# Assignment3

Phanisa Butsiri

2024-11-14

This R project is use to test for association between three SNPs like rs4244285 (CYP2C19*2),rs4986893 (CYP2C19*3*) and Rs662 (PON1. 192Q>R) with ADP-induced platelet aggregation level.

Use read_tsv fuction for read data PlateletHW that in tsv type and assign to 'data' variable. Use '%>%" for forward the data to use in next line. I found outliers data in ADP column that show in negative values, then I decided to change negative values by take absolute to values use mutate fuction to select data in ADP columns and take absolute. Save data frame after change values as clean_data.tsv and save to clean_data folder.

```
data <- read_tsv("raw_data/PlateletHW.tsv")
```

```
## 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.
```

```
df <- data %>%
  mutate(ADP_abs = abs(ADP))

write_tsv(df, "clean_data/clean_data.tsv")
```

Call new data that cleaned as clean_data variable. Take log to values to nomalize data, prepare for ploting in linear graph.

```
clean_data <- read_tsv("clean_data/clean_data.tsv")
```

```
## Rows: 211 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (3): PON1.192Q>R, CYP2C19*2, CYP2C19*3
## dbl (9): IID, ADP, Resistance, rs4244285, rs4986893, rs662, AGE, SEX, ADP_abs
##
## 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$ADP_log <- log(clean_data$ADP)
```

```
## Warning in log(clean_data$ADP): NaNs produced
```

Compare each SNPs with ADP(log) in clean_data data frame.

```
log_liner_A <- lm(ADP_log ~ rs4244285, data =clean_data)
log_liner_B <- lm(ADP_log ~ rs4986893, data =clean_data)
log_liner_C <- lm(ADP_log  ~ rs662, data =clean_data)
```

Call summary function to show data that will provide stat value for log_liner_A like Min, Med, Max,t-values.

```
summary(log_liner_A)
```

```
##
## Call:
## lm(formula = ADP_log ~ rs4244285, data = clean_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
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.07075,    Adjusted R-squared:  0.06619
## F-statistic: 15.53 on 1 and 204 DF,  p-value: 0.0001115
```

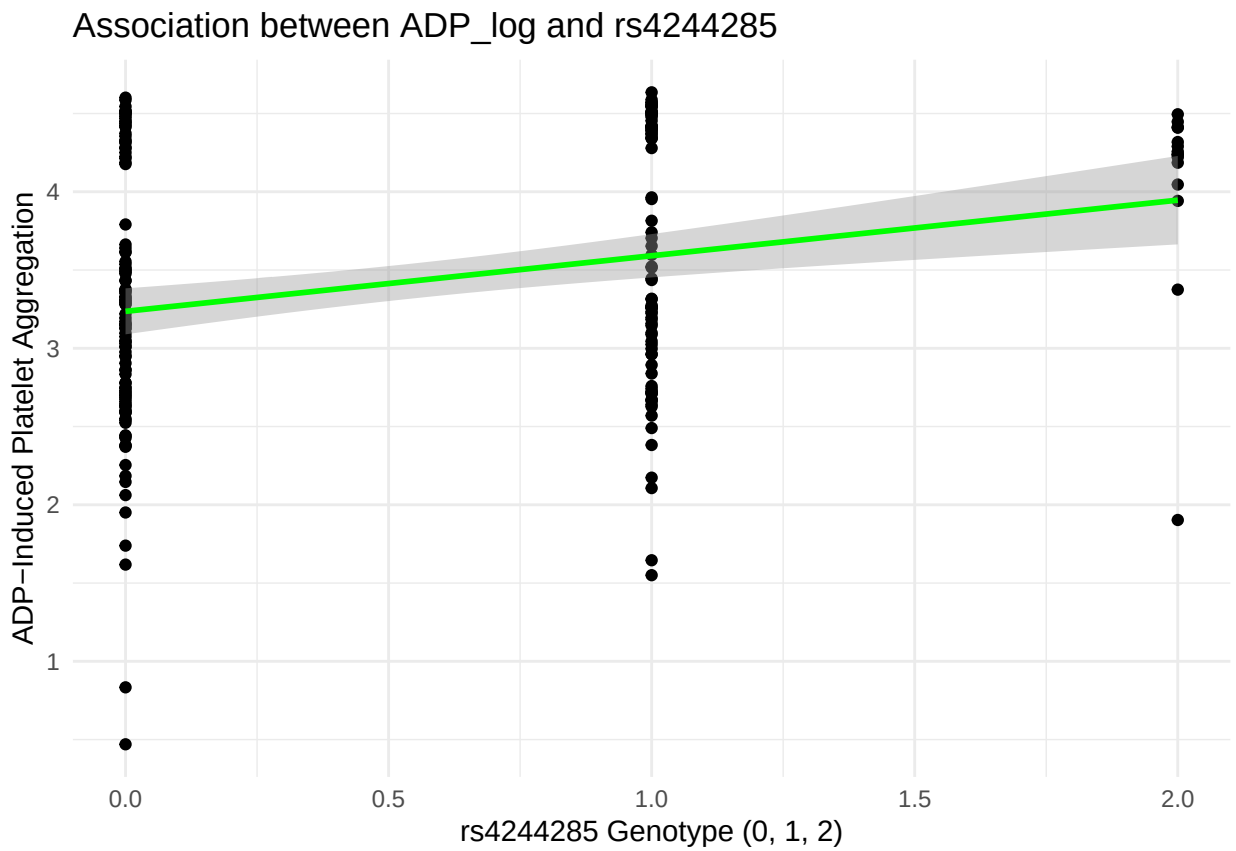
Use ggplot to show the linear regression of rs4244285 and ADP(log) with function lm and assign color as "green", x axis as re4244285 and y axis as ADP.

```
ggplot(clean_data, aes(x = rs4244285, y = ADP_log)) +
  geom_point() +
  geom_smooth(method = "lm", color = "green") +
  labs(title = "Association between ADP_log and rs4244285",
       x = "rs4244285 Genotype (0, 1, 2)",
       y = "ADP-Induced Platelet Aggregation")+theme_minimal()
```

