Finding the right distribution for highly skewed zero-inflated clinical data

Cordula Eggerth (00750881)
Viktoria Kittler (01606264)

UK Biostatistik
Dr. Robin Ristl & Dr. Andreas Baierl
WS 2018/19

Agenda

1 Problemstellung

2 Zugrundeliegende Daten & klinische Variablen

3 Analysephasen & Resultate

4 Fragen/kritische Anmerkungen zum Paper

5 Conclusio

Agenda

Probleme bei Studien

- **Ausgangslage: diskrete, sehr schiefe („highly skewed“) Verteilungen mit nur nicht-negativen Werten & großer Anzahl von Nullen in der abhängigen Variable**
  - Hier: klinische Studie über Kinder mit elektrophysiologischen Störungen („EP disorder“), wobei die Mehrheit dieser Kinder ohne operative Eingriffe behandelt wurde
- **Problem: verzerrte Schätzungen und irreführende Inferenz aus den Daten, falls die Verteilung (v.a. die große Anzahl von Nullen) nicht angemessen in der statistischen Modellierung berücksichtigt wird**
  - z.B. Berücksichtigung der großen Anzahl von Nullen („Zero Inflation“) mittels jeweiliger Zero-Inflated Distribution

Umgang mit solchen Situationen

- **Ziel**:
  - Optimales Modell für derartige Daten finden und Zero-Inflated Modelle als Alternative in Betracht ziehen
- **Vorgangsweise**:
  - Simulation von Zero-Inflated, Skewed Count Daten
  - Fitten mit Poissonmodell, Negativem Binomialmodell, Zero-Inflated Poissonmodell (ZIP) und Zero-Inflated Negative Binomialmodell (ZINB)
  - Anwendung der Modelle auf die erwähnten klinischen Daten über Kinder mit EP Disorder
  - Wahl des optimalen Modells (hier: ZIP Modell) basierend auf den Daten

Agenda

1 Problemstellung

2 Zugrundeliegende Daten & klinische Variablen

3 Analysephasen & Resultate

4 Fragen/kritische Anmerkungen zum Paper

5 Conclusio

Studienpopulation

Studie mit 286 Kindern:

- Im Alter von 8 bis 18 Jahren
- Zum Zeitpunkt ihres ambulanten Kardiologiebesuchs
- Diagnose: EP (Electropyhsiological) Disorder (bzgl. Herzrhythmus)
- Katheter-Intervention / Herzschrittmacher
- 147 Männer und 139 Frauen
- Genehmigung der Studie durch Cincinnati Children‘s Hospital Medical Center

Klinische Variablen

4 klinische Variablen zur Modellierung der Count-Daten:

- **Herzbedingter Krankenhausaufenthalt:**
  - Binäre Variable: Besuche pro Jahr – max. 2 oder mehr als 2
- **Herzblock (= deutlich verlangsamter Herzschlag)**
  - Binäre Variable: Vorhandensein eines Herzblocks – Ja oder Nein
- **Frühgeburt**
  - Binäre Variable: vor 37. Schwangerschaftswoche geboren – Ja oder Nein
- **Zeit seit letztem Krankenhausaufenthalt**
  - Stetige Variable: Monate seit dem letzten KH-Aufenthalt

Agenda

1 Problemstellung

2 Zugrundeliegende Daten & klinische Variablen

3 Analysephasen & Resultate

4 Fragen/kritische Anmerkungen zum Paper

5 Conclusio

Überblick über die statistische Analyse

3 Phasen:

- **Simulationsphase:**
  - „highly skewed“ Daten werden simuliert, um klinische Daten zu imitieren
- **Modellierungsphase (4 Regressionsmethoden):**
  - Poisson und Zero-Inflated Poisson (ZIP)
  - Negative Binomial und Zero-Inflated Negative Binomial (ZINB)
- **Anwendungsphase (Implementierung in SAS v9.2)**
  - Modellierungsmethoden auf tatsächliche klinische Daten anwenden und Ergebnisse bewerten/vergleichen, um optimales Modell zu finden
  - Kriterien: MSE, Bias, Überdeckungswahrscheinlichkeit

1. Simulationsphase

„Mixture Distribution Framework“ in 2 Teilen:

- **Null-Teil:**
  - Binomialverteilung (n,p_0) – Schätzung der WSK des Auftretens von Nullen
  - p_0 durch logistische Verteilung basierend auf den Zero-Inflated Regressoren
- **Positiver Count-Teil:**
  - Poissonverteilung(λ) – Schätzung des Zählteils
  - Bedingte Mittelwert durch Fitten eines Poissonmodells basierend auf den Regressoren des positiven Zählteils

1. Simulationsphase

→ Antwort auf Frage der Leser: Welche Daten lagen den Simulationen zugrunde?

