TIME SERIES ANALYSIS OF DEATH RATES FROM TOBACCO SMOKING

NAME: PRITAM SAHA

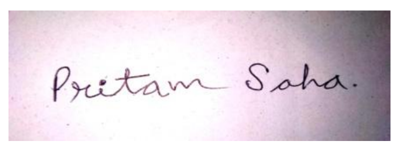
SEM: 6

ROLL: **422**

CIN: 17-300-4-07-0422

SUPERVISOR: DR. SURABHI DASGUPTA

I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.



Signature

TABLE OF CONTENTS

INTRODUCTION

MOTIVATION BEHIND THE PROJECT:-

Smoking of tobacco is one of the India's largest health problems.Over the course of 20^{th} century, nearly hundred million people died prematurely as a result of smoking around the globe,mostly in rich countries.The burden of health due to smoking tobacco is now moving from high to low to middle income countries such as India.Estimates have suggested that nearly 1 billion people could die from smoking tobacco over the course of 21^{st} century around the globe.

Smoking is a major risk factor for India's leading causes of death which includes lung and other forms of cancer,respiratory disease and heart disease. More people have died from secondhand smoke than road accidents globally in 2017.One out of five adults in the world smoke tobacco.Men are more probable to smoke as compared to women.So,as to promote awareness amongst people so that they can quit smoking,I am doing this project.I am doing this project as I cannot watch my countrymen dying from diseases caused as a result of smoking.

This will facilitate the government,policy makers to **launch new policies** getting the successful experience of the earlier policies.This will also help the government to impose more taxes on cigarettes,promote bans on tobacco advertising and to support our fellow countrymen so that they can quit smoking so that death rates from tobacco smoking falls significantly.Moreover,this project will also help other researchers to get an in depth knowledge about the topic so that they can further continue research into this field.This is the motivation behind my work.

Throughout this project,I have used a phrase as deaths from tobacco smoking.By this phrase,I mean to say that smoking is the major factor which degrades the health of persons as a result of which people have died.

OBJECTIVE OF THE PROJECT:-

The objective of my work is to predict death rates per 1 lakh persons from smoking(direct and secondhand) with time.

Secondly,to predict death rates per 1 lakh persons from tobacco smoking with time for different age groups.

Thirdly,to observe different measures taken by the government so as to reduce smoking.

DATA AND TERMINOLOGIES

This project is based on yearly data collected from IHME for 28 years from 1990-2017 of India on various parameters like death rates per 1 lakh persons from smoking,etc.

The data used in this project are **secondary** in nature.

A)Data on total annual no. of deaths in World due to various reasons from 1990-2017:-

Here data on total annual no. of deaths in the world is given with respect to various risk factors from 1990-2017.This data is measured across both sexes and all age groups.

Time(in Years)	Unsafe water source (deaths)	Poor sanitation (deaths)	Indoor air pollution (deaths)	Child wasting (deaths)	Secondhand smoke (deaths)
1990	2111659.077	1638021.199	2708904.82	3365308.624	1247230.879
1991	2095066.505	1622958.94	2677805.722	3296648.253	1251918.229
1992	2053362.286	1588275.164	2648544.39	3203517.655	1255581.16
1993	2013224.699	1553936.146	2604092.494	3115719.489	1260815.712
1994	1973406.044	1519967.074	2563321.579	3039123.021	1260994.206
1995	1931065.508	1484244.305	2521547.506	3031073.947	1253577.032
1996	1877778.83	1440235.417	2477283.308	2924737.715	1241763.527
1997	1840394.52	1409441.51	2456653.199	2823961.948	1232722.03
1998	1808119.007	1382608.295	2419886.895	2727432.701	1218980.197
1999	1766645.451	1347957.663	2357044.544	2625624.006	1204890.482
2000	1727905.469	1315192.214	2314854.156	2528264.401	1195523.541
2001	1682357.015	1277426.292	2262796.548	2415167.658	1183889.971
2002	1638128.167	1240534.972	2218646.043	2304567.114	1178246.291
2003	1596385.848	1205163.982	2172350.595	2118224.642	1171247.834
2004	1561987.644	1174027.895	2109307.172	2020585.348	1152014.441
2005	1536826.824	1149262.178	2057468.406	1929828.236	1141569.517
2006	1523674.684	1132528.956	2002416.138	1850101.089	1117579.421
2007	1503844.158	1109838.767	1961085.424	1769622.401	1107576.361
2008	1491461.978	1089920.305	1927176.022	1698621.025	1110817.943
2009	1489789.499	1077637.575	1897357.343	1640429.923	1115495.907
2010	1483731.657	1061903.157	1869568.125	1594289.594	1126799.656
2011	1431184.379	1012279.945	1832880.494	1512021.885	1134562.113
2012	1374545.908	957813.2057	1824696.213	1424618.771	1141945.31

Time(in Years)	Unsafe water source (deaths)	Poor sanitation (deaths)	Indoor air pollution (deaths)	Child wasting (deaths)	Secondhand smoke (deaths)
2013	1322082.215	902564.8107	1771869.513	1342889.805	1144785.279
2014	1282539.799	854076.1423	1744033.091	1261561.866	1164110.806
2015	1245322.727	810246.0167	1705654.026	1193359.088	1191774.354
2016	1220145.311	779071.6073	1696332.02	1127761.579	1209977.188
2017	1232368.284	774240.9861	1640599.784	1077101.961	1220044.162

Time(in Years)	Alcohol use (deaths)	Drug use (deaths)	High blood sugar (deaths)	Obesity (deaths)
1990	1800478.308	250864.6759	3436970.409	2159329.452
1991	1864042.538	267682.9469	3508798.395	2222986.643
1992	1939783.304	287034.8138	3597104.065	2303362.649
1993	2025045.603	306575.9508	3721781.68	2428019.231
1994	2097212.994	325420.8238	3815828.882	2524269.751
1995	2144887.033	342556.7162	3886724.12	2586034.531
1996	2168153.976	353115.8404	3948468.265	2632155.462
1997	2199073.322	361142.9969	4033219.042	2688066.122
1998	2236838.89	372086.7041	4117958.719	2754377.071
1999	2292873.113	386344.769	4224079.759	2860457.523
2000	2349425.559	399074.8586	4324898.2	2941771.575
2001	2385905.324	405362.8774	4438142.543	3015683.059
2002	2443612.781	414358.0392	4611053.543	3121412.686
2003	2494699.766	423709.2334	4787898.524	3217542.881
2004	2518388.904	433106.5906	4880129.558	3262967.782
2005	2562559.799	447740.285	5016361.945	3357385.922
2006	2544467.646	452485.8341	5056877.019	3388547.812
2007	2562325.476	458612.3074	5135848.389	3465994.585
2008	2621392.287	467883.4028	5258887.049	3586561.415
2009	2628366.75	474179.7575	5361133.733	3688360.783
2010	2677117.845	481085.006	5480476.47	3801353.307
2011	2652024.866	488733.8298	5597446.696	3895077.956
2012	2667932.256	496855.4929	5727642.572	4010870.864
2013	2673153.393	508271.6435	5842881.612	4122520.196
2014	2706466.692	526924.9122	6002241.778	4272118.459
2015	2773307.329	549156.8439	6218325.333	4462996.032
2016	2817609.639	572922.967	6399256.514	4614666.262
2017	2842854.196	585348.1802	6526028.193	4724346.293

Time(in Years)	High blood pressure (deaths)	Smoking (deaths)	Air pollution (outdoor & indoor)(deaths)
1990	6884458.978	5681577.793	4693071.312
1991	6988580.174	5769043.287	4720094.144
1992	7115388.804	5869243.876	4745836.577
1993	7329202.966	6023154.99	4768144.772
1994	7469316.4	6115916.982	4773589.913
1995	7537548.158	6170797.452	4755628.878
1996	7568246.535	6180919.478	4749140.8
1997	7626124.418	6208120.345	4766895.258
1998	7688173.701	6244660.28	4760687.913
1999	7817683.344	6298194.798	4726253.31
2000	7932073.943	6351837.677	4709245.489
2001	8029451.543	6382733.902	4702796.73
2002	8188714.686	6444008.309	4716666.14
2003	8325151.221	6500340.707	4725277.204
2004	8321914.696	6450954.751	4673825.644
2005	8440626.756	6457486.751	4650017.131
2006	8416624.125	6375847.554	4615828.129
2007	8503613.801	6383033.915	4626412.075
2008	8699254.174	6447767.153	4671511.435
2009	8863423.503	6495294.182	4713605.404
2010	9081676.32	6569515.346	4761102.293
2011	9239375.173	6615962.057	4804868.946
2012	9410997.059	6663081.086	4770981.619
2013	9535296.958	6673225.404	4767978.051
2014	9756852.926	6764715.825	4775123.184
2015	10070692.4	6936294.573	4875230.894
2016	10293777.65	7046703.251	4863821.567
2017	10440818.48	7099111.294	4895475.998

B)<u>Data on total annual no. of deaths in India due to various reasons from 1990-2017:-</u>

Here data on total annual no. of deaths in India is given with respect to various risk factors from 1990-2017.This data is measured across both sexes and all age groups.

Time(in Years)	Unsafe water source (deaths)	Poor sanitation (deaths)	Indoor air pollution (deaths)	Child wasting (deaths)	Secondhand smoke (deaths)
1990	807723.2259	636517.8157	691699.0043	861136.0094	189345.941
1991	811265.9069	638456.6346	693852.4155	847509.4954	193888.807
1992	809832.0949	636020.6287	693823.3194	832721.1036	198873.2296
1993	798193.5484	625142.5128	680356.2421	810052.0576	198535.6007
1994	782064.1082	611345.708	670049.3665	791943.3607	199021.2449
1995	770997.9855	601637.2419	659136.639	774324.8941	199278.0944
1996	757174.1992	590231.3711	653262.9633	748657.0839	201058.4099
1997	756259.4763	589696.9919	671861.3869	721795.7415	208476.4526
1998	750757.1052	585826.0562	668996.9391	691683.7054	209418.0484
1999	741045.3527	577967.4818	642748.1845	661414.8974	204088.1613
2000	734455.8641	572125.8089	629573.6696	639991.2898	203119.5113
2001	718835.9458	559478.171	619160.9953	606418.2663	201657.8166
2002	700750.5468	544821.641	604227.0178	572178.5584	196926.6555
2003	683466.4598	529869.4559	582220.5003	536602.1188	190095.6575
2004	668881.3042	515958.8538	552103.8981	506246.0389	180280.3493
2005	660257.8942	505825.6925	545624.6479	479780.2725	178291.8333
2006	654674.2677	497423.077	546837.6094	454782.7761	179203.3433
2007	650617.7678	489606.9735	549219.8389	433088.7519	179616.9358
2008	648163.4067	481544.8127	548852.731	413313.9589	180241.3537
2009	651194.0599	476656.9839	547817.3069	393356.2292	181238.9327
2010	656580.0425	473093.9055	550363.8975	376323.4114	185621.6379
2011	645210.7307	457656.6462	551432.3215	356571.0368	191643.6048
2012	625665.6091	434398.4699	563339.6875	339512.2694	196719.5125
2013	604731.7691	405370.4932	542067.3894	313727.0655	201799.6806
2014	588806.9726	377367.6562	525296.8903	278530.0649	207183.405
2015	573767.0403	352445.5677	511055.8769	254844.6087	213749.7075
2016	560745.6633	332744.6846	500517.1481	227107.4988	220874.508
2017	569679.1687	328719.9721	481737.838	207270.626	225812.8421

Time(in Years)	Alcohol use (deaths)	Drug use (deaths)	High blood sugar (deaths)	Obesity (deaths)
1990	292070.0358	17754.50585	363235.9547	91428.55886
1991	304697.5352	18603.01624	378733.637	95150.8897
1992	315478.5302	19473.64209	396105.2574	100634.2987
1993	325331.1102	20093.12451	404204.2891	103813.7454
1994	336492.8354	21102.17313	414214.7379	108392.9611
1995	343967.3817	22350.91212	423015.2223	113125.4519
1996	357996.7502	23896.57401	443759.5986	122574.2745
1997	387325.1328	26507.92501	485639.0388	139406.1912
1998	401047.4965	28833.3329	506349.6818	151680.5014
1999	406780.9301	31394.26873	511206.9373	160323.108
2000	414822.5469	33806.23936	532994.6972	170359.5009
2001	420310.4604	36037.65874	558776.98	180547.2216
2002	422241.1291	38061.6177	582755.3125	191670.2351
2003	419865.6404	39705.85385	602031.2274	201314.9586
2004	410729.8668	40918.78197	602934.6191	204946.3389
2005	424280.6097	43331.79292	633729.4427	218602.2973
2006	438599.8811	44639.0041	660654.8444	235257.9104
2007	458732.734	45516.98592	681582.2988	253635.6101
2008	481435.6544	46385.43663	702289.2291	272985.5592
2009	499326.6286	47183.7636	718453.2273	292622.8435
2010	513561.9412	47516.42296	745662.196	316329.6758
2011	529265.394	47812.61356	785955.1261	342353.4287
2012	542264.3012	47523.62792	824381.9132	367972.8718
2013	548202.5881	48241.76263	856495.5224	394982.8867
2014	549037.7964	49079.22732	889190.1701	425987.4331
2015	554102.9818	49960.72654	925496.8489	459958.7446
2016	571410.2651	51902.0096	972142.3106	499701.1215
2017	580499.0008	53281.22336	1003224.995	532151.0213

