HW3_summary

Nattha Bunyamani

2024-11-10

According to the “PlateletHW.tsv” dataset, which contains 211 observed records with 11 variables. There are two continuous variables: ADP-induced platelet aggregation level (ADP), which measures the level of platelet response, and Age, measured in years. There are 5 categorical variables: Clopidogrel resistance, marked as “1” for those resistant to drug and “0” for those not resistant to drug. Sex is coded as “0” for male and “1” for female. Also, the genetic data for three SNPs: rs4244285 (CYP2C19*2), rs4986893 (CYP2C19*3), and rs662 (PON1 192Q>R). Each genotype is coded as 0, 1, or 2 depending on the number of minor alleles exactly for each SNP:

- For rs4244285 (CYP2C19*2): 0 represents “GG”, 1 represents “AG”, and 2 represents “AA”.
- For rs4986893 (CYP2C19*3): 0 represents “AA”, and 1 represents “AG”.
- For rs662 (PON1 192Q>R): 0 represents “AA”, 1 represents “AG”, and 2 represents “GG”.

To examine the association between the three SNPs and the level of ADP-induced platelet aggregation, we can divide the analysis into two main steps: data cleaning by handling the outliers and test of association

1. Data cleaning

In this step, starting with loading some necessary packages, setting the location of the current working directory, and then loading the raw data. After that, use the summary function to see the statistics of ADP values in the raw data. As you can see in the summery, that the minimum of the value is -8.721, which is negative value and cannot thus be used for analysis due to the fact that ADP is measured as a percentage increase in platelet aggregation compared to a baseline level, so the appearance of a negative value would be abnormal.

```
library(tidyverse)
```

```
## -- 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
setwd("../HW3/")

data <- read.table("raw_data/PlateletHW.tsv", header = TRUE, sep = "\t")
```

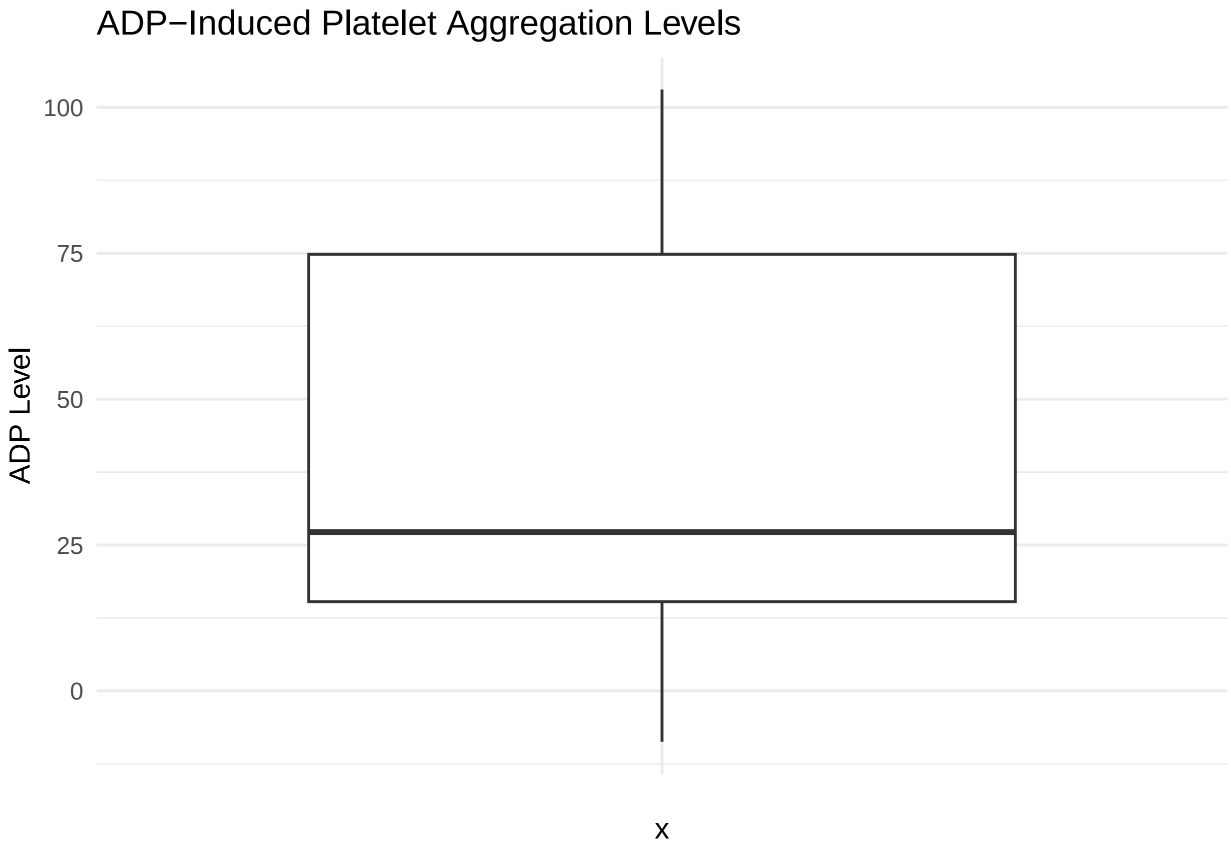
```r
summary(data$ADP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  -8.721  15.281  27.200  40.853  74.810 103.053
```

