PSY2014 – Kvantitativ metode

Forelesning 2: Bivariat 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.

Idag

- Bivariat regresjon
  - Linjer i planet
  - Målefeil
  - Tolkning av ustandardiserte og standardiserte koeffesienter
  - Minste kvadraters estimat
  - Kvadratsummer
  - Vurdering av underliggende antagelser

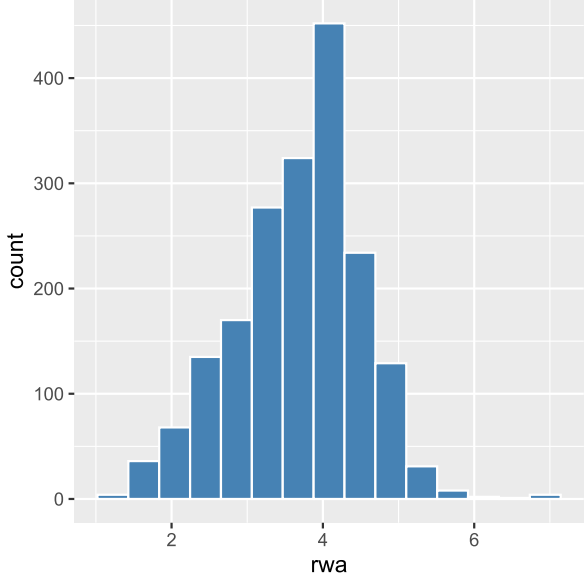
Eksempel: RWA og Åpenhet til erfaringer

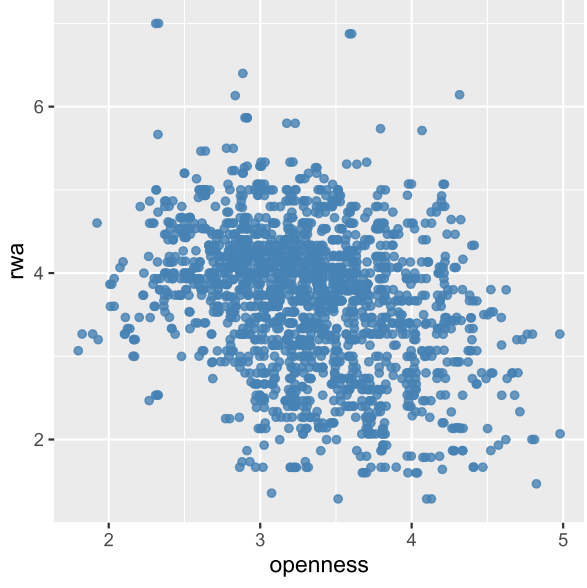
- **Right-Wing Authoritarianism (RWA)** skalaen var utviklet av Altemeyer i 1981.
- Folk med høye skårer på RWA underkaster seg i høy grad autoriteter de oppfatter som etablerte og legitime. De holder seg til samfunnskonvensjoner og normer, og er fiendtlige og straffende mot de som ikke holder seg til dem. De verdsetter enhetlighet og er for å bruke gruppeautoritet, inkludert tvang, for å oppnå den.
- Vi skal se på RWA skårer i et utvalg av 1,895 individer (803 menn, 1092 kvinner) hvor snittalderen er 63.1 år.
- Åpenhet til nye erfaringer (Openness) er en av de fem domenene i femfaktormodellen for personlighet.

RWA PLOTS

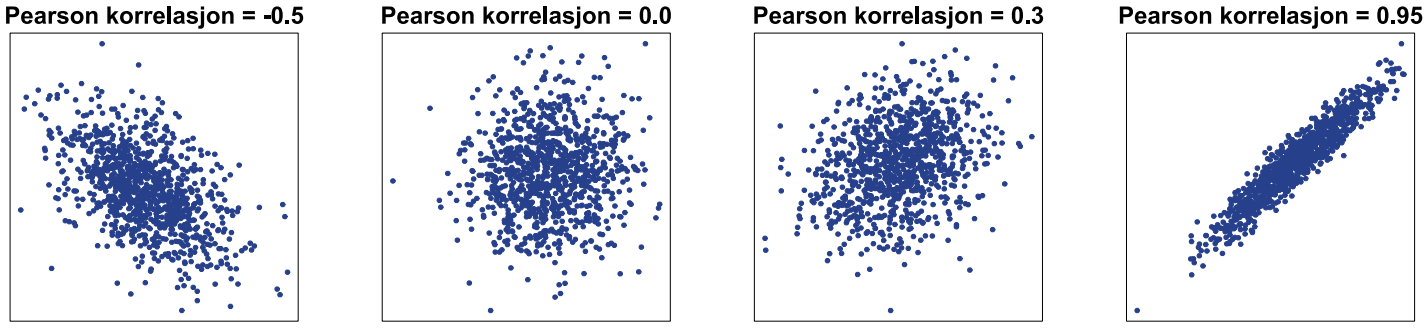
```
# Histogram of rwa
ggplot(dat, aes(x=rwa))+
  geom_histogram(color="white", fill="steelblue", bins=15)

# Scatter plott of rwa vs openness
ggplot(dat, aes(x=Openness, y=rwa))+
  geom_point(color="steelblue", fill="steelblue", position="jitter")
```





Pearson korrelasjon



$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Pearson-korrelasjon:

- utgjør en *standardisering* av kovariansen.
- er definert som kovariansen delt på den største verdien denne kovariansen kan ta (produktet av variablenes standardavvik).
- er den vanligste statistikken brukt for å bedømme styrken av forholdet mellom to variabler.

Korrelasjon mellom RWA og Openness

```
sd(dat$rwa)
```

```
## [1] 0.8433743
```

```
sd(dat$openness)
```

```
## [1] 0.5273023
```

```
cov(dat$rwa,dat$openness)
```

```
## [1] -0.1346597
```

Hva er korrelasjonen?

1. ca. -0.5
2. ca. -0.3
3. ca. 0.2
4. ca. 0.4

Hva er andel forklart varians?

1. ca. 0.01
2. ca. 0.09
3. ca. 0.15
4. ca. 0.30

Korrelasjon mellom RWA og Openness

```
R←cor(dat$rwa,dat$openness)
R %>% round(2)
```

```
## [1] -0.3
```

```
R^2 %>%  round(2)
```

```
## [1] 0.09
```

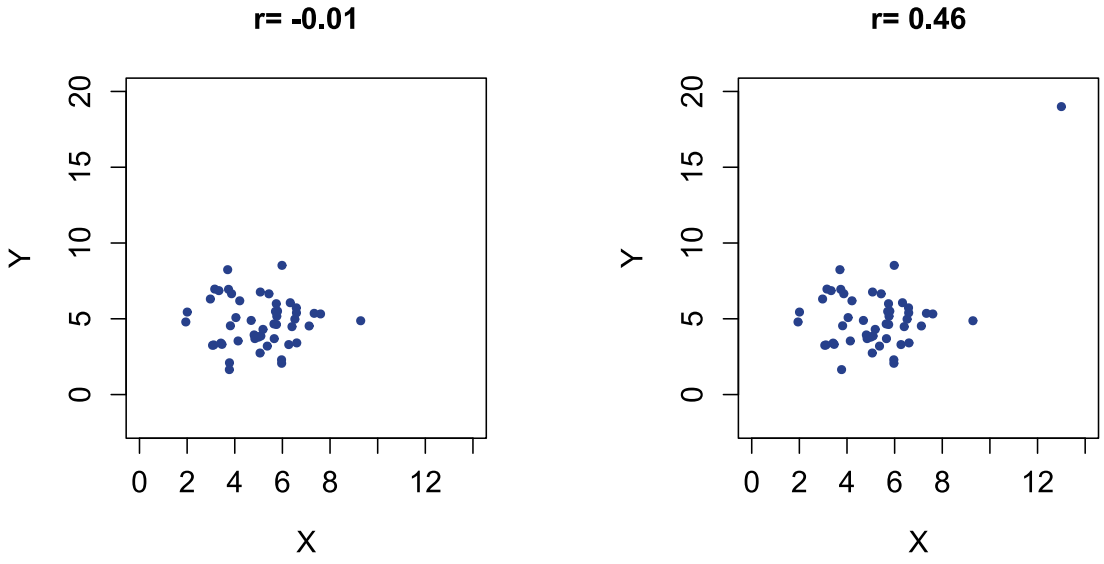
Korrelasjon er lettere å tolke enn kovarians:

$$r_{rwa,o} = \frac{s_{rwa,o}}{(s_{rwa})(s_o)} = \frac{-0.135}{(0.843)(0.527)} \approx -0.30$$

Kvadrering av korrelasjonen gir andel forklart varians:

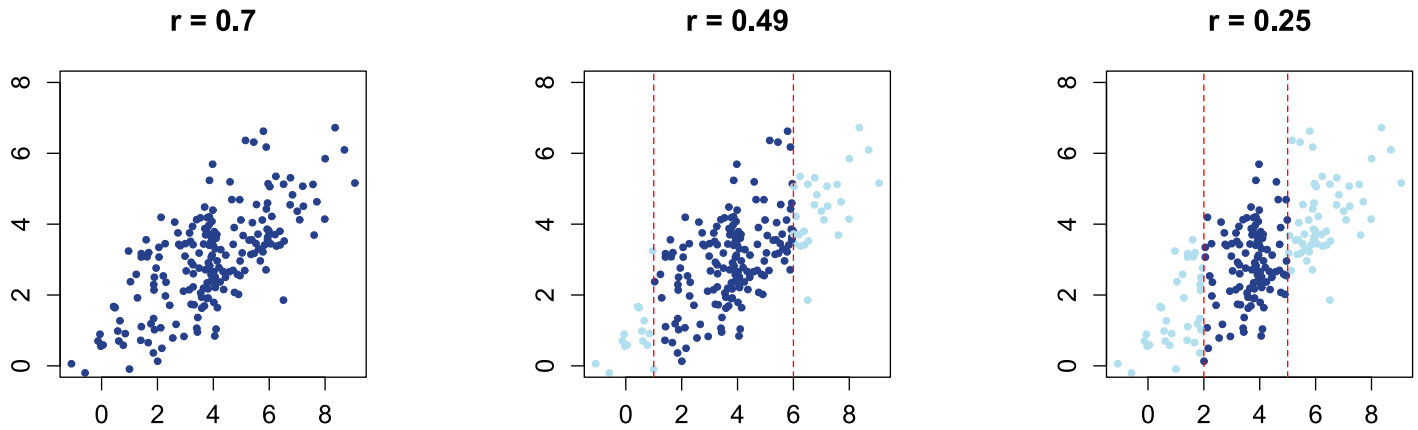
$$r^2 = -0.30^2 = 0.09$$

1. Korrelasjoner er ikke robobuste



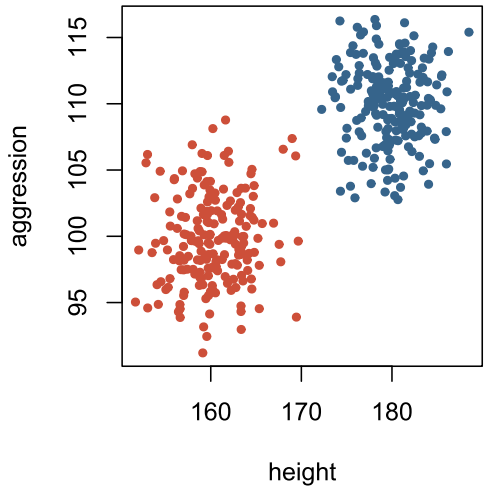
Uteliggere kan ha en veldig stor effekt på korrelasjoner.

2. Korrelasjoner er påvirket av range

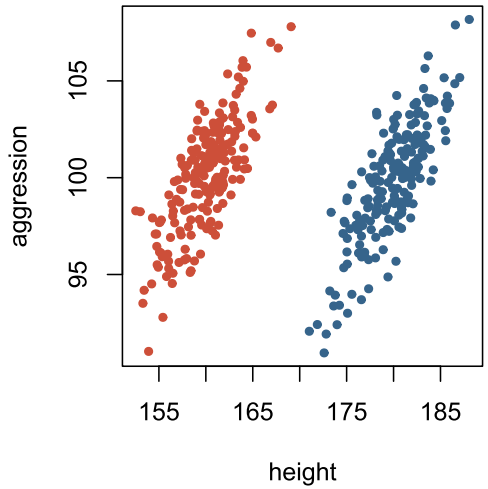


Typisk synker korrelasjonen ettersom rangen til variablene minker.

3. Bakenforliggende variabler kan påvirke korrelasjonen

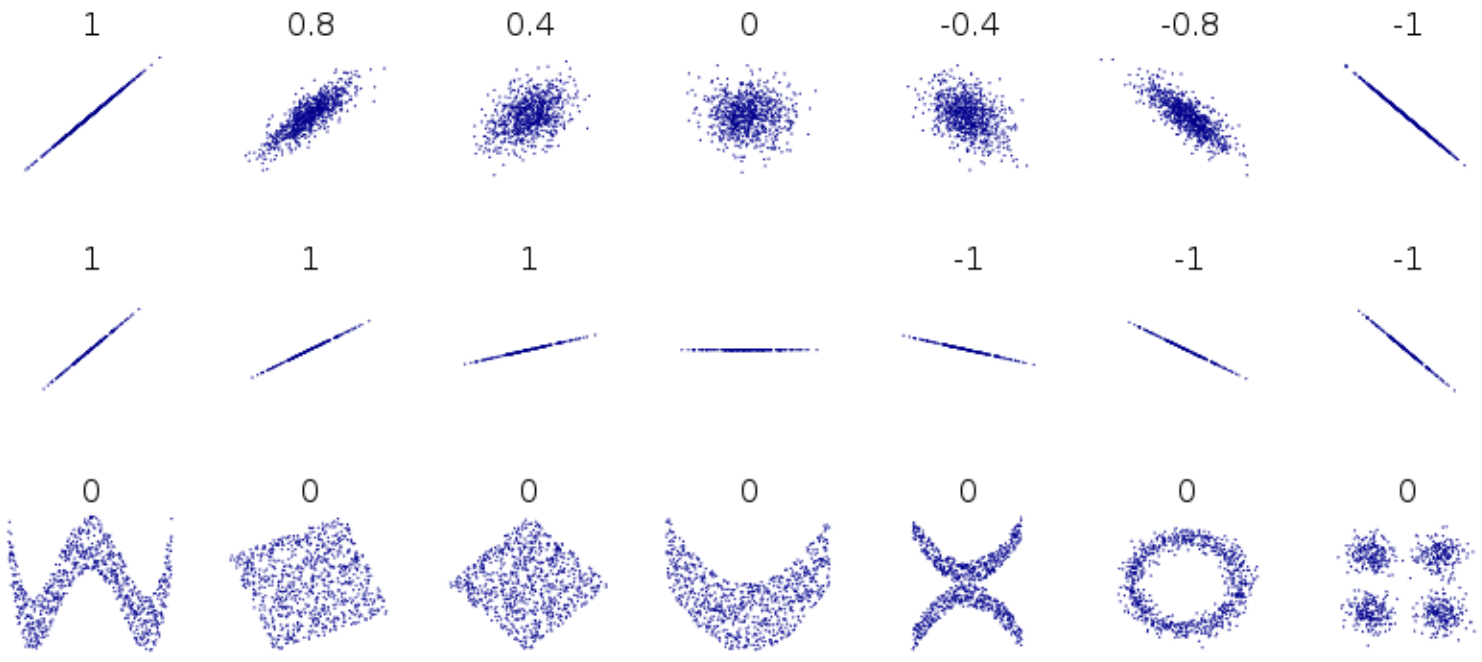


r (Alle) = 0.81
r (Menn)=0.01
r (Kvinner)=0.10



r (Alle) = 0.19
r (Menn)=0.80
r (Kvinner)=0.76

4. Korrelasjon og linearitet



Perason korrelasjoner er kun et mål på grad av «lineært» forhold mellom variablene.

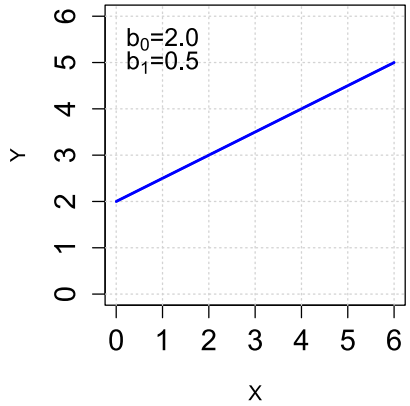
[1] See Wikipedia .

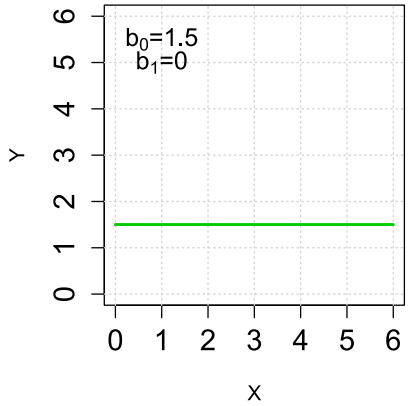
Oppsummerende om Pearson korrelasjon

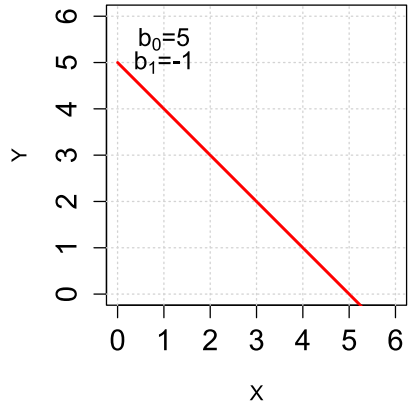
- Kvantifiserer grad av *lineær* assosiasjon mellom to variabler.
- Tar verdier mellom -1 og 1.
- Er upåvirket av lineær transformasjon av variabler.
    - F.eks. deling og multiplisering
- Krever at begge variablene er kontinuerlige (ikke kategoriske).

