8307 NL

MATHEMATISCH CENTRUM,
2e Boerhaavestraat 49,
Amsterdam-O,
Statistische Afdeling,
Rapport S 168[B] (M 71),

Kleinste kwadraten-schattingen voor de regressie-coëfficiënten van een meervoudige lineaire regressie-vergelijking[1)]

1. Inleiding

We onderstellen dat n waarnemingen $y_1,\ldots,y_n$ [2)] verricht worden, welke aan de volgende betrekkingen voldoen:

$$
\begin{aligned}
\underline{y}_1 &= \alpha + \beta_1 x_{11} + \ldots + \beta_k x_{k1} + \underline{v}_1 ,\\
\underline{y}_2 &= \alpha + \beta_1 x_{12} + \ldots + \beta_k x_{k2} + \underline{v}_2 ,\\
&\vdots\\
\underline{y}_n &= \alpha + \beta_1 x_{1n} + \ldots + \beta_k x_{kn} + \underline{v}_n ,
\end{aligned}
\tag{1,1}
$$

waarin $\alpha, \beta_1,\ldots,\beta_k$ onbekende regressie-coëfficiënten zijn, x_{ij} $(i=1,\ldots,k\ ;\ j=1,\ldots,n)$ bekende waarden van de onafhankelijke variabelen en $\underline{v}_j$ $(j=1,\ldots,n)$ niet waarneembare onafhankelijke stochastische grootheden met verwachting $\mathcal{E}\,\underline{v}_j=0$ en gelijke (onbekende) spreidingen.

De kleinste kwadratenschattingen voor de coëfficiënten $\alpha, \beta_1, \ldots, \beta_k$ zijn nu die waarden $a, b_1,\ldots,b_k$ welke voor $\alpha, \beta_1,\ldots,\beta_k$ gesubstitueerd de volgende kwadratische vorm minimaliseren:

$$
Q = \sum_{j=1}^{n} (y_j - \alpha - \beta_1 x_{1j} - \ldots - \beta_k x_{kj})^2 . \tag{1,2}
$$

Zij voldoen dus aan de relaties:

$$
\left\{\frac{\partial Q}{\partial \alpha}\right\}_{\substack{\alpha=a\\ \beta=b}} = 0 \tag{1,3}
$$

en

$$
\left\{\frac{\partial Q}{\partial \beta_i}\right\}_{\substack{\alpha=a\\ \beta=b}} = 0 \quad ;\ i=1,\ldots,k , \tag{1,4}
$$

---

1) Dit memorandum is slechts bedoeld ter oriëntatie en streeft niet naar volledigheid of volledige exactheid.
2) Stochastische grootheden geven we met onderstreepte letters aan; denken we speciaal aan de door deze grootheden aangenomen waarden dan geven we deze met dezelfde letters, maar zonder onderstreping, aan.

waarbij $\beta=b$ betekent $\beta_1=b_1$, $\beta_2=b_2$, ..., $\beta_k=b_k$.

We zullen de uitdrukking voor a en $b_1,\ldots,b_k$ eerst afleiden voor het geval k=2 en daarna met behulp van matrixnotatie de algemene formules geven.

2. Kleinste kwadratenschattingen voor het geval k=2

De relaties (1,3) en (1,4) luiden nu

$$(2,1) \qquad \left\{\frac{\partial Q}{\partial \alpha}\right\}_{\substack{\alpha=a\\ \beta=b}} = -2 \sum_{j=1}^{n} (y_i-a-b_1x_{1j}-b_2x_{2j}) = 0$$

en

$$(2,2) \quad \begin{cases} \left\{\frac{\partial Q}{\partial \beta_1}\right\}_{\substack{\alpha=a\\ \beta=b}} = -2 \sum_{j=1}^{n} x_{1j}(y_j-a-b_1x_{1j}-b_2x_{2j}) = 0 , \\ \left\{\frac{\partial Q}{\partial \beta_2}\right\}_{\substack{\alpha=a\\ \beta=b}} = -2 \sum_{j=1}^{n} x_{2j}(y_j-a-b_1x_{1j}-b_2x_{2j}) = 0 . \end{cases}$$

