PSY2014 – Kvantitativ metode

Forelesning 3: Multippel regresjon

Nikolai Czajkowski

Opptak

- **Det gjøres opptak av forelesningen**
- Opptaket vil bli lagret på emnesiden til PSY2014 UiO, og en lenke vil tilgjengelig for de som følger kurset.
- Opptaket skal bli slettet etter 2023.

Multippel regresjon (MR)

- Partielle regresjonskoeffisienter
- Statistisk kontroll
- Supressoreffekter
- Forklart varians i MR
- Partiell korrelasjon
- Polynomisk regresjon
- Innflytelse

Introduksjon til multippel regresjon

Hvorfor ha mer enn en uavhengig variabel?

1. **Mer nøyaktig prediksjons**
   - Den avhengige variabelen er hovedfokus.
   - *F.eks. Predikere bedre hvem som dropper ut av skolen.*

2. **Isolere effekten av en uavhengig variabel (statistisk kontroll)**
   - Interessen er i en eller flere spesifikke uavhengige variabler.
   - *F.eks. Øker fattigdom risikoen for frafall?*

3. **Forstå den samlede effekten av flere uavhengige variabler**
   - Flere uavhengige variabler kan samlet påvirke utfallet på kompliserte måter.
   - *F.eks. Er effekten av fattigdom på frafall sterkere for barn av enslige foreldre (interaksjon).*
   - *F.eks. Fører fattigdom til frafall fordi foreldre i fattige familier har mindre tid til å følge med på barna (mediering).*

Multippel regresjon (A:11.1)

- Har vi to uavhengige variabler X_1 og X_2, refererer vi til person i's skårer på disse variablene som X_{1i} og X_{2i} (f.eks.utdanning og arbeidserfaring)

- Med p uanhengige variabler blir regresjonsuttrykket:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_p X_{pi} + \epsilon_i$$

- Parametrene (b_j) kalles *partielle regresjonskoeffisienter*.
  - De tolkes som *forventet endring i Y som følger av at X_j endres en enhet, men alle andre uavhengige variabler holdes konstant.*

- Den forventede verdien for Y for individ i basert på verdiene på de uavhengige variablene er gitt ved:

$$E(Y_i) = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_p X_{pi}$$

```
mod1←lm(Inntekt~Utdanning+Erfaring)
summary(mod1)
```

```
##
## Call:
## lm(formula = Inntekt ~ Utdanning + Erfaring)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.699 -14.802   0.624  14.908  67.199
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 301.8267     5.1944   58.11   <2e-16 ***
## Utdanning    20.0193     0.8835   22.66   <2e-16 ***
## Erfaring      4.8827     0.4275   11.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.03 on 247 degrees of freedom
## Multiple R-squared:  0.7225,    Adjusted R-squared:  0.7202
## F-statistic: 321.5 on 2 and 247 DF,  p-value: < 2.2e-16
```

$$\hat{Inntekt} = \hat{b}_0 + \hat{b}_1 \cdot Utdanning_i + \hat{b}_2 \cdot Erfaring_i$$

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 301.8267     5.1944   58.11   <2e-16 ***
## Utdanning    20.0193     0.8835   22.66   <2e-16 ***
## Erfaring      4.8827     0.4275   11.42   <2e-16 ***
## ---
```

- Hva er den forventede inntekt for et individ uten utdanning og uten erfaring?
- Hvis jeg holder erfaring konstant, hvor mye øker inntektene for hvert år utdanning?

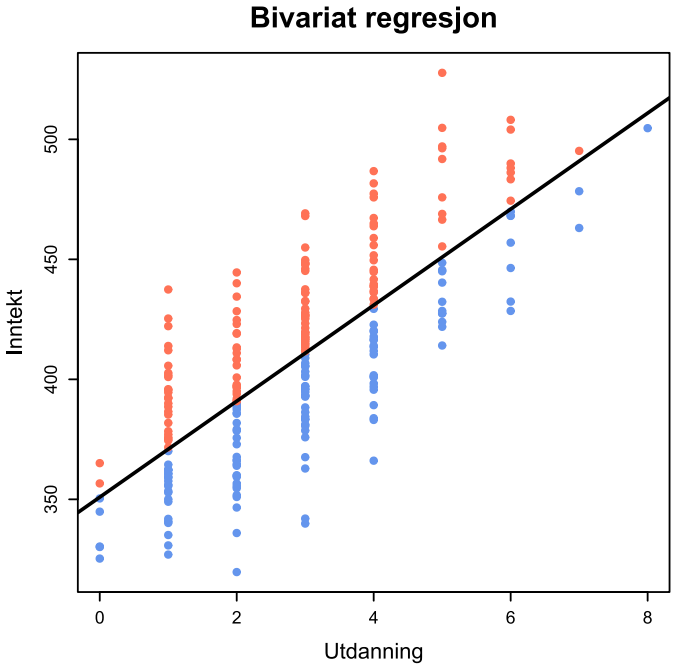
$$\hat{Inntekt} = \hat{b}_0 + \hat{b}_1 \cdot Utdanning_i + \hat{b}_2 \cdot Erfaring_i$$

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 301.8267     5.1944   58.11   <2e-16 ***
## Utdanning    20.0193     0.8835   22.66   <2e-16 ***
## Erfaring      4.8827     0.4275   11.42   <2e-16 ***
## ---
```

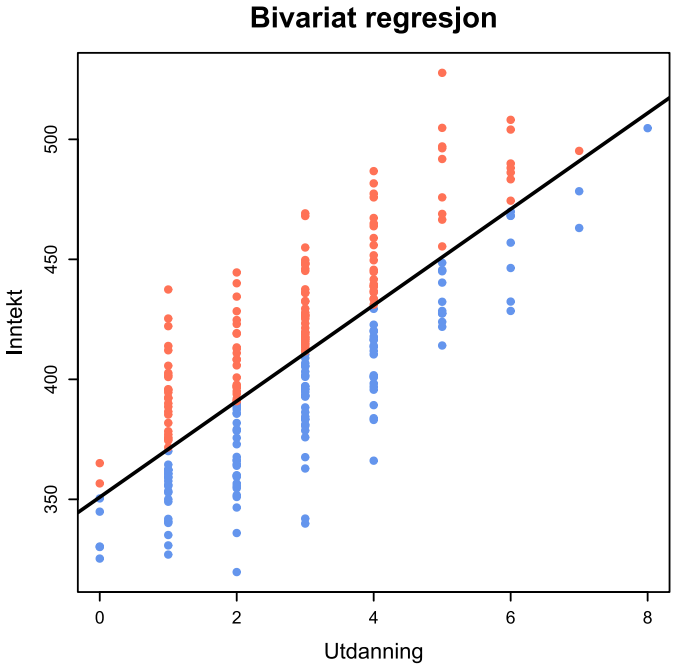
- Hva er den forventede inntekt for et individ uten utdanning og uten erfaring?
  - $\hat{Inntekt} = 301.8 + 20.0 \cdot 0 + 4.9 \cdot 0 = 301.8 = \hat{b}_0$

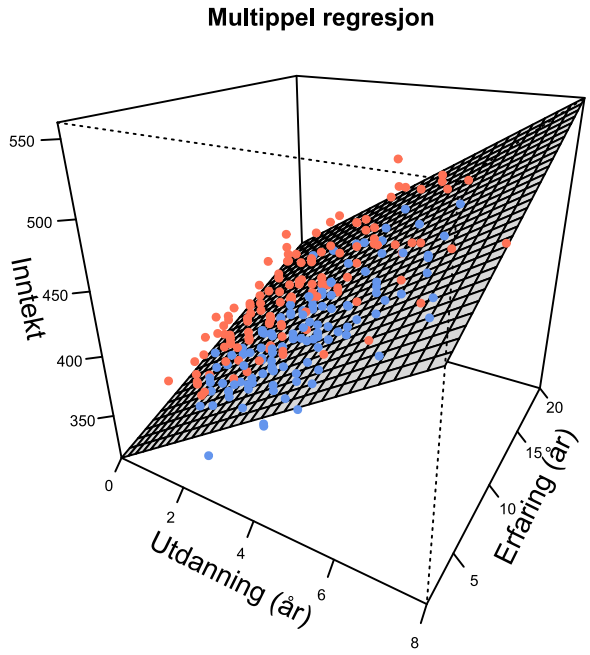
- Hvis jeg holder erfaring konstant, hvor mye øker inntektene for hvert år utdanning?
  - 20.0

