Time Series Modeling of Domestic Box Office Earnings from 1975-2019

Maria Escobar
Obehi Ikpea
James Miller
Eileen Ramirez del Rio

STA-4753: Time-Series Analysis

Dr. Jerome Keating

May 9th, 2022

INTRODUCTION

Film is a massive global industry, worth hundreds of billions of dollars globally. It is entertainment, art, entrepreneurship, as well as family bonding and nostalgia. Being able to quantify these intense feelings and meanings into economic trends can provide valuable insight into the industry as a whole, and possibly provide production companies with information on what time of the year to release films in order to optimize profits.

For this project, we have chosen a dataset that shows domestic box office gross earnings available from Box Office Mojo, an IMDb project. It is possible to classify the data by day, week, weekend, month, quarter, year, season, and even by holidays; for our project, we decided to utilize the quarterly data.

Variables that are given in Box Office Mojo are the year, the quarters within that year, the cumulative gross earnings of all released films up to that quarter, the gross change per year compared to the previous year, the number of releases in the quarter, the average gross earning of all films released that quarter, the name of the highest-grossing film in the quarter, the gross earnings of the highest-grossing film that quarter, and the percent of the cumulative gross earnings that the highest-earning film composes. The dataset sorted by quarter and calendar gross begins in 1972 and ends in 2022, but for our purposes, we have trimmed it to 1975-2019.

DATA

For the purposes of our project, we decided to utilize the following variables:

-**Time:** the amount of time elapsed (in quarters) since the beginning of the recording of the data at the beginning of the year 1972

-**Quarter 1 - Quarter 4:** a dummy variable denoting the quarter in which each of the respective time variable values falls into. The breakdown of the quarters by month is as follows:

- **Quarter 1:** January - March
- **Quarter 2:** April - June
- **Quarter 3:** July - September
- **Quarter 4:** October - December

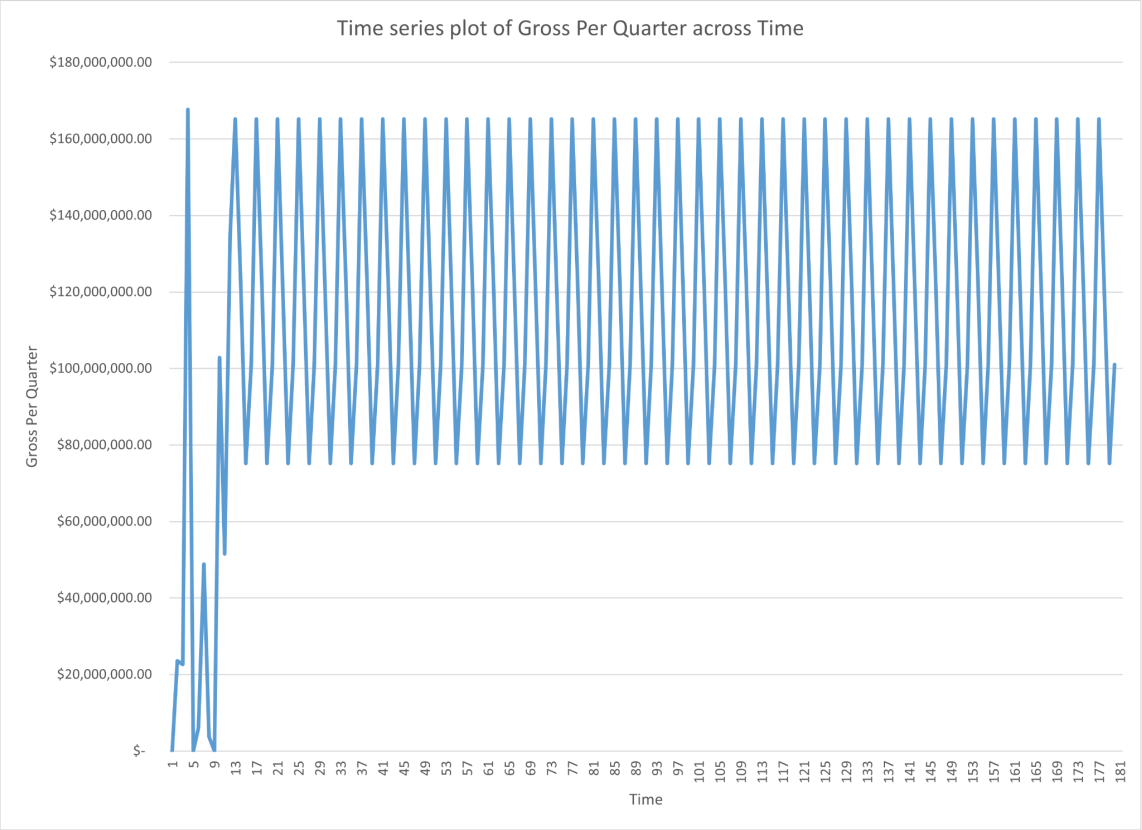
-**Gross per quarter:** the total revenue in USD of the respective quarter during its complete elapsed time of 3 months

We will also be utilizing both Microsoft Excel and SAS for our analysis.

