Time Series Analysis of Ames Housing Data

MATH 62.2 - Project Presentation

Brayden Jansen O. Ang   Miguel Antonio H. Germar
Andrea Mikaela S. Zialcita

Ateneo de Manila University

May 21, 2025

De Cock (2011)

Title: Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project

Author: Dean De Cock

Journal: Journal of Statistics Education, Vol. 19, no. 3

Year: 2011

DOI: https://doi.org/10.1080/10691898.2011.11889627

'AmesHousing' Package

Package 'AmesHousing'

January 20, 2025

Version 0.0.4

Title The Ames Iowa Housing Data

URL https://github.com/topepo/AmesHousing

BugReports https://github.com/topepo/AmesHousing/issues

Description Raw and processed versions of the data from De Cock (2011) <http://ww2.amstat.org/publications/jse> are included in the package.

License GPL-2

Encoding UTF-8

LazyData true

ByteCompile true

Depends R (>= 2.10)

Imports dplyr, magrittr

RoxygenNote 7.1.0.9000

Suggests covr

NeedsCompilation no

Author Max Kuhn [aut, cre],
Dmytro Perepolkin [ctb],
RStudio [cph]

Maintainer Max Kuhn <max@rstudio.com>

Repository CRAN

Date/Publication 2020-06-23 20:10:03 UTC

Ames Housing Dataset

	MS_SubClass	MS_Zoning	Lot_Frontage	Lot_Area	Street	Alley	Lot_Shape	Land_Contour	Utilities	Lot_Config	Land_Slope	Neighborhood	Condi
1	One_Story_1946_and_Newer_All_Styles	Residential_Low_Density	141	31770	Pave	No_Alley_Access	Slightly_Irregular	Lvl	AllPub	Corner	Gtl	North_Ames	Norm
2	One_Story_1946_and_Newer_All_Styles	Residential_High_Density	80	11622	Pave	No_Alley_Access	Regular	Lvl	AllPub	Inside	Gtl	North_Ames	Feedr
3	One_Story_1946_and_Newer_All_Styles	Residential_Low_Density	81	14267	Pave	No_Alley_Access	Slightly_Irregular	Lvl	AllPub	Corner	Gtl	North_Ames	Norm
4	One_Story_1946_and_Newer_All_Styles	Residential_Low_Density	93	11160	Pave	No_Alley_Access	Regular	Lvl	AllPub	Corner	Gtl	North_Ames	Norm
5	Two_Story_1946_and_Newer	Residential_Low_Density	74	13830	Pave	No_Alley_Access	Slightly_Irregular	Lvl	AllPub	Inside	Gtl	Gilbert	Norm
6	Two_Story_1946_and_Newer	Residential_Low_Density	78	9978	Pave	No_Alley_Access	Slightly_Irregular	Lvl	AllPub	Inside	Gtl	Gilbert	Norm
7	One_Story_PUD_1946_and_Newer	Residential_Low_Density	41	4920	Pave	No_Alley_Access	Regular	Lvl	AllPub	Inside	Gtl	Stone_Brook	Norm
8	One_Story_PUD_1946_and_Newer	Residential_Low_Density	43	5005	Pave	No_Alley_Access	Slightly_Irregular	HLS	AllPub	Inside	Gtl	Stone_Brook	Norm
9	One_Story_PUD_1946_and_Newer	Residential_Low_Density	39	5389	Pave	No_Alley_Access	Slightly_Irregular	Lvl	AllPub	Inside	Gtl	Stone_Brook	Norm
10	Two_Story_1946_and_Newer	Residential_Low_Density	60	7500	Pave	No_Alley_Access	Regular	Lvl	AllPub	Inside	Gtl	Gilbert	Norm
11	Two_Story_1946_and_Newer	Residential_Low_Density	75	10000	Pave	No_Alley_Access	Slightly_Irregular	Lvl	AllPub	Corner	Gtl	Gilbert	Norm
12	One_Story_1946_and_Newer_All_Styles	Residential_Low_Density	0	7980	Pave	No_Alley_Access	Slightly_Irregular	Lvl	AllPub	Inside	Gtl	Gilbert	Norm
13	Two_Story_1946_and_Newer	Residential_Low_Density	63	8402	Pave	No_Alley_Access	Slightly_Irregular	Lvl	AllPub	Inside	Gtl	Gilbert	Norm
14	One_Story_1946_and_Newer_All_Styles	Residential_Low_Density	85	10176	Pave	No_Alley_Access	Regular	Lvl	AllPub	Inside	Gtl	Gilbert	Norm

- Housing data from Ames, Iowa compiled by the Ames Assessor's Office
- Data is from 2006 to 2010
- Most houses are 3-bedroom houses, we focus on those only
- Many features, but we focus on year and month sold, as well as sale price

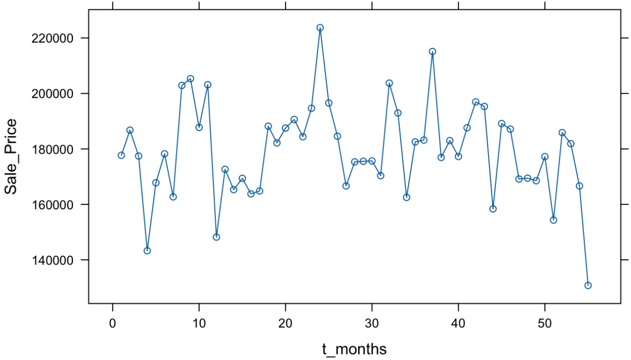
Statement of the Problem

We want to

1. Model the mean price of 3-bedroom houses in Ames as a time series.
2. Predict the mean sale price of 3-bedroom houses in Ames for the next 12 months.

