GROUP WORK PROJECT # 1
GROUP NUMBER: 5338

MScFE 610: FINANCIAL ECONOMETRICS

FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Shivansh Kumar	India	shivansh.business23@gmail.com	
Vaibhav Sandeep Somani	Ireland	vaibhavssomani@gmail.com	
Kshitiz Sharma	India	drkshitizsharma@gmail.com	

Statement of integrity: By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an “X” above).	
Team member 1	**Shivansh Kumar**
Team member 2	**Vaibhav Sandeep Somani**
Team member 3	**Kshitiz Sharma**

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed. **Note:** You may be required to provide proof of your outreach to non-contributing members upon request.

Data Collection & Processing

For our handbook we are going to use 5 years of daily return data from different assets such as NVIDIA, AMD, META, TESLA, AMAZON, S&P 500, Dow Jones Index, and Treasury yield index 30yrs.
We are using Yahoo Finance API to collect our data. After that, we calculate the daily returns and drop missing data, and properly format it.

Step 2:

The Four challenges that we have picked are:

- **Multicollinearity**
- **Skewness**
- **Sensitivity to outliers**
- **Overfitting**

Throughout this document we are going to address:-

Definition: Technical definition using formulas or equations
Description: Written explanation (1–2 sentences)
Demonstration: Numerical example using real-world data (or simulated data if not found)
Diagram: Visual example using real-world data (using same data as above)
Diagnosis: How to recognize or test that the problem exists
Damage: Clear statement of the damaged caused by the problem
Directions: Suggested models that can address this

For More detailed report please refer to handbook submitted or you can also go to this colab link:

Group_5338_Group Work Project 1 M3.ipynb

Multicollinearity

Definition

Multicollinearity is a phenomenon when two or more variables which are independent are highly correlated to each other. which can cause estimation issues to our regression coefficient and make our model unreliable (Alhassan Umar Ahmad Et. al.).

Let's take an example of a linear regression in which we have "n" independent variable and "c" observation and one dependent variable:

$$Y\_c = \quad \beta 0 + \beta X1 + \beta X2 + ... + \beta cXn + \epsilon\_c$$

where:

- Y is the dependent variable for c observations., X...X_n is the independent variable for c observations, β_0...β_c is the regression coefficient., ϵ_c is the idiosyncratic error observations.

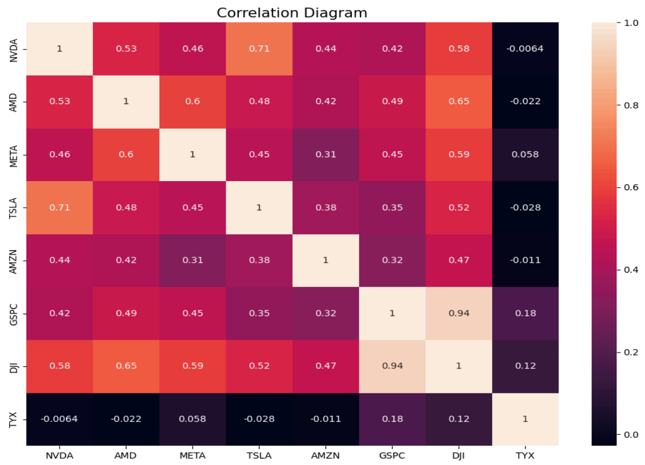
Description

We cannot control how financial data is generated, many times multiple financial variables move together and when two or more variables move together we call it multicollinearity. In our case, we are going to find if there is multicollinearity between NVIDIA, TESLA, META, AMAZON, AMD, S&P500, Dow Jones, and 30yr US treasury daily returns. Due to the current volatile rally in NVDIA, we want to analyze its peers as well as some tech companies with broader market indexes to understand the NVDIA market volatility better to hedge against any risk.

Demonstration & Diagram

Now we will try to determine if there is any multicollinearity between our financial variables. To achieve this we are going to plot a correlation matrix which is the best way to find multicollinearity, A correlation matrix provides us all the correlations of all our variables.

