Statistical Analysis of Capstone Project One

The purpose of this document is to explore some of the basic concepts of statistics using capstone project one dataset. We started using basic python condes to explore the descriptive statistics of dataset. Dataset has 10868 observations in all columns. Same number of observations also implies that there is no missing values in our dataset.

Ordinary Least- Squares (OLS) Regression technique is used for statistical learning of the dataset using the statsmodel python package. OLS uses squared error which has nice mathematical properties thereby making it easier to differentiate and compute gradient descent. This method is easy to analyze than more sophisticated models and computationally faster. We used Age-Adjusted Deaths as a dependent variable (y) and other attributes year, deaths, causes, log of deaths as independent variable(X). Age adjusted death rate is a measure that controls for the effects of age differences on health event rates. When comparing across geographic areas, age adjusting methods is used for the influence that population age distributions might have an health even rates (https://ibis.health.state.nm.us/resource/AARate.html.)

Checking the OLS Assumptions:

Hypothesis Testing: Since this is a Multiple Linear Regression, we need to know the importance of variables(significance) with respect to the hypothesis. To do this, we need to calculate the p value for each variable and if it is less than the desired cutoff (0.05 is the general cut off for 95% significance) then we can say with confidence that a variable is significant. Our OLS model shows all variables are significant as the p value is less than 0.05

R Square (Coefficient of Determination): - This metric explains the percentage of variance explained by covariates in the model. It ranges from 0-1. R square calculated for our model is 0.50 (approx.) meaning 50 percent variability of attributes of our sample.

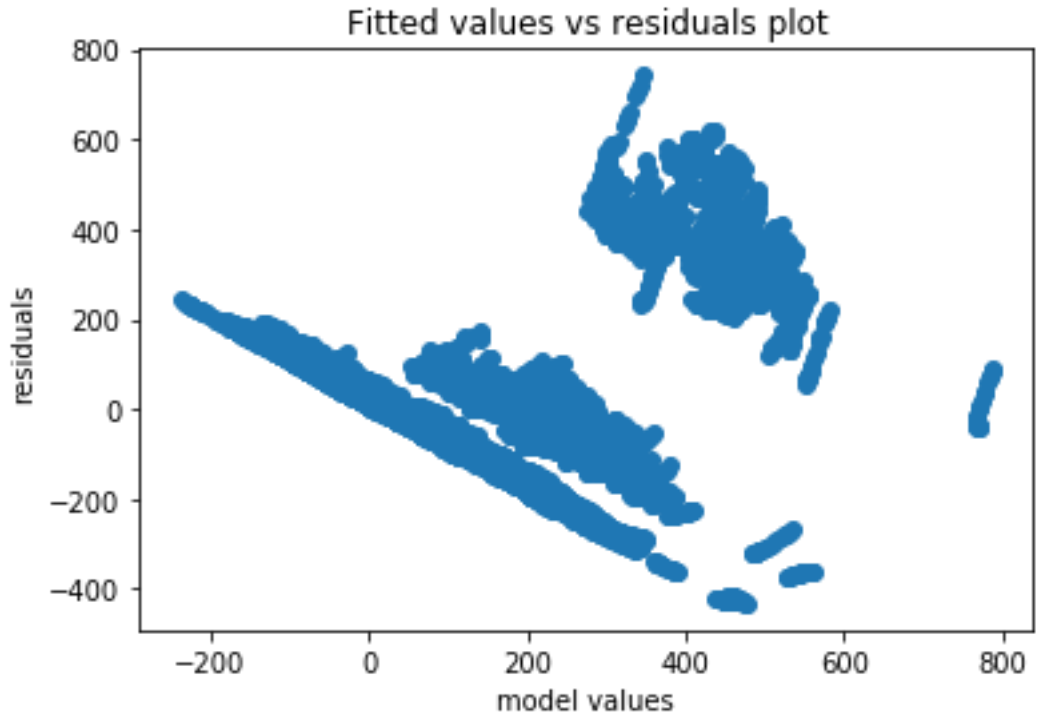
F Statistics - It evaluates the overall significance of the model. It is the ratio of explained variance by the model by unexplained variance. It compares the full model with an intercept only (no predictors) model. Its value can range between zero and any arbitrary large number. Naturally, higher the F statistics, better the model. Our model has f statistics of 2129 which implies that overall regression is meaningful.

