Final WriteUp

Stephen Robert, Gullet Cabdullahi, Kofi Owusu, Anisa Jeylani

Abstract:

The goal for this project was to use the data we found on the CDC website in 2021 Annual Survey to predict what were some of the explanatory variables that lead to heart disease in patients in the future. To find out the variables we will create a multi-logical model. We will also try to use the forward selection, backward elimination, and best subset to find the predictor variables for the heart disease (response variable). The final model will be constructed using the variables: Age, Stroke, BMI, Heart_Attack, HRT_Disease, Gen_Health and Sex. Since this was a random survey, the results can be used to infer to a larger population, probably to population of the world.

Introduction:

With 17.9 million deaths per year, cardiovascular diseases (CVDs) are the leading cause of death worldwide. Heart disease is a type of disease that affects the heart or blood vessels. Age, sex, or the patient's health are only a few examples of variables that may increase the chance of developing certain cardiac illnesses. Age, Stroke, BMI, Heart Attack, HRT Disease, Gen-Health, and Sex are the variables used in this example to predict heart disease. The ultimate objective of this study was to develop a model that may be used to forecast the causes of heart disease in individuals.

Data:

The Data set was created by the department of Center for Disease & Control (CDC,2021 Survey Data Information). This data set included 45782 cases in total (BRFSS,2021). This are the individuals that participated in the random survey that was done through phone. We chose to use the data set created in the year 2021, although there were a couple more data sets created in the previous years. This is because 2021 would be more recent and probably more accurate to study in. The response variable is the heart disease (Categorical) and the explanatory variables are: Age (Quantitative), BMI (Quantitative), Sex (Categorical), Stroke (Categorical), Gen_Health (Categorical), and Heart_Attack (Categorical).

```
BRFSS_Original <- read_csv("Group Project.csv", na = c("", NA, "N/A"))
```

```
New names:
Rows: 45782 Columns: 304
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(214): CTELENM1, PVTRESD1, COLGHOUS, STATERE1, CELPHON1, LADULT1, COLGSE... dbl
(90): ...1, _STATE, FMONTH, IDATE, IMONTH, IDAY, IYEAR, DISPCODE, SEQNO...
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
* `` -> `...1`
* `PHYSHLTH` -> `PHYSHLTH...35`
* `PHYSHLTH` -> `PHYSHLTH...271`
```

Renaming Variables:

```
BRFSS <- BRFSS_Original%>%
  rename(Sex = `_SEX`, Age = `_AGEG5YR`, Heart_Attack = `CVDINFR4`, HRT_Disease = `CVDCRHD
BRFSS <- BRFSS %>%
  rename(Stroke = `CVDSTRK3`)
```

naming the variables was only needed for some such as "CVDINFR4", but we renamed all of our variables anyway to make things easier to find, so we can just limit our variables to those we've renamed by using select().

Selecting Variables:

```
BRFSS <- BRFSS %>%
  select(Sex, Heart_Attack, HRT_Disease, Gen_Health, Stroke, BMI, Age)
```

The variables we chose are also the same variables we renamed; all but BMI are categorical, therefore we must use the skim() method to determine if they are the correct data type; if not, we must mutate them. We chose to use this variables in the study because we had a prediction that these variables would have a significant p value when compared to other variables in the data set.

Mutating Selected Variables:

```
BRFSS <- BRFSS %>%
  mutate(Sex = as_factor(Sex))

BRFSS <- BRFSS %>%
  mutate(Heart_Attack = as_factor(Heart_Attack))

BRFSS <- BRFSS %>%
  mutate(HRT_Disease = as_factor(HRT_Disease))

BRFSS <- BRFSS %>%
  mutate(Gen_Health = as_factor(Gen_Health))

BRFSS <- BRFSS %>%
  mutate(Stroke = as_factor(Stroke))

BRFSS <- BRFSS %>%
  mutate(BMI = as.numeric(BMI))
```

```
Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

