Experimental Design and Data Analysis - Assignment 1

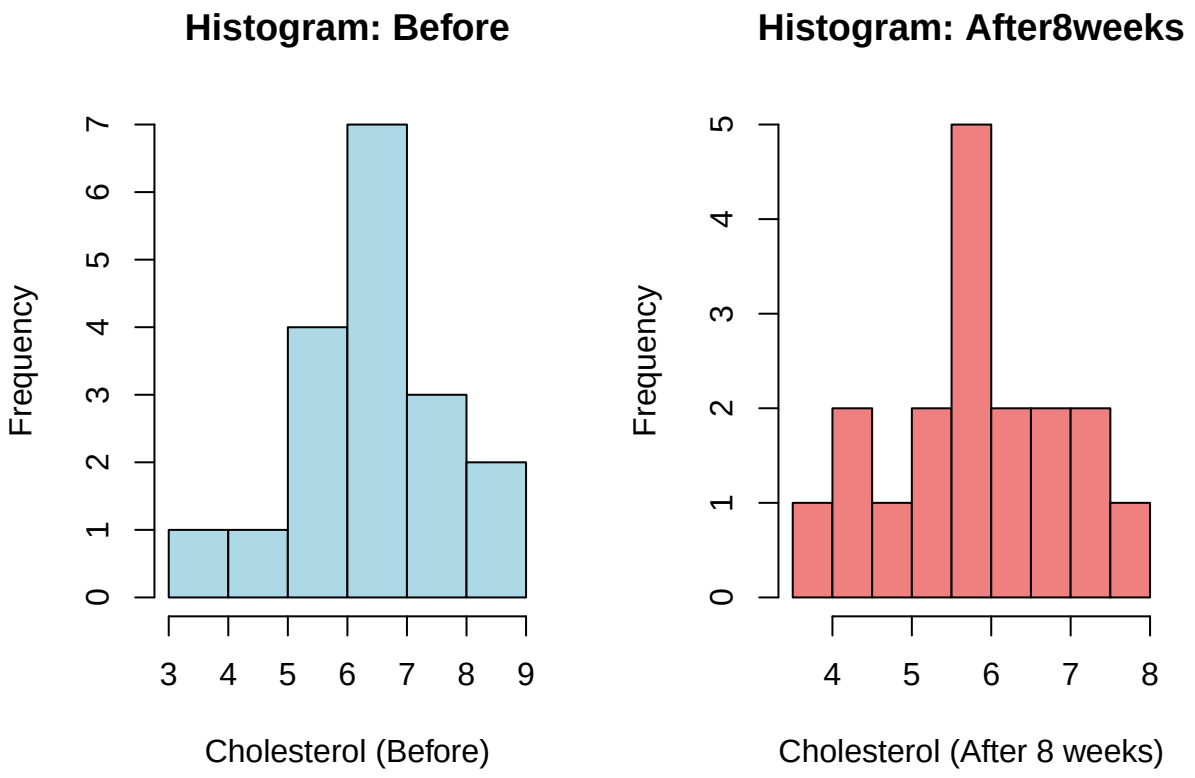
Group 5 - Ivana Malčić, Xuening Tang, Xiaoxuan Zhang

2025-02-23

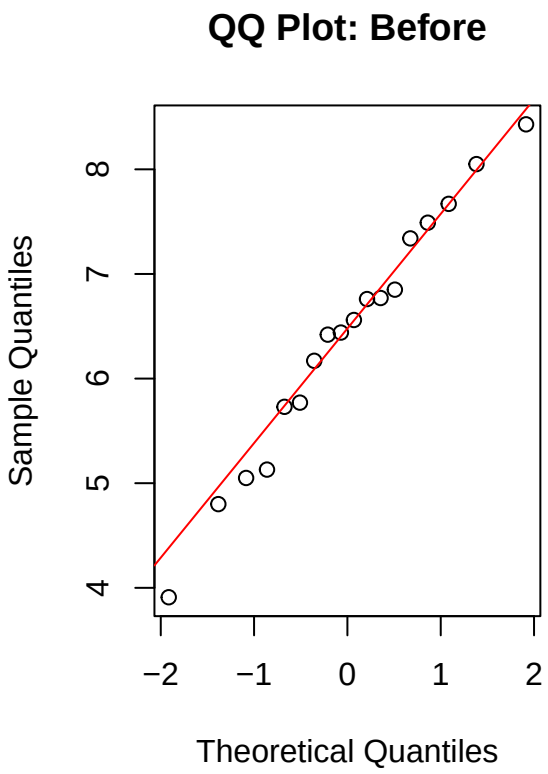
In order not to be bothered with rounding the numbers, set `options(digits=3)r options(digits=3)`.

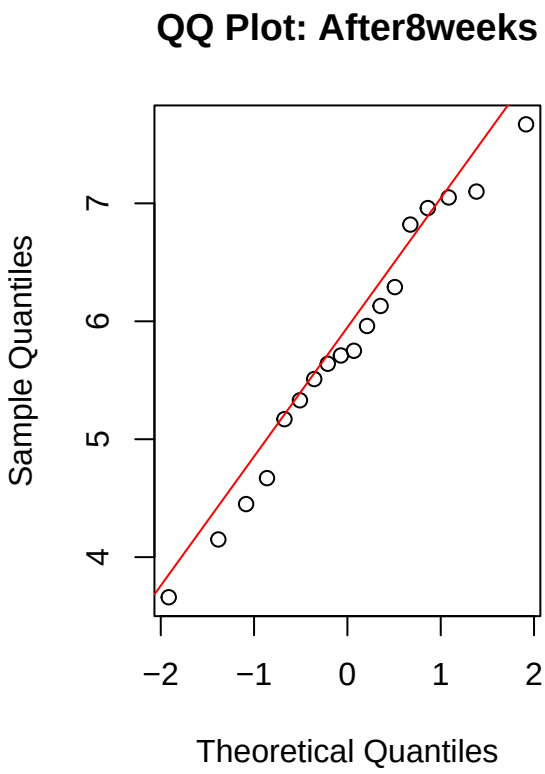
Exercise 1: Cholesterol

a) In this first section, both normality and variable correlation are explored using relevant plots and metrics. Firstly, the bell-like shape of the histograms indicates that the data is normally distributed.



The previous finding is further confirmed by the following QQ-plots where the data points seem relatively close to the reference line, again signaling normality.





Additional data exploration gives us further insight; the close mean and median signify symetric distribution, a feature which is also a common attribute of normality. Moreover, the skewness for both variables tells us that the left tail is slightly longer (distribution skewed to the left). Finally, kurtosis of 2.5 and 2.27 indicates a peaked distribution with less outliers and a more or less uniform distribution.

Table 1: Descriptive Statistics for Cholesterol Levels

Variable	Mean	Median	Skewness	Kurtosis
Before	6.41	6.50	-0.28	2.50
After8weeks	5.78	5.73	-0.17	2.27

After normality assesment, we turn to look at whether the two variables are correlated. For this we first utilize a simple scatterplot shown below which exhibits strong positive correlation visible by the densly clustered data points around the rising regression line.

Then, Pearson´s test is employed - the correlation coefficient of 0.991 indicates a strong and positive linear relationship between the two variables. Furthermore, the small p-value (<0.001) suggests this relationship is statistically significant, and therefore we can reject the null hypothesis of no correlation.

Table 2: Pearson Correlation Test Results

Statistic	Value
Correlation Coefficient	0.991
P-value	0
Confidence Interval	0.975 to 0.997

b) Now, our goal is to establish whether the low-fat margarine diet had any effect on cholesterol by utilizing 2 relevant test metrics. Since our data is paired, we first utilize a paired t-test. The large t-statistic and small p-value ($p < 0.001$) provide strong evidence against the null hypothesis of no difference. Additionally, the confidence interval suggests that the mean cholesterol level after 8 weekes lies somewhere between 0.54 and 0.718 with 95% confidence.

Table 3: Paired t-Test Results

Statistic	Value
t-statistic	14.946
Degrees of Freedom	17
P-value	0
Confidence Interval	0.54 to 0.718

Since our data are paired and normally distributed, the Mann-Whitney U test is not applicable in this scenario. However, we can apply the permutation test which is useful because it works well with small data volumes. The following permutation table reveals a similar trend as previously discussed with a statistically significant ($p<0.001$) average decrease in cholesterol levels by 0.629 units after the 8 week intervention.

Table 4: Permutation Test Results

Statistic	Value
Observed Mean Difference	0.629
Permutation Test P-value	0.000

c) Next, we are constructing a 97% CI and 97% bootstrapped CI, as opposed to our previously used 95% CI. As visible from *Table 5*, we can be 97% confident our true population parameter is encompased between the ranges of [5.16, 6.39] for normal CI and [5.23, 6.32] for the bootstrapped CI.

Table 5: 97% Confidence Intervals for Mean

	Method	Lower_Bound	Upper_Bound
	Normality (t-distribution)	5.16	6.39
1.5%	Bootstrap	5.23	6.32

d) Additionally, we use bootstrapping to come up with a 97% confidence interval for the maximum statistic for various candidate values of , helping us reject or not reject the hypothesis that the data follow a Uniform[3,] distribution. *Table 6* provides us with plausible candidate values for which we cannot reject the Null hypothesis. Kolmogorov-Smirnov test can also be applied in this case to test whether the data follows a uniform distribution.

Table 6: Non-Rejected Theta Values

Theta	Lower Bound	Upper Bound
7.7	6.72	7.70
7.8	6.81	7.80
7.9	6.90	7.90
8.0	6.98	8.00
8.1	7.07	8.10
8.2	7.13	8.20
8.3	7.20	8.30
8.4	7.29	8.39
8.5	7.36	8.50
8.6	7.45	8.60
8.7	7.52	8.70
8.8	7.57	8.79
8.9	7.68	8.90

Kolmogorov-Smirnov test can also be applied in this case to test whether the data follows a uniform distribution.

Table 7: Theta Values with Non-Rejected KS Test

Theta	P-Value
7.0	0.038
7.1	0.053
7.2	0.071
7.3	0.092
7.4	0.117
7.5	0.146
7.6	0.179
7.7	0.216
7.8	0.256
7.9	0.299
8.0	0.345
8.1	0.394
8.2	0.444
8.3	0.495
8.4	0.509
8.5	0.419
8.6	0.342
8.7	0.277
8.8	0.223
8.9	0.179
9.0	0.143
9.1	0.114
9.2	0.091
9.3	0.072
9.4	0.057
9.5	0.045
9.6	0.036

e) Finally, we are testing the following Null hypothesis: *Null hypothesis (H_0)*: The median cholesterol level after 8 weeks is 6. With the results presented below we can conclude there is not enough statistical evidence to conclude that the median cholesterol level after 8 weeks is less than 6. While 61.1% of the sample is below 6, this deviation could easily be due to random variation given the sample size ($p>0.1$).

Table 8: Median Test Results (H : median = 6)

Statistic	Value
Sample Size	18
Number < 6	11
Observed Proportion	0.611
p-value	0.24
95% CI	0.392 to 1

Subsequently, our second Null hypothesis goes as following: *Null hypothesis (H_0)*: the fraction of cholesterol levels below 4.5 is at most 0.25. Similarly, we also cannot reject this hypothesis because of the very high p-value ($p>0.1$) and a wide CI.

