Data Analytics - Assignment #3

Nishant Kumar (Sr.No: 21495)

September 2023

Effect of Smoking

1 Part 1:

To identify genes that respond differently to smoking in men versus women, you can perform a statistical analysis comparing gene expression data between these two groups while taking into account the interaction between smoking status and gender. This analysis involves a two-way analysis of variance (ANOVA) or regression analysis. (Smoking Status X Gender model vs. Smoking Status + Gender null model).
We computed A and A′ using:

$$h = AB + Error \tag{1}$$

h : gene expression
A : Alternative Hypothesis
A' : Null Hypothesis
B : Mean Vector

Alternative Hypothesis A :

$$\begin{bmatrix} h_1 \\ h_2 \\ . \\ . \\ . \\ . \\ h_{48} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ . & & & \\ . & & & \\ . & & & \\ . & & & \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} male_{nonsmoker} \\ male_{smoker} \\ female_{nonsmoker} \\ female_{smoker} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_{48} \end{bmatrix}$$

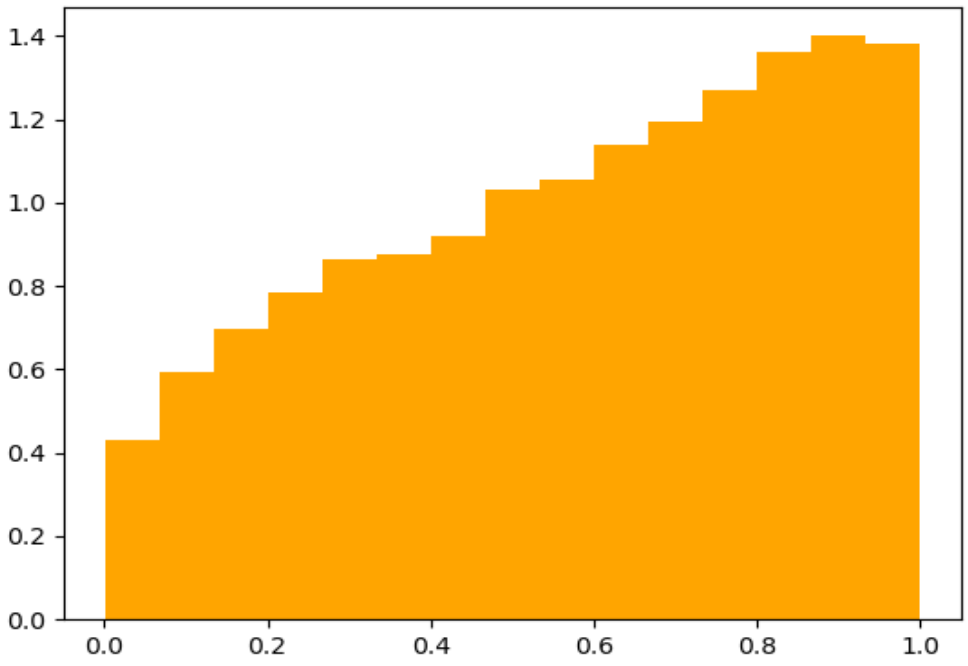
Null Hypothesis A':

$$\begin{bmatrix} h_1 \\ h_2 \\ . \\ . \\ . \\ . \\ h_{48} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ . & & & \\ . & & & \\ . & & & \\ . & & & \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} male \\ female \\ nonsmoker \\ smoker \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ . \\ \epsilon_{48} \end{bmatrix}$$

F-Statistic Computed by using formula below:

$$\text{F-statistic } \hat{f} = \frac{\vec{\hat{h}}^T (A(A^T A)^\dagger A^T - A'(A'^T A')^\dagger A'^T)\vec{\hat{h}}}{\vec{\hat{h}}^T (I - (A(A^T A)^\dagger A^T)\vec{\hat{h}}} * \frac{n - rank(A)}{rank(A) - rank(A')}$$

2 Part 2: Draw the histogram of p-values:



3 Part 3: Use an FDR cut-off of 0.05 to shortlist rows:

False Discovery Rate: [1.10554044e-04 1.31050111e-04 3.01869021e-04 ... 1.00000000e+00 1.00000000e+00 1.00000000e+00]

4 Part 4:Create a shortlist of gene symbols from these rows:

Refer genes-symbol-list.txt file

5 Part 5: Intersect with the following gene lists:

Xenobiotic metabolism, *Free Radical Response*, *DNA Repair*, *Natural Killer Cell Cytotoxicity.*

Genes Symbol Intersection with Xenobiotic ::
['CYB5R3', 'AOC1', 'SULT1A1', 'GRIN1', 'AS3MT']

Genes Symbol Intersection with Free Radical Response ::
['NFE2L2', 'DHFR', 'ADPRHL2']

Genes Symbol Intersetion with DNA Repair::
['MSH3', 'HMGB1', 'RAD17']

Genes Symbol Intersetion with NKCellCytotoxicity ::
['SHC2', 'PRKCB']

6 Part 6: Numbers of Intersection Counts:

Numbers of Genes that are intersecting with Xenobiotic Metabolism:: 5

Numbers of Genes that are intersecting with free Radical Response:: 3

Numbers of Genes that are intersecting with DNA Repair:: 3

Numbers of Genes that are intersecting with NKCellCytotoxicity:: 2

7 Part 7: Groups are as follows:

Female Smokers Up Genes::
{'GRIN1', 'CYB5R3', 'HMGB1', 'AOC1', 'PRKCB', 'SHC2', 'RAD17', 'MSH3'}

Female Smokers Down Genes::
{'GRIN1', 'SULT1A1', 'PRKCB', 'AS3MT'}

Male Smokers Up Genes::
{'GRIN1', 'AS3MT', 'HMGB1', 'SULT1A1', 'RAD17', 'MSH3'}

Male Smokers Down Genes::
{'CYB5R3', 'AOC1', 'PRKCB', 'SHC2'}