Bedingte lineare Regression

Worum geht es in diesem Modul?

- Beispiel: Das Verhältnismodell für geometrisch-optische Täuschungen I
- Bedingte lineare Regression
- Die bedingten Regressionen
- Eigenschaften des Residuums
- Einfache Spezialfälle
- Parametrisierungen der bedingten linearen Regression
- Dichotome Regressoren
- Einfache und bedingte lineare Regression
- Beispiel: Das Verhältnismodell für geometrisch-optische Täuschungen II

Beispiel: Baldwin-Figuren

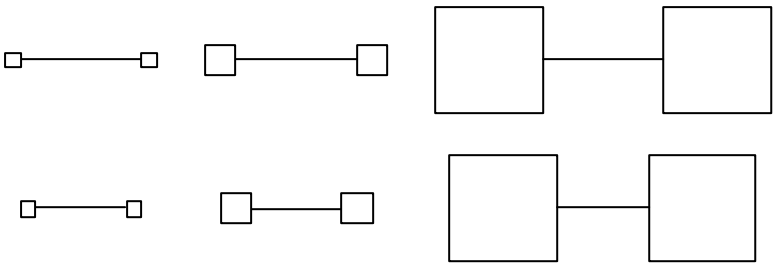

Abbildung 10.1. Sechs Baldwin-Figuren, die sich aus der Kombination von Linien zweier verschiedener Längen und Quadraten dreier verschiedener Größen ergeben.

Beispiel. Das Verhältnismodell für geometrisch-optische Täuschungen I

Tabelle 10.1. Kontext-Serienreiz-Kombinationen und die dabei vorkommenden konstanten Kontext-Serienreiz-Verhältnisse.

		Kontextreiz			
		1	2	4	8
Serienreiz	1	1/1	2/1		
	2	1/2	1/1	2/1	
	4		1/2	1/1	2/1
	8			1/2	1/1

Anmerkungen. Nur diejenigen Kontext-Serienreiz-Verhältnisse sind aufgeführt, in denen mindestens drei verschiedene Serienreize vorkommen. Nur bei diesen Verhältnissen kann die im Text formulierte Hypothese falsch sein.

Beispiel. Das Verhältnismodell für geometrisch-optische Täuschungen I

Bei der empirischen Überprüfung erwies sich das folgende Modell als das beste:

$$E(ln\ Y \mid lnX, Z) = g_0(Z) + g_1(Z) \cdot ln\ X.$$

Somit sollte das (stochastische) Potenzgesetz (in seiner logarithmierten Form) gelten, d.h.:

$$E_{Z=z}\ (ln\ Y \mid lnX) = g_0(z) + g_1(z) \cdot ln\ X.$$

Dabei sind $g_0(z)$ und $g_1(z)$ reelle Zahlen, die je nach Kontext-Serienreiz-Verhältnis z verschieden groß sein können.

Bedingte lineare Regression: Definition

Definition 10.1. Seien X und Y numerische Zufallsvariablen mit endlichen Erwartungswerten und Varianzen und Z eine Zufallsvariable, alle auf einem gemeinsamen Wahrscheinlichkeitsraum. Dann heißen die Regression $E(Y \mid X, Z)$ bzgl. Z *bedingt linear in* X und Y *von* X *bzgl.* Z *bedingt linear regressiv abhängig*, wenn zwei (beliebige) numerische Funktionen $g_0(Z)$ und $g_1(Z)$ von Z existieren, für die gilt:

$$E(Y \mid X, Z) = g_0(Z) + g_1(Z) \cdot X.$$

Im Fall $g_1(Z) = \gamma_0$, $\gamma_0 \in \mathbb{R}$, heißt Y *von* X *bzgl.* Z *partiell linear regressiv abhängig.*

Bedingte lineare Regression: Abbildung I

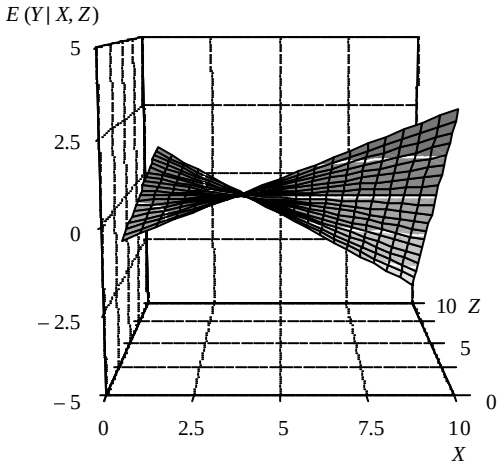

Abbildung 10.2. Darstellung einer bedingten linearen Regression mit

$g_0(Z) = -0.5 + 0.4 \cdot Z$ und $g_1(Z) = 0.15 - 0.1 \cdot Z$.

Die bedingten Regressionen

Der Schlüssel zum Verständnis der bedingten linearen regressiven Abhängigkeit liegt wieder in der Betrachtung der bedingten Regressionen von Y auf X bei jeweils gegebenem Wert z der Variablen Z. Für einen beliebigen festen Wert z von Z folgt nämlich:

$$E_{Z=z}(Y|X) = g_0(z) + g_1(z) \cdot X,$$

falls $E(Y|X) = g_0(Z) + g_1(Z) \cdot X$ gilt.

Die bedingten Regressionen: Abbildung 2

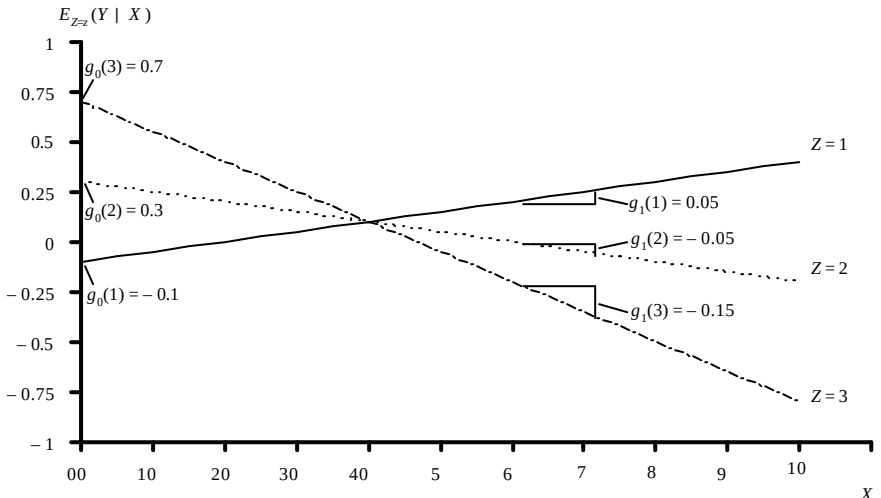

Abbildung 10.3. Graphen der bedingten linearen Regressionen von Y auf X für einige Werte z von Z.

Eigenschaften des Residuums

Für das Residuum:

$$e := Y - E(Y \mid X, Z)$$

gelten natürlich alle im Kapitel 6 behandelten Eigenschaften, insbesondere diejenigen, die schon im letzten Kapitel behandelt wurden:

$$E(e \mid X, Z) = E(e \mid X) = E(e \mid Z) = 0,$$

$$E[e \mid g_0(Z)] = E[e \mid g_1(Z)] = 0$$

$$E(e) = 0,$$

und

$$Cov(e, X) = Cov(e, Z) = Cov[e, g_0(Z)] = Cov[e, g_1(Z)] = 0.$$

Einfache Spezialfälle I

Ist die Regression $E(Y \mid X, Z)$ von Y auf X und Z mit der Regression $E(Y \mid X)$ von Y auf X identisch, d. h. gilt

$$E(Y \mid X, Z) = E(Y \mid X),$$

so sprechen wir von *bedingter regressiver Unabhängigkeit* des Regressanden Y von Z, gegeben X.

