Final_Report_Stats

Dane Dewees

March 11, 2018

Summary of datafile

Here we are looking at an actual RNA-seq dataset associated with a study looking at patients who underwent total knee replacement surgery (otherwise known as TKA). The data was collected and provided by Dr. Hans Dreyer, of the Human Physiology department at the University of Oregon. The data provides gene count data that was filtered for all protein encoding genes (exclusively mRNA) for 14 patients, randomly assigned for the EAA treatments. The total output of the file consisted of roughly sixty thousand genes per individual. Here were are looking at a 1-wk post surgey timeline of the tissue samples taken from the quadrecep muscle group of the patients (operative leg). The significance of this dataset was that patients were grouped in either reciveing EAA treatment (given essential amino acids) or not prior to undergoing surgery. Specifically, we're interested in the role of EAA's on patients who have undergone the surgery. Essential amino acids have been shown in previous studies to relive muscle atrophy and decrease in inflamtation post surgery of TKA. We want to fit a model to this data so that for future studies and future cohorts undergoing this treatment, we can predict whether or not this supplementation treatment was found/executed properly.

Goal of the report

The overall goal of this report is to construct a predictive logistic regression model utilizing Stan. When applied to each patients genecount files, realisticaly the model SHOULD identify whether or not these patients recieved EAA treatments prior to surgery. This model can essentially determine the efficacy of the treatment shown in mean expression levels across the genes given each individual. Due to the fact that the file was rather large and wide range of gene counts, I subsetted the file by just extracting 500 genes (randomly sampled) as shown below:

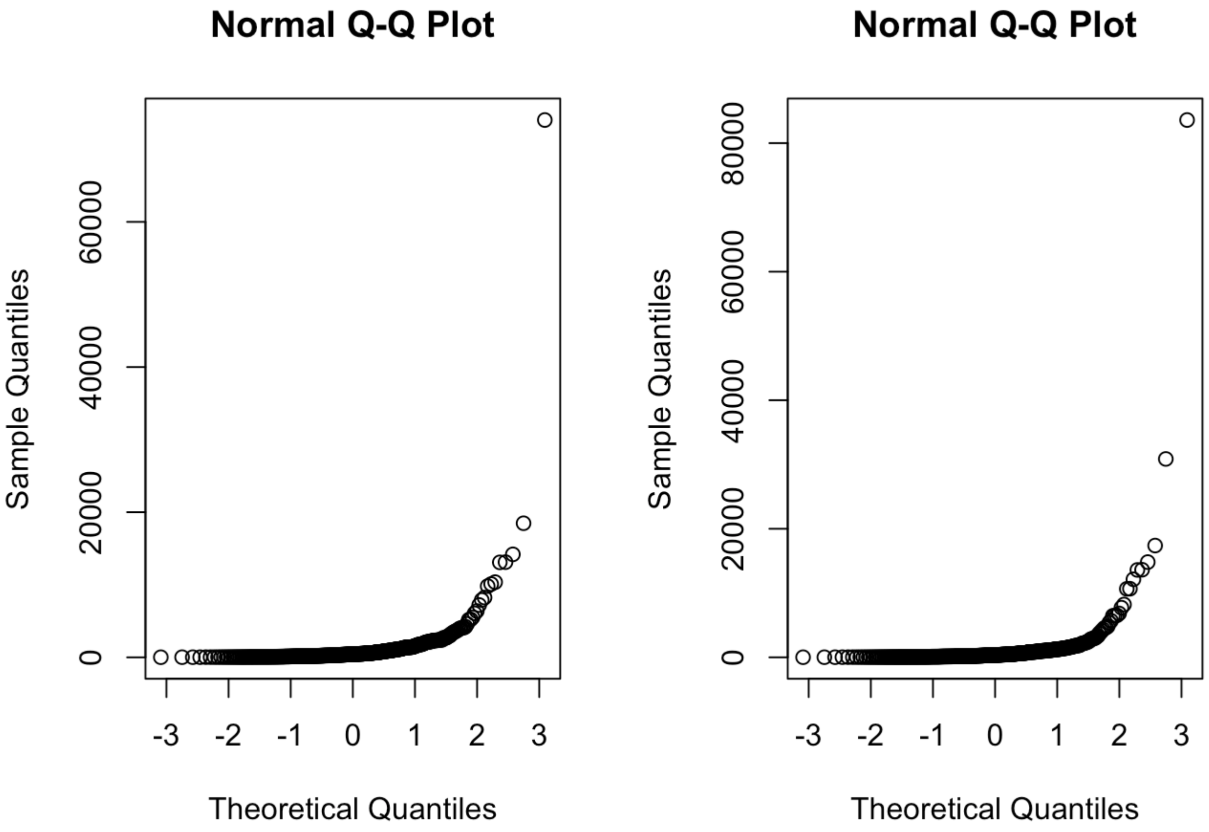
```
##randomly sampling the counts files to 500 genes
set.seed(151)
random_sample_500 <- subset_gene_df[sample(nrow(subset_gene_df), 500), ]
random_sample_500_2nd <- random_sample_500
treatment <- c(0,1,1,0,0,1,0,0,1,0,0,1,1,1)


## only keep values greater than 10 in terms of gene counts to speed up process
when running stan
Gene_ID <- random_sample_500$Gene_ID
random_sample_500$Gene_ID <- NULL
```

