Statistische Modellierung latenter Strukturen in den Lebens-, Sozial- und Wirtschaftswissenschaften

Überlick über Modelle für defizitäre Daten

Seminarleiter: Prof. Dr. Thomas Augustin
Betreuerin: Julia Plass

Referentin: Maria Schmelewa

Institut für Statistik
LMU

18.01.2014

Aufbau

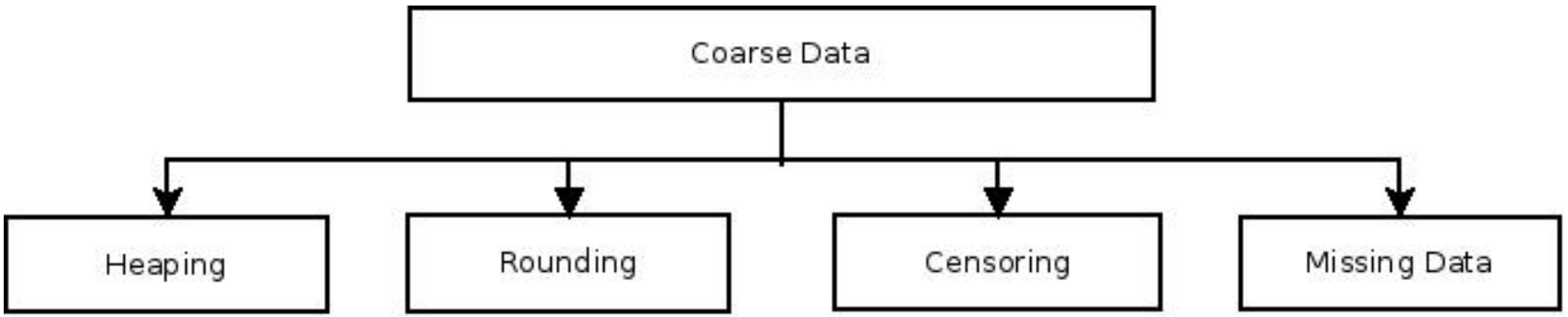
1. Übersicht: Defizitäre Daten
2. Runden
3. Heaping
4. Zensierung
5. Fehlende Werte

Übersicht

Coarsening ("Die Vergröberung")

Überbegriff für unvollständige Daten

Es gibt verschiedene Arten defizitärer Daten, welche alle unter den Begriff Coarse Data fallen. Die bekanntesten und sehr häufig vorkommenden sind hierbei



Heitjan und Rubin haben 1991 zu dem Begriff Coarse Data eine prägnante Definition niedergeschrieben.

Coarse Data: *"[It is the kind of] data [that] are neither entirely missing nor perfectly present. Instead, we observe only a subset of the complete-data samle space in which the true, unobservable data lie."(Heitjan and Rubin, 1991)*

Coarsening

Vor Betrachtung der einzelnen Arten werden allgemeine Likelihood Funktionen für die zwei Arten, dem stochastischen und nicht-stochastischen Coarsening aufgestellt. $f(y;\theta)$ stellt hierbei die Dichtefunktion der wahren Werte dar.

Likelihood beim stochastischen Coarsening:

$$L(\theta,\gamma;x) = \int_x f(y;\theta)k(x;y,\gamma)dy \qquad (1)$$

Der Unterschied zwischen den beiden Likelihood Funktionen ist, dass bei (2) die Stochastizität des Coarsenings $k(x;y,\gamma)$ nicht mit berücksichtigt wird.

Likelihood beim nicht-stochastischen Coarsening:

$$L(\theta;x) = \int_x f(y;\theta)dy \qquad (2)$$

Voraussetzung für die Verwendung von (2) ist Distinktheit der Parameter und Coarsening at Random bei den Beobachtungen der interessierenden Variable. Nur dann ist der Coarsening- Mechanismus ignorierbar und kann bei der Schätzung weggelassen werden.

