Lineær regression – Residualernes spredning

Allan Lind Jensen, Nærum Gymnasium

Denne artikel omhandler residualernes fordeling ved lineær regression. Situationen er som beskrevet på side 352 i MAT A2, 3. udgave. For givne x–værdier, x_i, $i = 1 \ldots n$, er y–værdierne givet ved uafhængige stokastiske variable $Y_i \sim N(ax_i + b, \sigma)$.

Der står på side 353 i samme bog, at residualerne har samme fordeling $N(0, \sigma)$. Det er effektivt forkert. Det passer jo, hvis vi beregner residualerne med modelligningen $y = ax + b$. Men den lineære regression finder jo med sandsynlighed 1 en linie, der passer bedre med de faktiske resultater. Derfor må residualerne have mindre spredning.

Men residualerne har ikke engang samme spredning. For de tre x–værdier (0, 1, 1000) er det meget billigt for linien at rette sig ind efter y–værdien i $x = 1000$. Residualspredningen i det punkt vil derfor være langt mindre end residualspredningen for $x = 0{,}1$.

Den værdi for hældningskoefficienten, som den lineære regression resulterer i, kan betragtes som en stokastisk variabel

$$A = \frac{n\sum x_i Y_i - \sum x_j \sum Y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{\sum\left(nx_i - \sum x_j\right)\sum Y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

Heraf følger, at A er normalfordelt. Eksplicit udregning giver, at

$$E(A) = a \quad \text{og} \quad Var(A) = \frac{\sigma^2}{n\sigma_x^2}$$

Estimatoren for konstantleddet kan også betragtes som en stokastisk variable, B. Residualerne kan også betragtes som en stokastisk variabel

$$R_k = Ax_k + B - Y_k$$

Det kan nemt udregnes, at residualerne er normalfordelte med middelværdi 0. Jeg har udregnet variansen et par gange, og resultatet er begyndt at konveregere, måske mod ligningen

$$\frac{Var(R_k)}{\sigma^2} = \frac{n-1}{n} - \frac{(x_k - \mu_x)^2}{(n-1)\cdot\sigma_x^2}$$

hvor variansen er udregnet ved at dividere med n - 1. Af ligningen fremgår, at spredningen på residualerne er størst tæt på middeltallet for førstekoordinaterne. Med x–værdierne (0, 0, 0, 1) forsvinder residualspredning på Y_4.
Konklusionen på dette er, at i opgave 7.D2.13 i de Vejledende Enkeltopgaver i Matematik stx A–niveau bør spørgsmålene c og d nok ikke stilles. Hverken til eksamen eller som hjemmeopgaver.

Har alt dette praktisk betydning? Det besvarer spørgsmål b i samme opgave sjovt nok. På QQplottet over residualerne kan man se, at punkterne i midten ligger over linjen, og punkterne ude i siderne ligger under linjen. Punkterne afviger systematisk fra linjen. Derfor tyder plottet på, at residualerne ikke er normalfordelt.