Modeling the Case-Shiller Index for Los Angeles housing using SARIMA modeling

Robert Iannuzzo

Abstract

In this paper we discuss techniques used to investigate, diagnose, model and ultimately predict data from the Case-Shiller Los Angeles Housing Index. The data contains housing price averages from Los Angeles from 1987 until now. After analyzing and adjusting the data we used techniques from SARIMA modeling to fit a model to the data and predict new values. We captured good fit for this data, suggesting prices that will continue to trend upwards. It also showed some limitations which we will discuss.

Introduction

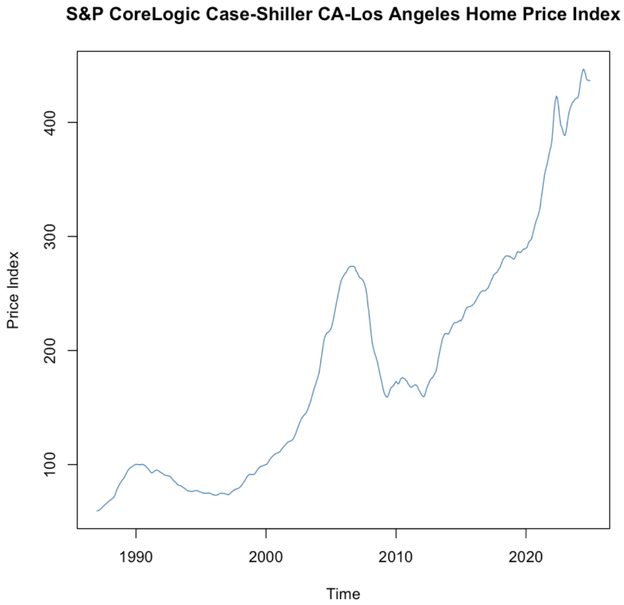
Los Angeles is the second-largest city in the United States and home to one of the most dynamic and expansive real estate markets in the world. Its unique combination of a sprawling urban landscape and a population of nearly four million residents makes real estate a critical component of the city's economy. As such, understanding and forecasting housing prices in Los Angeles holds value not only for prospective homebuyers and sellers but also for investors, policymakers, and real estate professionals seeking to make informed decisions.

To capture the movement of housing prices over time, we turn to the Case-Shiller Home Price Index, a widely respected benchmark for tracking residential real estate trends. The Case-Shiller index standardizes prices relative to the year 2000, assigning it a base value of 100. An index value of 200, for example, would indicate that home prices have doubled since that baseline year. In this project, we will work with monthly Case-Shiller index data from 1987 to the present, allowing us to observe fine-grained trends and seasonality in the housing market.

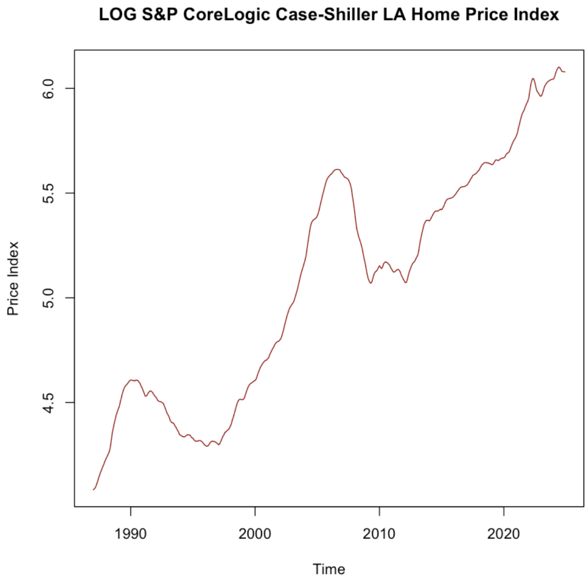
While this dataset offers valuable insights, it is important to note one limitation: the index only includes repeat sales of existing homes, thereby excluding newly developed properties. This constraint helps maintain data consistency but may introduce some bias, particularly in a city like Los Angeles where new construction plays a significant role in the housing landscape.

Model Fitting

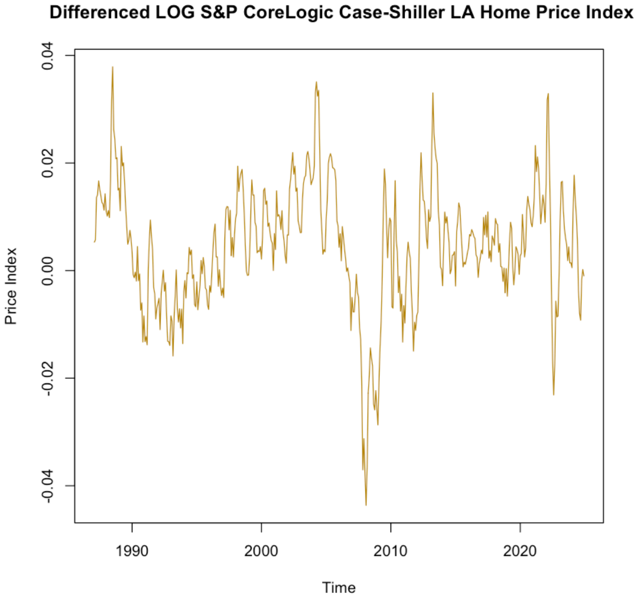
We start the model fitting just by looking at a plot of the original data.



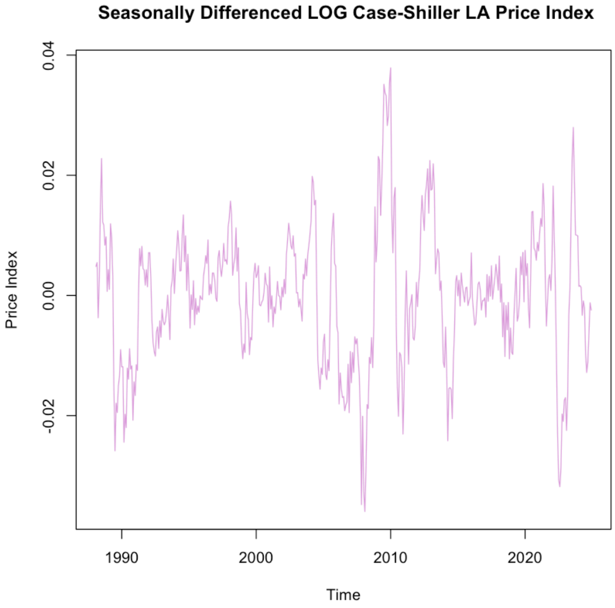
When we are looking to fit time series data with a SARIMA model we need to look for data has a few key properties: a stable mean and consistent variance. This is what is knows as being "stationary". This data does not currently have these properties but we can adjust it so it looks closer to what we are looking for. The first thing we will do is take the log of all of the data. This helps with keeping variance constant. This won't look dramatic now but it does help.



