Lineær regression, kvadratsummer og forklaringsgrad

Når man udfører lineær regression, søger man som bekendt den lineære funktion på formen $f(x)=a\cdot x+b$, som minimerer følgende kvadratsum:

$$SS_{res}=\sum_{n=1}^{n}\left(y_i-f(x_i)\right)^2 \tag{1}$$

Opgaven er altså at finde a og b, så summen bliver mindst mulig. Forskellen mellem punkternes y-værdier og funktionsværdierne for den lineære funktion i de tilhørende x-værdier kaldes for *residualerne* og beregnes med r_i. Vi har altså:

$$r_i=y_i-f(x_i) \tag{2}$$

Med denne definition har vi altså, at

$$SS_{res}=\sum_{n=1}^{n}r_i^2 \tag{3}$$

Sætning 1 (Lineær regression)

Givet en række datapunkter $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$. Den lineære funktion på formen $f(x)=a\cdot x+b$, som minimerer kvadratsummen (1) er givet ved:

$$a=\frac{\sum_{n=1}^{n}(x_i-\overline{x})\cdot(y_i-\overline{y})}{\sum_{n=1}^{n}(x_i-\overline{x})^2} \tag{4}$$

$$b=\overline{y}-a\cdot\overline{x} \tag{5}$$

hvor $\overline{x}$ og $\overline{y}$ er middelværdierne af henholdsvis x-værdierne og y-værdierne:

$$\overline{x}=\sum_{n=1}^{n}x_i \tag{6}$$

$$\overline{y}=\sum_{n=1}^{n}y_i \tag{7}$$

Bevis: Vi skal ikke give det her. Det kan findes mange steder.

□

I dette tillæg skal vi kigge på to andre kvadratsummer end (1), nemlig:

$$SS_{tot}=\sum_{n=1}^{n}\left(y_i-\overline{y}\right)^2 \tag{8}$$

$$SS_{reg}=\sum_{n=1}^{n}\left(f(x_i)-\overline{y}\right)^2 \tag{9}$$

Før vi kommer til en interessant identitet involverende de tre kvadratsummer, skal vi først se på nogle andre interessante hjælpesætninger.

Sætning 2

Middelværdien af residualerne er 0:

(10) $$\sum_{n=1}^{n} r_i \;=\; 0$$

Bevis:

$$\sum_{n=1}^{n} r_i \;=\; \sum_{n=1}^{n}\left(y_i - f(x_i)\right) \;=\; \sum_{n=1}^{n}\left(y_i - a\cdot x_i - b\right) \;=\; \sum_{n=1}^{n} y_i - a\cdot\sum_{n=1}^{n} x_i - n\cdot b$$
$$=\; n\cdot\overline{y} - a\cdot n\cdot\overline{x} \;- n\cdot\left(\overline{y} - a\cdot\overline{x}\right) \;=\; 0$$

hvor vi i fjerde lighedstegn har benyttet (5).

□

Sætning 3

(11) $$\sum_{n=1}^{n} f(x_i)\cdot(x_i - \overline{x}) \;=\; \sum_{n=1}^{n} y_i\cdot(x_i - \overline{x})$$

Bevis: Vi regner på venstresiden:

$$\sum_{n=1}^{n} f(x_i)\cdot(x_i - \overline{x}) \;=\; \sum_{n=1}^{n}(a\cdot x_i + b)\cdot(x_i - \overline{x})$$
$$=\; \sum_{n=1}^{n}(a\cdot x_i + \overline{y} - a\cdot\overline{x})\cdot(x_i - \overline{x})$$
$$=\; \sum_{n=1}^{n}\left(\overline{y} + a\cdot(x_i - \overline{x})\right)\cdot(x_i - \overline{x})$$
$$=\; \overline{y}\cdot\sum_{n=1}^{n}(x_i - \overline{x}) \;+\; a\cdot\sum_{n=1}^{n}(x_i - \overline{x})^2$$
$$=\; \overline{y}\cdot\sum_{n=1}^{n}(x_i - \overline{x}) \;+\; \sum_{n=1}^{n}(y_i - \overline{y})\cdot(x_i - \overline{x})$$
$$=\; \overline{y}\cdot\sum_{n=1}^{n}(x_i - \overline{x}) \;+\; \sum_{n=1}^{n} y_i\cdot(x_i - \overline{x}) \;-\; \overline{y}\cdot\sum_{n=1}^{n}(x_i - \overline{x})$$
$$=\; \sum_{n=1}^{n} y_i\cdot(x_i - \overline{x})$$

hvor vi har brugt følgende: 1. lighedstegn: (5). 4. lighedstegn: (4).

