Lineær Regression
A-niveau

Bo Markussen
Københavns Universitet

Anders Rønn-Nielsen
Copenhagen Business School

9. oktober, 2018

Forord

En måde at blive klogere på den omkringliggende verden er ved at indsamle data og bruge dette til at opnå en forståelse af eventuelle sammenhænge. En udfordring man ofte møder er, at data i mange situationer er behæftet med variation, eller **støj**, som vi også vil kalde det. Formålet med en statistisk analyse er, at adskille underliggende sammenhænge fra denne usikkerhed. I dette manuskript vil vi vise, hvorledes dette kan gøres i situationer, der samlet set går under betegnelsen **regressionsanalyse**.

Det grundlæggende eksempel kaldes for **simpel lineær regression**. Udgangspunktet for dette er:

- Sammenhørende par $(x_1, y_1), \ldots, (x_n, y_n)$ af tal. Her er n antallet af talpar, dette kunne f.eks. være $n = 10$. Disse talpar kan f.eks. tegnes som punkter i et koordinatsystem. For at lineær regression overhovedet giver mening, er det afgørende, at en sådan tegning viser en passende lineær sammenhæng.

- Hvis tegningen viser talpar, der synes at variere omkring en ret linje, så kan denne linje bruges som en overordnet beskrivelse af punkterne. En sådan simpel beskrivelse kaldes for en **model**. Vores model er altså en *ret linje*

$$\ell(x) = a \cdot x + b$$

Der er måske nogle læsere, som undrer sig over, hvorfor vi ikke lægger vægt på **sandsynlighedsregning** og **normalfordelingen** i forbindelse med lineær regression. Det kunne vi såmænd også godt havde gjort, og for en dybere matematisk analyse (som gives i statistikundervisningen på mange videregående uddannelser) er normalfordelingen heller ikke til at komme udenom. Det er dog vores håb, at en statistik analyse uden brug af sandsynlighedsregningen vil give en bedre forståelse af, hvordan den statistiske metode egentlig virker. Undervejs i dette dokument vil vi dog vise, hvordan de anvendte datasæt passer sammen med normalfordelingen.

I stedet for at beskrive sandsynligheder matematisk vil vi vise, hvorledes usikkerheden kan måles ved at tilsætte **variation** i form af tilfældig udvælgelse af tal fra en række af tal. For at man kan uddrage brugbare konklusioner skal sådanne tilfældige udvælgelser gentages mange gange, f.eks. 1000 gange. Dette kan gøres ved hjælp af en computer. I statistiksprog tales om **simulering**.

Et andet princip, som har været styrende for udformningen af dette manuskript, er, at vi bruger autentiske datasæt. Altså datasæt der er blevet indsamlet ude i virkeligheden for at beskrive og forstå virkelige fænomener. Det skal understreges, at god statistik ofte sker i samspil med viden og indsigt fra andre fagområder. Hvis der er en naturlig forklaring på og forståelse af en sammenhæng, så giver det nemlig bedre mening at lede efter den i tallene. Med de enorme mængder data, der er til rådighed i dag, risikerer man ellers blot at finde de såkaldte **spuriøse** sammenhænge (se [6] for underholdende eksempler på dette), der ikke er udtryk for nogen underliggende mekanismer.

Der er indsat øvelsesopgaver inde i teksten. Vi anbefaler, at man løser opgaverne undervejs, inden man læser videre i teksten.

I tilknytning til dette materiale er der adgang til instruktionsvideoer, der viser, hvordan databearbejdningen foregår i gængse matematiske værktøjsprogrammer.

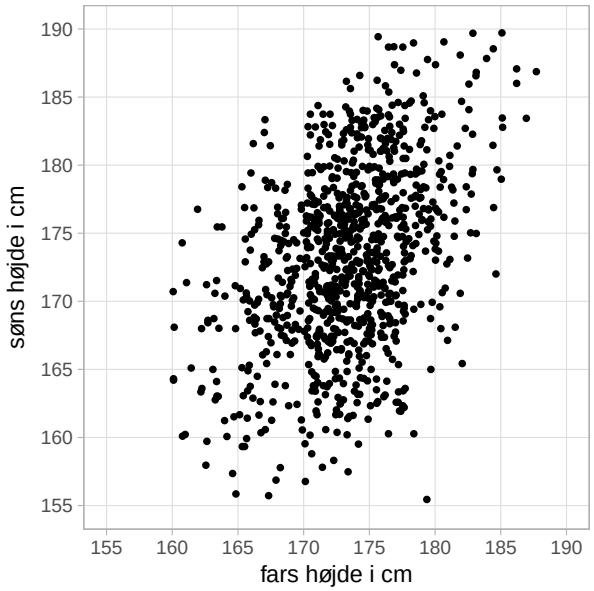
1 Simpel lineær regression

I dette afsnit vil vi se på et meget berømt datasæt, der blev indsamlet af Francis Galton tilbage i 1880'erne til en undersøgelse af Darwins arvelighedsteori. Historisk set var det analysen af dette datasæt, som medførte det umiddelbart besynderlige navn *"regressionsanalyse"*. Datasættet er altså interessant i matematikkens historie, men det illustrerer også på bedste vis mekanikken i regressionsanalysen. Og så kan vi ovenikøbet besvare det biologiske spørgsmål om i hvilken grad, en drengs højde kan forudses ud fra højden på hans far.

Datasættet indeholder sammenhørende målinger af fædres højder og deres førstefødte sønners højder som voksne. Der er målinger for i alt 952 par af fædre og sønner. De første 10 målinger (angivet i cm) ser ud som i skemaet herunder.

	fars højde	søns højde
1	186,9	183,4
2	184,6	172,0
3	185,0	179,0
4	182,1	165,4
5	179,4	155,4
6	178,4	160,3
7	179,7	165,0
8	179,7	168,7
9	176,5	160,3
10	173,4	157,5

For at få et overblik over alle 952 par af målinger, laver vi et plot, hvor hvert af de 952 par indtegnes som et punkt med faderens højde som x-værdi og sønnens højde som y-værdi. Dette plot kan ses på figur 1.



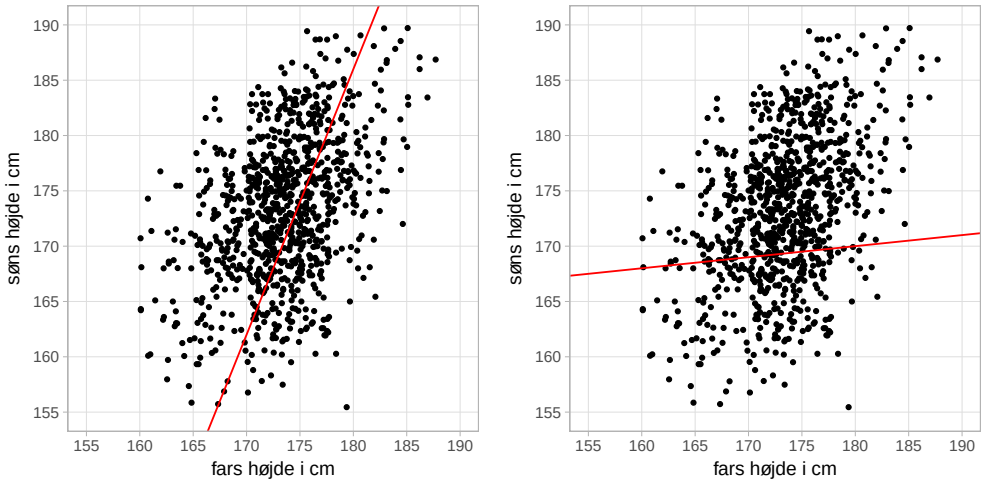
Figur 1: De 952 sammenhørende par af målinger for fædre og sønner.

Opgave 1. *Indlæs datasættet i et matematisk værktøjsprogram, og tegn et tilsvarende plot. Hvad er den mindste og største værdi af fædrenes højder? Hvad er den mindste og største værdi af sønnernes højder?*

Når man kigger på plottet, kunne det ved første øjekast godt se ud som om, punkterne ligger i en stor og tilfældig sky uden nogen nævneværdig sammenhæng mellem fædres og sønners højder. Ved nærmere eftersyn kan man imidlertid konstatere, at både øverste venstre hjørne og nederste højre hjørne er stort set tomme for punkter. De fleste af punkterne ligger i et retlinjet bælte fra plottets nederste venstre hjørne til det øverste højre hjørne. Dette er et tegn på, hvad vi vil kalde en **voksende sammenhæng** (betegnes også som en **positiv sammenhæng**) mellem fædres og sønners højder. Blandt de høje fædre er der en tendens til, at sønnerne er højere end gennemsnittet, mens der modsat er en tendens til, at sønner af relativt lave fædre selv er relativt lave.

Det er denne voksende sammenhæng, vi vil se nærmere på. Det første, vi vil gøre, er at forsøge at indtegne en ret linje på plottet, der passer "så godt som muligt" med punkterne.

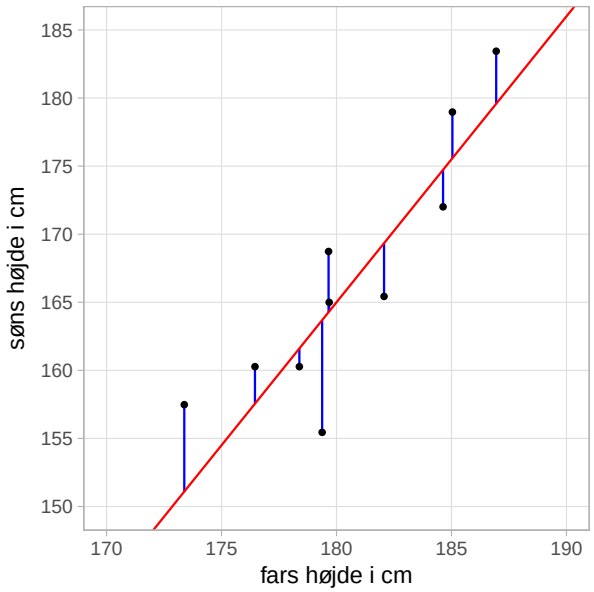
Opgave 2. *Prøv at indtegne forskellige linjer i plottet fra før. Overvej, hvad der skal til, for at en linje passer godt med punkterne.*



Figur 2: To valg af linjer til at beskrive punkternes voksende tendens.

To mere eller mindre vellykkede forsøg på at indtegne en god ret linje kan ses på figur 2. Umiddelbart kan man påstå, at linjen til venstre måske er lidt for stejl, mens hældningen på linjen til højre er for lille. Spørgsmålet er, hvilken af de to linjer der passer bedst, og om vi mon kan finde en, der

passer endnu bedre? Her får vi brug for en metode til at måle, hvor godt en given linje passer med punkterne.



Figur 3: De ti første punkter sammen med et (godt) bud på en linje (rød). De blå linjestykker markerer de lodrette afstande mellem punkterne og linjen. Det skal bemærkes, at 2 af de 10 fædre tilfældigvis har den samme højde (179,7 cm). Derfor er to af punkterne indtegnet lodret over hinanden.

For at gøre det lettere at overskue, vil vi til at starte med nøjes med at finde den linje, der passer bedst med de 10 første punkter på listen.

Opgave 3. *Lav et plot, der kun viser de 10 første punkter.*

På figur 3 er de 10 punkter indtegnet sammen med et forslag til en linje, der passer relativt godt med punkterne. Linjen har hældning $a = 2{,}1$ og skæring $b = -213$. Derudover er de lodrette afstande fra hvert datapunkt ind til linjen markeret med blå linjestykker. Vi vil bruge disse 10 lodrette afstande til at vurdere, hvor præcist den røde linje passer med datapunkterne. De 10 lodrette afstande udreges som

	fars højde	søns højde	lodret afstand
1	186,9	183,4	3,9
2	184,6	172,0	−2,7
3	185,0	179,0	3,5
4	182,1	165,4	−4,0
5	179,4	155,4	−8,3
6	178,4	160,3	−1,3
7	179,7	165,0	0,6
8	179,7	168,7	4,3
9	176,5	160,3	2,7
10	173,4	157,5	6,4

Opgave 4. *Den første lodrette afstand er udreget ved hjælp af formlen*

$$183{,}4 - \big(2{,}1 \cdot 186{,}9 - 213\big) = 3{,}9$$

Overvej, hvorfor formlen ser ud, som den gør. Kontroller, at vi har regnet rigtigt ved selv at udregne alle de 10 lodrette afstande.

Disse lodrette afstande kaldes i statistik også for **residualer**. De gængse værktøjsprogrammer har indbyggede kommandoer til at udregne og plotte residualerne hørende til en regression. Men her vil vi arbejde videre med vores egne beregninger for at forstå, hvad der ligger bag den statistiske metode. Sidst i materialet går vi lidt mere i dybden med hvilken information, vi kan aflede af residualerne.

Som mål for, hvor godt et givent valg af den røde linje passer med punkterne, vil vi bruge udtrykket

$$\sum_{\text{lodrette afstande}} (\text{størrelsen af lodret afstand})^2 \tag{1}$$

Det store græske bogstav $\sum$ kaldes for "*sigma*", og er matematisk notation for at lægge sammen. Med de 10 første par af fædres og sønners højder og valget af den røde linje med hældningen 2,1 og skæring −213, som på figur 3, bliver udtrykket altså

$$\begin{aligned} &3{,}9^2 + (-2{,}7)^2 + 3{,}5^2 + (-4{,}0)^2 + (-8{,}3)^2 \\ &+ (-1{,}3)^2 + 0{,}6^2 + 4{,}3^2 + 2{,}7^2 + 6{,}4^2 = 188{,}7 \end{aligned}$$

Vi vedtager nu, at den røde linje, som passer *"bedst muligt"* med datapunkterne, er den linje, som gør, at dette udtryk bliver *mindst muligt*. Intuitivt giver dette mål ret god mening: Hvis den røde linje passer godt med punkterne, bliver alle de lodrette afstande små, og summen bliver lille.

Omvendt bliver summen stor, hvis den røde linje passer dårligt med datapunkterne. I afsnit 1.4 vil vi argumentere for, hvorfor det giver mening at minimere de lodrette afstande, men indtil videre vil vi blot adoptere denne tilgang, som kaldes for **mindste kvadraters metode**.

Opgave 5. *Udregn selv summen af de kvadrerede lodrette afstande. Prøv at ændre på hældning og skæring, og udnyt værktøjsprogrammets muligheder for at få genberegnet talværdien. Prøv at finde værdier af hældning og skæring, der gør summen af de kvadrerede afstande så lille som muligt.*

Man kan selvfølgelig komme ret langt ved bare at "prøve sig frem". Imidlertid kan det også godt lade sig gøre at lave en matematisk beregning af hvilken linje, der passer bedst med en given samling af punkter. For at kunne gøre dette, vil vi introducere lidt notation. Vi vil formulere det generelt, således at vi kommer frem til et resultat, der også kan bruges, selvom der ikke er tale om 952 sammenhørende værdier af fædres og sønners højdemålinger.

Antag, at vi har n sammenhørende datapunkter, (x_i, y_i) for $i = 1, \ldots, n$. Det kunne altså f.eks. være de $n = 952$ målinger af fædres og sønners højde. Så ville x_i være den i'te fars højde, mens y_i ville være målingen hørende til den i'te søn. For eksempel var det første punkt i datasættet (186,9 , 183,4). Med den nye notation er så $x_1 = 186{,}9$ og $y_1 = 183{,}4$.

En linje, der har hældning a og skæring b, har forskriften

$$\ell(x) = a \cdot x + b$$

Den lodrette afstand mellem det i'te punkt (x_i, y_i) og linjen bliver (lige som før, men nu med den abstrakte notation a og b for hældning og skæring) $y_i - (a \cdot x_i + b)$, og derfor kan vi skrive summen af de kvadrerede lodrette afstande, altså formlen (1), på følgende måde

$$\sum_{i=1}^{n} \big(y_i - a \cdot x_i - b\big)^2. \tag{2}$$

Vi er interesserede i at finde *det* valg af hældningen a og skæringen b, som gør summen af de kvadrerede lodrette afstande mellem datapunkter og den rette linje $\ell(x)$ mindst mulig. For at bestemme den bedste hældning og skæring skal vi bruge gennemsnittet af x–værdierne og y–værdierne. Disse kaldes for $\bar{x}$ og $\bar{y}$, og beregnes via formlerne

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$\bar{y} = \frac{y_1 + \cdots + y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Opgave 6. *Eftervis i datasættet bestående af de 10 første punktpar, at* $\bar{x} = 180{,}6$ *og* $\bar{y} = 166{,}7$.

Nu kan vi komme med et matematisk udtryk for, hvordan den optimale hældning og den optimale skæring skal se ud. Vi dekorerer disse valg af a og b med tegnet " ˆ " for at angive, at dette er de optimale valg. Der gælder

$$\begin{aligned} \hat{a} &= \frac{\sum_{i=1}^{n}(y_i - \bar{y}) \cdot (x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \\ \hat{b} &= \bar{y} - \hat{a} \cdot \bar{x}. \end{aligned} \tag{3}$$

Det betyder, at udtrykket (2) bliver mindst muligt, netop når $a = \hat{a}$ og $b = \hat{b}$. Vi vil give et matematisk bevis for dette i afsnit 1.3.

Hvis vi skulle udregne tælleren i udtrykket for $\hat{a}$ for de første 10 punktpar, skulle vi regne

$$\begin{aligned} &(183{,}4 - 166{,}7) \cdot (186{,}9 - 180{,}6) + (172{,}0 - 166{,}7) \cdot (184{,}6 - 180{,}6) \\ &+ \cdots + (157{,}5 - 166{,}7) \cdot (173{,}4 - 180{,}6) \end{aligned}$$

og nævneren i udtrykket for $\hat{a}$ for de 10 punktpar er

$$(186{,}9 - 180{,}6)^2 + (184{,}6 - 180{,}6)^2 + \cdots + (173{,}4 - 180{,}6)^2$$

Opgave 7. *Udregn den optimale hældning* $\hat{a}$ *for de 10 første punktpar i datasættet. Udregn derefter den optimale skæring* $\hat{b}$ *– bemærk, at den udregnede værdi af* $\hat{a}$ *skal bruges i udtrykket for* $\hat{b}$.

