Längsschnittmodelle und Übergangsanalysen

Ulrich Rendtel

Institut für Statistik und Ökonometrie
FU Berlin

RatSWD Nachwuchsworkshop
Längsschnittanalysen auf Basis amtlicher Sozial- und Wirtschaftsdaten
FU-Berlin, 25.–28 August 2009

Freie Universität Berlin

Kalenderzeit und Prozesszeit

- Kalenderzeit: $t =$ Monat nach Beginn der (Panel-) Befragung. MZ–Panel (96–99): $t = 0$ (April 1996) $, \ldots, t = 36$ (April 1999)
- Prozesszeit: $t =$ Anzahl der Monate nach Ereignisbeginn, z.B. Arbeitslosigkeit. $T =$ Anzahl der Monate bis zum Ende der Arbeitslosigkeit.
- Episodendarstellung: $(t_{\text{begin}}, Z_{\text{begin}}), (t_{\text{end}}, Z_{\text{end}})$
  $Z_t \in \{E(\text{erwerbstätig}), U(\text{nemployed}), N(\text{ot employed})\}$
  $Z_t =$ Zustand zum Zeitpunkt t

Die Dauer einer Episode

Beispiel: Dauer einer Arbeitslosigkeitsepisode ("Unemployment Spell")
Zeit: Prozesszeit. $T =$ Zeit bis Ende einer Episode

- T stetige Zufallsgrösse: "Modell in stetiger Zeit"
- T diskrete Zufallsgrösse: " Modell in diskreter Zeit "
- Die Übergänge zwischen stetig und diskret sind fliessend:
  Im MZ gibt es retrospektive , monatliche Kalenderangaben, meistens als stetige Zeit analysiert.
  Im MZ-Panel meistens nur Angaben zum Zustand an 4 Befragungszeitpunkten. Kann nur als diskrete Zeit analysiert werden.

Freie Universität Berlin

Die Überlebenswahrscheinlichkeit (1/2)

Definition: $S(t) = P(T > t) = 1 - F(t)$
Der Kaplan-Meier (Product Limit) Schätzer von $S(t)$:

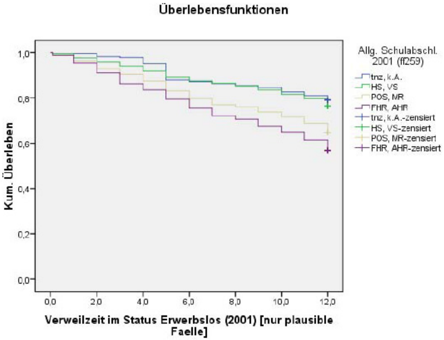
- $t_1 \leq t_2 \leq \ldots \leq t_n$ geordnete Menge von n Episodendauern
- $R_i =$ Anzahl Episoden unter Risiko im Intervall (t_{i-1}, t_i)
- $E_i =$Anzahl der Episoden, bei denen das Ereignis im Intervall (t_{i-1}, t_i) eingetreten ist.
- $$\hat{S}(t) = \prod_{t_i \leq t} \frac{R_i - E_i}{R_i}$$
- $\hat{S}(t)$ ist eine monoton fallende auf dem Intervall (t_{i-1}, t_i) konstante Sprungfunktion.

Die Überlebenswahrscheinlichkeit (2/2)

(Rechts-)Zensierungen: Ende der Episode wird nicht beobachtet, weil Abbruch der Beobachtung zum Zeitpunkt t_i: $E_i = 0$.
($\Rightarrow$ Konstanz von $\hat{S}(t)$).

Kaplan-Meier Schätzer werden für Gruppenvergleich genutzt (z.B. Männer vs. Frauen oder Altersgruppen)

Beispiel: Dauer von Arbeitslosigkeit nach Bildungsabschluss



Die Hazardrate

Die Hazardrate ist ein Mass für die Neigung, den Zustand im Intervall $(t, t + \Delta t)$ zu verlassen, wenn der Zustand bis zum Zeitpunkt t angedauert hat:

$$
\begin{aligned}
h(t) &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t (1 - F(t))} \\
&= \frac{f(t)}{1 - F(t)}
\end{aligned}
$$

Hierbei ist $f(t)$ die Dichte und $F(t)$ die Verteilungsfunktion von T.
Beispiel: Exponentialverteilung (=Verteilung ohne Gedächtnis)
$F(t) = 1 - e^{-\lambda t}$ und $f(t) = F'(t) = \lambda e^{-\lambda t}$:

$$
h(t) = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda
$$

Freie Universität Berlin

Das Proportional Hazard Modell (1/2)

