STA 141C Report

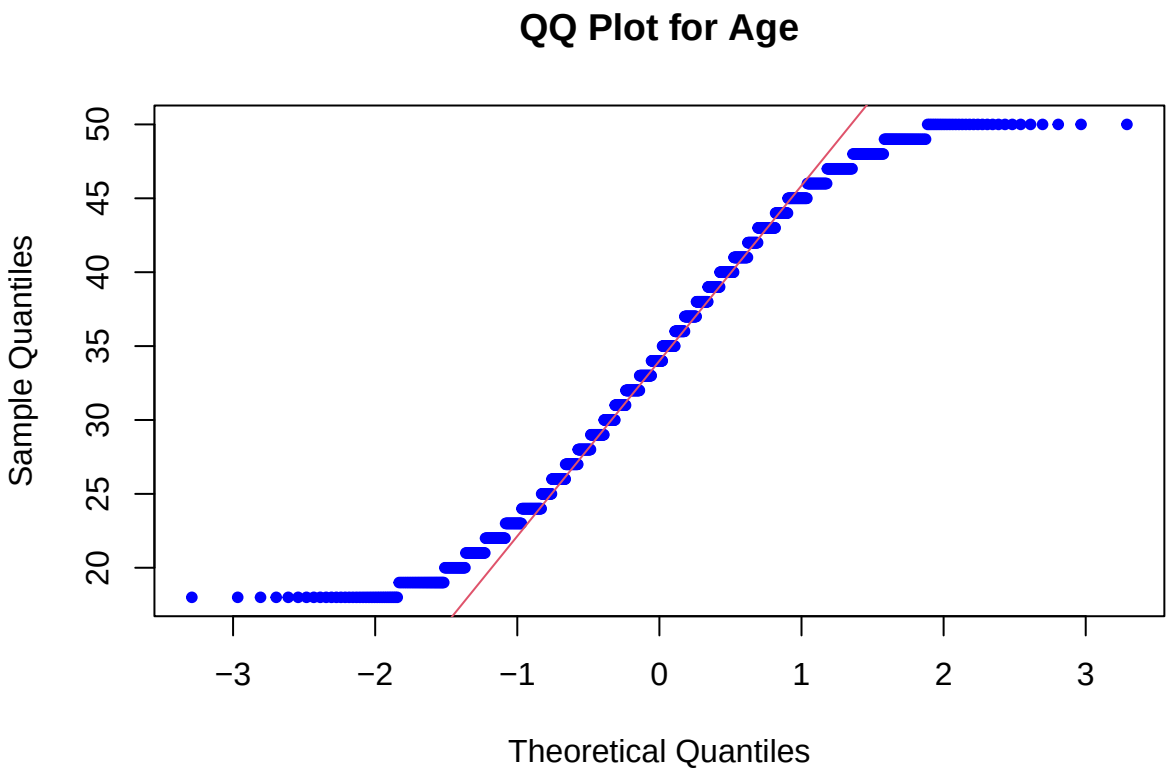
Juliet Lubin, Jessica Young, Hyeonwoo Shin, Suvethika Kandasamy

Background

The background of our report centers on Androgenetic Alopecia, commonly known as male or female pattern baldness. This condition significantly impacts millions of individuals in the United States, with an estimated 50 million men and 30 million women affected. Apart from its prevalence, Androgenetic Alopecia has profound psychological implications, leading to increased feelings of loss of self-confidence, low self-esteem, and heightened self-consciousness among those experiencing hair loss. Notably, this condition can onset as early as the teenage years and tends to escalate with age, highlighting the importance of understanding its causes and potential interventions.