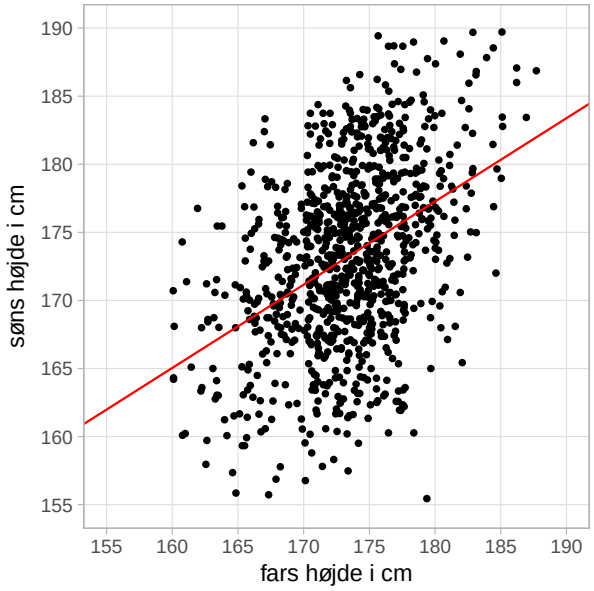
Indtegn denne optimale linje i plottet sammen med punkterne.

Når $\hat{a}$ og $\hat{b}$ udregnes på baggrund af *alle* 952 datapunkter med højdemålinger for fædre og sønner, fås $\hat{a} = 0{,}613$ og $\hat{b} = 67{,}0$. Altså linjen

$$\ell(x) = 0{,}613 \cdot x + 67{,}0$$

Linjen er indtegnet i et plot sammen med punkterne på figur 4. Vores umiddelbare forventning om, at der ville være en voksende sammenhæng mellem fædres og sønners højder, er altså blevet bekræftet: Hældningen er positiv!

Spørgsmålet er nu, hvad vi kan bruge linjen til? Det er jo ikke sådan, at punkterne ligger perfekt på linjen. Derimod ligger de i en tilsyneladende tilfældig sky omkring linjen. En måde at tænke på linjen er følgende: Vi forestiller os et nyt far–søn par, som ikke er med i undersøgelsen, hvor vi kender højden af faderen, mens sønnens højde er ukendt. Måske er sønnen et barn og dermed ikke udvokset endnu. Vi vil gerne prøve at forudsige sønnens højde.



Figur 4: Alle datapunkter indtegnet sammen med det bedste valg af ret linje.

Lad os antage, at faderens højde er 183 cm. Så er vores bedste bud på sønnens højde – baseret på datasættet – at udregne linjens værdi, når x–værdien er 183. Dette bud på sønnens højde er givet ved

$$\ell(183) = 0{,}613 \cdot 183 + 67{,}0 = 179{,}1$$

Vores gæt er altså, at en far, der er 183 cm høj, har en søn på 179,1 cm.

Opgave 8. *Antag, at en fars højde er 168 cm. Giv et bud på, hvor høj hans søn bliver. Prøv det samme, hvor faderens højde er 188 cm.*

Opgave 9. *Antag for et øjeblik, at linjen, der passede bedst med punkterne, havde hælding 1 og skæring 0. Altså at den så ud på følgende måde:*

$$\ell(x) = 1 \cdot x + 0 = x$$

Hvordan ville buddene på sønnernes højder være for de to far–højder 168 cm og 188 cm?

Hvis den bedste hældning var præcis 1, og den bedste skæring var 0, så ville buddet på sønners højde være nøjagtig deres fars højde. Men sådan er det altså ikke. I stedet er det sådan, at sønner af meget høje (højere end gennemsnittet) fædre ganske vist forventes at blive høje, men ikke helt så høje som deres fædre. Samtidigt forventes sønner af fædre, der er lavere end

gennemnsnittet, at blive relativt lave, men ikke helt så lave som deres fædre. Dette fænomen, som matematisk giver sig til udtryk ved, at $\hat{a}$ her ligger mellem 0 og 1, kaldes i den statistiske verden for *"regression towards the mean"*. Vi vil se nærmere på dette i afsnit 1.4.

Nu har vi set, hvordan den bedste rette linje kan bruges til at gætte på højden af sønnen i et "nyt" far–søn par. Vi har imidlertid ikke sagt noget om, hvor godt gættet er. Ovenfor gættede vi på, at en far med højden 183 cm ville få en søn, der var 179,1 cm høj. Det betyder ikke, at vi tror, at sønnen så faktisk får præcis højden 179,1 cm. Der er jo heller ingen af de andre punkter, der ligger nøjagtigtigt på linjen. Men vores bedste bud er 179,1 cm, og så kan vi få en ide om usikkerheden ved at kigge på, hvor langt alle de andre punkter ligger fra linjen. I det følgende vil vi give et matematisk mål for denne usikkerhed.

Den **gennemsnitlige kvadratiske afvigelse** fra den bedste rette linje betegnes ofte med symbolet $\hat{\sigma}^2$, og den har formel–udtrykket

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \hat{a} \cdot x_i - \hat{b} \right)^2$$

Bemærk, at der i denne formel divideres med $n-2$ og ikke med n, som man ellers umiddelbart ville gøre ved beregningen af gennemsnittet for de n kvadrerede lodrette afstande. Der er en matematisk forklaring på dette, som det dog vil føre for vidt at gå i dybden med her. Tager man kvadratroden, så fås

$$\hat{\sigma} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \hat{a} \cdot x_i - \hat{b} \right)^2} \tag{4}$$

Størrelsen $\hat{\sigma}$ kaldes **residualspredningen** og er altså et mål for, hvor langt punkterne i gennemsnit ligger fra den bedste rette linje.

Opgave 10. *Herunder ses et lille datasæt med 10 punktpar*

x	1	3	4	6	8	10	11	12	14	16
y	2,2	5,3	8,3	11,9	14,4	21,0	21,7	24,2	29,1	30,9

Indtegn de 10 punkter i et plot sammen med den bedste rette linje gennem punkterne. Beregn residualspredningen.

Gør det samme for dette datasæt

x	1	3	4	6	8	10	11	12	14	16
y	5,5	9,5	9,0	16,1	12,3	13,1	21,8	20,9	26,9	31,4

Diskuter, hvad størrelsen på $\hat{\sigma}$ betyder for punkternes afstand til den bedste rette linje (er punkterne generelt længere fra linjen, når $\hat{\sigma}$ er stor eller lille?).

Opgave 11. *Beregn $\hat{\sigma}$ ud fra de 10 første punktpar i far–søn–datasættet. Diskuter, hvad denne størrelse siger om usikkerheden på gættet for sønnens højde? Hvis $\hat{\sigma}$ er forholdsvis stor, er vi så mere eller mindre sikre på vores forudsigelse af sønnens højde?*

En rimelig konklusion på opgaverne ovenfor er, at $\hat{\sigma}$ siger noget om, hvor langt punkterne overordnet set ligger fra den bedste rette linje. Hvis punkterne generelt ligger langt fra linjen, er $\hat{\sigma}$ større, end den havde været, hvis punkterne lå tættere på linjen.

På den måde er $\hat{\sigma}$ et mål for usikkerheden på *"det bedste bud"*. En nyttig tommelfingerregel er, at punkterne i det store og hele opfylder, at den lodrette afstand ind til linjen højst er $2 \cdot \hat{\sigma}$. Hvad der menes med *"i det store og hele"*, vil blive præciseret senere.

Hvis vi regner $\hat{\sigma}$ for hele far–søn–datasættet, fås, at $\hat{\sigma} = 6{,}0$ (efter afrunding). Tommelfingerreglen siger så, at for de fleste af punkterne er den lodrette afstand mellem punktet og den bedste rette linje højst $2 \cdot 6 = 12$.

Vi efterprøver dette på det første punkt i datasættet, hvor $x_1 = 186{,}9$ og $y_1 = 183{,}4$. Den lodrette forskel mellem punktet og den bedste rette linje, der jo har forskriften $\ell(x) = 0{,}613 \cdot x + 67{,}0$, bliver nu

$$183{,}4 - \ell(186{,}9) = 183{,}4 - (0{,}613 \cdot 186{,}9 + 67{,}0) = 1{,}9$$

Forskellen er tydeligvis mindre end 12, så tommelfingerreglen havde ret i dette tilfælde! Vi kan gentage øvelsen for det andet punkt, hvor $x_2 = 184{,}6$ og $y_2 = 172{,}0$. Her fås

$$172{,}0 - \ell(184{,}6) = 172{,}0 - (0{,}613 \cdot 184{,}6 + 67{,}0) = -8{,}1$$

Også her er den lodrette forskel (numerisk) mindre end 12. Det er dog ikke alle punkter, der opfylder tommelfingerreglen om, at afstanden fra punkt til linje højst er 12. Hvis vi kigger på figur 4, kan vi se et punkt, hvor faderens højde er lige under 180 cm, og sønnens højde kun er lidt over 155 cm: Her ser det oplagt ud som om, at den lodrette forskel er noget større end 12. Ved at kigge datasættet igennem kan dette punkt findes allerede i den femte linje: $x_5 = 179{,}4$ og $y_5 = 155{,}4$. Her fås så, at den lodrette forskel er

$$155{,}4 - \ell(179{,}4) = 155{,}4 - (0{,}613 \cdot 179{,}4 + 67{,}0) = -21{,}5$$

hvilket er en meget større afstand end 12 (den er jo næsten dobbelt så stor). Det er altså ikke *alle* punkterne, der opfylder tommelfingereglen, men det

blev jo også kun præsenteret som en egenskab, der holdt *i det store og hele*! Tommelfingerreglen kan imidlertid præciseres: Generelt vil ca. 95% af punkterne højst ligge i afstanden $2 \cdot \hat{\sigma}$ fra den bedste rette linje. I vores tilfælde, hvor der er 952 punkter, vil vi så forvente, at cirka $0{,}95 \cdot 952 \approx 904$ af punkterne højst har afstanden 12 til linjen givet ved $\ell(x) = 0{,}613 \cdot x + 67{,}0$. Omvendt vil vi forvente, at cirka $0{,}05 \cdot 952 \approx 48$ af punkterne har en afstand til linjen, der er større end 12.

Opgave 12. *Undersøg hvor mange af punkterne, der opfylder, at den lodrette afstand mellem punktet og linjen $\ell(x) = 0{,}613 \cdot x + 67{,}0$ højst er 12.*

Undersøg også hvor mange af punkterne, der opfylder, at den lodrette afstand højst er 6.

Tommelfingerreglen kan også bruges som hjælp til at vurdere usikkerheden, når den bedste rette linje benyttes til at forudsige højden af sønnen i et nyt far–søn par, hvor faderens højde er kendt. Husk på, at vi gættede på, at en far på 183 cm ville få en søn på 179,1 cm. Nu kan vi sige, at den rigtige søn–højde med stor rimelighed kommer til at være højst 12 cm fra vores gæt. Vi er altså temmelig sikre på, at den rigtige søn–højde vil ligge et eller andet sted mellem $179{,}1 - 12 = 167{,}1$ cm og $179{,}1 + 12 = 191{,}1$ cm. Samlet siger vi, at vores bedste bud er 179,1 cm, og så udstyrer vi vores bud med **prædiktionsintervallet** [167,1 , 191,1], hvor vi i hvert fald er temmelig sikre på, at den rigtige søn–højde kommer til at ligge.

Opgave 13. *I opgave 8 blev der givet et bud på sønnens højde, hvis faderen er hhv. 168 cm høj og 188 cm høj. Find prædiktionsintervallerne hørende til hvert af disse to bud.*

1.1 Korrelation og forklaringsgrad

Vi har set, at i datasættet med fædres og sønners højder er der en sammenhæng mellem disse størrelser: Hvis faderen er høj, forventer vi også, at sønnen er relativt høj, og hvis faderen er lav, forventer vi tilsvarende, at sønnen er det. Sammenhængen er dog ikke deterministisk i den forstand, at sønnens højde ikke kan forudsiges præcist ud fra faderens højde.

I dette afsnit vil vi give et mål for, hvor godt man kan forudsige y–værdierne ud fra x–værdierne. Hertil definerer vi størrelsen ρ, som vi kalder **korrelationen** mellem x–erne og y–erne. Den er defineret ved udtrykket

$$\rho = \frac{\sum_{i=1}^{n}(y_i - \bar{y}) \cdot (x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \cdot \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}. \tag{5}$$

Bemærk, at denne formel minder ret meget om formeludtrykket for $\hat{a}$. Forskellen er, at der i nævneren både indgår et udtryk om x–værdierne og et udtryk om y–værdierne samt kvadratroden af disse.

Matematisk kan man vise, at ρ altid er et tal i intervallet fra -1 til 1. Hvis $\rho = 1$, ligger alle datapunkterne (x_i, y_i) perfekt på en ret linje med positiv hældning, og tilsvarende hvis $\rho = -1$, ligger alle datapunkterne perfekt på en ret linje med negativ hældning. Man kan sige, at i disse tilfælde er det muligt at forudsige y–værdierne helt præcist ud fra x–værdierne. Hvis $\rho = 0$, så er der derimod ingen lineær sammenhæng mellem x og y, hvilket betyder, at man ikke kan forbedre sit gæt på y–værdierne ved at bruge x–værdierne i formlen for en ret linje. For datasæt indsamlet i virkeligheden vil korrelationen som regel ligge et sted strengt mellem -1 og 1, og det er også meget sjældent, at korrelationen præcist bliver 0. Jo tættere korrelationen er på -1 eller 1, jo mere velegnet er den rette linje til at forudsige y–værdierne ud fra x–værdierne.

Opgave 14. *Herunder ses et lille datasæt med 10 punkter*

x	1	3	4	6	8	10	11	12	14	16
y	3	7	9	13	17	21	23	25	29	33

Indtegn de 10 punkter i et plot, og prøv at udregne korrelationen ρ. Overvej, hvordan værdien af korrelationen passer sammen med, hvordan punkterne ligger i plottet.

Gør det samme for disse to andre datasæt:

x	3	5	6	8	9	10	12	14	18	21
y	15,2	14,1	11,4	8,0	6,3	4,9	2,1	$-1{,}3$	$-7{,}0$	$-11{,}4$

og

x	1	4	7	8	9	11	14	16	19	22
y	4,3	3,5	0,8	$-11{,}2$	3,1	$-0{,}3$	$-13{,}2$	12,8	$-6{,}7$	$-22{,}1$

Vi vil nu introducere et andet, men nært beslægtet, mål for, hvor godt y–værdierne kan forudsiges ved hjælp af x–værdierne. Dette mål kaldes **forklaringsgraden**, og har den matematiske notation R^2. For at vi kan give en matematisk definition af R^2, er det praktisk at have et navn for prædiktionen af y_i ud fra x_i. Lad os kalde denne for $\hat{y}_i$, altså

$$\hat{y}_i = \hat{a} \cdot x_i + \hat{b}.$$

Opgave 15. *Gør rede for, at gennemsnittet af y–erne og gennemsnittet af $\hat{y}$–erne er ens. Vink: Brug, at $\hat{b} = \bar{y} - \hat{a} \cdot \bar{x}$, hvor $\bar{y}$ og $\bar{x}$ er gennemsnittet af henholdsvis y–erne og x–erne.*

Forklaringsgraden R^2 er defineret som kvadratet på korrelationen mellem de observerede værdier y_i og de prædikterede værdier $\hat{y}_i$. For simpel lineær regression gælder dermed, at

$$R^2 = \frac{\left(\sum_{i=1}^{n}(y_i - \bar{y}) \cdot (\hat{y}_i - \bar{y})\right)^2}{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right) \cdot \left(\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2\right)}, \tag{6}$$

hvor vi har brugt formlen for korrelation (mellem y_i og $\hat{y}_i$), og at gennemsnittet for $\hat{y}$-erne er lig med $\bar{y}$. Vi vil vise i afsnit 1.2, at det faktisk gælder, at forklaringsgraden er lig kvadratet på korrelationen ρ mellem x og y, altså at $R^2 = \rho^2$, hvor ρ er givet ved udtrykket i (5). Det betyder, at R^2 altid er et tal mellem 0 og 1. Hvis $R^2 = 1$, så ligger alle datapunkterne (x_i, y_i) perfekt på en ret linje. Hvis derimod $R^2 = 0$, så kan man (præcis som når $\rho = 0$) ikke bruge den rette linje til at forudsige noget om y–værdierne ud fra de tilhørende x–værdier. Afhængigt af hvor tæt tallet er på 1, kan vi sige, at den rette linje er mere eller mindre velegnet til at forudsige y–værdierne ud fra x–værdierne.

Opgave 16. *Udregn alle de 952 prædikterede y–værdier for far–søn–datasættet. Udregn forklaringsgraden R^2 ved hjælp af formel (6), og kontroller, at vores påstand om, at $R^2 = \rho^2$, faktisk passer i dette tilfælde.*

Man kunne nu med god ret spørge om, hvorfor vi overhovedet indfører begrebet forklaringsgrad, når det viser sig, at udtrykket er så nært beslægtet med korrelationen mellem x–værdierne og y–værdierne. Kunne vi ikke bare have nøjedes med korrelationen? Svaret er, at netop i denne situation havde korrelationen ρ mellem x og y været tilstrækkelig til at beskrive, hvor godt x-erne kan forudsige y-erne. Imidlertid kan udtrykket (6) også anvendes i den mere generelle multilineære regressionsmodel, som vi vil studere i afsnit 5. Korrelationen i sig selv kan derimod ikke umiddelbart generaliseres til dette tilfælde.

1.2 Forskel mellem "lodret" og "vandret" variation

I simpel lineær regression forsøger vi at forudsige y–værdierne ud fra x–værdierne vha. en ret linje. Men hvad sker der, hvis man bytter om på rollen af variablene, således at vi forsøger at forudsige x-erne ud fra y-erne? I dette afsnit dykker vi lidt dybere ned i matematikken bag dette spørgsmål. Det er

dog ikke nødvendigt at læse dette for at forstå resten af materialet, og hvis man ønsker det, kan man fortsætte læsningen ved afsnit 1.3.

Men for dem, der har mod på matematikken, vil vi starte med at se på definitionen af R^2 i formel (6). Hvis vi bruger $\hat{b} = \bar{y} - \hat{a} \cdot \bar{x}$, så fås følgende omskrivning

$$\hat{y}_i = \hat{a} \cdot x_i + \hat{b} = \hat{a} \cdot (x_i - \bar{x}) + \bar{y}$$

af prædiktionen $\hat{y}_i$. Indsættes dette i formel (6), så fås

$$\begin{aligned} R^2 &= \frac{\left(\sum_{i=1}^n (y_i - \bar{y}) \cdot \left(\hat{a} \cdot (x_i - \bar{x}) + \bar{y} - \bar{y} \right) \right)^2}{\left(\sum_{i=1}^n \left(y_i - \bar{y} \right)^2 \right) \cdot \left(\sum_{i=1}^n \left(\hat{a} \cdot (x_i - \bar{x}) + \bar{y} - \bar{y} \right)^2 \right)} \\ &= \frac{\left(\sum_{i=1}^n (y_i - \bar{y}) \cdot (x_i - \bar{x}) \right)^2 \cdot \hat{a}^2}{\left(\sum_{i=1}^n (y_i - \bar{y})^2 \right) \cdot \left(\sum_{i=1}^n (x_i - \bar{x})^2 \right) \cdot \hat{a}^2} \\ &= \frac{\left(\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y}) \right)^2}{\left(\sum_{i=1}^n (x_i - \bar{x})^2 \right) \cdot \left(\sum_{i=1}^n (y_i - \bar{y})^2 \right)} \\ &= \rho^2, \end{aligned}$$

hvor ρ er korrelationen mellem x–erne og y–erne som givet i formeludtrykket (5). Idet korrelationen ρ ikke ændres, hvis der byttes om på rollen af x–erne og y–erne, så gælder dette også for R^2. Det betyder, at y–erne forklarer lige så meget om x–erne (hvis der bruges en ret linje), som x–erne forklarer om y–erne!

