

Vorlesung 'Deskriptive Statistik'

11. und 12. Sitzung

Dr. Uwe Altmann

Institut für Psychologie
Friedrich-Schiller-Universität Jena

Jena, 07.01.2013

1 Beschreibung von Abhängigkeiten mit Regressionsmodellen

- Allgemeines
- Begriffliches zum Regressionsmodell
- Bedeutung der Modellparameter
- Prinzip der kleinsten Quadrate
- Eigenschaften der Regressionsgeraden
- Modellfit

Zusammenhänge geben nicht über die „Wirkrichtung“ Auskunft, im Sinne von Variable X beeinflusst Variable Y oder umgekehrt.

Regressionen beschreiben hingegen **Abhängigkeiten** zwischen zwei oder mehr (!) Variablen.

Anhand der **zeitlichen Ordnung** der Variablen oder anhand theoretischer Erwägungen wird hierbei die **Wirkrichtung festgelegt**, in dem einige Variablen als **abhängige Variablen** (auch Regressanden, Response oder Outcome) klassifiziert werden und andere Variablen als **unabhängige Variablen** (auch Regressoren oder Einflußgrößen).

Beispiel: Die Messung der Intensität von Ängsten **nach** einer Therapie könnte man als **abhängige** Variable und die Intensität von Ängsten **vor** der Therapie als **unabhängige** Variable auffassen.

Eine **Regression** ist eine **Funktion** der unabhängigen (Zufalls-) **Variablen**, z. B. eine Funktion der Variablen X :

$$f(X) \tag{1}$$

Im Idealfall ist $Y = f(X)$, d. h. bei Kenntnis von X können wir die Ausprägung von Y exakt angeben. Wegen Messfehlern, nicht beachteten Variablen etc. trifft $Y = f(X)$ so gut wie nie zu.

Die Funktion/Regression kann wie folgt spezifiziert werden, z. B.

$$f(X) = b_0 + b_1 \cdot X \quad \text{oder} \tag{2}$$

$$f(X) = b_0 + b_1 \cdot X + b_2 \cdot X^2 \tag{3}$$

Ein solches **Regressionsmodell** ist eine **vereinfachte Abbildung der Wirklichkeit**.

1 Beschreibung von Abhängigkeiten mit Regressionsmodellen

- Allgemeines
- Begriffliches zum Regressionsmodell
- Bedeutung der Modellparameter
- Prinzip der kleinsten Quadrate
- Eigenschaften der Regressionsgeraden
- Modellfit

Anstoß vieler Diskussionen ist das jeweils verwendete Regressionsmodell. Sind alle Umstände, welche die Ausprägungen der abhängigen Variablen (hier Y) erklären können, im Modell enthalten? Fehlen Variablen (z. B. Z)? Ist X^2 zu modellieren? Etc.

Aus forschungspraktischen Gründen ist dies oft nicht möglich. Die Psychologie verwendet verschiedene Strategien, um dieses Defizit auszugleichen.

Am bekanntesten ist die **Randomisierung**. Dabei werden die Probanden den Untersuchungsbedingungen/Gruppen **zufällig** zugewiesen. Patienten der **Kontrollgruppe** bekommen z. B. ein Placebo und Patienten der **Experimentalgruppe** ein Medikament. Durch die Randomisierung wird erreicht, dass eventuell wirksame Drittvariablen in beiden Gruppen gleich verteilt und deshalb nicht mehr im Modell zu berücksichtigen sind.

Ein anderer Ansatz besteht in der Verwendung von **Adjustierungsverfahren**. Eine **Randomisierung** ist hier **nicht notwendig**. Die Outcomemittelwerte der Kontroll- und Experimentalgruppe werden vor einem Vergleich über die Verteilung der eventuell verfälschenden Drittvariablen adjustiert.

Der Einfachheit halber beschäftigen wir uns mit Regressionen, die

- eine abhängige und eine unabhängige Variable aufweisen und
- linear sind (d. h. keine quadratischen Terme etc. enthalten).

Da wir uns im Bereich der deskriptiven Statistik befinden, betrachten wir **Messwerte**, z. B. die **der metrischen Variablen X und Y** : x_1, x_2, \dots, x_N und y_1, y_2, \dots, y_N .

