Brugen af R^2 i gymnasiet

Per Bruun Brockhoff, DTU Compute
Ernst Hansen, KU Matematik
Claus Thorn Ekstrøm, KU Biostatistik

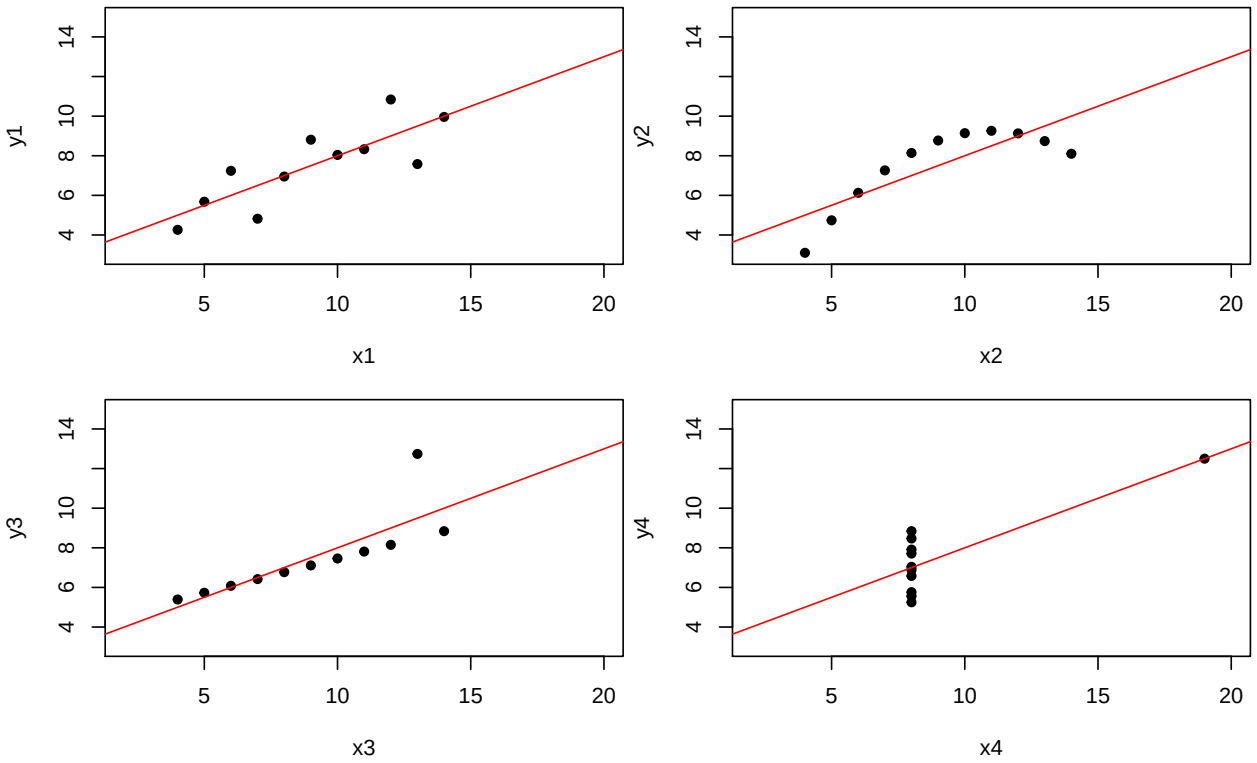
17 januar 2017

Der lader til at være en vis forvirring blandt og uenighed mellem forskellige faggrupper omkring R^2-værdien, også kaldet "forklaringsgraden" eller "determinationskoefficienten". Uenigheden omkring brugen og nytten af R^2 som et mål til at bekrive en statistisk model optræder ikke kun i gymnasiet: globalt set skaber brugen af R^2 tilsvarende gnidninger. Den anvendes rigtig meget i visse miljøer. Man kan imidlertid finde en del fagstatistikere, der vil tænde advarselslampen overfor forskellige over- og fejlfortolkninger af R^2-værdien, som det er let at lade sig besnære af, og som mange miljøer uden tvivl gør sig skyldige i engang imellem.

For en fagstatistiker kan det derfor være fristende simpelthen at fraråde brugen af R^2 i det hele taget for at undgå, at folk fejlfortolker resultatet og/eller misbruger størrelsen. Med dette indspark håber vi at kunne bidrage til den fælles forståelse for hvad R^2 kan og ikke kan gøre for os, og pege på et alternativ, der i mange faglige sammenhænge kunne være en mere direkte størrelse at beregne.