Opgave 17. *Det sidste af de tre datasæt, der blev regnet på i opgave 14, var følgende*

x	1	4	7	8	9	11	14	16	19	22
y	4,3	3,5	0,8	−11,2	3,1	−0,3	−13,2	12,8	−6,7	−22,1

Udregn forklaringsgraden for dette datasæt. Prøv nu at bytte rundt på x og y og udregn forklaringsgraden igen. Dvs, prøv at regne forklaringsgraden for følgende datasæt:

x	4,3	3,5	0,8	−11,2	3,1	−0,3	−13,2	12,8	−6,7	−22,1
y	1	4	7	8	9	11	14	16	19	22

Er de to forklaringsgrader ens?

Når nu forklaringsgraden bliver den samme, selvom man bytter rundt på rollerne for x–erne og y–erne, så kunne man tro, at det faktisk var helt lige meget, hvad der tildeles rollen som x, og hvad der tildeles rollen som y. I

far–søn–datasættet ville det betyde, at vi (på en måde) ville få de samme resultater, hvis vi byttede rundt på fædre og sønner og i stedet så fædrenes højder som et resultat af sønnernes. Biologisk ville denne måde nok ikke give så meget mening (vi gør flere overvejelser omkring dette i afsnit 1.4), da fædrene jo i sagens natur er kommet først, men matematisk kan man jo lige så godt bruge sønnernes højder som x–værdier og fædrenes som y–værdier.

Det viser sig imidlertid, at der er en forskel på, hvad man tildeler rollen som x, og hvad man tildeler rollen som y. Lad os illustrere dette på far–søn–datasættet. Her fandt vi frem til, at linjen givet ved

$$\ell(x) = 0{,}613 \cdot x + 67{,}0$$

passede bedst med alle punkterne i den forstand, at den gjorde summen af de kvadrerede lodrette afvigelser mindst mulig. Vi brugte bl.a. denne linje til at forudsige en søns højde ud fra faderens højde – simpelthen ved at sætte faderens højde ind i formlen. Man kunne også regne "omvendt" og forsøge at besvare spørgsmålet: "Hvilken far–højde ville give en forudsagt søn–højde på 180 cm?" Her må det blive et spørgsmål om at løse ligningen

$$180 = 0{,}613 \cdot x + 67{,}0$$

Opgave 18. *Løs ligningen, og vis derved, at hvis faderens højde er $x = 184{,}3$ cm, så er den forventede højde for sønnen 180 cm.*

Vi kunne gøre dette mere generelt ved at finde faderens højde x, hvis den forventede søn–højde skal være y. Det gør vi helt tilsvarende ved at isolere x i ligningen herunder

$$y = 0{,}613 \cdot x + 67{,}0$$

Her fås det let, at

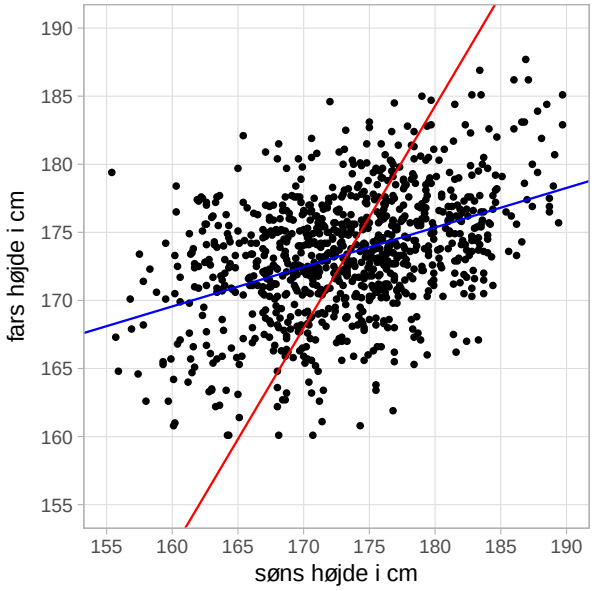
$$x = 1{,}631 \cdot y - 109{,}3$$

Opgave 19. *Regn efter og vis, at x bliver som angivet ovenfor.*

Denne funktion af far–værdien, y, er altså funktionsforskriften for en ret linje. Den kan vi skrive som

$$\ell(y) = 1{,}631 \cdot y - 109{,}3 \tag{7}$$

Det vil sige en ret linje, som har hældning 1,631 og skæring $-109{,}3$. På figur 5 er denne rette linje indtegnet sammen med alle søn–far–punkterne. Altså, hvor søn–værdierne er brugt som førstekoordinat, mens far–værdierne er brugt som andenkoordinat (her bruges ikke længere x– og y–værdi for at undgå forvirring). Punkterne og den røde linje er sådan set helt identisk med,



Figur 5: Datapunkter med søn–værdier som x og far–værdier som y med besdste rette linje (blå) og den rette linje fra (7) indtegnet som rød.

hvad der blev vist på figur 4; der er bare byttet rundt på akserne. Og linjen viser som hidtil sammenhængen mellem faderens højde og sønnens forventede højde.

En anden måde at beskrive sammenhængen mellem søn–værdierne og far–værdierne ville være simpelthen at finde den bedste rette linje gennem alle punkterne i figur 5. Altså at finde den linje, der gør summen af lodrette afstande fra punkt til linje mindst mulig. Hvis vi gør det, får vi linjen

$$\tilde{\ell}(y) = 0{,}290 \cdot y + 123{,}2$$

Denne linje er indtegnet med blåt på figur 5. Her ser vi noget ret interessant: Der er virkelig stor forskel på den røde og den blå linje! Den korte og intuitive forklaring på dette er, at den blå linje er fundet ved at gøre summen af de kvadrerede lodrette afstande mindst mulig, mens den røde linje er fundet ved at gøre summen af de kvadrerede *vandrette* afstande mindst mulig.

Opgave 20. *Overvej, hvorfor det giver mening at påstå, at den røde linje er den, som gør summen af de kvadrerede vandrette afstande mellem punkter og linje mindst mulig. Her kan det være en hjælp at tænke på, at det er den samme linje (pånær ombytning af akserne) som den røde linje i figur 4.*

I resten af afsnittet vil vi prøve at give en mere matematisk forklaring på forskellen mellem den røde og den blå linje. Vi får brug for to nye størrelser,

nemlig *spredningerne* σ_x og σ_y af x–erne og y–erne. Disse størrelser er defineret ved

$$\sigma_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \sigma_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Spredningerne måler, hvor meget x–erne og y–erne varierer omkring deres middelværdi. Der er en matematisk begrundelse for at dividere med $n-1$ og ikke med n, som man måske umiddelbart ville gøre, i definitionen af σ_x og σ_y. Vi vil ikke komme nærmere ind på dette i nærværende undervisningsmateriale, men blot notere os, at denne detalje er mindre væsentlig for os, idet det først og fremmest er forholdet $\frac{\sigma_x}{\sigma_y}$, der er vigtigt for os. Udfra formel (3) og (5) følger nemlig, at der gælder

$$\hat{a} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma_y}{\sigma_x} \cdot \rho.$$

Opgave 21. *Prøv selv at udregne σ_x, σ_y og ρ for far–søn–datasættet, og kontroller, at $\hat{a}$ kan regnes med den ovenstående formel.*

Hvis vi indsætter dette i den generelle formel for den bedste rette linje, $\ell(x) = \hat{a} \cdot x + \hat{b}$, fås

$$\begin{aligned} \ell(x) &= \hat{a} \cdot x + \hat{b} \\ &= \left(\bar{y} - \hat{a} \cdot \bar{x}\right) + \hat{a} \cdot x \\ &= \bar{y} + \hat{a} \cdot (x - \bar{x}) \\ &= \bar{y} + \frac{\sigma_y}{\sigma_x} \cdot \rho \cdot (x - \bar{x}). \end{aligned} \tag{8}$$

Ved, som vi gjorde for far–søn–linjen, at sætte højresiden lig med y

$$y = \bar{y} + \frac{\sigma_y}{\sigma_x} \cdot \rho \cdot (x - \bar{x})$$

og så isolere x, får vi med lidt omskrivninger, at

$$x = \bar{x} + \frac{\sigma_x}{\sigma_y} \cdot \frac{1}{\rho} \cdot (y - \bar{y})$$

Opgave 22. *Regn selv efter, og se, om du kan opnå samme resultat!*

Altså har vi nu et udtryk for, hvordan x–værdierne afhænger af y–værdierne. Sammenhængen er givet ved linjen

$$\ell(y) = \bar{x} + \frac{\sigma_x}{\sigma_y} \cdot \frac{1}{\rho} \cdot (y - \bar{y}) \tag{9}$$

og vi kan tænke på denne linje som den oprindelige linje $\ell(x)$, hvor vi har byttet rundt på akserne.

Opgave 23. *Kontroller, at for far–søn–datasættet, bliver $\ell(y)$ i formel (9) bare til den linje, vi fandt i (7):*

$$\ell(y) = 1{,}631 \cdot y - 109{,}3.$$

Hvis vi nu i stedet direkte havde fundet den bedste rette linje for alle de ombyttede punkter (y_i, x_i), så ville vi ifølge formel (8) med en ombytning af x og y få en ret linje givet ved

$$\tilde{\ell}(y) = \bar{x} + \frac{\sigma_x}{\sigma_y} \cdot \rho \cdot (y - \bar{y}). \tag{10}$$

Vi ser direkte, at forskellen mellem hældningen på linjen i (9) og hældningen på linjen i (10) er, at den sidstnævnte er ρ^2 gange større end den førstnævnte. De to hældninger kan kun blive ens, hvis $R^2 = \rho^2 = 1$, hvad der ville svare til, at alle punkterne ligger præcist på en ret linje.

Opgave 24. *Kontroller, at det passer med far–søn–datasættet, at linjen*

$$\tilde{\ell}(y) = 0{,}290 \cdot y + 123{,}2$$

har en hældning, der er R^2 gange større (bemærk, at R^2 er mindre end 1) end hældningen på linjen

$$\ell(y) = 1{,}631 \cdot y - 109{,}3.$$

1.3 Udledning af $\hat{a}$ og $\hat{b}$

I dette afsnit vil vi give argumentet for, at den bedste rette linje opnås ved at vælge a og b som de udtryk for $\hat{a}$ og $\hat{b}$, der er angivet i formel (3).

Resten af notatet kan sagtens læses uden at læse dette afsnit – selvom du selvfølgelig opfordres til at læse det. Det vigtigste er at forstå, at formeludtrykkene for hældningen og skæringen hørende til den bedste rette linje er udledt ved matematiske argumenter, og at der altså ikke er tale om at indtegne den linje, som man subjektivt føler ser bedst ud sammen med datapunkterne.

For at komme frem til hovedresultatet får vi brug for følgende lille hjælperesultat (i matematiksprog kaldes dette for et *lemma*), som du nok kender fra emnet parabler. Det handler om, ved hvilken værdi en parabel, der "vender benene opad", bliver mindst mulig.

I lemmaet vælger vi at lade z betegne funktionsvariablen i stedet for det sædvanlige x for at undgå sammenblanding med observationsværdierne x_i.

Lemma. *Lad $f(z) = c_1 \cdot z^2 + c_2 \cdot z + c_3$ være funktionsudtrykket for en parabel med $c_1 > 0$. Så er $f(z)$ mindst mulig, netop når $z = -\frac{c_2}{2c_1}$.*

Husk på, at vi ønsker at finde a og b, så udtrykket

$$\sum_{i=1}^{n} \left(y_i - b - a \cdot x_i\right)^2$$

bliver mindst muligt. Til at starte med vil vi glemme a–parameteren og så finde det b, der gør summen mindst. Vi opfatter altså udtrykket som en funktion $f(b)$ af b. Vores tilgang vil være at omskrive $f(b)$, så det bliver funktionsudtrykket for en parabel. Først laver vi følgende omskrivning

$$f(b) = \sum_{i=1}^{n} \left(y_i - b - a \cdot x_i\right)^2 = \sum_{i=1}^{n} \left(b^2 + 2 \cdot (a \cdot x_i - y_i) \cdot b + (a \cdot x_i - y_i)^2\right)$$

Opgave 25. *Det væsentlige i omskrivningen ovenfor er, at*

$$\left(y_i - b - a \cdot x_i\right)^2 = b^2 + 2 \cdot (a \cdot x_i - y_i) \cdot b + (a \cdot x_i - y_i)^2$$

Prøv selv at eftervise dette. Prøv også at udregne både $\sum_{i=1}^{n} \left(y_i - b - a \cdot x_i\right)^2$ og $\sum_{i=1}^{n} \left(b^2 + 2 \cdot (a \cdot x_i - y_i) \cdot b + (a \cdot x_i - y_i)^2\right)$, når $n = 2$, og (x_i, y_i) er givet som de første 2 punktpar i datasættet med fædre og sønner.

Næste skridt i argumentrækken er at skrive videre om på $f(b)$ til følgende

$$f(b) = n \cdot b^2 + \left(\sum_{i=1}^{n} 2 \cdot (a \cdot x_i - y_i)\right) \cdot b + \left(\sum_{i=1}^{n} (a \cdot x_i - y_i)^2\right).$$

Opgave 26. *Prøv at eftervise dette, når $n = 2$ ved at skrive sumtegnene ud, således at*

$$\sum_{i=1}^{2} \left(b^2 + 2 \cdot (a \cdot x_i - y_i) \cdot b + (a \cdot x_i - y_i)^2\right)$$
$$= \left(b^2 + 2 \cdot (a \cdot x_1 - y_1) \cdot b + (a \cdot x_1 - y_1)^2\right) + \left(b^2 + 2 \cdot (a \cdot x_2 - y_2) \cdot b + (a \cdot x_2 - y_2)^2\right),$$

og så skrive om på dette ved at ombytte rækkefølgen.

Hvis vi nu skriver

- c_1 i stedet for n
- c_2 i stedet for $\sum_{i=1}^{n} 2 \cdot (a \cdot x_i - y_i)$

- c_3 i stedet for $\sum_{i=1}^{n}(a \cdot x_i - y_i)^2$,

så står der bare, at

$$f(b) = c_1 \cdot b^2 + c_2 \cdot b + c_3\,,$$

hvor vi endda ved, at $c_1 > 0$ (det var jo bare et andet navn for n). Sådan et funktionsudtryk kan vi godt finde minimum for: Lemmaet siger, at b så skal vælges som $-\frac{c_2}{2c_1}$. Hvis vi indsætter, hvad c_1 og c_2 betyder, fås, at det optimale b, som vi har valgt at kalde $\hat{b}$, skal vælges som

$$\hat{b} = -\frac{c_2}{2c_1} = -\frac{\sum_{i=1}^{n} 2 \cdot (a \cdot x_i - y_i)}{2n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - a \cdot x_i) = \bar{y} - a \cdot \bar{x}.$$

Opgave 27. *Prøv selv at eftervise det sidste lighedstegn for $n = 2$ ved at skrive udtrykket ud*

$$\frac{1}{2}\sum_{i=1}^{2}(y_i - a \cdot x_i) = \frac{y_1 - a \cdot x_1 + y_2 - a \cdot x_2}{2}$$

og så bytte lidt rundt på rækkefølgen. Husk på, at $\bar{x}$ og $\bar{y}$ er gennemsnittet af x–erne hhv. y–erne.

Hidtil har vi haft fastholdt a. Det betyder at ligegyldigt hvilket tal, vi vælger at sætte ind i stedet for a, så er det bedst at vælge $\hat{b}$ som $\bar{y} - a \cdot \bar{x}$. Bemærk, at udtrykket for $\hat{b}$ afhænger af a. Hvis vi indsætter formeludtrykket $\hat{b} = \bar{y} - a \cdot \bar{x}$ i stedet for b, så får den rette linje forskriften

$$\ell(x) = a \cdot x + \hat{b} = a \cdot x + \bar{y} - a \cdot \bar{x} = \bar{y} + a \cdot (x - \bar{x}).$$

Næste skridt bliver at bruge linjen $\ell(x) = \bar{y} + a \cdot (x - \bar{x})$ i udtrykket for summen af de kvadrerede lodrette afstande $y_i - \ell(x_i)$:

$$\sum_{i=1}^{n}\Big(y_i - \big(\bar{y} + a \cdot (x_i - \bar{x})\big)\Big)^2 = \sum_{i=1}^{n}\Big((y_i - \bar{y}) - a \cdot (x_i - \bar{x})\Big)^2$$

og så overveje, hvordan a skal vælges for at gøre udtrykket mindst muligt. Vi skal altså finde minimumspunktet for følgende funktion af a

$$\begin{aligned} f(a) &= \sum_{i=1}^{n}\Big((y_i - \bar{y}) - a \cdot (x_i - \bar{x})\Big)^2 \\ &= \Big(\sum_{i=1}^{n}(x_i - \bar{x})^2\Big) \cdot a^2 + \Big(-2 \cdot \sum_{i=1}^{n}(y_i - \bar{y}) \cdot (x_i - \bar{x})\Big) \cdot a + \Big(\sum_{i=1}^{n}(y_i - \bar{y})^2\Big). \end{aligned}$$

Opgave 28. *Prøv selv at eftervise denne omskrivning ved først at vise, at*

$$\Big((y_i - \bar{y}) - a \cdot (x_i - \bar{x})\Big)^2 = (x_i - \bar{x})^2 \cdot a^2 - 2 \cdot (y_i - \bar{y}) \cdot (x_i - \bar{x}) \cdot a + (y_i - \bar{y})^2$$

Hvis vi nu skriver

- c_1 i stedet for $\sum_{i=1}^n (x_i - \bar{x})^2$
- c_2 i stedet for $-2 \cdot \sum_{i=1}^n (y_i - \bar{y}) \cdot (x_i - \bar{x})$
- c_3 i stedet for $\sum_{i=1}^n (y_i - \bar{y})^2$,

så står der bare, at

$$f(a) = c_1 \cdot a^2 + c_2 \cdot a + c_3 \,.$$

Opgave 29. *Overvej, hvorfor det nødvendigvis må gælde, at $c_1 > 0$; medmindre at alle x_i'erne er helt ens.*

Nu giver lemmaet, at $f(a)$ er mindst mulig, hvis vi vælger a til at være

$$\hat{a} = -\frac{c_2}{2c_1} = -\frac{-2 \cdot \sum_{i=1}^n (y_i - \bar{y}) \cdot (x_i - \bar{x})}{2 \cdot \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (y_i - \bar{y}) \cdot (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Alt i alt har vi fundet frem til, at den bedste rette linje er givet ved

$$\ell(x) = \bar{y} + \frac{\sum_{i=1}^n (y_i - \bar{y}) \cdot (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot (x - \bar{x}) = \hat{a} \cdot x + \hat{b},$$

hvor $\hat{a}$ og $\hat{b}$ er valgt som beskrevet i formel (3).

