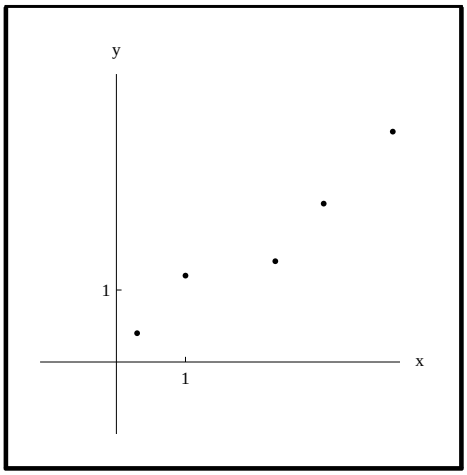
Regression

I dette projekt ser vi på, hvordan man kan opstille en matematisk model for sammenhængen mellem to størrelser ud fra målinger af dem.

Hvad er regression?

Når man i de naturvidenskabelige fag udfører et eksperiment, får man ofte et datasæt bestående af samhørende målinger af to størrelser. Lad os kalde størrelserne x og y. Så kan vi se vores datasæt som en samling **punkter**, hvor punkternes koordinater svarer til samhørende målinger. Disse punkter kalder vi **målepunkter**. Vi kan derfor visualisere vores datasæt ved at tegne målepunkterne ind i et koordinatsystem, hvorved vi får noget i stil med følgende figur:



Formålet med eksperimentet er ofte at efterprøve en teoretisk sammenhæng mellem x og y, som stammer fra en matematisk model for det, man undersøger. Denne sammenhæng er næsten altid en af de fire, vi så på i kapitlet om sammenhænge. I det kapitel så vi, at vi altid kan oversætte en af de fire typer sammenhænge til en lineær sammenhæng. Vi kan derfor i det følgende gå ud fra, at vi ønsker at efterprøve en lineær sammenhæng mellem x og y.

I praksis ligger målepunkterne ikke præcis på en ret linje. De vil ligesom på figuren ovenfor ligge lidt spredt. Dette skyldes de forskellige fejlkilder, som man bruger en del tid på at diskutere i de naturvidenskabelige fag. Det er altså (ikke nødvendigvis) et tegn på, at teorien/modellen er forkert.

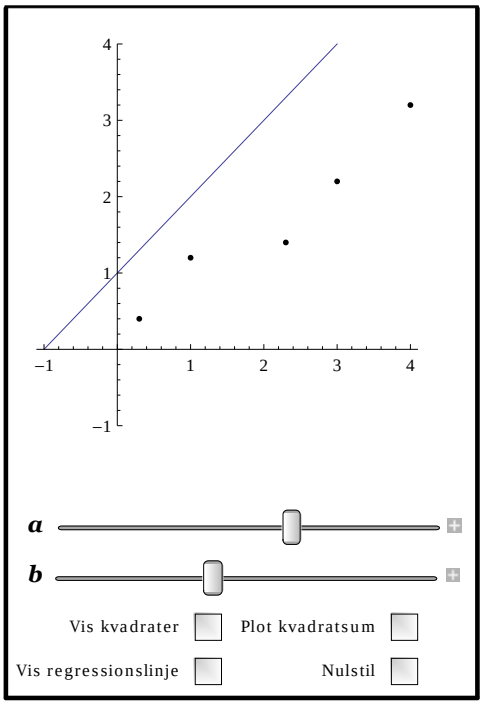
Vores mål i dette kapitel er at bestemme den linje, der passer bedst med målepunkterne. Denne linje kalder man også **'bedste rette linje'**, **'tendenslinjen'** eller **'regressionslinjen'**. Vi starter med at definere, hvad vi vil mene med at en linje 'passer bedst'. Det fører til **mindste kvadraters metode**. Vi ønsker desuden at kunne sige noget om, hvor sikre vi er på, at der er en lineær sammenhæng mellem x og y. Til det formål indfører vi **forklaringsgraden**.

Generelt bruger man ordet **regression**, når man bestemmer den kurve af en given type, som bedst passer med nogle punkter. At udføre regression handler altså om at **tilpasse kurver til punkter**.

