Kapittel 2
Utforske data—
Sammenhenger

2.1 Sammenhenger

2.2 Spredningsplott

2.3 Korrelasjon

2.4 Minste-kvadraters regresjon

2.5 Forsiktighetsregler for korrelasjon og regresjon

2.7 Spørsmålet om kausalitet

2.1 Sammenhenger

- Hva er en sammenheng (assosiasjon) mellom variable?
- Forklarings- og respons-variable
- Viktige karakteristikker til et datasett

Mye interessant statistikk involverer sammenhenger mellom par av variable

Det er en **sammenheng** mellom to variable målt på de samme individene (de er **assosiert)** hvis det å kjenne verdien av en av variablene sier deg noe du ellers ikke ville visst om verdien til den andre variabelen.

En **responsvariabel** måler utfallet av en studie. En **forklaringsvariabel** forklarer eller forårsaker forandringer i respons-variabelen.

Viktige aspekter ved et datasett når man skal undersøke assosiasjonen mellom to variabler

- ✓ Individer: Identifiser individene og hvor mange de er i datasettet.
- ✓ For å undersøke assosiasjonen mellom to variable må man
  - ✓ registrere begge variablene for hvert individ
  - ✓ vite hvilke målinger av de to variablene som hører til hvert individ
- ✓ Klassifiser hver variabel som kategorisk eller kvantitativ.
- ✓ Når det er relevant, klassifiser hver variabel som forklarings- eller respons-variabel.

Hvordan påvirker mengden av et bestemt legemiddel konsentrasjonen av legemiddelet i blodet?

- Eksempel på et planlagt eksperiment:
  - Ulike mus inntok ulike mengder legemiddel, konsentrasjonen ble målt etter 1 time
  - Responsvariabel: Konsentrasjon i blodet
  - Forklaringsvariabel: Mengde legemiddel
- Planlagt eksperiment: Bestemmer verdier av forklaringsvariabelen (mengde legemiddel) og undersøker på effekten på responsvariabelen (konsentrasjon i blodet)

Eksempel - observasjonsstudie

- Individer: 40-åringer innkalt til helseundersøkelse
- Observerte variabler:
  - Høyde/vekt - BMI
  - Blodtrykk
  - Røyking
  - Kolesterol
- Formålet er å studere hvordan risikoen for hjerte- og karsykdom avhenger av de observerte variablene

Når du undersøker (om det er) en sammenheng mellom to variable, dukker det opp et viktig spørsmål:

Er formålet bare å undersøke hvordan sammenhengen er, eller ønsker du å vise at en av variablene kan forklare variasjonen i den andre?

Kausalitet er årsakssammenheng

- Sammenheng mellom to variable, hvor den ene (responsen) er en konsekvens av den andre (forklaringsvariabelen)

Kausalitet

- Kausale sammenhenger er ofte det egentlige målet med studien, men det er vanskelig å bevise:
- Er sammenhengen er forårsaket av andre variable (confoundere = sammenblandende variable = 'lurkende' variable)?
- Hvilken vei går egentlig årsakssammenhengen?

Selv om vi finner en assosiasjon eller sammenheng så betyr ikke det at vi har påvist årsak!

2.2 Spredningsplott

- Spredningsplott
- Tolke spredningsplott
- Kategoriske variable i spredningsplott

Et spredningsplott er det nyttigste grafiske verktøyet for å vise sammenhengen mellom to kvantitative variable

Et **spredningsplott** viser sammenhengen mellom to kvantitative variable målt på de samme individene.

Hvert individ tilsvarer ett punkt i grafen.

Verdiene til en variabel vises langs den horisontale aksen, og verdiene til den andre vises langs den vertikale aksen.

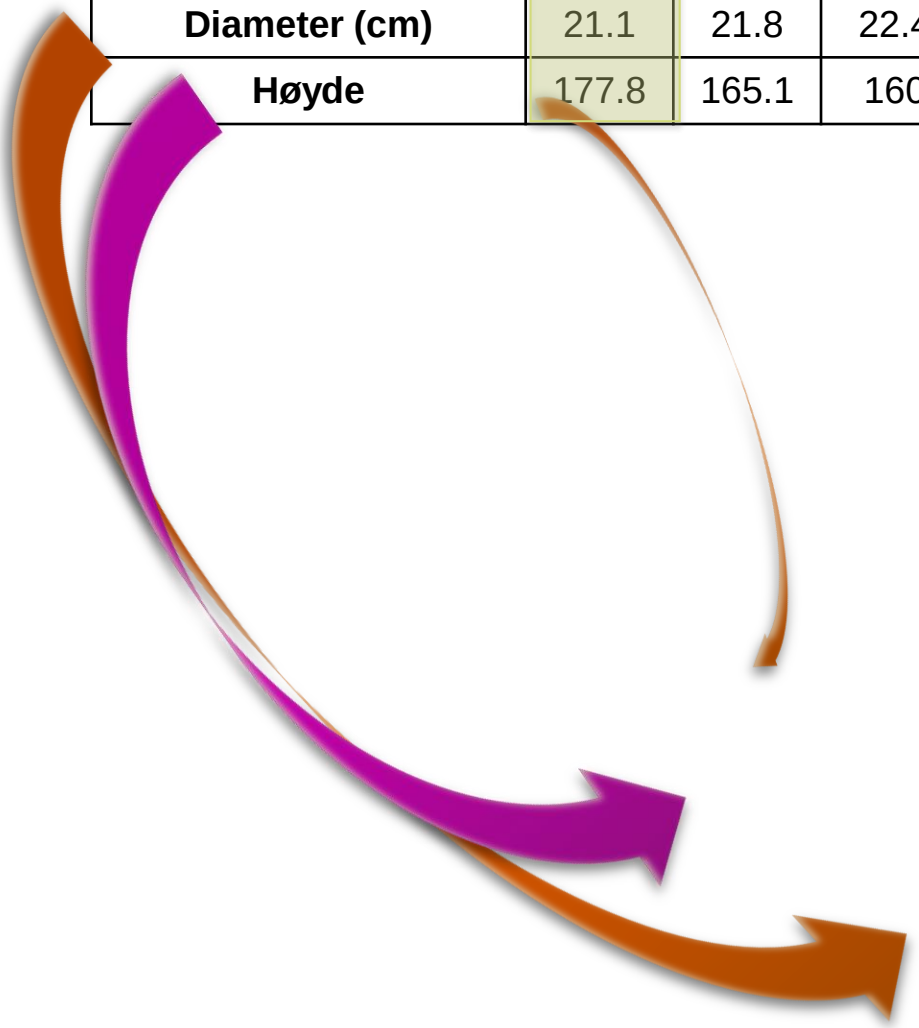
Å lage et spredningsplott

1. Bestem hvilken variabel som skal være langs hvilken akse.
   Hvis det er en naturlig oppdeling i forklarings- og responsvariable, velg forklaringsvariabelen langs x-aksen og responsvariabelen langs y-aksen.
2. Gi navn og skala til aksene.
3. Plott de individuelle data-verdiene.

Spredningsplott 2

Eksempel: Lag et spredningsplott for sammenhengen mellom diameter (i 137 cm høyde) og høyde for 31 felte kirsebærtrær ('trees' fra R-pakken 'datasets').

Diameter (cm)	21.1	21.8	22.4	26.7	27.2	27.4	…	52.3
Høyde	177.8	165.1	160	182.9	205.7	210.8	…	221



For å tolke et spredningsplott følger du de grunnleggende strategiene fra kapittel 1: Se etter mønstre og viktige avvik fra disse mønstrene.

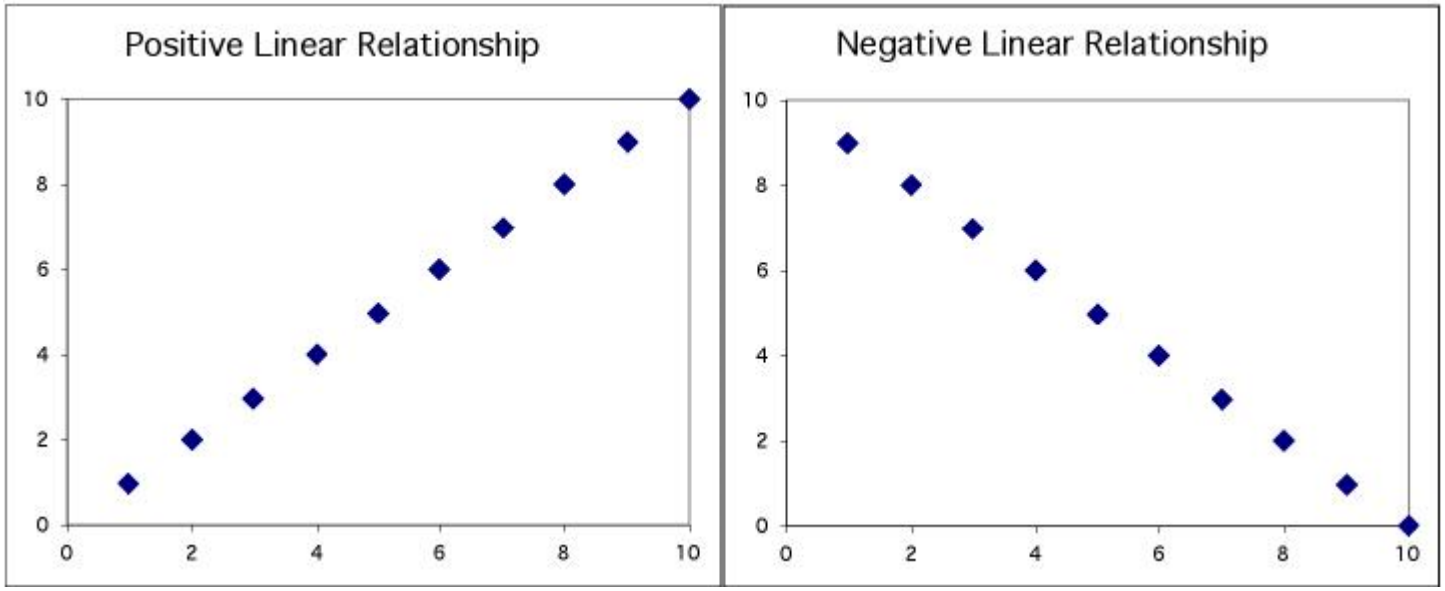
Som i enhver grafisk fremstilling av data, se etter det *overordnede mønsteret* og etter påfallende *avvik* fra dette mønsteret.

- Du kan beskrive det overordnede mønsteret til et spredningsplott ved **formen**, **retningen**, og **styrken** til sammenhengen.
- En viktig type avvik er en **uteligger**: en individuell verdi som faller utenfor det overordnede mønsteret til sammenhengen.

