NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA

FINAL REPORT

Study on the prediction of U.S Airline available seat miles and passenger revenue using various forecasting models

---

Under the guidance of

Dr.Savitha Bhat

Assistant Proffesor
School Of Management

Submitted by

Mansi S Sheth

16MT25

Contents

1 Dataset

The data for analysis of different models was obtained from here. The data description is provided here.and here.

It consists of data on Monthly U.S Airline Available Seat Miles and Revenue from 1978-1 to 2004-1.Number of observations-300.I converted the monthly data into quarterly data (taking the average of 3 months data for each quarter) and choose Available Seat Miles in Miles as my target variable. After conversion total Number of observations - 100. I took the latest 88 points for analysis.For time series analysis, seasonality and trend analysis I took the 76 points for training and next 12 points for testing. For regression analysis both time and Revenue are taken as independent variables.

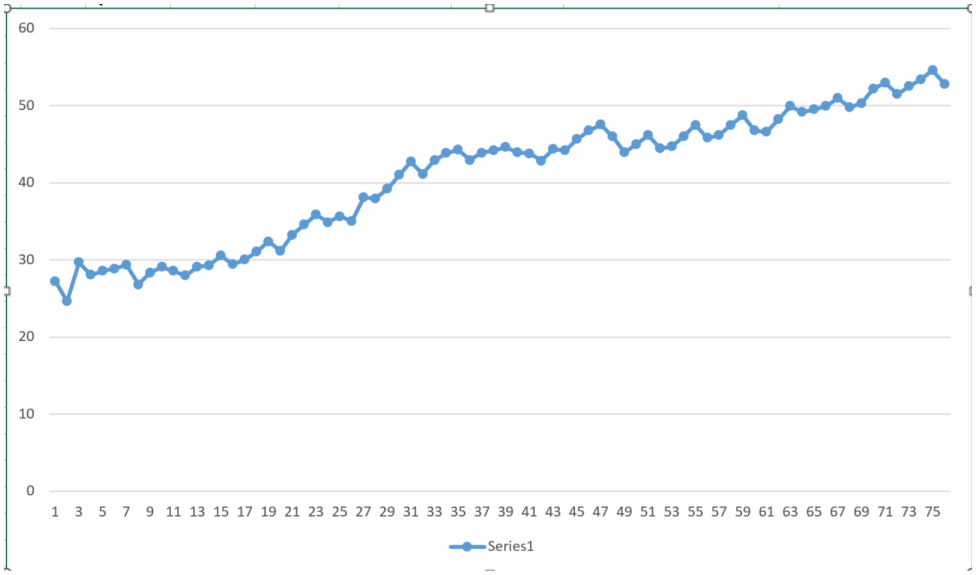

Figure 1: Line plot of target variable.

2 Proposed Methodology

Smoothing techniques,time series analysis and regression analysis were applied. For seasonality analysis a season was taken as one quarter. For regression analysis time and revenue both were considered as independent variables.

2.1 Smoothing techniques.

2.1.1 Moving Average

A technique that averages a number of recent actual values updated as new values become available.In my current dataset I took 4 point Moving average.

$$F_t = MA_n = \frac{A_{t-n} + \dots A_{t-2} + A_{t-1}}{n}$$

where

F_t = Forecast for time period t

Ma_n = n period moving average

A_{t-1} = Actual value in period t-1

n = Number of periods (data points) in the moving average

2.1.2 Exponential smoothing.

It is a weighted averaging method based on previous forecast plus a percentage of the forecast error.

