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, enten direkte fra hjemmesiden, eller ved at gemme data på egen computer først. Evt kan man efterfølgende give variablene nogle lables, således at output bliver mere selvforklarende.

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.

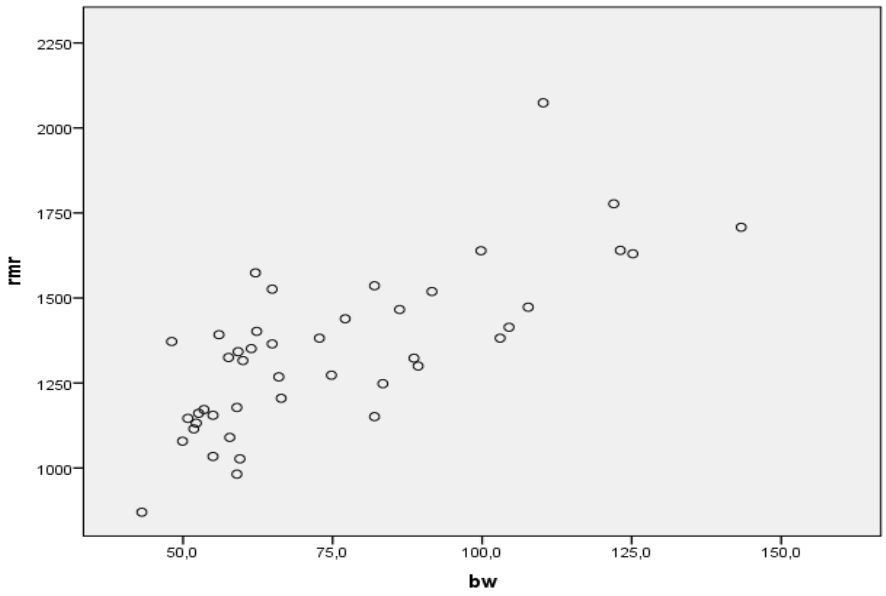
Da opgaven går ud på at konstruere normalområder for stofskiftet, som funktion af kropsvægten, må vi udnævne stofskiftet til at være responsen (Y), medens kropsvægten `bw` er den forklarende variable, altså X.

Vi plotter ved at gå i `Graph/Chart Builder/Scatter` og i den fremkomne boks at trække `rmr` over på Y-aksen, og `bw` over på X-aksen. Koden bliver hermed

```
GGRAPH
/GRAPHDATASET NAME="graphdataset" VARIABLES=bw rmr MISSING=LISTWISE REPOR
TMISSING=NO
/GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
SOURCE: s=userSource(id("graphdataset"))
DATA: bw=col(source(s), name("bw"))
DATA: rmr=col(source(s), name("rmr"))
GUIDE: axis(dim(1), label("bw"))
GUIDE: axis(dim(2), label("rmr"))
```

```
ELEMENT: point(position(bw*rmr))
END GPL.
```

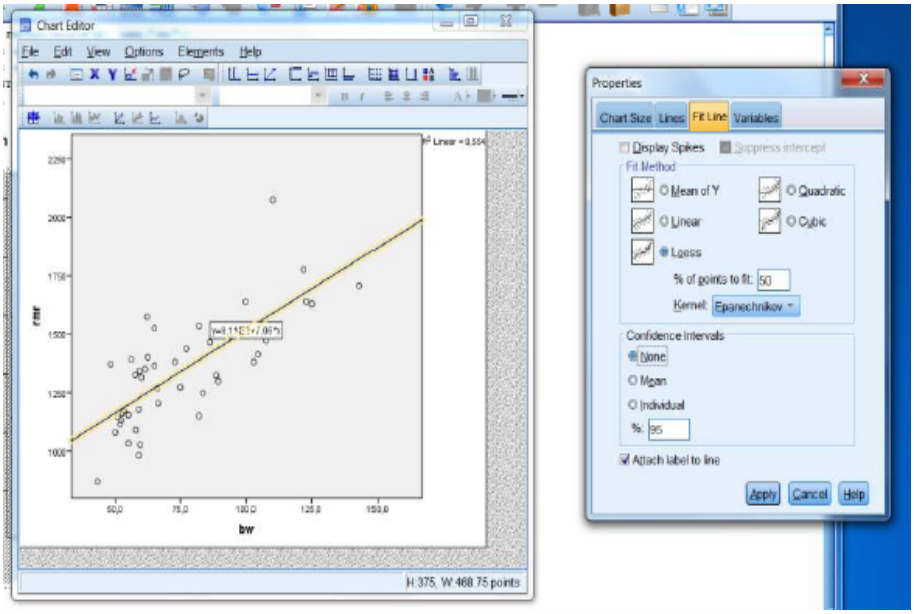
og det fremkomne plot ser således ud:



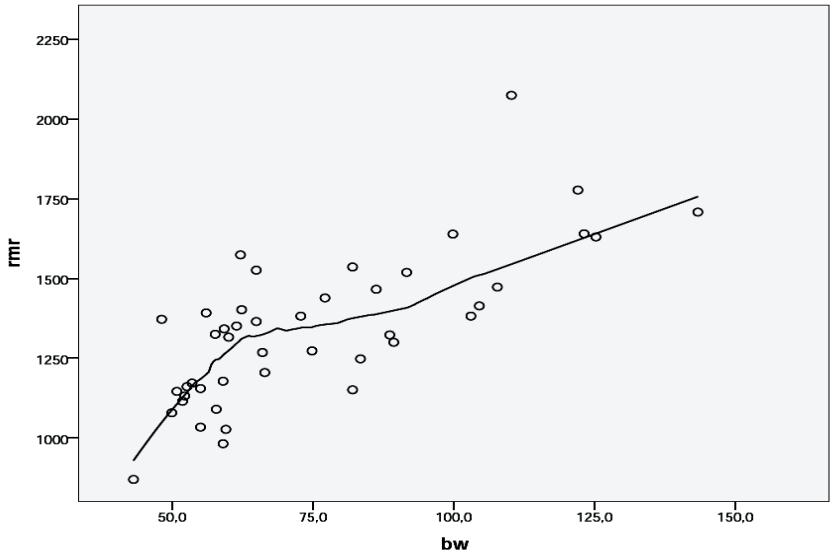
Spørgsmål 2.

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 dobbeltklikke på grafen, klikke på ikonet `Add Fit Line at Total` og derefter i `Properties`-boksen afkrydse `Loess/Apply`:



og herved får vi

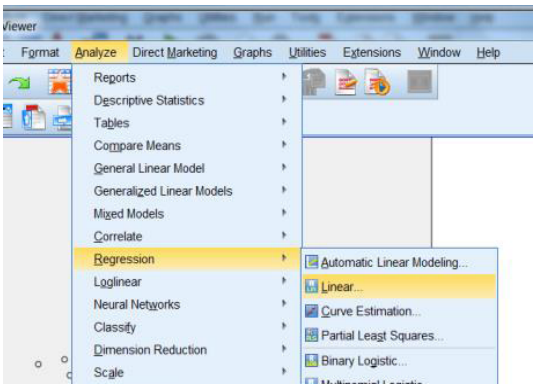


Baseret på denne figur og det faktum, at vi har ret få observationer i det høje område, vil vi fortsætte uden transformation.

