Hjemmeopgave - Kursus i basal statistik

Nils Hoyer (hold 9)

02-11-2016

Jeg bruger R til at løse opgaven. Jeg har indlæst de nødvendige libraries (ggplot2, dplyr, psych, knitr).

Forberedelser

Jeg starter med at indlæse data i R og gemme dem i "dataset"

```
dataset <- read.csv("http://publicifsv.sund.ku.dk/~lts/basal16_2/hjemmeopgave/hjemmeopgave.txt",
header=TRUE, sep=" ", dec=".")
```

Herefter kontrollerer jeg at R har kodet variablerne korrekt.

```
str(dataset)
```

```
## 'data.frame':    80 obs. of  4 variables:
##  $ idnr      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ bambuterol: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ke        : int  731 2334 1738 1645 1430 1230 1130 1082 995 977 ...
##  $ fer       : num  6.5 2 5 2.5 4.5 5 5 5 7 6.5 ...
```

Da både idnr og bambuterol er kategoriske variable omkoder jeg dem til faktorer i datasettet.

```
dataset$idnr <- as.factor(dataset$idnr)
dataset$bambuterol <- as.factor(dataset$bambuterol)
```

Jeg fjerner observation 1 og 51.

```
dataset <- filter(dataset, idnr!= 1)
dataset <- filter(dataset, idnr!= 51)
```

Jeg transformerer allerede nu variablerne "ke" og "fer" til senere brug.

```
dataset$logke <- log10(dataset$ke)
dataset$logfer <- log10(dataset$fer)
```

Jeg deler datasettet op i et med bambuterol og et uden bambuterol.

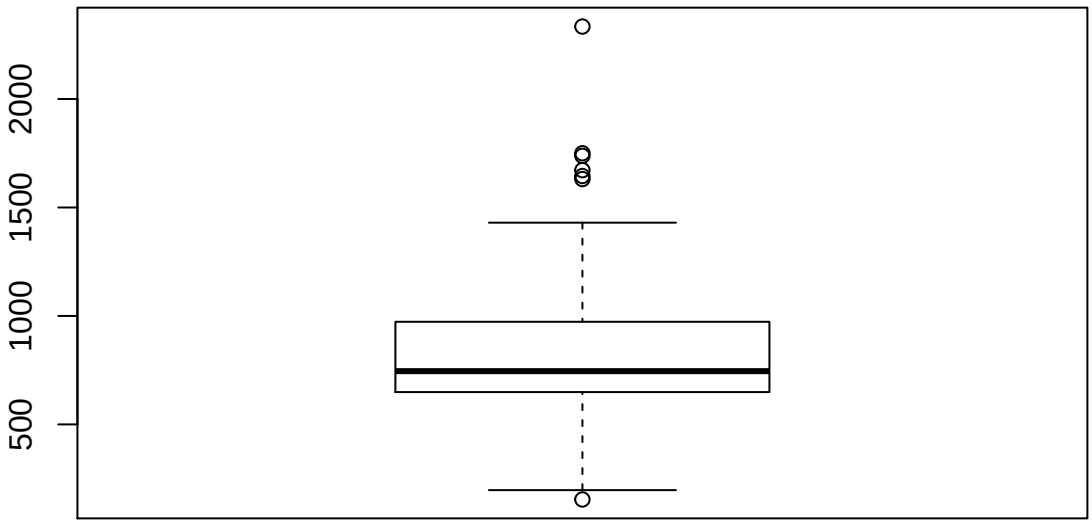
```
udenbamb <- filter(dataset, bambuterol == 0)
kunbamb <- filter(dataset, bambuterol == 1)
```

Spørgsmål 1 - Beskriv fordelingen af aktiviteten af kolinesterase i denne gruppe.

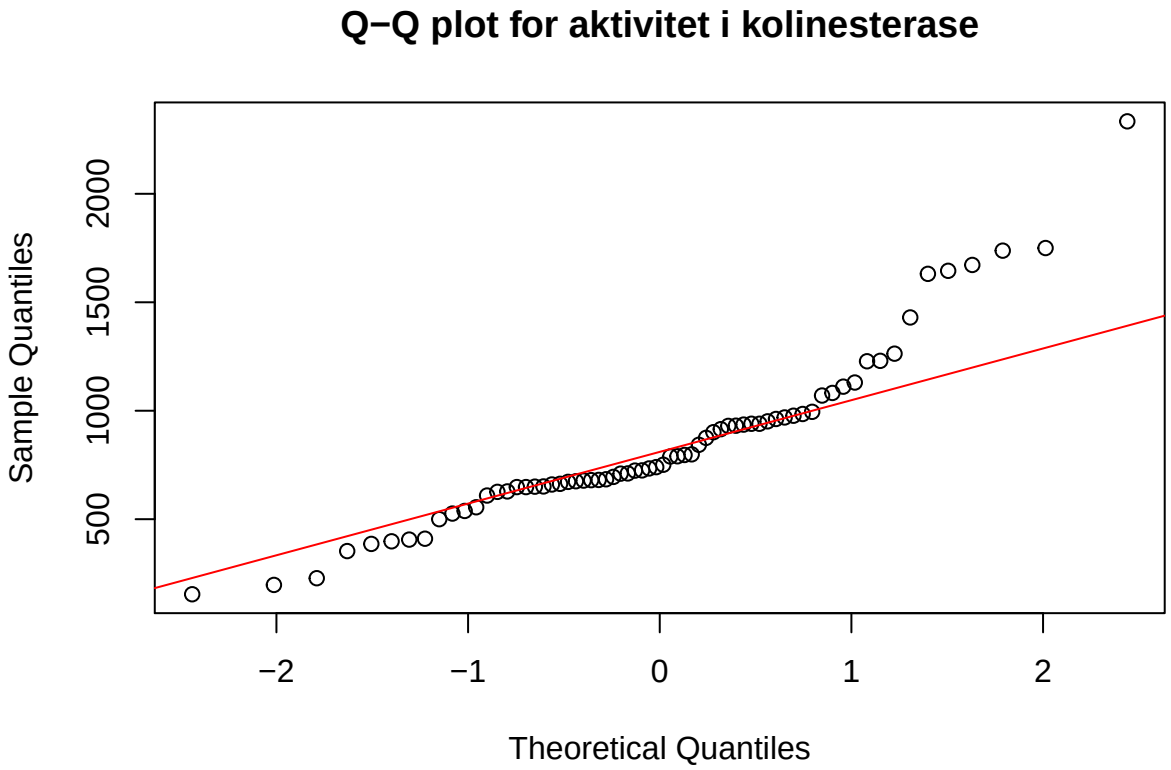
(a) Lav først en grafisk illustration.

Jeg bruger datasettet for deltagere uden bambuterol. Jeg undersøger om variabeln "ke" er normalfordelt. Boxplottet har nogle outliers opadtil, derfor undersøger jeg også med et Q-Q plot (fraktildiagram).

```
boxplot(udenbamb$ke)
```

