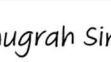
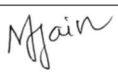
# Multivariate Regression Analysis for Forecasting Nvidia (NVDA) Stock Price

December 2023


Prepared by: Group 1

| Group Member Name | Signature |
|---|---|
| Anugrah Singh |  |
| Sagar Marathe |  |
| Atheesh Krishnan |  |
| Snigdha Jadhav |  |
| Mehul Jain |  |


*Note - We as a group have neither given nor received help (apart from the instructor) to complete this assignment.*

# Abstract

This paper delves into the dynamic and intricate world of stock price analysis and prediction, with a specific focus on NVIDIA Corporation (NVDA), a prominent entity in the technology sector. Since its initial public offering (IPO) on January 22, 1999, at $12.00 per share, NVIDIA has exhibited remarkable growth in its stock price. As of the current trading period, NVIDIA's shares are valued at approximately $452.05, with a substantial trading volume of about 44 million shares daily.

The core objective of this paper is to evaluate the efficacy of multivariable regression in modeling and predicting NVIDIA's stock price. Multivariable regression, a statistical technique, is utilized to determine the relationship between a dependent variable (in this case, NVIDIA's stock price) and multiple independent variables. By applying multivariable regression analysis, we aim to provide a comprehensive understanding of the variables that significantly impact NVIDIA's stock price.

After running the original regression model, using the results of this model the results for joint and individual significance for the individual variable. A new fitted model was created with the significant variables. Gauss Markov assumption were then analyzed for this fitted model and recommendations were provide in case of violations of the model.

# Contents

# 1. Introduction

In the realm of financial markets, the prediction of stock prices has perennially captivated the interest of investors, analysts, and researchers alike. The inherent challenge lies in the complexity of the task, as stock prices are influenced by a myriad of factors ranging from macroeconomic indicators to company-specific news. This paper delves into the intricate task of predicting stock prices, focusing on NVIDIA Corporation, a titan in the technology sector. Our choice to scrutinize NVIDIA's stock is not arbitrary; it is grounded in the company's pivotal role and influence within rapidly evolving industries.

Established in 1993, NVIDIA Corporation has emerged as a formidable force in the global technology market. Renowned for its groundbreaking contributions to manufacturing graphics processors, mobile technologies, and desktop computers, NVIDIA's ascendancy in the high-end GPU market, especially in computer gaming, and more crucially in Artificial Intelligence, has been particularly noteworthy.

Our analysis of NVIDIA's stock is predicated on its historical performance, characterized by significant growth interspersed with notable volatility. This fluctuation in stock value is attributable to a confluence of factors, including evolving market trends, the company’s overall performance, technological advancements, and changes in the global economic landscape. NVIDIA's proactive involvement in sectors such as artificial intelligence, gaming, and high-performance computing has further contributed to the dynamic nature of its stock and its exponential growth in recent years. Additionally, the stock reacts to broader market movements and is sensitive to specific corporate events, such as new product launches, financial earnings reports, and strategic initiatives, all of which have been quite strong for Nvidia.

By exploring the various factors influencing NVIDIA’s stock in a highly dynamic and influential market sector, this project endeavors to contribute valuable perspectives to the field of financial analysis.

## 2. Objective

The aim of the paper is to model and analyze the relation of the price of the stock of Nvidia with various factors influencing the price at Nvidia trades at exchanges (ticker: NVDA)

The paper is structured into several key phases, each vital for the comprehensive econometric analysis:

1. **Stock Price Analysis and Prediction:** To delve into the dynamic field of stock price analysis, with a specific focus on NVIDIA Corporation, evaluating its growth since the IPO and current market performance.
2. **Multivariate Regression Application:** Utilize multivariate regression analysis to model NVIDIA's stock price, identifying and quantifying the impact of multiple influencing factors.
3. **Data Collection and Optimization:** Systematically gather, clean, and optimize relevant data, ensuring its suitability for robust econometric analysis.
4. **Model Efficacy Evaluation:** Assess the efficacy of the regression model in accurately predicting NVIDIA's stock price, underlining its practical applicability in financial forecasting.
5. **Assumption Validation and Correction:** Conduct thorough residual analysis and checks for Gauss-Markov assumptions, including linearity, independence, homoskedasticity, and multicollinearity, and apply appropriate corrective measures for any deviations.
6. **Insightful Interpretation:** Provide insightful interpretations of the results, highlighting the key variables that significantly influence NVIDIA's stock price and their implications for stakeholders in the technology sector.

These objectives aim to encapsulate the comprehensive approach of the paper, balancing theoretical analysis with practical implications in the field of stock price prediction and econometrics.

## 3. Variable Definition

The variables are selected keeping in mind the necessary factors that can influence the price of the stock.

a. Fundamental parameters:

Share Price: The dependent variable “y” represents Nvidia's share price at a given date.

Revenue Growth: Revenue growth refers to an increase in revenue over a quarter. Represented by “x1”.

Gross Margin %: Refers to the percentage of a company's revenue that it keeps after subtracting direct expenses. Represented by “x2”.

Operating Margin: Refers to the percentage of revenue a company generates that can be used to pay the company's investors and taxes. Represented by “x3”.

Net Income Margin %: Refers to the percentage of net income generated from a company's revenue. Represented by “x4”.

Return on Equity %: Measure of a company's net income divided by its shareholders' equity. Represented by “x5”.

Inventory Turnover: Indicates how efficiently a company uses its inventory by dividing the cost of goods sold by the average inventory value during the period. Represented by “x6”.

Total Debt/Equity: Indicates the company's financial leverage and is calculated by the value of total debt and financial liabilities against the total shareholder's equity. Represented by “x7”.

EV/EBITDA: Financial ratio that measures a company's return on investment. Represented by “x8”.

Basic EPS: Calculated as a company's profit divided by the outstanding shares of its common stock. Represented by “x9”.

P/NTM EPS: Calculated by taking the current stock price and dividing it by the earnings per share (EPS) forecasted for the next 12 months. Represented by “x10”.

Dividend: Categorical data indicating whether dividend was given in that period or not. Represented by “x11”.

Research & Development Expenditure: Refers to the direct expenditures relating to a company's efforts to develop, design, and enhance its products, services, technologies, or processes. Represented by “x12”.

Capital Expenditure: Refers to the capital used by a company to acquire or upgrade fixed assets. Represented by “x13”.

Free Cash Flow: Refers to the amount cash available to the firm before accounting for its financial obligations. Represented by “x14”.

Volume: Refers to the Volume of NVDA stock trades at that date. Represented by “x15”.

Market Capitalization: Measurement of a company's size. Refers to the total value of the company’s outstanding shares. Represented by “x16”.

b. Other parameters:

Semi-conductor growth: Refers to the growth rate of semiconductors across the period. Represented by "x17".

VIX: Refers to the CBOE Volatility Index, a popular measure of the stock market's expectation of volatility based on S&P 500 index options. Represented by "x18".

