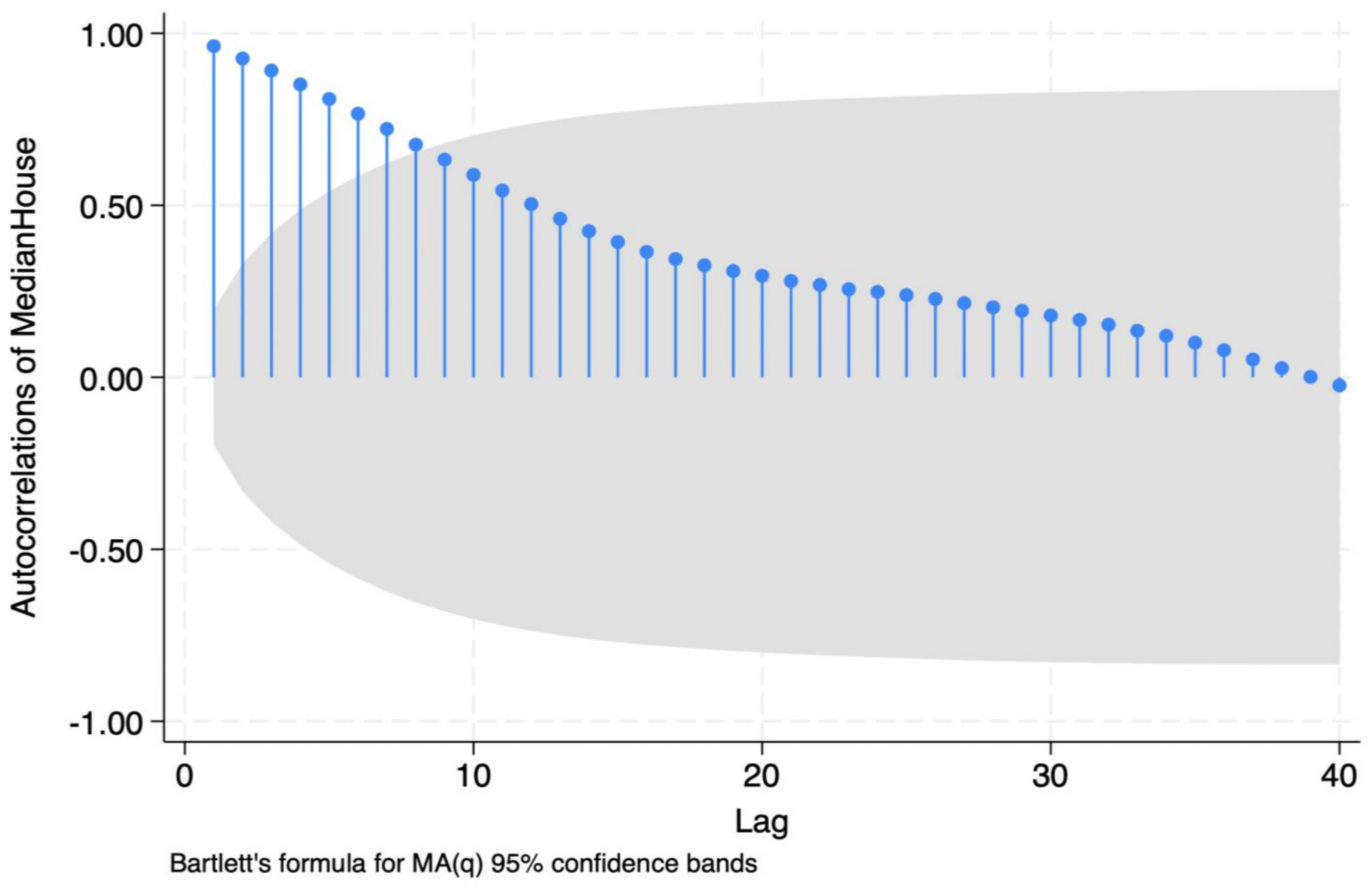
# Partial Autocorrelation




- Lag 1 has a very strong and significant spike
- This suggests a strong relationship between current house prices and prices from the previous period.
- Lags 2 and beyond fluctuate within the 95% confidence band, with no clearly significant spikes.
  - This tells us we can us an Auto recursive model with one or possibly two lags

