PROJECT

Janaki Ramya Namburu

Department of Health Informatics, Rutgers University

BINF5040E001– BIOMEDICAL INFORMATION PROCESS

Dr. Mohamed Ahmed

**Title:**

Exploratory Data Analysis and Predictive Modeling with Synthetic Splicing Dataset

**Introduction:**

Gene splicing is a critical biological process in eukaryotic cells that ensures the generation of functional messenger RNA (mRNA) molecules from the precursor RNA. The primary purpose of gene splicing is to remove non-coding regions (introns) and join the coding regions (exons) to create a mature mRNA molecule that can be translated into a protein.

Splicing is mediated by a highly dynamic and complex spliceosome composed of small nuclear ribonucleoproteins and splicing factors. These splicing factors play pivotal roles in the recognition of splice sites and the assembly and activity of spliceosomes. Aberrations in splicing, often resulting from mutations in splice sites or dysregulation of splicing factors, are linked to various diseases, including cancer and neurogenerative disorders.

This project aims to provide insights into the intricate relationships between splicing factors and the splicing event.

**Methodology:**

It begins with obtaining basic insights into the data and generating summary statistics to understand its structure and characteristics. This is followed by a data-cleaning process, which includes identifying and handling missing values and outliers. The datatypes of variables are examined and modified as needed to facilitate analysis. Pairwise correlations and correlation plots are used to explore the relationships between variables, providing an initial understanding of their interactions. Data visualizations are created to examine data distribution and explore the relationships between splicing factors and the splicing event. The data is split into training and testing sets to develop predictive models. A multivariate linear regression model is constructed, and predictions are generated based on the trained model. Finally, the model's performance is assessed using appropriate evaluation metrics.

**Results:**

**DATA EXPLORATION:**

A table of numbers with black text

Description automatically generated with medium confidence

**Table 1**

The data consists of 5 variables and 100 observations. Table 1 presents the summary statistics for the data. Among the variables, the coefficient of variation is relatively high for the splicing event, suggesting considerable variability in its data. The variable splicing factor 2 exhibits moderate positive skewness, which is statistically significant (skew.2SE=1.3). Additionally, it has heavier tails (leptokurtic) compared to a normal distribution. In contrast, the remaining variables are normally distributed, with slightly lighter tails (platykurtic).

A screen shot of a computer code

Description automatically generated

**Table 2**

Table 2 summarizes the missing values in the dataset, showing that no data points are missing.

A diagram of a graph

Description automatically generated with medium confidence

**Figure 1**

A graph of a number of objects

Description automatically generated with medium confidence

**Figure 2**

Figure 1 illustrates the box plot for the variables, highlighting the presence of outliers in splicing factor 1 and splicing factor 2, whereas Figure 2 represents the distribution of data after the removal of these outliers.

A table of numbers with black text

Description automatically generated with medium confidence

**Table 3**

Table 3 represents the summary statistics after the removal of outliers. We can notice that the coefficient of variation for splicing event decreased significantly. The skewness and kurtosis of splicing factor 2 also showed notable changes, with skewness becoming statistically insignificant and kurtosis shifting to platykurtic from leptokurtic distribution.

**Correlation:**

A number and digits on a white background

Description automatically generated

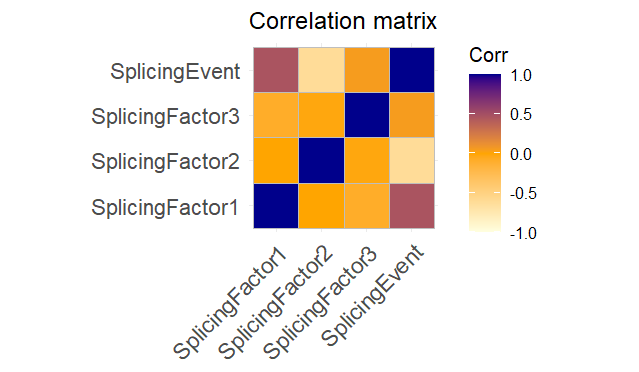
**Table 4**

Table 4 shows the correlation matrix between the variables, indicating that the Splicing event is positively correlated with splicing factor 1 and negatively correlated with splicing factor 2. These relationships are further visualized in the Correlation plot (Figure 3) and the heatmap (Figure 4).