Et eksempel: Anscombes data

Et klassisk eksempel, der viser, hvorfor R^2 i sig selv er problematisk, er Anscombes fire datasæt vist nedenfor (Anscombe 1973). Det er den samme bedste rette linje, der går gennem punkterne i alle fire figurer (hældning 0.5 og skæring 3). Desuden har alle 4 datasæt samme $R^2 = 0.667 = 66.7\%$, men det er klart, at de modeller, der er givet ved de fire rette linjer ikke beskriver data lige godt. I den øverste højre figur er sammenhængen mellem x og y åbenlyst ikke-lineær, og sammenhængen i figuren i nederste højre hjørne giver det slet ikke mening at modellere som en ret linje.



Fra dette simple eksempel burde det være åbenlyst, at det ikke giver mening at bruge *værdien* af R^2 alene til at vurdere om en model er god til at beskrive data.

Helt kort og overordnet: R^2 *kan* være OK som et led i at sætte tal på værdien af en statistisk model, men vær varsom!

- R^2 kan i visse situationer fortælle noget relevant om data/situationen — i mangel af en bedre betegnelse, så vil vi i det efterfølgende referere til sådanne situationer som "relevante tilfælde".
- Der er vigtige og centrale begrænsninger i hvad man kan uddrage *alene* af en R^2-værdi, selv inden for de relevante situationer.
- En R^2-værdi bør *aldrig* stå helt alene — kombiner *altid* med visualisering/plot af data. Man kan huske og indprente sig mantraen: *"Man skal tegne før man må regne"*

Det sidste punkt er måske det vigtigste. *Hvis* man vælger at bruge R^2 som et led i at vurdere en model, så *skal* man vide for det første vide, at det ikke er nok blot at udregne værdien. Der skal noget mere eller noget andet til.

Hvad er R^2?

Definitionen af R^2 værdien fremgår i de fleste lærebøger, og er også udførlig beskrevet på eksempelvis Wikipedia, og vi vil ikke gengive formlen her. Flere tekniske detaljer er præsenteret i Brockhoff, Ekstrøm, and Hansen (2017).

For det første kan en R^2-værdi beregnes for såvel den mest simple lineære model med en y-variabel og netop een x-variabel som for mere generelle modeller med flere x-input, de såkaldte multiple lineære regressionsmodeller, og herunder således også de såkaldte polynomielle regressionsmodeller, hvor en ikke-lineær sammensstruktur mellem y og x kan håndteres. Bemærk, at man således godt kan modellere en ikke-lineær relation mellem x og y med en lineær model. Der findes naturligvis også egentlig ikke-linære modeller, men selvom R^2 kan defineres for sådanne ikke-lineære regressionsmodeller, så har den ikke længere sin sædvanlige fortolkning som "forklaringsgrad". Der er yderligere matematiske finurligheder forbundet med såvel beregningen og fortolkningen af sådanne størrelser i forbindelse med egentlig ikke-lineære modeller, altså f.eks. modeller, hvor klassiske ikke-lineære funktioner som logaritme-, eksponential-, sinus - og cosinus-funktioner indgår, og/eller hvor eventuelt flere ukendte elementer af modelfunktionen indgår på en ikke-lineær måde. Det afholder ikke nødvendigvis statistisk software af forskellige slags at anføre en eller anden variant af en R^2-værdi for sådanne modeller. Der er faktisk mange eksempler på, at den fulde indsigt i forskellige statistiske metoders betydning og begrænsning ikke har forplantet sig helt ud i alle hjørner af verden (se Ekstrøm, Hansen, and Brockhoff 2017).

Vi vil i formuleringerne fokusere på det første simple setup i dette notat, altså netop en x variabel og y-variabel men vil ind imellem påpege, hvordan mange af betragtningerne enten kan anvendes direkte eller i tilpasset form til de multiple lineære setups.

1. R^2 er et relativt mål for hvor tæt punkterne gennemsnitlig ligger på den bedste rette linje i et plot af data fra to variable, x og y (målt ved lodrette y-afstande). R^2 giver en værdi mellem 0 (eller 0 %) og 1 (svarende til 100%), hvor 0 svarer til situationen, hvor der ikke er nogen form for lineær sammenhæng mellem x og y, og værdien 1 opnås, når alle punkterne ligger præcist på en ret linje.

