

12.1.5 Nichtlineare Zusammenhänge

Regressionsgeraden liefern selbstverständlich nicht immer hinreichende Beschreibungen bivariater Abhängigkeitsbeziehungen. Besteht zwischen dem Regressanden Y und dem Regressor X ein ausgeprägt nichtlinearer empirischer Zusammenhang kann die Bestimmung von Y durch eine Regressionsgerade kaum zufriedenstellend gelingen.

Beispiel

B-12.3 Windkraftanlage.

In Beispiel 12.1 wurde der Zusammenhang zwischen Kosten und Output eines industriellen Produktionsprozesses funktional beschrieben. Jetzt wird ein energietechnischer Produktionsprozess betrachtet. Hierbei interessiert der Zusammenhang zwischen dem Gleichstromoutput Y (in Kilowatt) einer Windkraftanlage und der Windgeschwindigkeit X (in Stundenkilometer). Die nachfolgende Tabelle weist 25 gemessene Wertepaare der Variablen X und Y aus (verändert nach Joglekar et al. 1989):

x_V	3.94	4.35	4.67	4.91	5.47	5.79	6.36	6.60	7.40	8.05
y_V	1.23	5.00	6.53	5.58	10.57	11.37	11.44	11.94	15.62	15.82
x_V	8.77	9.33	9.66	9.98	10.22	11.27	11.91	12.63	13.12	14.16
y_V	15.01	17.37	18.22	18.66	19.30	18.00	20.88	21.79	21.66	21.12
x_V	14.65	15.37	15.61	16.09	16.42					
y_V	23.03	22.94	23.86	22.36	23.10					

Im Streudiagramm der Daten (Abbildung 12.5) wird ein nichtlinearer Zusammenhang zwischen Windgeschwindigkeit und Stromoutput der Windkraftanlage deutlich, der durch eine Regressionsgerade nicht zufriedenstellend beschrieben werden kann. Im Falle eines linearen Zusammenhangs würde der Stromertrag proportional zur Windgeschwindigkeit variieren. Tatsächlich sinkt bei steigender Windgeschwindigkeit der Zusatzertrag (Grenzertrag) der Stromproduktion.

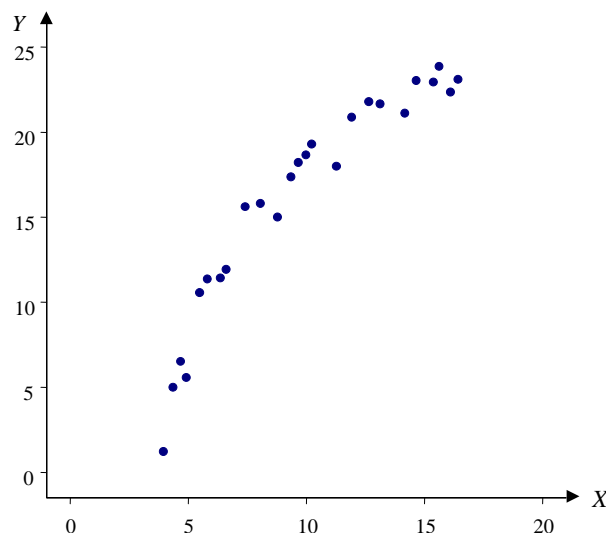


Abb. 12.5: Streudiagramm

Nichtlineare Zusammenhänge lassen sich vielfach durch *Datentransformationen* in lineare überführen. Die *Linearisierung* erlaubt es uns, Regressionskurven an ein Streudiagramm anzupassen, ohne dass wir hierzu den uns bereits bekannten methodischen Rahmen der linearen Regression verlassen müssen.

Ein besonders einfacher Ansatz geht von Regressionsgleichungen der Form

$$y_v = \hat{y}_v + u_v = a + b \cdot T(x_v) + u_v \quad (v=1, \dots, n)$$

aus. Die Regressorvariable X wird durch eine geeignet gewählte Funktion $T: \mathbb{R} \rightarrow \mathbb{R}$ transformiert. Die Transformation soll eine Maßstabsänderung des Regressors so herbeiführen, dass der Zusammenhang zwischen den Werten $x_v^* = T(x_v)$ der transformierten Variable $X^* = T(X)$ und den Werten y_v des Regressanden Y linear ist.

