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, og viser samtidig, hvordan man kan tilføje en label til variabelnavnet, således at output bliver mere selvfoklarende:

```
FILENAME navn URL "http://staff.pubhealth.ku.dk/~lts/basal/data/rmr.txt";

data rmrdata;
infile navn firstobs=2;
input bw rmr;

label bw='body weight'
      rmr='resting metabolic rate';
run;
```

Spørgsmål 1.

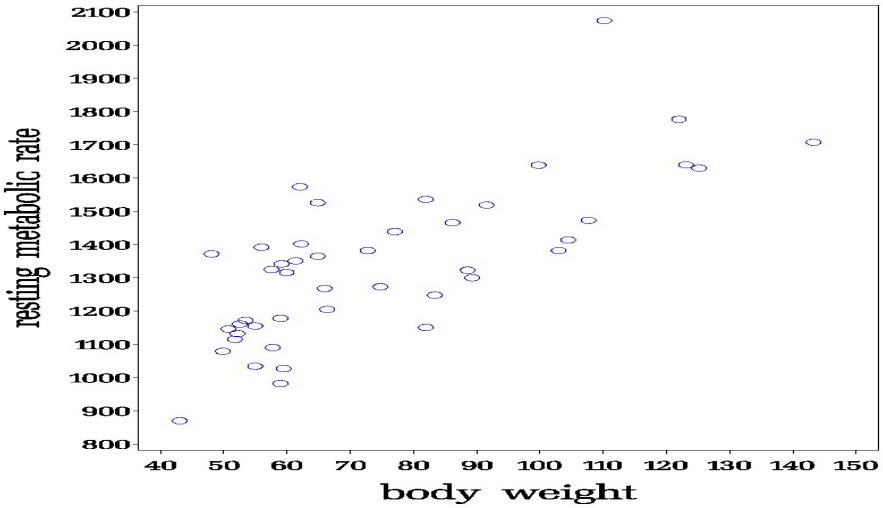
Lav en tegning af de sammenhørende værdier af kropsvægt og stofskifte. Overvej i denne forbindelse, hvad der bør være X- hhv. Y-akse.

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 (idet man dog sagtens kan undlade linie 3-7 i nedenstående, hvis blot man husker at sætte et semikolon i linie 2.

```
proc gplot data=rmrdata;
  plot rmr*bw
 /  haxis=axis1 vaxis=axis2 frame;
axis1 offset=(3,3) label=(H=3) value=(H=2) minor=NONE;
axis2 offset=(1,1) value=(H=2) minor=NONE
      label=(A=90 R=0 H=3);
symbol1 v=circle c=blue i=none h=2 r=1;
run;
```

hvorved vi får



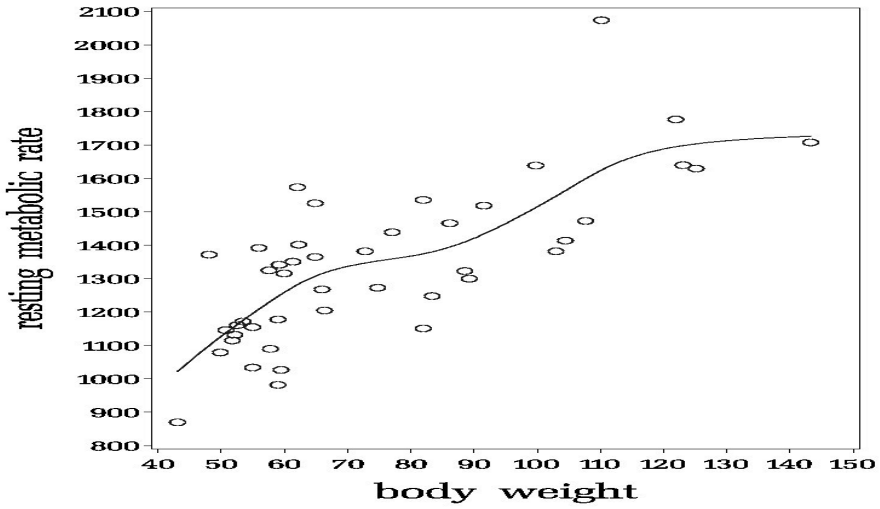
Spørgsmål 2.

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 at ændre `symbol`-sætningen til f.eks.

```
symbol1 v=circle c=blue i=sm60 h=2 r=1;
```

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.

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

I SAS gør vi dette ved at skrive

```
proc glm data=rmrdata;
model rmr = bw / solution clparm;
run;
```

hvorved vi får outputtet

```
The GLM Procedure

Number of Observations Read          44
Number of Observations Used          44

Dependent Variable: rmr   resting metabolic rate

                                  Sum of
Source                     DF       Squares   Mean Square  F Value  Pr > F
Model                       1   1300241.179   1300241.179    52.15  <.0001
Error                      42   1047230.708     24934.064
Corrected Total            43   2347471.886

R-Square     Coeff Var      Root MSE      rmr Mean
0.553890      11.78537      157.9052      1339.841

Source                     DF     Type I SS   Mean Square  F Value  Pr > F
bw                          1   1300241.179   1300241.179    52.15  <.0001

Source                     DF   Type III SS   Mean Square  F Value  Pr > F
bw                          1   1300241.179   1300241.179    52.15  <.0001

                                  Standard
Parameter         Estimate           Error    t Value    Pr > |t|
Intercept      811.2266745     76.97550341      10.54      <.0001
bw               7.0595278      0.97759784       7.22      <.0001

Parameter       95% Confidence Limits
Intercept      655.8838195  966.5695295
bw               5.0866555    9.0324001
```

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 kcal pr. døgn	*hældning* β kcal pr. døgn pr. kilo	spredning s 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 kommer med direkte i outputtet vha option `clparm`

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 SAS finde dette estimat til os ved at tilføje en `estimate`-sætning:

```
proc glm data=rmrdata;
    model rmr = bw / solution clparm;
estimate 'rmr ved 70 kg' intercept 1 bw 70;
run;
```

hvorved vi får

Parameter	Estimate	Standard Error	t Value	Pr > |t|
rmr ved 70 kg	1305.39362	24.2783526	53.77	<.0001

Parameter	95% Confidence Limits	
rmr ved 70 kg	1256.39792	1354.38932

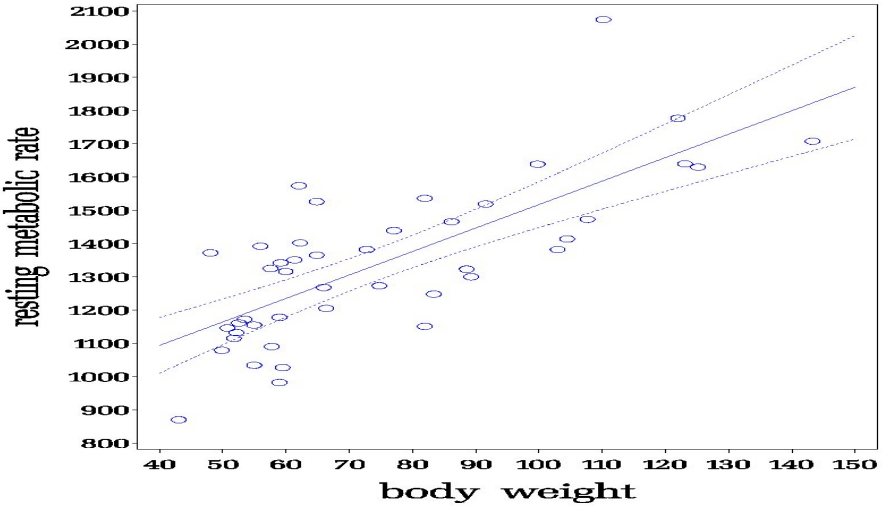
Svaret er altså en forventet værdi på 1305.39, med et konfidensinterval på (1256,1354).

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 fås ved at ændre symbol-sætning i plotteordrerne til

```
symbol1 v=circle c=blue i=rlclm95 h=2 r=1;
```

hvorved vi får figuren



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

Spørgsmål 4.

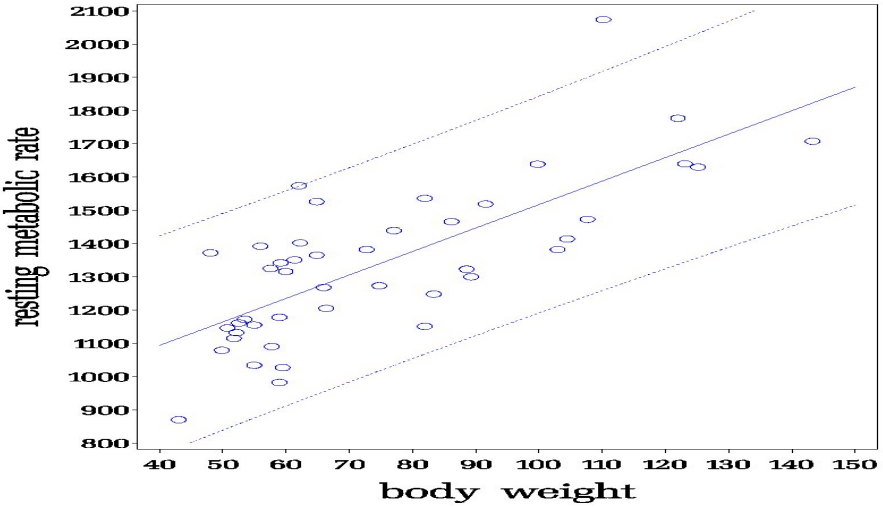
I ambulatoriet ser vi en kvinde, der vejer 80 kg.

- **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

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

  Et 95% prediktionsinterval kan (approksimativt) fås ved at lægge $\pm 2 \times 157.9$, 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 nedenstående figur, hvor der er indlagt 95% prediktionsgrænser (fås ved at ændre symbol-sætningen til `i=rlcli95`).



Vi finder altså grænserne

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

- **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

$$\frac{250}{157.9} = 1.58$$

Dette tal er tæt på 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:

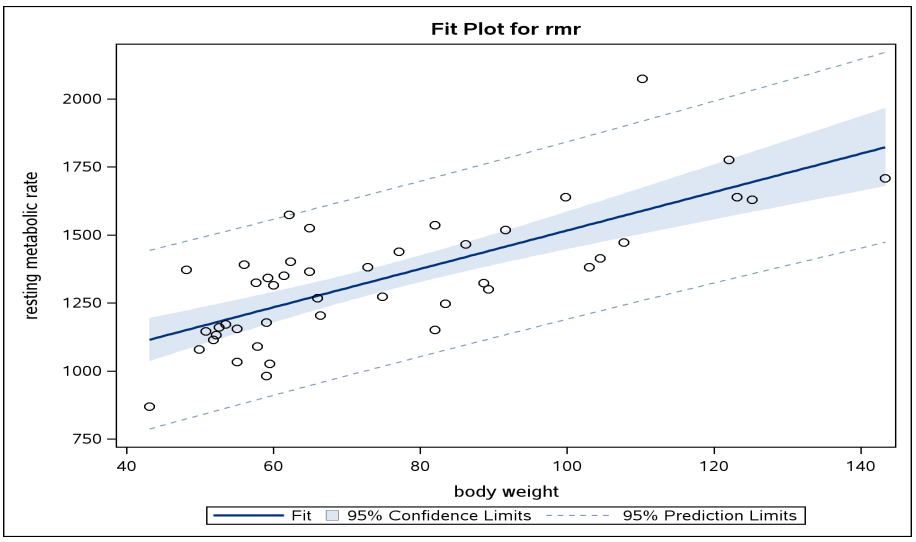
- Medtage flere personer i undersøgelsen:
  Dette ville næppe hjælpe, idet spredningen omkring regressionslinien er et udtryk for den biologiske variation i stofskiftet for kvinder med samme kropsvægt - og denne ændrer sig jo ikke af, at der bliver flere kvinder.
- Fjerne de yderligt liggende observationer:
  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*.

Bemærk, at konfidens- og prediktionsgrænser kan fås på samme (pæne) tegning ved at skrive:

```
ods html;
ods graphics on;
proc glm data=rmrdata;
  model rmr=bw;
```

```
run;
ods graphics off;
ods html close;
```

hvorved man får nedenstående figur (`FitPlot.png`). Hvis man ikke benytter ods-systemet, kan man kun få dem frem enkeltvis ved at benytte symbolsætninger, som beskrevet ovenfor (`i=rlclm95` for konfidensgrænser, og `i=rlcli95` for prediktionsgrænser).



Spørgsmål 5.

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

For at få fuld kontrol over modelkontrollen, anvender vi en `output`-sætning i regressionsopsætningen:

```
proc glm data=rmrdata;
  model rmr=bw / solution clparm;
  output out=ny p=yhat r=resid press=uresid
                student=stresid rstudent=ustresid cookd=cook;
run;
```

idet vi så får dannet et nyt datasæt, hvis første observationer ser således ud:

Obs	bw	rmr	bw70	yhat	resid	uresid	stresid	ustresid	cook
1	49.9	1079	-20.1	1163.50	-84.497	-88.631	-0.54805	-0.54343	0.00735
2	50.8	1146	-19.2	1169.85	-23.851	-24.973	-0.15456	-0.15275	0.00056
3	51.8	1115	-18.2	1176.91	-61.910	-64.702	-0.40081	-0.39677	0.00362
4	52.6	1161	-17.4	1182.56	-21.558	-22.497	-0.13947	-0.13783	0.00042
5	57.6	1325	-12.4	1217.86	107.145	110.935	0.69044	0.68607	0.00843
6	61.4	1351	-8.6	1244.68	106.318	109.572	0.68353	0.67913	0.00715
7	62.3	1402	-7.7	1251.04	150.965	155.440	0.97011	0.96942	0.01395
8	64.9	1365	-5.1	1269.39	95.610	98.217	0.61369	0.60908	0.00513

Relevante modelkontroltegninger kunne nu være:

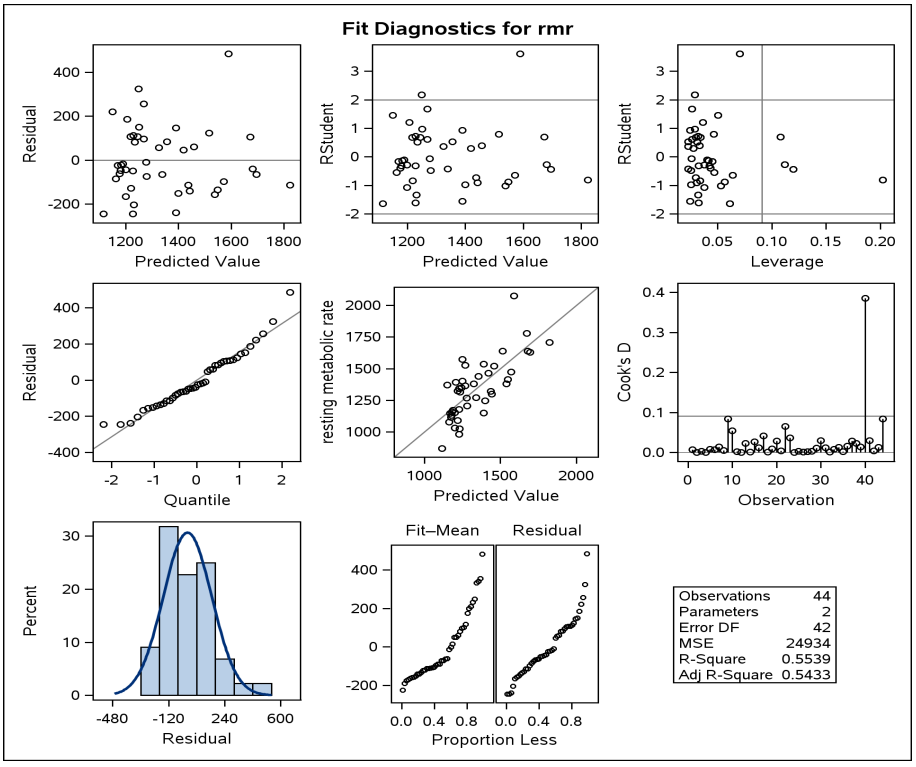
- Scatterplot af residualer (en af de 4 slags) mod predikterede værdier (`yhat`)
- Scatterplot af residualer (en af de 4 slags) mod værdier af den forklarende variabel (`bw`)
- Histogram over residualer
- Fraktildiagram over residualer

Vi kan også lave alle disse (og flere til) på en gang ved at benytte `ods`-systemet

```
ods graphics on;
proc glm PLOTS=(RESIDUALS DIAGNOSTICS) data=rmrdata;
model rmr = bw / solution clparm;
run;
ods graphics off;
```

hvorved vi får dannet et residualplot (residualer mod kovariat)

og en stor samlig diagnostiske figurer



Når vi fokuserer på de 4 ovennævnte plots kan vi konstatere, at der ikke synes at være nogen kritiske afvigelser fra den opstillede model.