Das häufig nach Cox benannte Proportional Hazard Modell dient der Modellierung des Einflusses von Kovariaten x auf die Hazardfunktion:

$$h(t, x) = h_0(t)e^{x'\beta}$$

Hierbei ist h_0 die aus den Daten geschätzte "Baseline Hazardfunktion". Für unterschiedliche Werte von x ergibt sich:

$$\frac{h(t, x_1)}{h(t, x_2)} = \frac{h_0(t)e^{x_1'\beta}}{h_0(t)e^{x_2'\beta}} = e^{(x_1 - x_2)'\beta}$$

Damit ist der zeitliche Verlauf der beiden Hazardfunktionen proportional zueinander (Modellname!)
Interpretation β_{x_p}: Veränderung von x_p um eine Einheit verändert die Hazardfunktion um den Faktor $e^{\beta_{x_p}}$

Freie Universität Berlin

Das Proportional Hazard Modell (2/2)

Parametrische (Hazard-)Ratenmodelle erhält man, wenn der Verteilungstyp von T vorgegeben wird und die Verteilungsparameter durch $x'\beta$ modelliert werden.
Beispiel: Exponential-Verteilung mit $\lambda = x'\beta$.
Weitere typische Wartezeitverteilungen sind: Gamma-Verteilung, Weibull-Verteilung (Hazard-Funktion ist Polynom!)

Modelle in diskreter Zeit

In diesen Fällen kann T nur bestimmte Werte annehmen, meistens $t = 1, 2, 3, \ldots$ d.h. gleichabständige Zeitpunkte (Monate, Jahre). Die Hazardfunktion lautet in diesem Fall ($\Delta t = 1$!):

$$h(t) = P(T = t + 1 | T > t)$$

d.h. zwischen t und $t + 1$ wird der Zustand gewechselt. Zustandswechsel wird durch Y_t angezeigt:

$$Y_t = \begin{cases} 1, & \text{Zustandswechsel zwischen t und t+1 ;} \\ 0, & \text{sonst.} \end{cases}$$

Damit gilt $h(t) = P(Y_t = 1)$.

Freie Universität Berlin

Das Proportional Odds (Logit) Modell

Das **Proportional Odds Modell** parametrisiert das logarithmierte Chancenverhältnis (Odds):

$$log \frac{P(Y_t = 1|x)}{P(Y_t = 0|x)} = x'\beta$$

Odds Ratio für x_1 und x_2:

$$\frac{\frac{P(Y_t=1|x_1)}{P(Y_t=0|x_1)}}{\frac{P(Y_t=1|x_2)}{P(Y_t=0|x_2)}} = e^{(x_1-x_2)'\beta}$$

Interpretation β_{x_p}: Veränderung von x_p um eine Einheit verändert das Chancenverhältnis um den Faktor $e^{\beta_{x_p}}$

Freie Universität Berlin

Das diskrete Proportional Hazard Modell (1/2))

Die Wartezeiten werden nur diskret gemessen. Mögliche Werte sind $T = 1, 2, \ldots t_{\max} = K$. Die genaue Wartezeit T^* ist unbekannt. Koppelung der beobachteten diskreten Grösse $T \in \{1, 2, \ldots, K\}$ und der nicht beobachteten Grösse $T^* = -x'\beta + \varepsilon$ über ein Schwellenwertmodell:

$$T = k \quad \Leftrightarrow \quad \alpha_{k-1} < T^* \leq \alpha_k$$

Hierbei sind $\alpha_0 = -\infty < \alpha_1 < \ldots < \alpha_K = +\infty$ konstante, aber unbekannte Schwellen und ε folgt einer Fehlerverteilung mit Verteilungsfunktion F.

Freie Universität Berlin

Das Schwellenwertmodell: Konstruktion

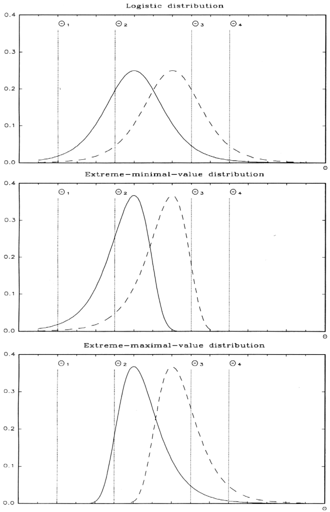


FIGURE 3.1. Densities of the latent response for two subpopulations with different values of x (logistic, extreme–minimal–value, extreme–maximal–value distributions).

