

Autocorrelation Analysis

Data Background:

This dataset contains time-based information on Data for Soft Drink Sales from 1960 to 1979. The goal is to predict the Annual Regional Concentrate Sales from Annual Advertising Expenditures and Annual Regional Population in a region.

Data Glimpse:

Year	Sales	Expenditure	Population
1960	3083	75	825000
1961	3149	78	830445
1962	3218	80	838750
1963	3239	82	842940
1964	3295	84	846315
1965	3374	88	852240
1966	3475	93	860760
1967	3569	97	865925
1968	3597	99	871640
1969	3725	104	877745
1970	3794	109	886520
1971	3959	115	894500
1972	4043	120	900400
1973	4194	127	904005
1974	4318	135	908525
1975	4493	144	912160
1976	4683	153	917630
1977	4850	161	922220
1978	5005	170	925910
1979	5236	182	929610

Total Number of Rows: 20

Total Number of Columns: 4

Column Details:

- Year: Year of data recording.
- Sales: Annual Regional Concentrate Sales (in units).
- Expenditure: Annual Advertising Expenditures (in thousands of dollars)
- Population: Annual Regional Population.

Main Dependent Variable: Sales (in units).

Using SPSS software EViews, we have analysed the data:

Descriptive Statistics:

	SALES	EXPENDIT...	POPULATION
Mean	3914.950	114.8000	880662.0
Median	3759.500	106.5000	882132.5
Maximum	5236.000	182.0000	929610.0
Minimum	3083.000	75.00000	825000.0
Std. Dev.	663.0818	32.98899	33993.82
Skewness	0.540345	0.603044	-0.128336
Kurtosis	2.086083	2.154859	1.666992
Jarque-Bera	1.669279	1.807424	1.535658
Probability	0.434031	0.405063	0.464019
Sum	78299.00	2296.000	17613240
Sum Sq. Dev.	8353871.	20677.20	2.20E+10
Observations	20	20	20

Inferences:

- The variable sales is slightly right skewed, ranging between 3083 to 5236 units.
- The variable expenditure is slightly right skewed, ranging between 75 to 182 thousand dollars.
- The variable population is slightly left skewed, ranging between 825000 to 929610 residents in the area.
- There is no missing data.

Correlation Analysis:

	SALES	EXPENDIT...	POPULATION
SALES	1.000000	0.999546	0.964438
EXPE...	0.999546	1.000000	0.957756
POPU...	0.964438	0.957756	1.000000

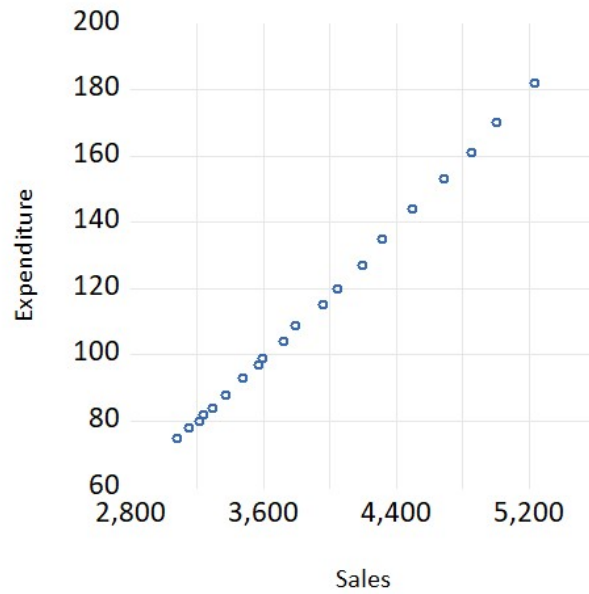
Inferences:

- Variables sales and expenditure have very high positive linear correlation, having correlation coefficient 0.99.
- Variables sales and population have very high positive linear correlation, having correlation coefficient 0.96.
- Variables expenditure and population have very high positive linear correlation, having correlation coefficient 0.95.

