En god ret linie II

Af Peter B. Yde, Silkeborg Amtsgymnasium.

Et indlæg i LMFK brugte nogle særegent kommenterede brudstykker fra min artikel i LMFK nr. 3, 2003. Bl.a. var der klippet en tredjedel af nedenstående tekst under billedet af C. F. Gauss:

Carl Friedrich Gauss (1777-1855) opstillede i begyndelsen af 1800-tallet to uafhængige sæt af betingelser, under hvilke det, der blev til standarden inden for beregning af en tilnærmelseslinie, var retfærdiggjort. Kan det hænge sammen med hans påfølgende anerkendelse, at metoden blev standard?

Min artikel var allerede mere end lang nok. Så selvom Gauss har leveret de bedste underbyggelser af mindste kvadraters metode, udelod jeg en udbygning af billedteksten. Men i stedet for blot at rynke på næsen af den brug, der blev gjort af de nævnte brudstykker, skal jeg da gerne levere udbygningen nu.

Den ældste retfærdiggørelse fra Gauss' hånd er indsendt i 1809-10. Den bruges bl.a. til at beskrive asteroiden Pallas' bane ud fra et antal observationer, der "overbestemmer" banen, altså har flere ligninger end ubekendte.

Gauss antager, at alle observationerne har samme spredninger (eller varianser). Dernæst kræver han, at sandsynlighedsfordelingerne for fejlene skal have maksimum for fejlen 0. Næste antagelse er, at iagttagelserne er uafhængige, så den samlede fejls sandsynlighed bliver produktet af de enkelte observationers sandsynligheder.

Endelig argumenterer Gauss for, at sandsynlighedsfordelingen for fejlen på den enkelte observation med fornuft kan sættes til

$$\Phi(v) = \frac{1}{\sqrt{2\pi}\,s}\, e^{-\frac{v^2}{2s^2}}$$

Dvs. at sandsynlighedsfordelingen antages at være normalfordelingen – dvs. den, vi i dag stadig ofte kalder Gauss-funktionen! ν er fejlen – nøjere bestemt den usystematiske fejl. s er spredningen. Gauss konstaterer, at funktionen har den "mangel", at den giver en sandsynlighed større end 0 for alle værdier af fejlen, men fysiske forhold begrænser mulige værdier af fejlen.

Ved multiplikation af sandsynlighederne for de enkelte observationer fås i kraft af uafhængigheden af observationerne, at den samlede sandsynlighed er proportional med

$$e^{-\frac{1}{2s}(v^2+v'^2+v''^2+\ldots)} \qquad (1)$$

Gauss brugte ikke indices (men dog undtagelsesvist superscripter, som vi i dag bruger indices). ν, ν', ν'', … er fejlene på de enkelte målinger; de er udtrykt ved et sæt parametre, hvis antal er mindre end antallet af observationer.

Gauss opsøger derpå maksimum for funktionen (1), hvilket han umiddelbart konstaterer fås i minimum for kvadratsummen $\nu^2+\nu'^2+\nu''^2+\ldots$ Fremstillingen er således en af de ældste, der benytter "the method of maximum likelihood". Denne optræder allerede i 1700-tallet, men finder dog først sin nuværende form med eksplicitte likelihoodfunktioner i 1920'erne.

Senere i skriftet kommer Gauss ind på minimering af alternative summer så som $\nu^4+\nu'^4+\nu''^4+\ldots$; $\nu^6+\nu'^6+\nu''^6+\ldots$ eller $|\nu|+|\nu'|+|\nu''|+\ldots$ Overvejelserne optræder som en kommentar til den ældste offentliggørelse af mindste kvadraters metode. Den er leveret af A.M. Legendre og udkom i 1806. Gauss

argumenterer på dette sted for minimering af $v^2 + v'^2 + v''^2 + \ldots$ med, at denne sum involverer de simpleste beregninger!

I maximum likelihood metoden er det først og fremmest Gauss' valg af Gauss-funktionen som sandsynlighedsfordeling, der sikrer potensen 2 (kvadraterne) i summen, der minimeres!

Maximum likelihood metoden støttes af, at Gauss-funktionen har en særlig position. Den ses i de mange eksempler på omtrentlige normalfordelinger, der findes. Og den kommer til udtryk i Den centrale Grænseværdisætning. Gauss var i 1820'erne bekendt med en forløber til den sætning. Den var publiceret af P.S. Laplace en halv snes år tidligere. Da Laplaces arbejde er en genoptagelse af et arbejde af A. de Moivre fra 1733, betegnes sætningen de Moivre-Laplace sætningen.

Men denne særlige position bringer ikke Gauss-funktionen så højt op på en piedestal, at den er hævet over diskussion som antagelse i begrundelsen for mindste kvadraters metode. Sandsynligheden for fejl svarende til negative værdier af den afhængige variable Y_i er ofte 0. Og ofte må sandsynlighedsfordelingen formodes ikke at være symmetrisk om middelværdien. Tænk i begge disse eksempler f. eks. på Ohms lov.

Finurligt nok fandt Gauss da heller ikke sin første begrundelse for mindste kvadraters metode tilstrækkelig god. Der er jo også en "sværm" af antagelser i den. Det lykkedes Gauss at opstille en anden begrundelse, hvor en af antagelserne faldt bort. Og hvilken? – den om Gauss-fordelingen! Til gengæld måtte han indføre et kriterium, som kan diskuteres.

Gauss kunne nøjes med at antage tre ting: Middelværdien af fejlen for en given værdi af den uafhængige variable er 0 uanset værdien. Variansen af fejlen for en given værdi af den uafhængige variable er uafhængig af værdien. Iagttagelserne er uafhængige. Disse forudsætninger er næsten identiske med antagelserne i den første udledning.

