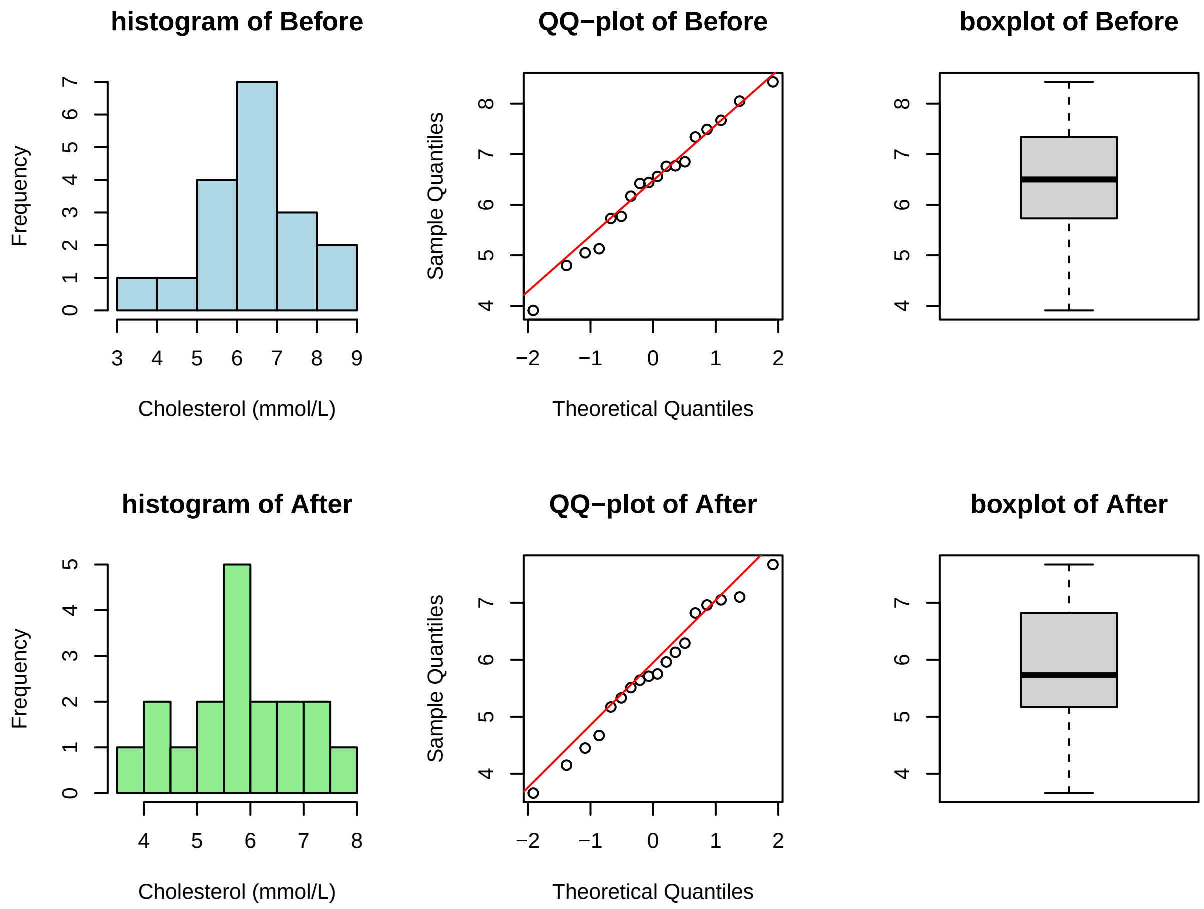
# Assignment1

Yinghao Luo, Zheyuan Zhang, Yujie Cao

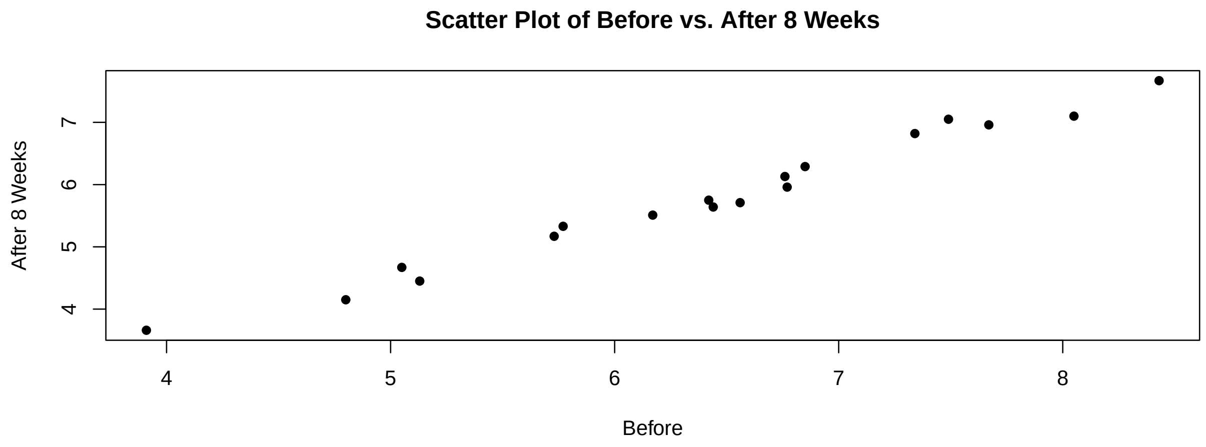
## Exercise 1

**(a)** We draw several figures to check the characteristics of the dataset. According to the Q-Q plots and from the shape of histogram, we can assume that the data Before and After 8 weeks both follow a normal distribution, since the points are approximately on a straight line,although the sample size is relatively small.Besides,we also check the scale,symmetry and outliers of the dataset through boxplot.