- **Daten – Reflexion:**
  - 1 Intercept, 3 Kovariate, 1 kategorielle Surrogatvariable („Herzblock“), 1 stetige Surrogatvariable („Zeit“) für beide Teile des „mixture“–Modells
  - Wahren Werte der schätzbaren Parameter beruhten auf den Schätzungen, die aus den echten Daten erhalten wurden
  - Kategoriale Kovariate – Herzblock (Y/N)
  - Stetige Kovariate – Zeit seit dem letzten Krankenhausaufenthalts
- **3 Simulationsdatensätze mit jeweils n = 1 000 000 generiert:**
  - Anfangswerte für 1. Simulation: Pediatric Cardiac Quality of Life Inventory Dataset (um möglichst nahe an den tatsächlichen klinische Bedingungen zu liegen)
  - 2 weitere Simulationen mit leicht abgeänderte Variablen der 1. Simulation

1. Simulationsphase

- **Simulationsergebnisse:**

Parameter	True Value
First Simulation	
Zeros	
Intercept (α_{00})	1.56
Covariate 1 (α_{01})	-1.37
Covariate 2 (α_{02})	-0.04
Positives	
Intercept (α_{10})	0.21
Covariate 1 (α_{11})	0.72
Covariate 2 (α_{12})	-0.02
Second Simulation	
Zeros	
Intercept (α_{00})	1.66
Covariate 1 (α_{01})	-1.27
Covariate 2 (α_{02})	-0.03
Positives	
Intercept (α_{10})	0.31
Covariate 1 (α_{11})	0.82
Covariate 2 (α_{12})	-0.01
Third Simulation	
Zeros	
Intercept (α_{00})	1.76
Covariate 1 (α_{01})	-1.17
Covariate 2 (α_{02})	-0.02
Positives	
Intercept (α_{10})	0.41
Covariate 1 (α_{11})	0.92
Covariate 2 (α_{12})	-0.009

1. Simulationsphase

Geschätzte Mittelwerte und Standardfehler pro Modell (Verteilung):

M (SD) … mean (standard deviation)

Parameter	True Value	Poisson M (SD)	Poisson M (eSE)	NB M (SD)	NB M (eSE)	ZIP M (SD)	ZIP M (eSE)	ZINB M (SD)	ZINB M (eSE)
First Simulation									
Zeros									
Intercept (α_{00})	1.56					1.55 (0.17)	0.17	1.55 (0.17)	0.17
Covariate 1 (α_{01})	-1.37					-1.37 (0.20)	0.19	-1.37 (0.20)	0.20
Covariate 2 (α_{02})	-0.04					-0.04 (0.08)	0.08	-0.04 (0.08)	0.08
Positives									
Intercept (α_{10})	0.21	-1.56 (0.14)	0.09	-1.56 (0.14)	0.12	0.19 (0.13)	0.14	0.18 (0.13)	0.13
Covariate 1 (α_{11})	0.72	1.69 (0.15)	0.10	1.70 (0.15)	0.14	0.73 (0.14)	0.14	0.73 (0.14)	0.13
Covariate 2 (α_{12})	-0.02	0.003 (0.06)	0.03	0.004 (0.06)	0.04	-0.02 (0.04)	0.04	-0.02 (0.04)	0.04
Second Simulation									
Zeros									
Intercept (α_{00})	1.66					1.66 (0.16)	0.16	1.66 (0.16)	0.16
Covariate 1 (α_{01})	-1.27					-1.27 (0.19)	0.19	-1.27 (0.19)	0.19
Covariate 2 (α_{02})	-0.03					-0.03 (0.08)	0.08	-0.03 (0.08)	0.08
Positives									
Intercept (α_{10})	0.31	-1.54(0.14)	0.09	-1.54 (0.14)	0.12	0.30 (013)	0.13	0.29 (0.13)	0.12
Covariate 1 (α_{11})	0.82	1.76(0.16)	0.10	1.76 (0.16)	0.14	0.82 (0.14)	0.14	0.82 (0.14)	0.13
Covariate 2 (α_{12})	-0.01	0.01(0.06)	0.03	0.01 (0.06)	0.05	-0.01 (0.04)	0.04	-0.01 (0.04)	0.04
Third Simulation									
Zeros									
Intercept (α_{00})	1.76					1.76 (0.16)	0.16	1.76 (0.16)	0.16
Covariate 1 (α_{01})	-1.17					-1.17 (0.18)	0.18	-1.17 (0.18)	0.18
Covariate 2 (α_{02})	-0.02					-0.02 (0.08)	0.08	-0.02 (0.08)	0.08
Positives									
Intercept (α_{10})	0.41	-1.52(0.14)	0.09	-1.53 (0.14)	0.13	0.40 (0.13)	0.13	0.39 (0.13)	0.12
Covariate 1 (α_{11})	0.92	1.82(0.16)	0.10	1.82 (0.16)	0.18	0.93 (0.14)	0.14	0.93 (0.14)	0.13
Covariate 2 (α_{12})	-0.009	0.006(0.06)	0.03	0.006 (0.07)	0.08	-0.008 (0.03)	0.03	-0.008 (0.03)	0.03