Distinktheit: Keine funktionale Beziehung zwischen θ und γ.

CAR: $P(y_{obs}|y_{true},\gamma)$ gleich für alle y_{obs}.

Ignorierbarkeit: Der Schätzer behält seine günstigen Eigenschaften, auch wenn Coarsening-Mechanismus bei der Schätzung nicht berücksichtigt wird.

Runden

Definition: Das Runden von Werten mit gleichem Rundungsintervall, üblicherweise auf die nächste ganze Zahl oder Dezimalstelle.
$\rightarrow$ Diskretisierung von stetigen Variablen
Auswirkungen: Verzerrung der geschätzten Momente und Regressionskoeffizienten.
Die beobachteten Werte X_i^* werden wie folgt modelliert. $round(\frac{X_i}{h})$ stellt dabei die Funktion des Rundens dar. Der Rundungsfehler wird als δ_i definiert. h beschreibt hierbei den Abstand zwischen zwei Punkten, in welchen der wahre Wert X_i liegt.
Modellierung:

$$X_i^* = h\, round \frac{X_i}{h} \quad \& \quad \delta_i = X_i^* - X_i$$

Likelihood: Da Nicht-stochastizität angenommen wird, lautet die Likelihood Fkt.:

$$\prod_i \int_{x_i^* - \frac{h}{2}}^{x_i^* + \frac{h}{2}} f(y;\theta)dy$$

$f(y;\theta)$ beschreibt dabei die Dichtefunktion der wahren Werte.
Mögliche Verfahren: Sheppard Korrektur

Heaping

Beispiel: Dauer der Arbeitslosigkeit von 14-29 jährigen aus dem italienischen LSF für die Lombardei

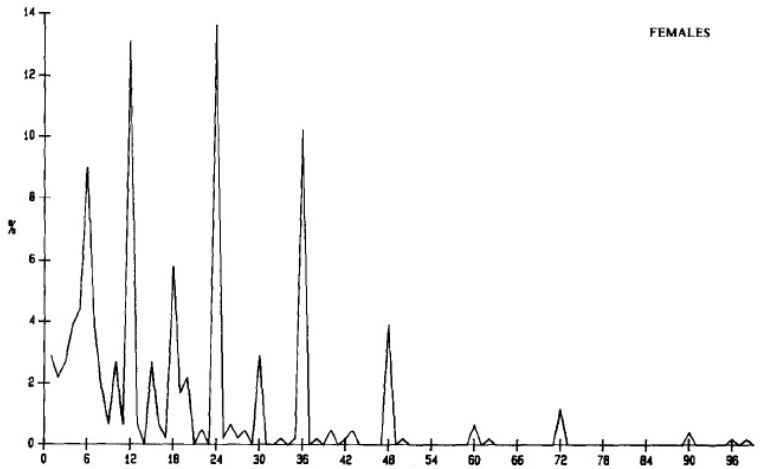


Fig. 1. Percentage distribution of unemployed individuals aged 14–29 years by reported unemployment duration (in months) at initial survey, Italian LFS, matched data for Lombardy, 1986.I–II: males, $N = 267$ and females, $N = 411$.

Abbildung : N. Torelli und U.Trivellato, 1993

Bei dieser Grafik erkennt man große Peaks bzw. abnormale Häufungen bei den Monaten 6, 12, 18, 24 etc.

Heaping

Einen allgemeineren Fall als das Runden stellt Heaping dar.

Definition: „Digit preference“ - Abnormale Häufung bzw. Konzentration von Beobachtungen an bestimmten Werten.

Daten enthalten sowohl exakte als auch unter unterschiedlichem Genauigkeitsgrad gerundete Werte.

Eine große Schwierigkeit besteht somit unter anderem darin zu Unterscheiden, welche Beobachtungen geaheapt und welche die wahren Werte sind.

