Stochastik für die Informatik, Vorlesung 17

Inhalt

- Maximum Likelihood Schätzung
- Lineare Regression
- Simulation von Zufallsvariablen

Lernziele

- Maximum Likelihood Schätzer berechnen können
- Die Grundprinzipien der linearen Regression kennen
- Simulationsmethoden kennen

Vorkenntnisse Differentialrechnung, Logarithmengesetze, lineare Gleichungssysteme, Zufallsvariablen, Verteilungsfunktion

Maximum Likelihood Schätzung (MLE)

Grundidee: Gegeben Daten $x_1, ..., x_n$ einer Verteilung mit einem unbekannten Parameter. Unter allen (theoretisch möglichen) Werten des Parameters ist der Maximum Likelihood-Schätzer derjenige, für welchen die Wahrscheinlichkeit, die gemessenen Werte $x_1, ..., x_n$ zu finden maximal wird.

Wir betrachten zwei verschiedene Situationen:

- Die zugrundeliegenden Zufallsvariablen $X_1, ..., X_n$ sind diskret, und haben Verteilung p_θ in Abhängigkeit von einem unbekannten Parameter θ
- Die zugrundeliegenden Zufallsvariablen $X_1, ..., X_n$ haben Dichte f_θ in Abhängigkeit von einem unbekannten Parameter θ.

Maximum Likelihood Schätzung (MLE)

Beispiel: Diskrete Verteilung

- Gegeben: $(x_1, ..., x_n)$
- Annahme: Dies sind Realisierungen von n unabhängigen identisch verteilten Zufallsvariablen $X_1, ..., X_n$ mit unbekanntem Parameter θ.
- Sei $p_\theta(x)$ die Verteilung der $X_1, ..., X_n$, mit unbekanntem θ.
- Betrachte

$$\mathbb{P}_\theta(X_1 = x_1, ..., X_n = x_n) = p_\theta(x_1) \cdot ... \cdot p_\theta(x_n).$$

Grundidee:

- Bestimme θ so, dass diese Wahrscheinlichkeit für das vorgegebene $x = (x_1, ..., x_n)$ maximal ist (z.B. duch Ableiten).
- (Beispiel 8.8)

Maximum Likelihood Schätzung (MLE)

Beispiel: Diskrete Verteilung

- Gegeben: $(x_1, ..., x_n)$
- Annahme: Dies sind Realisierungen von n unabhängigen identisch verteilten Zufallsvariablen $X_1, ..., X_n$ mit unbekanntem Parameter θ.
- Sei $p_\theta(x)$ die Verteilung der $X_1, ..., X_n$, mit unbekanntem θ.
- Betrachte

$$\mathbb{P}_\theta(X_1 = x_1, ..., X_n = x_n) = p_\theta(x_1) \cdot ... \cdot p_\theta(x_n).$$

Grundidee:

- Bestimme θ so, dass diese Wahrscheinlichkeit für das vorgegebene $x = (x_1, ..., x_n)$ maximal ist (z.B. duch Ableiten).
- (Beispiel 8.8)

Likelihood-Funktion

(Def. 8.9) Seien $x_1, ..., x_n \in \mathbb{R}$ Messwerte, welche als Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen $X_1, ..., X_n$ aufgefasst werden können. Die Likelihood-Funktion ist die Funktion $L((x_1, ..., x_n); \theta)$ von $\mathbb{R}^n \times \mathbb{R}$ nach $[0, 1]$, für welche gilt:

- Falls die X_i eine diskrete Verteilung p_θ mit Parameter θ besitzt, so ist

$$L((x_1, ..., x_n); \theta) = p_\theta(x_1) \cdot ... \cdot p_\theta(x_n) = \prod_{i=1}^{n} p_\theta(x_i).$$

- Falls die X_i eine Dichte f_θ mit Parameter θ besitzt, so ist

$$L((x_1, ..., x_n); \theta) = f_\theta(x_1) \cdot ... \cdot f_\theta(x_n) = \prod_{i=1}^{n} f_\theta(x_i).$$

Maximum Likelihood Schätzung (MLE)

(Def. 8.10) Seien $x_1, ..., x_n$ Messwerte, welche Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen sind, welche entweder der diskreten Verteilung p_θ folgen, oder die Dichte f_θ haben. Der Maximum Likelihood-Schätzer für θ ist $\theta_* = \theta_*(x_1, ..., x_n)$, welches die Likelihood-Funktion $L((x_1, ..., x_n); \theta)$ unter allen $\theta \in \mathbb{R}$ maximiert. Formal gesprochen ist

$$\theta_*(x_1, ..., x_n) = \text{argmax}_{\theta \in \mathbb{R}} L((x_1, ..., x_n); \theta).$$

- Der Maximum Likelihood-Schätzer wird dabei manchmal auch kurz als MLE bezeichnet, für das englische "maximum likelihood estimator".