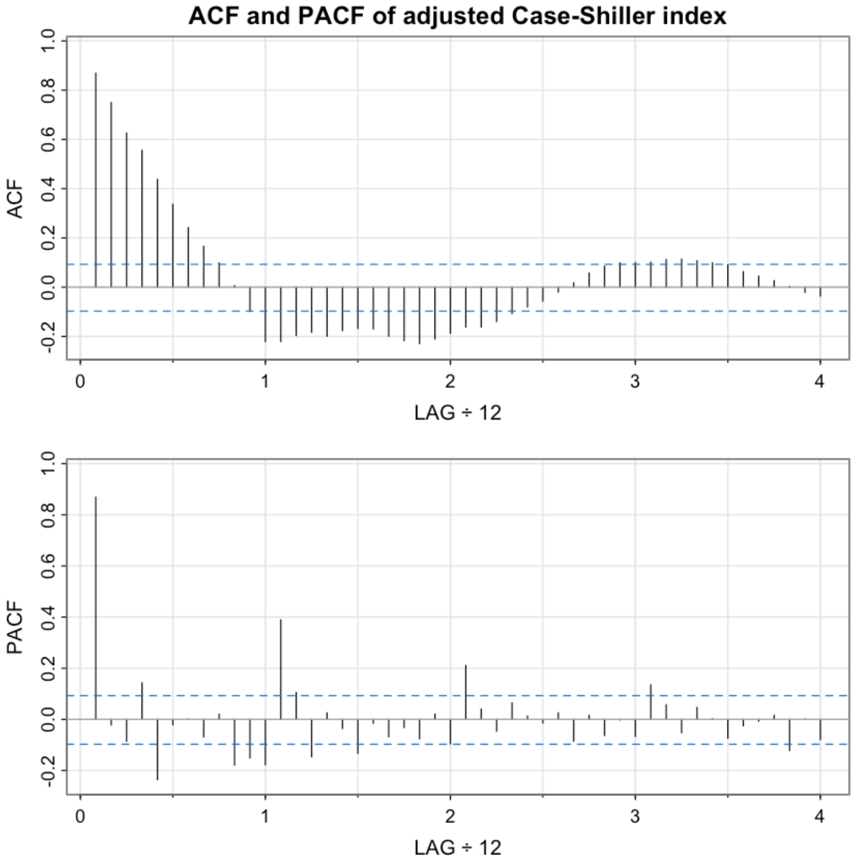
Next, we will take the difference of the data. As in, instead of looking at the raw data points we look at the difference of one point to the next. This will give us a stable mean. You should notice that now most of the values are centered around 0.



This data is looking a lot better, but we need to do one more thing. Housing prices are seasonal (people tend to buy in summer months so prices go up because demand is higher). We will take a seasonal difference now to account for this. This means subtracting the value from a year ago as well.



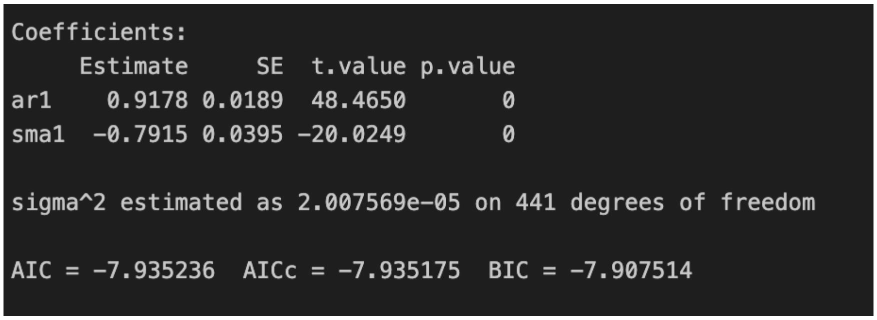
Now our data is looking well adjusted for SARIMA modeling and we will continue on to the next step. We want to now look at the autocorrelation (ACF) and partial autocorrelation (PACF) plots for this data to glean insight into how to start the model building process. These plots measure how closely correlated time points are to previous time points (the index which says how was away the time point is is known as “lag”).



Our SARIMA model will contain two parts: the “ARIMA” part which measures how much our current time point will relate previous time points (AR, which stands for “autoregressive”) and also how much it is related to random (white noise) processes (MA, which stands for “moving average”). The “S” part measures the seasonality of the data: how much our time point is influenced by the season it is in. The SARIMA model also accounts for the differencing that we have already done.

The initial model we would use for this is (1,1,0)x(0,1,1). The ACF tails off while there is clear cutoff after lag 1 in PACF. This would indicate an autoregressive (AR) model. There are lag spikes at 12,24,36 in ACF which tells me there is an SMA term. There are spikes in PACF but they aren't at 12,24,36 so we won't add an SAR term for now. The way we look to see if the model is fitting our data well is a two-fold process. We look at the statistical significance of the coefficients of our terms in our chose SARIMA model and if the p-value for the coefficient is less than 0.05 then we deem that coefficient “significant”. We also do statistical diagnostics on the residuals of our data to make sure they’re behaving randomly (if they are not then that means there’s some discrete behavior in the data that our model is not accounting for, and therefore we could find a better model). We will show this information for each model we talk about in this paper. We also want to note that all of our models will have middle terms equal to 1 because our data is differenced and seasonally differenced always.