Dataset Background

Our data source is Kaggle's "Hair Health Prediction" dataset. It includes individual records, each with an ID number and data on variables such as Genetics, Hormonal Changes, Medical Conditions, Medications/Treatments, Nutritional Deficiencies, Stress, Age, Poor Health Care Habits, Environmental Factors, Smoking, Weight Loss, and Baldness status (coded 0 for absent, 1 for present). These variables are mostly categorical, except for age.The dataset comprises 999 individuals, evenly split between those experiencing hair loss (502) and those who don't (497). Notably, there are no missing values; any absence is marked as "No Data". Therefore, we did not have to do too much data processing.



From this QQ plot, we can see that the age distribution is not normally distributed as there are points on the bottom left and top right that do not follow the QQ line. This visualization further verifies our Shapiro-Wilk normality test result.

This histogram represents the distribution of ages within the dataset, differentiated by hair loss status. We can see the distribution of ages is similar for both categories and it relatively even spreads across the age range. There is no immediately apparent age that significantly stands out with a higher prevalence of hair loss. The representation of individuals with and without hair loss appears to be relatively balanced across most age bins.

Methodologies

In this project, we will employ logistic regression models, random forest models, and linear discriminant analysis (LDA) to address our research question.

Research Question: What are the primary factors contributing to hair loss, and how do they collectively influence the likelihood of experiencing hair loss? Can we create a model to predict hair loss?

We will utilize logistic regression to create a classification model. Our aim is to analyze the influence of various factors contributing to hair loss. The categorical variables such as Genetics, Hormonal Changes, Medical Conditions, Medications/Treatments, Nutritional Deficiencies, Stress, Age, Poor Health Care Habits, Environmental Factors, Smoking, and Weight Loss will be used to predict the binary variable of hair loss (0 representing absence, 1 representing presence). Through combined models and also those with only one variable to gain some insight.

We will include Random Forest, an ensemble learning technique that constructs multiple decision trees during training time and outputs the mode of the classes of the individual trees for prediction. It's particularly effective for classification tasks where the relationship between predictors and the outcome is complex and non-linear. By aggregating the predictions of numerous trees, it reduces the risk of overfitting and enhances generalization.

Additionally, we will employ LDA to further understand the collective influence of these factors on the likelihood of experiencing hair loss. LDA will help us create a predictive model based on the variables mentioned above.

Model Evaluation & Expected Outcome: To assess the performance of our models, we will use confusion matrices. These matrices will provide insights into the accuracy of our predictions and help evaluate the effectiveness of our classification model in identifying individuals at risk of hair loss.By leveraging logistic regression, Random Forest, LDA, we aim to develop a comprehensive understanding of the primary factors contributing to hair loss and their collective impact on an individual's likelihood of experiencing hair loss. Ultimately, our goal is to create a predictive model that can assist in identifying individuals susceptible to hair loss, thereby enabling early intervention and tailored treatment strategies.

Logistic Regression

```
##
## predictions  0  1
##           0 48 60
##           1 36 55
```

```
## [1] "Accuracy: 0.517587939698492"
```

```
##           predictions
##              0   1
##   High     184 137
##   Low      179 148
##   Moderate 136 215
```

```
## [1] "Accuracy: 0.332332332332332"
```

```
##           predictions
##             0  1
##   High     44 27
##   Low      30 24
##   Moderate 41 33
```

```
## [1] "Accuracy: 0.341708542713568"
```

To do some further analysis, we fitted logistic regression models and predicted their accuracies with confusion matrices. We first split the entire dataset into 80% training and 20% testing and this is our resulting confusion matrix. We can see that the accuracy is around 52% which is a bit low so we wanted to fit more models to try to improve the accuracy score. We decided to fit a multinomial logistic regression model for hair loss with other variables as predictors and interaction terms. For example, we wanted to use the variable stress as a predictor because it had three subcategories. We can see that the accuracy is around 33% which is much lower. Using stress as an interaction term with hair loss, we can see that the accuracy is around 34%. Both accuracy scores for stress were low, so stress is not a definitive factor for hair loss.

```
##      predictions
##         0   1
##   No  312 215
##   Yes 187 285
```

```
## [1] "Accuracy: 0.597597597597598"
```

```
##      predictions
##        0  1
##   No  68 19
##   Yes 67 45
```

```
## [1] "Accuracy: 0.5678391959799"
```