Then we also plot the scatter figures,we can see that there is approximately linear relation between these two values.

```r
plot(data$Before, data$After8weeks,
     main = "Scatter Plot of Before vs. After 8 Weeks",
     xlab = "Before", ylab = "After 8 Weeks",
     col = "black", pch = 16)
```



The correlation of the two columns can also be calculated by Pearson as normality is assumed. The result is that the value of correlation is 0.991, which can infer that there is significant correlation.

**(b)** First,the data are paired as it is an experiment with two numerical outcomes per experimental unit.Therefore, the permutation test is applicable since permutation test is suitable for paired samples. However,Mann-Whitney test is not applicable, because it is utilized to compare two independent samples.

We already know that the data stems from a approximately normal distribution,so we can conduct **paired t-test** as follows:

```
t.test(data$Before,data$After8weeks, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  data$Before and data$After8weeks
## t = 15, df = 17, p-value = 3e-11
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.540 0.718
## sample estimates:
## mean difference
##           0.629
```

The p-value is 3e-11, so we can know taht $H_0$ can be rejected. $H_0$ means no significant difference between Before and After8weeks. In this case, there is significant difference between Before and After8weeks.

Then,as the sample size is 18,relatively small,we prefer non-Parametric tests. we can use **permutation test** as follows: The test statistic we choose is mean difference.

```
mean_difference = function(x,y) {mean(x-y)}
original_diff <- mean(data$Before - data$After8weeks)
B=1000; tstar=numeric(B)
for (i in 1:B) {
  md_star=t(apply(cbind(data$Before,data$After8weeks),1,sample))
  tstar[i]=mean_difference(data$Before,data$After8weeks)
}
myt = mean_difference(data$Before,data$After8weeks);
pl=sum(tstar<myt)/B;pr=sum(tstar>myt)/B;p=2*min(pl,pr); p
```

```
## [1] 0
```

In the permutation test, we can reject $H_0$ and conclude that the diet with low fat margarine indeed has a significant effect.

**(c)** First, we construct the a 97%-CI for $\mu$ based on normality. As $X_1, \ldots, X_{18} \sim N(\mu, \sigma^2)$ , we can calculate the t-confidence interval of level $1 - \alpha$ for $\mu$.We use mean instead of median as the estimator because the outlier do not exists from conclusion in (a).

```
x = data$After8weeks;n = length(x);x_mean = mean(x);x_sd = sd(x) ;alpha = 0.03;t = qt(1-alpha/2, df = n
CI_normal = c(
  x_mean - t * (x_sd / sqrt(n)),
  x_mean + t * (x_sd / sqrt(n)))
cat("97% CI based on normality: [", round(CI_normal[1], 3),
    ",", round(CI_normal[2], 3), "]\n")
```

```
## 97% CI based on normality: [ 5.16 , 6.39 ]
```

Bootstrap CI is an alternative to determine CI's for non-normal sample, we implement bootstrap CI as follows:

```
B = 1000
x_mean = mean(data$After8weeks)
Tstar = numeric(B)
for(i in 1:B){
  Xstar=sample(data$After8weeks,replace=TRUE)
  Tstar[i]=mean(Xstar)
}
Tstar15 = quantile(Tstar, 0.015)
Tstar985 = quantile(Tstar, 0.985)
CI_bootstrap = c(2*x_mean-Tstar985,2*x_mean-Tstar15)
cat("Bootstrap 97% CI: [", round(CI_bootstrap[1], 3),
    ",", round(CI_bootstrap[2], 3), "]\n")
```

```
## Bootstrap 97% CI: [ 5.25 , 6.31 ]
```

As bootstrap CI will always yield a different interval when repeating,we run the chunk several time. And the conclusion is that the difference between the two confidence intervals is slight. From our perspective, this is be because the column after8weeks data steams form a normal distribution.

**(d)** We use the maximum of the sample as the test statistic and set B as 1000 to run the bootstrap test.We repeated the test a few times to see if the result is stable.

```
n=length(data$After8weeks)
t=max(data$After8weeks)
B=1000;
tstar=numeric(B)
theta_values = seq(3, 12, by = 0.1)
```

```
p_values = sapply(theta_values, function(theta) {
  tstar = numeric(B)
  for (i in 1:B){
    xstar = runif(n, min = 3, max = theta)
    tstar[i] = max(xstar)
  }
  pl = sum(tstar < t) / B
  pr = sum(tstar > t) / B
  p = 2 * min(pl, pr)
  return(p)
})
theta_valid = theta_values[p_values > 0.05]
cat("The range of theta for H0 cannot be rejected: [", min(theta_valid), max(theta_valid),  "]\n")
```

```
## The range of theta for H0 cannot be rejected: [ 7.7 8.7 ]
```

Kolmogorov-Smirnov test can not be applied as it aims at two independent samples,but our data is paired.

**(e)** Since we are verifying the median now,the **Sign test** can be applied as follows:

```
n = length(data$After8weeks)
below_6 = sum(data$After8weeks < 6);
binom.test(below_6,n,p=0.5,alt="l")
```

```
##
##  Exact binomial test
##
## data:  below_6 and n
## number of successes = 11, number of trials = 18, p-value = 0.9
## alternative hypothesis: true probability of success is less than 0.5
## 95 percent confidence interval:
##  0.000 0.801
## sample estimates:
## probability of success
##                  0.611
```

We can conclude that $H_0$ cannot be rejected, since the p-value is higher than 0.05. We can set the original hypothesis $H_0$ as the percentage of cholesterol levels below 4.5 is less than 25%, while $H_1$ means the percentage of cholesterol levels below 4.5 is more than or equal to 25%.

```
below_4.5 = sum(data$After8weeks <= 4.5)
n = length(data$After8weeks)
binom.test(below_4.5,n,p=0.25,alt="g")
```

```
##
##  Exact binomial test
##
## data:  below_4.5 and n
## number of successes = 3, number of trials = 18, p-value = 0.9
## alternative hypothesis: true probability of success is greater than 0.25
## 95 percent confidence interval:
##  0.047 1.000
## sample estimates:
## probability of success
##                  0.167
```

Therefore, we cannot reject the original hypothesis $H_0$ since the p-value is larger than 0.05. We can conclude that the percentage of cholesterol levels below 4.5 is at most 25%. ## Exercise 2 **(a)** Before we perform relevant ANOVA models,we need to check whether the data meets the necessary assumptions.For example,we need to check normality for numeric columns.

```
df <- read.table("crops.txt", header = TRUE)
df$County <- as.factor(df$County);df$Related <- as.factor(df$Related)
shapiro.test(df$Crops);shapiro.test(df$Size)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$Crops
## W = 1, p-value = 0.3
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$Size
## W = 0.9, p-value = 0.001
```

From the result,the normaliy is doubtful for Crops.