Log-likelihood

(Def. 8.11) Die Log-Likelihood-Funktion ist definiert als

$$l((x_1, ..., x_n); \theta) = \ln L((x_1, ..., x_n); \theta).$$

(Satz 8.24) Die Funktion $l((x_1, ..., x_n); \theta)$ ist genau dann maximal (in θ), wenn $L((x_1, ..., x_n); \theta)$ maximal ist.

- $l(x_1, ..., x_n; \theta)$ ist oft einfacher zu maximieren als $L(x_1, ..., x_n; \theta)$.

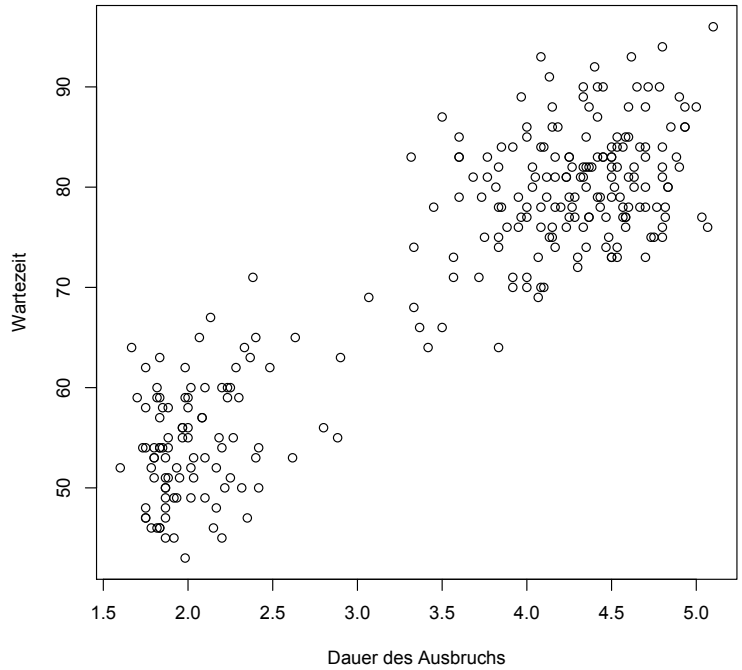
Maximum Likelihood Schätzung: Vorgehen

Das allgemeine Vorgehen zur Bestimmung des MLE wird somit:

1. Likelihood-Funktion $L(x_1, ..., x_n; \theta)$ aufstellen (nach Definition 10.8), abhängig davon ob die Verteilung diskret ist oder eine Dichte besitzt
2. Die log-Likelihood-Funktion $l(x_1, ..., x_n; \theta)$ durch logarithmieren von $L(x_1, ..., x_n; \theta)$ berechnen und so weit wie möglich vereinfachen (Logarithmengesetze!)
3. Den Parameter θ so bestimmen, dass $l((x_1, ..., x_n); \theta)$ für das vorgegebene $x = (x_1, ..., x_n)$ maximal ist. Meistens geschieht das durch Ableiten nach θ.
4. Überprüfen, dass es sich beim Ergebnis tatsächlich um ein Maximum handelt (z.B. durch Test mit der 2. Ableitung).

(Beispiel 8.9)

Beispiel 8.10: Korrelationen zwischen Daten



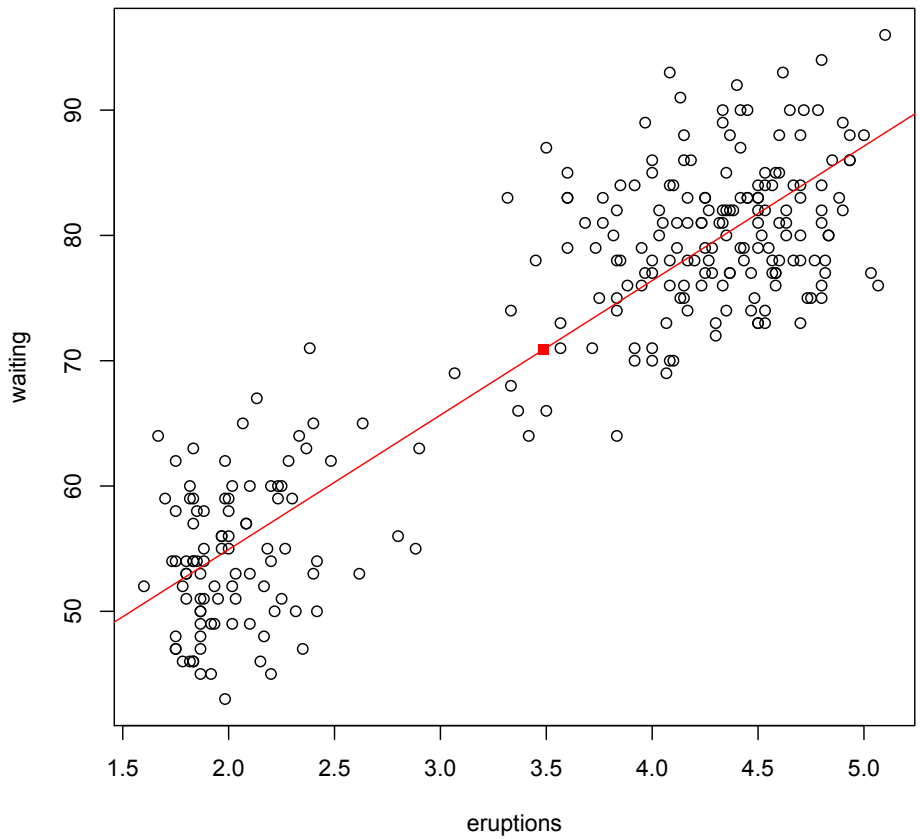
- Daten von zwei gleichzeitig gemessenen Größen: Paare von Messwerten $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$.
- Grundannahme: Realisierungen zweier Zufallsvariablen X und Y.

