Anhang B

Regression

Dieser Anhang rekapituliert die in der Analysis und Statistik wohlbekannte **Methode der kleinsten Quadrate**, auch **Regression** genannt, zur Bestimmung von Ausgleichsgeraden (Regressionsgeraden) und allgemein Ausgleichpolynomen. Die Darstellung folgt im wesentlichen [Heuser 1988].

(Physikalische) Meßdaten zeigen selten exakt den gesetzmäßigen Zusammenhang der gemessenen Größen, da sie unweigerlich mit Fehlern behaftet sind. Will man den Zusammenhang der gemessenen Größen dennoch (wenigstens näherungsweise) bestimmen, so steht man vor der Aufgabe, eine Funktion zu finden, die sich den Meßdaten möglichst gut anpaßt, so daß die Meßfehler „ausgeglichen“ werden. Natürlich sollte dazu bereits eine Hypothese über die Art des Zusammenhangs vorliegen, um eine Funktionenklasse wählen und dadurch das Problem auf die Bestimmung der Parameter einer Funktion eines bestimmten Typs reduzieren zu können.

Erwartet man z.B. bei zwei Größen x und y einen linearen Zusammenhang (z.B. weil ein Diagramm der Meßpunkte einen solchen vermuten läßt), so muß man die Parameter a und b der Gerade $y = g(x) = a + bx$ bestimmen. Wegen der unvermeidlichen Meßfehler wird es jedoch i.a. nicht möglich sein, eine Gerade zu finden, so daß alle gegebenen n Meßpunkte (x_i, y_i), $1 \leq i \leq n$, genau auf dieser Geraden liegen. Vielmehr wird man versuchen müssen, eine Gerade zu finden, von der die Meßpunkte möglichst wenig abweichen. Es ist daher plausibel, die Parameter a und b so zu bestimmen, daß die Abweichungsquadratsumme

$$F(a,b) \;=\; \sum_{i=1}^{n} (g(x_i) - y_i)^2 \;=\; \sum_{i=1}^{n} (a + bx_i - y_i)^2$$

minimal wird. D.h., die aus der Geradengleichung berechneten y-Werte sollen (in der Summe) möglichst wenig von den gemessenen abweichen. Die Gründe für die Verwendung des Abweichungsquadrates sind i.w. die gleichen wie die in Abschnitt 3.3 angeführten: Ersten ist die Fehlerfunktion durch die Verwendung des Quadrates überall (stetig) differenzierbar, während die Ableitung des Betrages, den man alternativ verwenden könnte, bei 0 nicht existiert/unstetig ist. Zweitens gewichtet das Quadrat große Abweichungen von der gewünschten Ausgabe stärker, so daß vereinzelte starke Abweichungen von den Meßdaten tendenziell vermieden werden.[1]

Eine notwendige Bedingung für ein Minimum der oben definierten Fehlerfunktion $F(a, b)$ ist, daß die partiellen Ableitungen dieser Funktion nach den Parametern a und b verschwinden, also

$$\begin{aligned} \frac{\partial F}{\partial a} &= \sum_{i=1}^{n} 2(a + bx_i - y_i) &&= 0 \qquad \text{und} \\ \frac{\partial F}{\partial b} &= \sum_{i=1}^{n} 2(a + bx_i - y_i)x_i &&= 0 \end{aligned}$$

gilt. Aus diesen beiden Gleichungen erhalten wir nach wenigen einfachen Umformungen die sogenannten **Normalgleichungen**

$$\begin{aligned} na + \left(\sum_{i=1}^{n} x_i\right) b &= \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right) a + \left(\sum_{i=1}^{n} x_i^2\right) b &= \sum_{i=1}^{n} x_i y_i, \end{aligned}$$

also ein lineares Gleichungssystem mit zwei Gleichungen und zwei Unbekannten a und b. Man kann zeigen, daß dieses Gleichungssystem eine eindeutige Lösung besitzt, es sei denn, die x-Werte aller Meßpunkte sind identisch (d.h., es ist $x_1 = x_2 = \ldots = x_n$), und daß diese Lösung tatsächlich ein Minimum der Funktion F beschreibt [Heuser 1988]. Die auf diese Weise bestimmte Gerade $y = g(x) = a + bx$ nennt man die **Ausgleichsgerade** oder **Regressionsgerade** für den Datensatz $(x_1, y_1), \ldots, (x_n, y_n)$.