We also check variance equality using leveneTest before applying anova model.As we can see p>0.05,thus we can apply anova models.

```
library(car)
```

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

```
leveneTest(Crops ~ County, data = df);leveneTest(Crops ~ Related, data = df);
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2     0.4   0.67
##       27
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1    1.09    0.3
##       28
```

```
leveneTest(Crops ~ County:Related, data = df)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  5     0.3   0.91
##       24
```

**ANOVA model** Without taking Size into account,there are two main factors:Country and Related,we also include the interaction factor:County+Related.

### Tests the independent effects

```
anova_model1 <- aov(Crops ~ County + Related, data = df) ;summary(anova_model1)
```

```
##             Df   Sum Sq Mean Sq F value Pr(>F)
## County       2 8.84e+06 4420721    0.82   0.45
## Related      1 2.38e+06 2378957    0.44   0.51
## Residuals   26 1.40e+08 5396286
```

### Interaction model

```
anova_model2 <- aov(Crops ~ County*Related, data = df) ;summary(anova_model2)
```

```
##                Df   Sum Sq Mean Sq F value Pr(>F)
## County          2 8.84e+06 4420721    0.76   0.48
## Related         1 2.38e+06 2378957    0.41   0.53
## County:Related  2 1.50e+06  748786    0.13   0.88
## Residuals      24 1.39e+08 5783578
```

Since for both model,all p-values are large (>0.05), we do not reject the null hypotheses, meaning there is no strong evidence that County or Related significantly impact Crops. As there is no significance for the interaction model,we choose the additive model to estimate the data.

```
new_farm <- data.frame(
  County = factor(3, levels = c(1, 2, 3)),
  Related = factor("no", levels = c("no", "yes"))
)
predicted_crops <- predict(anova_model1, newdata = new_farm, interval = "confidence");print(predicted_c
```

```
##    fit  lwr  upr
## 1 7760 6017 9504
```

The predicted crop yield for a farm in County 3 with no related factor is 7,760.However,the result may not be trustful.

Here we do the post-hoc check using TukeyHSD.

```
TukeyHSD(anova_model2)
```

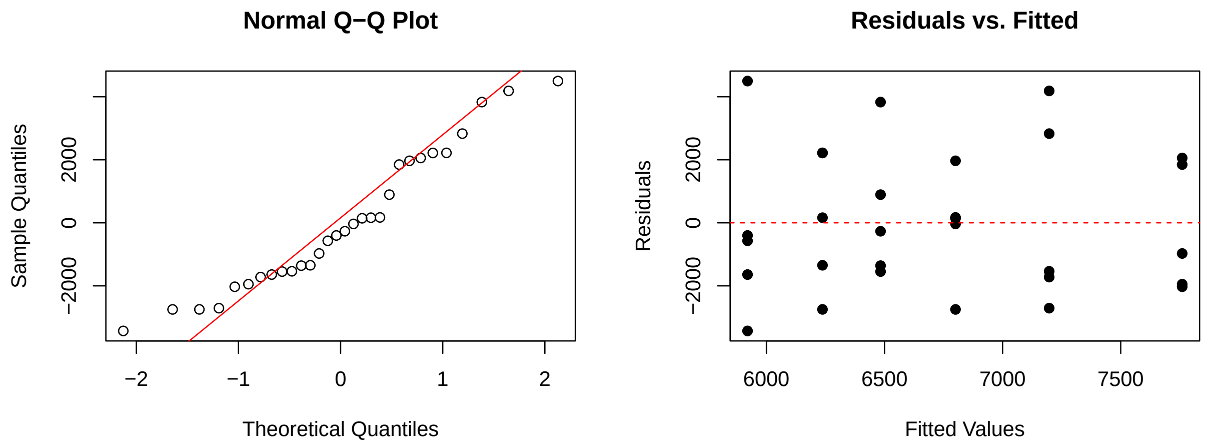
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Crops ~ County * Related, data = df)
##
## $County
##     diff   lwr  upr p adj
## 2-1 -317 -3003 2369 0.953
## 3-1  960 -1726 3646 0.650
## 3-2 1277 -1409 3963 0.472
```

```
##
## $Related
##         diff   lwr  upr p adj
## yes-no -563 -2376 1249 0.527
##
## $‘County:Related‘
##               diff   lwr  upr p adj
## 2:no-1:no      93 -4610 4796 1.000
## 3:no-1:no     851 -3852 5554 0.993
## 1:yes-1:no   -362 -5065 4341 1.000
## 2:yes-1:no  -1090 -5792 3613 0.978
## 3:yes-1:no    706 -3997 5409 0.997
## 3:no-2:no     758 -3945 5461 0.996
## 1:yes-2:no   -455 -5158 4248 1.000
## 2:yes-2:no  -1183 -5885 3520 0.969
## 3:yes-2:no    613 -4090 5316 0.998
## 1:yes-3:no  -1213 -5916 3490 0.965
## 2:yes-3:no  -1941 -6644 2762 0.795
## 3:yes-3:no   -145 -4848 4558 1.000
## 2:yes-1:yes  -728 -5430 3975 0.997
## 3:yes-1:yes  1068 -3635 5771 0.980
## 3:yes-2:yes  1796 -2907 6499 0.842
```

The Tukey test results confirm that the differences in means are not statistically significant.

Examining residuals can help assess if the model assumptions hold. We can see that normality is doubtful for residuals.Also,we can see from the right picture that it shows non-random pattern,which suggests that the model is not well-fitted and may omit variables.