Bivariat regresjon

Rette linjer







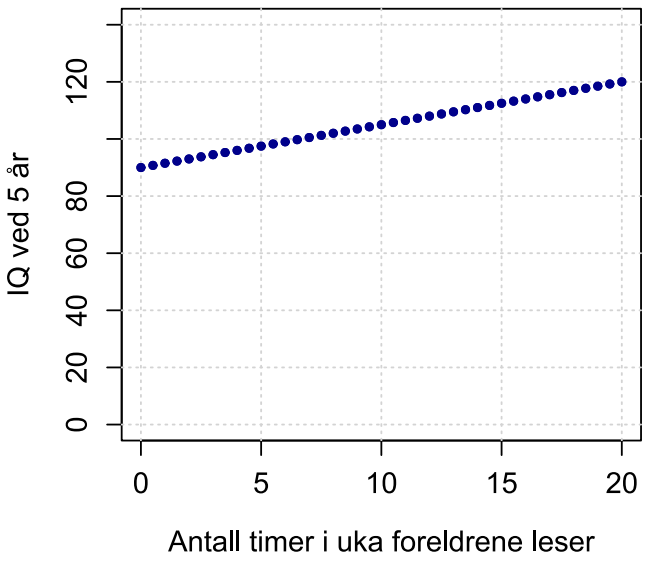
$$Y = b_0 + b_1 X$$

En rett linje er unikt definert av to parametere:

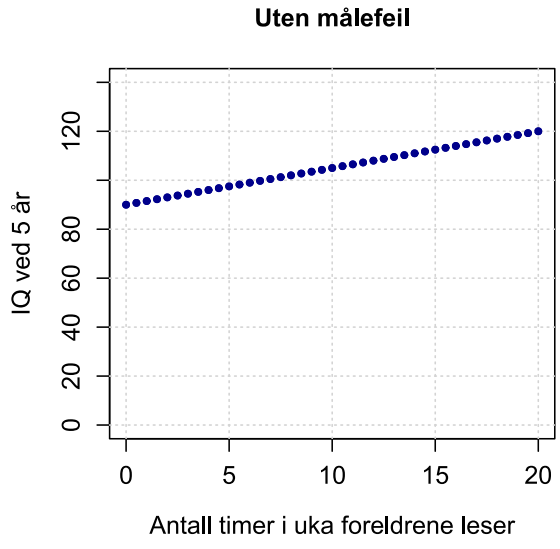
- b_0: konstantleddet / skjæringstallet (intercept)
  - Y-verdien når X er lik 0.
- b_1: stigningstallet (slope)
  - Hvor mye endrer Y verdien seg når X øker med en enhet.

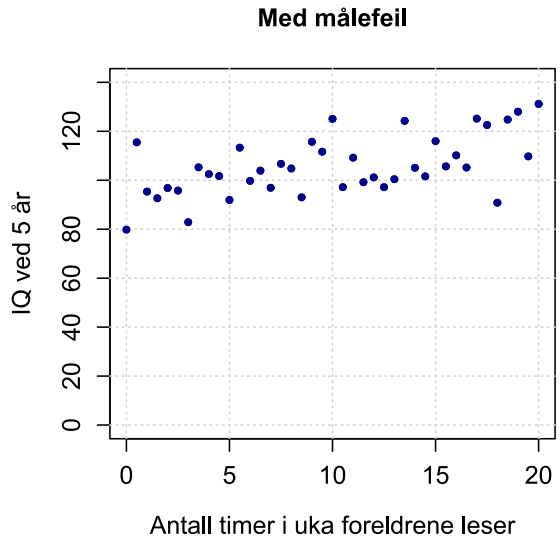
Lesing og IQ

$$IQ_i = b_0 + b_1 \cdot LESING_i$$



- Hva er b_0?
- Hva er b_1?
- Hva ville b_1 vært om det ikke var et forhold?
- Er det realistisk at alle punktene faller nøyaktig lang en rett linje?
- Blir du smartere av å bli lest til, ifølge dette?





$$IQ_i = 90 + 1.5 \cdot LESING_i$$

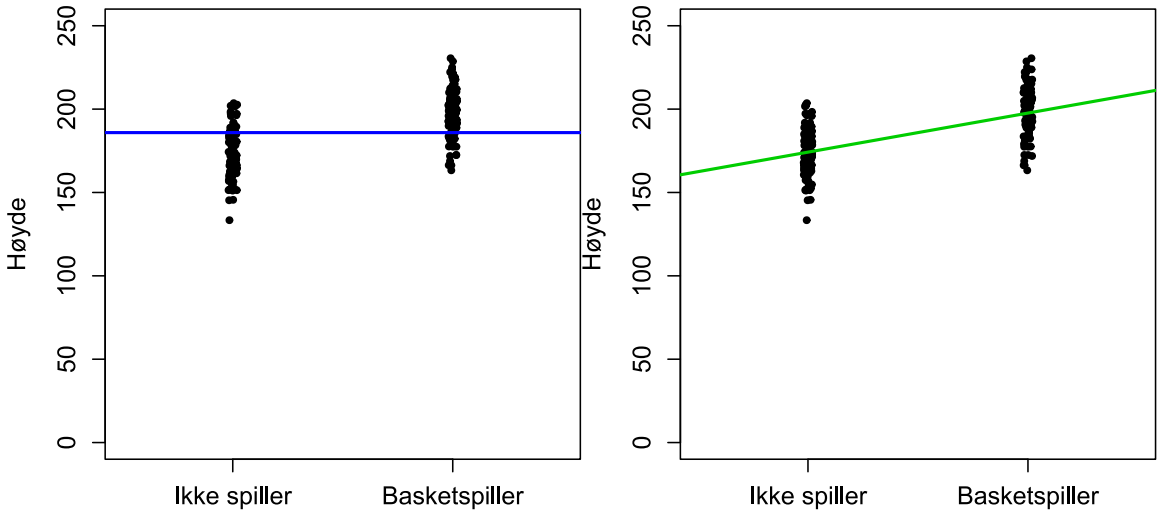
$$IQ_i = 90 + 1.5 \cdot LESING_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

Vår modell av IQ består av en *systematisk og lineær* sammengeng med lesing, og en stokastisk (tilfeldig) komponent.

Tankeeksperiment

Jeg tenker på noen, hva er høyden på denne personen?



- En **regresjonsfunksjon** beskriver hvordan gjennomsnittet til en avhengig variabel følger verdiene til en uavhengig variabel.

$$E(Y|X) = b_0 + b_1 X$$

$E(Y|X)$ leses "Forventet verdi av Y *gitt* den observerte verdien av X"

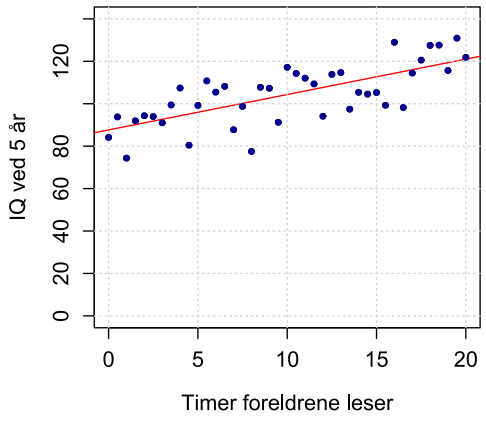
I **bivariat regresjon** har vi en avhengig variabel (Y) og en uavhengig (X).

Vi tenker at prosessen som ligger bak de observerte dataene er :

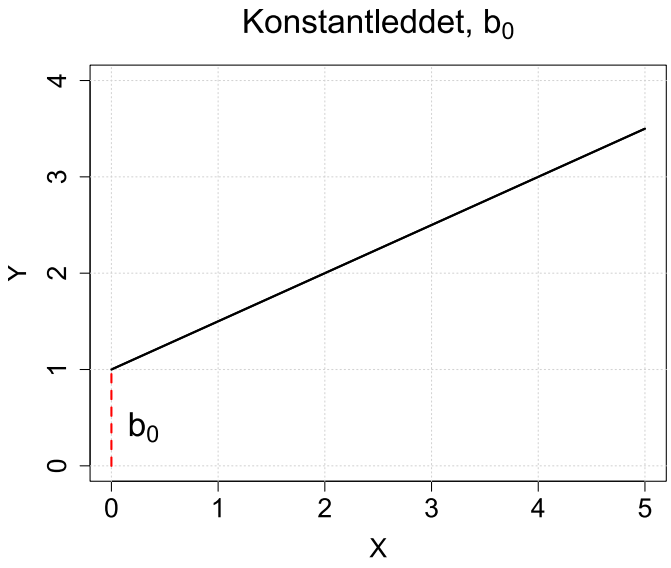
$$Y_i = b_0 + b_1 \cdot X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

Den *forventede* Y verdien gitt en X verdi basert på de estimerte regresjonskoeffisientene er gitt ved:

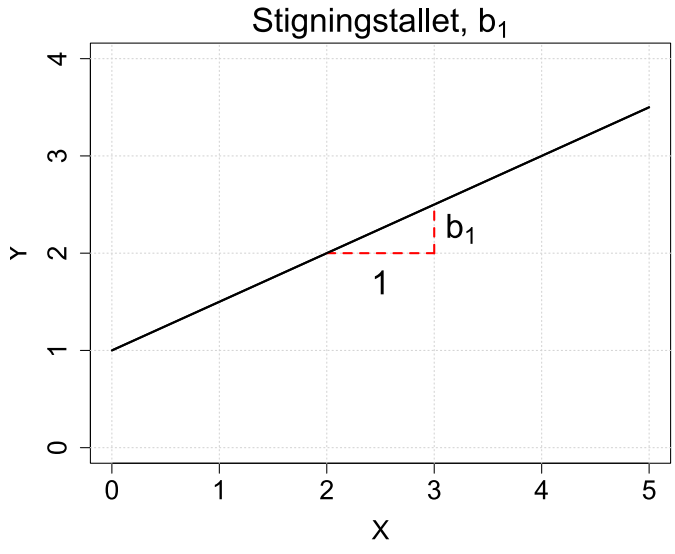
$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 \cdot X_i$$



- b_0 er den "sanne" verdien i populasjonen, mens den estimerte verdien basert på et utvalg indikeres med en "hatt" $\hat{b}_0$.
- I uttrykket til den forventede verdien til en gitt person inngår ikke ϵ_i, fordi den forventede verdien er 0.



Konstantleddet: Verdien til $\hat{Y}$ når X er 0.

