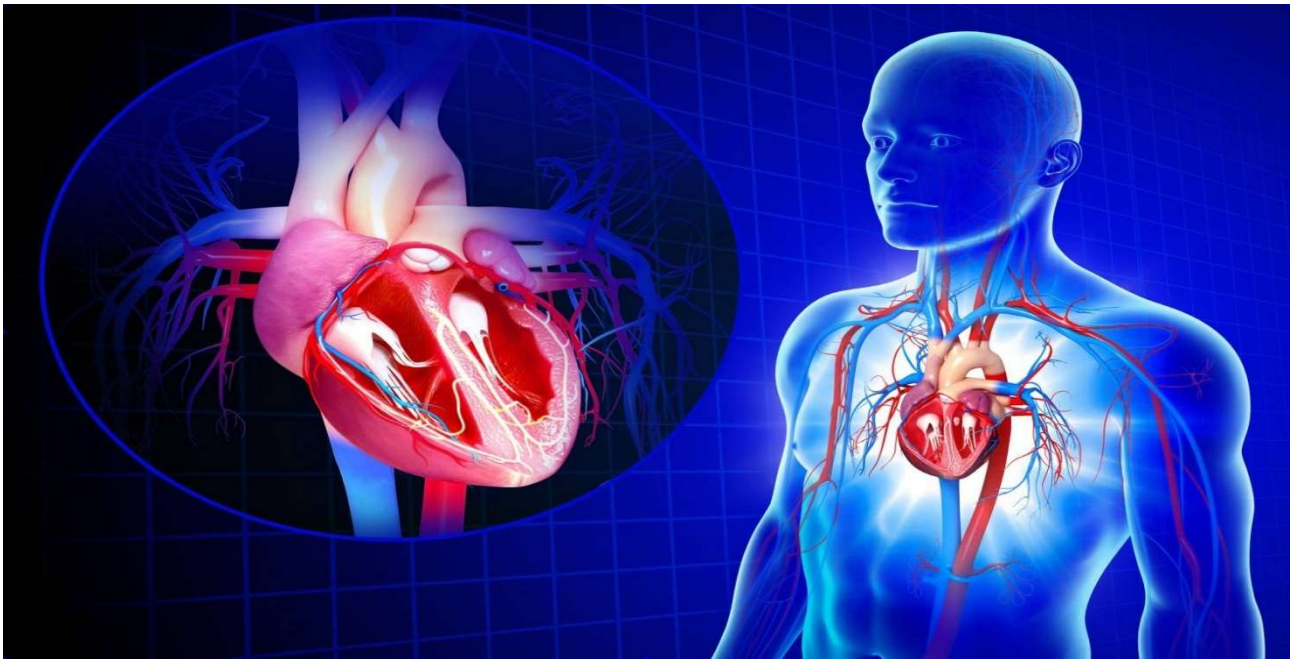

STAT 410 Project Report

Topic:

Cardiovascular Risk Prediction

Modeling method:

Binary Logistic & Probit & Complementary Log-Log Regression Model

Reported by:

Bushra Abukarn

Submitted to:

Dr. Olga Korosteleva

Date:

11/27/2023

Introduction

Sadly, each year we lose around 17.9 million lives to cardiovascular disease as reported by the World Health Organization [1], highlighting a significant health challenge. One way to reduce the likelihood of developing this disease is to enhance awareness of its risks and contributing factors. Utilizing the available routine clinical data to build a regression analysis model, aiming to predict cardiovascular disease risk. According to the National Library of Medicine, machine learning significantly improves the accuracy of predicting cardiovascular risk [2]. The goal is to spread information about the risks associated with cardiovascular diseases to individuals and subgroups and take a proactive stance.

Background

Cardiovascular disease (CVD) is a general term that describes a disease of the heart or blood vessels [3]. According to the World Health Organization Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels. Also, heart attacks and strokes are usually acute events and are mainly caused by a blockage that prevents blood from flowing to the heart or brain.

Data description

The Dataset Search through Google helped me to find the data that I was looking for from Kaggle website [4]. The dataset contains 17 variables and 3087 rows, form the 17 variables I used 7 as predictors and 1 as a response. The response variable is Heart Disease (Yes/N0) and the predictors are Exercise (Yes/No), Depression (Yes, No), Sex (Male, Female), Age (18-24:80+), Smoking History (Yes/No), BMI, and Alcohol Consumption.

Results

The probit model has smaller values in all three criteria AIC, BIC, & AICc. Thus, the probit model has the best fitted model.