Als Transformationsfunktionen bieten sich *Potenzfunktionen* der Form

$$T_p(x) = \begin{cases} (x+c)^p & (p>0) \\ \ln(x+c) & (p=0) \\ -(x+c)^p & (p<0) \end{cases}$$

an. Die *Konstante* c dient der Lageverschiebung, so dass $x_v + c > 0$ für alle $v=1, \dots, n$ erfüllt ist. Die Lageverschiebung wird also nur dann notwendig, wenn der Datensatz nicht-positive Messwerte des Regressors enthält. Aus statistischer Sicht sind nur streng monotone Transformationen brauchbar; d.h. es muss gelten: $x_i < x_j \Rightarrow T_p(x_i) < T_p(x_j)$. Diese Bedingung erfüllt beispielsweise eine quadratische Transformation ($p=2$) nur, falls alle Daten positiv sind. Darüber hinaus sind Potenzen negativer Zahlen nicht für alle Exponenten p definiert. Zum Beispiel existiert die Quadratwurzel ($p=0.5$) negativer Werte in der Menge der reellen Zahlen nicht. Der *natürliche Logarithmus* wird als Grenzfall einer Potenzfunktion für $p \rightarrow 0$ betrachtet. Bei *Potenzen mit negativen Exponenten* p ist die Multiplikation mit -1 notwendig, weil ansonsten die Ordnung der Daten umgekehrt würde. Für zwei positive Beobachtungen gilt ja z.B. $x_i < x_j \Rightarrow x_i^{-1} > x_j^{-1}$.

Bei der *Auswahl des Exponenten* p beschränkt man sich in der Regel auf die Werte $\dots, -2, -1.5, -1, -0.5, 0.0, +0.5, +1.5, +2, \dots$. Die Wahl hängt vom Muster des Streudiagramms ab und kann durch Probieren erfolgen. Eine Orientierungshilfe bietet das *Auswahldiagramm* von Mosteller & Tukey (1977). In dem Diagramm sind vier typische Muster von Streudiagrammen durch Kurven symbolisiert. Zeigt das Streudiagramm ein Muster, das mit einem der linken Muster im Diagramm vergleichbar ist, sind versuchsweise Exponenten $p < 1$ zu wählen. Man nutzt solange hintereinander die Transformationen $x^{0.5}, \ln x, x^{-0.5}, \dots$, bis eine Linearisierung erreicht ist. Zeigt das Streudiagramm ein Muster, das mit einem der rechten Muster im Diagramm vergleichbar ist, führt man solange hintereinander die Transformationen $x^{1.5}, x^2, x^{2.5}, \dots$ mit $p > 1$ durch, bis eine Linearisierung erreicht ist. Dabei ist vereinfachend unterstellt, dass nur positive Messwerte vorliegen.

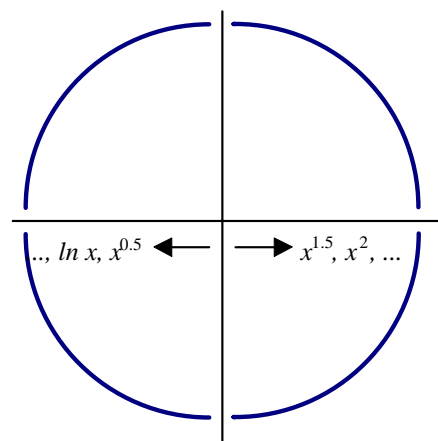


Abb. 12.6: Auswahldiagramm für Potenztransformationen

Beispiel

B-12.3 Windkraftanlage.

Das Muster des Streudiagramms in Abbildung 12.5 ist vergleichbar mit dem oberen linken Typ im Auswahldiagramm. Wir führen für die Messwerte des Regressors nacheinander die Potenztransformationen

(a) $T_{0.5}(x_v) = x_v^{0.5}$, (b) $T_0(x_v) = \ln x_v$, (c) $T_{-0.5}(x_v) = -x_v^{-0.5}$, (d) $T_{-1}(x_v) = -x_v^{-1}$
 durch Transformation (d) bewirkt schließlich eine Linearisierung des Streudiagramms (siehe Abb. 12.7).

