2. Korrelation, lineare Regression und multiple Regression

2.1 Korrelation

2.2 Lineare Regression

2.3 Multiple lineare Regression

2.4 Nichtlineare Zusammenhänge

2.1 Korrelation

2.1 Beispiel: Arbeitsmotivation

- Untersuchung zur Motivation am Arbeitsplatz in einem Chemie-Konzern
- 25 Personen werden durch Arbeitsplatz zufällig ausgewählt und verschiedene Variablen gemessen.
- y: Motivation (Einschätzung durch Experten)
  x: Leistungsstreben (Fragebogen)
- **Frage**: besteht ein Zusammenhang zwischen der Variablen "Motivation" und der Variablen "Leistungsstreben"
- Beachte: es werden auch noch weitere Variablen gemessen (Ehrgeiz, Kreativität, Hierarchie, Lohn, Arbeitsbedingungen, Lernpotential, Vielfalt, Anspruch)

Daten

x	20	30	15	39	5	6	12	0	35
y	32	14	12	27	20	13	17	8	22
x	8	34	26	32	26	12	36	27	26
y	19	25	23	17	22	19	27	26	20
x	13	19	25	30	18	21	11		
y	11	24	19	19	22	24	17		

2.2 Der Korrelationskoeffizient von Pearson

- Daten $(x_1, y_1), \ldots, (x_n, y_n)$
- Maß für die (lineare) Abhängigkeit zwischen x und y: **Korrelationskoeffizient von Pearson**

$$\hat{\rho}_{X,Y} = \frac{s_{x,y}}{s_{x,x} s_{y,y}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_\cdot)(y_i - \overline{y}_\cdot)}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x}_\cdot)^2 \sum_{i=1}^{n}(y_i - \overline{y}_\cdot)^2}}$$

- Dabei ist:
  - $\overline{x}_\cdot = \frac{1}{n}\sum_{i=1}^{n} x_i$: Mittelwert der Daten x_i
  - $\overline{y}_\cdot = \frac{1}{n}\sum_{i=1}^{n} y_i$: Mittelwert der Daten y_i
  - $s^2_{x,x} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_\cdot)^2$: Varianz der Daten x_i
  - $s^2_{y,y} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y}_\cdot)^2$: Varianz der Daten y_i
  - $s^2_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_\cdot)(y_i - \overline{y}_\cdot)$: Kovarianz zwischen den Daten x_i, y_i

2.3 Eigenschaften des Korrelationskoeffizienten

(1) $-1 \leq \hat{\rho}_{X,Y} \leq 1$

(2) $\hat{\rho}_{X,Y} = 1$ genau dann, wenn ein exakter linearer Zusammenhang

$$y_i = b_0 + b_1 x_i$$

mit $b_1 > 0$ besteht (ohne Störgrößen).

(3) $\hat{\rho}_{X,Y} = -1$ genau dann, wenn ein exakter linearer Zusammenhang

$$y_i = b_0 + b_1 x_i$$

mit $b_1 < 0$ besteht (ohne Störgrößen).

(4) Der Korrelationskoeffizient ist invariant bzgl. linearer Transformationen, d.h.

$$\left.\begin{array}{ll} \tilde{x}_i = a_0 + a_1 x_i & i = 1, \ldots, n \\ \tilde{y}_i = c_0 + c_1 y_i & i = 1, \ldots, n \end{array}\right\} \Rightarrow \rho_{\tilde{X},\tilde{Y}} = \rho_{X,Y}$$

(4) Der Korrelationskoeffizient von Pearson ist ein deskriptives Maß für den **linearen** Zusammenhang in der Stichprobe $(x_1, y_1), \ldots, (x_n, y_n)$

2.4 Beispiel: Korrelationskoeffizient für die Daten aus Beispiel 2.1

- Variablen
  x: Leistungsstreben
  y: Motivation
- Korrelationskoeffizient von Pearson

$$\hat{\rho}_{x,y} = 0.5592$$

- **Fragen**:
  - Wie genau ist diese Schätzung?
  - Ist die Korrelation von 0 verschieden (Unkorreliertheit zwischen den Merkmalen Leistungsstreben und Motivation)?

2.5 Signifikanztest für Korrelation

- ρ bezeichne die Korrelation des Merkmals X mit dem Merkmal Y einer Population
- $(x_1, y_1), \ldots, (x_n, y_n)$ ist eine Stichprobe (unabhängige Beobachtungen) aus einer (bivariat) normalverteilten Grundgesamtheit
- Ein Test zum Niveau α für die Hypothese "die Merkmale sind unkorreliert"

$$H_0 : \rho = 0$$

lehnt die Nullhypothese zu Gunsten der Alternative $H_1 : \rho \neq 0$ ab, falls

$$\left| \frac{\sqrt{n-2}\hat{\rho}_{x,y}}{\sqrt{1-\hat{\rho}^2_{x,y}}} \right| > t_{n-2,1-\alpha/2}$$

gilt.

2.6(a) Beispiel: Arbeitsmotivation (Fortsetzung von Beispiel 2.1)

- $n = 25$; $\hat{\rho}_{x,y} = 0.5592$; $t_{23,0.95} = 1.7139$
- $$\left| \frac{\sqrt{n-2}\ \hat{\rho}_{x,y}}{\sqrt{1-\hat{\rho}_{x,y}^2}} \right| = 3.2355 > 1.7139$$
- Die Nullhypothese $H_0 : \rho = 0$ (keine Korrelation zwischen den Merkmalen) wird zum Niveau 10% verworfen.
- p-Wert: 0.0037

SPSS Output für Korrelationskoeffizient

Korrelationen

		Motivation	Leistungsstreben
Motivation	Korrelation nach Pearson	1,000	,559**
	Signifikanz (2-seitig)		,004
	N	25	25
Leistungsstreben	Korrelation nach Pearson	,559**	1,000
	Signifikanz (2-seitig)	,004	
	N	25	25

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

2.7 Konfidenzintervall für Korrelation

- ρ: Korrelation zwischen Merkmal x und Merkmal y einer Population
- $(x_1, y_1), \ldots, (x_n, y_n)$: Stichprobe (unabhängige Beobachtungen) aus einer (bivariat) normalverteilten Grundgesamtheit
- Mathematische Statistik: $\hat{\rho}_{x,y}$ ist "näherungsweise" (d.h. bei großem Stichprobenumfang) normalverteilt mit Erwartungswert ρ und Varianz

$$\gamma^2 = Var(\hat{\rho}_{x,y}) \approx (1 - \rho^2)^2/n$$

- $(1 - \alpha)$-Konfidenzintervall für den Korrelationskoeffizienten

$$\left(\hat{\rho}_{x,y} - \hat{\gamma} z_{1-\alpha/2}, \hat{\rho}_{x,y} + \hat{\gamma} z_{1-\alpha/2}\right)$$

Hier bezeichnet $\hat{\gamma} = (1 - \hat{\rho}^2_{x,y})/\sqrt{n}$ einen Schätzer für die Standardabweichung von $\hat{\rho}_{x,y}$ und $z_{1-\alpha/2}$ das $(1 - \alpha/2)$ Quantil der Standardnormalverteilung (Tabelle, Software)

2.6(b) Beispiel: Arbeitsmotivation (Fortsetzung von Beispiel 2.1)

- $n = 25; \qquad \hat{\rho}_{x,y} = 0.5592$
- $z_{0.95} = 1.6449, \quad \hat{\gamma} = 0.1328$
- $\Rightarrow$ 90% Konfidenzintervall für den Korrelationskoeffizient

$$[0.2739, 0.7541]$$

2.8 Hinweise zur Interpretation von Korrelationen

- Annahme: man hat eine signifikante Korrelation zwischen dem Variablen x und y gefunden
- Folgende Interpretationen sind möglich
  (1) x beeinflusst y kausal
  (2) y beeinflusst x kausal
  (3) x und y werden von weiteren Variablen kausal beeinflusst
  (4) x und y beeinflussen sich wechselseitig kausal
- Die Korrelation zwischen zwei Variablen ist eine **notwendige** aber keine **hinreichende** Voraussetzung für einen kausalen Zusammenhang
- Der Korrelationskoeffizient gibt **keine** Information welche der vier Interpretationen zutrifft (in "vielen" Fällen wird das der Typ (3) sein)
- Korrelationen sollten ohne Zusatzinformation nicht interpretiert werden!

Beispiel

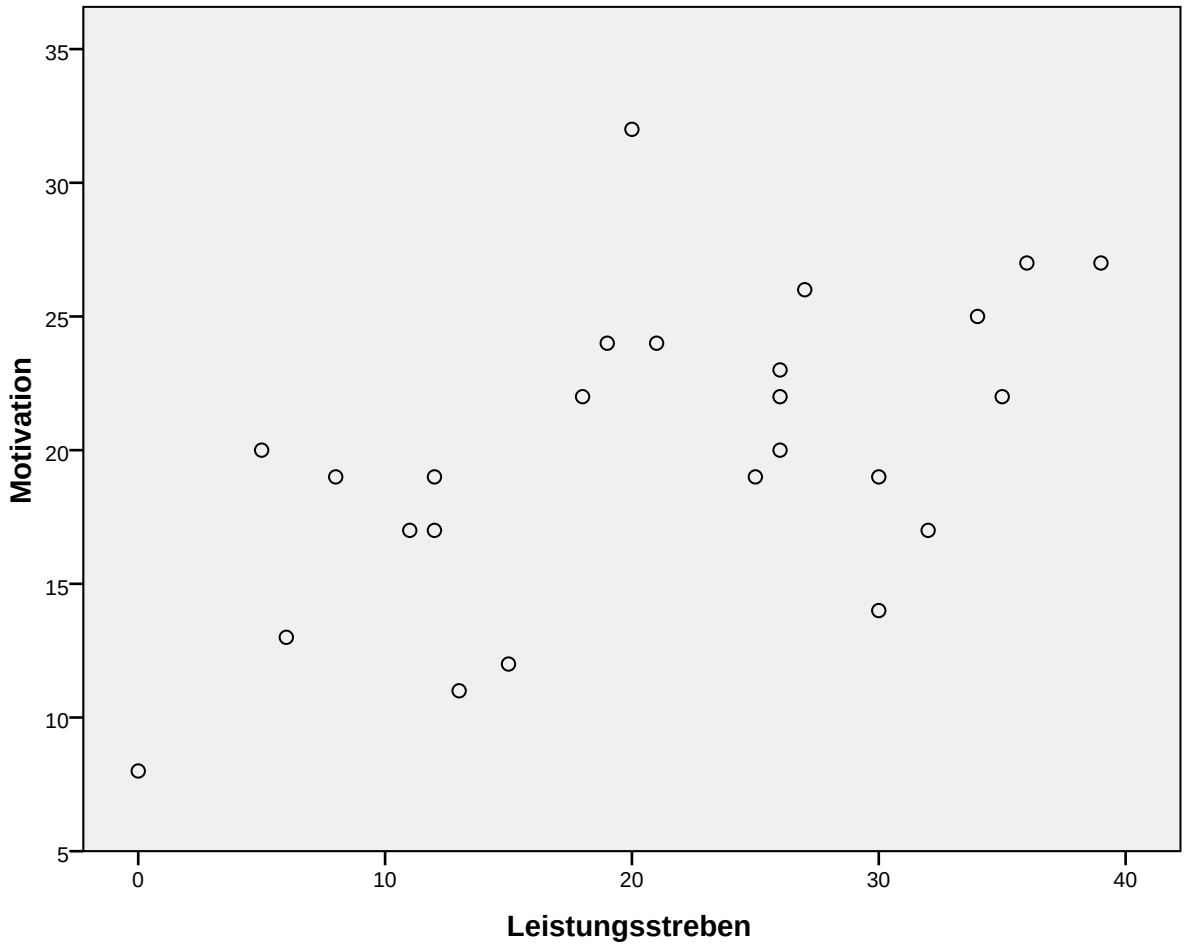
- Annahme: man hat eine signifikante Korrelation zwischen den Merkmalen "Ehrlichkeit" und "Häufigkeit des Kirchgangs" gefunden
- Folgende Interpretationen sind möglich
  - Die in der Kirche vermittelten Werte haben einen positiven Einfluß auf das Merkmal "Ehrlichkeit"
  - "Ehrliche" Menschen fühlen sich durch die in der Kirche vermittelten Inhalte eher angesprochen und gehen aus diesem Grund häufiger zur Kirche
  - Die allgemeine familiäre und außerfamiliäre Sozialisation beeinflußt **beide** Merkmale

2.2 Lineare Regression

2.9 Beispiel: (Fortsetzung von Beispiel 2.1)

- Untersuchung zur Motivation am Arbeitsplatz in einem Chemie-Konzern
- 25 Personen werden zufällig ausgewählt und verschiedene Variablen gemessen.
- y: Motivation (Einschätzung durch Experten)
  x: Leistungsstreben (Fragebogen)
- Kann man y aus x "vorhersagen"?

Streudiagramm für die Daten aus Beispiel 2.9



2.9 Beispiel: (Fortsetzung von Beispiel 2.1)

- Untersuchung zur Motivation am Arbeitsplatz in einem Chemie-Konzern
- 25 Personen werden zufällig ausgewählt und verschiedene Variablen gemessen.
- y: Motivation (Einschätzung durch Experten)
  x: Leistungsstreben (Fragebogen)
- **Frage**: besteht ein **funktionaler** Zusammenhang zwischen der Variablen "Motivation" und der Prädiktorvariablen "Leistungsstreben" (Kann man y aus x "vorhersagen"?)

  Genauer: Gesucht ist Funktion f, die aus der Prädiktorvariablen *Leistungsstreben* (x) eine Vorhersage für die abhängige Variable (y) *Motivation* liefert:
  Motivation = f(Leistungsbereitschaft)
- Beachte: es werden auch noch weitere Variablen gemessen (Ehrgeiz, Kreativität, Hierarchie, Lohn, Arbeitsbedingungen, Lernpotential, Vielfalt, Anspruch)

Regression

- **Ausgangslage:** Von Interesse ist der Zusammenhang zwischen verschiedenen Variablen. Im einfachsten Fall betrachtet man, wie im Beispiel der Arbeitsmotivation, den Zusammenhang zwischen zwei Variablen.
- Daten: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Annahme: Es existiert ein kausaler Zusammenhang der Form $y = f(x)$ zwischen der abhängigen Variablen y und der Prädiktorvariablen x.
  Weitere Annahme: Die Funktion f hat eine bestimmte Form.
  Beispiele:
  - **Lineare** Regression (der Zusammenhang ist also durch eine Gerade beschreibbar): $y = b_0 + b_1 x$
  - **Quadratische** Regression (der Zusammenhang ist also durch eine Parabel beschreibbar): $y = b_0 + b_1 x + b_2 x^2$
  - usw.
- **Beachte:** Der Zusammenhang ist in der Regel nicht exakt zu beobachten. Mathematisches Modell

$$Y = b_0 + b_1 x + \varepsilon$$

  Dabei bezeichnet ε eine **zufällige** Störgröße. Diese Modell bezeichnet man als **Lineare Regression**.

2.10 Das Modell der linearen Regression

- Daten $(x_1, y_1), \ldots, (x_n, y_n)$
- y_i ist Realisation einer Zufallsvariablen Y_i (unter der Bedingung x_i). Für den Zusammenhang zwischen den Variablen Y_i und x_i gilt:

$$Y_i = b_0 + b_1 x_i + \varepsilon_i \qquad i = 1, \ldots, n$$

- ε_i bezeichnet hier eine zufällige "Störung" und es wird angenommen, dass die Störungen **unabhängig** und **normalverteilt** sind mit Erwartungswert 0 und Varianz $\sigma^2 > 0$
- Deutung: es wird ein linearer Zusammenhang zwischen x und y postuliert, der noch zufälligen Störungen unterliegt

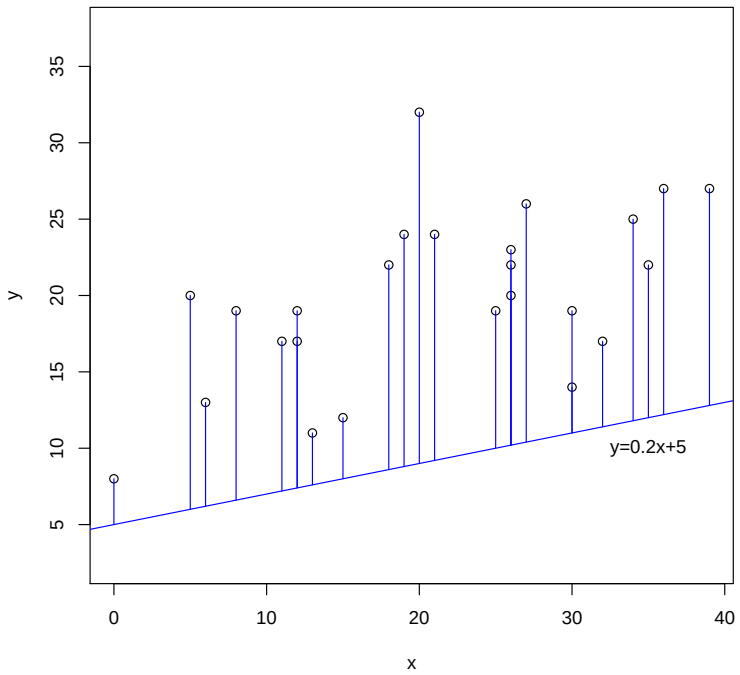
Idee der Schätzung bei (linearer) Regression

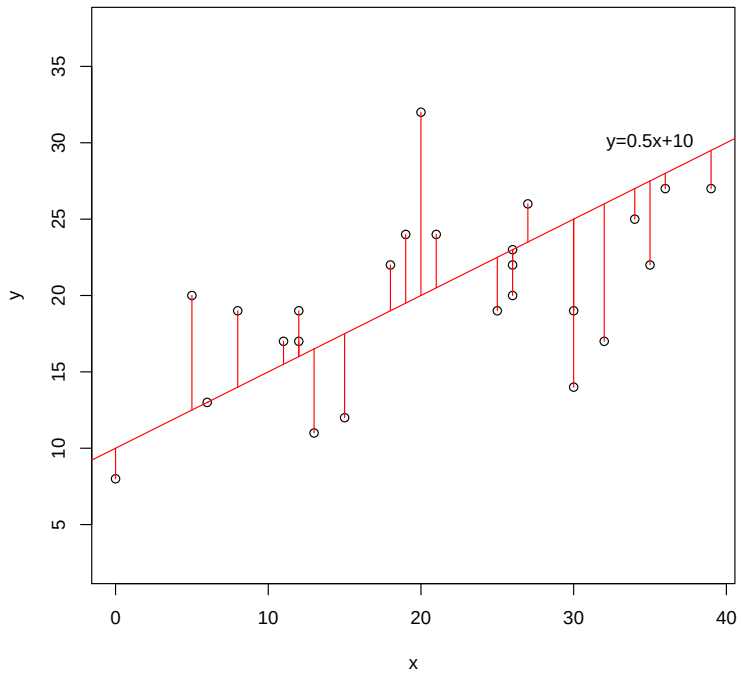
- Daten $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Annahme: Es existiert ein linearer Zusammenhang

$$Y = b_0 + b_1 x + \varepsilon$$

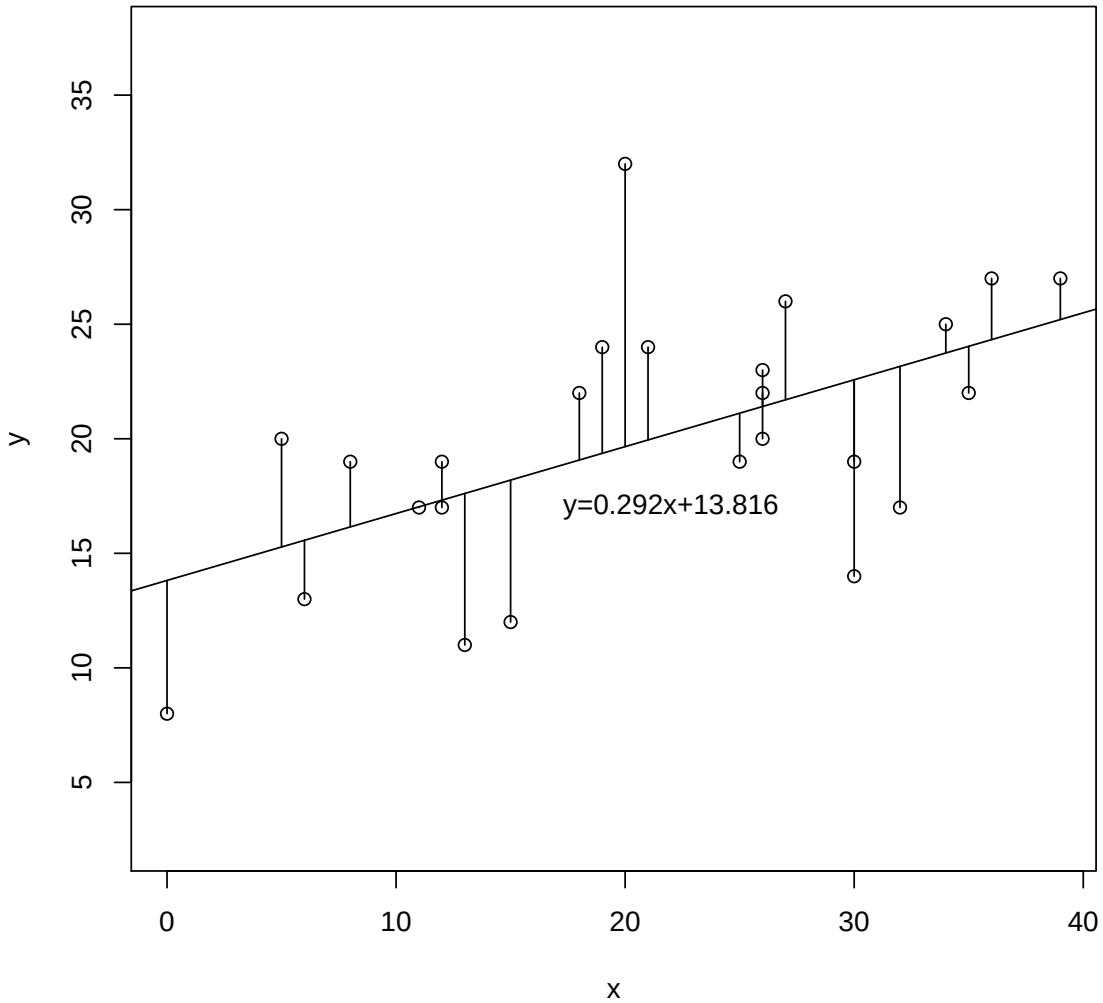
- **Gesucht:** Diejenige Gerade, die den Zusammenhang zwischen Y und x am besten beschreibt.
- **Idee:** Bestimme die Gerade so, dass die Summe der quadratischen (vertikalen) Abstände zwischen den y-Koordinaten der Datenpunkte und den entsprechenden Punkten auf der geschätzten Geraden minimal wird **Methode der kleinsten Quadrate**

Beispiel: Verschiedene Geraden mit senkrechten Abständen zu den Daten





Beispiel: Verschiedene Geraden mit senkrechten Abständen zu den Daten: die Lösung durch die Methode der kleinsten Quadrate



2.11 Die Methode der kleinsten Quadrate

- Bestimme die Gerade so, dass die Summe der quadrierten senkrechten Abstände zwischen Gerade und Daten minimal wird
  - Datum an der Stelle x_i: y_i
  - Wert der Geraden an der Stelle x_i: $b_0 + b_1 x_i$
  - Differenz: $y_i - (b_0 + b_1 x_i)$

- Minimiere

  $h(b_0, b_1) = \sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i)\right)^2$

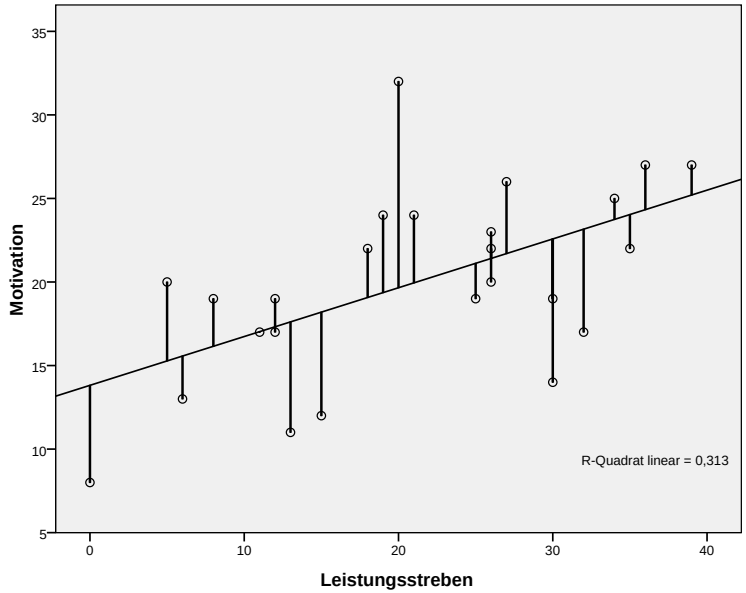
  bzgl. der Wahl der Parameter b_0 und b_1.

