Vorlesung Statistik II

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Sommersemester 2026

Verfahren zur Parameterschätzung

Es gibt verschiedene Methoden, um Schätzfunktionen zu gewinnen

Wir werden einige häufige Verfahren besprechen, insbesondere

- Moment-Methode (method of moments)
- Methode der kleinsten Quadrate (ordinary least squares, OLS)
- Maximum-Likelihood-Schätzung (maximum likelihood estimation, MLE)
- Bayessche Schätzverfahren (Bayesian parameter estimation)

Moment-Methode

Die Moment-Methode ist ein Verfahren zur Schätzung von Populationsparametern, bei dem die unbekannten Momente (Erwartungswert, Varianz, ...) der Population mit den Stichprobenmomenten (Mittelwert, Stichprobenvarianz, ...) gleich gesetzt werden

Schätzungen für die Populationsparameter resultieren dann, wenn man die hierbei resultierenden Gleichungen nach den Parametern auflöst

Als Beispiele betrachten wir

- die Schätzung des Parameters p einer Bernoulli-Verteilung
- die Schätzung des Lernparameters α im Alles-Oder-Nichts-Modell

Schätzung des Parameters p einer Bernoulli-Verteilung

Wir betrachten eine einmalige Ziehung aus einer Urne, in der sich eine unbekannte Anzahl roter und schwarzer Kugeln befindet

Die Zufallsvariable X nehme den Wert 1 an, wenn die gezogene Kugel rot ist und 0 sonst

Die Zufallsvariable X ist dann Bernoulli-verteilt mit dem zunächst unbekannten Parameter p, der den Anteil der roten Kugeln in der Urne angibt

Wir erheben nun eine Stichprobe $X_1, \ldots, X_n$ vom Umfang n, d. h. wir führen eine n-fach wiederholte unabhängige Ziehung (mit Zurücklegen) durch

Schätzung des Parameters p einer Bernoulli-Verteilung

Nach der Moment-Methode bestimmen wir den Erwartungswert der Bernoulli-verteilten Zufallsvariablen X, der gegeben ist durch

$$E(X) = 1 \cdot p + 0 \cdot (1 - p) = p$$

und ersetzen ihn durch das entsprechende Stichprobenmoment, also durch das arithmetische Mittel der $X_1, \ldots, X_n$

Damit erhalten wir

$$\frac{1}{n} \sum_{i=1}^{n} X_i = \hat{p}$$

Die nach der Moment-Methode resultierende Schätzfunktion des Parameters p einer Bernoulli-Verteilung ist also die relative Häufigkeit der Ziehung einer roten Kugel

Schätzung des Lernparameters α im Alles-Oder-Nichts-Modell

Die Zufallsvariable T beschreibe die beim Lernen nach dem Alles-Oder-Nichts-Modell bei einem einzelnen Item auftretenden Fehler (s. Statistik I)

Für die erwartete Anzahl der Fehler $E(T)$ hatten wir berechnet

$$E(T) = \frac{1-g}{\alpha}$$

Dabei ist

- g die Ratewahrscheinlichkeit, die üblicherweise als bekannt vorausgesetzt wird (Kehrwert der Anzahl der Antwortalternativen)
- α der Lernparameter, der die Wahrscheinlichkeit angibt, dass in einem Durchgang tatsächlich gelernt wird

Schätzung des Lernparameters α im Alles-Oder-Nichts-Modell

In einem Experiment zum Paarassoziationslernen wird nun eine Liste von n Items in wiederholten Durchgängen dargeboten, wobei eine jeweils erfolgende Rückmeldung über die Korrektheit der gegebenen Antwort das Lernen der Items ermöglicht

Das Experiment wird beendet, wenn die gesamte Liste in zwei aufeinander folgenden Durchgängen fehlerfrei reproduziert wurde

Die Stichprobe $T_1, \ldots, T_n$ gibt dann für jedes dieser n Items die Anzahl der auftretenden Fehler an

Schätzung des Lernparameters α im Alles-Oder-Nichts-Modell

Im Rahmen der Moment-Methode ersetzen wir nun in der Gleichung

$$E(T) = \frac{1-g}{\alpha}$$

den Erwartungswert $E(T)$ durch das arithmetische Mittel $\bar{T} = 1/n \cdot \sum_{i=1}^{n} T_i$ und erhalten

$$\bar{T} = \frac{1-g}{\hat{\alpha}}$$

Auflösen nach $\hat{\alpha}$ liefert dann die Schätzfunktion

$$\hat{\alpha} = \frac{1-g}{\bar{T}}$$

Schätzung des Lernparameters α im Alles-Oder-Nichts-Modell

Wir betrachten im Folgenden einen fiktiven Datensatz mit $n = 10$ und $g = 0.5$ (zwei Antwortalternativen)