Geometrisk representasjon



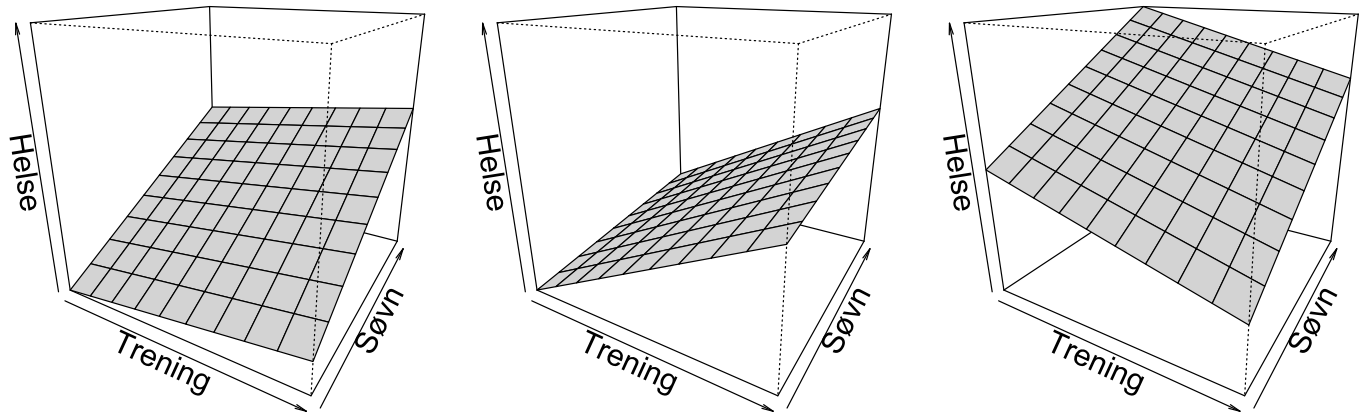
Geometrisk representasjon





Med to uanhengige variabler utgjør regresjonsmodellen (for den forventede Y-verdien) et plan i tredimensjonalt rom.

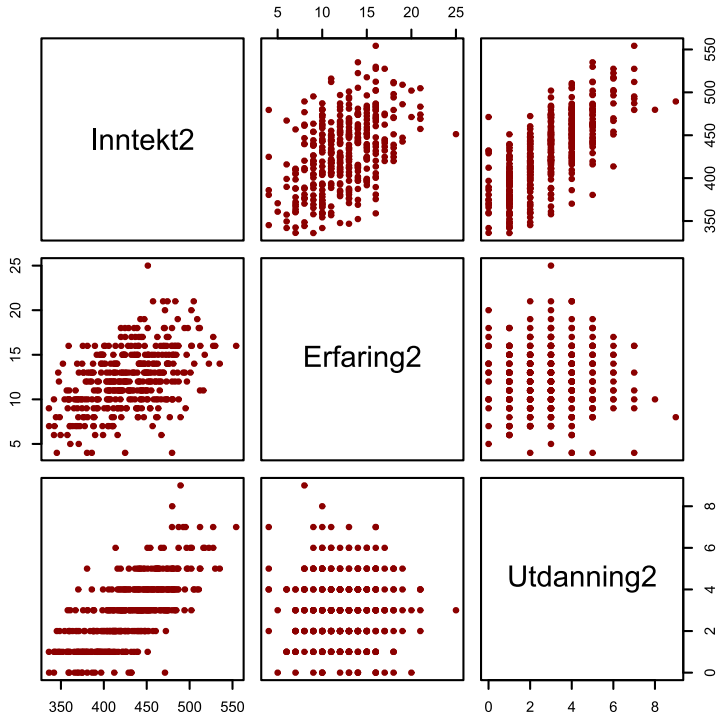
Hvilket regresjonsplan korresponderer til utskriften?




```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.0002582  0.0015005    6665   <2e-16 ***
## Søvn         0.9999792  0.0001627    6146   <2e-16 ***
## Trening     -0.5000364  0.0001185   -4220   <2e-16 ***
## ---
```

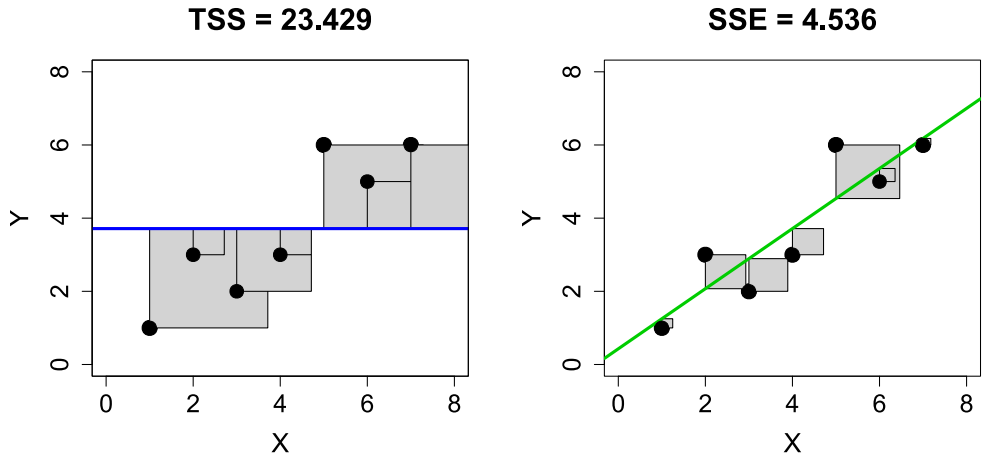
Matrix scatterplot

```
pairs(cbind(Inntekt2, Erfaring2, Utdanning2), pch=20, col="darkred")
```



R^2 i multippel regresjon (A11.2)

Forklart varians (Coefficient of determination) (A: p274)



$$r^2 = 1 - \frac{SSE}{TSS}$$

- r^2 er et mål for *nedgangen i proporsjonen* av uforklart varians når vi går fra en modell til en annen.

Justert r^2

- Du kan forklare varians ved tilfeldighet, særlig dersom du har du mange uavhengige variabler og et lite utvalg.
  - F.eks. Spør du 5 mennesker om de er for eller mot medlemskap i EU, kan det hende at alle som er for er blonde, mens alle som er imot har brunt hår. Kan hårfarge forklare 100% av variansen i oppfatning?

- **Justert r^2** er en statistikk som korrigerer r^2 for utvalgsstørrelse og antall uavhengige variabler i modellen.
  - Når n (antall observasjoner) er stor og p (antall uavhengige variabler) er liten, så er r^2 og justert r^2 praktisk talt identiske.

$$Justert \quad r^2 = 1 - \frac{(n-1)(1-r^2)}{n-p-1}$$

- r^2 øker alltid om du legger til flere uavhengige variabler, men justert r^2 kan synke.

R^2 NÅR PREDIKTORENE ER UKORRELERTE

```
cor(Utdanning2, Erfaring2)
```

```
## [1] 0.009907241
```

```
summary(lm(Inntekt2~Utdanning2))
```

```
## Multiple R-squared:  0.5114,    Adjusted R-squared:  0.5102
## F-statistic: 416.6 on 1 and 398 DF,  p-value: < 2.2e-16
```

```
summary(lm(Inntekt2~Erfaring2))
```

```
## Multiple R-squared:  0.2126,    Adjusted R-squared:  0.2106
## F-statistic: 107.5 on 1 and 398 DF,  p-value: < 2.2e-16
```

```
summary(lm(Inntekt2~Utdanning2+Erfaring2))
```

```
## Residual standard error: 22.28 on 397 degrees of freedom
## Multiple R-squared:  0.7175,    Adjusted R-squared:  0.7161
```

Når de uavhengige variablene er ukorrelerte forklarer de «separate proporsjoner» av variansen i den avhengige, og forklart varians i en modell med begge er summen av hva de kan forklare enkeltvis.

R^2 NÅR PREDIKTORENE ER KORRELERTE

```
cor(Alder2, Erfaring2)
```

```
## [1] 0.7579404
```

```
summary(lm(Inntekt2~Alder2))
```

```
## Multiple R-squared:  0.1298,    Adjusted R-squared:  0.1276
## F-statistic: 59.36 on 1 and 398 DF,  p-value: 1.055e-13
```

```
summary(lm(Inntekt2~Erfaring2))
```

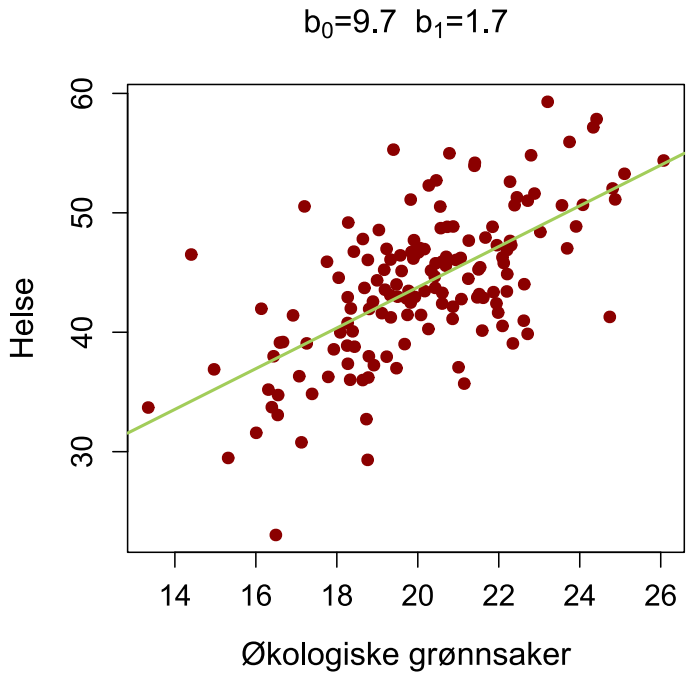
```
## Multiple R-squared:  0.1905,    Adjusted R-squared:  0.1885
## F-statistic: 93.68 on 1 and 398 DF,  p-value: < 2.2e-16
```

```
summary(lm(Inntekt2~Alder2+Erfaring2))
```

```
## Multiple R-squared:  0.1926,    Adjusted R-squared:  0.1885
## F-statistic: 47.34 on 2 and 397 DF,  p-value: < 2.2e-16
```

Når to uavhengige variabler er korrelerte, vil den variansen de kan forklare samlet være *mindre* enn summen av hva de kan forklare enkeltvis.

