

12. Bivariate Datenanalyse

*Während einer nur Zahlen im Kopf hat,
kann er nicht auf den Kausalzusammenhang kommen*

Anonymus

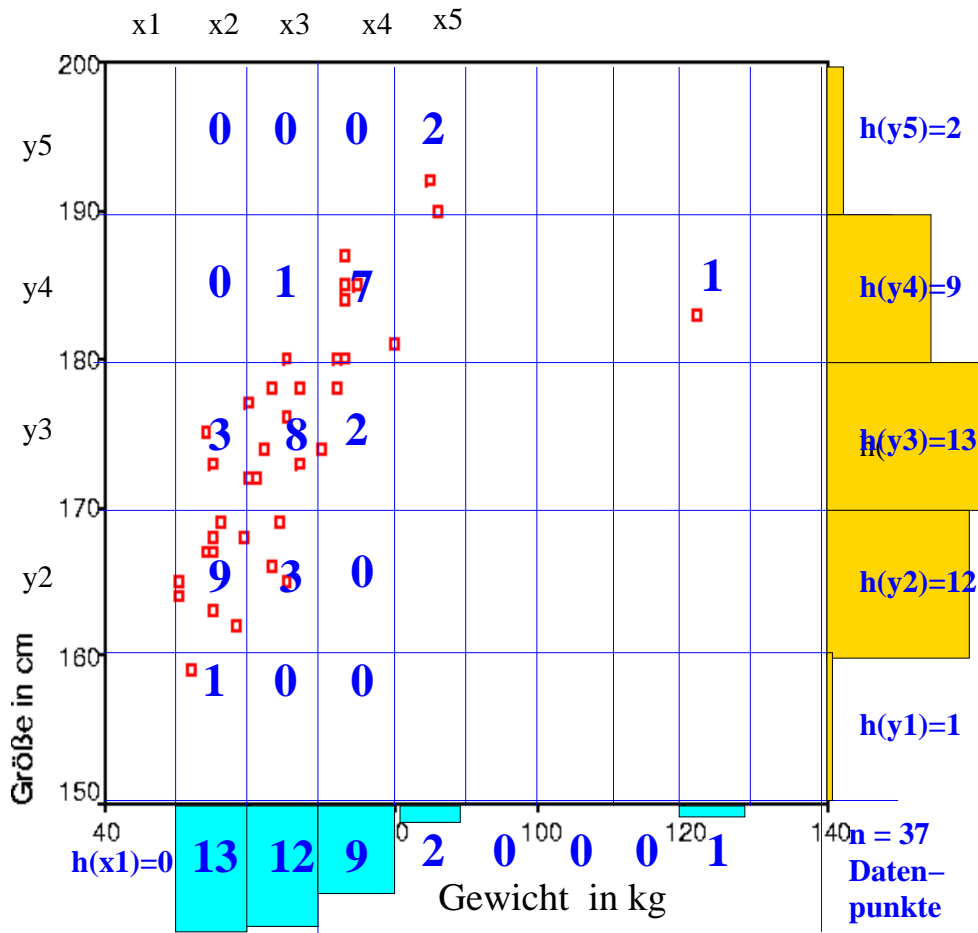
In den Kapiteln 4-11 wurden **univariate** Daten betrachtet:

- Von **univariaten Daten** spricht man, wenn bei der Datenerhebung nur ein Merkmal erfasst wurde bzw. man sich nur für ein Merkmal interessiert.
- Interessiert man sich für die Zusammenhänge zwischen mehreren Merkmalen (z.B. Verkehrsdichte *und* Geschwindigkeit), benötigt man die Mittel der **multivariaten** Datenanalyse (\Rightarrow Kap. 12-15).