2. R^2 er også den kvadrerede (Pearson) korrelationskoefficient mellem x og y (set som en procent), der også bruges i statistik til at beskrive, hvor tæt/stramt punkterne i et plot ligger omkring den bedste rette linje. På hjemmesiden http://guessthecorrelation.com kan man spille sig til en forståelse af, hvad forskellige punktskyer svarer til i korrelation. Kvadrerer man korrelationskoefficienterne i spillet får man således "forklaringsgrader" — og det kan ses som et R^2-spil i stedet. (Alle punktskyer i spillet svarer til "relevante situationer").

3. Når en R^2 naturligvis aldrig i praksis antager værdien 100% (virkelige data vil ikke falde eksakt på en linie), skyldes det faktisk to ting:

    - Den lineære model vil i praksis aldrig være en 100% korrekt model for den virkelighed man forsøger at modellere.

    - Selv hvis den var, så er der variation imellem individuelle y-værdier — flere observationer med samme x-værdi vil variere (f.eks. vil forskellige personer med samme højde (x) typisk have forskellige vægte (y))

4. R^2 er et samlet mål for (summen af) de to slags afvigelser svarende til de to netop nævnte fænomener, men skelner ikke mellem de to, se eksemplet med Anscombes data ovenfor.

5. Hvis man har sikret sig at ens data i situationen ikke er for "mærkelige" *og* også har sikret sig at ikke-lineariteten enten overhovedet ikke kan ses eller er så lille, at den bliver irrelevant, så kan man fint fortolke på R^2-værdien. (Der er dog stadig grænser for *hvad* den kan bruges til).

6. Når den så er relevant, kan man fortolke tallet som den del af y-variationen som x via den statistiske model (den rette linje) kan "forklare" f.eks. vil en vis procentdel af vores vægtforskellighed kunne forklares af vores højdeforskelligheder.

7. For lineære modeller med flere x'er: Alle punkter ovenfor gælder stadig med følgende tilpasninger. Generelt: Erstat "linje" med "hyperplan". I punkt 2: R^2 er den kvadrerede korrelationskoefficient mellem de estimerede modelværdier og y.

Hvad er R^2 så IKKE?

Foruden problemet vist i Ascombes eksempel ovenfor er der andre punkter, man skal være opmærksom på, hvis man har tænkt sig at bruge R^2:

1. R^2 er *ikke* et mål for den direkte kvantitative sammenhæng mellem x og y. R^2 siger altså intet om linjens skæring og hældning, som er de værdier, der beskriver den aktuelle sammenhæng i den relevante kontekst.

2. Ordet "forklaring" i "forklaringsgrad" kan *ikke* forstås som "kausalitet"/"årsags-sammenhæng" — det er alene et mål for den kvantitative sammenhæng. Det kræver helt andre overvejelser omkring den pågældende situation at forsøge at fortolke et resultat kausalt.

3. R^2-tallet kan i sig selv *ikke* fortælle om en lineær model er "korrekt":

    - En lille R^2 kan godt være udtryk for en korrekt lineær gennemsnitssammenhæng, der beskriver et system med en stor variation.
    - En høj R^2 kan godt stadigvæk levne rum for at der ville være en statistisk endnu højere R^2-værdi, hvis man fik fat i den "korrekte" ikke-lineære sammenhæng i en situation.

4. Der findes *ingen* meningsfulde globale kriterier for hvad der er "acceptable" R^2-værdier på tværs af fagområder. En R^2 værdi på 0.65 kan være tilfredsstillende i nogle situationer, mens en R^2-værdi på 0.95 kan være den ønskede grænse i et konkret tilfælde for et andet fagområde. Igen betyder det, at talværdien alene ikke giver os tilstrækkelig information til at vurdere kvaliteten af en model.

5. R^2 er transformations-afhængig: hvis man eksempelvis anvender en log-transformation på y-værdierne vil lineariteten og derved parameterfortolkningen samt R^2 værdien ændre sig.

6. R^2 er ikke en sandhed skåret i granit: R^2 er, som alt andet man beregner, behæftet med statistisk usikkerhed, som der dog for netop R^2s vedkommende ikke er så stor

tradition for at kigge på. Som i alle andre sammenhænge gælder der, at usikkerheden vil være større jo mindre datamængder, der er til rådighed.