Maß für den Grad der Abhänigigkeit?

Linearer Zusammenhang

Beispiel Geysir: Daten $(x_i, y_i)_{i=1,\ldots,n}$ ergeben eine Punktwolke im $\mathbb{R}^2$.

Rotes Quadrat: Schwerpunkt $(\bar{\mu}_x, \bar{\mu}_y)$.



Korrelationen zwischen Daten

(Def. 8.12) Seien $x = (x_1, ..., x_n)$ und $y = (y_1, ..., y_n)$ gegeben. Die empirische Kovarianz ist definiert als

$$c_{xy} := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{\mu}_x)(y_i - \bar{\mu}_y),$$

wobei $\bar{\mu}_x$ das empirische Mittel der x_i und $\bar{\mu}_y$ das empirische Mittel der y_i ist.
Der empirische Korrelationskoeffizient ist definiert als

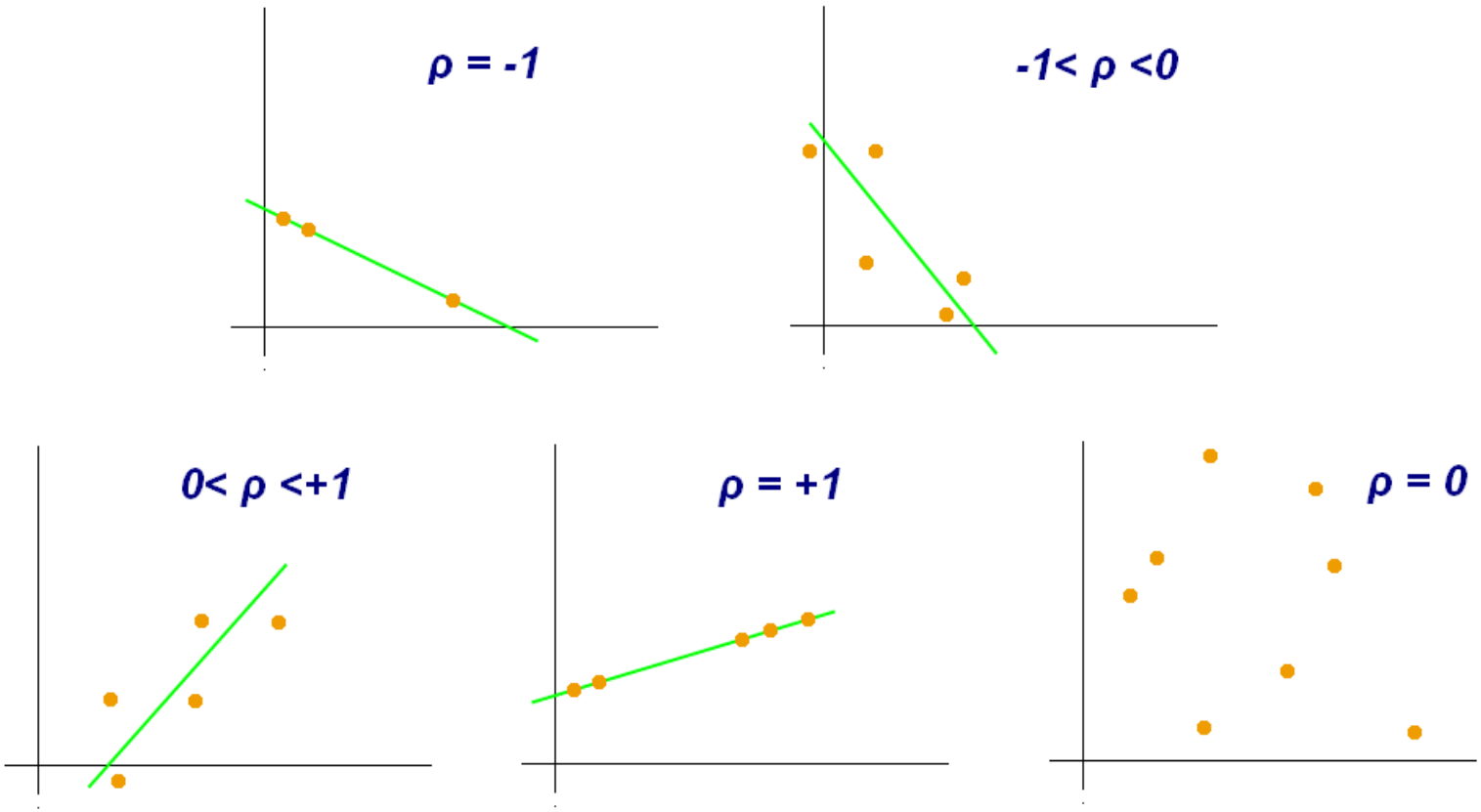
$$r_{xy} = \frac{c_{xy}}{\bar{\sigma}_x \bar{\sigma}_y},$$

