Time Series Forecasting of Harvard University Total Yearly Cost.

Abdirashid Chorshanbiyev
s314800@studenti.polito.it

July 17, 2024

Introduction

The "harvard fees.csv" dataset is the dataset about the yearly total expenses for students at Harvard University, covering the period from 1985 to 2016. This study aims to construct a time series analysis to project the future financial requirements for attending Harvard from 2017 to 2023. The predictive outcomes will then be aligned with actual costs available on Harvard's official site to assess the precision and effectiveness of the forecast model.

Preprocessing

First, we need to check whether our time series is stationary or not.We need to manipulate our data so that it resembles a time series with stationarity. A time series has stationarity if it has constant standard deviation from a mean. If it is not stationary, we will make some transformations to meet our time series stationarity requirement. By looking at the shape of our time series, there is a clear upward trend in data but no obvious seasonality. The upward trend in our data implies that it is non-stationary.

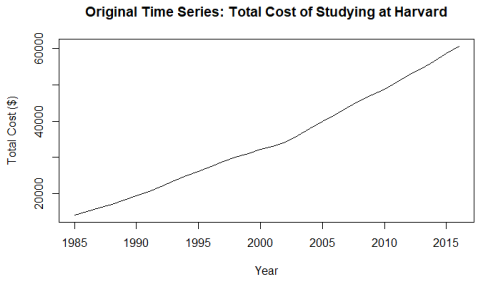

Figure 1: Total Cost of Studying at Harvard between the years 1985 and 2016

Differencing

Differencing is a method used in time series analysis to make non-stationary time series data stationary. This technique involves subtracting the previous observation from the current observation. The differenced series is the change between consecutive observations in the original series and can be written as

$$y'_t = y_t - y_{t-1}$$

Again we will plot our time series data after differencing , after first differencing data is looking to be slightly more stationary. It appears that the differenced values are lower in the first half of the series and higher in the second half.

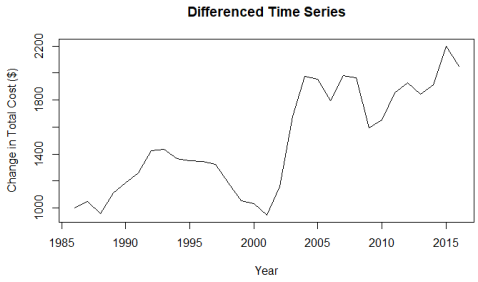

Figure 2: Time series data after first differencing

We tried to make the time series more stable around its mean by applying second differencing.The graph below is more telling of a time series with stationarity. It has an obvious mean of around 0 and doesn't have any clear trends other than spiking up and down. So it has constant variance from inspection.

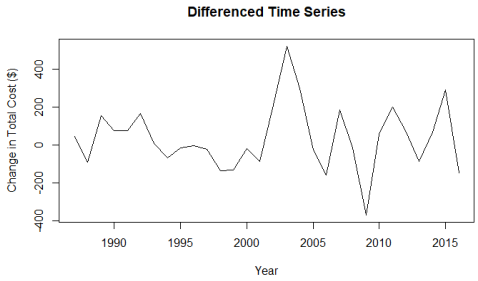

Figure 3: Time series data after second differencing

Now we plot the ACF for the first and second differenced data. The ACF tells us how correlated the values in a time series are with each other. On the y-axis we plot the correlation coefficient, and on the x-axis we plot the amount of lag which is measured in units of time for the time series. We have measures of tuition fee for each year in our data, so 1 value of

lag is one year in the past. To clarify, for a lag of 0 the ACF will always have a value of 1. The data points will always perfectly correlate with themselves with 0 errors.

By looking at ACF plots of first differenced data, the plot shows significant spikes at lags 1, 2, and 3. This pattern suggests the presence of short-term correlations in the data. After lag 3, the autocorrelations drop off, indicating that the data does not have long-term dependencies. This result tells me that if we use first differencing in our ARIMA model, we should put q=1, 2 and also try other values.

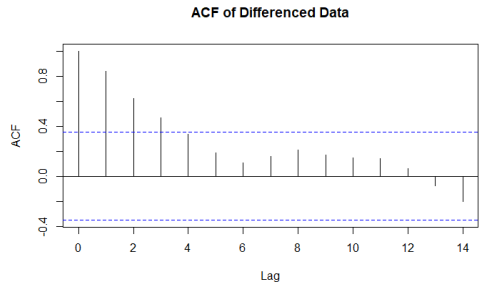

Figure 4: ACF plot of first differenced data

From the second differenced data's ACF plot, it is visible that our data is probably over-differenced. I say this because after a lag of 1 and 2 we immediately have a negative ACF value. A negative ACF value in this context means that if we were to have our value increasing at one point in time, it is highly likely that at the next period of time our value will be decreasing. Knowing the nature of our dataset, this does not make sense since the values seem to be rising at a constant rate.In addition, almost all lags lie inside the boundary which indicates the time series is white noise, but we are confident that there is a clear upward trend and data is predictable So the ACF graphs suggest using the first difference for ARIMA, or a value of d=1.