Time(in Years)	High blood pressure (deaths)	Smoking (deaths)	Air pollution (outdoor & indoor)(deaths)
1990	705723.705	650770.8855	1018675.716
1991	725433.7672	668334.501	1031705.573
1992	749610.512	680410.727	1043888.599
1993	754994.8513	686448.7975	1034362.609
1994	765792.6243	694103.4948	1030296.89
1995	773606.5719	694732.7861	1025386.182
1996	798854.3238	711212.5966	1032461.343
1997	857176.0917	753639.8494	1072920.003
1998	876937.6622	758057.4011	1079256.954
1999	870808.3317	733003.1458	1050957.123
2000	889423.5155	736787.6271	1047632.182
2001	914124.8132	741364.9802	1049419.489
2002	931005.9979	734441.5011	1041326.218
2003	939947.7434	721521.8294	1025084.858
2004	923018.4536	694001.9989	992397.2448
2005	958299.338	704923.0376	1000265.586
2006	1006446.461	725157.8233	1025864.615
2007	1057808.897	750616.2341	1053544.246
2008	1111001.093	776942.1913	1079994.953
2009	1161158.401	797954.8122	1102674.686
2010	1216791.476	818069.5945	1132044.371
2011	1273709.847	841311.0006	1164662.899
2012	1325460.109	861356.2188	1175900.521
2013	1367356.26	868037.6879	1193266.266
2014	1410189.561	866247.8497	1199984.922
2015	1457130.632	871785.3241	1217470.883
2016	1514945.058	889962.3527	1233018.222
2017	1544920.203	894371.5186	1240528.7

C)Data on Death rates from smoking(direct and secondhand) per 1 lakh persons in India from 1990-2017:-

Here data on annual no. of deaths in India attributed to tobacco smoking is given per 1 lakh people(i.e.,death rate per 1 lakh persons) from 1990-2017.This data is measured across both sexes and it is age-standardized.

Time(in years)	Death rates from smoking
1990	156.9408092
1991	156.5483017
1992	154.8799562
1993	151.8475104
1994	149.0566922
1995	144.9653522
1996	143.5536735
1997	147.3048448
1998	144.3947784
1999	135.8288612
2000	132.8161362
2001	129.8831663
2002	124.7560472
2003	118.7341926
2004	110.7721547
2005	108.5732692
2006	107.6633678
2007	107.1722505
2008	106.9479988
2009	105.8872626
2010	104.7873801
2011	103.9503852
2012	102.8220895
2013	100.4306155
2014	97.21092971
2015	94.86843007
2016	93.90535518
2017	91.57628926

D) Data on Death rates from smoking(direct and secondhand) per 1 lakh persons in India across various age groups in India from 1990-2017:-

Here data on annual no. of deaths in India attributed to tobacco smoking is given per 1 lakh people (i.e.,death rate per 1 lakh persons) from 1990-2017.This data is measured across both sexes and all age groups.Here we have assumed that death rates of children aged between 0-14 years of age is assumed to be negligible and hence it is not being considered in this study.

Time(in Years)	15-49 years old (deaths per 1 lakh)	50-69 years old (deaths per 1 lakh)	70+ years old (deaths per 1 lakh)
1990	20.05070494	374.5320832	1283.416056
1991	19.85882712	374.2345494	1286.267683
1992	19.52996266	371.460697	1277.476035
1993	19.09748188	365.2955608	1255.492826
1994	18.7463728	359.651158	1235.473953
1995	18.23334384	349.6202835	1204.890678
1996	18.20896797	349.7348207	1187.657961
1997	18.53801648	360.8216406	1222.194178
1998	17.87809911	350.4964251	1209.896129
1999	17.82120779	323.4886746	1139.615024
2000	17.38992336	312.6734745	1125.834448
2001	16.59708623	301.7117278	1111.536694
2002	15.99428515	286.8034797	1070.632114
2003	15.1767201	271.8931752	1017.326474
2004	14.20620638	251.3670288	946.0874652
2005	14.07204222	245.0442592	928.6858147
2006	14.1983584	241.9301627	922.6267509
2007	14.43665485	243.1756948	911.6012713
2008	14.38511569	243.1310545	911.5438021
2009	14.16527878	241.78007	904.6452026
2010	13.90524612	240.7431525	895.6579767
2011	13.7574839	240.5725405	890.0318104
2012	13.39582151	240.084739	881.8904254
2013	12.78304811	234.7234184	865.2924971
2014	12.17601144	226.6438053	842.2033639
2015	11.75433515	221.5791466	823.8045173
2016	11.55095428	218.9877839	820.6510985
2017	11.16055845	212.6358239	804.5366993

E) Data on Age-standardized share of cancer death rates from smoking(direct and secondhand) per 1 lakh persons in India(expressed in percentage) from 1990-2016:-

Here the data denotes age-standardized cancer death rate per 1 lakh persons and the share of the cancer death rate attributed to smoking(direct and secondhand) amongst all other causes of cancer deaths(expressed in percentage) in India from 1990-2016.

Time(in Years)	Share of cancer death rate
1990	10.2
1995	10.1
2000	10.2
2005	10.4
2006	10.3
2010	10.4
2016	10.8

F) Data on share of tobacco retail price which is tax in India(in percentage) from 2007-2014:-

Here the data denotes the share of retail price of tobacco which the government gets as a tax in India(in percentage) from 2007-2014.The taxes which are being considered here are GST,excise tax,import duty,etc.Only the most popular brand of cigarettes is being considered here.Cigarettes from unorganized markets are not being considered here because of lack of data.Share of tobacco retail price which is tax(in %) is a categorical variable divided into 4 categories:-

i)<=25% of tobacco price is tax,ii)26-50% of tobacco price is tax,iii)51-75% of tobacco price is tax,iv)>75% of tobacco price is tax.

Time(in Years)	Share of tobacco retail price which is tax(in %)
2007	26-50
2010	26-50
2012	26-50
2014	51-75

The variables mentioned in the points A)-E) are **metric** variables measured in **ratio** scale whereas the variable used in the point F),i.e.,share of tobacco retail price which is tax(in %) is a **categorical variable** measured in **ordinal** scale.

METHODOLOGY

1) **Collection** of Data from the website

   https://ourworldindata.org/smoking#what-can-we-do-to-reduce-smoking

2)Scrutiny,Summarization and Graphical representation of the data.

3)Analysis of the data by **fitting of mathematical trend curve** to predict for future.

4)Run test for the detection of autocorrelation in the residual data.

5)Fitting of AR(2) model in the residual data and to predict for future.

6)Graphical representation of observed and fitted data.

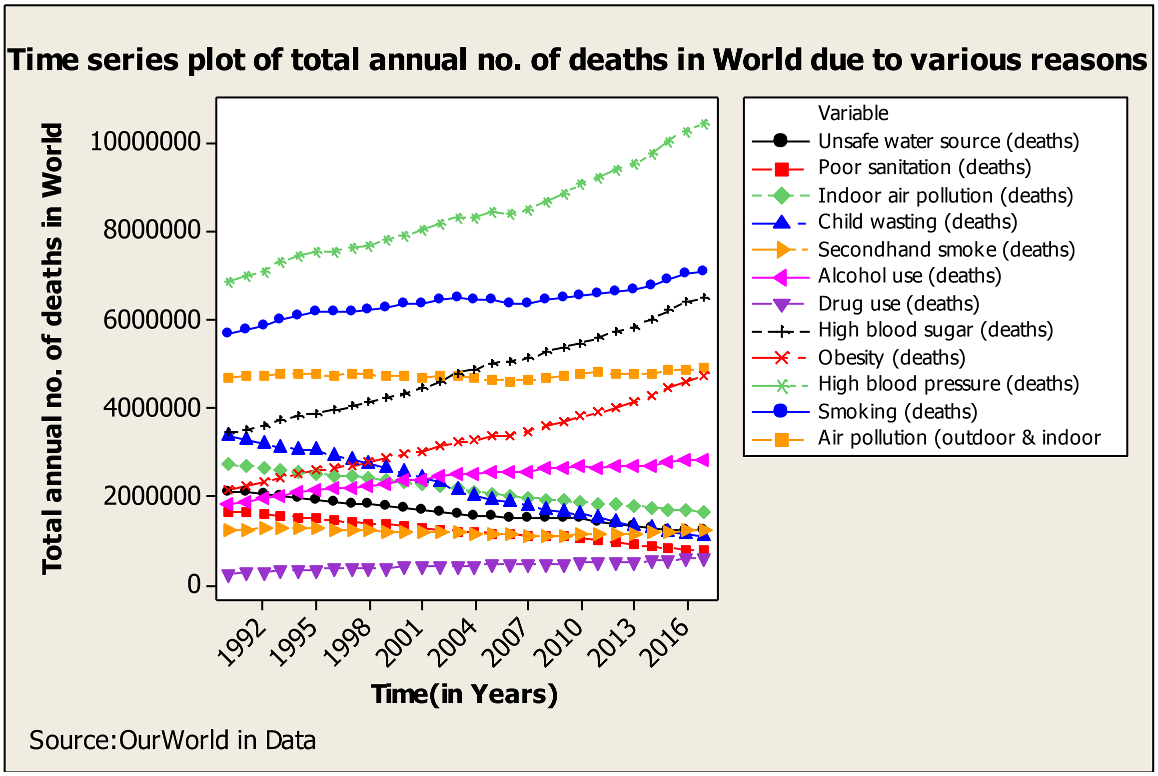
This will facilitate the government,policy makers to **launch new policies** getting the successful experience of the earlier policies.This will also help the government to impose taxes on cigarettes,promote bans on tobacco advertising and to support our fellow countrymen so that they can quit smoking so that death rates from tobacco smoking falls significantly.Moreover,this project will also help other researchers to get an in depth knowledge about the topic so that they can further continue research into this field.

DISCUSSION AND RESULTS

1) SCRUTINY, SUMMARIZATION AND GRAPHICAL REPRESENTATION OF THE DATA:-

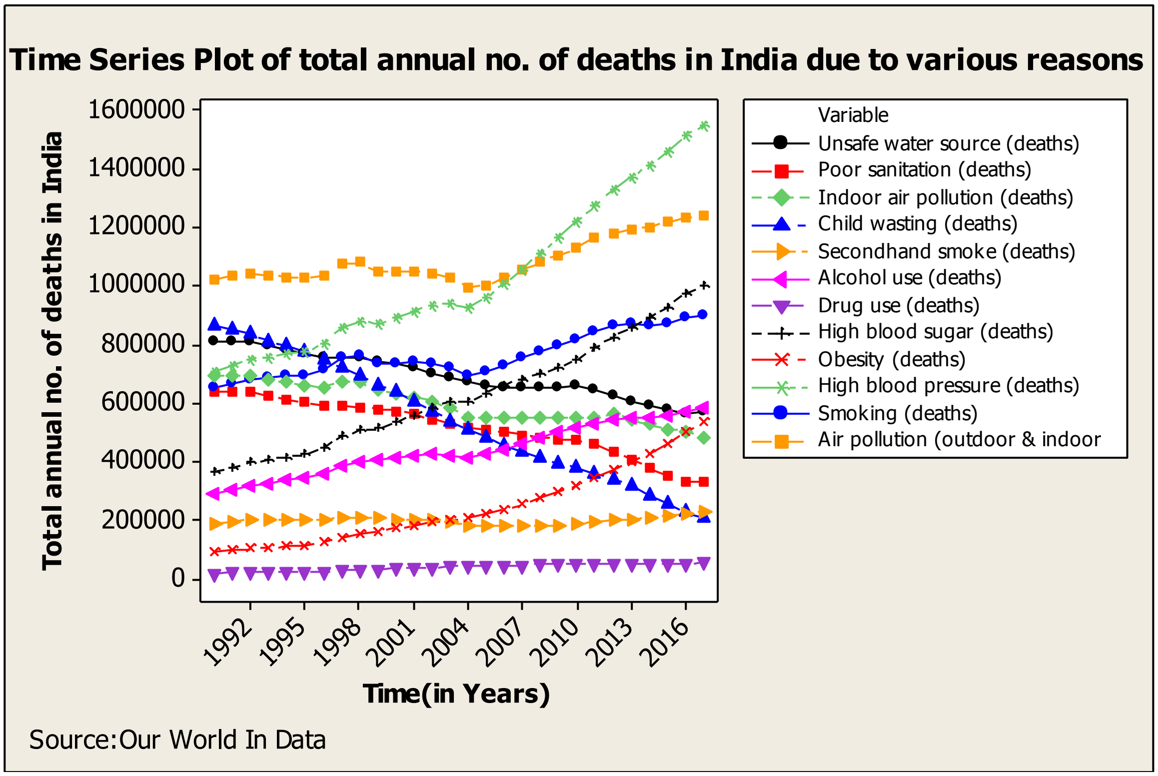
First,we plot the time series data on different parameters and then we note down the observations that follows.