Table 9: Fraction Test Results (H : fraction below 4.5 is 25%)

Statistic	Value
Sample Size	18
Number < 4.5	3
Observed Proportion	0.167
p-value	0.865
95% CI	0.047 to 1.000

Exercise 2

Section a To study the effect of County and Related on the variable Crops, we propose the following hypotheses:
H_{AB} There is no interaction effect of *County and Related*.
H_A There is no main effect of factor *County*.
H_B There is no main effect of factor *Related*.

```
library(dplyr)
crop_data <- read.delim("crops.txt", sep = " ") # read the dataset

cropframe = data.frame(
  crops = crop_data$Crops,
  county = factor(crop_data$County),
  related = factor(crop_data$Related)) # turn variables "county" and "related" into factors; store rele

cropframe <- cropframe %>% mutate(related = recode(related, 'yes' = '1', 'no' = '0')) # recode the "rel
```

```
cropframe # display the frame
```

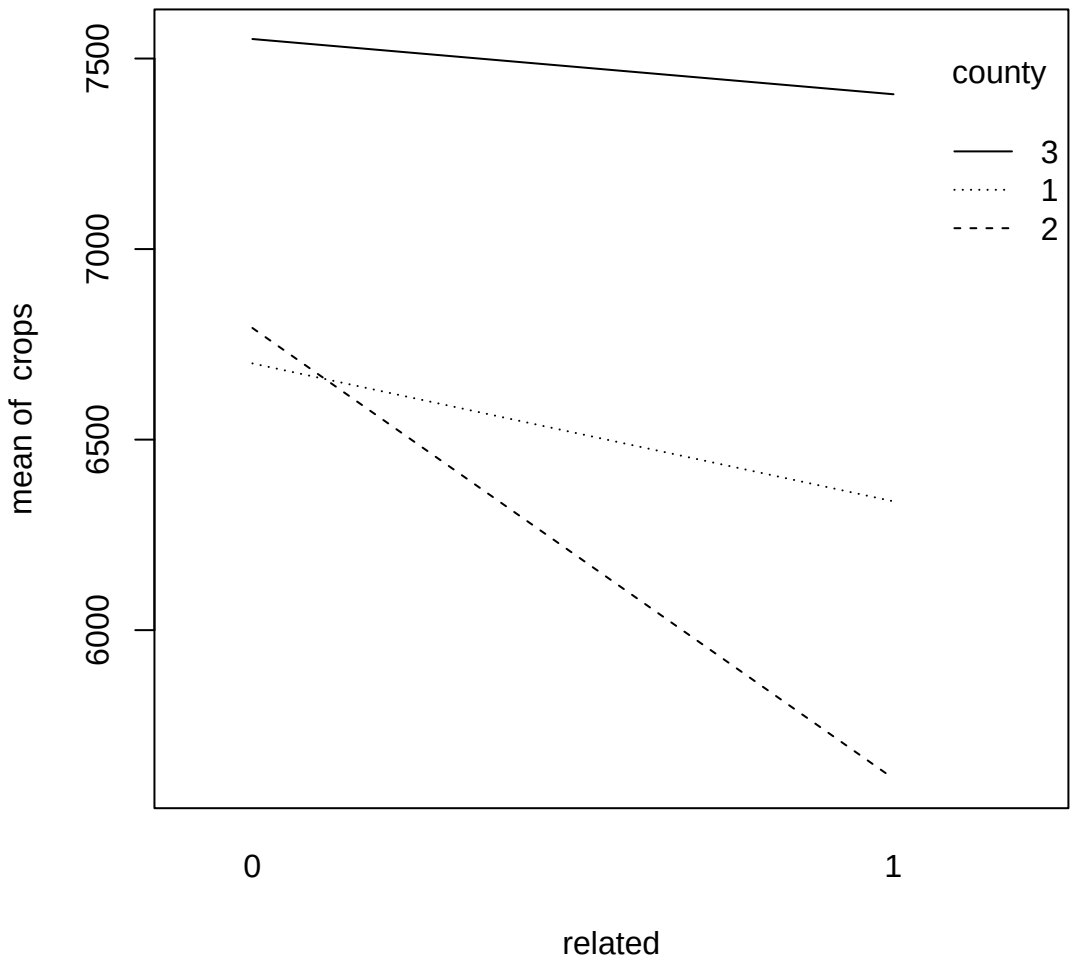
```
##    crops county related
## 1   6399      1       1
## 2   8456      1       1
## 3   8453      1       1
## 4   4891      1       1
## 5   3491      1       1
## 6   6944      1       0
## 7   6971      1       0
## 8   4053      1       0
## 9   8767      1       0
## 10  6765      1       0
## 11  2490      2       1
## 12  5349      2       1
## 13  5518      2       1
## 14 10417      2       1
## 15  4278      2       1
## 16  4936      2       0
## 17  7376      2       0
## 18  6216      2       0
## 19 10313      2       0
## 20  5124      2       0
## 21  4489      3       1
```

```
## 22 10026      3       1
## 23  5659      3       1
## 24  5475      3       1
## 25 11382      3       1
## 26  5731      3       0
## 27  6787      3       0
## 28  5814      3       0
## 29  9607      3       0
## 30  9817      3       0
```

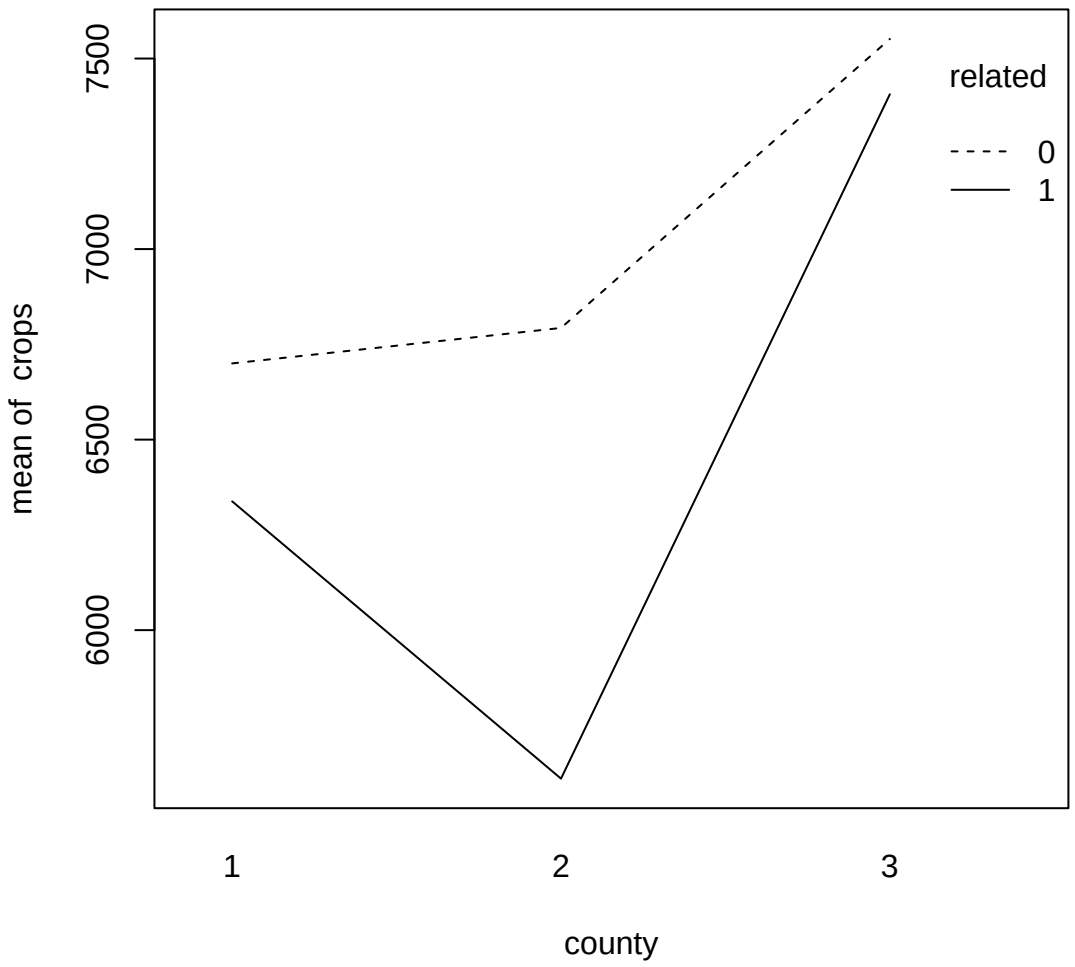
```
crops = cropframe$crops
county = cropframe$county
related = cropframe$related
```

Before conducting the ANOVA test, we first plotted two interaction plots to get a first glimpse of the potential interaction effect. Based on the two interaction plots, it seems there is little interaction effect, as the lines are parallel in general. We then conduct a two-way ANOVA to confirm our observation.

```
interaction.plot(related,county,crops) # fix county
```




```
interaction.plot(county,related,crops) # fix related
```




```
is.factor(county) # check if county and related are factors
```

```
## [1] TRUE
```

```
is.factor(related)
```

```
## [1] TRUE
```

```
cropanov=lm(crops~county*related); anova(cropanov) # conduct the two-way ANOVA test
```

```
## Analysis of Variance Table
##
## Response: crops
##                Df   Sum Sq Mean Sq F value Pr(>F)
## county          2 8.84e+06 4420721    0.76   0.48
```

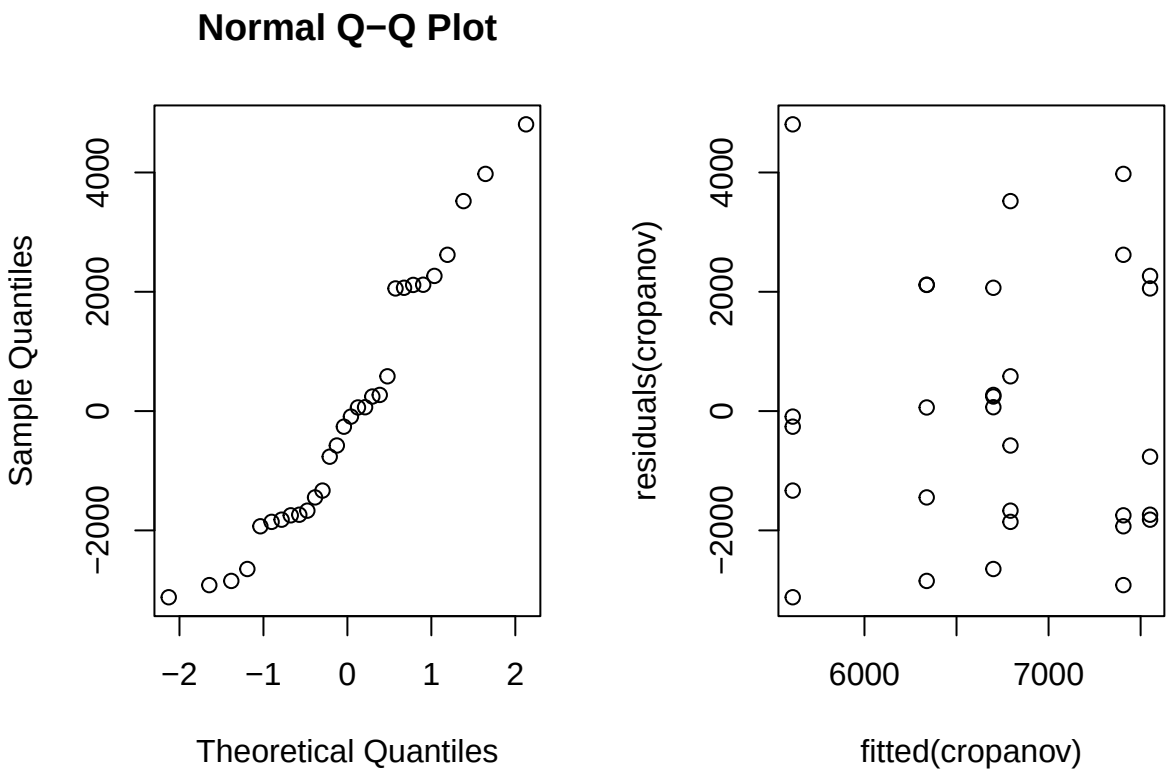
```
## related         1 2.38e+06 2378957    0.41   0.53
## county:related  2 1.50e+06  748786    0.13   0.88
## Residuals      24 1.39e+08 5783578
```

```
summary(cropanov)
```

```
##
## Call:
## lm(formula = crops ~ county * related)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3120  -1745   -177   2064   4807
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6700       1076    6.23  1.9e-06 ***
## county2                93       1521    0.06     0.95
## county3               851       1521    0.56     0.58
## related1             -362       1521   -0.24     0.81
## county2:related1     -821       2151   -0.38     0.71
## county3:related1      217       2151    0.10     0.92
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2400 on 24 degrees of freedom
## Multiple R-squared:  0.0839, Adjusted R-squared:  -0.107
## F-statistic: 0.44 on 5 and 24 DF,  p-value: 0.816
```

Result shows that there is no interaction effect between *Related* and *County* on *Crops*. None of the p values for *County*, *Crops* and *County:Related* are significant ($p = 0.477$; $p = 0.527$; $p = 0.879$). To make sure that this result is valid, we plot a Q-Q plot and residual plot. Based on the two plots, the assumption of normality is met: Q-Q plot gives a straight line in general, and the residuals display no pattern.

