Multifactorial Analysis of Baldness Prevalence in Individuals

Bala Yaswanth Kumar. S[1], Harsha Kamineni[1], Rakesh Reddy Tippa[1], Sai Meghana Imadabattini[1], Yashita Raga Saranam[1]

[1]Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202, United States of America
bkumars@iu.edu, Hskamine@iu.edu, rtippar@iu.edu, saimad@iu.edu, ysaranam@iu.edu

Abstract: Hair loss, a condition affecting a substantial segment of the population, has emerged as a mounting concern in recent times. This research endeavor embarks on an extensive exploration of the multifarious determinants that influence the manifestation of baldness among individuals. By considering a comprehensive array of factors encompassing genetics, hormonal dynamics, environmental influences, and lifestyle choices, the study aims to shed light on the intricate interplay of variables that contribute to this condition. Leveraging a robust dataset and employing rigorous statistical analysis techniques, the overarching objective is to identify significant predictors that can enhance our understanding of baldness. This research provides medical professionals with insights to develop better intervention strategies for hair loss, enhancing understanding and management of the condition, and contributing significantly to the collective knowledge base.
Keywords: Hair loss, genetics, environmental influences.

1 Project Scope

1.1 Introduction

Hair loss, a condition affecting millions globally, poses challenges not only to individuals' self-esteem but also to the medical community striving to comprehend its intricate causative factors. In this presentation, we delve into a comprehensive analysis of various determinants influencing the presence of baldness in individuals. Environmental factors, including exposure to pollutants and toxins, have been shown to impact hair health by disrupting hair growth cycles and damaging the hair follicle structure (Sibbald C. 2023).

Lifestyle choices such as diet, stress levels, and smoking also significantly influence hair health. Nutritional deficiencies can lead to weakened hair structure and increased shedding, while stress has been linked to conditions such as telogen effluvium, where hair prematurely enters the resting phase and subsequently falls out (O'Connor, K et., al 2021).

1.2 Aim

The aim of this study is to explore the collective influence of various factors, including genetics, hormonal changes, medical conditions, medications and treatments, nutritional deficiencies, stress levels, age, poor hair care habits, environmental factors, smoking habits, and weight loss, on the presence of baldness in individuals.
Based on our aim, we formulated the two hypotheses:

Null Hypothesis (H0):

None of the factors such as genetics, hormonal changes, medical conditions, medications and treatments, nutritional deficiencies, stress levels, age, hair care habits, environmental factors, smoking habits, and weight loss have an effect on the presence of baldness in individuals.

Alternate Hypothesis (H1):

At least one of the factors such as genetics, hormonal changes, medical conditions, medications and treatments, nutritional deficiencies, stress levels, age, hair care habits, environmental factors, smoking habits, and weight loss does have an effect on the presence of baldness in individuals.

1.3 Purpose

This purpose of our research is to uncover the diverse factors behind hair loss, including genetics, hormones, environment, and lifestyle, to offer insights for improved management strategies and understanding among medical professionals.

2 Methodology

2.1 Steps of the Project

Our project is dedicated to analyzing the multitude of factors affecting baldness. We utilized R for statistical analyses and for data visualization to enhance our understanding of these complex interactions.

We divided our methodology into five key stages, each crucial for advancing our research:

1. Data Collection: We initiated our data analysis by sourcing and collecting the relevant dataset needed for our research objectives.
2. Data Extraction and Cleaning: Once acquired, the data was carefully extracted from its original source and stored for future use and analysis.
3. Data Analysis: Our rigorous analysis involved statistical techniques to unveil meaningful insights from the data, identifying patterns and correlations between physical attributes and tumor classification.
4. Data Visualization: Using ‘R’, we visualized the results of our statistical tests and regression models to clearly communicate the findings.
5. Feature Selection and Model Refinement: Following initial analyses, we applied feature selection techniques to refine our models. This allowed us to focus on the most significant predictors and conduct a deeper analysis with these refined models.