A grid with blue and red circles

Description automatically generated

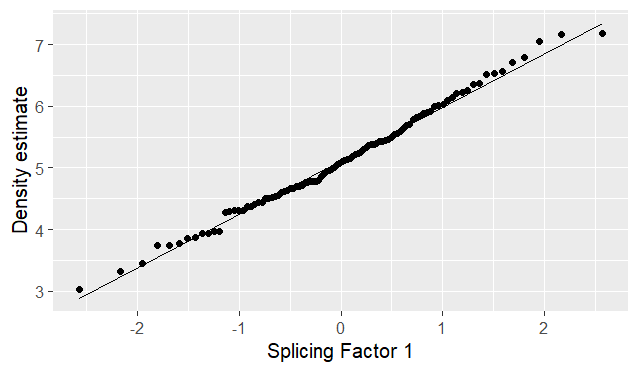
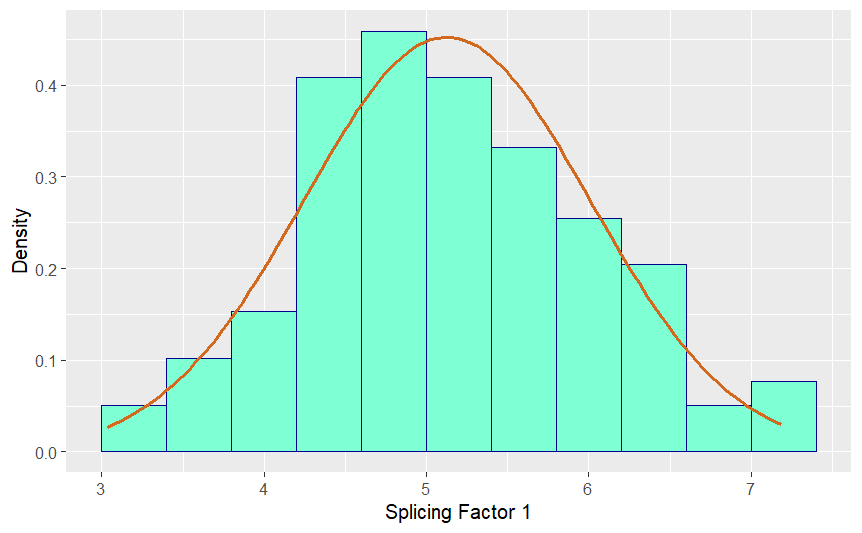
**Figure 3**



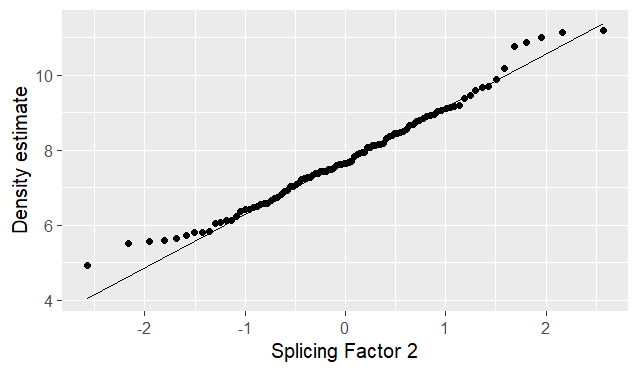
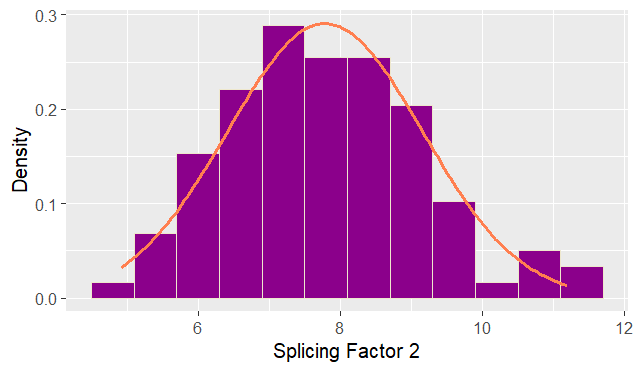
**Figure 4**

**DATA VISUALIZATION:**

**Data distribution:**



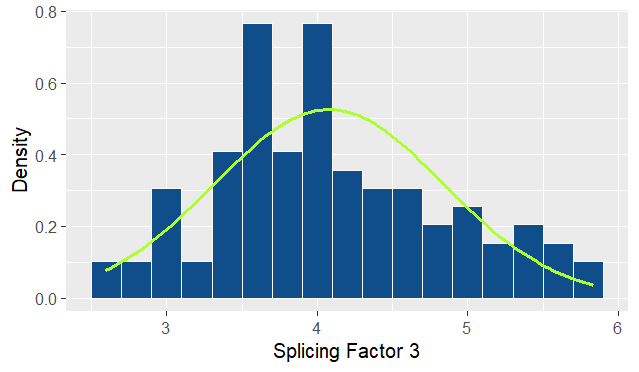
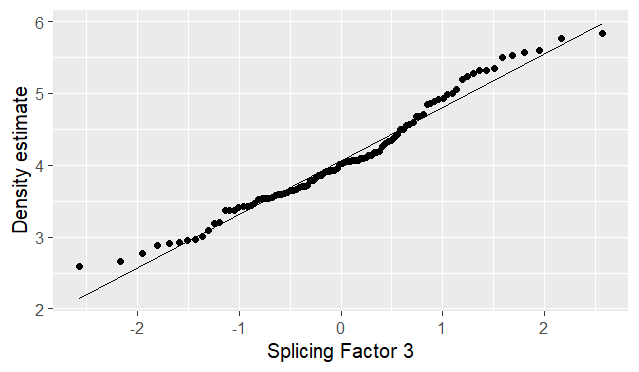
***Figure 5 Figure 6***



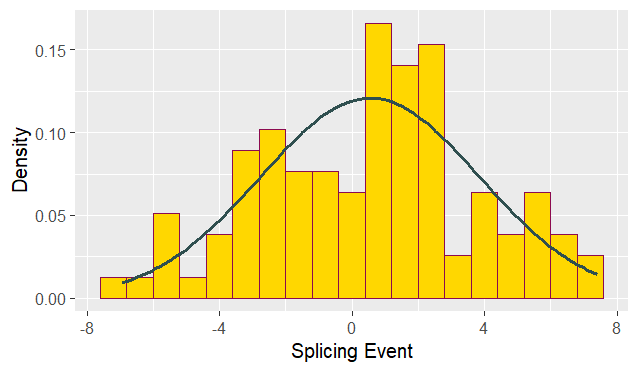
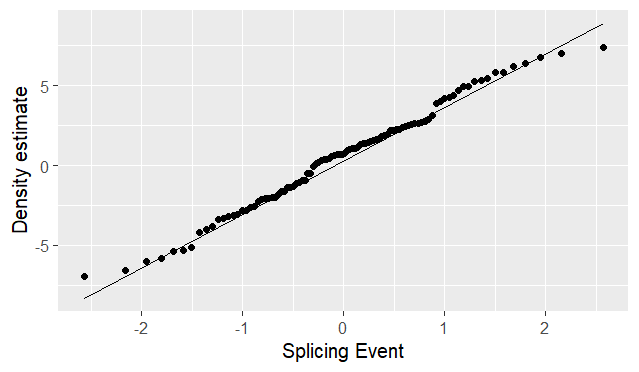
***Figure 7 Figure 8***

Figures 5 and 6 illustrate the distribution of data for splicing factor 1. The histogram in Figure 5, overlaid with a density curve, is symmetric and bell-shaped, indicating that the data follows a normal distribution. The Q-Q plot in Figure 6 further supports this, as most points align closely with the diagonal line, though slight deviations are observed at the tails.

