Test 3 M349R Fall 2020 (60% of Exam 3) ***Make you have 5 problems (Good luck!)***

Name:\_Adhrit Srivastav\_\_\_\_ UTeid:\_\_Ams22362\_\_\_\_\_\_\_\_\_

**[1]** A math student, Eva, conducted a study on college basketball players. She asked players to continually perform three different activities as quickly as they could: layups (jumping, one-handed shot), free-throw shots (standing shot from fixed distance), and running drills. After performing each activity for five minutes, Eva recorded the player’s heart rate in beats per minute (bpm). She had five players, and she made each of them do all three activities on three different days. The days and order of activities were randomized for each player.

[a] Did Eva conduct an experiment or an observational study? Explain your answer. (6 pts)

Experimental study for the three different activities because she actively made the basketball players perform the activities. Observational study for the five players because she just observed how the activities would affect their heart rate in beats per minute.

[b] Describe: (3 pts each)

Units: The units are the basketball players.

Factor(s): The factors are the three different activities and the players.

Treatments: The treatments are the three different activities.

Response variable: The response variable are the players’ heart rates in beats per minute (bpm).

[c] Is Eva’s study balanced? (6 pts)

1. **Yes.**
2. No.
3. There is no way to tell from the information given.
4. This question is nonsensical.

[d] In an ANOVA model, a large statistic is an indication that (choose all that apply) (6 pts)

1. group-to-group variation and unit-to-unit variation are approximately equal
2. **group-to-group variation is large compared to unit-to-unit variation**
3. the experiment was randomized
4. we can conclude cause and effect
5. the sum of squares is large

[e] Multiple comparisons are a way to control (choose all that apply) (6 pts)

1. **Type I error rate**
2. Type II error rate
3. both Type I and Type II error rate
4. neither Type I nor Type II error rate

[f] A student, Eva, conducted a study on college basketball players. She asked players to continually perform three different activities as quickly as they could. After performing each activity for five minutes, Eva recorded the player’s heart rate in beats per minute (bpm). She had five players, and she made each of them do all three activities on three different days. The days and order of activities were randomized for each player. Eva wants to see if heart rate depends on activity and player. Write down the theoretical model that Eva will fit. (6 pts)

Observed Value = Grand Average + Activity Effect + Player Effect + Residuals

(bpm)

[g] Continue with Eva’s work. Discuss the conditions for the ANOVA model and how well you think the conditions are met for this model based on the residual plots and other information given. The plots are on the next page. (12 pts)

model4 <- aov(rate ~ factor(player) + factor(activity), data=Eva)  
summary(model4)

## Df Sum Sq Mean Sq F value Pr(>F)  
## factor(player) 4 17744 4436 178.9 <2e-16 \*\*\*  
## factor(activity) 2 8943 4472 180.3 <2e-16 \*\*\*  
## Residuals 83 2058 25   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

plot(model4) **#plots on the next page**

The Q-Q Norm plot appears straight thus the data appears normal. There is a slight curve to the residuals and there are no pronounced clusters, so it likely passes the linearity condition. The data likely passed the equal variance condition because there does not appear to be much fanning out. The data is also independent. There appears to be a significant difference in the average bpm due to the player and activity because their p-values of <2e-16 are less than alpha and thus significant.