Next Forecast = Previous *Forecast* + α (Actual – Previous Forecast)

Where, (Actual – Previous) = the forecast error and

α = weightage of the error.

$$F_t = F_{t-1} + \alpha (A_{t-1} - F_{t-1}) \quad \text{OR}$$

$$F_t = (1- \alpha) F_{t-1} + \alpha A_{t-1}$$

Where F_t = Forecast for period t

F_{t-1} = Forecast for the previous period (i.e. period t-1)

α = Smoothing constant ($0 <= \alpha <= 1.0$)

A_{t-1} = Actual demand or sales for the previous period

For the current dataset to illustrate exponential smoothing the alpha values taken were 0.1 and 0.4. It gave better results when alpha was choosen to be 0.4

2.2 Trend Analysis.

Analysis of trend involves developing an equation that will suitably describe trend.The trend component may be linear, or it may not. Some commonly encountered non-linear

trend types are parabolic, exponential, growth curve.Here I tried to fit a linear trend curve.

Linear Trend Line [Can be derived using Calculus in the same way as derived for an OLS regression line with t == x & F_t = y]

$$b = \frac{n\sum(ty) - \sum t\sum y}{n\sum t^2 - (\sum t)^2}$$

$$a = \frac{\sum y - b\sum t}{n}$$

Where,

n = Number of periods

y = Value of the time series

The values of coefficients from above equations was found out to be b=0.371459695 and a=26.61902105.The trend equation is given as Average seat mile = 26.619 + Time*0.37145

2.3 Seasonality analysis.

The regular repeating movements (upward or downward) in series values that can be tied to recurring events is called Seasonal Variations in time series. Each quarter was taken as one season.Multiplicative and Additive seasonality was computed.Multiplicative seasonality gave poor results.

Quater	Multiplicative Seasonal Index	Additive Seasonal Index
Q1	0.9921896	-0.348796296
Q2	1.0031407	0.124166667
Q3	1.024975	0.964328704
Q4	0.9827692	-0.653888889

3 Interpretation of Results.

Since the dataset was divided into train and test data , the errors were computed for test data.To compute Mean absolute difference, Mean square error, Mean absolute percentage error the following equations were used.

$$\text{Mean Absolute Deviation (MAD)} = \frac{\sum |\text{Actual} - \text{forecast}|}{n}$$

$$\text{Mean Square Error (MSE)} = \frac{\sum (\text{Actual} - \text{forecast})^2}{n - 1}$$

$$\text{Mean Absolute Percent Error (MAPE)} = \frac{\sum (|\text{Actual} - \text{forecast}| / \text{Actual} * 100)}{n}$$