Das diskrete Proportional Hazard Modell (2/2)

Mit $F(x) = 1 - \exp(-\exp(x))$ erhält man:

$$h(x) = \frac{f(x)}{1 - F(x)} = \frac{\exp(-\exp(x))\exp(x)}{\exp(-\exp(x))} = \exp(x)$$

Damit erhält man mit
$P(T \leq k|x) = P(T^* \leq \alpha_k|x) = F(\alpha_k + x'\beta)$ für den Wert der Hazardfunktion von ε:

$$h(\alpha_k + x'\beta) = \exp(\alpha_k + x'\beta)$$

Für unterschiedliche Kovariatenwerte x_1 und x_2 erhält man:

$$\frac{h(\alpha_k + x_1'\beta)}{h(\alpha_k + x_2'\beta)} = \exp(\alpha_k + x_1'\beta - (\alpha_k + x_2'\beta)) = \exp((x_1 - x_2)'\beta)$$

Interpretation β_{x_p}: Veränderung von x_p um eine Einheit verändert die Hazardfunktion um den Faktor $e^{\beta_{x_p}}$

Ein alternatives diskretes Hazardratenmodell

$$h(t, x_t) = \frac{e^{\beta_t + x_t'\beta}}{1 + e^{\beta_t + x_t'\beta}} \quad t = 1, 2, \ldots, t_{\max}$$

Eigenschaften:

- Zu jedem Zeitpunkt wird ein Logitmodell für die Hazardrate benutzt.
- Der Einfluss der Kovariaten auf die Hazardraten ist zu jedem Zeitpunkt gleich (Restriktiv!).
- Die Konstante des Logitmodells variiert über die Zeitpunkte.
- Die Anzahl der beobachteten Zeitpunkte ist durch $t_{\max}$ beschränkt.

Diskrete parametrische Wartezeitmodelle

Bisher wurde nur die Hazardrate modelliert. Diskrete parametrische Wartezeitverteilungen sind die

- Geometrische Verteilung (Warten bis zum ersten Erfolg)
- Negative Binomialverteilung (Warten bis zum k. Erfolg)

Parametrisiere z.B. den Erfolgsparameter p (oder besser dessen Logit) durch Kovariaten!
Behandlung von Rechtszensierungen: $Z_i =$ Zensierungsindikator für Episode i. Bei unabhängigen Episodenlängen T_i erhält man für die Likelihood:

$$L = \prod_{Z_i=0} P(T_i = t_i) \prod_{Z_i=1} P(T_i > t_i)$$

Rechtszensierungen kommen bei Panels mit kurzen Laufzeiten häufig vor!

Freie Universität Berlin

Übergangsanalysen (1/3)

Bisher betrachtet: Dauer eines Zustands (Episode,Spell).
Jetzt: Wechsel zwischen Zuständen

- zwischen festem Zeitintervall (t_a, t_e)
- Sequenz von Zuständen ($t = 1, 2, 3, 4$ im MZ-Panel)
- Zustandsraum Z ist geordnet
  (z.B. gross, mittel, klein (Betriebsgrösse),
  Ausbildungsabschluss = keinen, Sek **I**, Sek **II**, Tertiär)
- Zustandsraum ist ungeordnet, z.B.
  Erwerbstätig,**U**nemployed,**N**icht erwerbstätig
- Z_t = Zustand in Welle t (Kalenderzeit)

Freie Universität Berlin

Übergangsanalysen (2/3)

Schätzung von $P(Z_{t+1} = b | Z_t = a)$.

- Basis ist die Übergangsmatrix $(n_{a,b})_{a \in Z, b \in Z}$,
  d.h. Zeilen =Zustand in t_a, Spalten = Zustand in t_e.
- ML-Schätzer unter Multinomialmodell ist: $\frac{n_{a,b}}{n_a}$
  (Reihenprozente !)
  wobei $n_{a,b}$ = Element (a, b) der Übergangsmatrix,
  n_a = Element a der Marginaltabelle für Z_{t_a}.

