# SDS 385: Final Project

Evan Ott
UT EID: eao466

December 7, 2016

## Contents

## Methods

We tried two general methods for trying to identify markers for periodontal disease: logistic regression with a LASSO penalty and the off-the-shelf `DESeq2` package in R. Within each framework, we also tried a master data set where each gene is treated separately, along with a "grouped" data set, where similar genes were counted together into biologically-sensible groups.

### Data

The master data can be thought of as a matrix $X$ with dimension $N \times P$ where $N$ samples were taken of $P$ genes from across the microbiome. $X_{i,j}$, determined using RNA-seq, represents the count of how many times gene $j$ was expressed for sample $i$. Along with this information, we also know $Y_i$, whether sample $i$ is in the control or treatment group. (Note: We also know which patient the sample comes from, but in our analysis, we disregarded this and treated all samples as independent – a simplifying assumption that we take with a grain of salt)

The grouped data $\tilde{X}$ is then an $N \times \tilde{P}$ matrix, where $\tilde{P} < P$ is the number of groups of genes. $Y$ remains unchanged.

Here, we have $N = 27$ samples, $P = 152448$ observed genes, and $\tilde{P} = 2227$ groups of genes.

### Logistic Regression

For this problem, perhaps without much surprise to those with topic familiarity, the logistic regression model failed pretty extravagantly. Using a pre-processing step common in using this manner of count data, we applied the following log transformation:

$$Z_{i,j} = \begin{cases} 0 & X_{i,j} = 0 \\ \log_2(X_{i,j} + 0.1) & X_{i,j} > 0 \end{cases}$$

Essentially, keep the zeros, but scale the rest of the data on a log scale (adding a small constant to ensure a count of one is not represented as a zero.

We used stochastic gradient descent to solve this objective:

$$\begin{aligned} &\underset{\alpha\in\mathbb{R},\ \beta\in\mathbb{R}^p}{\text{minimize}} \sum_i \|y_i - \hat{y}_{\alpha,\beta}(Z_{i,\cdot})\|_2^2 + \lambda\|\beta\|_1 \\ &\hat{y}_{\alpha,\beta}(x) = \frac{1}{1 + \exp\left(-\alpha - x^\top \beta\right)} \end{aligned}$$

that is, logistic regression with a LASSO penalty on the non-intercept parameters.

However, on both the "master" and "grouped" data, we were unable to produce a model that had any semblance of good predictive power. We used 3-fold cross validation to select the penalty $\lambda$ using the held-out data. In particular, we considered the mean 0-1 error (number of correct predictions when $\hat{y}$ is rounded to 0 or 1) and the mean of the absolute error ($\sum_i |y_i - \hat{y}_{\alpha,\beta}(Z_{i,\cdot})|$). We used 3 folds simply because we had so few samples $N = 27$. Consequently, we used the "eye test," as it were, to identify reasonable values of $\lambda$ from plotting these cross-validated error estimates. In the grouped case, see Figure 1

