R Notebook

Det er lidt et Hack, men for at sikre eksamensnummer på hver side kan det findes i url'en nederst til venstre

Opgave 1

Eksmanesnummer : 132

```
data1 <- read.table(file = "nov24opg1.txt", header = T)
```

```
range(log(data1$StorageTime))
```

```
## [1] 0.000000 7.521318
```

```
range((data1$StorageTime))
```

```
## [1]    1 1847
```

Opgave 1.1

Eksmanesnummer : 132

Data:
Par $(x_1, y_1), \ldots, (x_n, y_n)$ bestående af kvantitative kontinuerte data, både for vores responsvariabel logaritmen af fedtmålinger i gram og den forklarende variabel logaritmen af opbevaringstid i fryser målt i dage.

Statistisk model:

$$y_i = \mu_i + \epsilon_i$$

hvor $\epsilon_i \sim N(0, \sigma^2)$

Vi antager $y_1, \ldots, y_n$ er uafhængihed, og at alle y_i er normaltfordelt med middelværdi $\mu_i = \alpha + \beta x_i$ (ret linje) med spredning σ.

De ukendte populationsparametre
Skæringen α, hældningen β og spredningen σ er alle ukendte parametre, som vi kan estimere: $\hat{\alpha}, \hat{\beta}, \hat{\sigma}$

```
mod11 <- lm(log(Fat) ~ log(StorageTime), data=data1)
summary(mod11)
```

```
##
## Call:
## lm(formula = log(Fat) ~ log(StorageTime), data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3601 -1.2098  0.0161  1.2098  4.3791
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.75080    0.20735   8.444  < 2e-16 ***
## log(StorageTime) -0.16045    0.03923  -4.090 4.77e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.664 on 755 degrees of freedom
## Multiple R-squared:  0.02168,    Adjusted R-squared:  0.02038
## F-statistic: 16.73 on 1 and 755 DF,  p-value: 4.771e-05
```

Ud fra summary kan vi se estimater skæring, hældning og residualspredningen

$$\hat{\alpha} = 1.75, \hat{\beta} = -0.16, \hat{\sigma} = 1.664$$

Skæringsparameteren kan give os et estimat for median værdien for fedtmålingen ved en nedfysningstid på en dag

```
exp(1.75080)
```

```
## [1] 5.759208
```

Medianværdien for fedtmålingen på Teju'er som har været nedfrusset i en dag er 5.76.

Opgave 1.2

Eksmanesnummer : 132

Vi ønsker at teste hypotsen $H_0 : \beta = 0$, hvilket er et udtryk for at nedfrysningstiden ikke har nogen indflydelse på fedtmængden.

Dette kan klares med en t-test størrelse, som kan aflæses direkte ud fra summary

Med en t-værdi på -4.090 og en p-værdi på 4.77e-05 < 0.001 < 0.05 afviser vi nulhypotesen om at der ikke er nogen påvirkning. Det betyder at 0 ikke ligger i 95% konfidensintervallet for den sande middelværdi

```
confint(mod11)
```

```
##                      2.5 %      97.5 %
## (Intercept)       1.3437404  2.15785279
## log(StorageTime) -0.2374621 -0.08344521
```

Vi ser, ligesom vi også så det i estimatatet, at log(storagetime) har en negativ indfyldelse på fedtmængden på logaritmisk skala. Udfra konfidensintervallet for log(storagetime) (-0.24, -0.08)

Opgave 1.3

Eksmanesnummer : 132

Med en kontinuert og en kategorisk forklarende variable fittes ANCOVA model.

```
mod12 <- lm(log(Fat) ~ log(StorageTime) + factor(Year), data=data1)
```

Begge er varianter af linære modeller, man kan anse den mod11 til at være en delmodel/nested-model af mod12. Begge modeller er på formen

$$y_i = \mu_i + \epsilon_i$$

hvor $\epsilon_i \sim N(0, \sigma^2)$

Dette ses mere tydeligt hvis vi opskriver den:

$$y_i = \alpha_{\texttt{grp(i)}} + \beta \cdot x_i + \epsilon_i$$

hvor ϵ_i `iid.` $N(0, \sigma^2)$

Skæringsparameteren afhænger af gruppen og hældningen er den samme for alle grupper. For for dermed en statisktisk model med flere rette parallele linjer.

For at teste om nedfrysnings tiden har en påvirkning skal vi teste hypotesen $H_0 : \beta = 0$

Dette kan gøres med en f-teststørrelse, hvor vi tester modellen op imod en 1-sidet Anova. Hvilket vi kan fordi den 1-sidet anova er en også er en nested model af ANCOVA

Vi anvender en F-test til at teste hele hypotesen: * Fullmodel: Ankova med log(storagetime) som den kontinuerte forklarende variabel * Nulmodel: 1-sidet anova med uden log(storagetime), altså kun Year som forklarende variabel

```
fullModel <- mod12
nullModel <- lm(log(Fat) ~ factor(Year), data=data1)

anova(nullModel, fullModel)
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	750	1954.603	*NA*	*NA*	*NA*	*NA*
2	749	1951.261	1	3.34145	1.28263	0.2577731

2 rows

Vi afviser ikke hypotsen om at nedfrysningstiden har en signifikant indflydelse på den målte fedtmængde, når der også tages højde for ind fangningsåret. Med en f-testtørrelse på 1.28 og en p-værdi på 0.26 forkastes hypotesen ikke på et 95% signifikantnivue.

Dette må altså betyde at 95%-konfindensintervallet for det sande ukendte gennemsnit indenholder 0.