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



In the next step, we remove the interaction and apply an additive model. The code and results are shown below:

```
cropanov2=lm(crops~county+related,data=cropframe); anova(cropanov2) # additive model
```

```
## Analysis of Variance Table
##
## Response: crops
##           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
```

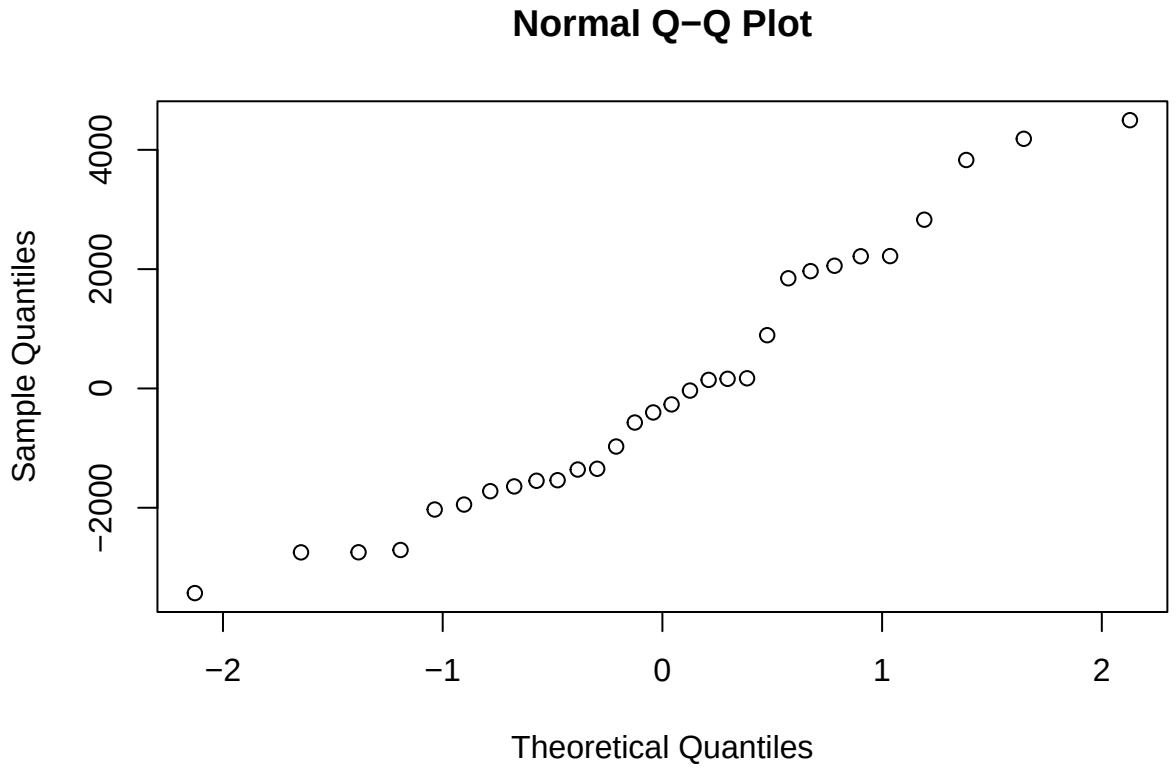
```
summary(cropanov2)
```

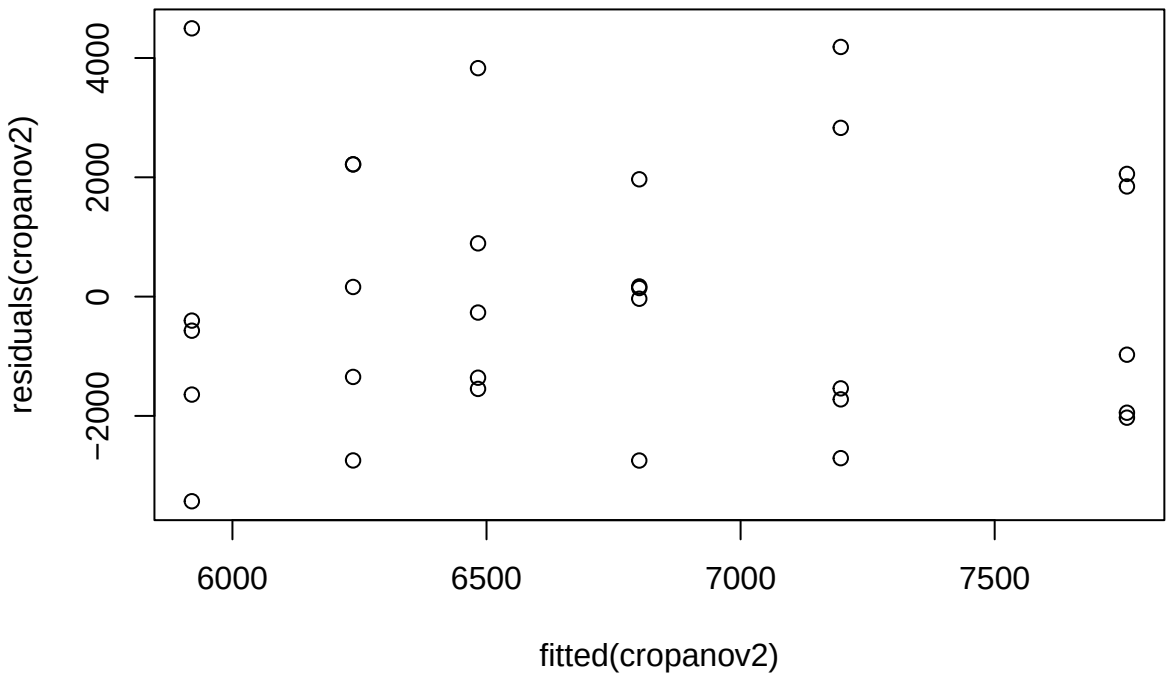
```
##
## Call:
## lm(formula = crops ~ county + related, data = cropframe)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3430  -1618   -335   1936   4497
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)     6801        848    8.02  1.7e-08 ***
## county2         -317       1039   -0.31     0.76
## county3          960       1039    0.92     0.36
## related1        -563        848   -0.66     0.51
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2320 on 26 degrees of freedom
## Multiple R-squared:  0.0741, Adjusted R-squared:  -0.0328
## F-statistic: 0.693 on 3 and 26 DF,  p-value: 0.565
```

The result of the additive model shows that neither of the factors has a significant main effect on Crops. The p-values are 0.4518 and 0.5126 for *County* and *Related* respectively, and are larger than the 0.05 significance level in both cases. Therefore, we fail to reject none of our hypotheses. The normality assumption of this ANOVA test is also met based on the following Q-Q plot and residual plot:

```
qqnorm(residuals(cropanov2)); plot(fitted(cropanov2),residuals(cropanov2))
```





Summary for the decisions to the null hypotheses:

Hypothesis	Decision
H_{AB}	not reject
H_A	not reject
H_B	not reject

The mathematical formula for a two-way ANOVA model is:

$$Y_{ijk} = \mu_{ij} + e_{ijk}$$

where

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

μ is the overall mean

α_i is the main effect of level i of the factor *County*, i = 1,2,3

β_j is the main effect of level j of the factor *Related*, j = 0,1

$\gamma_i j$ is the interaction effect of levels i, j of factor *County* and *Related*, which is 0 in this case, since there is no significant interaction effect.

We apply the model **cropanov2** for prediction.Therefore, the crops in *County 3* for which there is no related is:

$$Crops = Intercept + County3 + Related0 = 6800.6 + 959.7 + 0 = 7760.3$$

Therefore, the predicted value of the Crops is 7760.3.

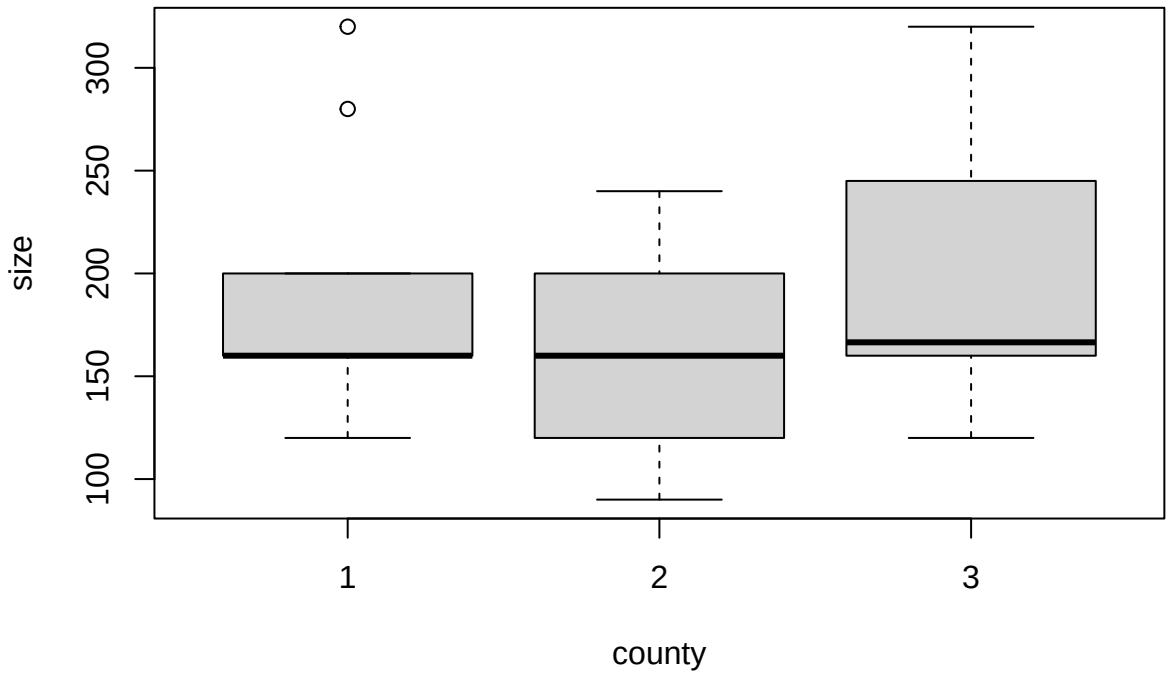
Section b We now add the variable *Size*. Since it is a numerical variable, it should be treated as an exploratory variable. We want to find out how *Size* influences the effect of *Related* or *County* on *Crops* in our model.

