Statistik II für Betriebswirte
Vorlesung 8

Dr. Andreas Wünsche

TU Bergakademie Freiberg
Institut für Stochastik

2. Dezember 2019

7. Regressionsanalyse

7.1. Lineare Regression

- Während bei der Korrelationsanalyse eine **qualitative Analyse** von Zusammenhängen zwischen Merkmalen im Vordergrund stand, führt man bei der Regressionsanalyse eine **quantitative Analyse** von derartigen Zusammenhängen durch.
- Insbesondere sucht man im Rahmen einer Regressionsanalyse, z.B. auf der Basis von Beobachtungen $(x_1, y_1), \ldots (x_n, y_n)$, nach einem konkreten funktionalen Zusammenhang, der die Abhängigkeit eines Merkmals Y von einer Merkmalsgröße X beschreibt. (einfache Regression)
- Im Beispiel 6.4 kann man z.B. die Frage stellen, ob ein funktionaler Zusammenhang zwischen den Variablen Jahresumsatz (Y) und der Prädiktorvariablen Verkaufsfläche (X) besteht?
- Gesucht ist also eine Funktion f, die aus der Prädiktorvariablen Verkaufsfläche (X) eine Vorhersage für die abhängige Variable Jahresumsatz (Y) liefert.

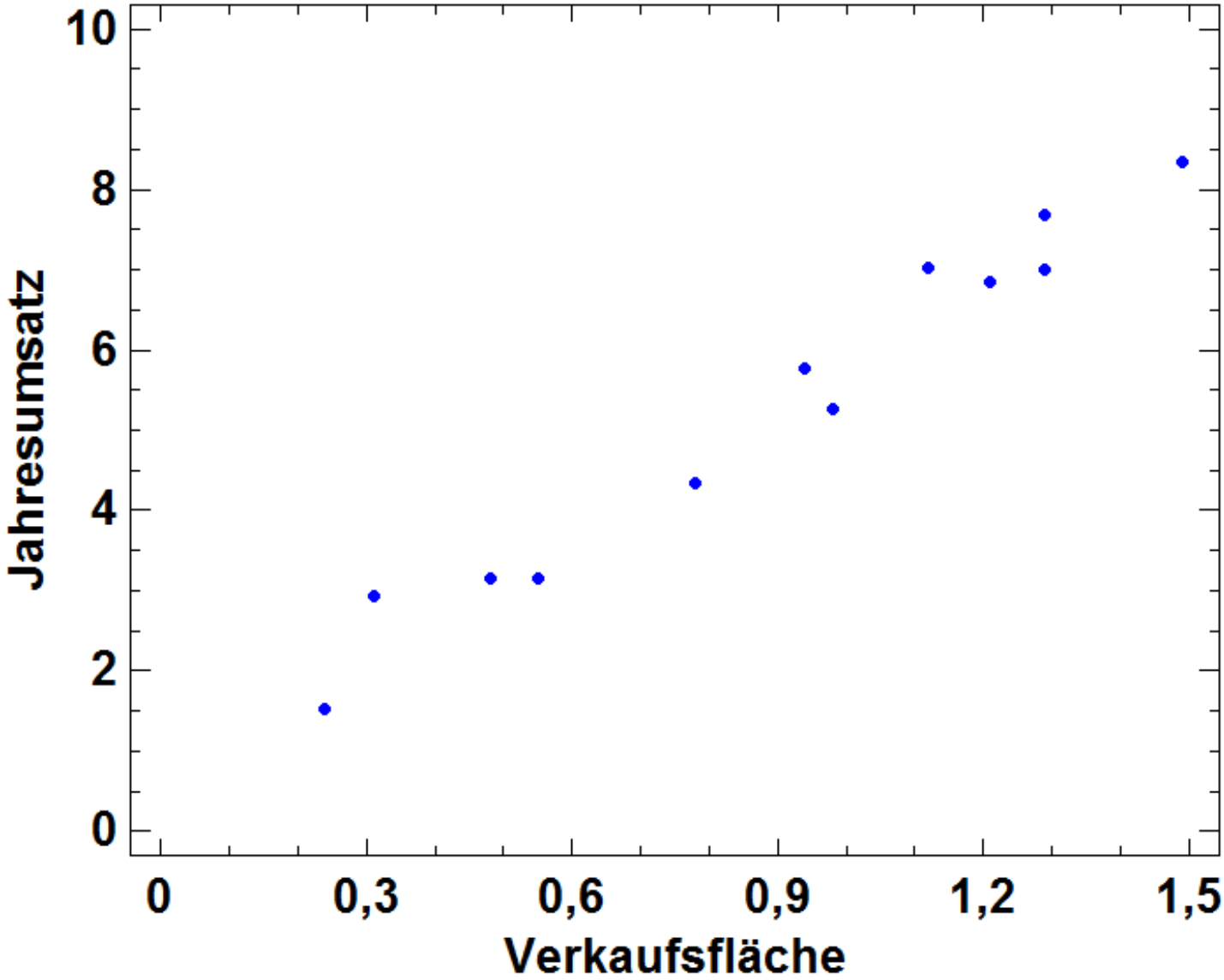
Regression

- 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 Variable y und der Prädiktorvariable x.
  Weitere Annahme : Die Funktion f hat eine bestimmte Form.
  Beispiele:
  - Lineare Regression (der Zusammenhang wird also durch eine Gerade beschrieben): $y = b_0 + b_1 x$,
  - Quadratische Regression (der Zusammenhang wird durch eine Parabel beschrieben): $y = b_0 + b_1 x + b_2 x^2$,
  - usw.
- Beachte: Der Zusammenhang ist in der Regel nicht exakt zu beobachten. Das Modell (**Lineare Regression**) lautet:

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

Dabei bezeichnet ε eine zufällige Störgröße.

Streudiagramm für die Daten aus Beispiel 6.4



Die Methode der kleinsten Quadrate

- 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 ist diejenige Gerade, die den Zusammenhang zwischen Y und x am besten beschreibt.
- Bestimme die Gerade so, dass die Summe der quadrierten senkrechten Abstände zwischen der Gerade und den 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:

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

Die Regressionsgerade

- Die Lösung des Extremwertproblems liefert Schätzer für die Steigung und den Achsenabschnitt der Geraden:

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

  und damit die Gleichung der geschätzten Regressionsgeraden

$$\hat{y} = \hat{b}_0 + \hat{b}_1\, x\,.$$

- Der Wert der geschätzten Regressionsgerade an der Stelle x_i ist

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

- Die Abweichungen $y_i - \hat{y}_i$ nennt man Residuen.
- Die Summe der Residuen ist Null, $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0\,.$
- Die Regressionsgerade verläuft durch $(\overline{x}, \overline{y})$, den Schwerpunkt .

Weitere Bezeichnungen und Bemerkungen

- Eine mögliche andere Parametrisierung ist $a = b_0$ und $b = b_1$.
  $\Longrightarrow$ Modellgleichung: $Y = a + bx + \varepsilon$.

- Ist eine funktionale Abhängigkeit der Größe Y von der Größe X gesucht, nennt man

X	und	Y	unter anderem auch:
Regressor	und	Regressand	,
Einflussgröße	und	Wirkungsgröße	,
unabhängige Variable	und	abhängige Variable	,
Prädiktorvariable	und	Zielvariable	,
exogene Variable	und	endogene Variable	.