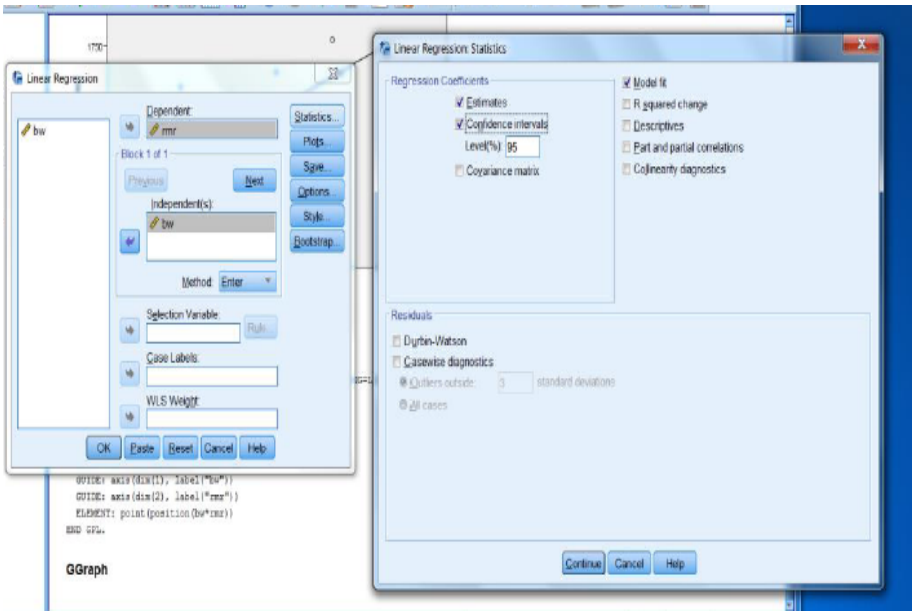
Vi vil derfor foretage en sædvanlig lineær regression med `rmr` (**Y**) som respons og `bw` (**X**) som forklarende variabel, uden at transformere nogen af dem.

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

I SPSS gør vi dette ved at gå ind i menuen `Analyze/Regression/Linear`, og i boksen sætte `rmr` som `Dependent` og `bw` som `Independent(s)` (et uheldigt navn til forklarende variable...)



Vi skal også huske at gå ind i `Statistics` og afkrydse `Parameter Estimates` og `Confidence intervals`



Den tilhørende kode er

```
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS CI(95) R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT rmr
/METHOD=ENTER bw.
```

og vi får outputtet

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,744[a]	,554	,543	157,905

a. Predictors: (Constant), bw

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1300241,179	1	1300241,179	52,147	,000[b]
	Residual	1047230,708	42	24934,064		
	Total	2347471,886	43			

a. Dependent Variable: rmr

b. Predictors: (Constant), bw

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	811,227	76,976		10,539	,000
	bw	7,060	,978	,744	7,221	,000

Coefficients[a]

Model		95,0% Confidence Interval for B Lower Bound	95,0% Confidence Interval for B Upper Bound
1	(Constant)	655,884	966,570
	bw	5,087	9,032

a. Dependent Variable: rmr

Vi ser af ovenstående output, at effekten af `bw` er stærkt signifikant, idet et test af hældningen $\beta = 0$ giver T=7.221, svarende til en P-værdi, der angives

som 0.000. Samme P-værdi får man (naturligvis) ved at anvende F-testet, idet F=7.221^2=52.147$\sim$ 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
811.227 (76.976)	7.060 (0.978)	157.905

En simpel **aflæsning** angiver standard error på hældningen til 0.978, altså s.e.($\hat{\beta}$)=0.978. 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)

$$7.060 \pm 2.018 \times 0.978 = (5.086, 9.034)$$

men det behøver vi jo slet ikke, for det kommer med direkte i outputtet, fordi vi har afkrydset `Confidence intervals`.

Fortolkningen af dette interval (5.087, 9.032) 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 (aflæses under
`Std Error of the Estimate`, under `Model Summary`) til 157.9, som også angivet i den lille tabel ovenfor.

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

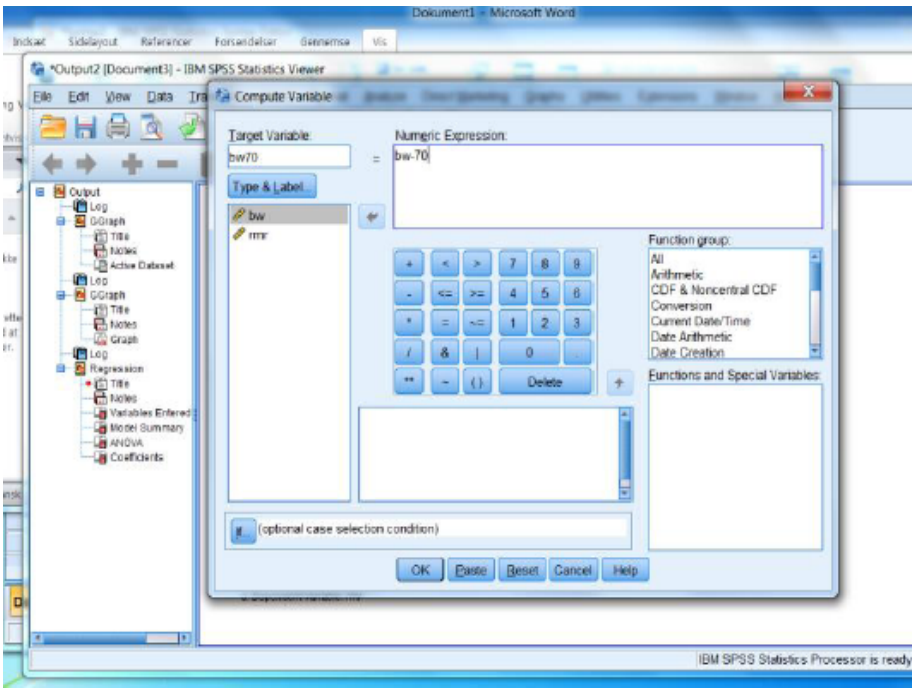
Spørgsmål 3.

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

$$811.227 + 7.060 \times 70 = 1305.43$$

Vi kan også lade SPSS finde dette estimat ved at snyde den til at tro, at det er interceptet. Dette gør vi ved at flytte nulpunktet hen i 70 kg, dvs. ved at benytte en ny X-variabel, der er kropsvægt minus 70. Vi definerer altså `bw70=bw-70` ved at gå ind i `Transform/Compute`, og herefter gentages regressionen med denne nye X-variabel.