(1) ANCOVA test: *Size * County*

H_{AB} There is no interaction effect between *Size* and *County* on *Crops*

We first get a glimpse of the distribution of size in different counties. It seems the distributions are different in different counties. We need to confirm our observation through a two-way ANCOVA model.

```
size = crop_data$Size
boxplot(size~county)
```




```
cropanov3=lm(crops~size*county,data=cropframe);anova(cropanov3)
```

```
## Analysis of Variance Table
##
## Response: crops
```

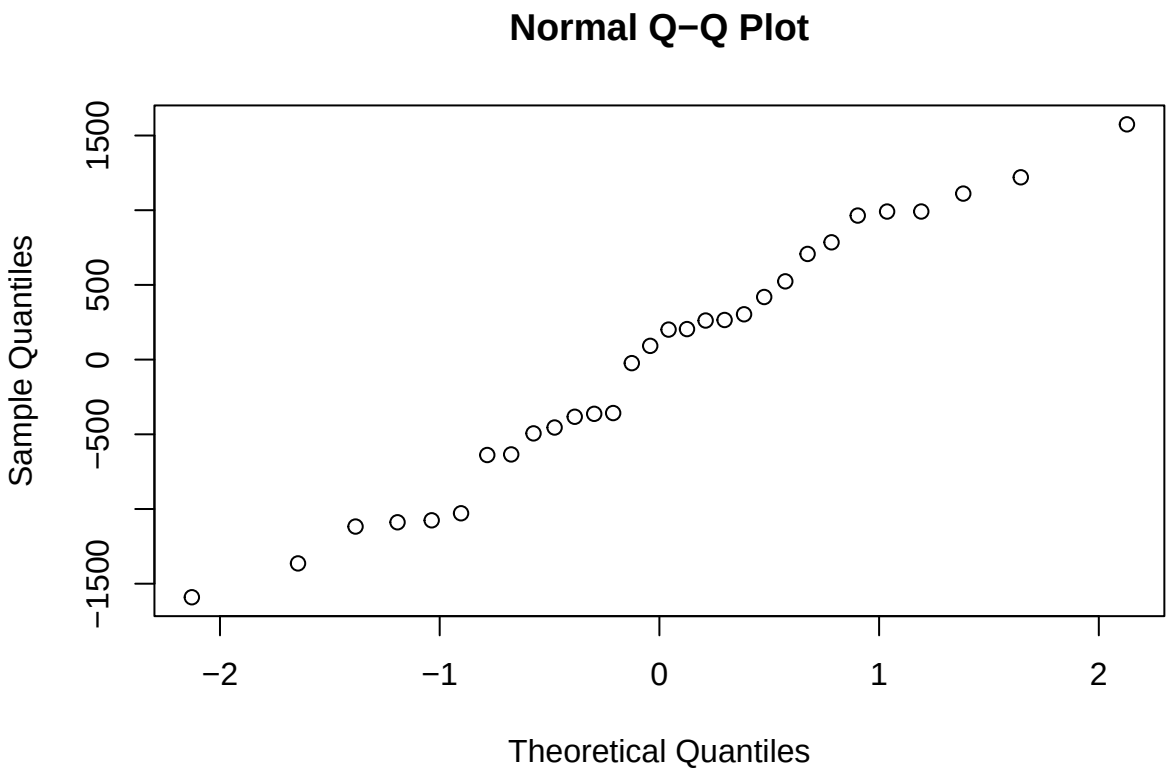
```
##             Df   Sum Sq  Mean Sq F value  Pr(>F)
## size         1 1.20e+08 1.20e+08  138.67 1.8e-11 ***
## county       2 7.67e+05 3.84e+05    0.44  0.6461
## size:county  2 1.05e+07 5.25e+06    6.09  0.0073 **
## Residuals   24 2.07e+07 8.62e+05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

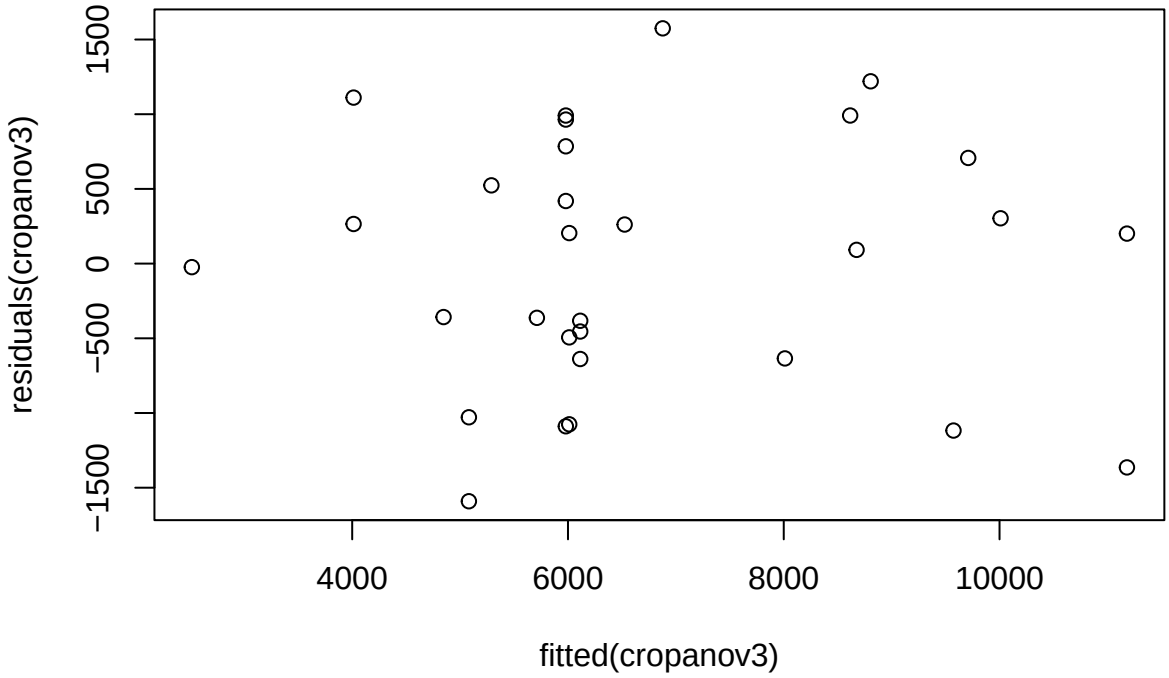
```
summary(cropanov3)
```

```
##
## Call:
## lm(formula = crops ~ size * county, data = cropframe)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -1591   -600    146    661   1575
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2386.54     913.22    2.61   0.0152 *
## size            22.46       4.70    4.78  7.3e-05 ***
## county2      -4370.41    1413.44   -3.09   0.0050 **
## county3      -1340.39    1285.69   -1.04   0.3075
## size:county2    27.51       7.89    3.49   0.0019 **
## size:county3     9.21       6.31    1.46   0.1574
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 929 on 24 degrees of freedom
## Multiple R-squared:  0.863,  Adjusted R-squared:  0.835
## F-statistic: 30.3 on 5 and 24 DF,  p-value: 1.25e-09
```

Based on the result, there is a significant interaction effect between *Size* and *County* on *Crops (p-value = 0.007)*. Summary of the ANOVA model shows that the effect mainly lies on the combination of *size:county 2 (p-value = 0.002)*, while size:county 3 is not significant (p-value = 0.157). Meanwhile, *county 2* also has a significant main effect under the influence of *Size* (p-value = 0.005). Q-Q plot and residual plot show that the assumption of normality is met in this case:

```
qqnorm(residuals(cropanov3)); plot(fitted(cropanov3),residuals(cropanov3))
```


Normal Q−Q Plot
Sample Quantiles
−1500
−500
0
500
1500
−2
−1
0
1
2
Theoretical Quantiles

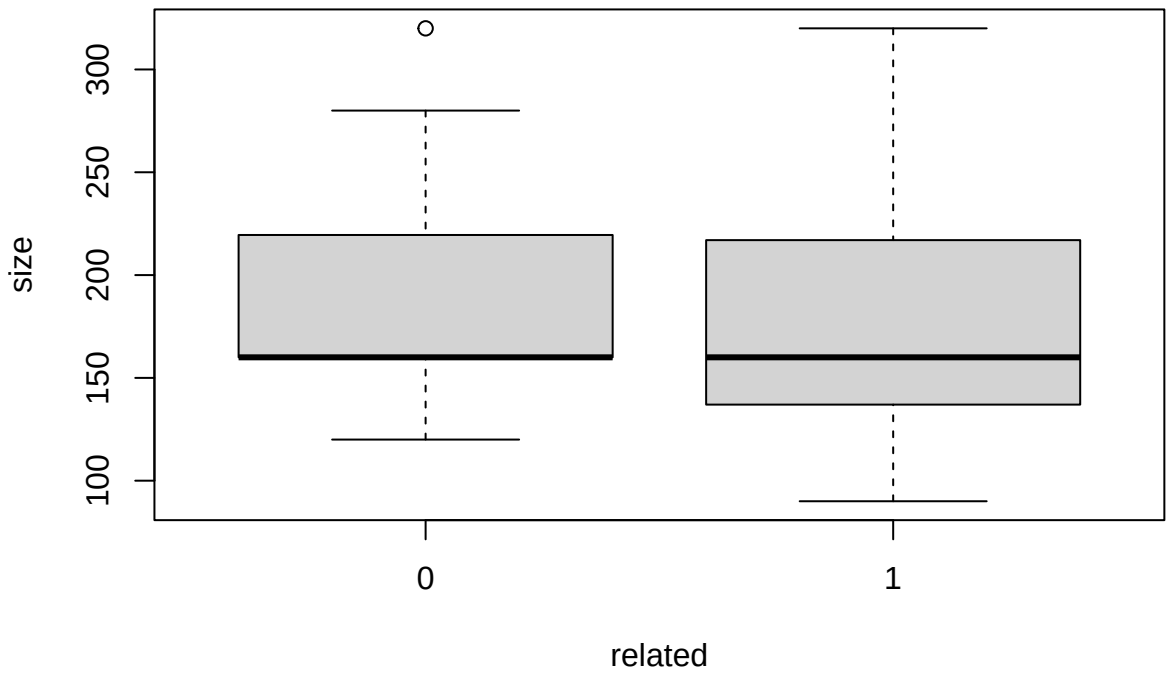


We **reject the null hypothesis** that there is no interaction effect between *Size* and *County*

2. ANCOVA test: *Size * Related*

H_{AB} There is no interaction effect between *Size* and *Related* on *Crops*

```
boxplot(size~related)
```




```
cropanov4=lm(crops~size*related,data=cropframe);anova(cropanov4)
```

```
## Analysis of Variance Table
##
## Response: crops
##              Df   Sum Sq  Mean Sq F value  Pr(>F)
## size          1 1.20e+08 1.20e+08  105.53 1.2e-10 ***
## related       1 1.38e+06 1.38e+06    1.22    0.28
## size:related  1 1.11e+06 1.11e+06    0.98    0.33
## Residuals    26 2.95e+07 1.13e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

