STK1110 Høsten 2026

Enkel lineær regresjon

Tilsvarer Avsnitt 12.1-12.6

Oppsummering så langt

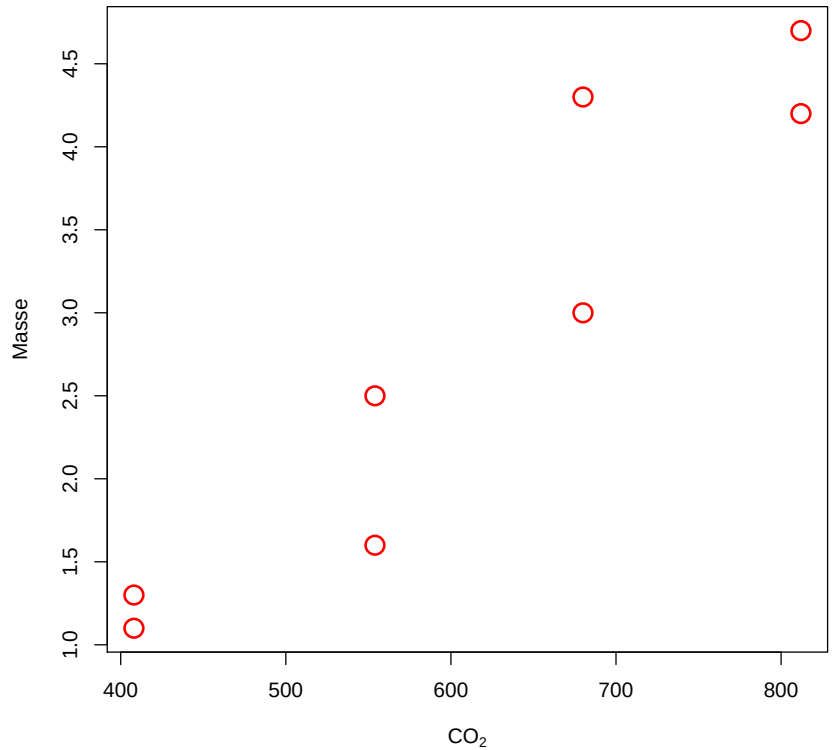
- Inferens for et utvalg
- Inferens for to utvalg
  - Forskjeller (variasjon) i respons forklares med ulike gruppetilhørigheter
- Variansanalyse, inferens for flere utvalg
  - Forskjeller (variasjon) i respons forklares med ulike gruppetilhørigheter
- Grupper: Kategorisk/kvalitativ forklaringsvariabel (faktor)
- Regresjon: Tilsvarende, men (også) kvantitativ forklaringsvariable
  - Punktestimering
  - Konfidensintervall
  - Hypotesetesting

Plan for forelesninger om regresjon

- Avsnitt 12.1 og 12.2 Enkel lineær regresjonsmodell - Repetisjon - Tirsdag 14/10
- Avsnitt 12.3 Egenskaper minste kvadraters estimatorer, inferens enkel lineær regresjon - Onsdag 15/10
- Avnsitt 12.4 Inferens forventet og ny respons - Onsdag 15/10
- Avsnitt 12.5 Korrelasjon - Tirsdag 21/10
- Avsnitt 12.6 Modellsjekk - Residualanalyse - Tirsdag 21/10
- Avsnitt 12.7 Multippel lineær regresjon - Onsdag
- Avsnitt 12.8 Polynomiell regresjon og interaksjon - Onsdag
- Avsnitt 12.9 Matriseformulering regresjon - Uke 44 -
- Avsnitt 12.10 Logistisk regresjon binære responser - Uke 44

Eksempel: CO_2-konsentrasjon og vekst av furutrær

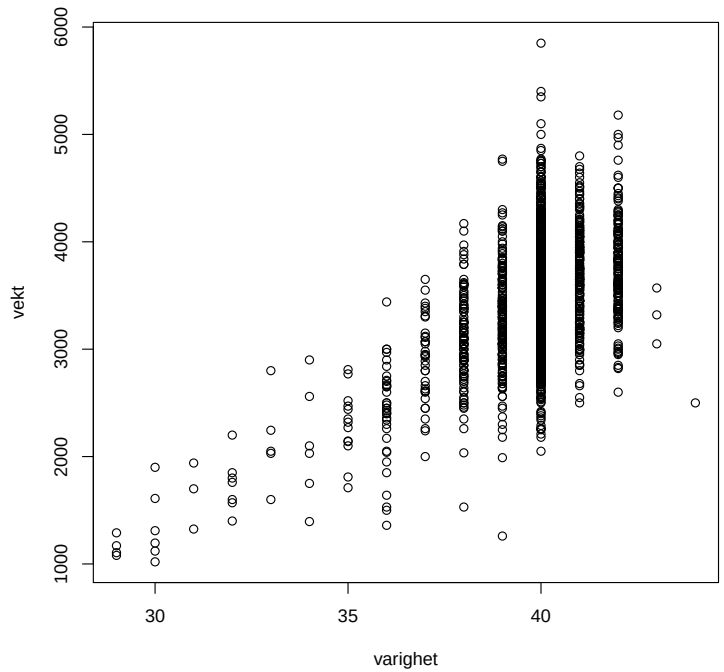
CO_2	Masse
408	1.1
408	1.3
554	1.6
554	2.5
680	3.0
680	4.3
812	4.2
812	4.7



Vil kvantifisere sammenhengen mellom CO_2-konsentrasjon og masse?

Eksempel: fødselsvekt og lengden av svangerskap

Figuren viser vekten til et utvalg av 2116 nyfødte gutter:



Hvordan henger fødselsvekt sammen med varighet av svangerskapet?

Eksempel: kroppsvekt og vekt av lillehjernen

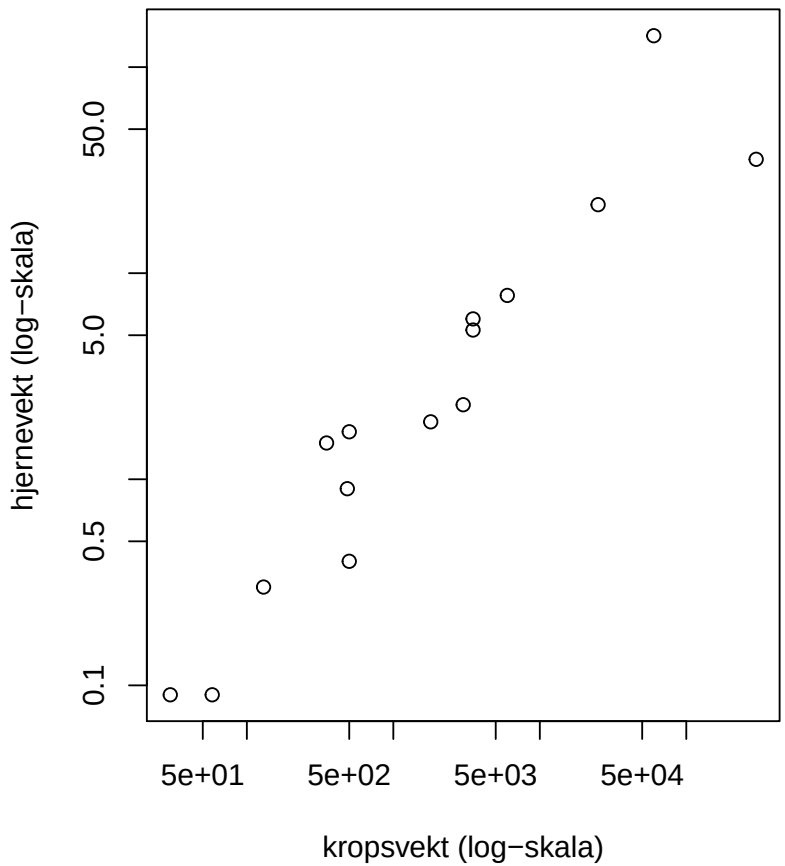
Species	**Cerebellum** Weight (grams)	Body Weight (grams)
Mouse	0.09	58
Bat	0.09	30
Flying Fox	0.3	130
Pigeon	0.4	500
Guinea Pig	0.9	485
Squirrel	1.5	350
Chinchilla	1.7	500
Rabbit	1.9	1,800
Hare	2.3	3,000
Cat	5.3	3,500
Dog	6.0	3,500
Macaque	7.8	6,000
Sheep	21.5	25,000
Bovine	35.7	300,000
Human	142	60,000
Sultan, F. and Braitenberg, V. Shapes and sizes of different mammalian cerebella. A study in quantitative comparative neuroanatomy. *J. Hirnforsch.*, 34:79-92, 1993.		