A)Time series plot of total annual no. of deaths in World due to various reasons from 1990-2017:-



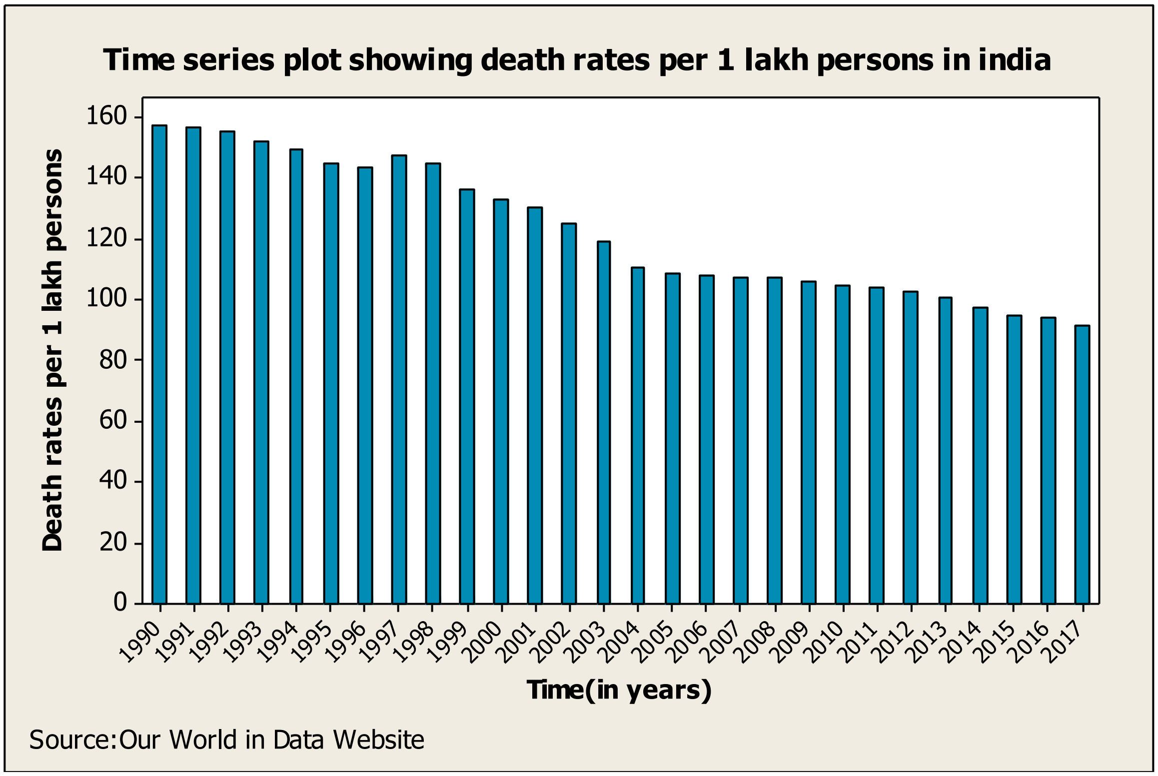
Observations:-From the above plot,it can be seen that total annual no. of deaths from smoking in the world(represented by blue dots)   is increasing with time.Moreover,deaths due to smoking is the 2^{nd} leading risk factors for death in a worldwide basis.

B) <u>Time series plot of total annual no. of deaths in India due to various reasons from 1990-2017:-</u>



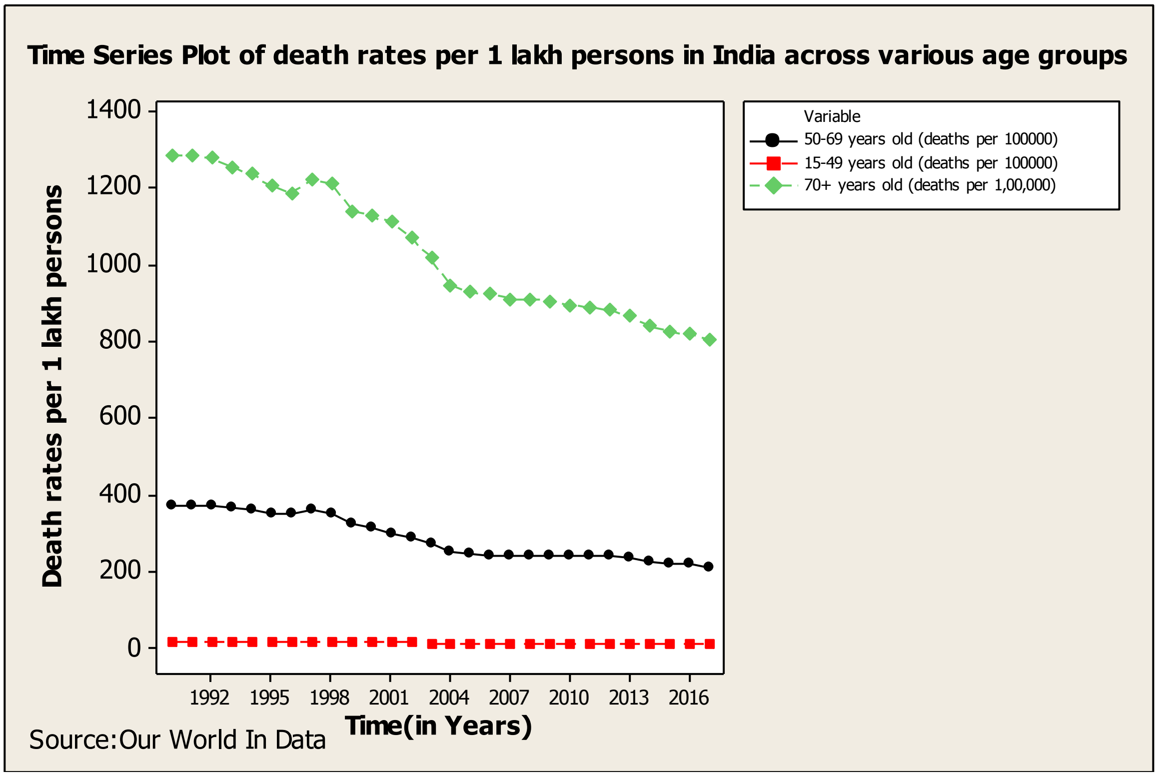
<u>Observations:-</u> From the above plot,it can be seen that total annual no. of deaths from smoking in India(represented by blue dots) is increasing with time. Moreover,deaths due to smoking is one of the leading risk factors for death in India.

C)Time series plot showing death rates from smoking(direct and secondhand) per 1 lakh persons in India from 1990-2017:-



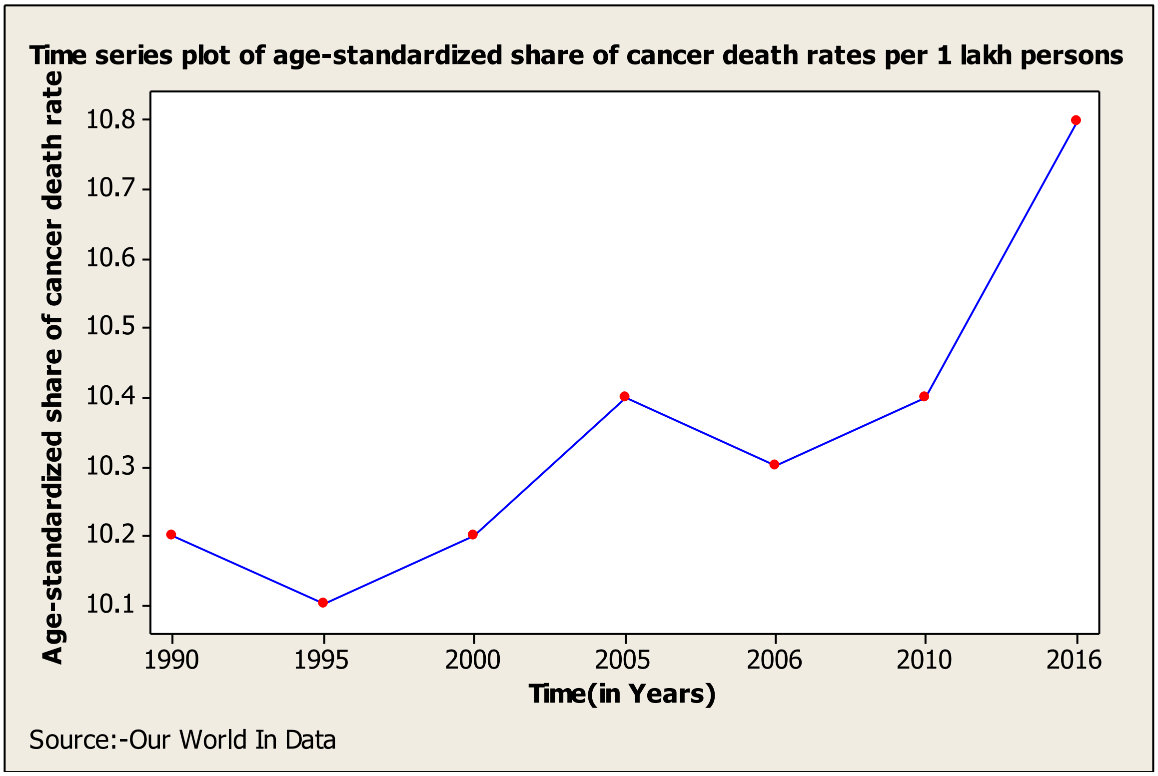
Observations:- From the above plot,it can be seen that death rates from tobacco smoking(direct and secondhand) per 1 lakh persons in India is seen to be decreasing over time.

D) <u>Time series plot of death rates from smoking(direct and secondhand) per 1 lakh persons in India across various age groups in India from 1990-2017:-</u>



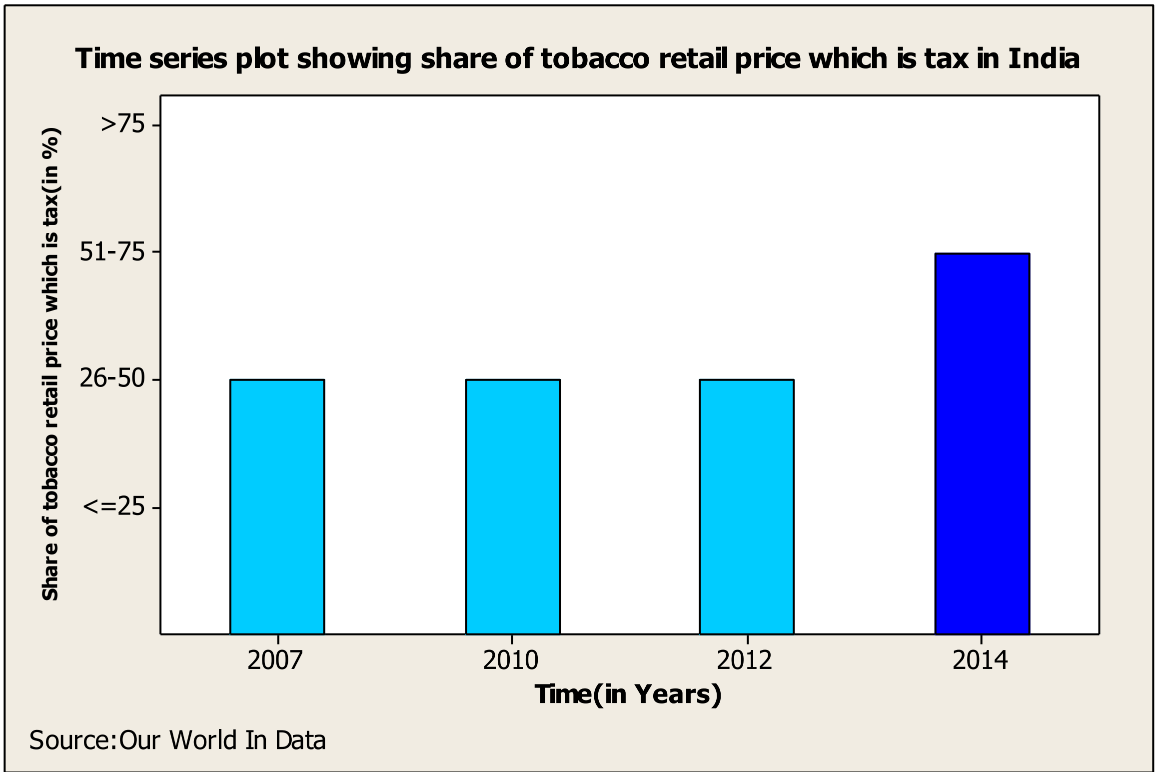
Observations:-From the above plot,it can be seen that death rates per 1 lakh persons in India across various age groups is seen to be decreasing over time.Moreover,death rates for persons with 70+ years of age is rapidly decreasing over time as compared to the persons belonging to other age groups.

