Sammenhenger

I dette kapitlet er emnet måter å studere sammenheng mellom to variable.

Som tidligere er grunnleggende spørsmål:

Hva er enhetene som skal studeres?

Hva er de interessante variablene? Hvordan er de målt, konstruert og registrert?

Er variablene kvantitative eller kategoriske?

To variable er assosierte hvis kunnskap om verdien på den ene bidrar til mer kunnskap om den andre enn hvis man ikke hadde denne informasjonen.

Et fundamentalt skille er mellom respons og forklaringsvariable.

Responsvariable: En slik variabel måler resultatet av en studie.

Forklaringsvariable: Slike variable forklarer effekten på resposvariablen.

De dataene vi vil se på har formen (x_1, y_1),...,$(x_n\ y_n)$,
Der y betegner verdien til responsvariabelen og x verdien til forklaringsvariabelen.

Eksempel y kalsiumopptak - x tilført kalsium
y tid for kjemisk reaksjon - x temperatur
y vekt - x høyde
y prisstigning - x arbeidsledighet

Skillet mellom respons og forklaringsvariable er viktig og har mange betegnelser. Responsvariablen kalles også avhengige variabel eller endogene variabel. Forklaringsvariable kalles uavhengige variable, kovariater og eksogene variable.

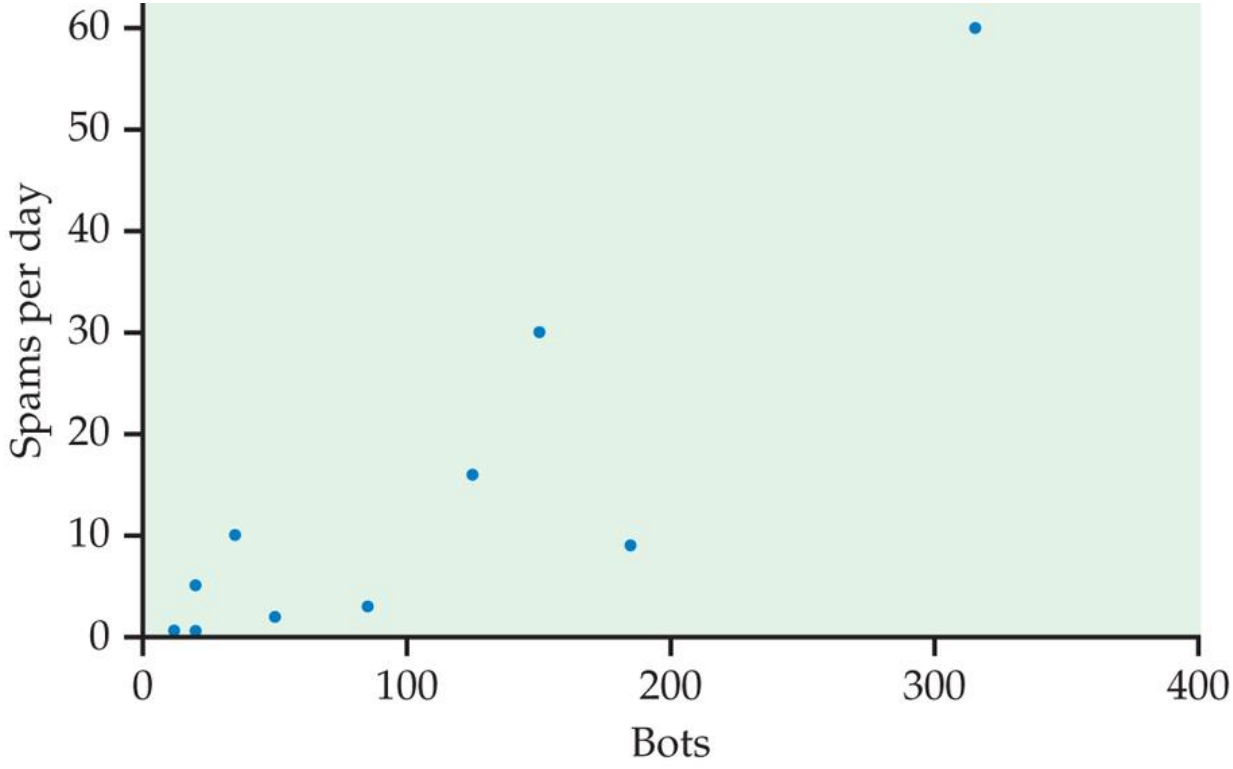
Strategien vi følger for å analysere sammenhenger er den samme som ble fulgt for å undersøke fordelingen til en variabel.

Start med en grafisk framstilling av data. Dernest finn det almmene mønsteret og eventuelle avvik fra dette. På grunnlag av det man finner kan resultatet sammenfattes i numeriske oppsummeringer.

Spredningsplott

Eksempel Data fra n=10 store botnets,
Respons y: Antall spam pr dag
Forklaringsvariabel x: Antall bots i hvert botnet.

Spredningsplot: Plot y_i mot x_i for hver enhet i=1,..,n



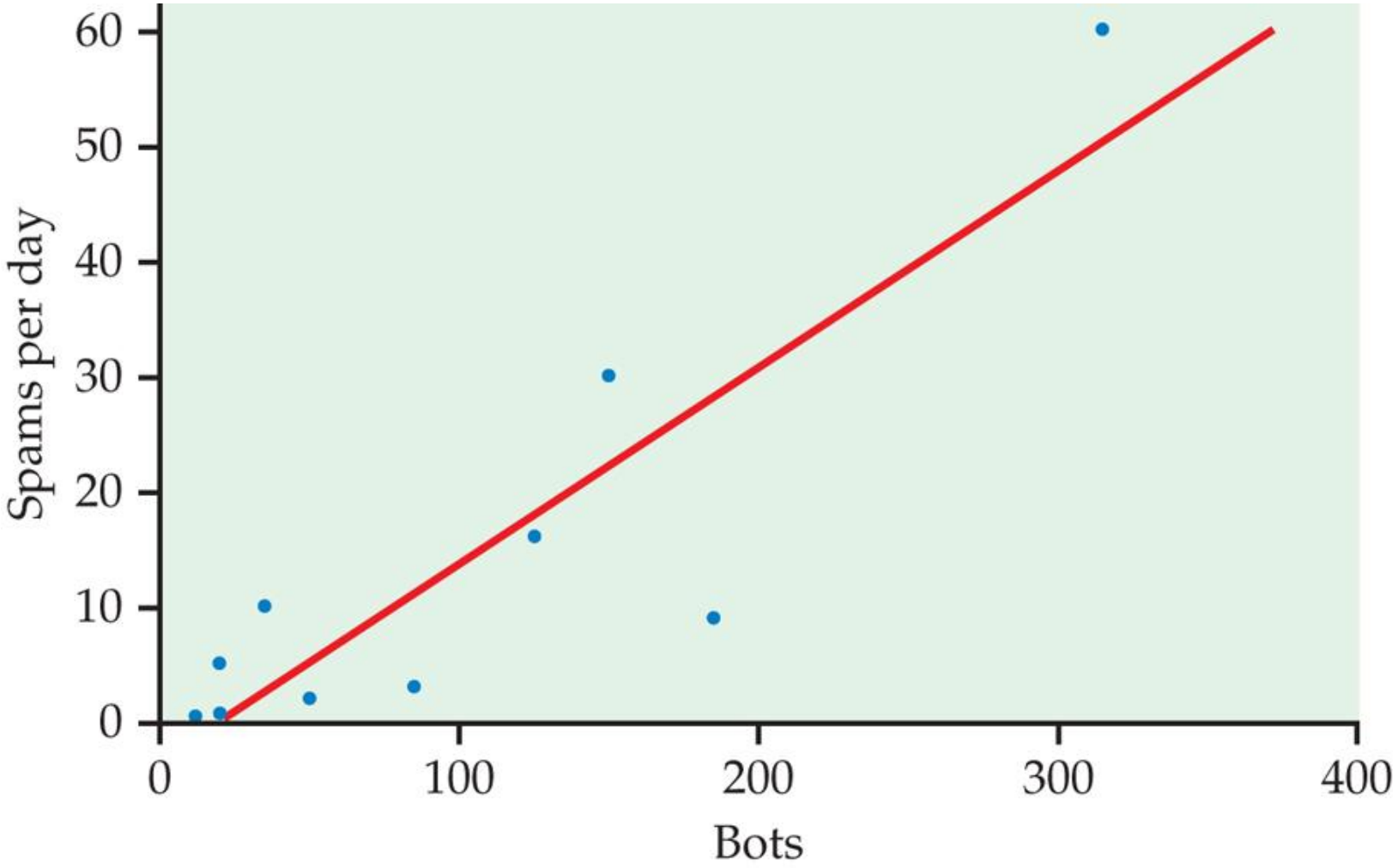
Forklaringsvariablen skal alltid plottes langs den horisontale aksen og responsvariabelen langs den vertikale aksen.

- Se etter almenne mønstre og eventuelle avvik i spredningsplottet.

- Beskriv form, retning og styrke på sammenhengen.

- Identifiser outliere som klart skiller seg ut.

Å supplere spredningsplottet med en linje kan være til hjelp ved vurderingen av spredningsplottet.


Spams per day
60
50
40
30
20
10
0
0
100
200
300
400
Bots

Sammenhengen kan være:

Positiv sammenheng eller assosiasjon: Store verdier av x forekommer stort sett sammen med store verdier av y, og små verdier av x forekommer stort sett sammen med små verdier av y.

Negativ sammenheng eller assosiasjon: Store verdier av x forekommer stort sett sammen med små verdier av y, og små verdier av x forekommer stort sett sammen med store verdier av y.

Eksempel Gjeld i 2007 versus gjeld i 2006 for 24 OECD land

Microsoft Excel - GovDebt200...

