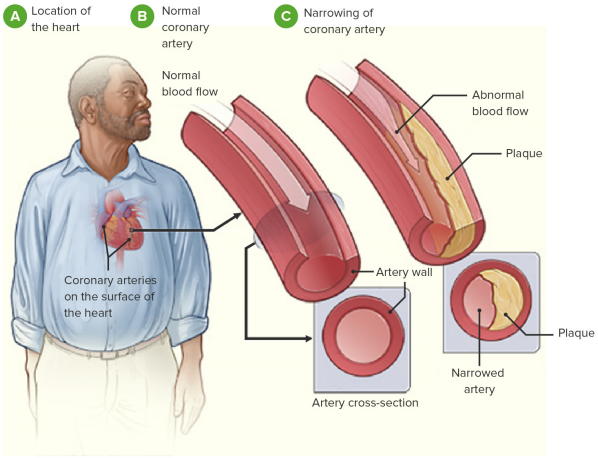
IT2020E : APPLIED STATISTICS

# Analysis on the Western Collaborative Group Study Dataset on Epidemiology




*Team:*

| | |
|---|---|
| Nguyen Manh Duong | 20210243 |
| Huynh Sang | 20214930 |
| Nguyen Song Hao | 20210532 |
| Pham Trung Hieu | 20214899 |

*Supervisor:*
Dr. Nguyen Linh Giang

Semester 2022.2
Jul, 2023

# Contents

**Abstract**

The Western Collaborative Group Study (WCGS) dataset is a valuable resource in epidemiology for understanding the risk factors associated with coronary heart disease (CHD). This longitudinal study, conducted in the 1960s and 1970s, followed a cohort of initially healthy men aged 39-59 years, collecting extensive data on medical history, physical examinations, laboratory tests, and lifestyle factors. Through the analysis of this dataset, our project aims to investigate the relationships between smoking, cholesterol levels, blood pressure, obesity, and the development of CHD. By exploring these associations, we hope to contribute to the existing knowledge base on cardiovascular health and provide insights for future prevention and intervention strategies.

Our project focuses on the tasks of Data Analysis, Statistical Model (Logistic Regression) and Survival Analysis. Through those tasks, we are able to draw some conclusions and test our prior assumptions about the data:

# 1 Data Analysis

## 1.1 Structures of WCGS

WCGS is composed of 10 features corresponding to each medical aspect when considering a patient's health status. Those features are respectively indicated as follows:

| Column | Meaning |
|---|---|
| id | ID |
| age | age, in years - no decimal places |
| height | height, in inches |
| weight | weight, in pounds |
| sbp | systolic blood pressure, in mmHg |
| dbp | diastolic blood pressure, in mmHg |
| chol | total cholesterol, mg/dL |
| behpat | behavioral pattern: A1, A2, B3, B4 |
| nicgs | number of cigarettes smoked per day |
| dibpat | behavioral pattern: A or B |

Table 1: Columns explained

In the original dataset, there is some missing information in cholesterol and arcus fields, which accounts for 0.38% and 0.06% of total observations. Since the missed observations are not significant in the whole dataset, we dropped these observations from the dataset. Before analysis, we accomplished some data preprocessing methods to clean data and convert it to our familiar form. Importantly, weight and height are simultaneously converted to kilograms (kg) and centimeters (cm), respectively

The following figure shows the descriptive analysis of the dataset. At first glance, the height is around its mean value (177.23 cm) with a slight variance (6 cm) indicating the fact that there is a huge range in western resident's heights. In contrast, a significantly large variance was observed in their weight, which ranges from 35.41 kg to 145.28 kg with a standard variation of

| | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|
| id | 10479.727157 | 5877.192558 | 2001.00 | 3741.75 | 11405.50 | 13114.25 | 22101.00 |
| age0 | 46.278553 | 5.522564 | 39.00 | 42.00 | 45.00 | 50.00 | 59.00 |
| height (cm) | 177.231079 | 6.422476 | 152.40 | 172.72 | 177.80 | 182.88 | 198.12 |
| weight (kg) | 77.158798 | 9.580260 | 35.41 | 70.37 | 77.18 | 82.63 | 145.28 |
| sbp0 | 128.626904 | 15.120601 | 98.00 | 120.00 | 126.00 | 136.00 | 230.00 |
| dbp0 | 82.007931 | 9.725147 | 58.00 | 76.00 | 80.00 | 86.00 | 150.00 |
| chol0 | 226.345541 | 43.338322 | 103.00 | 198.00 | 223.00 | 252.00 | 645.00 |
| behpat0 | 2.523160 | 0.799141 | 1.00 | 2.00 | 2.00 | 3.00 | 4.00 |
| ncigs0 | 11.602792 | 14.520504 | 0.00 | 0.00 | 0.00 | 20.00 | 99.00 |
| dibpat0 | 0.503807 | 0.500065 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |

Table 2: Basic descriptive statistics of WCGS

just under 10 kg. Interestingly, the follow-up time for each subject seems to randomly fluctuate since the number of days goes from 18 days to approximately 10 years.