E)<u>Time series plot of age-standardized share of cancer death rates(in percentage) from smoking(direct and secondhand) per 1 lakh persons in India from 1990-2016:-</u>



<u>Observations:-</u> From the above plot,it can be seen that age-standardized share of cancer death rates <u>from smoking(direct and secondhand)</u> per 1 lakh persons in India(expressed in percentage) is seen to be increasing over time.This means that the share of cancer deaths from tobacco smoking is increasing over time amongst all cancer types.

F) <u>Time series plot showing share of tobacco retail price which is tax in India(in percentage) from 2007-2014:-</u>



Observations:- From the above plot,it can be seen that the share of tobacco retail price which is tax in India is seen to be increasing over time.This is a good measure taken by the government of India(of increasing tobacco taxes over time) which will help in reducing tobacco consumption because of increased prices of tobacco as a result of which death rate from tobacco smoking will decrease rapidly over time.

2)ANALYSIS OF THE DATA BY FITTING OF MATHEMATICAL TREND CURVE TO PREDICT FOR FUTURE:-

In each of the following cases,we have fitted an appropriate **mathematical trend curve**, and extended it up to the year 2030 **for prediction purposes**. In the subsequent graphs,the fitted trend lines are represented by the **Red color** and the forecasted values are represented by the **Green color**.

A)Let us consider the data on death rates per 1 lakh persons from smoking from year t=1990(1)2017. Here,the data being yearly data,so it doesn't contain seasonality and includes trend and irregular fluctuations only. Thus, by observing the graph,we choose an additive time series model as

$$X_t = T_t + I_t$$

Now we choose a linear trend curve as $T_t = \alpha + \beta t$. So we have

$$X_t = \alpha + \beta t + e_t$$

Where e_t's are random errors with mean 0 and finite variance σ^2.

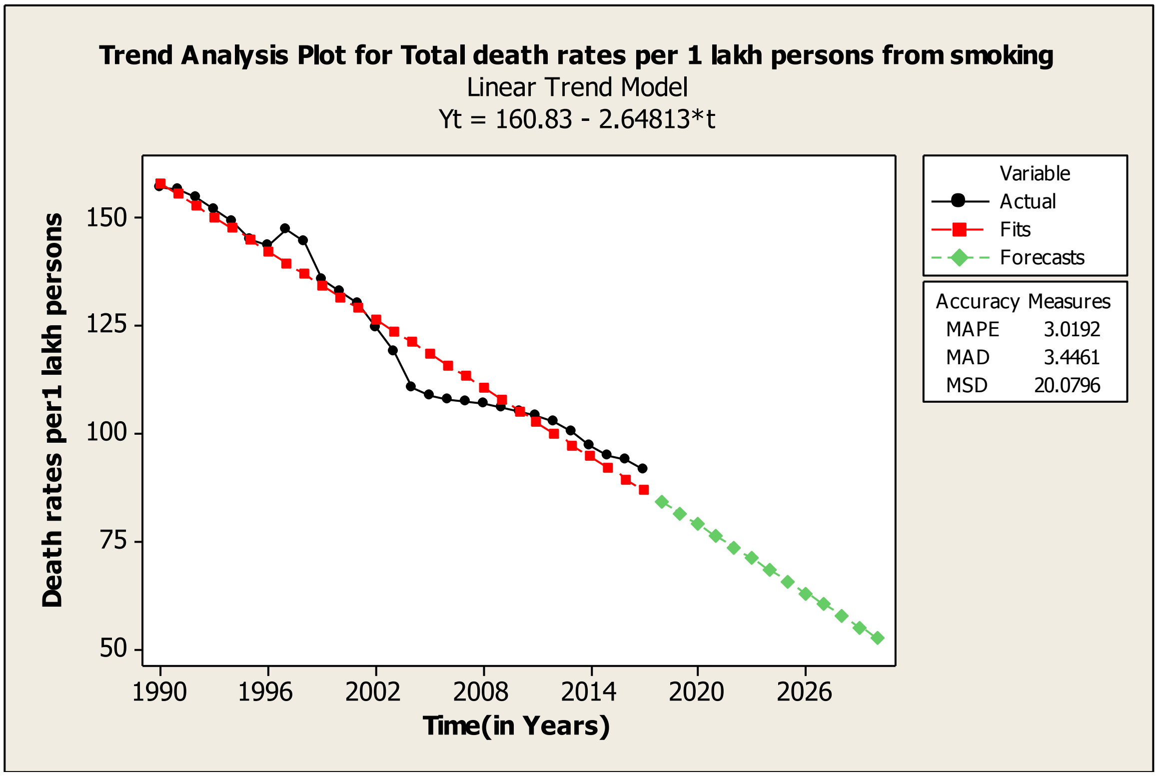
If a & b are the LSE's of α & β,then the fitted trend equation is given by

$$T_t = a + bt$$

Through Minitab, we have obtained the fitted trend equation as

$$T_t = 160.83 - 2.64813t$$

Here,we have checked for the quadratic trend curve but that didn't give a significant decrease in the MAPE value where MAPE denotes Mean Absolute Percent Error.



The fitted trend values are then given in the following table:-

Time(in Years)	Fitted death rates per 1 lakh persons
1990	158.1811758
1991	155.5330414
1992	152.8849069
1993	150.2367725
1994	147.588638
1995	144.9405036
1996	142.2923691
1997	139.6442347
1998	136.9961002
1999	134.3479658
2000	131.6998313
2001	129.0516969
2002	126.4035624
2003	123.7554279
2004	121.1072935
2005	118.459159
2006	115.8110246
2007	113.1628901
2008	110.5147557
2009	107.8666212
2010	105.2184868
2011	102.5703523
2012	99.92221789
2013	97.27408343
2014	94.62594898
2015	91.97781453
2016	89.32968008
2017	86.68154563

The forecasted trend values are then given in the following table:-

Time(in Years)	Forecasted Death rates per 1 lakh persons
2018	84.0334
2019	81.3853
2020	78.7371
2021	76.089
2022	73.4409
2023	70.7927
2024	68.1446
2025	65.4965
2026	62.8483
2027	60.2002
2028	57.5521
2029	54.9039
2030	52.2558

B)i) Let us consider the data on death rates per 1 lakh persons from smoking aged between 15 to 49 years old from year t=1990(1)2017. Here,the data being yearly data,so it doesn't contain seasonality and includes trend and irregular fluctuations only. Thus, by observing the graph,we choose an additive time series model as

$$X_t = T_t + I_t$$

Now we choose a linear trend curve as $T_t = \alpha + \beta t$. So we have

$$X_t = \alpha + \beta t + e_t$$

Where e_t's are random errors with mean 0 and finite variance σ^2.

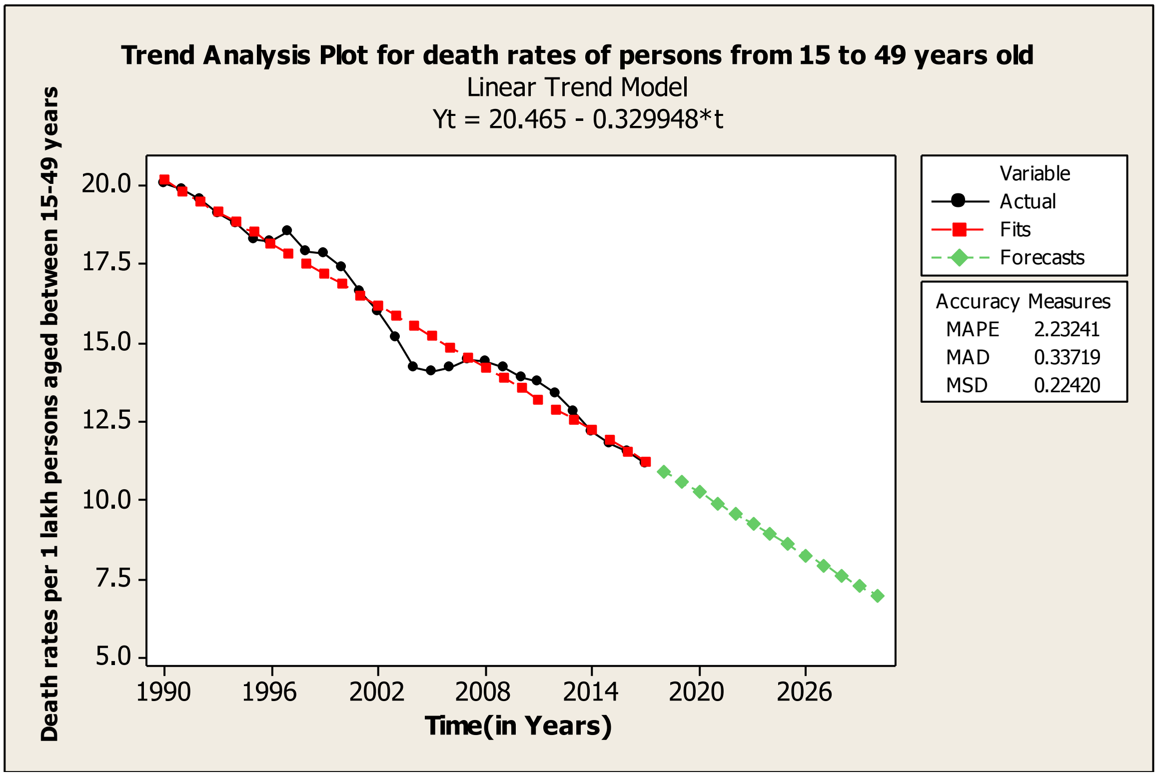
If a & b are the LSE's of α & β,then the fitted trend equation is given by

$$T_t = a + bt$$

Through Minitab, we have obtained the fitted trend equation as

$$T_t = 20.465 - 0.329948t$$

Here,we have checked for the quadratic trend curve but that didn't give a significant decrease in the MAPE value.



The fitted trend values are then given in the following table:-

Time(in Years)	Fitted death rates per 1 lakh persons aged between 15-49 years old
1990	20.13530222
1991	19.80535421
1992	19.47540621
1993	19.1454582
1994	18.81551019
1995	18.48556218
1996	18.15561417
1997	17.82566616
1998	17.49571815
1999	17.16577014
2000	16.83582213
2001	16.50587412
2002	16.17592611
2003	15.8459781
2004	15.51603009
2005	15.18608208
2006	14.85613407
2007	14.52618606
2008	14.19623805
2009	13.86629004
2010	13.53634204
2011	13.20639403
2012	12.87644602
2013	12.54649801
2014	12.21655
2015	11.88660199
2016	11.55665398
2017	11.22670597

The forecasted trend values are then given in the following table:-

Time(in Years)	Forecasted Death rates per 1 lakh persons aged between 15-49 years old
2018	10.8968
2019	10.5668
2020	10.2369
2021	9.9069
2022	9.577
2023	9.247
2024	8.9171
2025	8.5871
2026	8.2572
2027	7.9272
2028	7.5973
2029	7.2673
2030	6.9374

ii)Let us consider the data on death rates per 1 lakh persons from smoking aged between 50 to 69 years old from year t=1990(1)2017. Here,the data being yearly data,so it doesn't contain seasonality and includes trend and irregular fluctuations only. Thus, by observing the graph,we choose an additive time series model as

$$X_t = T_t + I_t$$

Now we choose a linear trend curve as $T_t = \alpha + \beta t$. So we have

$$X_t = \alpha + \beta t + e_t$$

Where e_t's are random errors with mean 0 and finite variance σ^2.