- Lösung dieses Extremwertproblems liefert Schätzer für Achsenabschnitt und Steigung der Geraden:

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x.})(y_i - \overline{y.})}{\sum_{i=1}^{n}(x_i - \overline{x.})^2}, \quad \hat{b}_0 = \overline{y.} - \hat{b}_1 \overline{x.}$$

  - $\overline{x.} = \frac{1}{n}\sum_{i=1}^{n} x_i$: Mittelwert der Prädiktorvariablen
  - $\overline{y.} = \frac{1}{n}\sum_{i=1}^{n} y_i$: Mittelwert der abhängigen Variablen

Beispiel Arbeitsmotivation: Streudiagramm und Regressionsgerade für die Daten aus Beispiel 2.1



- Schätzer: $\hat{b}_0 = 13.82,\ \hat{b}_1 = 0.29$
- **Fragen:**
  - Wie genau sind diese Schätzungen?
  - Besteht ein (signifikanter) Einfluß des Leistungsstrebens auf die Motivation

$$H_0 : b_1 = 0$$

  - Wie gut beschreibt das lineare Regressionsmodell die Situation?

Die Genauigkeit der Schätzer für die Parameter

- Beachte: vor der Datenerhebung sind $\hat{b}_0$ und $\hat{b}_1$ zufällig
- Mathematische Statistik (allegmeines lineares Modell) liefert Schätzer für die Varianzen von $\hat{b}_0$ und $\hat{b}_1$
-

Schätzer für die Varianz von $\hat{b}_0$:
$$\hat{s}^2_{b_0} = \frac{S^2_{y|x}}{n} \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \overline{x}.)^2}$$

Schätzer für die Varianz von $\hat{b}_1$:
$$\hat{s}^2_{b_1} = \frac{S^2_{y|x}}{n} \frac{1}{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x}.)^2}$$

Dabei bezeichnet

$$S^2_{y|x} = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - (\hat{b}_0 + \hat{b}_1 x_i))^2.$$

die **Residualvarianz** (Schätzer für die Varianz der Störgrößen)

- Je größer der Stichprobenumfang n, desto genauer sind die Schätzungen!

Fortsetzung von Beispiel 2.1: Schätzer für die Daten der Arbeitsmotivation

- Schätzer für die Parameter

$$\begin{aligned} \hat{b}_0 &= 13.82 \\ \hat{b}_1 &= 0.292 \\ S^2_{y|x} &= 22.737 \end{aligned}$$

- Schätzer für die Varianz von $\hat{b}_0$ und $\hat{b}_1$

$$\begin{aligned} \hat{s}^2_{b_0} &= 4.5158 \\ \hat{s}^2_{b_1} &= 0.0081 \end{aligned}$$

- Je größer der Stichprobenumfang n, desto genauer sind die Schätzungen!

SPSS Output: Schätzer und Standardabweichungen bei linearer Regression in Beispiel 2.1

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	13,816	2,125		6,501	,000
	Leistungsstreben	,292	,090	,559	3,235	,004

a. Abhängige Variable: Motivation

2.12 Konfidenzintervalle bei linearer Regression

- Modellannahme: lineare Regression

$$Y_i = b_0 + b_1 x_i + \varepsilon_i \ (i = 1, \ldots, n)$$

- Rechtfertigung der Normalverteilungs- und Unabhängigkeitsannahme für $\varepsilon_1, \ldots, \varepsilon_n$
- Bestimmung der Schätzer $\hat{s}^2_{b_0}$ und $\hat{s}^2_{b_1}$ für die Varianzen von $\hat{b}_0$ und $\hat{b}_1$. Damit ist dann

$$\Longrightarrow \qquad (\hat{b}_0 - t_{n-2,1-\alpha/2}\hat{s}_{b_0}, \hat{b}_0 + t_{n-2,1-\alpha/2}\hat{s}_{b_0})$$

ein $(1-\alpha)$ -Konfidenzintervall für b_0 und

$$\Longrightarrow \qquad (\hat{b}_1 - t_{n-2,1-\alpha/2}\hat{s}_{b_1}, \hat{b}_1 + t_{n-2,1-\alpha/2}\hat{s}_{b_1})$$

ein $(1-\alpha)$ -Konfidenzintervall für b_1. Hier ist $t_{n-2,1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der t-Verteilung mit $n-2$ Freiheitsgraden (tabelliert oder mit Software verfügbar)

Beispiel 2.13: Konfidenzbereiche im Beispiel 2.1 (Arbeitsmotivation)

- $n = 25, \qquad t_{23,0975} = 2.0687$
- Für das Beispiel der Arbeitsmotivation (vgl. Beispiel 2.1) ergibt sich als 95% Konfidenzintervall für

$$b_0: \quad [9.420, 18.212]$$
$$b_1: \quad [0.105, 0.479]$$

- **Frage:** Besteht ein (signifikanter) Einfluß der Prädikorvaraiblen x auf die abhängige Variable Y? Mathematische Formulierung: $H_0 : b_1 = 0$

SPSS Output: Konfidenzintervalle bei linearer Rgeression in Beispiel 2.1

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten			95%-Konfidenzintervall für B	
		B	Standardfehler	Beta	T	Signifikanz	Untergrenze	Obergrenze
1	(Konstante)	13,816	2,125		6,501	,000	9,420	18,212
	Leistungsstreben	,292	,090	,559	3,235	,004	,105	,479

a. Abhängige Variable: Motivation

2.14 F-test für die Hypothese $H_0 : b_1 = 0$

- Modellannahme: lineare Regression

$$Y_i = b_0 + b_1 x_i + \varepsilon_i \ (i = 1, \ldots, n)$$

- Rechtfertigung der Normalverteilungs- und Unabhängigkeitsannahme für $\varepsilon_1, \ldots, \varepsilon_n$
- Hypothesen

$$H_0 : b_1 = 0 \quad , \quad H_1 : b_1 \neq 0$$

- Die Nullhypothese $H_0 : b_1 = 0$ wird zu Gunsten der Alternative $H_1 : b_1 \neq 0$ verworfen, falls

$$F_n = \frac{S^2_{reg}}{S^2_{y|x}} = \frac{\frac{1}{1}\sum_{i=1}^{n}(\overline{y_\cdot} - (\hat{b}_0 + \hat{b}_1 x_i))^2}{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{b}_0 + \hat{b}_1 x_i))^2} > F_{1;n-2,1-\alpha}$$

gilt

- $F_{1;n-2,1-\alpha}$ bezeichnet das $(1-\alpha)$-Quantil der F-Verteilung mit $(1, n-2)$ Freiheitsgraden

Motivation des F-Tests: Zerlegung der Varianz

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y}_\cdot)^2}_{\text{Gesamtvarianz}} = \underbrace{\sum_{i=1}^{n}(y_i - (\hat{b}_0 + \hat{b}x_i))^2}_{\text{Residualvarianz}} + \underbrace{\sum_{i=1}^{n}(\overline{y}_\cdot - (\hat{b}_0 + \hat{b}_1 x_i))^2}_{\text{Varianz der Regression}}$$

- Bezeichnungen:

$$S^2_{reg} = \frac{1}{1}\sum_{i=1}^{n}(\overline{y}_\cdot - (\hat{b}_0 + \hat{b}_1 x_i))^2$$

heißt Varianz der Regression (diese hat 1 Freiheitsgrad) und

$$S^2_{y|x} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{b}_0 + \hat{b}_1 x_i))^2.$$

ist die **Residualvarianz** (diese hat $n-2$ Freiheitsgrade).

- Andere Interpretationen:
  - Schätzung für die Varianz der Größen ε_i
  - durch das lineare Regressionsmodell nicht erklärbare Varianz

Motivation des F-Tests: Zerlegung der Varianz

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y}_.)^2}_{\text{Gesamtvarianz}} = \underbrace{\sum_{i=1}^{n}(y_i - (\hat{b}_0 + \hat{b}x_i))^2}_{\text{Residualvarianz}} + \underbrace{\sum_{i=1}^{n}(\overline{y}_. - (\hat{b}_0 + \hat{b}_1 x_i))^2}_{\text{Varianz der Regression}}$$

$$= 1 \cdot S^2_{reg} + (n-2) \cdot S^2_{y|x}$$

Beachte:

- Bei dem F-Test für die Hypothese $H_0 : b_1 = 0$ bildet man den Quotienten aus der Varianz der Regression und der Residualvarianz
- Man untersucht also, welcher Anteil der Gesamtvarianz durch die Varianz der Regression erklärbar ist

2.15 Varianzanalyse (ANOVA; analysis of variance)

Art der Abweichung	Freiheits-grade (df)	Quadrat-summe	F-Quotient schätzer	
Regression	1	$\sum_{i=1}^{n}(\overline{y}_. - \widehat{y}_i)^2$	$F_n = S^2_{reg}/S^2_{y\|x}$	
Fehler	$n-2$	$\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$	—	
Total	$n-1$	$\sum_{i=1}^{n}(y_i - \overline{y}_.)^2$	—	—

Bezeichnung:

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$$

SPSS Output: F-Test bei linearer Regression in Beispiel 2.1

ANOVA[b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	238,015	1	238,015	10,468	,004[a]
	Residuen	522,945	23	22,737		
	Gesamt	760,960	24			

a. Einflußvariablen : (Konstante), Leistungsstreben

b. Abhängige Variable: Motivation

Beachte:

- $F_{25} = 10.468, \qquad F_{1,23,0.95} = 4.2793$
- Da $F_{25} = 10.468 > 4.2793$ wird die Nullhypothese $H_0 : b_1 = 0$ zu Gunsten der Alternative $H_1 : b_1 \neq 0$ zum Niveau 5% verworfen (p-Wert: 0.004)

Modellgüte: "wie geeignet" ist das Modell für die Beschreibung der Daten

- Maß für Modellanpassung: **Residualvarianz** (Summe der quadrierte Abstände von der Regressionsgerade):

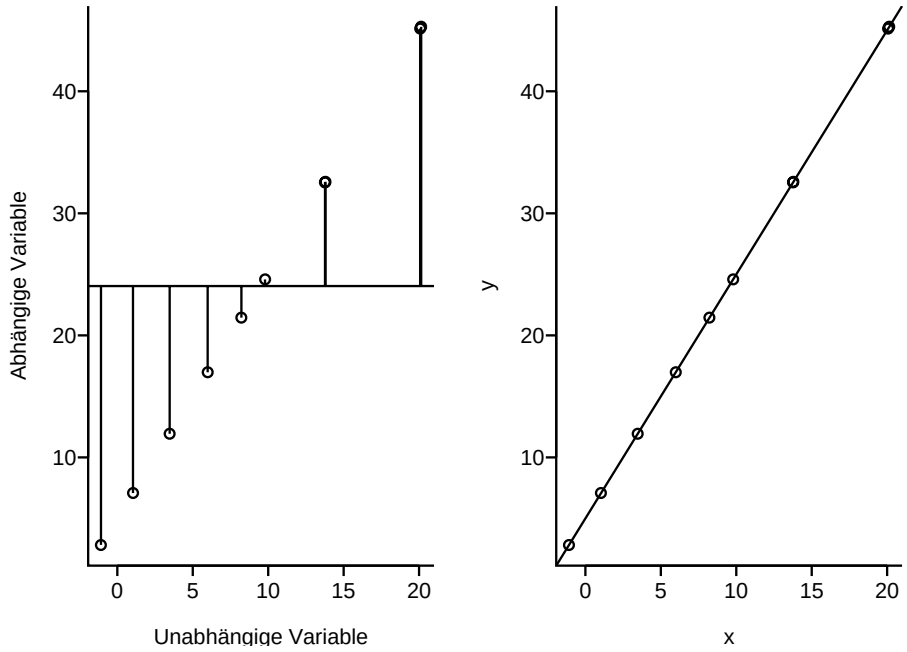
$$S^2_{y|x} = \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - (\hat{b}_0 + \hat{b}_1 x_i) \right)^2$$

- Beachte: $S^2_{y|x}$ ist ein Schätzer für die Varianz der Meßfehler
- Je **kleiner** $S^2_{y|x}$, desto **"besser"** ist das (lineare) Regressionsmodell
- Streuung der Daten ohne die "Information", dass ein lineares Modell vorliegt:

$$\sum_{i=1}^{n} (y_i - \overline{y}_\cdot)^2$$

- Man untersucht welchen Anteil der Streuung $\sum_{i=1}^{n} (y_i - \overline{y}_\cdot)^2$ man durch das lineare Modell erklären kann

Varianzzerlegung: ein extremes Beispiel



Beachte:

- Die Grafik zeigt eine extreme Situation.
- Die Streuung der Daten ist durch das lineare Regressionsmodell zu 100% erklärbar! $\sum_{i=1}^{n}(y_i - \overline{y}_.)^2 = \sum_{i=1}^{n}(\overline{y}_. - (\hat{b}_0 + \hat{b}_1 x_i))^2$
- Residualvarianz (durch das lineare Regressionsmodell nicht erklärbare Varianz) $= 0$

2.16 Beispiel: Arbeitsmotivation (Fortsetzung von Beispiel 2.1):

$$\sum_{i=1}^{25}(y_i - \overline{y}_.)^2 = 760.96$$

$$\sum_{i=1}^{25}(\overline{y}_. - (\hat{b}_0 + \hat{b}_1 x_i))^2 = 238.04$$

$$R^2 = \frac{\sum_{i=1}^{25}(\overline{y}_. - (\hat{b}_0 + \hat{b}_1 x_i))^2}{\sum_{i=1}^{25}(y_i - \overline{y}_.)^2} = 0.313$$

d.h. 31,3 % der Varianz der Variablen *Motivation* können durch die Prädiktorvariable *Leistungsstreben* erklärt werden.

2.17 Modellgüte: das Bestimmtheitsmaß

- Die Größe

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - (\hat{b}_0 + \hat{b}_1 x_i))^2}{\sum_{i=1}^{n}(y_i - \overline{y}_.)^2} = \frac{\sum_{i=1}^{n}(\overline{y}_. - (\hat{b}_0 + \hat{b}_1 x_i))^2}{\sum_{i=1}^{n}(y_i - \overline{y}_.)^2}$$

  ist ein Maß für die Güte der Regression und heißt **Bestimmtheitsmaß**.

- Beachte: man kann zeigen, dass R^2 genau das Quadrat der Korrelation ist.

- **Je "besser" das Modell ist, desto kleiner ist die Residualvarianz, bzw. desto größer R^2!**

- Das Bestimmtheitsmaß R^2 liegt immer zwichen 0 und 1

Zusammenhang zwischen Bestimmtheitsmaß und F-Test

- Ist F_n die Statistik für den F-Test aus 2.14 und R^2 das Bestimmtheitsmaß , dann gilt:

$$R^2 = \frac{\frac{1}{n-2} F_n}{1 + \frac{1}{n-2} F_n}$$

- In anderen Worten: die Statistik F_n des F-Test aus 2.5 kann aus dem Bestimmtheitsmaß berechnet werden (und umgekehrt)
- Im Beispiel des Zusammenhangs zwischen *Motivation* und *Leistungsstreben* ist

$$F_n = 10.468 \quad \Longrightarrow \quad R^2 = \frac{\frac{10.468}{23}}{1 + \frac{10.468}{23}} = 0.313$$

C.a. 31.3% der Variation der Variablen *Motivation* können durch die die Variable *Leistungsstreben* erklärt werden

Vorhersagen: es gibt zwei unterschiedliche

2.18 Vorhersage für den **Wert der Geraden** an einer Stelle x

- Schätzung für den Wert der Geraden $y(x) = b_0 + b_1 x$ an der Stelle x:

$$\widehat{y}(x) = \hat{b}_0 + \hat{b}_1 x$$

- $(1 - \alpha)$-Konfidenzintervall für $y(x)$

$$(\widehat{y}(x) - t_{n-2;\alpha/2} \cdot \hat{s}_{y(x)}, \widehat{y}(x) + t_{n-2;\alpha/2} \cdot \hat{s}_{y(x)})$$

wobei

$$\hat{s}^2_{y(x)} = S^2_{y|x}\left(\frac{1}{n} + \frac{(x - \overline{x}.)^2}{\sum_{i=1}^{n}(x_i - \overline{x}.)^2}\right)$$

den Schätzer für die Varianz von $\widehat{Y}(x)$ bezeichnet

Vorhersagen: es gibt zwei unterschiedliche

2.19 Vorhersage für **eine neue Beobachtung** an einer Stelle x

- Schätzer für eine neue Beobachtung $\tilde{Y}(x) = b_0 + b_1 x + \varepsilon$ an der Stelle x:

$$\widehat{y}(x) = \hat{b}_0 + \hat{b}_1 x$$

- $(1-\alpha)$-Konfidenzintervall für $y(x)$

$$(\widehat{y}(x) - t_{n-2;\alpha/2} \cdot \tilde{s}_{y(x)}, \widehat{y}(x) + t_{n-2;\alpha/2} \cdot \tilde{s}_{y(x)})$$

wobei

$$\tilde{s}^2_{y(x)} = S^2_{y|x}\left(1 + \frac{1}{n} + \frac{(x - \overline{x}.)^2}{\sum_{i=1}^{n}(x_i - \overline{x}.)^2}\right)$$

den Schätzer für die Varianz von $\widehat{y}(x) + \varepsilon$ bezeichnet

- **Beachte:** Diese Varianz wird bei wachsendem Stichprobenumfang **nicht** beliebig klein!

2.20 Beispiel (Fortsetzung von Beispiel 2.1)

(1) Gesucht ist ein 90% Konfidenzintervall für den Wert der Geraden an der Stelle $x = 16$

- $t_{23,0.95} = 1.714$,
  $S^2_{y|x} = 22.737, \hat{s}^2_{y(x)} = 1.116, \hat{y}(16) = \hat{b}_0 + 16\hat{b}_1 = 18.49$
- Das 90% Konfidenzintervall für den Wert der Geraden an der Stelle 16 ist gegeben durch

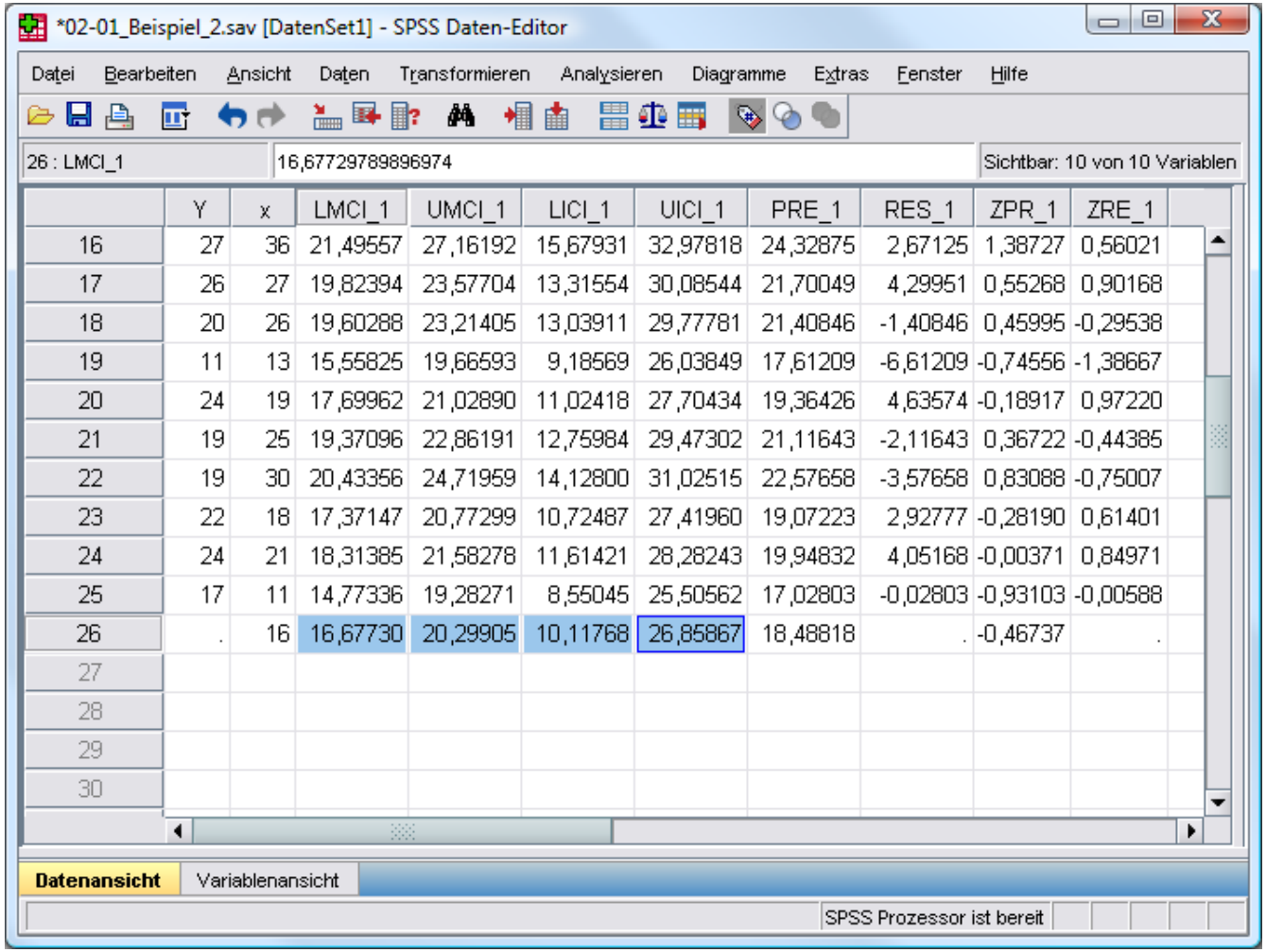
$$[16.677, 20.299]$$

(2) Gesucht ist ein 90% Konfidenzintervall für eine neue Beobachtung der Stelle $x = 16$

- $t_{23,0.95} = 1.714$,
  $S^2_{y|x} = 22.737, \hat{s}^2_{\tilde{y}(x)} = 23.85, \hat{y}(16) = \hat{b}_0 + 16\hat{b}_1 = 18.49$
- Das 90% Konfidenzintervall für eine neue Beobachtung an der Stelle 16 ist gegeben durch

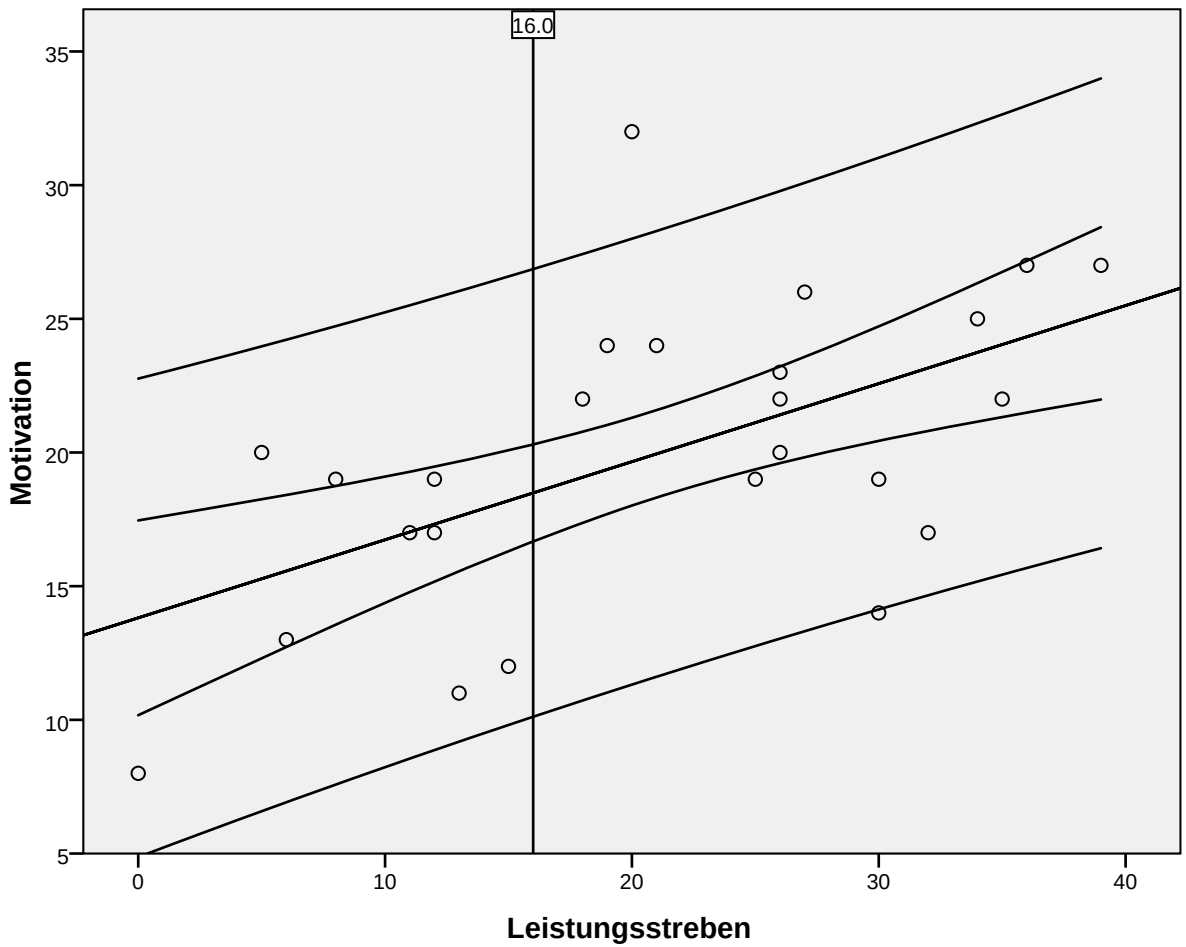
$$[10.118, 26.859]$$

SPSS Output: Vorhersagen bei linearer Regression in Beispiel 2.1 (schwierig)



	Y	x	LMCI_1	UMCI_1	LICI_1	UICI_1	PRE_1	RES_1	ZPR_1	ZRE_1
16	27	36	21,49557	27,16192	15,67931	32,97818	24,32875	2,67125	1,38727	0,56021
17	26	27	19,82394	23,57704	13,31554	30,08544	21,70049	4,29951	0,55268	0,90168
18	20	26	19,60288	23,21405	13,03911	29,77781	21,40846	-1,40846	0,45995	-0,29538
19	11	13	15,55825	19,66593	9,18569	26,03849	17,61209	-6,61209	-0,74556	-1,38667
20	24	19	17,69962	21,02890	11,02418	27,70434	19,36426	4,63574	-0,18917	0,97220
21	19	25	19,37096	22,86191	12,75984	29,47302	21,11643	-2,11643	0,36722	-0,44385
22	19	30	20,43356	24,71959	14,12800	31,02515	22,57658	-3,57658	0,83088	-0,75007
23	22	18	17,37147	20,77299	10,72487	27,41960	19,07223	2,92777	-0,28190	0,61401
24	24	21	18,31385	21,58278	11,61421	28,28243	19,94832	4,05168	-0,00371	0,84971
25	17	11	14,77336	19,28271	8,55045	25,50562	17,02803	-0,02803	-0,93103	-0,00588
26	.	16	16,67730	20,29905	10,11768	26,85867	18,48818	.	-0,46737	.
27										
28										
29										
30										

SPSS Output: Konfidenzintervalle für Vorhersagen bei linearer Regression in Beispiel 2.1



2.21 Residuenanalyse

- Unter der Modellannahme des linearen Regressionsmodells gilt: die Größen

$$\varepsilon_i = Y_i - b_0 - b_1 x_i$$

sind unabhängig und normalverteilt mit Erwartungswert 0 und Varianz $\sigma^2 > 0$.

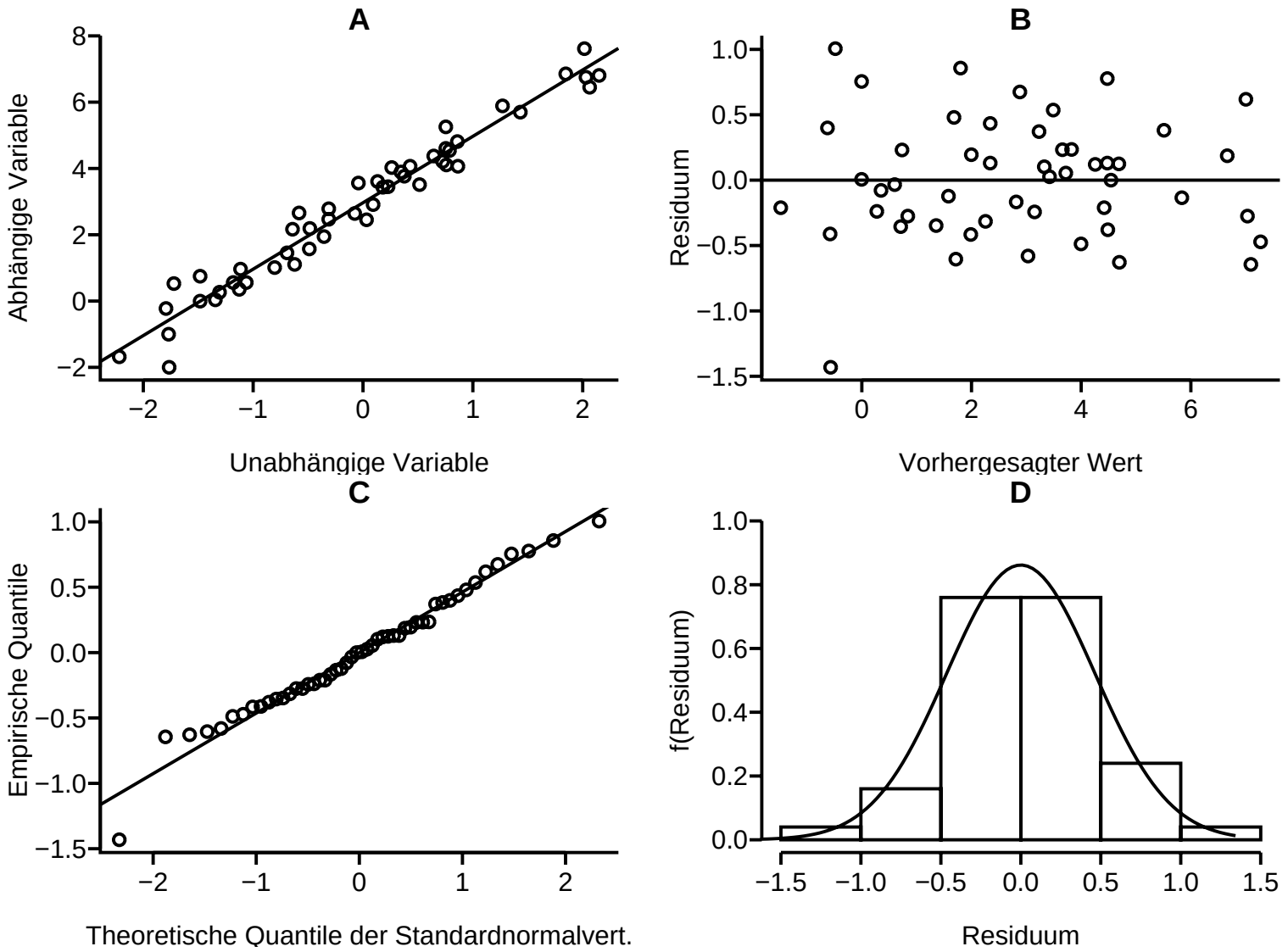
- Das bedeutet, dass diese Eigenschaften auch "näherungsweise" für die Residuen

$$\hat{\varepsilon}_i = y_i - \hat{b}_0 - \hat{b}_1 x_i$$

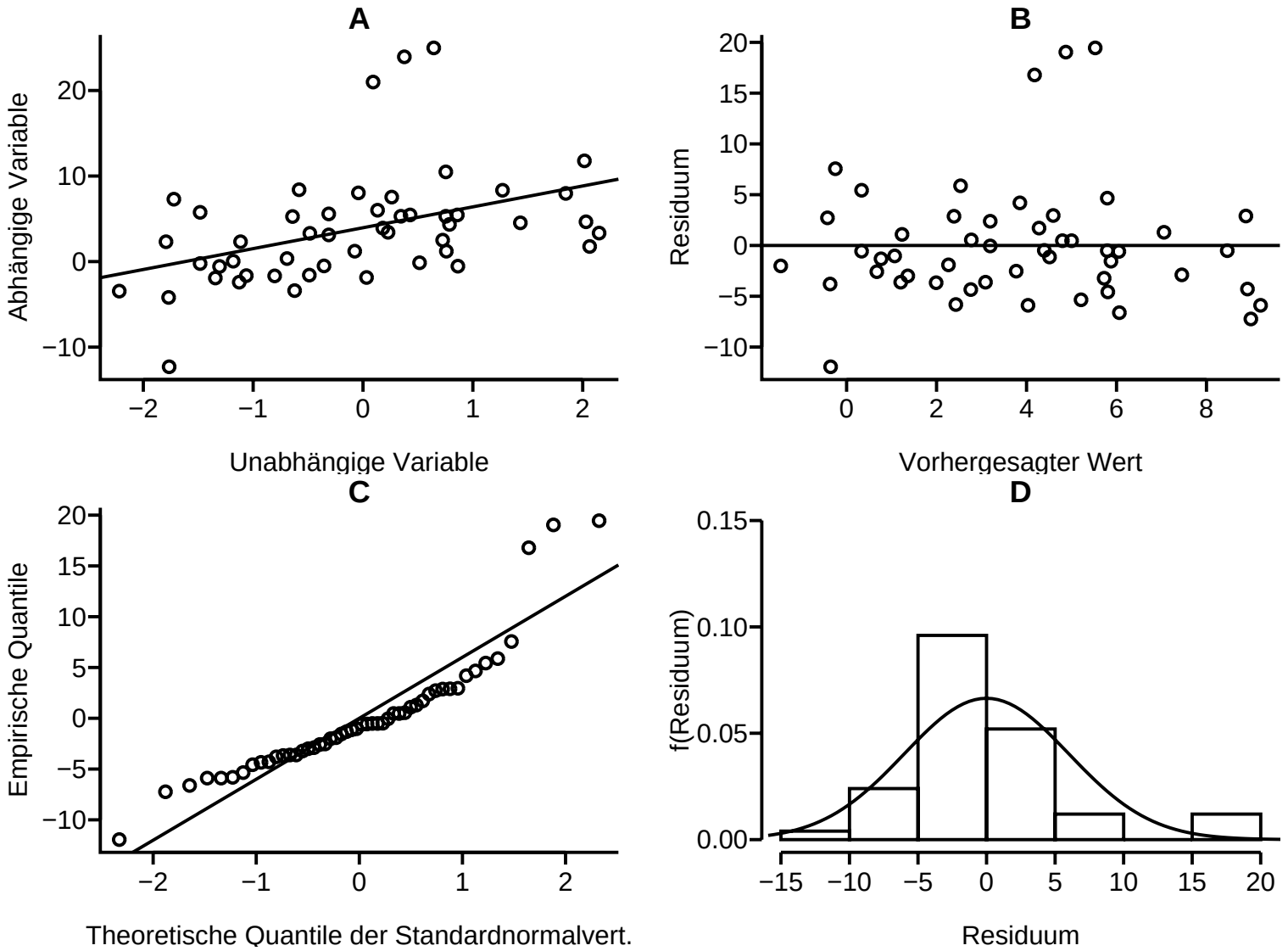
erfüllt sein sollte, falls die Modellannahme zutrifft.

- Residuenanalyse ist ein deskriptives Verfahren für die Überprüfung der Annahmen an $\varepsilon_1, \ldots, \varepsilon_n$ mit 4 Teilschritten (oft werden auch nicht alle gemacht):
  - A: Das Streudiagramm der Daten mit der Regressionslinie
  - B: Ein Streudiagramm der Residuen gegen die vorhergesagten Werte
  - C: Normalverteilungs-QQ-Plot der Residuen
  - D: Histogramm der Residuen mit angepasster Normalverteilungsdichte

Residuenanalyse bei “erfüllten” Voraussetzungen

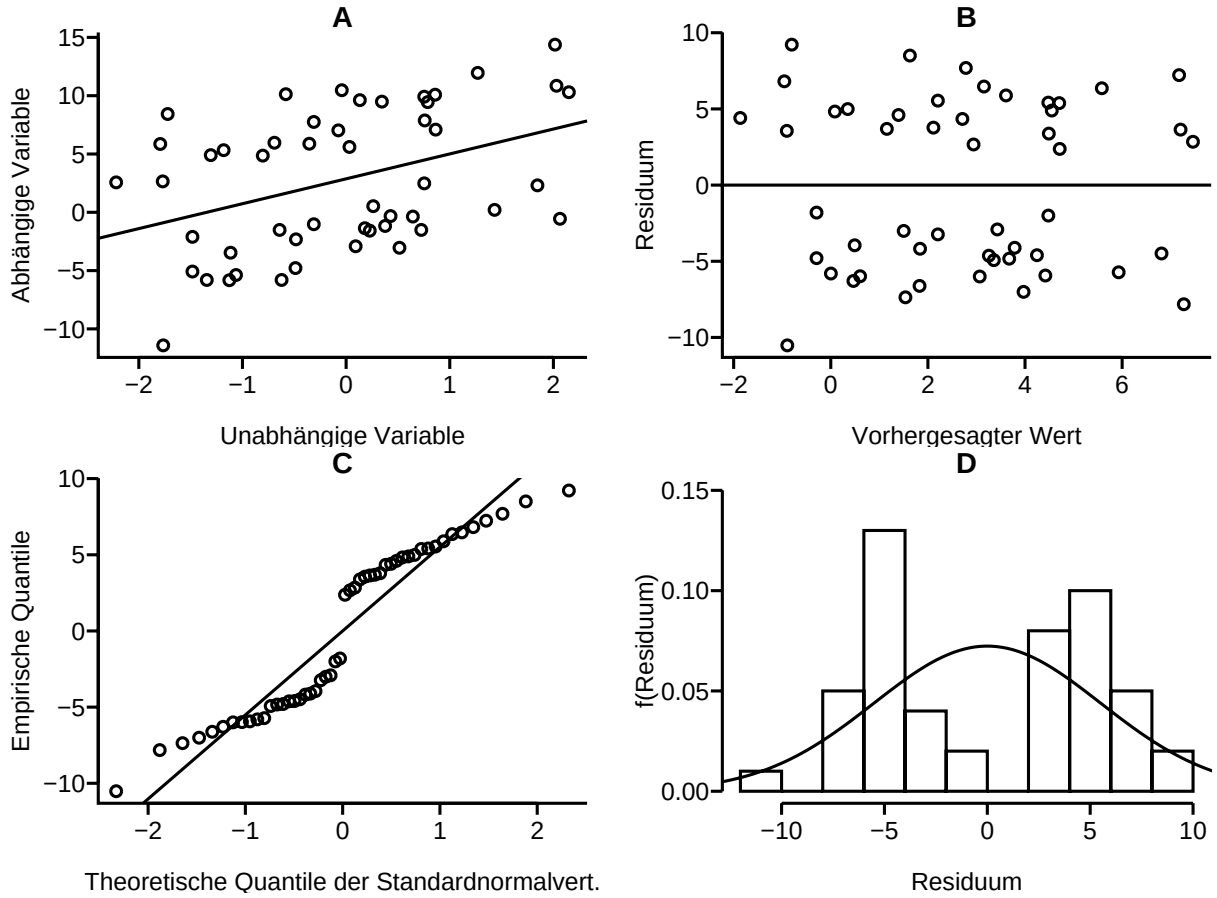


Residuenanalyse bei “Abweichungen” von der Normalverteilung (Ausreißer)

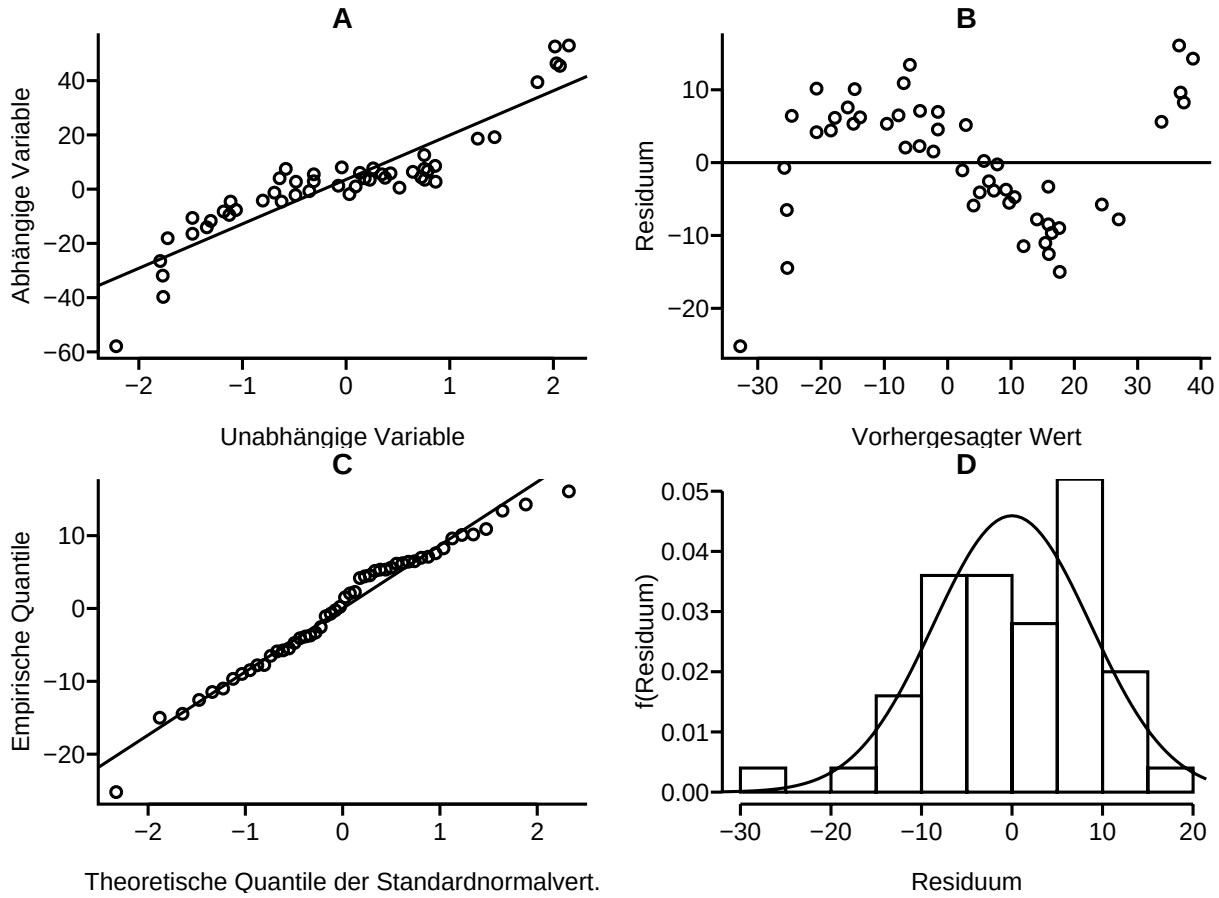


Residuenanalyse bei Stratifizierung

Beachte: verschiedene Untergruppen (**Strata**) können ebenfalls zu Abweichungen von den Modellannahmen führen. Für die Strata können dann unterschiedliche Regressionsgleichungen gelten.

