

Lærebogens formel (4.9)

Anvendes især i kapitlerne 5-6.

Formel (4.9) i lærebogen siger

$$(4.9) \quad h_{ii} = \frac{1}{n} + (\mathbf{x}_i - \bar{\mathbf{x}})(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})^T.$$

Under formelen i lærebogen angives at vektorerne og matricerne er af dimension p , dvs. at også første pladsen, der repræsenterer interceptet medtages. Imidlertid er første element i vektoren $\mathbf{x}_i - \bar{\mathbf{x}}$ lig med nul, da første element i $\mathbf{x}_i$ et 1-tal. Desuden er første element i $\bar{\mathbf{x}}$ som et gennemsnit af n ettaller også et 1-tal. I matrixmultiplikationen "skrælles" første søjle og første række derfor af $\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}^{-1}$. Første søjle i matricen $\tilde{\mathbf{X}}$ består af ettaller, mens de øvrige elementer er af formen x_{ij} fratrasket gennemsnittet af de n observationer af den j -te forklarende variabel. Pånær første søjle er søjlesummerne i matricen $\tilde{\mathbf{X}}$ derfor lig med nul.

Alt i alt gælder, at $p \times p$ matricen $\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$ er lig med blok matricen

$$\begin{pmatrix} n & 0 \\ 0 & \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \end{pmatrix},$$

hvor matricen $\tilde{\mathbf{X}}$ betegner det samme som ovenfor dog uden den første søjle af 1-taller. Matricen $\tilde{\mathbf{X}}$ betegner altså nu en $(p-1) \times (p-1)$ matrix. Vælges at se bort fra det første nul i vektoren $\mathbf{x}_i - \bar{\mathbf{x}}$, kan matrixmultiplikationen opfattes om en multiplikation af vektorer og matricer af dimension $p-1$, således at det udelukkende er en relation mellem de egentlige forklarende variable. I det følgende fastholdes dimensionen af matricen $\tilde{\mathbf{X}}$ og vektoren $\mathbf{x}_i - \bar{\mathbf{x}}$ til $p-1$.

Da $\tilde{\mathbf{X}}$ er matricen bestående af de forklarende variable fratrasket deres gennemsnit, er $\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$ (pånær at der ikke er divideret med n eller $n-1$) lig med den empiriske kovariansmatrix for de forklarende variable. Diagonalelementerne er altså "de empiriske varianser", og uden for diagonalen står "kovarianserne" mellem de forklarende variable ganget med faktoren n eller $n-1$.

Formlen (4.9) for h_{ii} udtrykker derfor, at h_{ii} er afstanden mellem den i 'te observation af de forklarende variable og gennemsnittet af de forklarende variable, idet afstanden er målt med den

empiriske varians fuldstændigt som afstand i et Wald test. Men da de forklarende variable på ingen måde antages at være normalfordelte, eller blot stokastiske variable, kan der ikke af denne analogi udledes noget fordelingsresultat. Størrelserne h_{ii} kan altså ikke danne grundlag for et χ^2 test.

Matricen $\mathbf{X}$ har formen

$$\mathbf{X} = \begin{pmatrix} 1 \\ \cdot \\ \cdot & \mathbf{X}_{(1)} \\ \cdot \\ 1 \end{pmatrix} = (\mathbf{1} | \mathbf{X}_{(1)}) ,$$

hvor $\mathbf{X}_{(1)}$ altså blot er de sidste $p - 1$ søjler i $\mathbf{X}$. Der gælder

$$\mathbf{X}^T \mathbf{X} = \begin{pmatrix} \mathbf{n} & \mathbf{1}^T \mathbf{X}_{(1)} \\ \mathbf{X}_{(1)}^T \mathbf{1} & \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \end{pmatrix} .$$

Denne blokmatrix inverteres nu ved hjælp af formel (A5.1) side 304 i lærebogens appendiks 5. Med notationen fra lærebogens side 304 gælder

$$\mathbf{E} = \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} - \mathbf{X}_{(1)}^T \mathbf{1} \mathbf{1}^T \mathbf{X}_{(1)} / \mathbf{n} .$$

Desuden gælder om $\mathbf{F}$

$$\mathbf{F} = \mathbf{1}^T \mathbf{X}_{(1)} / \mathbf{n} = \bar{\mathbf{x}}^T ,$$

idet

$$(1) \quad \mathbf{X}_{(1)}^T \mathbf{1} = \begin{pmatrix} \sum_{i=1}^n x_{i2} \\ \cdot \\ \cdot \\ \sum_{i=1}^n x_{ip} \end{pmatrix} = n \begin{pmatrix} \bar{x}_2 \\ \cdot \\ \cdot \\ \bar{x}_p \end{pmatrix} = n \bar{\mathbf{x}},$$