```
BRFSS <- BRFSS %>%
  mutate(Age = as_factor(Age))
```

```
BRFSS <- BRFSS %>%
  mutate(Heart_Attack = fct_recode(Heart_Attack,
                                   `Yes` = "1",
                                   `No` = "2",
                                   `Don't know/Not sure` = "7",
                                   `Refused` = "9"))
BRFSS <- BRFSS %>%
  mutate(Sex = fct_recode(Sex,
                                   `Male` = "1",
                                   `Female` = "2"))
BRFSS <- BRFSS %>%
  mutate(HRT_Disease = fct_recode(HRT_Disease,
                                   `Yes` = "1",
                                   `No` = "2",
                                   `Don't know/Not sure` = "7",
```

```
                                `Refused` = "9"))
```

```
Warning: Unknown levels in `f`: 7, 9
```

```
BRFSS <- BRFSS %>%
  mutate(Gen_Health = fct_recode(Gen_Health,
                                 `Excellent` = "1",
                                 `Very good` = "2",
                                 `Good` = "3",
                                 `Fair` = "4",
                                 `Poor` = "5",
                                 `Don't know/Not Sure` = "7",
                                 `Refused` = "9"
                                 ))
BRFSS <- BRFSS %>%
  mutate(Stroke = fct_recode(Stroke,
                                `Yes` = "1",
                                 `No` = "2",
                                 `Don't know/Not sure` = "7",
                                 `Refused ` = "9"))
BRFSS <- BRFSS %>%
  mutate(Age = case_when(
    Age == 1 ~ 21,
    Age==2~27,
    Age==3~32,
    Age==4~37,
    Age==5~43,
    Age==6~47,
    Age==7~52,
    Age==8~57,
    Age==9~62,
    Age==10~67,
    Age==11~72,
    Age==12~74,
    Age==13~80,

  ))%>%
  mutate(HRT_Binary=case_when(HRT_Disease=="Yes"~1,
                              HRT_Disease=="No"~0))%>%
  filter(Heart_Attack%in%c("Yes","No"),
         Gen_Health%in%c("Good","Fair","Poor","Very good"),
```

```
        Stroke%in%c("Yes","No"))%>%
  filter(BMI!="NA")%>%
  mutate(BMI=ifelse(BMI==9999,NA,BMI/100))
```

The categorical values were mutated, and factored into outputted strings for values using the BRFSS Codebook (BRFSS, 2021) for interpreting which values meant what. We can skip factoring and leave BMI for the second chunk since it is already a quantitative numeric value. We did experience a difficulty as a result, which is covered below.

Regression Model:

```
m1<-glm(HRT_Binary
        ~Age+Stroke+Gen_Health+Sex+BMI+Heart_Attack, data=BRFSS,family=binomial)
summary(m1)
```

```
Call:
glm(formula = HRT_Binary ~ Age + Stroke + Gen_Health + Sex +
    BMI + Heart_Attack, family = binomial, data = BRFSS)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3491  -0.7233   0.1750   0.7379   2.9308

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.143950   0.125630  -17.07   <2e-16 ***
Age             0.066067   0.001135   58.23   <2e-16 ***
StrokeNo       -0.733940   0.055257  -13.28   <2e-16 ***
Gen_HealthGood  0.619059   0.033613   18.42   <2e-16 ***
Gen_HealthFair  1.305688   0.039781   32.82   <2e-16 ***
Gen_HealthPoor  1.764498   0.056466   31.25   <2e-16 ***
SexFemale      -0.471839   0.028081  -16.80   <2e-16 ***
BMI             0.022980   0.002124   10.82   <2e-16 ***
Heart_AttackNo -2.796809   0.049686  -56.29   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 49482  on 35924  degrees of freedom
Residual deviance: 31560  on 35916  degrees of freedom
  (380 observations deleted due to missingness)
AIC: 31578