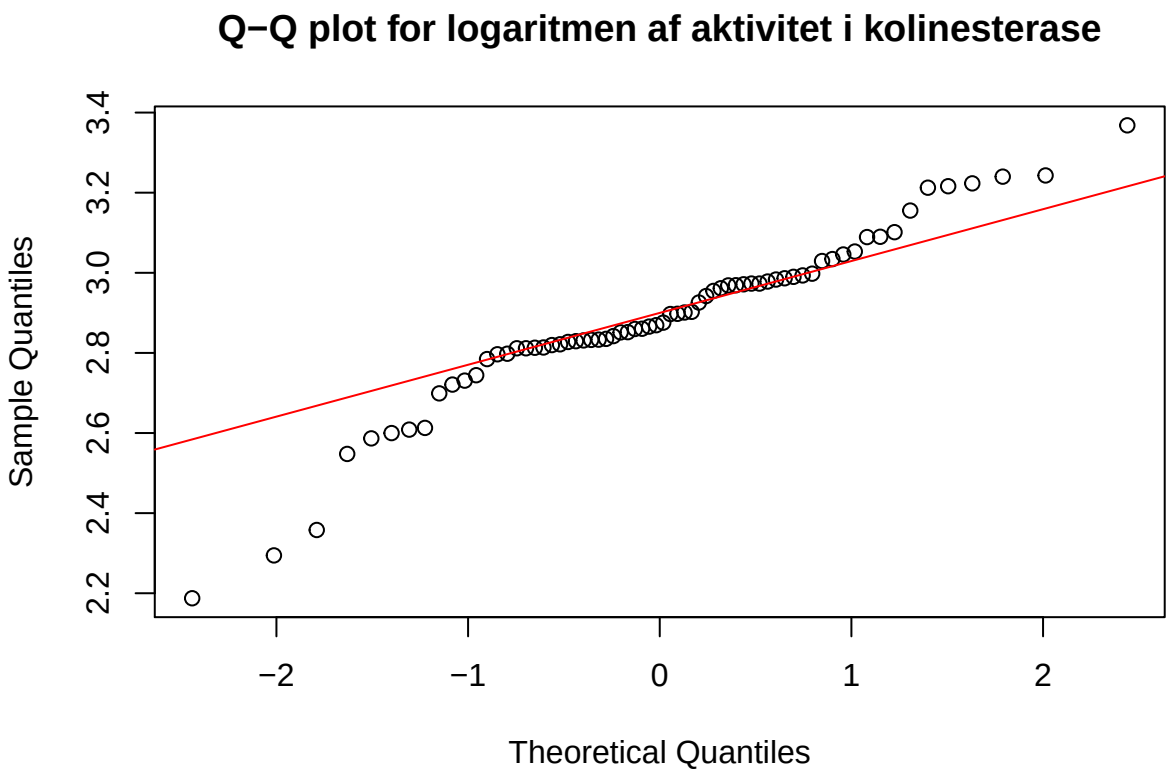



```
qqnorm(udenbamb$ke, main = "Q-Q plot for aktivitet i kolinesterase",
    xlab = "Theoretical Quantiles",
    ylab = "Sample Quantiles")
qqline(udenbamb$ke, col=2)
```



Da jeg ser at ke ikke er normalfordelt plotter jeg log-transformationen af "ke", hvilket kun ser marginalt bedre ud.

```
qqnorm(udenbamb$logke, main = "Q-Q plot for logaritmen af aktivitet i kolinesterase",
    xlab = "Theoretical Quantiles",
    ylab = "Sample Quantiles")
qqline(udenbamb$logke, col=2)
```



(b) Udregn dernæst passende valgte summary statistics, som om du skulle lave en "Tabel 1" til en artikel, og forklar kort hvorfor du vælger netop disse.

```
summary(udenbamb$ke)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   154.0   649.5   745.5   850.5   971.0  2334.0
```

Denne variabel er ikke er normalfordelt, hvilket kan ses på fraktildiagrammet ovenfor, og også ved at median og gennemsnit er langt fra hinanden. Derfor er det mest korret at rapportere den som median med kvartilgrænserne:

Tabel 1	Median	IQR
Kolinesterase	745.5	649.5-971.0

Evt. kan man også angive hele range, dvs. den mindste og den største værdi.

(c) Kan man sige, at det er usædvanligt lavt med en kolinesterase aktivitet på 200? Hvor mange har en værdi under dette?

```
sum(udenbamb$ke < 200)
```

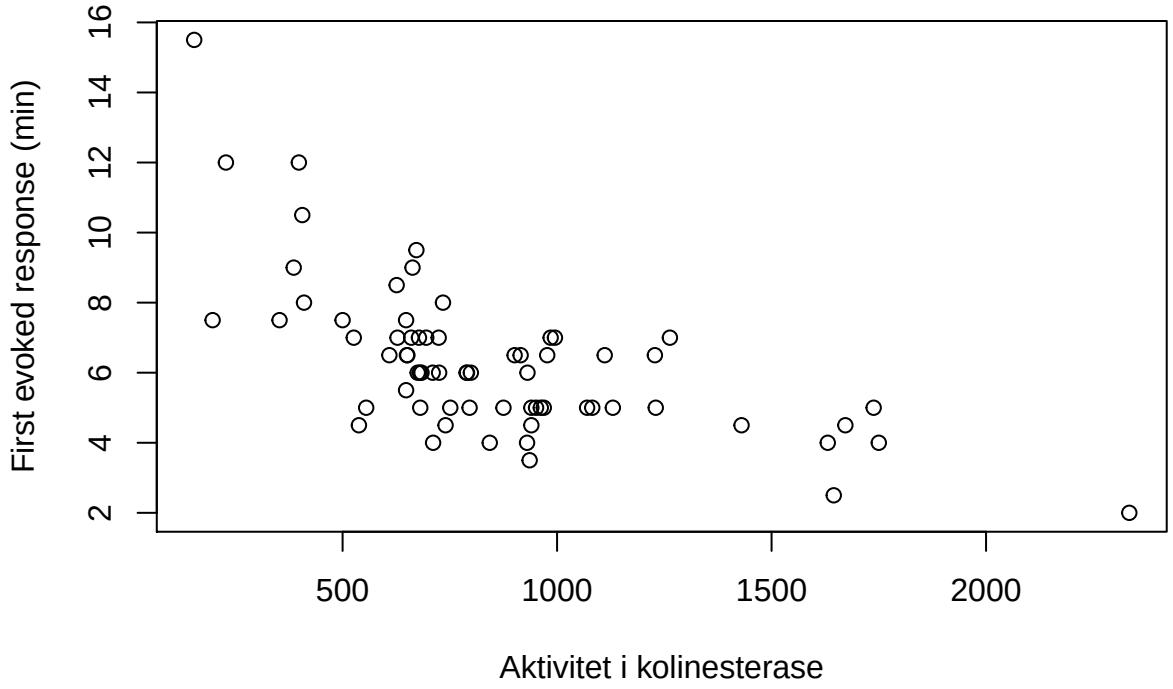
```
## [1] 2
```

Der er 2 observationer af kolinesteraseaktivitet under 200. Da data ikke er normalfordelt, kan jeg ikke beregne referenceværdier.

Spørgsmål 2. Vi skal nu se på relationen mellem kolinesterase og first evoked response for den normale gruppe:

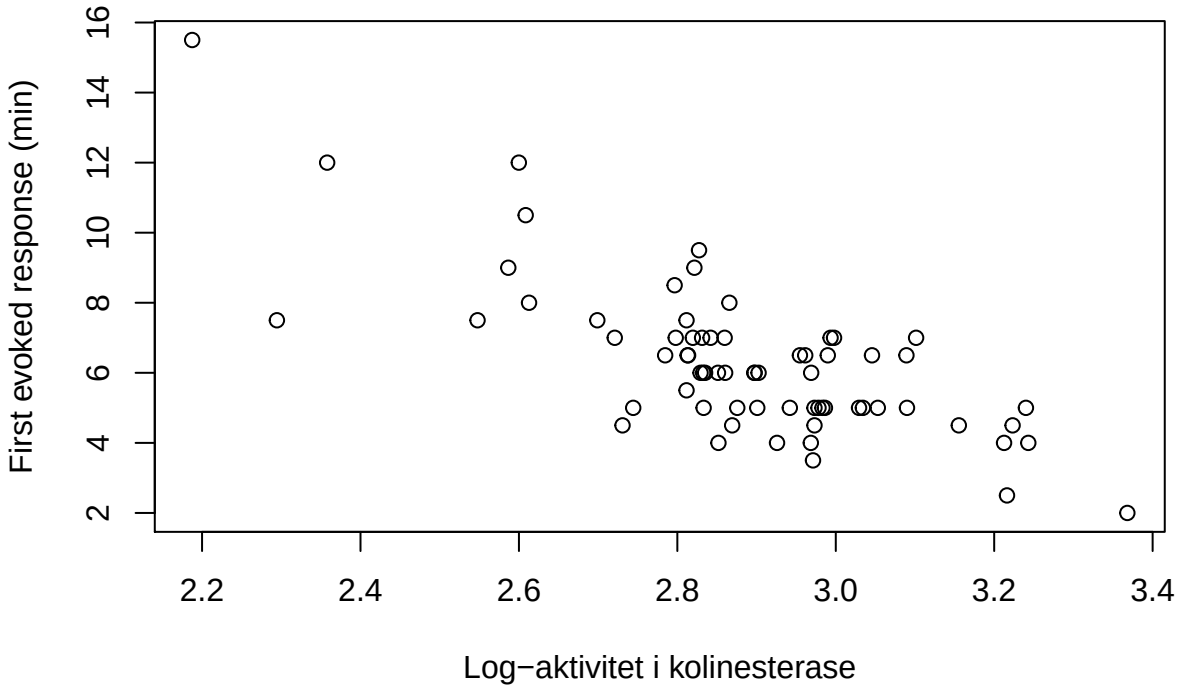
(a) Lav et scatterplot af first evoked response mod kolinesterase for denne gruppe. Ser det rimeligt lineært ud?

```
plot(udenbamb$ke, udenbamb$fer,
       xlab="Aktivitet i kolinesterase",
       ylab="First evoked response (min)")
```



