Lineær regression og residualer

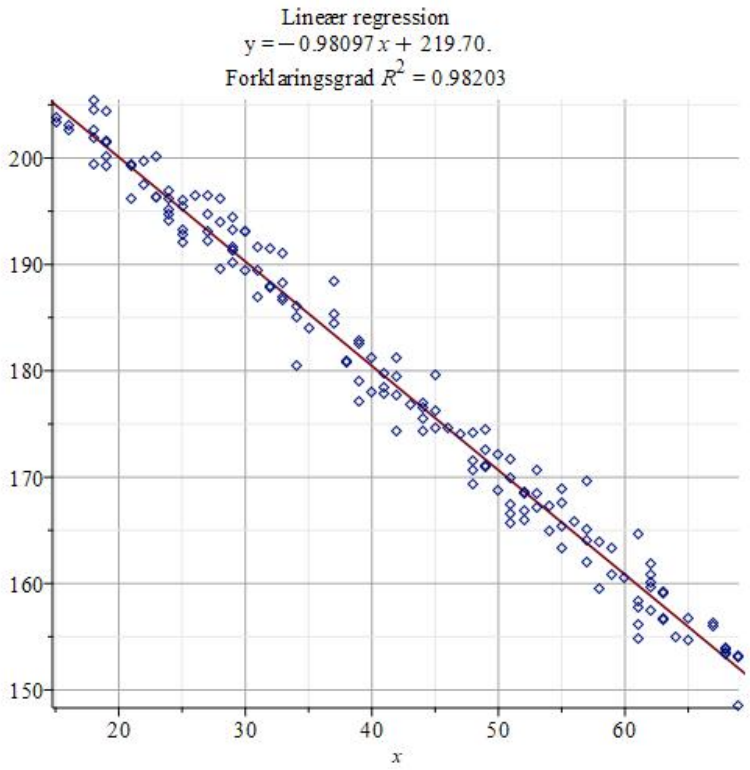
Kvalitet af lineær model

Vi har tidligere set på residualplots for at afgøre kvaliteten af en lineær model på et datasæt. Vi har nu et bedre værktøj til at afgøre, hvor god en regressionsmodel passer på et datasæt - fraktilplottet.

Eksempel 1.1. 155 personer har målt deres makspuls. Sammenhængen mellem deres alder og makspuls fremgår af dette datasæt. Vi laver lineær regression i Maple som vi plejer:

```
with(Gym):
alder := [...]
puls := [...]
LinReg(alder,puls)
```

Resultatet af dette kan ses på Fig. 1.

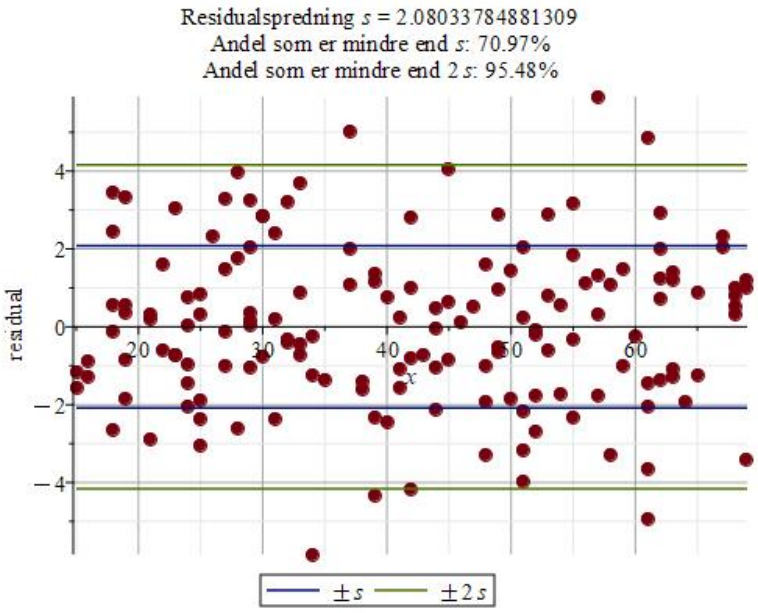


Figur 1: Lineær regression på pulsdata i Maple

Af Fig. 1 ser regressionen ud til at passe godt til datasættet, men vi fortsætter vores regressionsanalyse. Vi laver først et residualplot som vi plejer. Dette gøres i Maple ved at skrive

```
with(Gym):
alder := [...]
puls := [...]
plotResidualer(alder,puls,LinReg)
```

Resultatet af dette kan ses på Fig. 2

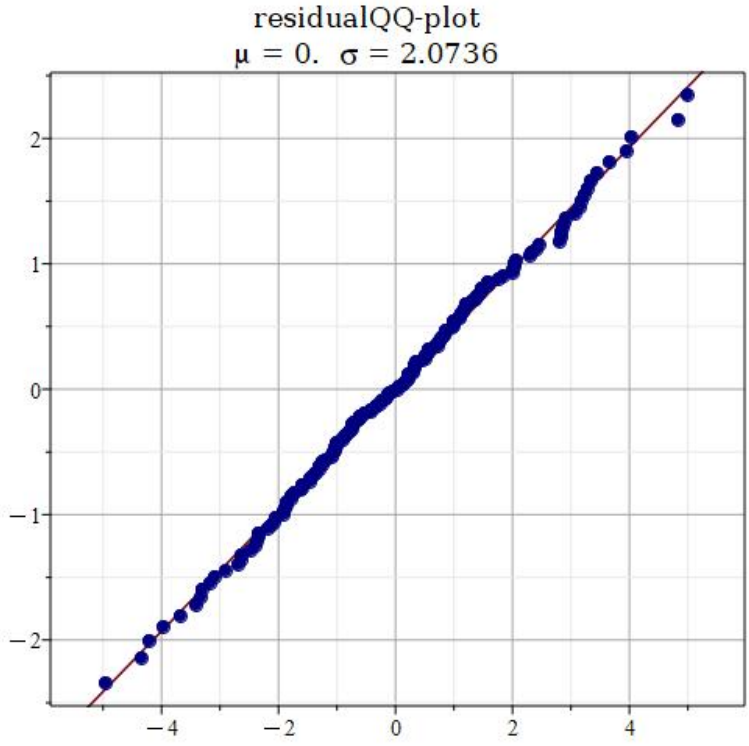


Figur 2: Residualplot for lineær regression på pulsdata i Maple

Igen ser resultatet fint ud og residualerne ser ud til at være normalfordelte. Vi afslutter med at lave et fraktilplot af residualerne. Dette gøres i Maple som

```
with(Gym):
alder := [...]
puls := [...]
residualQQplot(alder,puls,LinReg)
```

Resultatet af dette kan ses på Fig. 3



Figur 3: Fraktilplot for residualerne af pulsdata i Maple

Residualerne ser ud til at følge den rette linje pænt. Derfor antager vi, at residualerne er normalfordelte og dermed at en lineær model beskriver sammenhængen mellem alder og makspuls godt - i hvert fald i vores datas tilfælde. Til slut vil vi gerne bestemme et 95%-konfidensinterval for parametrene a og b i vores lineære model

$$y = ax + b.$$

Dette gøres i Maple ved at skrive

```
with(Gym):
alder := [...]
puls := [...]
testLin(alder,puls)
```

Dette giver os et 95%-konfidensinterval for a på

$$[-1.00216, -0.95978]$$

og for b

$$[218.75654, 220.64216].$$

Opgave 1

Prisen på en volatil aktie i en tidsperiode på 401 sekunder kan findes i dette datasæt.

i) Lav lineær regression på datasættet.

ii) Lav et residualplot og overvej, om residualerne ser ud til at være normalfordelte.

iii) Lav et fraktilplot og overvej, om residualerne ser ud til at være normalfordelte.

iv) Brug modellen til at afgøre, hvornår prisen er 680kr. Bestem sandsynligheden for, at prisen til dette tidspunkt er under 678 kr.

v) Bestem et 95%-konfidensinterval for a og b i den lineære model.

Opgave 2

250 frø er sået på forskellige tidspunkter og højden på planten er målt efter et antal dage. Sammenhængen mellem alderen og højden kan ses i dette datasæt.

i) Lav lineær regression på datasættet.

ii) Lav et residualplot og overvej, om residualerne ser ud til at være normalfordelte.

iii) Lav et fraktilplot og overvej, om residualerne ser ud til at være normalfordelte.

iv) Hvad er sandsynligheden for, at planten er under 12cm efter 100 dage?

v) Brug modellen til at afgøre, hvornår planterne er 50cm høje.

vi) Bestem et 95%-konfidensinterval for a og b i den lineære model.

Opgave 3

Et studie på 413 personer skal afklare, om et bestemt kosttilskud kan øge den muskelstyrkende effekt af vægttræning. Effekten af træningen måles ved at tage

et vægtet gennemsnit af ydelsen på en række forskellige øvelser. I dette datasæt fremgår ydelsesforskellen sammenlignet med gennemsnittet af en kontrolgruppe, der har udført samme træningsprogram men uden kosttilskud.

i) Lav en lineær regression på datasættet, og afgør, om residualerne er normalfordelte.

ii) Bestem et 95%-konfidensinterval for hældningen a og afgør, om vi kan konkludere, at kosttilskudet har en virkning.