MatStat Eksamen 2021

Eksamens nummer: 1

Eksamensdato 24. juni, 2021

Opgave 2

Vi kan starte med at indlæse data:

```
library(tidyverse)
vaccine <- read_csv("covid19vaccine.txt", col_types = cols(U = col_factor(), T = col_factor
()))
```

Opgave 2.1

(Primære besvarelse af opgaven her i R)

Vi er i delopgave 2.1 og 2.2 blevet bedt om kun at bruge datasættet for `vaccine_uge5` , som er givet ved

```
vaccine_uge5 <- filter(vaccine, U == 5)
```

Vi kan yderligere subsette vores datasæt ved;

```
vaccine_uge5Ind <- subset(vaccine_uge5, select = c(3,5))
head(vaccine_uge5Ind)
```

```
## # A tibble: 6 x 2
##   T           X
##   <fct>   <dbl>
## 1 KK    6175.
## 2 VZ    1742.
## 3 KK    1081.
## 4 PL       3.38
## 5 PL       5.89
## 6 KK    1174.
```

Vi kan på basis af dette fitte to `lm` modeller ud fra faktoren `T` , med respons hhv. X, $log(X)$. Dette kan gøres i R ved følgende:

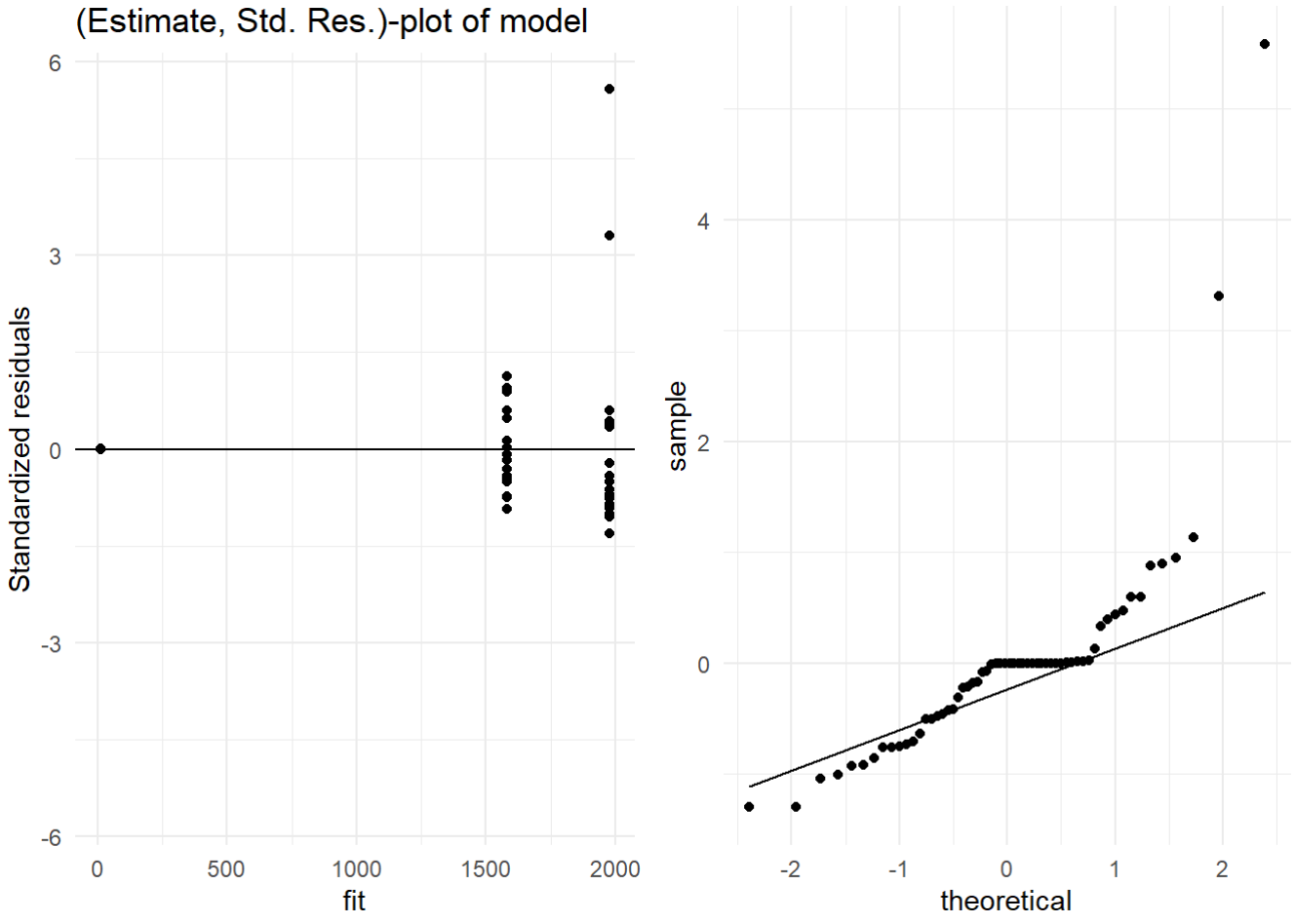
```
model <- lm(X~T,data=vaccine_uge5Ind)
logmodel <- lm(log(X)~T,data=vaccine_uge5Ind)
```

Følgende hjemmelavede plotfunktion benyttes til at plotte hhv. std. res. og QQplot for hver af de to modeller.

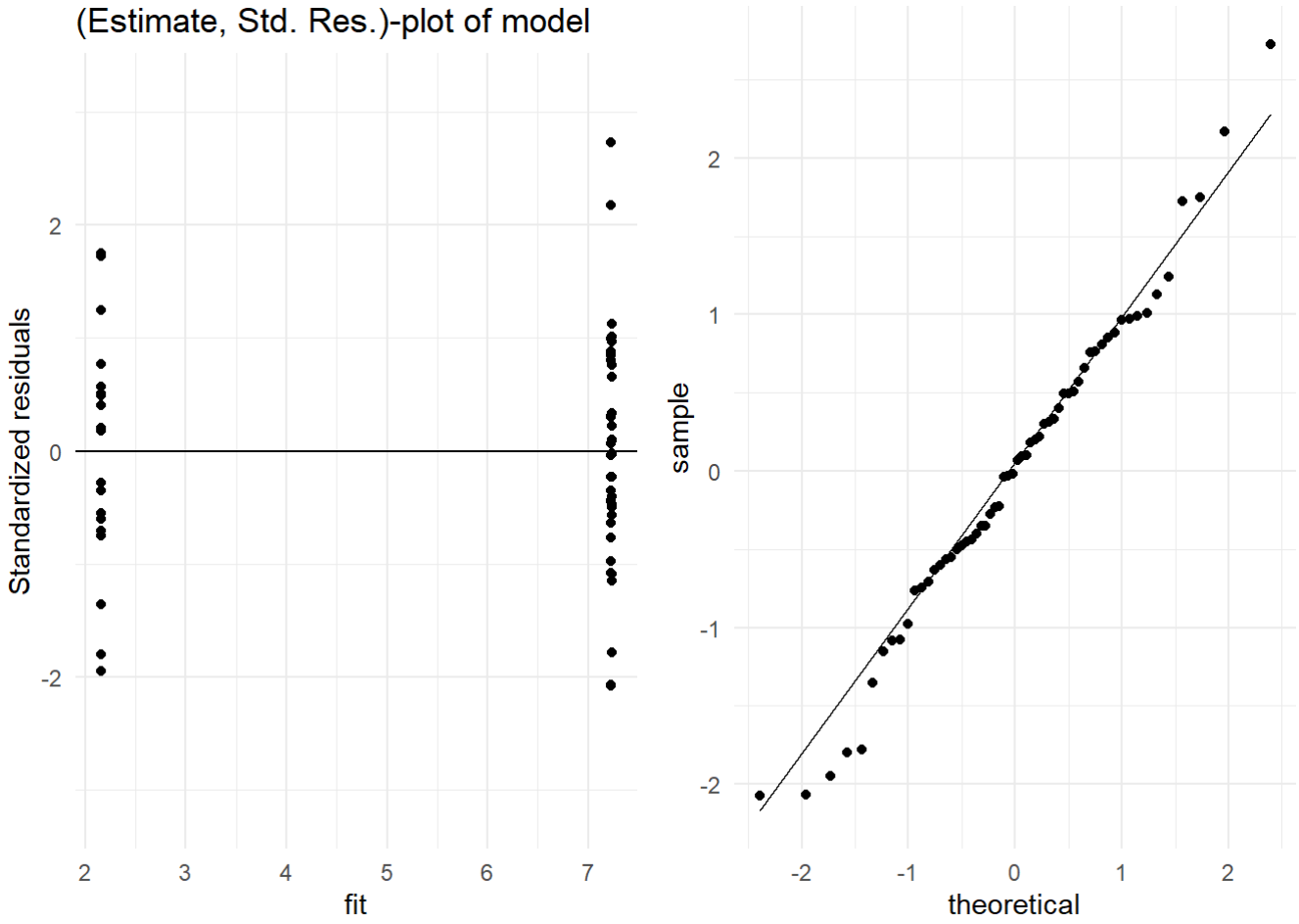
```
library(gridExtra)
Stdresplot <- function(model, main = paste("(Estimate, Std. Res.)-plot of", deparse(substitut
e(model))), ylab ="Standardized residuals", ...) {
 fit <- fitted(model)
 rst <- rstandard(model)
 qplot(fit, rst, main = main, ylab = ylab, ylim = c(-max(3.2,max(abs(rst))), max(3.2,max(abs
(rst)))) )+geom_hline(yintercept = 0) #Largest symmetric interval (around 0) of (-3.2,3.2) or
(-Largest absolute rst, largest absolute rst)
}
QQplotdraw <- function(model, main = paste("Normal QQ-plot of", deparse(substitute(model))),
 xlab = "Theoretical Quantiles", ylab ="Sample Quantiles", ...) {
   rst <- rstandard(model)
   #dataname <- getCall(lm_LT)$data
   ggplot(data = eval(getCall(model)$data), main = main, xlab = xlab, ylab = ylab) + geom_qq
() + geom_qq_line() + aes(sample = rst)
} #main, xlab, ylab call do not work for some reason
StdresQQPlot <- function(model,...) {
   p1 <- Stdresplot(model,...)
   p2 <- QQplotdraw(model,...)
   #library(gridExtra)
   grid.arrange(p1,p2, ncol = 2)
}
```

Således at vi kan plotte:

```
StdresQQPlot(model)
```




```
StdresQQPlot(logmodel)
```



I forhold til modelhypotesen for et ensidet faktor-analyse, kan det på overstående plot ses, at det log-transformerede plot passer langt bedre til antagelserne. Specifikt ser vi at vi ved at log-transformerer får tæmmet de 'outliers' vi har på std. res. plottet for den ikke-transformerede model, således at vi i højere grad kan være tilfredse med linearitets antagelsen i et-faktor modellen. - hertil ser vi at varians homogeniteten ser bedre ud i `logmodel` og specielt ser vi fra QQplottet at det er meget mere rimeligt med normalfordelingsantagelsen, når vi først har log-transformeret. - Vi vil derfor arbejde videre med `logmodel` .

Opgave 2.2

(Primære besvarelse af opgaven her i R)

For at analyserer om der er nogen forskel i effekt ved hver af de to vacciner kan vi kigge på summary;

```
summary(logmodel)
```

```
##
## Call:
## lm(formula = log(X) ~ T, data = vaccine_uge5Ind)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.43294 -0.39569  0.01726  0.47106  1.88359
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.227355   0.158596  45.571   <2e-16 ***
## TVZ          0.006945   0.224289   0.031    0.975
## TPL         -5.073055   0.224289 -22.618   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7093 on 57 degrees of freedom
## Multiple R-squared:  0.923,  Adjusted R-squared:  0.9203
## F-statistic: 341.5 on 2 and 57 DF,  p-value: < 2.2e-16
```

Vi har `KK` som intercept parameter, og tester i anden ligne forskellen mellem `KK` og `VZ`. Det er her værd at bemærke den enormt store p værdi på `0.975` understøtter hypotesen om at der ikke er signifikant forskel mellem logeffekten af VZ og KK vaccinen - hvilket også er understøttet af estimatet for forskellen som er ret lavt som 0.0069452.

Opgave 2.3

(Se besvarelses forklaring på papir)

Vi laver tabel for forskellige faktorer;

```
table(vaccine$S, vaccine$U) #"fuld rang": 10 niveauer
```

```
##
##       1  2  3  4  5
##   PL 20 20 20 20 20
##   VA 40 40 40 40 40
```

```
table(vaccine$T, vaccine$U) #"fuld rang": 15 niveauer
```

```
##
##       1  2  3  4  5
##   KK 20 20 20 20 20
##   VZ 20 20 20 20 20
##   PL 20 20 20 20 20
```

```
table(vaccine$T, interaction(vaccine$U,vaccine$S)) #"fuld rang": 15 niveauer
```

```
##
##      1.PL 2.PL 3.PL 4.PL 5.PL 1.VA 2.VA 3.VA 4.VA 5.VA
##   KK    0    0    0    0    0   20   20   20   20   20
##   VZ    0    0    0    0    0   20   20   20   20   20
##   PL   20   20   20   20   20    0    0    0    0    0
```

Opgave 2.4

(Primære besvarelse af første del af opgaven her i R, anden del i hånden)

Vi skal afgøre om $\mathbb{G}$ er et ortogonalt design. Vi har igen som i opgave 2.3 kun to ikke-trivielle sammenligninger at teste; henholdsvis om vi har geometrisk ortogonalitet mellem S og U og mellem T og $S \times U$. Vi tester den første i R ved den følgende selvlavede funktion;

```
SamhBalanceEquation <- function(x) {
   #Cannot be used with tablesum function,
   # !!! requires connected components (sammenhængende design). !!!
   sx <- sum(x)
   testy <- x
   rs <- apply(x,1,sum)
   cs <- apply(x,2,sum)
   nuco <- ncol(x)
   nuro <- nrow(x)
   for (i in 1:nuro) {
      for (j in 1:nuco) {
         if (x[i,j]==rs[i]*cs[j]/sx) {
            testy[i,j] <- TRUE
         } else {
         testy[i,j] <- FALSE
       }
      }
   }
   if (sum(testy) == nuco*nuro) {
      TRUE
   } else {
     FALSE
   }
}
SamhBalanceEquation(table(vaccine$S, vaccine$U)) #TRUE
```

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

Funktionen tester om sammenhængende designs opfylder balanceligningen, hvilket funktionen siger at tabellen gør. Per L13.11 i EH vil S og U dermed være parvist geometrisk ortogonale. Vi har set i opgave 2.3 at tabellen for T og $S \times U$ har balancerede komponenter, således at vi ved Sætning 14.8, Eksempel 14.9 i EH kan konkludere at T og $S \times U$ også er parvist geometrisk ortogonale.

Tilsammen kan vi dermed konkluderer at vi har med et geometrisk ortogonalt design at gøre.

Opgave 2.5

(Primære besvarelse af opgaven her i R)

Vi har i opgave 2.2 set at når vi udelukkende betragter antistofresponset i slukningen af perioden, er der ikke signifigant forskel mellem de to vacciner. Vi kan undersøge om der i stedet er forskel mellem vaccinerne, eksempelvis i forhold til, hvor hurtigt de kommer til deres antistofrespons.

Vi subsetter derfor vores model til kun at være de to ikke-placebo vacciner;

```
vaccinevacciner <- subset(vaccine, vaccine$T %in% c("VZ","KK"))
```

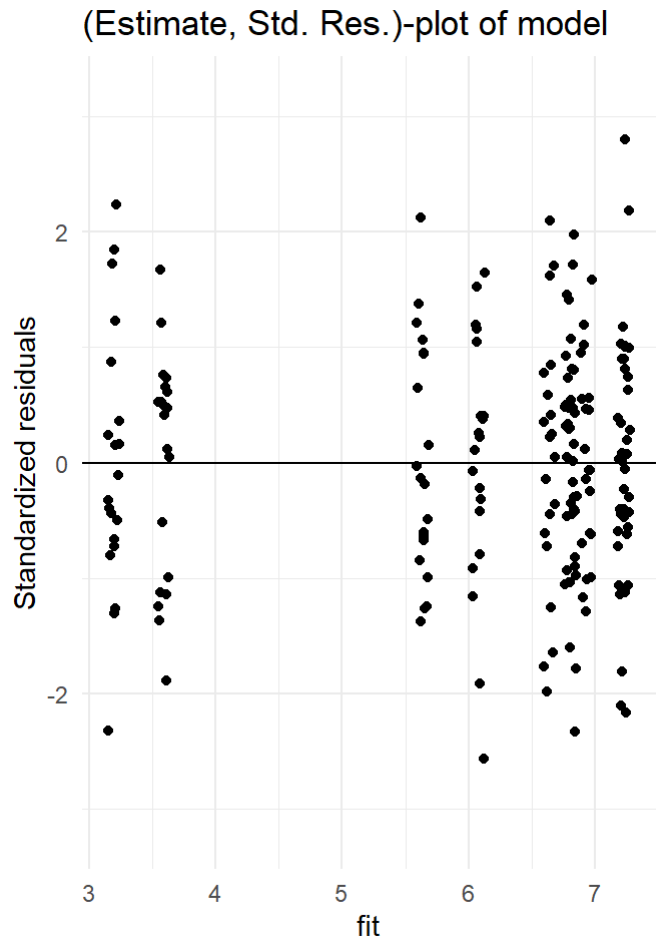
Vi opretter modellen i R

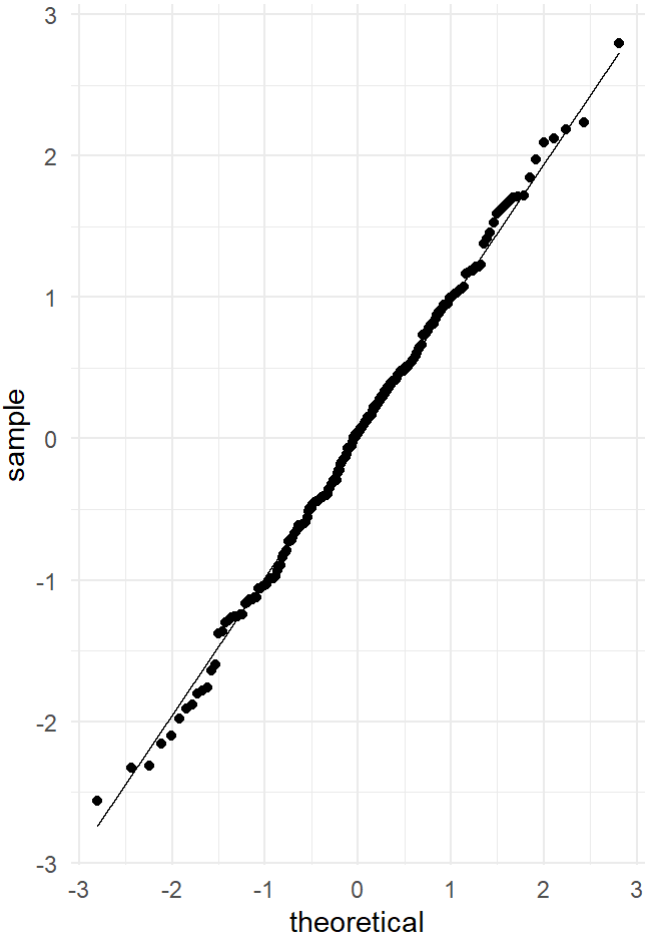
```
interacmodel <- lm(log(X)~Id+U+U:T, data = vaccinevacciner)
summary(interacmodel)
```

```
##
## Call:
## lm(formula = log(X) ~ Id + U + U:T, data = vaccinevacciner)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71058 -0.44529  0.03116  0.43622  1.87662
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.242002   0.175954  18.425  < 2e-16 ***
## Id          -0.001548   0.002782  -0.557  0.57848
## U2           2.443139   0.217992  11.207  < 2e-16 ***
## U3           2.885532   0.217992  13.237  < 2e-16 ***
## U4           3.443905   0.217992  15.798  < 2e-16 ***
## U5           4.032577   0.217992  18.499  < 2e-16 ***
## U1:TVZ       0.394548   0.218043   1.809  0.07196 .
## U2:TVZ       1.162915   0.218043   5.333 2.73e-07 ***
## U3:TVZ       0.852026   0.218043   3.908  0.00013 ***
## U4:TVZ       0.175261   0.218043   0.804  0.42253
## U5:TVZ       0.004313   0.218043   0.020  0.98424
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6893 on 189 degrees of freedom
## Multiple R-squared:  0.8123, Adjusted R-squared:  0.8024
## F-statistic: 81.79 on 10 and 189 DF,  p-value: < 2.2e-16
```

Og kan ved at tegne std. res. og qq plot se, at modellen på log-niveau passer antagelser ganske pænt;

```
StdresQQPlot(interacmodel)
```





Vi laver dermed modellen om at der kan være en selvstående individ effekt, en selvstående ugeeffekt, og at der sammentidigt er en interaktionseffekt mellem uge og hvilken vaccine man er blevet givet.

```
anova(interacmodel)
```

```
## Analysis of Variance Table
##
## Response: log(X)
##            Df Sum Sq Mean Sq  F value  Pr(>F)
## Id          1   0.32   0.315   0.6636  0.4163
## U           4 365.76  91.440 192.4236 < 2e-16 ***
## U:T         5  22.62   4.523   9.5184 4.1e-08 ***
## Residuals 189  89.81   0.475
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I `U:T` linjen i anova testet, tester vi modellen om at der ikke er en interaktionseffekt mellem behandling og uge, mod modellen hvor der ér en interaktionseffekt.

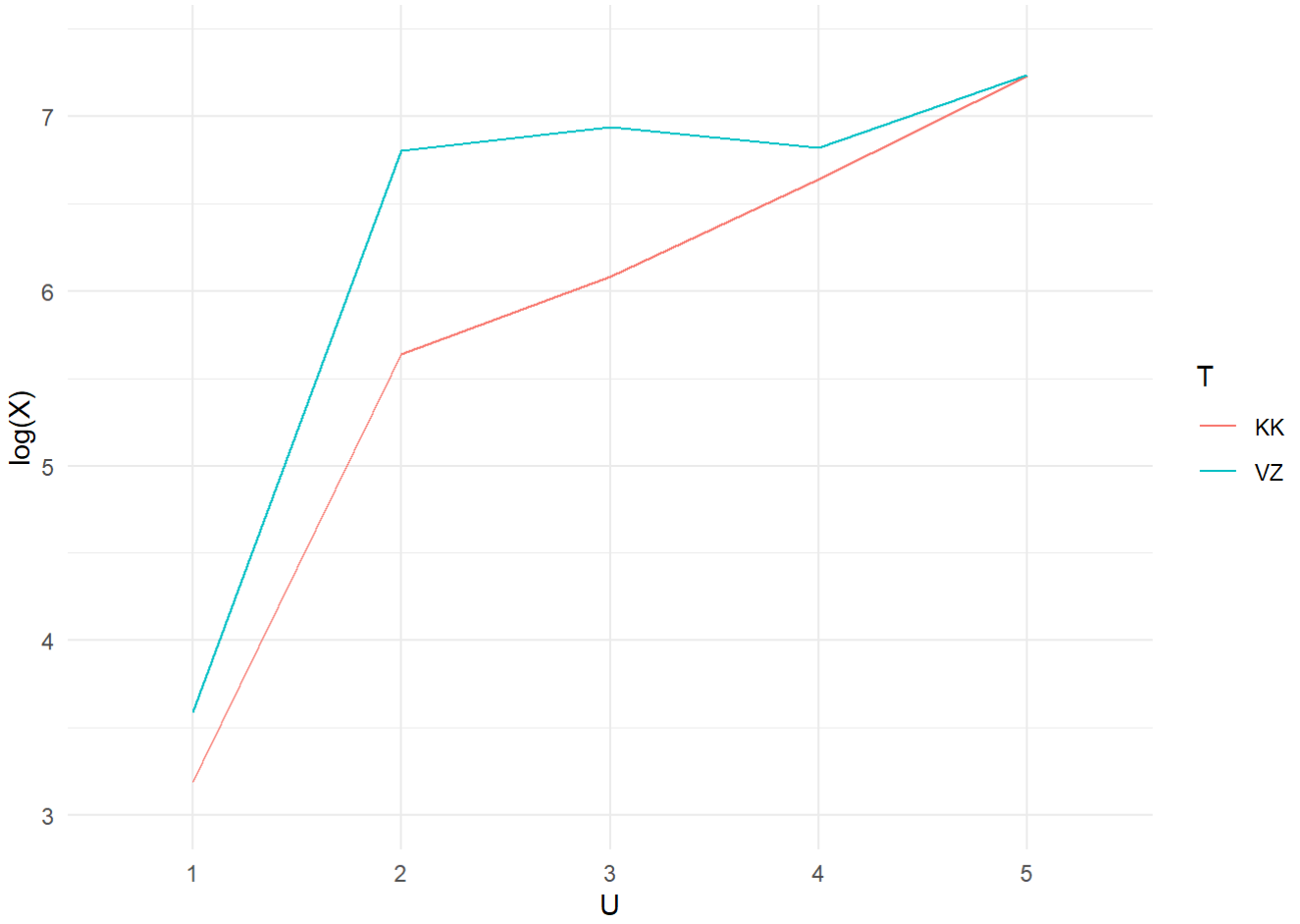
Vi ser en meget lav p værdi således at der i modellen er evidens for at forkaste nul-hypotesen om at der ikke er interaktionseffekt. Vi kan dermed sige at der forekommer evidens for at der er en klar forskel i hvordan de to vacciner virker henover tid. Sammenligner vi med opgave 2.2 vil forskellen i effekten dog efter fem uger være tilstrækkeligt lille til, at vi ikke statistisk kan bemærke forskel.

Vi kan visualiserer denne effekt over tid;

```
ggplot(vaccinevacciner, aes(x = U, y = log(X), color = T)) + stat_summary(aes(group = T), geo
m = "line", fun = "mean")
```

```
## Warning: Ignoring unknown parameters: fun
```

```
## No summary function supplied, defaulting to `mean_se()
```



Opgave 3

Opgave 3.1

Ikke løst. #### **Opgave 3.2** Ikke løst.