2. Modellierungsphase (Zusammenfassung)

- **Fitten der Modelle:**
  - Poisson Distribution
  - Negative Binomial Distribution
  - Zero-Inflated Poisson Distribution (ZIP)
  - Zero-Inflated Negative Binomial Distribution (ZINB)
- **Bewertungskriterien:**
  - Bias
  - MSE
  - Überdeckungswahrscheinlichkeit

2. Modellierungsphase (Wahrscheinlichkeitsverteilungen)

→ Antwort auf Frage der Leser: Tatsächliche Unterschiede in den Verteilungen?

Poisson

x … Regressoren

β… Koeffizientenvektor

Y_i … 0,1,2,…

$$p(y_i \mid X = x) = \frac{e^{-\mu}\mu^{y_i}}{y_i!}$$

$$with \quad \mu = e^{x'\beta}$$

Negative Binomial

α … Dispersionsparameter

Y_i … 0,1,2,…

$$P(y_i|\mu,\alpha) = \frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{\Gamma(y_i+1)\Gamma\frac{1}{\alpha}}\left(\frac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}}\left(\frac{\mu}{\frac{1}{\alpha}+\mu}\right)^{y_i}$$

- Auch Pascalverteilung genannt
- Beschreibt Anzahl der notwendigen Versuche, um in Bernoulliprozess eine gewisse Anzahl von Erfolgen zu erreichen

2. Modellierungsphase (Probability Density Functions)

Zero-Inflated Poisson

Y_i … 0,1,2,…

$$\Pr\left(Y_i = 0\right) = \pi_i + (1-\pi_i)\, e^{-\mu_i}$$

$$\Pr\left(Y_i = y_i\right) = (1-\pi_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$$

where

$$\pi_i = \frac{e^{z_i'\gamma}}{1+e^{z_i'\gamma}} \quad ; \quad \mu_i = e^{x'\beta}$$

Zero-Inflated Negative Binomial

Y_i … 0,1,2,…

$$\Pr\left(Y_i = 0\right) = \pi_i + (1-\pi_i)\left(1+\alpha\mu_i\right)^{-\frac{1}{\alpha}}$$

$$\Pr\left(Y_i = y_i\right) = (1-\pi_i)\frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{y_i!\,\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha}+\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\mu_i}{\frac{1}{\alpha}+\mu_i}\right)^{y_i}$$

π … Wahrscheinlichkeit von „extra“ Nullen

$Pr(Y_i=0)$ … generiert „strukturelle“ Nullen (d.h. nur Null oder Nicht-Null möglich)

$Pr(Y_i=y_i)$ … generiert Nullen nach Poisson- bzw. Negativbinomialverteilung

(d.h. die nicht-negativen Counts können Null sein, müssen aber nicht)