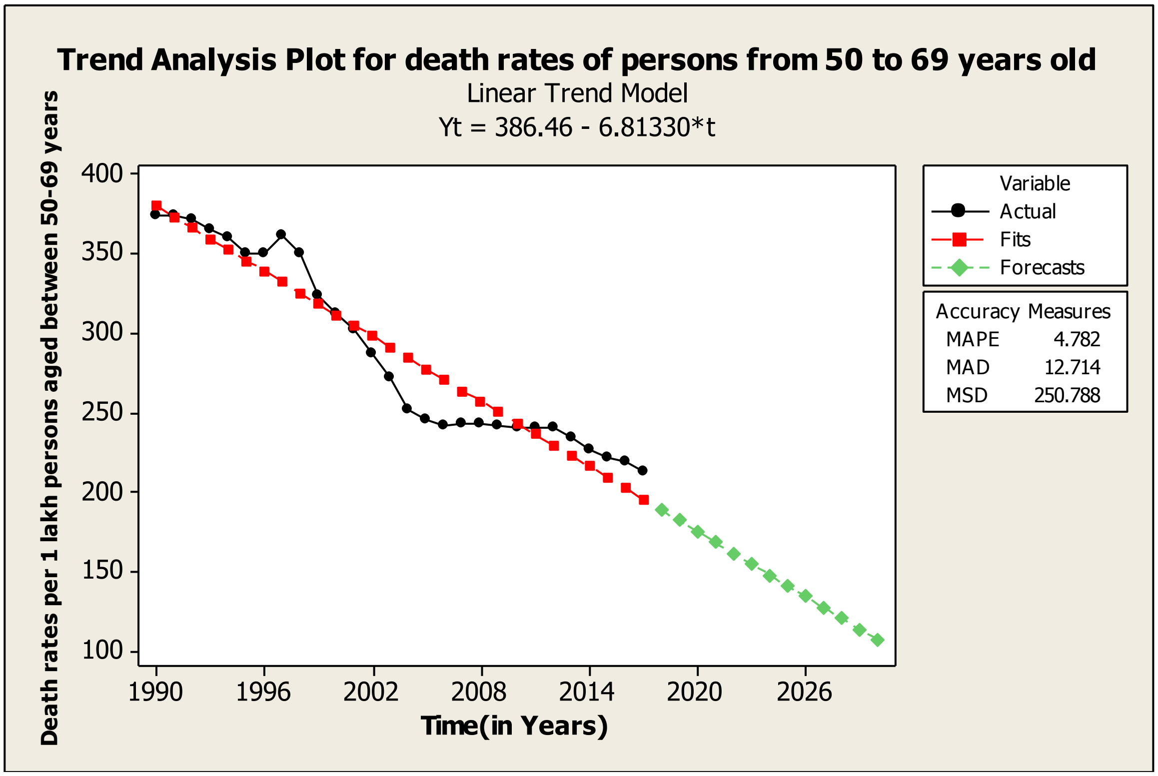
If a & b are the LSE's of α & β,then the fitted trend equation is given by

$$T_t = a + bt$$

Through Minitab, we have obtained the fitted trend equation as

$$T_t = 386.46 - 6.81330t$$

Here,we have checked for the quadratic trend curve but that didn't give a significant decrease in the MAPE value.



The fitted trend values are then given in the following table:-

Time(in Years)	Fitted death rates per 1 lakh persons aged between 50-69 years old
1990	379.651613
1991	372.8383095
1992	366.025006
1993	359.2117024
1994	352.3983989
1995	345.5850954
1996	338.7717918
1997	331.9584883
1998	325.1451848
1999	318.3318812
2000	311.5185777
2001	304.7052742
2002	297.8919707
2003	291.0786671
2004	284.2653636
2005	277.4520601
2006	270.6387565
2007	263.825453
2008	257.0121495
2010	243.3855424
2011	236.5722389
2012	229.7589354
2013	222.9456318
2014	216.1323283
2015	209.3190248
2016	202.5057212
2017	195.6924177

The forecasted trend values are then given in the following table:-

Time(in Years)	Forecasted death rates per 1 lakh persons aged between 50-69 years old
2018	188.879
2019	182.066
2020	175.253
2021	168.439
2022	161.626
2023	154.813
2024	147.999
2025	141.186
2026	134.373
2027	127.559
2028	120.746
2029	113.933
2030	107.119

iii)Let us consider the data on death rates per 1 lakh persons from smoking aged 70+ years from year t=1990(1)2017. Here,the data being yearly data,so it doesn't contain seasonality and includes trend and irregular fluctuations only. Thus, by observing the graph,we choose an additive time series model as

$$X_t = T_t + I_t$$

Now we choose a linear trend curve as $T_t = \alpha + \beta t$. So we have

$$X_t = \alpha + \beta t + e_t$$

Where e_t's are random errors with mean 0 and finite variance σ^2.

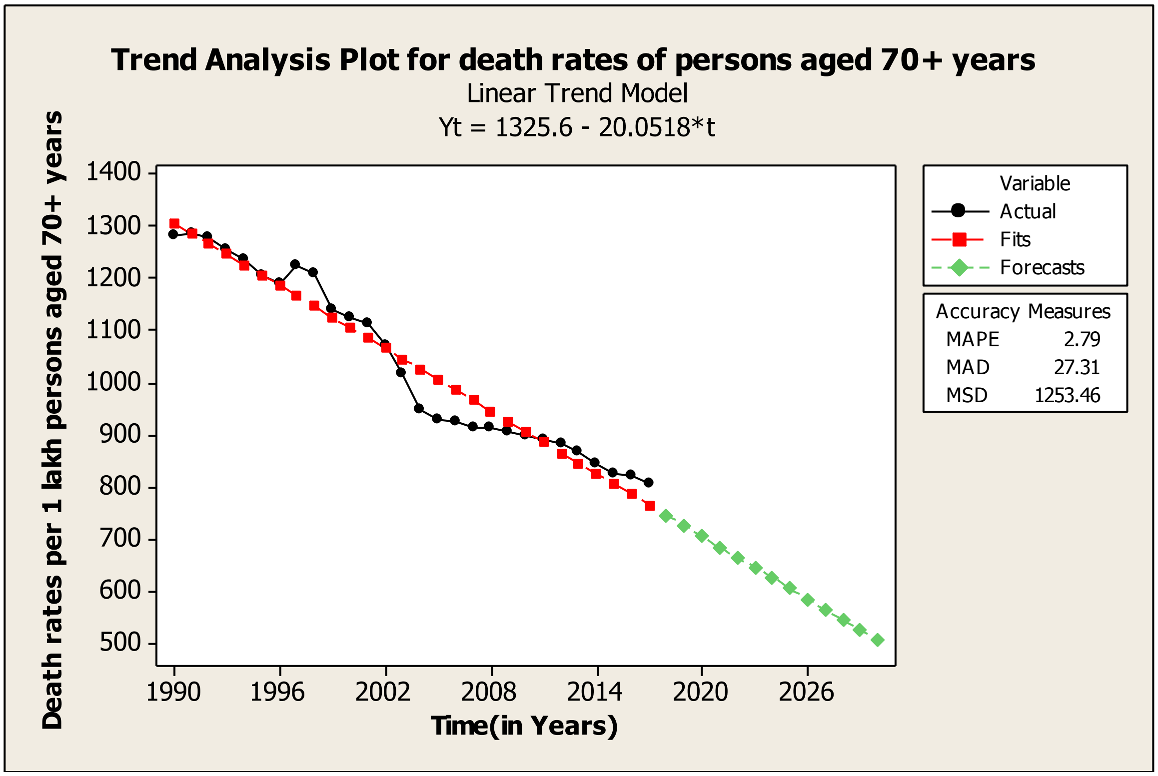
If a & b are the LSE's of α & β,then the fitted trend equation is given by

$$T_t = a + bt$$

Through Minitab, we have obtained the fitted trend equation as

$$T_t = 1325.6 - 20.0518t$$

Here,we have checked for the quadratic trend curve but that didn't give a significant decrease in the MAPE value.



The fitted trend values are then given in the following table:-

Time(in Years)	Fitted death rates per 1 lakh persons aged 70+ years old
1990	1305.590603
1991	1285.538836
1992	1265.487069
1993	1245.435302
1994	1225.383535
1995	1205.331768
1996	1185.280001
1997	1165.228234
1998	1145.176467
1999	1125.1247
2000	1105.072933
2001	1085.021166
2002	1064.969399
2003	1044.917632
2004	1024.865865
2005	1004.814098
2006	984.7623306
2007	964.7105635
2008	944.6587965
2009	924.6070295
2010	904.5552625
2011	884.5034954
2012	864.4517284
2013	844.3999614
2014	824.3481943
2015	804.2964273
2016	784.2446603
2017	764.1928932

The forecasted trend values are then given in the following table:-

Time(in Years)	Forecasted death rates per 1 lakh persons aged 70+ years old
2018	744.141
2019	724.089
2020	704.038
2021	683.986
2022	663.934
2023	643.882
2024	623.831
2025	603.779
2026	583.727
2027	563.675
2028	543.623
2029	523.572
2030	503.52

3)**<u>RUN TEST FOR THE DETECTION OF AUTOCORRELATION:-</u>**

A) Let us consider the data on **death rates per 1 lakh persons from smoking from year t=1990(1)2017**.

Previously,we have fitted a trend equation to the above data and obtained the fitted values.Now,on subtracting the fitted values from the original data,we get $\hat{I}_t = X_t - \hat{T}_t$,which is now suppose to comprise of the residuals only as we have already eliminated the evolutive part.Now we want to check whether this stationary time series is random or not.For this,we need to perform a non-parametric test called runs test(also known as Geary Test).

The values of $\hat{I}_t$ are given in **section 1 of appendix**.

Here we note down the signs (+ve or -ve) of the residuals $\hat{I_t}$ .We now define a run as uninterrupted sequence of 1 sign(+ve or -ve).We further define the length of a run as the no. of signs in a run.

Let N_1=No. of + ve residuals, N_2 = No. of -ve residuals,N = Total no. of obs and R = No.of runs.

To test H_0: Successive observations are not autocorrelated against H_1: $Not\ H_0$

Here we are assuming $N_1>=10$ and $N_2>=10$ under H_0.

$$R \sim N(\mu_R\ , \sigma_R^2) \text{ asymptotically under } H_0$$

where μ_R =$(2N_1N_2+N)/N$ and $\sigma_R^2=2N_1N_2(2N_1N_2-N)/(N^2(N-1))$.

The appropriate test statistic is then given by

$$T=(R-\mu_R)/\sigma_R \sim N(0,1)$$

We reject H_0 $against$ H_1 at level α iff $|T_{obs}| > \tau_{\alpha/2}$.

Note that the test is 2-sided as too many runs indicate that the residuals changes sign frequently and hence there is –ve autocorrelation and on the other hand too small runs indicate that there is +ve autocorrelation.

From the data we have got N_1= 18,N_2 =10,N =28,R =4.

So, $|T_{obs}| = 4.148305$ and $\tau_{\alpha/2} = 1.959964$.

Since $|T_{obs}| > \tau_{\alpha/2}$,so we reject H_0 $against$ H_1 at 5% level of significance and conclude on the basis of the given data that **the residual series is autocorrelated**.

The r code for the above calculation is given in **section 2 of appendix**.

B) Let us consider **the data on death rates per 1 lakh persons from smoking from year t=1990(1)2017 for different age groups**.

Previously,we have fitted a trend equation to the above data for different age groups and obtained the fitted values.Now,on subtracting the fitted values from the original data,we get $\hat{I}_t = X_t - \hat{T}_t$,which is now suppose to comprise of the residuals only as we have already eliminated the evolutive part.Now we want to check whether these 3 stationary time series are random or not.For this,we need to perform a non-parametric test called runs test(also known as Geary Test).

The values of $\hat{I}_t$ for different age groups are given in **section 3 of appendix**.

Here we note down the signs (+ve or -ve) of the residuals $\hat{I}_t$ .We now define a run as uninterrupted sequence of 1 sign(+ve or -ve).We further define the length of a run as the no. of signs in a run.

Let N_{1i}=No. of + ve residuals for the i^{th} age group, N_{2i} = No. of -ve residuals for the i^{th} age group,N_i = Total no. of obs for the i^{th} age group and R_i= No.of runs for the i^{th} age group, i=1(1)3.

To test H_{0i}: Successive observations are not autocorrelated against H_{1i}: $Not\ H_{0i}$,i=1(1)3.

Here we are assuming N_{1i}>=10 and N_{2i}>=10 under H_{0i}, i=1(1)3.

$$R_i \sim N(\mu_{Ri}, \sigma_{Ri}^2) \text{ asymptotically under } H_{0i}$$

where $\mu_{Ri} = (2N_{1i}N_{2i}+N_i)/N_i$ and

$\sigma_{Ri}^2 = 2N_{1i}N_{2i}(2N_{1i}N_{2i}-N_i)/(N_i^2(N_i-1))$,i=1(1)3.

The appropriate test statistics are then given by

$$T_i=(R_i-\mu_{Ri})/\ \sigma_{Ri}\sim N(0,1),\ i=1(1)3.$$

We reject H_{0i} $against$ H_{1i} at level α iff $|\ T_{iobs}| > \ \tau_{\alpha/2}$, i=1(1)3.

Note that the tests are 2-sided as too many runs indicate that the residuals changes sign frequently and hence there is –ve autocorrelation and on the other hand too small runs indicate that there is +ve autocorrelation.

From the data we have got

N_{11}= 14,N_{21} =14,N_1 =28,R_1 =7.

N_{12}= 17,N_{22} =11,N_2 =28,R_2 =4.

N_{13}= 18,N_{23} =10,N_3 =28,R_3 =6.