Items	Durchgang t										
	1	2	3	4	5	6	7	8	9	10	...
WAV-1	1	0	0	0	0	0	0	0	0	0	
CID-1	1	0	0	0	0	0	0	0	0	0	
KOL-2	0	0	0	0	0	0	0	0	0	0	
NAZ-2	0	1	0	0	0	0	0	0	0	0	
PED-2	0	0	0	1	1	1	0	0	0	0	
VEL-1	0	1	1	0	0	0	0	0	0	0	...
GUF-2	1	0	1	0	0	0	0	0	0	0	
ZEK-1	0	1	1	1	0	1	1	0	0	0	
XOS-1	0	1	1	1	0	0	1	1	0	0	
RIT-2	1	0	0	0	0	0	0	0	0	0	

Schätzung des Lernparameters α im Alles-Oder-Nichts-Modell

Die für $T_1, \ldots, T_n$ beobachteten Werte lauten

$$T_1 = 1, \quad T_2 = 1, \quad T_3 = 0, \quad T_4 = 1, \quad T_5 = 3,$$
$$T_6 = 2, \quad T_7 = 2, \quad T_8 = 5, \quad T_9 = 5, \quad T_{10} = 1$$

Mit im Durchschnitt $\bar{T} = 2.1$ Fehler pro Item erhält man als Schätzwert dann

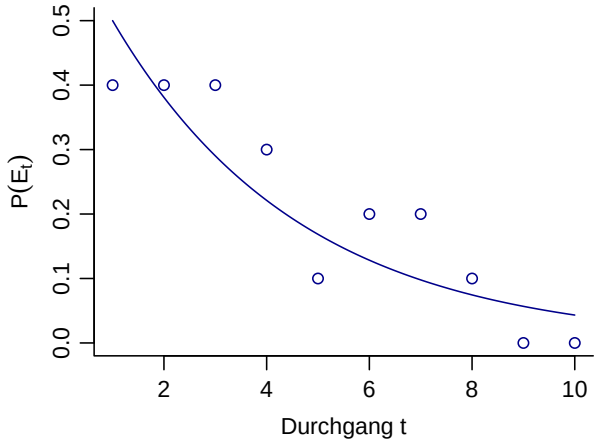
$$\hat{\alpha} = \frac{1 - g}{\bar{T}} = \frac{0.5}{2.1} = 0.238$$

Schätzung des Lernparameters α im Alles-Oder-Nichts-Modell

Mittlere Lernkurve

$$P(E_t) = (1 - g) \cdot (1 - \alpha)^{t-1}$$

mit $\hat{\alpha} = 0.238$ und relative Fehlerhäufigkeit



Eigenschaften der Moment-Methode

Vorteil der Moment-Methode ist die einfache Berechnung der Schätzfunktionen

- Beispielsweise kann die Ableitung von Maximum-Likelihood-Schätzern kompliziert werden und ist manchmal analytisch gar nicht möglich

Die Eigenschaften der Schätzfunktionen sind, sofern sie bekannt sind, weniger optimal als bei Maximum-Likelihood-Schätzern

- weniger (MSE-)effizient
- nicht notwendig suffizient

Methode der kleinsten Quadrate

Die Methode wurde von Carl Friedrich Gauß (1777-1855) im Jahre 1795 entwickelt



Wir hatten die Methode bereits bei der Bestimmung der Geradengleichung im Rahmen der Regressionsanalyse kennengelernt

Beispiel: Regressionsanalyse

Bei der Anwendung auf die Regressionsanalyse hat man Datenpunkte (x_i, y_i), $i = 1, \ldots, n$, deren Zusammenhang durch eine lineare Gleichung

$$\hat{y}_i = a + b\, x_i$$

vorhergesagt werden soll

Die Quadratsumme der Residuen

$$Q(a, b) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - b\, x_i)^2$$

wird dabei als Funktion der beiden Parameter a und b aufgefasst

Beispiel: Regressionsanalyse

Zur Bestimmung geeigneter Schätzwerte gilt es die Quadratsumme der Residuen $Q(a, b)$ zu minimieren

Bei einem Minimum dieser Funktion müssen die (partiellen) Ableitungen von $Q(a, b)$ nach a bzw. b verschwinden

