Statistical Inference Practice in Heart Attack Risk

Your Name Here

2023-10-27

Abstract

The risk of heart disease is a growing concern for cardiovascular health, which claims millions of lives around the world each year. Therefore, how to control the threat of this disease to human health has become a very important issue. Doctors are focused on how to prevent heart attacks in patients at a time when people are curious about a healthier lifestyle for the heart. Therefore, this data analysis report conducted Multilevel regression based on a multiple Logistic regression model to explore the effects of various clinical indicators and lifestyle habits on the risk of heart disease. According to the analysis results, some life suggestions are put forward for people. However, this model is not accurate enough, due to interpretability concerns, and machine learning and deep learning methods are not presented in this report. In the future, we will explore the machine learning model with high interpretability and better fitting prediction effect for further research.

Keywords:

Multilevel Regression Model, Heart Attack Risk, Health Information.

Introduction

According to WHO's 2019 Global Health Estimates Report, heart disease remains the number one killer of deaths from disease and injury in the early 2000s. Every year, millions of people around the world are killed by heart disease, which is a growing threat to people's health and security. Therefore, it is very important to study the risk of heart disease and people's clinical indicators and lifestyle habits. Clinicians were curious: How could they tell if a patient was at risk for cardiovascular disease based on cholesterol, Triglycerides and blood pressure? People are also concerned: What kind of routine can protect them from heart disease? Therefore, this data analysis report uses multiple Logistic regression models to explore these issues, hoping to make some contributions to reducing the incidence and mortality of heart disease.

Data

Data Collection Process

This is a clinical diagnostic data on heart disease risk collected from 20 countries in the northern and southern hemispheres and continents, including Argentina, Australia, Brazil, Canada, China, Colombia, France, Germany, India, Italy, Japan, New Zealand, Nigeria, South Africa, South Korea, Spain, Thailand, UK, Vietnam and USA. This data set comes from kaggle and could be downloaded in website(https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset/data). The data set includes information on more than 8,000 patients, including their medical tests and lifestyle habits. Multiple Logistic regression models can be used to determine the relationship between heart disease risk and patients' medical indicators and lifestyle habits.

Data Summary

```
data <- read.csv("heart_attack_prediction_dataset.csv")
```

Do the data wrangling firstly. After observation, it can be found that there are many character data in the data set, which needs to be modified column by column. For example, in the gender column, if male is 1 and female is 0, the meaning of this variable in the regression will be expressed as the chance of heart disease caused by gender differences.

```
data$Sex <- ifelse(data$Sex == "Male", 1, 0)
```

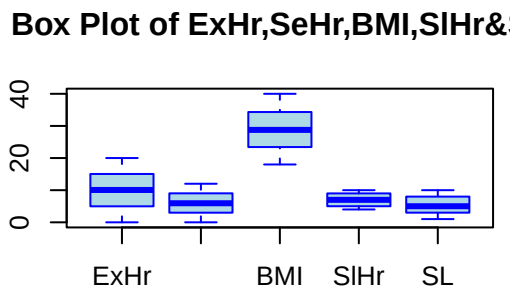
Similarly, modify other columns into a form that could be easily analysed.

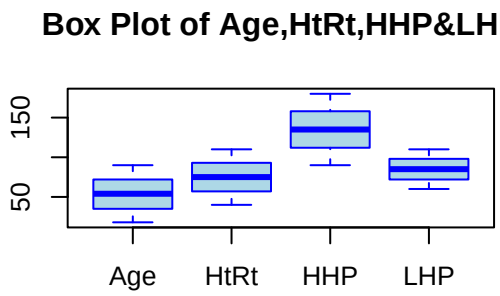
```
data$Blood.Pressure <- as.character(data$Blood.Pressure)
split_values  <- strsplit(data$Blood.Pressure, "/")
data$High.Blood.Pressure <- as.numeric(sapply(split_values, function(x) as.numeric(x[1])))
data$Low.Blood.Pressure <- as.numeric(sapply(split_values, function(x) as.numeric(x[2])))
data <- data[, -which(names(data) == "Blood.Pressure")]

data$Unhealthy.Diet <- ifelse(data$Diet == "Unhealthy", 1, 0)
data$Healthy.Diet <- ifelse(data$Diet == "Healthy", 1, 0)
data <- data[, -which(names(data) == "Diet")]
```

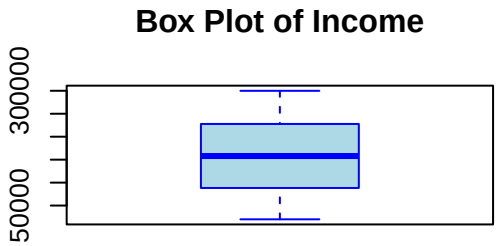
```
y <- data$Heart.Attack.Risk
```

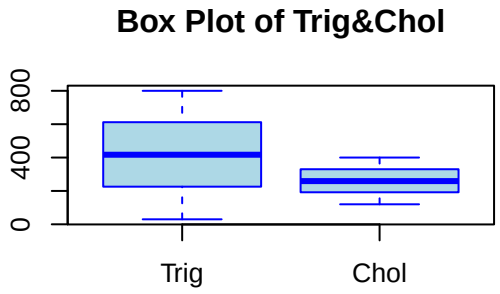
Use box plot to present the data distribution. From these figures, it is interesting that the data distributions all look like normal distribution, which is a good news for linear models application.

```
par(mfrow=c(2, 2))
boxplot(data.frame(ExHr = data$Exercise.Hours.Per.Week,
                   SeHr = data$Sedentary.Hours.Per.Day,
                   BMI = data$BMI,
                   SlHr = data$Sleep.Hours.Per.Day,
                   SL = data$Stress.Level
                   ),
        main="Box Plot of ExHr,SeHr,BMI,SlHr&SL", col="lightblue", border="blue", notch=FALSE, horizonta
boxplot(data.frame(Age = data$Age,
                   HtRt = data$Heart.Rate,
                   HHP = data$High.Blood.Pressure,
                   LHP = data$Low.Blood.Pressure
                   ),
        main="Box Plot of Age,HtRt,HHP&LHP", col="lightblue", border="blue", notch=FALSE, horizontal=FA
boxplot(data.frame(Income = data$Income),
        main="Box Plot of Income", col="lightblue", border="blue", notch=FALSE, horizontal=FALSE)
boxplot(data.frame(Trig = data$Triglycerides, Chol = data$Cholesterol),
        main="Box Plot of Trig&Chol", col="lightblue", border="blue", notch=FALSE, horizontal=FALSE)
```