```
par(mfrow = c(1,2))
residuals1 <- resid(anova_model1)
fitted_values1 <- fitted(anova_model1)
qqnorm(residuals1);qqline(residuals1, col="red")
plot(fitted_values1,residuals1 , main = "Residuals vs. Fitted",
     xlab = "Fitted Values", ylab = "Residuals", pch = 19)
abline(h = 0, col = "red", lty = 2)
```



```r
par(mfrow = c(1,1))
```

**(b)** ANOVA models only assumes that crops only depend on categorical factors,put size into perspective, we consider different ANCOVA models:

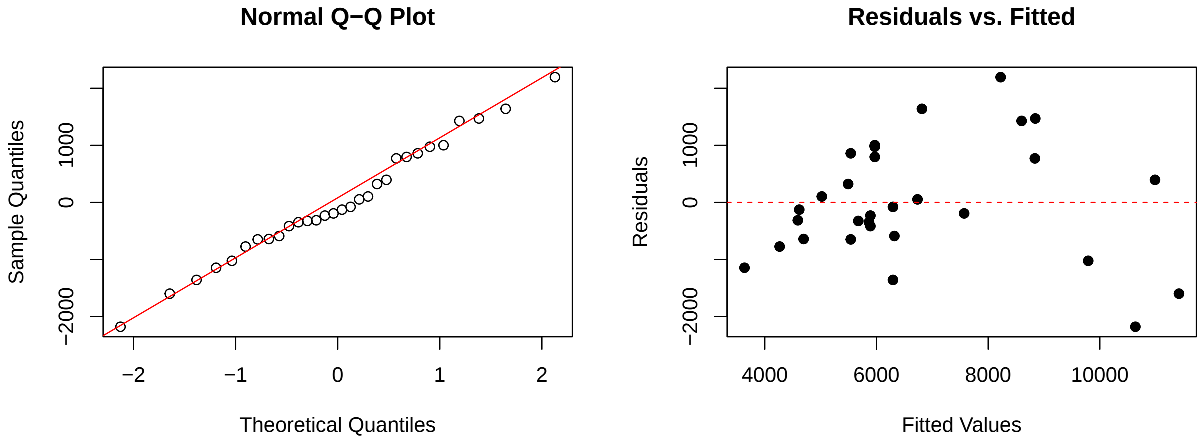
**Tests the independent effects**

```r
ancova_main <- aov(Crops ~ County + Related + Size, data = df);summary(ancova_main)
```

```
##             Df   Sum Sq  Mean Sq F value  Pr(>F)    
## County       2 8.84e+06 4.42e+06    3.71   0.039 *  
## Related      1 2.38e+06 2.38e+06    2.00   0.170    
## Size         1 1.10e+08 1.10e+08   92.68 6.9e-10 ***
## Residuals   25 2.98e+07 1.19e+06                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We check the model assumptions.Here we can see from the residuals that they are approximately normally distributed.

```r
par(mfrow = c(1,2))
residuals_main <- resid(ancova_main)
fitted_values_main <- fitted(ancova_main)
qqnorm(residuals_main);qqline(residuals_main, col="red")
plot(fitted_values_main,residuals_main , main = "Residuals vs. Fitted",
     xlab = "Fitted Values", ylab = "Residuals", pch = 19)
abline(h = 0, col = "red", lty = 2)
```



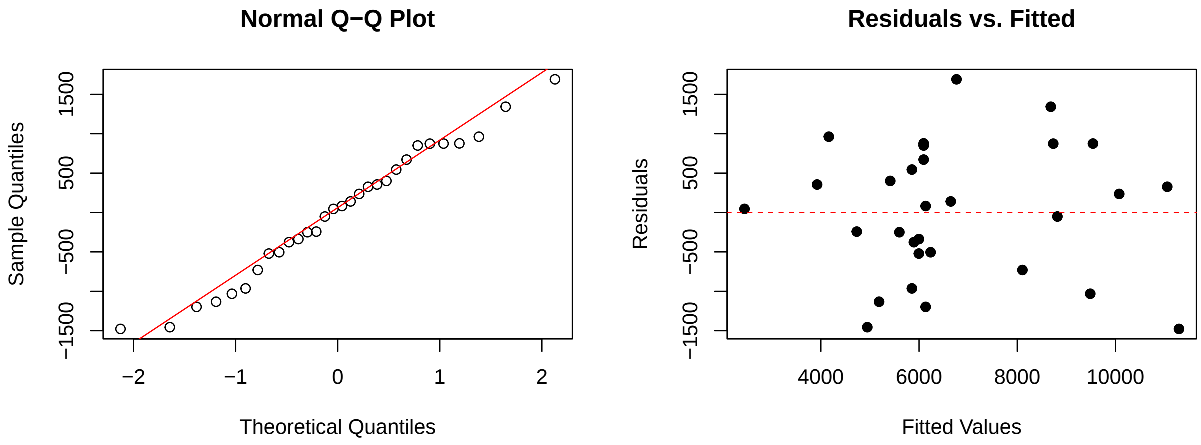
```r
par(mfrow = c(1,1))
```

**Size × County Interaction**

```r
ancova_county_size <- aov(Crops ~ County * Size + Related, data = df);summary(ancova_county_size)
```

```
##               Df   Sum Sq  Mean Sq F value Pr(>F)
## County         2 8.84e+06 4.42e+06    5.01  0.016 *
## Size           1 1.11e+08 1.11e+08  126.47  8e-11 ***
## Related        1 1.38e+06 1.38e+06    1.57  0.223
## County:Size    2 9.53e+06 4.76e+06    5.40  0.012 *
## Residuals     23 2.03e+07 8.82e+05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow = c(1,2))
residuals_county_size <- resid(ancova_county_size)
fitted_values_county_size <- fitted(ancova_county_size)
qqnorm(residuals_county_size);qqline(residuals_county_size, col="red")
plot(fitted_values_county_size,residuals_county_size , main = "Residuals vs. Fitted",
     xlab = "Fitted Values", ylab = "Residuals", pch = 19)
abline(h = 0, col = "red", lty = 2)
```



```
par(mfrow = c(1,1))
```

