Regressions- & Korrelationsanalysator

Mathematische Dokumentation

Deutsch

Wolfgang Böhme

wolfgang.boehme@economic-modeling.de

www.economic-modeling.de

6. November 2026

Dokumentationen dieser Webseite wurden mit Unterstützung der KI Claude (Anthropic) entwickelt.

Inhaltsverzeichnis

1 Deskriptive Statistik **3**
- 1.1 Lagemaße . . . 3
  - 1.1.1 Arithmetisches Mittel . . . 3
  - 1.1.2 Median . . . 3
  - 1.1.3 Quantile . . . 3
- 1.2 Streuungsmaße . . . 3
  - 1.2.1 Varianz . . . 3
  - 1.2.2 Standardabweichung . . . 3
  - 1.2.3 Interquartilsabstand (IQR) . . . 4
- 1.3 Formparameter . . . 4
  - 1.3.1 Schiefe (Skewness) . . . 4
  - 1.3.2 Kurtosis (Wölbung) . . . 4

2 Korrelationsanalyse **4**
- 2.1 Pearson-Korrelation . . . 4
- 2.2 Spearman-Rangkorrelation . . . 5

3 Einfache Lineare Regression **5**
- 3.1 Modell . . . 5
- 3.2 Parameterschätzung (OLS) . . . 5
- 3.3 Bestimmtheitsmaß (R^2) . . . 5
- 3.4 Fehlermaße . . . 5
  - 3.4.1 Root Mean Square Error (RMSE) . . . 5
  - 3.4.2 Mean Absolute Error (MAE) . . . 5

4 Polynomiale Regression **6**
- 4.1 Modell vom Grad p . . . 6
- 4.2 Matrixschreibweise . . . 6
- 4.3 Normalengleichungen . . . 6

5 Multiple Lineare Regression **6**
- 5.1 Modell . . . 6
- 5.2 Adjustiertes R^2 . . . 6
- 5.3 Multikollinearität: Variance Inflation Factor (VIF) . . . 6

6 Nichtlineare Regressionsmodelle **7**
- 6.1 Exponentialmodell . . . 7
- 6.2 Potenzmodell . . . 7
- 6.3 Logarithmisches Modell . . . 7
- 6.4 Logistische Funktion . . . 7
- 6.5 Gompertz-Funktion . . . 7
- 6.6 Gauß-Funktion . . . 7

7 Residuenanalyse **7**
- 7.1 Residuen . . . 7
- 7.2 Standardisierte Residuen . . . 8
- 7.3 Leverage (Hebelwerte) . . . 8

7.4 Cook's Distance . . . 8
7.5 Q-Q Plot . . . 8

8 Modellselektion **8**
8.1 Akaike Information Criterion (AIC) . . . 8

9 Beispiel 1: Konsumfunktion USA 1960–2008 **9**
9.1 Datensatz . . . 9
9.2 Deskriptive Statistik . . . 9
9.3 Korrelationsanalyse . . . 10
9.4 Einfache lineare Regression: Keynesianische Konsumfunktion . . . 10
9.5 Residuenanalyse . . . 10
9.6 Multiple Regression . . . 11
9.7 Modellvergleich: Wachstum über die Zeit . . . 11

10 Beispiel 2: Engel-Kurve **11**
10.1 Potenzmodell und Einkommenselastizität . . . 11
10.2 Multiple Regression mit der Haushaltsgröße . . . 12

11 Beispiel 3: Cobb-Douglas-Produktionsfunktion **12**
11.1 Linearisierung und multiple Regression . . . 12
11.2 Intensive Form . . . 12

12 Zusammenfassung: Modellvergleich **13**

13 Literatur **13**

14 Anhang: Numerische Methoden **13**
14.1 Gauß-Elimination . . . 13
14.2 Approximation der inversen Fehlerfunktion . . . 14

Zusammenfassung

Diese Dokumentation beschreibt die mathematischen Grundlagen und Formeln des Regressions- und Korrelationsanalysators. Es werden alle implementierten statistischen Verfahren formal definiert, einschließlich deskriptiver Statistiken, Korrelationsanalysen, linearer und nichtlinearer Regressionsmodelle sowie diagnostischer Verfahren. Drei ökonomische Beispiele – Konsumfunktion (echte Daten), Engel-Kurve und Cobb-Douglas-Produktionsfunktion – illustrieren die Anwendung der Methoden; sie sind im Tool als Beispieldatensätze enthalten.

