Forecasting using Extrapolative and Explanatory Methods

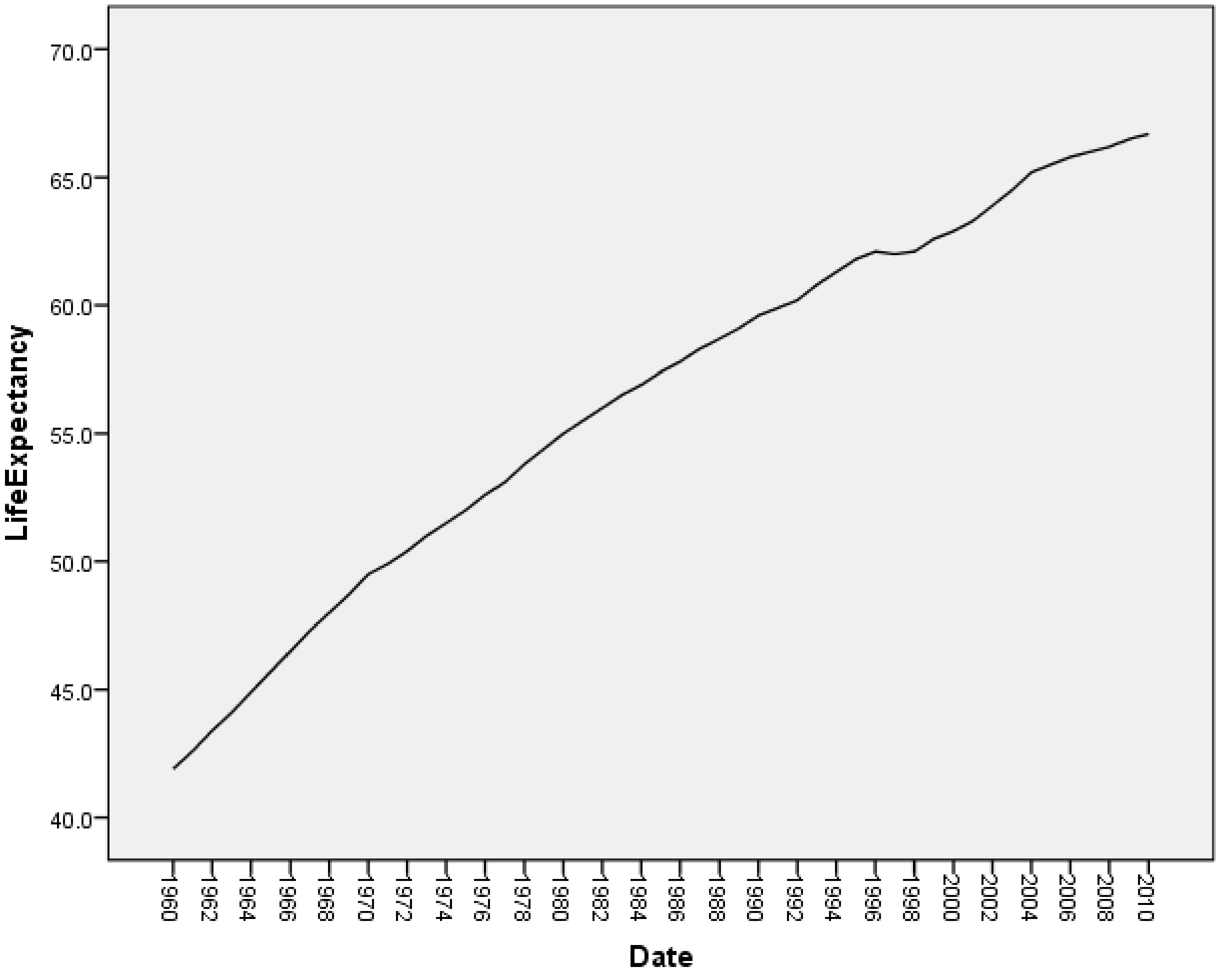
By

Amit Behura

Objective: We are forecasting the value of life expectancy from the data collected from gapminder website. We will be using multiple forecasting model and comparing respective accuracy to find the best fit model.

Data: Data we have used is collected from Gapminder website. With 1960-2010 as the data points to build the forecasting models and another 7 data points from 2011-2017 to test the accuracy of these models. For Bivariate regression and multivariate regression we have used child mortality rate for dependent variable and respective dummy variable for any structural instability.

Sequence Chart:



We can observe a very linear growth of LifeExpectancy over the year with some instability round 1996 which will be examined further.