3 Data Collection

The dataset, sourced from Kaggle, consists of 999 observations and 12 variables related to genetic predispositions and lifestyle choices. Each entry in the dataset corresponds to an individual case, detailing the presence or absence of each factor and culminating in the outcome variable of Hair Loss, represented as a binary indicator.

Link for dataset: https://www.kaggle.com/datasets/amitvkulkarni/hair-health?resource=download

Types and Summary of the Variables:

<table>
<tr><th colspan="2">Categorical Variables</th><th>Numerical Variables</th></tr>
<tr>
<td>A) Nominal Variables:
Genetics : Yes, No,
Hormonal Changes : Yes, No
Medical Conditions : Eczema, Dermatosis, Ringworm, etc.
Medications/Treatments : Antibiotics, Antifungal Cream, Accutane, etc.
Nutritional Deficiencies : Magnesium deficiency, Protein deficiency, Biotin Deficiency, etc.
Poor Hair Care Habits : Yes, No
Environmental Factors : Yes, No
Smoking : Yes, No
Weight Loss : Yes, No</td>
<td>B) Ordinal Variables:
Stress : Low, Moderate, High
Numerical Variables</td>
<td>Discrete variables
Age
Hair Loss: 0 or 1
We don’t have any Continuous Variables .</td>
</tr>
</table>

4 Data Extraction and Storage

4.1 Data Extraction

The first step in the data extraction and storage stage was identifying the data sources to extract the data from. Once the data source was identified, which is a Kaggle dataset in the case of our project, the next step was to extract the required data and store it on a relational database such as ‘R’.

4.2 Data Import

The hair health prediction data, stored in a CSV file, was first imported into R utilizing the "read.csv" command, which facilitated the process of reading and parsing the contents of the CSV file, thereby enabling access to the information contained within for subsequent analysis and manipulation within the R environment.

After loading, the initial steps in exploring the modified dataset included examining the first few rows through the command "head(data)," enabling a quick glimpse into its contents. Subsequently, a thorough summary was generated using "summary(data)," offering detailed statistics and distributions of the dataset's variables. Further exploration delved into understanding the dataset's overall dimensions with "dim(data)," shedding light on the number of attributes (columns) and instances (rows) present. Additionally, the structure of the dataset was scrutinized via "str(data),"

providing a comprehensive overview of variable types, their respective attributes, and any potential missing values or anomalies, thereby facilitating a deeper understanding of the dataset's composition and characteristics.

4.3 Data Cleaning

After meticulously verifying the absence of null values and identifying a solitary duplicate entry within the dataset, we promptly removed it to uphold data integrity. Our meticulous approach continued as we shifted our focus towards the crucial task of identifying and managing outliers. Recognizing that age was the only numerical column available, we concentrated our outlier detection efforts specifically on this variable. Employing the robust Interquartile Range (IQR) method, complemented by the visual insights provided by box plots, we systematically scrutinized the age data for any anomalies. Our thorough analysis culminated in the reassuring finding that no outliers were detected within the age column. This rigorous process ensures the integrity and reliability of our dataset for subsequent analyses and interpretations.

```r
# Check for null values in the dataframe
null_values <- is.na(data)
# Check if there are any null values overall
if (any(null_values)) {
  cat("There are null values in the dataset.\n")
  # Print the count of null values for each column
  print(colSums(null_values))
} else {
  cat("There are no null values in the dataset.\n")
}
```

```
There are no null values in the dataset.
```

Fig 1: Code to verify Null values

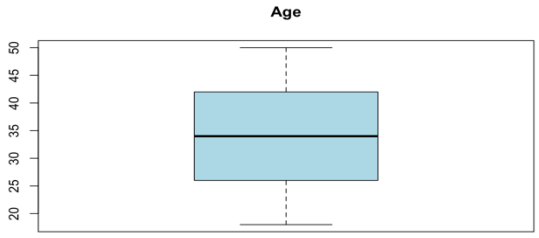

Fig 2: Box Plot visualization

After conducting an assessment of categorical variables, it was observed that certain columns contained entries labeled as 'No Data.' To address this, the 'No Data' entries were substituted with the mode of their respective columns, representing the most frequently occurring value. Subsequently, the revised unique values post-replacement was meticulously documented for further analysis.