To ensure the analysis is reliable and to avoid drawing a fault conclusion due to using abnormal values for analysis. We filtered outliers and negative values from the data. Boxplot by ggplot function has been used here to visualize the distribution of ADP levels, helping us identify any possible outliers. We used quartiles to filter outliers because they capture the central interval range of the data while excluding extreme values.

As a result, we found out 5 outliers, which are all negative values, and removed them from the dataset. After filtering, we confirmed with the summary function that the minimum value is now 1.60, meaning all values are within the defined range and with no more negative values left. The outliers-free cleaned dataset was saved as "PlateletHW_cleaned.tsv" in the clean_data directory.

```r
ggplot(data, aes(x = "", y = ADP)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 16, outlier.size = 5) +
  labs(title = "ADP-Induced Platelet Aggregation Levels", y = "ADP Level") +
  theme_minimal()
```



```r
Q1 <- quantile(data$ADP, 0.25)
Q3 <- quantile(data$ADP, 0.75)
IQR <- Q3 - Q1
lower_outlier <- Q1 - 1.5 * IQR
upper_outlier <- Q3 + 1.5 * IQR

cat("Lower Outlier Bound:", lower_outlier, "\n")
```

```
## Lower Outlier Bound: -74.01193
```

```r
cat("Upper Outlier Bound:", upper_outlier, "\n")
```

```
## Upper Outlier Bound: 164.1025
```

```r
outliers <- data %>%
  filter(ADP < lower_outlier | ADP > upper_outlier | ADP < 0)

print(outliers)
```

```
##   IID        ADP Resistance rs4244285 rs4986893 rs662 AGE SEX PON1.192Q.R
## 1 101 -7.8443945          0         1         0     0  63   0         A A
## 2 147 -0.7848469          0         0         0     1  68   0         A G
## 3 166 -4.1416578          0         1         0     2  77   0         G G
## 4 190 -8.7208368          0         1         0     1  77   1         A G
## 5 197 -8.6706725          0         0         0     2  54   0         G G
##   CYP2C19.2 CYP2C19.3
## 1       A G       A A
## 2       G G       A A
## 3       A G       A A
## 4       A G       A A
## 5       G G       A A
```

```r
cleaned_data <- data %>%
  filter(ADP >= 0 & ADP >= lower_outlier & ADP <= upper_outlier) %>%
  drop_na()

summary(cleaned_data$ADP)
```

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

```r
write.table(cleaned_data, "clean_data/PlateletHW_cleaned.tsv", sep = "\t", row.names = FALSE)
```