Tabellen viser kroppsvekt og vekten av lillehjernen for en del arter.

Kan vi finne en sammenheng mellom vekt av lillehjernen og kroppsvekten?

Kroppsvekt og hjernevekt, illustrasjon

Etter å log-transformere både kroppsvekt og lillehjernevekt illustreres sammenhengen ved spredningsplott



og altså som ganske lineær.

Lineær regresjon

- Vi har par av observasjoner (x_i, y_i), $i = 1, \ldots, n$, og vi vil studere hvordan y_i-ene avhenger av x_i-ene.
- Vi antar at y_i er observert verdi av den stokastiske variabelen Y_i, mens x_i er gitt (på forhånd), og vi har:

$$Y_i = f(x_i) + \varepsilon_i, \tag{1}$$

  der $f(\cdot)$ er en eller annen funksjon og ε_i-ene er tilfeldige feil, som regel uif med $\mathrm{E}(\varepsilon_i) = 0$ og $\mathrm{V}(\varepsilon_i) = \sigma^2$.
- Her er
  - Y_i responsvariabel (alt. avhengig variabel, "target variable")
  - x_i forklaringsvariabel (alt. kovariat, prediktor, uavhengig variabel, "feature").

Lineær regresjon (forts.)

- I regresjon oppfatter vi de observerte x_i−ene som gitte verdier
- Men de kan være utfall av stokastiske X_i, i så fall er analysen betinget på $X_i = x_i$
- I (1) ser vi på den betingede fordelingen til $Y_i|X_i = x_i$.

- Vi vil konsentrere oss om lineære funksjoner $f(\cdot)$.
- Det gir den lineære regresjonsmodellen

$$Y_i = f(x_i) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{2}$$

der β_0 og β_1 er ukjente parametere.

  - Noen ganger må vi transformere x_i-ene og/eller y_i-ene for å få en lineær sammenheng.

- Antagelser på ε_i:
  1. $E[\varepsilon_i] = 0$
  2. $V[\varepsilon_i] = \sigma^2$
  3. $\varepsilon_1, ..., \varepsilon_n$ er uavhengige
  4. $\varepsilon_i \sim N(0, \sigma^2)$

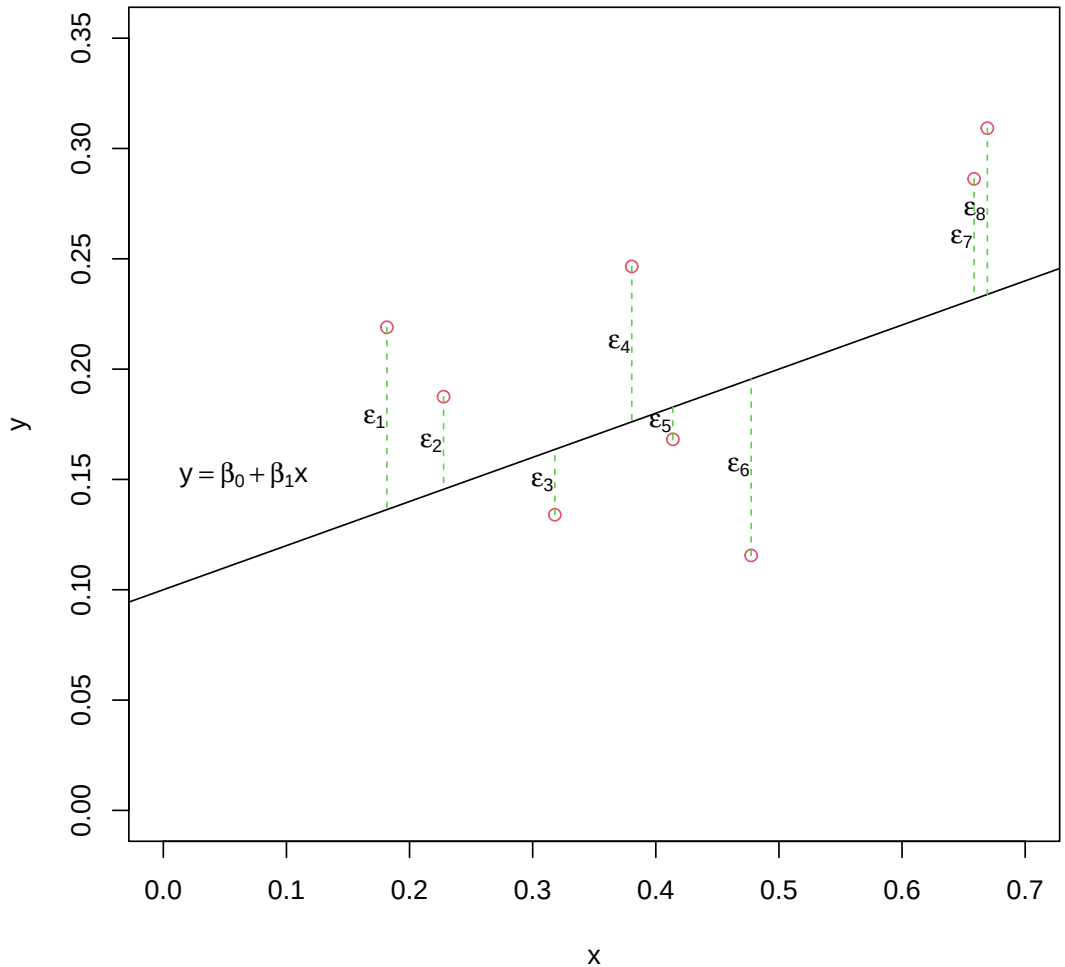
Fortolkning av β_0 og β_1

- I modellen $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ er $\mathrm{E}(Y_i|x_i) = \beta_0 + \beta_1 x_i$.
- Det betyr at
  - β_0 er forventet respons når $x_i = 0$
  - β_1 er forventet endring i respons når x_i øker med 1.
- $x_i = 0$ kan være en lite relevant verdi
- Omformulering av modell: $Y_i = \beta_0 + \beta_1(x_i - x_0) + \varepsilon_i$ der x_0 er en relevant verdi, f.eks. $\bar{x}$.
- Fortolkningen av β_1 er da den samme,
- mens β_0 nå er forventet respons når $x = x_0$.
- Estimatoren for β_0 er nå $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1(\bar{x} - x_0)$.

Eksempel

Fødselsvekt og svangerskapslengde.

