Supplement til kapitel 10, regression

14. Mere om lineære sammenhænge

Et af de vigtigste statistiske problemer er vurderingen af *sammenhænge* mellem variable. Især inden for økonomien spiller det en stor rolle at beskrive forskellige størrelsers gensidige afhængighed. Hvis der er almindelig vækst i et samfund, stiger prisen på huse f.eks. Samtidig stiger befolkningens indkomster. Men hvordan stiger prisen i sammenligning med indkomsterne?

Inden for driftsøkonomien har man mange sammenhænge. F.eks. stiger lønudgifterne ofte i takt med en virksomheds omsætning, enten ved udvidelse af medarbejderstaben eller ved lønforhøjelser. Men også omkostningerne vokser i takt med omsætningen. For økonomer er det en vigtig opgave at beskrive disse forskellige sammenhænge.

Ofte er det muligt at beskrive sammenhænge som lineære sammenhænge. Dvs. at den ene størrelse *y* er en lineær funktion af den anden størrelse *x*. Som f.eks. i følgende talmateriale.

Eksperiment 1: Lineær regression

I dette eksperiment vil vi udregne koefficienterne a og b ud fra en lineær regression, hvor vi benytter et talmateriale fra de olympiske lege i Atlanta 1996. Tallene er ordnet efter antal opnåede point, som er givet efter, at guld gav tre point, sølv gav to point, og bronze gav et point. USA havde f.eks. 44 gange 3 plus 32 gange 2 plus 25 gange 1, i alt 221 point. Derudover er opgivet bruttonationalprodukt per capita (BNP). Opskriv følgende tal i et regneark.

Land	BNP per capita i mio. dollar (X)	Medaljepoints olympiaden 1996 (Y)	$X \cdot Y$	X^2
USA	33.900	221		
Rusland	4.200	136		
Tyskland	22.700	123		
Kina	3.800	104		
Frankrig	23.300	74		
Italien	21.400	71		
Australien	22.200	68		
Syd-Korea	13.300	56		
Cuba	1.700	51		
Ukraine	2.200	43		

a. Beregn produkterne af ethvert par af *X*- og *Y*-værdierne samt X^2-værdierne i de to manglende søjler.
b. Find summen af tallene i søjlerne: anden søjle $\sum X$, tredje søjle $\sum Y$, fjerde søjle $\sum X \cdot Y$, femte søjle $\sum X^2$.
   Bemærk, at $N = 10$, antallet af datapar.
c. Indsæt disse værdier i formlerne til at finde a og b:

$$a=\frac{N\cdot\sum X\cdot Y-\sum X\cdot\sum Y}{N\cdot\sum X^2-\left(\sum X\right)^2}\ \text{ og }\ b=\frac{\sum X^2\cdot\sum Y-\sum X\cdot\sum X\cdot Y}{N\cdot\sum X^2-\left(\sum X\right)^2}$$

Dvs. find regressionslinjen. Vi vil her ikke argumentere for, at disse formler giver de bedste estimater på *a* og *b*.

d. Brug regnearket til at tegne et plot af datapunkterne sammen med regressionslinjen.
e. Hvor godt passer regressionslinjen med olympiadedata?
f. Hvilken betydning har punktet, hvor regressionslinjen skærer y-aksen?
g. Hvilken betydning kan du give hældningen af regressionslinjen?
h. Brug dit regneark til at finde regressionslinjen, og sammenlign linjen med punkterne fra data.

Ved regressionsanalyse møder man generelt tre hovedproblemer:

1. At bestemme linjens position og hældning.
2. At vurdere, om en ret linje kan beskrive punkterne.
3. At vurdere, om linjens hældning kan have en på forhånd oplyst værdi.

15. Residualanalyse

For at vurdere, i hvilken grad regressionsmodellen beskriver de observerede variable, kan man studere *residualerne* (se s. 247 og s. 26i *Statistik*).

