

Multiple Lineare Regression

Marc Röttig
marc@roettig.org

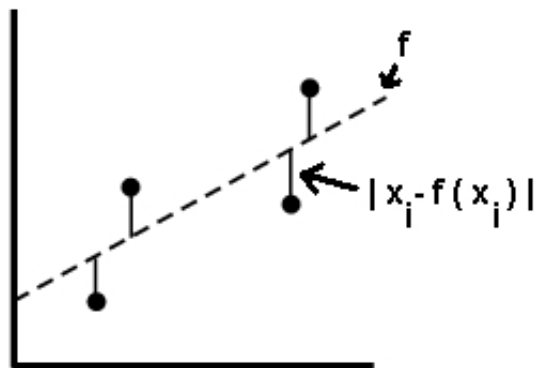
17. August 2005

Einleitung

Jeder experimentelle Meßvorgang bestimmt Daten unter bestimmten Kontrollbedingung, die variiert werden, um anhand der Meßdaten des Experiments, Aufschluß über das, dem Experiment zugrundeliegenden System zu erhalten. Experimentelle Meßdaten unterliegen in der Regel jedoch immer einem gewissen Grad an Rauschen, das bei jedem Meßvorgang unvermeidlich in die Meßwerte einfließen muß. Man bezeichnet die vom Experimentator variierten Parameter als die *ungebundenen Variablen* \mathbf{x}_i und die Meßdaten als die *gebundenen Variablen* \mathbf{y}_i . Um nun anhand der ermittelten Meßdaten ein mögliches Modell für das untersuchte System zu ermitteln, benötigt man zuerst eine Hypothese, welchem funktionellen Zusammenhang die Eingabe (*ungebundene Variablen*) an das System, und die darauf erfolgte Antwort (*gebundene Variable*) des Systems gehorchen. Mögliche funktionelle Zusammenhänge können beispielsweise linearer, polynomieller oder exponentieller Art sein. Hat man sich für einen funktionellen Zusammenhangsmodell entschieden, so gilt es dieses Modell, samt seiner noch zu bestimmenden Parameter, möglichst gut an die experimentellen Daten anzupassen. Das Rauschen der Meßergebnisse erschwert dieses Fitting jedoch ein wenig, da sich aufgrund der Meßfehler, die durch das Rauschen entstanden sind, die Modelle nicht exakt auf die Meßdaten fitten lassen. Man kann jedoch mittels des *Least-Squares*-Ansatzes ein Fitting berechnen, das den Fehler des Fittings an die Daten minimiert.

Lineare Least-Squares Methoden

Hat man sich nun für eine Funktion $f(\mathbf{x})$ entschieden, welche die gemessene Antwort \mathbf{y}_i des System auf Eingaben \mathbf{x}_i modellieren soll, so lautet die Forderung aller *Linear Least-Squares*-Methoden ¹, daß die Summe der Residualfehlerquadrate $\sum_{i=1}^n [y_i - f(x_i)]^2$ minimiert werden soll.



¹Linear meint hier die Linearität der Parameter, d.h. die ausgleichende Funktion hat die Form $\sum \alpha_i f_i(x)$, wobei die α_i lineare Skalare sind, also bspw. nicht die Form $a \cdot b$ annehmen können. Die Funktionen f_i können beliebige, auch nicht-lineare, Funktionen der unabhängigen Variablen sein (bspw. $\sin(x)$).

Einfacher eindimensionaler Least-Squares-Fit ($y = a_0 + a_1x$)

Im Fall der Wahl einer linearen Ausgleichsfunktion $\mathbf{f}(\mathbf{x}) = \mathbf{a}_0 + \mathbf{a}_1\mathbf{x}$ folgt für den summierten Residualquadratfehler \mathcal{E}

$$\mathcal{E} = \sum_{i=1}^n [y_i - f(x_i)]^2 = \sum_{i=1}^n [y_i - (a_0 + a_1x_i)]^2 = \text{Min} !$$

der, per Definition des Least-Squares Fit, zu minimieren ist.