Parametrene for regressionslinjen opfattes som linearkombinationer af de stokastiske variable Y_i'erne. Disse linearkombinationer er altså selv stokastiske variable, bestemt ved mindste kvadraters metode. Deres middelværdier vises at være, hvad de gerne skulle være, nemlig værdierne fundet ved mindste kvadraters metode. I denne egenskab betegnes de som middelrette (eller middelværdirette). Man kan imidlertid også danne andre middelrette linearkombinationer af Y_i'erne. Men de ved mindste kvadraters metode bestemte parametre er dem, der har de mindste varianser!

Dette er Gauss' anden underbygning af mindste kvadraters metode. Den blev publiceret i begyndelsen af 1820'erne. Det fremgår af senere skrifter, at Gauss er bedre tilfreds med denne anden "støttepille" for regressionsmetoden! Dette er taget til efterretning både i en fransk og en tysk oversættelse af Gauss' skrifter om mindste kvadraters metode fra 1800-tallet. I begge disse værker bringes den sidste underbygning først, just med henvisning til Gauss' egen opfattelse. Denne deles i øvrigt af flere nutidige matematikere.

Gauss skrev på latin. Det kan jeg ikke læse, men tysk. Imidlertid skal der mere end god tid til at tyde selv det tyske skrift om den anden begrundelse. Det er der flere årsager til: For det første brugte Gauss metoden til at filtrere et rimeligt datasæt ud af store overbestemte sæt af måletal, så han arbejdede med lineære funktioner af mange variable. Og for det andet var hverken matrix- eller vektor notationen med indices skabt endnu; Gauss brugte et sandt mylder af bogstaver for at forklare sig. Og endelig opstillede han parallelt med regressionsmetoden også Gauss-eliminationen(!) i skrifterne.

Der skete da også det, at Gauss' anden begrundelse gik i glemmebogen. Den blev genskabt af russeren A.A. Markov (1856-1922) og tillagt ham alene i nogle år. I dag kaldes den Gauss-Markov sætningen.

Begrebet varians som bedste mål for fejlens minimering er opstillet af Gauss! Dermed putter Gauss selv potenserne 2 (kvadraterne) på fejlene ind! Om det skrev Gauss i 1821-skriftet, at variansen synes "mest egnet til at definere og måle usikkerheden på iagttagelserne"… "Hvis nu en eller anden ville indvende, at den fastlæggelse var truffet vilkårligt uden tvingende nødvendighed, så medgiver vi gerne dette"… "I den uendelige

mangfoldighed af funktioner, der har den egenskab [altid at være positive], synes den simpleste at fortjene at blive foretrukket frem for andre, og det er [variansen]"… "[Om variansen er det bedste udtryk for fejlene] er alene overladt til fri bedømmelse"!

Som matematiklærere måske vil erindre, deler enhver matematisk kompetent nybegynder i spredningsteori netop denne skepsis.

I mange praktiske eksempler er der ikke kun usikkerhed på den afhængige variable, men også på den uafhængige. F.eks. er spredningerne positive både på strømstyrke og spændingsforskel ved Ohms lov. Det strider mod antagelsen i standard regression.

En anden ofte brudt forudsætning er, at spredningen er uafhængig af værdien af den uafhængige variable. F. eks. vil spredningerne ved Ohms lov ofte være større, jo større strømstyrken er, mens spredningen i forhold til strømstyrken måske er nærmere ved at være konstant.

Mylderet af forudsætninger i både maximum likelihood princip begrundelsen og Gauss-Markov sætningen er altså stort. Det er så omfattende, at en eller flere af antagelserne normalt er overtrådt væsentligt i anvendelserne. Derfor kan man roligt udskifte normalfordelingen med

$$\varphi(v) = c \cdot e^{-\frac{|v|^k}{d}}$$

hvor k godt må afvige betydeligt fra 2. c og d er positive konstanter, med hvilke $\varphi(v)$ normeres og gives en bestemt spredning. k skal være større end 1, men kunne godt være $k = \frac{1+\sqrt{5}}{2} \approx 1,62$ – hvis man nu gerne vil more sig med dette tal. Likelihood funktionen bliver så

$$e^{-\frac{1}{d}(|v|^k + |v'|^k + |v''|^k + \ldots)}$$

Svarende hertil skal summen $|v|^k + |v'|^k + |v''|^k + \ldots$ minimeres! For f. eks. $k = \frac{1+\sqrt{5}}{2}$ fås lige så gode resultater som for $k = 2$! Men beregningerne er naturligvis langt mere snørklede. Så det er fornuftigt at bruge $k = 2$, som oven i købet har sin særstilling udtrykt i Den centrale Grænseværdisætning og – knyttet hertil – de mange praktiske eksempler på nogenlunde normalfordelte fejl.

De mange antagelser betyder også, at standard regressions metoden oftest ikke er bedre funderet end Rikkes raske rette linje – som jeg i spøg kaldte linjen tegnet på frihånd i min artikel i LMFK nr. 3/2003. Nu og da er den endda dårligere og dette af ransagelige årsager.

Det er jo ganske pudsigt. Men jeg spøger ikke længere.

Litteratur

Her er et lille udpluk af min anvendte litteratur:
(1) Carl Friedrich Gauss, *Abhandlungen zur Methode der kleinsten Quadrate* (In deutscher Sprache), 1887, Neudruck 1964, Physica-Verlag, Würzburg.
(2) Esben Høg, H. J. Juhl, *Statistik for økonomer: Regressionsmodeller*, 3. udg. Systime, 1989.