### **Group #1 - Predictive Analysis of Near-Earth Comets using NASA JPL’s Small-Body Database**

Troy Allen, Roy Gabriel, Nattakorn Kittisut, Corey Smith & **Moe Kyaw Thu (Team Lead)**

**Introduction and Research Questions**:

The National Aeronautics and Space Administration (NASA) tracks numerous astronomical phenomena and celestial bodies to gain insights into the workings of the universe and to predict potentially catastrophic events. In particular, NASA’s Jet Propulsion Laboratory (JPL) focuses substantial effort on tracking, categorizing, and documenting near-Earth asteroids and comets to perform various science analyses and predict possible Earth impact. This study will attempt to identify algorithmic correlations between various scientific parameters related to near-Earth comets (NECs) based on regression techniques learned in this course.

Based on this topic, this study posits the following research questions:

1. *Which orbital elements best predict the minimum distance between the comet and Earth (MOID)? Are all input parameters significant? Is the MOID impacted more by the orbit eccentricity (e) or its critical angles (i, w, Node)?*
2. *Do the minimum (q) and maximum (Q) distances between the comet and the Sun have an impact on the orbital period (P)? Additionally, are these minimum and maximum distance values correlated or do they depend on some or all other orbital elements?*
3. *Is there any relationship between the two response variables, MOID and P? Is there a way to predict the likelihood and/or frequency of a Near-Earth comet sighting? Provide justifications using the results of the regression analysis.*

**Data and Methodology**

In light of the research questions, this study uses the dataset titled “[Near-Earth Comets - Orbital Elements](https://data.nasa.gov/Space-Science/Near-Earth-Comets-Orbital-Elements/b67r-rgxc/about_data)” from NASA’s data repository [Ref. 1-2]. This dataset provides J2000 heliocentric ecliptic orbital elements for NECs. In total, it has 17 attributes (columns) along with 160 observations (See appendix for Codebook). The attributes provided in the dataset include essential orbital elements such as the semi-major axis, eccentricity, inclination, the longitude of the ascending node, the argument of perihelion, mean anomaly, and so on. Collectively, they are all crucial for understanding the trajectories and behavior of these comets as they pass close to the Earth, Sun, and other celestial bodies during their periodic lifespan.

Based on this dataset and the goals of the regression analysis, the methodology in this study involves several key steps. First and foremost, data cleaning and preprocessing, including the use of parsing, are performed to ensure the quality and integrity of the dataset. This includes handling missing values, detecting outliers, standardizing data format, and visualizing initial relationships.

In the second stage following data preprocessing, linear regression models are developed to understand the relationship between the response and the predictor variables to analyze the three identified research questions and serve as a basis for the higher fidelity approaches. Model diagnostics, particularly detecting leverage points, determining normality, and testing for non-constant residual variance are conducted to evaluate the performance and accuracy of the regression models. Additionally, various visualization techniques, particularly the use of Q-Q plots, are utilized to assess the normality of residuals and the overall goodness-of-fit of the models.

The final stage will involve applying inferential analysis techniques to research question I to understand the relationship between the response and the predictor variables. In particular, this approach will focus on the use of stepwise regression and regularization methods, such as LASSO and Ridge regressions. These inferential techniques will help to uncover the patterns and trends within the dataset as well as feature importance and handling overfitting. The research question I is chosen since it includes all the predictor variables to explain the response variable, which may also lead to overfitting issues.