Initial Dataset Transformations

1. Filter 3-bedroom houses, keep year and month sold and sale price
2. Convert year and month sold to a time variable
3. Take average monthly sale price of 3-bedroom houses
4. Make $\{ X_t \}$ a time series object



Testing Stationarity

We first test if $\{ X_t \}$ is stationary.
H_0 : the data is not stationary (there is a unit root)
H_1 : the data is stationary (there is no unit root)

Augmented Dickey-Fuller Test

```
adf.test(x)
```

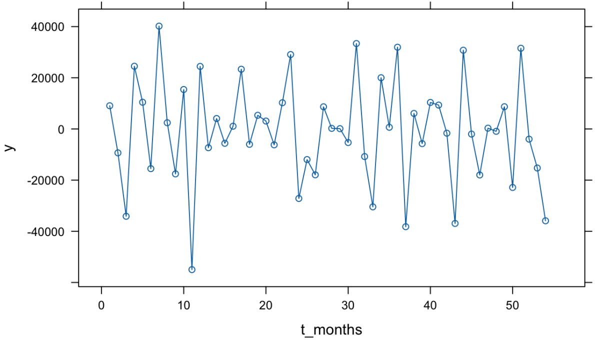
This gives us p-value $= 0.2192 > 0.05$, so we do not reject H_0, and we say that $\{ X_t \}$ is not stationary.

We take

$$Y_t = \nabla X_t = X_t - X_{t-1}.$$

Differencing

```
y = diff(x)
```



Testing Stationarity

We then test if $\{ Y_t \}$ is stationary.
H_0 : the data is not stationary (there is a unit root)
H_1 : the data is stationary (there is no unit root)

Augmented Dickey-Fuller Test

```
adf.test(y)
```

This gives us p-value $< 0.01 < 0.05$, so we reject H_0, and we say that $\{ Y_t \}$ is stationary.

Testing Autocorrelation

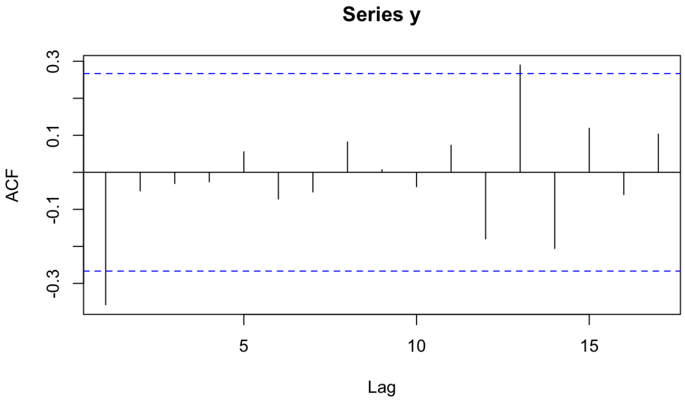
We then test if $\{ Y_t \}$ is autocorrelated.
H_0 : the data is uncorrelated
H_1 : the data is autocorrelated

Ljung-Box Test

```
Box.test(y, type = "Ljung", lag = 2)
```

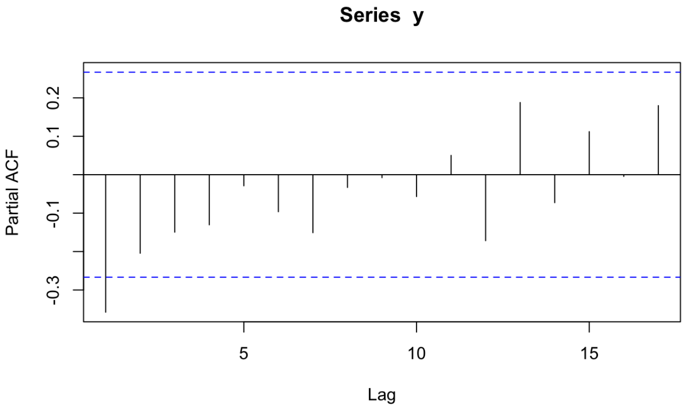
This gives us p-value $= 0.02408 < 0.05$, so we reject H_0, and we say that $\{ Y_t \}$ is autocorrelated. We can now proceed to modeling.

Identifying MA Order



The ACF cuts off at lag 1. So, we consider MA with lag order 1. By parsimony, we ignore lag 13.

Identifying AR Order



The PACF cuts off at lag 1. So, we consider AR with lag order 1.

