Opgave-1

2024-09-20

♣

Initialize

```r
# Import libraries
library("ggplot2")
library("tidyverse")
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ lubridate 1.9.3     ✓ tibble    3.2.1
## ✓ purrr     1.0.2     ✓ tidyr     1.3.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```r
library("isdals")

# Load data
data("tartar")
```

Opgave 1 (HS.18)

Opgave 1

```r
head(tartar, n = 3)
```

```
##     treat index
## 1 Control  0.49
## 2 Control  1.05
## 3 Control  0.79
```

Vi har en kategorisk forklarende variabel og en kvantitativ kontinuert variabel. Derfor bruger vi Anova til at analysere påvirkningen på gennemsittet givet variablen treat.

```
# refactor to make control group the primary group
tartar$treat <- relevel(factor(tartar$treat), ref="Control")
m1 <- lm(index ~ treat , data = tartar)
```

Opgave 2

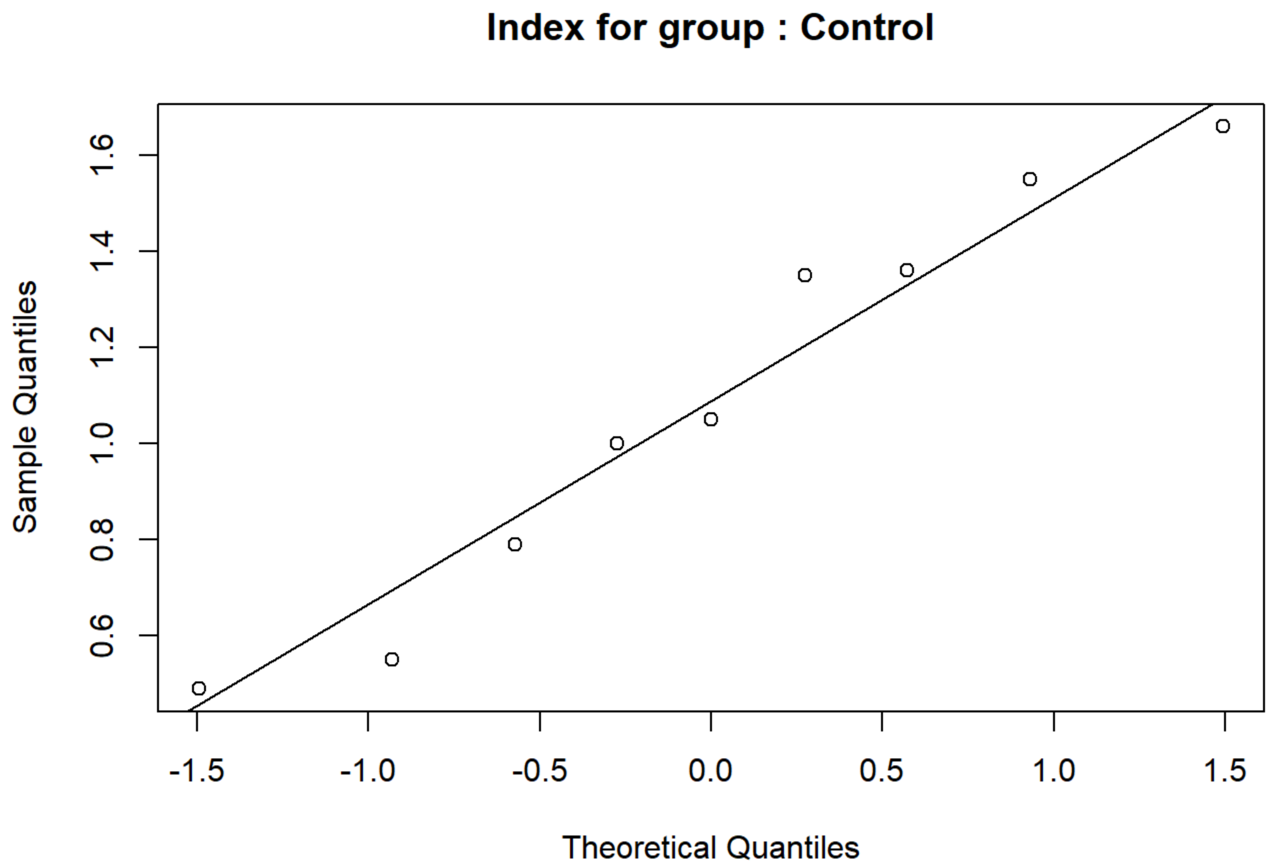
Model antagelse: Data: $(y_1, x_1), \ldots, (y_2 6, x_2 6)$ hvor responsvariablen y et index på hvor høj en grad hunden har problemer med tandsten og og den diskrete forklarende variable x angiver de forskellige typer af behandlinger. Svarende til to behandlinger og en kontrol gruppe.

