Assignment3

Sorawan Tiratrakoonwichaya

2024-11-04

The study “Effects of Cytochrome P450 2C19 and Paraoxonase 1 Polymorphisms on Antiplatelet Response to Clopidogrel Therapy in Patients with Coronary Artery Disease” (Tresukosol, et al. PLoS One. 2014) collected data on IID, ADP-induced platelet aggregation level (ADP), Clopidogrel resistance (Resistant; 1 = resistance, 0 = not resistance), three SNPs including rs4244285 (CYP2C19*2; 0 = GG, 1 = AG, 2 = AA), rs4986893 (CYP2C19*3; 0 = AA, 1 = AG, 2 = GG), and rs662 (PON1. 192Q>R; 0 = AA, 1 = AG, 2 = GG), age, and sex (0 = male, 1 = female) of the participants and saved in the file “PlateletHW.tsv”, where age and ADP are quantitative variables and the rest are qualitative variables.

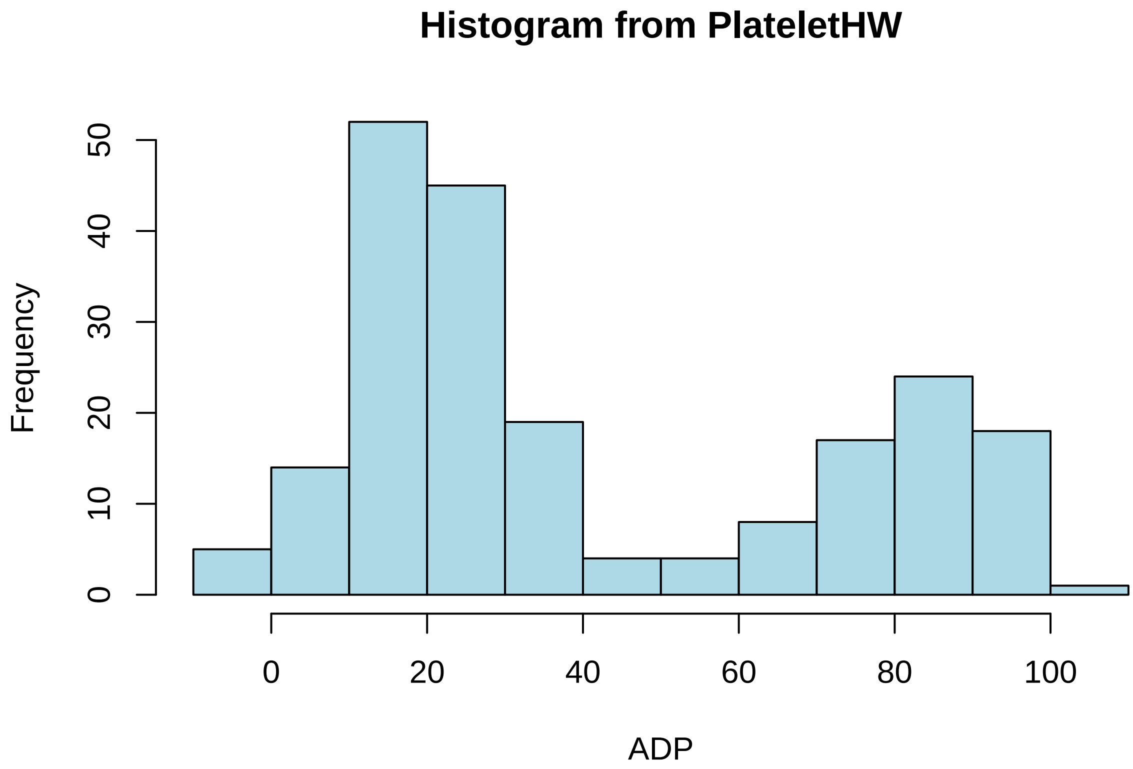
From this data, I want to test for an association between all three SNPs and ADP-induced platelet aggregation levels. I use linear regression to test the association. I also tested the relationships by adding age and sex because I think these two variables may be confounding factors. I started by checking for outliers in the data to deal with the abnormal data. I found some ADP data that was less than 0. I cleaned the data by absolute-valuing to be greater than 0 and saved the data in a new file named “clean_data.tsv” and ensured that there were no outliers in the data.

check for outliers in data

```
summary(raw_data$ADP)
```

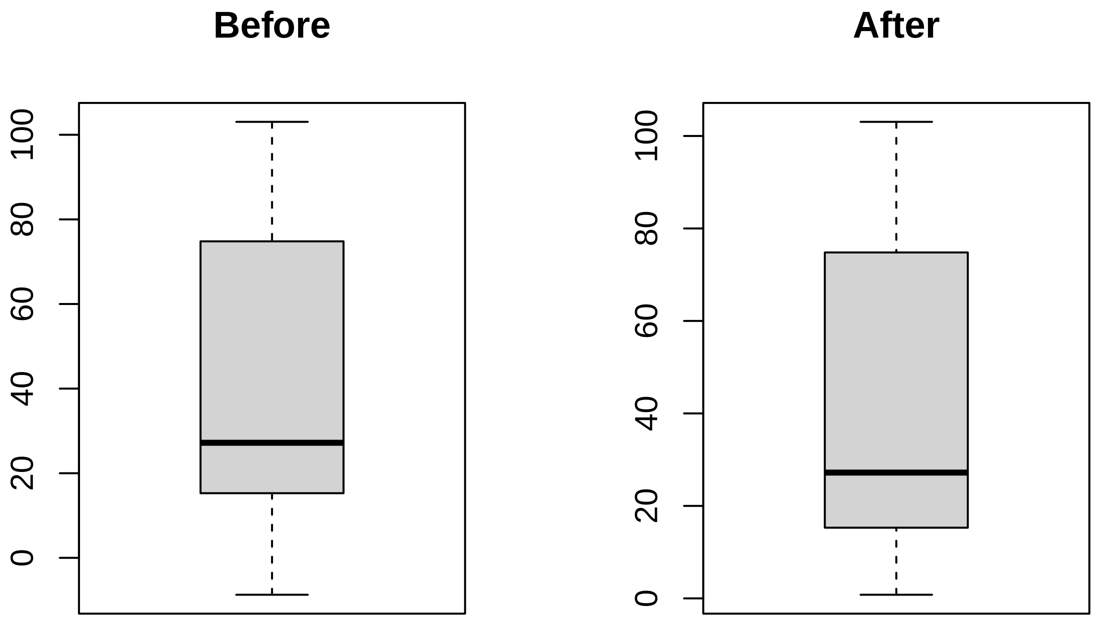
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -8.721  15.281  27.200  40.853  74.810 103.053
```

```
hist(raw_data$ADP, main = "Histogram from PlateletHW" , xlab = "ADP", col = "lightblue")
```



create clean data and compare the data before and after cleaned.