Extrapolative methods:

1. Naïve Forecasting

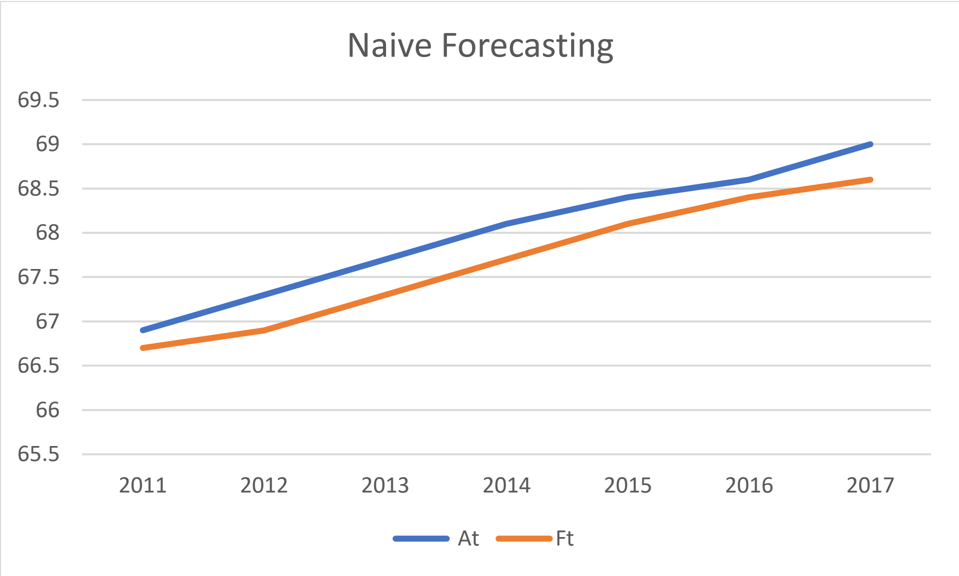
Using excel, we have performed the naïve forecasting for test data points as follows:

Year	LifeExpectancy	Predicted

2011	66.9	66.7
2012	67.3	66.9
2013	67.7	67.3
2014	68.1	67.7
2015	68.4	68.1
2016	68.6	68.4
2017	69	68.6

We have calculated various accuracy parameters for the same as follows:

MAD	0.328571
MSE	0.115714
RMSE	0.340168
MAPE (%)	0.483053



Note: At = Actual Value and Ft = Forecasted Value

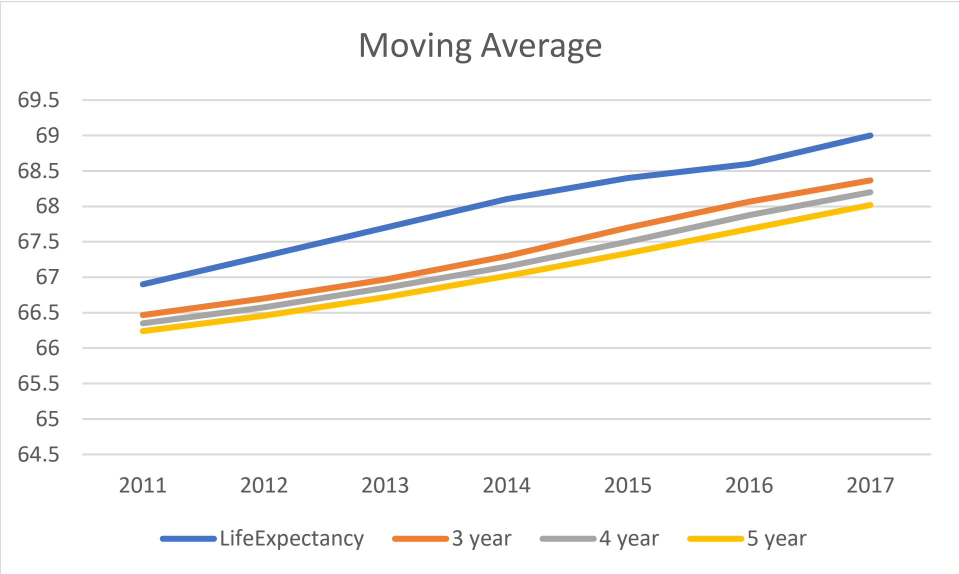
2. Moving Average

We have used 3 MA models (3 year, 4 year, and 5 year) to forecasting these data points as follows:

Year	LifeExpectancy	3 year	4 year	5 year
2011	66.9	66.5	66.4	66.2
2012	67.3	66.7	66.6	66.5
2013	67.7	67.0	66.9	66.7
2014	68.1	67.3	67.2	67.0
2015	68.4	67.7	67.5	67.3
2016	68.6	68.1	67.9	67.7
2017	69	68.4	68.2	68.0

And the respective accuracy parameters are as follows:

	3 year	4 year	5 year
MAD	0.6333	0.7857	0.9314
MSE	0.4144	0.6327	0.8855
RMSE	0.6438	0.7954	0.9410
MAPE (%)	0.9308	1.1546	1.3685



We can observe from graph as well as MAPE 3 year Moving Average have highest accurate results followed by 4 year and 5 year.