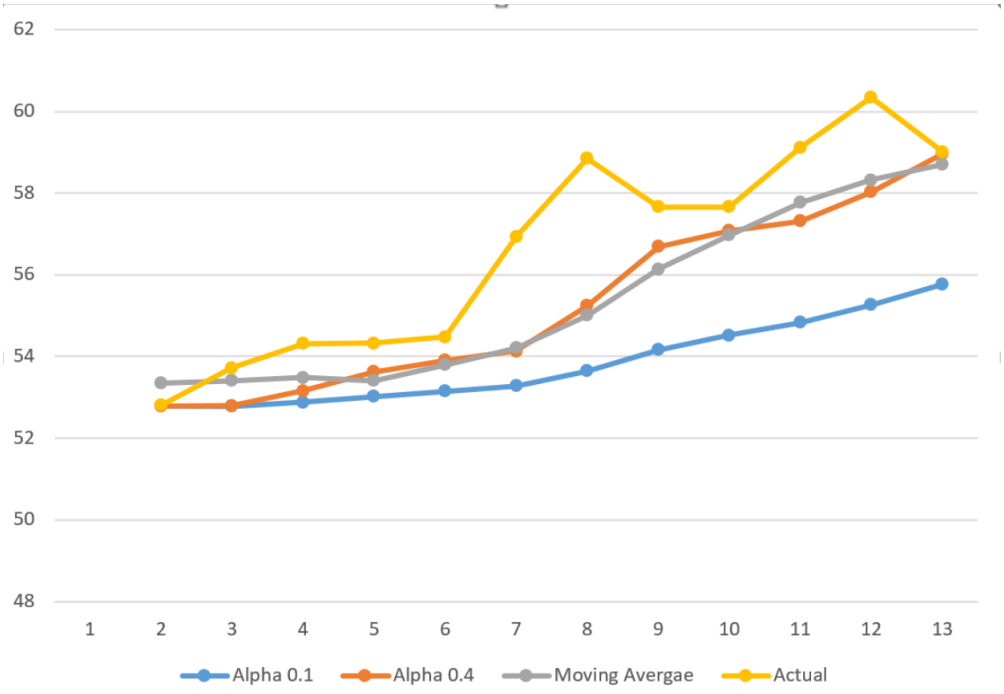

Figure 2:Line plot of actual and predicted values using smoothing techniques.

The exponential smoothing using alpha 0.1 gave the worst fit.

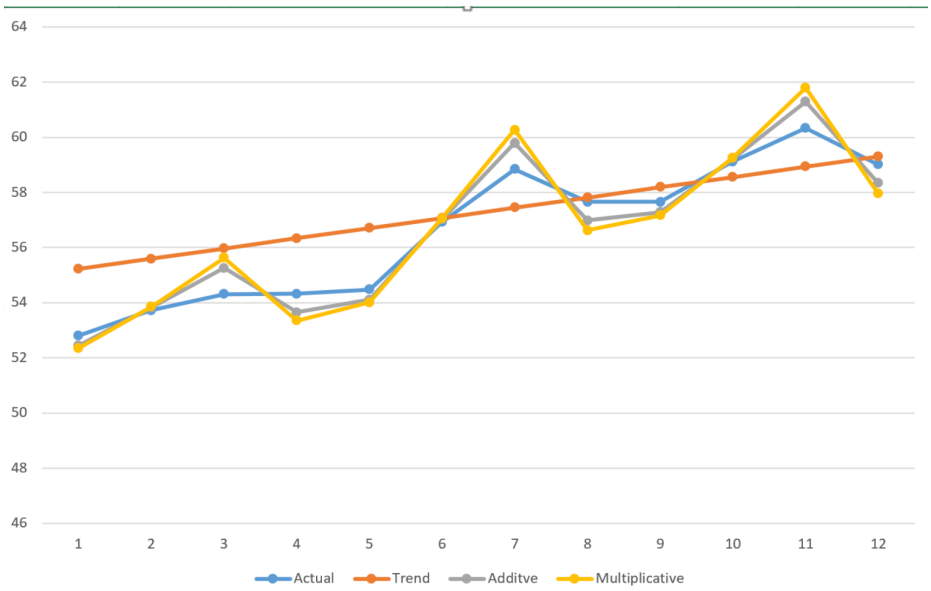

Figure 3:Line plot of actual and predicted values using trend and seasonality analysis

The error analysis for all 6 methods is given in below table. Additive seasonality gave the best fit.

	MAD	**MAPE**	**MSE**
Moving Average(4)	3.933333958	11.08294757	6.84226002
Exponential Smoothing a=0.1	8.279751693	41.16822197	14.29530807
Exponential Smoothing a=0.4	3.874411228	11.20829649	6.73340841
Trend	3.665465095	8.566185146	6.61904951
Additive Seasonality	1.568385417	1.492735898	2.756850173
Multiplicative Seasonality	2.272216054	3.253669297	3.983698248

4 Pearson's Correlation

Pearson r correlation is the most widely used correlation statistic.A Pearson correlation is a number between -1 and 1 that indicates the extent to which two variables are linearly related. The Pearson correlation is also known as the "product moment correlation coefficient" (PMCC) or simply "correlation".

$$r = \frac{N \sum xy - \sum (x)(y)}{\sqrt{N \sum x^2 - \sum (x^2)][N \sum y^2 - \sum (y^2)]}}$$

r= Pearson r correlation coefficient between x and y

N= number of observations

xi= value of x (for ith observation)

yi= value of y (for ith observation)

	Pearson's Correaltion
Quarter and Available seat mile	0.96861373
Quarter and Revenue	0.974637314
Available Seat Mile and Revenue	0.955760773

Quarter and revenue show good correlation (r is close to 1).