Nasdaq 100 Close: Refers to the closing price of NASDAQ 100 at that date. Represented by "x19".

AMD Share Price: Refers to AMD's share price at a given date. Represented by "x20".

World Inflation: Refers to the rate of increase in prices across the world over a given period. Represented by "x21".

## 4. Regression Model

The model is structured around the principles of multivariate regression, utilizing a dataset with a sample size of 95. This dataset includes 21 dependent variables, which are crucial in shaping the model's analysis and predictions. The primary response or dependent variable under consideration is the "Share Price," which serves as a pivotal metric in the financial analysis domain. Share Price refers to the current market price of a single share of a company's stock, encapsulating the market's valuation of a share at a given point in time. This metric is a critical indicator for investors, analysts, and traders, as it reflects the collective market perception of a company's value and potential. The model integrates the Share Price with 21 explanatory or independent variables, facilitating a robust and multifaceted analysis of stock market trends and patterns.

The multivariate regression model developed is as follows:

$$y = \beta 0 + \sum_{i=1}^{n} \beta_i x_i \, , where \; n = 21$$

## 5. Model Results:

Upon conducting the regression analysis of the model using RStudio, a comprehensive summary of the model's performance and statistical insights has been generated. This summary provides detailed information regarding the efficacy and characteristics of the model, based on the regression analysis conducted. The summary includes key statistical metrics and interpretations that are crucial for understanding the model's accuracy, reliability, and areas for potential improvement.

```
Call:
lm(formula = y ~ . - y, data = nvda)

Residuals:
     Min       1Q   Median       3Q      Max
-1.26597 -0.29269  0.00697  0.24347  1.12006

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.943e+00  8.978e-01   4.392 3.73e-05 ***
x1           1.041e+00  4.901e-01   2.125  0.03701 *
x2          -4.259e+00  1.631e+00  -2.611  0.01096 *
x3          -3.356e+02  5.256e+02  -0.638  0.52515
x4           3.342e+02  5.257e+02   0.636  0.52690
x5           4.430e+00  8.567e-01   5.171 1.96e-06 ***
x6          -1.969e-01  6.257e-02  -3.147  0.00239 **
x7           6.190e-01  3.828e-01   1.617  0.11023
x8           4.827e-03  3.027e-03   1.595  0.11506
x9          -7.264e-01  4.039e-01  -1.799  0.07623 .
x10         -4.264e-03  7.346e-03  -0.580  0.56341
x11          3.863e-01  3.429e-01   1.126  0.26370
x12          1.407e-03  5.965e-04   2.359  0.02098 *
x13          3.341e-04  1.139e-03   0.293  0.77009
x14          6.346e-05  2.690e-04   0.236  0.81415
x15         -3.024e-10  1.351e-10  -2.238  0.02827 *
x16          4.027e-04  1.378e-06 292.281  < 2e-16 ***
x17         -4.040e-02  1.855e-02  -2.178  0.03265 *
x18         -2.856e-03  8.846e-03  -0.323  0.74769
x19         -1.412e-04  8.838e-05  -1.597  0.11451
x20         -1.929e-02  7.085e-03  -2.722  0.00811 **
x21         -3.485e-01  2.252e-01  -1.548  0.12604
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5071 on 73 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 1.638e+05 on 21 and 73 DF,  p-value: < 2.2e-16
```

Figure 1 – Regression output

## 6. Observations:

1. The model is as follows:

$$y = 3.943 + 1.041 \cdot x1 - 4.259 \cdot x2 - 3.356e^{2} \cdot x3 + 3.342e^{2} \cdot x4$$
$$+ 4.430 \cdot x5 - 1.969e^{-1} \cdot x6 + 6.190e^{-1} \cdot x7 + 4.827e^{-3} \cdot x8 - 7.264e^{-1} \cdot x9$$
$$- 4.264e^{-3} \cdot x10 + 3.863e^{-1} \cdot x11 + 1.407e^{-3} \cdot x12 + 3.341e^{-4} \cdot x13 +$$
$$6.346e^{-5} \cdot x14 - 3.024e^{-10} \cdot x15 + 4.027e^{-4} \cdot x16 - 4.040e^{-2} \cdot x17 -$$
$$2.856e^{-3} \cdot x18 - 1.412e^{-4} \cdot x19 - 1.929e^{-2} \cdot x20 - 3.485e^{-1} \cdot x21$$

2. R squared ($R^2$) typically has a value between 0 through 1. In our model, $R^2$ has a value of 1, indicating that predictions are identical to observed values.

3. Hypothesis testing of the predicted model:

i) Testing for Joint Significance:

$H_0$: $\beta_1 = \beta_2 = \ldots\ldots = \beta_{21}$

$H_1$: Not all $\beta_i$ are the same for i = 1,2,…21 (At least one value of $\beta_i \neq 0$)

The regression model gives a very small p-value of $2.2e^{-16}$. At significance level $\alpha = 0.05$, we reject $H_0$. There is enough evidence that atleast one independent variable $x_i$ where i = 1,2…..,21 is significant in the model.

ii) Since $H_0$ is rejected during the joint testing, we will test for the significance of individual variables:

$H_0$: $\beta_i = 0$

$H_1$: $\beta_i$ = 1,2,…21 (At least one value of $\beta_i \neq 0$)

The variables having p-values less than the significance level $\alpha = 0.05$, we accept the null hypothesis and make that variable significant.

Therefore, for independent variables x1, x2, x5, x6, x12, x15, x16, x17 and x20, we accept the hypothesis of significance level $\alpha = 0.05$ as the p-value is less than 0.05. We say that these variables are significant at 95% confidence level.

**Fitted Model:**

The new model was developed using only significant independent variables at 95% confidence level, which were x1, x2, x5, x6, x12, x15, x16, x17 and x20.

The model is as follows:

$$y = 3.465 + 1.343 \cdot x1 - 4.145 \cdot x2 + 3.196 \cdot x5 - 2.026e^{-1} \cdot x6 + 1.220e^{-3} \cdot x12 - 3.001e^{-10} \cdot x15 + 4.016e^{-4} \cdot x16 - 4.746\, e^{-2} \cdot x17 - 2.575e^{-2} \cdot x20$$

Regression output for the fitted model is as follows:

```
Call:
lm(formula = y ~ x1 + x2 + x5 + x6 + x12 + x15 + x16 + x17 +
    x20, data = nvda)

Residuals:
     Min       1Q   Median       3Q      Max
-1.69444 -0.32213 -0.03255  0.27767  1.30341

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.465e+00  6.539e-01   5.299 9.02e-07 ***
x1           1.343e+00  4.516e-01   2.974  0.00383 **
x2          -4.145e+00  9.203e-01  -4.504 2.11e-05 ***
x5           3.196e+00  4.647e-01   6.878 9.65e-10 ***
x6          -2.026e-01  6.336e-02  -3.198  0.00194 **
x12          1.220e-03  3.739e-04   3.264  0.00158 **
x15         -3.001e-10  1.212e-10  -2.477  0.01523 *
x16          4.016e-04  8.069e-07 497.736  < 2e-16 ***
x17         -4.746e-02  1.705e-02  -2.783  0.00663 **
x20         -2.575e-02  4.083e-03  -6.307 1.23e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5439 on 85 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 3.323e+05 on 9 and 85 DF,  p-value: < 2.2e-16
```

Figure 2 – Regression output for the fitted model

**Observations:**

Based on the regression output of the original model, we find that only the following independent variables are significant:

| | |
|---|---|
| ***Revenue Growth*** | ***Gross Margin*** |
| ***Return on Equity*** | ***Inventory Turnover*** |
| ***R&D Expenditure*** | ***Volume*** |
| ***Market Capitalization*** | ***Semiconductor Industry Growth*** |
| ***AMD Share Price*** | |

We further observe that the $R^2$ is very close to 1, indicating that the observed values are almost identical to the predictors and that the model is a good fit.