5 Exploratory Data Analysis

Through a meticulous examination of summary statistics, we delved deeply into essential metrics such as the mean, median, and other pertinent measures. This exhaustive analysis afforded us intricate insights into the central tendencies and distributional attributes present within the dataset, offering a rich understanding of its underlying structure and characteristics.

5.1 Checking for Normality

To evaluate the normality of the dataset, we utilized a comprehensive method involving a histogram, Q-Q plot, and the Shapiro-Wilk test. This multifaceted approach enabled us to thoroughly examine the distribution traits, providing insights into whether the data conforms to a normal distribution pattern. Significantly, our analysis indicated that the age variable exhibits deviations from a normal distribution.

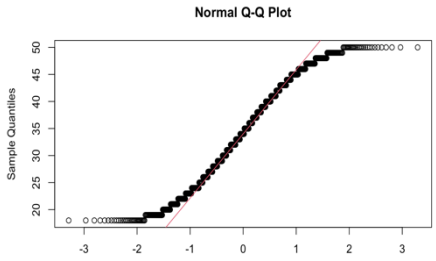

Fig 3: Q-Q Plot visualization

```
        Shapiro-Wilk normality test

data:  data$Age
W = 0.95763, p-value < 2.2e-16
```

Fig 4: Shapiro-Wilk test

To address the non-normal distribution observed in specific dataset columns, we employed a log transformation strategy on these variables. The objective of this transformation was to move the variables towards a more normal distribution, thereby enhancing the robustness of subsequent statistical analyses. We then visualized the effects of the log transformation through histograms depicting the distribution of selected columns before and after the transformation. This visual comparison offered clarity on how the log transformation influenced the variables, ultimately achieving a more normalized distribution for our analysis.

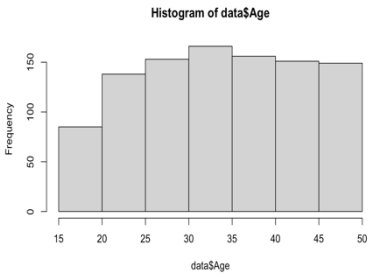

Fig 5: Histogram before Transformation

```r
columns_to_transform <- c(
  "Age"
)
for (col in columns_to_transform) {
  data[[col]] <- ifelse(data[[col]] > 0, log(data[[col]] + 1), 0)
}
par(mfrow = c(3, 3))
for (col in columns_to_transform) {
  hist(data[[col]], main = col, col = "lightblue", border = "black")
}
```

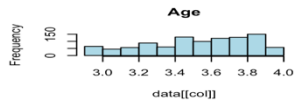

Fig 6: Histogram after Transformation

5.2 Chi-square test:

In our investigation into potential connections between factors like genetics, hormonal changes, and medical conditions with hair loss, we conducted Chi-square tests. These tests evaluate whether there is a significant association between categorical variables. However, our analysis revealed that all p-values derived from the Chi-square tests were above the conventional threshold of 0.05. Consequently, we failed to reject the null hypothesis for each test, indicating no statistically significant relationship between variables such as genetics, hormonal changes, and medical conditions, and the occurrence of hair loss within our dataset. While these results suggest no direct link in our sample, it's important to consider the limitations of our analysis and the possibility of other variables impacting hair loss not accounted for in our study.

6 Multiple Logistic regression Model

Before proceeding with the logistic regression model, we conducted an assessment of the dataset's balance. This involved examining the distribution between instances of hair loss and non-hair loss cases. Our analysis indicated that the dataset was balanced, with an equitable distribution between the two categories. This finding assures

the reliability of our model, as an imbalanced dataset could skew the predictive performance of the logistic regression analysis.