Anova: Responsvariablen $y_1, \ldots, y_n$ er uafhænige og normaltfordelt $ y_i N(_i, ^2) $ med sammen spredning for alle grupper. Derudover afhænger middelværdien $\mu_i = \alpha_{g_{(i)}}$ af gruppen $g(i)$ Vi antager at alle restled $e_1, .., e_n$ har samme fordeling $e_i \sim iid.\ N(0, 1)$

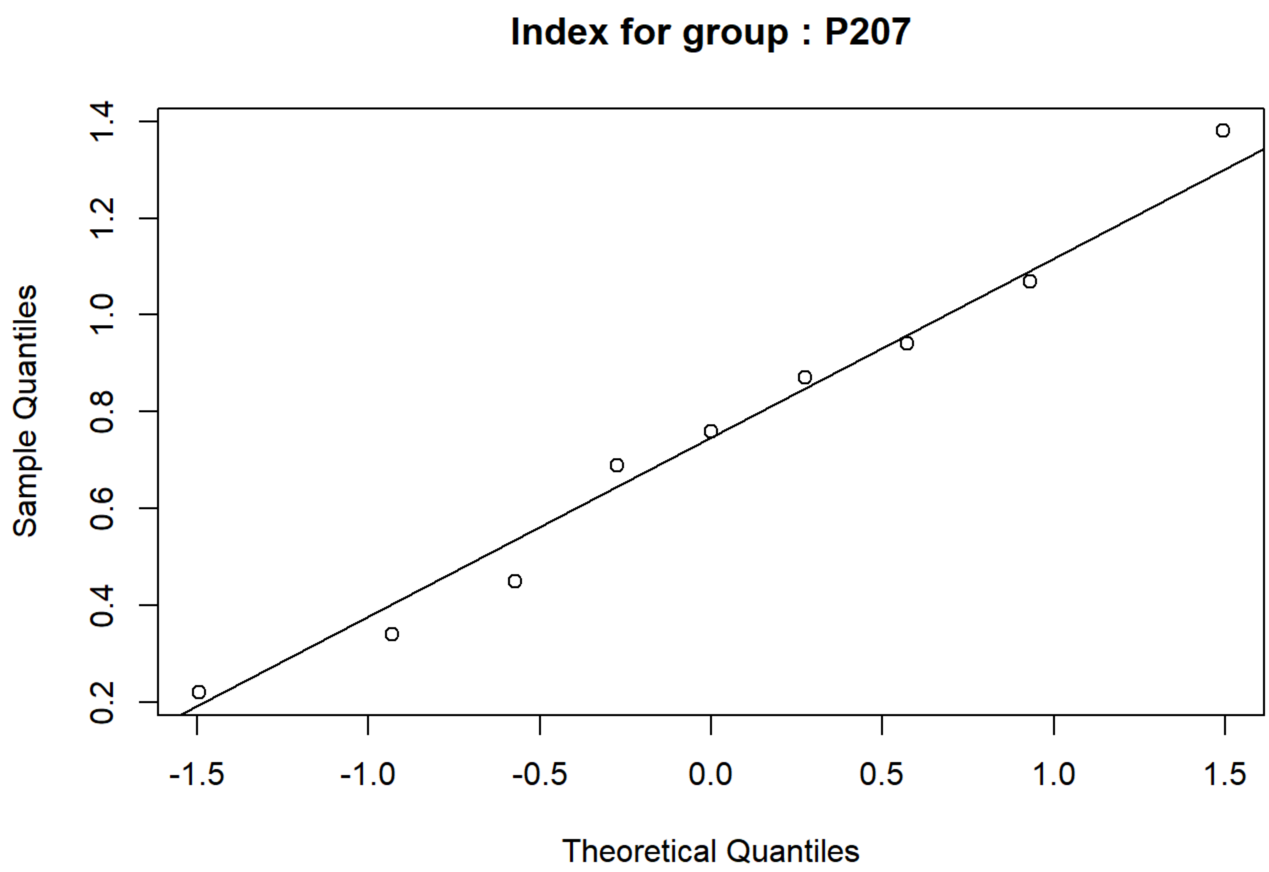
Plots - QQ-Plot og histogram til normal fordeling - Boxplot til analyse af forskel mellem data