wobei $\bar{\sigma}_x = \sqrt{\bar{\sigma}_x^2}$ die empirische Standardabweichung von x ist (und analog $\bar{\sigma}_y$ für y).

- $-1 \leq r_{x,y} \leq 1$,
- Beispiel 8.11: Im Beispieldatensatz:
  $c_{x,y}$=`cov(x,y)`=13.97781, $r_{x,y}$ =`cor(x,y)`=0.9008112.

Linearer Zusammenhang ($\rho = r$)

(Quelle: Wikipedia)



Lineare Regression

Grundidee: Finde die Gerade $y = ax + b$, welche durch den Punkt $(\bar{\mu}_x, \bar{\mu}_y)$ verläuft, und welche die Abstände der Punkte von der Geraden minimiert.

- Meistens: Summe der quadratischen Abstände in $y-$Richtungminimieren

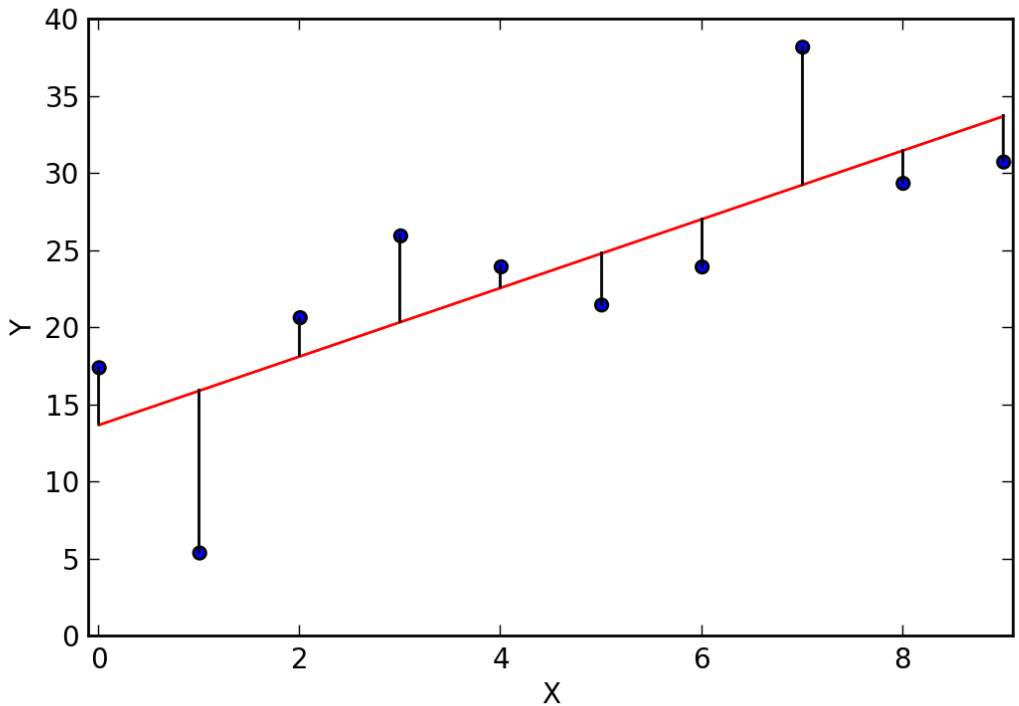


Bild: http://work.thaslwanter.at/

Lineare Regression

Gegeben: $x = (x_1, ..., x_n), y = (y_1, ..., y_n)$

Gesucht: Die Zahlen $a, b \in \mathbb{R}$, so dass die Gerade mit der Gleichung

$$y = ax + b$$

zwei Bedingungen erfüllt:

- $(\bar{\mu}_x, \bar{\mu}_y)$ liegt auf der Geraden, also $\bar{\mu}_y = a \cdot \bar{\mu}_x + b$,
- Der Ausdruck

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

ist minimal, wobei $\hat{y}_i := ax_i + b$ die $y-$Koordinaten des Punktes auf der Geraden ist, welcher $x-$Koordinate x_i hat.