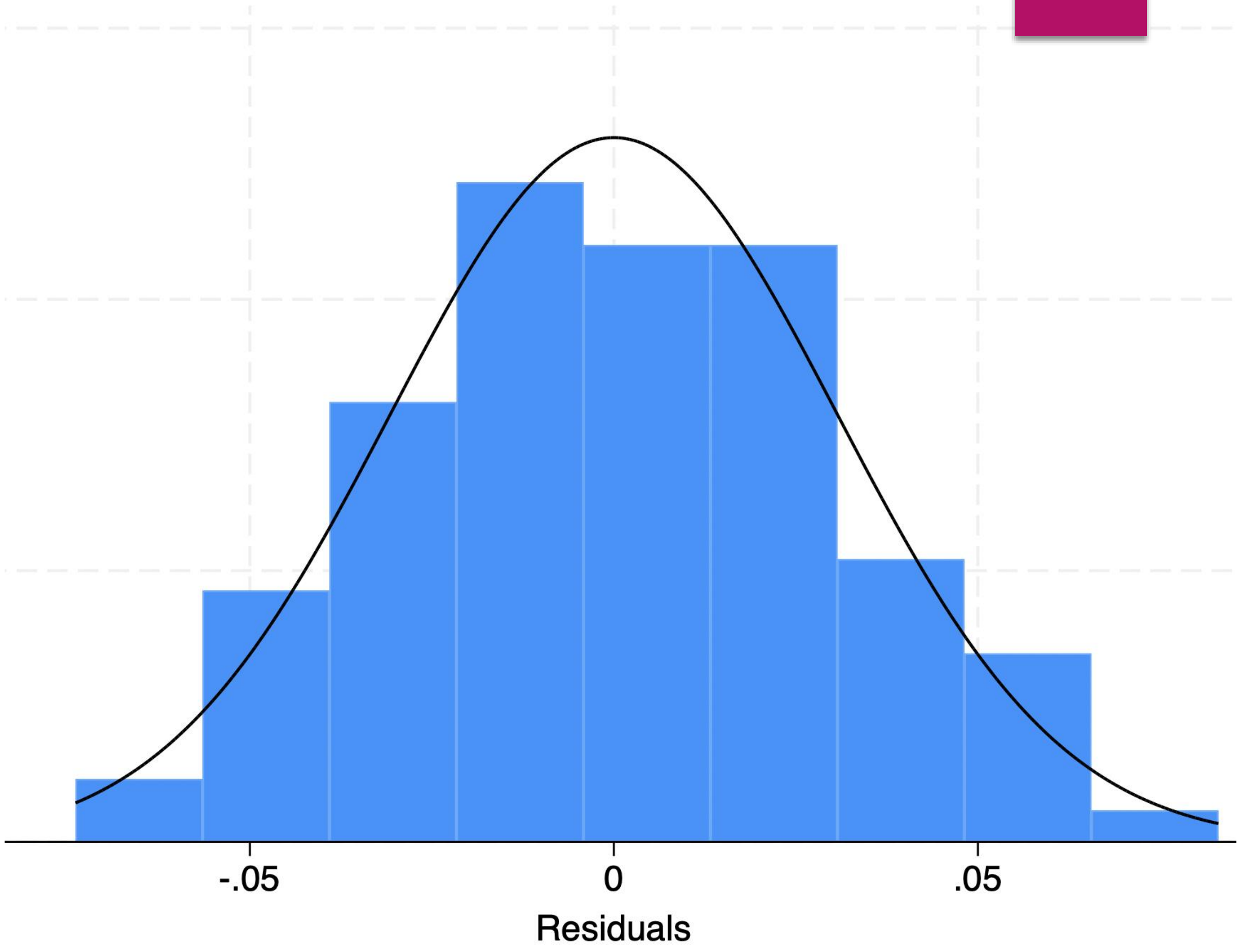
# Correlation Function

- Auto Regressive 1 Model
- **$y_t = \alpha + \beta_1 y_{t-1} + \varepsilon_t$**
- Auto regressive 2 model
- $y_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \varepsilon_t$



# Residuals Test

- Residual Distribution (Bell Curve)
- Residuals appear approximately normally distributed
- Supporting the model's assumptions of homoscedasticity and unbiasedness.



# Testing Goodness of Fit

- The regression on the right explains:
- Goodness of fit
- How statistically significant each variable is
- How well the model fits with historical data, which will be used for forecasting
- We observe an R-squared of 0.95, meaning the model fits very well with past data.
- Out of the four independent variables:
- Real GDP and Real Interest Rate are statistically significant
- Unemployment Rate is partially significant
- Disposable Income is not statistically significant

```
. reg MedianHousePrice RealGDP YRIR RealDispInc UnempRate

      Source |       SS           df       MS      Number of obs   =        96
-------------+----------------------------------   F(4, 91)        =    540.92
       Model |  5.2012e+11         4  1.3003e+11   Prob > F        =    0.0000
    Residual |  2.1875e+10        91   240387721   R-squared       =    0.9596
-------------+----------------------------------   Adj R-squared   =    0.9579
       Total |  5.4199e+11        95  5.7052e+09   Root MSE        =     15504

------------------------------------------------------------------------------
MedianHous~e | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     RealGDP |   27.36538   3.464804     7.90   0.000     20.48298    34.24779
        YRIR |   8195.817   2858.681     2.87   0.005     2517.397    13874.24
 RealDispInc |    1.66997   3.894857     0.43   0.669    -6.066685    9.406625
   UnempRate |  -2094.298   1244.944    -1.68   0.096    -4567.227     378.631
       _cons |  -236821.7   24851.31    -9.53   0.000    -286185.8   -187457.6
------------------------------------------------------------------------------
```