Das Regressionsmodell lautet:

$$\begin{array}{rclcl} \text{Messwert} & = & \text{Wert der Regression} & + & \text{Residuum} \\ y_i & = & f(x_i) & + & e_i \\ y_i & = & b_0 + b_1 \cdot x_i & + & e_i \end{array}$$

$b_0 + b_1 \cdot x_i$ ist der **Anteil am Messwert** von Y , **der** durch die unabhängigen Variablen/Regression **erklärt werden kann**.

e_i ist der „**unerklärte**“ **Rest** (auch Residuum, Fehler oder Error) bei Person Nr. i . Residuen (e_1, e_2, \dots, e_N) kann man auch als Realisierungen einer Residualvariablen (ε) auffassen.

Das Regressionsmodell lautet:

$$y_i = b_0 + b_1 \cdot x_i + e_i \quad (4)$$

y_i ist der Messwert der Variablen Y der Person Nr. i und x_i der Messwert der Variablen X der Person Nr. i . Y ist die abhängige und X die unabhängige Variable. Beide Variablen seien metrisch.

b_0 und b_1 nennt man (Regressions-) **Koeffizienten** (auch Modellparameter). Es sind feste Werte, die auf Grundlage der Daten zu schätzen sind.

b_0 bezeichnet man als **Achsenabschnittskoeffizient** (auch intercept).

b_1 ist der **Steigungskoeffizient** (auch slope) für X . Der **Betrag** von b_1 gibt die **Stärke** der Abhängigkeit und das **Vorzeichen** von b_1 die **Art und Weise** der Abhängigkeit an.

1 Beschreibung von Abhängigkeiten mit Regressionsmodellen

- Allgemeines
- Begriffliches zum Regressionsmodell
- Bedeutung der Modellparameter
- Prinzip der kleinsten Quadrate
- Eigenschaften der Regressionsgeraden
- Modellfit

Das Regressionsmodell lautet:

$$\begin{aligned}y_i &= b_0 + b_1 \cdot x_i + e_i \\y_i &= \underbrace{f(x_i)}_{\text{Regression}} + e_i\end{aligned}$$

Was bedeuten die Parameter b_0, b_1 der Regression

$$f(x_i) = b_0 + b_1 \cdot x_i \quad ? \quad (5)$$

Tipp: I. d. R. sind es Werte der Regression unter der Bedingung, dass die unabhängigen Variablen den Wert 0 oder 1 annehmen.

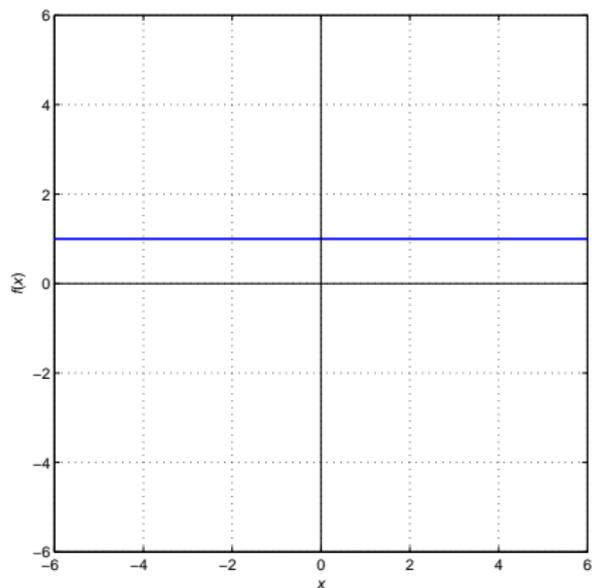
Im obigen Modell ist b_0 der Wert der Regression, falls $x_i = 0$ ist.