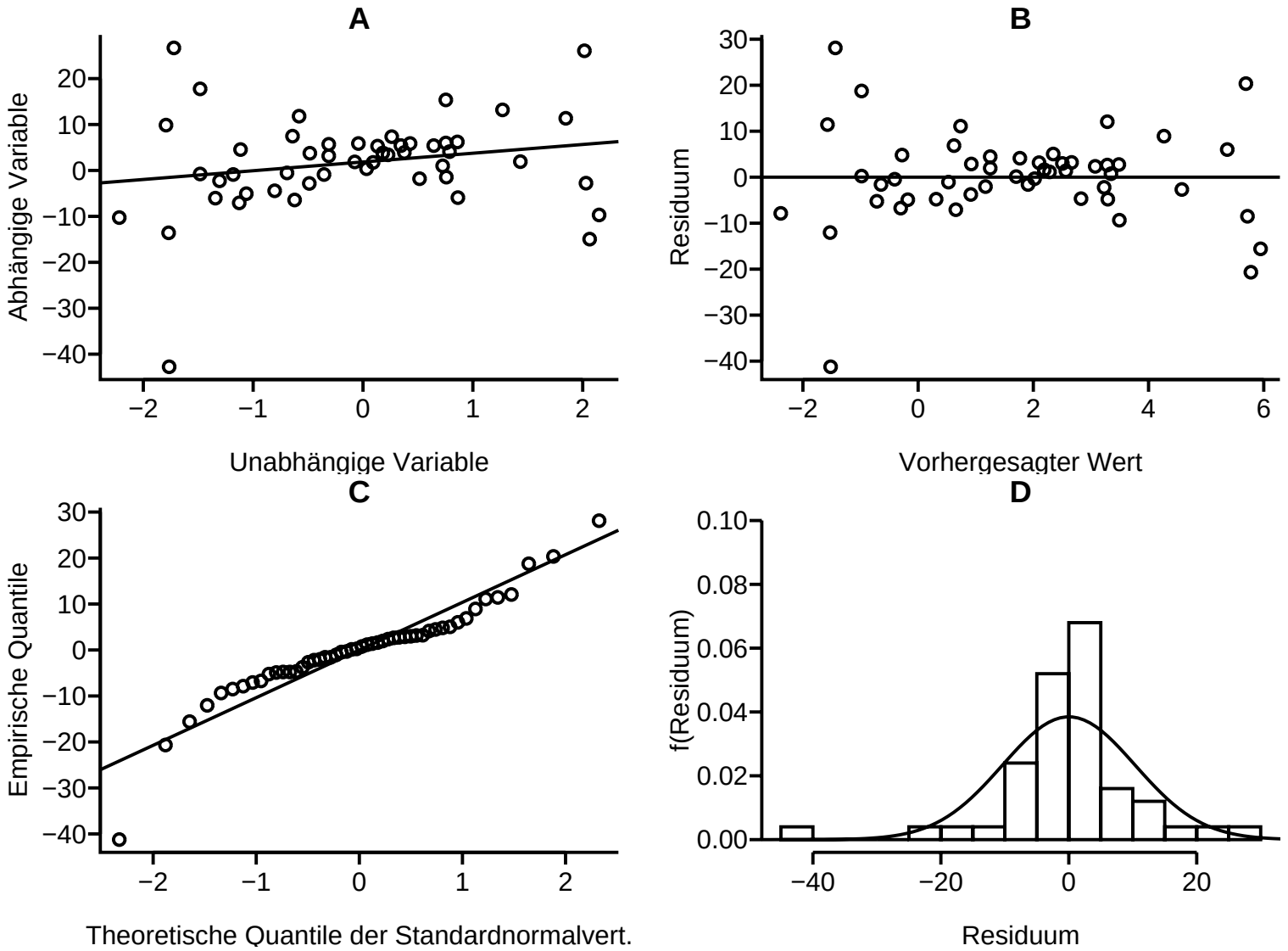


Residuenanalyse bei falscher Modellannahme

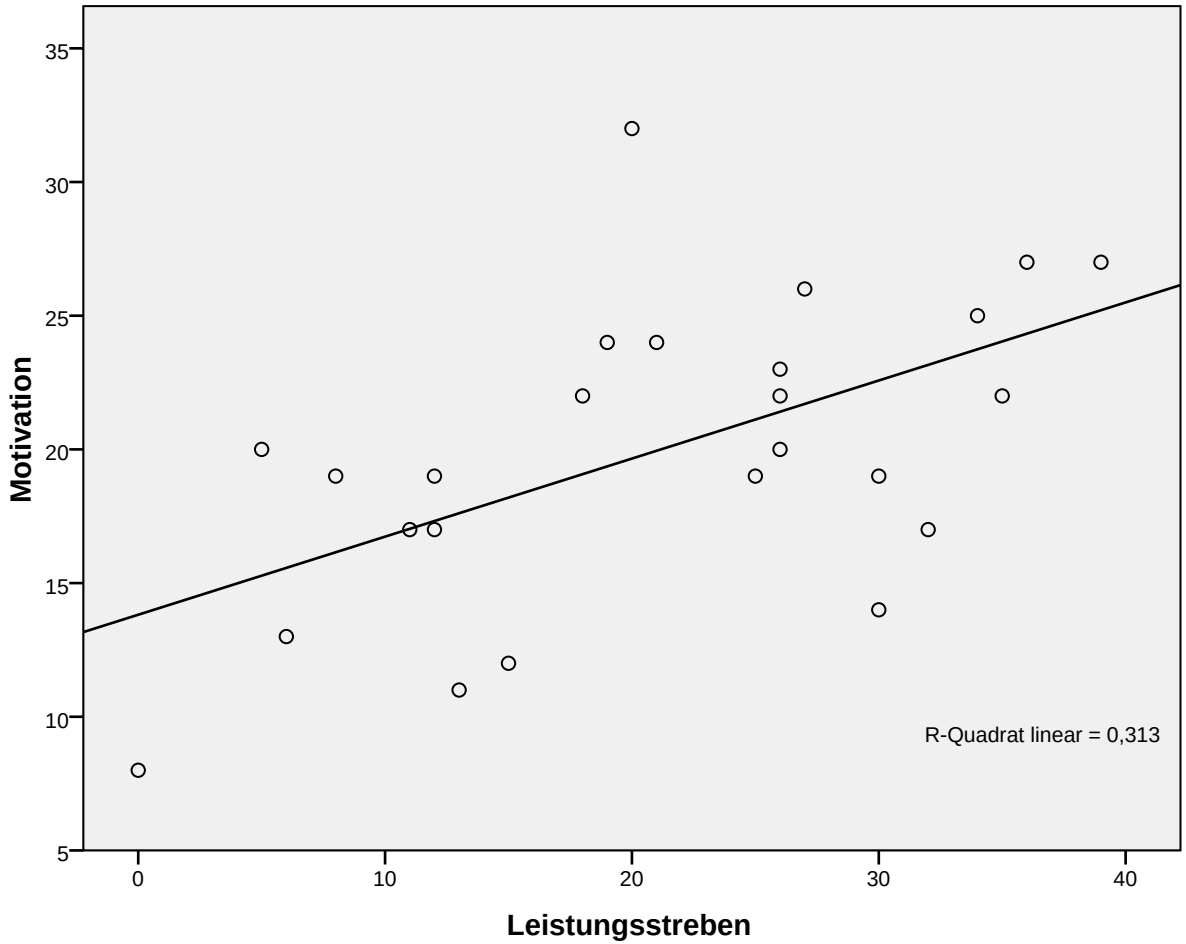


Statt des linearen Modells wäre ein Polynom 3. Grades die bessere Annahme für die Beschreibung des funktionalen Zusammenhangs!

Residuenanalyse bei ungleichen Varianzen (Heteroskedastizität)

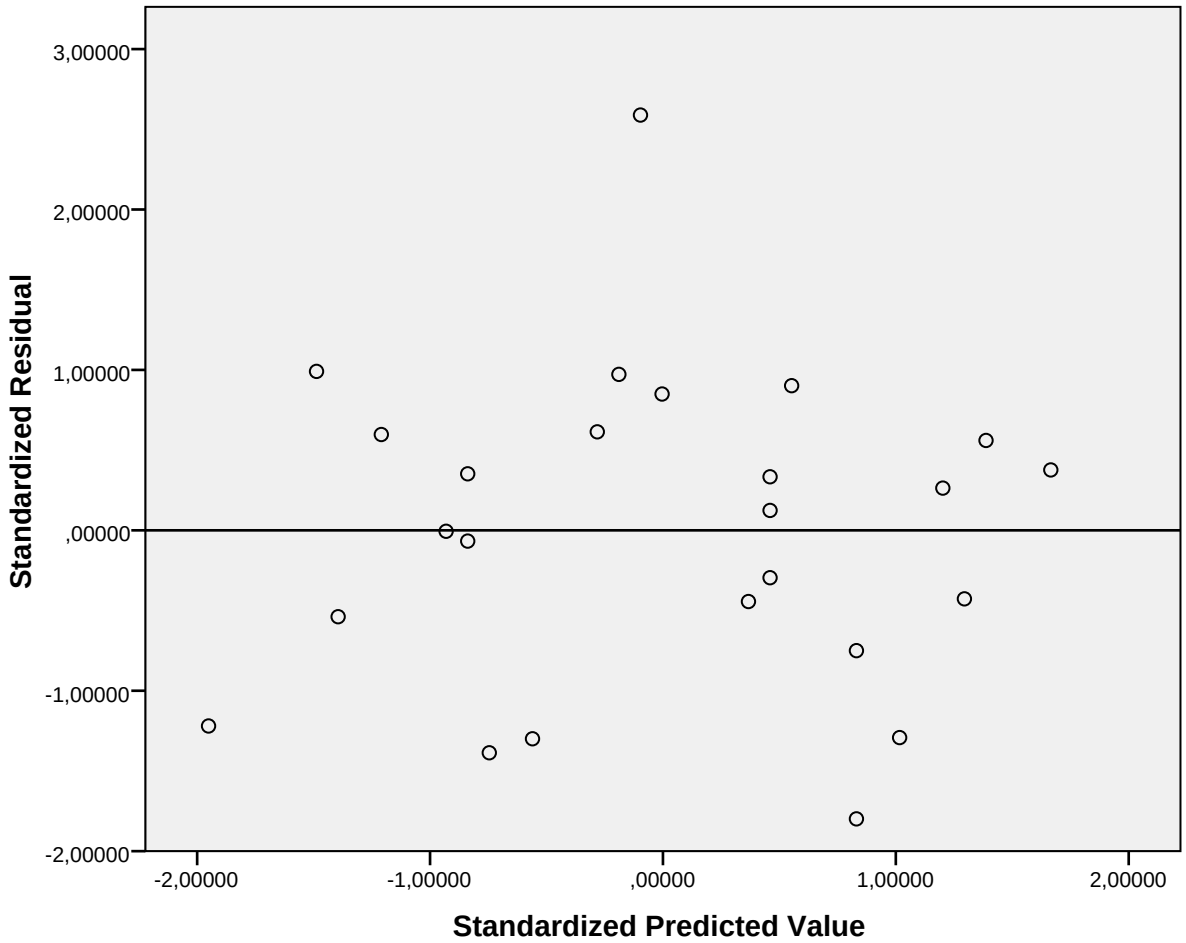


SPSS Output: Residuenanalyse in Beispiel 2.1



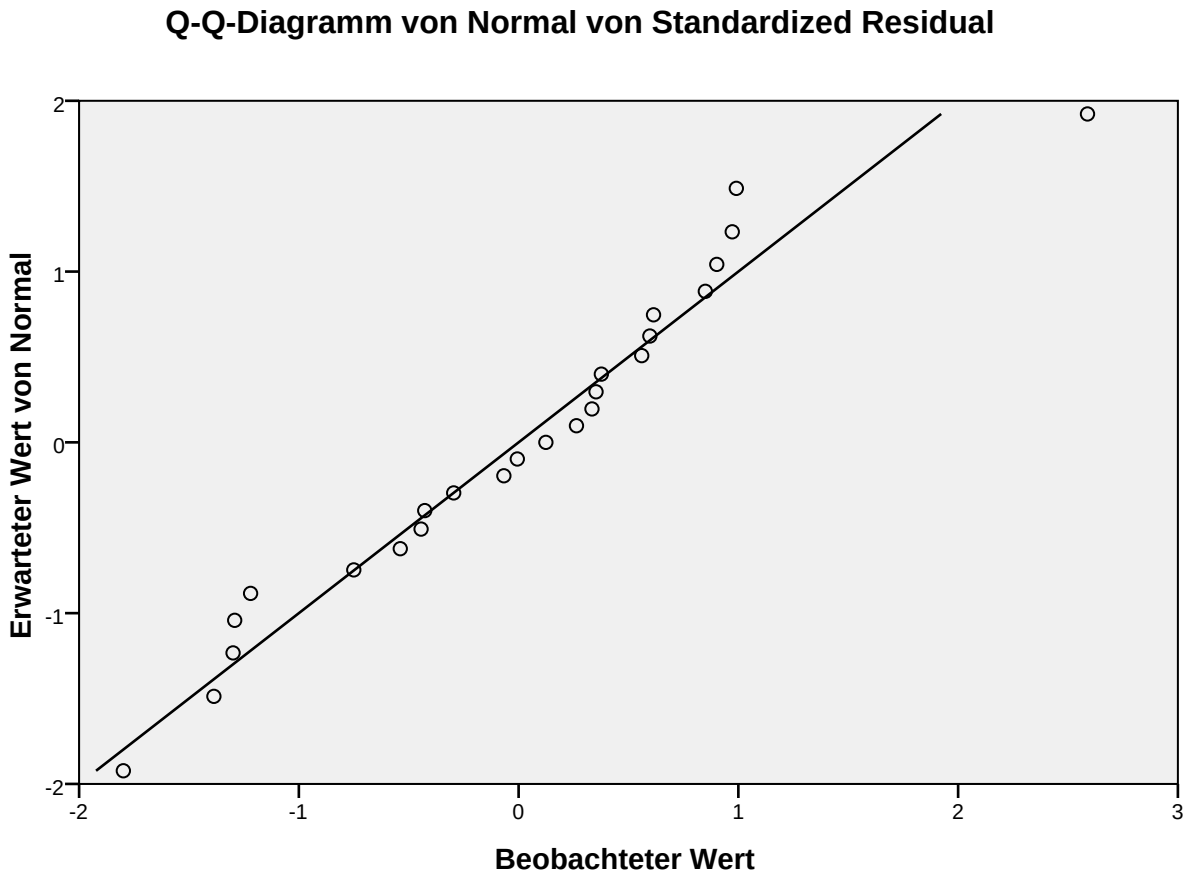
Streudiagramm und geschätzte Regressionsgerade im Beispiel der Arbeitsmotivation

SPSS Output: Residuenanalyse in Beispiel 2.1



Streudiagramm der Residuen gegen die vorhergesagten Werte im Beispiel der Arbeitsmotivation

SPSS Output für Residuenanalyse



QQ-Plot im Beispiel der Arbeitsmotiovation

Korrelation und lineare Regression

Es besteht ein enger Zusammenhang zwischen linearer Regression und Korrelation

- Ist $\hat{b}_1$ die Schätzung im linearen Regressionsmodell und $\hat{\rho}_{x,y}$ der Korrelationskoeffizient von Pearson, dann gilt:

$$\hat{\rho}_{x,y} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x}_.)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_.)^2}} \cdot \hat{b}_1$$

- Ist R^2 das Bestimmtheitsmaß und $\hat{\rho}_{x,y}$ der Korrelationskoeffizient von Pearson, dann gilt:

$$\hat{\rho}_{x,y}^2 = R^2$$

2.3 Multiple lineare Regression

2.22 Beispiel: "Arbeitsmotivation mit mehreren Prädiktoren"

y: Motivation (Einschätzung der Arbeitsmotivation durch Experten)

Prädiktoren: *Eigenschaften*

- $x1$: Ehrgeiz (Fragebogen)
- $x2$: Kreativität (Fragebogen)
- $x3$: Leistungsstreben (Fragebogen)

Prädiktoren: *Rahmenbedingungen*

- $x4$: Hierarchie (Position in der Hierarchie des Unternehmens)
- $x5$: Lohn (Bruttolohn pro Monat)
- $x6$: Arbeitsbedingungen (Zeitsouveränität, Kommunikationsstruktur usw.)

Prädiktoren: *Inhalte der Tätigkeit*

- $x7$: Lernponential (Lernpotential der Tätigkeit)
- $x8$: Vielfalt (Vielfalt an Teiltätigkeiten)
- $x9$: Anspruch (Komplexität der Tätigkeit)

Daten

i	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1	32	36	30	20	20	3100	34	29	69	66
2	14	30	11	30	7	2600	39	16	47	36
3	12	19	15	15	8	3200	42	13	32	17
4	27	42	16	39	13	2500	43	15	63	49
5	20	14	22	5	22	3700	42	29	38	62
6	13	12	16	6	11	2600	36	17	39	51
7	17	17	20	12	11	2500	41	18	44	55
8	8	4	5	0	16	3800	23	9	31	33
9	22	32	20	35	20	3500	25	21	40	55
10	19	15	13	8	13	3100	29	21	57	56
11	25	38	5	34	21	3600	59	27	53	67
12	23	24	6	26	9	2600	45	31	54	62

Daten

i	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
13	17	28	11	32	10	2600	30	7	45	26
14	22	36	4	26	16	2500	52	23	56	64
15	19	18	26	12	6	2500	40	17	54	55
16	27	40	27	36	12	2500	42	29	44	62
17	26	30	28	27	18	3000	38	34	43	64
18	20	27	11	26	10	2600	35	19	46	55
19	11	18	23	13	11	2800	42	18	31	43
20	24	32	18	19	15	2700	48	23	51	53
21	19	33	9	25	6	2400	38	23	37	65
22	19	33	22	30	5	2600	36	30	39	39
23	22	27	28	18	17	4000	45	23	52	54
24	24	30	32	21	11	2700	44	20	41	47
25	17	37	8	11	2	2300	32	20	44	41

2.23 Das Modell der multiplen linearen Regression

- Daten $(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)$
- Es gibt k unabhängige Variablen: $\mathbf{x_i} = (x_{1i}, ..., x_{ki})$
- y_i ist Realisation einer Zufallsvariablen Y_i (unter der Bedingung $\mathbf{x_i}$). Für den Zusammenhang zwischen der Variablen Y und dem Vektor $\mathbf{x_1}$ gilt (im Beispiel ist $k = 9$):

$$\begin{aligned} Y_i &= b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki} + \varepsilon_i \\ &= b_0 + \sum_{j=1}^{k} b_j x_{ji} + \varepsilon_i. \end{aligned}$$

- ε_i bezeichnet hier eine zufällige "Störung" und es wird angenommen, dass die Störungen $\varepsilon_1, \ldots, \varepsilon_n$ **unabhängig** und **normalverteilt** sind mit Erwartungswert 0 und Varianz $\sigma^2 > 0$
- Deutung: es wird ein linearer Zusammenhang zwischen $\mathbf{x}$ und Y postuliert, der noch zufälligen Störungen unterliegt

2.24 Schätzung bei multipler linearer Regression

- **Methode der kleinsten Quadrate**: Minimiere

$$\sum_{i=1}^{n}(y_i - b_0 - b_1 x_{1i} - \cdots - b_k x_{ki})^2$$

bzgl. der Wahl von $b_0, \ldots, b_k$

- Mathematische Statistik (allgemeines lineares Modell) liefert Schätzer

$$\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_k$$

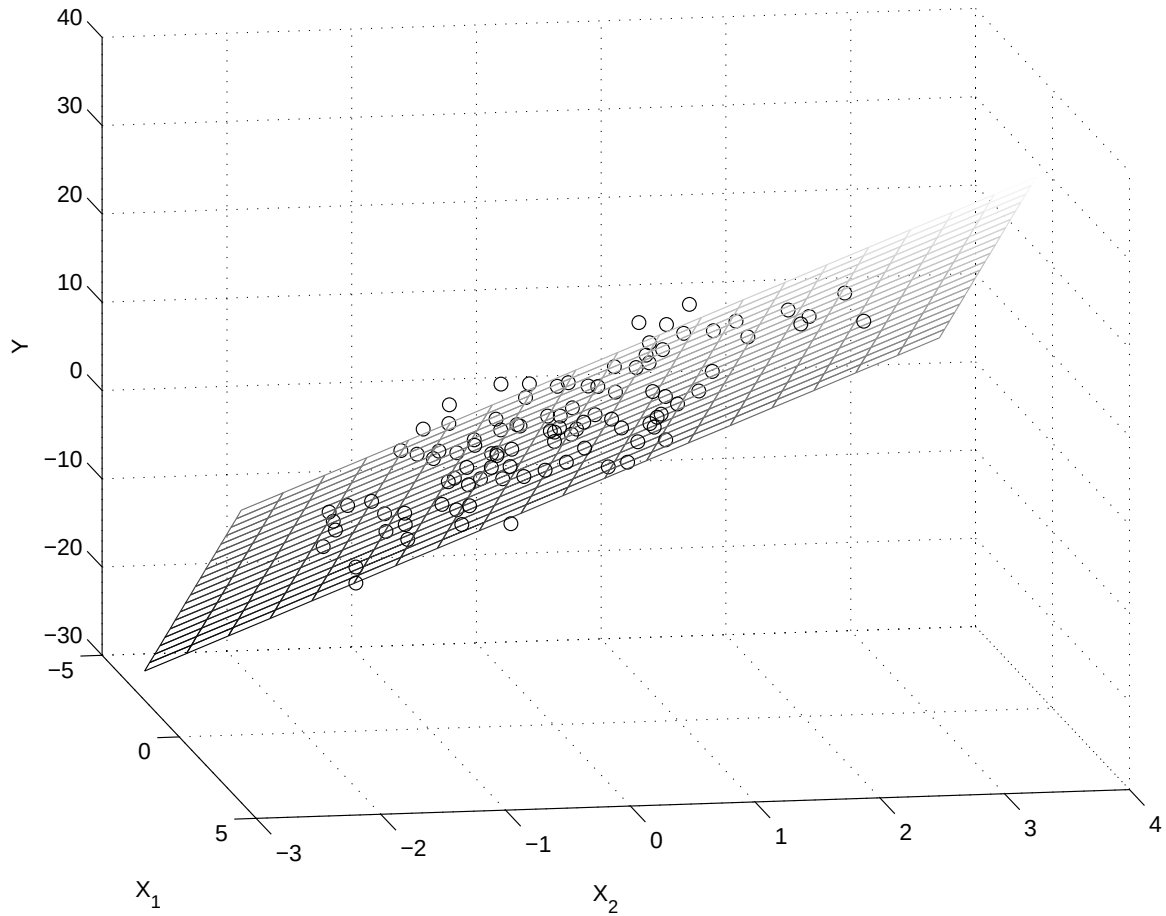
für die Parameter $b_0, \ldots, b_k$ (Formeln sind kompliziert)

- Schätzer für die Varianz der Meßfehler

$$S^2_{y|x} = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{b}_0 - \hat{b}_1 x_{1i} - \cdots - \hat{b}_k x_{ki})^2$$

Streudiagramm bei multipler linearer Regression ($k = 2$)

Regressionsfläche: $\widehat{y}(\mathbf{x}) = 3.24 + 4.5x_1 + 5.27x_2$.