File Edit View Insert Format Tools
Data Window Help

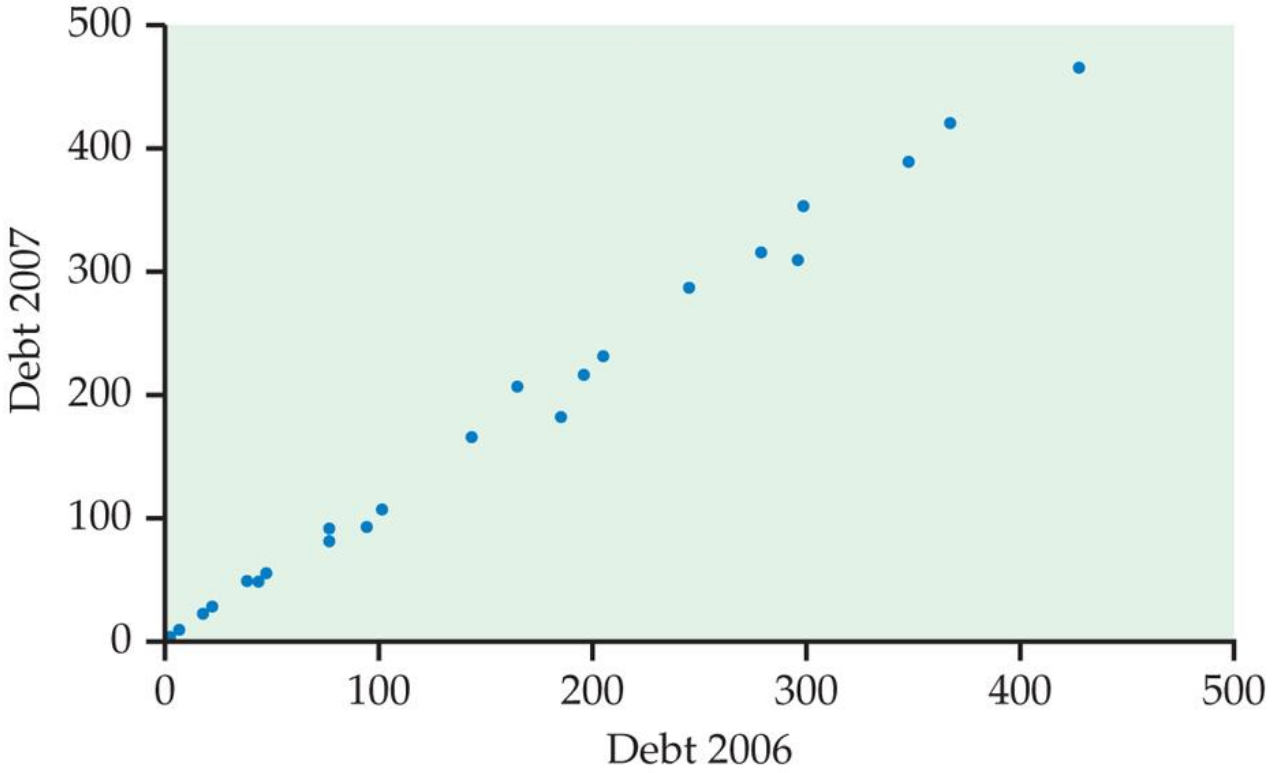
E25

	A	B	C
1	Country	Debt2006	Debt2007
2	Luxembourg	0.65	0.78
3	Iceland	4.04	4.92
4	Slovak Republic	18.44	22.79
5	New Zealand	21.44	27.82
6	Norway	43.03	49.05
7	Czech Republic	38.44	49.36
8	Australia	43.91	49.45
9	Ireland	47.30	55.29
10	Finland	77.58	82.54
11	Hungary	76.76	90.24
12	Denmark	94.29	92.74
13	Switzerland	101.28	107.49
14	Portugal	142.97	166.06
15	Sweden	185.01	182.11
16	Poland	164.41	205.97
17	Mexico	195.66	216.76
18	Austria	204.51	231.56
19	Turkey	244.89	286.96
20	Korea	295.07	309.34
21	Netherlands	278.67	316.07
22	Greece	297.93	352.80
23	Canada	347.68	389.83
24	Belgium	366.91	420.74
25	Spain	427.06	464.99
26			

Sheet1

Draw AutoShapes

NUM



Positiv sammenheng mellom gjeld i 2006 og 2007.

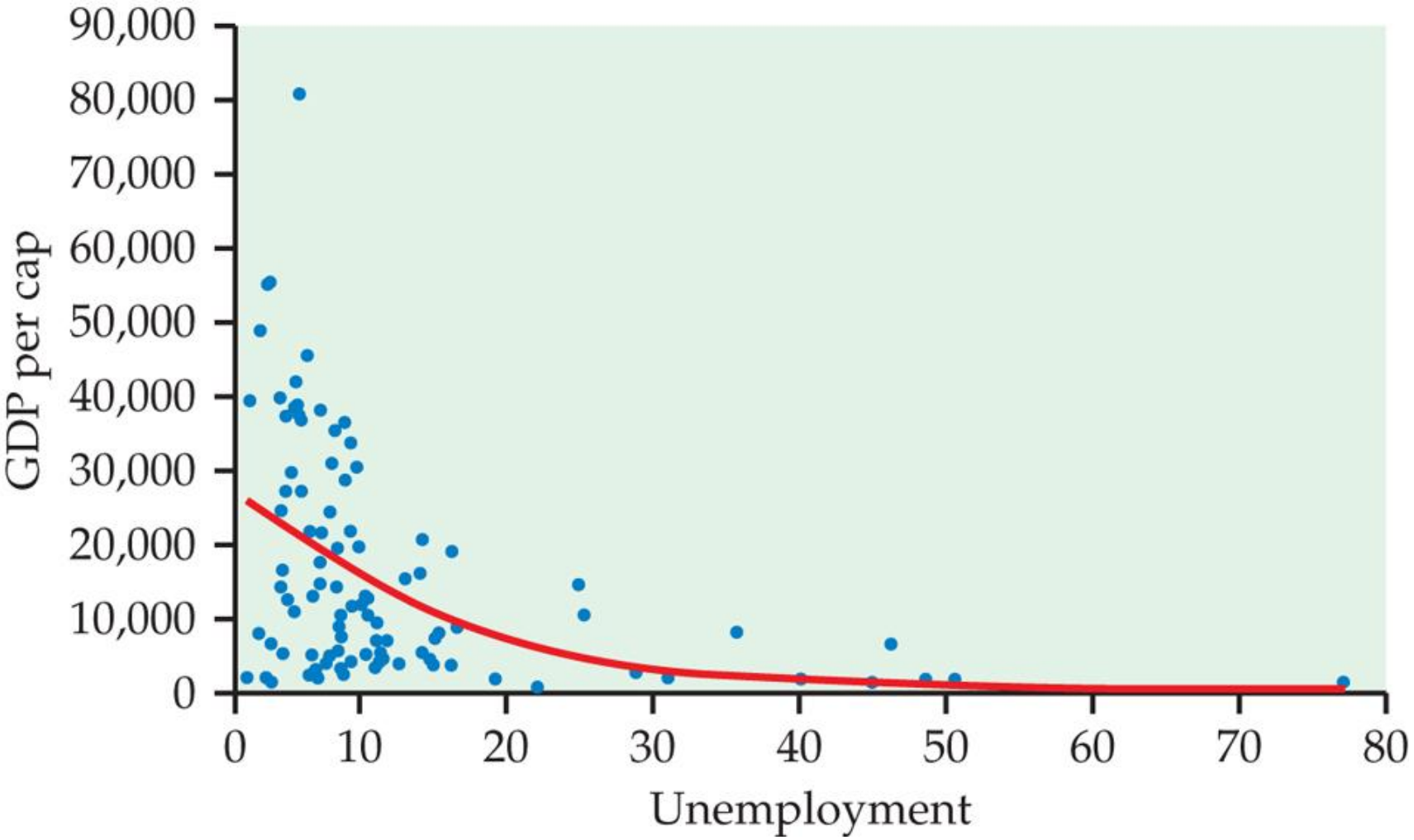
Andre mulige plot:

Nedbetalt gjeld, gjeld 2007 – gjeld 2006 mot gjeld 2006.

Tilbakebetalinsrate $\frac{\text{gjeld 2007} - \text{gjeld 2006}}{\text{gjeld 2006}}$ mot gjeld 2006.

Spredningsplott kan vise sammenhenger som er klart ikke-lineære.

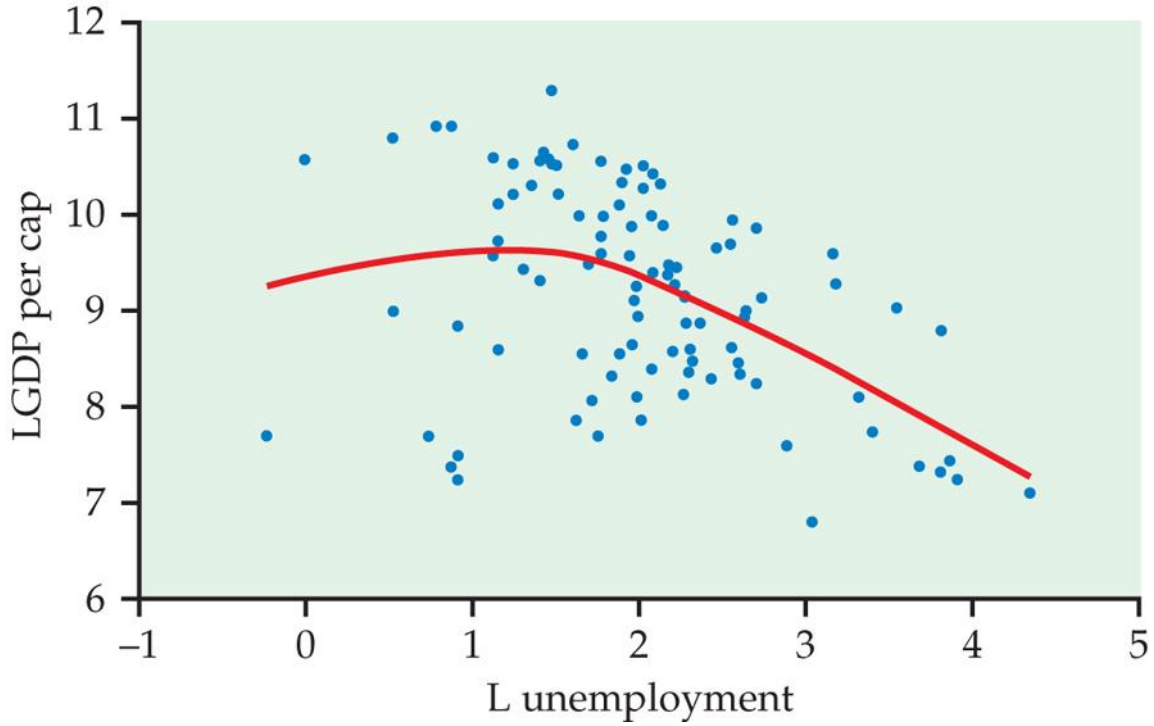
Eksempel Brutto nasjonalprodukt pr innbygger mot arbeidsløshetsrate.



