# Gene Expression in ALS Patients

Sean Coursey

3/8/2022

## Introduction

This project was completed for the *Data Science: Capstone* course in HarvardX's *Data Science Professional Certificate* program on edX. The general goal of this project was to use data sets published on Kaggle by the Answer ALS Research Program to analyze whether there is a significant difference in gene expression in patients with ALS and between ALS patients with differing ALS symptoms, and to model such a difference if it existed. The datasets are publicly available through the End ALS Kaggle Challenge, they include classification information about the patients along with gene expression data in the form of transcript counts normalized via DEseq to account for cross-trial variation. Working with this data proved difficult because of the small number of rows but large number of columns–the data simultaneously resisted providing statistically strong evidence because of the few test subjects while also making code run slowly because of the huge number of columns. Despite these difficulties, this author considers the analysis a mild success and concludes that it provides considerable, though not conclusive, evidence for a systematic difference in gene expression both between control patients and ALS patients and between ALS patients with differing onset sites, and that it provides at least one meaningful model of these differences.

## Methods

The methods used for analyzing the control versus case data and the bulbar versus limb onset site data were very similar, so this report will describe their methods simultaneously and later differentiate their results. For each dataset, the patient status was encoded as a binary factor (control versus case or bulbar versus limb) and the expressions of 53,859 genes were provided as normalized transcript counts. The datasets had a similar number of patients, with the control versus case dataset having 163 and the bulbar versus limb onset site dataset having 112. The datasets also provided a patient identification number, but this was not used in the analysis.

The first step in cleaning the data was to remove the columns for any genes whose transcriptions were were recorded, on avereage, fewer than 10 times. These genes were considered as practically not transcribed at all and so not interesing. Next, the patient ID was removed. Than the data was partitioned into training and testing sets to prepare for investigation and model creation.

The next step was to investigate for genes which were expressed differently between the two patient status groups. Using the train set, a Z-test was performed on the expression distributions of each gene for the two patient status groups. The Z-test value for two distributions is described by the equation:

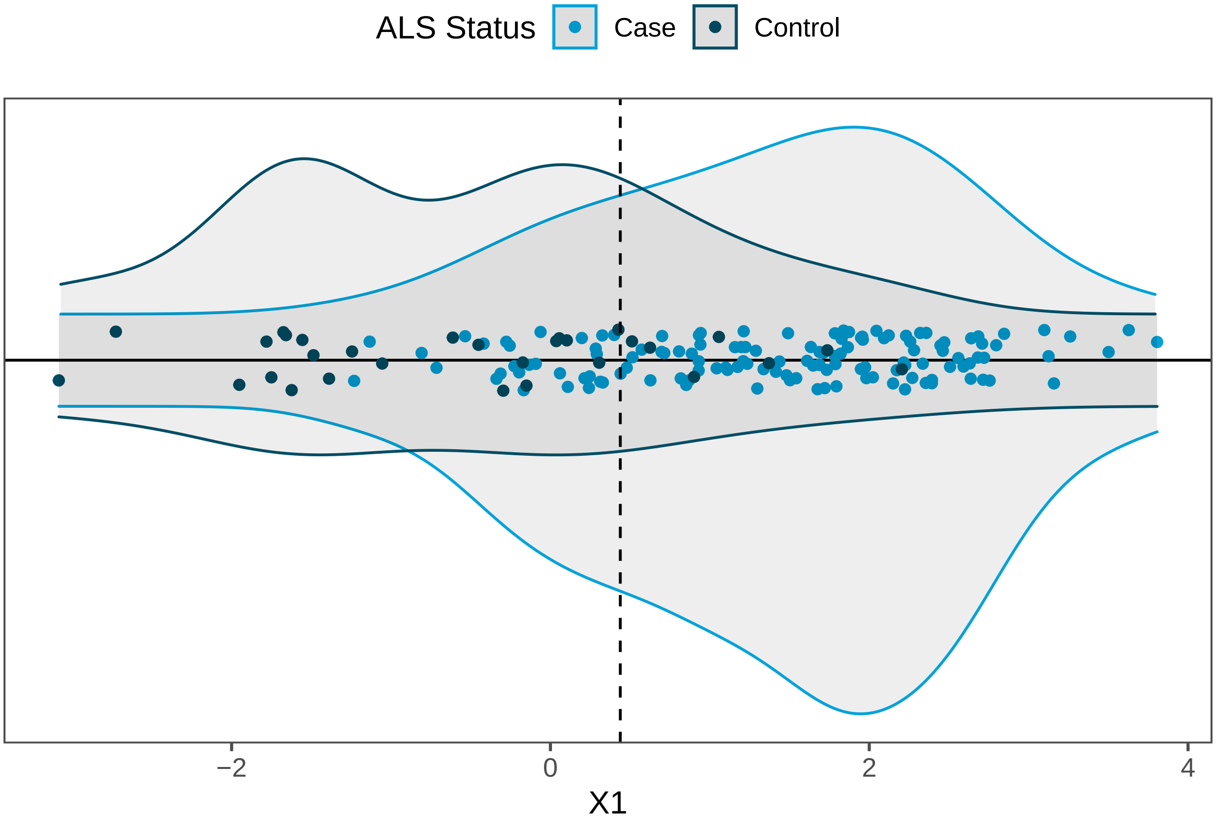
$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

