Statistik in Geodäsie, Geoinformation und Bauwesen

Bearbeitet von
Wilhelm Benning

erweitert, überarbeitet 2011. Taschenbuch. XII, 324 S. Paperback
ISBN 978 3 87907 512 6
Format (B x L): 17 x 24 cm
Gewicht: 592 g

6 Regressionsanalyse

6.1 Lineares Modell

6.1.1 Modelldefinition

Ein *lineares statistisches Modell* ist ein System von n Zufallsvariablen Y_i, $i = 1, \ldots, n$, der Form

$$\begin{aligned} Y_1 &= \beta_0 + x_{11}\,\beta_1 + x_{12}\,\beta_2 + \ldots + x_{1p}\,\beta_p + e_1 \\ Y_2 &= \beta_0 + x_{21}\,\beta_1 + x_{22}\,\beta_2 + \ldots + x_{2p}\,\beta_p + e_2 \\ &\vdots \\ Y_n &= \beta_0 + x_{n1}\,\beta_1 + x_{n2}\,\beta_2 + \ldots + x_{np}\,\beta_p + e_n\,, \end{aligned} \tag{6.1}$$

wobei

1. $e_i\ (i = 1, \ldots, n)$ Zufallsvariablen mit den Erwartungswerten $E(e_i) = 0$
2. $\beta_0 + x_{i1}\,\beta_1 + x_{i2}\,\beta_2 + \ldots + x_{ip}\,\beta_p\ (i = 1, \ldots, n)$ Linearformen mit (im Allgemeinen) unbekannten Parametern $\beta_0, \ldots, \beta_p$ (Anzahl $u = p + 1$) und vorgegebenen Koeffizienten $x_{i1}, x_{i2}, \ldots, x_{ip}\ (i = 1, \ldots, n)$

sind.

Wird zusätzlich eine Hilfsvariable x_0 mit $x_{i0} = 1$ eingeführt, ergibt sich mit dem Zufallsvektor $\mathbf{y}$, der Koeffizientenmatrix $\mathbf{X}$, dem Parametervektor $\boldsymbol{\beta}$ und dem Zufallsvektor $\mathbf{e}$

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{10} & x_{11} & \ldots & x_{1p} \\ x_{20} & x_{21} & \ldots & x_{2p} \\ \vdots & & & \vdots \\ x_{n0} & x_{n1} & \ldots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Gl. (6.1) in Matrizenschreibweise zu:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad E(\mathbf{e}) = \mathbf{0} \tag{6.2}$$

Das Wort „linear“ drückt aus, dass zwischen den u unbekannten Modellparametern β_j , $j = 0, \ldots, p$, eine lineare Beziehung unterstellt wird. Man geht davon aus, dass zwischen den beobachtbaren Größen Y_i, $x_{i1}, \ldots, x_{ip}$ ein funktionaler, modellierbarer Zusammenhang vorliegt, der sich „im Prinzip“ durch eine lineare Funktionsgleichung beschreiben lässt. In der Realität (d. h. bei den Beobachtungen dieser Größen) kommt es jedoch zu Abweichungen von diesem funktionalen Zusammenhang. Diese Abweichungen werden in das Modell einbezogen, in dem vorausgesetzt wird, dass die lineare funktionale Beziehung durch eine hinzuaddierte (nicht beobachtbare) Zufallsvariable e_i, als *Störvariable* bezeichnet, überlagert („gestört“) wird.

Die Zufallsvariable Y_i wird als *abhängige Variable* (auch *Zielvariable, endogene Variable, Regressand*) bezeichnet, $x_{i1}, \ldots, x_{ip}$ als *unabhängige Variablen* (auch *exogene Variablen, Regressoren*).

Um die u unbekannten Modellparameter β_j statistisch schätzen zu können, benötigt man je eine Datenreihe der Länge n für den Regressanden (Messwerte y_i, $i = 1, \ldots, n$) und für jeden der p Regressoren (Werte x_{ij}, $i = 1, \ldots, n\,, j = 1, \ldots, p$). Als Realisierungen der Gl. (6.1) erhält man bei $p = 1$ die Wertepaare $y_1, x_1\,;\ y_2, x_2\,; \ldots;\ y_n, x_n$, woraus sich die *einfache lineare* Regression

$$\begin{aligned} \hat{y}(x_i) &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\ \hat{e}_i &= y_i - \hat{y}(x_i)\,, \quad i = 1, \ldots n \end{aligned} \tag{6.3}$$

berechnen lässt. Hierbei sind $\hat{\beta}_0$ der *Achsabschnitt* auf der y-Achse und $\hat{\beta}_1$ der *Regressionskoeffizient*, d. h. die Steigung der Regressionsgeraden. Die Realisierungen $\hat{e}_i$ der (nicht beobachtbaren) Störvariablen werden als *Residuen* bezeichnet.