Jeg synes at den buer lidt for meget, derfor plotter jeg "fer" mod "log(ke)".

```
plot(udenbamb$logke, udenbamb$fer, xlab="Log-aktivitet i kolinesterase", ylab="First evoked response (m
```

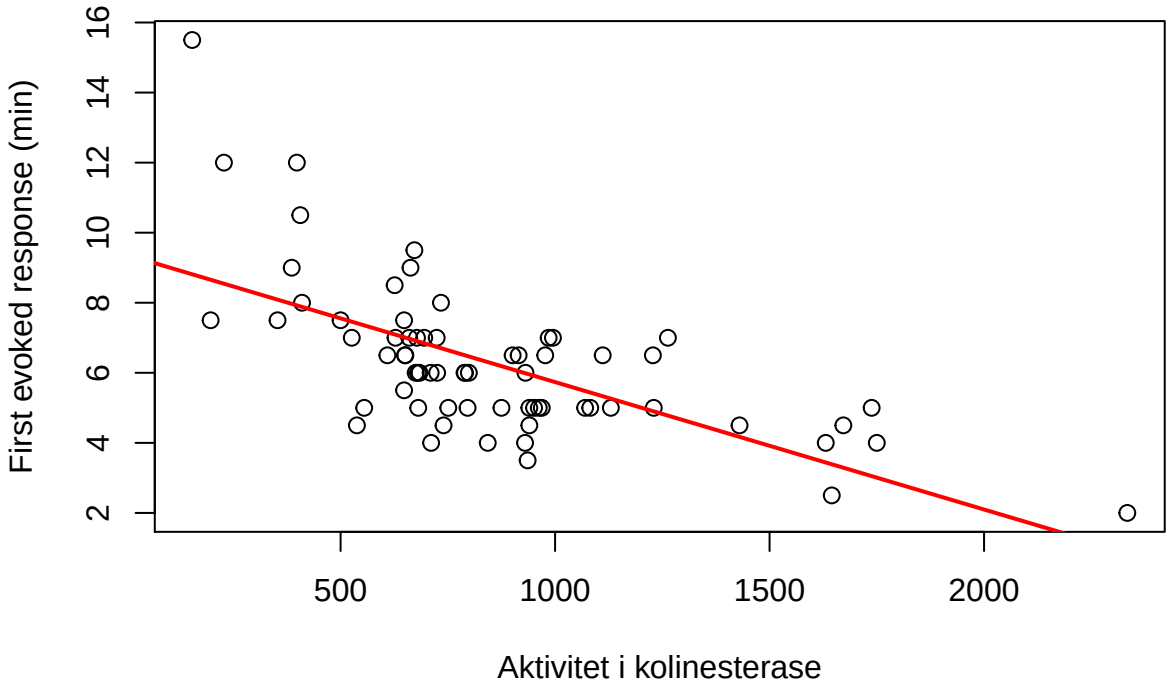


Dette plot viser en mere lineær trend. Dog ser det ud som om spredningen er større for lave værdier på aktivitet i kolinesterase end for høje. Der er også meget færre observationer for små værdier på kolinesterase-aktivitet. Så det er ikke perfekt.

Uanset svaret på spm. 2a ønskes nedenstående spørgsmål besvaret for de utransformerede data:

(b) Lav en lineær regression, og udfør passende modelkontrol og diagnostics. Sørg også for at få en figur af fittet med.

```
plot(udenbamb$ke, udenbamb$fer,
        xlab="Aktivitet i kolinesterase",
        ylab="First evoked response (min)",
        col="black", cex=1.1, pch=21)
abline(lm(fer~ke, data=udenbamb), lwd=2, col=2)
```




```
fit <- lm(fer~ke, data=udenbamb)
summary(fit)
```

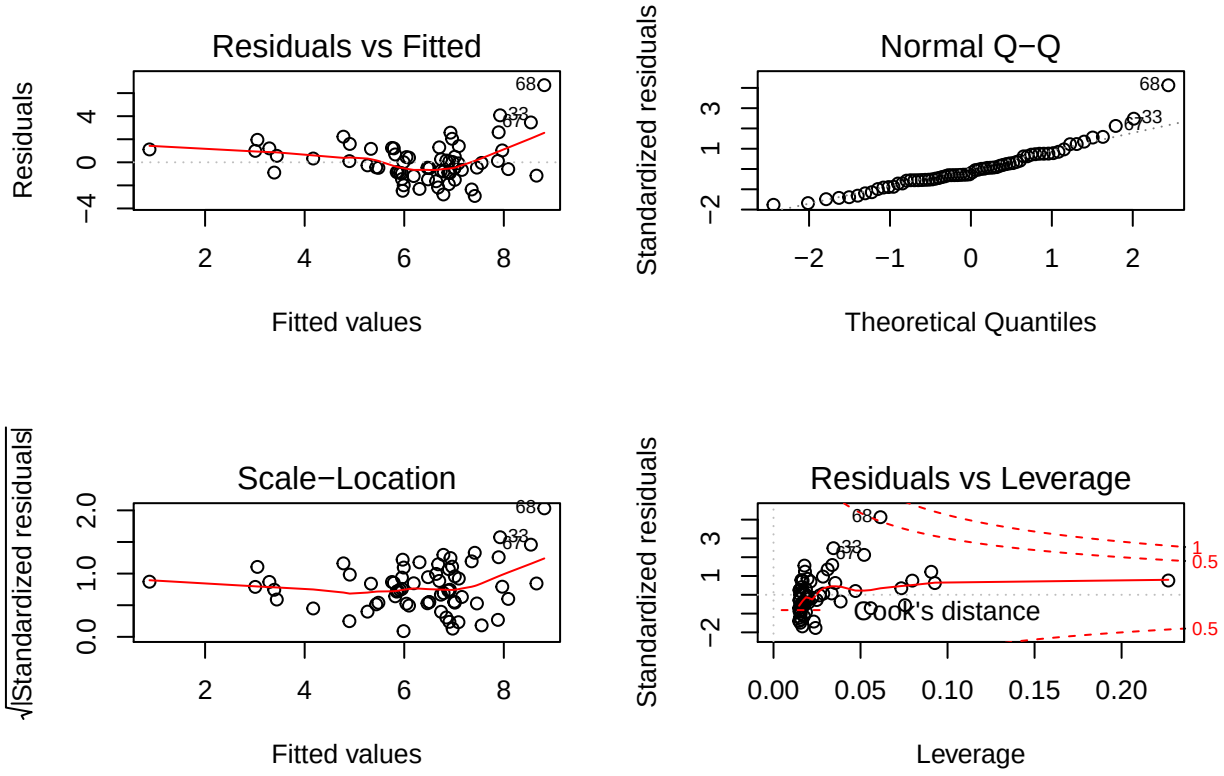
```
##
## Call:
## lm(formula = fer ~ ke, data = udenbamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9160 -0.9031 -0.3499  1.0023  6.6871
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.373062   0.485710  19.298  < 2e-16 ***
## ke          -0.003638   0.000519  -7.009 1.56e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.672 on 66 degrees of freedom
## Multiple R-squared:  0.4267, Adjusted R-squared:  0.418
## F-statistic: 49.13 on 1 and 66 DF,  p-value: 1.562e-09
```

```
confint(fit, level = 0.95)
```

```
##                    2.5 %       97.5 %
## (Intercept)  8.403310593 10.342814348
## ke          -0.004673774 -0.002601376
```

Dette tolkes som at First evoked response falder med 0.0036 min med 95% konfidensintervallet (0.0026 til 0.0047 min) for hver enhed stigning i aktivitet i kolinesterase.

```
par(mfrow=c(2,2))
plot(fit)
```



Nogen indvendinger mod modelantagelserne? I så fald hvilke?