ANALYSIS

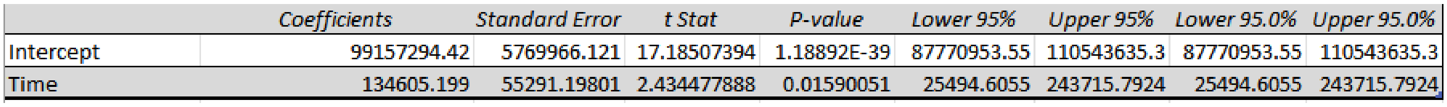
CONSTRUCTING THE MODEL

After importing the dataset, we made a plot of the Gross earnings to check for patterns in the dataset.



As we can see from the graph of the observed values obtained directly from Box Office Mojo, there seems to be a clear trend of periodicity in the data set, having a general cycle of increment and decrement. For these reasons, we can assume that our time series would be not strictly stationary, as its statistical properties are changing constantly through time. For our analysis and forecasting, we chose to do additive decomposition, as it would allow us to break down the time series into its systematic components (in our case, trend, seasonality, and [white] noise).

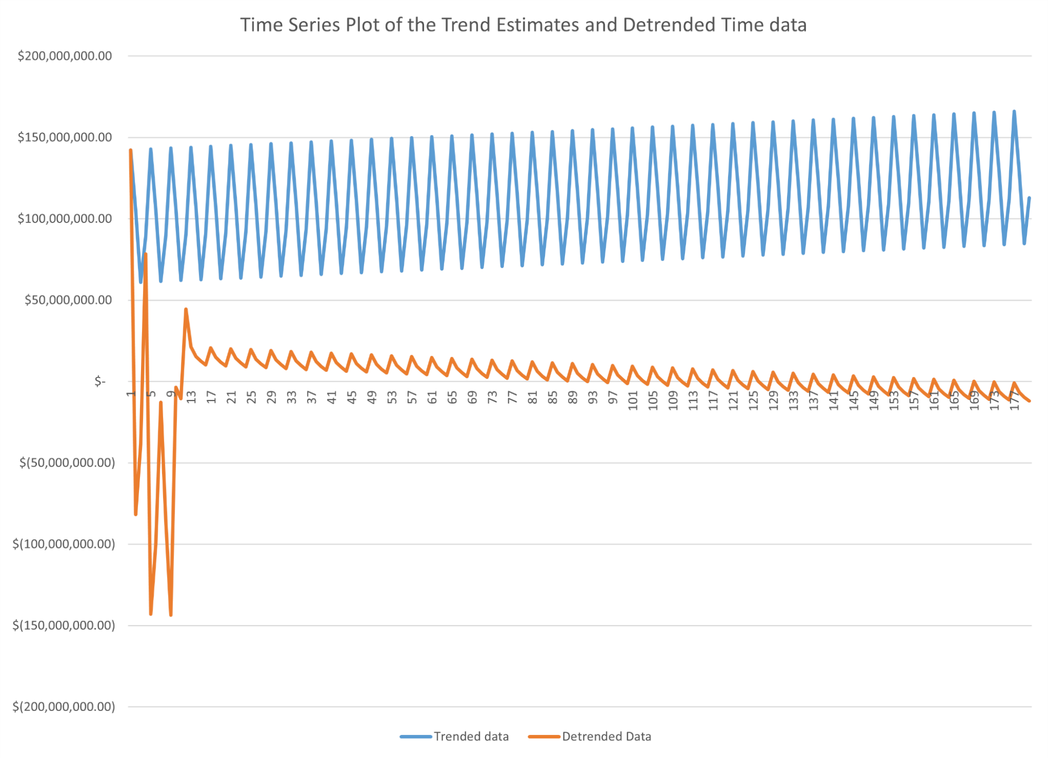
Our first step was to model the trend with a regression equation, where we used the gross earnings as our dependent variable and time as our regressor variable. Our output for this initial regression was as follows:



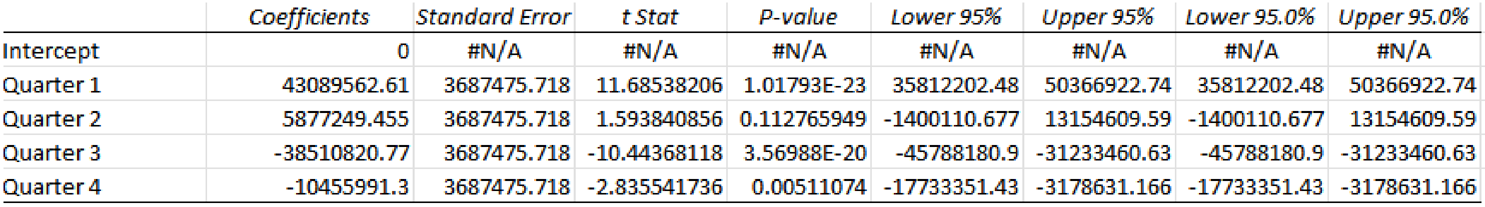
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	99157294.42	5769966.121	17.18507394	1.18892E-39	87770953.55	110543635.3	87770953.55	110543635.3
Time	134605.199	55291.19801	2.434477888	0.01590051	25494.6055	243715.7924	25494.6055	243715.7924

The regression equation for the trend is denoted as: $\widehat{y}_t = 99157294.42 + 134605.198t + \widehat{S}_t$

After getting the regression equation for the trend, we detrended the time series by subtracting the original time series from the predicted values in order to extract the seasonal and error terms of the model. A plot of the predicted values vs. detrended data is below.



We then performed regression analysis on the detrended time series (residuals) to estimate the quarterly effects, which gives the following output:



	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Quarter 1	43089562.61	3687475.718	11.68538206	1.01793E-23	35812202.48	50366922.74	35812202.48	50366922.74
Quarter 2	5877249.455	3687475.718	1.593840856	0.112765949	-1400110.677	13154609.59	-1400110.677	13154609.59
Quarter 3	-38510820.77	3687475.718	-10.44368118	3.56988E-20	-45788180.9	-31233460.63	-45788180.9	-31233460.63
Quarter 4	-10455991.3	3687475.718	-2.835541736	0.00511074	-17733351.43	-3178631.166	-17733351.43	-3178631.166

From this output, we obtain the following formula denoting the seasonality of our model:

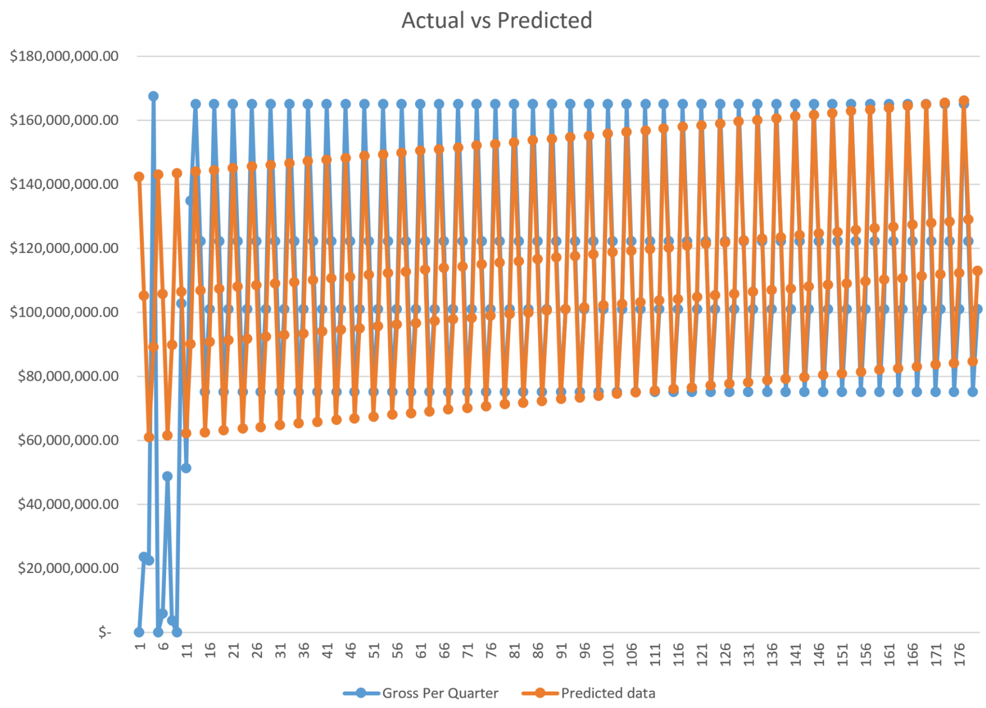
$$\widehat{S}_t = 43089562.61Q_1 + 5877249.455Q_2 - 38510820.77Q_3 - 10455991.3Q_4$$

By combining our 2 previously obtained equations, we are able to get the following final formula:

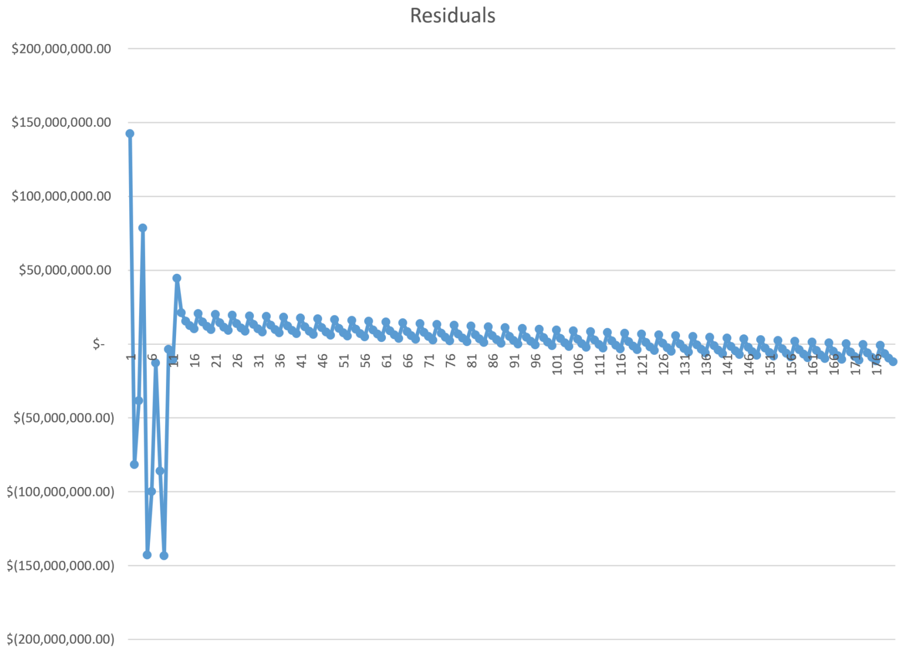
$$\widehat{y}_t =$$

$$99157294.42 \ + 134605.198t + 43089562.61Q_1 + 5877249.455Q_2 - 38510820.77Q_3 - 10455991.3Q_4 + \widehat{R}_t$$

Our intercept is $99157294.42, the average revenue for the quarter previous to the beginning of the data recording (therefore, the total revenue between October - December of 1971). T denotes the amount of time that has passed since the beginning of the data recording (as t increases by one unit, the average gross per quarter increases by $134605.198). Q1 through Q4 denotes quarters 1 through 4, which as stated before, denotes the quarter in which each of the respective time variable values falls into (the average gross per quarter increases/decreases by $43089562.61, $5877249.455, -$38510820.77, and -$10455991.3 for quarters 1 through 4 respectively). By utilizing this final formula, we are able to compare our observed and predicted values:



We are able to see that for the most part, our model seems to be a good prediction for the observed values, as well as modeling its cyclical/seasonal nature. We are also able to obtain the residuals between our observed and predicted values, which can be seen in the following graph:



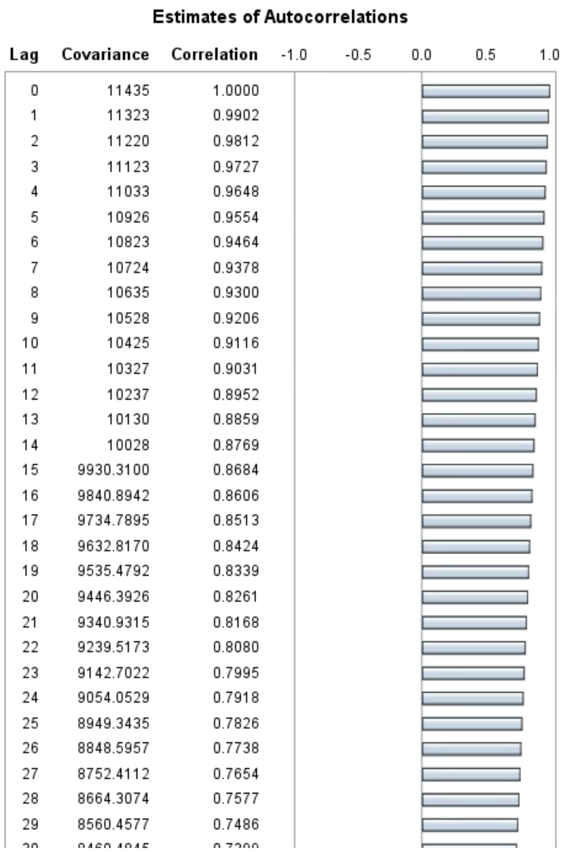
As we can observe, our residuals also seem to follow a type of pattern and thus point us to the possibility of there being an autocorrelation component to the residual component of our model. This is further proven by the Durban-Watson statistic obtained from our residuals, which can be obtained by using the following formula:

$$d = \sum_{t=2}^{T}(e_t - e_{t-1})^2 / \sum_{t=1}^{T} {e_t}^2$$

When using this formula, we obtain a Durbin-Watson statistic of 0.699473698, which would indicate that there is a very high autocorrelation between our residuals. Due to this, we must also fit an ARIMA model to our residual values.