Similarly, Figure 8&9 represents the distribution of data for splicing factor 2. The histogram in Figure 8 is also symmetrical and bell-shaped, indicative of a normal distribution. However, the Q-Q plot in Figure 9 shows inward deviations at the tails, suggesting a platykurtic distribution.

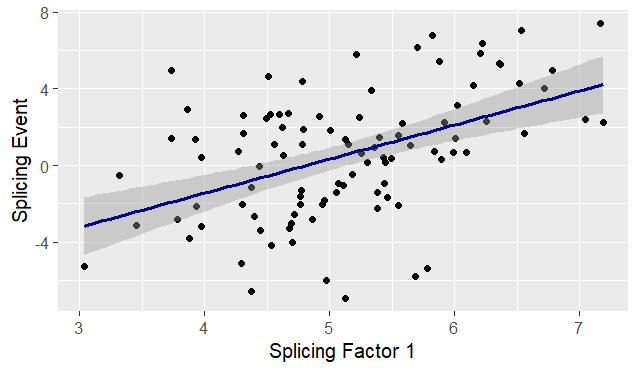
***Figure 9 Figure 10***

***Figure 11 Figure 12***

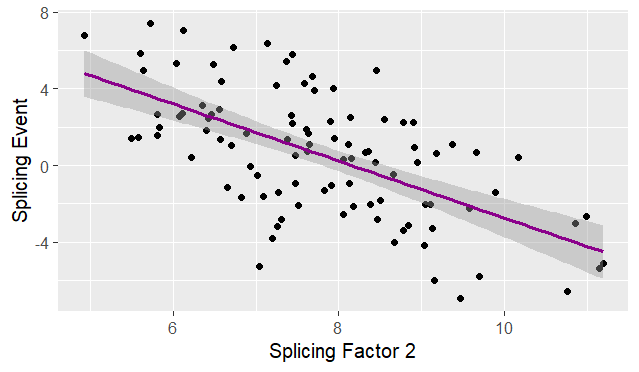
Figures 9 and 10 illustrate the data distribution for splicing factor 3, while figures 11 and 12 represent the data distribution for splicing event. The histograms are symmetric and bell-shaped in both cases, and the Q-Q plots show points closely aligning with the diagonal line. These observations confirm that the data for both variables follow a normal distribution with only slight deviations at the tails.

**Relationship between Splicing factors and Splicing event:**



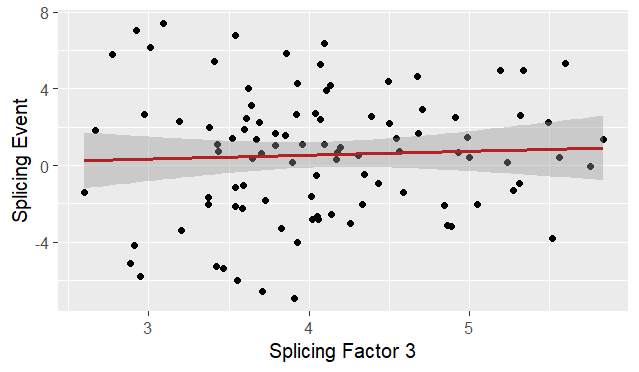
**Figure 13**

Figure 13 illustrates the Scatter plot with a regression line, highlighting a positive linear correlation between splicing factor 1 and splicing event. This indicates that as the splicing factor 1 increases, the splicing event also increases linearly.



**Figure 14**

Figure 14 illustrates the Scatter plot with a regression line, highlighting a negative linear correlation between splicing factor 2 and splicing event. This indicates that as the splicing factor 2 increases, the splicing event also decreases linearly.

