Case Study for Gene Expression Data

Kaylee Hodgson

Brigham Young University

Abstract. The analysis in this paper uses multiple linear regression (MLR) to create a model for the purpose of determining the malignancy scores of cancerous tumors based on gene expression. I overcome the obstacle of the high dimensionality in the dataset, or the large amount (5150) of genes measured, by using the Lasso shrinkage method, and bootstap confidence intervals. The MLR analysis finds that there are only a couple of genes that are significant to tumor malignancy scores.

1 Introduction and Motivation

Cancer affects many people throughout the world, and much has and is being done to try to learn more about this mysterious disease. Cancer is caused by massive, uncontrolled cell growth, and can occur in many different areas of the body. Although cancer is relatively widespread, there has been little understanding gained over the years on how to identify, prevent, or cure the disease. However, scientists believe that one potential avenue to understanding cancerous tumors better is to study gene expression in the form of gene expression profiling. This allows thousands of genes to be monitored at the same time to evaluate the gene expressions, and specifically in the case of a tumors, to evaluate which genes are regulating growth and could therefore, if altered, be the cause of the tumor.

This analysis will employ regression analysis in an attempt to identify the genes that are most highly associated with aggressive tumors. The purpose of this analysis is to create a model that informs us on the average malignancy scores of tumors given gene expression data. Regression analysis is a useful tool for inference because it allows us to evaluate average values of malignancy by just inputting values of gene expression into a single best-fit equation. The inferential findings can help future researchers to better understand how changes in the expression of certain genes impacts tumor growth. The significant findings, while small, are invaluable to the progression of tumor research because they identify genes that significantly contribute to growth regulation, and that are associated with cancerous tumors when altered.

2 Data Summary and Diagnostics

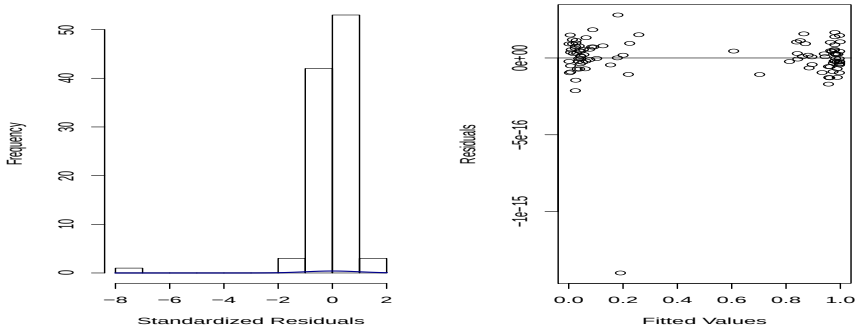
The data for this analysis is pulled from a Gene Expression dataset which has compiled information from 102 cancer patients on measures of 5150 genes. This analysis will use these measures in an attempt to identify what genes tend to affect the malignancy score of a tumor. The malignancy scores are measured between 0 and 1, with 1 identifying the most invasive tumors.

The main issue with employing regression analysis on this data is large number of genes, with only a small sample of cancer patients. A regression analysis is unsolvable (has infinite solutions) if there are more variables (in this case 5150 genes) than subjects (102 cancer patients). Because of this, the lasso shrinkage method is employed to make the regression model solvable. This will be discussed and implemented in the next section.

However, before addressing the issue of high-dimension variables, I first test the data for the assumptions of the linear model, including testing whether there is a linear relationship between malignancy scores and each of the genes, whether the standardized residuals of the malignancy scores are normally distributed, whether the malignancy measures are each independent of each other, and whether there is equal variance of the observed malignancy values around the fitted values of the model.The large amount of genes (explanatory variables) in the dataset make evaluating linearity of the model difficult. In order to address this, I randomly draw 50 of the genes from the dataset, and fit a model against these 50 selected genes. I repeat this randomization and model refitting method ten times in an attempt to more comprehensively evaluate the linearity of these variables.