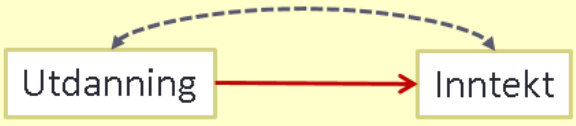
Økologisk mat



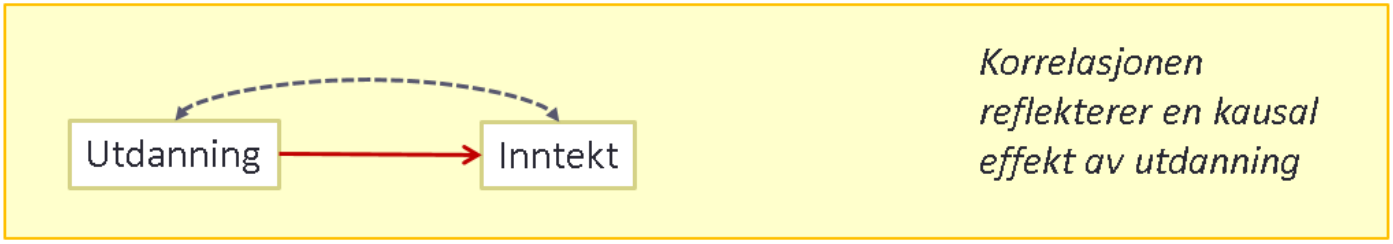
- Er det en helsegevinst av å spise organisk mat? (mat produsert uten pesticider eller syntetisk gjødsel).
- Hvis vi plotter folks inntak av økologisk mat mot deres helse, ser det ut til at høyere nivåer av økologisk mat er forbundet med bedre helse.

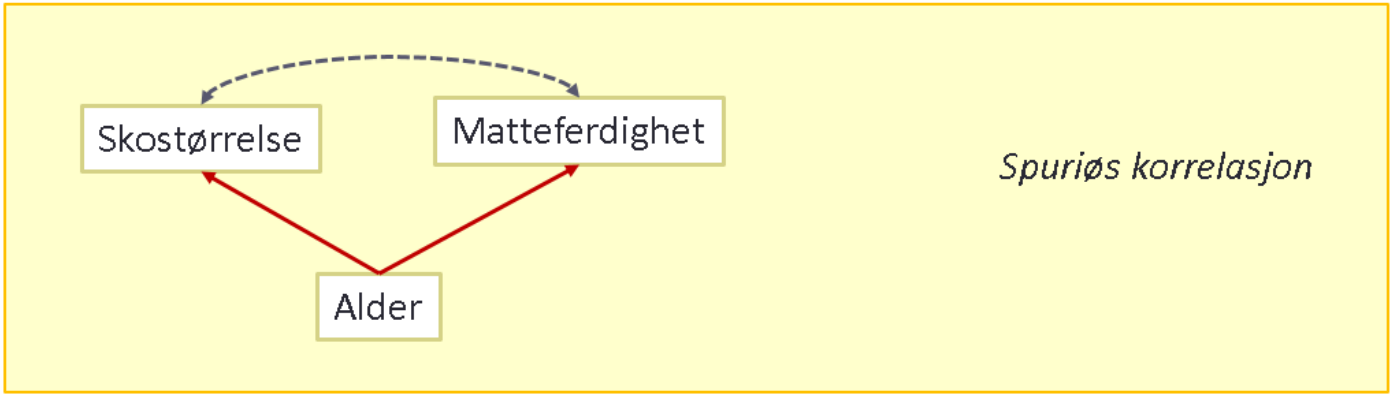
Statistisk kontroll

Problemet med tredjevariabler



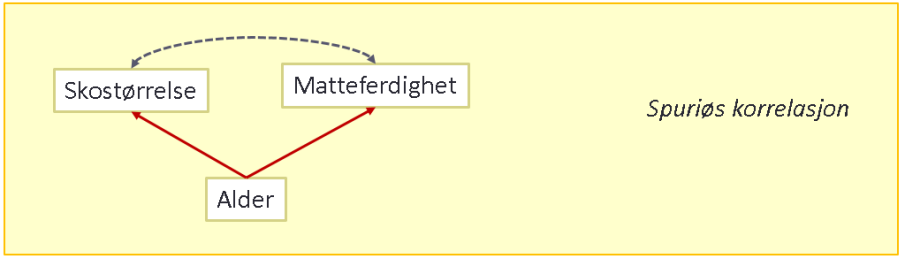
Korrelasjonen reflekterer en kausal effekt av utdanning





Utenforliggende tredjevariabler kan skape en spuriøs korrelasjon.

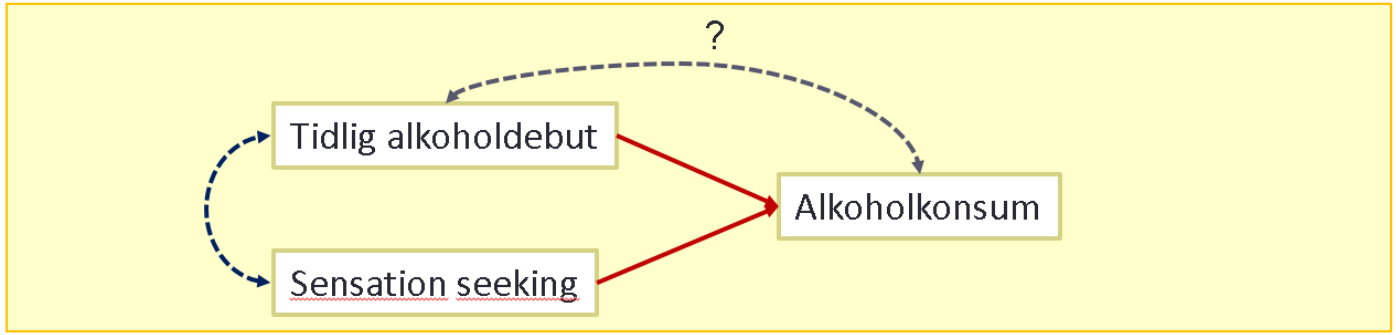
Spuriøs: Falsk, brukt hovedsakelig innen epidemiologi for å beskrive en relasjon som *virker kausal*, men som ikke er det.



- **Utenforliggende variabel** (extraneous): andre variabler enn den uavhengige variabelen i et eksperiment som påvirker (eller er assosiert med) den avhengige variabelen.
  - F.eks. for inntekt; kjønn, personlighet, ...
  - Dette er variabler som (du kanskje ikke har målt), men som står for noe av den uforklarte «erroren» rundt regresjonslinjen.

- **En konfunderende variabel** er en utenforliggende variabel som er assosiert med (korrelerer med) både den avhengige og den uavhengige variabelen du er interessert i.
  - Dette skaper kan skape en *tilsynelatende* kausal sammenheng mellom eksponeringen og utfallet.

Konfundering (2)



Konfunderende variabler korrelerer altså med både eksponeringen og utfallet.

- Er du interessert i sammenhengen mellom tidlig alkoholdebut og senere alkoholkonsum er sensation seeking er en mulig konfunderende variabel.