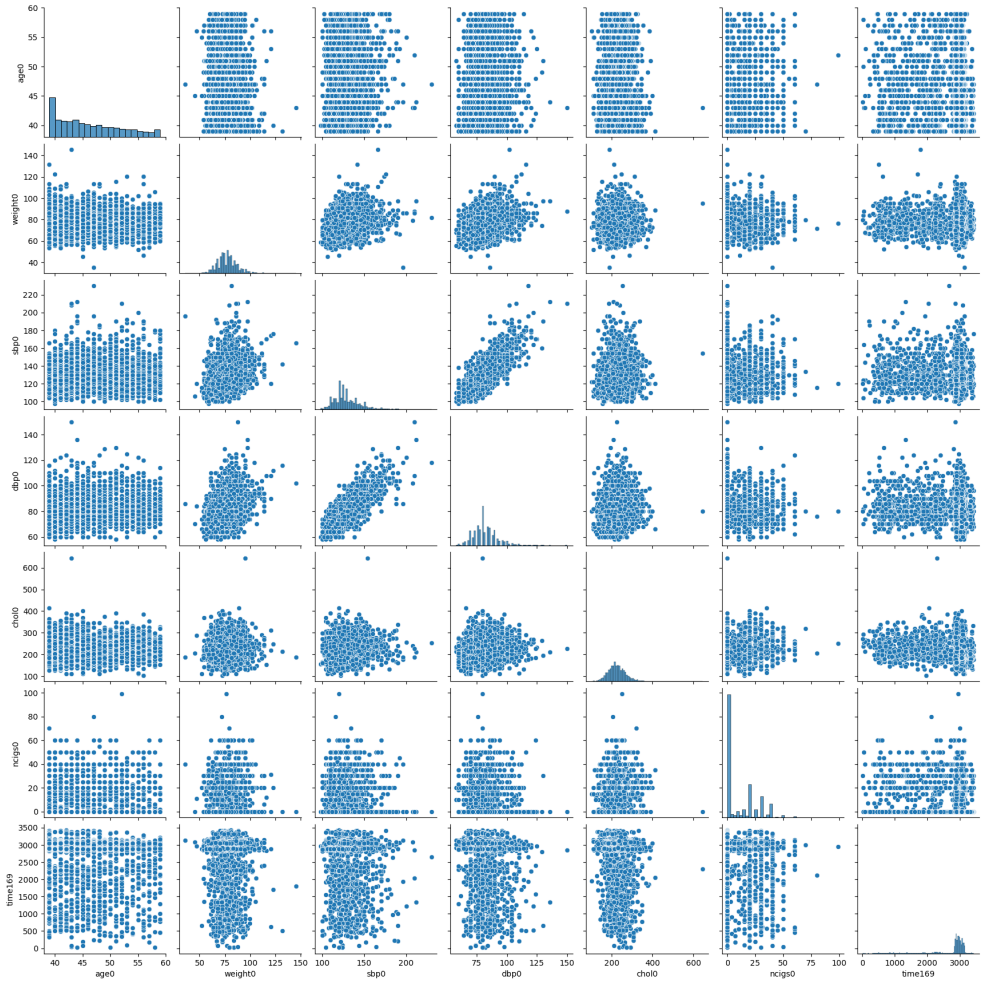



Figure 1: Correlation between numerical features

Figure 1 shows the linear correlations between numerical features in the dataset. A significant covariate trend was observed in weight, height, systolic blood pressure (SBP), diastolic blood pressure (DBP), and cholesterol level (CHOL). Specifically, SBP and DBP are strongly correlated with each other. It is reasonable in a medical sense since diastolic and systolic blood is basically one flow of blood, just different positions in the human body. Also, CHOL seems to be normally distributed with a similar bell shape comparing the Gaussian distribution. One more interesting thing is that most of the numerical features are likely to be uniformly distributed across the follow-up time (in days). This observation might be one of the pieces of evidence for the assumption that health status does not depend on the follow-up time.

## 1.2 Additional variables

Firstly, from the data set WCGS, we create two new variables:

1. The first new variable is Smoker. Each person (row) has the attribute Smoker equals to 1 if that person smokes, which means the attribute ncigs is greater than zero. Otherwise, the attribute Smoker of the person is 0.

2. The second new variable is BMI. This attribute is calculated by using two attributes weight and height (by using the formula 1)

$$BMI = 703 * \frac{weight}{height^2} \tag{1}$$

## 1.3 Pairplot graph

Then, we plot the pairplot graph (image 2) for all 15 attributes in the data to see into data in more detail. Looking to the plot, we can hypothesize that: Height, weight, cholesterol, sbp, dbp, BMI are bell-shaped. These variables could be normal, but we need to check with tests later on.

