UCLA College | Physical Sciences Statistics & Data Science

Team 21: Diabetes Research Project

Authors: Francis Chan, Evan Lum, Dylan Mochizuki, Gina Pak, Elise Pham, Khang Thai
University of California, Los Angeles

Introduction

This research analyzes the health demographics and genetic predispositions that contribute to diabetes risk.

This leads us to our research:
1. What factors contribute to determining whether an individual has diabetes?

2. Is there a combined effect of BMI and HBA1C on diagnosed individuals?

3. Does family history impact the relationship between age and risk of diabetes?

Methodology

This dataset contains 9,538 simulated medical records related to diabetes diagnosis with roughly one-third being confirmed positive. We investigate the factors that contribute to diabetes diagnosis and examine how they interacts with diabetes risk.

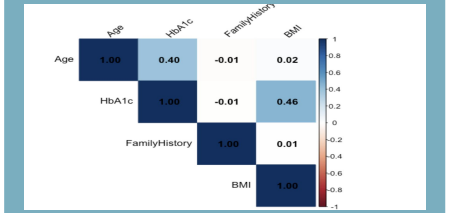
An observational study design will be used with a secondary dataset containing medical records of individuals. We will use statistical modeling to identify relationships between variables and diabetes diagnosis.

We began our analysis by summarying key characteristics of the dataset, implementing visualizations and bivariate analysis to explore the relationships between individual variables.

Approach

1. Perform Logistic Regression
2. Perform AIC alongside Analysis of Deviance (ANOVA)
3. Conclude Age, HbA1c, Family History, and BMI contribute to diabetes risk the most
4. Visualize how features interact together

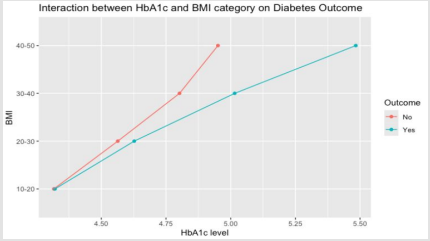
Performance



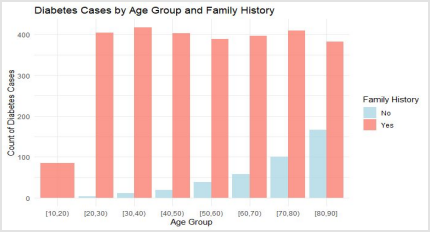
After performing logistic regression, our model yielded high p-values, suggesting that it was not statistically significant. This led us to use the Akaike information criterion (AIC), a forward stepwise regression model to iteratively remove variables based on the AIC value.

Based on this, we discovered that the best-fit model was the model with Age, BMI, HbA1c (blood sugar test), and FamilyHistory as predictor variables. We further validated this model using ANOVA to compare the full and reduced model. Finally, we assessed predictive performance of our model by implementing 5-fold cross-validation with above average accuracy.

Discussion



The plot models the interaction between a person's HbA1c and BMI based on whether they have diabetes. When compared at the same BMI, those diagnosed with diabetes tend to have a higher HbA1c levels. This result is consistent as high BMI is a significant indicator for elevated Hemoglobin AIC.



When looking directly at only those with diabetes, it is noticeable that family history plays a large role with diagnosis. Those who don't have a genetic disposition are often diagnosed at a later age.

Conclusion

- Best performing features to determine risk of diabetes are Age, BMI, HbA1c, and Family History
- Family History has the strongest correlation amongst all variables
- Those diagnosed with diabetes have a higher HbA1c level than those who do not, when compared at the same BMI
- The number of diabetes cases is consistent amongst all age groups of those who have a family history

Limitations

- Missing Confounders i.e. physical activity, dietary habits, socioeconomic status, etc.
- Potential measurement errors from dataset
- Risk of model selection bias from AIC

Acknowledgements

Tech. Diabetes Dataset, Version 1. Retrieved February 2025 from www.kaggle.com/datasets/asinow/diabetes-dataset.

We would like to acknowledge Professor Vivian Lew for her guidance and support throughout the Stats 140XP course.

```
AIC: 1351

Number of Fisher Scoring iterations: 20
```

```r
vif(final_model)
```

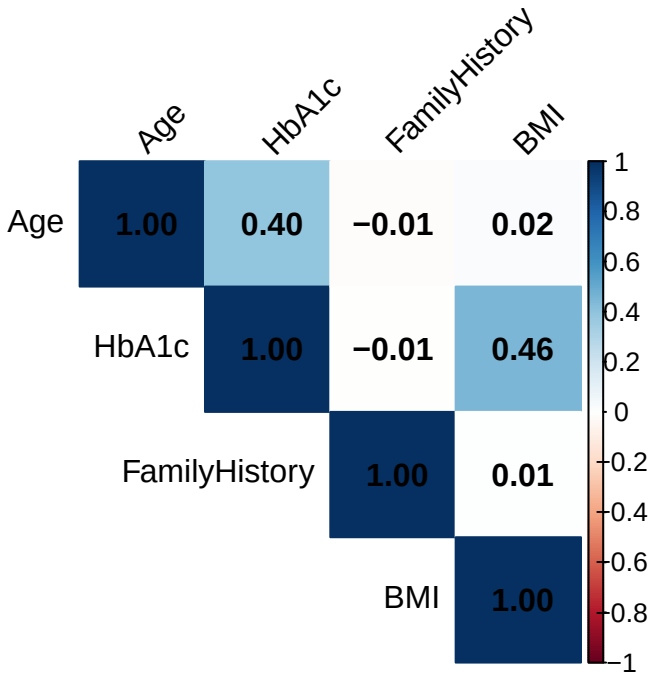
```
                Age                HbA1c factor(FamilyHistory)
           1.115542             1.002088                1.000001
                BMI
           1.117578
```

Create correlation heatmap

```r
cor_matrix <- cor(final_model_dataset, use = "pairwise.complete.obs")
diag(cor_matrix) <- 1

#png(
#  "correlation_heatmap_large.png",
 # width = 2000,
 # height = 2000,
 # res = 120)

heatmap <- corrplot(
  cor_matrix,
  method = "color",
  type = "upper",
  addCoef.col = "black",
  tl.col = "black",
  tl.srt = 45,
  diag = TRUE)
```




```
#dev.off()
heatmap
```

```
$corr
                     Age        HbA1c FamilyHistory         BMI
Age           1.00000000  0.396354303  -0.012887442 0.021793701
HbA1c         0.39635430  1.000000000  -0.007980574 0.459851892
FamilyHistory -0.01288744 -0.007980574   1.000000000 0.008037471
BMI           0.02179370  0.459851892   0.008037471 1.000000000

$corrPos
           xName         yName x y         corr
1            Age           Age 1 4  1.000000000
2          HbA1c           Age 2 4  0.396354303
3          HbA1c         HbA1c 2 3  1.000000000
4  FamilyHistory           Age 3 4 -0.012887442
5  FamilyHistory         HbA1c 3 3 -0.007980574
6  FamilyHistory FamilyHistory 3 2  1.000000000
7            BMI           Age 4 4  0.021793701
8            BMI         HbA1c 4 3  0.459851892
9            BMI FamilyHistory 4 2  0.008037471
10           BMI           BMI 4 1  1.000000000
```

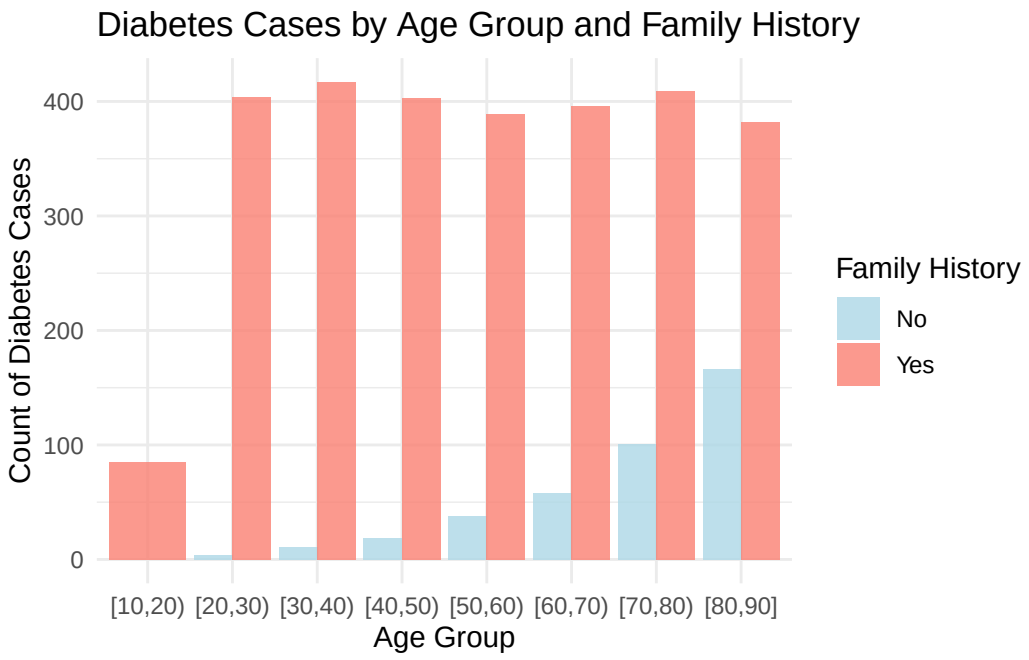
```
$arg
$arg$type
[1] "upper"
```

Visualization

```r
temp_data <- dataset %>%
  mutate(AgeGroup = cut(Age, breaks = seq(10, 90, by = 10), include.lowest = TRUE, right = FA
         FamilyHistory = factor(FamilyHistory, levels = c(0, 1), labels = c("No", "Yes")))

diabetes_data <- temp_data %>% filter(Outcome == 1)

ggplot(diabetes_data, aes(x = AgeGroup, fill = FamilyHistory)) +
  geom_bar(position = "dodge", alpha = 0.8) +
  labs(title = "Diabetes Cases by Age Group and Family History",
       x = "Age Group",
       y = "Count of Diabetes Cases",
       fill = "Family History") +
  scale_fill_manual(values = c("No" = "lightblue", "Yes" = "salmon")) +
  theme_minimal()
```



```r
temp_dataset <- dataset %>%
  mutate(BMI_category = cut(BMI,
                            breaks = c(10, 20, 30, 40, 50),
                            labels = c("10-20", "20-30", "30-40", "40-50"),
                            include.lowest = TRUE))

temp_dataset$Outcome <- factor(temp_dataset$Outcome, levels = c(0, 1), labels = c("No", "Yes"



ggplot(data=temp_dataset,
       mapping = aes(x = HbA1c,
                     y = factor(BMI_category),
                     group = Outcome,
                     color = Outcome)) +
  stat_summary(fun = "mean", geom = "point") +
  stat_summary(fun = "mean", geom = "line") +
  labs(title = "Interaction between HbA1c and BMI category on Diabetes Outcome",
      x = "HbA1c level",
      y = "BMI",
      color = "Outcome")
```