DA6823

Time Series Project

Name: Collin Real (yhi267)

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. **Be sure that you cut and paste your answers to each of the questions for the project. If you talk about something in a table or plot, that table or plot needs to be in your report!!! If the question says plot something, cut and paste that plot into your report.** In previous semesters I have had students talk about the plot but not display it – that makes no sense.

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**. Note that you should avoid time series data sets that have huge spikes in them (they are hard to model using GRETL) and also avoid data sets where the data plot looks like a straight line going up or down – those aren’t very interesting.

A green line graph with numbers

Description automatically generated

**It looks like the mean is trending upwards, suggesting that the data is non-stationary.**

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?

A graph with green and orange lines

Description automatically generated

**The ACF plot illustrates a consistent trend, further suggesting non-stationary data.**

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:** Since the p-value (< 0.01) is less than the significance level (assuming 0.05), we reject the null hypothesis and conclude a non-constant mean.

**ADF test:** Since the p-value (< 0.01) is less than the significance level (assuming 0.05), we reject the null hypothesis and conclude a constant mean.

A white paper with black text

Description automatically generated A screenshot of a computer code

Description automatically generated

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

**Both of our plots showed evidence of a trend in the data. The KPSS test indicated evidence of a non-constant mean, but the ADF test suggested the opposite; therefore, I need to difference the nat\_demand variable.**

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 and proceed to the steps below (a,b,c). Be sure to cut and paste your supporting evidence (unit root tests, plots, etc.) below.** If you got luck and concluded that your data set already has a constant mean then you can skip all of step 6 and move on using your data set without differencing!
   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.

A green and white sound wave

Description automatically generated

**The differenced dataset looks like it got rid of the trend.**

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

A graph with green and red lines

Description automatically generated

**The ski slope pattern is gone, and the spikes are alternating between positive and negative correlations, suggesting a constant mean.**

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

A screenshot of a computer

Description automatically generated A white text with black text

Description automatically generated

**Both tests suggest evidence of a constant mean, so the trend disappeared.**

**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 any autoregressive and/or moving average processes in the data set and what they are. Use the discussion in class as well as 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>

**A graph with green and orange lines

Description automatically generated**

**Since there are significant spikes for both positive and negative correlations, our PACF plot suggests both moving average and autoregressive processes. I believe there are at least first and second order processes for both AR and MA processes.**

