4 Korrelasjon og regresjon

Overskriftene viser til delkapitlene i boka. Først kommer en oppsummering av hele kapittelet. Under hver overskrift står gjennomregnede eksempler med løsning, og deretter oppgaver du skal løse uten hjelpemidler. Fasiten står bakerst. Ha linjal klar.

Formelark

KORRELASJON

Beskriv sammenhengen Retning, form, styrke og uteliggere i et punktdiagram.

Korrelasjonskoeffisient $-1 \leqslant r \leqslant 1$. $|r| < 0{,}3$ svak, 0,3–0,7 moderat, $> 0{,}7$ sterk.

Korrelasjon og årsak Årsak, omvendt årsak, bakenforliggende variabel eller tilfeldighet.

> - r måler bare lineære sammenhenger og påvirkes av uteliggere.

REGRESJON

Residual

$$y - (ax + b)$$

Regresjonslinja Minst mulig sum av kvadrerte residualer. Går gjennom $(\bar{x}, \bar{y})$.

Tolke a: endring i y når x øker med 1. b: verdien når $x = 0$.

Forklaringsgrad

$$R^2 = r^2$$

> - Interpolering er tryggere enn ekstrapolering.
> - Se på residualplottet: mønster betyr at en rett linje ikke passer.

4.1.1 Punktdiagram

4.1 Er sammenhengen trolig positiv, negativ eller ingen?

a) Høyde og skostørrelse.

b) Antall timer på mobilen og antall timer søvn.

c) Fødselsdag i måneden og karakter i matte.

d) Avstand til skolen og tid på skoleveien.

4.2 Tegn et punktdiagram og beskriv sammenhengen:

Timer trening per uke	0	1	2	3	4	5	6
Hvilepuls	78	75	72	70	66	64	60

4.3 Hvilken variabel bør stå langs den vannrette aksen?

a) Temperatur og salg av is.

b) Antall timer øving og antall riktige på en prøve.

4.4 Hva betyr det at en sammenheng er sterk?

4.5 Beskriv en sammenheng som er sterk, men ikke lineær.

4.6 Seks elever har disse verdiene (timer lekser, poeng): $(1, 10)$, $(2, 14)$, $(3, 15)$, $(4, 20)$, $(5, 22)$, $(6, 8)$. Tegn et punktdiagram. Hvilket punkt er en uteligger?

4.7 Hvorfor kan du ikke bruke punktdiagram for kjønn og høyde?

4.8 I et punktdiagram over utdanning og medianinntekt i kommunene ligger Oslo langt til høyre, men midt på i høyden. Hva betyr det?

4.9 Hva blir utskriften?

Python

```
x = [1, 2, 3, 4]
y = [5, 3, 4, 1]
for a, b in zip(x, y):
    print(a, b)
```

4.10 Hva blir utskriften?

Python

```
h = [160, 170, 180]
a = [158, 172, 179]
print([y - x for x, y in zip(h, a)])
```

4.11 Skisser tre punktdiagrammer: ett med sterk positiv, ett med svak negativ og ett med ingen sammenheng.

4.12 Hvorfor er det lurt å holde over punktene, eller skrive navn ved dem, når man leser et punktdiagram over kommuner?

4.1.2 Korrelasjonskoeffisienten

EKSEMPEL *Uten Oslo*

Korrelasjonen mellom utdanning og medianinntekt er $r = 0{,}434$ med alle kommunene og $r = 0{,}454$ uten Oslo. Forklar hvorfor r øker når Oslo tas ut.

Løsning

Oslo har mye høyere utdanning enn alle de andre, men bare middels inntekt. Punktet ligger langt under der linja gjennom de andre kommunene ville vært. Det trekker korrelasjonen ned. Uten Oslo passer en rett linje litt bedre, og r øker. Her er endringen liten fordi det er mange kommuner, men med få punkter kan én uteligger endre r kraftig. ∎

4.13 Hva er det største og minste r kan være?

4.14 Beskriv sammenhengen med ord.

a) $r = 0{,}95$ b) $r = -0{,}35$ c) $r = 0{,}02$ d) $r = -0{,}78$

4.15 Sorter fra svakest til sterkest: $0{,}6$, $-0{,}9$, $0{,}1$, $-0{,}4$.

4.16 Hva betyr $r = -1$?

4.17 To variabler har $r = 0$. Betyr det at de ikke henger sammen? Forklar.

4.18 Matche diagram og r: Et diagram har punkter tett langs en synkende linje, et annet en spredt sky som så vidt stiger, et tredje punkter tett langs en stigende linje. Verdiene er $0{,}2$, $0{,}97$ og $-0{,}9$.

