Zweifache lineare Regression

- Beispiel: Intelligenz, Bleibelastung und beruflicher Status
- Multiple lineare Regression mit zwei Regressoren
- Die bedingten Regressionen
- Eigenschaften des Residuums
- Identifikation der Regressionskoeffizienten
- Dichotome Regressoren
- Einfache und zweifache Regression und multiple lineare Regression mit zwei Regressoren
- Lineare Quasi-Regression

Beispiel: Intelligenz, Bleibelastung und beruflicher Status

In diesem Einführungsbeispiel befassen wir uns mit der Frage, ob eine erhöhte Bleibelastung der Umwelt zu einer Verminderung der Intelligenzleistungen bei Kindern führt.

In der Stichprobe fand sich zwischen dem *Logarithmierten Bleigehalt* (X) und dem *Verbalen Intelligenzquotienten* (Y) eine negative Korrelation von –.14. Wenn wir davon ausgehen, dass die Abhängigkeit des Regressanden Y vom Regressor X *linear* regressiv ist, dann lässt sich diese Abhängigkeit durch die Gleichung

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

beschreiben, wobei $\alpha_1 < 0$. Der Verbale IQ ist dann also negativ linear regressiv abhängig vom Bleigehalt der Zähne.

Beispiel: Intelligenz, Bleibelastung und beruflicher Status

3

Neben den beiden Variablen X und Y wurde aber auch der *Berufliche Status der Eltern* der Kinder erfragt, der mit Z bezeichnet sei. Es stellt sich nun die Frage, ob bei gegebenem Wert z von Z (also bei festem beruflichen Status) noch eine lineare regressive Abhängigkeit der Variablen Y von X besteht. Die Daten der Stolberg-Studie legen nun nahe, dass Y von X bezüglich Z partiell linear regressiv *unabhängig* ist, d. h., dass die Gleichung

$$E(Y \mid X, Z) = \beta_0 + \beta_1 X + \beta_2 Z$$

gilt, wobei der Koeffizient

$$\beta_1 = 0.$$

Zweifache lineare Regression: Definition

4

Definition 9.1

Seien X, Y und Z jeweils eindimensionale numerische Zufallsvariablen mit endlichen Erwartungswerten und Varianzen auf dem gleichen Wahrscheinlichkeitsraum. Dann heißt die Regression $E(Y \mid X, Z)$ *linear in* (X, Z), wenn gilt:

$$E(Y \mid X, Z) = \beta_0 + \beta_1 X + \beta_2 Z, \quad \beta_0, \beta_1, \beta_2 \in \mathbb{R}.$$

Zweifache lineare Regression : Abbildung I

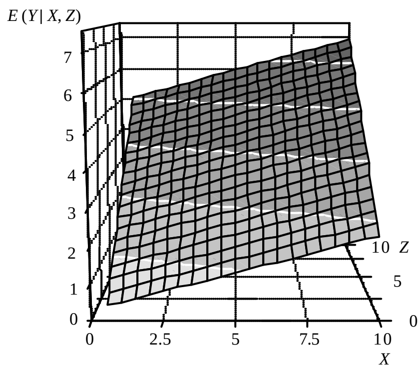

Abbildung 9.1. Regressionsebene der Regression $E(Y|X, Z) = 0.3 + 0.2 \cdot X + 0.5 \cdot Z$. Jeder Punkt auf der Ebene ist (bei kontinuierlichen Regressoren X und Z) ein Wert $E(Y|X = x, Z = z)$ der Regression $E(Y|X, Z)$.

$$E_{Z=z}(Y \mid X) = (\beta_0 + \beta_2 z) + \beta_1 X,$$

Die bedingten Regressionen

$$E_{Z=z}(Y \mid X) = (\beta_0 + \beta_2 z) + \beta_1 X,$$

Zweifache lineare Regression: Abbildung II

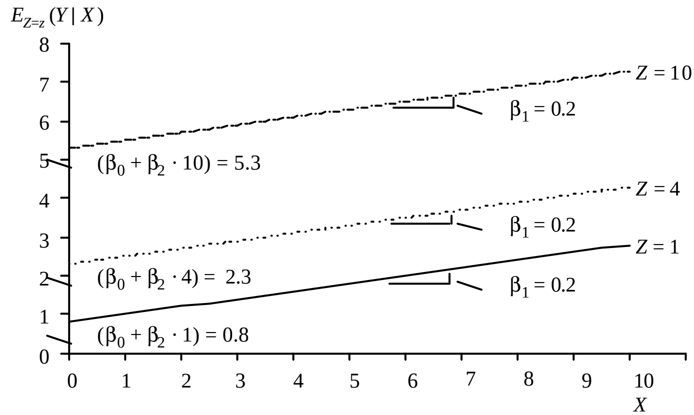

Abbildung 9.2. Die bedingten Regressionsgeraden bei bzgl. Z partieller linearer regressiver Abhängigkeit des Regressanden Y vom Regressor X. Dabei wird wie in Abbildung 9.1 die Regression $E(Y \mid X, Z) = 0.3 + 0.2 \cdot X + 0.5 \cdot Z$ zugrunde gelegt.