## 7. Gauss Markov Assumptions

In order to make a valid inference from our Multivariate regression model, we check for Gauss-Markov assumptions.

Analysis of Gauss-Markov Assumptions for the fitted model:

**I] Linearity in terms of parameters (β1, β2,…..β21):**

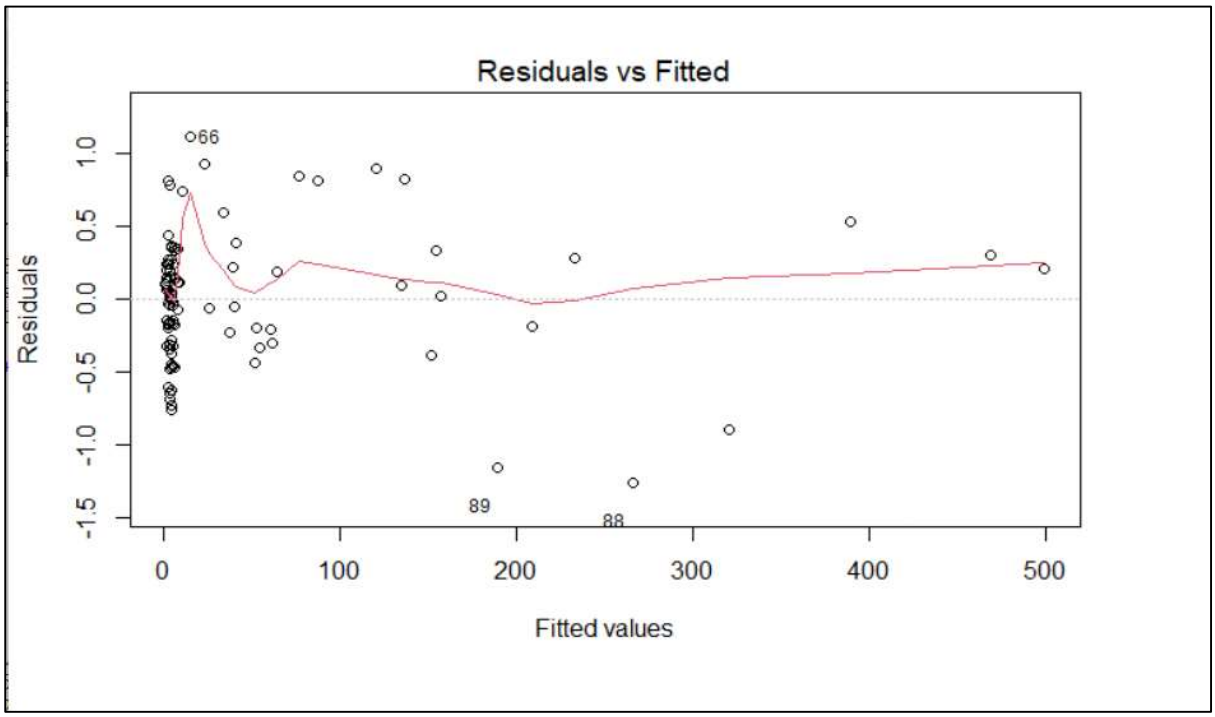
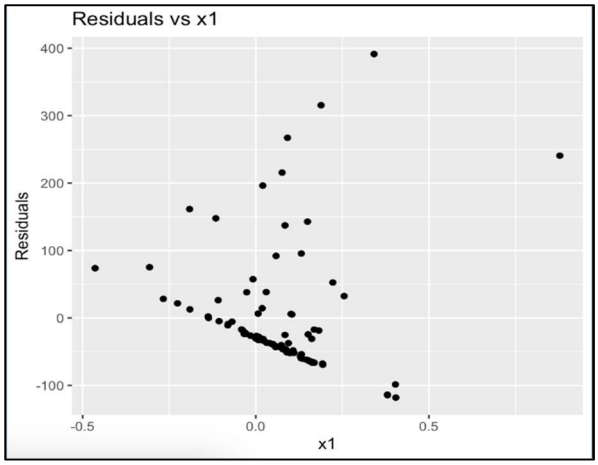
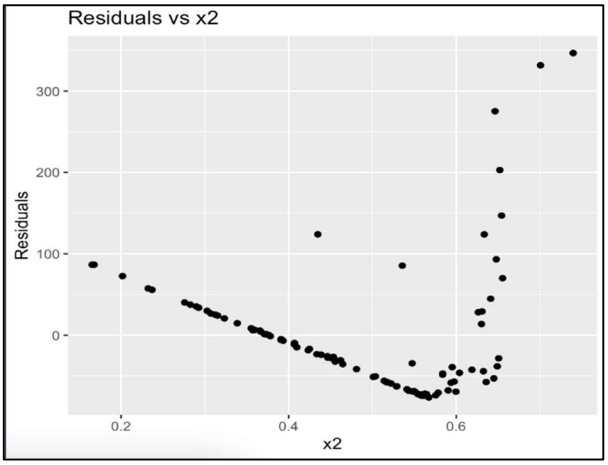