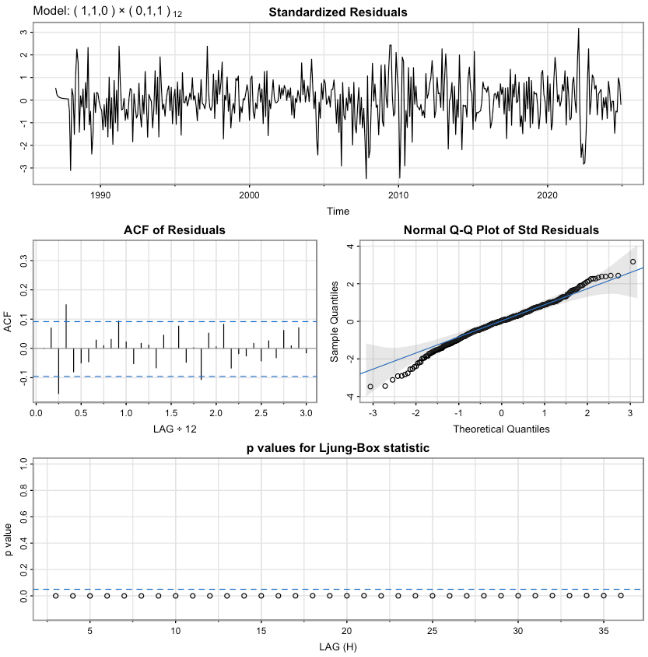
Our first model like we said is the (1,1,0)x(0,1,1) model.



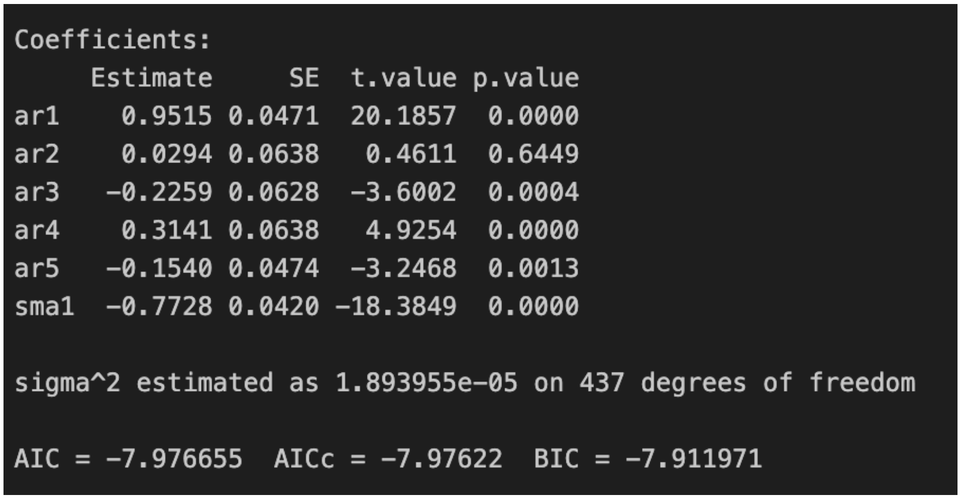
```
Coefficients:
     Estimate     SE  t.value p.value
ar1    0.9178 0.0189  48.4650       0
sma1  -0.7915 0.0395 -20.0249       0

sigma^2 estimated as 2.007569e-05 on 441 degrees of freedom

AIC = -7.935236  AICc = -7.935175  BIC = -7.907514
```



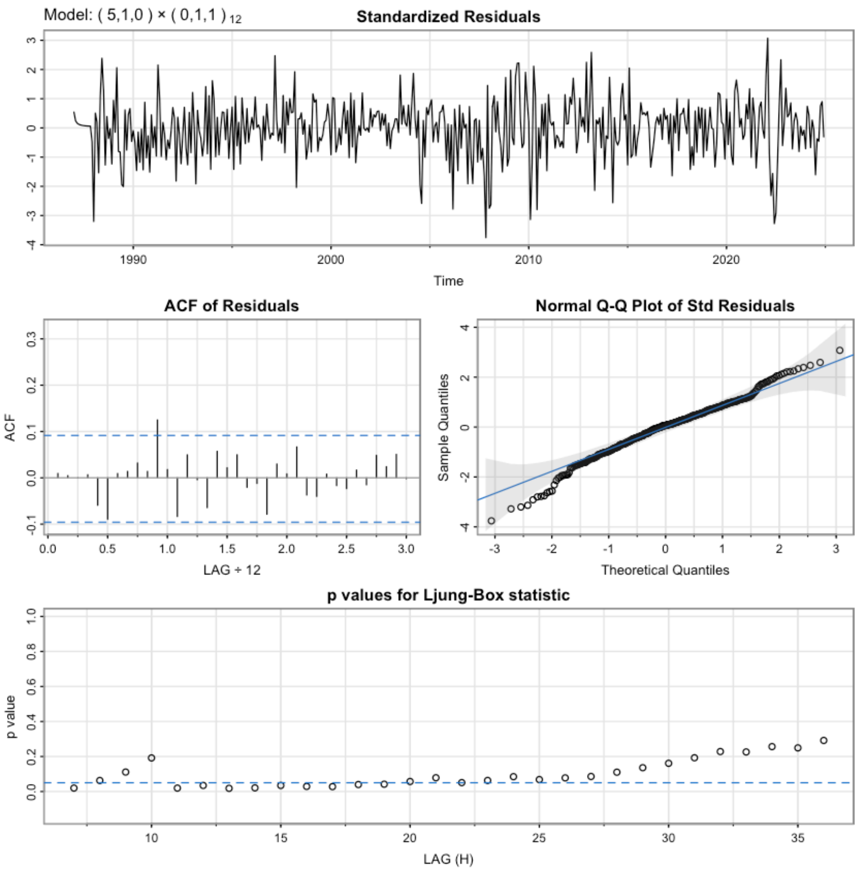
We can tell immediately this is not the model we want to use. Our coefficients may be significant but the residual diagnostics are not showing white noise behavior. Two things I am looking at to confirm this are the Ljung-Box statistic p-values being below the blue dotted line and the ACF spikes being above the blue dotted line: we want neither of those things. This means our model is underfit and we need to add more terms it. We will continue by adding more AR terms, which is the first coordinate of the first parentheses set, since we suspect this is a strongly autoregressive model. The next model will be (5,1,0)x(0,1,1).