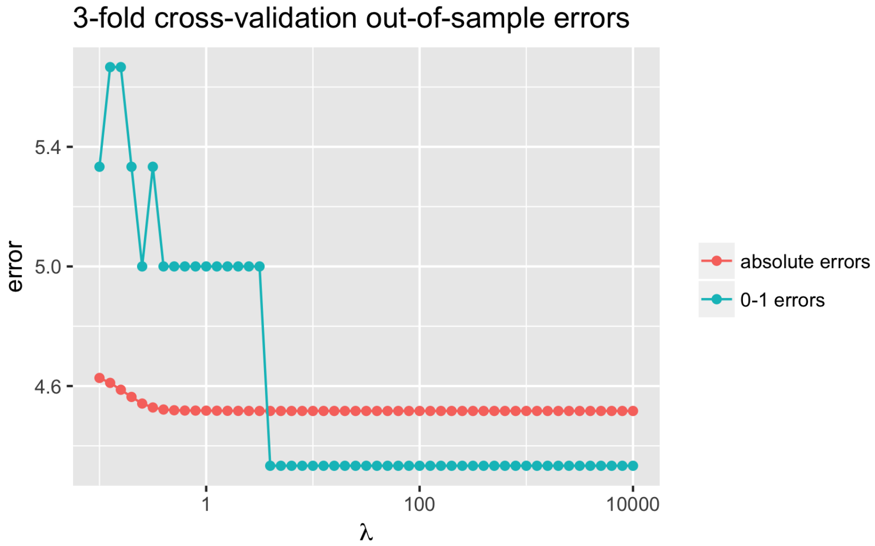

Figure 1: Error using the grouped data versus the LASSO penalty $\lambda$. In blue, the mean 0-1 error on the held-out set. In red, the mean absolute error on the held-out set.

By the "eye test" – we used only 3 folds, so it seems foolish to use the "most parsimonious within one standard error" trick simply by having a not-believable estimate of the standard error – the absolute error seems to not improve for $\lambda > 10^{-0.4}$ while the 0-1 error does not improve for $\lambda > 10^{+0.6}$. Using the more restrictive value $\lambda = 10^{+0.6}$, running on the entire grouped data set shows that the predicted values $\hat{y}$ are all below 0.5, indicating that the rounded predictions would all be zero! See Figure 2

Out of $\tilde{P} = 2227$ possible gene groups, this process produced non-zero coefficients for only 4. For the $P = 152448$ total genes, the same process (for $\lambda = 10^{+0.5}$) produced 25 non-zero coefficients. These are included in the tables below, but should not be viewed as particularly reliable, for the reasons outlined above.

On the grouped data, we also tried using the raw counts. The cross-validated error plot was certainly more interesting, but the results were similar, where the predictions using the entire grouped data set were abysmal.

### DESeq2

The R package `DESeq2` is a standard for analyzing this kind of data. It uses the model below to identify which genes are more indicative of a control or treatment sample:

$$\begin{aligned} X_{i,j} \sim& \texttt{Negative-Binomial}(\mu_{i,j}, \alpha_i) \\ \mu_{i,j} =& s_j q_{i,j} \\ \log_2(q_{i,j}) =& u_{j,\cdot} \beta_i \end{aligned}$$

where $u_{j,\cdot}$ comes from the $N \times 3$ design matrix (first column all ones, second column as an indicator for being in the control group, third column for being in the treatment group), and $\beta_i$ is the quantity of interest, with $\gamma_i = \beta_{i,3} - \beta_{i,2}$ being the "log2 fold change." We won't discuss the other parameters here as they are not of interest in this analysis. For more information, see §4.1 of (Love et al., 2014).

It's a little unclear exactly how `DESeq2` handles multiple testing. It says that it uses the Benjamini-Hochberg procedure to create adjusted $p$-values (such that checking $\text{padj}_i < \alpha$ is equivalent to the Benjamini-Hochberg procedure with FDR controlled at $\alpha$) (Benjamini and Hochberg, 1995). Any gene (gene group) with such an adjusted $p$-value is included in Table 2 (Table 1). However, this adjusted $p$-value is not always available – in cases where a gene/group is filtered by "independent filtering" (having a low mean normalized count), it is set to `NA`. See §1.5.3 of (Love et al., 2014) for more.

So, we implemented our own, standard version of the Benjamini-Hochberg procedure on the non-adjusted $p$-values (that are the result of a Wald test for the log2 fold change). While in principle, the genes identified in