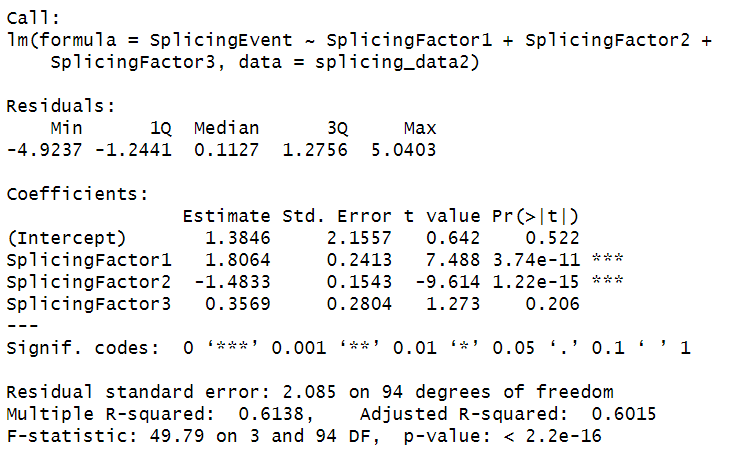
****

**Figure 15**

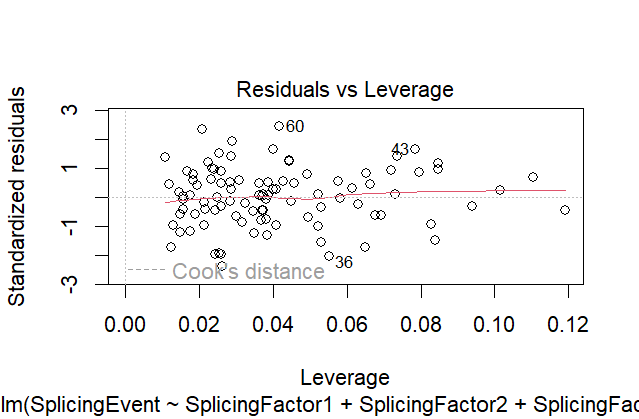
Figure 15 illustrates the Scatter plot with a horizontal regression line, indicating no significant correlation between splicing factor 3 and splicing event. The lack of slope suggests that changes in splicing factor 3 do not affect splicing event.

**PREDICTIVE MODELLING:**

The data was split into train (70%) and test (30%) datasets. Multiple linear regression was performed using a train dataset.

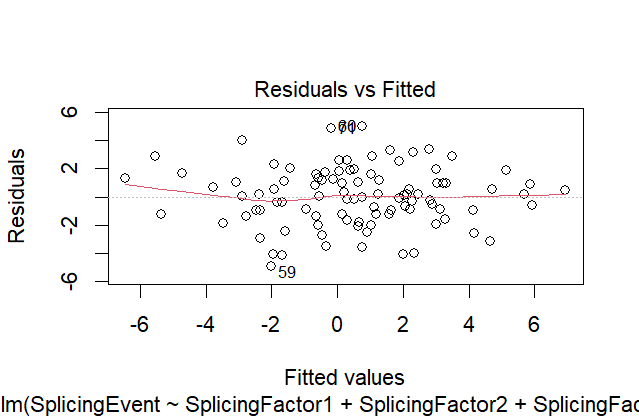


The estimates show that splicing factor 1 is positively correlated with the splicing event, whereas splicing factor 2 is negatively correlated. With the p-value <0.001, these correlations are strongly statistically significant. The high F-value and significance of the p-value also suggest that the model is a good fit. R2 is 0.61, indicating that the predictor variables explain 61% of the variability. The Mean squared error is 4.17, indicating a moderate degree of error in the model’s prediction.



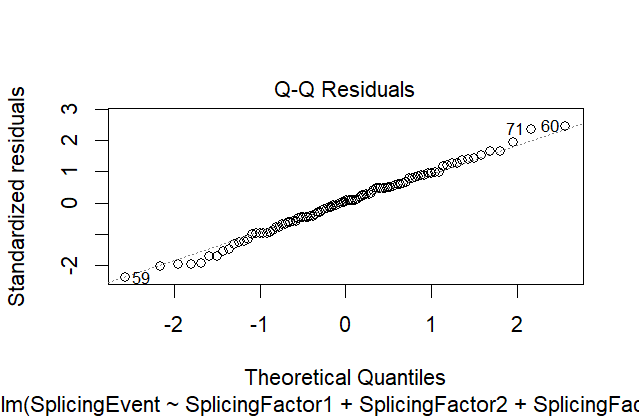
**Figure 16**

Figure 16 plot shows if any influential outliers are present in the data. We can see that no such cases are present in our model.