Lineare Regression

(Satz 8.2) Die beiden Bedingungen auf der vorigen Folien legen a und b eindeutig fest, und zwar als

$$a = \frac{c_{xy}}{\bar{\sigma}_x^2}, \quad b = \bar{\mu}_y - a \cdot \bar{\mu}_x.$$

- Rollen von x und y sind nicht symmetrisch!
- (Beweis)

Beispiel 8.12: Lineare Regression

R-Befehl: `lm(y~ x)`

```
> reg<-lm(y x)
> summary(reg)
```

Ausgabe:

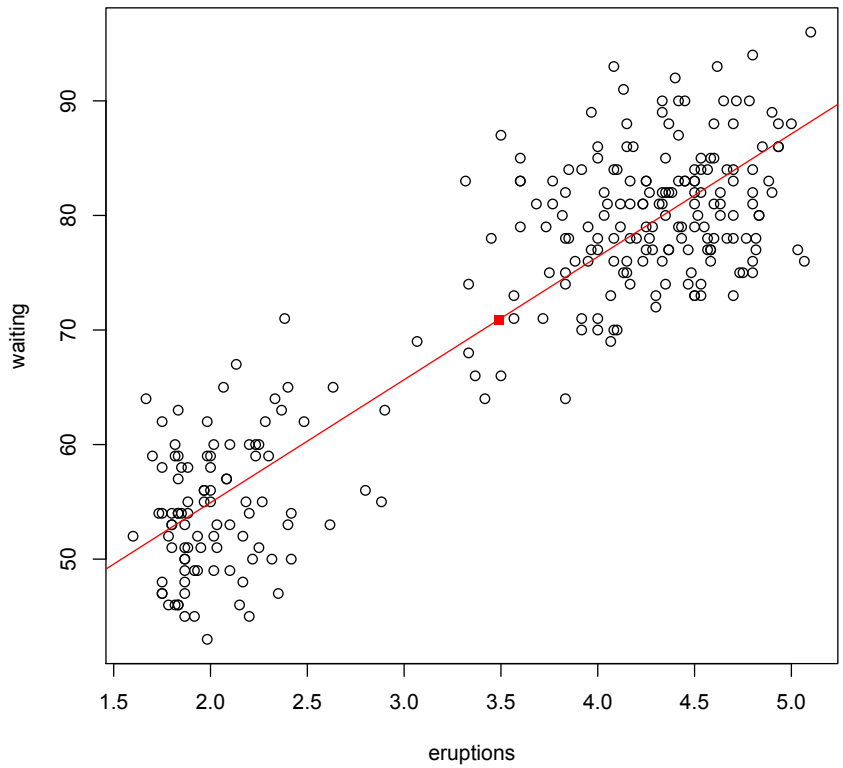
```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-12.0796  -4.4831   0.2122  | 3.9246  15.9719

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.4744     1.1549   28.98   <2e-16 ***
x            10.7296     0.3148   34.09   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.914 on 270 degrees of freedom
Multiple R-squared: 0.8115, Adjusted R-squared: 0.8108
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

Parameter: $a = 10.7296, b = 33.4744$

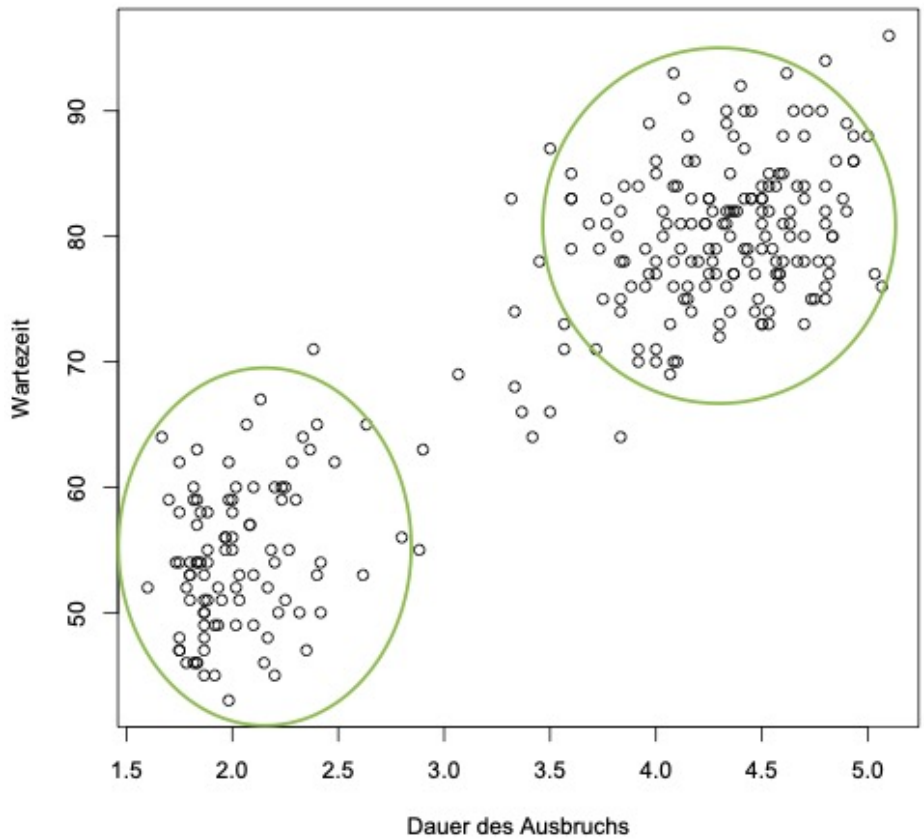
Andere Angaben: Informationen über Fehler/Güte des Fit

Beispiel 8.20

Plot:



Linearer Zusammenhang oder nicht?



Einteilung in zwei Cluster z.B. mittels EM-Algorithmus, basierend auf der Maximum-Likelihood-Schtzung. Statische Methode der Clusteranalyse.

