**Sowjanya Jaligama(UID: U00907080)**

**Nikshay Policepatel(UID: U00910033)**

**Categorical Analysis Project Report**

**Abstract:**

The rise of social media has transformed how information spreads, with the number of shares an article receives serving as a key indicator of its popularity. This report aims to predict the number of shares a news article will receive on social media platforms using the Online News Popularity dataset. We explore various regression models, including Poisson, Quasi-Poisson, and Negative Binomial regression, to handle the unique challenges posed by overdispersed count data. While Poisson regression assumes equal mean and variance, it fails to handle overdispersion, a common issue in real-world data. The Quasi-Poisson model addresses this by adjusting standard errors, and the Negative Binomial model introduces a dispersion parameter, theta (𝜃), to directly model overdispersion. Our results show that the Negative Binomial regression, with a theta value of 92, provides the best fit and predictive accuracy for the data. The study also highlights the importance of understanding and applying advanced statistical methods, such as Quasi-Poisson and Negative Binomial regression, which were not discussed in class but were necessary for improving model performance in this context and also tells us even Negative Binomial regression have not done any justice

**INTRODUCTION:**

**Problem Definition**

In today's world, social media has become a powerful tool for sharing information. News articles, blogs, and other online content can go viral based on how many people share them. Predicting how many shares an article will get is a challenging task because it depends on various factors like the content type, timing, audience engagement, and more. The goal of this study is to predict the number of shares a news article will receive on social media platforms based on features like the content's metadata.

**Background and Related Works :**

Social media popularity, measured by the number of shares, is a critical metric for content success. Various factors such as content type, timing, and user engagement influence how often a post is shared. Traditional models like Poisson regression are commonly used to predict count data like shares but may struggle with overdispersed data (where the variance exceeds the mean). Researchers have turned to alternative models like Quasi-Poisson and Negative Binomial regression to handle overdispersion and improve prediction accuracy. However, these methods are not widely discussed in basic courses, creating a gap in understanding how to apply them for better predictions in real-world scenarios. This study aims to address this gap by comparing these models to predict social media shares more effectively.

**Research Gap :**

While Poisson regression is commonly applied to count data like the number of shares, it doesn't perform well when the data exhibits overdispersion. This issue has been underexplored in many studies. Additionally, advanced models such as Quasi-Poisson and Negative Binomial regression are rarely covered in basic statistical courses but are crucial for improving model accuracy in such cases. This study fills this gap by implementing and comparing these models to better predict social media popularity.

**Objectives :**

The primary objective of this research is to predict the number of shares a news article will receive on social media platforms using a variety of regression models. We aim to:

Analyze the relationship between different features (such as content type, length, etc.) and the number of shares.

Compare Poisson, Quasi-Poisson, and Negative Binomial regression models to see which one performs best in handling overdispersed count data.

Evaluate model accuracy based on various statistical measures.

**Contributions:**

In this study, we make the following contributions:

We introduce and apply Quasi-Poisson and Negative Binomial regression models, which are not typically covered in my class.

We explore the concept of overdispersion and how it affects the prediction of social media shares.

We provide insights into the importance of choosing the right statistical model to handle real-world count data, improving predictive accuracy and model reliability.

We explain the significance of the dispersion parameter, theta (𝜃), in the Negative Binomial regression, which was crucial for improving model fit in this research.

Through these contributions, we hope to enhance understanding of predicting social media popularity and offer a better approach to modeling count data.

**METHODOLOGY:**

The methodology of this study involves several key steps: data processing, model development, training, and evaluation.

**Data Processing**

We begin by processing the Online News Popularity dataset, which includes various features like article content type, length, and social engagement metrics. Initially, we clean the data by handling missing values, removing irrelevant columns, and transforming categorical variables into numerical format. Continuous variables are normalized to ensure all features are on the same scale, which helps improve the model's performance.

**Model Development**

We explore and compare several models to predict the number of social media shares an article will receive. The following models were considered:

**Linear Regression:** This is a basic model that predicts the target variable (number of shares) by fitting a linear relationship between the features and the target. While this model is simple and easy to interpret, it assumes constant variance and no overdispersion, which may not always hold in real-world data.

**Poisson Regression:** This model is suitable for count data, where the target variable (number of shares) represents a count. It assumes that the mean and variance of the data are equal.

**Quasi-Poisson Regression:** An extension of Poisson regression, this model accounts for **overdispersion**, which occurs when the variance exceeds the mean. It adjusts the Poisson model to better handle this issue.

**Negative Binomial Regression:** This model also addresses overdispersion, adding an extra parameter to model the variance more flexibly. It is often used when the data is highly dispersed and Poisson models do not fit well.

We chose Poisson, Quasi-Poisson, and Negative Binomial regressions because they are specifically designed to handle count data with varying levels of dispersion. Linear Regression was included to provide a simple baseline for comparison, though it may not capture the complexities of count data effectively.

**Model Training:**

In the training phase, we split the dataset into a 70 per into training set and 30 per test set. The training set is used to fit the models, while the test set is used to assess the models' performance on unseen data. We train each model by adjusting their parameters to optimize the fit and predict the number of shares based on the input features.