The probit fitted model is: $\Phi^{-1}(\pi)$ = -2.871117 -.267145 Exercise + .3113835 Depression + .412838 Male + .258597 Smoking - .009876 Alcohol – 2.72266 (25-29 years old) + .115133 (30-34 years old) - .106887 (35-39 years old) + .242569 (40-44 years old) + .406810 (45-49 years old) + .553118 (50-54 years old) + .773307 (55-59 years old) + 1.029560 (60-64 years old) + .990837 (65-69 years old) + 1.350416 (70-74 years old) + 1.323566 (75-79 years old) + 1.572687 (80+ years old) + .014392 BMI.

Since we have a good amount of data, most of the predictors are significant. The significant predictors at 5% level are: Sex. Male, Exercise, Depression, Smoking, Age, and BMI.

Interpretation of Estimated Regression Coefficient: Males have a 0.412838 higher z-score of estimated probability of developing cardiovascular disease than females. Engaging in regular exercise decreases the z-score by 0.267145. Conversely, depression raises the z- score by 0.311835, and smoking increases the z-score by 0.258597. Besides, young people have a lower z-score than the elderly. Additionally, for each unit increase in BMI, the estimated z-score rises by 0.014392.

The predicted response probability of developing a cardiovascular disease to a 55-year-old man who is happy, exercises regularly and does not smoke, but drink rarely is .05479635, around 5.5%.

Conclusion

Certainly, early detection of diseases, especially cardiovascular diseases, can significantly mitigate potential consequences and better manage situations. The ongoing development of tools for disease detection has a great role in providing optimal assistance to everyone. Moreover, this project gave me the opportunity to apply the techniques that I learned during this semester, which is regression analysis, to analyze the dataset. These techniques enhance our ability to understand and address health-related challenges effectively.

I. SAS Code & Output.

```
proc import out=CVD
 datafile="C:/Users/bushr/OneDrive/Desktop/CVD_reduced.csv"
 dbms=csv
 replace;
 getnames=yes;
 run;
proc genmod;
    class Exercise(ref="No") Depression(ref="No") Smoking_history(ref="No") Sex(ref="Female") Age_Category(ref="18-24");
    model Heart_Disease(event="Yes") = Exercise Depression Smoking_history Sex Alcohol_Consumption Age_CAtegory BMI / dist=binomial link=probit;
run;
```

Full Log Likelihood		-744.6764	
AIC (smaller is better)		1527.3528	
AICC (smaller is better)		1527.6007	
BIC (smaller is better)		1642.0108	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.8711	0.3241	-3.5063	-2.2359	78.49	<.0001
Exercise	Yes	1	-0.2671	0.0790	-0.4219	-0.1123	11.44	0.0007
Exercise	No	0	0.0000	0.0000	0.0000	0.0000	.	.
Depression	Yes	1	0.3118	0.0921	0.1313	0.4924	11.46	0.0007
Depression	No	0	0.0000	0.0000	0.0000	0.0000	.	.
Smoking_History	Yes	1	0.2586	0.0730	0.1156	0.4016	12.55	0.0004
Smoking_History	No	0	0.0000	0.0000	0.0000	0.0000	.	.
Sex	Male	1	0.4128	0.0760	0.2639	0.5618	29.52	<.0001
Sex	Female	0	0.0000	0.0000	0.0000	0.0000	.	.
Alcohol_Consumption		1	-0.0099	0.0047	-0.0190	-0.0008	4.50	0.0338
Age_Category	25-29	1	-0.2723	0.4633	-1.1803	0.6357	0.35	0.5567
Age_Category	30-34	1	0.1151	0.3565	-0.5835	0.8138	0.10	0.7467
Age_Category	35-39	1	-0.1069	0.3812	-0.8540	0.6402	0.08	0.7792
Age_Category	40-44	1	0.2426	0.3286	-0.4016	0.8867	0.54	0.4605
Age_Category	45-49	1	0.4068	0.3191	-0.2186	1.0322	1.63	0.2023
Age_Category	50-54	1	0.5531	0.3068	-0.0483	1.1545	3.25	0.0714
Age_Category	55-59	1	0.7733	0.2906	0.2037	1.3429	7.08	0.0078
Age_Category	60-64	1	1.0296	0.2817	0.4774	1.5817	13.36	0.0003
Age_Category	65-69	1	0.9908	0.2851	0.4321	1.5496	12.08	0.0005
Age_Category	70-74	1	1.3504	0.2796	0.8024	1.8984	23.33	<.0001
Age_Category	75-79	1	1.3236	0.2895	0.7562	1.8910	20.90	<.0001
Age_Category	80+	1	1.5727	0.2849	1.0142	2.1312	30.46	<.0001
Age_Category	18-24	0	0.0000	0.0000	0.0000	0.0000	.	.
BMI		1	0.0144	0.0057	0.0031	0.0257	6.27	0.0123
Scale		0	1.0000	0.0000	1.0000	1.0000		

```
proc genmod;
 model Heart_Disease = / dist=binomial link=probit;
run;
```