We decided to fit the multinomial logistic regression model for all variables and found that weight loss seems to be the biggest predictor of hair loss due to its highest accuracy compared to other predictor variables. As seen in the confusion matrix on the left using weight loss as a predictor, we can see that the accuracy is around 60% which is the highest accuracy score achieved this far. As seen in the confusion matrix on the right with weight loss as an interaction term, the accuracy is around 57%.

Random Forest

```
##
## Call:
##  randomForest(formula = as.factor(Hair.Loss) ~ ., data = hair,      ntree = 500)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 48.25%
## Confusion matrix:
##     0   1 class.error
## 0 254 248   0.4940239
## 1 234 263   0.4708249
```

After logistic regression analyses, we implemented two random forest models to assess their predictive power regarding hair loss. This would lead to better performance than single models like logistic regression. We used 500 trees to provide a good balance between computational efficiency and model performance. Based on the output, the OOB error rate was approximately 48.95%. It serves as an unbiased estimate of the model error rate. We think the OOB error is relatively high and the confusion matrix shows a class error of around 47% for the 'No' class and 50% for the 'Yes' class. This error distribution suggests that the model does not favor one class over the other. It implies we need more improvement for the model.

```
##
## Call:
##  randomForest(formula = as.factor(Hair.Loss) ~ Age + Smoking +      Weight.Loss, data = hair, ntree = 500)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 1
##
##         OOB estimate of  error rate: 46.35%
## Confusion matrix:
##     0   1 class.error
## 0 302 200   0.3984064
## 1 263 234   0.5291751
```

For the second model, we focused on a subset of predictors: Age, Smoking, and Weight Loss. By using fewer predictors, this model is simpler and potentially more interpretable. There was a slight improvement with an OOB error rate of approximately 46.75%. This indicates that the selected features capture a good portion of the variance in hair loss without overcomplicating the model. The class error for the 'No' class decreased to about 40%. However, the error for the 'Yes' class remained high at around 52%. This suggests the model became more accurate at predicting the absence of hair loss but did not improve much on predicting its presence.

```
## Generalized Linear Model
##
## 999 samples
##   3 predictor
##   2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 899, 900, 898, 900, 899, 899, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.5385461  0.07718002
```

After our random forest models, we decide to use cross-validation on our logistic regression model. This will mitigate overfitting and provide a more accurate estimate of model performance on an independent dataset. The cross-validated

logistic regression model incorporated Age, Smoking, Weight Loss, and a quadratic term for age (I(Age^2)) as predictors. As explained above, We selected these factors based on their potential biological relevance to hair loss and the insights gained from previous models. The model achieved an average accuracy of approximately 53.66% across the 10 folds with around 0.073 Kappa statistic which is not that high. It suggests that there might not be a strong relationship. This cross-validation ensures that this estimate is less likely to be due to random chance in the data partitioning. However, the results also imply that additional variables or more complex models may be needed to adequately predict hair loss.