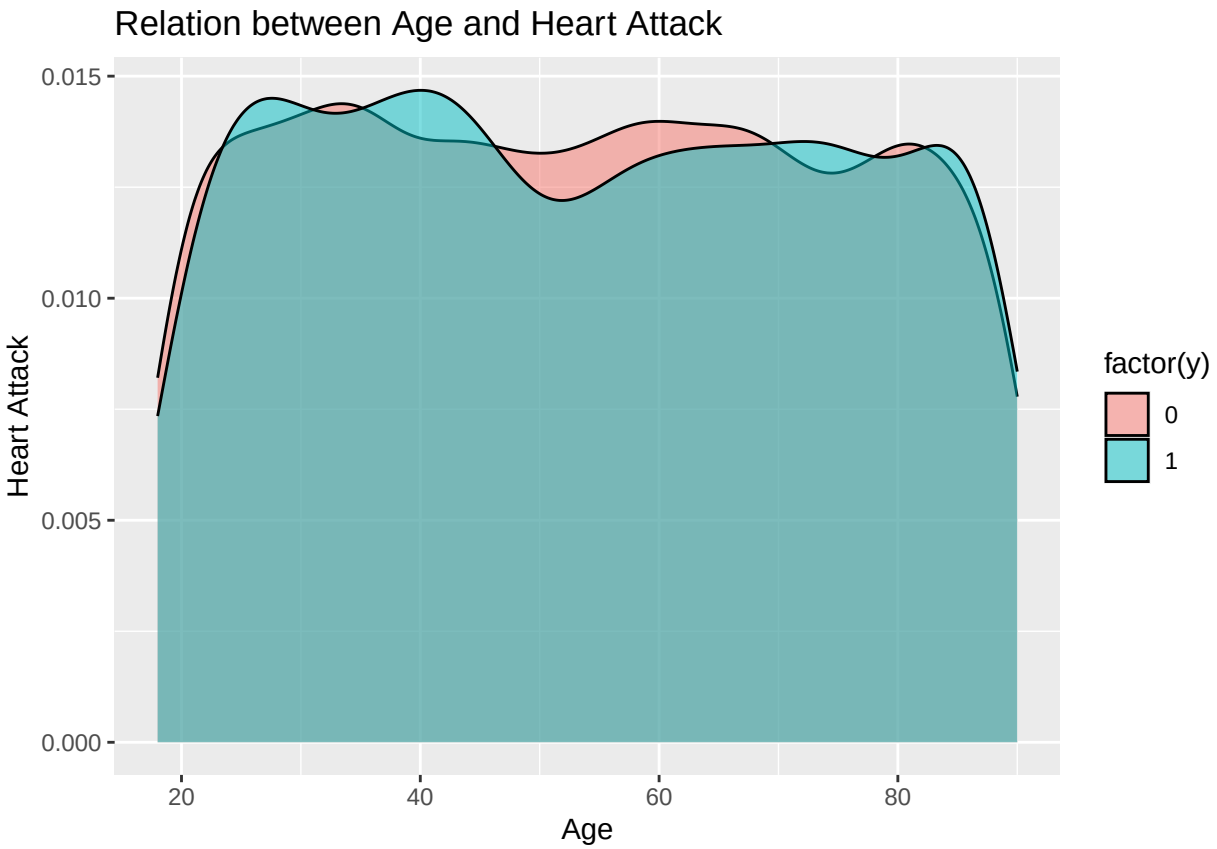






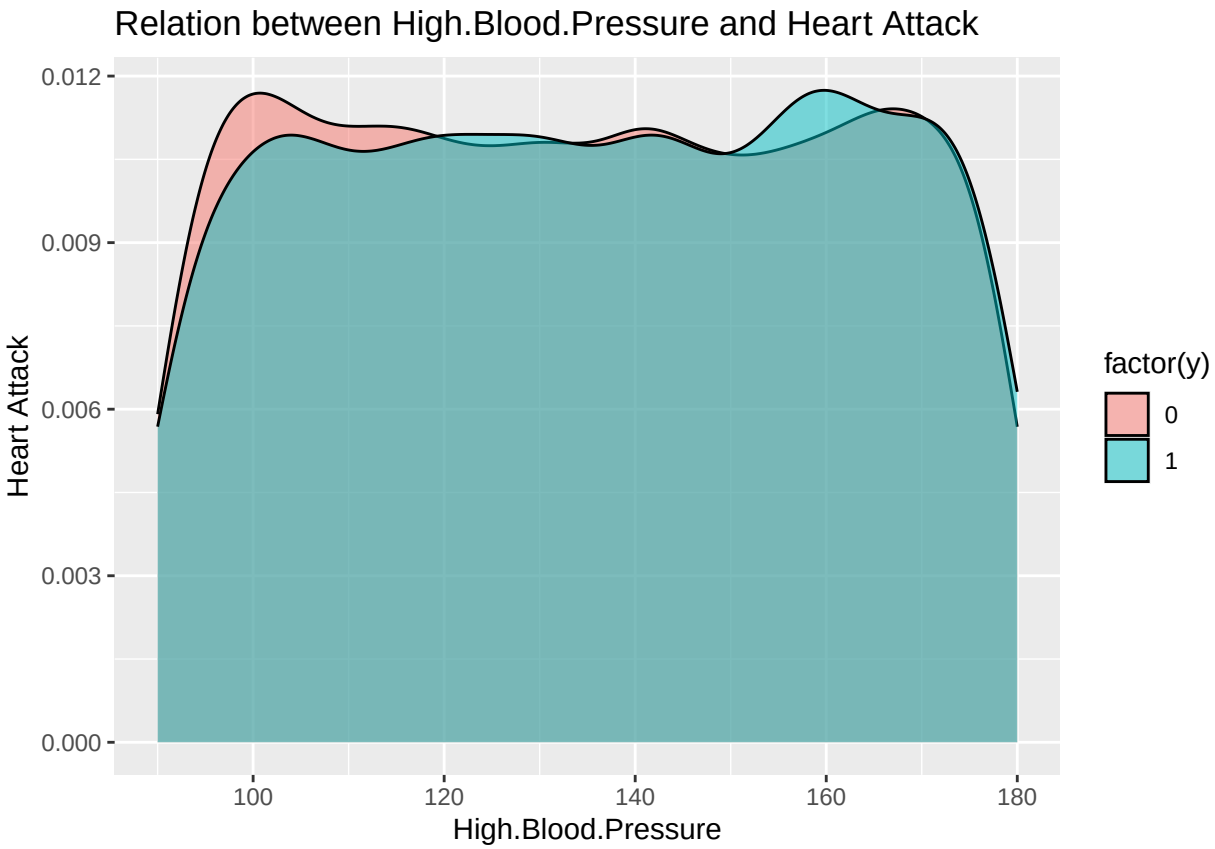



```
ggplot(data, aes(x = Age, fill = factor(y))) +
  geom_density(alpha = 0.5)+
  labs(title = "Relation between Age and Heart Attack", x = "Age", y = "Heart Attack")
```



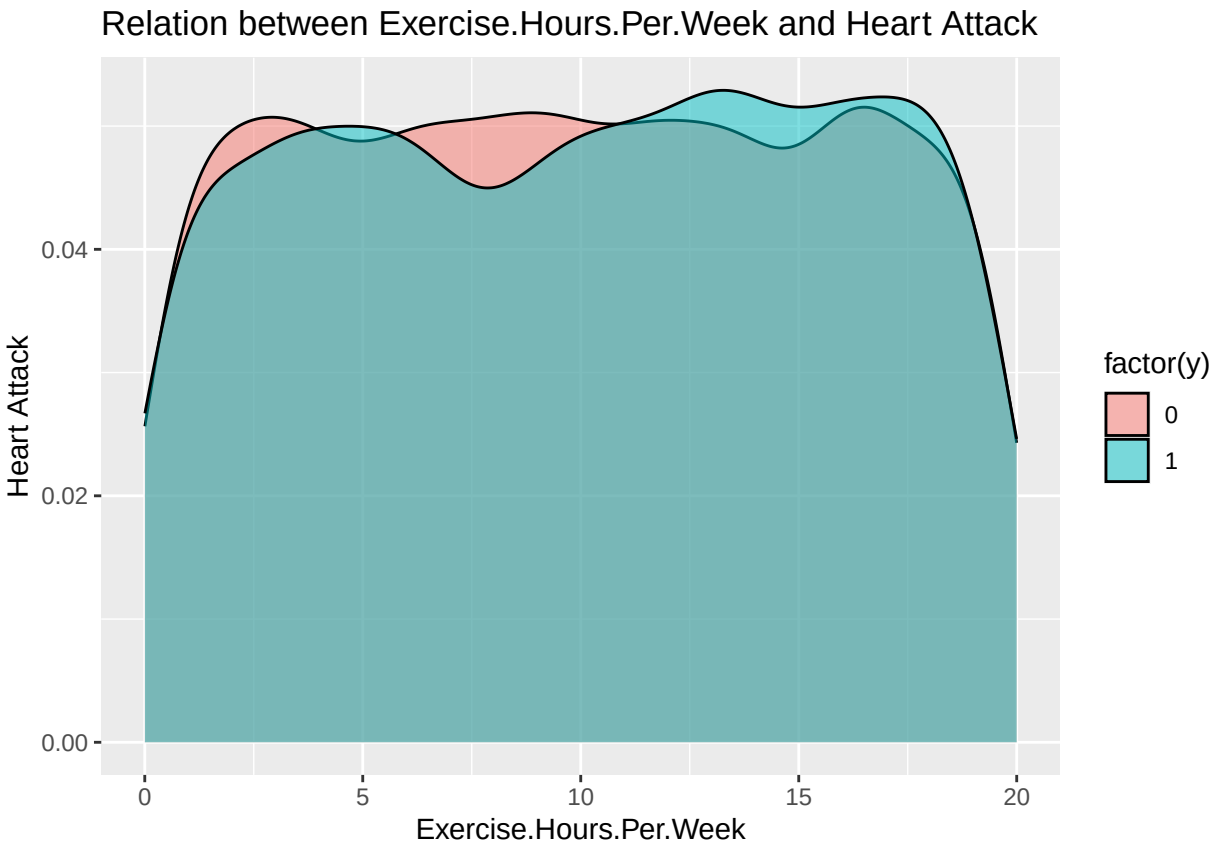
From the Relation between Age and Heart Attack, it appears that middle-aged people are less likely to be at risk of heart disease than younger and older people.

```
ggplot(data, aes(x = High.Blood.Pressure, fill = factor(y))) +
  geom_density(alpha = 0.5)+
  labs(title = "Relation between High.Blood.Pressure and Heart Attack", x = "High.Blood.Pressure", y =
```



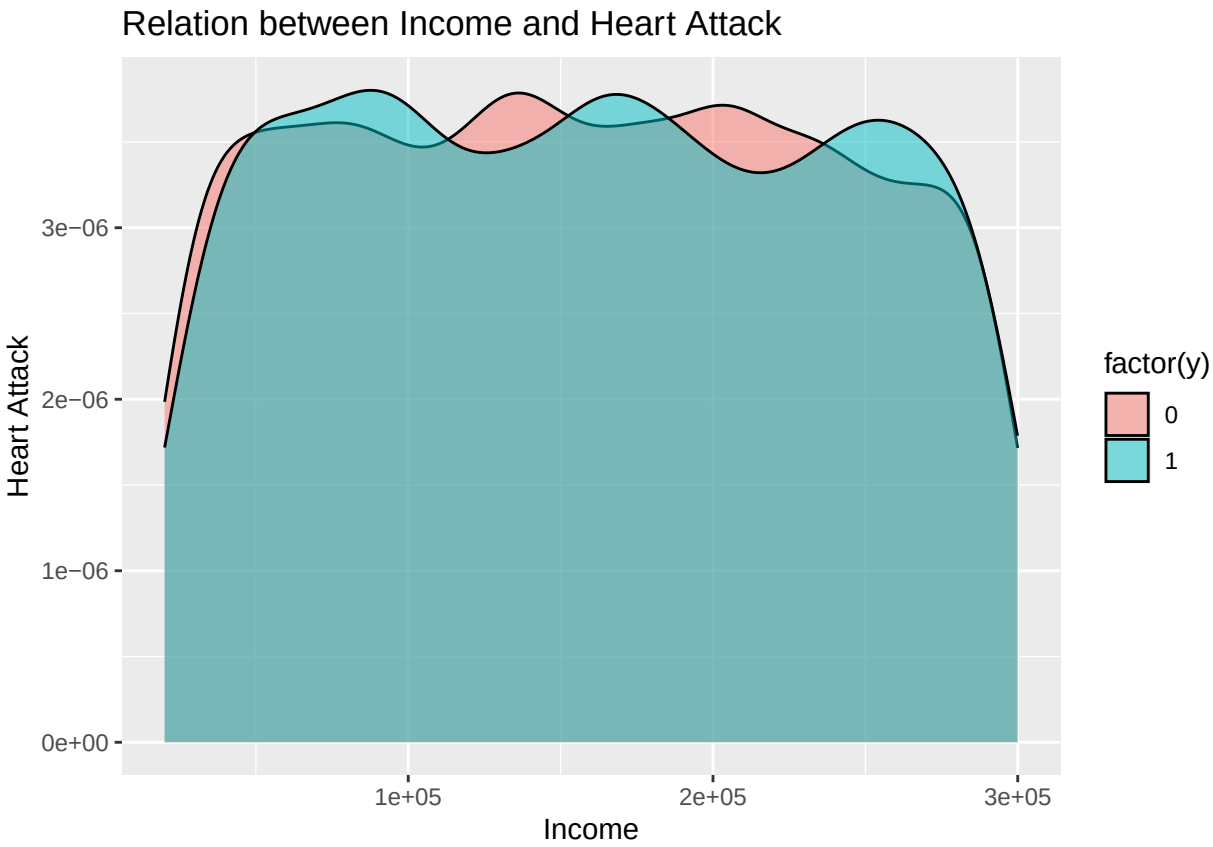
From the Relation between High.Blood.Pressure and Heart Attack, it seems that if a person's High.Blood.Pressure is lower than 120, he or she is relatively free of the risk of heart attack.

```
ggplot(data, aes(x = Exercise.Hours.Per.Week, fill = factor(y))) +
  geom_density(alpha = 0.5)+
  labs(title = "Relation between Exercise.Hours.Per.Week and Heart Attack", x = "Exercise.Hours.Per.Week
```



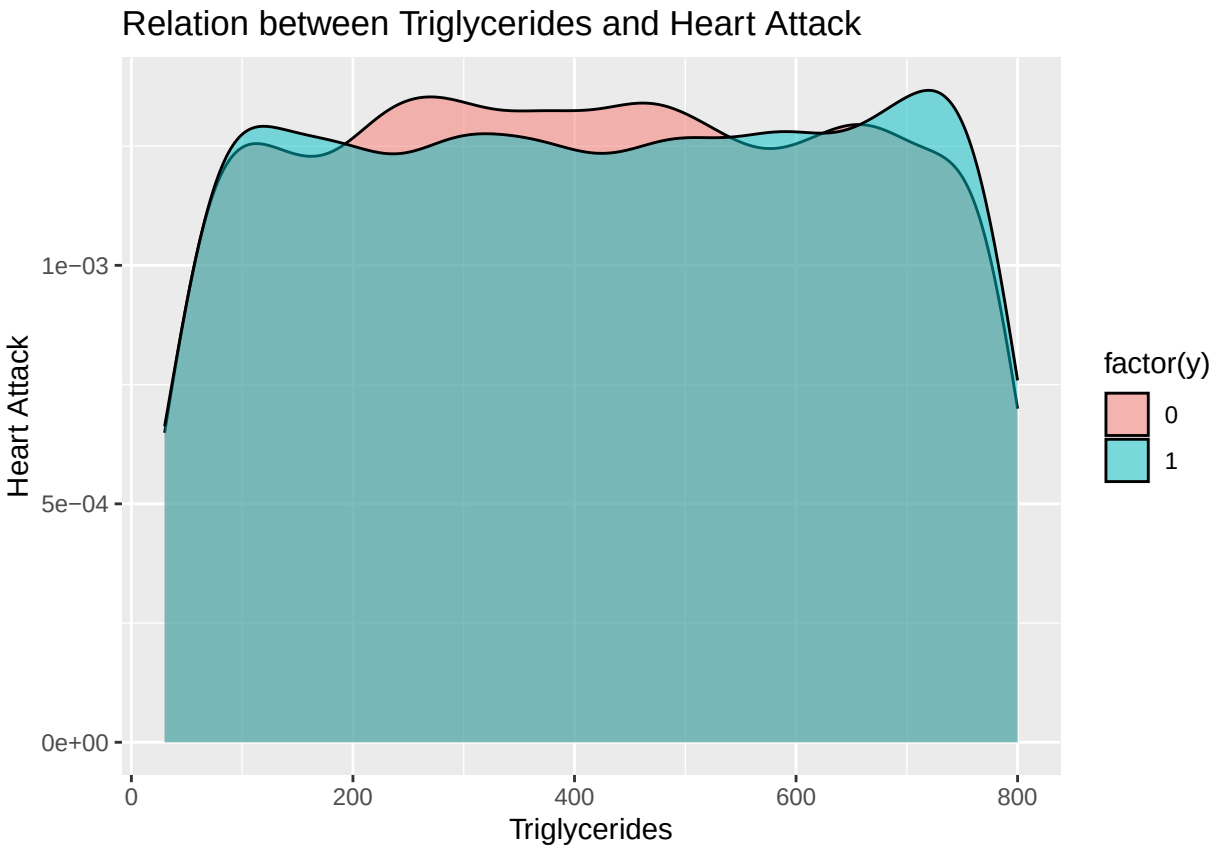
It is often said that active exercise contributes to physical and mental health. But from this figure, we found that if the Exercise.Hours.Per..Week is more than 10, it seems to be more likely to cause heart disease.

```
ggplot(data, aes(x = Income, fill = factor(y))) +
  geom_density(alpha = 0.5)+
  labs(title = "Relation between Income and Heart Attack", x = "Income", y = "Heart Attack")
```



There did not appear to be a significant relationship between income and heart disease risk.

```
ggplot(data, aes(x = Triglycerides, fill = factor(y))) +
  geom_density(alpha = 0.5)+
  labs(title = "Relation between Triglycerides and Heart Attack", x = "Triglycerides", y = "Heart Attac
```



According to medical wisdom, high levels of plasma triglycerides can be associated with an increased risk of cardiovascular disease. This picture seems to confirm it.

Methods

A Multilevel Regression Model is a statistical model used to analyze hierarchical data. It is also known as a Hierarchical Linear Model (HLM), Mixed-Effects Model, or Random Effects Model. This model is suitable for situations where research data contains multiple hierarchies, where observations (individuals, groups, or other units) belong to different groups at different levels and may be influenced by their respective levels.

The basic idea of a multilevel regression model is to consider multiple levels of variability simultaneously when analyzing data. These levels can be formalized by the following formula:

First: Let's divide the data set into 2 parts by the variable Hemisphere. Namely, a single data point belongs to Northern Hemisphere or Southern Hemisphere and the information of Continent and Country are ignored here. Here Hemisphere of a single data point equals to 1 means the patient lives in Northern Hemisphere. Thus the β estimated for it denotes the difference of impact in heart attack between 2 half Hemispheres.

$$Y = BX_1 + \epsilon \quad (1)$$

```
data$Hemisphere<- ifelse(data$Hemisphere == "Northern Hemisphere", 1, 0)
X_Hemisphere <- data[, -which(names(data) %in%  c("Patient.ID","Country","Continent","Heart.Attack.Risk

model_Hemisphere <- glm(y ~ ., data = cbind(X_Hemisphere,y), family = binomial(link = "logit"))
```

Second: Let's divide the data set into 6 parts by the variable Continent. Namely, a single data point belongs to its continent and the information of Country are ignored here. One patient is assumed to live in North

America by default, so no binary variable for North America is set in X_2. Hemisphere remains here for some continents straddle the northern and southern hemispheres.

$$Y = BX_2 + \epsilon \quad (2)$$

```
data$Africa <- ifelse(data$Continent == "Africa", 1, 0)
data$Asia <- ifelse(data$Continent == "Asia", 1, 0)
data$Australia <- ifelse(data$Continent == "Australia", 1, 0)
data$Europe <- ifelse(data$Continent == "Europe", 1, 0)
data$SouthAmerica <- ifelse(data$Continent == "South America", 1, 0)

X_Continent <- data[, -which(names(data) %in% c("Patient.ID","Country","Continent","Heart.Attack.Risk"))]

model_Continent <- glm(y ~ ., data = cbind(X_Continent,y), family = binomial(link = "logit"))
```

Third: Let's move the hierarchy down to the national level, where different patients come from different countries. Patients in each country were observed to be from the same hemisphere or continent. This is especially true in countries that straddle the northern and southern hemispheres, such as Colombia and Brazil. So here we delete the two variables of Continent and Hemisphere. And one patient is assumed to live in United States by default, so no binary variable for US is set in X_3.

$$Y = BX_3 + \epsilon \quad (3)$$

```
data$Argentina <- ifelse(data$Country == "Argentina", 1, 0)
data$Australia1 <- ifelse(data$Country == "Australia", 1, 0)
data$Brazil <- ifelse(data$Country == "Brazil", 1, 0)
data$Canada <- ifelse(data$Country == "Canada", 1, 0)
data$China <- ifelse(data$Country == "China", 1, 0)
data$Colombia <- ifelse(data$Country == "Colombia", 1, 0)
data$France <- ifelse(data$Country == "France", 1, 0)
data$Germany <- ifelse(data$Country == "Germany", 1, 0)
data$India <- ifelse(data$Country == "India", 1, 0)
data$Italy <- ifelse(data$Country == "Italy", 1, 0)
data$Japan <- ifelse(data$Country == "Japan", 1, 0)
data$NewZealand <- ifelse(data$Country == "New Zealand", 1, 0)
data$Nigeria <- ifelse(data$Country == "Nigeria", 1, 0)
data$SouthAfrica <- ifelse(data$Country == "South Africa", 1, 0)
data$SouthKorea <- ifelse(data$Country == "South Korea", 1, 0)
data$Spain <- ifelse(data$Country == "Spain", 1, 0)
data$Thailand <- ifelse(data$Country == "Thailand", 1, 0)
data$UK <- ifelse(data$Country == "United Kingdom", 1, 0)
data$Vietnam <- ifelse(data$Country == "Vietnam", 1, 0)

X_Country <- data[, -which(names(data) %in%  c("Patient.ID","Country","Continent","Heart.Attack.Risk",
                    "Africa","Asia","Australia","Europe","SouthAmerica",
                    "Hemisphere"))]

model_Country <- glm(y ~ ., data = cbind(X_Country,y), family = binomial(link = "logit"))
```

Results

After constructing the three multiple logistic regression models, the summary() function is used to obtain the model results. Through observation and analysis, it is found that it is difficult to get accurate fitting because

of the quality of variables. Some machine learning methods, such as random forest algorithms and support vector machine algorithms, have also been tried, but the results are not good enough. The final choice is to present only the results of multiple regression analysis and conduct the appropriate analysis.

```
summary(model_Hemisphere)
```

```
##
## Call:
## glm(formula = y ~ ., family = binomial(link = "logit"), data = cbind(X_Hemisphere,
##     y))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0962  -0.9523  -0.9079   1.4059   1.5819
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -7.685e-01  2.815e-01  -2.730  0.00633 **
## Age                             1.174e-03  1.176e-03   0.998  0.31807
## Sex                             4.406e-02  5.842e-02   0.754  0.45073
## Cholesterol                     5.010e-04  2.765e-04   1.812  0.07003 .
## Heart.Rate                     -4.569e-04  1.087e-03  -0.421  0.67411
## Diabetes                        7.827e-02  4.703e-02   1.664  0.09610 .
## Family.History                 -2.770e-03  4.468e-02  -0.062  0.95056
## Smoking                        -9.477e-02  9.579e-02  -0.989  0.32248
## Obesity                        -5.706e-02  4.467e-02  -1.277  0.20148
## Alcohol.Consumption            -6.027e-02  4.549e-02  -1.325  0.18519
## Exercise.Hours.Per.Week         3.988e-03  3.861e-03   1.033  0.30158
## Previous.Heart.Problems         4.774e-03  4.467e-02   0.107  0.91488
## Medication.Use                  1.064e-02  4.467e-02   0.238  0.81166
## Stress.Level                   -3.205e-03  7.817e-03  -0.410  0.68177
## Sedentary.Hours.Per.Day        -3.849e-03  6.442e-03  -0.598  0.55013
## Income                          2.410e-07  2.771e-07   0.870  0.38452
## BMI                            -1.426e-04  3.534e-03  -0.040  0.96781
## Triglycerides                   8.971e-05  9.983e-05   0.899  0.36887
## Physical.Activity.Days.Per.Week -4.398e-03  9.782e-03  -0.450  0.65298
## Sleep.Hours.Per.Day            -1.876e-02  1.124e-02  -1.669  0.09502 .
## Hemisphere                      5.487e-02  4.680e-02   1.172  0.24108
## High.Blood.Pressure             1.490e-03  8.474e-04   1.758  0.07877 .
## Low.Blood.Pressure             -1.171e-03  1.522e-03  -0.769  0.44178
## Unhealthy.Diet                  2.321e-02  5.499e-02   0.422  0.67297
## Healthy.Diet                    5.331e-02  5.455e-02   0.977  0.32842
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11434  on 8762  degrees of freedom
## Residual deviance: 11410  on 8738  degrees of freedom
## AIC: 11460
##
## Number of Fisher Scoring iterations: 4
```

For model_Hemisphere, the summary report found that people living in the Northern Hemisphere were significantly more likely to have heart disease than those living in the southern hemisphere.

```
summary(model_Continent)
```

```
##
## Call:
## glm(formula = y ~ ., family = binomial(link = "logit"), data = cbind(X_Continent,
##     y))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1066  -0.9520  -0.9041   1.4028   1.6183
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -7.716e-01  2.956e-01  -2.610  0.00904 **
## Age                             1.132e-03  1.177e-03   0.962  0.33625
## Sex                             4.199e-02  5.847e-02   0.718  0.47262
## Cholesterol                     5.005e-04  2.767e-04   1.809  0.07046 .
## Heart.Rate                     -4.945e-04  1.087e-03  -0.455  0.64923
## Diabetes                        8.063e-02  4.707e-02   1.713  0.08673 .
## Family.History                 -4.333e-03  4.471e-02  -0.097  0.92281
## Smoking                        -9.039e-02  9.587e-02  -0.943  0.34572
## Obesity                        -5.800e-02  4.469e-02  -1.298  0.19431
## Alcohol.Consumption            -6.023e-02  4.552e-02  -1.323  0.18575
## Exercise.Hours.Per.Week         3.960e-03  3.863e-03   1.025  0.30534
## Previous.Heart.Problems         3.164e-03  4.471e-02   0.071  0.94357
## Medication.Use                  1.262e-02  4.470e-02   0.282  0.77760
## Stress.Level                   -3.579e-03  7.822e-03  -0.458  0.64727
## Sedentary.Hours.Per.Day        -4.109e-03  6.450e-03  -0.637  0.52413
## Income                          2.408e-07  2.772e-07   0.869  0.38498
## BMI                            -2.064e-04  3.535e-03  -0.058  0.95344
## Triglycerides                   9.022e-05  9.987e-05   0.903  0.36628
## Physical.Activity.Days.Per.Week -4.086e-03  9.787e-03  -0.417  0.67635
## Sleep.Hours.Per.Day            -1.924e-02  1.125e-02  -1.711  0.08707 .
## Hemisphere                      1.259e-01  6.489e-02   1.941  0.05229 .
## High.Blood.Pressure             1.482e-03  8.478e-04   1.748  0.08048 .
## Low.Blood.Pressure             -1.167e-03  1.523e-03  -0.766  0.44348
## Unhealthy.Diet                  2.395e-02  5.504e-02   0.435  0.66348
## Healthy.Diet                    5.425e-02  5.461e-02   0.993  0.32050
## Africa                          3.265e-02  1.043e-01   0.313  0.75438
## Asia                           -9.618e-02  8.187e-02  -1.175  0.24007
## Australia                       6.584e-02  1.188e-01   0.554  0.57949
## Europe                         -8.572e-02  8.693e-02  -0.986  0.32410
## SouthAmerica                    4.428e-02  1.005e-01   0.441  0.65949
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11434  on 8762  degrees of freedom
## Residual deviance: 11404  on 8733  degrees of freedom
## AIC: 11464
##
## Number of Fisher Scoring iterations: 4
```

