Renjun Zhu(26332920)
Stat 153: Fall 2017
Professor: Brillinger
Dec. 8th, 2017

A Time Series Perspective of Ballet, "Shall We Dance?"

I. Question:

The scientific question motivating my work is: can we use any time series model to predict the appreciation of ballet performance over the next two years?

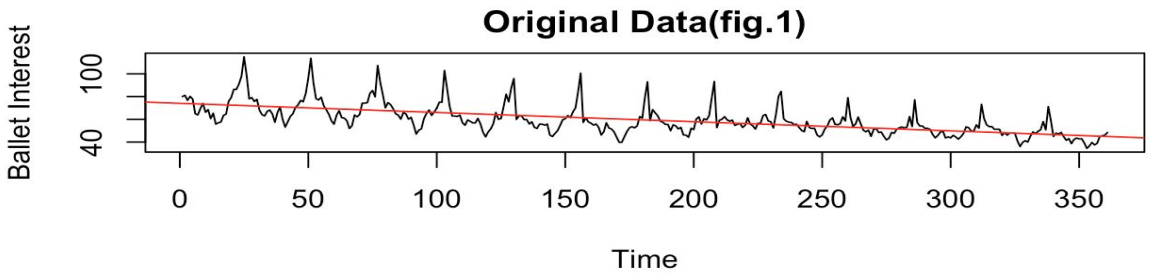
II. Introduction:

I am new to dance, and I am entranced by the show of ballet. I appreciate the complete performance, especially the synchronicity of ballerinas with the flow of the music, but I wonder if such an attraction will continue to appeal to people over time. My impetus for this question comes from the feeling that every time I attend a ballet, the people around me are considerably older. I often feel that I do not belong there, because I do not see many people from my generation. However, in the moment of the performance, I forget about my surroundings and immerse myself in it, and I imagine that the producers of the shows also find themselves doing this. Over the last few years, I have noticed no change in the marketing strategies that they use. This leaves no room for understanding the audience and making sure the productions attract new blood. When the closing act approaches, and I am brought back to real life. I feel afraid that one day, this cultural attraction might die out. I would like to use my knowledge from time series analysis to see if my worries are justified, that is, the culture is on a collision course with its own end, or if I am only overthinking the problem.

III. Data:

People usually see a live performance when they have free time such as over the weekends, during the summer, or on winter breaks, so I decided to take a weekly dataset of

people showing interest in ballet from Google Trends. Google Trend data is derived from search results and other relevant data. Google is nearly ubiquitously used for search and therefore its results might be a fair representation of interest. Since people from different regions might have different cultural standards for live performances, I chose to focus on the general interests of ballet performances in the United States. I used the keyword, "ballet", in Google Trend to get three consecutive five-year spans of weekly comma-separated data files which in total consists 726 values from January, 2004 to November, 2017. I evaluate the data in a bi-weekly fashion because, in most cases, the same ballet performance will not repeat for more than two weeks. The reduction in data also potentially removes some bias and noise.



V. Method:

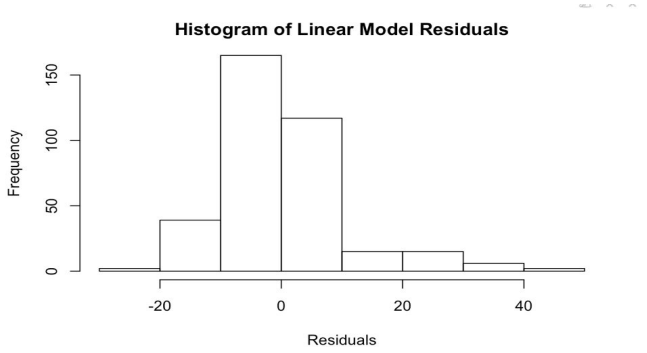
I first plot the data directly to see what it looks like. From fig.1 above, I found my data is obviously yearly seasonal, since there are 13 mountain-shaped ups and downs across the 13 years in a continuous curve. I also notice that there is a slightly downward trend, which I highlighted in red.