Sammenhengen er først lineær så ikke-lineær.

Ved ikke-lineære sammenhenger kan transformasjoner være en ide.

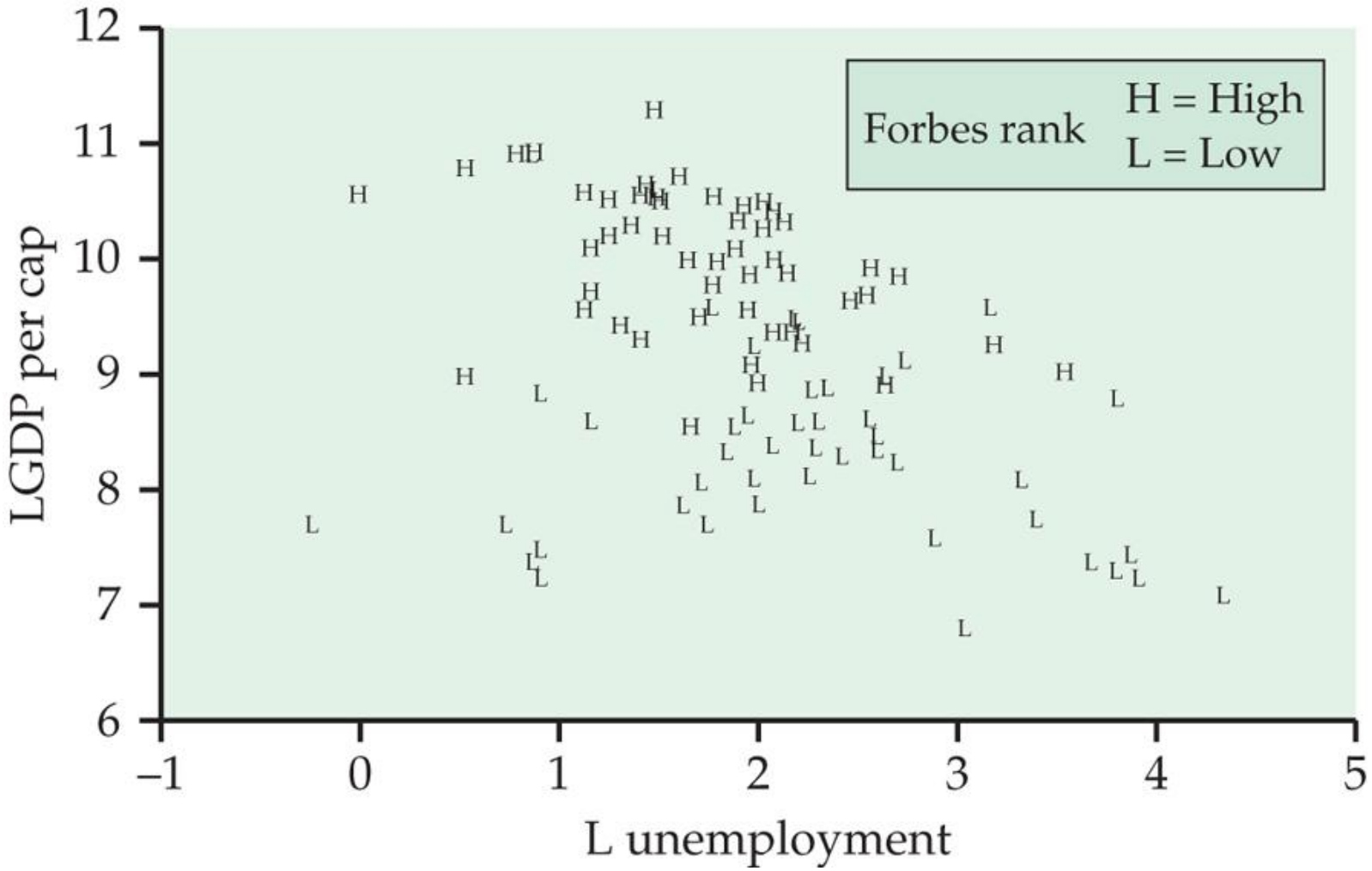
Den vanligste transformasjonen er log transformen som består i å ta (den naturlige) logaritmen til variablene.



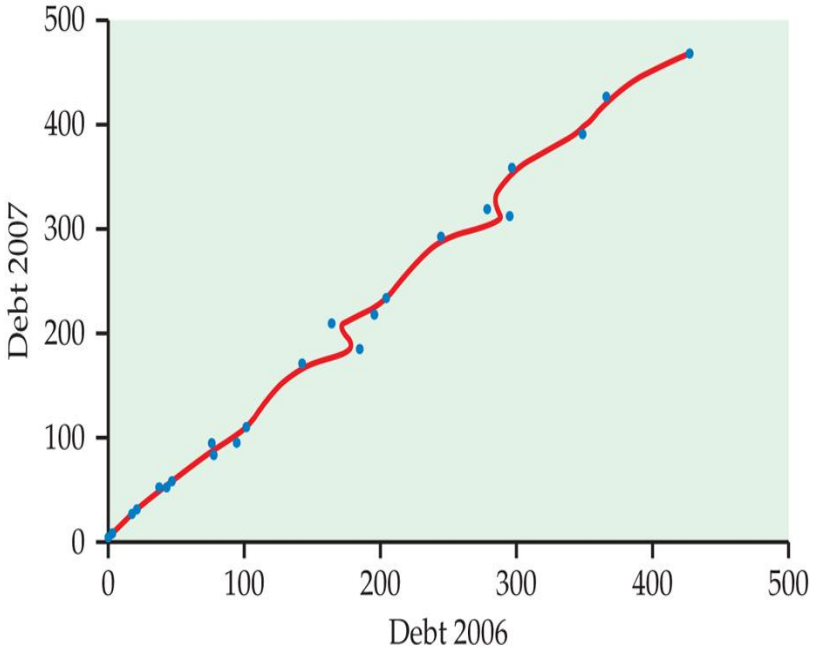
Sammenhengen er først essensielt flat til 1.6, som svarer til 5% arbeidsløshet (log 5 = 1.609438) deretter lineært avtagende.

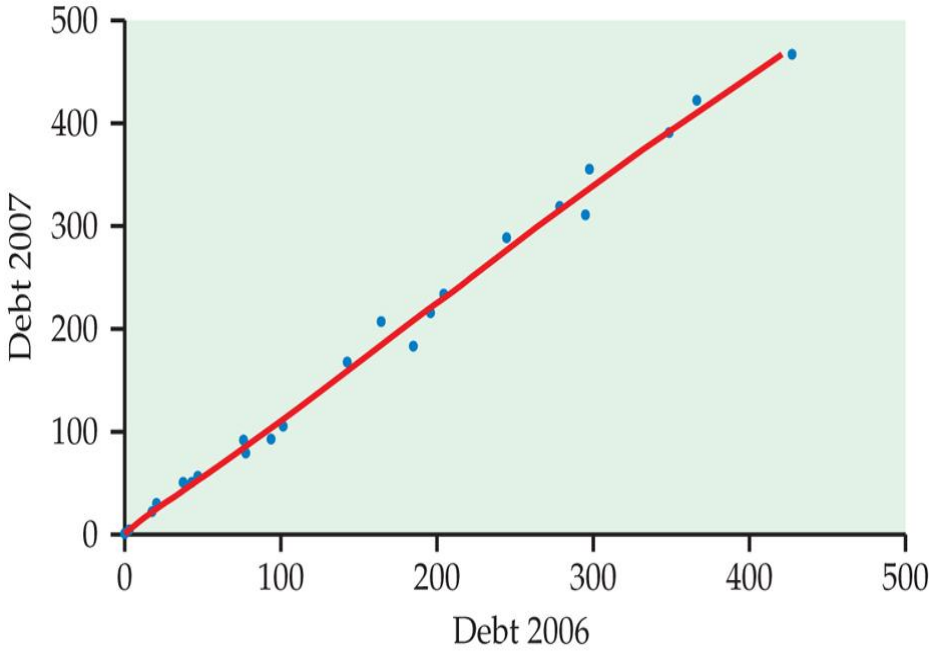
Kategoriske variable kan inkluderes i spredningsplott ved å bruke ulike plottesymboler for klassene enhetene tilnører.

Eksempel For dataene om bruttonasjonalprodukt kan landene inndeles etter hvodan Forbes.com vurderer dem som «Best contries for Business».


Forbes rank
H = High
L = Low
LGDP per cap
L unemployment
12
11
10
9
8
7
6
–1
0
1
2
3
4
5

Mange statistikkpakker har rutiner for å tilpasse glatte kurver til datapunktene. Graden av tilpasning bestemmes av en størrelse som kalle båndbredde.





Kategoriske forlaringsvariable kan tas hensyn til ved å sammenligne fordelingene i hver av klassene enhetene deles opp i. Dette kan gjøres ved rygg mot rygg («Back to back» stilk-og-blad plott eller ved å legge boksplott for verdiene i hver klasse ved siden av hverandre.

Korrelasjon

Plott og andre grafiske fremstillinger er nyttige verktøy, men visuelle inntrykk kan være misledende. Derfor er det nødvendig også å se på numeriske oppsummeringer av sammenhengene som supplerer plottene.

Anta at vi observerer n par av verdier, $(x_1, y_1),\ldots,(x_n, y_n)$ for enhetene, f.eks. høyde og vekt for n voksende kvinner.

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

der s_x er standardavvik for verdiene av forklaringsvariablene, x_i, og s_y er standardavvik for verdiene av responsene, y_i.

$\frac{x_i - \bar{x}}{s_x}$ er standardiserte x-verdier.

$\frac{y_i - \bar{y}}{s_y}$ er standardiserte y-verdier.