Da Minimierungsproblemen stetiger Funktionen stets über die Bestimmung von Nullstellen der Ableitungen erfolgen kann, lautet der weitere Schritt für die Bestimmung der Parameter a_0, a_1 :

$$\begin{aligned} \frac{d\mathcal{E}}{da_0} &= -2 \sum_{i=1}^n [y_i - (a_0 + a_1x_i)] = 0 \\ \frac{d\mathcal{E}}{da_1} &= -2 \sum_{i=1}^n x_i [y_i - (a_0 + a_1x_i)] = 0 \end{aligned}$$

was sich umformen lässt zu

$$\begin{aligned} \sum_{i=1}^n y_i &= a_0 \sum_{i=1}^n 1 + a_1 \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i y_i &= a_0 \sum_{i=1}^n x_i + a_1 \sum_{i=1}^n x_i^2 \end{aligned}$$

was folgendem zweireihigem Gleichungssystem entspricht

$$\begin{bmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

In expliziter Form ergeben sich die Darstellung von a_0 und a_1 (z.B. schnell mittels Cramerscher Regel) zu

$$\begin{aligned} a_0 &= \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \\ a_1 &= \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \end{aligned}$$

Quadratischer eindimensionaler Least-Squares-Fit ($y = a_0 + a_1x + a_2x^2$)

Nimmt man als Ausgleichsfunktion $f(\mathbf{x}) = \mathbf{a}_0 + \mathbf{a}_1\mathbf{x} + \mathbf{a}_2\mathbf{x}^2$ an, so folgt damit für den summierten Residualquadratfehler \mathcal{E}

$$\mathcal{E} = \sum_{i=1}^n [y_i - f(x_i)]^2 = \sum_{i=1}^n [y_i - (a_0 + a_1x_i + a_2x_i^2)]^2 = \text{Min} !$$

den es zu minimieren gilt.

Um nun die Werte der Parameter a_0, a_1 und a_2 zu berechnen, bestimmt man wieder die Nullstellen der Ableitungen (\rightarrow Kettenregel) von \mathcal{E} nach den 3 Parametern

$$\begin{aligned} \frac{d\mathcal{E}}{da_0} &= -2 \sum_{i=1}^n [y_i - (a_0 + a_1x_i + a_2x_i^2)] = 0 \\ \frac{d\mathcal{E}}{da_1} &= -2 \sum_{i=1}^n x_i [y_i - (a_0 + a_1x_i + a_2x_i^2)] = 0 \\ \frac{d\mathcal{E}}{da_2} &= -2 \sum_{i=1}^n x_i^2 [y_i - (a_0 + a_1x_i + a_2x_i^2)] = 0 \end{aligned}$$

Nun lassen sich die 3 obigen Gleichungen umformen zu

$$\begin{aligned} \sum_{i=1}^n y_i &= a_0 \sum_{i=1}^n 1 + a_1 \sum_{i=1}^n x_i + a_2 \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i y_i &= a_0 \sum_{i=1}^n x_i + a_1 \sum_{i=1}^n x_i^2 + a_2 \sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n x_i^2 y_i &= a_0 \sum_{i=1}^n x_i^2 + a_1 \sum_{i=1}^n x_i^3 + a_2 \sum_{i=1}^n x_i^4 \end{aligned}$$

und somit in Matrixschreibweise zu folgendem Gleichungssystem

$$A \cdot \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \underbrace{\begin{bmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^4 \end{bmatrix}}_A \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n x_i^2 y_i \end{bmatrix}$$

das sich mittels Matrixinversion von A leicht lösen lässt². Die Lösung für die Parameter a_0, a_1, a_2 ergibt sich zu

$$A^{-1} \cdot \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n x_i^2 y_i \end{bmatrix} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}$$

² $A = X^T X \rightarrow$ siehe übernächster Abschnitt.

Polynomieller eindimensionaler Least-Squares-Fit

Analog zum obigen Least-Squares Fit mit ein Polynom 2. Grades ergibt sich die Lösung für den Fit mit einem Polynom m.-ten Grades $\mathbf{f}(\mathbf{x}) = \mathbf{a}_0 + \mathbf{a}_1\mathbf{x} + \dots + \mathbf{a}_{m-1}\mathbf{x}^{m-1} + \mathbf{a}_m\mathbf{x}^m$ durch Lösung des folgenden $(m + 1) \times (m + 1)$ -Gleichungssystems:

$$\begin{bmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n x_i & \dots & \sum_{i=1}^n x_i^m \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \dots & \sum_{i=1}^n x_i^{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_i^m & \sum_{i=1}^n x_i^{m+1} & \dots & \sum_{i=1}^n x_i^{2m} \end{bmatrix} \begin{bmatrix} a_0 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \\ \vdots \\ \sum_{i=1}^n x_i^m y_i \end{bmatrix}$$