Er data normalt fordelt?

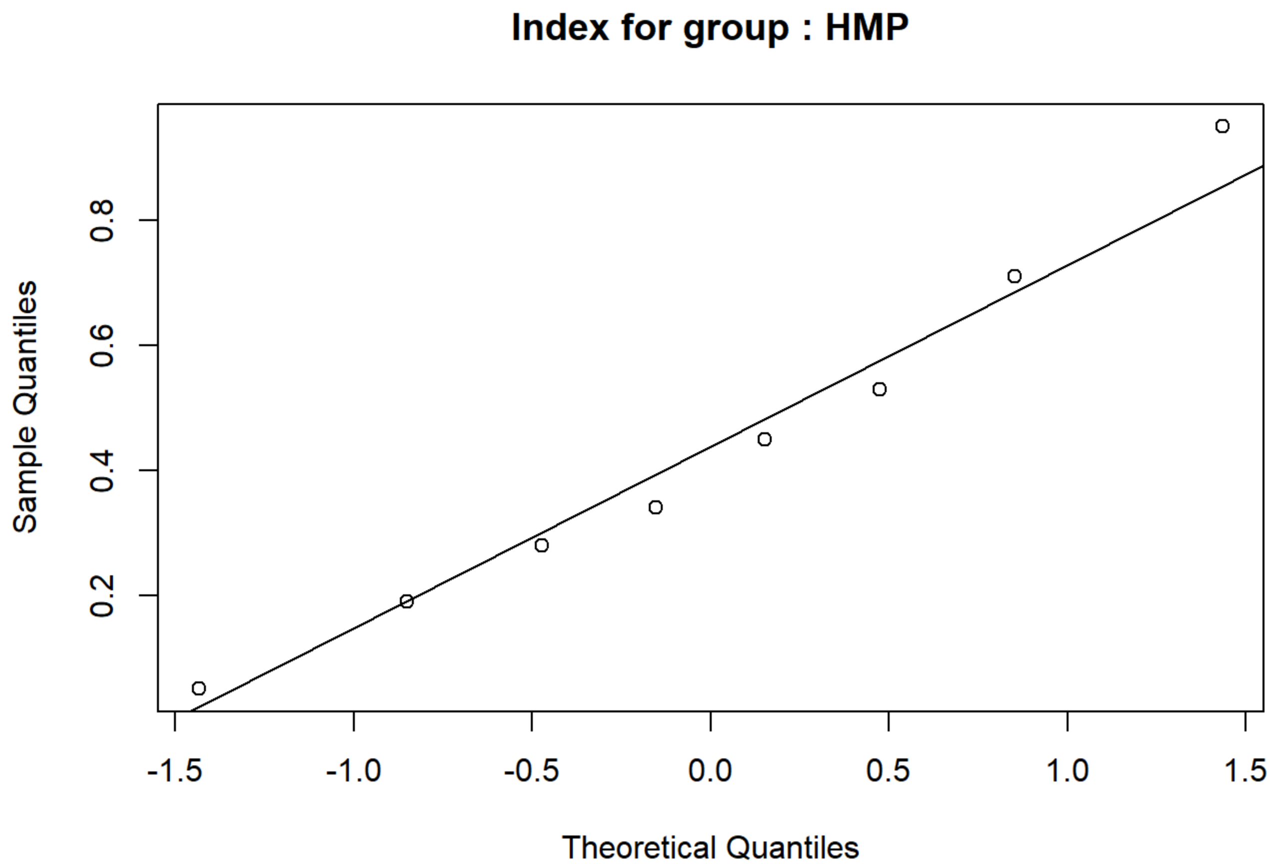
```
tartar_contol <- subset(tartar, treat=="Control")
tartar_P207 <- subset(tartar, treat=="P2O7")
tartar_HMP <- subset(tartar, treat=="HMP")

qqnorm(tartar_contol$index, main = "Index for group : Control")
abline(a = mean(tartar_contol$index), b = sd(tartar_contol$index))
```




```
qqnorm(tartar_P207$index, main = "Index for group : P207")
abline(a = mean(tartar_P207$index), b = sd(tartar_P207$index))
```