I used added variable plots to test the linear relationship between the different genes and the malignancy score. These relationships all appeared to be linear. The plots in Figure 1 were used to test the linear model assumptions of normal distribution, independence, and equal variance for just the first 50 genes that were drawn, in order to give a visualization of the data trends. I find that the dataset appears to meet all of the assumptions of the linear model, with the exception of the normality assumptions, which are shown in Figure 1. However, the normality assumption only appears to be broken by a single outlier, which can be seen in the histogram. Besides that, the rest of the histogram appears to be fairly normal. I determine that this is sufficient to continue with a linear analysis. Additionally, with the use of the shrinkage method employed in the next section, the only assumption that has to be met to run that is the linearity assumption.

Fig. 1. Tests of Linear Assumptions



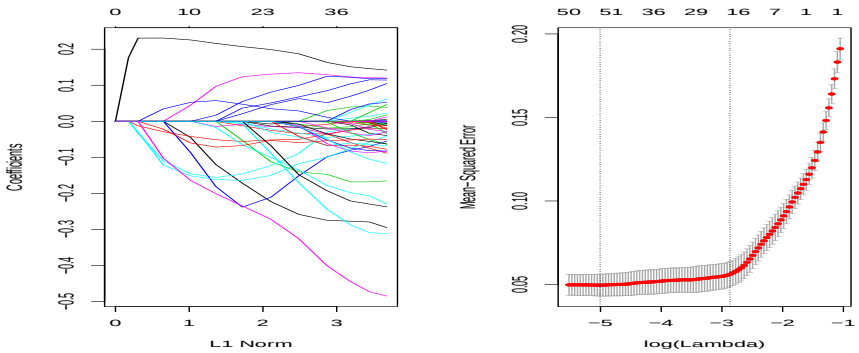
3 Lasso Shrinkage Method

The large amount of genes that are measured in this dataset make the regression model impossible to solve because the sample size of 102 patients is smaller than the number of genes, 5150, being measured for each of the patients. There are multiple methods used to deal with this high dimensionality problem, including variable dimension reduction and shrinkage. I choose to employ a shrinkage method in this analysis because, while dimension reduction can be useful, ease of interpretability of the regression output is largely lost in an analysis that employs dimension reduction. The two main shrinkage methods used are ridge regression and lasso. While ridge regression can be useful, this method does not employ variable subset selection, so the resulting model can also be difficult to interpret. Instead, lasso is employed, which performs variable subset selection by shrinking the least important coefficients to zero. This property makes the resulting model much easier to interpret, with fewer coefficients in the model. The lasso coefficients chosen for the model minimize the following function:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j|, \tag{1}$$

where λ is the shrinkage parameter and $\sum_{j=1}^{p}|\beta_j|$ is the shrinkage penalty used for the lasso method. The value for λ is chosen using a cross-validation method, where a grid of λ values is chosen and the test error for each value of λ is calculated. The solution path and cross-validated MSE as functions of λ are both plotted in Figure 2. The first plot shows how different coefficients (the estimates that inform us how tumor malignancy changes as the different gene expression values change) are zeroed out depending on the shrinkage parameter that is chosen. The values at the top of the first plot indicate how many coefficients are not zeroed out depending on the value of lambda. The second plot shows the different mean squared error (MSE) estimates for different shrinkage parameter values. The shrinkage parameter that gives the lowest MSE is chosen for the lasso method.

Fig. 2. Cross Validation Output for λ



Using this 10-fold cross-validation method, I find that the best value for shrinkage parameter in terms of the mean squared error (MSE) is $\lambda = 0.032$. This is the value used in the lasso method to find the best subset and the parameter estimates. The model chosen using the lasso method includes 28 of the genes, a much lower number than the original 5150 genes.

4 Bootstrap Method

Once the lasso algorithm is run and coefficients are estimated for the 28 genes included (all others are zeroed out), the confidence intervals are estimated for each of the remaining coefficients using bootstrap sampling. The bootstrap algorithm samples with replacement from the set of observations for patients in the dataset. These samples are compiled into a new dataset that is again run through the lasso algorithm (with the same optimal tuning parameter) and estimates of the coefficients are stored for that sample. This process is repeated 10,000 times, and the confidence intervals are calculated from the quantiles of the parameter estimates. Each time this process is repeated, different coefficients are zeroed out, so the estimates are recorded for all coefficients, but only the coefficients that were not zeroed out in the lasso method with the actual dataset are included in the quantile calculations. The results for the model are recorded and discussed in the next section.