```
clean_data$ADP <- abs(raw_data$ADP)
par(mfcol=c(1,2))
boxplot(raw_data$ADP, main = "Before")
boxplot(clean_data$ADP, main = "After")
```

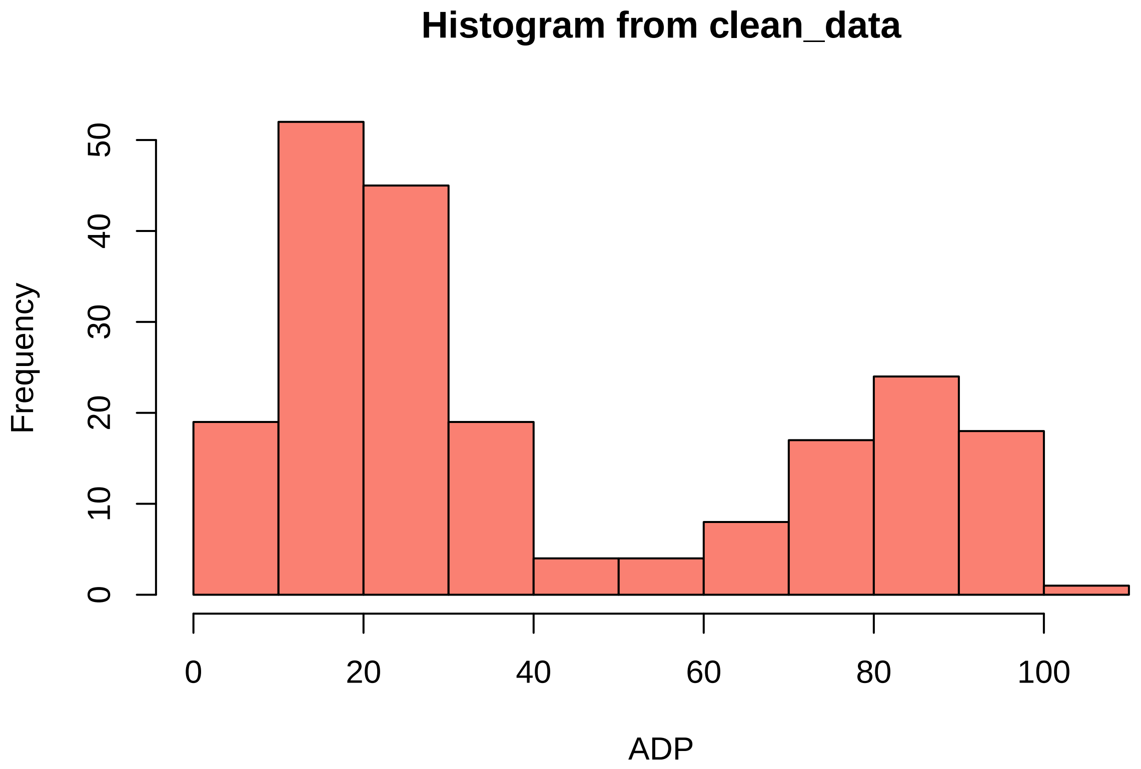


check for outliers in clean data

```
summary(clean_data$ADP)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
##   0.7849  15.2810  27.2005  41.1392  74.8096 103.0535
```

```
hist(clean_data$ADP, main = "Histogram from clean_data" , xlab = "ADP", col = "salmon")
```



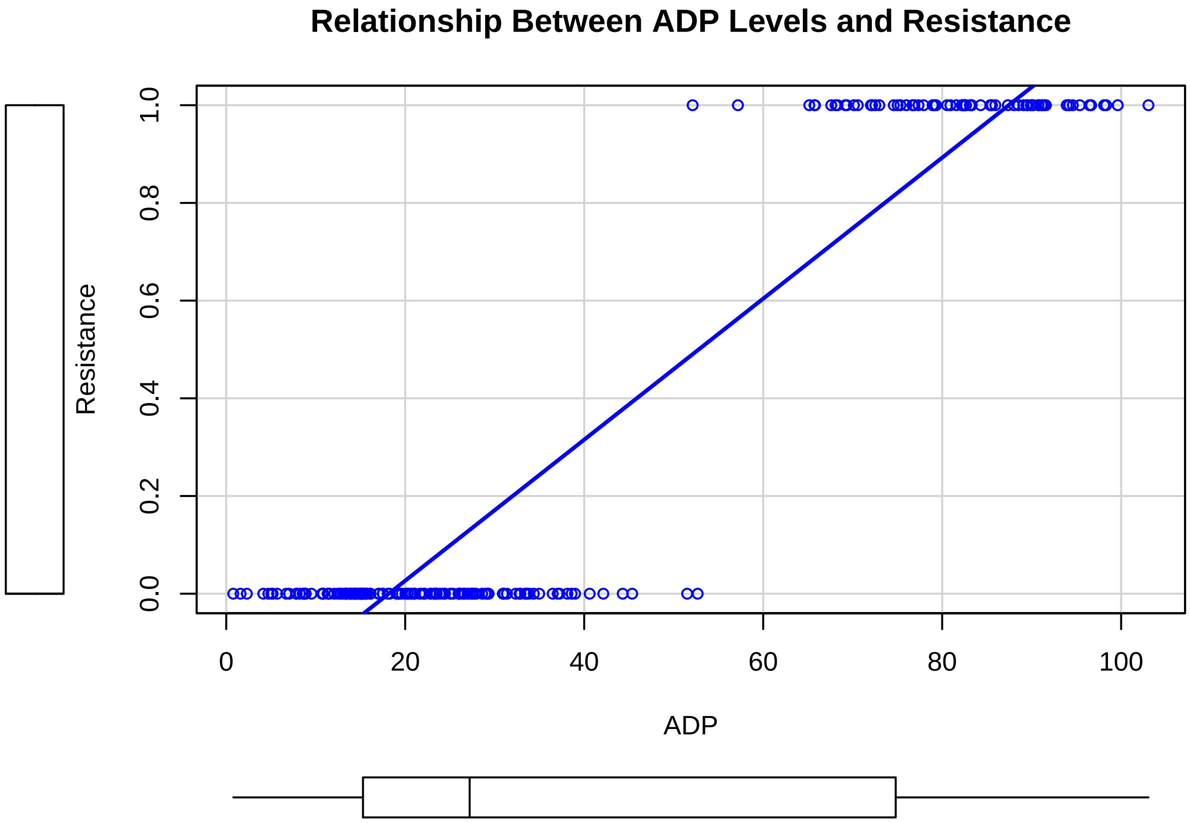
Then I did some preliminary correlation between ADP and resistance, rs4244285, rs4986893, and rs662 genotypes by making a scatter plot between ADP and Resistance and a boxplot between ADP and each SNP. From the graph of the relationship between ADP and Resistance, it can be seen that higher ADP levels correlate with increased drug resistance. From the graph of the relationship between ADP and rs4244285, it can be seen that people with genotype AA have higher ADP levels than genotypes AG and GG. From the graph of the relationship between ADP and rs4986893, it can be seen that people with genotype AG have higher ADP levels than genotype AA. And from the graph of the relationship between ADP and rs4986893, it can be seen that no matter what genotype people have, they will have similar ADP levels.

Checking The Relationship Between ADP Levels and Resistance, rs4244285, rs4986893, and rs662 Genotype

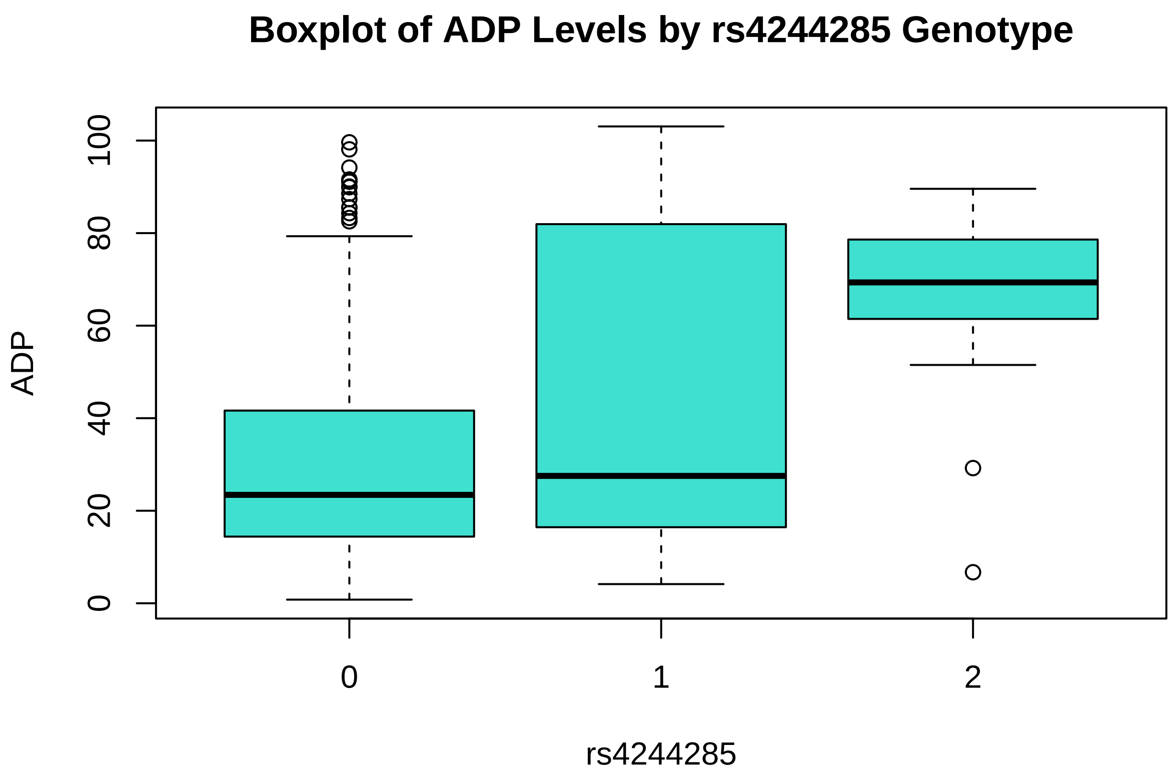
```
library(car)
```

