



Vorlesung
Wahrscheinlichkeit und Regression
WS 2005/06

Kapitel 6
Regression

Prof. Dr. Rolf Steyer

**Lehrstuhl für Methodenlehre
und Evaluationsforschung**



Bedingter Erwartungswert

Definition 1.

Ist Y eine reellwertige Zufallsvariable mit endlich vielen Werten $y_1, \dots, y_n$ und ist $P(X = x) > 0$, dann ist der *bedingte Erwartungswert* von Y gegeben $X = x$ die mit den bedingten Wahrscheinlichkeiten $P(Y = y_i \mid X = x)$ gewichtete Summe ihrer Werte:

$$E(Y \mid X = x) := \sum_{i=1}^n y_i \cdot P(Y = y_i \mid X = x).$$

Kann Y nur die beiden Werte 1 und 0 annehmen, dann folgt aus obiger Gleichung:

$$\begin{aligned} E(Y \mid X = x) &= 1 \cdot P(Y = 1 \mid X = x) + 0 \cdot P(Y = 0 \mid X = x) \\ &= P(Y = 1 \mid X = x). \end{aligned}$$



Bedingter Erwartungswert: Rechenregeln

B. Rechenregeln

Sind X und Y sowie Y_1 und Y_2 numerische Zufallsvariablen auf $\langle \Omega, \mathcal{F}, P \rangle$ mit endlichen Erwartungswerten sowie α und β reelle Zahlen, dann gelten:

$$E(\alpha \mid X = x) = \alpha$$

$$E(\alpha \cdot Y \mid X = x) = \alpha \cdot E(Y \mid X = x)$$

$$E(\alpha \cdot Y_1 + \beta \cdot Y_2 \mid X = x) = \alpha \cdot E(Y_1 \mid X = x) + \beta \cdot E(Y_2 \mid X = x)$$

$$E(Y \mid X = x) = \sum_z E(Y \mid X = x, Z = z) \cdot P(Z = z \mid X = x)$$



Regression

Definition 2.

Die Regression $E(Y | X)$ kann als diejenige Funktion von X definiert werden, deren Werte die *bedingten Erwartungswerte* $E(Y | X = x)$ von Y gegeben $X = x$ sind. Dabei kann der Regressor X durchaus beliebige Werte annehmen, die nicht unbedingt Zahlen sein müssen.



Regression: Rechenregeln

Box 6.2. Rechenregeln für Regressionen

$$E(\alpha | X) = \alpha, \quad \alpha \in \mathbb{R}$$

$$E(\alpha Y | X) = \alpha E(Y | X), \quad \alpha \in \mathbb{R}$$

$$E(\alpha Y_1 + \beta Y_2 | X) = \alpha E(Y_1 | X) + \beta E(Y_2 | X), \quad \alpha, \beta \in \mathbb{R}$$

$$E[E(Y | X)] = E(Y)$$

$$E[f(X) | X] = f(X), \quad \text{falls } f(X) \text{ numerisch ist}$$

$$E[E(Y | X) | f(X)] = E[Y | f(X)]$$

$$E[f(X) \cdot Y | X] = f(X) \cdot E(Y | X), \quad \text{falls } f(X) \text{ numerisch ist}$$



Residuum

Definition 3.

Das *Residuum* ε bezüglich einer Regression $E(Y | X)$ ist definiert als *Abweichung der Zufallsvariablen Y von ihrer Regression $E(Y | X)$ auf X .*

In Formeln: $\varepsilon := Y - E(Y | X)$.

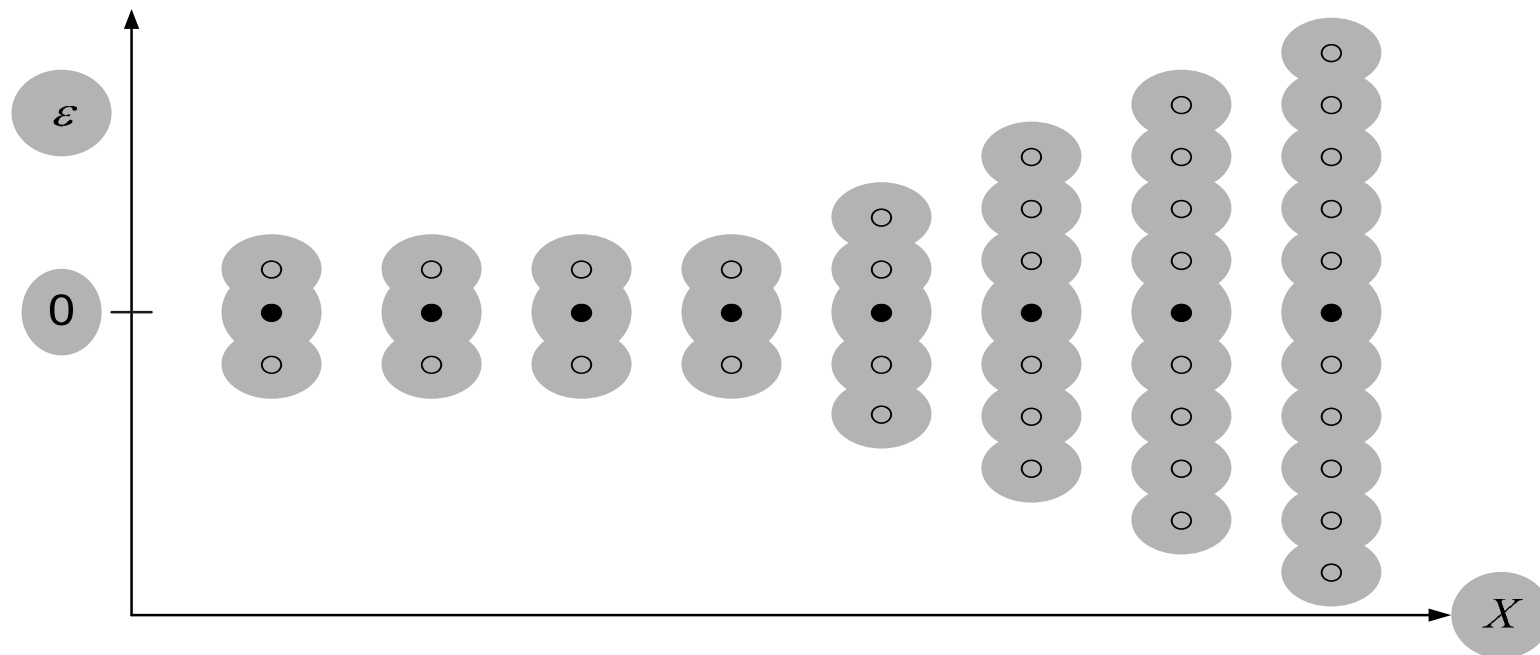


Abbildung 1: Die Regressionen von Y und ε auf einen numerischen Regressor X . In diesem Beispiel sind die bedingten Varianzen des Residuums von X abhängig. Das Zeichen $\circ$ markiert die Werte der Regression $E(\varepsilon | X)$ und die Werte des Residuums ε .



Eigenschaften des Residuums

Box 6.3. Die wichtigsten Eigenschaften des Residuums

- (i) $Y = E(Y | X) + \varepsilon$
- (ii) $E(\varepsilon) = 0$
- (iii) $E(\varepsilon | X) = 0$
- (iv) $E[\varepsilon | f(X)] = 0$
- (v) $E[\varepsilon | E(Y | X)] = 0$
- (vi) $Cov(\varepsilon, X) = 0$, falls X numerisch ist
- (vii) $Cov(\varepsilon, X_i) = 0$, $i = 1, \dots, m$, falls $X = (X_1, \dots, X_m)$ numerisch ist
- (viii) $Cov[\varepsilon, f(X)] = 0$, falls $f(X)$ numerisch ist
- (ix) $Cov[\varepsilon, E(Y | X)] = 0$
- (x) $Var(Y) = Var[E(Y | X)] + Var(\varepsilon)$



Determinationskoeffizient

Ein Begriff, der unmittelbar auf den oben behandelten Eigenschaften des Residuums ε [insb. auf Regel (x)] basiert, ist der des Determinationskoeffizienten, der für jede numerische Zufallsvariable Y mit endlichem Erwartungs- wert $E(Y)$ und endlicher Varianz $Var(Y)$ definiert ist, und zwar durch:

$$R_{Y|X}^2 = \frac{Var[E(Y|X)]}{Var(Y)}, \quad \text{falls } Var(Y) > 0$$



Determinationskoeffizient (fortgesetzt)

Der Determinationskoeffizient lässt sich als der durch X determinierte Varianzanteil von Y interpretieren. Wie man sehen kann, addiert er sich mit dem Residualvarianzanteil von Y zu 1 auf, falls $Var(Y) > 0$:

$$1 = \frac{Var(Y)}{Var(Y)} = \frac{Var[E(Y|X)]}{Var(Y)} + \frac{Var(\varepsilon)}{Var(Y)}$$



Multiple Korrelation

Die Wurzel aus dem Determinationskoeffizienten heißt *multiple Korrelation*. Für diese gilt auch:

$$R_{Y|X} = \text{Kor}[Y, E(Y|X)]$$



Der allgemeine Regressionsbegriff

Definition 4.

Seien $\langle \Omega, \mathcal{F}, P \rangle$ ein Wahrscheinlichkeitsraum, Y eine numerische Zufallsvariable auf $\langle \Omega, \mathcal{F}, P \rangle$ mit endlichem Erwartungswert [d. h. $-\infty < E(Y) < +\infty$?] und X eine diskrete Zufallsvariable auf $\langle \Omega, \mathcal{F}, P \rangle$ mit den Werten $x_1, \dots, x_n$, für die jeweils $P(X = x_i) > 0$ gilt.

Dann heißt die Zufallsvariable

$$E(Y | X) := \sum_{i=1}^n E(Y | X = x_i) \cdot I_{X=x_i}$$

Regression von Y auf X oder *bedingte Erwartung* von Y unter X .