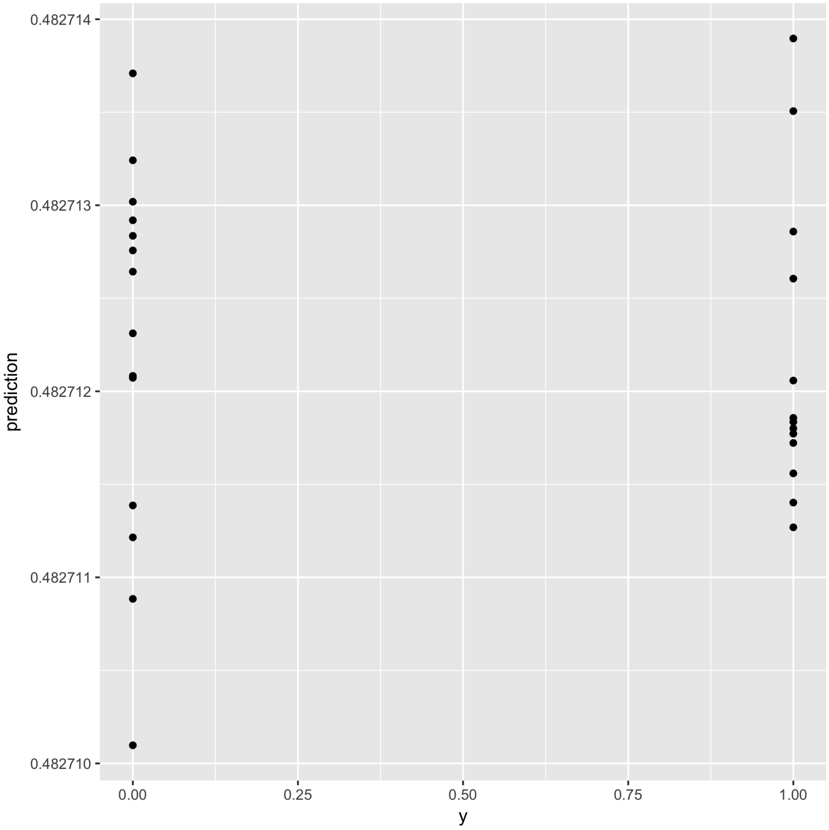

Figure 2: Predictions for $\lambda = 10^{+0.6}$ on the grouped data set versus actual value (0 for control, 1 for treatment).

this manner need not have been a subset of those identified in the default setting, in practice they are. These are seen in the "Our B-H" column in the tables.

Finally, we also looked into the "independent filtering" which is described in §3.8 and §4.7 of (Love et al., 2014). This essentially tries to determine a cutoff for the mean of the normalized counts where smaller values are excluded from the analysis. Turning this off (in general) will likely yield fewer rejections once DESeq2's internal Benjamini-Hochberg procedure is applied – more "noise" data will enter in, padding the "signal" data, making it more difficult for a particular signal to make it past the Benjamini-Hochberg procedure (if the significance level is kept constant). However, if we are happy with fewer rejections (perhaps just a smaller set of genes to consider during an initial investigation), this becomes useful. The results from turning off the filtering (but using DESeq2's specialized Benjamini-Hochberg procedure) are in the "No Filtering" column in the tables.

# Results

Tables 1 and 2 represent the high-level results of this investigation. They indicate genes/groups of genes identified by logistic regression (that should not be regarded highly for reasons outlined above) and by DESeq2. The variants applied to the latter indicate smaller subsets of genes/groups of genes to consider when trying to investigate any biological basis for periodontal disease. In particular, dpig_c_1_8 and raer_c_9_1389 seem indicative of positive and negative markers, respectively for disease based on turning off DESeq2's independent filtering (Table 2).

# References

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

Michael Love, Simon Anders, and Wolfgang Huber. Differential analysis of count data–the deseq2 package. *Genome Biology*, 15:550, 2014. URL https://www.bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf.

| Gene group | LR | `DESeq2` | Our B-H | No Filtering |
|---:|---|---|---|---|
| 1.1.1.36 |  | - |  |  |
| 1.14.-.- |  | - |  |  |
| 1.14.14.- |  | - |  |  |
| 1.14.14.10 |  | - | - | - |
| 1.2.1.39 |  | - | - | - |
| 1.2.99.8 |  | - | - | - |
| 1.3.99.5 |  | - | - | - |
| 1.4.99.- | Y |  |  |  |
| 2.1.1.140 |  | - | - | - |
| 2.7.7.73 |  | - |  |  |
| 2.7.13.3 | Y |  |  |  |
| 3.1.3.21 |  | - | - | - |
| 3.4.11.- |  | + | + | + |
| 3.4.11.9 | Y |  |  |  |
| 3.4.24.- | Y |  |  |  |
| 3.5.1.54 |  | - | - | - |
| 3.6.3.9 |  | + | + | + |
| 4.2.1.153 |  | - |  |  |
| 5.1.1.7 |  | - | - | - |
| 5.3.3.- |  | - | - | - |
| 5.4.99.26 |  | - | - | - |
| 6.-.-.- |  | - |  |  |
| 6.3.3.3 |  | - | - | - |
| 6.4.1.6 |  | - |  | - |