Übliche Werte bei Zeitangaben: Vielfaches von 4 (bei Wochen), vielfaches von 6 oder 12 (bei Monaten), vielfaches von 10(bei Jahren).

Entstehung: Retrospektiv erhobene Verweildauern, retrospektive Befragung zu Alter, Mengenangabe etc.

Auswirkungen: Über- bzw. Unterschätzung von Parametern, Bias in den geschätzten Regressionskoeffizienten.

Heaping

Da es meist um Zeitangaben geht, werden die wahren Werte als t_i bezeichnet. Die beobachteten Werte ${}_H t_i$ stehen mit den wahren Werten wie folgt im Zusammenhang. p_i sind Werte von Bernoulli-verteilten Zufallsvariablen P_i, welche die Ausprägung 1 mit der Wahrscheinlichkeit $G(t_i;\gamma)$ annehmen. $G(t_i;\gamma)$ gibt dabei die Wahrscheinlichkeit für Heaping an und wird als „Heaping Funktion“ bezeichnet.
Modellierung: $h_{(m)}$ sind die geheapten Werte in aufsteigender Reihenfolge.

$$ {}_H t_i = t_i + d_i p_i \text{ mit } d_i = h_{(m)} - t_i $$

Likelihood: Da der Heaping Prozess stochastisch ist, lautet die Likelihood Fkt.:

$$ \prod_{i=1}^{I} [f(t_i;\theta)(1 - G(t_i;\gamma))] \prod_{j=1}^{J} \int_{{}_l t_j}^{{}_u t_j} f(y;\theta) G(y;\gamma) dy $$

Der erste Teil schätzt die wahren Werte t_i, der zweite Teil die geheapten, für welche gilt: ${}_H t_i \neq t_i$. Die Intervallgrenzen, in welchen die geheapten Werte liegen, werden mit ${}_l t_j$ und ${}_u t_j$ angegeben.
Mögliche Verfahren:

- **Maximum-Likelihood-Schätzung**
- Glättung mit penalisierter Likelihood

Zensierung

Definition: Wenn nur ein Teil der Daten, d.h. nicht alle Werte einer statistischen Variable bekannt sind.

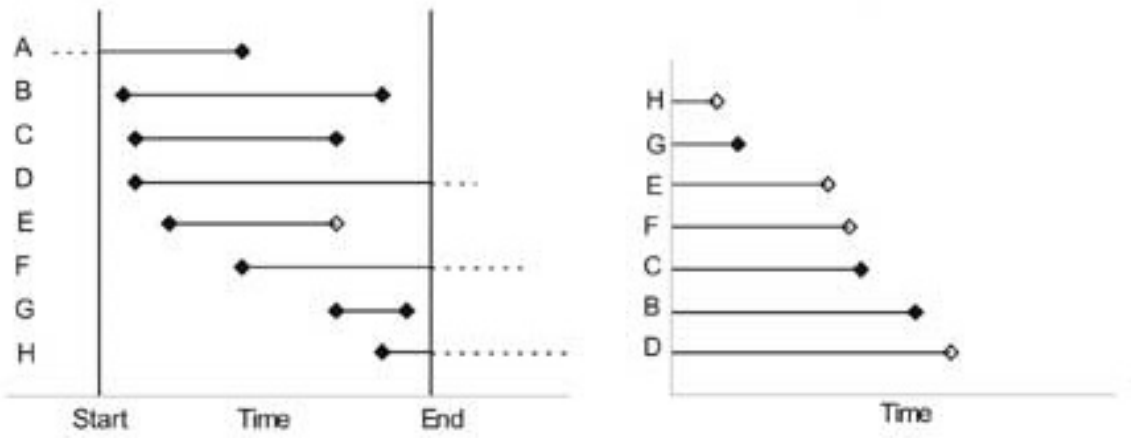


Abbildung : http://www.indianpediatrics.net/sep2010/sep-743-748.htm

Entstehung: Bei Analyse von Lebensdauern

- Soziologie: Dauer bis Scheidung.
- Medizin: Überlebensdauer von Krebspatienten.
- Technik: Ausfall von technischen Geräten.