```
confint(mod12)
```

```
##                       2.5 %      97.5 %
## (Intercept)       1.4524706  3.14103772
## log(StorageTime) -0.1492417  0.04004327
## factor(Year)2013 -1.1812638  0.64381454
## factor(Year)2014 -1.1737779  0.44788417
## factor(Year)2015 -1.8467720 -0.34033739
## factor(Year)2016 -1.8092771 -0.23461658
## factor(Year)2017 -4.4524504 -1.88765162
## factor(Year)2018 -2.2483291 -0.70012288
```

Vi ser at 0 ligger i intervallet (-0.15 0.04). Vi kan ikke bekræfte vores hypotese, kun undlade at afvise den. Men det er værd at bemærke at en ANCOVA med hældning lig 0, udgør en model med flere vandrette linjer, en for hver gruppe, hvilket er det samme som en 1-sidet ANOVA.

Opgave 1.4

Eksmanesnummer : 132

Først og fremmest vælger vi en 2-sidet ANOVA, fordi vi har to kategoriske forklarende variable og 1 kontinuert responsvariabel.

Modellen med vekselvirkning udtrykker, at den forventede fedtmængde afhænger både af årstal og køn, uden antagelsen om, at de to kategoriske variable virker additivt. Vi antager altså ikke på forhånd, at der skal være samme forskel på den forventede fedtmængde for hanner og hunner for hvert år.

Den additivte antagelse, **som vi ikke laver**, svare til $\beta_3 = 0$ i den statisktiske model:

$$y_i = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e_i$$

hvor $e_i \sim N(0, \sigma^2)$

```
# Jeg ved godt der ikke er nogen grund til at factor Sex da datatypen er char/string, # Men nu h
ar jeg gjordt det ...
mod14 <- lm(log(Fat) ~
              factor(Year) + factor(Sex) +
              factor(Year) * factor(Sex)
              , data=data1)
summary(mod14)$coefficients
```

```
##                                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)                   2.4205006  0.5092910  4.7526869 2.410408e-06
## factor(Year)2013             -0.6139920  0.7036866 -0.8725362 3.831978e-01
## factor(Year)2014             -0.6060208  0.5774817 -1.0494200 2.943261e-01
## factor(Year)2015             -1.3049490  0.5309726 -2.4576582 1.421197e-02
## factor(Year)2016             -1.1892991  0.5470663 -2.1739579 3.002329e-02
## factor(Year)2017             -3.3797740  0.9527962 -3.5472161 4.137608e-04
## factor(Year)2018             -1.7033466  0.5339084 -3.1903346 1.480798e-03
## factor(Sex)M                 -0.7842200  0.7399826 -1.0597816 2.895883e-01
## factor(Year)2013:factor(Sex)M  0.7517663  0.9493527  0.7918724 4.286878e-01
## factor(Year)2014:factor(Sex)M  0.6168444  0.8242096  0.7484072 4.544514e-01
## factor(Year)2015:factor(Sex)M  0.5309803  0.7665094  0.6927250 4.886986e-01
## factor(Year)2016:factor(Sex)M  0.2299796  0.7951093  0.2892427 7.724763e-01
## factor(Year)2017:factor(Sex)M  0.4683111  1.3094931  0.3576278 7.207234e-01
## factor(Year)2018:factor(Sex)M  0.3600426  0.7704019  0.4673438 6.403909e-01
```

Estimatet for den forventet log-fedtmængde for en teju af hunkøn i år 2014 er 1.81

```
2.4205006 + -0.6060208
```

```
## [1] 1.81448
```

Estimatet for den forventet log-fedtmængde for en teju af hankøn i år 2018 er 0.55

```
2.4205006 + (-1.7033466) + (-0.7842200) + 0.6168444
```

```
## [1] 0.5497784
```

Opgave 1.5

Eksmanesnummer : 132

Spørgsmålet kan omfumulere til at udersøge, om der er samme forskel på log(Fat) for SEX = M og SEX = F for alle år. Hvilket er svare til at teste nulhypotesen $H_0 : \beta_3 = 0$.

Vi anvender en F-test til at teste hypotesen: * Fullmodel: 2-sidet ANOVA med vekselvirkning * Nulmodel: 2-sidet ANOVA uden vekselvirkning

```
fullmodel <- mod14
nullmodel <- lm(log(Fat) ~ factor(Year) + factor(Sex), data=data1)