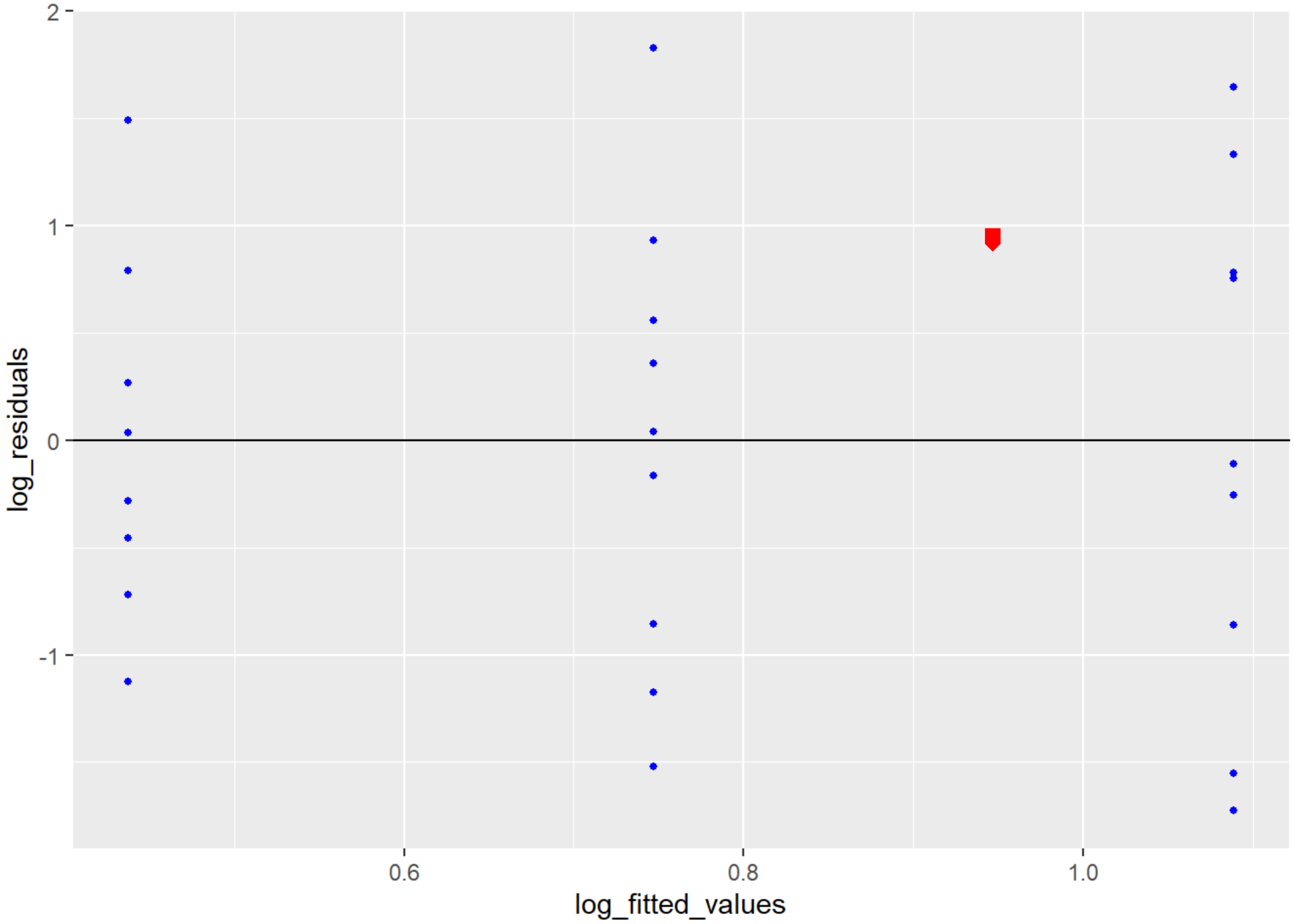



```
qqnorm(tartar_HMP$index, main = "Index for group : HMP")
abline(a = mean(tartar_HMP$index), b = sd(tartar_HMP$index))
```