Lineær regresjon: illustrasjon støyledd ε_i



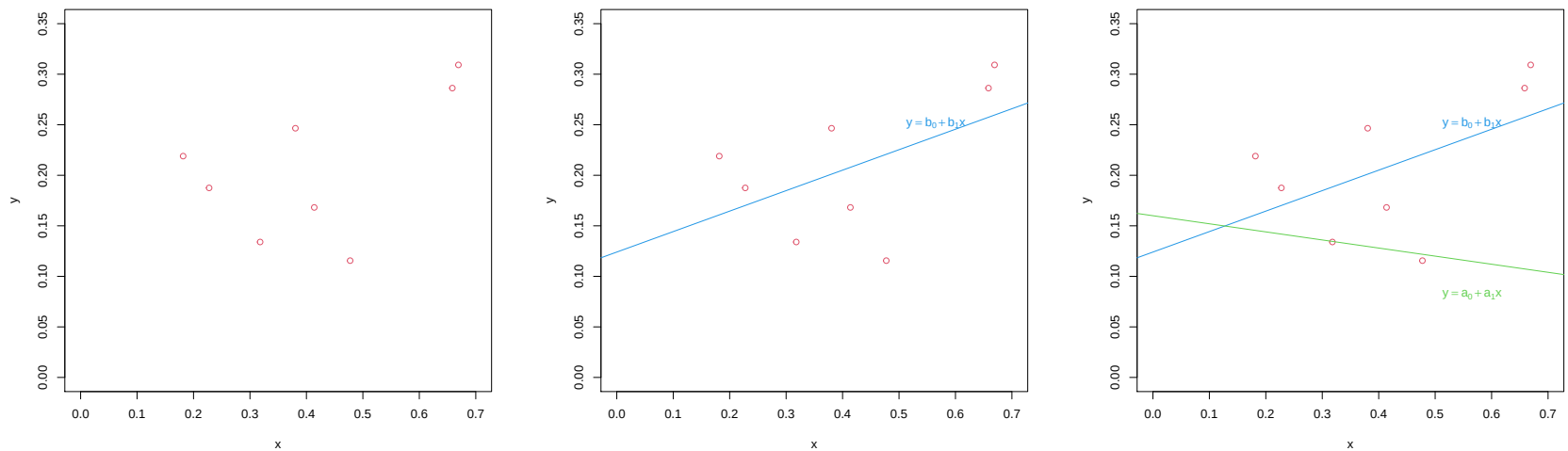
Støyledd $\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$ kan skyldes

- Tilfeldig variasjon
- Observasjonsfeil
- Modellfeil (avvik fra linearitet)

Minste kvadraters metode

Når vi skal estimere modellparameterne β_0, β_1 og σ^2, prøver vi å finne

- den rette linja $b_0 + b_1 x$
- som "passer best" til de observerte punktene (x_i, y_i).



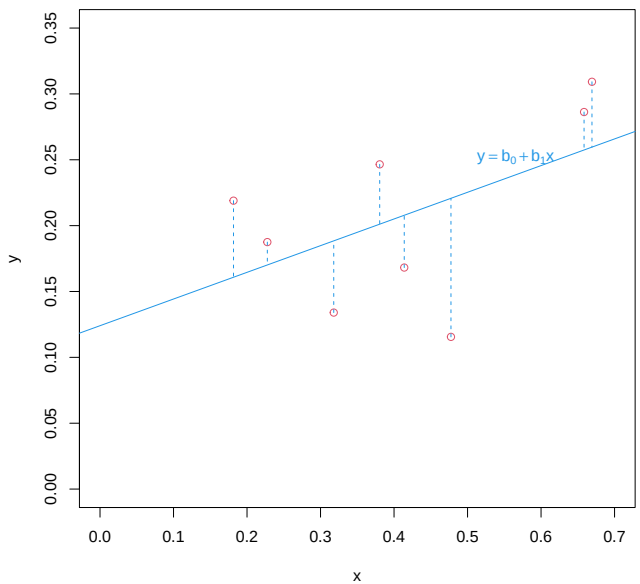
Vi må da definere hva vi mener med "passer best".

Minste kvadraters metode (forts.)

La

$$g(b_0, b_1) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2,$$

som er summen av kvadratavvik mellom y_i og $b_0 + b_1 x_i$ i vertikal retning.



Minste kvadraters estimatorer får man ved å minimere $g(b_0, b_1)$ m.h.p. b_0 og b_1.

Minste kvadraters metode (forts.)

Minste kvadraters estimater (MKE) er gitt ved

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

og

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Viser dette på tavla

Vi skal videre vise at hvis $\varepsilon_i \sim N(0, \sigma^2)$, så er minste kvadraters estimatene også maksimum likelihood-estimatene.

- Men MKE har gode egenskaper også under mindre restriktive antagelser!

Utledning MKE

- Ønsker å minimere

$$g(b_0, b_1) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2,$$

- som er en glatt funksjon av (b_0, b_1)

$$\begin{aligned}
\frac{\partial}{\partial \beta_0} g(b_0, b_1) &= -2 \textstyle\sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i)) \\
&= -2[n\bar{y} - nb_0 - nb_1 \bar{x}] \\
\frac{\partial}{\partial \beta_1} g(b_0, b_1) &= -2 \textstyle\sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i)) x_i \\
&= -2[\textstyle\sum_{i=1}^{n} y_i x_i - nb_0 \bar{x} - b_1 \sum_{i=1}^{n} x_i^2]
\end{aligned}$$

Setter lik 0 og løser mhp b_0 og b_1
Løsningene $\hat{\beta}_0 = b_0$ og $\hat{\beta}_1 = b_1$, gis altså fra

$$\begin{aligned}
0 &= \bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} \\
0 &= \textstyle\sum_{i=1}^{n} y_i x_i - n\hat{\beta}_0 \bar{x} - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2
\end{aligned}$$

Alternativ utledning av minste kvadrater

Vi kan reformulere modellen til

$$Y_i = \alpha_0 + \beta_1(x_i - \bar{x}) + \varepsilon_i$$

der $\alpha_0 = \beta_0 + \beta_1\bar{x}$. Uttrykt ved α_0 og β_1 får vi kvadratsummen som

$$\sum_{i=1}^{n}(y_i - \alpha_0 - \beta_1(x_i - \bar{x}))^2$$

og denne har partiell-deriverte

$$\frac{\partial \sum_{i=1}^{n}(y_i-\alpha_0-\beta_1(x_i-\bar{x}))^2}{\partial \alpha_0} = -2\sum_{i=1}^{n}(y_i - \alpha_0 - \beta_1(x_i - \bar{x}))$$

$$= -2n(\bar{y} - \alpha_0)$$

$$\frac{\partial \sum_{i=1}^{n}(y_i-\alpha_0-\beta_1(x_i-\bar{x}))^2}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \alpha_0 - \beta_1(x_i - \bar{x}))(x_i - \bar{x})$$

$$= -2(\sum_{i=1}^{n} y_i(x_i - \bar{x}) - \beta_1 \sum_{i=1}^{n}(x_i - \bar{x})^2)$$

siden $0 = \sum_{i=1}^{n}(x_i - \bar{x}) = \beta_1 \sum_{i=1}^{n}(x_i - \bar{x}) = \alpha_0 \sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

Alternativ utledning MKE, forts.

Men siden $0 = \bar{y}\sum_{i=1}^{n}(x_i - \bar{x})$ blir også

$$\sum_{i=1}^{n} y_i(x_i - \bar{x}) = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$$

og den andre partiell-deriverte satt lik 0 gis som

$$\hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}),$$

altså $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

Den første partiell-deriverte gir direkte $\hat{\alpha}_0 = \bar{y}$ og dermed siden $\beta_0 = \alpha_0 + \beta_1\bar{x}$ blir

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Utledning av at MKE=MLE når $\varepsilon_i \sim N(0, \sigma^2)$

- Anta $\varepsilon_i \overset{uif}{\sim} N(0, \sigma^2)$,
- Da er $Y_i \overset{uavh}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$
- Likelihood

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 x_i)^2\right) \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2\right) \end{aligned}$$

Videre

$$\ell(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2$$

- Maksimering av $\ell(\beta_0, \beta_1, \sigma^2)$ mhp (β_0, β_1) er ekvivalent med minimering av

$$g(\beta_0, \beta_1) = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 x_i))^2,$$

Residualer og tilpassede (predikerte) verdier

- De tilpassede (eller predikerte) verdiene er

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \ldots, n$$

og residualene er

$$e_i = \hat{\varepsilon}_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n.$$

- Residualkvadradsummen defineres nå som

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

- Denne kan brukes til å estimere σ^2:

$$s^2 = \frac{SSE}{n-2} = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Hvorfor dele på $n-2$?

- For $Y_1, ..., Y_n \overset{uif}{\sim} N(\mu, \sigma^2)$ så brukte vi

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

  Merk: $Y_i = \mu + \varepsilon_i$ med $\varepsilon_i \overset{uif}{\sim} N(0, \sigma^2)$

- For $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ med $\varepsilon_i \overset{uif}{\sim} N(0, \sigma^2)$ så bruker vi

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

- I utgangspunkt n "frihetsgrader"
  - For det første tilfellet "bruker vi opp" 1 frihetsgrad for å estimere μ, har $n-1$ igjen til σ^2
  - For det andre tilfellet "bruker vi opp" 2 frihetsgrader for å estimere β_0, β_1, har $n-2$ igjen til σ^2
- I begge tilfeller blir s^2 forventningsrett

Estimering av σ^2

- Estimatoren S^2 får en ved å bytte ut y_i-er med Y_i-er, og denne er forventningsrett for σ^2.