□

Sætning 4

(12) $$\sum_{n=1}^{n} r_i \cdot (x_i - \overline{x}) \;=\; 0$$

Bevis: Fås straks ved at trække venstresiden fra højresiden i (11) fra sætning 3:

$$0 \;=\; \sum_{n=1}^{n} \big(y_i - f(x_i)\big) \cdot (x_i - \overline{x}) \;=\; \sum_{n=1}^{n} r_i \cdot (x_i - \overline{x})$$

□

Sætning 5

(13) $$\sum_{n=1}^{n} r_i \cdot x_i \;=\; 0$$

Bevis: Vi bruger resultatet i sætning 4:

$$0 \;=\; \sum_{n=1}^{n} r_i \cdot (x_i - \overline{x}) \;=\; \sum_{n=1}^{n} r_i \cdot x_i \;-\; \overline{x} \cdot \sum_{n=1}^{n} r_i \;=\; \sum_{n=1}^{n} r_i \cdot x_i$$

hvor vi for at få sidste lighedstegn har benyttet sætning 2.

□

Sætning 6 (Sætning med kvadratsummer)

(14) $$SS_{tot} \;=\; SS_{reg} + SS_{res}$$

Bevis: Vi regner på venstresiden:

$$\begin{aligned}
SS_{tot} \;&=\; \sum_{n=1}^{n} (y_i - \overline{y})^2 \\
&=\; \sum_{n=1}^{n} \big[\, y_i - f(x_i) + f(x_i) - \overline{y} \,\big]^2 \\
&=\; \sum_{n=1}^{n} \big(y_i - f(x_i)\big)^2 \;+\; \sum_{n=1}^{n} \big(f(x_i) - \overline{y}\big)^2 \;+\; 2 \cdot \sum_{n=1}^{n} \big(y_i - f(x_i)\big) \cdot \big(f(x_i) - \overline{y}\big) \\
&=\; SS_{res} \;+\; SS_{reg} \;+\; 2 \cdot \sum_{n=1}^{n} r_i \cdot (a \cdot x_i + b - \overline{y}) \\
&=\; SS_{res} \;+\; SS_{reg} \;+\; 2a \cdot \sum_{n=1}^{n} r_i \cdot x_i + 2 \cdot (b - \overline{y}) \cdot \sum_{n=1}^{n} r_i \\
&=\; SS_{res} \;+\; SS_{reg}
\end{aligned}$$

hvor vi i 4. lighedstegn har udnyttet definitionerne (1), (9) samt (2). For at få sidste lighedstegn har vi udnyttet hjælpesætningerne 2 og 5.

□

Definition 7 (Forklaringsgraden)

Givet følgende sæt af datapunkter: $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$. *Forklaringsgraden* R^2 for den regressionslinje, som hører til datapunkterne, er defineret ved:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{15}$$

Vi skal give en fortolkning af denne størrelse snart, men først skal vi give et par andre formler for forklaringsgraden.

Sætning 8

Der gælder følgende alternative formler for forklaringsgraden R^2 for regressionslinjen, som hører til sættet af datapunkter $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$:

$$R^2 = \frac{SS_{reg}}{SS_{tot}} \tag{16}$$

$$R^2 = a^2 \cdot \frac{\sum_{n=1}^{n}(x_i-\overline{x})^2}{\sum_{n=1}^{n}(y_i-\overline{y})^2} = a^2 \cdot \frac{\overline{x^2}-\overline{x}^2}{\overline{y^2}-\overline{y}^2} \tag{17}$$

hvor a er hældningskoefficienten for regressionslinjen.

