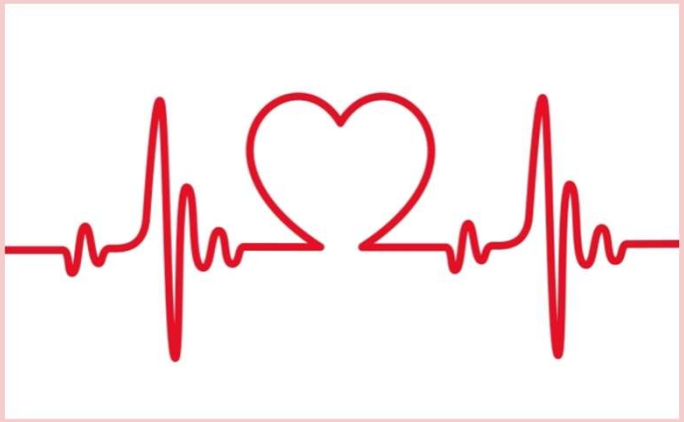
Personal Key Indicators Of Heart Disease



Ethan Huang

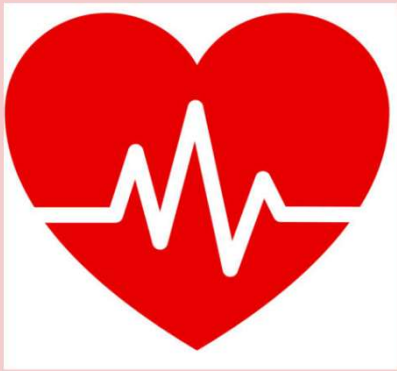
STAT 495

Background

Each year, the CDC conducts telephone surveys asking adults about their health status through the Behavioral Risk Factor Surveillance System. It contains questions about demographics, underlying conditions, and habits. Over 400,000 adults are interviewed each year.

The Dataset

The data, found through Kaggle, is a compilation of the 319,796 surveys administered in 2020. It highlights the 20 top variables influencing a person's presence or absence of heart disease. The explanatory variables include BMI, Smoking, Alcohol, Stroke, and many more variables.

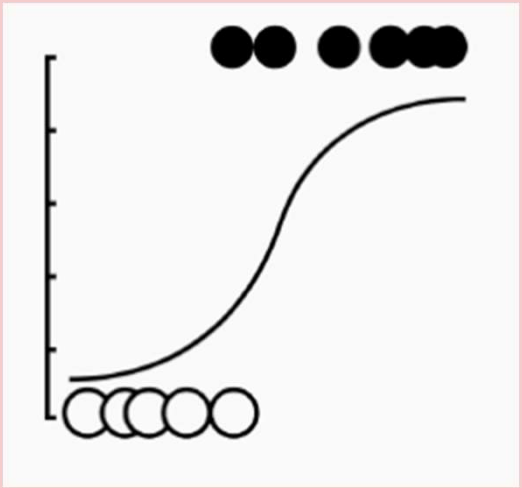


Changes to Dataset

- Reduced explanatory variables from 19 to 11
- Converted Age from categorical to numeric
  - AgeCategory 30-34 → 32
- Condensed Diabetes to “Yes” and “No”
  - Gestational Diabetes → “Yes”
  - Borderline Diabetes → “No”

Statistical Method: Binary Logistic Model

We use a binary logistic model to regress all the explanatory variables on Heart Disease. Heart Disease is a binary response variable containing only two values: Yes and No.



SAS Code

```
proc import out=heartdisease datafile="C:\Users\colle\Downloads\archive (6)\heart_2020_modified.csv"
dbms=csv replace;

proc genmod data=heartdisease;
class Smoking (ref="No") AlcoholDrinking (ref="Ye") Sex (ref= "Female") Race (ref= "Asian") Diabetic (ref= "No")
Stroke (ref="No") PhysicalActivity  (ref= "Yes") Asthma (ref = "No");
model HeartDisease(event="Yes")= Smoking AlcoholDrinking Sex Age Race Diabetic Stroke PhysicalActivity Asthma BMI SleepTime/
            dist=binomial link=logit;
run;

data prediction;
input Smoking $ AlcoholDrinking $ Sex $ Age Race $ Diabetic $ Stroke $ PhysicalActivity $ Asthma $ BMI SleepTime;
cards;
Yes Ye Male 80 Nativ Yes Yes No Yes 35 3
;

data heartdisease2;
set heartdisease prediction;
run;

proc genmod data=heartdisease2;
class Smoking AlcoholDrinking Sex Race Diabetic Stroke PhysicalActivity Asthma;
model HeartDisease(event="Yes")= Smoking AlcoholDrinking Sex Age Race Diabetic Stroke PhysicalActivity Asthma BMI SleepTime/
            dist=binomial link=logit;
output out=outdata p=pred_probdisease;
run;

proc print data=outdata(firstobs=319796)noobs;
var pred_probdisease;
run;
```