Kapitel 15: Simulation von Zufallsvariablen

(Satz 15.1) Sei X uniform auf $[0,1]$ verteilt. Sei F eine invertierbare Verteilungsfunktion. Dann ist

$$Y := F^{-1}(X)$$

eine Zufallsvariable mit Verteilungsfunktion F.

Anwendung dieses Satzes: Simulation von Zufallsvariablen: Sobald man eine auf $[0,1]-$gleichverteilte Zufallsvariable simulieren kann, kann man im Prinzip jede Zufallsvariable mit bekannter (invertierbarer) Verteilungsfunktion simulieren (also z.B. Exponentialverteilung).

- Diskrete Verteilungen, Normalverteilung?
- und wie simuliert man eine $[0,1]-$gleichverteilte Zufallsvariable?

Bernoulli aus gleichverteilt

(Satz 15.1) Sei U eineauf $[0,1]$ gleichverteilte Zufallsvariable. Sei $p \in (0,1)$. Dann ist

$$X(U) := 1_{\{U \leq p\}}$$

eine Bernoulli-verteilte Zufallsvariable mit Parameter p.

- (Beweis)
- (15.1, 15.2)

Normal aus gleichverteilt

(Beispiel 15.3) Seien U und V unabhängige, auf $[0,1]$ gleichverteilte Zufallsvariablen. Dann sind

$$(X, Y) := \sqrt{-2 \log U}(\cos(2\pi V), \sin(2\pi V))$$

zwei unabhängige, standardnormalverteilte Zufallsvariablen.

- Box-Muller-Methode
- (ohne Beweis)
- Bem. allgemeine Normalverteilung

