Multippel lineær regresjon, 12.7 + 12.8

Flere forklaringsvariabler per enhet (individ)

Responsen Y_i

Forklaringsvariabel 1 x_{i1} $\quad i = 1, \cdots, n$

— " — 2 x_{i2}

⋮

— " — k x_{ik}

Modell: $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$

der $\varepsilon_i \sim N(0, \sigma^2)$ uavh. ($E\varepsilon_i = 0$, $Var(\varepsilon_i) = \sigma^2$)

Ex) Fødselsvekten, Y_i = fødselsvekt

x_{i1} = Svangerskapslengde

x_{i2} = Mors alder

x_{i3} = Antall tidligere fødsler

x_{i4} = Barnets kjønn $= \begin{cases} 1 & \text{Gutt} \\ 0 & \text{Jente} \end{cases}$

Ex) Kaffesalg, Y_i = salg

x_{i1} = Ant. automater $= x_i$

$x_{i2} = x_{i1}^2$

$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$ Polynomiell regresjon

okt 30-12:15

Formål
- Multifaktorielle fenomener
- Mer presise prediksjoner
- Lettere å påvise signifikans.
- Kan løse opp i konfundering hvis lurkende variablen er kjent.

Fortolkning av β_j

Endring i forventning $\mu_i = EY_i$ hvis x_{ij} endres til $x_{ij}+1$, når samtidig de øvrige forklaringsvariablene holdes konstant.

Dette kan ikke alltid holde, f.eks. ved polynomiell regresjon.

okt 30-12:27

Estimering: MKE = minste kvadraters estimatoren.

$f(b_0, b_1, \dots, b_k) = \sum (Y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik})^2$

MKE $\hat{\beta}^T = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)$ er de verdien av $(b_0, b_1, \dots, b_k)$ som gjør $f(b_0, b_1, \dots, b_k)$ minst mulig.

Normalligningene

$$\frac{\partial f}{\partial b_0} = -2 \sum (Y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik}) = 0$$

$$\frac{\partial f}{\partial b_1} = -2 \sum (Y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik}) x_{i1} = 0$$

⋮

$$\frac{\partial f}{\partial b_k} = -2 \sum (Y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik}) x_{ik} = 0$$

Entydig løsning hvis $k+1 \le n$ og lign. lineært uavh.

okt 30-12:36

Ligningene er enklere å uttrykke med vektorer og matriser:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \text{og} \quad \mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = E\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{bmatrix} EY_1 \\ \vdots \\ EY_n \end{bmatrix} = \begin{pmatrix} \beta^T x_1 \\ \vdots \\ \beta^T x_n \end{pmatrix}$$

der $x_i^T = (1, x_{i1}, x_{i2}, \cdots, x_{ik})$ og $\beta^T = (\beta_0, \beta_1, \cdots, \beta_k)$

og $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$

og designmatrise

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & & & \\ \vdots & & & \\ 1 & x_{n1} & & x_{kn} \end{pmatrix} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}$$

Ser da at $\mu = X\beta$

Vi kan da skrive modellen

$Y = \mu + \varepsilon = X\beta + \varepsilon$

okt 30-12:55

På matriseform blir normallign.

$-2 X^T [Y - \mu] = 0$ (Sjekk dette)

og siden $\mu = X\beta$ kan omforme dette til

$X^T Y = X^T \mu = X^T X \beta$

Løsning (MKE) for β er altså

$\hat{\beta} = (X^T X)^{-1} X^T Y$

såfremt $X^T X$ er invertebar $\Leftrightarrow$ X har full rang

Videre er, med $\hat{\mu} = X\hat{\beta}$,

$$f(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k) = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik})^2 = (Y - \hat{\mu})^T (Y - \hat{\mu}) = \|Y - \hat{\mu}\|^2 = SSE$$

Forv. rett estimator for σ^2: $s^2 = \frac{SSE}{n-k-1}$

Bruker $k+1$ frihetsgrader til å estimere β.

okt 30-13:16

La videre $\bar{Y} = (\bar{Y}, \bar{Y}, \cdots, \bar{Y})$ (n elementer)

Da er

$SST = \sum (Y_i - \bar{Y})^2 = (Y - \bar{Y})^T (Y - \bar{Y}) = \|Y - \bar{Y}\|^2$

og vi kan uttrykke

SSR = kvadratsummen svarende til regresjonen $= SST - SSE = \cdots = \|\bar{Y} - \hat{Y}\|^2$ ← Oppgave

Ut fra dette defineres

R^2 = Forklart andel av variasjon $= 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$

Fortolkning av multippel R^2

= (korrelasjon med Y_i og $\hat{Y}_i$)^2

okt 30-13:26

Forventningsrett het av $\hat{\beta}$. Lin. transf. Extra 7.4 og 7.5

$$E[\hat{\beta}] = E[(X^T X)^{-1} X^T Y] = (X^T X)^{-1} X^T E Y$$

$$= (X^T X)^{-1} X^T \mu = (X^T X)^{-1} X^T X \beta = \beta$$

På tilsv. måte finnes varians (kovarians for $\hat{\beta}$.

Merk også at $\hat{\beta}_j$ er lineær komb av Y_i og hvis disse er normale blir også $\hat{\beta}_j$ normalfordelt.

La se_j = standard feil for $\hat{\beta}_j$. Da blir

$$t_j = \frac{\hat{\beta}_j - \beta_j}{se_j} \sim t_{n-k-1}$$

og vi kan teste $H_0: \beta_j = 0$ $H_a: \beta_j \neq 0$ ved $|\hat{\beta}_j / se_j|$. Forkaster hvis denne er større enn

$$t_{\alpha/2} = (1-\alpha/2)100 \text{ persentil i } t_{n-k-1}$$

Et $(1-\alpha)100\%$ KI for β_j gis ved

$$\hat{\beta}_j \pm t_{\alpha/2}\, se_j$$

Dessuten $F = \frac{SSR/k}{SSE/(n-k-1)} \sim F_{k,n-k-1}$

under $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$

Alt. H_a: Minst en av $\beta_j \neq 0$

Justert R^2: $R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$

okt 30-13:33