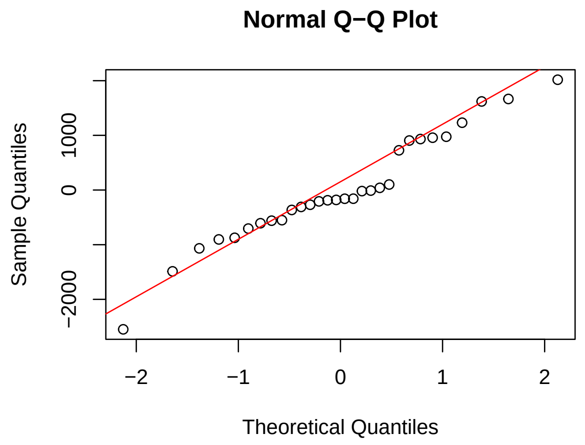
### Size × Related Interaction

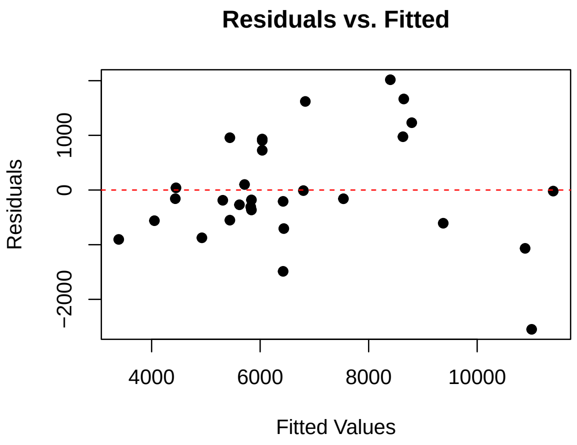
```
ancova_related_size <- aov(Crops ~ County + Related * Size, data = df);summary(ancova_related_size)
```

```
##               Df   Sum Sq  Mean Sq F value   Pr(>F)
## County         2 8.84e+06 4.42e+06    3.73   0.039 *
## Related        1 2.38e+06 2.38e+06    2.01   0.169
## Size           1 1.10e+08 1.10e+08   93.21 9.7e-10 ***
## Related:Size   1 1.35e+06 1.35e+06    1.14   0.296
## Residuals     24 2.85e+07 1.19e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow = c(1,2))
residuals_related_size <- resid(ancova_related_size)
fitted_values_related_size <- fitted(ancova_related_size)
```

```
qqnorm(residuals_related_size);qqline(residuals_related_size, col="red")
plot(fitted_values_related_size,residuals_related_size , main = "Residuals vs. Fitted",
     xlab = "Fitted Values", ylab = "Residuals", pch = 19)
abline(h = 0, col = "red", lty = 2)
```





```
par(mfrow = c(1,1))
```

```
shapiro.test(residuals(ancova_main));shapiro.test(residuals(ancova_related_size))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(ancova_main)
## W = 1, p-value = 1

##
##  Shapiro-Wilk normality test
##
## data:  residuals(ancova_related_size)
## W = 1, p-value = 0.3
```

```
shapiro.test(residuals(ancova_county_size))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(ancova_county_size)
## W = 1, p-value = 0.7
```

We can see from p-values that Size × County interaction model has the lowest Residual Sum of Squares,from QQplot the normaliry is doubtful. But using Shapiro-Wilk normality test suggests that normality assumption holdsl.Thus we choose the Size × County interaction model.

**(c)** We can review the ANCOVA Model Results.

Significant County ($p < 0.05$) → County influences on Crops,meaning crop yields differ across counties..

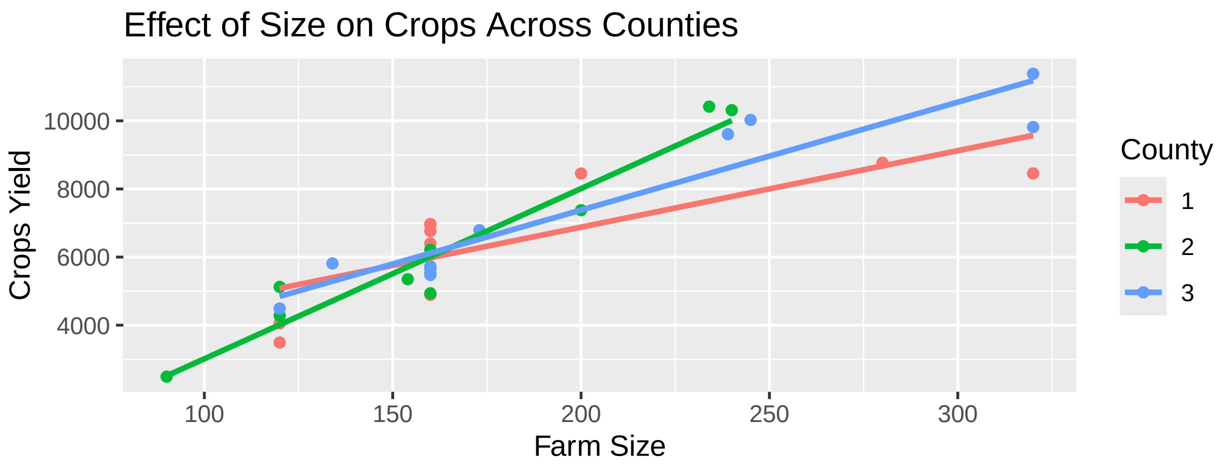
Significant Size ($p < 0.05$) → Farm Size has a strong effect on Crops,which makes sense—larger farms generally produce more.