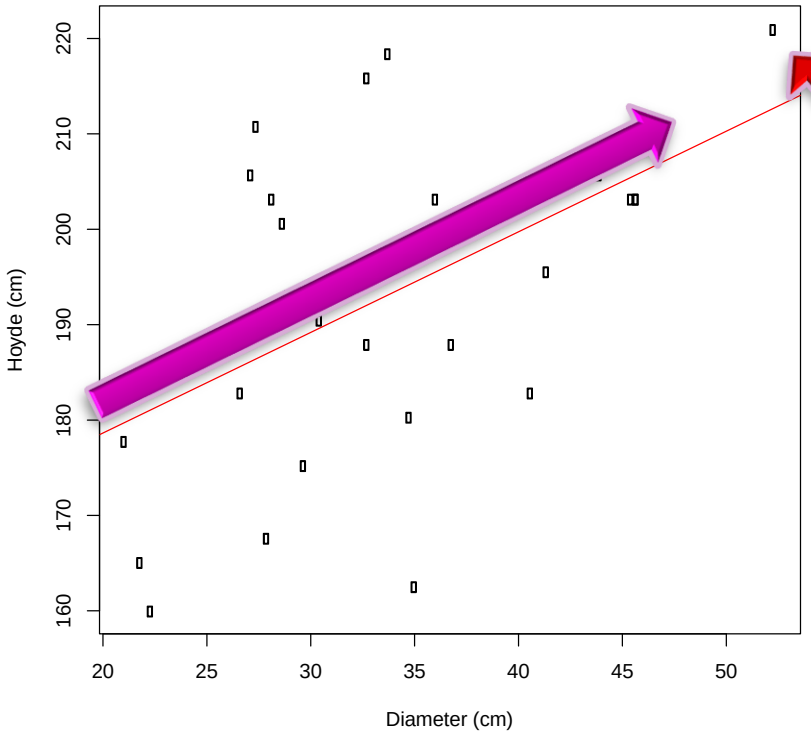
Positiv og negativ lineær sammenheng

To variable har en **positiv sammenheng** når over gjennomsnittlige verdier av en variabel tenderer til å høre sammen med over gjennomsnittlige verdier av den andre, og når under gjennomsnittlige verdier også tenderer til å forekomme sammen.

To variable har en **negativ sammenheng** når over gjennomsnittlige verdier av en variabel tenderer til å høre sammen med under gjennomsnittlige verdier av den andre og vice versa.



Eksempel: Spredningsplott av diameter og høyde for 31 felte kirsebærtrær



Uteligger

- ✓ Det er en mulig uteligger

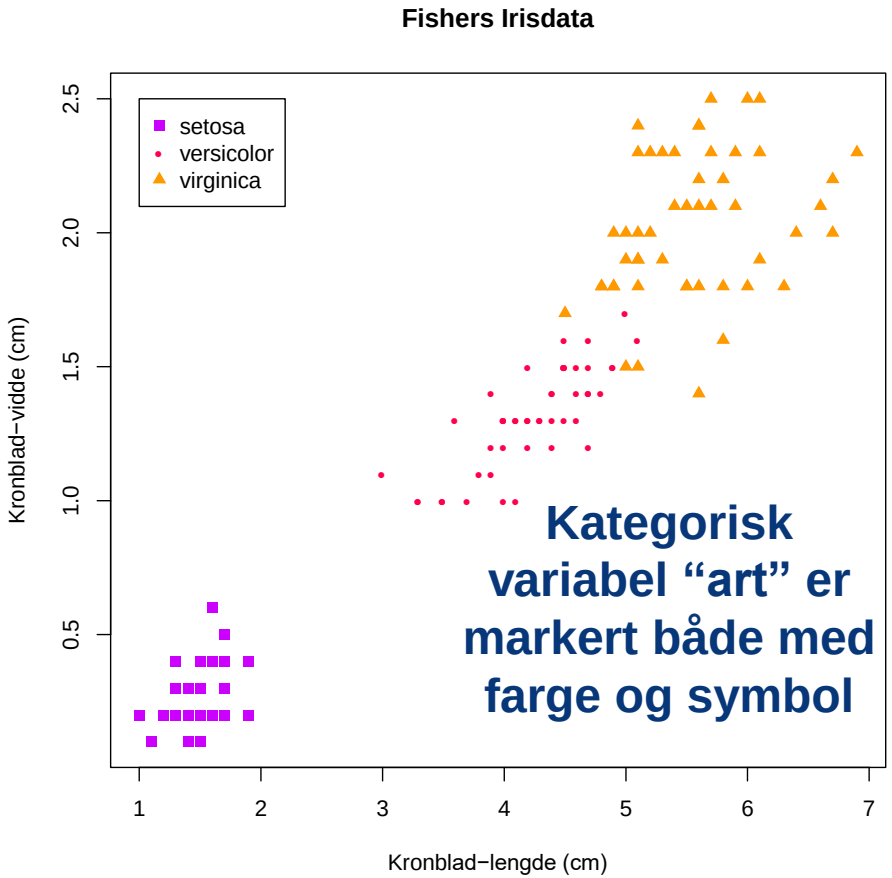
Styrke **Retning** **Form**

- ✓ Det er en svak, positiv, lineær sammenheng mellom diameter og høyde på trærne.
- ✓ Det ser ut som om trær med større omkrets tenderer til å være høyere enn trær med mindre omkrets.

Sammenhenger innenfor ulike kategorier

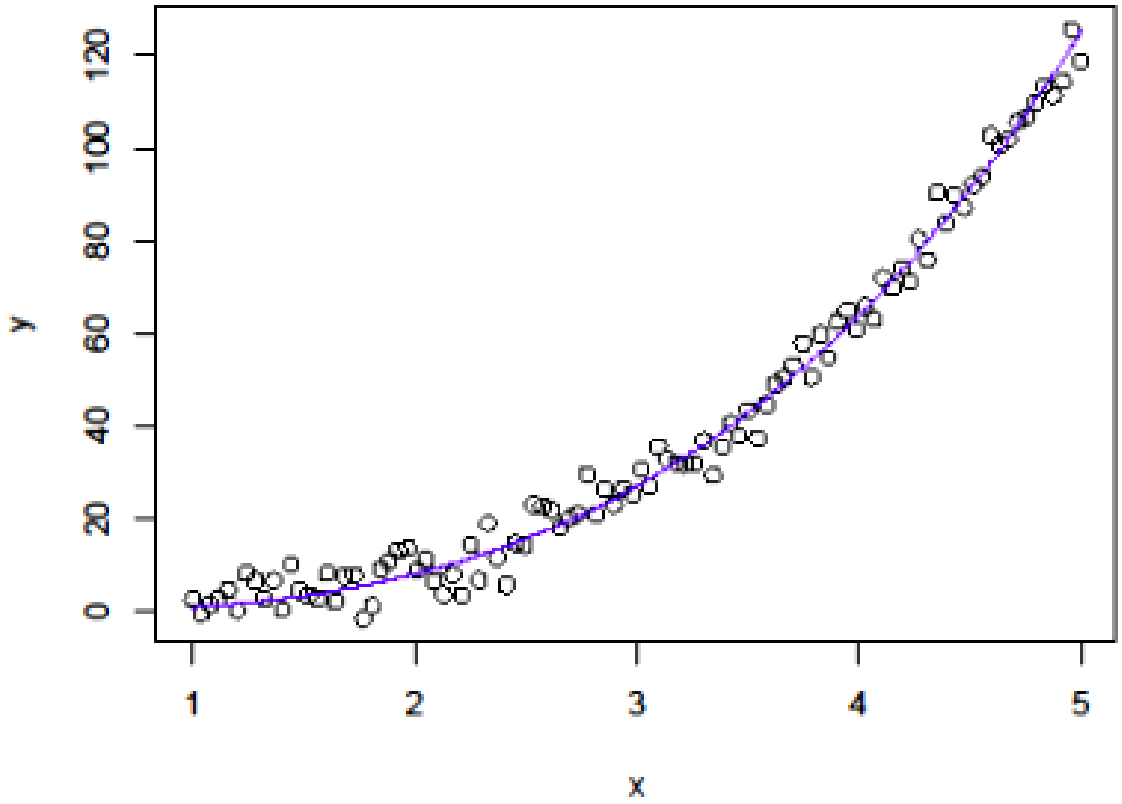
- For å se på sammenhenger innenfor ulike verdier av en *kategorisk variabel*, bruk forskjellige farger og/eller plott-symboler for hver kategori:

  - 'Fishers Irisdata' (hentet fra R-pakken 'datasets'): 50 målinger av lengde og bredde på begerblad og kronblad for hver av 3 arter av Iris (Iris setosa, Iris virginica og Iris versicolor).
  - Plotter sammenhengen mellom kronblad-lengde og kronblad-bredde for de tre artene.

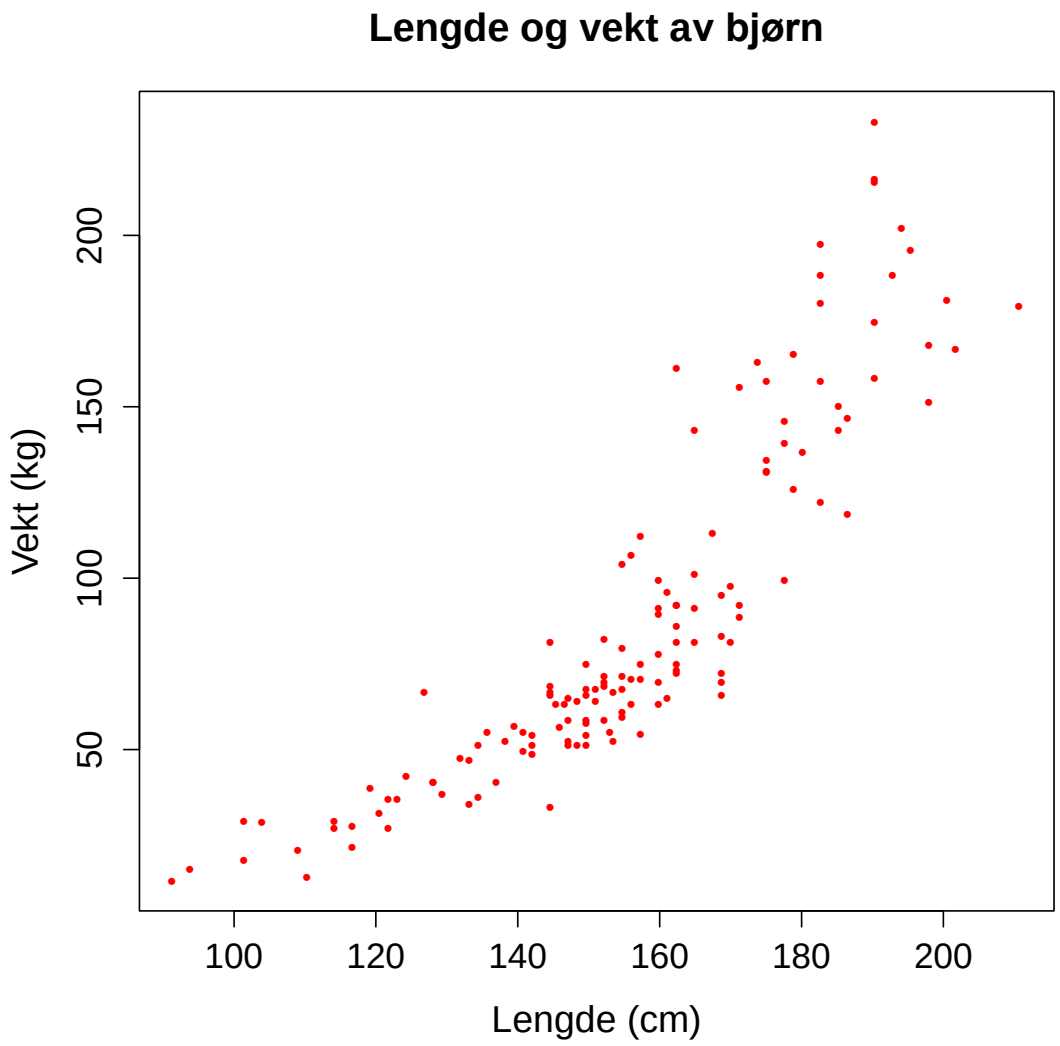


Det finnes andre former for sammenhenger enn lineære

- Spredningsplottet under er et eksempel på en ikke-lineær form.
- Merk kurvaturen i sammenhengen mellom *x* og *y*.