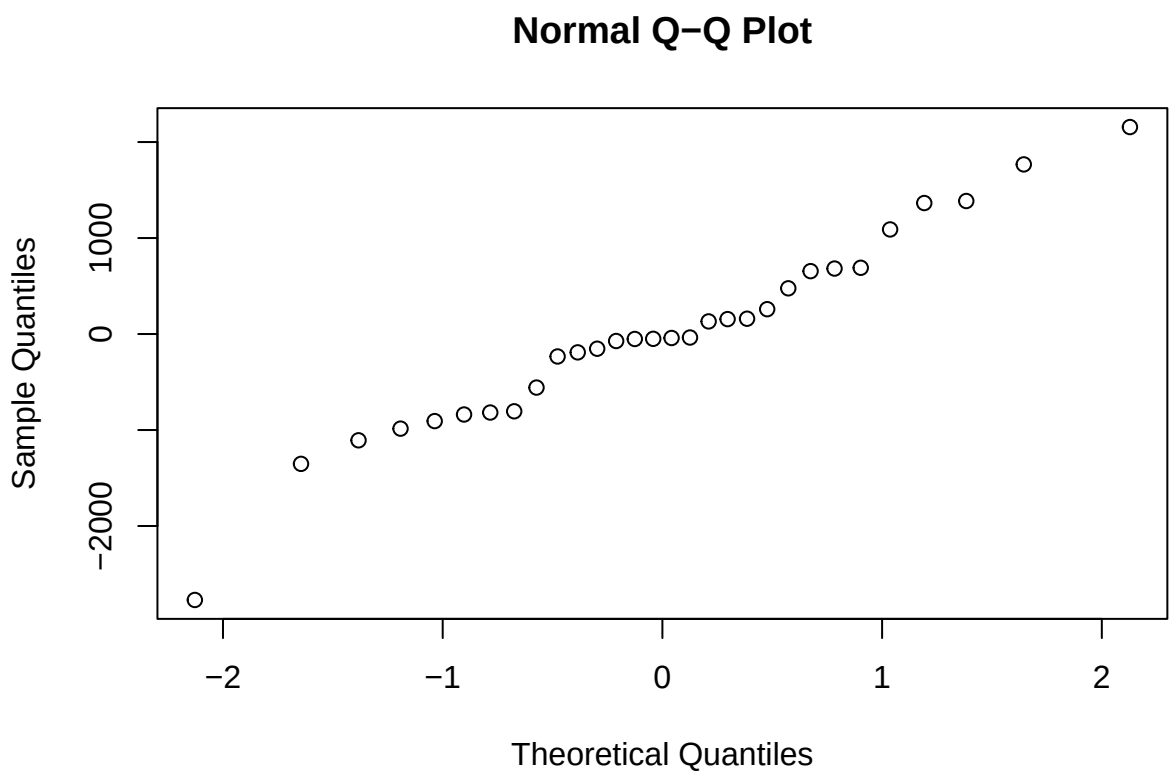
```
summary(cropanov4)
```

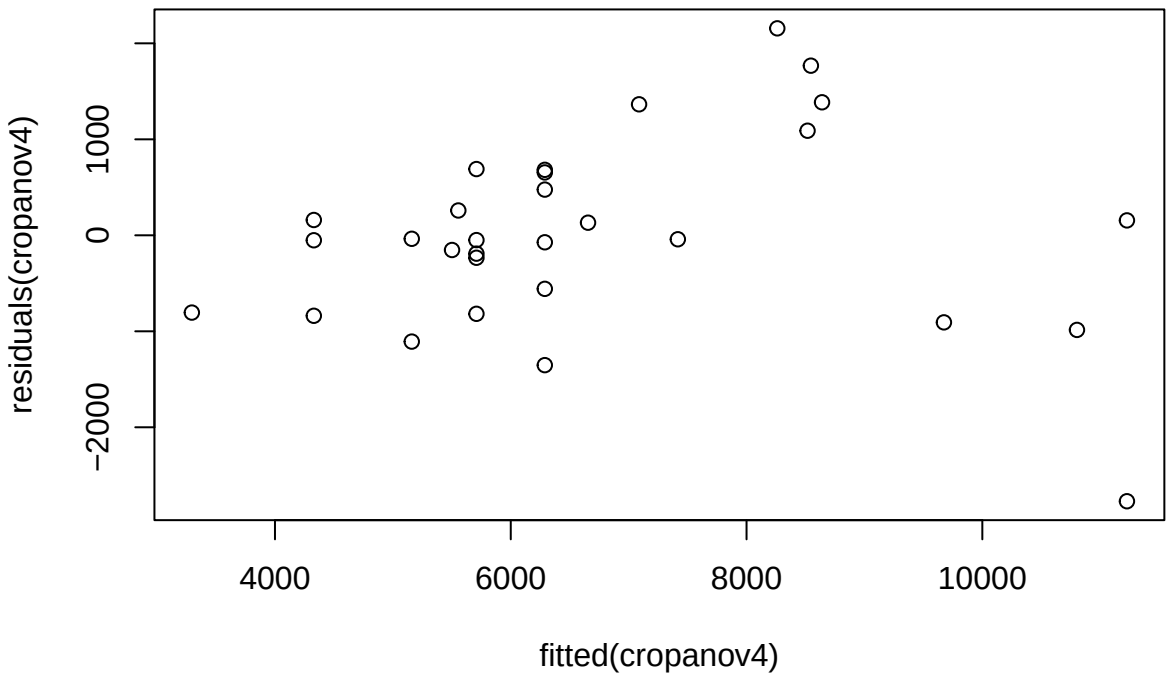
```
##
## Call:
## lm(formula = crops ~ size * related, data = cropframe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2770.6  -743.2   -45.6   610.5  2156.1
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1774.98     940.16    1.89     0.07 .
## size             28.21       4.84    5.83  3.8e-06 ***
## related1      -1583.66    1227.31   -1.29     0.21
## size:related1     6.27       6.33    0.99     0.33
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1060 on 26 degrees of freedom
## Multiple R-squared:  0.806,  Adjusted R-squared:  0.783
## F-statistic: 35.9 on 3 and 26 DF,  p-value: 2.15e-09
```

In the box plot, the distribution of *Size* does not differ much for different *Related* values. The interaction effect is also not significant according to the result of the ANCOVA test (p=0.331). Therefore, we **cannot reject the null hypothesis** in this case. The assumption of normality is met in general, though one can argue that there are more residuals on the lower fitted value side.

```
qqnorm(residuals(cropanov4)); plot(fitted(cropanov4),residuals(cropanov4))
```





We then conduct two ANCOVA tests without interaction. We investigate the main effect of *Related* under the influence of *Size (cropanov5)*, and the main effect of *Size* under the influence of *Related (cropanov6).*

```
cropanov5=lm(crops~size+related,data=cropframe);anova(cropanov5)# related on the second place
```

```
## Analysis of Variance Table
##
## Response: crops
##           Df   Sum Sq  Mean Sq F value  Pr(>F)
## size       1 1.20e+08 1.20e+08  105.59 7.9e-11 ***
## related    1 1.38e+06 1.38e+06    1.22    0.28
## Residuals 27 3.06e+07 1.13e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(cropanov5)
```

```
##
## Call:
## lm(formula = crops ~ size + related, data = cropframe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2410.4  -765.2   -47.1   618.2  2292.6
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1092.78     640.63    1.71     0.10 .
## size           31.88       3.12   10.23  8.6e-11 ***
## related1     -429.29     388.78   -1.10     0.28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1060 on 27 degrees of freedom
## Multiple R-squared:  0.798,  Adjusted R-squared:  0.783
## F-statistic: 53.4 on 2 and 27 DF,  p-value: 4.13e-10
```

```
cropanov6=lm(crops~related+size,data=cropframe);anova(cropanov6)# size on the second place
```

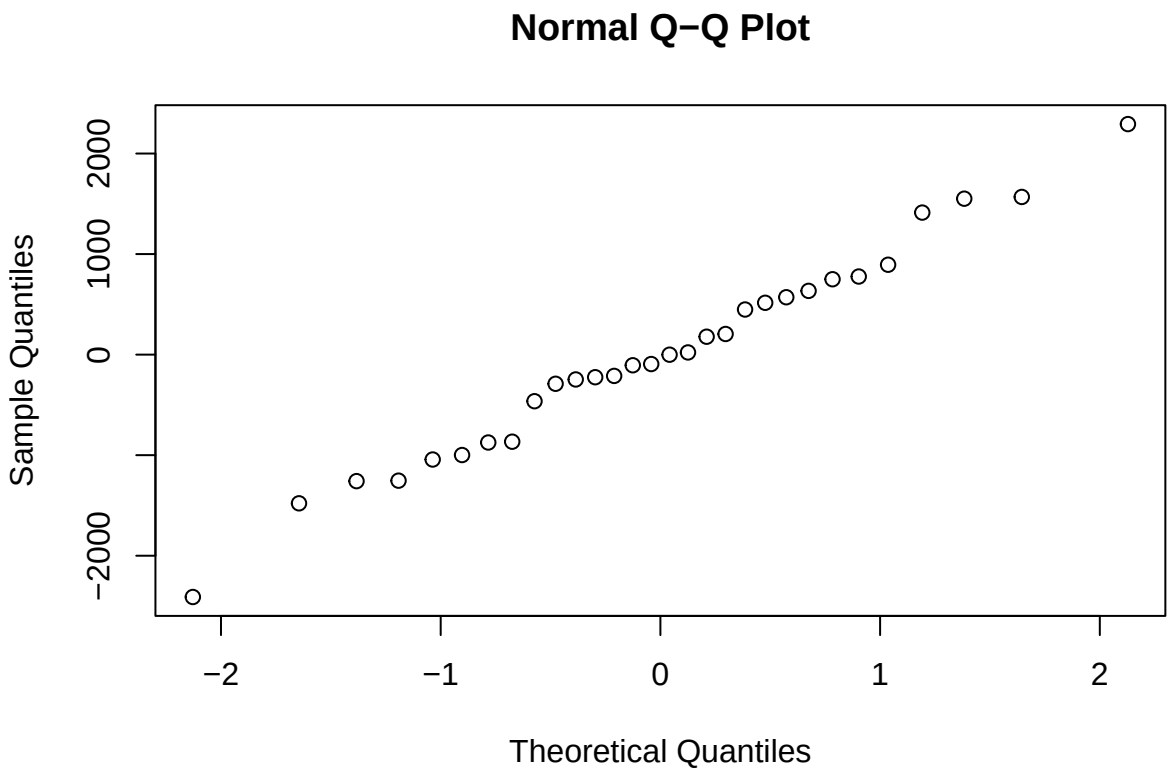
```
## Analysis of Variance Table
##
## Response: crops
##           Df   Sum Sq  Mean Sq F value  Pr(>F)
## related    1 2.38e+06 2.38e+06     2.1    0.16
## size       1 1.19e+08 1.19e+08   104.7 8.6e-11 ***
## Residuals 27 3.06e+07 1.13e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

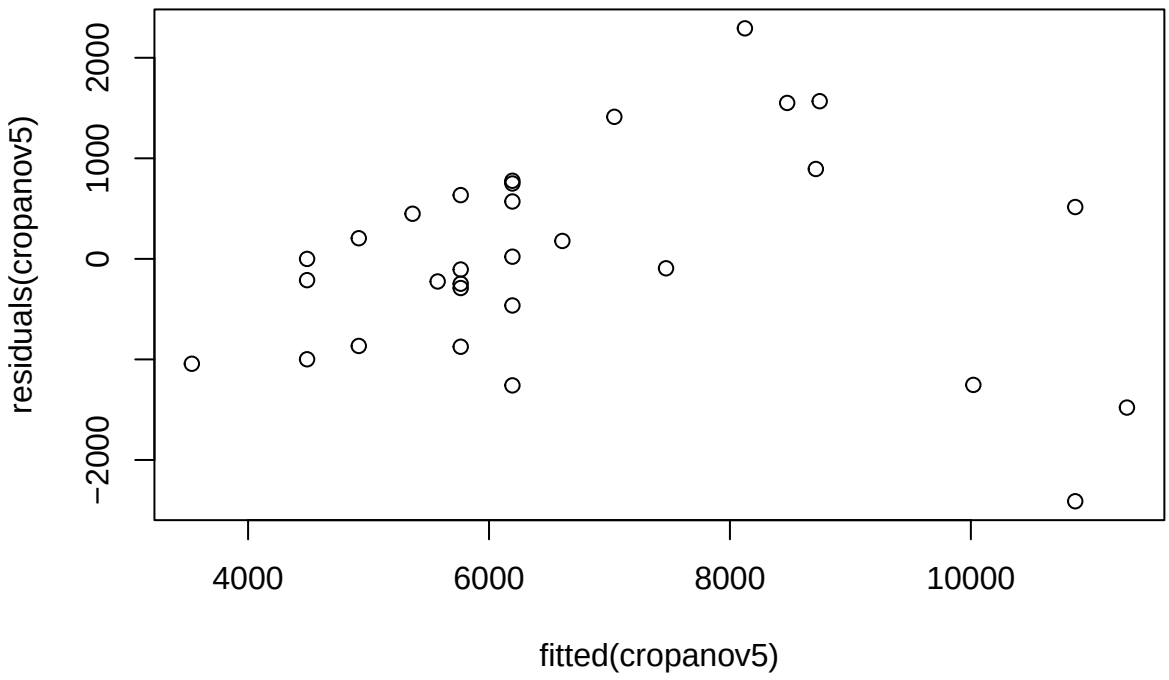
```
summary(cropanov6)
```