(1). Linear Trend:

The trend is linear when fitting a linear model to our raw data, it gives a negative slope with estimates of -2.1163 and an intercept at 4315. Meaning that every half month, the people's interests in ballet performance are regressing a small amount. This can be explained by a possible interest shift towards other dance such as Korean pop dance and Zumba. If I type either

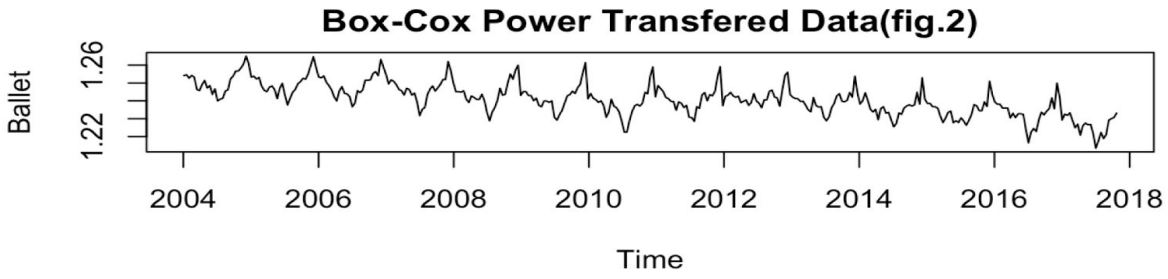
of these keywords into Google Trend, both sets of data have a steady upward growing random walk tendency. The estimates of the parameters are significant given that their P-values are sufficiently less than 0.05, and the F-statistics corresponds with a P-value less than 0.5 also indicate the significance of a linear trend. However, by looking at R-squared, only 39% of the residuals can be explained. The long right tail of the residuals histogram suggest that we need to find a better model that can explain these results more in detail.

```
Call:
lm(formula = Y_ts ~ time(Y_ts), na.action = NULL)

Residuals:
    Min      1Q  Median      3Q     Max
-20.563  -6.464  -1.572   3.419  43.590

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4315.1728   278.6367   15.49   <2e-16 ***
time(Y_ts)    -2.1163     0.1386  -15.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.52 on 359 degrees of freedom
Multiple R-squared:  0.3939,    Adjusted R-squared:  0.3922
F-statistic: 233.3 on 1 and 359 DF,  p-value: < 2.2e-16
```

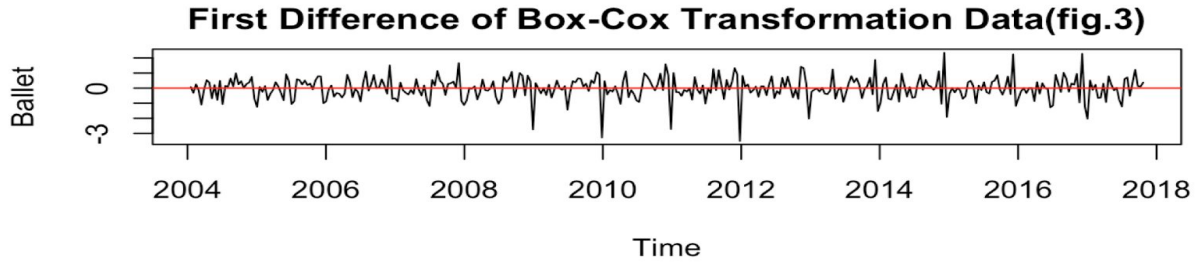


(2). Seasonal ARIMA Model:

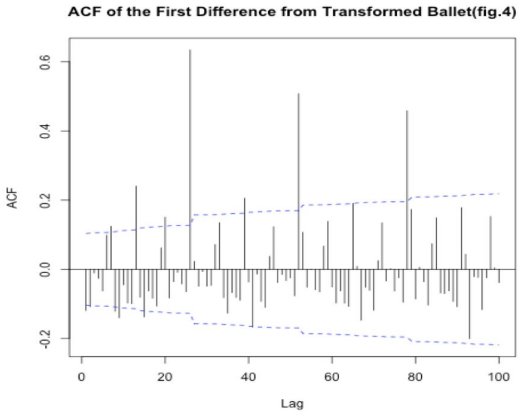
We know from earlier, the data has a linear trend. The KPSS test for stationarity, P-value is 0.01, indicates that we reject the null hypothesis that the data is stationary. However, the current data might also have been affected by extreme observations such as low interests in the middle and observations clustered between 220 to 240. To stabilize the data, we take a Box-Cox transformation with a -0.7700342 lambda value, and subsequently the graph looks like fig.2.

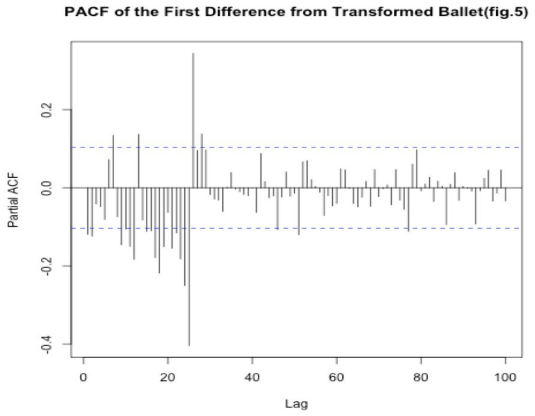


We apply the Dickey-Fuller test for stationarity, and it produces a 0.962 P-value, which suggests stationarity after taking the difference (see fig.3). Now, we see our data bouncing around a zero-mean.



To fit the above data into an ARIMA model, I evaluated the ACF(fig.4), PACF(fig.5) and EACF table. From the ACF, there is a temporary "cut off" at lags 26, 52, and, 78 with most agreement of "tails off" after these lags, and PACF also indicates a "cut off" after lag 26, so it might be just an AR model with seasonal period of 26. However, most of lags before lag 26 in the PACF have significant correlations. Since our data starts in the winter season, these can be explained by the negative correlation of seasonal interests within a year right after Christmas. We can conclude some of the questions with introspection: Ask yourself a simple question, if you have visited a ballet performance, will you visit again within the same month? It is extremely unlikely for the general audience, and the normal spring ballet season starts two months later.





The huge periodic “spikes” are due to the existence of the Christmas season at the end of the year, and I do not think I should treat these spikes as “outliers” to prewhiten the series because these “spikes” are indeed a part of ballet culture.

Fun fact: The Nutcracker ballet is annually performed around December, and this is where the “spikes” of our data came from. This makes sense because people have days off during the Christmas season. Additionally, this time coincides with frequent performances of the Nutcracker, a traditional Christmas ballet performance that attracts entire families.

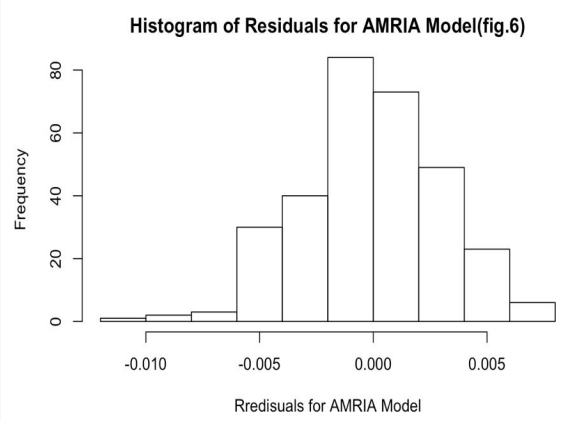
To continue the investigation, I analyzed the EACF below. It appears to indicate parameter combinations such as ARIMA(1,1,1)x(1,0,0)[26], ARIMA(2,1,2)x(1,0,0)[26], and several others. In order to compare AIC values and find the smallest one, I use the auto.arima function in R which suggests an ARIMA(2,1,0)x(1,0,0)[26] model with AIC= -3102.26. Notice that all parameter estimates are significantly different from zero, so I think there is no overfitting. However, this model is tentative since it completely ignores the MA part which I cannot infer from the EACF table.

