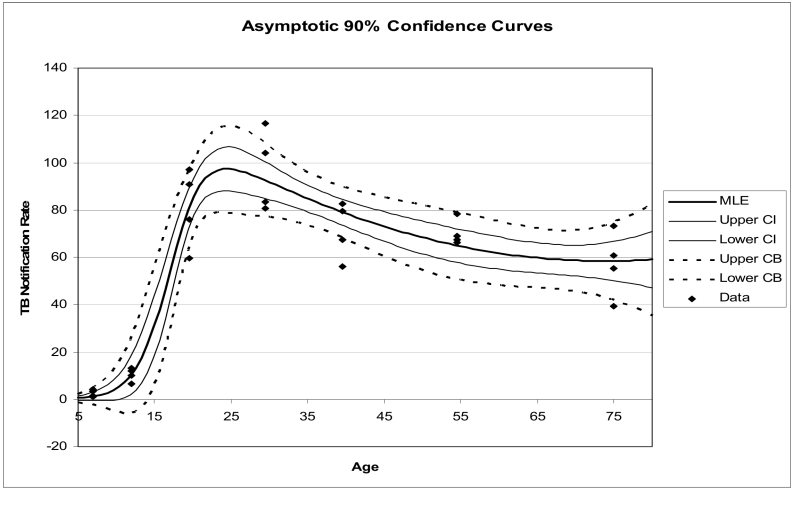
Bootstrapping Ansätze zur Bestimmung von Konfidenzbändern für Verteilungsfunktionen

Dennis Richter

4. Januar 2021

Lehrstuhl IV
Informatik

Motivation und Problemstellung



Motivation und Problemstellung

- Konfidenzintervalle sind visuelle Hilfsmittel zur Interpretation der Genauigkeit einzelner Schätzwerte
- Konfidenzbänder werden für simultane Schätzungen benötigt
- Es gibt verschiedene Ansätze zur Bestimmung von Konfidenzbändern (analytische Verfahren, aber auch Bootstrap-Verfahren)
- In der Literatur gibt es viele theoretische Diskussionen über die Verwendung von Bootstrap zur Bestimmung von Konfidenzbändern, aber wenige Implementierungen und empirische Ergebnisse

Grundlagen

- Regressionsfunktion $\eta(x, \theta)$:
  gesucht ist θ_0 mit $y_j = \eta(x_j, \theta_0) + \epsilon_j, j = 1, 2..., n$ und $\epsilon \sim N(0, \sigma^2)$
- Schätzer $\hat{y}(x) = \eta(x, \hat{\theta})$ für den "wahren Wert" $E(y|x) = \eta(x, \theta_0)$
- Konfidenzbereich für den Schätzer:
  $P\left(\theta_L \leq \theta_0 \leq \theta_U\right) \geq 1 - \alpha$
  z.B. $\theta_L, \theta_U = \hat{\theta} \mp z_{\alpha/2}\sqrt{\mathbf{V}(\hat{\theta})}$
- Konfidenzintervall für einen Punkt der Regressionsfunktion:
  $\forall x : P\left(y_L(x) \leq \eta(x, \theta_0) \leq y_U(x)\right) \geq 1 - \alpha$
  z.B. $y_L(x), y_U(x) = \eta(x, \hat{\theta}) \mp z_{\alpha/2}\sqrt{\left(\frac{\partial\eta(x,\theta)}{\partial\theta}\right)^T_{\hat{\theta}} \mathbf{V}(\hat{\theta}) \left(\frac{\partial\eta(x,\theta)}{\partial\theta}\right)_{\hat{\theta}}}$
- Konfidenzband für die Regressionsfunktion:
  $P\left(\forall x : y_L(x) \leq \eta(x, \theta_0) \leq y_U(x)\right) \geq 1 - \alpha$
  z.B. $y_L(x), y_U(x) = \eta(x, \hat{\theta}) \mp \sqrt{\chi^2_p(a) \left(\frac{\partial\eta(x,\theta)}{\partial\theta}\right)^T_{\hat{\theta}} \mathbf{V}(\hat{\theta}) \left(\frac{\partial\eta(x,\theta)}{\partial\theta}\right)_{\hat{\theta}}}$