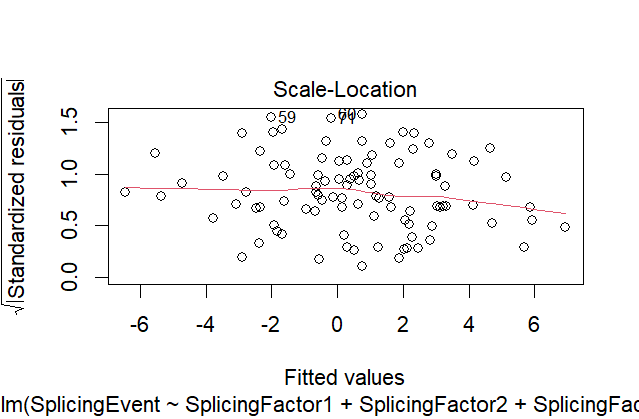
**Figure 17**

Figure 17 shows that data appears homoscedastic, and the residuals are almost linear. From this, we can say our data met the assumptions of linearity and homoscedasticity.



**Figure 18**

Figure 18 illustrates that residuals are normally distributed as the points align with the diagonal line.



**Figure 19**

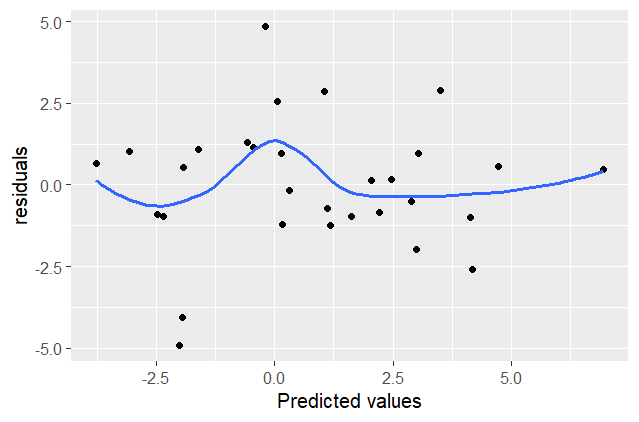
Figure 19 shows the red line, which is flat and horizontal with equally and randomly spread data points. This means the model assumption of homoscedasticity is met.

**Prediction of test data:**

The predicted values of the Splicing event and the residuals were added as new columns in the test data.

The cross-validation correlation between predicted values and true values in the test dataset is 0.80, indicating strong predictive power, with the predicted and actual values being closely aligned in most cases.

The R2 for the test data is 0.64, which is almost similar to the 0.61 of the training dataset. The shrinkage (R2 of train data - R2 of test data) is -0.032 in R2, reflects a small reduction in the model’s predictive power, suggesting the model remains effective. The mean of residuals is 0.0037, close to zero, indicating a good model fit, while the mean squared error is 3.81, which shows a slight improvement over the previous model.



**Figure 20**

The scatter plot (Figure 20) of residuals versus predicted values shows a horizontal linear pattern around zero. This is a good sign, suggesting that the model has captured the underlying relationships well and that no obvious patterns are left in the errors.

**CONCLUSION:**

Among the three splicing factors, Splicing Factor 1 is positively associated, whereas Splicing Factor 2 is negatively associated, and Splicing Factor 3 does not correlate with the splicing event. Similarly, a strong positive correlation of 0.80 was observed between predicted and true values in the cross-validation, indicating good model performance. Additionally, a minimal loss of predictive power was observed when moving from training to test data set. However, the mean residuals close to zero (0.0037) and a slight decrease in mean squared error (3.81) suggest the model is reasonably accurate and well-fitted.

No noticeable difference in the model's performance was observed when retaining the outliers in the data.

For future recommendations, consider including interacting variables like environmental factors and additional splicing factors that may further explain the variability in the splicing event.