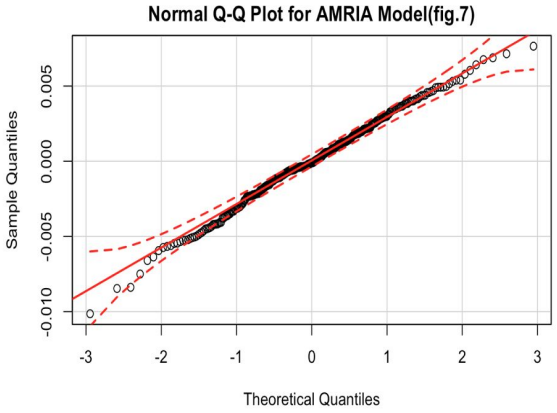
```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x o o o o x x x o o  o  x  o
1 x o o o o o x o o o o  o  x  x
2 x x o o o o o o x o o  o  o  x
3 x o x o o o o o x o o  x  o  x
4 x o x x o o o o o o o  o  o  x
5 x x o x o o o o o o o  o  x  x
6 x x x o o o o o o o o  x  x  x
7 x x x o x x o o o o o  o  o  x
```

```
Series: newY
ARIMA(2,1,0)(1,0,0)[26]

Coefficients:
          ar1      ar2    sar1
      -0.3385  -0.2685  0.7499
s.e.   0.0518   0.0513  0.0343

sigma^2 estimated as 9.841e-06:  log likelihood=1555.13
AIC=-3102.26   AICc=-3102.15   BIC=-3086.72

Training set error measures:
                      ME        RMSE         MAE           MPE      MAPE      MASE
Training set -2.680791e-05 0.003119561 0.002431852 -0.002656336 0.1960694 0.7456881
                    ACF1
Training set -0.04810039
```

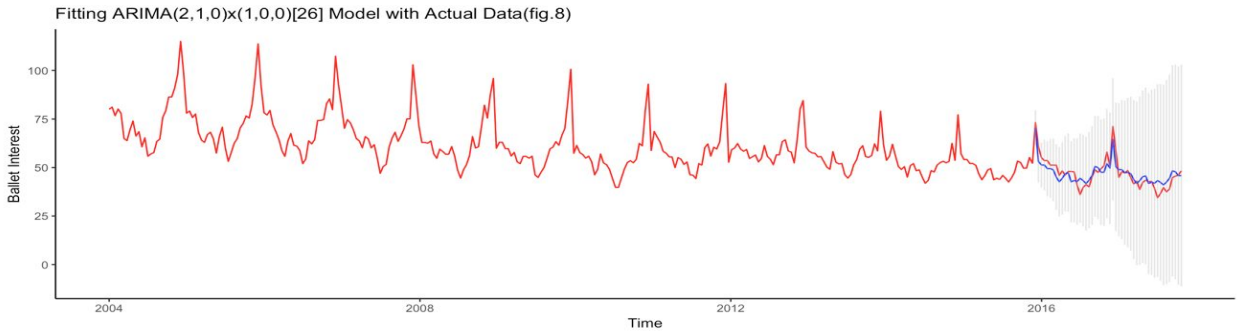
We need to diagnose the suggested model from auto.arima. Luckily, the Ljung-Box test has a P-value 0.3581, which indicates that we do not have significant autocorrelations. Checking normality from the histogram of residuals in fig.6, we see an almost bell-shaped curve, and the QQ-plot in fig.7 indicates a normal distribution with only 1 point outside of the 95% confidence

level. The Shapiro-Wilk test gives a P-value of 0.4149, hence normality is confirmed. The residuals of the seasonal ARIMA model is independent and normally distributed, so the ARIMA(2,1,0)x(1,0,0)[26] model fits the ballet data.





