12/13/2023

Final Project 5301_001

Statistic Theory and Application

Applied Statistics and Data Science

Jani Shariff Shaik Vishnu Sai Muppalla
UNIVERSITY OF TEXAS AT ARLINGTON

Two-Way ANOVA : The Impact of Exercise and Cholesterol on Systolic Blood Pressure.

Introduction :

The project aimed to investigate the combined influence of exercise and cholesterol levels on systolic blood pressure. The study involved conducting a Two-Way Analysis of Variance (ANOVA) to explore the interplay between these two factors and their impact on systolic blood pressure.

Problem Statement :

The primary objective was to determine whether exercise (active) and cholesterol levels significantly contribute to variations in systolic blood pressure (ap_hi).

Data :

The dataset used for the analysis was obtained from Kaggle, containing information on individuals' demographic details, health metrics, and lifestyle factors.

```r
cardio_data <- read.csv('C:/Users/janis/Downloads/5301
Datasets/cardio_train.csv', sep=";")
str(cardio_data)
```

```
## 'data.frame':    70000 obs. of  13 variables:
##  $ id         : int  0 1 2 3 4 8 9 12 13 14 ...
##  $ age        : int  18393 20228 18857 17623 17474 21914 22113 22584 17668
19834 ...
##  $ gender     : int  2 1 1 2 1 1 1 2 1 1 ...
##  $ height     : int  168 156 165 169 156 151 157 178 158 164 ...
##  $ weight     : num  62 85 64 82 56 67 93 95 71 68 ...
##  $ ap_hi      : int  110 140 130 150 100 120 130 130 110 110 ...
##  $ ap_lo      : int  80 90 70 100 60 80 80 90 70 60 ...
##  $ cholesterol: int  1 3 3 1 1 2 3 3 1 1 ...
##  $ gluc       : int  1 1 1 1 1 2 1 3 1 1 ...
##  $ smoke      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ alco       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ active     : int  1 1 0 1 0 0 1 1 1 0 ...
##  $ cardio     : int  0 1 1 1 0 0 0 1 0 0 ...
```

```r
summary(cardio_data)
```

```
##        id             age            gender          height     
##  Min.   :    0   Min.   :10798   Min.   :1.00   Min.   : 55.0  
##  1st Qu.:25007   1st Qu.:17664   1st Qu.:1.00   1st Qu.:159.0  
##  Median :50002   Median :19703   Median :1.00   Median :165.0  
##  Mean   :49972   Mean   :19469   Mean   :1.35   Mean   :164.4  
##  3rd Qu.:74889   3rd Qu.:21327   3rd Qu.:2.00   3rd Qu.:170.0  
##  Max.   :99999   Max.   :23713   Max.   :2.00   Max.   :250.0  
##      weight           ap_hi             ap_lo            cholesterol   
##  Min.   : 10.00   Min.   : -150.0   Min.   :  -70.00   Min.   :1.000  
##  1st Qu.: 65.00   1st Qu.:  120.0   1st Qu.:   80.00   1st Qu.:1.000
##  Median : 72.00   Median :  120.0   Median :   80.00   Median :1.000
##  Mean   : 74.21   Mean   :  128.8   Mean   :   96.63   Mean   :1.367
##  3rd Qu.: 82.00   3rd Qu.:  140.0   3rd Qu.:   90.00   3rd Qu.:2.000
##  Max.   :200.00   Max.   :16020.0   Max.   :11000.00   Max.   :3.000
##       gluc           smoke               alco             active
##  Min.   :1.000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:1.0000
##  Median :1.000   Median :0.00000   Median :0.00000   Median :1.0000
##  Mean   :1.226   Mean   :0.08813   Mean   :0.05377   Mean   :0.8037
##  3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :3.000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##      cardio
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.4997
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

As we see in the above the data has there are 70000 rows and 13 columns but from the problem statement we only choose 3 columns or variables those are ap_hi (Systolic Pressure) But according to the requirement we only consider 3 columns.

Data Preprocessing:

Removing the unnecesary columns, and changing the column cholesterol to factor and ap_hi to numeric.