- Der Name „Regression“ („Rückschritt“) geht auf GALTON zurück. Ausgangspunkt war damals eine Untersuchung der Größe der Söhne (Variable Y) im Zusammenhang mit der Größe der Väter (Variable X) von PEARSON. Galton schrieb damals: „Each peculiarity in a man is shared by his kinsmen but on the average in a less degree.“

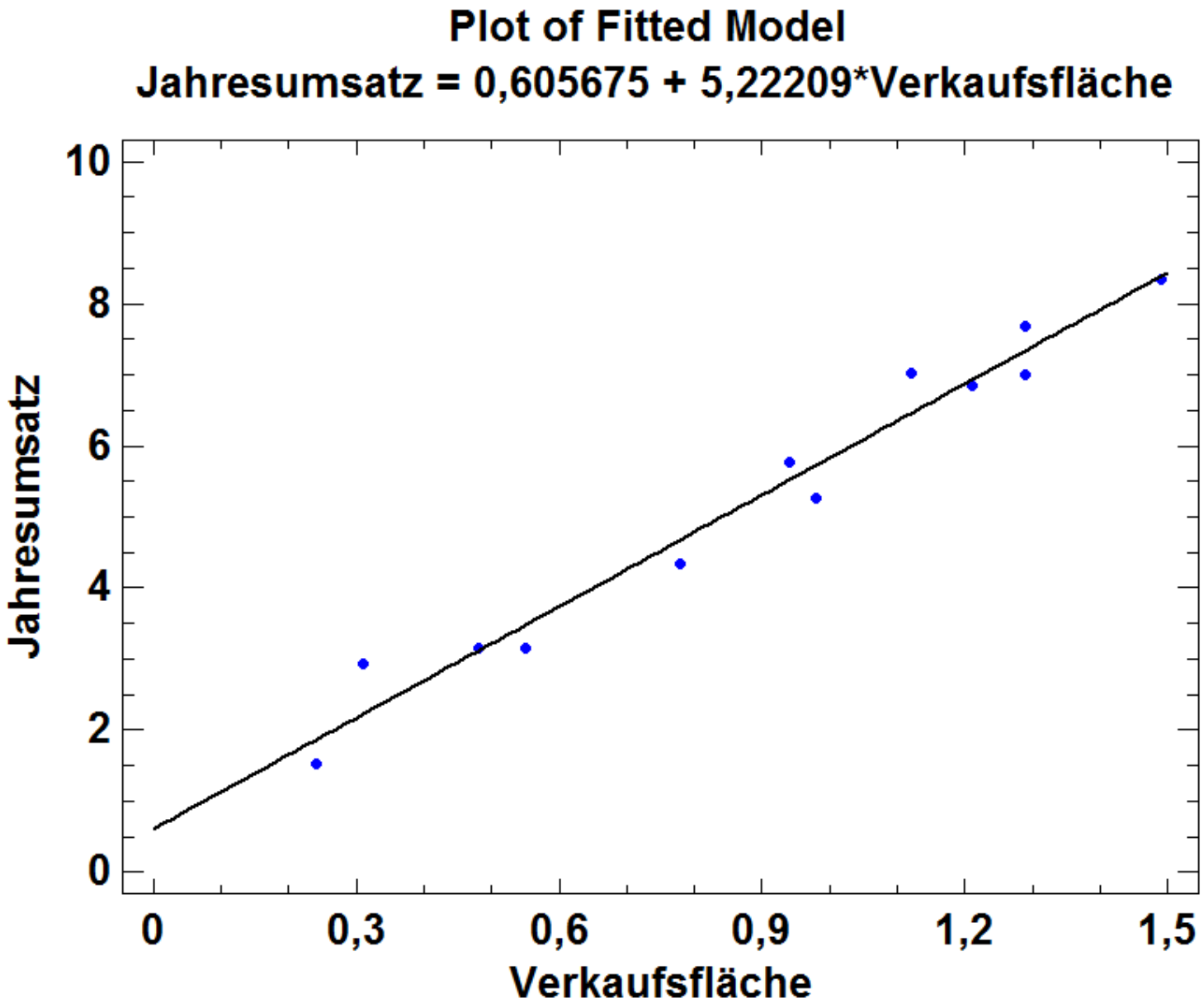
Beispiel 7.1 Jahresumsatz und Verkaufsfläche

- Fortsetzung vom Beispiel 6.4: Daten aus Bleymüller et al, Statistik für Wirtschaftswissenschaftler, 2004, Kap. 20.
- i Filiale
  x_i Verkaufsfläche in Tsd. qm
  y_i Jahresumsatz in Mio. €

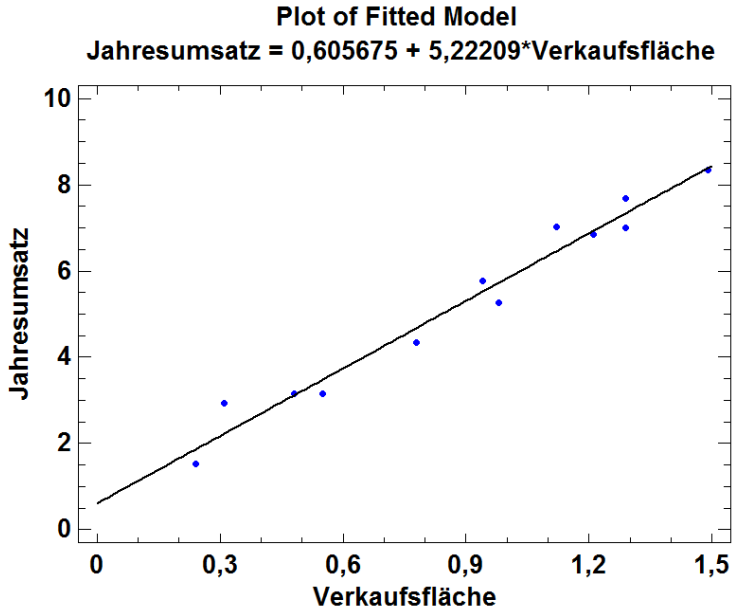
i	1	2	3	4	5	6
x_i	0.31	0.98	1.21	1.29	1.12	1.49
y_i	2.93	5.27	6.85	7.01	7.02	8.35
i	7	8	9	10	11	12
x_i	0.78	0.94	1.29	0.48	0.24	0.55
y_i	4.33	5.77	7.68	3.16	1.52	3.15

- Berechnung der Regressionsgeraden in Statgraphics unter:
  Relate → One Factor → Simple Regression
  (Beziehungen → Ein Faktor → Einfache Regression).

Regressionsgerade im Beispiel 7.1 (Statgraphics)



Weitere Fragen zur Regression im Beispiel 7.1



- Schätzer: $\hat{b}_0 = 0.606\,, \hat{b}_1 = 5.222$
- **Fragen**:
  - Wie genau sind diese Schätzungen?
  - Besteht ein (signifikanter) Einfluss der Verkaufsfläche auf den Jahresumsatz?

$$H_0 \ : b_1 = 0$$

  - Wie gut beschreibt das lineare Regressionsmodell die Situation?