4.19 Hvordan kan én uteligger påvirke r?

4.20 Korrelasjonen mellom årstall og tid på papiravis er $-0{,}98$ (SSB, 1991–2025). Tolk tallet.

4.21 To sammenhenger har $r = 0{,}9$. I den ene øker y med 5 per x, i den andre med 0,5. Hvilken er sterkest?

4.22 Hva blir utskriften?

Python

```
r = -0.65
if abs(r) > 0.7:
    print("sterk")
elif abs(r) >= 0.3:
    print("moderat")
else:
    print("svak")
```

4.23 Hva blir utskriften?

Python

```
for r in [0.5, -0.8, 0.2]:
    print(r ** 2)
```

4.24 En elev får $r = -1{,}2$. Hva kan du si?

4.1.3 Korrelasjon og årsak

EKSEMPEL *Hvilken forklaring?*

Foreslå en rimelig forklaring på korrelasjonen.

a) Is-salg og antall drukningsulykker per måned.

b) Antall brannbiler som rykker ut og skadene etter brannen.

c) Søvn og karakterer blant elever: de som sover mer, får bedre karakterer.

d) Utdanning og medianinntekt i kommunene (4.1.1).

Løsning

a) Bakenforliggende variabel: varme og sommer gir både is-salg og bading.

b) Bakenforliggende variabel: store branner gir både mange brannbiler og store skader. Det er ikke brannbilene som lager skadene.

c) Det kan være årsak (søvn hjelper læringen), men også bakenforliggende variabler (motivasjon, helse, rutiner hjemme) og omvendt årsak (stress over dårlige karakterer gir dårlig søvn).

d) Delvis årsak (utdanning gir ofte høyere lønn), men også bakenforliggende variabler: kommuner nær store byer har både mange med høyere utdanning og mange godt betalte jobber.

■

4.25 Nevn fire mulige forklaringer på at to variabler korrelerer.

4.26 Foreslå en bakenforliggende variabel.

a) Antall solbriller solgt og antall mygg.

b) Antall lærere og antall elever med dårlige karakterer i en kommune.

c) Skostørrelse og leseferdighet hos barn.

4.27 Elever som har mange venner, er mer fornøyd med skolen. Gi et eksempel på omvendt årsak.

4.28 Hvorfor korrelerer to tidsserier med trend ofte sterkt, selv om de ikke har noe med hverandre å gjøre?

4.29 En forsker undersøker 50 variabler i et lite datasett og finner én sterk korrelasjon. Hvorfor skal vi være skeptiske?

4.30 Hva er den sikreste måten å finne ut om X er årsak til Y?

4.31 Hvilke setninger påstår årsak?

a) «Folk som trener, har lavere blodtrykk.»

b) «Trening senker blodtrykket.»

c) «Det er en sammenheng mellom trening og blodtrykk.»

4.32 En undersøkelse viser at elever som spiser middag med familien, har bedre karakterer. Skriv om påstanden «Familiemiddag gir bedre karakterer» slik at den stemmer med hva undersøkelsen kan vise.

4.33 Tid på TV og tid på internett hadde $r = -0{,}92$ fra 2000 til 2025. Hvorfor er det ikke bevis for at internett har tatt over TV-tiden?

4.34 Hva blir utskriften?

Python

```
r = [0.8, -0.3, -0.65, 0.1]
print([x for x in r if abs(x) > 0.6])
```

4.35 Beskriv et eksperiment som kan avgjøre om musikk under lekser påvirker konsentrasjonen.

4.36 Kommuner med mange med høyere utdanning har ofte høyere medianinntekt. Nevn en bakenforliggende variabel som kan forklare noe av sammenhengen.

4.2.1 Regresjonslinja

EKSEMPEL *Hvilken linje er best?*

Fire punkter: $(1,2)$, $(2,3)$, $(3,5)$ og $(4,4)$. Regn ut summen av de kvadrerte residualene for linjene $y = x + 1$ og $y = x + 0{,}5$. Hvilken passer best?

Løsning

$y = x + 1$: residualene er $2-2=0$, $3-3=0$, $5-4=1$ og $4-5=-1$. Summen av kvadratene er $0+0+1+1=2$.

$y = x + 0{,}5$: residualene er $0{,}5$, $0{,}5$, $1{,}5$ og $-0{,}5$. Summen av kvadratene er $0{,}25+0{,}25+2{,}25+0{,}25=3$.