Figure : 1



Diagnosis

Now let's analyze the diagram (Figure 1):

1. We can see in the matrix , NVDA daily returns have a positive correlation with AMD, TSLA and DJI (Dow Jones Index) daily returns, which means these three variables can explain NVDA daily returns pretty well.
2. No we will analyze independent variables, TSLA, AMZN, TYX have very low correaltion with other independent varaibles.
3. After analysis we can see that GSCP and DJI daily retruns , AMD daily returns META and DJI daily returns are highly correlated.

Damage

Now we will try to understand what damage will be done to our model if we have multicollinearity :

- If our independent varibales are highly correlated, the std errors, variances and covariances of the coefficients from the regression may become large and due to that the confidence intervals for coefficient estimates will be wide making our estimates less precise.
- Our estimates are less likely to be significant as we have high std errors. we may get high R^2 and Adjusted R^2 because our model is good. but the collinearity issue makes it harder to isolate the individual impact from the independent varibales in the model.
- As we discussed it is harder to seperate each independent varibles impact in the model, our model can still be used to predict if the new data have same colliner issue as the sample data on which is model is build.

To understand further we will run our model and see how multicollinearity affects our model results.

First we will run a simple test regression model just for comparision purpose in this model we will only use three independent varaible i.e GSCP, DJI, TYX to understant its effect on NVDA daily returns and after that we will run one more model which contains more independent varaibles which are peer of NVDA or from a similar sector. **(You can find model result and outputs in the handbook)**

After comparision between the results of these 2 models to understand the effect of multicollinerity we found that :

- In the above reults summary table we can see that in test_model DJI coefficients are (0.106091) which is significant in the model but in the main model i.e "model" it is just (0.050476) which is not significant at all.
- In the results of the "model" we can see all the p-values are larger than 0.05 which means no independent variable estimates are significant.
- The R^2 in second model is (0.592) which is higher than first model (0.485) this is because we have more independent value and the adjusted R^2 of the first model also higher (0.589) compared to first model's adjusted R^2 (0.484)

One more statstical method that we use is **variance inflation factor(VIF)** which can be used to determine if there is any multicollinerity issue in independent varaible or not (Tay).

After running the VIF statistical method we found that GSPC and DJI have a higher value than 5, which suggests that these variables have multicollinearity issues. **(You can find more results and outputs in the handbook)**

Skewness

Definition:

Skewness refers to the asymmetry of a probability distribution (Sharma). It describes how the data points are distributed around the center of the distribution (Mudholkar and Hutson).

Formula for Skewness:

Fisher - Pearson Standarized Moment Coefficient:

$$\sum_{i=1}^{n} \frac{(X_i - \bar{X})^3/n}{s_X^3}$$

where, X¯ = Mean of X, S = Standard Deviation of X

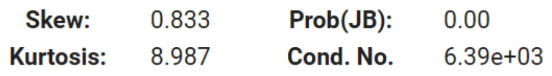
Description:

A distribution can have three types of skewness:

- Right (Positive) Skew: Mean > Median > Mode
- Left (Negative) Skew: Mean < Median < Mode
- Zero Skew: Mean = Median = Mode

Demonstration:

Test Model Summary



Skew:	0.833	**Prob(JB):**	0.00
Kurtosis:	8.987	**Cond. No.**	6.39e+03

From the above model summary, we can notice that the skewness is 0.833, which implies that it is moderately right (positively) skewed.