With low values indicating that the two distributions are similar or identical while high values indicate significant difference between the distribution means. These Z-test values were used to filter the data, and progressively more extreme filters were used while looking at the t-SNE and PCA graphs they produced to see if any grouping was visually apparent. A t-SNE plot is a plot designed to maintain proximity (though not distance) relations in high dimensional data in two dimensions while a PCA plot analyzes the two most important axes in the data distribution. Picking the top 32 genes produced clearly distinct distributions for the patient status groups in both the t-SNE and PCA plots.

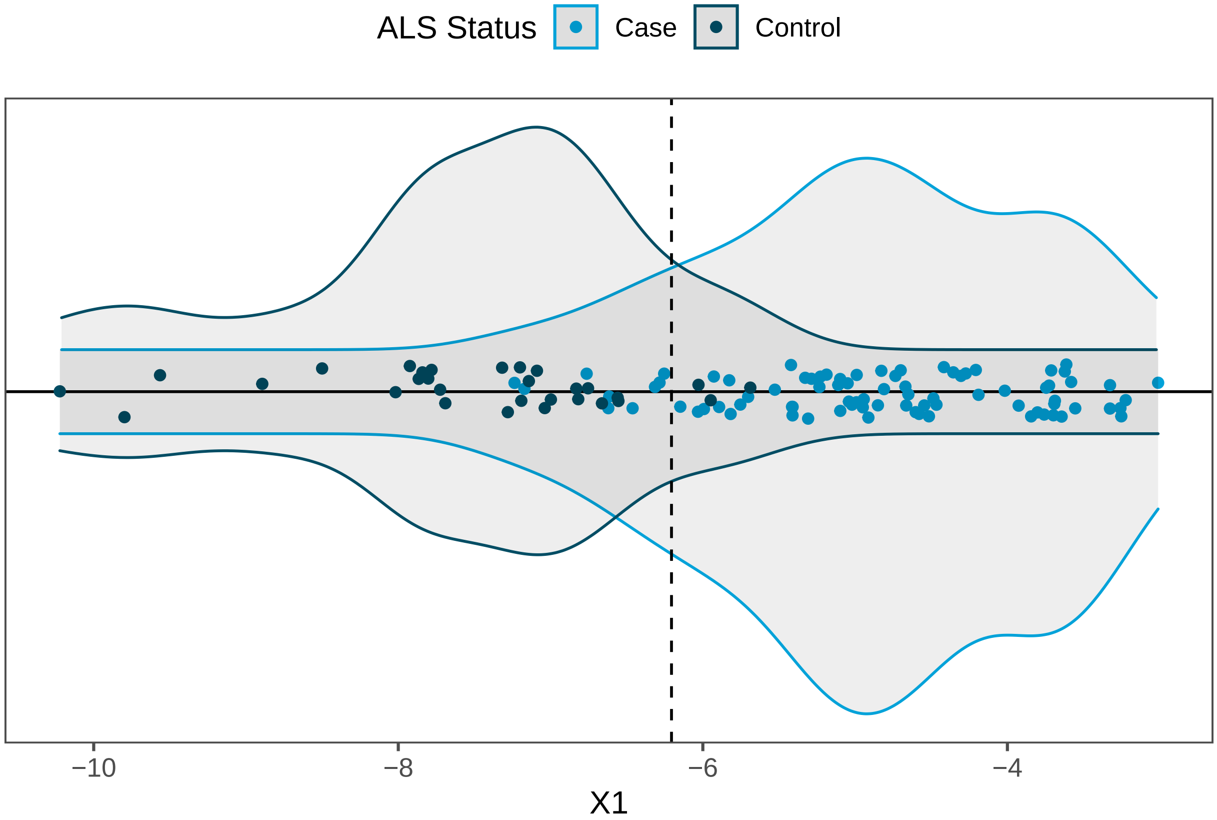
Two models were made for each data set–a linear discriminant analysis model and a classification tree model. These models were chosen for their comprehensibility. Considering the minor actual predictive power of possible models because of the few patients in the datasets, comprehensibility was chosen as more important than accuracy so that the models could hopefully still provide some insight. Initially there was a great difference between the specificity and sensitivity in the models due to an imbalance in each dataset between the two patient status groups (for example, there was an approximately 1:5 ratio of controls to cases). A weighted training set was made adding enough synthetic rows to even the balance of the two status groups. These rows were made using using randomized combinations of gene observations from the minority status group in the training set. Using weighted training data fixed the major imbalance between the sensitivity and specificity of the models which was initially observed. After training the models, a Monte Carlo method was defined for carrying out the whole process, from calculating Z-test values to creating models, on random data and seeing how the models so derived compared in balanced accuracy to the models developed on the actual data.

