**Heart disease analysis using logistic model**

**Issue**

Different issues will be determined during the analysis of the data. Age groups will be used to compare how different individuals are educated on heart health. It’s achieved through grouping the age column into different groups and the show how different groups are educated.

The most affected age group will be looked at during the analysis. The health data can be used to show how different ages are affected by the heart disease. Through grouping of the data into different age groups, the data show how the different age sets are affected by the diseases using the model.

 The health data will be used to clearly depict the most popular signs and symptoms in the data and among different age groups as well as age sets. Different age groups can be used to clearly indicate the most common symptoms among the different age sets. Using bar chart different age set can be used to indicate the different symptoms which have been showing up as symptom of the heart disease.

**Findings**

From the data analysis it indicates there is much correlation between the coughing signal symptoms and the individual from all different type of gender. The correlation changes with change in the median. If the median is assumed to be two the correlation ranking shift to the edema. It clearly indicates that there is so much correlation between different symptoms and different age groups as well as age sets for the health heart disease data set.

As the data changes to mean during analyses the correlation in the data will shift to the cough and the edam ani the data. Throughout the data number of the relation in the mean and the symptoms the data will change but the correlation among the data the different ages groups will remain the same.

**Discussion**

The data has different correlation depending on the analysis of the data. For the example using the mean the there is large correlation on the edam symptom as well as age . the probability of seeking treatment with two day change on different occasion depending the age of the different individuals. The number of delayed treatment sought increase through up to the number of delayed treatment sought where is equal to on one. The number of null delayed day is lest is lest than the number of the delayed status as well as the number of delayed status is equal to one.

**Appendix A: data collection**

The health heart dataset used was collected by during a health heart study. The data was aimed at seeking the delays and factors associated with delays in heart associated disease.

Individuals were asked about their demographic characteristic, medical history and the symptoms they experience during different times. The delay in seeking medical treatment was defined as number of days between the onset of symptoms and the first visited to a health care provider

 The survey was conducted use a standardized question and entered into an electronic database. The database was then exported as excel filer and imported to r Studio using R libraries.

Variables created in data set includes

Age,Gender,Ethnicity,Marital,Livewith,Education, palpitations,orthopnea,chestpain,nausea

,cough,fatigue,dyspnea.edema,PND,Tightshoes,Weightgain,DOE.

The delay variable was transformed into binary variable for the model development

**Appendix B:Results**

The data is summarized depending on different groups. The data is first of all summarized in term of age and the number of females and males available in data. It’s the further summarize in term of different symptoms.

Number of males in % is 55.13%

Number of females in % is 44.83%

The age means 74.25

Minimum age is 41

Maximum age is 96

p.value=n/a

**Appendix C:Data and Code**

library(na.tools)
library(ggplot2)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
## filter, lag

## The following objects are masked from 'package:base':
##
## intersect, setdiff, setequal, union

library(openintro)

## Loading required package: airports

## Loading required package: cherryblossom

## Loading required package: usdata

library(readxl)

## Warning: package 'readxl' was built under R version 4.2.3

library(ggpubr)

## Warning: package 'ggpubr' was built under R version 4.2.3

options(repos = "http://cran.rstudio.com")
install.packages("dplyr", repos = "http://cran.rstudio.com")

## Warning: package 'dplyr' is in use and will not be installed

library(dplyr)

library(knitr)

# Load the data
healthdata<-read\_excel("C:/Users/Osman/Desktop/health/Heart Health Data.xls")

# Descriptive statistics
mydesc <- summary(healthdata$Age)

# Create a table of results
agesumarrytable <- data.frame(
 "Variable" = c("Age"),
 "Mean" = round(mean(healthdata$Age), 2),
 "Standard Deviation" = round(sd(healthdata$Age), 2),
 "Minimum" = min(healthdata$Age),
 "Maximum" = max(healthdata$Age),
 "p-value" = NA
)
View(agesumarrytable)

# Format the table
# Calculate the number of males and females
n\_males <- sum(healthdata$Gender == 1)
n\_females <- sum(healthdata$Gender == 2)

# Calculate the percentage of males and females
pct\_males <- n\_males / nrow(healthdata) \* 100
pct\_females <- n\_females / nrow(healthdata) \* 100

# Print the results
cat("Number of males:", n\_males, "\n")

## Number of males: 224

cat("Number of females:", n\_females, "\n")

## Number of females: 182

cat("Percentage of males:", round(pct\_males, 2), "%\n")

## Percentage of males: 55.17 %

cat("Percentage of females:", round(pct\_females, 2), "%\n")

## Percentage of females: 44.83 %

View(healthdata)
ModelhealthData<-ifelse(healthdata$delaydays<=2,1,0)
datamodel<-glm(ModelhealthData ~Age +Gender+Ethnicity+ Marital+ Livewith + Education+ palpitations+orthopnea+chestpain+nausea+cough+fatigue+dyspnea+edema+PND+tightshoes+weightgain+DOE, data = healthdata, family = binomial())
summary(datamodel)

##
## Call:
## glm(formula = ModelhealthData ~ Age + Gender + Ethnicity + Marital +
## Livewith + Education + palpitations + orthopnea + chestpain +
## nausea + cough + fatigue + dyspnea + edema + PND + tightshoes +
## weightgain + DOE, family = binomial(), data = healthdata)
##
## Deviance Residuals:
## Min 1Q Median 3Q Max
## -2.1462 -1.1038 0.6974 1.0924 1.9241
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.497772 1.188270 0.419 0.67529
## Age 0.013161 0.009351 1.407 0.15929
## Gender -0.093782 0.217231 -0.432 0.66595
## Ethnicity -0.053748 0.187852 -0.286 0.77479
## Marital 0.074639 0.177241 0.421 0.67367
## Livewith -0.208708 0.264574 -0.789 0.43020
## Education 0.009505 0.077872 0.122 0.90285
## palpitations 0.140329 0.126346 1.111 0.26671
## orthopnea -0.041271 0.116699 -0.354 0.72360
## chestpain 0.126625 0.126599 1.000 0.31721
## nausea -0.092194 0.135773 -0.679 0.49712
## cough -0.314724 0.112493 -2.798 0.00515 \*\*
## fatigue -0.187713 0.139209 -1.348 0.17752
## dyspnea 0.093580 0.134549 0.696 0.48674
## edema -0.223813 0.123652 -1.810 0.07029 .
## PND -0.171321 0.111453 -1.537 0.12425
## tightshoes 0.129220 0.131284 0.984 0.32498
## weightgain 0.204976 0.113305 1.809 0.07044 .
## DOE -0.209921 0.124091 -1.692 0.09071 .
## ---
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 555.84 on 400 degrees of freedom
## Residual deviance: 522.49 on 382 degrees of freedom
## (5 observations deleted due to missingness)
## AIC: 560.49
##
## Number of Fisher Scoring iterations: 4

avg\_delay <- mean(healthdata$delaydays, na.rm = TRUE)
healthdata$BinaryDelay <- ifelse(healthdata$delaydays > avg\_delay, 0, 1)
healthmodel <- glm(BinaryDelay ~ Age + edema + DOE, data = healthdata, family = binomial())
summary(healthmodel)

##
## Call:
## glm(formula = BinaryDelay ~ Age + edema + DOE, family = binomial(),
## data = healthdata)
##
## Deviance Residuals:
## Min 1Q Median 3Q Max
## -1.7836 -1.4181 0.7627 0.8665 1.0500
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.4101421 0.7601768 1.855 0.0636 .
## Age -0.0009691 0.0091756 -0.106 0.9159
## edema -0.2135413 0.0959974 -2.224 0.0261 \*
## DOE -0.1233503 0.1083387 -1.139 0.2549
## ---
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 492.53 on 402 degrees of freedom
## Residual deviance: 485.96 on 399 degrees of freedom
## (3 observations deleted due to missingness)
## AIC: 493.96
##
## Number of Fisher Scoring iterations: 4

ggplot(healthdata, aes(x = factor(BinaryDelay), fill = factor(BinaryDelay))) +
 geom\_bar() +
 scale\_fill\_discrete(name = "Delay Status", labels = c("Delayed", "Not Delayed")) +
 xlab("Delay Status") +
 ylab("Count") +
 ggtitle("Proportion of Delayed and Not Delayed Cases")



median\_delays <- median(healthdata$delaydays)

# Create a new variable for whether the delay is 2 days or less (1) or more than 2 days (0)
healthdata$BinaryDelay <- ifelse(healthdata$delaydays <= 2, 1, 0)

# Build the logistic model
healthmodel <- glm(BinaryDelay ~ Age + cough + edema + weightgain + DOE, data = healthdata, family = binomial())

# Check which factors are most useful in predicting the outcome
summary(healthmodel)

##
## Call:
## glm(formula = BinaryDelay ~ Age + cough + edema + weightgain +
## DOE, family = binomial(), data = healthdata)
##
## Deviance Residuals:
## Min 1Q Median 3Q Max
## -1.7062 -1.1144 0.7643 1.0989 1.8435
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.068221 0.721755 -0.095 0.92470
## Age 0.011614 0.008713 1.333 0.18258
## cough -0.345986 0.107459 -3.220 0.00128 \*\*
## edema -0.175230 0.100512 -1.743 0.08127 .
## weightgain 0.215648 0.108684 1.984 0.04724 \*
## DOE -0.218431 0.105982 -2.061 0.03930 \*
## ---
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 558.65 on 402 degrees of freedom
## Residual deviance: 536.01 on 397 degrees of freedom
## (3 observations deleted due to missingness)
## AIC: 548.01
##
## Number of Fisher Scoring iterations: 4

healthdata$BinaryDelayforoneday <- ifelse(healthdata$delaydays <= 1, 1, 0)
healthmodel3 <- glm(BinaryDelayforoneday ~ Age + cough + edema + weightgain + DOE, data = healthdata, family = binomial())
summary(healthmodel3)

##
## Call:
## glm(formula = BinaryDelayforoneday ~ Age + cough + edema + weightgain +
## DOE, family = binomial(), data = healthdata)
##
## Deviance Residuals:
## Min 1Q Median 3Q Max
## -1.4229 -0.9064 -0.7281 1.2472 2.1020
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.957203 0.782019 -1.224 0.22095
## Age 0.016328 0.009465 1.725 0.08451 .
## cough -0.341243 0.115629 -2.951 0.00317 \*\*
## edema -0.275115 0.108136 -2.544 0.01095 \*
## weightgain 0.131737 0.116174 1.134 0.25681
## DOE -0.215040 0.110528 -1.946 0.05171 .
## ---
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 516.65 on 402 degrees of freedom
## Residual deviance: 491.83 on 397 degrees of freedom
## (3 observations deleted due to missingness)
## AIC: 503.83
##
## Number of Fisher Scoring iterations: 4