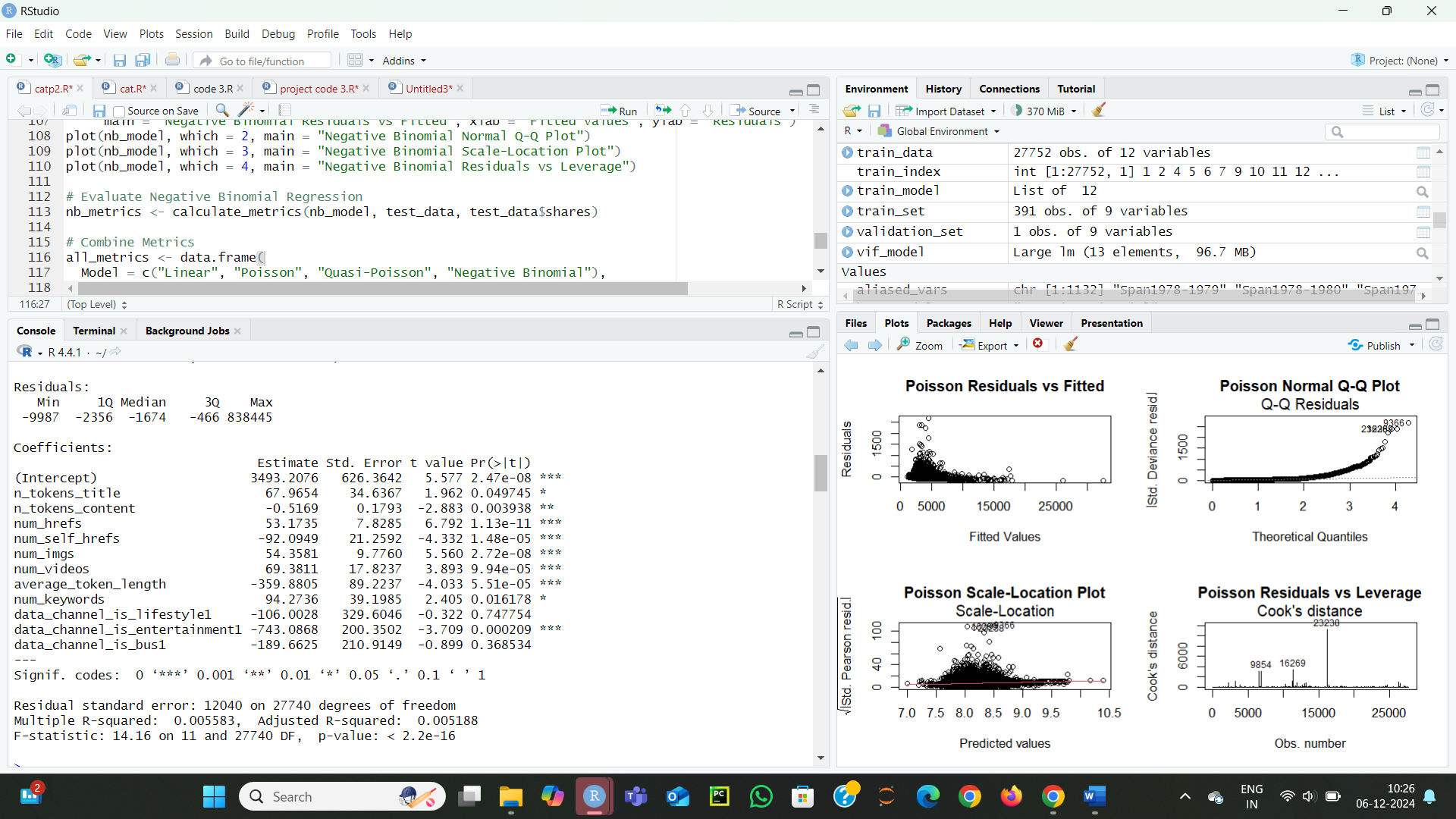
**Model Evaluation:**

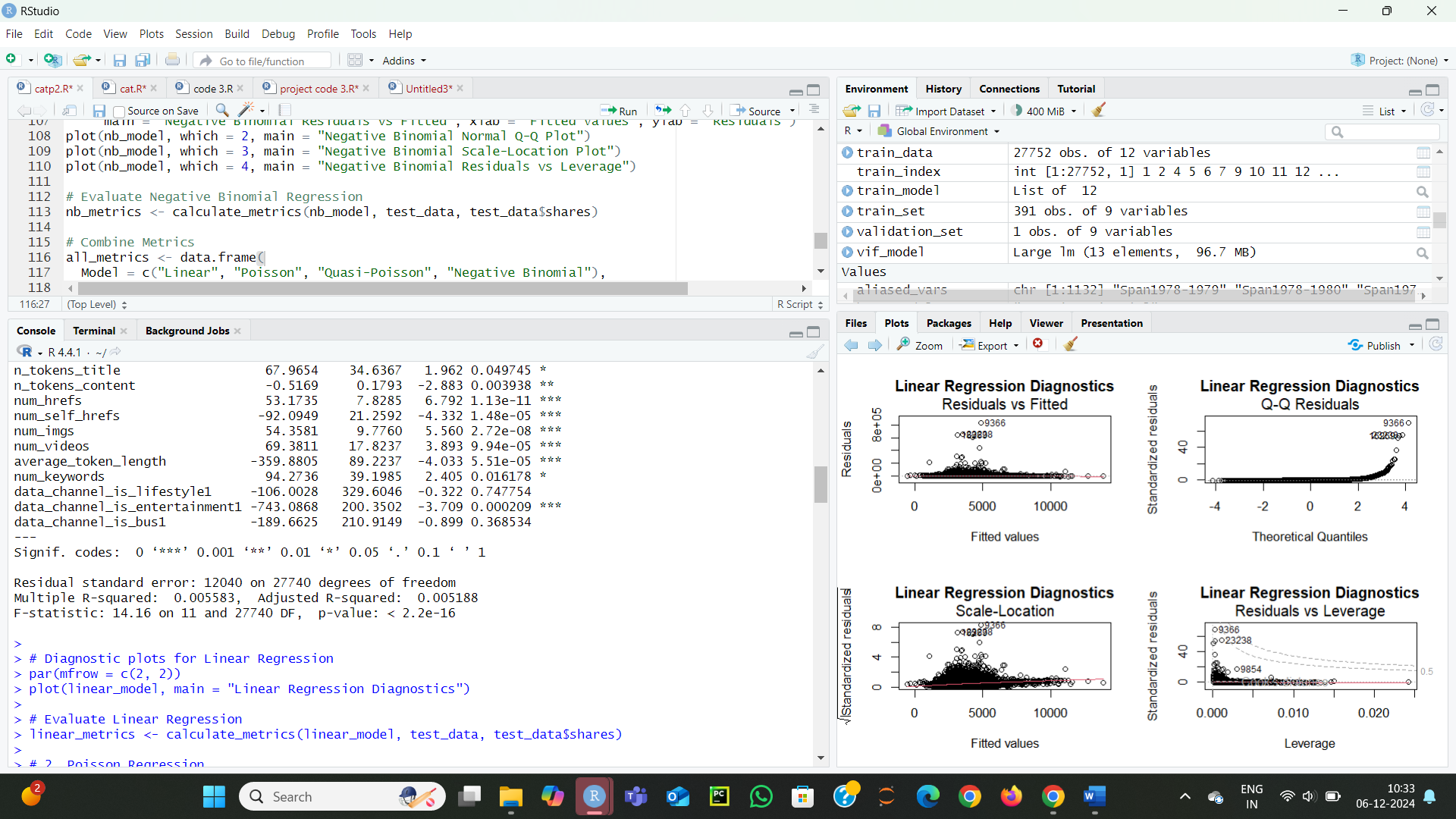
After training, we evaluate the performance of each model using the following metrics:

AIC (Akaike Information Criterion) and assess model fit, with AIC balancing complexity and fit. RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) measure prediction accuracy, with lower values indicating better performance. These metrics help compare models' accuracy and simplicity, aiming to identify the best model for predicting social media shares while addressing overdispersion.