Shapiro _test

Shapiro W: 0.9232006669044495 , Shapiro p-value 8.000548907699045e-25

As per the Normality Test using the Shapiro Test, the p-value is way less than 0.05, thus we can reject null hypotheses. So, they are not normally distributed

Diagram:

Histogram for Distribution

HistogPlotNormal(residuals, "Normal Disribution - Residuals from OLS Model")

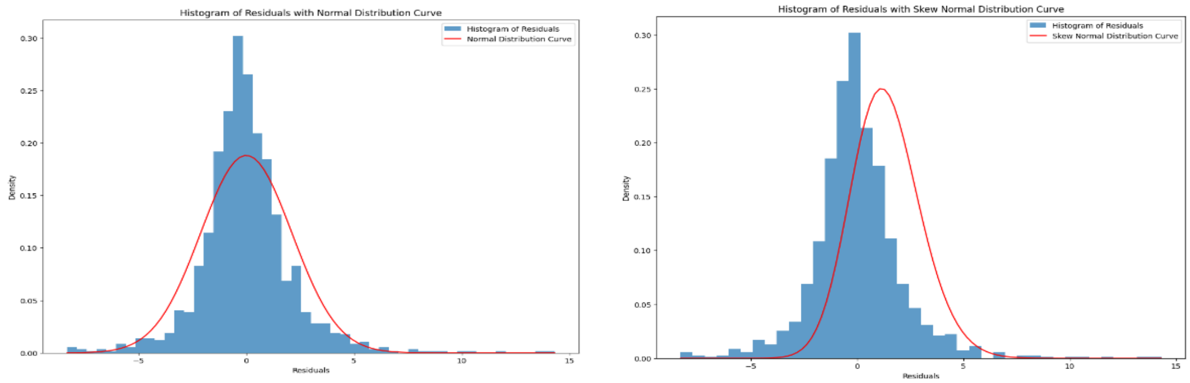


Figure : 2

From the above graph, we can see that the tail on the right side is longer, indicating that it is right skewed.

Diagnosis:

There are various ways to check if Skewness exists in the data:

1. **Visual Inspection** - Plotting histograms, plots to check if it is skewed by analyzing the plot and checking for asymmetry

2. **Using the statistical packages** -

For example: As used above,

The summary gives the skewness coefficient which can be used to analyze.

If skewness > 0, right skewed

skewness < 0, left skewed

skewness = 0, no skewness

Other ways includes checking the Normality:

1. **Q -Q Plot Analysis**

2. **Shapiro Test for Normality on the residuals:** If p- value is greater than 0.05, we can reject the Null hypothesis of normality, if not, it is normally distributed.

This indicates if the distribution is normally distributed, if not, we can conclude that it is asymmetrical.

Damage:

1. **Impact on statistical models and tests:** Generally, the models and tests assume the data distribution to be normal, skewness violates these assumptions, thus, impacting the results. To address skewness, we need to use skew-normal, skew-t distribution.

2. **Outliers is one of the reasons behind the skewness.** Thus, the data distribution is distorted due to presence of skewness, making the analysis challenging (Wikipedia Contributors)

Sensitivity to outliers

Definition

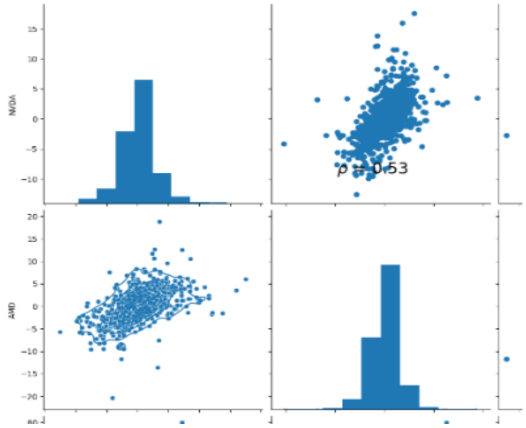
Some data points are extreme that do not confirm to the distribution pattern of majority of the data, are called outliers. In the measurement of centrality, mode and median are not sensitive to outliers but mean is highly sensitive and statistical measures based on mean score are thus highly sensitive too (“Sensitivity to Outliers”). Ignoring outliers can lead to significant incorrect estimation (Chambers et al.). Outliers that can influence the model performance are influential points.

Description

An outlier shows the deviation from the general trend of the other points. In that case the regression model will not make correct estimation. In OLS (Orinary Least Square) outliers will pull the regression line closer to outlier.

