STK1100 våren 2023

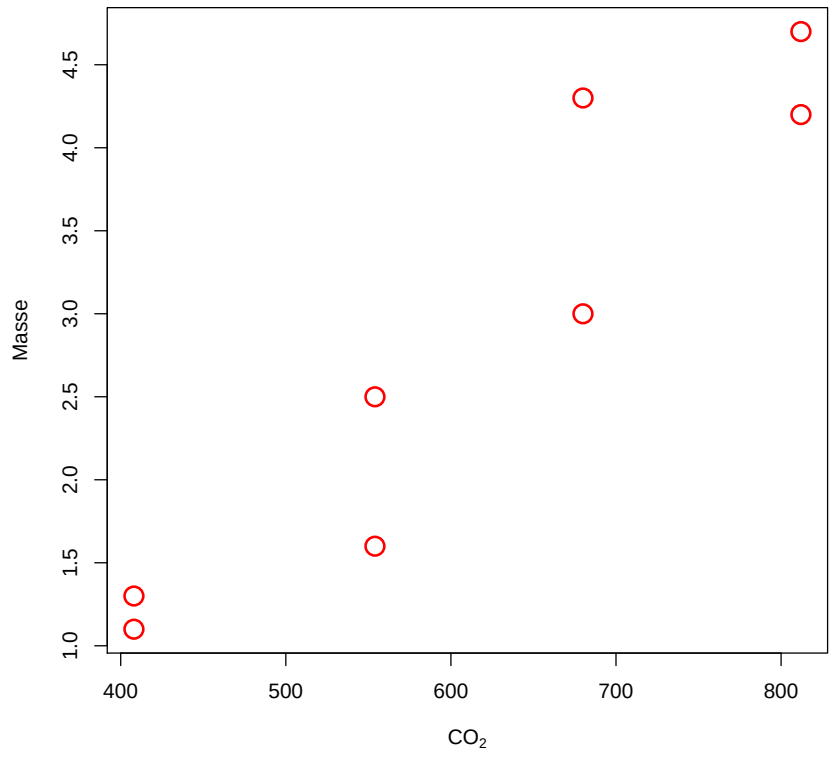
Lineær regresjon

Svarer til avsnitt 12.1-12.2

Matematisk institutt
Universitetet i Oslo

Eksempel: CO_2-konsentrasjon og vekst av furutrær

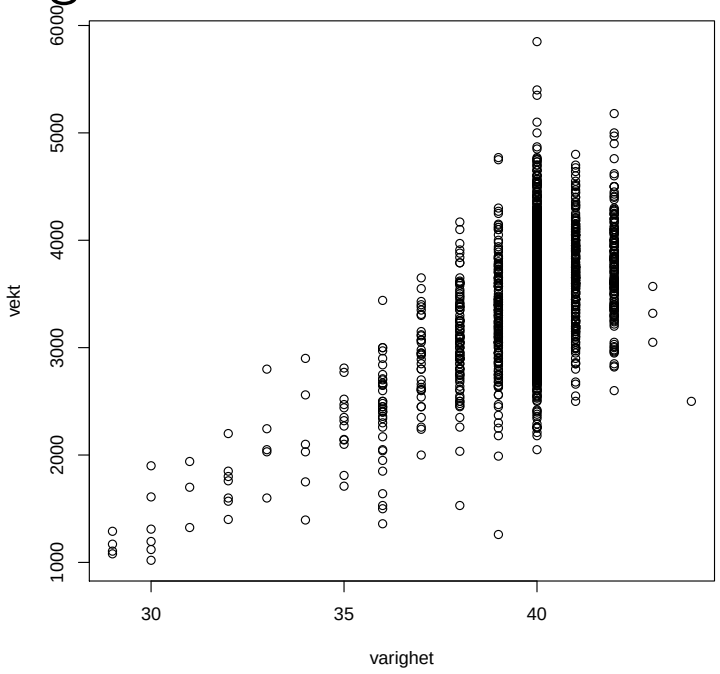
CO_2	Masse
408	1.1
408	1.3
554	1.6
554	2.5
680	3.0
680	4.3
812	4.2
812	4.7



Kan vi finne sammenhengen mellom CO_2-konsentrasjon og masse?

Eksempel: fødselsvekt og lengden av svangerskap

Figuren viser vekten til et utvalg av 2116 nyfødte gutter:



Kan vi finne en sammenheng som beskriver hvordan fødselsvekten henger sammen med varigheten 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?

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").

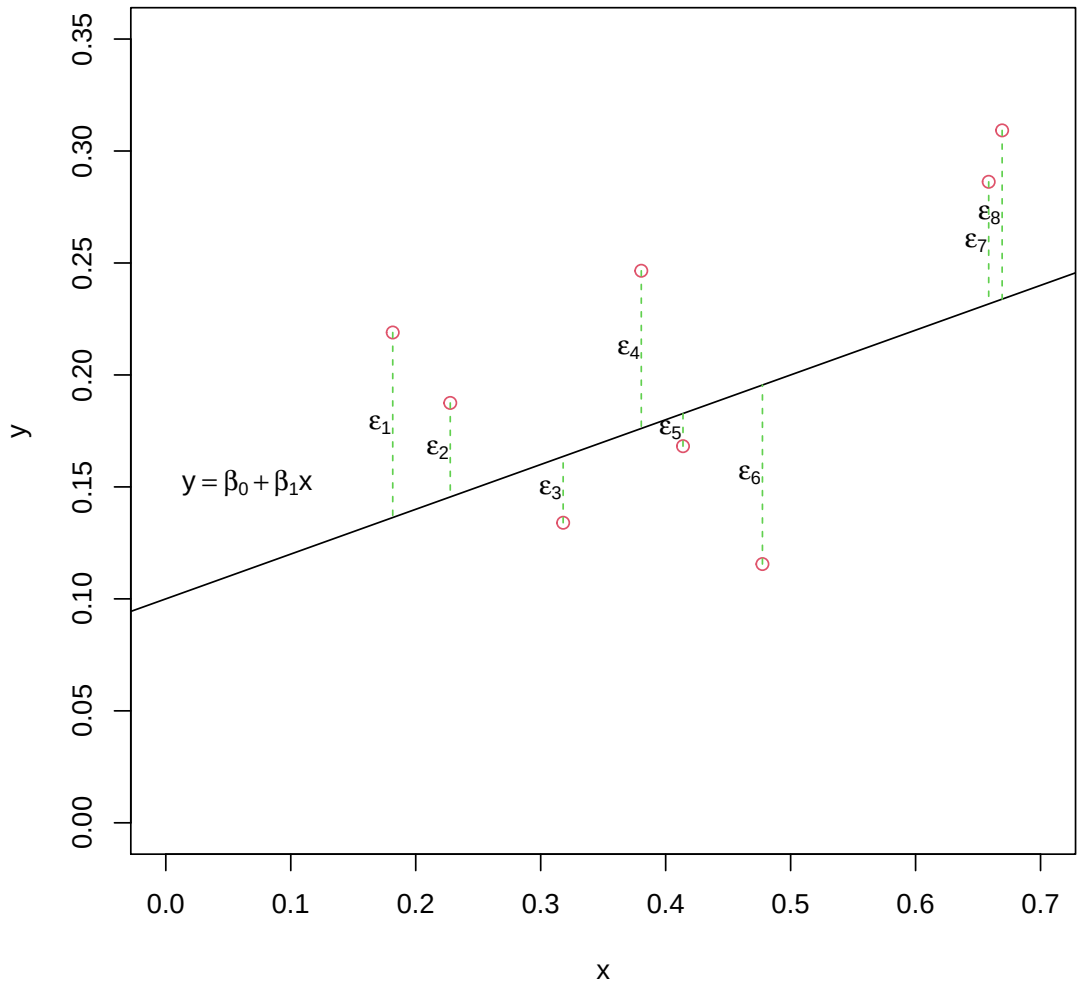
Lineær regresjon (forts.)

- Ofte er x_i den observerte verdien av den stokastisk variabelen 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**

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

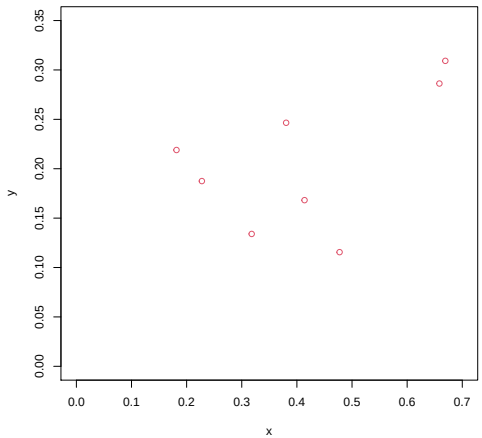
  der β_0 og β_1 er ukjente parametere.
- Noen ganger må vi transformere x_i-ene og/eller y_i-ene for å få en lineær sammenheng.
- Vi vil som regel anta at $\varepsilon_i \sim N(0, \sigma^2)$.

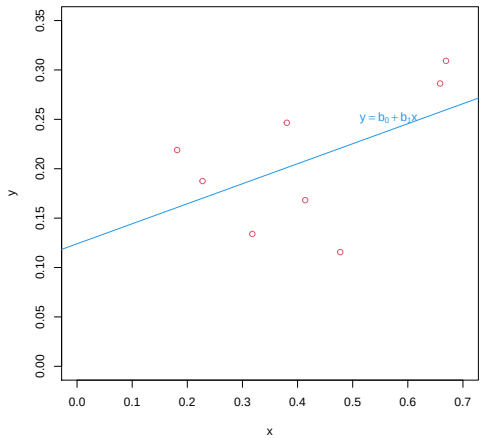
Lineær regresjon: illustrasjon

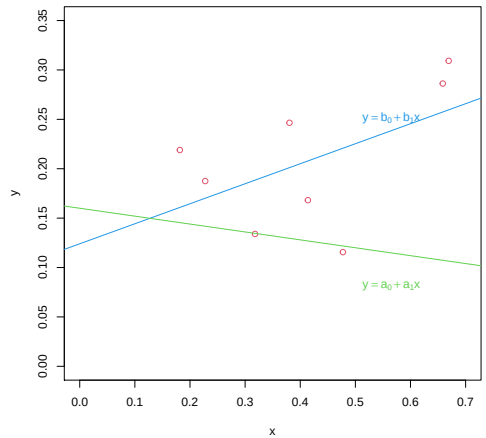


Minste kvadraters metode

- Når vi skal estimere modellparameterne β_0, β_1 og σ^2, vil vi prøve å finne den rette linja $b_0 + b_1 x$ som "passer best" til de observerte punktene (x_i, y_i).







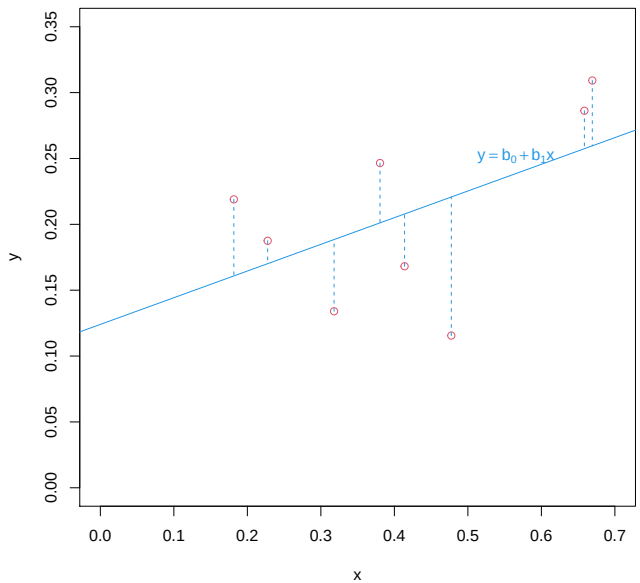
- Vi må da definere hva vi mener med "passer best".

Minste kvadraters metode (forts.)

- La

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

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



- Minste kvadraters estimater får en ved å minimere $f(b_0, b_1)$ m.h.p. b_0 og b_1.

Minste kvadraters metode (forts.)

- Minste kvadraters estimater er gitt ved

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

- Minste kvadraters metode forutsetter at ε_i-ene er uif med $\mathrm{E}(\varepsilon_i) = 0$ og $\mathrm{V}(\varepsilon_i) = \sigma^2$, men ikke nødvendigvis at de er normalfordelt.
- Hvis $\varepsilon_i \sim N(0, \sigma^2)$, så er minste kvadraters estimater de samme som maksimum likelihood-estimatene.

Residualer

- 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 er

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

- Denne kan brukes til å estimere σ^2:

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

Kvadratsummer og variasjon

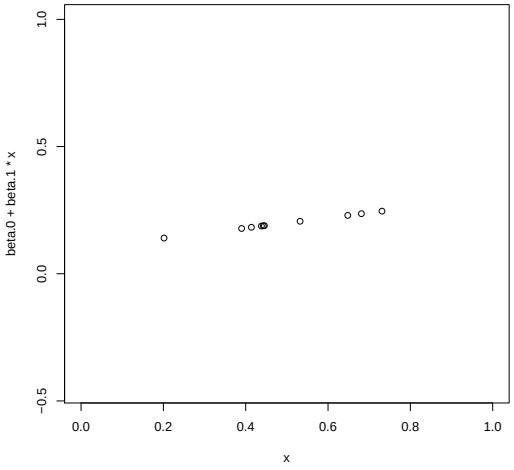
- Estimatoren S^2 får en ved å bytte ut y_i-er med Y_i-er, og denne er forventningsrett for σ^2.
- La nå

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad \text{og} \quad SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

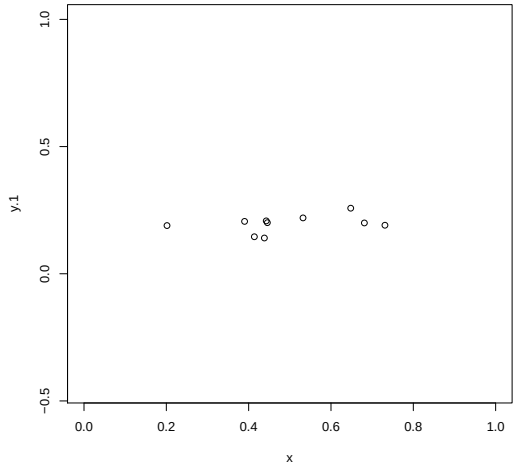
- Da kan det vises at $SST = SSR + SSE$.
- SSR og SSE kan relateres til hvor mye av variasjonen i y_i-ene som kan forklares av modellen og hvor mye som er tilfeldig støy.
- Et mål for hvor mye av variasjonen som kan forklares av modellen er

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

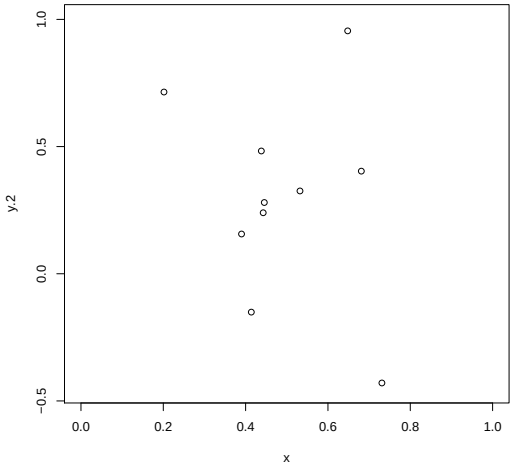
Kvadratsummer og variasjon (forts.)



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.

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.
- Det er ikke alltid $x = 0$ er en relevant verdi, og da kan en i stedet formulere modellen som $Y_i = \beta_0 + \beta_1(x_i - x_0) + \varepsilon_i$, 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.