1.4 Matematikken i biologien

Ud fra et biologisk synspunkt er det forventeligt, at en søns højde er positivt korreleret med faderens højde. Sønnen og faderen deler nemlig nøjagtig halvdelen af deres gener. Den anden halvdel af generne har sønnen arvet fra sin mor, mens faderens øvrige gener alene deles med farmor og farfar (cirka en fjerdedel for hver). Derudover kan en sammenhæng mellem fædres og sønners højder skyldes kulturelle og sociologiske forhold. Hvis mænd og kvinder foretrækker en partner af sammenlignelig højde som dem selv, så vil moderens gener ligne faderens gener mht. højde, hvormed det genetiske aspekt vil blive yderligere forstærket. Videre kunne der f.eks. være sammenhæng mellem faderens og sønnens opvækstvilkår, men idet der er gået mange år fra, at faderen voksede op til, at hans sønner vokser op, så er dette formodentlig af mindre betydning.

Alt dette er ikke bare snak, men det har en konsekvens for, hvordan vi kvantificerer sammenhængen mellem sønnens og faderens højde. Hvis vi ønsker at forudsige, hvor høje sønner en mand har, alene ud fra hans højde, så giver diskussionen ovenfor, at vi indirekte leder efter effekten af den genetiske fællesmængde mellem faderen og sønnen. Men hvis faderen f.eks. er usædvanlig høj, så kunne dette også skyldes nogle af de gener, som faderen har til fælles med sin mor eller far, men som *ikke* er gået i arv til sønnen. Hvis vi ønsker at beskrive sammenhængen mellem sønnens og faderens højde med en ret linje *og* bruge denne til at forudsige sønnernes højde ud fra deres fars højde, så har det således følgende *matematiske konsekvenser:*

1. Faderen har den højde, han nu en gang har, og det er kun en del af faderens afvigelse fra gennemsnitshøjden af fædre, der kan forventes overført til sønnen. Resten af afvigelsen skal *"føres tilbage"* (på engelsk: *"to regress"*) mod gennemsnittet.

2. Den bedste rette linje er den, som på passende vis minimerer den *lodrette* afstand til datapunkterne, altså afvigelsen mellem den faktiske højde af sønnerne og forudsigelsen ud fra deres fars højde.

Begrebet "regression towards the mean" er altså et matematisk fænomen, der optræder, når en egenskab (genetisk, sociologisk, økonomisk, eller andet) delvist deles af og har indflydelse på to forskellige enheder (søn og far), og man forsøger at prædiktere (forudsige) den ene ud fra den anden.

Opgave 30. *Vi forestiller os et nyt far-søn par, som ikke er med i undersøgelsen, hvor vi kender sønnens højde, mens faderens højde er ukendt. Diskuter, om man kan bruge den samme* "bedste rette linje" *som før til at forudsige faderens højde. Dette er et svært spørgsmål, men man kan eventuelt tage udgangspunkt i begrebet "regression towards the mean" – vil man f.eks. forvente at en far er lige så høj som sin søn, hvis sønnen er meget høj?*

2 Statistisk model

I de forrige afsnit kom vi frem til, at der godt kan siges at være en lineær sammenhæng mellem fædres og sønners højder. Det skal ikke forstås sådan, at hvis man kender faderens højde, så ved man også præcist, hvor høj sønnen er. Derimod vil det være i orden at sige, at hvis man kender faderens højde, så kan den rette linje bruges til at sige i hvilket område af y–værdier, vi kan *forvente* at finde sønnens højde.

Nu vil vi vende hele tankegangen omkring datapunkterne på hovedet. I stedet for bare at kigge på de observerede datapunkter vil vi prøve at lave en

slags matematisk beskrivelse af, hvordan sønnernes højder er fremkommet som resultat af fædrenes højder. Bemærk, at der tales om en *matematisk* beskrivelse. Vi er altså ikke ude efter at give en detaljeret biologisk forklaring på, hvorfor den enkelte søns højde er blevet det eksakte tal, der er observeret.

Vi tænker os, at sønnernes værdier, altså tallene $y_1, \ldots, y_{952}$, er frembragt ud fra fædrenes værdier, $x_1, \ldots, x_{952}$ efter følgende opskrift

$$y_i = a \cdot x_i + b + r_i \tag{11}$$

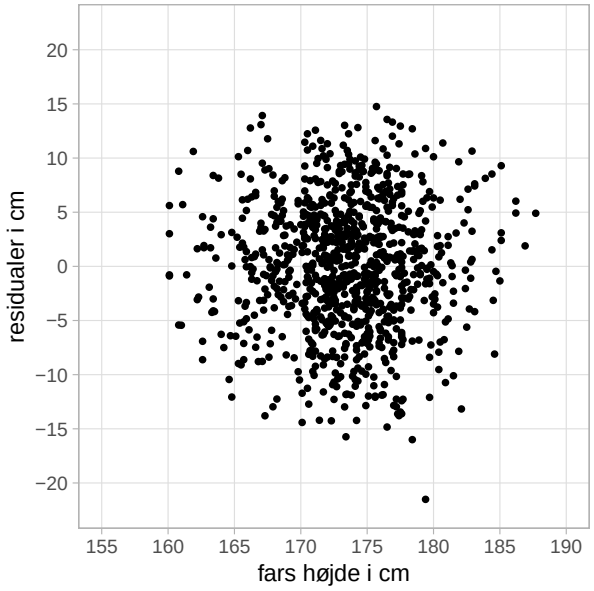
for alle $i = 1, \ldots, 952$. Her er a og b henholdsvis hældning og skæring for en ret linje $\ell(x) = a \cdot x + b$, og $r_1, \ldots, r_{952}$ er afvigelserne fra linjen. Størrelserne r_i svarer til **residualerne**, dvs. de lodrette afstande, og de bestemmes altså som den fejl, der bliver begået, hvis man forsøger at udregne y–værdierne ud fra x–værdierne ved brug af den rette linje $\ell(x) = a \cdot x + b$. Umiddelbart virker det måske som ren snyd at opskrive sammenhængen mellem x–erne og y–erne på denne måde. Vi har jo bare introduceret en række tal $r_1, \ldots, r_{952}$, der får ligningen til at passe for alle punkterne! Imidlertid er det en rigtig nyttig skrivemåde, da vi så får en notation for forskellen mellem datapunkterne og linjen.

Vi forestiller os, at de enkelte tal r_i er tilfældige. I statistikken kaldes tallene r_i også for **støjled**. De repræsenterer den "støj", der kommer til udtryk i data, når vi forsøger at beskrive data med en model. Residualerne kan være både positive og negative, og hvis et af r_i'erne er positivt, har det ingen indflydelse på, om de andre er positive eller negative. Denne antagelse om residualerne, dvs. støjleddene, er ensbetydende med at sige, at alle punkterne er placeret tilfældigt omkring linjen. Nogle ligger over, mens andre ligger under. Nogle ligger langt fra linjen, og andre ligger tættere på, og det hele skal være tilfældigt i den forstand, at der ikke er noget system i, hvordan punkterne varierer omkring linjen. For at se, om dette er tilfældet kan man f.eks. optegne residualerne r_i mod x_i–erne (alternativt kan man optegne r_i–erne mod $\hat{y}_i$–erne). En sådan tegning kaldes for et **residualplot**, og residualplottet for far-søn-datasættet findes på figur 6.

Videre tænker vi os, at linjen $\ell(x) = a \cdot x + b$ er ukendt for os. Den repræsenterer den underliggende biologiske sammenhæng mellem fædres og sønners højde. Vi kender ikke denne sammenhæng præcist, men er selvfølgelig interesserede i at sige noget om den.

Hele den ovenstående beskrivelse af, hvordan y–værdierne er blevet lavet ud fra x–værdierne og passende tilfældigheder, er det, der i statistiksprog kaldes **en statistisk model**.

Opgave 31. *Som vi vil se lidt nærmere på i afsnit 3, vil støjleddene ofte opføre sig, som om de er normalfordelte med middelværdi 0 og den samme*



Figur 6: Residualplot for den statistiske modellering af sønners højde ud fra fadernes højde.

spredning. Udfør selv en simulering ud fra en selvvalgt statistisk model med normalfordelte støjled: Vælg en ret linje og standardafvigelsen på støjleddene.

Vi har allerede udviklet en metode til at give vores bedste bud på den ukendte underliggende rette linje $\ell(x) = a \cdot x + b$. Nemlig ved at udregne hældningen vha. formlen for $\hat{a}$ og skæringen vha. formlen for $\hat{b}$. I statistiksprog kaldes a og b for modellens **parametre**, og $\hat{a}$ og $\hat{b}$ for **estimater** for parametrene. Vi kan ikke være sikre på, at estimaterne giver os præcis de rigtige, men ukendte, parametre a og b, som "blev brugt", da naturen frembragte sønnernes højder ud fra fædrenes via den rette linje og de tilfældige afvigelser. Men estimaterne $\hat{a}$ og $\hat{b}$ er altså vores bedste bud på hældningen og skæringen.

I næste afsnit vil vi prøve at give et bud på, hvor præcist dette bud er.

2.1 Usikkerhed på parameterestimaterne

I forrige afsnit gav vi en beskrivelse af, hvordan vi forestiller os, at naturen har "frembragt" de 952 far–søn–højder: Der er en, for os, ukendt ret linje $\ell(x) = a \cdot x + b$, hvor vi altså ikke kender værdien af a og b. Hver enkelt y_i er så lavet ud fra x_i ved formlen

$$y_i = a \cdot x_i + b + r_i$$

hvor r_i er et støjled, som naturen vælger helt tilfældigt og uafhængigt af, hvor stor eller lille x_i er. Når vi har vores datasæt, som for far–søn–datasættet er de 952 punkter (x_i, y_i), så kan vi ikke regne os baglæns frem til de rigtige værdier a og b, men vi kan give et godt bud ved at udregne $\hat{a}$ og $\hat{b}$.

Da naturen "tilsatte" tilfældighed, da den lavede y–værdierne ud fra x–værdierne, skal vi ikke regne med, at $\hat{a}$ og $\hat{b}$ bliver præcist de rigtige a og b, men forhåbentlig er de ikke så langt fra. Hvis vi havde et nyt datasæt, altså hvis Francis Galton også havde indsamlet højder fra nogle andre par af fædre og sønner, og hvis vi på baggrund af det nye datasæt udregnede $\hat{a}$ og $\hat{b}$, kunne vi heller ikke regne med, at de nye $\hat{a}$ og $\hat{b}$ ville være helt magen til de værdier, vi har udregnet på baggrund af det oprindelige datasæt. Der er altså en vis usikkerhed på $\hat{a}$ og $\hat{b}$, som kommer af hvilke målinger, vi tilfældigvis har fået med i undersøgelsen.

Omvendt har vi en fornemmelse af, at de fundne værdier af hældning og skæring nok ikke ville have været meget anderledes. Disse to tal er jo udtryk for nogle biologiske og sociologiske sammenhænge, som ikke afhænger af nøjagtig hvilke fædre og sønner, der blev undersøgt. Hvis vi gerne vil bruge $\hat{a}$ og $\hat{b}$ til at sige noget om disse sammenhænge, så er det nødvendigt at kunne sige noget præcist om, hvor meget disse størrelser kan forventes at variere, hvis man i stedet regner på baggrund af andre sæt af datapunkter.

Det er denne usikkerhed, vi vil prøve at beskrive i dette afsnit. Principielt kunne man finde ud af, hvor meget $\hat{a}$ og $\hat{b}$ varierer ved simpelthen at indsamle et nyt datasæt og beregne de tilhørende værdier af $\hat{a}$ og $\hat{b}$. For at få et godt billede af variationen ville det dog ikke være tilstrækkeligt blot at indsamle ét nyt datasæt. Derimod ville man være nødt til at indsamle adskillige nye datasæt. I praksis er dette ofte ikke muligt, og hvis vi vitterligt indsamlede adskillige nye datasæt, så vil det nok være bedre at samle disse datasæt til ét enkelt og større datasæt med henblik på at opnå endnu mere præcise værdier af $\hat{a}$ og $\hat{b}$.

Noget overraskende viser det sig, at man kan aflure variationen ved at tilføje ny variation. Dette kræver dog temmelig meget beregningskraft, men det er ikke noget problem med en computer ved hånden. Den næste opgave giver et indtryk af, hvordan dette kan gribes an, selvom metoden foreslået i opgaven ikke vil kunne løse vores problem.

Opgave 32. *Find ud af, hvor meget $\hat{a}$ og $\hat{b}$ varierer, når de er baseret på højdemålinger for 10 par af fædre og sønner. Dette gøres ved tilfældigt at udvælge 10 par fra datasættet en masse gange (hver gang udvælges 10 par uden tilbagelægning) og hver gang beregne $\hat{a}$ og $\hat{b}$. Se derefter på histogrammer for de udregnede værdier af $\hat{a}$ og $\hat{b}$ for at få en ide om, hvor meget de ændrer sig for hver gang, der benyttes et nyt datasæt med 10 punktpar.*

Opgavens resultat giver kun en ide om, hvor præcise $\hat{a}$ og $\hat{b}$ er, hvis de udregnes på baggrund af 10 datapunkter, men metoden kan ikke bruges til at undersøge variationen af $\hat{a}$ og $\hat{b}$, når disse baseres på hele det datasæt, der er til rådighed: Der er nemlig kun én måde at udtage 952 datapunkter ud fra 952 datapunkter, når udtagelsen sker uden tilbagelægning som beskrevet i opgaven. Dermed er der ikke nogen variation at tilføje via udvælgelse af delmængden. Der er imidlertid en anden måde, hvorpå vi kan generere nye datasæt, der strukturelt set ligner det originale datasæt samtidig med, at variationen afsløres. Den metode, vi vil beskrive i det følgende, kaldes **omrøring** (og på engelsk er den kendt som **bootstrap**). Til dette får vi brug for lidt ekstra notation. Vi lader $\hat{r}_1, \ldots, \hat{r}_n$ være de støjled, der opstår, hvis man prøver at forudsige y–værdien ud fra x–værdierne ved brug at den bedste rette linje $\ell(x) = \hat{a} \cdot x + \hat{b}$. Eller sagt på en anden måde: $\hat{r}_i$ er den lodrette afstand mellem y_i og værdien af linjen $\ell(x) = \hat{a} \cdot x + \hat{b}$ udregnet i punktet x_i. Bemærk, at det er disse lodrette afstande, dvs. residualerne, som indgår i formlen (4) for residualspredningen. Det eneste nye er, at vi har fundet på en notation for disse værdier. Der gælder altså, at

$$\hat{r}_1 = y_1 - (\hat{a} \cdot x_1 + \hat{b}), \qquad \ldots, \qquad \hat{r}_n = y_n - (\hat{a} \cdot x_n + \hat{b})$$

Disse n ligninger kan hurtigt omskrives til

$$y_1 = \hat{a} \cdot x_1 + \hat{b} + \hat{r}_1, \qquad \ldots, \qquad y_n = \hat{a} \cdot x_n + \hat{b} + \hat{r}_n$$

og viser altså, hvordan datapunkterne $(x_1, y_1), \ldots, (x_n, y_n)$ er sammenkædet med støjleddene $\hat{r}_1, \ldots, \hat{r}_n$.

Husk på, at vores model i formel (11) siger, at naturen kommer fra x_i til y_i ved at bruge den rette linje og så tilsætte noget tilfældig støj. I vores forsøg på at lave nye datasæt vil ideen være i passende forstand at erstatte støjledene $\hat{r}_i$ i ligningerne ovenfor med nye tilfældige støjled. Da vi ikke har en masse helt nye og tilfældige støjled til rådighed, vil vi løse dette ved på passende tilfældig vis at trække fra de støjled, vi allerede har.

Vi vil beskrive fremgangsmåden ved at se på den lille udgave af far–søn–datasættet, som kun består af 10 datapunkter. Vi lader altså for et øjeblik som om, at de 10 datapunkter udgør *hele* datasættet, og at vi ikke har mere data at trække fra. Datasættet med 10 far–søn–par, vi vil kigge på, er følgende

	x_i	y_i
1	183,9	187,8
2	172,1	167,4
3	169,5	172,4
4	176,5	183,7
5	179,7	168,7
6	175,2	168,3
7	166,5	169,9
8	177,0	170,5
9	172,4	170,8
10	176,5	183,5

Opgave 33. *Lav et plot, der viser disse 10 punkter og vis, at den bedste rette linje gennem punkterne har hældning* $\hat{a} = 0{,}83$ *og skæring* $\hat{b} = 28{,}3$.

I opgave 33 blev det vist, at den bedste rette linje gennem datasættets 10 punkter er givet ved

$$\ell(x) = 0{,}83 \cdot x + 28{,}3$$

Ud fra dette kan vi regne de 10 støjled $\hat{r}_1, \ldots, \hat{r}_{10}$. Resultatet kan ses i det nedenstående skema

	x_i	y_i	$\hat{r}_i$
1	183,9	187,8	6,01
2	172,1	167,4	−4,53
3	169,5	172,4	2,63
4	176,5	183,7	8,09
5	179,7	168,7	−9,58
6	175,2	168,3	−6,22
7	166,5	169,9	2,63
8	177,0	170,5	−5,53
9	172,4	170,8	−1,39
10	176,5	183,5	7,89

Opgave 34. *Prøv selv at regne efter, at de 10 støjled ser ud som angivet i skemaet.*

Næste skridt er, at vi fra listen med de 10 støjled

$$6{,}01 \quad -4{,}53 \quad 2{,}63 \quad 8{,}09 \quad -9{,}58 \quad -6{,}22 \quad 2{,}63 \quad -5{,}53 \quad -1{,}39 \quad 7{,}89 \tag{12}$$

udtrækker 10 tal tilfældigt med tilbagelægning:

$$7{,}89 \quad -4{,}54 \quad 2{,}63 \quad 2{,}63 \quad 8{,}09 \quad -9{,}58 \quad -4{,}54 \quad 6{,}01 \quad 6{,}01 \quad 7{,}89$$

Vi vælger at kalde disse 10 nye støjled for $s_1, \ldots, s_{10}$. Nu bruger vi så listen med nye støjled til at lave en liste med 10 nye søn–værdier, som vi nu vælger at kalde $z_1, \ldots, z_{10}$ ved at bruge formlen

$$z_i = \hat{a} \cdot x_i + \hat{b} + s_i \tag{13}$$

hvor altså $\hat{a} = 0{,}83$ og $\hat{b} = 28{,}3$ er bedste hældning og skæring udregnet på baggrund af de (oprindelige) 10 datapunkter. Ved at bruge formlen for alle 10 datapunkter, fås nu

	x_i	s_i	z_i
1	183,9	7,89	189,7
2	172,1	−4,54	167,4
3	169,5	2,63	172,4
4	176,5	2,63	178,2
5	179,7	8,09	186,4
6	175,2	−9,58	164,9
7	166,5	−4,54	162,7
8	177,0	6,01	182,0
9	172,4	6,01	178,2
10	176,5	7,89	183,5

De 10 par $(x_1, z_1), \ldots, (x_{10}, z_{10})$ udgør nu vores nye datasæt, og i dette datasæt kan vi finde et bud på den bedte rette linje: Vi udregner simpelthen $\hat{a}$ og $\hat{b}$ på baggrund af de 10 nye datapunkter. Her fås, at $\hat{a} = 1{,}53$ og $\hat{b} = -90{,}5$. Bemærk, at disse to tal, som er regnet på baggrund af et datasæt, som minder en del om det oprindelige, ser noget anderledes ud end de to værdier, 0,83 og 28,3, vi fik i første omgang.

