Assignment 3: Effects of Smoking

Anushka Amte *M.Tech CSA* 22486

Abstract

ANOVA, or Analysis of Variance, is a test used to determine differences between research results from three or more unrelated samples or groups. The goal is to identify genes that show differential responses to smoking exposure between men and women by comparing the interaction between Smoking Status and Gender. I perform hypothesis testing by using F statistics to calculate the p values.

1. Data

We have gene data of around 40,000 men and women. The data is in log form due to some extremely small values, so the first step is to exponentiate the data to the log base to make it linear again.
I consider the following 48 samples for our hypothesis testing:

- 12 Male Non-smokers (106-117)
- 12 Male Smokers (118-129)
- 12 Female Non-Smokers (130-141)
- 12 Female Smokers (142-153)

This ensures that the data distribution for the sampling is not skewed towards one group.

2. Hypothesis Modelling

Now, I model out two hypotheses: Null and Alternate.
The goal here is to find whether the effects of smoking are related to the gender of a person, for which gene data is available.

- **Null Hypothesis:** The effects of smoking and gender are independent
- **Alternate Hypothesis:** The effects of smoking and gender are correlated

There are four underlying distributions here over which I am modeling my hypothesis:

- Male + Smoker
- Non-Smoker + Male
- Female + Smoker
- Female + Non-Smoker

The Null hypothesis tells us that the means of these four distributions will be additive, i.e. the mean of the Male + Smoker distribution can be given as $\mu_{Male} + \mu_{Smoker}$.
Whereas the alternative hypothesis says that these means may be arbitrary.

Assumption: I assume that all the four distributions have similar variance.

3. Constructing Matrices

Let the null hypothesis means be denoted by $\mu_m, \mu_f, \mu_s, \mu_{ns}$ i.e. Male, Female, Smoker, Non-smoker respectively.
The alternative hypothesis means be $\mu_{ms}, \mu_{mns}, \mu_{fs}, \mu_{fns}$.
I now construct two matrices $A_{null} \in \{0,1\}^{48x4}$ and $A \in \{0,1\}^{48x4}$ to model these two hypotheses. The four columns are basically the means that I am estimating here. According to our samples, I set the values of the corresponding columns of A and A_{null} as 1. This means for rows 1-12 in A_{null} the columns of Male and Non Smoker will be 1 and so on and so forth.
For the A matrix or the design matrix the Column of Male Smoker will be 1 for the rows 12-24.

4. Estimating Values

The null hypothesis can be posed as a linear regression problem, where the means can be estimated by it's closed-form least squares solution.(University, n.d.)

$$\min_y (X - A_{null}y)^T (X - A_{null}y) \tag{1}$$

where X is the input features and y are the means. The closed form solution to y is given as:

$$numerator = I - \left(A_{\text{null}} \cdot \left(A_{\text{null}}^T A_{\text{null}} \right)^+ \cdot A_{\text{null}}^T \right)$$

[+ is the pseudo inverse] as the rank of A_{null} is 3.
Similarly, we can compute this for the alternate hypothesis means as follows:

$$\min_{y}(X - Ay)^T(X - Ay) \tag{2}$$

$$denominator = I - \left(A \cdot \left(A^T A\right)^+ \cdot A^T\right) \tag{3}$$

The rank of A is 4.
The F statistics is finally calculated as,

$$F_{stats} = scalingfactor * \left(\frac{numerator}{denominator} - 1\right)$$

$$ScalingFactor = \frac{48 - rank(A)}{rank(A) - rank(A_{null})}$$

4.1. p Values

The p values are given by the cumulative distribution function (CDF) of the F-distribution: We need the right-sided distribution, there we do as,

$$p = 1 - CDF(F_{stats}) \tag{4}$$

5. Results

The final results I obtain after calculating the p-values for all the rows is presented in the below histogram 1.

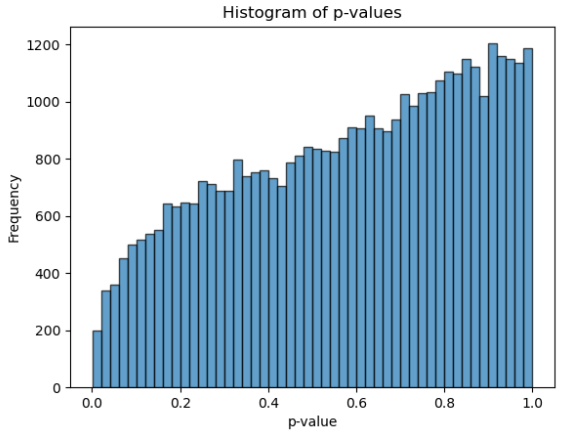

Figure 1. Histogram

I also saved the rows whose p-values were less than 0.05, there were 702 such ”interesting” rows.

6. Conclusion

The plot generated shows that majority of the rows had p values closer to 1. This tells us that there is no strong evidence for widespread differential gene expression based on the interaction between smoking status and gender.The high p-values imply that, for most genes, the differences observed are likely due to random variation rather than a true interaction effect. As a result, only a small subset of genes may exhibit meaningful differences in expression based on smoking status and gender.

References

University, S. Anova, n.d. URL https://web.stanford.edu/class/stats191/Chapter13/ANOVA.html#/title-slide. Accessed: 2024-09-29.

A. Appendix

The zip file contains a interesting_rows.csv which contains rows whose p-values where ¡ 0.05.
Some parts of the code were generated using Github Copilot.