Bevis: (16) fås direkte af sætning 6:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{tot}-SS_{res}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}}$$

For at vise formlerne i (17), bemærker vi først følgende:

$$f(x_i)-\overline{y} = a\cdot x_i + b - \overline{y} = a\cdot x_i + (\overline{y}-a\cdot\overline{x}) - \overline{y} = a\cdot(x_i-\overline{x}) \tag{18}$$

Vi regner videre på (16):

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = \frac{\sum_{n=1}^{n}\left(f(x_i)-\overline{y}\right)^2}{\sum_{n=1}^{n}\left(y_i-\overline{y}\right)^2} = \frac{\sum_{n=1}^{n}a^2\cdot(x_i-\overline{x})^2}{\sum_{n=1}^{n}\left(y_i-\overline{y}\right)^2} = a^2\cdot\frac{\sum_{n=1}^{n}(x_i-\overline{x})^2}{\sum_{n=1}^{n}\left(y_i-\overline{y}\right)^2} \tag{19}$$

hvorved den første identitet er vist. For at vise den sidste i (17), omskriver vi først tælleren i brøken:

$$\sum_{n=1}^{n}(x_i-\overline{x})^2 = \sum_{n=1}^{n}\left(x_i^2-2\cdot\overline{x}\cdot x_i+\overline{x}^2\right) = \sum_{n=1}^{n}x_i^2-2\cdot\overline{x}\cdot\sum_{n=1}^{n}x_i+\overline{x}^2\cdot\sum_{n=1}^{n}1$$
$$= \sum_{n=1}^{n}x_i^2-2\cdot\overline{x}\cdot n\cdot\overline{x}+\overline{x}^2\cdot n = \sum_{n=1}^{n}x_i^2-n\cdot\overline{x}^2 = n\cdot\left(\overline{x^2}-\overline{x}^2\right)$$

Nævneren omskrives helt på samme vis, hvorefter vi kan regne videre på (19):

$$R^2 = a^2\cdot\frac{\sum_{n=1}^{n}(x_i-\overline{x})^2}{\sum_{n=1}^{n}(y_i-\overline{y})^2} = a^2\cdot\frac{n\cdot\left(\overline{x^2}-\overline{x}^2\right)}{n\cdot\left(\overline{y^2}-\overline{x}^2\right)} = a^2\cdot\frac{\overline{x^2}-\overline{x}^2}{\overline{y^2}-\overline{x}^2}$$

Hermed er den sidste identitet vist.

□

Vi er rede til at give en fortolkning af forklaringsgraden direkte ud fra definition 17. Den *totale kvadratsum* (Total Sum of Squares) SS_{tot}, den *residuale kvadratsum* SS_{res} (Residual Sum of Squares) og *regressions-kvadratsummen* SS_{reg} (Regression Sum of Squares) skal i spil her. SS_{tot} beskriver y-koordinaternes variation i forhold til middelværdien af disse. Den er faktisk n gange den empiriske varians af y-værdierne. SS_{res} beskriver y-koordinaternes variation i forhold til regressionslinjens y-værdier, altså kvadratsummen af residualerne. Bemærk, at det er den størrelse, som er minimeret ved regressionsprocessen! Den sidste størrelse SS_{reg} handler om variationen af de y-værdier, som regressionslinjen forudsiger (altså $f(x_i)$'erne) set i forhold til middelværdien af y-værdierne. Den kaldes også for *den forklarende kvadratsum* (Explained sum of Squares), fordi den handler om den naturlige variation, som regressionslinjen kan forklare.

Hvis vi kigger på definition 7 for forklaringsgraden, så kan brøken SS_{res}/SS_{tot} fortolkes som den brøkdel af den totale variation, som *ikke* kan forklares, stammende fra residualerne. Brøken skal trækkes fra 1 for at få forklaringsgraden. Det er derfor ikke mærkeligt, at man tolker forklaringsgraden som den brøkdel af den totale variation, som *kan forklares*. Det stemmer samtidigt fint med den alternative formel (16) for forklaringsgraden.