Bei $p > 1$ lässt sich mit den Wertetupeln $y_1, x_{11}, x_{12}, \ldots, x_{1p}\,;\quad y_2,\ x_{21},\ x_{22},\ \ldots,\ x_{2p}\,;\ \ldots\ ;\ y_n, x_{n1}, x_{n2},\ \ldots,\ x_{np}$ die *multiple lineare Regression*

$$\begin{aligned} \hat{y}(x_i) &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \beta_p x_{ip} \\ \hat{e}_i &= y_i - \hat{y}(x_i)\,, \quad i = 1, \ldots n \end{aligned} \tag{6.4}$$

berechnen.

6.1.2 Linearisierung und Gauß-Newton-Verfahren

Vielfach lassen sich die Anwendungen der Praxis nicht mit einem linearen Modell beschreiben, sondern nur durch Regressionsfunktionen, bei denen die Parameter nichtlinear eingehen. Mit der Notation von Gl. (6.2) lässt sich eine *nichtlineare Regressionsfunktion* darstellen in der Form:

$$\mathbf{y} = \mathbf{g}(\boldsymbol{\beta}) + \mathbf{e}, \quad E(\mathbf{e}) = \mathbf{0}, \tag{6.5}$$

mit $\mathbf{y} = (Y_1, \ldots, Y_n)'$, $\quad \mathbf{g}(\boldsymbol{\beta}) = (g_1(\boldsymbol{\beta}), \ldots, g_n(\boldsymbol{\beta}))'$, $\quad \mathbf{e} = (e_1, \ldots, e_n)'$.

Damit die Schätzmethoden der linearen Modelle angewandt werden können, müssen die nichtlinearen Ansätze linearisiert werden. Dazu werden ausgehend von einer bekannten Startlösung $\boldsymbol{\beta}^{(0)}$ für die Parameter iterativ Näherungen $\boldsymbol{\beta}^{(1)}$, $\boldsymbol{\beta}^{(2)}$,..., konstruiert. Diese Folge erhält man mit dem *Gauß-Newton-Verfahren* durch Linearisieren von $\mathbf{g}(\boldsymbol{\beta})$. Die *Taylorentwicklung* um $\boldsymbol{\beta}^{(k)}$ wird nach dem linearen Term abgebrochen. Voraussetzung ist natürlich, dass $\mathbf{g}(\boldsymbol{\beta})$ wenigstens einmal stetig differenzierbar nach $\boldsymbol{\beta}$ ist. Im ersten Iterationsschritt zeigt sich die Parameterdarstellung $\beta_0 \approx \beta_0^{(0)} + \Delta\beta_0^{(0)}$, $\beta_1 \approx \beta_1^{(0)} + \Delta\beta_1^{(0)}, \ldots,\ \beta_p \approx \beta_p^{(0)} + \Delta\beta_p^{(0)}$, mit den bekannten Näherungswerten $\beta_j^{(0)}$ und den unbekannten Korrektionen $\Delta\beta_j^{(0)}$.

Allgemein gilt:

$$\mathbf{g}(\boldsymbol{\beta}) \approx \mathbf{g}(\boldsymbol{\beta}^{(k)}) + \mathbf{Z}_k \cdot \boldsymbol{\Delta\beta}^{(k)} \tag{6.6}$$

mit

$$\mathbf{Z}_k = \begin{pmatrix} \dfrac{\partial g_1(\boldsymbol{\beta}^{(k)})}{\partial \beta_0} & \dfrac{\partial g_1(\boldsymbol{\beta}^{(k)})}{\partial \beta_1} & \dots & \dfrac{\partial g_1(\boldsymbol{\beta}^{(k)})}{\partial \beta_p} \\ \dfrac{\partial g_2(\boldsymbol{\beta}^{(k)})}{\partial \beta_0} & \dfrac{\partial g_2(\boldsymbol{\beta}^{(k)})}{\partial \beta_1} & \dots & \dfrac{\partial g_2(\boldsymbol{\beta}^{(k)})}{\partial \beta_p} \\ \vdots & & & \vdots \\ \dfrac{\partial g_n(\boldsymbol{\beta}^{(k)})}{\partial \beta_0} & \dfrac{\partial g_n(\boldsymbol{\beta}^{(k)})}{\partial \beta_1} & \dots & \dfrac{\partial g_n(\boldsymbol{\beta}^{(k)})}{\partial \beta_p} \end{pmatrix} \tag{6.7}$$

und $\quad k = 0, 1, 2, \dots$

Identifiziert man in Gl. (6.2)

$$\begin{aligned} \mathbf{y} &\quad \text{mit} \quad \mathbf{y} - \mathbf{g}(\boldsymbol{\beta}^{(k)}) \\ \boldsymbol{\beta} &\quad \text{mit} \quad \boldsymbol{\Delta\beta}^{(k)} \\ \mathbf{X} &\quad \text{mit} \quad \mathbf{Z}_k \end{aligned} \tag{6.8}$$

ergibt sich eine Darstellung im linearen Modell.