Number of Fisher Scoring iterations: 5
```

```
exp(coef(m1))
```

```
   (Intercept)            Age       StrokeNo Gen_HealthGood Gen_HealthFair
    0.11719103     1.06829822     0.48001407     1.85717876     3.69022661
Gen_HealthPoor      SexFemale            BMI Heart_AttackNo
    5.83864120     0.62385398     1.02324606     0.06100443
```

Reject the null hypothesis, because the p value for all the explanatory variables are less than 0.05. The p value for the variables is/are 2e-16 . Hence all the explanatory variables are good predators for the heart disease.

Compared to the reference group, Stroke-yes, the odds of a person in the Stroke-No group having heart disease are multiplied by a factor of exp(-0.7550)=0.47. That is, the odds are lower. People who have had a stroke probably have poor health compared to people who haven't had a stroke. Therefore this makes sense.

Compared to the reference group, Sex-Male, the odds of a person in the Sex-Female group having heart disease are multiplied by a factor of exp(-5.019)=0.61. That is, the odds are lower. This does make sense, because male gender do have a higher risk of getting diseases like diabetes and cholesterol, which a factors that can lead to heart disease.

Compared to the reference group, Gen_Health-Excellent, the odds of a person in the Gen_Health-Good group having heart disease are multiplied by a factor of exp(0.619059)=1.85718. That is, the odds are higher. People with excellent general health are less likely to go through a heart disease when compared to the individuals with good general health. This is mainly because they are more likely to have a balanced diet, spend good number of hours working out and so on. Hence the prediction makes sense.

Compared to the reference group, Gen_Health Excellent, the odds of a person in the Gen_Health-Fair group having heart disease are multiplied by a factor of exp(1.305688)=3.690227. That is, the odds are higher. People with excellent general health are less likely to go through a heart disease when compared to the individuals with fair general health. This is mainly because they are more likely to have a balanced diet, spend good number of hours working out and so on. All this daily activities they do will help them overcome from getting other diseases that would lead to heart disease. Hence the prediction makes sense.