Übergangsanalysen (2/3)

Table 4: **Bias estimates for flows between labour force states based on the SOEP data (unweighted results).**

Flows		E			U			N		
from 96 to		FULL	IMMO	Δ	FULL	IMMO	Δ	FULL	IMMO	Δ
	97	91.02	91.16	-0.14	4.92	4.86	0.06	4.05	3.97	0.08
E	98	87.82	88.03	-0.21	6.32	6.04	**0.28**	5.86	5.93	-0.07
	99	87.01	86.37	**0.64**	6.04	6.30	-0.26	6.96	7.33	-0.37
	97	32.83	30.85	**1.98**	48.39	49.83	**-1.44**	18.78	19.32	-0.54
U	98	34.92	31.79	**3.13**	40.13	41.20	-1.07	24.95	27.01	-2.06
	99	41.37	37.46	**3.91**	28.91	29.10	-0.19	29.71	33.44	-3.73
	97	12.74	11.64	**1.10**	5.48	4.97	0.51	81.77	83.39	-1.62
N	98	19.66	16.07	**3.59**	5.09	4.40	**0.69**	75.25	79.54	-4.29
	99	25.89	21.13	**4.76**	4.53	3.71	**0.82**	69.58	75.15	-5.57

Δ = estimate of absolute bias

Boldface figures: Significant differences $\hat{P}_{ALL} - \hat{P}_{IMMO}$

Das Multinomiale Logitmodell (1/2)

Modelliert wird der Einfluss von Kovariaten auf einen Übergang $Z_t = a$ (fest) nach Z_{t+1}, d.h. Verteilung auf der a-Zeile der Übergangsmatrix.

- Zustandsraum $(1, 2, \ldots, K)$ ist nicht geordnet.
- $$P(Z_{t+1} = k | Z_t = a, x) = \frac{\exp(x'\beta_k)}{\sum_{l=1}^{K} \exp(x'\beta_l)} \quad k = 1, 2, \ldots, K$$
- Schätzung auf Basis aller Beobachtungen mit $Z_t = a$.
- Einer der Parametervektoren kann auf 0 gesetzt werden (meist $\beta_K = 0$)

Freie Universität Berlin

Das Multinomiale Logitmodell (2/2)

- Interpretation über die Odds:

$$\begin{aligned}\frac{P(Z_{t+1}=k|Z_t=a,x)}{P(Z_{t+1}=l|Z_t=a,x)} &= \frac{\exp(x'\beta_k)}{\exp(x'\beta_l)} \\ &= \exp(x'(\beta_k-\beta_l))\end{aligned}$$

- Interpretation über Random Utility Model:

$$\begin{aligned}U_k &= x'\beta_k+\epsilon_k \\ &= \text{ Nutzen bei Wahl der Alternative k und Vorliegen von x}\end{aligned}$$

Gewählte Alternative $Z_{t+1}=k$ maximiert den Nutzen über die Alternativen U_l. Verteilung der ϵ_k folgt Extremwertverteilung mit $F(x)=\exp(-\exp(-x))$. Die $\epsilon_l \quad (l=1,2,\ldots,K)$ sind unabhängig.

Das Kumulative Logitmodell (1/2)

Jetzt: Zustandsraum ist geordnet.
Koppelung des beobachteten Zustands $Z_{t+1} = Z \in \{1, 2, \ldots, K\}$ und nicht beobachteter Propensities (Neigungen) $Z^* = -x'\beta + \varepsilon$ über ein Schwellenwertmodell:

$$Z = k \quad \Leftrightarrow \quad \alpha_{k-1} < Z^* \leq \alpha_k$$

Hierbei sind $\alpha_0 = -\infty < \alpha_1 < \ldots < \alpha_K = +\infty$ konstante, aber unbekannte Schwellen und ε folgt einer Fehlerverteilung mit logistischer Verteilungsfunktion $F(\varepsilon) = \exp(\varepsilon)/(1 + \exp(\varepsilon))$.

$$P(Z \leq k|x) = F(\alpha_k + x'\beta) \quad k = 1, \ldots, K$$

Freie Universität Berlin

Das Kumulative Logitmodell (2/2)

Interpretation über Odds-Ratio:

$$\frac{\frac{P(Z \leq k | x_1)}{P(Z > k | x_1)}}{\frac{P(Z \leq k | x_2)}{P(Z > k | x_2)}} = \frac{\exp(\alpha_k) \exp(x_1' \beta)}{\exp(\alpha_k) \exp(x_2' \beta)}$$
$$= \exp((x_1 - x_2)' \beta))$$

Odds-Ratio hängt nicht von k ab (restriktiv!).