Preliminary analysis

We looked at the median and MAD while ingnoring mean/SD due to needing a robust solution. Factoring in outliers was determined while scaling the data, which was important to look at when running our Stan models (i.e., run time issues, etc.).

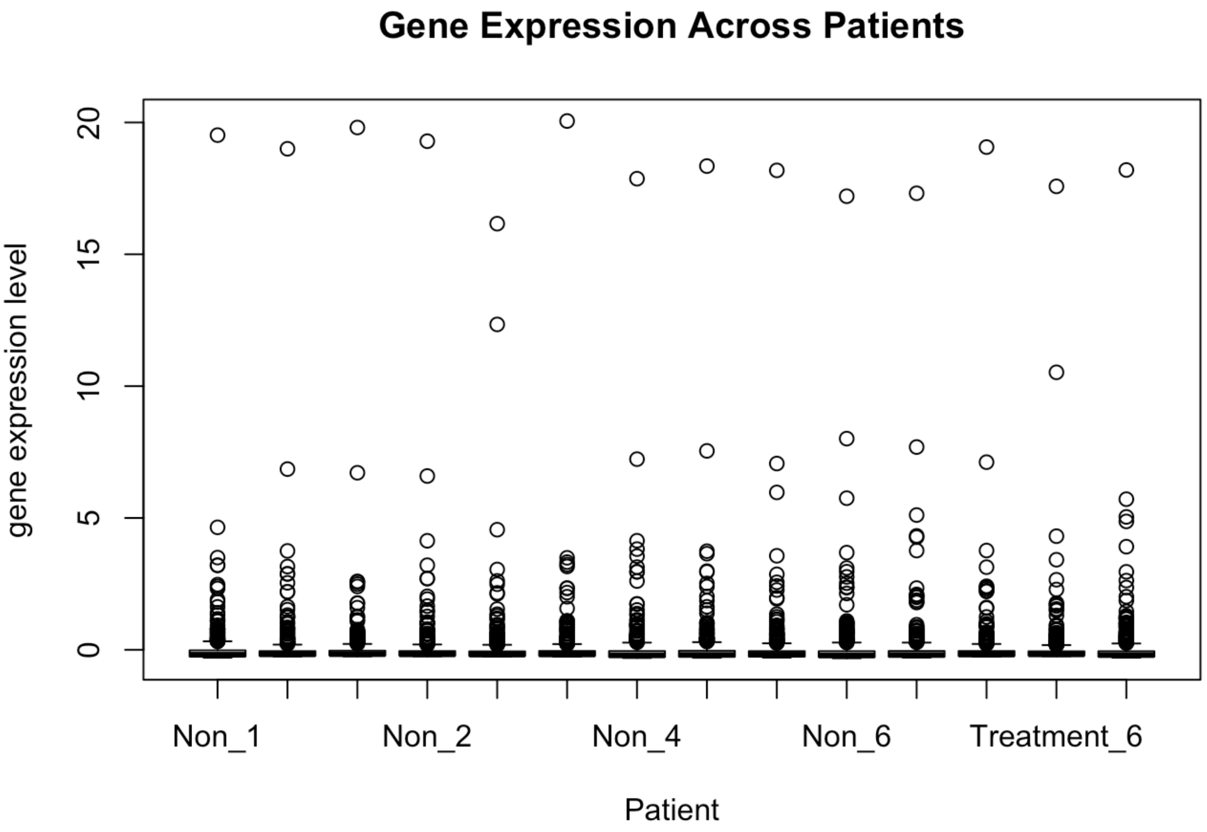
As you can see, the Q_Q plot shows non-linearirty which indicates there is no gaussian noise. This is then used in how we wrote our first model to determine the difference between treatment and non-treatment patients. Using a scaled function, and given the range of outliers in the normal dataset, I scaled the data in order to reduce the influence the outliers had on the median coverage.