So, $|T_{1obs}| = 3.081316$ and $\tau_{\alpha/2} = 1.959964.$

So, $|T_{2obs}| = 4.188958.$

So, $|T_{3obs}| = 3.30662.$

Since $|T_{1obs}| > \tau_{\alpha/2}$,so we reject H_{01} $against$ H_{11} at 5% level of significance and conclude on the basis of the given data that **the residual series corresponding to the age group 15-49 years is autocorrelated**.

Since $|T_{2obs}| > \tau_{\alpha/2}$,so we reject H_{02} $against$ H_{12} at 5% level of significance and conclude on the basis of the given data that **the residual series corresponding to the age group 50-69 years is autocorrelated**.

Since $|T_{3obs}| > \tau_{\alpha/2}$,so we reject H_{03} $against$ H_{13} at 5% level of significance and conclude on the basis of the given data **that the residual series corresponding to the persons aged 70+ years is autocorrelated**.

A **similar r code** for the above calculation is given in **section 2 of appendix**.

4)<u>**FITTING OF AR(2) MODEL IN THE RESIDUAL DATA:-**</u>

A) Let us consider the data on **death rates per 1 lakh persons from smoking from year t=1990(1)2017**.Previously,we have obtained data on residual series,i.e.,$\hat{I}_t$ and found out that the residual series is autocorrelated.We shall assume here that I_t follows AR(2) scheme given by

$$I_t = \delta_1 I_{t-1} + \delta_2 I_{t-2} + v_t$$

Here v_t's are random errors with mean 0 and finite variance σ^2.Here δ_1 and δ_2 are unknown constants called parameters.Here I_{t-1} denotes the lagged value of I_t with time lag 1 and I_{t-2} denotes the lagged value of I_t with time lag 2.

Replacing I_t by the residuals $\hat{I}_t$,we get

$$\hat{I}_t = \delta_1 \hat{I}_{t-1} + \delta_2 \hat{I}_{t-2} + u_t \quad \text{where } u_t \text{ denotes the disturbance term.}$$

Here one can obtain the estimates of δ_1 and δ_2 by regressing $\hat{I}_t$ on $\hat{I}_{t-1}$ and $\hat{I}_{t-2}$ without an intercept term.

If $\hat{\delta}_1$ & $\hat{\delta}_2$ are the LSEs of δ_1 & δ_2,then the fitted residual equation is then given by

$$I_t = \hat{\delta}_1 I_{t-1} + \hat{\delta}_2 I_{t-2}$$

By Minitab we have obtained the fitted residual equation as

$$I_t = 1.2858\ I_{t-1} - 0.4914\ I_{t-2}$$

The fitted values of residuals are then given in the following table:-

Time(in Years)	Fitted residuals
1990	-1.668805991
1991	-0.566122692
1992	1.914885915
1993	2.066404121
1994	1.09083322
1995	1.096200719
1996	-0.689378585
1997	1.609587457
1998	9.230340779
1999	5.749244017
2000	-1.731195359
2001	0.707716404
2002	0.520614108
2003	-2.526933227
2004	-5.64684338
2005	-10.8218304
2006	-7.633192695
2007	-5.618895274
2008	-3.699464399
2009	-1.642665055
2010	-0.792549769
2011	0.418237941
2012	1.986283513
2013	3.050601863
2014	2.633844676
2015	1.772828432
2016	2.446648549
2017	4.463139889

The forecasted values of residuals are then given in the following table:-

Time(in Years)	Forecasted residuals
2018	4.04545
2019	2.79664
2020	1.60821
2021	0.69372
2022	0.1018
2023	-0.20997
2024	-0.32
2025	-0.30829
2026	-0.23917
2027	-0.15605
2028	-0.08313
2029	-0.03022
2030	0.00199

Note that here AR(2) model is used instead of AR(1) model as the residual sum of squares obtained from fitting AR(2) model is significantly less than those of AR(1) model.

B)Let us consider the data on **death rates per 1 lakh persons from smoking from year t =1990(1)2017 across various age groups**.Previously,we have obtained data on residual series,i.e.,$\hat{I}_t$ for different age groups and found out that the residual series are autocorrelated.We shall assume here that I_t follows AR(2) scheme given by

$$I_t = \delta_1 I_{t-1} + \delta_2 I_{t-2} + v_t$$

Here v_t's are random errors with mean 0 and finite variance σ^2.Here δ_1 and δ_2 are unknown constants called parameters.Here I_{t-1} denotes the lagged value of I_t with time lag 1 and I_{t-2} denotes the lagged value of I_t with time lag 2.

Replacing I_t by the residuals $\hat{I}_t$,we get

$$\hat{I}_t = \delta_1 \hat{I}_{t-1} + \delta_2 \hat{I}_{t-2} + u_t \quad \text{where } u_t \text{ denotes the disturbance term.}$$

Here one can obtain the estimates of δ_1 and δ_2 by regressing $\hat{I}_t$ on $\hat{I}_{t-1}$ and $\hat{I}_{t-2}$ without an intercept term.

If $\hat{\delta}_1$ & $\hat{\delta}_2$ are the LSEs of δ_1 & δ_2,then the fitted residual equation is then given by

$$I_t = \hat{\delta}_1 I_{t-1} + \hat{\delta}_2 I_{t-2}$$

By Minitab we have obtained the fitted residual equation of group 1 as

$$I_t = 1.2112 I_{t-1} - 0.5275 I_{t-2}$$

By Minitab we have obtained the fitted residual equation of group 2 as

$$I_t = 1.4022 I_{t-1} - 0.5645 I_{t-2}$$

By Minitab we have obtained the fitted residual equation of group 3 as

$$I_t = 1.1907 I_{t-1} - 0.4284 I_{t-2}$$

Here groups 1,2,3 refers to the age groups of persons aged 15-49,50-69,70+ years old respectively.

The fitted values of residuals are then given in the following table:-

Time(in Years)	Fitted residuals for group 1	Fitted residuals for group 2	Fitted residuals for group 3
1990	-0.098325539	-6.496672498	-22.25210431
1991	-0.033536503	-2.681417324	-14.95768693
1992	0.109393289	4.847880498	10.36799077
1993	0.037873665	6.834010534	13.96306198
1994	-0.086889089	5.462609912	6.839158744
1995	-0.058434038	6.735806427	7.705805496
1996	-0.269025446	1.564117574	-4.848203825
1997	0.197669171	13.09501686	3.020422745
1998	0.834678809	34.28465339	66.81084126
1999	0.087387877	19.25535502	52.65632663
2000	0.592182344	-7.079807019	-10.47387843
2001	0.325402758	-1.291585436	18.51280551
2002	-0.181809313	-4.849645154	22.67739586
2003	-0.268124122	-13.8589587	-4.617315742
2004	-0.714812419	-20.64333345	-35.2789764
2005	-1.233469308	-35.3013597	-81.98104995
2006	-0.658430689	-26.87253122	-56.8956059
2007	-0.209064001	-21.96222863	-41.36980599
2008	0.238533564	-12.74995249	-36.61699839
2009	0.276002756	-7.807855174	-16.67678767
2010	0.262512335	-3.969268933	-9.581322578
2011	0.289112556	1.047156346	-2.041877797
2012	0.472901822	7.101065612	10.39441819
2013	0.338385239	12.22116165	18.39586648
2014	0.012546844	10.6864225	17.40565312
2015	-0.173881914	8.09108776	12.30933297
2016	-0.138821998	11.25796329	15.57875938
2017	0.062867118	16.19105844	34.99154921

The forecasted values of residuals are then given in the following table:-

Time(in Years)	Forecasted residuals for group 1	Forecasted residuals for group 2	Forecasted residuals for group 3
2018	-0.077114	14.4547	32.4401
2019	-0.05851	10.7044	21.3423
2020	-0.030192	6.8505	11.5141
2021	-0.005705	3.5634	4.5664
2022	0.009016	1.1297	0.5042
2023	0.01393	-0.4275	-1.3559
2024	0.012116	-1.2372	-1.8306
2025	0.007328	-1.4935	-1.5987
2026	0.002484	-1.3959	-1.1194
2027	-0.000856	-1.1143	-0.6479
2028	-0.002348	-0.7745	-0.2919
2029	-0.002392	-0.457	-0.07
2030	-0.001659	-0.2037	0.0417

Note that here AR(2) model is used instead of AR(1) model as the residual sum of squares obtained from fitting AR(2) model is significantly less than those of AR(1) model.

5) GRAPHICAL REPRESENTATION OF OBSERVED AND FITTED DATA:-

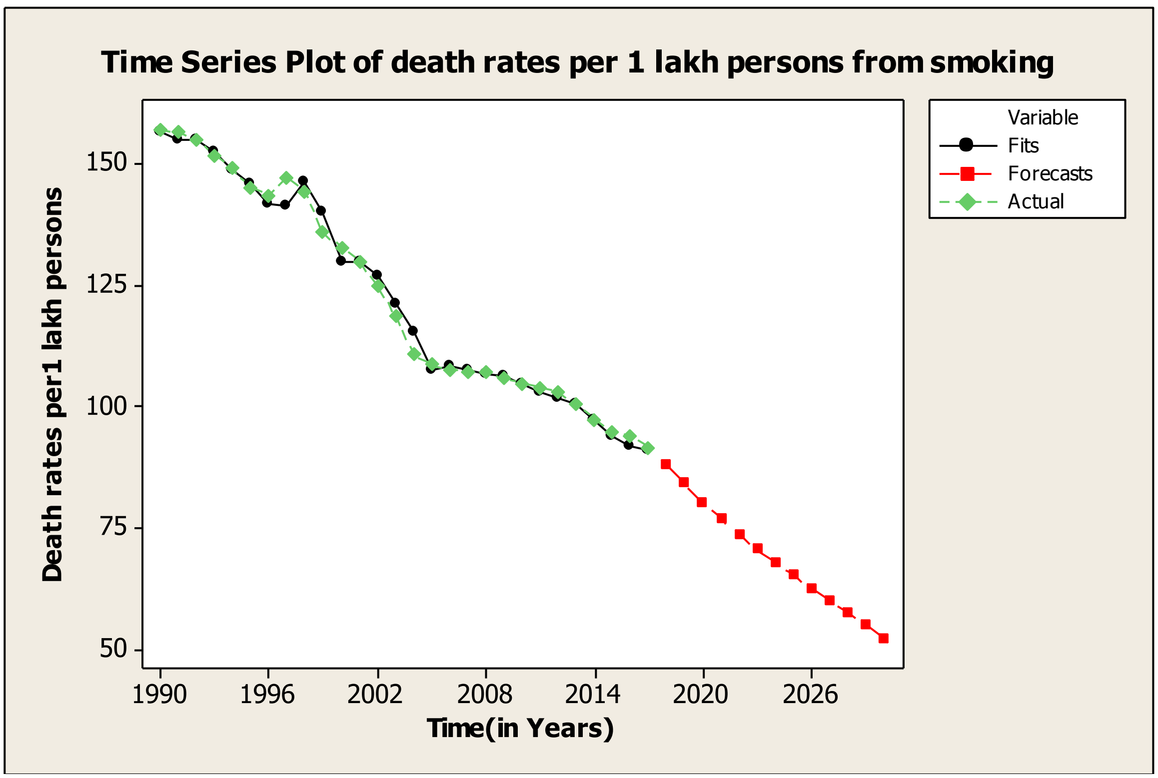
In each of the following cases,we have fitted an appropriate model and extended it up to the year 2030 **for prediction purposes**. In the subsequent graphs,the fitted values are represented by **Black color**, the forecasted values are represented by the **Red color** and the observed values are represented by the **Green color.**

A) Let us consider **the observed data** on **death rates per 1 lakh persons from smoking from year t=1990(1)2017.**Previously we have estimated the trend and irregular component of the above time series.So, after adding them,we will get **the predicted death rates** per 1 lakh persons **from smoking** and the **corresponding data** is given in **section 4 of appendix**.

Since the correlation coefficient between the observed death rate and the predicted death rate from smoking comes out to be 0.996,so we can say that **our fitting is more or less good**.

