STAT 7995 Dr. Inman DNA

Dominic Matriccino

September 2022

1 Executive Summary

The purpose of this report is to discover the variables that are most responsible for the variation in our response variable lLR. Variables that contribute more to this variation would indicate extra importance in the field of probabilistic genotyping. 7 variables were initially identified for us to analyze and out of those 7, 5 were determined to be significantly responsible for variance in lLR when using ANOVA. These variables include the presence or absence of stutter, the number of contributors, the true contributor, the total amount of DNA and the amount of DNA of the true contributor. Interaction terms were also considered and several of these terms were found to be significant especially those that included either total DNA or amount of DNA. One variable that was not determined to be statistically significant were the per loci likelihood ratios. The problem with these variables is that they all add up to the ultimate response variable lIR. This introduces a challenge in that these variables will be considered significant when contained in a model due to the fact that they always sum up to lLR. The relationship that these variables have with each individual loci LR is a question that remains unanswered. Certainly this information would be useful and quite interesting to determine, but this was not an objective of this analysis. Consequently, the loci LRs were removed from any ANOVA model and should be studied individually to determine their significance in probabilistic genotyping. The value of this analysis is in the variables that were found to play a significant role in the overall likelihood ratio, as well as the variables that were determined not be to significant. Beyond this, individual loci likelihood ratios should be analyzed on their own to determine if specific loci have a greater influence on the overall likelihood ratio.

2 Introduction

2.1 General Background

Our problem begins with the issue of DNA profiling with respect to criminal justice. DNA can provide some of the strongest evidence in a criminal case and has been used to both overturn and convict thousands of people. Criminal cases that include DNA usually involve a pool of possible suspects being identified as potential suspects in a crime. DNA is gathered from both the crime scene and from the individuals and then an attempt is made to match those together using mathematical models. These models can tell us the likelihood that an individual from the pool of suspects is the actual perpetrator through this process of DNA profiling. These models use many variables to determine the strength of evidence for or against the suspect. Currently, these models rely on two different scoring models EFM and likeLTD to calculate the likelihood ratios. However, minimal research has been done regarding the variability of these models with respect to their parameters. Certain variables must be more or less responsible for the variation in the LRs and we would like to determine exactly what they are. Some of these variables include sample ID, replicate set ID, the sum of the DNA signature LRs, the scoring model used to calculate the LR,the presence or absence of stutter, the number of contributors and the true contributor. Whether or not these variables contribute significant variation to our response variable would be very useful and could inform others in

the field of probabilistic genotyping. More informed scientists will create better and more accurate DNA profiling models that can serve society in many ways.

2.2 Objectives

Dr. Inman would like to know which variables are responsible for the variation in the calculation of likelihood ratios. 7 variables were identified by Dr. Inman as potentially being significant. These include, total DNA, the number of contributors, dose/ratio of contributors, presence of stutter, replicates, per Locus LRs and finally the different math models. His tangential question is regarding whether there is a single metric to determine the variation exhibited by one replicate set. In other words, when these experiments are conducted again how can the variability be compared to previous experiments.

3 Approach to Project

The first step in our analysis began with data cleaning. We were told that the frequency column should be disregarded as it was not important in our analysis, so it was removed. Variables C1-C4 and D1-D4 were also removed because they provided redundant information given that we already have the total amount of DNA and the amount of DNA from the true contributor. Variables such as the number of contributors and the replicate set identification number were converted into factors. Lastly, given that we have the total sum of the likelihood ratios for all DNA signatures, removal of the DNA signatures would help avoid misleading results. The DNA signatures would be directly related to the overall likelihood ratios, so removing these variables would provide us with a clearer analysis and less of a chance of receiving misleading or incorrect results.

After data cleaning, our analysis of variance could begin. Our problem is clearly an analysis of variance question because we would like to know which variables contribute the most variation in the response variable lLR. Before ANOVA was conducted, there were 3 assumptions that needed to be considered, independence, normality, and homogeneity of variance. Dr. Inman confirmed to us that his trials were independent, so we did not need to dive deeper into that assumption. The next assumption is normality, which can be confirmed with a qqplot. After running our first ANOVA model with our 7 variables and interaction terms, we looked at the residual plots and determined that the normality assumption was violated. The tails of the qqplot deviated significantly from the perfect normality line, which displays the theoretical distribution of data points if it were truly normal. We can say from this that our data are not normal at some of the more extreme quantiles. This issue was resolved by using a transformation that would shift our data into a more normal configuration. A square root transformation was decided on, which resulted in a qqplot that did not violate the normality assumption nearly as much, albeit still some.(See Figure 1).