Genauigkeit der Schätzer für die Parameter

- Beachte: Vor der Datenerhebung sind $\hat{b}_0$ und $\hat{b}_1$ zufällig.
- Die mathematische Statistik (allgemeines lineares Modell) liefert Schätzer für die Varianzen von $\hat{b}_0$ und $\hat{b}_1$

Schätzer für die Varianz von $\hat{b}_0$:

$$s_{b_0}^2 = \frac{s_{Rest}^2}{n} \cdot \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Schätzer für die Varianz von $\hat{b}_1$:

$$s_{b_1}^2 = \frac{s_{Rest}^2}{n} \cdot \frac{1}{\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Dabei ist

$$\hat{\sigma}^2 = s_{Rest}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{b}_0 + \hat{b}_1 x_i))^2$$

der Schätzer für die Varianz der Störgrößen.

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

Streuungszerlegung

- Es gilt die Streuungszerlegung $\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}$ mit

  $\mathrm{SST} = \sum_{i=1}^{n} (y_i - \overline{y})^2$, der Totalvariabilität (Totalvarianz);

  $\mathrm{SSE} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$, der „erklärten“ Variabilität (erklärte Varianz);

  $\mathrm{SSR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, der Restvariabilität (Residualvarianz).

- Das Verhältnis $B = \dfrac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \dfrac{\mathrm{SSR}}{\mathrm{SST}}$ heißt Bestimmtheitsmaß.
- Es gelten $0 \leq B \leq 1$ und $B = r_{X,Y}^2$ mit dem gewöhnlichen empirischen Korrelationskoeffizienten $r_{X,Y}$.
- **Je besser das Modell ist, desto kleiner ist die Residualvarianz, bzw. desto größer ist das Bestimmtheitsmaß B!**

Das stochastische Modell

- Weiterführende statistische Aussagen, wie Konfidenzintervalle oder statistische Tests, basieren auf einem geeigneten stochastischen Modell.
- Üblicherweise nimmt man in dieser Situation an, dass

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

gilt, wobei die Werte x_i (zunächst) deterministische, einstellbare Werte sind und die zufälligen Störgrößen durch unabhängige normalverteilte Zufallsgrößen ε_i („zufällige Fehler“) mit $\mathbf{E}\varepsilon_i = 0$ und $\mathbf{Var}\varepsilon_i = \sigma^2$ (unbekannt, aber konstant) verursacht werden.

- Unter diesen Bedingungen sind $\hat{b}_0$ bzw. $\hat{b}_1$ erwartungstreue und konsistente Schätzfunktionen für die Modellparameter b_0 bzw. b_1.
- Die Standardabweichung σ der Fehler kann geschätzt werden durch

$$\hat{\sigma} = s_{Rest} = \sqrt{\frac{\mathrm{SSR}}{n-2}}\,.$$

Konfidenzintervalle zum Niveau $1-\alpha$ für die Parameter

- Ein Konfidenzintervall für b_0 ist
  $[\hat{b}_0 - s_{b_0} t_{n-2;1-\alpha/2} \quad ; \quad \hat{b}_0 + s_{b_0} t_{n-2;1-\alpha/2}]$.

- Ein Konfidenzintervall für b_1 ist
  $[\hat{b}_1 - s_{b_1} t_{n-2;1-\alpha/2} \quad ; \quad \hat{b}_1 + s_{b_1} t_{n-2;1-\alpha/2}]$.

- Ein Konfidenzintervall für die Fehlervarianz σ^2 ist
  $$\left[\frac{(n-2)\hat{\sigma}^2}{\chi^2_{n-2;1-\alpha/2}} \quad ; \quad \frac{(n-2)\hat{\sigma}^2}{\chi^2_{n-2;\alpha/2}}\right] = \left[\frac{\mathrm{SSR}}{\chi^2_{n-2;1-\alpha/2}} \quad ; \quad \frac{\mathrm{SSR}}{\chi^2_{n-2;\alpha/2}}\right].$$

Konfidenzintervalle im Beispiel 7.1

Mit $\hat{b}_0 = 0.605675$, $s_{\hat{b}_0} = 0.288656$, $\hat{b}_1 = 5.22209$, $s_{\hat{b}_1} = 0.296079$ (vgl. Statgraphics-Ergebnisse auf Folie 17) lauten die Konfidenzintervalle zum Konfidenzniveau 95% für:

$$b_0: \quad [-0.038\ ;\ 1.2494] \qquad \text{und} \qquad b_1: \quad [4.5618\ ;\ 5.8823]\,.$$

Mit $\mathrm{SSR} = 1.66612$ ist die Punktschätzung für σ^2, der Varianz der Fehler:

$$\hat{\sigma}^2 = \frac{1}{n-2}\mathrm{SSR} = \frac{1}{10}1.66612 = 0.166612.$$

(vgl. Statgraphics-Ergebnisse auf Folie 19)

Damit ist das Konfidenzintervall für σ^2 zum Konfidenzniveau 95%:

$$[0.0814\ ;\ 0.5127].$$

Tests für die Parameter b_0 und b_1

- **Hypothesen:** $H_0 : b_0 = b_{0_0}$, $H_A : b_0 \neq b_{0_0}$;
  bzw. $H_0 : b_1 = b_{1_0}$, $H_A : b_1 \neq b_{1_0}$.
- **Testgrößen:** $T_{b_0} = \dfrac{\hat{b}_0 - b_{0_0}}{s_{b_0}}$ bzw. $T_{b_1} = \dfrac{\hat{b}_1 - b_{1_0}}{s_{b_1}}$
- Die Testgrößen sind unter H_0 $t-$verteilt mit $n-2$ Freiheitsgraden.
- **Kritischer Bereich (Niveau α):** $K = \{t \in \mathbb{R} : |t| > t_{n-2;1-\alpha/2}\}$.
- Analog können einseitige Tests durchgeführt werden.

t-Tests im Beispiel 7.1 mit Statgraphics

Simple Regression - Jahresumsatz vs. Verkaufsfläche

Dependent variable: Jahresumsatz (Mio Euro)
Independent variable: Verkaufsfläche (1000 qm)
Linear model: Y = a + b*X

Coefficients

	Least Squares	*Standard*	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
Intercept	0,605675	0,288656	2,09826	0,0623
Slope	5,22209	0,296079	17,6375	0,0000

- Test mit $H_0 : b_0 = 0$ gegen $H_A : b_0 \neq 0$; $\hat{b}_0 = 0.605675$, $p = 0.0623 > 0.05 = \alpha \Rightarrow H_0$ wird nicht abgelehnt, d.h., man kann nicht darauf schließen, dass der Koeffizient b_0 signifikant von 0 verschieden ist.
- Test mit $H_0 : b_1 = 0$ gegen $H_A : b_1 \neq 0$; $\hat{b}_1 = 5.22209$, $p = 0.0000 < 0.05 = \alpha \Rightarrow H_0$ wird abgelehnt, d.h., der Koeffizient b_1 ist signifikant von 0 verschieden.