```
## Loading required package: carData
```

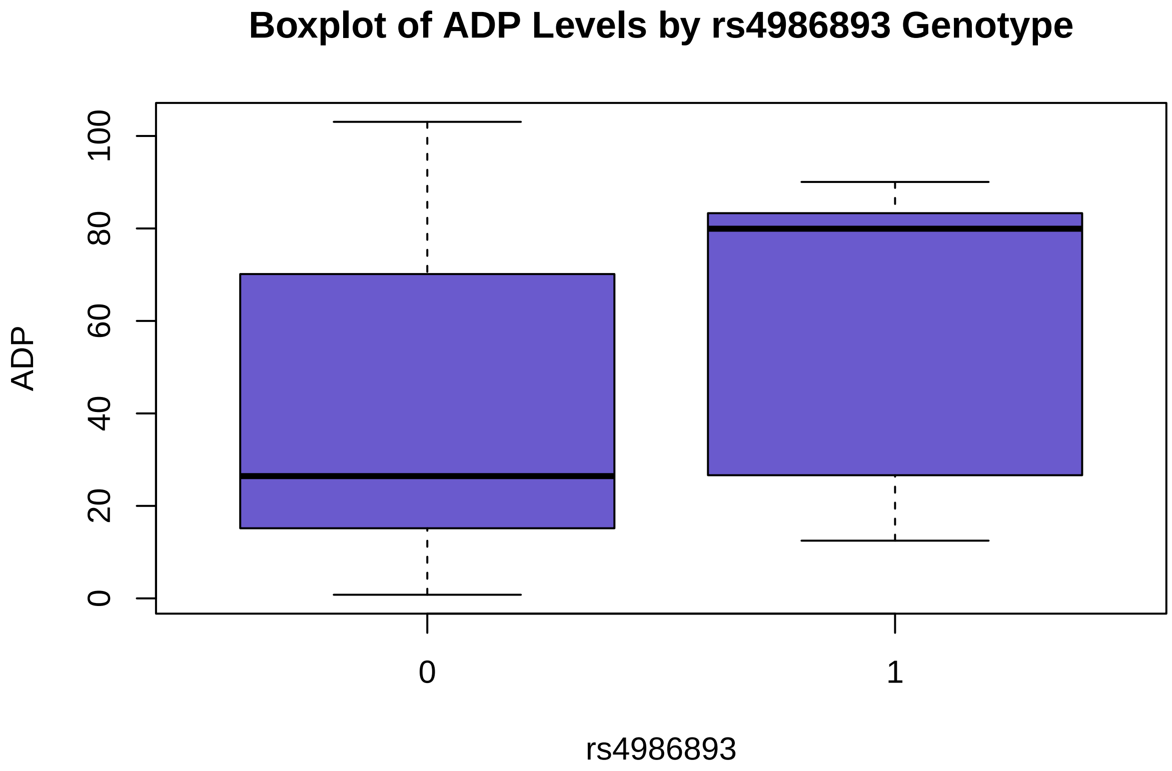
```
scatterplot(Resistance ~ ADP, data = clean_data, reg.line = lm, smooth = FALSE,
            main = "Relationship Between ADP Levels and Resistance")
```



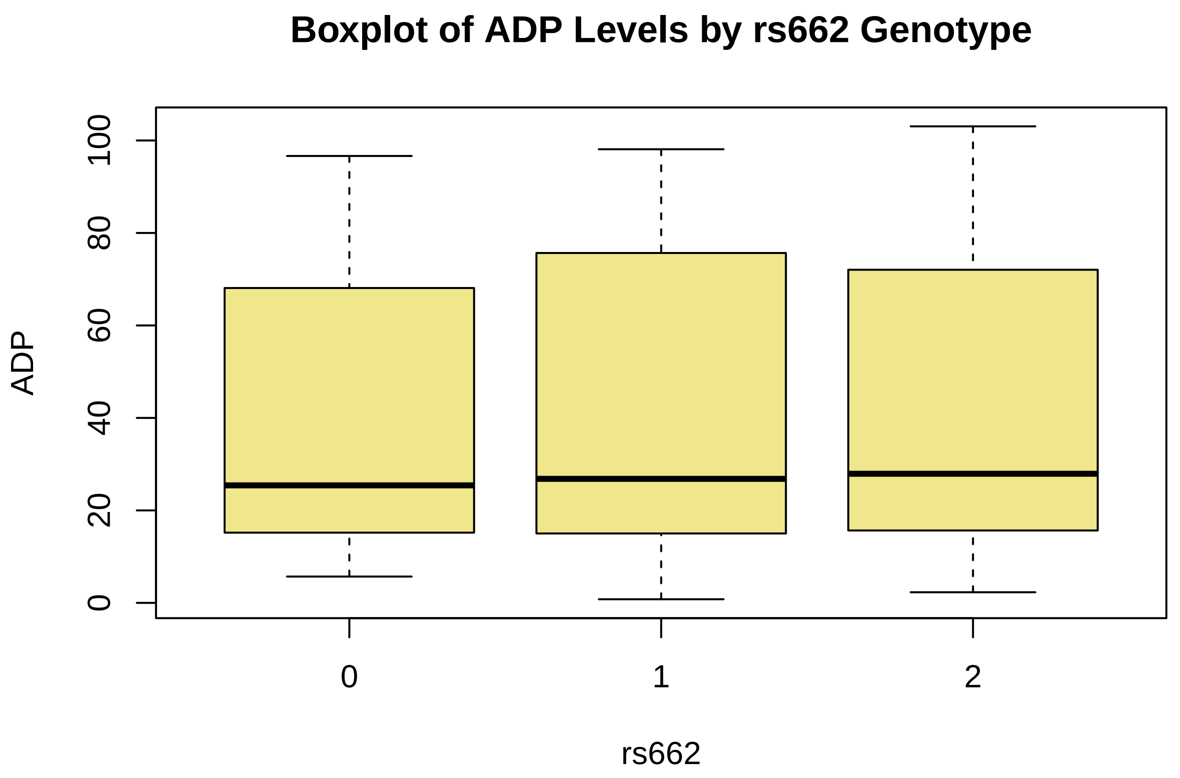
```
boxplot(clean_data$ADP ~ clean_data$rs4244285, main = "Boxplot of ADP Levels by rs4244285 Genotype",
        xlab = "rs4244285", ylab = "ADP", col = "turquoise")
```



```
boxplot(clean_data$ADP ~ clean_data$rs4986893, main = "Boxplot of ADP Levels by rs4986893 Genotype",
        xlab = "rs4986893", ylab = "ADP", col = "slateblue")
```



```
boxplot(clean_data$ADP ~ clean_data$rs662, main = "Boxplot of ADP Levels by rs662 Genotype",
        xlab = "rs662", ylab = "ADP", col = "khaki")
```



Next step, I will do a Linear regression test between ADP and each SNP. I checked the Quantile-Quantile plot (QQ plot) of each relationship and found that I should Normalize ADP first to make the data distribution closer to a normal distribution, I chose to take the logarithm of the ADP data.

Linear regression test

Normalize ADP by taking log

```
clean_data$ADP_log <- log(clean_data$ADP)
View(clean_data)
```

QQ plot of relationship between ADP and rs4244285 before and after normalize ADP

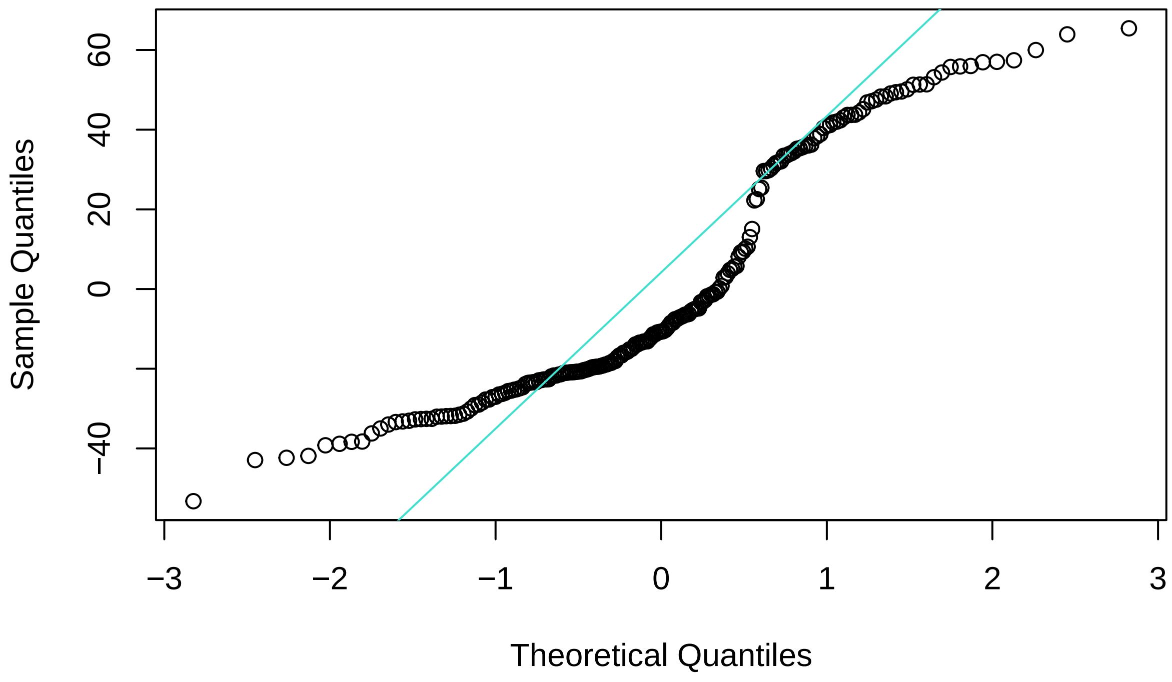
```
linear_1 <- lm(ADP ~ rs4244285, data = clean_data)
summary(linear_1)
```

```
##
## Call:
## lm(formula = ADP ~ rs4244285, data = clean_data)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.25 -22.27 -10.70  30.64  65.45
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   34.175      2.711  12.607  < 2e-16 ***
## rs4244285     12.889      3.281   3.928 0.000116 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 29.79 on 209 degrees of freedom
## Multiple R-squared:  0.06875,    Adjusted R-squared:  0.06429
## F-statistic: 15.43 on 1 and 209 DF,  p-value: 0.0001163
```

```
qqnorm(linear_1$residuals,
       main = "QQ Plot: Relationship of ADP and rs4244285 (Unnormalized ADP)")
qqline(linear_1$residuals, col = "turquoise")
```

QQ Plot: Relationship of ADP and rs4244285 (Unnormalized ADP)

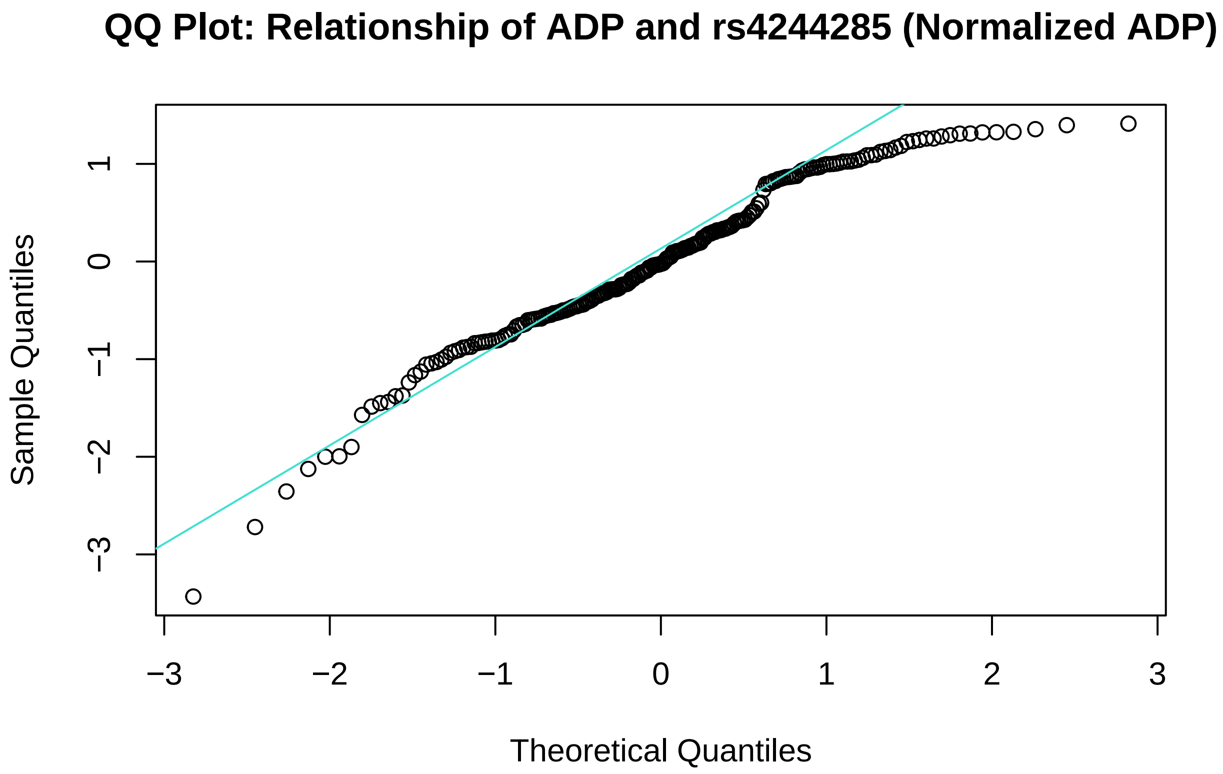


```
linear_log_1 <- lm(ADP_log ~ rs4244285, data = clean_data)
summary(linear_log_1)
```

```
##
## Call:
## lm(formula = ADP_log ~ rs4244285, data = clean_data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4317 -0.5467 -0.0235  0.8128  1.4120
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.18940    0.07873   40.51  < 2e-16 ***
## rs4244285    0.35644    0.09530    3.74 0.000238 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8651 on 209 degrees of freedom
## Multiple R-squared:  0.06273,    Adjusted R-squared:  0.05825
## F-statistic: 13.99 on 1 and 209 DF,  p-value: 0.0002375
```