The **forecasted data** on death rates from smoking **from year 2018 to 2030** is given in **section 5 of appendix.**

The **time series plot** showing **observed,predicted and forecasted data** on death rates per 1 lakh persons from smoking is given in the diagram below:-



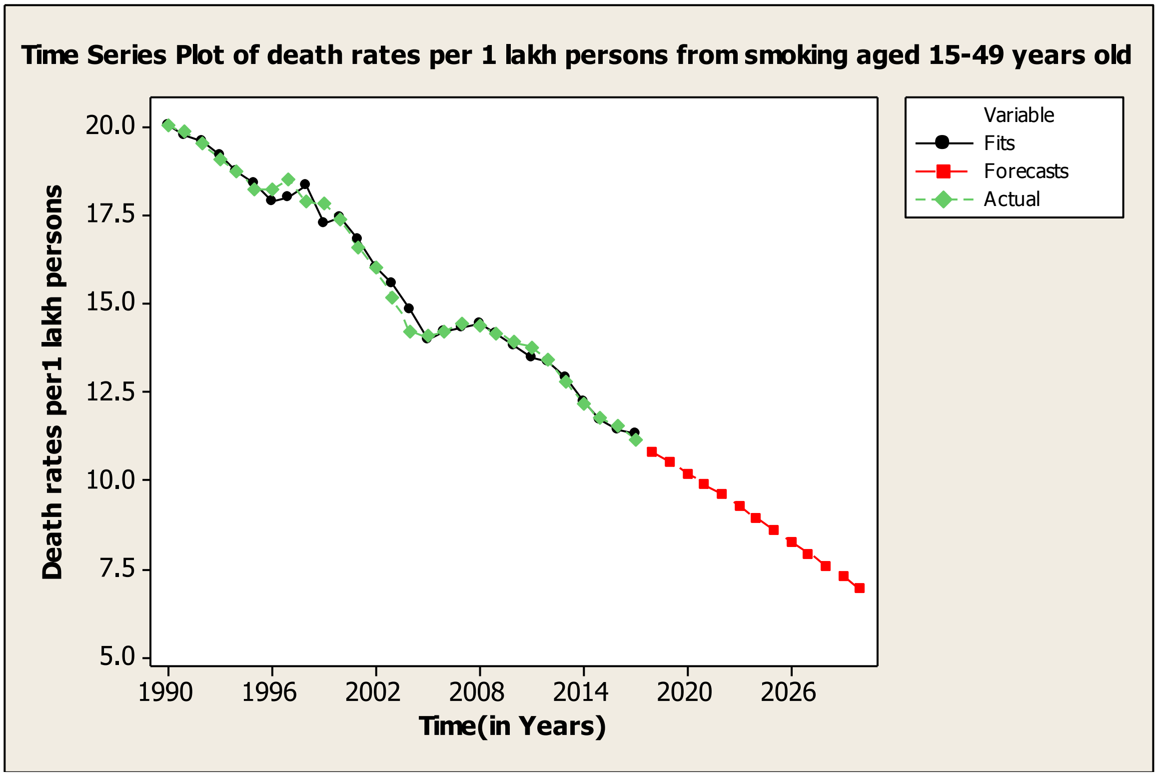
__**Observations:-**__ From the above plot,it can be seen that death rates from smoking(direct and secondhand) per 1 lakh persons in India is seen to be decreasing over time.Moreover,we can also say that our fitting is more or less good.

B)i)Let us consider the data on death rates per 1 lakh persons from smoking aged between 15 to 49 years old from year t=1990(1)2017. Previously we have estimated the trend and irregular component of the above time series.So, after adding them,we will get **the predicted death rates** per 1 lakh persons **from smoking aged between 15 to 49 years old** and the **corresponding data** is given in **section 6 of appendix**.

Since the correlation coefficient between the observed death rate and the predicted death rate per 1 lakh persons from smoking of persons aged between 15 to 49 years old comes out to be 0.996,so we can say that **our fitting is more or less good**.

The **forecasted data** on death rates per 1 lakh persons from smoking **aged between 15 to 49 years old from year 2018 to 2030** is given in **section 7 of appendix.**

The **time series plot** showing **observed,predicted and forecasted data** on death rates per 1 lakh persons from smoking aged between 15 to 49 years old is given in the diagram below:-



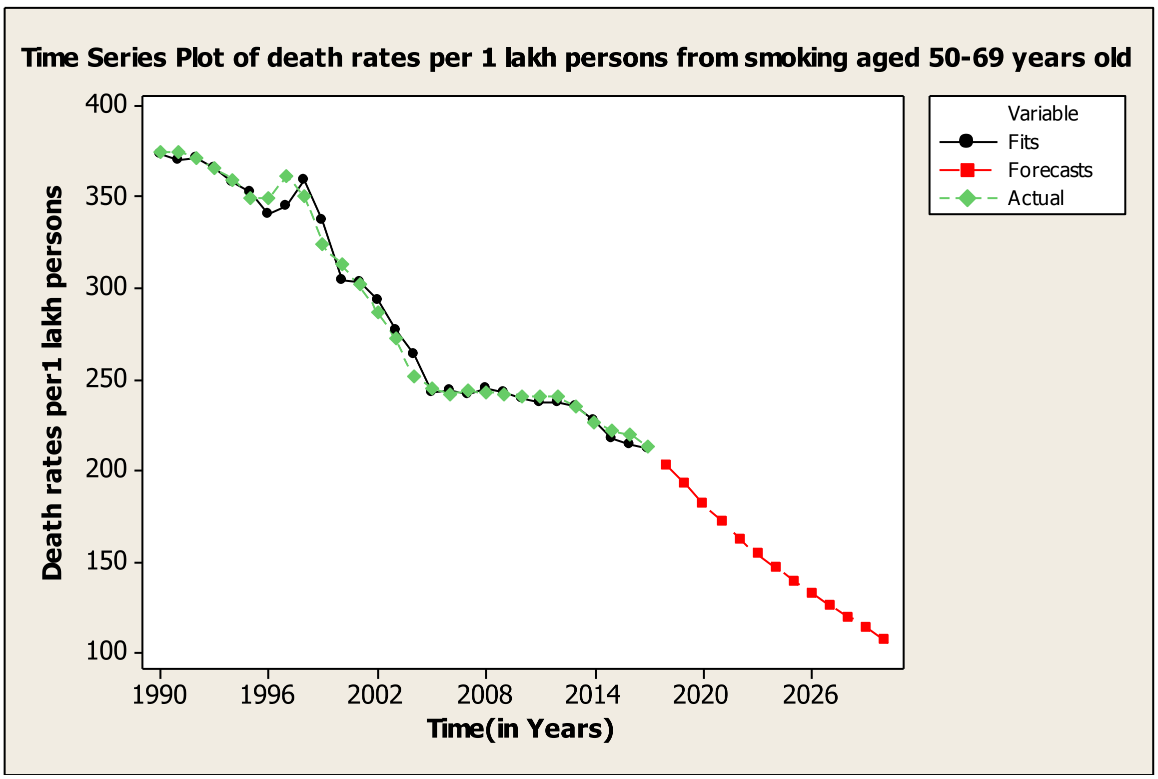
<u>Observations:-</u> From the above plot,it can be seen that death rates from smoking per 1 lakh persons in India aged 15-49 years old is seen to be decreasing over time.Moreover,we can also say that our fitting is more or less good.

ii)Let us consider the data on death rates per 1 lakh persons from smoking **aged between 50 to 69 years old from year t=1990(1)2017**. Previously we have estimated the trend and irregular component of the above time series.So, after adding them,we will get **the predicted death rates** per 1 lakh persons **from smoking aged between 50 to 69 years old** and the **corresponding data** is given in **section 8 of appendix**.

Since the correlation coefficient between the observed death rate and the predicted death rate per 1 lakh persons from smoking of persons aged between 50 to 69 years old comes out to be 0.994,so we can say that **our fitting is more or less good**.

The **forecasted data** on death rates per 1 lakh persons from smoking **aged between 50 to 69 years old from year 2018 to 2030** is given in **section 9 of appendix.**

The **time series plot** showing **observed,predicted and forecasted data** on death rates per 1 lakh persons from smoking aged between 50 to 69 years old is given in the diagram below:-



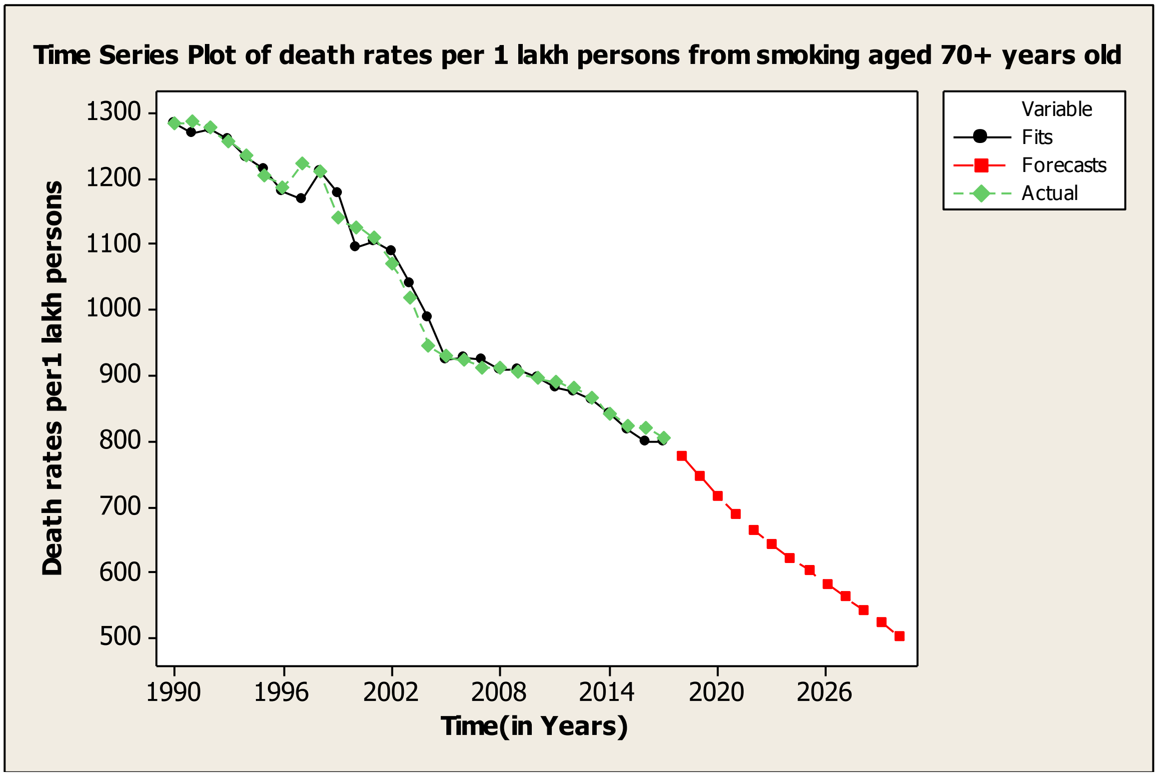
<u>Observations:-</u> From the above plot,it can be seen that death rates from smoking per 1 lakh persons in India aged 50-69 years old is seen to be decreasing over time.Moreover,we can also say that our fitting is more or less good.

iii)Let us consider the data on death rates per 1 lakh persons from **smoking aged 70+ years old from year t=1990(1)2017.** Previously we have estimated the trend and irregular component of the above time series.So, after adding them,we will get **the predicted death rates** per 1 lakh persons **from smoking aged 70+ years old** and the **corresponding data** is given in **section 10 of appendix**.

Since the correlation coefficient between the observed death rate and the predicted death rate per 1 lakh persons from smoking of persons aged 70+ years old comes out to be 0.994,so we can say that **our fitting is more or less good**.

The **forecasted data** on death rates per 1 lakh persons from smoking **aged 70+ years old from year 2018 to 2030** is given in **section 11 of appendix.**

The **time series plot** showing **observed,predicted and forecasted data** on death rates per 1 lakh persons from smoking aged 70+ years old is given in the diagram below:-



<u>**Observations:-**</u> From the above plot,it can be seen that death rates from smoking per 1 lakh persons in India aged 70+ years old is seen to be decreasing over time.Moreover,we can also say that our fitting is more or less good.

CONCLUSION

Though the total no. of yearly deaths from smoking in the World and in India is increasing over time because of exponential increase in population,yet the **death rate per 1 lakh persons from smoking in India is decreasing over time**.Moreover,we have also seen that the **death rate per 1 lakh persons from smoking in India for different age groups is decreasing over time**.

In this project,I have done trend analysis and residual analysis of the **death rates per 1 lakh persons from smoking in India** and have fitted an appropriate model to the above data for forecasting purposes.After analysis,**we conclude that the death rate per 1 lakh persons from smoking in India is decreasing over time overall and also among various age groups.**

This has been possible due to the various measures taken by the government like imposing huge taxes on cigarettes,promote bans on tobacco advertising and to support our fellow countrymen educationally and through other means so that they can quit smoking so that death rates from tobacco smoking falls significantly.Finally,I would like to conclude by telling that this project might draw the attention of policy makers so that they can improvise and create new policies so that death rates from tobacco smoking falls significantly.

LIMITATIONS

There are some limitations in doing this project, they are as follows:

1) Yearly data was being considered for analysis of death rates per 1 lakh persons from smoking.Cyclical fluctuations were not being considered here in this project.

2) Here, for simplicity of calculation we have assumed that the error terms of our chosen models are homoscedastic.

3)In this project,I have used run test. In run test, we have assumed that the no. of + ve residuals and the no. of -ve residuals should be individually >=10.If the no. of + ve residuals or the no. of -ve residuals is less than 10,then run test would become invalid.In that case,we can use other non-parametric tests such as sign test,difference test etc.

