MKT6971

Time Series Project

Name: **Rudy Martinez**

The objective of this project is for you to practice what you have learned about time series analysis and interpreting data. I suggest you use GRETL for this project.

1. Select a scientific, biomedical, business or other issue that appeals to you and go looking online for relevant time series data sets. The good news here is that there are tons of free and interesting time series data sets online. If you have problems locating them let me know and I will help. **Be sure that it looks like there is little or no seasonality to it.**

**Kaggle Dataset:** [House Property Sales Time Series](https://www.kaggle.com/htagholdings/property-sales)

This is a dataset (raw\_sales.csv) that contains house sales data ranging from 2007-2019. I wrote the below python script to create a subset of the data that focuses on 3-bedroom house sales and converts the “datesold” column to datetime. I then made additional transformations to the data file to display the monthly average sale price over time.

Graphical user interface

Description automatically generated

1. Plot out your time series variable. Tell me using your Mark I eyeball whether or not you think the time series data set is stationary in terms of **constant mean** and also **constant variance**.



Using Mark I Eyeball, if we draw an imaginary line through the center of the data, we can see that there is a trend / there is a **non-constant mean**. This also implies that the **variance is not constant**.

1. Plot the ACF for the time series data set. Looking at ACF, does it look like there may be a trend or non-constant mean for each time series?

****

It looks like the ACF exhibits a ski slope behavior which indicates that there is a trend / non-constant mean.

Autocorrelation function for averageprice

\*\*\*, \*\*, \* indicate significance at the 1%, 5%, 10% levels

using standard error 1/T^0.5

LAG ACF PACF Q-stat. [p-value]

1 0.8050 \*\*\* 0.8050 \*\*\* 99.1674 [0.000]

2 0.7463 \*\*\* 0.2791 \*\*\* 184.9714 [0.000]

3 0.7370 \*\*\* 0.2440 \*\*\* 269.2113 [0.000]

4 0.6710 \*\*\* -0.0181 339.5256 [0.000]

5 0.6310 \*\*\* 0.0221 402.1406 [0.000]

6 0.6262 \*\*\* 0.0972 464.2343 [0.000]

7 0.6235 \*\*\* 0.1225 526.2115 [0.000]

8 0.6138 \*\*\* 0.0755 586.7036 [0.000]

9 0.5612 \*\*\* -0.1121 637.6244 [0.000]

10 0.5628 \*\*\* 0.0583 689.2022 [0.000]

11 0.5328 \*\*\* -0.0332 735.7618 [0.000]

12 0.5107 \*\*\* 0.0461 778.8593 [0.000]

13 0.4945 \*\*\* -0.0128 819.5523 [0.000]

14 0.4916 \*\*\* 0.0432 860.0685 [0.000]

15 0.4614 \*\*\* -0.0528 896.0261 [0.000]

16 0.4819 \*\*\* 0.1301 935.5431 [0.000]

17 0.4773 \*\*\* 0.0328 974.6010 [0.000]

18 0.4495 \*\*\* -0.0453 1009.4941 [0.000]

19 0.4424 \*\*\* -0.0006 1043.5594 [0.000]

20 0.4197 \*\*\* -0.0562 1074.4506 [0.000]

21 0.3719 \*\*\* -0.0799 1098.8983 [0.000]

1. Now let’s examine the time series data set using unit root tests. First use the KPSS test for the time series data set and tell me if the test suggests if there is a constant mean or not. Then see if you can confirm your KPSS evaluation using the Augmented Dickey Fuller (ADF) or the ADF-GLS test and tell me what the ADF test suggests is the case.

KPSS test for averageprice

T = 150

Lag truncation parameter = 4

Test statistic = 2.67319

10% 5% 1%

Critical values: 0.349 0.462 0.737

P-value < .01

**Null:** There is a constant mean / no trend.

**Alternative:** There is a non-constant mean / there is a trend.

From the results, the p - value is less than 0.01 which is less than the significance value of 0.05; therefore, we can reject the null hypothesis and conclude that there is a **non-constant mean / there is a trend.**

Augmented Dickey-Fuller test for averageprice

testing down from 13 lags, criterion AIC

sample size 143

unit-root null hypothesis: a = 1

test with constant

including 6 lags of (1-L)averageprice

model: (1-L)y = b0 + (a-1)\*y(-1) + ... + e

estimated value of (a - 1): -0.0442435

test statistic: tau\_c(1) = -1.03145

asymptotic p-value 0.7442

1st-order autocorrelation coeff. for e: -0.044

lagged differences: F(6, 135) = 9.296 [0.0000]

**Null:** There is a non-constant mean / there is a trend.

**Alternative:** There is a constant mean / no trend.

From the results above, the P-value is 0.7442 which is greater than the significance value 0.05; therefore, we cannot reject the null hypothesis and can conclude that there is a **trend thus non-constant mean.**

1. Summarize the results of steps 2 through 4 and tell what your decision is regarding constant mean in the time series data set.

