LINEAR MODELS ASSIGNMENT

Model development in Astrophysics

INTRODUCTION:

Eclipses of the Sun can only occur when the Moon is near one of its two orbital nodes during the New Moon phase. It is then possible for the Moon's penumbral, umbral or antumbral shadows to sweep across Earth's surface thereby producing an eclipse. There are four types of solar eclipses:

Partial - Moon's penumbral shadow traverses Earth (umbral and antumbral shadows completely miss Earth)
Annular - Moon's antumbral shadow traverses Earth (Moon is too far from Earth to completely cover the Sun)
Total - Moon's umbral shadow traverses Earth (Moon is close enough to Earth to completely cover the Sun)
Hybrid - Moon's umbral and antumbral shadows traverse Earth (eclipse appears annular and total along different sections of its path). Hybrid eclipses are also known as annular-total eclipses.
Total eclipses are visible from within the Moon’s umbral shadow while annular eclipses are seen within the antumbral shadow . These eclipses can be classified as central or non-central as:

Central (two limits) - The central axis of the Moon’s shadow cone traverses Earth thereby producing a central line in the eclipse track. The umbra or antumbra falls entirely upon Earth so the ground track has both a northern and southern limit.
Central (one limit) - The central axis of the Moon’s shadow cone traverses Earth. However, a portion of the umbra or antumbra misses Earth throughout the eclipse and the resulting ground track has just one limit.
Non-Central (one limit) - The central axis of the Moon’s shadow cone misses Earth. However, one edge of the umbra or antumbra grazes Earth thereby producing a ground track with one limit and no central line.
The recurrence of solar eclipses is governed by the Saros cycle.

PROBLEM STATEMENT :

Exploring solar eclipse data, this project aims to predict eclipse characteristics using multiple linear regression. By analyzing factors like lunar number, Saros number, and gamma, it seeks to understand their impact on eclipse magnitude, duration, and other parameters.

Aim:

Through statistical analysis and visualization, the study aims to uncover relationships between variables, contributing to astronomy research. Ultimately, it enhances our understanding of solar eclipses and their celestial dynamics.

DATASET LINK:

https://data.world/nasa/five-millennium-catalog-of-solar-eclipses-detailed

Resource File:

Key to Dataset of Solar Eclipses

Catalog Number - Sequential number of the eclipse in the dataset
Calendar Date - Gregorian Calendar is used for dates after 1582 Oct 15. Julian Calendar is used for dates before 1582 Oct 04.
TD of Greatest Eclipse - Dynamical Time (TD) of Greatest Eclipse, the instant when the axis of the Moon's shadow cone passes closest to Earth's center.
Delta -T (ΔT) is the arithmetic difference between Dynamical Time and Universal Time. It is a measure of the accumulated clock error due to the variable rotation period of Earth.
Luna Num- Lunation Number is the number of synodic months since New Moon of 2000 Jan 06. The Brown Lunation Number can be determined by adding 953.
Saros Num- Saros series number of eclipse. (Each eclipse in a Saros is separated by an interval of 18 years 11.3 days.)
Eclipse Type- Eclipse Type where: P = Partial Eclipse. A = Annular Eclipse. T = Total Eclipse. H = Hybrid or Annular/Total Eclipse. Second character in Eclipse