```
cardio_data <- cardio_data[, !names(cardio_data) %in% c("id")]
cardio_data <- cardio_data[, !names(cardio_data) %in% c("age")]
cardio_data <- cardio_data[, !names(cardio_data) %in% c("gender")]
cardio_data <- cardio_data[, !names(cardio_data) %in% c("height")]
cardio_data <- cardio_data[, !names(cardio_data) %in% c("weight")]
cardio_data <- cardio_data[, !names(cardio_data) %in% c("ap_lo")]
cardio_data <- cardio_data[, !names(cardio_data) %in% c("gluc")]
cardio_data <- cardio_data[, !names(cardio_data) %in% c("smoke")]
cardio_data <- cardio_data[, !names(cardio_data) %in% c("alco")]
cardio_data <- cardio_data[, !names(cardio_data) %in% c("cardio")]
# Convert cholesterol to factor
cardio_data$cholesterol <- as.factor(cardio_data$cholesterol)

cardio_data <- cardio_data[, !names(cardio_data) %in% c("id")]# Convert ap_hi to numeric
cardio_data$ap_hi <- as.numeric(cardio_data$ap_hi)
summary(cardio_data)
```

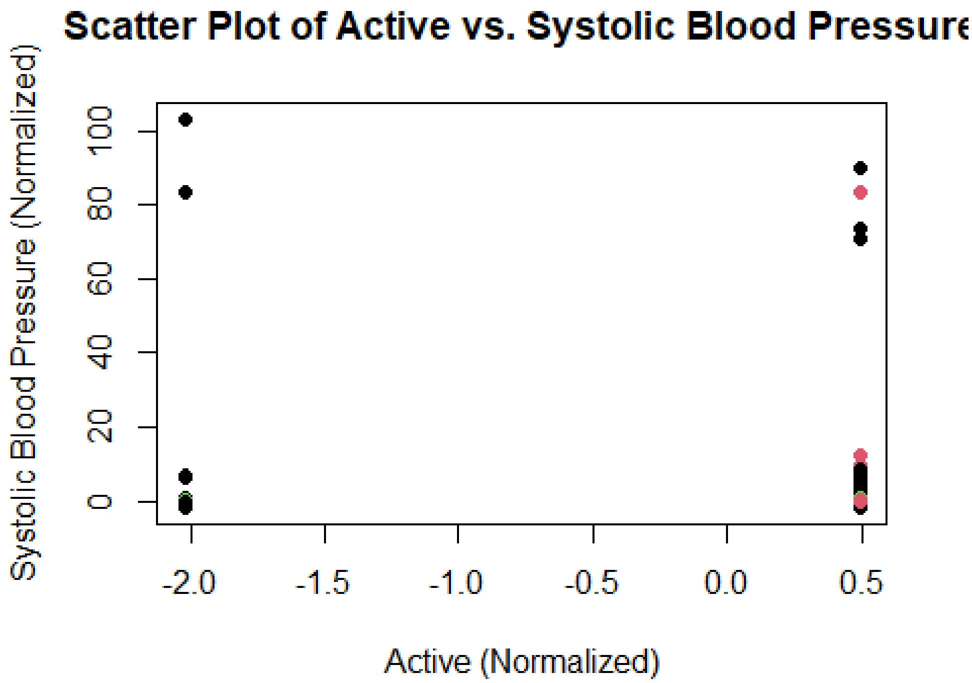
```
##      ap_hi          cholesterol     active
##  Min.   : -150.0   1:52385     Min.   :0.0000
##  1st Qu.:  120.0   2: 9549     1st Qu.:1.0000
##  Median :  120.0   3: 8066     Median :1.0000
##  Mean   :  128.8                Mean   :0.8037
##  3rd Qu.:  140.0                3rd Qu.:1.0000
##  Max.   :16020.0                Max.   :1.0000
```

Plots for exploratory Data analysis:

From the Plots and previous analysis we say that the data is not normally distributed but as the dataset is large we can proceed with Two way anova analysis. But we normalize the data before performing the analysis

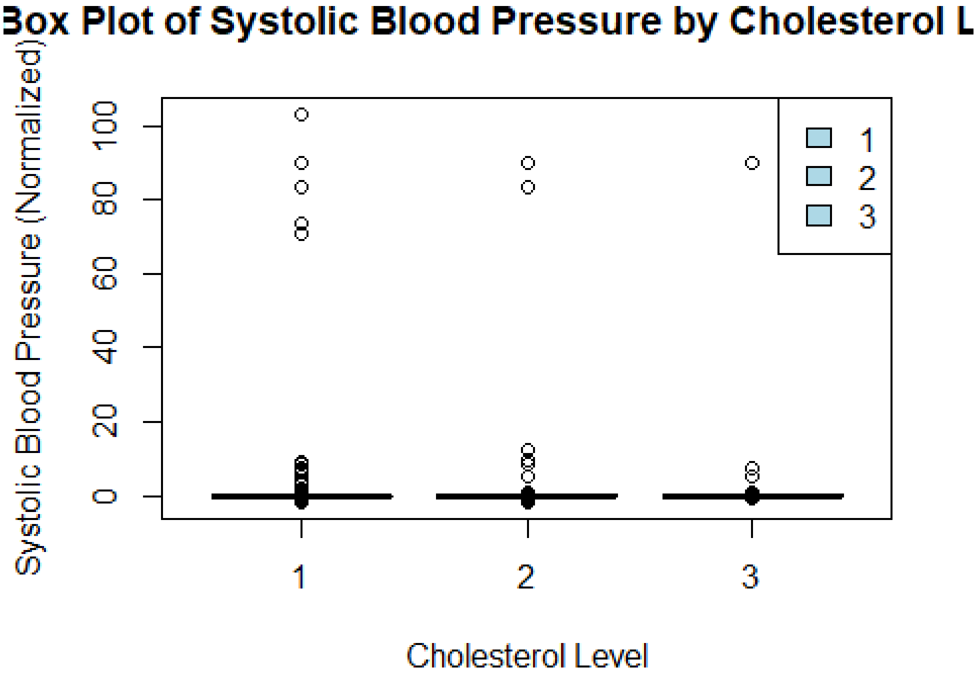
```
 # Normalize the data using z-score normalization
transformed_data <- scale(cardio_data[, c("active", "ap_hi")])