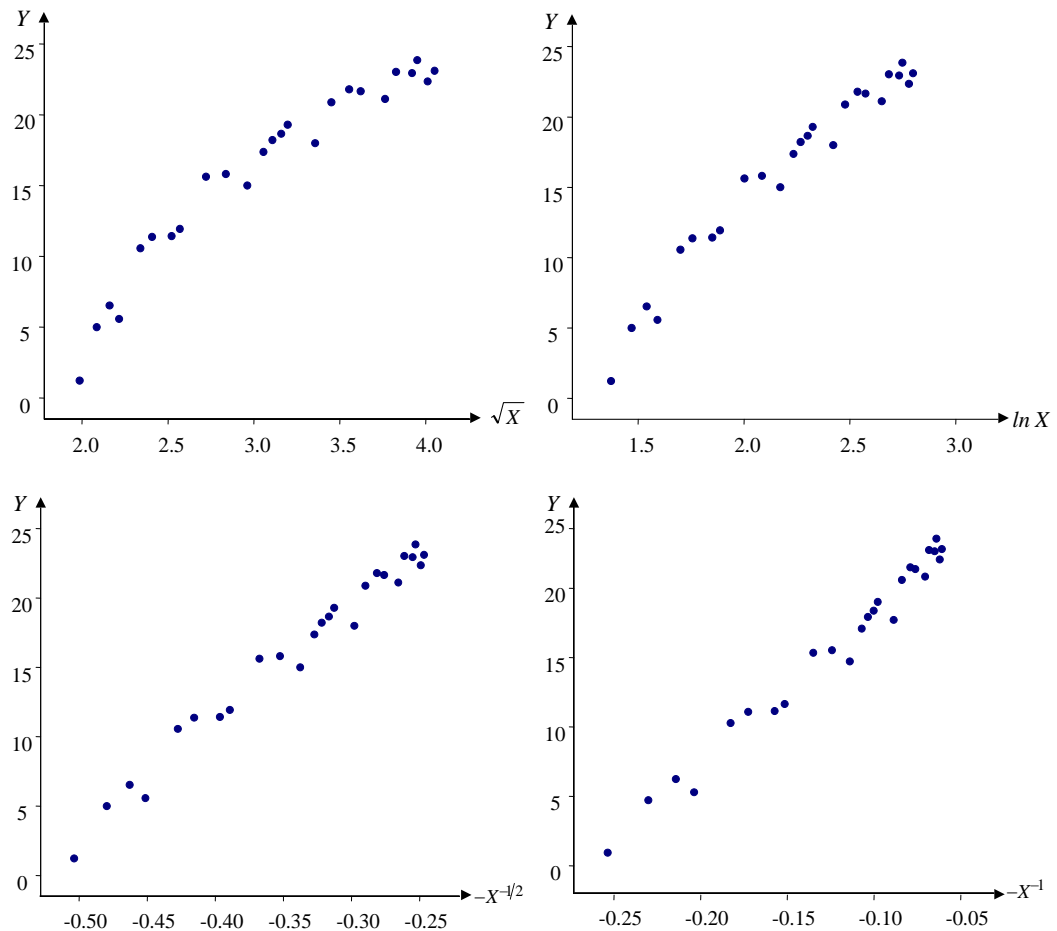


Abb. 12.7: Linearisierung des Streudiagramms durch Potenztransformationen

Im Folgenden sei $x_v^* = -x_v^{-1}$ ($v = 1, \dots, n$) bzw. $X^* = -X^{-1}$. Mit

$$\sum_v x_v^* = -3.0673, \quad \sum_v y_v = 402.4000, \quad \sum_v x_v^{*2} = 0.4567, \quad \sum_v y_v^2 = 7498.1492,$$

$$\sum_v x_v^* y_v = -40.4046$$

und

$$\overline{x^*} = -0.1227, \quad \tilde{s}_{X^*}^2 = 0.0032, \quad \bar{y} = 16.0960, \quad \tilde{s}_Y^2 = 40.8448,$$

$$\tilde{s}_{X^*Y} = 0.3587, \quad r_{X^*Y} = 0.9900$$

folgt

$$\hat{b} = \frac{\tilde{s}_{X^*Y}}{\tilde{s}_{X^*}^2} = 111.6007 \quad \text{und} \quad \hat{a} = \bar{y} - \hat{b} \cdot \overline{x^*} = 29.7886.$$

Wir erhalten die Regressionsfunktion

$$\hat{y} = 29.7886 + 111.6007 x^* \quad \text{bzw.} \quad \hat{y} = 29.7886 - 111.6007 \frac{1}{x}$$

mit einem Bestimmtheitsmaß von $R^2 = r_{X^*Y}^2 = 0.9801$.