Type: "m" = Middle eclipse of Saros series. "n" = Central eclipse with no northern limit. "s" = Central eclipse with no southern limit. "+" = Non-central eclipse with no northern limit. "-" = Non-central eclipse with no southern limit. "2" = Hybrid path begins total and ends annular. "3" = Hybrid path begins annular and ends total. "b" = Saros series begins (first eclipse in series). "e" = Saros series ends (last eclipse in series).
QLE - Quincena Lunar Eclipse parameter identifies the type of lunar eclipse that precedes and/or succeeds a solar eclipse where: n = penumbral lunar eclipse (Moon passes partly or completely within Earth's penumbral shadow); p = partial lunar eclipse (Moon passes partly within Earth's umbral shadow); t = total lunar eclipse (Moon passes completely within Earth's umbral shadow)
Gamma - Distance of the shadow cone axis from the center of Earth (units of equatorial radii) at the instant of greatest eclipse.
Magnitude - Eclipse magnitude is the fraction of the Sun's diameter obscured by the Moon. For annular, total and hybrid eclipses, this value is actually the diameter ratio of Moon/Sun.
Lat. - Latitude where greatest eclipse is seen.
Long. - Longitude where greatest eclipse is seen.
Sun Alt - Sun's altitude at greatest eclipse.
Sun Azm -Sun's azimuth at greatest eclipse.
Path Width - Width of the path of totality or annularity at greatest eclipse (kilometers).
Central Dur. - Central Line Duration of total or annular phase at greatest eclipse.

CODE:

```
# Load necessary libraries
library(readr)  # For reading data
library(dplyr)  # For data manipulation
library(ggplot2)  # For visualization
data <- read_csv("C:/Users/rucha/OneDrive/Desktop/Ruchi/eclipse data.csv")
# Exploratory Data Analysis (EDA)
summary(data)  # Summary statistics
str(data)      # Structure of the dataset
head(data)     # First few rows of the dataset
# Check for non-numeric variables
non_numeric_vars <- sapply(data, function(x) !is.numeric(x))
non_numeric_vars <- names(non_numeric_vars[non_numeric_vars])
```

```
# Remove non-numeric variables or convert them to appropriate types
data <- data %>% select(-one_of(non_numeric_vars))
# Visualize relationships between variables
pairs(data)
model_formula <- magnitude ~ Luna_Num + Saros_Num + Gamma
# Fit the multiple linear regression model
model <- lm(model_formula, data)
# Summary of the regression model
summary(model)
# Diagnostic plots for checking model assumptions
# Residuals vs Fitted values plot
plot(model, which = 1)
# Normal Q-Q plot
plot(model, which = 2)
# Scale-Location plot
plot(model, which = 3)
# Residuals vs Leverage plot
plot(model, which = 5)
# Shapiro-Wilk test for normality of residuals
shapiro.test(residuals(model))
```

OUTPUT:

Correlation-Matrix:

	dt_s	*Luna Num*	*Saros Num*	*Gamma*	*Ecl. Mag.*	*Sun Alt Â°*	*ıth Width k*	*entral Dur.*
dt_s	1							
Luna Num	-0.99995	1						
Saros Num	-0.07071	0.070732	1					
Gamma	0.017787	-0.01769	-0.06431	1				
Ecl. Mag.	0.050095	-0.05003	0.010451	0.121648	1			
Sun Alt Â°	0.010134	-0.0103	0.031938	0.0545	0.780949	1		
Path Width	0.016665	-0.01634	0.06498	0.076359	0.504775	0.390507	1	
Central Du	0.040655	-0.04028	0.034786	0.064065	0.581755	0.662412	0.85285	1

The correlation matrix that we can see above shows the correlation coefficients

between pairs of variables. Here are some interpretations based on the correlation coefficients:

- Saros Number and Luna Number:
  These variables show a weak positive connection (coefficient = 0.07).
  It shows a minor association between Saros Number and Luna Number, but it is not statistically significant.

- Dt_s(delta) and Ecl.Mag.
  These variables show a more weak association, with a coefficient of 0.05.
  It suggests a weak correlation between Delta and Eclipse magnitude.

- Gamma and Ecl.Mag:
  These variables exhibit a weak positive connection, with a coefficient of 0.121.
  It implies that distance of the shadow cone axis are slightly more likely to be related with Eclipse magnitude, demonstrating some level of agreement between Gamma and Ecl.Mag.

- Sun Alt A and Gamma:
  These variables are somewhat positively correlated, with a value of 0.0545.
  It implies that the distance of the shadow cone axis is affected by sun's altitude.