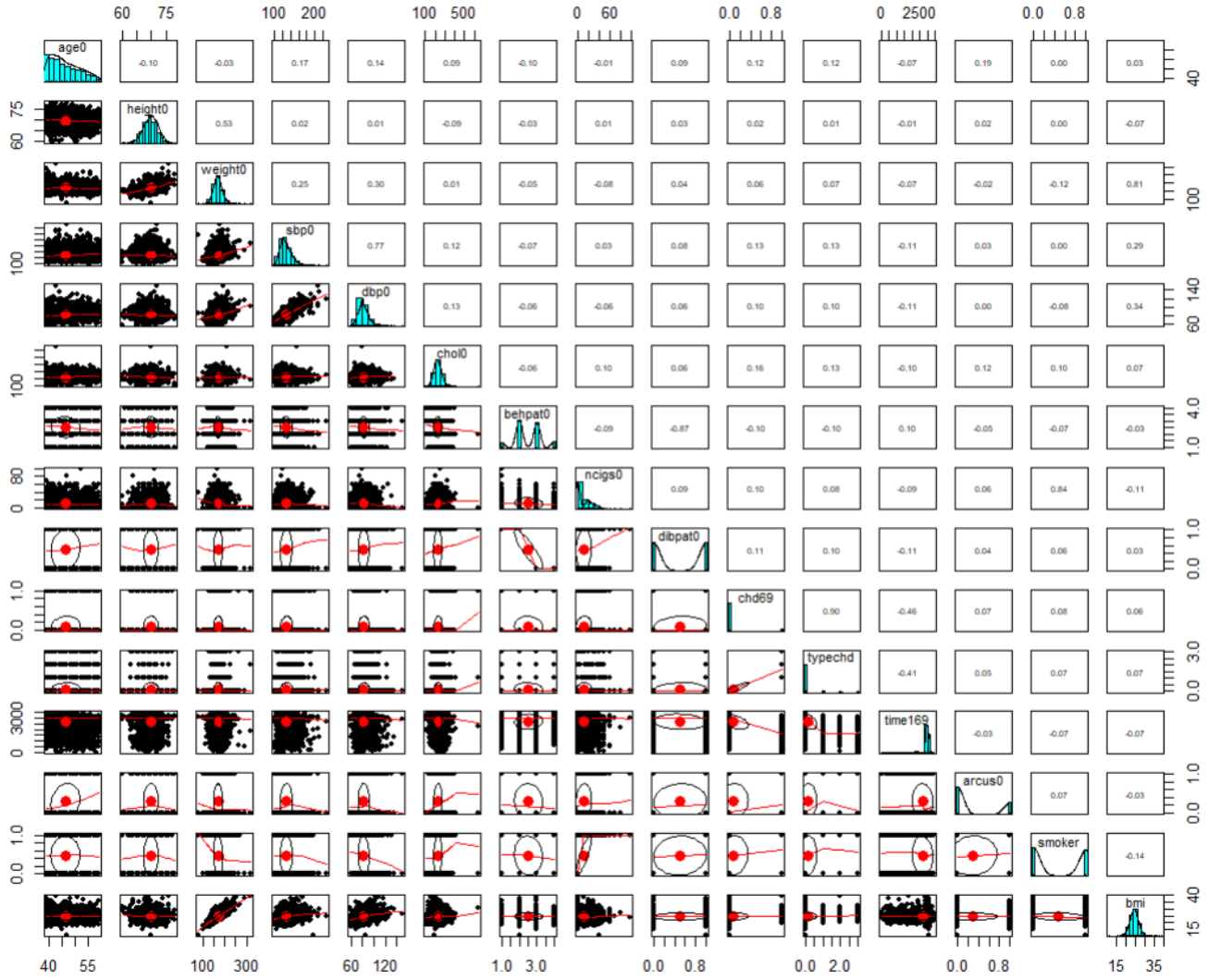



Figure 2: The pairplot of 15 attributes.

We can also inspect correlation through the heatmap:

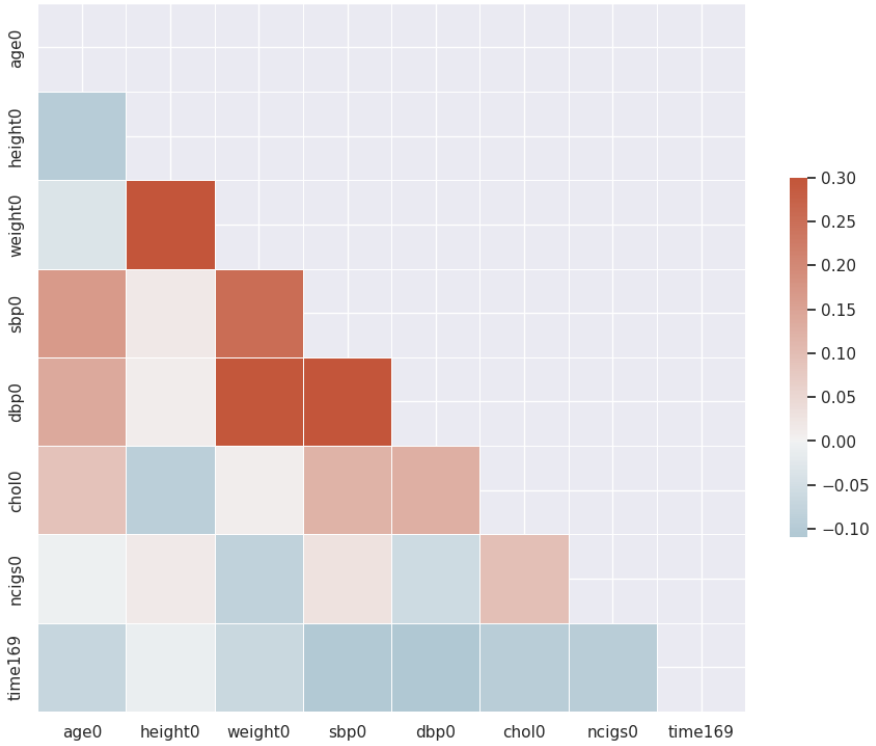



Figure 3: Correlation between numerical features

From Figure 3, we can draw the following conclusions:

- SBP is slightly correllated on weight and age .
- DBP is correllated on age, weight, and SBP.
- Weight and height are strongly correllated on each other with large positive covariance.

As we run the models in the next step, we will see that if we have both weight and height in a model, or both sbp and dbp, one of them will have less statistical significance.

## 1.4 Variable plots and Normality Check

### 1.4.1 BMI

Looking at the figure 4, we can see that almost the BMI values is in the range from 22 to 27.

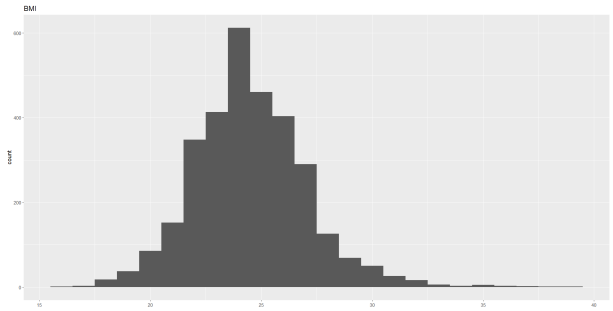



Figure 4: The distribution of BMI

The histogram demonstrated a bell-shaped curve, with the peak occurring in the 24-25 range(count $> 600$). This indicates that a significant proportion of individuals falls within the "normal weight" category according to the WHO classification. Individuals within this range are generally considered to have a healthy weight for their height. So that, it demonstrates a positive aspect of overall health and a lower risk of weight-related health complications.