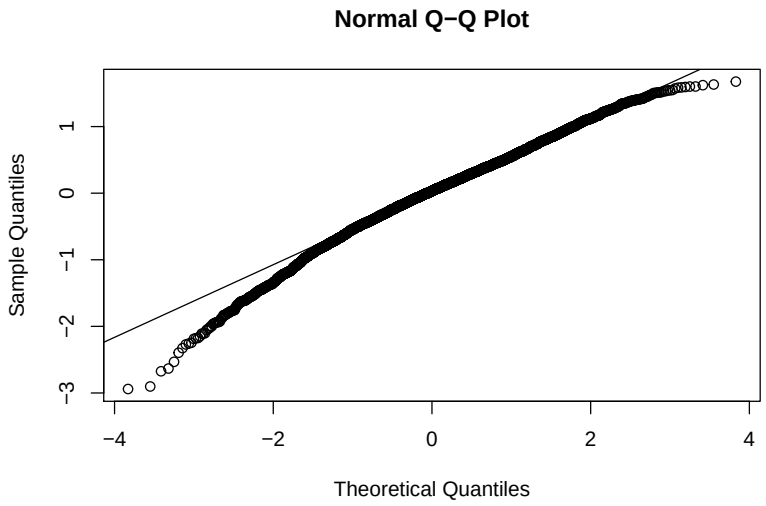

Figure 1: QQ Plot for Normality Assumption

Our last assumption is the constant variance assumption which can be confirmed through looking at a plot of residuals. When looking for homogeneity of variance you do not want to see any obvious patterns or shapes of the residuals. (See Figure 2). In our graph the points seem to fall randomly with the vast majority being within plus or minus 1 residual. The mean of the residuals should also be very close to zero meaning that all the residuals essentially cancel out to zero. The mean of our residuals is 1.409115e-17 which gives us confidence in this assumption. The only issue with this graph are the residuals on each end. There appears to be a small pattern starting at the larger amounts of the fitted values. This is likely due to the square root transformation because the transformation has a greater effect on larger values. Given the large sample size we would expect there to be some outliers and therefore problematic aspects of the residual plots. Considering the complexity of the data, a much larger violation could have been observed.

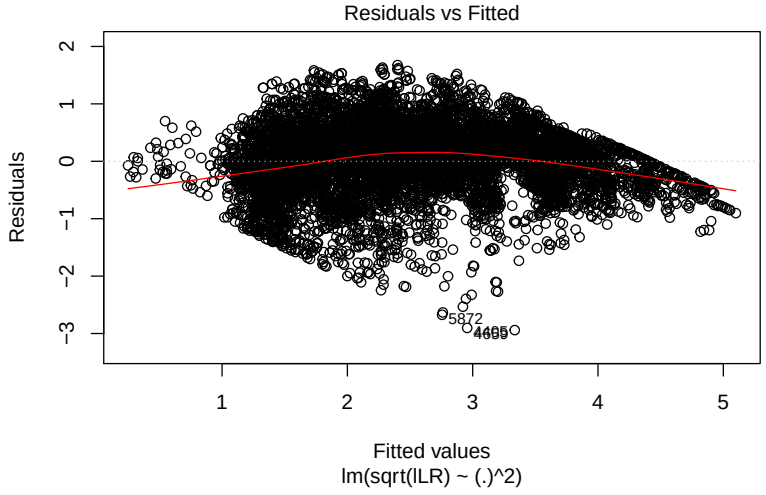

Figure 2: Residual Plot for Constant Variance Assumption

One final graphical representation of our residuals is the residual histogram which shows us the distribution of our residuals. (See Figure 3). Normality of the residuals is an assumption in many analyses and gives

us extra confidence in this analysis given the minor violations of other assumptions. We can see from the histogram a clear normal distribution with only minimal outliers in the negative direction. After considering all the assumptions we have reasonable confidence in the validity of our analyses and the results and recommendations forthcoming.