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). På figuren ovenfor over Cooks afstand (i midten til højre), ser vi også klart, hvordan denne observation skiller sig ud fra de andre. Vi kan prøve at foretage analysen uden denne observation. Vi udelukker den ved i regressionsanalysekaldet at skrive `where rmr<2000;`, hvorefter regressionsanalysen gentages.

Vi finder da (lettere beskåret output)

```
The GLM Procedure

Number of Observations Read          43
Number of Observations Used          43

Dependent Variable: rmr   resting metabolic rate

                                  Sum of
Source                     DF       Squares   Mean Square  F Value  Pr > F
Model                       1   1001600.370   1001600.370    51.70  <.0001
Error                      41    794347.304     19374.324
Corrected Total            42   1795947.674

R-Square     Coeff Var      Root MSE      rmr Mean
0.557700      10.52276      139.1917      1322.767

                               Standard
Parameter         Estimate           Error    t Value    Pr > |t|
Intercept      852.2483969     68.79650101      12.39      <.0001
bw               6.3533738      0.88363001       7.19      <.0001

Parameter       95% Confidence Limits
Intercept      713.3110445  991.1857493
bw               4.5688468    8.1379008
```

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.

En liste over ændringer i parameterestimater ved udeladelse af hver enkelt observation ser ikke ud til at kunne konstrueres med `proc glm`, men kan kan med `proc reg;`:

```
proc reg data=rmrdata;
  model rmr=bw / clb influence;
run;
```

Herved får vi outputtet (bl.a.)

Dependent Variable: rmr resting metabolic rate

Output Statistics

Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS Intercept	DFBETAS bw
1	-84.4971	-0.5434	0.0466	1.0850	-0.1202	-0.1078	0.0861
2	-23.8507	-0.1528	0.0450	1.0975	-0.0331	-0.0294	0.0233
3	-61.9102	-0.3968	0.0431	1.0883	-0.0843	-0.0740	0.0580
4	-21.5578	-0.1378	0.0418	1.0941	-0.0288	-0.0250	0.0194
5	107.1445	0.6861	0.0342	1.0620	0.1290	0.1036	-0.0747
6	106.3183	0.6791	0.0297	1.0576	0.1188	0.0869	-0.0575
7	150.9647	0.9694	0.0288	1.0326	0.1669	0.1187	-0.0766
8	95.6100	0.6091	0.0265	1.0587	0.1006	0.0651	-0.0381
9	-245.4923	-1.6365	0.0614	0.9852	-0.4187	-0.3948	0.3324
10	221.2100	1.4565	0.0502	0.9988	0.3349	0.3053	-0.2478
11	-47.7340	-0.3056	0.0424	1.0909	-0.0643	-0.0563	0.0438
12	-16.9114	-0.1080	0.0402	1.0928	-0.0221	-0.0190	0.0146
13	-165.5007	-1.0704	0.0379	1.0322	-0.2124	-0.1786	0.1343
14	-44.5007	-0.2842	0.0379	1.0864	-0.0564	-0.0474	0.0357
15	185.4398	1.2027	0.0364	1.0160	0.2337	0.1933	-0.1432
16	-129.2674	-0.8298	0.0339	1.0506	-0.1555	-0.1243	0.0893
17	-245.7388	-1.6119	0.0324	0.9591	-0.2949	-0.2296	0.1611
18	-49.7388	-0.3168	0.0324	1.0792	-0.0580	-0.0451	0.0317
19	112.8493	0.7223	0.0322	1.0572	0.1316	0.1020	-0.0713
20	-204.2686	-1.3265	0.0318	0.9965	-0.2404	-0.1849	0.1284
21	81.2017	0.5179	0.0312	1.0692	0.0930	0.0706	-0.0485
22	324.3766	2.1753	0.0290	0.8686	0.3759	0.2690	-0.1747
23	256.6100	1.6826	0.0265	0.9432	0.2779	0.1797	-0.1054
24	-9.1555	-0.0580	0.0257	1.0769	-0.0094	-0.0058	0.0032
25	-74.9793	-0.4766	0.0255	1.0650	-0.0771	-0.0466	0.0253
26	56.8397	0.3604	0.0229	1.0672	0.0552	0.0215	-0.0047
27	-66.2794	-0.4204	0.0227	1.0646	-0.0641	-0.0200	0.0002
28	83.4837	0.5303	0.0229	1.0594	0.0812	0.0180	0.0074
29	145.8920	0.9341	0.0247	1.0315	0.1486	0.0044	0.0417
30	-239.1080	-1.5592	0.0247	0.9588	-0.2480	-0.0074	-0.0696
31	-151.9913	-0.9745	0.0255	1.0286	-0.1577	0.0035	-0.0521
32	46.2420	0.2937	0.0276	1.0747	0.0495	-0.0060	0.0209
33	-113.7008	-0.7270	0.0299	1.0544	-0.1277	0.0252	-0.0627

```
34 -141.6425   -0.9092   0.0307   1.0402  -0.1618    0.0354  -0.0824
35   61.1206    0.3897   0.0334   1.0777   0.0725   -0.0205   0.0410
36  123.2324    0.7957   0.0465   1.0674   0.1758   -0.0816   0.1257
37 -156.3580   -1.0180   0.0530   1.0542  -0.2409    0.1244  -0.1821
38 -134.9473   -0.8773   0.0564   1.0714  -0.2144    0.1154  -0.1656
39  -98.5378   -0.6405   0.0640   1.0990  -0.1675    0.0971  -0.1345
40  484.8134    3.6128   0.0705   0.6496   0.9953   -0.6046   0.8194  <-----
41  104.5109    0.6964   0.1078   1.1489   0.2421   -0.1702   0.2151
42  -40.2545   -0.2675   0.1119   1.1774  -0.0949    0.0673  -0.0847
43  -65.0796   -0.4350   0.1198   1.1812  -0.1605    0.1158  -0.1445
44 -114.8570   -0.8110   0.2022   1.2741  -0.4082    0.3234  -0.3846


Sum of Residuals                          0
Sum of Squared Residuals            1047231
Predicted Residual SS (PRESS)       1155764
```