Four scatter plots are titled Residuals vs Fitted; Normal Q-Q; Scale-Location; and Constant Leverage: Residuals vs Factor Levels. All data in the scatter plot are approximate.
"The graph titled Residuals vs Fitted correlates Residuals and Fitted values. The graph has Fitted Values on the horizontal axis ranging from 80 to 140 in increments of 10 and Residuals on the vertical axis ranging from negative 10 to 15 in increments of 5. The regression line starts at (80, negative 6), increases up to (98, 2) and gradually decreases to (144, negative 4). The graph shows many points scattered around the regression line. Three points are labeled as 66(92, 13), 65(92, 12), 67(79, negative 11).
The graph titled Normal Q-Q correlates Theoretical Quantiles and Standardized residuals. The graph has Theoretical Quantiles on the horizontal axis ranging from negative 2 to 2 in increments of 1 and Standardized residuals on the vertical axis ranging from negative 2 to 3 in increments of 1. A dotted regression line having a positive slope extends from (negative 3, negative 2) to (3, 2.1). The graph shows many points along the regression line. Three points are labeled as 66 (2.9, 2.7), 65 (2, 2.2), and 67 (negative 3, negative 2.5).
The graph titled Scale-Location correlates Fitted values and standard deviation of Standardized residuals. The graph has Fitted values on the horizontal axis ranging from 80 to 140 in increments of 10 and standard deviation of Standardized residuals on the vertical axis ranging from 0.0 to 1.5 in increments of 0.5. The regression line starts from (80, 1.3) decreases to (110, 0.8) and gradually ends at (144, 0.7). The graph shows many points around the regression line. Three points are labeled as 66(92, 1.7), 65(92, 1.5), and 67(80, 1.5).
The graph titled Constant Leverage: Residuals vs Factor levels correlates Factor Level Combinations and Standardized residuals. The graph has Factor Level Combinations on the horizontal axis ranging from 1 to 5 in increments of 1 and Standardized residuals on the vertical axis ranging from negative 2 to 3 in increments of 1. The regression line starts from (0, negative 0.1), increases to (2.5, 0) and slightly decreases to (5.3, negative 0.2). The graph shows many points around the regression line. The points are divided in between six vertical lines on the graph. Three points are labeled as 66(3.6, 2.7), 65(3.6, negative 2.4), and 67(4, negative 2.2). Factor level combinations 1 is labeled as factor (player)."

[h] Eva is wondering if it would make more sense to use a reexpression of her response (heart rate), so she makes a Tukey plot for nonadditivity, seen below. What reexpression of the response variable (if any) is indicated by this plot? (6 pts)

plot(model.resid~comp.value, ylab="residuals", xlab="comparison value")  
Tukey <- lm(model.resid~comp.value); Tukey

##   
## Call:  
## lm(formula = model.resid ~ comp.value)  
##   
## Coefficients:  
## (Intercept) comp.value   
## 0.000 -1.564

abline(reg=Tukey)

A scatter plot graph correlates comparison value and residuals.
The graph has comparison value along the horizontal axis ranging from negative 2 to 3 and residuals along the vertical axis ranging from negative 5 to 5. The regression line having a negative slope extends from (negative 3, 4) to (3, negative 5). The graph shows many points scattered around the regression line. A point lies significantly at (negative 1, 5).

If the points do fall near the line, it indicates that the reexpression will fit an additive model well. For Eva’s Tukey plot for nonadditivity, the plot shows wide scatter about the fitted line. Thus, there is no reexpression of the response variable indicated by the plot.

**[2]** Problem 2 (Overlays)

A study of two surgical methods compares recovery times, in days, for two treatments, the standard and the new method. Three randomly chosen patients got the new treatment; the remaining three patients got the standard. Here are the results:

New procedure 16, 20, 24

Standard 28, 33, 35

[a] Fit a one-way additive model “days = treatment + error” and write a conclusion (14 points)

Observed value (recovery days) = Grand mean + Treatment effect + Residuals

H\_0: alpha\_1 = alpha\_2 = 0; H\_a: at least one alpha does not equal 0.

If we fit a one-way additive model, we get a p-value of .0182 which is less than the alpha, so we reject the null hypothesis. Thus the type of treatment has a significant impact on the days of recovery.

