Statistical Matching

Seminar "Statistische Herausforderungen im Umgang mit fehlenden bzw. fehlerbehafteten Daten"

Katrin Hummrich

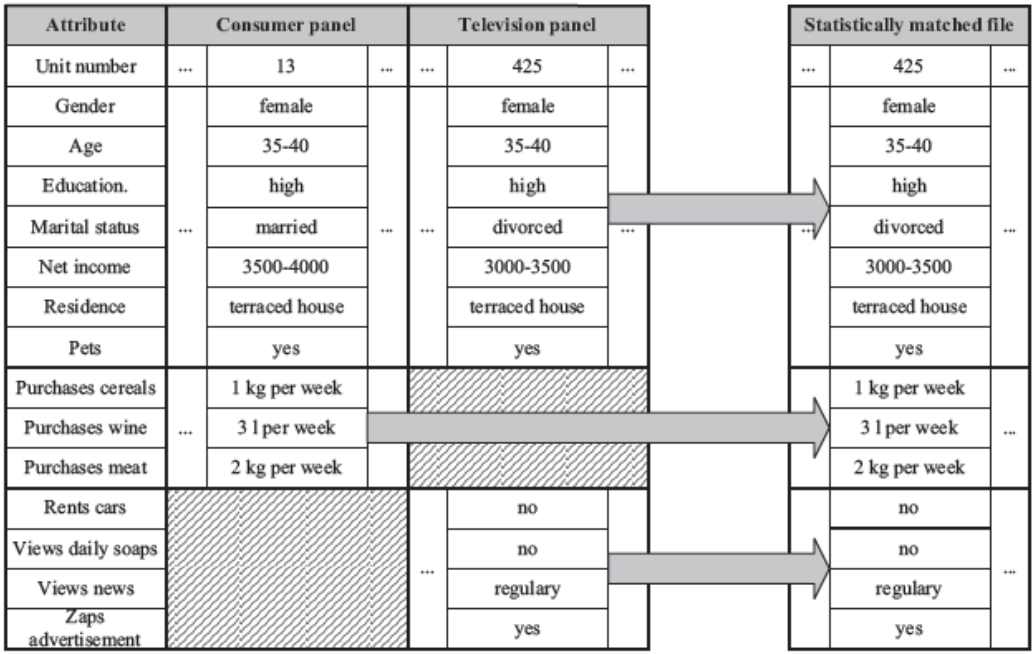
Statistik Institut der LMU

05. Dezember 2014

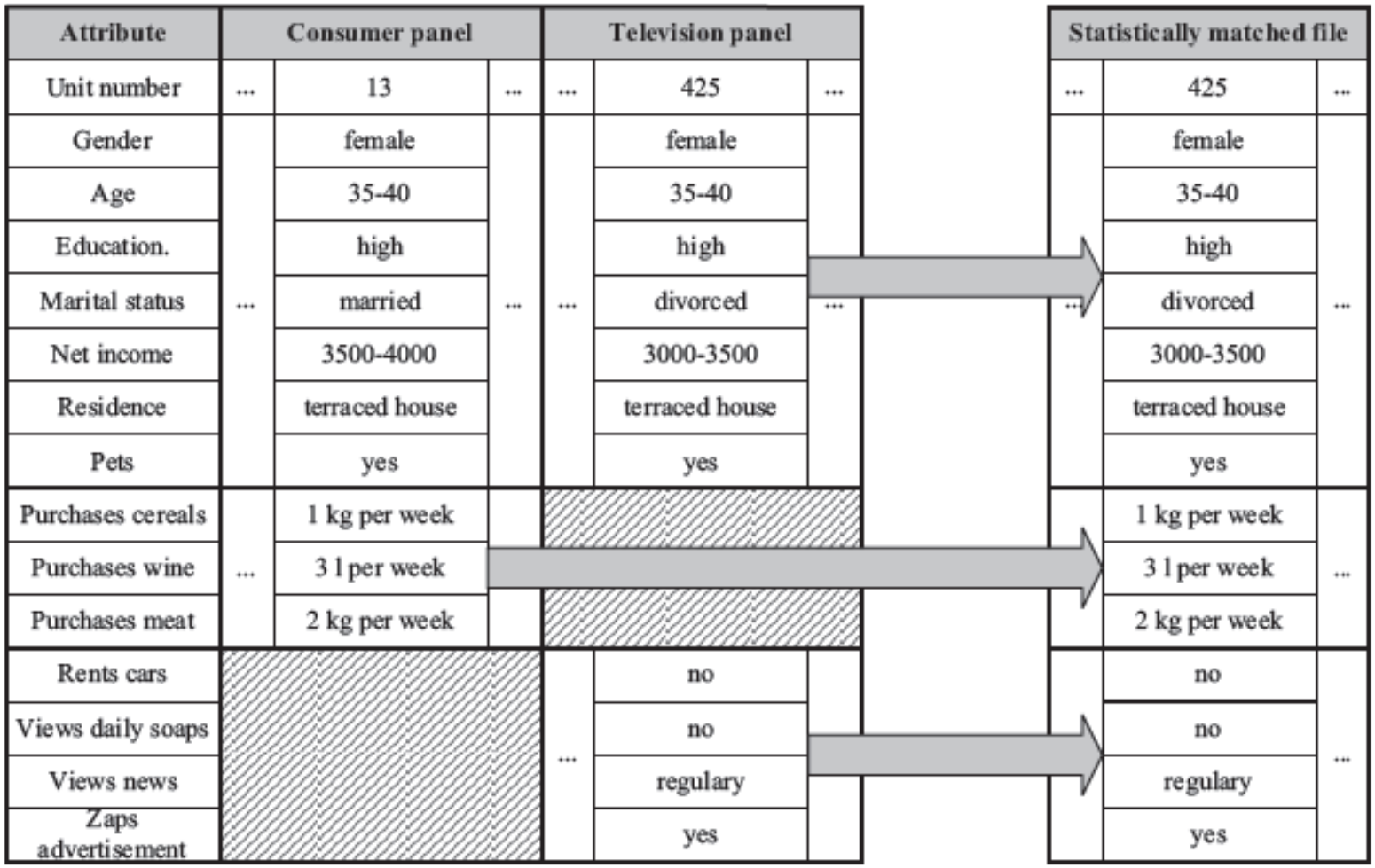
Inhaltsverzeichnis

Worum geht es beim statistischen Matching?

- KEIN finden von statistischen Zwillingen innerhalb eines Datensatzes zum Beispiel um Treatment-Evaluationsproblematik zu lösen
- Zusammenführen von zwei oder mehr Datensätzen
- KEIN Record Linkage
- hier existieren nicht dieselben Objekte $\rightarrow$ Zusammenführen anhand von Matchingvariablen



Worum geht es beim statistischen Matching? - Beispiel



Attribute	Consumer panel	Television panel	Statistically matched file
Unit number	13	425	425
Gender	female	female	female
Age	35-40	35-40	35-40
Education.	high	high	high
Marital status	married	divorced	divorced
Net income	3500-4000	3000-3500	3000-3500
Residence	terraced house	terraced house	terraced house
Pets	yes	yes	yes
Purchases cereals	1 kg per week		1 kg per week
Purchases wine	3 l per week		3 l per week
Purchases meat	2 kg per week		2 kg per week
Rents cars		no	no
Views daily soaps		no	no
Views news		regulary	regulary
Zaps advertisement		yes	yes

Abb. : Matchingbeispiel aus der Wirtschaft (Noll (2009), S. 11)

Warum statistisches Matching?

Gründe für statistisches Matching können sein:

- Zeit
- Geld
- mehr fehlende Werte bei zu langen Fragebögen
- niedrigere Teilnahmebereitschaft bei zu vielen Umfragen

Ziel des statistischen Matchings: möglichst viel Information aus den bereits vorhanden Datenquellen schöpfen

(Vgl. D'Orazio et al. (2006), S. 1)

Die Datensituation I

Der Einfachheit halber werden lediglich zwei Datensätze A und B betrachtet. A enthält die Variablen $\mathbf{X}$ und $\mathbf{Y}$. B enthält die Variablen $\mathbf{X}$ und $\mathbf{Z}$. Außerdem gilt:

- $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ haben die gemeinsame Dichte $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$, wobei f aus der Verteilungsfamilie $\mathcal{F} = \{f\}$
- $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{Y}$ und $\mathbf{z} \in \mathcal{Z}$
- Die Zufallsvariablen haben die die Dimensionen P, Q und R $\mathbf{X} = (X_1, ..., X_p)^T$, $\mathbf{Y} = (Y_1, ..., Y_q)^T$ und $\mathbf{Z} = (Z_1, ..., Z_r)^T$
- Annahme: n_A Beobachtungen aus A und n_B Beobachtungen aus B sind i.i.d und stammen aus einer Verteilung mit Dichte $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$
- A enthält die beobachteten Werte $(\mathbf{x}_a, \mathbf{y}_a) = (x_{a1}, ..., x_{ap}, y_{a1}, ..., y_{aq})$ und B die beobachteten Werte $(\mathbf{x}_b, \mathbf{z}_b) = (x_{b1}, ..., x_{bp}, z_{b1}, ..., z_{br})$