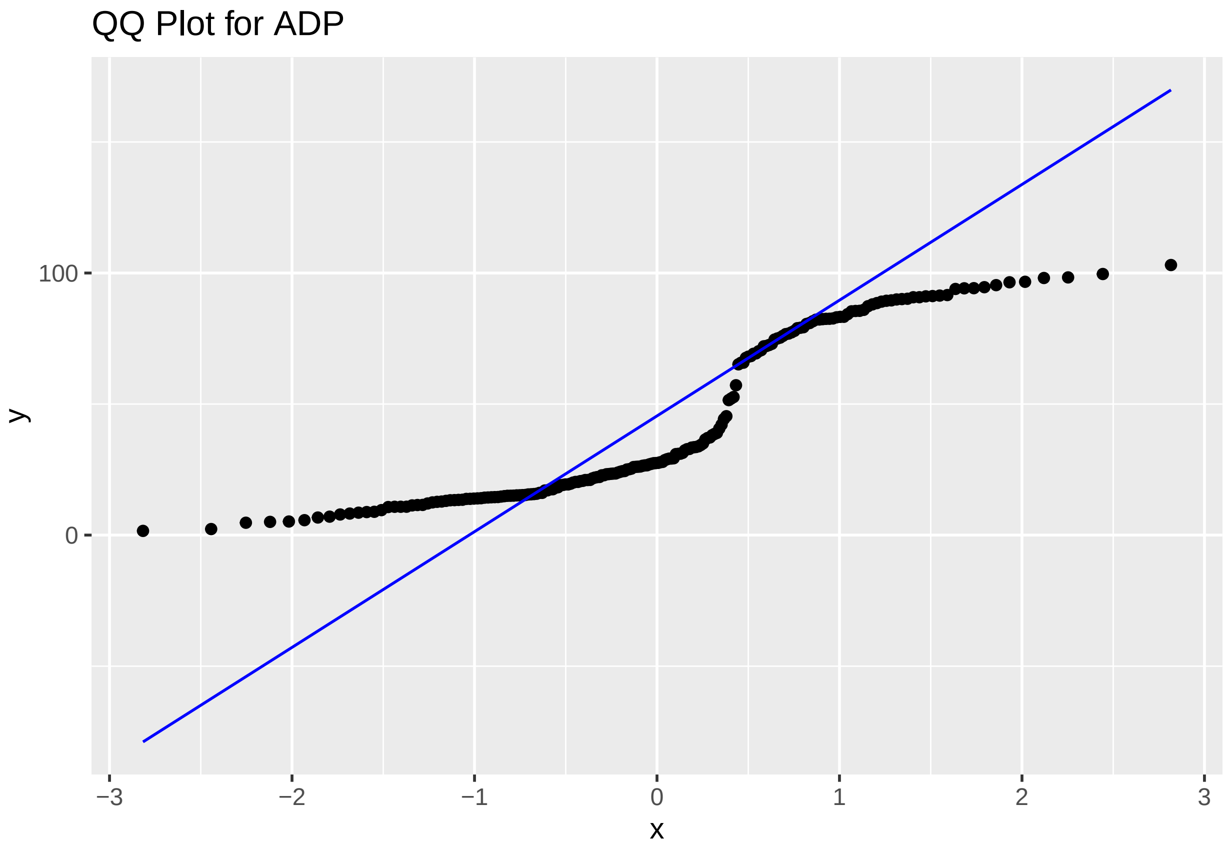
2. Test of Association

To investigate the association between the three SNPs and ADP-induced platelet aggregation levels, start by loading the cleaned dataset and assigning it to the cleaned_data variable. Before linear regression, it's important to check whether the ADP values follow a normal distribution using a QQ plot. The first QQ plot shows that the values of ADP are not normally distributed, with the deviations from the theoretical line in the tails. This kind of skewness in data may affect the assumptions of linear regression and furthermore impact the reliability of results.

To resolve this problem, we apply a log transformation to the ADP values to reduce skewness and making the distribution closer to normal. After transformation, the second QQ plot shows an improved alignment with the theoretical line in the middle range of the data, although there is still some deviation at the tails. The log transformation doesn't fully normalize the data; however, it does reduce skewness and makes the distribution more suitable for regression analysis.