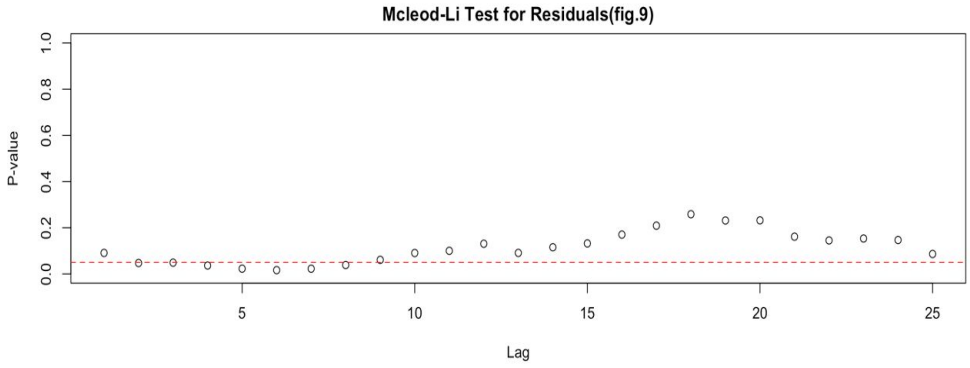
Let us visualize the fitting of ARIMA(2,1,0)x(1,0,0)[26] Model from the year 2016 to 2017 with 52 points, and see how well they can match up with the actual data over these years. The blue curve in fig.8 is the fit from our model, the gray area is the confidence interval, and the red curve is our raw data. Clearly, the model does a good job describing the real world since it almost completely coincides with our actual data.



(3). ARCH/GARCH Effect:

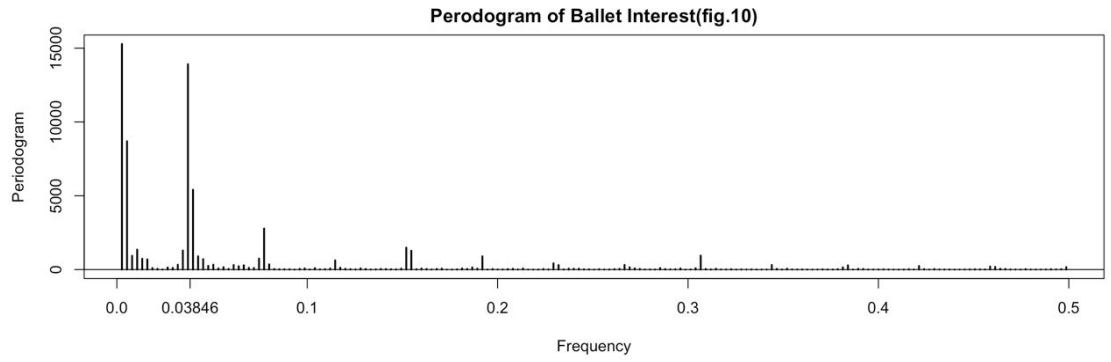
At this point I am suspicious of a potential GARCH effect from the pattern of alternating quiet and volatile periods as seen in fig.3. The McLeod-Li test in fig.8 shows that most of the

lags are above significance level, and only 5 of them are slightly below. I do not think this supports a strong evidence for conditional heteroscedasticity, so here I decide to ignore them. However, This result might change when I use weekly data instead of bi-weekly, because GARCH effect generally applies to a larger data set.