Uit (2,1) volgt:

$$a = \sum_{j=1}^{n} y_j/n - b_1 \sum_{j=1}^{n} x_{1j}/n - b_2 \sum_{j=1}^{n} x_{2j}/n =$$

$$= \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 .$$

Substitutie in (2,2) geeft:

$$\begin{cases} \sum_{j=1}^{n} x_{1j}\left[y_j-\bar{y}-b_1(x_{1j}-\bar{x}_1) -b_2(x_{2j}-\bar{x}_2)\right]= 0 , \\ \sum_{j=1}^{n} x_{2j}\left[y_j-\bar{y}-b_1(x_{1j}-\bar{x}_1) -b_2(x_{2j}-\bar{x}_2)\right]= 0 , \end{cases}$$

of:

$$\begin{cases} b_1 \sum_{j=1}^{n}(x_{1j}-\bar{x}_1)^2+b_2 \sum_{j=1}^{n}(x_{1j}-\bar{x}_1)(x_{2j}-\bar{x}_2) = \sum_{j=1}^{n}(x_{1j}-\bar{x}_1)(y_j-\bar{y}), \\ b_1 \sum_{j=1}^{n}(x_{1j}-\bar{x}_1)(x_{2j}-\bar{x}_2)+b_2\sum_{j=1}^{n} (x_{2j}-\bar{x}_2)^2 = \sum_{j=1}^{n}(x_{2j}-\bar{x}_2)(y_j-\bar{y}). \end{cases}$$

(Immers sommen als $\sum_{j=1}^{n} \bar{x}_1(y_j-\bar{y})$ zijn nul en dus mag $\sum_{j=1}^{n} x_{1j}(y_j-\bar{y})$ vervangen worden door $\sum_{j=1}^{n} (x_{1j}-\bar{x}_1)(y_j-\bar{y})$, enz.). De oplossing van deze lineaire vergelijkingen kunnen wij schrijven in de vorm:

$$b_1 = \frac{\begin{vmatrix} \sum_{j=1}^{n} (x_{1j}-\bar{x}_1)(y_j-\bar{y}) & \sum_{j=1}^{n} (x_{1j}-\bar{x}_1)(x_{2j}-\bar{x}_2) \\ \sum_{j=1}^{n} (x_{2j}-\bar{x}_2)(y_j-\bar{y}) & \sum_{j=1}^{n} (x_{2j}-\bar{x}_2)^2 \end{vmatrix}}{\begin{vmatrix} \sum_{j=1}^{n} (x_{1j}-\bar{x}_1)^2 & \sum_{j=1}^{n} (x_{1j}-\bar{x}_1)(x_{2j}-\bar{x}_2) \\ \sum_{j=1}^{n} (x_{1j}-\bar{x}_1)(x_{2j}-\bar{x}_2) & \sum_{j=1}^{n} (x_{2j}-\bar{x}_2)^2 \end{vmatrix}}$$

en

$$b_2 = \frac{\begin{vmatrix} \sum_{j=1}^{n} (x_{1j}-\bar{x}_1)^2 & \sum_{j=1}^{n} (x_{1j}-\bar{x}_1)(y_j-\bar{y}) \\ \sum_{j=1}^{n} (x_{1j}-\bar{x}_1)(x_{2j}-\bar{x}_2) & \sum_{j=1}^{n} (x_{2j}-\bar{x}_2)(y_j-\bar{y}) \end{vmatrix}}{\begin{vmatrix} \sum_{j=1}^{n} (x_{1j}-\bar{x}_1)^2 & \sum_{j=1}^{n} (x_{1j}-\bar{x}_1)(x_{2j}-\bar{x}_2) \\ \sum_{j=1}^{n} (x_{1j}-\bar{x}_1)(x_{2j}-\bar{x}_2) & \sum_{j=1}^{n} (x_{2j}-\bar{x}_2)^2 \end{vmatrix}}$$

Voor het geval k=1 zouden we op deze manier de bekende uitdrukking:

$$b_1 = \frac{\sum_{j=1}^{n} (x_{1j}-\bar{x}_1)(y_j-\bar{y})}{\sum_{j=1}^{n} (x_{1j}-\bar{x}_1)^2}$$

gevonden hebben.

3. Kleinste kwadratenschattingen voor het algemene geval

Ook in het algemene geval vinden wij voor a, als oplossing van

$$(3,1) \qquad \left(\frac{\partial Q}{\partial \alpha}\right)_{\substack{\alpha=a\\ \beta=b}} = -2 \sum_{j=1}^{n} (y_j - a - b_1 x_{1j} - \ldots - b_k x_{kj}) = 0 \ ,$$

de uitdrukking

$$a = \sum_{j=1}^{n} y_j/n - b_1 \sum_{j=1}^{n} x_{1j}/n - \ldots - b_k \sum_{j=1}^{n} x_{kj}/n = \bar{y} - b_1\bar{x}_1 - \ldots - b_k\bar{x}_k .$$

Substitueren we dit in $\left(\frac{\partial Q}{\partial \beta_i}\right)_{\substack{\alpha=a\\ \beta=b}} = 0$ $(i=1,\ldots,k)$ dan ontstaat:

$$\sum_{j=1}^{n} x_{ij}\left[y_j - \bar{y} - b_1(x_{1j} - \bar{x}_1) - \ldots - b_k(x_{kj} - \bar{x}_k)\right] = 0 \quad (i=1,\ldots,k) \ ,$$

of

$$(3,2) \quad b_1 \sum_{j=1}^{n} (x_{1j} - \bar{x}_1)(x_{ij} - \bar{x}_i) + \ldots + b_k \sum_{j=1}^{n} (x_{kj} - \bar{x}_k)(x_{ij} - \bar{x}_i) =$$

$$= \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)(y_j - \bar{y}) \qquad (i=1,\ldots,k) \ .$$

We zullen nu om overzichtelijker formules te verkrijgen de volgende matrix-notatie invoeren:

$$y \underset{\sim}{\text{def}} \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix} ; \quad X \underset{\sim}{\text{def}} \begin{pmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & \ldots & x_{k1} - \bar{x}_k \\ x_{12} - \bar{x}_1 & x_{22} - \bar{x}_2 & \ldots & x_{k2} - \bar{x}_k \\ \vdots & \vdots & & \vdots \\ x_{1n} - \bar{x}_1 & x_{2n} - \bar{x}_2 & \ldots & x_{kn} - \bar{x}_k \end{pmatrix} ; \quad v \underset{\sim}{\text{def}} \begin{pmatrix} v_1 - \bar{v} \\ v_2 - \bar{v} \\ \vdots \\ v_n - \bar{v} \end{pmatrix} ;$$

$$\beta \underset{\sim}{\text{def}} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \qquad \text{en} \qquad b \underset{\sim}{\text{def}} \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix} .$$

Nu volgt uit (1,1) dat

$$\bar{y} = \alpha + \beta_1\bar{x}_1 + \ldots + \beta_k\bar{x}_k + \bar{v} \ ,$$

zodat we voor (1,1) ook mogen schrijven:

$$(\underline{y}_j - \overline{\underline{y}}) = \beta_1(x_{1j} - \overline{x}_1) + \ldots + \beta_k(x_{kj} - \overline{x}_k) + \underline{v}_j - \overline{\underline{v}} \; ; \quad j = 1, \ldots, n \; .$$

Met de bovengedefinieerde matrices is dit te schrijven als

$$\underline{y} = X\beta + \underline{v} \; . \tag{3,3}$$

De kwadratische vorm welke we moeten minimaliseren heeft nu de gedaante

$$Q = (\underline{y} - X\beta)' \; (\underline{y} - X\beta) \tag{3,4}$$ [3]

en de lineaire vergelijkingen waaraan de kleinste kwadratenschattingen moeten voldoen gaan over in:

$$(X'X)\, \underline{b} = X'\, \underline{y} \; .$$

Hierin is (X'X) een kxk-matrix. Is deze niet singulier dan geeft

$$\underline{b} = (X'X)^{-1} X'\, \underline{y} \tag{3,5}$$

de gezochte kleinste kwadratenschattingen voor $\beta_1, \ldots, \beta_k$. Om te bereiken dat (X'X) een inverse bezit, moeten we eisen dat X de rang k bezit, hetgeen overeenkomt met de eis dat de regressievectoren $(x_{11} - \overline{x}_1, \ldots, x_{1n} - \overline{x}_1)$, $(x_{21} - \overline{x}_2, \ldots, x_{2n} - \overline{x}_2)$, ... , $(x_{k1} - \overline{x}_k, \ldots, x_{kn} - \overline{x}_k)$ lineair onafhankelijk zijn, dus niet via lineaire betrekkingen in elkaar uitgedrukt kunnen worden.
Dit betekent o.a. ook dat $n > k$ is.

De kleinste kwadratenschattingen zijn de beste zuivere lineaire schattingen d.w.z.: ze bezitten van alle mogelijke zuivere lineaire schattingen de kleinste variantie.
Het minimum van Q, dat voor de waarden b van β bereikt wordt, is:

$$\begin{aligned} Q_{min} &= (y - Xb)'(y - Xb) = y'y - b'X'y - y'Xb + b'X'Xb = \\ &= y'y - y'X(X'X)^{-1}Xy - y'X(X'X)^{-1}X'y + \\ &\qquad + y'X(X'X)^{-1} X'X(X'X)^{-1}X'y = \\ &= y'y - y'X(X'X)^{-1} X'y = y'y - b'X'y \; . \end{aligned} \tag{3,6}$$

---

3) A', de getransformeerde van matrix A, ontstaat uit A door rijen en kolommen te verwisselen, dus $\begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{pmatrix}' = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}$.

4. Verwachtingen, varianties van $\underline{b}_1,\ldots,\underline{b}_k$

Uit de uitdrukking (3,5) voor $\underline{b}$ zien we dat de $\underline{b}$'s steeds lineaire combinaties van $\underline{y}_1,\ldots,\underline{y}_n$ zijn. Daar in (3,3) $X\underline{v}_j = 0$ is en dus $\mathcal{E}\,\underline{y} = X\beta$, geldt voor de verwachting van b:

$$\mathcal{E}\,\underline{b} = (X'X)^{-1}X'\,\mathcal{E}\,\underline{y} = (X'X)^{-1}X'X\beta = \mathcal{I}\beta = \beta \ .$$

De schattingen $\underline{b}_i$ zijn dus zuiver. Dit geldt ook, indien de v_i niet onderling onafhankelijk zijn en/of verschillende spreidingen bezitten, doch wel $\mathcal{E}\,\underline{v}_i = 0$ is $(i=1,\ldots,n)$.

Zijn de grootheden $\underline{y}_1,\ldots,\underline{y}_n$ wel onderling onafhankelijk verdeeld met dezelfde variantie σ^2, dan kunnen we als volgt de covariantiematrix van $\underline{b}_1,\ldots,\underline{b}_k$ bepalen. We maken weer gebruik van het feit dat $\sum_{j=1}^{n}(x_{1j}-\bar{x}_1)(y_j-\bar{y}) = \sum_{j=1}^{n}(x_{1j}-\bar{x}_1)\,y_j$ is, dus dat

$$X'y = \begin{pmatrix} \sum_{j=1}^{n}(x_{1j}-\bar{x}_1)(y_j-\bar{y}) \\ \vdots \\ \sum_{j=1}^{n}(x_{kj}-\bar{x}_k)(y_j-\bar{y}) \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n}(x_{1j}-\bar{x}_1)\,y_j \\ \vdots \\ \sum_{j=1}^{n}(x_{kj}-\bar{x}_k)\,y_j \end{pmatrix}$$