```
##
## Call:
## lm(formula = crops ~ related + size, data = cropframe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2410.4  -765.2   -47.1   618.2  2292.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1092.78     640.63    1.71     0.10 .
## related1     -429.29     388.78   -1.10     0.28
## size           31.88       3.12   10.23  8.6e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1060 on 27 degrees of freedom
## Multiple R-squared:  0.798,  Adjusted R-squared:  0.783
## F-statistic: 53.4 on 2 and 27 DF,  p-value: 4.13e-10
```

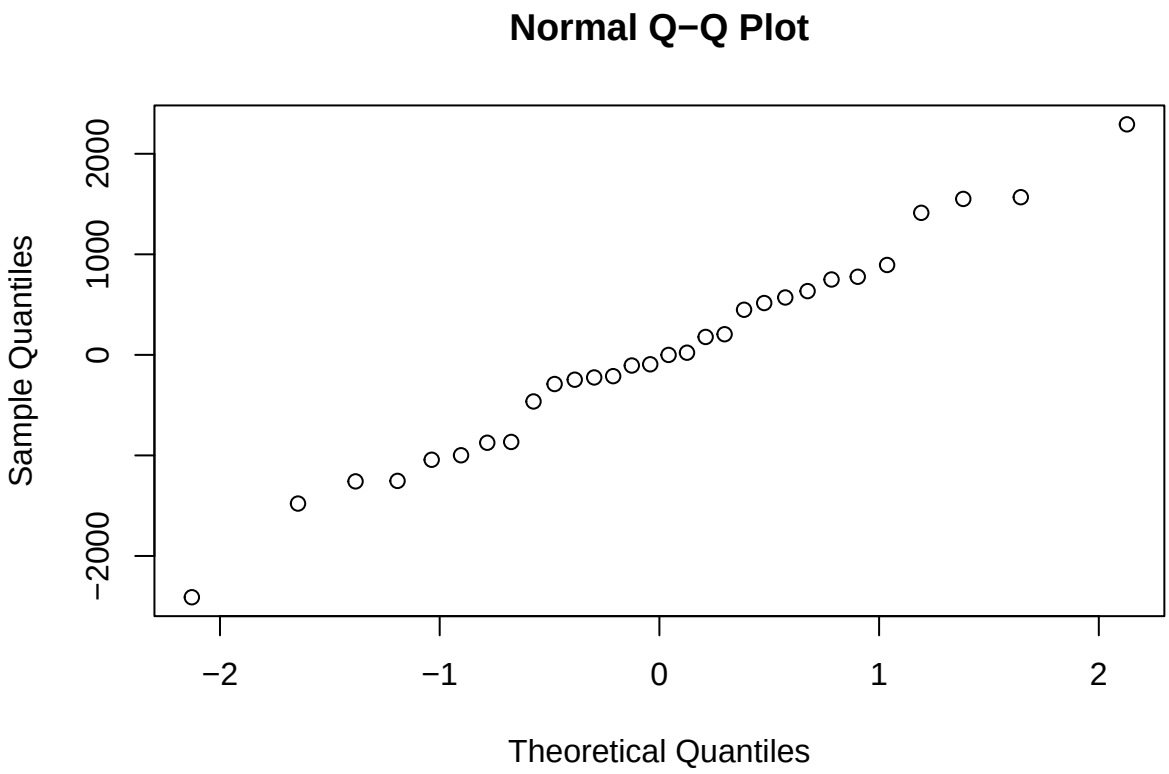
Results show that *Related* does not have a significant main effect on *Crops* under the influence of *Size*, but *Size* has a significant main effect under the influence of *Related*. The normality assumption of both of the ANCOVA tests are met in general, although there seems to be more residuals in the area with a lower fitted score:

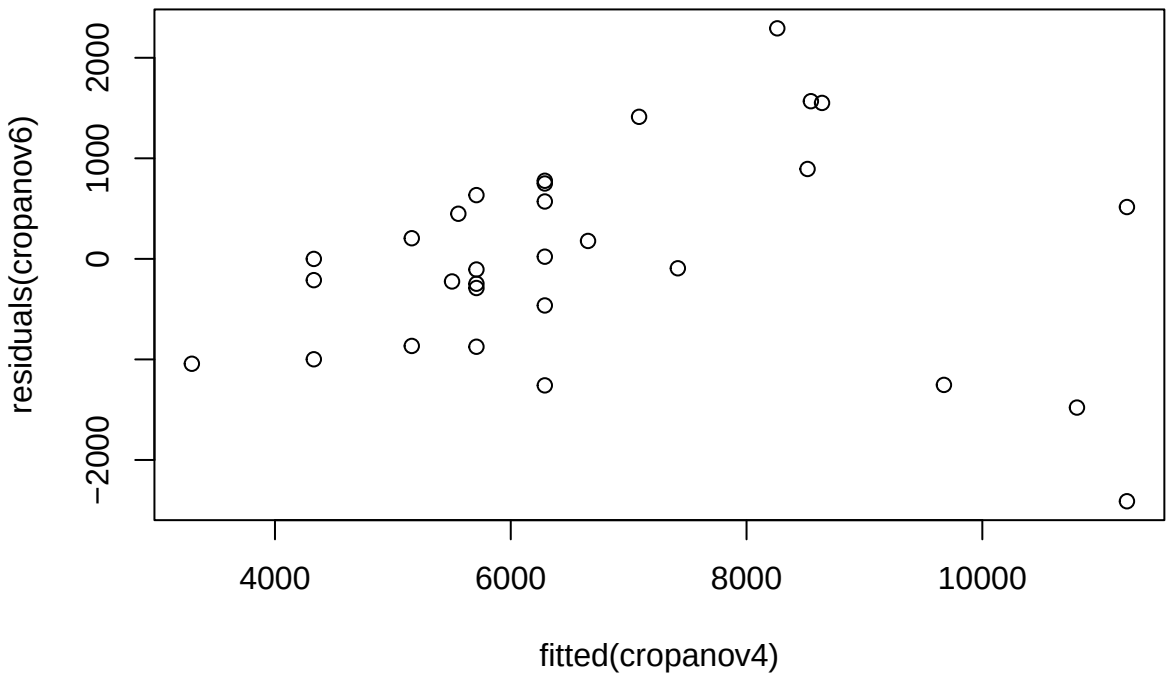
```
qqnorm(residuals(cropanov5)); plot(fitted(cropanov5),residuals(cropanov5))
```






```
qqnorm(residuals(cropanov6)); plot(fitted(cropanov4),residuals(cropanov6))
```


Normal Q−Q Plot
Sample Quantiles
−2000
0
1000
2000
−2
−1
0
1
2
Theoretical Quantiles



Summary of this part:

(1) There is a significant interaction effect between *Size* and *County* on *Crops*. The combination *Size:County 2* is making the most contribution.

(2) No significant interaction effect is found for *Size * Related*.

(3) *Related* has no significant main effect on *Crops* under the influence of *Size*, while *Size* has a significant main effect on *Crops* under the influence of *Related*.

Section c Based on our findings in part (b), we now include all factors (*Related* and *County*) and the exploratory variable (*Size*) together in the same model. We conduct a full ANCOVA test.

$$Crops \sim County + Related + Size + (County \times Size) + (Related \times Size) + (County \times Related)$$

We already know there is no interaction between *Related and Size*, and *County and Related*, so we drop the last two terms and get:

$$Crops \sim County + Related + Size + (County \times Size)$$

```
cropanov7=lm(crops ~ county+related+size+county*size, data = crop_data);anova(cropanov7)# size on the s
```

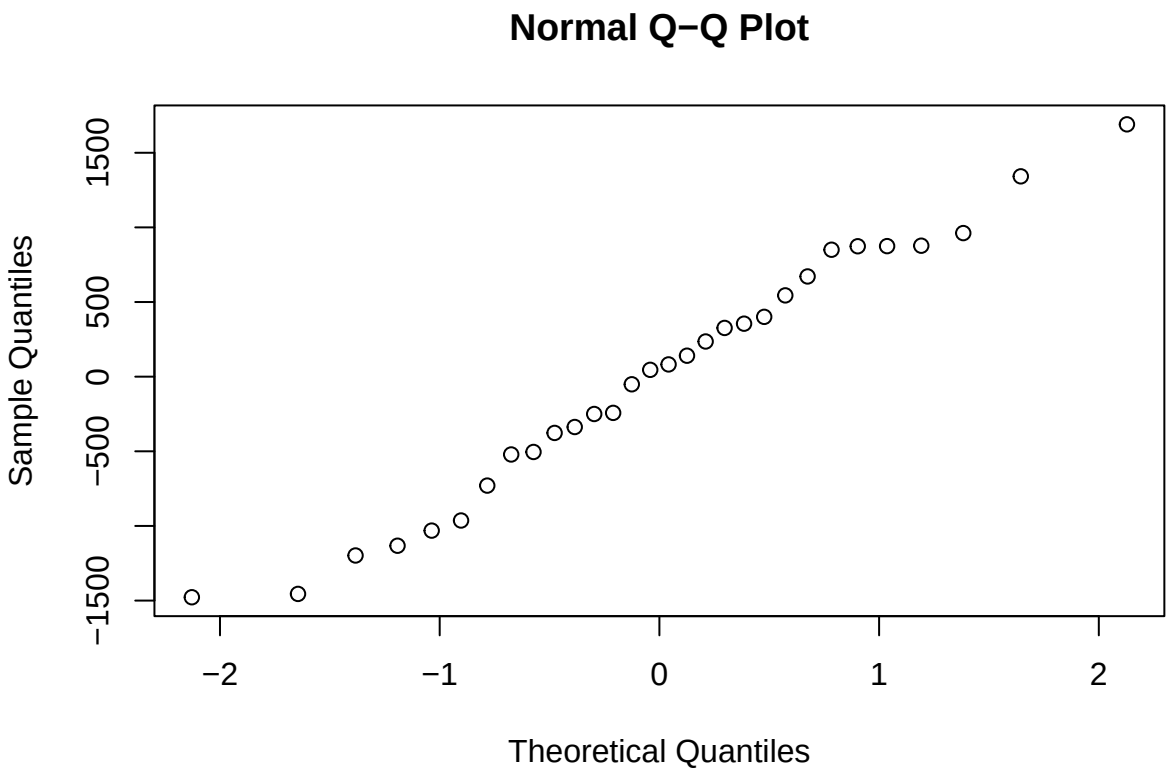
```
## Analysis of Variance Table
##
```

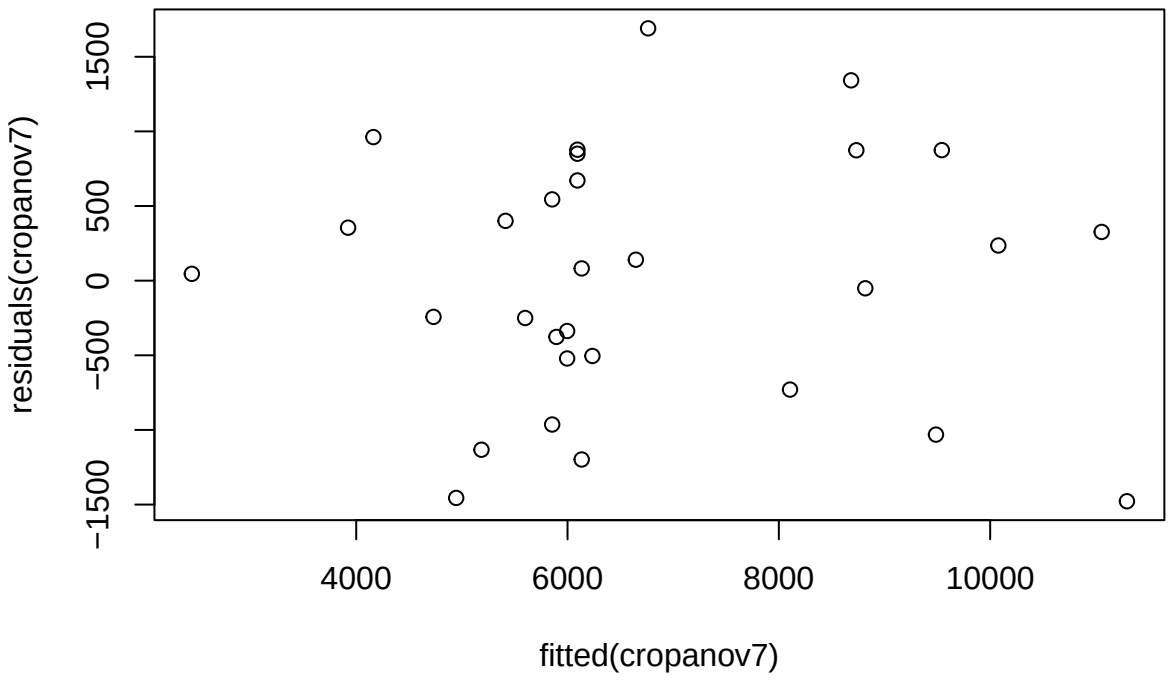
```
## Response: crops
##             Df   Sum Sq  Mean Sq F value  Pr(>F)
## county       2 8.84e+06 4.42e+06    5.01   0.016 *
## related      1 2.38e+06 2.38e+06    2.70   0.114
## size         1 1.10e+08 1.10e+08  125.33 8.7e-11 ***
## county:size  2 9.53e+06 4.76e+06    5.40   0.012 *
## Residuals   23 2.03e+07 8.82e+05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(cropanov7)
```