Bei zwei Merkmalen spricht man von **bivariaten Daten**. Wie im eindimensionalen Fall gibt es grundsätzlich zwei Analysemethoden:

- Analyse *nichtklassierter Daten*: Man nimmt die einzelne Elemente der statistischen Reihe direkt als Ausgangspunkt der Untersuchungen \Rightarrow **Streudiagramm** bzw. **Scatter-Plot**.
- Analyse *klassierter Daten*: Vor der Analyse fasst man jeweils mehrere Merkmalsausprägungen in (Merkmalswerte-)Klassen zusammen \Rightarrow **Kreuztabelle**.

zu 12: Beispiele für die Darstellung



- **Streudiagramm** (nur für kardinalskalierte Daten)
- **Kreuztabelle** (auch für nominalskalierte Daten)

Übergang Streudiagramm \rightarrow Kreuztabelle: Rasterung ("Klassierung") in Intervallen mit den Klassenmitten x_i und y_j und Ermittlung der absoluten Häufigkeiten $h(x_i, y_j)$ ("zweidimensionale Strichlisten").

12.1. Nichtklassierte bivariate Daten

Wie bisher univariate Größen bezüglich beider Merkmalsausprägungen:

- $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$
- $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i,$
- $s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2,$
- $s_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2.$

Zusätzlich neue Analysemöglichkeiten:

- **Kovarianz:**

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- **Korrelationskoeffizient:**

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- **Verschiebungssatz der Kovarianz:**

$$n s_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}$$

12.2. Kreuztabelle I: Charakteristische Größen

- $x_i, i = 1, \dots, n$ bzw. $1, \dots, K_x$
 - Nichtklassierte Daten: n Verschiedene Ausprägungen des Merkmals X
 - Klassierte Daten: Klassenmitten der Einteilung des Merkmals X in K_x Klassen (Klassenmitten-“Stern” oft weggelassen)

- $y_j, j = 1, \dots, n$ bzw. $1, \dots, K_y$

Analoges für die Größe Y

- $h(x_i, y_j)$

Absolute Häufigkeit des Ausprägungspaares (x_i, y_j) bzw. der Wertepaare, bei denen X in die Klasse i und Y in die Klasse j fällt.

Achtung! Manchmal wird $h(x_i, y_j)$ auch als h_{ij} geschrieben. Fasst man dies als Matrixelement auf, so bezeichnet, entgegen der üblichen Konvention, i die *Spalte* und j die *Zeile*, zumindest, wenn x die horizontale und y die vertikale Achse bezeichnet.

- $f(x_i, y_j) = \frac{h(x_i, y_j)}{n}$ mit $n = \sum_{i=1}^{K_x} \sum_{j=1}^{K_y} h(x_i, y_j)$

Relative Häufigkeit, wobei n die Zahl der Wertepaare, d.h. die Zahl der Merkmalsträger, ist.

12.2. Kreuztabelle II: Randhäufigkeiten

- **Spaltensumme** $h(x_i) = \sum_{j=1}^{K_y} h(x_i, y_j)$

Absolute Randhäufigkeit der Merkmalsausprägung i bzw. der Klasse i des Merkmals X

- $f(x_i) = \frac{h(x_i)}{n}$

Relative Randhäufigkeit (Randhäufigkeitsverteilung) von X

Die Randhäufigkeiten für das zweite Merkmal Y sind analog:

- Absolute Randhäufigkeit:

$$\text{Zeilensumme } h(y_j) = \sum_{i=1}^{K_x} h(x_i, y_j)$$

- Relative Randhäufigkeit (Randhäufigkeitsverteilung)

$$f(y_j) = \frac{h(y_j)}{n}$$

Zu 12.2, I,II: Allgemeines Schema

Absolute Häufigkeiten und Randhäufigkeiten:

$y_j \backslash x_i$	x_1	x_2	\dots	x_{K_x}	Σ
y_{K_y}	$h(x_1, y_{K_y})$	$h(x_2, y_{K_y})$	\dots	$h(x_{K_x}, y_{K_y})$	$h(y_{K_y})$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
y_2	$h(x_1, y_2)$	$h(x_2, y_2)$	\dots	$h(x_{K_x}, y_2)$	$h(y_2)$
y_1	$h(x_1, y_1)$	$h(x_2, y_1)$	\dots	$h(x_{K_x}, y_1)$	$h(y_1)$
Σ	$h(x_1)$	$h(x_2)$	\dots	$h(x_{K_x})$	n

Relative Häufigkeiten und Randhäufigkeiten:

$y_j \backslash x_i$	x_1	x_2	\dots	x_{K_x}	Σ
y_{K_y}	$f(x_1, y_{K_y})$	$f(x_2, y_{K_y})$	\dots	$f(x_{K_x}, y_{K_y})$	$f(y_{K_y})$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
y_2	$f(x_1, y_2)$	$f(x_2, y_2)$	\dots	$f(x_{K_x}, y_2)$	$f(y_2)$
y_1	$f(x_1, y_1)$	$f(x_2, y_1)$	\dots	$f(x_{K_x}, y_1)$	$f(y_1)$
Σ	$f(x_1)$	$f(x_2)$	\dots	$f(x_{K_x})$	1

12.2. Kreuztabelle III: Unabhängigkeit

- **Relative bedingte Häufigkeiten** von X unter der Bedingung $Y = y_j$:

$$f(x_i|y_j) = \frac{h(x_i, y_j)}{h(y_j)}$$

- **Relative bedingte Häufigkeiten** von Y unter der Bedingung $X = x_i$:

$$f(y_j|x_i) = \frac{h(x_i, y_j)}{h(x_i)}$$

Beliebig skalierte Merkmale X und Y sind **empirisch unabhängig**, falls die relativen Häufigkeiten von X nicht von den Werten y_j und die relativen Häufigkeiten von Y nicht von x_i abhängen, also alle bedingten relativen Häufigkeiten von Y bzw. von X gleich sind. Sie sind dann gleich den relativen Randhäufigkeiten:

$$f(x_i|y_j) = f(x_i) \quad \text{für alle } y_j, j = 1, \dots, K_y$$

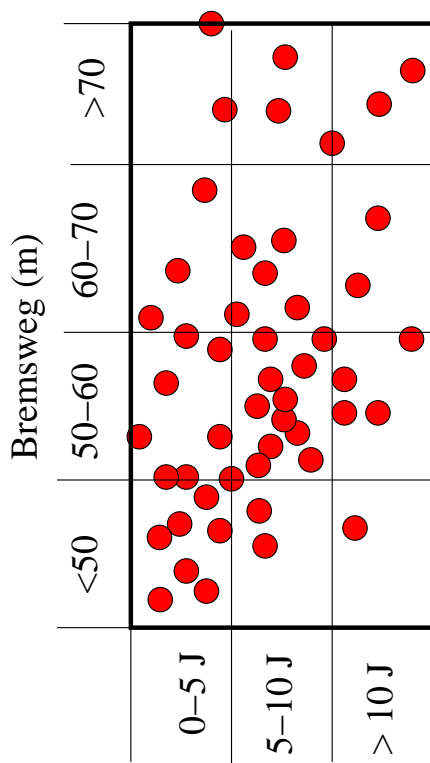
$$f(y_j|x_i) = f(y_j) \quad \text{für alle } x_i, i = 1, \dots, K_x$$

Damit folgt auch

$$f(x_i, y_j) = f(x_i)f(y_j)$$

12.3 Beispiel

Es werden von 50 Kfz verschiedenen Alters die Bremswege von 100 km/h zum Stopp gemessen. Aus der Urliste liegen das Streudiagramm Alter vs. Bremsweg, eine Klasseneinteilung sowie die resultierende Kreuzta-
belle zur Auswertung vor.



Alter des Fahrzeugs

	$h(x1)=20$	$h(x2)$	$h(x3)$	
$y4$	2	2	3	$h(y4)$
$y3$	3	5	2	$h(y3)$
$y2$	5	11	4	$h(y2)$
$y1$	10	2	1	$h(y1)=13$
	0-5 J	5-10 J	> 10 J	$n=50$
	$x1$	$x2$	$x3$	

$x =$ Alter des Fahrzeugs

12.3(b) Beispiel (Forts.)