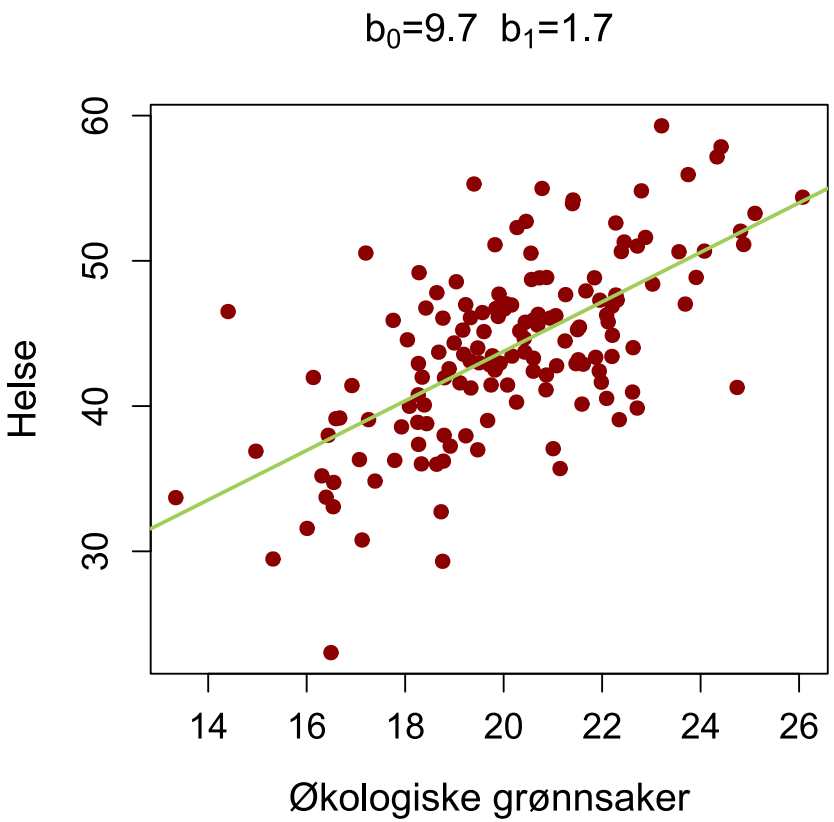
I eksperimentelle studier fjerner vi effekten av konfunderende variabler ved å holde dem *konstante*.

- F.eks. vi inkluder bare deltagere med samme sensation seeking skåre.
- Har alle samme skåre på sensation seeking, kan dette ikke forklare hvorfor folk varierer i alkoholkonsum.

I observasjonelle studier trenger vi også en måte å holde en «konfunderende variabel konstant».

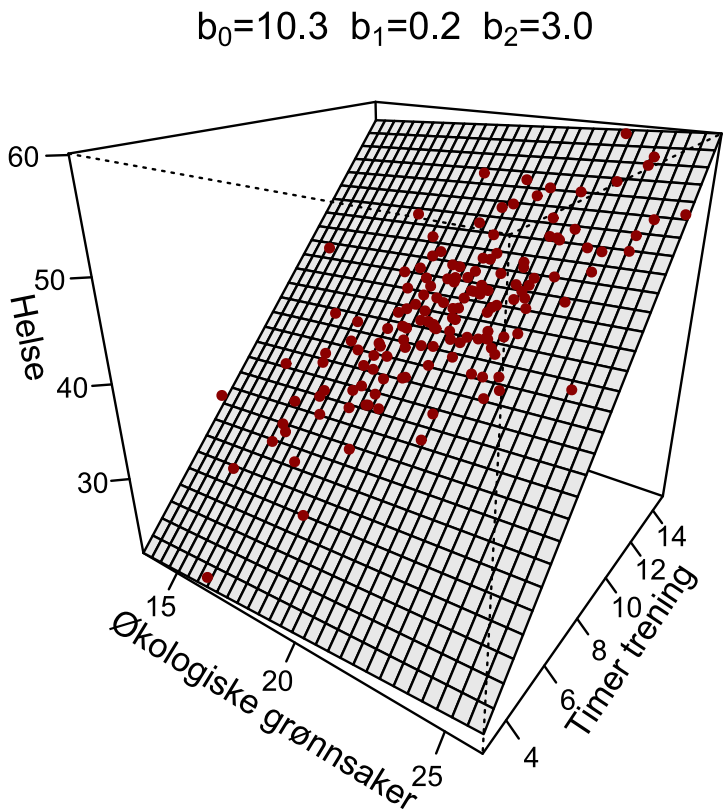
- **Dårlig løsning:** Bare se på effekten innad i et subsett av deltagere som har samme skåre på sensation seeking.
  - Dårlig, fordi vi kaster bort det meste av dataene.
- **Bedre løsning:** Inkluder både sensation seeking og alder for alkoholdebut som uavhengige variabler i regresjonsmodellen.
  - For å se hvorfor dette hjelper oss å holde sensation seeking konstant hjelper det å ha en geometrisk modell for regresjonsplanet.

Tilbake til økologisk mat



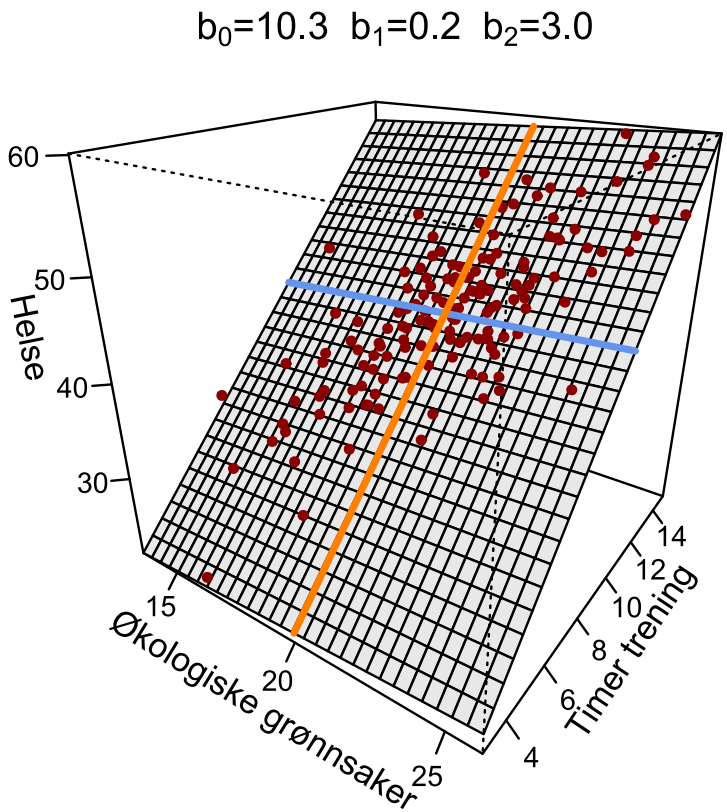
Er det en helsegevinst av å spise økologisk mat? (mat produsert uten pesticider eller syntetisk gjødsel).

Økologisk mat



Folk som spiser mer økologisk mat trener typisk også mer.

- Blant folk som trener like mye, f.eks. 20, er de som spiser mer økologisk mat sunnere?



I

- Oransje: Folk som spiser 20 økologiske grønnsaker og trener fra 4-14 timer.
- Blå: Folk som trener 10 timer i måneden og spiser fra 14-26 økologiske grønnsaker.

begge tilfeller ser vi hva modellen predikerer når den andre andre variabelen *konstant*.

Supressorer

- I alle eksemplene vi har sett så langt blir effekten av en uavhengig variabel mindre når jeg kontrollerer for en konfounder.
    - F.eks. av høyde på matteferdighet forsvinner når jeg kontrollerer for alder.

- Noen ganger er det lite/ingen assosiasjon mellom to variabler *inntil* du kontrollerer for en tredje.
    - Den tredje variabelen kalles da en supressor. Den «suppress» (fortrenger) assosiasjonen mellom de første to.

Avdekking av supressoreffekt

```
mod1←lm(lykke~fritid)
summary(mod1)
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.32577    0.53983  63.586  < 2e-16 ***
## fritid       0.13810    0.05043   2.739  0.00635 **
## ---
```

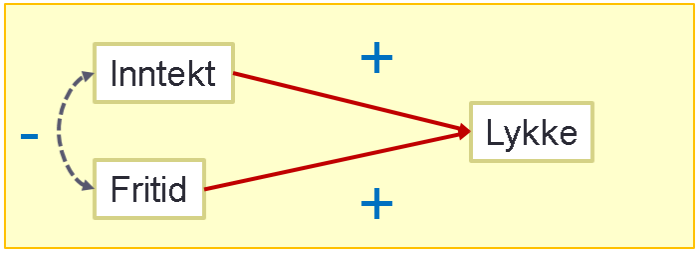
```
mod2←lm(lykke~fritid+inntekt)
summary(mod2)
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.511441   1.663516  11.128  < 2e-16 ***
## fritid       0.341395   0.050987   6.696 4.96e-11 ***
## inntekt      0.027405   0.002749   9.968  < 2e-16 ***
## ---
```

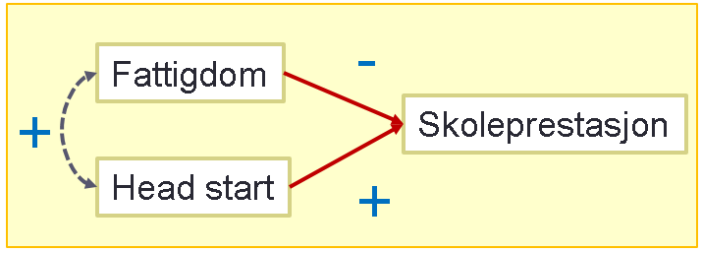
Over: I dette tenkte eksemplet mer enn dobler størrelsen på sammenhengen mellom fritid og lykke seg når man legger til inntekt.

