My research question is “Does air pollution cause the respiratory disease?"

My dataset

1. Air pollution data: <https://data.epa.gov.tw/dataset/stat_p_116/resource/fb4dbfff-f6e9-4bdf-ac16-5914422a60f9>
2. Annual report from 1981 to 2020 for cause of death <https://dep.mohw.gov.tw/dos/cp-5069-62791-113.html>

Step1. Load the data:

1. Download the data to the folder and rename the column to correct name.
2. Change the data frame to correct type

Step2. Check the missing data and interpolate:

1. Interpolate missing by properly method :

air pollution by linear interpolation <https://machinelearningmastery.com/resample-interpolate-time-series-data-python/>

cause of death by cubic spline

<https://rdo.psu.ac.th/sjstweb/journal/36-2/36-2-15.pdf>

1. Define the independent variable and dependent variable.
2. Combine all cause of death that is related to respiratory and change to percentage. (The total of death is different. Use percentage will be better for the changing of the number.)
3. Create a heatmap to check for correlations among variables. (by bokeh and seaborn)

Step3. Check Assumptions

1. Check whether it is suitable to use ANOVA (dependent variable is normally distributed, the observations are independent of each other)
2. Use QQ plot to check whether the data is normal distribution.
3. Use histogram to plot the dependent variable.
4. Test by Shapiro and find all of my data are normal distribution.
5. ANOVA also assumes homogeneity of variance. My data is normal distribution, so I cannot use Levene’s test. I analyzed by Bartlett’s test and found my data do not contain equal variance. This means I cannot use ANOVA.

Step4. Try another way to solve the problem

1. I did different regression line and plot p-value, R, R2, in the graph.
2. Every air pollution index has negative correlation with respiratory diseases. Only PM total, PM 2.5, SO2 have the higher R2.(It is higher than 0.75)
3. The standard error is too high for these three data, so I cannot explain the less air pollution cause the less respiratory diseases.
4. We need more data and other research to make sure it is true.