F-Test für die Hypothese $H_0 : b_1 = 0$

- Es besteht also ein signifikanter Einfluss der Verkaufsfläche auf den Jahresumsatz.
- Die Hypothesen

$$H_0 : \ b_1 = 0 \qquad \text{gegen} \qquad H_A : b_1 \neq 0$$

können auch mit dem F-Test getestet werden. Dieser Test spielt z.B. im Modell der multiplen parameterlinearen Regression eine eigenständige Rolle.

- Testgröße:

$$T = \frac{\frac{1}{1}\mathrm{SSE}}{\frac{1}{n-2}\mathrm{SSR}} = \frac{\mathrm{MSE}}{\mathrm{MSR}}$$

- Falls $H_0 : b_1 = 0$ gilt ist $T \sim F_{1,n-2}$ und damit ist der kritische Bereich:

$$K = \{t \ : \quad t > F_{1,n-2;1-\alpha}\}.$$

F-Test im Beispiel 7.1 mit Statgraphics

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	51,8297	1	51,8297	311,08	0,0000
Residual	1,66612	10	0,166612		
Total (Corr.)	53,4959	11			

$$t = \frac{\frac{1}{1}51.8297}{\frac{1}{12-2}1.66612} = \frac{51.8297}{0.166612} = 311.08 > 4.96 = F_{1,10;0.95}$$

$\Longrightarrow$ H_0 wird abgelehnt. (Gleiches Ergebnis wie beim t-Test.)

- Zusammenhang zum t-Test: Ist $t \sim t_{n-2}$, dann ist $t^2 \sim F_{1,n-2}$. Hier: $17.6375^2 = 311.08$.
- Zusammenhang zum Bestimmtheitsmaß B: Ist t die Realisierung der Testgröße des F-Tests, dann gilt:

$$\frac{\frac{1}{n-2}t}{1+\frac{1}{n-2}t} = \frac{\frac{1}{10}311.08}{1+\frac{1}{10}311.08} = 0.96885 = \frac{51.8297}{53.4959} = B.$$

Konfidenzintervalle für die Regressionsgerade

- Häufig möchte man jedoch Konfidenzintervalle für den **Wert der Regressionsgerade** an einer Stelle x (oder für ein Intervall von $x-$Werten) bestimmen, d.h. für $\mathbf{E}Y(x) = b_0 + b_1 x$.
- Ein solches Konfidenzintervall zum Niveau $1-\alpha$ kann berechnet werden durch

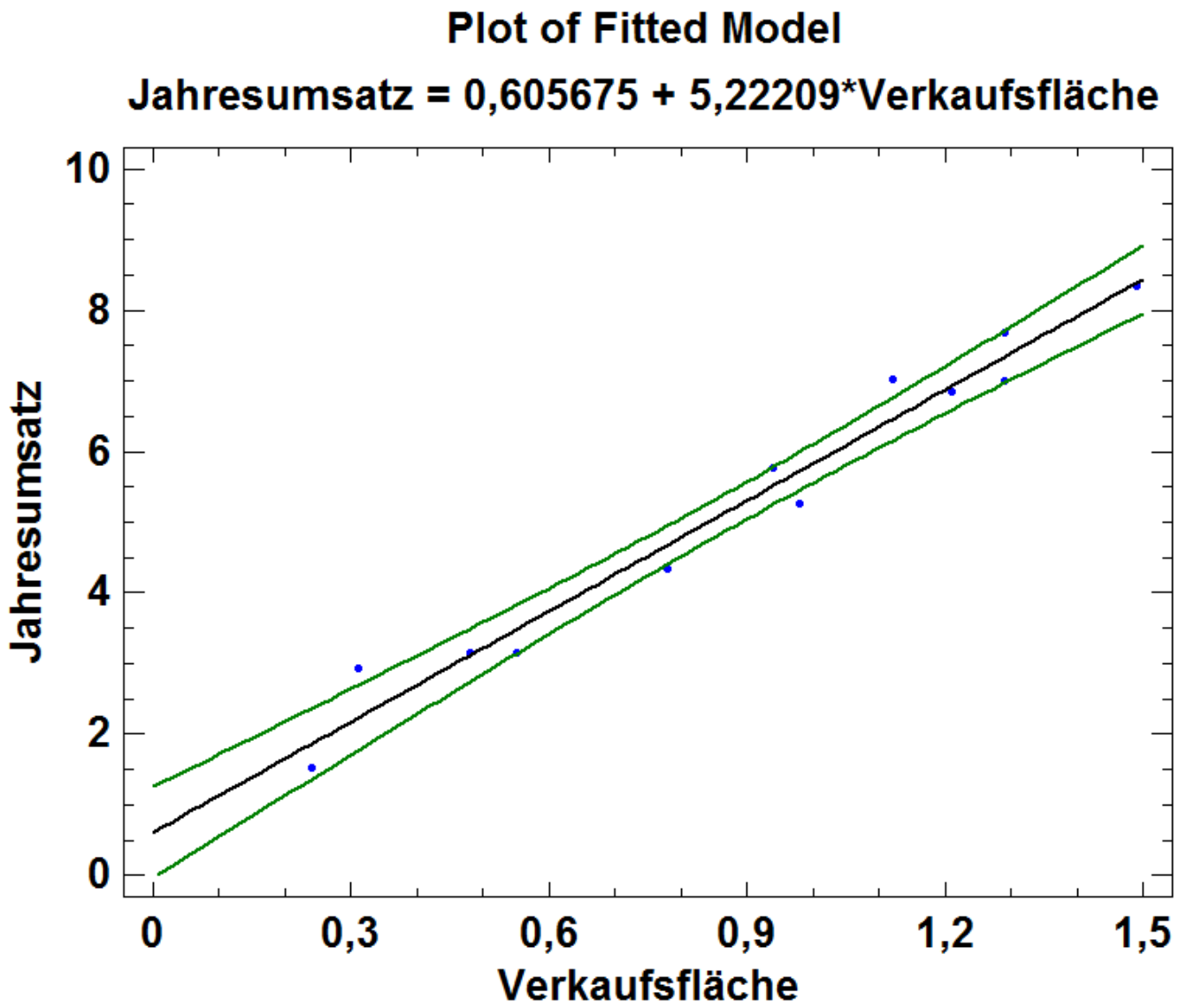
$$[\,\hat{y}(x) - d(x)\,;\, \hat{y}(x) + d(x)\,]$$

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

$$d(x) = \hat{\sigma} \cdot t_{n-2,1-\alpha/2} \sqrt{\frac{1}{n} + \frac{(\overline{x} - x)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}\,.$$

- Für unterschiedliche Werte x erhält man unterschiedliche Abstände zwischen der oberen und unteren Grenze. Für alle $x-$Werte betrachtet ergibt sich ein Konfidenzstreifen (Konfidenzschlauch), der an der Stelle $x = \overline{x}$ am schmalsten ist.