Vi ser at observation nr. 40 er i stand til at ændre hældningsestimatet med 0.82 standard errors.

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 $\beta = 2$.

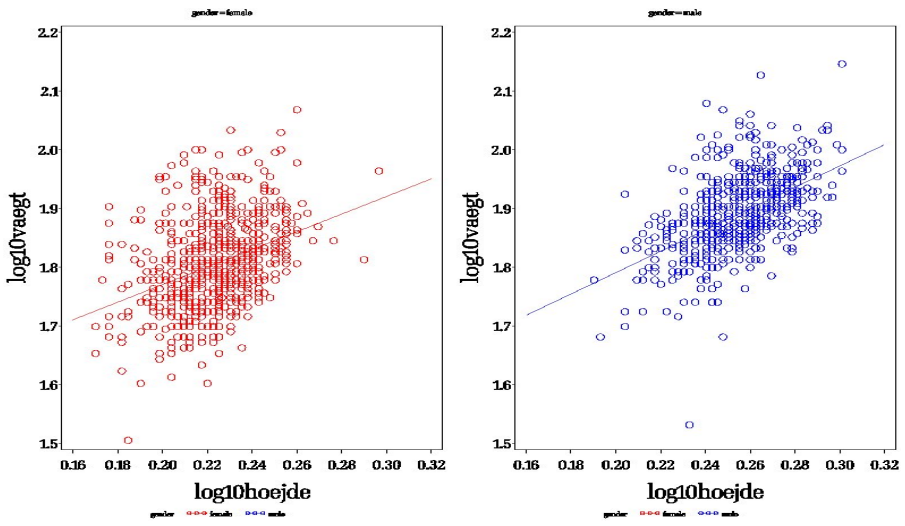
   I de næste spørgsmål skal vi se, om dette ser fornuftigt ud i Sundby-materialet.

2. **Hent nu `sundby` ind fra `sasuser`, hvis den findes der fra tidligere, eller fra filen på hjemmesiden.**

   Det har vi prøvet mange gange før, så ingen kommentarer her.

