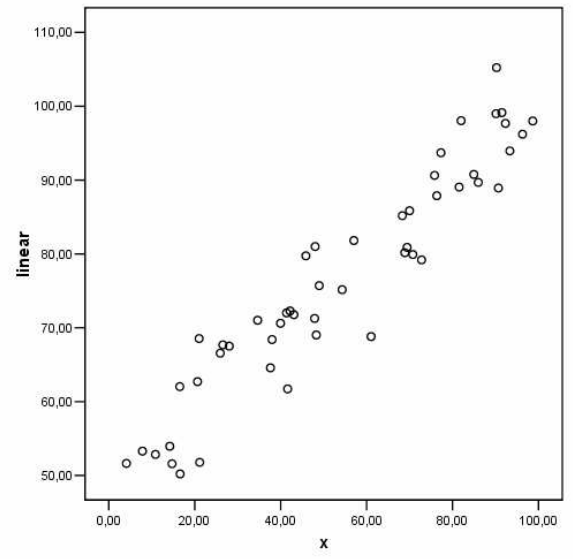
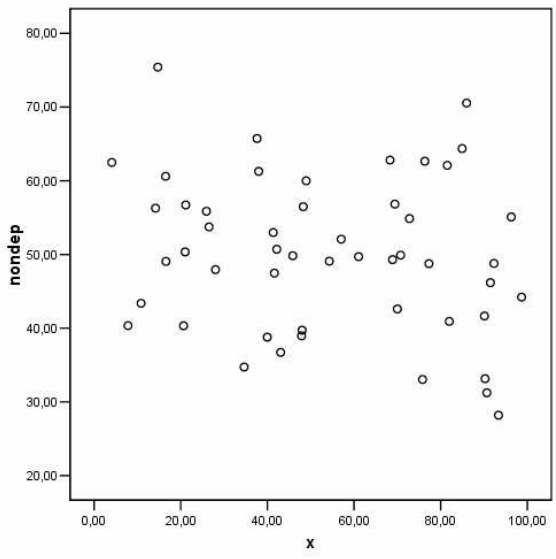
Regressionsanalysen

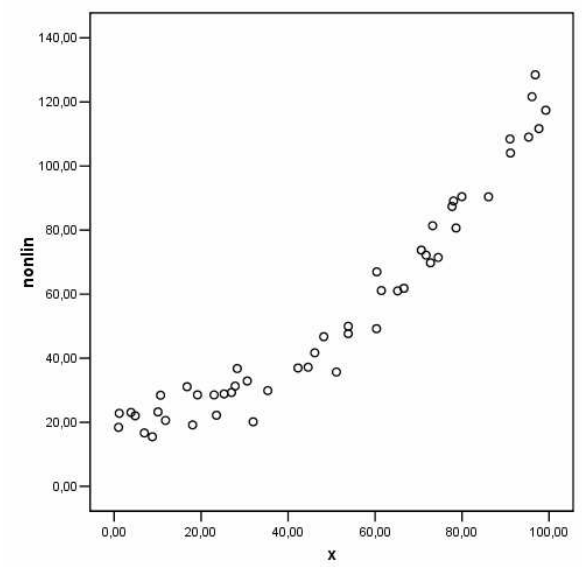
Zusammenhänge von Variablen



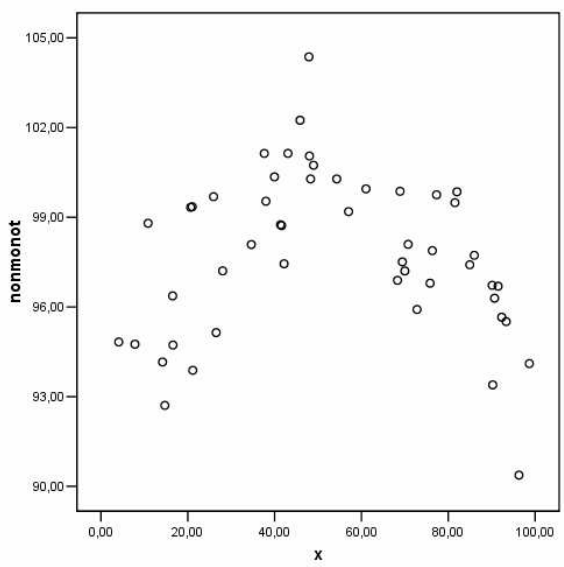
linearer Zusammenhang
(*„Idealfall“*)



