



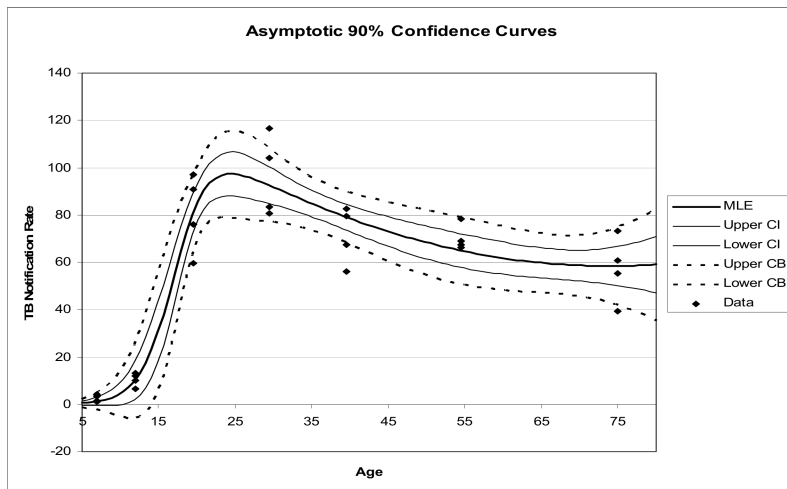
Bootstrapping Ansätze zur Bestimmung von Konfidenzbändern für Verteilungsfunktionen

Dennis Richter

4. Januar 2021

Lehrstuhl IV
Informatik

Motivation und Problemstellung



Motivation und Problemstellung

- Konfidenzintervalle sind visuelle Hilfestellung zum interpretieren der Daten
- die Intervalle gelten jeweils immer nur für einen Messwert
- manchmal möchte man solche Schätzungen für die ganze Regressionsfunktion haben, dann benötigt man Konfidenzbänder
- Es gibt viele Verschiedene Methoden zu Bestimmung von Konfidenzbändern
- neben analytischen Verfahren, kann auch Bootstrap eingesetzt werden -> einfacher, aber ähnlich gut
- zwar theoretische ansätze in der Literatur aber wenig tatsächliche Implementierungen und damit Empirischer Vergleich zu herkömmlichen Methoden
- (Aufgabe ist also vorstellung der Methoden, umsetzung in Omnet und empirischer Vergleich zu anderen CBs)

Grundlagen

- Regressionsfunktion $\eta(x, \theta)$:
gesucht ist θ_0 mit $y_j = \eta(x_j, \theta_0) + \epsilon_j, j = 1, 2, \dots, n$ 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:

$$P(\theta_L \leq \theta_0 \leq \theta_U) \geq 1 - \alpha$$
 z.B. $\theta_L, \theta_U = \hat{\theta} \mp z_{\alpha/2} \sqrt{\mathbf{V}(\hat{\theta})}$
- Konfidenzintervall für einen Punkt der Regressionsfunktion:

$$\forall x : P(y_L(x) \leq \eta(x, \theta_0) \leq y_U(x)) \geq 1 - \alpha$$
 z.B. $y_L(x), y_U(x) = \eta(x, \hat{\theta}) \mp z_{\alpha/2} \sqrt{\left(\frac{\partial \eta(x, \theta)}{\partial \theta}\right)_{\hat{\theta}}^T \mathbf{V}(\hat{\theta}) \left(\frac{\partial \eta(x, \theta)}{\partial \theta}\right)_{\hat{\theta}}}$
- Konfidenzband für die Regressionsfunktion:

$$P(\forall x : y_L(x) \leq \eta(x, \theta_0) \leq y_U(x)) \geq 1 - \alpha$$
 z.B. $y_L(x), y_U(x) = \eta(x, \hat{\theta}) \mp \sqrt{\chi_p^2(a) \left(\frac{\partial \eta(x, \theta)}{\partial \theta}\right)_{\hat{\theta}}^T \mathbf{V}(\hat{\theta}) \left(\frac{\partial \eta(x, \theta)}{\partial \theta}\right)_{\hat{\theta}}}$

Grundlagen

Basic-Sampling-Methode:

```

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

```

Bootstrap-Methode:

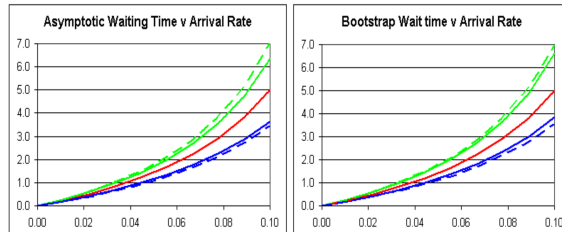
```

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

```

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.

Lösungsansätze

In den Papern werden 2 Lösungsansätze vorgestellt, die Bootstrap zur Vereinfachung der erwähnten analytischen Verfahren einsetzen

- Parametric Bootstrap:

- setzt voraus, dass $\hat{\theta}$ als normalverteilt angenommen werden kann, also $\hat{\theta} \sim N(\theta_0, \mathbf{V}(\theta_0))$
- verzichtet allerdings auf lineare Approximation von $\eta(\mathbf{x}, \theta)$ durch die Delta-Methode

- Non-Parametric Bootstrap:

- keine Verteilungsannahme über $\hat{\theta}$
- und auch keine lineare Approximation von $\eta(\mathbf{x}, \theta)$
- rechenintensiv, da geschachteltes Bootstrap

Es gibt allerdings auch andere analytische Verfahren zur Bestimmung von Konfidenzbändern, bei denen Bootstrap zum Einsatz kommen kann.

Anwendungs 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

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

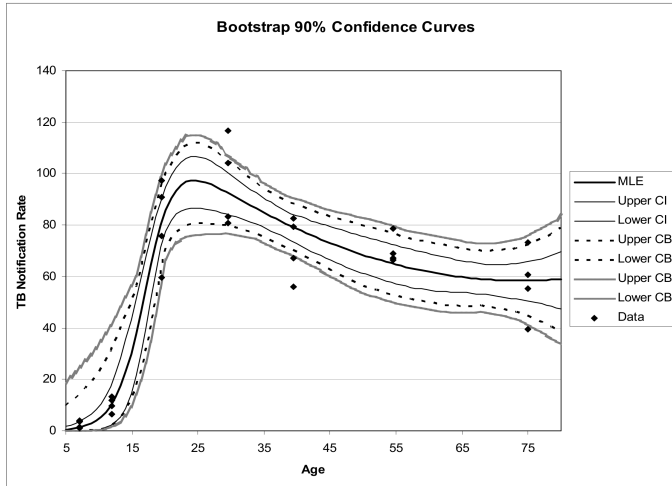
Abbildung: Morocco Pulmonary TB notifications per 100,000

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

Anwendungs Beispiel



Umsetzung

- genauere Recherche zu Konfidenzbändern und Bootstrapping-Ansätze zur Bestimmung von Konfidenzbändern (mehr dimensionale Parameter)
- Recherche zu Parameterstudien, Auswertung und Darstellung in Kontext von OMNeT++
- Erstellung einfacher Beispiele als Grundlage (z.B. für die Evaluation)
- Implementierung der Verfahren in C++ und Integration in die OMNeT++ IDE
- Anwendung der Verfahren anhand der erstellten Beispiele

Geplante Evaluation



Zeitplan

