Opgavebesvarelse, Resting metabolic rate

I filen `'rmr.txt'` findes sammenhørende værdier af kropsvægt (`bw`, **i kg**) og hvilende stofskifte (`rmr`, **kcal pr. døgn**) for 44 kvinder (Altman, 1991 og Owen et.al., *Am. J. Clin. Nutr.*, **44**, 1986).

Filen indeholder 45 linier, først en linie med variabelnavnene (`bw` og `rmr`) og derefter 44 datalinier, hver med disse to oplysninger.

Vi ønsker at konstruere normalområder for stofskiftet, som funktion af kropsvægten.

Vi starter med at indlæse data direkte fra hjemmesiden:

```
rmr <-
read.table("http://publicifsv.sund.ku.dk/~lts/basal/data/rmr.txt",
header=T)
```

Spørgsmål 1.

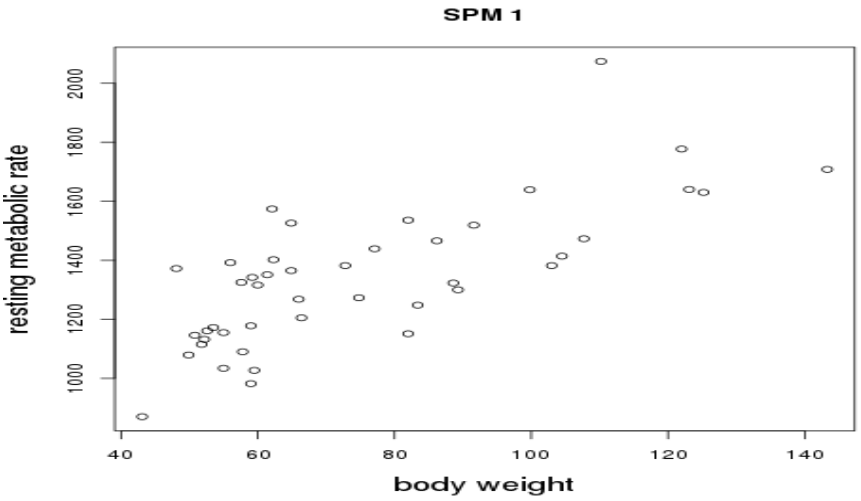
Lav en tegning af de sammenhørende værdier af kropsvægt og stofskifte. Overvej i denne forbindelse, hvad der bør være X- hhv. Y-akse.

Da opgaven går ud på at konstruere normalområder for stofskiftet, som funktion af kropsvægten, må vi udnævne stofskiftet til at være responsen (Y), medens kropsvægten `bw` er den forklarende variable, altså X.

Vores plotte-kode bliver derfor som vist nedenfor (hvor der er tilføjet overskrift og aksetekster i halvanden størrelse):

```
plot(rmr$bw, rmr$rmr, main="SPM 1", ylab="resting metabolic rate",
xlab="body weight", cex.lab=1.5)
```

Herved får vi figuren



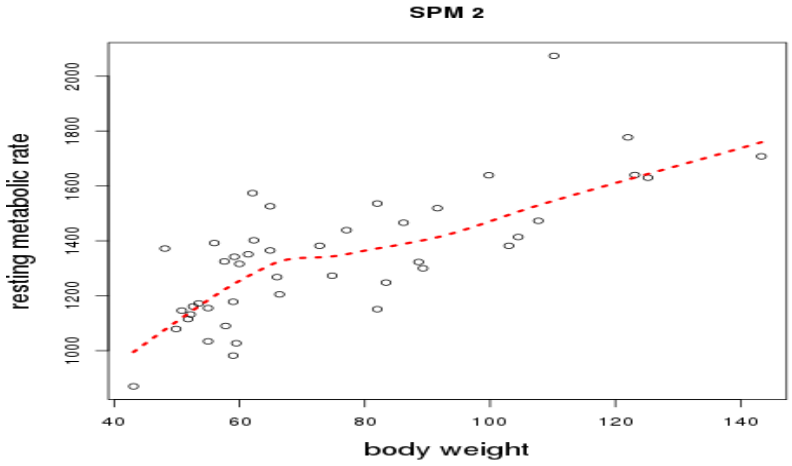
Spørgsmål 2.

Opstil en rimelig model (evt. ved først at transformere data på passende vis, hvis du synes, modelforudsætningerne halter), og estimer parametrene i denne, med tilhørende spredninger (spredninger på parameterestimater = standard errors).

Figuren ovenfor ser jo rimelig lineær ud, omend en vis affladning synes at forekomme, således at vi for høje kropsvægte ikke finder helt det høje stofskifte, som vi ville forvente i en lineær model. Vi kan illustrere dette ved at indlægge en blød kurve (smoother) på scatter plottet ved i stedet at skrive

```
with(rmr, scatter.smooth(bw, rmr, main="SPM 2",
  ylab="resting metabolic rate", xlab="body weight", cex.lab=1.5,
  lpars = list(col = "red", lwd = 3, lty = 3)))
```

hvorved vi får figuren



Baseret på denne figur og det faktum, at vi har ret få observationer i det høje område, vil vi fortsætte uden transformation.

Vi vil derfor foretage en sædvanlig lineær regression med `rmr` (**Y**) som respons og `bw` (**X**) som forklarende variabel, uden at transformere nogen af dem.

Vi gør dette ved at skrive

```
model1 = lm(rmr ~ bw, data=rmr)
summary(model1)
```

hvorved vi får outputtet

```
> summary(model1)

Call:
lm(formula = rmr$rmr ~ rmr$bw)

Residuals:
    Min      1Q  Median      3Q     Max
-245.74 -113.99  -32.05  104.96  484.81

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 811.2267    76.9755  10.539 2.29e-13 ***
rmr$bw        7.0595     0.9776   7.221 7.03e-09 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 157.9 on 42 degrees of freedom
Multiple R-squared:  0.5539,Adjusted R-squared:  0.5433
F-statistic: 52.15 on 1 and 42 DF,  p-value: 7.025e-09
```

Vi ser af ovenstående output, at effekten af `bw` er stærkt signifikant, idet et test af hældningen $\beta = 0$ giver T=7.22, svarende til en P-værdi, der er mindre end 0.0001. Samme P-værdi får man (naturligvis) ved at anvende F-testet, idet F=7.22^2=52.15$\sim$ F(1,42). Bemærk, at der ikke i outputtet findes noget test for linearitet!

Parameterestimaterne ses at være:

afskæring α	*hældning* β	spredning s
kcal pr. døgn	kcal pr. døgn pr. kilo	kcal pr. døgn
811.23(76.98)	7.06(0.98)	157.9

En simpel **aflæsning** angiver standard error på hældningen til 0.9776, altså s.e.($\hat{\beta}$)=0.9776. Dette spredningsestimat har naturligvis samme enheder som selve hældningsestimatet, altså kcal pr. døgn pr. kilo.

Et konfidensinterval for hældningen konstrueres efter den sædvanlige formel: estimat $\pm$ ca.2 $\times$ s.e.(estimat). De '*ca. 2*' skal i virkeligheden være en 97.5% fraktil i en t-fordeling med 42 frihedsgrader, nemlig 2.018. Med de ovenfor angivne værdier finder vi altså (ved brug af lommeregner)

$$7.0595 \pm 2.018 \times 0.9776 = (5.0866, 9.0324)$$

men det behøver vi jo slet ikke, for det kan vi få R til at gøre ved at skrive:

```
> confint(model1)
                 2.5 %   97.5 %
(Intercept) 655.883819 966.5695
rmr$bw        5.086656   9.0324
```

Fortolkningen af dette interval (5.09,9.03) er:

1. De værdier af hældningen, der ikke ville give en signifikant teststørrelse ved test på 5% niveau
   (altså hvis vi f.eks. testede om hældningen var 9).

2. Intervallet fanger den sande hældning med 95% sandsynlighed.

Aflæsning fra ovenstående regressionsanalyseoutput giver ligeledes estimatet for spredningen omkring regressionslinien til 157.9.

Dette spredningestimat har de samme enheder som vores oprindelige observationer, altså kcal pr. døgn.

Spørgsmål 3.

Hvad er det forventede hvilestofskifte for kvinder på 70 kg?
Hvis det tilsvarende hvilestofskifte for mænd (på 70 kg) vides at være skønnet til 1406 kcal/døgn (med CI på 1360-1452), kan vi så konkludere, at der er forskel på hvilestofskiftet hos mænd og kvinder på 70 kg?
(Dette spørgsmål kan evt. besvares løseligt ud fra en tegning med indlagte konfidensgrænser).

Ud fra estimaterne som angivet i tabellen ovenfor kan vi nu på en passende valgt lommeregner lave regnestykket

$$811.23 + 7.06 \times 70 = 1305.43$$

Vi kan også lade R finde dette estimat til os ved at prediktere ud fra en ny data frame, bestående af en enkelt `bw`-værdi:

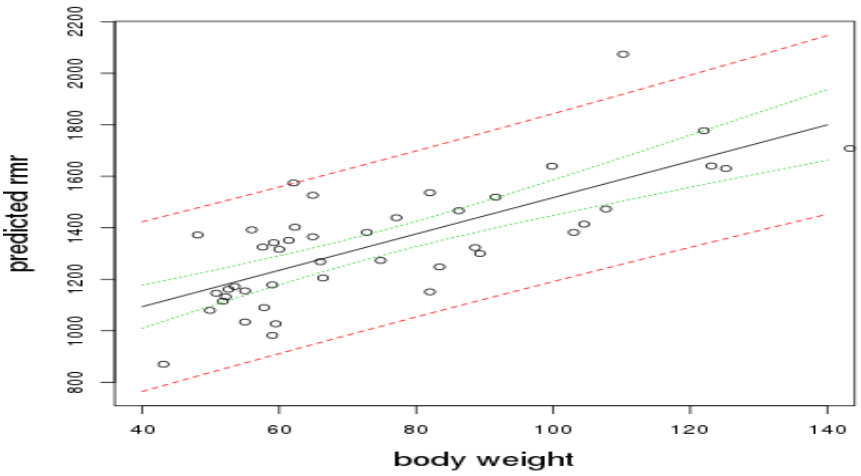
```
> bw70 <- data.frame(bw=70)
> predict(model1, newdata=bw70, interval="confidence")
       fit      lwr      upr
1 1305.394 1256.398 1354.389
```

Svaret er altså en forventet værdi på ca. 1305, med et konfidensinterval på (1256, 1354).

Hvis det tilsvarende hvilestofskifte for mænd vides at være skønnet til 1406 kcal/døgn (med CI på 1360-1452), kan vi se, at de to konfidensintervaller ikke overlapper, og dermed kan vi konkludere, at mænd på 70 kg har et højere hvilestofskifte end tilsvarende kvinder.

Dette spørgsmål kunne også være besvaret ved blot at se på en figur med indtegnede konfidensgrænser, som kan fås ved at skrive:

```
ny = data.frame(bw=40:140)
predict(model1, ny, se.fit = TRUE)
predl <- predict(model1, ny, interval = "prediction")
confl <- predict(model1, ny, interval = "confidence")

matplot(ny$bw, cbind(predl, confl[,-1]), lty = c(1,2,2,3,3),
        col = c(1,2,2,3,3), type = "l", cex.lab=1.5,
        ylab = "predicted rmr", xlab="body weight")
points(rmr$bw, rmr$rmr)
```



I denne figur kan vi for `bw=70` aflæse de relevante oplyninger, omend ikke med den helt store nøjagtighed.

Spørgsmål 4.

I ambulatoriet ser vi en kvinde, der vejer 80 kg.

- **Hvad er dit bedste gæt på hendes hvilestofskifte (i mangel af yderligere information)? Og indenfor hvilke grænser vil du med 95% sikkerhed tippe hendes hvilestofskifte at ligge, forudsat at hun er rask.**

  Ligesom ovenfor kan vi udfra estimaterne udregne det forventede hvilestofskifte til

  $$811.23 + 7.06 \times 80 = 1376.03$$

  Prediktionsintervallet kan (approksimativt) fås ved at lægge $\pm 2 \times 157.9$ til de estimerede 1376.0, idet de 157.9 er spredningen omkring regressionslinien. Formelt er det kun for en kvinde med gennemsnitsvægt, at dette gælder eksakt, men der er ikke den store forskel, som det fremgår af ovenstående figur.

  Vi finder altså grænserne

  $$1376.03 \pm 2 \times 157.9 = (1060.2, 1691.8)$$

  Hvis vi benytter R til at udregne disse grænser, får vi

```
> bw80 <- data.frame(bw=80)
> predict(model1, newdata=bw80, interval="prediction")
       fit      lwr      upr
1 1375.989 1053.564 1698.414
```

  dvs. en anelse bredere end vores håndregning: (1053.6, 1698.4).

- **Hvis hun viser sig at have et hvilestofskifte på 950 kcal/døgn, hvilken diagnose ville du så stille, og hvorfor?**

  Da 950 kcal/døgn ikke ligger i det ovenfor udregnede interval, må vi diagnosticere denne kvinde til at have et for lavt stofskifte i forhold til sin vægt.

- **Er det muligt (med 95% sikkerhed) at forudsige en enkelt kvindes stofskifte udfra kropsvægten med en nøjagtighed på +/- 250 kcal/døgn ?**

For en kvinde med gennemsnitsvægt $\bar{X}$, har prediktionsintervallet en bredde på ca. 2*2*157.9 = 631.6, og da dette er mere end 2*250=500, må svaret være benægtende: Vi kan ikke foretage så nøjagtig en prediktion med 95% grænser.

Vi kan også regne den anden vej, altså bestemme procentdækningen for et interval, der går 250 kcal. pr. døgn til hver side. Vi skal så finde ud af, hvor mange spredninger (spredningen var 157.9 kcal. pr. døgn), de 250 kcal/døgn svarer til, hvilket svarer til ratioen

```
> 250/157.9
[1] 1.583281
```

og dermed til 94.3%-fraktilen, idet

```
> pnorm(250/157.9)
[1] 0.9433212
```

Da 250 kcal/døgn således svarer til 94.3%-fraktilen i den normerede normalfordeling (se evt. en tabel), således at der er 5.7% tilbage i hver hale. Det må betyde, at området i midten omfatter $100 - 2 \times 5.7 = 88.6$ % af fordelingen, og dermed af fremtidige observationer.

Man kunne så overveje, hvordan man kunne gøre disse grænser smallere, altså hvordan man kan nedbringe variationen omkring regressionslinien. Der kunne tænkes flere muligheder:

- Medtage flere personer i undersøgelsen:
  Dette ville næppe hjælpe, idet spredningen omkring regressionslinien er et udtryk for den biologiske variation i stofskiftet for kvinder med samme kropsvægt - og denne ændrer sig jo ikke af, at der bliver flere kvinder.
- Fjerne de yderligt liggende observationer:
  Man må **ikke** bare fjerne observationer uden gyldig grund, og det er **ikke** gyldig grund, at observationen ligger langt fra regressionslinien!! Man kan fjerne observationer, hvis de adskiller sig fra de øvrige ud fra *andre* kriterier end det observerede stofskifte, f.eks. kropsvægten eller oplysninger om helt specielle forhold (at kvinden måske var professionel atlet el.lign.). I et sådant tilfælde skal man huske at fjerne *samtlige* kvinder, der opfylder dette kriterium, idet det da skal betragtes som et eksklusionskriterium.

– Inddrage yderligere information til at forbedre prediktionen:
Dette er absolut en farbar vej, og man kan ofte finde adskillige nyttige forklarende variable, som kan nedbringe variationen. Her kunne det måske være alternative mål for kropsbygning eller oplysninger om fysisk aktivitet i hverdagen. Analysen ville da blive en *multipel regression*.

Spørgsmål 5.

Vurder om modellens forudsætninger kan siges at være opfyldt. Suppler evt. med teoretiske overvejelser omkring stofskifte.

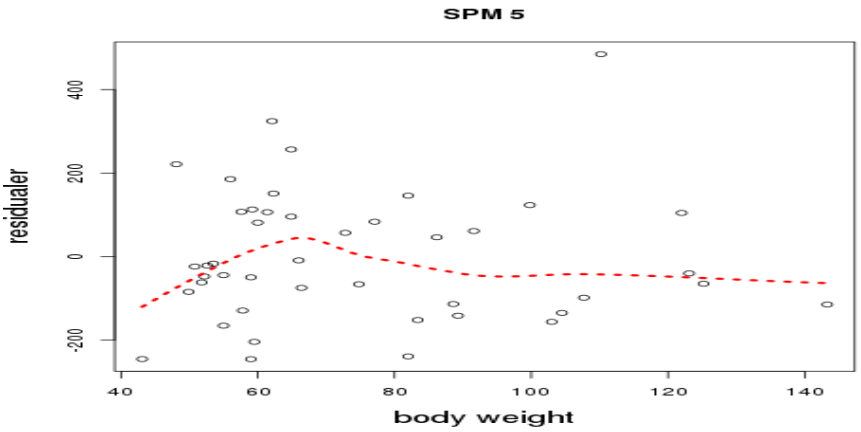
Relevante modelkontroltegninger kunne være:

1. Scatterplot af residualer (en af de 4 slags) mod værdier af den forklarende variabel (`bw`)
2. Scatterplot af residualer (en af de 4 slags) mod predikterede værdier
3. Fraktildiagram for residualer
4. Histogram for residualer

Det første af disse udføres ved at skrive:

```
with(rmr, scatter.smooth(bw, resid(model1), main="SPM 5",
ylab="residualer", xlab="body weight", cex.lab=1.5,
lpars = list(col = "red", lwd = 3, lty = 3)))
```

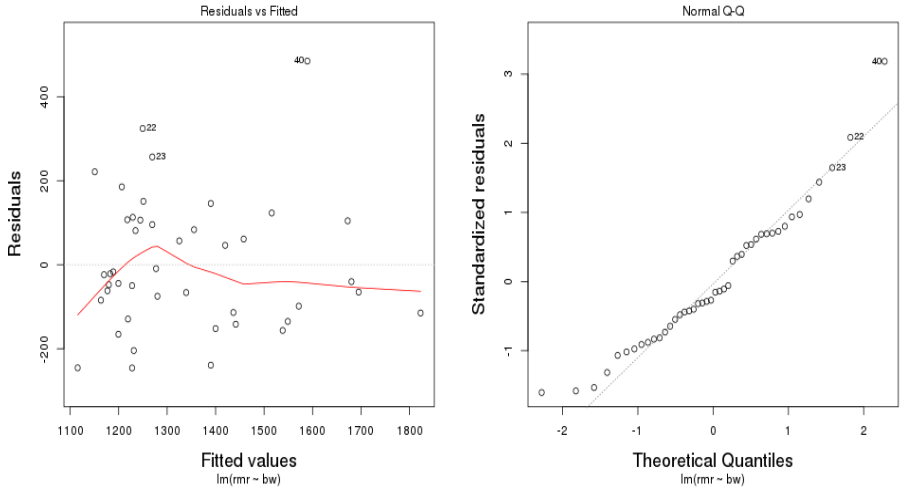
hvorved vi får dannet et residualplot (residualer mod kovariat), med en udglattet kurve



som viser en smule tegn på en bue i det lave område.

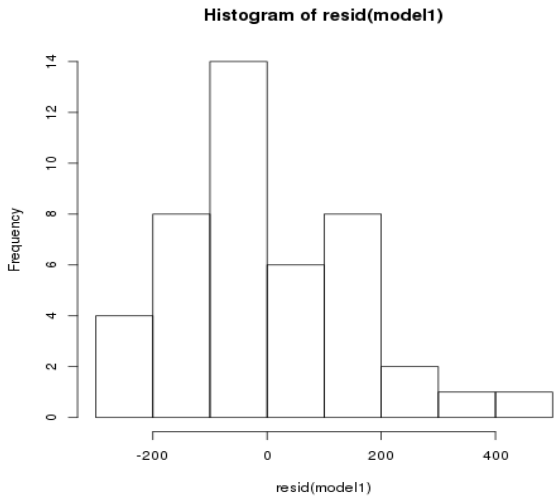
Desuden kan vi R lave en del automatiske plots, heriblandt nr. 2 og 3 ovenfor:

```
par(mfrow=c(1,2))
plot(model1,which=1, cex.lab=1.5)
plot(model1,which=2, cex.lab=1.5)
```



og vi kan supplere med et histogram over residualerne:

```
hist(resid(model1))
```



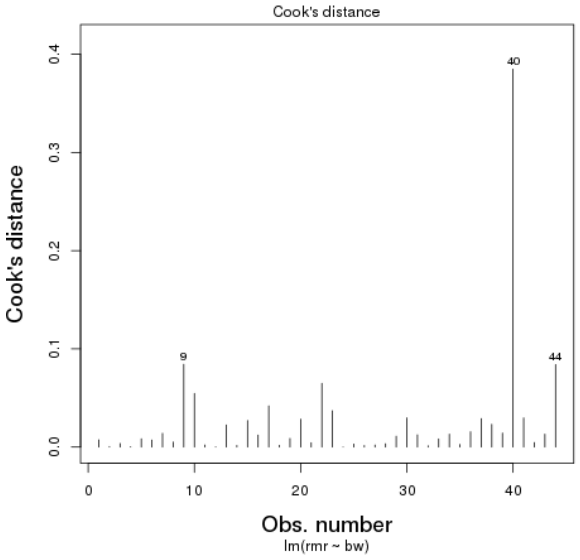
Ingen af de ovenfor viste tegninger tilføjer tvivl om modelantagelserne (ud over den allerede omtalte tendens til bue i den lave ende).

Spørgsmål 6.

Overvej, om der muligvis er enkelte observationer, der har en speciel kraftig indflydelse på estimationen.

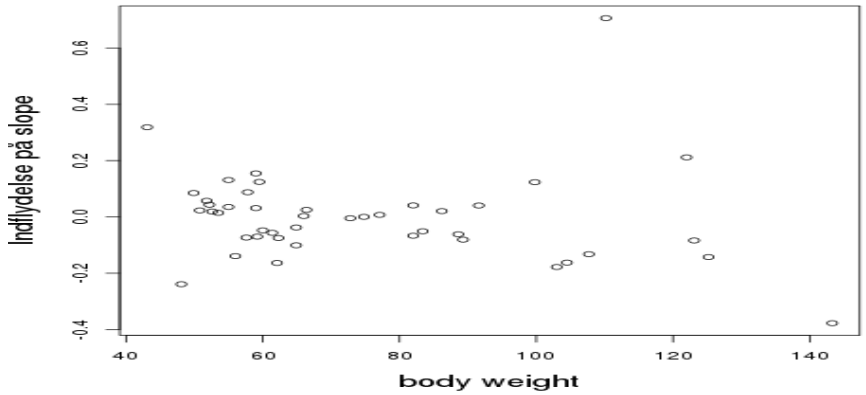
Udfra det oprindelige scatter plot, ser den mest suspekte observation ud til at være den med den højeste `rmr` (hvilket ses at være observation nr. 40 med en `rmr`-værdi på over 200 kcal. pr. døgn). Vi kan få et plot af Cooks afstand ved at skrive

```
plot(model1,which=4, cex.lab=1.5)
```



og på denne figuren ser vi også klart, hvordan denne observation skiller sig ud fra de andre. Vi kan også spalte Cooks afstand ud i to dele, og kun se på indflydelsen på hældningsestimatet (estimat nr.2), da det er det, der har vores primære interesse:

```
plot(rmr$bw,influence(model1)$coefficients[,2], cex.lab=1.5,
     ylab="Indflydelse p{\aa} slope", xlab="body weight")
```



Vi kan også prøve at foretage analysen uden denne observation. Vi udelukker den ved i regressionsanalysekaldet at benytte en reduceret data frame:

```
> rmr2 = rmr[rmr$rmr<2000,]
> model2 = lm(rmr ~ bw, data=rmr2)
> summary(model2)
```

som giver os

```
Call:
lm(formula = rmr ~ bw, data = rmr2)

Residuals:
    Min      1Q  Median      3Q     Max
-256.08  -91.22  -25.44  103.61  327.21

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 852.2484    68.7965   12.39 1.92e-15 ***
bw            6.3534     0.8836    7.19 8.90e-09 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 139.2 on 41 degrees of freedom
Multiple R-squared:  0.5577,Adjusted R-squared:  0.5469
F-statistic:  51.7 on 1 and 41 DF,  p-value: 8.896e-09
```

og tillige

```
> confint(model2)
                 2.5 %     97.5 %
(Intercept) 713.311044 991.185749
bw            4.568847   8.137901
```

Vi ser, at hældningsestimatet ændrede sig fra 7.06(0.98) til 6.35(0.88), hvilket set i lyset af standard error ikke ligefrem er alvorligt, men dog værd at bemærke.

Hvorvidt en sådan ændring er betydningsfuld, må afhænge af formålet med studiet. Man kan f.eks. se på ændringen i prediktionsgrænserne og vurdere, om de er store. Hvis de er det, må man konkludere, at der ikke er information nok til at udtale sig om normalområder for stofskiftet.

Som diskuteret under spørgsmål 4, kan man **ikke** bare smide den pågældende kvinde ud af materialet. Der er ikke nogen objektiv grund til dette, og det er i hvert fald ikke en god grund i sig selv, at hun tillægges meget vægt i estimationen!

Opgavebesvarelse, Definition af BMI

Vi skal se på rimeligheden i definitionen af body mass index

$$\text{BMI} = \frac{\text{vægt i kg}}{(\text{højde i m})^2}$$

og skal hertil benytte Sundby-data.

1. **Transformer ovenstående teoretiske relation (altså selve formlen) med logaritmen, dvs. find et teoretisk udtryk for logaritmen til BMI.**

   For nemheds skyld benyttes her den naturlige logaritme log (tidligere kaldet `ln`), idet man så undgår at skulle skrive fodtegn hele tiden. Formlerne er dog nøjagtigt de samme, hvis man benytter en hvilkensomhelst anden logaritme.

   $$\begin{aligned} \log(\text{BMI}) &= \log(\text{vægt i kg}) - \log((\text{højde i m})^2) \\ &= \log(\text{vægt i kg}) - 2\log(\text{højde i m}) \end{aligned}$$

   Hvis alle havde samme BMI, hvilken lineær relation ville vi så forvente at finde mellem de logaritmerede værdier af vægt og højde?

   Nu lader vi som om bmi er konstant (vi kunne definere $\alpha = \log(\text{BMI})$), og så omarrangerer vi ovenstående formel, så vi får

   $$\log(\text{vægt i kg}) = \alpha + 2\log(\text{højde i m})$$

   Hvis vi nu indfører betegnelserne

   $$\begin{aligned} Y &= \log(\text{vægt i kg}) \\ X &= \log(\text{højde i m}) \end{aligned}$$

   kan vi skrive denne relation som

   $$Y = \alpha + 2X$$

   altså en ret linie med afskæring α og hældning $\beta = 2$.

   I de næste spørgsmål skal vi se, om dette ser fornuftigt ud i Sundby-materialet.

2. **Læs nu datasættet `sundby` ind igen.**

   Det har vi prøvet mange gange før, og vi indfører i samme omgang labels for kønnet, samt nye variabelnavne:

```
su <-
read.table(
"http://publicifsv.sund.ku.dk/~lts/basal/data/sundby_lille.txt",
header=T,na.strings=c("."))

su$gender <- factor(su$kon, levels=1:2, labels=c("MALE", "FEMALE"))

su$vaegt=su$v75;
su$hoejde=su$v76/100;
su$bmi=su$vaegt/su$hoejde^2;

su$v75 = NULL
su$v76 = NULL
su$v17 = NULL
su$v24af = NULL
```

3. **Tegn et scatterplot af logaritmen til vægt (vægt er `v75`) overfor logaritmen til højde (højde er `v76`), for hvert køn for sig, og vurder forudsætningerne for at foretage en lineær regression.**

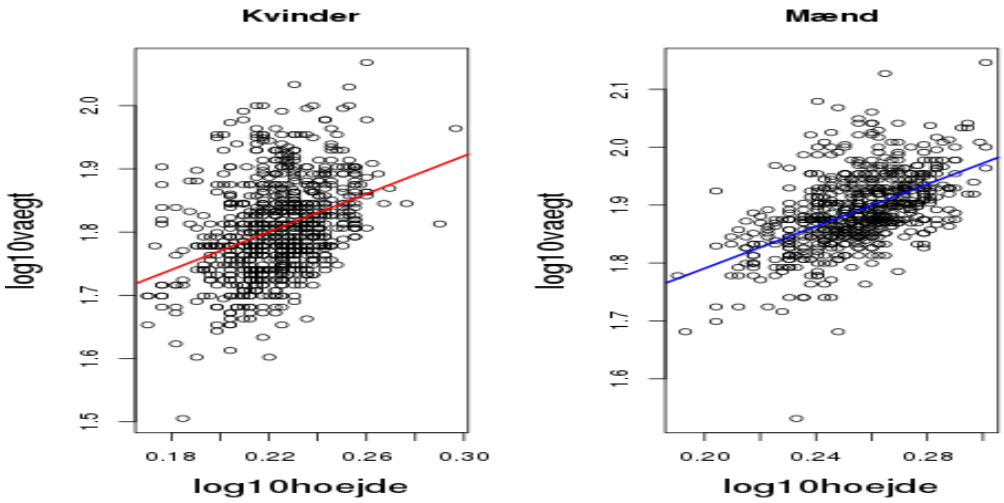
   Vi skal først have defineret logaritmer til både højde og vægt:

```
su$log10hoejde=log10(su$hoejde)
su$log10vaegt=log10(su$vaegt)
```

   og så er vi klar til at tegne, og vi indlægger med det samme en linie i plottene

```
su.F = su[su$gender=="FEMALE",]
```

```
su.M = su[su$gender=="MALE",]

par(mfrow=c(1,2))

plot(su.F$log10hoejde, su.F$log10vaegt, cex.lab=1.5,
  main="Kvinder", ylab="log10vaegt", xlab="log10hoejde")
model.F = lm(log10vaegt ~ log10hoejde, data=su.F)
abline(model.F, col="red",lwd=2)

plot(su.M$log10hoejde, su.M$log10vaegt, cex.lab=1.5,
  main="Mænd", ylab="log10vaegt", xlab="log10hoejde")
model.M = lm(log10vaegt ~ log10hoejde, data=su.M)
abline(model.M, col="blue",lwd=2)
```



En lille teknisk sidebemærkning:
På ovenstående scatterplots er der indlagt regressionslinier. De fleste synes, at disse linier er lidt for flade, fordi man visuelt vil være tilbøjelig til at lægge linien svarende til storcirklen i den ellipse, som visuelt kan lægges uden om punktsværmen. Men regressionslinien skal minimere de kvadratisk *lodrette* afstande til linien, hvilket svarer til, at den skal gå igennem de punkter på ellipsen, som har lodrette tangenter.

Forudsætningerne for at foretage lineær regression er (bortset fra uafhængighed mellem observationerne):

- linearitet
- varianshomogenitet, dvs. konstant spredning, uafhængig af højden
- normalfordelte residualer

Visuelt synes der at være en svag krumning opad ved de store højder, men dette kan ikke bekræftes ved at tilføje et andengradsled i regressionen i spørgsmål 4. Vi kan derfor ikke afvise lineariteten som en fornuftig beskrivelse. De to øvrige antagelser synes visuelt ikke at give nogen problemer.

4. **Fit en lineær relation, med log(vægt) som respons og log(højde) som kovariat, for et af kønnene (eller hvert køn for sig). Hvordan passer resultatet med definitionen af bmi?**

De simple lineære regressionsanalyser giver nedenstående resultater, først for kvinder

```
> summary(model.F)

Call:
lm(formula = log10vaegt ~ log10hoejde, data = su.F)

Residuals:
      Min        1Q    Median        3Q       Max
-0.242352 -0.048705 -0.008375  0.039828  0.217266

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.47039    0.03055   48.13   <2e-16 ***
log10hoejde  1.50043    0.13636   11.00   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06921 on 786 degrees of freedom
  (39 observations deleted due to missingness)
Multiple R-squared:  0.1335,Adjusted R-squared:  0.1324
F-statistic: 121.1 on 1 and 786 DF,  p-value: < 2.2e-16
```

og så for mænd:

```
> summary(model.M)

Call:
lm(formula = log10vaegt ~ log10hoejde, data = su.M)

Residuals:
     Min       1Q   Median       3Q      Max
-0.31920 -0.03314 -0.00379  0.02901  0.21864

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.4276     0.0300   47.58   <2e-16 ***
log10hoejde   1.8160     0.1175   15.46   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05563 on 626 degrees of freedom
  (19 observations deleted due to missingness)
Multiple R-squared:  0.2762,Adjusted R-squared:  0.275
F-statistic: 238.8 on 1 and 626 DF,  p-value: < 2.2e-16
```

De interessante størrelser er her:

Køn	intercept $\hat{\alpha} \approx \log(\text{BMI})$	$10^{\text{intercept}}$ $10^{\hat{\alpha}} \approx \text{BMI}$	hældning "ca. 2 ??"
Kvinder	1.47039 (1.41041, 1.53036)	29.54 (25.73, 33.91)	1.50043 (1.23275, 1.76810)
Mænd	1.42756 (1.36864, 1.48648)	26.76 (23.37, 30.65)	1.81599 (1.58524, 2.04674)

Vi ser, at hældningen for mænd med lidt god vilje godt kan passe med et 2-tal, hvorimod vi for kvindernes vedkommende finder et noget lavere estimat.

5. **Hvordan kunne man forklare en evt. afvigelse fra det forventede?**

Der *er* jo faktisk en afvigelse fra den forventede hældning på 2, i hvert fald for kvindernes vedkommende. Der kunne være flere forklaringer på dette:

- Tilfældigheder:
  Næppe, da vi har at gøre med et stort datamateriale

- "Fejl" i estimationen:
  Faktisk er dette en rimelig indvending, idet der kan være målefejl på såvel højde som vægt. En målefejl på højde vil give sig udslag i en fladere linie, altså en lavere hældning, hvorimod en målefejl på vægt blot ville indgå som en del af residualspredningen.

  Man kan korrigere for en sådan målefejl i kovariaten, ved at dividere hældningen med den såkaldte *reliability coefficient*, der afspejler forholdet mellem målefejl og variation i samplet. Da denne koefficient formentlig her vil være meget tæt på 1, kan målefejlen således ikke være forklaringen på en hældning, der er mindre end 2.

- Misvisende definition af BMI:
  Det er velkendt, at BMI ikke dur til børn, men vi har at gøre med et voksen-materiale her. Alligevel kunne man godt forestille sig, at definitionen af BMI var noget "ad-hoc", altså at man havde indført størrelsen ud fra nemheds-betragtninger, og ikke så meget fordi den nødvendigvis afspejlede den bedste måde til at måle kropsbygning.

  Hvis BMI var et godt mål, ville vi forvente, at det var uafhængigt af højde. Vi kan undersøge dette, dels ved at plotte bmi mod højde og dels ved at teste (Spearman) korrelation lig 0 (Spearman fordi vi blot ønsker at vide, om der *er* nogen sammenhæng, og derfor ligeså godt kan slippe for fordelingsantagelsen).

  Først må vi have defineret BMI:

```
su$bmi=su$vaegt/su$hoejde^2;
su$log10bmi=log10(su$bmi)

su.F = su[su$gender=="FEMALE",]
su.M = su[su$gender=="MALE",]
```

  og derefter udregner vi Spearman korrelationer for hvert køn for sig:

```
with(su.F,cor.test(log10hoejde, log10bmi, method="spearman"))
```

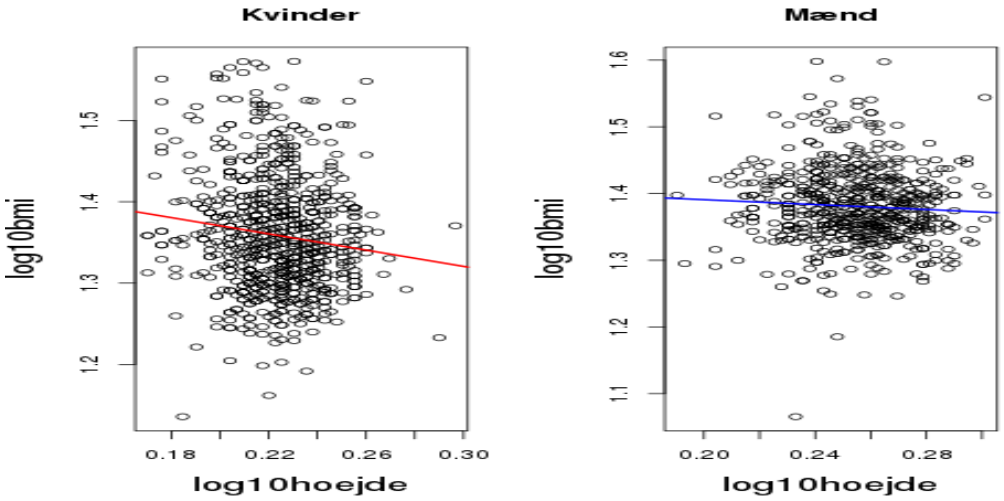
```
with(su.M,cor.test(log10hoejde, log10bmi, method="spearman"))
```

hvorved vi får

```
> with(su.F,cor.test(log10hoejde, log10bmi, method="spearman"))

Spearman's rank correlation rho

data:  log10hoejde and log10bmi
S = 91537000, p-value = 0.0005712
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1224537
```

```
> with(su.M,cor.test(log10hoejde, log10bmi, method="spearman"))

Spearman's rank correlation rho

data:  log10hoejde and log10bmi
S = 45234000, p-value = 0.01632
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.09580746
```

Selv om det kan være svært at se på figurerne nedenfor, viser Spearman korrelationerne, eller rettere, de tilhørende P-værdier, at der *er* evidens for en negativ sammenhæng mellem BMI og højde, for begge køn ($P = 0.0006$ hhv $P = 0.016$). Dette *kan* naturligvis skyldes, at høje mennesker rent faktisk er tyndere end lave mennesker, men det kunne jo også skyldes, at metoden til normering med højden ikke var optimal.


Kvinder
log10bmi
log10hoejde
1.2
1.3
1.4
1.5
0.18
0.22
0.26
0.30
Mænd
log10bmi
log10hoejde
1.1
1.2
1.3
1.4
1.5
1.6
0.20
0.24
0.28