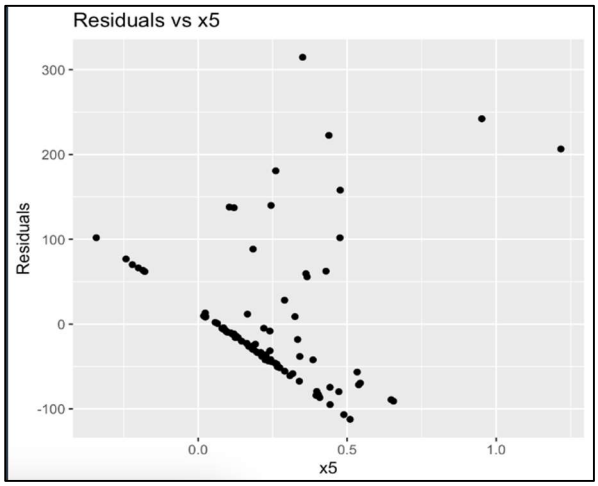

Figure 3 – Residuals vs Fitted Graph

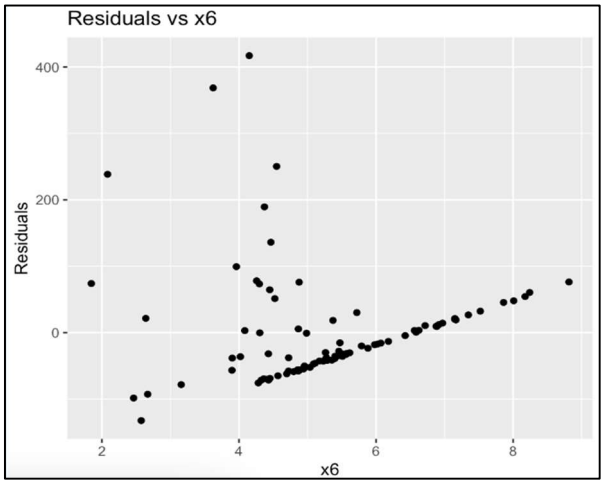
**Inference:** As observed from Figure 3, the red line, which represents the average trend of the residuals, should ideally be flat and close to zero across all levels of fitted values if the linearity assumption holds. However, in this graph, the red line shows a clear non-linear pattern, especially at the lower end of the fitted values, suggesting that the linearity assumption may be violated.

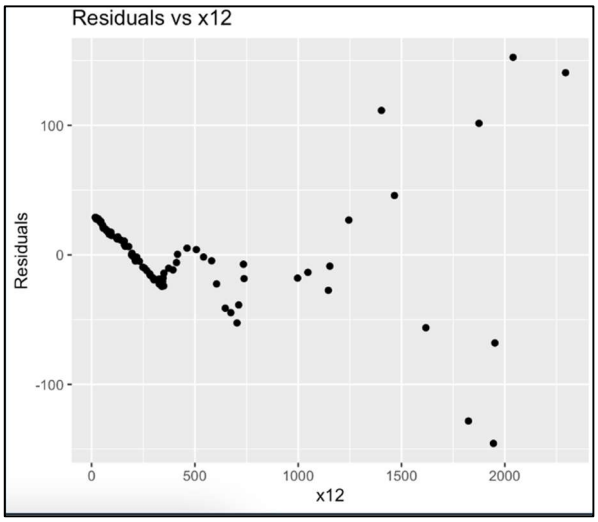
**II] Independence and Randomness of error terms.**

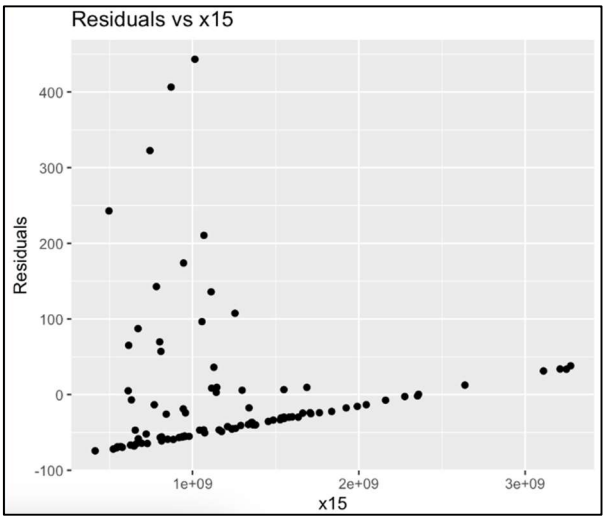


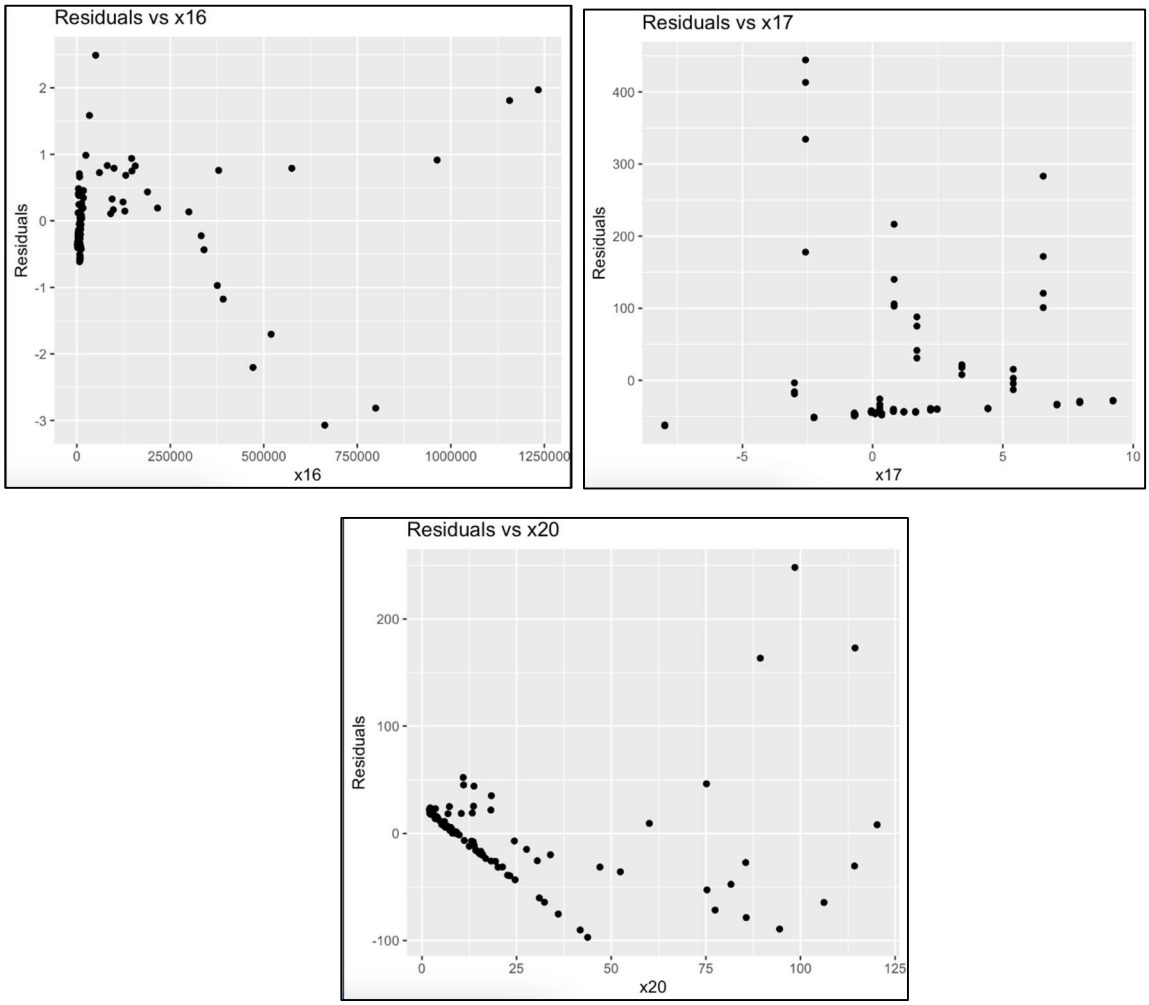



Figure 4 – Residuals plots for independent variables.