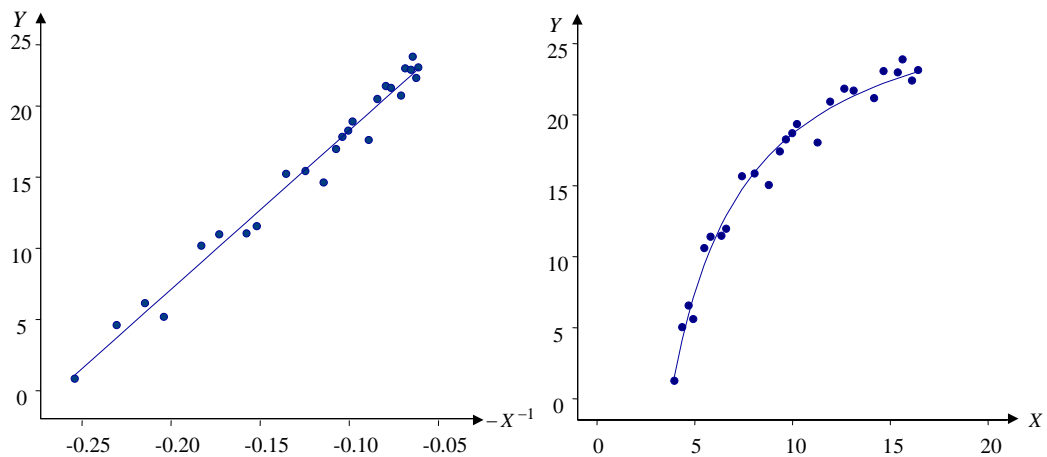


Abb. 12.8: Streudiagramm und Regressionsfunktion (mit und ohne Achsen-Transformation)

Die Regressionsgerade der Regression von Y bezüglich X^* entspricht einer Regressionskurve der Regression von Y bezüglich X . Die Regressionsfunktion

$$\text{Stromoutput} = 29.7886 - 111.6007 \frac{1}{\text{Windgeschwindigkeit}}$$

ist erst für Windgeschwindigkeiten ab ca. 3.75 km/h sinnvoll interpretierbar, da ein negativer Stromoutput nicht möglich ist. Mit wachsender Windgeschwindigkeit steigt der prognostizierte Stromoutput \hat{y} , wobei der Ertragszuwachs (Grenzertrag) mit jedem zusätzlichen Stundenkilometer monoton sinkt. Für $x \rightarrow \infty$ strebt \hat{y} gegen $\hat{a} = 29.7886$ Kilowatt. Das Absolutglied \hat{a} der Regressionsfunktion kann somit als ein empirischer Näherungswert für die Obergrenze des Stromoutputs der Windkraftanlage interpretiert werden. Hierbei ist allerdings zu berücksichtigen, dass Windkraftanlagen bei extrem hohen Windgeschwindigkeiten aus Sicherheitsgründen abgeschaltet werden müssen. Die Obergrenze wird also voraussichtlich nie erreicht.

Die Linearisierung nichtlinearer Zusammenhänge ist aufgrund ihrer Einfachheit attraktiv. Aus der Einfachheit resultieren aber natürlich auch Grenzen. Ist in einem Streudiagramm ersichtlich, dass der nichtlineare Zusammenhang Wendepunkte und/oder lokale Extrema aufweist, dann ist eine Linearisierung nicht möglich. In diesem Fall ist man gezwungen, eine hinreichend flexible nichtlineare Funktion $f(x)$ als Regressionsfunktion auszuwählen und an die Originaldaten anzupassen. Häufig verwendete Funktionen sind beispielsweise

$$f(x) = a + bx + cx^2 \quad (\text{quadratische Funktion}),$$

$$f(x) = a + bx + cx^2 + dx^3 \quad (\text{kubische Funktion}),$$

$$f(x) = a + be^{cx} \quad (\text{modifizierte Exponentialfunktion}),$$

$$f(x) = e^{a+be^{cx}} \quad \text{mit } b, c < 0 \quad (\text{Gompertz-Funktion}),$$

$$f(x) = \frac{c}{1 + e^{a+bx}} \quad \text{mit } c > 0, b < 0 \quad (\text{logistische Funktion}) \text{ u.s.w.}$$

Die Bestimmung der Funktionsparameter a, b, c, d, \dots kann jeweils mit Hilfe der Methode der Kleinsten Quadrate durch Minimierung der Residuenquadratsumme erfolgen. Weitreichende Möglichkeiten ergeben sich auch durch die *Glättung der Daten* mittels Spline-Funktionen, sogenannten Kernschätzern oder dergleichen. Man spricht in diesem Zusammenhang von *nichtparametrischer Regression* (siehe z.B. Fahrmeir et al. 2007)

B-12.3 Windkraftanlage.

