

6. und einen Check auf Unabhängigkeit (ohne nichtparametrischen Test!)

Lösung:

(1) $n = 50$; z.B. $f(x_1, y_1) = h(x_1, y_1)/n = 10/50 = 0.2 =$ Anteil der bis 5 Jahre alten Kfz mit Bremswegen kleiner 50 m.

(2) Randhäufigkeiten $h(x_i)$ bezüglich Merkmal "Bremsweg":

$h(y_1) = \sum_{i=1}^3 h(x_i, y_1) = 13$: Zahl der Kfz mit Bremswegen < 50 m.

$h(y_2) = \sum_{i=1}^3 h(x_i, y_2) = 20$: Zahl der Kfz mit Bremswegen zwischen 50 und 60 m.
etc.

Rekulative Häufigkeiten: $f(y_1) = h(y_1)/n = 0.26 =$ Anteil der Kfz mit Bremswegen < 50 m, etc

Randhäufigkeiten $h(x_i)$ bezüglich Merkmal "Alter":

$h(x_1) = \sum_{j=1}^4 h(x_1, y_j) = 20$: Zahl der unter 5 Jahre alten Kfz

$f(x_1) = h(x_1)/n = 0.4$: Anteil der unter 5 Jahre alten Kfz, etc.

(3) Randhäufigkeitsverteilung bezüglich Alter: Diese ist gegeben durch den Punkt $(x_1^u, 0)$ sowie die Punkte $(x_i^o, F(x_i))$ mit

- x_1^u die Klassenuntergrenze der ersten Altersklasse,
- x_i^o die Klassenobergrenze der i -ten Altersklasse,
- $F(x_i) = \sum_{k=1}^i f(x_k)$ die relative Summenhäufigkeit,

also hier durch die Punkte

$$(0, 0), (5, 0.4), (10, 0.8), (15, 1).$$

Analog ist die Verteilungsfunktion bezüglich des Bremswegs durch die Punkte

$$(40, 0), (50, 0.26), (60, 0.66), (70, 0.86), (80, 1)$$

gegeben.

Um Verteilungsfunktionen zu bestimmen, braucht man auch die jeweils "außen" liegenden Grenzen der Randklassen, z.B. x_1^u und x_n^o . Falls es sich um "offene Randklassen" wie im Beispiel handelt, und nichts anderes gesagt ist, werden diese näherungsweise als geschlossenen Klassen mit gleicher Klassenbreite wie die anderen Klassen behandelt, also z.B. $x_1^u = 0$ und $x_3^o = 15$.

(4) Bedingte relative Häufigkeit der Bremswegklassen für über 10 Jahre alte Fahrzeuge: $i = 3$.

$$f(y_j|x_3) = h(y_j, x_3)/h(x_3)$$

also

$$f(y_1|x_3) = 1/10 = 0.1,$$

$$f(y_2|x_3) = 4/10 = 0.4,$$

$$f(y_3|x_3) = 2/10 = 0.2,$$

$$f(y_4|x_3) = 3/10 = 0.3$$

Bedingte relative Häufigkeit der Bremswegklassen für die unter 10 Jahre alten Kfz:

$$f(y_j|x = x_1 \text{ oder } x_2) = \frac{h(y_j, x_1) + h(y_j, x_2)}{h(x_1) + h(x_2)}$$

also

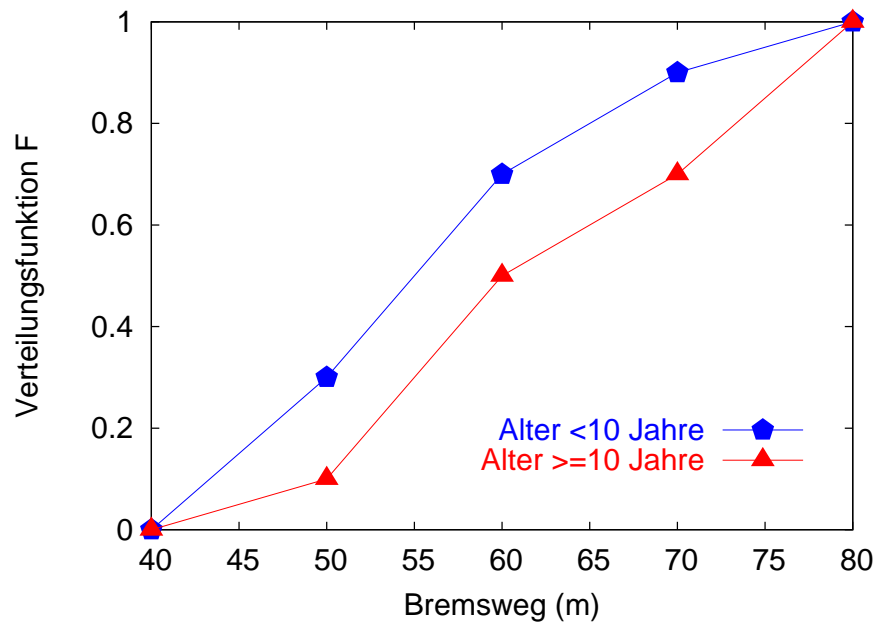
$$f(y_1|x = x_1 \text{ oder } x_2) = 12/40 = 0.3,$$