- $S^2_{ml} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ er ML estimatoren for σ^2
  - Kan finnes ved å maksimere $\ell(\beta_0, \beta_1, \sigma^2)$ mhp σ^2.

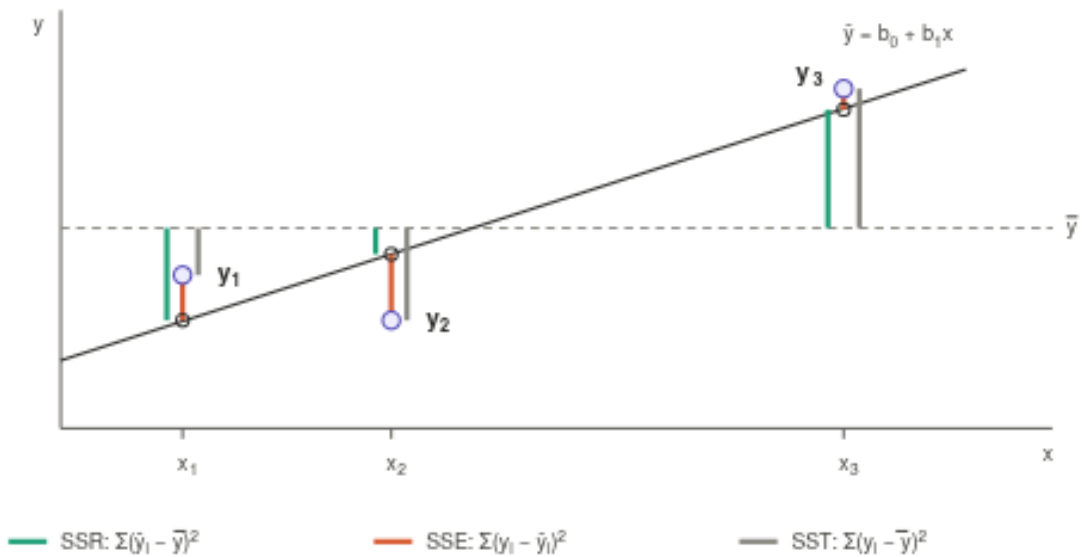
- Foretrekker oftest S^2
  - Liten forskjell hvis n er stor

$$S^2 = \frac{n}{n-2} S^2_{ml}$$

Kvadratsummer og variasjon

La nå

$$
\begin{aligned}
SST &= \sum_{i=1}^{n}(y_i - \bar{y})^2 && \text{Total sum of squares} \\
SSR &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 && \text{Regression sum of squares} \\
SSE &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 && \text{Error sum of squares}
\end{aligned}
$$



Kvadratsummer og variasjon

La nå

$$\begin{aligned} SST &= \sum_{i=1}^{n}(y_i - \bar{y})^2 && \text{Total sum of squares} \\ SSR &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 && \text{Regression sum of squares} \\ SSE &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 && \text{Error sum of squares} \end{aligned}$$

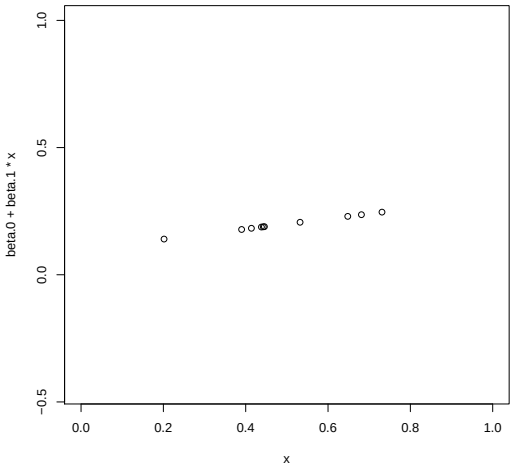
Kan vises: $SST = SSR + SSE$ der altså

- SSR: Variasjon forklart av modell
- SSE: Variasjonen forklart av tilfeldig støy.

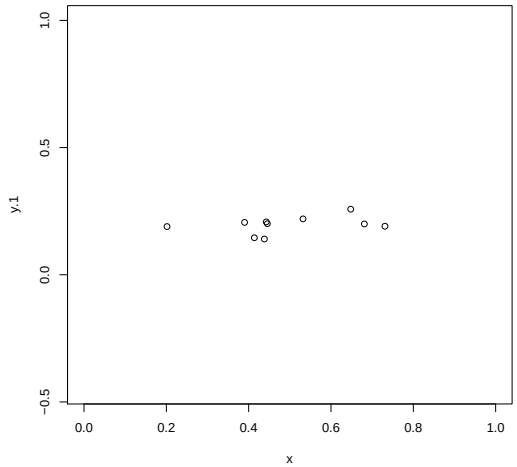
Vi definerer på denne bakgrunn følgende mål på variasjonen forklart av modellen:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

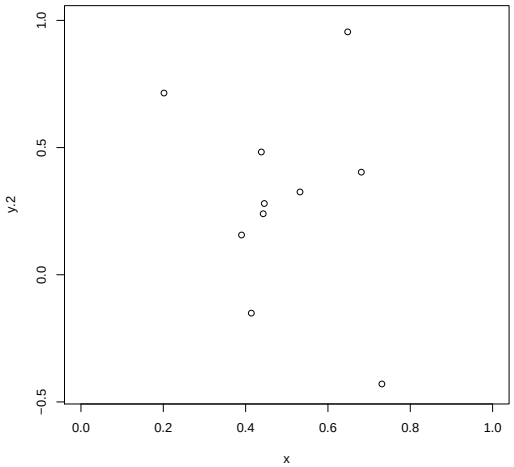
Kvadratsummer og variasjon (forts.)



All variasjon forklart



Mye variasjon forklart



Lite variasjon forklart

Eksempel

CO_2 og vekst av trær.

Eksempel

Fødselsvekt og svangerskapslengde.

Eksempel

Kroppsvekt og vekt av lillehjernen.

Vil vise SST = SSR + SSE

$$
\begin{aligned}
SST &= \sum_{i=1}^{n}(y_i - \bar{y})^2 && \text{Total sum of squares} \\
SSR &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 && \text{Regression sum of squares} \\
SSE &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 && \text{Error sum of squares}
\end{aligned}
$$

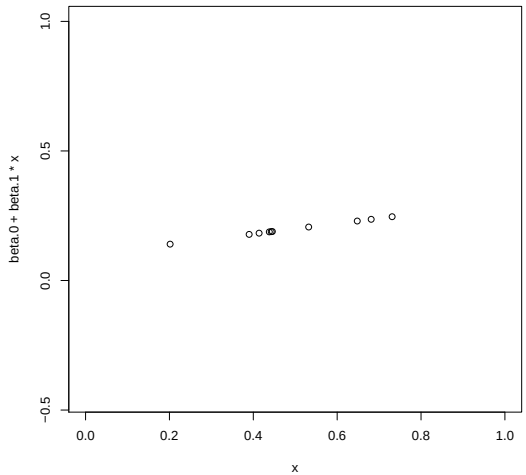
Vi begynner med

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 = \sum_{i=1}^{n}(\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2$$

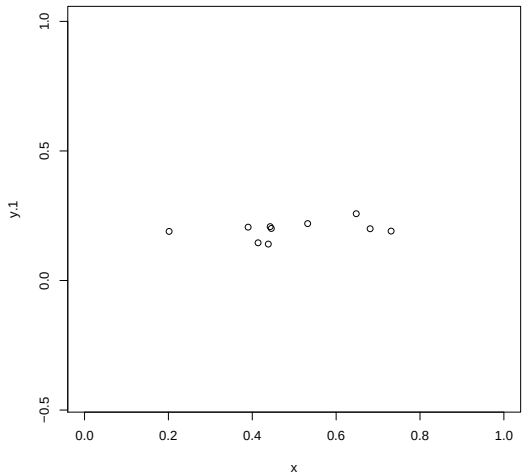
altså

$$SSR = \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$$

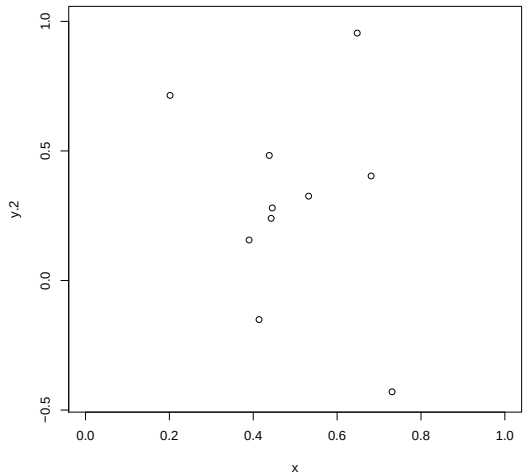
Kvadratsummer og variasjon (forts.)



