

# Assignment 1 - Report

Name, Eleni Liarou, Zoë Azra Blei, group 20

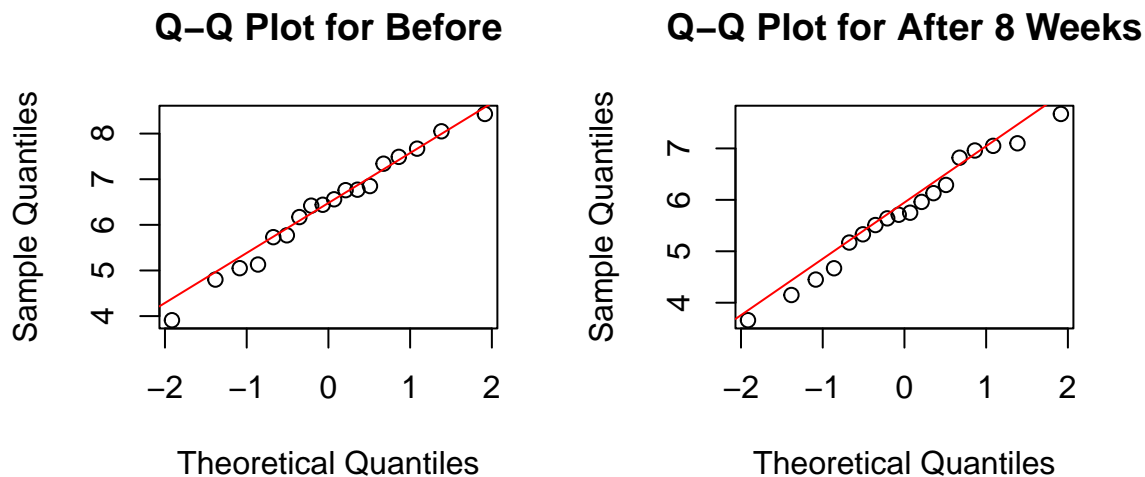
23 February 2025

## Exercise 1: Cholesterol

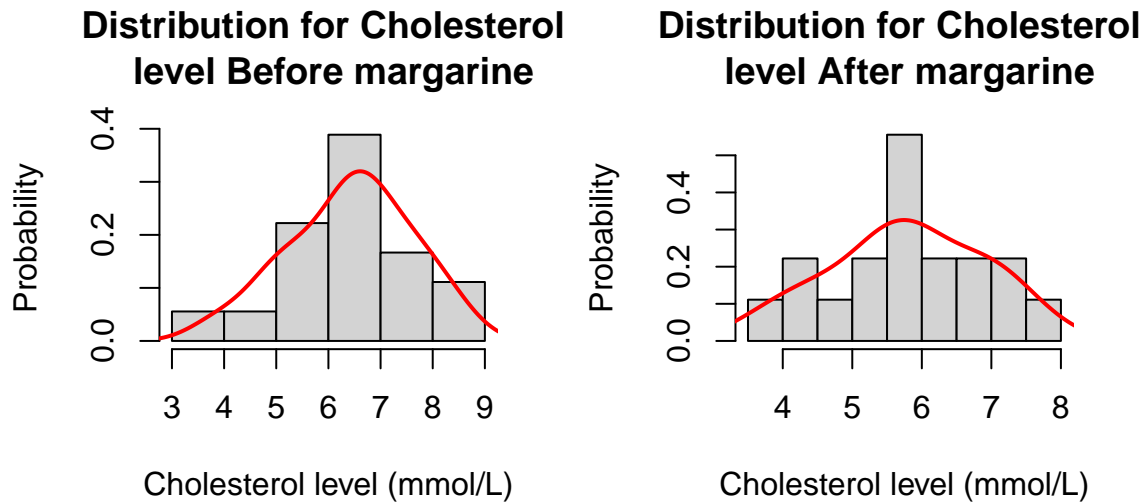
First we load and read the necessary data set

```
data = read.delim("cholesterol.txt", sep=' ')
```

**Section a** In order to investigate the normality of the data set, Q-Q plots are created below for both the *Before* and *After8weeks* columns. As in both plots the data points closely follow the diagonal red line, the data is approximating a normal distribution. While some minor deviations may be present in the tails, the overall pattern suggests that the normality assumption is reasonable.



To further explore the normality assumption, histograms below are plotted for both *Before* and *After8weeks*. The histograms exhibit a roughly bell-shaped distribution, which supports the assumption of normality.



However, to address normality more formally, a Shapiro-Wilk test is conducted, as this test is suitable to test for normality for small data sets. The null hypothesis is as follows:

H0: The data is normally distributed.

The W-statistic measures how closely the data aligns with a normal distribution, ranging from 0 to 1, where values closer to 1 indicate a stronger likelihood of normality. Considering the results for *Before* and *After8weeks*, both W-values are close to 1. Additionally, with a 95% confidence level, both p-values exceed 0.05, meaning that we fail to reject H0. These findings provide strong evidence that the data in both columns can be considered to be normally distributed.