3. Anwendungsphase (Überblick)

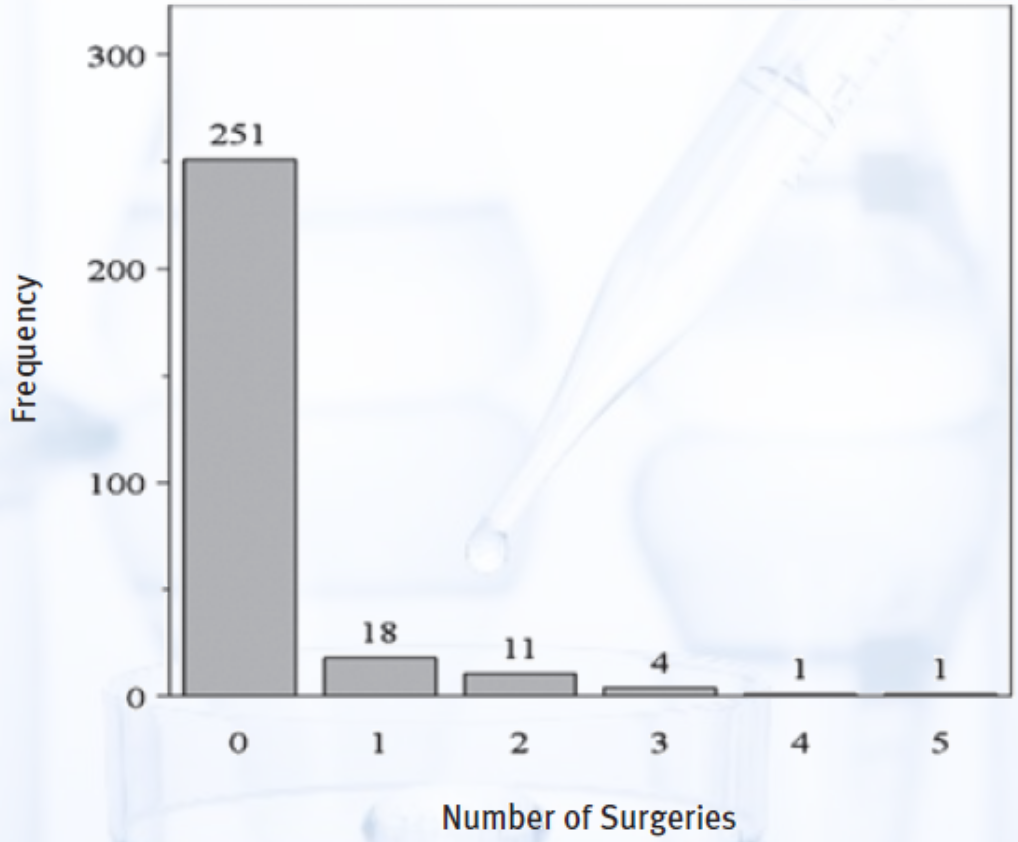
- Deskriptive Statistiken
- Anwendung der vorgestellten Modelle auf klinische Daten

- **Poissonness Plot:** zur Untersuchung, ob Daten einer Poissonverteilung unterliegen
- **Lagrange Multiplier Test**: zur Überprüfung auf Overdispersion
- **Van-den-Broeck Score Test**: zur Überprüfung auf Zero-Inflation
- **Vuong Test**: zum Vergleich von „non-nested models“
- **AIC** für jedes Modell
- **Vorhergesagte und beobachtete WSKT des Count** für jedes Modell

3. Anwendungsphase

Deskriptive Statistik (Histogramm)

Anzahl von operativen Eingriffen an Kindern mit EP Disorder, die bisher keine Transplantation hatten