[b] For the data above decompose the response value as a sum of grand mean + treatment effects + residuals. (14 points)

|  |  |
| --- | --- |
| 16 | 28 |
| 20 | 33 |
| 24 | 35 |

= Grand Mean

|  |  |
| --- | --- |
| 26 | 26 |
| 26 | 26 |
| 26 | 26 |

+ Treatment Effects

New Procedure Standard

|  |  |
| --- | --- |
| -6 | 6 |
| -6 | 6 |
| -6 | 6 |

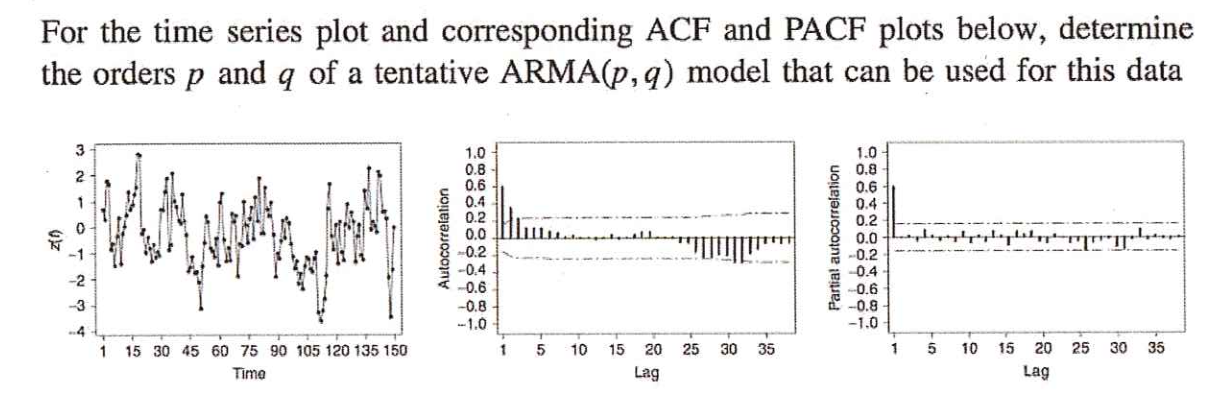
+ Residuals

|  |  |
| --- | --- |
| -4 | -4 |
| 0 | 1 |
| 4 | 3 |

Make sure (double check) that the sum of square residuals and the sum of square treatment effects is the same as the Anova table from part [a]

Problem 3

Part [a] (4 points for explain the process of reading correlograms and 2 points for backshift notation)

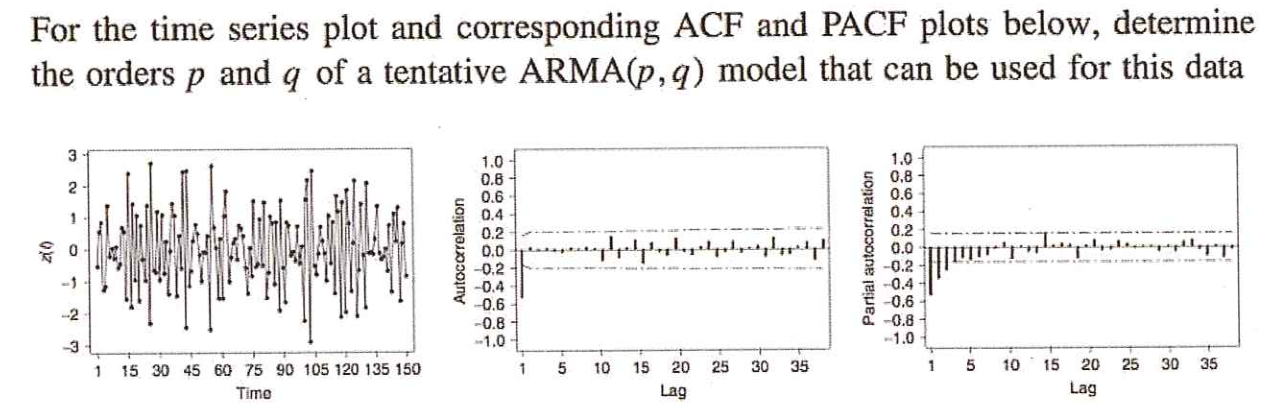


Explain how you picked an Arima model:

AR(1) have ACFs that exponentially decrease. Higher-order AR processes are often a mixture of exponentially decreasing and dampening sinusoidal components. Use AR model for geometric/sinusoidal decay in ACF. Use MA model for geometric/sinusoidal decay in PACF. I chose an ARMA(2, 0) model because in the ACF there are significant autocorrelations at lags 1 and 2 (lag 3 is barely significant), suggesting that one value is related to the next and the next. I chose AR(2) because of the mixture of exponentially decreasing sinusoidal components, which we know is a characteristic of higher-order AR models. There are also two significant lags at the beginning of the ACF. The PACF cuts after lag 1 and doesn’t have the exponential/geometric/sinusoidal decay, so I chose a MA(0) model.

Y\_t = delta + Phi\_1\*Y\_t-1 + Phi\_2\*Y\_t-2 + epsilon\_t

Part [b] (4 points for explain the process of reading correlograms and 2 points for backshift notation)

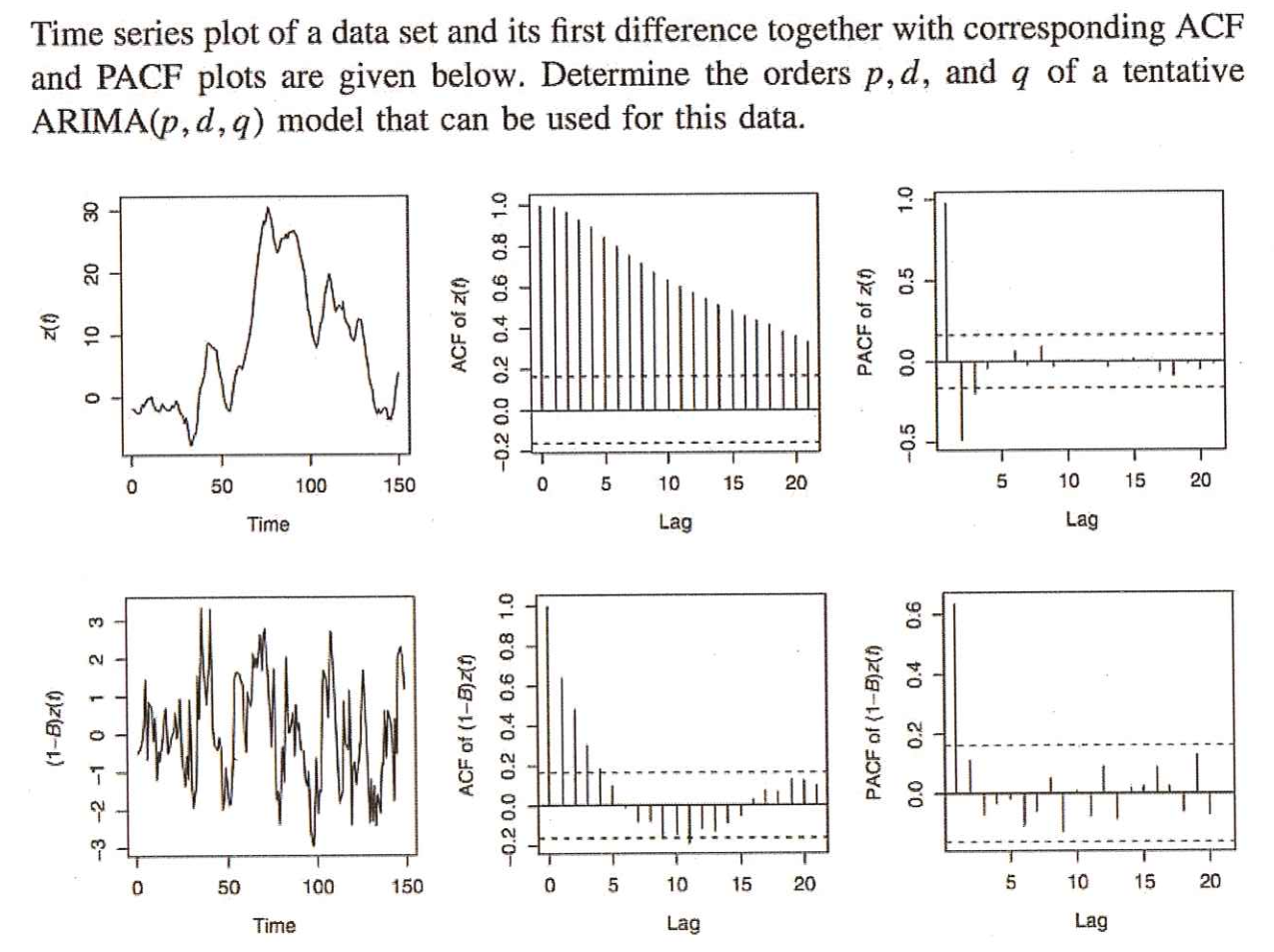


Explain how you picked an Arima model:

AR(1) have ACFs that exponentially decrease. Higher-order AR processes are often a mixture of exponentially decreasing and dampening sinusoidal components. Use AR model for geometric/sinusoidal decay in ACF. Use MA model for geometric/sinusoidal decay in PACF. I chose an ARMA(0, 1) model because the ACF cuts after lag 1 which indicates one value is likely not related to the next values. The ACF does not have the exponential/geometric/sinusoidal decay, so I chose a AR(0) model. The series values might also be related to the residuals at previous times based on the time series plot. The PACF does have the exponential/geometric/sinusoidal decay, so I chose a MA(1) model.

Y\_t = delta + epsilon\_t + theta\_1\*epsilon\_t-1

Part [c] (4 points for explain the process of reading correlograms and 2 points for backshift notation)