kein Zusammenhang



nichtlinearer monotoner Zusammenhang (i.d.Regel berechenbar über Variablen-transformationen mittels linearer Regr.)



nichtlinearer nichtmonotoner Zusammenhang (erfordert spezielle Programme zur nichtlinearen Regression)

Ziel der Regression

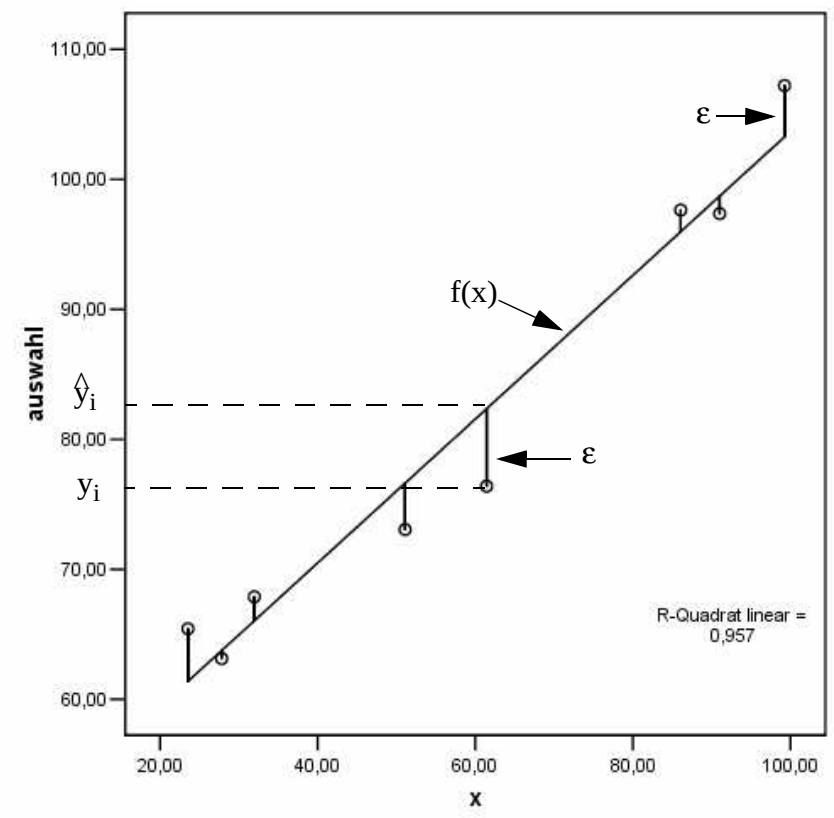
Beschreibung und Beziehung zwischen einer abhängigen Variablen y und einer oder mehreren unabhängigen Variablen x_1, x_2,... (Prädiktoren) durch eine Funktion

$$y = f(x_1, x_2, \ldots; b_0, b_1, b_2, \ldots) + \varepsilon$$

(b_0, b_1, b_2,... sind die Parameter der Regressionsfunktion und heißen Regressionskoeffizeinten, ε sind die Abweichungen zwischen beobachteten y-Wert und Funktionswert $y = f(\ldots)$

$$\varepsilon = y - f(x_1, x_2, \ldots; b_0, b_1, b_2, \ldots)$$

und heißen Residuen



Primäre Aufgabe der Regression:

Bestimmung der Parameter b_0, b_1, b_2,.. wobei ein Funktionsmodell, d.h. eine Klasse von Funktionen, die von b_0, b_1, b_2, ... abhängen, vorgegeben werden, z.B.