$$f(0) = b_0 + b_1 \cdot 0 = b_0 \quad (6)$$

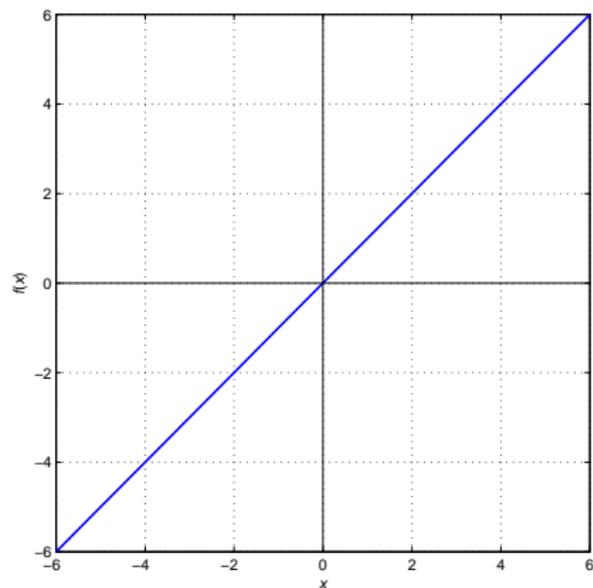
b_1 ist die Differenz der Regressionswerte der Bedingungen $x_i = 1$ und $x_i = 0$