Fortsetzung von Beispiel 2.22: Schätzer im multiplen linearen Regressionsmodell

- Ergebnisse für die Schätzer im multiplen linearen Regressionsmodel

$$\begin{aligned} \hat{b}_0 &= -3.842 & \hat{b}_1 &= 0.193 \\ \hat{b}_2 &= 0.153 & \hat{b}_3 &= 0.049 \\ \hat{b}_4 &= 0.246 & \hat{b}_5 &= 0.000 \\ \hat{b}_6 &= -0.031 & \hat{b}_7 &= 0.165 \\ \hat{b}_8 &= 0.206 & \hat{b}_9 &= -0.053 \end{aligned}$$

- **Fragen:**
  - Wie genau sind diese Schätzungen?
  - Besteht ein (signifikanter) Einfluß der unabhängigen Merkmale auf die Motivation

$$\begin{aligned} H_0 : b_1 &= 0 \\ H_0 : b_2 &= 0 \\ &\vdots \end{aligned}$$

  - Wie gut beschreibt das multiple lineare Regressionsmodell die Situation?

Genauigkeit der Schätzung bei multipler linearer Regression

- Schätzer $\hat{s}_{b_0}, \ldots, \hat{s}_{b_k}$ für die Standardfehler von $\hat{b}_0, \ldots, \hat{b}_k$ sind verfügbar (Allgemeines lineares Modell $\rightarrow$ Formeln kompliziert)
- Anmerkung: Für wachsenden Stichprobenumfang konvergieren die Schätzer $\hat{s}_{b_j}$ gegen 0 *"je größer der Stichprobenumfang, desto genauer die Schätzungen"*
- Damit erhält man Konfidenzintervalle für $b_0, \ldots, b_k$, z.B.

$$(\hat{b}_0 - t_{n-k-1,1-\alpha/2}\ \hat{s}_{b_0}, \hat{b}_0 + t_{n-k-1,1-\alpha/2}\ \hat{s}_{b_0})$$

ist $(1-\alpha)$-Konfidenzintervall für b_0

Fortsetzung von Beispiel 2.22: Schätzer für den Standardfehler der Schätzer im multiplen linearen Regressionsmodell

- Ergebnisse für den Standardfehler der Schätzer im multiplen linearen Regressionsmodell

$$\begin{array}{llll} \hat{s}_{b_0} = & 5.052 & \hat{s}_{b_1} = & 0.081 \\ \hat{s}_{b_2} = & 0.049 & \hat{s}_{b_3} = & 0.065 \\ \hat{s}_{b_4} = & 0.148 & \hat{s}_{b_5} = & 0.001 \\ \hat{s}_{b_6} = & 0.054 & \hat{s}_{b_7} = & 0.098 \\ \hat{s}_{b_8} = & 0.052 & \hat{s}_{b_9} = & 0.058 \end{array}$$

- Wegen $t_{15,0.975} = 2.1314$ ist

$$[-0.089, 0.188]$$

ein 95%-Konfidenzintervall für den Parameter b_3. Man beachte:

  - $0.049 + 2.1314 \cdot 0065 \approx 0.188$)
  - $n = 25;\ k = 9 \Rightarrow n - k - 1 = 15$

2.25 Konfidenzintervalle für multiple lineare Regression

- Modellannahme: multiple lineare Regression

$$Y_i = b_0 + \sum_{j=1}^{k} b_j x_{ji} + \varepsilon_i \quad (i = 1, \ldots, n)$$

- Rechterfertigung der Normalverteilungs- und Unabhängigkeitsannahme
- Schätzer $\hat{s}_{b_j}$ für den Standardfehler von $\hat{b}_j$

$$\Longrightarrow \qquad (\hat{b}_j - t_{n-k-1,1-\alpha/2}\hat{s}_{b_j}, \hat{b}_j + t_{n-k-1,1-\alpha/2}\hat{s}_{b_j})$$

  ist ein $(1-\alpha)$ -Konfidenzintervall für b_j $(j = 0, \ldots k)$
- $t_{n-k-1,1-\alpha/2}$; $(1-\alpha/2)$-Quantil der t-Verteilung mit $n-k-1$ Freiheitsgraden (Tabelle oder Software)
- Anmerkung: Für wachsenden Stichprobenumfang konvergieren die Schätzer $\hat{s}_{b_j}$ gegen 0 *"je größer der Stichprobenumfang, desto kleiner die Konfidenzintervalle"*

2.26 Beispiel: Konfidenzintervalle für die Parameter in Beispiel 2.22 (Arbeitsmotivation)

$\hat{b}_j$	Merkmal	Schätzung	$\hat{s}_{b_j}$	Konfidenzintervall
$\hat{b}_0$	—	-3.842	5.052	[-14.609, 6.926]
$\hat{b}_1$	Ehrgeiz	0.193	0.081	[0.020, 0.365]
$\hat{b}_2$	Kreativität	0.153	0.049	[0.049, 0.258]
$\hat{b}_3$	Leistungsstreben	0.049	0.065	[-0.089, 0.188]
$\hat{b}_4$	Hierarchie	0.246	0.148	[-0.069, 0.561]
$\hat{b}_5$	Lohn	0.000	0.001	[-0.004, 0.002]
$\hat{b}_6$	Arbeitsbdg.	-0.031	0.054	[-0.147, 0.085]
$\hat{b}_7$	Lernpotential	0.165	0.098	[-0.044, 0.373]
$\hat{b}_8$	Vielfalt	0.206	0.052	[0.095, 0.316]
$\hat{b}_9$	Anspruch	0.053	0.058	[-0.070, 0.177]

SPSS Output: Schätzer, Standardabweichung und Konfidenzintervalle im Beispiel 2.22 (Arbeitsmotivation mit mehreren Prädiktoren)

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	95%-Konfidenzintervall für B	
		B	Standard fehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	-3,842	5,052		-,760	,459	-14,609	6,926
	x1	,193	,081	,337	2,381	,031	,020	,365
	x2	,153	,049	,234	3,127	,007	,049	,258
	x3	,049	,065	,095	,761	,458	-,089	,188
	x4	,246	,148	,235	1,664	,117	-,069	,561
	x5	,000	,001	-,077	-,589	,564	-,004	,002
	x6	-,031	,054	-,045	-,576	,573	-,147	,085
	x7	,165	,098	,199	1,683	,113	-,044	,373
	x8	,206	,052	,354	3,973	,001	,095	,316
	x9	,053	,058	,124	,920	,372	-,070	,177

a. Abhängige Variable: Y

2.27 Vorhersage der multiplen linearen Regression

- Modellannahme: multiple lineare Regression

  $$Y_i = b_0 + \sum_{j=1}^{k} b_j x_{ji} + \varepsilon_i \quad (i = 1, \ldots, n)$$

- Rechterfertigung der Normalverteilungs- und Unabhängigkeitsannahme
- Vorhersage für den Wert der multiplen Regression an der Stelle $\mathbf{x} = (x_1, \ldots, x_k)$ (im Beispiel ist $k = 9$)

  $$\hat{y}(\mathbf{x}) = \hat{b}_0 + \sum_{j=1}^{k} \hat{b}_j x_j$$

- In Beispiel 2.22 ergibt sich z.B. als Vorhersage der multiplen linearen Regression an der Stelle

  $$\begin{aligned} x_1 &= 21,\ x_2 = 30,\ x_3 = 15,\ x_4 = 11,\ ,\ x_5 = 2900, \\ x_6 &= 41,\ x_7 = 25,\ x_8 = 55,\ x_9 = 54 \end{aligned}$$

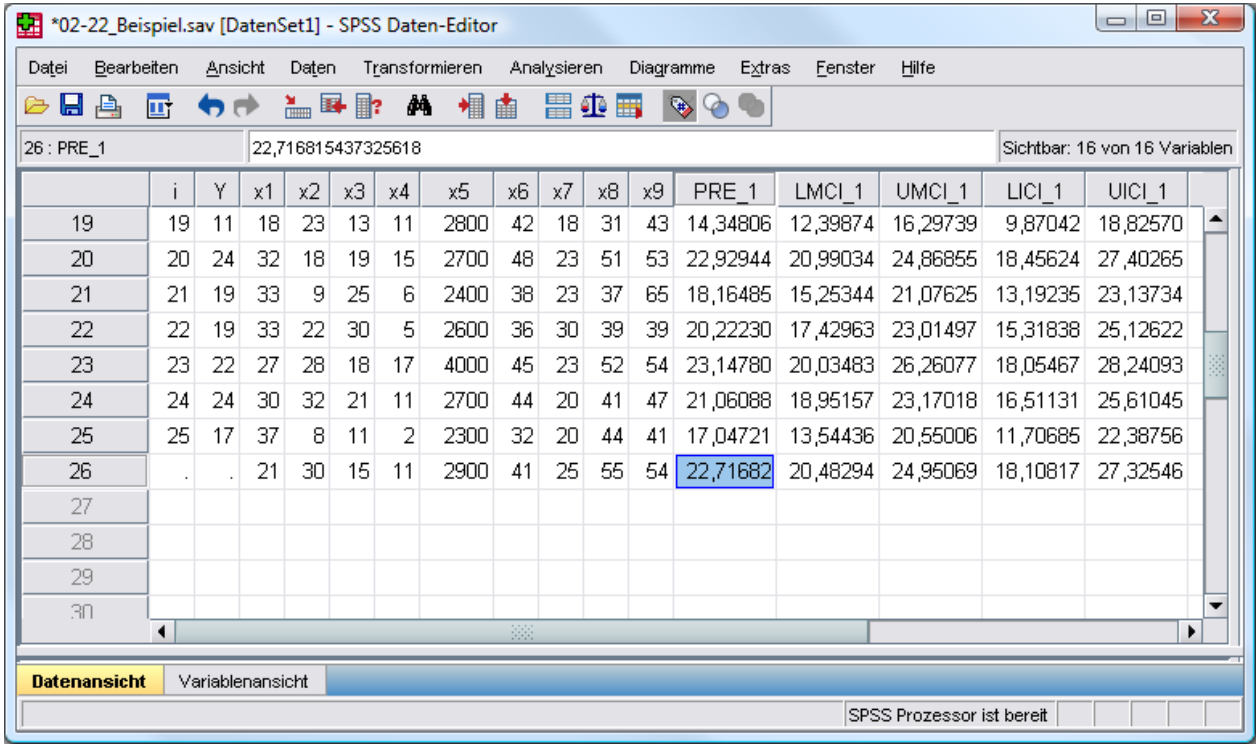
  der Wert $\hat{y}(\mathbf{x}) = 22.717$

Vorhersage der multiplen linearen Regression

Beachte: Wie in Abschnitt 2.18 und 2.19 gibt es zwei Vorhersagen:

- Vorhersage für *den Wert der multiplen Regression* an der Stelle $\mathbf{x} = (x_1, \ldots, x_k)$ (im Beispiel ist $k = 9$)
- Vorhersage für *den Wert einer neuen Beobachtung* an der Stelle $\mathbf{x} = (x_1, \ldots, x_k)$ (im Beispiel ist $k = 9$)
- Für beide Vorhesagen, kann man den Standardfehler bestimmen (Formeln kompliziert) und Konfidenzbereiche angeben (vgl. Abschnitt 2.18 und 2.19 für den Fall $k = 1$)

SPSS Output: Vorhersage bei der multiplen linearen Regression (schwierig)

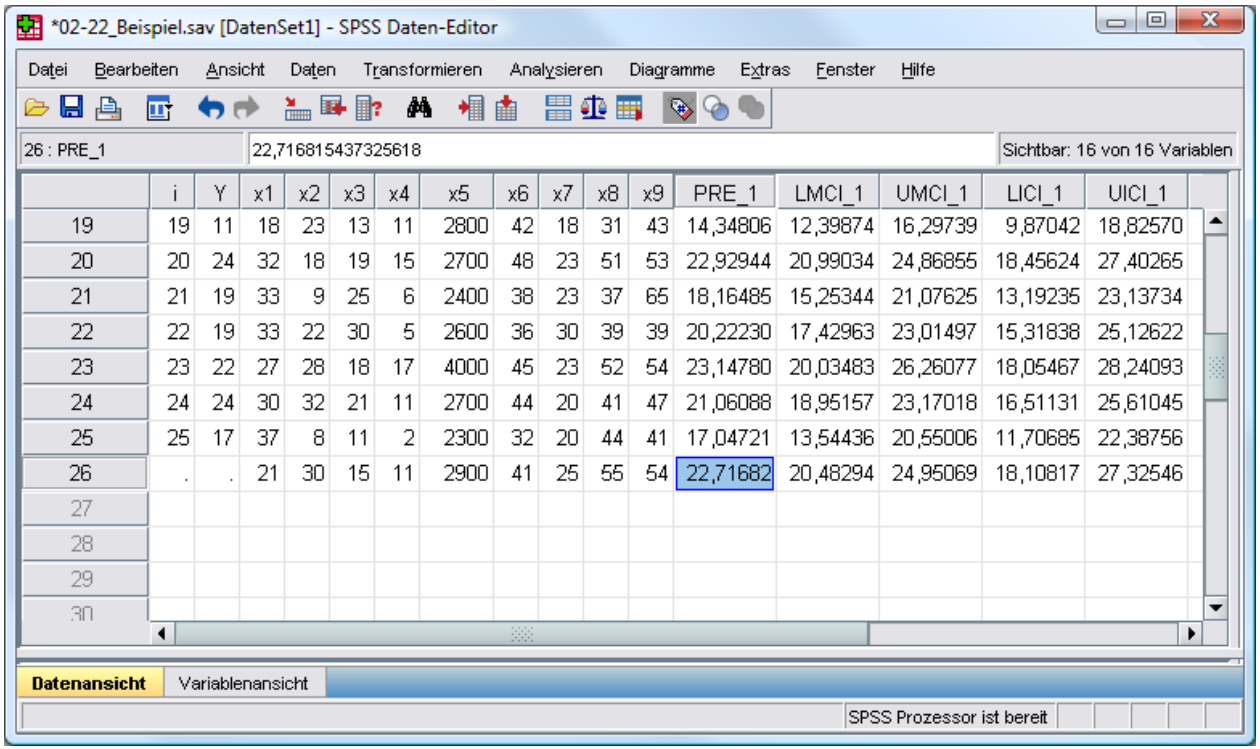


	i	Y	x1	x2	x3	x4	x5	x6	x7	x8	x9	PRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1
19	19	11	18	23	13	11	2800	42	18	31	43	14,34806	12,39874	16,29739	9,87042	18,82570
20	20	24	32	18	19	15	2700	48	23	51	53	22,92944	20,99034	24,86855	18,45624	27,40265
21	21	19	33	9	25	6	2400	38	23	37	65	18,16485	15,25344	21,07625	13,19235	23,13734
22	22	19	33	22	30	5	2600	36	30	39	39	20,22230	17,42963	23,01497	15,31838	25,12622
23	23	22	27	28	18	17	4000	45	23	52	54	23,14780	20,03483	26,26077	18,05467	28,24093
24	24	24	30	32	21	11	2700	44	20	41	47	21,06088	18,95157	23,17018	16,51131	25,61045
25	25	17	37	8	11	2	2300	32	20	44	41	17,04721	13,54436	20,55006	11,70685	22,38756
26	.	.	21	30	15	11	2900	41	25	55	54	22,71682	20,48294	24,95069	18,10817	27,32546
27																
28																
29																
30																

Beispiel:

- Schätzung für den Wert der "Ebene" an der Stelle $\mathbf{x} = (18, 23, 13, 11, 2800, 42, 18, 31, 43)$: 14.348
- Schätzung für eine weitere Beobachtung an der Stelle $\mathbf{x} = (18, 23, 13, 11, 2800, 42, 18, 31, 43)$: 14.348

SPSS Output: Konfidenzintervalle für Vorhersagen bei mulitpler linearer Regression



	i	Y	x1	x2	x3	x4	x5	x6	x7	x8	x9	PRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1
19	19	11	18	23	13	11	2800	42	18	31	43	14,34806	12,39874	16,29739	9,87042	18,82570
20	20	24	32	18	19	15	2700	48	23	51	53	22,92944	20,99034	24,86855	18,45624	27,40265
21	21	19	33	9	25	6	2400	38	23	37	65	18,16485	15,25344	21,07625	13,19235	23,13734
22	22	19	33	22	30	5	2600	36	30	39	39	20,22230	17,42963	23,01497	15,31838	25,12622
23	23	22	27	28	18	17	4000	45	23	52	54	23,14780	20,03483	26,26077	18,05467	28,24093
24	24	24	30	32	21	11	2700	44	20	41	47	21,06088	18,95157	23,17018	16,51131	25,61045
25	25	17	37	8	11	2	2300	32	20	44	41	17,04721	13,54436	20,55006	11,70685	22,38756
26	.	.	21	30	15	11	2900	41	25	55	54	22,71682	20,48294	24,95069	18,10817	27,32546
27																
28																
29																
30																

- Konfidenzintervall für den Wert der "Ebene" an der Stelle $\mathbf{x} = (18, 23, 13, 11, 2800, 42, 18, 31, 43)$: $[12.399, 16.297]$
- Konfidenzintervall für eine weitere Beobachtung an der Stelle $\mathbf{x} = (18, 23, 13, 11, 2800, 42, 18, 31, 43)$: $[9.870, 18.826]$

2.28 Bestimmtheitsmaß bei multipler linearer Regression

- Modellvohersage:

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{1i} + \ldots \hat{b}_k x_{ki} = \hat{b}_0 + \sum_{j=1}^{k} \hat{b}_j x_{ji}$$

- Residuum $\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{b}_0 + \sum_{j=1}^{k} \hat{b}_j x_{ji})$
- **Beachte:** Die Werte der abhängigen Variable zerfallen in Modellvorhersage ($\hat{y}$) und Residuum ($\hat{\varepsilon}$), d.h.

$$y_i = \hat{y}_i + \hat{\varepsilon}_i \qquad i = 1, \ldots, n$$

- Die Güte der Modellanpassung wird (wieder) durch das **Besimmtheitsmaß** R^2 beschrieben werden (Anteil erklärter Varianz)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_.)^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y}_.)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_.)^2} \; .$$

Beispiel: das Bestimmtheitsmaß für das Beispiel 2.22 (Arbeitsmotivation)

In Beispiel 2.22 ist

- $n = 25;\ k = 9$
- $\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = 53.651$
- $\sum_{i=1}^{n}(y_i - \overline{y}_.)^2 = 790.96$
- $$R^2 = 1 - \frac{53.651}{790.96} = 92.95$$

D.h. 92.95% der Varianz der Variablen Motivation werden durch das multiple lineare Regressionsmodell erklärt

2.29 Statistische Tests bei der multiplen linearer Regression. Zwei "wichtige" Fragestellungen:

- **Frage A:** Hat mindestens eine der Prädiktorvariablen $x_1, \ldots, x_k$ einen Einfluß auf die abhängige Variable y **(Gesamttest auf Signifikanz)**.
- Mathematische Formulierung der Hypothese:

  Nullhypothese:

  $$H_0 : b_j = 0 \text{ für alle } j \in \{1, 2, \ldots, k\}$$

  Alternative:

  $$H_1 : b_j \neq 0 \text{ für mindestens ein } j \in \{1, 2, \ldots, k\}$$

- **Frage B:** Hat die Prädiktorvariablen x_j (z.B. Ehrgeiz) einen Einfluß auf die abhängige Variable y.
- Mathematische Formulierung der Hypothese:

  Nullhypothese: $H_0 : b_j = 0$
  Alternative: $H_1 : b_j \neq 0$

2.29(A) Gesamttest auf Signifikanz

- Nullhypothese: $H_0 : b_j = 0$ für alle $j \in \{1, 2, \ldots, k\}$

  Alternative: $H_1 : b_j \neq 0$ für mindestens ein $j \in \{1, 2, \ldots, k\}$

  (1) Bestimme

$$S^2_{reg} = \frac{1}{k} \sum_{i=1}^{n} (\hat{y}_i - \overline{y}_.)^2$$

  die Varianz der Regression, und

$$S^2_{y|\mathbf{x}} = \frac{1}{n-k-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

  die **Residualvarianz**

- **Beachte**: man geht genau wie im linearen Regressionsmodell vor!

2.29(A) Gesamttest auf Signifikanz

(2) H_0 wird zu Gunsten der Alternative H_1 verworfen, falls

$$F_n = \frac{S^2_{reg}}{S^2_{y|x}} > F_{k;n-k-1;1-\alpha}$$

gilt (oder der entsprechende p-Wert kleiner als α ist). Dabei bezeichnet $F_{k;n-k-1;1-\alpha}$ das $(1-\alpha)$ Quantil der F-Verteilung mit $(k, n-k-1)$ Freiheitsgraden,

- **Beachte:** Wird H_0 durch diesen Test verworfen, dann bleibt aber noch unklar, "welches der Merkmale signifikant ist"

2.29(B) Tests für die Signifikanz einzelner Merkmale

Nullhypothese:

$$H_0 : b_j = 0$$

Alternative:

$$H_1 : b_j \neq 0$$

- Die Nullhypothese H_0 wird zu Gunsten der Alternative H_1 verworfen, falls

$$T_n = \left| \frac{\hat{b}_j}{\hat{s}_{b_j}} \right| > t_{n-k-1;1-\alpha/2}$$

gilt (oder der entsprechende p-Wert kleiner als α ist). Dabei ist

  - $t_{n-k-1;1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der t-Verteilung mit $n-k-1$ Freiheitsgraden
  - $\hat{s}_{b_j}$ der Standardfehler von $\hat{b}_j$

- **Beachte:** Werden mehrere Hypothesen gestestet, ist das Niveau entsprechend anzupassen (vgl. Abschntt 1.18).