Bestimmen Sie:

1. Die relativen Häufigkeiten $f(x_i, y_j)$,
2. die Randhäufigkeiten $h(x_i)$ bezüglich Merkmal "Alter" sowie $h(y_j)$ bezüglich Merkmal "Bremsweg" sowie die relativen Randhäufigkeiten $f(x_i)$, $f(y_j)$,
3. die Verteilungsfunktionen der Randhäufigkeiten,
4. die bedingten relativen Häufigkeiten sowie die Verteilung der Bremswege für 0-10 Jahre alte und für 10-15 Jahre alte Fahrzeuge,
5. die Altersverteilung der Fahrzeuge mit Bremswegen unter 50 m,
6. die arithmetischen Mittel der Bremswege der 0-5, 5-10 und der 10-15 Jahre alte Fahrzeuge,
7. und einen Check auf Unabhängigkeit (ohne nichtparametrischen Test!)

13. Regressionsanalyse

Als **(Einfach-)Regression** bezeichnet man die Annäherung eines Streudiagramms $\{(x_i, y_i), i = 1, \dots, n\}$ oder einer Kreuztabelle durch eine **Regressionsfunktion** $\hat{y}(x; a, b, \dots)$, so dass der “Fehler” zwischen den Daten und der Regressionsfunktion möglichst klein wird. Die Regressionsfunktion enthält neben der unabhängigen Variablen x i.A. mehrere Parameter a, b, \dots , welche zur Minimierung variiert werden.

- Die Regressionsanalyse wird auf *einseitig gerichtete Abhängigkeiten* angewendet, z.B. $x =$ unabhängige und $y =$ abhängige Koordinate. *Beispiel:* Trendbestimmung bei Zeitreihen mit der Zeit als unabhängigen Koordinate \Rightarrow Kap. 16-21
- Der “Fehler” ist üblicherweise durch die **Summe der Abweichungsquadrate** der abhängigen Koordinate y_i der Datenpunkte von der Regressionsfunktion $\hat{y}(x_i)$ an der Stelle der unabhängigen Koordinate quantifiziert. Wegen der Unterscheidung der Koordinaten in abhängigen und unabhängigen wird als Abweichung also der “vertikale” Abstand $y_i - \hat{y}(x_i)$ verwendet und *nicht* etwa der geometrisch kürzeste Abstand zwischen Datenpunkten und Regressionsfunktion.
- Regression ist insbesondere dann sinnvoll, wenn man zu einem Zusammenhang zwischen zwei Variablen
 - “verrauschte” Daten in Form eines Scatter-Plots (Streudiagramm) vorliegen hat,
 - ein als Regressionsfunktion formulierbares Modell des Zusammenhangs vorliegt oder vermutet wird.

Die Regression liefert dann Aussagen über die Werte der “Modellparameter” a, b , etc!

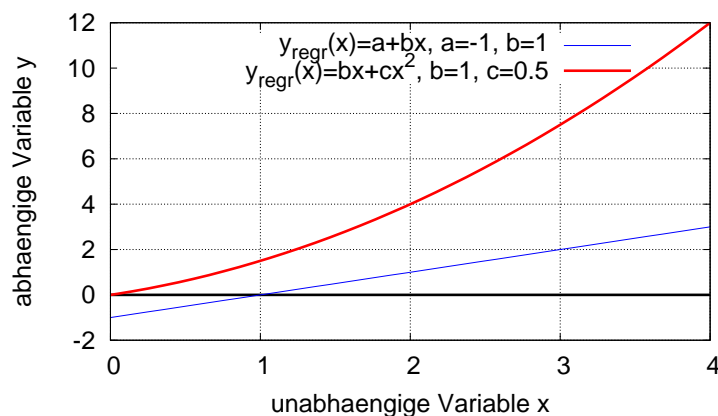
13.1 Beispiele für Regressionsfunktionen