Den tilhørende kode er

```
COMPUTE bw70=bw-70.
EXECUTE.
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS CI(95) R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT rmr
/METHOD=ENTER bw70.
```

og det nye output (beskåret, således at kun de dele, der har ændret sig, er vist nedenfor) bliver:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	1305,394	24,278		53,768	,000
	bw70	7,060	,978	,744	7,221	,000

Coefficients[a]

Model		95,0% Confidence Interval for B	
		Lower Bound	Upper Bound
1	(Constant)	1256,398	1354,389
	bw70	5,087	9,032

a. Dependent Variable: rmr

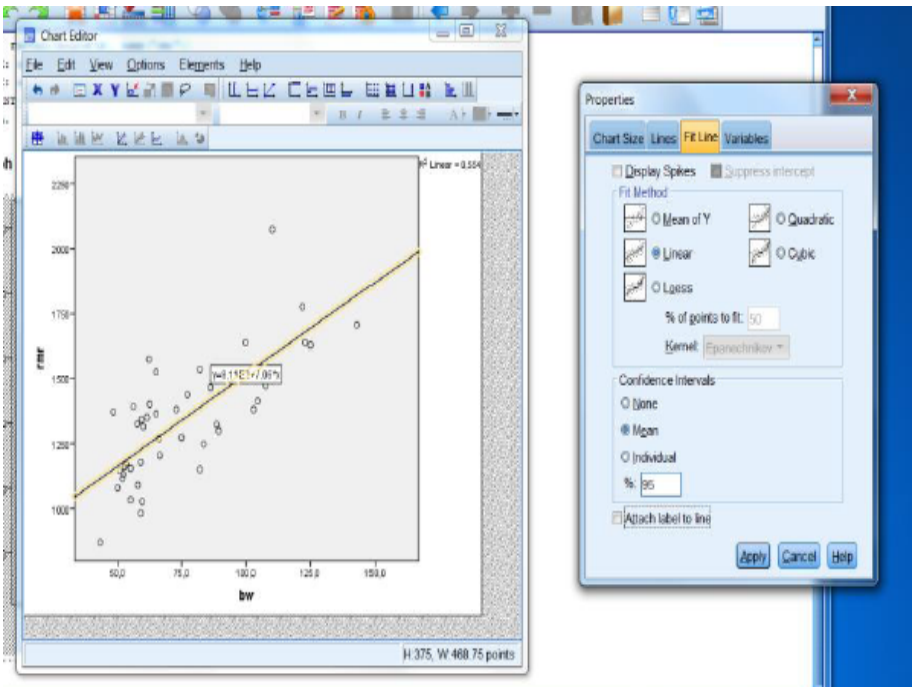
Svaret er altså en forventet værdi på 1305.394, med et konfidensinterval på $(1256.398, 1354.389) \approx (1256, 1354)$.

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.

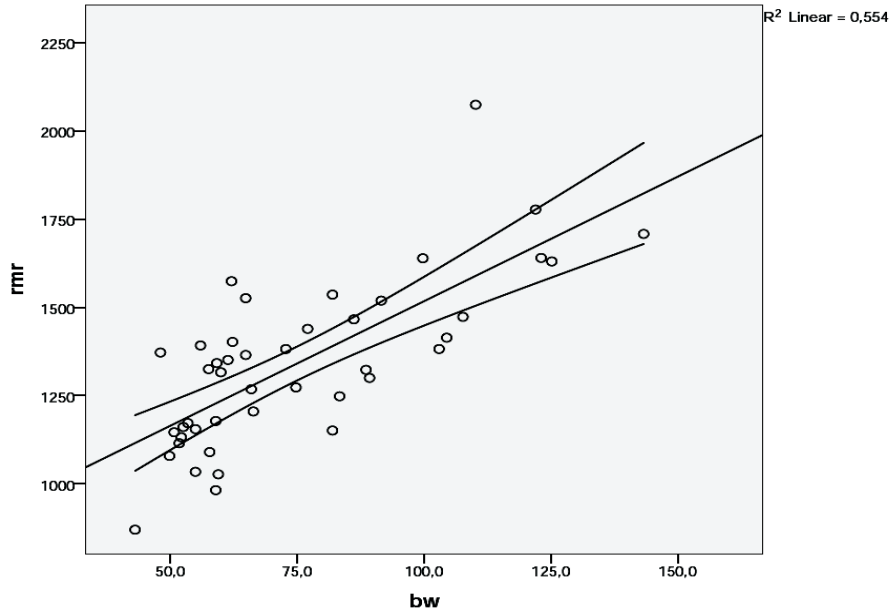
Bemærk, at det omvendte ikke gælder:
Selv om konfidensintervallerne overlapper, kan der godt være signifikans!

Dette spørgsmål kunne også være besvaret ved blot at se på en figur med indtegnede konfidensgrænser, der kan fås ved at dobbeltklikke på scatterplottet fra spørgsmål 1, og benytte `Add Fit Line at Total` og derefter i `Properties`-boksen afkrydse `Linear` og `Confidence Intervals/Mean`.

For at undgå boksen med regressionslinien hen over figuren, kan man evt. fjerne fluebenet i `Attach label to line`:



Herved får vi 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.

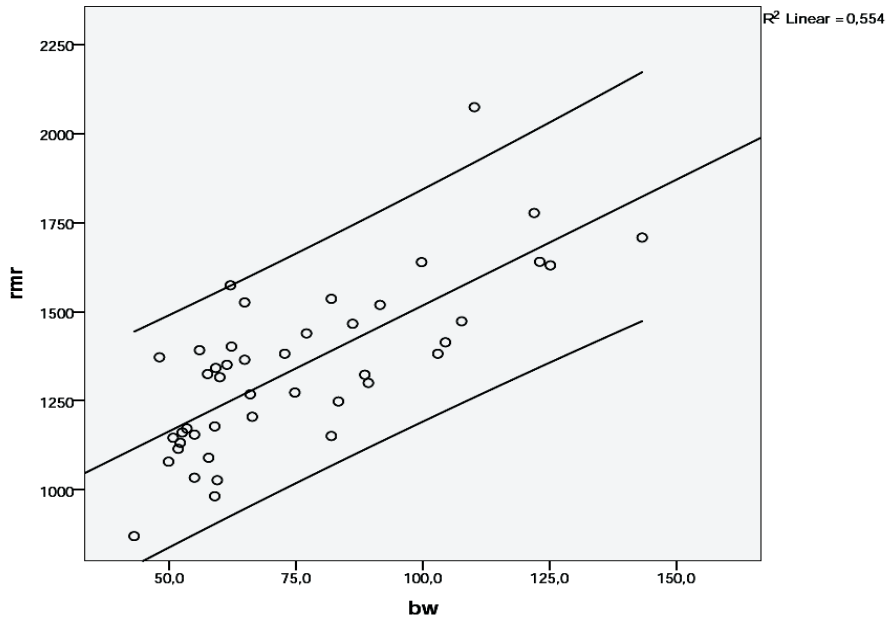
- **Hvad er dit bedste gæt på hendes hvilestofskifte (i mangel af yderligere information)? Og indenfor hvilke grænser vil du med 95% sikkerhed tippe hendes hvilestofskifte at ligge, forudsat at hun er rask.**

  Ligesom ovenfor kan vi udfra estimaterne udregne det forventede hvilestofskifte til

  $$811.227 + 7.060 \times 80 = 1376.03$$

  Et 95% prediktionsinterval kan (approksimativt) fås ved at lægge $\pm 2 \times 157.9$ til omkring estimatet, 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 afkrydse `Linear` og `Confidence Intervals/Individual`, og igen fjerne fluebenet i `Attach label to line`)



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.