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

```
## Warning: Removed 5 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

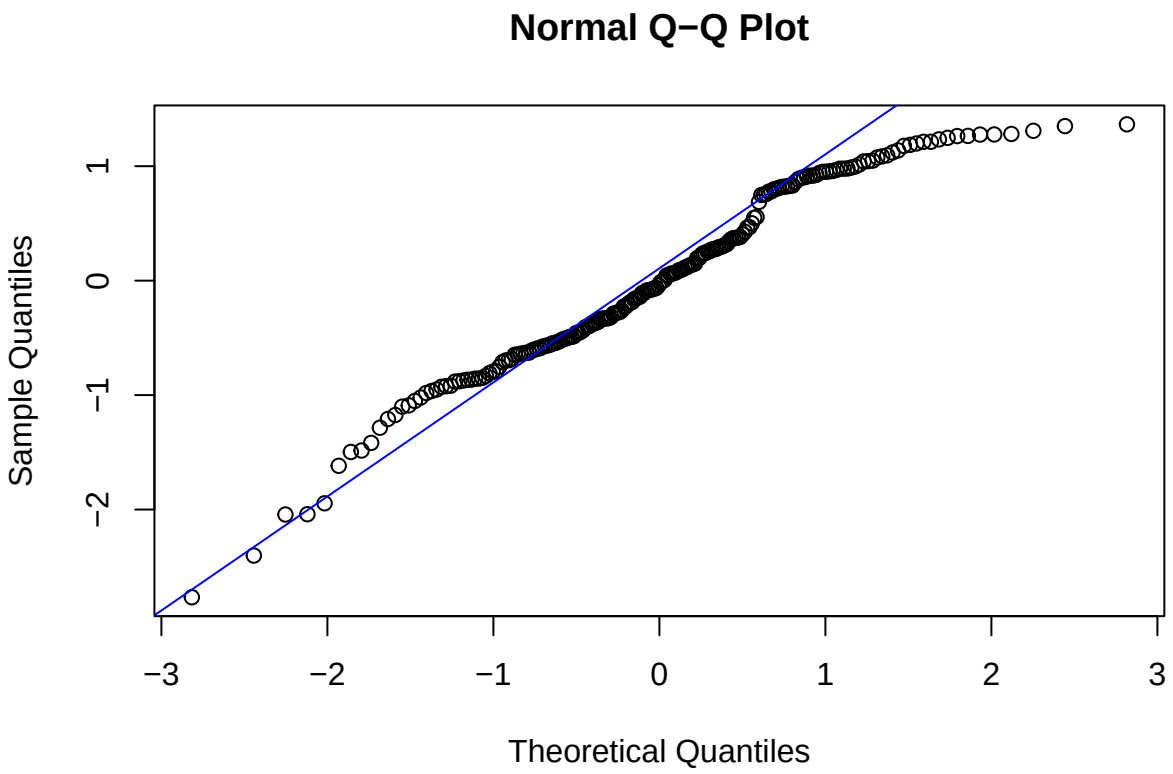
```
## Warning: Removed 5 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