In summary, the Mark I Eyeball, the correlogram, the KPSS test, and the ADF test all imply that there is a non-constant mean / there is a trend. Therefore, we must perform differencing to find out if the trend recedes.

Graphical user interface, text, application, email

Description automatically generated

1. Review the decision in step #5. If the test suggests that there is a non-constant mean then use differencing to create a new differenced variable for the time series data set. Otherwise proceed to steps a through c below:
   1. Plot out the data for the new differenced data set. Tell me if it looks like the differencing got rid of the trend or non-constant mean.



The plot for the new differenced dataset is showed above. The plot seems to have gotten rid of the trend or non-constant mean by Mark I Eyeball.

* 1. Plot the ACF for the differenced time series. Tell me if this new ACF plot looks like there now is no trend.

The ski slope behavior has definitely been reduced. There are also flip flops in the plot. This indicates that there is no longer a trend / there is a constant mean.



Autocorrelation function for d\_averageprice

\*\*\*, \*\*, \* indicate significance at the 1%, 5%, 10% levels

using standard error 1/T^0.5

LAG ACF PACF Q-stat. [p-value]

1 -0.4213 \*\*\* -0.4213 \*\*\* 26.9879 [0.000]

2 -0.0900 -0.3253 \*\*\* 28.2279 [0.000]

3 0.1140 -0.0997 30.2315 [0.000]

4 -0.0701 -0.1085 30.9943 [0.000]

5 0.0719 0.0242 31.8030 [0.000]

6 -0.0745 -0.0577 32.6764 [0.000]

7 0.0380 0.0012 32.9056 [0.000]

8 0.0990 0.1256 34.4683 [0.000]

9 -0.1687 \*\* -0.0529 39.0419 [0.000]

10 0.1139 0.0459 41.1428 [0.000]

11 -0.0358 -0.0137 41.3515 [0.000]

12 -0.0406 -0.0391 41.6216 [0.000]

13 0.0222 -0.0610 41.7029 [0.000]

14 0.0331 0.0311 41.8851 [0.000]

15 -0.1326 -0.1703 \*\* 44.8381 [0.000]

16 0.0655 -0.0856 45.5633 [0.000]

17 0.0360 -0.0117 45.7845 [0.000]

18 -0.0820 -0.0942 46.9384 [0.000]

19 0.0543 -0.0057 47.4485 [0.000]

20 0.0708 0.1238 48.3229 [0.000]

21 -0.1328 -0.0494 51.4217 [0.000]

* 1. Apply the KPSS test and the ADF or ADF-GLS test to the differenced data – does the trend disappear?

KPSS test for d\_averageprice

T = 149

Lag truncation parameter = 4

Test statistic = 0.10825

10% 5% 1%

Critical values: 0.349 0.462 0.737

P-value > .10

**Null:** There is a constant mean / no trend.

**Alternative:** There is a non-constant mean / there is a trend.

From the results, the p - value is greater than 0.1 which is greater than the significance value of 0.05; therefore, we do not reject the null hypothesis and conclude that there is a **constant mean / there is not a trend.**

Augmented Dickey-Fuller test for d\_averageprice

testing down from 13 lags, criterion AIC

sample size 143

unit-root null hypothesis: a = 1

test with constant

including 5 lags of (1-L)d\_averageprice

model: (1-L)y = b0 + (a-1)\*y(-1) + ... + e

estimated value of (a - 1): -2.93539

test statistic: tau\_c(1) = -7.9433

asymptotic p-value 7.324e-13

1st-order autocorrelation coeff. for e: -0.051

lagged differences: F(5, 136) = 5.642 [0.0001]

