

Herleitung der Parameter-Gleichungen für die einfache lineare Regression

Uwe Ziegenhagen

16. März 2009

1 Einführung

(Wikipedia) Die Regressionsanalyse ist ein statistisches Analyseverfahren mit dem Ziel, Beziehungen zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen festzustellen.

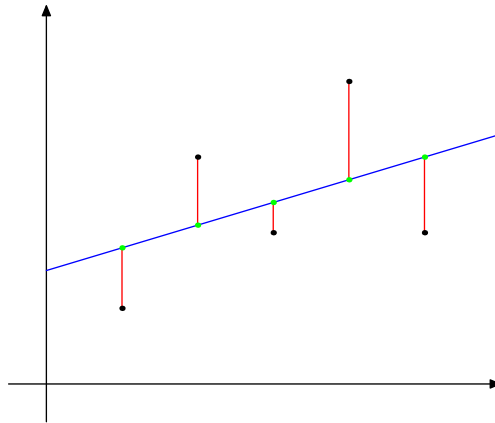
Allgemein wird eine metrische Variable Y betrachtet, die von einer zweiten Variablen x abhängt. Üblicherweise ist $x = (x_1, \dots, x_n)^T$ ein n -dimensionaler Vektor, wobei die einzelnen x -Werte untereinander unabhängig sind. Im eindimensionalen Fall spricht man von einer einfachen linearen Regressionsanalyse, in Dimensionen größer gleich zwei von einer multiplen Regressionsanalyse.

Bei der einfachen linearen Regression liegen Daten der Form $(x_i, y_i), i = 1, \dots, n$ vor. Als Modell wählt man

$$Y_i = b + ax_i + \epsilon_i \quad (1)$$

man nimmt somit einen linearen Zusammenhang zwischen x_i und Y_i an. Die Daten y_i werden als Realisierungen der Zufallsvariablen Y_i angesehen, die x_i sind nicht stochastisch, sondern Messstellen. Ziel der Regressionsanalyse ist in diesem Fall die Bestimmung der unbekannt Parameter a und b .

Die Vorgehensweise bei der linearen Regression veranschaulicht folgende Grafik. Gegeben sind Wertepaare x_i, y_i , als schwarze Punkte eingezeichnet. Grün sind die Werte (\hat{x}, \hat{y}) die durch die lineare Regressionsfunktion errechnet werden. Die roten Linien symbolisieren die Abweichungen $\hat{y}_i - y_i$ dieser durch die Gleichung bestimmten Punkte von den wahren Punkten. Aufgabe bei der Bestimmung der Parameter ist es nun, a und b so zu wählen, dass die Summe der quadrierten Abweichungen QS (also $\sum_{i=1}^n y_i - \hat{y}_i$) minimal wird.



2 Herleitung der Gleichungen

$$QS(a, b) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2)$$

$$= \sum_{i=1}^n (y_i - [ax_i + b])^2 \quad (3)$$

Da wir die optimalen Werte für die Minimierung dieser Quadratsumme erhalten wollen, bilden wir die partiellen Ableitungen nach a und b . Vorher können wir jedoch Gleichung 2 vereinfachen. Mit Hilfe der 2. Binomischen Formel¹ lösen wir 3 auf:

$$QS(a, b) = \sum_{i=1}^n (y_i^2 - 2y_i(ax_i + b) + (ax_i + b)^2) \quad (4)$$

Da der Term $(ax_i + b)^2$ der 1. Binomischen Formel² entspricht, lösen wir auch diesen auf und vereinfachen:

$$QS(a, b) = \sum_{i=1}^n (y_i^2 - 2ax_iy_i - 2by_i + a^2x_i^2 + 2abx_i + b^2) \quad (5)$$

Ausgehend von Gleichung 5 bilden wir jetzt die partiellen Ableitungen nach a und b :

$$\frac{\partial QS(a, b)}{\partial a} = \sum_{i=1}^n (-2x_iy_i + 2ax_i^2 + 2bx_i) \quad (6)$$

$$= 2 \sum_{i=1}^n x_i(-y_i + ax_i + b) \quad (7)$$

¹ 2. Binomische Formel: $(s - t)^2 = s^2 - 2st + t^2$

² 1. Binomische Formel: $(s + t)^2 = s^2 + 2st + t^2$

$$\frac{\partial \text{QS}(a, b)}{\partial b} = \sum_{i=1}^n (-2y_i + 2ax_i + 2b) \quad (8)$$

$$= 2 \sum_{i=1}^n (ax_i + b - y_i) \quad (9)$$

Wenn wir Gleichung 9 nullsetzen und auflösen, erhalten wir

$$2 \sum_{i=1}^n ax_i + 2 \sum_{i=1}^n b - 2 \sum_{i=1}^n y_i = 0 \quad (10)$$

$$2 \sum_{i=1}^n ax_i + 2nb - 2 \sum_{i=1}^n y_i = 0 \quad (11)$$

$$2nb = 2 \sum_{i=1}^n y_i - 2 \sum_{i=1}^n ax_i \quad (12)$$

Auflösen nach b (durch 2 und n teilen) gibt:

$$b = \frac{\sum_{i=1}^n y_i}{n} - \frac{\sum_{i=1}^n ax_i}{n} \quad (13)$$

$$= \frac{1}{n} \sum_{i=1}^n y_i - a \frac{1}{n} \sum_{i=1}^n x_i \quad (14)$$

$$= \bar{y} - a\bar{x} \quad (15)$$

Setzen wir nun $b = \bar{y} - a\bar{x}$ in Gleichung 7 ein, erhalten wir

$$2 \sum_{i=1}^n x_i (ax_i + (\bar{y} - a\bar{x}) - y_i) = 0 \quad (16)$$

Durch Ausmultiplizieren und Vereinfachen ergibt sich:

$$0 = \sum_{i=1}^n x_i (ax_i + (\bar{y} - a\bar{x}) - y_i) \quad (17)$$

$$= \sum_{i=1}^n ((ax_i^2 - ax_i\bar{x}) - x_i y_i + x_i \bar{y}) \quad (18)$$

$$\sum_{i=1}^n (ax_i^2 - ax_i\bar{x}) = \sum_{i=1}^n x_i y_i + \sum_{i=1}^n x_i \bar{y} \quad (19)$$

$$a \sum_{i=1}^n (x_i^2 - x_i\bar{x}) = \sum_{i=1}^n x_i y_i + \bar{y} \sum_{i=1}^n x_i \quad (20)$$

$$a = \frac{\sum_{i=1}^n x_i y_i + \bar{y} \sum_{i=1}^n x_i}{\sum_{i=1}^n (x_i^2 - x_i \bar{x})} = \frac{\sum_{i=1}^n x_i y_i + \bar{y} n \bar{x}}{\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i} \quad (21)$$

Über alternative Formeln zu Varianz und Kovarianz³ erhalten wir

$$a = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} = \frac{n \left(\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} \right)}{n \left(\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 \right)} = \frac{n \text{Cov}(x, y)}{n \text{Var}(x)} = \frac{\text{Cov}(x, y)}{\text{Var}(x)} \quad (22)$$

3 Beispiel




Für unser Beispiel vom Anfang hier die numerische Bestimmung der Parameter. Für \bar{x} erhalten wir 3, für \bar{y} 2.4, die Summe der $(x - \bar{x})(y - \bar{y})$ ergibt 3, die Summe der $(x - \bar{x})^2$ 10. Durch Einsetzen dieser Werte erhalten wir dann als Parameterwert für b 1.5, als Parameterwert für a 0.3, sodass die Formel unseres linearen Modells

$$y = 0.3 \cdot x + 1.5$$

lautet.

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|-----|-----|---------------|---------------|------------------------------|-------------------|
| | x | y | $x - \bar{x}$ | $y - \bar{y}$ | $(x - \bar{x})(y - \bar{y})$ | $(x - \bar{x})^2$ |
| 1 | 1 | 1 | -2 | -1.4 | 2.8 | 4 |
| 2 | 2 | 3 | -1 | 0.6 | -0.6 | 1 |
| 3 | 3 | 2 | 0 | -0.4 | 0.0 | 0 |
| 4 | 4 | 4 | 1 | 1.6 | 1.6 | 1 |
| 5 | 5 | 2 | 2 | -0.4 | -0.8 | 4 |

4 Quelldateien

- \LaTeX 
- Metapost 
- Metapost (kompiliert) 

³Verschiebungssatz:

$$\text{Cov}(X, Y) = E((X - E(X))(Y - E(Y))) = E(X, Y) - E(X)E(Y)$$

$$\text{Var}(X) = E((X - E(X))^2) = E(X^2) - (E(X))^2$$