From this graph, linear is going up, that mean 2 recessive gene, is the most induce platelet aggregation level.

Use qqnorm to plot the data to graph and qqline to plot line as blue color to graph.

```
qqnorm(log_liner_A$residuals)
qqline(log_liner_A$residuals, col = "blue")
```




Call summary function for log_liner_B

```
summary(log_liner_B)
```

```
##
## Call:
## lm(formula = ADP_log ~ rs4986893, data = clean_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
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.03405,    Adjusted R-squared:  0.02932
## F-statistic: 7.191 on 1 and 204 DF,  p-value: 0.007926
```

From p-value equal to 0.0001115. That less than 0.05, so we reject null Hypothesis. Then rs4244285 is significant to induce platelet aggregation level.

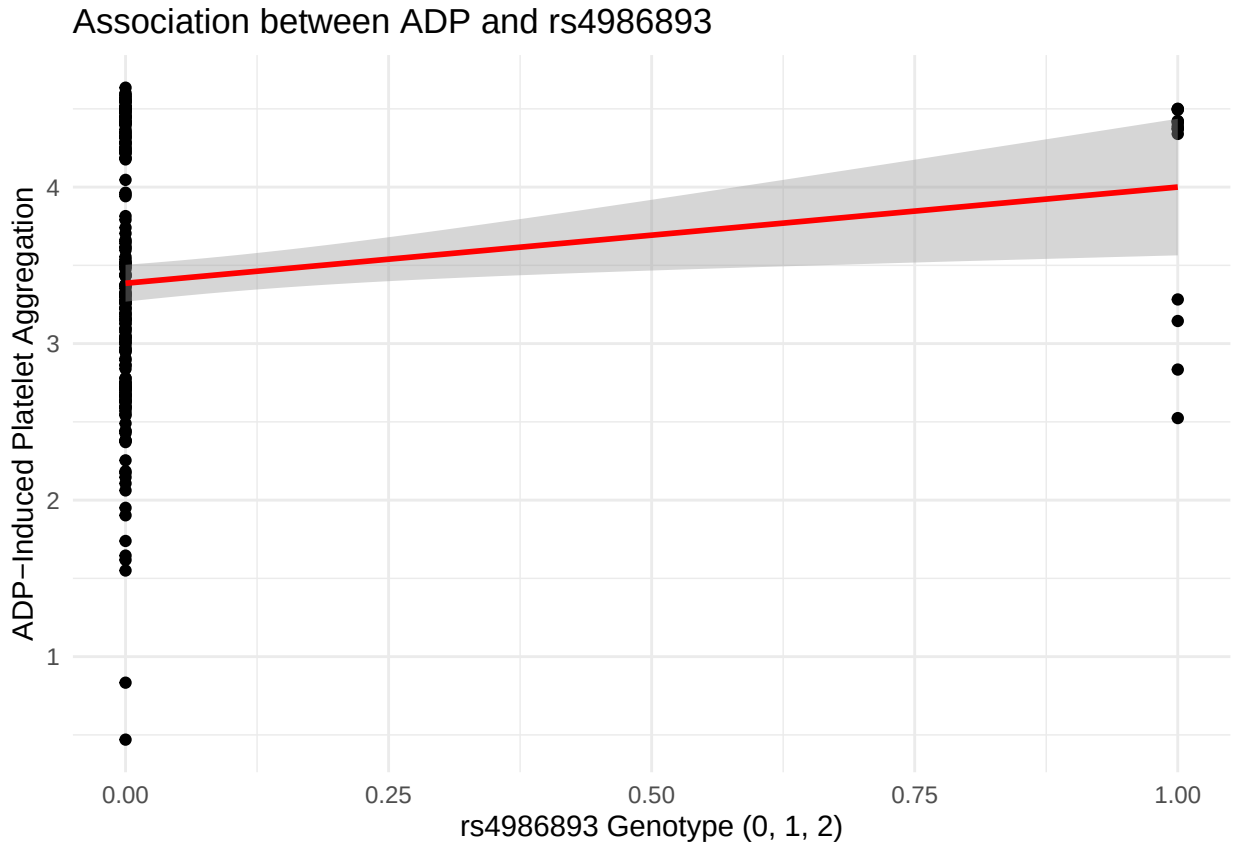
Use ggplot to show the linear regression of rs4986893 and ADP_log with function lm and assign color as "red, x axis as rs4986893 and y axis as ADP.

```
ggplot(clean_data, aes(x = rs4986893 , y = ADP_log)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Association between ADP and rs4986893",
       x = "rs4986893 Genotype (0, 1, 2)",
       y = "ADP-Induced Platelet Aggregation") +theme_minimal()
```