```r
qqnorm(linear_log_1$residuals,
       main = "QQ Plot: Relationship of ADP and rs4244285 (Normalized ADP)")
qqline(linear_log_1$residuals, col = "turquoise")
```

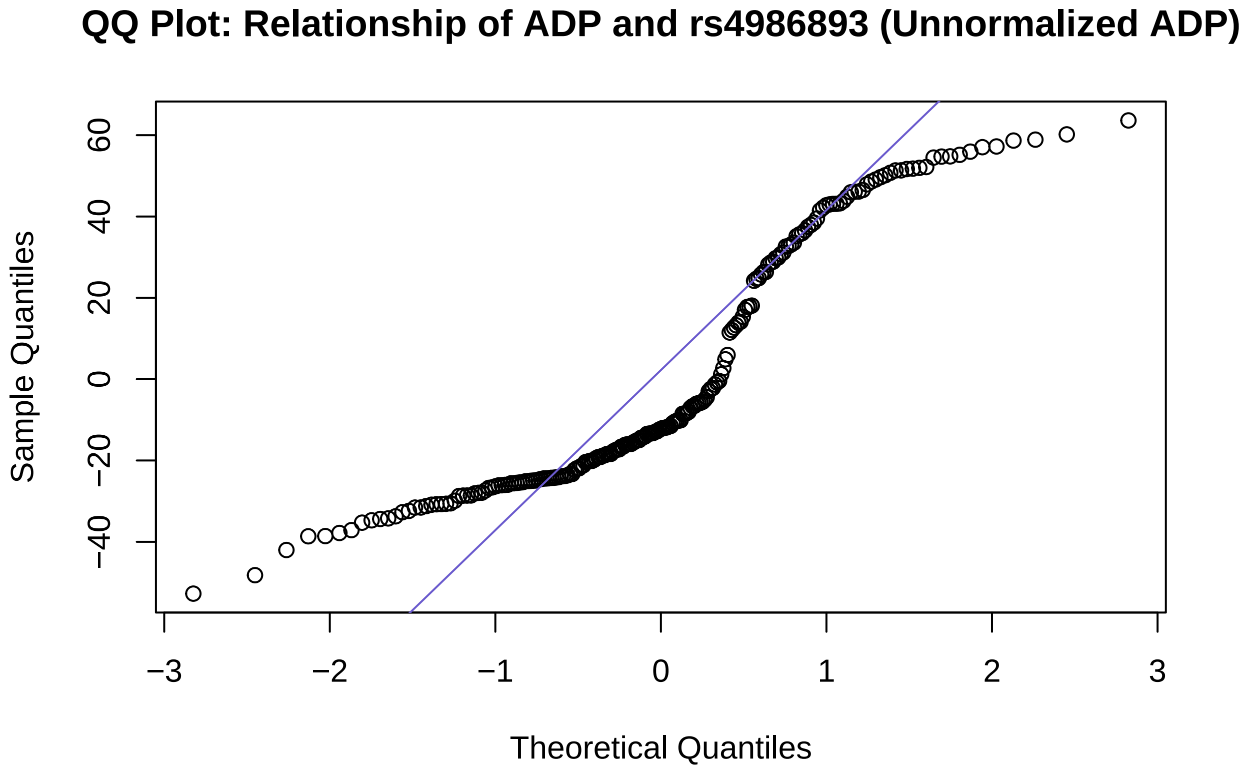


QQ plot of relationship between ADP and rs4986893 before and after normalize ADP

```r
linear_2 <- lm(ADP ~ rs4986893, data = clean_data)
summary(linear_2)
```

```
##
## Call:
## lm(formula = ADP ~ rs4986893, data = clean_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52.74 -24.32 -12.23  28.77  63.63
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   39.428      2.151  18.333  < 2e-16 ***
## rs4986893     25.792      8.349   3.089  0.00228 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 30.19 on 209 degrees of freedom
## Multiple R-squared:  0.04367,    Adjusted R-squared:  0.03909
## F-statistic: 9.543 on 1 and 209 DF,  p-value: 0.00228
```

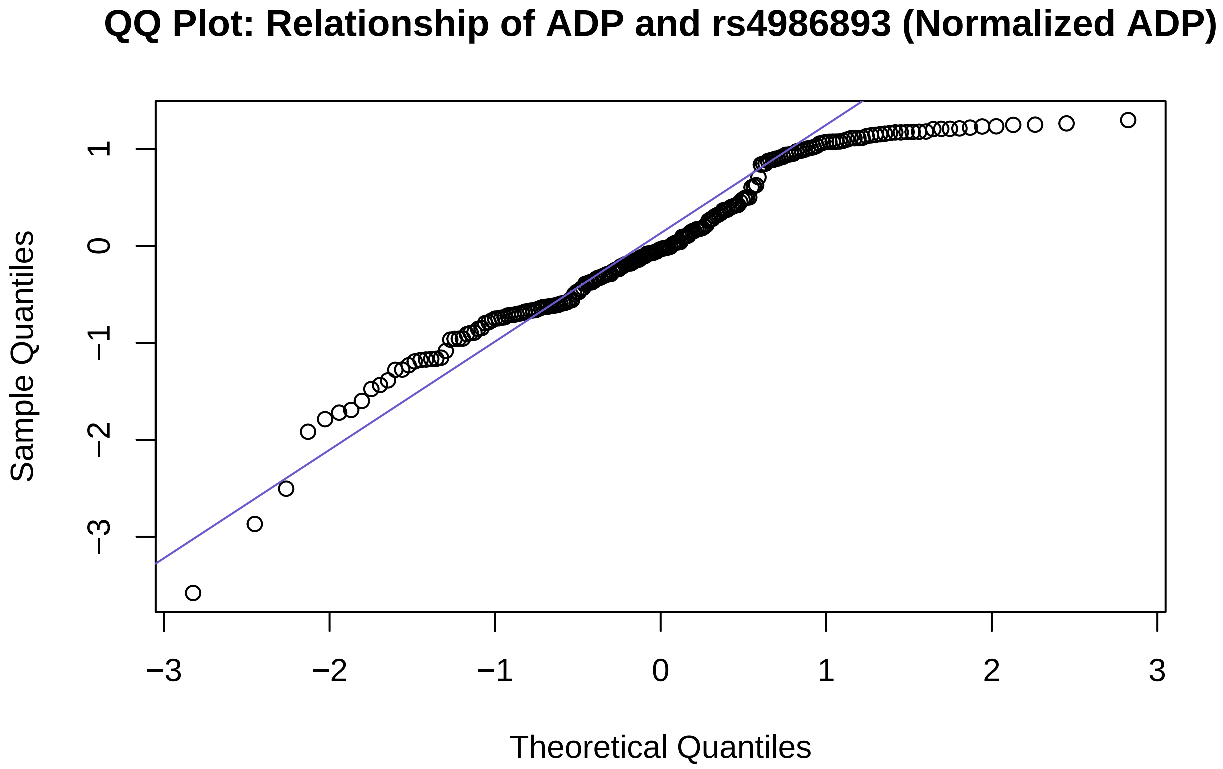
```r
qqnorm(linear_2$residuals,
       main = "QQ Plot: Relationship of ADP and rs4986893 (Unnormalized ADP)")
qqline(linear_2$residuals, col = "slateblue")
```



```r
linear_log_2 <- lm(ADP_log ~ rs4986893, data = clean_data)
summary(linear_log_2)
```