Checking for multicollinearity for regression: - The variance inflation factor is used to check multicollinearity on our model. Vif value equal to 1 no multicollinearity. Table 1 shows the value of Vif .Vif value equal to 1 show no multicollinearity.

	VIF	Features
0	1.325250	Year
1	1.076213	Deaths
2	1.399952	Age-adjusted Death Rate

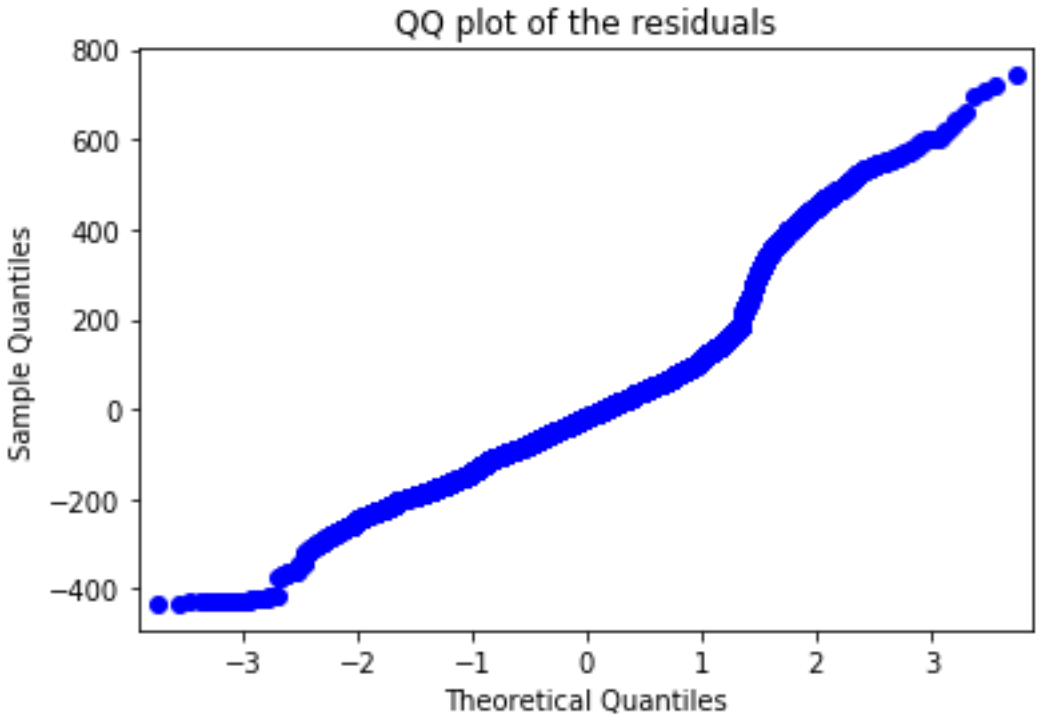
Residual vs. Fitted Values Plot

Ideally, this plot should not show any pattern. But if you see any shape (curve, U shape), it suggests non-linearity in the data set. In addition, if you see a funnel shape pattern, it suggests your data is suffering from heteroskedasticity, i.e. the error terms have non-constant variance



Normality Q-Q Plot

As the name suggests, this plot is used to determine the normal distribution of errors. It uses standardized values of residuals. Ideally, this plot should show a straight line. If you find a curved, distorted line, then your residuals have a non-normal distribution (problematic situation).



References:

- https://ibis.health.state.nm.us/resource/AARate.html.)
- https://www.statsmodels.org/stable/generated/statsmodels.stats.outliers_influence.variance_inflation_factor.html
- https://statisticalhorizons.com/multicollinearity