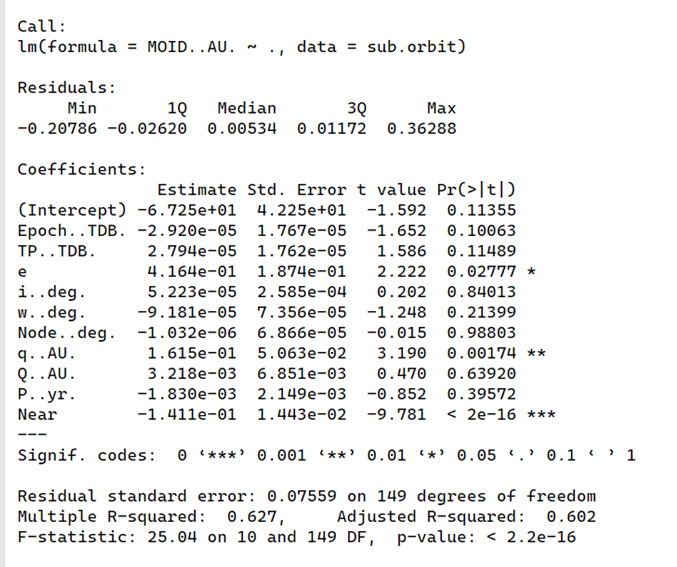
**Data Cleaning and Preprocessing**

To ensure data quality, this study begins by cleaning and preprocessing the dataset. First and foremost, it identifies the missing values and finds that the non-gravitational acceleration parameters (A1, A2, A3) and the time span of the data arc used (DT) have a large number of empty cells within the original dataset. Due to their nature, they are omitted from the finalized dataset since they could cause issues in the analysis and moreover, eliminating rows based on NaN values will also lead to the omission of a large number of observations. Subsequently, an outlier diagnosis is performed to identify any problematic observations which is followed by exploratory data analysis to understand the finalized dataset’s characteristics. Lastly, all the remaining columns are then standardized to prepare for regression modeling requirements.

