FAKULTÄT FÜR MATHEMATIK UND NATURWISSENSCHAFTEN
INSTITUT FÜR PHYSIK
Dr. rer. nat. Norbert Stein

TECHNISCHE UNIVERSITÄT ILMENAU

Lineare Regression – Mathematische Grundlagen

Bei erwarteten linearen Zusammenhängen zwischen gewonnenen Messwerten $x_1, x_2, \ldots, x_N$ sowie $y_1, y_2, \ldots, y_N$ interessieren häufig die Parameter einer ausgleichenden Geraden $y = a\,x + b$ durch die Messpunkte. Diese kann optional auch ohne Absolutglied durch den Koordinatenursprung verlaufen.

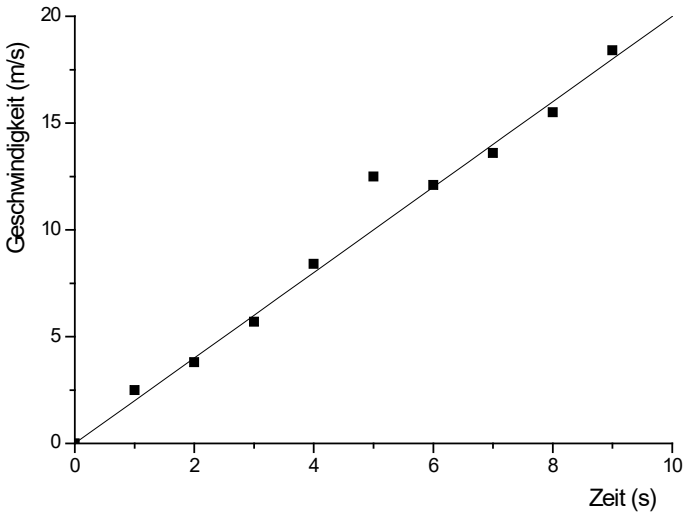

Abb.1: Spezialfall einer ausgleichenden Gerade durch den Koordinatenursprung

1. Allgemeine Gerade

Vereinfachend für die weiteren Berechnungen wird angenommen, dass die Abszissenwerte x_i ohne Abweichungen gegeben sind und die Ordinatenwerte y_i mit dem gleichen Gewicht betrachtet werden können. Unter Anwendung der Methode der kleinsten Quadrate geht man zunächst von der Summe der Abweichungsquadrate S aus, die durch Variation der Geradenkoeffizienten a und b minimiert wird. Mit σ^2, der Varianz der Einzelmessung, erhält man:

$$S = \frac{1}{\sigma^2}\sum_{i=1}^{N}\left(a\,x_i + b - y_i\right)^2 . \tag{1.1}$$

Nullsetzen der partiellen Ableitungen von S nach den Unbekannten a und b liefert die Normalgleichungen

$$\frac{\partial S}{\partial a} = \frac{2}{\sigma^2}\sum_{i=1}^{N}\left(a\,x_i + b - y_i\right)x_i = 0 \tag{1.2}$$

sowie

$$\frac{\partial S}{\partial b} = \frac{2}{\sigma^2}\sum_{i=1}^{N}\left(a\,x_i + b - y_i\right) = 0 . \tag{1.3}$$

Zur Vereinfachung der Schreibweise für die weiteren Rechenschritte werden nach Gauß die Summenzeichen $\sum_{i=1}^{N} z_i$ durch gleichbedeutende Klammerausdrücke $[z]$ ersetzt, sodass sich das zu lösende Gleichungssystem folgendermaßen darstellt:

$$\begin{pmatrix} N & [x] \\ [x] & [x^2] \end{pmatrix} \cdot \begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} [y] \\ [yx] \end{pmatrix} . \tag{1.4}$$

Die Lösung erfolgt mithilfe der Cramerschen Regel (Determinantenmethode), indem zunächst die Nennerdeterminante $D = N[x^2] - [x]^2$ berechnet wird. Man erhält folgende Schätzwerte für die Koeffizienten der Ausgleichsgerade:

$$\hat{a} = \frac{N[yx] - [y][x]}{D} \tag{1.5}$$

und

$$\hat{b} = \frac{[y][x^2] - [yx][x]}{D}. \tag{1.6}$$

$\hat{a}$ wird auch als Regressionskoeffizient bezeichnet. Für die empirische Standardabweichung der Einzelmessung $s(y_i)$ erhält man unter Berücksichtigung der Tatsache, dass zwei Unbekannte gesucht wurden:

$$s(y_i) = \sqrt{\frac{\sum_{i=1}^{N} \left(\hat{a}\, x_i + \hat{b} - y_i\right)^2}{N-2}}. \tag{1.7}$$

