Technical University of Denmark

Lineær regression – lidt mere tekniske betragtninger om R^2 og et godt alternativ

Brockhoff, Per B.; Ekstrøm, Claus Thorn; Hansen, Ernst

Published in:
LMFK-Bladet

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

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.

Lineær regression

– lidt mere tekniske betragtninger om R^2 og et godt alternativ

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

Dette ekstra lille notat om den såkaldte R^2–værdi, som kan beregnes i forbindelse med lineær regression, skal ses i sammenhæng med vores ikke–tekniske notat om samme emne . Ud over at få defineret tingene matematisk præcist, vil vi foreslå spredningen σ som et godt alternativ. De to hænger nært sammen, måler for så vidt det samme, R^2 på en relativ måde og σ på en absolut måde. Spredningen σ kan ses i ret direkte sammenhæng med usikkerhedsbetragtninger mere generelt, som vi i det store billede mener er ret vigtige.

Definition af R^2 i den simple lineære regressionssituation

Lad os lige minde om hvad vi overhovedet taler om. Den simpleste forekomst af R^2 optræder i den lineære regressionsmodel

$$y_i = \alpha + \beta x_i + \varepsilon_i,\; i = 1,\ldots,n$$

Til hver måling y_i er der knyttet en kovariat x_i, og man kan ønske at undsøger om kovariaten har en lineær påvirkning af målingen og i givet fald at kvantificere og fortolke sammenhængen og måske at benytte den til at forudsige y–værdien for nye x–værdier. Parametrene α og β er ukendte, og analysen af regressionsmodellen fokuserer normalt på at estimere dem. De tilbageværende størrelser $\varepsilon_1,\ldots,\varepsilon_n$ er såkaldte støjvariable, der skal redde modellen fra at kollapse i mødet med virkeligheden, hvor parrene (x_i, y_i) jo aldrig ligger præcis på en matematisk ret linje. Den sædvanlige antagelse om støjvariablene er, at de er uafhængige, og at de er normalfordelte med middelværdi 0 og samme varians σ^2 (endnu en parameter i modellen). I denne ramme defineres R^2 ved formlen

$$R^2 = \frac{SS_{xy}^2}{SS_{xx}\cdot SS_{yy}} \qquad (1)$$

hvor SS_{xy}, SS_{xx} og SS_{yy} er nogle af de standard beregningsstørrelser, man alligevel ofte regner ud i forbindelse med estimation af de tre parametre α, β og σ^2:

$$\widehat{\beta} = \frac{SS_{xy}}{SS_{xx}} \qquad (2)$$

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x} \qquad (3)$$

hvor

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad (4)$$

$$SS_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 \qquad (5)$$

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \qquad (6)$$

Disse resultater er også velkendte fra mindste kvadraters metode, og giver modellens estimerede hældning og skæring (på baggrund af de tilgængelige data). Med de estimerede parametre kan vi bruge modellen til at udregne de forventede værdier, $\widehat{y}_i$, der beskriver, hvad vi i gennemsnit forventer at observere for en given x_i–værdi:

$$\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}x_i$$

Matematik

Med disse størrelser kan vi estimere variansen, der er baseret på forskellen mellem de reelle observationer, y_i, og de forventede observationer (på baggrund af modellen og de tilhørende x_i'er)

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (7)$$

R^2 for mere komplicerede modeller

Ovenfor er R^2 defineret i det simple lineære regressionssetup, men R^2 kan også benyttes for mere komplicerede modeller med flere forklarende variable, for eksempel P forklarende variable, dvs. $\mathbf{x}_i = (x_{i1}, \ldots, x_{iP})$, så længe man stadig er inden for klassen af lineære modeller. En lineær model refererer til, at sammenhængen mellem y og $\mathbf{x}$ kan skrives i et lineært ligningssystem

$$y_i = \alpha + \sum_{p=1}^{P}\beta_p x_{ip} + \varepsilon_i,\ i = 1,\ldots,n$$

og vil derfor også dække specialtilfælde som eksempelvis polynomial regression

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i,\ i = 1,\ldots,n$$

Bemærk, at man således godt kan modellere en ikke–lineær relation mellem x og y med en lineær model. Der findes naturligvis også egentlige ikke–linære modeller, men selvom R^2 kan defineres for sådanne ikke–lineære regressionsmodeller, så har den ikke længere sin sædvanlige fortolkning som "forklaringsgrad". Formlen nedenfor gælder ikke længere, og summen af residualerne er ikke længere nul. Detaljerne i disse yderligere (ikke–linære) udfordringer ved forståelsen og brugen af R^2 er ikke berørt nærmere hverken her eller i vores ikke-tekniske notat.

Lidt flere detaljer om R^2 for linære modeller

R^2 er også den kvadrerede korrelation mellem y–værdier og de forventede y–værdier i modellen for de x'er man har med, $\hat{y}_i$–værdierne. Denne definition gælder også for de mere generelle lineære modeller med flere x–variable.

R^2 er givet ved y–variationer, og dem findes der to/tre af – to af dem summerer til den tredje:

$$SST = SSM + SSE$$

hvor

$$SST = SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$SSM = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Bemærk at SS(Total) = SST udtrykker y–variationen uden nogen x–indblanding, og bemærk, at hvis man dividerer SST med n – 1, så har man den klassiske beregning af en stikprøvevarians anvendt på y–data. SSM er variationen givet ud fra x'erne, og SSE er den såkaldte restvariation, der udtrykker forskellen mellem linje-værdierne $\hat{y}_i$ og data y_i. Det er denne SSE–værdi man har minimeret, når man har fundet den bedste rette linje ved hjælp af mindste kvadraters metode – den er så lille som det er muligt med de data man har. Nu kan man så skrive præcist hvad R^2 faktisk er på en lidt anden vis end ovenfor:

$$R^2 = \frac{(SST - SSE)}{SST} = 1 - \frac{SSE}{SST}$$

Så R^2 er givet ved forholdet mellem disse to y–variationer: y–variationen, som den nu engang kommer, og restvariationen i y, når man har fjernet det, som x kan forklare gennem linjen (eller en mere generelle model, hvis en sådan er i spil – det gør *ingen* forskel for udtrykkene her). Og heraf fortolkningen: *Forklaringsgrad*

Heraf kan man også linke til teori omkring variation og varianser/spredninger: En matematiker vil vide, at en varians (som teoretisk begrebsmæssigt er et integral) ikke er transformationsinvariant: Anvender man en ikke–lineær transformation af skala/data, vil disse tal naturligvis ændre sig på en ikke–simpel måde. Selvom det ikke fremgår så direkte af ovenstående, så afhænger R^2 naturligvis også af x–værdierne – det er gennem $\hat{y}_i$–værdierne denne afhængighed kan ses. Hvis enten alle y–værdierne er ens, så SST = 0, og punkterne ligger eksakt på en horisontal linje, eller alle x–værdierne er ens, så er R^2 ikke defineret.

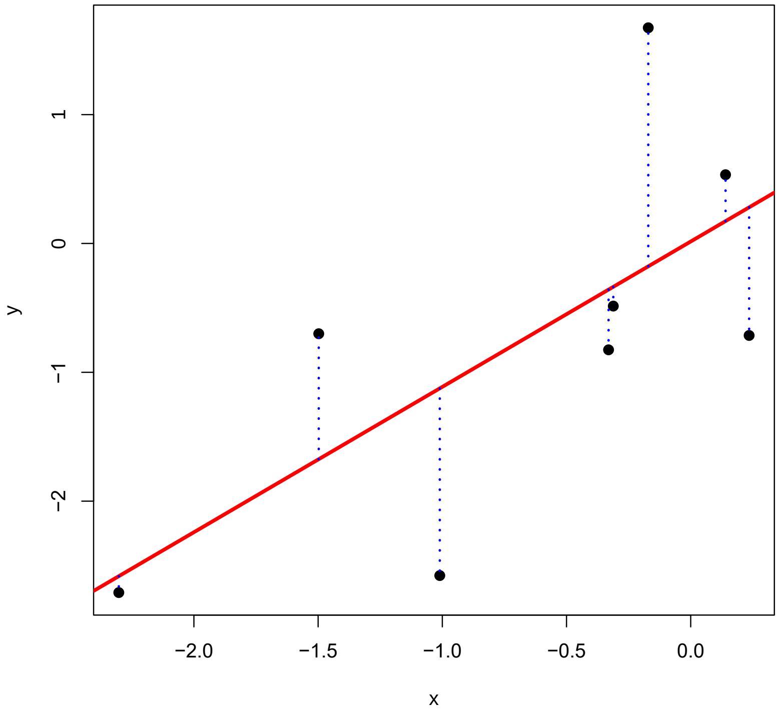
Lineær regression kan have forskellige formål — stikord

Der kan være forskellige årsager til, at man laver lineær regression, og derfor kan fokus også ændre sig lidt. Formålet kan eksempelvis være at

- **afdække** om der er en "sammenhæng" mellem to variable (hermed underforstået, at vi leder efter en *lineær* sammenhæng). Nogle ville kalde dette for "korrelationsanalyse", og fokusere på korrelationen og ikke på selve linjen (og måske lave et hypotese–test for om korrelation = 0). I denne situation – der oftest bliver brugt i socio– og samfundsfagssammenhænge – er man udelukkende interesseret i at vurdere, om der er en sammenhæng, og derfor opfattes x'erne og y'erne i modellen i princippet symmetrisk: hvis man laver en tilsvarende lineær regressionsanalyse, hvor man modellerer x som lineær funktion af y, så opnår man samme resultat.

- **kvantificere** den underliggende lineære relation mellem middelværdierne af y og x for fortolkningens skyld eller evt. at kunne "interpolere", altså skønne/estimere eksempelvis middelvægten for personer af en højde man ikke lige fik med i stikprøven (men stadig inden for range af data – man skal være varsom med at ekstrapolere). Her kan man beregne linjen, og kombinere med konfidensintervaller (Bemærk: hypotese–testet for hælding = 0 er det samme som for korrelation = 0).

- bruge modellen til at **prædiktere** nye cases, der kommer til — altså beregne $\hat{y}_{\text{ny}}$ for en konkret x_{ny}–værdi (beregne linjen, og kombinere med prædiktionsintervaller, der også direkte involverer spredningen).



Figur 1
De absolutte vertikale afstande (de blå stiplede linjer) måler, hvor langt den lineære regressionsmodel (den røde linje) ligger fra de observerede data, og de har samme skala som y. Spredningen $\hat{\sigma}$ udtrykker den gennemsnitlige værdi af disse.

Formålet kan/bør måske også ses i en lidt større sammenhæng som at besvare: *"Hvad er den rette model?"* Her vil målet være at finde den rette model, dernæst kvantificere elementerne i denne model, og så kan vi evt til sidst enten "estimere" eller "prædiktere", hvis vi ønsker.

Et alternativ til R^2

R^2 er som vist et relativ mål for hvor tæt modellen ligger på data. Dette anvendes ofte i situationer, hvor skalaen på variablerne ikke i sig selv betyder så meget, fx i samfundsfag, sociologi, psykologi, etc., hvor det kan være forskellige spørgeskema–skalaer, der er i brug. Taler vi om anvendelser inden for teknik og naturvidenskab, vil der ofte være ret konkrete skalaer for såvel *x* som *y*. I sådanne tilfælde kan følgende alternativ være en god ide.

Vi giver herunder et forslag til at benytte et absolut mål for afvigelsen mellem model og data fremfor det relative mål som R^2 faktisk er. Men før dette bør man gøre sig klart, at det i langt de fleste tilfælde er selve linjen, der vil være det mest interessante i en konkret sammenhæng. Derfor er estimation/beregning af linjen helt centralt. Dernæst vil det mest relevante være at kvantificere usikkerheden i bestemmelsen af linjen, noget vi typisk ville gøre ved at beregne stikprøveusikkerhederne for afskæring og hældning, for derefter evt. at udtrykke disse i konfidensintervaller for disse to størrelser (i praksis er det oftest hældningen, der udtrykker noget spændende). Den centrale størrelse der indgår i formlerne for disse usikkerheder er netop det absolutte mål, der præsenteres nu (se Figur 1).

Hvis man bruger *SSE* absolut set i stedet for relativt, *SSE*/*SST*, og beregner:

$$\hat{\sigma} = \sqrt{\frac{SSE}{(n-2)}}$$

så har man faktisk estimeret "den underliggende spredning" for *y*–værdierne (for en fastholdt *x*–værdi), som desuden er en parameter i den klassiske formelle statistiske model, der kan ligge bagved:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \;\; \varepsilon_i \sim N\left(0, \sigma^2\right)$$

Man har således på denne vis kvantificeret den gennemsnitlige afstand mellem *y*–værdier og linjen direkte, og det vil være klart for de fleste med teknisk/naturvidenskabelig baggrund, at det er et tal, der kommer med den samme fysiske enhed som *y*–værdien kommer med fra starten. På den vis bliver fx skalaafhængigheden meget direkte tydelig for enhver, og tallet har en rigtig god fortolkning.

Det er klart, at da tallet her for en given total *y*–variation er ækvivalent med R^2–værdien, så er det hverken mere eller mindre "rigtigt" at beregne end R^2, og hvis man forsøger at bruge σ til at besvare de spørgsmål, vi har anført ovenfor, løber man ind i samme problemer som med R^2.

Men måske det 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 ud fra dette tal end man kan være med R^2.

En lille krølle er følgende: Skulle man nu alligevel få tanken, at man gerne i tillæg vil fortolke (residual)spredningen relativt til den spredning, som *y*–værdierne har uden indblanding af *x*'erne:

$$\hat{\sigma}_y = \sqrt{\frac{SS_{yy}}{n-1}}$$

Altså tilbage til den relative fortolkning som R^2 egentlig har, og fx beregne $\hat{\sigma}^2 / \hat{\sigma}_y^2$, så har man faktisk beregnet den såkaldte "Adjusted R^2", som mange software-pakker helt standard i tillæg vil beregne for sådanne modeller, ellere rettere $1-R^2_{\text{adj}}$:

$$\begin{aligned}\frac{\hat{\sigma}^2}{\hat{\sigma}_y^2} &= \frac{SSE/(n-P-1)}{SST/(n-1)} \\ &= \frac{(n-1)}{(n-P-1)}\frac{SSE}{SST} \\ &= \frac{(n-1)}{(n-P-1)}\left(1-R^2\right) = 1-R^2_{adj}\end{aligned}$$

hvor P er antallet af x–variabler i modellen. Og på flere måder er den adjustede R^2 faktisk at foretrække frem for den "almindelige". For simple lineære regressioner (P = 1) med stort n, er der ikke nogen væsentlig forskel på de to, idet $(n-1)/(n-2)$ således vil være tæt på 1.

Usikkerhedsbetragtninger

Et af de helt centrale budskaber i "statistik som fagområde" er, at alt vi beregner på og uddrager af data er behæftet med en eller anden form for usikkerhed/variation. Faktisk er dette mere centralt end det klassiske hypotesetest, som ofte som metode lidt uhensigtsmæssigt kan blive synonym med "statistik", se også Ekstrøm et al. (2017). Dette gælder således ligeledes beregningsstørrelser i regressionssammenhænge. Den beregnede spredning $\hat{\sigma}$ indgår centralt i de relevante usikkerhedsberegninger for såvel afskæring $\hat{\alpha}$, hældning $\hat{\beta}$ og linjeberegninger:

$$\begin{aligned}\hat{\sigma}_{\hat{\alpha}} &= \hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{SS_{xx}}} \\ \hat{\sigma}_{\hat{\beta}} &= \hat{\sigma}\sqrt{\frac{1}{SS_{xx}}} \\ \hat{\sigma}_{\hat{\alpha}+x_0\hat{\beta}} &= \hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_{xx}}} \\ \hat{\sigma}_{\hat{\alpha}+x_0\hat{\beta}+\epsilon_0} &= \hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_{xx}}}\end{aligned}$$

Det er naturligvis hverken hensigten her at forklare og bevise alle disse formler eller at man som gymnasieelev i Danmark skal lære detaljerne af dette. I mange såkaldte "non–calculus" baserede statistikkurser på indledende universitetsniveau for mange studieretninger verden over angives disse og lignende formler ligeledes uden bevis. Det kræver dog ikke andet end nogle lineære varians–regneregler eller, om man vil, lineære fejlophobningsbetragtninger. Alle disse spredninger kaldes også nogen gange for "standard errors" eller "stikprøvespredninger" – de udtrykker hvor meget en beregnet størrelse forventes at variere "fra stikprøve–til–stikprøve", altså hvor usikkert bestemt den egentlig er.

Vi viser formlerne her for at understrege den fundamentale betydning af spredningen, som indgår på samme måde i alle formlerne. Og alle usikkerhedsformlerne er udvidede versioner af den samme og helt fundamentale usikkerhedsformel for et simpelt stikprøvegennemsnit, fx udtrykt ved y: (uden indblanding af x)

$$\hat{\sigma}_{\bar{y}} = \frac{\hat{\sigma}_y}{\sqrt{n}} = \hat{\sigma}_y\sqrt{\frac{1}{n}}$$

De fleste indledende statistikkurser vil introducere de grundlæggende statistiske begreber som hypotesetests og konfidensintervaller i dette simpleste af alle setups. Konfidensintervaller er det konkrete statistiske redskab man kan tage i anvendelse for at formalisere usikkerhedsbetragtninger ved hjælp af sandsynlighedsteori. Igen er det ikke hensigten at give en udtømmende gennemgang her, men fx bliver $(1-\alpha)$ konfidensintervallerne for α og β

$$\begin{aligned}\hat{\alpha} &\pm t_{1-\alpha/2}\hat{\sigma}_{\hat{\alpha}} \\ \hat{\beta} &\pm t_{1-\alpha/2}\hat{\sigma}_{\hat{\beta}}\end{aligned}$$

hvor $t_{1-\alpha/2}$ er $(1-\alpha/2)$–fraktilen for en t–fordeling med $n-2$ frihedsgrader. t–fordelingen kan løst siges at være "en version af" standardnormalfordelingen, der tager højde for at variansen, der indgår er estimeret fra data, og altså i sig selv er behæftet med usikkerhed. Hvis den ikke var det, ville normalfordelingen kunne anvendes for alle usikkerhedsberegningerne, idet lineære transformationer af normalfordelinger igen er normalfordelinger. I praksis vil disse konfidensintervalformler ofte tilnærmelsesvis have formen:

$$\bar{y} \pm 2\hat{\sigma}_{\bar{y}}$$

altså hvor man har et estimat for en parameter (her gennemsnittet) plus/minus 2 gange usikkerheden på estimatet. Ved at bruge normalfordelingen kan man se, at disse grænser omtrentlig vil svare til "95 % konfidens", og denne fundamentale relation kan være god at kommunikere ud.

Og som en sidste lille perspektiverende gymnasiekrølle: Det er standard metodik i indledende statistikkurser, at man under visse normalfordelingsforudsætninger kan kvantificere usikkerheden i sprednings– og variansberegninger i sig selv ved brug af χ^2–fordelingen. Alstå den samme fordeling som anvendes til det klassiske "χ^2–test". Det er en matematisk/sandsynlighedsteoretisk konsekvens af at kvadrere normalfordelinger. Og pudsigt nok er der ingen global tradition for tilsvarende at kvantificere usikkerheden i en R^2–beregning, selvom den, som sammenhængene ovenfor viser, på helt samme måde er behæftet med usikkerhed. Og dermed er der uden tvivl en større risiko for at denne usikkerhed bliver glemt i skyndingen, end tilsvarende for $\hat{\sigma}$.

Referencer

Brockho Per Bruun, Hansen Ernst, Ekstrm Claus Thorn. *Brugen af R^2 i gymnasiet*, LMFK–bladet. 2/2017.

Ekstrm Claus Thorn, Hansen Ernst, Brockho Per Bruun. Statistik i gymnasiet, LMFK–bladet. 2/2017.