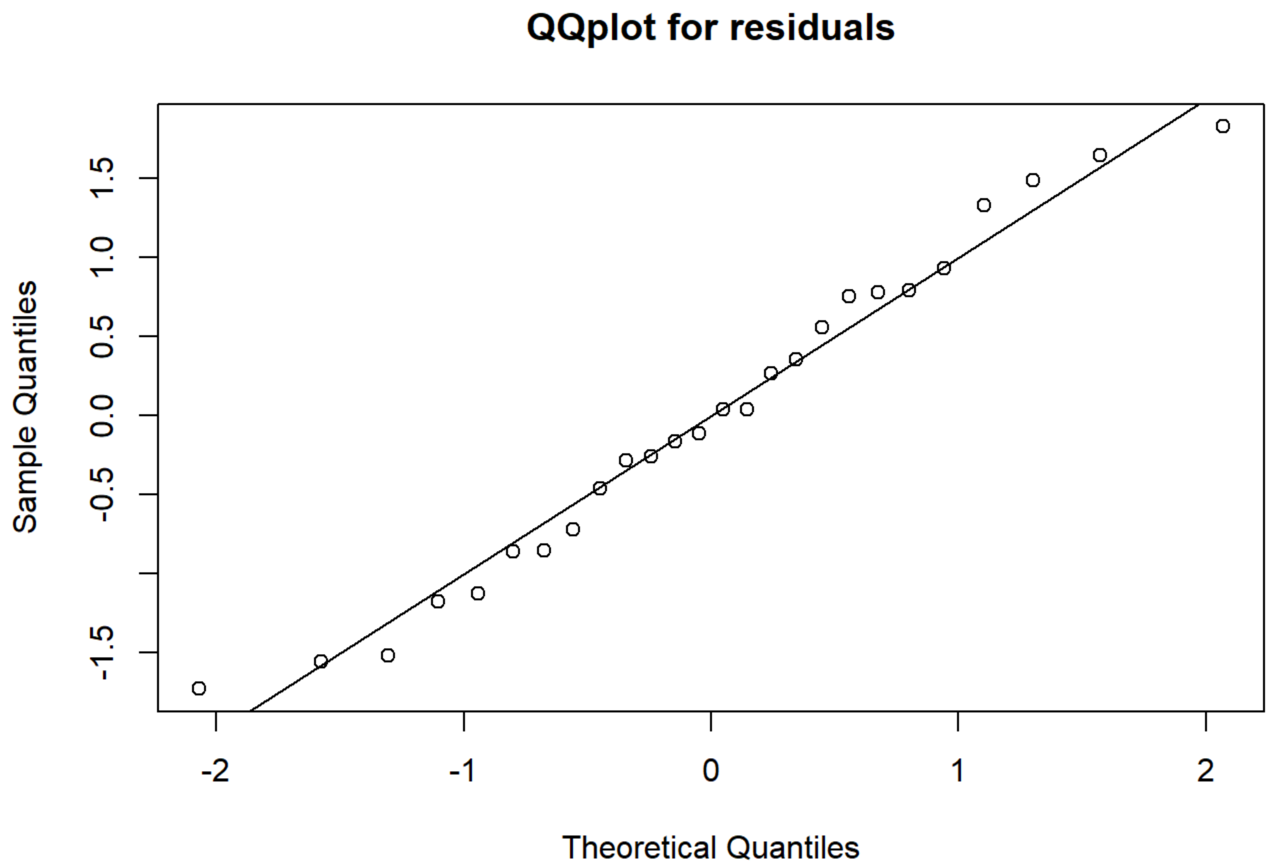


Vi kan se på QQ-plottet at der ikke er nogen systematisk afvigelse fra identitets linjen, hvilket tyder på at data falder i de kvartiler, som der forventes ved en normalfordeling.

```
m2 <- lm(index ~ treat, data = tartar)

tartar$log_fitted_values <- fitted(m2)
tartar$log_residuals <- rstandard(m2)

ggplot(data = tartar, aes(x = log_fitted_values, y = log_residuals)) +
  geom_point(color = "blue", size = 1) + # Scatter
  geom_abline(slope = 0, intercept = 0) # Line
```




```
qqnorm(tartar$log_residuals, main = "QQplot for residuals")
abline(a = 0, b = 1)
```