Korrelasjonen r er summen av produktet av disse verdiene delt på (n-1).

Hvis det er en positiv sammenheng mellom x og y-verdiene, vil også de standariserte verdiene ha en slik sammenheng. Ved en positiv sammenheng blir derfor produktene stort sett positive og summen blir derfor positiv. Altså en positiv verdi for r, som indikerer en positiv sammenheng.

Merk:

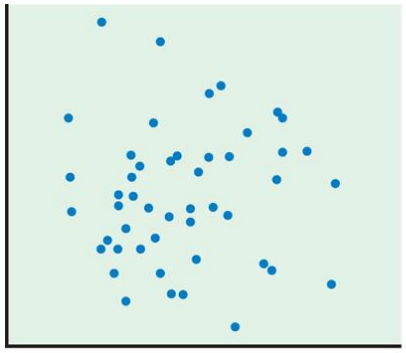
Positive verdier indikerer positiv sammenheng, og negativ r indikerer negativ sammenheng.

Korrelasjonen r er symmetrisk, så det skilles ikke mellom respons og forklaringsvariable.

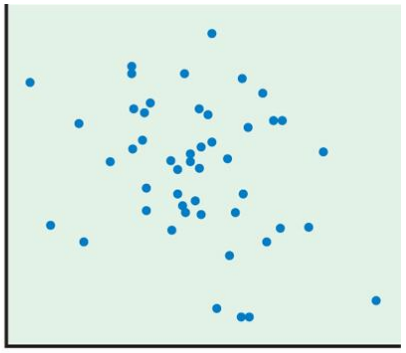
Korrelasjon r er definert for kvantitative variable.

Korrelasjon r er uavhengig av skala. Lengder i kilometer gir samme resultat som lengder i meter.

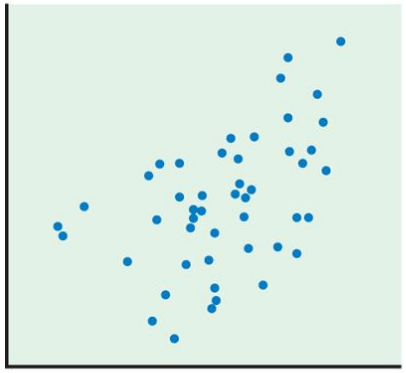
Korrelasjonsen r ligger mellom -1 og 1, ±1 svarer til rette linjer.



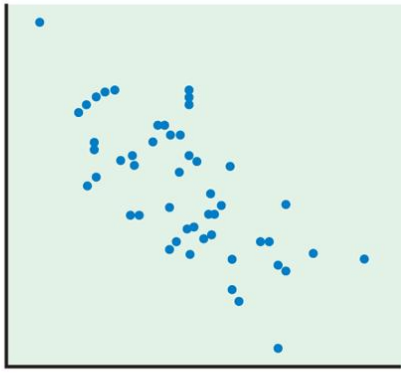
Correlation $r = 0$



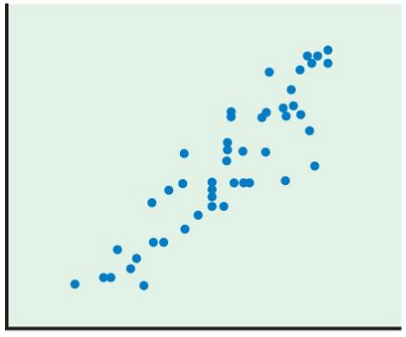
Correlation $r = -0.3$



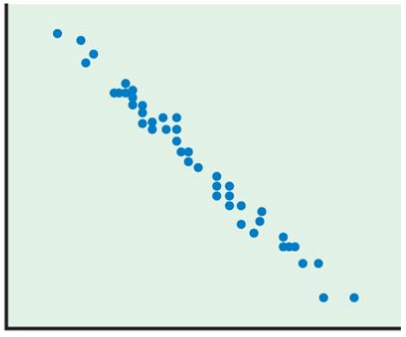
Correlation $r = 0.5$



Correlation $r = -0.7$



Correlation $r = 0.9$



Correlation $r = -0.99$

Merk også:

Korrelasjon måler kun lineær sammenheng,
for ikke-lineære (kurvede) sammenhenger er den ikke et godt mål.

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

$$x_i = -1.0, -0.9, \cdots, 0.9, 1.0$$

Da ligger dataene langs en parabel, men korrelasjonen *r* blir lik 0.

Dessuten: Korrelasjonen er ikke robust mot outliere (på samme måte som gj.sn. og std.avvik.)

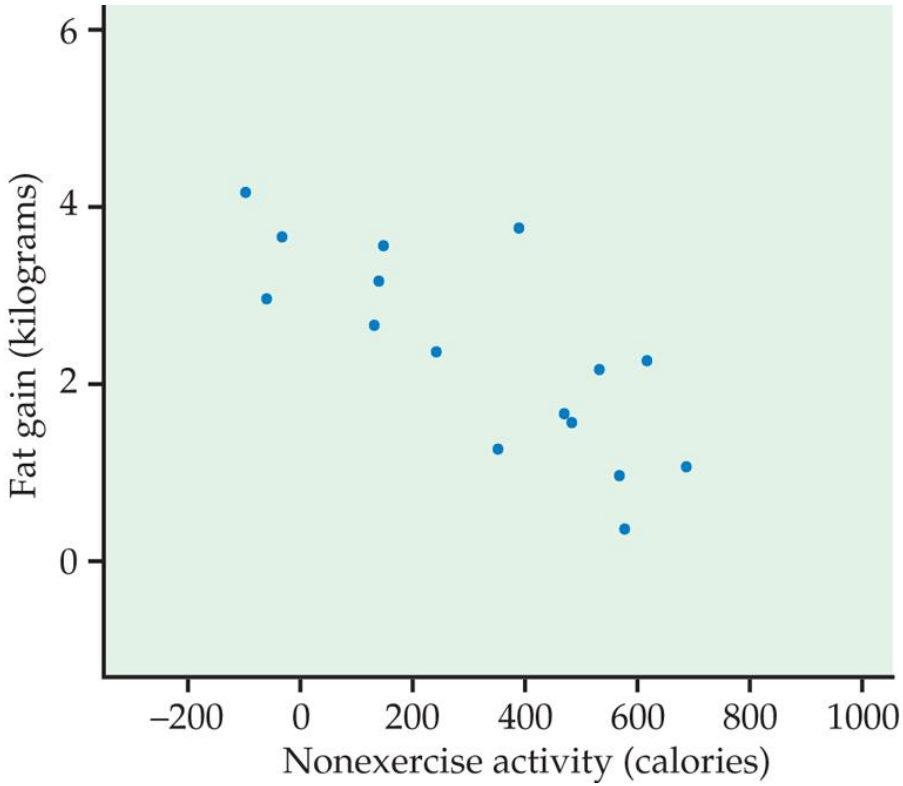
Minste kvadraters regresjon

En regresjonslinje oppsummerer sammenhengen mellom de to variablene i et spredningsplott.

Regresjonslinje: En rett linje som beskriver hvordan responsvariabelen y endres når forklaringsvariabelen x endres.

Regresjonslinjer brukes ofte til å predikere verdien på responsvariabelen.

Eksempel Spredningsplottet nedenfor viser sammenheng mellom *endring* i vekt i kg og *endring* i aktivitet som ikke er knyttet til trening, NEA, målt i kalorier, i en studie av 16 unge voksne. Rastløse personer har typisk høyt NEA.



Vi observerer en negativ sammenheng mellom vektøkning og NEA.
Korrelasjonen mellom vektøkning og NEA er lik -0.7786
Sammenhengen ser ut til å være nokså lineær.

Lineær eller rettlinjet sammenheng:

$$y = b_0 + b_1 x$$

Her kalles koeffisenten b_1 for stigningsforholdet
og koeffisenten b_0 er linjas skjæringspunkt (med y-aksen).

Vi referer ofte til b_0 som konstantleddet.

Å tilpasse en rett linje til et datasett innebærer å finne den linja som har så liten avstand til punktene i spredningsplottet som mulig – i henhold til kriteriet ”minste kvadrater løsning” som forklares senere.

I eksemplet blir denne linja

$$\text{vektøkning} = 3.505 - 0.00344\,\text{NEA}$$

dvs. $b_0 = 3.505$ og $b_1 = -0.00344$

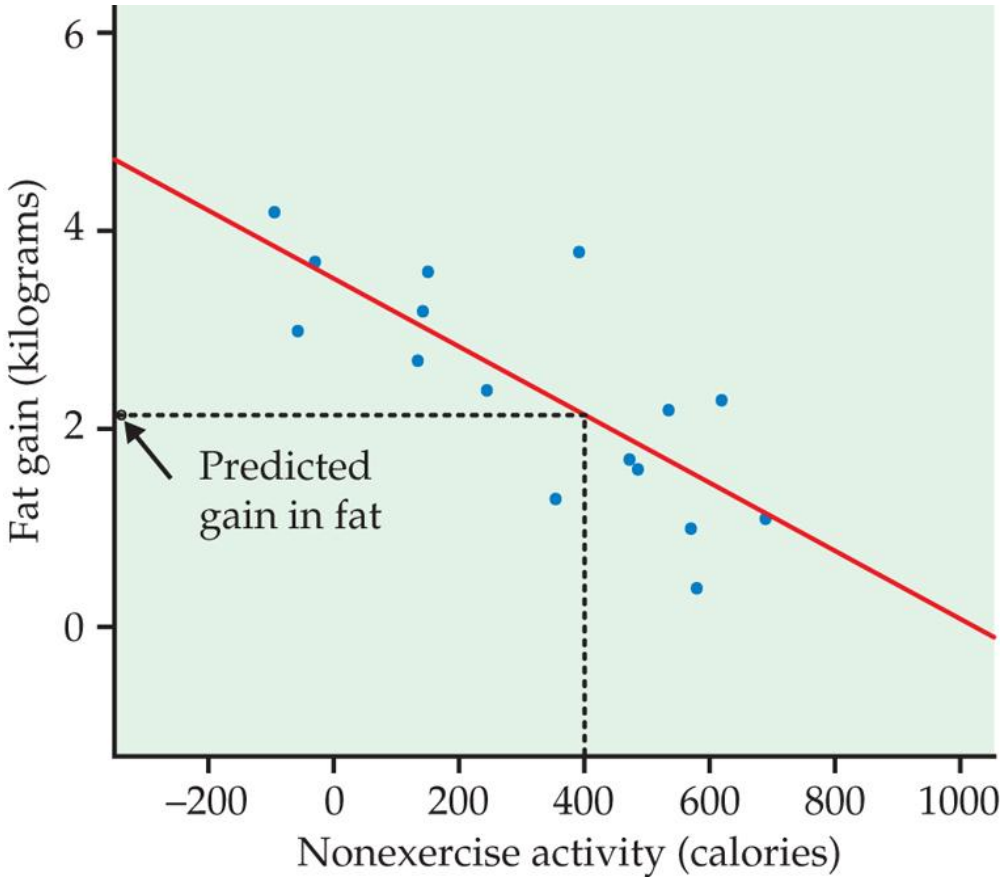
Spesielt er stigningsforholdet b_1 en viktig størrelse.
Den måler endringen i linja $y = b_0 + b_1 x$
når x endres med en enhet.