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

```
## Warning: Removed 5 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

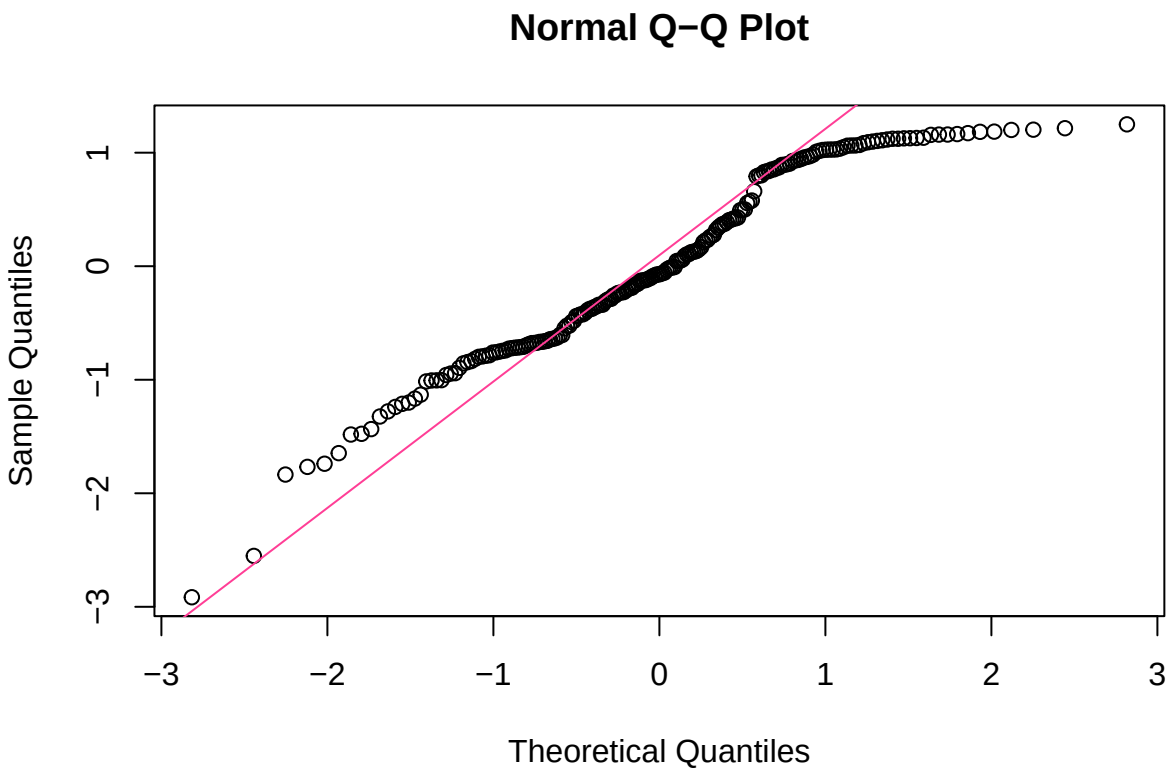
```
## Warning: Removed 5 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




From this graph, the linear is slightly going up, that mean the rs4986893 that have recessive gene can induce platelet aggregation level than the normal one.