Die Darstellung von Zustandssequenzen über hierarchische loglineare Modelle (1/4)

Der Zustandsraum von $Z_1 = A, Z_2 = B$ und $Z_3 = C$ besitze 3 Elemente. Eine Zustandssequenz (A, B, C) spannt eine $3 \times 3 \times 3$ Kontingenztabelle auf.
Loglineare Modelle bestimmen die Erwartungswerte $\mu_{a,b,c}^{A,B,C}$ der Kontingenztabelle über:

$$\log(\mu_{a,b,c}^{A,B,C}) = \beta_0 + \beta_a^A + \beta_b^B + \beta_c^C + \beta_{a,b}^{A,B} + \beta_{b,c}^{B,C} + \beta_{a,c}^{A,C} + \beta_{a,b,c}^{A,B,C}$$

β_a^A heißt Haupteffekt der Variablen A (Bez. A).
$\beta_{a,b}^{A,B}$ heisst (2-er) Interaktionseffekt von A und B (Bez. A*B).
$\beta_{a,b,c}^{A,B,C}$ heisst (3-er) Interaktionseffekt von A,B u. C (Bez. A*B*C).

Freie Universität Berlin

Die Darstellung von Zustandssequenzen über hierarchische loglineare Modelle (2/4)

Ein loglineares Modell heisst **hierarchisch**, wenn zu jedem Interaktionsterm höherer Ordnung auch alle Interaktionsterme niederer Ordnung im Modell enthalten sind.
Durch Weglassen von Termen höherer Ordnung lassen sich Aussagen über Unabhängigkeit und bedingte Unabhängigkeit formulieren.

- Gemeinsame Unabhängigkeit:

  $$\underline{\text{Def.}}: \quad \pi_{a,b,c}^{A,B,C} = \pi_a^A \, \pi_b^B \, \pi_c^C \qquad \text{für alle } a, b, c$$

  Modelldarstellung : $A + B + C$

- C unabhängig von A und B:

  $$\text{Def.}: \quad \pi_{a,b,c}^{A,B,C} = \pi_{ab}^{AB} \, \pi_c^C \qquad \text{für alle } a, b, c$$

  Modelldarstellung : $A + B + A * B + C$

Die Darstellung von Zustandssequenzen über hierarchische loglineare Modelle (3/4)

- Bedingte Unabhängigkeit: A und C unabhängig bei gegebenen B Werten

$$\begin{aligned}
\pi_{ac|b}^{AC|B} &= \frac{\pi_{abc}^{ABC}}{\pi_{b}^{B}} \\
&= \pi_{a|b}^{A|B} \, \pi_{c|b}^{C|B} \\
&= \frac{\pi_{ab}^{AB}}{\pi_{b}^{B}} \, \frac{\pi_{cb}^{CB}}{\pi_{b}^{B}}
\end{aligned}$$

Modelldarstellung: $A + B + A * B + C + B * C$

Die Darstellung von Zustandssequenzen über hierarchische loglineare Modelle (4/4)

- Markov Kette für $Z_1 = A, Z_2 = B, Z_3 = C$ und $Z_4 = D$:

$$\begin{aligned} \pi_{abcd}^{ABCD} &= \pi_{d|cba}^{D|CBA} \pi_{c|ba}^{C|BA} \pi_{b|a}^{B|A} \pi_{a}^{A} \\ &= \pi_{d|c}^{D|C} \pi_{c|b}^{C|B} \pi_{b|a}^{B|A} \pi_{a}^{A} \end{aligned}$$

Modelldarstellung: $A + B + C + D + A * B + B * C + C * D$

Fehlende Daten im MZ-Panel

Der MZ ist eine Flächenstichprobe: Personen, die aus den gezogenen Flächen (Auswahlbezirke) wegziehen, verlassen das Panel. Information nach Wegzug/ vor Zuzug fehlt.
Umfang: Pro Jahr ca 12 Prozent aller Personen, über 3 Wiederbefragungen ca 30 Prozent.
Fragestellung: Erzeugt die Beschränkung auf immobile Personen einen Auswertungsbias?
Antwort in vielen Fällen: Ja! (Analyseinstrument SOEP, weil vergleichbarer Fragebogen und Weiterverfolgung mobiler Personen.

Präzisierung der Annahmen über das Zustandekommen fehlender Werte

$R =$ Responseindikator,
$Y =$ Merkmal, das von Ausfallen betroffen ist,
$X =$ Kovariatenvektor, der immer beobachtet wird.

- Missing Completely at Random (MCAR): $P(R|Y,X) = P(R)$
- Missing at Random (MAR): $P(R|Y,X) = P(R|X)$
- Missing not at Random (MNAR): $P(R|Y,X) \neq P(R|X)$

Was kann man tun?