Skjæringspunktet b_0 har fortolkning som estimert verdi av y når x=0.

Regresjon kan brukes til *prediksjon* av responsen for en gitt verdi av forklaringsvariablen.

Eksempel For en person som har en økning i NEA på 400 kalorier vil den predikerte vektøkningen være

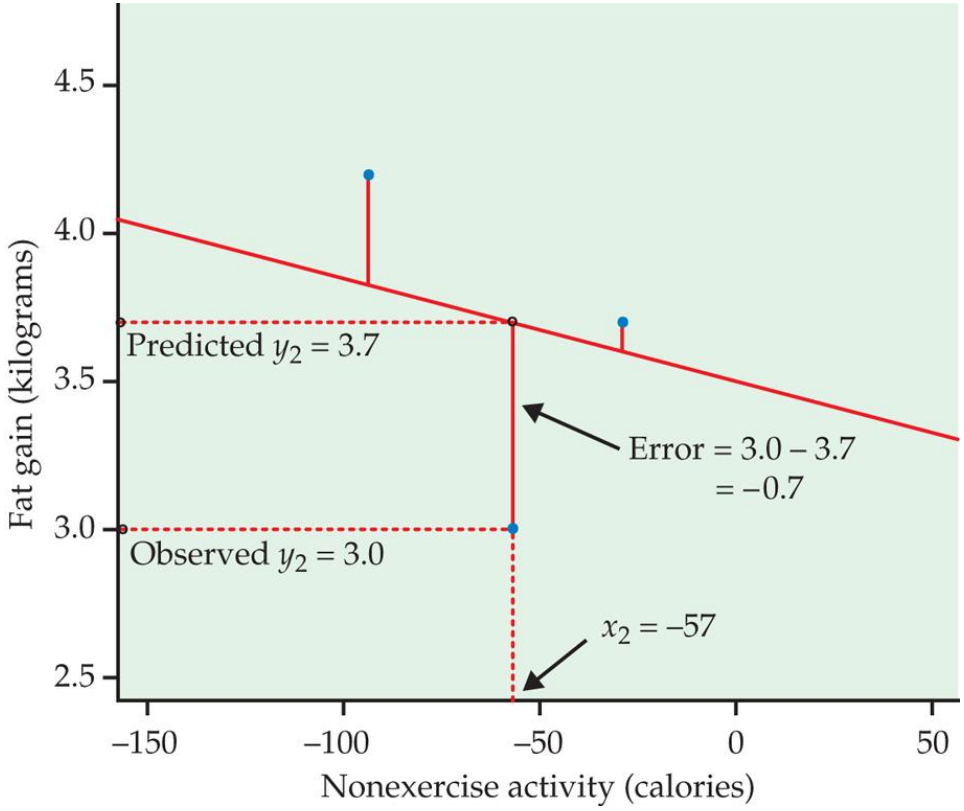
$$3.505 - 0.00344 \times 400 = 2.13 \text{ kilo}$$



Prediksjon utenfor det området der man har verdier for forklaringsvariabelen kalles ekstrapolasjon. Det skal man være forsiktig med og ikke gjøre uten grundig vurdering. Grunnen er at innenfor det områdetet som dekkes av spredningsplottet kan man se om sammenhengen er noenlunde lineær og dermed fanges opp av regresjonslinja. Utenfor dette området har man ikke noe data, og vet derfor heller ikke noe om sammenhengen mellom variblene. Den kan være lineær med andre koeffesienter eller ikke-lineær.

Hittil har vi ikke sagt noe om hvordan regresjonslinja blir lagd. Det er en framgangsmåte som brukes mye.

Kriterium: Minste kvadraters regresjon av y på x er den linja som minimerer summen av kvadratetene av den vertikale avstanden fra datapunktene $(x_1, y_1),\ldots,(x_n, y_n)$ til linja.



Anta at $y = b_0 + b_1 x$ er regresjonslinja. Med denne får vi prediksjoner $\hat{y}_i = b_0 + b_1 x_i$ når forklaringsvariablen er lik x_i.

Da er $(x_i, \hat{y}_i)$ det punktet på regresjonslinja som ligger rett over eller under (x_i, y_i).

Avstanden mellom de to punkten er

$$| y_i - \hat{y}_i | = | y_i - b_0 - b_1 x_i |$$

og den kvadrerte avstanden blir

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

slik at summen av de kvadrerte avstanden blir

$$\sum (y_i - b_0 - b_1 x_i)^2$$

Vi kan tenke oss ulike kandidater for regresjonslinja $y = b_0 + b_1 x$, dvs. ulike verdier av (b_0, b_1). Disse vil gi ulike kvadratsummer

$$S(b_0, b_1) = \sum (y_i - b_0 - b_1 x_i)^2$$

Minste kvadraters metode består i å velge de verdiene av (b_0, b_1) som gjør $S(b_0, b_1)$ minst mulig.

Løsningen er faktisk relativt enkel og gis ved formlene

$$b_1 = r \frac{s_y}{s_x} \quad \text{og} \quad b_0 = \bar{y} - b_1 \bar{x}$$

Her er som før

$\bar{x}$ gjennomsnittet av $x_1, \ldots, x_n$

$\bar{y}$ gjennomsnittet av $y_1, \ldots, y_n$

s_x standardavviket for $x_1, \ldots, x_n$

s_y standardavviket for $y_1, \ldots, y_n$

r korrelasjonen for $(x_1, y_1), \ldots, (x_n, y_n)$

Eksempel Vekt og aktivitet.

Her er $\bar{x}$ = 324.8 kalorier og s_x = 257.66 kalorier og
$\bar{y}$= 2.388 kg og s_y = 1.1389 kg. Korrelasjonen er r = -0.7786.

Det gir stigningsforholdet

$$b_1 = r\frac{s_y}{s_x} = -0.7786\,\frac{1.1389}{257.66} = -0.00344$$ kr pr kalori

og skjæringspunkt

$$b_0 = \bar{y} - b_1\bar{x} = 2.388 - (-0.00344)(324.8) = 3.505$$ kg.

Regresjonslinja er $\hat{y}$ = 3.505 – 0.00344 x.

Minitab

Regression Analysis: Fat gain versus NEA

The regression equation is
Fat gain = 3.51 - 0.00344 NEA

Predictor	Coef	SE Coef	T	P
Constant	3.5051	0.3036	11.54	0.000
NEA	-0.0034415	0.0007414	-4.64	0.000

S = 0.739853 R-Sq = 60.6% R-Sq(adj) = 57.8%

Merk

- Stigningsforholdet $b_1 = r\frac{s_y}{s_x}$ måler endringsraten. Korrlasjonen r er uavhengig av valg av skala som vi vet, men b_1 er ikke det. Angir vi verdien på responsvariabelen i gram i stedet for kilo i eksemplet med vekt og fysisk aktivitet ville b_1 bli 1000 ganger større.

- Regresjon er en av de viktigste metoder for en statistikker, og minste kvadrater er den vanligste formen for tilpasning.

- Siden $b_1 = r\frac{s_y}{s_x}$ er stigningsforholdet vil en endring på ett standardavvik i forklaringsvariabelen svare til en endring på r standardavvik i responsvariabelen siden

$$b_0 + r\frac{s_y}{s_x}(x + s_x) = b_0 + b_1 x + r\, s_y = (b_0 + b_1 x) + r\, s_y$$