$f(x; b_0, b_1) = b_0 + b_1 x$ lineare Funktion

$f(x; b_0, b_1, b_2) = b_0 + b_1 x + b_2 x^2$ Parabel (Polynom 2. Grades)

$f(x, a, b) = a e^{-bx}$ Exponentialfunktion

Beispiel

Untersuchung der Beziehung zwischen Gewicht und Größe bei normalgewichtigen Personen zur Ermittling des „Normalgewichts“:

Gewicht = b_0 + b_1*Größe

durch ein lineares Funktionsmodell ergibt (beispielsweise) b_0=-85.2 und b_1=0.91, d.h. pro cm Körpergröße nimmt das Gewicht um 0.91 kg zu.

Berechnungsmethoden:

üblicherweise: kleinste Quadrate-Schätzung (LS, least squares)

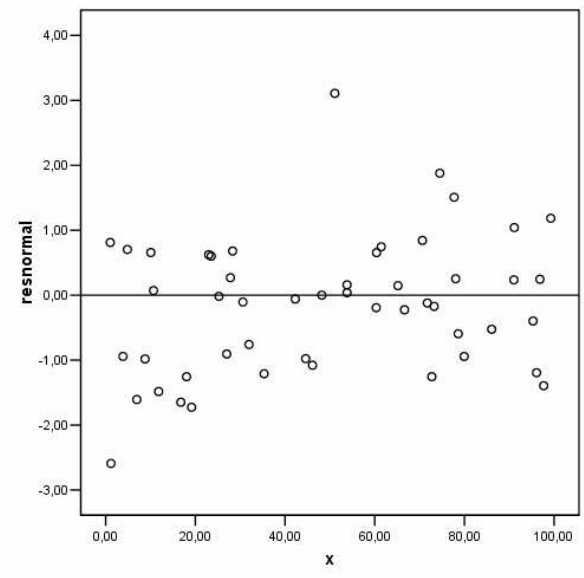
Minimierung von $\sum \varepsilon_i^2$

Voraussetzungen:

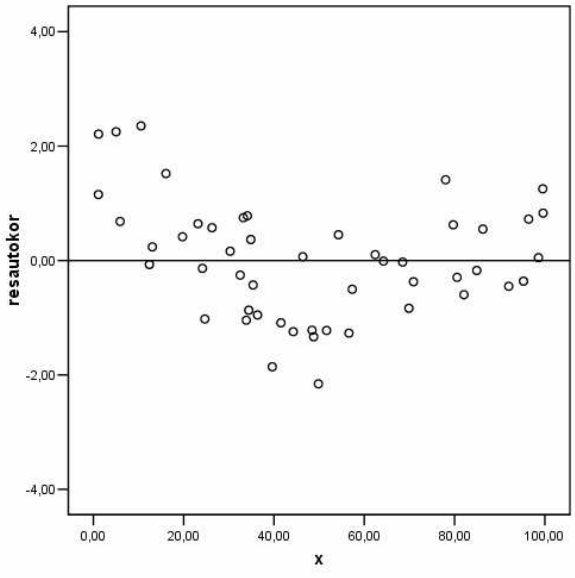
y metrisch

x metrisch oder dichotom (ordinal möglich, aber schwierig zu interpretieren)

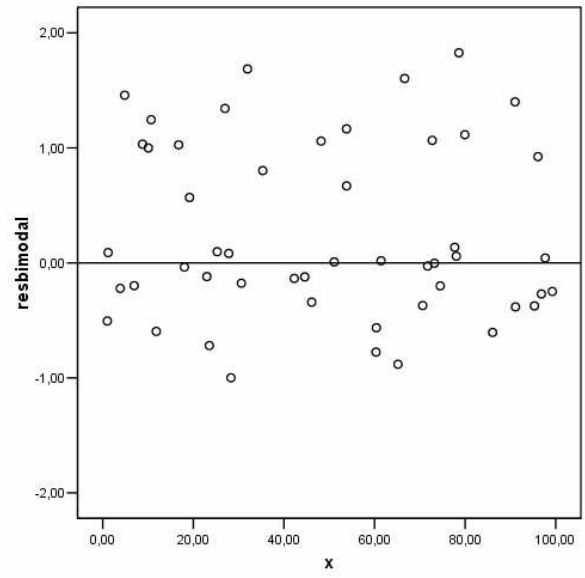
ε_i unabhängig, normalverteilt, homoskedastisch (konstante Varianz)



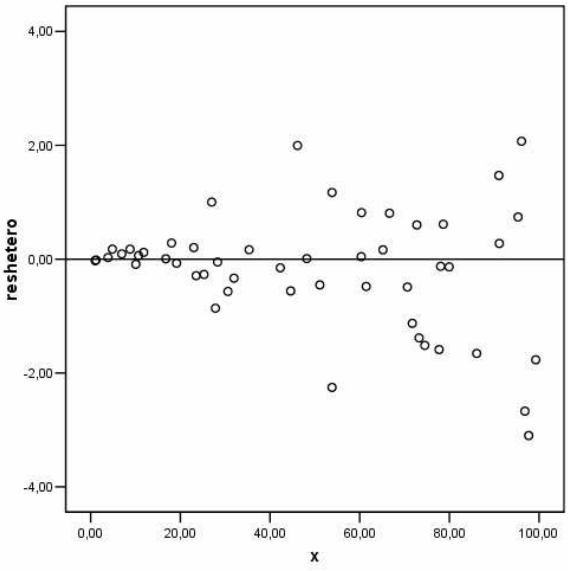
ε_i unabhängig, normalverteilt und homoskedastisch



ε_i autokorreliert (nicht unabhängig)



ε_i bimodal verteil (nicht normalverteilt)



ε_i heteroskedastisch (nicht konstante Varianz)

Die Voraussetzungen sind prüfbar z.B. über die Ausgabe des o.a. Streudiagramms (x-y-Punktewolke) oder auch über Histogramme der Residuen:

- im Falle nur eines Prädiktors x:
  y: standardisierte Residuen (ZRESID)
  x: Prädiktor x

- im Falle mehrerer Prädiktoren x_1, x_2,..

y: standardisierte Residuen (ZRESID)
x: standardisierte vorhergesagte Werte (ZPRED)

Die Kleinste-Quadrate-Schätzmethode ist zwar mathematisch die einfachste Methode, hat aber den Nachteil, sehr empfindlich auf Extremwerte, insbes. Ausreißer, zu reagieren. Deswegen wurden sog. „robuste Schätzmethoden“ entwickelt (u.a. L_1 -, Huber-, Hampel-Methode), die unempfindlicher reagieren

(z.B. Minimierung von $\sum |\varepsilon_i|$).

Eine neuere robuste Methode ist die Least Median Squares-Methode, bei der nicht die o.a. Residuenquadratsumme sondern der Median der Residuenquadrate minimiert wird,

also Minimierung von $median(\varepsilon_i^{\,2})$.

Wichtige Begriffe und Tests:

Regressionskoeffizienten

$b_0, b_1, b_2, ...$

geben an, in welchem Maße sich y verändert, wenn sich jeweils eine der unabhängigen Variablen $x_1, x_{2,}$ ... um eine Einheit vergrößert, dabei alle übrigen unverändert bleiben.

$\beta_1, \beta_2, ...$

sind die standardisierten Regressionskoeffizienten $\beta_i = b_i \frac{s_x}{s_y}$

Diese erlauben einen Vergleich des Einflusses verschiedener Variablen sowie verschiedener Regressionsmodelle.

t-test eines Regressionskoeffizienten

$H_0: b_i = 0 \quad H_1: b_i =/= 0$

prüft, ob dieser Parameter/Prädiktor (wichtiger) Bestandteil des Funtionsmodells ist, sollte i.a. signifikant sein. (sinnvoll bei multiplen Regressionsanalysen).