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

**Model 1: ARMA, using observations 2015-01-04:2020-06-26 (T = 2001)**

**Function evaluations: 127**

**Evaluations of gradient: 34**

**Estimated using AS 197 (exact ML)**

**Dependent variable: d\_nat\_demand**

**Standard errors based on Hessian**

**coefficient std. error z p-value**

**-----------------------------------------------------------**

**const 0.278992 5.39860 0.05168 0.9588**

**phi\_1 −0.693421 0.0290346 −23.88 4.64e-126 \*\*\***

**theta\_1 0.412033 0.0291849 14.12 2.94e-45 \*\*\***

**theta\_2 −0.795469 0.0207173 −38.40 0.0000 \*\*\***

**theta\_3 −0.763683 0.0298222 −25.61 1.25e-144 \*\*\***

**theta\_4 0.00717183 0.0246889 0.2905 0.7714**

**theta\_5 0.389563 0.0207332 18.79 9.23e-79 \*\*\***

**Mean dependent var 0.978277 S.D. dependent var 2055.406**

**Mean of innovations 1.234910 S.D. of innovations 1633.185**

**R-squared 0.369208 Adjusted R-squared 0.367627**

**Log-likelihood −17644.44 Akaike criterion 35304.89**

**Schwarz criterion 35349.70 Hannan-Quinn 35321.34**

**Real Imaginary Modulus Frequency**

**-----------------------------------------------------------**

**AR**

**Root 1 -1.4421 0.0000 1.4421 0.5000**

**MA**

**Root 1 -1.0926 0.0000 1.0926 0.5000**

**Root 2 1.3528 0.0000 1.3528 0.0000**

**Root 3 -0.7500 0.9271 1.1925 0.3582**

**Root 4 -0.7500 -0.9271 1.1925 -0.3582**

**Root 5 1.2213 0.0000 1.2213 0.0000**

**-----------------------------------------------------------**

**Test for autocorrelation up to order 7**

**Ljung-Box Q' = 415.751,**

**with p-value = P(Chi-square(1) > 415.751) = 2.053e-92**

**Model 2: ARMA, using observations 2015-01-04:2020-06-26 (T = 2001)**

**Function evaluations: 83**

**Evaluations of gradient: 26**

**Estimated using AS 197 (exact ML)**

**Dependent variable: d\_nat\_demand**

**Standard errors based on Hessian**

**coefficient std. error z p-value**

**-----------------------------------------------------------**

**const 0.207649 3.54132 0.05864 0.9532**

**phi\_1 −0.478648 0.0460433 −10.40 2.60e-25 \*\*\***

**phi\_2 −0.524760 0.0422794 −12.41 2.26e-35 \*\*\***

**theta\_1 0.0535882 0.0470861 1.138 0.2551**

**theta\_2 −0.145956 0.0449409 −3.248 0.0012 \*\*\***

**theta\_3 −0.716886 0.0204646 −35.03 7.69e-269 \*\*\***

**Mean dependent var 0.978277 S.D. dependent var 2055.406**

**Mean of innovations 1.492703 S.D. of innovations 1654.371**

**R-squared 0.355472 Adjusted R-squared 0.354180**

**Log-likelihood −17670.24 Akaike criterion 35354.49**

**Schwarz criterion 35393.70 Hannan-Quinn 35368.89**

**Real Imaginary Modulus Frequency**

**-----------------------------------------------------------**

**AR**

**Root 1 -0.4561 -1.3029 1.3804 -0.3036**

**Root 2 -0.4561 1.3029 1.3804 0.3036**

**MA**

**Root 1 -0.6390 -0.9434 1.1394 -0.3448**

**Root 2 -0.6390 0.9434 1.1394 0.3448**

**Root 3 1.0744 0.0000 1.0744 0.0000**

**-----------------------------------------------------------**

**Test for autocorrelation up to order 7**

**Ljung-Box Q' = 595.485,**

**with p-value = P(Chi-square(2) > 595.485) = 4.921e-130**

**Model 3: ARMA, using observations 2015-01-04:2020-06-26 (T = 2001)**

**Function evaluations: 118**

**Evaluations of gradient: 30**

**Estimated using AS 197 (exact ML)**

**Dependent variable: d\_nat\_demand**

**Standard errors based on Hessian**

**coefficient std. error z p-value**

**-----------------------------------------------------------**

**const 0.210333 3.79336 0.05545 0.9558**

**phi\_1 −0.574194 0.0429400 −13.37 8.82e-41 \*\*\***

**theta\_1 0.254302 0.0451158 5.637 1.73e-08 \*\*\***

**theta\_2 −0.787072 0.0209187 −37.63 8.30e-310 \*\*\***

**theta\_3 −0.620655 0.0334366 −18.56 6.51e-77 \*\*\***

**theta\_4 0.130163 0.0311904 4.173 3.00e-05 \*\*\***

**theta\_5 0.327453 0.0201257 16.27 1.60e-59 \*\*\***

**theta\_6 −0.139848 0.0197536 −7.080 1.45e-12 \*\*\***

**Mean dependent var 0.978277 S.D. dependent var 2055.406**

**Mean of innovations 1.337233 S.D. of innovations 1614.888**

**R-squared 0.383748 Adjusted R-squared 0.381893**

**Log-likelihood −17622.02 Akaike criterion 35262.04**

**Schwarz criterion 35312.45 Hannan-Quinn 35280.55**

**Real Imaginary Modulus Frequency**

**-----------------------------------------------------------**

**AR**

**Root 1 -1.7416 0.0000 1.7416 0.5000**

