Gene Expression Analysis Using 2-Way ANOVA

Kanishk Aman (22237)
Indian Institute of Science (IISc), Bangalore

Abstract

This report presents an analysis of gene expression data using a two-way ANOVA framework. The goal is to identify genes that exhibit differential responses to smoking based on gender. We explore the interaction effect between 'Smoking Status' (Smokers and Non-Smokers) and 'Gender' (Males and Females), using two-way ANOVA.

Contents

1 Introduction

Gene expression analysis is a powerful tool in genomics to understand the biological mechanisms underlying various conditions. In this study, we aim to identify genes that respond differently to smoking in males and females using a two-way ANOVA with an interaction effect ('Smoking Status x Gender'). The **Null hypothesis** assumes no interaction effect between smoking status and gender, while the **Alternative hypothesis** suggests a significant interaction.

2 Data Description and Preprocessing

The dataset consists of gene expression values for 48 samples split into 4 groups:

- **Male Non-Smokers** (12 samples)
- **Male Smokers** (12 samples)
- **Female Non-Smokers** (12 samples)
- **Female Smokers** (12 samples)

The provided gene expression data was in logarithmic form, so an exponentiation transformation was applied to convert the values back to the original scale. I have also written a line of code (commented) in my python script, which gives the result without the use of converted values.

3 Methodology

3.1 Two-Way ANOVA Framework

The analysis uses a two-way ANOVA framework with matrix-based rank operations to detect interaction effects between the two factors ('Smoking Status' and 'Gender'). Two design matrices were constructed:

1. **Additive Model** (N): Considers the main effects of 'Smoking Status' and 'Gender'.
2. **Interaction Model** (D): Extends the Additive Model by incorporating an interaction term.

3.2 Mathematical Framework

Given the response vector $\mathbf{Y}$ (gene expression values), the matrices are defined as follows:

$$N = \begin{bmatrix} 1 & S_i & G_i \end{bmatrix}, \quad D = \begin{bmatrix} 1 & S_i & G_i & S_i G_i \end{bmatrix}$$

Where:

- S_i is the 'Smoking Status' (0 = Non-Smoker, 1 = Smoker).

- G_i is the 'Gender' indicator (1 = Male, 0 = Female).

The hypothesis tests are performed using the difference in residual sum of squares:

$$\text{SS}_{\text{Interaction}} = \text{trace}\left((\mathbf{P_D} - \mathbf{P_N})\mathbf{Y}\mathbf{Y}^T\right)$$

$$\text{SS}_{\text{Residual}} = \text{trace}\left((\mathbf{I} - \mathbf{P_D})\mathbf{Y}\mathbf{Y}^T\right)$$

Where $\mathbf{P_N}$ and $\mathbf{P_D}$ are the projection matrices for N and D.

4 Implementation

The implementation is divided into several steps:

1. Loading & preprocessing the data using exponentiation to revert the logarithmic scale.
2. Constructing the design matrices N and D for additive and interaction models.
3. Calculating the projection matrices $\mathbf{P_N}$ and $\mathbf{P_D}$.
4. Looping through each gene's expression probe, computing the F-statistic, and deriving the p-values.

5 Results and Conclusion

5.1 Histogram of Interaction p-values

The histogram of the interaction p-values is shown in the given Figures. It indicates the distribution of p-values for the interaction effect between Smoking Status and Gender. I have attached both the results I got, (with and without using exponentiated values) while running my code.

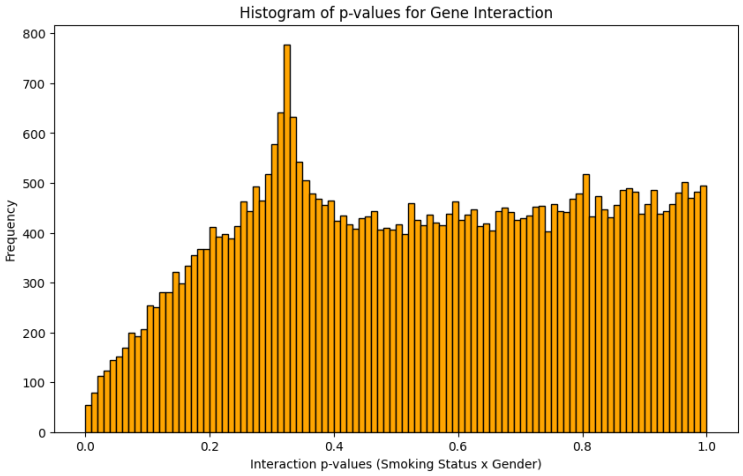

Figure 1: Histogram of p-values (using preprocessed values)

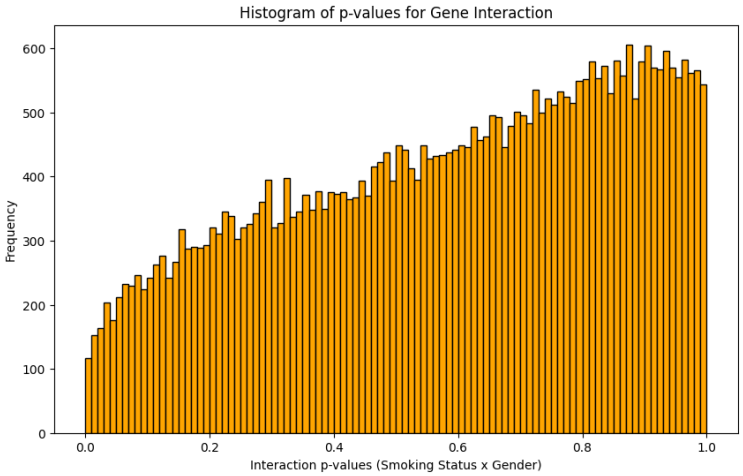

Figure 2: Histogram for the same, but without using exponentiated values

5.2 Conclusion

The histogram of p-values suggests that most genes do not show a strong gender-specific response to smoking. The spike around p-values of 0.3–0.4 indicates some genes may exhibit subtle interaction effects, though not highly significant. Further analysis could focus on these genes for more detailed investigation.