3. Exponential Smoothing

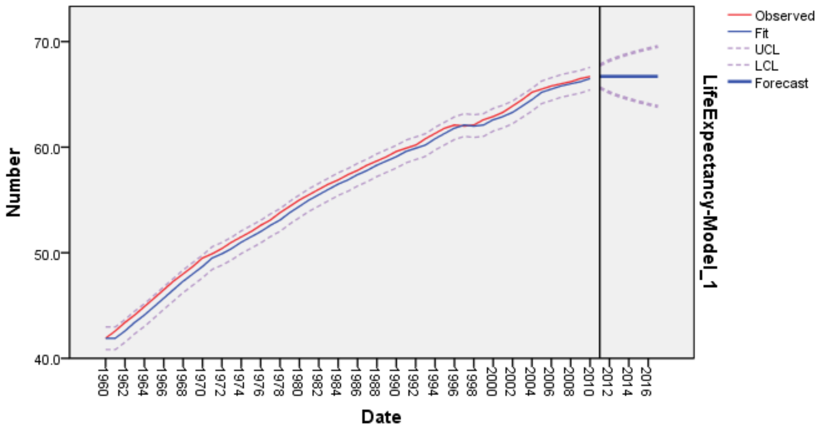
We have used SPSS to do Exponential Smoothing (using Holt's Linear Trend, Damped Trend and Simple Exponential Smoothing Models). Since we don't have any seasonality in our data due to annual form of data, we have not used winter method and other methods.

Forecasted values are as follows:

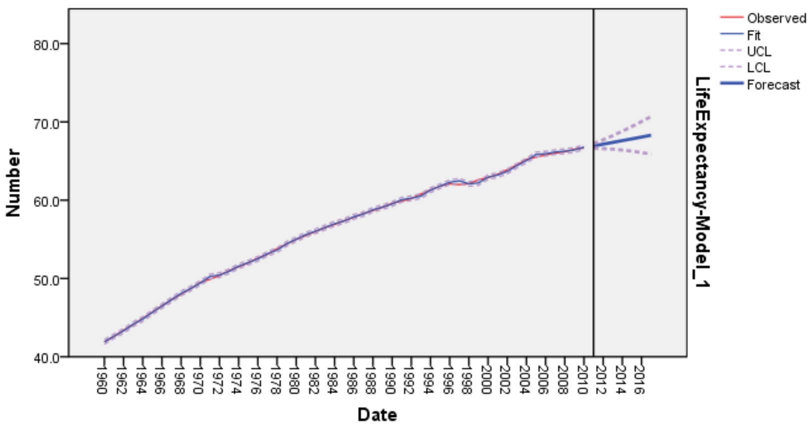
	Actual	Forecasted		
Year	LifeExpectancy	Simple	Holt's Linear Trend	Damped Trend
2011	66.9	66.7	66.9	66.9
2012	67.3	66.7	67.2	67.2
2013	67.7	66.7	67.4	67.4
2014	68.1	66.7	67.6	67.6
2015	68.4	66.7	67.8	67.9
2016	68.6	66.7	68.1	68.1
2017	69	66.7	68.3	68.3

Respective Forecast Plot are as follow:

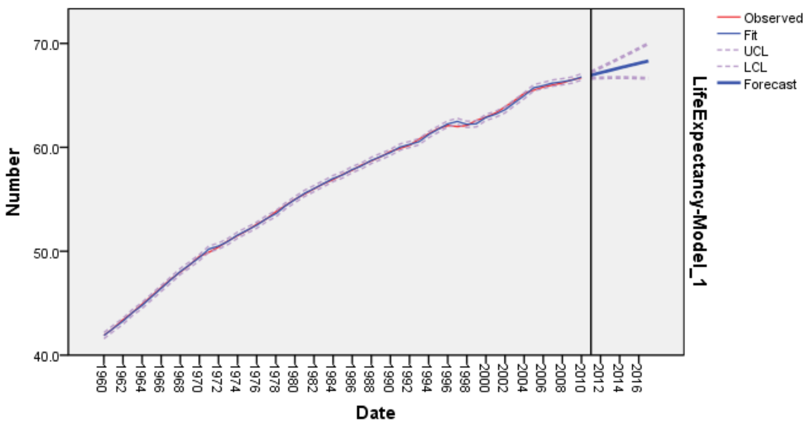
I. Simple Exponential Smoothing



II. Holt's Linear Trend



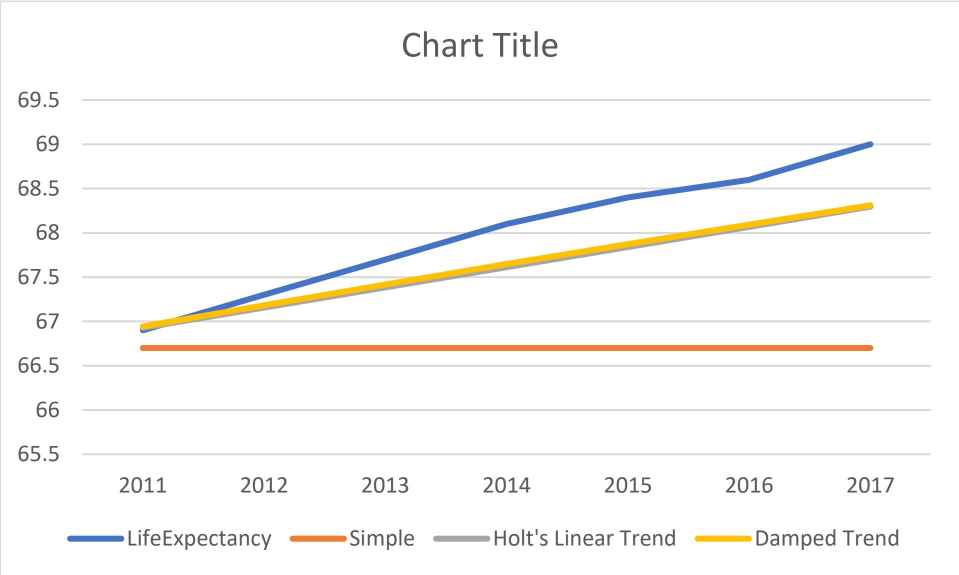
III. Damped Trend



Respective Accuracy parameters are as follows:

	Simple	Holt's Linear Trend	Damped Trend
MAD	1.3000	0.3960	0.3743
MSE	2.1643	0.2074	0.1873
RMSE	1.4712	0.4554	0.4327

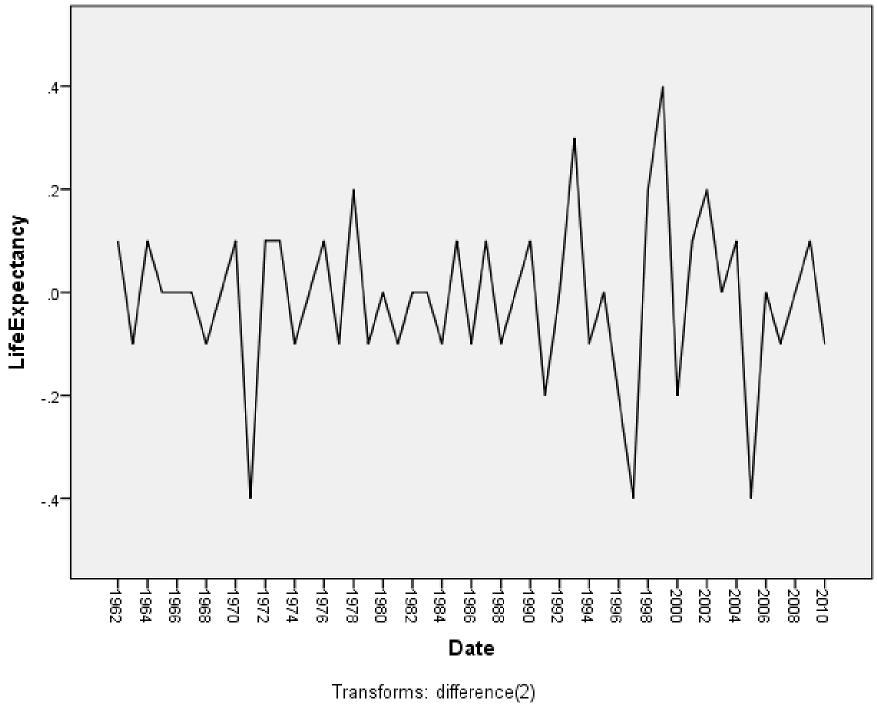
MAPE (%)	1.9017	0.5791	0.5473



From graph as well as MAPE we can observe that, Damped Trend provides highest accurate result followed by Holt's Linear trend then Simple model.

