Assignment 3: The Effect of Smoking

Tarun Kumar Sahu
SR No. 23156

September 30, 2024

1 Introduction

This report presents a mathematical analysis of the gene expression data using a two-way ANOVA framework. The aim is to investigate the effects of **gender** and **smoking status** on gene expression levels. We further analyze how significant genes are identified and how they intersect with predefined gene lists.

2 Dataset Description

The dataset used for the analysis consists of gene expression data generated from the white blood cells of 48 individuals. The key features of the dataset are as follows:

- The dataset contains a total of 48 columns of data, each representing an individual, along with some auxiliary columns.
- Auxiliary columns included: Probe Name, Gene Symbol, Entrez Gene ID.
- Each gene is identified by a Gene Symbol or Entrez Gene ID, and a single gene can have multiple probes associated with it.
- The dataset contains a total of 41,094 probes.
- The 48 data columns are organised as follows:
    1. **12 Male Non-Smokers**: Columns B-M (Columns 2-13).
    2. **12 Male Smokers**: Columns N-Y (Columns 14-25).
    3. **12 Female Non-Smokers**: Columns Z-AK (Columns 26-37).
    4. **12 Female Smokers**: Columns AL-AW (Columns 38-49).

3 Gene Expression Analysis Using Two-Way ANOVA

The two-way ANOVA is a statistical method used to assess the impact of two independent variables on a dependent variable. In this case, the independent variables are **gender** (male and female) and **smoking status** (smoker and non-smoker), while the dependent variable is gene expression.

For each gene, the data is divided into four groups:

- Male Non-Smokers (Columns B-M)
- Male Smokers (Columns N-Y)
- Female Non-Smokers (Columns Z-AK)
- Female Smokers (Columns AL-AW)

3.1 Mathematical Formulation of Two-Way ANOVA

Let X_{ijk} represent the expression level of the i-th individual in the j-th gender group ($j = 1, 2$ for male, female) and k-th smoking group ($k = 1, 2$ for non-smoker, smoker). The total number of observations per gene is $n = 48$.

We decompose the variation in gene expression into:

- **Main effect of gender**, measured by the difference between the mean gene expression in males and females.
- **Main effect of smoking**, measured by the difference between the mean gene expression in smokers and non-smokers.
- **Interaction effect**, which measures whether the effect of one factor (gender) changes depending on the level of the other factor (smoking status).

The total variation in gene expression is captured by the **total sum of squares** (SS):

$$SS_{\text{Total}} = \sum_{i,j,k} (X_{ijk} - \mu_{\text{Grand}})^2$$

where μ_{Grand} is the overall mean of the gene expression data across all groups.

3.2 Between-Groups Sum of Squares (Main Effects)

The main effects are calculated by partitioning the total variation into **between-group** variation due to gender and smoking status.

Gender Effect: The between-group sum of squares for gender (SS_A) measures how much the gene expression varies between males and females:

$$SS_A = n_B \cdot \left((\mu_{\text{Male}} - \mu_{\text{Grand}})^2 + (\mu_{\text{Female}} - \mu_{\text{Grand}})^2 \right)$$

where n_B is the number of observations in each smoking group, and μ_{Male}, μ_{Female} are the mean gene expression levels for males and females, respectively.

Smoking Effect: Similarly, the between-group sum of squares for smoking (SS_B) measures how much the gene expression varies between smokers and non-smokers:

$$SS_B = n_G \cdot \left((\mu_{\text{Non-Smoker}} - \mu_{\text{Grand}})^2 + (\mu_{\text{Smoker}} - \mu_{\text{Grand}})^2 \right)$$

where n_G is the number of observations in each gender group, and $\mu_{\text{Non-Smoker}}$, μ_{Smoker} are the mean gene expression levels for non-smokers and smokers, respectively.

3.3 Interaction Effect Sum of Squares

The interaction between gender and smoking status is measured by the interaction sum of squares ($SS_{\text{Interaction}}$). This effect captures how much the effect of gender depends on smoking status and vice versa:

$$SS_{\text{Interaction}} = \sum_{i,j,k} \left(\mu_{jk} - \mu_j - \mu_k + \mu_{\text{Grand}} \right)^2$$

where μ_{jk} is the mean gene expression for the j-th gender and k-th smoking group, and μ_j, μ_k are the marginal means for gender and smoking, respectively.

