Assignment3_Jittin

Jittin Pomprakay

2024-11-10

Introduction

This statistical analysis aims to examine the association between ADP-induced platelet aggregation levels and Clopidogrel resistance with three single nucleotide polymorphisms (SNPs): rs4244285 (CYP2C192), rs4986893 (CYP2C193), and rs662 (PON1.192Q>R). The analysis also considers two potential confounding factors, age and sex, which are included in the models to adjust for their influence. Linear regression is used to test the association between ADP (as a continuous outcome) and the SNPs.

The dataset provided in PlateletHW.tsv includes 11 variables: IID, ADP, Resistance, rs4244285, rs4986893, rs662, AGE, SEX, PON1.192Q>R, CYP2C192, and CYP2C193. Among these variables, ADP is continuous, while the rest are categorical. The coding for the SNPs is as follows: rs4244285 (CYP2C192) is coded as 0 = GG, 1 = AG, 2 = AA, while both rs4986893 (CYP2C193) and rs662 (PON1.192Q>R) are coded as 0 = AA, 1 = AG, 2 = GG. For the sex variable, 0 = male and 1 = female, and for drug resistance, 0 = not resistant and 1 = resistant.

Import Data

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.2

## Warning: package 'ggplot2' was built under R version 4.4.2

## Warning: package 'dplyr' was built under R version 4.4.2

## Warning: package 'lubridate' was built under R version 4.4.2

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(dplyr)
library(ggplot2)
platelet_data <- read_delim("RawData/PlateletHW .tsv", delim = "\t", escape_double = FALSE, trim_ws = TR
```

```
## Rows: 211 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (3): PON1.192Q>R, CYP2C19*2, CYP2C19*3
## dbl (8): IID, ADP, Resistance, rs4244285, rs4986893, rs662, AGE, SEX
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Clean Data

Initially, I decided to clean the data due to negative values in the ADP column, likely caused by systematic errors. Since platelet aggregation represents a physical or biochemical process, the measured value cannot logically be negative. A negative value would imply a reversal of aggregation, which is not biologically feasible under normal conditions. To reduce the risk of misinterpreting true errors, I filtered the data to include only values greater than zero.

```r
platelet_data_clean <- platelet_data %>%
  filter(ADP >= 0)
```

IQR method

After eliminating the negative values, I removed the outliers from the data using the IQR method.

```r
total_rows <- nrow(platelet_data_clean)
Q1 <- quantile(platelet_data_clean$ADP, 0.25, na.rm = TRUE)
Q3 <- quantile(platelet_data_clean$ADP, 0.75, na.rm = TRUE)
IQR_value <- Q3 - Q1
lower_bound <- Q1 - 1.5 * IQR_value
upper_bound <- Q3 + 1.5 * IQR_value
platelet_data_clean_filtered <- platelet_data_clean %>%
filter(ADP >= lower_bound & ADP <= upper_bound)

num_outliers <- total_rows - nrow(platelet_data_clean_filtered)

cat("Number of outliers by IQR method:", num_outliers, "\n")
```

```
## Number of outliers by IQR method: 0
```

```r
summary(platelet_data_clean$ADP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.60   15.68   27.52   41.99   75.26  103.05
```

Before and After Clean Data Box Plot

“‘{r}, echo=TRUE, results=“show”} platelet_data$Status <- “Before”platelet_data_cleanStatus <- “After”

combined_data <- rbind(platelet_data %>% select(Status, ADP), platelet_data_clean %>% select(Status, ADP))

combined_data$Status <- factor(combined_data$Status, levels = c(“Before”, “After”))

ggplot(combined_data, aes(x = Status, y = ADP, fill = Status)) + geom_boxplot() + labs(title = “Box Plot: Before and After Cleaning”, x = “Data Status”, y = “ADP Values”) + theme_minimal() + scale_fill_manual(values = c(“Before” = “red”, “After” = “blue”))

```
‘‘‘ r
#Write a new clean data
write_tsv(platelet_data_clean, "CleanData/PlateletHW_clean.tsv")
```

The box plot compares ADP values before and after data cleaning, showing minimal changes in the distribution, median, and range. This indicates that the cleaning process effectively removed outliers without significantly altering the central tendency or variability of the data set, preserving its overall integrity.

Linear Equation

Before testing the linear regression, the ADP values should follow a normal distribution. For this study, I decided to take the logarithm of the cleaned ADP values to achieve normalization.

```
platelet_data_clean$ADP_log<-log(platelet_data_clean$ADP)
```

Then, I test the regression with three SNPs individually by taking the log ADP. I also plot the Q-Q plot to visualize the data distribution.

Single Linear Regression

rs4244285 vs ADP_log

“‘{r}, echo=TRUE, results=“show”} liner_logA <- lm(ADP_log ~ rs4244285, data = platelet_data_clean)

qqnorm(liner_logA$residuals)qqline(liner_logA$residuals,col=“blue”)

summary(liner_logA)

```
### rs4986893 vs ADP_log

‘‘‘{r}, echo=TRUE, results="show"}
liner_logB <- lm(ADP_log ~ rs4986893, data = platelet_data_clean)

qqnorm(liner_logB$residuals)
qqline(liner_logB$residuals,col="red")

summary(liner_logB)
```

rs662 vs ADP_log

“‘{r}, echo=TRUE, results=“show”} linear_logC <- lm(ADP_log ~ rs662, data = platelet_data_clean)

qqnorm(linear_logC$residuals)qqline(linear_logC$residuals,col=“green”)

summary(linear_logC)

```
### Multiple Linear Regression
```

```
‘‘‘{r}, echo=TRUE, results="show"}}
linear_logABC <- lm(ADP_log ~ rs4244285 + rs4986893 + rs662, data = platelet_data_clean)

qqnorm(linear_logABC$residuals)
qqline(linear_logABC$residuals,col="yellow")

summary(linear_logABC)
```

Confounding Factor

“‘{r}, echo=TRUE, results=“show”}} linear_CF <- lm(ADP_log ~ rs4244285 + rs4986893 + rs662 + SEX + AGE, data = platelet_data_clean) summary(adjusted_model)

qqnorm(linear_CF$residuals)qqline(linear_CF$residuals,col=“purple”)

summary(linear_CF)

“‘

Summary

The linear regression model evaluated the relationship between `ADP_log` and the predictors `rs4244285`, `rs4986893`, `rs662`, SEX, and AGE, while accounting for the potential confounding effects of SEX and AGE. Among the predictors, `rs4244285` (Estimate = 0.355223, $p < 0.001$) and `rs4986893` (Estimate = 0.595210, $p < 0.01$) showed statistically significant positive associations with `ADP_log`, indicating these SNPs play an important role in ADP-induced platelet aggregation. In contrast, `rs662` ($p = 0.78448$) was not significant, suggesting no meaningful contribution. Although SEX ($p = 0.57625$) and AGE ($p = 0.28820$) were not significant predictors of `ADP_log` themselves, their inclusion in the model slightly adjusted the coefficients of `rs4244285` and `rs4986893`, supporting their role as potential confounders. The model explained 11.27% of the variation in `ADP_log` ($R^2 = 0.1127$) with an adjusted R^2 of 0.09053, indicating a modest fit. These findings highlight the importance of considering SEX and AGE as potential confounders when analyzing the effects of SNPs on platelet aggregation.