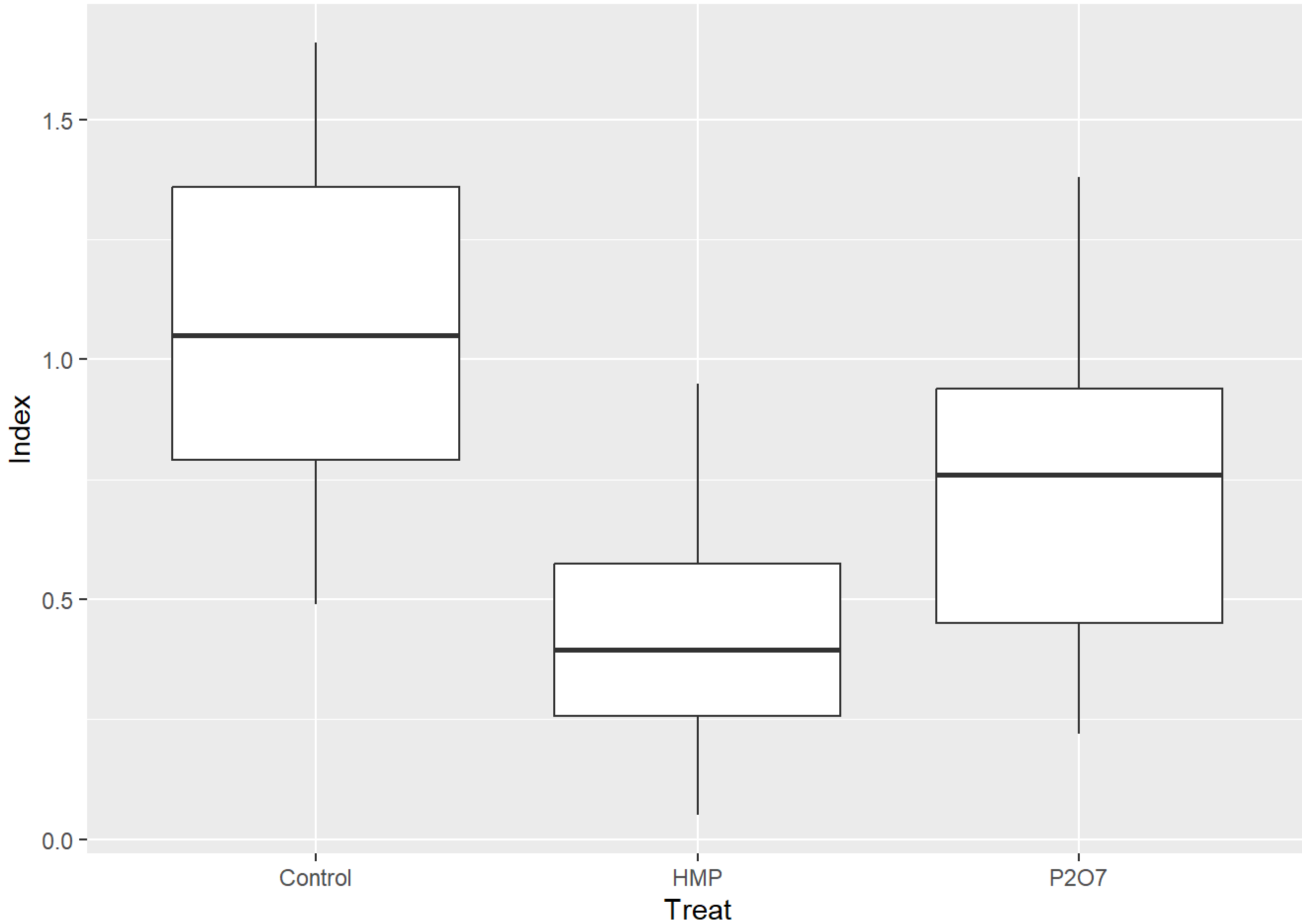


Antagelsen om at $e_1, .., e_n$ er standard normalt fordelt holder, der er ca. lige mange punkter over og under den vandrette linje på residualplottet. Og i QQ-plotet ser vi ingen systematisk afvigelse fra den ligefrem propotionalle linje. Herved validere vi at antagelse om spredningen er den samme for alle grupper.

Afhænger middelværi af gruppen

Boxplot

```
## Boxplot
ggplot(data = tartar, aes(x = treat, y = index)) +
  geom_boxplot(outliers = TRUE) +
  xlab("Treat") + ylab("Index")
```



Plottet viser at der er en markant between-group variation, hvilket tyder på at den forklarende variabel har en påvirkning af på værdien af responsvariablen.

Dertil ser vi en større witihn-group variation i control og P207, end vi ser for HMP.

Opgave 3

Vi bruger LM til at udregne estimatet for den forventede værdi $\hat{\mu}$

```
m2 <- lm(index ~ treat -1 , data = tartar)
summary(m2)
```

```
##
## Call:
## lm(formula = index ~ treat - 1, data = tartar)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59889 -0.28437 -0.01319  0.26861  0.63333
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## treatControl   1.0889     0.1226   8.881 6.83e-09 ***
## treatHMP       0.4375     0.1301   3.364  0.00268 **
## treatP207      0.7467     0.1226   6.090 3.27e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3678 on 23 degrees of freedom
## Multiple R-squared:  0.8469, Adjusted R-squared:  0.827
## F-statistic: 42.42 on 3 and 23 DF,  p-value: 1.547e-09
```