### 1.4.2 Cholesterol

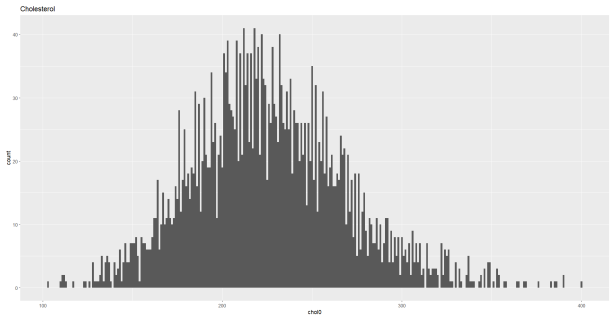



Figure 5: Histogram on Cholesterol

Looking at figure 6, the analysis of cholesterol distribution within the studied population indicated the highest frequency of values in the 200-240 range, which falls within the borderline high cholesterol category. This highlights the potential increased risk of developing coronary heart disease.

We can see that there are people with extremely high cholesterol level through the box plot:

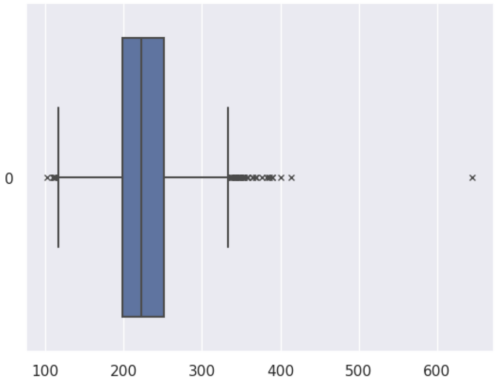



Figure 6: Boxplot on Cholesterol

As at the time of the survey, these people did not have coronary heart disease yet, so these values are extremely unusual. Observations from box plot depicted that there were 1.64% of total patients whose cholesterol levels were out of the interquartile range. As those data points were distant from the reasonable range of cholesterol levels, we removed those observations and then run Kolmogorov-Smirnov Test to check for normality.

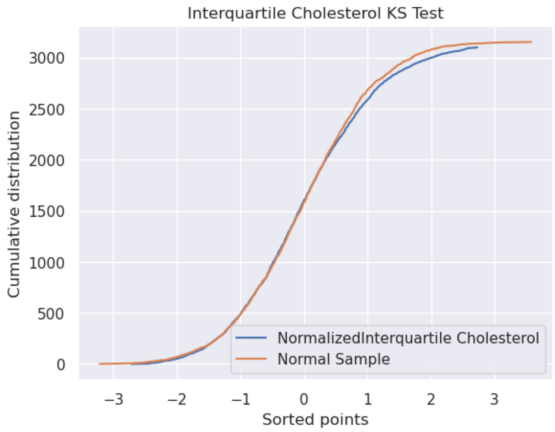



Figure 7: KS Test of Interquartile Cholesterol

Kolmogorov-Smirnov Test as shown in Figure 6 indicated that p-values and D statistics were 0.036 and 0.025 respectively. With a p-value smaller than 0.05 but greater than 0.01, we did have moderate statistical evidence but not strong enough to totally reject the null hypothesis that the Cholesterol level feature comes from the Normal distribution. With more data collected in further research, we expect to make a more reliable conclusion with strong statistical evidence about cholesterol level feature's normality.

### 1.4.3 DBP and SBP

Diastolic and systolic blood pressure are considered to be two of the most crucial features to evaluate patients' health status. In the conventions of medical fields, diastolic and systolic blood pressures are blood pressure but from different places in the human body. Thus, one reasonable assumption about DBP and SBP is their correlation.

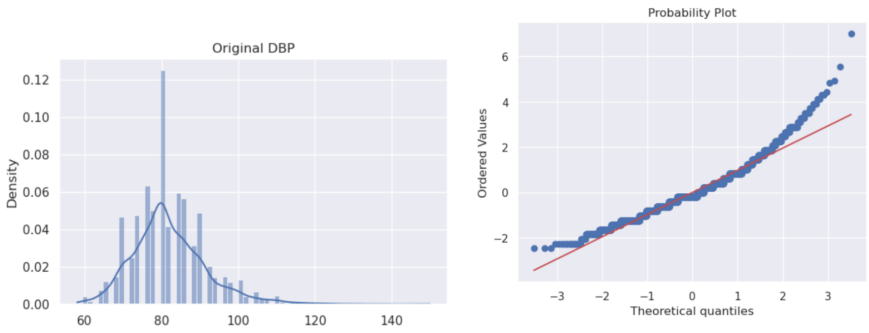



Figure 8: Density of Diastolic blood pressure

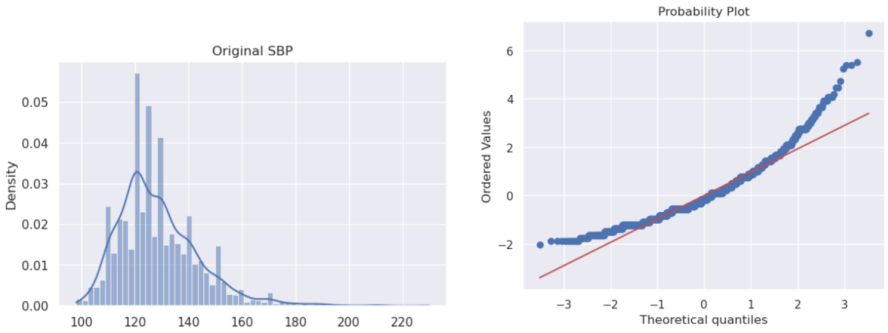



Figure 9: Density of Systolic blood pressure

As our prior assumption about correlated blood pressures, the aforementioned figures 8, 9 depict that systolic and diastolic blood pressure come from a slightly similar distribution which is a skewed bell-shaped curve. Quantile-quantile plots (QQ plots) also show the similarity between those two types of features in terms of normality. In addition, the p-values from their KS Test (after being normalized) are both approximately 0 with significantly high D statistics being around 0.12. Thus, there is significant evidence enough to conclude that DBP and SBP were not normally distributed.