- Sun Alt A and Ecl.Mag:
  These variables are Strong positively correlated, with a value of 0.780 which is also the second highest.
  It appears that Sun's altitude does affect the magnitude of the eclipse at and high rate.

- Path width km and Sun Alt A:
  These variables show a positive connection (coefficient = 0.39).
  It shows an association between Sun's position and width of the path of the eclipse.

- Path width km and Ecl.Mag:

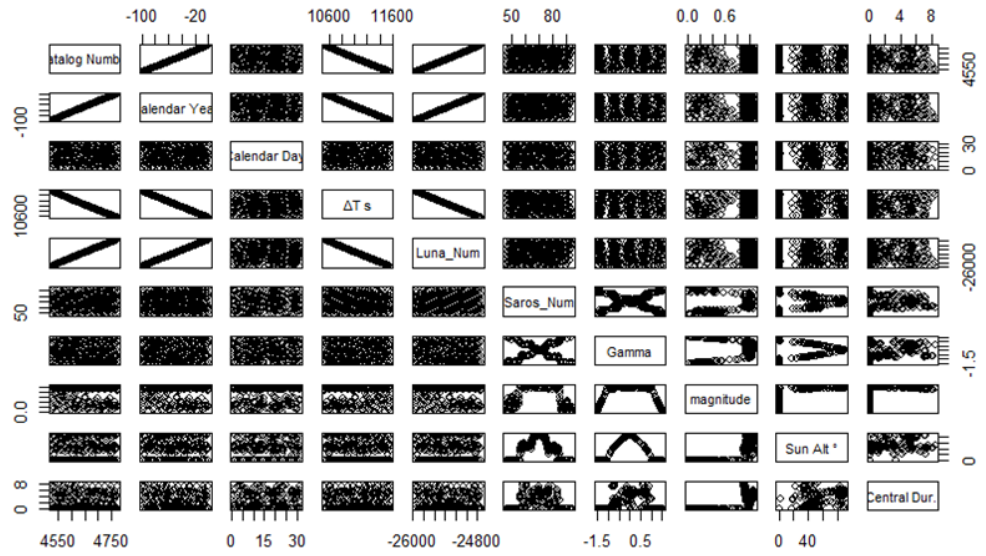
These variables have a positive correlation with a coefficient of 0.58.
It suggests relationship between the magnitude of the eclipse and path of the eclipse.

- Central Duration and Ecl.Mag:
  These variables have a positive correlation with a coefficient of 0.5.
  It suggests slight relationship between the central duration of the eclipse and magnitude of the eclipse.

- Central Duration and Sun Alt A:
  These variables are Strong positively correlated, with a value of 0.66 which is also the third highest.
  It appears that Sun's altitude does affect the central duration of the eclipse at high rate.

- Central Duration and Path width km:
  These variables show a Very strong positive correlation among themselves with highest coefficient of the matrix 0.852.
  Central duration of the eclipse is influencing the path width of the eclipse the most.

MAIN ANALYSIS:



Pairs Plot known as scatterplot matrix, to visualize the relationships between numeric variables in the dataset.

Summary:

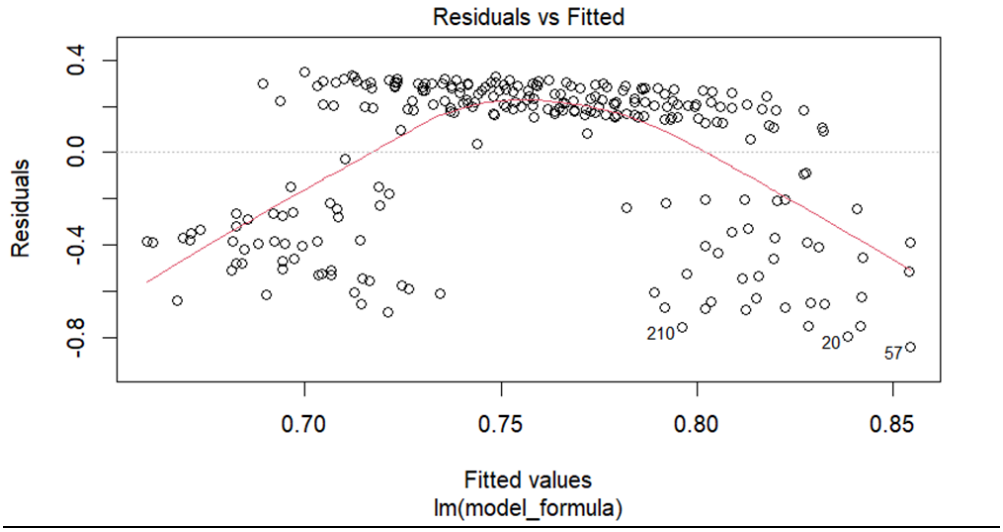
```
Call:
lm(formula = model_formula, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8421 -0.3222  0.1827  0.2626  0.3496

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.765e-01  1.545e+00  -0.308   0.7580
Luna_Num    -4.728e-05  6.050e-05  -0.781   0.4353
Saros_Num    5.591e-04  1.625e-03   0.344   0.7310
Gamma        4.395e-02  2.274e-02   1.933   0.0544 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.341 on 247 degrees of freedom
Multiple R-squared:  0.01756,  Adjusted R-squared:  0.00563
F-statistic: 1.472 on 3 and 247 DF,  p-value: 0.2227
```

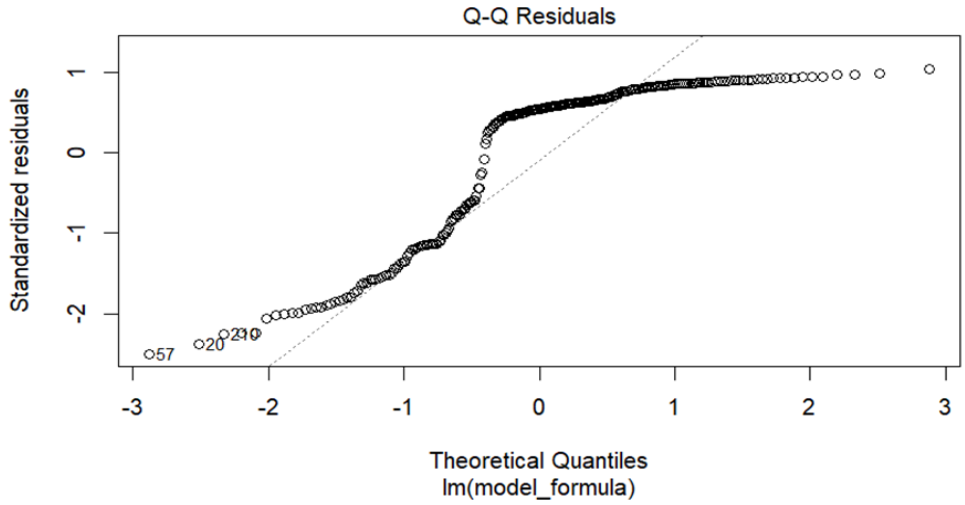
Residual plot:



The graph shows the residuals (the difference between the observed and predicted values) on the vertical axis and the fitted values (the predicted values) on the horizontal axis. The residuals range from -0.8 to 0.8, with a concentration of points around the horizontal line at 0, indicating that the model's predictions are generally