We kunnen dan volgens (3,5) $b_1,\ldots,b_k$ schrijven als:

$$\begin{cases} b_1 = A_{11}\,y_1 + \ldots + A_{1n}\,y_n \ , \\ \vdots \\ b_k = A_{k1}\,y_1 + \ldots + A_{kn}\,y_n \ , \end{cases}$$

waarin de kxk-matrix $A = (X'X)^{-1}X'$ is. Voor de covariantiematrix van $\underline{b}_1,\ldots,\underline{b}_k$ geldt dan, daar $\sigma\{\underline{y}_j,\underline{y}_\ell\} = 0$ $(j\neq\ell)$ en $\sigma^2\{\underline{y}_j\}=\sigma^2$ is

$$(4,1)\qquad (\sigma^2\{\underline{b}_i,\underline{b}_k\}) = \Big(\sum_{j=1}^{n} A_{ij}A_{kj}\,\sigma^2\{\underline{y}_j\}\Big) =$$

$$= \sigma^2.\Big(\sum_{j=1}^{n} A_{ij}A_{kj}\Big) = \sigma^2.AA' =$$

$$= \sigma^2.(X'X)^{-1}\,X'X\,(X'X)^{-1} =$$

$$= \sigma^2.\,(X'X)^{-1} \ .$$

Voor het geval k=1 is dus:

$$\sigma^2\{\underline{b}_1\} = \sigma^2/\sum_{j=1}^{n}(x_{1j}-\bar{x}_1)^2 \ .$$

Voor het geval k=2 is, wanneer we

$$Z \underline{\underline{\text{def}}} \begin{vmatrix} \sum_{j=1}^{n} (x_{1j}-\bar{x}_1)^2 & \sum_{j=1}^{n} (x_{1j}-\bar{x}_1)(x_{2j}-\bar{x}_2) \\ \sum_{j=1}^{n} (x_{1j}-\bar{x}_1)(x_{2j}-\bar{x}_2) & \sum_{j=1}^{n} (x_{2j}-\bar{x}_2)^2 \end{vmatrix} = X'X$$

stellen:

$$\sigma^2\{\underline{b}_1\} = \sigma^2 . \sum_{j=1}^{n} (x_{2j}-\bar{x}_2)^2 \; / \; |Z| \quad ,$$

$$\sigma^2\{\underline{b}_2\} = \sigma^2 . \sum_{j=1}^{n} (x_{1j}-\bar{x}_1)^2 \; / \; |Z| \quad ,$$

$$\sigma^2\{\underline{b}_1, \underline{b}_2\} = -\sigma^2 . \sum_{j=1}^{n} (x_{1j}-\bar{x}_1)(x_{2j}-\bar{x}_2) \; / \; |Z| \quad .$$

En in 't algemeen wanneer we de minor van het element in de i^e rij en k^e kolom van Z door $|Z_{ik}|$ voorstellen:

$$(4,2) \qquad \sigma^2\{\underline{b}_i, \underline{b}_k\} = \sigma^2 . |Z_{ik}| \; / \; |Z| \quad .$$

Men kan aantonen dat

$$(n-k-1)^{-1} \underline{Q}_{min}$$

een zuivere schatting voor σ^2 is (J. van YZEREN (1954)).

Zijn $\underline{v}_1, \ldots, \underline{v}_n$ onderling onafhankelijk normaal verdeeld met gelijke spreidingen, dan komen de kleinste kwadratenschattingen voor $\alpha, \beta_1, \ldots, \beta_k$ overeen met de meest aannemelijke schattingen (maximum likelihood estimates). De meest aannemelijke schatting voor σ^2 is dan $n^{-1}\underline{Q}_{min}$, deze is asymptotisch equivalent met de zuivere schatting $(n-k-1)^{-1} \underline{Q}_{min}$. De grootheid Q_{min}/σ^2 heeft nu een χ^2-verdeling met (n-k-1) vrijheidsgraden en is onafhankelijk verdeeld van $\underline{a}, \underline{b}_1, \ldots, \underline{b}_k$.

Literatuur

H.B. MANN (1949) Analysis and design of experiments, hoofdstuk 4. New York, Dover Publications, Inc.

H. CRAMÉR (1951) Mathematical methods of statistics, hoofdstuk 37. Princeton University Press.

J. VAN IJZEREN (1954) De theoretische zijde van de methode der kleinste kwadraten, Statistica 8, pp. 21-45.