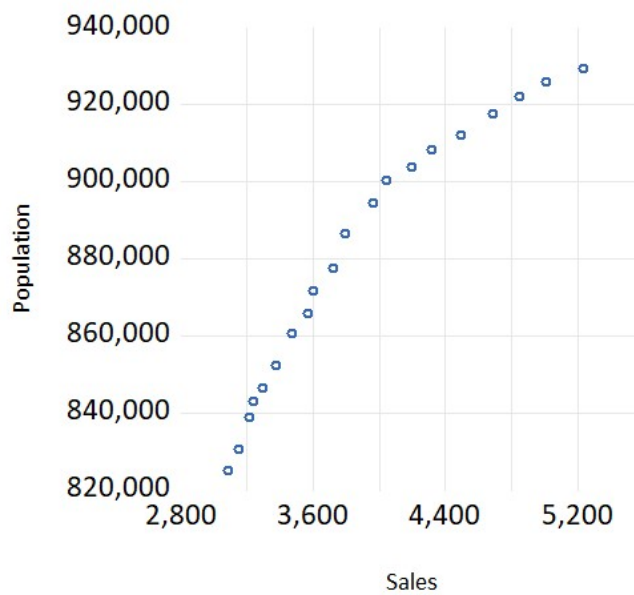
Since the variables are highly correlated with each other, we may infer the problem of autocorrelation. However, to confirm the same, we need to run a few more tests as follows:

- Plot scatter plots and check for correlation.
- Build a regression model and check for very high R^2 value (above 0.9).
- Run the Durbin-Watson test and check for result value between 0-1.7 and 2.3-4.

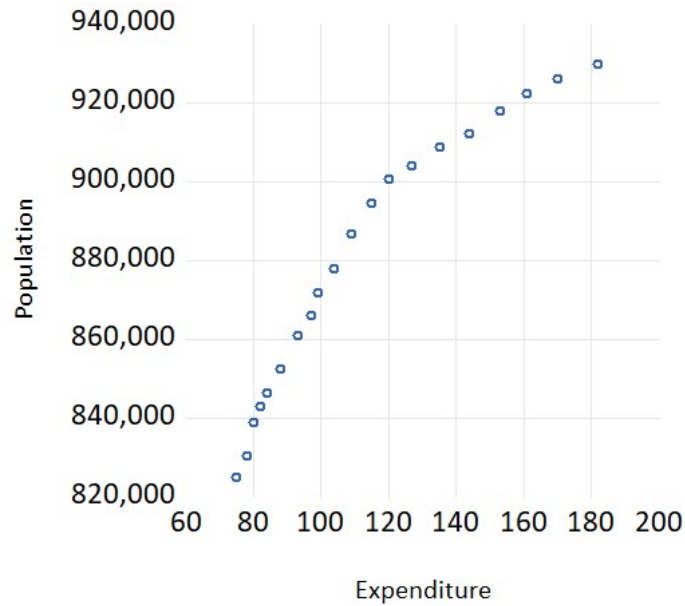
Scatter Plots:



Inference: the variables expenditure and sales have very high positive linear correlation.



Inference: the variables sales and population have very high positive linear correlation.



Inference: the variables expenditure and population have very high positive linear correlation.

Looking at the scatter plots, we may confirm autocorrelation with some degree of confidence. However, to confirm our inference, we need to:

- Build a regression model and check for very high R^2 value (above 0.9).
- Run the Durbin-Watson test and check for result value between 0-1.7 and 2.3-4.

Regression Model:

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	320.3396	217.3278	1.473993	0.1588
EXPENDITURE	18.43421	0.291532	63.23226	0.0000
POPULATION	0.001679	0.000283	5.933585	0.0000
R-squared	0.999704	Mean dependent var	3914.950	
Adjusted R-squared	0.999669	S.D. dependent var	663.0818	
S.E. of regression	12.05574	Akaike info criterion	7.954440	
Sum squared resid	2470.794	Schwarz criterion	8.103800	
Log likelihood	-76.54440	Hannan-Quinn criter.	7.983596	
F-statistic	28730.40	Durbin-Watson stat	3.059322	
Prob(F-statistic)	0.000000			

From the above model, we get the estimate equation:

$$\text{Sales} = 320.3396 + (18.43421)(\text{Expenditure}) + (0.001679)(\text{Population})$$

Inferences:

- The model has very good explanatory power, having R^2 value 0.99.
- The independent variables expenditure and population are statistically significant, having p-values 0.000 and 0.000.
- The intercept has a p-value of 0.1588 and is statistically insignificant.
- The Durbin-Watson test returns a value of 3.059.

From the inferences made using the high value of R^2 , correlation matrix, and within range (2.3-4) return value of the Durbin-Watson test, we can confirm the presence of autocorrelation.