- **Jährliche Gesamtkosten** (Fixkosten+laufende Kosten) für Kfz in Abhängigkeit der Kilometerleistung:

$$\hat{y}(x; b, c) = a + bx$$

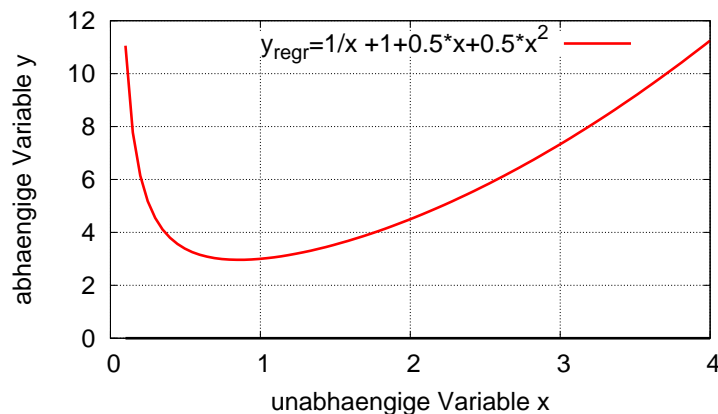
- **Anhalteweg** (Bremsweg + in der Reaktionszeit zurückgelegter Weg) von Autos in Abhängigkeit der Geschwindigkeit:

$$\hat{y}(x; b, c) = bx + cx^2$$



- **Treibstoffverbrauch** in Abhängigkeit der Geschwindigkeit:

$$\hat{y}(x; a_{-1}, a_0, a_1, a_2) = \frac{a_{-1}}{x} + a_0 + a_1x + a_2x^2$$



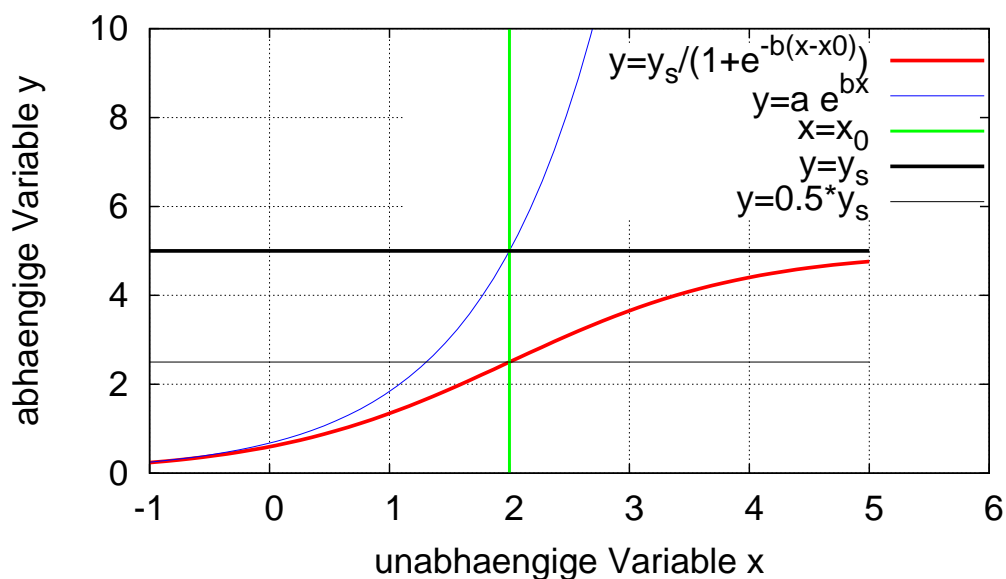
13.1 (b) Beispiele für Regressionsfunktionen II

- **Volkswirtschaftliche Größen** wie das BIP als Funktion der Zeit, oder: **Wertentwicklung** von z.B. Kfz als Funktion der Zeit:

$$\frac{d\hat{y}}{dx} = by \quad \Rightarrow \quad \hat{y}(x; a, b) = ae^{bx}$$

- **Gesättigte Wachstumsprozesse** wie Zahl der Kfz, Waschmaschinen, Mobiltelefone etc als Funktion der Zeit:

$$\frac{d\hat{y}}{dx} = by \left(1 - \frac{\hat{y}}{y_s} \right) \quad \Rightarrow \quad \hat{y}(x; y_s, b, x_0) = \frac{y_s}{1 + e^{-b(x-x_0)}}$$



- **Nachfrage-Preis-Relation** bei konstanter Preis-Elastizität:

$$\frac{x d\hat{y}}{y dx} = b = \text{const.} \quad \Rightarrow \quad \hat{y}(x; a, b) = ax^b$$