We will examine the outlier in scatter plot

Figure 3: Histogram, Correlation, and Scatter Plot Graph Matrix For Independent Variables and Dependent Variable



This Figure provides a summary of the relationships of variables in NVDA stock analysis. The matrix is split into three parts.

As you can see in this figure the scatter plot shows a matrix for two-way combinations of all variables in the model. the graphs diagonally in the matrix are histograms of all the variables, And the upper right triangle of the matrix is the correlation values of two-way combinations of all variables.

The scatter plot NVDA and AMD, NVDA and TSLA, NVDA and DJI is high correlated (>.5). The data points also make oval shape more or less. These signs suggest these variables are good choices for modelling NVDA. NVDA is negatively correlated with TYX.

Demonstration

Our goal here is to identify influnetial points in the dataset and understand why are they extreme (a special event or erroneous data input) as they can darastically alter the regression result. For this We are going to use Cook's distance (Cook's D) test to identify influential points in the dataset. it is a method to calculate the prediction difference from the model with a data point and without a data point. It helps us visualize the Cook's D among all data points. On the horizontal axis of the plot is hat value. On the vertical axis of the plot is studentized residuals.After plotting the method indicated the size of the Cook's D for each point. The bigger the bubble, the higher the Cook's D for a data point.

From the influence plot, we can see that Point 1249 has the highest Cook's D values. Among the points, Point 1249 has a Cook's D of .256777 which is less than 1, so there is no influential point in either the test model or complete model. **(You can find more results and outputs in the handbook).**

Diagram

As you can see in Figure 4 we have plotted scatter plot diagram and an influence plot diagram of our variables now we will to find the heteroskedasticity issue.

For deep analysis we will plot scatter plot for OLS Fitted Value and OLS Residuals to check the presence of heteroskedasticity issue

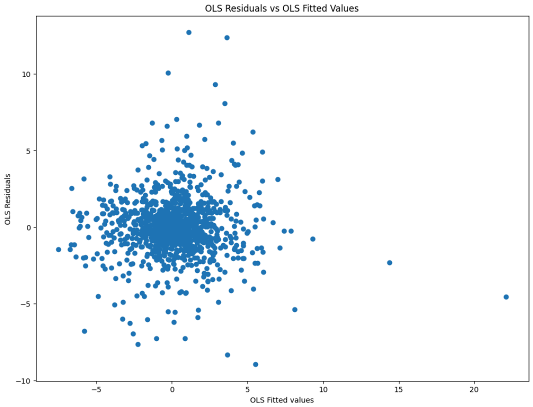
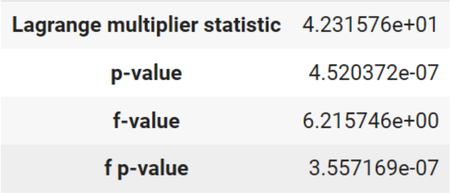


Figure : 4

Diagnosis

Figure 5 shows that there is more variation in residuals when fitted values are positive than when fitted values are negative. The scatter plot indicates there might be a heteroskedasticity issue. We run a Breusch-Pagan test to double-check the hypothesis.

The test result from Figure 4 also confirms the existence of heteroskedasticity because the p-value is less than 0.05. Based on the above information, we will run the model with **weighted least square**, with the weight generated.



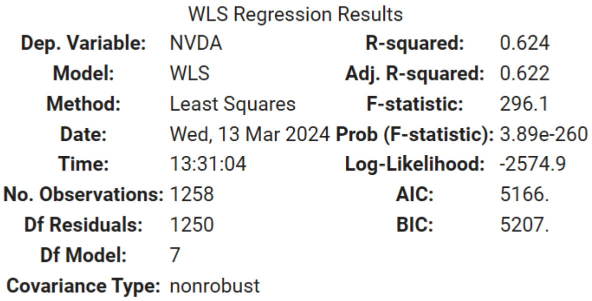
Lagrange multiplier statistic	4.231576e+01
p-value	4.520372e-07
f-value	6.215746e+00
f p-value	3.557169e-07

Figure 5 : Breusch-Pagan Test

Damage

We will analyse the damage by the presence of extreme events. The first model is built by "Team member 1 Multicollinerity : Step 2".