### 1.5 Data cleaning

Suppose that after consulting with clinical experts, we want to ensure that:

- All age values are between 39 and 59 years
- All bmi are between 15 and 50
- All sbp are between 80 and 250 mm Hg
- All dbp are between 50 and 200 mm Hg
- All values of sbp-dbp are at least 10 and no more than 90 mm Hg
- All values of chol are between 100 and 400 mg/dl

So that, we filter out the rows that don't satifiy these constraint. After filtering, the number of data decrease from 3154 to 3130.

## 2 Statistical Model

Within what the project covers, we focus on *Logistic Regression*. We are going to inspect the relationships between Coronary Heart Disease event variable versus multiple independent variables.

## 2.1 Simple models

### 2.1.1 Single independent variable models

```
Call:
glm(formula = chd69 ~ bmi, family = binomial, data = wcgs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6732  -0.4262  -0.4002  -0.3710   2.4296

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.49247    0.61424  -7.314 2.59e-13 ***
bmi          0.08295    0.02444   3.393  0.00069 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1748.0  on 3129  degrees of freedom
Residual deviance: 1736.9  on 3128  degrees of freedom
AIC: 1740.9

Number of Fisher Scoring iterations: 5
```

Figure 10: CHD vs BMI

Regressing CHD event on BMI shows us that: One unit increase in BMI is associated with an increase in the log-odds of the response variable (coronary heart disease) by 0.08295.

To interpret this in a more intuitive manner, we can say that for every one-unit increase in BMI, the odds of having coronary heart disease increase by a factor of exp(0.08295), or approximately 1.086.

The p-value of the coefficient of BMI is very small, hence there is strong evidence to suggest that there is a relationship between BMI and the probability of having coronary heart disease, i.e., the coefficient for bmi is significantly different from zero.

Similarly, one could also interpret the model CHD vs Age: an increase in age results in an increase of 0.07555 in log-odds of CHD.

```
Call:
glm(formula = chd69 ~ age0, family = binomial, data = wcgs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6195  -0.4512  -0.3630  -0.3252   2.4941

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.01119    0.55613 -10.809  < 2e-16 ***
age0         0.07555    0.01144   6.607 3.93e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1748.0  on 3129  degrees of freedom
Residual deviance: 1704.8  on 3128  degrees of freedom
AIC: 1708.8

Number of Fisher Scoring iterations: 5
```

Figure 11: CHD vs Age

### 2.1.2 Multiple independent variables models

Here we will try to regress CHD on Age, BMI and Cholesterol level.

```
Call:
glm(formula = chd69 ~ age0 + bmi + chol0, family = binomial,
    data = wcgs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0300  -0.4446  -0.3444  -0.2651   2.7867

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.331487   0.919611 -11.235  < 2e-16 ***
age0          0.069772   0.011672   5.978 2.26e-09 ***
bmi           0.074420   0.025528   2.915  0.00355 **
chol0         0.011660   0.001512   7.711 1.25e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1748.0  on 3129  degrees of freedom
Residual deviance: 1635.8  on 3126  degrees of freedom
AIC: 1643.8

Number of Fisher Scoring iterations: 5
```

Figure 12: CHD vs Age, BMI and Cholesterol

For this multiple logistic regression model, one-unit increase in each independent variable produces the following effects on the log odds of the response variable (CHD):

- age: log-odds of CHD increase by 0.069772 units for every one-unit increase in age, holding all other variables constant.
- bmi: log-odds of CHD increase by 0.074420 units for every one-unit increase in BMI, holding all other variables constant.
- cholesterol level: log-odds of CHD increase by 0.011660 units for every one-unit increase in cholesterol level, holding all other variables constant.

**Likelihood Ratio Test**

We can perform a Likelihood Ratio (Deviance) Test to check if any subset of the coeffients are equal to zero.

- Null hypothesis $H_0$: The set of coefficient $\beta_{i1} = \beta_{i2} = ... = 0$
- Alternate hypothesis $H_1$: $\exists \beta_{ij} \neq 0$

```
Model:
chd69 ~ age0 + bmi + chol0
       Df Deviance    AIC    LRT  Pr(>Chi)
<none>      1635.8 1643.8
age0    1   1671.3 1677.3 35.454 2.612e-09 ***
bmi     1   1644.1 1650.1  8.285  0.003997 **
chol0   1   1694.6 1700.6 58.780 1.764e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 13: Drop 1 table for the above model

In our test, we only try to drop one variable at a time. The p-value for each part indicates that the $\beta$'s are significant at $\alpha = 1\%$ confidence level

**Multicollinearity Test**

We are having multiple independent variables, hence we would want to check if there is multicollinearity in these variables. We refer to the Variance Inflation Factor table:

| age | bmi | chol |
|---|---|---|
| 1.000666 | 1.000322 | 1.000985 |

If the VIF value of a variable is above 5, then there might exist some multicollinearity in the variables. In this case, this is not a concerning problem because all the VIF are approximate 1, which is good for us.

## 2.2 Model selection

Using the package bestglm from R, we can find the best model when accounting for the variables: age, bmi, height, weight, sbp, dbp, smoker, cholesterol, behavior pattern, corneal arcus event. The function selects the best GLM using the Bayesian Information Criterion (BIC) value using exhaustive search. The best model found by this method is:

```
BIC
BICq equivalent for q in (0.0135960877809465, 0.570117117588597)
Best Model:
               Estimate  Std. Error    z value     Pr(>|z|)
(Intercept) -9.93702915 0.866044981 -11.474034 1.781571e-30
age0         0.05872591 0.012036449   4.879006 1.066216e-06
sbp0         0.02134786 0.004140877   5.155395 2.530971e-07
smoker       0.56912864 0.140039073   4.064070 4.822428e-05
chol0        0.01043598 0.001544608   6.756390 1.414724e-11
behpat0     -0.35522372 0.088166523  -4.029009 5.601247e-05
```

Figure 14: Best model found by exhaustive search

Realize that this model's coefficients has very small p-value, which is favourable for us. We can compare this model to other models:

| variables | model1 | model2 | model3 | model4 | model5 |
|---|---|---|---|---|---|
| age | 0.0587*** | 0.0602*** | 0.0630*** | 0.0595*** | 0.0595 |
| sbp | 0.0213*** | 0.0188*** | 0.0186*** | 0.0189*** | 0.0194** |
| dbp | | | | | -0.0010 |
| smoker | 0.5691*** | 0.6130*** | 0.6104*** | 0.6000*** | 0.5984*** |
| chol | 0.0104*** | 0.0103*** | 0.0106 | 0.0104*** | 0.0104*** |
| behpat | -0.3552*** | -0.3498*** | -0.3435*** | -0.3425*** | -0.3429*** |
| height | | | 0.0156 | 0.0129 | 0.0127 |
| weight | | | 0.0078* | 0.0080* | 0.0081* |
| arcus | | | | 0.2019 | 0.2011 |
| bmi | | 0.0535* | | | |
| intercept | -9.9370*** | -11.0208*** | -12.3248*** | -12.0579*** | -12.0291*** |

We can note that, when sbp and dbp are in the same model, the coefficient for dbp is not statistically significant. This is probably because of sbp and dbp are highly correlated, as stated above. The same can be said about height and weight, plus height is a weaker indicator of coronary heart disease event.

We would like to perform log-likelihood test of model 2 to 5 against model 1

| Log-likelihood vs model1 | model2 | model3 | model4 | model5 |
|---|---|---|---|---|
| Pr(>Chisq) | 0.04831* | 0.02121* | 0.02176* | 0.04653* |

We can conclude that the interaction terms of model 2 to 5 are not significant at a 2% significant level.

## 2.3 Assumptions Testing

Logistic regression is based on a few assumptions:

- Dependent variable is binary
- The observation are independent
- The sample size is sufficiently large
- There are no extreme outliers
- There is no multicollinearity among independent variables
- There is a linear relationship between the independent variables and the logit of the dependent variable

The first four assumptions are inspected in our data cleaning step, hence there is little need to perform these tests again. We will use Variance Inflation Factor (VIF) to check for assumption number 5, and Box-Tidwell Test to check for assumption number 6.

**Multicollinearity Test**