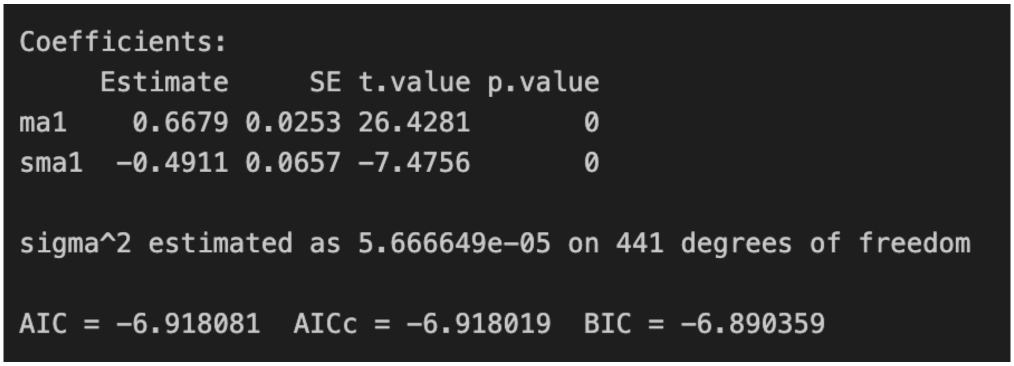
```
Coefficients:
     Estimate     SE  t.value p.value
ar1    0.9515 0.0471  20.1857  0.0000
ar2    0.0294 0.0638   0.4611  0.6449
ar3   -0.2259 0.0628  -3.6002  0.0004
ar4    0.3141 0.0638   4.9254  0.0000
ar5   -0.1540 0.0474  -3.2468  0.0013
sma1  -0.7728 0.0420 -18.3849  0.0000

sigma^2 estimated as 1.893955e-05 on 437 degrees of freedom

AIC = -7.976655  AICc = -7.97622  BIC = -7.911971
```



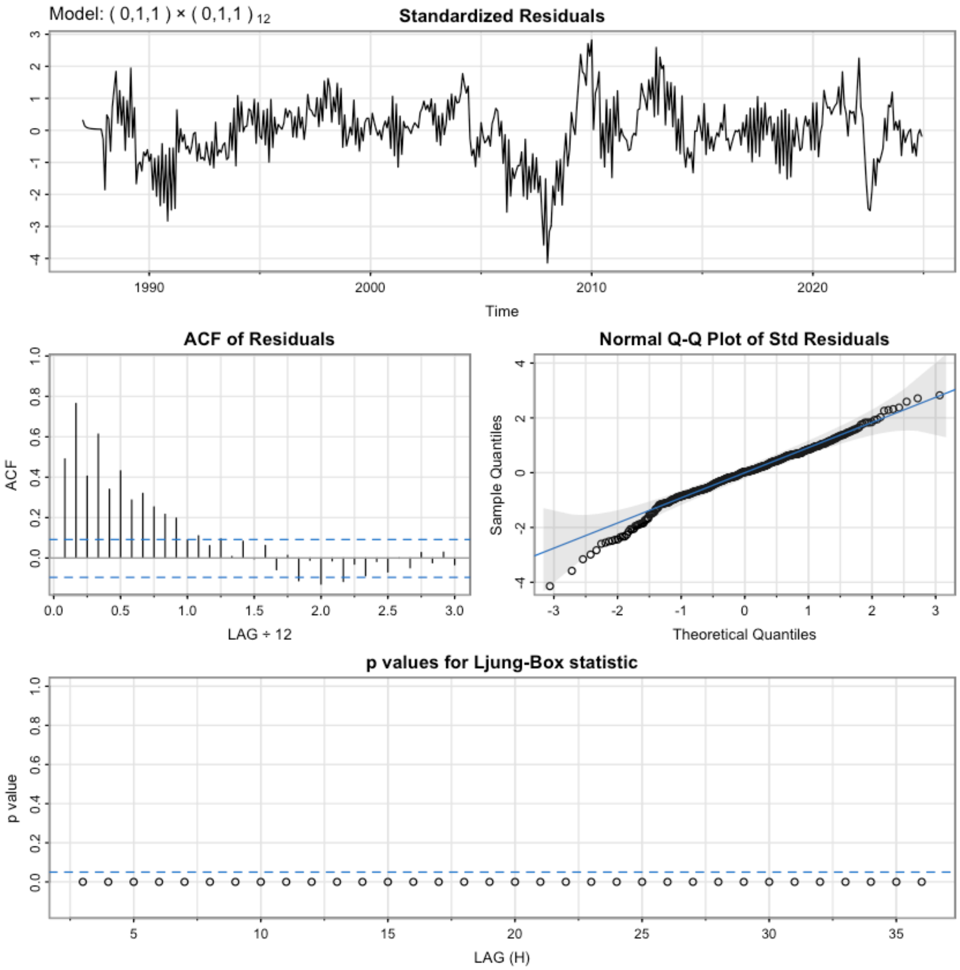
This model is definitely better than the last one since we are seeing some activity in the Ljung-Box statistic p-value plot. This suggests to me that more AR terms was a good thing and we will keep them. We would like to investigate now whether or not we need any MA terms, the third number of the first parentheses set. We will look at (0,1,1)x(0,1,1) now.