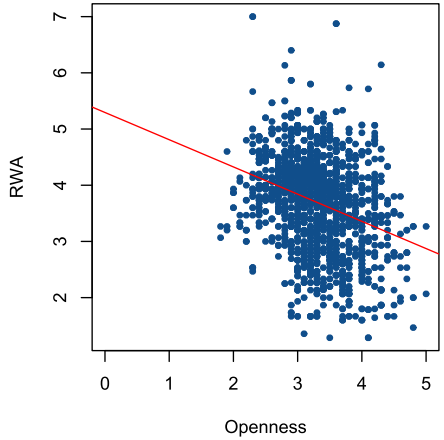


Stigningstallet: Endringen i Y når X endrer seg med 1 enhet.

Bivariat regresjon i R

```
m1 ← lm (rwa~openness, data=dat)
summary(m1)
```

```
##
## Call:
## lm(formula = rwa ~ openness, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4358 -0.5329  0.0857  0.4982  3.3242
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.29431    0.11809   44.83   <2e-16 ***
## openness    -0.48430    0.03522  -13.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.804 on 1873 degrees of freedom
## Multiple R-squared:  0.09169,    Adjusted R-squared:  0.0912
## F-statistic: 189.1 on 1 and 1873 DF,  p-value: < 2.2e-16
```



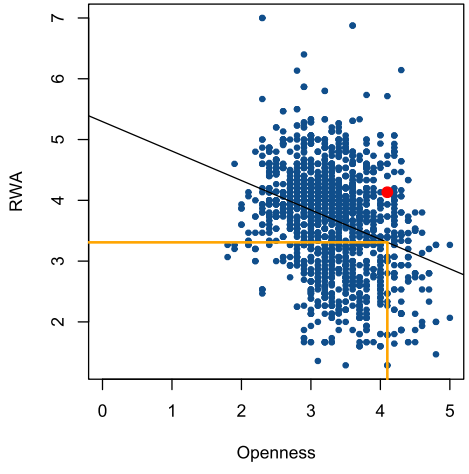
Med de estimerte regresjonskoeffisientene kan du sette opp uttrykket for den *predikerte* skåren for den uavhengige variabelen (prediction equation).

$$\begin{aligned}\hat{Y}_i &= \hat{b}_0 + \hat{b}_1 \cdot X_i \\ &= 5.29 + (-0.48) \cdot X_i\end{aligned}$$

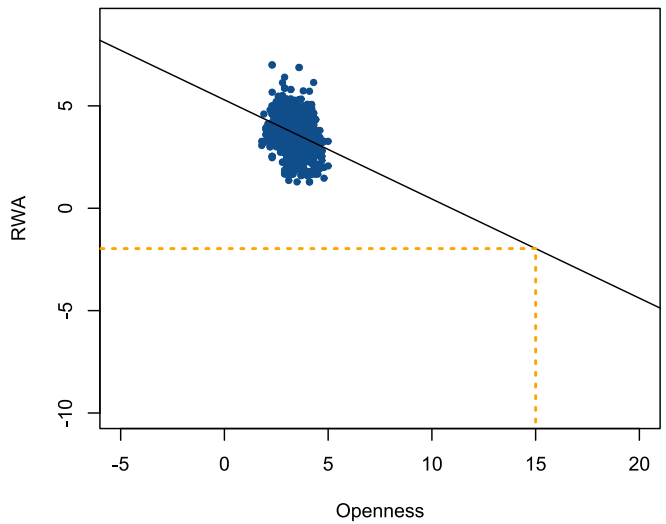
Openness	RWA	Fitted	Residual
2.5	3.80	4.08	-0.28
4.1	4.13	3.31	0.82
3.0	4.93	3.84	1.09
3.3	4.13	3.70	0.44
3.6	4.40	3.55	0.85
3.1	4.27	3.79	0.47
3.2	2.27	3.74	-1.48
2.3	3.93	4.18	-0.25

Hva er forventet RWA for en person med Openness skåre på 4.1?

$$\hat{Y}_i = 5.29 + (-0.48) \cdot 4.1 = 3.322$$



Ekstrapolering



Ekstrapolering: Prediksjon utenfor rangen av den uavhengige variabelen.

Husk at modellen bare er gyldig i et begrenset interall, vi kan f.eks. ikke bruke den til å predikere forventet RWA skåre for noen med Openness på 15.

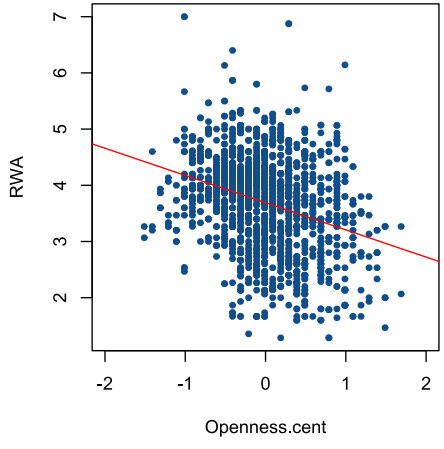
Sentrering av uavhengige variabler

Sentrering av uavhengige variabler innebærer å trekke en konstant verdi fra skårene, ofte gjennomsnittet. Dette endrer skalaen, men påvirker ikke enhetene eller stigningstallet.

- Sentrering kan blandt annet gjøre b_0 mer tolkbar.

```
dat ← dat %>% mutate(openness_c = openness-mean(openness))
m2 ← lm(rwa~openness_c, data=dat)
summary(m2)
```

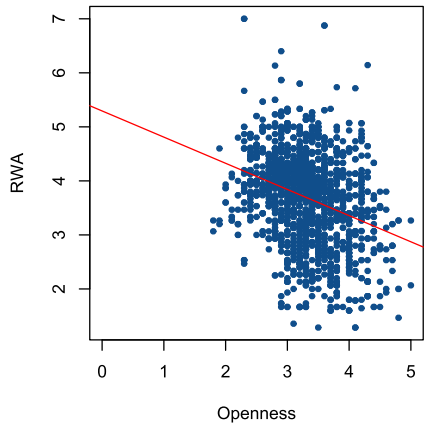
```
...
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.69076    0.01857  198.78   <2e-16 ***
## openness_c  -0.48430    0.03522  -13.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.804 on 1873 degrees of freedom
## Multiple R-squared:  0.09169,    Adjusted R-squared:  0.0912
## F-statistic: 189.1 on 1 and 1873 DF,  p-value: < 2.2e-16
...
```



Standardiserte regresjonskoeffisienter

Logikken bak standardiserte koeffisienter

- $b_0 = 5.29$
- $b_1 = -0.48$
- $sd(rwa) = 0.843$
- $sd(openness) = 0.527$



Hvor mye endrer RWA seg når Openness øker 1 enhet?

$$b_1 \cdot 1 = -0.48 \cdot 1 = -0.48$$

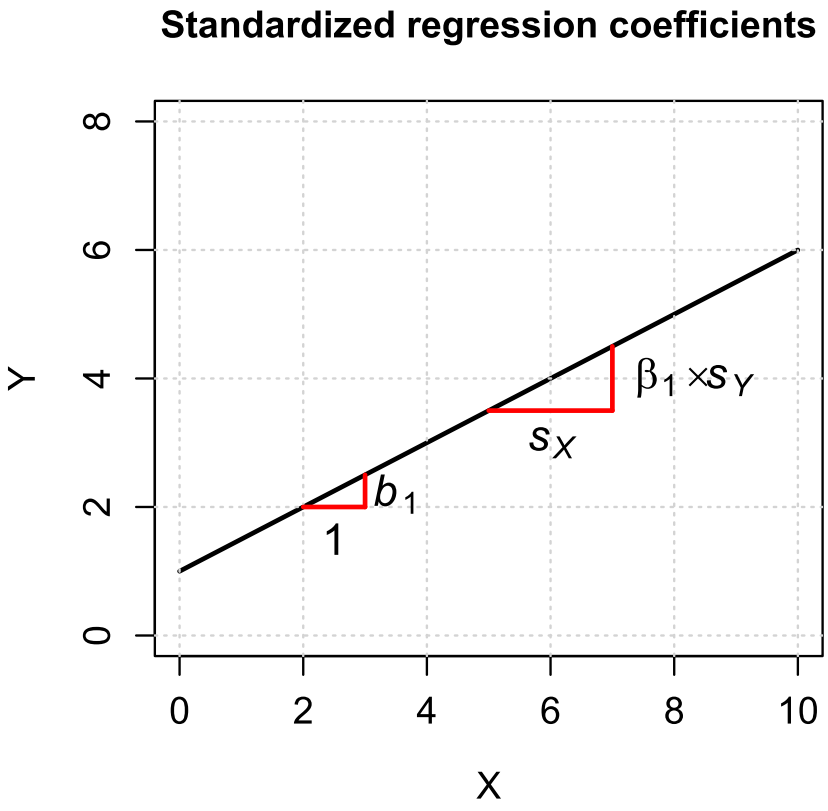
Hvor mye endrer RWA seg når Openness øker 0.527 enheter (ett sd)?

$$b_1 \cdot 0.527 = -0.48 \cdot 0.527 \approx -0.253$$

Hvor mange standardavvik i RWA tilsvarer dette?

$$\frac{-0.253}{0.843} = -0.30 = \beta_1$$

Standardiserte vs ustandardiserte koeffisienter (A:9.4)



$$\beta_1 = b_1 \cdot \frac{s_x}{s_y}$$

b_1: Antall enheter endring i Y per enhet økning av X.

β_1: Antall standardavvik endring i Y når X øker tilsvarende ett standardavvik.