```
> vif(model1)
    age0     sbp0   smoker    chol0  behpat0
1.032939 1.031892 1.011698 1.009664 1.009321
```

Figure 15: Variance Inflation Factor for model 1

Refering to the VIF table above, we can see that there is no sign of multicollinearity.

### 2.3.1 Box-Tidwell Test

The Box-Tidwell test is used to check for linearity between the predictors and the logit. This is done by adding log-transformed interaction terms between the continuous independent variables and their corresponding natural log into the model.

- Null hypothesis $H_0$: linearity between logit and independent variable

We inspect linearity of logit of the response variable versus the numeric independent variables, one by one, in Model 1:

| | chd - age | chd - chol | chd - sbp | chd - bmi |
|---|---|---|---|---|
| Pr(>\|t\|) | 0.3318 | 0.696 | 0.2441 | 0.4613 |

All the p-values are above 0.05, hence we do not have enough evidence to reject the null hypothesis. We can comfortably say that there is linearity between the logit of CHD event and the independent variables in our best model.

# 3 Survival Analysis

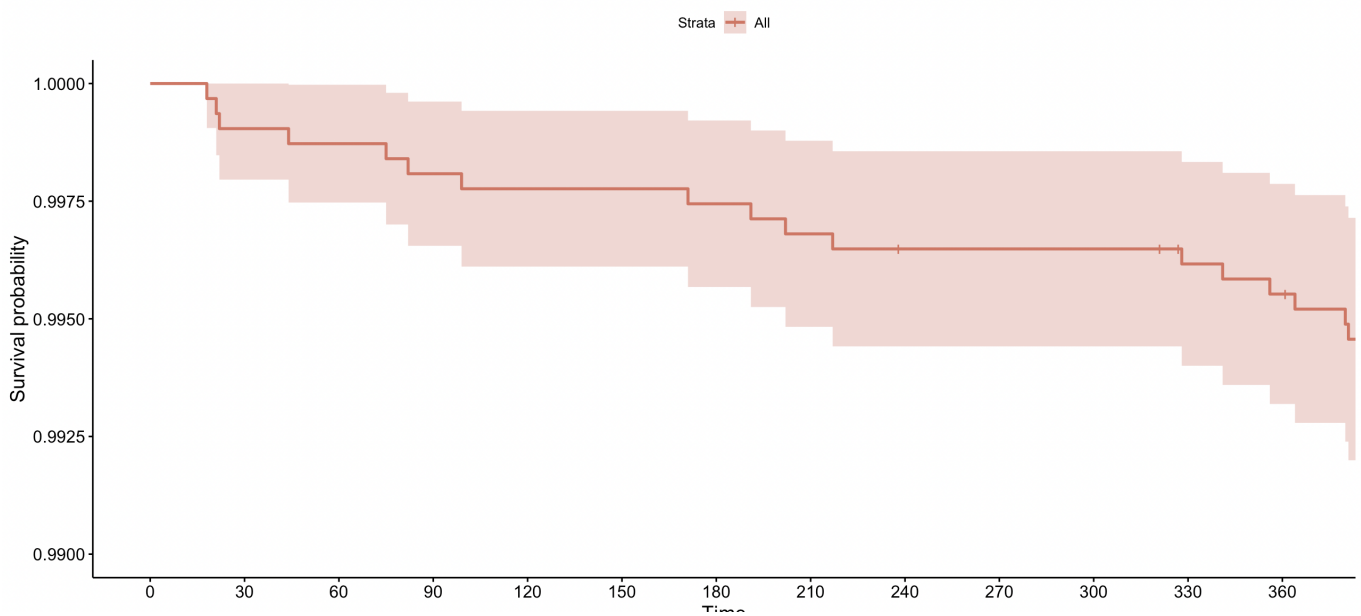
First, we discover the information contained in how long they were followed without developing heart disease.

In R, we combine this information by creating a variable of the survival object type. We create it with the command Surv , which is in the survival package, and evaluate how long until the event or the end of follow-up and whether an event or impairment occured at that point of time.

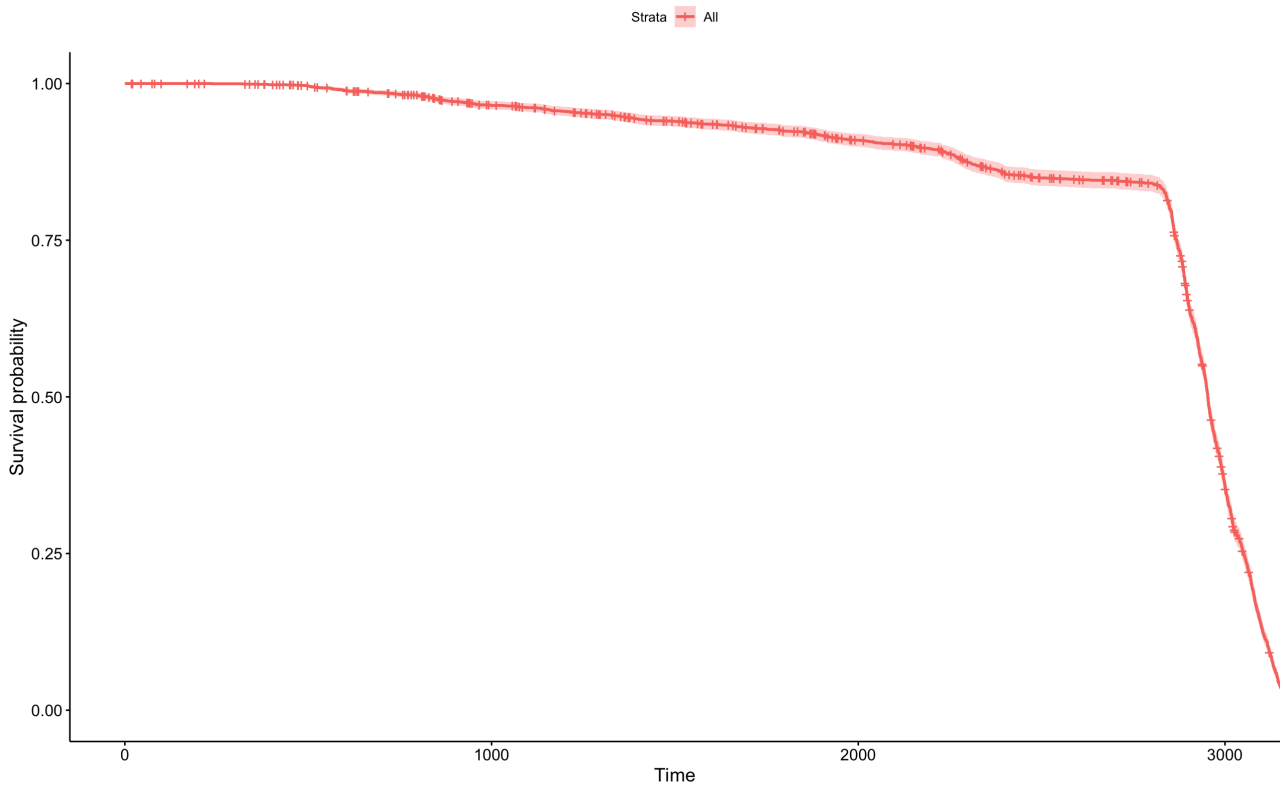
```
> survival <- Surv(wcgs$time169, wcgs$chd69)
> head(survival)
[1] 1664+ 3071+ 3071+ 3064+ 1885  3102+
```

The first measurement is 1664. The plus indicates that it was followed for 1664 days without an event and therefore the measurement is impaired. However, the fifth measurement, 1885 has no plus indicating that the man developed heart disease after 1885 days.

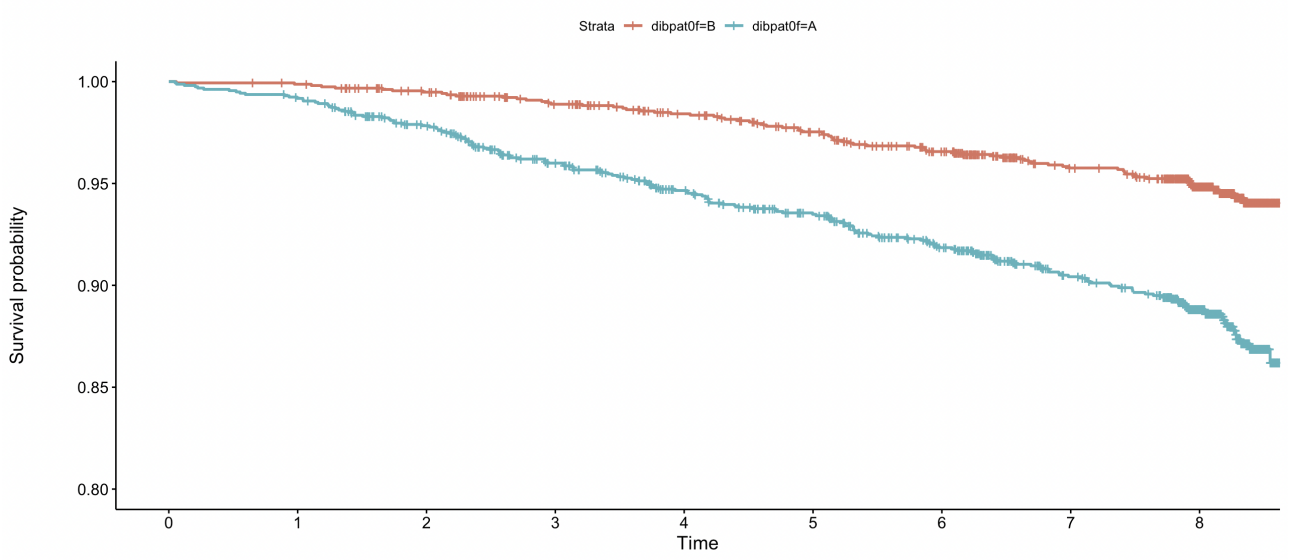
After that, we perform the Kaplan Meier graph to display the probability of surviving beyond a certain time point.




Above is the Kaplan-Meier chart for the first 365 days. From the graph, we can see that no one has had heart disease at the start of the study so the survival probability is 100%. The first drop occurs precisely when the first man got a heart disease. There the lines falls from 1.00 to 0.997 and the next fall will be at when the next person had a heart attack.



If we inspect the full follow-up timeline, there is a plummet at around day 3000. This is due to the fact that this is near the end of the follow-up sequence, which is only 10 years.




Here we change the scale to year and look at time to heart disease according to the personalities of A and B.

In this figure, there appears to be a large difference in the incidence of heart disease by personality type. The blue curve, which represents personality type A, is all below the red curve, which represents personality type B.

As a result, we hypothesize that the probability of not having CHD is lower for people of personality type B. We test this hypothesis using log-rank test.