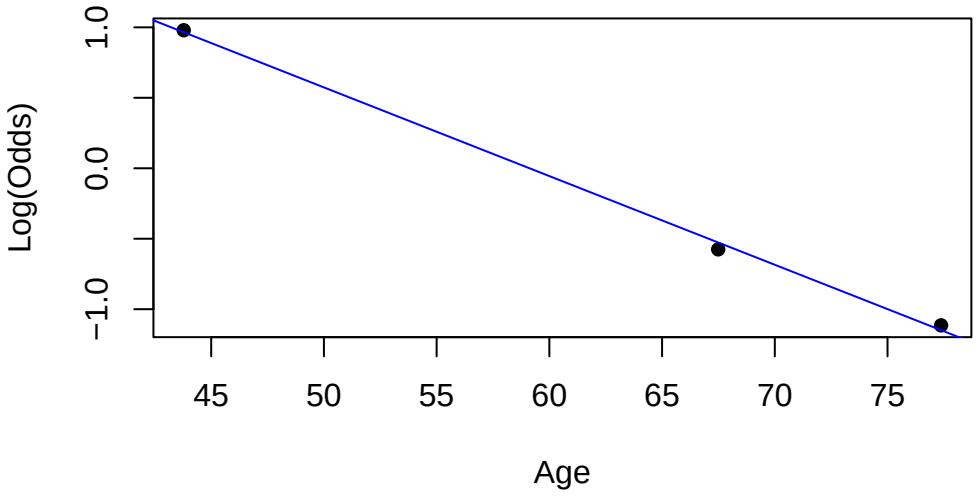
Compared to the reference group, Gen_Health Excellent, the odds of a person in the Gen_Health Poor group having heart disease are multiplied by a factor of exp(1.764498)=5.838641. That is, the odds are higher. People with excellent general health are less likely to go through a heart disease when compared to the individuals with fair general health. This is mainly because they are more likely to have a balanced diet, spend good number of hours working out and so on. All this daily activities they do will help them overcome from getting other diseases that would lead to heart disease. Individuals that are having poor health in general would be either obese or having other health issues. We also expect this group to have the highest possibility of getting the heart disease when compared to other group. Hence the prediction makes sense.

A one unit increase in age is associated with multiplying the odds of having a heart disease by a factor of exp(0.066067)=1.068298. This means that the odd ratio goes higher. This makes sense,because as the individual get older, they become weaker. For example, they will not have enough energy to work out or do other sort of activities that would help them prevent from getting heart disease. Also because individuals tend to get more stress as they get older because they have more responsibilities, like from work and life.

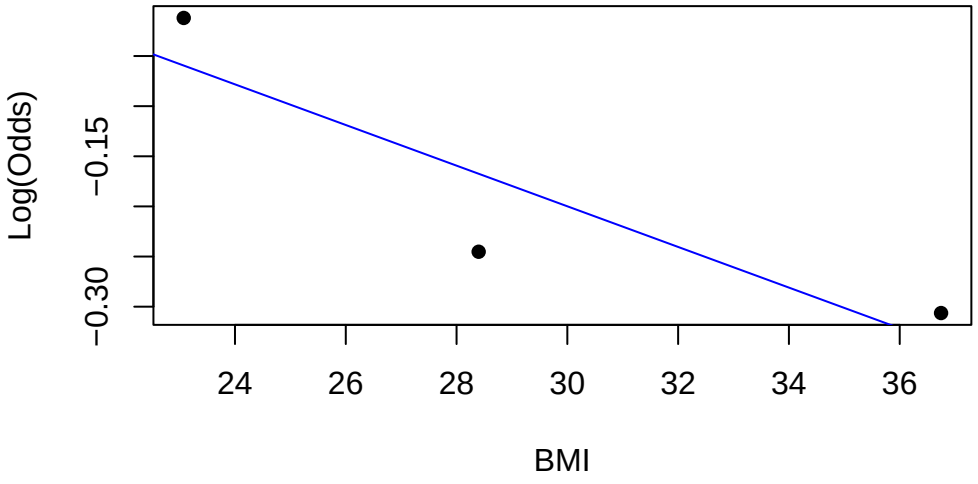
A one unit increase in BMI is associated with multiplying the odds of having a heart disease by a factor of exp(0.022980)=1.023246. This means that the odd ratio goes higher. This makes sense, because BMI is related to the weight and height of the individual. Many individuals in the United States tend to have a higher BMI(obese). Obesity is one of the major issues that lead to heart disease. Since majority of people consume fast food and have unbalanced meals, probability of individuals getting obesity at a young age is higher. Not only obesity even other diseases that can lead to heart diseases in the future. In fact we expected BMI to be much more strongly related to the response variable(heart disease).

Graphs:

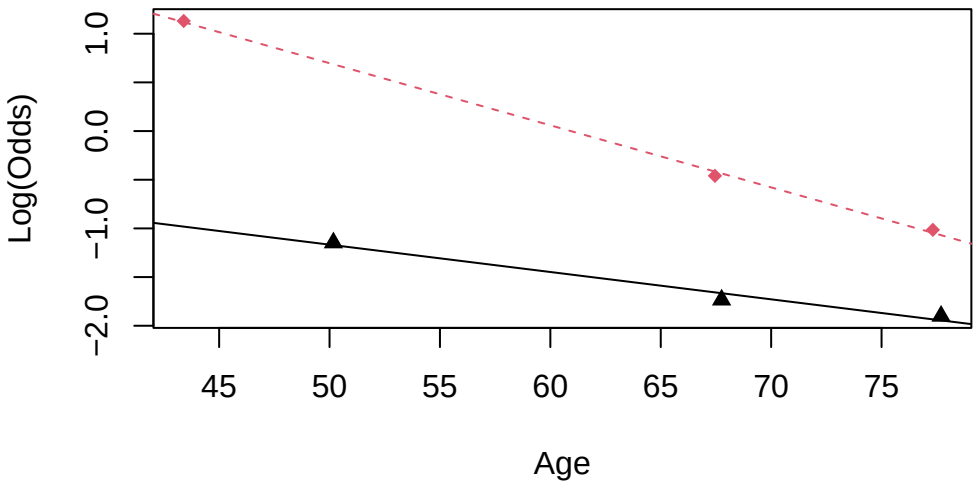
```
emplogitplot1(HRT_Disease~Age,data=BRFSS)
```



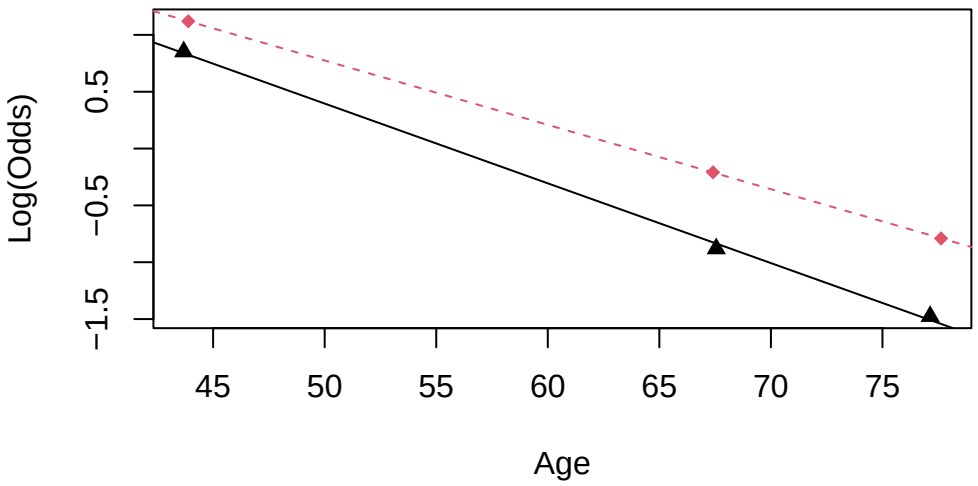