MODELING THE RESIDUALS

Using the ARIMA model, we performed our analysis on the residuals in three stages. The first thing we did was analyze the correlation properties of the residuals using the AUTOCORRELATION statement in SAS. For this model, we decided to use 40 lags, as it was the best candidate given our total number of observations. Our initial output is as follows:

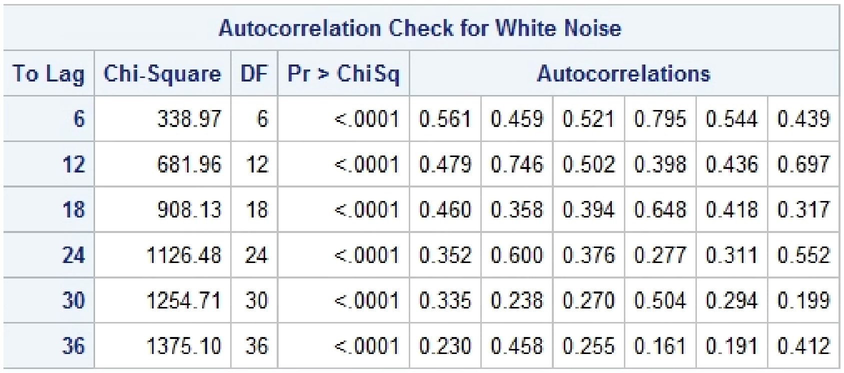


```
                  Estimates of Autocorrelations

Lag   Covariance   Correlation   -1.0   -0.5   0.0   0.5   1.0

  0        11435        1.0000
  1        11323        0.9902
  2        11220        0.9812
  3        11123        0.9727
  4        11033        0.9648
  5        10926        0.9554
  6        10823        0.9464
  7        10724        0.9378
  8        10635        0.9300
  9        10528        0.9206
 10        10425        0.9116
 11        10327        0.9031
 12        10237        0.8952
 13        10130        0.8859
 14        10028        0.8769
 15    9930.3100        0.8684
 16    9840.8942        0.8606
 17    9734.7895        0.8513
 18    9632.8170        0.8424
 19    9535.4792        0.8339
 20    9446.3926        0.8261
 21    9340.9315        0.8168
 22    9239.5173        0.8080
 23    9142.7022        0.7995
 24    9054.0529        0.7918
 25    8949.3435        0.7826
 26    8848.5957        0.7738
 27    8752.4112        0.7654
 28    8664.3074        0.7577
 29    8560.4577        0.7486
 30    8460.4945        0.7399
```

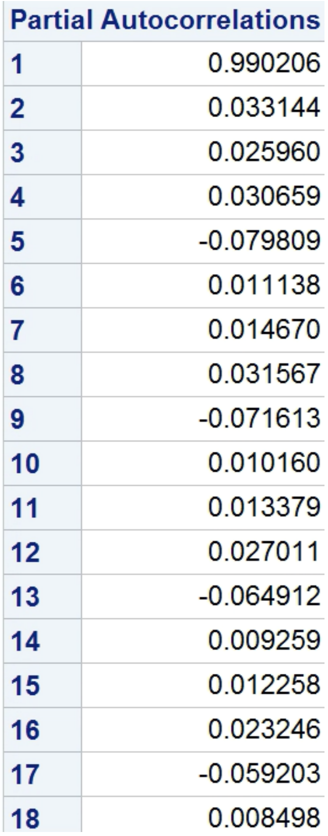
As we can see from the autocorrelation estimates, our residuals appear to follow an exponential decay pattern, proving our point of time series being non-stationary in nature.



Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	338.97	6	<.0001	0.561	0.459	0.521	0.795	0.544	0.439
12	681.96	12	<.0001	0.479	0.746	0.502	0.398	0.436	0.697
18	908.13	18	<.0001	0.460	0.358	0.394	0.648	0.418	0.317
24	1126.48	24	<.0001	0.352	0.600	0.376	0.277	0.311	0.552
30	1254.71	30	<.0001	0.335	0.238	0.270	0.504	0.294	0.199
36	1375.10	36	<.0001	0.230	0.458	0.255	0.161	0.191	0.412

From the Autocorrelation check for white noise table, we observe that the p-value for the first six lags is less than 0.001, allowing us to reject the null hypothesis that none of the autocorrelations of the series up to a given lag are significantly different from 0 and thus further proving our point of time series being non-stationary in nature. We also looked at the partial autocorrelations between the lags, as these are commonly used to identify the order of the autoregressive model.