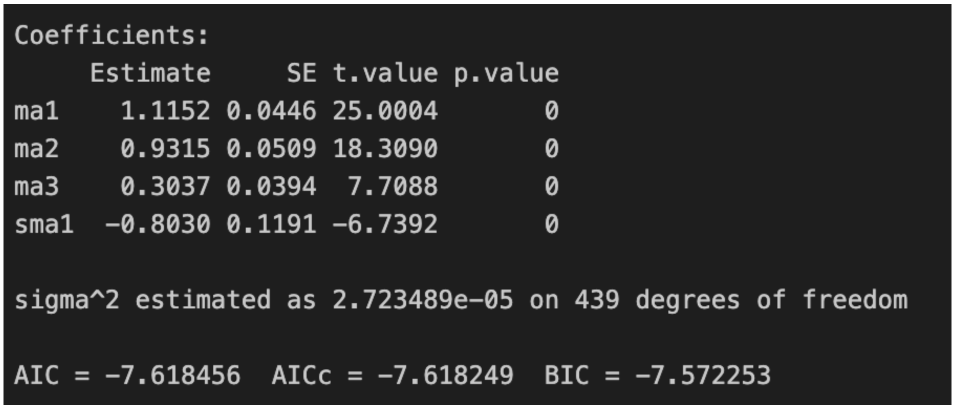
```
Coefficients:
     Estimate     SE t.value p.value
ma1    0.6679 0.0253 26.4281       0
sma1  -0.4911 0.0657 -7.4756       0

sigma^2 estimated as 5.666649e-05 on 441 degrees of freedom

AIC = -6.918081  AICc = -6.918019  BIC = -6.890359
```



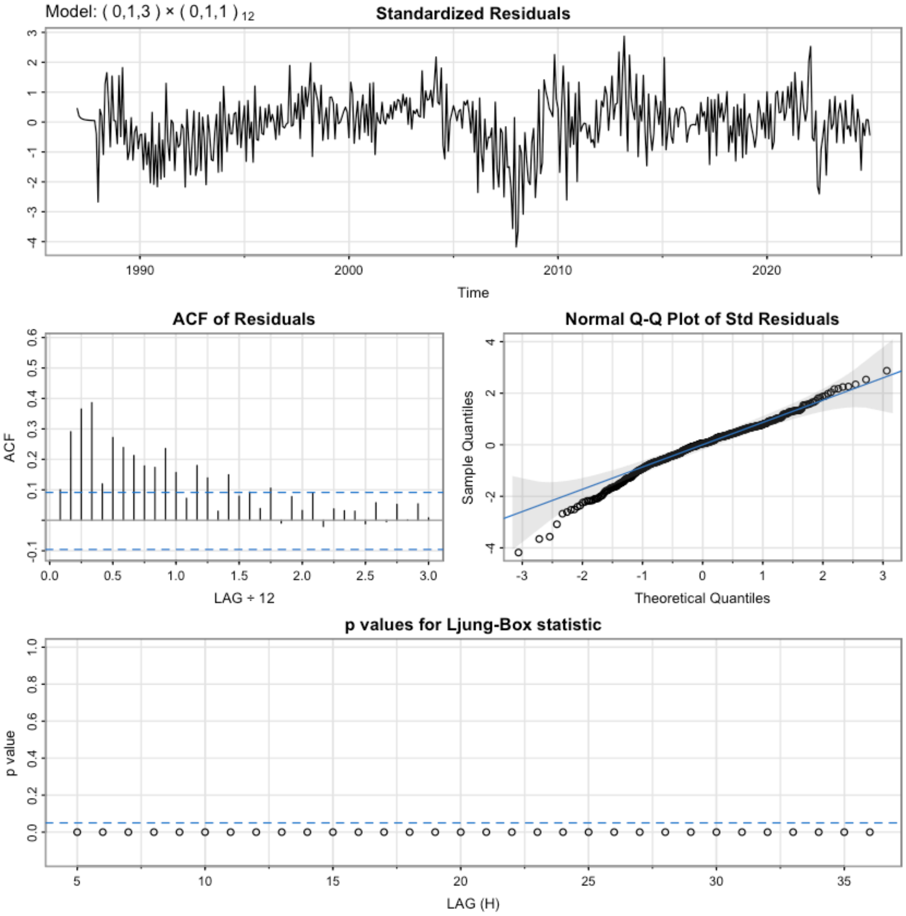
It is very clear from the plot of the ACF of the residuals that this will not be a good model. However, the coefficient of the MA term is significant which says that we may want some of these terms in our model. Let's add more and see what happens. We will try (0,1,3)x(0,1,1).



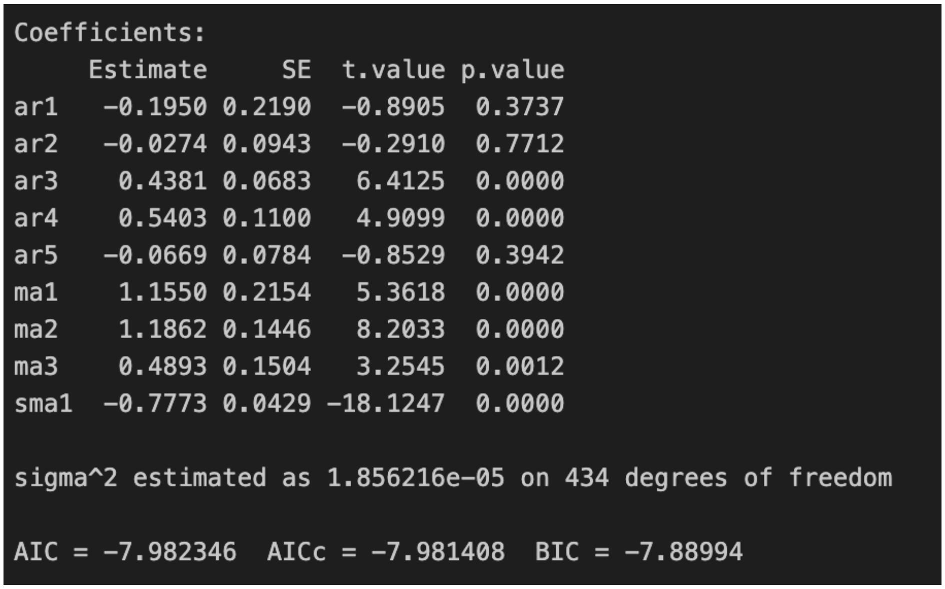
```
Coefficients:
     Estimate     SE t.value p.value
ma1    1.1152 0.0446 25.0004       0
ma2    0.9315 0.0509 18.3090       0
ma3    0.3037 0.0394  7.7088       0
sma1  -0.8030 0.1191 -6.7392       0

sigma^2 estimated as 2.723489e-05 on 439 degrees of freedom

AIC = -7.618456  AICc = -7.618249  BIC = -7.572253
```