Linear Discriminant Analysis

Given LDA's assumption of predictor variable normality, we conducted transformations on the Age variable to align with this requirement. In selecting categorical variables, we opted to include only those capable of binary representation. Variables such as Medical Conditions and Nutritional Deficiencies presented challenges due to their numerous and diverse categories, making it difficult to consolidate them into fewer groups.

```
## Call:
## lda(Hair.Loss ~ Genetics + Hormonal.Changes + Poor.Hair.Care.Habits +
##     Environmental.Factors + Smoking + Weight.Loss + Age, data = train_data)
##
## Prior probabilities of groups:
##        No       Yes
## 0.5181477 0.4818523
##
## Group means:
##     GeneticsYes Hormonal.ChangesYes Poor.Hair.Care.HabitsYes
## No    0.4782609           0.5072464                0.5144928
## Yes   0.5506494           0.5220779                0.4753247
##     Environmental.FactorsYes SmokingYes Weight.LossYes      Age
## No                 0.5265700  0.5628019      0.4589372 34.97826
## Yes                0.5116883  0.4831169      0.4961039 33.73766
##
## Coefficients of linear discriminants:
##                                   LD1
## GeneticsYes               1.01896951
## Hormonal.ChangesYes       0.21036020
## Poor.Hair.Care.HabitsYes -0.51479040
## Environmental.FactorsYes -0.29446891
## SmokingYes               -1.11184435
## Weight.LossYes            0.51229352
## Age                      -0.05120366
```

We can see that the coefficients between the groups are very similar which probably means that they are not very different.

```
## [1] "Accuracy: 0.48"
```

```
##       Predicted
## Actual No Yes
##    No  47  41
##    Yes 63  49
```

As expected after seeing the coefficients, the accuracy is relatively low and this model doesn't accurately predict hair loss. We are better off randomly guessing the class.

Conclusion

In conclusion, we wanted to figure out whether it was one variable or a combination of variables that affected hair loss. It was difficult to come to a conclusion because the accuracy of the models were weak. Based on the confusion matrices and the accuracy rate of 0.60, which was higher than the rate of other predictors, weight loss seems to be the most promising

predictive variable. In general the reason why the accuracy of the models were weak was because of limited predictor variables, numerous categorical variables, and the lack of training data. It was also relatively hard to identify which out of each categorical variable was the most influential to hair loss. Variables such as Medical Conditions, Nutritional Deficiencies, and Medication Treatments had so many different data entries that it was hard to pinpoint what was actually contributing to the hair loss people were facing. We definitely needed more information about the categorical predictors. For example, a list of detailed environmental factors would be much better than having a list of yeses and nos. If we had more numerical predictors, we would not have to heavily depend on binary variables and our model accuracy would improve. More information on categorical variables, especially weight loss since it was our most promising predictive variable, would give us more insight to whether or not those variables are actually affecting hair loss. Another tactic that we could implement in the future is future engineering. By transforming existing features and creating new features or using interactions to capture more complex relationships, we would be able to get higher accuracy rates and we would be able to better predict which variables affect hair loss.

Discussion

Both logistic regression and LDA models demonstrated a relatively low accuracy rate, hovering around 50-60%. Several factors likely contributed to these findings. Firstly, the models may have been limited by the constrained number of predictor variables, hindering their ability to capture the complexity of underlying relationships within the data. Moreover, the dataset contained a substantial amount of categorical predictor variables, which posed challenges in effectively utilizing them within the models. Additionally, the size of the training dataset may have been insufficient for the models to learn representative patterns adequately. Moreover, the need for more detailed information on categorical predictors, such as nuanced data on environmental factors, could potentially enhance the models' predictive capabilities. Lastly, the complexity of certain variables, such as medical conditions and medication treatments, presented challenges in accurately capturing their nuances, thereby affecting the models' predictive performance.

To overcome these limitations and potentially enhance model performance, several strategies can be explored. Seeking out a more comprehensive dataset with detailed categorical variables, particularly those related to weight loss, could provide richer information for model learning. Additionally, incorporating more numerical predictor values and implementing feature engineering techniques to create new features or capture complex relationships may improve the models' ability to capture underlying patterns in the data. By addressing these challenges, logistic regression and LDA models can offer better insights into the factors influencing hair loss and potentially improve their predictive accuracy.

Citations

Ho CH, Sood T, Zito PM. Androgenetic Alopecia. [Updated 2024 Jan 7]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK430924/

Williamson, D., Gonzalez, M., & Finlay, A. Y. (2001). The effect of hair loss on quality of life. Journal of the European Academy of Dermatology and Venereology, 15(2), 137-139.