Multippel lineær regresjon

$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$ uavh.

MKE $\hat{\beta} = (X^T X)^{-1} X^T Y \sim N(\beta, (X^T X)^{-1}\sigma^2)$

hvis modellen holder.

Modellsjekk

Predikerte verdier $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}$

Residualer $e_i = Y_i - \hat{Y}_i$

Stand. residualer $e_i^* = \frac{e_i}{s\sqrt{1-h_{ii}}}$ der $(h_{11}, \ldots, h_{nn}) = \mathrm{diag}(X(X^TX)^{-1}X^T) = \mathrm{diag}(H)$

1. Linearitet — plott $(\hat{Y}_i, e_i)$ evt. (x_{ij}, e_i)
2. Konstant varians — plott $(\hat{Y}_i, e_i)$ evt. $(\hat{Y}_i, \sqrt{|e_i^*|})$
3. Uavhengighet — plott (e_{i-1}, e_i) evt. (i, e_i)
4. Normalfordelte Y_i og ε_i — Q-Q-plott, Histogram, boksplott
5. Ekstreme og innflytelsesrike observasjoner. — plott (h_{ii}, e_i^*)

nov 14-14:13

Typer forklaringsvariable

A. Polynomiell regresjon

$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \varepsilon_i$

Sjelden både $x_i^4, x_i^5, \ldots, x_i^k$ gir vesentlig mer informasjon.

$x_i, x_i^3, x_i^5 \ldots$ er sterkt korrelert

$x_i^2, x_i^4, x_i^6, \ldots$ ″

Prøv med $\sqrt{x_i} = x_i^{1/2}$, $\frac{1}{x_i} = x_i^{-1}$, $\ln(x_i)$

B. Kategoriske forkl. var. – Dummy, indikatorvar.

$x_{i1} = \begin{cases} 1 & \text{Jente} \\ 0 & \text{Gutt} \end{cases}$

$E[Y_i \mid x_{i1}] = \beta_0 + \beta_1 x_{i1} = \begin{cases} \beta_0 + \beta_1 = \mu_1 & \text{Jente} \\ \beta_0 = \mu_0 & \text{Gutt} \end{cases}$ Toutvalg.

$\Leftrightarrow$ Toutvalg t-test

Kan ha kategoriske forkl. var. med K nivåer

Introduserer K-1 indikator var. på forkl. var

$x_{i1} = \begin{cases} 1 & \text{Nivå 1} \\ 0 & \text{Ellers} \end{cases}$

$x_{i2} = \begin{cases} 1 & \text{Nivå 2} \\ 0 & \text{Ellers} \end{cases}$

$\vdots$

$x_{ik-1} = \begin{cases} 1 & \text{Nivå k-1} \\ 0 & \text{ellers} \end{cases}$

$E[Y_i \mid x_{i1}, \ldots, x_{ik-1}] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{k-1} x_{ik-1}$

$= \begin{cases} \beta_0 + \beta_1 = \mu_1 & \text{nivå 1} \\ \beta_0 + \beta_2 = \mu_2 & \text{″ 2} \\ \vdots & \\ \beta_0 + \beta_{k-1} = \mu_{k-1} & \text{″ k-1} \\ \beta_0 = \mu_K & \text{″ K} \end{cases}$

Har da at nivå K er et referansenivå.

MKE for μ_j blir $\bar{Y}_j$ = gj.sn. Y_i med nivå j.

$\Rightarrow$ MKE for β_j-ene.

Dette blir det samme som 1-veis variansanalyse (Kap. 11.1)

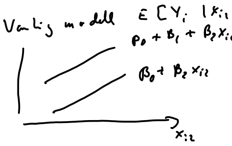
Ex) Yrke Ex) Farve på blomst.

nov 14-14:51

C. Interaksjon

La $x_{i1} = \begin{cases} 1 & \text{Gr. 1} \\ 0 & \text{Gr. 2} \end{cases}$, x_{i2} = skalar forkl. var.

Vanlig modell $E[Y_i \mid x_{i1}, x_{i2}] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$

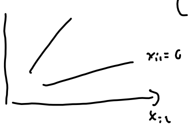


Parallelle linjer.

Kan mer generelt ha at

$E[Y_i \mid x_{i1}, x_{i2}] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}$

$= \begin{cases} \beta_0 + \beta_2 x_{i2} & \text{hvis } x_{i1} = 0 \\ (\beta_0 + \beta_1) + (\beta_2 + \beta_3) x_{i2} & \text{hvis } x_{i1} = 1 \end{cases}$



Ikke parallelle linjer.

nov 14-15:22

Hva skjer hvis vi utelater en forkl. var?

Anta at $E[Y_i \mid x_{i1}, x_{i2}] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$

samt at (x_{i1}, x_{i2}) er bivariat normalfordelte,

med forv. μ_1 og μ_2, std.avvik σ_1 og σ_2 og korrelasjon ρ

Da har vi, ved lov om dobbeltforventning, $EY = E\,E(Y|X)$,

$E[Y_i \mid x_{i1}] = E\{E[Y_i \mid x_{i1}, x_{i2}] \mid x_{i1}\}$

$= E[\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \mid x_{i1}]$

$= \beta_0 + \beta_1 x_{i1} + \beta_2 E[x_{i2} \mid x_{i1}]$

Bivariat normalf. $\rightarrow$ $= \beta_0 + \beta_1 x_{i1} + \beta_2 [\mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x_{i1} - \mu_1)]$

$= \gamma_0 + \gamma_1 x_{i1}$ der $\boxed{\gamma_1 = \beta_1 + \beta_2 \rho \frac{\sigma_2}{\sigma_1}}$ (*)

Faktisk i en lineær regn. mod

$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$

får vi ved å analysere

$Y_i = \gamma_0 + \gamma_1 x_{i1} + \varepsilon_i'$

at MKE for γ_1 blir $\hat{\gamma}_1 = \hat{\beta}_1 + r_{12} \frac{s_2}{s_1} \hat{\beta}_2$

der $\hat{\beta}_1, \hat{\beta}_2$ = MKE fra full modell (*), s_j = std.avvik for x_{ij}

og r_{12} = korrelasjon mellom x_{i1} og x_{i2}.

nov 14-15:48