Use qqnorm to plot the data to graph and qqline to plot line as violetred color to graph.

```
qqnorm(log_liner_B$residuals)
qqline(log_liner_B$residuals, col = "violetred1")
```




Call summary function for log_liner_C

```
summary(log_liner_C)
```

```
##
## Call:
## lm(formula = ADP_log ~ rs662, data = clean_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
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.0003684,  Adjusted R-squared:  -0.004532
## F-statistic: 0.07518 on 1 and 204 DF,  p-value: 0.7842
```

From p-value equal to -0.-4532. That more than 0.05, so we accept null Hypothesis. Then rs662 is not significant to induce platelet aggregation level.

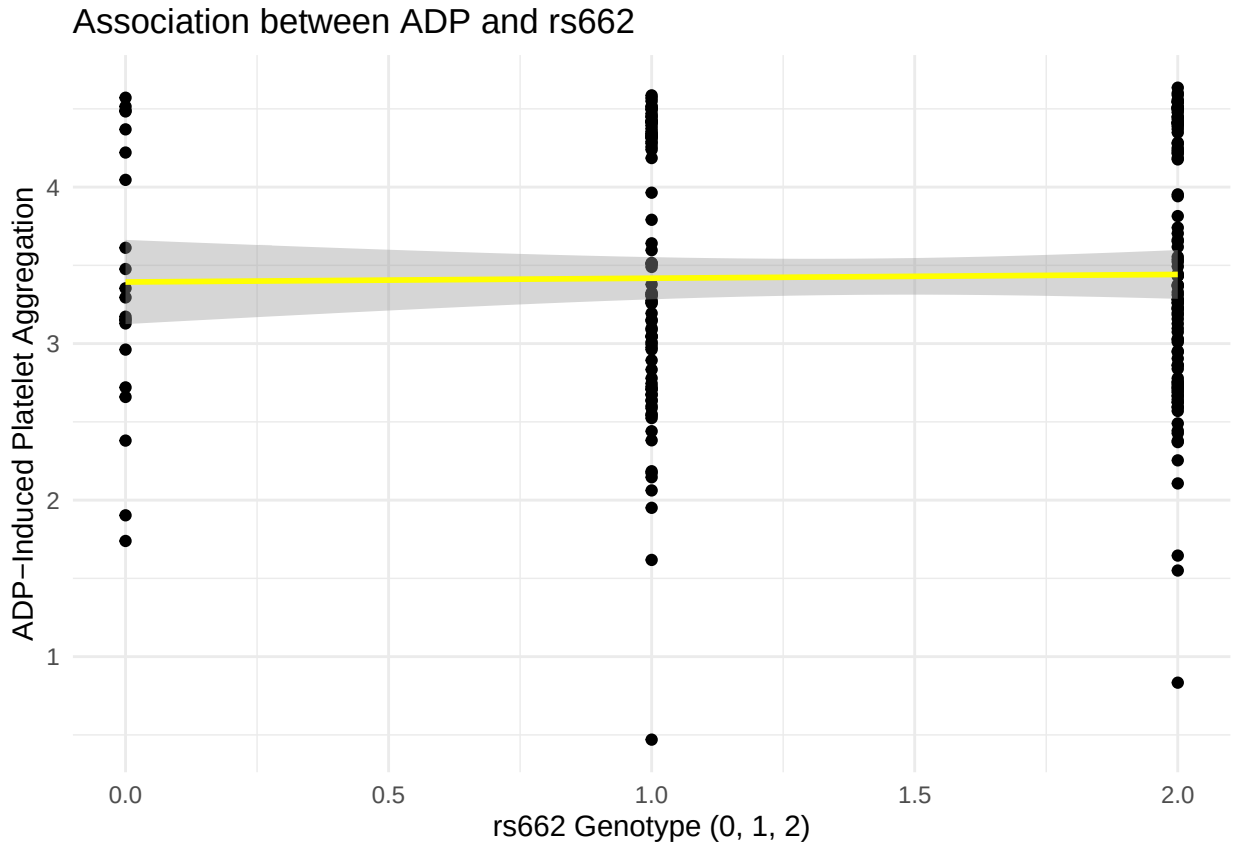
Use ggplot to show the linear regression of rs662 and ADP_log with function lm and assign color as “yellow”, x axis as rs4986893 and y axis as ADP.

```
ggplot(clean_data, aes(x = rs662 , y = ADP_log)) +
  geom_point() +
  geom_smooth(method = "lm", color = "yellow") +
  labs(title = "Association between ADP and rs662",
       x = "rs662 Genotype (0, 1, 2)",
       y = "ADP-Induced Platelet Aggregation") +theme_minimal()
```