(Vgl. D'Orazio et al. (2006), S. 3)

Die Datensituation II

Datensatz der Vereinigung $A \cup B$ mit $n_A + n_B$ i.i.d. Beobachtungen aus $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ hat folgende charakteristische Eigenschaften:

- das Auftreten von fehlenden Daten und daraus resultierenden Mechanismen
- der Mangel an gemeinsamen Informationen über **X**, **Y** und **Z**

(Vgl. D'Orazio et al. (2006), S. 4)

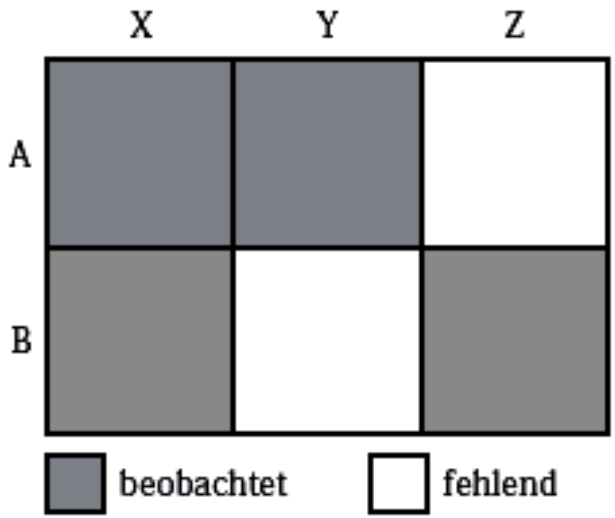


Abb. : Datensituation beim statistischen Matching (Meinfelder (2013), S. 85)

Die Datensituation III

Das statistische Matching befasst sich mit der zweiten Eigenschaft. Um die Besonderheit des nicht-vollständig beobachteten Datensatzes darstellen zu können, noch ein paar Notationen:

- die Zufallsvariable $\mathbf{R} = (\mathbf{R}_x, \mathbf{R}_y, \mathbf{R}_z)$ mit $\mathbf{R}_x = (R_{X_1}, ..., R_{X_p})^T$, $\mathbf{R}_y = (R_{Y_1}, ..., R_{Y_q})^T$ und $\mathbf{R}_z = (R_{Z_1}, ..., R_{Z_r})^T$ gibt Auskunft darüber welche Beobachtungen vorhanden sind bzw. fehlen
- für Missing Mechanismus interessant bedingte Dichte $h(\mathbf{r}_x, \mathbf{r}_y, \mathbf{r}_z | \mathbf{x}, \mathbf{y}, \mathbf{z})$
- unter MCAR gilt

$$h(\mathbf{r}_x, \mathbf{r}_y, \mathbf{r}_z | \mathbf{x}, \mathbf{y}, \mathbf{z}) = h(\mathbf{r}_x, \mathbf{r}_y, \mathbf{r}_z) \tag{1}$$

(Vgl. D'Orazio et al.(2006), S. 6 und vgl. Meinfelder(2013), S. 84)

Die Datensituation IV

- wegen der Symmetrie von Unabhängigkeit gilt auch $\phi(\mathbf{x}, \mathbf{y}, \mathbf{z}|\mathbf{r}_x, \mathbf{r}_y, \mathbf{r}_z) = \phi(\mathbf{x}, \mathbf{y}, \mathbf{z})$
- wegen der Struktur von A und B gibt es nur zwei mögliche Beobachtungsmuster für $\mathbf{R}$ und da für die $n_A + n_B$ Beobachtungen aus $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ i.i.d angenommen wird, gilt:

$$\phi(\mathbf{x}, \mathbf{y}, \mathbf{z}|\mathbf{1}_P, \mathbf{1}_Q, \mathbf{0}_R) = \phi(\mathbf{x}, \mathbf{y}, \mathbf{z}|\mathbf{1}_P, \mathbf{0}_Q, \mathbf{1}_R) = f(\mathbf{x}, \mathbf{y}, \mathbf{z}), \quad (2)$$

  dh. es liegt MCAR vor
- damit lässt sich die beobachtete Stichprobenverteilung der $n_A + n_B$ Beobachtungseinheiten berechnen durch

$$\prod_{a=1}^{n_A} f_{\mathbf{XY}}(\mathbf{x}_a, \mathbf{y}_a) \prod_{b=1}^{n_B} f_{\mathbf{XZ}}(\mathbf{x}_b, \mathbf{z}_b) \quad (3)$$

