Lineær Regression
B-niveau

Anders Rønn-Nielsen
Copenhagen Business School

Bo Markussen
Københavns Universitet

14. september, 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 beregnelsen **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 tal par, 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.

Et andet princip, som har været styrende for udformningen af dette manuskript, er, at vi bruger rigtige 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 tilrådighed 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.

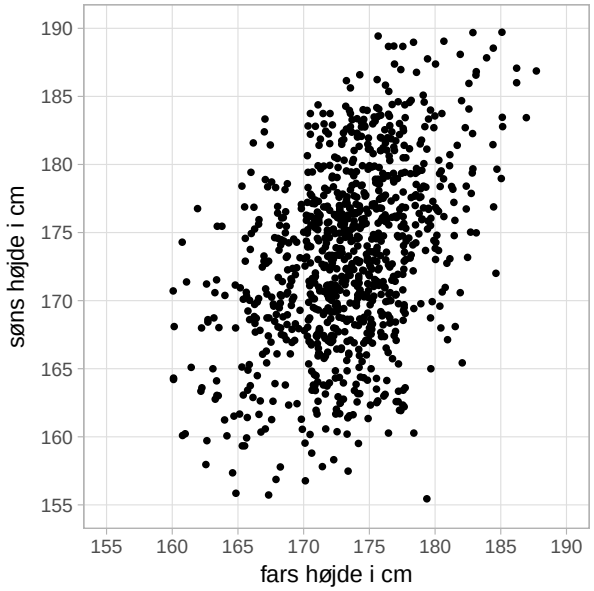
1 Simpel lineær regression

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.

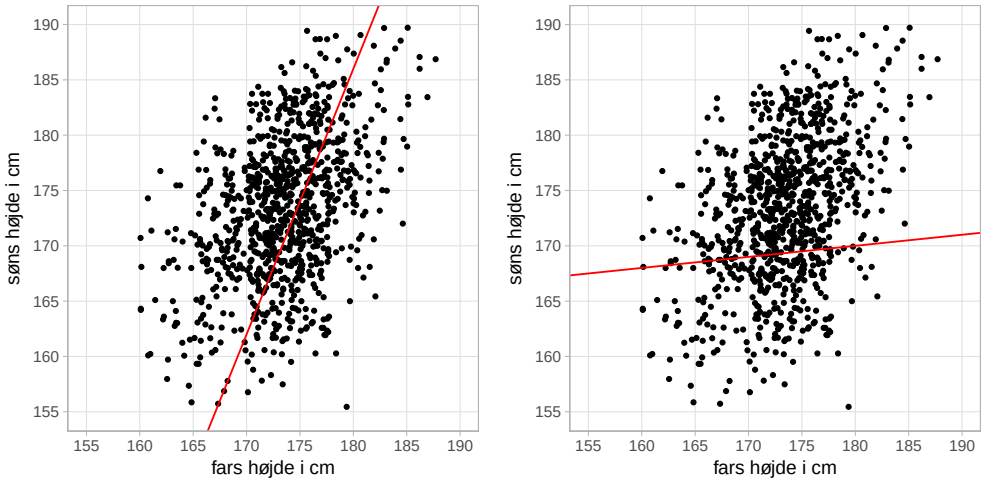
Opgave 1. *Indlæs datasættet i et matematisk værktøjsprogram, og tegn et tilsvarende plot. Hvad er den mindste og største værdi af fædrenes højder? Hvad er den mindste og største værdi af sønnernes højder?*

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.*

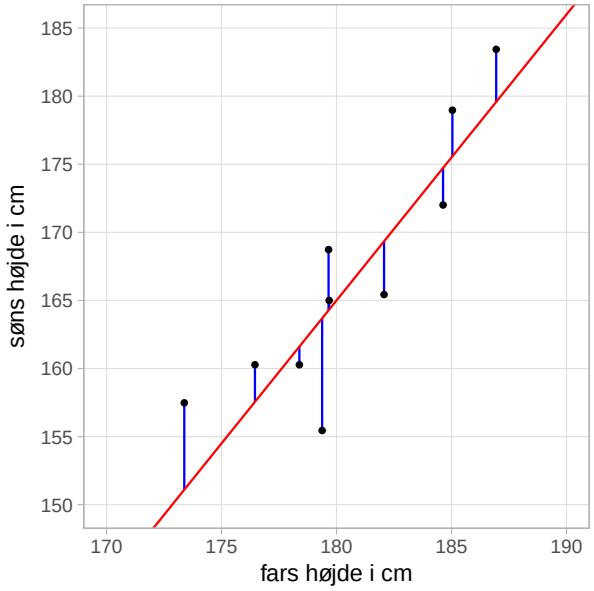


Figur 2: To valg af linjer til at beskrive punkternes voksende tendens.

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.

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.*



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.

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.

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 rød 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 1.1 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 " $\hat{}$ " 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 undlade at komme med et matematisk bevis for dette men blot

bemærke, at beviset kan gennemføres ved at bruge formlen for toppunktet for en parabel. Beviset kan findes i det tilsvarende notat, som er rettet mod det gymnasiale A–niveau.

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

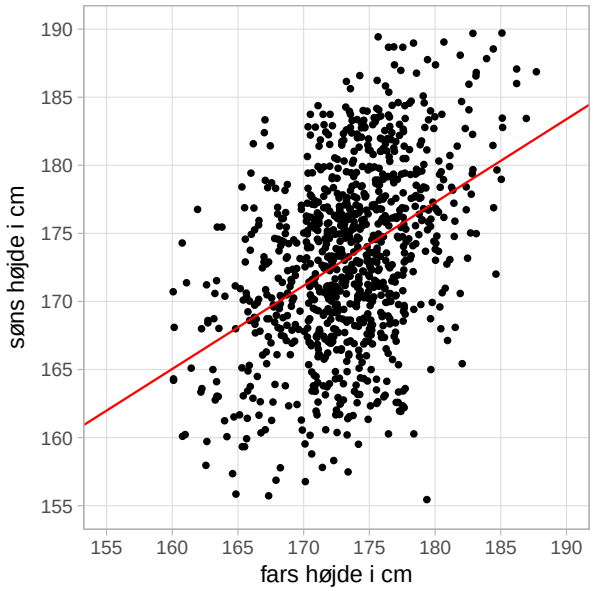
$$(183{,}4 - 166{,}7) \cdot (186{,}9 - 180{,}6) + (172{,}0 - 166{,}7) \cdot (184{,}6 - 180{,}6) \\ + \cdots + (157{,}5 - 166{,}7) \cdot (173{,}4 - 180{,}6)$$

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.



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

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.

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 tilsyneladende 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 1.1.

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 nøjagtigt 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 formeludtrykket

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

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 i, hvor langt fra linjen punkterne ligger. 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 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 14. *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

$$y_i = a \cdot x_i + b + r_i \tag{5}$$

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.

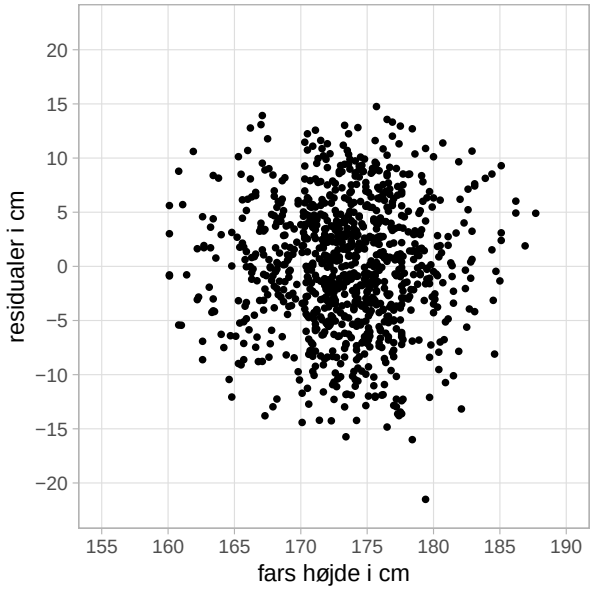
Vi forestiller os, at de enkelte tal r_i er tilfældige. I statistikken kaldes tallene r_i også for **støjled**. De repræsenterer den "støj", der kommer til udtryk i data, når vi forsøger at beskrive data med en model. Residualerne kan være både positive og negative, og hvis et af r_i'erne er positivt, 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 5.

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 15. *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 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



Figur 5: Residualplot for den statistiske modellering af sønners højde ud fra fadernes højde.

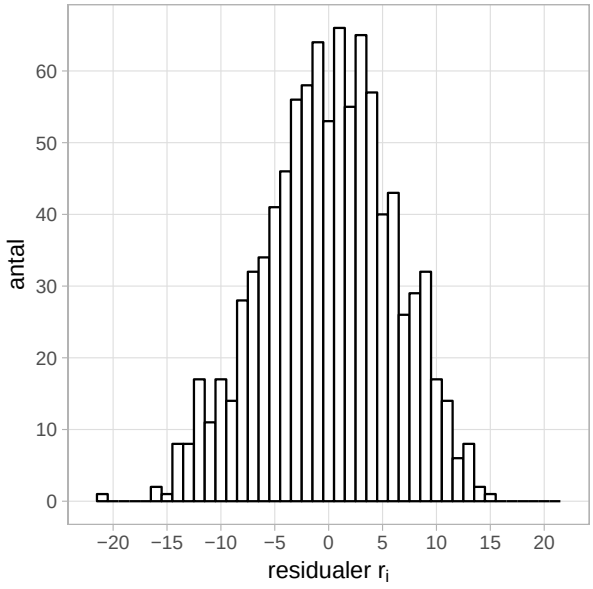
rigtige, men ukendte, parametre a og b, som "blev brugt", da naturen frembragte sønnernes højder ud fra fædrenes via den rette linje og de tilfældige afvigelser. Men estimaterne $\hat{a}$ og $\hat{b}$ er altså vores bedste bud på hældningen og skæringen.

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 resiudalspredning betyder, at residualerne generelt set er numerisk størrere (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 6: 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 6 er der tegnet et histogram over alle residualværdierne.

Opgave 16. *Prøv selv at tegne et histogram over residualernes værdier. Bemærk, at det ikke nødvendigvis kommer til at ligne histogrammet på figur 6 fuldstændigt. Det kommer an på hvor brede søjlerne i histrogrammet er lavet. På figur 6 er histrogrammet konstrueret, så alle søjler har bredde 1.*

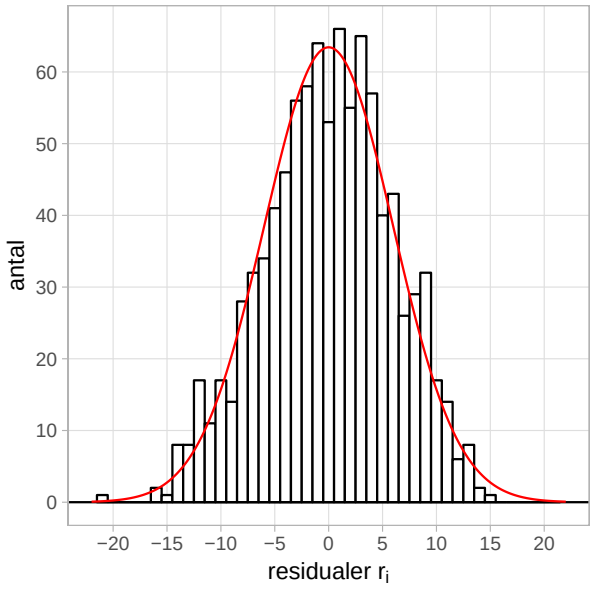
Histogrammet har – pånær nogle enkelte svipsere – overordnet set en "klokkeagtig" form, der er symmetrisk omkring 0. De fleste residualværdier er altså relativt tæt på 0, der er ca. lige mange på den negative side som på den positive side, og dem på den negative side er fordelt nogenlunde som dem på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

Opgave 17. *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 virker 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 7



Figur 7: Histogram over alle residualerne sammen med grafen for funktionen $\varphi(x)$. Grafen for $\varphi(x)$ er indtegnet med rødt.

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 det, der kaldes **normalfordelingen**.

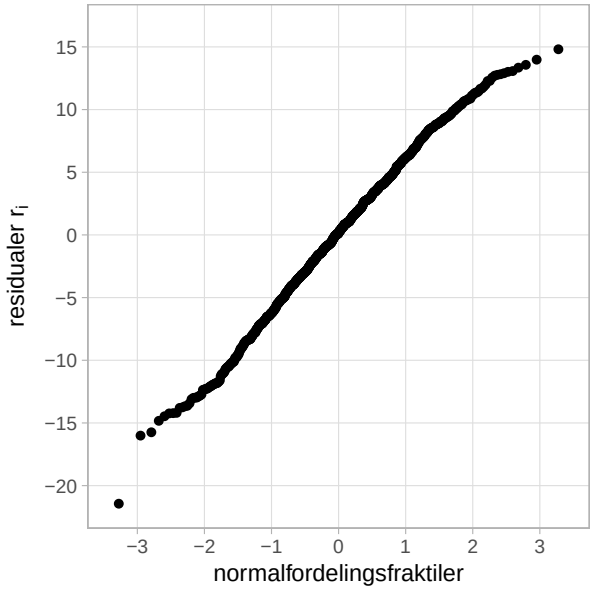
Mere generelt siges en samling af talværdier at følge normalfordelingen, hvis et histogram over dem nogenlunde følger en funktion $\varphi(x)$ på formen

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

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



Figur 8: Normalfordelingsfraktildiagram for residualerne for far–søn–datasættet.

valgt eksponentialfunktion. Det gøres ved at tegne et **normalfordelingsfraktildiagram**. Hvis punkterne på sådan en tegning stort set ligger på en ret linje, så indikerer dette, at residualerne godt kan antages at være normalfordelte. Et normalfordelingsfraktildiagram kan let tegnes i et matematisk værktøjsprogram.

På figur 8 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 (5) 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 $f_1 = \exp(r_1)$, ..., $f_n = \exp(r_n)$, 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. Selv om 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 18. *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.

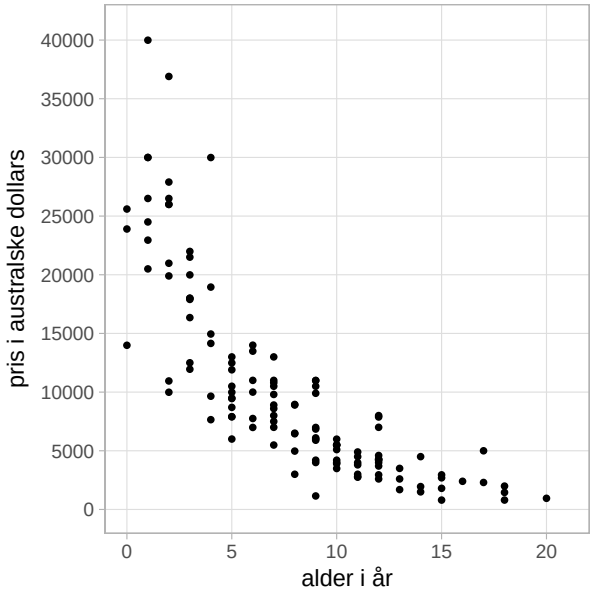
4.1 Eksempel på en eksponentiel regression

Figur 9 viser sammenhængen mellem pris og alder for 124 brugte biler af mærket *Mazda* solgt i Melbourne, Australien, i året 1991.

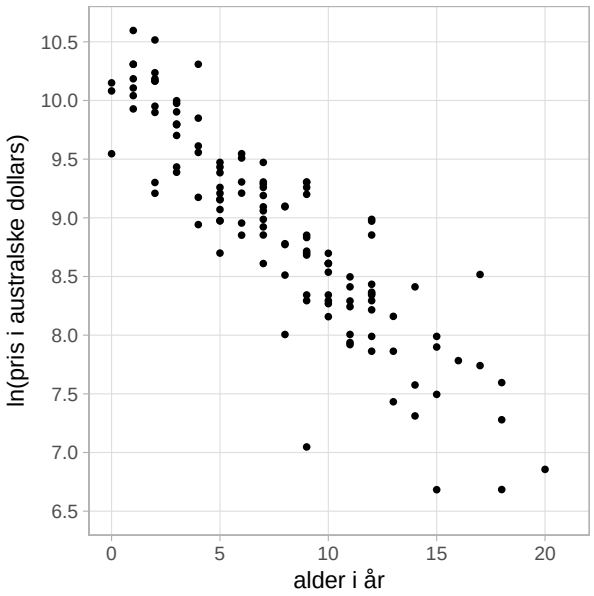
Som det tydeligt fremgår af plottet, så vil en ret linje ikke kunne beskrive sammenhængen mellem y_i og x_i, hvor

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

Men hvis vi laver et plot af logaritmen af prisen mod alderen, så synes der at være en pæn lineær sammenhæng. Se figur 10. Resultatet af en simpel



Figur 9: 124 sammenhørende par af pris og alder for brugte Mazda'er solgt i Melbourne i 1991.



Figur 10: 124 sammenhørende par af ln(pris) og alder for brugte Mazda'er solgt i Melbourne i 1991.

lineær regression er, at den bedste rette linje til beskrivelse af ln(pris) ud fra

alder er

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

Opgave 19. *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.*

Opgave 20. *Tegn et normalfordelingsfraktildiagram for residualerne fra den lineære regression af* ln(*pris*) *på alder. Overvej, om residualerne kan antages at være normalfordelte.*

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 følgende to opgaver omhandler fortolkningen af henholdsvis skæringen og hældningen.

Opgave 21. *Argumentér 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 22. *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 brugt Mazda halveres hver gang, bilen er 4 år og 2,5 måneder ældre.

Litteratur

[1] Susanne Ditlevsen og 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] A. 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`.

[4] Bo Markussen og Anders Rønn-Nielsen (2018). Lineær Regression – A-Niveau.

[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`.