Significant Related ($p > 0.05$) → Whether the landlord and tenant are related does not significantly impact the crops.

Significant County × Size ($p < 0.05$) → This suggests that the effect of farm size on crop yield depends on the county. The figure blow confirms this conclusion,as the slopes are different.

```r
library(ggplot2)
ggplot(df, aes(x = Size, y = Crops, color = County)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Effect of Size on Crops Across Counties",
       x = "Farm Size", y = "Crops Yield")
```

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



**(d)**

```r
new_farm <- data.frame(
  County = factor(2, levels = c(1, 2, 3)),
  Related = factor("yes", levels = c("no", "yes")),
  Size = 165
)
predicted_crops <- predict(ancova_county_size, newdata = new_farm, interval = "confidence")
print(predicted_crops)
```

```
##    fit  lwr  upr
## 1 6141 5428 6855
```

We estimate the error variance.

```r
error_variance <- sum(residuals(ancova_county_size)^2) / df.residual(ancova_county_size)
print(error_variance)
```

```
## [1] 881623
```

## Exercise 3

### (a)

```
library(MASS)
blocks = rep(1:6, each = 4)

random_N = unlist(tapply(rep(c(1, 1, 0, 0), 6), blocks, sample))
random_P = unlist(tapply(rep(c(1, 1, 0, 0), 6), blocks, sample))
random_K = unlist(tapply(rep(c(1, 1, 0, 0), 6), blocks, sample))

N_matrix <- matrix(random_N, nrow = 6, byrow = TRUE)
P_matrix <- matrix(random_P, nrow = 6, byrow = TRUE)
K_matrix <- matrix(random_K, nrow = 6, byrow = TRUE)

random_result = cbind(blocks, N_matrix, P_matrix, K_matrix)
```

```
## Warning in cbind(blocks, N_matrix, P_matrix, K_matrix): number of rows of
## result is not a multiple of vector length (arg 1)
```

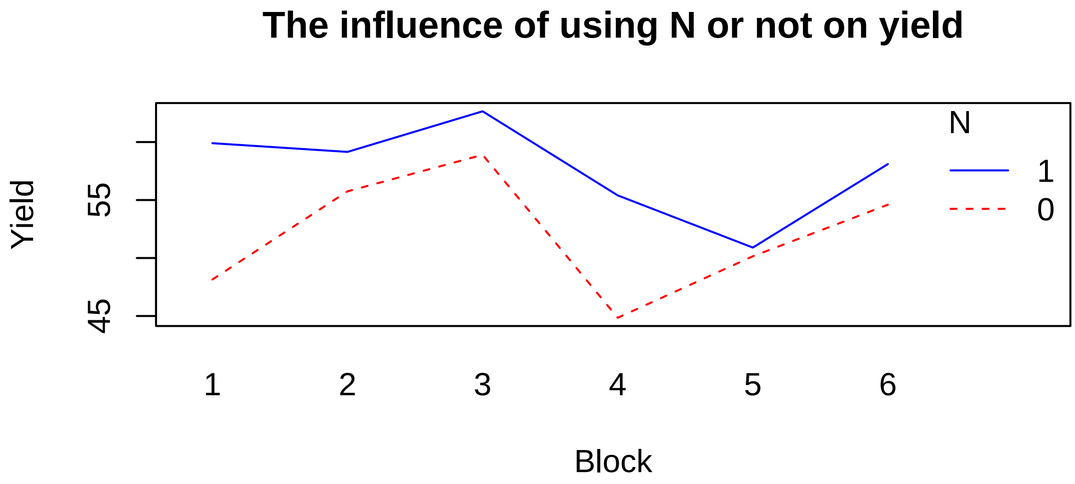
```
colnames(random_result) <- c("Block", "N1", "N2", "N3", "N4",
                             "P1", "P2", "P3", "P4", "K1", "K2", "K3", "K4")

random_result
```

```
##      Block N1 N2 N3 N4 P1 P2 P3 P4 K1 K2 K3 K4
## [1,]     1  0  1  0  1  0  0  1  1  0  0  1  1
## [2,]     1  1  0  0  1  0  1  0  1  0  1  1  0
## [3,]     1  0  1  0  1  0  1  0  1  0  1  0  1
## [4,]     1  0  1  1  0  1  0  0  1  0  1  1  0
## [5,]     2  0  1  0  1  1  0  1  0  1  1  0  0
## [6,]     2  1  1  0  0  1  0  1  0  1  1  0  0
```

