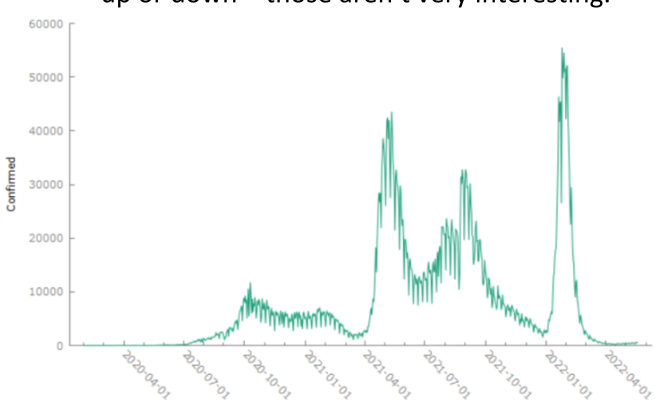
DA6823
Time Series Project

Name: Moneeb Abu-Esba

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.

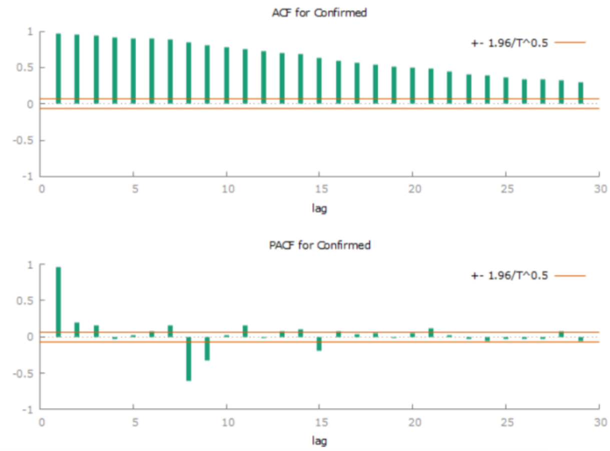
1. Select a scientific, biomedical, business or other issue that appeals to you and go looking online for relevant time series data sets. The good news here is that there are tons of free and interesting time series data sets online. If you have problems locating them let me know and I will help. **Be sure that it looks like there is little or no seasonality to it.**
   Latest Covid-19 Confirmed Cases Kerala
   https://www.kaggle.com/datasets/anandhuh/covid19-confirmed-cases-kerala/

2. 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.



The data set does not have a constant mean or variance.

3. 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?



Theres a trend in the ACF. No trend at PACF

but just use the ACF for trend diagnostics

4. 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:
T = 841
Lag truncation parameter = 6
Test statistic = 2.50575

	10%	5%	1%
Critical values:	0.348	0.462	0.743

No - hnull: constant mean, halt: nonconstant mean, so p<.05 so reje

P-value < .01
There is a constant mean as P-value is less than 0.05.

ADF:
Augmented Dickey-Fuller test for Confirmed
testing down from 20 lags, criterion AIC
sample size 820
unit-root null hypothesis: a = 1

test with constant
including 20 lags of (1-L)Confirmed
model: (1-L)y = b0 + (a-1)*y(-1) + ... + e
estimated value of (a - 1): -0.0210893
test statistic: tau_c(1) = -2.94654
asymptotic p-value 0.0402
1st-order autocorrelation coeff. for e: -0.005
lagged differences: F(20, 798) = 50.502 [0.0000]

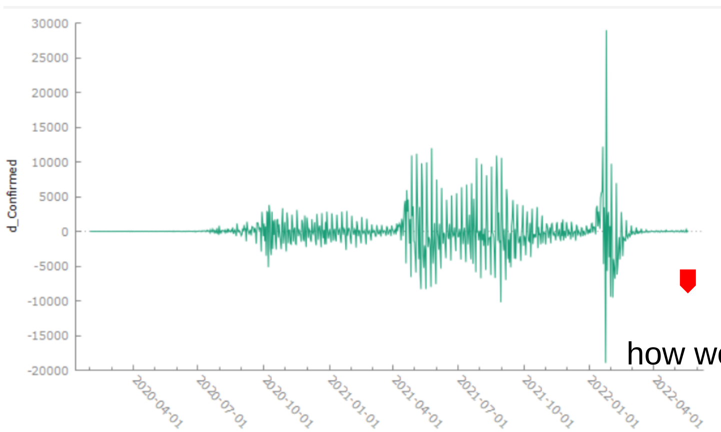
The p-value < 0.05. Accept the null hypothesis and there is evidence of multiple means.

no! p < .05 so reject null hypothesis of non-constant mean and accept alternative hypothesis of cons

5. Summarize the results of steps 2 through 4 and tell what your decision is regarding constant mean in the time series data set.
   Steps 2-4 show that there are multiple means.

6. 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!

   a. 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.