All variasjon forklart



Mye variasjon forklart



Lite variasjon forklart

Eksempel

CO_2 og vekst av trær.

Eksempel

Fødselsvekt og svangerskapslengde.

Eksempel

Kroppsvekt og vekt av lillehjernen.

Vil vise SST = SSR + SSE,forts

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2$$

som igjen kan uttrykkes

$$\begin{aligned} SST &= \sum_{i=1}^{n}(y_i - \hat{y}_i + \bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{\beta}_1(x_i - \bar{x}))^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 = SSE + SSR \end{aligned}$$

siden

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \hat{y}_i)(x_i - \bar{x}) &= \sum_{i=1}^{n}(y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x}))(x_i - \bar{x}) \\ &= \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) - \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 - \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 = 0 \end{aligned}$$

Forklart andel av variasjon R^2 = kvadratet av korrelasjon

Vi definerer korrelasjonskoeffisienten (eg. Pearsons momentkorrelasjon) ved

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Da har vi at

- $-1 \leq r \leq +1$
- Hvis $r = +1$ ligger (x_i, y_i) på en rett stigende linje
- Hvis $r = -1$ ligger (x_i, y_i) på en rett avtagende linje
- $r^2 = \frac{SSR}{SST}$

God øving å vise:

$$r^2 = \frac{[\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})]^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSR}{SST}$$

Avsnitt 12.3: Egenskaper til $\hat{\beta}_0$ og $\hat{\beta}_1$

- Vår modell er altså

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{uif}{\sim} N(0, \sigma^2)$$

så $Y_i \overset{uavh}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$ og $\bar{Y} \sim N(\beta_0 + \beta_1 \bar{x}, \sigma^2/n)$.

- Minste kvadraters estimatorer er gitt ved

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \quad \text{og} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

- Disse kan omskrives til, med $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$,

$$\hat{\beta}_0 = \sum_{i=1}^n \left(\frac{1}{n} - \bar{x} \frac{x_i - \bar{x}}{S_{xx}} \right) Y_i \quad \text{og} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{S_{xx}}$$

dvs. som lineær-kombinasjoner av Y_i-ene.

Egenskaper for $\hat{\beta}_0$ og $\hat{\beta}_1$: viser på tavla

Vi ønsker å etablere

- $\hat{\beta}_1$ og $\hat{\beta}_0$ er forventningsrette, $\mathrm{E}(\hat{\beta}_j) = \beta_j$
- $\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$
- $\hat{\beta}_1 \sim \mathrm{N}(\beta_1, \frac{\sigma}{\sqrt{S_{xx}}})$

Dessuten gjelder

- $\mathrm{Var}(\hat{\beta}_0) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})$
- $\hat{\beta}_0 \sim \mathrm{N}(\beta_0, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}})$

I tillegg har vi for $s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{SSE}{n-2}$ at

- $\mathrm{E}(s^2) = \sigma^2$
- $(n-2)\frac{s^2}{\sigma^2} \sim \chi^2_{n-2}$ og uavhengig av $\hat{\beta}_1$ og $\hat{\beta}_0$

som ikke vises, men som leder til (bl.a.) $\frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{S_{xx}}} \sim t_{n-2}$

Egenskaper til $\hat{\beta}_0$ og $\hat{\beta}_1$ (forts)

- Modell: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Minste kvadraters estimatorer er gitt ved

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \quad \text{og} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

- Anta x_i'ene faste. Anta (bare) $E[\varepsilon_i] = 0$ som medfører

$$E[Y_i] = \beta_0 + \beta_1 x_i, \quad E[\bar{Y}] = \beta_0 + \beta_1 \bar{x}$$

Da er

$$\begin{aligned} E[\hat{\beta}_1] &= E\left[\frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] = \frac{\sum_{i=1}^n (x_i - \bar{x}) E[Y_i - \bar{Y}]}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})[\beta_0 + \beta_1 x_i - (\beta_0 + \beta_1 \bar{x})]}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1 \end{aligned}$$

og

$$E[\hat{\beta}_0] = E[\bar{Y} - \hat{\beta}_1 \bar{x}] = E[\bar{Y}] - E[\hat{\beta}_1]\bar{x} = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0$$

- Merk: Kun brukt at $E[\varepsilon_i] = 0$! - og hverken konstant varians, uavhengighet eller normalfordeling

Egenskaper til $\hat{\beta}_0$ og $\hat{\beta}_1$ - forts

- Anta nå også $V[\varepsilon_i] = \sigma^2$ samt at ε_i-ene er uavhengige
- Har

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- Da er

$$\begin{aligned} V[\hat{\beta}_1] &= V\left[\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2 V[Y_i]}{[\sum_{i=1}^n (x_i - \bar{x})^2]^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2}{[\sum_{i=1}^n (x_i - \bar{x})^2]^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

- Kan også vise $V[\hat{\beta}_0] = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right]$ (Neste slide)

$V(\hat{\beta}_0)$ samt en sammenheng

Proposisjon: $\bar{Y}$ og $\hat{\beta}_1$ er generelt ukorrelerte og uavhengige hvis Y_i-ene er uavhengige og normalfordelte.

Bevis: Merk. kan skrive $\hat{\beta}_1 = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) Y_i$ siden $\sum_{i=1}^{n} \bar{Y}(x_i - \bar{x}) = 0$.

$$\begin{aligned}
\mathrm{Cov}(\bar{Y}, \hat{\beta}_1) &= \tfrac{1}{nS_{xx}} \mathrm{Cov}(\textstyle\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} (x_i - \bar{x}) Y_i) \\
&= \tfrac{1}{nS_{xx}} \textstyle\sum_{i=1}^{n} \mathrm{Cov}(Y_i, (x_i - \bar{x}) Y_i) \qquad \text{(Se Kap. 6)} \\
&= \tfrac{1}{nS_{xx}} \textstyle\sum_{i=1}^{n} (x_i - \bar{x}) \mathrm{Cov}(Y_i, Y_i) = \tfrac{\sigma^2}{nS_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) = 0
\end{aligned}$$

Hvis Y_i-ene er simultant normalfordelt impliserer ukorrelerthet at de også er uavhengige.

Vi får da $\mathrm{V}(\hat{\beta}_0) = \mathrm{V}(\bar{Y} - \hat{\beta}_1 \bar{x}) = \mathrm{V}(\bar{Y}) + \bar{x}^2 \mathrm{V}(\hat{\beta}_1) = \sigma^2 (\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})$

Konfidensintervaller og hypotesetester for β_0 og β_1 (forts.)

- Det kan vi bruke til å vise at $\mathrm{E}(\hat{\beta}_0) = \beta_0$, $\mathrm{E}(\hat{\beta}_1) = \beta_1$, $\mathrm{V}(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2$ $\mathrm{V}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$.
- Da ϵ_i-ene er normalfordelt, er også Y_i-ene det, og siden $\hat{\beta}_0$ og $\hat{\beta}_1$ er lineærkombinasjoner av Y_i-ene, må $\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2\right)$ og $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$.
- La $S_e^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}$
- Det kan vises at $\hat{\beta}_0$ og $\hat{\beta}_1$ er uavhengige av S_e^2, samt at $\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2_{n-2}$.
- Dermed er

$$\frac{\hat{\beta}_0 - \beta_0}{S_{\hat{\beta}_0}} \sim t_{n-2} \quad \text{og} \quad \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \sim t_{n-2},$$

  der $S_{\hat{\beta}_0} = S_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$ og $S_{\hat{\beta}_1} = \frac{S_e}{\sqrt{S_{xx}}}$.
- Dette kan vi bruke til å lage konfidensintervaller for og teste hypoteser vedrørende β_0 og β_1 (som regel er vi mest interessert i β_1).

Konfidensintervaller for β_0 og β_1 (forts.)

- For $j = 0, 1$ har vi:

$$\mathrm{P}\left(-t_{\alpha/2,n-2} \leq \frac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}} \leq t_{\alpha/2,n-2}\right) = 1 - \alpha$$

$$\Rightarrow \quad \mathrm{P}\left(\hat{\beta}_j - t_{\alpha/2,n-2} s_{\hat{\beta}_j} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2,n-2} s_{\hat{\beta}_j}\right) = 1 - \alpha.$$

- Et $100 \cdot (1 - \alpha)\%$ konfidensintervall for β_j er da gitt ved

$$\hat{\beta}_j \pm t_{\alpha/2,n-2} s_{\hat{\beta}_j}.$$

Eksempel

CO_2-konsentrasjon og vekst av furutrær.

Eksempel

Fødselsvekt og lengden av svangerskap.

Hypotesetester vedrørende β_0 og β_1

- Vi kan være interessert i å teste hypoteser av typen
  1. $H_0 : \beta_j \leq \beta_{j0}$ mot $H_a : \beta_j > \beta_{j0}$
  2. $H_0 : \beta_j \geq \beta_{j0}$ mot $H_a : \beta_j < \beta_{j0}$
  3. $H_0 : \beta_j = \beta_{j0}$ mot $H_a : \beta_j \neq \beta_{j0}$

  for en gitt verdi av β_{j0} (som oftest $\beta_{j0} = 0$).
- Vi bruker testobservatoren

$$T = \frac{\hat{\beta}_j - \beta_{j0}}{S_{\hat{\beta}_j}}.$$

- Vi forkaster H_0 ved signifikansnivå α dersom
  1. $T \geq t_{\alpha,n-2}$
  2. $T \leq -t_{\alpha,n-2}$
  3. $T \leq -t_{\alpha/2,n-2}$ eller $T \geq t_{\alpha/2,n-2}$.
- Med observert $T = t$ og T_{n-2} en t-fordelt variabel med $n-2$ frihetsgrader fås p-verdier
  1. $\mathrm{P}(T_{n-2} \geq t)$
  2. $\mathrm{P}(T_{n-2} \leq t)$
  3. $\mathrm{P}(T_{n-2} \leq -|t|) + \mathrm{P}(T_{n-2} \geq |t|)$

Kvadratsummer

- Husk at $SST = SSE + SSR$, der $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$, $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ og $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$.
- Dette minner om enveis ANOVA, og faktisk kan vi teste $H_0 : \beta_1 = 0$ mot $H_a : \beta_1 \neq 0$ med den samme F-testen.
  Vi kan lage følgende variansanalysetabell:

Kilde til var.	Df	SS	MS	f
Regresjon	1	SSR	$MSR = \frac{SSR}{1} = SSR$	$\frac{MSR}{MSE} = \frac{SSR}{SSE/(n-2)}$
Feil	$n-2$	SSE	$MSE = \frac{SSE}{n-2}$	
Total	$n-1$	SST		

- F-testen over er da ekvivalent med den tosidige t-testen vi har sett på tidligere.

Eksempel

CO_2-konsentrasjon og vekst av trær.

Eksempel

Fødselsvekt og svangerskapslengde.

Konfidensintervall for $\mu_{Y|x^*} = \mathrm{E}(Y|x^*)$

- Vi vil estimere $\mu_{Y|x^*} = \mathrm{E}(Y|x^*) = \beta_0 + \beta_1 x^*$ for en gitt verdi x^*, samt lage et konfidensintervall for denne.
- En estimator for $\mu_{Y|x^*}$ er

$$\hat{\mu}_{Y|x^*} = \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*.$$

- Denne er normalfordelt med forventning $\mathrm{E}(\hat{\mu}_{Y|x^*}) = \beta_0 + \beta_1 x^* = \mathrm{E}(Y|x^*)$, altså forventningsrett, og har varians $\sigma^2_{\hat{Y}} = \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right)$.

Variansen til $\hat{Y}$ finner vi ved

$$\begin{aligned} \mathrm{V}(\hat{Y}) &= \mathrm{V}(\hat{\mu}_{Y|x^*}) = \mathrm{V}(\hat{\beta}_0 + \hat{\beta}_1 x^*) = \mathrm{V}(\bar{Y} + (x^* - \bar{x})\hat{\beta}_1) \\ &= \mathrm{V}(\bar{Y}) + \mathrm{V}((x^* - \bar{x})\hat{\beta}_1) = \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right) \end{aligned}$$

Konfidensintervall for $\mu_{Y|x^*}$, forts

- Da er $\frac{\hat{Y}-\mu_{Y|x^*}}{s_{\hat{Y}}} = \frac{\hat{\beta}_0+\hat{\beta}_1 x^*-(\beta_0+\beta_1 x^*)}{s\sqrt{\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}} \sim t_{n-2}$.
- Et $100 \cdot (1-\alpha)\%$ konfidensintervall for $\mu_{Y|x^*}$ er dermed gitt ved

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} s \sqrt{\frac{1}{n} + \frac{(x^*-\bar{x})^2}{S_{xx}}}.$$

Prediksjonsintervall for ny Y med $x = x^*$

- Vi vil nå lage et prediksjonsintervall for en ny Y med $x = x^*$, altså et intervall som med stor sannsynlighet inneholder den nye Y-en.
- Vi har $Y = \beta_0 + \beta_1 x^* + \varepsilon$, med $\varepsilon \sim N(0, \sigma^2)$.
- Vi kan anta at den nye Y-en er uavhengig av tidligere observasjoner $Y_1, \ldots, Y_n$.
- Vi predikerer Y med $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$.
- Prediksjonsfeilen er da $Y - \hat{Y}$, som er normalfordelt med forventning 0 og varians
  $\text{Var}(Y - \hat{Y}) = \text{Var}(Y) + \text{Var}(\hat{Y}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right)$.
- Det betyr at

$$\frac{Y - \hat{Y}}{S\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}.$$

Prediksjonsintervall for Y med $x = x^*$ (forts.)

Vi får

$$\mathrm{P}\left(-t_{\alpha/2,n-2} \leq \frac{Y-\hat{Y}}{s\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}} \leq t_{\alpha/2,n-2}\right) = 1-\alpha$$

$$\Rightarrow \quad \mathrm{P}\left(\hat{Y} - t_{\alpha/2,n-2}s\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}} \leq Y \leq \hat{Y} + t_{\alpha/2,n-2}s\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}\right)$$

$$= 1-\alpha$$

Et $100 \cdot (1-\alpha)\%$ prediksjonsintervall for Y er dermed gitt ved

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2}s\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}$$

der $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$.

Prediksjonsintervall for Y med $x = x^*$ (forts.)

- Det ekstra leddet i standardavviket gjør at prediksjonsintervallet for Y er mye videre enn konfidensintervallet for $\mathrm{E}(Y|x^*)$.
- Videre vil lengden til konfidensintervallet gå mot 0 når $n \to \infty$, mens lengden til prediksjonsintervallet går mot $2z_{\alpha/2}\sigma$.
- Det skyldes at Y fortsatt er usikker selv om parameterne i modellen er kjent.

Eksempel

CO_2 og vekst av trær.

Eksempel

Fødselsvekt og svangerskapslengde.

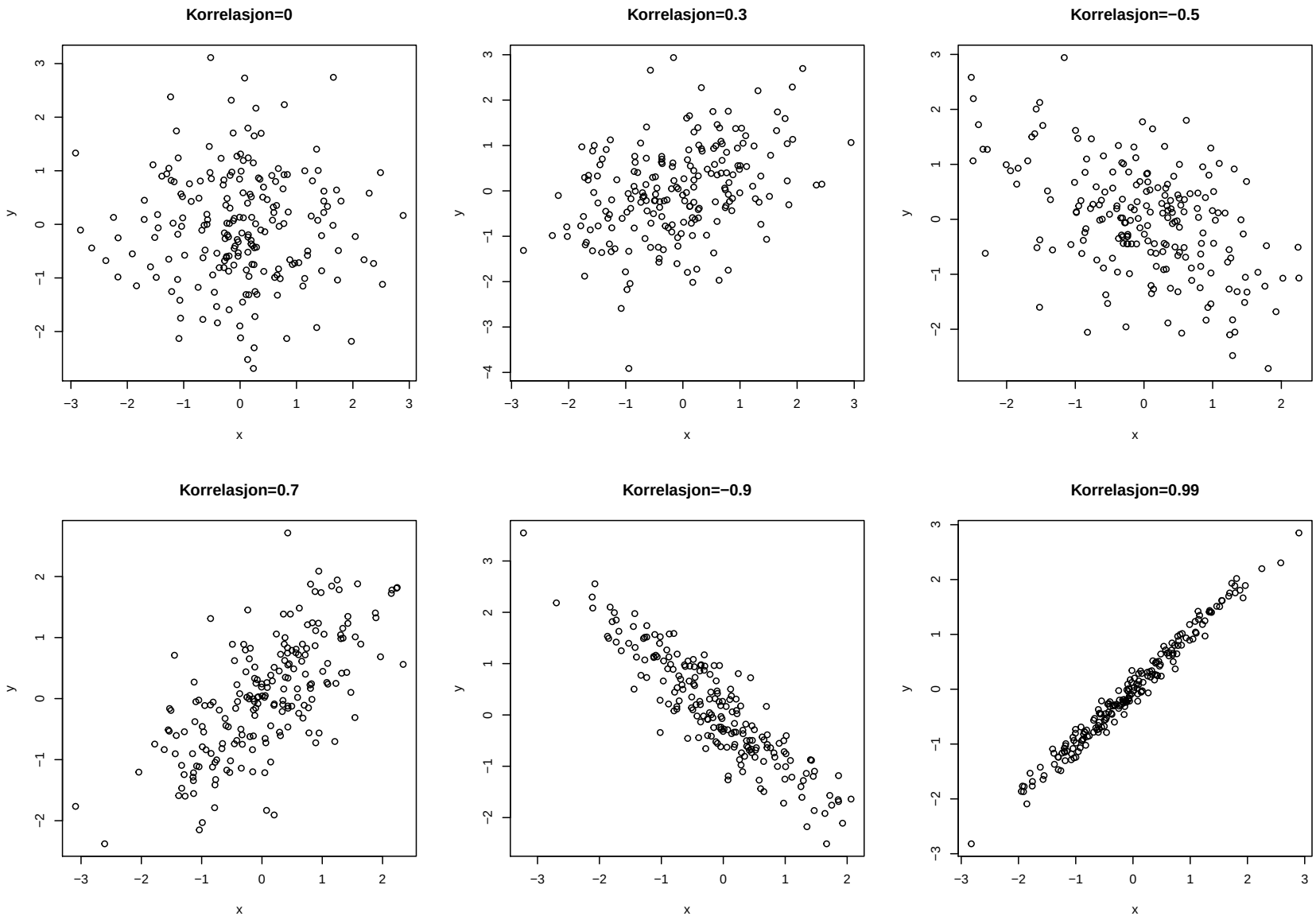
Avsnitt 12.5: Korrelasjonskoeffisienten

- Anta at vi har par av observasjoner (x_i, y_i), $i = 1, \ldots, n$, men nå er vi interessert i å se på samvariasjonen mellom x_i-ene og y_i-ene.
- Vi ser da på både x_i og y_i som observasjoner av stokastiske variabler X_i og Y_i, $i = 1, \ldots, n$.
- Ett mål på samvariasjonen er **den empiriske korrelasjonskoeffisienten**, gitt ved

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

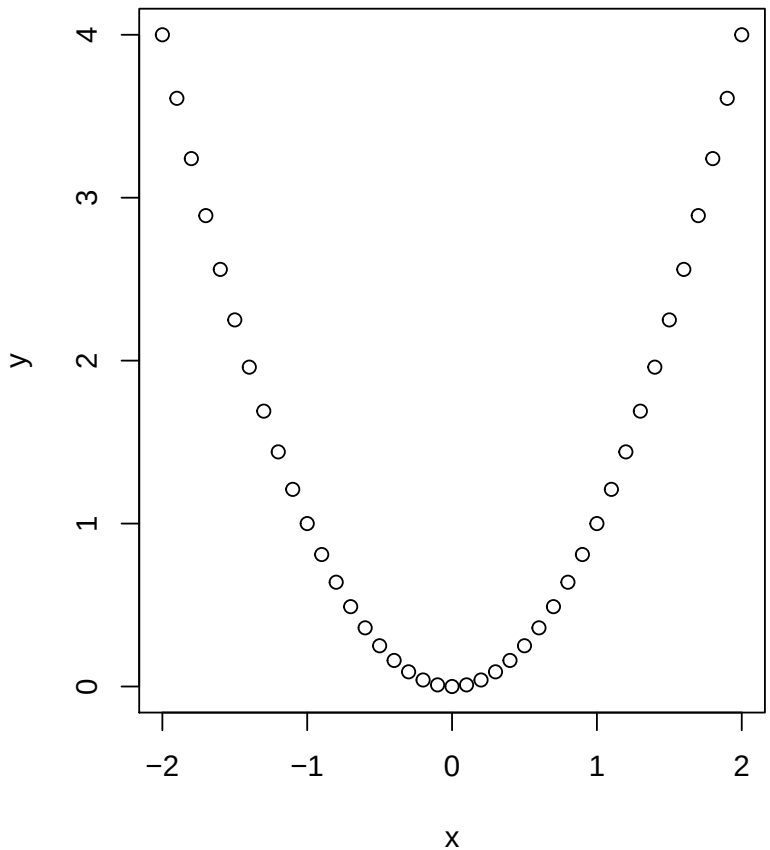
- Korrelasjonskoeffisienten har følgende egenskaper:
  1. r er symmetrisk i x og y
  2. r avhenger ikke av måleenhetene til x og y
  3. $-1 \leq r \leq 1$
  4. $r = 1$ ($r = -1$) hvis og bare hvis punktene (x_i, y_i) ligger på en rett linje med positivt (negativt) stigningstall.
  5. $r^2 = SSR/SST$.

Illustrasjon korrelasjon



Men: Korrelasjon måler spesifikt lineær sammenheng

Så f.eks. hvis x tar verdiene $-2, -1.9, \ldots, 1.9, 2$ og $y = x^2$
så er korrelasjonen $r = 0$



Korrelasjon og lineær regresjon

- Anta at $(X_1, Y_1), \ldots, (X_n, Y_n)$ er uavhengige par av stokastiske variabler fra en bivariat fordeling med korrelasjon ρ, dvs.

$$\rho = \text{Cor}(X_i, Y_i) = \frac{\text{Cov}(X_i, Y_i)}{\sigma_1 \sigma_2},$$

der $\text{Cov}(X_i, Y_i) = \text{E}((X_i - \mu_1)(Y_i - \mu_2))$, $\mu_1 = \text{E}(X_i)$, $\mu_2 = \text{E}(Y_i)$, $\sigma_1 = \text{sd}(X_i)$ og $\sigma_2 = \text{sd}(Y_i)$.

- Da er

$$\hat{\rho} = R = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

en moment (evt. plug-in) estimator for ρ.

Korrelasjon og lineær regresjon (forts.)

- Anta nå at $(X_1, Y_1), \ldots, (X_n, Y_n)$ er uavhengige par fra en bivariat normalfordeling med tetthet

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp^{-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_1}{\sigma_1}\right)^2+\left(\frac{y-\mu_2}{\sigma_2}\right)^2-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right)}.$$

- Da er

$$[Y|X=x] \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1), \sigma_2^2(1-\rho^2)\right).$$

- Vi har:

$$\mu_{Y|X=x} = \mathrm{E}(Y|X=x) = \underbrace{\mu_2 - \rho\frac{\sigma_2}{\sigma_1}\mu_1}_{\beta_0} + \underbrace{\rho\frac{\sigma_2}{\sigma_1}}_{\beta_1} x = \beta_0 + \beta_1 x.$$

- Det betyr at hvis observasjonene kommer fra en binormal fordeling, så er lineær regresjon riktig modell for å studere hvordan Y varierer med varierende $X = x$.

Hypotesetester for korrelasjonen til normalfordelte data

- Anta at $(X_1, Y_1), \ldots, (X_n, Y_n)$ er uavhengige par fra en bivariat normalfordeling.
- Vi vil nå teste hypoteser av typen
  1. $H_0 : \rho \leq \rho_0$ mot $H_a : \rho > \rho_0$
  2. $H_0 : \rho \geq \rho_0$ mot $H_a : \rho < \rho_0$
  3. $H_0 : \rho = \rho_0$ mot $H_a : \rho \neq \rho_0$.
- For $\rho_0 = 0$ kan vi bruke testobservatoren $T = \frac{R\sqrt{n-2}}{1-R^2} = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$.
- Ut fra formelen for $\mu_{Y|X=x}$ ser vi at når $\rho = \rho_0 = 0$, er $\beta_1 = 0$, og da er $T \sim t_{n-2}$.
- Vi får da følgende tester med signifikansnivå α:
  1. Forkast H_0 dersom $T \geq t_{\alpha,n-2}$
  2. Forkast H_0 dersom $T \leq -t_{\alpha,n-2}$
  3. Forkast H_0 dersom $T \leq -t_{\alpha/2,n-2}$ eller $T \geq t_{\alpha/2,n-2}$.

Tester for korrelasjonen til normalfordelte data (forts.)

- Når $\rho_0 \neq 0$, kan en i stedet bruke en testobservator basert på $V = \frac{1}{2} \ln\left(\frac{1+R}{1-R}\right)$.
- Det kan vises at $V \overset{\text{tiln.}}{\sim} N\left(\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{\sqrt{n-3}}\right)$.
- Vi kan dermed bruke testobservatoren

$$Z = \frac{V - \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{1/\sqrt{n-3}} \overset{\text{tiln.}}{\sim} N(0,1).$$

- Vi får da følgende tester med signifikansnivå α:
  1. Forkast H_0 dersom $Z \geq z_\alpha$
  2. Forkast H_0 dersom $Z \leq -z_\alpha$
  3. Forkast H_0 dersom $Z \leq -z_{\alpha/2}$ eller $Z \geq z_{\alpha/2}$.

Konfidensintervall for korrelasjonen til normalfordelte data

- Et tilnærmet $100 \cdot (1-\alpha)\%$ konfidensintervall for $\mu_V = \frac{1}{2} \ln\left(\frac{1+\rho}{1-\rho}\right)$ er gitt ved

$$\left(v - \frac{z_{\alpha/2}}{\sqrt{n-3}}, v + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right) = (v_L, v_U),$$

  der $v = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) = h(r)$.

- Vi har $r = h^{-1}(v) = \frac{e^{2v}-1}{e^{2v}+1}$.
- Dermed er et tilnærmet $100 \cdot (1-\alpha)\%$ konfidensintervall for ρ gitt ved

$$(h^{-1}(v_L), h^{-1}(v_U)) = \left(\frac{e^{2v_L}-1}{e^{2v_L}+1}, \frac{e^{2v_U}-1}{e^{2v_U}+1}\right).$$

Eksempel

Eks. 12.16 i boka (Leonardo da Vinci's mann)

Inferens for korrelasjon uten normalfordeling

Intervallet basert på $Z = \frac{\frac{1}{2}\log\left(\frac{1+R}{1-R}\right) - \frac{1}{2}\log\left(\frac{1+\rho_0}{1-\rho_0}\right)}{1/\sqrt{n-3}}$ fungerer ikke alltid bra uten normalfordelte data.

Alternativer er da

- Bootstrapping! (som vanlig)
- Spearman korrelasjon

Spearman-korrelasjon baseres på ranger til X_i-ene og Y_i-ene.

- Rangen til minste X_i er lik $R_i = 1$
- Rangen til neste minste X_i er lik $R_i = 2$
- Dette fortsetter til største X_i som gis rang $R_i = n$
- Tilsvarende angis rangene S_i til Y_i-ene

Spearman-korrelasjonen er da korrelasjonen for parrene (R_i, S_i):

$$r_{SP} = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(S_i - \bar{S})^2 \sum_{i=1}^{n}(R_i - \bar{R})^2}}$$

og måler egentlig monotonitet heller enn linearitet.

Avsnitt 12.6: Sjekk av modellantakelser

- I den lineære regresjonsmodellen

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{uif}{\sim} N(0, \sigma^2)$$

gjøres følgende antakelser, som bør undersøkes.

1. Sammenhengen er lineær: $\mathrm{E}(Y|x) = \beta_0 + \beta_1 x$.
2. Variansen er lik for alle verdier av x: $\mathrm{V}(\varepsilon_i) = \sigma^2$.
3. Feilleddene ε_i er normalfordelt.
4. Feilleddene ε_i er uavhengige.

- Vi skal se på måter å kontrollere de første tre antakelsene på, mens den siste blir behandlet i seinere kurs.
- Disse metodene brukes også for multiple lineære regresjonsmodeller (Avsnitt 12.7 - 12.9).

Sjekk av modellantakelser (forts.)

- Da β_0 og β_1 er ukjent, kan vi ikke undersøke feileddene $\varepsilon_i = Y_i - \beta_0 - \beta_1 x_i$ direkte, men ser i stedet på residualene

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad i = 1, \ldots, n.$$

- For å undersøke om sammenhengen er lineær kan en plotte e_i-ene mot x_i-ene eller mot $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ -ene.
- Det skal ikke være noe systematisk mønster i dette plottet i den forstand at man får øye på en kurvatur gjennom plottet

- Når modellantakelsene er oppfylt, så er

$$\mathrm{E}(Y_i - \hat{Y}_i) = 0 \quad \text{og} \quad \mathrm{V}(Y_i - \hat{Y}_i) = \sigma^2 \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right).$$

- Siden disse variansene avhenger av σ^2 og ikke er helt like benyttes ofte standardiserte residualer gitt ved

$$e_i^* = \frac{y_i - \hat{y}_i}{s\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}}}, \quad i = 1, \ldots, n.$$

Sjekk av modellantakelser (forts.)

- For å undersøke om variansen til feilleddene er lik, kan en (også) plotte e_i^*-ene mot $\hat{y}_i$-ene.
- Alternativt benyttes i R plottet av $\sqrt{|e_i^*|}$-ene mot $\hat{y}_i$
- Det skal ikke være noe systematisk mønster i dette plottet.

  i den forstand at spredningen øker (eller avtar) med $\hat{y}_i$.

- Antakelsen om normalfordeling kan en undersøke ved å lage et normalfordelingsplott av de standardiserte residualene.
- Hvis n er ganske stor, har det ikke så mye å si om en standardiserer residualene.

Eksempel

CO_2 og vekst av trær.

Eksempel

Svangerskapsvarighet og fødselsvekt

Konsekvenser av modellavvik

- Det er mest viktig å sjekke lineariteten $\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i$
- Hvis dette svikter i betydelig grad har vi tilpasset en ikke-adekvat modell

- Hvis så lineariteten holder rimelig bra, men variansen synes å endres med $\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i$ så har kan vi ikke stole på variansestimater for $\hat{\beta}_j$ pause og tester/konfidensintervall kan være misvisende
- Tilsvarende kan inntreffe hvis det er avhengighet mellom ε_i-ene (og Y_i-ene).

- Mindre avvik fra normalfordelte feil ε_i er ikke kritisk siden bare dette betyr at $t -$ *korreksjonen* ikke er helt adekvat.
- Men det er viktig å sjekke konsekvenser av *uteliggere*