**Null:** There is a non-constant mean / there is a trend.

**Alternative:** There is a constant mean / no trend.

From the results above, the P-value is 7.324e-13 which is smaller than the significance value 0.05; therefore, we reject the null hypothesis and can conclude that there is not a **trend thus a constant mean.**

**Note: From this point onward through step 9, if the time series was differenced, use the differenced time series data set for all the rest of the questions. Otherwise you can use the undifferenced data set.**

1. Plot the PACF for the time series data set. Using the combined information from the ACF you plotted earlier along with the information in the PACF, tell me if you see autoregressive and/or moving average processes in the data set. To help with interpretation you may want to refer to online resources – here is a decent resource from Duke University [**https://people.duke.edu/~rnau/411arim3.htm**](https://people.duke.edu/~rnau/411arim3.htm) or Penn State https://onlinecourses.science.psu.edu/stat510/node/64



From the PACF plot, the green bars do not go above the red significance line (positive). Therefore, we do not see any evidence of autoregressive processes. The green bars do go below the red significance line (negative). This suggests moving average processes at lags of order 1, 2, and 15.

1. For your time series data set, experiment with different ARIMA models for them. Try at least four models. As you try them, list out the results of the various models and
   1. Comment on how each one is working and compare it to the previous model using various metrics such as AIC, BIC (Schwartz) and Box Leung Most students end up creating a small table with these statistics across the models tried so it is easy to compare them.

**MODEL 1: ARIMA (1,0,2)**

Model 2: ARMA, using observations 2007:03-2019:07 (T = 149)

Dependent variable: d\_averageprice

Standard errors based on Hessian

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Coefficient* | *Std. Error* | *z* | *p-value* |  |
| const | 1916.73 | 852.272 | 2.249 | 0.0245 | \*\* |
| phi\_1 | −0.611743 | 0.276731 | −2.211 | 0.0271 | \*\* |
| theta\_1 | −0.00272805 | 0.268980 | −0.01014 | 0.9919 |  |
| theta\_2 | −0.477563 | 0.162953 | −2.931 | 0.0034 | \*\*\* |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mean dependent var | 2316.934 |  | S.D. dependent var | 37663.89 |
| Mean of innovations | 760.8480 |  | S.D. of innovations | 31594.51 |
| R-squared | 0.292221 |  | Adjusted R-squared | 0.282525 |
| Log-likelihood | −1755.470 |  | Akaike criterion | 3520.941 |
| Schwarz criterion | 3535.960 |  | Hannan-Quinn | 3527.043 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | *Real* | *Imaginary* | *Modulus* | *Frequency* |
| AR |  |  |  |  |  |
|  | Root 1 | -1.6347 | 0.0000 | 1.6347 | 0.5000 |
| MA |  |  |  |  |  |
|  | Root 1 | -1.4499 | 0.0000 | 1.4499 | 0.5000 |
|  | Root 2 | 1.4442 | 0.0000 | 1.4442 | 0.0000 |

**MODEL 1 Interpretation**

Since the p-value for phi\_1 is < 0.05, we can say that we found the first order auto regressive process in the model. Since the p-value for theta\_1 is not significant, we do not find moving average in the model. Since the p-value for theta\_2 is < 0.05, we can say that we found moving average in the model.

|  |  |
| --- | --- |
| Adjusted R-squared | 0.282525 |
| Akaike criterion | 3520.941 |
| Schwarz criterion | 3535.960 |
| Hannan-Quinn | 3527.043 |

**Ljung-Box:**

Test for autocorrelation up to order 12

Ljung-Box Q' = 8.20132,

with p-value = P(Chi-square(9) > 8.20132) = 0.5140

**Null:** There is no evidence of serial autocorrelation in the residuals

**Alternative:** There is serial autocorrelation in the residuals.

Because the p-value is greater than 0.05, we accept the null; therefore, there is no evidence of serial autocorrelation.