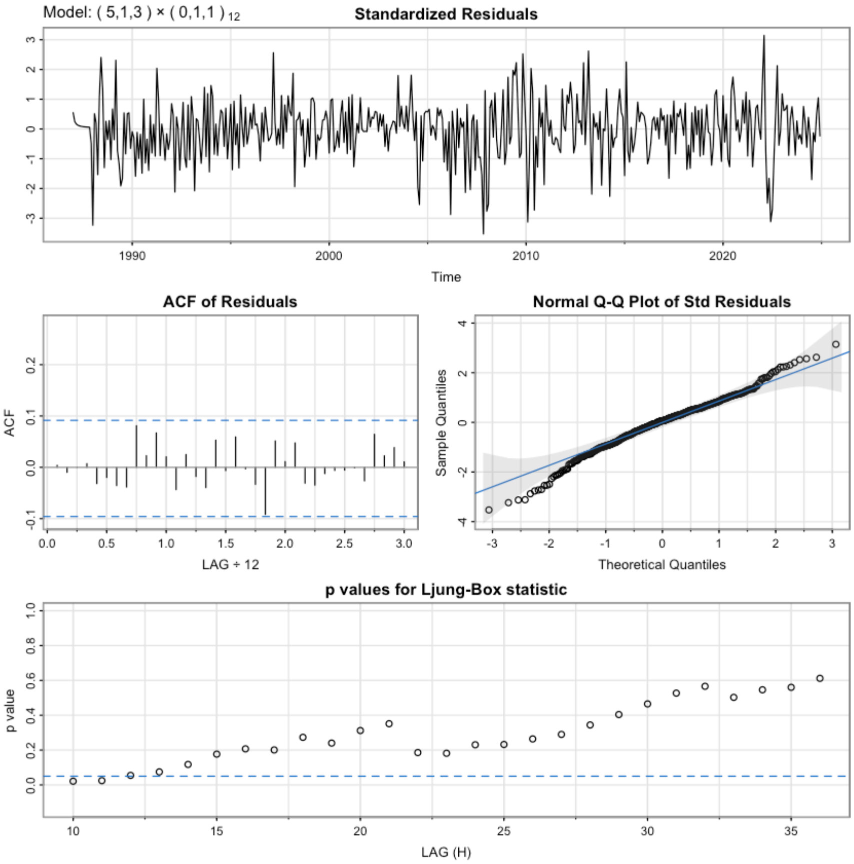
Again, this model is no good for the same reason as the last one, but, yet again, we can see the MA terms are significant. Let's try combing our best model so far with this one and see if we can get a more balanced model. This model will be (5,1,3)x(0,1,1).



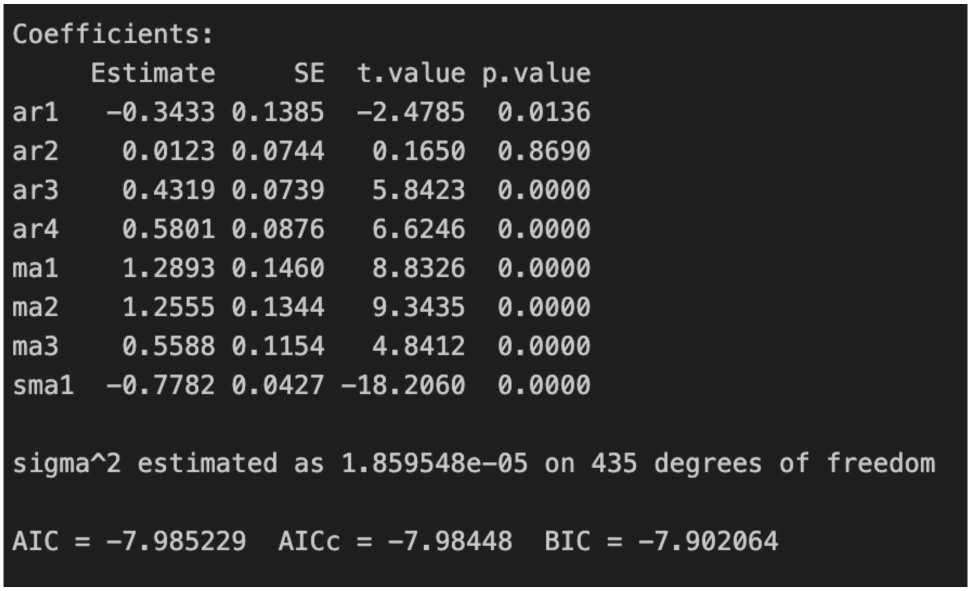
```
Coefficients:
     Estimate     SE  t.value p.value
ar1   -0.1950 0.2190  -0.8905  0.3737
ar2   -0.0274 0.0943  -0.2910  0.7712
ar3    0.4381 0.0683   6.4125  0.0000
ar4    0.5403 0.1100   4.9099  0.0000
ar5   -0.0669 0.0784  -0.8529  0.3942
ma1    1.1550 0.2154   5.3618  0.0000
ma2    1.1862 0.1446   8.2033  0.0000
ma3    0.4893 0.1504   3.2545  0.0012
sma1  -0.7773 0.0429 -18.1247  0.0000

sigma^2 estimated as 1.856216e-05 on 434 degrees of freedom

AIC = -7.982346  AICc = -7.981408  BIC = -7.88994
```