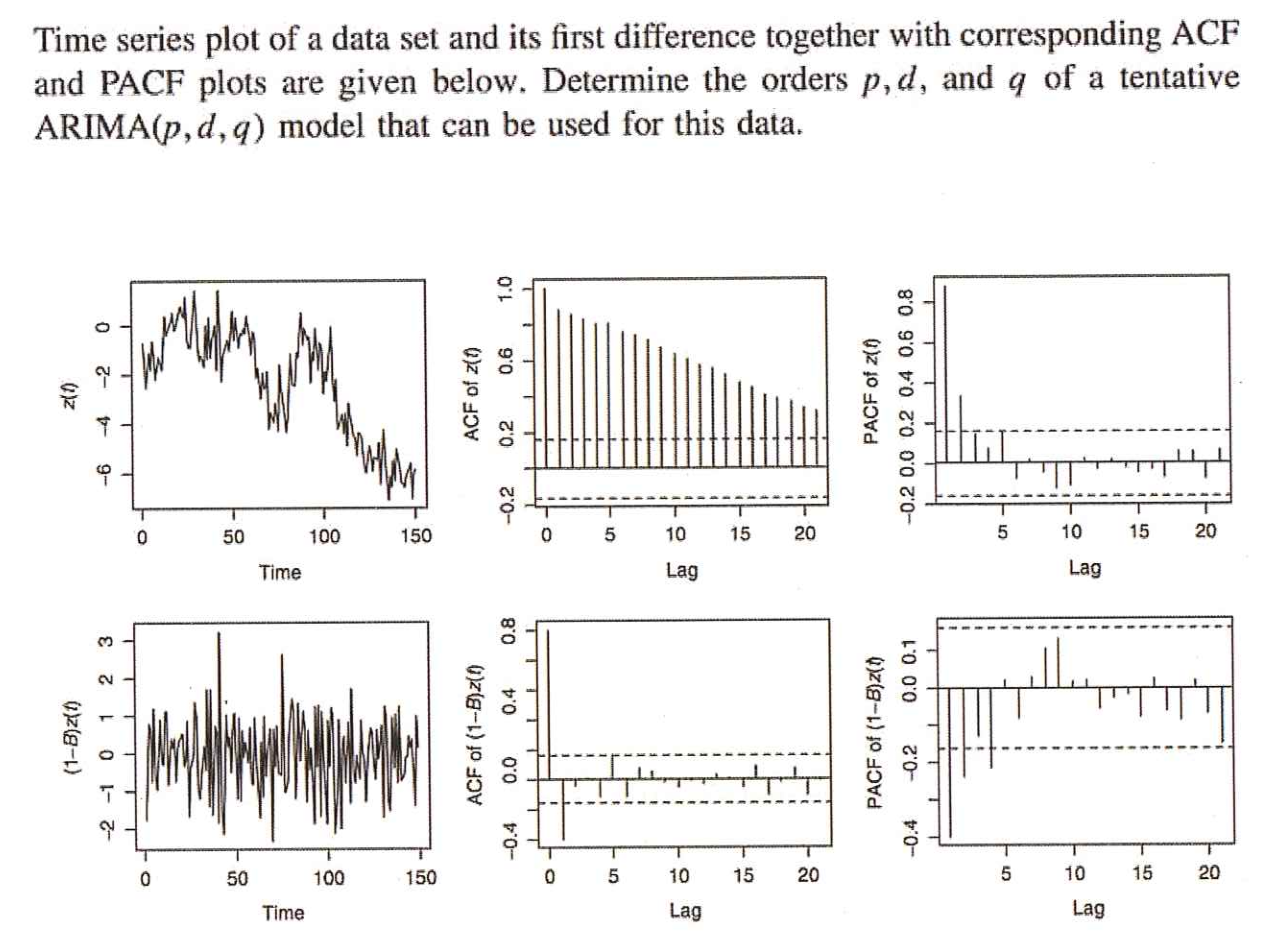


Explain how you picked an Arima model:

AR(1) have ACFs that exponentially decrease. Higher-order AR processes are often a mixture of exponentially decreasing and dampening sinusoidal components. Use AR model for geometric/sinusoidal decay in ACF. Use MA model for geometric/sinusoidal decay in PACF. I am choosing an ARIMA(3,1,0) model because the slow linear decay in the initial ACF signals that the data is non-stationary. We must try some sort of differencing, thus d = 1. I chose AR(3) because in the ACF there are significant autocorrelations at lags 1, 2, and 3 (lag 4 is barely significant), suggesting that one value is related to the next and the next. I chose AR(3) because of the mixture of exponentially decreasing sinusoidal components, which we know is a characteristic of higher-order AR models. There are also two significant lags at the beginning of the ACF. The PACF cuts after lag 1 and doesn’t have the exponential/geometric/sinusoidal decay, so I chose a MA(0) model.

Delta\_1\*Y\_t = delta + epsilon\_t + Phi\_1\*Delta\_1\*Y\_t-1 + Phi\_2\*Delta\_1\*Y\_t-2 + Phi\_3\*Delta\_1\*Y\_t-3

Part [d] (4 points for explain the process of reading correlograms and 2 points for backshift notation)