```
##
## Call:
## lm(formula = ADP_log ~ rs4986893, data = clean_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5803 -0.6227 -0.0348  0.8844  1.2972
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.33804    0.06257  53.353  < 2e-16 ***
## rs4986893    0.66218    0.24289   2.726  0.00695 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8781 on 209 degrees of freedom
## Multiple R-squared:  0.03434,    Adjusted R-squared:  0.02972
## F-statistic: 7.433 on 1 and 209 DF,  p-value: 0.006949
```

```r
qqnorm(linear_log_2$residuals,
       main = "QQ Plot: Relationship of ADP and rs4986893 (Normalized ADP)")
qqline(linear_log_2$residuals, col = "slateblue")
```

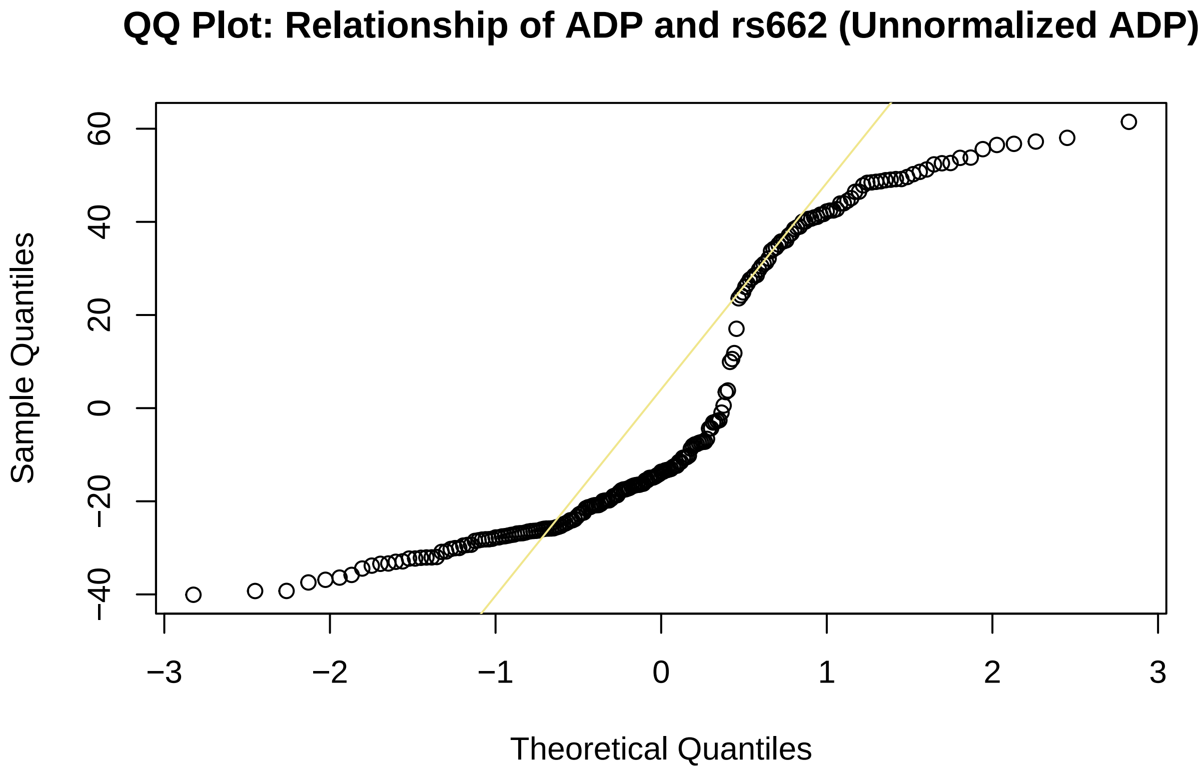


QQ plot of relationship between ADP and rs662 before and after normalize ADP

```
linear_3 <- lm(ADP ~ rs662, data = clean_data)
summary(linear_3)
```

```
##
## Call:
## lm(formula = ADP ~ rs662, data = clean_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -40.07 -25.85 -13.66  33.95  61.48
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   40.137      4.911   8.173 2.86e-14 ***
## rs662          0.719      3.177   0.226    0.821
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.86 on 209 degrees of freedom
## Multiple R-squared:  0.0002449,  Adjusted R-squared:  -0.004539
## F-statistic: 0.0512 on 1 and 209 DF,  p-value: 0.8212
```

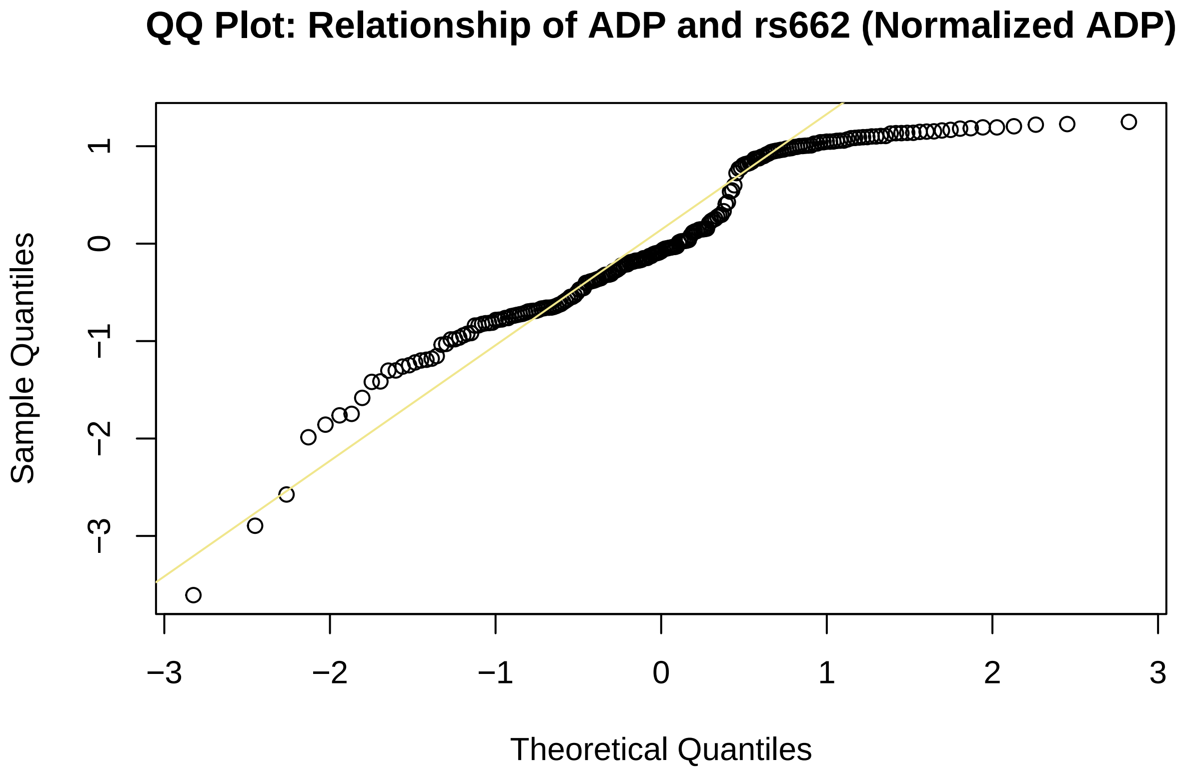
```
qqnorm(linear_3$residuals,
       main = "QQ Plot: Relationship of ADP and rs662 (Unnormalized ADP)")
qqline(linear_3$residuals, col = "khaki")
```



```
linear_log_3 <- lm(ADP_log ~ rs662, data = clean_data)
summary(linear_log_3)
```

```
##
## Call:
## lm(formula = ADP_log ~ rs662, data = clean_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6073 -0.6565 -0.0787  0.9437  1.2492
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.32192    0.14211  23.376   <2e-16 ***
## rs662        0.04310    0.09195   0.469     0.64
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8932 on 209 degrees of freedom
## Multiple R-squared:  0.00105,    Adjusted R-squared:  -0.003729
## F-statistic: 0.2197 on 1 and 209 DF,  p-value: 0.6397
```

