Fodgængerversion af lineær regression

Carl Winsløw, IND, Københavns Universitet

I gymnasiet er lineær regression som bekendt blevet standard allerede fra C–niveauet, utvivlsomt fordi det i mange sammenhænge er nyttigt og bekvemt at udføre beregningerne på CAS ud fra et datasæt. Nogle lærebøger giver alligevel formlen for den bedste rette linie $y = ax + b$ (eller rettere, formlerne for a og b, givet et datasæt bestående af punkter eller talpar, som linien skal tilnærme). Disse formler kan give et indtryk af at CAS–værktøjet ikke udfører trylleri, men "bare sætter ind" i nogle relativt banale formler. Enkelte lærebøger giver også en slags bevis for formlerne, der dog ikke er helt tilfredsstillende, fordi man naturligvis mangler de nødvendige forudsætninger vedr. funktioner af to variable, herunder arten af kritiske punkter.

Meningen med denne note er at give en matematisk begrundelse for formlerne på basis af elementær bogstavregning. Noten er skrevet for lærere og skal utvivlsomt uddybes, hvis man vil dele den med elever; men principielt er den på C–niveau. Ideen er ikke ny (se fx Key, 2005), men så vidt jeg har kunnet konstatere, er den heller ikke alment kendt.

Regression for fodgængere

Matematikopgaven er flg.: For et datasæt $(x_1, y_1), \ldots, (x_n, y_n)$ skal vi bestemme α og β så kvadratsummen

$$S(\alpha,\beta) = \sum_{k=1}^{n} (y_k - \alpha x_k - \beta)^2$$

bliver mindst mulig. For kortheds skyld indfører vi flg. notationer:

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k; \quad \bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k;$$

$$z_k = x_k - \bar{x} \quad k = 1, \ldots, n; \quad w_k = y_k - \bar{y} \quad k = 1, \ldots, n;$$

$$A = \sum_{k=1}^{n} z_k^2; \quad B = \sum_{k=1}^{n} z_k w_k; \quad C = \sum_{k=1}^{n} w_k^2; \quad D = \bar{y} - \beta - \alpha\bar{x}$$

Bemærk, at $\sum_{k=1}^{n} z_k = \sum_{k=1}^{n} w_k = 0$, og at vi (idet vi antager at ikke alle x_k er ens) har $A \neq 0$.

Vi finder nu:

$$\begin{aligned} S(\alpha,\beta) &= \sum_{k=1}^{n} (\bar{y} + w_k - \alpha(\bar{x} + z_k) - \beta)^2 \\ &= \sum_{k=1}^{n} (D + w_k - \alpha z_k)^2 \\ &= nD^2 + \sum_{k=1}^{n} (w_k - \alpha z_k)^2 + 2D\left(\sum_{k=1}^{n} w_k - \alpha\sum_{k=1}^{n} z_k\right) \\ &= nD^2 + \alpha^2 A + C - 2\alpha B \\ &= nD^2 + A\left(\alpha^2 - 2\frac{\alpha B}{A} + \frac{B^2}{A^2} - \frac{B^2}{A^2}\right) + C \\ &= nD^2 + A\left(\alpha - \frac{B}{A}\right)^2 + \left(C - \frac{B^2}{A}\right). \end{aligned}$$

Vi bemærker, at det sidste led i summen ikke afhænger af α og β. De to første led er positive, og de afhænger af α og β. Vi kan derfor minimere kvadratsummen ved at bestemme α og β, så de to første led begge bliver 0. Disse værdier kalder vi a og b, og de er:

$$a = \frac{B}{A} \quad \text{dvs.} \quad a = \frac{\sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^{n} (x_k - \bar{x})^2}, \text{ og}$$

$$0 = D = \bar{y} - b - a\bar{x} \quad \text{dvs.} \quad b = \bar{y} - a\bar{x}.$$

Forklaring af forklaringsgraden

De fleste gængse matematikværktøjer giver sammen med regressionslinien også dennes "forklaringsgrad". Betydningen af dette tal er næppe helt klart for eleverne, der ofte tilegner sig et forenklet syn på, hvad man kan slutte ud fra tallets værdi. Meningen med dette afsnit er at vise, at udledningen ovenfor også kan bruges til at kaste lys over forklaringsgraden.

Vi antager nu også, at ikke alle y_k er ens, så $C \neq 0$. Vi definerer $R = B / \sqrt{AC}$ som kaldes *korrelationskoefficienten* for datasættet. Den celebre *forklaringsgrad* er så $R^2 = B^2 / AC$. Det forklarer ikke navnet! Men vi har

$$R^2 = \frac{B^2 A}{A^2 C} = \frac{a^2 A}{C} = \frac{\sum_{k=1}^{n} (ax_k - a\bar{x})^2}{C} = \frac{\sum_{k=1}^{n} (ax_k + b - \bar{y})^2}{\sum_{k=1}^{n} (y_k - \bar{y})^2}.$$

Det sidste udtryk er ophav til navnet. Tælleren måler hvor langt "modelværdierne" af y ligger fra gennemsnittet af de observerede værdier; nævneren, hvor langt de observerede værdier af y ligger fra samme gennemsnit. Man kan så sige, at R^2 er et mål for, hvor stor en del af den observerede afvigelse fra gennemsnittet, som modellen $y = ax + b$ "forklarer". Det er oplagt at R^2 er et ikke–negativt tal, som er 1, hvis alle datapunkter ligger nøjagtigt på linien. Desuden er

$$R^2 = \frac{B^2 / A}{C} = \frac{B^2 / A}{B^2 / A + S(a,b)} \leq 1.$$

Altså $R^2 \in [0,1]$ og dermed $R \in [-1,1]$.

Reference

Key, E. (2005). *A painless approach to least squares*. The College Mathematical Journal 36 (1), 65–67.