Daher wird die Funktion $Q(a, b)$ partiell nach ihren Argumenten a und b abgeleitet

$$\frac{\partial Q(a, b)}{\partial a} \qquad \frac{\partial Q(a, b)}{\partial b}$$

Die resultierenden Ausdrücke werden dann gleich Null gesetzt, wodurch sich zwei Gleichungen in zwei Unbekannten ergeben, die es zu lösen gilt

Beispiel: Regressionsanalyse

Ausgehend von den Gleichungen

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

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

hatten wir folgende Schätzfunktionen abgeleitet

$$\hat{b} = \frac{s_{xy}}{s_x^2} \quad \text{und} \quad \hat{a} = \bar{y} - \hat{b}\,\bar{x}$$

Eigenschaften

Die auf diese Weise abgeleiteten Schätzfunktionen $\hat{a}$ und $\hat{b}$ haben viele wünschenswerte Eigenschaften

- Unter gewissen Voraussetzungen, die im Zuge der Besprechung der Regressionsanalyse noch genauer spezifiziert werden, kann man zeigen, dass die nach der Methode der kleinsten Quadrate ermittelten Schätzfunktionen optimal sind
- Sie sind *Best Linear Unbiased Estimators* (BLUE), d. h. sie sind
  – erwartungstreu
  – MSE-effizient
- Diese Eigenschaften folgen aus dem Satz von Gauß-Markow

Beispiel: Erwartungswert

Die Methode lässt sich allgemein zur Gewinnung von Schätzfunktionen einsetzen

Wir betrachten als Beispiel die Ableitung einer Schätzfunktion für den Erwartungswert $\mu = E(X)$

Für die Zufallsvariable X liege eine einfache Stichprobe $X_1, \ldots, X_n$ vom Umfang n vor

Den Ausgangspunkt der Ableitung bildet wiederum die Summe der quadratischen Abweichungen

$$Q(\mu) = \sum_{i=1}^{n} (X_i - \mu)^2,$$

die als Funktion des Parameters μ aufgefasst wird

Beispiel: Erwartungswert

Ableitung der Quadratsumme nach dem Parameter μ ergibt

$$\frac{dQ(\mu)}{d\mu} = \sum_{i=1}^{n}[2 \cdot (X_i - \mu) \cdot (-1)]$$

Nullsetzen liefert dann

$$\sum_{i=1}^{n}[2 \cdot (X_i - \hat{\mu}) \cdot (-1)] = 0$$

$$n \cdot \hat{\mu} - \sum_{i=1}^{n} X_i = 0$$

Beispiel: Erwartungswert

Auflösen nach dem Parameter schließlich ergibt die Schätzgleichung

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Damit ist, ausgehend von der Quadratsumme $Q(\mu)$, das arithmetische Mittel $\bar{X}$ die nach der Methode der kleinsten Quadrate abgeleitete Schätzfunktion für den Erwartungswert

Maximum-Likelihood-Schätzung

Das *Maximum-Likelihood-Prinzip* wurde durch R. A. Fisher vorgeschlagen und zeichnet sich gegenüber anderen Verfahren hinsichtlich vieler Aspekte aus

- Es ist sehr breit anwendbar, auch in komplexen Situationen
- Die resultierenden Schätzfunktionen haben viele wünschenswerte Eigenschaften

Im Folgenden werden wir das Maximum-Likelihood-Prinzip für den Fall der Schätzung eines allgemeinen Parameters ϑ betrachten

- Der Parameter ϑ kann dabei auch ein Vektor von Parametern sein (z. B. $\vartheta = (\mu, \sigma)$)

Maximum-Likelihood-Schätzung

Den Ausgangspunkt bildet als statistisches Modell der Population eine Zufallsvariable X, deren Verteilung von einem unbekannten Parameter (oder Parametervektor) ϑ abhängt

Es liege eine einfache Stichprobe $X_1, \ldots, X_n$ (unabhängig identisch verteilte Zufallsvariablen) vom Umfang n vor, mit Realisierungen $x_1, \ldots, x_n$

Hierzu betrachten wir dann

- im diskreten Fall die Wahrscheinlichkeit
- im stetigen Fall die Wahrscheinlichkeitsdichte

eben die vorliegenden Daten $x_1, \ldots, x_n$ zu erhalten, gegeben den Parameter ϑ