Mindste kvadraters metode

Introduktion til metoden

I dette afsnit ser vi på den mest almindelige form for regression, nemlig lineær regression med **mindste kvadraters metode**. Hele afsnittet tager udgangspunkt i følgende figur:



Hvordan skal jeg bruge figuren?

1. Åbn dette afsnit i linkbogen, så du kan se figuren og forklaringen samtidigt.
2. Vælg først dit bud på den bedste rette linje ved at vælge hældning og begyndelsesværdi med skyderne.
3. Begrund dit valg: hvorfor netop den linje? Prøv at formulere din begrundelse på en måde, så du kan være sikker på, at alle vil finde frem til præcis samme linje som dig, hvis de hører din begrundelse.
4. Du fandt sandsynligvis ud af, at det er svært at forklare sin begrundelse særlig præcist i 3. Den metode som NSpire, Maple, Excel og alle andre regneprogrammer bruger til at bestemme den bedste rette linje kaldes **mindste kvadraters metode**. Klik på 'Vis kvadrater'. Metoden fungerer ved at minimere det samlede areal af de kvadrater, du nu ser. Kvadraternes samlede areal kaldes **kvadratsummen**.
   Tænk over, hvorfor det er en fornuftig metode at **minimere kvadratsummen**. Fx ved at overveje følgende spørgsmål:
   a. Hvad svarer sidelængden af kvadraterne til i forsøget?
   b. Hvorfor sætte sidelængderne i anden? Hvilken betydning har det for punkter, der ligger langt fra linjen?
5. Forsøg nu at gøre kvadratsummen endnu mindre ved at ændre på skyderne. Husk at du kan holde Alt nede for at finjustere og Alt og Shift tasterne nede for at justere ekstra fint.
   Når du mener at have fundet den hældning og den begyndelsesværdi, der gør kvadratsummen mindst, så gå videre til næste trin.
6. Klik nu på 'Vis regressionslinje'. Den røde linje er den linje, der har den mindste kvadratsum. Fandt du den samme?
7. Uanset om du fandt regressionslinjen eller ej, så kan du nok godt se at ovenstående er en besværlig og langsommelig proces. I de næste afsnit ser vi på, hvordan man kan bestemme formler for den optimale hældning og begyndelsesværdi, så man slipper for at prøve sig frem på denne måde.
8. Vi kan også bruge figuren til at forstå, hvor idéen til bestemmelse af formlerne kommer fra:
   Klik først på 'Nulstil' og så på 'Plot kvadratsum'. Denne tredimensionelle graf kan du **rotere** ved at trække i den. Du kan **zoome** ved at holde Alt-tasten nede, mens du trækker, og du kan **flytte kameraet** ved at holde shift-tasten nede mens du trækker.
   Det smarte ved denne graf er, at vi kan visualisere kvadratsummen og hurtigt se, hvad der er den optimale hældning og begyndelsesværdi: Det er dem, der svarer til det punkt, som svarer til den laveste kvadratsum. Altså det punkt, der ligger i bunden af 'skålen'.
   Ved at flytte på skyderne kan du flytte det blå punkt, der svarer til dit valg af hældning (a) og begyndelsesværdi (b). Punktet flytter sig langs den røde kurve, når du ændrer på a og langs den blå, når du ændrer på b.
   Ved at eksperimentere med forskellige kameravinkler og værdier for a og b, kan man overbevise sig om, at det punkt, der svarer til den optimale hældning og begyndelsesværdi er det eneste punkt, der **både ligger i bunden af den røde og den blå kurve.** Det er denne egenskab, vi bruger, når vi senere udleder formlerne for regressionslinjens hældning og begyndelsesværdi ved brug af differentialregning.
   Vi oversætter altså problemet med at bestemme den bedste rette linje til problemet med at **bestemme minimum for en funktion**. Vi er så heldige, at det at bestemme minimum for en funktion er en af standardteknikkerne i differentialregning.

Proportionalitetsregression 2.2

Som opvarmning til generel lineær regression, ser vi først på det simplere problem med at bestemme den rette linje gennem (0, 0), som bedst tilnærmer et datasæt $\{(x_i, y_i)\}_{i=1}^{n}$. Vi ser altså i første omgang bort fra friheden til at vælge begyndelsespunktet. Med 'bedst tilnærmer' menes hele tiden den linje, hvis **kvadratsum er minimal**.

I rigtig mange tilfælde er det faktisk den mest relevante form for regression, da rigtig mange sammenhænge i naturvidenskab er proportionalitetssammenhænge.

Problemet løses ved brug af følgende sætning:

S	**æ**	**t**	**n**	**i**	**n**	**g**

Den linje gennem (0, 0), der bedst tilnærmer et datasæt $\{(x_i, y_i)\}_{i=1}^{n}$, har en hældning, a_{reg}, givet ved:

$$a_{\text{reg}} = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$$

Bevis 1 (differentialregning)

Vi ser på kvadratsummen (summen af kvadraterne på afvigelserne) som funktion af linjens hældning a:

$$S(a) = \Sigma(y_i - ax_i)^2 = \Sigma y_i^2 - 2a\Sigma y_i x_i + a^2 \Sigma x_i^2$$

For at finde den værdi af a, der gør $S(a)$ mindst, løser vi ligningen $S'(a_{\text{reg}}) = 0$:

$$S'(a_{\text{reg}}) = 0 \Leftrightarrow 2\Sigma y_i x_i + 2a_{\text{reg}}\Sigma x_i^2 = 0 \Leftrightarrow a_{\text{reg}} = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$$

Vi kunne også have brugt toppunktsformlen, da $S(a)$ er et andengradspolynomium.

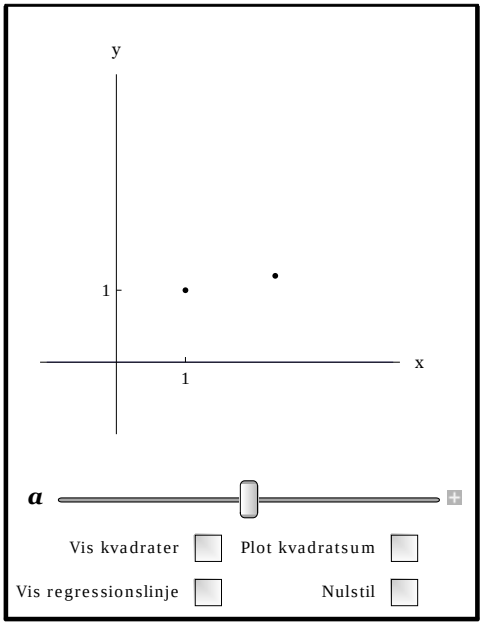
Bevis 2 (vektorprojektion)

Problemet med at finde den bedste rette linje gennem (0, 0) er ækvivalent med at finde det tal a_{reg}, som minimerer $\|ax - y\|$, hvor $\boldsymbol{x}$ er vektoren $(x_1, x_2, \ldots, x_n)$ og $\boldsymbol{y}$ er vektoren $(y_1, y_2, \ldots, y_n)$. Dette indser man ved at se på summen af kvadratet på afvigelserne fra bevis 1 og sammenligne med formlen for længden af en vektor. Men den længde er jo netop mindst, når $a\boldsymbol{x}$ er projektionen af $\boldsymbol{y}$ på $\boldsymbol{x}$ ifølge vores sætning om projektionsvektorer. Vi vælger altså:

$$a_{\text{reg}} = \frac{\boldsymbol{y} \cdot \boldsymbol{x}}{\boldsymbol{x}^2} = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$$

Illustration

Følgende figur viser regressionslinjen for forskellige datasæt og illustrerer også oversættelsen til andengrads-polynomier.



Hvordan skal jeg bruge figuren?

1. Du kan trække i punkterne for at flytte dem.
2. Klikker du ikke på et punkt, skaber du et nyt punkt (svarende til en ny måling).
3. Klik først på 'Vis regressionslinje' og lav så en håndfuld punkter, som ligger nogenlunde på en ret linje.
   a. Træk et af punkterne rundt for at få en fornemmelse for, hvilken betydning de enkelte punkter har for regressionslinjen.
   b. Undersøg om der er forskel på at flytte et punkt vandret og lodret. Har den ene retning større betydning for regressionslinjen end den anden? Hvad svarer en vandret/lodret ændring til i et forsøg?

c. Har punkterne med stor x-koordinat større betydning end dem med lille x-koordinat? (Prøv at flytte et punkt lodret både omkring $x = 4$ og omkring $x = 0$. Hvor sker der mest med linjens hældning?) Kan du finde en teoretisk forklaring på denne forskel? (Måske kan det hjæpe at klikke på 'Tegn kvadrater').

d. Hvis et punkt ligger langt fra den linje, som de andre punkter med god tilnærmelse ligger omkring, hvad sker der så med regressionslinjen? Hvad bør det få af betydning for dine fysik/kemi/biologi-forsøg?

4. Klik nu på 'Plot kvadratsum'. Frem kommer grafen for kvadratsummen som funktion af a, altså den som vi i bevis 1 kaldte $S(a)$. Bemærk at grafen er en parabel. Det blå punkt svarer til hældningen, som man kan vælge med skyderen.

Regressionslinje gennem et punkt

2.2.1

Med en lille smule mere arbejde kan vi tvinge regressionslinjen gennem et vilkårligt punkt i stedet for (0, 0):

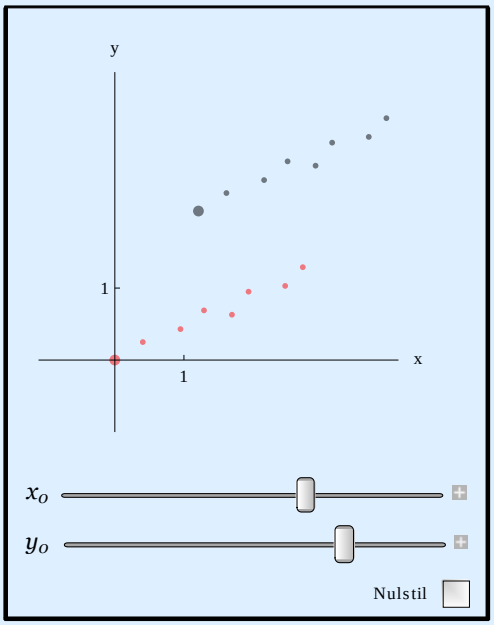
L e m m a

Den linje gennem (x_0, y_0), der bedst tilnærmer et datasæt $\{(x_i, y_i)\}_{i=1}^n$, har en hældning, a_{reg}, og begyndelsesværdi, b_{reg}, givet ved:

$$a_{reg} = \frac{\Sigma(x_i - x_0)(y_i - y_0)}{\Sigma(x_i - x_0)^2}, \quad b_{reg} = y_0 - a_{reg}x_0$$

Bevis

På nedenstående figur kan man se, at datasættet $\{(x_i - x_0, y_i - y_0)\}_{i=1}^n$ svarer til en forskydning af datasættet $\{(x_i, y_i)\}_{i=1}^n$ med x_0 enheder til venstre og y_0 enheder nedad. Det store sorte punkt er (x_0, y_0) og det store røde er (0, 0).



Hældningen af den linje gennem (x_0, y_0), der bedst tilnærmer $\{(x_i, y_i)\}_{i=1}^n$, har derfor samme hældning som den linje gennem (0, 0), der bedst tilnærmer$\{(x_i - x_0, y_i - y_0)\}_{i=1}^n$. Fra vores sætning om proportionalitetsregression, får vi derfor som ønsket:

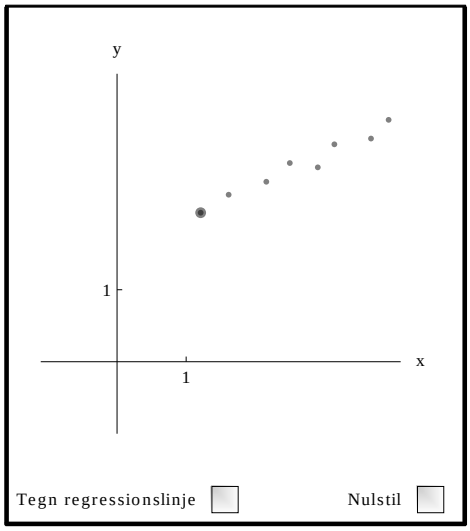
$$a_{reg} = \frac{\Sigma(x_i - x_0)(y_i - y_0)}{\Sigma(x_i - x_0)^2}$$

Da regressionslinjen går i gennem (x_0, y_0) kan vi bruge vores formel for begyndelsesværdien til at bestemme b_{reg}:

$$b_{reg} = y_0 - a_{reg}x_0$$

Illustration

Følgende figur illustrerer regression gennem et vilkårligt punkt (det store). Du kan flytte punkterne, lave nye og fjerne punkter. For at fjerne punkter skal du holde Alt (PC) eller Cmd (Mac) tasten nede.



Lineær regression

2.3

I dette afsnit bestemmer vi den rette linje, som bedst tilnærmer et datasæt $\{(x_i, y_i)\}_{i=1}^n$. I forhold til sidste afsnit er det nye altså, at vi nu kan vælge linjens begyndelsesværdi (skæring med y-aksen) frit. Den følgende sætning løser vores problem. Bemærk, at $\tilde{x}$ betyder gennemsnittet af x-værdierne $\left(\tilde{x} = \frac{\Sigma x_i}{n}\right)$.

Sætning **3**

Den linje, der bedst tilnærmer et datasæt $\{(x_i, y_i)\}_{i=1}^n$, har en hældning a_{reg} og begyndelsesværdi b_{reg} givet ved:

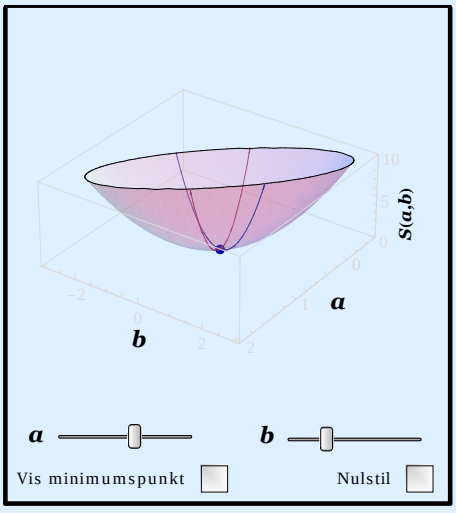
$$a_{\text{reg}} = \frac{\Sigma(x_i - \tilde{x})(y_i - \tilde{y})}{\Sigma(x_i - \tilde{x})^2}, \quad b_{\text{reg}} = \tilde{y} - a_{\text{reg}}\tilde{x}$$

Bevis 1 (differentialregning)

Vi ser på kvadratsummen (summen af kvadraterne på afvigelserne) som funktion af linjens hældning a og begyndelsesværdi b:

$$S(a, b) = \Sigma(y_i - (ax_i + b))^2 = \Sigma y_i^2 + a^2\Sigma x_i^2 + nb^2 - 2\left(a\Sigma y_i x_i + b\Sigma y_i\right) + 2ab\Sigma x_i$$

Ved at se på grafen for $S(a, b)$:



kan vi se, at $(a_{\text{reg}}, b_{\text{reg}})$ svarer til det eneste punkt, som ligger i bunden af både den røde og den blå kurve, som er graferne for funktionerne $S(\diamond, b)$ og $S(a, \diamond)$. a_{reg} er derfor minimumsstedet for $S(\diamond, b_{\text{reg}})$ og b_{reg} er minimumsstedet for $S(a_{\text{reg}}, \diamond)$. Da både $S(\diamond, b_{\text{reg}})$ og $S(a_{\text{reg}}, \diamond)$ kun afhænger af en variabel, ved vi hvordan vi bestemmer deres minimumssteder: Differentier og sæt lig 0. Vi har derfor følgende ligningssystem:

$$0 = \frac{dS}{da}(a_{\text{reg}}, b_{\text{reg}}) = 2a_{\text{reg}}\Sigma x_i^2 - 2\Sigma y_i x_i + 2b_{\text{reg}}\Sigma x_i$$
$$0 = \frac{dS}{db}(a_{\text{reg}}, b_{\text{reg}}) = 2nb_{\text{reg}} - 2\Sigma y_i + 2a_{\text{reg}}\Sigma x_i$$

Vi kunne nu bruge en af vores teknikker til løsning af to ligninger med to ubekendte til at få følgende løsning:

$$a_{\text{reg}} = \frac{n\Sigma y_i x_i - \Sigma y_i \Sigma x_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}, \quad b_{\text{reg}} = \frac{\Sigma y_i - a_{\text{reg}}\Sigma x_i}{n}$$

Derefter ville det bare være et spørgsmål om symbolmanipulation at komme frem til det ønskede.

I stedet for den ret tekniske vej, vælger vi i stedet følgende mere elegante:

Af den anden ligning i ligningssystemet følger:

$$b_{\text{reg}} = \frac{\Sigma y_i - a_{\text{reg}}\Sigma x_i}{n} = \tilde{y} - a_{\text{reg}}\tilde{x} \;\Leftrightarrow\; \tilde{y} = a_{\text{reg}}\tilde{x} + b_{\text{reg}}$$

Punktet $(\tilde{x}, \tilde{y})$ ligger derfor på regressionslinjen. Dermed giver lemmaet om regressionslinjer gennem et punkt det ønskede:

$$a_{\text{reg}} = \frac{\Sigma(x_i - \tilde{x})(y_i - \tilde{y})}{\Sigma(x_i - \tilde{x})^2}$$

Bevis 2 (vektorregning)

a_{reg} og b_{reg} er de værdier af a og b som minimerer $\|\boldsymbol{y} - (a\boldsymbol{x} + b\boldsymbol{1})\|^2 = \|(\boldsymbol{y} - a\boldsymbol{x}) - b\boldsymbol{1}\|^2$. Ifølge sætningen om vektorprojektion er vektoren $b_{\text{reg}}\boldsymbol{1}$ derfor projektionen af $\boldsymbol{y} - a_{\text{reg}}\boldsymbol{x}$ på $\boldsymbol{1}$. Derfor er:

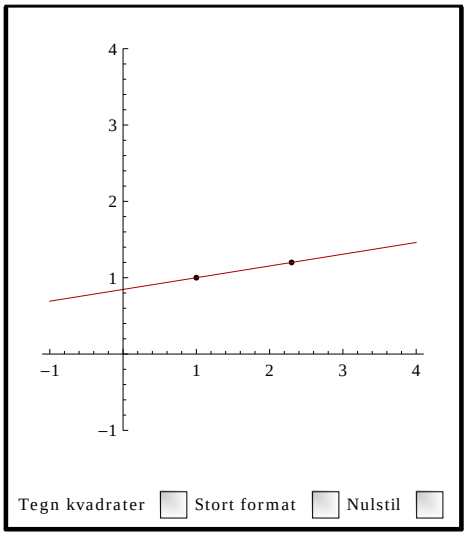
$$b_{\text{reg}} = \frac{(\boldsymbol{y} - a_{\text{reg}}\boldsymbol{x}) \cdot \boldsymbol{1}}{\boldsymbol{1}^2} = \frac{\Sigma(y_i - a_{\text{reg}}x_i)}{\Sigma 1^2} = \frac{\Sigma y_i - a_{\text{reg}}\Sigma x_i}{n} = \tilde{y} - a_{\text{reg}}\tilde{x}$$

Dvs. $\tilde{y} = a_{\text{reg}}\tilde{x} + b_{\text{reg}}$, hvilket betyder, at punktet $(\tilde{x}, \tilde{y})$ ligger på regressionslinjen. Dermed giver lemmaet om regressionslinjer gennem et punkt det ønskede:

$$a_{\text{reg}} = \frac{\Sigma(x_i - \tilde{x})(y_i - \tilde{y})}{\Sigma(x_i - \tilde{x})^2}$$

Illustration

Følgende figur viser regressionslinjen for forskellige datasæt:



Hvordan skal jeg bruge figuren?

1. Du kan trække i punkterne for at flytte dem.
2. Klikker du ikke på et punkt, skaber du et nyt punkt (svarende til en ny måling).

3. Prøv at lave en håndfuld punkter, som ligger nogenlunde på en ret linje.
   a. Træk et af punkterne rundt for at få en fornemmelse for, hvilken betydning de enkelte punkter har for regressionslinjen.
   b. Undersøg om der er forskel på at flytte et punkt vandret og lodret. Har den ene retning større betydning for regressionslinjen end den anden? Hvad svarer en vandret/lodret ændring til i et forsøg?
   c. Har punkterne med stor x-koordinat større betydning end dem med lille x-koordinat? Sammenlign med proportionalitetsregression. Prøv at forklare forskellen.
   d. Hvis et punkt ligger langt fra den linje, som de andre punkter med god tilnærmelse ligger omkring, hvad sker der så med regressionslinjen? Hvad bør det få af betydning for dine fysik/kemi/biologi-forsøg?
4. Flytter man på et punkt, opdager man typisk, at både hældning og begyndelsesværdi ændrer sig. Er det altid tilfældet? Prøv at eksperimentere dig frem til en ændring af et punkt, der kun resulterer i en ændring af begyndelsesværdien. Prøv derefter kun at ændre på hældningen. Kan du give en teoretisk forklaring? Prøv at tage udgangspunkt i formlerne for regressionslinjens hældning og begyndelsesværdi.

Forklaringsgraden

Vores mål i dette afsnit er at indføre en størrelse, som vi kan bruge som indikator på, hvor sikre vi kan være på, at der er en lineær sammenhæng mellem de målte størrelser x og y. Vi kan naturligvis ikke håbe på at kunne sige noget om en årsagssammenhæng mellem de to størrelser - vi kan kun udtale os om hvorvidt målepunkterne med tilnærmelse kan siges at ligge på en ret linje.

Vi har følgende yderligheder:

På den ene side: Regressionslinjen er en vandret linje. I så fald tyder det på, at der ingen sammenhæng er mellem x og y, fordi vores bedste bud på en sammenhæng er en vandret linje, svarende til at y slet ikke afhænger af x.

På den anden side: Regressionslinjen er skrå og punkterne ligger præcis på den. I så fald er vi helt overbeviste om, at der er en lineær sammenhæng mellem størrelserne.

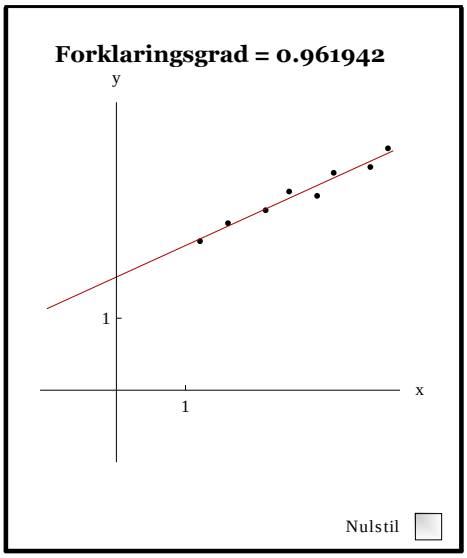
I det første tilfælde vil vi altså være 0% sikre på, at der er en lineær sammenhæng mellem størrelserne og i det andet ville vi være 100% sikre.

En størrelse som opfylder, at den er 0% = 0, når regressionslinjen er vandret og 100% = 1, når den er skrå og punkterne ligger på den er den såkaldte **forklaringsgrad** $(\boldsymbol{r^2})$, som er defineret på følgende måde:

Definition 1

$$r^2 = \text{forklaringsgrad} = 1 - \frac{\text{kvadratsum(regressionslinje)}}{\text{kvadratsum(vandret linje: } y = \tilde{y})} = 1 - \frac{\sum(y_i - (a_{\text{reg}}x_i + b_{\text{reg}}))^2}{\sum(y_i - \tilde{y})^2}$$

Følgende figur kan bruges til at undersøge forklaringsgraden for forskellige datasæt:



Hvordan skal jeg bruge figuren?

1. Du kan flytte på punkterne, tilføje punkter eller fjerne punkter. For at fjerne et punkt skal du holde Alt (PC) eller Cmd (Mac) tasten nede når du klikker på punktet.
2. Undersøg om det passer, at forklaringsgraden er 0, når regressionslinjen er vandret, og 1, når den er skrå og punkterne ligger præcis på den.
3. Eksperimentér dig frem til en grænse for, hvor stor du mener, forklaringsgraden skal være, før man med rimelighed kan sige, at der er tale om en lineær sammenhæng.

Af: Jan Agentoft Nielsen, lektor, ph.d. - Rødkilde gymnasium

version: 15.01.12