2.30(A) Test auf Signifikanz im multiplen Regressionsmodel in Beispiel 2.22

- **Frage**: "Hat einer der 9 Prädiktorvariablen einen Einfluß auf die abhängige Variable?"
- Mathematische Hypothesen:

  $H_0 : b_j = 0$ für alle $j = 1, \ldots, 9$

  $H_1 : b_j \neq 0$ für mindestens ein $j \in \{1, \ldots, 9\}$
- $F_n = 21.972, \qquad F_{9,15,0.95} = 2.5876$
- Da $F_n > 21.972 > 2.5876$ ist, wird die Nullhypothese zum Niveau 5% verworfen.

2.30(B) Beispiel: Test auf Signifikanz eines Merkmals im multiplen linearen Regressionsmodell in Beispiel 2.22

- **Frage:** "Hat die Prädiktorvariable Ehrgeiz (x_1) einen Einfluß auf die abhängige Variable Motivation (Signifikanz des Regressionskoeffizienten b_1)?"
- Mathematische Hypothesen:

$$H_0 : b_1 = 0; \qquad H_1 : b_1 \neq 0$$

- $\hat{b}_1 = 0.193, \qquad \hat{s}_{b_1} = 0.081, \qquad t_{25-10,0.975} = 2.13$

$$\Rightarrow T_{25} = 2.381$$

- Da

$$T_{25} = 2.381 > 2.13$$

wird die Nullhypothese H_0 zu Gunsten der Alternative $H_1 : b_1 \neq 0$ verworfen (zum Niveau 5%)

SPSS Output: Der Test 2.29(A) für das Beispiel 2.22 (Arbeitsmotivation)

ANOVA[b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	707,309	9	78,590	21,972	,000[a]
	Residuen	53,651	15	3,577		
	Gesamt	760,960	24			

a. Einflußvariablen : (Konstante), x9, x5, x2, x3, x6, x8, x7, x4, x1

b. Abhängige Variable: Y

SPSS Output: Der Test 2.29(B) für das Beispiel 2.22 (Arbeitsmotivation)

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	95%-Konfidenzintervall für B	
		B	Standard fehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	-3,842	5,052		-,760	,459	-14,609	6,926
	x1	,193	,081	,337	2,381	,031	,020	,365
	x2	,153	,049	,234	3,127	,007	,049	,258
	x3	,049	,065	,095	,761	,458	-,089	,188
	x4	,246	,148	,235	1,664	,117	-,069	,561
	x5	,000	,001	-,077	-,589	,564	-,004	,002
	x6	-,031	,054	-,045	-,576	,573	-,147	,085
	x7	,165	,098	,199	1,683	,113	-,044	,373
	x8	,206	,052	,354	3,973	,001	,095	,316
	x9	,053	,058	,124	,920	,372	-,070	,177

a. Abhängige Variable: Y

2.31 Das Problem der Multikollinearität

Beispiel: Betrachte in dem Beispiel der "Arbeitsmarktmotivation" ein multiples lineares Regressionsmodell mit 3 Prädiktorvariablen

$$Y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + \varepsilon_i \qquad i = 1, \ldots, 25$$

(Y: Motivation, x_1 : Ehrgeiz x_2: Kreativität, x_3: Leistungsstreben)

- Schätzer für die Modellparameter

i	$\hat{b}_i$	$\hat{s}_{b_i}$	p-Wert
0	5.54	2.62	
1	0.39	0.14	0.008
2	0.23	0.09	0.020
3	0.001	0.12	0.994

- Bestimmtheitsmaß $R^2 = 0.7861$
- Beachte: nur für den Koeffizient b_3 (Leistungsstreben) kann keine Signifikanz (zum Niveau 5%) nachgewiesen werden

Korrelationsmatrix für die Prädiktoren

	Motivation	Ehrgeiz	Kreativität	Leistungsstreben
Motivation	1			
Ehrgeiz	.71	1		
Kreativität	.38	.05	1	
Leistungsstreben	.56	.82*	-.02	1

Beachte: Der Test 2.5 liefert eine signifikante Korrelation (zum Niveau 1%) zwischen den Variablen Leistungsstreben und Ehrgeiz (SPSS)

- Beachte: Es gibt eine signifikante Korrelation zwischen den Variablen Leistungsstreben und Ehrgeiz
  - Beide Variablen tragen weitgehend identische Information
  - Im Beispiel ist die Variable Leistungsstreben redundant und wird nicht für die Vorhersage der abhängigen Variablen Motivation benötigt
  - Die Variable Ehrgeiz ist stärker mit der Variablen Motivation korreliert als die Variable Leistungsstreben (aus diesem Grund ist der entsprechende Koeffizient auch signifikant)

- Für die Bestimmtheitsmaße in den multiplen linearen Regressionsmodellen mit drei bzw. zwei Variablen erhält man

  $R^2 = 0.786179$ für Modell mit den Prädiktoren x_1, x_2, x_3

  $R^2 = 0.786178$ für Modell mit den Prädiktoren x_1, x_2

SPSS Output: Multikollinearität; Schätzer im Modell mit 3 Parametern

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	95%-Konfidenzintervall für B	
		B	Standard fehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	5,539	2,618		2,116	,046	,095	10,983
	x1	,393	,135	,688	2,913	,008	,112	,674
	x2	,225	,089	,343	2,528	,020	,040	,410
	x3	,001	,123	,002	,008	,994	-,255	,257

a. Abhängige Variable: Y

SPSS Output: Multilkollinearität; Korrelationsmatrix

Korrelationen

		Y	x1	x2	x3
Y	Korrelation nach Pearson	1,000	,708**	,379	,559**
	Signifikanz (2-seitig)		,000	,061	,004
	N	25	25	25	25
x1	Korrelation nach Pearson	,708**	1,000	,053	,818**
	Signifikanz (2-seitig)	,000		,802	,000
	N	25	25	25	25
x2	Korrelation nach Pearson	,379	,053	1,000	-,016
	Signifikanz (2-seitig)	,061	,802		,939
	N	25	25	25	25
x3	Korrelation nach Pearson	,559**	,818**	-,016	1,000
	Signifikanz (2-seitig)	,004	,000	,939	
	N	25	25	25	25

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

2.32 Das Problem der Suppressionseffekte

Beispiel: Betrachte in dem Beispiel 2.22 der "Arbeitsmarktmotivation" ein multiples lineares Regressionsmodell mit 3 anderen Prädiktorvariablen

$$Y_i = b_0 + b_4 x_{4i} + b_5 x_{5i} + b_6 x_{6i} + \varepsilon_i \qquad i = 1, \ldots, 25$$

(Y: Motivation, x_4: Hierarchie, x_5: Lohn, x_6: Arbeitsbedingungen)

- Schätzungen für die Modellparameter

i	$\hat{b}_i$	$\hat{s}_{b_i}$	p-Wert
0	25.08	8.40	0.007
4	0.88	0.26	0.002
5	-0.01	0.003	0.016
6	0.13	0.12	0.308

Korrelationsmatrix für die Variablen Motivation, Hierarchie, Lohn und Arbeitsbedingungen

	Motivation	Hierarchie	Lohn	Arbeitsbedingungen
Motivation	1			
Hierarchie	.42*	1		
Lohn	-.04	.72**	1	
Arbeitsbedingungen	.35	.16	-.06	1

Beachte:

- Zwischen der Prädiktorvariablen Lohn (x_5) und der abhängigen Variablen Motivation liegt keine signifikante Korrelation vor
- Dennoch bekommt diese Variable im multiplen Regressionsmodell ein signifikantes Gewicht; d.h. die Hypothese $H_0 : b_5 = 0$ wird zum Niveau 5% verworfen (p-Wert: 0.016).
- Man spricht von einem **Suppressionseffekt**.

- Grund für diesen **scheinbaren** Widerspruch: Korrelationen sind **bivariate Maße** für Zusammenhänge (zwischen zwei Merkmalen). Das Modell der multiplen Regression untersucht aber den Zusammenhang zwischen der Variablen Motivation und dem (3-dimensionalen) Prädiktor (x_4, x_5, x_6):
  - Motivation ist stark mit der Variablen *Hierarchie* korreliert
  - *Lohn* ist ebenfalls stark mit der Variablen *Hierarchie* korreliert
  - Prädiktorvariable *Lohn* wird in der multiplen linearen Regression benötigt, um "unerwünschte" Varianzanteile der Variablen *Hierarchie* zu kompensieren
- Bestimmtheitsmaße für verschiedene Modelle

  $R^2 = 0.664282$ für Modell mit x_4, x_5, x_6

  $R^2 = 0.509720$ für Modell mit x_4, x_6

SPSS Output: Suppressionseffekte; Schätzer im Modell mit 4 Parametern

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	95%-Konfidenzintervall für B	
		B	Standard fehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	25,076	8,398		2,986	,007	7,612	42,539
	x4	,884	,257	,843	3,444	,002	,350	1,419
	x5	-,007	,003	-,632	-2,612	,016	-,013	-,001
	x6	,125	,120	,179	1,045	,308	-,124	,375

a. Abhängige Variable: Y

SPSS Output: Suppressionseffekte; Schätzung der Korrelationsmatrix

Korrelationen

		Y	x4	x5	x6
Y	Korrelation nach Pearson	1,000	,419*	-,038	,354
	Signifikanz (2-seitig)		,037	,856	,082
	N	25	25	25	25
x4	Korrelation nach Pearson	,419*	1,000	,717**	,163
	Signifikanz (2-seitig)	,037		,000	,435
	N	25	25	25	25
x5	Korrelation nach Pearson	-,038	,717**	1,000	-,060
	Signifikanz (2-seitig)	,856	,000		,777
	N	25	25	25	25
x6	Korrelation nach Pearson	,354	,163	-,060	1,000
	Signifikanz (2-seitig)	,082	,435	,777	
	N	25	25	25	25

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

2.33 Merkmalselektionsverfahren

- **Ziel**: mit möglichst wenig Prädiktorvariablen eine gute Vorhersage der abhängigen Variablen zu erzielen.
- **Prinzip**: untersuche wie sich durch Weglassen einzelner Variablen das Bestimmtheitsmaß R^2 verändert.

  Typische Selektionsprozeduren:

  (A) Rückwärtsverfahren
  (B) Vorwärtsverfahren
  (C) Schrittweise Verfahren

- Beachte: es handelt sich um explorative Verfahren, die hauptsächlich der Modellbildung dienen (Interpretation nicht einfach).

2.34 Das Rückwärtsverfahren

- Betrachte das vollständige Modell (mit allen Prädiktorvariablen) und berechne das Bestimmtheitsmaß R^2
- Entferne sukzessive diejenigen Variablen, die zu dem geringsten Rückgang des Bestimmtheitsmaßes führen würden
- Das Verfahren wird abgebrochen, falls sich bei dem Entfernen einer Variablen das Bestimmtheitsmaß "signifikant" verkleinert

2.35 Beispiel: Variablenselektion mit dem Rückwärtsverfahren (vgl. Beispiel 2.22)

Schritt	Prädiktorvariablen	t-Wert	Ausgeschlossene Variablen	R^2
1	Ehrgeiz	2.38		.929
	Kreativität	3.13		
	Leistungsstreben	.76		
	Hierarchie	1.66		
	Lohn	-.59		
	Arbeitsbedingungen	-.58		
	Lernpotential	1.68		
	Vielfalt	3.97		
	Anspruch	.92		
2	Ehrgeiz	2.38	Arbeitsbedingungen	.928
	Kreativität	3.28		
	Leistungsstreben	.79		
	Hierarchie	1.66		
	Lohn	-.57		
	Lernpotential	1.66		
	Vielfalt	4.04		
	Anspruch	.91		

Beispiel: Rückwärtsverfahren - Fortsetzung

Schritt	Prädiktorvariablen	t-Wert	Ausgeschlossene Variablen	R^2
3	Ehrgeiz	2.54	Arbeitsbedingungen	.926
	Kreativität	3.43	Lohn	
	Leistungsstreben	.88		
	Hierarchie	2.11		
	Lernpotential	1.59		
	Vielfalt	4.17		
	Anspruch	1.35		
4	Ehrgeiz	5.40	Arbeitsbedingungen	.923
	Kreativität	3.38	Lohn	
	Hierarchie	2.31	Leistungsstreben	
	Lernpotential	1.55		
	Vielfalt	4.12		
	Anspruch	1.31		
5	Ehrgeiz	5.18	Arbeitsbedingungen	.916
	Kreativität	3.16	Lohn	
	Hierarchie	2.84	Leistungsstreben	
	Lernpotential	3.31	Anspruch	
	Vielfalt	5.04		

SPSS Output: Rückwärtssverfahren im Beispiel der Arbeitsmotivation

Aufgenommene/Entfernte Variablen[b]

Modell	Aufgenommene Variablen	Entfernte Variablen	Methode
1	x9, x5, x2, x3, x6, x8, x7, x4, …	.	Eingeben
2	.	x6	Rückwärts (Kriterium: Wahrscheinlichkeit von F-Wert für Ausschluß >= ,100).
3	.	x5	Rückwärts (Kriterium: Wahrscheinlichkeit von F-Wert für Ausschluß >= ,100).
4	.	x3	Rückwärts (Kriterium: Wahrscheinlichkeit von F-Wert für Ausschluß >= ,100).
5	.	x9	Rückwärts (Kriterium: Wahrscheinlichkeit von F-Wert für Ausschluß >= ,100).

a. Alle gewünschten Variablen wurden aufgenommen.

b. Abhängige Variable: Y

SPSS Output: Rückwärtssverfahren im Beispiel der Arbeitsmotivation

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers	Änderungsstatistiken				
					Änderung in R-Quadrat	Änderung in F	df1	df2	Änderung in Signifikanz von F
1	,964[a]	,929	,887	1,891	,929	21,972	9	15	,000
2	,963[b]	,928	,892	1,851	-,002	,332	1	15	,573
3	,963[c]	,926	,896	1,814	-,001	,327	1	16	,575
4	,961[d]	,923	,897	1,803	-,003	,783	1	17	,389
5	,957[e]	,916	,894	1,837	-,007	1,713	1	18	,207

a. Einflußvariablen : (Konstante), x9, x5, x2, x3, x6, x8, x7, x4, x1

b. Einflußvariablen : (Konstante), x9, x5, x2, x3, x8, x7, x4, x1

c. Einflußvariablen : (Konstante), x9, x2, x3, x8, x7, x4, x1

d. Einflußvariablen : (Konstante), x9, x2, x8, x7, x4, x1

e. Einflußvariablen : (Konstante), x2, x8, x7, x4, x1

SPSS Output: Rückwärtssverfahren im Beispiel der Arbeitsmotivation: ANOVA

ANOVA[f]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	707,309	9	78,590	21,972	,000[a]
	Residuen	53,651	15	3,577		
	Gesamt	760,960	24			
2	Regression	706,120	8	88,265	25,752	,000[b]
	Residuen	54,840	16	3,427		
	Gesamt	760,960	24			
3	Regression	705,000	7	100,714	30,596	,000[c]
	Residuen	55,960	17	3,292		
	Gesamt	760,960	24			
4	Regression	702,422	6	117,070	35,999	,000[d]
	Residuen	58,538	18	3,252		
	Gesamt	760,960	24			
5	Regression	696,852	5	139,370	41,306	,000[e]
	Residuen	64,108	19	3,374		
	Gesamt	760,960	24			

a. Einflußvariablen : (Konstante), x9, x5, x2, x3, x6, x8, x7, x4, x1

b. Einflußvariablen : (Konstante), x9, x5, x2, x3, x8, x7, x4, x1

c. Einflußvariablen : (Konstante), x9, x2, x3, x8, x7, x4, x1

d. Einflußvariablen : (Konstante), x9, x2, x8, x7, x4, x1

e. Einflußvariablen : (Konstante), x2, x8, x7, x4, x1

f. Abhängige Variable: Y

SPSS Output: Rückwärtssverfahren im Beispiel der Arbeitsmotivation: Koeffizienten

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten			95%-Konfidenzintervall für B	
		B	Standardfehler	Beta	T	Signifikanz	Untergrenze	Obergrenze
1	(Konstante)	-3,842	5,052		-,760	,459	-14,609	6,926
	x1	,193	,081	,337	2,381	,031	,020	,365
	x2	,153	,049	,234	3,127	,007	,049	,258
	x3	,049	,065	,095	,761	,458	-,089	,188
	x4	,246	,148	,235	1,664	,117	-,069	,561
	x5	,000	,001	-,077	-,589	,564	-,004	,002
	x6	-,031	,054	-,045	-,576	,573	-,147	,085
	x7	,165	,098	,199	1,683	,113	-,044	,373
	x8	,206	,052	,354	3,973	,001	,095	,316
	x9	,053	,058	,124	,920	,372	-,070	,177
2	(Konstante)	-4,737	4,706		-1,007	,329	-14,713	5,238
	x1	,187	,079	,326	2,376	,030	,020	,353
	x2	,157	,048	,239	3,285	,005	,056	,258
	x3	,050	,063	,096	,790	,441	-,084	,185
	x4	,240	,144	,228	1,660	,116	-,066	,545
	x5	,000	,001	-,073	-,572	,575	-,004	,002
	x7	,157	,095	,190	1,655	,117	-,044	,358
	x8	,205	,051	,352	4,040	,001	,097	,312
	x9	,052	,057	,121	,914	,374	-,068	,172
3	(Konstante)	-7,154	2,027		-3,529	,003	-11,431	-2,877
	x1	,193	,076	,338	2,540	,021	,033	,354
	x2	,159	,046	,244	3,431	,003	,061	,258
	x3	,055	,062	,105	,885	,389	-,076	,185
	x4	,172	,081	,164	2,113	,050	,000	,344

a. Abhängige Variable: Y

2.36 Das Vorwärtsverfahren

- Bestimme diejenige Prädiktorvariable, die mit der abhängigen Variablen am stärksten korreliert ist und berechne das Bestimmtheitsmaß R^2
- Ist R^2 signifikant, wird diese Variable in das Modell aufgenommen
- Füge sukzessive diejenigen Variablen zu dem Modell hinzu, die zu dem größten Anstieg des Bestimmtheitsmaßes führen
- Das Verfahren bricht ab, falls sich bei Hinzunahme einer neuen Variablen das Bestimmtheitsmaß R^2 "nicht signifikant" vergrößert

SPSS Output: Vorwärtssverfahren im Beispiel der Arbeitsmotivation

Aufgenommene/Entfernte Variablen[a]

Modell	Aufgenommene Variablen	Entfernte Variablen	Methode
1	x1	.	Vorwährts- (Kriterium: Wahrscheinlichkeit von F-Wert für Aufnahme <= ,050)
2	x9	.	Vorwährts- (Kriterium: Wahrscheinlichkeit von F-Wert für Aufnahme <= ,050)
3	x2	.	Vorwährts- (Kriterium: Wahrscheinlichkeit von F-Wert für Aufnahme <= ,050)
4	x8	.	Vorwährts- (Kriterium: Wahrscheinlichkeit von F-Wert für Aufnahme <= ,050)
5	x4	.	Vorwährts- (Kriterium: Wahrscheinlichkeit von F-Wert für Aufnahme <= ,050)

a. Abhängige Variable: Y

SPSS Output: Vorwärtssverfahren im Beispiel der Arbeitsmotivation

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers	Änderungsstatistiken				
					Änderung in R-Quadrat	Änderung in F	df1	df2	Änderung in Signifikanz von F
1	,708[a]	,501	,479	4,065	,501	23,059	1	23	,000
2	,863[b]	,744	,721	2,973	,244	20,980	1	22	,000
3	,906[c]	,820	,795	2,552	,076	8,876	1	21	,007
4	,944[d]	,891	,869	2,039	,070	12,879	1	20	,002
5	,955[e]	,913	,890	1,869	,022	4,810	1	19	,041

a. Einflußvariablen : (Konstante), x1

b. Einflußvariablen : (Konstante), x1, x9

c. Einflußvariablen : (Konstante), x1, x9, x2

d. Einflußvariablen : (Konstante), x1, x9, x2, x8

e. Einflußvariablen : (Konstante), x1, x9, x2, x8, x4

SPSS Output: Vorwärtssverfahren im Beispiel der Arbeitsmotivation: ANOVA

ANOVA[f]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	380,968	1	380,968	23,059	,000[a]
	Residuen	379,992	23	16,521		
	Gesamt	760,960	24			
2	Regression	566,456	2	283,228	32,035	,000[b]
	Residuen	194,504	22	8,841		
	Gesamt	760,960	24			
3	Regression	624,244	3	208,081	31,962	,000[c]
	Residuen	136,716	21	6,510		
	Gesamt	760,960	24			
4	Regression	677,797	4	169,449	40,751	,000[d]
	Residuen	83,163	20	4,158		
	Gesamt	760,960	24			
5	Regression	694,596	5	138,919	39,773	,000[e]
	Residuen	66,364	19	3,493		
	Gesamt	760,960	24			

a. Einflußvariablen : (Konstante), x1

b. Einflußvariablen : (Konstante), x1, x9

c. Einflußvariablen : (Konstante), x1, x9, x2

d. Einflußvariablen : (Konstante), x1, x9, x2, x8

e. Einflußvariablen : (Konstante), x1, x9, x2, x8, x4

f. Abhängige Variable: Y

SPSS Output: Vorwärtssverfahren im Beispiel der Arbeitsmotivation: Koeffizienten

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	95%-Konfidenzintervall für B	
		B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	9,088	2,406		3,778	,001	4,111	14,064
	x1	,404	,084	,708	4,802	,000	,230	,579
2	(Konstante)	,063	2,642		,024	,981	-5,415	5,542
	x1	,320	,064	,560	4,983	,000	,187	,454
	x9	,221	,048	,515	4,580	,000	,121	,321
3	(Konstante)	-2,101	2,380		-,883	,387	-7,052	2,849
	x1	,319	,055	,558	5,776	,000	,204	,433
	x9	,203	,042	,474	4,862	,000	,116	,290
	x2	,183	,061	,279	2,979	,007	,055	,310
4	(Konstante)	-6,502	2,263		-2,873	,009	-11,224	-1,781
	x1	,253	,048	,442	5,286	,000	,153	,352
	x9	,150	,037	,350	4,101	,001	,074	,226
	x2	,192	,049	,293	3,908	,001	,089	,294
	x8	,190	,053	,327	3,589	,002	,080	,301
5	(Konstante)	-6,833	2,080		-3,285	,004	-11,186	-2,479
	x1	,271	,045	,474	6,076	,000	,178	,364
	x9	,116	,037	,271	3,147	,005	,039	,193
	x2	,177	,045	,271	3,903	,001	,082	,272
	x8	,181	,049	,311	3,706	,001	,079	,283
	x4	,181	,083	,173	2,193	,041	,008	,354

a. Abhängige Variable: Y

2.37 Das schrittweise Verfahren

- Rückwärts- und Vorwärtsverfahren werden kombiniert!
- Man führt ein Vorwärtsverfahren durch, wobei in jedem Schritt untersucht wird, ob bei Entfernen einer bereits aufgenommenen Variable das Bestimmtheitsmaß signifikant abnehmen würde.