Zensierung

Arten: Da es sich hier wieder um Zeitangaben handelt, werden die wahren Werte als t_i und die beobachteten als t_i^* bezeichnet.

- **Rechtszensierte Daten:** Ereignis bis zum Ende nicht beobachtet.
  - Typ1: Feste, vorher festgelegte Beobachtungsdauer.
  - Typ2: Untersuchung beendet, wenn vorher festgelegte Zahl von Lebensdauern unzensiert beobachtet.
  - Random Censoring: Zensierungszeiten unabhängig von T_i.
- **Linkszensierte Daten:** Ereignis bereits vor Beginn eingetreten.
- **Intervallzensierte Daten:** Ereignis zwischen ZP a und b eingetreten.
- **Trunkierung:**
  - Rechtstrunkierung: Nur Personen, die das interessierende Ereignis erlebt haben, gelangen in Studie.
  - Linkstrunkierung: Verursacht durch weiteres Ereignis. Falls dieses zum ZP $T \geq Y$ eintritt, gelangt Individuum nicht in Studie.

Zensierung

Da zumeist Rechtszensierung von Interesse ist, wird diese im Folgenden betrachtet. Beobachtete Werte werden über das Minimum aus den latenten Größen T_i und C_i modelliert. δ_i ist dabei eine Indikatorfunktion bezüglich der Zensierung.

Modellierung: Zwei latente Größen

1. Wahre Lebensdauer $T_i \overset{iid}{\sim} F$
2. Maximale Beobachtungsdauer $C_i(\textit{Zensierungszeit}) \overset{iid}{\sim} G$

$T_i^* = min(T_i, C_i)$ & $\delta_i = I\{T_i \leq C_i\}$ $\delta_i = 1$, wenn nicht zensiert, $\delta_i = 0$, sonst

Ziel: Schätzung von Hazardraten und Survivorfunktionen

Es gibt drei verschiedene Arten die Harzardraten und Survivorfunktionen zu schätzen, unter anderem parametrisch. Da auch hier von Stochastizität ausgegangen wird, besteht die Likelihood aus zwei Teilen, dem für unzensierte und für zensierte Daten.

- Parametrisch $\rightarrow$ basierend auf Verteilungsannahme $G(t_i)$ Survivorfkt. von C_i, **Likelihood:** $\prod_{i=1}^{n}[f(t_i;\theta)^{\delta_i} S(t_i;\theta)^{1-\delta_i}][g(t_i;\gamma)^{1-\delta_i} G(t_i;\gamma)^{\delta_i}]$ $S(t_i)$ von T_i.
- Semiparametrisch $\rightarrow$ Spline Schätzung, Likelihood-basiert
- Nichtparametrisch $\rightarrow$ Ohne Verteilungsannahme, rein Datenbasiert z.B. Kaplan-Meier-Schätzer, Nelson-Aalen-Schätzer

Fehlende Werte

Definition: ”Werte werden dann als fehlend bezeichnet, wenn als existierend angenommene und als Reaktion auf einen Reiz hervorrufbare Werte, deren Beobachtung intendiert ist, als Reaktion auf die Reizvorgabe nicht beobachtet werden und auch nicht ohne Unsicherheit aus anderen Informationen ableitbar sind.”(M. Spieß, 2008) Es gibt folgende drei Arten von fehlenden Werten:

- **Item-Nonresponse:** Einzelne Werte für beobachtete Einheiten fehlen.
- **Unit-Nonresponse:** Ganze Einheiten nicht beobachtet.
- **Panelattrition:** Ausscheiden im Verlauf bei Panelstudien.

Entstehung: Fehler bei Messung, falsch eingetragene oder unlesbare Werte, Übersehen einer Frage, Verweigerung einer Angabe etc.