**Inference:** As observed from Figure 4, the residual plots show some form of pattern or structure between independent variables and residuals, which may be non-linear. Thus, we cannot assume that the Independence and randomness assumption holds true. Therefore, we conduct statistical analysis using the Durbin-Watson Test to test the assumption further.

**Durbin-Watson Test**

Further, we perform the Durbin-Watson (DW) test to analyze the randomness of independent variables.

Hypothesis for D-W test:

$H_0$: First-order autocorrelation does not exist (i.e., they are independent)

$H_1$: First-order autocorrelation exists (i.e., they are dependent)

```
> durbinWatsonTest(fitted_model, alternative = "positive")
 lag Autocorrelation D-W Statistic p-value
   1       0.7000329     0.2696282       0
 Alternative hypothesis: rho > 0
```

**Inference:** As observed from above, the p-value is significantly small. Therefore, we reject the null hypothesis ($H_0$). This implies that a first-order autocorrelation exists among residuals i.e., they are dependent. Hence, the assumption is violated.

The DW Statistic always assumes a value between 0 and 4. A value of DW = 2 indicates that there is no autocorrelation. When the value is below 2, it indicates a positive autocorrelation and a value higher than 2 indicates a negative serial correlation.

Further, the DW Statistic is compared to the lower (DL) and upper critical values (DU) to test positive autocorrelation.

For our fitted model, considering 95 observations and 9 independent variables, the $D_L$ and $D_U$ are as follows:

$D_L = 1.48861$ and $D_U = 1.85164$

Since the DW Statistic < $D_L$, there is statistical evidence that the data is positively autocorrelated.

**III] Zero Conditional Mean**

As we use residuals to estimate errors, the Zero Conditional Mean assumption is considered as holding for any dataset, and therefore, this assumption is not violated.

$$\sum_{i=1}^{n} \hat{\epsilon} = 0, \qquad \text{so } \hat{\epsilon} = 0$$

**IV] Homoskedasticity of Error Terms**

As can be observed from the residual plots from Figure 4, the assumption of Homoskedasticity of Error Terms is violated. This implies that the variance of errors differs across observations.

Therefore, we perform statistical analysis of Heteroskedasticity. We can broadly distinguish heteroskedasticity into two kinds: Conditional and unconditional.

Unconditional Heteroskedasticity occurs when the heteroskedasticity of error variance is not correlated with the independent variable in the multivariate regression. This does not create significant problems for statistical inference.

Conditional Heteroskedasticity occurs when the heteroskedasticity of error variance is correlated with the value of independent variables in the multivariate regression. Further, conditional heteroskedasticity can be of linear or quadratic type. We could test for linear conditional heteroskedasticity through the Breusch Pagan Test and quadratic conditional heteroskedasticity through the White Test.

We perform the Breusch Pagan (BP) Test to test for linear conditional heteroskedasticity.

Hypothesis testing for BP test:

$H_0$: there is no linear conditional heteroskedasticity in the model.

$H_1$: there is linear conditional heteroskedasticity in the model.

```
> bptest(fitted_model)

        studentized Breusch-Pagan test

data:  fitted_model
BP = 29.148, df = 9, p-value = 0.0006118
```

**Inference:** Since the p-value < significance level $\alpha = 0.05$, we reject the null hypothesis. There is enough evidence that linear conditional heteroskedasticity exists.

**V] Multicollinearity – No exact linear relationship between independent variables.**

Multicollinearity occurs when two or more independent variables (or a combination of independent variables) are highly (but not perfectly) correlated with each other. The magnitude of pairwise correlation among the independent variables to assess multicollinearity is generally inadequate. Although a very high pairwise correlation between the independent variables can indicate multicollinearity, it is not a necessary condition for multicollinearity, and a low pairwise correlation among the independent variables does not mean that multicollinearity is not a problem.

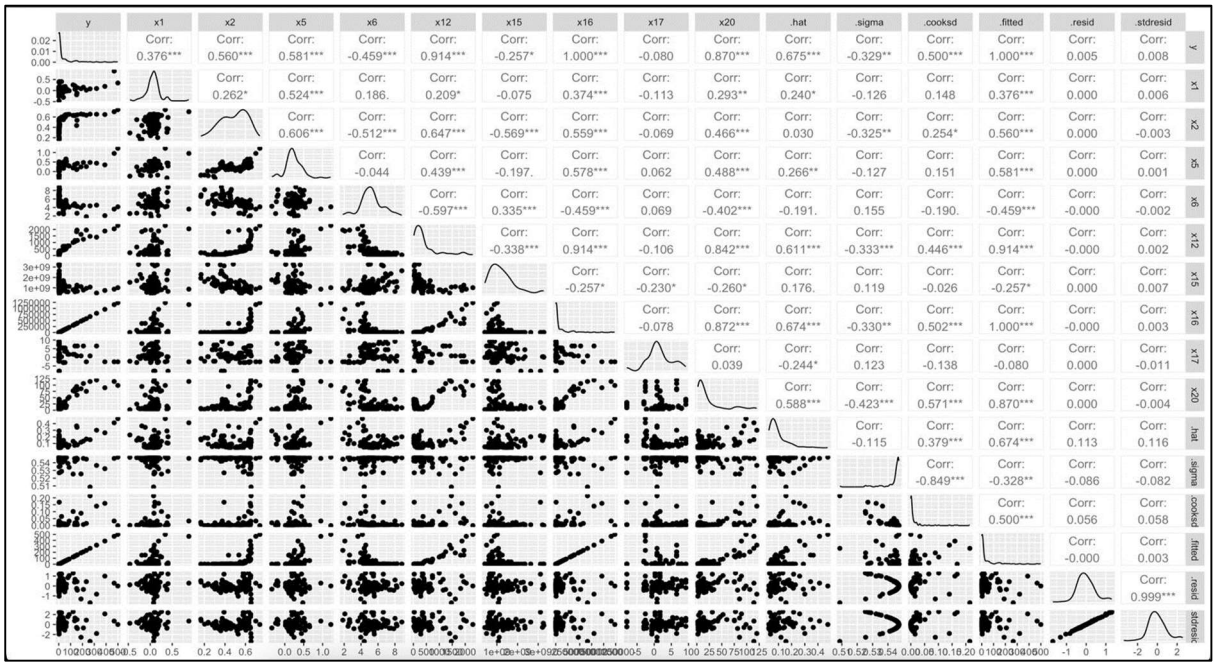



Figure 5 -ggpairs graph

To comprehensively assess multicollinearity, we examine the Variance Inflation Factor (VIF) to consider the combined effect of independent variables.

The general range of VIF output is:

VIF = 1: No correlation between the independent variable and the other variables.