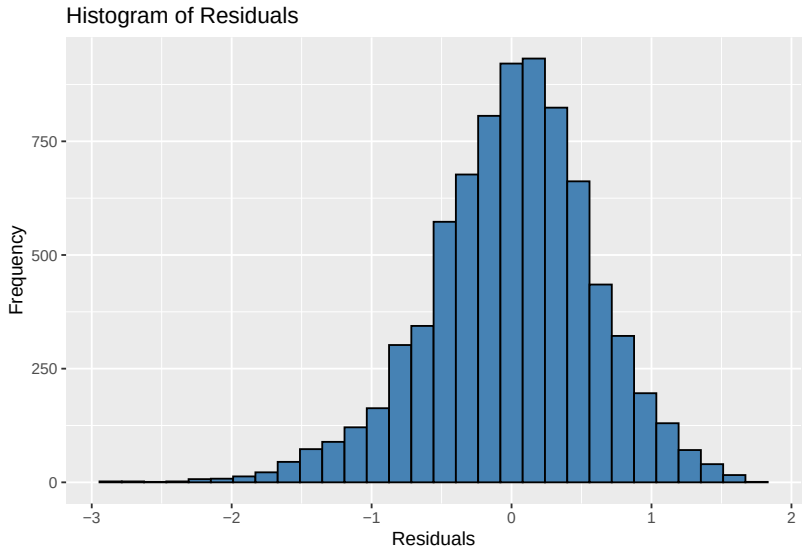

Figure 3: Residual Histogram

4 Results

The results of our ANOVA test (see appendix) gave us insight into the variables that contribute the most to the response variable lLR. Looking at our individual terms we can see that there are five variables that are considered significant. St, NC, TC, TotDNA and amtDNA. The two that are insignificant are Rep and Pr. Analyzing this first from the insignificant predictors, we can see that the mathematical model that is used to calculate the likelihood ratios is an insignificant predictor of the overall LR. This is a good outcome because we would hope that the two methods would provide likelihood ratios that are similar to each other and this result seems to prove that. Regardless of the model that is chosen we know that it will not have a significant effect on the overall likelihood ratio. The other insignificant variable is Rep which is also a good outcome because we want our replicate sets to have minimal variation and this tells us that regardless of the replicate set this variable does not contribute to the variation in the overall likelihood ratio.

Of our significant variables we have that the true contributor and number of contributors significantly influence the overall likelihood ratio. This makes sense because there should significant differences in the likelihood ratio when the true contributors DNA is found versus when it is not. The number of contributors likely influences the overall likelihood ratio due the difficulty in profiling someone when there are many more sources of DNA in a sample. The fewer the number of contributors the easier it is to profile the exact individual who is responsible for a crime, thus making more of a difference in the overall likelihood when the number of contributors is low. The next pair is the total amount of DNA and the amount of DNA of the true contributor. This makes intuitive sense because changes in the amount of DNA would likely cause changes in the calculation of the likelihood ratio. Having larger amounts of DNA to calculate the ratios would seem to provide more accurate results which is why it is significant. The last variable to be significant is the presence of stutter. This is potentially a concerning result because this means that the choice of the lab to process the DNA with a stutter will affect the overall likelihood ratio. This is concerning because someone could be convicted of a crime if the stutter is absent, but then be exonerated if the stutter is present (or vice versa). This seemingly arbitrary decision could have major implications in whether or not someone is convicted.

Finally, some of the interaction terms that seem the most significant include information on the contributors (NC and TC) or information on the amount of DNA(TotDNA and amtDNA). The most significant interaction term is TotDNA and amtDNA. This goes to show the importance of the amount of DNA when the likelihood ratios are being calculated. Clearly, the amount of DNA that is processed has a very significant effect on the variation of the overall likelihood ratio.

5 Conclusions and Appropriate Recommendations

From our analysis we can conclude that certain variables are indeed more important than others with respect to the overall likelihood ratio. Variables that are identified as significant in ANOVA should be considered and studied further to determine the nature of the relationship. Variables with the largest F-statistics should be considered the most important. The single variable that contributes the most variation in lLR is TotDNA and the interaction that contributes the most variation is TC:NC. Out of the other single predictor variables NC, TC, St, AmtDNA all play a significant role in the variation of lLR. Out of the other interaction terms our significant interaction variables include Pr:TC, St:TotDNA, NC:TC, NC:TotDNA, NC:amtDNA, TC:TotDNA, TC: amtDNA.