Hvordan få standardiserte koeffisienter fra R

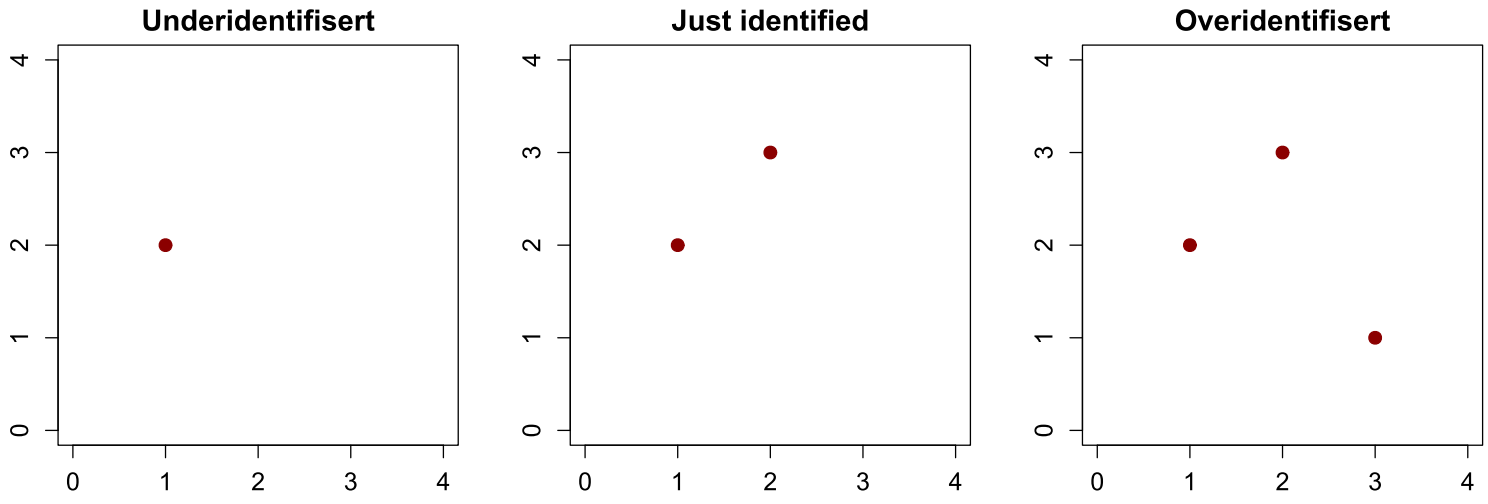
- I motsetning til SPSS skriver ikke R ut standardiserte regresjonskoeffisienter som standard.
- En rekke pakker tilbyr en utregning av β, f.eks. `lm.beta`.

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

m3 ← lm (rwa~openness, data=dat)
lm.beta(m3)
```

```
##
## Call:
## lm(formula = rwa ~ openness, data = dat)
##
## Standardized Coefficients::
## (Intercept)    openness
##   0.0000000  -0.3028011
```

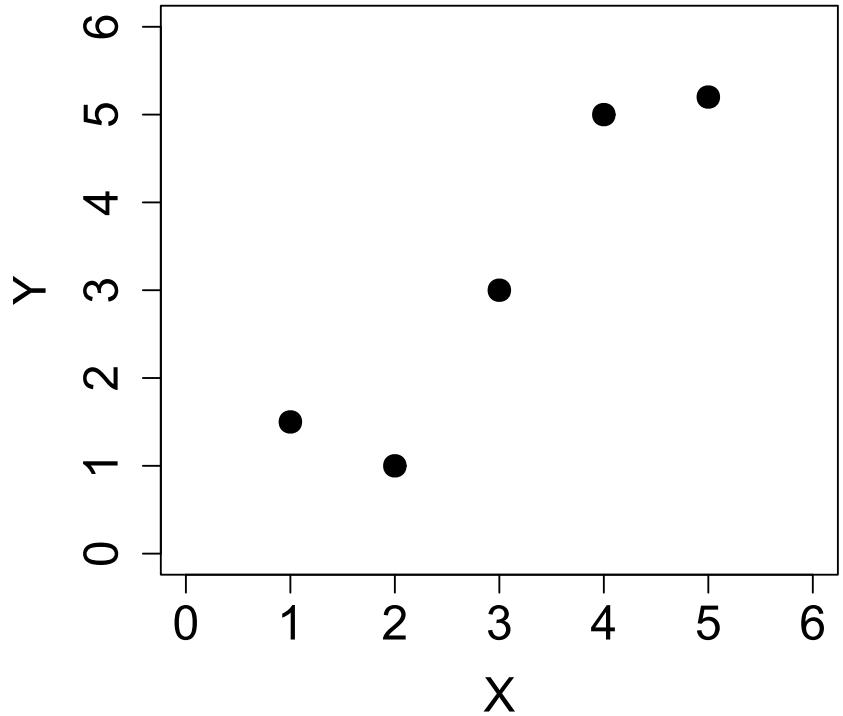
Estimering av koeffisienter



$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 \cdot X_i$$

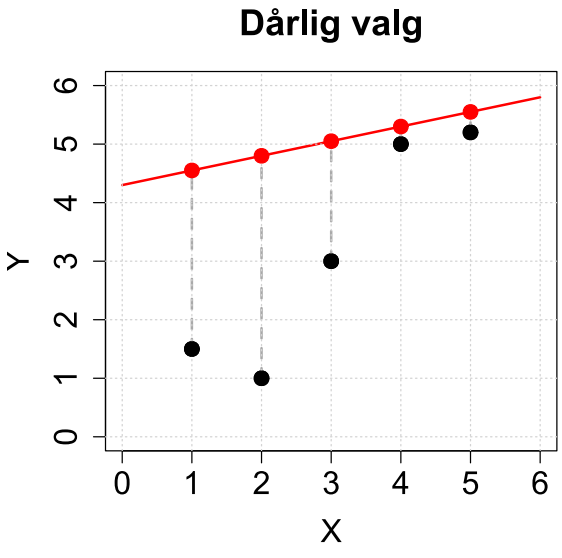
- I underidentifiserte systemer er det ikke nok informasjon til å unikt estimere verdier for modell parametrene.
- I akkurat identifisert (just identified) systemer er det ett unikt sett med estimater.
- I overidentifiserte systemer (slik vi typisk har), kan vi ikke finne en løsning som passer perfekt.

Tilpasse en linje til data



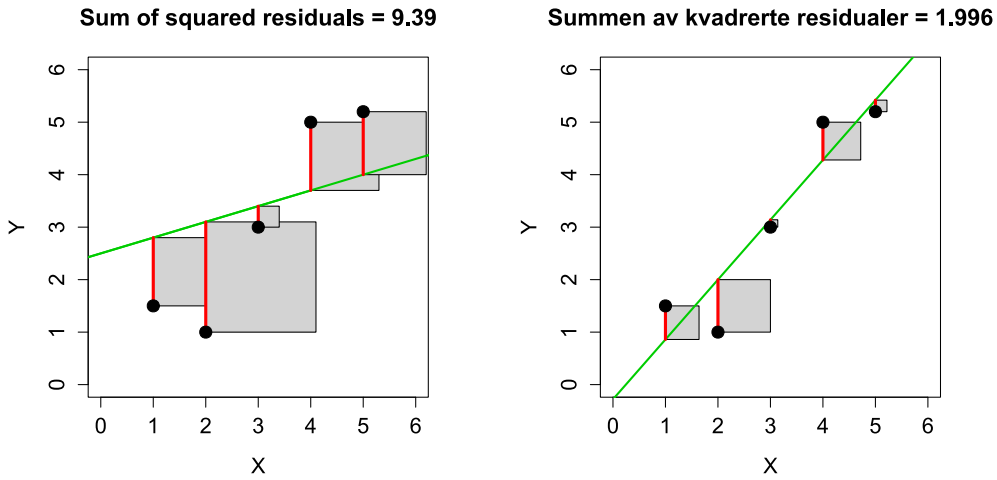
Anta at du har disse fem punktene. Hvordan finner du linjen som passer best?

Tilpasse en linje til data (2)



$$Residual_i = Y_i - \hat{Y}_i$$

Residual: Differansen mellom den observerte og den predikerte verdien.



Det minste kvadraters estimatet for b_0 og b_1 er verdiene $\hat{b}_0$ og $\hat{b}_1$ som minimerer summen av de kvadrerte residualene:

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - (\hat{b}_0 + \hat{b}_1 X_i))^2$$

Normal equations (A: 9.2)

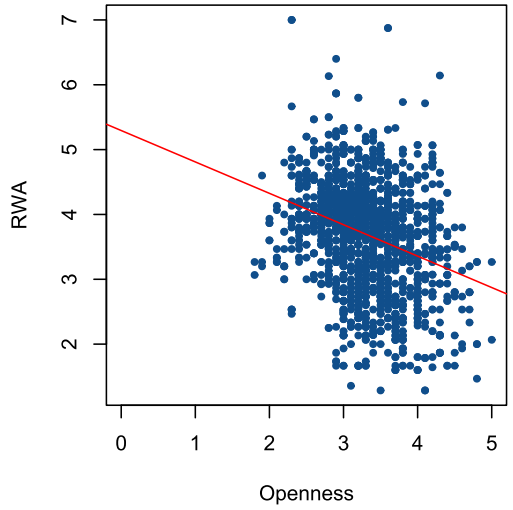
$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \cdot \bar{X}$$

$$\hat{b}_1 = \frac{s_{XY}}{s_X^2}$$

Det kan vises at minste kvadraters estimater av regresjonskoeffisientene kan regnes ut ved disse uttrykkene (kalles ofte *normal equations*).

Eksempel på bruk av normal equations for Openness (X) og RWA (Y)

- $\bar{Y} = 3.691$
- $\bar{X} = 3.311$
- $s_{X,Y} = -0.135$
- $s_X^2 = 0.278$

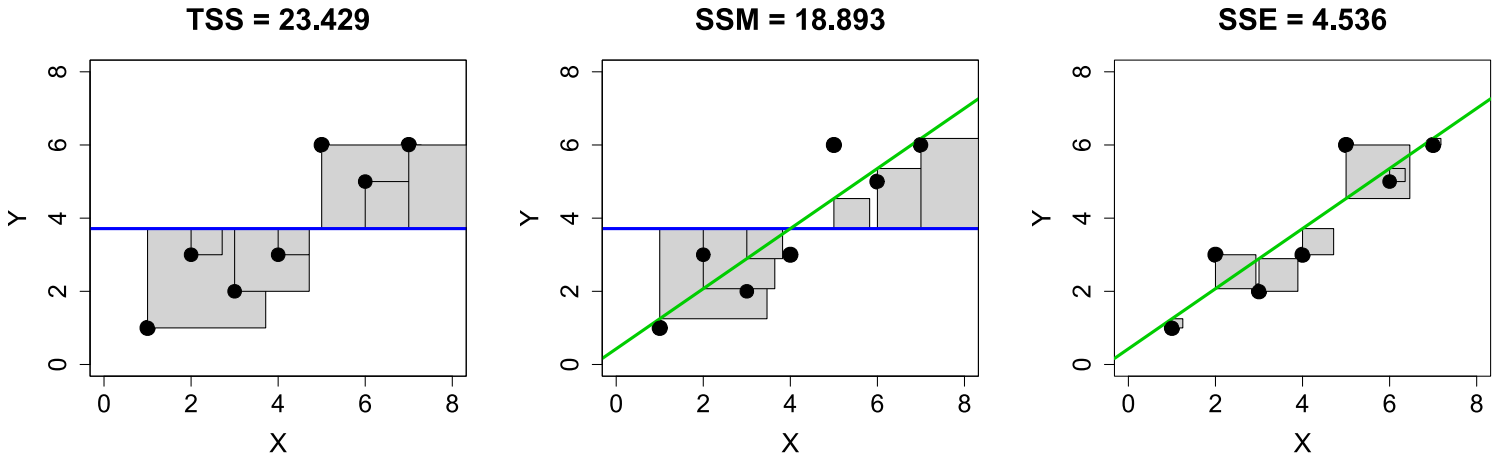


$$\hat{b}_1 = \frac{s_{XY}}{s_X^2} = \frac{-0.135}{0.278} \approx -0.48$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \cdot \bar{X} = 3.691 - (-0.486) \cdot 3.311 \approx 5.29$$

Kvadratsummer (Sum of squares)

Sum of squares



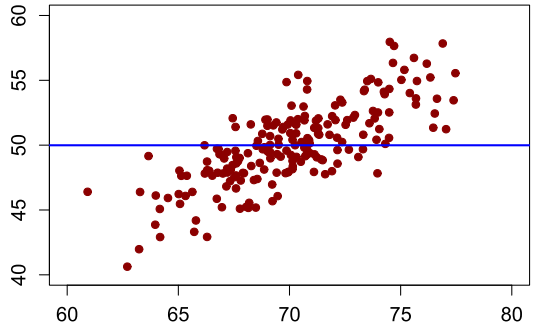
$$TSS = SSM + SSE$$

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

Betinget varians er mindre enn marginal varians (A:9.3)

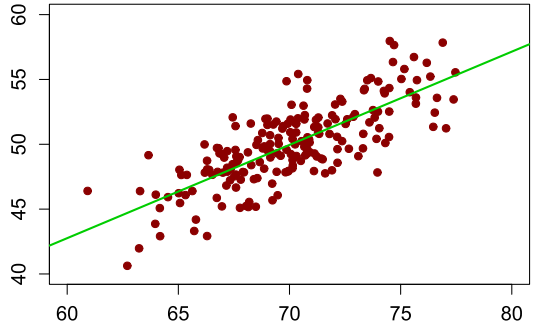
Marginal varians: Varians rundt gjennomsnittet. Dette er varians vi ikke kan forklare før vi tilpasser regresjonsmodellen.

- Her 2057.4

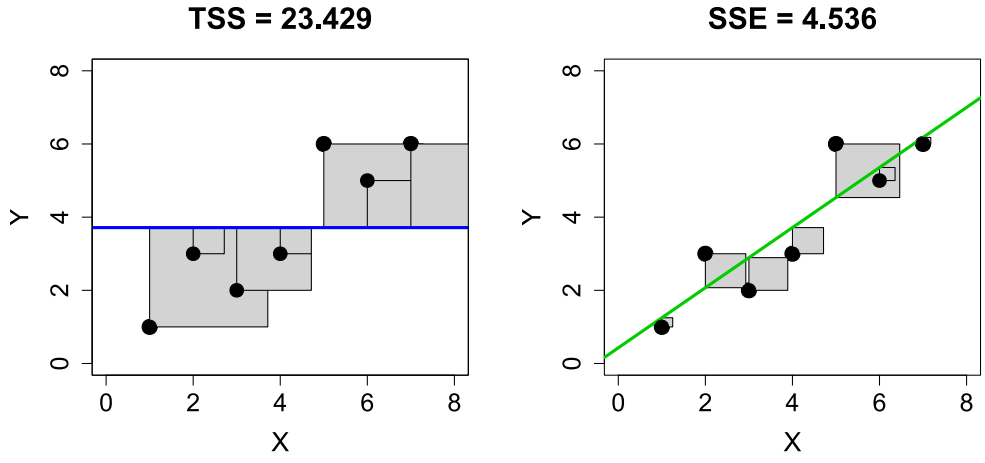


Betinget varians: Variasjon rundt regresjonslinjen. Dette er varians vi ikke kan gjøre rede for etter at vi har tilpasset regresjonsmodellen.

- Her 970.2.



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



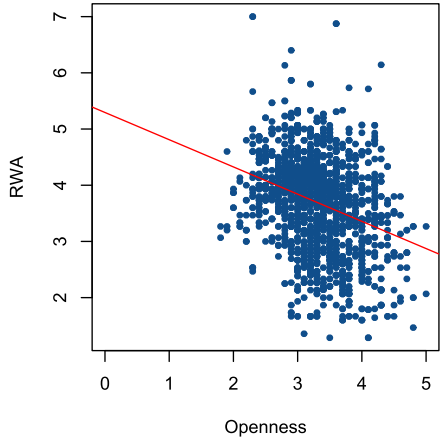
- Kall summen av kvadrerte residualer i den naive modellen til venstre E_1, og summen av kvadrerte residualer i modellen til høyre E_2.
- r^2 er et mål på *nedgangen i proporsjonen* av uforklart varians når vi går fra en modell til en annen.

$$r^2 = \frac{E_1 - E_2}{E_1} = \frac{TSS - SSE}{TSS} = \frac{TSS}{TSS} - \frac{SSE}{TSS} = 1 - \frac{SSE}{TSS}$$

Forklart varians i RWA