Wir unterstellen nun die Gültigkeit der Modellannahmen des klassischen linearen Modells für die Regression von Y bezüglich X^* . Für die KQ-Schätzer \hat{A} und \hat{B} erhalten wir die geschätzten Varianzen

$$s_{\hat{A}}^2 = \frac{\sum_{v=1}^n x_v^{*2}}{n \cdot \sum_{v=1}^n (x_v^* - \bar{x}^*)^2} \cdot s_U^2 = \frac{0.4567}{25 \cdot 0.0803} \cdot 0.8868 = 0.2016 ,$$

$$s_{\hat{B}}^2 = \frac{1}{\sum_{v=1}^n (x_v^* - \bar{x}^*)^2} \cdot s_U^2 = \frac{1}{0.0803} \cdot 0.8868 = 11.0370 ,$$

wobei

$$s_U^2 = \frac{RSS}{n-2} = \frac{20.397}{23} = 0.8868 \quad \text{mit} \quad RSS = \sum_{v=1}^n \hat{u}_v^2 = n \tilde{s}_Y^2 (1 - r_{X^*Y}^2) = 20.3970$$

der Schätzwert der Störvariablenvarianz σ^2 ist. Mit den Standardfehlern $s_{\hat{A}} = 0.4490$, $s_{\hat{B}} = 3.3222$ und dem Quantil $t_{[1-\alpha/2; n-2]} = t_{[0.975; 23]} = 2.0687$ folgen zum Konfidenzniveau $1 - \alpha = 0.95$ die Schätzintervalle

$$[\hat{a} - 2.0687 \cdot s_{\hat{A}}, \hat{a} + 2.0687 \cdot s_{\hat{A}}] = [28.8600, 30.7174] ,$$

$$[\hat{b} - 2.0687 \cdot s_{\hat{B}}, \hat{b} + 2.0687 \cdot s_{\hat{B}}] = [104.7281, 118.4733]$$

für die Modellparameter a, b . Keiner der beiden realisierten Konfidenzintervalle schließt die Null ein, so dass die Schätzwerte signifikant von Null verschieden sind ($\alpha = 0.05$). Zusammenfassend notieren wir:

$$\hat{y} = \underset{(0.449)}{29.7886} - \underset{(3.3222)}{111.6007} \cdot \frac{1}{x} , \quad R^2 = 0.9801 .$$

Die Abbildung 12.11 zeigt das Streudiagramm der Daten, die der Analyse zugrundeliegen, die geschätzte Regressionsfunktion sowie Prognoseintervalle zum 95%-Konfidenzniveau

$$[\hat{y}_0 - 2.0687 \cdot s_{\hat{y}_0 - y_0}, \hat{y}_0 + 2.0687 \cdot s_{\hat{y}_0 - y_0}]$$

mit

$$s_{\hat{y}_0 - y_0}^2 = s_U^2 \cdot \left(1 + \frac{1}{n} + \frac{(x_0^* - \bar{x}^*)^2}{\sum_{v=1}^n (x_v^* - \bar{x}^*)^2} \right) = 0.8868 \cdot \left(1.04 + \frac{(x_0^* + 0.1227)^2}{0.0803} \right)$$

für Werte x_0 des Regressors X im Intervall $[3.94, 16.42]$ und $x_0^* = -x_0^{-1}$.

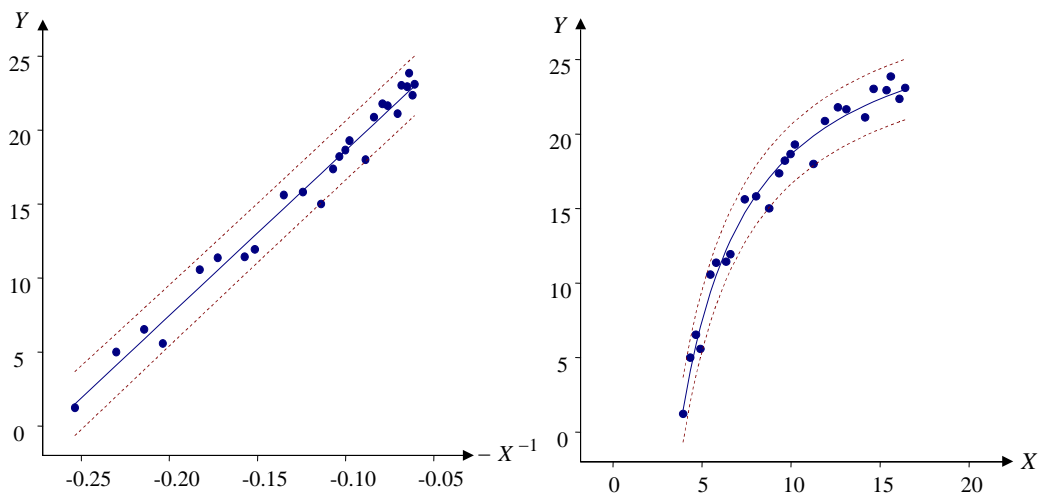


Abb.12.11: Streudiagramm der Daten sowie Punkt- und Intervallprognosen des Gleichstromoutputs (mit und ohne Achsen-Transformation)