Assignment 2
Cheyanne Durham
Analysis Report
Github Link: https://github.com/cheyannewdurham/CPBS7602

Data

The data for this assignment is a csv file containing transcript parts per million. This also incorporates the meta data of subject phenotypes (age and sex) into the analysis.

Data Preparation

The data was scaled using scikit preprocessing's standard scaler around 0. One hot encoding was used for the Sex (male or female) values to make them boolean values, tissue and age were recorded as ordinals with numbers representing the decades they were taken from. Samples with missing values for sex and age were other retained, giving us a total of 881 samples to work with.

Linear Regression Results

This first analysis focuses on analyzing how well age, tissue (blood or brain), and sex predict gene expression. When sorting by, DCLK2, PLEK, MIR99AHG, EPN2, and NCAM1 all had extremely low p-values (e -189 to -236) and a reasonably large effect size considering the scale. All genes besides PLEK were positively correlated with the predictors (present in brain tissue or heavily influenced by sex and age) and PLEK was negatively correlated with the predictors (likely due to the fact it is present in blood). MIR99AHG, for example, is commonly over expressed in the nervous system and digestive system and under expressed in other systems (respiratory, reproductive) during cancer progression. By effect size, PLEK is present again with SIT1, ARAP1, P2RY10, and BCLA2A1. PLEK is a platelet gene that likely would be expressed more in the blood, explaining it's negative correlation to brain tissue. Many of the top genes are related to cytoskeletal organization. The extremely low p-values indicate a very high confidence and do cause some concern regarding overfitting. Given the nature of this data, oversampling of certain sexes or patients with particular pathologies may present a challenge (possibly data leakage) in the logarithmic regression.

In conclusion, the gene expression seemed to successfully cluster by tissue type being the main predictor (due to the nature of the reported genes).

Logarithmic Regression Model

To determine to performance of the logarithmic model, testing data was used. According to the classification report for male predictions, the model has high precision (0.98) which is the true positive predictions out of all positive predictions; high recall (minimal false negatives); F1-Score (0.91) which shows a good balance between precision and recall through the mean. For the female predictions, we have a 0.86 precision, 0.96 recall, and 0.91 F1-score. Since the males were overrepresented in the population, the lower precision score for females makes sense. To account for the diverse feature set, the training sets were stratified and the L1 normalization approach was used along with balanced class weighting to penalize the over-representation of one sex or other feature class in the data set.

Logarithmic Regression Results

The results from the model are a bit suspicious. While our p-values and effect sizes appear reasonable, the coefficients of 0 are potentially an artifact of the logarithmic model. It makes it very difficult to determine validity. However, in both p-value significance and effect size, some of the top genes were Y-linked (KDM5D, KINC00278 specifically). UC_338 was a very interesting find as well due to it's connections colorectal cancer. This cancer has a slightly higher incidence rate in men and it is possibly that our sample set had underlying patters of this patient population.

In conclusion, it is to immediately apparent which features cause the clustering for some genes, but many of them are Y-linked. This indicates our model had a stronger performance, but there is likely some underlying confounding patters in the data preventing a better performance (since we got some apparently non-sex-linked genes.