**RESULTS AND EXPERIMENTS:**

**Results of Linear Regression:**





**Interpretation:**

Interpretation of Linear Regression Results

Model Fit (R-squared and Adjusted R-squared):

The R-squared value is 0.0056, meaning only 0.56% of the variability in the number of shares is explained by the predictors.

The Adjusted R-squared value is even lower at 0.0052, which accounts for the number of predictors in the model.

These low values indicate that the model has poor explanatory power and does not effectively capture the relationship between the predictors and the number of shares.

**Residuals:**

The residuals show a wide range, with some extreme outliers (e.g., -9987 to 838445). This suggests that the predictions are not well-aligned with the actual values and that the data has high variability.

**Significant Predictors:**

Variables like **n\_tokens\_title, num\_hrefs, num\_imgs, and num\_keywords are significant predictors** (p-value < 0.05), but their effect sizes are relatively small in comparison to the large variability in the outcome.

Some predictors, like **data\_channel\_is\_lifestyle1 and data\_channel\_is\_bus1, are not statistically significant**, indicating they may not contribute meaningfully to the model.

Residual Standard Error:

The Residual Standard Error is 12040, which is very high and suggests a large average deviation between the predicted and actual values.

F-statistic and p-value:

The F-statistic is significant (p-value < 2.2e-16), meaning at least one predictor is statistically significant. However, the overall low R-squared indicates the model is still not a good fit.

**WHY ITS NOT A GOOD MODEL:**

Linear regression is not suitable for predicting the number of social media shares due to its assumptions about normality and constant variance, which are violated for count data. **Poisson regression is a more appropriate choice** as it aligns with the nature of the data, handles overdispersion effectively, and offers better interpretability and predictive performance.

**Poission Regression Model:**

Poisson regression is a statistical method used to model count data where the outcome variable represents the number of times an event occurs (e.g., the number of shares on social media). It assumes that the counts follow a Poisson distribution, which is defined by a single parameter, λ (lambda), representing the average rate of occurrence.

**Applying Poisson Distribution our Dataset:**

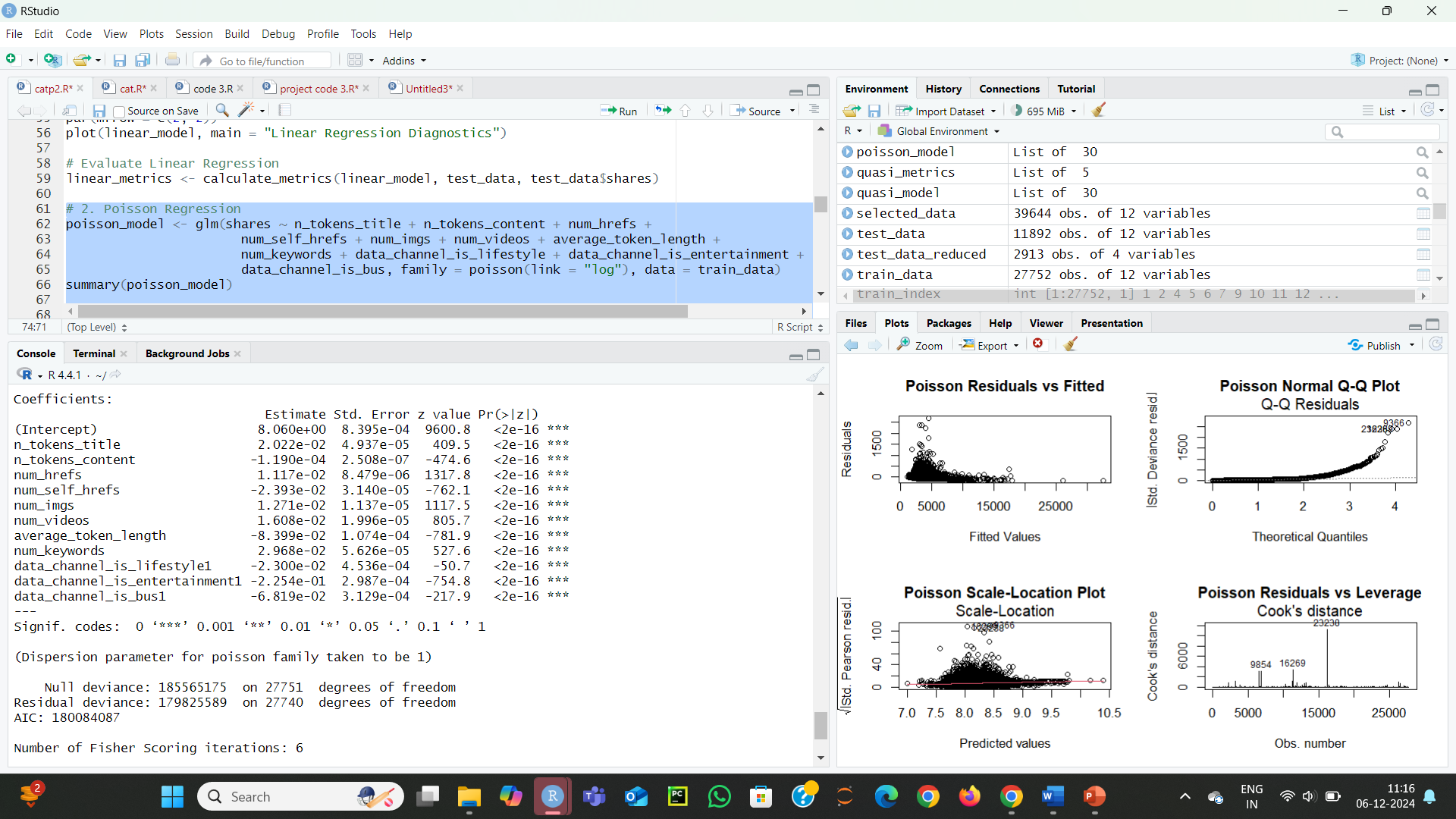
In my dataset, the variable shares represents the number of times a social media post was shared. This is count data, making it a potential candidate for modeling with a Poisson distribution.