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

```
## Warning: Removed 5 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

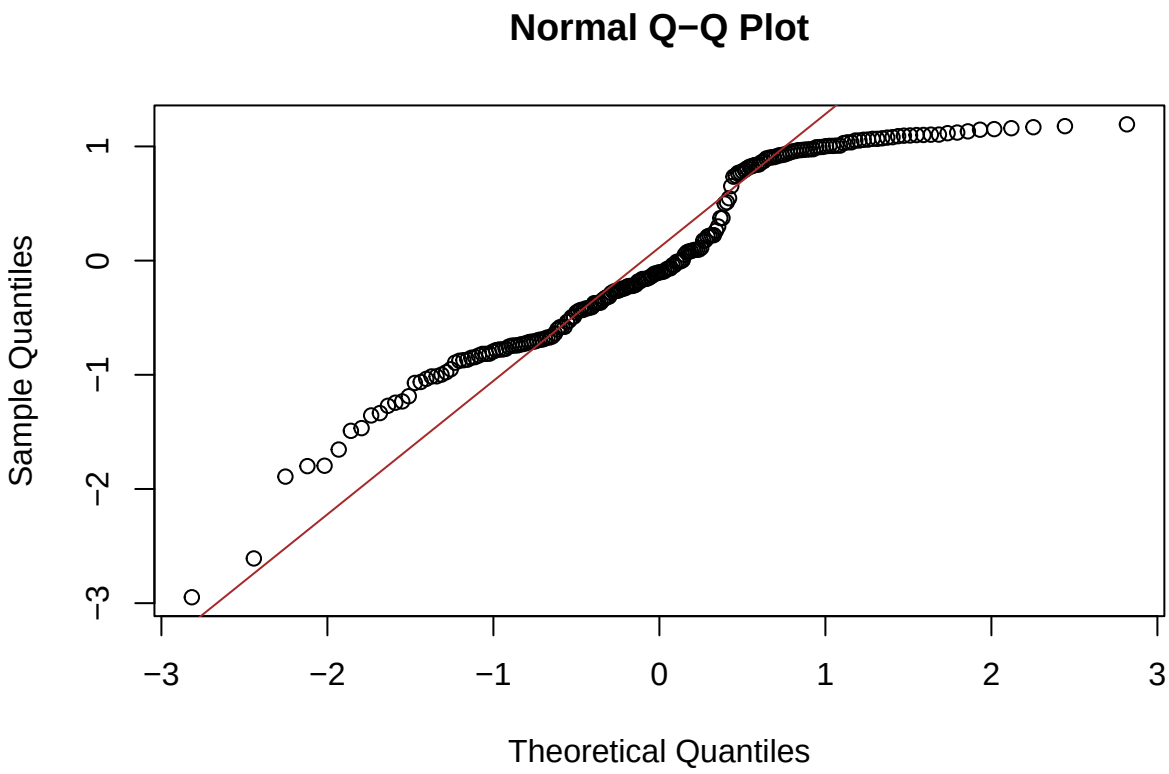
```
## Warning: Removed 5 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




From this graph, linear is constant values, so no matter the rs662 have recessive gene or not. It will do not induce platelet aggregation level.

Use qqnorm to plot the data to graph and qqline to plot line as brown color to graph.

```
qqnorm(log_liner_C$residuals)
qqline(log_liner_C$residuals, col = "brown")
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−3
−2
−1
0
1
2
3