



Vorlesung
Wahrscheinlichkeit und Regression
WS 2005/06

Kapitel 7
Einfache lineare Regression

Prof. Dr. Rolf Steyer

**Lehrstuhl für Methodenlehre
und Evaluationsforschung**



Das Potenzgesetz von Stevens I

- (a) Wie hängt die *Länge der Urteilslinie* von der Länge der Reizlinie ab?
- (b) Wie hängt das *Ausmaß der Urteilsfehler* von der Länge der Reizlinie ab?

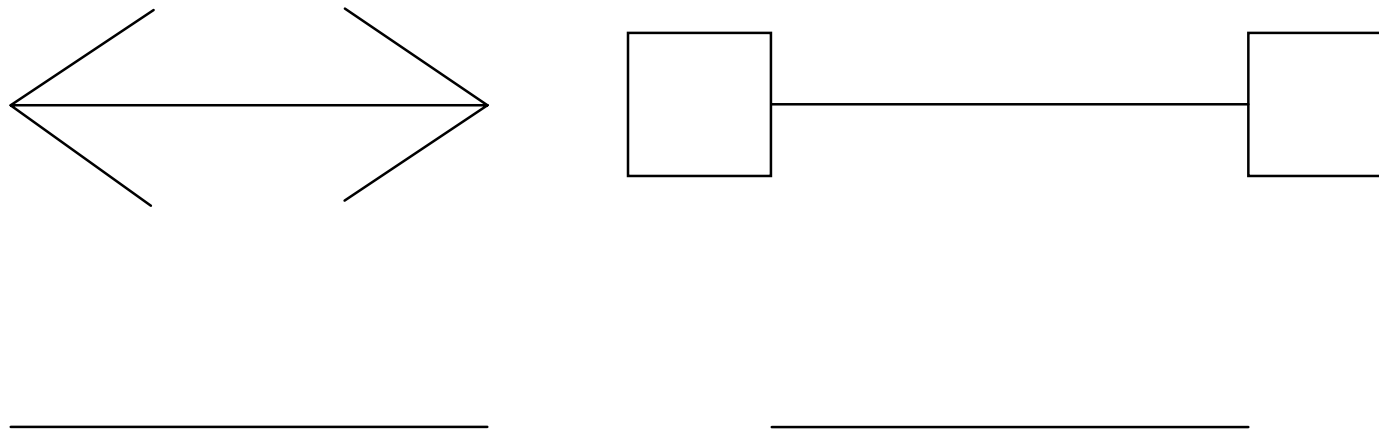


Abbildung 8.1. Müller-Lyer-Figur (links) und Baldwin-Figur (rechts). Unterhalb der Figuren sind zum direkten Vergleich jeweils Referenzlinien gleicher Länge ohne die entsprechenden Kontextreize abgebildet.



Einfache lineare Regression: Definition

Definition 1. Seien X und Y numerische Zufallsvariablen auf einem gemeinsamen Wahrscheinlichkeitsraum, beide mit positiver und endlicher Varianz. Die Regression $E(Y | X)$ des Regressanden Y auf den Regressor X heißt *linear in X* , wenn

$$E(Y | X) = \alpha_0 + \alpha_1 \cdot X,$$

wobei α_0 und α_1 reelle Zahlen sind. Der Regressand Y heißt dann auch *linear regressiv abhängig* vom Regressor X .



Regressive Unabhängigkeit

Lineare Regression

$$E(Y | X) = \alpha_0 + \alpha_1 \cdot X.$$

Den Fall $\alpha_1 = 0$ betrachten wir als Spezialfall der reg-linearen Abhängigkeit, bei dem wir auch von *regressiver Unabhängigkeit* sprechen. Es gilt dann nämlich:

$$E(Y | X) = E(Y) = \alpha_0.$$



Einfache lineare Regression: Residuum

Die Variable Y wird bei einer linearen Regression als Summe einer linearen Funktion der Variablen X und der Residualvariablen $\varepsilon = Y - E(Y | X)$ dargestellt, wie man aus der Umstellung der Gleichung

$$E(Y | X) = \alpha_0 + \alpha_1 \cdot X,$$

ersehen kann:

$$Y = \alpha_0 + \alpha_1 \cdot X + \varepsilon.$$



Lineare Regression: Folgerungen

Wichtige Folgerungen sind:

$$\alpha_0 = E(Y) - \alpha_1 \cdot E(X),$$

und, falls die Varianz $Var(X)$ größer Null ist,

$$\alpha_1 = Cov(X, Y) / Var(X).$$

Außerdem

$$E(\varepsilon \mid X = x) = 0.$$

Eine weitere Folgerung ist

$$E(Y \mid X = x) = \alpha_0 + \alpha_1 \cdot x.$$

Für das Residuum $\varepsilon := Y - E(Y \mid X)$ gelten natürlich alle im letzten Kapitel ausführlich besprochenen Eigenschaften.



Lineare Regression: Determinationskoeffizient

Im Falle einer linearen Regression lässt sich der Determinationskoeffizient durch die Gleichungen

$$R_{Y|X}^2 = \alpha_1^2 \frac{\text{Var}(X)}{\text{Var}(Y)} = \frac{\text{Cov}(X, Y)^2}{\text{Var}(X) \text{Var}(Y)} = \text{Kor}(X, Y)^2$$

berechnen.



Parametrisierung einer Regression I

Eine Regression $E(Y | X)$ kann auf verschiedene Arten *parametrisiert*, d. h. durch verschiedene Gleichungen dargestellt werden, in denen verschiedene Parameter vorkommen. Ein einfaches Beispiel ist, dass man den Regressor X durch den Regressor $-X$ ersetzt. In diesem Fall würde man anstelle des Steigungskoeffizienten α_1 den Steigungskoeffizienten $-\alpha_1$ erhalten. Der entscheidende Sachverhalt ist dabei, dass sich die Regression $E(Y | X)$ selbst nicht ändert, obwohl die Gleichung mit den darin vorkommenden Parametern anders aussieht. Die Regression $E(Y | X)$ verändert sich durch die Auswechslung von X durch $-X$ deswegen nicht, weil sie definitionsemäßig eine Zufallsvariable ist, deren Werte die bedingten Erwartungswerte $E(Y | X = x)$ sind. Diese sind aber identisch, egal, ob ich X oder $-X$ als Regressor betrachte, d. h. es gilt: $E(Y | X = x) = E(Y | X = -x)$ für alle Werte x von X . Auch wenn sich die Gleichungen, und die darin vorkommenden Parameter ändern, gilt also: $E(Y | X) = E(Y | -X)$.



Parametrisierung einer Regression II

Das Entsprechende gilt selbst für einen Übergang von einem Regressor X zu $\ln X$ [$\ln$ steht für den „natürlichen Logarithmus“ (s. Regelbox 1)]. In diesem Fall ändert sich im allgemeinen sogar der Gleichungstyp, denn wenn die Regression $E(Y | X)$ eine lineare Funktion von X ist, so kann sie im allgemeinen nicht zugleich eine lineare Funktion von $\ln X$ sein und umgekehrt, wenn $E(Y | \ln X)$ eine lineare Funktion von $\ln X$ ist, so kann sie im allgemeinen nicht zugleich eine lineare Funktion von X sein. Dennoch bleibt die Regression identisch, d. h. es gilt: $E(Y | X) = E(Y | \ln X)$. Was sich ändert ist lediglich die *Parametrisierung* der Regression, d. h. ihre Darstellung durch eine Gleichung bestimmten Typs (z. B. linear oder nichtlinear) oder durch bestimmte Parameter in verschiedenen Gleichungen des selben Typs. Ein anderes Beispiel werden wir im nächsten Abschnitt über dichotome Regressoren kennen lernen.



Dichotomer Regressor I

Im allgemeinen muss die Regression $E(Y | X)$ nicht unbedingt eine lineare Funktion von X sein. Nimmt der Regressor X jedoch nur zwei verschiedene reelle Werte an, z. B. $X = 1$ für die Experimentalbedingung und $X = 0$ für die Kontrollbedingung, so gilt Gleichung $E(Y | X) = \alpha_0 + \alpha_1 \cdot X$ immer. Man spricht in diesem Zusammenhang auch von einem *saturierten Modell*. Für die beiden Regressionskoeffizienten gelten im dichotomen Fall mit der Kodierung der Werte von X mit 0 und 1 außer den allgemeinen Gleichungen auch:

$$\alpha_0 = E(Y | X = 0)$$

und

$$\alpha_1 = E(Y | X = 1) - E(Y | X = 0).$$



Dichotomer Regressor II

Man beachte jedoch, dass diese Interpretation der beiden Regressionskoeffizienten von der Kodierung der Werte von X mit 0 und 1 abhängt. Kodiert man stattdessen die beiden Bedingungen z. B. mit -1 und $+1$, so kommt man zu einer anderen Interpretation der beiden Regressionskoeffizienten:

$$\alpha_0 = [E(Y \mid X = 1) + E(Y \mid X = -1)] / 2$$

und

$$\alpha_1 = [E(Y \mid X = 1) - E(Y \mid X = -1)] / 2.$$