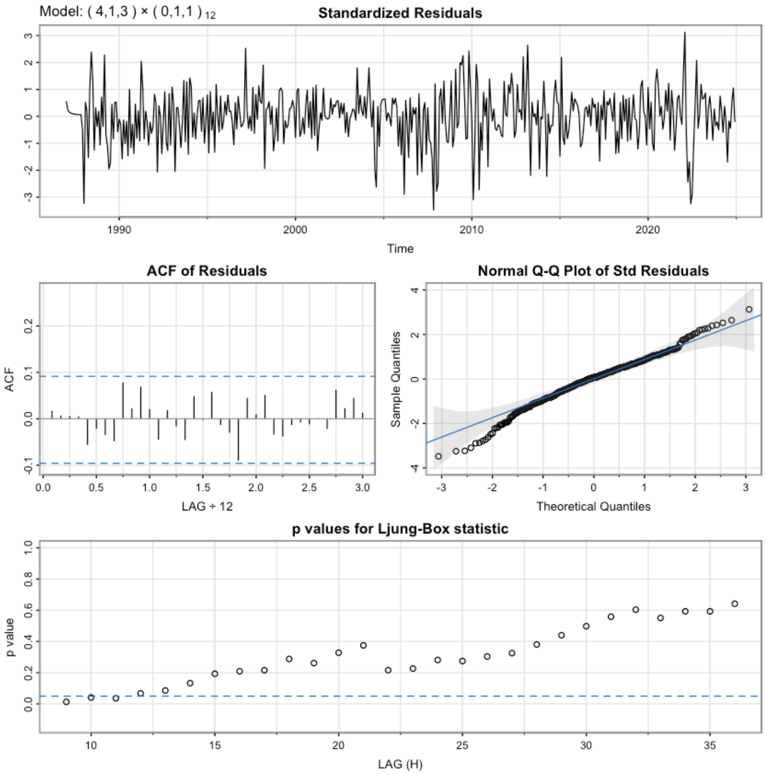
This model is looking very good, our best one yet. We can see the residuals are passing the white noise test quite comfortably. The only issues now is the AR5 terms is not statistically significant so we have to remove it. This give us the (4,1,3)x(0,1,1) model.



```
Coefficients:
     Estimate     SE  t.value p.value
ar1   -0.3433 0.1385  -2.4785  0.0136
ar2    0.0123 0.0744   0.1650  0.8690
ar3    0.4319 0.0739   5.8423  0.0000
ar4    0.5801 0.0876   6.6246  0.0000
ma1    1.2893 0.1460   8.8326  0.0000
ma2    1.2555 0.1344   9.3435  0.0000
ma3    0.5588 0.1154   4.8412  0.0000
sma1  -0.7782 0.0427 -18.2060  0.0000

sigma^2 estimated as 1.859548e-05 on 435 degrees of freedom

AIC = -7.985229  AICc = -7.98448  BIC = -7.902064
```



This model is looking excellent. All but one of the coefficients are statistically significant and since it is a middle coefficient we will not be worried about it. We can't expect to find a perfect model and this one looks much better than the other ones we found.

Fitting and Diagnostics

We now have a candidate model that we are looking to use. To make sure this is the right model we should compare the AIC and BIC values of our model with other models that we found to make sure it's the right model. We are looking for the lowest AIC and BIC values attached to our model.

```
(1,1,0)x(0,1,1): AIC = -7.935236  AICc = -7.935175  BIC = -7.907514

(5,1,0)x(0,1,1): AIC = -7.976655  AICc = -7.97622  BIC = -7.911971

(0,1,1)x(0,1,1): AIC = -6.918081  AICc = -6.918019  BIC = -6.890359

(0,1,3)x(0,1,1): AIC = -7.618456  AICc = -7.618249  BIC = -7.572253

(5,1,3)x(0,1,1): AIC = -7.982346  AICc = -7.981408  BIC = -7.88994

(4,1,3)x(0,1,1): AIC = -7.985229  AICc = -7.98448  BIC = -7.902064
```

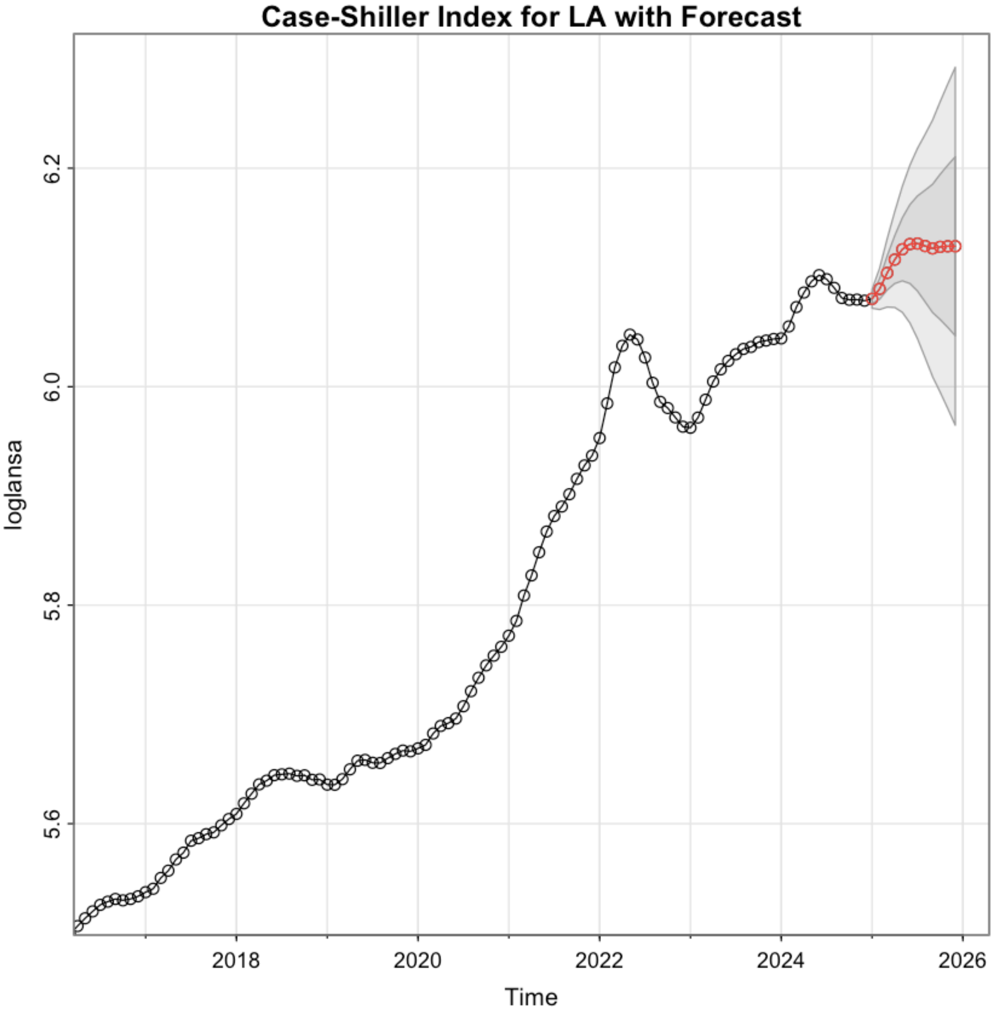
The lowest AIC value is the AIC of our chosen model, which is what we want. The first two models have a lower BIC, however. This might cause some concern but this is minimally important. They are close enough to each other while we have ample evidence we should be choosing our model over the first two based on coefficient significance and residual diagnostics. Our model has one coefficient that is not significant but we are willing to trade that off for the positives it has.