```
##
## Call:
## lm(formula = crops ~ county + related + size + county * size,
##     data = crop_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1477.6  -517.1    63.9   639.6  1690.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2461.01     929.76    2.65   0.0144 *
## county2      -4214.05    1447.24   -2.91   0.0079 **
## county3      -1284.81    1302.58   -0.99   0.3342
## related1      -239.10     347.92   -0.69   0.4988
## size            22.70       4.77    4.76  8.4e-05 ***
## county2:size    26.59       8.09    3.29   0.0032 **
## county3:size     8.92       6.40    1.39   0.1768
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 939 on 23 degrees of freedom
## Multiple R-squared:  0.866,  Adjusted R-squared:  0.831
## F-statistic: 24.8 on 6 and 23 DF,  p-value: 5.85e-09
```

```
qqnorm(residuals(cropanov7)); plot(fitted(cropanov7),residuals(cropanov7))
```


Normal Q−Q Plot
Sample Quantiles
−1500
−500
0
500
1500
−2
−1
0
1
2
Theoretical Quantiles



Result of a full ANCOVA test also confirms what we found in section B in general. There is a significant main effect of variable *Size (p = 0.000)* and *County 2 (p = 0.008)*. The interaction effect of *Size:County* is therefore also significant *(p = 0.012)*. Different from section B, *County* now also has a slightly significant main effect *(p = 0.016)*, when *County, Related* and *Size* are all included into the model. The assumption of normality is met according to the Q-Q plot and residual plot.

Based on the summary, we can derive several conclusions:

- There is a significant difference between crop values in these three counties. County 2 yields significantly fewer crops than County 1 (Estimate = -4214.050).
- Size has a significantly positive effect on crops, with a larger land results in more crops (Estimate = 22.704).
- The positive effect of size is more prominent in County 2, as there is a significant interaction effect. They yield a higher crops value than Size: County1 (Estimate = 26.590).
- Value of crops does not depend on the relation of landlord and tenant in all three counties, ,as the difference between different relations is not statistically significant.

The following table summarizes all ANOVA tests we've run in this exercise: (the ones with bold fonts are used for predictions)

Name	Test Type	R-model
cropanov	Two-way ANOVA model	$ Crops (County×Related) $
cropanov2	Additive model	$ Crops (County+Related) $

Name	Test Type	R-model
cropanov3	Two-way ANCOVA model	$ Crops (County×Size) $
cropanov4	Two-way ANCOVA model	$ Crops (Size×Related) $
cropanov5	Additive model	$ Crops (Size+Related) $
cropanov6	Additive model	$ Crops (Related+Size) $
cropanov7	Full ANCOVA model	$Crops \sim County + Related + Size + (County \times Size)$

Section d We will apply model **cropanov7** to make the prediction. The mathematical formula for a full ANCOVA is:

$$Y_{ijk} = \mu_{ij} + e_{ijk}$$

where

$$\mu_{ijk} = \mu + \alpha_i + \beta_j + \delta_k + \gamma_{ik}$$

μ is the overall mean

α_i is the main effect of level i of the factor *County*, i = 1,2,3

β_j is the main effect of level j of the factor *Related*, j = 0,1

δ_k is the main effect of the exploratory variable Size, k = 1…n

$\gamma_i k$ is the interaction effect of levels i, k of factor *County* and exploratory variable *Size*.

According to this equation, the crops from County 2 of size 165, and related landlord and tenant is therefore:

$$Crops = Intercept + County2 + Related1 + Size165 + County2 * Size165$$

$$= 2461.014 - 4214.050 - 239.099 + 22.704 * 165 + 26.590 * 165 = 6141.378$$

So the final crops value is 6141.378

The error variance is given by:

$$\hat{\sigma}^2 = \frac{\text{RSS}}{\text{df}}$$

According to the summary of the **cropanov7,** we then have:

$$\hat{\sigma}^2 = \frac{20277325}{23} = 881623$$

The error variance is therefore 881623

Exercise 3: Yield of peas

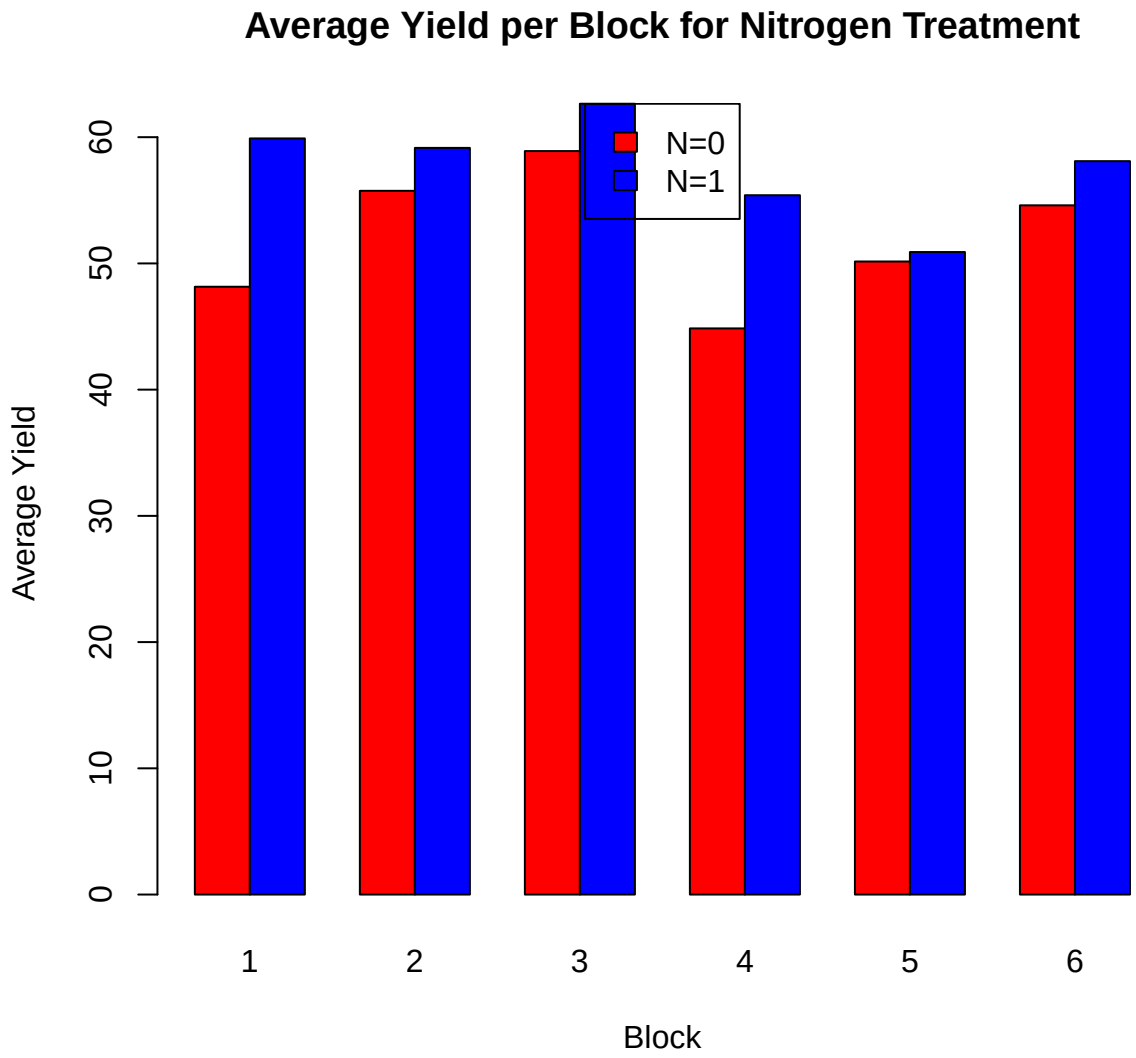
Section a

```r
library('MASS')

set.seed(123)  # add random seed for reproduce

# initial params
I <- 6  # blocks
J <- 4  # plots per block

# initial data frame
randomized_design <- data.frame(block = rep(1:I, each = J), plot = rep(1:J, times = I))

# for each block b, put (N, P, K) on each 2 plots randomly
for (b in 1:I) {
  plots <- sample(1:J, J, replace = FALSE)  # randomly reorder plots in each block

  # put N in the header 2 plots
  randomized_design$N[randomized_design$block == b] <- ifelse(plots %in% plots[1:2], 1, 0)

  # randomly put P in 2 plots
  randomized_design$P[randomized_design$block == b] <- ifelse(plots %in% sample(plots, 2), 1, 0)

  # randomly put K in 2 plots
  randomized_design$K[randomized_design$block == b] <- ifelse(plots %in% sample(plots, 2), 1, 0)
}

# print the plots
print(randomized_design)
```

```
##    block plot N P K
## 1      1    1 1 0 0
## 2      1    2 1 1 1
## 3      1    3 0 1 0
## 4      1    4 0 0 1
## 5      2    1 1 1 0
## 6      2    2 1 1 1
## 7      2    3 0 0 1
## 8      2    4 0 0 0
## 9      3    1 1 1 1
## 10     3    2 1 0 0
## 11     3    3 0 0 0
## 12     3    4 0 1 1
## 13     4    1 1 1 0
## 14     4    2 1 1 1
## 15     4    3 0 0 1
## 16     4    4 0 0 0
## 17     5    1 1 0 0
## 18     5    2 1 0 1
## 19     5    3 0 1 0
## 20     5    4 0 1 1
## 21     6    1 1 0 1
## 22     6    2 1 1 0
## 23     6    3 0 1 1
## 24     6    4 0 0 0
```

Section b

```r
# combine those yield in the same block same N, and calc its mean
yield_matrix <- tapply(npk$yield, list(npk$block, npk$N), mean)