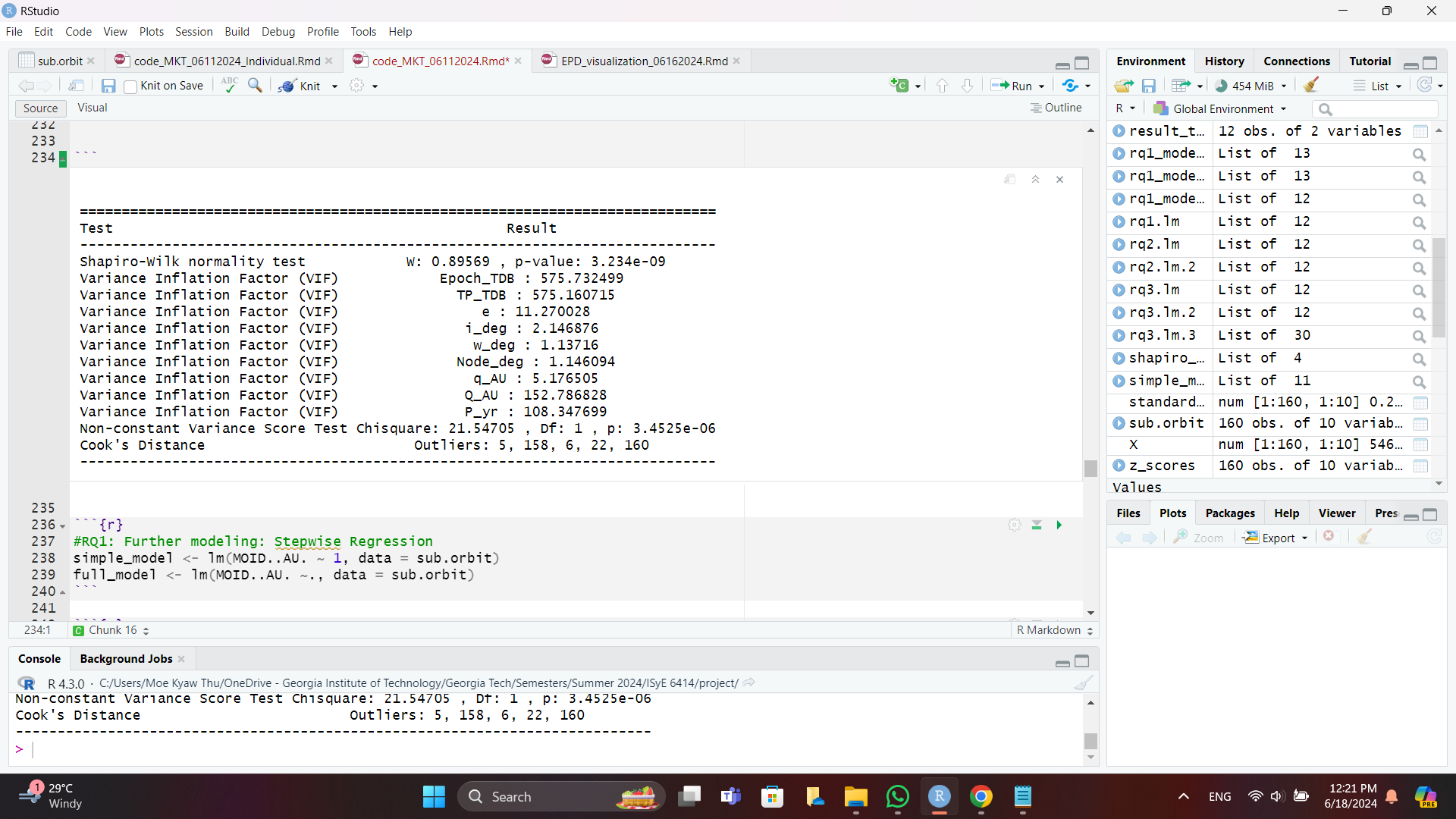
**Regression Modeling Results**

*Research Question I*

This study identifies the perihelion distance (q..AU.), longitude of ascending node (Node..deg.), argument of perihelion (w..deg.), and eccentricity (e) variables are significant for the minimum distance between the comets and the Earth (MOID..AU). To ensure the robustness of the model, diagnostic tests are conducted, namely normality, non-constant variance, multicollinearity, and outliers. The residuals were found to not be normally distributed, indicating issues with the normality assumption. There is evidence of heteroskedasticity, suggesting nonconstant variance. Multicollinearity is also detected among the remaining predictor variables, epoch (Epoch..TDB.), time of perihelion (TP..TDB.), aphelion distance (Q..AU.), and orbital period (P..yr.). Additionally, Cook’s Distance identifies five outlier observations at points 5, 158, 6, 22, and 160. To enhance model performance, stepwise and LASSO regression models are also employed, which is further discussed in the optimal regression section.



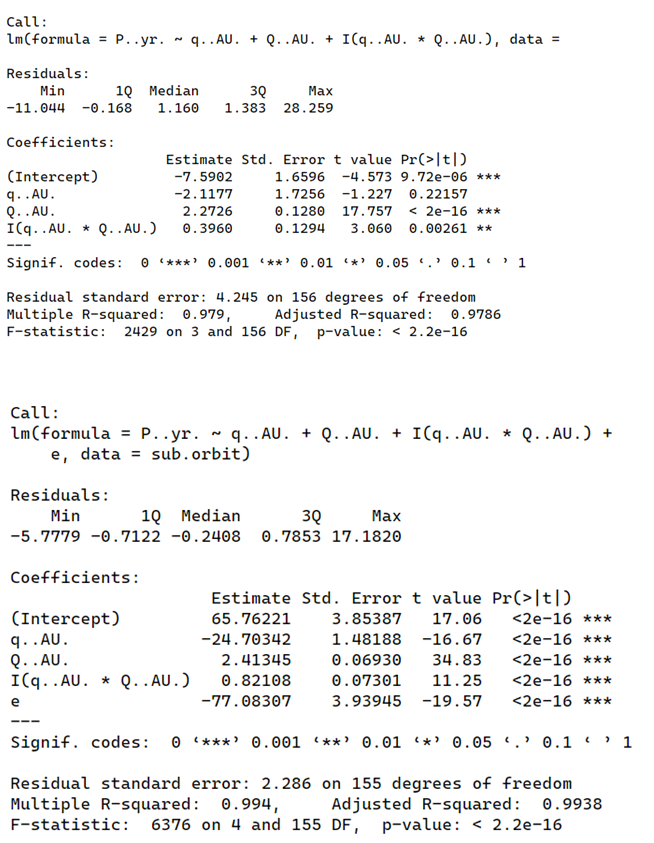