- Ignorieren (nur vollständige Fälle analysieren) bei MCAR
- Kontrollvariablen (z.B. Altersgruppen) bei MAR
- Gewichtungsansatz (z.B. Kehrwert der Mobilitätsw.-keit) bei MAR
- Schätzen der fehlenden Werte (Multiple Imputation) bei MAR
- Selektionsmodelle bei MNAR

Kontrollvariablen (1/2)

Missing at Random (MAR) impliziert:

$$
\begin{aligned}
P(Y|X,R) &= \frac{P(Y,X,R)}{P(X,R)} = \frac{P(R|Y,X)P(Y,X)}{P(X,R)} \\
&= \frac{P(R|X)P(Y,X)}{P(X,R)} = \frac{P(Y,X)}{P(X)} = P(Y|X)
\end{aligned}
$$

Beispiel: Übergang von Arbeitslosigkeit in die Erwerbstätigkeit (1996/1999) (Kontrolle nach Alter)

	FULL	IMMO	Δ
Alter ≤ 30	65.69	64.05	1.64
Alter > 30	30.28	28.78	1.50
Insgesamt	41.37	37.46	3.89

Freie Universität Berlin

Kontrollvariablen (2/2)

Die Kontrollvariablen müssen einen statistischen Zusammenhang zu dem Ereignis "Umzug" haben

	FULL	IMMO	Δ
Ost	41.13	36.44	4.69
West	42.82	39.64	3.18
Insgesamt	41.37	37.46	3.89

Freie Universität Berlin

Gewichtungsansatz (1/4)

- Gewichte jede Beobachtung mit dem Kehrwert der Immobilitätswahrscheinlichkeit
- Beispiel: Übergang von Arbeitslosigkeit in die Erwerbstätigkeit (1996/1999)

	FULL	IMMO	GEW
X=Alter	41.37	37.46	40.39
X=Region	41.37	37.46	37.84

Freie Universität Berlin

Gewichtungsansatz (2/4)

- Verallgemeinerung des Gewichtungsansatzes auf Schätzung von Modellen
- Beispiel: Schätzung eines Logit-Modells

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = X_i'\beta, \quad \text{mit} \quad \pi_i = P(Y_i = 1|X_i)$$

wobei das Untersuchungsmerkmal:

$$Y_i = \begin{cases} 1 & \text{Person } i \text{ wird erwerbstätig} \\ 0 & \text{Person } i \text{ wird nicht erwerbstätig} \end{cases}$$

Vollständige Daten:

$$U^{com}(\beta) = \sum_{i=1}^{n} X_i(Y_i - \pi_i) = 0, \quad \text{mit} \quad \pi_i = \frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}$$

Gewichtungsansatz (3/4)

- Annahme: Die erklärenden Variablen X aus dem Untersuchungsmodell reichen nicht aus, um den Einfluss von Y auf Mobilität zu entfernen
- Annahme: Es gibt weitere Variablen Z, die den Einfluss von Y auf Mobilität herausnehmen
- Gewichte die Schätzgleichung des Modells mit dem Kehrwert der Immobilitätswahrscheinlichkeit auf Basis von X und Z
- Ausweg: gewichtete Schätzgleichung

$$U(\beta) = \sum_{i=1}^{n} R_i \, \frac{1}{\hat{P}(R_i = 1|X_i, Z_i)} \, (Y_i - \pi_i) = 0$$

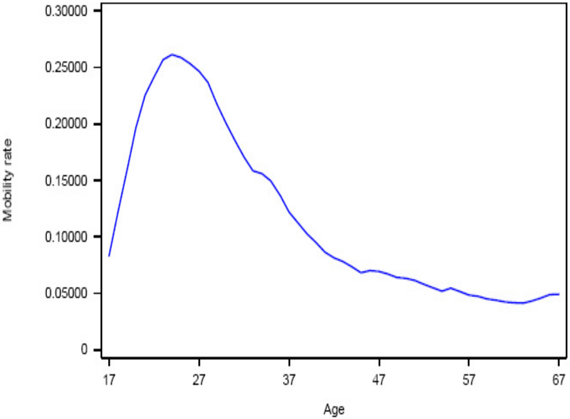
wobei R_i Mobilitätsindikator mit

$$R_i = \begin{cases} 1 & \text{falls, Person } i \text{ immobil} \\ 0 & \text{falls, Person } i \text{ mobil} \end{cases}$$

Gewichtungsansatz (4/4)