Supressorvariabler (2)

Figurene under viser to måter supressorvariabler (inntekt og fattigdom) kan skapes.



De uavhengige variablene er negativt korrelert, men positivt forbundet med lykke.



De uavhengige variablene er positivt korrelerte, men har motsatt effekt på den avhengige.

Standardiserte regresjonskoeffisienter (A11.7)

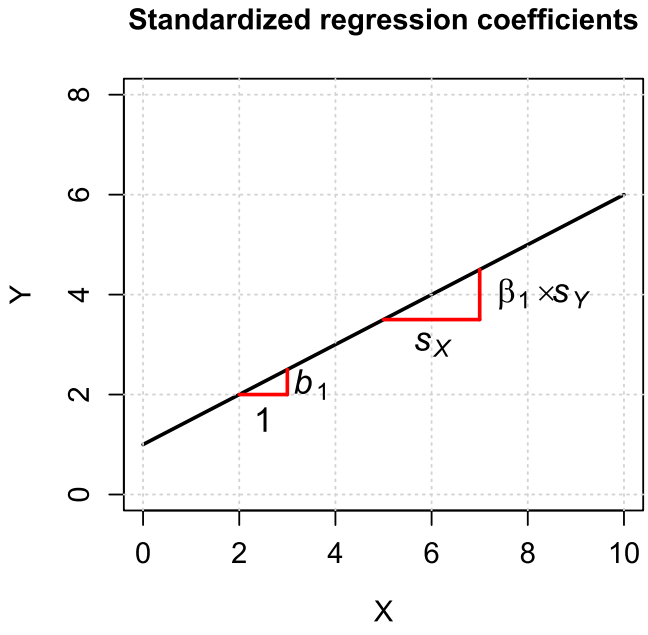
Sammenlikning av stigningskoeffisienter

I modellen under er (høyere) utdanning målt i år, og inntekt målt i 1000 NOK. Hvilken uavhengig variabel er sterkest assosiert med lykke?

```
mod2←lm(lykke~utdanning+inntekt)
summary(mod2)
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.068076   1.267740  18.196  < 2e-16 ***
## utdanning    0.685353   0.074623   9.184  < 2e-16 ***
## inntekt      0.019779   0.002446   8.087 3.43e-15 ***
## ---
```

Sammenlikning av stigningskoeffisienter (2)



Det gir som regel ikke mening å *sammenlikne* de ustandariserte koeffisientene, fordi variablene kan ha veldig ulik skala.

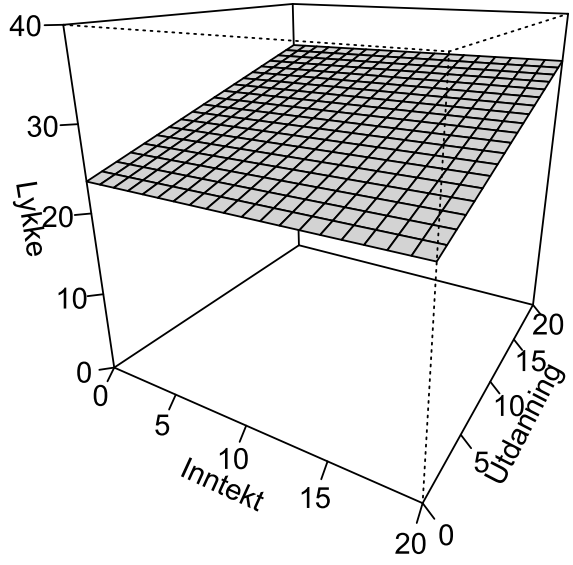
For å kunne sammenlikne regresjonskoeffisienter må vi sette dem på en liknende skala, der "en enhet økning" er sammenliknbare størrelser.

```r
# install.packages("lm.beta")
library(lm.beta)

mod2←lm(lykke~utdanning+inntekt)
summary(lm.beta(mod2))
```

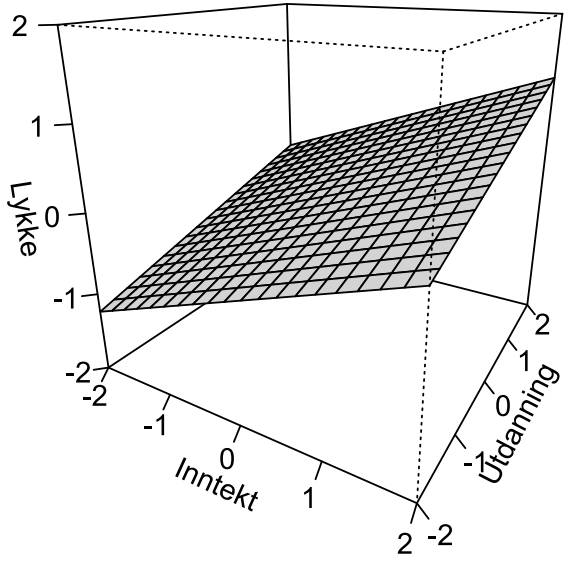
```
## Coefficients:
##              Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept) 23.068076     0.000000   1.267740  18.196  < 2e-16 ***
## utdanning    0.685353     0.335716   0.074623   9.184  < 2e-16 ***
## inntekt      0.019779     0.295600   0.002446   8.087 3.43e-15 ***
## ---
```

```
## Coefficients:
##              Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept) 23.068076     0.000000   1.267740  18.196  < 2e-16 ***
## utdanning    0.685353     0.335716   0.074623   9.184  < 2e-16 ***
## inntekt      0.019779     0.295600   0.002446   8.087 3.43e-15 ***
## ---
```

Ustandardisert enheter



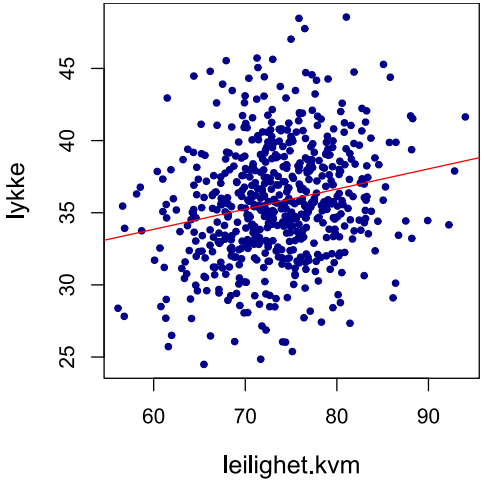
Standardisert enheter



Partiell korrelasjon (A 11.6)

Partiell korrelasjon: korrelasjonen mellom to variabler der den effekten som overlapper med en eller flere andre variabler er fjernet.

- $r_{yx_1 \cdot x_2}$, korrelasjonen mellom y og x_1 kontrollert for x_2.



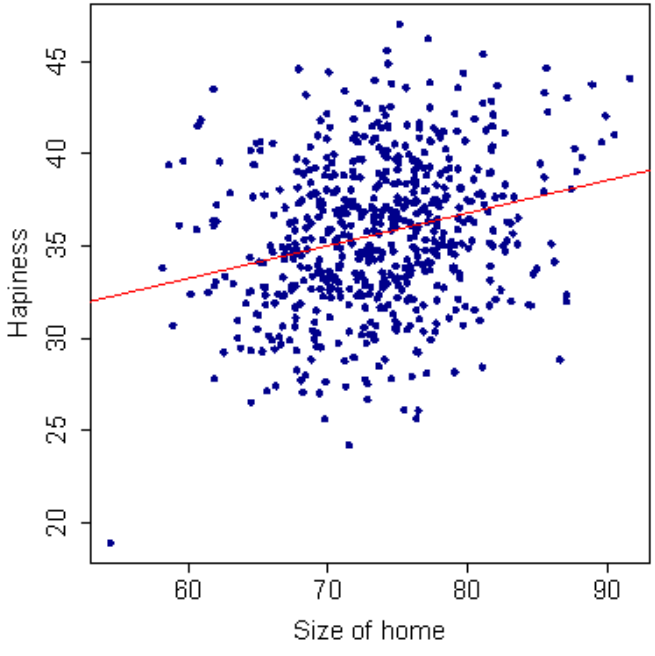

```
cor(leilighet.kvm, lykke)
```

```
## [1] 0.2085761
```

Det er rimelig deler av korrelasjoen mellom lykke og størrelsen på leiligheten skyldes inntekt. Hvordan kan jeg kontrollerer for inntekt?