# Scatter plots for exploratory data analysis
plot(transformed_data[, "active"], transformed_data[, "ap_hi"],
     main = "Scatter Plot of Active vs. Systolic Blood Pressure",
     xlab = "Active (Normalized)", ylab = "Systolic Blood Pressure
(Normalized)",
     pch = 19, col = cardio_data$cholesterol)
```



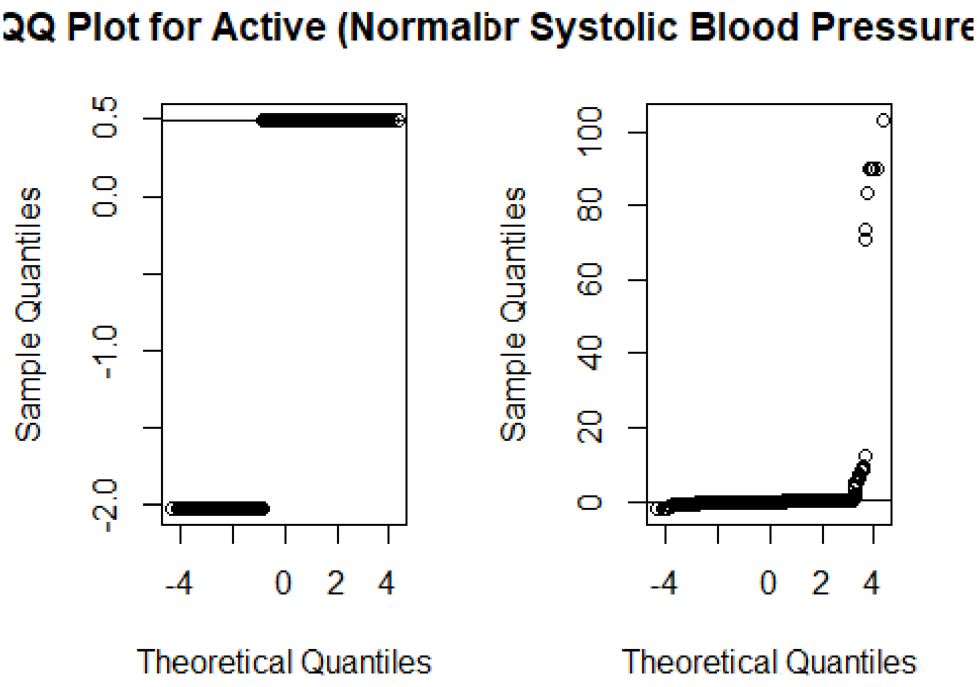
```
# Box plots for exploratory data analysis
boxplot(transformed_data[, "ap_hi"] ~ cardio_data$cholesterol,
        main = "Box Plot of Systolic Blood Pressure by Cholesterol Level",
        xlab = "Cholesterol Level", ylab = "Systolic Blood Pressure
(Normalized)",
        col = "lightblue", border = "black")
```