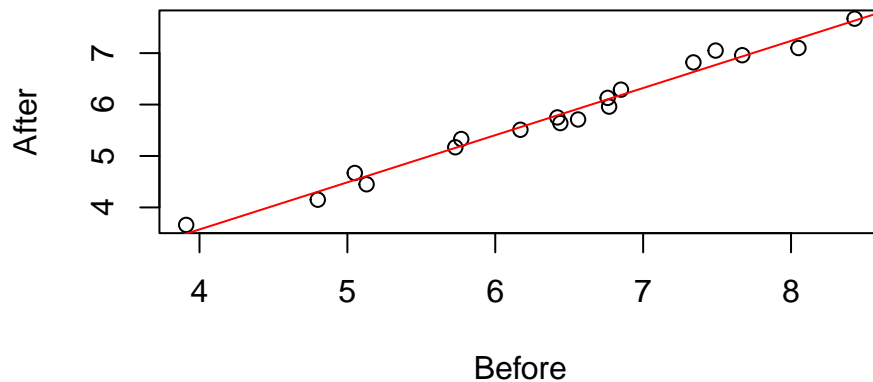
```
shapiro_before = shapiro.test(data$Before)
shapiro_after = shapiro.test(data$After8weeks)
```

**Shapiro-Wilk Test for Before:** W-statistic = 0.982 , p-value = 0.967

**Shapiro-Wilk Test for After 8 Weeks:** W-statistic = 0.977 , p-value = 0.918

In order to investigate the relationship between the columns of *Before* and *After8weeks* a scatter plot is created below. The scatter plot demonstrates a strong positive correlation between 'Before' and 'After8weeks' cholesterol levels. The data points align closely with the red regression line, suggesting that individuals with higher cholesterol levels before the diet intervention also tend to have higher cholesterol levels after 8 weeks. This indicates that while cholesterol levels may have decreased, there remains a strong relationship between pre- and post-diet measurements.

## Regression for Before and After8weeks



To quantify this correlation, the Pearson's correlation coefficient is calculated below. A high Pearson correlation (close to 1) indicates a strong positive relationship between the two columns. The correlation coefficient exhibits a value of approximately 0.99, confirming the strong positive relationship. Additionally, the p-value is smaller than 0.05, indicating that the correlation is statistically significant.

```
cor_result = cor.test(data$Before, data$After8weeks, method = "pearson")
```

### Pearson Correlation Test:

Correlation coefficient ( $r$ ) = 0.991 , p-value = 2.321e-15 , 95% CI: [ 0.975 , 0.997 ]

**Section b** Since the cholesterol levels were measured on the same individuals at different times, the data is paired, making a paired t-test appropriate. However, since the t-test assumes normality of the mean difference, we first assess this assumption by applying a Shapiro-Wilk test.  $H_0$ : The data is normally distributed.

The W-statistic (0.99) is close to 1, and the p-value exceeds 0.05, meaning we fail to reject  $H_0$ , providing strong evidence that the mean difference follows a normal distribution.

```
difference = data$Before - data$After8weeks  
shapiro_result = shapiro.test(difference)
```

**Shapiro-Wilk Test for Mean Difference:** W-statistic = 0.985 , p-value = 0.987

The paired t-test shows strong evidence against the null hypothesis ( $t = 14.946$ ,  $p = 1.639e-11$ ), indicating a significant cholesterol reduction after 8 weeks on the margarine diet. The mean difference of 0.629 mmol/L supports this, with a 95% confidence interval [0.556, Inf] confirming that the reduction is at least 0.556 mmol/L. These results suggest a clear effect of the diet on lowering cholesterol levels.

```
t_test_result = t.test(data$Before, data$After8weeks, paired = TRUE, alternative = "greater")
```

### Paired T-Test Results:

T-statistic = 14.946 , p-value = 1.639e-11 , 95% Confidence Interval: [ 0.556 , Inf ]

The permutation test ( $p = 1.0e-05$ ) provides strong evidence to reject the null hypothesis, confirming a significant cholesterol reduction of 0.629 mmol/L after 8 weeks. This supports the effectiveness of the margarine diet. A Mann-Whitney U-test is not applicable as it requires independent samples, while this data is paired.

```
diff = data$Before - data$After8weeks
n_permutations = 100000
observed_mean = mean(diff)

permute_test = function(diff) {
  permuted_diff <- diff * sample(c(-1, 1), length(diff), replace = TRUE)
  return(mean(permuted_diff))
}

set.seed(42)
permute_distr = replicate(n_permutations, permute_test(diff))
rounded_observed_mean = round(observed_mean, 3)
rounded_p_value = formatC(mean(abs(permute_distr) >= abs(observed_mean)), format = "e", digits
```

Observed mean difference: 0.629 Permutation test p-value: 1.000e-05

**Section c** To estimate the mean cholesterol level after 8 weeks, we compute a 97% confidence interval using both the t-test and bootstrapping. The t-test CI, assuming normality, is [5.164, 6.394], while the bootstrap CI, based on resampling, is [5.230, 6.320]. The substantial overlap suggests that both methods provide similar estimates. The slight differences arise because the t-test relies on the sample standard deviation, whereas bootstrapping does not assume normality. Since both approaches yield consistent results, we can confidently conclude that cholesterol levels remain moderate after 8 weeks on the diet.

```
n = length(data$After8weeks)
sample_mean = mean(data$After8weeks)
sample_sd = sd(data$After8weeks)
critical_value = qt(1-0.015, df=17)
standard_error = sample_sd / sqrt(n)
left_bound = sample_mean - critical_value * standard_error
right_bound = sample_mean + critical_value * standard_error

bootstrap_ci = function(x, conf_level = 0.97, B = 10000) {
  alpha = 1 - conf_level
  Bstats = lapply(1:B, FUN = function(i) {
    boot_sample = sample(x, size = length(x), replace = TRUE)
```

```

    mean(boot_sample)
  } )
Bstats = unlist(Bstats)
ci = round(quantile(Bstats, prob = c(alpha/2, 1-alpha/2)), 3)
return(ci)
}