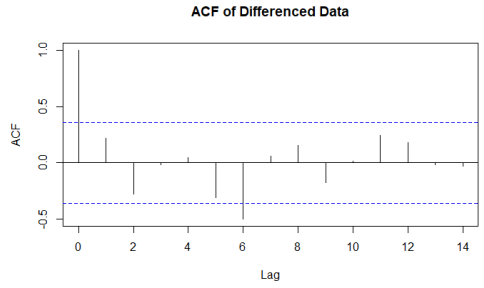

Figure 5: ACF plot of second differenced data

PACF (Partial Autocorrelation Function) Plot shows the correlation between a time series and its lags, controlling for other lagged variables, helps identify direct relationships between observations at different time lags and useful for determining the order of the AR

(Autoregressive) component in ARIMA models. Above we mentioned that we use first differencing so for this reason we plot PACF plot of first differenced data. Looking at the PACF plot for the first differenced data only the first lagged value has a significant correlation with current values. Since the correlation for other lagged values is so low, this tells us that the correlation between other lagged values and the current value in our ACF plot can simply be explained by only the first lagged value. Let's try using p=1,2 values for the MA part.

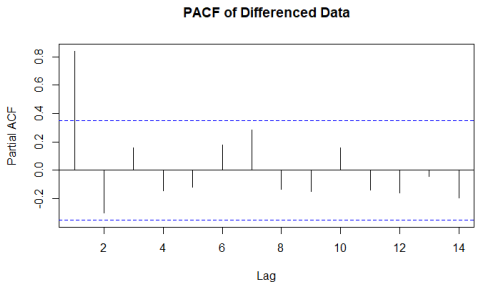

Figure 6: PACF plot of first differenced data

Modelling

In order to choose an appropriate model for our time series, different models should be checked, and their outputs should be compared in order to select the best among them. We are seeking to train an ARIMA(p,d,q) model, which is essentially an ARMA(p,q) model enhanced by an integration factor. Integration needs to be specified if the data is non-stationary, but in our scenario, we're setting d=0 because our data have already been differenced.

$$X_t - \varphi_1 X_{t-1} - \ldots - \varphi_p X_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} + c$$

Where ε_t is a centered (zero mean) white noise. This process is called an ARMA(p,q) sequence. In our case, it is not easy to select an appropriate model because of the size of the small dataset. For this reason, we will try different models to select the best among them.

We selected the optimal model by looking at Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).Lower AIC and BIC values indicate better model fit, balancing goodness-of-fit with model complexity. We tried various models by trying various values of p and q. The table provided encapsulates the BIC and AIC values for a series of models. From the table we choose ARMA(2,1) because of smaller AIC and BIC values and less parameters compared to higher order ARMA.

Model	AIC	BIC
ARMA(2,1)	407.9	413.64
ARMA(2,2)	408.15	415.32
ARMA(3,3)	407.24	417.28

Table 1: Model comparison based on AIC and BIC values

Forecasting

The fitted model is used for forecasting future values of Harvard tuition fee in next 7 years from 2017 to 2023. Below we can see how much our model makes predictions closer to actual values. As we see from graph ,prediction error increases for more distant years in this forecast widening difference between actual and predicted values. But this is natural because as we predict further into the future, there are more unknown factors that can influence the outcome, leading to increased uncertainty. But overall, predicted values are close to actual values. We also measured how precise our model is by several error metrics such as RMSE error 132.6643 and MAPE around 2.85 percent which is not a bad result.The gray area is 95 percent confidence interval of the model's predictions. On inspection the model does not perform too poorly at all. It captures the general trend, and the actual population is well within the prediction's confidence interval

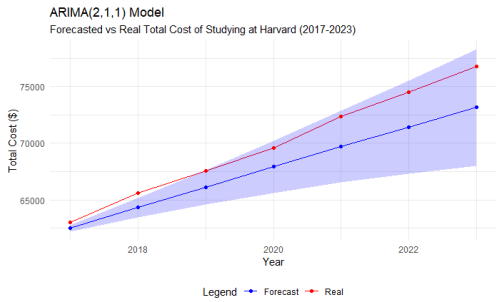

Figure 7: Comparision between actual and forecasted values from 2017 to 2023 years

Conclusion

To effectively apply an ARIMA model to a time series, initial preprocessing steps are necessary to ensure the data is stationary. The most intricate aspect of this modeling is not merely selecting from various models but rather determining the appropriate values for the parameters p, d, and q, which underpin the model's assumptions. Once models are tailored to fit the data, they can be employed to predict future observations. The selection of model for the dataset is then made by evaluating error measurements and observing visual trends.