anova(nullmodel, fullmodel)
```

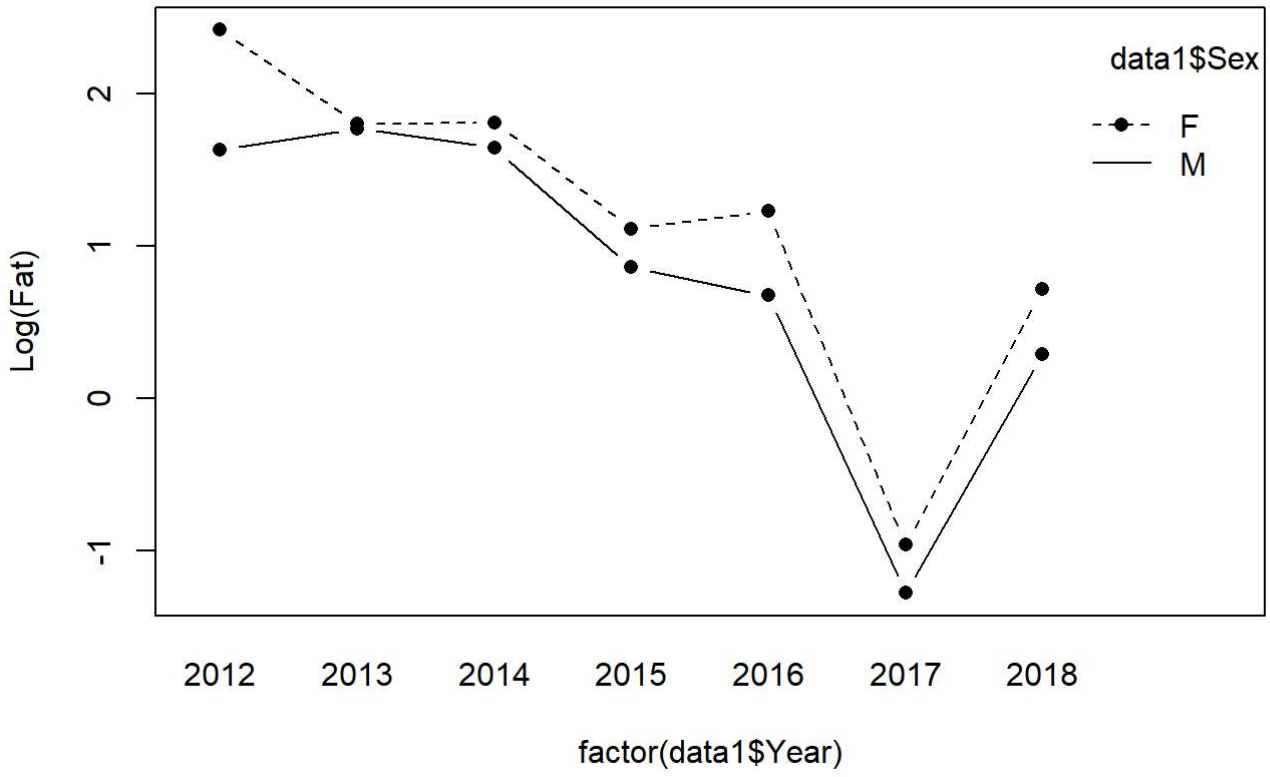
Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	749	1931.662	*NA*	*NA*	*NA*
2	743	1927.173	6	4.488364	0.2884064	0.9425395

2 rows

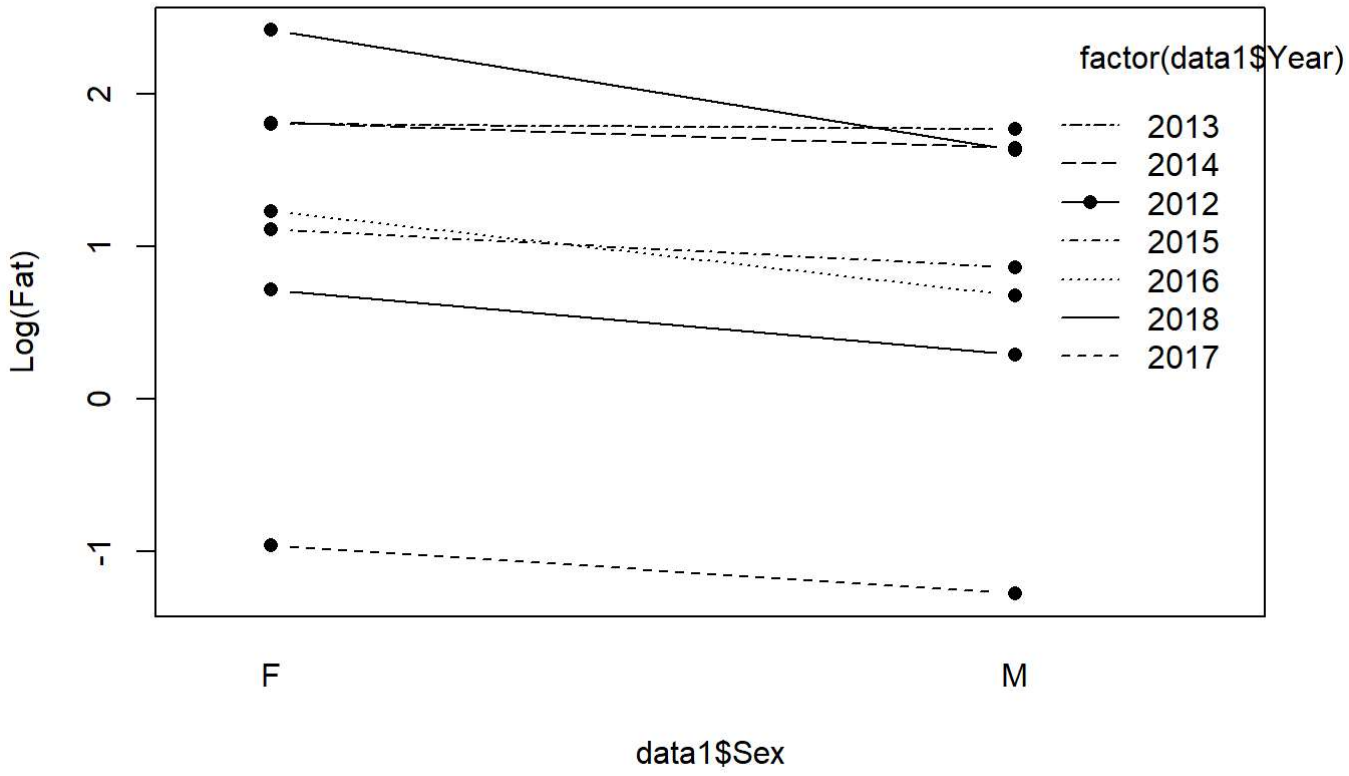
Vi afviser ikke hypotesen $ H_0 : _3 = 0$ på et signifikantnivue på 95% med en f-teststørrelse på 0.2884 og 0.9425. Data taler altså for at forskellen på hanner og hunner ikke afhænger af indfangningsåret.

Resulatet kontrolleres grafisk:

```
interaction.plot(factor(data1$Year), data1$Sex, log(data1$Fat), type="b", pch=16, ylab = "Log(Fa
t)")
```




```
interaction.plot(data1$Sex, factor(data1$Year), log(data1$Fat), type="b", pch=16, ylab = "Log(Fa
t)")
```



Profilerne i vekselvirkningsgraferne ser ganske parallelle ud, så plottene tyder umiddelbatrt på at der ikke er vekselvirkning mellem indfangningsår og køn i deres beskrivelse af højden.

Opgave 1.6

Eksmanesnummer : 132

Her giver det bedst mening at tage udgangspunkt i den additive model, både baseret på konklussionen i opgave 1.5, samt at vi i den additative ikke behøver at udvælge et år, men kan udføre undersøgelsen for et vilkårligt år.

```
mod16 <- nullmodel
exp(confint(mod16))
```

```
##                       2.5 %     97.5 %
## (Intercept)      4.41002491 19.0467692
## factor(Year)2013 0.33108364  2.0393421
## factor(Year)2014 0.33209000  1.6608686
## factor(Year)2015 0.16634690  0.7442256
## factor(Year)2016 0.15613248  0.7387442
## factor(Year)2017 0.01190359  0.1527137
## factor(Year)2018 0.10106885  0.4556223
## factor(Sex)M     0.55856959  0.8868876
```

95%-konfidensintervallet for forskellen i fedtprocenten på hanner og hunner for Teju'er som er fanget indenfor det samme år er (0.559, 0.887)

Opgave 2

Eksmanesnummer : 132

```
data2 <- read.table(file = "nov24opg2.txt", header = T)
head(data2, n = 3)
```

	Fat <dbl>	Weight <dbl>	Length <dbl>
1	54.479	1030	29.4
2	21.896	1560	34.5
3	10.885	880	29.5

3 rows

```
range(data2$Fat)
```

```
## [1]  0.016 72.028
```

```
range(data2$Weight)
```

```
## [1]   48.3 1780.0
```

```
range(data2$Length)
```

```
## [1] 11.3 37.4
```

Opgave 2.1

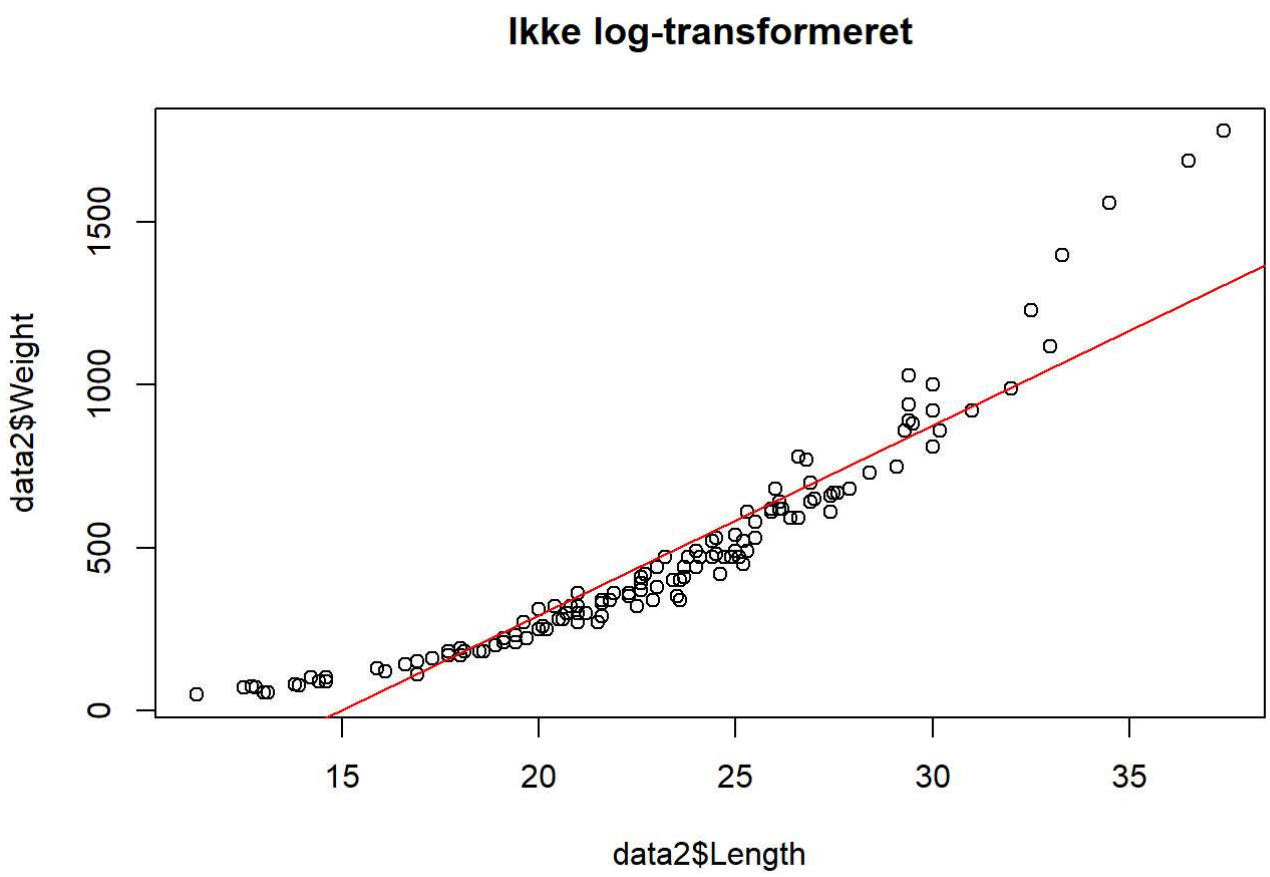
Eksmanesnummer : 132