Konfidenzstreifen im Beispiel (Statgraphics)



Prognoseintervalle für $Y(x)$

- Berechnet man ein zufälliges Intervall, welches mit einer vorgegebenen Wahrscheinlichkeit $1-\alpha$ eine Realisierung von $Y(x) = b_0 + b_1 x + \varepsilon$ überdeckt (Vorhersage für **eine neue Beobachtung** an einer Stelle x), erhält man ein sogenanntes Prognoseintervall für $Y(x)$ zum Niveau $1-\alpha$.
- Unter den gemachten Voraussetzungen berechnet man

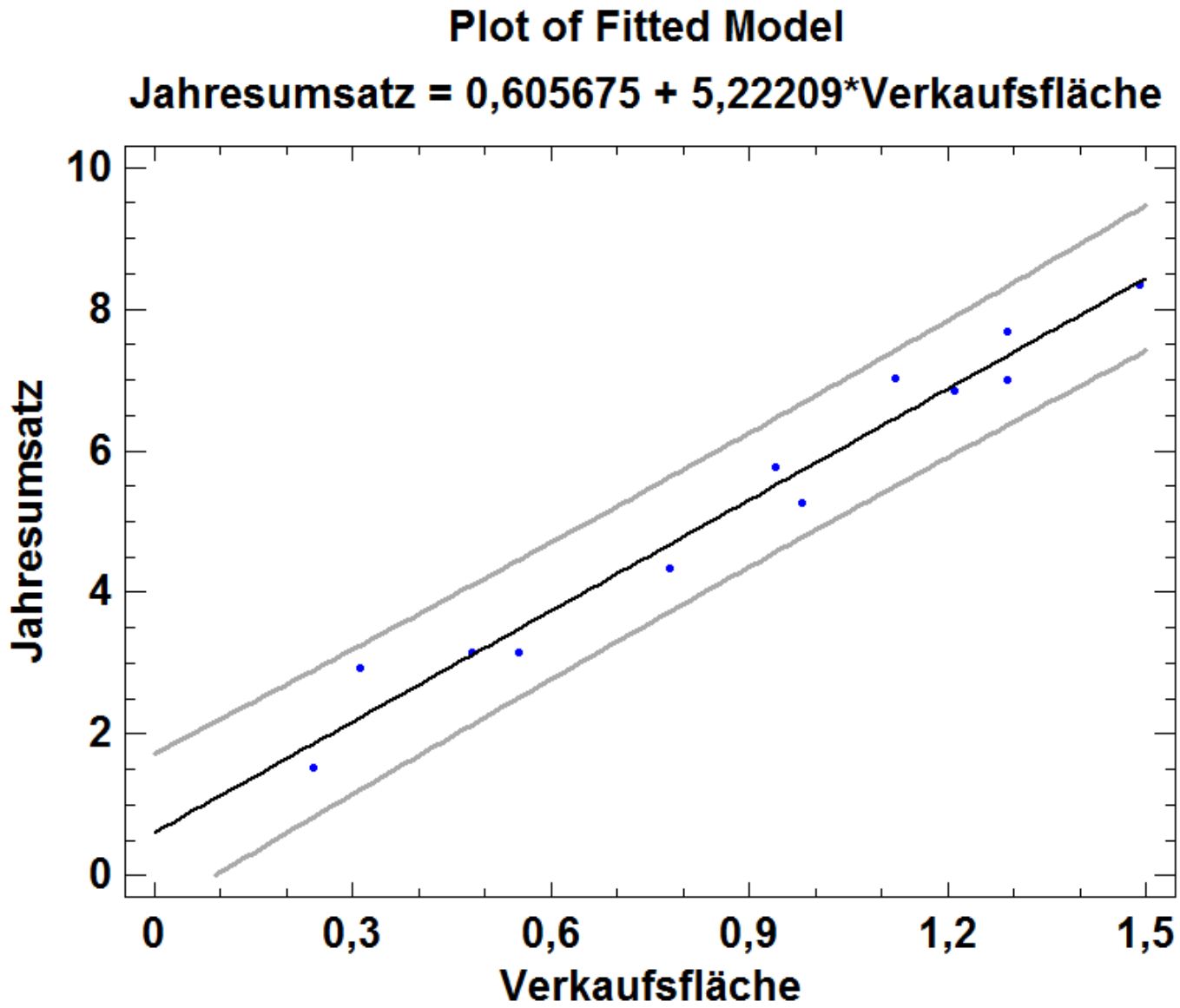
$$[\,\hat{y}(x) - d(x)\,;\, \hat{y}(x) + d(x)\,]$$

mit $\hat{y}(x) = \hat{b}_0 + \hat{b}_1\, x$ und

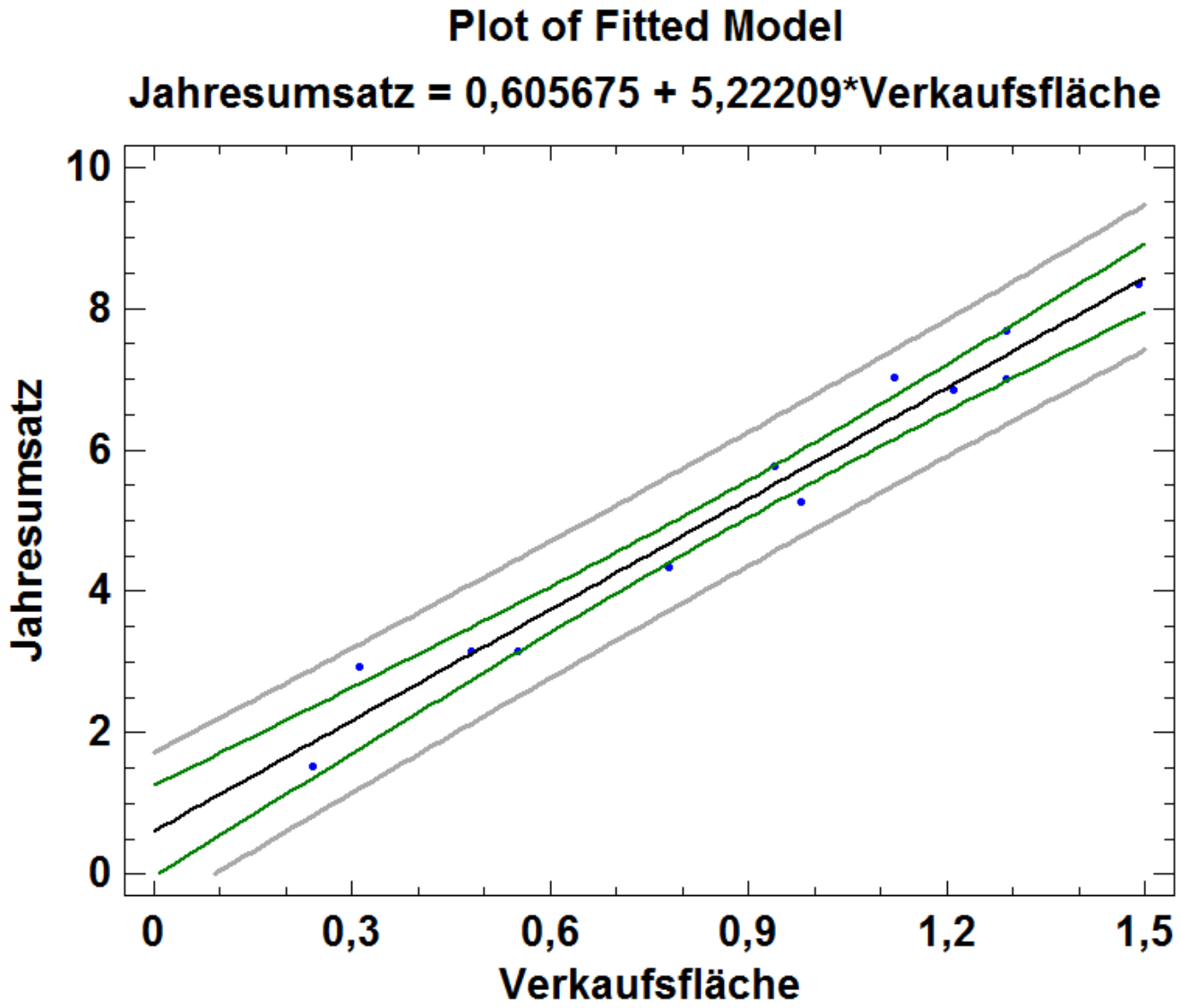
$$d(x) = \hat{\sigma} \cdot t_{n-2,1-\alpha/2} \sqrt{1 + \frac{1}{n} + \frac{(\overline{x} - x)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}\,.$$