```
emplogitplot1(HRT_Disease~BMI,data=BRFSS)
```



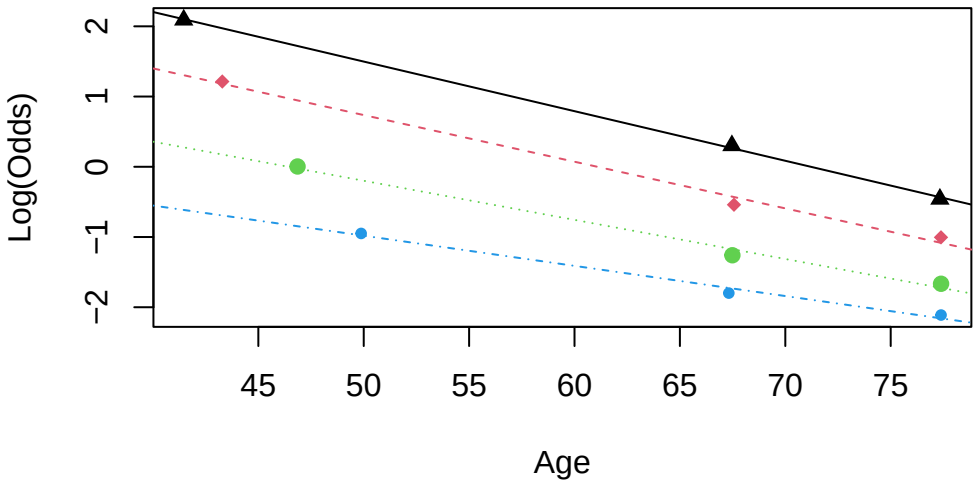
```
emplogitplot2(HRT_Disease~Age+Stroke,data=BRFSS)
```



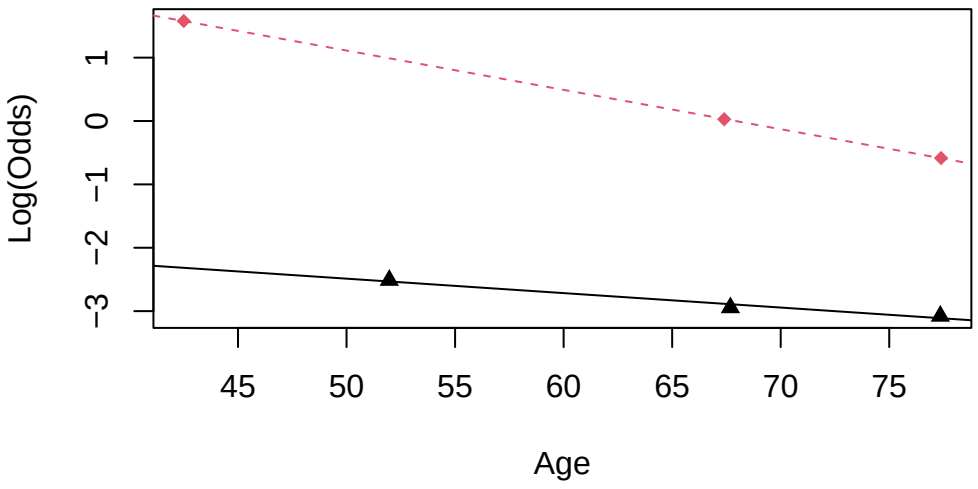