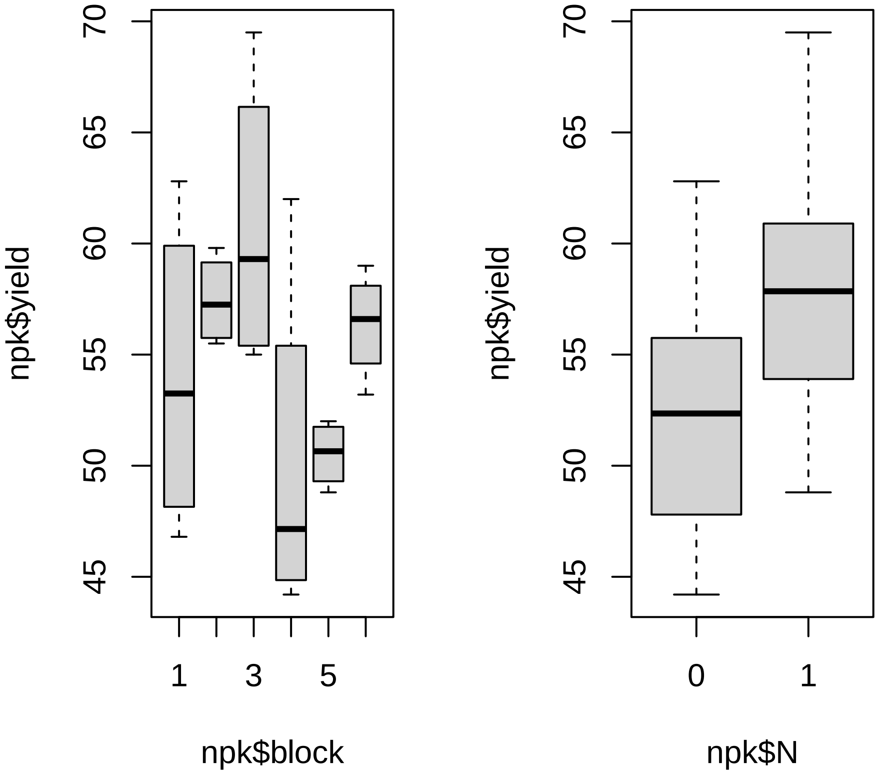
**(b)** By generating the **interaction plot**, we can know that the mean yield of blocks using N is always higher than the blocks without using N. As interaction shows up as nonparallel curves,it looks like interaction seems to be not present.

```
interaction.plot(npk$block, npk$N, npk$yield, xlab = 'Block', ylab = "Yield",
trace.label = "N", col = c("red", "blue"),
main = "The influence of using N or not on yield")
```



**(c)** First we look at the boxplot.

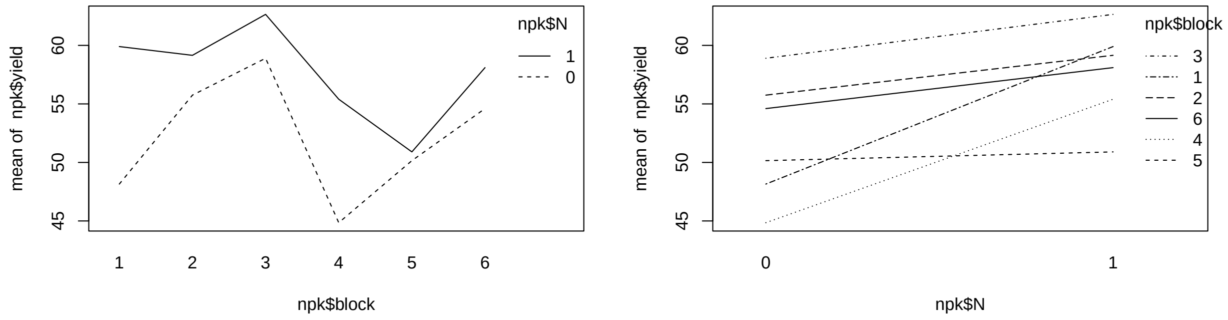
```r
par(mfrow = c(1,2))
boxplot(npk$yield~npk$block);boxplot(npk$yield~npk$N);
```



From the interaction, we

can observe that the lines seem parallel, so interaction might be no present.

```
par(mfrow = c(1,2))
interaction.plot(npk$block, npk$N, npk$yield);interaction.plot(npk$N, npk$block, npk$yield)
```



```
npk$block = as.factor(npk$block); npk$N = as.factor(npk$N);
npkaov = lm(yield~block*N, data=npk);anova(npkaov)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value Pr(>F)
## block      5    343    68.7    3.36   0.04 *
## N          1    189   189.3    9.26   0.01 *
## block:N    5     99    19.7    0.96   0.48
## Residuals 12    245    20.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the result, there is no evidence for interaction between block and N, since the p-value of it is 0.48($>$ 0.05). As the result of no interaction, we can remove interaction term from the model and fit the additive model.

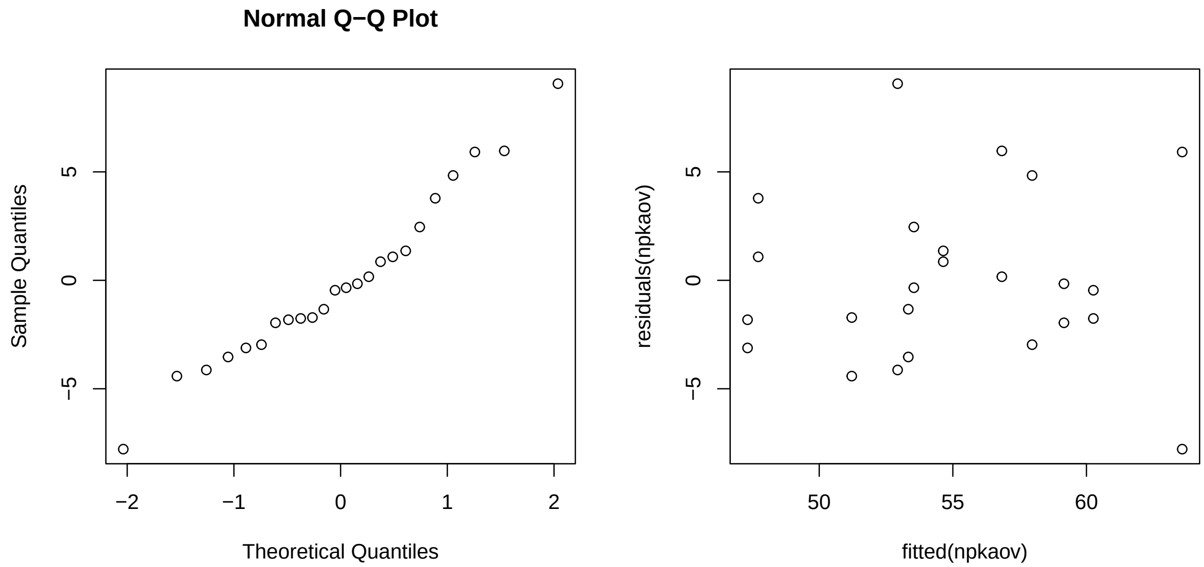
```
npk$block = as.factor(npk$block); npk$N = as.factor(npk$N);
npkaov = lm(yield~block+N, data=npk);anova(npkaov)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value Pr(>F)
## block      5    343    68.7    3.40 0.0262 *
## N          1    189   189.3    9.36 0.0071 **
## Residuals 17    344    20.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Then we can observe the p-value of each factor, it is known that both factors have a main effect in the additive model, since the p-values are smaller than 0.05.