***Fig. 1.*** *RQ I Regression Results*

**

***Fig. 2.*** *Model Diagnostics Results*

*Research Question II*

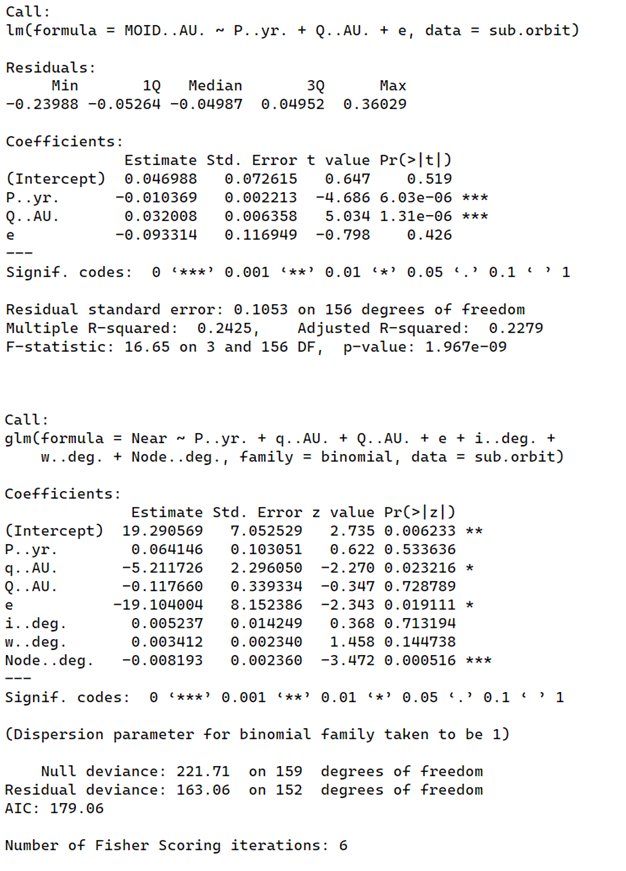
In order to tackle the second research question, this study conducts a regression analysis between P..yr. and the minimum (q..AU.) and maximum (Q..AU.) comet distances, including interaction terms between these two. Initially, it finds that, except for q..AU., the variables, including the interaction terms, are significant. In order to understand the model more in-depth, correlation tests are conducted with all the variables; this reveals that eccentricity variable, e, has the highest correlation with both q..AU. and Q..AU.. By considering e, a second round of regression analysis is performed, and the results now show that all variables including q..AU. are now significant predictors. Similar to research question I, model diagnostics are subsequently performed. The residuals for both models are not normally distributed. There is also evidence of heteroskedasticity in both models. Multicollinearity is detected between Q..AU. and the interaction term in both models. Additionally, Cook’s Distance identifies two outlier observations at points 159 and 123.