- Bei Betrachtung beliebiger $x-$Werte erhält man wieder einen Streifen um die Regressionsgerade, den Prognosestreifen. Er ist breiter als der zugehörige Konfidenzstreifen zum selben Niveau.

Prognosestreifen im Beispiel (Statgraphics)



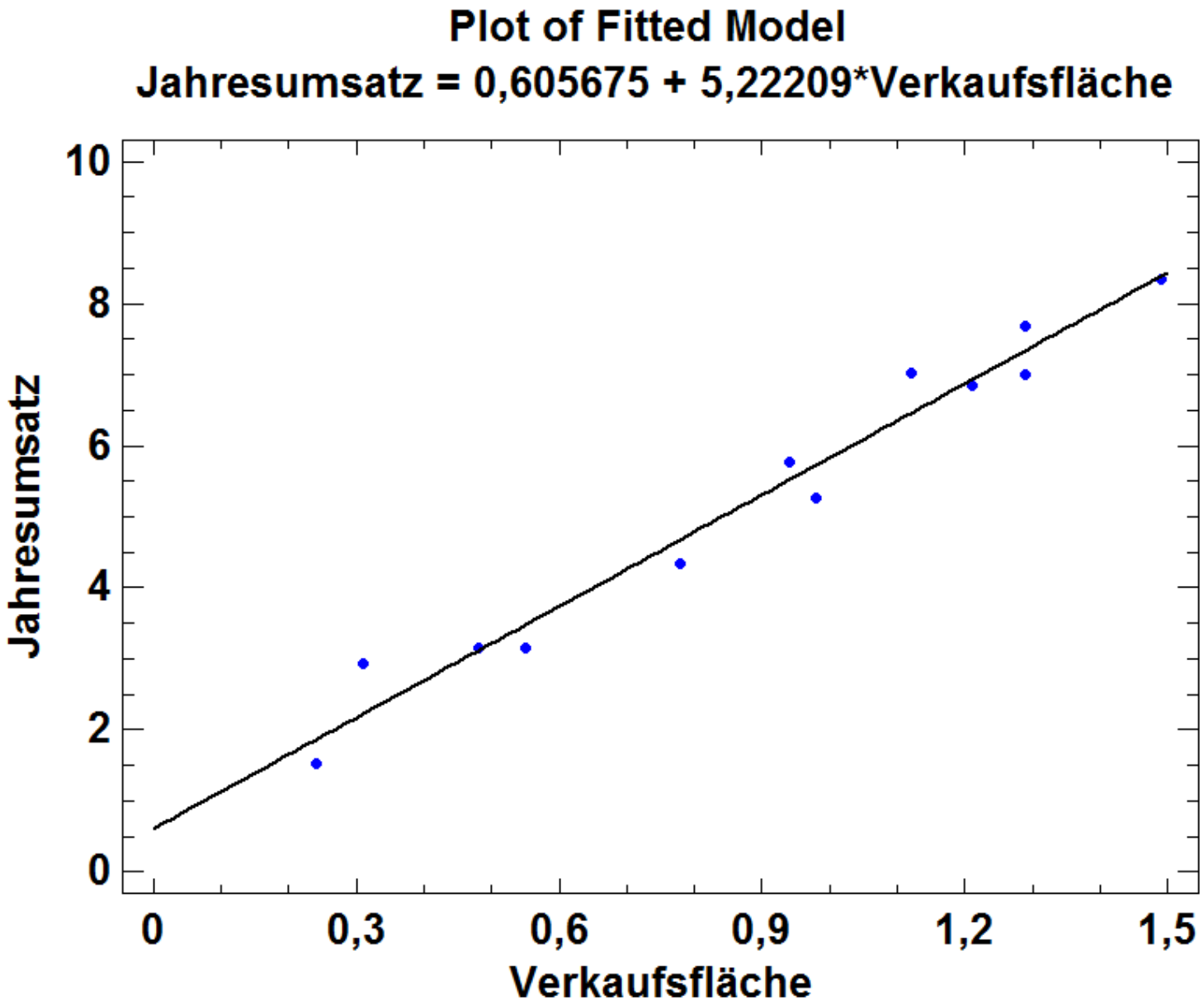
Konfidenz- und Prognosestreifen im Beispiel (Statgraphics)



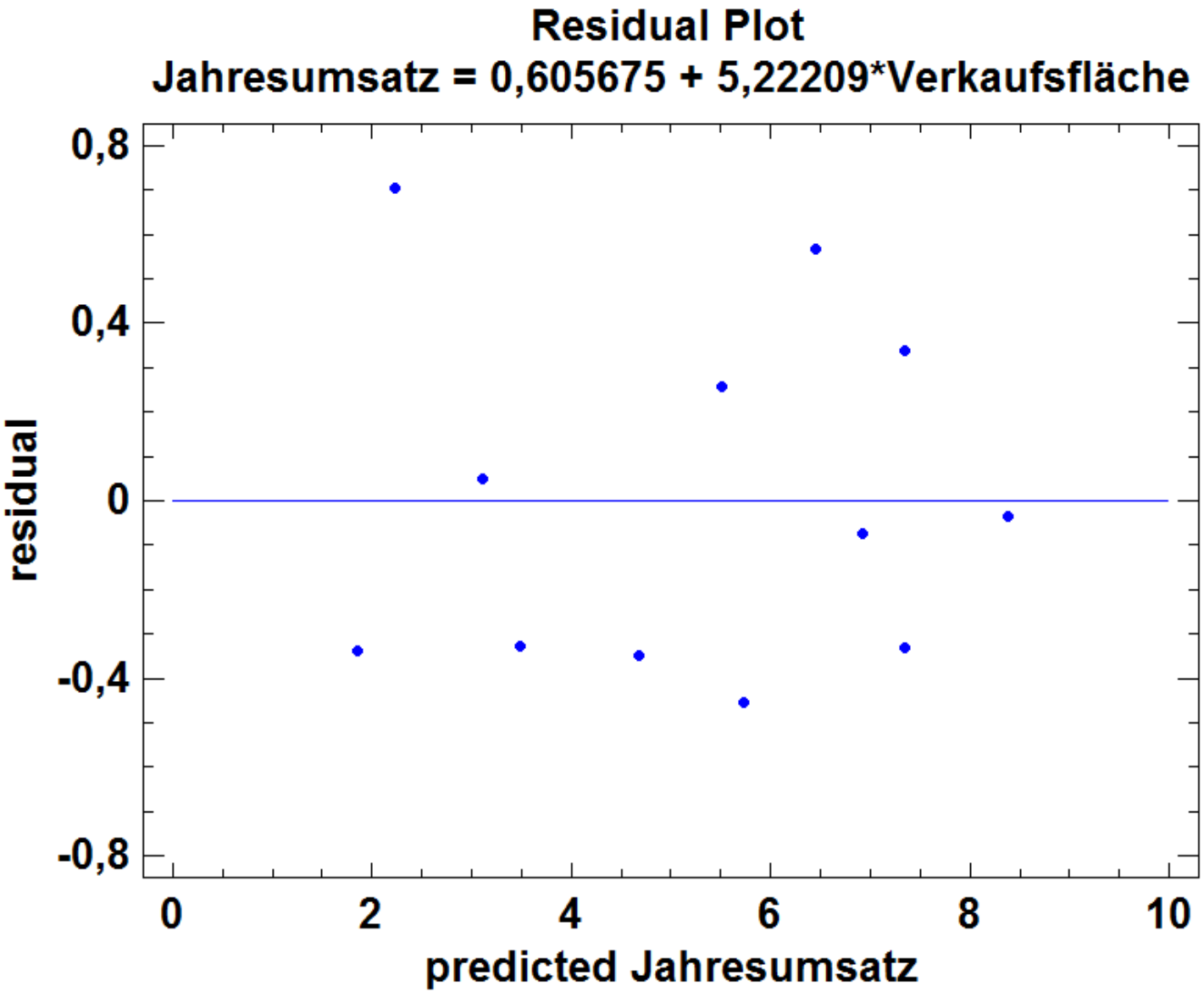
Residualanalyse zur Überprüfung des Modells

- Ist der Modellansatz, insbesondere die Normalverteilungsannahme der zufälligen Fehler $\varepsilon_i,\ i = 1, \ldots, n,$ richtig, dann sind die Residuen $\hat{\varepsilon}_i = Y_i - \hat{y}_i$ näherungsweise unabhängig und identisch normalverteilt.
- Diese Eigenschaft kann anschaulich grafisch (Residualanalyse) überprüft oder durch Anwendung statistischer Tests untersucht werden.
- Die Residualanalyse ist ein deskriptives Verfahren zur Überprüfung der Modellannahmen an $\varepsilon_1; \ldots; \varepsilon_n$. Mögliche Teilschritte sind dabei:
  - A: Streudiagramm der Daten mit der Regessionsgerade,
  - B: Streudiagramm der Residuen gegen die vorhergesagten Werte $\hat{y}_i$ (oder z.B. auch gegen die Fallnummern der x_i-Werte),
  - C: Normalverteilungs-Q-Q-Plot der Residuen,
  - D: Histogramm der Residuen mit angepasster Normalverteilungsdichte.

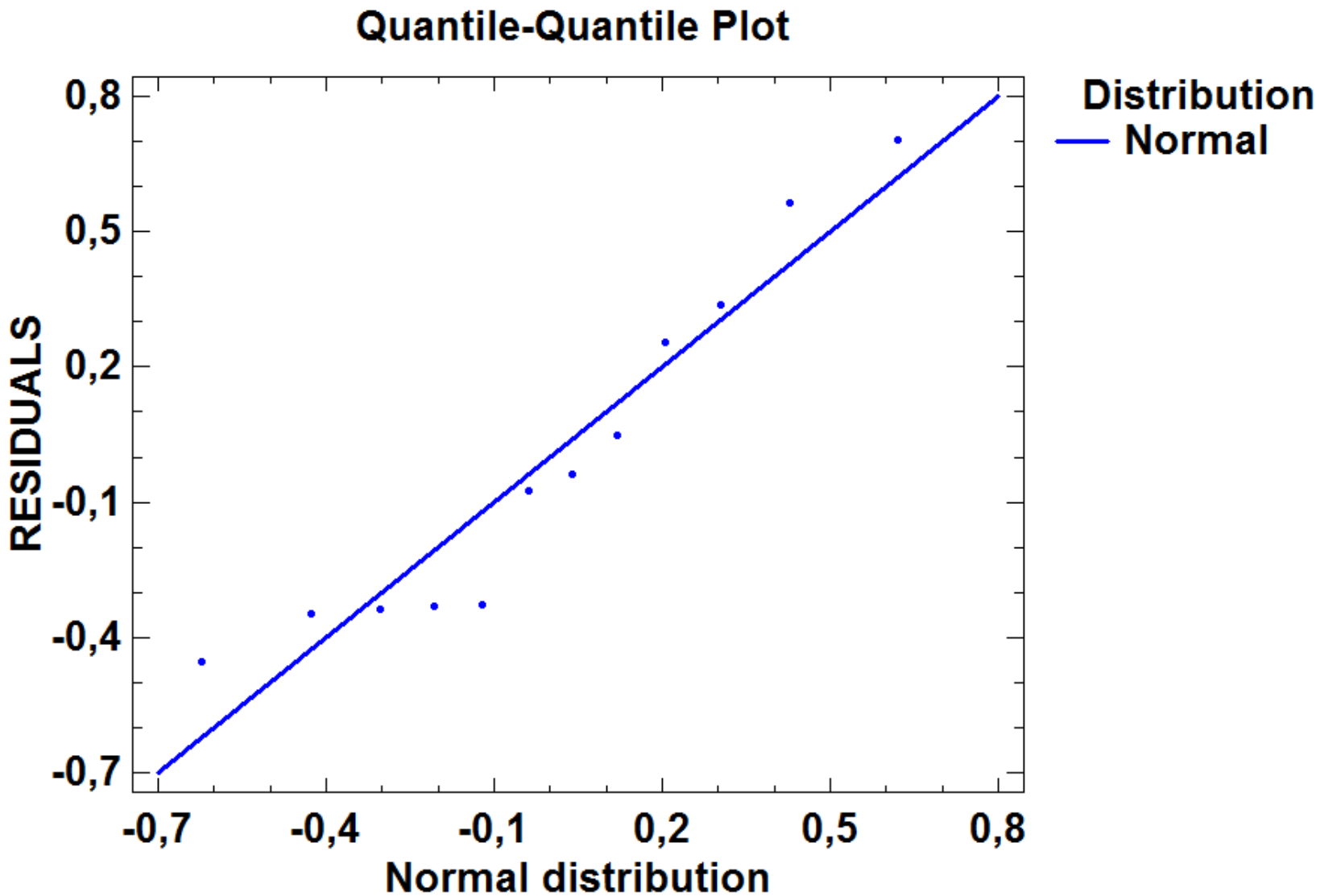
Streudiagramm und Regressionsgerade im Beispiel 7.1



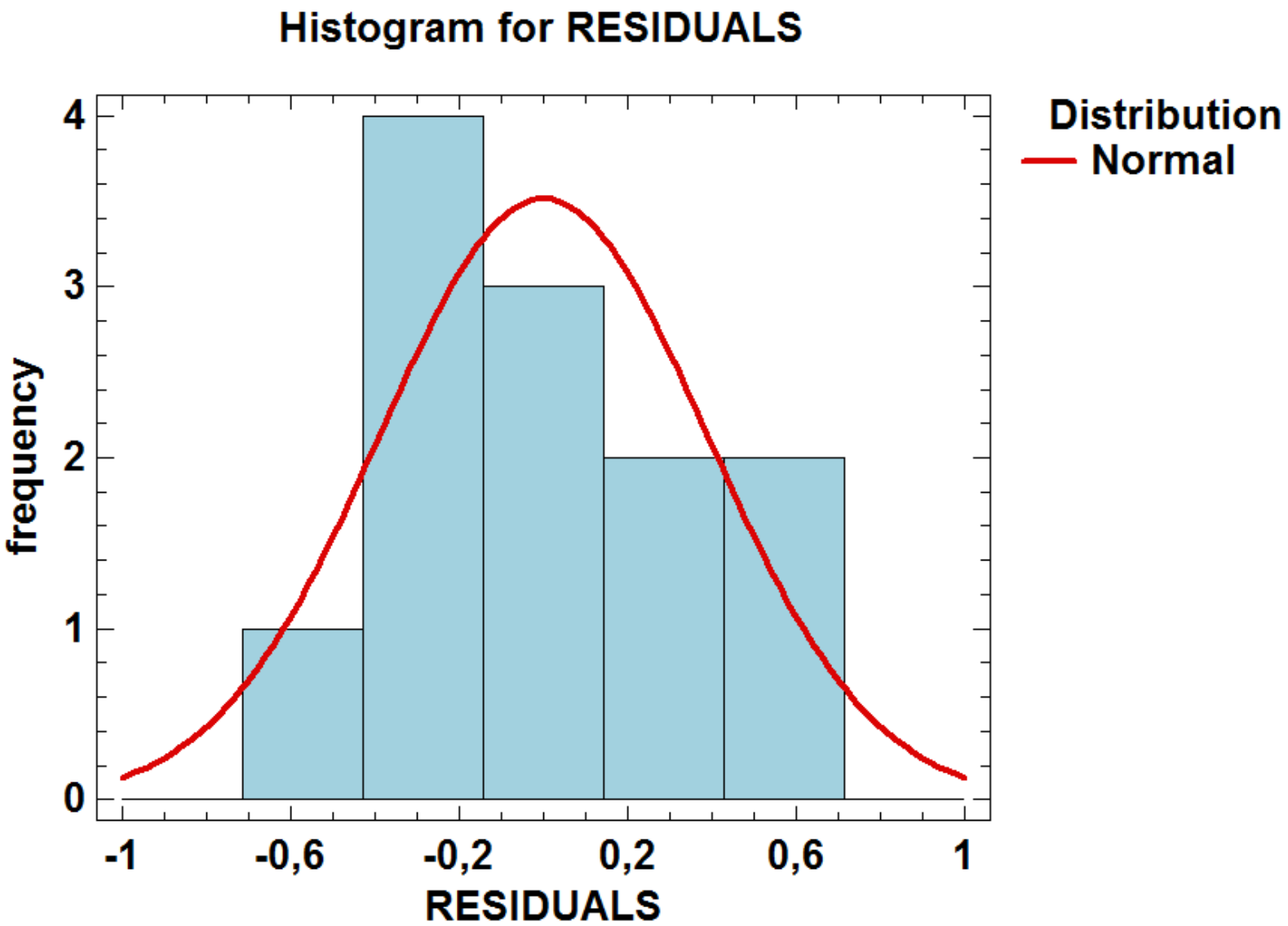
Streudiagramm der Residuen im Beispiel 7.1



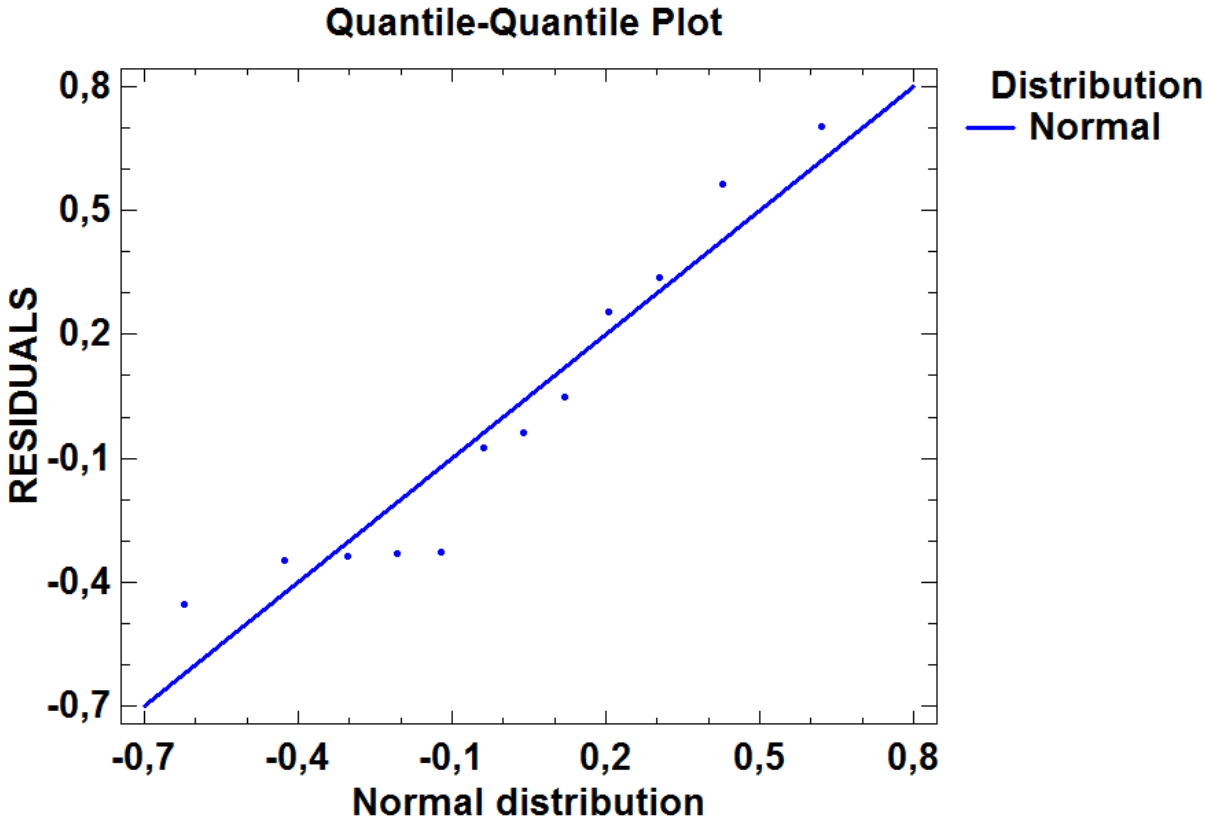
Normalverteilungs-Q-Q-Plot im Beispiel 7.1 (Statgraphics)



Histogramm der Residuen im Beispiel 7.1 (Statgraphics)



Shapiro-Wilk-Test im Beispiel 7.1 (Statgraphics)



Tests for Normality for RESIDUALS

Test	*Statistic*	*P-Value*
Shapiro-Wilk W	0,906018	0,18088