$$f(1) - f(0) = [b_0 + b_1 \cdot 1] - [b_0 + b_1 \cdot 0] = b_1 \quad (7)$$

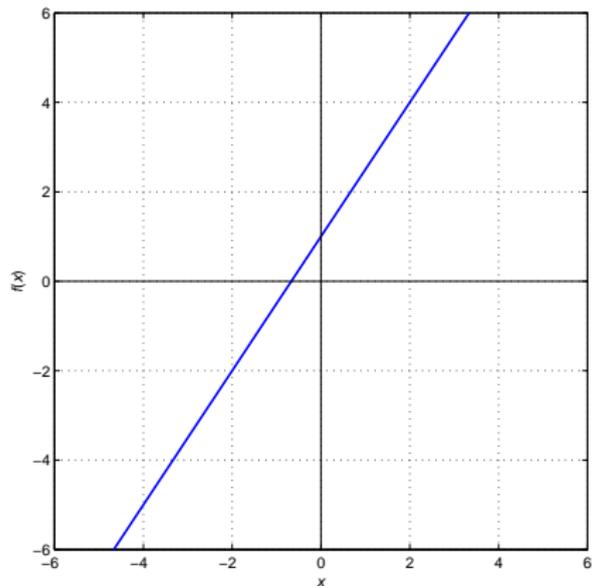
Beispiel mit $b_0 = 1$ und $b_1 = 0$



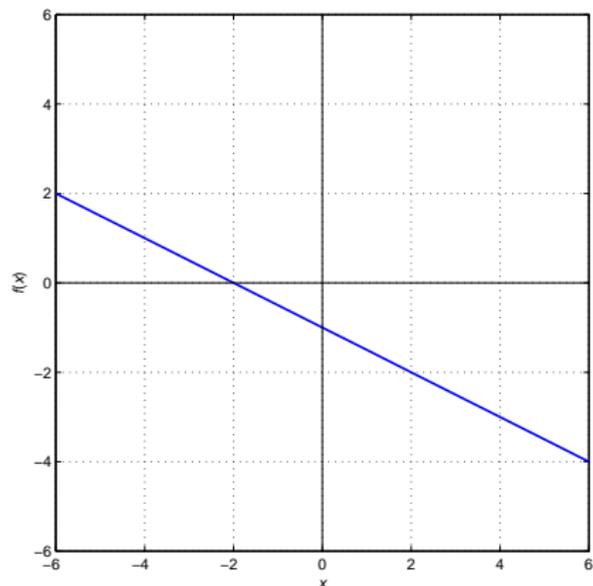
Beispiel mit $b_0 = 0$ und $b_1 = 1$



Beispiel mit $b_0 = 1$ und $b_1 = 1.5$



Beispiel mit $b_0 = -1$ und $b_1 = -0.5$

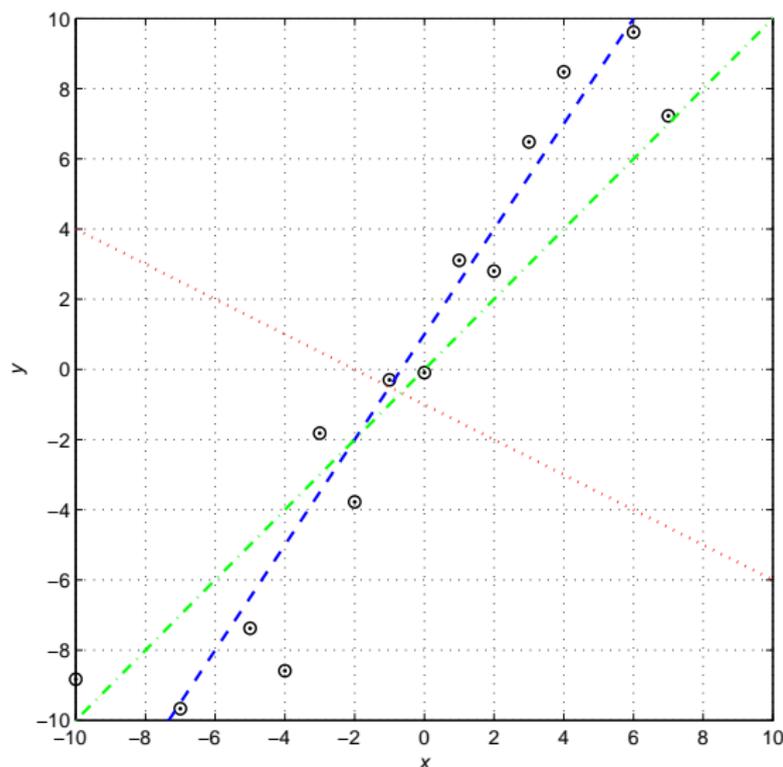


1 Beschreibung von Abhängigkeiten mit Regressionsmodellen

- Allgemeines
- Begriffliches zum Regressionsmodell
- Bedeutung der Modellparameter
- Prinzip der kleinsten Quadrate
- Eigenschaften der Regressionsgeraden
- Modellfit

Für **einen** gegebenen **Datensatz** lassen sich **beliebige Kombinationen** von b_0 und b_1 angeben.

Welche Kombination (welche Linie) ist die **beste** unter allen **möglichen**?



Welche Kombination der Parameterwerte ist die beste unter allen möglichen?

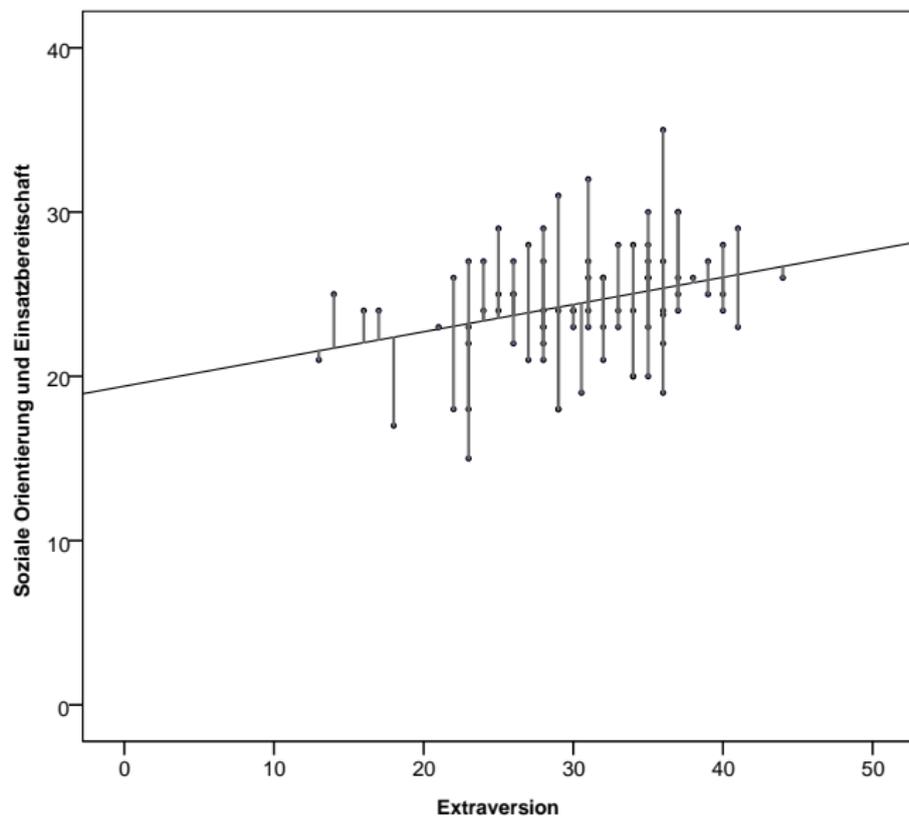
Um dies zu beantworten, muss man ein **Kriterium definieren**, anhand dessen bessere und schlechte Wertekombinationen von b_0 und b_1 erkennbar sind.