[1] Man beachte allerdings, daß dies auch ein Nachteil sein kann. Enthält der gegebene Datensatz „Ausreißer“ (das sind Meßwerte, die durch zufällig aufgetretene, unverhältnismäßig große Meßfehler sehr weit von dem tatsächlichen Wert abweichen), so wird die Lage der berechneten Ausgleichsgerade u.U. sehr stark von wenigen Meßpunkten (eben den Ausreißern) beeinflußt, was das Ergebnis unbrauchbar machen kann.

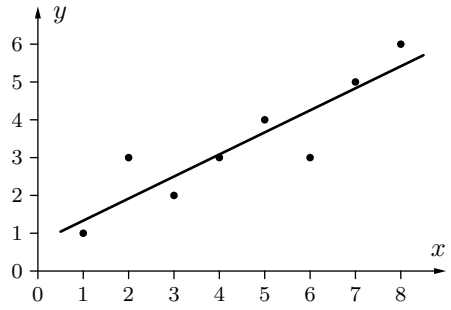

Abbildung B.1: Beispieldaten und mit der Methode der kleinsten Quadrate berechnete Ausgleichsgerade.

Zur Veranschaulichung des Verfahrens betrachten wir ein einfaches Beispiel. Gegeben sei der aus acht Meßpunkten $(x_1, y_1), \ldots, (x_8, y_8)$ bestehende Datensatz, der in der folgenden Tabelle gezeigt ist [Heuser 1988]:

x	1	2	3	4	5	6	7	8
y	1	3	2	3	4	3	5	6

Um das System der Normalgleichungen aufzustellen, berechnen wir

$$\sum_{i=1}^{8} x_i = 36, \qquad \sum_{i=1}^{8} x_i^2 = 204, \qquad \sum_{i=1}^{8} y_i = 27, \qquad \sum_{i=1}^{8} x_i y_i = 146.$$

Damit erhalten wir das Gleichungssystem (Normalgleichungen)

$$\begin{aligned} 8a &+ 36b &= 27, \\ 36a &+ 204b &= 146, \end{aligned}$$

das die Lösung $a = \frac{3}{4}$ und $b = \frac{7}{12}$ besitzt. Die Ausgleichsgerade ist also

$$y = \frac{3}{4} + \frac{7}{12}x.$$

Diese Gerade ist zusammen mit den Datenpunkten, von denen wir ausgegangen sind, in Abbildung B.1 dargestellt.

Das gerade betrachtete Verfahren ist natürlich nicht auf die Bestimmung von Ausgleichsgeraden beschränkt, sondern läßt sich mindestens auf Ausgleichspolynome erweitern. Man sucht dann nach einem Polynom

$$y = p(x) = a_0 + a_1 x + \ldots + a_m x^m$$

mit gegebenem, festem Grad m, das die n Meßpunkte $(x_1, y_1), \ldots, (x_n.y_n)$ möglichst gut annähert. In diesem Fall ist

$$F(a_0, a_1, \ldots, a_m) \;=\; \sum_{i=1}^{n} (p(x_i) - y_i)^2 \;=\; \sum_{i=1}^{n} (a_0 + a_1 x_i + \ldots + a_m x_i^m - y_i)^2$$

zu minimieren. Notwendige Bedingung für ein Minimum ist wieder, daß die partiellen Ableitungen nach den Parametern a_0 bis a_m verschwinden, also

$$\frac{\partial F}{\partial a_0} = 0, \quad \frac{\partial F}{\partial a_1} = 0, \quad \ldots \quad , \frac{\partial F}{\partial a_m} = 0$$

gilt. So ergibt sich das System der Normalgleichungen [Heuser 1988]

$$\begin{array}{rcccccc}
n a_0 & + & \left(\sum_{i=1}^{n} x_i\right) a_1 & + \ \ldots \ + & \left(\sum_{i=1}^{n} x_i^m\right) a_m & = & \sum_{i=1}^{n} y_i \\
\left(\sum_{i=1}^{n} x_i\right) a_0 & + & \left(\sum_{i=1}^{n} x_i^2\right) a_1 & + \ \ldots \ + & \left(\sum_{i=1}^{n} x_i^{m+1}\right) a_m & = & \sum_{i=1}^{n} x_i y_i \\
\vdots & & & & & & \vdots \\
\left(\sum_{i=1}^{n} x_i^m\right) a_0 & + & \left(\sum_{i=1}^{n} x_i^{m+1}\right) a_1 & + \ \ldots \ + & \left(\sum_{i=1}^{n} x_i^{2m}\right) a_m & = & \sum_{i=1}^{n} x_i^m y_i,
\end{array}$$

aus dem sich die Parameter a_0 bis a_m mit den üblichen Methoden der linearen Algebra (z.B. Gaußsches Eliminationsverfahren, Cramersche Regel, Bildung der Inversen der Koeffizientenmatrix etc.) berechnen lassen. Das so bestimmte Polynom $p(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m$ heißt **Ausgleichspolynom** oder **Regressionspolynom** m-ter Ordnung für den Datensatz $(x_1, y_1), \ldots, (x_n, y_n)$.