# Testing Goodness of Fit

- The regression on the right is a Dickey-Fuller Test, used to determine whether the data is stationary or non-stationary.
- The issue with the first regression is that it may produce a misleading R-squared if the data is non-stationary.
- That model assumes stationarity — meaning the data does not trend up or down over time.
- The Dickey-Fuller test reports an R-squared of 0.118, which is much lower than the 0.95 in the first regression.
- This lower R-squared is acceptable because the model is now correctly accounting for the lack of stationarity in the data.

| Source | SS | df | MS |
|---|---|---|---|
| Model | 943689313 | 4 | 235922328 |
| esidual | 7.0503e+09 | 90 | 78336333.9 |
| Total | 7.9940e+09 | 94 | 85042120.9 |

| | | |
|---|---|---|
| Number of obs | = | 95 |
| F(4, 90) | = | 3.01 |
| Prob > F | = | 0.0221 |
| R-squared | = | 0.1181 |
| Adj R-squared | = | 0.0789 |
| Root MSE | = | 8850.8 |

| x1 | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| x2 | 1.824283 | 8.208943 | 0.22 | 0.825 | -14.48421 | 18.13278 |
| x3 | 6183.485 | 3713.644 | 1.67 | 0.099 | -1194.317 | 13561.29 |
| x4 | 1.01554 | 3.121386 | 0.33 | 0.746 | -5.185638 | 7.216717 |
| x5 | -2068.787 | 1940.485 | -1.07 | 0.289 | -5923.899 | 1786.325 |
| _cons | 2397.111 | 1240.03 | 1.93 | 0.056 | -66.42584 | 4860.648 |

# Log Regression

- We see a higher r-squared due to the nature of the logged values
- A one percent increase change.
- Increase decrease interest rate by –6 percent

| Source | SS | df | MS |
|---|---|---|---|
| Model | 7.1444216 | 5 | 1.42888432 |
| Residual | .198591266 | 94 | .002112673 |
| Total | 7.34301286 | 99 | .074171847 |

| | | |
|---|---|---|
| Number of obs | = | 100 |
| F(5, 94) | = | 676.34 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.9730 |
| Adj R-squared | = | 0.9715 |
| Root MSE | = | .04596 |

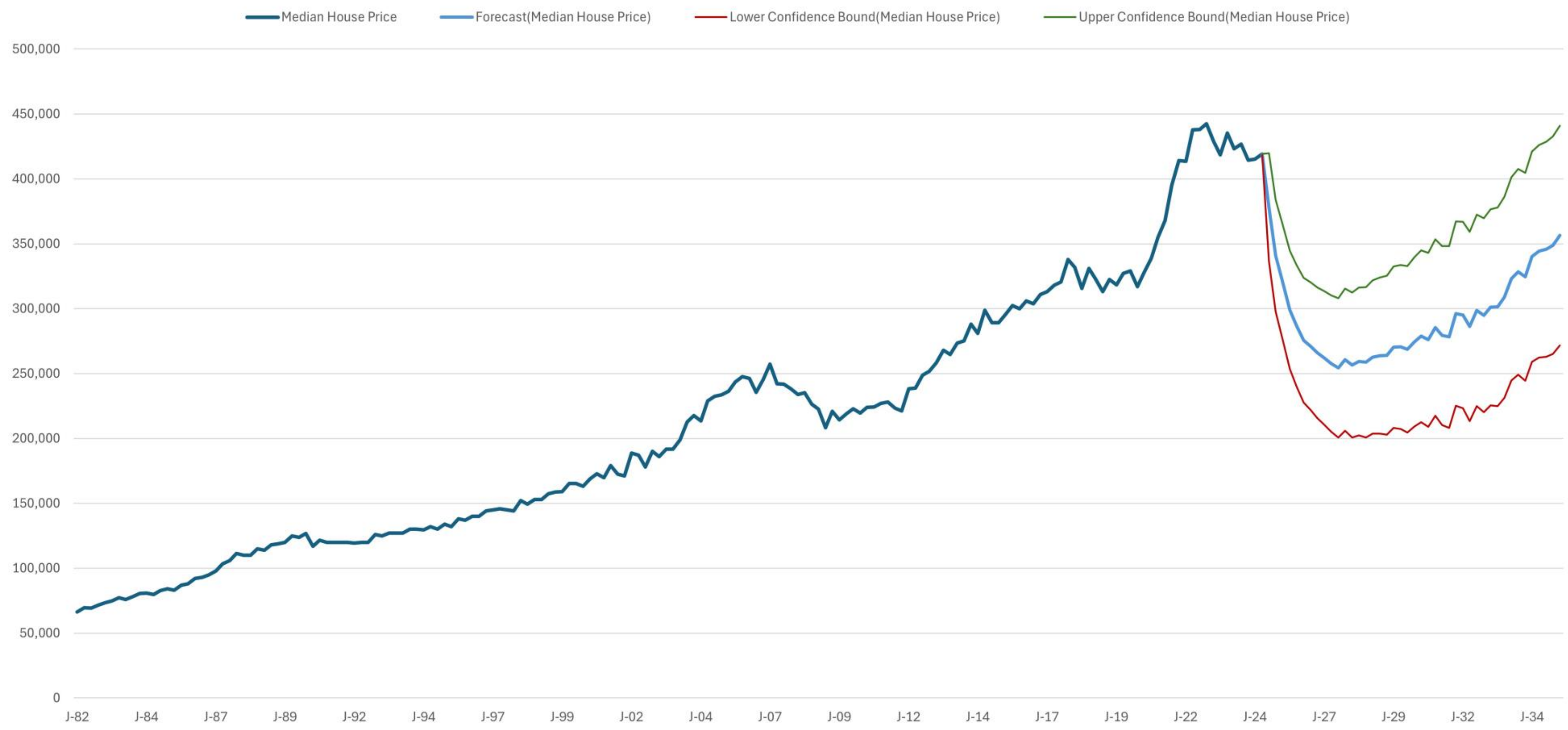
| MedianHouse | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| IR | -.0062989 | .0055655 | -1.13 | 0.261 | -.0173494 | .0047516 |
| Inflation | .1966653 | .0674938 | 2.91 | 0.004 | .0626548 | .3306759 |
| UNEMP | -.0178097 | .003252 | -5.48 | 0.000 | -.0242667 | -.0113528 |
| GDPln | .7153314 | .0363035 | 19.70 | 0.000 | .64325 | .7874129 |
| Dispincln | .0709509 | .0164111 | 4.32 | 0.000 | .0383664 | .1035355 |
| _cons | 4.881442 | .2666577 | 18.31 | 0.000 | 4.351987 | 5.410897 |

# Regression Continued

- While the original regression using level data yielded a high $R^2$ of 0.95, this was likely due to **shared trends** across non-stationary variables, leading to a **spurious relationship**.
- After applying the Dickey-Fuller test, We found that the variables were non-stationary, so we transformed the data using **log-differencing**.
  This led to a lower $R^2$ (~0.13), which is **normal** in models of changes rather than levels.
- The lower $R^2$ reflects the **reduction of artificial trend correlations** and a shift to modeling short-term fluctuations, which are more variable and harder to predict.
- Therefore, although the $R^2$ is lower, the model is **statistically more reliable**.

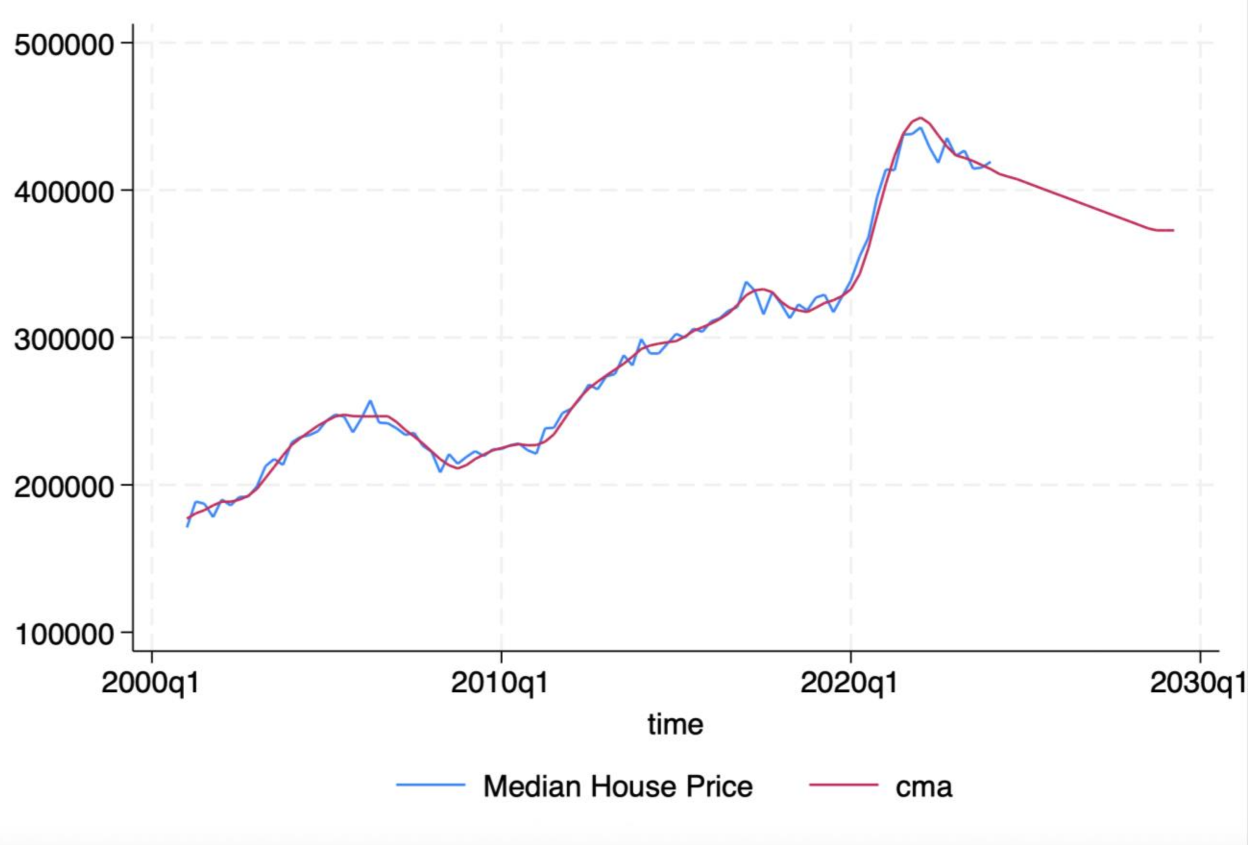

# Forecasting

# 95 Percent Confidence Interval Forecast



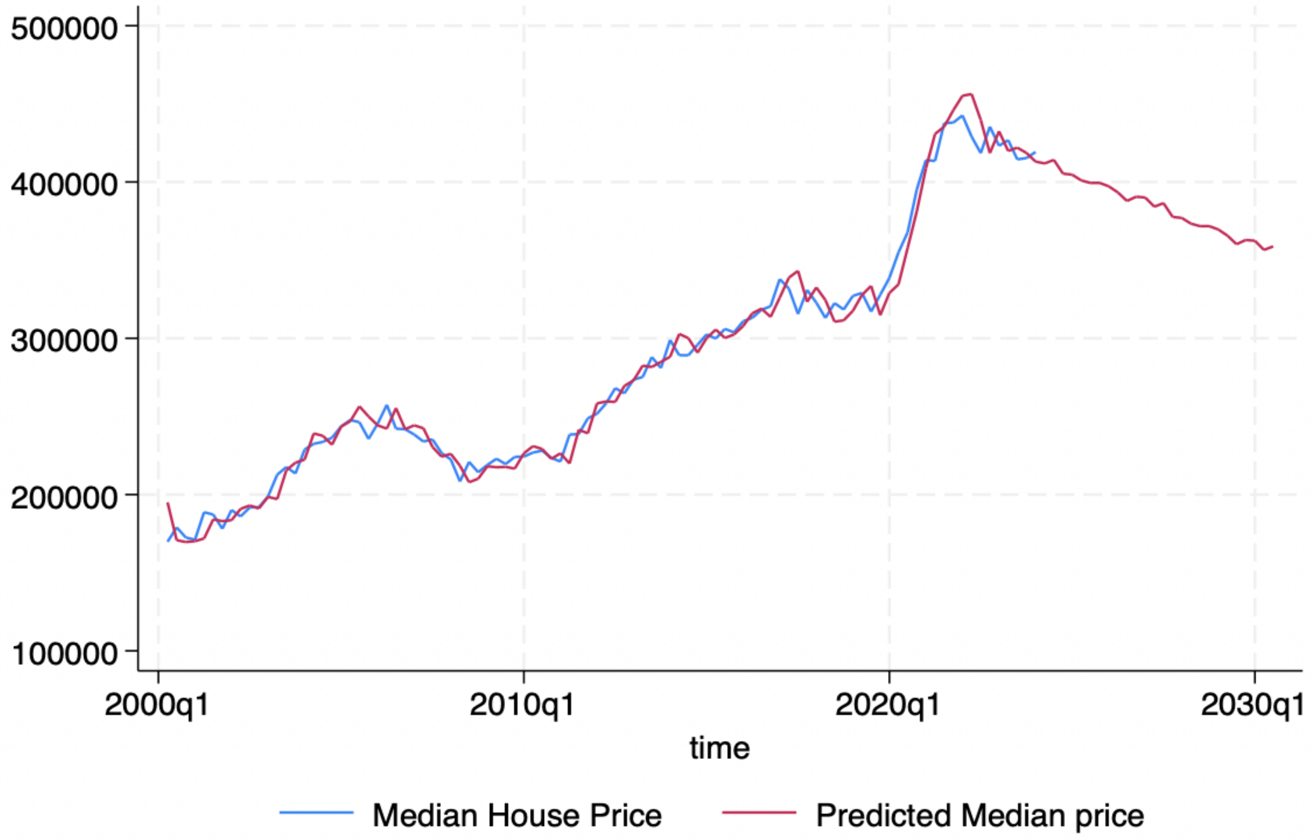

Data range: Jan 1982-2024

# Double Exponential Smoothing with a Centered Moving Average

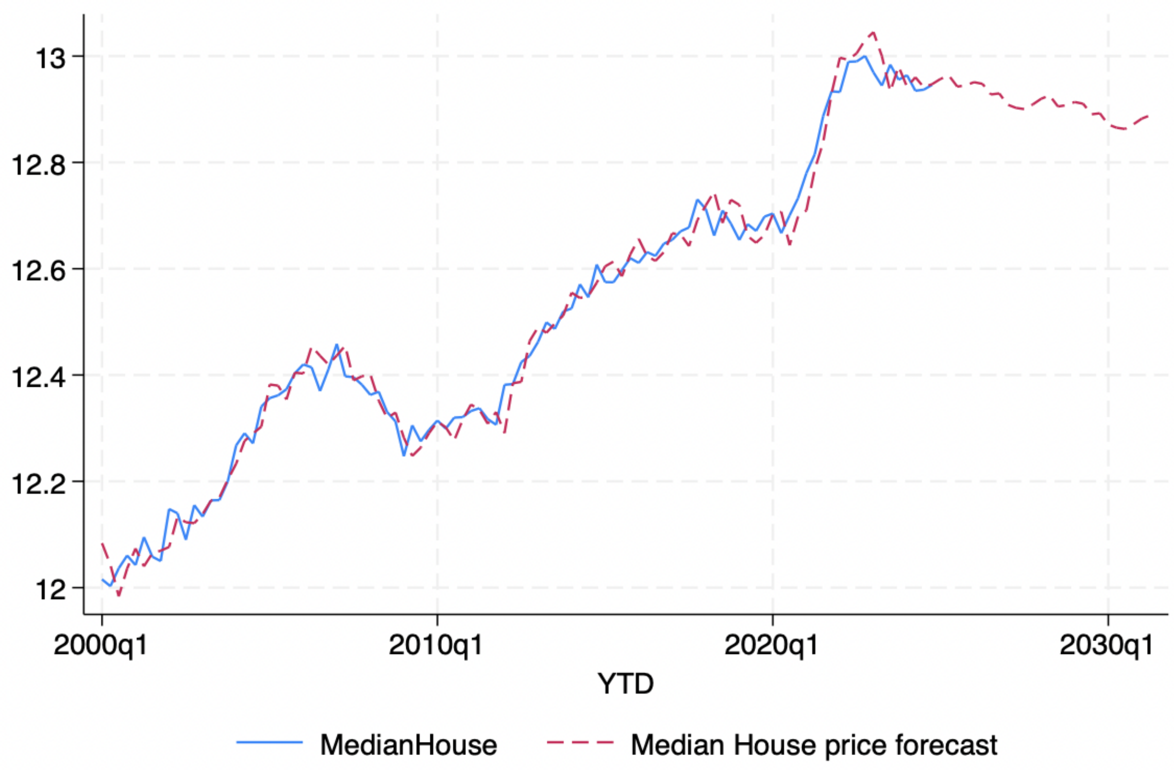



Data range: 2000 to 2024

# Holt Winters Additive



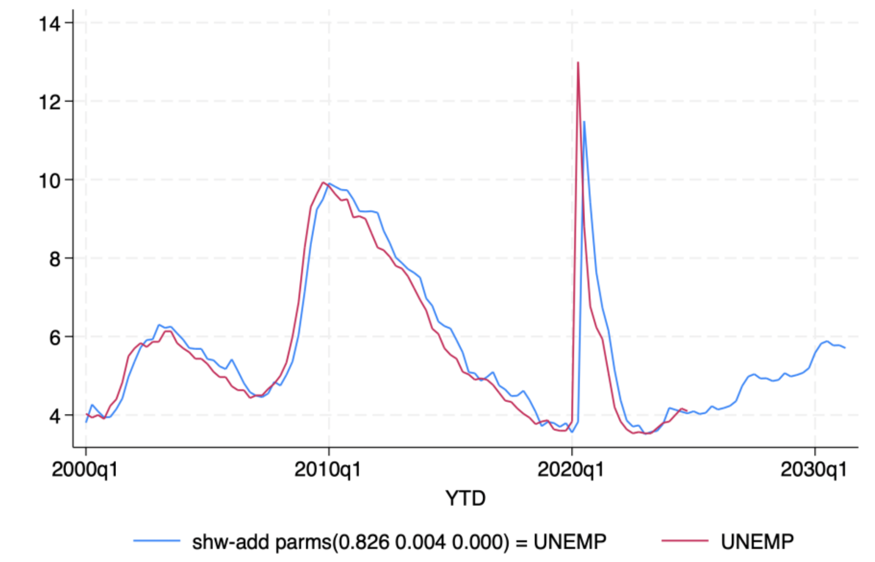

Data range: 2000 to 2024




Data range: 2000 to 2024