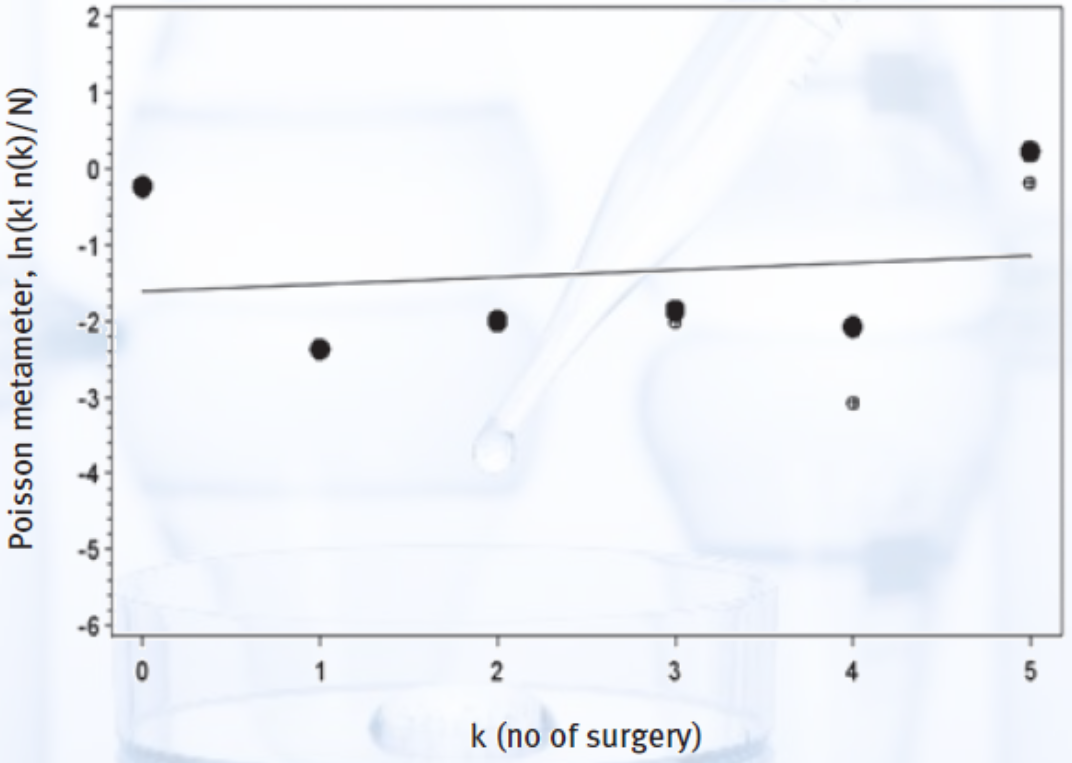
3. Anwendungsphase

- **Poissonness Plot (für die klinischen Daten):**

  Daten liegen nicht auf der Gerade

  → Beobachtete Häufigkeit ≠ Erwartete Häufigkeit

  → Daher kein guter Fit mit Poissonverteilung



3. Anwendungsphase

- **Lagrange Multiplier Test:**
  - Unter H0: „Poissonverteilung ohne Overdispersion“, ~ χ^2 mit 1 Freiheitsgrad
  - Hier: Teststatistik = 4.48 (p-value = 0.03)

    → H0 verwerfen

    → Verwerfe Hypothese, dass Verteilung der Poissonverteilung ohne Overdispersion entspricht

    → Grund: Overdispersion wegen großer Anzahl an Nullen

3. Anwendungsphase

- **Van-den-Broeck Score Test:**
  - Unter H0: „keine Zero-Inflation“, ~ χ^2 mit 1 Freiheitsgrad
  - Vergleicht tatsächliches Auftreten von Nullen mit der Anzahl an Nullen, die vom betrachteten Modell vorhergesagt werden (anhand von Log-Likelihood-Estimates (LL))
  - Signifikantes Testresultat als Hinweis auf Overdispersion (wegen großer Anzahl an Nullen)
    - Poisson: LL = -108.64
    - Negative Binomial: LL = -105.10
    - ZIP: LL = -90.63
    - ZINB: LL = -90.65

3. Anwendungsphase

- **Vuong Test:**
  - Vergleicht vorhergesagte Wahrscheinlichkeiten von zwei „non-nested“ Modellen (z.B. Poisson vs. ZIP oder Negative Binomial vs. ZINB)
  - Resultat (für diese Studie):
    - ZIP hat „besseren“ Fit als Poisson
    - ZINB hat „besseren“ Fit als Negative Binomial

3. Anwendungsphase

- **Zusammenfassung der Resultate:**

CLINICAL DATA: MODEL SELECTION

Test statistic (*p-value*)	Poisson	NB	ZIP	ZINB	Vuong test
Log-likelihood	-108.64	-105.10	-90.63	-90.65	
AIC	231.28	220.20	205.27	205.28	
Vuong test (Poisson vs. ZIP)					3.54 (<0.01)
Vuong test (NB vs. ZINB)					3.85 (<0.01)
Estimated proportion of zeros (%)	0.771	0.783	0.875	0.870	
Estimated dispersion Parameter	NA	0.67	NA	0.001	

AIC (= Akaike Information Criterion): kleiner für ZIP und ZINB
→ Deutet auf besseren Fit der ZIP oder ZINB für die Daten hin

Insgesamt: Entscheidung der Autoren für ZIP Model

3. Anwendungsphase

- **Zusammenfassung der Resultate:**

CLINICAL DATA: RESULTS FROM ZERO-INFLATED POISSON MODEL

Risk Factor	β Coefficient	SE	*t*-value	*p*-value
Logistic Portion of the Model				
Intercept	0.01	2.14	0.01	0.99
Hospitalization	2.25	2.33	0.97	0.33
Heart block	0.12	1.27	0.10	0.92
Prematurity	6.99	3.71	1.88	0.05
Time	-0.08	0.03	-2.48	0.01
Heart block x prematurity	-7.49	3.91	-1.91	0.05
Poisson Portion of the Model				
Intercept	-2.19	0.66	-3.32	< 0.01
Hospitalization	1.81	0.53	3.40	< 0.01
Heart block	1.48	0.52	2.87	< 0.01
Prematurity	2.65	1.04	2.54	0.01
Time	-0.01	0.01	-2.32	0.02
Heart block x prematurity	-2.26	1.13	-2.01	0.04

