Linjär regression

Vi har en verklighet som vi vill försöka beskriva (eller approximera) med hjälp av matematiska modeller.

T. ex. en individs konsumtion styrs av hennes inkomst, familjeförhållanden, besparingar osv. Det kan beskrivas med hjälp av följande matematiska modell

$$kons = b_1 ink + b_2 fam + b_3 besp + \varepsilon,$$

där $b_1 ink + b_2 fam + b_3 besp$ är den förklarande (deterministiska) delen av modellen och ε den icke förklarande (slumpmässiga) delen av modellen, som ibland kallas den naturliga variationen (av konsumtion).

Enkel linjär regression

Inom regressionsanalysen har man ett antal oberoende (förklarande) variabler som styr (eller förklarar) värdet på en beroende variabel. Ofta betecknas

- oberoende var X och
- beroende (av värdet på X) var Y

Två begrepp

Kausalitet: Ett faktiskt beroende mellan variabler i verkligheten, och riktning på beroendet.

Ex. Dotterns kroppslängd beror (till viss del) på moderns kroppslängd, ej tvärtom!

Korrelation: Linjärt numeriskt samband, ej nödvändigtvis förankrat i verkligheten. Korrelationskoefficienten ρ mäter detta samband.

Ex. Antalet poäng i statistik (för män) är högt korrelerad med skägglängd.

Vi har observationspar (x_i, y_i). Utifrån dessa kan vi anpassa en rät linje.

- Räta linjens ekvation

$$y = a + bx$$

Exempel AJÅ kap 2.1:

En oljedistributör vill prognostisera villaägares oljeförbrukning. Tio kedjehusägare avläser oljeförbrukning under varsin månad.

Hur skall vi "lägga linjen", dvs hur skall vi skatta interceptet a och lutningskoefficienten b. Eftersom a och b är skattningar från stickprovet skrivs räta linjens ekvation i det här sammanhanget som

$$\hat{y} = a + bx$$

Minsta Kvadrat Metoden

Den innebär att man "lägger linjen" så att kvadratsumman

$$Q = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - bx_i)^2$$

minimeras. Genom partiell derivering av Q på a och b, respektive, får man följande formler för regressionskoefficienterna

$$\begin{aligned} a &= \bar{y} - b\bar{x} \\ b &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \end{aligned}$$

Exemplet i AJÅ:

För att kunna utföra beräkningar behövs följande summor

$$\begin{array}{lll} n = 10 & \sum x = 63.2 & \sum x^2 = 874.24 \\ \sum y = 3320 & \sum y^2 = 1412000 & \sum xy = 8990.5 \end{array}$$

De skattade regressionskoefficienterna blir

$$\begin{aligned} b &= \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{8990.5 - \frac{63.2 \cdot 3320}{10}}{874.24 - \frac{63.2^2}{10}} \\ &= -25.2559 \\ a &= \bar{y} - b\bar{x} = \frac{3320}{10} - (-25.3)\frac{63.2}{10} = 492. \end{aligned}$$

Regressionslinjen har alltså ekvationen

$$\hat{y} = 492 - 25.3x.$$

Tolkning av koefficienterna:

- $b = -25.3$: När utomhustemperaturen ökar en grad C minskar oljeförbrukningen i *genomsnitt* med 25.3 liter. OBS!! Tolkningen gäller i intervallet som vi har observerat data, dvs för temperaturer mellan -5 och 15 grader C. Om man vill tolka oljeförbrukning för andra temperaturer utanför observerade intervallet, måste man vara försiktig och fråga sig om tolkningen blir relevant.

- $a = 492$: Den *genomsnittliga* oljeförbrukningen då utomhustemperaturen är noll grader C. Notera att ofta är tolkningen av interceptet a meningslös.

ANOVA (Analysis of Variance)

Avvikelsen från en observation y_i till medelvärdet $\bar{y}$ kan delas in i avvikelsen från en observation y_i till den anpassade regressionslinjen $\hat{y}_i$, och avvikelsen från den anpassade regressionslinjen $\hat{y}_i$ till medelvärdet $\bar{y}$, dvs

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}).$$

Om man kvadrerar och summerar vänster och höger led var för sig och därefter förenklar får vi följande samband

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2.$$

- $\sum (y_i - \bar{y})^2$ är den totala variationen y har kring $\bar{y}$ och brukar betecknas SST.

- $\sum (\hat{y}_i - \bar{y})^2$ är den variation i y som förklaras av det linjära sambandet med x (utomhustemperaturen) och brukar betecknas SSR.

- $\sum (y_i - \hat{y}_i)^2$ är den variation y har kring $\hat{y}$ (regressionslinjen). Alltså, den variation som återstår och ej förklaras av det linjära sambandet mellan y och x. Den brukar betecknas SSE.

Dessa variationsorsaker kan redovisas i en $ANOVA-$ tablå:

Variati.-orsak	Kvadrat-summa (SS)	Frihets-gr. (fg)	Medelkvadr.-summa (MS)
Regr (R)	$SSR = \sum (\hat{y}_i - \bar{y})^2$	1	$MSR = SSR/$
Res (E)	$SSE = \sum (y_i - \hat{y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Tot (T)	$SST = \sum (y_i - \bar{y})^2$	$n-1$	

Medelkvadratsumman

$$MSE = \sum (y_i - \hat{y}_i)^2 = \frac{SSE}{n-2}$$

är *residualvariansen*, som också betecknas s_e^2. *Residualspridningen* ges av

$$s_e = sqrt\left(s_e^2\right).$$

För oljeförbrukningsdatat blir

$$\begin{aligned} SSR &= \sum (\hat{y}_i - \bar{y})^2 = b \sum (x_i - \bar{x})(y_i - \bar{y}) \\ &= b\left[\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}\right] \\ &= -25.2559\left[8990.5 - \frac{63.2 \cdot 3320}{10}\right] \\ &= 302866 \\ SST &= \sum (y_i - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n} \\ &= 1412000 - \frac{3320^2}{10} = 309760 \\ SSE &= SST - SSR = 309760 - 302866 = 6894 \end{aligned}$$

$ANOVA-$ tablån blir då:

Variations-orsak	SS	f.g.	MS
Regr (R)	302866	1	302866
Res (E)	6894	$n-2$	862
Tot (T)	309760	$n-1$	

Residualvariansen är alltså

$$s_e^2 = MSE = \frac{SSE}{n-2} = 862,$$

och residualspridningen

$$s_e = sqrt\left(s_e^2\right) = 29.4.$$

Determinationskoefficienten (förklaringsgraden)

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

mäter hur stor del av den totala variationen för den beroende variabeln som förklaras av det linjära sambandet (av de oberoende förklarande variablerna).

I oljeexemplet blir förklaringsgraden

$$R^2 = \frac{SSR}{SST} = \frac{302866}{309760} = 0.98$$