```

### Results:

Sample mean: 5.779, Sample standard deviation: 1.102, Critical value: 2.368, 97% CI for mu: [ 5.164 , 6.394 ]

Bootstrap test with 97% CI: [ 5.23 , 6.32 ]

**Section d** For the bootstrap test, we consider the following null hypothesis:

H0: The data follows a uniform distribution with a minimum of 3 and a maximum of theta.

```

set.seed(42) # Ensure reproducibility
T_max_observed = max(data$After8weeks)
n = length(data$After8weeks)
theta_values = 3:12
p_values = numeric(length(theta_values))
Bootstraps = 100000
for (i in seq_along(theta_values)) {
  theta = theta_values[i]
  t_star = numeric(Bootstraps)
  for (j in 1:Bootstraps) {
    x_star = runif(n, min = 3, max = theta)
    t_star[j] = max(x_star)
  }
  pl = sum(t_star < T_max_observed) / Bootstraps
  pr = sum(t_star > T_max_observed) / Bootstraps
  p_values[i] = 2 * min(pl, pr)
}

```

**Results: Table 1** Theta values and their p-values

	3	4	5	6	7	8	9	10	11	12
P-Value	0	0	0	0	0	0.58	0.021	0.001	0	0

To generate bootstrap samples from a Uniform(3, theta) distribution, we iterate over theta values from 3 to 12 in steps of 1, generating 100,000 bootstrap samples for each. The maximum value from each re-sample is recorded, and the p-value is computed by comparing the observed maximum ( $T\_max\_observed$ ) with the bootstrap distribution ( $t\_star$ ).

```

accepted_theta <- c(8, 9, 10, 11)
ks_results <- list()

for (theta in accepted_theta) {
  ks_test_result <- ks.test(data$After8weeks, "punif", min = 3, max = theta)
  ks_results[[as.character(theta)]] <- list(
    theta = theta,
    D_statistic = round(ks_test_result$statistic, 3),
    p_value = formatC(ks_test_result$p.value, format = "e", digits = 3)
  )
}

```

**Kolmogorov-Smirnov Test for Uniform(3, 8 ):** D-statistic: 0.212 , P-value: 3.451e-01  
**Kolmogorov-Smirnov Test for Uniform(3, 9 ):** D-statistic: 0.261 , P-value: 1.429e-01  
**Kolmogorov-Smirnov Test for Uniform(3, 10 ):** D-statistic: 0.359 , P-value: 1.401e-02  
**Kolmogorov-Smirnov Test for Uniform(3, 11 ):** D-statistic: 0.432 , P-value: 1.438e-03

For  $\theta = 9$ ,  $\theta = 10$ , and  $\theta = 11$  we get a  $p$ -value  $< 0.05$ , thus we reject  $H_0$  indicating that the observed maximum for these values is significantly different from the expected maximum under  $U(3, \theta)$ . For  $\theta = 8$ , we fail to reject  $H_0$ , meaning that the data could be from a uniform distribution.

**Section e** To test whether the median of the cholesterol level is less than 6, we will use a Wilcoxon signed-rank test, because it is a non-parametric test that does not assume normality, making it suitable for small sample sizes and skewed data. The null hypothesis is:

$H_0$ : The median cholesterol level is 6 or greater.

The test yields a  $p$ -value of 0.223, meaning we fail to reject  $H_0$ . This suggests that the median cholesterol level after 8 weeks is not significantly less than 6 at the 5% significance level.

```

wilcox_result = wilcox.test(data$After8weeks, mu = 6, alternative = "less", exact = FALSE)

```

**Wilcoxon Signed-Rank Test:** V-statistic: 67.5 ,  $p$ -value: 2.230e-01

To determine whether the fraction of cholesterol levels below 4.5 after 8 weeks exceeds 25%, a binomial test is conducted. The null hypothesis is:

$H_0$ : The fraction of cholesterol levels below 4.5 is at most 25%.

The test yields a  $p$ -value of 0.865, meaning we fail to reject  $H_0$  at the 5% significance level. This suggests that there is no significant evidence that the proportion of cholesterol levels below 4.5 is greater than 25%.

```

count_below_4.5 = sum(data$After8weeks < 4.5)
binom_result = binom.test(count_below_4.5, length(data$After8weeks), p = 0.25, alternative = ">")

```

**Exact binomial test**

Number of successes: 3 , Number of trials: 18 ,  $p$ -value: 0.865 , 95% CI: [ 0.047 1 ], Estimated probability of success: 0.167

## Exercise 2: Crops

First we load the necessary data

```
crops_data <- read.table("crops.txt", header=TRUE)
crops_data$County <- as.factor(crops_data$County)
crops_data$Related <- as.factor(crops_data$Related)
```