We check the model assumptions.The Q-Q plot shows the normality.

```r
par(mfrow = c(1,2))
qqnorm(residuals(npkaov)); plot(fitted(npkaov),residuals(npkaov))
```



The inclusion of Block factor is sensible for this model,because the block effect is significant in both models ($p < 0.05$ ), which means that different blocks have different average yields.

In this scenario,it’s not suitable to use Friedman test,as from the dataset we can see that it is a replicated Complete Block Design(N>1).But actually we can do the aggregrate operation to meet the requirement.

**(d)** From the results, we can know the pair values of each pair are all higher than 0.05, which means that there is no evidence for interaction in this pairs (N:block, P:block and K:block). Here, we can see N, K and block all present main effects, but we still cannot conclude this now.

```r
model_N_block <- lm(yield ~ N * block + P + K, data = npk)
model_P_block <- lm(yield ~ P * block + N + K, data = npk)
model_K_block <- lm(yield ~ K * block + N + P, data = npk)
anova(model_N_block);anova(model_P_block);anova(model_K_block)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value Pr(>F)
## N          1    189   189.3   13.36 0.0044 **
## block      5    343    68.7    4.85 0.0164 *
## P          1      8     8.4    0.59 0.4590
## K          1     95    95.2    6.72 0.0268 *
## N:block    5     99    19.7    1.39 0.3066
## Residuals 10    142    14.2
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
## Analysis of Variance Table
##
```

```
## Response: yield
##           Df Sum Sq Mean Sq F value Pr(>F)
## P          1      8     8.4    0.50 0.4966
## block      5    343    68.7    4.07 0.0282 *
## N          1    189   189.3   11.21 0.0074 **
## K          1     95    95.2    5.64 0.0389 *
## P:block    5     71    14.3    0.85 0.5473
## Residuals 10    169    16.9
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value Pr(>F)
## K          1     95    95.2    5.60 0.0395 *
## block      5    343    68.7    4.04 0.0288 *
## N          1    189   189.3   11.14 0.0075 **
## P          1      8     8.4    0.49 0.4980
## K:block    5     70    14.1    0.83 0.5583
## Residuals 10    170    17.0
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```r
additive_model <- lm(npk$yield ~ npk$N + npk$block + npk$P + npk$K)
anova(additive_model)
```

```
## Analysis of Variance Table
##
## Response: npk$yield
##           Df Sum Sq Mean Sq F value Pr(>F)
## npk$N      1    189   189.3   11.82 0.0037 **
## npk$block  5    343    68.7    4.29 0.0127 *
## npk$P      1      8     8.4    0.52 0.4800
## npk$K      1     95    95.2    5.95 0.0277 *
## Residuals 15    240    16.0
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Therefore,we can conclude that P, K and block present significant effects in additive model, while P has no main effect in the additive model. Because in the first anova test, we cannot see interaction in each pair, so the additive model is better than the previous model with interaction terms.

**(e)** According to the results in **d**, we know P and K has the main effects in the model. The combination of P and K can contribute to the yield most.

**f** Our main question of interest is whether nitrogen $N$ has an effect on *yield*. The mixed model processes the block variable as a random effect. Then, we shouldn create a model without N factor. Therefore, the Pr(>Chisq) is 0.0012, which means that the effect of N factor is significant. The results are the same as the fixed model.

```r
library(lme4)
```

```
## Loading required package: Matrix
```

```r
mixed_model <- lmer(npk$yield ~ npk$N + npk$P + npk$K + (1|npk$block), REML=FALSE)
summary(mixed_model)
```

```
## Linear mixed model fit by maximum likelihood  [’lmerMod’]
## Formula: npk$yield ~ npk$N + npk$P + npk$K + (1 | npk$block)
##
##      AIC      BIC   logLik deviance df.resid
##    151.0    158.1    -69.5    139.0       18
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.9706 -0.6831  0.0554  0.7124  1.4716
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  npk$block (Intercept) 11.0     3.31
##  Residual              13.3     3.65
## Number of obs: 24, groups:  npk$block, 6
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)    54.65       2.01   27.15
## npk$N1          5.62       1.49    3.77
## npk$P1         -1.18       1.49   -0.79
## npk$K1         -3.98       1.49   -2.67
##
## Correlation of Fixed Effects:
##        (Intr) npk$N1 npk$P1
## npk$N1 -0.370
## npk$P1 -0.370  0.000
## npk$K1 -0.370  0.000  0.000
```

```r
mixed_without_N <- lmer(npk$yield ~ npk$P + npk$K + (1|npk$block), REML=FALSE)
anova(mixed_model, mixed_without_N)
```

```
## Data: NULL
## Models:
## mixed_without_N: npk$yield ~ npk$P + npk$K + (1 | npk$block)
## mixed_model: npk$yield ~ npk$N + npk$P + npk$K + (1 | npk$block)
##                 npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)
## mixed_without_N    5 160 165  -74.7      150
## mixed_model        6 151 158  -69.5      139  10.5  1     0.0012 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```