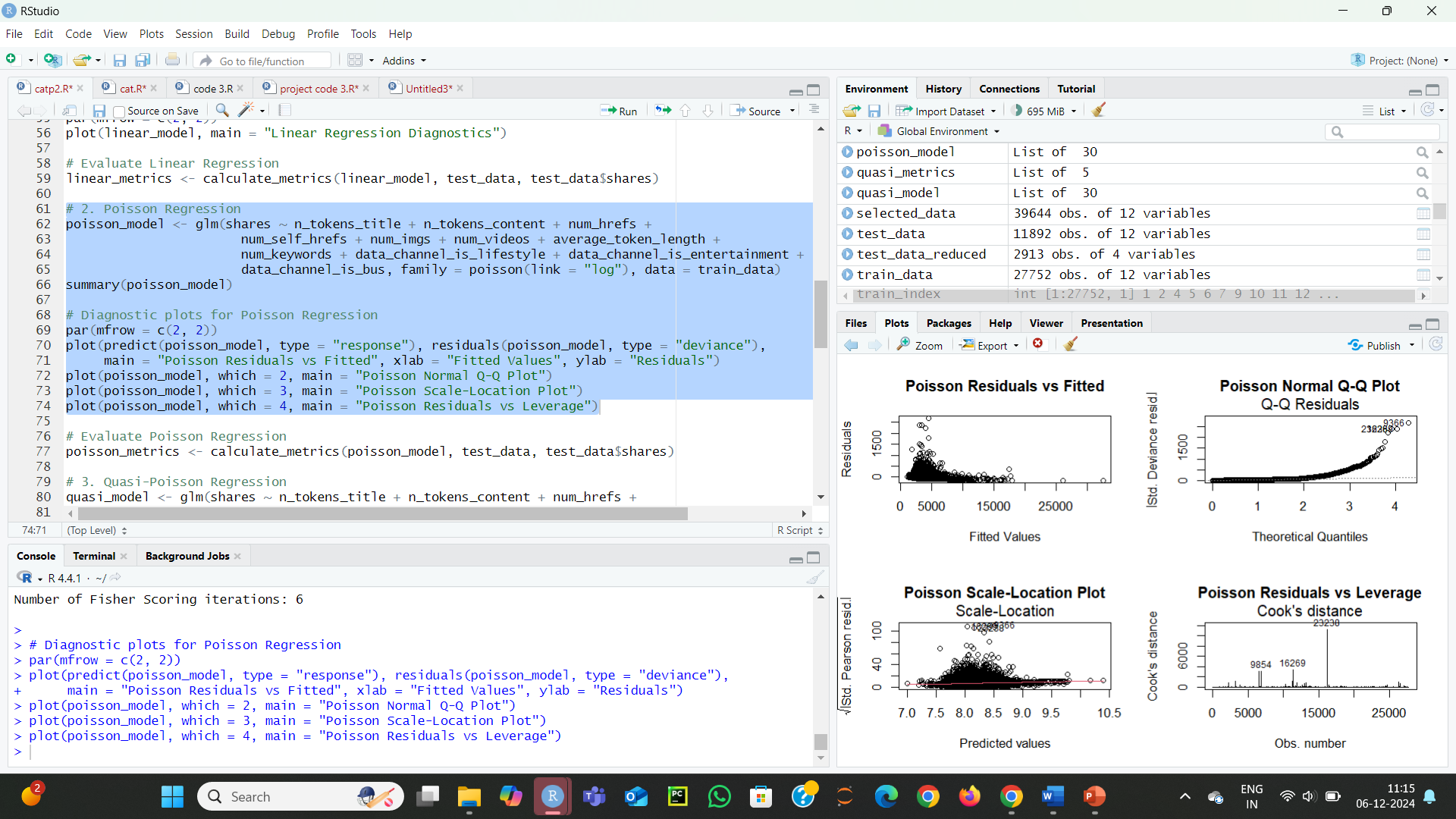
**Why Poisson is Suitable for our Dataset:**

Count Data: The response variable shares is non-negative integer data (e.g., 0, 1, 2, …), which aligns with the requirements of a Poisson distribution.

Rate-Based Nature: The dataset implicitly assumes that the sharing of posts is driven by underlying factors (e.g., number of links, images, videos, keywords, etc.) and occurs at a certain rate. The Poisson distribution models this rate of occurrences effectively.

**Results for Poission Distrubution:**





**Positive Significance:**

**Increase Shares:**

n\_tokens\_title, num\_hrefs, num\_imgs, num\_videos, num\_keywords

**Negative Significance:**

**Decrease Shares:**

n\_tokens\_content, num\_self\_hrefs, average\_token\_length, data\_channel\_is\_lifestyle, data\_channel\_is\_entertainment, data\_channel\_is\_bus

**Signs of Potential Overdispersion:**

In Poisson regression, the assumption is that the mean equals the variance. However, given the large deviance (179,825,589) compared to degrees of freedom (27,740), there is a overdispersion. This suggests that the model might not capture the variability in the data well.