Vi forventer at en hund som ikke er blevet behandlet har et index på 1.0889 Vi forventer at en hund som er blevet behandlet med P207 harr et index på 0.7467

Opgave 4

$$H_0 : \alpha_1 = \alpha_2 = \alpha_2$$

- *fuld model*: ensidet anova med 3 gruppegennemsnit
- *nulmodel*: model hvor data opfattes som en sample fra en gruppe, altså hvor alle gruppegennemsnit antages at være ens

```
n_groups <- 3
lenght <- nrow(tartar)
```

```
fullModel <- lm(index ~ treat , data = tartar)
nulModel <- lm(index ~ 1 , data = tartar)
anova(nulModel, fullModel)
```

```
## Analysis of Variance Table
##
## Model 1: index ~ 1
## Model 2: index ~ treat
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     25 4.9166
## 2     23 3.1120  2    1.8046 6.6684 0.005198 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
1 - pf(6.6684, 3-1, 28-5)
```

```
## [1] 0.005198563
```

Vi anvender F-test til at sige noget om gennemsnittet på tværs af alle grupper, ved at sammenligne mellem-grupe variationen med indenfor-gruppe variationen. Da p-værdien for f-statistic er mindre end 0.05 forkastes nul-hypotesen på et signifikansnivue på 95%, for at alle middelværdier er nes.

Der er altså mindst en af grupperne, hvis middelværedi afviger fra kontrol gruppen.

Opgave 5

```
confint(m1)
```

```
##                  2.5 %      97.5 %
## (Intercept)  0.8352441  1.34253368
## treatHMP    -1.0211365 -0.28164124
## treatP2O7   -0.7009301  0.01648568
```

Vi kan med 95% sikkerhed konkludere at den sande forskel ligger i intervallet (-1.0211365, -0.28164124). Da 0 ikke indgår i intervallet er der evidens for at HMP behandlingen virker.