```
> survdiff(Surv(time169,chd69)~dibpat0f,data=wcgs)
Call:
survdiff(formula = Surv(time169, chd69) ~ dibpat0f, data = wcgs)

               N Observed Expected (O-E)^2/E (O-E)^2/V
dibpat0f=B 1553       77      128      20.1        41
dibpat0f=A 1577      174      123      20.8        41

 Chisq= 41  on 1 degrees of freedom, p= 2e-10
```

Referring to the figure below, we see that:

- At 1 year, the percentage of people of type B behavior not getting CHD is 0.999, and of type A is 0.992
- At 5 year, the percentage of people of type B behavior not getting CHD is 0.975, and of type A is 0.935

The p-value of the log-rank test is very small, hence we can comfortably conclude our hypothesis.

```
> summary(kmfit.2, times=c(365.25,5*365.25))
Call: survfit(formula = Surv(time169, chd69) ~ dibpat0f, data = wcgs)

                dibpat0f=B
 time n.risk n.event survival  std.err lower 95% CI upper 95% CI
  365   1549       2    0.999 0.000911        0.997        1.000
 1826   1422      35    0.975 0.004006        0.968        0.983

                dibpat0f=A
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  365   1562      13    0.992 0.00228        0.987        0.996
 1826   1338      86    0.935 0.00634        0.923        0.947
```

# 4 Conclusion

In this project, we conducted an analysis of the Western Collaborative Group Study (WCGS) dataset on epidemiology to investigate the risk factors associated with coronary heart disease (CHD). Through our analysis, we aimed to explore the relationships between smoking, cholesterol levels, blood pressure, obesity, and the development of CHD.

We began by performing data analysis to understand the structure and characteristics of the dataset. We examined the basic descriptive statistics of the variables, explored correlations between numerical features, and conducted variable plots to assess normality and distributions. This initial analysis provided us with insights into the dataset and helped us identify variables of interest for further investigation.

Next, we employed logistic regression to model the relationship between CHD and several independent variables. We started with simple models, examining the impact of individual variables such as age, BMI, and cholesterol on the likelihood of developing CHD. We then built a multiple logistic regression model, including age, BMI, cholesterol level, smoker status, behavioral pattern, height, weight, and corneal arcus event as independent variables. The best model selection process helped us identify the most significant predictors of CHD and evaluate their statistical significance.

We used the Kaplan-Meier method to estimate the survival probability over time and examined the differences in survival curves based on personality types.

Based on our analysis, we found that BMI, age, smoker status, and cholesterol level were significant predictors of CHD. Individuals with higher BMI, older age, and higher cholesterol levels had a higher likelihood of developing CHD. Additionally, personality type A was associated with a lower risk of CHD compared to personality type B.

# 5 Further Research

More research is needed in order to come to more conclusions about our dataset. In this setting, we accomplished statistical analyzing some important features and predictive model selection focusing on Logistic Regression. For further research, we are planning to:

- Survey the linearity and correlation of more features
- Fitting predictive models and assessing performances


# Acknowledgement

We would like to thank our supervisor, Dr. Nguyen Linh Giang, for being supportive and always open to our queries. This capstone project is a valuable experience for all of us to learn and develop ourselves.


# References


https://online.stat.psu.edu/stat462/node/207/
https://rdrr.io/cran/epitools/man/wcgs.html
https://thomaselove.github.io/431-notes/17-wcgs.html
https://cpb-us-w2.wpmucdn.com/voices.uchicago.edu/dist/6/1366/files/2017/01/Statistics-I
pdf
https://www.statisticshowto.com/kolmogorov-smirnov-test/
https://www.statology.org/assumptions-of-logistic-regression/ https://bggj.is/SurvivalAn
introduction-to-survival-analysis.html https://thomaselove.github.io/431-notes/
17-wcgs.html