(20) $$R^2 = 1-\frac{SS_{res}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}}$$

Hvis forklaringsgraden i et eksempel fås til $R^2=0{,}85$, så siger vi, at regressionslinjen kan forklare 85% af variationen i y-værdierne. Der er altså 15% af variationen, som regressionslinjen *ikke* kan forklares. Vi ser samtidigt, at hvis $R^2=1$, så ligger alle punkterne på regressionslinjen, mens $R^2=0$ betyder, at $SS_{res}=SS_{tot}$ og $SS_{reg}=0$. I øvrigt vil forklaringsgraden altid være et tal mellem 0 og 1. Kvadratsummerne er nemlig alle *ikke-negative*, hvorfor (14) i sætning 6 giver os, at SS_{tot} er mindst lige så stor som henholdsvis SS_{res} og SS_{reg}, hvorved $0\le R^2\le 1$ følger. Lige et forbehold: Hvis alle datapunkterne har samme y-værdier, så er forklaringsgraden slet ikke defineret!

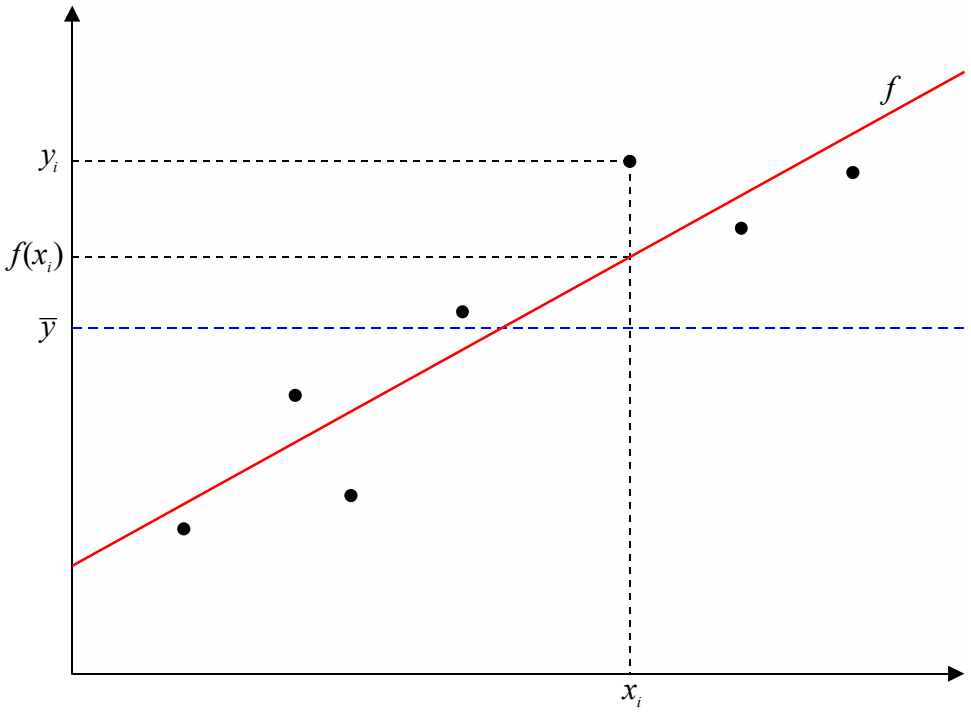
Lad os et øjeblik betragte de størrelser vi summer over i kvadratsummerne i et grafisk perspektiv: $y_i - \overline{y}$, $f(x_i) - \overline{y}$ og $y_i - f(x_i)$, hvor den sidste er det i'te residual r_i. Der gælder klart at den første størrelse er summen af de to sidste:

(21) $$y_i - \overline{y} = \left(f(x_i) - \overline{y}\right) + \left(y_i - f(x_i)\right)$$

Det mest interessante er imidlertid, at hvis vi kvadrerer hvert af de tre led og summerer fra $i = 1$ til n, så gælder lighedstegnet også. Det var det, vi viste i sætning 6:

(22) $$\sum_{n=1}^{n}\left(y_i - \overline{y}\right)^2 = \sum_{n=1}^{n}\left(f(x_i) - \overline{y}\right)^2 + \sum_{n=1}^{n}\left(y_i - f(x_i)\right)^2$$

der er det samme som: $SS_{tot} = SS_{reg} + SS_{res}$.