```
mod21 <- lm(Weight ~ Length, data=data2)
mod22 <- lm(log(Weight) ~ Length, data=data2)
mod23 <- lm(log(Weight) ~ log(Length), data=data2)
```

Data og ret linje

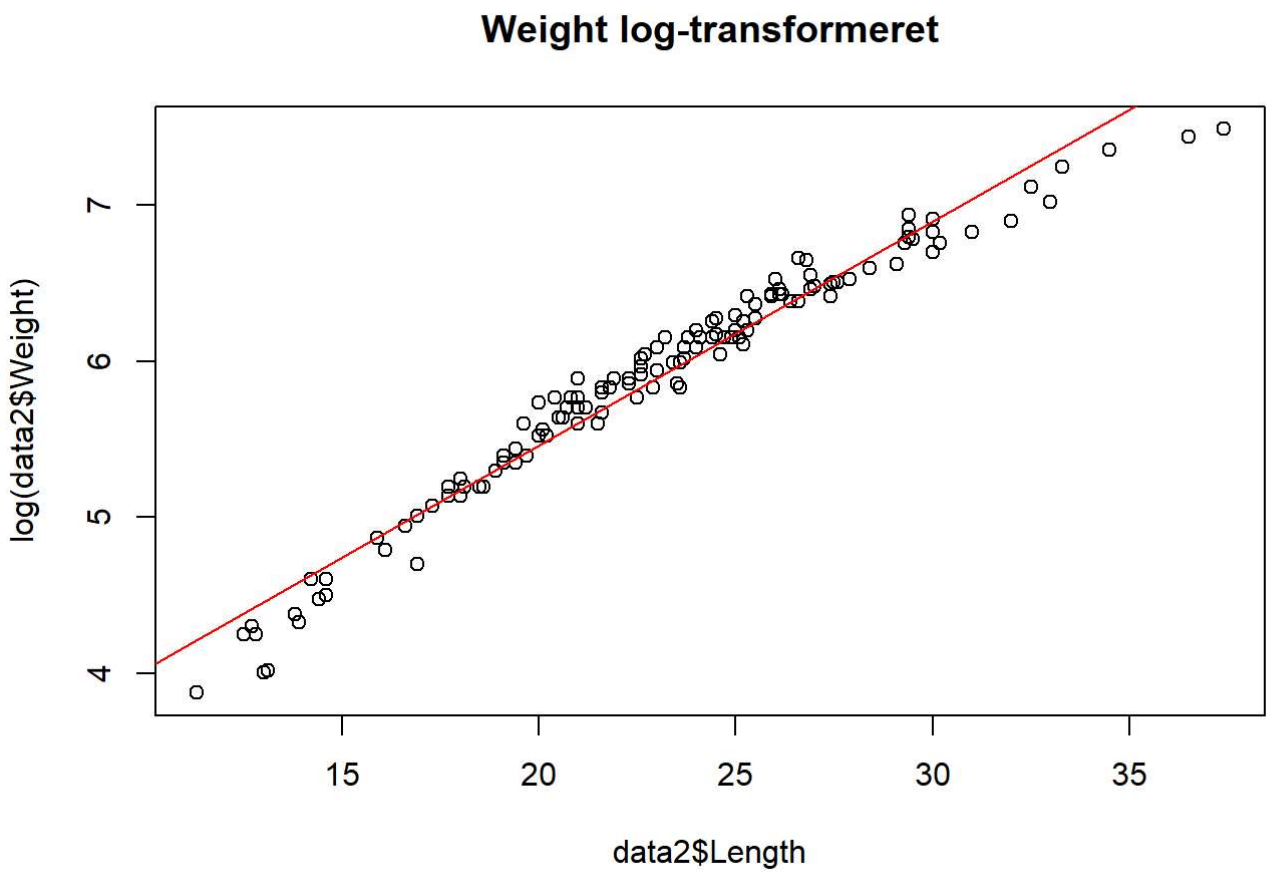
Vi starter med at plotte modellerne op imod data.

```
plot(data2$Length, data2$Weight, main = "Ikke log-transformeret")
abline(mod21[1], mod21[2], col = 'red')
```



Her ses vi en krumning i datapunkterne. Der er systematiske afvigelse mellem linjen og punkterne.

```
plot(data2$Length, log(data2$Weight), main = "Weight log-transformeret")
abline(mod22[1], mod22[2], col = 'red')
```



Dette plot ser meget pænere ud, man ville udfra det her plot godt kunne argumentere for at vi har fundet en god model. Der er ikke nogen markant systematisk afvigelse. Dog ser vi at overvejde mange punkter ligger over linjen i midten og under linjen ved de højeste og laveste værdier.

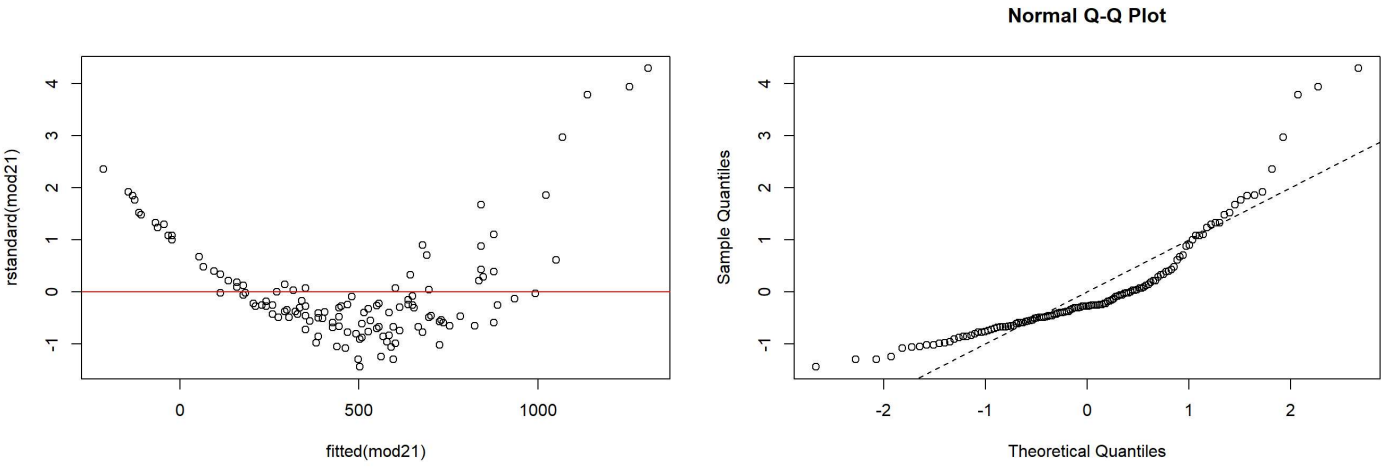
```
plot(log(data2$Length), log(data2$Weight), main = "Weight og Lenght log-transformeret")
abline(mod23[1], mod23[2], col = 'red')
```

Dette er det pæneste af plotsene. Vi ser ikke nogen systematisk afvigelse fra linjen ved nogen intervaller af log(Lenght)

Residualer og normalfordelings antagelse

```
plot(fitted(mod21), rstandard(mod21))
abline(0,0, col = 'Red')

qqnorm(rstandard(mod21))
abline(0,1, lty=2)
```

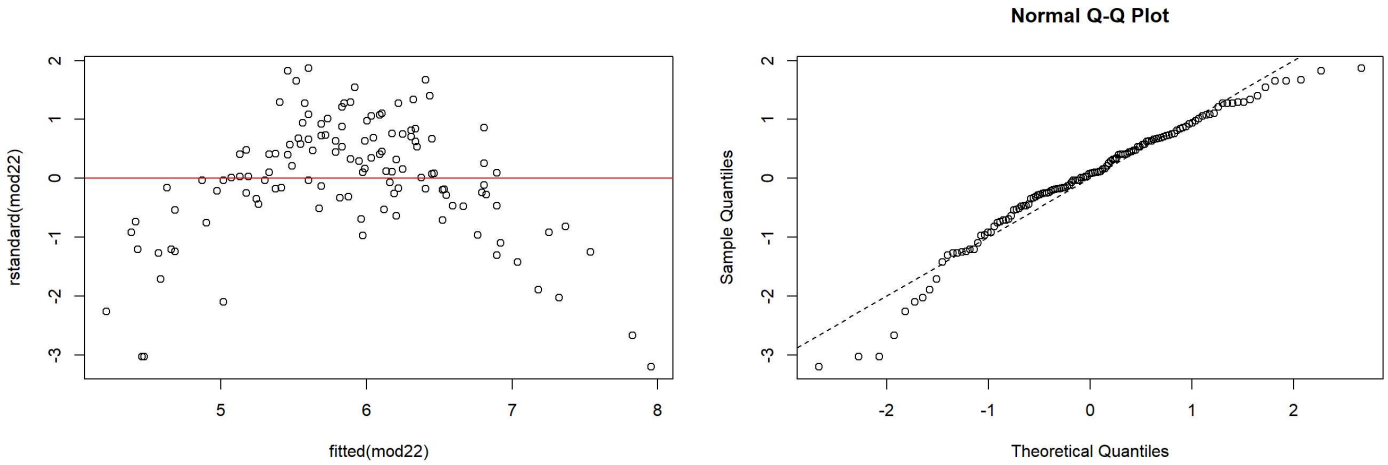


Uden logtransformation :

Her ser en tydeligt koveks krumning r residualplottet, og der er tydelig vertikal systematisk afvigelse omkring den vandrette linje. I qq-plottet ses også en tydelig systematisk afvigelse ved en konveks krumning.