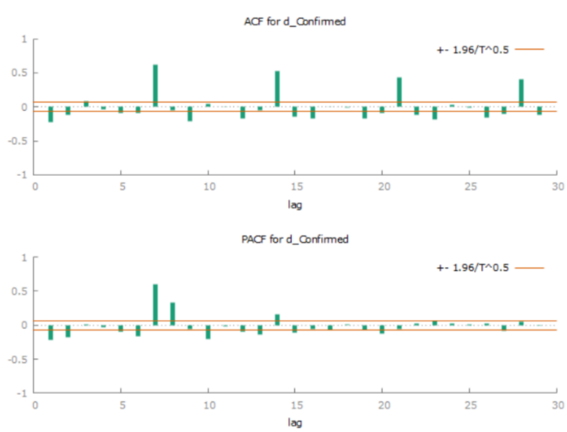


that spike is helping things here...

how would it show trend? draw a straight line through the da

It passes the eyeball test and shows trend.

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



No Trend and constant mean

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

KPSS test for d_Confirmed (including seasonals)

T = 840

Lag truncation parameter = 6

Test statistic = 0.0631644

	10%	5%	1%
Critical values:	0.348	0.462	0.743

no - evidence there is a constant mean -5

P-value > .10

P-Value is greater than 0.05 theres evidence for more than 1 mean.

ADF:
Augmented Dickey-Fuller test for d_Confirmed
testing down from 20 lags, criterion AIC
sample size 819
unit-root null hypothesis: a = 1

test with constant
including 20 lags of (1-L)d_Confirmed
model: (1-L)y = b0 + (a-1)*y(-1) + ... + e
estimated value of (a - 1): -1.03399
test statistic: tau_c(1) = -7.81024
asymptotic p-value 1.784e-12
1st-order autocorrelation coeff. for e: 0.001
lagged differences: F(20, 797) = 46.884 [0.0000]

P-value is less than 0.05 there is more than 1 mean.

no - means there is a constant mean -5

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.

7. 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** or Penn State https://onlinecourses.science.psu.edu/stat510/node/64
   I see an autoregressive and moving average in the ACF as there's a pattern of few lines going down and one larger one going up.

where? what lags for which processes? -2

8. 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

- **a.** 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!**
- **b.** Plot the observed versus fitted data for the time series data set **for each model.**
- **c.** Pick one of the models as your favorite and tell me why you like that one the best.
- **d.** Forecast your model out 6 time periods and graph the time series including the forecast. How well does the forecast seem to work?

Ljung-Box Q' = 359.637,
with p-value = P(Chi-square(5) > 359.637) = 1.472e-075

Function evaluations: 20
Evaluations of gradient: 8

Model 1: ARMA, using observations 2020-02-01:2022-05-20 (T = 840)
Estimated using AS 197 (exact ML)
Dependent variable: d_Confirmed
Standard errors based on Hessian

	coefficient	std. error	z	p-value	
const	0.664650	59.3515	0.01120	0.9911	
phi_1	0.184061	0.0835186	2.204	0.0275	**
theta_1	−0.451808	0.0724425	−6.237	4.47e-010	***

Mean dependent var	0.664286	S.D. dependent var	2641.530
Mean of innovations	−0.017349	S.D. of innovations	2546.930
R-squared	0.069235	Adjusted R-squared	0.068124
Log-likelihood	−7779.774	Akaike criterion	15567.55
Schwarz criterion	15586.48	Hannan-Quinn	15574.80

	Real	Imaginary	Modulus	Frequency
AR				
Root 1	5.4330	0.0000	5.4330	0.0000
MA				
Root 1	2.2133	0.0000	2.2133	0.0000

LM test for autocorrelation up to order 7 -
Null hypothesis: no autocorrelation
Test statistic: Chi-square(5) = 359.637

Test for ARCH of order 7 -

Null hypothesis: no ARCH effect is present
Test statistic: LM = 248.925
with p-value = P(Chi-square(7) > 248.925) = 4.69328e-50

9. 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.

Function evaluations: 128

Evaluations of gradient: 26

Model 2: ARIMA, using observations 2020-02-02:2022-05-20 (T = 839)

Estimated using AS 197 (exact ML)

Dependent variable: (1-L) d_Confirmed

Standard errors based on Hessian

	coefficient	std. error	z	p-value	
const	−0.128028	0.301382	−0.4248	0.6710	
phi_1	−0.217599	0.0336783	−6.461	1.04e-010	***
theta_1	−1.00000	0.00315755	−316.7	0.0000	***

Mean dependent var	0.002384	S.D. dependent var	4126.325
Mean of innovations	13.92732	S.D. of innovations	2577.357
R-squared	0.048030	Adjusted R-squared	0.046893
Log-likelihood	−7784.019	Akaike criterion	15576.04
Schwarz criterion	15594.97	Hannan-Quinn	15583.29

Real Imaginary Modulus Frequency

AR

Root 1	-4.5956	0.0000	4.5956	0.5000

MA

Root 1	1.0000	0.0000	1.0000	0.0000

Moneeb - you did not answer any part of question 8

8a. Where is your comparison table? Where are your four models?

8b. Where are your plots?

8c. Which model is the best model - what is your decision?

8d. Where is your forecast and plot? How well does it work?

points off for question 8 -20