$y = x + 1$ passer best av de to. Regresjonslinja for de fire punktene er $y = 0{,}8\,x + 1{,}5$, med kvadratsum $1{,}8$. Ingen linje gir mindre.

■

4.37 Hva er en residual?

4.38 Linja er $y = 2x + 3$. Regn ut residualen for punktene $(1,6)$, $(2,6)$ og $(4,12)$.

4.39 Regn ut summen av de kvadrerte residualene for linja $y = x$ og punktene $(1,2)$, $(2,2)$, $(3,4)$.

4.40 Hvilken linje passer best for punktene $(0,0)$, $(1,2)$, $(2,3)$: $y = x + 0{,}5$ eller $y = 1{,}5\,x$?

4.41 Hva betyr det at regresjonslinja er laget med minste kvadraters metode?

4.42 Hva er summen av residualene for en regresjonslinje?

4.43 Hvorfor kvadrerer vi residualene i stedet for bare å summere dem?

4.44 Regresjonslinja går gjennom $(\bar{x}, \bar{y})$. Punktene har $\bar{x} = 4$ og $\bar{y} = 10$, og stigningstallet er 2. Hva er konstantleddet?

4.45 Tegn punktene $(1, 3)$, $(2, 4)$, $(3, 6)$, $(4, 7)$ og legg en linje på øyemål. Les av stigningstallet og konstantleddet for linja di.

4.46 Hva blir utskriften?

Python

```
x = [1, 2, 3]
y = [3, 5, 6]
a, b = 1.5, 1.7
for xi, yi in zip(x, y):
    print(round(yi - (a * xi + b), 1))
```

4.47 Hva blir utskriften?

Python

```
res = [0.5, -1, 0.5]
print(sum(res), sum(r ** 2 for r in res))
```

4.48 Hvilken Python-funksjon kan du bruke for å finne regresjonslinja, og hva gir den ut?

4.2.2 Tolke og predikere

EKSEMPEL *Tolke linjene*

Tolk stigningstallet og konstantleddet.

a) Armspenn: $y = 0{,}977\,x + 2{,}43$, der x er høyden i cm.

b) Medianinntekt i kommunene: $y = 4{,}29\,x + 541{,}2$, der x er andelen med høyere utdanning i prosent og y medianinntekten i tusen kroner.

c) Papiravis: $y = -1{,}08\,t + 42{,}84$, der t er år etter 1991.

Løsning

a) For hver centimeter høyere en elev er, er armspennet i gjennomsnitt 0,977 cm lengre, altså nesten like mye. Konstantleddet 2,43 cm er armspennet til en person som er 0 cm høy. Det gir ingen mening.

b) For hvert prosentpoeng høyere andel med høyere utdanning er medianinntekten i gjennomsnitt omtrent 4290 kr høyere. Konstantleddet, 541 200 kr, er inntekten i en kommune der ingen har høyere utdanning. Ingen kommune er i nærheten (lavest er 16,4 %), så det gir ikke mening.

c) Tiden på papiravis har i gjennomsnitt sunket med 1,08 minutter per år. Konstantleddet er verdien på linja i 1991 ($t = 0$): omtrent 43 minutter. Her gir det mening, fordi 1991 er med i dataene. (Den faktiske verdien var 39 minutter.)

EKSEMPEL *Papiravisen i 2010 og 2035*

Bruk $y = -1{,}08\,t + 42{,}84$, der t er år etter 1991.

a) Prediker tiden på papiravis i 2010. Faktisk verdi var 23 minutter.

b) Prediker tiden i 2035. Er prediksjonen til å stole på?

Løsning

a) $y = -1{,}08 \cdot 19 + 42{,}84 \approx 22{,}3$ minutter. Det er interpolering, og prediksjonen er nær den faktiske verdien.

b) $y = -1{,}08 \cdot 44 + 42{,}84 \approx -4{,}7$ minutter. Det er ekstrapolering, og svaret er umulig: Tid kan ikke være negativ. Nedgangen har dessuten flatet ut de siste årene. Linja passer for perioden 1991–2025, men kan ikke forlenges langt.

4.49 En regresjonslinje for hvor mye en plante vokser, er $y = 1{,}8\,x + 12$, der x er antall uker og y høyden i cm. Tolk 1,8 og 12.