```
plot(fitted(mod22), rstandard(mod22))
abline(0,0, col = 'Red')

qqnorm(rstandard(mod22))
abline(0,1, lty=2)
```



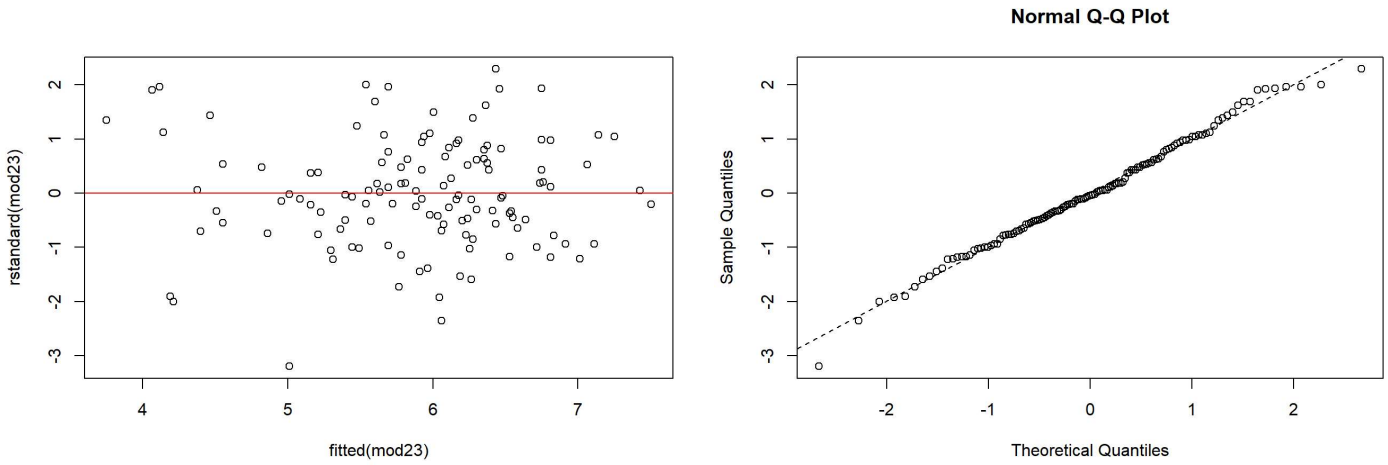
Log transformeret weight :

I qq-plottet er den systematiske afvigelse ikke ligeså markant, men den ses udenfor intervallet ca. (-1.5, 1.5).

I residualplottet ser vi dog en tydelig kokav krumning omkring 0. Altså en markant systematisk vertikal afvigelse.

```
plot(fitted(mod23), rstandard(mod23))
abline(0,0, col = 'Red')

qqnorm(rstandard(mod23))
abline(0,1, lty=2)
```



Log transformeret weight og længde :

Residualplot
Vi ser ikke nogen vertikal systematisk afvigelse omkring 0, der er ca. samme mængde punkter over og under linjen. Vi ser dertil også at ca. 95% af datapunkterne ligger indenfor 2 og -2, ligesom vi forventer ved en normalfordeling. Der er helle ikke nogle tegn på variansinhomoginitet.

QQ-plot
Der er ikke nogen systematisk afvigelse fra linjen i QQ-plottet, hvilket betyder at data ligger i de kvartiler, som vi forventer ved en normalfordelingen.

Ud fra dette, kan vi se at antagelsen om normalfordelte residulaler er korrekt. Og det bekræfter os i at middelværdi antagelsen er korrekt (Altså at vi har valgt den rigtige statistiske model).

Da ser vi at model mod23 hvor vi har logtransformeret både den forklarende og den kontinuerte varible er den mest velegnede af de tre modeller til beskrive sammenhængen mellem vægten og længden.

Opgave 2.2

Eksmanesnummer : 132

Den linære regressionsmodel $ y_i = _i + _i $ er er et ret linje hvor middelværdien kan beskrives som:
$\mu_i = \alpha + \beta x_i$

Forskernes gæt svare til at hældningen er lig 3. Derfor kan vi teste hypotesen $H_0 : \beta = 3$

Vi anvender en F-test til at teste hypotesen: * Fullmodel: log transformeret linær regressionsmodel * Nulmodel: log transformeret linær regressionsmodel med hældning = 3

```
fullmodel <- lm(log(Weight) ~ log(Length), data=data2)
nullmodel <- lm(log(Weight) ~ offset(3 * log(Length)), data=data2)

anova(nullmodel, fullmodel)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	129	1.383285	*NA*	*NA*	*NA*	*NA*
2	128	1.237426	1	0.1458591	15.08775	0.0001637019