$$f(y_2|x = x_1 \text{ oder } x_2) = 16/40 = 0.4,$$

$$f(y_3|x = x_1 \text{ oder } x_2) = 8/40 = 0.2,$$

$$f(y_4|x = x_1 \text{ oder } x_2) = 4/40 = 0.1.$$

Die Verteilungsfunktion wird wieder wie oben aus den Punkten $(y_1^u, 0)$ sowie (y_j^o, F_j) , $j = 1..4$ berechnet, wobei j die zu den relativen Häufigkeiten $f(y_1|x_3)$ bzw. $f(y_j|x = x_1 \text{ oder } x_2)$ gehörige Summenhäufigkeit ist. Verbindet man die Punkte linear, ergibt sich als Grafik:



(5) Altersverteilung (x_i^o, F_i) für Bremswege < 50 m:

$$(0, 0), (5, 10/13), (10, 12/13), (15, 1).$$

- (6) z.B. $f(x_1, y_1) = \frac{1}{5}$, $f(x_1)f(y_1) = \frac{2}{5} + \frac{13}{50} = \frac{26}{250} \neq \frac{1}{5} \Rightarrow$ Bremsweg und Alter sind empirisch abhängig. Ob die Abhängigkeit allerdings *signifikant* ist, wird erst in der Test-Theorie am Ende von Statistik II untersucht.

Aufgabe zu Kap. 19.3 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 folgende Transformation der unabhängigen Variablen:

$$x = 1/z$$

(i) Lineare Regression der transformierten Daten

Mit folgender Transformation der unabhängigen Variablen:

$$x = \frac{1}{z}$$

wird die Regressionsfunktion $\hat{y}(z) = a + bz$ linear und man erhält *in der neuen unabhängigen Variable* z die üblichen Ausdrücke für die Koeffizienten:

$$a = \bar{y} - b\bar{z}, \quad b = \frac{s_{zy}}{s_z^2}.$$

Nach Rücktransformation $z = \frac{1}{x}$ erhält man für a und b folgende Ausdrücke:

$$a = \bar{y} - \frac{1}{n} \left(\sum_{i=1}^n \frac{1}{x_i} \right) b, \tag{1}$$

$$b = \frac{\sum_{i=1}^n \frac{y_i}{x_i} - \bar{y} \sum_{i=1}^n \frac{1}{x_i}}{\sum_{i=1}^n \frac{1}{x_i^2} - \frac{1}{n} \left(\sum_{i=1}^n \frac{1}{x_i} \right)^2} \tag{2}$$

(ii) Direkte Regression

Da $\hat{y}(x)$ quasilinear ist, ist dies ohne Probleme möglich:

$$F = \sum_{i=1}^n \left(y_i - a - \frac{b}{x_i} \right)^2 \stackrel{!}{=} \min!$$

$$\Rightarrow$$

$$\frac{\partial F}{\partial a} = 2 \sum_{i=1}^n \left(y_i - a - \frac{b}{x_i} \right) (-1) = 0,$$

$$\frac{\partial F}{\partial b} = 2 \sum_{i=1}^n \left(y_i - a - \frac{b}{x_i} \right) \left(\frac{-1}{x_i} \right) = 0.$$

Daraus ein lineares Gleichungssystem für a und b :

$$\begin{aligned} na + \left(\sum_{i=1}^n \frac{1}{x_i} \right) b &= \sum_{i=1}^n y_i, \\ \left(\sum_{i=1}^n \frac{1}{x_i} \right) a + \left(\sum_{i=1}^n \frac{1}{x_i^2} \right) b &= \sum_{i=1}^n \frac{y_i}{x_i}. \end{aligned}$$

Daraus als Ergebnis die Koeffizienten der direkten Regression:

$$a = \bar{y} - \frac{1}{n} \left(\sum_{i=1}^n \frac{1}{x_i} \right) b, \quad (3)$$

$$b = \frac{\sum_{i=1}^n \frac{y_i}{x_i} - \bar{y} \sum_{i=1}^n \frac{1}{x_i}}{\sum_{i=1}^n \frac{1}{x_i^2} - \frac{1}{n} \left(\sum_{i=1}^n \frac{1}{x_i} \right) \left(\sum_{i=1}^n \frac{1}{x_i} \right)} \quad (4)$$

Die Ausdrücke (3) und (1) für den Koeffizienten a sowie (4) und (2) für b sind identisch! Dies ist insofern anschaulich, als

- einerseits bei der Regressionsrechnung ja unsymmetrisch nur die Fehlerquadratsumme F bezüglich der *abhängigen* Variablen y minimiert wird,
- andererseits eine quasilineare Regressionsfunktion durch Transformation ausschließlich der *unabhängigen* Variablen zur linearen Funktion wird. Im quasilinearen Fall wird die Fehlerquadratsumme F also durch die Transformation gar nicht beeinflusst!

Bei Transformation der *abhängigen* Koordinate sind beide Methoden i.A. *nicht* äquivalent! (vgl. "Exkurse 2")

Aufgabe zu Kap. 19.4

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

Lösung:

- (1) Damit eine Regressionsfunktionen $\hat{y}(x)$ eine von x unabhängige Elastizität beschreibt, muss gelten:

$$\epsilon_{yx}(x) = \frac{x \, d\hat{y}}{\hat{y} \, dx} = C = \text{const.}$$

“Bruchrechnen” mit den Differentialquotienten ergibt

$$\frac{d\hat{y}}{\hat{y}} = C \frac{dx}{x}$$

Nun wird auf beiden Seiten das unbestimmte Integral gebildet:

$$\ln(\hat{y}) = C \ln(x) + D$$

mit der Integrationskonstante D .

Auflösen nach \hat{y} ergibt die gesuchte Klasse von Regressionsfunktionen (mit der neuen Konstanten $A = e^D$):

$$\hat{y}(x) = e^D e^{C \ln(x)} = A \left(e^{\ln(x)} \right)^C = \underline{\underline{Ax^C}}$$

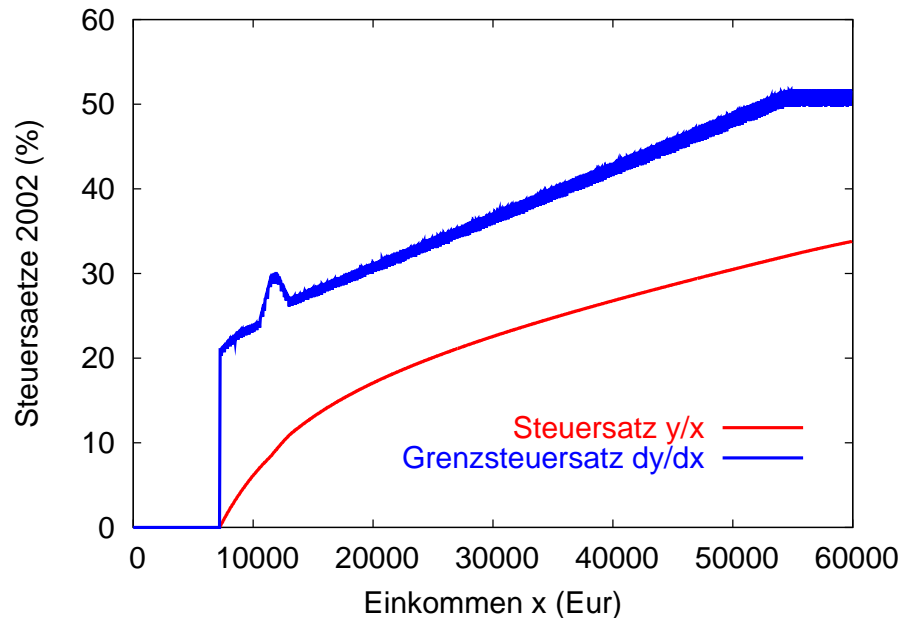
Die Elastizität ist dabei durch $\epsilon_{yx} = C = \text{const.}$ gegeben. Zum Beispiel

$$\hat{y}(x) = 7x: \epsilon_{yx} = 1,$$

$$\hat{y}(x) = \frac{0.4}{x}: \epsilon_{yx} = -1.$$

- (2) Sei x das Einkommen und $\hat{y}(x)$ die Regressionsfunktion für die bezahlte Lohnsteuer y . Ohne Absetzungsmöglichkeit wäre $\hat{y}(x) = y(x)$ die Steuer aus der Lohnsteuertabelle, die tatsächliche Regressionsfunktion wird wohl niedriger liegen.

Die **Grenzfunktion** $g(x) = \frac{dy}{dx}$ kann als der bei gegebenen Einkommen im Mittel gültige **Grenzsteuersatz** interpretiert werden. Dieser liegt i.A. höher als der **Steuersatz** \hat{y}/x selbst. Das heißt, die Elastizität $\frac{x}{\hat{y}} \frac{d\hat{y}}{dx} = \frac{x}{\hat{y}} g(x)$ ist größer als 1 oder progressiv, was leider einer **progressiven Besteuerung** entspricht :(



<http://home.t-online.de/home/parmentier.ffm/steuer.htm?steuer01.htm>