L&R Uddannelse
EGMONT

Om at forstå forklaringsgraden og om de fælder man kan gå i.

Tallet R^2 kaldes for forklaringsgrad, hvilket er en lidt ulyksaligt navngivning af dette begreb, da det giver anledning til en række misforståelser, der florerer i andre fag, hyppigst ser vi det i samfundsfag. Det kan ofte være ganske kompetente videnskabsfolk fra et andet felt end matematisk statistik, der anvender statistik og giver en forkert fortolkning af R^2, fx i retning af, at hvis tallet $R^2 = 0.35$, så skulle det betyde, at 35% af sammenhængen mellem de to variable er forklaret. Selve denne sætning er meningsløs, og når man hører det bør man bede om en uddybning. Hvad er det % af? Og hvordan kan forskellige variable bidrage med hver sin % til en given variabelsammenhæng? Og hvordan skulle man lægge sådanne procenter sammen. Fx afhænger fertilitet af mange ting, alder, genetik, hormoner, forurening i miljøet, måske kost, måske sygdomme man har haft, måske vaccinationer man har fået osv osv. Det er meningsløst at tale om, at hver af disse variable og mange flere bidrager med eb bestemt % til fertiliteten, hvordan skulle vi lægge sådanne faktorer sammen, og hvad er 100% ?

Man kan illustrere det meningsløse i sådanne påstande gennem følgende eksempel og øvelse:

Nedenfor ses fire datasæt præsenteret af statistikeren Francis Anscombe i 1973

x	10	8	13	9	11	14	6	4	12	7	5
y_1	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68
y_2	9.14	8.14	8.74	8.77	9.26	8.1	6.13	3.1	9.13	7.26	4.74
y_3	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73
x_4	8	8	8	8	8	8	8	19	8	8	8
y_4	6.58	5.76	7.71	8.84	8.47	7.04	5.25	12.5	5.56	7.91	6.89

a) Udfør lineær regression på fire datasæt (x, y_1), (x, y_2), (x, y_3) og (x_4, y_4), og undersøg, om der kan være tale om en lineær sammenhæng.
b) Bestem et residualplot for hver af sammenhængene.
c) Bestem R^2 for hver af sammenhængene.
d) Beskriv forskellene mellem de fire sammenhænge og deres datasæt.

I *Hvad er matematik? A* handler kapitel 9 om regressionsmodeller. Heri forklares hvordan man finder en given lineær regression, dvs den bedste rette linje der minimerer summen af afstandskvadraterne mellem modellens tal og de empiriske værdier. Kvadratet på denne sum af afstandskvadrater svarer via Pythagoras sætning til længden af en vektor i et rum med, lige så mange dimensioner som antallet af dataværdier. Og længden af den vektor, vi her betragter, svarer til afstanden fra et punkt, der repræsenterer dataværdierne til en hyperplan. Det er denne afstand vi minimerer, når vi udfører lineær regression. Betragter vi vinklen mellem denne vektor og hyperplanen, så viser det sig, at cosinus til denne vinkel præcis bliver lig med korrelationskoefficienten r !. Derfor ligger r altid mellem -1 og 1.

Man kan yderligere vise, at for lineære sammenhænge, så er $R^2 = r^2$, dvs R^2 ligger altid mellem 0 og 1. Når vinklen er 0, er cosinus 1. Det svarer til, at i det 2-dimensionelle rum ligger punkterne på en ret linje. Er vinklen tæt ved 0 er cosinus tæt ved 1. Dette svarer til at punkterne med god tilnærmelse ligger på en ret linje. Men cosinus er jo ikke lineær. Så man kan ikke udlede en lineær sammenhæng mellem tallet R^2 og kvaliteten af den lineære sammenhæng. Man kan blot sige, at for et tal tæt ved 1 er der matematisk tale om en god sammenhæng, og omvendt hvis tallet er tæt ved 0.

Men man skal som sagt passe på med at fortolke forklaringsgraden for firkantet ud fra det navn den har fået. Der er også stor forskel i fagenes traditioner på, hvordan man fortolker forklaringsgraden. Forskellen bunder for en stor del i, hvilke *modeller*, man arbejder med. Vi kan opdele i to hovedgrupper alt efter om modellerne:

- bruges i en simpel konktekst, hvor der er styr på de variable, der indgår, i form af variabelkontrol, og hvor man fra fagets teori har en formodning om, at sammenhængen kan begrundes teoretisk.
- bruges i en kompleks kontekst, hvor man ikke har styr på de mange variable, der indgår, og hvor der heller ikke i fagets teori er forventninger til en bestemt simpel sammenhæng mellem to af de typisk mange hundrede variable man fokuserer på.

Det første er typisk situationen i fysik eller kemi, hvor man forsøger at eftervise en eksakt naturlov, der udmønter sig i en lineær sammenhæng mellem to veldefinerede variable, der kan isoleres og varieres systematisk. I så fald er summen af de kvadratiske afvigelser alene knyttet til måleusikkerheden. Hvis residualerne spreder sig som forventet i forhold til måleusikkerheden er den lineære model derfor en god model. I et typisk omhyggeligt udført skoleforsøg af denne type forventer man derfor erfaringsmæssigt forklaringsgrader på over 95% og hvis forklaringsgraden når helt op over 99% har man eftervist sammenhængen med stor succes.

Det andet er derimod typisk situationen i biologi og samfundsfag: Et kondital afspejler fx et stort antal faktorer i organismen og det samme gælder en bestemt løbetid. Der er derfor ingen grund til at forvente en simpel entydig sammenhæng mellem sådanne to variable. Der er heller ingen variabelkontrol: De andre faktorer varierer tilfældigt fra forsøgsperson til forsøgsperson, så selv om der for den enkelte forsøgsperson godt kan være en rimelig simpel sammenhæng mellem løbetiden og konditallet, vil den kollektive sammenhæng i en større gruppe sagtens kunne føre til en betydelig spredning på grund af variationerne i de øvrige faktorer. Og selv om der er en simpel sammenhæng behøver den ikke lige netop være lineær. Fx kunne det være at en omvendt proportionalitet rent faktisk gav en bedre beskrivelse af sammenhængen mellem løbetiden og konditallet. Når vi fokuserer på en lineær sammenhæng er det altså alene som en deskriptiv model af en kompleks sammenhæng. I sådanne tilfælde, hvor der ikke er nogen bagved liggende teori til at guide os, vil man typisk være tilfreds med en forklaringsgrad, som den vi fandt på 45%. Hvis forklaringsgraden når op over 75% vil man ofte opfatte det som en stor "succes" med modellen.