Opgave 35. *Prøv selv at frembringe 10 "nye" støjled ved at trække med tilbagelægning fra listen i (12). Regn derefter de tilhørende 10 søn–værdier ved at bruge formlen (13). Tegn et plot af de 10 datapunkter bestående af de 10 (oprindelige) far–værdier sammen med de 10 nye søn–værdier. Regn også hældning og skæring for den bedste rette linje gennem punkterne. Prøv at gentage alt dette ved at frembringe endnu 10 nye støjled, og se så, hvor meget de udregnede $\hat{a}$ og $\hat{b}$ ændrer sig.*

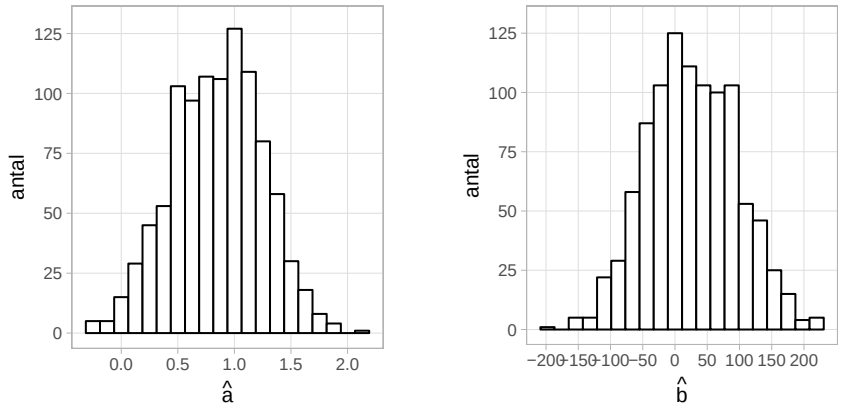
Nu gentager vi det hele 1000 gange:

1. Vi udtrækker 10 nye støjled fra listen (12). Udtrækningen sker med tilbagelægning.

2. Ud fra de 10 nye støjled regnes 10 nye søn–værdier med formlen (13), hvor $\hat{a} = 0{,}83$ og $\hat{b} = 28{,}3$ er hældning og skæring udregnet på baggrund af de oprindelige 10 punkter.

3. Derved fås en liste med 10 far–søn–par, hvor søn–værdierne er de nye værdier udregnet i punkt 2. Ud fra denne liste udregnes hældning $\hat{a}$ og skæring $\hat{b}$ for den bedste rette linje gennem punkterne.

Samlet giver dette en liste på 1000 hældninger og 1000 skæringer, som kan hjælpe os med at vurdere usikkerheden på de oprindelige værdier $\hat{a} = 0{,}83$ og $\hat{b} = 28{,}3$. Vi kan tegne histogrammer over disse $\hat{a}$–værdier og $\hat{b}$–værdier. Disse histogrammer kan ses på figur 7.



Figur 7: Histogrammer over $\hat{a}$–værdierne og $\hat{b}$–værdierne, der er opnået via omrøringsmetoden lavet på baggrund af det lille datasæt med 10 observationspunkter.

En tilbundsgående matematisk analyse (under passende forudsætninger) af denne metode kan godtgøre, at disse histogrammer vil give et retvisende billede af, hvor meget $\hat{a}$ og $\hat{b}$ ville variere, hvis vi vitterlig var i stand til at indsamle massevis af nye datasæt af samme størrelse som det aktuelle datasæt. I stedet for at gå i detaljer med dette (hvilket også vil være alt for krævende på gymnasialt niveau), vil vi bare benytte resultaterne til at udtale os om usikkerheden på $\hat{a}$ og $\hat{b}$.

Hvis vi først kigger på histogrammet for $\hat{a}$, kan vi se, at stort set alle værdierne ligger mellem ca. 0 og ca. 1.7. Det kan vi bruge til at sige, at vores bedste bud på a – baseret på de første 10 målepunkter – er den oprindeligt udregnede værdi 0,83, og vi er i hvert fald rimelig sikre på, at a ligger mellem 0 og 1,7.

Tilsvarende kan vi ud fra histogrammet for $\hat{b}$ se, at værdierne generelt ligger mellem -150 og 200, så vores konklusion om b er, at vores bedste bud på værdien – baseret på de første 10 målepunkter – er den oprindeligt udregnede værdi 28,3, og at vi i hvert fald er rimelig sikre på, at b ligger mellem -150 og 200.

Nu kan man sige, at det måske ikke er et specielt præcist resultat, at vi bare kan sige, at a ca. ligger 0 og 1,7, og at b ca. ligger mellem -150 og 200. Omvendt skal vi huske, at dette er, hvor præcist vi kan sige det, når vi kun har haft 10 datapunkter til vores rådighed.

Det ville være meget mere interessant at undersøge, hvor præcist vi kan give et bud på a og b, når vi har alle 952 datapunkter til vores rådighed. For at gøre dette, bruger vi præcist den samme fremgangsmåde som før. Husk på, at den bedste rette linje gennem alle punkterne er blevet regnet til at være

$$\ell(x) = 0{,}613 \cdot x + 67{,}0$$

Derudfra kan vi regne de 952 støjled $\hat{r}_1, \ldots, \hat{r}_{952}$ med formlen

$$\hat{r}_i = y_i - (0{,}613 \cdot x_i + 67{,}0)$$

Opgave 36. *Lav selv en liste med alle disse støjled. Det første støjled skulle meget gerne være ca. (afhængigt af afrundinger) 1,89, mens det sidste skal være 2,77.*

Herefter bruger vi den samme opskrift som før og gentager følgende 1000 gange:

1. Vi udtrækker 952 nye støjled fra listen $\hat{r}_1, \ldots, \hat{r}_{952}$ af støjled fra opgave 36. Udtrækningen sker med tilbagelægning.

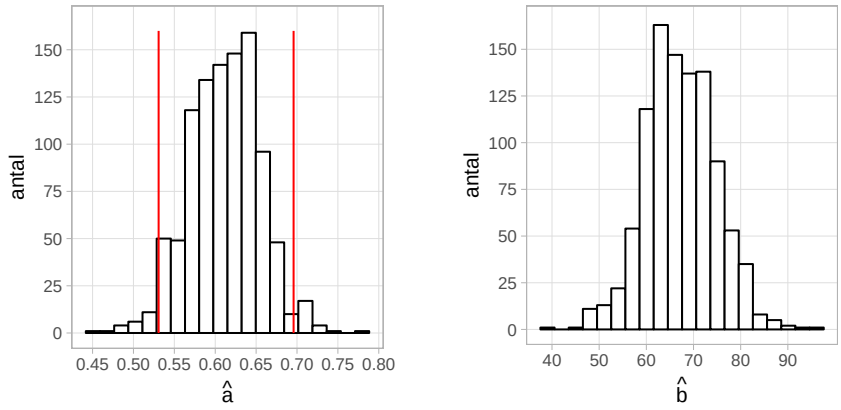
2. Ud fra de 952 nye støjled regnes 952 nye søn–værdier med formlen (13), hvor $\hat{a} = 0{,}613$ og $\hat{b} = 67{,}0$ er hældning og skæring udregnet på baggrund af de oprindelige 952 punkter.

3. Derved fås en liste med 952 far–søn–par, hvor søn–værdierne er de nye værdier udregnet i punkt 2. Ud fra denne liste udregnes hældning $\hat{a}$ og skæring $\hat{b}$ for den bedste rette linje gennem punkterne.

Dette giver altså igen en liste med 1000 hældninger og 1000 skæringer. På figur 8 er tegnet histogrammer over alle disse $\hat{a}$–værdier og $\hat{b}$–værdier.

Opgave 37. *Prøv selv at fremstille lignende histogrammer ved at følge proceduren beskrevet ovenfor.*

Fra histogrammet for $\hat{a}$ kan vi nu se, at langt de fleste værdier ligger mellem 0,50 og 0,72. Det vil sige, at vores bedste bud på det rigtige a er 0,613, som er hældningen på den bedste rette linje gennem de observerede punkter, og at vi i hvert fald er temmelig sikre på, at a ligger mellem 0,50 og 0,72. Bemærk, at vi nu er noget mere sikre på, hvor det rigtige a ligger, end vi var før, hvor vi kun havde 10 datapunkter til vores rådighed.



Figur 8: Histogrammer over $\hat{a}$–værdierne og $\hat{b}$–værdierne, der er opnået via omrøringsmetoden lavet på baggrund af det fulde datasæt. De røde linjer på histogrammet til venstre indrammer området for de 95% midterste $\hat{a}$–værdier.

Opgave 38. *Diskuter, om det virker rimeligt, at vi kan give et mere præcist bud på a, når vi har et større datamateriale til vores rådighed!*

Fra histogrammet for $\hat{b}$ ser vi, at langt de fleste værdier ligger mellem 50 og 85, hvilket vi kan bruge til at konkludere, at vores bedste bud på b er 67,0, og at vi i hvert fald er rimelig sikre på, at det rigtige b ligger mellem 50 og 85. Bemærk, at også dette interval er meget smallere, end hvad vi så, da vi vurderede usikkerheden på $\hat{b}$ beregnet ud fra kun 10 datapunkter.

Opgave 39. *I opgave 8 skulle man bestemme det bedste bud på gennemsnitshøjden for sønner, hvis fædre er 168 cm høje, og for sønner, hvis fædre er 188 cm høje. Brug nu omrøringsmetoden til at fremstille histogrammer, der viser usikkerheden på disse estimater.*

De to histogrammer i figur 8 synes, pånær tilfældig variation der kommer fra omrøringen, at være symmetriske omkring estimaterne $\hat{a}$ og $\hat{b}$. Lad os kigge på histogrammet for $\hat{a}$. Det interval, der indeholder de 95% miderste estimater for a efter omrøringen, altså hvor man fjerner de 2,5% mindste og de 2,5% største værdier, kaldes for et **95% konfidensinterval for** a. For den omrøring, som vi har lavet, går dette fra 0,53 til 0,70 og er indrammet med røde linjer på histogrammet. Vi bemærker, at man kan finde eksplicitte matematiske formler for sådanne konfidensintervaller. Disse formler involverer såkaldte t-fordelinger, som fx beskrevet i lærebøgerne [1] og [5]. Disse bøger er dog skrevet for universitetet, og det vil blive for omfattende at komme nærmere ind på disse emner i dette notat.

2.2 Er der en sammenhæng?

I den statistiske model

$$y_i = a \cdot x_i + b + r_i$$

beskriver tallet a sammenhængen mellem x–erne og y–erne. Hvis $a > 0$, så vil y_i typisk være større, når x_i er større. Og hvis $a < 0$, så vil y_i typisk være mindre, når x_i er større. Men hvis $a = 0$, så står der bare $y_i = b + r_i$ tilbage. Det betyder, at der ikke længere er nogen samvariation mellem x–erne og y–erne. I eksemplet med højden af fædre og deres sønner svarer dette til, at man ikke kan forbedre sit gæt på en søns højde ved at kende faderens højde: Simpelthen fordi sønnens højde udelukkende afhænger af b og naturens tilfældige støj, hvorimod faderens højde tilsyneladende ikke længere spiller en rolle.

Svaret på spørgsmålet om, hvorvidt a er 0 eller ej, er derfor af kvalitativ betydning for fortolkningen af datasættet: Hvis svaret er, at a *ikke* er 0, ved vi jo, at fædrenes højde faktisk har en betydning for deres sønners højde. Man kan imidlertid ikke afgøre, om $a = 0$ ved at beregne $\hat{a}$, og se om denne størrelse er 0! Grunden til dette er, at estimatet $\hat{a}$ afhænger af det specifikke datasæt, vi har til rådighed. Som vi så i forrige afsnit, er der en vis usikkerhed på $\hat{a}$: Hvis vi fik et andet datasæt, ville $\hat{a}$ blive lidt anderledes, selv om den underliggende virkelighed (altså det rigtige a, som naturen har brugt til at lave begge datasæt) forbliver den samme. I praksis er det forøvrigt også sådan, at $\hat{a}$ beregnet udfra et rigtigt datasæt *næsten altid* er forskellig fra 0.

Vi vil formulere vores spørgsmål som en såkaldt **statistisk hypotese**. Den skriver vi på følgende måde

$$H_0 \ : \ a = 0$$

Bemærk, at H_0 svarer til situationen, at der ikke er sammenhæng mellem x–værdierne og y–værdierne. Hvis hypotesen viser sig *ikke* at være sand, ender vi med den **alternative** hypotese, som siger, at a er forskellig fra 0. Den alternative hypotese skriver vi derfor som

$$H_1 \ : \ a \neq 0$$

Vores tilgang til de to mulige hypoteser H_0 og H_1 kan illustreres med situationen ved en retssag. H_0 svarer til, at den anklagede er uskyldig, mens H_1 svarer til, at den anklagede er skyldig. Som i en retssag lader vi tvivlen komme den anklagede til gode, så medmindre vi har stærke beviser for, at den anklagede er skyldig, så vælger vi at acceptere H_0 (frikende den anklagede). Vi forkaster altså kun H_0, hvis vi har rigtigt gode beviser for, at H_0 er forkert.

Opgave 40. *I det foregående afsnit blev der vist en metode til at undersøge usikkerheden på estimatet $\hat{a}$. Diskuter, om (og i givet fald hvordan) dette kan bruges til at vurdere, om $a = 0$ er i overensstemmelse med datasættet.*

Lad os et øjeblik forestille os, at hypotesen H_0 om, at det bagvedliggende a er lig 0, faktisk er rigtig. Så er der altså ingen sammenhæng, og x–værdierne må siges at være helt unyttige til at forudsige y–værdierne. Så må værdien $\hat{a} = 0{,}613$ bare være et resultat af tilfældigheder, og hvis Francis Galton havde målt på et helt andet udvalg af fædre og sønner, havde den estimerede hældning måske været negativ. Det er altså et spørgsmål om, hvor stort $\hat{a}$–estimatet kan blive som resultat af tilfældigheder, hvis det ægte bagvedliggende a er 0. Desværre har vi ikke massevis af fædre–sønner–datasæt til vores rådighed, hvor vi ved, at den rigtige parameter faktisk er 0. Hvis vi havde haft det, kunne vi ellers for hvert datasæt have estimeret hældningsparameteren og så derudfra have vurderet, om $\hat{a} = 0{,}613$ er en urimeligt høj værdi.

Med et lille trick (med den samme overordnede tankegang som i det forrige afsnit) kan vi imidlertid alligevel lave en masse datasæt, der næsten ligner nye datasæt, men som faktisk er lavet ud fra det datasæt, vi allerede har. Hvis $a = 0$, og fædrenes højder ikke har betydning for sønnernes, så gør det jo heller ikke nogen forskel for datasættet, hvis vi bytter tilfældigt rundt på fædrene, så de bliver matchet med helt andre sønner!

For at illustrere, hvad vi mener, viser tabellen herunder til venstre de oprindelige 10 første datapar, mens tabellen til højre viser de samme 10 datamålinger, men med nye parringer, idet der er byttet tilfældigt rundt på fædremålingerne.

fars højde	søns højde	fars højde	søns højde
186,9	183,4	179,4	183,4
184,6	172,0	178,4	172,0
185,0	179,0	186,9	179,0
182,1	165,4	173,4	165,4
179,4	155,4	176,5	155,4
178,4	160,3	179,7	160,3
179,7	165,0	184,6	165,0
179,7	168,7	179,7	168,7
176,5	160,3	185,0	160,3
173,4	157,5	182,1	157,5

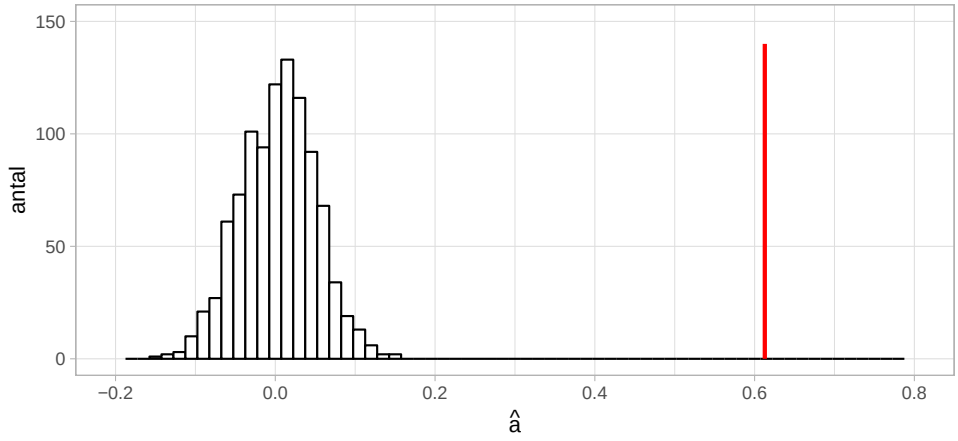
Opgave 41. *Prøv selv at lave tilfældige ombytninger på fædremålingerne for de 10 første datapunkter – ligesom det er gjort ovenfor.*

I vores analyse bytter vi ikke kun rundt på de 10 første fædremålinger, men på dem allesammen. Ved at lave en tilfældig ombytning af alle fædrenes målinger og så estimere den bedste rette linje hørende til de nye punkter, kunne man få hældningen $\hat{a} = 0,031$. Altså et tal, der tilsyneladende er meget mindre end det oprindeligt estimerede. Men dette tal er måske også bare en tilfældighed, så 0,613 alligevel ikke er usædvanligt stort. Vi gentager derfor processen 10 gange, hvilket giver tallene

−0,033	−0,073	−0,053	0,018	0,032
0,007	0,005	0,085	0,060	−0,030

Opgave 42. *Prøv også selv at bytte tilfældigt rundt på fædremålingerne i hele datasættet (ikke længere bare de 10 første datapunkter), og udregn hældningen $\hat{a}$ for den bedste rette linje. Prøv at gentage ombytningen, og regn hældningen igen.*

Hvis vores hypotese H_0 om, at $a = 0$ havde været rigtig, skulle det oprindeligt estimerede $\hat{a} = 0,613$ have været af samme størrelse som disse tal: Alle ombytninger af fædrenes højder skulle i princippet være lige så gode (eller dårlige, om man vil) til at forudsige sønnernes højder som i det oprindelige datasæt. Dog er det ret tydeligt, at alle tallene er *meget* tættere på 0 end 0,613. Så meget tyder på, at hypotesen ikke kan være rigtig - simpelthen fordi, 0,613 er et usædvanligt stort tal.