Partiell korrelasjon: korrelasjonen mellom to variabler der den effekten som overlapper med en eller flere andre variabler er fjernet.

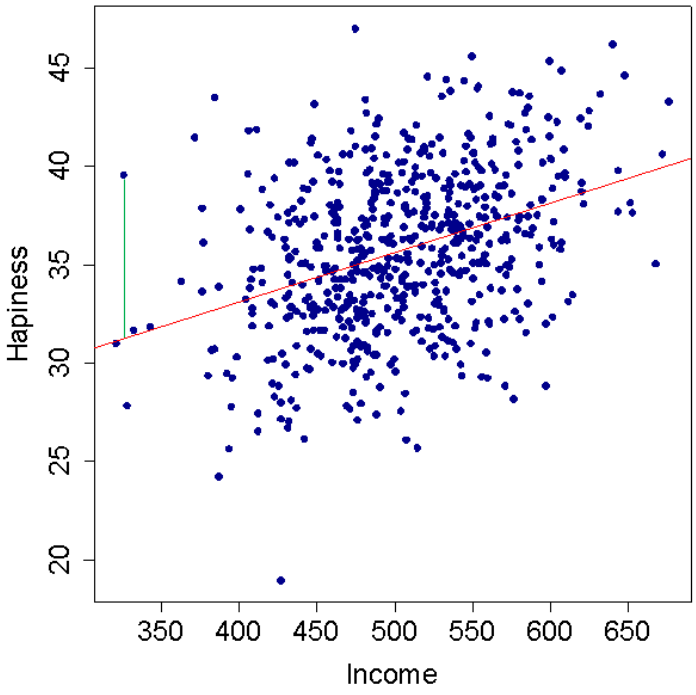
- $r_{yx_1 \cdot x_2}$, korrelasjonen mellom y og x_1 kontrollert for x_2.



Korrelasjonen mellom størrelsen på leiligheten og lykke er r=0.20

Det er rimelig at noe av dette skyldes inntekt. Hvordan kan jeg finne korrelasjonen mellom husstørrelse og lykke, mens jeg kontrollerer for inntekt.

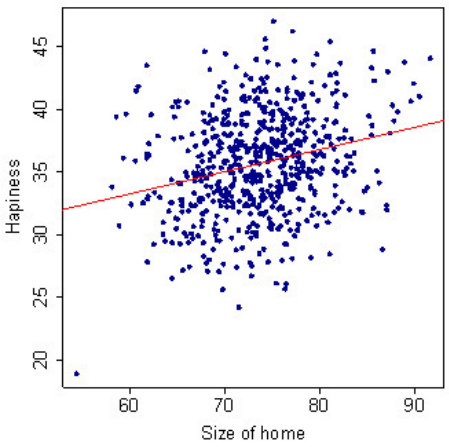
Partiell korrelasjon (2)



Et gitt individ har lykke 40, mens forventet verdi utifra modellen er 31. Følgelig er «lykke residualen» 9, dette er en verdi som per definisjon ikke kan forklares av inntekt.

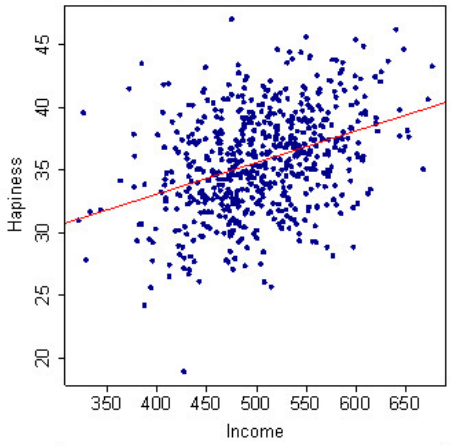
Jeg trekker derfor forventet lykke basert på inntekt fra den observerte lykken for hver enkelt person, og får en ny «verdi lykke for lykke» hvor effekten av inntekt har blitt fjernet.

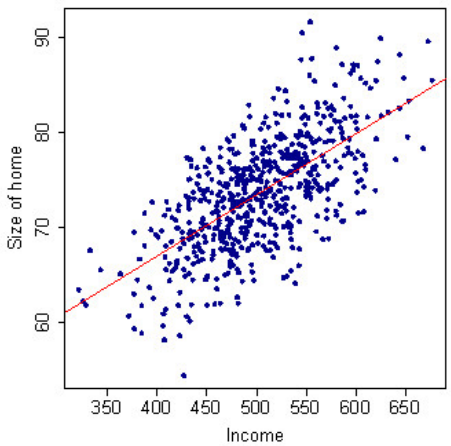
Partiell korrelasjon (3)



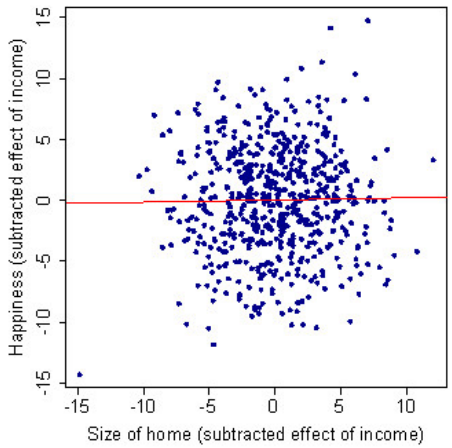
Korrelasjon mellom hus og lykke r=0.20

Få *residual lykke*, kontrollert for inntekt.





Få *residual husstørrelse*, kontrollert for inntekt.



Perason-korrelasjonen mellom de to residualene gir den partielle korrelasjonen mellom leilighetsstørrelse og lykke, kontrollert for inntekt, r = 0.01

Partiell korrelasjon i R

```
cor(leilighet.kvm, lykke)
```

```
## [1] 0.2085761
```

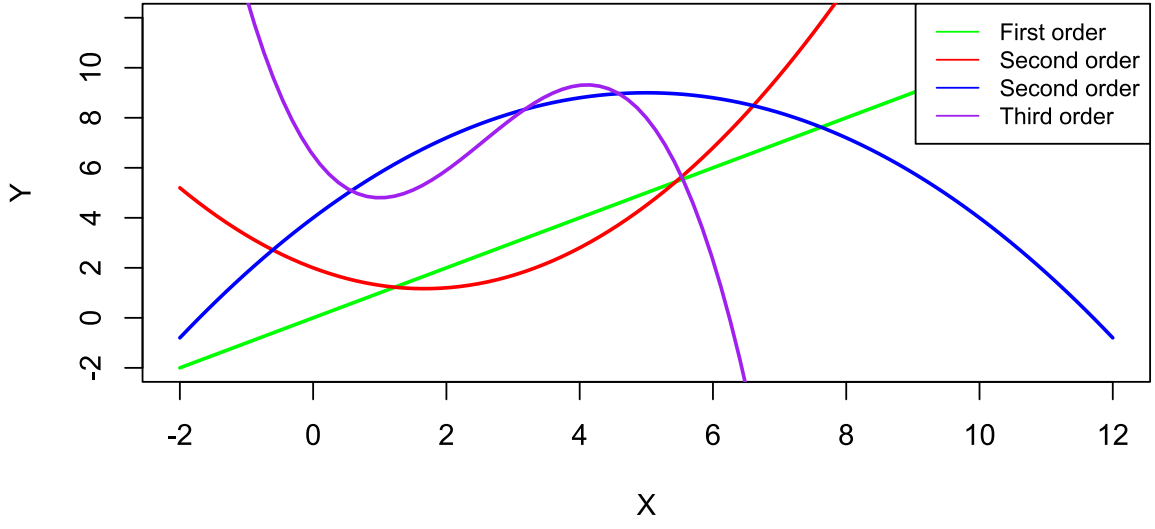
```
# install.packages("ppcor")
library(ppcor)

# pcor.test(VAR1, VAR2, KONTROLLERT_FOR)
pcor.test(leilighet.kvm, lykke, inntekt)
```

```
##     estimate   p.value statistic   n gp  Method
## 1 0.01347995 0.7419738 0.3293935 600  1 pearson
```

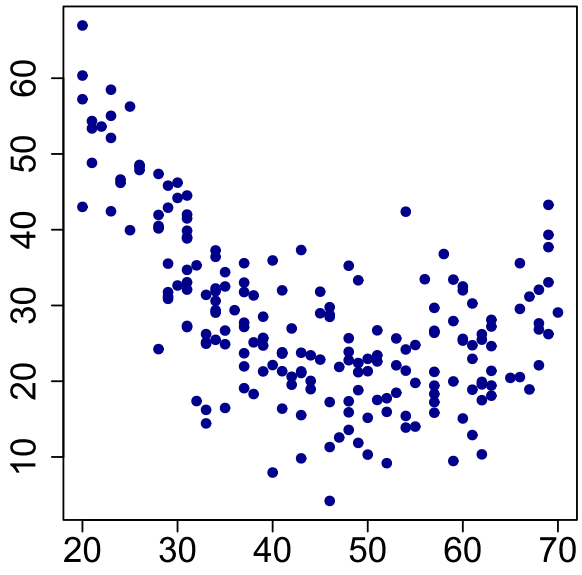
Polynomisk regresjon

- Lineær regresjon er mer fleksibel enn man skulle tro, og kan modellere ikke-lineære relasjoner også.
- En utbredt måte å regresjon til å se på ikke lineære sammenhenger er ved å modellere en *polynomisk regresjonsfunksjon*.