Jeg synes at der er flere problemer med modellen:

- Der er få observationer for lave estimerede værdier. Derfor er modellen usikker i denne ende
- Spredningen for residualerne stiger med høje estimerede værdier - der er ikke varianshomogenitet
- Fraktildiagrammet viser at residualerne ikke er normalfordelede, men har hængekøjeform
- Punkt 68 har helt klart den største Cooks værdi, dvs. den har stor indflydelse på linien. Denne kan skævvride hele kurven. Såfremt det er en fejlmåling, bidrager den altså til en forkert model. Hvis det er en reel måling, er det ikke nødvendigvis et problem.

(c) Giv en forståelig fortolkning af hældningsestimatet.

First evoked response falder med 0.0036 min (eller 0.2 sek) for hver enhed stigning i aktivitet i kolinesterase. Dette estimat kan selvfølgelig være lidt forkert, men vi kan med 95% sikkerhed sige at det ligger i intervallet 0,0026 til 0,0047 min.

(d) Hvad er middelværdien af first evoked response for personer med kolinesterase aktivitet på 500? Og hvad er konfidensintervallet for denne?

```r
predict(fit, newdata = data.frame(ke = 500),level=.95,interval="confidence")
```

```
##        fit      lwr      upr
## 1 7.554275 7.010503 8.098047
```

Den estimerede værdi for en kolinesterase aktivitet på 500 er således 7.55 min (95% konfidensinterval 7.01 - 8.10 min).

(e) Angiv et 95% prediktionsinterval for patienter med en kolinesterase aktivitet på 500. Er det usædvanligt at se sådan en person have en first evoked response på kun 5 minutter?

```r
predict(fit, newdata = data.frame(ke = 500),level=.95,interval="prediction")
```

```
##        fit      lwr      upr
## 1 7.554275 4.172794 10.93576
```

Prediktionsgrænserne for en kolinesterase aktivitet på 500 går altså fra 4.17 - 10.94. 5 minutter er dermed indenfor prediktionsgrænserne hvor 95% af alle individer vil befinde sig. Vi kan sige at det *ikke* er usædvanligt at se denne værdi.

(f) Hvis vi nu vovede en ekstrapolation af relationen fra spørgsmål 2b, hvad ville vi så gætte på, at middelværdien af first evoked response ville være, når kolinesterase var helt nede på 50?

Og hvad med usikkerheden på sådan en ekstrapolation?

```r
predict(fit, newdata = data.frame(ke = 50))
```

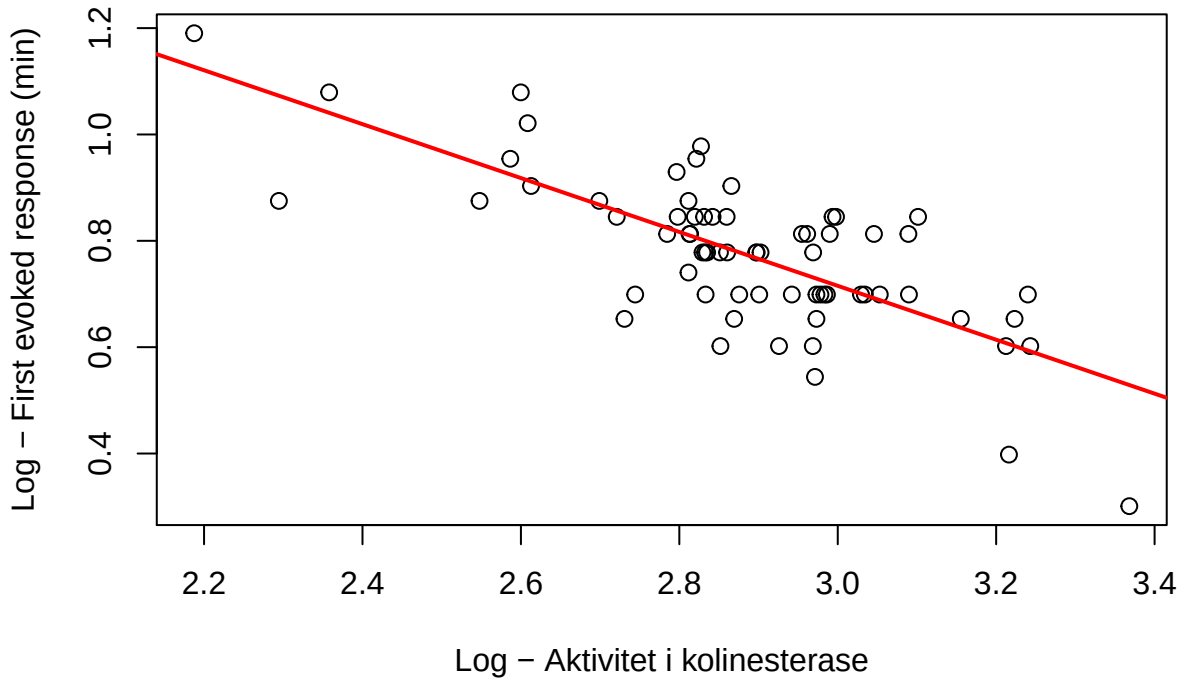
```
##        1
## 9.191184
```

Ifølge modellen ville en kolinesterase aktivitet på 50 give en First evoked response på 9.19 min. Vi har i vores dataset kun 2 observationer med en kolinesterase aktivitet under 200, og denne slags ekstrapolation er naturligvis fuldstændig uforsvarlig. Vi er ikke sikre på at modellen stadig er lineær for punkter udenfor vores observationer.

Spørgsmål 3. Foretag nu en logaritmetransformation af såvel kolinesterase som first evoked response, og svar på de samme 6 delspørgsmål som ovenfor i spm. 2, samt

(a) Lav et scatterplot af first evoked response mod kolinesterase for denne gruppe. Ser det rimeligt lineært ud?

```r
plot(udenbamb$logke, udenbamb$logfer,
        xlab="Log - Aktivitet i kolinesterase",
        ylab="Log - First evoked response (min)",
        col="black", cex=1.1, pch=21)
abline(lm(logfer~logke, data=udenbamb), lwd=2, col=2)
```



Det ser fornuftigt lineært ud. Jeg fortsætter med de transformerede data.

(b) Lav en lineær regression, og udfør passende modelkontrol og diagnostics. Sørg også for at få en figur af fittet med.

```r
logfit <- lm(logfer~logke, data=udenbamb)
summary(logfit)
```

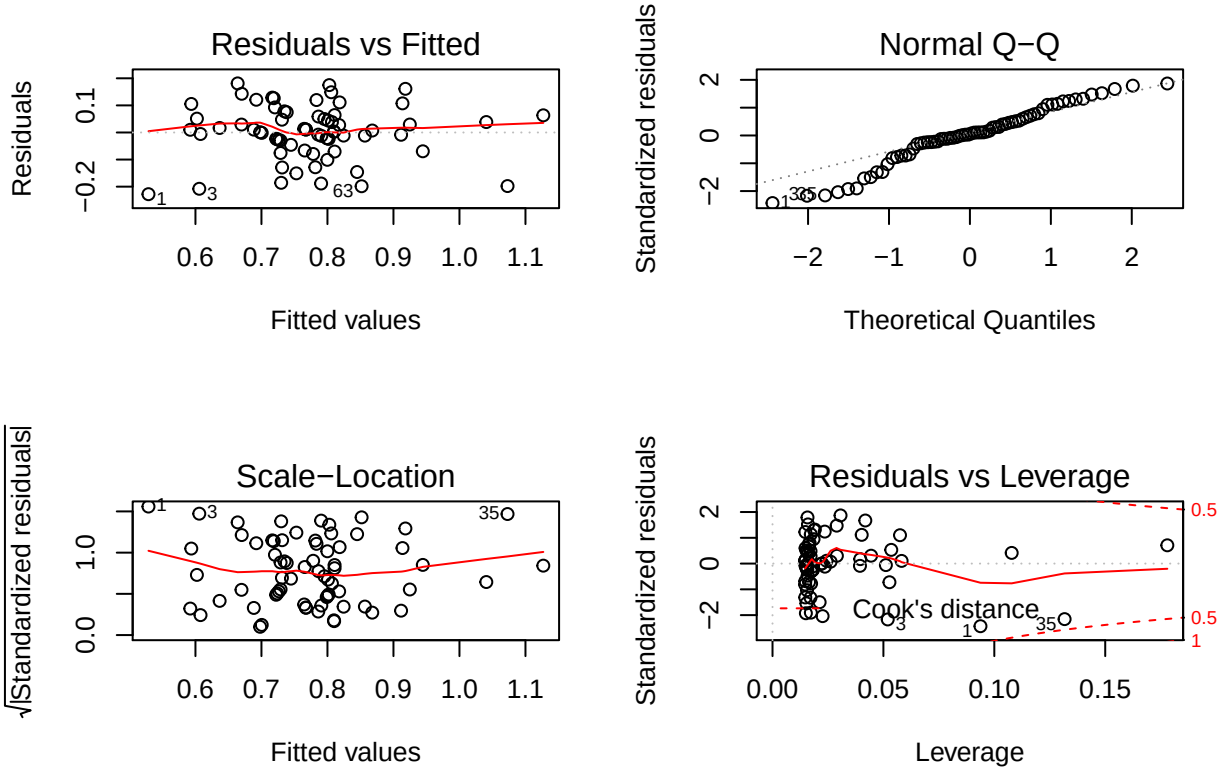
```
##
## Call:
## lm(formula = logfer ~ logke, data = udenbamb)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.227780 -0.033922  0.005154  0.059843  0.181148
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.23549    0.16525  13.528  < 2e-16 ***
## logke       -0.50672    0.05715  -8.866 7.53e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.09847 on 66 degrees of freedom
## Multiple R-squared:  0.5436, Adjusted R-squared:  0.5367
## F-statistic:  78.6 on 1 and 66 DF,  p-value: 7.53e-13
```

```
confint(logfit, level = 0.95)
```

```
##                  2.5 %     97.5 %
## (Intercept)  1.9055639  2.5654199
## logke       -0.6208315 -0.3926074
```

```
par(mfrow=c(2,2))
plot(logfit)
```



Der er pænere varianshomogenicitet end for de utransformerede data. Der er signifikant regression, som er negativ. Der er dog stadig tyndt med observationer sv.t. de høje værdier for first evoked response (de lave kolinesterase aktiviteter). Residualerne ser stadig ikke helt normalfordelte ud (men pænere end for modellen baseret på utransformerede data).

(c) Giv en forståelig fortolkning af hældningsestimatet.

Da vi har en log-log model (baseret på 10-logaritmen), gælder følgende: en stigning på en faktor 10 på x svarer til en faktor 10 opløftet til beta på y.

```
10^(-0.50672)
```

```
## [1] 0.3113723
```

Derfor gælder: For hver ti-dobling af x, ændres y med faktoren 0.31, dvs. det falder med 69%. Her ville det evt. være relevant at bruge en anden logaritme, fx. en log2, som vil kunne angive hvordan y ændres for hver fordobling af x.

(d) Hvad er middelværdien af first evoked response for personer med kolinesterase aktivitet på 500? Og hvad er konfidensintervallet for denne?

```
a <- log10(500)
logfit500 <- predict(logfit, newdata = data.frame(logke = a),level=.95,interval="confidence")
logfit500
```

```
##         fit       lwr       upr
## 1 0.8678712 0.8360475 0.8996949
```

```
10^c(logfit500)
```

```
## [1] 7.376855 6.855633 7.937704
```

Middelværdien for first evoked response er således 7.38 min for en kolinesterase aktivitet på 500 (95% konfidensinterval: 6.86 - 7.94 min)

(e) Angiv et 95% prediktionsinterval for patienter med en kolinesterase aktivitet på 500. Er det usædvanligt at se sådan en person have en first evoked response på kun 5 minutter?

```
logfit500 <- predict(logfit, newdata = data.frame(logke = a),level=.95,interval="prediction")
logfit500
```

```
##         fit       lwr      upr
## 1 0.8678712 0.6687202 1.067022
```

```
10^c(logfit500)
```

```
## [1]  7.376855  4.663589 11.668693
```

95% prædiktionsinterval for en kolinesterase aktivitet på 500 går fra 4.68 - 11.67 min. Bland de 95% "mest almindelige" med en aktivitet på 500 vil man altså også forvente at finde en med en first evoked response på 5 minutter.

(f) Hvis vi nu vovede en ekstrapolation af relationen fra spørgsmål 2b, hvad ville vi så gætte på, at middelværdien af first evoked response ville være, når kolinesterase var helt nede på 50?

Og hvad med usikkerheden på sådan en ekstrapolation?

```r
b <- log10(50)
logfit50 <- predict(logfit, newdata = data.frame(logke = b),level=.95,interval="confidence")
logfit50
```

```
##        fit      lwr      upr
## 1 1.374591 1.237312 1.511869
```

```r
10^c(logfit50)
```

```
## [1] 23.69140 17.27079 32.49894
```

Ifølge denne model vil vi forvente en gennemsnitlig værdi på 23.69 minutter first evoked response for en kolinesterase aktivtitet på 50. Det er naturligvis stadig forkert at ekstrapolere på trods af at data er logaritmetransformeret.

(g) Hvilke forskelle ses på fittene for den utransformerede og den logaritmetransformerede relation?

Hvilken en vil du foretrække, og hvorfor?

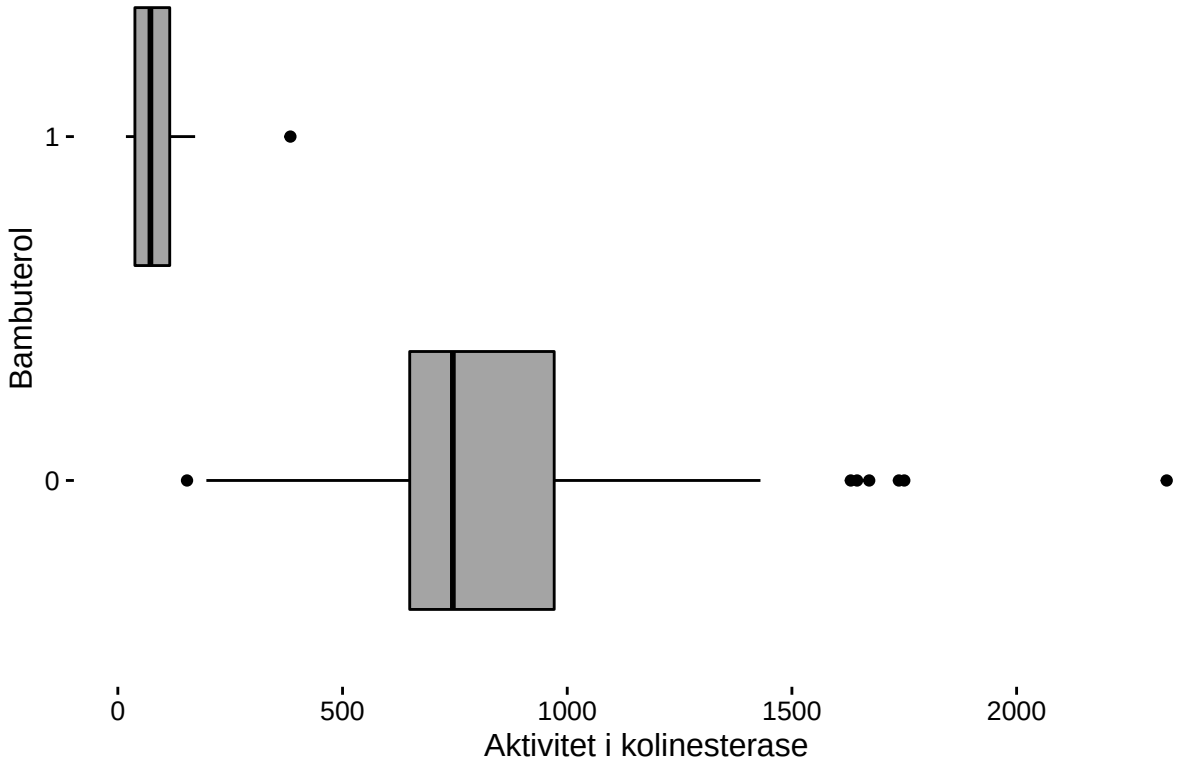
Jeg synes at den logaritmetransformerede model opfylder langt flere krav om linearitet og varianshomogenicitet. Derfor vil jeg vælge denne.

Spørgsmål 4. Vi skal nu sammenligne kolinesterase for de to grupper:

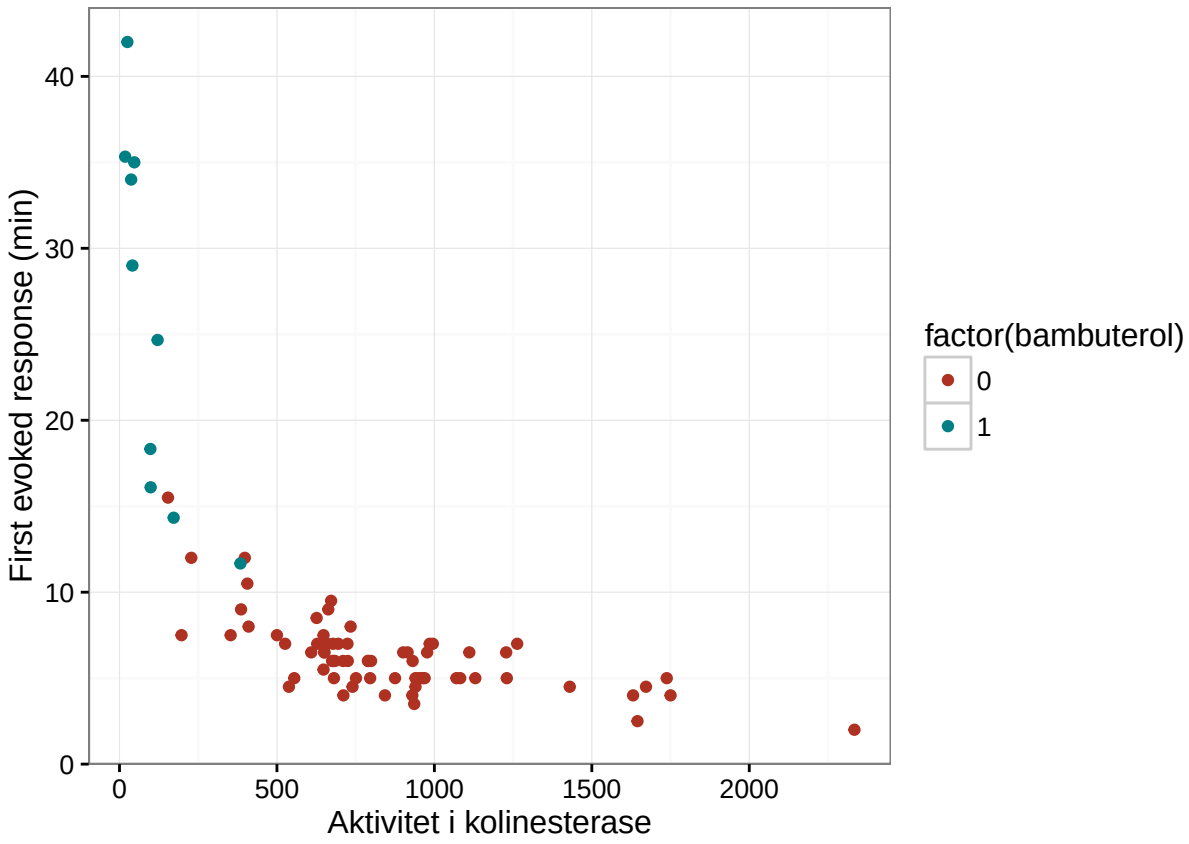
(a) Lav en illustration til sammenligning af kolinesterase for den normale gruppe og bambuterol-gruppen, på passende skala.

Jeg starter med et boxplot og et scatterplot over de to grupper.

```r
ggplot(dataset, aes(x=bambuterol, y=ke)) +
    geom_boxplot(fill='#A4A4A4', color="black") +
    xlab("Bambuterol") +
    ylab("Aktivitet i kolinesterase") +
    theme_classic() +
    coord_flip()
```




```
ggplot(dataset, aes(x=ke, y=fer, color=factor(bambuterol))) +
    geom_point(shape=19) +
    xlab("Aktivitet i kolinesterase") +
    ylab("First evoked response (min)") +
    theme_bw() +
    scale_colour_hue(l=40)
```



Der ses meget stor forskel i både gennemsnit og spredning samt i fordelingen på x-aksen. Der er nok også større forskel i spredningen end jeg ville forvente bare pga. forskellige gruppestørrelser.

(b) Er der overhovedet noget overlap mellem de to fordelinger?

Ikke ret meget overlap.

```
summary(udenbamb$ke)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   154.0   649.5   745.5   850.5   971.0  2334.0
```

```
summary(kunbamb$ke)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    18.0    38.0    72.5   104.2   115.5   384.0
```

Derved kan vi se at grupperne overlapper i halerne, men det nedre kvartil for gruppen uden bambuterol er langt fra den øvre kvartil for gruppen med bambuterol.

(c) Giv et estimat for den procentuelle reduktion af kolinesterase hos bambuterol patienter, sammenlignet med de normale. Husk at angive konfidensinterval også.

Jeg beregner den gennemsnitlige reduktion af kolinesterase aktivitet med en t-test. Da data ikke er normalfordelt (som set i spørgsmål 1), bruger jeg logaritmen for aktiviteten i kolinesterase.

```r
t.test(logke ~ bambuterol, data=dataset)
```

```
##
##  Welch Two Sample t-test
##
## data:  logke by bambuterol
## t = 7.9151, df = 9.7117, p-value = 1.54e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7490121 1.3392383
## sample estimates:
## mean in group 0 mean in group 1
##        2.883705        1.839580
```

Differensen mellem gennemsnittene er altså 1.044 (95% konfidensinterval: 0.749 - 1.339) på en logaritmeskala. Dette er signifikant forskelligt fra 0, som kan ses på den meget lave p-værdi. Når jeg tilbagetransformerer får jeg den procentuelle reduktion. Jeg oplyfter til negative værdier for at få faktoren for bambuterol/normale i stedet for normale/bambuterol.

```r
10^(-1.044)
```

```
## [1] 0.09036495
```

```r
10^(-0.749)
```

```
## [1] 0.1782379
```

```r
10^(-1.339)
```

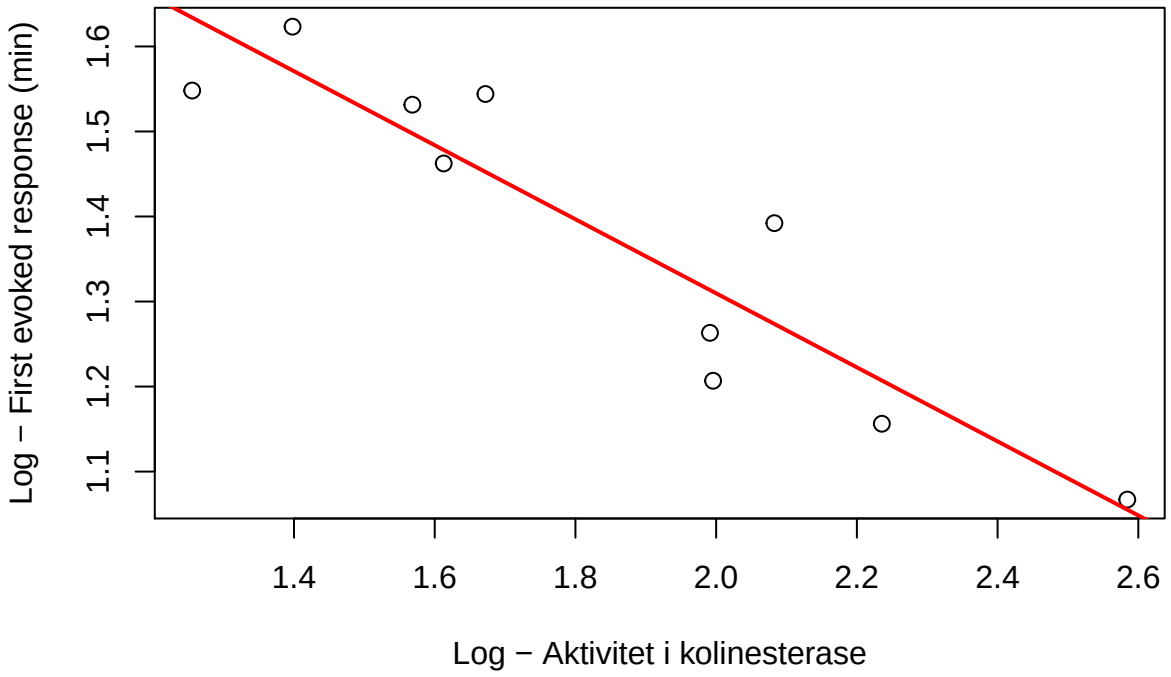
```
## [1] 0.04581419
```

Aktiviteten i kolinesterase i bambuterol-gruppen er altså 91% lavere (95% konfidensinterval 84%-95%) sammenlignet med personerne uden bambuterol.

Spørgsmål 5. Selv om der kun er 10 personer i bambuterolgruppen, ser vi et øjeblik på denne alene:

(a) Udfør en lineær regression på log-log skala som i spm. 3

```r
plot(kunbamb$logke, kunbamb$logfer,
        xlab="Log - Aktivitet i kolinesterase",
        ylab="Log - First evoked response (min)",
        col="black", cex=1.1, pch=21)
abline(lm(logfer~logke, data=kunbamb), lwd=2, col=2)
```




```
logfit_bamb <- lm(logfer~logke, data=kunbamb)
summary(logfit_bamb)
```

```
##
## Call:
## lm(formula = logfer ~ logke, data = kunbamb)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.104708 -0.050701 -0.001987  0.047066  0.118581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.18044    0.12115  17.997 9.32e-08 ***
## logke       -0.43540    0.06444  -6.757 0.000144 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07912 on 8 degrees of freedom
## Multiple R-squared:  0.8509, Adjusted R-squared:  0.8323
## F-statistic: 45.65 on 1 and 8 DF,  p-value: 0.0001441
```

```
confint(logfit_bamb, level = 0.95)
```

```
##                  2.5 %     97.5 %
## (Intercept)  1.9010620  2.4598253
## logke       -0.5840037 -0.2868063
```

(b) Predikter værdien af first evoked response for en kolinesteraseværdi på 50, som i spørgsmål 2f.

Hvordan svarer det til prediktionen fra de normale?

```
b <- log10(50)
logfit50_bamb <- predict(logfit_bamb, newdata = data.frame(logke = b),level=.95,interval="prediction")
logfit50_bamb
```

```
##        fit     lwr      upr
## 1 1.440704 1.24821 1.633197
```

```
10^c(logfit50_bamb)
```

```
## [1] 27.58695 17.70964 42.97318
```

Ved en kolinesterase aktivitet på 50 forventer jeg en gennemsnitlig værdi på first evoked response på 27,59 min (95% konfidensinterval: 17,71 - 42,97). Da denne værdi for aktivitet nu ligger inde i mine observerede værdier, kan jeg bedre stole på dette estimat. Jeg har et meget bredt konfidensinterval pga. de få observationer. Den estimerede værdi er højere end det tilsvarende estimat for de normale (23.69 min).

(c) Ser hældningerne for de to grupper ud til at være rimeligt ens?

På en log-log skala er hældningen for bambuterol-gruppen -0.44 og for de resterende deltagere -0.51. På scatterplottet hvor der er tegnet en linie ind for hver gruppe (se næste spørgsmål) synes jeg også, at hældningerne ser nogenlunde ens ud. Jeg ville dog teste for interaktion (dvs. forskellig hældning for de to gruppers regressionslinier) før jeg udtaler mig om dette.

Spørgsmål 6. Sammenlign nu de to lineære relationer på log-log skala ved at bygge en model for samtlige personer, inkluderende to forskellige linier. Husk passende illustrationer til analyserne.

Jeg vælger at tillade interaktion, dvs. forskellig effekt af kolinesterase på first evoked response i de to grupper. Dette giver to hældningsestimater.

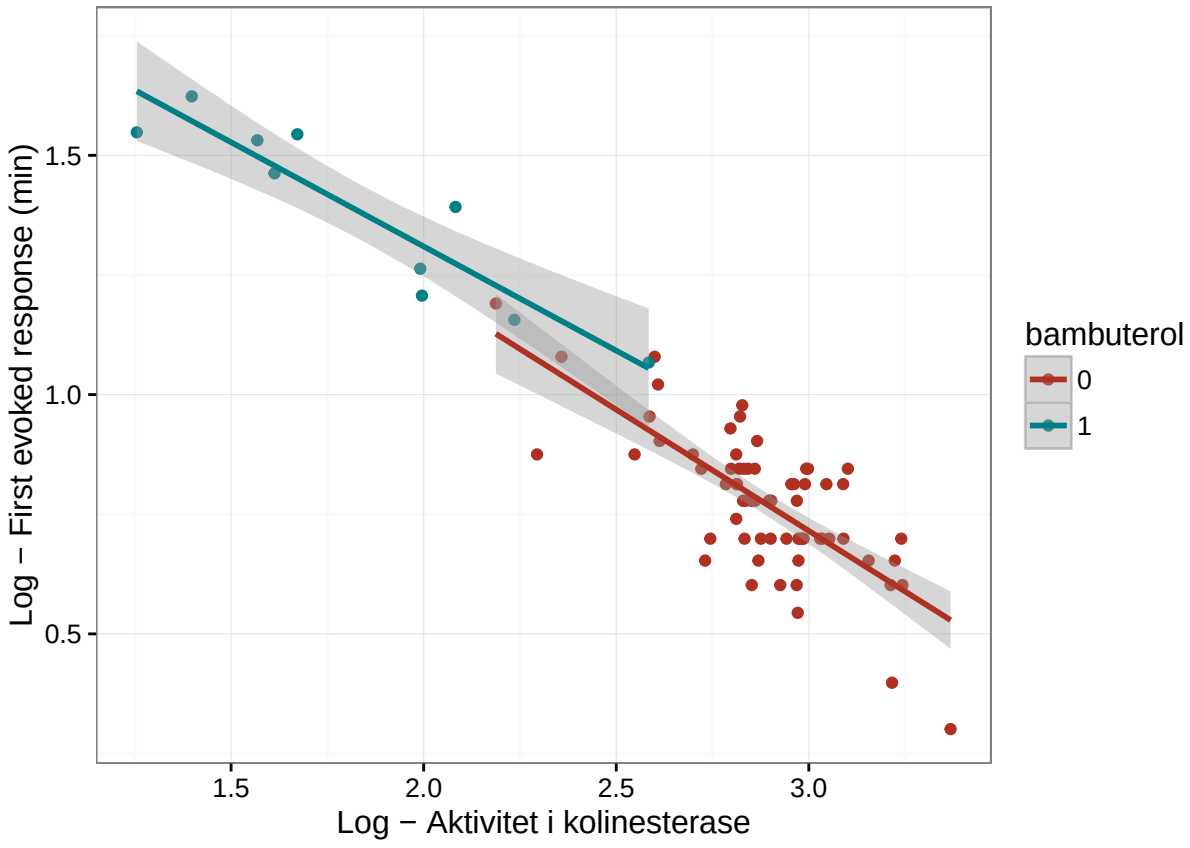
```
fit2 <- lm(formula = logfer ~ logke*bambuterol, data=dataset)
summary(fit2)
```

```
##
## Call:
## lm(formula = logfer ~ logke * bambuterol, data = dataset)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.227780 -0.049156  0.005154  0.056887  0.181148
##
## Coefficients:
```

```
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.23549    0.16205  13.795  < 2e-16 ***
## logke              -0.50672    0.05605  -9.041 1.39e-13 ***
## bambuterol1        -0.05505    0.21937  -0.251    0.803
## logke:bambuterol1   0.07131    0.09657   0.738    0.463
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09656 on 74 degrees of freedom
## Multiple R-squared:  0.8601, Adjusted R-squared:  0.8544
## F-statistic: 151.6 on 3 and 74 DF,  p-value: < 2.2e-16
```

Jeg tegner to regressionslinier (med og uden bambuterol).

```
ggplot(dataset, aes(x=logke, y=logfer, color=bambuterol)) +
    geom_point(shape=19) +
    xlab("Log - Aktivitet i kolinesterase") +
    ylab("Log - First evoked response (min)") +
    scale_colour_hue(l=40) +
    theme_bw() +
    geom_smooth(method=lm, se=TRUE, fullrange=FALSE)
```



(a) Er der evidens for, at kolinesterase har en forskellig effekt på first evoked response for de to grupper?

Hvad kaldes det, hvis dette er tilfældet?

Da p-værdien for effekten af bambuterol er 0.458 er der ikke signifikant forskellig effekt af kolinesterase-aktivitet afhængig af bambuterol-gruppen. Der er altså ikke sikkert forskellig hældning mellem linierne. Vi kan også sige at der *ikke* er tegn på interaktion.

(b) Kvantificer forskellen på de to grupper ved en kolinesterase aktivitet på 200. Er denne signifikant forskellig fra 0?

Da der ikke var tegn på interaktion i modellen, bruger jeg en additiv model.

```
fit2 <- lm(formula = logfer ~ logke + bambuterol, data=dataset)
summary(fit2)
```

```
##
## Call:
## lm(formula = logfer ~ logke + bambuterol, data = dataset)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.239415 -0.040372  0.004641  0.057233  0.176268
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.16622    0.13174  16.443   <2e-16 ***
## logke       -0.48270    0.04550 -10.608   <2e-16 ***
## bambuterol1  0.10122    0.05762   1.757   0.0831 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09627 on 75 degrees of freedom
## Multiple R-squared:  0.859,  Adjusted R-squared:  0.8553
## F-statistic: 228.5 on 2 and 75 DF,  p-value: < 2.2e-16
```

Først beregnes den estimerede værdi sv.t. en kolinesterase aktivitet på 200 i gruppen uden bambuterol

```
c <- log10(200)
predict1 <- predict(fit2, newdata = data.frame(logke = c, bambuterol = "0"),level=.95,interval="confide
predict1
```

```
##        fit       lwr      upr
## 1 1.055519 0.9978059 1.113231
```

```
10^c(predict1)
```

```
## [1] 11.363670  9.949606 12.978704
```

Derefter beregnes den estimerede værdi sv.t. en kolinesterase aktivitet på 200 i gruppen med bambuterol

```
predict2 <- predict(fit2, newdata = data.frame(logke = c, bambuterol = "1"),level=.95,interval="confide
predict2
```

```
##       fit      lwr      upr
## 1 1.15674 1.083068 1.230412
```

```
10^c(predict2)
```

```
## [1] 14.34631 12.10789 16.99856
```

Da konfidensintervallerne overlapper hinanden, kan jeg ikke forkaste nulhypotesen om at der ikke er forskel i first evoked response mellem de to grupper ved en kolinesterase aktivitet på 200.

Jeg forsøger at svare mere præcist ud fra modellen: Da jeg har valgt en additiv model er linierne parallelle, og derfor er forskellen i y mellem bambuterol-gruppen og normal-gruppen den samme, for alle værdier på x. Der kan læses fra modellen at forskellen er 0,1 (log-skala) og denne er ikke signifikant forskellig fra 0 p=0,08. Ved tilbagetransformering får vi at bambuterol-gruppen har en first evoked responstid som ligger 26% over de normale (for den samme kolinesterase aktivitet):

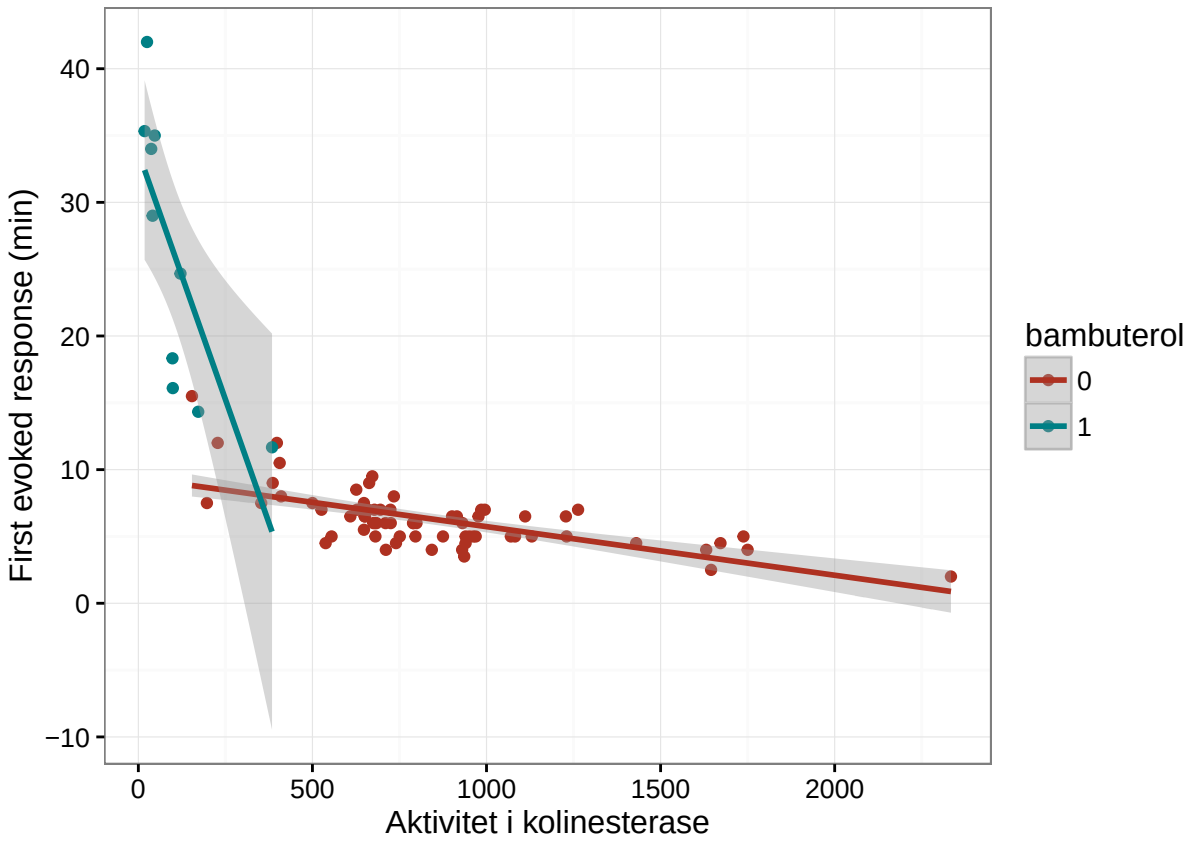
```
10^0.10122
```

```
## [1] 1.262467
```

Spørgsmål 7. Hvordan ville sammenligningen af de lineære relationer tage sig ud på den oprindelige (utransformerede) skala?

(Dette behøver I kun svare på ud fra en passende figur).

```
ggplot(dataset, aes(x=ke, y=fer, color=bambuterol)) +
    geom_point(shape=19) +
    xlab("Aktivitet i kolinesterase") +
    ylab("First evoked response (min)") +
    scale_colour_hue(l=40) +
    theme_bw() +
    geom_smooth(method=lm, se=TRUE, fullrange=FALSE)
```



Da hældningerne er meget forskellig kan transformering helt klart anbefales. Udover at hældningen er forskellig, er det interval for kolinesterase aktivitet forskellig i grupperne med kun et meget lille overlap. Derfor ville man kunne forvente at modellen bliver påvirket mere af bambuterol gruppen i den lave ende af kolinesterase aktivitet og i den høje ende af de resterende deltagere. Derudover er konfidensgrænserne for bambuterol gruppen brede (pga. det lille antal observationer), og der er altså meget mere usikkerhed omkring værdierne i denne ende af skalaen.