**MODEL 2: ARIMA (1,0,3)**

Model 4: ARMA, using observations 2007:03-2019:07 (T = 149)

Dependent variable: d\_averageprice

Standard errors based on Hessian

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Coefficient* | *Std. Error* | *z* | *p-value* |  |
| const | 1980.72 | 956.244 | 2.071 | 0.0383 | \*\* |
| phi\_1 | −0.452474 | 0.383791 | −1.179 | 0.2384 |  |
| theta\_1 | −0.176762 | 0.384123 | −0.4602 | 0.6454 |  |
| theta\_2 | −0.372863 | 0.233033 | −1.600 | 0.1096 |  |
| theta\_3 | 0.0773671 | 0.0974010 | 0.7943 | 0.4270 |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mean dependent var | 2316.934 |  | S.D. dependent var | 37663.89 |
| Mean of innovations | 556.9573 |  | S.D. of innovations | 31540.00 |
| R-squared | 0.294232 |  | Adjusted R-squared | 0.279630 |
| Log-likelihood | −1755.202 |  | Akaike criterion | 3522.405 |
| Schwarz criterion | 3540.429 |  | Hannan-Quinn | 3529.728 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | *Real* | *Imaginary* | *Modulus* | *Frequency* |
| AR |  |  |  |  |  |
|  | Root 1 | -2.2101 | 0.0000 | 2.2101 | 0.5000 |
| MA |  |  |  |  |  |
|  | Root 1 | 4.7238 | 0.0000 | 4.7238 | 0.0000 |
|  | Root 2 | -1.6071 | 0.0000 | 1.6071 | 0.5000 |
|  | Root 3 | 1.7026 | 0.0000 | 1.7026 | 0.0000 |

**MODEL 2 Interpretation**

Since the p-values for phi\_1, theta\_1, theta\_2, and theta\_3 are above the significance level of 0.05, we do not find moving average in the model for first, second, and third order.

|  |  |
| --- | --- |
| Adjusted R-squared | 0.279630 |
| Akaike criterion | 3522.405 |
| Schwarz criterion | 3540.429 |
| Hannan-Quinn | 3529.728 |

**Ljung-Box:**

Test for autocorrelation up to order 12

Ljung-Box Q' = 6.45212,

with p-value = P(Chi-square(8) > 6.45212) = 0.5967

**Null:** There is no evidence of serial autocorrelation in the residuals

**Alternative:** There is serial autocorrelation in the residuals.

Because the p-value is greater than 0.05, we accept the null; therefore, there is no evidence of serial autocorrelation.

**MODEL 3: ARIMA (2,0,1)**

Model 5: ARMA, using observations 2007:03-2019:07 (T = 149)

Dependent variable: d\_averageprice

Standard errors based on Hessian

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Coefficient* | *Std. Error* | *z* | *p-value* |  |
| const | 2008.64 | 1001.11 | 2.006 | 0.0448 | \*\* |
| phi\_1 | −0.117450 | 0.216132 | −0.5434 | 0.5868 |  |
| phi\_2 | −0.143346 | 0.140057 | −1.023 | 0.3061 |  |
| theta\_1 | −0.519274 | 0.215969 | −2.404 | 0.0162 | \*\* |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mean dependent var | 2316.934 |  | S.D. dependent var | 37663.89 |
| Mean of innovations | 482.8497 |  | S.D. of innovations | 31614.06 |
| R-squared | 0.290873 |  | Adjusted R-squared | 0.281159 |
| Log-likelihood | −1755.542 |  | Akaike criterion | 3521.084 |
| Schwarz criterion | 3536.104 |  | Hannan-Quinn | 3527.187 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | *Real* | *Imaginary* | *Modulus* | *Frequency* |
| AR |  |  |  |  |  |
|  | Root 1 | -0.4097 | -2.6093 | 2.6412 | -0.2748 |
|  | Root 2 | -0.4097 | 2.6093 | 2.6412 | 0.2748 |
| MA |  |  |  |  |  |
|  | Root 1 | 1.9258 | 0.0000 | 1.9258 | 0.0000 |

**MODEL 3 Interpretation**

Since the p-values for phi\_1 and phi\_2 are above the significance level of 0.05, we do not find moving average in the model for first and second order. Since the p-values for theta\_1 is < 0.05 we can say that we found moving average in the model.

|  |  |
| --- | --- |
| Adjusted R-squared | 0.281159 |
| Akaike criterion | 3521.084 |
| Schwarz criterion | 3536.104 |
| Hannan-Quinn | 3527.187 |

**Ljung-Box:**

Test for autocorrelation up to order 12

Ljung-Box Q' = 7.00182,

with p-value = P(Chi-square(9) > 7.00182) = 0.6369

**Null:** There is no evidence of serial autocorrelation in the residuals

**Alternative:** There is serial autocorrelation in the residuals.

Because the p-value is greater than 0.05, we accept the null; therefore, there is no evidence of serial autocorrelation.