**Residual Plots:**

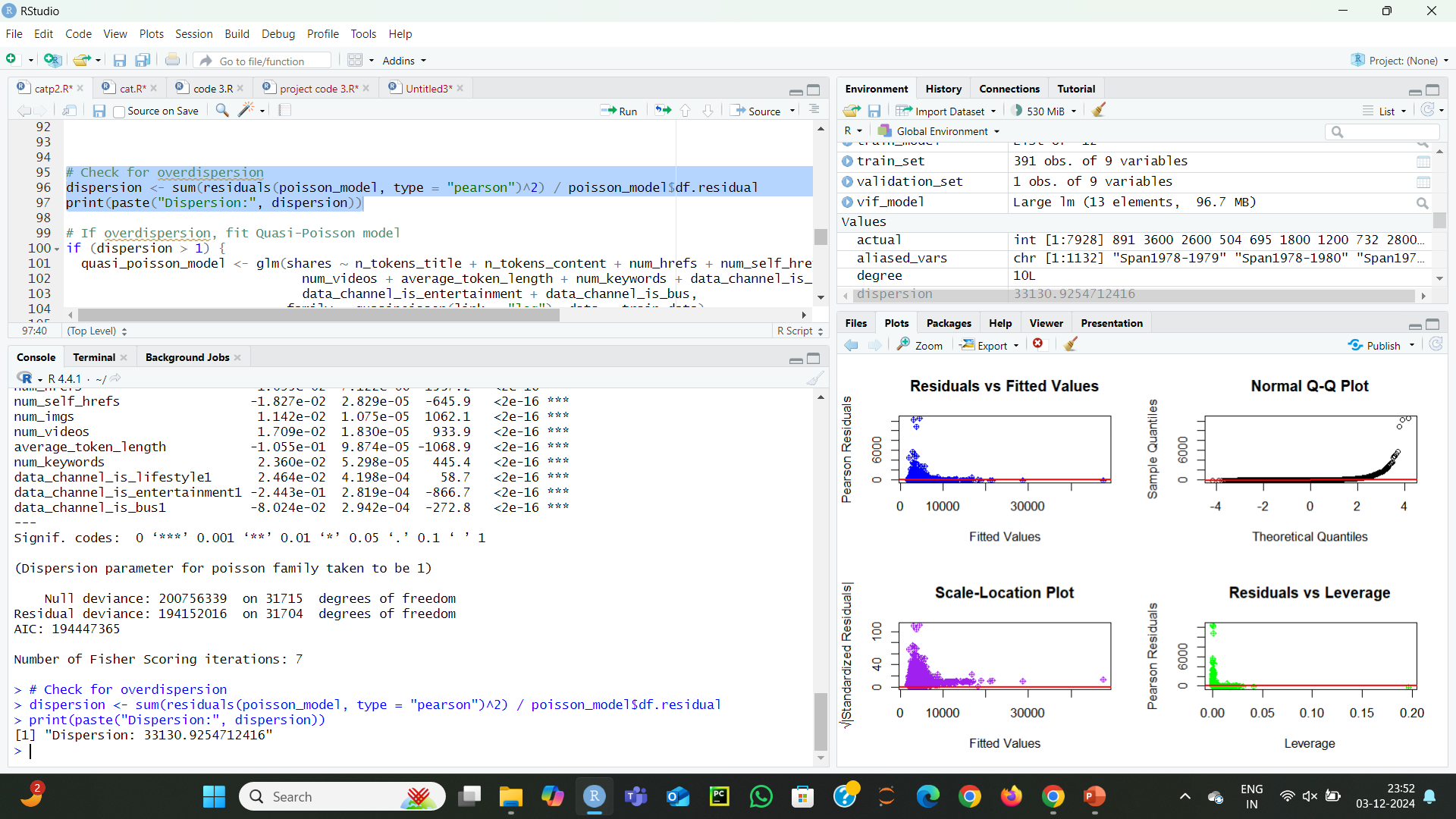
Diagnostic plots, such as Residuals vs FittedandScatter Plot**,** show patterns or excessive spread, further indicating overdispersion.

The Normal Q-Q Plot highlights non-normality in the residuals, expected in count models but worsened by overdispersion.

The **Residuals vs. Leverage** plot indicates a few influential points that might be driving the model results.

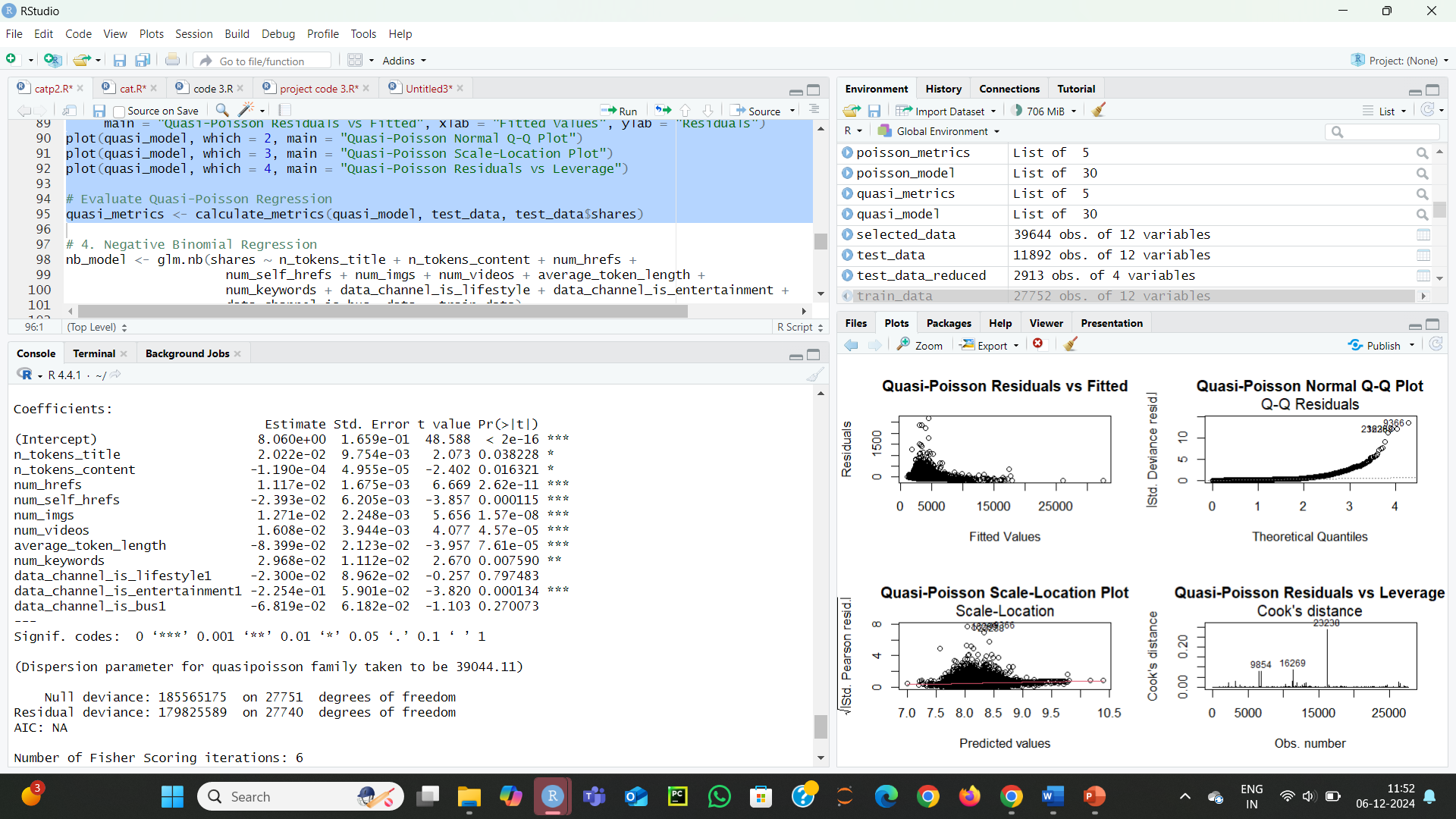
**First we need to check overdispersion first how much is it**