Q-Q plot



plot illustrating non-normally distributed data

Here you can see from the boxplot as well as the histogram plots that the distribition is not normal. This is why we presented cauchy in our model, and as support material for the Q-Q plot in terms of either having gaussian noise present or not. From these plots, we can then start thinking about using a logistic regression model. Using this model with stan can help illustrate and make use of our predictor vairables of choice (both continuous and categorical in terms our patents and treatment groups). Since we want to predict our dependent variables in multiple categories (i.e., treatment/non-treatment), loogistic regression model is ideal.

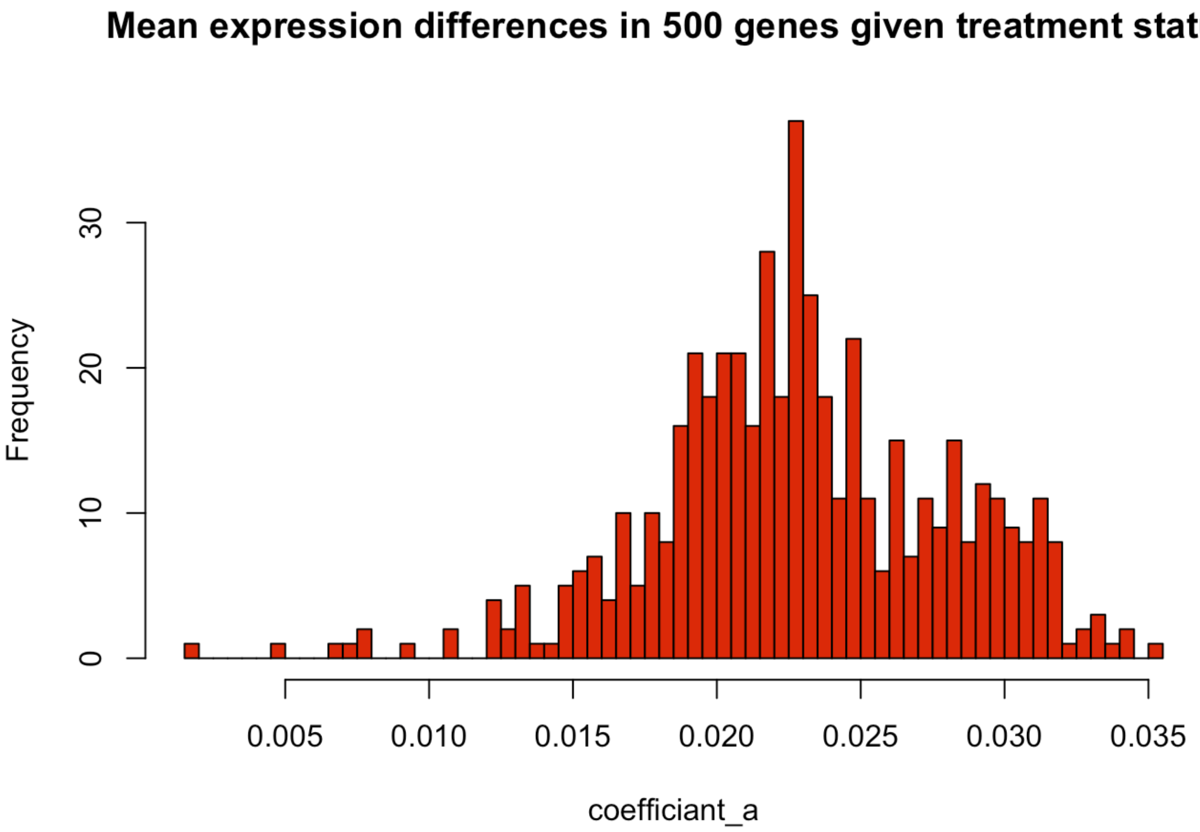


Stan Model

We then wanted to see how much gene expression differs between treatment and non-treatment patients. In particular, we wanted to look at individual variation. Given that there was poisson distribution, as shown in the previous plots, we applied cauchy in our model. Using cauchy was important to use since our variance is not yet defined across all patients. We fitted a robust model with separate mean expression values for treatment and non-treatment patients, and used the inference on the priors of those means to answer this question of inter-individual variation.

Mean_Expected_Diff Plot

After running the model, I wanted to plot the mean expression difference across each gene. First I needed to extract the posterior from the model ran previously and then calculate the mean vector levels across all genes. Here you can see the histogram showing the spread