

Kapitel 11

Lineær regression

Peter Tibert Stoltze
stat@peterstoltze.dk

Elementær statistik
F2011

1 / 1

Indledning

- ▶ Vi modellerer en afhængig variabel (responset) på baggrund af en uafhængig variabel (stimulus), som vi i princippet kan vælge frit
- ▶ Vi vælger en simpel model:

$$y = ax + b + \varepsilon$$

hvor a er hældningen, b er skæringen med y -aksen og ε er den del af responset, der ikke kan forklares med modellen

- ▶ Vi antager at $\varepsilon \sim N(0, \sigma^2)$ og uafhængige
- ▶ Vi bestemmer estimater for a og b ud fra data og kan herefter forudsige (prediktere) værdier af y som

$$\hat{y} = \hat{a}x + \hat{b}$$

2 / 1

Liniens ligning

- ▶ En linie kan defineres ved to punkter eller ét punkt og en hældning
- ▶ Liniens ligning i et almindeligt (x, y) koordinat-system er givet ved

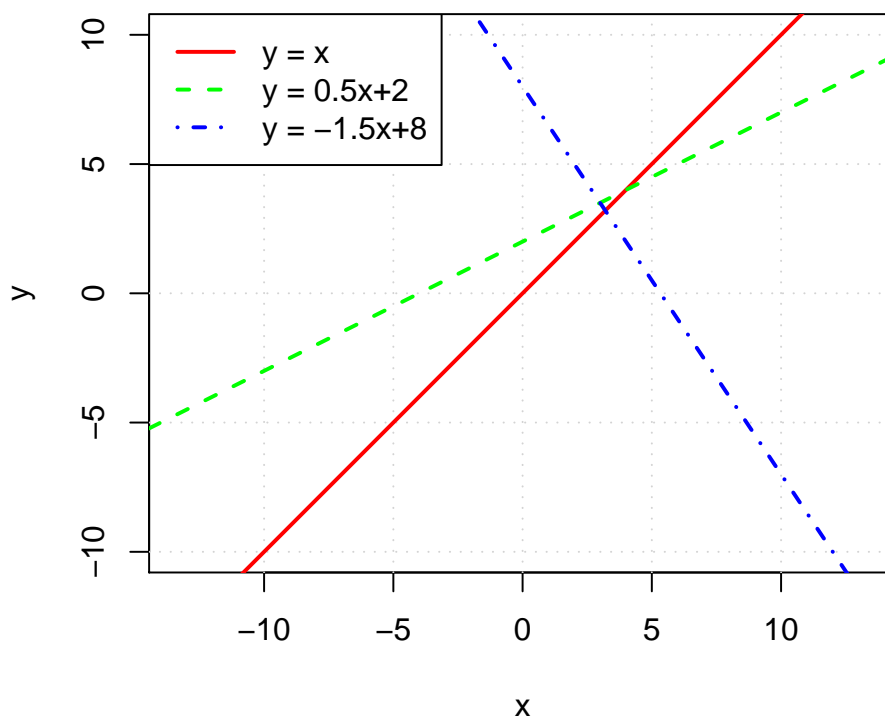
$$y = ax + b$$

hvor a er hældning og b er skæring med y -aksen

- ▶ Regressionen af x på y går altså igennem punkterne $(0, b)$ og $(1, a + b)$

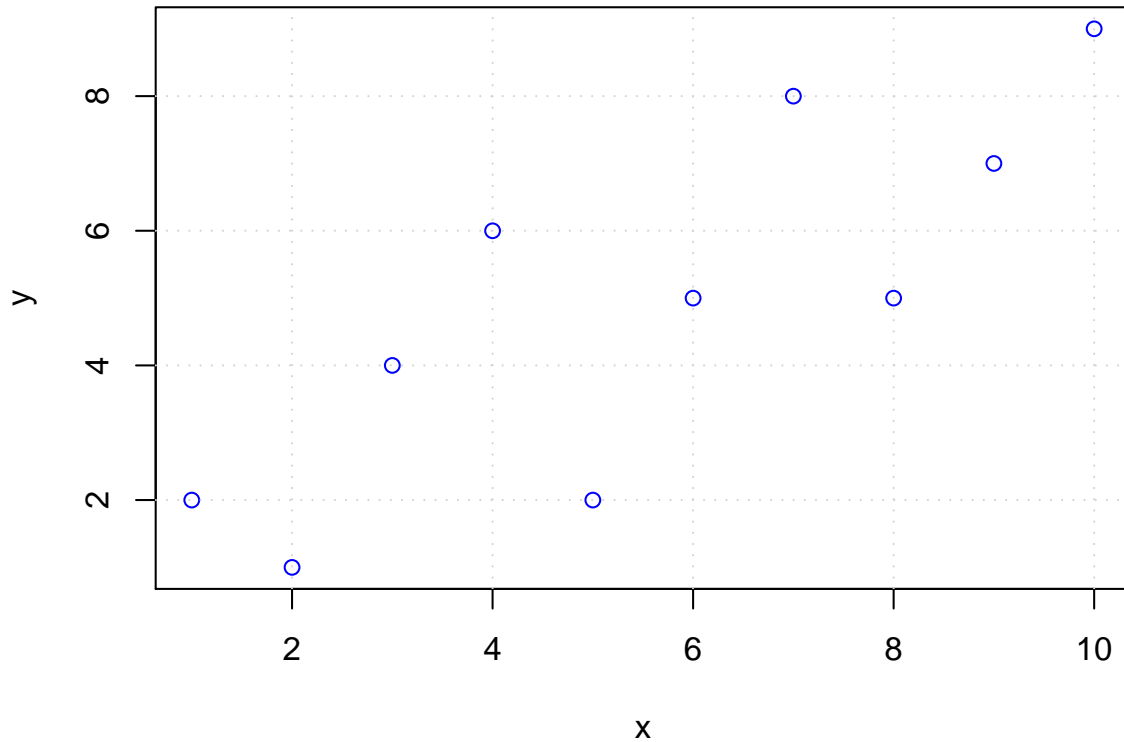
3/1

Liniens ligning



4/1

Eksempel med $\rho = 0,81$ og $n = 10$



5/1

Bestemmelse af regressionslinien

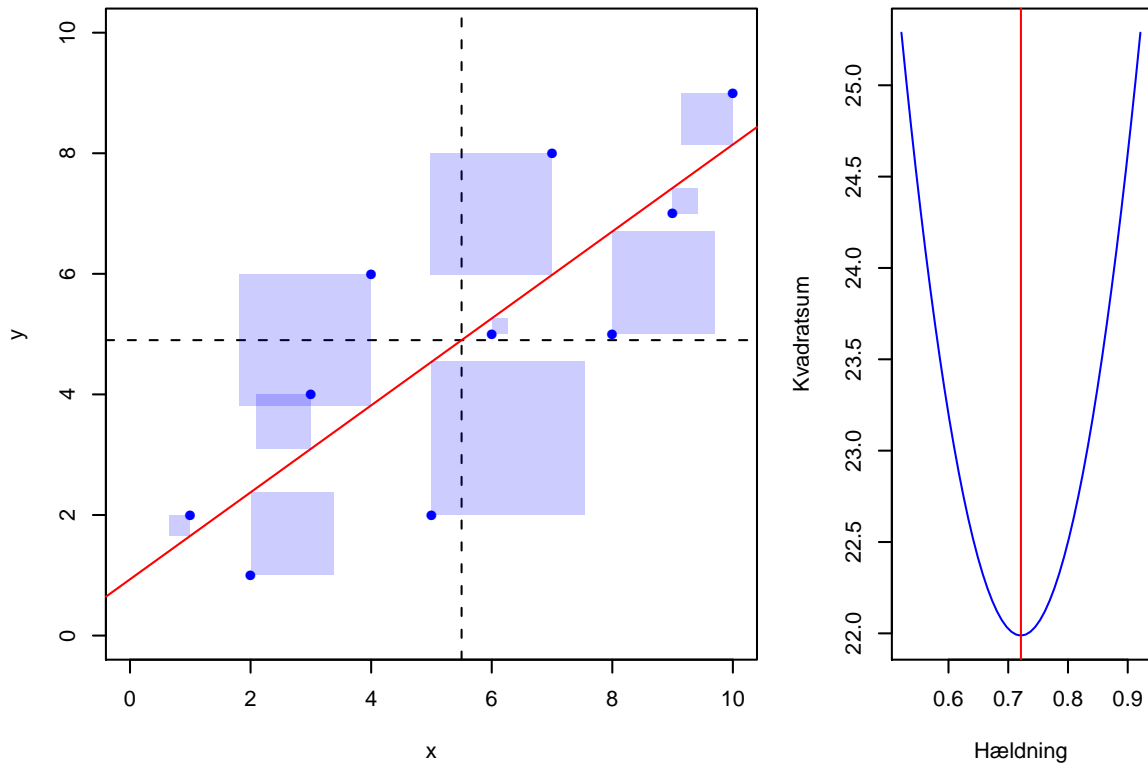
- ▶ Vi benytter **mindste kvadraters metode** så summen af kvadraterne på de lodrette afstande fra datapunkterne til regressionslinien bliver så lille som muligt:

$$\sum (y - \hat{y})^2 = \sum (y - ax - b)^2 = \min!$$

- ▶ Det viser sig, at a og b skal vælges så regressionslinien passerer gennem (\bar{x}, \bar{y}) , der også kaldes datamaterialets massemidtpunkt

6/1

Bestemmelse af regressionslinien (fortsat)



7 / 1

Bestemmelse af regressionslinien (fortsat)

- Man kan vise, at summen af de kvadratiske afvigelser er minimal for

$$a = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$

henholdsvis

$$b = \frac{\sum_{i=1}^n y_i - a \sum_{i=1}^n x_i}{n}$$

8 / 1

Beregning — eksempel

i	x	y	x^2	y^2	xy
1	1	2	1	4	2
2	2	1	4	1	2
3	3	4	9	16	12
4	4	6	16	36	24
5	5	2	25	4	10
6	6	5	36	25	30
7	7	8	49	64	56
8	8	5	64	25	40
9	9	7	81	49	63
10	10	9	100	81	90
Σ	55	49	385	305	329

9 / 1

Beregning — eksempel (fortsat)

- Først beregner vi hældningen a

$$\hat{a} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} = \frac{10 \cdot 329 - 55 \cdot 49}{10 \cdot 385 - 55^2} = 0,721$$

- Dernæst skæringen b

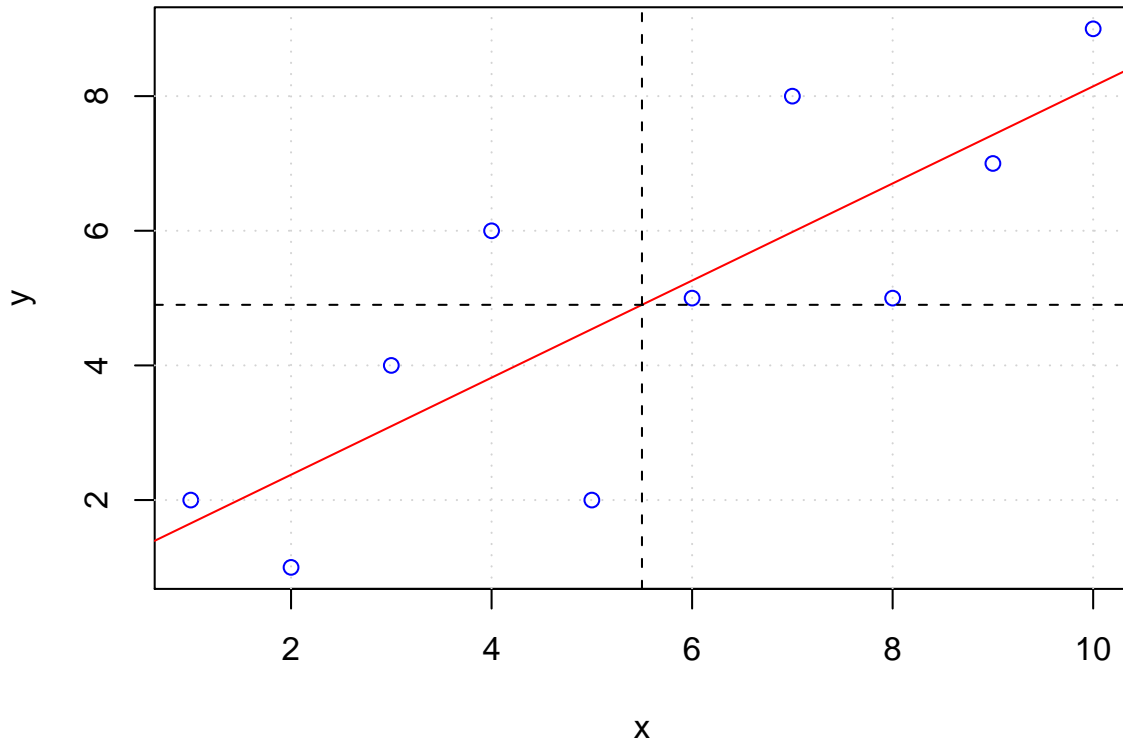
$$\hat{b} = \frac{\sum_{i=1}^n y_i - a \sum_{i=1}^n x_i}{n} = \frac{49 - 0,721 \cdot 55}{10} = 0,933$$

- Den samlede regressionslinie hedder derfor

$$\hat{y} = 0,721x + 0,933$$

10 / 1

Beregning — eksempel (fortsat)



11 / 1

Den omvendte regression

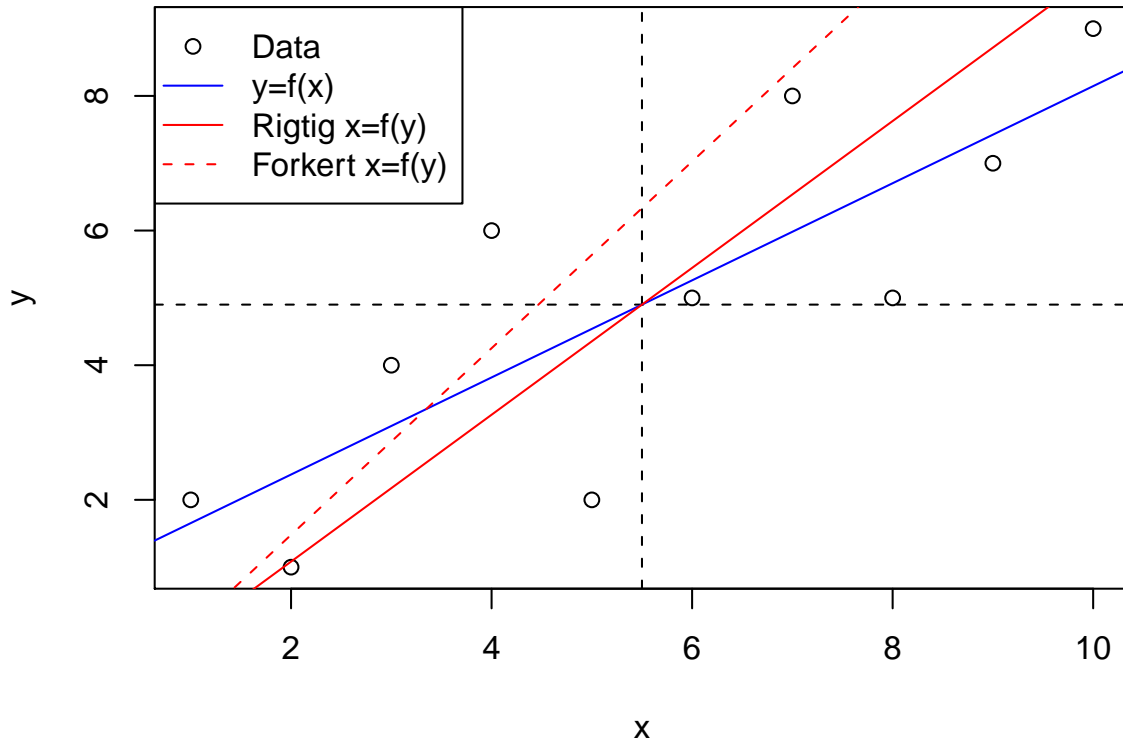
- ▶ Det er fristende at flytte lidt rundt på ligningen:

$$y = ax + b \Leftrightarrow x = \frac{1}{a}y - \frac{b}{a}, \quad a \neq 0$$

- ▶ Beregningen er sådan set lovlig, men højresiden udtrykker ikke regressionen af y på x
- ▶ For at få den omvendte regression må man lade x og y bytte roller og starte beregningen af a og b helt forfra...

12 / 1

Den omvendte regression (fortsat)



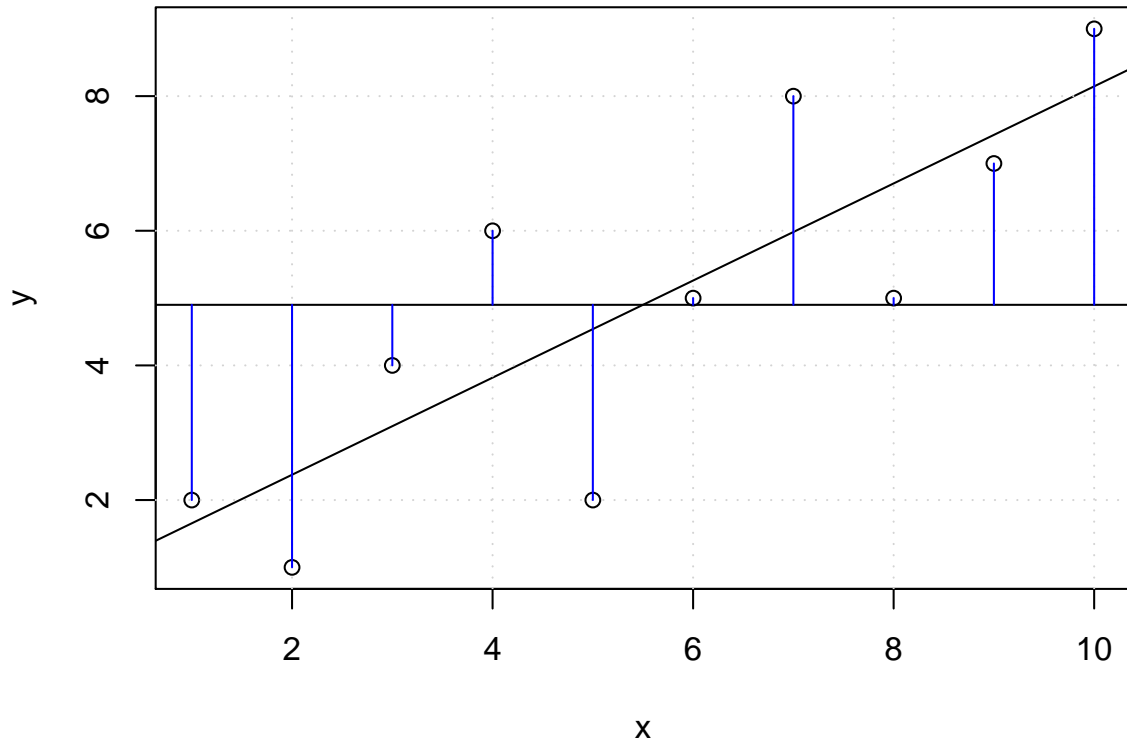
13 / 1

Variationskilder

- ▶ De observerede værdier af y har en varians, der er sammensat af et bidrag fra
 - ▶ model (variation på x eller $\hat{y} - \bar{y}$) og
 - ▶ rest (residualen $y - \hat{y}$)
- ▶ Ved at se på de to bidrag i forhold til hinanden kan vi vurdere, hvor sikkert vi kan forudsige y ud fra x

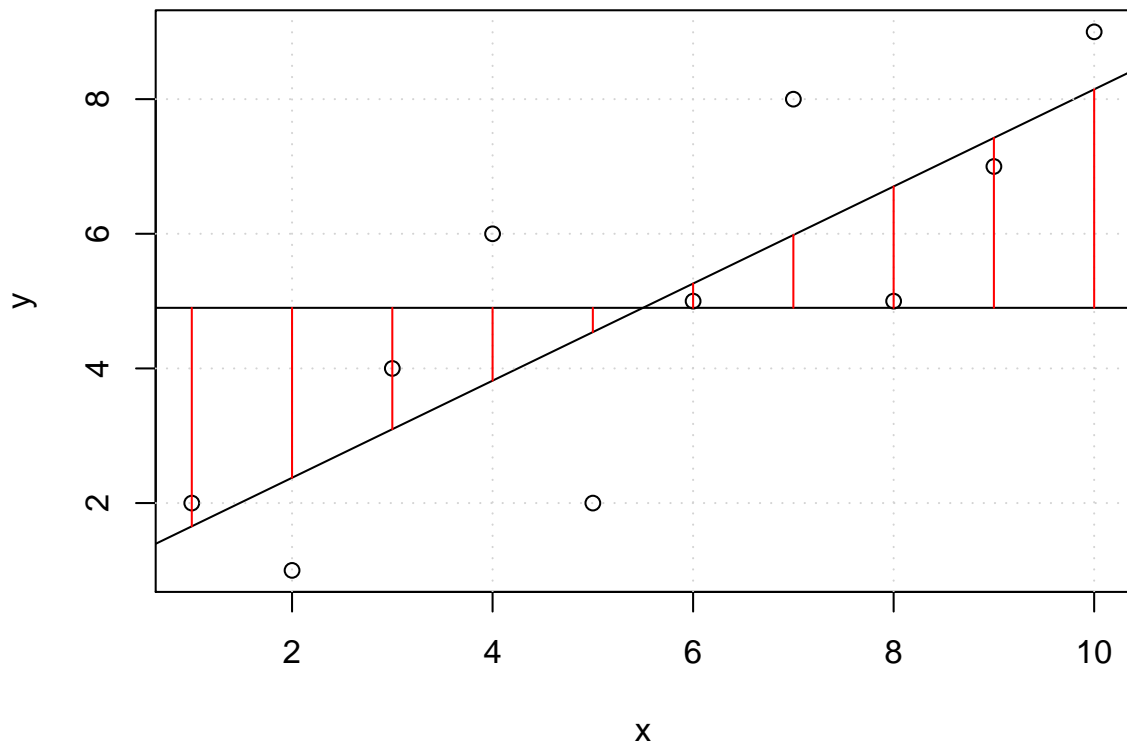
14 / 1

Variationskilder: Totalvarians



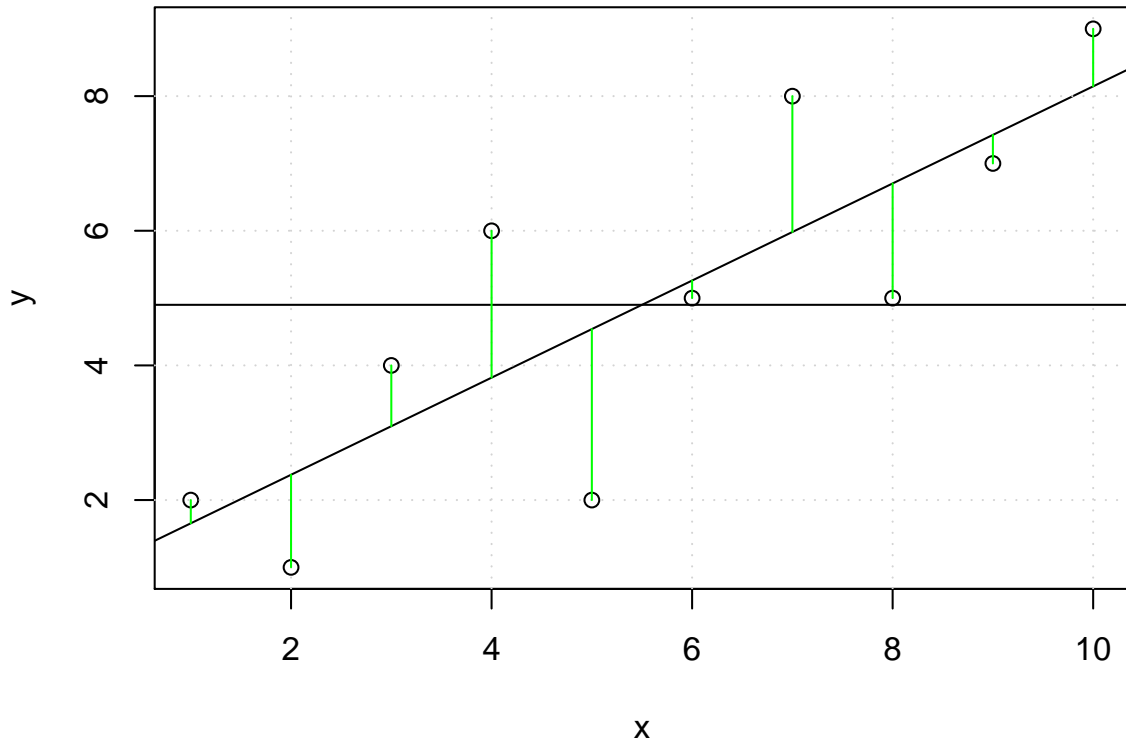
15 / 1

Variationskilder: Modelvarians



16 / 1

Variationskilder: Restvarians



17 / 1

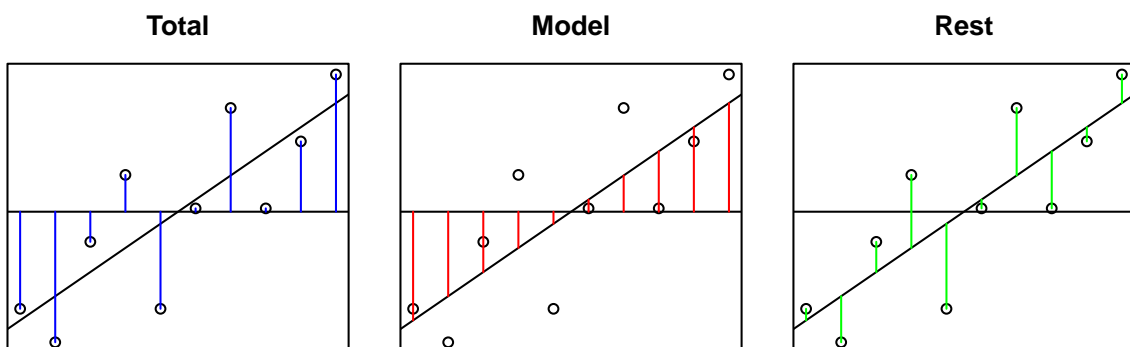
Variationskilder: Samlet

- ▶ Den totale varians kan henføres til model og rest

$$y_i - \bar{y} = (y_i - \hat{y}) + (\hat{y} - \bar{y}) \Rightarrow \dots \Rightarrow s_y^2 = s_{\hat{y}}^2 + s_e^2$$

- ▶ Andelen af total varians, der kan forklares med modellen, er

$$r^2 = \frac{s_{\hat{y}}^2}{s_y^2}$$



18 / 1

Variationskilder — eksempel

x	y	\hat{y}	$y - \bar{y}$	$(y - \bar{y})^2$	$\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$	$y - \hat{y}$	$(y - \hat{y})^2$
1	2	1,65	-2,90	8,41	-3,25	10,53	0,35	0,12
2	1	2,38	-3,90	15,21	-2,52	6,37	-1,38	1,89
3	4	3,10	-0,90	0,81	-1,80	3,25	0,90	0,82
4	6	3,82	1,10	1,21	-1,08	1,17	2,18	4,76
5	2	4,54	-2,90	8,41	-0,36	0,13	-2,54	6,45
6	5	5,26	0,10	0,01	0,36	0,13	-0,26	0,07
7	8	5,98	3,10	9,61	1,08	1,17	2,02	4,07
8	5	6,70	0,10	0,01	1,80	3,25	-1,70	2,90
9	7	7,42	2,10	4,41	2,52	6,37	-0,42	0,18
10	9	8,15	4,10	16,81	3,25	10,53	0,85	0,73
Σ			0	64,90	0	42,91	0	21,99

19 / 1

Variationskilder — eksempel (fortsat)

- ▶ Bemærk at $\sum(y - \bar{y})^2 = \sum(\hat{y} - \bar{y})^2 + \sum(y - \hat{y})^2$
- ▶ Vi kan nu beregne alle tre varianser:

$$s_{total}^2 = \frac{\sum(y - \bar{y})^2}{n - 1} = \frac{64,90}{10 - 1} = 7,21$$

$$s_{model}^2 = \frac{\sum(\hat{y} - \bar{y})^2}{n - 1} = \frac{42,91}{10 - 1} = 4,77$$

$$s_{rest}^2 = \frac{\sum(y - \hat{y})^2}{n - 1} = \frac{21,99}{10 - 1} = 2,44$$

- ▶ Andelen af den totale variation, der kan forklares ved modellen er

$$r^2 = \frac{s_{model}^2}{s_{total}^2} = \frac{4,77}{7,21} = 0,66$$

20 / 1

Variationskilder - ekstra

- ▶ Med de allerede beregnede kvadratsummer kan vi med en variansanalyse teste om modellen er signifikant — i vores tilfælde svarende til om hældningen er signifikant forskellig fra nul

Kilde	DF	SAK	MS	F
Model	1	42,91	42,91	15,61
Rest	8	21,99	2,75	
Total	9	64,90		

- ▶ Teststørrelsen vurderes i en F -fordeling med (1,8) frihedsgrader og vi finder

$$P(F > 15,61) = 0,0042$$

hvilket betyder at hældningen er signifikant forskellig fra nul ($p = 0,42$ pct.)

21 / 1

Multipel lineær regression

- ▶ Samme princip som for simpel lineær regression, nemlig mindste kvadrater
- ▶ Beregning foretages altid med computer og passende software (Excel *kan* bruges)
- ▶ Pas på med fortolkning af parameterestimerer hvis de forklarende variable er korrelerede

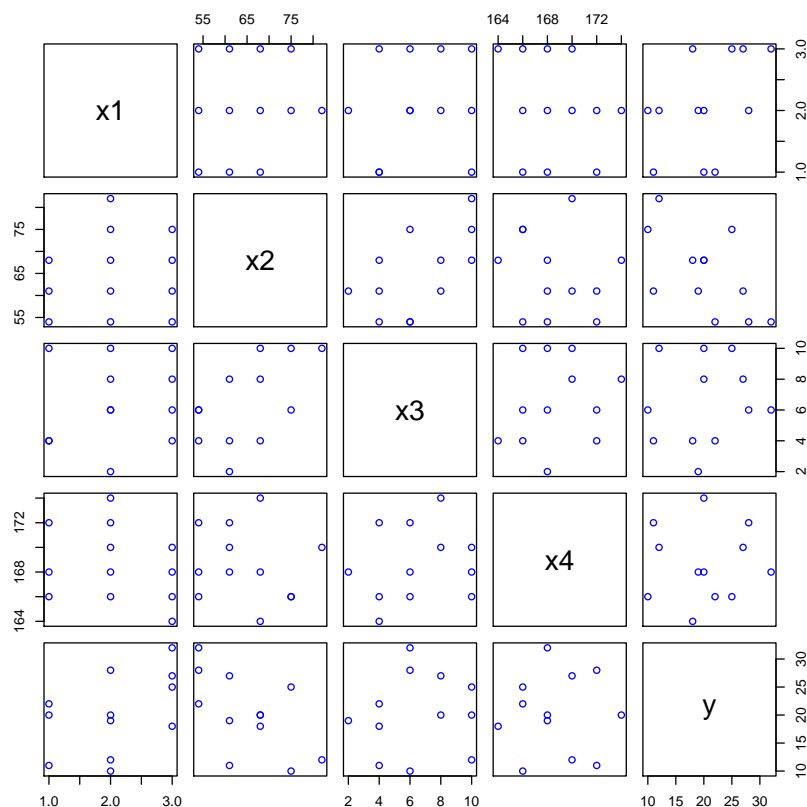
22 / 1

Multipel lineær regression — eksempel

Uddannelse	Alder	Mdr. Efter	Højde	Score
x_1	x_2	x_3	x_4	y
3	54	6	168	32
3	61	8	170	27
3	68	4	164	18
3	75	10	166	25
2	54	6	172	28
2	61	2	168	19
2	68	8	174	20
2	75	6	166	10
2	82	10	170	12
1	54	4	166	22
1	61	4	172	11
1	68	10	168	20

23 / 1

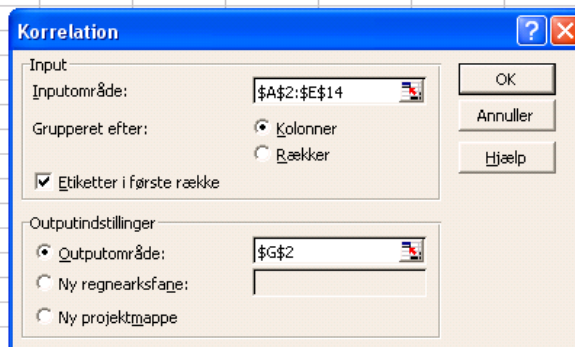
Multipel lineær regression — eksempel



24 / 1

Multipel lineær korrelation — eksempel

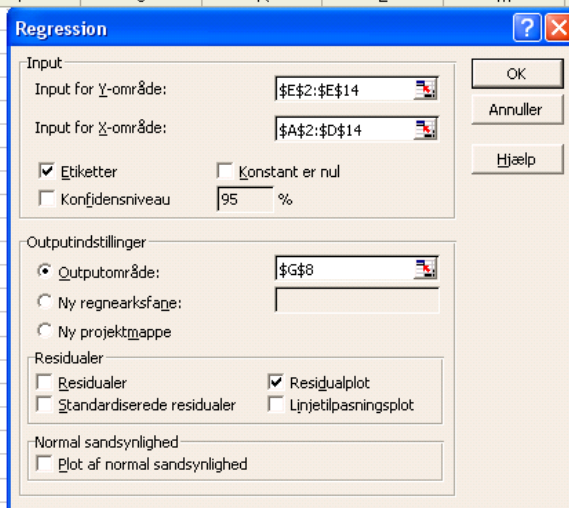
	A	B	C	D	E	F	G	H	I	J	K	L
1	Uddannels	Alder	Mdr. Efter	Højde	Score							
2	x1	x2	x3	x4	y							
3	3	54	6	168	32		x1	1				
4	3	61	8	170	27		x2	0,124	1			
5	3	68	4	164	18		x3	0,148	0,575	1		
6	3	75	10	166	25		x4	-0,255	-0,154	0,179	1	
7	2	54	6	172	28		y	0,471	-0,600	0,106	0,023	1
8	2	61	2	168	19							
9	2	68	8	174	20							
10	2	75	6	166	10							
11	2	82	10	170	12							
12	1	54	4	166	22							
13	1	61	4	172	11							
14	1	68	10	168	20							



25 / 1

Multipel lineær regression — eksempel

	E	F	G	H	I	J	K	L	M	N
1	Score									
2	y									
3	32									
4	27									
5	18									
6	25									
7	28									
8	19									
9	20									
10	10									
11	12									
12	22									
13	11		RESUMEOUTPUT							
14	20									
15										
16			<i>Regressionsstatistik</i>							
17			Multipel R	0,9642						
18			R-kvadreret	0,9297						
19			Justeret R-kvadreret	0,8895						
20			Standardfejl	2,3177						
21			Observationer	12						
22			ANOVA							
23				<i>fg</i>	<i>SK</i>	<i>MK</i>	<i>F</i>	<i>Signifikans F</i>		
24			Regression	4	497,06	124,27	23,13	0,039%		
25			Residual	7	37,60	5,37				
26			I alt	11	534,67					
27										
28				<i>Koefficienter</i>	<i>Standardfejl</i>	<i>t-stat</i>	<i>P-værdi</i>	<i>Nedre 95%</i>	<i>Øvre 95%</i>	<i>Nedre 95%</i>
29			Skæring	109,73	46,22	2,37	0,05	0,45	219,01	
30			x1	4,11	0,93	4,41	0,00	1,90	6,31	
31			x2	-0,81	0,10	-8,26	0,00	-1,04	-0,58	
32			x3	1,74	0,34	5,12	0,00	0,94	2,54	
33			x4	-0,33	0,26	-1,28	0,24	-0,95	0,28	



26 / 1