What tells you that you have a heart disease?*

An analysis of Heart Disease indicators

Puyu Liu (1005092971)

27 April 2022

Abstract

This paper analyzed the data of personal key indicators of heart disease. The result indicates that age, sex, BMI, etc are correlated to having a heart disease. Based on the analysis, this paper consists of a linear regression analysis, with the use of the programming language R. Keywords: Heart Diseases, Physical health, Mental Health, kaggle, age & sex, stroke, Smoking, Alcohol Drinking

1 Introduction

As heart diseases have caused most deaths for people in the US, this dataset include about 20 variables that involve several risk factors of heart disease. Around half of all Americans have at least one heart disease indicator: blood pressure, high cholesterol, smoking, alcohol drinking, diabetes and obesity (high BMI), physical and mental health. Hopefully we can derive a regression analysis to use the logistic regression model to make future predictions for population with heart disease risk factors. Computational developments have a great effect on detecting the early stage of some typical heart diseases.

We used the programming language R (R Core Team 2020).The paper starts with an overview of the data with 300 thousands observations. And also, we will look into the correlation of every variable then try to conduct a regression analysis. Before that, the data should be properly cleaned and outliers and leverage should be well taken care of. After processing the data, we will be using the ANOVA test to determine our variables in our model.

2 Data

In this paper we discuss the key indicators of heart disease from BMI, smoking, and other categorical variables. The data set (CDC survey data of 400k adults related to their health status 2020)that we used is provided by the CDC (*Centers for Diseases Control and Prevention* 2020) whom surveyed by telephone to gather data from residents across the US and I retrieved the organized data set from the Kaggle website. The dataset that we used surveys 400,000 adults whom are asked a series of questions related to their health status, for example "Have you smoked more than 100 cigarettes in your life?" etc. Many variables were documented by the CDC but only relevant variables related to increasing a risk of heart disease were used and cleaned up within our data set. After cleaning, only roughly 320,000 observations were noted and had relevancy towards directly or indirectly influencing heart disease.

The packages we used include knitr (Xie 2021), janitor (Firke 2021), dplyr (Wickham et al. 2021) and tidyverse (Wickham et al. 2019)

*Code and data are available at: https://github.com/liupuyu2/Puyu-Liu.

3 Model

Y indicates if this person have heart disease. It follows Bernoulli Distribution with the probability p_i. pi is the probability of having heart disease. p/1-p is the odds ratio, which is the probability of getting heart disease over the probability of not getting heart disease, and so the log $\frac{p_i}{1-p_i}$ is the log odds ratio. We include 7 independent variables in our model. X1 is BMI, which is the Body Mass Index. It is calculated by dividing a person's weight in kg by his height in meters. X2 represents whether a person smokes, which is obtained by surveys. X3 stands for whether a person drink alcohol. X4 is Mental health, research found that it is closely related to several kinds of heart diseases. Sleep time is also included in this model. Age category is a dummy variable which consists of 12 different adult age categories.

$$Y_i \sim Bernuolli(p_i)$$

$$log(\frac{p_i}{1-p_i}) = \beta_0 + X_1\beta_1 + X_1\beta_1 + X_1\beta_1 + X_1\beta_1 + X_1\beta_1 + X_1\beta_1 + X_1\beta_1$$

4 Results

```
estimate=fit$coefficients
p = summary(fit)$coefficients[,4]
ta=cbind(estimate,p)
colnames(ta) = c("Estimate","p-value")
knitr::kable(ta,digits = 5)
```

	Estimate	p-value
(Intercept)	-5.98809	0.00000
BMI	0.03243	0.00000
SmokingYes	0.57084	0.00000
AlcoholDrinkingYes	-0.37447	0.00000
PhysicalHealth	0.03660	0.00000
MentalHealth	0.01149	0.00000
AgeCategory25-29	0.06893	0.57822
AgeCategory30-34	0.41169	0.00020
AgeCategory35-39	0.54394	0.00000
AgeCategory40-44	0.99139	0.00000
AgeCategory45-49	1.37618	0.00000
AgeCategory50-54	1.85191	0.00000
AgeCategory55-59	2.16650	0.00000
AgeCategory60-64	2.48385	0.00000
AgeCategory65-69	2.76151	0.00000
AgeCategory70-74	3.08740	0.00000
AgeCategory75-79	3.33027	0.00000
AgeCategory80 or older	3.64533	0.00000
SleepTime	-0.03055	0.00000

From the table above, as age increases, we see that the coefficients for these variables becomes higher and this would mean that the people in a higher age group will be most at risk to heart disease. The significant factor (P-value) is low in younger age groups and suddenly increase in older age groups which tells us that while people are getting older, aging becomes more related to heart diseases.