4.50 Bruk $y = 1{,}8\,x + 12$ til å predikere høyden etter 6 uker.

4.51 Hva er forskjellen på interpolering og ekstrapolering?

4.52 Dataene for planten gjelder uke 0 til 10. Er det interpolering eller ekstrapolering å predikere for uke 4? For uke 40?

4.53 Regresjonslinja for papiravis er $y = -1{,}08\,t + 42{,}84$, der t er år etter 1991 (SSB). Prediker tiden i 2000.

4.54 Hvorfor er det lurt å bruke $t =$ år etter et startår i stedet for selve årstallet?

4.55 En regresjonslinje for månedslønn (kr) ut fra antall år i jobben er $y = 1200x + 38\,000$. Tolk stigningstallet.

4.56 Når gir konstantleddet ikke mening? Gi et eksempel.

4.57 Linja $y = -0{,}45\,x + 280$ gir bruktpris (tusen kr) ut fra kjørelengde (tusen km). Dataene går fra 10 til 200 tusen km. Prediker prisen ved 50 og ved 800 tusen km, og kommenter.

4.58 Hva blir utskriften?

Python

```
a = 10.55
b = -11.0
for t in [5, 25, 30]:
    print(2000 + t, round(a * t + b))
```

4.59 Hva blir utskriften?

Python

```
a = -1.08
b = 42.84
t = 0
while a * t + b > 0:
    t = t + 1
print(1991 + t)
```

4.60 Løs $-1{,}08\,t + 42{,}84 = 20$ for å finne når linja for papiravis gir 20 minutter.

4.2.3 Hvor god er modellen?

EKSEMPEL *Tolk R^2*

En regresjonslinje for prisen på brukte biler ut fra kjørelengde har $r = -0{,}7$.

a) Hva er R^2?

b) Hva betyr det?

Løsning

a) $R^2 = (-0{,}7)^2 = 0{,}49$

b) Kjørelengden forklarer 49 % av variasjonen i bruktprisen. Resten, 51 %, skyldes andre ting, som merke, alder, utstyr og tilstand. ■

4.61 Regn ut R^2 for $r = 0{,}6$, $r = -0{,}9$ og $r = 0{,}1$.

4.62 $R^2 = 0{,}64$. Hva kan r være?

4.63 Tolk $R^2 = 0{,}49$ for en modell som predikerer karakter ut fra antall timer øving.

4.64 Hva skal du se etter i et residualplott?

4.65 Residualene for en modell er: negative for små x, positive for middels x og negative for store x. Hva betyr det?

4.66 En modell har $R^2 = 0{,}97$, men residualplottet viser en tydelig bue. Er modellen god? Forklar.

4.67 Høyden forklarer 87 % av variasjonen i armspenn i en klasse. Hvor mye forklares ikke, og hva kan det skyldes?

4.68 Er R^2 alltid positiv? Forklar.

4.69 Hva blir utskriften?

Python

```
r = -0.7
print(round(r ** 2, 2), round(1 - r ** 2, 2))
```

4.70 Hva blir utskriften?

Python

```
res = [2, -1, -2, 1]
print(sum(1 for r in res if r > 0))
```

4.71 Skisser et residualplott for en modell der spredningen øker med x. Hva betyr det for prediksjonene?

4.3.1 Undersøkelse: Henger det sammen?

4.72 Hva bør rapporten fra en undersøkelse av en sammenheng inneholde?

4.73 Hvordan kan du lage variabelen «personer per husholdning» fra datasettene om kommunene?

4.74 Personer per husholdning og medianinntekt i kommunene har $r = 0{,}83$. Hva er R^2, og hva betyr det?

4.75 Regresjonslinja er $y = 357{,}9\,x - 104{,}1$ (tusen kr), der x er personer per husholdning. Prediker medianinntekten i en kommune med 2,0 personer per husholdning.

4.76 Hvorfor er det rimelig at husholdninger med flere personer har høyere inntekt?

4.77 Hvorfor kan vi ikke legge sammen at utdanning forklarer 19 % og personer per husholdning 69 %, og si at de til sammen forklarer 88 %?

4.78 Hvilken begrensning har det at hver verdi i inntektsdatasettet er en median for en hel kommune?

4.79 Hva blir utskriften?

Python

```
folk = [3000, 12000]
hush = [1400, 5000]
for f, h in zip(folk, hush):
    print(round(f / h, 2))
```

4.80 Formuler et eget spørsmål om en sammenheng i kommunedataene, og si hva som er x og y.

4.81 En gruppe skriver: «Høyere utdanning gir høyere inntekt i kommunene, $r = 0{,}43$.» Hva bør de endre?

Fasit

4.1.1 Punktdiagram

4.1
a) Positiv b) Negativ c) Ingen d) Positiv

4.2
Trening langs den vannrette aksen, puls langs den loddrette. Negativ, lineær og sterk sammenheng uten uteliggere.

4.3
a) Temperatur b) Antall timer øving (den forklarende variabelen).

