Fall 2023 DATA 220 Mathematical Methods for Data Analytics

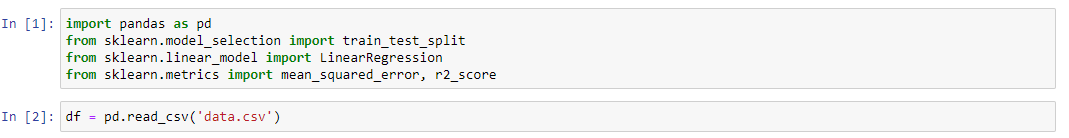
Homework – 5

Name :- Prayag Nikul Purani

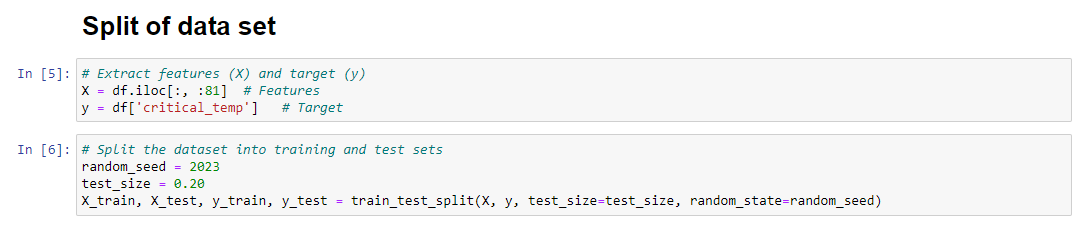
SJSU Id :- 017416737

**Problem 1:**

**Importing the file**

****

**Splitting the dataset into test and train data.**

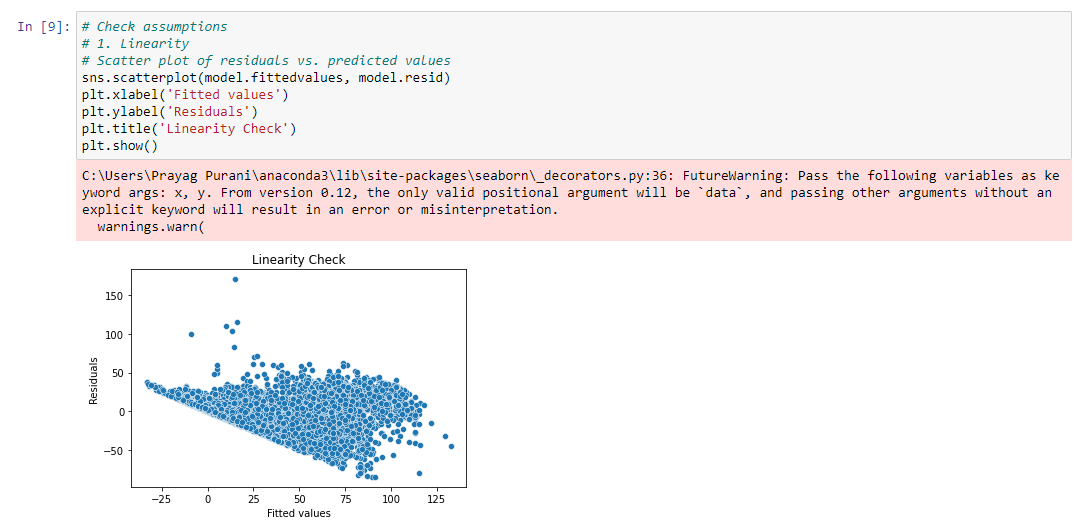
****

1. **Five Assumptions**

## Linearity:

There should be a linear connection between the dependent variable (target) and the independent variables (features). This implies that adjustments to the independent variables ought to cause a corresponding adjustment to the dependent variable.