5 Discussion

5.1 First discussion point

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Appendix

```
fit
```

```
##
## Call:  glm(formula = y ~ BMI + Smoking + AlcoholDrinking + PhysicalHealth +
##     MentalHealth + AgeCategory + SleepTime, family = binomial,
##     data = dat)
##
## Coefficients:
##            (Intercept)                     BMI             SmokingYes
##                -5.98809                 0.03243                 0.57084
##     AlcoholDrinkingYes          PhysicalHealth            MentalHealth
##                -0.37447                 0.03660                 0.01149
##       AgeCategory25-29        AgeCategory30-34        AgeCategory35-39
##                 0.06893                 0.41169                 0.54394
##       AgeCategory40-44        AgeCategory45-49        AgeCategory50-54
##                 0.99139                 1.37618                 1.85191
##       AgeCategory55-59        AgeCategory60-64        AgeCategory65-69
##                 2.16650                 2.48385                 2.76151
##       AgeCategory70-74        AgeCategory75-79  AgeCategory80 or older
##                 3.08740                 3.33027                 3.64533
##              SleepTime
##                -0.03055
##
## Degrees of Freedom: 319794 Total (i.e. Null);  319776 Residual
## Null Deviance:       186900
## Residual Deviance: 158300    AIC: 158400
```

```
exp(fit$coefficients)
```

```
##            (Intercept)                     BMI             SmokingYes
##              0.00250845              1.03296055              1.76975126
##     AlcoholDrinkingYes          PhysicalHealth            MentalHealth
##              0.68765356              1.03728107              1.01155258
##       AgeCategory25-29        AgeCategory30-34        AgeCategory35-39
##              1.07136516              1.50936430              1.72277781
##       AgeCategory40-44        AgeCategory45-49        AgeCategory50-54
##              2.69496772              3.95975429              6.37198998
##       AgeCategory55-59        AgeCategory60-64        AgeCategory65-69
##              8.72771914             11.98736404             15.82366364
##       AgeCategory70-74        AgeCategory75-79  AgeCategory80 or older
##             21.91994435             27.94588138             38.29559387
##               SleepTime
##              0.96991103
```

```
summary(fit)
```

```
##
## Call:
## glm(formula = y ~ BMI + Smoking + AlcoholDrinking + PhysicalHealth +
##     MentalHealth + AgeCategory + SleepTime, family = binomial,
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -1.7708  -0.4705  -0.2940  -0.1437   3.4219
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -5.9880903  0.0977014 -61.290  < 2e-16 ***
## BMI                     0.0324290  0.0010164  31.905  < 2e-16 ***
## SmokingYes              0.5708390  0.0135875  42.012  < 2e-16 ***
## AlcoholDrinkingYes     -0.3744701  0.0325093 -11.519  < 2e-16 ***
## PhysicalHealth          0.0366029  0.0006530  56.055  < 2e-16 ***
## MentalHealth            0.0114864  0.0008261  13.904  < 2e-16 ***
## AgeCategory25-29        0.0689337  0.1239856   0.556 0.578224
## AgeCategory30-34        0.4116886  0.1108434   3.714 0.000204 ***
## AgeCategory35-39        0.5439380  0.1060324   5.130 2.90e-07 ***
## AgeCategory40-44        0.9913862  0.0996432   9.949  < 2e-16 ***
## AgeCategory45-49        1.3761820  0.0960052  14.334  < 2e-16 ***
## AgeCategory50-54        1.8519118  0.0926413  19.990  < 2e-16 ***
## AgeCategory55-59        2.1665041  0.0911290  23.774  < 2e-16 ***
## AgeCategory60-64        2.4838531  0.0902589  27.519  < 2e-16 ***
## AgeCategory65-69        2.7615065  0.0899289  30.708  < 2e-16 ***
## AgeCategory70-74        3.0873969  0.0897579  34.397  < 2e-16 ***
## AgeCategory75-79        3.3302698  0.0901367  36.947  < 2e-16 ***
## AgeCategory80 or older  3.6453348  0.0897387  40.622  < 2e-16 ***
## SleepTime              -0.0305509  0.0043188  -7.074 1.51e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 186906  on 319794  degrees of freedom
## Residual deviance: 158329  on 319776  degrees of freedom
## AIC: 158367
##
## Number of Fisher Scoring iterations: 7
```

```r
set.seed(3080)

prob <- fit %>% predict(dat, type = "response")
head(prob)
```

```
##          1          2          3          4          5          6
## 0.08235427 0.13044769 0.27656995 0.11344550 0.03084835 0.21454395
```

References

CDC survey data of 400k adults related to their health status, 2020 annual. 2020. *Personal Key Indicators of Heart Disease*. https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease.

Centers for Diseases Control and Prevention. 2020. https://www.cdc.gov/heartdisease/risk_factors.html.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.