across mean expression levels. There is little to no significance that came out of the probability factor of MAD being greater than one. By looking at the mean value across all samples, which was ‘1.3e06’, we can assume that none of the genes are effected by treatment given expression levels. This is not ideal but recomendations toward scaling the data could be suffice in terms of mean expression differences across both groups.



Stan Model ~ Predictor of Treatment influence on gene expression

Here I was selecting 10 genes to generate a 'treatment test' for the patients that underwent TKA in order to fit a robust logistic regression model given those 10 genes. Given the selection of these ten genes, we can predict the infleunce of mean expression levels across treatment vs non-treatment individuals in the study. This can help determine whether or not the expression levels infleunces the outcome from the treatment of EAA's to the patients prior to the surgery.

Methods behind analysis/STAN model

When validating the model, we need to approach the equations used to support the reasoning behind the top-10 gene expression levels. Here we applied: $(y_i = c_0 + c_1 * x_{1i} + ... + c_{10} * x_{10i})$ when looking at the top 10 gene expression levels.

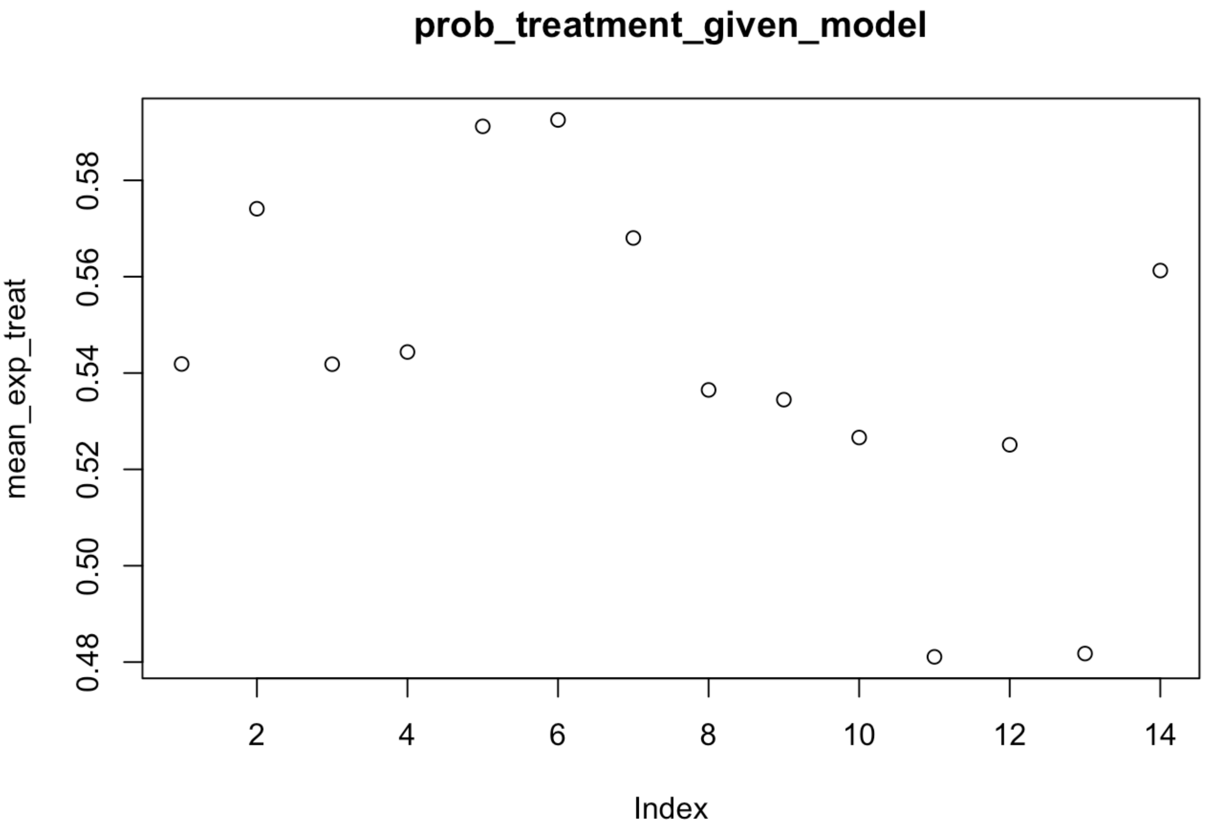
From there, we can return probability values that illustrates the liklihood of either recieving drug treatment or not. We can do this by using the logit function of y, where $(1/(1 + e^- y))$ can be applied to STAN in order to find the liklihood of treatment given from the use of our robust logistical regression model.