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.**

```
proc gplot data=sundby gout=plotud1 uniform; by gender;
  plot log10vaegt*log10hoejde=gender
 /  haxis=axis1 vaxis=axis2 frame;
axis1 offset=(3,3) label=(H=3) value=(H=2) minor=NONE;
axis2 offset=(1,1) value=(H=2) minor=NONE
      label=(A=90 R=0 H=3);
symbol1 v=circle c=red i=rl h=2 r=1;
symbol2 v=circle c=blue i=rl h=2 r=1;
run;
```



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 output (her fra `proc reg`, og en del beskåret):

```
gender=female

The REG Procedure
Model: MODEL1
Dependent Variable: log10vaegt

Number of Observations Read                         827
Number of Observations Used                         788
Number of Observations with Missing Values           39

Root MSE              0.06921    R-Square     0.1335
Dependent Mean        1.80546    Adj R-Sq     0.1324
Coeff Var             3.83324

                        Parameter Estimates

                      Parameter       Standard
Variable       DF      Estimate          Error    t Value    Pr > |t|
Intercept       1       1.47039        0.03055      48.13      <.0001
log10hoejde     1       1.50043        0.13636      11.00      <.0001

              Parameter Estimates

Variable       DF       95% Confidence Limits
Intercept       1        1.41041        1.53036
```

```
log10hoejde     1        1.23275        1.76810


gender=male

The REG Procedure
Model: MODEL1
Dependent Variable: log10vaegt

Number of Observations Read                         647
Number of Observations Used                         628
Number of Observations with Missing Values           19

Root MSE              0.05563    R-Square     0.2762
Dependent Mean        1.88997    Adj R-Sq     0.2750
Coeff Var             2.94340

                        Parameter Estimates

                      Parameter       Standard
Variable       DF      Estimate          Error    t Value    Pr > |t|
Intercept       1       1.42756        0.03000      47.58      <.0001
log10hoejde     1       1.81599        0.11750      15.45      <.0001

              Parameter Estimates

Variable       DF       95% Confidence Limits
Intercept       1        1.36864        1.48648
log10hoejde     1        1.58524        2.04674
```

og de interessante størrelser er her:

Køn	intercept $\hat{\alpha} \approx \log(\text{BMI})$	$10^{\text{intercept}}$ $10^{\hat{\alpha}} \approx$ 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:

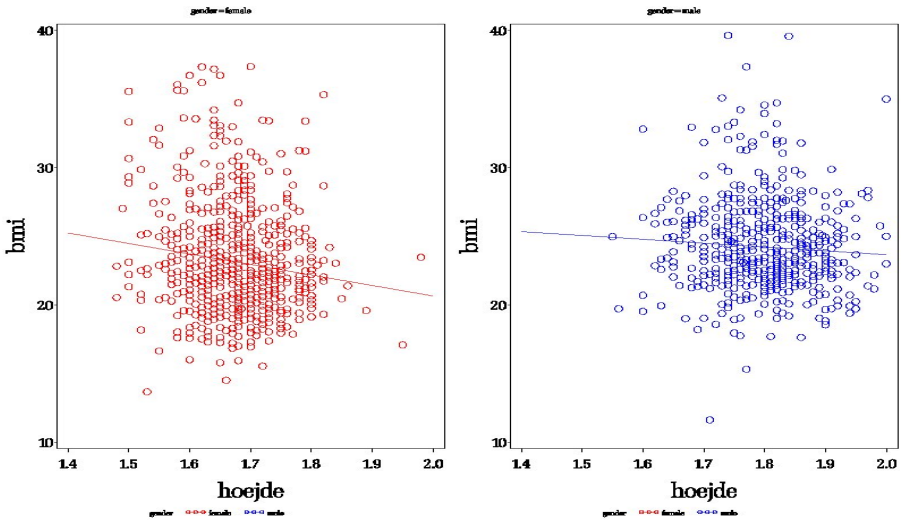
- Tilfældigheder:
  Næppe, da vi har at gøre med et stort datamateriale

- "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).



```
gender=female

The CORR Procedure
   2  Variables:    bmi      hoejde

Spearman Correlation Coefficients
   Prob > |r| under H0: Rho=0
      Number of Observations

                  bmi        hoejde
bmi           1.00000      -0.12245
                            0.0006
                  788           788

hoejde       -0.12245       1.00000
               0.0006
                  788           805


gender=male

The CORR Procedure
   2  Variables:    bmi      hoejde

Spearman Correlation Coefficients
   Prob > |r| under H0: Rho=0
      Number of Observations

                  bmi        hoejde
bmi           1.00000      -0.09579
                            0.0163
                  628           628

hoejde       -0.09579       1.00000
               0.0163
                  628           632
```

Selv om det kan være svært at se på figurerne, 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.