Figur 9: Histogram over alle $\hat{a}$–værdierne, der er opnået via 1000 gentagelser af ombytningsmetoden. Den røde linje markerer $\hat{a} = 0,613$ fra det oprindelige datasæt.

For at være helt sikre laver vi 1000 estimerede $\hat{a}$–værdier i stedet for blot 10. Dette gøres ved at lave 1000 ombytninger af x–værdierne og for hver

ombytning at udregne hældningen på den bedste rette linje. En liste med disse 1000 tal ville blive meget lang, så vi har i stedet tegnet et histogram, der viser, hvordan tallene fordeler sig. Dette histogram ses på figur 9, hvor den lodrette røde linje markerer den oprindelige $\hat{a}$–værdi på 0,613. Af histogrammet fremgår det endnu mere tydeligt: Den observerede værdi på 0,613 er helt urealistisk stor, *hvis* hypotesen H_0 at $a = 0$ skulle være rigtig. Det "normale" område for $\hat{a}$–værdierne, hvis $a = 0$ var rigtig, er mellem $-0{,}2$ og 0,2, og der er *ingen* af de andre $\hat{a}$–værdier, der er bare tilnærmelsesvist så langt fra 0, som 0,613. Der er nu ingen tvivl om vores konklusion: Hypotesen H_0 kan ikke være sand, og forkastes derfor. I stedet må vi konkludere, at den alternative hypotese H_1 er sand, og derved har vi påvist, at fædrenes højder har betydning for sønnernes højder.

I denne situation var der ikke meget tvivl om konklusionen, da den oprindeligt estimerede værdi af $\hat{a}$ var så meget længere fra 0 end alle de andre. Nu kan man så overveje, hvor meget anderledes billedet skulle have set ud for, at vi ikke havde forkastet H_0. En almindelig praksis i statistik er, at hvis mindst 5% af $\hat{a}$–værdierne fremkommet ved ombytning havde været numerisk (vi ser altså bort fra fortegnet) større end 0,613, så havde vi accepteret H_0, og hvis færre end 5% af dem havde været numerisk større, havde vi forkastet H_0. I vores situation var der (præcis) 0% af $\hat{a}$–værdierne, som var numerisk større end den oprindeligt estimerede værdi på 0,613, og derfor forkastede vi hypotesen. Jo mindre procentdelen er, jo stærkere er konklusionen om, a ikke er 0, og at der altså er en sammenhæng mellem x–værdier og y–værdier.

Metoden, vi har fulgt til at teste, om hypotesen H_0 er sand, kan beskrives med den følgende algoritme. Vælg f.eks. $N = 1000$ (eller som et tilsvarende stort tal) og gør derefter følgende:

1. Beregn $\hat{a}$ ud fra $(x_1, y_1), \ldots, (x_n, y_n)$.

2. For hvert $j = 1, \ldots, N$ lav en tilfældig omrokering af $x_1, \ldots, x_n$. Kald denne for $z_1, \ldots, z_n$, og beregn den bedste hældning $\hat{a}_j$ hørende til punkterne $(z_1, y_1), \ldots, (z_n, y_n)$.

3. Lad tallet $p \in [0, 1]$ være andelen af gange, hvor $|\hat{a}_j|$ er mindst ligeså stor som $|\hat{a}|$. Altså antallet af gange, hvor $\hat{a}_j$ ligger mindst lige så langt væk fra 0, som det oprindelige $\hat{a}$ gør.

I statistisk jargon kaldes tallet p for en ***p*-værdi**. Hvis p er lille, så er det få af $\hat{a}_j$'erne, der er mindst lige så langt fra 0 som $\hat{a}$ (det svarer til at sige, at $|\hat{a}_j|$ er mindst lige så stor som $|\hat{a}|$). Generelt kan det om $\hat{a}$ siges, at jo længere værdien er fra 0, jo stærkere er sammenhængen mellem y–erne og x–erne. Konklusionen bliver dermed, at hvis p er lille, så indikerer dette, at der er

en stærkere sammenhæng mellem y–erne og x–erne i det originale datasæt end i de datasæt, hvor vi har brudt sammenhængen ved at lave en tilfældig omrokering af x'erne. Altså

$$\text{lille p-værdi} \implies \text{indikation for samvariation mellem } y\text{–erne og } x\text{–erne.}$$

Jo mindre p er, jo stærkere er denne indikation. Som nævnt ovenfor har man historisk set ofte brugt $p < 0{,}05$ som kriterie for, at der er samvariation mellem y–erne og x–erne. Denne grænse mellem *små* og *store* p-værdier, historisk set altså 0,05, kaldes for **signifikansniveauet**. Det skal dog bemærkes, at det klassiske valg af 0,05 som signifikansniveau er helt arbitrært.

Opgave 43. *Prøv at bruge ombytningsmetoden til at teste, om det er muligt, at $a = 0$, hvis der bare kigges på det lille datasæt med 10 datapunkter, som vi også benyttede i forrige afsnit:*

	x_i	y_i
1	183,9	187,8
2	172,1	167,4
3	169,5	172,4
4	176,5	183,7
5	179,7	168,7
6	175,2	168,3
7	166,5	169,9
8	177,0	170,5
9	172,4	170,8
10	176,5	183,5

Følg altså beskrivelsen for dette lille datasæt:

1. *Beregn $\hat{a}$ udfra $(x_1, y_1), \ldots, (x_{10}, y_{10})$ (vi har faktisk allerede udregnet, at $\hat{a} = 0,83$ i forrige afsnit).*

2. *For hvert $j = 1, \ldots, N$ lav en tilfældig omrokering af $x_1, \ldots, x_{10}$. Kald denne for $z_1, \ldots, z_{10}$, og beregn den bedste hældning $\hat{a}_j$ hørende til punkterne $(z_1, y_1), \ldots, (z_{10}, y_{10})$.*

3. *Lad tallet $p \in [0, 1]$ være andelen af gange, hvor $\hat{a}_j$ ligger mindst lige så langt væk fra 0 som 0,83. Altså andelen af gange, hvor det enten gælder, at $\hat{a}_j \leq -0{,}83$ eller at $\hat{a}_j \geq 0{,}83$.*

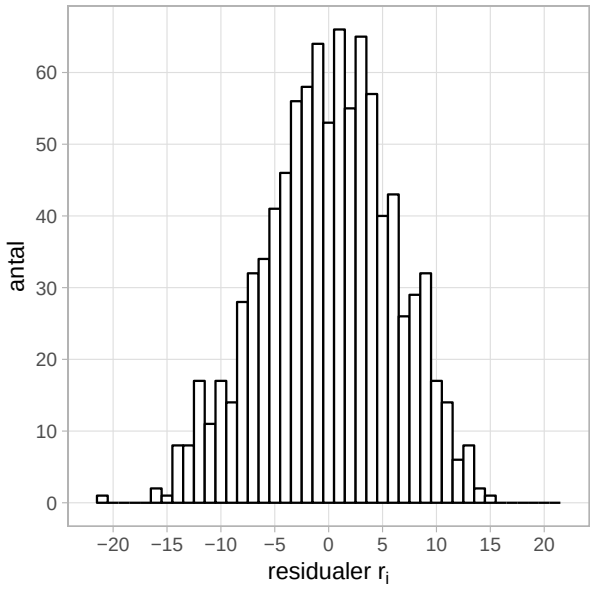
Denne gang viser det sig, at p bliver større end 5%. Overvej, hvordan det kan være, og hvad det siger om vores hypoteser H_0 og H_1.

3 Normalfordelingen

I afsnit 1 kiggede vi på sammenhængen mellem fædres og sønners højder, og vi så, at der var en voksende tendens, som kunne beskrives ved en ret linje. Vi regnede os frem til den linje, der passede bedst med punkterne ved at vælge den linje, som gjorde summen af de kvadrerede residualer mindst mulig. Husk på, at residual bare er et andet navn for den lodrette afstand mellem punktet og den rette linje. Resultatet blev linjen

$$\ell(x) = 0{,}613 \cdot x + 67{,}0$$

Vi udregnede endvidere residualspredningen $\hat{\sigma}$, som kan forstås som et mål for, hvor store residualerne er (regnet numerisk). Så en stor residualspredning betyder, at residualerne generelt set er numerisk større (enten positive eller negative), altså at punkterne generelt set ligger længere væk fra den rette linje. Vi vil se lidt nærmere på, hvordan residualerne samlet set opfører sig. Vi vil studere det, der kaldes residualernes **fordeling**.



Figur 10: Histogram over alle residualerne, der fremkommer ved at regne de lodrette afstande mellem de 952 punkter for far–søn–målingerne og den bedste rette linje gennem punkterne.

På figur 10 er der tegnet et histogram over alle residualværdierne.

Opgave 44. *Prøv selv at tegne et histogram over residualernes værdier. Bemærk, at det ikke nødvendigvis kommer til at ligne histogrammet på figur 10 fuldstændigt. Det kommer an på, hvor brede søjlerne i histogrammet*

er lavet. På figur 10 er histrogrammet konstrueret, så alle søjler har bredde 1.

Histogrammet har – pånær nogle enkelte svipsere – overordnet set en "klokkeagtig" form, der er symmetrisk omkring 0. De fleste residualværdier er altså relativt tæt på 0, der er ca. lige mange på den negative side som på den positive side, og dem på den negative side er fordelt nogenlunde som dem på den positive – bare spejlvendt. Vi kan også bemærke, at langt de fleste residualer ligger mellem -12 og 12.

Opgave 45. *Tænk over, hvordan det passer med den tommelfingerregel, der blev omtalt i afsnit 1: De fleste lodrette afstande må forventes at være numerisk mindre end 2 gange residualspredningen.*

Nu laver vi et forsøg, der måske kan virke en lille smule arbitrært: Vi indtegner grafen for følgende funktion i histogrammet

$$\varphi(x) = 63 \cdot e^{-\frac{x^2}{2 \cdot \hat{\sigma}^2}}$$

Her er $\hat{\sigma}$ residualspredningen, som blev udregnet til at være 6,0. Konstanten 63 foran eksponetialfunktionen er udregnet ved $\frac{952}{\sqrt{2 \cdot \pi} \cdot \hat{\sigma}}$, men det vil vi ikke interessere os nærmere for i dette notat. Det væsentlige er, at funktionen er givet som en passende konstant gange eksponentialfunktionen. På figur 11 er grafen for funktionen $\varphi(x)$ indtegnet sammen med histogrammet. Det interessante er, at histogrammet og funktionen $\varphi(x)$ faktisk følger hinanden ret godt. Det tyder på, at residualerne tilnærmelsesvist er fordelt som i det, der kaldes **normalfordelingen**.

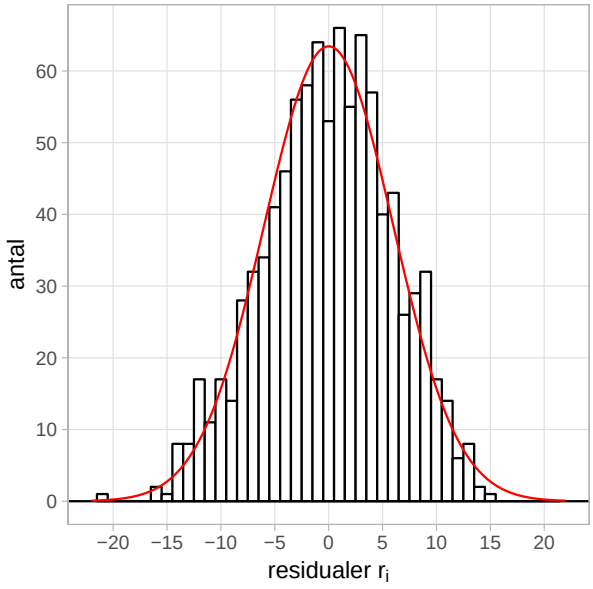
Mere generelt siges en samling af talværdier at følge normalfordelingen, hvis et histogram over dem nogenlunde følger en funktion $\varphi(x)$ på formen

$$\varphi(x) = K \cdot e^{-\frac{x^2}{2 \cdot C}} ,$$

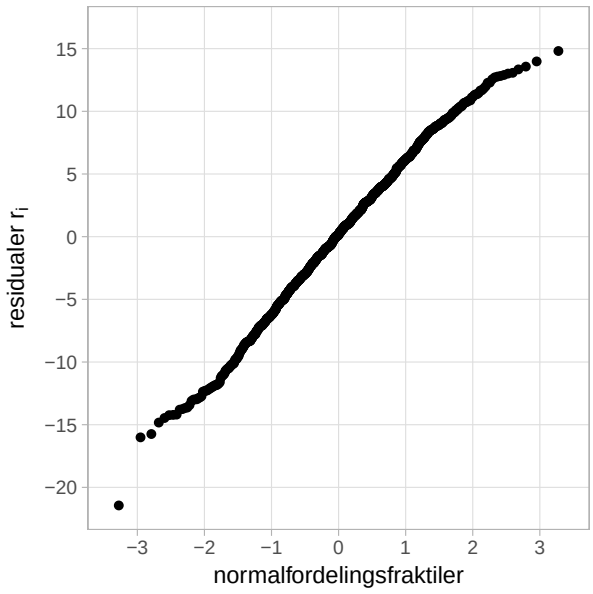
hvor K og C er passende konstanter. Det fungerede i vores histogram, hvis K blev sat til 63, og C blev valgt som $\hat{\sigma}^2 = 6^2 = 36$.

At residualerne kan siges tilnærmelsesvist at være normalfordelte, sker overraskende tit, og for meget forskelligartede datasæt. Dette er til stor glæde både for *statistikeren*, der så kan beskrive mange forskellige slags datasæt med den samme type af modeller, og for *matematikeren*, der kan give dybtliggende sandsynlighedsteoretiske argumenter for, hvorfor det er tilfældet.

Der findes en anden – og også lettere – måde at efterse, om residualerne er normalfordelte, end ved at tegne histogrammet sammen med en passende valgt eksponentialfunktion. Det gøres ved at tegne et **normalfordelingsfraktildiagram**. Hvis punkterne på sådan en tegning stort set ligger på en



Figur 11: Histogram over alle residualerne sammen med grafen for funktionen $\varphi(x)$. Grafen for $\varphi(x)$ er indtegnet med rødt.



Figur 12: Normalfordelingsfraktildiagram for residualerne for far–søn–datasættet.

ret linje, så indikerer dette, at residualerne godt kan antages at være normalfordelte. Et normalfordelingsfraktildiagram kan let tegnes i et matematisk værktøjsprogram.

På figur 12 er der tegnet et normalfordelingsfraktildiagram for residualerne for far–søn–datasættet. Punkterne i diagrammet ligger med enkelte undtagelser pænt på en ret linje, hvormed vi endnu engang kan konstatere, at residualerne er normalfordelte.

4 Eksponentiel regression og potensregression

Indtil videre har vi undersøgt statistiske metoder for situationen, hvor sammenhængen i datapunkterne $(x_1, y_1), \ldots, (x_n, y_n)$ kan beskrives som *"tilfældig variation omkring en ret linje"*. Vi vil nu undersøge muligheden for at udvide disse metoder til situationen, hvor sammenhængen enten kan beskrives ved en *ekponentialfunktion* eller en *potensfunktion*.

Antag, at alle y–værdierne er større end 0, således at vi kan tage logaritmen (i det følgende bruger vi den naturlige logaritme ln, men faktisk er det ligegyldigt hvilken logaritme, man vælger at bruge). Hvis vi bruger den statistiske model i ligning (11) mellem x–værdierne og logaritmen af y–værdierne, så får vi følgende sammenhæng mellem x–værdierne og y–værdierne

$$\ln(y_i) = a \cdot x_i + b + r_i$$

Hvis vi derefter tager eksponentialfunktionen på begge sider af lighedstegnet og introducerer parameteren $c = \exp(b)$ og fejlene $f_1 = \exp(r_1), \ldots, f_n = \exp(r_n)$, så får vi

$$\begin{aligned} y_i &= \exp(a \cdot x_i + b + r_i) \\ &= \exp(a \cdot x_i) \cdot \exp(b) \cdot \exp(r_i) \\ &= c \cdot \exp(a \cdot x_i) \cdot f_i \end{aligned}$$

Dette er nu en **statistisk model** for en eksponentiel sammenhæng mellem y–værdierne og x–værdierne, hvor fejlene f_i skal ganges på i stedet for at lægges til. Selvom det ser kompliceret ud, så ved vi allerede, hvordan vi kan regne i denne model. Nemlig ved at lave simpel lineær regression af $\ln(y_i)$'erne på x_i'erne.

Før vi afprøver dette på et datasæt, illustrerer følgende opgave, hvorledes samme tilgang kan bruges til at lave en **statistisk model** for en potenssammenhæng mellem y–værdierne og x–værdierne.

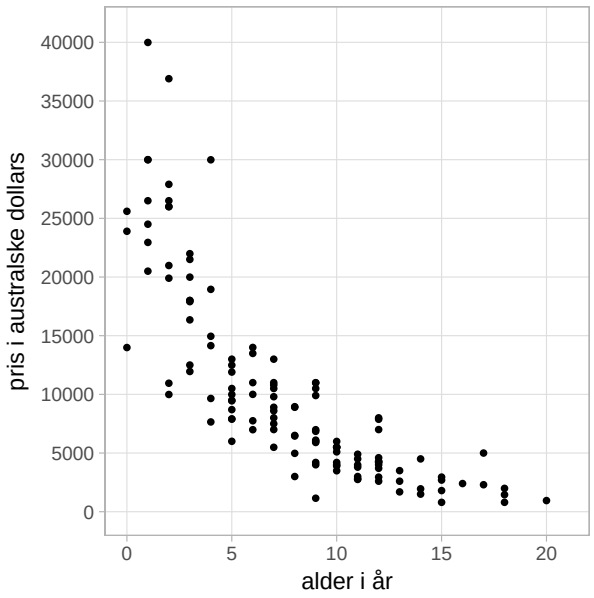
Opgave 46. *Antag, at alle x–værdierne og alle y–værdierne er større end 0. Antag videre, at sammenhængen mellem $\ln(y_i)$'erne og $\ln(x_i)$'erne kan beskrives via en simpel lineær regression*

$$\ln(y_i) = a \cdot \ln(x_i) + b + r_i$$

Hermed mener vi, at $\ln(y_i)$*'erne varierer omkring en ret linje mht.* $\ln(x_i)$*'erne. Vis, hvorledes dette fører til en potenssammenhæng mellem y–værdierne og x–værdierne.*

4.1 Eksempel på en eksponentiel regression

Figur 13 viser sammenhængen mellem pris og alder for 124 brugte biler af mærket *Mazda* solgt i Melbourne, Australien, i året 1991.



Figur 13: 124 sammenhørende par af pris og alder for brugte Mazda'er solgt i Melbourne i 1991.

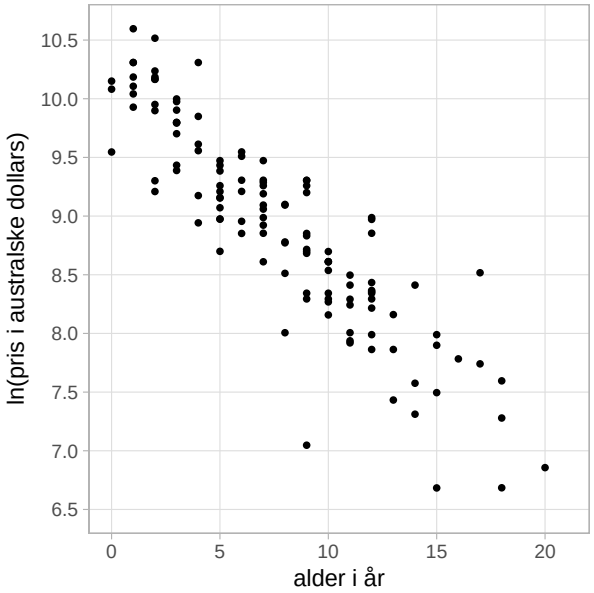
Som det tydeligt fremgår af plottet, så vil en ret linje ikke kunne beskrive sammenhængen mellem y_i og x_i, hvor

$$x_i = \text{alder for den } i\text{'te bil målt i år}$$
$$y_i = \text{pris for den } i\text{'te bil målt i australske dollars}$$

Men hvis vi laver et plot af logaritmen af prisen mod alderen, så synes der at være en pæn lineær sammenhæng. Se figur 14. Resultatet af en simpel lineær regression er, at den bedste rette linje til beskrivelse af ln(pris) ud fra alder er

$$\ell(\text{alder}) = -0{,}1647 \cdot \text{alder} + 10{,}188$$

I modsætning til eksemplet med Galtons højdemålinger, så har parametrene $\hat{a} = -0{,}1647$ og $\hat{b} = 10{,}188$ nogle interessante fortolkninger i sig selv. De



Figur 14: 124 sammenhørende par af ln(pris) og alder for brugte Mazda'er solgt i Melbourne i 1991.

følgende to opgaver omhandler fortolkningen af henholdsvis skæringen og hældningen.

Opgave 47. *Argumenter for, at den forventede pris på en helt ny brugt bil (altså en brugt bil med alder* 0 *år) er* $\exp(b)$*. Argumenter for, at vores estimat for prisen på en helt ny brugt bil af mærket Mazda i året 1991 er 26582 australske dollars.*

Opgave 48. *Vi har set, at der er en aftagende eksponentiel sammenhæng mellem prisen på en brugt Mazda og dens alder. I fysikkens verden kender man aftagende eksponentielle sammenhænge fra f.eks.* radioaktivt henfald*. En standardbeskrivelse af radioaktivt henfald er ved anvendelse af den såkaldte* halveringstid*, altså hvor lang tid der går, før radioaktiviteten er halveret. I situationen med prisen på brugte Mazda'er kan vi helt tilsvarende beregne halveringstiden* $T_{\frac{1}{2}}$*, altså hvor lang tid der går, før* prisen *er halveret. Argumenter for, at halveringstiden for prisen er givet ved formlen*

$$T_{\frac{1}{2}} = \frac{\ln(\frac{1}{2})}{a} = \frac{\ln(2)}{-a}$$

og at vores bedste bud på denne størrelse er

$$\hat{T}_{\frac{1}{2}} = \frac{\ln(2)}{-\hat{a}} = \frac{\ln(2)}{0{,}1647} = 4{,}21$$

Det betyder, at vores forståelse af prisudviklinen er, at prisen på en brugt Mazda halveres hver gang, bilen er 4 år og 2,5 måneder ældre.

I afsnit 3 så vi på fordelingen af residualerne, i afsnit 2.1 blev der udviklet en metode til at beskrive usikkerheden på parameterestimaterne $\hat{a}$ og $\hat{b}$, og i afsnit 2.2 så vi, hvorledes man statistisk kan undersøge, om der er en sammenhæng mellem y–værdierne og x–værdierne. I den sidste opgave i dette afsnit vil vi gøre brug af disse metoder.

Opgave 49. *Indlæs datasættet i dit værktøjsprogram, og tegn et plot af* $\ln(pris)$ *mod alder. Kontroller vores udregning* $\hat{a} = -0{,}1647$ *og* $\hat{b} = 10{,}188$ *ved selv at beregne hældningen og skæringen for den bedste rette linje. Afprøv videre følgende statistiske teknikker:*

- *Tegn et normalfordelingsfraktildiagram for residualerne fra den lineære regression af* $\ln(pris)$ *på alder. Overvej, om residualerne kan antages at være normalfordelte.*
- *Brug omrøringsmetoden til at vurdere usikkerheden på henholdsvis* $\hat{a}$ *og* $\hat{b}$*. Diskuter, hvordan disse usikkerhedsvurderinger kan oversættes til usikkerhedsvurderinger for henholdsvis prisen på en helt ny brugt Mazda og halveringstiden for prisudviklingen.*

5 Multilineær regression

Som eksempel på en såkaldt *multilineær regression* vil vi bruge et dataeksempel bestående af sammenhørende målinger af *timeløn* (i US dollars), *alder* (i år), *uddannelse* (i år), og *arbejdserfaring* (i år) fra 49 kvindelige håndværkere i USA i 1981. De første 10 målinger ser ud som i skemaet herunder.

	timeløn	**alder**	**uddannelse**	**erfaring**
1	12,05	28	16	6
2	6,00	26	17	3
3	12,00	31	14	10
4	10,61	57	15	33
5	10,00	18	16	0
6	7,78	28	16	6
7	10,28	32	15	10
8	15,00	35	18	11
9	12,00	48	17	24
10	8,56	35	14	14

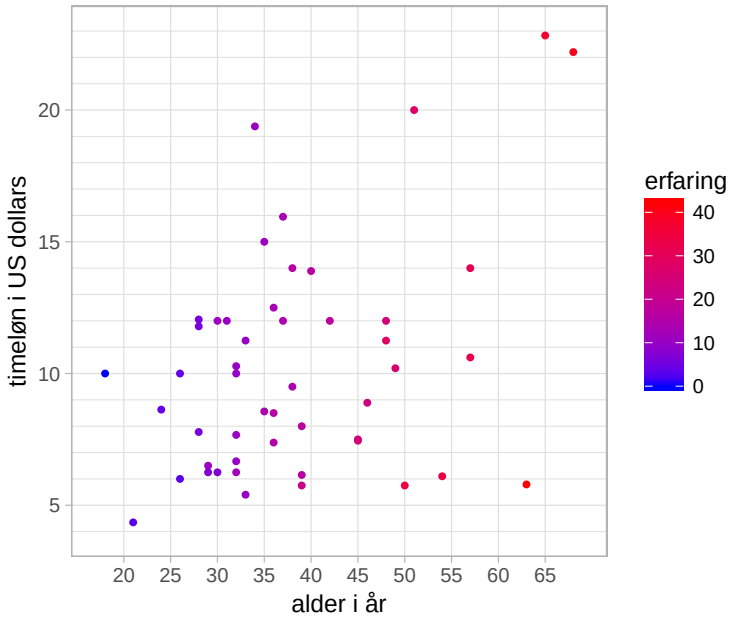
Vi vil undersøge i hvilket omfang, timelønnen kan forklares af alder, uddannelse og erfaring. Ud fra et matematisk synspunkt er det nye, at der nu bruges 3 variable (alder, uddannelse, erfaring) til at forklare lønnen. I de tidligere eksempler har vi kun en enkelt forklarende variabel:

- I afsnit 1 blev en søns højde beskrevet ved farens højde (én variabel).
- I afsnit 4 blev prisen på en brugt Mazda beskrevet ved dens alder (én variabel).

En multilineær regression laves ved samme fremgangsmåde som den simple lineære regression, men nu blot med mere end én forklarende variabel. I løneksemplet vælger vi således at beskrive timelønnen y_i for den i'te kvinde ud fra alderen, uddannelsen og erfaringen ved denne matematiske formel

$$a_1 \cdot \text{alder}_i + a_2 \cdot \text{uddannelse}_i + a_3 \cdot \text{erfaring}_i + b \tag{14}$$

Her er a_1, a_2, a_3, b parametre, som vi vil vælge således, at udtrykket (14) passer så godt som muligt med de observerede timelønninger. Men for at det



Figur 15: Scatterplot af timelønnen mod alderen for 49 kvindelige håndværkere i USA i 1981. Håndværkernes erfaring er visualiseret via farvekodning fra blå ($\sim$ ingen erfaring) til rød ($\sim$ lang erfaring).

overhovedet giver mening at lave en multilineær regression og dermed lede efter disse parameterværdier, så skal man først argumentere for, at den lineære

model i ligning (14) overhovedet er en fornuftig beskrivelse af timelønstallene. Dette kan f.eks. gøres ved at lave nogle scatterplots, såsom det i figur 15. Det er oplagt at have timelønnen på y-aksen, men de tre forklarende variable kan ikke være på x-aksen på en gang. Man kan dog f.eks. visualisere to forklarende variable på en gang ved at bruge den ene på x-aksen og den anden til en farvekodning af datapunkterne. Figur 15 viser en overordnet tendens til, at timelønnen vokser både med håndværkerens alder (de fleste af datapunkterne ligger i et bælte fra nederste venstre til øverste højre hjørne), og også med erfaringen (der er en farvegradient fra blå til rød langs diagonalen fra nederste venstre til øverste højre hjørne).

Opgave 50. *Indlæs datasættet i dit matematikprogram, og lav selv 3 plots med punkter, hvor timelønnen bruges som y–værdier, mens hver af variablene alder, uddannelse og erfaring bruges som x–værdier. Hvad sker der med timelønnen, hvis hver af variablene bliver større? Kan man f.eks. med god rimelighed tegne en ret linje mellem timelønnen og uddannelsen?*

Opgave 51. *Rent faktisk var der målinger på 52 kvinder i det oprindelige datasæt. Men vi har fjernet målingerne for 3 kvinder for at tydeliggøre den matematiske pointe, vi kommer med senere. Diskuter, om det er tilladeligt at fjerne målinger for at fremhæve en pointe.*

Næste skridt bliver at vælge a_1, a_2, a_3 og b på en sådan måde, at modellen for timelønningerne udtrykt i (14) kommer til at passe bedst muligt med de observerede timelønninger. Vores metode til at afgøre, hvad der "passer bedst" vil overordnet set være den samme som metoden til at finde den bedste rette linje i en simpel lineær regression, og den går ud på, at summen af de kvadrerede forskelle mellem de $n = 49$ timelønninger $y_1, \ldots, y_{49}$ og størrelsen udregnet ved formlen i (14), altså

$$\sum_{i=1}^{n} \big(y_i - a_1 \cdot \text{alder}_i - a_2 \cdot \text{uddannelse}_i - a_3 \cdot \text{erfaring}_i - b\big)^2,$$

bliver mindst mulig. Her er n antallet af observationer – i vores tilfælde er $n = 49$. Matematisk kan man vise, at der (i de fleste situationer) findes netop ét valg $\hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{b}$ af parametrene, som gør summen af de kvadrerede forskelle mindst muligt.

Matematisk set kan valget $\hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{b}$ af parametrene, der minimerer den kvadrerede fejl, findes ved at eliminere parametrene en ad gangen. Dette kan gøres ved at opfatte summen af de kvadrerede fejl som en parabel i en af parametrene, f.eks. a_1, og så derefter bruge *lemmaet* fra afsnit 1.3. I praksis er dette dog både besværligt og rodet. Minimeringsproblemet kan

også løses via teknikker fra, hvad der kaldes *"lineær algebra"*, hvilket giver en matematisk elegant beskrivelse. Begge tilgange ligger dog udenfor, hvad vi ønsker at gennemgå i dette manuskript. Heldigvis kan de matematiske værktøjsprogrammer udregne parameterestimaterne for os. For den multilineære regression på de 49 lønmålinger fås således

$$\hat{a}_1 = 0{,}641, \qquad \hat{a}_2 = 0{,}531, \qquad \hat{a}_3 = -0{,}547, \qquad \hat{b} = -13{,}42,$$

På samme måde som for den simple lineære regression kan vi tale om de prædikterede y–værdier ud fra x–værdierne. Altså den værdi, timelønnen skulle have været, hvis det skulle passe perfekt med modellen. For at spare plads senere, bruger vi igen notationen $\hat{y}_i$ for disse værdier, altså

$$\hat{y}_i = \hat{a}_1 \cdot \text{alder}_i + \hat{a}_2 \cdot \text{uddannelse}_i + \hat{a}_3 \cdot \text{erfaring}_i + \hat{b}$$

Opgave 52. *Udregn alle de 49 prædikterede y–værdier for timeløns-datasættet. Udregn også forskellene mellem de faktiske og de prædikterede timelønninger.*

Med notationen for prædikterede y–værdier kan vi opskrive et udtryk for residualspredningen

$$\hat{\sigma} = \sqrt{\frac{1}{n-4}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Igen skal residualspredningen forstås som en slags gennemsnitsværdi for, hvor langt de observerede y_i–værdier ligger fra de prædikterede værdier $\hat{y}_i$. Formel (6) for forklaringsgraden, R^2, fungerer stadigvæk. Vi kan altså udregne forklaringsgraden som

$$R^2 = \frac{\Big(\sum_{i=1}^{n}(y_i - \bar{y}) \cdot (\hat{y}_i - \bar{y})\Big)^2}{\big(\sum_{i=1}^{n}(y_i - \bar{y})^2\big) \cdot \big(\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2\big)}.$$

Opgave 53. *Udregn residualspredningen (det skulle meget gerne omtrentligt give, at $\hat{\sigma} = 3{,}42$).*

Opgave 54. *Hvis man skal være præcis, så skal man også angive enheden på estimaterne. Enheden for b-parameteren og for residualspredningen er* US dollars*. Men hvad er enheden for a_1, a_2 og a_3?*

Helt som for den simple lineære regression kan vi bruge $\hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{b}$ til at give vores bedste bud på, hvordan timelønnen ser ud for personer, som ikke er med i datasættet, men hvor vi kender alder, uddannelse og erfaring. Det er simpelthen et spørgsmål om at sætte de tre værdier ind i ligningen

$$\text{forventet } y = 0{,}641 \cdot \text{alder} + 0{,}531 \cdot \text{uddannelse} - 0{,}547 \cdot \text{erfaring} - 13{,}42$$

Opgave 55. *Beregn den forventede timeløn for en kvindelig håndværker i USA i 1981 i følgende 4 situationer:*

A:	*alder = 28 år,*	*uddannelse = 15 år,*	*erfaring = 6 år.*
B:	*alder = 29 år,*	*uddannelse = 15 år,*	*erfaring = 6 år.*
C:	*alder = 28 år,*	*uddannelse = 16 år,*	*erfaring = 6 år.*
D:	*alder = 28 år,*	*uddannelse = 15 år,*	*erfaring = 7 år.*

Den umiddelbare fortolkning af parameteren $\hat{a}_1$ er, at håndværkerens timeløn i gennemsnit stiger med 0,641 dollars per år ældre, hun er. Tilsvarende er den umiddelbare fortolkning af $\hat{a}_2$, at timelønnen stiger med 0,531 dollars per år længere uddannelse, hun har. Uddannelse synes altså at kunne betale sig. Så langt så godt, men hvad med $\hat{a}_3$? Kan det virkelig være rigtigt, at timelønnen i gennemsnit falder med 0,547 dollars per år erfaring, man har som håndværker? Nej, vel? Så hvad er der galt? Regner de gængse matematikprogrammer forkert? Eller kan multilineær regression ikke bruges til at beskrive timelønnen?

Svaret er, at der ikke er noget galt, men at det kontraintuitive fortegn på $\hat{a}_3$ i det konkrete eksempel er en artefakt af et samspil mellem *alder*, *uddannelse* og *erfaring*. Faktisk er det sådan, at man meget præcist kan forudsige alderen på kvinderne i undersøgelsen, hvis man ved, hvor lang deres uddannelse og erfaring er i år. Med god approksimation gælder nemlig

$$\text{alder} \approx 6 + \text{uddannelse} + \text{erfaring}. \tag{15}$$

Forklaringen på dette er helt ligetil: Man starter i skole, når man er 6 år gammel. Derefter uddanner man sig i nogle år (uddannelse er inkl. skoletiden), hvorefter man får sig et arbejde og får mere og mere arbejdserfaring år for år.

Opgave 56. *Efterprøv, hvor godt approksimationen fra (15) passer for personerne i datasættet: Udregn for personerne i datasættet værdien af*

$$6 + \textit{uddannelse} + \textit{erfaring}$$

og sammenlign med den faktiske alder.

Opgave 57. *Lav en multilineær regression i det matematiske værktøjsprogram med* alder *som y-værdien, og* uddannelse *og* erfaring *som de forklarende variable. Sammenlign resultatet med ligning* (15)*, og diskuter, hvorfor den multilineære regression ikke giver nøjagtig samme resultat.*

Indsættes approksimationen (15) i ligning (14), så fås følgende model for timelønnen

$$\begin{aligned} &a_1 \cdot \text{alder}_i + a_2 \cdot \text{uddannelse}_i + a_3 \cdot \text{erfaring}_i + b \\ &\quad \approx a_1 \cdot \big(6 + \text{uddannelse}_i + \text{erfaring}_i\big) + a_2 \cdot \text{uddannelse}_i + a_3 \cdot \text{erfaring}_i + b \\ &\quad = (a_1 + a_2) \cdot \text{uddannelse}_i + (a_1 + a_3) \cdot \text{erfaring}_i + (6 \cdot a_1 + b). \end{aligned}$$

Man kan derfor lave en multilineær regression for timelønnen, som har næsten samme forklaringsgrad, *uden* at *alder* bruges som forklarende variabel. Tilsvarende kan man lave multilineære regressioner med næsten samme forklaringsgrad, hvor man udelader henholdsvis *uddannelse* eller *erfaring*. Hvilken af disse 3 statistiske modeller, der giver den "korrekte" forklaring af timelønnen, er svært af afgøre ud fra et datasæt bestående af kun 49 målinger.

Opgave 58. *Lav en multilineær regression i det gængse matematikprogram med* timeløn *som y-værdien, og* alder *og* erfaring *som de forklarende variable. Diskuter, hvordan det kan være, at fortegnet på hældningen mht. erfaring nu er blevet positivt.*

Lad os opsummere dette afsnit med en oplistning af nogle, efter vores opfattelse, vigtige moraler:

- Multilineær regression er et yderst nyttigt værktøj, som tillader, at man kan inddrage flere forklarende variable på en gang.
- Men man skal passe på med fortolkningen af parametrene. Specielt hvis der er et samspil mellem de forklarende variable, kan man blive snydt. I det konkrete eksempel så vi således, at fortegnet på hældningen mht. *erfaring* ændrede sig alt efter hvilke andre forklarende variable, der blev medtaget i modellen.
- Angående opgave 51: Når statistik bruges til at finde sammenhænge i naturen, samfundet, eller andet, så må man naturligvis *ikke* selektere i data! Men når vi laver matematik for matematikkens egen skyld, så kan vi godt selektere i data. Hvis vi havde beholdt de 3 målinger, som blev fjernet, så var der gode argumenter for at modellere $\frac{1}{\sqrt{timeløn}}$ i stedet for timelønnen selv. Og i vores eksempel ønskede vi at undgå denne unødvendige komplikation.

6 Case study: Verdensrekorder

I dette afsnit vil vi lave en statistisk model for verdensrekorderne i løb. I modelleringsarbejdet får vi foruden logaritme–transformationer, fraktildiagrammer, og multilineær regression også brug for at studere residualplots.

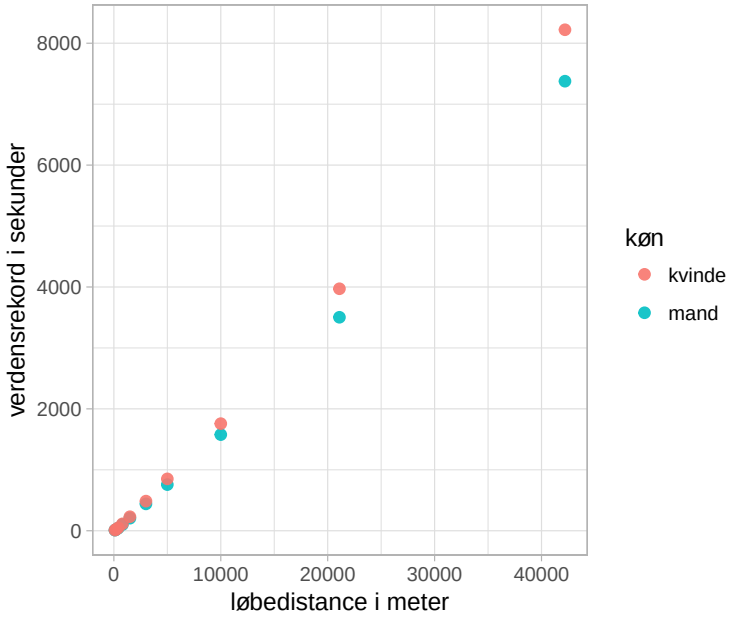
Verdensrekorderne i løb, som vi downloadede fra hjemmesiden for IAAF, se [3], d. 7. juni 2018, findes i skemaet herunder.

distance	**køn**	**rekord**
100	mand	9,58
200	mand	19,19
400	mand	43,03
800	mand	100,91
1500	mand	206,00
3000	mand	440,67
5000	mand	757,35
10000	mand	1577,53
21097,5	mand	3503
42195	mand	7377
100	kvinde	10,49
200	kvinde	21,34
400	kvinde	47,60
800	kvinde	113,28
1500	kvinde	230,07
3000	kvinde	486,11
5000	kvinde	851,15
10000	kvinde	1757,45
21097,5	kvinde	3971
42195	kvinde	8221

Distancen er angiver i meter, og rekorden er angivet i sekunder. Desuden skal det bemærkes, at distancerne op til 10 km løbes på bane, mens halv- og helmarathon løbes på landevej. Som altid er det en god ide at starte med at tegne datasættet, før man forsøger at lave en statistisk model. På figur 16 har vi tegnet verdensrekorden mod distancen, hvor punkterne er farvekodet efter, om rekorden er for mænd eller kvinder.

Opgave 59. *Prøv selv at fremstille et tilsvarende plot, hvor der kan ses forskel på punkterne for kvinder og mænd.*

Vi kan umiddelbart se, at mændene løber hurtigere end kvinderne. Inden for hvert køn synes punkterne at ligge på rette linjer, hvilket kunne få os til



Figur 16: Verdensrekorderne i løb, som de var den 7. juni 2018.

at foreslå den statistiske model:

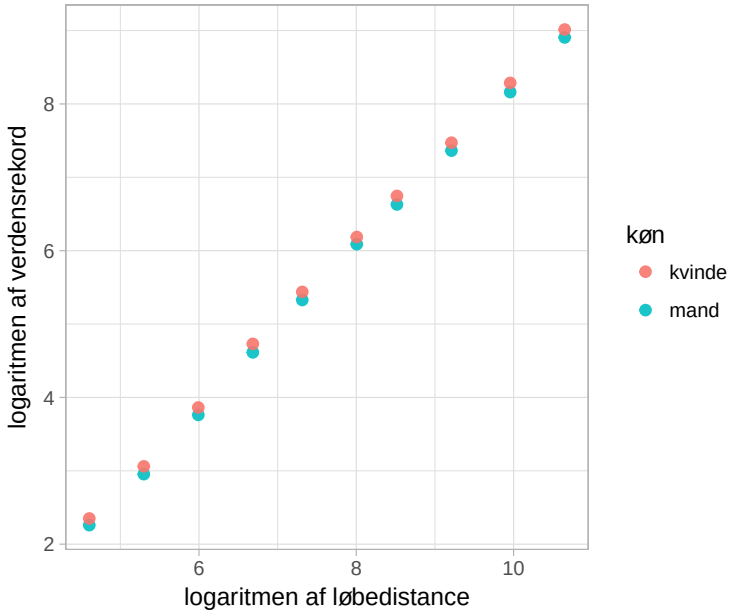
$$\begin{aligned} \text{For mænd:} \quad & \text{rekord}_i \approx a_{\text{mand}} \cdot \text{distance}_i + b_{\text{mand}}, \\ \text{For kvinder:} \quad & \text{rekord}_i \approx a_{\text{kvinde}} \cdot \text{distance}_i + b_{\text{kvinde}}, \end{aligned}$$

hvor parametrene a_{mand} og a_{kvinde} angiver den reciprokke løbehastighed (i enheden *sekunder per meter*) for de to køn. At mændene løber hurtigere end kvinderne, ville da give sig til udtryk ved, at $a_{\text{mand}} < a_{\text{kvinde}}$.

Men her er vi blevet snydt af tegningen, for at foreslå en statistisk model, hvor løbehastigheden kun afhænger af kønnet og ikke af løbedistancen, strider mod almindelig viden omkring løb!

Opgave 60. *Har du selv prøvet at løbe en 100 meter, en 400 meter, og en 1500 meter? Hvis ja, løb du så disse distancer med samme hastighed? Og hvis du har set verdensrekordholderne i TV eller på nettet, har du så lagt mærke til, at der er meget stor forskel i kropsbygningen af kort-, mellem- og langdistance løberne?*

Hvis man ser nøje efter på figur 16, så kan man faktisk godt se, at "punkterne" krummer opad svarende til, at løbehastigheden er langsommere, jo længere distancen er. Det vil vise sig, at dette kan beskrives ved en ret linje



Figur 17: Verdensrekorderne i løb, som de var den 7. juni 2018. Der er taget logaritmen af både distancen og verdensrekorden.

på logaritmisk skala! Altså

$$\begin{aligned} \text{For mænd:} \quad & \ln(\text{rekord}_i) \approx a \cdot \ln(\text{distance}_i) + b_{\text{mand}}, \\ \text{For kvinder:} \quad & \ln(\text{rekord}_i) \approx a \cdot \ln(\text{distance}_i) + b_{\text{kvinde}}. \end{aligned} \tag{16}$$

Vi bruger her metoden fra opgave 46, hvor vi så, at en simpel lineær regression af $\ln(y)$ mod $\ln(x)$ svarer til en potenssammenhæng mellem y og x. Bemærk, at vi i log-mod-log modellen har valgt at bruge samme hældning a for mænd og kvinder. Dette svarer til, at linjerne for mænd og for kvinder er parallelle, hvilket faktisk synes at være tilfældet i figur 17.

Opgave 61. *Prøv selv at fremstille en figur ligesom figur 17: Udregn først den naturlige logaritme af både distancer og rekorder, og optegn så* $\ln(\text{rekord}_i)$ *mod* $\ln(\text{distance}_i)$ *for både mænd og kvinder.*

Opgave 62. *På figur 17 vokser x-værdierne med stort set sammen størrelse fra en løbedistance til den næste. Dette er i modsætning til figur 16, hvor de korte løbedistancer klumper sig sammen i nederste venstre hjørne. Forklar hvorfor! Vink: Den næste løbedistance er stort set dobbelt så lang som den forrige.*

Næste skridt er at finde estimaterne $\hat{a}$, $\hat{b}_{\text{mand}}$ og $\hat{b}_{\text{kvinde}}$ for modellens parametre ud fra mindste kvadraters metode. Men her er der en matematisk udfordring. De formler for hældning og skæring, der blev udledt i afsnit 1.3, kan umiddelbart kun bruges for mænd og kvinder hver for sig. Men i ligning (16) har vi besluttet os for en model, hvor hældningen m.h.t. ln(distance) antages at være den samme for de to køn. Løsningen på den matematiske udfordring kan findes ved at bemærke en egenskab, der gælder for en simpel lineær regression. Hvis man

1. trækker gennemsnittet $\bar{x}$ fra alle x_i–målingerne,

2. trækker gennemsnittet $\bar{y}$ fra alle y_i–målingerne,

3. finder den bedste rette linje gennem de nye punkter,

så vil den bedste rette linje have skæring $b = 0$ og en hældning, som er den samme som hældningen på den bedste rette linje gennem de oprindelige punkter (x_i, y_i).

Opgave 63. *Gå tilbage til datasættet med højder for fædre og sønner. Træk gennemsnittet af fædrehøjderne fra alle fædrehøjderne, og træk gennemsnittet af sønhøjderne fra alle sønhøjderne. Plot de nye punktpar, find også den bedste rette linje gennem dem, og indtegn den i plottet.*

I opgave 63 så vi, at man ved at følge den ovenstående beskrivelse bare flytter hele punktskyen, så den ligger centreret omkring punktet (0,0). Dette ændrer ikke noget på hældningen af den bedste rette linje gennem punkterne.

Vi vil nu bruge samme teknik til at flytte punkterne

$$\big(\log(\text{distance}_i), \log(\text{rekord}_i)\big)$$

ind omkring (0,0) i passende forstand. Vi lader nu $\bar{y}_{\text{mand}}$ betegne gennemsnittet af $\log(\text{rekord}_i)$ for mændenes rekorder, $\bar{y}_{\text{kvinde}}$ betegne gennemsnittet af $\log(\text{rekord}_i)$ for kvindernes rekorder, og $\bar{x}$ betegne gennemsnittet af $\log(\text{distance}_i)$ (for alle værdierne samlet). Herefter gør vi følgende:

1. Vi trækker gennemsnittet $\bar{y}_{\text{mand}}$ fra alle $\log(\text{rekord}_i)$ for mændenes rekorder.

2. Vi trækker gennemsnittet $\bar{y}_{\text{kvinde}}$ fra alle $\log(\text{rekord}_i)$ for kvindernes rekorder.

3. Vi trækker gennemsnittet $\bar{x}$ fra alle $\log(\text{distance}_i)$–værdierne.

4. Vi finder hældningen af den bedste rette linje gennem *alle* de nye punkter.

Som led i udregningerne fås, at

$$\bar{y}_{\text{mand}} = 5{,}6067, \qquad \bar{y}_{\text{kvinde}} = 5{,}7148, \qquad \bar{x} = 7{,}6234,$$

og videre, at estimatet $\hat{a}$ for den fælles hældning bliver $\hat{a} = 1{,}1093$.

Opgave 64. *Prøv at gå gennem de 4 trin i beskrivelsen ovenfor, og prøv derved at efterregne værdien $\hat{a} = 1{,}1093$. Fremstil også et plot, der viser alle punkterne fra punkt 4.*

Næste skridt er at finde de to skæringer for henholdsvis mændenes og kvindernes linje. Her bruger vi samme formel som i afsnit 1.3 og får derved

$$\hat{b}_{\text{mand}} = \bar{y}_{\text{mand}} - \hat{a} \cdot \bar{x} = 5{,}6067 - 1{,}1093 \cdot 7{,}6234 = -2{,}8499,$$
$$\hat{b}_{\text{kvinde}} = \bar{y}_{\text{kvinde}} - \hat{a} \cdot \bar{x} = 5{,}7148 - 1{,}1093 \cdot 7{,}6234 = -2{,}7418.$$

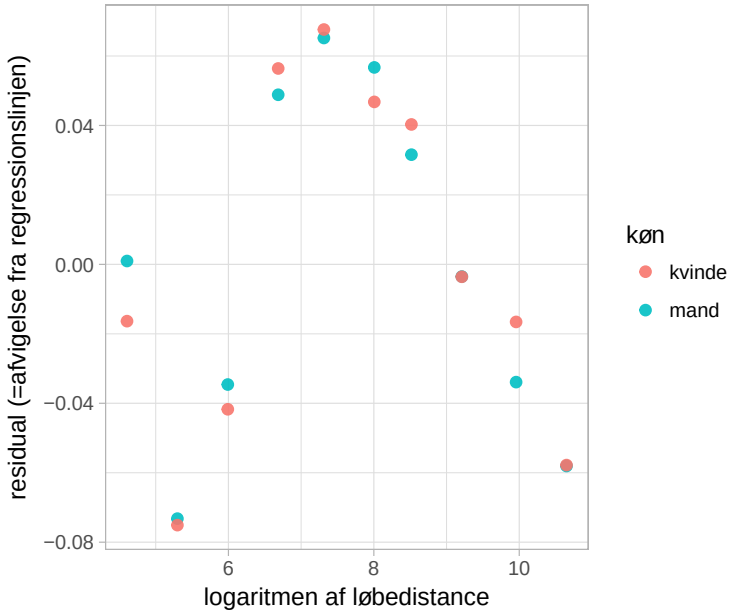
Alt i alt fik vi fundet estimaterne for parametrene i ligning (16). For at tage højde for de afrundinger der blev lavet undervejs, så afrunder vi yderligere til 2 decimaler, og får resultatet

$$\hat{a} = 1{,}11, \qquad \hat{b}_{\text{mand}} = -2{,}85, \qquad \hat{b}_{\text{kvinde}} = -2{,}74.$$

Opgave 65. *Træk vejret dybt. Den matematiske argumentation, der blev brugt i beregningen af $(\hat{a}, \hat{b}_{mand}, \hat{b}_{kvinde})$, er ikke simpel, og den vil være svær at forstå for de fleste. Det er helt ok, hvis du bare tager resultatet for gode varer, og gør dig klar til at læse videre. Men du er naturligvis også meget velkommen til at gennemtænke argumentationen endnu engang, og til at efterprøve vores udregninger.*

Ved hjælp af formel (6) kan vi beregne, at modellen (16) har en forklaringsgrad på $R^2 = 0{,}9999$. De to linjer med hældning 1,11 og skæring med y-aksen på henholdesvis $-2{,}85$ og $-2{,}74$ ligger således ganske tæt på punkterne på figur 17. Men faktisk er det alligevel ikke en god statistisk model! Der er nemlig en påfaldende systematik i residualerne, dvs. i afvigelsen af punkterne fra linjerne. For at se dette kan man tegne et **residualplot**, der viser afvigelserne fra linjerne for hver værdi af logaritmen af løbedistancen.

Residualplottet findes i figur 18. Vi ser et "knæk" omkring ln(distance) $\approx$ 7,1, svarende til distance ≈ 1200. Dette knæk svarer til, at hældningen mod ln(distance) på figur 17 er lidt stejlere, når distancen er kortere end 1200 meter, hvilket svarer til en fysiologisk forskel på kort- og langdistance løbere!



Figur 18: Residualplottet hørende til log-mod-log modellen i ligning (16).

Der synes dog ikke at være et udtalt behov for at inddele distance i 3 kategorier, hvor der også tales om mellemdistance. Derimod synes de 2 residualer for 100 meteren, altså ved x-værdien $\ln(100) = 4{,}6$ i figur 18, at være anderledes. Det har en fysisk forklaring, nemlig at accelerationsfasen betyder rigtigt meget i den meget korte 100 meter distance.

Af residualplottet kan vi konkludere, at vi i virkeligheden burde gøre modellen *endnu* mere kompliceret ved at tillade en form for knæk på de linjer, der beskriver mændenes og kvindernes rekorder som funktion af distancen. Til at opstille sådan en model kunne man benytte teknikker fra den multilineære regression. Det ville imidlertid blive for omfattende i dette notat, så vi vil nøjes med at konstatere, at modellen ville kunne gøres endnu bedre.

Opgave 66. *Gør rede for, at hvis verdensrekorderne i løb for mænd og kvinder kan beskrives ved modellen (16), så vil kvindernes verdensrekorder generelt set være* $\exp(b_{kvinde} - b_{mand})$ *gange så lange som mændenes verdensrekorder. Brug dette til at argumentere for, at uanset distancen så løber de bedste kvinder 11,6% langsommere end de bedste mænd.*

Litteratur

[1] Susanne Ditlevsen, Helle Sørensen (2015), "Introduktion til statistik", Institut for Matematiske Fag, Københavns Universitet.

[2] Francis Galton (1886), "Regression towards Mediocrity in Hereditary Stature", Journal of the Anthropological Institute, side 246–263.

[3] International Association of Atheletics Federations. Hjemmeside `http://www.iaaf.org` tilgået d. 7. juni 2018.

[4] Alan Lee (1994), "Data Analysis: An introduction based on R", Department of Statistics, University of Auckland. Datasæt kan downloades fra `http://www.statsci.org/data/oz/mazdas.html`.

[5] Michael Sørensen (2012), "En introduktion til sandsynlighedsregning", Institut for Matematiske Fag, Københavns Universitet.

[6] Tyler Vigen, "Spurious Correlations", 2015. Hjemmeside `http://www.tylervigen.com/spurious-correlations`.