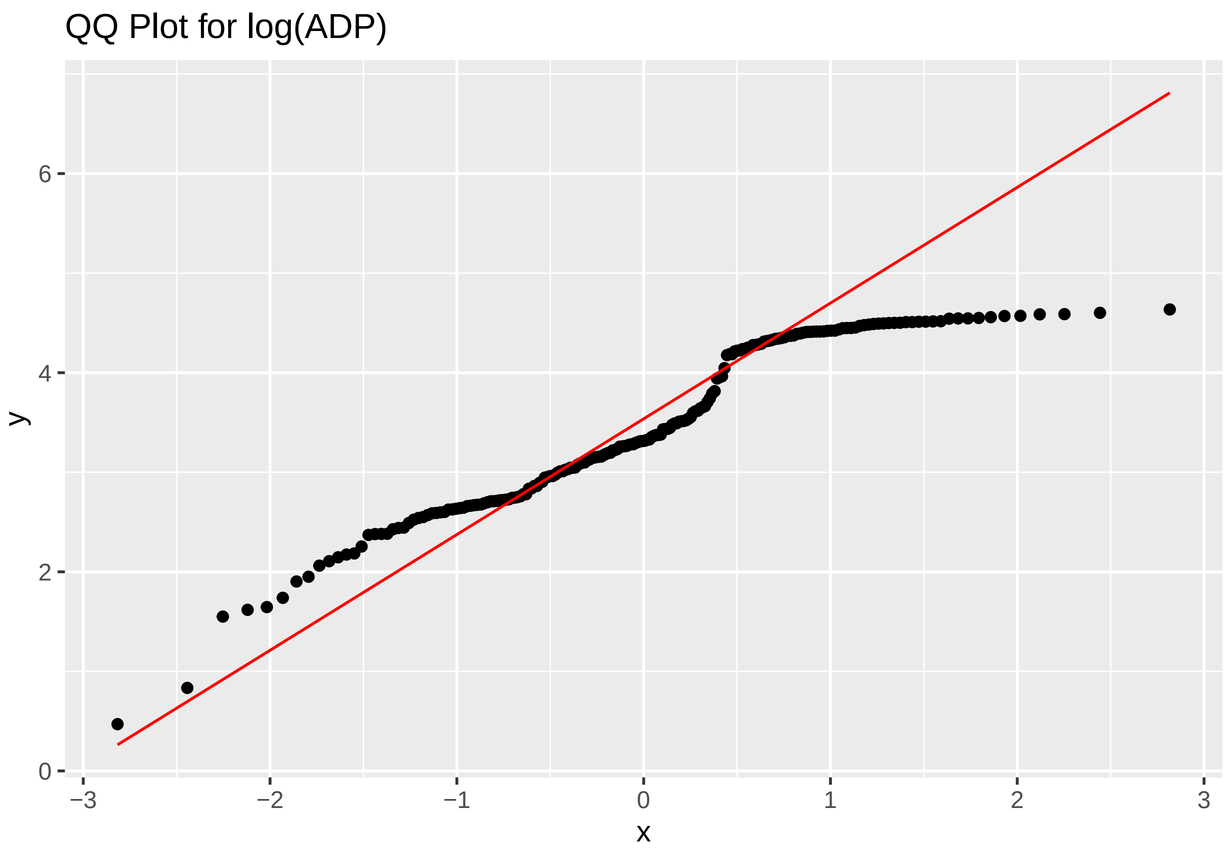
```r
cleaned_data <- read.table("clean_data/PlateletHW_cleaned.tsv", header = TRUE, sep = "\t")
```

```r
ggplot(cleaned_data, aes(sample = ADP)) +
  stat_qq() +
  stat_qq_line(color = "blue") +
  ggtitle("QQ Plot for ADP")
```



```r
cleaned_data$log_ADP <- log(cleaned_data$ADP)

ggplot(cleaned_data, aes(sample = log_ADP)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  ggtitle("QQ Plot for log(ADP)")
```



To determine the relationship between each SNP and log-transformed ADP levels independently, we used single linear regression models and examined the p-values of each model. These results show that both rs4244285 and rs4986893 significantly impact the levels of ADP by p-values 0.000111 and 0.00793, respectively. In contrast, rs662 does not show a statistically significant impact with ADP levels since its p-value is 0.784.

We can use boxplots to show the distribution of log-transformed ADP levels across genotypes for each SNP for further demonstration. For rs4244285 and rs4986893, boxplots indicated an increasing trend across the genotypes in both cases, which support the result of statistically significant associations from the regression models, while the trend for rs662 was flat across genotypes, supporting the result of non-statistical significance observed for this SNP in the regression analysis.

```
#Single Linear Regression for each SNP
model_rs4244285 <- lm(log_ADP ~ rs4244285, data = cleaned_data)
summary(model_rs4244285)
```

```
##
## Call:
## lm(formula = log_ADP ~ rs4244285, data = cleaned_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.76622 -0.56494 -0.02906  0.77925  1.36542
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.23596    0.07456  43.399  < 2e-16 ***
## rs4244285    0.35518    0.09013   3.941 0.000111 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.8121 on 204 degrees of freedom
## Multiple R-squared:  0.07075,    Adjusted R-squared:  0.06619
## F-statistic: 15.53 on 1 and 204 DF,  p-value: 0.0001115
```

```
model_rs4986893 <- lm(log_ADP ~ rs4986893, data = cleaned_data)
summary(model_rs4986893)
```