Grundlagen

Basic-Sampling-Methode:

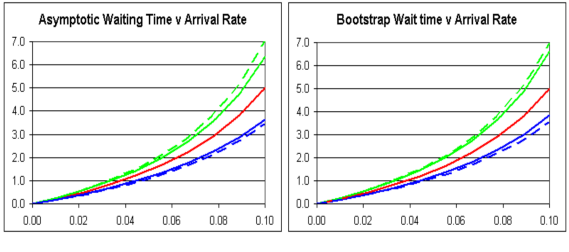
for $j = 0$ to B **do**
  for $i = 0$ to n **do**
    Draw sample y_{ij} from $F(.)$
  end for
  Calculate statistic $s_j = s(y_j)$
end for
Form the EDF $G_n(.|s)$

Bootstrap-Methode:

Require: Random sample $y = (y_1, y_2, ...y_n)$ from $F(.)$
Form the EDF $F_n(.|y)$
for $j = 0$ to B **do**
  for $i = 0$ to n **do**
    Draw sample y^*_{ij} from $F_n(.|y)$
  end for
  Calculate statistic $s^*_j = s(y^*_j)$
end for
Form the EDF $G_n(.|s*)$

Verwandte Arbeiten

- Cheng, Russell. (2005). Bootstrapping simultaneous confidence bands. 8 pp.-. 10.1109/WSC.2005.1574257.
- Cheng, Russell. (2015). Bootstrap confidence bands and goodness-of-fit tests in simulation input/output modelling. 16-30. 10.1109/WSC.2015.7408150.



Weitere:

- Govind, Nirmal & Roeder, Theresa. (2006). Estimating Expected Completion Times with Probabilistic Job Routing. 1804-1810. 10.1109/WSC.2006.322958.
- Wang, Xing & Wang, Xin & Sun, Zhaonan. (2009). Comparison on Confidence Bands of Decision Boundary between SVM and Logistic Regression. 272-277. 10.1109/NCM.2009.281.

Lösungsansatze

2 Ansätze werden vorgestellt, bei denen Bootstrap zur Vereinfachung der analytischen Verfahren verwendet wird

- Parametric Bootstrap:
  - setzt voraus, dass $\hat{\theta}$ als normalverteilt angenommen werden kann, d.h. $\hat{\theta} \sim N(\theta_0, \mathbf{V}(\theta_0))$
  - verzichtet jedoch auf die lineare Approximation von $\eta(x, \theta)$ durch die Delta-Methode
- Non-Parametric Bootstrap:
  - keine Verteilungsannahme über $\hat{\theta}$
  - und auch keine lineare Approximation von $\eta(x, \theta)$
  - rechenintensiv wegen verschachteltem Doppel-Bootstrap

Es gibt jedoch auch andere analytische Verfahren zur Bestimmung von Konfindenzbändern, bei denen Bootstrap verwendet werden kann

Anwendnungs Beispiel

Av. Age	Year	1980	1986	1993	2000
2	0-4	1.26	2.78	0.63	0.34
7	5-9	3.53	4.10	1.31	0.91
12	10-14	11.98	13.14	9.86	6.53
19.5	15-24	90.82	97.12	75.85	59.46
29.5	25-34	83.45	116.62	104.00	80.85
39.5	35-44	55.98	67.28	79.33	82.66
54.5	45-64	66.32	78.53	69.10	67.27
75	65+	39.42	55.35	60.76	73.20