VIF between 1 and 5: Typically, this is considered a moderate correlation but not severe enough to require attention.

VIF greater than 5: This often indicates that the variable has a significant multicollinearity with other independent variables.

VIF greater than 10: This value suggests a very high correlation and is a clear sign that the variable is involved in significant multicollinearity.

```
> vif(fitted_model)
      x1       x2       x5       x6      x12      x15      x16      x17      x20
1.739506 4.664135 3.439568 2.304449 12.203960 1.834597 12.000578 1.299386 4.915701
```

Inference: We see that for factors x12 and x16 have a VIF > 5. This could imply significant multicollinearity with each other. Hence, the multicollinearity assumption is violated.

**Inference:** Through the VIF command, we can see that the variables are multicollinear in nature. Hence the assumption is violated. Therefore, we reject $H_0$.

**VI] Normality of error terms**

The residual errors are normally distributed with mean = 0 and variance = σ^2.

To determine normality, we examine the histogram of the residuals; a bell-shaped curve suggests that the residuals approximate a normal distribution, with a mean of zero and constant variance.

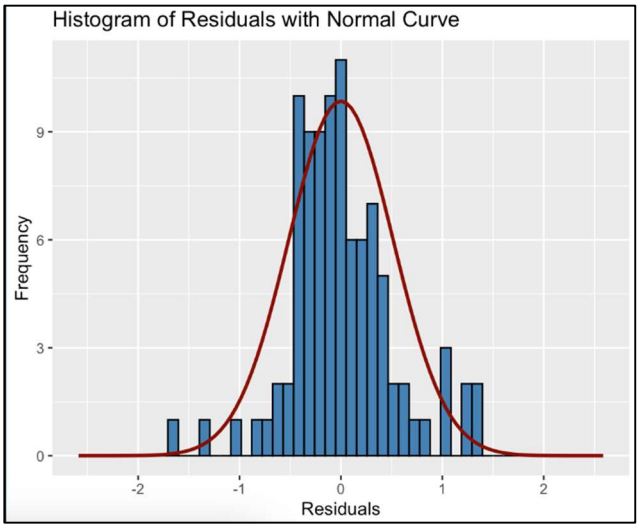



Figure 6 – Histogram graph of residuals

**Inference:** As can be seen from Figure 6, the residuals are roughly normally distributed in the histogram, with a few outliers; hence we can say that the null hypothesis is not violated, and the assumption holds.

# 8. Recommendations for violation of assumptions:

1. ASSUMPTION I: Parameters have a linear relationship.

   In addressing the violations of the linearity assumption, we recommend the following steps:

   A. Variable Transformation: Apply transformations like log, square root, or reciprocal to the variables to correct non-linear relationships [1]

   B. Model Expansion: Reassess the model to ensure all relevant variables are included, thus addressing any omitted variable bias [2]

   C. Non-linear Models: If transformations do not suffice, employ non-linear modeling techniques such as generalized additive models (GAMs) for greater flexibility [3]

2. ASSUMPTION II: Independence and Randomness of error terms.

   We recommend the following remediations for the violation of independence and randomness of error terms assumption:

   A. Time Series Adjustment: For time series data, consider using models that account for autocorrelation, such as ARIMA [4], or adding lagged variables to the model.

   B. Randomization: Ensure that the data collection process is randomized, if possible, to minimize systematic error and promote independence [5]

   C. Panel Data Techniques: For panel data, use fixed effects or random effects models to control for unobserved heterogeneity between groups or over time [6]

3. ASSUMPTION IV: Homoskedasticity of Error Terms.

   We recommend the following methods to address heteroskedasticity:

   A. The first method is computing robust standard errors (also known as White's standard errors), which corrects the standard error of linear regression model's estimated coefficients to account for conditional heteroskedasticity.

   B. The second method is the Generalized Least Squares method, which modifies the original equation to eliminate heteroskedasticity.

   C. Weighted Least Squares (WLS): Instead of ordinary least squares (OLS), using WLS can give different weights to different observations that inversely relate to the variance of errors.

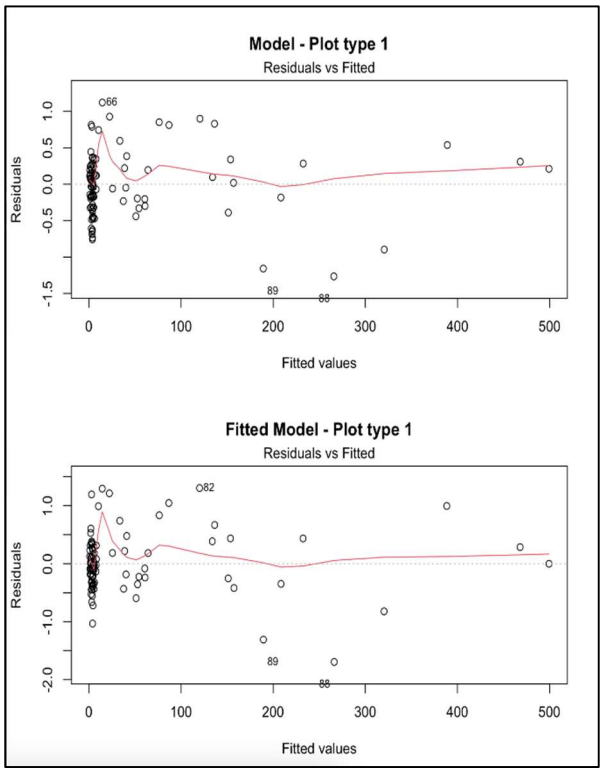
4. ASSUMPTION V: Multicollinearity:

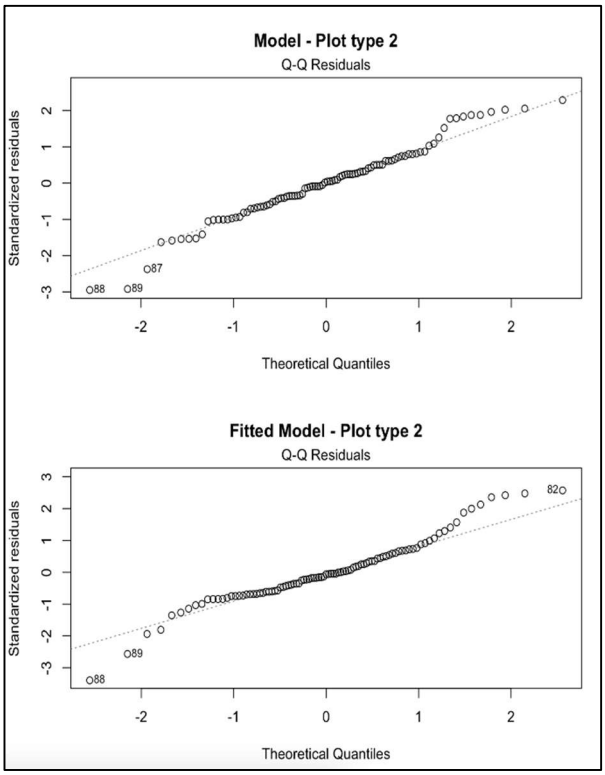
   A. Choosing not to address multicollinearity may be appropriate when the model's aim is prediction, not inference, and its predictive accuracy remains unaffected.