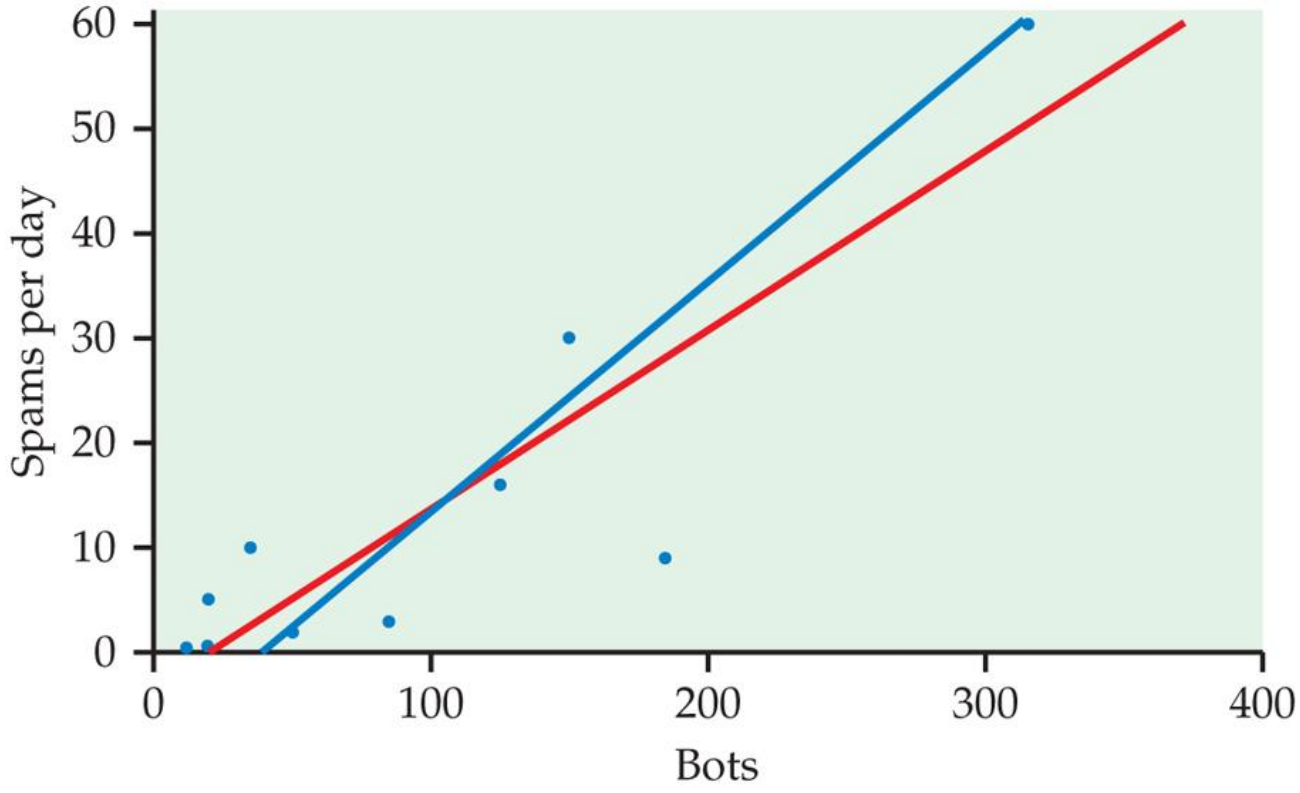
- Regresjonslinja går gjennom punktet ($\bar{x}$, $\bar{y}$) siden $b_0 + b_1 \bar{x} = \bar{y} - b_1 \bar{x} + b_1 \bar{x} = \bar{y}$. Sammen med at stigningningsforholdet er $b_1 = r \frac{s_y}{s_x}$ bestemmer dette regresjonslinja.

- I regresjon er skillet mellom respons- og forklaringsvariable sentralt. Som vi har sett er det både en ulik behandling og også en ulik tolkning av de to varablene.

Det er ikke alltid opplagt hvilken variabel som skal velges Som responsvariabel. I eksemplet med spam pr dag og antallet bots i et botnet kan man både forklare antallet spam per dag med størrelsen på botnettet eller omvent størrelsen med antallet spam pr dag.

Rød : Respons er antallet spam pr dag.

Blå: Respons er bots



Det er en nær sammenheng mellom korrelasjon og regresjon.
Vi har sett at stigningsforholdet kan uttrykkes ved korrelasjonen:

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

En annen sammenheng er at kvadratet av korrelasjon, r^2, måler styrken på sammenhengen mellom forklaringsvariabel og respons.

Mer spesifikt: r^2 er den delen av variasjonen i responsvariablen som er forklart ved hjelp av forklaringsvariabelen.

I eksempelet med vektøkning er $r = -0.7786$ og $r^2 = 0.606$.
Denne siste verdien rapporteres i utskrift fra statistikkpakker.

Minitab

Regression Analysis: Fat gain versus NEA

The regression equation is
Fat gain = 3.51 - 0.00344 NEA

Predictor	Coef	SE Coef	T	P
Constant	3.5051	0.3036	11.54	0.000
NEA	-0.0034415	0.0007414	-4.64	0.000

S = 0.739853 R-Sq = 60.6% R-Sq(adj) = 57.8%

Uttrykt på en litt annen måte: Variasjonen til observasjonene $y_1,\ldots,y_n$ er større enn variansjonen i de predikerte verdiene $\hat{y}_1,\ldots,\hat{y}_n$.

Vi har faktisk at

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

I vektøkningseksempelet kan man regne ut at variansen for $y_1,\ldots,y_n$ er lik 1.297
mens variansen for $\hat{y}_1,\ldots,\hat{y}_n$ blir lik 0.786

Dette stemmer med at

$$r^2 = \frac{\text{varians for } \hat{y}_1,\ldots,\hat{y}_n}{\text{varians for } y_1,\ldots,y_n} = \frac{0.786}{1.297} = 0.606$$

Forsiktighetsregler for korrelasjon og regresjon

Residualer er bestemt som differensen mellom observert og predikert verdi

Residual: $y - \hat{y}$

Eksempe En av personene i datasettet om vektøking og aktivitet hadde en verdi 2.7 kg på responsvariabelen og 135 kalorier på forklaringsvariabelen.
Da er residualet:

$2.7 - (\ 3.505 - \ 0.00344 \times 135) = 3.04$ kg

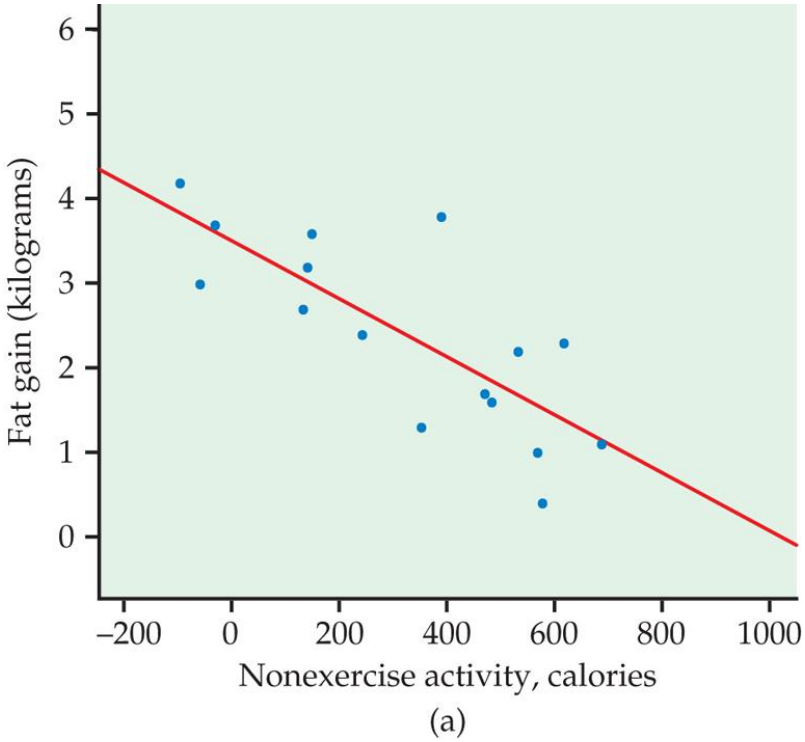
Merk

- Summen av residualer beregnet med minste kvadraters metode er lik null.

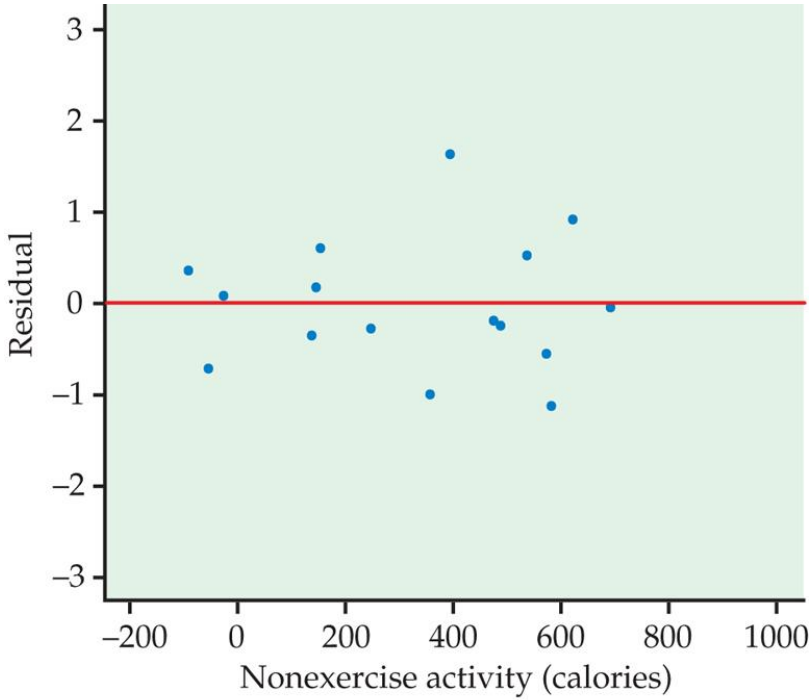
Residualer bidrar til å forstå hvor godt regresjonslinja tilpasser data og også å oppdage store avvik. Plott er nyttige.

Residualplott: Et residualplott er et spredningsplott av residualene mot forklaringsvariabelen.

For at en beskrivelse av data ved en regresjonslinje skal være tilfredsstillende skal det ikke være noe mønster i residualene.