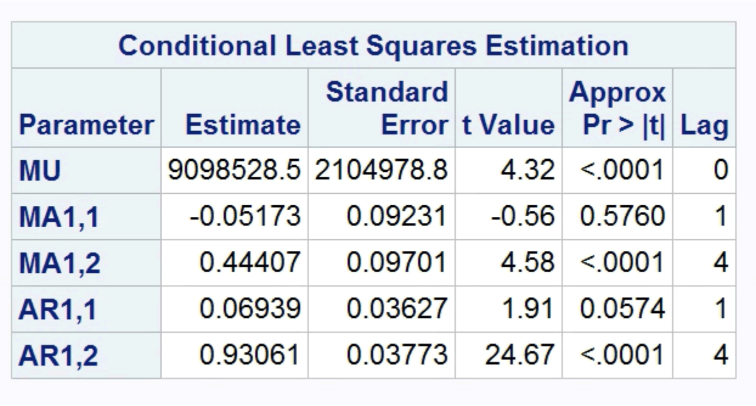


Partial Autocorrelations	
1	0.990206
2	0.033144
3	0.025960
4	0.030659
5	-0.079809
6	0.011138
7	0.014670
8	0.031567
9	-0.071613
10	0.010160
11	0.013379
12	0.027011
13	-0.064912
14	0.009259
15	0.012258
16	0.023246
17	-0.059203
18	0.008498

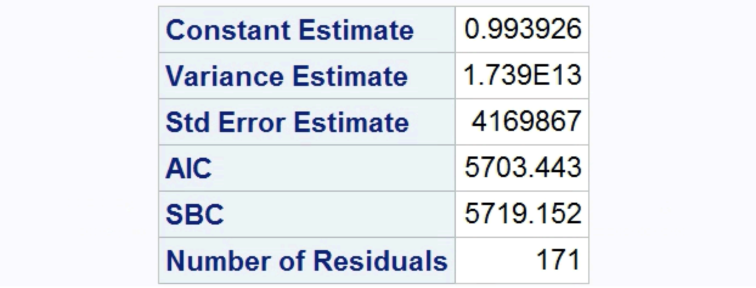
As we can observe from our partial autocorrelations, there seems to be a pattern of exponential decay, with a relatively higher peak every 4th lag. For this reason, we decided run an ARIMA statement and estimate the value of both p (for our AR model) and q (for our MA model) to be either 1 or 4 for the following reasons:

- The ACF of our residuals have an exponential decay trend to them, which is common for an AR(1) process
- As the data of our time series is collected quarterly, we could have an AR(4) process
- The order of the MA(q) process is decided similarly to that of the AR(p) process; thus, we also utilize 1 and 4 for this

Once we run this new process, we obtain the following diagnostics:

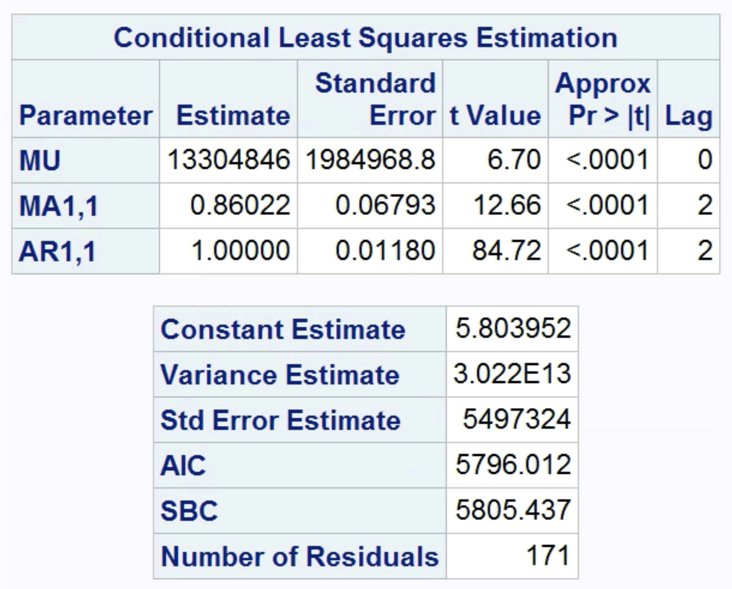


Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	9098528.5	2104978.8	4.32	<.0001	0
MA1,1	-0.05173	0.09231	-0.56	0.5760	1
MA1,2	0.44407	0.09701	4.58	<.0001	4
AR1,1	0.06939	0.03627	1.91	0.0574	1
AR1,2	0.93061	0.03773	24.67	<.0001	4



Constant Estimate	0.993926
Variance Estimate	1.739E13
Std Error Estimate	4169867
AIC	5703.443
SBC	5719.152
Number of Residuals	171

From the conditional least squares estimation, the only statistically significant parameters are MU, MA(2), and AR(2), with the other 2 parameters' p-values indicating that they add too little to the model to justify their inclusion. Given this information, we run the model once again with these parameters and obtain the following:

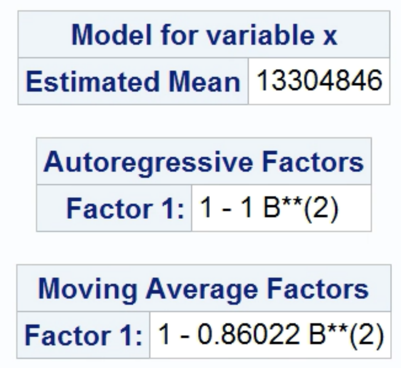


Conditional Least Squares Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	13304846	1984968.8	6.70	<.0001	0
MA1,1	0.86022	0.06793	12.66	<.0001	2
AR1,1	1.00000	0.01180	84.72	<.0001	2

Constant Estimate	5.803952
Variance Estimate	3.022E13
Std Error Estimate	5497324
AIC	5796.012
SBC	5805.437
Number of Residuals	171

Despite the AIC and SBC statistics being higher when running our ARIMA model with both p and q having a value of 2, the difference is so small that it is deemed insignificant. We are able to obtain the following results from this statement:



Model for variable x	
Estimated Mean	13304846

Autoregressive Factors	
Factor 1:	1 - 1 B**(2)

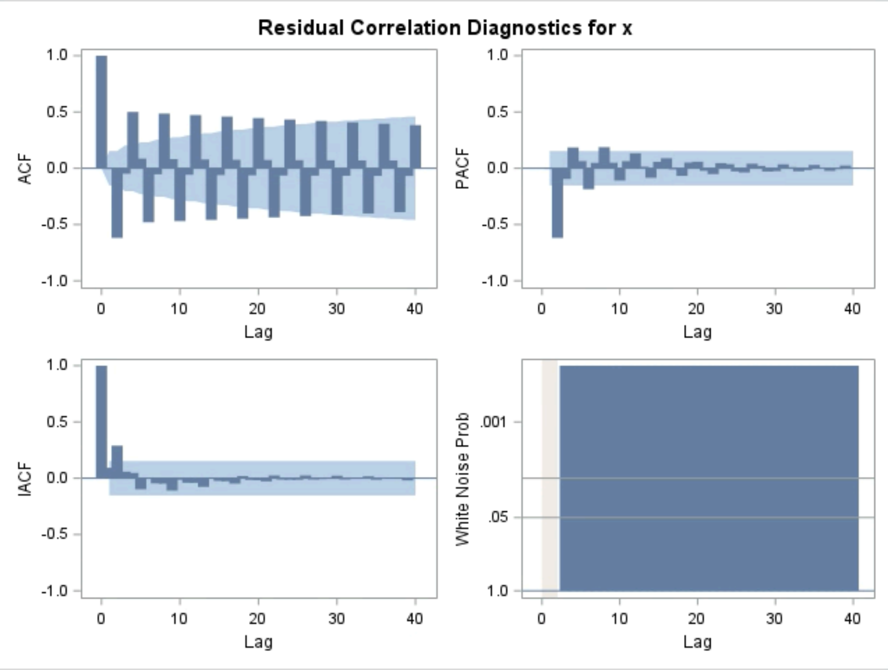
Moving Average Factors	
Factor 1:	1 - 0.86022 B**(2)

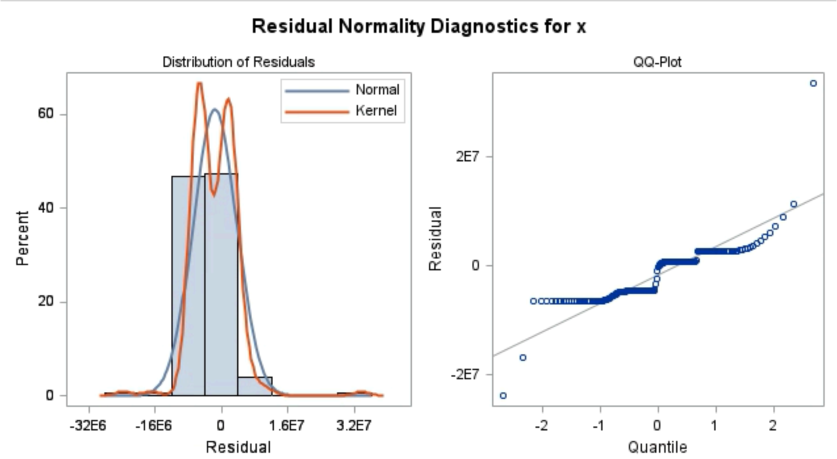
The final equation for the ARIMA model for the residuals of our time series would be as follows:

$$\widehat{R}_t = (1 - B)_{grossperquarter(t)} = 13304846 + \frac{1 - B^2}{1 - 0.86022B^2} a_t$$

DIAGNOSTICS AND ASSUMPTIONS CHECK

One of the main assumptions of time series is their stationary qualities; as our data set was not stationary (having a clear trend), we had to transform it to a stationary process in by obtaining the first difference (or residuals). We then observed that our residuals were highly autocorrelated as their Durbin-Watson statistic was 0.699, so the autocorrelation assumption wasn't met. Therefore, we decided to model the residuals using the ARIMA procedure to fix this problem. From this analysis, we got the Correlation Diagnostics and Normality diagnostics plot in which we can check if further assumptions are met.