```
emplogitplot2(HRT_Disease~Age+Sex,data=BRFSS)
```




```
emplogitplot2(HRT_Disease~Age+Gen_Health,data=BRFSS)
```



```
emplogitplot2(HRT_Disease~Age+Heart_Attack,data=BRFSS)
```



L-Linear: The graphs tend to be linear. Hence passed the linear test.

I-Independence: The sample is collected randomly through the United States. Hence the individuals are not related to each other. They are totally independent from one another. This also helps us check the condition of multicolinearity. In this case this condition is met. This is mainly because variables are not connected to each other. For example, BMI variable is not related to the Stroke variable, like they don't have a relation, one is not lead by the other one.

R- Randomness: The response variable(heart disease) meets the randomness factor, because of their inheritance. We can not predict which individual will get a heart disease and which one will not. Since this sample size meets both the Independence and randomness factors, the results can be used to represent the larger population. For example, for all the individuals across the world.

Since the conditions are met, we did not have to transform the model to make a new one. Every output of the code looks like the way we would like to see. Hence additional analysis is not required.

Conclusion:

All the variables used to predict heart disease for the individuals living in United States are significant, because they have a p value less than 0.05. Since this data set meets the conditions for linearity, Randomness, and Independence, I would say that the result can be used to predict for the larger population. For example, to predict the heart disease for people around the world. The factors that statisticians can work on this project, would be understanding and predicting why the p values are the same for all the predictors. Secondly, why does the result predict that women are having a less chance to get heart disease when compared to men. For example, is there a scientific reason behind this fact. I believe it would be more interesting a accurate to analysis the data if we could use more variable from the data set, that are randomly chosen, because from the explanatory variable we an predict that this variables have an impact on the heart disease, and the result have proven it. So using variety of variables would be more hard to predict. Selecting the variables we wanted to analysis was somehow difficult due to the fact that there about 50 variables. Making the data set shorter(reducing the number of variables) would be something the statisticians can work on in the future.

Bibliography:

"CDC - 2021 BRFSS Survey Data and Documentation." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 7 Dec. 2022, https://www.cdc.gov/brfss/annual_data/annu

"A Dynamic Visualization Tool of Local Trends in Heart Disease and Stroke Mortality in the United States." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 8 Sept. 2022, https://www.cdc.gov/pcd/issues/2022/22_0076.htm.

"Heart Disease Facts." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 14 Oct. 2022, https://www.cdc.gov/heartdisease/facts.htm.

Thomas, Jen. "Facts and Statistics on Heart Disease." *Healthline*, Healthline Media, 16 July 2020, https://www.healthline.com/health/heart-disease/statistics#Who-is-at-risk?