4.4
At punktene ligger tett rundt mønsteret (for eksempel en rett linje).

4.5
For eksempel TV-tid mot år, som økte og så sank, eller vekst hos et barn som er rask først og så flater ut. Punktene følger en tydelig kurve.

4.6
$(6, 8)$: alle de andre følger en positiv, omtrent lineær sammenheng.

4.7
Kjønn er en kategorisk variabel. Punktdiagram brukes for to numeriske variabler. Bruk heller boksplott for hver gruppe.

4.8
Oslo har svært høy andel med høyere utdanning, men bare middels medianinntekt for husholdningene. Oslo passer ikke med mønsteret til de andre kommunene.

4.9

```
1 5
2 3
3 4
4 1
```

4.10

```
[-2, 2, -1]
```

4.11
Sterk positiv: punktene tett langs en stigende linje. Svak negativ: spredte punkter som så vidt synker mot høyre. Ingen: en sky uten retning.

4.12
Da ser man hvilke enheter som er uteliggere eller skiller seg ut, og kan forsøke å forklare hvorfor.

4.1.2 Korrelasjonskoeffisienten

4.13
1 og -1.

4.14
a) Sterk positiv b) Moderat negativ c) Ingen lineær d) Sterk negativ

4.15
$0{,}1$, $-0{,}4$, $0{,}6$, $-0{,}9$

4.16
Alle punktene ligger nøyaktig på en rett linje som synker mot høyre.

4.17
Nei, bare at det ikke er noen lineær sammenheng. Det kan finnes en sterk sammenheng av en annen form, for eksempel en U-form.

4.18
Synkende og tett: $-0{,}9$. Spredt og svakt stigende: $0{,}2$. Stigende og tett: $0{,}97$.

4.19
Den kan gjøre r mye større eller mindre, særlig når det er få punkter. Et punkt langt ute i samme retning som mønsteret øker r, et punkt langt unna mønsteret senker r.

4.20
Svært sterk negativ lineær sammenheng: tiden på papiravis har sunket jevnt over tid.

4.21
De er like sterke. r måler hvor tett punktene ligger rundt linja, ikke hvor bratt linja er.

4.22

```
moderat
```

4.23

```
0.25
0.6400000000000001
0.04000000000000001
```

4.24
Det er en feil: r kan ikke være mindre enn -1.

4.1.3 Korrelasjon og årsak

4.25
Årsak, omvendt årsak, bakenforliggende variabel, tilfeldighet.

4.26
a) Sommer/sol b) Antall elever (kommunens størrelse) c) Alder

4.27
Elever som trives på skolen, blir lettere kjent med andre og får flere venner.

4.28
Begge endrer seg over tid. Alt som øker (eller synker) i samme periode, vil korrelere, uansett om det er en årsakssammenheng.

4.29
Når man leter blant mange par, finner man nesten alltid noen sterke korrelasjoner ved tilfeldighet, særlig med få observasjoner. Sammenhengen må testes på nye data.

4.30
Et randomisert eksperiment med kontrollgruppe.

4.31
Bare b).

4.32
For eksempel: «Elever som spiser middag med familien, har i gjennomsnitt bedre karakterer.»

4.33
Begge er tidsserier med trend, så de ville korrelert sterkt uansett. Mange andre ting har endret seg i samme periode. Korrelasjonen viser ikke hvor mye av TV-tiden som er flyttet til nettet.

4.34

```
[0.8, -0.65]
```

4.35
Fordel elever tilfeldig på to grupper som får samme oppgaver og tid. Den ene gruppen hører musikk, den andre ikke. Sammenlikn antall riktige.

4.36
For eksempel avstand til en stor by, eller næringslivet i kommunen.

4.2.1 Regresjonslinja

4.37
Observert verdi minus verdien på linja: $y - (ax + b)$.

4.38
$6 - 5 = 1$, $6 - 7 = -1$, $12 - 11 = 1$

4.39
Residualer $1, 0, 1$. Sum av kvadrater: 2.

4.40
$y = x + 0{,}5$: residualer $-0{,}5, 0{,}5, 0{,}5$, sum av kvadrater $0{,}75$. $y = 1{,}5\,x$: residualer $0, 0{,}5, 0$, sum $0{,}25$. $y = 1{,}5\,x$ passer best.