5 Regression Analysis.

In statistical modeling, regression analysis is used to estimate the relationships between two or more variables.Regression analysis helps understand how the dependent variable changes when one of the independent variables varies and allows to mathematically determine which of those variables really has an impact. Technically, a regression analysis model is based on the sum of squares, which is a mathematical way to find the dispersion of data points. The goal of a model is to get the smallest possible sum of squares and draw a line that comes closest to the data.

5.1 Regression analysis output: Summary Output

Multiple R. It is the Correlation Coefficient that measures the strength of a linear relationship between two variables. The correlation coefficient can be any value between -1 and 1, and its absolute value indicates the relationship strength. The larger the absolute value, the stronger the relationship:

1. 1 means a strong positive relationship
2. -1 means a strong negative relationship
3. 0 means no relationship at all

R Square. It is the Coefficient of Determination, which is used as an indicator of the goodness of fit. It shows how many points fall on the regression line. The R2 value is calculated from the total sum of squares, more precisely, it is the sum of the squared

deviations of the original data from the mean.Generally, R Squared of 95% or more is considered a good fit.

Adjusted R Square. It is the R square adjusted for the number of independent variable in the model. You will want to use this value instead of R square for multiple regression analysis.

Standard Error. It is another goodness-of-fit measure that shows the precision of your regression analysis - the smaller the number, the more certain you can be about your regression equation. While R2 represents the percentage of the dependent variables variance that is explained by the model, Standard Error is an absolute measure that shows the average distance that the data points fall from the regression line.

Observations. It is simply the number of observations in your model.

5.2 Regression analysis output: ANOVA

Basically, it splits the sum of squares into individual components that give information about the levels of variability within your regression model:

1. df is the number of the degrees of freedom associated with the sources of variance.
2. SS is the sum of squares. The smaller the Residual SS compared with the Total SS, the better your model fits the data.
3. MS is the mean square.
4. F is the F statistic, or F-test for the null hypothesis. It is used to test the overall significance of the model.
5. Significance F is the P-value of F.

The ANOVA part is rarely used for a simple linear regression analysis in Excel. The Significance F value gives an idea of how reliable (statistically significant) your results are. If Significance F is less than 0.05 (5%), your model is OK. If it is greater than 0.05, you'd probably better choose another independent variable.

5.3 Regression analysis output: coefficients

This section provides specific information about the components of the analysis.The most useful component in this section is Coefficients. It enables to build a linear regression equation in Excel.

6 Regression Results

The regression analysis was done first taking only time as independent variable and then only income as independent variable and then taking both as independent variables.

	A	B	C	D	E	F	G	H
1	SUMMARY OUTPUT							
2								
3	*Regression Statistics*							
4	Multiple R	0.982761797						
5	R Square	0.96582075						
6	Adjusted R Square	0.964884332						
7	Standard Error	1.586985942						
8	Observations	76						
9								
10	ANOVA							
11		*df*	*SS*	*MS*	*F*	*Significance F*		
12	Regression	2	5195.20895	2597.604475	1031.399377	3.03621E-54		
13	Residual	73	183.8522796	2.518524379				
14	Total	75	5379.06123					
15								
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	
17	Intercept	15.61619607	1.479292088	10.55653322	2.41002E-16	12.66797119	18.56442095	
18	Quarter	0.164413321	0.028211052	5.827975504	1.41297E-07	0.108188775	0.220637866	
19	Revenue	0.735376879	0.095765885	7.678902336	5.64021E-11	0.544515745	0.926238013	
20	Trend Equation	Available Seat Mile = 15.6161 + 0.1644*Quarter+ 0.73537*Revenue						
21								

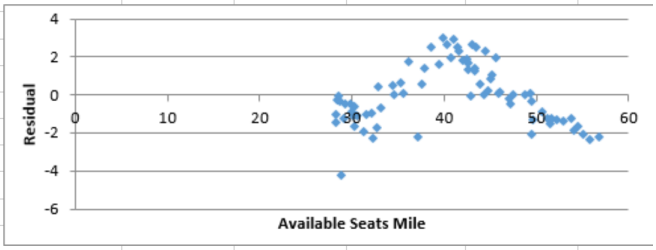

Figure 4:Regression output taking just time as independent variable and Available Seat mile ad dependent variable

R square value of 0.965 means that 96.5% of the variation of y-values around the mean are explained by the x-values.In other words, 96.5% of the values fit the model.

	A	B	C	D	E	F	G	H
1	SUMMARY OUTPUT							
2								
3	*Regression Statistics*							
4	Multiple R	0.955760773						
5	R Square	0.913478656						
6	Adjusted R Square	0.912309448						
7	Standard Error	1.926399658						
8	Observations	76						
9								
10	ANOVA							
11		*df*	*SS*	*MS*	*F*	*Significance F*		
12	Regression	1	2899.343361	2899.343361	781.2802859	4.55542E-41		
13	Residual	74	274.6151574	3.711015641				
14	Total	75	3173.958519					
15								
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	
17	Intercept	14.96215789	0.446343976	33.52158581	1.7461E-46	14.07279811	15.85151768	
18	Quarter	0.281551378	0.010072893	27.95139148	4.55542E-41	0.261480701	0.301622056	
19	Trend	Revenue = 14.96215 + 0.28155* Quarter						
20								

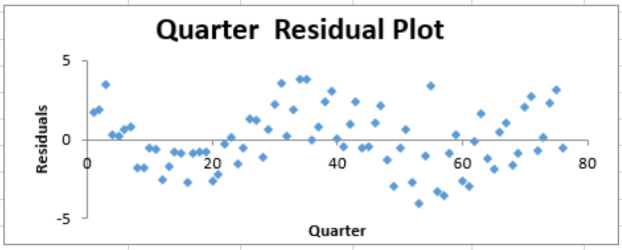

Figure 5:Regression output taking time as independent variable and revenue as dependent variable

R square value of 0.913 means that 91.3% of the variation of y-values around the mean are explained by the x-values.In other words, 91.3% of the values fit the model.

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.96861373					
5	R Square	0.938212558					
6	Adjusted R Square	0.937377593					
7	Standard Error	2.119275859					
8	Observations	76					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	5046.702797	5046.702797	1123.654376	1.74839E-46	
13	Residual	74	332.3584323	4.491330167			
14	Total	75	5379.06123				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	26.61902105	0.491033109	54.21023666	2.57664E-61	25.64061623	27.59742588
18	Quarter	0.371459695	0.011081418	33.52095429	1.74839E-46	0.349379488	0.393539901
19	Trend Equation	Available Seat Mile = 26.61902105+ 0.3714596*Quarter					
20							

Figure 6:Regression output taking time and income as independent variable

R square value of 0.938 means that 93.8% of the variation of y-values around the mean are explained by the x-values.In other words, 93.8% of the values fit the model.

1. All 3 models are statistically significant as their p values are less then 0.05 .
2. Time and expenditure shows good correlation coefficient.
3. All the 3 models are satistically significant (significance F).

7 References.

References

[1] **Dataset Link**

http://users.stat.ufl.edu/~winner/datasets.html

[2] **Datset Description**

http://users.stat.ufl.edu/~winner/data/air_rpm.dat

http://users.stat.ufl.edu/~winner/data/air_rpm.txt

[3] **Excel sheet Link**

https://1drv.ms/x/s!AjtcjLaiP7PzjnGyDeFcv6JvPfN-