Therefore, the Stan model implementation would index through each of the 10 genes across the 14 individuals and estimate $(y_i = c_0 + c_1 * x_{1i} + ... + c_{10} * x_{10i})$ using y = b0 + c .* to_vector(e[i]) and predicting the logit function using bernoulli_logit().

```
#Stan model
treatment_test_model <- "
data {
    int m; // Number of patients in sample
    matrix[m, 10] e; // matrix of 14x10 original expression values
    int p[m]; // treatment status dummy coded vector
}
parameters{
  real b0;        // intercept of equation
  vector[10] c;  // vector of 10 coefficients, 1 per gene

  real<lower=0> df_b;
  real<lower=0> df_a;

  real<lower=0> sigma_b;
  real<lower=0> sigma_a;

  real mu_b;
  real mu_a;
}
model {
  vector[10] y;
  for(i in 1:10){
     y = b0 + c .* to_vector(e[i]);
     p[i] ~ bernoulli_logit(y);
  }

  // Priors:
  b0 ~ student_t(df_b, mu_b, sigma_b);
  c ~ student_t(df_a, mu_a, sigma_a);

  // Priors on b
  mu_b ~ normal(0, 10);
  sigma_b ~ normal(0, 10);
  df_b ~ normal(0, 20);

  // Priors on c
  mu_a ~ normal(0, 10);
  sigma_a ~ normal(0, 10);
  df_a ~ normal(0, 20);
}
"
```

Validating Predictor Model ~ Plot - treatment outcome

As you can see from the plot below, after running our robust logistic regression on the highest gene exp_samples crossed with total samples, there is no clear mean expression difference shown between the two different sample types (drug treatment vs non-treatment). Values from the x-axis (0-7) are the individuals who recieved treatment and (8-14) are those who did not recieve treatment. As you can see from the range as well as clustering patterns between the two groups, there is not clear distinction between the groups in terms of whether treatment was given.



Here we see the top differentially expressed genes out of the 500 genes, and given our model from above, we can, in theory, predict the mean expression values for genes shown while either recieving drug treatment or not recieving drug treatment. Plugging those genes back into our model we should be getting clear outputs of either 1’s (recieved treatment) or 0’s (no-treatment). Given that our range for mean_exp_treatment being around 0.5, it shows that our predictor model fiven the top expressed genes was not efficient enough to determine whether or not treatment was given to each individual. Maybe adjusting the scaling during the preliminary analysis or factoring in the priors when running stan could optimize the predictor model for future use.