Weiter läßt sich die Methode der kleinsten Quadrate nicht nur verwenden, um, wie bisher betrachtet, Ausgleichspolynome zu bestimmen, sondern kann auch für Funktionen mit mehr als einem Argument eingesetzt werden. In diesem Fall spricht man von **multipler** oder **multivariater Regression**. Wir untersuchen hier beispielhaft nur den Spezialfall der **multilinearen Regression**, wobei wir uns außerdem zunächst auf eine Funktion mit zwei Argumenten beschränken. D.h., wir betrachten, wie man zu einem gegebenen Datensatz $(x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)$ eine Ausgleichsfunktion der Form

$$z = f(x, y) = a + bx + cy$$

so bestimmen kann, daß die Summe der Abweichungsquadrate minimal wird. Die Ableitung der Normalgleichungen für diesen Fall ist zu der Ableitung für Ausgleichspolynome völlig analog. Wir müssen

$$F(a,b,c) \;=\; \sum_{i=1}^{n}(f(x_i,y_i)-z_i)^2 \;=\; \sum_{i=1}^{n}(a+bx_i+cy_i-z_i)^2$$

minimieren. Notwendige Bedingungen für ein Minimum sind

$$\begin{aligned}
\frac{\partial F}{\partial a} &= \sum_{i=1}^{n} 2(a+bx_i+cy_i-z_i) &&= 0,\\
\frac{\partial F}{\partial b} &= \sum_{i=1}^{n} 2(a+bx_i+cy_i-z_i)x_i &&= 0,\\
\frac{\partial F}{\partial c} &= \sum_{i=1}^{n} 2(a+bx_i+cy_i-z_i)y_i &&= 0.
\end{aligned}$$

Also erhalten wir das System der Normalgleichungen

$$\begin{aligned}
na \;+\; \left(\sum_{i=1}^{n} x_i\right) b \;+\; \left(\sum_{i=1}^{n} y_i\right) c &= \sum_{i=1}^{n} z_i\\
\left(\sum_{i=1}^{n} x_i\right) a \;+\; \left(\sum_{i=1}^{n} x_i^2\right) b \;+\; \left(\sum_{i=1}^{n} x_i y_i\right) c &= \sum_{i=1}^{n} z_i x_i\\
\left(\sum_{i=1}^{n} y_i\right) a \;+\; \left(\sum_{i=1}^{n} x_i y_i\right) b \;+\; \left(\sum_{i=1}^{n} y_i^2\right) c &= \sum_{i=1}^{n} z_i y_i
\end{aligned}$$

aus dem sich a, b und c leicht berechnen lassen.

Im allgemeinen Fall der multilinearen Regression (Funktion mit m Argumenten) ist ein Datensatz $((x_{11},\ldots,x_{m1},y_1),\ldots,(x_{1n},\ldots,x_{mn},y_n))$ (oder auch dargestellt als $((\vec{x}_1,y_1),\ldots,(\vec{x}_n,y_n))$ mit einem Eingabevektor $\vec{x}_i$ und der zugehörigen Ausgabe y_i, $1 \le i \le n$) gegeben, für den eine lineare Ausgleichsfunktion

$$y = f(x_1,\ldots,x_m) = a_0 + \sum_{k=1}^{m} a_k x_k$$

gesucht ist. Zur Ableitung der Normalgleichungen stellt man in diesem Fall das zu minimierende Funktional bequemer in Matrixform dar, nämlich als

$$F(\vec{a}) = (\mathbf{X}\vec{a}-\vec{y})^\top(\mathbf{X}\vec{a}-\vec{y}),$$

wobei

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1m} \\ \vdots & & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nm} \end{pmatrix} \quad \text{und} \quad \vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