However, people on other continents are not significantly more exposed to heart disease than North Americans, according to model_Continent's summary report.

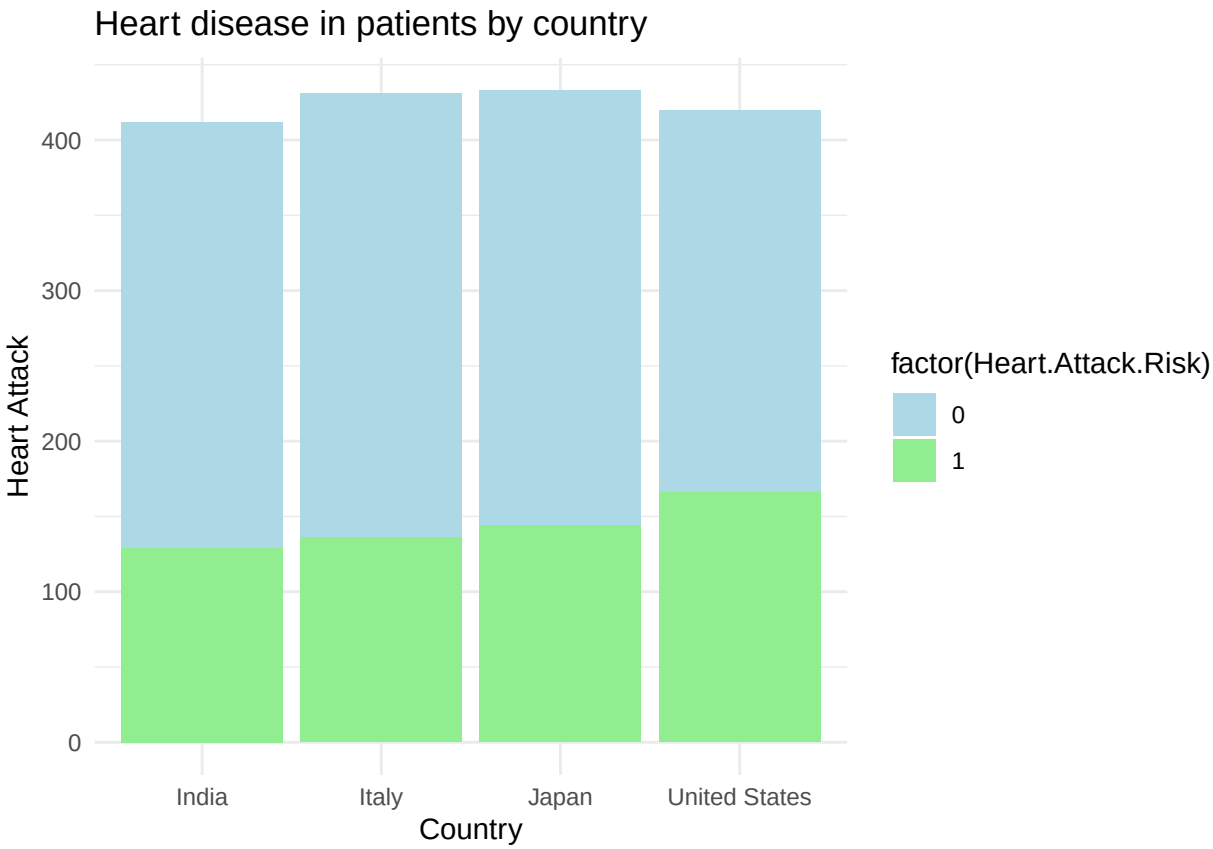
```
summary(model_Country)
```

```
##
## Call:
## glm(formula = y ~ ., family = binomial(link = "logit"), data = cbind(X_Country,
##     y))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1297  -0.9543  -0.8951   1.3928   1.6385
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -5.740e-01  2.985e-01  -1.923   0.0545 .
## Age                              1.132e-03  1.179e-03   0.960   0.3371
## Sex                              4.326e-02  5.855e-02   0.739   0.4600
## Cholesterol                      4.839e-04  2.770e-04   1.747   0.0807 .
## Heart.Rate                      -4.591e-04  1.089e-03  -0.422   0.6732
## Diabetes                         8.057e-02  4.713e-02   1.710   0.0873 .
## Family.History                  -7.844e-03  4.478e-02  -0.175   0.8609
## Smoking                         -8.983e-02  9.598e-02  -0.936   0.3493
## Obesity                         -5.756e-02  4.474e-02  -1.287   0.1982
## Alcohol.Consumption             -5.974e-02  4.559e-02  -1.310   0.1901
## Exercise.Hours.Per.Week          4.106e-03  3.868e-03   1.062   0.2884
## Previous.Heart.Problems          5.329e-03  4.479e-02   0.119   0.9053
## Medication.Use                   1.249e-02  4.476e-02   0.279   0.7801
## Stress.Level                    -3.942e-03  7.836e-03  -0.503   0.6149
## Sedentary.Hours.Per.Day         -4.204e-03  6.458e-03  -0.651   0.5151
## Income                           2.551e-07  2.776e-07   0.919   0.3582
## BMI                             -4.532e-04  3.540e-03  -0.128   0.8981
## Triglycerides                    8.317e-05  1.000e-04   0.831   0.4057
## Physical.Activity.Days.Per.Week -3.644e-03  9.804e-03  -0.372   0.7101
## Sleep.Hours.Per.Day             -1.947e-02  1.126e-02  -1.729   0.0838 .
## High.Blood.Pressure              1.506e-03  8.492e-04   1.774   0.0761 .
## Low.Blood.Pressure              -1.114e-03  1.525e-03  -0.730   0.4652
## Unhealthy.Diet                   2.904e-02  5.512e-02   0.527   0.5983
## Healthy.Diet                     5.876e-02  5.470e-02   1.074   0.2827
## Argentina                       -1.005e-01  1.385e-01  -0.726   0.4680
## Australia1                      -7.542e-02  1.399e-01  -0.539   0.5899
## Brazil                          -1.752e-01  1.399e-01  -1.253   0.2103
## Canada                          -1.439e-01  1.412e-01  -1.019   0.3082
## China                           -1.626e-01  1.417e-01  -1.148   0.2511
## Colombia                        -6.681e-02  1.414e-01  -0.472   0.6366
## France                          -1.812e-01  1.411e-01  -1.284   0.1990
## Germany                         -1.392e-01  1.384e-01  -1.006   0.3145
## India                           -3.447e-01  1.462e-01  -2.357   0.0184 *
## Italy                           -3.512e-01  1.443e-01  -2.435   0.0149 *
## Japan                           -2.732e-01  1.431e-01  -1.910   0.0562 .
## NewZealand                      -1.941e-01  1.423e-01  -1.364   0.1726
## Nigeria                          1.925e-02  1.392e-01   0.138   0.8900
## SouthAfrica                     -2.324e-01  1.435e-01  -1.620   0.1053
## SouthKorea                       2.985e-02  1.424e-01   0.210   0.8340
```

```
## Spain                          -1.941e-01  1.425e-01  -1.362   0.1732
## Thailand                       -7.727e-02  1.414e-01  -0.546   0.5848
## UK                             -1.823e-01  1.403e-01  -1.299   0.1939
## Vietnam                        -1.972e-01  1.429e-01  -1.380   0.1676
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11434  on 8762  degrees of freedom
## Residual deviance: 11391  on 8720  degrees of freedom
## AIC: 11477
##
## Number of Fisher Scoring iterations: 4
```