Abbildung: Morocco Pulmonary TB notifications per 100,000

Parameter	MLE
θ_1	10.17
θ_2	153.79
θ_3	17.26
θ_4	-2.58
θ_5	0.455
θ_6	0.01746

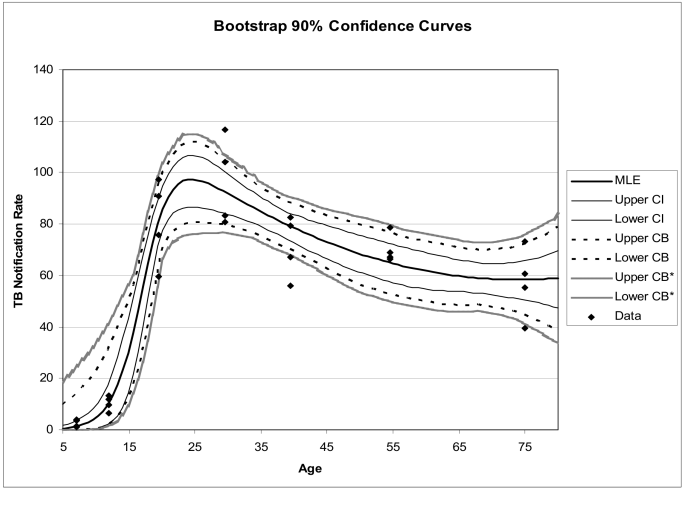
Abbildung: MLE's for the Morocco TB Model

Als Modell wurde gewählt:

$$y_j = (\theta_2 + \theta_4 x_j + \theta_6 x_j^2)\frac{exp(\theta_5(x_j - \theta_3))}{1 + exp(\theta_5(x_j - \theta_3))} + \epsilon_j$$

wobei $\epsilon_j \sim N(0, \theta_1^2)$

Anwendnungs Beispiel



Umsetzung

- 2 Wochen: Vertiefende Recherche der Bootstrap-Ansätzen
- 1 Woche: Recherche zu Parameterstudien, Auswertung und Darstellung in Kontext von OMNeT++
- 1 Woche: Erstellung von mindestens 2 einfachen Beispielen in Form einer OMNeT++ Simulation
- 2 Wochen: Implementierung der Verfahren in C++ (mindestens die von Cheng (2005) vorgestellten Ansätze)
- 2 Wochen: Anwendung der Verfahren und empirische Bewertung

Zudem:

- Die implementierung soll wiederverwendbar sein (über Git) und nach Möglichkeit in die IDE integriert
- Zwischen den Arbeitspaketen sind jeweils Schreibphasen geplant

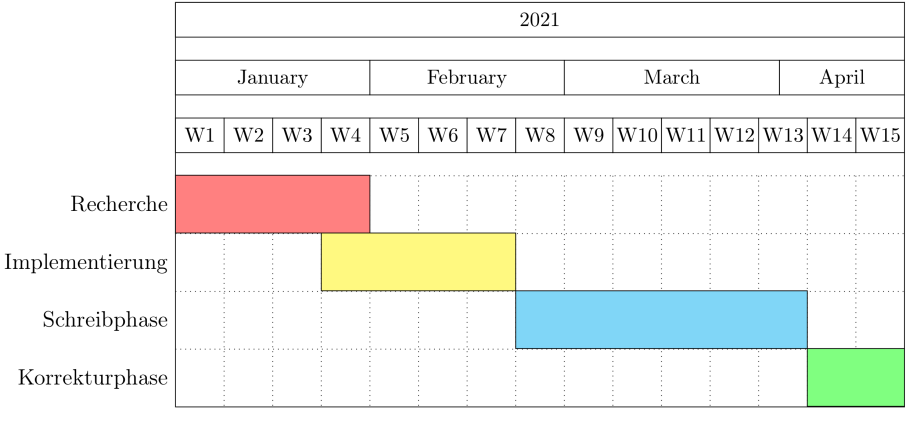
Geplante Evaluation

Die Verfahren werden anhand der zuvor erstellten Beispiele getestet:

- Datenerhebung durch die Simulation
- Bestimmung eines statistischen Modells und der Regressionsgerade (Modellanpassung)
- Auswertung mit analytischen Verfahren (durch C++-Libraries oder in R)
- Auswertung mit den implementierten Bootstrap-Verfahren
- Vergleich der Konfidenzbänder

Es wird erwartet, dass die Bootstrap-Ansätze ähnliche Konfidenzbänder liefern wie die analytischen Methoden und sich diesen annähern

Zeitplan


2021
January
February
March
April
W1
W2
W3
W4
W5
W6
W7
W8
W9
W10
W11
W12
W13
W14
W15
Recherche
Implementierung
Schreibphase
Korrekturphase