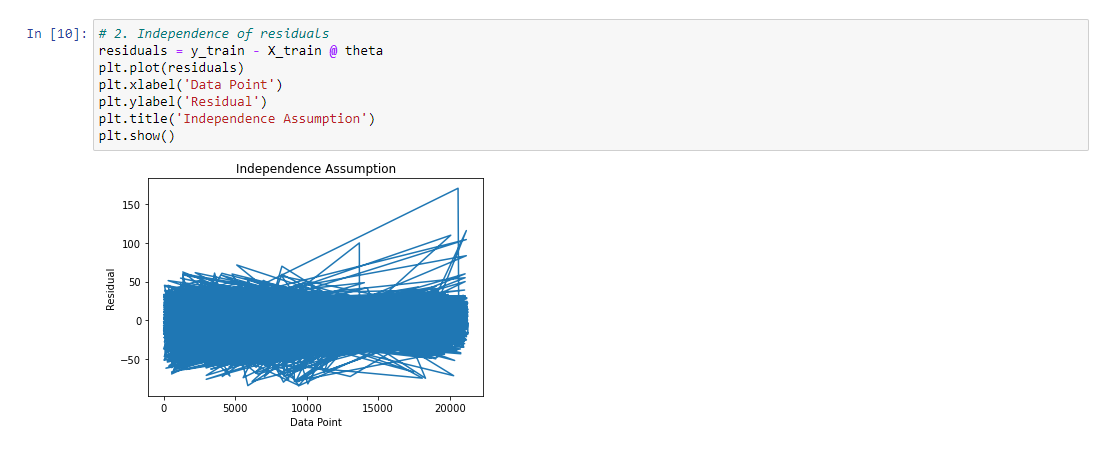
Interpretation: To verify that the connection is generally linear, plot each independent variable against the dependent variable. This will allow you to verify the assumption.



## Independence:

The disparities between the observed and expected values, or residuals, need to be unrelated to one another. The residuals shouldn't show any consistent trends.

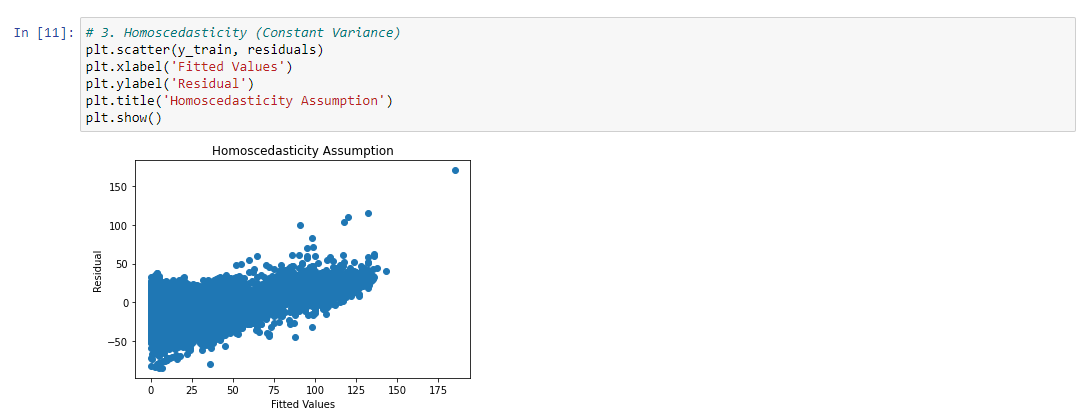
Interpretation: Look for residual autocorrelation. Finding patterns can be aided by plotting residuals versus time or any other pertinent variable.



## Homoscedasticity:

All levels of the independent variables should have a consistent variance in the residuals. Stated otherwise, there should be an approximately constant residual spread.

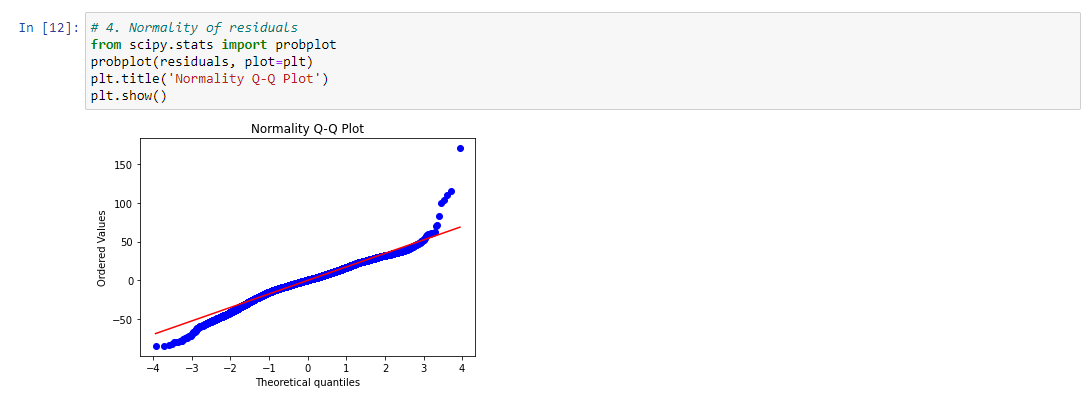
Plot the residuals against the expected values to interpret the results. The assumption is satisfied if the residual spread stays mostly constant. A discernible pattern, like the form of a funnel, might point to heteroscedasticity.



## Normality:

A normal distribution should be seen in the residuals. For larger sample sizes, this assumption is not essential, but for lower sample sizes, normalcy is necessary to draw reliable statistical conclusions.

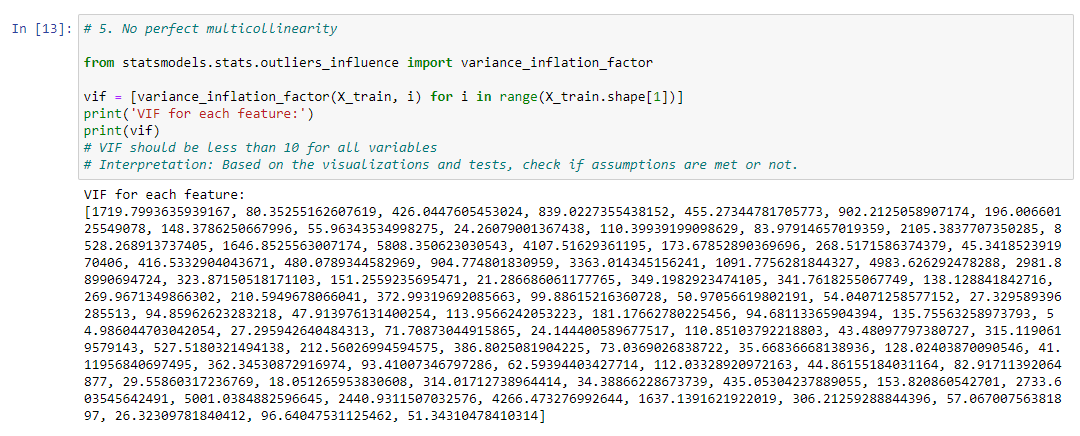
Interpretation: Plot the residuals as a Q-Q plot or as a histogram. The assumption is deemed met if the distribution is roughly normal.



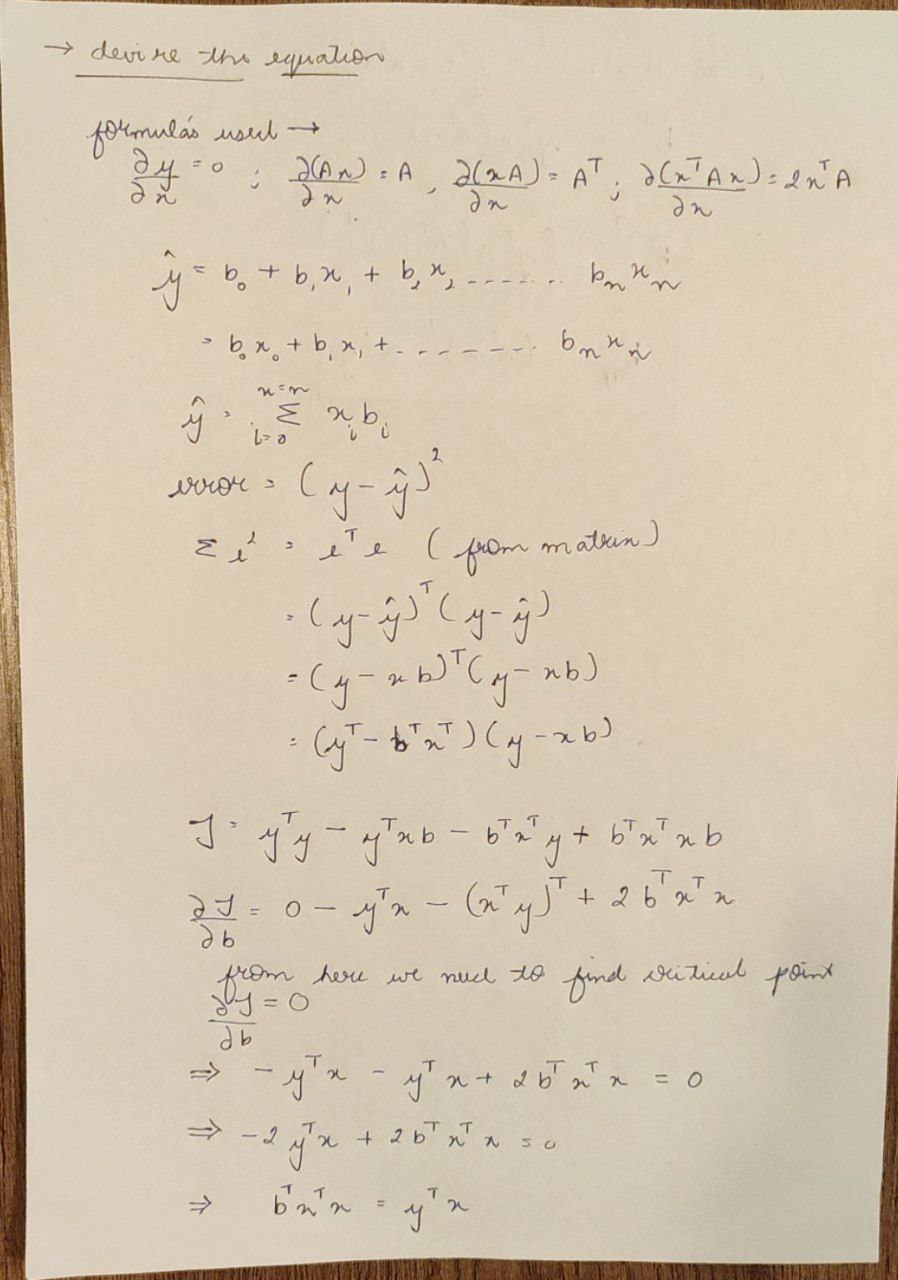
## No perfect Multicollinearity:

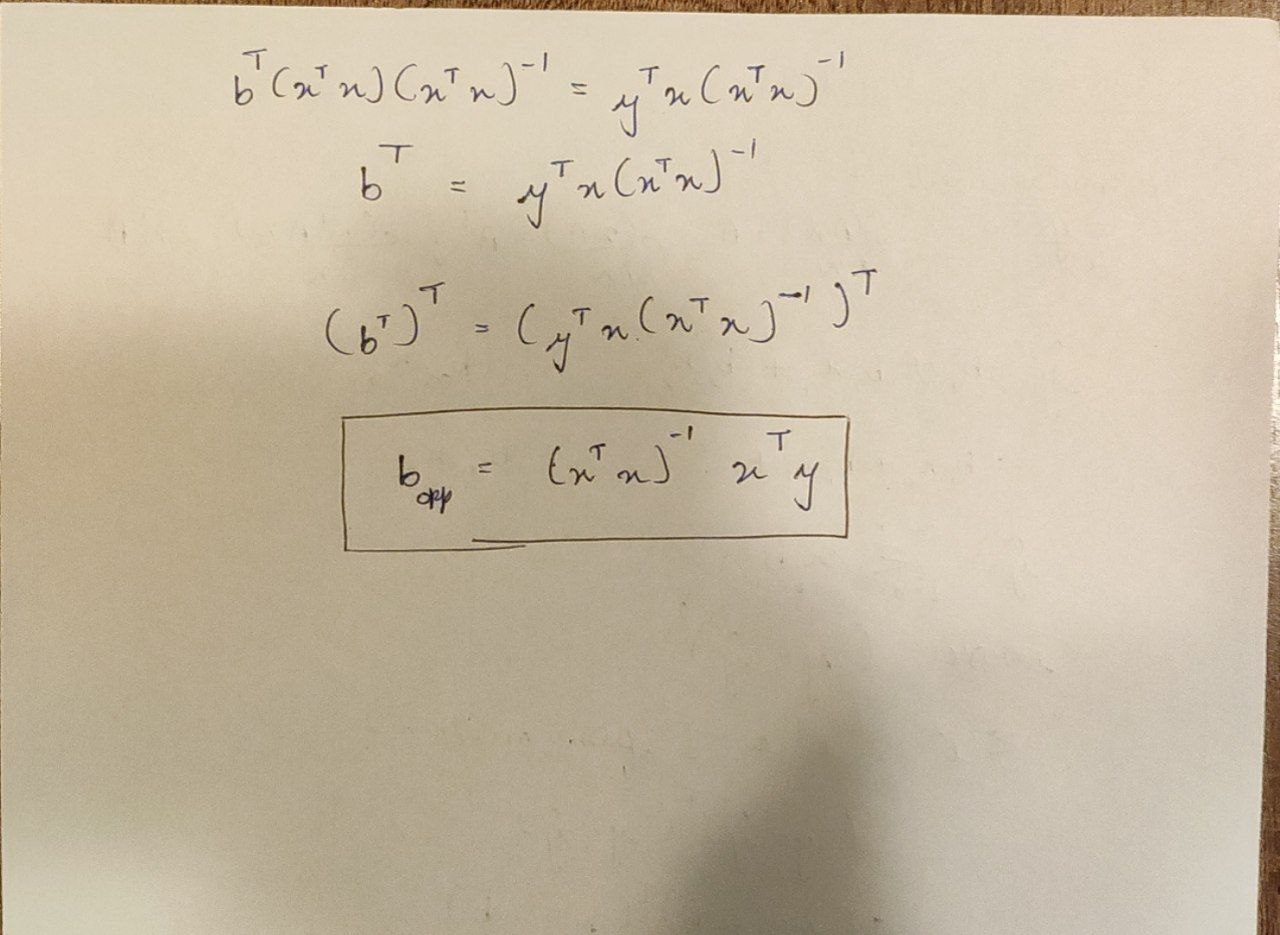
The independent variables shouldn't have perfect multicollinearity. That is to say, no independent variable ought to be a perfect linear combination of other variables.