SPSS Output: Das schrittweise Verfahren im Beispiel der Arbeitsmotivation

Aufgenommene/Entfernte Variablen[a]

Modell	Aufgenommene Variablen	Entfernte Variablen	Methode
1	x1	.	Schrittweise Auswahl (Kriterien: Wahrscheinlichkeit von F-Wert für Aufnahme <= ,050, Wahrscheinlichkeit von F-Wert für Ausschluß >= ,100).
2	x9	.	Schrittweise Auswahl (Kriterien: Wahrscheinlichkeit von F-Wert für Aufnahme <= ,050, Wahrscheinlichkeit von F-Wert für Ausschluß >= ,100).
3	x2	.	Schrittweise Auswahl (Kriterien: Wahrscheinlichkeit von F-Wert für Aufnahme <= ,050, Wahrscheinlichkeit von F-Wert für Ausschluß >= ,100).
4	x8	.	Schrittweise Auswahl (Kriterien: Wahrscheinlichkeit von F-Wert für Aufnahme <= ,050, Wahrscheinlichkeit von F-Wert für Ausschluß >= ,100).
5	x4	.	Schrittweise Auswahl (Kriterien: Wahrscheinlichkeit von F-Wert für Aufnahme <= ,050, Wahrscheinlichkeit von F-Wert für Ausschluß >= ,100).

a. Abhängige Variable: Y

SPSS Output: Das schrittweise Verfahren im Beispiel der Arbeitsmotivation

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers	Änderungsstatistiken				
					Änderung in R-Quadrat	Änderung in F	df1	df2	Änderung in Signifikanz von F
1	,708[a]	,501	,479	4,065	,501	23,059	1	23	,000
2	,863[b]	,744	,721	2,973	,244	20,980	1	22	,000
3	,906[c]	,820	,795	2,552	,076	8,876	1	21	,007
4	,944[d]	,891	,869	2,039	,070	12,879	1	20	,002
5	,955[e]	,913	,890	1,869	,022	4,810	1	19	,041

a. Einflußvariablen : (Konstante), x1

b. Einflußvariablen : (Konstante), x1, x9

c. Einflußvariablen : (Konstante), x1, x9, x2

d. Einflußvariablen : (Konstante), x1, x9, x2, x8

e. Einflußvariablen : (Konstante), x1, x9, x2, x8, x4

SPSS Output: Das schrittweise Verfahren im Beispiel der Arbeitsmotivation: ANOVA

ANOVA[f]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	380,968	1	380,968	23,059	,000[a]
	Residuen	379,992	23	16,521		
	Gesamt	760,960	24			
2	Regression	566,456	2	283,228	32,035	,000[b]
	Residuen	194,504	22	8,841		
	Gesamt	760,960	24			
3	Regression	624,244	3	208,081	31,962	,000[c]
	Residuen	136,716	21	6,510		
	Gesamt	760,960	24			
4	Regression	677,797	4	169,449	40,751	,000[d]
	Residuen	83,163	20	4,158		
	Gesamt	760,960	24			
5	Regression	694,596	5	138,919	39,773	,000[e]
	Residuen	66,364	19	3,493		
	Gesamt	760,960	24			

a. Einflußvariablen : (Konstante), x1

b. Einflußvariablen : (Konstante), x1, x9

c. Einflußvariablen : (Konstante), x1, x9, x2

d. Einflußvariablen : (Konstante), x1, x9, x2, x8

e. Einflußvariablen : (Konstante), x1, x9, x2, x8, x4

f. Abhängige Variable: Y

SPSS Output: Das schrittweise Verfahren im Beispiel der Arbeitsmotivation: Koeffizienten

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten			95%-Konfidenzintervall für B	
		B	Standardfehler	Beta	T	Signifikanz	Untergrenze	Obergrenze
1	(Konstante)	9,088	2,406		3,778	,001	4,111	14,064
	x1	,404	,084	,708	4,802	,000	,230	,579
2	(Konstante)	,063	2,642		,024	,981	-5,415	5,542
	x1	,320	,064	,560	4,983	,000	,187	,454
	x9	,221	,048	,515	4,580	,000	,121	,321
3	(Konstante)	-2,101	2,380		-,883	,387	-7,052	2,849
	x1	,319	,055	,558	5,776	,000	,204	,433
	x9	,203	,042	,474	4,862	,000	,116	,290
	x2	,183	,061	,279	2,979	,007	,055	,310
4	(Konstante)	-6,502	2,263		-2,873	,009	-11,224	-1,781
	x1	,253	,048	,442	5,286	,000	,153	,352
	x9	,150	,037	,350	4,101	,001	,074	,226
	x2	,192	,049	,293	3,908	,001	,089	,294
	x8	,190	,053	,327	3,589	,002	,080	,301
5	(Konstante)	-6,833	2,080		-3,285	,004	-11,186	-2,479
	x1	,271	,045	,474	6,076	,000	,178	,364
	x9	,116	,037	,271	3,147	,005	,039	,193
	x2	,177	,045	,271	3,903	,001	,082	,272
	x8	,181	,049	,311	3,706	,001	,079	,283
	x4	,181	,083	,173	2,193	,041	,008	,354

a. Abhängige Variable: Y

2.38 Bemerkung zu den verschiedenen Merkmalselektionsverfahren

- **Beachte**: Verschiedene Verfahren liefern verschiedene Ergebnisse (es gibt kein richtig oder falsch!)
- Beispiel (Arbeitsmotivation)

Rückwärtsverfahren	Vorwärtsverfahren	Schrittweises Verfahren
Ehrgeiz	Ehrgeiz	Ehrgeiz
Kreativität	Kreativität	Kreativität
Hierarchie	Hierarchie	Hierarchie
Lernpotential	Anspruch	Anspruch
Vielfalt	Vielfalt	Vielfalt
$R^2 = .916$	$R^2 = .913$	$R^2 = .913$

2.4 Nichtlineare Zusammenhänge

Nichtlineare Zusammenhänge

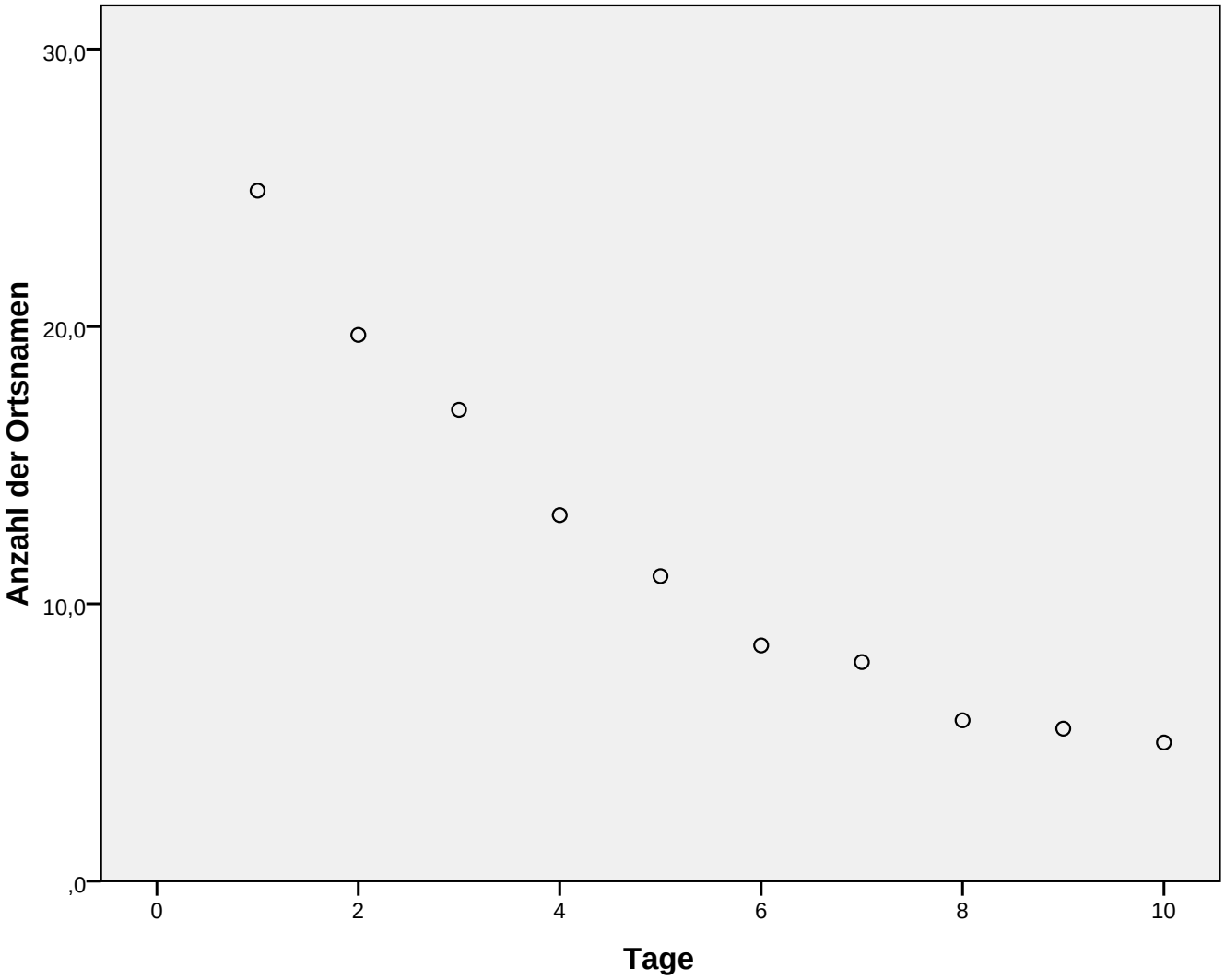
- Die (multiplen) linearen Regressionsmodelle beruhen auf der Annahme, dass der Zusammenhang zwischen jeder Prädiktorvariable und der abhängigen Variablen linear ist, d.h., durch eine Gerade beschrieben werden kann
- Diese Annahme muss nicht immer erfüllt sein. Zusammenhänge zwischen Variablen können im Grunde beliebige Form haben
- Man spricht in diesen Fällen von **nichtlinearen Zusammenhängen**

2.39 Beispiel: Gedächtnistest

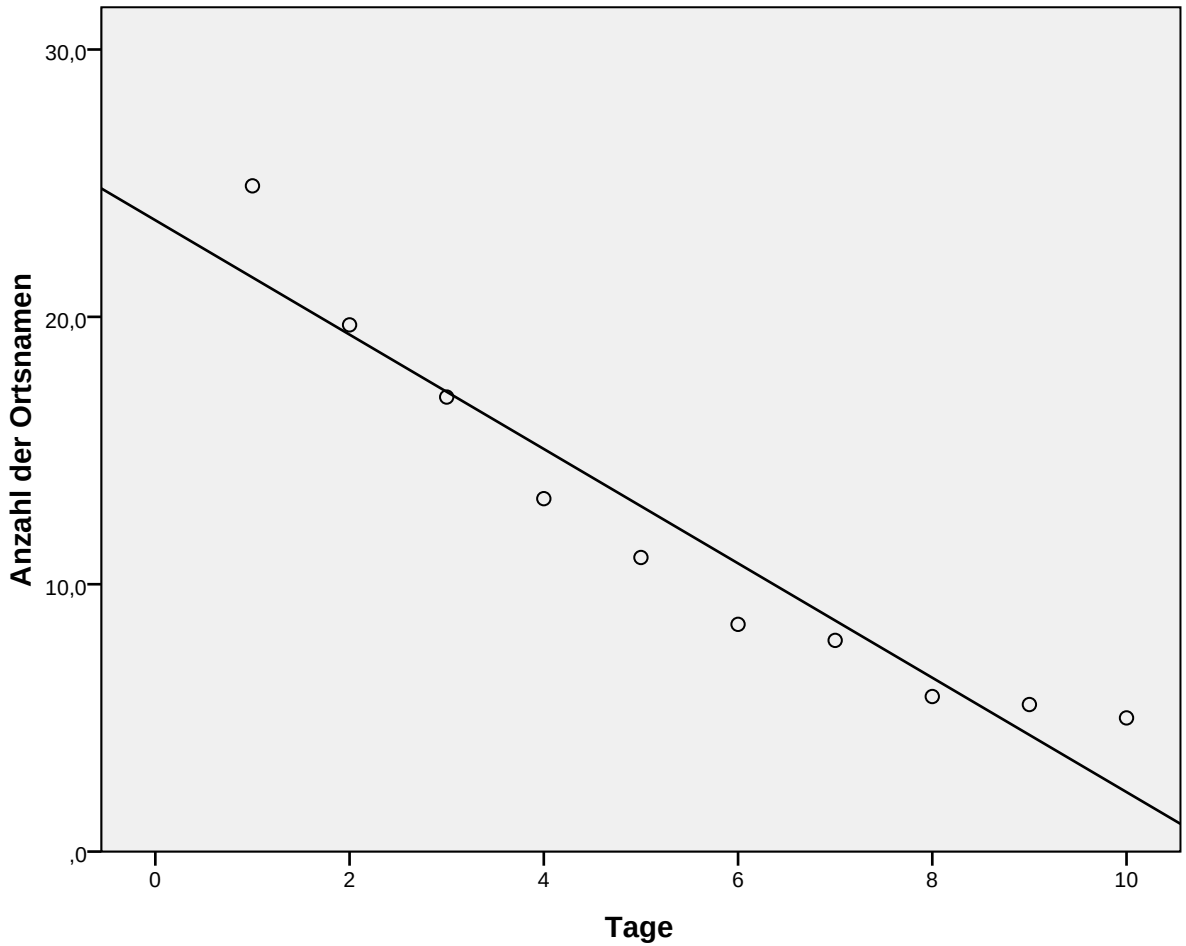
- Mehrere Personen machen einen Gedächtnistest
- 30 Ortsnamen (aus Mongolei) werden vorgegeben
- $y(x)$: Anzahl der Ortsnamen, die nach x Tagen noch im Gedächtnis geblieben sind (Mittelwerte)

x	1	2	3	4	5	6	7	8	9	10
$y(x)$	24.9	19.7	17.0	13.2	11.0	8.5	7.9	5.8	5.5	5.0

Das Streudiagramm für die Daten aus Beispiel 2.39 (Gedächtnistest)



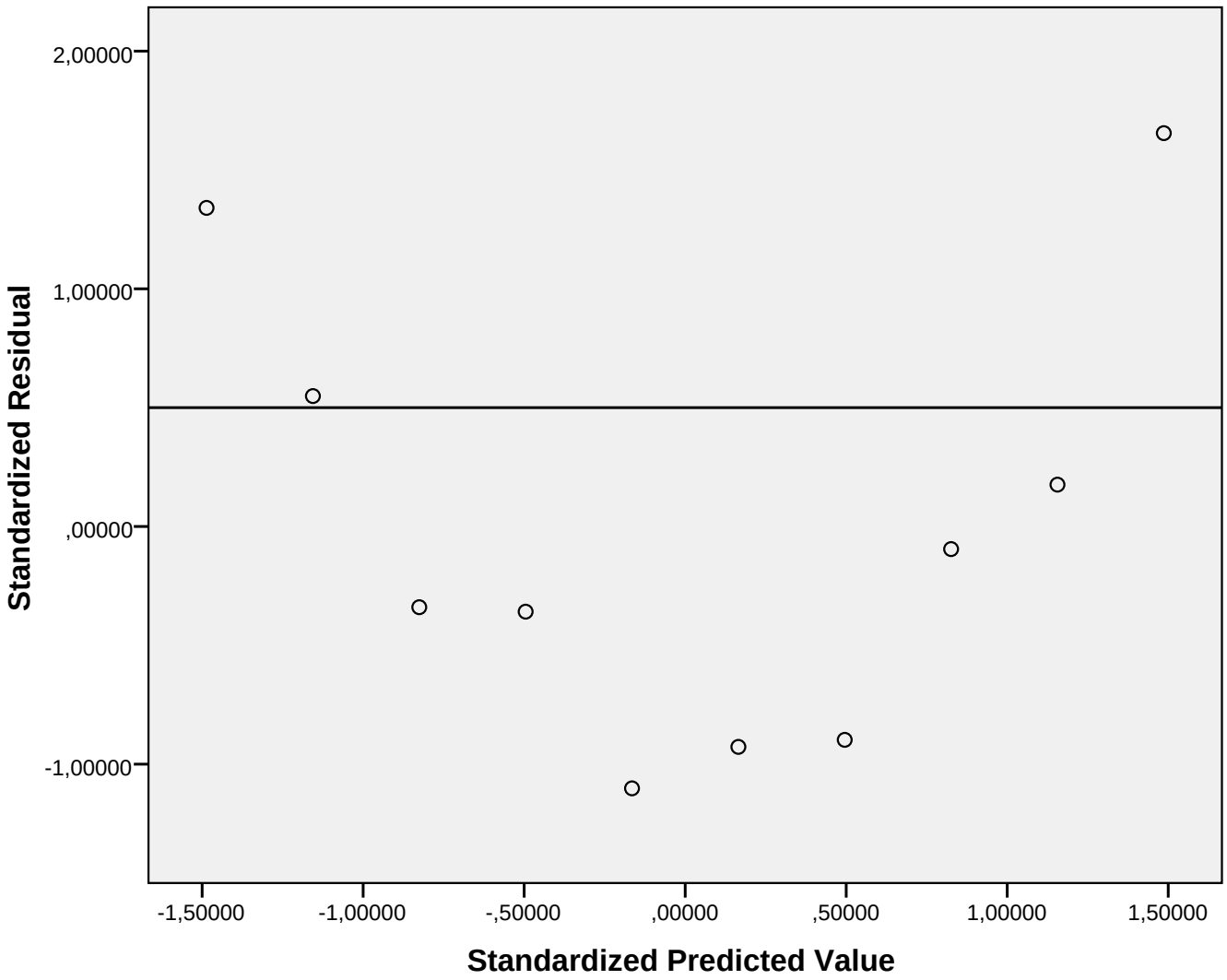
Lineare Regression für die Daten aus Beispiel 2.39 (Gedächtnistest)



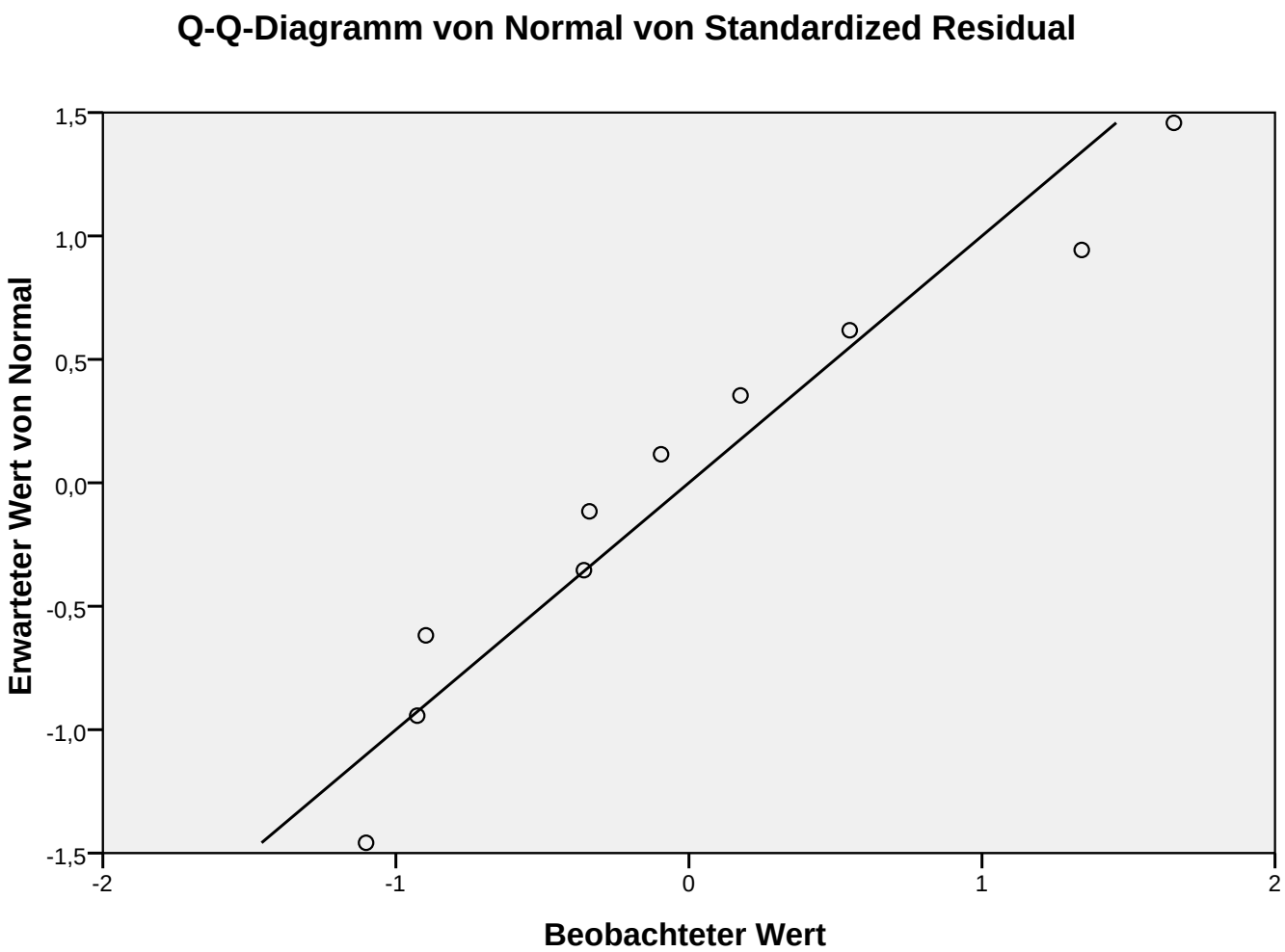
Die Gleichung der geschätzten Geraden:

$$y = 10.579 - 0.429x$$

Residuenanalyse bei linearer Regression für die Daten aus Beispiel 2.39 (Gedächtnistest)



QQ - Plot bei linearer Regression für die Daten aus Beispiel 2.39 (Gedächtnistest)



Beachte:

- Ein lineares Regressionmodell ist für die Beschreibung des Zusammenhangs ungeeignet!
- Quadratisches Regressionsmodel

$$Y_i = b_0 + b_1 x_i + b_2 x_i^2 + \varepsilon_i$$

- Schätzung der Parameter mit der Methode der kleinsten Quadrate und die entsprechenden Standardfehler

$$\begin{array}{lrlrlr} \hat{b}_0 = & 29.088 & \hat{b}_1 = & -4.876 & \hat{b}_2 = & 0.249 \\ \hat{s}_{b_0} = & 0.558 & \hat{s}_{b_1} = & 0.233 & \hat{s}_{b_2} = & 0.021 \end{array}$$