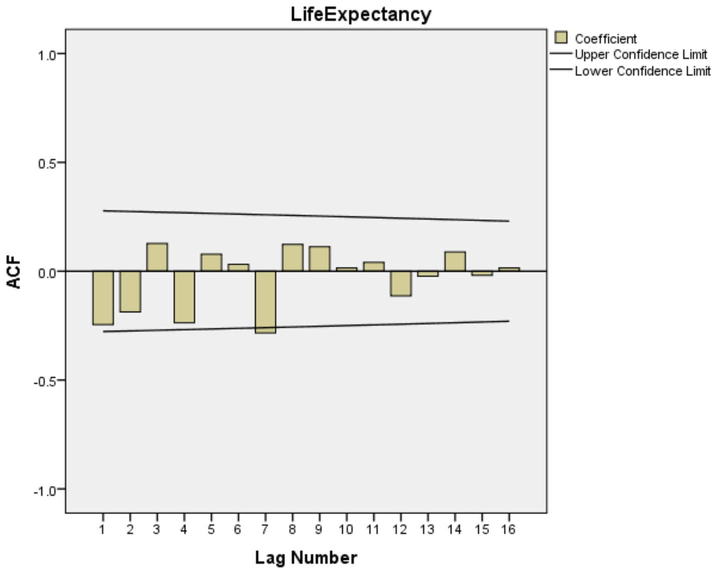
4. ARIMA (autoregressive integrated moving average)

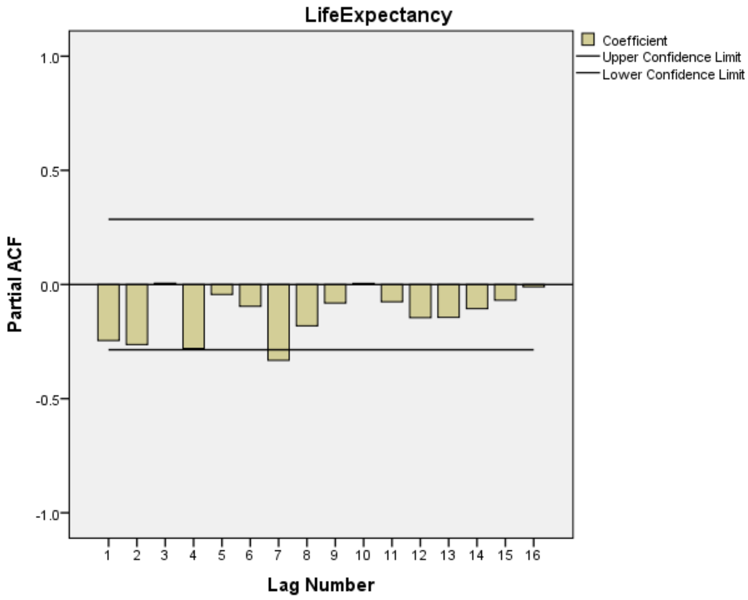
Here first we had to check for stationary, from first diagram it can be observed that the trend line is not stationary. So taking difference = 2, we achieved stationary.



Then we check for ACF and PACF graph to come with required for of ARIMA.

ACF and PACF graph are as follows:

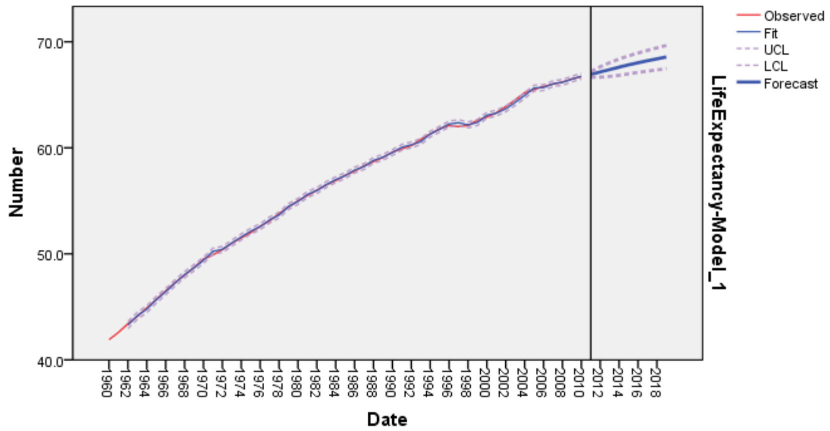




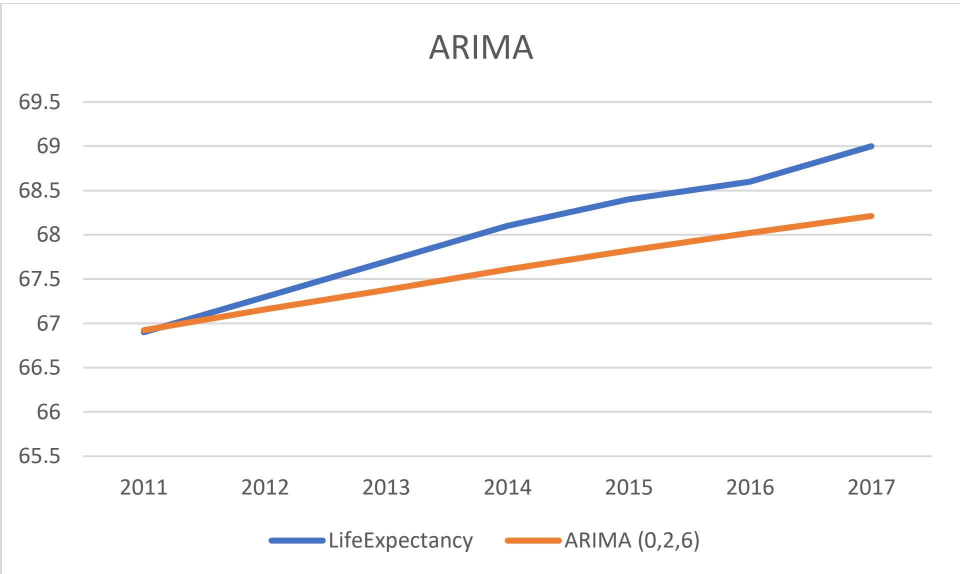
We can observe similar graphs where as ACF reduced to zero more compared to PACF. And from ACF graph we can observe one spike rest are essentially zero. So we decided to go with Moving Average models only.