Varianzanalyse:

$$F = \frac{\sum (\hat{y}_i - \bar{y})^2}{\frac{1}{n-2}\sum (y_i - \hat{y}_i)^2} = \frac{Regressionsstreuung}{Residuenstreuung}$$

sollte immer signifikant sein, sonst unpassendes Modell gewählt oder kein Zusammenhang vorhanden
(bei einfacher linearer Regression ist dieser Test mit dem o.a. t-Test identisch)

Mutiples R und R^2:

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - y_i)^2} = 1 - \frac{Residuenstreuung}{Gesamtstreuung}$$

Anteil der durch die Regression erklärten Streuung von y
(wichtigstes Maß für die Güte der Regression)

Signifikanzüberprüfung erfolgt durch o.a. Varianzanalyse.

Bei einfacher linearer Regression ist R=r, (r=Produkt Moment Korrelation). R ist die Korrelation der y_i mit den y_i. Somit ist R ein allgemeinerer Korrelationskoeffizient als r, insbesondere auch für nicht-lineare Zusammenhänge.

Adjusted R und R^2:

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

wobei p die Anzahl der Variablen in der Regression und n die Anzahl der Fälle ist.

Während das R^2 mit zunehmender Prädiktorzahl immer ansteigt, werden beim R^2_{adj} die Anzahl der Prädiktoren berücksichtigt, so dass Variablen, die keinen Erklärungsbeitrag leisten (d.h. die Residuenquadratsumme nicht reduzieren) zu einer Verkleinerung des R^2_{adj} führen.

Einfache Regression

d.h. nur ein Prädiktor x, elementar berechenbar:

$$b_1 = \frac{s_{xy}}{s_x^2} \qquad b_0 = \bar{y} - b_1\bar{x}$$

(mit s_{xy} Kovarianz)

Nicht-lineare Regression

je nach Funktionsmodell schwierig lösbar.

- Polynomregression leicht berechenbar, Überprüfung des Polynomgrads möglich, berechenbar auch mit Hilfe multipler linearer Regression, durch Variablentransformationen x -> x^2, x -> x^3 u.s.w.
- einfache hyperbolische und exponentielle Regressionsmodelle lassen sich durch geeignete Transformationen „linearisieren“,

Beispiel:

$y = a\ b^x$

ergibt durch Logarithmieren eine lineare Regression

$\log y = \log a + x*\log b$

bzw. unter Verwendung neuer Symbole

$y' = a' + x*b'$ (mit $y'=\log y$, $a'=\log a$, $b'=\log b$)

die berechnet und daraus a und b ermittelt werden können.

Vergleich von Regressionsmodellen

Der Vergleich von 2 Regressionsmodellen ist statistisch (mit Signifikanztest) nur dann möglich, wenn eine Funktion die andere als Spezialfall enthält, z.B. lineare R. und Polynom-R. mit Polynom 2. Grades, das die lineare R. als Spezialfall (b_2=0) enthält. Das „größere“ Funktionsmodell bringt (natürlich) eine bessere Anpassung, d.h. eine kleinere Residuenquadratsumme. Die Differenz kann statistisch überprüft werden und besagt, ob der „zusätzliche“ Koeffizient statistisch notwendig ist.

Multiple Regression

d.h. mehrere x-Variablen

$$y = f(x_1, x_2, \ldots; b_0, b_1, b_2, \ldots) + \varepsilon$$

z.B.

$y = b_0 + b_1 x_i + b_2 x_2 ... + b_k x_k$ multiple lineare Regression)

$y = ax^2_1 + bx_1 x_2 + cx_2^2 + dx_2^3 + e$ spez. Polynom 3. Grades)

Körperoberfläche = p_1 *Gewicht^{p2}*Größe^{p3}

wichtig (bei „größeren" Modellen):

Anzahl der Fälle mindestens	>= Anzahl Parameter + 2
sinnvoll	>= 2 * Anzahl Parameter

Multiple lineare Regression:

(relativ) leicht berechenbar, wird daher auch gewählt, wenn der Zusammenhang zwischen einer x-Variable und y nicht-linear ist.

z.B.

$RR_{diff} = b_0 + b_1$Alter + b_2*Geschlecht + $b_3 RR_{Ruhe}$ + b4*Gewicht

Die Regressionskoeffizienten b_0, b_1, ... sind davon abhängig, welche anderen Variablen ebenfalls im Modell enthalten sind.

Schrittweise (multiple lineare) Regression

Prinzip: Das Regressionsmodell wird nacheinander immer um eine unabhängige Variable mehr erweitert, und zwar i.d. Regel die Variable, die das R^2 am meisten vergrößert und damit die Vorhersage am meisten verbessert:

0. Schritt:	Modell nur mit absolutem Glied b_0
1. Schritt	Modell mit nur 1 x-Variable größter R^2-Anstieg hat die Variable, die größtes r mit y hat (da in diesem Fall r =R ist)
2. Schritt	Modell mit 2 (bzw. mehreren) x-Variablen
und folgende:	1. x-Variable bleibt vom 1. Schritt, für alle Variablen (die noch nicht im Modell sind) wird der R^2-Anstieg ausgerechnet und in einen F-Wert umgerechnet), die Variable mit dem größten F-Wert wird in das Modell genommen, vorausgesetzt der F-Wert ist hinreichend groß, (i.d. Regel F 4.0, was etwa einem bei 5% signifikanten F-Wert entspricht).

Wenn kein F-Wert groß genug ist, hört die Prozedur auf. Vielfach wird das $R^2_{adjusted}$ verwendet, das die Anzahl der Regressionsparameter berücksichtigt, indem für jede tatsächliche Variable ein „Malus" vom R^2 abgezogen wird, so dass unbedeutende Variablen keinen R^2-Anstieg verursachen können, wenn schon mehrere Variablen im Modell enthalten sind.

Warnung für die Interpretation bei der schrittweisen Regression

- Wenn $r_{yx_1} \approx r_{yx_2}$, dann können dennoch nicht beide Variablen gleichzeitig in das Modell aufgenommen werden und die Auswahl ist dann zufällig.
- Wenn $r_{x_1 x_2} \approx 1$, kann nach Aufnahme von x_1 in das Modell x_2 kaum noch einen R^2-Anstieg bringen und bleibt daher aus dem Modell, auch wenn sie hoch mit y korreliert.
- Wenn $r_{yx_2} \approx 0$, kann dennoch x_2 in das Modell aufgenommen werden („Suppressor"-Variable x_2).

Multicollinearität

Eine Multicollinearität der unabhängigen Variablen liegt vor, wenn sich eine der x-Variablen aus den anderen x-Variablen (annähernd genau) linear berechnen lässt:

$$x_1 \sim c_0 + c_2 x_2 + c_3 x_3 + \ldots$$

Inhaltlich bringt die Hinzunahme einer solchen Variable x keinen R^2-Anstieg und keinen Informationsgewinn, wenn die anderen Variablen x_2, x_3 ... im Modell enthalten sind.

Mathematisch ist eine solche x-Variable kaum bearbeitbar und wird aus der Analyse ausgeschlossen.

Zur Überprüfung auf Multicollinearität wird für alle Variablen die Toleranz berechnet:

$$Toleranz = 1 - R^2_{x_1 x_2 x_3 \ldots}$$

R^2 ist die multiple Korrelation von x_1 mit den anderen x_2, x_3, ...

Prüfung der Multicollinearität in SPSS: Erstellung eines multiplen linearen Regressionsmodells mit den Variablen x_1, x_2, x_3,... als Prädiktoren und einer beliebigen anderen Variablen (ohne fehlenden Werten) als y-Variable. SPSS errechnet dann für alle x_1, x_2, x_3,...die Toleranz.

Sinnvolles Vorgehen:

- Bei großen Variablenzahl und hoch korrelierender Variablen bildet man Variablengruppen, die dann als ganzes bei der schrittweisen Analyse in die Regression genommen werden.

- Zur Überprüfung der Qualität der Ergebnisse teilt man die Fälle in 2 Gruppen und führt 2 Analysen zur gegenseitigen Überprüfung durch (sog. Kreuzvalidierung).

Nominale Prädiktoren

Eine poylchtotome Variable (nominal skalierte Variable mit mehr als 2 Ausprägungen) muss dichotomisiert werden. Wenn diese m Ausprägungen hat, werden daraus m-1 dichotome Variablen erzeugt, die zusammen als Variablengruppe in die Regression eingehen. Für die Dichtomisierung gibt es u.a. die folgenden beiden Methoden:

- Dummy Variable Coding:
  m-1 der m Ausprägungen werden ausgewählt und erzeugen jeweils eine dichtome Variable d_j mit

  d_j = 1 wenn Merkmal die Ausprägung j hat, sonst 0.

  Die Regressionskoeffizienten für jede der d_j (entspricht jeweils einer Ausprägung) geben die Veränderung von y in der Gruppe dieser Ausprägung gegenüber der Gruppe an, die nicht für die Dichotomisierung verwendet wurde.

- Effect-Coding:
  m- 1 der m Ausprägungen werden ausgewählt und erzeugen jeweils eine dichotome Variable d_j mit

  d_j = 1 wenn Merkmal die Ausprägung j hat,
  d_j = -1 wenn Merkmal die nicht-verwendete Ausprägung hat,
  d_j = 0 sonst

  Die Regressionskoeffizienten für jede der d_j geben die Veränderung von y in der Gruppe dieser Ausprägung gegenüber dem Durchschnitt aller Gruppen an. Für die Gruppe der Ausprägung, die nicht für die Dichotomisierung verwendet wurde, ist der Regressionskoeffizient berechenbar als Summe aller m -1 Regressionskoeffizienten, jedoch mit umgekehrtem Vorzeichen.

Prädiktoren mit nicht-linearem Einfluss

Normalerweise wird ein quadratischer oder kubischer Einfluss gewählt. Prinzipiell wird der Prädiktor vor der Re-

gression geeignet transformiert. Wird das Abhängigkeitsverhältnis durch eine mehrgliedrige Funktion beschrieben (z.B. $a_1x + a_2x^2$) , so müssen mehrere x-Variablen erzeugt werden und zusammen als Variablengruppe in die Regression eingehen.

Beispiel:

Wenn zwischen y und x ein quadratischer Zusammenhang besteht, bildet man $x_1=x$ und $x_2=x^2$ und führt eine Regression mit x_1 und x_2 (als Variablengruppe bei schrittweisen Analysen) durch.

Optimal Subset Regression/Selection

Ziel: Optimale Vorhersage (Schätzung von y) mit möglichst wenig Variablen. Die Lösung wird durch Probieren gefunden.

Methode: Für k=1, 2, 3, ... Prädiktoren werden alle k-elementigen Teilmengen aller möglicher Prädiktoren in einer Regression untersucht und diejenigen aus ihnen bestimmt, für die die Zielgröße (R^2, $R^2_{adjusted}$ oder Mallows´s C_p) maximal wird. Die so gefundenen k unabhängigen Variablen brauchen nicht mit denen aus dem k-ten Schritt einer schrittweisen Regression identisch zu sein! $R^2_{adjusted}$ und Mallows´s C_p werden ab einen bestimmten k wieder kleiner, so daß sich daraus auch eine optimale Anzahl von Prädiktoren/Regressoren bestimmen läßt.