(4). Naïve Harmonic Model:

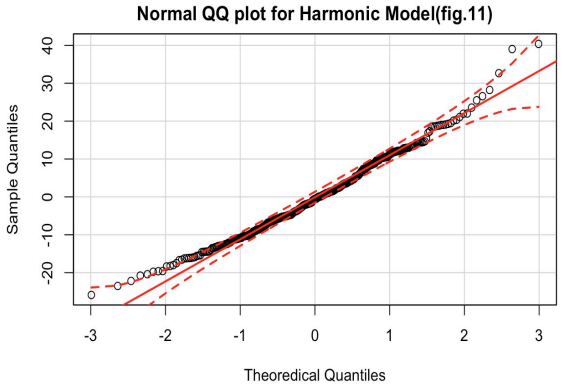
Any time series data might have "hidden periodicities". Since my ballet data is identified as a seasonal data with period 26, we can also examine this in the Periodogram of my raw data.



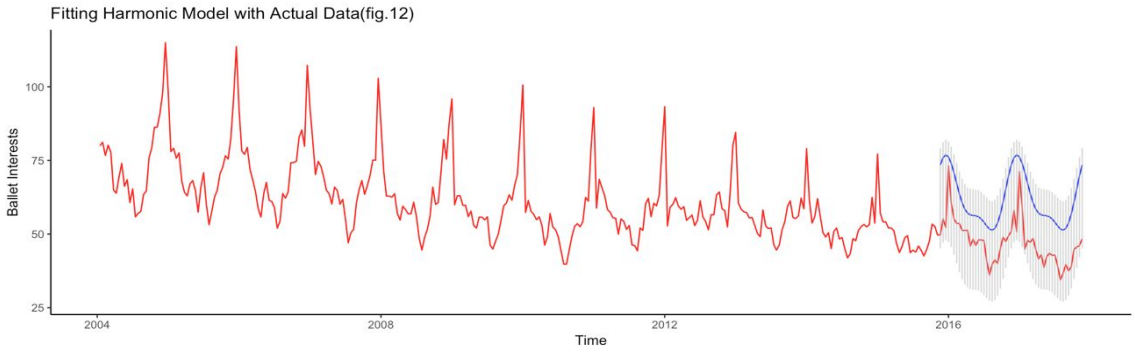
Here, we see two clear prominent peaks. One frequency is very close to zero, and can be explained by our downward sloping trend that gives rise to about half of the points with their own short periods, which further supports our ARIMA(2,1,0)x(1,0,0)[26] model with the pure AR part. The other frequency is closer to 0.03846, which captures about another half of our

observations with the biweekly period of 26. Besides these two dominant bars, there are a few smaller ones with frequencies spread out. This suggests a fit with two or three cosine-sine combinations for a harmonic model. Comparing R-squared values, two cosine-sine fit gives 0.41 and three cosine-sine fit gives 0.42. Of course, the more combinations of cosine-sine pairs, the more we can capture in little detail. Since they do not largely differ, I decided to fit a two cosine-sine combination. Below is a summary with the estimated parameters, notice that all are significant, except the second cosine estimate. It is interesting when comparing the R-squared with the linear model, there is a 2% improvement due to the naïve harmonic fit when explaining the residuals, so the shape bests the trend in terms of fit? Checking with the QQ plot, most of residuals are within the 95% confidence band, but there are some "irregularities" on the tails as behaviors may be affected by the downward linear trend.

```
Call:
lm(formula = har_predY ~ har)

Residuals:
    Min      1Q  Median      3Q     Max
-24.301  -7.019  -0.723   7.060  38.202

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   61.6252     0.5757 107.041  < 2e-16 ***
harcos(2*pi*t)  10.3946     0.8163  12.733  < 2e-16 ***
harcos(4*pi*t)   1.2091     0.8149   1.484    0.139
harsin(2*pi*t)  -3.8402     0.8120  -4.729 3.45e-06 ***
harsin(4*pi*t)  -4.2834     0.8135  -5.266 2.64e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.15 on 306 degrees of freedom
Multiple R-squared:  0.4097,    Adjusted R-squared:  0.402
F-statistic:  53.1 on 4 and 306 DF,  p-value: < 2.2e-16
```

