Messfehler (Heaping) in Arbeitslosigkeitsdaten

Sebastian Kaiser

10.06.2004

Überblick

1.Einleitung I

Ziel:

- Modellierung von Arbeitslosigkeitsdauern nach retrospektivischer Befragung

Problem:

- Fehlerbehaftete Daten aufgrund der Umfragen

1.Einleitung II

Periodische Umfragen

i) In regelmäßigen Zeitabständen Umfragen durchführen

ii) Aktuelle Arbeitslose nach Dauer der Arbeitssuche befragen

iii) Zusammenfügen von Ergebnissen derselben Person falls vorhanden (Longitudinalcharakter)

2. Heaping

Def: Anormale Häufung von Anteilen an bestimmten Zeitpunkten

- Retrospektivische Fragestellung
- Skalenabhängigkeit
- Spezielle Form des Rundens

2. Heaping: Retrospektivische Fragestellungen

- Erinnerungseffekte
- Teleskopeffekte

$\Longrightarrow$ Heaping in allen retrospektivischen Umfragen vorhanden

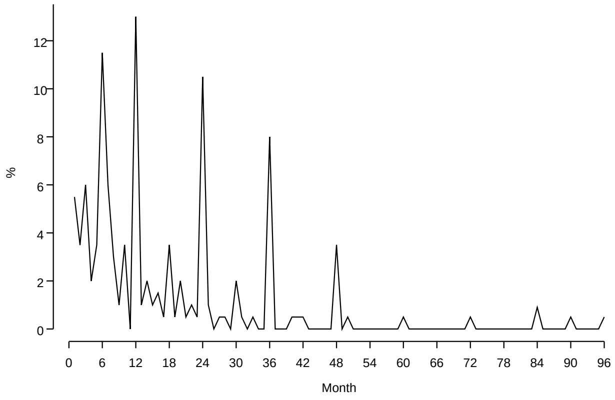
2. Heaping: Skalenabhängigkeit

- Umso größer die Zeitspanne, desto größer die Tendenz zum Runden
- Skala Wochen $\Longrightarrow$ Häufung bei 4,8,12... (Monate)
- Skala Monate $\Longrightarrow$ Häufung bei 6,12,18... (Jahre)

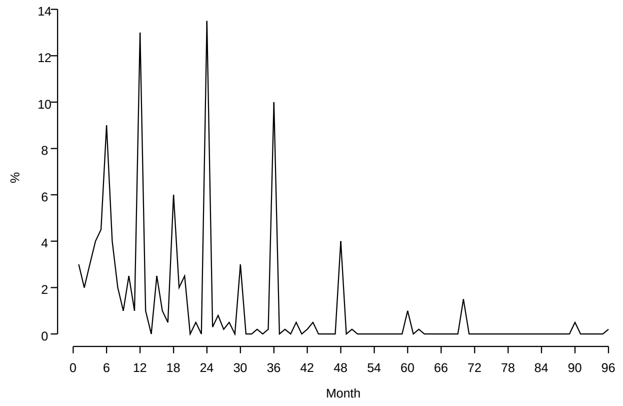
$\Longrightarrow$ Skala wird auf die um 1 höhere Ebene projeziert

- Abhilfe: z.B. anderes Fragebogendesign

Beispiel: Arbeitslosigkeitsdauern der Männer



Beispiel: Arbeitslosigkeitsdauern der Frauen



2. Heaping: Mögliche andere Ursachen für Spikes

i) Spikes repräsentieren wahre Begebenheit (z.b. Ende der Arbeitslosenunterstützung)

ii) Periodische Zyklen auf dem Arbeitsmarkt (Saisonarbeit)

iii) Fehler bei der Stichproben- oder Datenauswahl

$\Longrightarrow$ Heaping Effekt und andere Anormalitäten gleichzeitig beheben

3. Modell zum Heaping Prozeß

Zu Beginn sucht man ein Modell, das die Heaping Werte ${}_Ht_i$ auf die tatsächlichen Werte t_i zurückführt:

$$ {}_Ht_i = t_i + k_i y_i \tag{1}$$

mit $k_i = h_{(m)} - t_i$,$h_{(m)} \epsilon H$ und $|h_{(m)} - t_i| \to \min$

- mit $k_i = h_{(m)} - t_i$,$h_{(m)} \epsilon H$ und $|h_{(m)} - t_i| \to \min$
- $H = \{h_{(1)} < h_{(2)} < \ldots < h_{(M)}\}$ Punktmenge
- y_i Bernoulli verteilt mit $p(t_i) = G(t_i, \gamma)$
- $G(\cdot, \gamma)$ Heaping Funktion

3.1 Likelihood

Als nächstes sucht man die Likelihood zum ZeitdauerModell $f(t, \vartheta)$

$$\prod_{i=1}^{I} [f(t_i, \vartheta)(1 - G(t_i, \gamma))] \prod_{j=1}^{J} \int_{{}_lt_j}^{{}_ut_j} f(z, \vartheta) G(z, \gamma) dz \tag{2}$$