13.1 (c) Arten von Regressionsfunktionen

- **Einfache Regression**: nur eine unabhängige Variable; **mehrfache** oder multiple Regression: mehrere unabhängige Variable
- **Lineare Regression**: Linear in den unabhängigen Variablen, z.B. $\hat{y}(x) = a + bx$
- **Quasilineare Regression**: Zwar nichtlinear in den Variablen, aber linear in den Koeffizienten, z.B. $\hat{y}(x) = a + b/x$
- **Nichtlineare Regression**: Nichtlinear in Variablen und Koeffizienten, z.B. $\hat{y}(x) = ae^{bx}$ oder $\hat{y}(x) = ax^b$

13.2 Lineare einfache Regressionsfunktion

Minimiere die Fehlerquadratsumme

$$F = \sum_{i=1}^n e_i^2 \equiv \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (a + bx_i))^2$$

Die notwendige Bedingung: $F \stackrel{!}{=} \text{Extremum!}$, also $\partial F / \partial a = 0$, $\partial F / \partial b = 0$, ergibt

$$a = \bar{y} - b\bar{x},$$

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

Oder:

$$\hat{y}(x) = \bar{y} + b(x - \bar{x}), \quad b = \frac{s_{xy}}{s_x^2}$$

\bar{x}	$= \frac{1}{n} \sum_{i=1}^n x_i$	Mittelwert unabh. Var.,
\bar{y}	$= \frac{1}{n} \sum_{i=1}^n y_i$	Mittelwert abh. Var.,
s_x^2	$= \frac{1}{n} \sum (x_i - \bar{x})^2$	empirische Varianz,
s_{xy}	$= \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$	empirische Kovarianz

Aufgabe: Zeigen Sie dies alles !

13.2(b) Aufgaben zur linearen Regression

- (1) Von 4 Kfz sind das Alter und die Bremswege bei einer Vollbremsung von 100 km/h zum Stillstand gegeben:

Alter (Jahre)	4	7	11	2
Bremsweg (m)	50	80	70	45

- (a) Bestimmen Sie die Koeffizienten der linearen Regression
(b) Extrapolieren Sie den erwarteten mittleren Bremsweg für 15 Jahre alte Fahrzeuge
(c) Zeichnen Sie ein Streudiagramm der Daten und zeichnen Sie die Regression ein.
- (2) Es soll die mittlere Geschwindigkeit auf einer Autobahns pur bei "freier Fahrt" anhand einer Regression von Punkten eines "Fundamentaldiagramms" (Fluss-Dichte-Diagramms) ermittelt werden:

Verkehrsdichte ρ (Fz/km/Spur)	20	18	4	7	22
Verkehrsfluss Q (Fz/km/h)	1900	1400	500	900	2000

- (a) Für diesen Sachverhalt muss man bei einer linearen Regression die Regressionskonstante $a = 0$ setzen. Warum? Wie sieht die Bedingung für den verbleibenden Parameter b der Regressionsgeraden $\hat{y} = bx$ nun aus?
(b) Berechnen Sie die Regression. Welcher Geschwindigkeit entspricht sie?
(c) Plotten Sie das Fundamentaldiagramm und die Regressionsgerade

13.3 Nichtlineare einfache Regressionsfunktionen

Es gibt zwei Möglichkeiten der Analyse:

- Direktes Ableiten der Fehlersumme F und Nullsetzen: Bei quasi-linearer Regressionsfunktion erhält man ein lineares Gleichungssystem für die Koeffizienten.
- Variablentransformation, so dass sich *in den transformierten Variablen* wieder eine lineare Regression ergibt.

Die Variablentransformation verläuft nach folgendem Schema (Details in [folien8_trafo.pdf](#)):

1. Transformiere $x \rightarrow u$ und/oder $y \rightarrow w$, so dass aus $\hat{y}(x)$ die lineare Regression $\hat{w}(u) = \tilde{a} + \tilde{b}u$ hervorgeht.
2. Bestimme \tilde{a} und \tilde{b} mit linearer Regression.
3. Bestimme ursprüngliche Regressionsfunktion $\hat{y}(x)$ durch Rücktransformation.

Hinweis: Diese beiden Verfahren sind nur dann äquivalent, wenn man lediglich die *unabhängige* Variable (x) transformiert, aber die abhängige (y), bezüglich derer die Fehlerquadrate gebildet werden, unverändert lässt.

Beispiel und Aufgabe: Leiten Sie für die *hyperbolische Regression* $\hat{y}(x) = a + b/x$ die Bestimmungsgleichungen für die Parameter auf beide Arten her. Verwenden Sie für die Transformations-Methode die Transformation $x = 1/u$.

13.4 Grenzfunktion und Elastizität

Als *Elastizität* bezeichnet man die Nachgiebigkeit bzw. Flexibilität der abhängigen Variable bzw. seiner Regressionsfunktion \hat{y} bei Änderung der unabhängigen Variablen x . Dies wird charakterisiert durch die

Grenzfunktion (“absolute Elastizitätsfunktion”):

$$g(x) = \frac{\partial \hat{y}}{\partial x}$$

sowie durch die **(relative) Elastizitätsfunktion:**

$$\epsilon_{yx}(x) = \frac{x}{\hat{y}} \frac{\partial \hat{y}}{\partial x}$$

Als **Elastizität:** bezeichnet man die Elastizitätsfunktion an einer bestimmter Stelle x_0

- $|\epsilon_{yx}(x_0)| < 1 \Rightarrow$ unterproportional elastisch
z.B. X =Preis von, Y =Nachfrage nach Grundnahrungsmitteln, Heizöl, Benzin etc: Die *Preiselastizität* dieser zum Leben notwendigen Produkte ist unterproportional.
- $|\epsilon_{yx}(x_0)| > 1 \Rightarrow$ überproportional elastisch

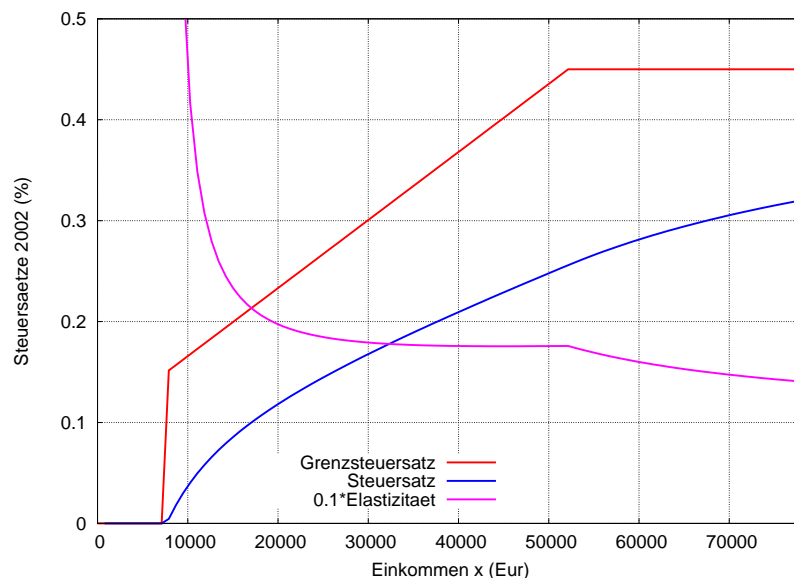
13.4(b) Beispiele und Aufgaben

Beispiel: Elastizität einer durch die hyperbolische Regression beschriebenen Preisrelation, z.B. kilometerbezogene Haltungskosten eines Kfz in Abhängigkeit der Fahrleistung (wie kommt man darauf?):

$$\epsilon_{yx}(x) = \frac{x}{a + \frac{b}{x}} \left(-\frac{b}{x^2} \right) = \frac{-b}{ax + b}$$

Aufgaben:

1. Welche Klasse von Regressionsfunktionen hat eine von x unabhängige Elastizität?
2. Erläutern Sie Grenzfunktion und Elastizitätsfunktion im Zusammenhang mit der Lohnsteuertabelle (sic!). Bringen Sie dabei die Begriffe "Grenzsteuersatz und "progressive Besteuerung" unter.



3. Für welche Preiselastizität sind die gesamten Ausgaben für ein Produkt unabhängig von dessen Preis?