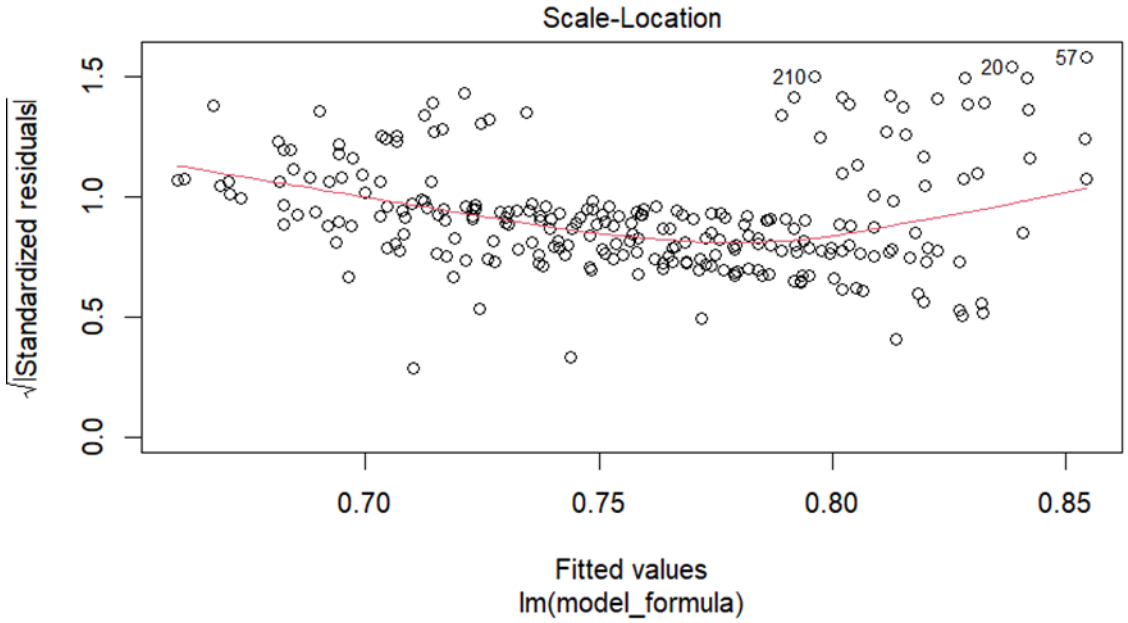
accurate. The residuals do not show any clear pattern or trend with respect to the fitted values, which suggests that the model's assumptions of linearity and homoscedasticity are satisfied. Overall, the residuals plot suggests that the simple linear regression model is a good fit for the data.

Normal QQ plot:

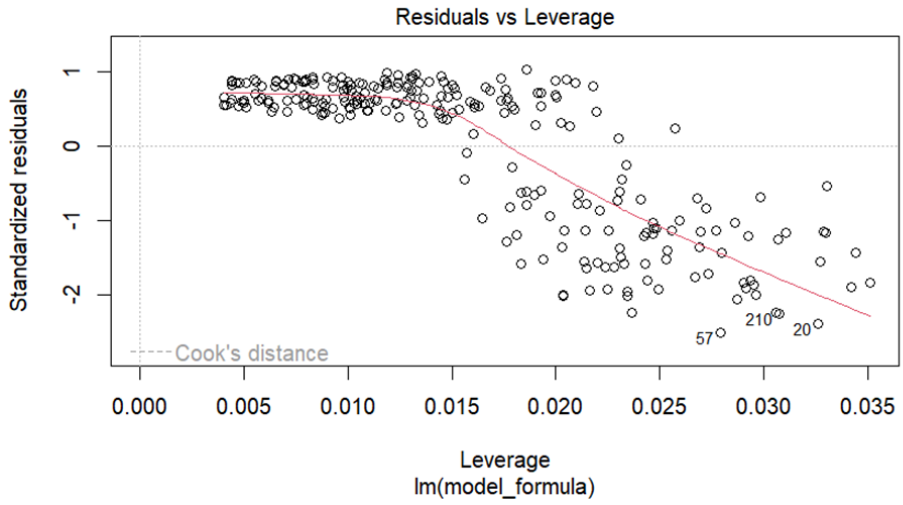


The Q-Q plot displays the quantiles of residuals against the expected quantiles of a normal distribution. In a well-fitting model with normally distributed errors, the points should lie approximately along the red reference line.