```
qqnorm(linear_log_3$residuals,
       main = "QQ Plot: Relationship of ADP and rs662 (Normalized ADP)")
qqline(linear_log_3$residuals, col = "khaki")
```

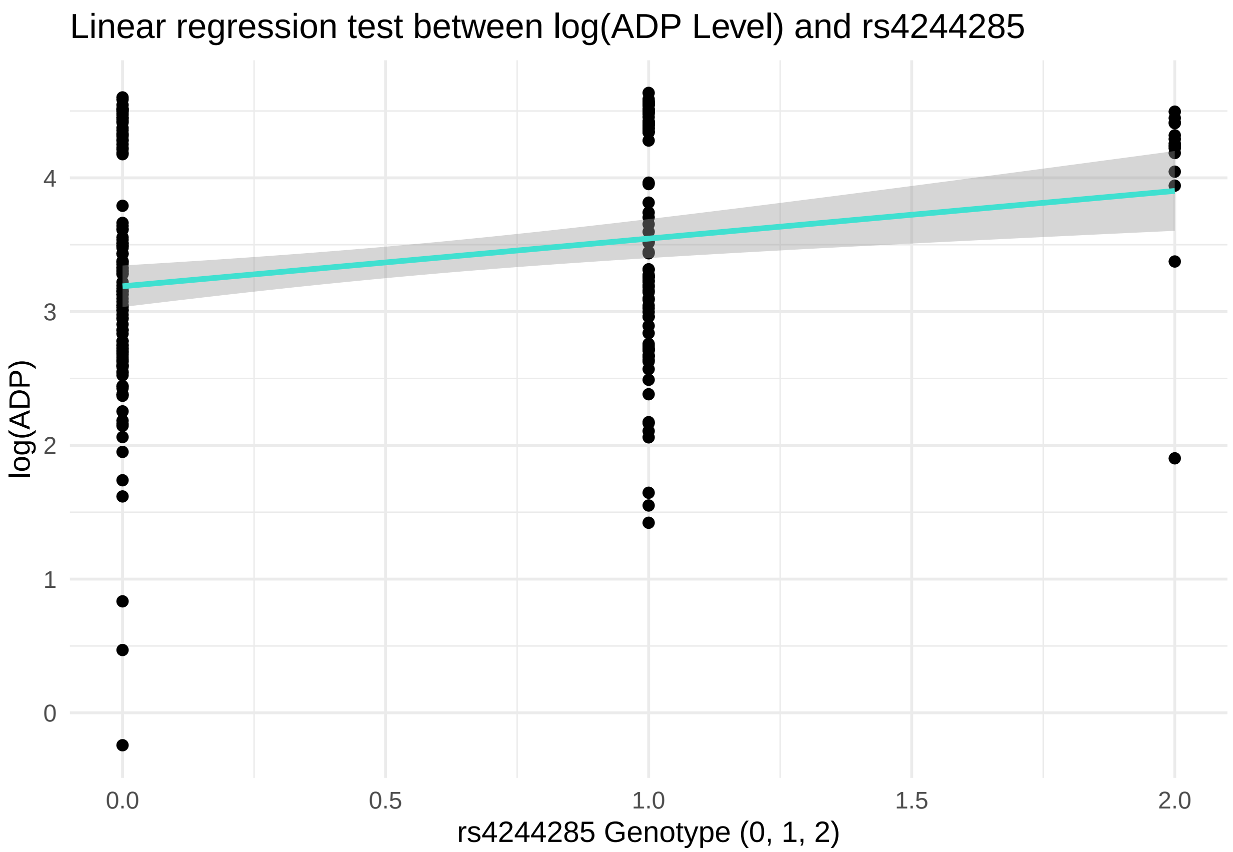


From checking the Quantile-quantile plot (QQ plot), I did a Linear regression test between log(ADP) and each SNP instead. From the graph, it can be seen that the relationship of ADP with each SNP tends in the same direction as the preliminary relationship before.

Linear regression test between log(ADP) and rs4244285

```r
library(ggplot2)
ggplot(clean_data, aes(x = rs4244285, y = ADP_log)) +
  geom_point() +
  geom_smooth(method = "lm", color = "turquoise") +
  labs(title = "Linear regression test between log(ADP Level) and rs4244285",
       x = "rs4244285 Genotype (0, 1, 2)",
       y = "log(ADP)") +
  theme_minimal()
```

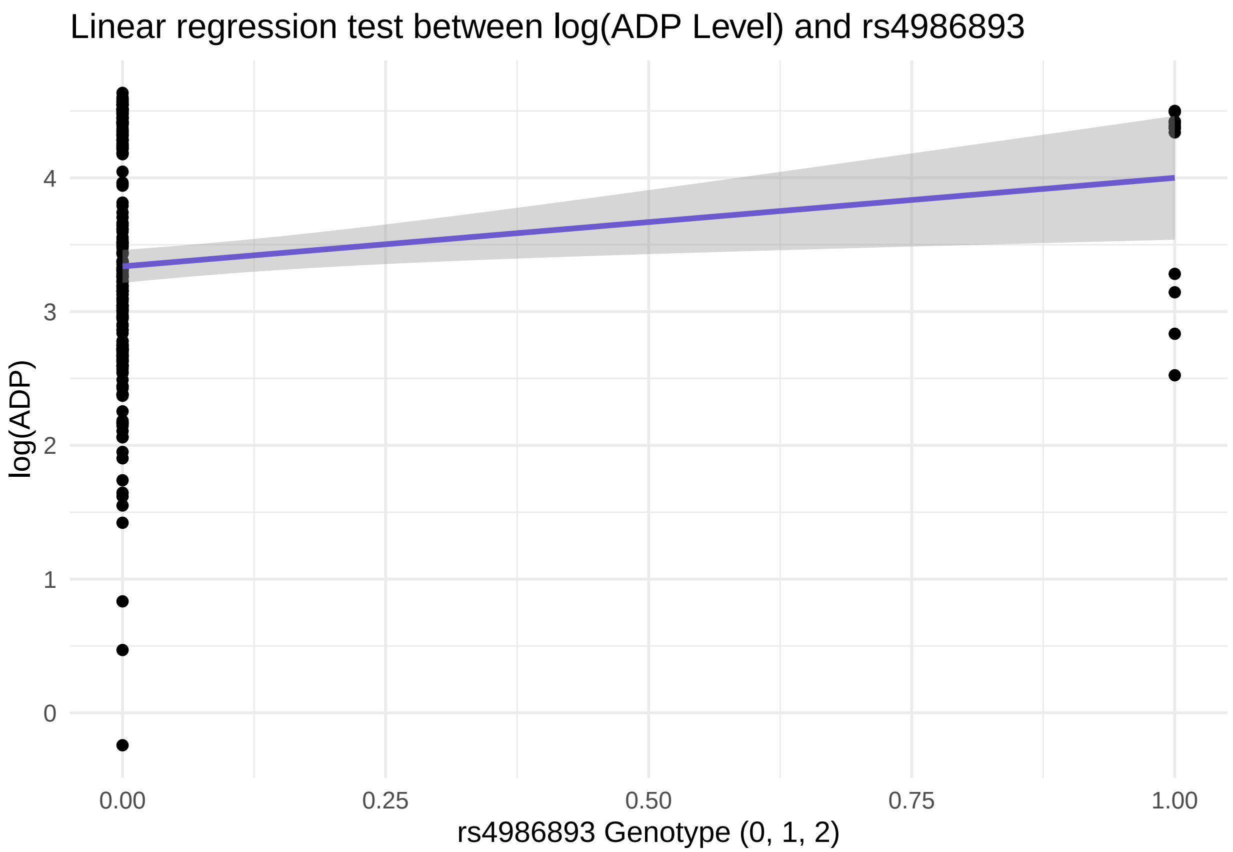
```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



Linear regression test between log(ADP) and rs4986893

```
library(ggplot2)
ggplot(clean_data, aes(x = rs4986893, y = ADP_log)) +
  geom_point() +
  geom_smooth(method = "lm", color = "slateblue") +
  labs(title = "Linear regression test between log(ADP Level) and rs4986893",
       x = "rs4986893 Genotype (0, 1, 2)",
       y = "log(ADP)") +
  theme_minimal()
```

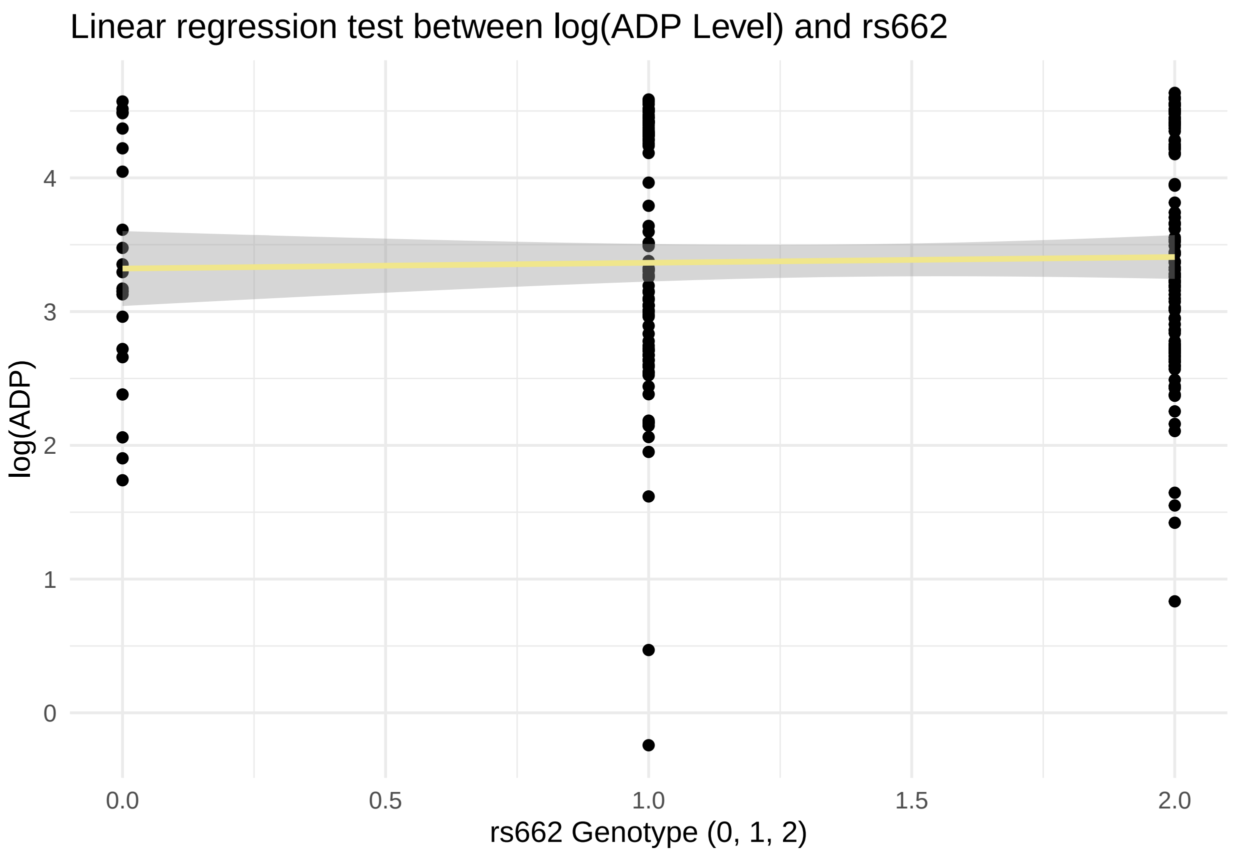
```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



Linear regression test between log(ADP) and rs662

```
library(ggplot2)
ggplot(clean_data, aes(x = rs662, y = ADP_log)) +
  geom_point() +
  geom_smooth(method = "lm", color = "khaki") +
  labs(title = "Linear regression test between log(ADP Level) and rs662",
       x = "rs662 Genotype (0, 1, 2)",
       y = "log(ADP)") +
  theme_minimal()
```

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



When I added the data of sex and age to each association, the Linear regression test of rs42244285 and rs4986893 were significantly associated with ADP-platelet aggregation (P value = 0.000203, 0.00701, P value < 0.05/3). On the other hand, the Linear regression test of rs662 are not statistically significant to ADP (P value = 0.608, P value > 0.05/3). In addition, the confounding variables (sex and age) were not significant. However, the R-square values were very low, and some graphs had negative R-squared values, indicating that the chosen model could not predict or explain the variability of the data.

Linear regression test that added sex and age to each relationship

```
snp_list <- c("rs4244285", "rs4986893", "rs662")

results_list <- list()

for (snp in snp_list) {
  model_sum <- lm(as.formula(paste("ADP_log ~ AGE + SEX +", snp)), data = clean_data)
  results_list[[snp]] <- summary(model_sum)
}
print(results_list[["rs4244285"]]) # significant (p-value < 0.05/3)
```

```
##
## Call:
## lm(formula = as.formula(paste("ADP_log ~ AGE + SEX +", snp)),
##     data = clean_data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4329 -0.5847  0.0234  0.7691  1.3790
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.642427   0.369440   9.859  < 2e-16 ***
## AGE         -0.006644   0.005629  -1.180 0.239184
## SEX         -0.047234   0.134027  -0.352 0.724883
## rs4244285    0.360830   0.095387   3.783 0.000203 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.8652 on 207 degrees of freedom
## Multiple R-squared:  0.07152,    Adjusted R-squared:  0.05806
## F-statistic: 5.315 on 3 and 207 DF,  p-value: 0.001507
```

```
print(results_list[["rs4986893"]])# significant (p-value < 0.05/3)
```

```
##
## Call:
## lm(formula = as.formula(paste("ADP_log ~ AGE + SEX +", snp)),
##     data = clean_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5707 -0.5926 -0.0460  0.8435  1.3307
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.787300   0.372706  10.162  < 2e-16 ***
## AGE         -0.006749   0.005720  -1.180  0.23941
## SEX         -0.006994   0.136411  -0.051  0.95916
## rs4986893    0.663689   0.243691   2.723  0.00701 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.879 on 207 degrees of freedom
## Multiple R-squared:  0.04167,    Adjusted R-squared:  0.02778
## F-statistic:     3 on 3 and 207 DF,  p-value: 0.03161
```

```
print(results_list[["rs662"]]) # not significant (p-value > 0.05/3)
```

```
##
## Call:
## lm(formula = as.formula(paste("ADP_log ~ AGE + SEX +", snp)),
##     data = clean_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6018 -0.6330 -0.0758  0.9179  1.2493
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.758708   0.392094   9.586   <2e-16 ***
## AGE         -0.006572   0.005838  -1.126    0.262
## SEX         -0.024108   0.139273  -0.173    0.863
## rs662        0.047704   0.092747   0.514    0.608
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8941 on 207 degrees of freedom
## Multiple R-squared:  0.008599,   Adjusted R-squared:  -0.005769
## F-statistic: 0.5985 on 3 and 207 DF,  p-value: 0.6167
```

From all the test results, it was concluded that the genotype of CYP2C19*2 (rs42244285) and CYP2C19*3 (rs4986893) significantly affected the ADP-induced platelet aggregation level, while the genotype of PON1. 192Q>R (rs662), sex, and age had no significant effect on the ADP-induced platelet aggregation level.