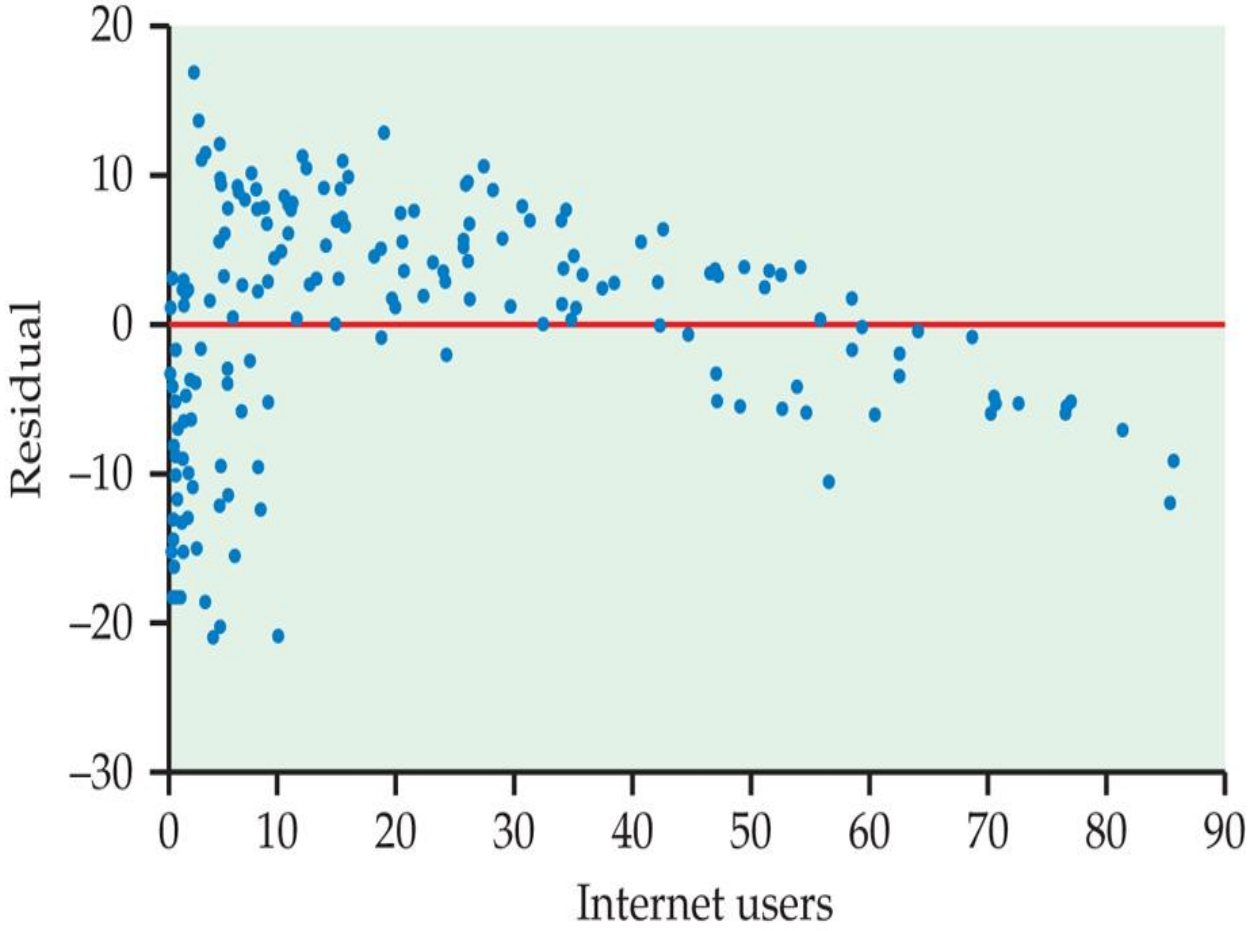

Fat gain (kilograms)
6
5
4
3
2
1
0
–200
0
200
400
600
800
1000
Nonexercise activity, calories

(a)


Residual
3
2
1
0
–1
–2
–3
–200
0
200
400
600
800
1000
Nonexercise activity (calories)

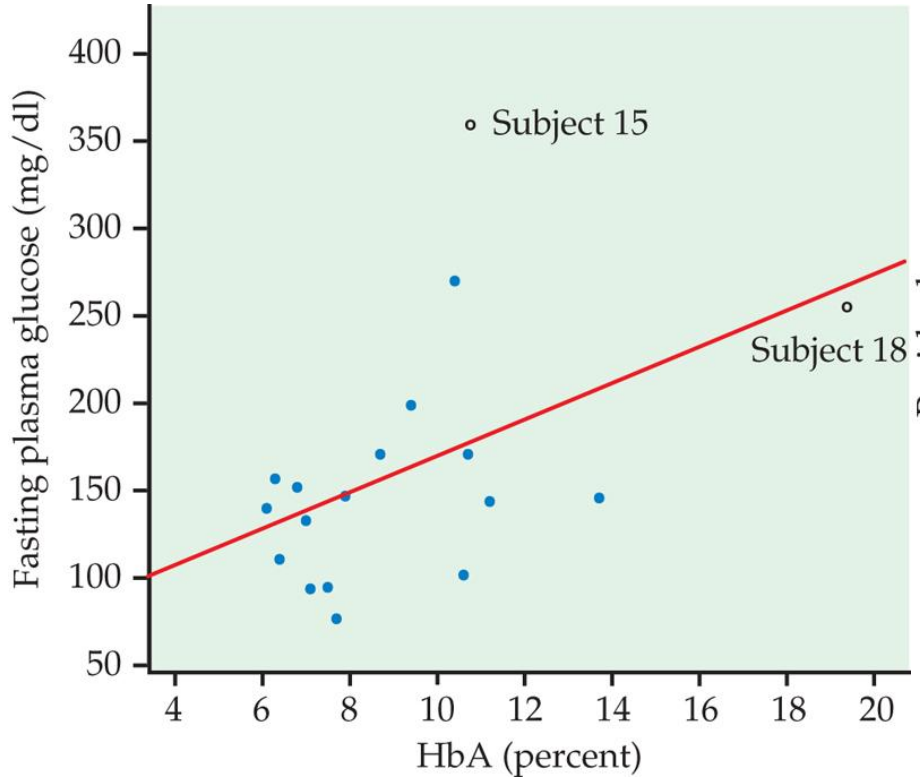
(b)

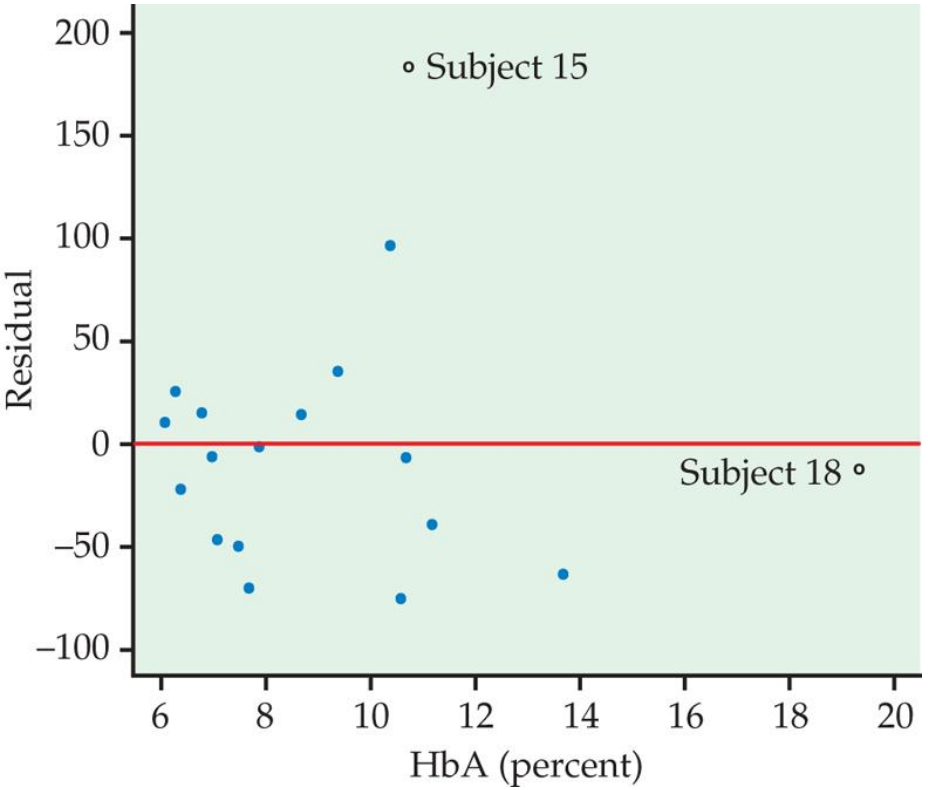
Eksempel Følgende residualplott er for data om fødselsrate (fødsler pr 1000 innbygger) og internettbruk (antall brukere pr 100 innbygger) i 181 land. Plottet viser at residualene svarende til små verdier av forklaringsvariabelen tenderer til å være positive. Dette er et eksempel på systematisk avvik.



Både det almenne mønsteret og spesielle avvik er av interesse.

Eksempel n= 18 målinger av glukose i blodet. To typer målinger
FPG, måles av paienten med glukosemeter
HbA1c, måles ved medisinsk checkup





Positiv sammenheng, men adskillig variasjon. To enheter spesielle. 15 med stort residual, 18 med høy verdi på forklaringsvariabelen.

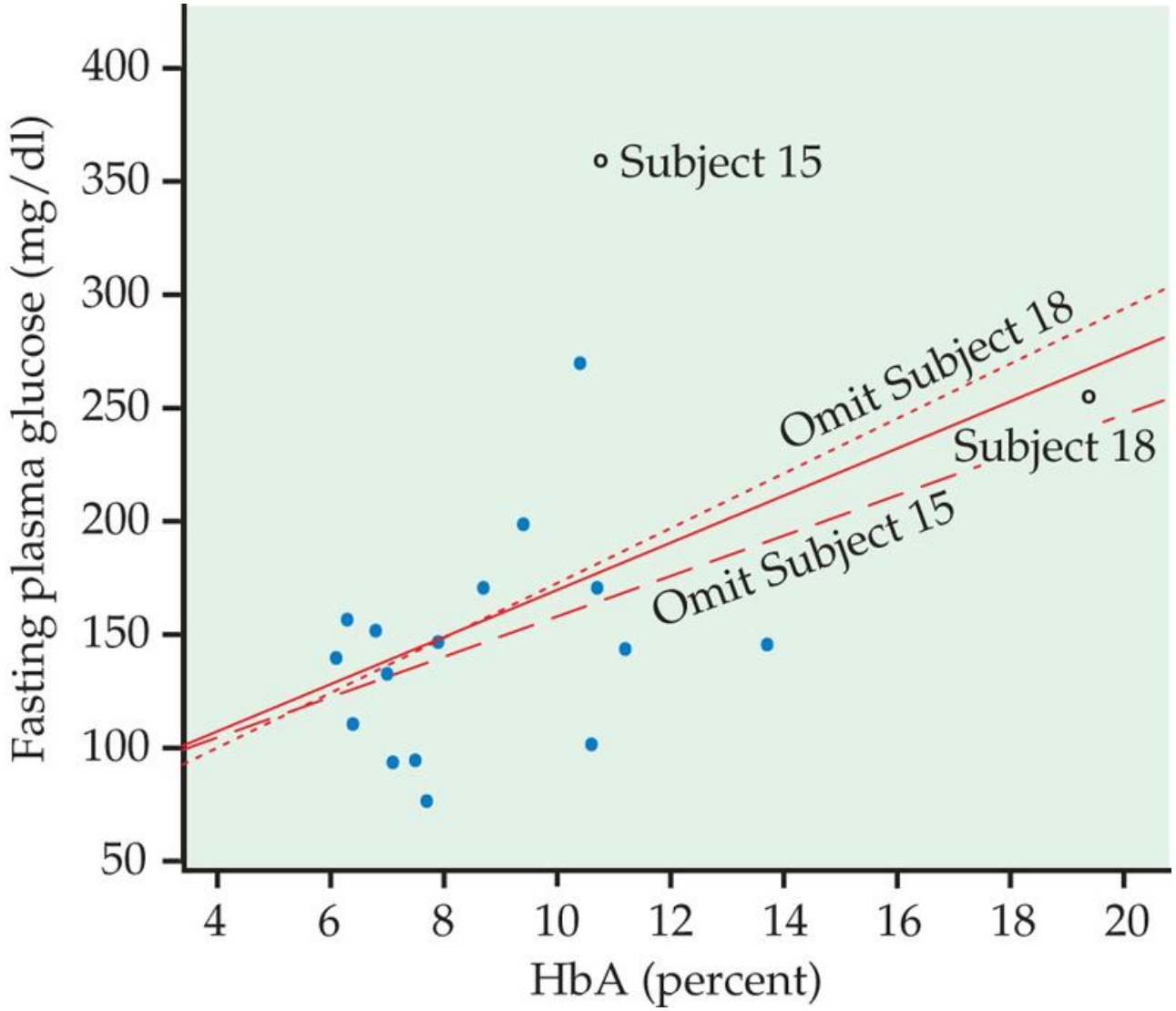
Outlier: En observasjon som skiller seg ut fra det totale bildet.