R Code

```
heartdisease <- read.csv("C:\\Users\\colle\\Downloads\\archive (6)\\heart_2020_modified.csv")

HeartDisease.rel <- relevel(as.factor(heartdisease$HeartDisease),ref="No")
AlcoholDrinking.rel <- relevel(as.factor(heartdisease$AlcoholDrinking),ref="Yes")
Race.rel <- relevel(as.factor(heartdisease$Race),ref="Asian")
PhysicalActivity.rel <- relevel(as.factor(heartdisease$PhysicalActivity),ref="Yes")

summary(fitted.model<- glm(HeartDisease.rel ~ BMI + Smoking + AlcoholDrinking.rel + Stroke +
 Sex + Age + Race.rel + Diabetic + PhysicalActivity.rel +
 SleepTime + Asthma ,data=heartdisease, family=binomial(link=logit)))

print(predict(fitted.model,type="response",data.frame(Race.rel="NativeAmerican",BMI=35,Smoking="Yes",
AlcoholDrinking.rel="Yes",Stroke="Yes",Sex="Male",Age=80,Diabetic="Yes",
PhysicalActivity.rel="No",SleepTime=3,Asthma="Yes")))
```

- We set references for certain categorical variables
- Fit the model. Family=Binomial, Link=Logit
- Use the predict function to find the predicted value

SAS Output

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-8.0927	0.0918	-8.2727	-7.9128	7771.71	<.0001
Smoking	Yes	1	0.4920	0.0139	0.4648	0.5192	1254.08	<.0001
Smoking	No	0	0.0000	0.0000	0.0000	0.0000	.	.
AlcoholDrinking	No	1	0.2841	0.0329	0.2195	0.3486	74.46	<.0001
AlcoholDrinking	Ye	0	0.0000	0.0000	0.0000	0.0000	.	.
Sex	Male	1	0.6639	0.0140	0.6365	0.6914	2242.59	<.0001
Sex	Female	0	0.0000	0.0000	0.0000	0.0000	.	.
Age		1	0.0608	0.0006	0.0597	0.0620	10569.5	<.0001
Race	Black	1	0.3164	0.0714	0.1765	0.4563	19.65	<.0001
Race	Hispa	1	0.4298	0.0720	0.2886	0.5710	35.59	<.0001
Race	Nativ	1	0.6885	0.0827	0.5263	0.8507	69.24	<.0001
Race	Other	1	0.5754	0.0762	0.4260	0.7248	57.00	<.0001
Race	White	1	0.4875	0.0665	0.3572	0.6177	53.80	<.0001
Race	Asian	0	0.0000	0.0000	0.0000	0.0000	.	.
Diabetic	Yes	1	0.7407	0.0158	0.7098	0.7716	2205.51	<.0001
Diabetic	No	0	0.0000	0.0000	0.0000	0.0000	.	.
Stroke	Yes	1	1.3448	0.0219	1.3019	1.3877	3781.18	<.0001
Stroke	No	0	0.0000	0.0000	0.0000	0.0000	.	.
PhysicalActivity	No	1	0.3320	0.0149	0.3028	0.3612	496.87	<.0001
PhysicalActivity	Yes	0	0.0000	0.0000	0.0000	0.0000	.	.
Asthma	Yes	1	0.4879	0.0186	0.4515	0.5244	688.82	<.0001
Asthma	No	0	0.0000	0.0000	0.0000	0.0000	.	.
BMI		1	0.0208	0.0011	0.0186	0.0229	350.38	<.0001
SleepTime		1	-0.0531	0.0045	-0.0619	-0.0444	141.18	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

R Output

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -8.0927301  0.0917920 -88.164  < 2e-16 ***
BMI                      0.0207536  0.0011087  18.719  < 2e-16 ***
SmokingYes               0.4919958  0.0138928  35.414  < 2e-16 ***
AlcoholDrinking.relNo    0.2840615  0.0329195   8.629  < 2e-16 ***
StrokeYes                1.3447983  0.0218695  61.492  < 2e-16 ***
SexMale                  0.6639444  0.0140200  47.357  < 2e-16 ***
Age                      0.0608215  0.0005916 102.814  < 2e-16 ***
Race.relBlack            0.3163944  0.0713646   4.433 9.27e-06 ***
Race.relHispanic         0.4298358  0.0720419   5.966 2.42e-09 ***
Race.relNativeAmerican   0.6885143  0.0827403   8.321  < 2e-16 ***
Race.relOther            0.5754208  0.0762118   7.550 4.34e-14 ***
Race.relWhite            0.4874692  0.0664553   7.335 2.21e-13 ***
DiabeticYes              0.7406995  0.0157719  46.963  < 2e-16 ***
PhysicalActivity.relNo   0.3319632  0.0148922  22.291  < 2e-16 ***
SleepTime               -0.0531171  0.0044703 -11.882  < 2e-16 ***
AsthmaYes                0.4879486  0.0185914  26.246  < 2e-16 ***
---
```

Analysis

- All Predictors are significant
- Smoking: exp(0.4920)*100%=163.558%
- SleepTime:(exp(-0.0531)-1)*100=-5.172%
- Interpretation of Smoking: Smokers are 163.558% more likely than non-smokers to develop heart disease
- Interpretation of SleepTime. For one extra hour of sleep, the estimated odds of heart disease decrease by 5.172%

Predicted Probability for High Risk of Heart Disease

- Gender=Male
- Age=80
- Race= Native American
- Smoker, Drinker
- No Physical Activity
- Has Asthma, Diabetes, Stroke
- BMI=35
- SleepTime= 3 Hours

pred_probdisease
0.88992

1
0.8899156

Thanks to Dr. Olga and STAT 495 Classmates