The recommendations from this analysis are rather clear. Many of the variables in the dataset need to be carefully considered because they play a role in the response variable lLR. Since the likelihood ratio is used to evaluate evidence in favor of a conviction, extreme care must be given to ensure the results are independent, accurate and consistent. The consequences of a false conviction are often devastating. DNA evidence is, however, not the only source of evidence used in a trial. Probablistic genotyping combined with other sources of evidence such as eyewitnesses or video evidence would provide the best evidence in a criminal trial. While some of these variables can be understood to have an affect on the lLR, further analysis should be done to confirm the consistency and relationship of these results.

Additionally, we were unable to determine whether specific loci contribute to the overall likelihood ratio. To analyze this we would need to make the loci LRs the response variable and conduct an ANOVA with the remaining variables. Using the loci as predictor variables in this analysis would be problematic because we know that they will always relate to the response variable. This would give us misleading information about the significance of these loci and lead us to the wrong conclusion that loci LRs contribute to variation in lLR. lLR is just the sum of all the loci LRs, so if we are interested in what causes variation in likelihood ratio we do not want to look at other likelihood ratios that simply sum to lLR.

Finally, the tangential question is concerned with variability of many replicate sets. When further experiments are conducted there will be many more replicate sets and it will be useful to understand if two groups differ from each other. This appears to be question regarding the difference in variance of two groups which can be answered using a chi-sq test for variance. If these experiments are conducted over time a population variance could be estimated and from there tests could be conducted as the many variables are tweaked. For example, if the scoring systems are updated it would be useful to know if this would result in a change of variance. A significant difference in variance would indicate that the change in scoring system did result in real measurable change in the data. As probabilistic genotyping improves, measuring the standard deviation of the likelihood ratios will be extremely useful and could be a useful tool that helps improve the field immensely.

6 Appendix

Results from our ANOVA analysis in R

```
##                  Df Sum Sq Mean Sq  F value   Pr(>F)
## Rep               4    1.4     0.4    1.022 0.394138
## Pr                1    0.0     0.0    0.138 0.710548
## St                1    1.7     1.7    4.935 0.026353 *
## NC                2  739.6   369.8 1055.347  < 2e-16 ***
## TC                7  773.9   110.6  315.501  < 2e-16 ***
## TotDNA            1 1726.0  1726.0 4925.646  < 2e-16 ***
## amtDNA            1  388.7   388.7 1109.165  < 2e-16 ***
## Rep:Pr            4    0.4     0.1    0.305 0.874509
## Rep:St            4    0.2     0.0    0.121 0.975086
## Rep:NC            8    5.8     0.7    2.073 0.034870 *
## Rep:TC           28   13.6     0.5    1.383 0.086178 .
## Rep:TotDNA        4    2.5     0.6    1.783 0.129126
## Rep:amtDNA        4    0.2     0.0    0.133 0.970152
## Pr:St             1    0.1     0.1    0.167 0.682896
## Pr:NC             2    1.8     0.9    2.504 0.081844 .
## Pr:TC             7    9.8     1.4    4.001 0.000223 ***
## Pr:TotDNA         1    1.4     1.4    3.915 0.047899 *
## Pr:amtDNA         1    0.2     0.2    0.448 0.503121
## St:NC             2    0.8     0.4    1.187 0.305155
## St:TC             7    2.4     0.3    0.968 0.452288
## St:TotDNA         1    3.9     3.9   11.140 0.000849 ***
## St:amtDNA         1    0.0     0.0    0.010 0.919388
## NC:TC             5  315.0    63.0  179.791  < 2e-16 ***
## NC:TotDNA         2    6.5     3.2    9.270 9.52e-05 ***
## NC:amtDNA         1   16.3    16.3   46.563 9.55e-12 ***
## TC:TotDNA         7   67.4     9.6   27.458  < 2e-16 ***
## TC:amtDNA         5   86.2    17.2   49.177  < 2e-16 ***
## TotDNA:amtDNA     1 1827.6  1827.6 5215.598  < 2e-16 ***
## Residuals      7686 2693.2     0.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 133 observations deleted due to missingness
```