We also verified the same using Auto ARIMA by R to determine efficient model with respect to Lower normalised BIC value.

After multiple iteration, we reached ARIMA (0,2,6) to have highest accuracy. Results are as follows:



Year	LifeExpectancy	ARIMA (0,2,6)
2011	66.9	66.9
2012	67.3	67.2
2013	67.7	67.4
2014	68.1	67.6
2015	68.4	67.8
2016	68.6	68.0
2017	69	68.2



Accuracy parameter are as follows:

MAD	0.4169
MSE	0.2357
RMSE	0.4855
MAPE	0.6094

Explanatory Methods

We are using explanatory methods to generate single equation models to forecast LifeExpectancy.

I. Bivariate Regression

We have observed from earlier exercises Child Mortality rate to be the strongest independent variable to explain variability across life expectancy value.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	74.516	7.060		10.554	.000
	BabiesPerWomen	3.896	2.636	.573	1.478	.146
	CO2perPerson	-4.944	3.582	-.219	-1.380	.174
	ChildMortalityper1000	-.218	.041	-1.733	-5.348	.000
	IncomePerPerson$	.000	.001	.045	.413	.682
	GDPgrowth	-.020	.043	-.009	-.478	.635
2	(Constant)	74.339	6.984		10.645	.000
	BabiesPerWomen	3.790	2.600	.557	1.458	.152
	CO2perPerson	-3.747	2.088	-.166	-1.795	.079
	ChildMortalityper1000	-.215	.040	-1.707	-5.422	.000
	GDPgrowth	-.016	.041	-.007	-.382	.704
3	(Constant)	74.063	6.882		10.761	.000
	BabiesPerWomen	3.862	2.569	.568	1.503	.139
	CO2perPerson	-3.736	2.068	-.165	-1.806	.077
	ChildMortalityper1000	-.216	.039	-1.714	-5.503	.000
4	(Constant)	84.039	1.848		45.469	.000
	CO2perPerson	-6.229	1.252	-.276	-4.975	.000
	ChildMortalityper1000	-.158	.007	-1.253	-22.599	.000

a. Dependent Variable: LifeExpectancy

We did bivariate regression with LifeExpectancy as dependent variable and ChildMortalityper1000 as independent variable.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	74.999	.411		182.654	.000	74.173	75.824
	ChildMortalityper1000	-.125	.003	-.990	-48.721	.000	-.130	-.119

a. Dependent Variable: LifeExpectancy

Forecast Single Equation:

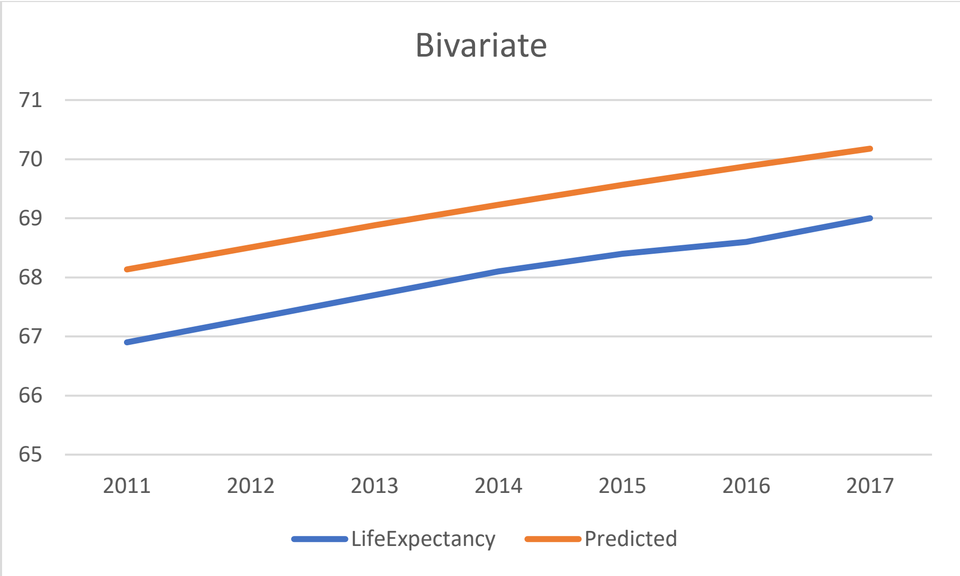
LifeExpectancy = 74.999 + (-.125)*ChildMortalityper1000