Allgemeiner mehrdimensionaler Least-Squares-Fit ($f(\vec{x}) = \sum_{i=1}^m \alpha_i \phi_i$)

Entscheidet man sich, als Modellierungsfunktion eine Linearkombination von beliebigen mehrdimensionalen skalarwertigen Funktionen $\mathbf{f}(\vec{x}) = \sum_{i=0}^m \mathbf{a}_i \phi_i(\vec{x})$ zu verwenden (bspw. $f(\vec{x}) = a_2 \sin(x_1) + a_1 \cos(x_2) + a_0$), wobei $\vec{x}_i = (x_{i1}, \dots, x_{in})$ die i.-te Messung des Systems mit der zugehörigen Antwort y_i ist, so lässt sich die Lösung des Least-Squares-Fit mittels der Designmatrix X angeben :

$$X = \begin{bmatrix} \phi_0(\vec{x}_1) & \phi_1(\vec{x}_1) & \dots & \phi_m(\vec{x}_1) \\ \phi_0(\vec{x}_2) & \phi_1(\vec{x}_2) & \dots & \phi_m(\vec{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\vec{x}_n) & \phi_1(\vec{x}_n) & \dots & \phi_m(\vec{x}_n) \end{bmatrix}$$

Mithilfe dieser Designmatrix X lässt sich, klassisch nach Gauß, das Gleichungssystem formulieren, das es zu lösen gilt, um die Parameter des Modells zu bestimmen :

$$X^T X a = X^T y \iff \underbrace{\begin{bmatrix} \sum_{i=1}^n \phi_0(\vec{x}_i) \phi_0(\vec{x}_i) & \sum_{i=1}^n \phi_0(\vec{x}_i) \phi_1(\vec{x}_i) & \dots & \sum_{i=1}^n \phi_0(\vec{x}_i) \phi_m(\vec{x}_i) \\ \sum_{i=1}^n \phi_1(\vec{x}_i) \phi_0(\vec{x}_i) & \sum_{i=1}^n \phi_1(\vec{x}_i) \phi_1(\vec{x}_i) & \dots & \sum_{i=1}^n \phi_1(\vec{x}_i) \phi_m(\vec{x}_i) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n \phi_m(\vec{x}_i) \phi_0(\vec{x}_i) & \sum_{i=1}^n \phi_m(\vec{x}_i) \phi_1(\vec{x}_i) & \dots & \sum_{i=1}^n \phi_m(\vec{x}_i) \phi_m(\vec{x}_i) \end{bmatrix}}_{\mathbb{R}((m+1) \times (m+1))} \begin{bmatrix} a_0 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \phi_0(\vec{x}_i) \\ \sum_{i=1}^n y_i \phi_1(\vec{x}_i) \\ \vdots \\ \sum_{i=1}^n y_i \phi_m(\vec{x}_i) \end{bmatrix}$$

Dieses Gleichungssystem lässt sich durch Matrixinversion der $(m + 1) \times (m + 1)$ Matrix $X^T X$ leicht nach a auflösen.

Ist speziell $\mathbf{f}(\vec{x}) = \mathbf{a}_0 + \mathbf{a}_1 \mathbf{X}_1 + \dots + \mathbf{a}_m \mathbf{X}_m$ also mit $\phi_0 = 1$ und $\phi_i(x_1, \dots, x_n) = X_i = x_i$ so ist die Designmatrix X

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}$$

Analyse der Regressions-Ergebnisse

Bestimmtheitsmaß R^2

Die Lösung a des Least-Squares-Fit der linearen Funktion \mathbf{f} an die Daten war gegeben durch

$$a = (X^T X)^{-1} X^T y$$

die Prognosewerte \hat{y}_i ergeben sich nun zu

$$\hat{y} = X \cdot a$$

und die sogenannten Residuen e_i zu

$$e = y - \hat{y}$$

von y , die definiert ist als

Die Gesamtvariation der Regression ist dann gegeben durch

$$SS_T = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n e^2 \quad \text{mit} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

wobei \bar{y} gerade dem Mittelwert der y -Werte entspricht. Hieraus kann man den Schätzer für die Varianz der Grundgesamtheit zu

$$\sigma^2 \cong S_R^2 = \frac{SS_T}{n - m - 1} \quad \sigma \cong S_R = \sqrt{\frac{SS_T}{n - m - 1}}$$

bestimmen.

Betrachtet man nun die Gesamtvariation SS_T , so kann man folgende Zerlegung³ der Gesamtvarianz SS_T durchführen

$$\begin{aligned} SS_T &= \sum_{i=1}^n (y_i - \bar{y})^2 = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{\text{kausale Variation}} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{\text{nichtkausale Variation (zufällig)}} = SS_R + SS_E \\ &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n e_i^2 \end{aligned}$$

wobei SS_R der **Sum-of-Squares-Regression** ist, also die, durch die Regression (also das Modell) beschriebene bzw. erklärte Varianz. SS_E ist der **Sum-of-Squares-Error**, welcher als Restvariabilität betrachtet wird, die zufällig ist und durch das Modell nicht erklärt werden kann.

³Die Varianzzerlegung ist nur gültig falls die Modellebene nicht durch den Ursprung geht, also $a_0 \neq 0$ gilt.

Um nun die Güte des Modells, also die Güte der Anpassung an die Meßdaten, zu bestimmen kann man nun den Anteil der erklärten Variabilität SS_R an der Gesamtvariabilität SS_T der y-Werte betrachten und erhält das Bestimmtheitsmaß R^2 und den Korrelationskoeffizienten R

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \quad , \quad R = \sqrt{\frac{SS_R}{SS_T}} = \sqrt{1 - \frac{SS_E}{SS_T}}$$

Das Bestimmtheitsmaß R^2 ist ein Maß für die Güte des Fits der Regression an die Meßdaten. Liegt ein perfekter Fit an die Daten vor ist $SS_E = 0$ und somit $R^2 = 1$. Je besser die Güte der Anpassung, desto näher liegt R^2 bei 1, je schlechter die Güte der Anpassung desto näher liegt R^2 bei 0. Ein hoher Wert von R^2 sagt jedoch noch nichts darüber aus, ob die gewählten Variablen die Meßdaten sinnvoll erklären. So lassen sich beispielsweise mit hinreichend vielen Variablen beliebige Meßdaten perfekt modellieren, jedoch liegt dann aufgrund der vielen Variablen ein Overfitting vor. Man sagt es fehlt dann an Generalisierungsfähigkeit des Modells ⁴. Da R^2 mit steigender Variablenzahl m wächst, führt man folgende Normierung von R^2 durch und erhält R^2_{adj} .

$$R^2_{adj} = 1 - \frac{n-1}{n-m-1}(1-R^2) = 1 - \frac{\frac{SS_E}{n-m-1}}{\frac{SS_T}{n-1}} = 1 - \frac{MSS_E}{MSS_T}$$

um R^2 an die Anzahl der erklärenden Variablen anzupassen.

F-Test für die Signifikanz der Regression

Ein weitere Methode zur Bestimmung der Güte einer Regression ist der F-Test, der überprüft, ob die erklärenden Variablen X_i einen signifikanten Einfluß auf y_i haben. Wir benötigen folgende Werte der Regression

Variationsquelle	Freiheitsgrade		
Regression	m	SS_E	$MSS_E = \frac{SS_E}{m}$
Residualfehler	n-m-1	SS_R	$MSS_R = \frac{SS_R}{n-m-1}$
Gesamt	n-1	SS_T	

Der Wert für den F-Test ergibt sich dann zu $F^* = \frac{MSS_E}{MSS_R} = \frac{R^2}{1-R^2} \frac{n-m-1}{m}$, der mittels $\mathbf{F}^{-1}_{m,n-m-1}(1-\alpha)$ auf Signifikanz mit dem Konfidenzniveau α getestet wird ⁵. Es wird also getestet ob

$$H_0: a_0 = a_1 = \dots = a_m = 0 \quad H_A: \text{mind. ein } a_i \neq 0$$

$F^* \geq \mathbf{F}^{-1}_{m,n-m-1}(1-\alpha)$	verwerfe H_0 , Regressionskoeffizienten sind signifikant mit dem Konfidenzniveau α
$F^* < \mathbf{F}^{-1}_{m,n-m-1}(1-\alpha)$	bleibe bei H_0 , daß alle Koeffizienten $a_i = 0$

⁴Es ist daher gemäß Occam's Razor eine möglichst geringe Anzahl an erklärenden Variablen erwünscht.