$$E(Y|X) = b_0 + b_1 \cdot X + b_2 \cdot X^2 + \ldots + b_p \cdot X^p$$



Alder og lykke



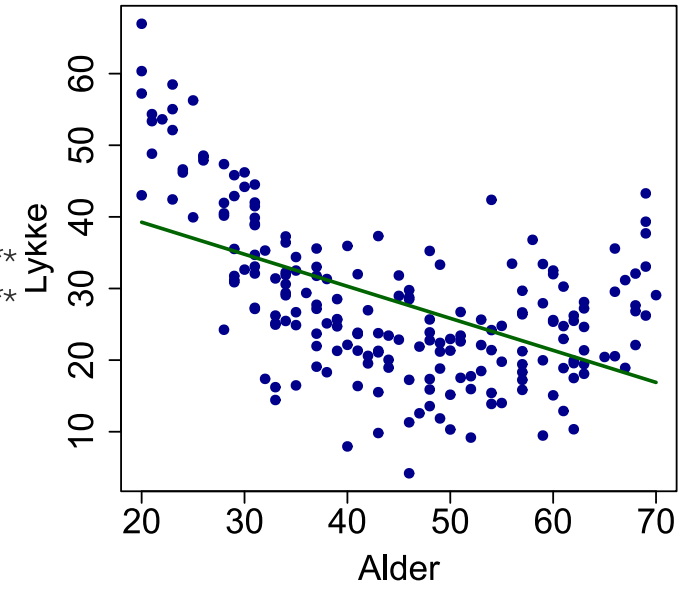
Lykke	Alder	Alder_sq
15.95492	52	2704
27.64124	68	4624
46.19217	24	576
33.46233	56	3136
21.95981	37	1369
32.09236	31	961
28.49370	46	2116
46.21149	30	900

Over er et tenkt forhold mellom lykke og alder.

En lineær modell (polynom av grad 1)

```
summary(lm(Lykke~Alder))
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 48.19619    2.30885  20.875   <2e-16 ***
## Alder       -0.44727    0.04964  -9.011   <2e-16 ***
```



En modell med polynom av grad 1 er forventet lykke gitt ved

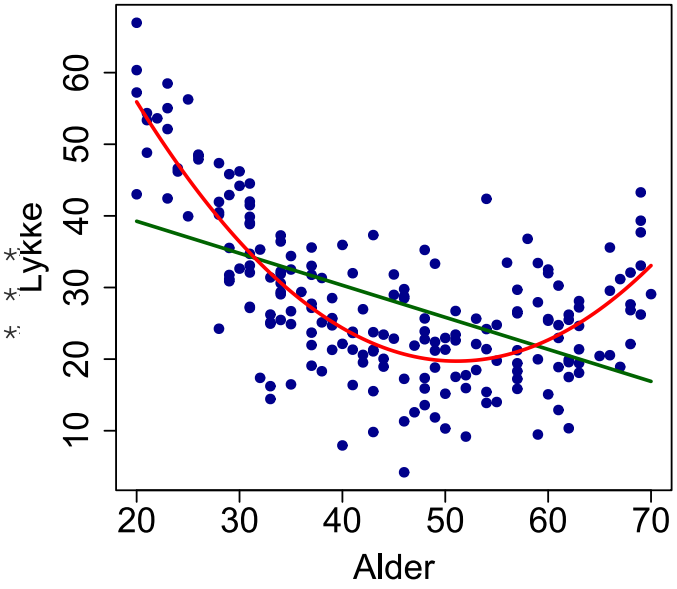
$$E(Lykke) = b_0 + b1 \cdot Alder$$

$$E(Lykke) = 48.20 - 0.45 \cdot Alder$$

En kvadratisk modell (polynom av grad 2)

```
summary(lm(Lykke~Alder+I(Alder^2)))
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 117.441728   5.318759   22.08   <2e-16 *
## Alder        -3.824112   0.249037  -15.36   <2e-16 *
## I(Alder^2)    0.037411   0.002731   13.70   <2e-16 *
```



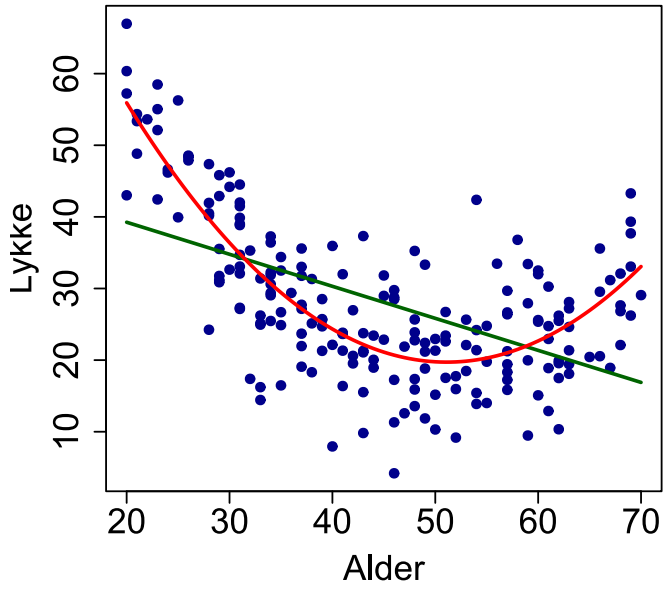
En modell med polynom av grad 2 er forventet lykke gitt ved

$$E(Lykke) = b_0 + b1 \cdot Alder + b2 \cdot Alder^2$$

$$E(Lykke) = 117.44 - 3.82 \cdot Alder + 0.04 \cdot Alder^2$$

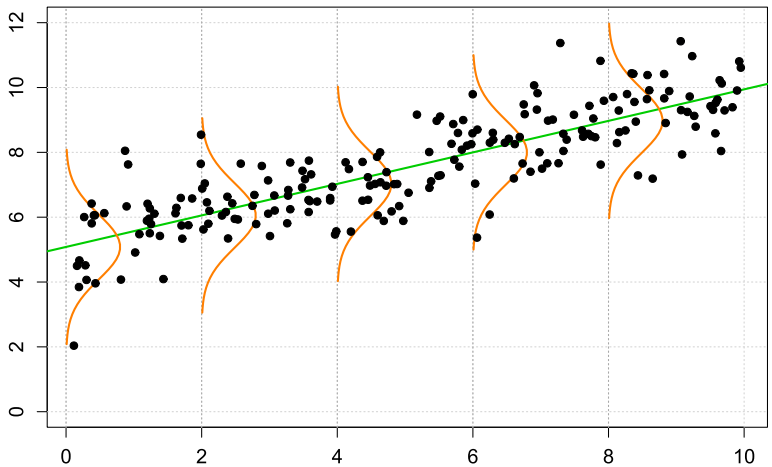
Obs ved polynomiske modeller

- Vær kritisk til modeller med høy polynomer, de kan utgjøre en overtilpasning til dataene.
- Modellen kan ha begrenset gyldighet utenfor rangen der vi har observerte verdier av den uavhengige variabelen (unngå *ekstrapolering*).
- Koeffisientene har ikke lenger tolkningen *partielle stigningsgrader*. Det er ikke mulig å se på alder, men "holde alder^2 konstant".



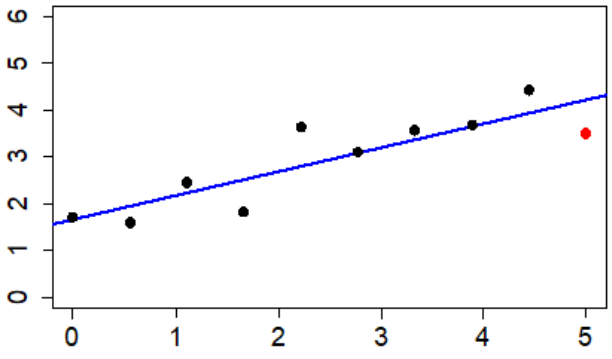
Innflytelse

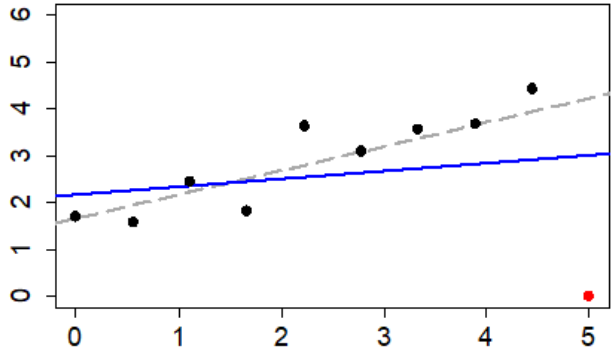
$$Y_i = b_0 + b_1 \cdot X_i + \epsilon_i \qquad \epsilon_i \; N(0, \sigma^2)$$



1. Forholdet mellom X og Y er lineært.
2. Feilvariansen (spredningen rundt regresjonslinjen) er normalfordelt, og har samme varians for alle nivåer av X.
3. **Det er ingen ekstreme uteliggere.**
4. Observasjonene er uavhengige.
5. Den uavhengige variabelen er målt uten feil.