***Fig. 3.*** *RQ II Regression: Initial Results (Top) & Final Results (Bottom)*

*Research Question III*

Like the previous two research questions, the third and final research question conducts regression analysis between MOID..AU. and P..yr., with results showing high statistical significance. Using correlation tests, the new model indicates that MOID..AU. has a high correlation with P, Q..AU., and e, while P..yr. is highly correlated with Q and e. A second regression shows that, except for e, the other two variables are good predictors for MOID..AU. Due to the nature of the research question, in particular the last part, the third regression is conducted using a logistic regression approach. A new variable, Near, is created as 1 if the values MOID..AU. are less than 0.05, or otherwise 0. In the new logistic regression model, where Near is the response variable and MOID..AU. excluded, the results show that the variables, e, Node..deg., and q..AU. are significant predictors. Like the previous two research questions, model diagnostics are conducted for all three models. The residuals for none of the three models are normally distributed, indicating issues with normality. There is evidence of heteroskedasticity in all three models. Multicollinearity is detected between Q..AU. and P..yr. in the two models tested for it (no testing for model 1 since there’s only one variable). Additionally, Cook’s Distance identifies two common outlier observations at 123 and 98 in all three models.

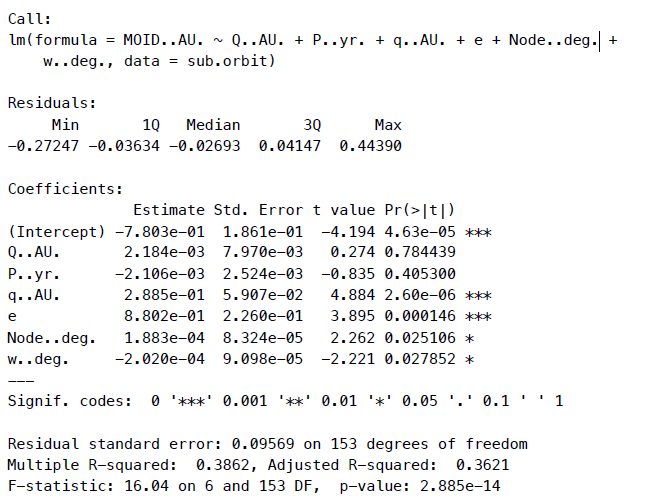


# ***Fig. 4.*** *RQ III Regression: Second Model Results (Top) & Logistic Results (Bottom)*

# **Optimal Regression Model for Research Question I**

## *Stepwise Regression: Forward Stepwise Regression*

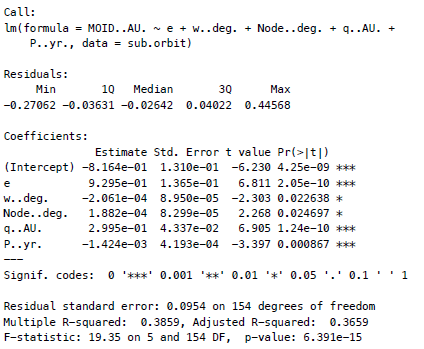
Given that research question I includes a large number of predictor variables, forward stepwise regression is applied, starting with a simple model and incrementally adding predictors based on the AIC criterion. After running a number of models, the final model includes variables Q..AU., P..yr., q..AU., e, Node..deg., and w..deg with the lowest AIC score of -744.09 while having an adjusted R-squared of 36.32 percent. However, despite the forward process, the significant predictors are identified as q..AU., e, Node..deg., and w..deg, which demonstrates that they are significant in predicting the minimum distance between the comet and Earth (MOID). The output is illustrated below:



***Fig. 5.*** *Forward Stepwise Regression Output*

### *Stepwise Regression: Backward Stepwise Regression*

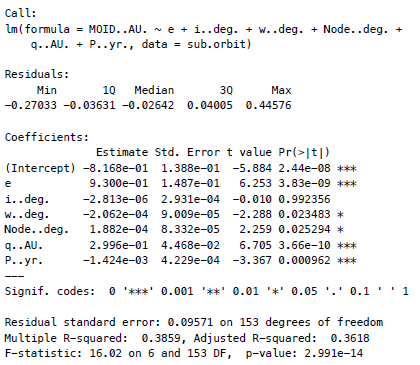
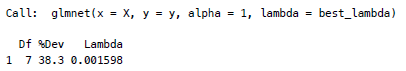
Retrospectively, backward stepwise regression is applied, starting with a full model and incrementally removing predictors based on the AIC criterion. After running a number of models, the final model includes variables e, w..deg., Node..deg., q..AU., and P..yr with the lowest AIC score of -746.01 while having an adjusted R-squared of 36.59 percent. Unlike the forward process, all predictor variables are significant here, which also demonstrates that they are significant in predicting the minimum distance between the comet and Earth (MOID). The output is illustrated below:



***Fig.6.*** *Backward Stepwise Regression*

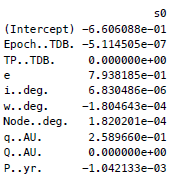
### *LASSO Regression*

After performing stepwise regression, this study further uses LASSO regression with cross-validation to select the optimal lambda. After applying LASSO, the final model includes P..yr., e, i..deg, w..deg, Node..deg, and q..AU. More significantly, the optimal lambda is 0.001597953, and the model achieves an adjusted R-squared of 36.18 percent. The significant predictors are identified as P..yr., e, w..deg, Node..deg, and q..AU. These selected variables are found to be significant in predicting the minimum distance between the comet and Earth (MOID). The output is illustrated below:



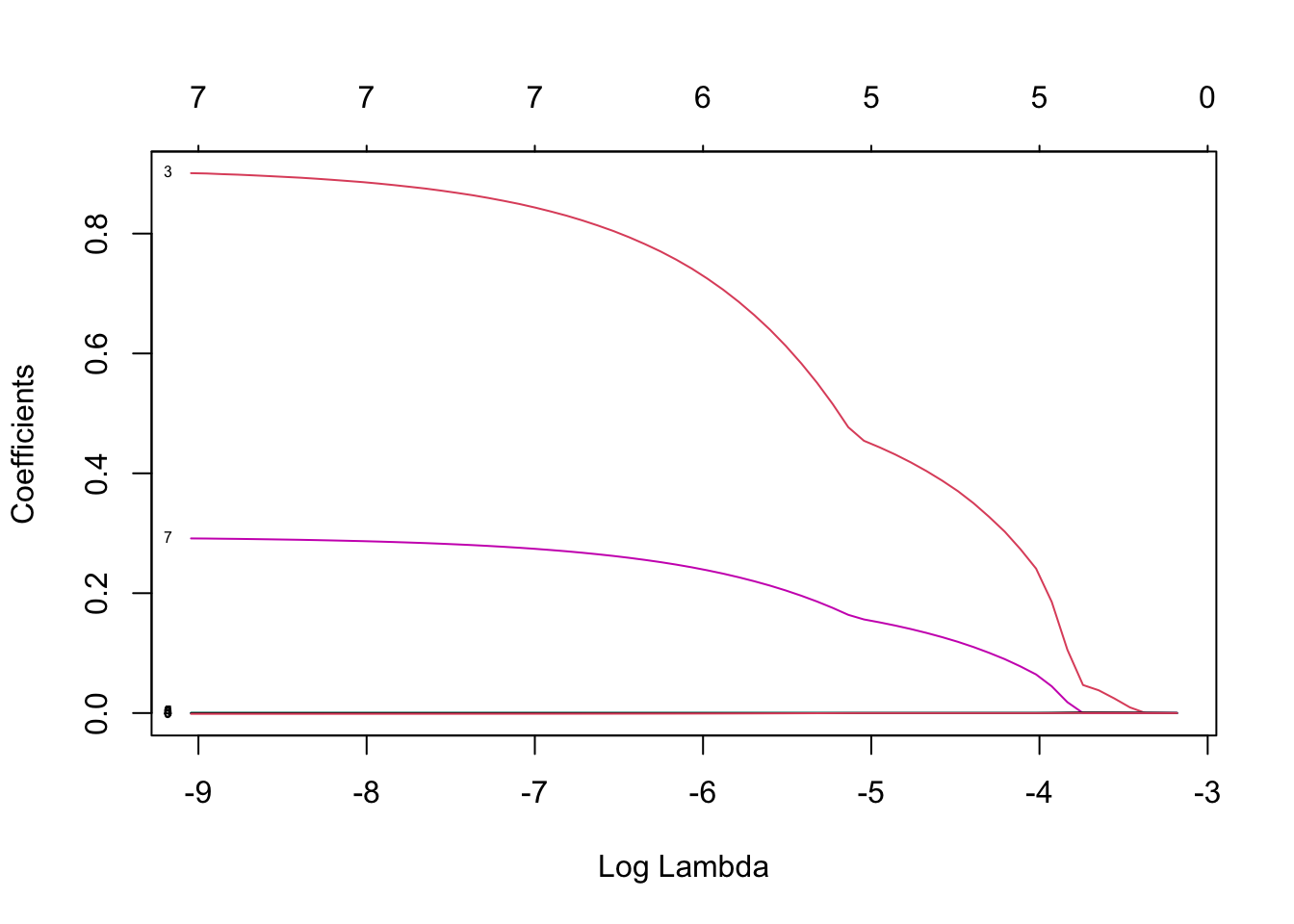
***Fig. 7.*** *LASSO Regression Output.*