Konfidenzbereiche und Tests

- Man geht wie in 2.12 und 2.14 bzw. 2.29 vor
- 90% Konfidenzintervall für b_2 (man beachte: das Modell hat 3 Parameter)

$$t_{10-3,0.95} = 1.8946 \quad \hat{b}_2 = 0.249 \quad \hat{s}_{b_2} = 0.021$$

$$\Rightarrow [\hat{b}_2 - t_{7,0.95}\ \hat{s}_{b_2}, \hat{b}_2 + t_{7,095}\ \hat{s}_{b_2}] = [0.2092, 0.2888]$$

ist 90% Konfidenzintervall für b_2

- Die Hypothese $H_0.b_2 = 0$ wird (zum Niveau 10%) verworfen, fall

$$\left| \frac{\hat{b}_2}{\hat{s}_{b_2}} \right| > t_{10-3,0.95}$$

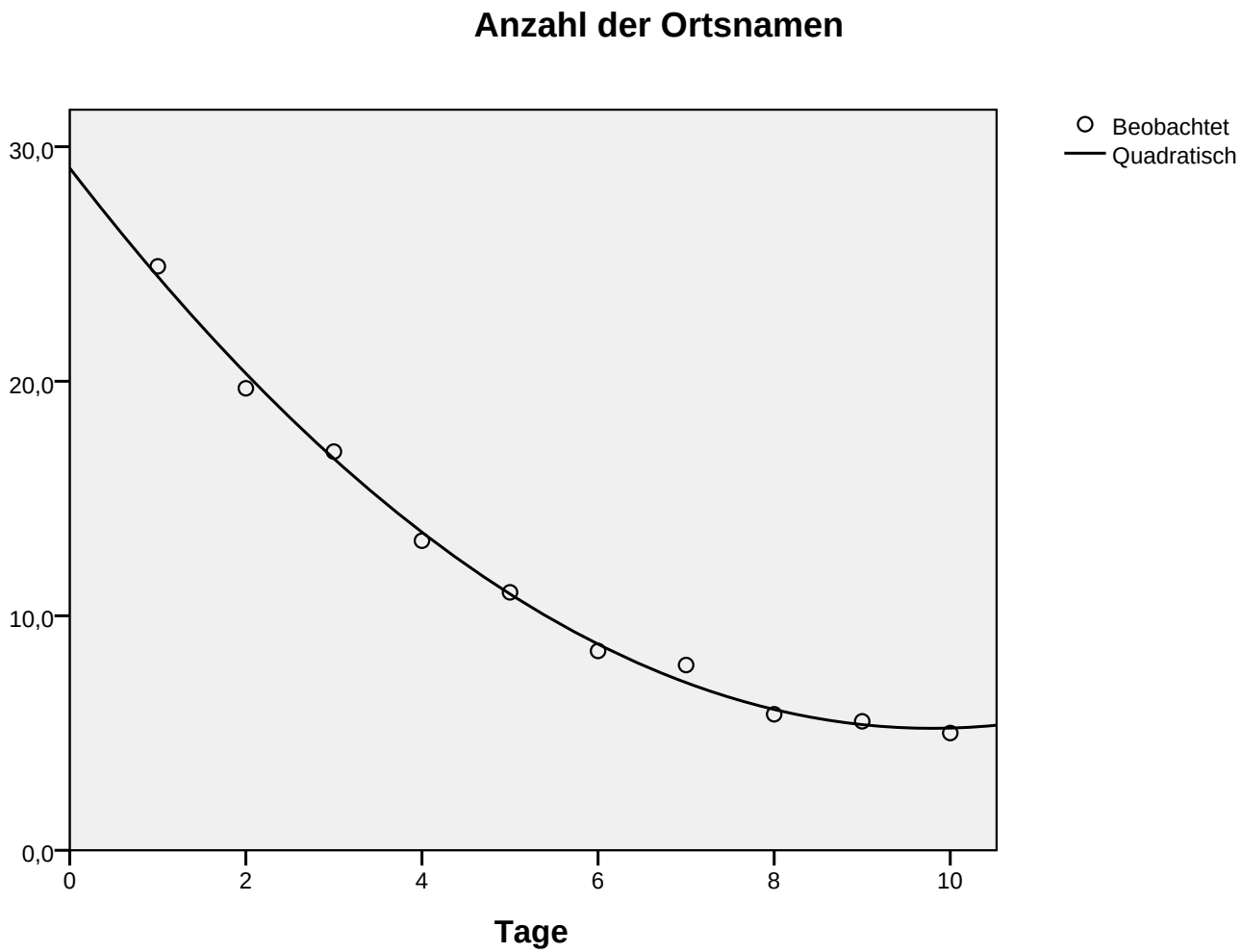
gilt (im Beispiel wird also H_0 abgelehnt)

SPSS-Output: Schätzer für quadratische Regression

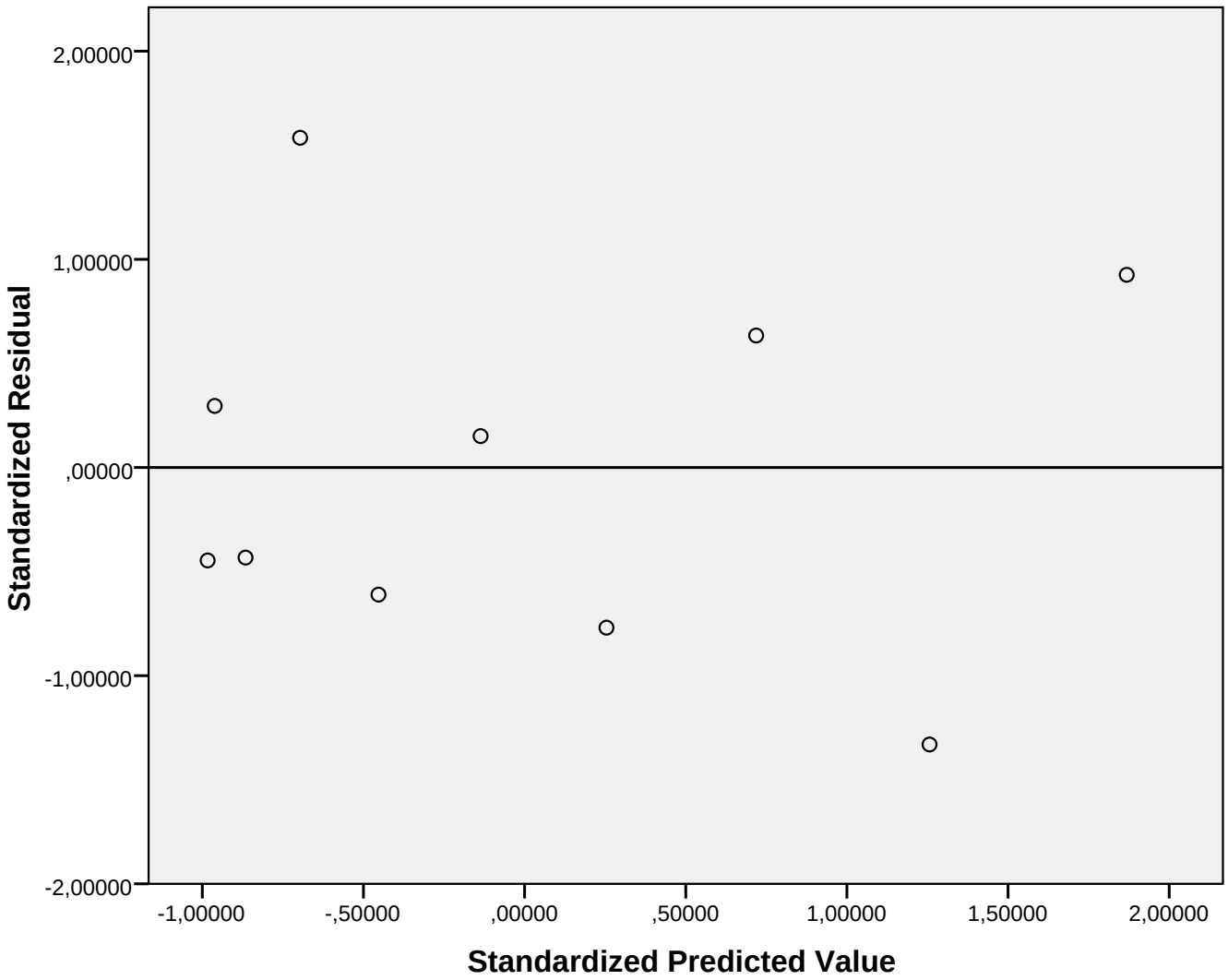
Koeffizienten

	Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
	B	Standardfehler	Beta	t	Sig.
Tage	-4,876	,233	-2,183	-20,927	,000
Tage ** 2	,249	,021	1,257	12,055	,000
(Konstante)	29,088	,558		52,136	,000

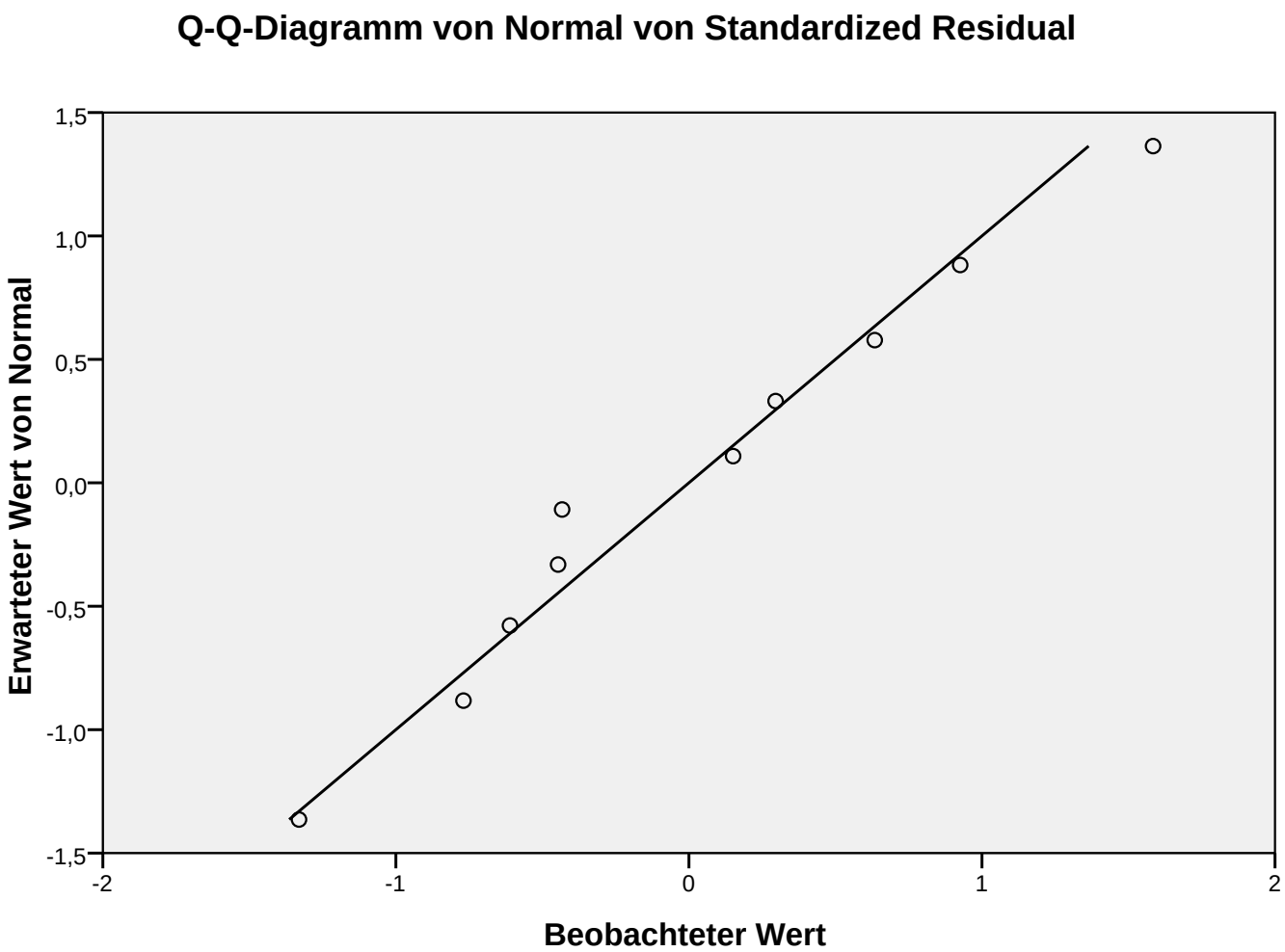
Streudiagramm für die Daten aus Beispiel 2.39 mit der geschätzten Parabel



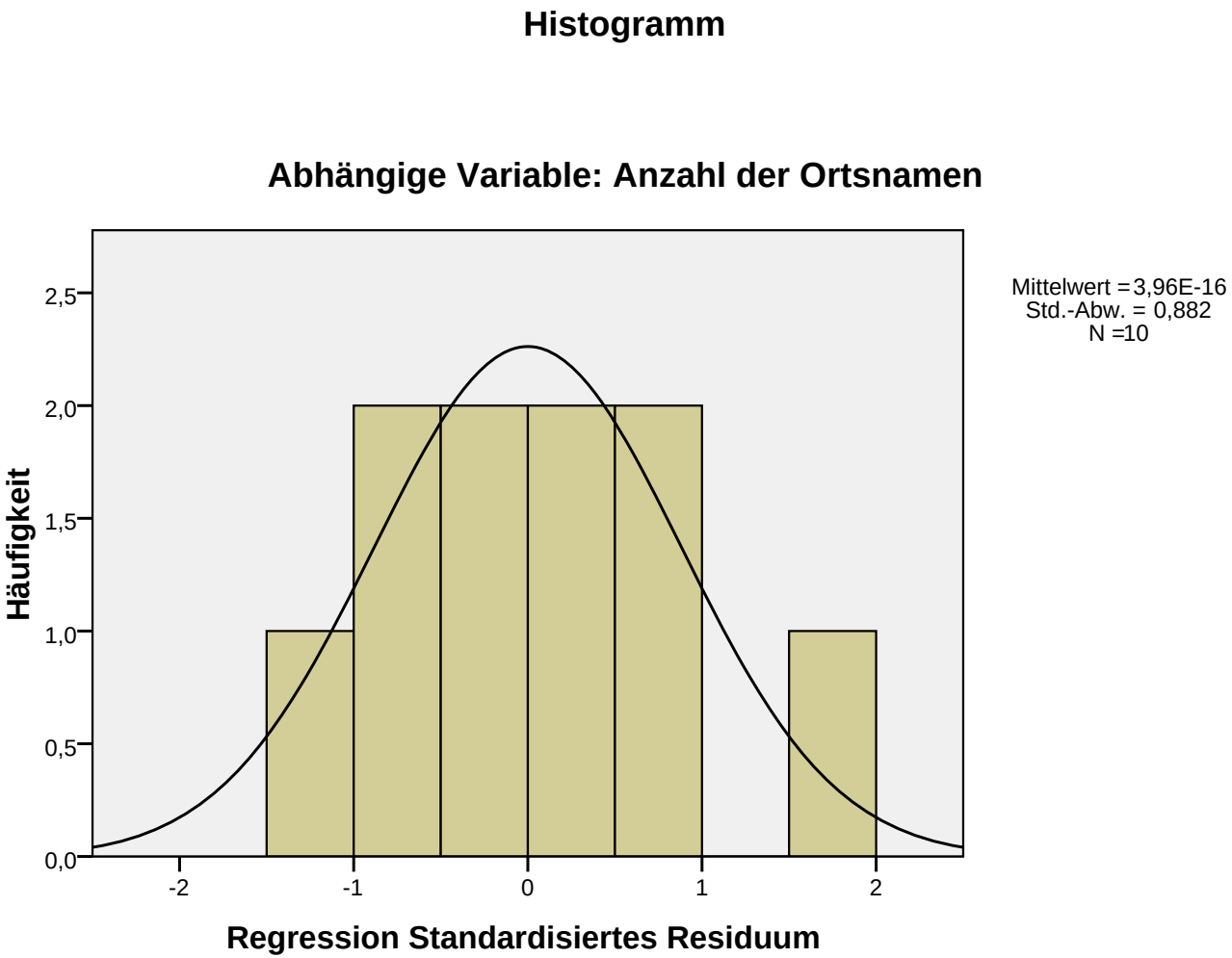
SPSS-Output: Residuenanalysefür für die Daten aus Beispiel 2.39 bei quadratischer Regression



SPSS-Output: QQ-Plot für die Daten aus Beispiel 2.39 bei quadratischer Regression



SPSS-Output: Histogramm für die Residuen aus Beispiel 2.39 bei quadratischer Regression



2.40 Polynomiale Regressionsmodelle

Modelle zur polynomialen Regression

Ordnung	Modell
0.	$Y = b_0 + \varepsilon$
1.	$Y = b_0 + b_1 x^1 + \varepsilon$
2.	$Y = b_0 + b_1 x^1 + b_2 x^2 + \varepsilon$
$\vdots$	$\vdots$
k.	$Y = b_0 + b_1 x^1 + b_2 x^2 + \ldots + b_k x^k + \varepsilon$

Beachte:

- In der Regel werden nur Modelle von niedrigem Grad verwendet ($k \leq 3$)!
- Schätzung der Parmeter erfolgt mit der Methode der kleinsten Quadrate
- Konfidenzintervalle, Tests und Residuenanalyse werden wie bei der linearen bzw. multiplen Regression durchgeführt (Allgemeines lineares Modell)

2.41 Mehrdimensionale Polynome

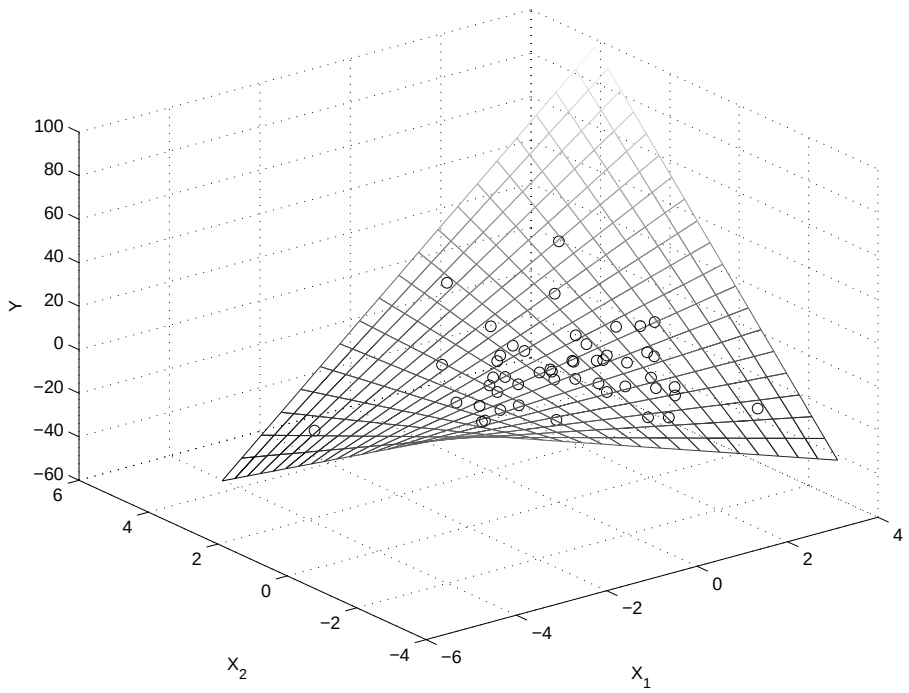
- Sind mehrere Prädiktorenvariable verfügbar, so können neben Potenzen auch Produkte von zwei oder mehr Variablen in die Regressionsgleichung aufgenommen werden
- **Beispiele:**

$$Y(\mathbf{x}) = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2 + \varepsilon$$

$$Y(\mathbf{x}) = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2 + b_{02} x_1^2 + b_{20} x_2^2 + \varepsilon$$

$$Y(\mathbf{x}) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_{120} x_1 x_2 + b_{103} x_1 x_3 + b_{023} x_2 x_3 + b_{123} x_1 x_2 x_3 + \varepsilon$$

3D-Streudiagramm mit der geschätzten Funktion



Die geschätzte Funktion ist:

$$\widehat{y}(\mathbf{x}) = 2.23 + 3.52x_1 + 5.77x_2 + 3.96x_1x_2$$

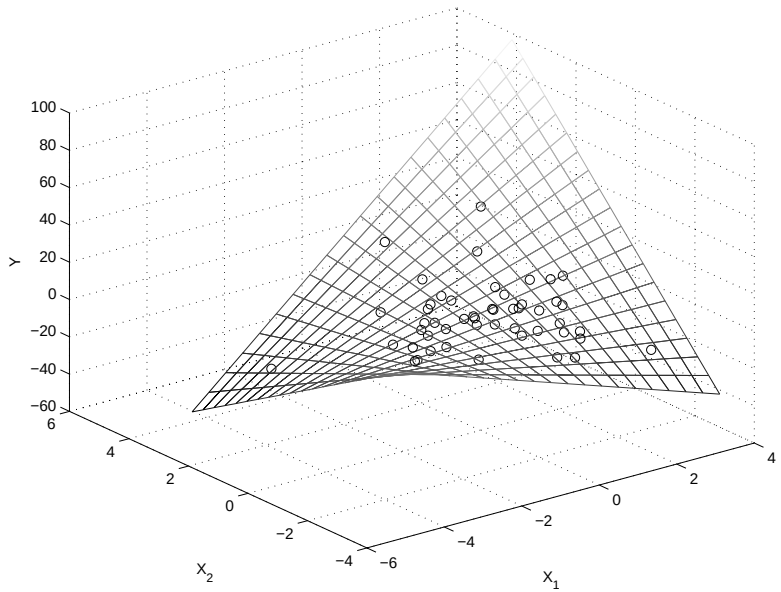
.

3D-Streudiagramm mit der geschätzten Funktion

Polynomiale Terme und Produkte der Prädiktoren können natürlich auch gemeinsam vorkommen.

Beispiel:

$$y(\mathbf{x}) = b_0 + b_{11}x_1 + b_{12}x_1^2 + b_{21}x_2 + b_{23}x_2^3 + b_{11;21}x_1x_2 + \varepsilon.$$



Die angepasste Funktion hat die Form

$$\widehat{y}(\mathbf{x}) = 1 + 2.15x_1 + 6.59x_1^2 + 1.66x_2 + 3.07x_2^3 + 3.76x_1x_2$$