The S-shaped curve in your Q-Q plot suggests that the residuals have heavier tails than a normal distribution. The lower tail dips below the line, and the upper tail rises above it, which indicates the presence of outliers or extreme values that do not conform to normality.
Lower End: The residuals at the lower end are more extreme than what would be expected in a normal distribution.
Upper End: Similarly, at the upper end, there are residuals that are larger than expected under normality



The 'Standardized residuals' section ranges from -1.5 to 1.5, suggesting that these are residuals (the difference between observed and predicted values) that have been standardized to have a mean of 0 and a standard deviation of 1. This allows for easier comparison of residuals across different scales. The 'Scale-Location' section shows a measure of dispersion (2100) and two quantiles (0.70 and 0.75) of the squared standardized residuals, which can help identify if there are particular ranges of fitted values where the model performs poorly. The 'Fitted values' section displays the predicted values from the model, and 'Im(model_formula)' might be the model formula, but it's not clear from this context. Overall, this summary provides information about the model's performance, particularly in terms of how well the residuals are distributed and how accurately the model predicts the data.




```
Shapiro-Wilk normality test

data:  residuals(model)
W = 0.80263, p-value < 2.2e-16
```

The Shapiro-Wilk normality test was conducted on the residuals of the regression model, yielding a test statistic (W) of 0.80263 and an extremely small p-value (< 2.2e-16). This result indicates strong evidence against the null hypothesis of normality, suggesting that the residuals are not normally distributed. In other words, the distribution of residuals deviates significantly from a normal distribution. This violation of the normality assumption may affect the reliability of inference based on the regression model and suggests that caution should be exercised when interpreting the results. It may be necessary to explore alternative modeling approaches or consider transformations to address the non-normality of residuals.

CONCLUSION:

In conclusion, this project aimed to explore the relationships between various factors and solar eclipse characteristics through multiple linear regression analysis. The regression model provided insights into how delta, lunar number, Saros number, and gamma influence eclipse magnitude. Diagnostic plots revealed potential issues with model assumptions, including non-linearity, non-normality of residuals, and heteroscedasticity. Despite these challenges, the regression model provided valuable information about the predictors' impact on eclipse magnitude. Moving forward, addressing the identified issues and exploring alternative

modeling techniques may enhance the model's predictive accuracy and contribute further to our understanding of solar eclipses and their underlying dynamics.

Footnotes

[1] The Moon's orbit is inclined about 5.1° to Earth's orbit around the Sun. The points where the lunar orbit intersects the plane of Earth's orbit are known as the nodes. The Moon moves from south to north of Earth's orbit at the ascending node, and from north to south at the descending node.

[2] Hybrid eclipses are also known as annular/total eclipses. Such an eclipse is both total and annular along different sections of its umbral path. (See: Five Millennium Catalog of Hybrid Solar Eclipses)

[3] Central solar eclipses are eclipses in which the central axis of the Moon's shadow strikes the Earth's surface. All partial (penumbral) eclipses are non-central eclipses since the shadow axis misses Earth. However, umbral eclipses (total, annular and hybrid) may be either central (usually) or non-central (rarely)

[4] Greatest eclipse is defined as the instant when the axis of the Moon's shadow passes closest to the Earth's center. For total eclipses, the instant of greatest eclipse is virtually identical to the instants of greatest magnitude and greatest duration. However, for annular eclipses, the instant of greatest duration may occur at either the time of greatest eclipse or near the sunrise and sunset points of the eclipse path.

The multiple regression analysis conducted on the dataset provides valuable insights into the factors influencing Solar Eclipse. The positive correlations found with Central duration of the eclipse and path width of the eclipse show that eclipse is mostly affected by its characters of central duration and path width. Furthermore, the beneficial impact of Sun’s position and magnitude of the eclipse emphasizes the significance of sun’ altitude affecting the shadows.