How do various factors such as genetics, hormonal changes, medical conditions, medications and treatments, nutritional deficiencies, stress levels, age, poor hair care habits, environmental factors, smoking habits, and weight loss collectively influence the presence of baldness in individuals?

Null Hypothesis (H0): None of the factors such as genetics, hormonal changes, medical conditions, medications and treatments, nutritional deficiencies, stress levels, age, hair care habits, environmental factors, smoking habits, and weight loss have an effect on the presence of baldness in individuals.

Alternate Hypothesis (H1): At least one of the factors such as genetics, hormonal changes, medical conditions, medications and treatments, nutritional deficiencies, stress levels, age, hair care habits, environmental factors, smoking habits, and weight loss does have an effect on the presence of baldness in individuals.

Rationale:

In alignment with our research inquiry, we initiated a multiple logistic regression analysis where hair loss served as the outcome variable, while the remaining variables functioned as predictors. Initially, none of the independent variables demonstrated a significant association with the dependent variable, prompting us to pursue feature selection to enhance model efficacy.

Following feature selection, our regression analysis identified a notable association between hair loss and the medical condition of thyroid problems. Subsequently, we sought to explore potential interactions between smoking habits and medical conditions. To facilitate this examination, we transformed the smoking variable into two binary variables, "smoking_yes" and "smoking_no," utilizing one-hot encoding techniques.

Upon incorporating the interaction terms into the logistic regression model, we observed significant associations between hair loss and various factors, including smoking_yes, the presence of eczema, thyroid problems, and the interaction between smoking_yes and eczema. Consequently, we reject the null hypothesis and assert that at least one of the factors encompassing genetics, hormonal changes, medical conditions, medications and treatments, nutritional deficiencies, stress levels, age, hair care habits, environmental factors, smoking habits, and weight loss indeed influences the presence of baldness in individuals. This comprehensive analysis underscores the multifaceted nature of hair loss determinants and underscores the importance of considering interactions between variables in understanding its etiology.

7 Conclusion

Our comprehensive analysis of factors influencing hair loss, initially inconclusive, culminated in the rejection of the null hypothesis. Through refined logistic regression, we identified significant associations between thyroid problems, smoking habits, and their interaction with eczema, emphasizing the multifactorial nature of baldness determinants. This insight aids in developing more effective intervention strategies and enhances our understanding of hair loss etiology.

8 References

O'Connor, K., & Goldberg, L. J. (2021). Nutrition and hair. *Clinics in Dermatology*, *39*(5), 809–818. https://doi.org/10.1016/j.clindermatol.2021.05.008

Sibbald, C. (2023). Alopecia areata: an updated review for 2023. *Journal of Cutaneous Medicine and Surgery*, *27*(3), 241–259. https://doi.org/10.1177/12034754231168839

9 Appendix

9.1 Chi-square test interpretation

```
Chi-square test for Genetics :

        Pearson's Chi-squared test with Yates' continuity correction

data:  data[[variable]] and data$Hair.Loss
X-squared = 1.47, df = 1, p-value = 0.2253

There is no significant association between Genetics and Hair Loss.

Chi-square test for Hormonal.Changes :

        Pearson's Chi-squared test with Yates' continuity correction

data:  data[[variable]] and data$Hair.Loss
X-squared = 0.037549, df = 1, p-value = 0.8464

There is no significant association between Hormonal.Changes and Hair Loss.

Chi-square test for Medical.Conditions :

        Pearson's Chi-squared test

data:  data[[variable]] and data$Hair.Loss
X-squared = 9.6405, df = 9, p-value = 0.3804

There is no significant association between Medical.Conditions and Hair Loss.

Chi-square test for Medications...Treatments :

        Pearson's Chi-squared test

data:  data[[variable]] and data$Hair.Loss
X-squared = 3.5955, df = 9, p-value = 0.936

There is no significant association between Medications...Treatments and Hair Loss.
```

Fig 7: Chi-Square Interpretation of some variables

9.2 Evaluation of Dataset Balance

```
```{r}
# Count occurrences of each unique value in the "Hair.Loss" column
hair_loss_counts <- table(data$Hair.Loss)

# Print the counts
print(hair_loss_counts)
```
```

```
  0   1
502 496
```

Fig 8: Image showing Balanced Dataset

9.3 Multiple logistic regression

```r
model <- glm(Hair.Loss ~ Age + Genetics + Hormonal.Changes + Medical.Conditions +
             Nutritional.Deficiencies + Stress + Poor.Hair.Care.Habits + Environmental.Factors +
             Smoking + Weight.Loss,
             data = data, family = binomial(link = "logit"))
summary(model)
```

Fig 9: Code for regression model

```
Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                               1.395870   0.860978   1.621   0.1050
Age                                      -0.383736   0.229691  -1.671   0.0948 .
GeneticsYes                               0.178125   0.129506   1.375   0.1690
Hormonal.ChangesYes                       0.027219   0.130344   0.209   0.8346
Medical.ConditionsAndrogenetic Alopecia   0.022111   0.273590   0.081   0.9356
Medical.ConditionsDermatitis             -0.290810   0.275124  -1.057   0.2905
Medical.ConditionsDermatosis             -0.253769   0.277133  -0.916   0.3598
Medical.ConditionsEczema                 -0.285445   0.299394  -0.953   0.3404
Medical.ConditionsPsoriasis              -0.343891   0.269541  -1.276   0.2020
Medical.ConditionsRingworm               -0.350788   0.296326  -1.184   0.2365
Medical.ConditionsScalp Infection        -0.091169   0.286686  -0.318   0.7505
Medical.ConditionsSeborrheic Dermatitis   0.023152   0.280317   0.083   0.9342
Medical.ConditionsThyroid Problems       -0.420170   0.274665  -1.530   0.1261
Nutritional.DeficienciesIron deficiency   0.121329   0.296573   0.409   0.6825
Nutritional.DeficienciesMagnesium deficiency  0.262387   0.283258   0.926   0.3543
Nutritional.DeficienciesOmega-3 fatty acids   0.080164   0.283404   0.283   0.7773
Nutritional.DeficienciesProtein deficiency    0.216818   0.285899   0.758   0.4482
Nutritional.DeficienciesSelenium deficiency   0.147029   0.292045   0.503   0.6147
Nutritional.DeficienciesVitamin A Deficiency  0.170549   0.274336   0.622   0.5342
Nutritional.DeficienciesVitamin D Deficiency  0.051519   0.272978   0.189   0.8503
Nutritional.DeficienciesVitamin E deficiency -0.130173   0.291354  -0.447   0.6550
Nutritional.DeficienciesZinc Deficiency       0.007747   0.269590   0.029   0.9771
StressLow                                     0.009224   0.160648   0.057   0.9542
StressModerate                                0.154500   0.158435   0.975   0.3295
Poor.Hair.Care.HabitsYes                     -0.155395   0.129994  -1.195   0.2319
Environmental.FactorsYes                     -0.030598   0.129750  -0.236   0.8136
SmokingYes                                   -0.232494   0.130027  -1.788   0.0738 .
Weight.LossYes                                0.205730   0.129210   1.592   0.1113
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1383.5  on 997  degrees of freedom
Residual deviance: 1361.8  on 970  degrees of freedom
AIC: 1417.8

Number of Fisher Scoring iterations: 4
```

Fig 10: Interpretation showing no association.

9.4 Feature Selection

```r
data$Hair.Loss <- factor(data$Hair.Loss)
p_values <- sapply(names(data), function(variable) {
  if (is.factor(data[[variable]]) && variable != "Hair.Loss") {
    chi_square_result <- chisq.test(data[[variable]], data$Hair.Loss)
    return(chi_square_result$p.value)
  } else {
    return(NA)
  }
})
# Select the top 5 variables
top_5_features <- names(sort(p_values, decreasing = FALSE))[1:5]

# Print the top 5 features
print(top_5_features)
```

```
[1] "Smoking"             "Weight.Loss"         "Genetics"            "Poor.Hair.Care.Habits"
[5] "Stress"
```

Fig 11: Code and Interpretation of Feature Selection

9.5 Interaction Between Variables

```
# Convert 'Smoking' to one-hot encoding
data$Smoking_Yes <- as.integer(data$Smoking == "Yes")
data$Smoking_No <- as.integer(data$Smoking == "No")

# Rename the columns
colnames(data)[colnames(data) == "Smoking_Yes"] <- "Smoking_Yes"
colnames(data)[colnames(data) == "Smoking_No"] <- "Smoking_No"

```

```{r}
data <- subset(data, select = -c(Smoking))
```

```{r}
data
```

Environmental.Factors <fctr>	Weight.Loss <fctr>	Hair.Loss <fctr>	Genetics_Yes <int>	Genetics_No <int>	Smoking_Yes <int>	Smoking_No <int>
Yes	No	0	1	0	0	1
Yes	No	0	0	1	0	1
Yes	Yes	0	0	1	0	1
Yes	No	0	1	0	0	1
Yes	No	1	0	1	1	0
Yes	Yes	1	1	0	0	1
No	No	1	1	0	0	1
No	No	0	1	0	1	0

Fig 12: One-hot encoding for Interaction

9.6 Multiple logistic regression

```
Call:
glm(formula = Hair.Loss ~ Smoking_Yes + Medical.Conditions +
    Smoking_Yes * Medical.Conditions, family = binomial(link = "logit"),
    data = data)

Coefficients:
                                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                                         0.6931     0.2739   2.531  0.01137 *
Smoking_Yes                                        -0.7949     0.3781  -2.102  0.03552 *
Medical.ConditionsAndrogenetic Alopecia            -0.5534     0.4105  -1.348  0.17759
Medical.ConditionsDermatitis                       -0.6931     0.4024  -1.722  0.08500 .
Medical.ConditionsDermatosis                       -0.5447     0.3866  -1.409  0.15886
Medical.ConditionsEczema                           -1.1180     0.4151  -2.694  0.00707 **
Medical.ConditionsPsoriasis                        -0.7357     0.4002  -1.838  0.06600 .
Medical.ConditionsRingworm                         -0.3567     0.4351  -0.820  0.41232
Medical.ConditionsScalp Infection                  -0.6931     0.3979  -1.742  0.08151 .
Medical.ConditionsSeborrheic Dermatitis            -0.4418     0.3996  -1.106  0.26883
Medical.ConditionsThyroid Problems                 -0.8416     0.3866  -2.177  0.02950 *
Smoking_Yes:Medical.ConditionsAndrogenetic Alopecia  0.8714     0.5465   1.594  0.11085
Smoking_Yes:Medical.ConditionsDermatitis             0.6518     0.5493   1.187  0.23536
Smoking_Yes:Medical.ConditionsDermatosis             0.2587     0.5530   0.468  0.63986
Smoking_Yes:Medical.ConditionsEczema                 1.2198     0.5833   2.091  0.03651 *
Smoking_Yes:Medical.ConditionsPsoriasis              0.8375     0.5451   1.537  0.12441
Smoking_Yes:Medical.ConditionsRingworm               0.1708     0.5954   0.287  0.77424
Smoking_Yes:Medical.ConditionsScalp Infection        0.4501     0.5717   0.787  0.43108
Smoking_Yes:Medical.ConditionsSeborrheic Dermatitis  0.6237     0.5548   1.124  0.26093
Smoking_Yes:Medical.ConditionsThyroid Problems       0.5787     0.5341   1.084  0.27855
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1383.5  on 997  degrees of freedom
Residual deviance: 1365.2  on 978  degrees of freedom
AIC: 1405.2

Number of Fisher Scoring iterations: 4
```

Fig 13: Logistic regression showing association with outcome variable after feature selection.