Beispiel 6.1: *Linearisierung einer nichtlinearen Regressionsfunktion*
Bei der polaren Punktbestimmung oder bei der Punktbestimmung durch Bogenschnitt werden jeweils von einem Festpunkt F aus die horizontale Strecke s zum Neupunkt N gemessen, um mit dieser Strecke und weiteren Messelementen die Koordinaten y_N, x_N zu berechnen. Die Koordinaten y_F, x_F des Festpunktes werden als gegeben vorausgesetzt.
Da die Erwartungswerte von Messgrößen (hier die als Zufallsvariable mit großem Buchstaben bezeichnete Strecke S) als Funktion der gesuchten Parameter (hier y_N und x_N) und gegebener Koeffizienten dargestellt werden, ergibt sich die nichtlineare Regressionsfunktion

$$\begin{aligned} E(S) &= \sqrt{(y_N - y_F)^2 + (x_N - x_F)^2} \qquad \text{bzw.} \\ S &= \sqrt{(y_N - y_F)^2 + (x_N - x_F)^2} + e\,, \end{aligned}$$

wobei e die Störvariable ist. Da mit $\Delta y, \Delta x$ üblicherweise die Koordinatenunterschiede zwischen zwei Punkten bezeichnet werden, sollen im Falle der Koordinatenausgleichung die unbekannten Koordinatenzuschläge mit $\delta y, \delta x$ bezeichnet werden. Die linearisierte Beobachtungsgleichung zur nichtlinearen Regressionsfunktion lautet damit

$$\begin{aligned} S - s^{(0)} &= \begin{pmatrix} \dfrac{\partial s^{(0)}}{\partial y_N} & \dfrac{\partial s^{(0)}}{\partial x_N} \end{pmatrix} \begin{pmatrix} \delta y_N^{(0)} \\ \delta x_N^{(0)} \end{pmatrix} + e \\ &= \begin{pmatrix} \dfrac{y_N^{(0)} - y_F}{s^{(0)}} & \dfrac{x_N^{(0)} - x_F}{s^{(0)}} \end{pmatrix} \begin{pmatrix} \delta y_N^{(0)} \\ \delta x_N^{(0)} \end{pmatrix} + e \\ &= \frac{y_N^{(0)} - y_F}{s^{(0)}} \cdot \delta y_N^{(0)} + \frac{x_N^{(0)} - x_F}{s^{(0)}} \cdot \delta x_N^{(0)} + e\,, \end{aligned}$$

wobei der Näherungswert für die Strecke

$$s^{(0)} = \sqrt{(y_N^{(0)} - y_F)^2 + (x_N^{(0)} - x_F)^2}$$

sowie die partiellen Ableitungen mithilfe der Näherungskoordinaten $y_N^{(0)}, x_N^{(0)}$ berechnet werden. Werden im anschließenden Ausgleichungsprozess für die jeweilige Zufallsvariable S deren Realisierungen, also die Messwerte s_i, eingesetzt, dann bilden die Streckendifferenzen $s_i - s_i^{(0)}$ zusammen mit den entsprechenden Differenzen anderer Messelemente (wie z. B. Richtungsdifferenzen $r_i - r_i^{(0)}$) nach der Notation der Gl. (6.2) den Beobachtungsvektor $\mathbf{y}$. Die partiellen Ableitungen sind in der Koeffizientenmatrix $\mathbf{X}$ zusammengefasst und die $\delta y_N^{(0)}, \delta x_N^{(0)}$ stellen den unbekannten Parametervektor $\boldsymbol{\beta}$ dar. Mit den sich ergebenden Näherungskoordinaten $y_N^{(1)}, x_N^{(1)}$ kann eine weitere Iteration erfolgen.

6.2 Klassisches und allgemeines lineares Regressionsmodell

6.2.1 Modellbeschreibung

Durch unterschiedliche Annahmen über $\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \mathbf{e}$ im linearen Modell Gl. (6.2) ergeben sich unterschiedliche Regressionsmodelle. Die Annahmen für das *klassische* und das *allgemeine lineare Regressionsmodell* lauten:

$\mathbf{y}$: beobachtbare Zufallsvariablen Y_i

$\mathbf{X}$: beobachtbare vorgegebene (deterministische) Variablen. Diese Variablen sind keine Zufallsvariablen, d. h. die Werte x_{ij} lassen sich systematisch oder „kontrolliert" variieren. Bei wiederholter Messung der Zufallsvariable Y_i können die Werte x_{ij} konstant gehalten werden.

$\boldsymbol{\beta}$: fester, unbekannter Parametervektor

$\mathbf{e}$: nicht beobachtbare Zufallsvariable mit $E(\mathbf{e}) = \mathbf{0}$ und

- beim *klassischen linearen Regressionsmodell*:

$$\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}, \quad \text{d. h. } var(e_i) = \sigma^2, \quad cov(e_i, e_k) = 0, \quad i \neq k \tag{6.9}$$

  σ^2 ist ein weiterer, im Allgemeinen unbekannter Parameter.

- beim *allgemeinen linearen Regressionsmodell*:

$$\begin{aligned} \boldsymbol{\Sigma} &= \begin{pmatrix} var(e_1) & cov(e_1,e_2) & \dots & cov(e_1,e_n) \\ cov(e_2,e_1) & var(e_2) & \dots & cov(e_2,e_n) \\ \vdots & & \ddots & \vdots \\ cov(e_n,e_1) & cov(e_n,e_2) & \dots & var(e_n) \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{pmatrix} \end{aligned} \tag{6.10}$$