1 Deskriptive Statistik

Sei $X = \{x_1, x_2, \ldots, x_n\}$ eine Stichprobe mit n Beobachtungen.

1.1 Lagemaße

1.1.1 Arithmetisches Mittel

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

1.1.2 Median

Der Median ist definiert als:

$$\text{Median}(X) = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{falls } n \text{ ungerade} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{falls } n \text{ gerade} \end{cases} \tag{2}$$

wobei $x_{(i)}$ die i-te Ordnungsstatistik bezeichnet (sortierte Werte).

1.1.3 Quantile

Das p-Quantil ($0 \leq p \leq 1$) ist definiert durch lineare Interpolation:

$$Q_p = x_{(k)} + (x_{(k+1)} - x_{(k)}) \cdot r \tag{3}$$

wobei $h = p(n-1)$, $k = \lfloor h \rfloor$ und $r = h - k$.

1.2 Streuungsmaße

1.2.1 Varianz

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{4}$$

1.2.2 Standardabweichung

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{5}$$

1.2.3 Interquartilsabstand (IQR)

$$\text{IQR} = Q_{0.75} - Q_{0.25} \tag{6}$$

1.3 Formparameter

1.3.1 Schiefe (Skewness)

Die Schiefe misst die Asymmetrie der Verteilung:

$$\gamma_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3 \tag{7}$$

Interpretation:

- $\gamma_1 \approx 0$: symmetrische Verteilung
- $\gamma_1 > 0$: rechtsschiefe Verteilung (rechter Schwanz länger)
- $\gamma_1 < 0$: linksschiefe Verteilung (linker Schwanz länger)

1.3.2 Kurtosis (Wölbung)

Die Kurtosis misst die „Schwere der Ausläufer:

$$\gamma_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} \tag{8}$$

Interpretation (Exzesskurtosis):

- $\gamma_2 \approx 0$: mesokurtisch (wie Normalverteilung)
- $\gamma_2 > 0$: leptokurtisch (schwere Ausläufer, spitzer Gipfel)
- $\gamma_2 < 0$: platykurtisch (leichte Ausläufer, flacher Gipfel)

2 Korrelationsanalyse

Seien $X = \{x_1, \ldots, x_n\}$ und $Y = \{y_1, \ldots, y_n\}$ zwei Variablen.

2.1 Pearson-Korrelation

Der Pearson-Korrelationskoeffizient misst die lineare Abhängigkeit:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{9}$$

Äquivalente Formel:

$$r_{XY} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \cdot \sqrt{n \sum y_i^2 - (\sum y_i)^2}} \tag{10}$$

Es gilt: $-1 \leq r_{XY} \leq 1$

2.2 Spearman-Rangkorrelation

Für Ränge $\text{rang}(x_i)$ und $\text{rang}(y_i)$:

$$\rho_{XY} = r_{\text{rang}(X),\text{rang}(Y)} \tag{11}$$

Dies ist der Pearson-Korrelationskoeffizient der Ränge, robust gegenüber Ausreißern.

3 Einfache Lineare Regression

3.1 Modell

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n \tag{12}$$

wobei $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

3.2 Parameterschätzung (OLS)

Minimierung der Fehlerquadratsumme:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{13}$$

Schätzer:

$$\hat{\beta}_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \tag{14}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{15}$$

3.3 Bestimmtheitsmaß (R^2)

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{16}$$

wobei $\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ die totale Quadratsumme ist.
Interpretation: R^2 gibt den Anteil der erklärten Varianz an ($0 \leq R^2 \leq 1$).