Korrelationskoefficienten

Definition 9 (Den empiriske korrelationskoefficient)

Den *empiriske korrelationskoefficient*, som beskriver samvariationen mellem x- og y-værdierne i sættet af datapunkter: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, er defineret ved:

(23) $$\rho = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\cdot\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Sætning 10

Der gælder følgende formel for den empiriske korrelationskoefficient:

$$\rho = a \cdot \frac{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}}{\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}} \tag{24}$$

hvor a er hældningskoefficienten for regressionslinjen.

Bevis: Vi omskriver udtrykket for definitionen af ρ fra definition 9 og udnytter udtrykket for hældningskoefficienten for regressionslinjen fra sætning 1:

$$\rho = \frac{\sum_{i=1}^{n}(x_i-\overline{x})\cdot(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})\cdot(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\cdot\frac{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}}{\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

$$= a \cdot \frac{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}}{\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

□

Sætning 11

Forklaringsgraden for datasættet $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ er kvadratet på den empiriske korrelationskoefficient:

$$R^2 = \rho^2 \tag{25}$$

Bevis: Fremgår direkte ved at sammenligne udtrykkene i sætning 8 og 10.

□

Definition 12 (Empiriske spredninger)

De *empiriske spredninger*, af x- og y-værdierne i sættet $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ af datapunkter er defineret ved:

$$s_x = \sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-\overline{x})^2} \quad \text{og} \quad s_y = \sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{n}(y_i-\overline{y})^2} \tag{26}$$

Sætning 13

For hældningskoefficienten a for linjen, som fremkommer ved lineær regression af datapunkterne $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ gælder følgende formel:

(27) $$a = \rho \cdot \frac{s_y}{s_x}$$

Bevis: Det ses umiddelbart af sætning 10 og definition 12.

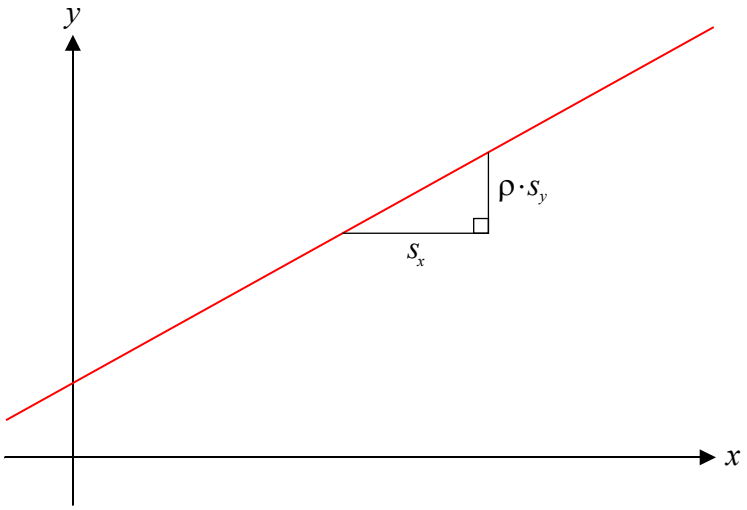
□

Bemærkning 14

Sætning 13 ses også at gælde, selv om man ikke skulle anvende spredningerne, som er korrigeret for bias, altså selv om der i definition (26) benyttes $1/n$ i stedet for $1/(n-1)$.

□

Sætning 13 kan grafisk illustreres således:



En ændring på én spredning (standardafvigelse) i x afstedkommer altså en ændring i y på korrelationskoefficienten ganget med en spredning (standardafvigelse) i y. Desuden skal det tilføjes, at den empiriske korrelationskoefficient er et tal mellem –1 og 1: $-1 \le \rho \le 1$.