Uteliggere er av betydning

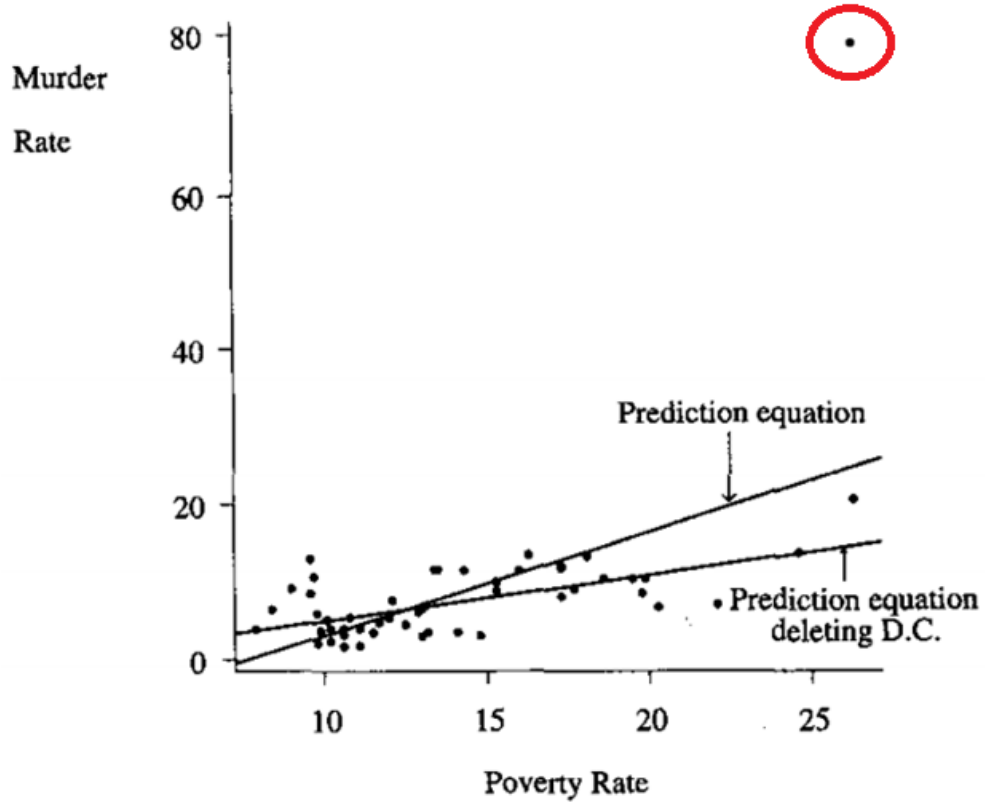




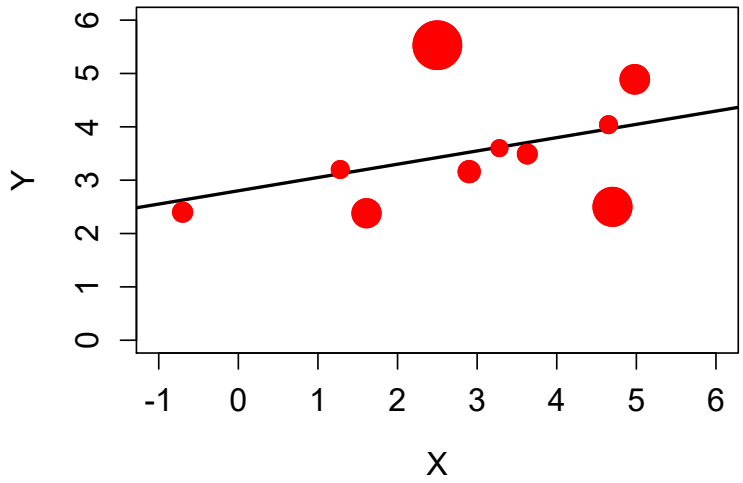
- En **uteligger** er en observasjon som ligger lang i verdi fra de øvrige. uavhengige variabler.
  - F.eks. Det er ikke uvanlig å være 150cm eller veie 120 kg., men individer som både er 150cm høy og veier 110 kg er sjeldne.
  - Uteliggere blir identifisert ved ulike distansemål.

- Det vanligste målet på avstand fra modellen er *residualet*, $Y_i - \hat{Y}_i$.
- Dette er nyttig for å identifisere uteliggere på den *avhengige* variabelen.

Et reelt eksempel fra Agresti (2009)

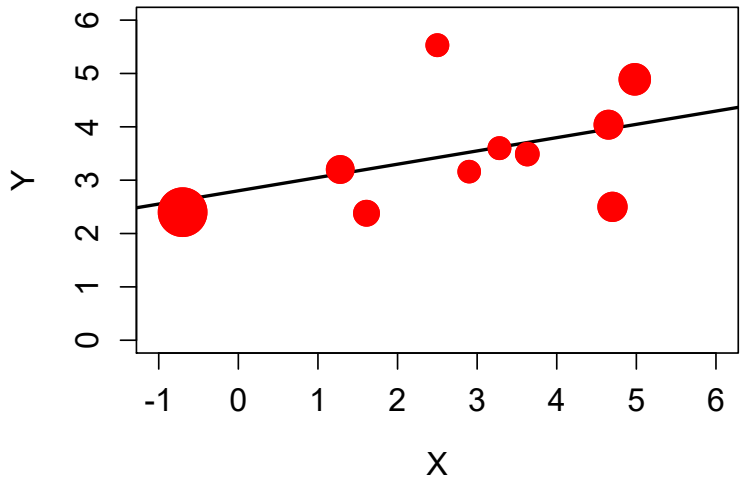


Residuals



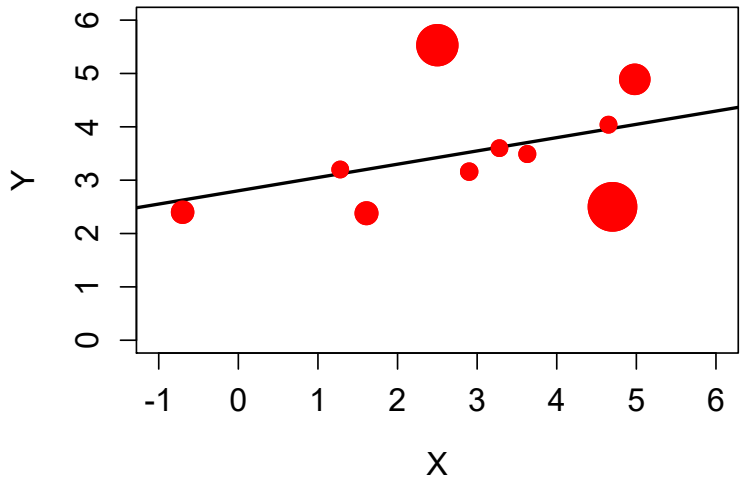
- Det vanligste målet på distanse (avvik) er *residualet*, $Y_i - \hat{Y}_i$.
- Størrelsen på residualene hjelper oss å avdekke uteliggere på den *avhengige variabelen*

Leverage



- En statistikk hjelper til med å identifisere observasjoner som har verdier langt fra gjennomsittet på den uavhengige variabelen.
  - Nyttig for å finne potensielle uteliggere blant de uavhengige variablene.
  - Observasjoner med høy leverage trekker regresjonslinjen sterkt mot seg.

Influens



- Mål på influens kombinerer avstand og leverage for å identifisere særlig innflytelsesrike observasjoner.
    - Det vanligste målet på innflytelse er Cook's D, som kvantifiserer grad av endring du ville sett i stigningstallet til linjen dersom en observasjon var slettet. Observasjoner med verdier > 1 bør granskes.

- Andre vanlige influens statistikker; **DFFIT** and **DFBETA**.

Neste uke

Statistisk inferens i regresjon

- Teststatistikker og samplingfordelinger
- Standardfeil
- t-fordelinger og F-fordelinger
- P-verdier
  - Signifikans av individuelle prediktorer
  - Signifikans i forskjeller mellom modeller
- Konfidensintervaller