# Holt Winters Additive( Log Transformed)

# Suggested Causes of Housing Price Decreases




# Forecasting Approaches

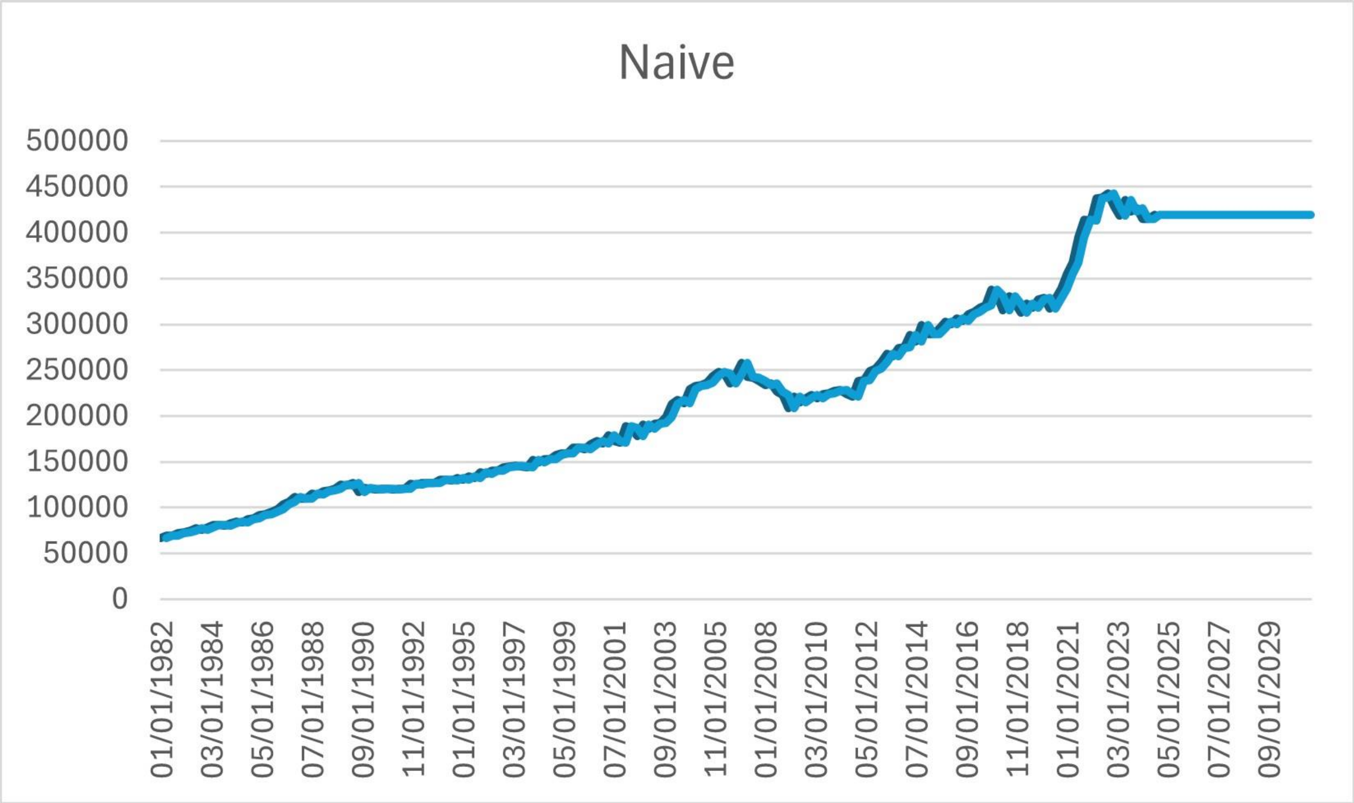
# Naive Approach

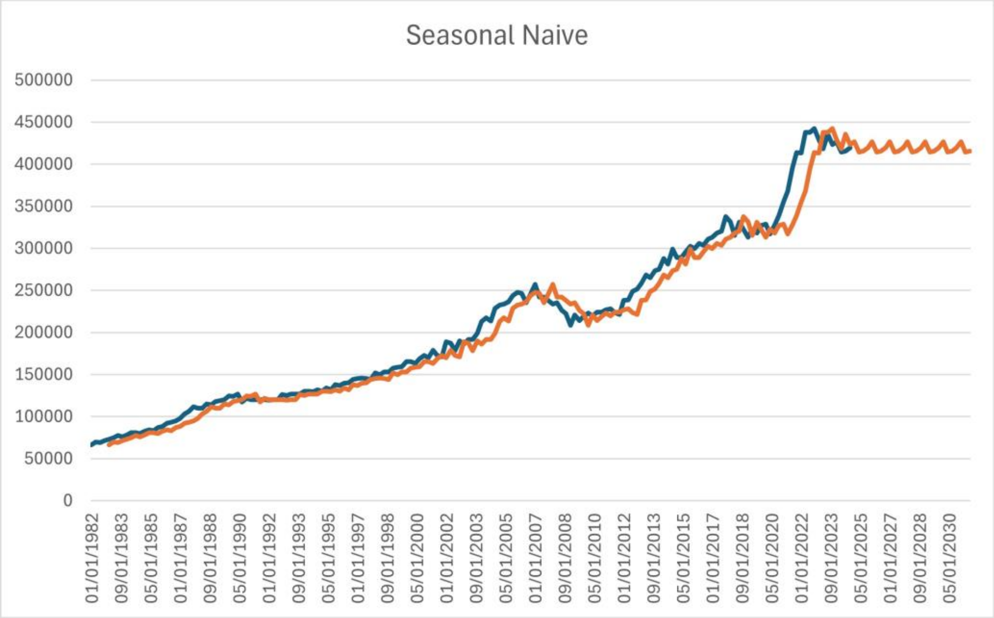
The naive approach simply uses the previous value as the next forecast. The naive approach has validity but is very simple not showing the trend. Looking at figure ….