Interpretation: For every independent variable, look up the variance inflation factor (VIF). More than ten VIF values are frequently seen as suggestive of multicollinearity.



1. **Derive the equation**

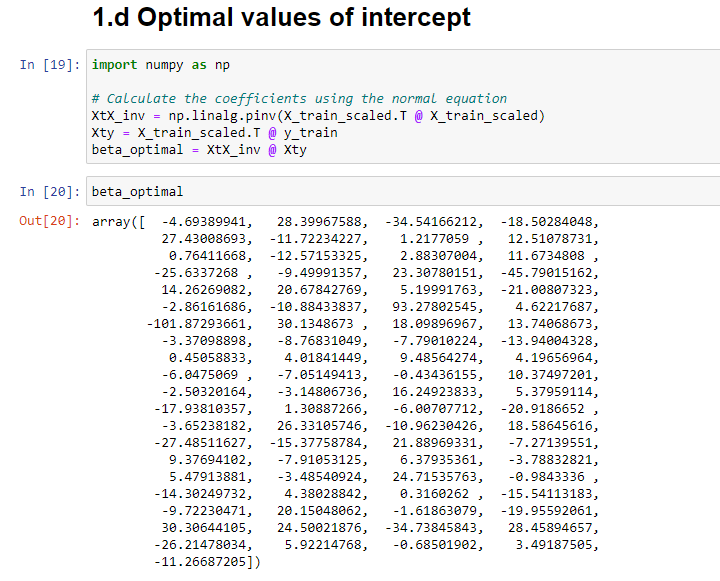
****

****

1. **Standardization technique**

****

1. **Optimal value**

****

1. **Find y predict**

****

1. **R2 and MSE**

****

**References: -**

<https://gocardless.com/en-us/guides/posts/multiple-linear-regression-mlr-definition/#:~:text=Multiple%20linear%20regression%2C%20shortened%20to,variables%20and%20single%20dependent%20variables>.

<https://www.investopedia.com/terms/m/mlr.asp>

<https://online.stat.psu.edu/stat462/node/83/>

<https://www.fireblazeaischool.in/blogs/assumptions-of-linear-regression/>

<https://www.statology.org/multiple-linear-regression-assumptions/>

<https://usq.pressbooks.pub/statisticsforresearchstudents/chapter/multiple-regression-assumptions/>

<https://clas.ucdenver.edu/marcelo-perraillon/sites/default/files/attached-files/week_7_diagnostics_0.pdf>

<http://www.ce.memphis.edu/7012/L15_MultipleLinearRegression_I.pdf>

<https://stats.stackexchange.com/questions/250730/what-is-the-mathematical-relationship-between-r2-and-mse>

<https://vitalflux.com/mean-square-error-r-squared-which-one-to-use/>

<https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/>

<https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e>