# plot bars
barplot(t(yield_matrix), beside = TRUE, col = c("red", "blue"),
        main = "Average Yield per Block for Nitrogen Treatment",
        xlab = "Block", ylab = "Average Yield")
legend("top", legend = c("N=0", "N=1"), fill = c("red", "blue"))
```



This plot illustrates that the average yields for soil treated by N are higher than for untreated soil. What's more, each block and treatment tend to have a similar change.

Meanwhile, we here have assigned treatments randomly to each soil within a block, which reduces the variation and get more precise results.

Section c

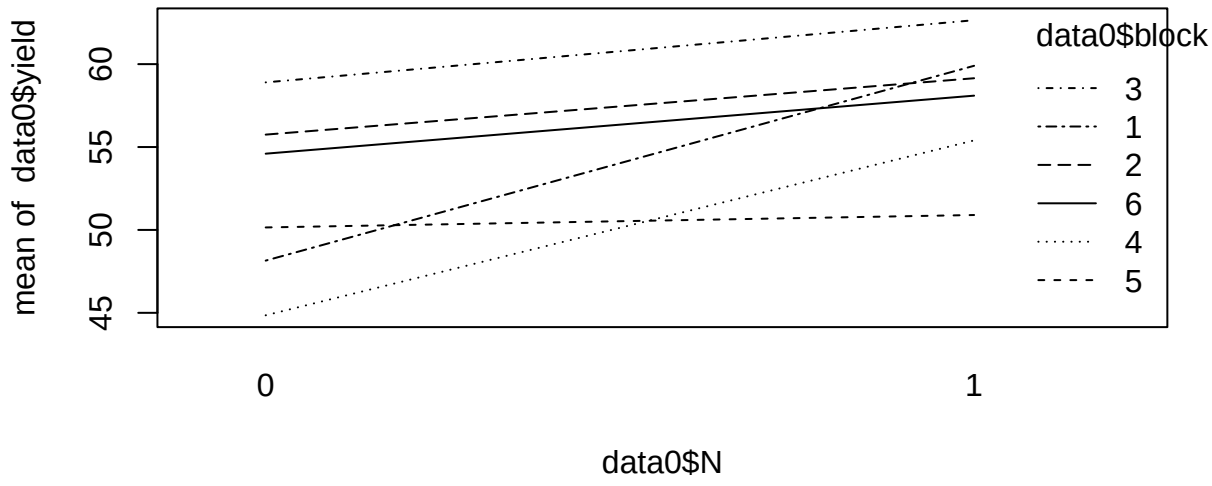
```
data0 = npk
data0$block = as.factor(data0$block)
data0$N = as.factor(data0$N)
```

```
# Two-Way ANOVA
model2way = lm(yield~N*block, data=data0)
anova(model2way)
```

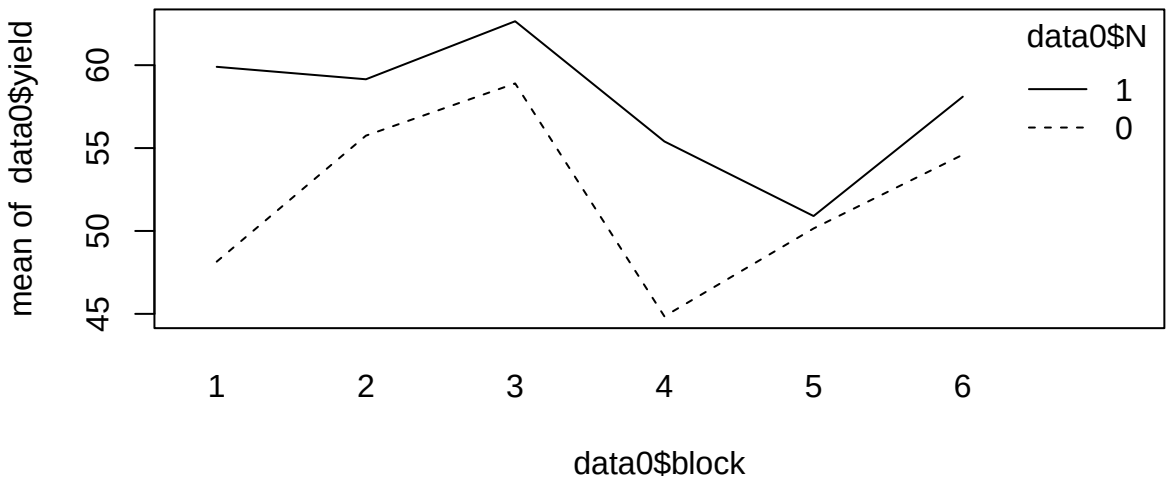
```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value Pr(>F)
## N          1    189   189.3    9.26   0.01 *
## block      5    343    68.7    3.36   0.04 *
## N:block    5     99    19.7    0.96   0.48
## Residuals 12    245    20.4
## ---
## Signif. codes: 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

$p > 0.05$, which means there is no significant evidence of interaction effect.

```
interaction.plot(data0$N,data0$block,data0$yield)
```




```
interaction.plot(data0$block,data0$N,data0$yield)
```



Interaction plot also display parallel lines, indicating no interaction.

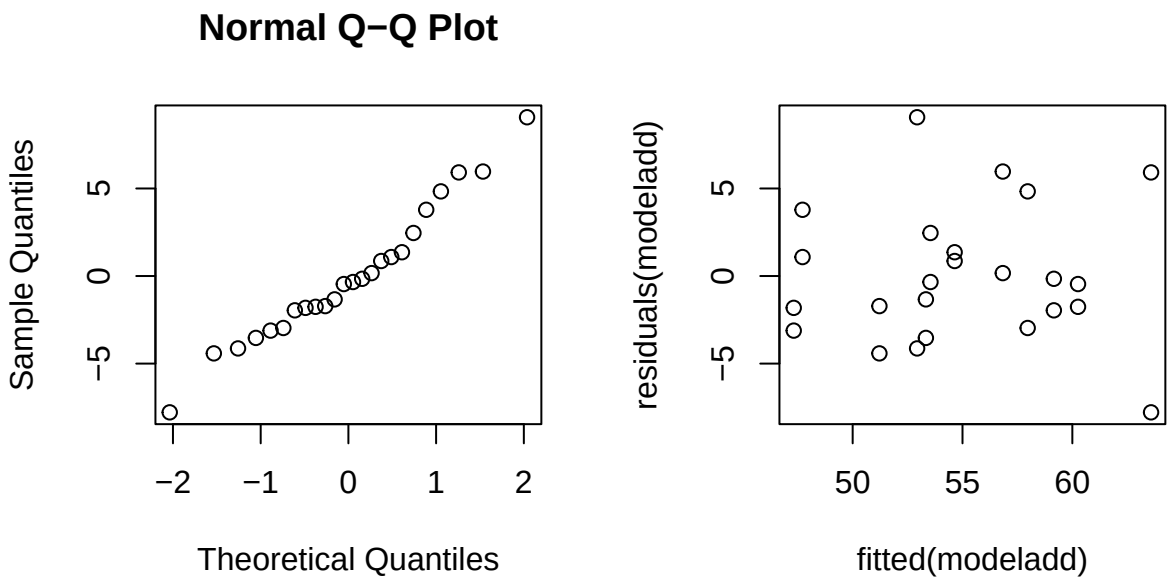
So, we have to try the "additive" model:

```
modeladd <- lm(yield ~ N + block, data = data0)
anova(modeladd)
```

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

In both cases $p < 0.05$, so both factors have a main effect in the "additive" model.

```
# Diagnostics:
par(mfrow=c(1,2))
qqnorm(residuals(modeladd)); plot(fitted(modeladd),residuals(modeladd))
```



From QQPlot, we can tell that the curve more or less straight, so it is likely normal. Meanwhile, there is no significant pattern in the fitted plot, which is good and means the residual is independent and identical.

- **Was it sensible to include factor block into this model?** From the results showed in "additive" model, the p_value of block is $0.007095 < 0.05$, and the N is the first order in our model, so it makes sense to include the block.
- **Can we also apply the Friedman test for this situation?** No, because each block has more than one same value N, meanwhile, the treatments are not completely randomized.

Section d

```
pairwiseP <- lm(yield ~ block*P + N + K, data = data0)
pairwiseK <- lm(yield ~ block*K + P + N, data = data0)
pairwiseN <- lm(yield ~ block*N + K + P, data = data0)
```

```
anova(pairwiseP); anova(pairwiseK); anova(pairwiseN)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value Pr(>F)
## block      5    343    68.7    4.07 0.0282 *
## P          1      8     8.4    0.50 0.4966
## N          1    189   189.3   11.21 0.0074 **
## K          1     95    95.2    5.64 0.0389 *
## block:P    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)
## block      5    343    68.7    4.04 0.0288 *
## K          1     95    95.2    5.60 0.0395 *
## P          1      8     8.4    0.49 0.4980
## N          1    189   189.3   11.14 0.0075 **
## block:K    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
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value Pr(>F)
## block      5    343    68.7    4.85 0.0164 *
## N          1    189   189.3   13.36 0.0044 **
## K          1     95    95.2    6.72 0.0268 *
## P          1      8     8.4    0.59 0.4590
## block:N    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
```

No interaction effect for either of the three, thus we do an additive model:

```
modeladd2 <- lm(yield ~ block + N + P + K, data = data0); anova(modeladd2)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value Pr(>F)
## block      5    343    68.7    4.29 0.0127 *
## N          1    189   189.3   11.82 0.0037 **
## P          1      8     8.4    0.52 0.4800
## 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
```

$p < 0.05$ for N and K, showing a main effect. $p > 0.05$ for P, so we can conclude that there is no significant effect.

We conclude that the best model is the additive ANOVA model. Additive ANOVA model provides an overall indication of the effects of each factor N, P, K and block. From the p values we can tell additive model are better than the pairwise model. Furthermore, the pairwise models only focus on the interaction between two factors at a time, lacking the control for others.

Section e

```r
model_d = lm(yield ~ block + N + K + P, data = data0)
summary(model_d)
```

```
##
## Call:
## lm(formula = yield ~ block + N + K + P, data = data0)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.000 -1.708 -0.083  2.246  6.483
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    53.80       2.45   21.96  8.1e-13 ***
## block2          3.43       2.83    1.21   0.2448
## block3          6.75       2.83    2.39   0.0307 *
## block4         -3.90       2.83   -1.38   0.1883
## block5         -3.50       2.83   -1.24   0.2351
## block6          2.32       2.83    0.82   0.4241
## N1              5.62       1.63    3.44   0.0037 **
## K1             -3.98       1.63   -2.44   0.0277 *
## P1             -1.18       1.63   -0.72   0.4800
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4 on 15 degrees of freedom
## Multiple R-squared:  0.726,  Adjusted R-squared:  0.58
## F-statistic: 4.97 on 8 and 15 DF,  p-value: 0.00376
```

From the summary we can see block 3 is the best in all blocks, and N1 is better than N0, which means N treated is better, while P and K are prefered to be untreated. So, the best combination is **(3, 1, 0, 0)** for (block, N, P, K), leading the largest yield.

Section f

```r
library('MASS')
library('lme4')
data0 = npk

model_mixed <- lmer(yield ~ N+P+K+(1|block),data=data0,REML=FALSE)
model_fixed <- lm(yield ~  N + P + K + block, data = data0)
anova(model_mixed,model_fixed)
```

```
## Data: data0
## Models:
## model_mixed: yield ~ N + P + K + (1 | block)
## model_fixed: yield ~ N + P + K + block
##             npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)
## model_mixed    6 151 158  -69.5      139
## model_fixed   10 143 155  -61.7      123  15.6  4     0.0035 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model comparison results are:

- AIC: The fixed effects model (AIC = 143.39) is lower than the mixed effects model (AIC = 151.03), suggesting better model fit.
- BIC: The fixed effects model (BIC = 155.17) is also lower, reinforcing the AIC results.
- Log-likelihood: The fixed effects model has a higher log-likelihood (-61.695 vs. -69.514), meaning it fits the data better.
- Chi-square test: $^2 = 15.639$, p = 0.003544 (significant at $p < 0.05$), indicating that treating block as a fixed effect is more appropriate.

The fixed effects model (lm(yield ~ block + N + P + K)) provides a better fit than the mixed effects model.