Praktische Umsetzung

- Information über Befragungsstatus einer Person im Datensatz vorhanden (Variable PERKL + VERLUSTE bzw. GEWINNE)
- Alter wichtigste Variable für Mobilität



Beispiel Gewichtungsansatz

Gewichte jede Beobachtung mit dem Kehrwert der Immobilitätswahrscheinlichkeit aus einem Logit Modell mit vielen erklärenden Variablen.
Beispiel: Übergang von Arbeitslosigkeit in die Erwerbstätigkeit (1996/1999)

	FULL	IMMO	GEW
X=Alter	41.37	37.46	40.39
X=Region	41.37	37.46	37.84
Logit Modell	41.37	37.46	40.76

Selektionsmodelle (1/4)

- $Z_{t_1} = A$ Erwerbsstatus zur Zeit t_1 (vollständig beobachtet), $Z_{t_2} = B$ Erwerbsstatus zur Zeit t_2 (unvollständig beobachtet)
- Von Interesse ist: $P(B|A)$

	$R=1$			R=0
		t_2		
t_1	E	U	N	
E	$n(EE)$	$n(EU)$	$n(EN)$	$n(E.)$
U	$n(UE)$	$n(UU)$	$n(UN)$	$n(U.)$
N	$n(NE)$	$n(NU)$	$n(NN)$	$n(N.)$

$$L = \prod_{i \in R=1} P(A,B)P(R=1|A,B) \times \prod_{i \in R=0} \sum_{B} P(A,B)P(R=0|A,B)$$

Freie Universität Berlin

Selektionsmodelle (2/4)

$$P(R=1|B,A) = \begin{cases} P(R=1|A) & \text{MAR;} \\ P(R=1|B) & \text{Restricted NMAR;} \\ P(R=1|A,B,A*B) & \text{Unrestricted NMAR.} \end{cases}$$

Das unrestringierte Selektionsmodell ist nicht schätzbar, da insgesamt 17 freie Parameter
(A (2), $B|A$ (3*2),$R|A,B,A*B$(9))
mit 12 Kontingenztabellenfeldern
($A*B|R=1$ (9) $A|R=0$ (3)) geschätzt werden müssen.

Setzen von Restriktionen

Beispiel: Veränderung zwischen den Zuständen **e**mployed (=1), **u**nemployed (=2) und **n**ot employed (=3).
Hypothese: Es gibt drei Gruppen mit hoher, mittlerer und niedriger Mobilität! Die Mobilität hängt von den Übergängen $a \longrightarrow b$ ab!

- Gruppe hohe Mobilität: $A = 2(u), B = 1(e)$ und $A = 3(n), B = 1, 2(e, u)$
- Gruppe niedrige Mobilität: $A = 2(u), B = 3(n)$ und $A = 3(n), B = 3(n)$
- Gruppe mittlere Mobilität: $A = 1(e)$ und $B = 1, 2, 3(e, u, n)$ und $A = 2(u), B = 2(u)$

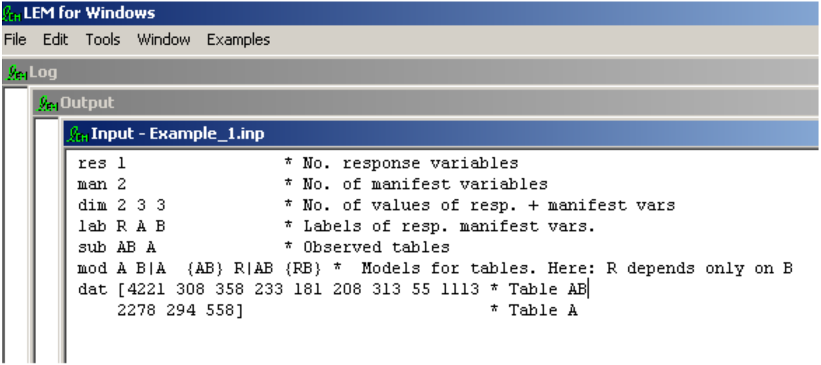
LEM: A useful program

LEM stands for: **L**oglinear and event history analysis with missing data using the **EM** algorithm.