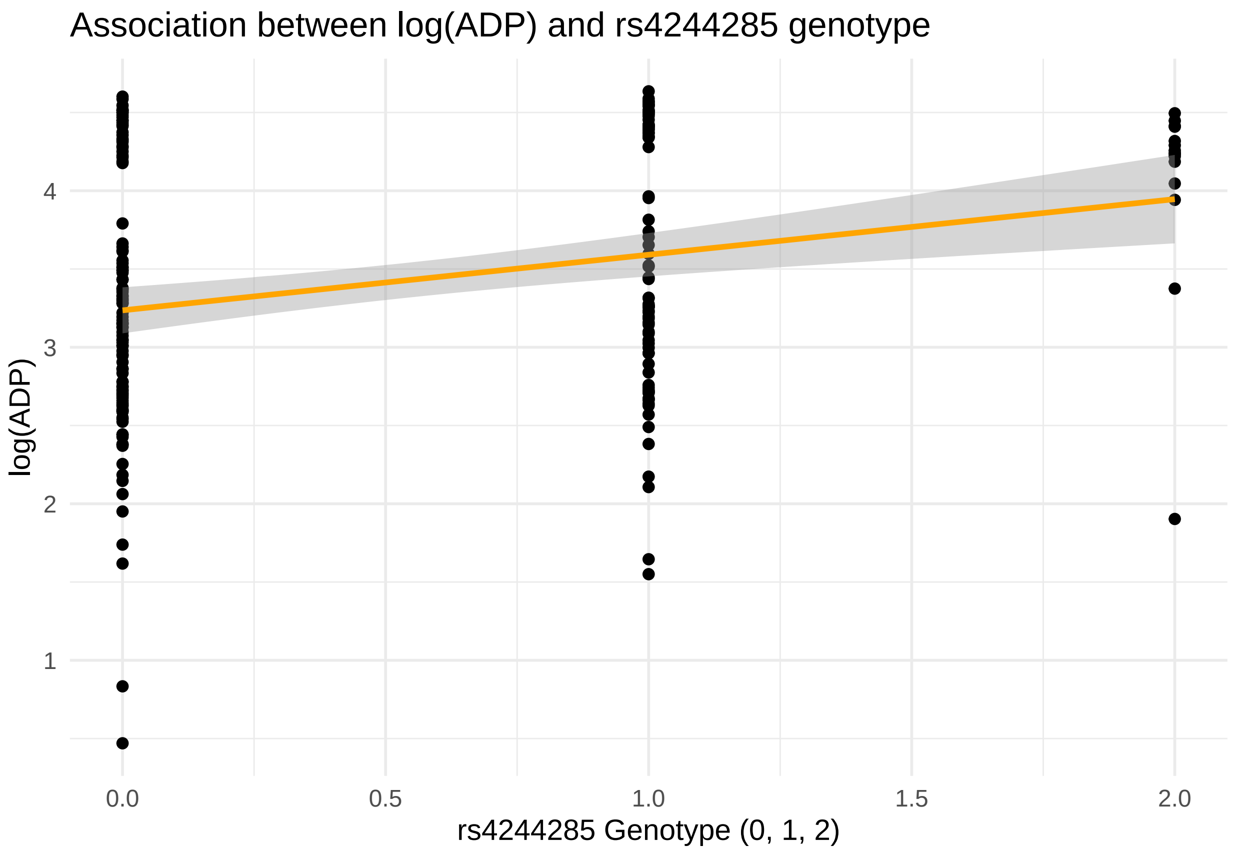
```
##
## Call:
## lm(formula = log_ADP ~ rs4986893, data = cleaned_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.91583 -0.65367 -0.07058  0.84795  1.24968
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.38557    0.05975  56.660  < 2e-16 ***
## rs4986893    0.61465    0.22921   2.682  0.00793 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.828 on 204 degrees of freedom
## Multiple R-squared:  0.03405,    Adjusted R-squared:  0.02932
## F-statistic: 7.191 on 1 and 204 DF,  p-value: 0.007926
```

```
model_rs662 <- lm(log_ADP ~ rs662, data = cleaned_data)
summary(model_rs662)
```

```
##
## Call:
## lm(formula = log_ADP ~ rs662, data = cleaned_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9480 -0.6731 -0.1027  0.9032  1.1934
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.39356    0.13646  24.869   <2e-16 ***
## rs662        0.02416    0.08812   0.274    0.784
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.8423 on 204 degrees of freedom
## Multiple R-squared:  0.0003684,  Adjusted R-squared:  -0.004532
## F-statistic: 0.07518 on 1 and 204 DF,  p-value: 0.7842
```