**MODEL 4: ARIMA (1,0,1)**

Model 7: ARMA, using observations 2007:03-2019:07 (T = 149)

Dependent variable: d\_averageprice

Standard errors based on Hessian

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Coefficient* | *Std. Error* | *z* | *p-value* |  |
| const | 1923.72 | 865.948 | 2.222 | 0.0263 | \*\* |
| phi\_1 | 0.0471738 | 0.141005 | 0.3346 | 0.7380 |  |
| theta\_1 | −0.690004 | 0.117521 | −5.871 | <0.0001 | \*\*\* |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mean dependent var | 2316.934 |  | S.D. dependent var | 37663.89 |
| Mean of innovations | 750.4904 |  | S.D. of innovations | 31711.81 |
| R-squared | 0.286726 |  | Adjusted R-squared | 0.281874 |
| Log-likelihood | −1756.015 |  | Akaike criterion | 3520.031 |
| Schwarz criterion | 3532.046 |  | Hannan-Quinn | 3524.912 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | *Real* | *Imaginary* | *Modulus* | *Frequency* |
| AR |  |  |  |  |  |
|  | Root 1 | 21.1982 | 0.0000 | 21.1982 | 0.0000 |
| MA |  |  |  |  |  |
|  | Root 1 | 1.4493 | 0.0000 | 1.4493 | 0.0000 |

**MODEL 4 Interpretation**

Since the p-value for phi\_1 is above the significance level of 0.05, we do not find moving average in the model for first order. Since the p-values for theta\_1 is < 0.05 we can say that we found moving average in the model.

|  |  |
| --- | --- |
| Adjusted R-squared | 0.281874 |
| Akaike criterion | 3520.031 |
| Schwarz criterion | 3532.046 |
| Hannan-Quinn | 3524.912 |

**Ljung-Box:**

Test for autocorrelation up to order 12

Ljung-Box Q' = 9.01238,

with p-value = P(Chi-square(10) > 9.01238) = 0.5309

**Null:** There is no evidence of serial autocorrelation in the residuals

**Alternative:** There is serial autocorrelation in the residuals.

Because the p-value is greater than 0.05, we accept the null; therefore, there is no evidence of serial autocorrelation.

* 1. Plot the observed versus fitted data for the time series data set **for each model** and comment on how well the model seems to be working

**MODEL 1**

****

**MODEL 2**

****

**MODEL 3**

****

**MODEL 4**

****

* 1. Pick one of the models as your favorite and tell me why you like that one the best.

I will choose the MODEL 1 – ARIMA (1,0,2) as the best model because the adjusted R² is the highest of all.

* 1. Use the ARCH test to tell me if your best model has any issues with autoregressive conditional heteroscedasticity.

**Null**: There is no evidence of non-constant variance across time.

**Alternative**: There is non-constant variance across time.

Test for ARCH of order 12

coefficient std. error t-ratio p-value

------------------------------------------------------------

alpha(0) 8.05093e+08 3.02196e+08 2.664 0.0087 \*\*\*

alpha(1) −0.0429082 0.0893038 −0.4805 0.6317

alpha(2) 0.0210426 0.0893623 0.2355 0.8142

alpha(3) −0.000276371 0.0893351 −0.003094 0.9975

alpha(4) −0.0423561 0.0892880 −0.4744 0.6361

alpha(5) 0.0458817 0.0893942 0.5133 0.6087

alpha(6) −0.0408619 0.0880376 −0.4641 0.6434

alpha(7) −0.0287278 0.0878330 −0.3271 0.7442

alpha(8) −0.00913124 0.0644259 −0.1417 0.8875

alpha(9) −0.0320904 0.0642563 −0.4994 0.6184

alpha(10) −0.0121486 0.0634596 −0.1914 0.8485

alpha(11) 0.0179689 0.0623834 0.2880 0.7738

alpha(12) 0.0767981 0.0622983 1.233 0.2200

Null hypothesis: no ARCH effect is present

Test statistic: LM = 3.25676

with p-value = P(Chi-square(12) > 3.25676) = 0.993444

The p-value is 0.993444 which is more than the significance level of 0.05. Therefore, we do not reject the null and conclude that there is no evidence on non-constant variance across time.