Figure 6 : WLS regression result



WLS Regression Results

Dep. Variable:	NVDA	**R-squared:**	0.624
Model:	WLS	**Adj. R-squared:**	0.622
Method:	Least Squares	**F-statistic:**	296.1
Date:	Wed, 13 Mar 2024	**Prob (F-statistic):**	3.89e-260
Time:	13:31:04	**Log-Likelihood:**	-2574.9
No. Observations:	1258	**AIC:**	5166.
Df Residuals:	1250	**BIC:**	5207.
Df Model:	7		
Covariance Type:	nonrobust		

In the above results in we can see that we have 2 models one is OLS which is a Ordinary Least Square regression with 7 independent variables and all are broader market assets which are peer of NVDA or from a similar sector and a WLS which is Weighted Least Square model.

Now we will see the comparison between the results of these 2 models to understand the effect of heteroskedasticity which explains outlier events :

- In the results of OLS and WLS we can see that the p-values of 'AMD', 'META' and 'TYX' are larger than 0.05 which means these independent variables estimates are not significant.
- The R^2 in WLS model is (0.624) which is higher than OLS model (0.592) this is because we have more independent value and the adjusted R^2 of the OLS model also higher (0.622) compared to OLS model's adjusted R^2 (0.589)

Step 3 :

Directions

Direction for multicollinearity:

As we discussed the damage due to multicollinearity, In this section we will talk about the directions that we should take to counter these issues:

Direction 1 : We can drop independent variables which have high correlation, the thumb rule is to select a correlation between 0.8 and 0.9. we can drop any one of the correlated value and see if the model shows any improvement. but this approach has some issues as it only considers the relationship between two variables at one time, so if we have multicollinearity with more than two variables there will be an issue using this method.

Note: In our case, we have high multicollinearity between only 2 variables so we can we this approach in our case.

Direction 2 : The next approach we can take is to define one of the independent variables in the model as dependent variable and run a regression with the other independent variables. we can try to find if any of the independent variables can be used to explain another independent variable.

Direction 3 : At last our third approach which is one of the most effective approaches is pricipal component analysis.

Direction for skewness:

To address the mentioned damage caused by Skewness, we can do the following:

Direction 1 : Skew-Normal or Skew - t Distributions Model : We can use these distributions to handle the skewness effectively, without any transformation.

Direction 2: Log transformation: Transform skewed distribution to a normal distribution. It compresses the tail and makes the distribution symmetrical.

Direction 3: Remove outliers: Outliers that are irrelavant to the analysis can be removed.

Direction 4: Normalize: By using Mix-Max Scaling, scaling data to 0 to 1 range.

Direction 5: Box Cox transformation (Lai)

Directions for Sensitivity to outliers

Handling outliers in regression analysis is crucial to ensure robust and accurate model performance.

In our model we did not identify specific problem of outliers as Cooks's Distance is less than 1 in each case, so outliers are not present. However outliers are always required to be handled effectively and directions are:

Direction 1: We can exclude outliers from the dataset if outliers are not important data points in the study.

Direction 2: Rather than excluding outliers, we consider using robust regression methods which is less sensitive to outliers.

Ways to address heteroscedasticity:

Direction 3: Transform the Dependent Variable: One approach is to transform the dependent variable. Common transformations include taking the logarithm of the dependent variable. This can help stabilize the variance and make it more consistent across different levels of the predictor variables.

Direction 4: Weighted Least Squares Regression (WLS): WLS assigns different weights to each observation based on their variance. By giving more weight to observations with lower variance, WLS accounts for heteroscedasticity.

Step 4 :

Overfitting

Definition

The tendency to fit training data too closely to training set makes it complex where model captures noises and idiosyncrasies. Such a model works well with training data but fails on testing data because of overfitting. Such a model memorizes training data and loses focus from underlying patterns and loses its ability to generalize.