Forecasted values are as follows:

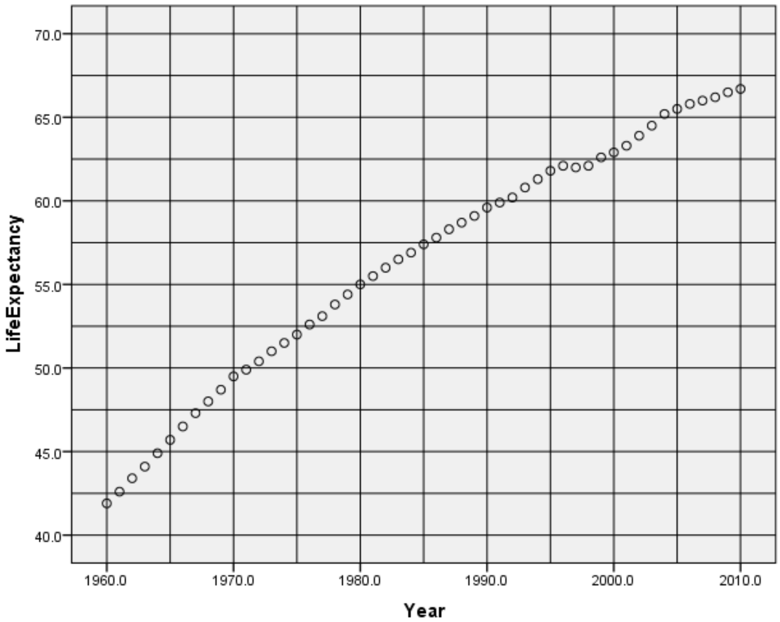
Year	LifeExpectancy	Predicted
2011	66.9	68.13
2012	67.3	68.51
2013	67.7	68.88
2014	68.1	69.23
2015	68.4	69.57
2016	68.6	69.88
2017	69	70.18

Accuracy parameters as follows:

MAD	1.1965
MSE	1.4336
RMSE	1.1973
MAPE (%)	1.7598



2. Multivariate Regression with dummy for structural instability



We can observe a kink in the graph at 1996, which might be due to economic restructuring during 1990-95 in India. So we took a dummy variable (D) with 0 as years before 1996 and 1 for years starting from 1996.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	77.125	.655		117.820	.000	75.808	78.441
	ChildMortalityper1000	-.135	.004	-1.075	-37.988	.000	-.143	-.128
	D	-1.740	.447	-.110	-3.896	.000	-2.638	-.842

a. Dependent Variable: LifeExpectancy

Forecast Single Equation:

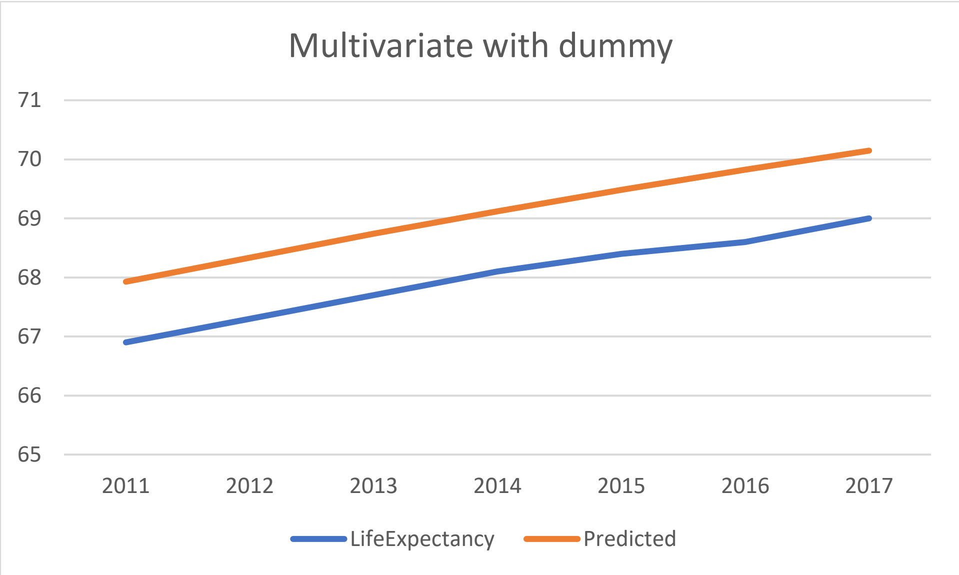
LifeExpectancy = 77.125 + (-.135)*ChildMortalityper100 + (-1.740)*D