Note: The logistic portion of the table provides results for the portion of the data that consist of always zero while the Poisson portion of the table consists of the sampling zeroes as well as positive integer portion of the data

Agenda

1 Problemstellung

2 Zugrundeliegende Daten & klinische Variablen

3 Analysephasen & Resultate

4 Fragen/kritische Anmerkungen zum Paper

5 Conclusio

Fragen / Kritische Anmerkungen

- **FRAGE:**

  Funktioniert die Datensimulation am Anfang wie das Bilden eines Training- und Test-Sets?

- **ANTWORT:**

  Ja, in etwa – um zu analysieren, welche Modellwahl für alle möglichen Daten dieser Form am besten geeignet ist. 286 vs. 1,000.000 Beobachtungen

Fragen / Kritische Anmerkungen

- **FRAGE:**

  ZIP und ZINB liefern einen besseren Fit und fast erwartungstreuen Schätzer.

  Liegt das an der Methode oder an den Daten?

- **ANTWORT:**

  In den Daten kommen im Vergleich zur reinen Poisson-/Negativen Binomial-verteilung sehr viele Nullen vor. Daher berücksichtigt eine Zero-Inflated Distribution solch eine Situation (im Allgemeinen) besser.

Fragen / Kritische Anmerkungen

- **FRAGE:**

  Was sind Non-Nested bzw. Nested Modelle?

- **ANTWORT:**
  - „Nested“ bedeutet hierarchisch
  - D.h. Modell A ist nested in Modell B, wenn Parameter in Modell A ein Subset der Parameter in Modell B sind
  - Bsp:

    Modell A hat 4 Parameter: β0, β1, β2, β3

    Modell B hat 7 Parameter: β0, β1, β2, β3, β4, β5

    Modell C hat 6 Parameter: β0, β1, β2, β4, β5

    Modell A ist nested in Modell B. Modell C ist nested in Modell B.
    C und A sind non-nested (weil jedes Modell Parameter enthält, die das andere nicht enthält).

Fragen / Kritische Anmerkungen

- **FRAGE:**

  Vuong-Test: wie ist Teststatistik oder p-Wert zu lesen, und wie kann man sagen welches Modell besser ist?

- **ANTWORT:**

  Testet die H0, ob 2 Modelle – egal ob „nested“ (hierarchisch) oder nicht – gleich nahe an der wahren Verteilung liegen. Er trifft aber keine Entscheidung, ob das hier bessere Modell auch generell das „wahre“ beste Modell ist.

Agenda

1 Problemstellung

2 Zugrundeliegende Daten & klinische Variablen

3 Analysephasen & Resultate

4 Fragen/kritische Anmerkungen zum Paper

5 Conclusio

Conclusio

- **Zero-Inflated Models**
  - Einsatz gut, wenn tatsächliche Anzahl an Nullen die Anzahl an Nullen, die von traditionellen Poisson-/Negativbinomialmodellen vorhergesagt würden, übersteigt
  - Traditionelle Modelle können zu verzerrter Schätzung und Inferenz führen
- **Mischung von Modellen**
  - Logistischer Teil (für Nullen)
  - Poisson-/Negativbinomialteil (für Count-Daten)
- **Anwendung von reiner Poissonverteilung (in dieser Studie)**
  - Würde Anzahl von Nullen unterschätzen (→ Overdispersion)
  - Annahme von Gleichheit von Erwartungswert und Varianz wäre verletzt

Literaturverzeichnis

- [1] **Baierl, A. 2018**. "*Modeling Non-Negative Data*." Unterlagen zum Kurs Biostatistik. S. 1-7.
- [2] **Gupta, R., et al. 2013**. "Finding the Right Distribution for Highly Skewed Zero-Inflated Clinical Data." In *Epidemiology Biostatistics and Public Health*. Vol. 10 No. 1, S. e8732-1 – e8732-15.
- [3] **Humphreys, B. 2013**. "*Dealing with Zeroes in Economic Data.*" University of Alberta – Department of Economics. 04.04.2013. S. 1-27.

Danke für die Aufmerksamkeit!