Free download + documentation from:
http://www.uvt.nl/faculteiten/fsw/organisatie/departementen/mto/software2.html

LEM: Example 1 with SOEP data

$$P(R|A, B) = P(R|B)$$




```
res 1                    * No. response variables
man 2                    * No. of manifest variables
dim 2 3 3                * No. of values of resp. + manifest vars
lab R A B                * Labels of resp. manifest vars.
sub AB A                 * Observed tables
mod A B|A  {AB} R|AB {RB} *  Models for tables. Here: R depends only on B
dat [4221 308 358 233 181 208 313 55 1113 * Table AB
    2278 294 558]                         * Table A
```

LEM: Example 2 with SOEP data

Medium mobility group: $A = 1(e)$ and $B = 1, 2, 3(e, u, n)$ and $A = 2(u), B = 2(u)$

High mobility group: $A = 2(u), B = 1(e)$ and $A = 3(n), B = 1, 2(e, u)$

Low mobility group: $A = 2(u), B = 3(n)$ and $A = 3(n), B = 3(n)$

Input - Example_2.inp

```
res 1
man 2
dim 2 3 3
lab R A B
sub AB A
mod A B|A
R|AB {fac(ABR,3)}        * 3 Restrictions for the ABR table
des [0 0 0 0 0 0 0 0 0   * No restrictions for the AB table
     1 1 1 2 1 3 2 2 3] * Parameters with 1 set to be equal, 2 and 3 similar
dat [4221 308 358 233 181 208 313 55 1113
     2278 294 558]
```

Table 11: **Estimation of flows between labour force states. Control by age. Correction of estimates by different model alternatives.** alt_1 **: transitions** $U \rightarrow N$ **attributed to the low mobility group .** alt_2 **: transitions** $U \rightarrow N$ **attributed to the mean mobility.** alt_3 **: Main effect model for B**

Transition	$U \rightarrow E$					$N \rightarrow E$				
	ALL	IMMO	alt_1	alt_2	alt_3	ALL	IMMO	alt_1	alt_2	alt_3
	Age≤30									
97	52.43	52.12	52.39	51.46	53.09	25.98	24.16	25.33	24.88	25.67
	(2.84)	(3.10)	(3.53)	(3.46)	(3.64)	(1.56)	(1.64)	(2.35)	(2.20)	(1.82)
98	55.09	56.02	57.78	55.29	57.64	37.86	33.33	37.50	35.89	37.42
	(2.95)	(3.59)	(4.50)	(4.22)	(5.04)	(1.80)	(2.06)	(4.45)	(3.78)	(2.68)
99	65.69	64.05	65.65	62.49	68.39	50.07	46.28	51.37	48.91	53.57
	(2.87)	(3.88)	(5.31)	(4.62)	(6.87)	(1.92)	(2.44)	(6.85)	(5.34)	(3.85)
	Age>30									
97	24.02	22.04	24.19	23.41	21.63	6.36	6.13	6.97	6.75	6.24
	(1.63)	(1.66)	(1.96)	(1.91)	(1.65)	(0.60)	(0.61)	(0.82)	(0.77)	(0.62)
98	25.90	23.25	27.45	24.98	22.66	10.13	8.81	11.14	10.14	9.19
	(1.74)	(1.81)	(2.48)	(2.31)	(1.84)	(0.81)	(0.80)	(1.46)	(1.21)	(0.85)
99	30.28	28.78	36.04	30.70	27.79	12.72	11.28	15.73	13.40	12.13
	(1.87)	(2.09)	(2.85)	(2.64)	(2.19)	(0.95)	(0.97)	(2.06)	(1.56)	(1.06)
	Total									
97	32.84	30.85	32.08	31.24	30.68	12.74	11.64	12.54	12.21	11.99
	(1.49)	(1.55)	(1.83)	(1.78)	(1.60)	(0.68)	(0.68)	(0.93)	(0.87)	(0.71)
98	34.92	31.79	34.55	32.26	31.57	19.66	16.07	18.48	17.26	16.89
	(1.57)	(1.72)	(2.41)	(2.16)	(1.86)	(0.86)	(0.87)	(1.71)	(1.39)	(0.95)
99	41.37	37.46	41.74	37.74	37.25	25.89	21.13	25.19	22.78	22.48
	(1.66)	(1.94)	(2.46)	(2.45)	(2.24)	(1.00)	(1.06)	(2.54)	(1.84)	(1.20)

Source: Authors' calculations, Data base: SOEP, Waves: 1996-1999
Standard error in parenthesis

Vorsicht ist geboten!!

- Trotz der Benutzung von mehr Tabellen ($A * B|R = 1$ **und** $A|R = 0$ und zusätzlicher Restriktionen vergrössert sich die Standardabweichung der Schätzwerte!
- Grund: Flache Likelihood!
- Oft substantielle Überkorrekturen!