```
m4 ← lm (rwa~openness, data=dat)
summary(m4)
```

```
##
## Call:
## lm(formula = rwa ~ openness, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4358 -0.5329  0.0857  0.4982  3.3242
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.29431    0.11809   44.83   <2e-16 ***
## openness    -0.48430    0.03522  -13.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.804 on 1873 degrees of freedom
## Multiple R-squared:  0.09169,    Adjusted R-squared:  0.0912
## F-statistic: 189.1 on 1 and 1873 DF,  p-value: < 2.2e-16
```

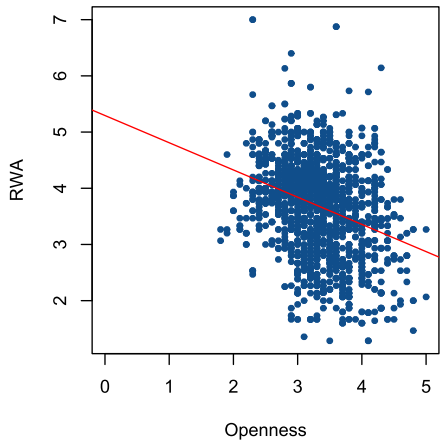


ANOVA() FUNKSJONEN

Du kan skrive ut SSM og SSE ved å bruke anova() funksjonen i R.

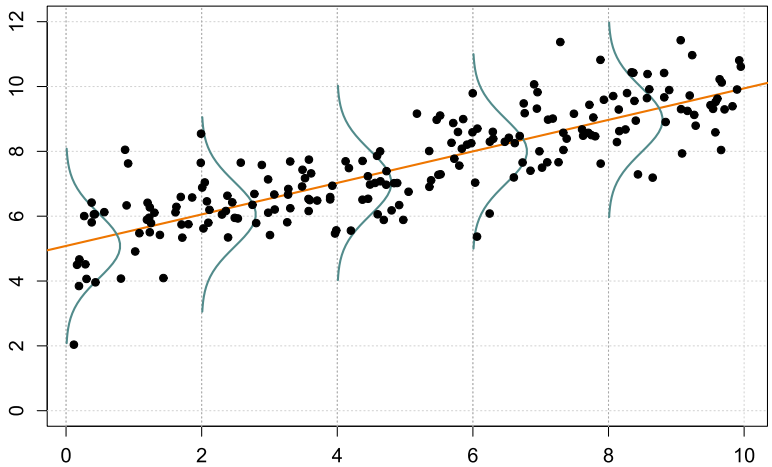
```
m4 ← lm (rwa~openness, data=dat)
anova(m4)
```

```
## Analysis of Variance Table
##
## Response: rwa
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## openness      1  122.22 122.215  189.07 < 2.2e-16 ***
## Residuals  1873 1210.72   0.646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Antagelser i regresjon

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



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.

Residualer

Residual: Differansen mellom den observerte og forventede verdien for en observasjon (Yi).