Table 1: Gene groups identified using the various strategies outlined. A "Y" in the first column indicates that the listed gene group had a non-zero coefficient in the logistic regression method – this is primarily recorded for posterity, not because the analysis is trusted. For the remaining columns, a "-" indicates that a high count in that gene group is indicative of being a control sample. A "+" indicates that a high count is indicative of being a treatment sample. A blank indicates that it is not present in that particular model.

| Gene | DESeq2 | B-H | No Filter |
|---|---|---|---|
| abau_c_1_3356 | - | | |
| aisr_c_20_1998 | - | | |
| amas2385_c_4_872 | - | | |
| aot170_c_17_1854 | - | | |
| aot171_c_140_1750 | - | | |
| aot171_c_3_100 | - | | |
| aot172_c_99_1663 | - | | |
| aot448_c_6_964 | + | | |
| apre_c_1_700 | - | | |
| cdip_c_1_1509 | - | | |
| chom_c_27_1386 | - | | |
| chom_c_55_2086 | - | | |
| cper_c_2_296 | + | | |
| cure_c_1_1578 | - | | |
| dpig_c_1_8 | + | | + |
| efae1080_c_1_913 | - | | |
| esak_c_1_3901 | - | | |
| fnucp_c_3_1442 | + | | |
| fper2555_c_2_652 | + | | |
| gmor_c_19_1592 | - | | |
| lcat_c_18_1871 | - | | |
| lgas_c_1_25 | - | | |
| lgoo_c_14_911 | - | | |
| lmon_c_1_308 | - | | |
| lot107_c_9_1346 | + | | |
| mlot_c_1_5349 | - | | |
| mneo_c_1_1226 | - | | |
| mneo_c_1_1611 | - | | |
| mot186_c_3_1980 | - | | |
| mtub_c_1_788 | + | | |
| nbac_c_3_267 | - | | |
| NCBIABIX_c_1_1199 | - | | |
| opro_c_21_2166 | - | | |
| pend_c_1_176 | - | | |
| peno_c_75_2748 | - | | |
| pmel_c_20_1104 | - | | |
| pmuls_c_6_923 | - | | |
| pot786_c_28_1621 | - | | |
| pple_c_7_1120 | - | | |
| pstu_c_1_2331 | - | | |
| pver_c_10_1377 | - | | |
| raer_c_1_313 | - | | |
| raer_c_17_1833 | - | | |
| raer_c_2_422 | - | | |
| raer_c_5_1016 | - | | |
| raer_c_9_1389 | - | - | - |
| rden_c_1_560 | - | | |
| rden1994_c_1_356 | - | | |
| sked_c_1_1556 | - | | |
| sked_c_1_1767 | - | | |
| smal_c_1_1545 | - | | |
| smit_c_1_1759 | - | | |
| smut_c_1_1468 | + | | |
| smut_c_1_1548 | + | | |
| smut_c_1_1794 | + | | |
| smut_c_1_371 | + | | |
| smut_c_1_836 | + | | |
| sot138_c_21_1472 | - | | |
| sot149_c_17_1639 | - | | |
| ssal_c_18_1887 | - | | |
| tmed_c_12_2247 | - | | |

Table 2: Gene identified using the various strategies outlined. The interpretation is identical to that in Table 1. The logistic regression data is not included here, because, like in the grouped data, it is a) a completely different set of genes and b) has seemingly no predictive power.