Pseudozufallszahlen

(Def. 15.1) Ein Pseudozufallszahlengenerator ist ein Algorithmus welcher ausgehend von einem Startwert eine deterministische Folge von Zahlen erzeugt, welche sich "wie eine Realisierung einer Folge von (uniform auf $[0, 1]$ verteilten) Zufallsvariablen" verhält.

- Überprüfung des zufälligen Verhaltens mit Hilfe von statistischen Tests, z.B. χ^2-Test auf uniforme Verteilung
- "Echte" Zufallszahlen: Mit Hilfe physikalischer Generatoren, z.B. durch Beobachtung radioaktiver Zerfälle, kosmisches Rauschen...

Pseudozufallszahlen

Beispiel 15.4: Der lineare Kongruenzgenerator

- Wähle Parameter $m \in \mathbb{N}, a \in \mathbb{Z} \setminus \{0\}, c \in \mathbb{Z}$.
- Wähle einen Startwert $x_0 \in \{0, ..., m\}$
- Definiere rekursiv

$$x_n := (ax_{n-1} + c) \bmod m$$

- Definiere für $n \in \mathbb{N}$

$$u_n := \frac{x_n}{m}$$

- Dann ist die Folge $(u_n)_{n \in \mathbb{N}_0}$ eine Folge von Pseudozufallszahlen (uniform auf $[0, 1]$).

Pseudozufallszahlen

Beispiel 15.4: Der lineare Kongruenzgenerator

Eigenschaften:

- Damit können höchstens m verschiedene Zahlen erzeugt werden
- Die erzeugte Folge ist periodisch