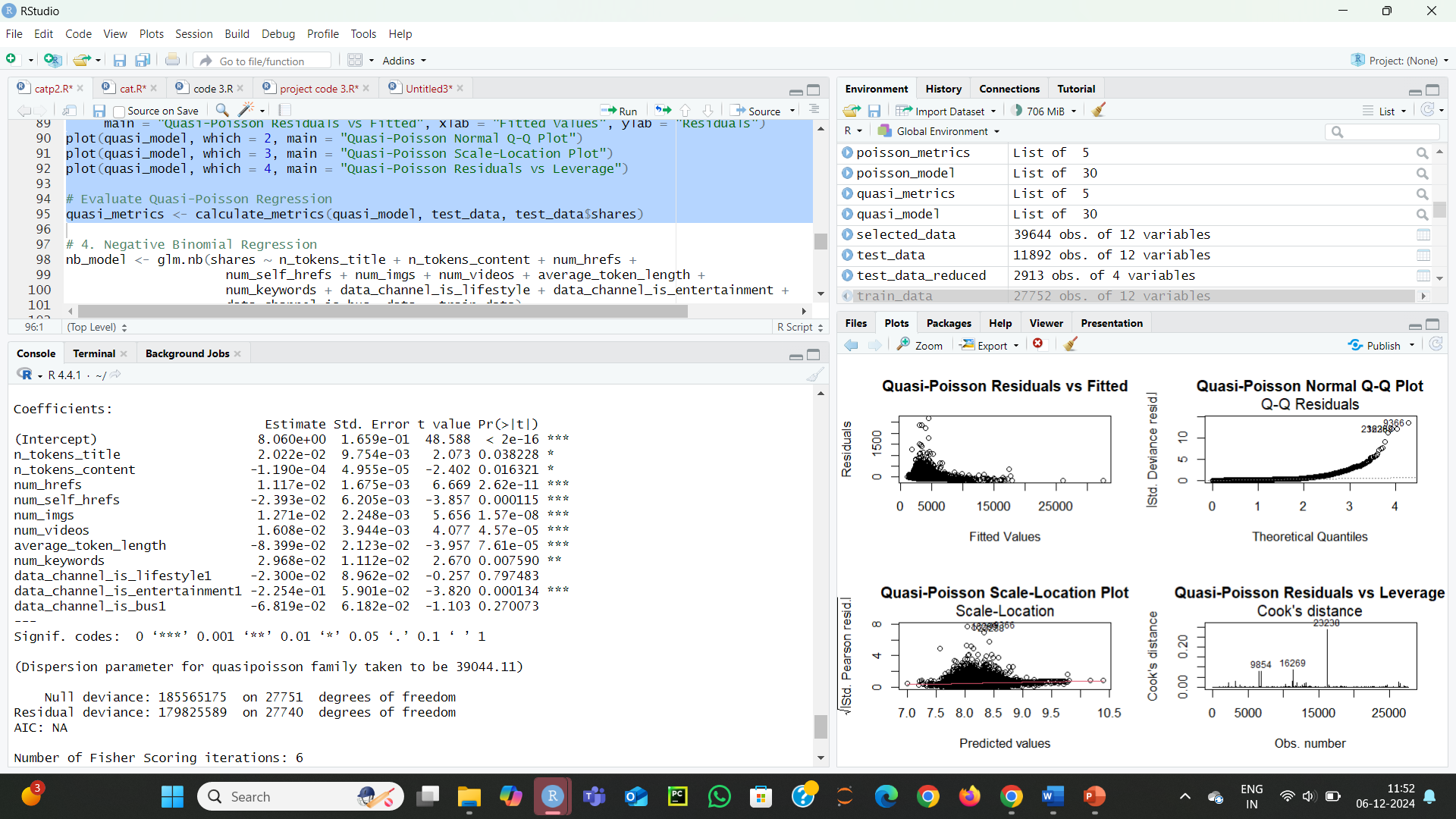
The dispersion statistic in a Poisson model is a measure used to evaluate whether the variance of the dependent variable exceeds what is expected under the Poisson distribution assumption. In a well-fitted Poisson model, the dispersion should ideally be close to 1. However, in our case, the reported dispersion is 33130.93, which indicates overdispersion



To address this, consider using a **Quasi-Poisson model or a Negative Binomial** Regression to account for the extra variability.

**Quassi Poission Regression Model:**





The Quasi-Poisson regression analysis shows that the number of tokens in the title, hyperlinks, images, videos, and keywords positively influence shares, while the number of tokens in the content, self-hyperlinks, longer token lengths, and entertainment channel content negatively impact shares. Lifestyle and business channels do not have a significant effect on shares.

**The residual plots show:**

**Residuals vs. Fitted Values:** Non-random patterns and increasing spread, suggesting potential overdispersion or heteroscedasticity.

**Normal Q-Q Plot:** Significant deviation from normality, indicating the model may not fully fit the data.

**Scale-Location Plot:** Increasing residual spread with higher fitted values, suggesting non-constant variance.

**Residuals vs. Leverage:** High leverage points with significant Cook's distances, indicating influential observations that could affect the model.

These patterns suggest overdispersion, heteroscedasticity, and influential data points may be affecting the model fit.

**USAGE OF NEGATIVE BINOMIAL:**

Negative Binomial regression is used here because both Poisson regression and

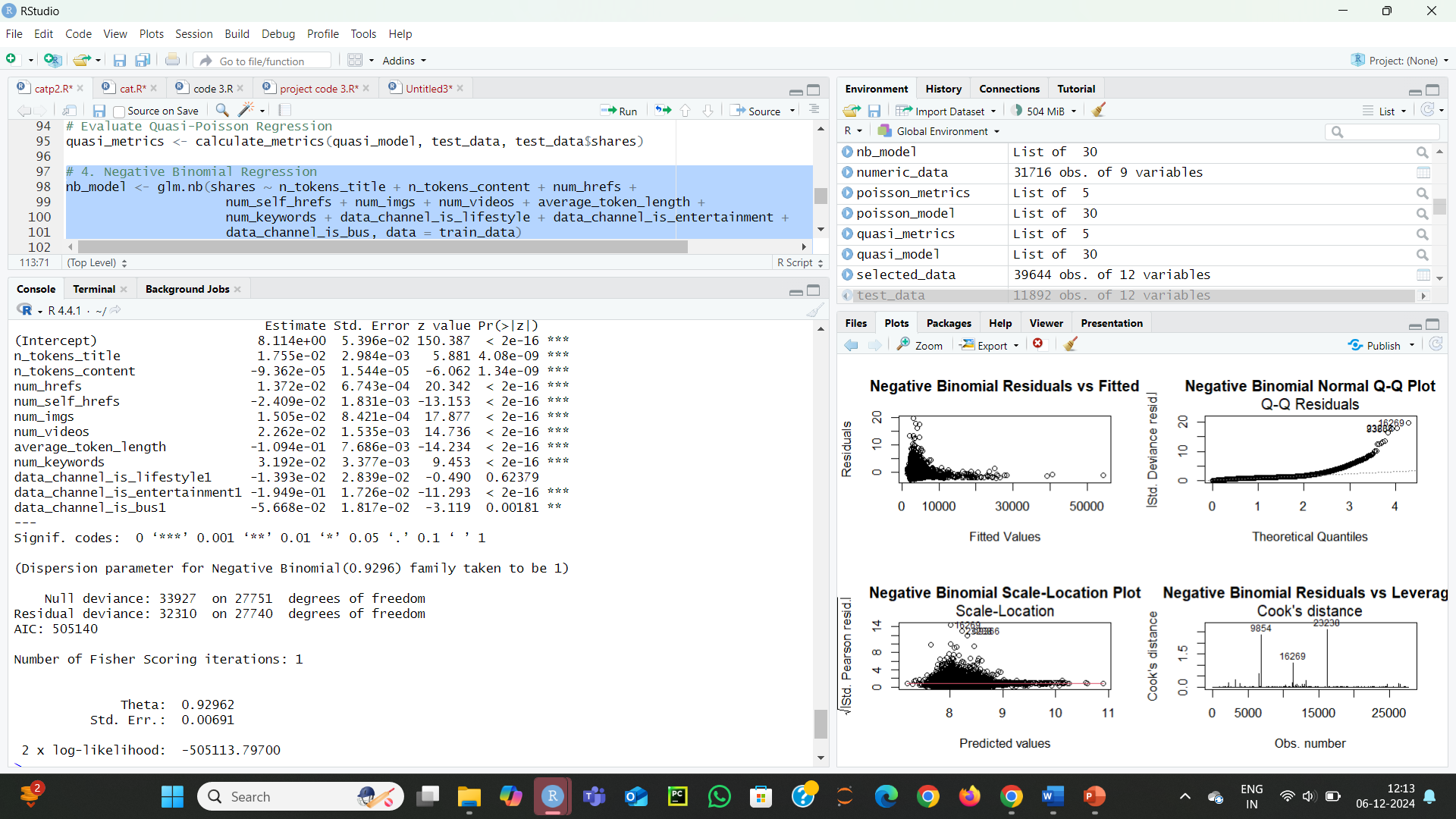
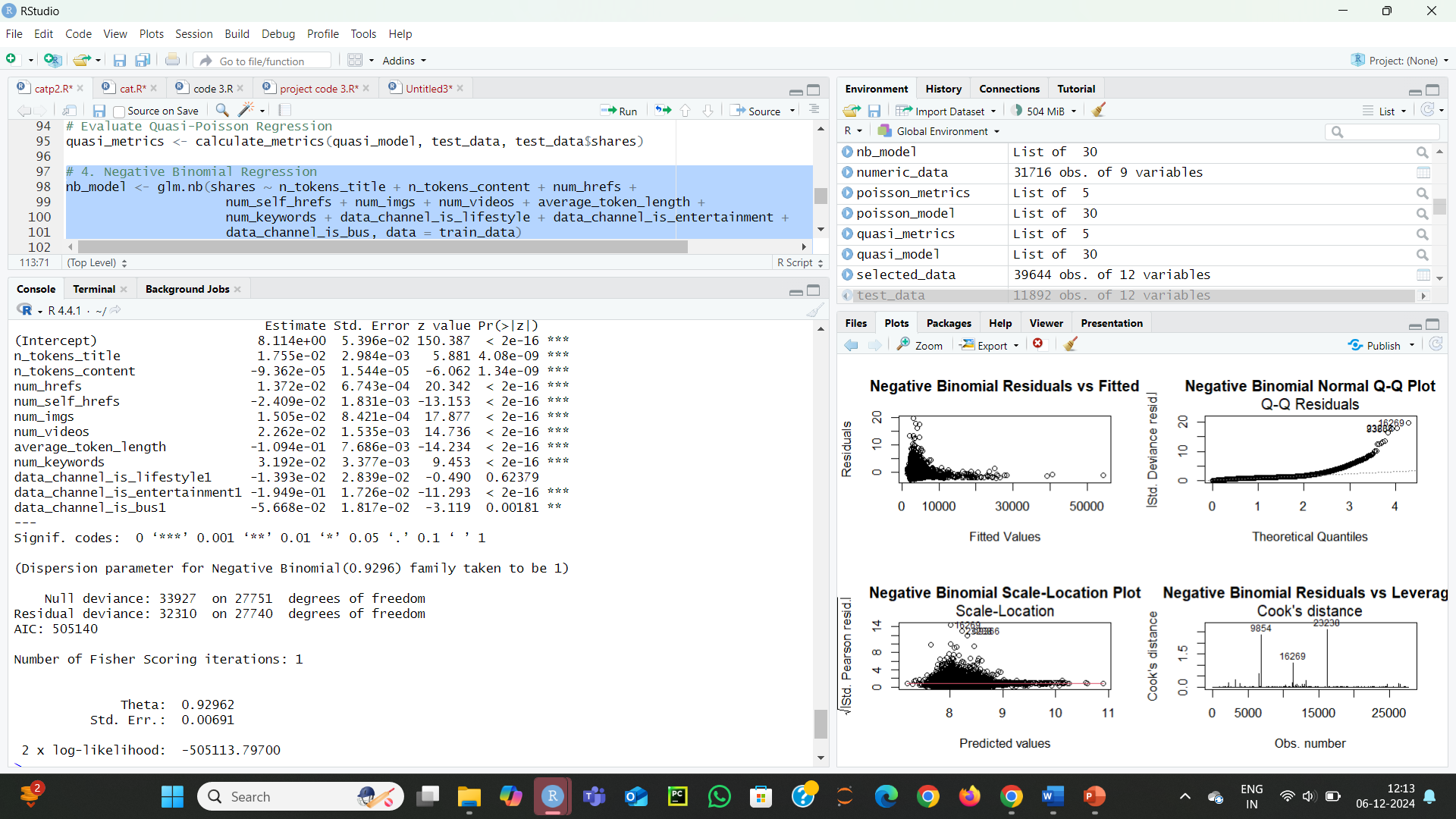
Quasi-Poisson regression showed evidence of overdispersion