5 Model Results

The model summary can be found in Table 1. Because there are still 28 coefficients estimated in the reduced model, not all are reported in the table. Only the significant coefficients appear. The entire model that was output from the lasso method is:

$$\text{Malignant Score}_i = \beta_0 + \textstyle\sum_{p=1}^{28} x_i\beta_p + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)\ ,$$

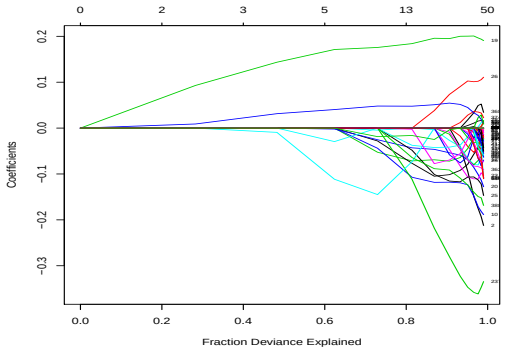
where β_0 is the intercept for the linear model, and all of the other β values indicate the slope, which tells us how much the malignancy score changes as that certain gene expression value changes by one unit. This equation is very useful in inferring the average malignancy scores from gene expression values, which is the goal of this analysis. This allows us to plug in values for the 28 genes and to evaluate tumor malignancy, a useful tool for scientists working to discover the genes that could potentially impact tumor malignancy the most. The two genes found to be significant to malignancy scores (X37639_at and X38087_s_at) are recorded in the table below. The confidence intervals in the table indicate that we can be 95% confident that the slopes of the two significant genes are within the lower and upper bounds, which do not contain zero. These are the two genes that should be evaluated more closely by scientists because they significantly impact the malignancy scores.

Table 1. Model Summary for Significant Terms

Coefficients	Estimate	Lower Bound	Upper Bound
β_0 (Intercept)	3.7407	1.7767	9.0457
β_{x37639_at}	0.2003	0.0724	0.2307
$\beta_{x38087\_s\_at}$	-0.0666	-0.1612	-0.0165

The root mean squared error (RMSE) for the model chosen using the lasso method is 0.21, which indicates that the values of malignancy scores found for this model are, on average, off by 0.21. This appears to be a small value, however, because the malignancy scores only fall between 0 and 1, this can be quite a large value to be off by in inference. Future researches and scientists should evaluate how to improve the accuracy of the model. While the R-squared value is not directly calculated from the lasso method, Figure 3 shows the percentage of deviance that is explained by the different possible models from the lasso method. The values at the top of the plot are the different numbers of variables or coefficients that could be chosen to fit in the model. The number of variables chosen from the optimal shrinkage parameter value was 28, as discussed above, so the percentage of deviance explained falls within 0.8 and 1.0. Even at the lower estimate, this plot indicates that the model explains at least 80% of the deviance in malignancy scores. This indicates that the model chosen does well at explaining deviance.

Fig. 3. Percentage of Deviance Explained



6 Conclusion

This paper employs lasso shrinkage as a useful method to employ regression analysis on this dataset, with such a large amount of observations on different genes, possible. I use the lasso method to find the new coefficient values once the optimal shrinkage parameter is selected, and employ bootstap confidence intervals to find whether each gene in the reduced model is significant. I find that only two genes are significant to tumor malignancy scores, and obtain a relatively large RMSE for this prediction model. However, I also find that the percentage of deviance explained in the model is high. While this is a useful basis for scientists and other researches to build off of, there is quite a bit more work to be done in order to more confidently predict how gene expression affects malignancy of tumors.

Future studies should evaluate how to create a better model to describe tumor malignancy. Additionally, scientists should begin to evaluate more closely the properties of the two genes that were significant in this model in order to determine the properties that make these impactful. This paper, although slightly inconclusive, provides a useful basis for scientists to continue analysis of gene expression data's impact on cancerous tumors.