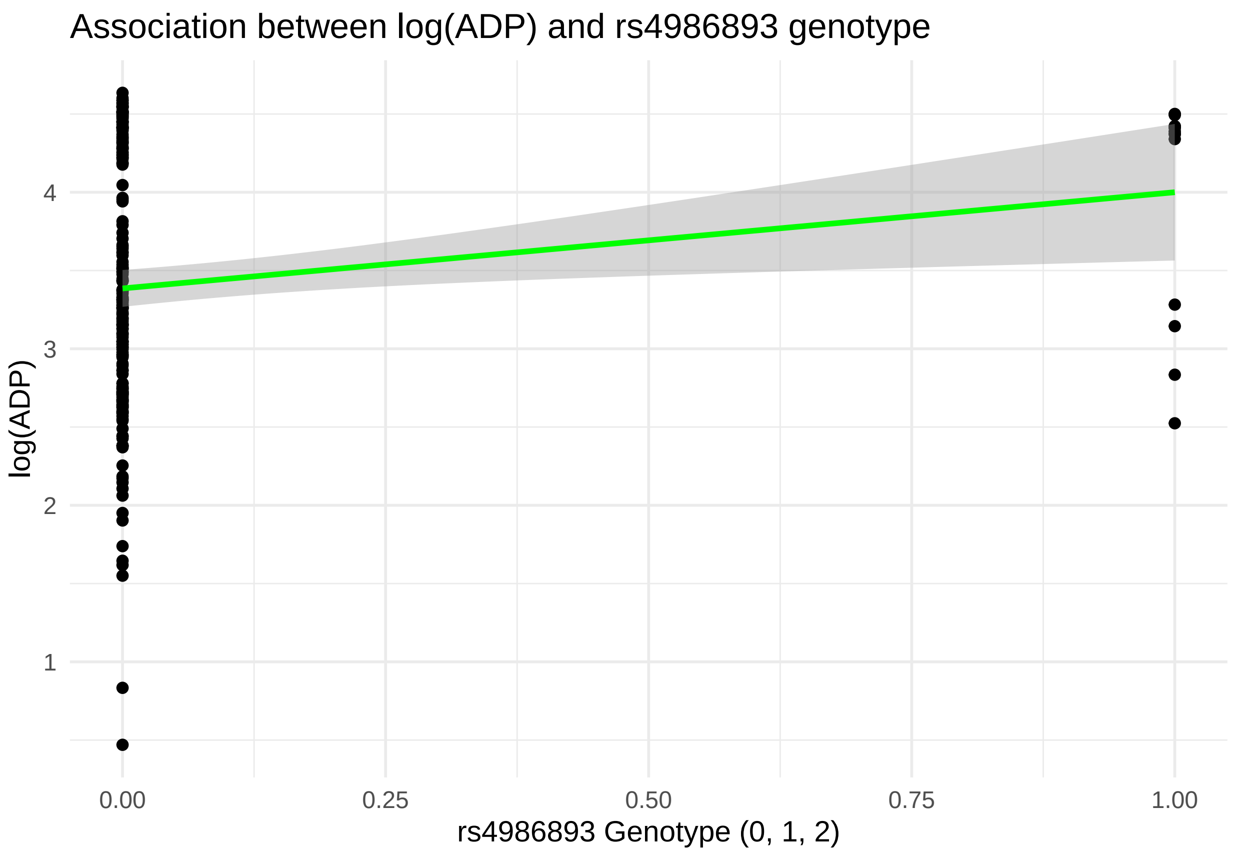
```
ggplot(cleaned_data, aes(x = rs4244285, y = log_ADP)) +
  geom_point() +
  geom_smooth(method = "lm", color = "orange") +
  labs(title = "Association between log(ADP) and rs4244285 genotype",
       x = "rs4244285 Genotype (0, 1, 2)",
       y = "log(ADP)") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



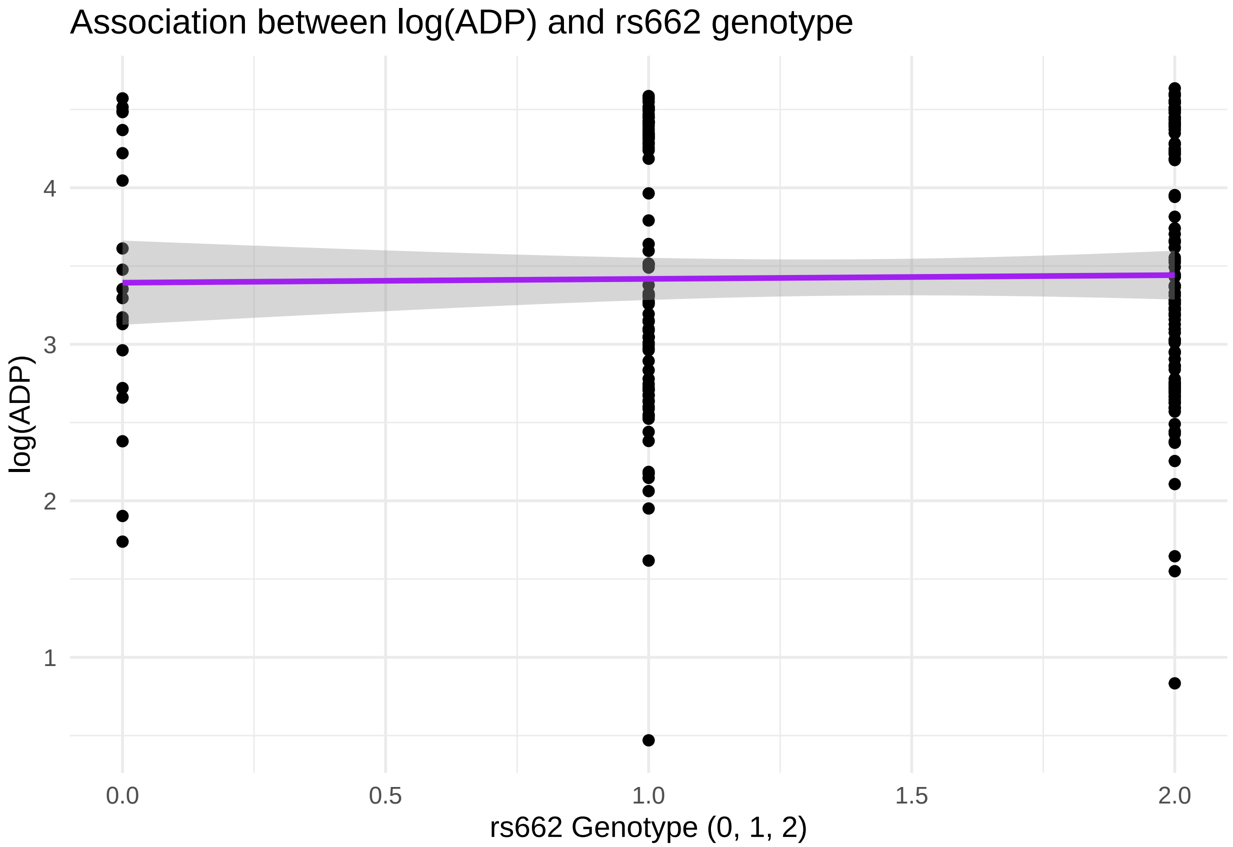
```
ggplot(cleaned_data, aes(x = rs4986893, y = log_ADP)) +
  geom_point() +
  geom_smooth(method = "lm", color = "green") +
  labs(title = "Association between log(ADP) and rs4986893 genotype",
       x = "rs4986893 Genotype (0, 1, 2)",
       y = "log(ADP)") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
ggplot(cleaned_data, aes(x = rs662, y = log_ADP)) +
  geom_point() +
  geom_smooth(method = "lm", color = "purple") +
  labs(title = "Association between log(ADP) and rs662 genotype",
       x = "rs662 Genotype (0, 1, 2)",
       y = "log(ADP)") +
  theme_minimal()
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