Vi betegner med y^* de beregnede y-værdier ud fra formlen:

$$y^*=a\cdot x+b$$

a og b bestemmes ud fra formlerne (her bruges *mindste kvadraters metode*):

$$a=\frac{\left(x_1-\bar{x}\right)\cdot\left(y_1-\bar{y}\right)+\ldots+\left(x_n-\bar{x}\right)\cdot\left(y_n-\bar{y}\right)}{\left[\left(x_1-\bar{x}\right)^2+\ldots+\left(x_n-\bar{x}\right)^2\right]}\ \text{ og }\ b=\bar{y}-a\cdot\bar{x}$$

Betegnelsen $\bar{x}$ og $\bar{y}$ er gennemsnittet af x'erne henholdsvis y'erne.

En y^*-værdi ligger altså på linjen. Svarende til $y_1, y_2, \ldots, y_n$ har vi dermed $y_1^*, y_2^*, \ldots, y_n^*$, hvor:

$$y_i^*=a\cdot x_i+b,\quad i=1,\ldots,n.$$

Residualerne z_i beregnes nu som:

$$z_i=y_i-y_i^*$$

Eller simpelthen den lodrette afstand mellem observationen og linjen.

Det må være sådan, at residualerne skal være relativt små og fordele sig jævnt og tilfældigt om 0, hvis modellen skal kunne accepteres. For at vurdere dette, kan man afbilde residualerne mod i, dvs. mod observationens nummer, eller mod x_i, dvs. set i relation til x-observationernes variation.

Det væsentlige er at få afsløret systematik. Vi må f.eks. forkaste modellen hvis residualerne vokser eller aftager med voksende værdi af x_i.

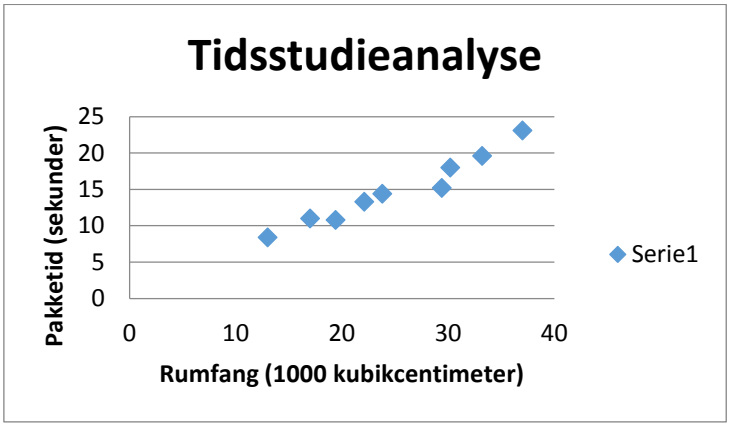
Nedenfor er en øvelse, som behandler ovenstående teori om residualer.

Øvelse 12: Tidsstudium

I forbindelse med en tidsstudieanalyse har man ladet trænede personer pakke nogle pakker af varierende størrelse. For ni pakker er i tabellen nedenfor vist pakkens størrelse i 1000 cm^3 (x) og pakketiden i sekunder (y).

Pakkenummer	Rumfang (1000 cm^3)	Pakketid (sekunder)
1	19,4	10,8
2	23,8	14,4
3	33,2	19,6
4	30,2	18,0
5	13,0	8,4
6	29,4	15,2
7	17,0	11,0
8	22,1	13,3
9	37,0	23,1

a. Undersøg sammenhængen mellem pakketid og rumfang grafisk, og bestem en regressionslinje.



b. Vis, at regressionslinjen bliver $y^* = 0{,}5757 \cdot x + 0{,}4682$.

Vi opstiller en tabel med værdierne y_i, y_i^* og z_i^*

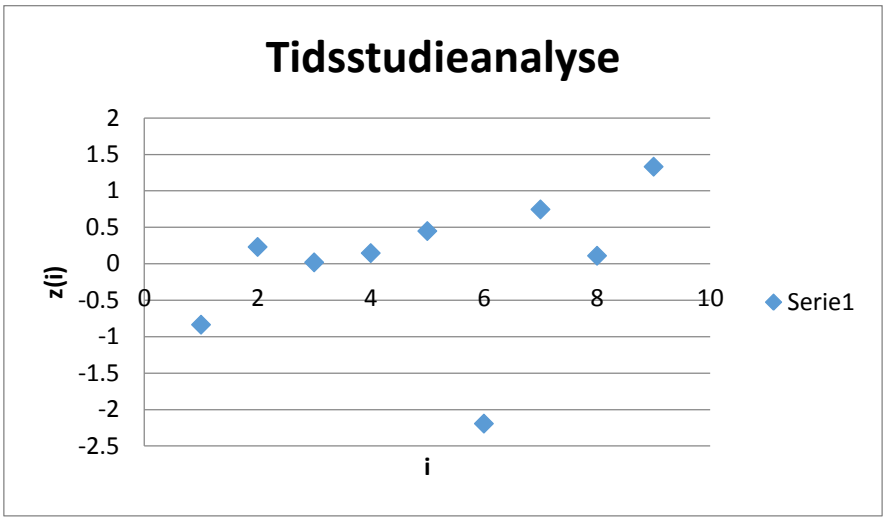
Pakke nummer	y_i	y_i^*	z_i
1	10,8	11,63	-0,84
2	14,4	14,17	0,23
3	19,6	19,58	0,02
4	18,0	17,85	0,15
5	8,4	7,95	0,45
6	15,2	17,39	-2,19

7	11,0	10,25	0,75
8	13,3	13,19	0,11
9	23,1	21,77	1,33

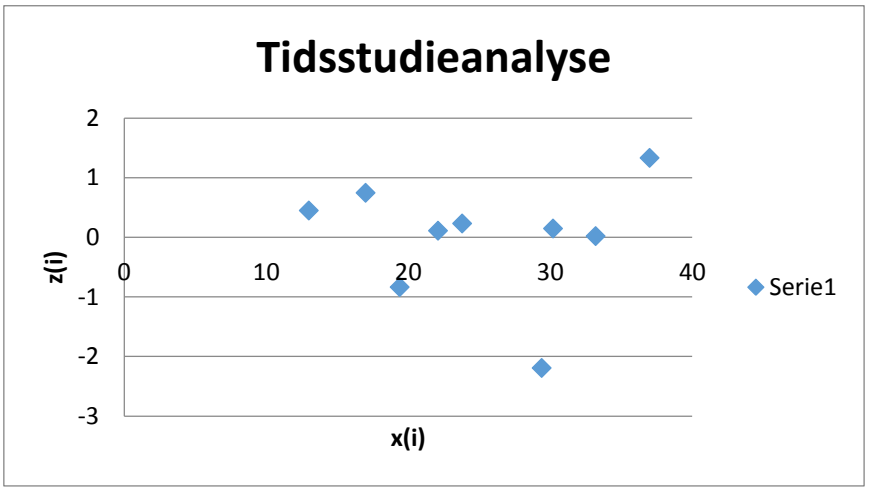
c. Kontroller, at tallene i tabellen stemmer overens med din beregning af regressionslinjen ovenfor.

d. Tegn et diagram som nedenfor, altså z_i-værdierne mod datanummer i eller x_i.

e.



Residualerne (z-værdierne) tegnet mod i.



Residualerne(z-værdierne) tegnet mod x_i.

Konklusion: Vi ser af de to figurer ovenfor, at residualerne er relativt små, idet den største kun er en femtedel af y'ernes gennemsnit. Der kan ikke spores nogen systematik, hverken med i eller med x_i. Derfor kan vi acceptere modellen.

Man kunne diskutere, om det ene punkt er en 'outlier' (se kapitel 3 i *Statistik*) og dermed påvirker regressionslinjen for meget i den ene retning.