### Section a

We want to investigate whether two factors County and Related (and possibly their interaction) influence the crops by performing relevant ANOVA model(s), without taking Size into account. So we create and test 3 separate Null Hypotheses with a two-way ANOVA: H\_(01): There is no significant difference in the mean Crops yield across different Counties. (the means of observations grouped by country are the same) H\_(02): There is no significant difference in the mean Crops yield between cases where the landlord and tenant are related versus not related. (the means of observations grouped by related are the same) H\_(03): There is no interaction effect between County and Related on Crops yield (there is no interaction between county and related)

```
model_a <- lm(Crops ~ County * Related, data = crops_data)
anova(model_a)
```

```
## Analysis of Variance Table
##
## Response: Crops
##           Df      Sum Sq Mean Sq F value Pr(>F)
## County      2    8841441  4420721   0.7644  0.4766
## Related     1    2378957  2378957   0.4113  0.5274
## County:Related  2    1497573   748786   0.1295  0.8792
## Residuals   24  138805865  5783578
```

From this table we can see the following: For County: The p-value (0.476) is greater than 0.05, meaning we fail to reject the null hypothesis H\_(01), suggesting that there is no significant effect of the County on the Crops variable. For related: The p-value (0.527) is also greater than 0.05, meaning we fail to reject the null hypothesis H\_(02), which means that there is no significant effect of whether the landlord and tenant are related on the Crops variable. For both: The p-value (0.879) is much greater than 0.05, meaning we fail to reject the null hypothesis H\_(03), implying there is no significant interaction between County and Related on the Crops variable.

```
summary(model_a)
```

```
##
## Call:
## lm(formula = Crops ~ County * Related, data = crops_data)
##
```

```

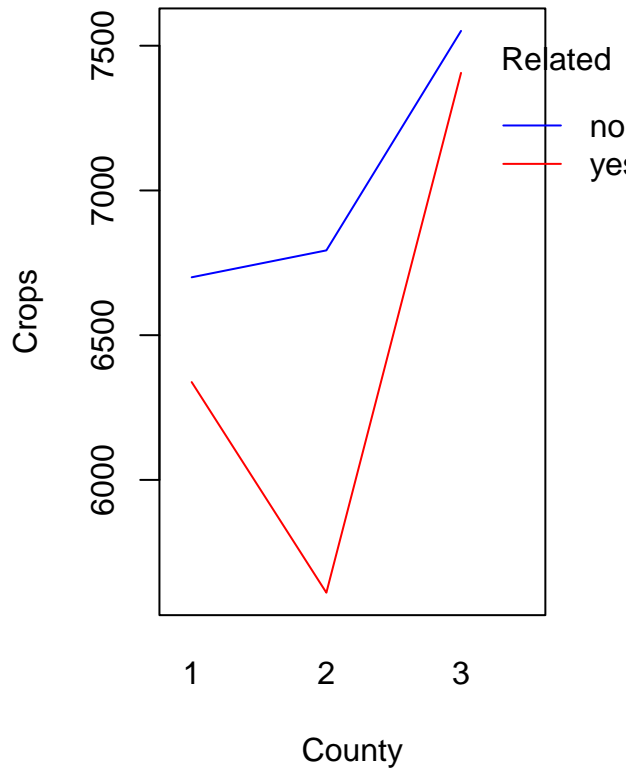
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3120.4 -1744.7  -176.9   2064.2  4806.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6700.0     1075.5   6.230 1.94e-06 ***
## County2           93.0     1521.0   0.061   0.952
## County3          851.2     1521.0   0.560   0.581
## Relatedyes       -362.0     1521.0  -0.238   0.814
## County2:Relatedyes -820.6     2151.0  -0.381   0.706
## County3:Relatedyes  217.0     2151.0   0.101   0.920
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2405 on 24 degrees of freedom
## Multiple R-squared:  0.08393,    Adjusted R-squared:  -0.1069
## F-statistic: 0.4398 on 5 and 24 DF,  p-value: 0.8163

```

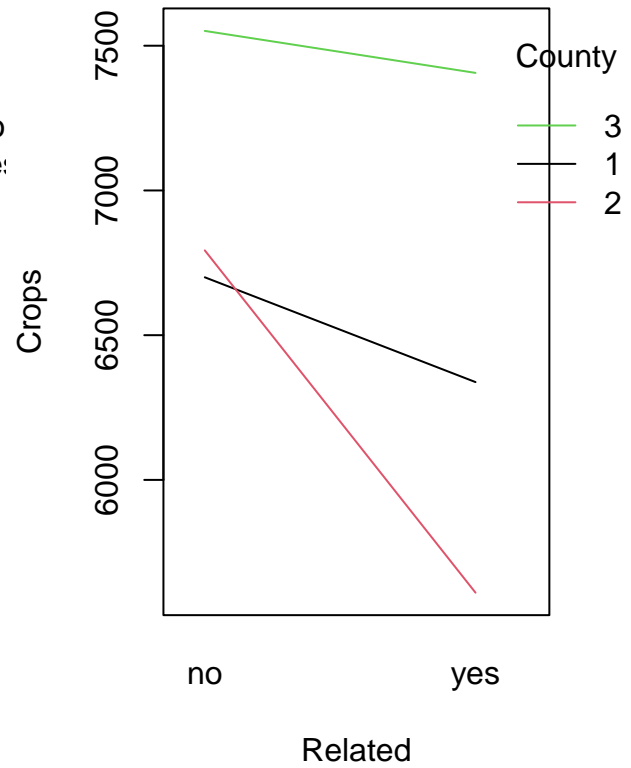
The above model summary table aligns with the ANOVA p-values as both show that none of the predictors(County, Related, or their interaction) are significant in either table.



**Interaction: County & Related**



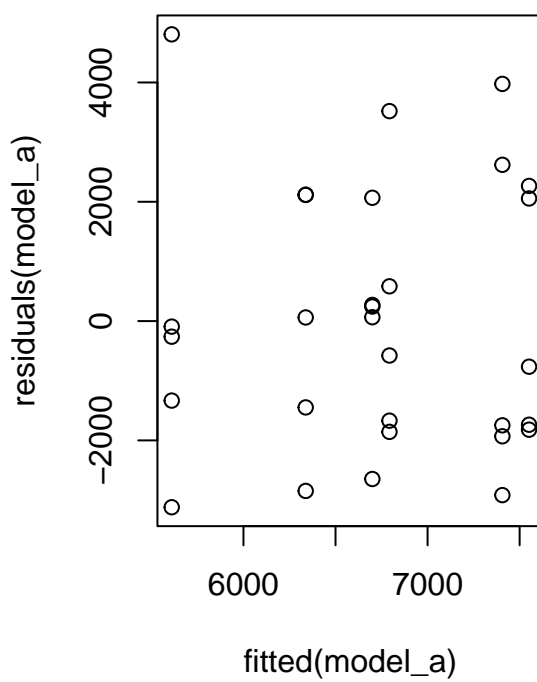
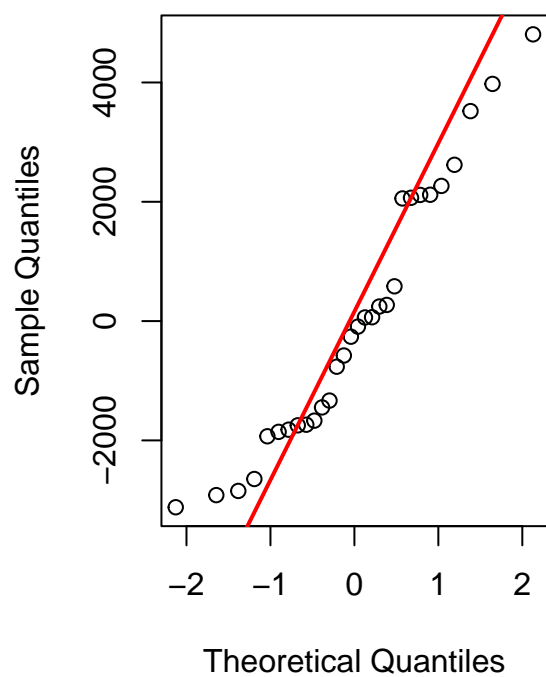
**Interaction: Related & County**



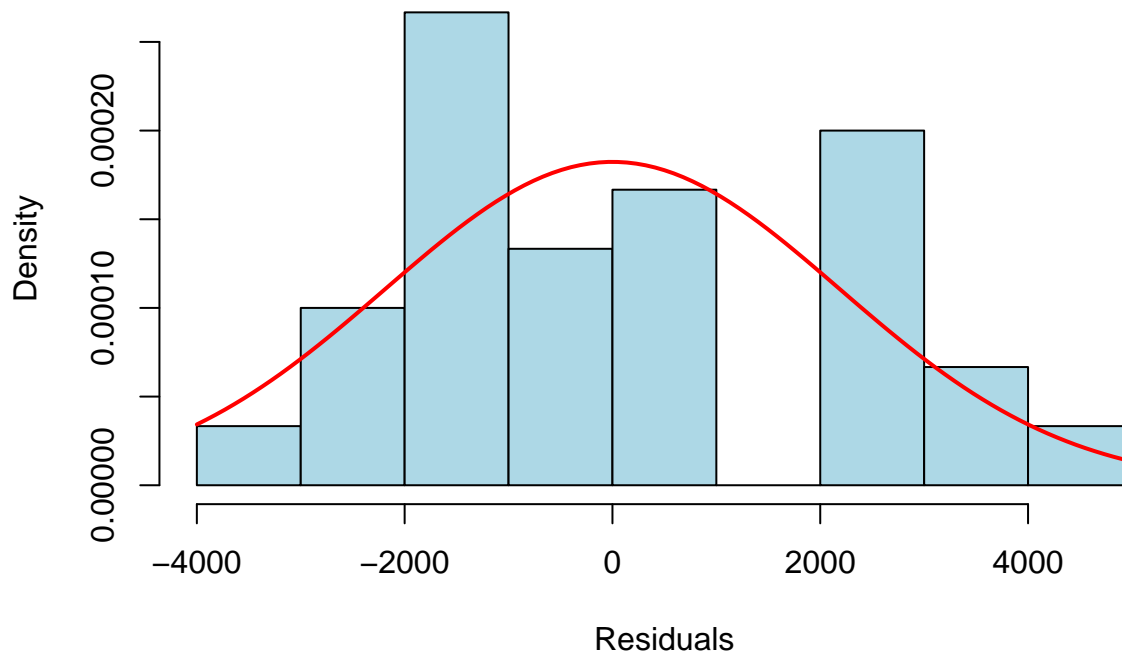
While the above plots suggest potential interaction effects (non-parallel lines), statistical analysis via ANOVA and linear regression indicates that these differences are not statistically significant ( $p > 0.05$ ). This suggests that observed variations in the plot may be due to random noise rather than meaningful effects of County or Related on Crops.

Finally, we have to check the model assumptions

**Normal Q-Q Plot**



## Histogram of Residuals



```
round(shapiro.test(residuals(model_a))$p.value, 3)
```

```
## [1] 0.099
```

```
round(shapiro.test(residuals(model_a))$statistic, 3)
```

```
##      W  
## 0.941
```

The Q-Q plot shows minor deviations from normality, particularly in the tails, but the overall trend follows the theoretical quantiles. The residuals vs. fitted plot suggests no strong patterns, indicating an approximately random distribution of residuals. Given the small sample size ( $n = 30$ ), results should be interpreted with caution, as minor departures from normality can impact statistical inference. The Shapiro-Wilk test ( $W = 0.941$ ,  $p = 0.099$ ) fails to reject the null hypothesis of normality at the 0.05 level.

We estimate the crops for County 3 when landlord and tenant are not related. We implement 2 different ways of the prediction. First way: We compute the raw mean of Crops for the subset of data where the County is “3” and the Related is “no”. This doesn’t account for any potential modeling (such as an interaction effect) or covariates and is a straightforward group mean estimate.

```
predicted_value <- with(crops_data, mean(Crops[County == "3" & Related == "no"], na.rm=TRUE))
cat("Estimated crops for County 3 (Landlord and Tenant NOT related):", predicted_value, "\n")
```

```
## Estimated crops for County 3 (Landlord and Tenant NOT related): 7551.2
```

Second way: We use the emmeans function to estimate the adjusted mean crops yield for a typical farm in County 3 where the landlord and tenant are not related, based on the fitted linear model (anova\_model). The emmeans function provides the estimated marginal means, which account for the effects of the factors (County and Related) while adjusting for any interactions.

```
emm_results <- emmeans::emmeans(model_a, ~ County * Related)
emm_summary <- as.data.frame(emm_results)
county3_not_related <- emm_summary[emm_summary$County == "3" & emm_summary$Related == "no", ]
cat("Estimated crops for County 3 (Landlord and Tenant NOT related):", county3_not_related$emmeans, "\n")
```

```
## Estimated crops for County 3 (Landlord and Tenant NOT related): 7551.2
```

## Section b

First, we ensure correct formatting of the data

```
crops_data$County <- as.factor(crops_data$County)
crops_data$Related <- as.factor(crops_data$Related)
crops_data$Size <- as.numeric(crops_data$Size)
crops_data$Crops <- as.numeric(crops_data$Crops)
```

We define 3 different models: Model\_county\_size: This model examines how crop yields are influenced by County, Related, and Size, with an additional focus on the interaction between County and Size. It does not include an interaction term for Related.

```
model_county_size <- lm(Crops ~ Size * County + Related, data = crops_data)
anova(model_county_size)
```

```
## Analysis of Variance Table
##
## Response: Crops
##           Df    Sum Sq   Mean Sq  F value    Pr(>F)
## Size       1 119569344 119569344 135.6241 4.01e-11 ***
## County     2   767179    383589   0.4351  0.65242
## Related    1  1381334   1381334   1.5668  0.22325
## Size:County 2   9528654   4764327   5.4040  0.01192 *
## Residuals 23  20277325    881623
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model\_related\_size: This model evaluates how County, Related, and Size affect crop yields, and it adds an interaction term for Related and Size to test if the effect of Size on crop yields depends on whether the landlord and tenant are related.

```
model_related_size <- lm(Crops ~ Size * Related + County, data = crops_data)
anova(model_related_size)
```

```
## Analysis of Variance Table
##
## Response: Crops
##           Df      Sum Sq   Mean Sq  F value    Pr(>F)
## Size       1 119569344 119569344 100.8587 4.521e-10 ***
## Related    1  1380585   1380585    1.1645  0.2913
## County     2   767927    383964    0.3239  0.7264
## Size:Related 1   1353666   1353666    1.1418  0.2959
## Residuals  24 28452313   1185513
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model\_additive: This model assumes that the effect of each factor on crop yield is independent of the others. It does not test for any interaction effects, only the individual contributions of County, Related, and Size to crop yields.

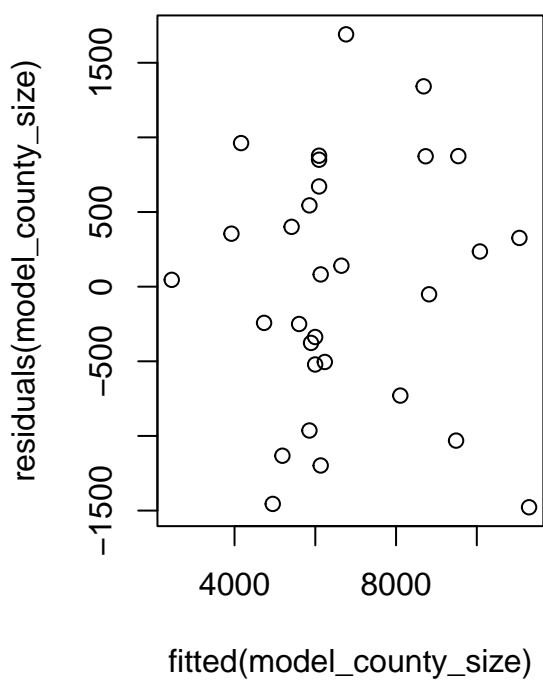
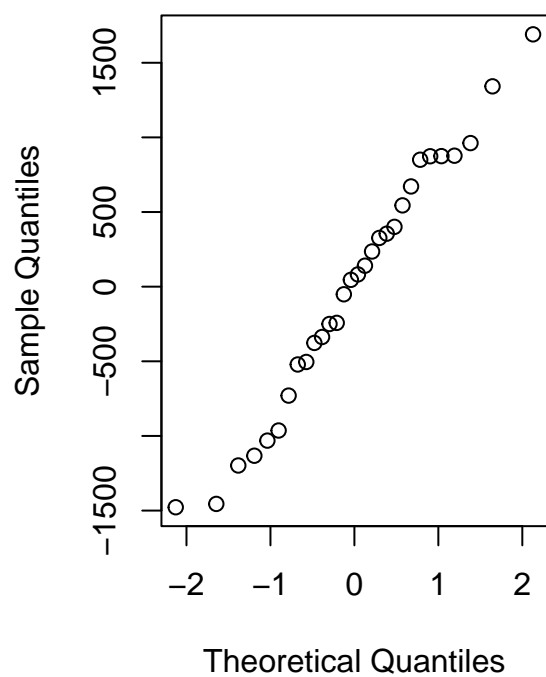
```
model_additive <- lm(Crops ~ Size + County + Related, data = crops_data)
anova(model_additive)
```

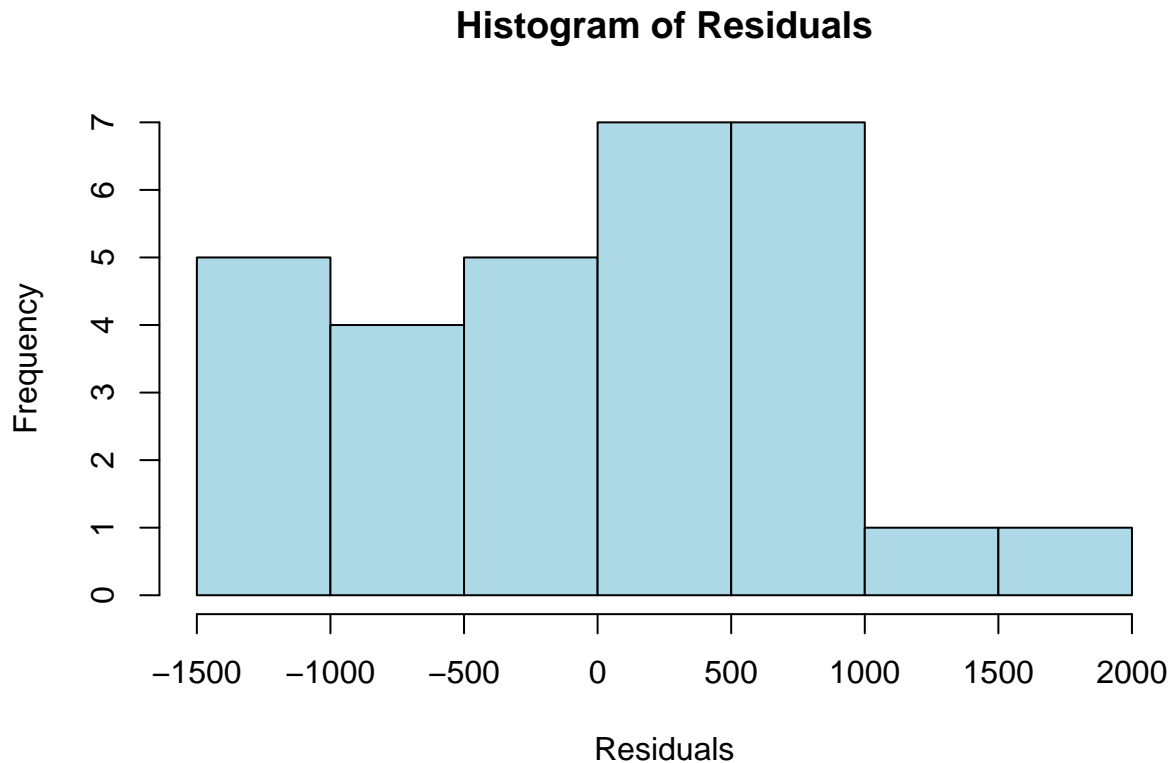
```
## Analysis of Variance Table
##
## Response: Crops
##           Df      Sum Sq   Mean Sq  F value    Pr(>F)
## Size       1 119569344 119569344 100.2897 3.114e-10 ***
## County     2   767179    383589    0.3217  0.7278
## Related    1  1381334   1381334    1.1586  0.2920
## Residuals  25 29805979   1192239
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have tested interaction models as well as purely additive. The interaction Size-Related and the individual effect of County and Related are insignificant(all 3 have p-values>0.5). Therefore, the best model is model\_county\_size, since it shows the significance of Size and of the interaction Size-County.

Finally, we can check this model's assumptions.

**Normal Q-Q Plot**





We check normality assumptions for the full model only, since the other 3 are subset of it.

```
round(shapiro.test(residuals(model_county_size))$p.value, 3)
```

```
## [1] 0.733
```

```
round(shapiro.test(residuals(model_county_size))$statistic, 3)
```

```
##      W
```

```
## 0.977
```

The Shapiro-Wilk normality test for the residuals of `model_county_size` returned a p-value of 0.733 > 0.05. This indicates that we fail to reject the null hypothesis, meaning there is no significant evidence to suggest that the residuals deviate from a normal distribution. From the QQ-plot, the points mainly follow a straight line and the histogram has a bell-like shape.

## Section c

```
summary(model_county_size)
```

```
##
## Call:
## lm(formula = Crops ~ Size * County + Related, data = crops_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1477.64  -517.10   63.89   639.62  1690.29
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2461.014    929.764   2.647  0.01441 *
## Size           22.704      4.765   4.764 8.38e-05 ***
## County2      -4214.050    1447.242  -2.912  0.00785 **
## County3      -1284.813    1302.578  -0.986  0.33422
## Relatedyes    -239.099     347.916  -0.687  0.49881
## Size:County2    26.590      8.091   3.286  0.00323 **
## Size:County3     8.916      6.398   1.394  0.17676
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 938.9 on 23 degrees of freedom
## Multiple R-squared:  0.8662, Adjusted R-squared:  0.8313
## F-statistic: 24.81 on 6 and 23 DF,  p-value: 5.851e-09
```

The coefficient for Size is 22.704 ( $p < 0.001$ ), meaning that for County 1 (the reference level), an increase of 1 unit in Size leads to an expected increase of 22.7 units in Crops, assuming all other factors remain constant. This effect is statistically significant, indicating that Size has a strong positive influence on Crops. County 2 has a negative coefficient of -4214.050 ( $p < 0.01$ ), meaning that, for the same Size, crops in County 2 are expected to be 4214 units lower than in County 1. County 3 has a coefficient of -1284.813, but it is not statistically significant ( $p = 0.334 > 0.05$ ), suggesting that the difference between County 3 and County 1 is not strong enough to be conclusive. The coefficient for Related is -239.099, but its p-value is 0.499, meaning that this effect is not statistically significant. This suggests that the relationship between the landlord and tenant does not significantly influence crop yields. Size:County2 has a coefficient of 26.590 ( $p = 0.003 < 0.05$ ), meaning that the effect of Size on Crops in County 2 is higher than in County 1. Specifically, the crop yield increase per unit Size is  $22.7 + 26.6 = 49.3$  in County 2. Size:County3 has a coefficient of 8.916, but it is not statistically significant ( $p = 0.176$ ), meaning we cannot confidently conclude that the Size effect in County 3 differs significantly from County 1. The model explains 86.6% of the variation in crops ( $R^2 = 0.866$ ), making it a strong explanatory model.

```
confint(model_county_size)
```

```
##              2.5 %      97.5 %
## (Intercept)  537.651576 4384.37722
## Size         12.845887  32.56210
## County2     -7207.896850 -1220.20244
## County3     -3979.399914  1409.77389
```



```
## Relatedyes      -958.817141  480.61942
## Size:County2     9.852682    43.32644
## Size:County3    -4.319019    22.15129
```

For Size, the CI does not include 0 confirming a strong and statistically significant positive effect of Size on Crops. The CI of County2 does not include 0 confirming statistical significance, while for County3 it does meaning there is no strong evidence that County3 differs significantly from County1. The CI of Related includes zero, so this effect is not statistically significant and we cannot conclude that relationship status affects crop yield. The CI of the Size-County2 interaction does not include 0 confirming significance, while the same for County3 includes zero, so we cannot confidently say that the effect of Size is different in County3 compared to County1.

## Section d

```
pred <- emmeans::emmeans(model_county_size, specs = ~ County * Related * Size,
                          at = list(County = "2", Size = 165, Related = "yes"))

summary(pred)
```

```
## County Related Size emmean SE df lower.CL upper.CL
## 2      yes      165  6141 345 23    5428    6855
##
## Confidence level used: 0.95
```

The predicted yield crops for a farm from County 2 of size 165, with related landlord and tenant is 6141, with a 95% CI: (5428, 6855)

```
pred_se <- summary(pred)$SE

# Compute the prediction variance
pred_var <- pred_se^2

# Print results
cat("Prediction Variance:", pred_var, "\n")
```

```
## Prediction Variance: 118896.9
```

```
cat("Residual Variance (sigma^2):", sigma(model_county_size)^2, "\n")
```

```
## Residual Variance (sigma^2): 881622.8
```

```
cat("Total Variance:", pred_var + sigma(model_county_size)^2)
```

```
## Total Variance: 1000520
```

The prediction variance is much smaller than the residual variance, which suggests that the model's coefficient estimates are relatively stable and that most of the uncertainty comes from residual variation rather than parameter estimation.

### Exercise 3

- a) Present an R-code for the randomization process to distribute soil additives over plots in such a way that each soil additive is received exactly by two plots within each block.
- b) Make a plot to show the average yield per block for the soil treated with nitrogen and for the soil that did not receive nitrogen, and comment.
- c) Conduct a full two-way ANOVA with the response variable *yield* and the two factors *block* and *N*. Was it sensible to include factor *block* into this model? Can we also apply the Friedman test for this situation?
- d) Investigate other possible models with all the factors combined, restricting to only one (pair-wise) interaction term of factors *N*, *P* and *K* with *block* in one model (no need to check the model assumptions for all the models). Test for the presence of main effects of *N*, *P* and *K*, possibly taking into account factor *block*. Give your favorite model and motivate your choice.
- e) For the resulting model from d), investigate how the involved factors influence *yield*. Which combination of the levels of the factors in the model leads to the largest yield?
- f) Recall the main question of interest. In this light, for the resulting model from d) perform a mixed effects analysis, modeling the block variable as a random effect by using the function *lmer*. Compare your results to the results found by using the fixed effects model. (You will need to install the R-package *lme4*, which is not included in the standard distribution of R.)