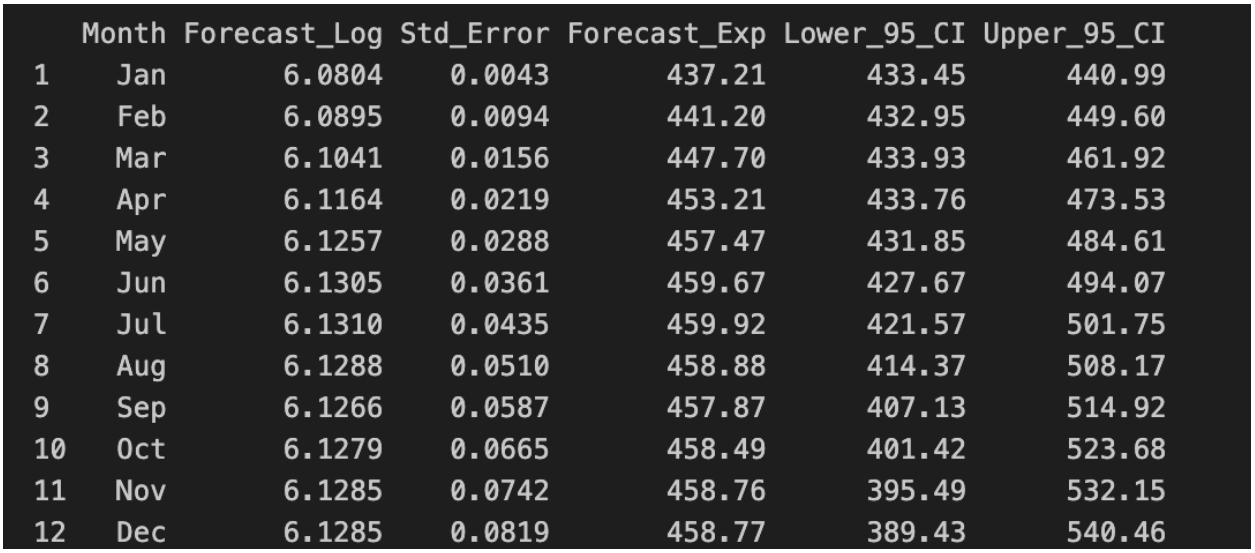
Forecasting

This is the equation of our final model with which we shall forecast. The left side is our new data point and the right side is out current data we are using to calculate the new data point.

$$
\begin{aligned}
x_t = {} & (-0.3433)\, x_{t-1} + (0.0123)\, x_{t-2} + (0.4319)\, x_{t-3} + (0.5801)\, x_{t-4} \\
& + (1.2893)\, \varepsilon_{t-1} + (1.2555)\, \varepsilon_{t-2} + (0.5588)\, \varepsilon_{t-3} \\
& - (0.7782)\, \varepsilon_{t-12} + \varepsilon_t
\end{aligned}
$$

This is our equation but we will have R actually do our forecasting for us.





```
   Month Forecast_Log Std_Error Forecast_Exp Lower_95_CI Upper_95_CI
1    Jan       6.0804    0.0043       437.21      433.45      440.99
2    Feb       6.0895    0.0094       441.20      432.95      449.60
3    Mar       6.1041    0.0156       447.70      433.93      461.92
4    Apr       6.1164    0.0219       453.21      433.76      473.53
5    May       6.1257    0.0288       457.47      431.85      484.61
6    Jun       6.1305    0.0361       459.67      427.67      494.07
7    Jul       6.1310    0.0435       459.92      421.57      501.75
8    Aug       6.1288    0.0510       458.88      414.37      508.17
9    Sep       6.1266    0.0587       457.87      407.13      514.92
10   Oct       6.1279    0.0665       458.49      401.42      523.68
11   Nov       6.1285    0.0742       458.76      395.49      532.15
12   Dec       6.1285    0.0819       458.77      389.43      540.46
```

We see in our graph that our predicted data does have an upward trend like we expected. In fact, the real data will have an even sharper upward trend since we were predicting the log transform of our original data. In the table, the actual predicted values are listed under the Forecast_Exp column since we had to exponentiate all our values to “un-log” them. You can also see we calculated the confidence value for ear of our predicted terms.

Discussion

Overall, our data modeling was a success. Often, modeling real world data gets very messy and our model is pretty close to the actual data, and it contains a lot of coefficients so it won’t be underfit. The main takeaway is that LA housing is not getting any cheaper which should surprise nobody. If LA housing prices suddenly start declining then some outside force, which our model couldn’t handle, is affecting it. It is part of the issues with our model: it basically has no ability to predict big swings in variance like the housing market crash or COVID. Perhaps some other type of model maybe me absorptive to these strange events happening. I mentioned this before, but this model also has issues predicting prices for ALL Los Angeles housing since it does not take into account new development, which is going to be a big part of LA’s future. This model is a good place to start if you’re interested at all in Los Angeles real estate.

Bibliography

S&P Dow Jones Indices LLC, S&P CoreLogic Case-Shiller CA-Los Angeles Home Price Index [LXXRSA], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/LXXRSA, March 20, 2025.

Appendix

https://github.com/card28/ATS-Final-project/blob/main/ATSfinalcode1.ipynb