Lineare Regression

Bernhard Möller

11. Juli 2013

Zusammenfassung

Messwerte innerhalb einer Messreihe sind stets mit Fehlern behaftet und lassen augenscheinlich nur den Schluss auf die Art des beobachteten Gesetzes zu, das durch die Messreihe beobachtet werden soll. Dadurch ergibt sich eine Streung der Messwerte um die erwarteten Werte. Häufig ist aber das dahinterstehende Gesetz nicht bekannt. In diesem Fall kann nur eine mathematische Schätzung auf Grundlage der Messdaten erfolgen. Im Folgenden soll die Methode der „Minimierung der Abweichungsquadrate“ am Beispiel eines linearen Zusammenhangs vorgestellt werden und der Zusammenhang der gesuchten Geradenparameter mit den ersten beiden Moduln einer statistischen Stichprobe, der Stichprobenvarianz (s_{xx}) und der Stichprobenkovarianz (s_{xy}), gefunden werden.

Die lineare Regression ist ein Spezialfall des allgemeinen Konzepts der Regressionsanalyse, mit der versucht wird, eine abhängige Variable durch eine oder mehrere unabhängige Variablen zu erklären − das Beiwort *linear* ergibt sich dabei daraus, dass die Regressionskoeffizienten (nicht unbedingt auch die Variablen selbst!) in diesem Fall in erster Potenz in das Regressionsmodell eingehen.

1 Einfache lineare Regression

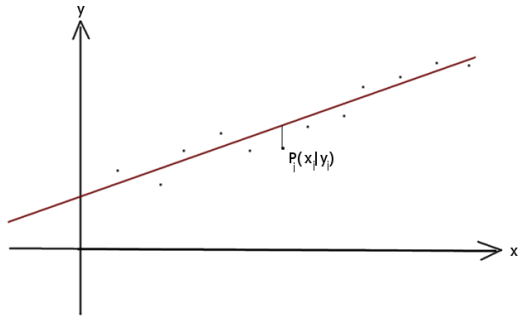
Ein Spezialfall von Regressionsmodellen sind lineare Modelle. Hierbei spricht man von der einfachen linearen Regression, und die Daten liegen in der Form $(y_i, x_i), i = 1, \ldots, n$ vor. Als Modell wählt man

$$Y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i \ ,$$

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, eine möglichst genaue Schätzung der unbekannten Parameter α_0 und α_1 zu finden. Diese seien mit a_0 und a_1 bezeichnet.
Die gesuchte Regressionsgerade hat also die Form $y(x) = a_0 + a_1 x$ zu besitzen.

1.1 Berechnung der Regressionsgeraden

Skizze:



Zur Berechnung der Parameter a_0 und a_1 verwenden wir die Methode der Minimierung der Abweichungsquadrate.
Die Messwerte y_i seien mit den Fehlern e_i behaftet. Ausgehend von der Annahme, es gäbe einen linearen, funktionalen Zusammenhang zwischen y_i und x_i, erhalten wir die Gleichung

$$y_i = a_0 + a_1 x_i + e_i \;\Leftrightarrow\; e_i = y_i - a_1 x_i - a_0.$$

Wir definieren die Funktion

$$\begin{aligned} f : \mathbb{R} \times \mathbb{R} &\longrightarrow \mathbb{R}_+ \\ (a_0, a_1) &\longmapsto f(a_0, a_1) := \sum_{i=1}^{n} (y_i - a_1 x_i - a_0)^2 \end{aligned}$$

f hängt von den Parametern a_0 und a_1 ab, da die Messwerte (x_i, y_i) fix sind und eine Minimierung der Abweichungsquadrate nur über die Variation der Geradenparameter erreicht werden kann.
Wir suchen jetzt das Minimum von f. Das notwendige Kriterium für ein Extremum einer Funktion diesen Typs lautet $\nabla f(a_0, a_1) = 0$. Damit der Gradient von f verschwindet, müssen alle partiellen Ableitungen verschwinden. Das heißt, wir erhalten folgende Gleichungen:

$$\frac{\partial f(a_0, a_1)}{\partial a_0} = 0 \;\Rightarrow\; -2 \sum_{i=1}^{n} (y_i - a_1 x_i - a_0) = 0 \tag{1}$$

$$\frac{\partial f(a_0, a_1)}{\partial a_1} = 0 \;\Rightarrow\; -2 \sum_{i=1}^{n} x_i (y_i - a_1 x_i - a_0) = 0 \tag{2}$$

Diese verwenden wir jetzt, um a_0 und a_1 zu bestimmen und beginnen mit Gleichung (1):

$$\begin{aligned} 0 &= -2\sum_{i=1}^{n}(y_i - a_1 x_i - a_0) \\ \Leftrightarrow\ 0 &= \sum_{i=1}^{n} y_i - a_1 \sum_{i=1}^{n} x_i - n a_0 \\ \Leftrightarrow\ a_0 &= \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{a_1}{n}\sum_{i=1}^{n} x_i \end{aligned}$$

Nun sind

$$\frac{1}{n}\sum_{i=1}^{n} y_i =: \bar{y} \quad \text{und} \quad \frac{1}{n}\sum_{i=1}^{n} x_i =: \bar{x}.$$

Daraus folgt

$$a_0 = \bar{y} - a_1 \bar{x}. \tag{3}$$

Wir setzen (3) jetzt in Gleichung (2), um a_1 zu berechnen:

$$\begin{aligned} 0 &= -2\sum_{i=1}^{n} x_i(y_i - a_1 x_i - a_0) \\ \Rightarrow 0 &= \sum_{i=1}^{n} x_i(y_i - a_1 x_i - \bar{y} + a_1 \bar{x}) \\ \Leftrightarrow 0 &= \sum_{i=1}^{n} x_i(y_i - \bar{y}) - a_1 \sum_{i=1}^{n} x_i(x_i - \bar{x}) \\ \Leftrightarrow a_1 &= \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})} \end{aligned} \tag{4}$$

Wir brauchen (4) jetzt nur noch in (3) einzusetzen und erhalten

$$a_0 = \bar{y} - \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})}\,\bar{x} \tag{5}$$

Die Regressionsgerade r ist also gegeben durch

$$\boxed{r(x) = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})}\,(x - \bar{x}) + \bar{y}}$$

1.2 Zusammehang mit s_{xx} und s_{xy}

Wir wollen nun die Gleichung der Regressionsgeraden r mithilfe der Stichprobenvarianz und -kovarianz ausdrücken.
Die Stichprobenkovarianz ist definiert durch

$$s_{xy} := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

und die Stichprobenvarianz durch

$$s_{xx} := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Nun gilt

$$\begin{aligned} s_{xy} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\Big(\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + \bar{x}\bar{y}\sum_{i=1}^{n} 1\Big) \\ &= \frac{1}{n}\Big(\sum_{i=1}^{n} x_i y_i - 2n\bar{x}\bar{y} + n\bar{x}\bar{y}\Big) = \frac{1}{n}\Big(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\Big) \end{aligned}$$

Analog zeigt man

$$s_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\Big(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\Big)$$

Dies sind die alternativen Darstellungen von Kovarianz und Varianz.
Andererseits zeigt man

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = n s_{xy}$$

und

$$\sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = n s_{xx}$$

Dies brauchen wir jetzt nur noch in die Gleichung der Regressionsgeraden einsetzen und erhalten

$$\boxed{r(x) = \frac{s_{xy}}{s_{xx}}(x - \bar{x}) + \bar{y}}$$