Er det sammenheng mellom lengde og vekt for 143 bjørner?



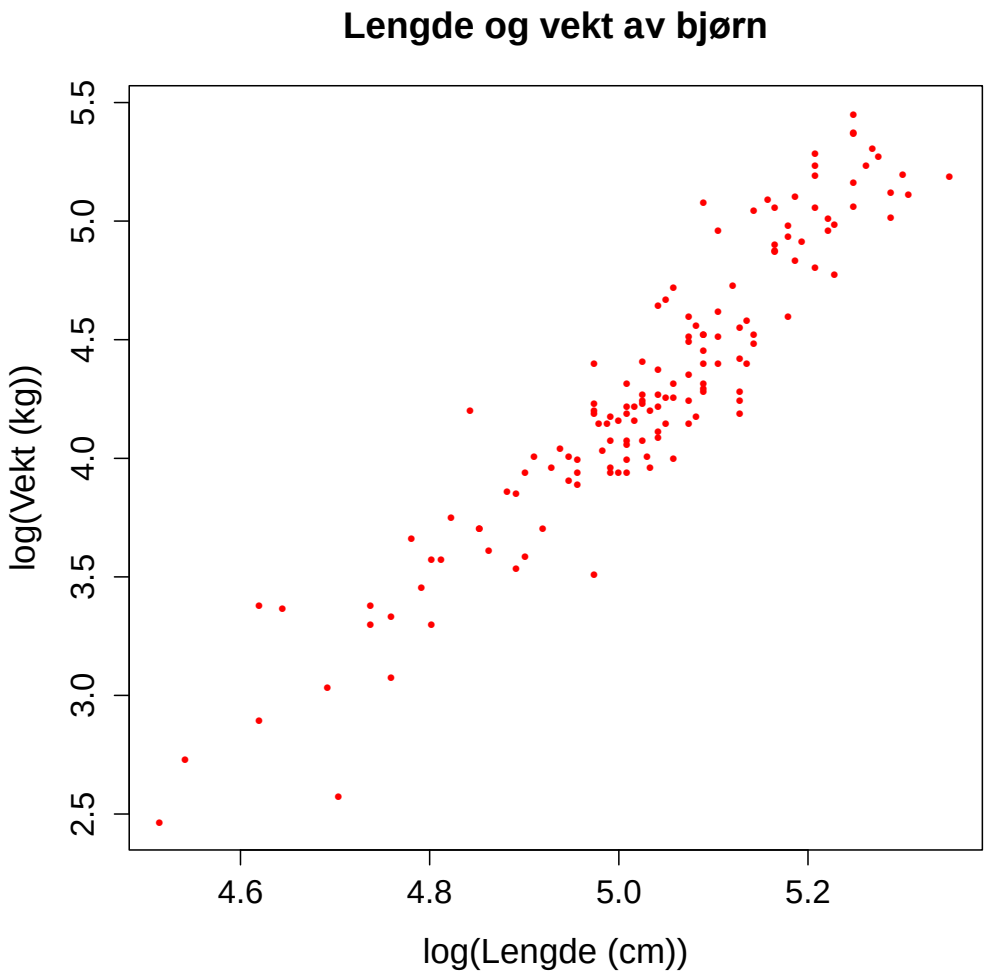
Spredningsplottet viser en *ikke-lineær* sammenheng

Data hentet fra programvaren Minitab

Ved ikke-lineære sammenhenger *kan* transformasjoner være en god idé

- Den vanligste transformasjonen er log-transformasjon:
  å ta (den naturlige) logaritmen til en eller begge variablene

- Etter log-transformasjon av begge variablene ser vi tendenser til en *lineær sammenheng*



2.3 Korrelasjon

- Korrelasjons-koeffisienten *r*
- Egenskaper til *r*

Et spredningsplott viser styrken, retningen og formen til sammenhengen mellom to kvantitative variable

- Lineære sammenhenger er viktige fordi en rett linje er et enkelt mønster som er ganske vanlig.
- Øynene våre er ikke alltid så gode til å bedømme hvor sterk en sammenheng er. Derfor bruker vi et **numerisk mål** som et tillegg til spredningsplottet vårt, som hjelper oss å **tolke styrken** av den lineære sammenhengen.

Korrelasjonen r måler styrken til den lineære sammenhengen mellom to kvantitative variable

Vi har data på variablene *x* og *y* for *n* individer. *x* har verdien x_i for individ nr *i*, og *y* har verdien y_i for individ nr *i*, *i=1*,...,*n*

Korrelasjon:

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Eksempel:

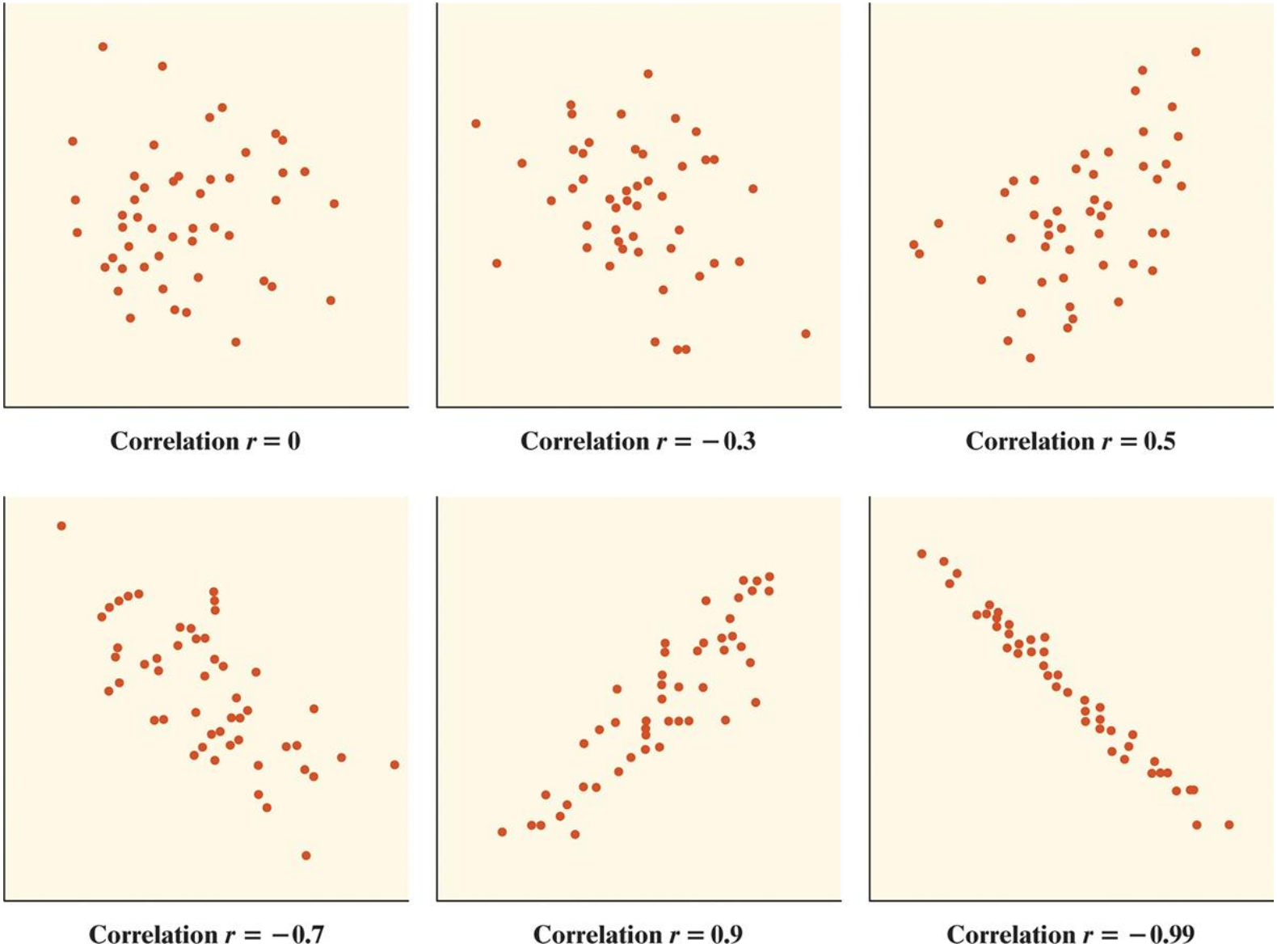
- Måler vekt (*x*) og høyde (*y*) for *n*=100 individer.
- Da er x_1 og y_1 vekt og høyde til individ nr 1, x_2 og y_2 vekt og høyde til individ nr 2, ... , x_i og y_i vekt og høyde til individ nr *i*,..., x_{100} og y_{100} vekt og høyde til individ nr 100.
- Gjennomsnitt og standardavvik er $\bar{x}$ og s_x for *x*-verdiene, og $\bar{y}$ og s_y for *y*-verdiene.

Vi sier at en lineær sammenheng er sterk hvis punktene ligger nær en rett linje og svak hvis de har stor spredning rundt en rett linje.

Egenskaper til korrelasjon

- r er alltid et tall mellom –1 og 1.
- $r > 0$ indikerer en *positiv* sammenheng, $r < 0$ indikerer en *negativ* sammenheng.
- Verdier av r nær 0 indikerer en veldig svak lineær sammenheng.
- Styrken til den lineære sammenhengen vokser når r beveger seg bort fra 0 mot –1 eller 1. De ekstreme verdiene $r = –1$ og $r = 1$ inntreffer bare dersom det er en perfekt lineær sammenheng.
- Korrelasjonen r er symmetrisk, dvs skiller ikke på forklarings og responsvariable.
- r har ingen enhet og forandrer seg ikke når vi endrer måleenhet for x eller y eller begge.