3.4 Fehlermaße

3.4.1 Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{17}$$

3.4.2 Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{18}$$

4 Polynomiale Regression

4.1 Modell vom Grad p

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \varepsilon_i \tag{19}$$

4.2 Matrixschreibweise

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{20}$$

wobei

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{pmatrix} \tag{21}$$

4.3 Normalengleichungen

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{22}$$

Die Lösung erfolgt durch Gauß-Elimination.

5 Multiple Lineare Regression

5.1 Modell

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{23}$$

5.2 Adjustiertes $\mathbf{R^2}$

Das adjustierte R^2 berücksichtigt die Anzahl der Parameter:

$$R^2_{\text{adj}} = 1 - \frac{(1 - R^2)(n-1)}{n - p - 1} \tag{24}$$

wobei p die Anzahl der Prädiktoren ist.

5.3 Multikollinearität: Variance Inflation Factor (VIF)

Für jeden Prädiktor x_j:

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \tag{25}$$

wobei R_j^2 das Bestimmtheitsmaß der Regression von x_j auf alle anderen Prädiktoren ist.

Interpretation:

- $\text{VIF}_j < 5$: keine bedenkliche Multikollinearität
- $5 \leq \text{VIF}_j \leq 10$: moderate Multikollinearität
- $\text{VIF}_j > 10$: hohe Multikollinearität (problematisch)

6 Nichtlineare Regressionsmodelle

6.1 Exponentialmodell

$$y = a \cdot e^{bx} \tag{26}$$

Linearisierung: $\ln(y) = \ln(a) + bx$

6.2 Potenzmodell

$$y = a \cdot x^b \tag{27}$$

Linearisierung: $\ln(y) = \ln(a) + b\ln(x)$

6.3 Logarithmisches Modell

$$y = a + b\ln(x) \tag{28}$$

6.4 Logistische Funktion

$$y = y_{\min} + \frac{L}{1 + e^{-k(x-x_0)}} \tag{29}$$

wobei:

- L: Kapazität (Differenz zwischen Maximum und Minimum)
- k: Wachstumsrate
- x_0: Wendepunkt

6.5 Gompertz-Funktion

$$y = a \cdot e^{-b \cdot e^{-cx}} \tag{30}$$

Asymmetrische Wachstumskurve, häufig in der Biologie.

6.6 Gauß-Funktion

$$y = a \cdot e^{-\frac{(x-b)^2}{2c^2}} \tag{31}$$

Symmetrische Glockenkurve mit:

- a: Amplitude (Maximum)
- b: Zentrum (Mittelwert)
- c: Breite (Standardabweichung)

7 Residuenanalyse

7.1 Residuen

$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n \tag{32}$$

7.2 Standardisierte Residuen

$$r_i = \frac{e_i}{s_e} \tag{33}$$

wobei $s_e = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$ die Standardabweichung der Residuen ist.

7.3 Leverage (Hebelwerte)

Für einfache lineare Regression:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \tag{34}$$

Es gilt: $0 \leq h_i \leq 1$ und $\sum_{i=1}^{n} h_i = 2$ (einfache Regression).
Grenzwert: $h_i > 2p/n$ deutet auf hohen Leverage hin (mit p = Anzahl Parameter).

7.4 Cook's Distance

$$D_i = \frac{r_i^2}{p} \cdot \frac{h_i}{1 - h_i} \tag{35}$$

Interpretation: $D_i > 1$ deutet auf einflussreichen Datenpunkt hin.

7.5 Q-Q Plot

Vergleicht die Quantile der standardisierten Residuen mit den theoretischen Quantilen der Standardnormalverteilung:

$$\Phi^{-1}\left(\frac{i - 0.5}{n}\right) \quad \text{vs.} \quad r_{(i)} \tag{36}$$

wobei $r_{(i)}$ die sortierten standardisierten Residuen sind und Φ^{-1} die Quantilfunktion der Standardnormalverteilung.