Even though both rs4244285 and rs4986893 have independent effects on ADP level, these independent effects may be interdependent due to the interaction between these SNPs. Using multiple linear regression allows us to analyze the influence of all three SNPs together on the ADP level and possible confounding between these SNPs. When including all three SNPs as predictor variables, the results are the same with the single linear regressions, which are both rs4244285 and rs4986893 are significantly related to ADP levels, while rs662 is not.

Further investigating by including more covariates of resistance, AGE, and SEX to ensure comprehensively determining all the effects of predictors. In this model, resistance was the only variable that showed statistical significance to ADP levels, while both rs4244285 and rs4986893 became nonsignificant. These results may indicate that resistance is a confounding variable, potentially acting to mediate the relationship of those SNPs with ADP levels.

To further explore this, logistic regression of the SNPs for resistance is considered. As a result, rs4244285 and rs4986893 are significantly associated with resistance. Thus, resistance is a mediator of the association between the ADP levels and these SNPs.

```
# Multiple linear regression including covariates
model_SNPs <- lm(log_ADP ~ rs4244285 + rs4986893 + rs662, data = cleaned_data)
summary(model_SNPs)
```

```
##
## Call:
## lm(formula = log_ADP ~ rs4244285 + rs4986893 + rs662, data = cleaned_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7184 -0.5387  0.0058  0.5767  1.3979
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.16603    0.13910  22.760  < 2e-16 ***
## rs4244285    0.35185    0.08896   3.955 0.000106 ***
## rs4986893    0.60163    0.22191   2.711 0.007283 **
## rs662        0.02206    0.08386   0.263 0.792779
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.8015 on 202 degrees of freedom
## Multiple R-squared:  0.1037, Adjusted R-squared:  0.09044
## F-statistic: 7.794 on 3 and 202 DF,  p-value: 5.983e-05
```

```
model_all_variables <- lm(log_ADP ~ rs4244285 + rs4986893 + rs662 + Resistance + AGE + SEX, data = clear
summary(model_all_variables)
```

```
##
## Call:
## lm(formula = log_ADP ~ rs4244285 + rs4986893 + rs662 + Resistance +
##     AGE + SEX, data = cleaned_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.38101 -0.21312  0.03225  0.25753  1.06386
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.773046   0.212762  13.034   <2e-16 ***
## rs4244285    0.046214   0.054461   0.849    0.397
## rs4986893    0.017245   0.133583   0.129    0.897
## rs662        0.031788   0.049502   0.642    0.522
## Resistance   1.457051   0.074353  19.596   <2e-16 ***
## AGE          0.001697   0.003110   0.546    0.586
## SEX         -0.072888   0.073989  -0.985    0.326
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4694 on 199 degrees of freedom
## Multiple R-squared:  0.6971, Adjusted R-squared:  0.688
## F-statistic: 76.35 on 6 and 199 DF,  p-value: < 2.2e-16
```

```
#Logistic regression to examine the association between Resistance and three SNPs
model_resistance <- glm(Resistance ~ rs4244285 + rs4986893 + rs662, family = binomial, data = cleaned_da
summary(model_resistance)
```

```
##
## Call:
## glm(formula = Resistance ~ rs4244285 + rs4986893 + rs662, family = binomial,
##     data = cleaned_data)
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.30683    0.40211  -3.250  0.00115 **
## rs4244285    0.97688    0.24857   3.930  8.5e-05 ***
## rs4986893    1.78111    0.62900   2.832  0.00463 **
## rs662       -0.04735    0.23674  -0.200  0.84149
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 264.05  on 205  degrees of freedom
## Residual deviance: 238.70  on 202  degrees of freedom
## AIC: 246.7
##
## Number of Fisher Scoring iterations: 4
```

Summary

In summary, staring from the cleaning step, five outliers, which were all negative values, were found and removed. The ADP values were log-transformed in order to best fit for linear regression analysis. The single linear regression was conducted for testing the association of log-transformed ADP with each SNP, indicating that rs4244285 and rs4986893 are statistically significant but rs662 is not significantly associated with it. For further investigation, the multiple linear regression was used, which yielded similar results. However, with the addition of more covariates, the results were altered; resistance was the only variable that showed statistical significance, and both SNPs became non-significant. This was supported by logistic regression showing that resistance is associated with both SNPs. These results imply that resistance is a confounder of the relationship between SNPs and ADP levels, meaning that the association between the SNPs and ADP depends on the presence of resistance. Thus, it is likely that the direct effect of the SNPs on ADP levels is mediated through resistance.