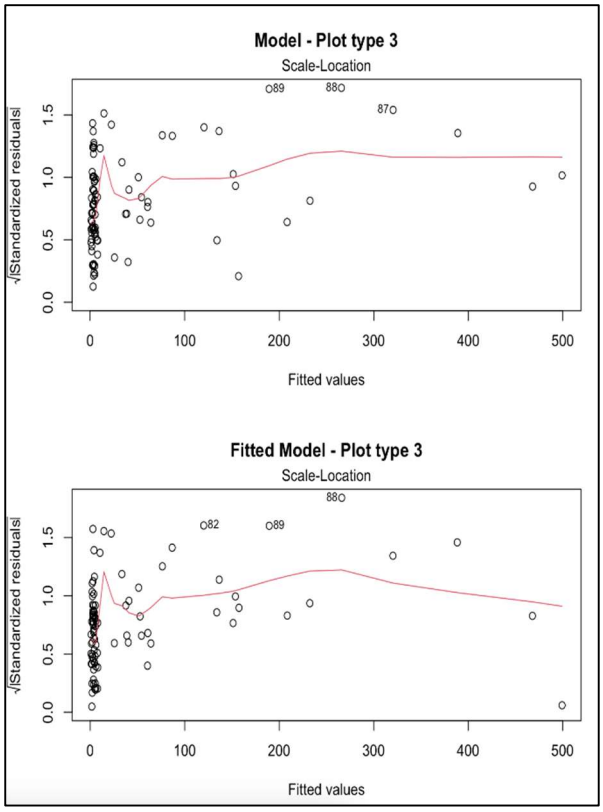
   B. Eliminate Redundant Variables: Removing a redundant variable to correct a specification error in the model.

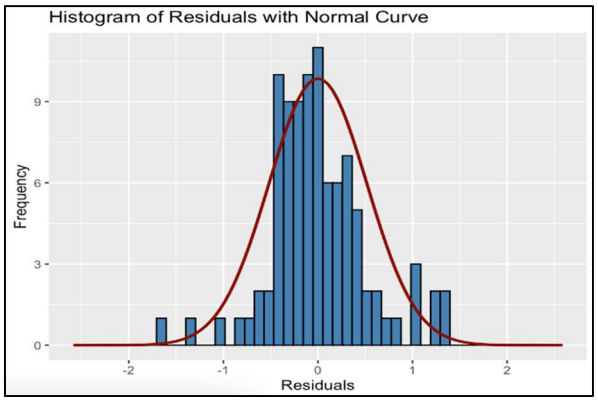
   C. Implement sophisticated regression models such as Lasso, Ridge, or Elastic Net regression, which uses qualities from both Lasso and Ridge regression.

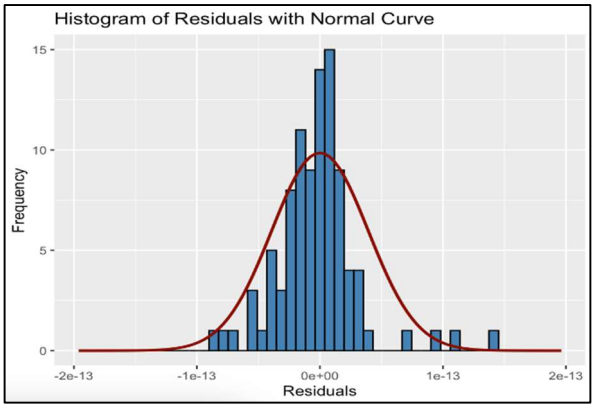
# 9. Original v/s Fitted Model Plots



## 10. Results:

The following is the multivariate regression equation for the fitted model:

$$\text{Share Price} = 3.465 + 1.343 \cdot \text{Revenue Growth} - 4.145 \cdot \text{Gross Margin} + 3.196 \cdot \text{Return on Equity \%} - 2.026e^{-1} \cdot \text{Inventory Turnover} + 1.220e^{-3} \cdot \text{R\&D Expenditure} - 3.001e^{-10} \cdot \text{Volume} + 4.016e^{-4} \cdot \text{Market Capitalization} - 4.746\, e^{-2} \cdot \text{Semi-conductor growth} - 2.575e^{-2} \cdot \text{AMD Share Price}.$$

The paper delved into the intricacies of econometrics, emphasizing the practical application of statistical concepts and the utilization of multivariate regression for model fitting.

The project encapsulated key learnings in econometric analysis, starting with thorough data collection and cleaning, highlighting the importance of a solid dataset foundation. We did this by compiling our data from multiple disparate sources. This was followed by selecting the significant factors based on the regression output of the original model. Finally, constructing and analyzing a multivariate regression model demonstrated the application of regression techniques in understanding complex relationships between stock prices and various market variables. These steps collectively underscored the depth and intricacy involved in econometric analysis, from data preparation to insightful model interpretation.

The findings culminate in a model where the identified significant variables effectively predict Nvidia's stock price. This paper aims to comprehensively understand statistical modeling and highlight the importance of assumption testing and correction in developing a reliable econometric model.

## 11. References and Citations

[1] Box, G. E. P., & Cox, D. R. (1964). Journal of the Royal Statistical Society. Series B (Methodological).

[2] Hastie, T., & Tibshirani, R. (1990). Generalized Additive Models. Chapman & Hall/CRC.

[3] Wooldridge, J. M. (2015). Introductory econometrics: A modern approach. Cengage Learning.

[4] Box, G. E. P., & Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control. Holden-Day.

[5] Fisher, R. A. (1935). The Design of Experiments. Edinburgh: Oliver and Boyd.

[6] Baltagi, B. H. (2005). Econometric Analysis of Panel Data (3rd ed.). John Wiley & Sons.