8 Modellselektion

8.1 Akaike Information Criterion (AIC)

$$\text{AIC} = n \ln(\text{RSS}/n) + 2k \tag{37}$$

wobei:

- $\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$: Residual Sum of Squares
- k: Anzahl der geschätzten Regressionskoeffizienten (linear, exponentiell, logarithmisch, Potenz: $k = 2$; Polynom Grad p: $k = p + 1$; multiple Regression mit p Variablen: $k = p + 1$)

Prinzip: Wähle das Modell mit dem kleinsten AIC (Balance zwischen Anpassungsgüte und Modellkomplexität). Das Tool sortiert die Modelle im Modellvergleich nach R^2 und zeigt das AIC zusätzlich an.

9 Beispiel 1: Konsumfunktion USA 1960–2008

Das Tool enthält drei ökonomische Beispieldatensätze (Auswahlliste neben *Beispiel laden*). Beispiel 1 wird hier ausführlich durchgerechnet, die Beispiele 2 und 3 in den folgenden Abschnitten. Alle Zahlen entsprechen den Ausgaben des Tools.

9.1 Datensatz

49 Jahreswerte 1960–2008: Konsum = reale private Konsumausgaben, Einkommen = reales verfügbares Einkommen, BIP = reales Bruttoinlandsprodukt (jeweils Mrd. US-Dollar, Preise von 2005), Zins = 3-Monats-Treasury-Bill-Satz (%), Arbeitslosenquote (%), Inflation (%). Quelle: Federal Reserve Bank of St. Louis (FRED) und U.S. Bureau of Labor Statistics (Datensatz „macrodata“ des Python-Pakets statsmodels), Jahresdurchschnitte aus Quartalsdaten.

Jahr	Konsum	Einkommen	Zins	Arbeitslosenquote
1960	1784.4	1964.0	2.71	5.58
1965	2241.8	2514.6	4.01	4.53
1970	2740.1	3108.5	6.11	5.00
1975	3214.1	3691.4	5.66	8.48
1980	3766.2	4295.0	11.68	7.18
1985	4540.4	5144.8	7.42	7.20
1990	5316.2	5896.3	7.38	5.60
1995	6079.0	6651.7	5.43	5.62
2000	7608.1	8161.6	5.80	3.95
2005	8819.0	9277.3	3.30	5.08
2008	9290.9	9911.1	1.15	5.80

Tabelle 1: Auszug aus Beispiel 1 (jedes fünfte Jahr; vollständige Daten im Tool)

9.2 Deskriptive Statistik

	Konsum (Y)	Einkommen (X)
Mittelwert	4821.1	5307.9
Standardabweichung	2241.1	2340.0
Median	4315.6	4989.0
$Q_{0.25}$ / $Q_{0.75}$	3019.5 / 6291.2	3406.5 / 6871.0
IQR	3271.7	3464.5
Schiefe	0.565	0.464
Kurtosis	−0.739	−0.806

Beide Reihen sind leicht rechtsschief (Mittelwert > Median): Wegen des Wachstums liegen in den späten Jahren höhere Werte weiter auseinander.

9.3 Korrelationsanalyse

Mit $n = 49$, $\sum x_i = 260085$, $\sum y_i = 236234$, $\sum x_i y_i = 1\,510\,639\,978.65$ und $\sum x_i^2 = 1\,648\,793\,107.72$:

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 256\,743\,654.4, \qquad \sum (x_i - \bar{x})^2 = 268\,299\,082.7,$$
$$r_{XY} = 0.9992, \qquad \rho_{XY} \text{ (Spearman)} = 0.9998.$$

Interpretation: Nahezu perfekter linearer Zusammenhang. Vorsicht: Bei trendbehafteten Zeitreihen ist eine hohe Korrelation zum Teil Folge des gemeinsamen Wachstums. Konsum korreliert schwach negativ mit dem Zins ($r = -0.29$) und der Arbeitslosenquote ($r = -0.14$).