The results further show that the predictor variables, which were selected, are shown in Figure 4 below. More significantly, the predictors with a coefficient of zero were dropped.



***Fig. 8.*** *LASSO Regression selected predictors.*

While the above output is important for feature selection, the LASSO regularization path plots also show that as the lambda increases, the absolute values of the coefficients shrink and tend towards zero, demonstrating the sparsity enforcement by shrinking less important coefficients to zero. As seen in Figure 7, at higher values of lambda (left side of the plot), more coefficients are non-zero, indicating that more variables are included in the model. As lambda decreases (move more to the right), fewer coefficients remain non-zero. This process effectively selects the most significant variables while penalizing less important ones. Moreover, the coefficients that remain non-zero even at higher values of *log(lambda)* (far left) are the most significant variables e.g. the red line representing a coefficient around 0.8 at the far left shows that the variable remains important even with a strong penalty. Finally, the point where lines start approaching zero significantly indicates the region of optimal lambda values.



**Fig. 9.** LASSO Regularization path plots.

**Discussion and Conclusion**

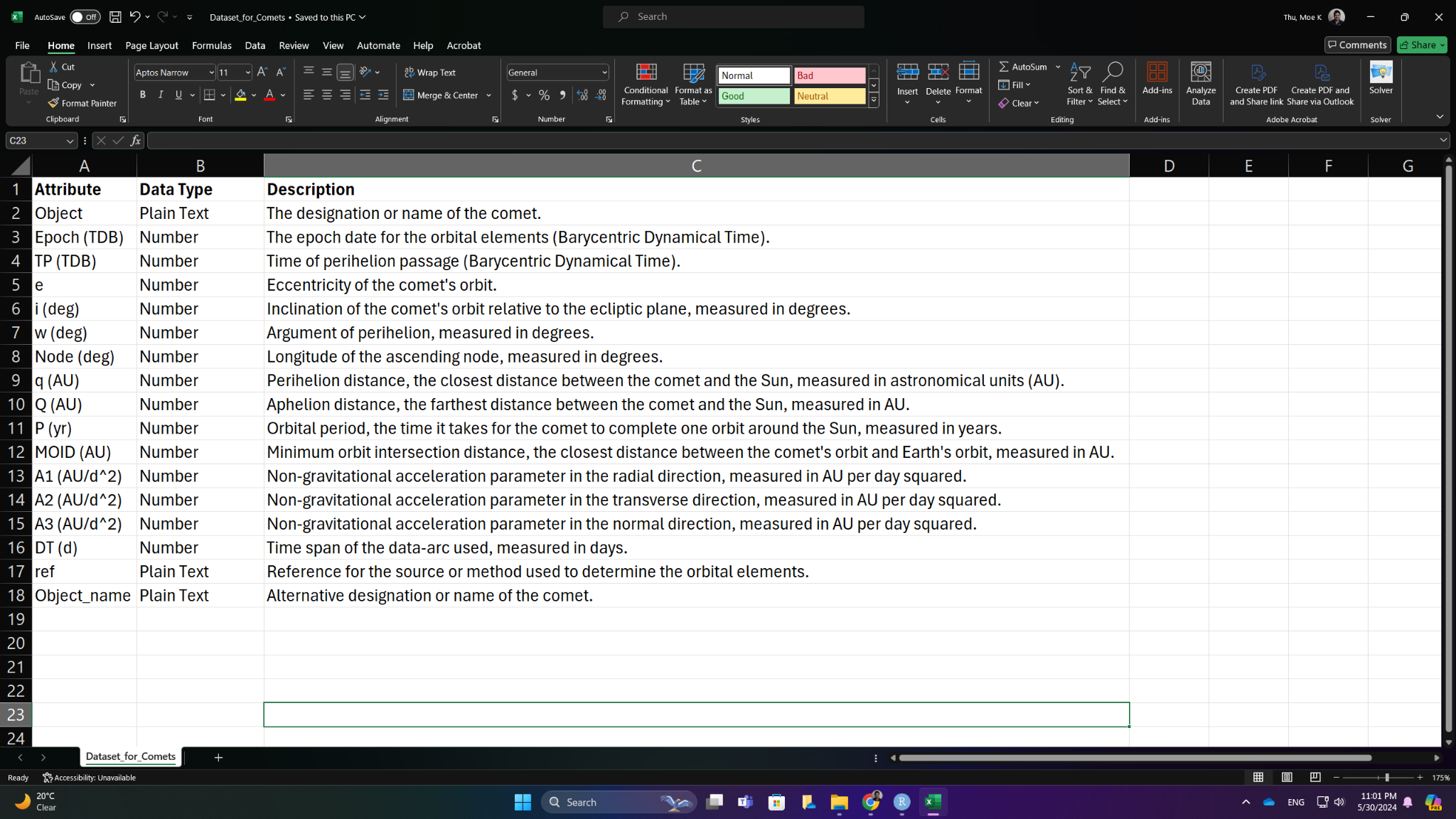
The culmination of these modeling tools allows researchers to adequately analyze relevant orbital parameters and their effect on NEC behavior, specifically MOID, with a relatively small dataset of observations and key features. Focusing on MOID allows the research team to determine the possible interactions with near-Earth space for scientific investigation and possible catastrophic events. The inclusion of the orbital period as a predictor and comparative variable provides additional insight into the frequency of these events. In conclusion, this study has shown firstly that orbital elements such as eccentricity (e), argument of perihelion (w..deg.), longitude of ascending node (Node..deg.), perihelion distance (q..AU.), and orbital period (P..yr.) are the best predictors for the distance between a comet and Earth (MOID..AU.). This is further supported by the stepwise and LASSO regression approaches. Secondly, the maximum distance (Q..AU.) and minimum distance (q..AU.) can significantly impact the orbital period (P..yr.) of the comet which informs how frequently the comet will enter near-Earth vicinity. Moreover, MOID..AU. are correlated with P..yr.; the coefficients are rather low at 0.0012, implying a weak relationship. However, by using the logistic regression model, this study has also shown that for a distance less than 0.05, the relationship between MOID..AU. and P..yr. are not significant at all which indicates that the minimum distance from Earth and the frequency in which it reaches its vicinity are not significantly correlated. Lastly, model diagnostics reveal issues with non-normality, heteroskedasticity, and multicollinearity, along with the presence of 3-4 general outliers.

**References**

1. Near Earth Object Program, “Near-Earth Comets - Orbital Elements | NASA Open Data Portal,” data.nasa.gov.
2. Solar System Dynamics, “Small-Body Database Query,” NASA Jet Propulsion Laboratory.
3. R Core Team, “R: A language and environment for statistical computing. R Foundation for Statistical Computing,” Vienna, Austria (2021).