The season naïve captures a seasonal trend to the data but still provided a repetitive result. Although these forecast methods are very simple, they proved grounds for more complex forecasting

The forecast simply cuts off at a certain value and stays level

# Naïve and Seasonal Naïve



# Forecasting Approaches cont.

- **Exponential smoothing**
- **$\hat{y}_{t+1|t} = \alpha Y_t + \alpha(1 - \alpha)Y_{t-1} + \alpha(1 - \alpha)^2 Y_{t-2} + \ldots$**
- **A verbal description of this expression is:**
- **New mean = old mean + (difference between new observation and old mean) / new sample size (= n + 1)**

- **Moving average**
- **$MA^{(2,1,1)}_t = (1/4) \times (Y_{t-2} + Y_{t-1} + Y_t + Y_{t+1})$**
- **$MA^{(1,1,2)}_t = (1/4) \times (Y_{t-1} + Y_t + Y_{t+1} + Y_{t+2})$**
- **$CMA_t = (1/2) \times [MA^{(2,1,1)}_t + MA^{(1,1,2)}_t]$**

# Forecasting Approaches cont.

- **Naive**
- $\hat{y}_{\square+\square|\square} = y_{\square}$
- **Drift**
- $\hat{y}_{\square+\square|\square} = y_{\square} + (h / (T - 1)) \times \Sigma_{\square=2}{}^{\wedge T} (y_{\square} - y_{\square-1})$
- $= y_{\square} + h \times ((y_{\square} - y_1) / (T - 1))$

- **Double exponential smoothing**
- $S_{2,\square} = \alpha \times S_{1,\square} + (1 - \alpha) \times S_{2,\square-1}$
- **Holt-winters Additive**
- $L_{\square} = L_{\square-1} + T_{\square-1} + \alpha e_{\square}$
- $T_{\square} = T_{\square-1} + \alpha\beta e_{\square}$
- $S_{\square} = S_{\square-\square} + \gamma e_{\square}$

# Stata Code Representation

- tsmooth hwinters sm1= Y, forecast(3)
- tssmooth ma madxpmhp = dexp2Mhp, window(2 1 1)
- madxpmhp(t) = (1/4) * [x(t-2) + x(t-1) + x(t) + x(t+1)]
- tssmooth ma ma2xpmhp = dexp2Mhp, window(1 1 2)
- ma2xpmhp(t) = (1/4) * [x(t-1) + x(t) + x(t+1) + x(t+2)]
- tssmooth shwinter ASHW_MHP= Median hp, period(x) additive replace forecast(x)

# Conclusion

- Price trends provides insight into consumer speculation and consumer confidence
- Housing prices can serve as a benchmark for the U.S. economy, and it can be a predictive tool
- Housing market is more than just supply and demand

# Citations


- Baum, Christopher F. *An Introduction to Modern Econometrics Using Stata*. Stata Press, 2006.
- Booth, Wayne C., et al. *The Craft of Research*. The University of Chicago Press, 2024.
- Deaton, Angus. "(PDF) Housing, Land Prices, and Growth." *Journal of Economic Growth*, Kluwer Academic Publishers, www.researchgate.net/publication/5149905_Housing_Land_Prices_and_Growth. Accessed 2 Apr. 2025.
- Hirano, Tomohiro, and Joseph E. Stiglitz. "Growth and Fluctuations Economies with Land Speculation." *NBER*, 24 Mar. 2025, www.nber.org/papers/w33589.
- Long, J. Scott. *The Workflow of Data Analysis Using Stata*. Stata Press, 2009.
- Sun, Qian, et al. "The Impact of Housing Prices and Land Financing on Economic Growth: Evidence from Chinese 277 Cities at the Prefecture Level and Above." *PLOS ONE*, Public Library of Science, journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0302631. Accessed 22 Mar. 2025.