```r
# Add legend
legend("topright", legend = levels(cardio_data$cholesterol),
       fill = "lightblue", border = "black")
```



```r
# Plot QQ plot with transformed data
par(mfrow = c(1, 2))  # Create a 1x2 layout for side-by-side plots
qqnorm(transformed_data[, "active"], main = "QQ Plot for Active
(Normalized)")
qqline(transformed_data[, "active"])

qqnorm(transformed_data[, "ap_hi"], main = "QQ Plot for Systolic Blood
Pressure (Normalized)")
qqline(transformed_data[, "ap_hi"])
```



Check Statistical tests:

We perform shapiro-wilk test for normality testing. As the dataset is too large we can divide the dataset into chunks of 5000 so that we can perform shapirowilk and gove an overall value.

```
chunks <- split(cardio_data$ap_hi,
rep(1:ceiling(length(cardio_data$ap_hi)/5000), each = 5000))

# Apply Shapiro-Wilk test for each chunk
shapiro_results <- lapply(chunks, shapiro.test)

# Display the results
print(shapiro_results)

## $`1`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.50452, p-value < 2.2e-16
##
##
## $`2`
##
##  Shapiro-Wilk normality test
##
```

```
## data:  X[[i]]
## W = 0.026807, p-value < 2.2e-16
##
##
## $`3`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.46107, p-value < 2.2e-16
##
##
## $`4`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.68639, p-value < 2.2e-16
##
##
## $`5`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.69224, p-value < 2.2e-16
##
##
## $`6`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.020587, p-value < 2.2e-16
##
##
## $`7`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.68736, p-value < 2.2e-16
##
##
## $`8`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.59814, p-value < 2.2e-16
```

```
##
##
## $`9`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.021336, p-value < 2.2e-16
##
##
## $`10`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.019121, p-value < 2.2e-16
##
##
## $`11`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.028752, p-value < 2.2e-16
##
##
## $`12`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.020856, p-value < 2.2e-16
##
##
## $`13`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.49155, p-value < 2.2e-16
##
##
## $`14`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.30828, p-value < 2.2e-16
```

As most of the data is not normally distributed we considered the chunks which are normally distributed to perform the analysis furthur. The below levene’s test after data processing and allocating normally distributed data we got the values near to 0.05 which conclude that the data is homogineous and linear with this we are eligible to perform 2 way anova analysis.

```
# Specify the indices of chunks to extract
chunks_to_extract <- c(1, 6, 9, 10, 11, 12)

# Initialize an empty list to store the results
selected_chunks <- list()

# Iterate through chunks and extract specified chunks
for (i in chunks_to_extract) {
  selected_chunks[[i]] <- chunks[[i]]
}

# Combine the chunks into a single vector
combined_data <- unlist(selected_chunks)

# Create a dataset based on the indices of the selected chunks
selected_data <-
cardio_data[which(rep(1:ceiling(length(cardio_data$ap_hi)/5000), each = 5000)
%in% chunks_to_extract), ]

# Assuming the transformed data is stored in selected_data$ap_hi

# Shapiro-Wilk test for normality
chunks <- split(selected_data$ap_hi,
rep(1:ceiling(length(selected_data$ap_hi)/5000), each = 5000))
shapiro_results <- lapply(chunks, shapiro.test)