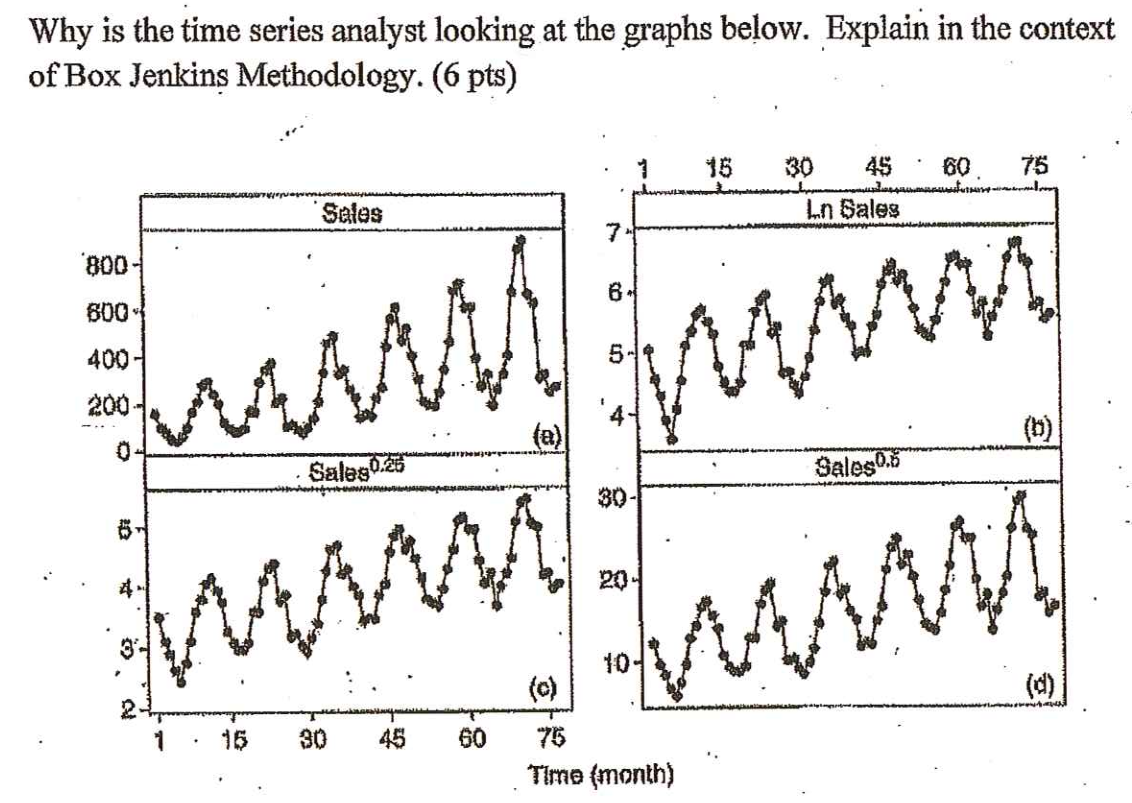


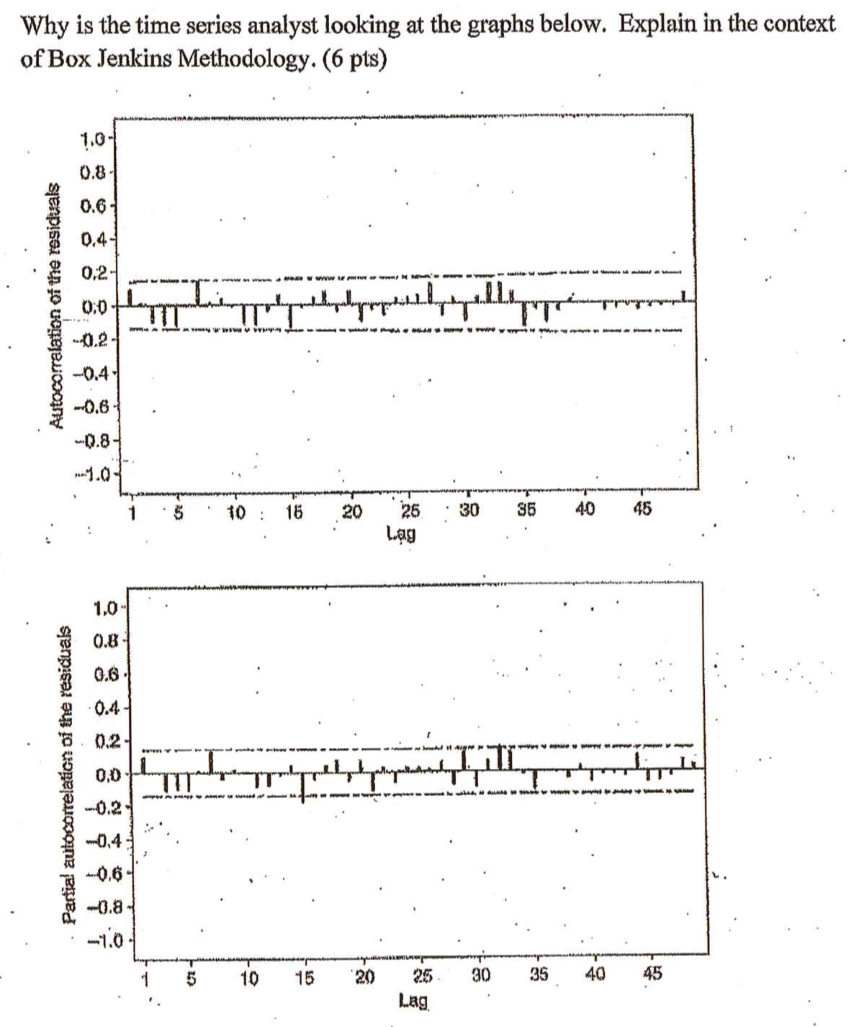
Explain how you picked an Arima model:

AR(1) have ACFs that exponentially decrease. Higher-order AR processes are often a mixture of exponentially decreasing and dampening sinusoidal components. Use AR model for geometric/sinusoidal decay in ACF. Use MA model for geometric/sinusoidal decay in PACF. I chose an ARIMA(0,1,1) model because the slow linear decay in the initial ACF signals that the data is non-stationary. We must try some sort of differencing, thus d = 1. The ACF does not have the exponential/geometric/sinusoidal decay, so I chose an AR(0) model. The series values might also be related to the residuals at previous times based on the time series plot. The PACF does have the exponential/geometric/sinusoidal decay, so I chose a MA(1) model.

Delta\_1\*Y\_t = delta + epsilon\_t + theta\_1\*epsilon\_t-1

Problem 4





The Box Jenkins Methodology starts by determining whether or not the time series is stationary and if there is any significant seasonality that needs to be modeled. The time series analyst is looking at the ACF and PACF functions to make sure that the data is stationary. There is no slow linear decay in the ACF, so the data is indeed stationary. AR(1) have ACFs that exponentially decrease. Higher-order AR processes are often a mixture of exponentially decreasing and dampening sinusoidal components. The time series analyst looks at the time series plots, ACF, and PACF to detect seasonality. From the time series plots we can see the data has a significant seasonality due to its sinusoidal nature. The different data transformations are to stabilize the data and confirms that there is significant seasonality in the data.

Problem 5 (36 points)

Go to the data set “series” under announcements in Canvas and use the four steps of Arima Nonseasonal Modeling in order to forecast ten periods (ten units). For full credit explain in detail each step in essay format and copy-and-paste every pertinent graph and R output after each explanation (make it flow like an essay).

(N.B) No auto.arima, but your reasoning. ***Enjoy the break! We had fun.***

First step is to look at the initial data. There is a slow linear decay in the ACF which indicates the data is non-stationary.

Graphical user interface

Description automatically generated with medium confidence

Next step is to try some sort of differencing. We run the plots again with the differences of the values and we see that the ACF and PACF now cut after lag 2 and are now stationary.

Graphical user interface, application, timeline

Description automatically generated

The following step is to determine a potential ARIMA model. I chose an ARIMA(0,1,1) model because we know we had to apply differencing to the model. The series values might also be related to the residuals at previous times based on the time series plot. The next step is to check the residuals from the ARIMA model.

Chart

Description automatically generated with medium confidence

The ACF has no significant autocorrelations and overall the residuals look good.

The final step is to forecast ten periods/units.

Chart, line chart

Description automatically generated