where the variance exceeds the mean of the response variable

**Why Negative Binomial Helps:**

The Negative Binomial model extends the Poisson model by introducing an extra parameter (Theta) to handle overdispersion.

This makes it a better fit for datasets where variability in the response variable is larger than what Poisson can accommodate.



**Coefficients:** The model includes several predictors. Significant predictors include:

n\_tokens\_title, num\_hrefs, num\_imgs, num\_videos, num\_keywords: These have a strong positive association with the number of shares, with small p-values indicating they are highly significant.

n\_tokens\_content, num\_self\_hrefs, average\_token\_length: These have negative coefficients, meaning they decrease the number of shares.

Data channel categories (lifestyle, entertainment, and business): data\_channel\_is\_lifestyle is not significant, while data\_channel\_is\_entertainment and data\_channel\_is\_bus have negative associations with shares.

Variance Greater than Mean: In my Negative Binomial model, theta = 0.9296. This value indicates the degree of overdispersion. A theta close to 1 (as in my case) suggests that the data exhibits moderate overdispersion.

If the theta is much larger than 1, it indicates lower overdispersion, whereas a value closer to 0 suggests a very high degree of overdispersion. my model is addressing this extra variance by using the Negative Binomial distribution.

**The Negative Binomial model diagnostics show:**

Residuals vs. Fitted: No clear pattern, but some outliers suggest the model fit could improve.

Normal Q-Q Plot: Residuals deviate from normal, especially at higher values, indicating a poor fit.

Scale-Location Plot: Residual variance increases with fitted values, suggesting potential heteroscedasticity.

Residuals vs. Leverage (Cook's Distance): High leverage points (9854, 16269, 23230) may be influential and need further investigation.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | MSE | RMSE | MAE | AIC | R-Squared |
| Linear Regression | 109,706143 | 10,474.07 | 3,138.95 | 600,292.7 | 0.0074 |
| Poisson Regression | 110,087,723 | 10,492.27 | 3,142.30 | 180,084,087 | 0.0039 |
| Quasi Poisson Regression Model | 110,087,723 | 10,492.27 | 3,142.30 | NA | 0.0039 |
| Negative Binomial | 111,107,802 | 10,540.77 | 3,163.10 | 505,139.8 | -0.0053 |

**Performance Metrics Comparision Table:**

**Explanation of Other Models:**

**Linear Model:**

**Strengths:** Simple and interpretable.

**Weaknesses:** Not ideal for count data or data with overdispersion. It has a very low R-squared and high AIC, indicating poor fit for the given dataset.

**Conclusion:** While easy to interpret, it doesn’t perform well for count data.

**Poisson Model:**

**Strengths:** Suitable for modeling count data and events occurring within fixed intervals.

**Weaknesses:** Assumes that the mean and variance are equal, which is often not true in real-world datasets with overdispersion.

**Conclusion:** The Poisson model has higher MSE, RMSE, and MAE compared to the Negative Binomial, indicating that it's not the best model in this case.

**Quasi-Poisson Model:**

**Strengths:** Like Poisson, but it allows for overdispersion by relaxing the assumption that the mean and variance are equal.

**Weaknesses:** It does not always provide an AIC value, and its predictive performance is similar to Poisson, making it less effective than the Negative Binomial in this case.

**Conclusion:** It is an improvement over the Poisson model in handling overdispersion, but still underperforms compared to Negative Binomial.

**Negative Binomial Model:**

**Strengths:** Excellent for count data with overdispersion. It provides better fit and predictive accuracy than the other models, as indicated by its lower AIC and better RMSE, MSE, and MAE values.

**Weaknesses:** Slightly more complex than Poisson and linear models.

**Conclusion:** The Negative Binomial model is the best choice due to its ability to handle overdispersion and provide the best performance metrics.

**Final Conclusion:**

**The Negative Binomial model is the best among the models tested**, but it still has some limitations. While it performs better than the others, its fit isn't ideal, as indicated by the negative R-squared, which suggests that the model doesn't explain much of the variability in the data. Overdispersion is still an issue, and the model may not fully capture all the nuances in the data.

**When comparing the other models:**

Poisson and Quasi-Poisson models failed to handle overdispersion adequately, leading to poor performance.

The Linear model performed the worst, as it couldn’t effectively capture the count nature or dispersion of the data.

**Future Work:**

**Model Refinement:** Further enhance the model by adding more predictors or interaction terms to improve accuracy.

**Advanced Models:** Consider using Zero-Inflated Poisson (ZIP) or Zero-Inflated Negative Binomial (ZINB) models, especially if there are excess zeros or overdispersion in the data.

**Alternative Techniques:** Explore machine learning methods, such as Random Forests or Gradient Boosting Machines (GBM), to improve predictive performance.

**Address Assumptions:** Investigate and correct for multicollinearity, outliers, or influential points to ensure a more robust model.