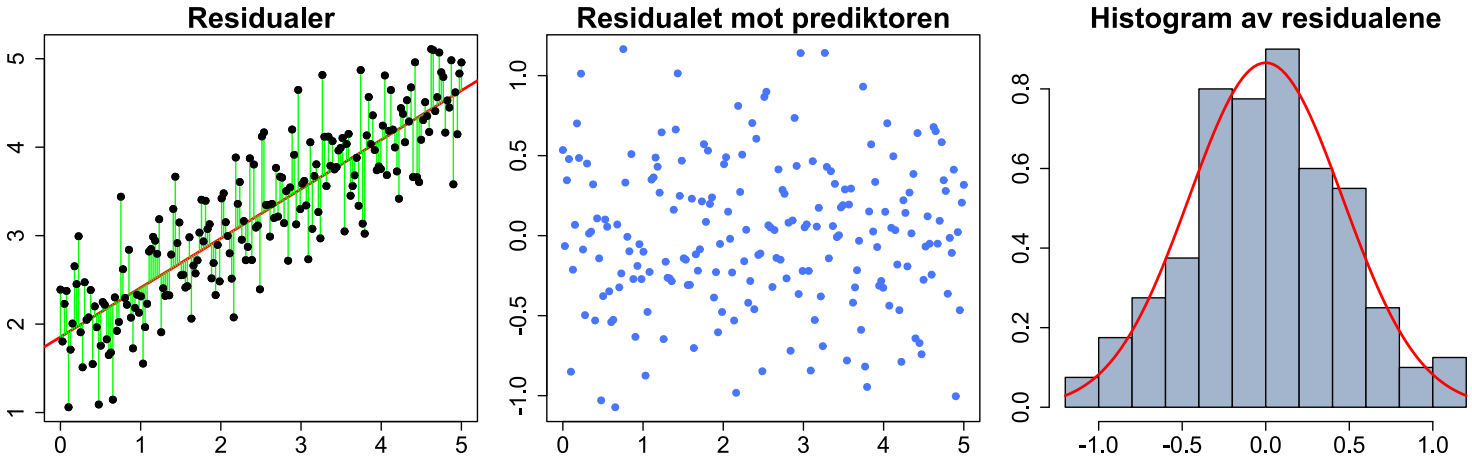
$$Residual = Y_i - \hat{Y}_i$$

Vi sjekker gjerne rimeligheten av de ulike antagelsene i lineær regresjon ved å inspisere diagnostiske plot av residualene.

Vi skal se på tre vanlige diagnostiske plot

1. Histogram av residualene.
2. Spredningsplot av residualene mot en uavhengig variabel.
3. QQ-plots.

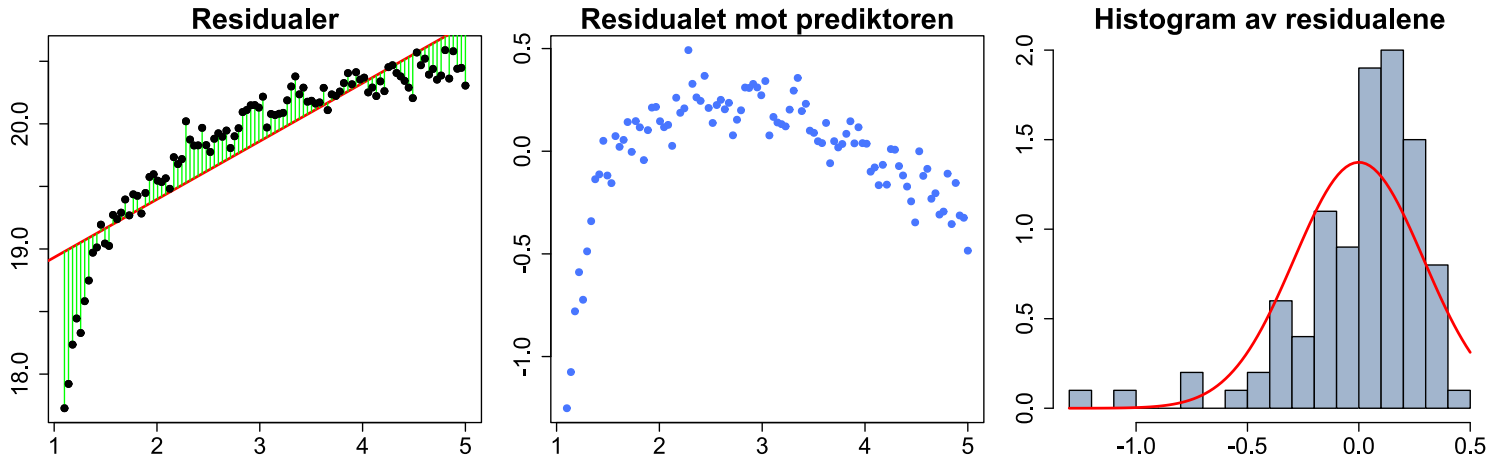
Ingen antagelser brutt



Du ønsker å se:

- Ingen tydelig struktur i scatter plottet av residualene.
- Tilnærmet normalfordelt histogram av residualene.

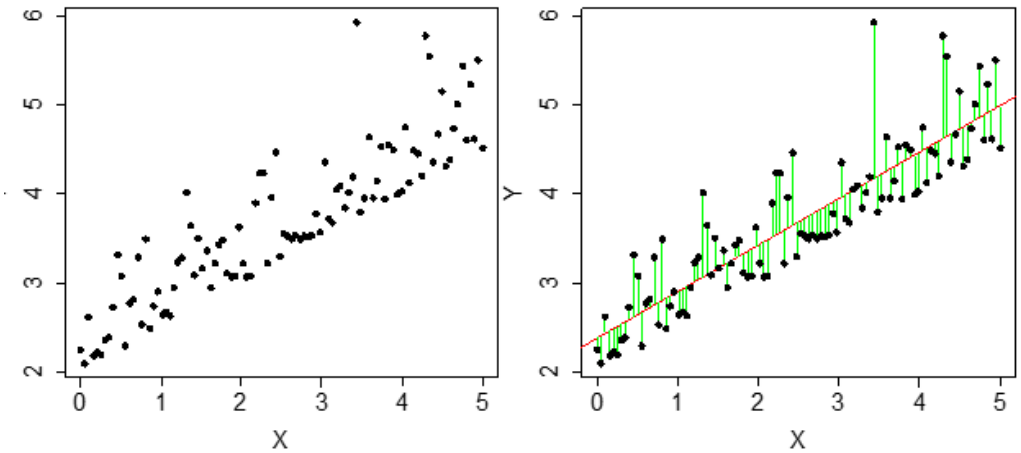
Brudd på antagelsen om linearitet



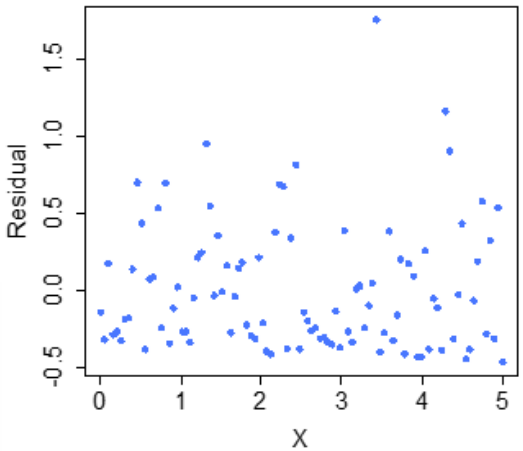
Dersom antagelsen om linearitet brutt kan hele modellen være uvalid.

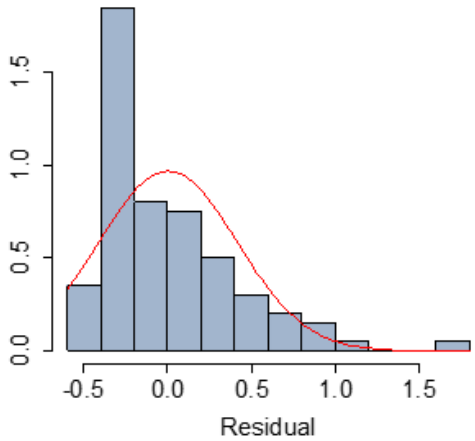
Brudd på antagelse av normalitet

Riktig signifikansnivåer (p-verdier) er basert på antagelse om normalfordelt feil.