3.4 Within-Groups Sum of Squares (Error Term)

The within-groups sum of squares (SS_{Within}) captures the variability within each group that is not explained by the main effects or interaction. It is computed as:

$$SS_{\text{Within}} = \sum (X_{ijk} - \mu_{jk})^2$$

where μ_{jk} is the mean gene expression within each of the four groups.

3.5 Mean Squares and F-Statistics

For each effect (gender, smoking, interaction), we calculate the mean square (MS) by dividing the sum of squares by the corresponding degrees of freedom (df):

$$MS_A = \frac{SS_A}{df_A}, \quad MS_B = \frac{SS_B}{df_B}, \quad MS_{\text{Interaction}} = \frac{SS_{\text{Interaction}}}{df_{\text{Interaction}}}, \quad MS_{\text{Within}} = \frac{SS_{\text{Within}}}{df_{\text{Within}}}$$

where $df_A = df_B = 1$, and $df_{\text{Interaction}} = 1$, and $df_{\text{Within}} = n - 4$ (the number of groups subtracted from the total number of observations).

The F-statistics for each effect are computed as:

$$F_A = \frac{MS_A}{MS_{\text{Within}}}, \quad F_B = \frac{MS_B}{MS_{\text{Within}}}, \quad F_{\text{Interaction}} = \frac{MS_{\text{Interaction}}}{MS_{\text{Within}}}$$

The p-values for gender, smoking, and interaction are derived from the F-distribution, indicating the significance of each effect.

4 Results and Visualizations

After executing the Python code, 58 plots will be generated in the ‘plots‘ folder:

- **57 Gaussian distribution plots**: These visualize the expression levels of the 57 significant genes identified from the intersection with biological pathways.
- **1 Histogram plot**: This displays the overall distribution of p-values for the gender, smoking, and interaction effects.

Please refer to the ‘plots‘ folder for these visualizations, which provide detailed insights into the gene expression analysis.

4.1 Significant Genes and Gene List Intersection

The number of significant genes identified was 3860. These genes were intersected with several predefined gene lists to investigate their involvement in specific biological pathways. The results of the intersection are as follows:

- **XenobioticMetabolism1.txt**: 18 intersecting genes

  {’HNF4A’, ’CYB5R3’, ’CYP2F1’, ’CES3’, ’CYP2C8’, ’AADAC’, ’AHRR’, ’CYP2S1’, ’ACY3’, ’CYP2A6’, ’UGT2B15’, ’SULT1A1’, ’GSTM4’, ’AS3MT’, ’AOC2’, ’DPEP1’, ’CES2’, ’CYP1A2’}

- **FreeRadicalResponse.txt**: 3 intersecting genes

  {’MPO’, ’UCP2’, ’BMP7’}

- **DNARepair1.txt**: 11 intersecting genes

  {’ABL1’, ’TNP1’, ’XAB2’, ’CIB1’, ’PNKP’, ’RAD9A’, ’POLE’, ’OGG1’, ’GADD45G’, ’SMC1A’, ’RAD51B’}

- **NKCellCytotoxicity.txt**: 25 intersecting genes

  {’MAP2K2’, ’PIK3R3’, ’KLRC2’, ’KIR2DS4’, ’ULBP2’, ’IFNG’, ’HLA-E’, ’CSF2’, ’CD244’, ’HLA-G’, ’KIR3DL2’, ’KIR2DL4’, ’KIR2DS1’, ’PRF1’, ’SH2D1B’, ’KIR2DL5A’, ’IFNA6’, ’SHC1’, ’PTPN6’, ’KIR3DL1’, ’KIR2DL2’, ’SH3BP2’, ’GZMB’, ’ULBP1’, ’HLA-C’}

The total number of unique intersecting genes across all lists was found to be 57.

5 Conclusion

The two-way ANOVA analysis revealed that several genes show significant expression changes based on gender, smoking status, and their interaction. A total of 3860 significant genes were identified, with 57 unique genes intersecting with various biological pathway gene lists. Further analysis of these genes could provide insights into their role in smoking-related health conditions.