In finance, this problem arises where model (e.g. Stock market prediction model) works well with historical data but is unable to provide accuracy on data in future market conditions (Melanie).

Description

While working with financial data we usually have a lot of variables, which makes our data high dimensional when on that high dimensional data we run regression and the model predict the training data accurately but fails on testing or new data.

This ability to predict model correctly i.e not only the relationship between endogenous and exogenous variables but also idiosyncratic random error is termed as **overfitting**.

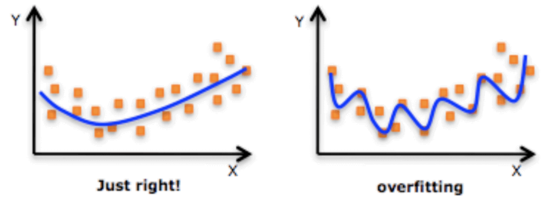


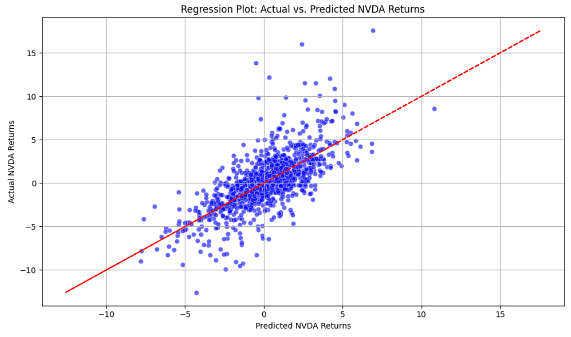
Figure : 7 (Source: Quora)

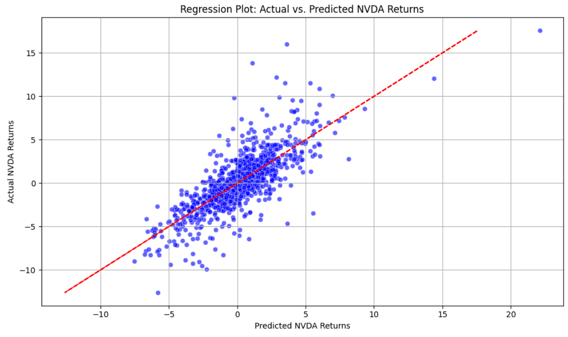
Here in the image we can see that there are two models with same data points, the left model is simple model with one variable but on the right we can we more than one variable which makes the model complicated and the models also captures noise and as we can make prediction of the data accurately hence there is overfitting in right model.

Demonstration & Diagram:

Now let's see the real life example of our data that we are using which is NVDIA, TESLA, META, AMAZON, AMD, S&P500, Dow jones, 30yr US tresury daily returns.

First we will use model build by (Team member 1 Multicollinearity : Step 2), we will plot the regression plot for that model :



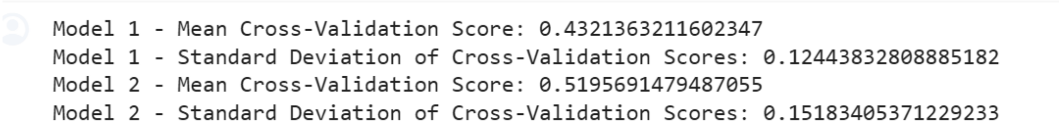


Here in the above figures we can see a simple regression plot from our test_model model an we can use model regression plot written by (Team member 3 : Sensitivity to outliers: step 2), Here we can see in these above diagram for the second figure that this regression is fitting the data more accurately, to verify this we need to do diagnosis which we will do in diagnosis section and if there is any overfitting in our data we will discuss the damage done due to overfitting as well as direction how to further improve it.