Determinationskoeffizient

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

$$R^2_{Y \mid X,\, Z} = \frac{\beta_1^2 Var(X) + \beta_2^2 Var(Z) + 2\beta_1\beta_2\, Cov(X, Z)}{Var(Y)}$$

Das Residuum und seine Eigenschaften

Für das Residuum

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

gelten, neben den bereit behandelten allgemeinen Eigenschaften, wie z.B.

$$E(e \mid X, Z) = 0 \text{ und } E(e) = 0,$$

insbesondere

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

und

$$Cov(e, X) = Cov(e, Z) = 0.$$

Identifikation

$$\beta_0 = E(Y) - \beta_1 E(X) - \beta_2 E(Z)$$

$$\beta_1 = \frac{Var(Z)\,Cov(X,Y) - Cov(X,Z)\,Cov(Y,Z)}{Var(X)\,Var(Z) - Cov(X,Z)^2}$$

$$= \frac{Std(Y)}{Std(X)} \cdot \frac{Kor(X,Y) - Kor(X,Z)\,Kor(Y,Z)}{1 - Kor(X,Z)^2}$$

$$\beta_2 = \frac{Var(X)\,Cov(Z,Y) - Cov(X,Z)\,Cov(Y,X)}{Var(X)\,Var(Z) - Cov(X,Z)^2}$$

$$= \frac{Std(Y)}{Std(Z)} \cdot \frac{Kor(Z,Y) - Kor(X,Z)\,Kor(X,Y)}{1 - Kor(X,Z)^2}$$

Identifikationen: Dichotome Regressoren

Im speziellen Fall, in dem X und Z dichotom (zweiwertig) sind, lassen sich die Koeffizienten β_0, β_1 und β_2 noch einfacher berechnen. In diesem Fall kann man die vier Gleichungen

$$E(Y \mid X = 1, Z = 1) \;\; = \beta_0 + \beta_1 + \beta_2$$

$$E(Y \mid X = 1, Z = 0) \;\; = \beta_0 + \beta_1$$

$$E(Y \mid X = 0, Z = 1) \;\; = \beta_0 + \beta_2$$

$$E(Y \mid X = 0, Z = 0) \;\; = \beta_0$$

für die vier bedingten Erwartungswerte $E(Y|X = x, Z = z)$ ableiten, und diese dann nach den unbekannten Koeffizienten β_0, β_1 und β_2 auflösen:

$$\beta_1 = E(Y \mid X = 1, Z = 0) - E(Y|X = 0, Z = 0)$$

und

$$\beta_2 = E(Y \mid X = 0, Z = 1) - E(Y \mid X = 0, Z = 0)$$

Einfache und zweifache Regression: I

Theorem 9.1. Unter den gleichen Voraussetzungen wie in Definition 9.1 gilt folgendes: Ist $E(Y \mid X, Z) = \beta_0 + \beta_1 X + \beta_2 Z$ erfüllt und ist $\beta_2 = 0$ oder ist Z von X regressiv unabhängig,

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

dann folgt

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

wobei $\alpha_0 = \beta_0 + \beta_2 E(Z)$ und $\alpha_1 = \beta_1$ und weiter

$$Cov(X, Z) = 0,$$

$$Var[E(Y|X, Z)] = \beta_1^2 \, Var(X) + \beta_2^2 \, Var(Z).$$

Einfache und zweifache Regression: II

Gilt außerdem noch

$$E(X|\,Z) = E(X),$$

dann folgen auch

$$Var[E(Y|\,X, Z)] = Var[E(Y\,|\,X)] + Var[E(Y\,|\,Z)],$$

$$R^2_{Y|X,Z} = R^2_{Y|X} + R^2_{Y|Z}\,,$$

$$R^2_{Y|X} = Kor(X,Y)^2$$

$$R^2_{Y|Z} = Kor(Y,Z)^2$$

Einfache und zweifache Regression: III

Kann man nicht voraussetzen, dass Z von X regressiv unabhängig ist, und gilt stattdessen

$$E(Z|\,X) = \gamma_0 + \gamma_1 X,$$

so folgt dennoch

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

wobei $\alpha_0 := \beta_0 + \beta_2\gamma_0$ und $\alpha_1 := \beta_1 + \beta_2\gamma_1$, also

$$E(Y|\,X) = (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1)\,X.$$

Multiple lineare Quasi-Regression: I

Definition 9.2. Unter den gleichen Voraussetzungen wie in Definition 9.1 handelt es sich bei der *zweifachen linearen Quasi-Regression*, die wir mit $Q(Y \mid X, Z)$ bezeichnen, um diejenige Linearkombination $\beta_0 + \beta_1 X + \beta_2 Z$ von X und Z, die folgendes erfüllt:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + n\,,$$

$$E(n) = 0\,,$$

und

$$Cov(n, X) = Cov(n, Z) = 0.$$

Multiple lineare Quasi-Regression: II

Definition 9.3. Unter den gleichen Voraussetzungen wie in Definition 9.1 können wir $Q(Y|X, Z)$ auch als diejenige Linearkombination von X und Z definieren, welche die folgende Funktion der reellen Zahlen b_0, b_1 und b_2, das *Kleinst-Quadrat-Kriterium* minimiert:

$$LS(b_0, b_1, b_2) = E[[Y - (b_0 + b_1 X + b_2 Z)]^2].$$

Diejenigen Zahlen b_0, b_1 und b_2, für welche die Funktion $LS(b_0, b_1, b_2)$ ein Minimum annimmt, seien mit β_0, β_1 und β_2 respektive, bezeichnet. Die *zweifache lineare Quasi-Regression* ist dann definiert durch::

$$Q(Y \mid X, Z) := \beta_0 + \beta_1 X + \beta_2 Z$$