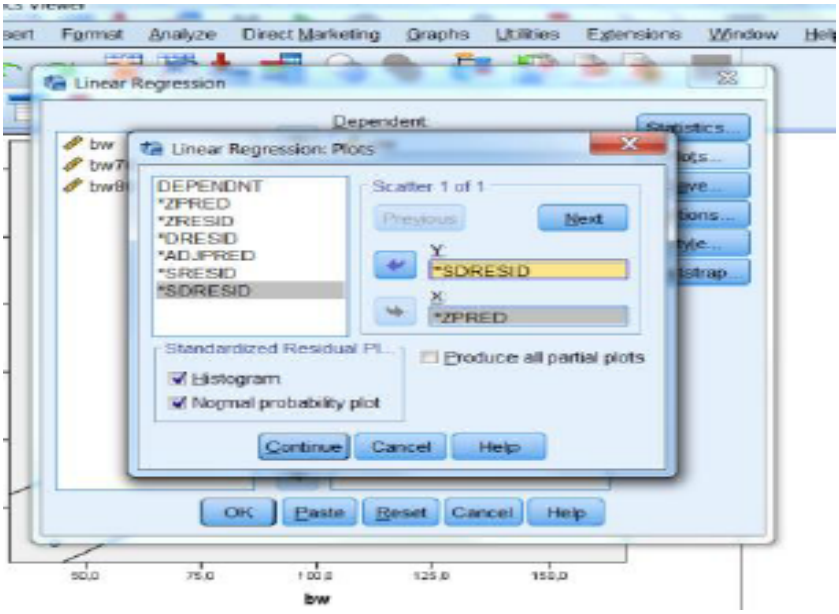
Man kunne så overveje, hvordan man kunne gøre disse grænser smallere, altså hvordan man kan nedbringe variationen omkring regressionslinien. Der kunne tænkes flere muligheder:

- 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.*

Spørgsmål 5.

Vurder om modellens forudsætninger kan siges at være opfyldt. Suppler evt. med teoretiske overvejelser omkring stofskifte.
Til dette spørgsmål (samt til det følgende) kan man med fordel gemme informationer fra regressionsanalysen ved at benytte `Save`-knappen i regressions-boksen, og der afkrydse de ekstra informationer, man ønsker gemt.

Nogle enkelte modelkontroltegninger kan fås frem direkte fra regressionsanalysen, ved at klikke på `Plots`

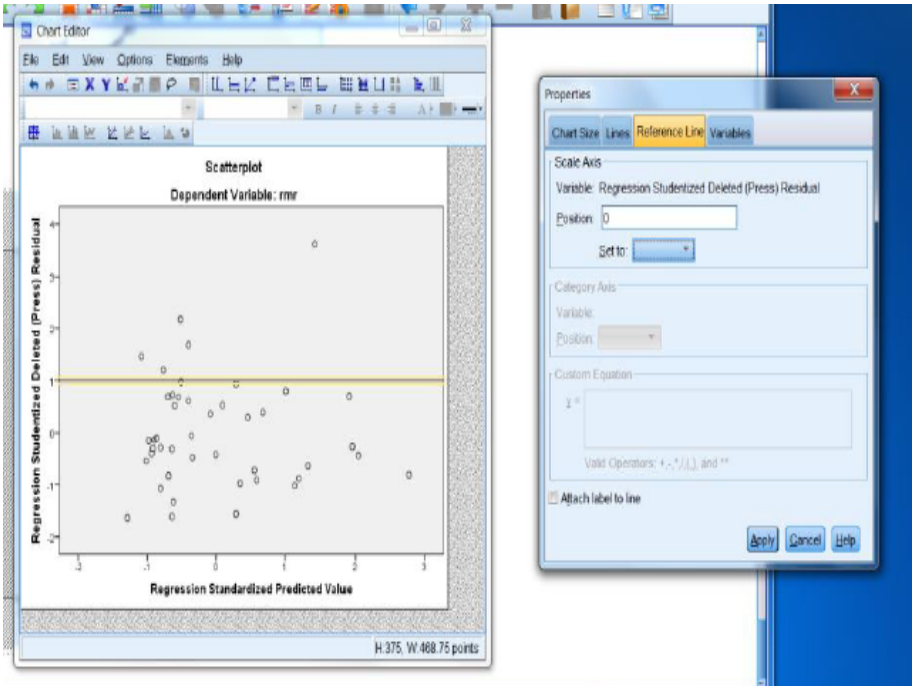


Betegnelserne er her:

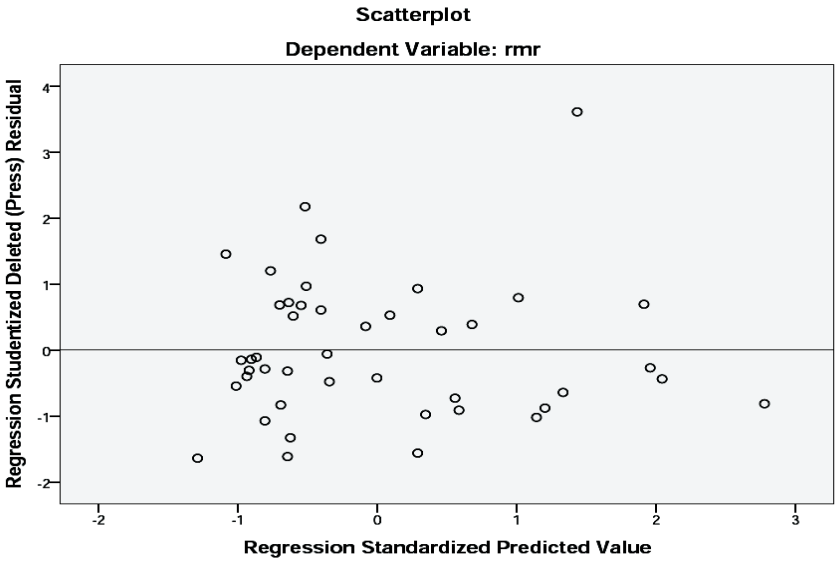
```
DEPENDNT: the dependent variable
*ZPRED Standardized predicted values
*ADJPRED Adjusted predicted values
*ZRESID Standardized residuals
*DRESID Deleted residuals
*SRESID Studentized residuals
*SDRESID Studentized deleted residuals
```

men hvad betegnelserne præcist dækker over, kan variere fra program til program.

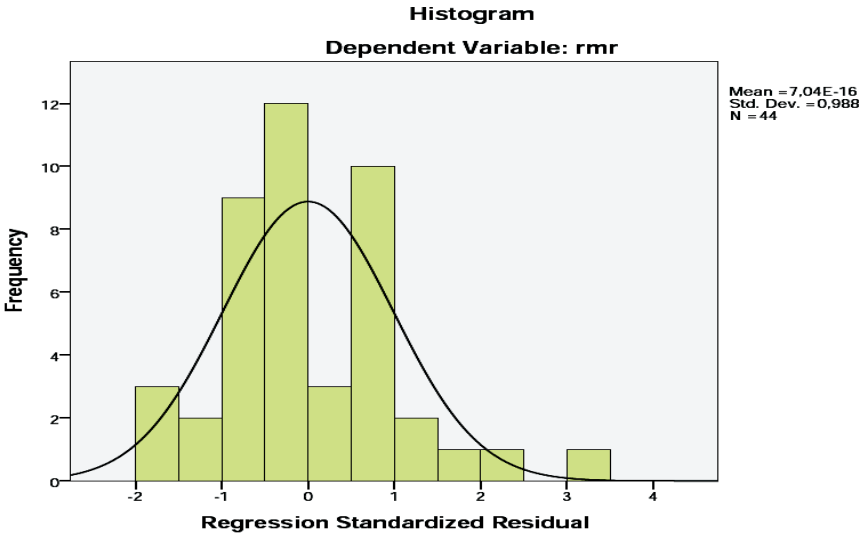
Afkrydsningerne ovenfor vil give et plot af residualer (student deleted), plottet mod fittede (=forventede=predikterede, godt nok standardiserede) værdier. På denne har vi efterfølgende indlagt en vandret linie i 0 ved at dobbeltklikke på grafen og klikke
`Add a reference line to the Y axis`, hvorefter der fremkommer en `Properties`-boks, hvor man skriver `0` i `Position`.