Det k'te element i $\mathbf{F}$ er derfor givet ved gennemsnittet af de k'te forklarende variable

$$F_k = \bar{x}_k.$$

Det (k,m)'te element i $\mathbf{E} = \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} - n \bar{\mathbf{x}} \bar{\mathbf{x}}^T$ er

$$(2) \quad E_{km} = \sum_{i=1}^n x_{ik} x_{im} - n \bar{x}_k \bar{x}_m = \sum_{i=1}^n (x_{ik} - \bar{x}_k)(x_{im} - \bar{x}_m),$$

og dermed lig med det (k,m)'te element i $\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$, hvor $\tilde{\mathbf{X}}$ er den p - 1 dimensionale centrerede designmatrix af de egentlige forklarende variable. Altså er

$$\mathbf{E} = \tilde{\mathbf{X}}^T \tilde{\mathbf{X}}.$$

Bemærk at der i disse udtryk for elementerne i hhv. $(p - 1) \times (p - 1)$ matricen $\mathbf{E}$ og $(p - 1)$ søjlen $\mathbf{F}$ er indiceret med $k, m = 2, \dots, p$. Blokinddelingen af $(\mathbf{X}^T \mathbf{X})^{-1}$ er nu ifølge (A5.1)

$$(\mathbf{X}^T \mathbf{X}^{-1}) = \begin{pmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{12}^T & \mathbf{G}_{22} \end{pmatrix},$$

hvor

$$\mathbf{G}_{11} = \frac{1}{n} + \bar{\mathbf{x}}(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \bar{\mathbf{x}}^T ,$$

$$\mathbf{G}_{12} = \mathbf{G}_{21}^T = - \bar{\mathbf{x}}(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} ,$$

og

$$\mathbf{G}_{22} = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} .$$

I disse formler er $\bar{\mathbf{x}}$, jf. (1) ovenfor, den $p - 1$ dimensionale søjlevektor, der består af gennemsnittene for de $p - 1$ egentlige forklarende variable. Matricen $(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1}$ er, jf. (2) ovenfor, den inverse af den empiriske kovariansmatrix for de $p - 1$ egentlige forklarende variable på nær en faktor n . Dermed er det i 'te diagonalelement i hatmatricen $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ givet ved

$$h_{ii} = \left(1 \mid \mathbf{x}_{i(1)} \right) \begin{pmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{12}^T & \mathbf{G}_{22} \end{pmatrix} \begin{pmatrix} 1 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{x}_{i(1)}^T \end{pmatrix}$$

hvor $\mathbf{x}_{i(1)}$ betegner i 'te observation af de forklarende variable uden det første ettal, som svarer til interceptet. Da gælder

$$\begin{aligned} h_{ii} &= \mathbf{G}_{11} + \mathbf{x}_{i(1)} \mathbf{G}_{12}^T + \mathbf{G}_{12} \mathbf{x}_{i(1)}^T + \mathbf{x}_{i(1)} \mathbf{G}_{22} \mathbf{x}_{i(1)}^T \\ &= \frac{1}{n} + \bar{\mathbf{x}}(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \bar{\mathbf{x}}^T - \mathbf{x}_{i(1)}(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \bar{\mathbf{x}} - \bar{\mathbf{x}}(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \mathbf{x}_{i(1)}^T + \mathbf{x}_{i(1)}(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \mathbf{x}_{i(1)}^T . \end{aligned}$$

Dette giver følgende udtryk for h_{ii} :

$$h_{ii} = \frac{1}{n} + (\mathbf{x}_{i(1)} - \bar{\mathbf{x}})(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1}(\mathbf{x}_{i(1)} - \bar{\mathbf{x}})^T ,$$

hvilket ses ved at gange højresiden i (4.9) ud. Dette er et udtryk i $p - 1$ dimensionale vektorer og matricer. Udvides med nuller til p dimensionale vektorer og matricer ses formel (4.9) som

beskrevet i indledningen

Potentialet h_{ii} er altså på nær det første led $1/n$ proportional med afstanden mellem den i 'te observationsvektor af egentlige forklarende variable og gennemsnittet af alle de n observationer af denne vektor. Afstanden er målt med den empiriske kovariansmatrix.