Interestingly, from the summary report of model_Country, it can be seen that Indians, Italians and Japanese are significantly less likely to get heart disease than Americans, and the two variables of India and Italy are very significant, which may be due to the different genetic genes and living habits of people in different countries. Moreover, regression parameters for most other countries are negative, and positive regression parameters (such as Nigeria and South Korea) are very insignificant, which may indicate that the United States, Nigeria and South Korea are the countries with the highest incidence of heart disease among these countries.

```
selected_countries <- c("Japan", "Italy", "India", "United States")
subset_data <- data[data$Country %in% selected_countries,]

ggplot(subset_data, aes(x = Country, fill = factor(Heart.Attack.Risk))) +
  geom_bar() +
  labs(title = "Heart disease in patients by country", x = "Country", y = "Heart Attack") +
  scale_fill_manual(values = c("0" = "light blue", "1" = "light green")) +
  theme_minimal()
```



Conclusions

To sum up, this data analysis report draws the following conclusions and puts forward some relevant recommendations.

Correlation analysis

According to the correlation analysis:

1. Adolescents and the elderly have higher risk of heart disease than middle-aged people.
2. There is a strong correlation between high blood pressure and heart disease.
3. Moderate exercise can help cardiovascular health, but too much exercise can lead to a more serious risk of heart disease.
4. The correlation between income and heart disease risk was not significant.
5. Too low or too high triglycerides may lead to an increased risk of heart disease.

Regression analysis

According to regression analysis:

1. Cholesterol, High blood pressure and family history have significant positive effects on heart disease risk.
2. People with diabetes are more likely to develop heart disease.

3. Adequate sleep time contributes to cardiovascular health.
4. There are no significant differences in heart disease risk between continents, but Indians, Italians and Japanese have significantly lower risk of heart disease than Americans.

Suggestions on people's life

1. young people should pay more attention to healthy living habits, the elderly should always pay attention to their own health indicators.
2. High blood pressure, high cholesterol and high triglycerides are red flags for heart attack.
3. Diabetes patients should be more concerned about their cardiovascular health.
4. Maintain adequate sleep more than 7 hours a day and exercise less than 10 hours a week.
5. Learn about Indian, Japanese and Italian lifestyles and eating habits.

Weaknesses

The sample size of the data is sufficient, but it does not fully reveal the impact of clinical indicators and lifestyle habits on heart disease risk. If you use dimensionality reduction methods such as PCA, or machine learning and deep learning methods, you can indeed get better accuracy for prediction, but these methods lack interpretability, so it is difficult to make recommendations for clinicians' diagnoses and people's lifestyle habits. Perhaps some more advanced interpretable machine learning methods can be applied to this dataset to get more accurate advice and help.

Ethics Statement

First, this report highlight the data sources. This study uses open data, which means that it relies on publicly available data resources that do not involve the collection of personal privacy or sensitive information. This decision was made out of respect for data privacy, while also contributing to the reproducibility and transparency of the research. We are committed to ensuring the legality and ethics of data in order to avoid infringement of data providers.

Second, this report focused on ethical issues in the model development and evaluation process. This study did not make P-value adjustments to prevent excessive pursuit of statistical significance from leading to spurious findings. This report emphasize the interpretability of the model rather than just focusing on the predictive accuracy of the model. This is to ensure that our research findings can be understood and trusted by the wider community, rather than just relying on a black box model.