2 rows

Vi afviser 0-hypotesen $H_0 : \beta = 3$ med et 95% signifinansnivue med f-testtørrelse på 15.088 en p-værdi på 0.0001637

Dette må betyde at 3 ikke ligger i 95%-konfidensintervallet for hældningen.

```
confint(mod23)
```

```
##                 2.5 %    97.5 %
## (Intercept) -4.076364 -3.640309
## log(Length)  3.067424  3.207440
```

Som forventtet ligger 3 ikke i intervallet 95% konfidenintervallet for hældningen (3.067, 3.207)

Opgave 2.3

Eksmanesnummer : 132

Vi anvender en Multilineær regression model med 2 kontinuerte variable.

Statistisk model
Udgangspunkt i to forklarende variable:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$$

hvor $e_i \sim N(0, \sigma^2)$

De ukendte populationsparametre
De partielle hældninger β_1, β_2, skæring med y aksen α og spredningen σ er populationsparametre som vi kan estimere: $, , , $

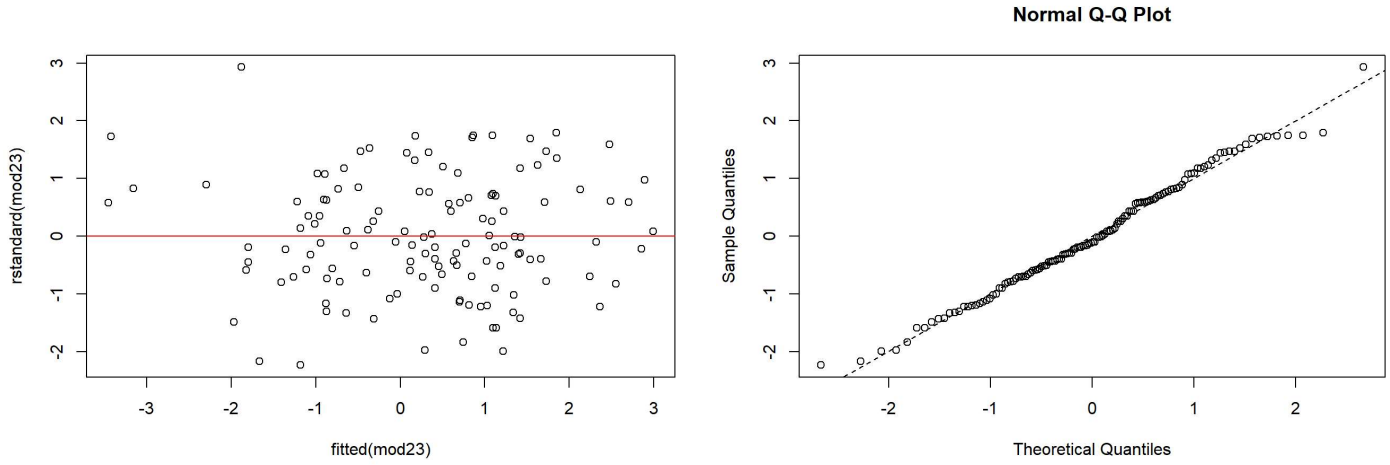
Skal vægt og længde log transformeres

Første skridt er at finde ud af om middelværdi antagelsen fungere for hvis vi log-transformere de forklarende variable

```
mod23 <-  lm(log(Fat) ~ log(Length) + log(Weight), data=data2)
```

```
plot(fitted(mod23), rstandard(mod23))
abline(0,0, col = 'Red')

qqnorm(rstandard(mod23))
abline(0,1, lty=2)
```



Residualplot
Vi ser nogen systematisk afvigelse i intervallet (-4. -3), men dette kan godt skyldes tilfældig variation. Vi ser dertil også at ca. 95% af datapunkterne ligger indenfor 2 og -2, ligesom vi forventer ved en normalfordeling. Der er helle ikke nogle tegn på variansinhomoginitet.

QQ-plot
Der er ikke nogen systematisk afvigelse fra linjen i QQ-plottet, udover en mindre afvigelse for de højeste værdier. Dette betyder at data nogenlunde ligger i de kvartiler, som vi forventer ved en normalfordelingen.

Ud fra dette, kan vi se at antagelsen om normalfordelte residulaler er korrekt. Og det bekræfter os i at middelværdi antagelsen er korrekt (Altså at vi har valgt den rigtige statistiske model).

prædiktion

```
# Vi behøver ikke log tranformere det nye input, den klare R :)
new_data <- data.frame(Length = 22.6, Weight = 386 )
interval <- predict(mod23, newdata = new_data, interval = "p")
exp(interval)
```

```
##        fit       lwr     upr
## 1 1.821056 0.1843393 17.9899
```

Median fedt mængden for en ny observation med længde 22.6 cm og vægt 386 g er 1.82 g.

Det hertil liggende prædiktions interval er (0.18, 17.99) hvilket vil sige at 95% af nye observationer med den angivet længde og vægt vil have en fedtmængde, som ligger indefor intervallet.

Opgave 3

Eksmanesnummer : 132

3.1 : C

3.2 : E

3.3 : A

3.4 : D

3.5 : A

3.6 : B