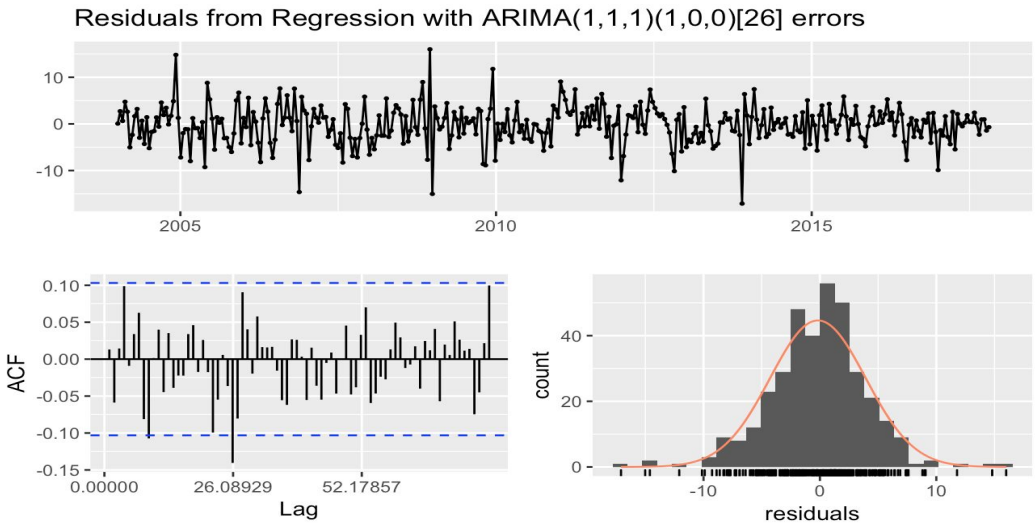


The fitting is static and not changing over time, so it can only capture the general shape, seen in fig.12. Of course, the Shapiro-Wilk test with P-value 0.001 rejects this naïve approach, which further suggests that we need to modify our model to capture the trend.



(5). Dynamic Harmonic Regression Model:

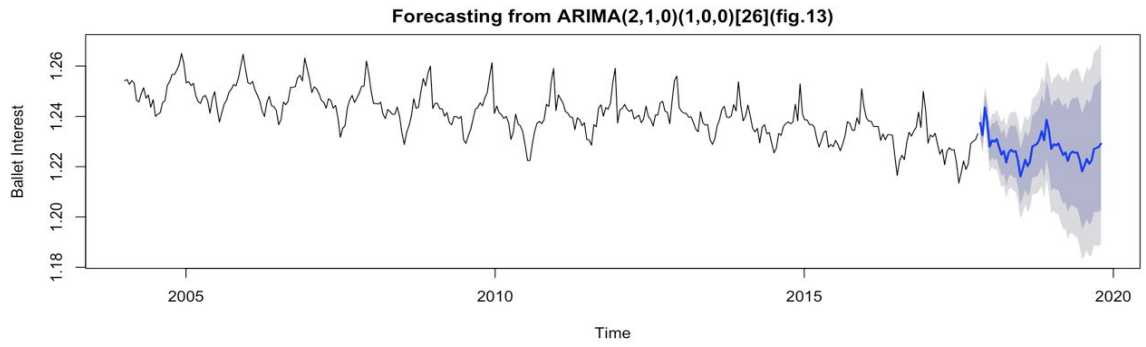
To improve on the naïve Harmonic Model, I updated to a Dynamic Harmonic Regression model with ARIMA errors under the assumption of unchanging seasonality: $Y_t = \beta_0 + \sum_i^n \beta_i x_{t,i} + \sum_{j=1}^m [A_i cos(2\pi f_j t) + B_i sin(2\pi f_j t)] + e_t$, which contains both the regression part and the harmonic part with e_t set to be an ARIMA process instead of white noise. I do not think fixing the seasonality is a large disadvantage compared to our ARIMA(2,1,0)x(1,0,0)[26] model earlier since we do not have a very long time series. Using auto.arima again accommodating with pairs of Fourier terms in R. We have the following results with an ARIMA(1,1,1)(1,0,0)[26] errors.