Beispiel: Bernoulli-Verteilung

Sei X Bernoulli-verteilt mit dem Parameter p

Dann betrachten wir die Wahrscheinlichkeit

$$\begin{aligned} f(x_1, \ldots, x_n \mid p) &= P(x_1, \ldots, x_n \mid p) \\ &= p^{x_1} \cdot (1-p)^{1-x_1} \cdot \ldots \cdot p^{x_n} \cdot (1-p)^{1-x_n} \\ &= p^{x_1+\ldots+x_n} \cdot (1-p)^{n-(x_1+\ldots+x_n)} \end{aligned}$$

die Daten $x_1, \ldots, x_n$ zu erhalten, gegeben der zugrunde liegende Parameter ist p

Beispiel: Normalverteilung

Sei X normalverteilt mit den Parametern μ und σ^2

Dann betrachten wir die Wahrscheinlichkeitsdichte

$$f(x_1, \ldots, x_n \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x_1-\mu)^2}{2\sigma^2}\right) \cdot \ldots \cdot \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right)$$

die Daten $x_1, \ldots, x_n$ zu erhalten, gegeben die zugrunde liegenden Parameter sind μ und σ^2

Likelihoodfunktion

Allgemein betrachten wir also eine Funktion

$$f(x_1, \ldots, x_n \mid \vartheta)$$

Diese Funktion f gibt die Wahrscheinlichkeitsdichte (bzw. im diskreten Fall die Wahrscheinlichkeit) an, mit der wir die beobachteten Daten erhalten $x_1, \ldots, x_n$, gegeben der zugrunde liegende Parameter ist ϑ

Da die Daten $x_1, \ldots, x_n$ ja gegeben und daher fest sind, können wir f als Funktion mit dem Argument ϑ ansehen

Wir betrachten also eine Funktion

$$L(\vartheta) = f(x_1, \ldots, x_n \mid \vartheta)$$

Beispiel: Bernoulli-Verteilung

Für die Funktion L gilt allgemein

$$L(p) = p^{x_1+\ldots+x_n} \cdot (1-p)^{n-(x_1+\ldots+x_n)}$$

Mit dem Stichprobenumfang $n = 5$ und

$$x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1, x_5 = 0$$

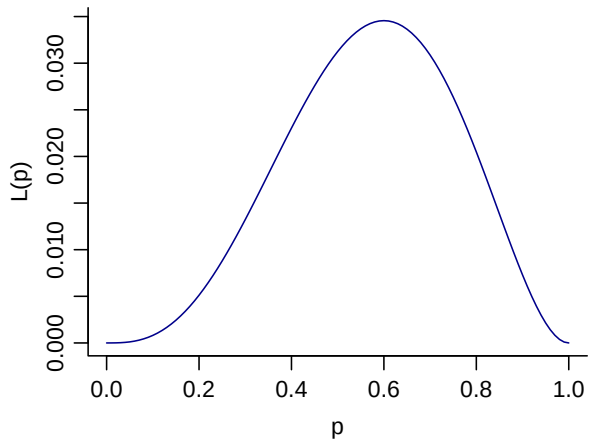
erhält man dann

$$L(p) = p^{1+0+1+1+0} \cdot (1-p)^{5-(1+0+1+1+0)} = p^3 \cdot (1-p)^2$$

Beispiel: Bernoulli-Verteilung

Grafik der Funktion $L(p)$ für die gegebenen Daten

```
curve(p^3*(1 - p)^2,
      xname = "p",
      from = 0, to = 1,
      ylab = "L(p)")
```



Beispiel: Bernoulli-Verteilung

Es zeigt sich, dass die Funktion $L(p)$ in Abhängigkeit vom Parameter p unterschiedliche Werte annimmt

Die Wahrscheinlichkeit, die Daten

$$x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1, x_5 = 0$$

zu beobachten, unterscheidet sich also deutlich, je nach dem Parameterwert p, der dem statistischen Modell der Population zugrunde gelegt wird

- Man beachte, dass das Ergebnis nicht von der beobachteten Reihenfolge der Datenwerte abhängt

Es ist sinnvoll den unbekannten Parameter so zu wählen, dass die Wahrscheinlichkeit der beobachteten Daten gegeben den Parameterwert möglichst groß wird

Maximum-Likelihood-Prinzip

Die Maximierung der "Wahrscheinlichkeit der Daten gegeben den Parameterwert" ist der Inhalt des Maximum-Likelihood-Prinzips