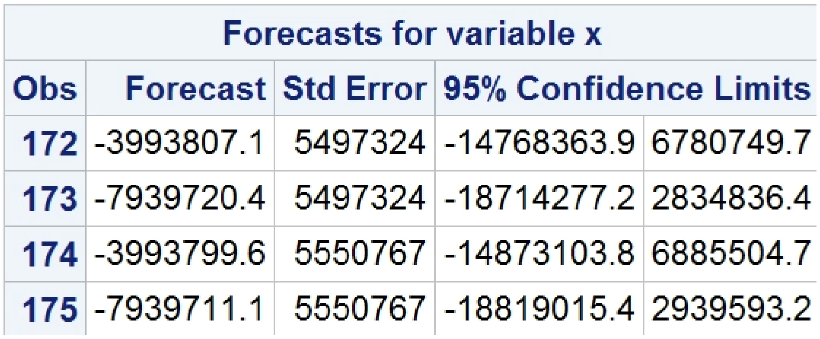
While the ACF vs. Lag plot does seem to appear to go slightly out of the error parameters, the PACF and IACF vs Lag plots seem to be better fitted with these orders for the AR(p) and MA(q). We are also able to see that our Q-Q plot is now closer to a straight line than our previous models, giving it some validity over them. The white noise indicates that the residuals of the problem are non-normal and thus are over-dispersed for normality (the variances are too big to follow a normal distribution), which might mean we should fit them to a different distribution or we could have converted the data by taking a log or a square root for example.

FORECAST OF FUTURE VALUES

Using the full model equation we obtained previously, we are able to forecast the gross per quarter revenue a year (or 4 quarters) into the future, utilizing the following times:

- **T : 181** => $99157294.42\ +\ 134605.198(181)\ +\ 43089562.61(1)\ +\widehat{R}_t = \$166,610,398.04\ +\widehat{R}_t$
- **T : 182** => $99157294.42\ +\ 134605.198(182)\ +\ 5877249.455(1)\ +\widehat{R}_t = \$129,532,690.08\ +\widehat{R}_t$
- **T : 183** => $99157294.42\ +\ 134605.198(183)\ -\ 38510820.77(1)\ +\widehat{R}_t = \$85,279,225.06 + \widehat{R}_t$
- **T : 184** =>$99157294.42\ +\ 134605.198(184)\ -\ 10455991.3(1)\ +\widehat{R}_t = \$113,468,659.73 + \widehat{R}_t$

Utilizing the ARIMA procedure we utilized for our residual model, we are able to obtain the following forecasted values for our residuals:



Forecasts for variable x

Obs	Forecast	Std Error	95% Confidence Limits	
172	-3993807.1	5497324	-14768363.9	6780749.7
173	-7939720.4	5497324	-18714277.2	2834836.4
174	-3993799.6	5550767	-14873103.8	6885504.7
175	-7939711.1	5550767	-18819015.4	2939593.2

We then add these values to our full model results for t = (181, 182, 183, 184):

- **T : 181** => $\$166610398.04\ +\widehat{R}_t\ =\ \$166,610,398.04\ -\ \$3,993,807.1\ = \$162,616,590.91$
- **T : 182** => $\$129,532,690.08\ +\widehat{R}_t\ = \$129,532,690.08\ -\ \$7,939,720.01\ =\ \$121,592,970.07$
- **T : 183** => $\$85,279,225.06 + \widehat{R}_t\ = \$85,279,225.06\ -\ \$3,993,799.6\ =\ \$\ 81,285,475.46$
- **T : 184** =>$\$113,468,659.73 + \widehat{R}_t\ = \$113,468,659.73\ -\ \$7,939,711.1\ = \$105,528,948.63$

APPENDIX

5.1 decomposition models: Stat 510. PennState: Statistics Online Courses. (n.d.). Retrieved May 4, 2022, from https://online.stat.psu.edu/stat510/lesson/5/5.1

Brillinger, D. (2000, November 17). Time Series: General. Berkley, California, USA; University of California, Berkely.

Domestic yearly box office. Box Office Mojo. (n.d.). Retrieved May 2, 2022, from https://www.boxofficemojo.com/year/?ref_=bo_nb_qy_secondarytab

Iordanova, T. (2022, February 8). *An introduction to non-stationary processes*. Investopedia. Retrieved May 4, 2022, from https://www.investopedia.com/articles/trading/07/stationary.asp#:~:text=When%20a%20time%20series%20is,variables%20do%20vary%20with%20time.

Radečić, D. (2022, March 23). *Time series from scratch - decomposing Time Series Data*. Medium. Retrieved May 4, 2022, from https://towardsdatascience.com/time-series-from-scratch-decomposing-time-series-data-7b7ad0c30fe7

SAS Institute Inc. (1993). The ARIMA Procedure. In *SAS*.