Forklaring af forklaringsgraden via variansanalyse

Christian Thybo, lektor emeritus

Denne note bygger videre på Carl Winsløvs artikel *Fodgængerversion af lineær regression*, som stod at læse i LMFK–bladet nr.1/2015. Læseren forudsættes derfor at kende denne artikel og jeg vil bruge Winsløws notation. Jeg har en fornemmelse af, at forklaringsgraden naturligt hører hjemme i den statistiske disciplin, som kaldes variansanalyse (til indføring heri kan fx konsulteres Sprent: *Statistics in Action*, Penguin Books(1977), Hoel: *Introduction to Mathematical Statistics*, Wiley(1947)). Men nu til sagen:

Winsløws første formel for R^2 forlænges med $\frac{1}{n}$. Dermed er

$$R^2 = \frac{\frac{1}{n}\sum_{k=1}^{n}\left((ax_k+b)-\bar{y}\right)^2}{\frac{1}{n}\sum_{k=1}^{n}\left(y_k-\bar{y}\right)^2}$$

Her er nævneren V_y variansen af de målte værdier y_k, $k = 1, 2, 3, \dots, n$.

Men tælleren kan også opfattes som en varians, nemlig variansen på *modelværdierne* $ax_k + b$, $k = 1, 2, 3, \dots, n$. Vi har nemlig $\bar{y} = a\bar{x} + b = \overline{ax+b}$.

Lad os kalde tælleren V_{model}. Altså har vi $\frac{V_{\text{model}}}{V_y} = R^2$ og dermed

$$V_y - V_{\text{model}} = V_y - V_y R^2 = V_y\left(1-R^2\right)$$

Det vises nu, at denne differens er lig med $\frac{1}{n}S(a,b)$.

Winsløws formel for $S(\alpha,\beta)$ ganget med $\frac{1}{n}$ lyder

$$\frac{1}{n}S(\alpha,\beta) = D^2 + \frac{1}{n}A\left(\alpha-\frac{A}{B}\right)^2 + \frac{1}{n}\left(C-\frac{B^2}{A}\right)$$

og denne størrelse er mindst, når $a = \dfrac{B}{A}$ indsættes for α og $b = \bar{y} - a\bar{x}$ for β.

Herved bliver (studer Winsløv!) $D = 0$ og det andet led ligeledes lig med 0, så vi står tilbage med

$$\frac{1}{n}S(a,b) = \frac{1}{n}\left(C - \frac{B^2}{A}\right) = \frac{1}{n}C\left(1 - \frac{B^2}{AC}\right) = V_y\left(1 - R^2\right)$$

idet, stadig ifølge Winsløw,

$$R^2 = \frac{B^2}{AC} \quad \text{og} \quad \frac{1}{n}C = \frac{1}{n}\sum_{k=1}^{n} w_k^2 = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \bar{y}\right)^2 = V_y$$

Hermed har vi sammenhængen

$$V_y = \frac{1}{n}S(a,b) + V_{\text{model}} \quad \text{eller} \quad \frac{\frac{1}{n}S(a,b)}{V_y} + \frac{V_{\text{model}}}{V_y} = 1$$

Dette er den grundlæggende ligning.

Det første led betragtes som *"støj"*, det andet som den del af den samlede varians, som modellen *"forklarer"*, altså *forklaringsgraden* R^2. Det er denne ligning, som knytter regressionsanalyse sammen med variansanalyse, og det er dér, R^2 naturligt hører hjemme.

Den historiske udredning af R og af R^2 vil jeg overlade til andre. Mon ikke R. A. Fisher har en finger med i spillet mht. forklaringsgraden og F. Galton mht. korrelationskoefficienten. De to størrelser er nok opstået i forskellige begrebsmæssige landskaber, og det kunne være interessant at vide, hvor meget disse oprindeligt har at gøre med den bedste rette linje.

Lektor Flemming Pedersen har omsat manuskriptet til den ønskede elektroniske form. Jeg er ham stor tak skyldig. Alle fejl og mangler er naturligvis mit ansvar.