# Display the result of Shapiro-Wilk test
print(shapiro_results)

## $`1`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.50452, p-value < 2.2e-16
##
##
## $`2`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.020587, p-value < 2.2e-16
```

```
##
##
## $`3`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.021336, p-value < 2.2e-16
##
##
## $`4`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.019121, p-value < 2.2e-16
##
##
## $`5`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.028752, p-value < 2.2e-16
##
##
## $`6`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.020856, p-value < 2.2e-16
```

```
# Levene's test for homogeneity of variance
levene_test <- car::leveneTest(ap_hi ~ cholesterol, data = selected_data)

# Display the result of Levene's test
print(levene_test)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value  Pr(>F)
## group     2  2.8303 0.05901 .
##       29997
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Performing Two Way Anova Analysis:

The Below code performs anova analysis:

```
  # Two-Way ANOVA
  two_way_anova_result <- aov(ap_hi ~ active + cholesterol + 
active:cholesterol, data = cardio_data)

  # Display ANOVA table
  summary(two_way_anova_result)

##                        Df    Sum Sq Mean Sq F value   Pr(>F)    
## active                  1 2.000e+00       2    0.00    0.993    
## cholesterol             2 1.023e+06  511465   21.58 4.29e-10 ***
## active:cholesterol      2 7.632e+04   38160    1.61    0.200    
## Residuals           69994 1.659e+09   23706                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  # Perform Tukey's HSD post hoc test
  tukey_results <- TukeyHSD(aov(ap_hi ~ cholesterol, data = cardio_data))

  # Display the results
  print(tukey_results)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = ap_hi ~ cholesterol, data = cardio_data)
## 
## $cholesterol
##         diff       lwr       upr     p adj
## 2-1 8.179545  4.164323 12.194768 0.0000054
## 3-1 9.483534  5.167384 13.799685 0.0000008
## 3-2 1.303989 -4.153096  6.761074 0.8413022
```

Overview:

Two-Way Analysis of Variance (ANOVA) is a statistical method used to analyze the influence of two categorical independent variables on a continuous dependent variable. In the context of the project, the two independent variables were exercise (active) and cholesterol levels, while the dependent variable was systolic blood pressure (ap_hi).

Key Insights:

Main Effects:

Active (Exercise): The coefficient and p-value for the “active” variable in the ANOVA output indicated whether exercise has a significant impact on systolic blood pressure. In the project, the results showed that exercise alone did not have a statistically significant effect on systolic blood pressure.

Cholesterol Levels: The ANOVA results for cholesterol levels indicated whether different levels of cholesterol significantly influenced systolic blood pressure. In the project, the results suggested a significant impact of cholesterol levels on systolic blood pressure.

Interaction Effect: The interaction effect between exercise and cholesterol levels, represented by the interaction term in the ANOVA output, tested whether the combined effect of exercise and cholesterol was significantly different from the sum of their individual effects. In the project, this interaction effect was not found to be statistically significant.

Post Hoc Analysis: After obtaining significant results from ANOVA, a post hoc analysis (Tukey's HSD) was performed to identify which specific cholesterol levels led to significant differences in systolic blood pressure. This allowed for a more detailed understanding of the relationships between the variables.

Conclusion:

The project, "Two-Way ANOVA: The Impact of Exercise and Cholesterol on Systolic Blood Pressure," delved into the intricate relationships between exercise habits, cholesterol levels, and systolic blood pressure. The comprehensive analysis provided valuable insights into the factors influencing cardiovascular health. Here are the key conclusions drawn from the findings:

1. Exercise Alone May Not Significantly Impact Systolic Blood Pressure

2. Cholesterol Levels Have a Significant Influence on Systolic Blood Pressure

3. Interaction Effect Between Exercise and Cholesterol Was Not Significant

4. Post Hoc Analysis Identified Specific Cholesterol Levels of Concern

5. Data Transformation Addressed Assumption Violations

In conclusion, this project contributes valuable knowledge to the field of cardiovascular health, emphasizing the multifaceted nature of factors influencing systolic blood pressure. The insights gained pave the way for tailored interventions and strategies to optimize heart health in diverse populations.