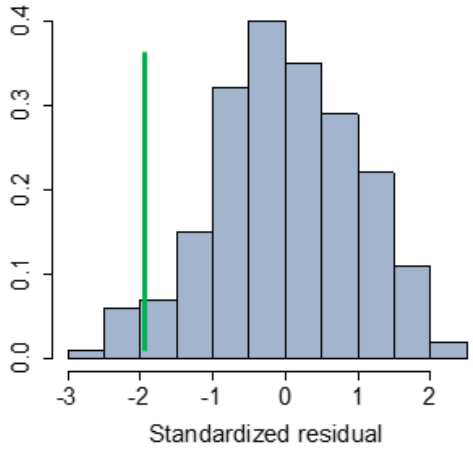


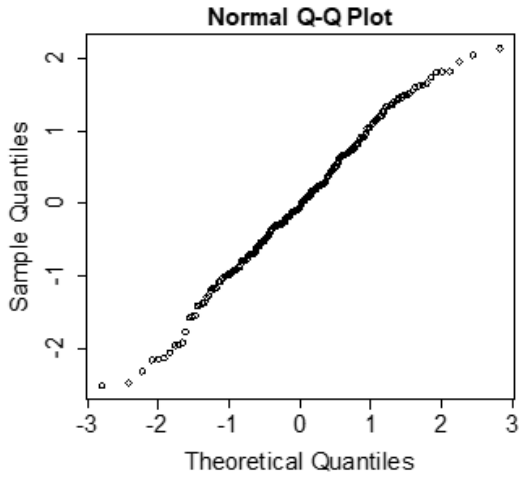
Eksempel på ikke-normalfordelt feil

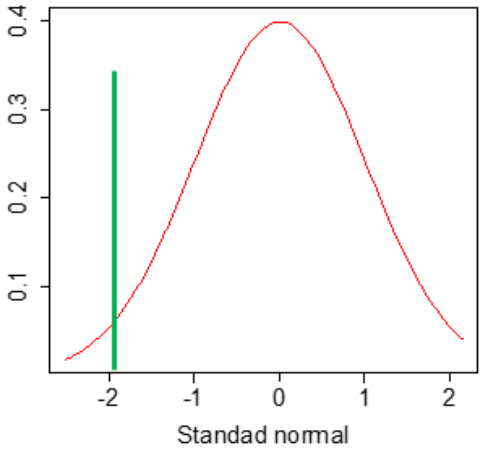




QQ-PLOT





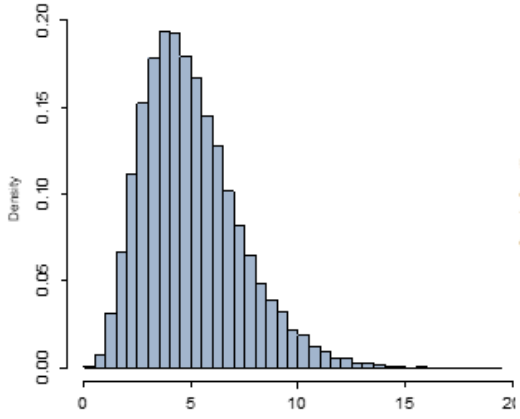


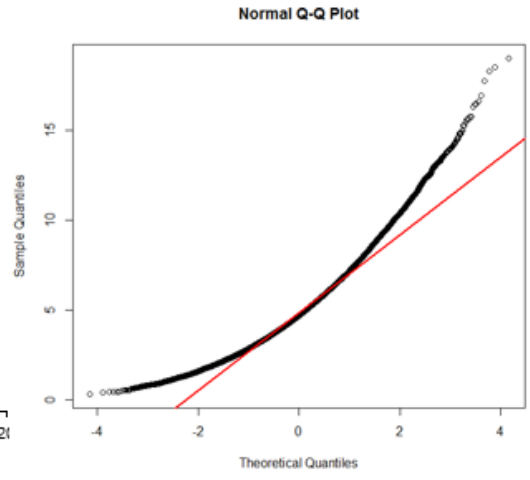
QQ plot: ("Q" star for "quantile") er en grafisk måte å sammenlikne to fordelinger ved å plotte kvantilene deres mot hverandre.

Vi vil typisk bruke den for å avgjøre om residualene kan sies å være normalfordelte.

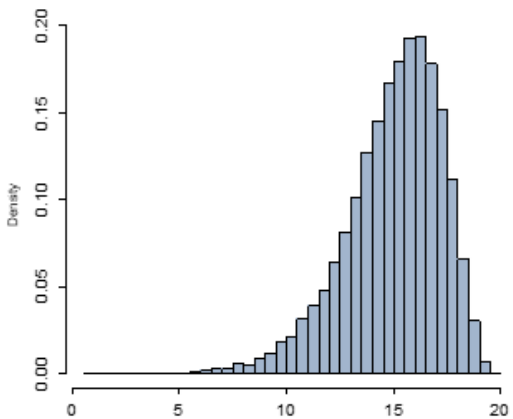
[1] http://en.wikipedia.org/wiki/Q%E2%80%93Q_plot

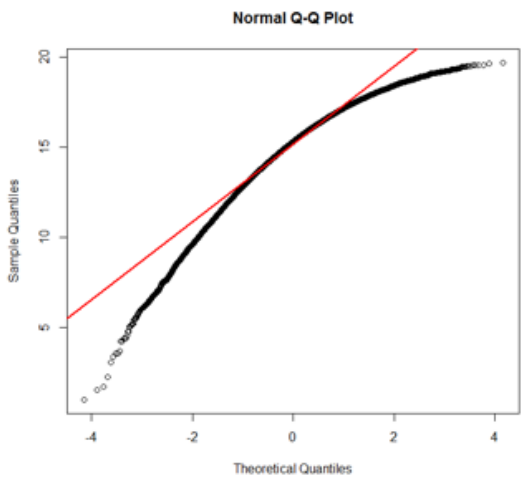
QQ-plot ved skjevhet





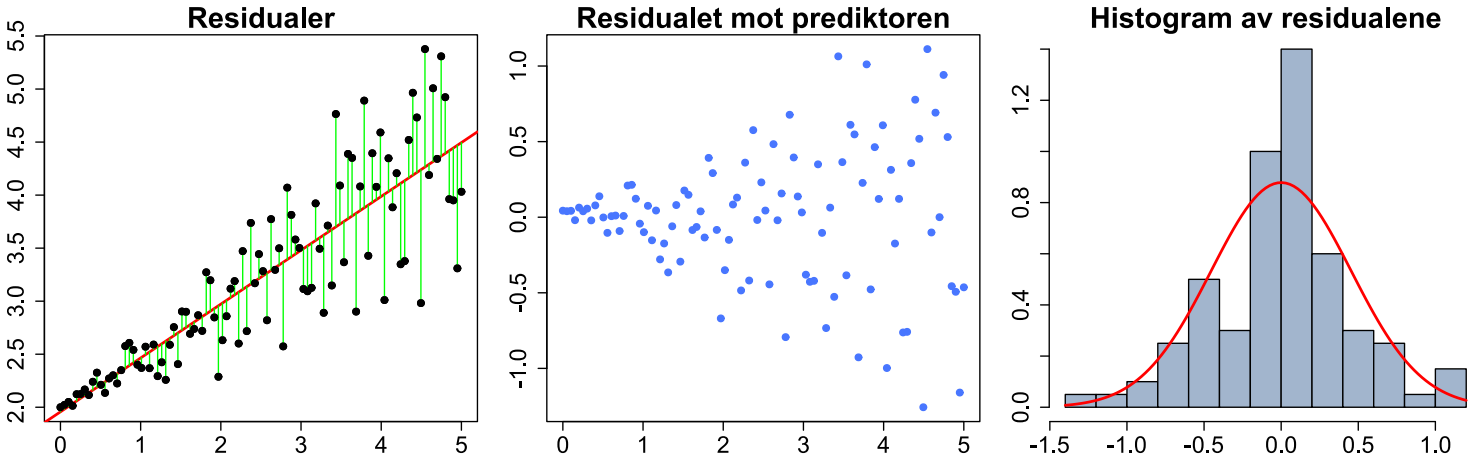
En slik krumming indikerer positiv skjevhet.





En slik krumming indikerer negativ skjevhet.

Brudd på antagelsen om konstant feilvarians



Er antagelsen om konstant feilvarians brutt vil du få valide koeffisienter, men trolig gale p-verdier (og konfidensintervaller).

AUGMENT-FUNKSJONEN()

Funksjonen `augment()` i `broom` pakken tar som input et modellobjekt og et datasett og legger til informasjon om hver observasjon i datasettet.

```
library(broom)

m1 ← lm (rwa~openness, data=dat)  # Kjør regresjonsanalyse
aug ← augment(m1, dat) # Augment funksjonen tar ut diverse
aug
```

```
## # A tibble: 1,875 x 10
##    .rownames   rwa openness openness_c .fitted .resid     .hat .sigma   .cooksd
##    <chr>     <dbl>    <dbl>      <dbl>   <dbl>  <dbl>    <dbl>  <dbl>     <dbl>
##  1 1           3.8      2.5    -0.811     4.08 -0.284 0.00180   0.804 0.000112
##  2 2           4.13     4.1     0.789     3.31  0.825 0.00173   0.804 0.000912
##  3 3           4.93     3      -0.311     3.84  1.09  0.000719  0.804 0.000664
##  4 4           4.13     3.3    -0.0110    3.70  0.437 0.000534  0.804 0.0000790
##  5 5           4.4      3.6     0.289     3.55  0.849 0.000694  0.804 0.000387
##  6 6           4.27     3.1    -0.211     3.79  0.474 0.000619  0.804 0.000108
##  7 7           2.27     3.2    -0.111     3.74 -1.48  0.000557  0.803 0.000942
##  8 8           3.93     2.3    -1.01      4.18 -0.247 0.00250   0.804 0.000118
##  9 10          2.5      3      -0.311     3.84 -1.34  0.000719  0.804 0.00100
## 10 11          1.67     3.8     0.489     3.45 -1.79  0.000992  0.803 0.00246
## #  ... with 1,865 more rows, and 1 more variable: .std.resid <dbl>
```

Eksempel: Residualer i RWA

```
m1 ← lm (rwa~openness, data=dat)  # Kjør regresjonsanalyse
aug ← augment(m1) # Ta ut diverse resultater

# Histogram of residuals
ggplot(aug, aes(x=.resid)) + geom_histogram(color="white")

# Scatter plot of residuals
ggplot(aug, aes(x=openness, y=.resid)) + geom_point()

# QQ plot of residuals
ggplot(aug, aes(sample=.resid)) +
  geom_qq() +
  geom_qq_line(color="red")
```

Forklart varians?

```
## Analysis of Variance Table
##
## Response: rwa
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## openness     2  100      50       24.5 < 2.2e-16 ***
## Residuals   98  200    2.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hvor mye av variansen i den avhengige variabelen kan forklares?

1. 0%
2. 25%
3. 33%
4. 50%
5. 75%

Konklusjon