Log Likelihood	-879.9704

```
data deviance_test;
 deviance = -2*(-879.9704 - (-744.6764));
 pvalue = 1 - probchi(deviance,17);
run;
proc print noobs;
run;
```

deviance	pvalue
270.588	0

II. R Code & Output:

```{r}
CVD1.data<- read.csv(file="C:/Users/bushr/OneDrive/Desktop/CVD_reduced.csv",header=TRUE, sep=",",
stringsAsFactors = TRUE)
Exercise.rel<- relevel(CVD1.data$Exercise, ref = "No")
Depression.rel<- relevel(CVD1.data$Depression, ref = "No")
Sex.rel<- relevel(CVD1.data$Sex, ref = "Female")
Smoking.rel<- relevel(CVD1.data$Smoking, ref = "No")

fitted.model<- glm(Heart_Disease ~ Exercise.rel + Depression.rel + Sex.rel + Smoking.rel +
Alcohol_Consumption + Age_Category + BMI,
                   data = CVD1.data, family = binomial(link = "probit"))
summary(fitted.model)
```

```{r}
n<- 3087
p<- 18
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))
```

```{r}
BIC(fitted.model)
```

```{r}
null.model<- glm(Heart_Disease ~ 1, data=CVD1.data, family=binomial(link=probit))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

```{r}
print(p.value<- pchisq(deviance,17,lower.tail = FALSE))
```

```
Call:
glm(formula = Heart_Disease ~ Exercise.rel + Depression.rel +
    Sex.rel + Smoking.rel + Alcohol_Consumption + Age_Category +
    BMI, family = binomial(link = "probit"), data = CVD1.data)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.871117   0.325727  -8.814  < 2e-16 ***
Exercise.relYes     -0.267145   0.079193  -3.373 0.000743 ***
Depression.relYes    0.311835   0.091841   3.395 0.000685 ***
Sex.relMale          0.412838   0.075768   5.449 5.07e-08 ***
Smoking.relYes       0.258597   0.072908   3.547 0.000390 ***
Alcohol_Consumption -0.009876   0.004640  -2.128 0.033319 *
Age_Category25-29   -0.272266   0.463815  -0.587 0.557194
Age_Category30-34    0.115133   0.355748   0.324 0.746214
Age_Category35-39   -0.106887   0.380769  -0.281 0.778930
Age_Category40-44    0.242569   0.328008   0.740 0.459590
Age_Category45-49    0.406810   0.318707   1.276 0.201800
Age_Category50-54    0.553118   0.305648   1.810 0.070349 .
Age_Category55-59    0.773307   0.289444   2.672 0.007547 **
Age_Category60-64    1.029560   0.280829   3.666 0.000246 ***
Age_Category65-69    0.990837   0.284435   3.484 0.000495 ***
Age_Category70-74    1.350416   0.278891   4.842 1.28e-06 ***
Age_Category75-79    1.323566   0.289077   4.579 4.68e-06 ***
Age_Category80+      1.572687   0.284848   5.521 3.37e-08 ***
BMI                  0.014392   0.005678   2.535 0.011248 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1759.9  on 3085  degrees of freedom
Residual deviance: 1489.4  on 3067  degrees of freedom
AIC: 1527.4
```

```
'log Lik.' 1525.576 (df=19)
```

```
[1] 1642.011
```

```
'log Lik.' 270.5881 (df=1)    'log Lik.' 2.203557e-46 (df=1)
```

Fitted model for a prediction response in R:

```r
print(predict(fitted.model, data.frame(Exercise.rel="Yes", Depression.rel="No", Sex.rel="Male",
Smoking.rel="No", Age_Category="55-59", Alcohol_Consumption=1, BMI=25.15), type="response"))
```

```
        1
0.05479635
```

References

[1] World Health Organization. “Cardiovascular Diseases.” World Health Organization, 2022, www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.

[2] Weng, Stephen F., et al. “Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data?” *PLOS ONE*, vol. 12, no. 4, 4 Apr. 2017, p. e0174944, journals.plos.org/plosone/article?id=10.1371/journal.pone.0174944, https://doi.org/10.1371/journal.pone.0174944. Accessed 17 Nov. 2019.

[3] World Health Organization. “Cardiovascular Diseases.” *World Health Organization*, World Health Organization, 11 June 2021, www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[4] “Cardiovascular Diseases Risk Prediction Dataset.” Www.kaggle.com, www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset.