Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

CAS Datenanalyse

Kapitel 4: Funktionale Form der Regression

Prof. Dr. Raúl Gimeno
FRM,CAIA, PRM

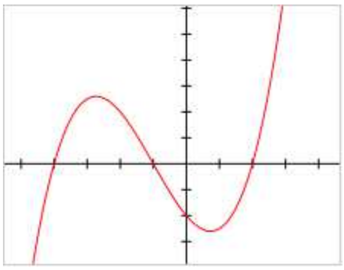
Transformation von Variablen

- Durch eine Transformation der Variablen y und x können viele gekrümmte, nichtlineare Beziehungen dargestellt werden und immer noch das lineare Regressionsmodell verwenden.
- Die Wahl einer algebraischen Form für die Beziehung bedeutet, Transformationen der ursprünglichen Variablen vorzunehmen.
- Die häufigsten Transformationen sind:

Potenzieren: $x^n = \underbrace{x \cdot x \ldots x \cdot x}_{\text{n Faktoren}}$

Quadratische Funktion: f(*x*) = ax^2 + bx + c

Kubische Funktion f(x) = ax^3 + bx^2 + cx + d →



Natürlicher Logarithmus: ln(x) oder lnx

Lösung der Exponentialfunktion y = exp(x) = e^x ⇔ lny = x

Logarithmusfunktion = Umkehrfunktion zur Exponentialfunktion

Die Bedeutung der logarithmischen Transformation folgt aus drei Eigenschaften.

Logarithmierung: Eigenschaften

1. Multiplikative Zusammenhänge können durch Logarithmierung additiv dargestellt werden, bzw. Exponentialfunktionen werden durch Logarithmierung zu linearen Funktionen:

Eigenschaften: 1. $\ln(xy) = \ln(x) + \ln(y)$ für $x,y > 0$

2. $\ln(x^{\alpha}) = \alpha \ln x$ 3. $\ln(x/y) = \ln x - \ln y$

Cobb-Douglas Produktionsfunktion: $Q = AK^{\alpha}L^{\beta}$

mit K: Kapital und L: Arbeit (labour)

Linearisierung: $\ln Q = \ln(A) + \alpha \ln(K) + \beta \ln(L)$

2. Die Differenz zwischen zwei logarithmierten Werten entspricht näherungsweise der relativen Änderung der ursprünglichen Werte

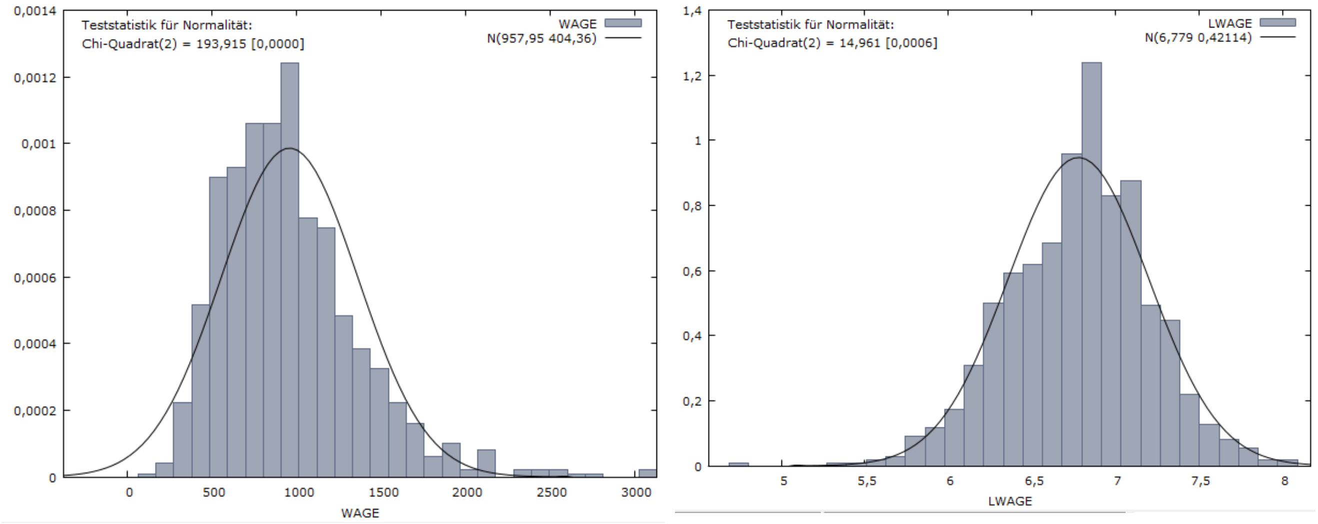
$$\ln x_2 - \ln x_1 \cong \frac{x_2 - x_1}{x_1} = \frac{\Delta x}{x}$$

Prozentuelle Änderung = relative Änderung x 100

x	$\Delta x/x$	$\%\Delta x$	ln(x)	$\ln x_2 - \ln x_1$
5			1.60943	0.995%
5.05	0.01	1%	1.61938	0.00995

Histogramm ohne und mit Logarithmen

3. Durch Logarithmierung werden kleine Werte gespreizt, grosse Werte gestaucht → Einfluss extremer Beobachtungen wird auf die Schätzung reduziert oder schiefe Verteilungen werden symmetrischer.



Verteilung der Monatslöhne ohne (links) und mit Logarithmierung (rechts) → starke Reduktion des Chi-Quadrat-Wertes!

Regressionen mit Logarithmen

Drei Konfigurationen mit logarithmischen Termen:

1. linear-logarithmische (lin-log) Spezifikation: $y_i = \beta_1 + \beta_2 \ln x_i + u_i$

→ die exogene Variable x wird durch Logarithmus transformiert.

→ nur der natürliche Logarithmus wird benutzt!

2. Logarithmisch-lineare (log-lin) Spezifikation: $\ln y_i = \beta_1 + \beta_2 x_i + u_i$

→ nur die endogene Variable y wird durch Logarithmus transformiert.

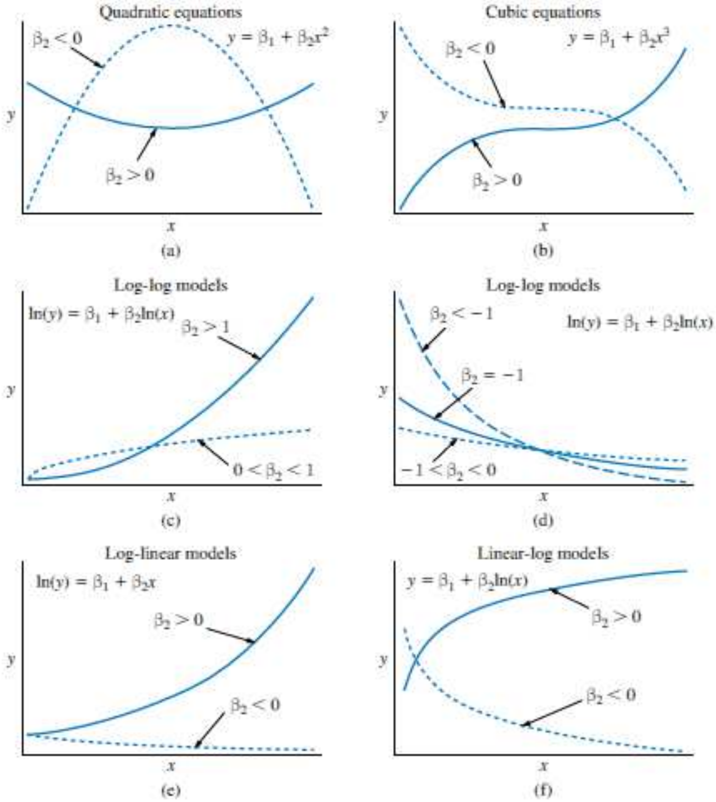
3. Log-Log Spezifikation: $\ln y_i = \beta_1 + \beta_2 \ln x_i + u_i$

→ endogene und exogene Variablen werden durch Logarithmen transformiert.

Verwendung von logarithmierten Variablen

- Geldbeträge und Bevölkerungszahlen werden oft logarithmiert in linearen Regressionsmodellen untersucht. Grund: Grosse Zahlen
- In Jahren gemessene Variablen (z.B. Bildungsjahre, Alter) erscheinen oft in ursprünglicher Form.
- Bei der Interpretation des Effektes von logarithmierten oder nicht-logarithmierten prozentualen Variablen muss zwischen prozentualen Veränderungen und Prozentpunktveränderungen (Differenz zwischen Prozenten) unterschieden werden.
- Logarithmierung kann nicht bei Variablen vorgenommen werden, die negative Werte annehmen.
- Die Prognose der ursprünglichen abhängigen Variable y ist schwieriger, falls diese logarithmiert eingeht.

Unterschiedliche Regressionsmodelle



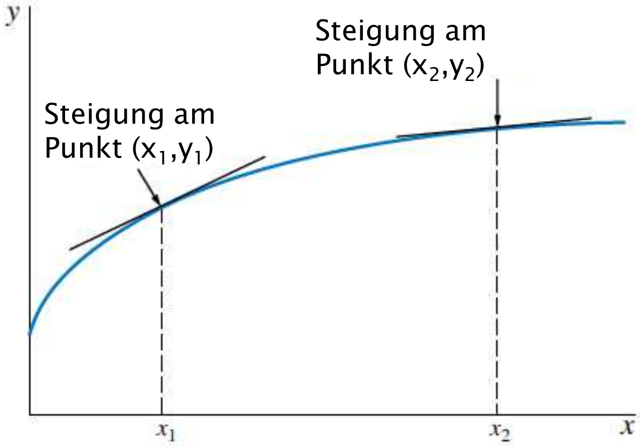
Interpretation der Koeffizienten: lineares Modell

Lineares Regressionsmodell: $y = \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$

Interpretation der Koeffizienten als partielle Ableitungen: $\frac{\partial y}{\partial x_i} = \beta_i$

→ β_i gibt an, um wie viel Einheiten sich y verändert, wenn x_i sich um eine Einheit verändert, ceteris paribus.

Die marginale Veränderung einer erklärenden Variable x wird durch die Steigung der Tangente an die Kurve an einem bestimmten Punkt gemessen.



Elastizität

Elastizität: Gibt die relative Änderung einer abhängigen Variable auf eine relative Änderung einer ihrer unabhängigen Variablen an.

Fragestellung: Um wie viel Prozent verändert sich die abhängige Variable y als Reaktion auf eine einprozentige Änderung der unabhäng.en Variable x?

Relative Änderung = Elastizität von y bezüglich *x* oder *x*-Elastizität von y.

Mathematisch: $\varepsilon_{y,x} = \frac{\Delta y / y}{\Delta x / x}$ → Verhältnis zweier relativer Änderungen

Andere Darstellung: $\varepsilon_{y,x} = \frac{d \ln y}{d \ln x}$

infinitesimale Änderung von lny

1. Ableitung: dlny/dy = 1/y
2. Diskreter Fall: ln(x+Δx)–lnx = Δx/x

$|\varepsilon| = 1$ → y ist proportional elastisch → relative Änderung von y ist gleich der von x

$|\varepsilon| > 1$ → y ist elastisch → y ändert sich relativ stärker als x

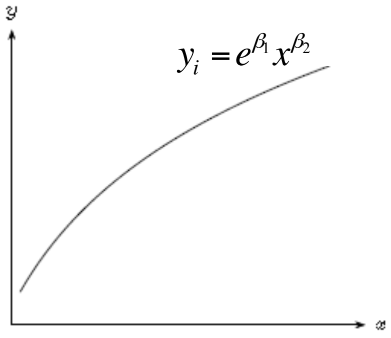
---

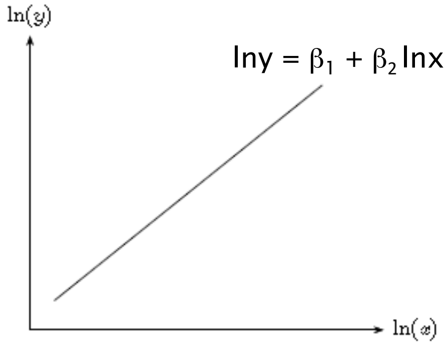
Log-log Modell

Exponentialfunktion: $y_i = b x_i^{\beta_2} \exp(\varepsilon_i)$

Linearisierung

Log-log Modell: $\ln y_i = \beta_1 + \beta_2 \ln x_i + u_i$ mit $\beta_1 = \ln b$ und $u_i = \ln \varepsilon_i$





$\Delta \ln y = \ln y_1 - \ln y_0 = \beta_2(\ln x_1 - \ln x_0)$
$= \beta_2 \, \Delta \ln x$

$(y_1 - y_0)/y_0 = \Delta y/y \approx \ln(y+\Delta y) - \ln y$

$\beta_2 = \Delta \ln y / \Delta \ln x$ → $\beta_2 = \frac{\Delta \ln y}{\Delta \ln x} \approx \frac{\Delta y / y}{\Delta x / x} = \varepsilon_{y,x}$

→ Koeffizient β_2 kann als Elastizität interpretiert werden!

Logarithmisches Modell: Interpretation der Koeffizienten

Log-log Modell: $\ln y_i = \beta_1 + \beta_2 \ln x_i + u$

Ableitung: $\frac{d\ln y}{d\ln x} = \beta_2$ $\quad \frac{d\ln y}{dy} = \frac{1}{y}$ $\quad \frac{d\ln x}{dx} = \frac{1}{x} \longrightarrow d\ln x = \frac{dx}{x}$

$$\beta_2 = \frac{d\ln y/dy}{d\ln x/dx} = \frac{100dy/y}{100dx/x}$$

← Prozentänderung in y

← Prozentänderung in x

β_2 gibt an, um wie viel Prozent sich y verändert, wenn x sich um ein Prozent verändert, ceteris paribus.

Vorteil: Die Koeffizienten können unmittelbar als **Elastizitäten** interpretiert werden!

Beispiel: Cobb-Douglas Produktionsfunktion

$Y = bL^{\alpha}K^{(1-\alpha)} \Rightarrow \ln y = \beta_1 + \beta_2 \ln L + \beta_3 \ln K$

Beispiel für Approximation

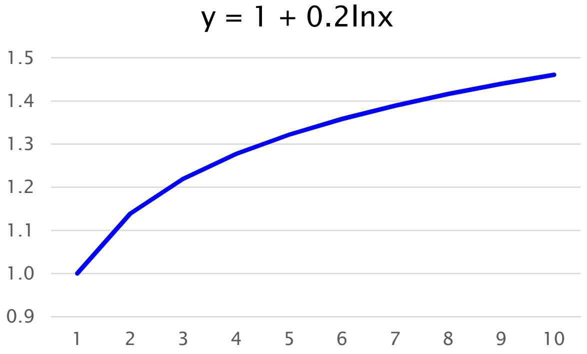
Funktion: $\ln y = 1 + \mathbf{0.2}\ln x \Leftrightarrow y = \exp[1+0.2\ln x]$

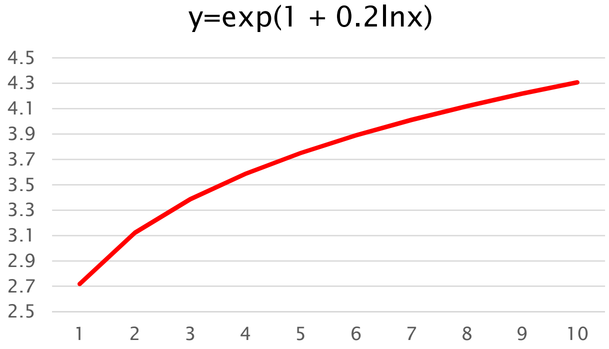
Frage: Wie verändert sich y, wenn x um ein Prozent zunimmt?

x	Δx/x	%Δx	exp(1+0.2lnx)	Δy/y	%Δy
5			3.75049		
5.05	0.01	1%	3.75796	0.00199	≈ 0.2%

→ β x100
0.2x100

Δx = 0.05





Beispiel: Lohngleichung für CEOs

Bestimmung des jährlichen Gehaltes von amerikanischen CEOs (salary) anhand des Unternehmensumsatzes (sales) und der Börsenkapitalisierung (mktval)

Lohngleichung: $\ln(\text{salary}) = \beta_1 + \beta_2 \ln(\text{sales}) + \beta_3 \ln(\text{mktval}) + u$

$$\Leftrightarrow \ln(y) = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

mit $x_2 = \ln(\text{sales})$, $x_3 = \ln(\text{mktval})$

Geschätztes Modell:

$\ln(\text{salary}) = 4.62 + 0.1621 \ln(\text{sales}) + 0.1067 \ln(\text{mktval})$

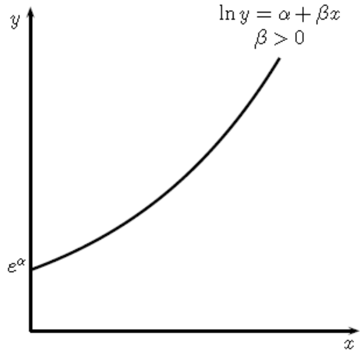
Interpretation β_2: Mit einer Steigerung der Umsätze um 1% ist eine Vergrösserung des Managergehaltes um ca. 0.16% zu erwarten, ceteris paribus.

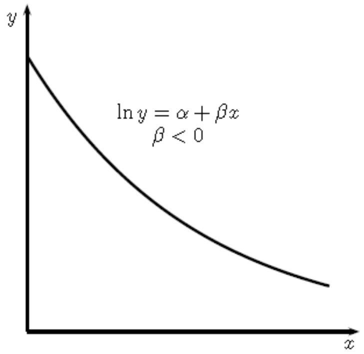
---

log-lin Modell

Logarithmisch-lineares (log-lin) Modell: $\ln y_i = \beta_1 + \beta_2 x_i + u_i$

Nur die abhängige Variable y wird logarithmiert.





Differenzen: $\Delta \ln y = \beta_2 \Delta x$

$$\beta_2 = \frac{\Delta \ln y}{\Delta x} \approx \frac{\Delta y / y}{\Delta x}$$

β_2: **Semielastizität** → gibt näherungsweise die relative Änderung von y an, wenn x um eine Einheit zunimmt.

Wenn *x* um eine Einheit zunimmt, ändert sich y um ungefähr $100\beta_2$ Prozent.

$$100\beta_2 \approx \frac{(\Delta y / y)100}{\Delta x} = \frac{\text{Prozentuale Änderung von y}}{\text{Absolute Änderung von x um 1 Einheit}}$$

Prozentuale Änderung von y durch eine Zunahme von x um eine Einheit:

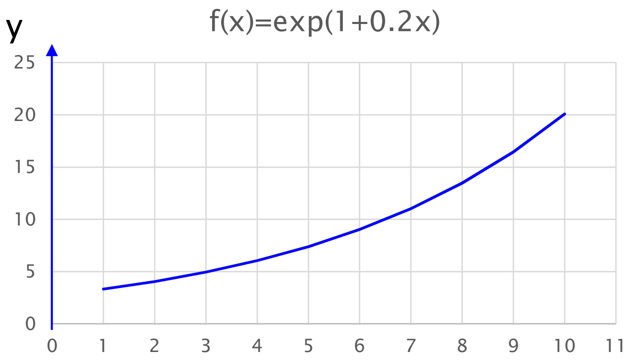
Genauere Approximation: $\%\Delta y \equiv 100(\Delta y/y) = 100\left[e^{\beta_2} - 1\right]$

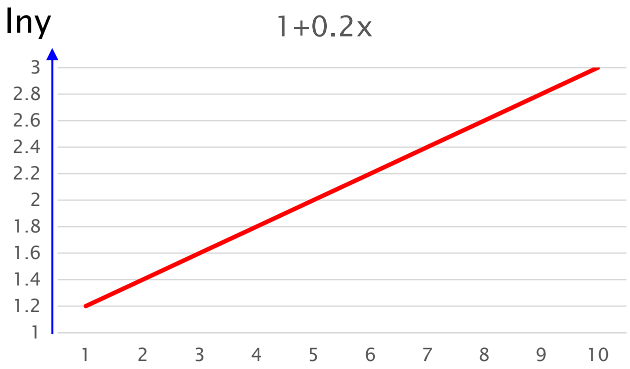
Approximation Beispiel

Funktion: y = exp(1 + 0.2x) ⇔ lny = 1 + 0.2x   β = 0.2

x	Δx/x	%Δx	exp(1+0.2x)	Δy/y	%Δy	100[exp(β -1)]
5			7.38905			
6	0.2	20%	7.4633	0.2214	22.14%	22.14%

Δx = 1





Diskrete Änderungen bei log-lin Modellen

Log-lin Modell: $\ln y_t = \beta_1 + \beta_2 x_t + u_t$

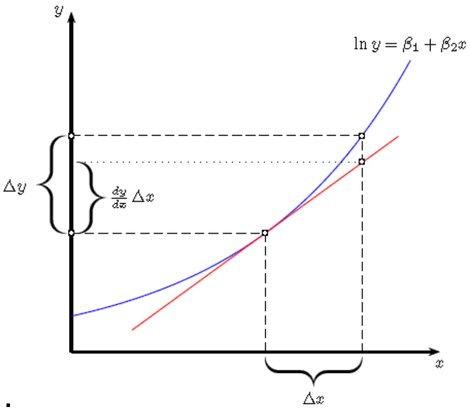
$$\Delta \ln y = \ln(y + \Delta y) - \ln y = \beta_1 + \beta_2(x + \Delta x) - (\beta_1 + \beta_2 x)$$

$$\ln\left(\frac{y+\Delta y}{y}\right) = \beta_2 \Delta x \quad \Leftrightarrow \quad \left(\frac{y+\Delta y}{y}\right) = \exp(\beta_2 \Delta x)$$

$$\Leftrightarrow \left(\frac{\Delta y}{y}\right) = \exp(\beta_2 \Delta x) - 1$$

$$\Leftrightarrow 100\left(\frac{\Delta y}{y}\right) = 100\left[\exp(\beta_2 \Delta x) - 1\right]$$

$\Delta x = 1 \rightarrow \%\Delta y = 100(e^{\beta_2} - 1)$



- Wenn β_2 klein ist → wenig gekrümmte Funktion
  → keiner grosser Unterschied zwischen $100\beta_2$ und $100[\exp(\beta_2 -1)]$
- $\beta_2 > 0.1$ → genauere Approximation: $100[\exp(\beta_2 -1)]$

Beispiel: Lohngleichung

Mit Hilfe eines linearen Regressionsmodells soll der Effekt der Ausbildungszeit in Jahren (educ), der Berufserfahrung in Jahren (exper) und der Betriebszugehörigkeit in Jahren (tenure) auf den Logarithmus des Stundenlohns (lnwage) untersucht werden:

Schätzung: $lnwage = 0.284 + 0.092 educ + 0.00412 exper + 0.022 tenure$

Nach der Kontrolle für educ und tenure hat die Berufserfahrung (exper) in Jahren keinen Effekt auf den Logarithmus des Stundenlohns.

$\beta_2 = 0.092 > 0 \rightarrow$ educ hat einen positiven Effekt

Eine Zunahme der Ausbildung um 1 Jahr, lässt den Stundenlohn um 9.2% ($=100\beta_2$) oder genauer um 9.64% $= 100(\exp(\beta_2)-1)$ steigen, wenn educ und tenure konstant bleiben (ceteris paribus).

Schätzung von y when lny die abhängige Variable ist

Regressionsmodell: $lny = \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$

x_i können auch Logarithmen darstellen $x_2 = \ln(sales)$

Schätzfunktion: $\widehat{lny} = b_1 + b_2 x_2 + \ldots + b_k x_k \quad \rightarrow \hat{y} = \exp(\widehat{lny})$

Annahme: u ist normalverteilt $\rightarrow u \approx N(0,\sigma^2)$

$E(y|\mathbf{x}) = \exp(\sigma^2/2)\exp(\beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$

$\mathbf{x}$ = Vektor der unabhängigen Variablen

Wenn $u \approx N(0,\sigma^2) \rightarrow E[\exp(u)] = \exp(\sigma^2/2) \rightarrow$ Verzerrung

Bessere Prognose: $\hat{y} = \exp(s_e^2/2)\exp(\widehat{lny})$

Adjustierungsfaktor $\rightarrow \exp(s_e^2/2) > 1 \quad$ da $\; s_e^2 > 0$

Für grosse Werte s_e^2 kann dieser Adjustierungsfaktor gross ausfallen

Schätzung von y when lny die abhängige Variable ist

Regressionsmodell: $\ln y = \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$

Annahme: u ist nicht normalverteilt aber unabhängig von x

$E(y|\mathbf{x}) = \alpha_0 \exp(\beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$

wobei $\alpha_0 = E[\exp(u)] > 1$

Geschätzte Residuen: $e_i = \ln y_i - b_1 - b_2 x_2 - \ldots - b_k x_k$

Schätzer für α_0: $a_0 = \frac{1}{N}\sum_{i=1}^{N} \exp(e_i)$

Beispiel: Lohngleichung

Mit Hilfe eines linearen Regressionsmodells soll der Effekt des Unternehmensumsatzes (sales), der Börsenkapitalisierung (mktval) und der Betriebszugehörigkeit in Jahren (ceoten) auf den CEO-Gehalt (wage) untersucht werden:

Schätzung: lwage = 4.504 + 0.163lsales + 0.109lmktval + 0.0117ceoten

lwage = ln(wage) lsales = ln(sales) lmktval = ln(mkval)

$a_0 = 1.136$

CEO-Gehalt für sales = 5000 ($5bn), mktval = 10’000 (10bn), ceoten = 10

lwage = 4.504 + 0.163ln(5000) + 0.109ln(10’000) + 0.0117(10) ≅ 7.013

$\widehat{wage}$ = exp(7.013) ≅ 1’110.983 x1000 = $1’110’983

Benutzung von a_0: 1.136x1’110’983 = $1’262’077

Log-log Modell: $R^2 = 0.318$

Interpretation: Das log-log Modell erklärt 31% der Varianz von lny

Bestimmtheitsmass

Modell 1: $\ln y = b_1 + b_2x + e$ Modell 2: $y = b_1 + b_2x + e$

Verwendung von R^2 für Modellvergleiche:

✓ die abhängige Variable y wurde nicht transformiert

✓ gleiche Anzahl erklärender Variablen.

⇒ nicht für einen Vergleich zwischen Modellen 1 und 2

Verwendung von $\overline{R}^2$ und Informationskriterien für Modellvergleiche:

✓ die abhängige Variable y wurde nicht transformiert

✓ unterschiedliche Anzahl erklärender Variablen

⇒ nicht für einen Vergleich zwischen Modellen 1 und 2

R^2 = Quadrat des Stichproben-Korrelationskoeffizienten zwischen y und $\hat{y}$.

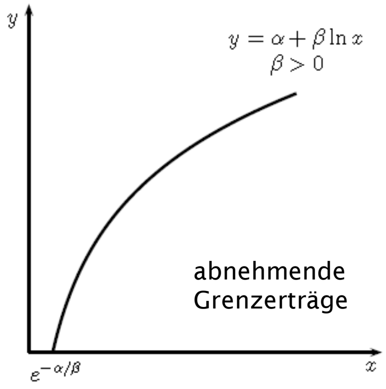
Wenn R^2 nicht direkt vergleichbar ist → Quadrat des Korrelationskoeffizienten zwischen y und $\hat{y}$ angeben.

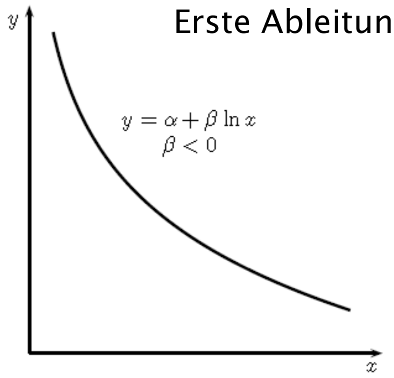
R^2 für Mehrfachregression

- Mehrfachregression: R^2 = Korrelationskoeffizienten$^2(y,\hat{y}) = r^2(y,\hat{y})$
- Korrelationskoeffizient für Lohnmodell: $r(y,\hat{y}) = 0.493$
- $r(y,\hat{y})^2 = 0.243 = R^2$
- Das log-log Modell erklärt 24% der Varianz des CEO-Gehaltes
- Konkurrierendes Modell: $wage = \beta_1 + \beta_2\, sales + \beta_3\, mktval + \beta_4\, ceoten + u$
- Schätzung: $\widehat{wage} = 613.4 + 0.019 sales + 0.0234\, mktval + 12.7\, ceoten$
- $R^2 = 0.201$ → Modell erklärt nur 20% der Varianz von y (CEO-Gehalt)
- Die Informationskriterien dürfen nicht zum Vergleichszweck herangezogen werden!
- Konsequenz: Log-log Spezifikation wird vorgezogen
- Vorteil: Bessere Interpretation der Parameter

Linear-logarithmisches Modell

Lin-log Modell: $y_i = \beta_1 + \beta_2 \ln x_i + u_i$





Erste Ableitung: $\frac{dy_i}{dx_i} = \frac{dy_i}{d(\ln x_i)} \frac{d(\ln x_i)}{dx_i} = \frac{\beta_2}{x_i}$

$$\Delta y = \beta_2 \Delta \ln(x)$$

$$\beta_2 = \frac{\Delta y}{\Delta \ln x} \approx \frac{\Delta y}{\Delta x / x}$$

$$dy_i = \beta_2 \frac{dx_i}{x_i} = \frac{\beta_2}{100}\left(100 \frac{dx_i}{x_i}\right)$$

$$\frac{\beta_2}{100} \approx \frac{dy}{(dx/x)100} = \frac{\text{absolute dy}}{\text{prozentuelle dx}}$$

Marginale Veränderung von y hängt auch vom Niveau der Variable x ab.

Eine Zunahme von x um ein Prozent führt ceteris paribus zu einer absoluten Änderung von y um 0.01 x β_2 Einheiten.

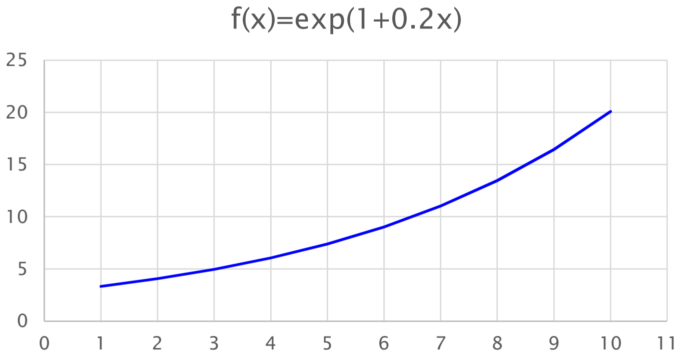
Approximation Beispiel

Funktion: $y = 1 + 0.2\ln x$

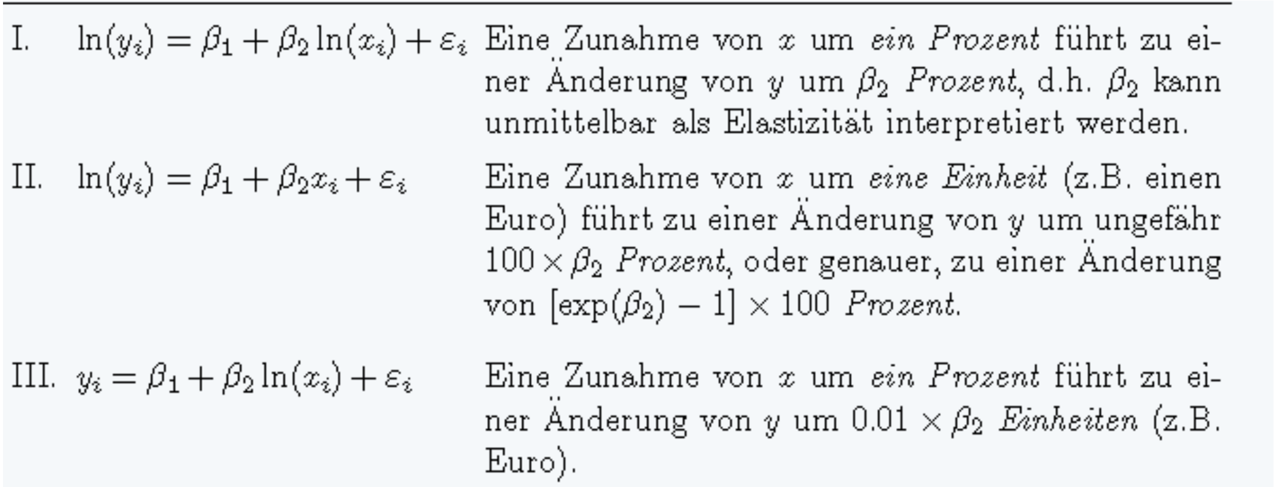
Frage: Wie verändert sich y, wenn x um ein Prozent zunimmt?

x	Δx/x	%Δx	1+0.2lnx	Δy	Δy%
5			1.32188		
5.05	0.05	1%	1.32387	0.00199	0.199%

β x100
→ 0.2%



Zusammenfassung



Spezifikation	Interpretation
I. $\ln(y_i) = \beta_1 + \beta_2 \ln(x_i) + \varepsilon_i$	Eine Zunahme von x um *ein Prozent* führt zu einer Änderung von y um β_2 *Prozent*, d.h. β_2 kann unmittelbar als Elastizität interpretiert werden.
II. $\ln(y_i) = \beta_1 + \beta_2 x_i + \varepsilon_i$	Eine Zunahme von x um *eine Einheit* (z.B. einen Euro) führt zu einer Änderung von y um ungefähr $100 \times \beta_2$ *Prozent*, oder genauer, zu einer Änderung von $[\exp(\beta_2) - 1] \times 100$ *Prozent*.
III. $y_i = \beta_1 + \beta_2 \ln(x_i) + \varepsilon_i$	Eine Zunahme von x um *ein Prozent* führt zu einer Änderung von y um $0.01 \times \beta_2$ *Einheiten* (z.B. Euro).

Das entscheidende Argument für die Wahl der Funktionsform sollte sein, welches Modell mit der Theorie konsistent ist und die Daten am besten abbildet.

MWD Test

Zwei konkurrierende Modelle für Rosennachfrage:
Lineares Modell 1: $y = \beta_1 + \beta_2 PR + \beta_3 PN + u$
Log-log Modell 2: $\ln y = \beta_1 + \beta_2 \ln PR + \beta_3 \ln PN + u$

- Schritt 1: Schätze Modell 1 und speichere die geschätzten $\hat{y}$ Werte, als yhat1 bezeichnet → «hat» steht für forecast und 1 für Modell 1
- Schritt 2: Schätze Modell 2 und speichere die geschätzten $\widehat{\ln y}$ Werte, als yhat2
- Schritt 3: Definiere die Variable z_1 = ln(yhat1) – yhat2
- Schritt 4: Regressiere y auf die Regressoren x und z_1
  Regressionsmodell: $y = \beta_1 + \beta_2 PR + \beta_3 PN + \beta_4 z_1 + u$

Hypothesentest:
H_0: y ist eine lineare Funktion der Regressoren PR und PN
H_1: lny ist eine linear Funktion der Regressoren lnPR und lnPN
Regel: Verwerfe H_0 wenn der Koeffizient von z_1 statistisch signifikant ist.

MWD Test, zweiter Teil

Zwei konkurrierende Modelle für Rosennachfrage:

Lineares Modell 1: $y = \beta_1 + \beta_2 PR + \beta_3 PN + u$

Log-log Modell 2: $\ln y = \beta_1 + \beta_2 \ln PR + \beta_3 \ln PN + u$

Hypothesentest:

H_0: y ist eine lineare Funktion der Regressoren PR und PN

H_1: lny ist eine linear Funktion der Regressoren lnPR und lnPN

- Schritt 5: Definiere die Variable z_2 = exp[yhat2] – yhat1
- Schritt 6: Regressiere lny auf die Regressoren lnx und z_2

Regression: $\ln y = \beta_1 + \beta_2 \ln PR + \beta_3 \ln PN + \beta_4 z_2 + u$

Regel: Verwerfe H_1 wenn der Koeffizient von z_2 statistisch signifikant ist.

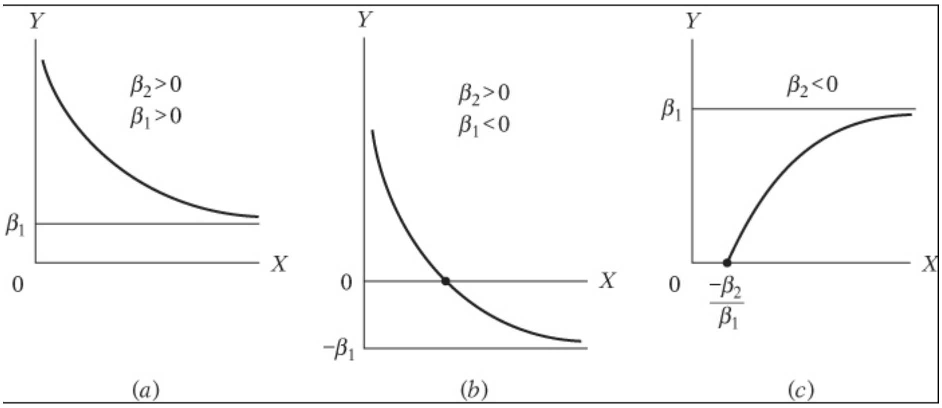
Falls H_0 wahr wäre → Variable z_1 = ln(yhat1) – yhat2 wird klein → sollte nicht statistisch signifikant sein.

Inverses Modell

Verwendung des Kehrwertes: $$y_i = \beta_1 + \beta_2 \frac{1}{x_i} + u_i$$

Beispiel: Kindersterblichkeit in Abhängigkeit der Alphabetisierungsrate und des pro-Kopf Bruttosozialproduktes

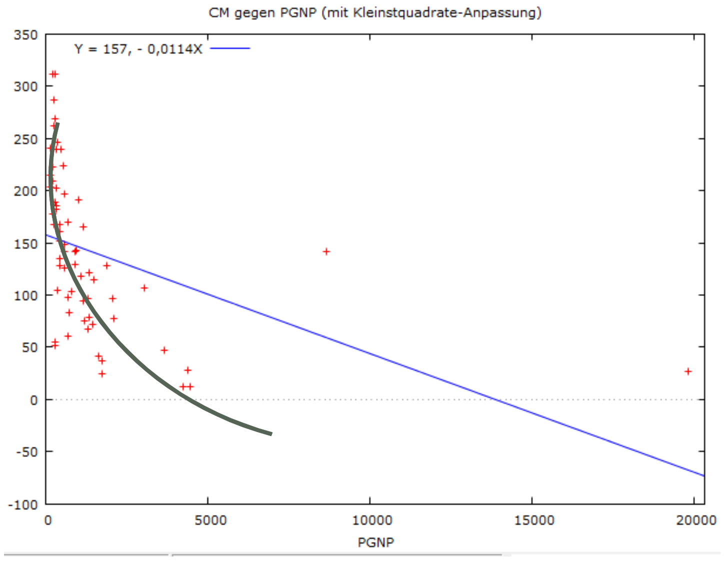
$$\frac{dy_i}{dx_i} = -\frac{\beta_2}{x_i^2}$$



Die marginale Veränderung von y hängt auch vom Ausgangsniveau x ab.

Inverses Modell: Beispiel

- Zusammengefasste Geburtenziffer: Durchschnittliche Zahl der lebendgeborenen […] Kinder, die eine Frau im Verlauf ihres Lebens
- […] zur Welt bringen würde, wenn die derzeitigen altersspezifischen Geburtenziffern unverändert blieben



$$\widehat{CM} = 81.794 + 27'237.17(1/PGNP)$$

Log-inverses Modell

Log-inverses Modell: $\ln y_i = \beta_1 + \beta_2 \frac{1}{x_i} + u_i$

Linearer Zusammenhang zwischen den Variablen $1/x_i$ und $\ln y_i$

Beispiel: Verkaufsumsatz in Abhängigkeit von den Werbeausgaben.

Exponentielle Form: $y_i = e^{\beta_i + \beta_2 (1/x_i)}$

$$\frac{dy_i}{dx_i} = -\frac{\beta_2}{x_i^2} e^{\beta_1 + \beta_1 (1/x_i)}$$

β_2-Wert negativ → positive marginale Veränderung

Die marginale Veränderung von y hängt auch vom Ausgangsniveau x ab.

Polynomen

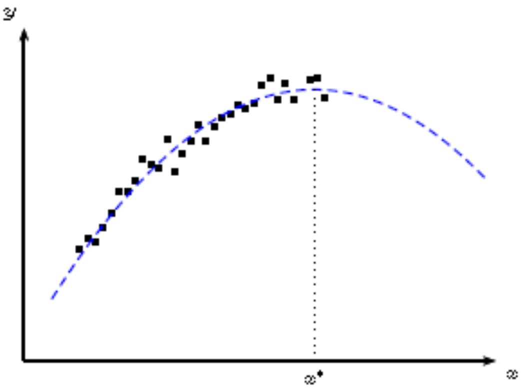
- Die Verwendung von Polynomen ermöglicht die Modellierung von Nichtlinearitäten. In den meistens Fällen beschränkt man sich auf quadratische Funktionsformen: $y = \beta_1 + \beta_2 x + \beta_3 x^2 + u$
- Diese Funktion ist linear in den Parametern β_i und kann deshalb mittels Regression geschätzt werden.
- Die Funktion ist nichtlinear in der erklärenden Variable x.
- Allgemeine Form für eine kubische Beziehung: $y = b_1 + b_2 x + b_3 x^2 + b_4 x^3$

Quadratische Modelle

Quadratisches Modell: $y = b_1 + b_2 x + b_3 x^2 + e$

Teste, ob es einen signifikanten Zusammenhang zwischen x und y gibt → F-Test: H_0: $b_2 = 0 = b_3 = 0$

Marginaler Effekt von x auf y ist nicht konstant: $\frac{d\hat{y}}{dx} = b_2 + 2b_3 x$



Achtung: b_2 misst den marginalen Effekt nur im Punkt $x = 0$ → nur selten von Interesse. Marginaler Effekt im Mittelwert von x oder in einem anderen gut interpretierbaren Punkt angeben!

Extremwert x*: Maximum oder Minimum wenn dy/dx = 0 → $x^* = -b_2/2b_3$

Quadratische Modelle unterstellen einen symmetrischen Verlauf→ haben in der Stichprobe einen guten Fit, aber liefern sehr schlechte Prognosen!

Quadrierte erklärende Variablen

- Zur Einbeziehung quadrierter erklärender Variablen: Wachsende oder sinkende (partielle) marginale Effekte können in Regressionsmodellen untersucht werden

Modell: $y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \ldots + \beta_k x_{k,t} + u_t$

- Interpretation: b_j = Veränderung von y, falls x_j um 1Einheit steigt ceteris paribus → (partielle) marginaler Effekt konstant und hängt nicht von x_j ab.
- Quadratisches Modell: $y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{2t}^2 + \beta_4 x_{3t} + \ldots + \beta_k x_{k-1,t} + u_t$
- Interpretation: b_2 beschreibt nicht allein die Veränderung von y bei einer Veränderung von x_2.
- Regressionsfunktion: $\hat{y}_t = b_1 + b_2 x_{2t} + b_3 x_{2t}^2 + b_4 x_{3t} + \ldots + b_k x_{k-1,t}$
- Falls $x_3, \ldots, x_{k-1}$ konstant gehalten werden: $\frac{\Delta \hat{y}}{\Delta x_2} \cong b_2 + 2b_3 x_2$

  Approximation: $\Delta\hat{y} \approx (b_2 + 2b_3x_2)\Delta x_2$
- Damit hängt der geschätzte (partielle) marginale Effekt von x_2 auf y auch von b_3 sowie den Werten von x_2 ab. Häufig ist b_2 positiv und b_3 negativ.

Beispiel: Lohngleichung

Mit Hilfe eines Regressionsmodells soll der Effekt der Berufserfahrung in Jahren (exper) und der quadrierten Berufserfahrung in Jahren (exper2) auf den Stundenlohn untersucht werden. N = 526

Regressionsfunktion:

Wage = 3.725 + 0.298exper – 0.00612exper^2

$$\frac{d\hat{y}}{dx_2} = b_2 + 2b_3x_2 = 0.298 - 2(0.00612)x_2$$

Damit ergibt sich ein geschätzter sinkender positiver Effekt von exper:

- Das erste zusätzliche Jahr an Berufserfahrung (ausgehend von exper = 0) führt zu einer geschätzten Erhöhung des Stundenlohnes um $0.298.
- Erhöhung der Berufserfahrung von ein auf zwei Jahre: geschätzte Steigerung des Stundenlohnes: 0.298 – 2 · 0.0061 · 1 = $0.286
- Erhöhung der Berufserfahrung von zehn auf elf Jahre: geschätzte Steigerung des Stundenlohnes: 0.298 – 2 · 0.0061 · 10 = $0.176

Beispiel: Lohngleichung

Schätzung: Wage = 3.725 + 0.298exper – 0.00612exper^2

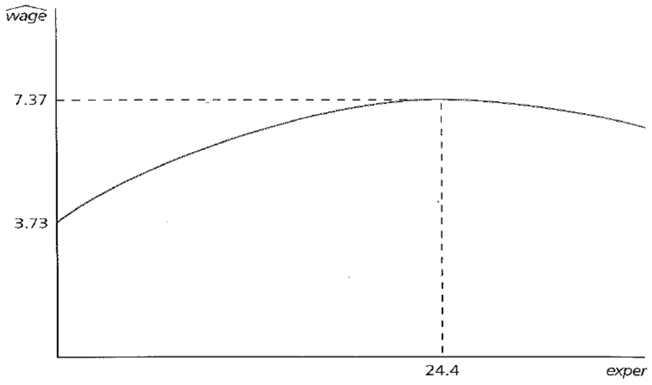
⇔ y = b_1 + $b_2$exper + $b_3\text{exper}^2$

- Falls b_2 positiv und b_3 negativ sind, ergibt sich immer ein positiver Wert von x_2 (exper), bei dem der geschätzte Effekt von x_2 auf y null ist.
- Vor diesem Punkt hat x_2 einen (mit x_2 sinkenden) positiven und nach diesem Punkt einen negativen geschätzten marginalen Effekt.
- Für diesen Wendepunkt (Scheitelpunkt) gilt: $x_2^* = \left|b_2/2b_3\right|$

- Beispiel: Wendepunkt des geschätzten Effektes von exper mit b_2 = 0.298 und b_3 = – 0.0061:
- $|x_2^*|$ = |0.298/2(– 0.0061)| = 24.43 Jahre

Lohngleichung: Interpretation

- Bei einer Berufserfahrung von mehr als 24 Jahren ergibt sich ein unerwarteter negativer geschätzter Effekt durch steigende Berufserfahrung.
- Falls nur wenige Personen in der Stichprobe eine solch hohe Berufserfahrung besitzen würden, kann dieses Ergebnis ignoriert werden (da dann lediglich der positive geschätzte Effekt eine Rolle spielen würde).
- Im Beispiel haben 28% der Personen eine derart hohe Berufserfahrung. Eine Erklärung wären verzerrte Schätzungen, da wichtige Faktoren nicht einbezogen werden, oder aber tatsächliche negative Effekte bei hohem exper.



Interaktionsmodelle

Als erklärende Variablen können auch Produkte einzelner erklärender Variablen verwendet werden.

$\hat{y} = b_1 + b_2 x_2 + b_3 x_3 + b_4 x_2 x_3$

x_2, x_3: Hauptterme $x_2 x_3$: Interaktionsterm

Der marginale Effekt von x_2 hängt vom Niveau von x_3 ab: $\frac{\partial \hat{y}}{\partial x_2} = b_2 + b_4 x_3$

Hat x_2 einen Einfluss auf y? → F-Test

Nullhypothese H_0: $\beta_2 = \beta_4 = 0$

Achtung: Der Koeffizient β_2 misst den marginalen Effekt von x_2 nur im Punkt $x_3 = 0$!

Änderung von x_3, wenn x_2 konstant gehalten wird: $\frac{\partial \hat{y}}{\partial x_3} = b_3 + b_4 x_2$

Der Koeffizient des Interaktionsterms ist die zweite Ableitung:

$$\frac{\partial \hat{y} / \partial x_3}{\partial x_2} = \frac{\partial^2 \hat{y}}{\partial x_2 \partial x_3} = b_4$$

b_4 gibt an, wie sich der marginale Effekt von x_2 ändert, wenn x_3 um eine Einheit zunimmt.

Interaktionsmodelle

$\hat{y} = b_1 + b_2 x_2 + b_3 x_3 + b_4 x_2 x_3$

Marginaler Effekt von x_2 gemessen im Mittelwert der anderen Variablen.

$$\frac{\partial \hat{y}}{\partial x_2} = b_2 + b_4 \bar{x}_3 \qquad (1)$$

Schätzung des marginalen Effektes im Mittelwert der anderen Variablen durch einfache Variablentransformation:

$$y = \alpha_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 (x_2 - \bar{x}_2)(x_3 - \bar{x}_3)$$

α_2 misst den marginalen Effekt von x im Mittelwert von x_3, d.h. gibt exakt den gleichen Wert als (1).

Durch Ausmultiplikation des Interaktionsterms haben wir $\alpha_2 = \beta_2 + \beta_4 \bar{x}_3$

Beispiel: Lohngleichung (1)

Lohngleichung: $\ln\text{wage} = b_1 + b_2\text{educ} + b_3\text{exper} + b_4\text{exper}^2 + b_5\text{exper}\times\text{educ}$

exper: Berufserfahrungsjahre

educ: Ausbildungsjahre

Alle Koeffizienten sind hoch signifikant auf 1%-Niveau

Durchschnittswerte:

$\overline{\text{exper}} = 19.5$

$\overline{\text{educ}} = 13.815$

y = Ln(wage)	Modell (1)	Modell (2)
const	1.44742	1.2482
educ	0.04235	**0.05677**
exper	0.0294	**0.03962**
exper^2	–0.0006	-0.0006
educ x exper	0.00074	
$(\text{educ}-\overline{\text{educ}})\text{x } (\text{exper}-\overline{\text{exper}})$		0.00074

Marginaler Effekt von educ:

$$\ln(\text{wage})/\partial\,\text{educ} = b_2 + b_5\text{exper}$$

Marginaler Effekt = 0.04235 + 0.00074(19.5) = **0.05677** = α_2 (reparametrisiertes Modell 2)

Für jemanden mit 19.5 Jahren Berufserfahrung bringt ein zusätzliches Ausbildungsjahr (educ) ungefähr einen um 5.67% höheren Stundenlohn.

Beispiel: Lohngleichung (2)

Lohngleichung: $\ln\text{wage} = b_1 + b_2\text{educ} + b_3\text{exper} + b_4\text{exper}^2 + b_5\text{exper}\times\text{educ}$

exper: Berufserfahrungsjahre

educ: Ausbildungsjahre

$\overline{\text{exper}} = 19.5$

$\overline{\text{educ}} = 13.815$

y = Ln(wage)	Modell (1)	Modell (2)
const	1.44742	1.2482
educ	0.04235	**0.05677**
exper	0.0294	**0.03962**
exper^2	–0.0006	-0.0006
educ x exper	0.00074	
$(\text{educ}-\overline{\text{educ}})\text{x } (\text{exper}-\overline{\text{exper}})$		0.00074

Marginaler Effekt von exper:

$$\ln(\text{wage})/\partial\,\text{exper} = b_3 + 2b_4\text{exper} + b_5\text{educ}$$

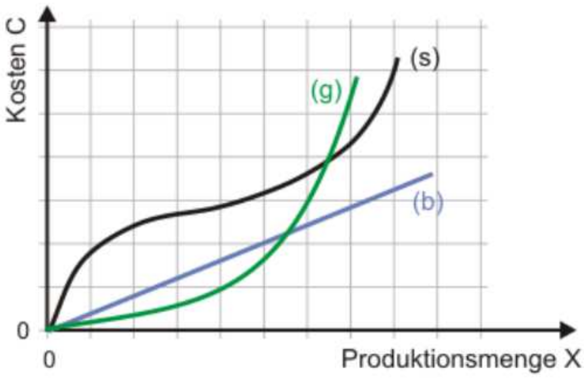
educ = 13.815 und exper = 0

Marginaler Effekt = 0.0294 + 0 + 0.00074 x13.815 ≈ **0.03962**

Marginaler Effekt nimmt ceteris paribus (c.p.) mit der Ausbildungsdauer zu, und ist c.p. desto höher, je geringer die Berufserfahrung ($b_4 < 0$) ist.

Kostenfunktion

Kostenverläufe bei unterschiedlichen Produktionstechnologien:
(s) ertragsgesetzlicher Kostenverlauf
(g) neoklassischer Kostenverlauf (Cobb-Douglas mit Homogenitätsgrad r < 1),
(b) lineare Technologie (Leontief oder Cobb-Douglas mit r = 1)



Mit polynomischen Modellen (quadratische oder kubische Modelle) kann man zwar manchmal einen sehr guten Fit in der Stichprobe erreichen, aber für Prognosen sind sie meistens ziemlich unbrauchbar.

RESET-Test

Schätzung einer Kostenfunktion

Modell 1: $Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_{1i}$ richtiges Modell

Modell 2: $Y_i = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^2 + u_{2i}$

Modell 3: $Y_i = \gamma_1 + \gamma_2 X_i + u_{3i}$

Schätzfunktionen:

Modell 1: $\hat{Y}_i = 141.767 + 63.4778 X_i - 12.9615 X_i^2 + 0.939 X_i^3$ $\overline{R}^2 = 0.9975$

Modell 2: $\hat{Y}_i = 222.383 - 8.025 X_i + 2.5417 X_i^2$ $\overline{R}^2 = 0.908$

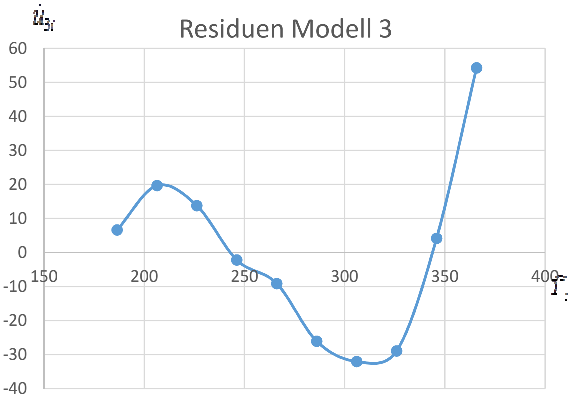
Modell 3: $\hat{Y}_i = 166.467 + 19.933 X_i$ $R^2 = 0.821$

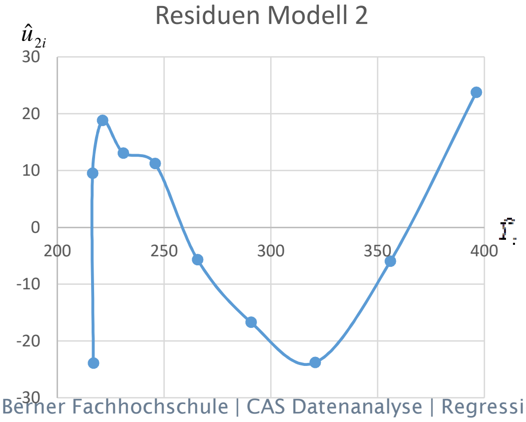
Modell 3 wird mit den i-ten Potenzen der Prognose für Y ergänzt:

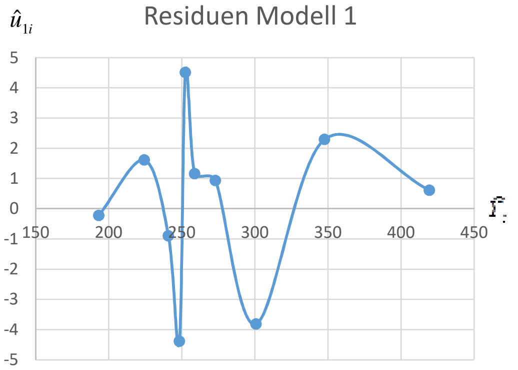
Erweitertes Modell 4: $Y_i = \gamma_1 + \gamma_2 X_i + \gamma_3 \hat{Y}_i^2 + \gamma_4 \hat{Y}_i^3 + u_i$

$\hat{Y}_i = 2140.22 + 476.6557 X_i - 0.9187 \hat{Y}_i^2 + 0.000119 \hat{Y}_i^3$ $R_e^2 = 0.9983$

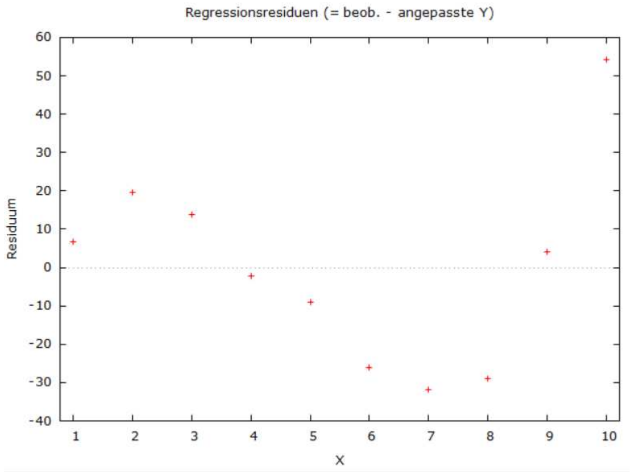
X	Modell 3	Modell 2	Modell 1
1	6.600	-23.899	-0.2215
2	19.667	9.503	1.613
3	13.734	18.823	-0.8975
4	-2.199	13.061	-4.387
5	-9.132	11.217	4.5105
6	-26.065	-5.709	1.161
7	-31.998	-16.717	0.9305
8	-28.931	-23.807	-3.815
9	4.136	-5.979	2.2905
10	54.203	23.767	0.613

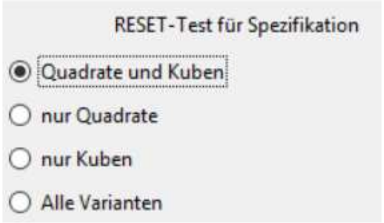

Residuen Modell 3


Residuen Modell 2


Residuen Modell 1

Residuengraph




RESET-Test für Spezifikation

● Quadrate und Kuben
○ nur Quadrate
○ nur Kuben
○ Alle Varianten

```
              Koeffizient       Std.-fehler     t-Quotient      p-Wert
  ---------------------------------------------------------------------
  const        2140,22          131,989           16,22      3,50e-06 ***
  X             476,552          33,3909           14,27      7,40e-06 ***
  yhat^2         -0,0918652       0,00619172      -14,84      5,90e-06 ***
  yhat^3          0,000118631     7,46258e-06      15,90      3,93e-06 ***

Teststatistik: F = 284,403480,
mit p-Wert = P(F(2,6) > 284,403) = 1,14e-006
```

RESET-Test: Ergebnisse

Der RESET-Test überprüft die Nullhypothese H_0: $\gamma_3 = \gamma_4 = 0$

$K^* = 4$ Anzahl Regressor im erweiterten Modell 4

$$Y_i = \gamma_1 + \gamma_2 X_i + \gamma_3 \hat{Y}_i^2 + \gamma_4 \hat{Y}_i^3 + u_i$$

Anzahl Restriktionen: $L = 2$ Anzahl Beobachtungen: $N = 10$

	Modell 3	Modell 4
R^2	0.84089	0.99833
S_{ee}	6202.533	64.7438

$F_c(0.95,2,6) = 5.14$

$F_e > F_c \rightarrow H_0$ verwerfen

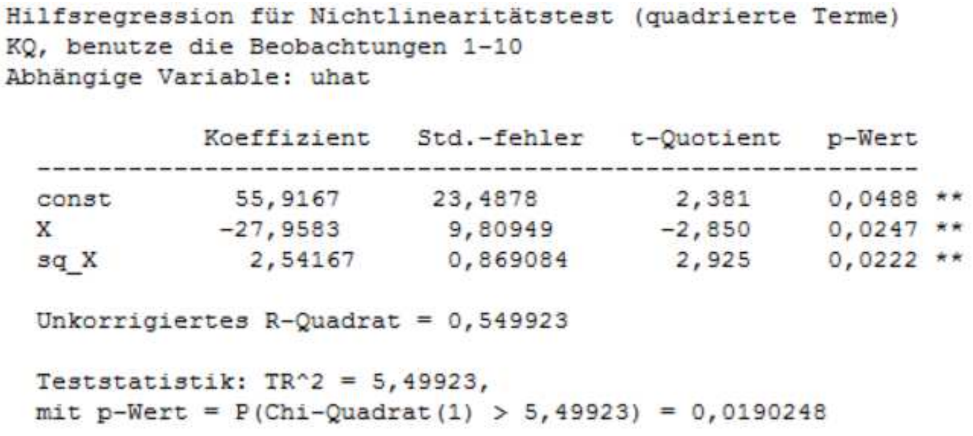
$$F_e = \frac{\left(R_e^2 - R^2\right)/L}{\left(1-R_e^2\right)/\left(N-K^*\right)} = \frac{(0.9983-0.8409)/2}{(1-0.9983)/(10-4)} = 284.403$$

$$F_e = \frac{\left(S_{ee} - S_{ee}^*\right)/L}{S_{ee}^*/\left(N-K^*\right)} = \frac{(6202.53-64.743)/2}{64.743/(10-4)} = 284.403$$

$P[F_{2,6} > 284.403] \cong 0$ ⇔ p-Wert nahe von Null →H_0 wird verworfen

---

Andere Nichtlinearitätstests

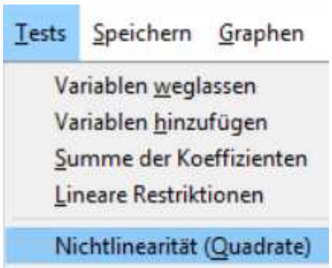


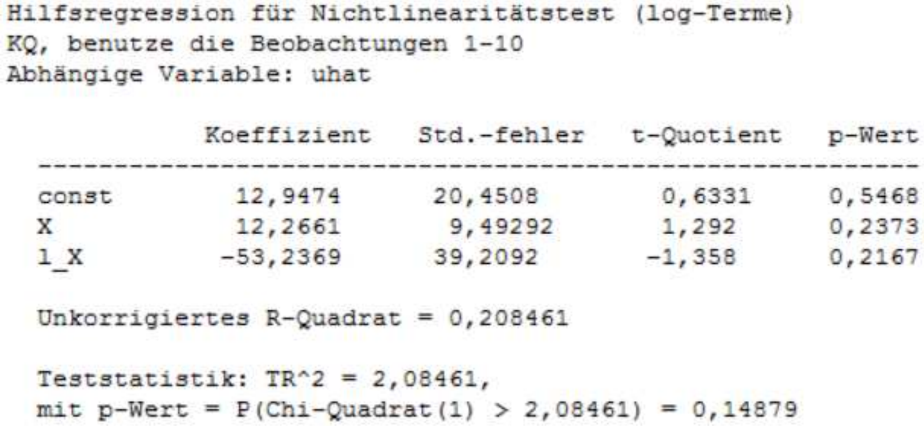
```
Hilfsregression für Nichtlinearitätstest (quadrierte Terme)
KQ, benutze die Beobachtungen 1-10
Abhängige Variable: uhat

             Koeffizient   Std.-fehler   t-Quotient    p-Wert
  ------------------------------------------------------------
  const        55,9167      23,4878         2,381      0,0488  **
  X           -27,9583       9,80949       -2,850      0,0247  **
  sq_X          2,54167      0,869084       2,925      0,0222  **

  Unkorrigiertes R-Quadrat = 0,549923

  Teststatistik: TR^2 = 5,49923,
  mit p-Wert = P(Chi-Quadrat(1) > 5,49923) = 0,0190248
```





```
Hilfsregression für Nichtlinearitätstest (log-Terme)
KQ, benutze die Beobachtungen 1-10
Abhängige Variable: uhat

             Koeffizient   Std.-fehler   t-Quotient    p-Wert
  -----------------------------------------------------------
  const        12,9474      20,4508        0,6331      0,5468
  X            12,2661       9,49292       1,292       0,2373
  l_X         -53,2369      39,2092       -1,358       0,2167

  Unkorrigiertes R-Quadrat = 0,208461

  Teststatistik: TR^2 = 2,08461,
  mit p-Wert = P(Chi-Quadrat(1) > 2,08461) = 0,14879
```