## Results

The linear discriminant analysis for the control versus case data achieved a balanced accuracy (the average of the sensitivity and specificity) of 0.8375. Using the Monte Carlo method with 100 random trials indicates that this level of balanced accuracy could be expected from a model derived from random data only around nine percent of the time–with a confidence interval of around ten percent. This effectively argues that there is a four-fifths probability that the difference in gene expression between the control and case groups is non-random–indicating some underlying correlation. Below is a graph of the control versus case data plotted along the axis of the linear discriminant analysis model. The jittered dots are the individual data points, the plot above the x-axis is the normalized density of each group (the density if there were equal numbers of patients in each status group) and below the x-axis is the unnormalized density (or count data). The vertical line divides the data into the two decision groups for the linear discriminant analysis model.



The linear discriminant analysis for the bulbar versus limb onset site data achieved a balanced accuracy of 0.9282. The Monte Carlo method indicated a p-value of zero percent with a confidence interval of fourteen percent, suggesting an 86% probability that the difference in gene expression between those with bulbar versus limb onset is non-random. Below is a graph of the bulbar versus limb onset data plotted along the axis of the linear discriminant analysis model. It is structured similarly to the plot above.



Unfortunately, the classification tree models proved insignificant. The Monte Carlo analysis indicated a p-value of 62% for the classification tree made for the control versus case data, with the bulbar versus limb onset data performing similarly poorly. While more accurate methods, like random forest methods, could easily provide better accuracy, the failure of the classification tree models is dissapointing because of their potential to illuminate specifically which genes in what combination indicate patient status.

## Conclusion

This analysis successful demonstrated the liklihood of a systematic difference bewteen gene expression in control versus case patients and bulbar versus limb onset patients. It also successfully created linear discriminant analysis models to describe this difference. Unfortunately, it failed to create meaningful classification tree models. The biggest struggle with this project was time. The data had many columns which slowed analysis, and the author was only able to begin a few weeks before the analysis was due. The second biggest struggle was extracting significant results from data with few rows and many possible columns to cherry-pick correlations from. Thankfully, the linear discriminant analysis models held up to scrutiny. The next step in this analysis would be to take into account the differences in cell-type composition of the tissue samples and see if that has explanatory value towards gene expression. It remains to be shown from this data that the differences in gene expression analyzed were not, in fact, due to a structural difference in cell-composition of the tissue samples between the patient status groups. And, as always, more data would be nice. A dataset with significantly more patients could draw much stronger conclusions, and perhaps create statistically significant classification trees.