- Unter den eingangs beschriebenen Bedingungen betrachtet man die Wahrscheinlichkeitsdichte (bzw. die Wahrscheinlichkeitsfunktion im diskreten Fall)

$$f(x_1, \ldots, x_n \mid \vartheta)$$

- Für gegebene Daten $x_1, \ldots, x_n$ bezeichnet man jede zu $f(x_1, \ldots, x_n \mid \vartheta)$ proportionale Funktion

$$L(\vartheta) = c \cdot f(x_1, \ldots, x_n \mid \vartheta)$$

  als *Likelihood* (mit einer beliebigen Konstante $c > 0$)

- Die Likelihood wird als Funktion des Parameters ϑ betrachtet

Maximum-Likelihood-Prinzip

Wegen der stochastischen Unabhängigkeit der zu den Realisierungen $x_1, \ldots, x_n$ gehörigen Zufallsvariablen $X_1, \ldots, X_n$ hat man

$$f(x_1, \ldots, x_n \mid \vartheta) = f(x_1 \mid \vartheta) \cdot \ldots \cdot f(x_n \mid \vartheta) = \prod_{i=1}^{n} f(x_i \mid \vartheta)$$

Daher gilt für die Likelihood

$$L(\vartheta) \propto \prod_{i=1}^{n} f(x_i \mid \vartheta),$$

wobei $\propto$ die oben angegebene Proportionalität kennzeichnet

Maximum-Likelihood-Prinzip

Das Maximum-Likelihood-Prinzip besagt dann

- Wähle für gegebene Daten $x_1, \ldots, x_n$ als Schätzung denjenigen Parameterwert $\hat{\vartheta}$, für den die Likelihood maximal ist, d. h.

$$L(\hat{\vartheta}) = \max_{\vartheta} L(\vartheta)$$

Dies bedeutet, dass

$$f(x_1, \ldots, x_n \mid \hat{\vartheta}) = \max_{\vartheta} f(x_1, \ldots, x_n \mid \vartheta)$$

Maximum-Likelihood-Schätzung

Es ist also derjenige Wert $\hat{\vartheta}$ (Maximum-Likelihood-Schätzer; maximum likelihood estimator, MLE) zu finden, für den die Likelihood $L(\vartheta)$ ein Maximum annimmt

Hierzu sind prinzipiell folgende Schritte zu absolvieren

- Ableitung der Likelihood $L(\vartheta)$ nach dem Parameter ϑ
- Nullsetzen der Ableitung
- Auflösen der resultierenden Gleichung nach dem Schätzer $\hat{\vartheta}$

Die Bildung der Ableitung der Likelihood wird dadurch erschwert, dass sie ein Produkt darstellt (erfordert "Produktregel" der Ableitung)

Log-Likelihood

Zur rechentechnischen Vereinfachung wird daher zunächst der (natürliche) Logarithmus der Likelihood gebildet, man betrachtet also die *Log-Likelihood*, $\log L(\vartheta)$

- Bei Bildung der Log-Likelihood geht die Produktdarstellung in eine Summe über, d. h. aus

$$L(\vartheta) = c \cdot \prod_{i=1}^{n} f(x_i \mid \vartheta)$$

wird

$$\log L(\vartheta) = \log c + \sum_{i=1}^{n} \log f(x_i \mid \vartheta)$$

- In dieser Summe kann jeder Summand getrennt abgeleitet werden (wobei die Konstante $\log c$ wegfällt!)

Log-Likelihood

Die Bildung der Log-Likelihood ist zulässig, da der Logarithmus eine streng monoton steigende Transformation ist, d. h. es gilt

$$L(\vartheta_1) < L(\vartheta_2) \quad \text{impliziert} \quad \log L(\vartheta_1) < \log L(\vartheta_2)$$

Der größte Wert der Likelihood ist auch der größte Wert der Log-Likelihood, d. h. wenn

$$L(\hat{\vartheta}) = \max_{\vartheta} L(\vartheta),$$

dann ist auch

$$\log L(\hat{\vartheta}) = \max_{\vartheta} \log L(\vartheta)$$

Log-Likelihood

Die Bestimmung des Maximum-Likelihood-Schätzers $\hat{\vartheta}$ erfolgt also durch

- Bildung der Log-Likelihood

$$\log L(\vartheta)$$

- Ableitung der Log-Likelihood nach dem Parameter ϑ

$$(\log L)'(\vartheta) = \frac{d \log L(\vartheta)}{d\vartheta}$$

- Nullsetzen der Ableitung

$$(\log L)'(\hat{\vartheta}) = 0$$

- Auflösen der resultierenden Gleichung nach dem Schätzer $\hat{\vartheta}$