- I Zahl der "non heaped"-Werte; J Zahl der "heaped"-Werte
- ${}_ut_j = [h_{(m)} + h_{(m+1)}]/2; {}_lt_j = [h_{(m)} + h_{(m-1)}]/2$für ${}_Ht_j = h_{(m)}$

Falls $G()$ auf einem Intervall ${}_l t_{j,u}\, t_j$ kann die Likelihood auch so lauten:

$$\prod_{i=1}^{I}[f(t_i,\vartheta)(1-G(t_i,\gamma))]\prod_{j=1}^{J}[F({}_u t_j,\vartheta)-F({}_l t_j,\vartheta)]G({}_H t_j,\gamma)dz \qquad (3)$$

3.2 Implementierung

Wenn man Modell (1) verwenden will, sind noch einige Veränderungen notwendig

i) Auswahl von H

ii) Bestimmung von $G(\cdot,\gamma)$

iii) Verfeinern von Modell(1)

iv) Likelihood für diskrete Zeitdauern

3.2 Implementierung: Auswahl von H und $G(\cdot,\gamma)$

- Bestimmung von H empirisches Problem
- $G(\cdot,\gamma)$ Funktion von $\mathbb{R}^+$ nach $[0,1]$
- Verteilungsfunktionen mit geringer Anzahl an Variablen günstig
- Funktion so auswählen, dass sie den hervorstechenden Eigenschaften des Heaping entspricht

3.2 Implementierung: Verfeinern von Modell(1)

Modell(1) ist eigentlich nur ein simples Runden. Es gibt deshalb viele Möglichkeiten das Modell zu verfeinern. Einfache sind:

a) H so bestimmen, dass bei großen Zahlen weiter Abstände auftreten

b) Modifizieren der Begrenzungsregel durch Definition asymmetrischer Intervalle $[{}_l t_{j,u}\, t_j]$

Ein interessanterer Weg ist die Erweiterung von Modell(1):

$${}_H t_i = t_i + y_i[q_i w_i + r_i(1-w_i)] \qquad (4)$$

mit $q_i = \bar{h}_{(m)} - t_i, \bar{h}_{(m)} \epsilon \bar{H}$ und $|\bar{h}_{(m)} - t_i| \to \min$
mit $k_i = \tilde{h}_{(m)} - t_i, \tilde{h}_{(m)} \epsilon \tilde{H}$ und $|\tilde{h}_{(m)} - t_i| \to \min$

- $H = \bar{H} \cup \tilde{H}$
- w_i Bernoulli verteilt (unabhängig von y_i) mit $p_w(t_i) = L(t_i, \gamma_2)$

Die Heaping-Funktion ist jetzt $G(\cdot, \gamma_1)L(\cdot, \gamma_2)$

3.2 Implementierung: Likelihood für diskrete Zeitdauern

Da es sich bei den Daten um diskrete Zeiten handelt, muss die Likelihood modifiziert werden. Dabei bleibt der erste Teil von Formel (2) erhalten und der zweite Teil lautet dann:

$$\prod_{j=1}^{J} [f(t_i, \vartheta)(1 - G(t_i, \gamma))] \int_{_lt_j}^{_ut_j} f(z, \vartheta)G(z, \gamma)dz \tag{5}$$

Für jeden Stichprobenplan müssen also die Likelihood und das Modell(1) entsprechend dem Plan angepasst werden. Dies geschieht empirisch.

4. Simulationen

Es wurde eine Monte Carlo Studie durchgeführt, mit folgendem Simulationsplan:

a) $f(t, \vartheta)$ Weibull-,Exponential- oder Log-Logistisch-verteilt. ϑ jeweils drei verschieden Werte

b) Es wird Modell(1) verwendet mit:
   - $G(t, \gamma)$ Verteilungsfunktion der Exponentialverteilung mit zwei verschiedenen Werten von γ
   - H als Vielfaches von 12 oder von 6

c) verschiedene Sample Zahlen: N=50, N=200 und N=500

4. Simulationen

Aus den erhaltenen simulierten Daten wurden dann γ und ϑ geschätzt unter folgenden Annahmen

(A) Verwendung der exakten Likelihood(2)

(B) Verwendung der Likelihood(3) (teilweise konstate Funktion $G(t, \gamma)$)

(C) Ignorieren des Heaping-Effekts (Einfaches Schätzen von ϑ)

$\Longrightarrow$ Es wird untersucht, ob sich A besser verhält als die simpleren Schätzer B,C

4. Simulationen

Ergebnisse und Schlüsse

- Ausgewählte Schätzer mit Standardabweichung im Anhang (anhangheaping.ps)
- Im exponentiellen Fall: B,C verhalten sich ähnlich wie A (Mittelwertsuffizienz der Exponentialverteilung)
- Bei Weibull und Log-Logistischer Verteilung:
  i) Schätzer bei B,c verschieden von A bei den meisten Schätzungen (dramatische Unterschiede bei einigen Log-Logistischen Modellen)
  ii) B beeinflußt von downward bias und C von upward bias
  iii) Bias abhängig von der Anzahl an heaping in den Daten

- Vergrößerung von H führt zu geringerem Bias von B,C (größere Anzahl von Spikes führt zu gleichmäßiger Verteilung der Daten)
- Erhöhung von N führt zu genauerer Unterscheidung und zu größeren Unterschieden von A,B,C

$\Longrightarrow$ Die Erscheinung der unbeständigen Schätzer hängt vom Auftreten und der Stuktur des Heaping und von der Verteilung der Daten ab.