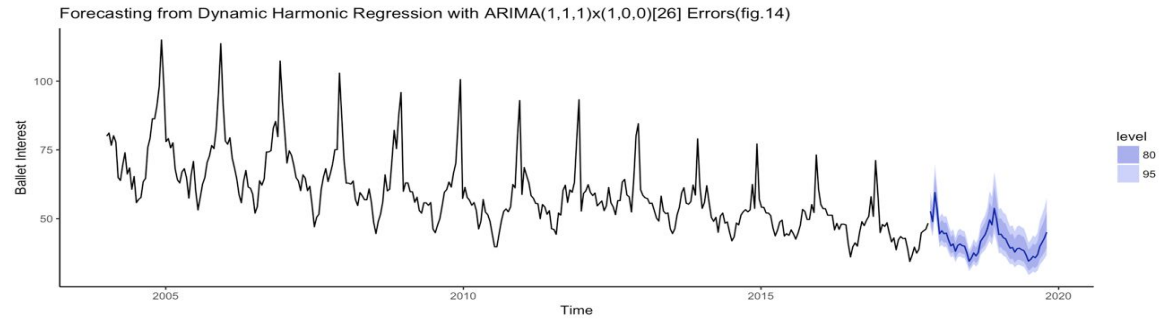


The P-value of 0.2857 from Ljung-Box test suggests that the residuals are uncorrelated with the only exception appearing in the ACF at lag 26, and this might be due to the slightly downward trend affecting the next year during the same season. The residuals look normal but with long tails, so let us keep this model for the moment, and judge it later after forecasting!

VI. Forecasting:

After comparing the model fittings, I think the ARIMA(2,1,0)x(1,0,0)[26] and Dynamic Harmonic Regression with ARIMA(1,1,1)(1,0,0)[26] errors can both serve for forecasting, see fig.13 and fig.14. Here, I use 361 bi-weekly training data for the past 13 years to predict next two years (2018 and 2019).





Notice that the Dynamic Harmonic Regression model makes the predictions within a finer confidence interval, and both models capture the shape and the trend of our raw data.

VII. Conclusion:

To answer my question concerning the general audience appreciation of ballet performances over the next two years, I think Dynamic Harmonic Regression Model with ARIMA(1,1,1)x(1,0,0)[26] did a slightly better job than ARIMA(2,1,0)x(1,0,0)[26] model in terms of the confidence regions, and I am glad we kept it. Statistics! They both can offset the insufficiency of linear model and naïve harmonic model by adequately capturing the decreasing tendency while preserving the seasonal shape in great detail. Who knows, at some point, ballet culture may suddenly bloom again. The seasonal patterns throughout the year is obviously peaking at Christmas every year, and the forecasts from both models justified my concerns on the subject of the general appreciation of ballet. I can still take heart that I will definitely be able to find my generation if I sit in Nutcracker performance this December.

For further investigation of ballet performances, I can search for other sources than Google Trends, which collect data from more than just searches about the general term of ballet. Sub-categories such as "ballet shoes" and "ballet tutus" and "ballet performances" all fall under the hierarchical category of "ballet" and some of these sub-categories may not predict anything about interest. Additionally I would love to have retrieved the full 13 year dataset in one query with a common scale. Instead I modified the raw data to map the sum of the weekly means to a monthly dataset containing all 13 years on a common scale. If I have access to more specific data on real attendance at performances, I might be able to judge whether the trends are in fact due to interest in performances.

Finally, thank you Professor Brillinger for your guidance during office hours and reading my paper this far!