Die obige Definition der bedingten linearen regressiven Abhängigkeit schließt auch denjenigen Spezialfall ein, in dem $g_0(Z)$ und $g_1(Z)$ konstante Funktionen von Z sind, also für alle Werte z von Z den gleichen Wert $g_0(Z) = \beta_0$ bzw. $g_1(Z) = \gamma_0$ annehmen. In diesem Fall gilt

$$E(Y \mid X, Z) = \beta_0 + \gamma_0 X = E(Y \mid X).$$

Einfache Spezialfälle II

Die multiple lineare Regression mit zwei Regressoren ist ein weiterer Spezialfall von $E(Y|X) = g_0(Z) + g_1(Z) \cdot X$, wobei

$$g_0(Z) = \beta_0 + \beta_1 Z, \quad \beta_0, \beta_1 \in \mathbb{R},$$

$$g_1(Z) = \gamma_0,$$

da dies folgende Gleichung impliziert:

$$E(Y|X, Z) = \beta_0 + \beta_1 Z + \gamma_0 X.$$

Einfache Spezialfälle III

Gilt sowohl $E(Y|X) = g_0(Z) + g_1(Z) \cdot X$ und ist

$$g_0(Z) = \beta_0 + \beta_1 Z, \quad \beta_0, \beta_1 \in \mathbb{R},$$

$$g_1(Z) = \gamma_0 + \gamma_1 Z, \quad \gamma_0, \gamma_1 \in \mathbb{R},$$

dann folgt

$$\begin{aligned} E(Y|X, Z) &= (\beta_0 + \beta_1 Z) + (\gamma_0 + \gamma_1 Z) \cdot X \\ &= \beta_0 + \beta_1 Z + \gamma_0 X + \gamma_1 XZ. \end{aligned}$$

Die Funktion $g_1(Z) = \gamma_0 + \gamma_1 Z$ heißt *lineare Modifikatorfunktion*.

Parametrisierungen: Allgemein

Im Allgemeinen spielt die Anzahl der Werte von Z keine Rolle, gleichgültig, ob Z nun ein- oder mehrdimensional ist. Zur Analyse einer bedingten linearen Regression mit verfügbaren PC-Programmen zur multiplen linearen Regression allerdings müssen die Funktionen $g_0(Z)$ und $g_1(Z)$ als lineare Funktionen vom Typ

$$g_0(Z) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_{k-1} Z_{k-1}$$
$$g_1(Z) = \gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2 + ... + \gamma_{k-1} Z_{k-1}$$

dargestellt werden, wobei jede der Variablen Z_1, Z_2, ..., Z_{k-1} eine (zunächst beliebige) Funktion von Z ist.

Parametrisierung als Polynome

Kann Z nur k verschiedene Zahlen als Werte annehmen, so lassen sich sowohl die Funktion $g_0(Z)$ als auch die Modifikatorfunktion $g_1(Z)$ immer als ein Polynom (k–1)-ten Grades darstellen:

$$g_0(Z) = \beta_0 + \beta_1 Z + \beta_2 Z^2 + ... + \beta_{k-1} Z^{k-1}$$

und

$$g_1(Z) = \gamma_0 + \gamma_1 Z + \gamma_2 Z^2 + ... + \gamma_{k-1} Z^{k-1} .$$

Die Koeffizienten β_i und γ_i, $i = 1, ..., k-1$, sind dabei reelle Zahlen. Wir nennen dies die *polynomiale Parametrisierung* der bedingten linearen Regression.

Parametrisierung durch Indikatorvariablen

Kann Z nur k verschiedene Werte annehmen, so lässt sich sowohl die Funktion

$$g_0(Z) = \beta_0 + \beta_1 I_1 + ... + \beta_j I_j + ... + \beta_{k-1} I_{k-1}$$

als auch die Funktion

$$g_1(Z) = \gamma_0 + \gamma_1 I_1 + ... + \gamma_j I_j + ... + \gamma_{k-1} I_{k-1}$$

als eine gewichtete Summe von Indikatorvariablen I_j darstellen, wobei jede Indikatorvariable I_j den Wert 1 annimmt, falls der j-te Wert von Z vorliegt und andernfalls den Wert 0.

Dichotome Regressoren I

Nimmt X bspw. nur die Werte 0 und 1 an, so sind die bedingten Regressionskoeffizienten $g_1(z)$ als bedingte wahre Mittelwertsunterschiede zwischen den beiden durch X repräsentierten Gruppen bei gegebenem $Z = z$ zu interpretieren. In Formeln:

$$g_1(z) = E_{Z=z}(Y \mid X = 1) - E_{Z=z}(Y \mid X = 0)$$

oder auch

$$g_1(z) = E(Y \mid X = 1, Z = z) - E(Y \mid X = 0, Z = z).$$

Beide Schreibweisen sind äquivalent.

Dichotome Regressoren II

Nehmen beide Regressoren X und Z jeweils nur zwei verschiedene reelle Werte an, so ist die folgende Gleichung allgemeingültig:

$$\begin{aligned} E(Y \mid X, Z) &= (\beta_0 + \beta_1 Z) + (\gamma_0 + \gamma_1 Z) \cdot X \\ &= \beta_0 + \beta_1 Z + \gamma_0 X + \gamma_1 Z \cdot X. \end{aligned}$$

Dies ist also ein saturierte Parametrisierung. Setzt man die Werte von X und Z in obige Gleichung ein folgt das lineare Gleichungssystem:

$$\begin{aligned} E(Y \mid X = 1, Z = 1) &= \beta_0 + \beta_1 + \gamma_0 + \gamma_1 \\ E(Y \mid X = 1, Z = 0) &= \beta_0 + \gamma_0 \\ E(Y \mid X = 0, Z = 1) &= \beta_0 + \beta_1 \\ E(Y \mid X = 0, Z = 0) &= \beta_0. \end{aligned}$$

Dichotome Regressoren III

Die Lösung dieses linearen Gleichungssystem lautet:

$$\begin{aligned} \gamma_0 &= E(Y \mid X = 1, Z = 0) - E(Y \mid X = 0, Z = 0) \\ \beta_1 &= E(Y \mid X = 0, Z = 1) - E(Y \mid X = 0, Z = 0) \\ \gamma_1 &= [E(Y \mid X = 1, Z = 1) - E(Y \mid X = 0, Z = 1)] \\ &\quad - [E(Y \mid X = 1, Z = 0) - E(Y \mid X = 0, Z = 0)] \end{aligned}$$

Demnach können die Regressionkoeffizienten wie folgt interpretiert werden:

$$\begin{aligned} g_0(0) &= E(Y \mid X = 0, Z = 0) = \beta_0 \\ g_1(0) &= E(Y \mid X = 1, Z = 0) - E(Y \mid X = 0, Z = 0) = \gamma_0 \\ g_0(1) &= E(Y \mid X = 0, Z = 1) = \beta_0 + \beta_1 \\ g_1(1) &= E(Y \mid X = 1, Z = 1) - E(Y \mid X = 0, Z = 1) = \gamma_0 + \gamma_1. \end{aligned}$$

Dichotome Regressoren: Abbildung

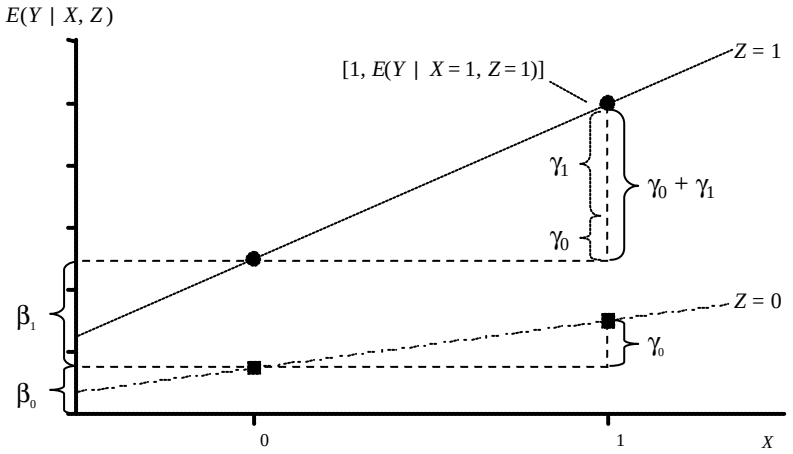

Abbildung 10.4. Bedingte lineare Regressionen bei dichotomen Regressoren mit Werten 0 und 1.

Einfache und bedingte lineare Regression

Wenn Y von X *bedingt* linear regressiv abhängig ist gegeben Z, folgt dann auch, dass Y von X (einfach) linear regressiv abhängig ist? Die Antwort lautet: im Allgemeinen nicht.

Aber falls

$$E[g_0(Z)|X] = E[g_0(Z)] \quad \text{und} \quad E[g_1(Z)|X] = E[g_1(Z)],$$

dann

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

wobei

$$\alpha_0 = E[g_0(Z)] \quad \text{und} \quad \alpha_1 = E[g_1(Z)].$$

Beispiel: Das Verhältnismodell für geometrisch-optische Täuschungen II

Für die drei Kontext-Serienreiz-Verhältnisse seien zunächst die folgenden beiden Indikatorvariablen definiert:

$$I_{1/2} := \begin{cases} 1, & \text{falls das Kontext-Serienreiz-Verhältnis 1/2 vorliegt} \\ 0, & \text{andernfalls,} \end{cases}$$

$$I_{1/1} := \begin{cases} 1, & \text{falls das Kontex-Serienreiz-Verhältnis 1/1 vorliegt} \\ 0, & \text{andernfalls,} \end{cases}$$

Die beiden Funktionen

$$g_0(Z) := \beta_0 + \beta_1 I_{1/2} + \beta_2 I_{1/1}$$

und

$$g_1(Z) := \gamma_0 + \gamma_1 I_{1/2} + \gamma_2 I_{1/1}$$

können für jedes der drei Kontext-Serienreiz-Verhältnisse andere Werte annehmen.

Beispiel: Das Verhältnismodell für geometrisch-optische Täuschungen II

Die zu konstruierende Gleichung für die Regression $E(ln\ Y \mid ln\ X, Z)$ ist dann also:

$$E(ln\ Y \mid ln\ X, Z) = \beta_0 + \beta_1 I_{1/2} + \beta_2 I_{1/1} + (\gamma_0 + \gamma_1 I_{1/2} + \gamma_2 I_{1/1})\ ln\ X\,,$$

$$= \beta_0 + \beta_1 I_{1/2} + \beta_2 I_{1/1} + \gamma_0\ ln\ X + \gamma_1\ I_{1/2}\ ln\ X + \gamma_2\ I_{1/1}\ ln\ X.$$

Ein interessanter Aspekt bei diesem Modell ist auch der folgende

$$Std_{Z=z}(Y \mid X) = exp[g_0(z)] + X^{\,g_1(z)} \cdot Std(d \mid Z = z)$$

Determinationskoeffizient

Der Determinationskoeffizient berechnet sich wie folgt:

$$R^2_{Y|X,Z} = \frac{Var[g_0(Z)] + Var[g_1(Z) \cdot X] + 2Cov[g_0(Z), g_1(Z) \cdot X]}{Var(Y)}$$

Kennwert für die Stärke der bedingten linearen regressiven Abhängigkeit des Regressanden Y von X oder der durch X zusätzlich zu Z erklärte Varianzanteil von Y:

$$R^2_{Y|X,Z} - R^2_{Y|Z}$$