Testing AR(1)

Testing AR(1)

```
Arima(y, order = c(1, 0, 0))
```

This gives us AIC $= 1222.36$, log-likelihood $= -608.18$, and the equation

$$\begin{aligned} Y_t &= -741.6501(1 + 0.3735) - 0.3735Y_{t-1} \\ &= -1020.2037 - 0.3735Y_{t-1} + Z_t, \end{aligned}$$

where $\{ Z_t \} \sim \mathrm{WN}(0, \sigma^2)$.

Testing MA(1)

Testing MA(1)

```
Arima(y, order = c(0, 0, 1))
```

This gives us AIC $= 1214.18$, log-likelihood $= -604.09$, and the equation

$$Y_t = -187.6784 - 0.8411 Z_{t-1} + Z_t.$$

Testing ARMA$(1, 1)$

Testing ARMA(1, 1)

```
Arima(y, order = c(1, 0, 1))
```

This gives us AIC $= 1213.50$, log-likelihood $= -602.75$, and the equation

$$\begin{aligned} Y_t &= -114.4641(1 - 0.2786) + 0.2786Y_{t-1} + Z_t \\ &= -82.5768 + 0.2786Y_{t-1} + Z_t - 0.9999985Z_{t-1}. \end{aligned}$$

Choosing ARMA$(1, 1)$

Since the ARMA$(1, 1)$ had the lowest AIC and the highest log likelihood, we choose the ARMA$(1, 1)$ model given by

$$\begin{aligned} Y_t &= -114.4641(1 - 0.2786) + 0.2786 Y_{t-1} + Z_t - Z_{t-1} \\ &= -82.5744 + 0.2786 Y_{t-1} + Z_t - Z_{t-1}. \end{aligned}$$

Grid search supports that ARMA$(1, 1)$ has the lowest AIC among all models ARMA(p, q), where $1 \leq p \leq 2$ and $1 \leq q \leq 2$.

Checking Causality

The AR polynomial is given by

$$\phi(B) = 1 - 0.2786B,$$

and the solution of $\phi(B) = 0$ is

$$\begin{aligned} B &= \frac{1}{0.2786} \\ &\approx 3.5894 \\ \implies |B| &> 1. \end{aligned}$$

So, the model is causal.

Checking Invertibility

The MA polynomial is given by

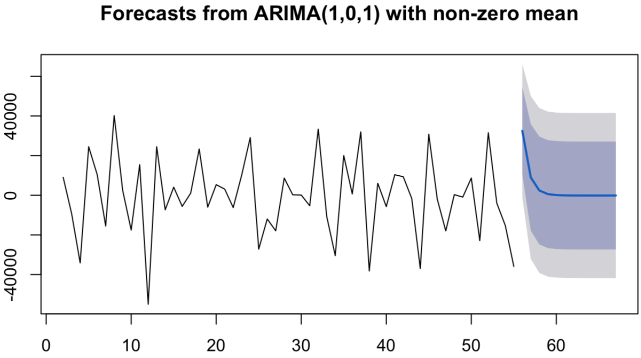
$$\theta(B) = 1 - 0.9999985B.$$

Note that R rounded off the coefficient of B to -1.0000. The solution of $\theta(B) = 0$ is then

$$\begin{aligned} B &= \frac{1}{0.9999985} \\ &> 1 \\ \implies |B| &> 1. \end{aligned}$$

So, the model is invertible.

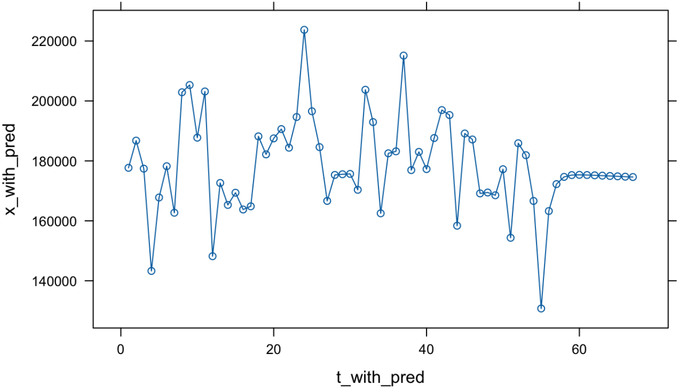
Forecasting



Note that $\{ Y_t \}$ is the differenced mean sale price per month.

Forecasting

So, we take the predicted values of Y_t for the next 12 months and apply them starting from the last X_t to get the predicted values of X_t for the next 12 months.



Results and Discussion

- Assuming that the forecast is accurate, the sale prices of 3-bedroom houses will sharply increase in first two months and then be generally stable over the next year
- Forecasted values follow the overall mean, but do not capture the volatility of the sale price
- Indicates that an ARMA model might not be the most suitable model

Appendix: Grid Search

The models arranged by increasing AIC are:

AR Order (p)	MA Order (q)	AIC
1	1	1213.5020
0	2	1213.9340
0	1	1214.1820
2	2	1215.4480
1	1	1215.4510
2	2	1217.4630
2	0	1221.3560
1	0	1222.3590