(Die Simulationen haben auch gezeigt, dass C für N=500 B dominiert. D.h. eine falsche Behandlung von Heaping ist schlimmer als das einfache Ignorieren.)

5. Eine empirische Analyse

$$h(t|x) = h_0(t,\vartheta)exp(\beta' x) \tag{6}$$

- $h_0(t,\vartheta)$ Baseline Hazard
- $F(t|x)$ dazugehörige Verteilungsfunktion

5. Eine empirische Analyse

Arbeitslosigkeitsdaten werden immer im Abstand von 3 Monaten erhoben. Deshalb kann man für das Ende der Arbeitslosigkeit nur ein Intervall $[t, t+3]$ angeben. Aus diesem Grund benützt man folgende Likelihood bei der Analyse:

$$\prod_{i=1}^{NO}(1 - c(t_i|x)) \prod_{j=1}^{ND}(t_j|x), c(t|x) = \frac{1 - F(t+3|x)}{1 - F(t|x)} \tag{7}$$

- $NO(ND)$Anzahl der Individuen mit kompletter(fortschreitender) Zeitdauer der Arbeitslosigkeit

5. Eine empirische Analyse

Es werden 4 verschieden Herangehensweisen verglichen:

(A) Das Modell mit Heaping-Prozeß wie bisher besprochen, mit

- $H = \{12s, s = 1, 2, \ldots\}$ und $\tilde{H} = \{6 + 12s, s = 1, 2, \ldots\}$
- $G(\cdot, \gamma_1)$ und $G(\cdot, \gamma_2)$ Verteilungsfunktion von Exponentialverteilungen

(B) Zwei Dummy-Variablen um das Heaping zu korrigieren:

- D1 $= 1$ für $t = 12s(s = 1, 2, \ldots)$
- D2 $= 1$ für $t = 6 + 12s(s = 1, 2, \ldots)$

(C) Ein Modell mit gleichverteilten Zufallsvariablen v_i, um die Spikes auf die Umgebung zu verteilen :

- $\underline{t_i} = t_i - 2 + v2_i$ für $t_i = 6s$, $s = 1, 2, \ldots, 6$
- $\underline{t_i} = t_i - 3 + v3_i$ für $t_i = 6s$, $s > 6$
- $\underline{t_i} = t_i$ sonst.

(D) Ein Modell, das Heaping einfach ignoriert.

(Ergebnisse und Parameter-Schätzer im Anhang)

5. Eine empirische Analyse: Ergebnisse

- Das Modell B scheint eher geeignet, um feste Effekte zu modellieren und schneidet eher schlecht ab
- C hängt sehr von den Zufallszahlen ab und muss von vorneherein schon sehr genau bestimmt werden
- Es bleibt also der Vergleich zwischen A und D, die beide sehr ähnliche Parameter-Schätzer erzeugen
- Da A sehr gute Parameterschätzungen liefert (vor allem für den Unterschied zwischen Männern und Frauen), muss man zwei Dinge berücksichtigen:

i) Untersuchungen haben gezeigt, dass der Stichprobenplan (wie oben besprochen) dem Modell D entgegen kommt.
ii) Heaping hängt von Messfehlern, aber auch von Modellfehlern ab. D.h. ein nur kleiner Effekt des Heaping kann auf einen Fehler des Modells hinweisen.

6.Anmerkungen

Da die Ergebnisse von Modell und Daten abhängen hier nur ein paar Anmerkungen:

a) Heaping hat Auswirkung auf alle retrospektivisch erhobenen Daten und die ergebnisse sind deshalb von breiterem Interesse

b) Beim Heaping Effekt spielt das Modell und die Art der Stichprobenerhebung eine wichtige Rolle. Auch deshalb ist es wichtig, bei der Formulierung des Modells sehr vorsichtig zu sein

Die Forschung mit Heaping ist noch nicht sehr weit fortgeschritten, aber spielt eine immer größere Rolle gerade in solchen Zeitdauermodellen. Es gibt noch viele Möglichkeiten, die hier dargestellten Modelle zu verbessern.

Ein Beispiel ist die Tatsache, dass jede diskret erhobene Variable als eine Art von heaping aufgefasst werden kann.

7.References

N.Torelli und U.Trivellato(1993) Modelling inaccuracies in Job-search duration data
Journal of Econometrics 59(1993) 187-211.

Amar Ahmad(2003)Heaping
Seminar Fehler in Variablen, Sommersemester 2003

T.Augustin und J.Wolff(2002)A bias analysis of Weibull model-s under heaped data
Statistical Papers(2004) 45,211-229

K. Petoussis,R.D. Gill und C. Zeelenberg(2004)Statistical Analysis of Heaped Duration Data