Outliere kan ha høye residualer men behøver nødvendigvis ikke ha det.

Innflytelsesrike observasjon: En observasjon hvor regresjonslinja blir helt annerledes hvis observasjonen fjernes.

Observasjoner som skiller seg ut med høye verdier på forklaringsvariabelen er ofte innflytelsrike.

I eksemplet med de to glukosemålingene er 18 ikke spesielt innflytelsesrik. Beregningene endres mer når observasjon 15 tas ut av materialet.


Fasting plasma glucose (mg/dl)
400
350
300
250
200
150
100
50
4
6
8
10
12
14
16
18
20
HbA (percent)
Subject 15
Omit Subject 18
Subject 18
Omit Subject 15

Merk En posititiv eller negativ sammenheng mellom to variabler kan skyldes at begge er relatert til en eller flere variable som det ikke er tatt hensyn til.

Underliggende (”lurkende”) variabel: En variabel som hverken er respons- eller forklaringsvariabel, men som likevel kan bety noe for tolkningen av sammenhengen mellom de to variablene.

Eksempel Flere studier viser at menn oftere får bypass kirurgi enn kvinner.

Sammenhengen kan skyldes at bypass kirurgi er en kirurgisk teknikk som fortrinnsvis bør anvendes for yngre, og at hjerteproblemer vanligvis inntreffer senere hos kvinner enn menn. Her er alder den underliggende variabelen.

To advarsler:

- Korrelasjon basert på gjennomsnitt er vanligvis høyere enn korrelasjon basert på individer.

- Pass på at forklaringsvariablene spenner ut et tilstrekkelig område.

Store datamengder er en utfordring. Eksplorativ utforsking av store datamengder kalles «data mining». Problemer som ekstrapolering, underliggende variable og sammenblanding av korrelasjon og årsak forsvinner ikke, tvert imot.

Sålangt har vi sett på sammenheng mellom to kvantitative variable.

Med en kategorisk og en kvantitativ variabel:
Lag boks-plott (evt. stilk-og-blad-plott) for den kvantitative variabelen etter hver verdi av kategoriske variablen.

Med to kategorisk variable:
Lag to-veis tabeller (kryss-tabeller)

Dataanalyse for toveistabeller

Toveistabeller kan brukes når begge veriablene er kategoriske.

Toveistabell: En tabell som angir antallet eller andelen av observasjonene som faller i hver kombinasjon av verdier av de to variablene.

Eksempel Kalsiuminntak, møter kravet, møter ikke kravet
Alder, 5-10 år, 11-13 år.

		Alder		
		5-10	11-13	Total
Møter kravet	nei	194	557	751
	ja	861	417	1278
	total	**1055**	**974**	**2029**

Simultan fordeling: Andelen i de ulike cellene i tabellen

		Alder		
		5-10	11-13	Total
Møter kravet	nei	0.10	0.27	0.37
	ja	0.42	0.21	0.63
	total	**0.52**	**0.48**	**1.00**

Sammenhenger mellom kategoriske variable beskrives ved de relevante prosentene beregnet fra antallene. Av de betingede fordelingene for de to alderskategoriene ser man at eldere barn sjeldnere møter kravet.

Betingede fordelinger: Andelen etter kollonne i tabellen (her)

		Alder		
		5-10	11-13	Total
Møter kravet	nei	0.18	0.57	0.37
	ja	0.82	0.43	0.63
	total	**1.00**	**1.00**	**1.00**

En annen betinget fordelinger er andelen etter rad i tabellen.

MINITAB utskrift fra en annen tabell

Minitab

	no	yes	All
men	5550	1630	7180
	77.30	22.70	100.00
	40.27	49.19	42.00
	32.46	9.53	42.00
women	8232	1684	9916
	83.02	16.98	100.00
	59.73	50.81	58.00
	48.15	9.85	58.00
All	13782	3314	17096
	80.62	19.38	100.00
	100.00	100.00	100.00
	80.62	19.38	100.00

Cell Contents --
Count
% of Row
% of Col
% of Tbl

Simpsons paradoks er et eksempel på underliggende (”lurkende”) variable i toveistabeller.

De observerte sammenhengene kan være misledende når det er en underliggene variabel og toveistabellen brytes ned etter hvilke verdier den bakenforliggende variabelen har.

172	118
28	82
200	**180**

162	19	10	99
18	1	10	81
180	**20**	**20**	**180**

Eksempel Utprøving av ny medisin. Her er kjønn underliggende (lurkende) variabel .

Alle

	Bedring	Ikke bedring	Antall	Suksess rate
Med med.	20	20	40	0.5= 20/40
Ikke med.	16	24	40	0.4= 16/40
Total	36	44	80	

Menn

Med med.	18	12	30	0.6=18/30
Ikke med.	7	3	10	0.7= 7/10
Total	25	15	40	

Kvinner

Med med.	2	8	10	0.2=2/10
Ikke med.	9	21	30	0.3=9/30
Total	11	29	40	

Kausalitet (årsakssammeheng)?

Ofte er årsakssammenhenger målet for undersøkelser.
Det er viktig å påpeke at korrelasjon ikke medfører kausalitet.

Eksempel på ulike x og y: Er dette årsakssammenhenger?

1. x mødres BMI, y døtres BMI

2. x funstig søtning i diett, y antall kreftsvulster i rotters urinblære

3. x studenters SAT score, y karakter første år på universitet

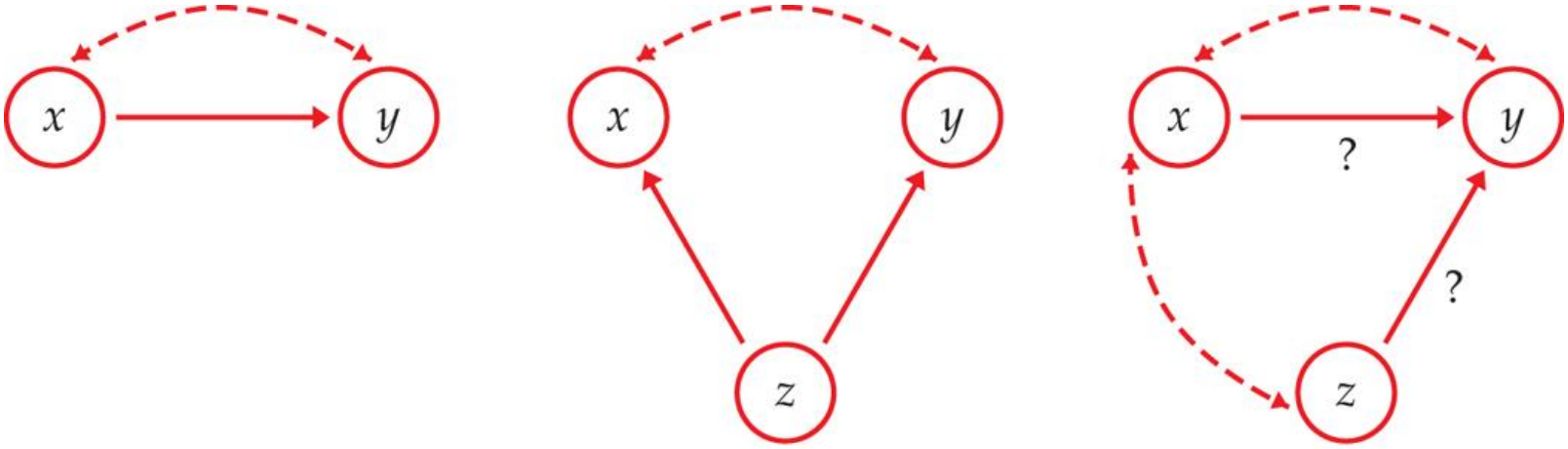
4. x månedlige investeringer i aksjefond, y månedlig avkastningsrate

5. x deltagelse i religiøs aktivitet, y levetid

6. x lengde på utdannelse, y lønn i kroner

Konfounding: to variabel er «confounded» hvis begge har effekter på responsvariabelen som ikke kan skilles.

I eksemplet er i følge MMC 1 og 2 årsakssammenhenger, 3 og 4 skyldes en felles underliggende variabel mens 5 og 6 er eksempler på konfounding.



Causation
(a)

Common response
(b)

Confounding
(c)

Merk:

- Variabler som er «confounded» kan være både forklaringsvariable, underliggende variable eller begge deler.

- Selv klare sammenhenger behøver ikke skyldes en årsakssammenheng.

Årsakssammenhenger?

Eksempel Kraftledninger og leukemi

Eksempel Røyking og lungekreft

Sammenheng sterk, konsistent fund i mange undersøkelser, høyere doser forbundet med økt forekomst, forskjell i tid, rimelig forklaring.