den Datensatz wiedergeben und $\vec{a} = (a_0, a_1, \dots, a_m)^\top$ der Vektor der zu bestimmenden Koeffizienten ist.[2] (Man beachte, daß die Einsen in der ersten Spalte der Matrix $\mathbf{X}$ dem Koeffizienten a_0 zugehören.) Wieder ist eine notwendige Bedingung für ein Minimum, daß die partiellen Ableitungen nach den Koeffizienten a_k, $0 \leq k \leq m$, verschwinden, was wir mit Hilfe des Differentialoperators ∇ (gesprochen: nabla) schreiben können als

$$\nabla_{\vec{a}}\, F(\vec{a}) = \frac{\mathrm{d}}{\mathrm{d}\vec{a}} F(\vec{a}) = \left(\frac{\partial}{\partial a_0} F(\vec{a}), \frac{\partial}{\partial a_1} F(\vec{a}), \dots, \frac{\partial}{\partial a_m} F(\vec{a}) \right) = \vec{0}.$$

Die Ableitung läßt sich am leichtesten berechnen, wenn man sich klar macht (wie man durch elementweises Ausschreiben leicht nachrechnet), daß sich der Differentialoperator

$$\nabla_{\vec{a}} = \left(\frac{\partial}{\partial a_0}, \frac{\partial}{\partial a_1}, \dots, \frac{\partial}{\partial a_m} \right)$$

formal wie ein Vektor verhält, der von links an die Summe der Fehlerquadrate „heranmultipliziert“ wird. Alternativ kann man die Ableitung komponentenweise ausschreiben. Wir verwenden hier jedoch die erstere, wesentlich bequemere Methode und erhalten

$$\begin{aligned} \vec{0} &= \nabla_{\vec{a}} \left(\mathbf{X}\vec{a} - \vec{y}\right)^\top (\mathbf{X}\vec{a} - \vec{y}) \\ &= \left(\nabla_{\vec{a}} \left(\mathbf{X}\vec{a} - \vec{y}\right)\right)^\top (\mathbf{X}\vec{a} - \vec{y}) + \left(\left(\mathbf{X}\vec{a} - \vec{y}\right)^\top \left(\nabla_{\vec{a}} \left(\mathbf{X}\vec{a} - \vec{y}\right)\right)\right)^\top \\ &= \left(\nabla_{\vec{a}} \left(\mathbf{X}\vec{a} - \vec{y}\right)\right)^\top (\mathbf{X}\vec{a} - \vec{y}) + \left(\nabla_{\vec{a}} \left(\mathbf{X}\vec{a} - \vec{y}\right)\right)^\top (\mathbf{X}\vec{a} - \vec{y}) \\ &= 2\mathbf{X}^\top (\mathbf{X}\vec{a} - \vec{y}) \\ &= 2\mathbf{X}^\top \mathbf{X}\vec{a} - 2\mathbf{X}^\top \vec{y}, \end{aligned}$$

woraus sich unmittelbar das System

$$\mathbf{X}^\top \mathbf{X}\vec{a} = \mathbf{X}^\top \vec{y}$$

[2] $\top$ bedeutet die Transponierung eines Vektors oder einer Matrix, also die Vertauschung von Zeilen und Spalten.

der Normalgleichungen ergibt. Dieses System ist offenbar lösbar, wenn $\mathbf{X}^\top\mathbf{X}$ invertierbar ist. Dann gilt

$$\vec{a} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\vec{y}.$$

Den Ausdruck $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ nennt man auch die (Moore-Penrose-) **Pseudoinverse** der Matrix $\mathbf{X}$ [Albert 1972]. Mit ihr kann man unmittelbar die Lösung der Regressionsaufgabe angeben.

Es dürfte klar sein, daß sich die Methode der kleinsten Quadrate auch auf Polynome in mehreren Variablen erweitern läßt. Am einfachsten geht man dabei ebenfalls von der oben verwendeten Matrixdarstellung aus, wobei man in die Matrix $\mathbf{X}$ als Eingangsgrößen auch die zu verwendenden Potenzprodukte der unabhängigen Variablen einträgt. Die Ableitung der Normalgleichungen kann dann einfach übernommen werden.

Ein Programm zur multipolynomialen Regression, das zur schnellen Berechnung der benötigten Potenzprodukte eine auf Ideen der dynamischen Programmierung beruhende Methode benutzt, findet man unter

`http://fuzzy.cs.uni-magdeburg.de/~borgelt/software.html#regress`

Wie sie sich unter bestimmten Umständen die Regression auch noch auf andere Funktionenklassen erweitern läßt, ist in Abschnitt 4.3 anhand der **logistischen Regression** gezeigt.