Kriterium A: Summe der Abweichungen von der Regressionsgeraden (mathematisch schlecht zu handhaben)

$$\sum_{i=1}^N |e_i| \rightarrow \min \quad (8)$$

Kriterium B: **Summe der quadrierten Abweichungen** von der Regressionsgeraden (**die übliche Methode**)

$$\sum_{i=1}^N (e_i)^2 \rightarrow \min \quad (9)$$



Welche Werte der Koeffizienten sind beim Regressionsmodell

$$y_i = b_0 + b_1 \cdot x_i + e_i$$

optimal im Sinne der Minimierung von

$$\sum_{i=1}^N (e_i)^2 \rightarrow \min \quad ?$$

Da

$$\begin{aligned} e_i &= y_i - (b_0 + b_1 \cdot x_i) \\ &= y_i - b_0 - b_1 \cdot x_i \end{aligned}$$

ist, gilt es, die Summe $S(b_0, b_1)$ zu minimieren.

$$S(b_0, b_1) = \sum_{i=1}^N (y_i - b_0 - b_1 \cdot x_i)^2 \rightarrow \min \quad (10)$$

Die optimalen Koeffizientenwerte erhält man, indem man die partielle Ableitung von $S(b_0, b_1)$ nach b_0 und partielle Ableitung von $S(b_0, b_1)$ nach b_1 bildet und anschließend die Nullstellen der resultierenden Gleichungen bestimmt.¹

Letztendlich erhält man als optimale Koeffizientenwerte

$$\hat{b}_1 = \frac{s_{xy}}{s_x^2} \quad (11)$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \cdot \bar{x} \quad (12)$$

Die Summe der quadrierten Abweichung $S(\hat{b}_0, \hat{b}_1)$ ist bei dieser Kombination von Werten minimal.

Merke: \hat{b}_0 ist der Schätzwert für b_0 und \hat{b}_1 der Schätzwert für b_1 .

¹Für Details sei auf Toutenburg (1998, 2. Aufl., S. 141ff.) verwiesen.

Aufgabe: Für die auf Folie 17 gezeigten Daten seien folgende Kennwerte bekannt: $\bar{x} = 30.51$, $s_x = 6.519$, $\bar{y} = 24.45$, $s_y = 3.549$ und $r_{xy} = 0.304$. Ermitteln Sie bitte folgende Werte:

- $s_{xy} =$
- $s_x^2 =$
- $\hat{b}_1 =$
- $\hat{b}_0 =$

Aufgabe: Bei der Analyse der Daten von Folie 17 erhält man folgenden SPSS Output.

Parameterschätzer

Abhängige Variable :Soziale Orientierung und Einsatzbereitschaft

Parameter	Regressionskoeffizient B	Standardfehler r	T	Sig.	95%-Konfidenzintervall	
					Untergrenze	Obergrenze
Konstanter Term	19,399	1,797	10,792	,000	15,822	22,975
NEO_E	,166	,058	2,876	,005	,051	,280

- 1 Geben Sie die Schätzwerte von b_0 und b_1 an und vergleichen Sie diese mit den von Ihnen berechneten Werten (Folie 20).
- 2 Wie sind die Schätzwerte zu interpretieren?

1 Beschreibung von Abhängigkeiten mit Regressionsmodellen

- Allgemeines
- Begriffliches zum Regressionsmodell
- Bedeutung der Modellparameter
- Prinzip der kleinsten Quadrate
- Eigenschaften der Regressionsgeraden
- Modellfit

Folgende Eigenschaften der Regressionsgeraden sollten Ihnen bekannt sein:

- 1 Das Vorzeichen des Anstiegs (slopes) gleicht dem Vorzeichen des Korrelationskoeffizienten von abhängiger und unabhängiger Variable.
- 2 Der vom Modell für den Mittelwert der unabhängigen Variablen (\bar{x}) vorhergesagte Wert ($f(\bar{x})$) ist der Mittelwert der abhängigen Variablen (\bar{y}).
- 3 Der Mittelwert der durch das Modell vorhergesagten Werte ($\bar{\hat{y}}$) ist gleich dem Mittelwert der unabhängigen Variablen (\bar{y}).
- 4 Die Summe der Abweichungen von Messwerten und vorhergesagten Werten (bzw. die Summe aller Residuen) ist Null.

Das Vorzeichen des Anstiegs (slopes) gleicht dem Vorzeichen des Korrelationskoeffizienten von abhängiger und unabhängiger Variable.

Korrelationskoeffizient und Kovarianz haben folgendes Verhältnis:

$$r = \frac{s_{xy}}{s_x \cdot s_y} \quad (13)$$

$$s_{xy} = r \cdot s_x \cdot s_y \quad (14)$$

Der Anstieg ist das Produkt des Korrelationskoeffizienten und der Wurzel des Verhältnisses beider Varianzen:

$$\hat{b}_1 = \frac{s_{xy}}{s_x^2} \quad (15)$$

$$= \frac{r \cdot s_x \cdot s_y}{s_x^2} \quad (16)$$

$$= r \cdot \frac{s_y}{s_x} \quad (17)$$

Da $\frac{s_y}{s_x}$ immer positiv ist, hat \hat{b}_1 das gleiche Vorzeichen wie r .

Der vom Modell für den Mittelwert der unabhängigen Variablen (\bar{x}) vorhergesagte Wert ($f(\bar{x})$) ist der Mittelwert der abhängigen Variablen (\bar{y}).

Die vom Modell vorhergesagten Werte ergeben sich zu

$$\hat{y}_i = f(x_i) = \hat{b}_0 + \hat{b}_1 \cdot x_i \quad (18)$$

Da $\hat{b}_0 = \bar{y} - \hat{b}_1 \cdot \bar{x}$ ist, erhält man für den Mittelwert der unabhängigen Variablen (\bar{x}) folgenden vorhersagten Wert:

$$f(\bar{x}) = \hat{b}_0 + \hat{b}_1 \cdot \bar{x} \quad (19)$$

$$= \bar{y} - \hat{b}_1 \cdot \bar{x} + \hat{b}_1 \cdot \bar{x} \quad (20)$$

$$= \bar{y} \quad (21)$$

Daraus folgt, dass der **Punkt mit Koordinaten (\bar{x}, \bar{y}) auf der Regressionsgeraden** liegt.

Der Mittelwert der durch das Modell vorhergesagten Werte ($\bar{\hat{y}}$) ist gleich dem Mittelwert der unabhängigen Variablen (\bar{y}).

$$\bar{\hat{y}} = \frac{1}{N} \sum_{i=1}^N (\hat{b}_0 + \hat{b}_1 \cdot x_i) \quad (22)$$

$$= \frac{1}{N} \sum_{i=1}^N \hat{b}_0 + \frac{1}{N} \sum_{i=1}^N \hat{b}_1 \cdot x_i \quad (23)$$

$$= \frac{1}{N} \sum_{i=1}^N (\bar{y} - \hat{b}_1 \cdot \bar{x}) + \frac{1}{N} \sum_{i=1}^N \hat{b}_1 \cdot x_i \quad (24)$$

$$= \frac{1}{N} \sum_{i=1}^N \bar{y} - \frac{1}{N} \sum_{i=1}^N \hat{b}_1 \cdot \bar{x} + \frac{1}{N} \sum_{i=1}^N \hat{b}_1 \cdot x_i \quad (25)$$

$$= \frac{1}{N} \cdot N \cdot \bar{y} \quad (26)$$

$$= \bar{y} \quad (27)$$

Die Summe der Abweichungen von Messwerten und vorhergesagten Werten (bzw. die Summe aller Residuen) ist Null.