- Entsprechend sollte m möglichst groß gewählt werden, auf jeden Fall deutlich größer als die Anzahl zu erzeugender Zufallszahlen
- Wahl von a nach zahlentheoretischen Überlegungen
- Wahl des Startwerts kann große Auswirkung auf die erzeugte Folge haben, deshalb ist die Methode geeignet um Zufallszahlen zu erzeugen.
- Reproduzierbarkeit durch Wahl des Startwerts gegeben

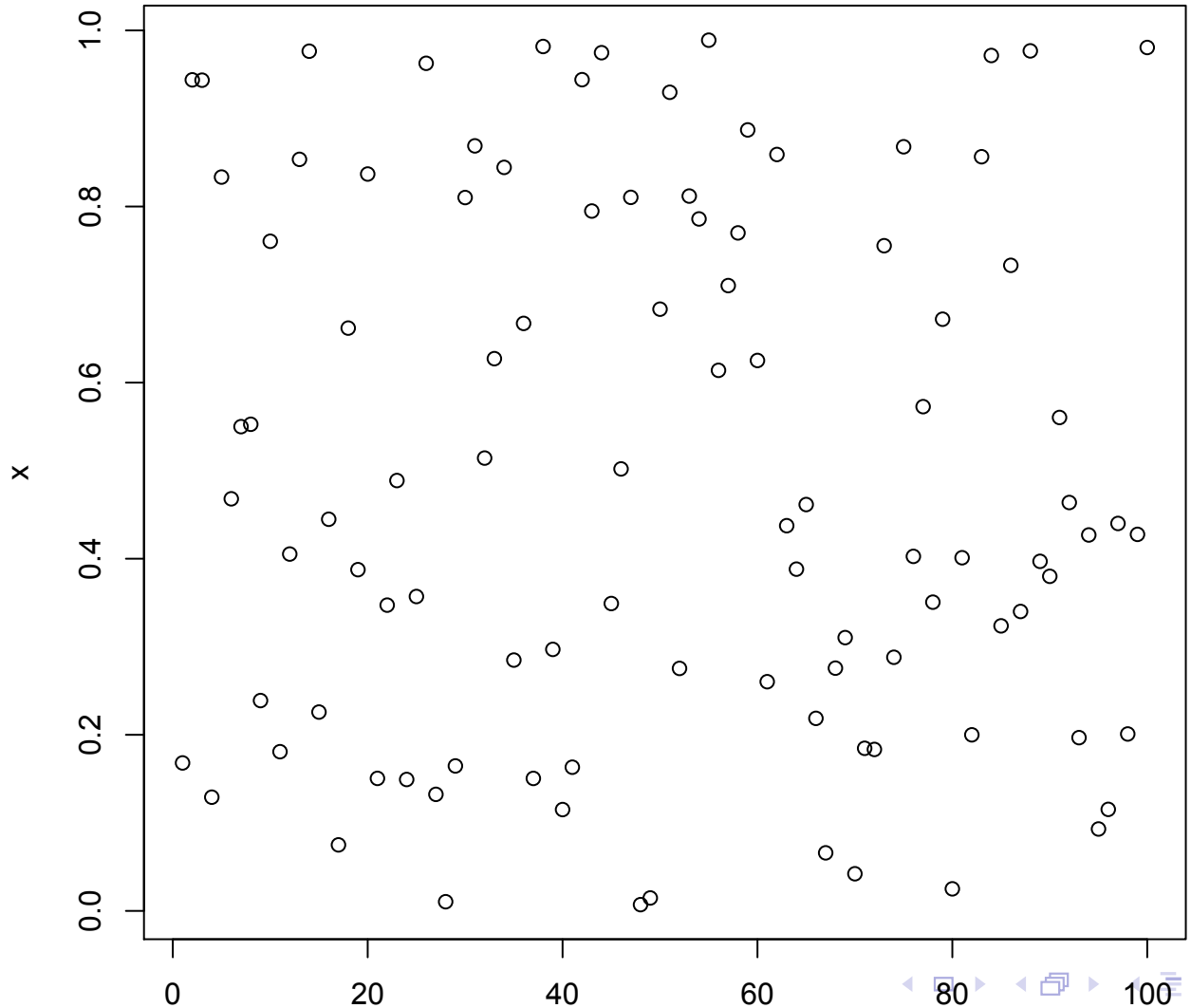
Pseudozufallszahlen

Das Software-Paket R bietet die Auswahl aus verschiedenen Pseudozufallszahlen-Generatoren, der Standard ist der sogenannte Mersenne-Twister (welcher ebenfalls ein rekursives Verfahren benutzt).

- Befehl: `runif`
- `runif(n)`: Vektor von n unabhängigen uniform $[0, 1]$ verteilten Zufallszahlen
- `runif(n, a, b)` : Vektor von n unabhängigen uniform $[a, b]$ verteilten Zufallszahlen
- Mit dem Befehl `set.seed(x)` kann der Startwert x_0 festgelegt werden.
- Wählt man zweimal denselben Startwert, so erhält man zweimal dieselbe Folge
- Simulation von anderen Verteilungen: `rnorm, rexp...`

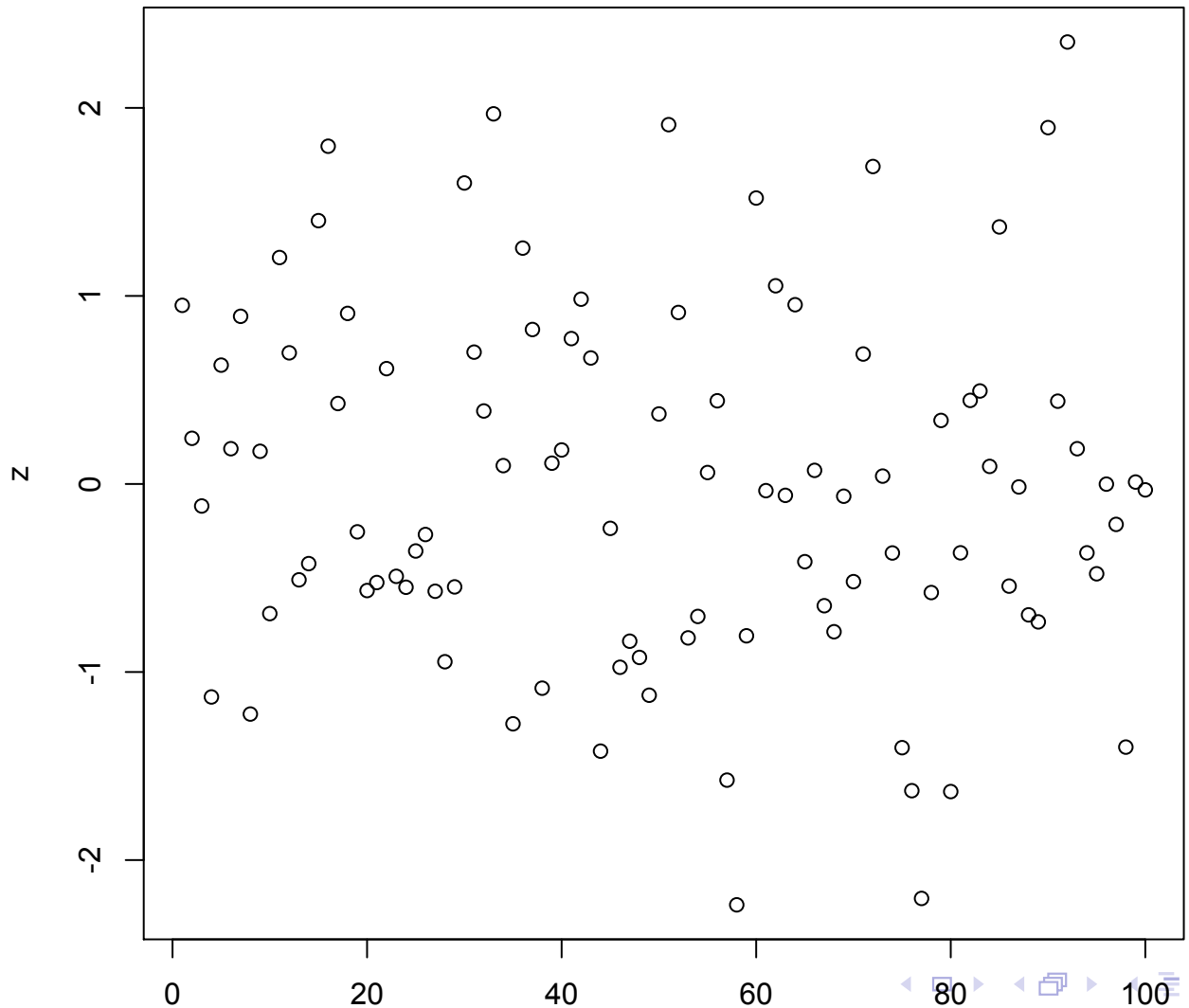
Pseudozufallszahlen

Beispiel: `x<-runif(100)`



Pseudozufallszahlen

Beispiel: `z<-rnorm(100)`



Pseudozufallszahlen

Beispiel: `z<-rnorm(1000)`