Sie muss bei der Angabe der Standardunsicherheiten der Größen $\hat{a}$ und $\hat{b}$ berücksichtigt werden. Hierzu wird das Gaußsche Fortpflanzungsgesetz für die Unsicherheiten, dessen Form für den Fall kleiner, unkorrelierter Abweichungen

$$s(f) = \sqrt{\left(\frac{\partial f}{\partial x_1}\right)^2 s(x_1)^2 + \left(\frac{\partial f}{\partial x_2}\right)^2 s(x_2)^2 + \ldots + \left(\frac{\partial f}{\partial x_N}\right)^2 s(x_N)^2} \tag{1.8}$$

ist, für jeden der gemessenen Werte y_i auf die Gleichungen (1.5) und (1.6) angewendet und man erhält zunächst für die partiellen Ableitungen:

$$\frac{\partial \hat{a}}{\partial y_i} = \frac{1}{D}\left(N x_i - [x]\right) \tag{1.9}$$

sowie

$$\frac{\partial \hat{b}}{\partial y_i} = \frac{1}{D}\left([x^2] - x_i[x]\right). \tag{1.10}$$

Mit der Standardabweichung $s(y_i)$ der Einzelmessung multipliziert und quadriert, erhält man nach Summation gemäß (1.8) die Standardabweichungen der Schätzwerte:

$$\Delta\hat{a} \equiv s(\hat{a}) = \frac{s(y_i)}{D}\sqrt{N^2[x^2] - N[x]^2} = s(y_i)\sqrt{\frac{N}{D}} \tag{1.11}$$

sowie

$$\Delta\hat{b} \equiv s(\hat{b}) = \frac{s(y_i)}{D}\sqrt{N[x^2]^2 - [x^2][x]^2} = s(y_i)\sqrt{\frac{[x^2]}{D}}. \tag{1.12}$$

Diese sollten bei der weiteren Auswertung von Versuchsergebnissen, die die Ergebnisse der Ausgleichsrechnung linearer Zusammenhänge verwenden, mitberücksichtigt werden.

Formal kann man durch beliebig gegebene „Punktwolken“ eine Ausgleichsgerade legen. Es ist daher sinnvoll, zumindest mit Hilfe einer graphischen Darstellung, den Grad des linearen Zusammenhangs zwischen den gegebenen Werten $x_1, x_2, \ldots, x_N$ und $y_1, y_2, \ldots, y_N$ zu beurteilen.

Vertauscht man die Koordinatenachsen und berechnet wieder die Ausgleichgerade, dann lässt sich ein Korrelationskoeffizient r_{xy} folgendermaßen definieren:

$$r_{xy} = \hat{a}\,\hat{a}'. \tag{1.13}$$

$\hat{a}'$ ist der Regressionskoeffizient der invertierten Geraden. Mit den schon oben berechneten Summen erhält für den Regressionskoeffizienten:

$$r_{xy} = \frac{N[yx]-[y][x]}{\sqrt{(N[y^2]-[y]^2)(N[x^2]-[x]^2)}}. \tag{1.14}$$

Sein Wertevorrat reicht von Null (keinerlei Korrelation) bis ± 1 (exakt linearer Zusammenhang).

Die gefundene Ausgleichsgerade verläuft übrigens durch den Schwerpunkt aller angegebenen Messpunkte, der folgendermaßen bestimmt werden kann:

$$P_S = (\bar{x}, \bar{y}) = \left(\frac{[x]}{N}, \frac{[y]}{N}\right). \tag{1.15}$$

Die Gleichungen (1.5) und (1.6) hätte man auch nach erfolgter Koordinatentransformation mit P_S als neuem Ursprung erhalten.

2. Gerade durch den Ursprung

Es kann vorkommen, dass Messwertpaare durch eine Ausgleichsgerade approximiert werden können, die genau durch den Koordinatenursprung geht. In einem solchen Fall vereinfacht sich die Summe der Abweichungsquadrate, weil nur der Regressionskoeffizient gesucht ist:

$$S = \frac{1}{\sigma^2}\sum_{i=1}^{N}(a x_i - y_i)^2. \tag{2.1}$$

Aus der Normalgleichung

$$\frac{\partial S}{\partial a} = \frac{2}{\sigma^2}\sum_{i=1}^{N}(a x_i - y_i)x_i = 0 \tag{2.2}$$

gewinnt man leicht den Schätzwert

$$\hat{a} = \frac{[yx]}{[x^2]}. \tag{2.3}$$

Zur Berechnung der Standardabweichung der Einzelmessung $s(y_i)$ geht man analog zu (1.7) vor und findet für die Gerade durch den Ursprung:

$$s(y_i) = \sqrt{\frac{\sum_{i=1}^{N}(\hat{a} x_i - y_i)^2}{N-1}}. \tag{2.4}$$

Ähnlich dem Vorgehen bei der allgemeinen Geraden muss für die Berechnung der Standardunsicherheit $s(\hat{a})$ die Gleichung (2.3) partiell nach jedem einzelnen y_i abgeleitet und dann nach (1.8) vorgegangen werden. Aus

$$\frac{\partial \hat{a}}{\partial y_i} = \frac{x_i}{[x^2]} \tag{2.5}$$

gewinnt man schließlich

$$\Delta\hat{a} \equiv s(\hat{a}) = \frac{s(y_i)}{[x^2]}\sqrt{[x^2]} = s(y_i)\sqrt{\frac{1}{[x^2]}}\,. \tag{2.6}$$

Für den Korrelationskoeffizienten nach (1.13) erhält man:

$$r_{xy} = \frac{[yx]}{\sqrt{[x^2][y^2]}}\,. \tag{2.7}$$

3. Senkrechte Abstände optimiert

Bei den vergangenen Betrachtungen war vorausgesetzt worden, dass die angegebenen Abszissenwerte „fehlerfrei“, d. h. ohne Abweichungen, gegeben sind, was experimentell nicht unbedingt der Realität entspricht. Es wird deshalb ein anderer Ansatz zum Auffinden des Regressionskoeffizienten gesucht, gleichzeitig aber unterstellt, dass die anzupassende Ausgleichsgerade wieder durch den Schwerpunkt P_S aller Punkte gemäß (1.15) verläuft. Die transformierte Geradengleichung lautet dann:

$$y = a(x - \overline{x}) + \overline{y}\,. \tag{3.1}$$

Weiterhin soll die Summe der Quadrate der senkrechten Abstände s_i bei Variation von a minimal werden (vgl. Abb. 2).

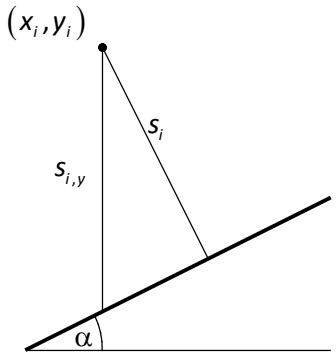

Abb. 2: Zur Definition des senkrechten Abstandes von der Geraden

Der senkrechte Abstand s_i des Punktes (x_i, y_i) auf die Ausgleichsgerade ergibt sich aus ihrem Steigungswinkel α:

$$s_i = s_{i,y}\cos\alpha = \frac{\left[y_i - y(x_i)\right]}{\pm\sqrt{1+\tan^2\alpha}} = \frac{\left[y_i - y(x_i)\right]}{\pm\sqrt{1+a^2}}\,. \tag{3.2}$$

die Summe der Abweichungsquadrate S wird dann mit der Varianz $\sigma^2 = \sigma_x^2 + \sigma_y^2$:

$$S = \frac{1}{\sigma^2}\sum_{i=1}^{N}\frac{\left[a(x_i - \overline{x}) - (y_i - \overline{y})\right]^2}{a^2+1}\,. \tag{3.3}$$

Die Normalgleichung zur Bestimmung von a findet man analog zu (1.2) und mit den Substitutionen $\tilde{x}_i = x_i - \overline{x}$ sowie $\tilde{y}_i = y_i - \overline{y}$:

$$\frac{\partial S}{\partial a} = \frac{-2a}{\left(a^2+1\right)^2}\sum_{i=1}^{N}\left(a\tilde{x}_i - \tilde{y}_i\right)^2 + \frac{2}{a^2+1}\sum_{i=1}^{N}\left(a\tilde{x}_i^2 - \tilde{y}_i\tilde{x}_i\right) = 0\,. \tag{3.4}$$

Hieraus folgt die quadratische Gleichung:

$$[\tilde{y}\tilde{x}]a^2 - \left([\tilde{y}^2] - [\tilde{x}^2]\right)a - [\tilde{y}\tilde{x}] = 0\,. \tag{3.5}$$

Für den Fall, dass die Summe $[\tilde{y}\,\tilde{x}]$ Null ist, wird der Schätzwert des Regressionskoeffizienten $\hat{a}$ auch zu Null gesetzt, ansonsten findet man

$$\hat{a} = \frac{1}{2}\left(P \pm \sqrt{P^2 + 4}\right), \tag{3.6}$$

wobei P mit

$$P = \frac{[\tilde{y}^2] - [\tilde{x}^2]}{[\tilde{y}\tilde{x}]} = \frac{N[y^2] - [y]^2 - N[x^2] + [x]^2}{N[yx] - [y][x]} \tag{3.7}$$

berechnet werden kann. Eine genauere Betrachtung zeigt, dass für das Vorzeichen vor der Wurzel in (3.6) das der Summe $[\tilde{y}\,\tilde{x}]$ gewählt werden muss. Das Absolutglied der Ausgleichsgeraden gewinnt man anschließend über die Rücktransformation:

$$\hat{b} = \overline{y} - \hat{a}\,\overline{x} = \frac{[y] - \hat{a}[x]}{N}\,. \tag{3.8}$$