Korrelasjon



Korrelasjon er ikke en komplett oppsummering av to-variabel-data

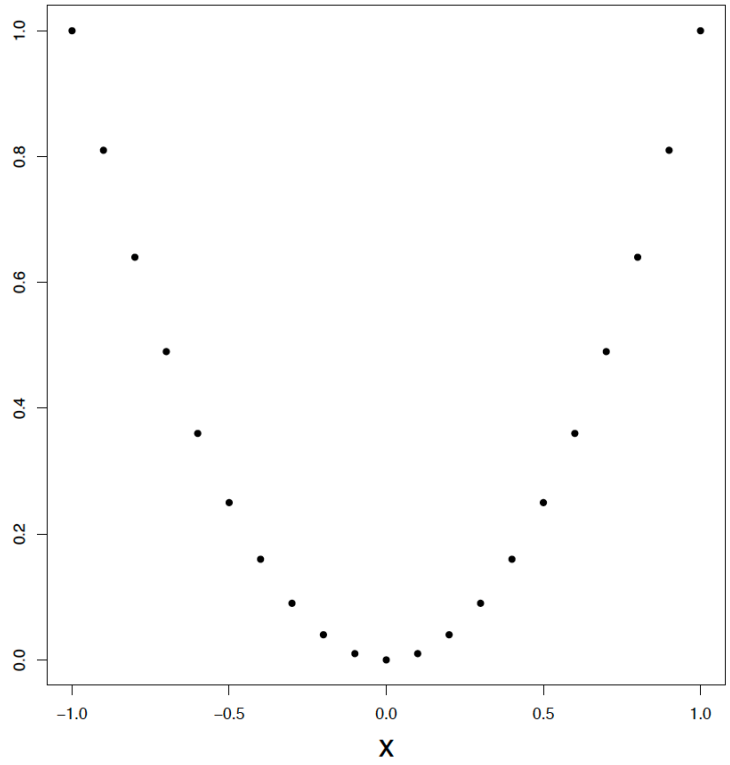
- Korrelasjon krever at begge variablene er kvantitative.
- Korrelasjon *beskriver* ***ikke*** *ikke-lineære sammenhenger* mellom variable, uavhengig av hvor sterk sammenhengene er.
- Korrelasjonen *r* er ikke robust mot uteliggere, og kan bli sterkt påvirket av noen få uteliggende observasjoner.

Korrelasjon måler bare lineær sammenheng

Eks: Anta at vi eksakt har $y_i = x_i^2$ samt at

x_1=-1.0, x_2=-0.9,...., x_{19}=0.9, x_{20}=1.0

Da ligger dataene perfekt (uten spredning) langs en parabel, men korrelasjonen $r = 0$.



Eksempeldatasett i R:
diameter og høyde til 31 kirsebærtrær

2.4 Minste-kvadraters regresjon

- Regresjonslinjer
- Minste-kvadraters regresjons-linje
- Prediksjoner
- Fakta om minste kvadraters regresjon
- Korrelasjon og regresjon

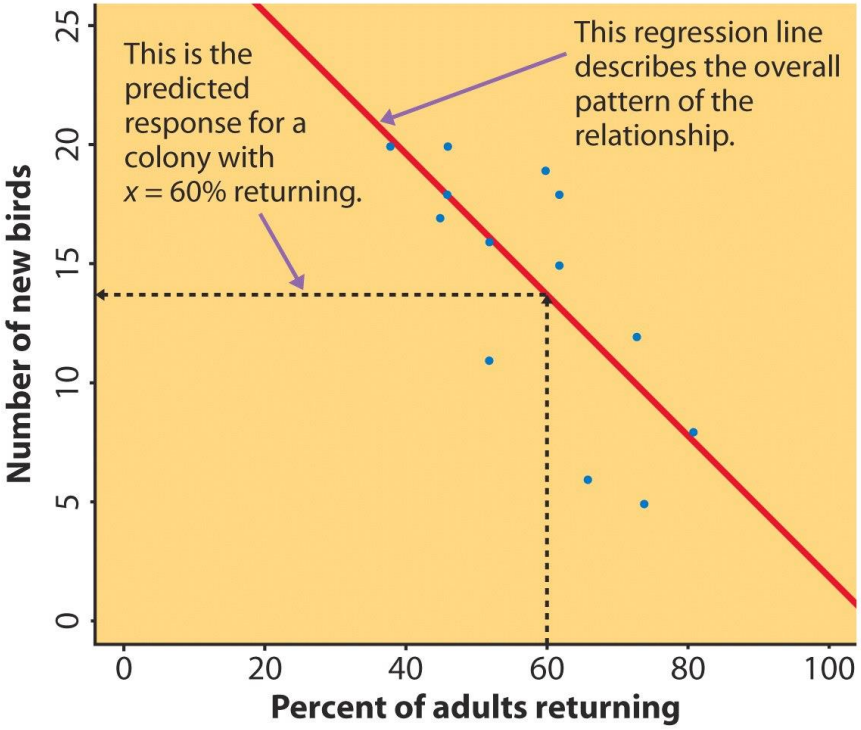
En regresjonslinje er en rett linje som beskriver hvordan en responsvariabel y endres når en forklaringsvariabel x endres.

- Vi kan bruke en regresjons-linje til å *predikere* verdien til *y* for en gitt verdi av *x*
- Krever en responsvariabel og en forklaringsvariabel

Hvis 60% av voksne fugler fra i fjor returnerer, hvor mange nye, voksne fugler predikeres å bli med i kolonien?

Eksempel: Gitt målinger av prosentandel av voksne fugler fra året før som returnerer til kolonien (x) og antall nye, voksne fugler som blir med i kolonien (y), for hver av 13 fuglekolonier,

Kan man bruke dette til å predikere (uobservert) antall nye, voksne fugler som blir med i en koloni basert på observert prosentandel av voksne fugler fra året før som returnerer til kolonien?



Å tilpasse en linje til data betyr å finne en linje som går så nær datapunktene som mulig

Når et spredningsplott viser et lineært mønster kan vi beskrive det overordnede mønsteret ved å tegne en rett linje gjennom punktene

Regresjons-ligning: $\hat{y} = b_0 + b_1 x$

- ***x*** er verdien til forklarings-variabelen.
- $\hat{y}$ **(“*y-hatt*”)** er den predikerte verdien av respons-variabelen for en gitt verdi av *x*.
- b_1 er **stigningstallet,** mengden *y* forandres for hver økning i *x* på en enhet.
- b_0 er **konstant-leddet,** verdien til *y* når *x = 0.*

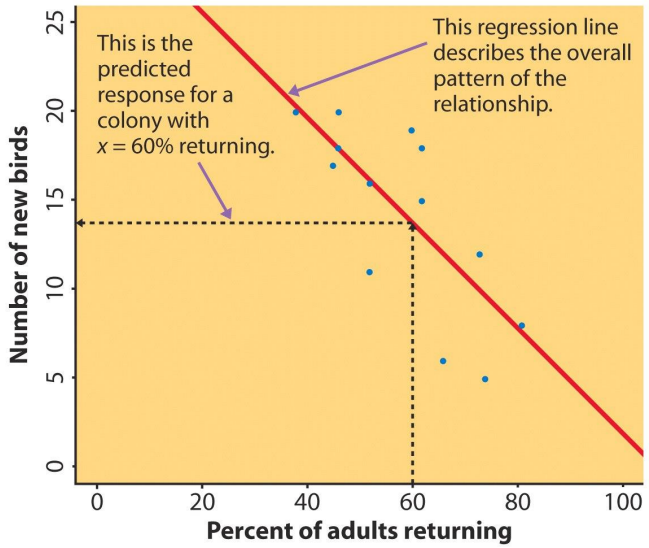
Regresjonslinje $\hat{y} = 31.9343 - 0.3040x$ for fuglekoloniene

- $b_0 = 31.9343$ er skjæringspunktet og beskriver antall nye fugler hvis ingen fugler fra i fjor returnerer
- $b_1 = -0.3040$ er stigningstallet og beskriver endringen (nedgangen) i antall nye fugler ved ett prosentpoengs økning %-andel returnerende fugler
- $\hat{y}$ er det predikerte antall nye fugler for kolonier med prosentandel ***x*** av returnerende fugler.

Anta vi vet at en koloni har 60% returnerende fugler. Hva ville vi ***predikere*** at antall nye fugler vil være for akkurat den kolonien?

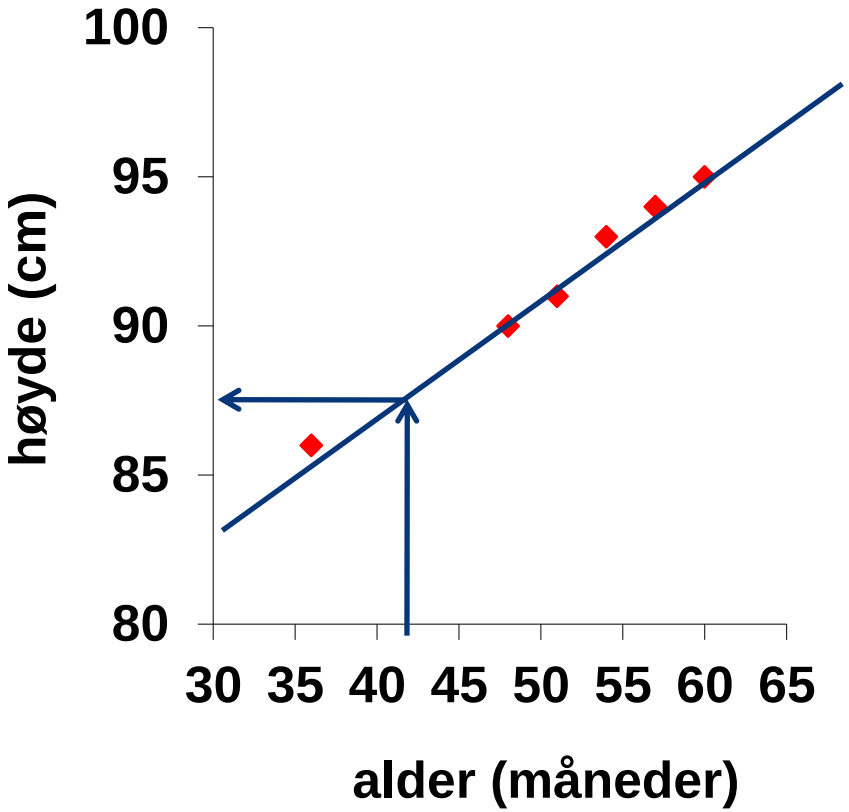
For kolonier med 60% returnering, ***predikerer*** vi antall nye fugler til å være