Forecasted values are as follows:

Year	LifeExpectancy	Predicted
2011	66.9	67.92727
2012	67.3	68.33329
2013	67.7	68.73932
2014	68.1	69.11828
2015	68.4	69.4837
2016	68.6	69.82206
2017	69	70.14688

Accuracy parameters are as follows:

MAD	1.081542
MSE	1.174743
RMSE	1.083856
MAPE (%)	1.589893



3. Multivariate Regression with Slope and intercept dummy

Coefficients[a]

Model	Unstandardized Coefficients	Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B

	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	77.330	.663		116.637	.000	75.996	78.663
ChildMortalityper1000	-.136	.004	-1.084	-37.774	.000	-.144	-.129
D	-3.765	1.477	-.239	-2.549	.014	-6.738	-.793
D_ChildMortality	.023	.016	.124	1.437	.157	-.009	.056

a. Dependent Variable: LifeExpectancy

Forecast Single Equation:

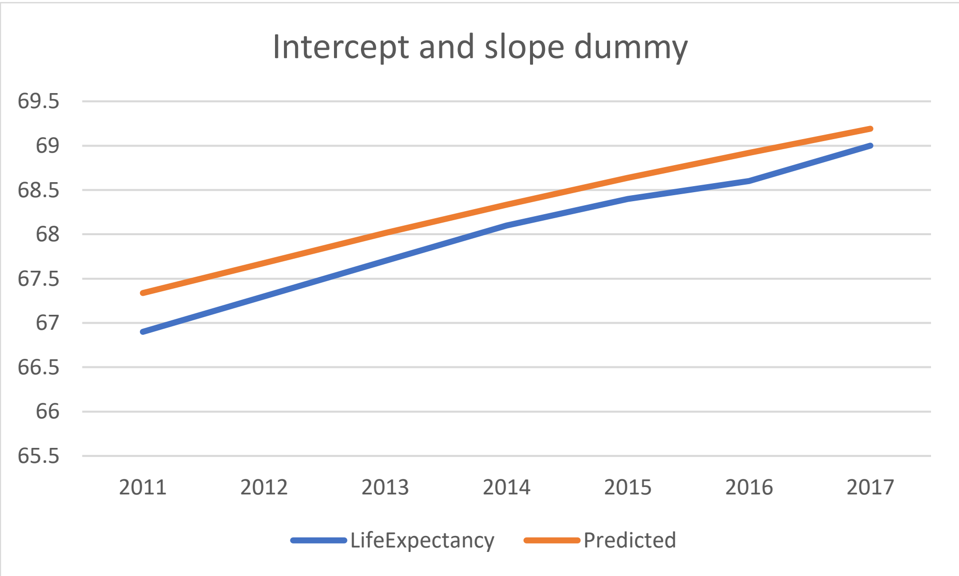
LifeExpectancy = 77.330 - 3.765* D - .136* ChildMortalityper1000 + .023* D_ChildMortalityper1000

Forecasted Value are as follows:

Year	LifeExpectancy	Predicted
2011	66.9	67.33745
2012	67.3	67.67647
2013	67.7	68.01549
2014	68.1	68.33191
2015	68.4	68.63702
2016	68.6	68.91954
2017	69	69.19075

Accuracy parameter are as follows:

MAD	0.301233
MSE	0.097297
RMSE	0.311925
MAPE (%)	0.444087



Comparison Models used for forecasting:

We compared respective MAPE for various models and give ranking to the models accordingly.

Model	**MAPE (%)**	**Rank**
Extrapolative methods		
1. Naive	0.4831	2
2. Moving Average		
I. 3 year	0.9308	6
ii. 4 year	1.1546	7
iii. 5 year	1.3685	8
3. Exponential Smoothing		
I. Simple	1.9017	
ii. Holt's Linear Trend	0.5791	4
iii. Damped Trend	0.5473	3
4. ARIMA (0,2,6)	0.6094	5
Explanatory Methods		
1. Bivariate Regression	1.7598	10
2. Multivariate regression with dummy	1.5899	9
3. Multivariate regression with slope and intercept dummy	0.4441	1

Conclusion

- In general Extrapolative methods provide accurate result compared to Explanatory Methods.
- But, Multivariate regression with slope and intercept dummy provided highest accurate result followed by Naïve method.
- Every Method used here are efficient to be used in forecasting as MAPE % is lower than 2%. Relative accuracy can be observed from rank column for these models used.

*Note: Calculations in excel and Results of SPSS are attached with the same name.