Opgave Eksamen 2021

Opgave 1

```
library("readxl")
?read.table
```

```
## starting httpd help server ... done
```

```
data1 <- read.table(file = "data/feb2021opg1.txt", header = 1)
head(data1, n = 30)
```

```
##         region kommune_id   dec   jan
## 1   Syddanmark        580  5.68  3.57
## 2  Nordjylland        851 11.22  4.28
## 3  Midtjylland        751 17.32  5.23
## 4   Syddanmark        492  1.34  2.68
## 5  Hovedstaden        165 27.66  9.74
## 6  Hovedstaden        201 15.60  5.27
## 7   Syddanmark        420  5.74  2.64
## 8  Hovedstaden        151 24.79  8.02
## 9   Syddanmark        530  4.51  2.03
## 10 Hovedstaden        400  5.39  0.94
## 11 Hovedstaden        153 26.39 11.74
## 12 Nordjylland        810  6.25  4.74
## 13 Hovedstaden        101 26.14  7.06
## 14 Hovedstaden        155 22.91  6.07
## 15 Hovedstaden        240 20.90  5.77
## 16  Syddanmark        561  7.50  4.56
## 17  Syddanmark        430  5.92  2.75
## 18  Syddanmark        563  6.02  0.29
## 19 Midtjylland        710 10.54  4.26
## 20    Sjælland        320 13.26  4.76
## 21 Hovedstaden        210 14.32  9.69
## 22  Syddanmark        607  6.19  3.93
## 23 Hovedstaden        147 25.77  6.35
## 24 Nordjylland        813  5.95  5.20
## 25 Hovedstaden        250 13.13  5.44
## 26 Hovedstaden        190 17.60  5.66
## 27 Hovedstaden        157 20.41  5.44
## 28 Hovedstaden        159 22.80  8.17
## 29 Hovedstaden        161 24.78  8.26
## 30    Sjælland        253 21.24  8.37
```

Vi laver en ny kollonne med difference imellem smittede in december og januar.

Vi opstiller en nul hypotese $ H_0: = 0 $ altså at der ikke er nogen forskel, samt en alternativ hypotese $ H_1: = 1 $

```
data1$fald <-  data1$dec - data1$jan


t.test(data1$fald, mu = 0)
```

```
##
##  One Sample t-test
##
## data:  data1$fald
## t = 13.834, df = 97, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  6.624328 8.843427
## sample estimates:
## mean of x
##  7.733878
```

Vi forkaster vores 0-hypoteser da vi for en p-værdi på 2.2e-16, hvilket praktisk talt er 0. Dette bliver understøttet af vores konfidens interval, da 0 ikke ligger i intervallet.

Opgave 2

```
m3 <- lm(fald ~ dec, data = data1)
summary(m3)
```

```
##
## Call:
## lm(formula = fald ~ dec, data = data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.218 -0.736  0.127  1.076  2.863
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.47123    0.29622  -4.967 2.96e-06 ***
## dec          0.72060    0.02009  35.877  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.466 on 96 degrees of freedom
## Multiple R-squared:  0.9306, Adjusted R-squared:  0.9299
## F-statistic:  1287 on 1 and 96 DF,  p-value: < 2.2e-16
```

$$\hat{\alpha} = -1.471, \quad \hat{\beta} = 0.720$$

De to ukende parameter skæring (alpah) og hældning (beta) er angivet overfor.

Opgave 3

Faldet $\hat{\beta}$ er det samme for alle værdier af x, så vi kan ikke konkludere ud fra den linære model at der er forskel i falde givet at der er mange smittede i december.

Opgave 4

```
predict(m3, newdata = data.frame(dec = 10), interval = "p")
```

```
##        fit      lwr      upr
## 1 5.734795 2.808736 8.660854
```

$$\alpha + \beta x = f(x) \Rightarrow -1.471 + 0.72 \cdot 10 = 5.735$$

Intervallet vil indeholde 95 % af nye/fremtidige målinger. Da 5.7 er indenfor intervallet er det ikke en uanmindelig værdi. Vores interval er lidt større end konfidensintervallet for samme værdi, dette er fordi der skal tages højde for e_i, som vi antager har en spredning på 1.