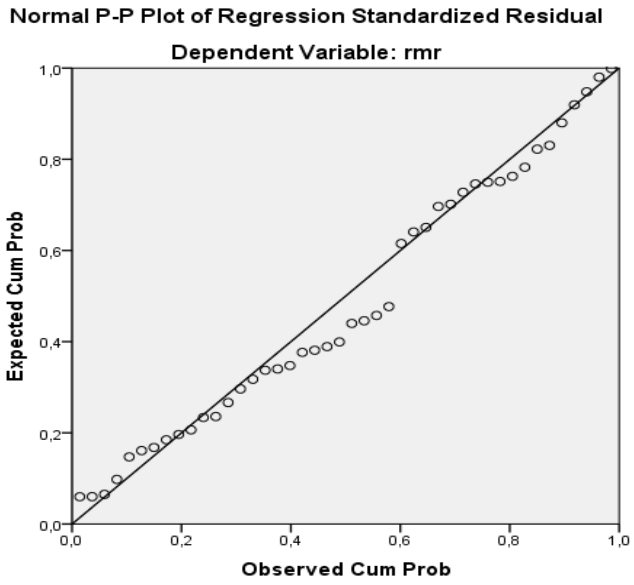


Hermed ser grafen således ud:



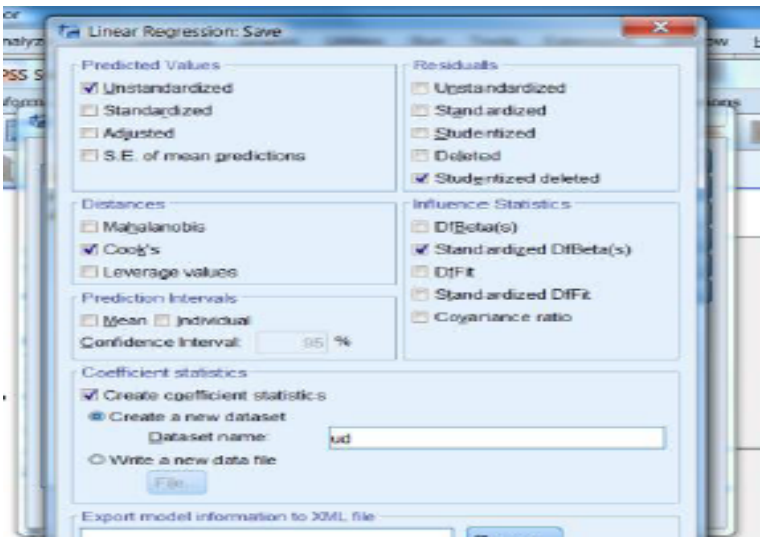
Endvidere fremkommer histogram (med overlejret normalfordelingskurve), samt et såkaldt PP-plot:



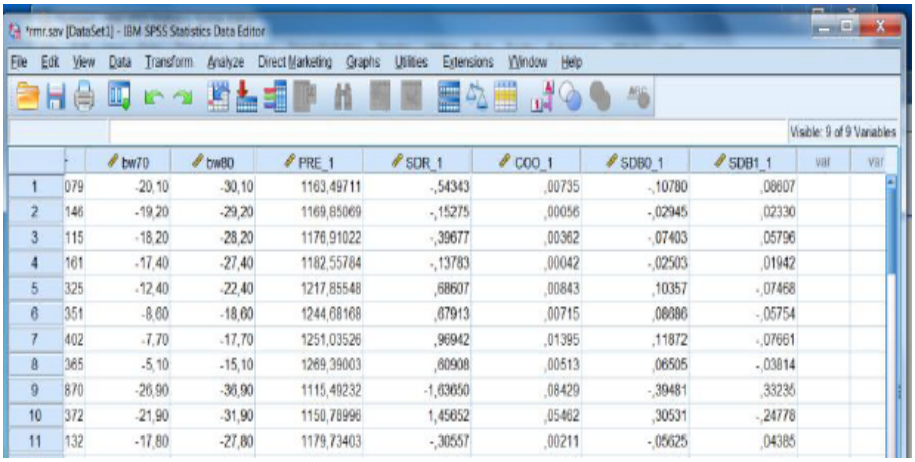


Det sidste af disse plot, det såkaldte PP-plot, er dog ikke så velegnet til at afsløre afvigelser fra normalitet af residualerne som det tilsvarende QQ-plot, og desuden mangler vi plot af Cooks afstand til vurdering af indflydelsesrige observationer.

For at få fuld kontrol over modelkontrollen, anvender vi derfor `Save`-knappen, og krydser det af, vi gerne vil bruge:



idet vi så får tilføjet disse størrelser til vores datasæt:



		bw70	bw80	PRE_1	SDR_1	COO_1	SDB0_1	SDB1_1	var	var
1	079	-20,10	-30,10	1163,49711	-,54343	,00735	-,10780	,08607		
2	146	-19,20	-29,20	1169,85069	-,15275	,00056	-,02945	,02330		
3	115	-18,20	-28,20	1176,91022	-,39677	,00362	-,07403	,05796		
4	161	-17,40	-27,40	1182,55784	-,13783	,00042	-,02503	,01942		
5	325	-12,40	-22,40	1217,85548	,68607	,00843	,10357	-,07468		
6	351	-8,60	-18,60	1244,68168	,67913	,00715	,08086	-,05754		
7	402	-7,70	-17,70	1251,03526	,96942	,01395	,11872	-,07661		
8	365	-5,10	-15,10	1269,39003	,60908	,00513	,06505	-,03814		
9	870	-26,90	-36,90	1115,49232	-1,63650	,08429	-,39481	,33236		
10	372	-21,90	-31,90	1150,78996	1,45652	,05462	,30531	-,24778		
11	132	-17,80	-27,80	1179,73403	-,30557	,00211	-,05625	,04385		

med betegnelserne

Variables Created or Modified	PRE_1	Unstandardized Predicted Value
	SDR_1	Studentized Deleted Residual
	COO_1	Cook's Distance
	SDB0_1	Standardized DFBETA for (Constant)
	SDB1_1	Standardized DFBETA for bw

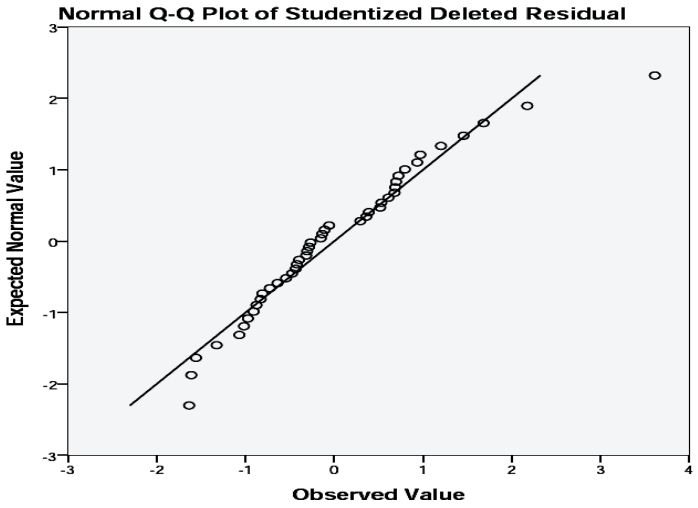
Relevante modelkontroltegninger kunne nu være:

- Scatterplot af residualer mod predikterede værdier:
  Dette har vi set ovenfor (godt nok med standardiserede predikterede værdier, dvs. med en X-akse, der ikke viser de faktisk predikterede værdier), og der synes ikke at være noget mønster, f.eks. i form af trompetfacon, og derfor må vi sige, at der ser ud til at være rimelig varianshomogenitet.

- Scatterplot af residualer mod værdier af den forklarende variabel (`bw`): Dette vil være identisk med det ovenstående, pånær X-aksen. Imidlertid skal vi her se efter buer som tegn på afvigelse fra lineariteten. Det synes der ikke at være.

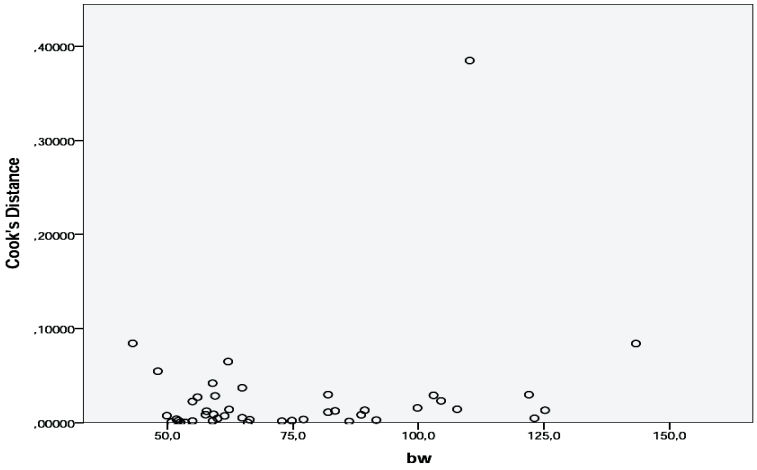
- Histogram over residualer:
  Dette har vi set ovenfor og må kontstatere, at det *ikke* er smukt. Med kun 44 observationer, er det dog svært at vurdere, og man vil i stedet foretrække et fraktildiagram (QQ-plot).

- Fraktildiagram over residualer (QQ-plot):
  Dette har vi endnu ikke set, men vi har gemt residualerne under navnet `SDR_1` og kan nu konstruere det ved at gå ind i
  `Analyze > Descriptive Statistics > Q-Q` og sætte residualerne over i `Variables`, hvorved man får



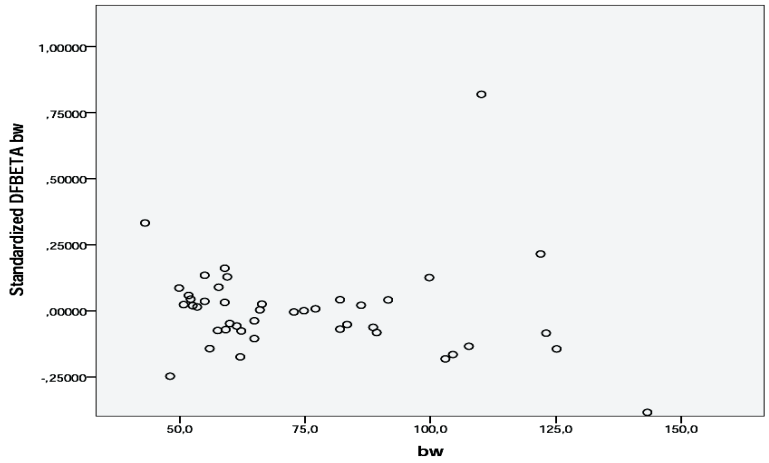
  Dette tyder ikke på nogen afvigelse af betydning fra normalfordelingen.

- Plot af Cooks afstand, f.eks. mod kropsvægten, `bw`, for at vurdere, om der er enkeltobservationer med stor indflydelse på resultaterne (brug blot `Graph/Chart builder/Scatter` som tidligere):



Tommelfingerreglerne mht Cooks afstand siger, at de er store, når de overstiger $\frac{4}{n} = \frac{4}{44} \approx 0.09$.

En figur over over ændringer i parameterestimater ved udeladelse af hver enkelt observation kan vi få ved at benytte den nydannede variabel `SDB1_1`:



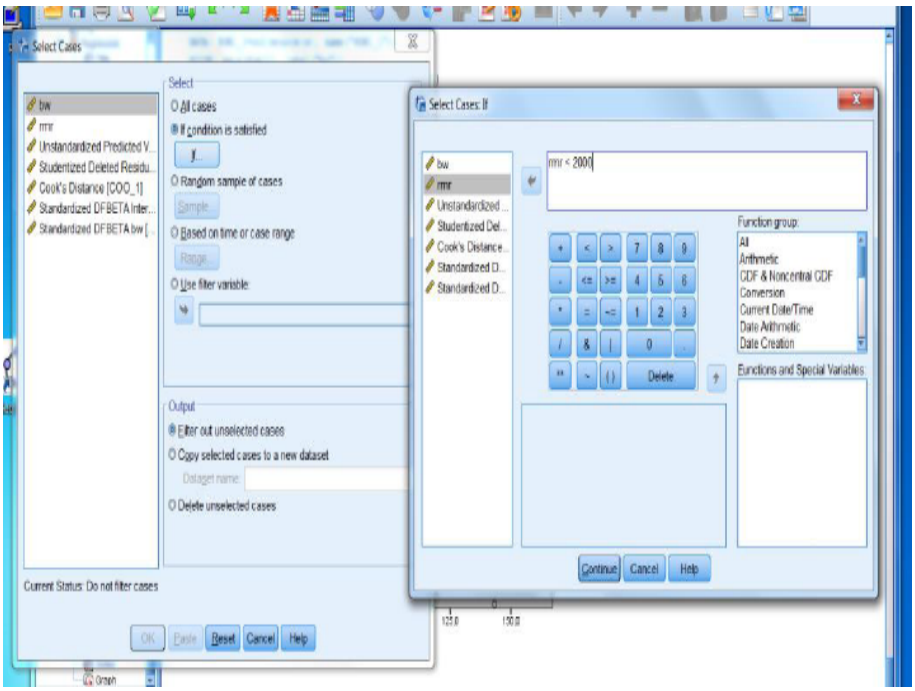
Tommelfingerreglerne her siger, at de er store, når de overstiger $\frac{2}{\sqrt{n}} = \frac{2}{\sqrt{44}} \approx 0.3$.

Begge disse plots viser tydeligt en enkelt observation (nr. 40) med stor indflydelse, se næste spørgsmål. Denne formår at ændre hældningsestimatet med 0.82 standard errors.

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 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 at gå ind i `Data/Select Cases`, afkrydse `If condition is satisfied` og så skrive `rmr < 2000`:



hvorefter regressionsanalysen gentages, og vi får

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,747[a]	,558	,547	139,192

a. Predictors: (Constant), bw

b. Dependent Variable: rmr

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1001600,370	1	1001600,370	51,697	,000[b]
	Residual	794347,304	41	19374,325		
	Total	1795947,674	42			

a. Dependent Variable: rmr

b. Predictors: (Constant), bw

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	852,248	68,797		12,388	,000
	bw	6,353	,884	,747	7,190	,000

Coefficients[a]

Model		95,0% Confidence Interval for B	
		Lower Bound	Upper Bound
1	(Constant)	713,311	991,186
	bw	4,569	8,138

a. Dependent Variable: rmr

Vi ser, at hældningsestimatet ændrede sig fra 7.06 (0.98) til 6.35 (0.88), hvil-

ket 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 $\beta = 2$.

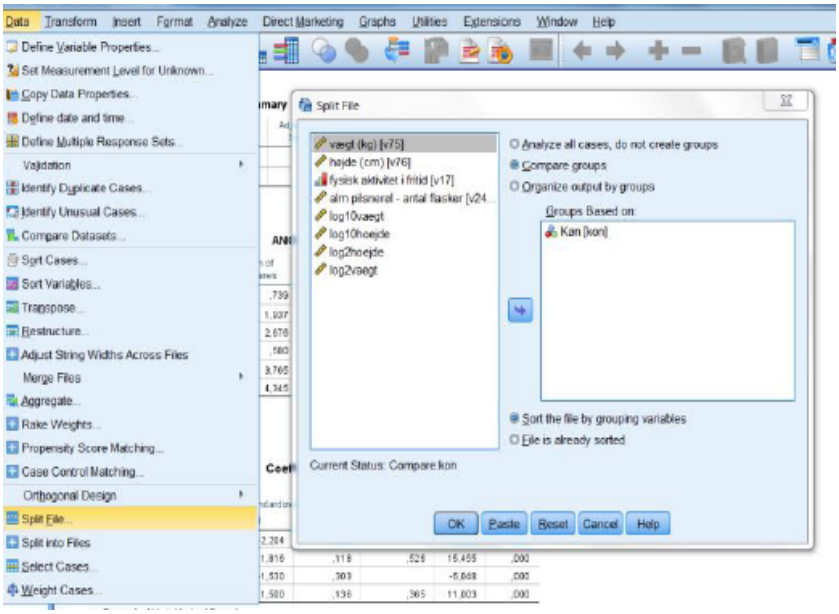
I de næste spørgsmål skal vi se, om dette ser fornuftigt ud i Sundby-materialet.

2. **Hent nu `sundby`-data ind (denne har I brugt tidligere), og 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.**

   Når man skal benytte logaritmetransformerede data, er der mange logaritmefunktioner at vælge imellem, men de vil alle give det samme resultat, når man transformerer resultaterne tilbage til den oprindelige skala.

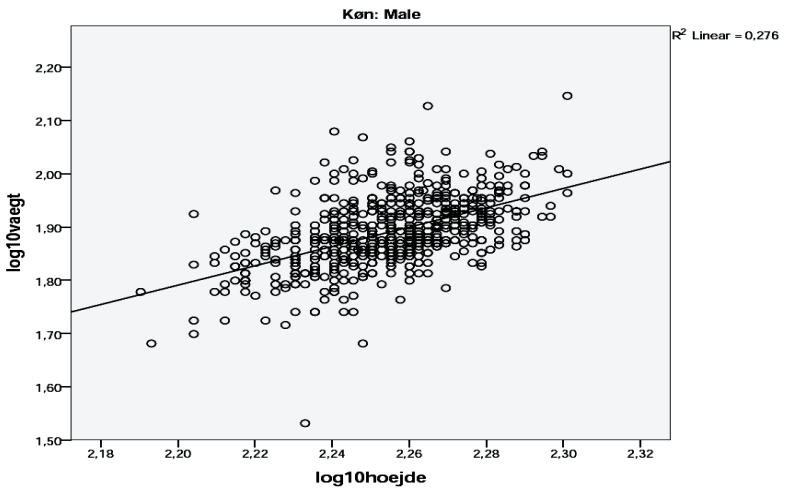
   Her har vi benyttet både 2-tals logaritmer og 10-tals logaritmer, blot for at vise, hvordan disse defineres. 10-tals logaritmen hedder i SPSS `lg10`, den naturlige logaritme hedder `ln`, medens 2-tals-logaritmen faktisk ikke har et selvstændigt navn, så der er man nødt til at benytte det lidt mere indviklede `ln(x)/ln(2)`. Man går altså ind i `Transform/Compute`, sætter f.eks. `log2vaegt` i `Target Variable` og `ln(v75)/ln(2)` i `Numeric Expression`.

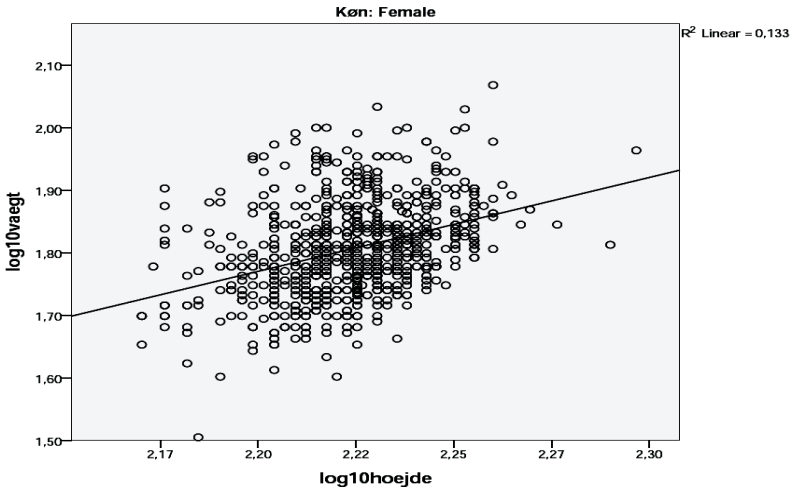
   Når vi skal plotte for hver køn for sig, skal vi først gå ind i `Data/Split File`, sætte flueben ved `Compare Groups` og sætte `kon` over i `Groups Based on`:



Herefter kan vi lave plots af logaritmetransformeret vægt (her `log10vaegt`) mod logaritmetransformeret hoejde (ligeledes `log10hoejde`) ved at gå ind i `Graph/Chart Builder/Scatter`, i den fremkomne boks trække `log10vaegt` over på Y-aksen, og `log10hoejde` over på X-aksen. Man kan få en regressionslinie med ved efterfølgende at dobbeltklikke på scatterplottet og benytte `Add Fit Line at Total`.

Herved får vi to plots:





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 3 (ikke vist). Vi kan derfor ikke afvise lineariteten som en fornuftig beskrivelse. De to øvrige antagelser synes visuelt ikke at give nogen problemer.

3. **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?**

Vi bibeholder vores `Split File`, da vi stadig skal se på hvert køn for sig.

For at fitte en regressionslinie, går vi ind i menuen `Analyze/Regression/Linear`, og i boksen sættes `log10vaegt` som `Dependent` og `log10hoejde` som `Independent(s)`. Desuden klikker vi `Statistics` og afkrydser `Parameter Estimates` og `Confidence intervals`, hvorved vi får outputtet:

Coefficients[a]

Køn	Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t
Male	1	(Constant)	-2,204	,265		-8,321
		log10hoejde	1,816	,118	,526	15,455
Female	1	(Constant)	-1,530	,303		-5,048
		log10hoejde	1,500	,136	,365	11,003

Coefficients[a]

Køn	Model		Sig.	95,0% Confidence Interval for B Lower Bound	95,0% Confidence Interval for B Upper Bound
Male	1	(Constant)	,000	-2,725	-1,684
		log10hoejde	,000	1,585	2,047
Female	1	(Constant)	,000	-2,126	-,935
		log10hoejde	,000	1,233	1,768

a. Dependent Variable: log10vaegt

De interessante størrelser er her:

Køn	hældning "ca. 2 ??"
Mænd	1.816 (1.585, 2.047)
Kvinder	1.500 (1.233, 1.768)

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.

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

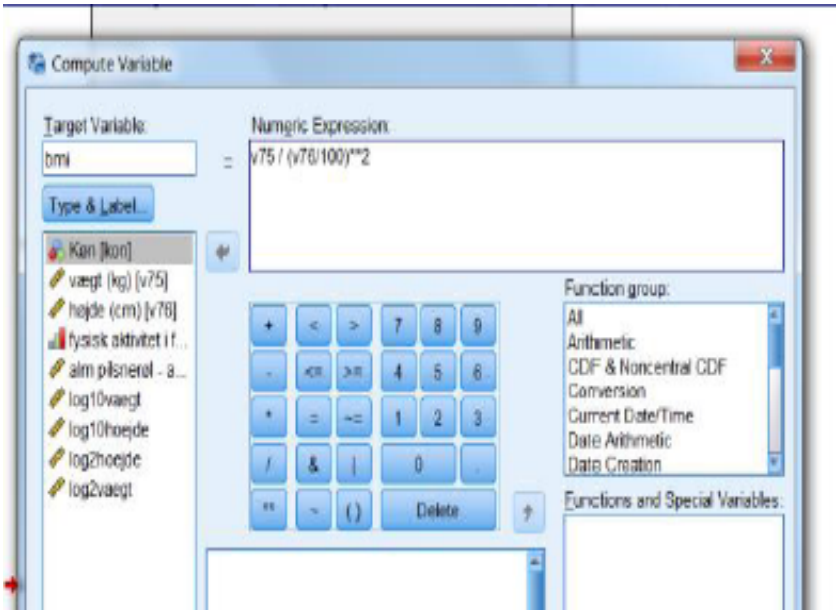
   Der *er* jo faktisk en afvigelse fra den forventede hældning på 2, i hvert fald for kvindernes vedkommende. Der kunne være flere forklaringer på dette:

   - 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.

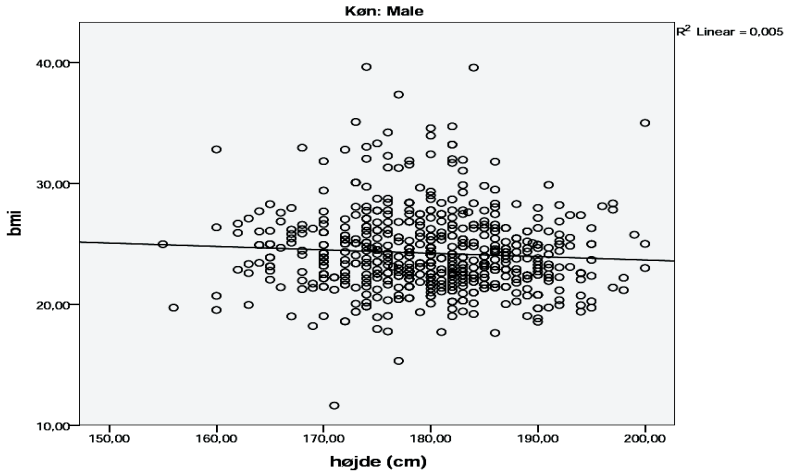
     Man kan korrigere for en sådan målefejl i kovariaten, ved at dividere hældningen med den såkaldte *reliability coefficient*, der afspejler forholdet mellem målefejl og variation i samplet. Da denne koefficient formentlig her vil være meget tæt på 1, kan målefejlen således ikke være forklaringen på en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

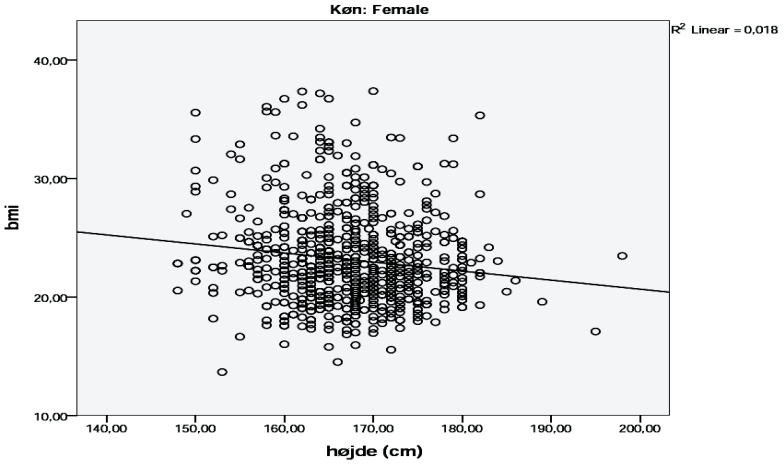
     Vi skal lige have udregnet BMI, og her skal vi huske, at formlen kræver, at højden er angivet i meter. Det er den ikke her, så derfor må vi ved definitionen (se billedet nedenfor) dividere denne med 100.

     Vi er gået ind i `Transform/Compute` og udregner altså `bmi` som `v75/(v76/100)**2`:



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).





Korrelationer udregnes ved at benytte `Analyze/Correlate/Bivariate` og afkrydse `Spearman`, hvis man også vil have denne med
Herved får vi

Correlations

	Køn			højde (cm)	bmi
Spearman's rho	Male	højde (cm)	Correlation Coefficient	1,000	-,096*
			Sig. (2-tailed)	.	,016
			N	632	628
		bmi	Correlation Coefficient	-,096*	1,000
			Sig. (2-tailed)	,016	.
			N	628	628
	Female	højde (cm)	Correlation Coefficient	1,000	-,122**
			Sig. (2-tailed)	.	,001
			N	805	788
		bmi	Correlation Coefficient	-,122**	1,000
			Sig. (2-tailed)	,001	.
			N	788	788

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

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 ($P = 0.016$ hhv $P = 0.001$). 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.