[7] https://r-coder.com/plot-r/

[8] https://reintech.io/term/using-ggplot2-for-data-visualization-in-r

[9] https://search.r-project.org/CRAN/refmans/car/html/vif.html

[10] https://godatadrive.com/blog/basic-guide-to-test-assumptions-of-linear-regression-in-r

[11] https://www.statisticshowto.com/gauss-markov-theorem-assumptions/

[12] https://economictheoryblog.com/2015/04/01/ols_assumptions/

[13] https://stats.stackexchange.com/questions/133625/violation-of-gauss-markov-assumptions

[14] https://nathanasmooha.com/wp-content/uploads/2014/02/Set-2-GLS.pdf

[15] https://finance.yahoo.com/quote/NVDA/history/

[16] https://www.capitaliq.com/ciqdotnet/Login-okta.aspx

[17] https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG

[18] https://www.statista.com/

[19] CFA Institute Investment Series – Quantitative Investment Analysis [4th Edition]

[20] https://investor.nvidia.com/home/default.aspx

# 12. Appendix

R-Code

```
# Importing required libraries
library("readxl")
library(memisc)
library(dplyr)
library(lmtest)
library(sjPlot)
library(sgof)
library(ggplot2)
library(foreign)
library(car)
library(hexbin)
library(lmtest)
library(GGally)

# Declaring working directory for the project
setwd("~/AK0/RBS/Academic/SEM 1/Econometrics")

# Read data from compiled source excel file
nvda <- read_excel("Final Variables (2).xlsx")
is.data.frame(nvda)
nvda = subset(nvda, select = -c(Date))

# ====================================
# IMPUTATION
# Imputing 0 values for blanks wherever required
for (column_name in names(nvda)) {

# Check if there are any NAs in the column
 if (any(is.na(nvda[[column_name]]))) {
   # for total debt/equity, impute with 0
   if (column_name == "Total Debt/Equity") {
     nvda[[column_name]][is.na(nvda[[column_name]])] <- 0
   }
 }
```

```
}

# ===================================
# Modifying column names for simplicity
colnames(nvda) <- c("y", "x1", "x2", "x3","x4","x5","x6", "x7", "x8",
                    "x9", "x10", "x11", "x12", "x13", "x14","x15","x16","x17",
                    "x18","x19","x20", "x21")
# create scatter plots for pairs of variables - checks gauss-markov assumption 2
ggpairs(nvda)
# preparing multivariate model
# creating OLS model with 21 independent variables
model <- lm(y ~ .-y, data = nvda)
# estimators of the parameters - beta
coef(model)
# predicted values of y we obtain from the model, as opposed to observed. (y_i^)
fitted(model)
# diff between obs and fitted y values
residuals(model)
# calculating ssr
deviance(model)
summary(model)
ggpairs(model)
coeftest(model)
plot_model(model)

# new model with only significant variables
fitted_model <- lm(y ~ x1+x2+x5+x6+x12+x15+x16+x17+x20, data=nvda)
summary(fitted_model)
ggpairs(fitted_model)
compare_models <- mtable(model, fitted_model)
compare_models
# ===================
# PRELIMINARY PLOTS
# ===================
# ggplot2 scatter plot between DEPENDENT VAR (y) and DIFFERENT X VARS

# Define the x variables (significant variables)
x_vars <- c("x1", "x2", "x5", "x6", "x12", "x15", "x16", "x17", "x20")
```

```
# Loop through each x variable and create a plot
for (x_var in x_vars) {
 p <- ggplot(nvda, aes_string(x = x_var, y = "y")) +
      geom_point() +
      ggtitle(paste("Scatter plot of y vs", x_var)) +
      xlab(x_var) +
      ylab("y")
 print(p)
}

# ========================
# INDIVIDUAL RESIDUAL PLOTS
x_vars <- c("x1", "x2", "x5", "x6", "x12", "x15", "x16", "x17", "x20")

# Loop through each x variable, fit a model, and create a diagnostic plot
for (x_var in x_vars) {
 # Fit the linear model
 model_formula <- as.formula(paste("y ~", x_var))
 indiv_fitted_model <- lm(model_formula, data = nvda)
  # Create a diagnostic plot (residual plot)
 plot(indiv_fitted_model, which = 1, col = "darkblue", main = paste("Residual Plot for model y
~", x_var))
}
#=========================
# INDIVIDUAL RESIDUAL VS X VARIABLE PLOTS
x_vars <- c("x1", "x2", "x5", "x6", "x12", "x15", "x16", "x17", "x20")
# Loop through each x variable, fit a model, and create a residual vs x variable plot
for (x_var in x_vars) {
 # Fit the linear model
 model_formula <- as.formula(paste("y ~", x_var))
 indiv_fitted_model <- lm(model_formula, data = nvda)
  # Extract residuals
 residuals <- indiv_fitted_model$residuals
  # Create a scatter plot of residuals vs x variable
 p <- ggplot(nvda, aes_string(x = x_var, y = residuals)) +
   geom_point() +
```

```
   ggtitle(paste("Residuals vs", x_var)) +
   xlab(x_var) +
   ylab("Residuals")
 print(p)
}
#====================
model <- lm(y ~ .-y, data = nvda)
fitted_model <- lm(y ~ x1+x2+x5+x6+x12+x15+x16+x17+x20, data=nvda)
summary(model)
summary(fitted_model)
#================================
#GAUSS MARKOV ASSUMPTIONS
#================================
#1.LINEARITY
#Residual vs Fitted plot would not have a pattern where the red line is approximately horizontal
at zero.
plot(model, 1)
#========================
#2.RESIDUAL ERROR - INDEPENDENCE AND RANDOMNESS
#plot residual vs fitted for individual significant variables - code above
#alternatively - dw test below
#Durbin-Watson test
#p-value < 0.05, we would reject the null hypothesis.
durbinWatsonTest(model)
par(mfrow=c(2,2))
plot(model)
durbinWatsonTest(fitted_model, alternative = "positive")
#========================
#3.MULTICOLLINEARITY
vif(model)
#========================
#5.CONSTANT VARIANCE OF RESIDUAL ERRORS/ HOMOSKEDASTICITY OF ERROR TERMS
#scale location plot
plot(fitted_model,3)
#check for heteroscedasticity
bptest(fitted_model)
#================================
```

```
#6.NORMALITY
# QQ PLOT
plot(fitted_model, 2)
# ================================
# HISTOGRAM OF RESIDUALS
residuals <- model$residuals
residuals_df <- data.frame(residuals = residuals)
# Calculate mean and standard deviation of residuals, used to plot the superimposed normal plot
mean_residuals <- mean(residuals)
sd_residuals <- sd(residuals)
# Range for x-axis limits, centered around the mean
x_min <- mean_residuals - (5 * sd_residuals)
x_max <- mean_residuals + (5 * sd_residuals)
# Calculate the number of observations and appropriate bin width
n <- length(residuals)
binwidth <- (x_max - x_min) / 50
# Calculate scaling factor to plot the normal curve
scaling_factor <- 1.3 * n * binwidth
# Create the histogram and superimpose the normal distribution curve
ggplot(residuals_df, aes(x = residuals)) +
 geom_histogram(aes(y = ..count..), fill = 'steelblue', color = 'black', binwidth = binwidth) +
 stat_function(fun = function(x) dnorm(x, mean = mean_residuals, sd = sd_residuals) *
scaling_factor,
               color = "darkred", size = 1) +
 labs(title = 'Histogram of Residuals with Normal Curve', x = 'Residuals', y = 'Frequency') +
 xlim(x_min, x_max)
# ============================
# COMPARISON GRAPHS
# ============================
plot_types <- c(1, 2, 3)
for (plot_type in plot_types) {
 # Plot for the first model
 par(mfrow = c(2, 1))  # Set up the graphics layout for two plots
 plot(model, which = plot_type, main = paste("Model - Plot type", plot_type))
 plot(fitted_model, which = plot_type, main = paste("Fitted Model - Plot type", plot_type))
}
```