$$31.9343 - (0.3040)(60) = \mathbf{13.69} \text{ fugler}$$



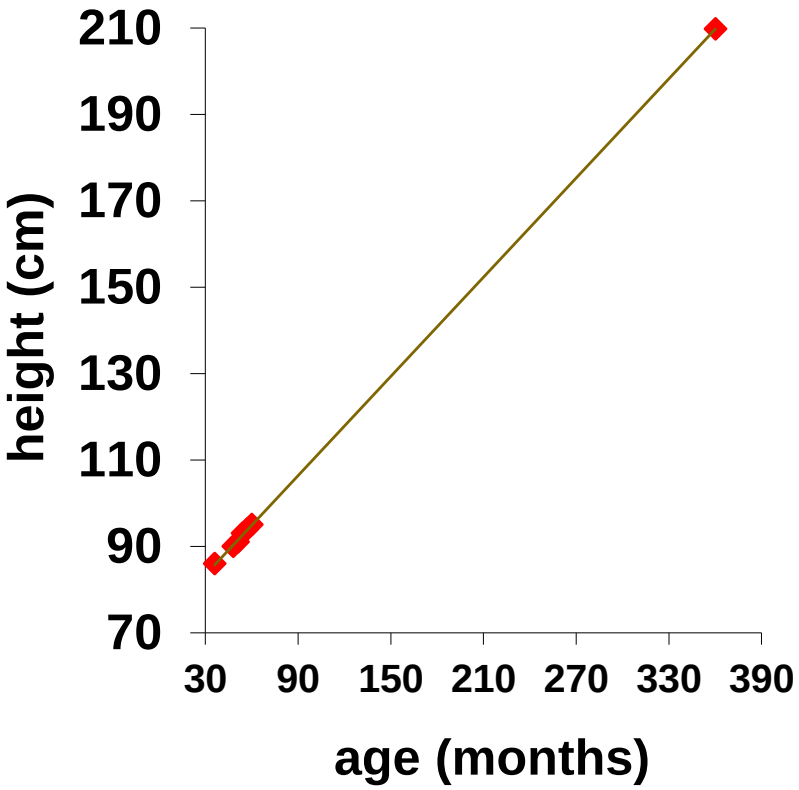
Å interpolere og ekstrapolere

- Høyden til ei jente ble plottet mot alderen hennes.
- Kan du gjette (predikere) hva høyden hennes var da hun var 42 måneder?
- Kan du predikere hva høyden hennes vil bli når hun er 30 år gammel (360 måneder)?

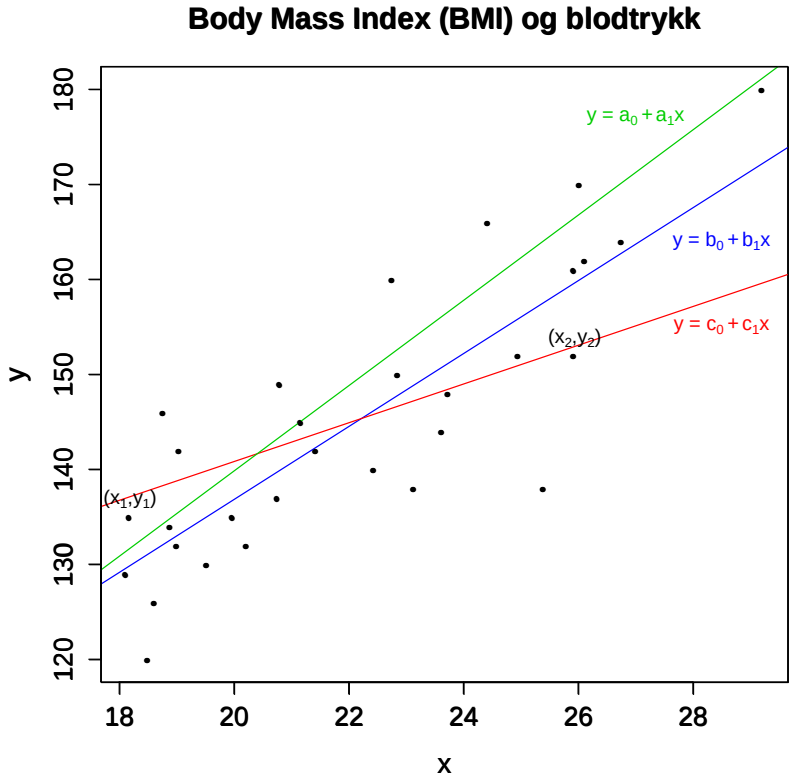


Å ekstrapolere er risikabelt

- Regresjonslinje:
  y-hatt* = 71.95 + 0.383 *x
- Høyde ved alder 42 måneder? ***y-hatt* = 88**
- Høyde ved alder 30 år?
  ***y-hatt* = 209.8**
- Hun predikeres til å bli 209.8 cm når hun blir 30 år gammel! *Høres det rimelig ut? Hva er galt?*



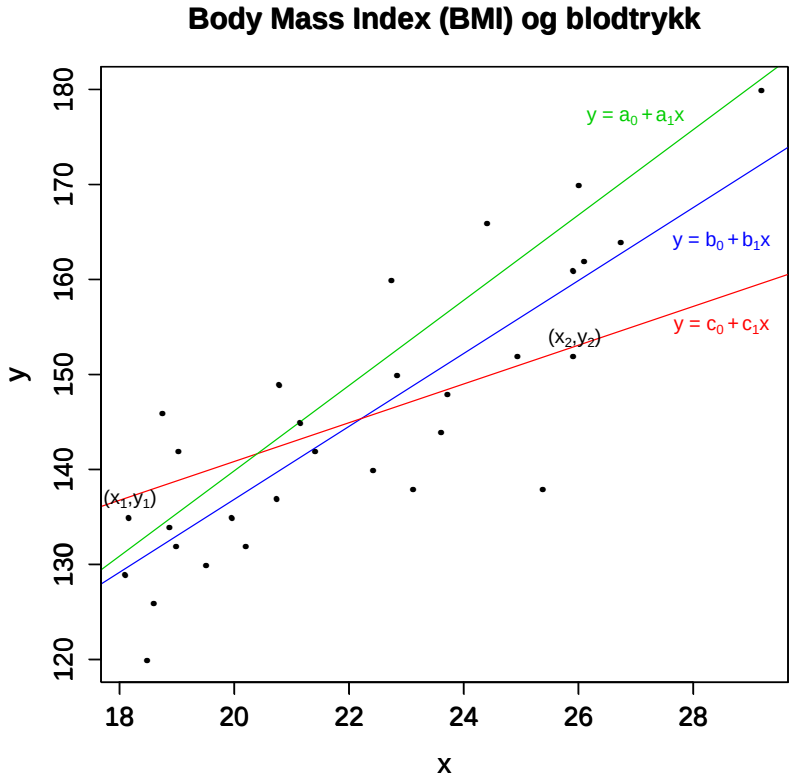
Hvilken linje er «best»?



- Eksempel: Kryssplot av målinger av BMI (*x*) og blodtrykk (*y*) fra 30 individer
- Ønsker å tilpasse en regresjons-linje til datapunktene
- Hvordan måle hva som er «best»?

Når vi ønsker å predikere y, søker vi at regresjonslinja går så nær datapunktene i den vertikale (y-) retninga som mulig

Regresjonslinja kan ikke gå gjennom alle punktene med mindre alle punktene ligger på en rett linje! Dermed blir vanligvis ikke den predikerte verdien $\hat{y}$ lik den observerte verdien y for hver x.



Minste kvadraters regresjonslinje

Anta at vi har data på en forklarings-variabel *x* og en respons-variabel *y.*

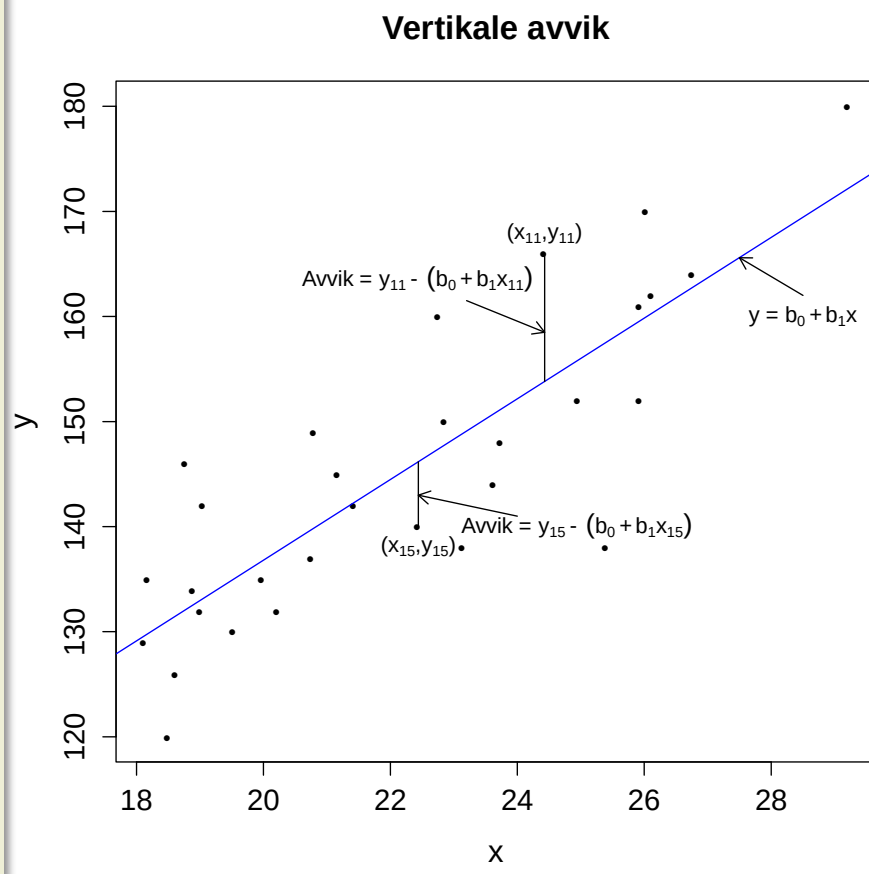
Minste kvadraters regresjonslinja av *y* på *x* er gitt av ligninga

$$\hat{y} = b_0 + b_1 x$$

som *minimerer summen av kvadratene av de vertikale avstandene/avvikene* til data-punktene fra linja, dvs som minimerer :

$$\sum (vertikale\ avvik)^2$$

$$= \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$



Minste kvadraters regresjonslinje 2

Ligninga for minste kvadraters-regresjonslinje (LSRL):

Minste kvadraters-regresjonslinja av *y* på *x*

$$\hat{y} = b_0 + b_1 x$$

som minimerer

$$\sum (vertikale\ avvik)^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

(kvadrat-summen av de vertikale avstandene til data-punktene fra linja)

har **stigningstall**

$$b_1 = r\frac{s_y}{s_x}$$

og **konstantledd**

$$b_0 = \bar{y} - b_1 \bar{x}$$

Her er som før $\bar{x}$ gjennomsnittet og s_x standardavviket til $x_1, x_2, \ldots, x_n$,
$\bar{y}$ gjennomsnittet og s_y standardavviket til $y_1, y_2, \ldots, y_n$,
og r korrelasjonen til $(x_1, y_1), \ldots (x_n, y_n)$

