DA6823

Time Series Project

Name: Emily Bates, xzz320

*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. *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.***

[Daily Births in 1959 in CA](https://www.kaggle.com/datasets/dougcresswell/daily-total-female-births-in-california-1959)

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

The mean appears to be slightly positive over time.

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 of a baby

Description automatically generated with medium confidence

The bars on the ACF plot for a stationary time series flip flop between positive and negative. A non-constant time series does not flip flop so much. My ACF plot is mostly positive without any flip flop. This suggests that the mean is non-constant– there is trend.

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 births

T = 365

Lag truncation parameter = 5

Test statistic = 1.82765

10% 5% 1%

Critical values: 0.348 0.462 0.741

P-value < .01

KPSS Null Hypothesis: There is no evidence that the data has more than one mean (it’s stationary).

KPSS Alternative Hypothesis: The data has more than one mean (it’s non-stationary)

Because the p-value is less than .05, we reject the null hypothesis, and there is evidence of more than one mean, meaning it’s non-constant.

Augmented Dickey-Fuller test for births

testing down from 16 lags, criterion AIC

sample size 358

unit-root null hypothesis: a = 1

test with constant

including 6 lags of (1-L)births

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

estimated value of (a - 1): -0.478271

test statistic: tau\_c(1) = -4.80829

asymptotic p-value 4.954e-05

1st-order autocorrelation coeff. for e: -0.011

lagged differences: F(6, 350) = 2.885 [0.0093]

Dickey Fuller null hypothesis: the data has more than one mean (it’s non-stationary)

Dickey Fuller alternative hypothesis: the data has one mean (it’s stationary).

Because my p-value is less than 0.05, we must reject the null hypothesis that the data has more than one mean, meaning the data is stationary.

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

I’ve concluded that the mean is non-constant because 3 of the 4 tests indicate a non-constant mean. These tests included the KPSS test, the ACF plot, and the time-series plot. The Augmented Dickey-Fuller test suggested that the mean could be constant, however, the other three tests outweigh the one constant mean result.

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 sound wave

Description automatically generated*

The differencing got rid of the trend in the data – the mean now looks constant.

* 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 of a number of birth rate

Description automatically generated with medium confidence*

I am seeing more flip flopping of the bars in the acf chart, meaning the mean is constant. It starts out a little weird though without flip flopping, but I think if the KPSS and Dickey-Fuller indicate a constant mean, it should be fine.

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

KPSS:

KPSS test for d\_births

T = 364

Lag truncation parameter = 5

Test statistic = 0.00946404

10% 5% 1%

Critical values: 0.348 0.462 0.741

P-value > .10

We can’t reject the null hypothesis, meaning the mean is constant.

Augmented Dickey-Fuller:

Augmented Dickey-Fuller test for d\_births

testing down from 16 lags, criterion AIC

sample size 357

unit-root null hypothesis: a = 1

test with constant

including 6 lags of (1-L)d\_births

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

estimated value of (a - 1): -4.31231

test statistic: tau\_c(1) = -12.6252

asymptotic p-value 9.719e-28

1st-order autocorrelation coeff. for e: -0.003

lagged differences: F(6, 349) = 17.583 [0.0000]

We reject the null hypothesis, meaning the mean is constant.

We can conclude that differencing the data succeeded in removing trend.

***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*](https://onlinecourses.science.psu.edu/stat510/node/64)

**A graph of a number of birth rate

Description automatically generated with medium confidence**

* I do not see any autoregressive processes in the dataset – none pass the positive 95% confidence interval line. There are multiple moving average processes in the data, specifically the first six bars.

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Adjusted R squared** | **AIC (Akaike)** | **BIC (Schwarz)** | **Hannan-Quinn** | **Lejung Box p-value** | **ARCH p-value** |
| (0,0,3) | 0.416394 | 2461.423 | 2480.909 | 2469.168 | 0.2366 | 0.194583 |
| (0,0,2) | 0.415830 | **2460.722** | 2476.311 | **2466.918** | **0.2969** | 0.268492 |
| (0,0,4) | 0.419401 | 2463.349 | 2486.732 | 2472.643 | 0.06876 | 0.147371 |
| (0,0,1) | 0.416438 | 2463.584 | **2475.275** | 2468.231 | 0.1301 | **0.438591** |
| (0,0,6) | 0.411900 | 2467.250 | 2498.427 | 2479.641 | 0.01813 | 0.179826 |
| (0,0,5) | 0.413544 | 2465.250 | 2492.530 | 2476.093 | 0.06070 | 0.179237 |
| (1,0,2) | **0.421923** | **2460.514** | 2480.000 | 2468.259 | 0.1462 | 0.212302 |
| (2,0,2) | 0.421122 | 2461.874 | 2485.256 | 2471.167 | 0.1352 | 0.260992 |

The adjusted R squared value indicates what percentage of the data can be explained by the model. The AIC,BIC, and Hannan-Quinn are all relative, so you’re looking for a lower value in comparison to the same metric from the other models. The Lejung Box p-value indicates if there is serial autocorrelation in the residuals or not (you want to fail to reject the null hypothesis – no serial autocorrelation in the residuals). The ARCH test p-value indicates if there is constant variance or not (you want to fail to reject the null hypothesis – there’s constant variance across data set)

**(**0,0,3)A graph showing a number of birth rate

Description automatically generated with medium confidence

(0,0,2)A graph showing a number of birth rate

Description automatically generated with medium confidence

(0,0,4)A graph showing a number of birth rate

Description automatically generated with medium confidence

(0,0,1) A graph showing a number of birth rate

Description automatically generated with medium confidence’

(0,0,6)A graph showing a number of birth rate

Description automatically generated with medium confidence

(0,0,5)A graph showing a number of birth rate

Description automatically generated with medium confidence

(1,0,2)A graph showing a number of birth rate

Description automatically generated with medium confidence

(2,0,2) A graph showing a number of birth rate

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.*

My favorite model is (0,0,2) because it’s tied for lowest AIC, has the lowest Hannan-Quinn, and has the lowest Lejung Box p-value. Both the Lejung box and ARCH p-values are greater than 0.05. For the Lejung Box, this means there isn’t significant serial autocorrelation in the residuals, and for the ARCH test, this means there is constant variance across the data set.

* 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 time of a baby

Description automatically generated with medium confidence

The forecast doesn’t seem to work that well since the last four time periods are equal.

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.*
2. Test for ARCH of order 7
3. coefficient std. error t-ratio p-value
4. ---------------------------------------------------------
5. alpha(0) 53.4702 8.69768 6.148 2.15e-09 \*\*\*
6. alpha(1) −0.0407045 0.0531494 −0.7658 0.4443
7. alpha(2) 0.00192709 0.0530899 0.03630 0.9711
8. alpha(3) −0.0360565 0.0530732 −0.6794 0.4974
9. alpha(4) −0.0422063 0.0530164 −0.7961 0.4265
10. alpha(5) −0.0327808 0.0530081 −0.6184 0.5367
11. alpha(6) −0.0586243 0.0530384 −1.105 0.2698
12. alpha(7) 0.119528 0.0530975 2.251 0.0250 \*\*
13. Null hypothesis: no ARCH effect is present
14. Test statistic: LM = 8.78462
15. with p-value = P(Chi-square(7) > 8.78462) = 0.268492