Zur Berechnung der Standardunsicherheiten der Schätzwerte $\hat{a}$ und $\hat{b}$ wird zunächst die Standardabweichung $s(s_i)$ der Einzelmessung unter Berücksichtigung von (3.2) bestimmt:

$$s(s_i) = \sqrt{\frac{\sum_{i=1}^{N}\left(\hat{a}\,x_i + \hat{b} - y_i\right)^2}{\left(\hat{a}^2 + 1\right)(N-2)}}\,. \tag{3.9}$$

Hieraus lassen sich die Standardabweichungen der Messwerte in x-Richtung

$$s(x_i) = s(s_i)\frac{\hat{a}}{\sqrt{\hat{a}^2 + 1}} \tag{3.10}$$

und in y-Richtung

$$s(y_i) = s(s_i)\frac{1}{\sqrt{\hat{a}^2 + 1}} \tag{3.11}$$

angeben. Zur Berechnung der Standardunsicherheit $s(\hat{a})$ des Regressionskoeffizienten muss (3.6) partiell nach jedem x_i abgeleitet, mit $s(x_i)$ multipliziert und gemäß dem Gaußschen Gesetz der Fortpflanzung von Unsicherheiten quadratisch aufsummiert werden. Hinzu kommt dieselbe Verfahrensweise für die Abweichungen in y-Richtung. Die Ausdrücke für die partiellen Ableitungen sind nicht mehr trivial und lauten:

$$\begin{aligned}
\frac{\partial \hat{a}}{\partial x_i} &= \frac{1}{2}\left(1 \pm \frac{P}{\sqrt{P^2 + 4}}\right)\cdot\frac{\partial P}{\partial x_i} \\
\frac{\partial P}{\partial x_i} &= \frac{-2\left(N x_i - [x]\right)\left(N[yx] - [y][x]\right) - \left(N[y^2] - [y]^2 - N[x^2] + [x]^2\right)\left(N y_i - [y]\right)}{\left(N[yx] - [y][x]\right)^2}
\end{aligned} \tag{3.12}$$

sowie

$$\frac{\partial \hat{a}}{\partial y_i} = \frac{1}{2}\left(1 \pm \frac{P}{\sqrt{P^2+4}}\right) \cdot \frac{\partial P}{\partial y_i}$$
$$\frac{\partial P}{\partial y_i} = \frac{2\left(N y_i - [y]\right)\left(N[yx]-[y][x]\right) - \left(N[y^2]-[y]^2 - N[x^2]+[x]^2\right)\left(N x_i - [x]\right)}{\left(N[yx]-[y][x]\right)^2} \tag{3.13}$$

Für die weiteren Rechenschritte ist ein Mathematikprogramm hilfreich, das wenigstens keine Schusselfehler macht. Nach einigen Umformungen erhält man schließlich die Standardunsicherheit des Regressionskoeffizienten:

$$\Delta\hat{a} \equiv s(\hat{a}) = \frac{P \pm \sqrt{P^2+4}}{2\left(N[yx]-[y][x]\right)} \sqrt{N\left\{\left(N[y^2]-[y]^2\right)\left[s(x_i)\right]^2 + \left(N[x^2]-[x]^2\right)\left[s(y_i)\right]^2\right\}}$$
$$= \frac{\sqrt{N}\,\hat{a}}{N[yx]-[y][x]} \sqrt{\left(N[y^2]-[y]^2\right)\left[s(x_i)\right]^2 + \left[N[x^2]-[x]^2\right]\left[s(y_i)\right]^2} \tag{3.14}$$

Die Berechnung der Standardunsicherheit $s(\hat{b})$ basiert auf (3.8) unter Berücksichtigung von $\Delta\hat{a}$, man erhält:

$$\Delta\hat{b} \equiv s\left(\hat{b}\right) = \frac{1}{N}\sqrt{N\left\{\left[\hat{a}\,s(x_i)\right]^2 + s(y_i)^2\right\} + \left[[x]s(\hat{a})\right]^2}\ . \tag{3.15}$$

4. Bemerkungen

Die hier vorgestellten mathematischen Grundlagen der linearen Regression sind im Praktikumsprogram *„PhysPract“* eingearbeitet und können im Bearbeitungsfenster für die lineare Regression, angepasst an den Satz von Messwerten, ausgewählt werden.

Ilmenau, den 13.03.2020