Diagnosis

To diagnose the issue of overfitting we will use cross-validation technique. we will perform cross validation on both the models and asses thir performance on test data. we will compare the mean cross-validation score as well as standard deviation of cross-validation scores and see if there is any overfitting issue:

Output mean and standard deviation of cross-validation scores for models :



```
Model 1 - Mean Cross-Validation Score: 0.4321363211602347
Model 1 - Standard Deviation of Cross-Validation Scores: 0.12443832808885182
Model 2 - Mean Cross-Validation Score: 0.5195691479487055
Model 2 - Standard Deviation of Cross-Validation Scores: 0.15183405371229233
```

Here we can see Mean and std dev. of cross-validation scores are higher in model_2 compared to model_1, which concludes that there is overfitting issue in the model_2. Now in the next section we will understand what damages it can do to our model.

Damage

The major damage that overfitting do to a model is that, it gives the model a perfect fit on the training data and when the model is used on the testing dataset the model prediction of the endogenous variable from testing data perform very poor.

The overfitted model try to model all the possible data in the training data set from the good, the bad and the ugly which not only includes signal but also idiosyncratic error. And the result of all this is that the model becomes garbage.

Directions

Now that we have understood what is overfitting, how is this an issue, why does it happen, what damages it can do, now we will discuss the further directions in this section on how to mitigate this issue of overfitting:

Direction 1: Regularization Techniques

- Penalized Regression: We can use Penalized Regression techniques like Ridge Regression, LASSO (Least Absolute Shrinkage and Selection Operator) Regression.

Direction 2: Simplify the model

- As we see model_1 has three input variables which are independent to each other in contrast to model_2 which has seven input variables (leads to multicollinearity) the standard deviation of the cross-validation score is less in model_1. The efficiency of simpler model is higher.

These regression techniques introduce penalty functions in the regression model which removes overfitting by shrinking the regression coefficients towards zero, it the value is too large. the penalty functions try to reduce the impact of an independent variable in the model if the coefficient of the exogenous variables is too large.

References

- Chambers, Ray, et al. "Outlier robust small area estimation." Journal of the Royal Statistical Society Series B: Statistical Methodology 76.1 (2014): 47-69.
- Et. al., Alhassan Umar Ahmad,. “A Study of Multicollinearity Detection and Rectification under Missing Values.” Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 1S, 11 Apr. 2021, pp. 399–418, https://doi.org/10.17762/turcomat.v12i1s.1880. Accessed 7 Oct. 2021.
- Lai, Cheryl. “Study Notes: Handling Skewed Data for Machine Learning Models.” Medium, 4 Nov. 2020, reinec.medium.com/my-notes-handling-skewed-data-5984de303725.
- Melanie. “Overfitting: What Is It? How Can I Avoid It?” Data Science Courses | DataScientest, 23 Sept. 2023, datascientest.com/en/overfitting-what-is-it-how-can-i-avoid-it. Accessed 13 Mar. 2024.
- "Team member 1 Multicollinearity: Step 2", OLS Model: Group_5338_Group Work Project 1 M3
- "Team member 3 Sensitivity to outliers : Step 2", OLS Model: Group_5338_Group Work Project 1 M3
- Mudholkar, Govind S., and Alan D. Hutson. “The Epsilon–Skew–Normal Distribution for Analyzing Near-Normal Data.” Journal of Statistical Planning and Inference, vol. 83, no. 2, Feb. 2000, pp. 291–309, https://doi.org/10.1016/s0378-3758(99)00096-8. Accessed 28 Apr. 2021.
- “Sensitivity to Outliers.” Cloud.sowiso.nl, cloud.sowiso.nl/courses/theory/116/1746/26513/en. Accessed 10 Mar. 2024.
- Sharma, Abhishek. “What Is Skewness in Statistics? | Statistics for Data Science.” Analytics Vidhya, 5 July 2020, www.analyticsvidhya.com/blog/2020/07/what-is-skewness-statistics/.
- Tay, Richard. "Correlation, variance inflation and multicollinearity in regression model." Journal of the Eastern Asia Society for Transportation Studies 12 (2017): 2006-2015.
- Wikipedia Contributors. “Skewness.” Wikipedia, Wikimedia Foundation, 22 Nov. 2019, en.wikipedia.org/wiki/Skewness.