**MA**

**Root 1 -1.1034 0.0000 1.1034 0.5000**

**Root 2 1.0856 0.0000 1.0856 0.0000**

**Root 3 -0.7276 0.8853 1.1459 0.3595**

**Root 4 -0.7276 -0.8853 1.1459 -0.3595**

**Root 5 1.9072 -0.9533 2.1322 -0.0738**

**Root 6 1.9072 0.9533 2.1322 0.0738**

**-----------------------------------------------------------**

**Test for autocorrelation up to order 8**

**Ljung-Box Q' = 403.086,**

**with p-value = P(Chi-square(1) > 403.086) = 1.173e-89**

**Model 4: ARMA, using observations 2015-01-04:2020-06-26 (T = 2001)**

**Function evaluations: 74**

**Evaluations of gradient: 17**

**Estimated using AS 197 (exact ML)**

**Dependent variable: d\_nat\_demand**

**Standard errors based on Hessian**

**coefficient std. error z p-value**

**-----------------------------------------------------------**

**const 0.481805 7.80315 0.06174 0.9508**

**phi\_1 −0.894778 0.0193217 −46.31 0.0000 \*\*\***

**phi\_2 −0.780006 0.0196381 −39.72 0.0000 \*\*\***

**phi\_3 −0.764313 0.0212115 −36.03 2.54e-284 \*\*\***

**phi\_4 −0.715536 0.0207969 −34.41 2.06e-259 \*\*\***

**phi\_5 −0.733649 0.0195476 −37.53 2.84e-308 \*\*\***

**phi\_6 −0.712357 0.0157143 −45.33 0.0000 \*\*\***

**theta\_1 0.498083 0.0219558 22.69 6.21e-114 \*\*\***

**Mean dependent var 0.978277 S.D. dependent var 2055.406**

**Mean of innovations 0.459457 S.D. of innovations 1304.099**

**R-squared 0.597244 Adjusted R-squared 0.596032**

**Log-likelihood −17195.35 Akaike criterion 34408.70**

**Schwarz criterion 34459.11 Hannan-Quinn 34427.21**

**Real Imaginary Modulus Frequency**

**-----------------------------------------------------------**

**AR**

**Root 1 -0.9639 -0.4438 1.0612 -0.4313**

**Root 2 -0.9639 0.4438 1.0612 0.4313**

**Root 3 0.6787 -0.8044 1.0525 -0.1385**

**Root 4 0.6787 0.8044 1.0525 0.1385**

**Root 5 -0.2297 -1.0357 1.0609 -0.2847**

**Root 6 -0.2297 1.0357 1.0609 0.2847**

**MA**

**Root 1 -2.0077 0.0000 2.0077 0.5000**

**-----------------------------------------------------------**

**Test for autocorrelation up to order 8**

**Ljung-Box Q' = 149.306,**

**with p-value = P(Chi-square(1) > 149.306) = 2.458e-34**

* 1. Construct a table with the identity of the model, the R square, the AIC, BIC(Schwartz), the Hannan-Quinn, Lejune-Box and a final column that notes the terms that are significant in the model. **Be sure to paste that table into your project report!**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Adj. R Square** | **AIC** | **BIC** | **Hannan-Quinn** | **Ljung-Box** | **Significant Terms** |
| (1,0,5) | 0.367627 | 35304.89 | 35349.7 | 35321.34 | 2.05E-92 | phi\_1, theta\_1, theta\_2, theta\_3, theta\_5 |
| (2,0,3) | 0.35418 | 35354.49 | 35393.7 | 35368.89 | 4.92E-130 | phi\_1, phi\_2, theta\_2, theta\_3 |
| (1,0,6) | 0.381893 | 35262.04 | 35312.45 | 35280.55 | 1.17E-89 | phi\_1, theta\_1, theta\_2, theta\_3, theta\_4, theta\_5, theta\_6 |
| (6,0,1) | 0.596032 | 34408.7 | 34459.11 | 34427.21 | 2.46E-34 | phi\_1, phi\_2, phi\_3, phi\_4, phi\_5, phi\_6, theta\_1 |

* 1. Plot the observed versus fitted data for the time series data set **for each model.**

**Model (1,0,5)**

A green and orange line graph

Description automatically generated

**Model (2,0,3)**A graph with orange and blue lines

Description automatically generated

**Model (1,0,6)**

A graph with orange and blue lines

Description automatically generated

**Model (6,0,1)**

A graph of a sound wave

Description automatically generated with medium confidence

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

**Model (6,0,1) is the best model because it has the highest adjusted R square and lowest AIC value.**

* 1. Forecast your model out 6 time periods and graph the time series including the forecast. How well does the forecast seem to work?

A graph of a graph showing the amount of time

Description automatically generated with medium confidence

**The forecasted time series looks consistent with the pre-forecasted time series values. The forecasted line follows the same pattern as the pre-forecasted line, fluctuating above and below 0.**

1. Test the time series data set you select for constant variance using the ARCH test (GRETL does this nicely). Note that we will not do anything about this issue for the moment, but it’s good to know.

**A screenshot of a computer

Description automatically generated**