7. Man kan skelne mellem situationer hvor man selv har bestemt x-værdierne, f.eks. et dosis-respons forsøg i kemi, og så en situation hvor såvel x som y er tilfældige udfald, f.eks. højde-vægt eksemplet, hvor man ville udtage mennesker tilfældigt, og dernæst måle såvel højde (x) som vægt (y). Lineær regression kan give fin mening i begge situationer, men R^2-værdien (eller tilsvarende korrelationskoefficienten, r) kan have en mere fundamental fortolkning i det sidste tilfælde end i det første. I det sidste kan det (hvis alt ellers er i orden) fortolkes som en grundlæggende biologisk størrelse. I det første kan man faktisk selv langt hen ad vejen bestemme R^2-værdien i de valg af x-værdier man gør: Jo større forskellighed og afstand mellem de selvvalgte x-værdier jo større vil R^2 blive, hvilket betyder, at den person, der laver forsøget kan gøre R^2 større simpelthen ved at sprede x-værdierne ud! Man kan ikke sige at R^2-værdien bliver decideret "forkert" - det er et tal, der er i en-til-en-relation med andre ganske fornuftige beregningsstørrrelser — blot er fortolkningen situationsafhængig.

8. For lineære modeller med flere x'er: Alle punkter ovenfor gælder uden anden tilpasning end at x skal læses og forstås i flertal. R^2 er et problematisk værktøj i forbindelse med modellering generelt, altså i valget mellem forskellige multiple modeller — en R^2-værdi vil altid stige, hvis en model gøres mere nuanceret (et matematisk faktum), så en stigning *alene* kan ikke bruges til noget. Kun når en sådan sammenligning kombineres med andre statistiske værktøjer kan det bruges til noget relevant. Se også bloggen sandsynligvis.dk for flere detaljer om dette (det er skrevet af statistikere, så "en god model" = "en tilstrækkelig korrekt model", uanset hvor stor variationen er, se diskussionen nedenfor).

Et godt alternativ til R^2: spredningen $\hat{\sigma}$

R^2 er som fortalt et relativ mål for hvor tæt modellen ligger på data. Dette anvendes ofte i situationer, hvor skalaen på variablene i sig selv ikke betyder så meget, f.eks. i samfundsfag, sociologi, psykologi, og så videre, hvor det kan være forskellige spørgeskemaskalaer, der er i brug. Taler vi om anvendelser indenfor teknik og naturvidenskab, vil der ofte være ret konkrete skalaer for såvel x som y. I sådanne tilfælde kan det være et godt alternativ at kigge specifikt på den mere direkte eller *absolutte* forskel mellem modellen og data, også kaldet "restspredningen" eller "residualspredningen", $\hat{\sigma}$, der udtrykker den gennemsnitlige (lodrette) afstand mellem datapunkterne og modellinjen. Beregningerne vil vi ikke vise her, men kan findes mange steder, f.eks. (Brockhoff, Ekstrøm, and Hansen 2017). Tallet vil også have en direkte fundamental fortolkning i den anvendte lineære regressionsmodel: i højde-vægt eksemplet, hvor vægten modelleres som en lineær funktion af højden, vil $\hat{\sigma}$ udtrykke vægtspredningen for mennesker med samme fastholdte højde. Dette tal vil typisk være noget mindre end vægtspredningen i populationen som helhed på tværs af alle højder. I Anscombes eksempel ovenfor bliver $\hat{\sigma}$ 1.237, som således kan fortolkes på samme skala og med samme fysiske enhed som y-data kommer med. Værdien for $\hat{\sigma}$ er iøvrigt — præcis som

for R^2 — det samme tal i alle fire tilfælde!

Tallet $\hat{\sigma}$ er således hverken mere eller mindre "rigtigt" eller "forkert" at beregne end R^2, og det kan hverken mere eller mindre benyttes til alle de ting, som vi berører ovenfor. Til gengæld har tallet en fortolkning, der kan være direkte relateret til den konkrete problemstilling, hvilket passer bedre i forhold til punkt 7 ovenfor, og så kan man — i modsætning til R^2 — ikke sådan lige påvirke $\hat{\sigma}$-tallet bare ved at ændre på x-værdierne. Måske vil $\hat{\sigma}$ for mange vil være et tal man lettere kan forholde sig til, og måske man i lidt mindre grad vil være fristet til at drage forhastede konklusioner, hvis man benytter $\hat{\sigma}$ som hvis man bruger R^2. Man kan sige, at det absolutte mål $\hat{\sigma}$ sådan set indgår i det relative mål, som R^2 faktisk er. Omend der faktisk er en lille finurlig men nydelig krølle på dette ræsonnement, se Brockhoff, Ekstrøm, and Hansen (2017). Det er iøvrigt så også et direkte eksempel på det fundamentale begreb varians og/eller spredning, som nok fortjener lidt større bevågenhed i uddannelsessystemet, herunder på gymnasieniveau (Ekstrøm, Hansen, and Brockhoff 2017).

"Omvendt" regression?

Der kan i konkrete tilfælde med to variable u og v opstå en overvejelse omkring, hvilken der skal tage rollen som x og hvilken som y. Den nysgerrige studerende kunne spørge sig selv og/eller sin lærer: "hvad sker der egentlig, hvis man vender det om, og ombytter rollerne for de to variable?" Man kan forholde sig til denne overvejelse på to niveauer: hvad der sker rent beregningsmæssigt, og hvad der i forhold til den kontekstspecifikke anvendelse er det mest relevante. R^2-værdien, og tilsvarende korrelationskoefficienten afhænger ikke af hvordan tingene vender, men selve estimatet for den bedste rette linje og $\hat{\sigma}$-beregningen vil give to forskellige ting. Det kræver muligvis lidt forståelsesmæssig tilvænning, men det giver faktisk god mening: Det er to forskellige ting at finde den bedste rette linje som beskriver vægt som en lineær funktion af højde, hvor man minimerer vægtafvigelser, og så at finde den bedste rette linje, som beskriver højde som funktion af vægt, hvor man minimerer højdeafvigelser. Der er præcise matematiske relationer mellem de to løsninger.

Folk med matematisk baggrund kender sikkert til muligheden for at finde en helt tredje beregningsvariant, der ligger præcis midt imellem de to andre, og som minimerer de vinkelrette afstande til den rette linje. Denne kommer som en konsekvens af en analysemetode, der også kaldes principal komponent analysis (PCA), som faktisk bruges i stor stil til at eksplorativ analyse af højdimensionale data og til dimensionsreduktion. Men PCA er faktisk ikke i sig selv en regressionsmetode, og den PCA-baserede linje er ikke det korrekte svar på nogen af de to oplagte fagspecifikke spørgsmål: Hvad er modellen for u som funktion af v eller hvad er modellen for v som funktion af u? Det korrekte svar på hvert disse spørgsmål er det tilsvarende valg af den "asymmetriske" beregning, hvor den ene får y-rollen, og den anden x-rollen.

Denne diskussion skal ses i forhold til punkt 7 ovenfor. Hvis man selv har bestemt x-værdierne, så har R^2, som beskrevet, ikke så god en fortolkning, og den kontekstspecifikke problemstilling, altså hvad der er x og hvad der er y er defineret fra starten. I den anden mere symmetriske (x,y)-situation, kan begge veje give teoretisk lige god mening, og det er således alene den

kontekstspecifikke betragtning, der skal afgøre hvilket spørgsmål man vil besvare, og så lave beregningerne og konklusionerne derefter.

Hvordan sikrer man sig at man er i et "relevant tilfælde"?

Der findes desværre ikke et og kun et tal, man kan beregne, der kan besvare dette spørgsmål med et klart ja eller nej. Det er en del af den kompleksitet man må vænne sig til omkring brugen af "statistisk ræssonering", se Ekstrøm, Hansen, and Brockhoff (2017). Der er mange redskaber, der forsøger at belyse forskellige aspekter af om en model er god, og på gymnasieniveau skal man finde en passende simpel måde at håndtere dette.

Det primære værktøj er visualisering af selve (x,y)-relationen: ser punktskyen nogenlunde lineær og "samlet" ud? Hvordan ser modelafvigelserne ud, når de plottes mod de forventede værdier, og/eller mod x-inputs: Er de tilstrækkelig uden struktur? Er der ingen enkeltafvigelser, der er helt ekstreme? Og ser de ellers ud til at følge en normalfordeling? Det sidste kunne vurderes i boxplots og histogrammer af afvigelserne.

Det er ikke nemt pædagogisk og præcist at indkredse denne del af den statistiske proces. Hvad angår undersøgelsen af om den lineære model er tilstrækkelig korrekt kan man også anvende modellering med mere komplekse modeller for helt konkret at vurdere om de mere komplekse modeller faktisk er nødvendige. Hvis ikke, kan man med mere ro i sindet anvende den lineære. Helt konkret kunne man tilpasse en mere generel funktion til data, og så plotte denne tilpasning sammen med nogle konfidensgrænser for sammenhængen, og derefter vurdere, om man med rimelighed kan antage, at den rette linie er at finde indenfor konfidensbåndene. Vi er med på, at dette ikke ligger inden for almindelig gymnasiepensum, men hvis man udvidede pensum til at omfatte multipel regressionsanalyse ville dette falde indenfor.

Kommunikationsudfordring

Vi tror, at en del af forklaringen på de gnidninger, der måske opstår mellem faggrupper, kan være af kommunikationsmæssig karakter. Måske forskellige folk lægger forskellige ting i ord som "en god model" versus "en dårlig model", og tilsvarende et begreb som en "korrekt model". Model, som begreb og som ord, kan naturligvis også betyde vidt forskellige ting afhængig af sammenhængen det indgår i. Man man forestille sig, at en matematiker/statistiker naturligt vil sætte lighedstegn mellem "en god model" og en "tilstrækkelig korrekt model", mens anvendere i de faglige miljøer, f.eks. samfundsfag eller andet kan forestilles at mene, at "en god model" er lig med en model, der *både* er tilstrækkelig korrekt *og* har en lille variation, så den ligger konkret tæt på data. Begge betragtninger giver på sin vis ganske god mening! Idet R^2 måler begge dele i et samlet mål, så er det på den ene side et fornuftigt mål for anvenderen, men på den anden side skelner målet ikke mellem de to bidrag, og matematikeren/statistikeren har dermed helt ret i, at R^2 *ikke* er noget godt mål for "korrektheden" af modellen, og at der dermed kan gemme sig nuancer nedenunder, som man kan overse uden denne nuancering.

At tale om en "tilstrækkelig korrekt model" er en typisk fagstatistisk terminologi og tankegang, hvor udgangspunktet ofte er at *alle modeller er forkerte, men nogen er brugbare* (Box and Draper 1987, 424). Det er en tankegang, der muligvis hos nogle naturvidenskabsfolk, der søger de "sande mekanismer" ="korrekte modeller" kan være lidt fremmed. Men det afspejler nok, at rigtig mange af de komplekse problemstillinger, som søges løst med statistiske og matematiske modeller i samfundsmæssige, industrielle og forskningsmæssige sammenhænge ikke lader sig løse med en enkelt universel, kendt og veldefineret sand/korrekt model. Der behøver dog ikke være nogen rigtig modstrid i de to tankegange. Det kan være særdeles fornuftigt at søge at beskrive fænomener med kendte modeller, hvad enten det er fysiske, kemiske, biologisk eller andre typer modeller. Man kan nogen gange beskrive en del af strukturerne i et fænomen med kendte og velunderbyggede modeller, og lade resten modelleres af mere empirisk baserede modeller for resterende struktur og variation. Så længe man ikke lader sig "teoriforblænde" af modeller, der alene på grund af diverse historiske årsager og begrænset information har tilkæmpet sig uretmæssige forskningsmæssige positioner.

Det er vigtigt at fortælle den samme historie.

Det vigtigste må være, at de studerende lærer noget, som 1) de forstår hvad måler, og som de 2) har kompetencen til at bruge (og vide, hvornår man ikke kan bruge). Det bør derfor tilstræbes, at de forskellige fagmiljøer — hvis man fortsat vælger at bruge R^2 som et led i statistikundervisningen i gymnasiet — fortæller den *samme* historie omkring R^2.

Desværre findes der ikke en simpel, objektiv måde at vurdere korrektheden af en statistisk model på, men det understreger blot vigtigheden af at alle faggrupper er i stand til at formidle alle de fordele og ulemper, der måtte være, ved den valgte metode.

Referencer

Anscombe, F. J. 1973. "Graphs in Statistical Analysis." *American Statistician* 27: 17–21.

Box, G. E. P., and N. R. Draper. 1987. *Empirical Model-Building and Response Surfaces*. John Wiley; Sons.

Brockhoff, Per Bruun, Claus Thorn Ekstrøm, and Ernst Hansen. 2017. "Lineær Regression: Lidt Mere Tekniske Betragtninger Om R^2 Og et Godt Alternativ." *LMFK-Bladet*.

Ekstrøm, Claus Thorn, Ernst Hansen, and Per Bruun Brockhoff. 2017. "Statistik I Gymnasiet." *LMFK-Bladet*.