QR6 – Galtons regressionslov – Regression towards the mean

Begrebet *regression* og den såkaldte lov om *regression towards the mean* stammer fra nogle artikler, som englænderen Francis Galton (1822-1911) skrev i 1888. Galton beskæftigede sig med mange forskellige emner og discipliner og blandt andet også statistik. Det var i statistikkens tidlige barndom, og nogle af pionererne havde meget bastante opfattelser af, at fx biologiske fænomener blev styret af lovmæssigheder næsten ligesom fysiske fænomener – man skulle blot have tilstrækkelig med data for at afdække disse love.

Galtons fortjeneste vart at han organiserede indsamling af en række store datasæt, bl.a. det som vi i lærebogssystemet simpelthen kalder for '*Galtons datasæt*'. Det er data over samhørende målinger af fædres og sønners højde. Galton lagde mærke til at høje mennesker godt nok fik høje børn, men i gennemsnit ikke helt så høje som de selv, og omvendt små mennesker fik små børn, men som dog i gennemsnit var lidt større end forældrene. Det at være høj, var for Galton en positiv egenskab, og derfor så han det som noget negativt, at de fik børn der var ikke helt så høje. Derfor indførte han begrebet *regression* til at beskrive denne konkrete sammenhæng. Regression betyder *tilbageskridt*. Det tænker man ikke videre over i brugen af ordet i dag, for siden Galton er regression kommet til at dække det fænomen, at man i en sky af datapunkter prøver at finde en (normalt lineær) sammenhæng.

Hans centrale artikel om emnet, som du kan finde *her*, gav han titlen *Regression towards mediocrity in Hereditary Stature*. Idag taler vi om *regression towards the mean*, hvilket betyder regression mod *middelværdien*, men Galtons betyder *regression mod middelmådighed*, som jo er en værdimæssig negativ betegnelse.

Begrebet *regression towards the mean* beskriver det *statistiske* fænomen, at "ekstreme" værdier (meget høje eller meget lave) vil have en tendens til at blive efterfulgt af knap så ekstreme. En præstation af en sportsudøver, der ikke er suverænt god, vil altid have et element af tilfældighed over sig. Skyder man en golfbold i hul over lang afstand, vil man nok præstere dårligere i næste slag. Rammer man i plet i bueskydning, vil man næppe gøre det igen næste gang. Det samme gælder selvfølgelig i spil, det er yderst sjældent nogen vinder i lotto to gange i træk. Har man slået tre seksere i streg, forventer de færreste nok at kunne gøre det igen.

Ekstreme præstationer efterfølges af mere middelmådige – men også omvendt: Er man totalt uheldig i sit skud, er man nok lidt bedre næste gang. Ligesom der ikke er mange supermænd i virkeligheden, er der heller ikke mange hvis præstationer giver helt vilkårlige resultater. Til en *multiple choice-test* kan man måske med sikkerhed svare på en del, men der vil også være en portion spørgsmål, hvor man er usikker, og så giver et kvalificeret gæt ud fra de 5 eller 6 muligheder der er. Præstationen er en kombination af noget der er sikkert og noget der er tilfældigt. Sådan er det også i naturen, fx med nedarvede egenskaber. Det er ven kombination af noget der er årsagsbestemt og noget der er tilfældigt. Og det tilfældige vil være underlagt '*regression towards the mean*'.

Men lige præcis denne del af en præstation – herunder det at lave et barn! – er ikke *årsagsbestemt*. Vi vil gerne se årsager og har sværere ved at acceptere tilfældigheder som en forklaring. Og her går det galt i mange tolkninger. Det gjorde det allerede fra starten hos Galton selv. Hvis hans triste konklusion om at vi bevæger os mod middelmådighed var korrekt ville spredningen på højderne i næste generation blive mindre. Men det er ikke tilfældet. I Galtons eget datasæt sker der faktisk det modsatte – en markant øgning af spredningen. Galton ledte efter årsager – til sin forkerte tese! – og forklarede, at den sørgelige tilstand med at det hele ender i middelmådighed skyldes, at vi ikke kun nedarver fra forældrene – for så ville høje forældre få høje børn påstod han fejlagtigt. Men vi arver også en stor portion af vores fra forfædre langt tilbage, bedsteforældre, oldeforældre osv. Og det er helt forkert. Arvemassen kommer udelukkende fra en mor og en far.

Et af de klassiske eksempler på en forkert tolkning af *'regression towards the mean'* har en psykolog, Daniel Kahnemann givet. Det stammer fra en periode, hvor han underviste israelske flyinstruktører i, hvordan man skulle tilrettelægge god undervisning:

I en undervisningssituation forklarede jeg flyinstruktørerne, at ros er mere effektiv end straf for at fremme indlæring. Da jeg var færdig med min entusiastiske tale, bad en af de mest erfarne instruktører i om ordet, og han holdt så en kort tale, som begyndte med at sige, at han ikke ville afvise at ros var godt i hunde - og fugleopdræt. Men ikke i oplæring af flykadetter. Han sagde, "ved mange lejligheder har jeg rost flyvekadetter når de har udført nogle fantastiske manøvrer i luften, men når de så prøver igen præsterer de dårligere. På den anden side har jeg ofte skreget til kadetter der har performet dårligt. Og generelt kan man sige, at de har gjort det bedre næste gang. Så lad være med at fortælle os, at ros fungerer og straf ikke gør, for det modsatte er tilfældet." Dette var et aha-øjeblik, hvor jeg forstod hvor udbredt misforståelsen om *regression towards the mean* er. Jeg arrangerede straks en demonstration, hvor hver deltager kastede to mønter på et mål bag ryggen uden nogen feedback. Vi målte afstande fra målet og kunne se, at de, der havde gjort det bedste, første gang, for det meste havde forværret deres andet forsøg og omvendt. Men jeg vidste godt at denne demonstration næppe ville ændre på en pervers opfattelse af forholdet mellem ros og straf.

Professionelle statistikere er nødt til at tage begrebet med i deres analyser af fx effekten af et nyt præparat. Noget er årsagsbestemt, men andre effekter er tilfældige. Og tager man ikke det i betragtning, kan et tilfældigt ekstremt resultat bliver tolket forkert son noget der kun tilskrives det nye præparat.