Fehlende Werte

Für die weitere Betrachtung und Analyse von Daten mit fehlenden Werten ist es wichtig zwischen drei Missingmechanismen zu unterscheiden. Falsche Annahmen bezüglich des Missingmechanismus kann zu verzerrten Ergebnissen führen. Hierbei ist anzumerken, dass die ersten beiden Mechanismen für die Schätzung zumeist ignorierbar sind.

Missingmechanismen:

x_{obs} und x_{mis} sind Vektoren mit den jeweils beobachteten und nicht beobachteten Werten. r ist ein „Response-Indikator“ Vektor mit Ausprägung 1, wenn x_{obs} und 0, wenn x_{mis}. $g(.)$ beschreibt hierbei die unbekannte Wahrscheinlichkeitsfunktion von r und γ einen unbekannten Parameter, welcher den Missingmechanismus steuert.

- MCAR („Missing completely at random“):
  $g(r|x_{obs}, x_{mis}; \gamma) = g(r; \gamma)$ (generell ignorierbar)
- MAR(„Missing at random“):
  $g(r|x_{obs}, x_{mis}; \gamma) = g(r|x_{obs}; \gamma)$ (meistens ebenfalls ignorierbar)
- NMAR(„Not missing at random“):
  $g(r|x_{obs}, x_{mis}; \gamma)$ *keine Vereinfachung möglich* (nicht ignorierbar - externe Informationen notwenig)

Fehlende Werte

Auch bei dieser Art von Coarse Data wird die Likelihood aufgezeigt, welche die Maximum-Likelihood-Schätzung verwendet wird. Auch für die meisten anderen Schätzmethoden ist MAR eine Grundvoraussetzung.

Likelihood

Der erste Teil der Likelihood stellt die Dichte für die beobachteten Werte der interessierenden Variable dar. Der zweite Teil beschreibt die Wahrscheinlichkeitsfunktion von r.

$$\prod_i f(x_{i\ obs}; \theta) \cdot g(r|x_{i\ obs}; \gamma)$$

Methoden: Voraussetzung wenigstens MAR

- **Maximum-Likelihood-Schätzung**
- EM Algorithmus
- Multiple Imputation

Auswirkungen: Kann je nach Missingmechanismus zu verzerrten Ergebnissen (Unter- oder Überschätzung von Mittelwerten) und Fehlinterpretation führen.

Zusammenfassung

- Viele verschiedene Arten von defizitären Daten, welche Spezialfälle von Coarsening darstellen.
  Häufigste: Runden, Heaping, Zensierung, fehlende Werte.
- Bei Ignorierbarkeit zumeist nicht problematisch.
- Falls Coarsening Mechanismus nicht ignorierbar, Verzerrung von geschätzten Parametern, Regressionskoeffizienten und Fehlinterpretation möglich.

Literatur

- Heitjan, D. F. und Rubin, D. B. (1991). Ignorability and Coarse Data. The annals of Statistics, 19(4):2244-2253.
- Plaß, J. (2013). Coarse categorical data under epistemic and ontologic uncertainty: Comparison and extension of some approaches. Institut für Statistik: Master Thesis.
- Schneeweiss, H., Komlos, J. und Ahmad, A.S. (2006).Symmetric and Asymmetric Rounding. Insitut für Statistik: Sonderforschungsbereich 386, Paper 479.
- Torelli, N. und Trivellato, U. (1993). Modelling inaccuracies in job-search duration data. Journal of Econometrics, 59:187-211.
- Indrayan, A, Basal A.K. (2010). The Methods of Survival Analysis for Clinicians. Indian Pediatr, 47: 743-748.
- Spieß, M. (2008). Missing-data-Techniken: Analyse von Daten mit fehlenden Werten. Lit Verlag.

Vielen Dank für Ihre Aufmerksamkeit!