⁵ $\mathbf{F} : \mathbb{R} \rightarrow [0, 1]$ (Verteilungsfunktion der Fisher F-Verteilung), $\mathbf{F}^{-1} : [0, 1] \rightarrow \mathbb{R}$

t-Test für die Signifikanz der Koeffizienten

Um die Signifikanz eines Koeffizienten a_i zu bestimmen, also die Wahrscheinlichkeit daß der geschätzte Koeffizienten a_i in Wahrheit 0 ist (also $\alpha_i = 0$), wendet man den t-Test an. Man entscheidet also zwischen

$$H_0: a_i = 0 \quad H_A: a_i \neq 0$$

Als Testwert berechnen wir $t^* = \frac{a_i - 0}{\hat{\sigma}_{X_i}} = \frac{a_i}{\hat{\sigma}_{X_i}}$ mit $\hat{\sigma}_{X_i} = \frac{S_R}{\sqrt{S_{X_i X_i}}}$,⁶ und entscheiden dann

$t^* \geq t_{1-\frac{\alpha}{2}; n-m-1}$	verwerfe H_0 , Regressionskoeffizient $a_i = 0$ ist signifikant von 0 verschieden mit dem Konfidenzniveau α (zweiseitig)
$t^* < t_{1-\frac{\alpha}{2}; n-m-1}$	bleibe bei H_0 , Regressionskoeffizient $a_i = 0$

Vertrauensintervalle für die Koeffizienten

Die Vertrauensintervalle für die Regressionskoeffizienten berechnen sich dann wie folgt

$$a_i \pm t_{1-\frac{\alpha}{2}; n-m-1} \cdot \hat{\sigma}_{X_i}$$

Beispiel

Gegeben sind folgende Daten aus *Correlation of Performance Test Scores with Tissue Concentration of Lysergic Acid Diethylamide in Human Subjects* (Wagner et al., Clinical Pharmacology and Therapeutics)

LSD Konzentration (X)	Punkte im Mathetest (Y)
1.17	78.93
2.97	58.20
3.26	67.47
4.69	37.47
5.83	45.65
6.00	32.92
6.41	29.97

die wir wie folgt fiten wollen : $Y = a_0 + a_1 X$. Die Designmatrix X sowie y ergeben sich dann zu

$$X = \begin{bmatrix} 1 & 1.17 \\ 1 & 2.97 \\ 1 & 3.26 \\ 1 & 4.69 \\ 1 & 5.83 \\ 1 & 6.00 \\ 1 & 6.41 \end{bmatrix} \quad y = \begin{bmatrix} 78.93 \\ 58.20 \\ 67.47 \\ 37.47 \\ 45.65 \\ 32.92 \\ 29.97 \end{bmatrix}$$

und die berechneten Ergebnis-Matrizen und Vektoren zu

$$X^T X = \begin{bmatrix} 7.000 & 30.330 \\ 30.330 & 153.891 \end{bmatrix} \quad X^T y = \begin{bmatrix} 350.610 \\ 1316.656 \end{bmatrix}$$

$$(X^T X)^{-1} = \begin{bmatrix} 0.978 & -0.193 \\ -0.193 & 0.044 \end{bmatrix} \quad a = \begin{bmatrix} 89.124 \\ -9.009 \end{bmatrix} = (X^T X)^{-1} \cdot X^T y$$

⁶ $S_{X_i X_i} = \sum_{i=1}^n (x_{ki} - \bar{x}_k)^2$ sowie $\hat{\sigma}_{X_i} = S_R \sqrt{(X^T X)^{-1}_{ii}}$

und das Perl Skript mlinreg.pl liefert das gleiche Ergebnis

Output of mlinreg.pl :

Coeff.	Estimate	Std.Err.	t	P(>t)
0	+89.124	7.048	+12.646	0.00005
1	-9.009	1.503	-5.994	0.00185

-- Sum of Squares --

SSR= 1824.30 SSE= 253.88 SST= 2078.18 S= 7.13
MSSR= 1824.30 MSSE= 50.78

-- Overall Significance --

R2= 0.878 R2adj= 0.853
F= 35.928 P(>F)= 0.00185