4.41
At den har den minste mulige summen av kvadrerte residualer blant alle rette linjer.

4.42
0.

4.43
Positive og negative residualer ville opphevet hverandre. Kvadrering gjør alle positive, og gir én entydig beste linje.

4.44
$10 = 2 \cdot 4 + b$ gir $b = 2$.

4.45
Omtrent $y = 1{,}4\,x + 1{,}5$.

Regresjonslinja er $y = 1{,}4\,x + 1{,}5$.

4.46

```
-0.2
0.3
-0.2
```

4.47

```
0.0 1.5
```

4.48
`np.polyfit(x, y, 1)`. Den gir stigningstallet a og konstantleddet b.

4.2.2 Tolke og predikere

4.49
Planten vokser i gjennomsnitt 1,8 cm per uke. 12 cm er høyden ved start ($x = 0$).

4.50
$1{,}8 \cdot 6 + 12 = 22{,}8$ cm

4.51
Interpolering: predikere innenfor dataene. Ekstrapolering: predikere utenfor dataene.

4.52
Uke 4: interpolering. Uke 40: ekstrapolering (og trolig urimelig, planten slutter å vokse).

4.53
$t = 9$: $-1{,}08 \cdot 9 + 42{,}84 \approx 33{,}1$ minutter.

4.54
Da blir konstantleddet verdien i startåret, som gir mening. Med årstallet blir det verdien i år 0.

4.55
For hvert år i jobben er lønna i gjennomsnitt 1200 kr høyere per måned.

4.56
Når $x = 0$ ligger langt utenfor dataene. For eksempel er armspennet til en person som er 0 cm høy, meningsløst.

4.57
50: 257,5 tusen kr (interpolering). 800: −80 tusen kr, umulig (ekstrapolering).

4.58

```
2005 42
2025 253
2030 306
```

4.59

```
2031
```

4.60
$t = \dfrac{22{,}84}{1{,}08} \approx 21{,}1$, altså rundt 2012.

4.2.3 Hvor god er modellen?

4.61
0,36, 0,81 og 0,01

4.62
0,8 eller $-0{,}8$.

4.63
Øvingstiden forklarer 49 % av variasjonen i karakterer. Resten skyldes andre ting.

4.64
Om residualene spres tilfeldig rundt 0 (linja passer), eller om det er et mønster, som en bue (ikke lineær) eller økende spredning.

4.65
Sammenhengen er ikke lineær; den krummer. Linja bommer systematisk.

4.66
Den fanger opp mye av variasjonen, men formen er feil. Den vil bomme systematisk, særlig ved ekstrapolering.

4.67
13 %. For eksempel ulik kroppsbygning og målefeil.

4.68
Ja, $R^2 = r^2$ er et kvadrat, så den er mellom 0 og 1.

4.69

```
0.49 0.51
```

4.70

```
2
```

4.71
Residualene ligger tett rundt 0 for små x og spres mer for store x (traktform). Prediksjonene er mer usikre for store x.

4.3.1 Undersøkelse: Henger det sammen?

4.72
Spørsmål og variabler, punktdiagram med regresjonslinje, r og R^2, tolkning av stigningstallet i konteksten, residualer, drøfting av årsak og begrensninger.

4.73
Del folketallet på antall husholdninger i hver kommune.

4.74
$R^2 \approx 0{,}69$: personer per husholdning forklarer omtrent 69 % av variasjonen i medianinntekt mellom kommunene.

4.75
$357{,}9 \cdot 2 - 104{,}1 = 611{,}7$ tusen kr

4.76
Flere voksne i husholdningen betyr ofte flere inntekter.

4.77
Variablene henger sammen med hverandre, så forklaringene overlapper.

4.78
Vi vet ikke om sammenhengen gjelder for hver enkelt husholdning, og kommunene teller likt uansett størrelse.

4.79

```
2.14
2.4
```

4.80
For eksempel: «Har kommuner med flere innbyggere høyere andel med høyere utdanning?» x: folketall, y: andel med høyere utdanning.

4.81
r viser en moderat sammenheng, ikke årsak. De bør skrive at det er en sammenheng, nevne mulige bakenforliggende variabler (for eksempel husholdningsstørrelse eller nærhet til byer), og at R^2 bare er 0,19.