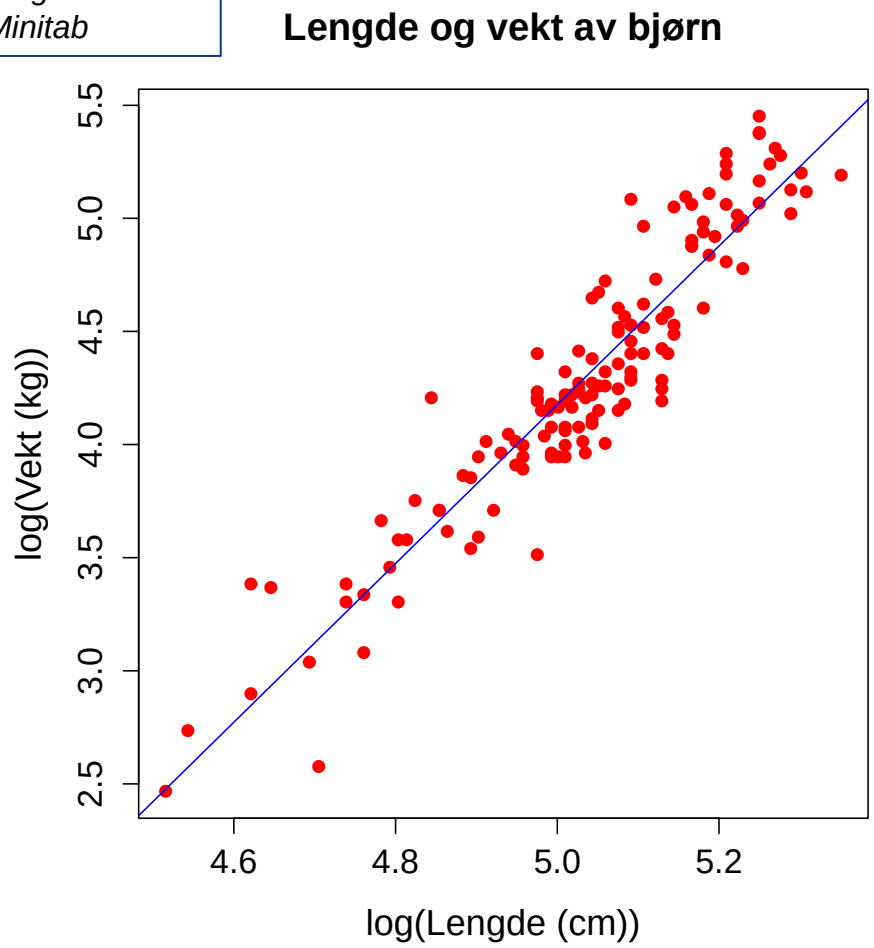
Minste kvadraters regresjonslinje for lengde av bjørn

- Log-lengde og log-vekt for 143 bjørner
- Ønsker å finne minste kvadraters-regresjonslinja:

- Fra R: $\bar{x} = 5.0353$, $\bar{y} = 4.2995, s_x = 0.1612, s_y = 0.6046, r = 0.9355$

- Formlene gir da:
- $b_1 = r\frac{s_y}{s_x} = 0.9355\frac{0.6046}{0.1612} = 3.5084$
- $b_0 = \bar{y} - b_1\bar{x} = 4.2995 - 3.5084 \cdot 5.0353 = -13.3663$

Data hentet fra programvaren Minitab



Regresjon i R: Lengde og vekt av bjørner

Data er lest inn i datasettet 'Bears.data' med variablene 'loglength' (logaritmen til lengde) og 'logweight' (logaritmen til vekt).

Beregningene gjøres ved kommandoen 'lm' (lineær modell) som følger:

```
> regr.Bears <- lm(logweight~loglength)

> regr.Bears


Call:
lm(formula = logweight ~ loglength)

Coefficients:
(Intercept)    loglength
    -13.366        3.508
```

Regresjon i R: Lengde og vekt av bjørner

Man får ut mer informasjon med kommandoen 'summary'

```
> summary(regr.Bears)

Call:
lm(formula = logweight ~ loglength)

Residuals:
     Min       1Q   Median       3Q      Max
-0.57529 -0.15137  0.00251  0.11904  0.58919

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.3663     0.5623  -23.77   <2e-16 ***
loglength     3.5084     0.1116   31.44   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2144 on 141 degrees of freedom
Multiple R-squared:  0.8751,    Adjusted R-squared:  0.8742
F-statistic: 988.2 on 1 and 141 DF,  p-value: < 2.2e-16
```

Vi skal se nærmere på slike utskrifter senere, merk nå kun

(multippel) R-square $0.8751 = (0.9355)^2 = r^2$ der r = korrelasjonen

Regresjon er veldig mye brukt, og minste-kvadrater er den vanligste metoden for å tilpasse en regresjonslinje til data.

- En endring på ett standardavvik i *x* tilsvarer en endring på *r* standardavvik i *y* (fordi $b_1 = r\frac{s_y}{s_x}$)
- Stigningstall og skjæringspunkt avhenger av skala
- Fortegn på stigningstall spesielt interessant
- Perfekt tilpasning hvis r = -1 eller 1
- Linja går alltid gjennom ($\bar{x}$, $\bar{y}$) (fordi $\hat{y} = b_0 + b_1 x$ og $b_0 = \bar{y} - b_1\bar{x}$ gir $\hat{y} = \bar{y}$ for $x = \bar{x}$)
- *Det er helt essensielt å skille mellom forklarings- og respons-variable*. Linja er basert på avstander i *y*-retning for gitte x-verdier, og dermed kan vi ikke bare rotere plottet for å få regresjonslinja med *x* som respons og *y* som forklaringsvariabel

Korrelasjon og regresjon

Minste kvadraters regresjon ser på avstandene til data-punktene fra linja bare i *y*-retning. Derfor har *x*- og *y*–variablene forskjellige roller i regresjon, hvis de bytter roller må man tilpasse en ny regresjonslinje!

Selv om korrelasjonen *r* er symmetrisk og ikke skiller på rollene til *x og y* er det en nær forbindelse mellom korrelasjon og regresjon.

Kvadratet av korrelasjonen, r^2, er andelen av variasjonen i *y*–verdiene som kan forklares av minste kvadraters regresjonslinja av *y* på *x*.

Med andre ord: I minste kvadraters regresjon er r^2 andel variasjon i respons-variabelen som kan forklares ved hjelp av forklaringsvariabelen

Variasjonen til observasjonene y_1,..., y_n er større enn variasjonen i de predikerte verdiene $\hat{y}_1$,..., $\hat{y}_n$. Vi har faktisk

$$r^2 = \frac{\text{varians for predikerte verdier } \hat{y}_1, \ldots, \hat{y}_n}{\text{varians for observerte verdier } y_1, \ldots, y_n}$$

r^2 = Andel forklart varians

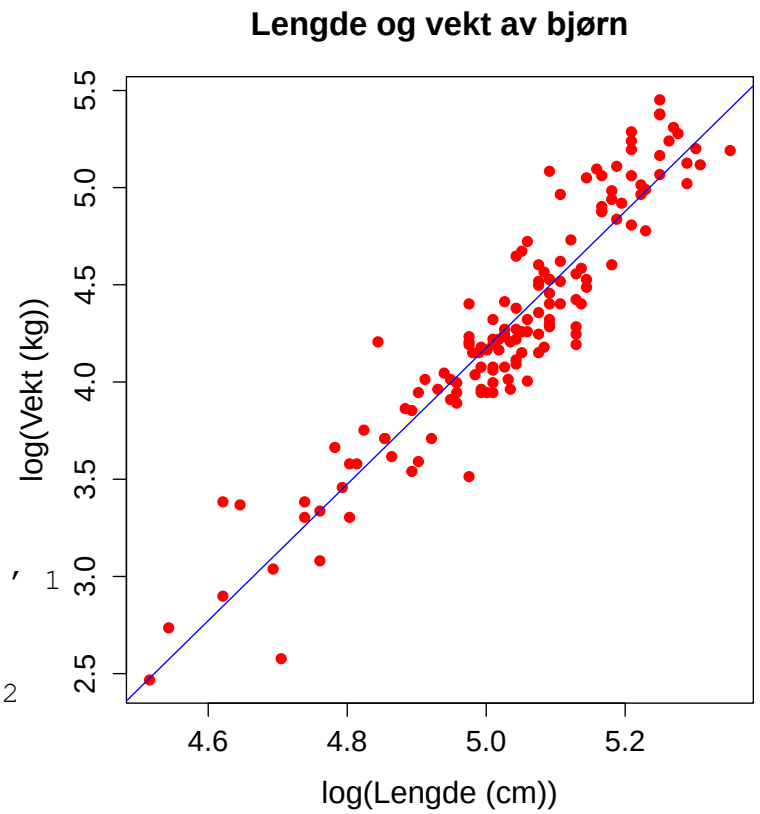
$$r^2 = \frac{\text{varians predikerte verdier } \hat{y}}{\text{varians observerte verdier } y}$$

Spesialtilfeller:

- Hvis all variabilitet i y kunne forklares ved x, ville de observerte y-ene ligget eksakt på regresjonslinja, dvs. samme varians for observerte og predikerte verdier. Så $r^2 = 1$
- Hvis det ikke er noen lineær sammenheng mellom x-er og y-er blir $r=0=b_1$, og $\hat{y}=b_0$ for alle verdier av x. Dermed er det ingen variasjon i predikerte verdier. Så $r^2 = 0$.

Eksempel r^2: Lengde og vekt av bjørner

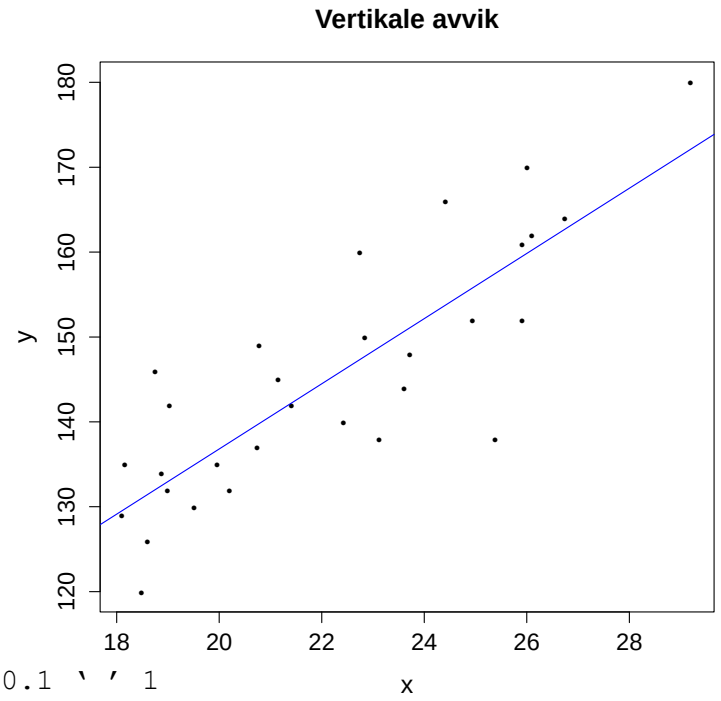
```
> summary(regr.Bears)

Call:

lm(formula = logweight ~ loglength)

Residuals:

     Min       1Q   Median       3Q      Max

-0.57529 -0.15137  0.00251  0.11904  0.58919

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) -13.3663     0.5623  -23.77   <2e-16 ***

loglength     3.5084     0.1116   31.44   <2e-16 ***

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2144 on 141 degrees of freedom

Multiple R-squared:  0.8751,     Adjusted R-squared:  0.8742

F-statistic: 988.2 on 1 and 141 DF,  p-value: < 2.2e-16
```



(multippel) R-square=0.8751 = $(0.9355)^2$ = r^2 der r = korrelasjonen. Dvs 87.5% av variasjonen i *log(vekt)* som kan forklares av *log(lengde)*

Eksempel r^2: BMI og blodtrykk

```
Call:
lm(formula = SBT ~ BMI, data = bt.data[-c(2, 9), ])

Residuals:
     Min       1Q   Median       3Q      Max
-19.5386  -5.3812  -0.4422   4.8117  13.9164

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  60.0188    10.6291   5.647 4.76e-06 ***
BMI           3.8394     0.4741   8.099 8.10e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.94 on 28 degrees of freedom
Multiple R-squared:  0.7008,    Adjusted R-squared:  0.6902
F-statistic: 65.59 on 1 and 28 DF,  p-value: 8.099e-09
```



(multippel) R-square=0.7008. Dvs 70.1% av variasjonen i *blodtrykk* som kan forklares av BMI

2.5 Forsiktighetsregler for korrelasjon og regresjon

- Residualer og residualplott
- Uteliggere og innflytelsesrike observasjoner
- Underliggende («lurkende») variable
- Korrelasjon og kausalitet

De vertikale avstandene mellom punktene og minste-kvadrater-regresjonslinja kalles residualer

En regresjonslinje beskriver det overordnede mønsteret til en lineær sammenheng mellom en forklarings-variabel og en respons-variabel. Avvik fra det overordnede mønsteret er også viktige.

Et **residual** er differansen mellom den observerte verdien av respons-variabelen og verdien som er predikert av regresjonslinja:

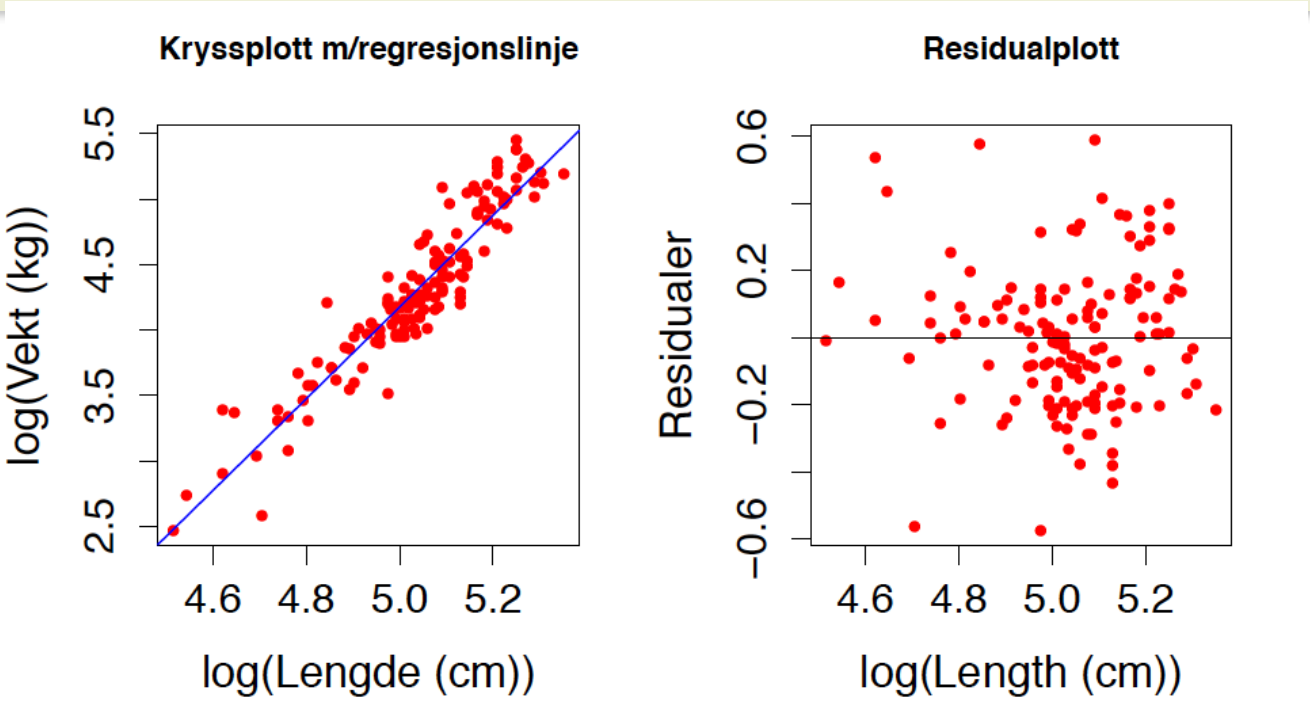
residual = observert *y* – predikert *y*

$$= y - \hat{y}$$

Et residualplott er et spredningsplott av residualene mot forklaringsvariabelen

Residualplott hjelper oss å vurdere hvor godt regresjonslinja passer til og beskriver dataene.

- Ideelt sett skal det være en "tilfeldig" spredning rundt *x*-aksen.
- *Mønstre* i residualene er tegn på at en lineær sammenheng ikke fanger mønsteret i dataene, og at lineær regresjon ikke er passende.
  Eksempel: **Lengde og vekt for 143 bjørner**



En uteligger er en observasjon som ligger utenfor det overordnede mønsteret til (de andre) observasjonene

Både store y-er i forhold til linja og ekstreme x-er er uteliggere.

Innflytelsesrike punkter: å fjerne slike punkter vil endre regresjonsligninga til linja markant.

- Uteliggere i ***y*-retning** har store residualer. ***De er innflytelsesrike hvis*** det er få punkter med lignende *x*-verdi som «holder linja på plass»

- Uteligger i ***x*-retning** trenger ikke å ha stort residual, men er ***ofte innflytelsesrik*** for minste-kvadrater-regresjonslinja, med mindre den ligger nær regresjonslinja beregnet uten denne

Uteliggere og innflytelsesrike punkter 1

Eksempel (2.28-22.9): Fastende plasma-glukose (FPG) og langtids-blodsukker (HbA1c)

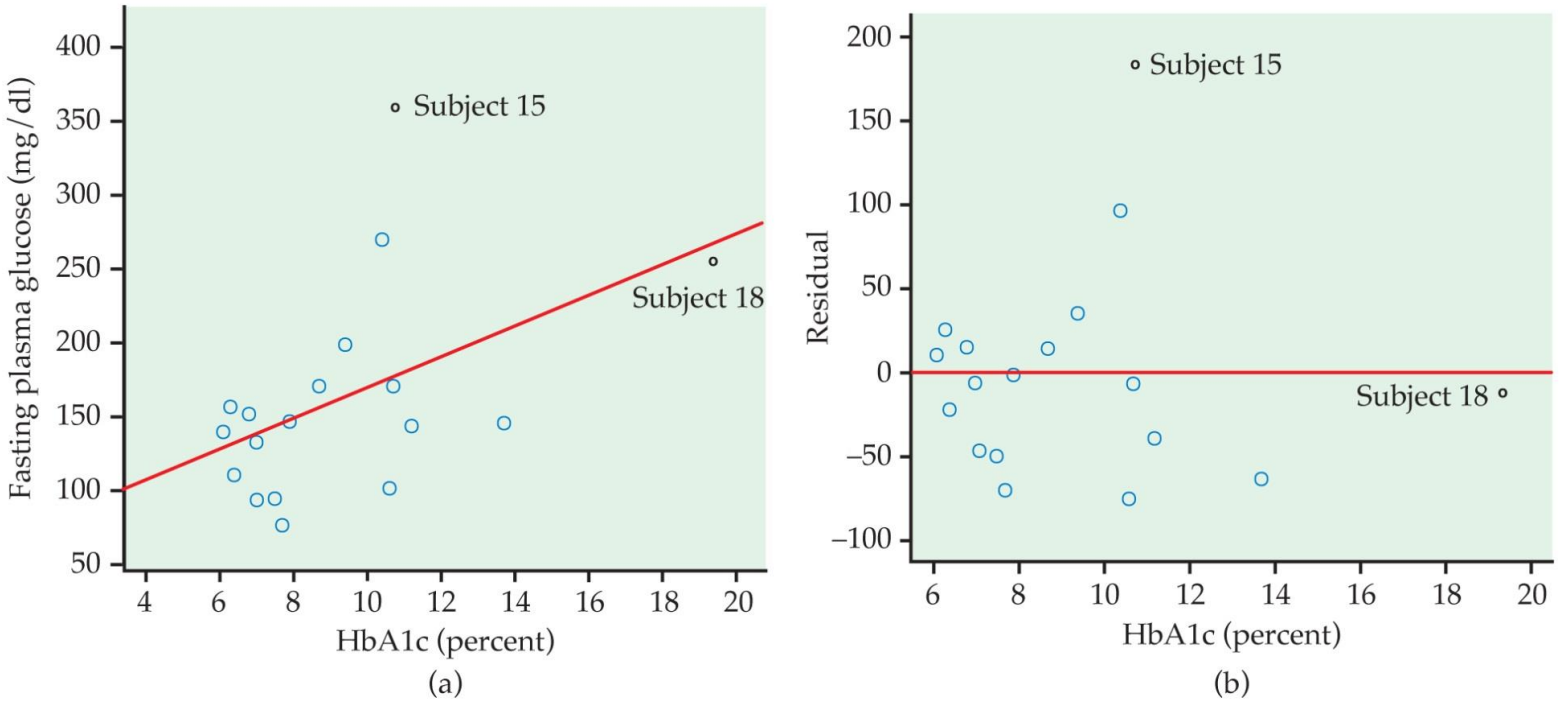


Figure 2.25
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

Uteliggere og innflytelsesrike punkter 2

Fastende plasma-glukose (FPG) og langtids-blodsukker (HbA1c):
Uteliggere i *x*- og *y*-retning: Ikke veldig innflytelsesrike

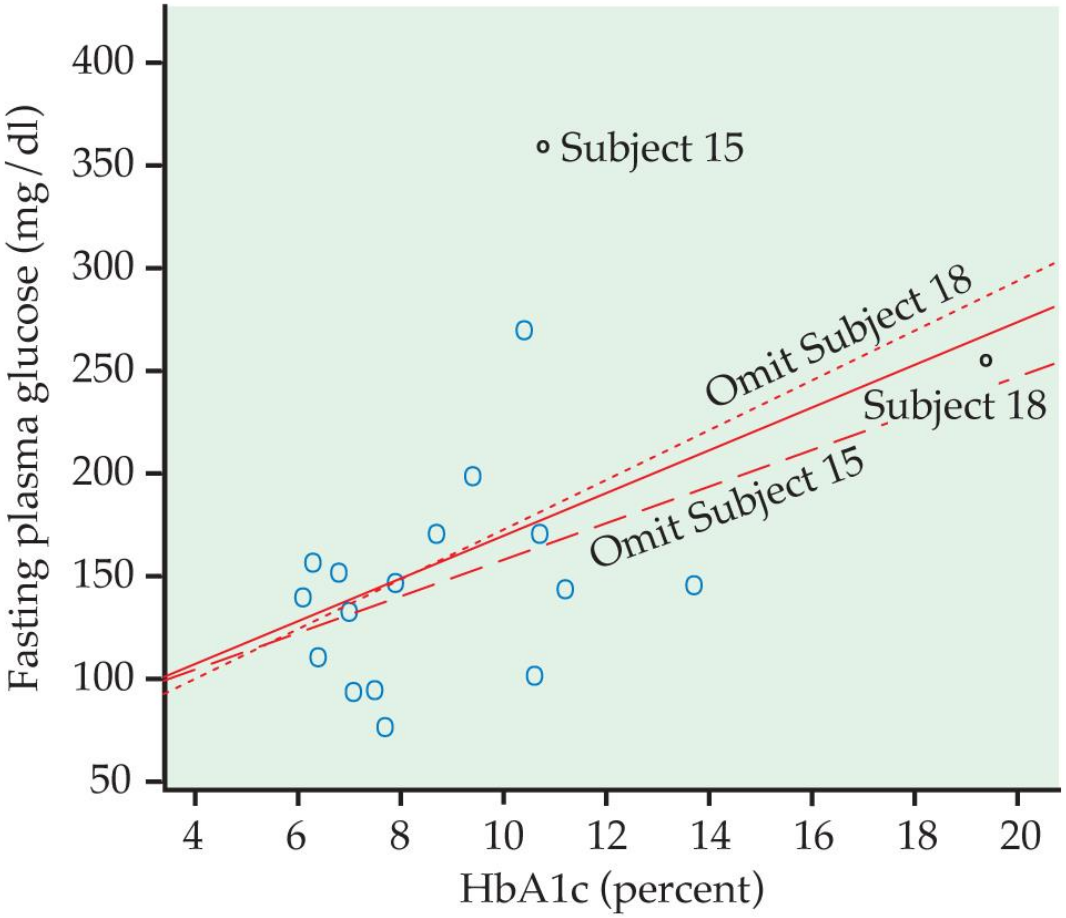


Figure 2.26
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

En **underliggende variabel («lurkende»)** er en variabel som ikke er verken forklarings- eller respons-variabel men likevel kan påvirke tolkningen av sammenhengen mellom forklarings- og respons-variablene.

Eksempel:

- Flere studier viser at menn med hjerteproblemer oftere får kirurgiske behandlinger slik som bypass-kirurgi enn kvinner.
- Skyldes denne sammenhengen som sees mellom kjønn og behandling forskjellsbehandling mellom kjønnene?
- Ikke nødvendigvis. Hjerteproblemer inntreffer vanligvis senere hos kvinner enn menn. Behandlinger slik som bypass-kirurgi kan være forbundet med stor risiko for eldre pasienter, slik at leger kan være forsiktige med å anbefale det for eldre pasienter.
- Den underliggende variabelen *alder* kan forklare sammenhengen mellom kjønn og behandling.

Forsiktighetsregler for korrelasjon og regresjon

- Begge beskriver lineære sammenhenger.
- Begge påvirkes av uteliggere.
- Plott alltid dataene før du tolker.
- Vær oppmerksom på ***ekstrapolering.***
  - Vær forsiktig med å predikere *y* når *x* er utenfor verdiområdet til de observerte *x*-ene.
- Vær oppmerksom på ***underliggende variable.***
  - Disse har en viktig effekt på sammenhengen mellom variablene i en studie men er ikke inkludert i studien.
- **Korrelasjon impliserer ikke kausalitet!**

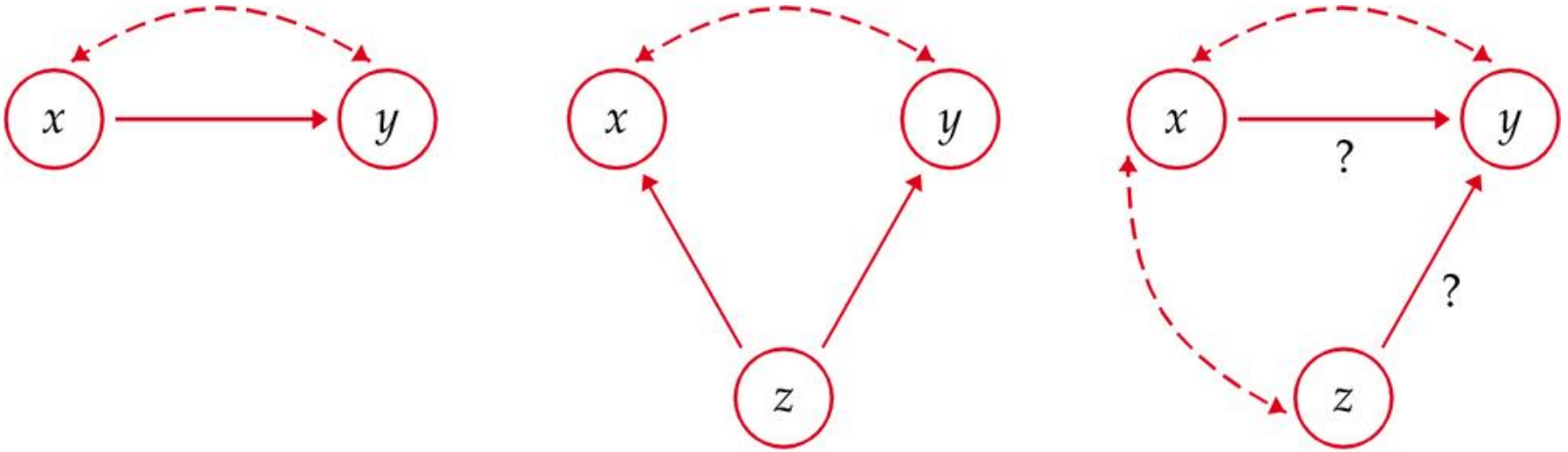
2.7 Spørsmålet om kausalitet

- Forklare sammenheng
- Kausalitet
- Felles respons
- Konfundering
- Fastslå kausalitet

Kausalitet = Årsakssammenheng, og Uansett hvor sterk en sammenheng er, impliserer IKKE styrken kausalitet.

Noen mulige forklaringer for en observert sammenheng

De stiplede linjene viser en observert sammenheng mellom to variable. De heltrukne pilene viser årsaks- og effekt-linker. *x* er forklaringsvariabel, *y* er responsvariabel, og *z* er en underliggende variabel.



“Vær oppmerksom på underliggende variable”

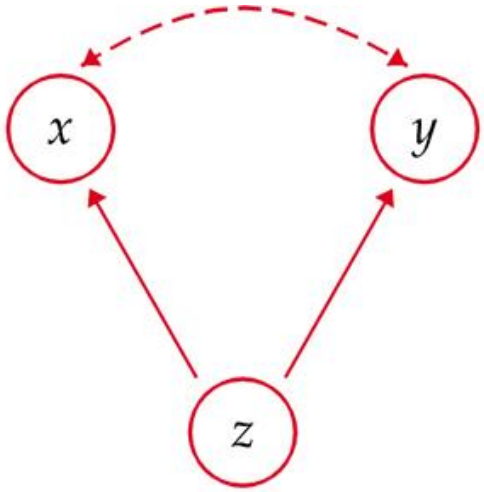
-et godt råd når man tenker på en sammenheng mellom to variable.
Når både *x* og *y* kan forandres som respons på endringer i den underliggende variabelen *z,* kan en observert sammenheng mellom variablene forklares av den underliggende variabelen.

De fleste studenter som har gode karakterer på videregående (*x*) har også gode karakterer (*y*) det første året de tar høyere utdanning.

- Den observerte sammenhengen mellom *x* og *y* kan forklares av en tredje underliggende variabel *z.* I dette eksempelet er “evner og kunnskap” den underliggende variabelen.

- Positiv korrelasjon kan skyldes *felles respons* på studenters evner og kunnskap.

- Både *x* og *y* forandres som respons på endringer i *z*. Dette lager en statistisk sammenheng selv det ikke er noen direkte årsakssammenheng.

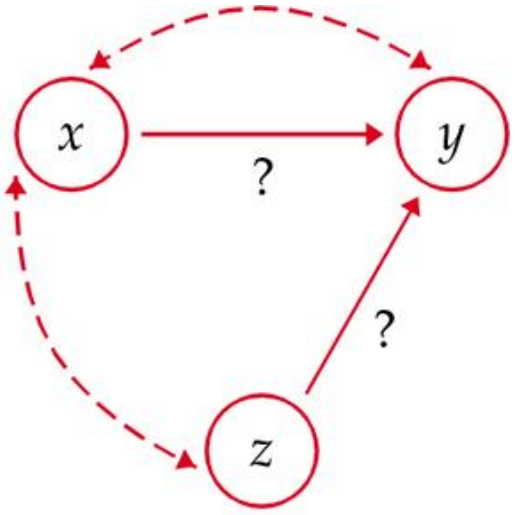


Common response

To variable er konfundert når man ikke kan skille mellom effektene de har på en respons-variabel.

De konfunderte variablene kan enten være forklarings-variable eller underliggende variable eller begge deler.

- Eksempel: Studier viser at religiøse personer lever lengre enn ikke-religiøse personer.
  x: Religiøsitet (ja/nei), *y*: levealder
- Studier viser at å være religiøs har sammenheng med en sunnere livsstil (*z)* (f.eks. mindre overvekt og røyking)



Confounding

Fastslå kausalitet

Det ser ut som om lungekreft har en sammenheng med røyking.

Hvordan vet vi at ikke begge disse variablene påvirkes av en uobservert tredje (underliggende) variabel?

For eksempel: Hva hvis det er en genetisk predisposisjon som forårsaker at visse personer både får lungekreft *og* blir avhengige av å røyke, men at røyking i seg selv ikke FORÅRSAKER lungekreft?

Vi kan evaluere om sammenhengen er kausal ved å bruke følgende kriterier:

1. Sammenhengen er sterk.
2. Sammenhengen er konsistent:
    - Sammenhengen ses i *repeterte studier.*
    - Sammenhengen skjer under *varierende betingelser.*
3. Høyere doser henger sammen med sterkere respons.
4. Påstått årsak kommer før effekten.
5. Påstått sammenheng er plausibel (en rimelig forklaring).

www.menti.com

KODE: