Regressioner og regressionsanalyse

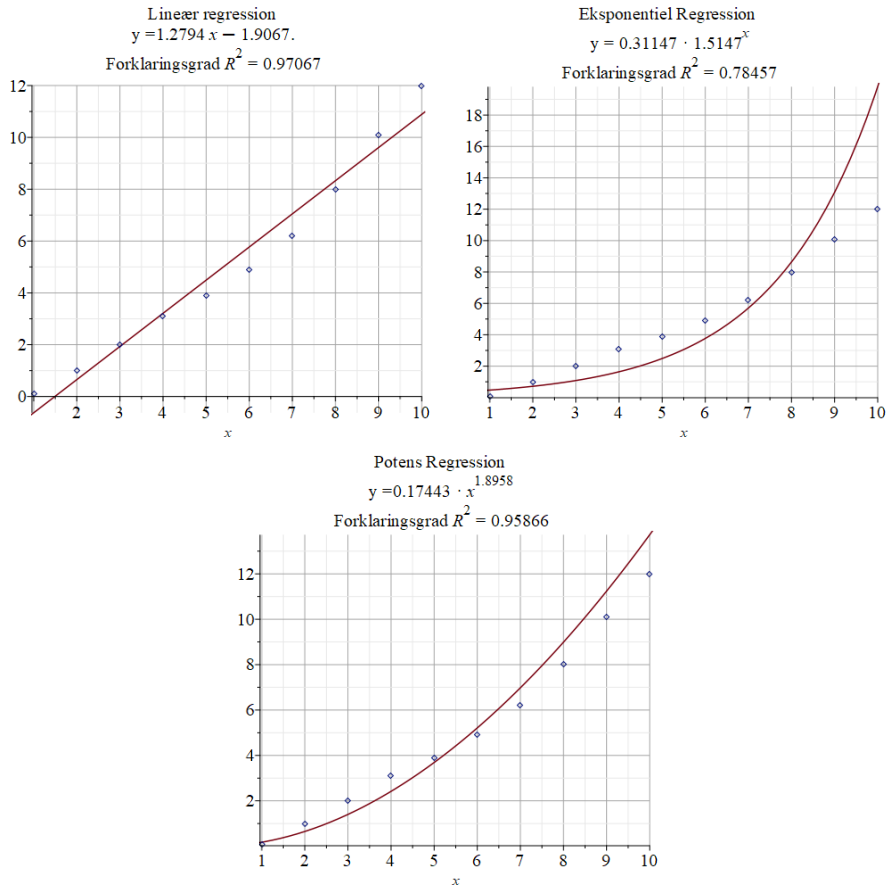
Regression

Vi skal i dag se på regressionsanalyse.

Eksempel 1.1. Har vi et datasæt

x	1	2	3	4	5	6	7	8	9	10
y	0.1	1	2	3.1	3.9	4.9	6.2	8.0	10.1	12.0

uden at vide, hvilken underlæggende sammenhæng, der skaber dataet, kan vi prøve at lave forskellige regressioner på det. Dette gøres i Maple og kan ses af Fig. 1



Figur 1: Lineær regression, eksponentiel regression og potensregression på datasæt

Hvordan afgør vi så, hvilken af modellerne der er bedst, hvis vi ikke har yderligere

information omkring kilden af dataet? Man kan ikke sammenligne korrelationskoefficienter mellem modeller. Derfor skal vi lave regressionsanalyse. En del af regressionsanalyse er blot at betragte de fittede modeller og se, hvilken model der ser ud til at passe bedst. I dette tilfælde er det den lineære model eller potensmodellen.

Regressionsanalyse

Til at vurdere, om en regression er god, skal vi bruge et værktøj til at afgøre, om det modellen rammer forbi blot er tilfældige fejl eller om vi rammer strukturelt forkert. Første tilfælde er ønskværdigt. Vi definerer nu residualerne til en regression.

Definition 1.2. Givet måledata $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ og regressionsmodel $f(x)$, så definerer vi det i'te residual som det, modellen f rammer forbi den målte værdi. Mere præcist

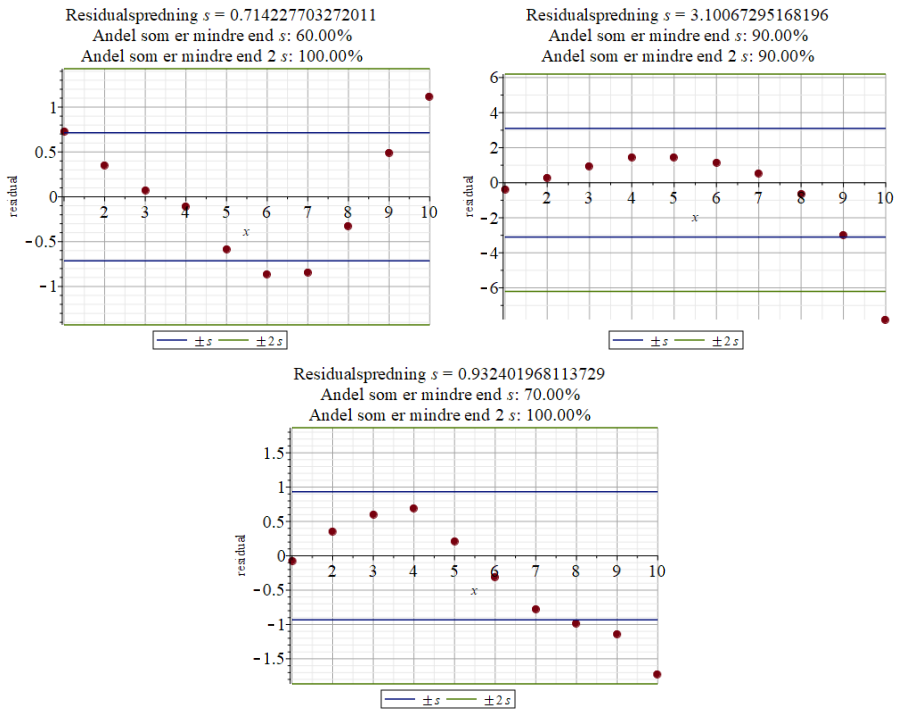
$$r_i = y_i - f(x_i).$$

Det er klart, at en god model $f(x)$ vil have små residualer. Udover dette så vil vi også have normalfordelte residualer. Det vil sige, at vi gerne vil have mange residualer tæt på 0 og kun få langt fra 0. Der må ikke være nogen yderligere struktur i residualerne. For at se, hvad dette betyder, ser vi på et eksempel.

Eksempel 1.3. Betragter vi regressionerne fra tidligere eksempel, kan vi finde residualerne for de tre modeller. Disse findes i Maple ved eksempelvis

```
plotResidualer(X,Y,LinReg)
```

i fald residualerne skal findes for lineær regression. Residualerne for de tre modeller kan ses af Fig. 2



Figur 2: Lineær regression, eksponentiel regression og potensregression på datasæt

Det er lineær regression og potensregression, der har mindst spredning i residualerne, men det ser ud til, at der er struktur i residualerne for alle tre tilfælde. Derfor kunne det tyde på, at ingen af de tre regressioner passer godt til datasættet.

Opgave 1

Lav regression på følgende datasæt og afgør med residualanalyse hvilken model der passer bedst.

1	2	3	4	5	6	7	8	9	10
1.45	14.96	7.22	26.63	45.49	66.88	103.90	157.63	162.63	252.61

Opgave 2

I dette datasæt findes antal bakterier i mia. sammenholdt med den forløbne tid i timer i en bestemt bakteriekoloni.

i) Afgør, hvilken regressionstype, der passer bedst på datasættet ved at bruge residualerne for modellen.

ii) Brug din valgte model til at bestemme, hvor mange bakterier der vil være efter 30 timer.

Opgave 3

I dette datasæt findes antal beboere i en by angivet efter antal forløbne år efter år 1900.

i) Afgør, hvilken regressionstype, der passer bedst på datasættet ved at bruge residualerne for modellen.

ii) I hvilket år vil der i følge din model være 8000 beboere i byen?

Opgave 4

I dette datasæt findes prisen på en bestemt aktie i en tidsperiode på én time.

i) Afgør, hvilken regressionstype, der bedst beskriver aktieprisen ved at anvende residualerne for modellen

ii) Hvornår vil aktiens pris være på 220kr ifølge modellen?

Opgave 5

I dette datasæt findes BNP per person (i US dollars) for et land fra år 1960 til 2021

i) Afgør, hvilken regressionstype, der bedst beskriver BNP ved at anvende residualerne for modellen

ii) Hvis vi antager, at der er et konstant befolkningstal på 10 millioner mennesker i landet, hvornår vil landets samlede BNP så være på 1 billion dollars?