9.4 Einfache lineare Regression: Keynesianische Konsumfunktion

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{256\,743\,654.4}{268\,299\,082.7} = 0.9569,$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 4821.10 - 0.9569 \cdot 5307.86 = -258.15$$

Regressionsgleichung:

$$\boxed{\widehat{\text{Konsum}} = -258.15 + 0.9569 \cdot \text{Einkommen}} \tag{38}$$

- Die Steigung ist die **marginale Konsumquote**: Von jedem zusätzlichen Dollar verfügbaren Einkommens werden rund 96 Cent konsumiert, rund 4 Cent gespart.
- Die durchschnittliche Konsumquote $\bar{y}/\bar{x} = 0.908$ liegt darunter; wegen $\hat{\beta}_0 < 0$ steigt sie mit dem Einkommen leicht an.

Gütemaße:

$$\text{SST} = 246\,097\,042, \quad \text{SSE} = 411\,133, \quad R^2 = 1 - \frac{411\,133}{246\,097\,042} = 0.9983,$$
$$\text{RMSE} = \sqrt{411\,133/49} = 91.60, \quad \text{MAE} = 76.48 \quad \text{(Mrd. US-Dollar)}.$$

Vorhersage für ein Einkommen von 10 000: $\hat{y} = -258.15 + 0.9569 \cdot 10\,000 = 9311.2$.

9.5 Residuenanalyse

Standardabweichung der Residuen $s_e = 91.60$. Beispiel Jahr 2005 (größte Cook's Distance):

$$\hat{y} = -258.15 + 0.9569 \cdot 9277.3 = 8619.58, \quad e = 8819.0 - 8619.58 = 199.42,$$
$$r = 199.42/91.60 = 2.18, \quad h = \tfrac{1}{49} + \tfrac{(9277.3 - 5307.86)^2}{268\,299\,082.7} = 0.079,$$
$$D = \tfrac{2.18^2}{2} \cdot \tfrac{0.079}{1 - 0.079} = 0.204.$$

Interpretation:

- Nur zwei Jahre haben $|r| > 2$ (1984: -2.19, 2005: 2.18); alle $D_i < 1$ – keine einflussreichen Ausreißer.

- Im Residualplot folgen positive und negative Residuen in längeren Phasen aufeinander: Die Residuen sind **autokorreliert** (Korrelation aufeinanderfolgender Residuen 0.82, Durbin-Watson-Statistik 0.35; beide Werte berechnet das Tool nicht). Typisch für Zeitreihen – die Standardannahme unabhängiger Fehler ist verletzt, R^2 wirkt deshalb besser, als es ist.

9.6 Multiple Regression

Mit Einkommen, Zins und Arbeitslosenquote als Regressoren:

$$\widehat{\text{Konsum}} = 56.84 + 0.9494 \cdot \text{Einkommen} - 18.21 \cdot \text{Zins} - 30.17 \cdot \text{Arbeitslosenquote} \quad (39)$$

$R^2 = 0.99938$, $R^2_{\text{adj}} = 0.99933$, RMSE = 56.00 (statt 91.60 bei der einfachen Regression). VIF: 1.08, 1.22, 1.14 – keine Multikollinearität.

Interpretation: Ein um einen Prozentpunkt höherer Zins senkt den Konsum bei gegebenem Einkommen um rund 18 Mrd. US-Dollar, ein Prozentpunkt mehr Arbeitslosigkeit um rund 30 Mrd. US-Dollar (Vorsichtssparen). Die marginale Konsumquote bleibt mit 0.95 nahezu unverändert.

9.7 Modellvergleich: Wachstum über die Zeit

Mit Y = Konsum und X = Jahr zeigt das Modell *Exponentiell*

$$\widehat{\text{Konsum}} = a \cdot e^{0.0340 \cdot \text{Jahr}}, \qquad R^2 = 0.9967, \quad (40)$$

ein durchschnittliches reales Konsumwachstum von $e^{0.034} - 1 \approx 3.5\,\%$ pro Jahr (a ist wegen der Jahreszahlen als X extrem klein und nicht interpretierbar).

10 Beispiel 2: Engel-Kurve

40 Haushalte (fiktive, realistische Werte): Einkommen = Haushaltsnettoeinkommen, Nahrung, Wohnen, Freizeit = monatliche Ausgaben (Euro), Haushaltsgröße (Personen), Nahrungsanteil (% des Einkommens).

10.1 Potenzmodell und Einkommenselastizität

Im Potenzmodell $y = a \cdot x^b$ ist $b = \dfrac{\mathrm{d}\ln y}{\mathrm{d}\ln x}$ die **Einkommenselastizität**. Ergebnisse mit X = Einkommen (Tab *Einfache Regression*):

Y	a	b	R^2	Einordnung
Nahrung	14.84	0.412	0.718	notwendiges Gut ($0 < b < 1$)
Wohnen	0.587	0.885	0.971	notwendiges Gut, nahe 1
Freizeit	0.00903	1.296	0.960	Luxusgut ($b > 1$)

Tabelle 2: Einkommenselastizitäten (Potenzmodell)

Engelsches Gesetz: Der Nahrungsanteil sinkt mit dem Einkommen ($r = -0.84$). Potenzmodell für den Anteil: $\widehat{\text{Nahrungsanteil}} = 1478.9 \cdot \text{Einkommen}^{-0.588}$, $R^2 = 0.848$ – es passt deutlich besser als die Gerade ($R^2 = 0.706$), da der Anteil konvex fällt. Erwartungsgemäß gilt $-0.588 \approx 0.412 - 1$.

10.2 Multiple Regression mit der Haushaltsgröße

Für Nahrung erklärt das Einkommen allein nur 72 % der Varianz. Mit der Haushaltsgröße:

$$\widehat{\text{Nahrung}} = 91.32 + 0.0538 \cdot \text{Einkommen} + 46.47 \cdot \text{Haushaltsgröße} \tag{41}$$

$R^2 = 0.916$, $R^2_{\text{adj}} = 0.911$, VIF $= 1.05$. Jede weitere Person erhöht die Nahrungsausgaben bei gegebenem Einkommen um rund 46 Euro im Monat; von jedem zusätzlichen Euro Einkommen gehen rund 5 Cent in Nahrung.

11 Beispiel 3: Cobb-Douglas-Produktionsfunktion

30 Industriebetriebe (fiktive, realistische Werte): Output = Wertschöpfung, Kapital = Kapitalstock (jeweils Tsd. Euro), Beschäftigte; dazu Produktivität (Output je Beschäftigten), Kapitalintensität (Kapital je Beschäftigten) und die natürlichen Logarithmen.

11.1 Linearisierung und multiple Regression

$$Y = A\,K^{\alpha}L^{\beta} \quad \Longleftrightarrow \quad \ln Y = \ln A + \alpha \ln K + \beta \ln L \tag{42}$$

Multiple Regression von ln_Output auf ln_Kapital und ln_Beschäftigte:

$$\widehat{\ln Y} = 0.933 + 0.3113 \cdot \ln K + 0.6920 \cdot \ln L, \qquad R^2 = 0.9982,\ R^2_{\text{adj}} = 0.9981. \tag{43}$$

- **Produktionselastizitäten:** 1 % mehr Kapital erhöht den Output um 0.31 %, 1 % mehr Beschäftigte um 0.69 %.
- **Skalenerträge:** $\alpha + \beta = 1.003 \approx 1$ – konstante Skalenerträge.
- $A = e^{0.933} = 2.54$ (totale Faktorproduktivität).
- **Multikollinearität:** VIF = 11.1 (Tool: „Hoch“), denn große Betriebe haben viel Kapital und viele Beschäftigte ($r(\ln K, \ln L) = 0.95$). Die Schätzung bleibt hier dank geringer Störgrößen stabil; in realen Daten werden α und β dadurch oft ungenau.

Fehlspezifikation durch ausgelassene Variable: Die einfache Regression von ln_Output nur auf ln_Kapital liefert die Steigung 0.918 – weit über $\alpha = 0.31$, weil ln_Kapital den Einfluss der fehlenden Beschäftigten mit aufnimmt.

11.2 Intensive Form

Bei konstanten Skalenerträgen gilt $Y/L = A\,(K/L)^{\alpha}$. Potenzmodell mit Y = Produktivität, X = Kapitalintensität:

$$\widehat{\text{Produktivität}} = 2.568 \cdot \text{Kapitalintensität}^{0.312}, \qquad R^2 = 0.862. \tag{44}$$

Der Exponent 0.312 stimmt mit α aus der multiplen Regression überein.

Modell	R^2	RMSE	AIC
Linear	0.99833	91.60	446.71
Polynomial Grad 2	0.99952	49.15	387.69
Polynomial Grad 3	0.99964	42.74	376.00
Polynomial Grad 6	0.99968	40.39	376.47
Exponentiell	0.94932	504.51	613.91
Logarithmisch	0.93672	563.76	624.79
Potenz ($b = 1.035$)	0.99832	91.75	446.87
Multiple Regression	0.99938	56.00	402.49

Tabelle 3: Vergleich der Regressionsmodelle für die Konsumfunktion

12 Zusammenfassung: Modellvergleich

Tab *Modellvergleich* für Beispiel 1 mit Y = Konsum, X = Einkommen (AIC wie im Tool, Abschnitt Modellselektion):

Empfehlung: Das Tool ordnet die Modelle nach R^2; dabei gewinnt fast immer das Polynom höchsten Grades. Das AIC bestraft zusätzliche Parameter, hier liegen die Polynome ab Grad 3 praktisch gleichauf. Ökonomisch sinnvoll sind dennoch die lineare Konsumfunktion (interpretierbare marginale Konsumquote) oder die multiple Regression mit Zins und Arbeitslosenquote, die mit nur vier Parametern einen großen Teil der Verbesserung erreicht. Polynome hohen Grades passen sich der Stichprobe an und eignen sich nicht für Prognosen außerhalb des Datenbereichs.

13 Literatur

- Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: Models, Methods and Applications*. Springer.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. Wiley.
- Seber, G. A. F., & Wild, C. J. (2003). *Nonlinear Regression*. Wiley.
- Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis*. Wiley.

14 Anhang: Numerische Methoden

14.1 Gauß-Elimination

Zur Lösung des linearen Gleichungssystems $\mathbf{Ax} = \mathbf{b}$:

Algorithmus:

1. Vorwärtselimination: Bringe $\mathbf{A}$ in obere Dreiecksform
2. Rückwärtssubstitution: Löse für $\mathbf{x}$

14.2 Approximation der inversen Fehlerfunktion

Für den Q-Q Plot wird die inverse Fehlerfunktion erf^{-1} benötigt:

$$\text{erf}^{-1}(x) \approx \text{sgn}(x)\sqrt{\sqrt{\left(\frac{2}{\pi a}+\frac{\ln(1-x^2)}{2}\right)^2-\frac{\ln(1-x^2)}{a}}-\left(\frac{2}{\pi a}+\frac{\ln(1-x^2)}{2}\right)} \tag{45}$$

mit $a = 0.147$ (Winitzki-Approximation).
Die theoretischen Quantile für den Q-Q Plot sind dann:

$$q_i = \sqrt{2}\cdot\text{erf}^{-1}\left(2\cdot\frac{i-0.5}{n}-1\right), \quad i = 1,\ldots,n \tag{46}$$