Die Abweichung des Messwertes y_i vom vorhergesagten Wert \hat{y}_i ist

$$\hat{e}_i = y_i - \hat{y}_i \quad (28)$$

$$= y_i - (\hat{b}_0 + \hat{b}_1 \cdot x_i) = y_i - \hat{b}_0 - \hat{b}_1 \cdot x_i \quad (29)$$

$$= y_i - (\bar{y} - \hat{b}_1 \cdot \bar{x}) - \hat{b}_1 \cdot x_i = y_i - \bar{y} + \hat{b}_1 \cdot \bar{x} - \hat{b}_1 \cdot x_i \quad (30)$$

Die Summe der Abweichungen ist

$$\sum_{i=1}^N \hat{e}_i = \sum_{i=1}^N (y_i - \bar{y} + \hat{b}_1 \cdot \bar{x} - \hat{b}_1 \cdot x_i) \quad (31)$$

$$= \sum_{i=1}^N y_i - \sum_{i=1}^N \bar{y} + \sum_{i=1}^N \hat{b}_1 \cdot \bar{x} - \sum_{i=1}^N \hat{b}_1 \cdot x_i \quad (32)$$

$$= N \cdot \bar{y} - N \cdot \bar{y} + \hat{b}_1 \cdot N \cdot \bar{x} - \hat{b}_1 \cdot N \cdot \bar{x} \quad (33)$$

$$= 0 \quad (34)$$

1 Beschreibung von Abhängigkeiten mit Regressionsmodellen

- Allgemeines
- Begriffliches zum Regressionsmodell
- Bedeutung der Modellparameter
- Prinzip der kleinsten Quadrate
- Eigenschaften der Regressionsgeraden
- Modellfit

Die Schätzwerte der Regressionskoeffizienten geben über die Stärke der Abhängigkeit (der Betrag des Slopes) und die Art und Weise der Abhängigkeit (das Vorzeichen des Slopes) Auskunft.

Abschließend bleibt die Frage zu klären: **Wie gut sagt die Regression die Messwerte der abhängigen Variablen vorher?**

Das „wie gut“ quantifiziert man in der Statistik mit **Fit-Indizes**, z. B. der Determinationskoeffizient (R^2), das Akaike Information Criterion (AIC) oder der Root Mean Square Error of Approximation (RMSEA).

Einige Fit-Indizes sind so konstruiert, dass

- der Wert Null anzeigt, dass die Regression nicht besser als zufälliges Raten ist, und
- der Wert eins eine perfekte Vorhersage (Messwert = vorhergesagter Wert) anzeigt.

Gleich der Dekomposition des Messwertes der abhängigen Variablen in Regression und Residuum

$$y_i = \underbrace{f(x_i)}_{\text{Regression}} + \underbrace{e_i}_{\text{Residuum}} \quad (35)$$

kann auch die Variation der abhängigen Variablen zerlegt werden in die Summe der Abweichungsquadrate der Regression und die Summe der Abweichungsquadrate der Residuen:

$$SS_{\text{total}} = SS_{\text{Regression}} + SS_{\text{Residuen}} \quad (36)$$

Die Summe der Abweichungsquadrate (engl. sum of squares, deshalb Symbol SS) der abhängigen Variablen ist

$$SS_{\text{total}} = \sum_{i=1}^N (y_i - \bar{y})^2 \quad (37)$$

Die Summe der Abweichungsquadrate der Regression ist

$$SS_{\text{Residuen}} = \sum_{i=1}^N (\hat{y}_i - \bar{y})^2 \quad (38)$$

Die Summe der Abweichungsquadrate der Residuen ist

$$SS_{\text{Residuen}} = \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (39)$$

Wir beschäftigen uns hier mit dem **Determinationskoeffizienten** (auch **Bestimmtheitsmaß**; Symbol R^2).

R^2 -Wert gibt den **Anteil der Variation der abhängigen Variablen an, der durch die Regression erklärt wird**.

$$R^2 = \frac{SS_{\text{Regression}}}{SS_{\text{total}}} \quad (40)$$

$$= \frac{SS_{\text{Regression}}}{SS_{\text{Regression}} + SS_{\text{Residuen}}} \quad (41)$$

Es gilt $0 \leq R^2 \leq 1$. $R^2 = 0$ ist gleichbedeutend mit raten (das Modell kann „nichts“ erklären) und $R^2 = 1$ gleichbedeutend mit einer perfekten Vorhersage.

Es kann gezeigt werden, dass der Determinationskoeffizient gleich dem Quadrat des Korrelationskoeffizienten ist (siehe z. B. Toutenburg, 1998, 2. Aufl., S. 150ff.)

$$R^2 = r^2 \quad (42)$$