**Appendix**

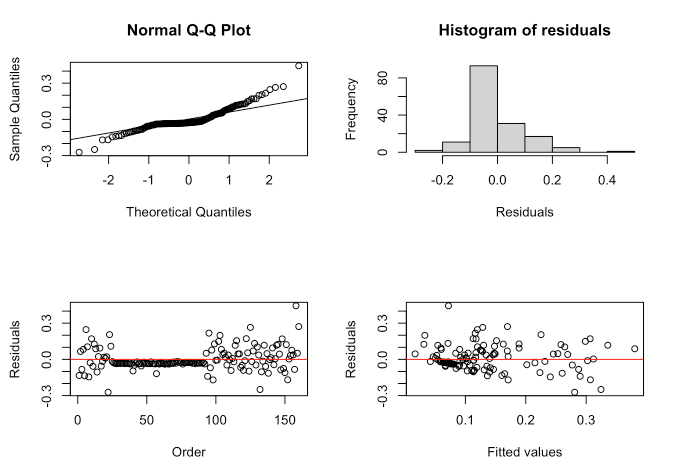
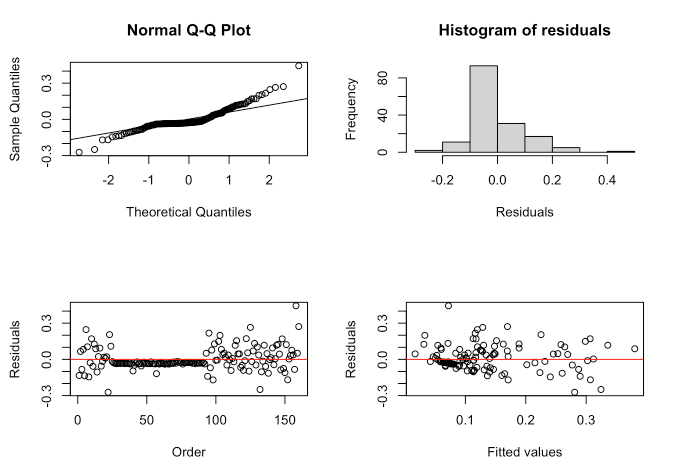
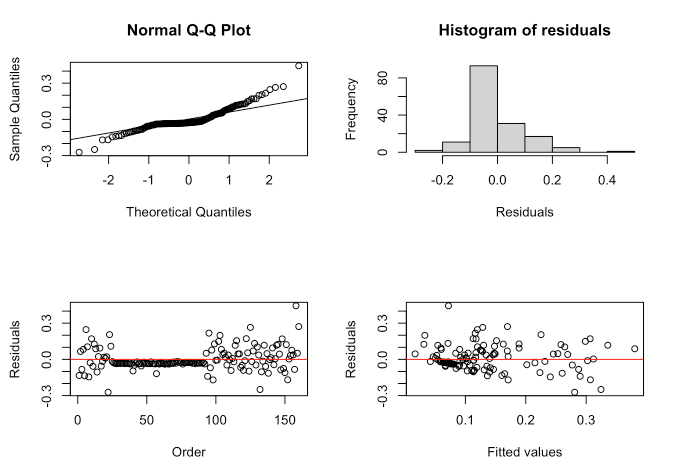
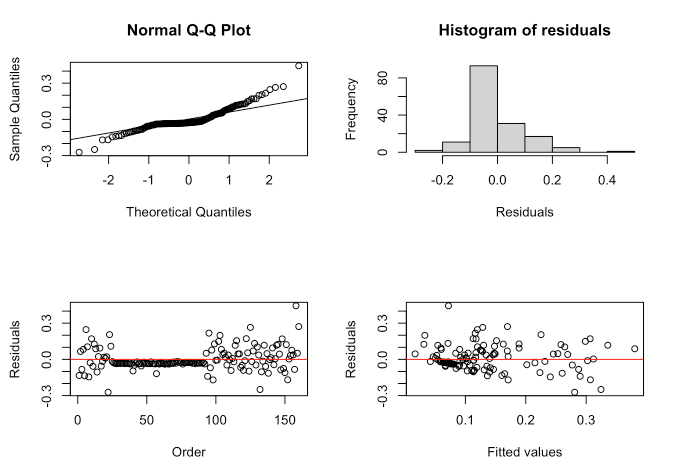
*Codebook*



***Fig. 10.*** *Codebook for Dataset*

## *Regression Assumption Visualizations for the LASSO Regression*

The visualization of the forward stepwise regression model, the backward stepwise regression model, and the LASSO regression model have similar results as seen in Figure 6 below.



***Fig. 11.*** *Regression Assumption Visualizations.*

**1. Normal Q-Q Plot:**

The points follow the reference line closely, indicating that the residuals are approximately normally distributed. Even with deviation at the tails, overall a good fit.

**2. Histogram of Residuals:**

The residuals are centered around zero, with a slight skew to the right and major residuals fall into the (-0.2 to 0.2). This can be implied that there are no major outliers or deviations from normality.

**3. Residuals vs. Order Plot:**

The residuals are randomly scattered around 0 and when more than 100, indicating that there is no discernible pattern over time and there is no autocorrelation in the residuals.

**4. Residuals vs. Fitted Values Plot**

The residuals are spread evenly around zero, without a clear pattern, which can imply that the model has captured the relationship between the predictors and the response variable well. There is no significant evidence of heteroskedasticity.

To interpret this, we can see that the forward stepwise regression model, the backward stepwise regression model, and the LASSO regression model fit the data well. The assumptions of normality, independence, and homoscedasticity of residuals are reasonably met. The model is appropriate for predicting the minimum distance between the comet and Earth (MOID) using the selected orbital elements.