4) Accuracy of this project is limited because of factors like limitation of time,money and manpower.

FURTHER SCOPES

The further scopes of this project are as follows:-

1)One can also account for cyclical fluctuations in the data used in this project for further analysis.

2)Consider the data on age-standardized share of cancer death rates from smoking(direct and secondhand) per 1 lakh persons in India(expressed in percentage) from 1990-2016.Here one can use interpolation to fill up the missing values and hence can fit a time series model to the new data for forecasting purposes.

3)One can assume the error terms of the models used in this project to be heteroscedastic and after adjusting for it,one can complete the entire project of mine giving it a new look.

4)Instead of performing a run test with the assumption that the no. of + ve residuals and the no. of -ve residuals should be individually >=10,one can perform a more generalized test of randomness such as sign test,difference test etc.which do not have the above assumption.

REFERENCES AND BIBLIOGRAPHY

1)**BOOK**:-

a)The Analysis of Time Series-An Introduction by Chris Chatfield

2)**WEBSITE:-**

a)https://ourworldindata.org/smoking#what-can-we-do-to-reduce-smoking

b)http://www.healthdata.org/

ACKNOWLEDGEMENT

I am thankful to St.Xavier's College and University of Calcutta for providing me an opportunity to devote myself in studying and carrying out this project.I would also like to thank Fr.Principal and the college authorities for their support and logistics.I would also like to thank our college librarian for providing me with some nice books so that I can complete my project without any hassle.I would also like to thank my dissertation guide Dr.Surabhi Dasgupta and other professors of our department for their continuous guidance,monitoring and inspiration throughout the course of this project.Lastly,I would like to thank my parents and friends for their extraordinary support in completing my project.

APPENDIX

SECTION 1:-

THE VALUES OF $\hat{I}_T$ ARE GIVEN IN THE TABLE BELOW:-

Time(in Years)	Residuals
1990	-1.240366616
1991	1.015260336
1992	1.995049287
1993	1.610737938
1994	1.46805419
1995	0.024848641
1996	1.261304392
1997	7.660610144
1998	7.398678195
1999	1.480895447
2000	1.116304898
2001	0.831469449
2002	-1.647515199
2003	-5.021235348
2004	-10.3351388
2005	-9.885889845
2006	-8.147656794
2007	-5.990639642
2008	-3.566756891
2009	-1.97935864
2010	-0.431106688
2011	1.380032863
2012	2.899871615
2013	3.156532066
2014	2.584980727
2015	2.890615539
2016	4.5756751
2017	4.894743632

SECTION 2:-

R CODE FOR RUN TEST FOR THE DETECTION OF AUTOCORRELATION:-

```
data=c(-
1.2404,1.0153,1.9950,1.6107,1.4681,0.0248,1.2613,7.6606,7.3987,1.4809,1.1163,0.8315,-
1.6475,-5.0212,-10.3351,-9.8859,-8.1477,-5.9906,-3.5668,-1.9794,-
0.4311,1.3800,2.8999,3.1565,2.5850,2.8906,4.5757,4.8947)

N1=length(data[data>0])

N2=length(data[data<0])

N=length(data)

R=4

mur=(2*N1*N2+N)/N

sigma2=2*N1*N2*(2*N1*N2-N)/(N*N*(N-1))

sigma=sqrt(sigma2)

Tobs=abs((R-mur)/sigma)

Tobs

Tcrit=qnorm(1-0.025)

Tcrit
```

SECTION 3:-

THE VALUES OF $\hat{I}_T$ FOR DIFFERENT AGE GROUPS ARE GIVEN IN THE TABLE BELOW:-

Time(in Years)	Residuals for age group of 15-49 years old	Residuals for age group of 50-69 years old	Residuals for age group of 70+ years old
1990	-0.084597284	-5.11952982	-22.17454708
1991	0.053472906	1.39623991	0.728846948
1992	0.054556455	5.43569104	11.98896598
1993	-0.047976316	6.08385837	10.05752401
1994	-0.069137386	7.2527591	10.09041804
1995	-0.252218337	4.03518813	-0.441089926
1996	0.053353803	10.96302886	2.377960106
1997	0.712350322	28.86315229	56.96594414
1998	0.382380961	25.35124032	64.71966217
1999	0.655437651	5.15679335	14.4903242
2000	0.55410123	1.154896781	20.76151523
2001	0.09121211	-2.993546389	26.51552826
2002	-0.181640961	-11.08849096	5.662715295
2003	-0.669258001	-19.18549193	-27.59115767
2004	-1.309823712	-32.8983348	-78.77839944
2005	-1.114039863	-32.40780087	-76.12828291
2006	-0.657775673	-28.70859384	-62.13557968
2007	-0.089531214	-20.64975821	-53.10929225
2008	0.188877636	-13.88109498	-33.11499442
2009	0.298988735	-8.418775949	-19.96182688
2010	0.368904084	-2.642389919	-8.897285752
2011	0.551089874	4.000301611	5.52831498
2012	0.519375493	10.32580364	17.43869701
2013	0.236550103	11.77778657	20.89253574
2014	-0.040538558	10.511477	17.85516957
2015	-0.132266838	12.26012183	19.50809001
2016	-0.005699699	16.48206266	36.40643824
2017	-0.06614752	16.94340619	40.34380607

SECTION 4:-

THE OBSERVED AND PREDICTED DATA ON DEATH RATES PER 1 LAKH PERSONS FROM SMOKING ARE GIVEN IN THE TABLE BELOW:-

Time(in Years)	Observed death rates per 1 lakh persons	Predicted death rates per 1 lakh persons
1990	156.9408092	156.5123698
1991	156.5483017	154.9669187
1992	154.8799562	154.7997928
1993	151.8475104	152.3031766
1994	149.0566922	148.6794712
1995	144.9653522	146.0367043
1996	143.5536735	141.6029905
1997	147.3048448	141.2538221
1998	144.3947784	146.226441
1999	135.8288612	140.0972098
2000	132.8161362	129.9686359
2001	129.8831663	129.7594133
2002	124.7560472	126.9241765
2003	118.7341926	121.2284947
2004	110.7721547	115.4604501
2005	108.5732692	107.6373286
2006	107.6633678	108.1778319
2007	107.1722505	107.5439949
2008	106.9479988	106.8152913
2009	105.8872626	106.2239562
2010	104.7873801	104.425937
2011	103.9503852	102.9885903
2012	102.8220895	101.9085014
2013	100.4306155	100.3246853
2014	97.21092971	97.25979366
2015	94.86843007	93.75064296
2016	93.90535518	91.77632863
2017	91.57628926	91.14468552

SECTION 5:-

THE FORECASTED DATA ON DEATH RATES PER 1 LAKH PERSONS FROM SMOKING IS GIVEN IN THE TABLE BELOW:-

Time(in Years)	Forecasted death rates per 1 lakh persons
2018	88.07885
2019	84.18194
2020	80.34531
2021	76.78272
2022	73.5427
2023	70.58273
2024	67.8246
2025	65.18821
2026	62.60913
2027	60.04415
2028	57.46897
2029	54.87368
2030	52.25779

SECTION 6:-

THE OBSERVED AND PREDICTED DATA ON DEATH RATES PER 1 LAKH PERSONS FROM SMOKING AGED 15-49 YEARS OLD ARE GIVEN IN THE TABLE BELOW:-

Time(in Years)	Observed death rates of persons aged 15-49 years old	Predicted death rates of persons aged 15-49 years old
1990	20.05070494	20.03697669
1991	19.85882712	19.77181771
1992	19.52996266	19.58479949
1993	19.09748188	19.18333186
1994	18.7463728	18.7286211
1995	18.23334384	18.42712814
1996	18.20896797	17.88658872
1997	18.53801648	18.02333533
1998	17.87809911	18.33039696
1999	17.82120779	17.25315802
2000	17.38992336	17.42800447
2001	16.59708623	16.83127688
2002	15.99428515	15.9941168
2003	15.1767201	15.57785398
2004	14.20620638	14.80121767
2005	14.07204222	13.95261277
2006	14.1983584	14.19770338
2007	14.43665485	14.31712206
2008	14.38511569	14.43477162
2009	14.16527878	14.1422928
2010	13.90524612	13.79885437
2011	13.7574839	13.49550658
2012	13.39582151	13.34934784
2013	12.78304811	12.88488325
2014	12.17601144	12.22909684
2015	11.75433515	11.71272007
2016	11.55095428	11.41783198
2017	11.16055845	11.28957309

SECTION 7:-

THE FORECASTED DATA ON DEATH RATES PER 1 LAKH PERSONS FROM SMOKING AGED 15-49 YEARS OLD ARE GIVEN IN THE TABLE BELOW:-

Time(in Years)	Forecasted death rates of persons aged 15-49 years old
2018	10.819686
2019	10.50829
2020	10.206708
2021	9.901195
2022	9.586016
2023	9.26093
2024	8.929216
2025	8.594428
2026	8.259684
2027	7.926344
2028	7.594952
2029	7.264908
2030	6.935741

SECTION 8:-

THE OBSERVED AND PREDICTED DATA ON DEATH RATES PER 1 LAKH PERSONS FROM SMOKING AGED 50-69 YEARS OLD ARE GIVEN IN THE TABLE BELOW:-

Time(in Years)	Observed death rates of persons aged 50-69 years old	Predicted death rates of persons aged 50-69 years old
1990	374.5320832	373.1549405
1991	374.2345494	370.1568922
1992	371.460697	370.8728865
1993	365.2955608	366.045713
1994	359.651158	357.8610088
1995	349.6202835	352.3209018
1996	349.7348207	340.3359094
1997	360.8216406	345.0535052
1998	350.4964251	359.4298382
1999	323.4886746	337.5872363
2000	312.6734745	304.4387707
2001	301.7117278	303.4136888
2002	286.8034797	293.0423255
2003	271.8931752	277.2197084
2004	251.3670288	263.6220302
2005	245.0442592	242.1507004
2006	241.9301627	243.7662253
2007	243.1756948	241.8632244
2008	243.1310545	244.262197
2009	241.78007	242.3909908
2010	240.7431525	239.4162735
2011	240.5725405	237.6193952
2012	240.084739	236.860001
2013	234.7234184	235.1667935
2014	226.6438053	226.8187508
2015	221.5791466	217.4101125
2016	218.9877839	213.7636845
2017	212.6358239	211.8834761

SECTION 9:-

THE FORECASTED DATA ON DEATH RATES PER 1 LAKH PERSONS FROM SMOKING AGED 50-69 YEARS OLD ARE GIVEN IN THE TABLE BELOW:-

Time(in Years)	Forecasted death rates of persons aged 50-69 years old
2018	203.3337
2019	192.7704
2020	182.1035
2021	172.0024
2022	162.7557
2023	154.3855
2024	146.7618
2025	139.6925
2026	132.9771
2027	126.4447
2028	119.9715
2029	113.476
2030	106.9153

SECTION 10:-

THE OBSERVED AND PREDICTED DATA ON DEATH RATES PER 1 LAKH PERSONS FROM SMOKING AGED 70+ YEARS OLD ARE GIVEN IN THE TABLE BELOW:-

Time(in Years)	Observed death rates of persons aged 70+ years old	Predicted death rates of persons aged 70+ years old
1990	1283.416056	1283.338499
1991	1286.267683	1270.581149
1992	1277.476035	1275.85506
1993	1255.492826	1259.398364
1994	1235.473953	1232.222694
1995	1204.890678	1213.037573
1996	1187.657961	1180.431797
1997	1222.194178	1168.248657
1998	1209.896129	1211.987308
1999	1139.615024	1177.781026
2000	1125.834448	1094.599054
2001	1111.536694	1103.533971
2002	1070.632114	1087.646795
2003	1017.326474	1040.300316
2004	946.0874652	989.5868882
2005	928.6858147	922.8330477
2006	922.6267509	927.8667247
2007	911.6012713	923.3407576
2008	911.5438021	908.0417981
2009	904.6452026	907.9302418
2010	895.6579767	894.9739399
2011	890.0318104	882.4616176
2012	881.8904254	874.8461466
2013	865.2924971	862.7958278
2014	842.2033639	841.7538475
2015	823.8045173	816.6057603
2016	820.6510985	799.8234196
2017	804.5366993	799.1844424

SECTION 11:-

THE FORECASTED DATA ON DEATH RATES PER 1 LAKH PERSONS FROM SMOKING AGED 70+ YEARS OLD ARE GIVEN IN THE TABLE BELOW:-

Time(in Years)	Forecasted death rates of persons aged 70+ years old
2018	776.5811
2019	745.4313
2020	715.5521
2021	688.5524
2022	664.4382
2023	642.5261
2024	622.0004
2025	602.1803
2026	582.6076
2027	563.0271
2028	543.3311
2029	523.502
2030	503.5617