(Vgl. D'Orazio et al. (2006), S. 6-7)

Vier wichtige Fragen

(a) Welche Annahmen für das gemeinsame Modell $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ können vernünftig in Betracht gezogen werden? $\rightarrow$ **Modellannahmen**

(b) Welcher Schätzer unter allen, die den Modellannahmen aus (a) entsprechen, ist für die Dichtefunktion $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ zu bevorzugen? $\rightarrow$ **Genauigkeit des Schätzers**

(c) Welche Methoden können verwendet werden um passende Werte für die fehlenden Variablen zu erzeugen, die zum gewählten Modell aus (a) und dem gewählten Schätzer aus (b) passen? $\rightarrow$ **Repräsentativität des gematchten Datensatzes**

(d) Welche Inferenzverfahren können auf den durch statistisches Matching erhaltenen Datensatz angewendet werden? $\rightarrow$ **Genauigkeit der Schätzer basierend auf dem gematchten Datensatz**

(Vgl. D'Orazio et al. (2006), S. 8)

Die bedingte Unabhängigkeitsannahme (CIA)

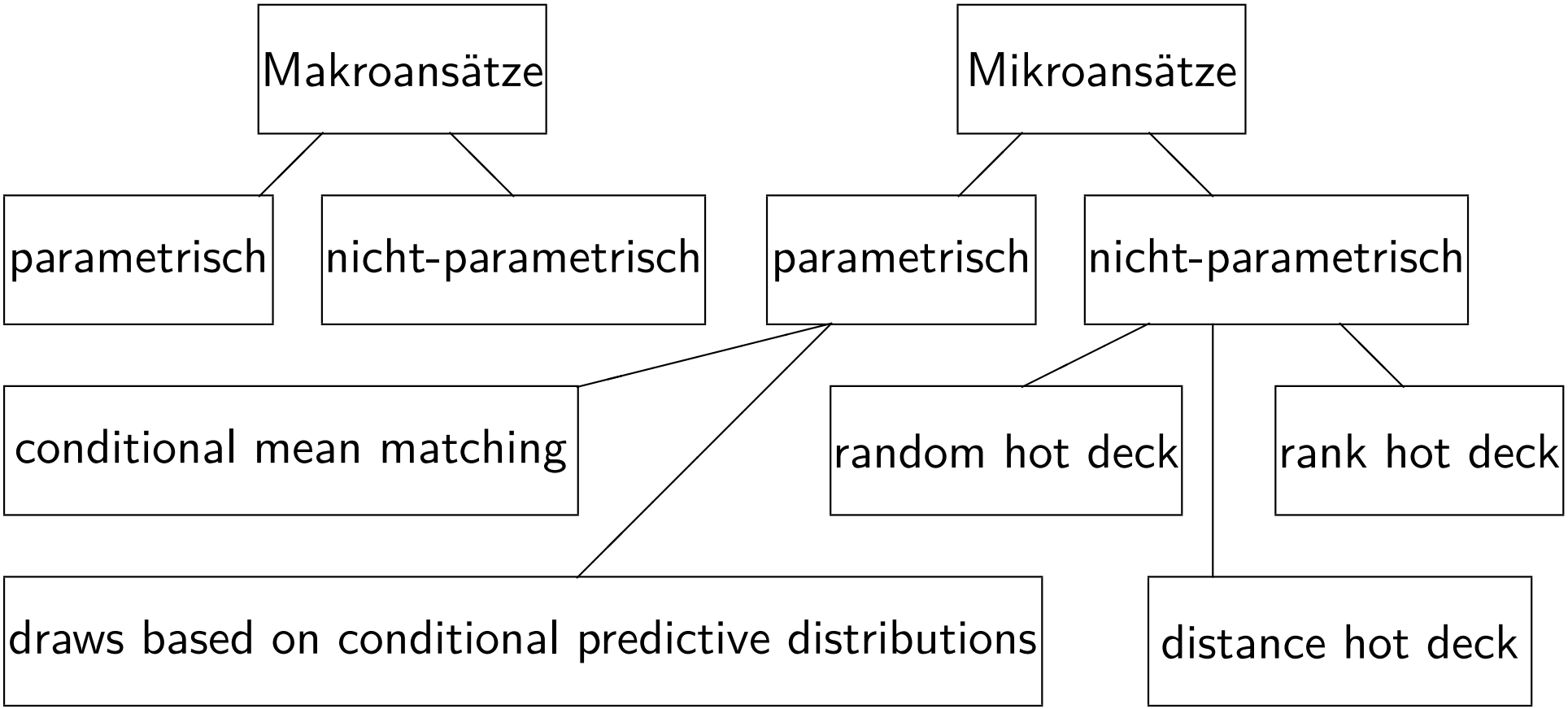
Eine sehr wichtige, aber durchaus restriktive Annahme, die es ermöglicht $A \cup B$ zu identifizieren und direkt zu schätzen und vielen Verfahren des statistischen Matchings zu Grunde liegt. Trifft diese zu gilt für die Dichte von $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$:

$$f(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}), \tag{4}$$

- in (4) werden lediglich Informationen über die marginale Verteilung von $\mathbf{X}$ und den paarweisen Beziehungen von $\mathbf{X}$ und $\mathbf{Y}$ sowie von $\mathbf{X}$ und $\mathbf{Z}$ gebraucht $\rightarrow$ die Informationen aus A und B reichen für die Schätzung der gemeinsamen Verteilung aus
- ACHTUNG: diese Annahme kann anhand $A \cup B$ **nicht** getestet werden!

(Vgl. D'Orazio et al. (2006), S. 13)

Matchingmethoden



Parametrischer Makroansatz

Ziel ist die Schätzung von $(\boldsymbol{\theta_X}, \boldsymbol{\theta_{Y|X}}, \boldsymbol{\theta_{Z|X}})$.

- $f(\mathbf{x}, \mathbf{y}, \mathbf{z}; \boldsymbol{\theta}) \in \mathcal{F}$ mit $\mathcal{F}$ als parametrische Verteilungsfamilie
- analog zu (4) und unter CIA $\mathcal{F}$ faktorisierbar in $\mathcal{F}_{\mathbf{X}}$, $\mathcal{F}_{\mathbf{Y}|\mathbf{X}}$ und $\mathcal{F}_{\mathbf{Z}|\mathbf{X}}$
- Verteilung von $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ durch Parametervektoren $\boldsymbol{\theta_X}$, $\boldsymbol{\theta_{Y|X}}$ und $\boldsymbol{\theta_{Z|X}}$ darstellbar als:

$$f(\mathbf{x}, \mathbf{y}, \mathbf{z}; \boldsymbol{\theta}) = f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta_X}) f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta_{Y|X}}) f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta_{Z|X}}), \tag{5}$$

- ML-Schätzung anhand $A \cup B$ möglich durch:

$$\begin{aligned} L(\boldsymbol{\theta}|A \cup B) &= \prod_{a=1}^{n_A} f_{\mathbf{XY}}(\mathbf{x}_a, \mathbf{y}_a; \boldsymbol{\theta}) \prod_{b=1}^{n_B} f_{\mathbf{XZ}}(\mathbf{x}_b, \mathbf{z}_b; \boldsymbol{\theta}) \\ &= \prod_{a=1}^{n_A} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}_a|\mathbf{x}_a; \boldsymbol{\theta}_{\mathbf{Y}|\mathbf{X}}) \prod_{b=1}^{n_B} f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}_b|\mathbf{x}_b; \boldsymbol{\theta}_{\mathbf{Z}|\mathbf{X}}) \qquad (6) \\ &\quad \cdot \prod_{a=1}^{n_A} f_{\mathbf{X}}(\mathbf{x}_a; \boldsymbol{\theta}_{\mathbf{X}}) \prod_{b=1}^{n_B} f_{\mathbf{X}}(\mathbf{x}_b; \boldsymbol{\theta}_{\mathbf{X}}). \end{aligned}$$

(Vgl. D'Orazio et al. (2006), S. 14)

Nicht-parametrischer Makroansatz

Ist dann vorzuziehen, wenn nicht ausreichend Informationen vorhanden sind, um $\mathcal{F}$ einer parametrischen Verteilungsfamilie zuzuordnen.

- Möglichkeit 1: nutze $f(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x})$ und schätze die Einzeldichten nicht-parametrisch durch Kerndichteschätzung oder Nächste-Nachbarn-Klassifikation (kNN von k nearest neighbour)
- Möglichkeit 2 wenn $\mathbf{X}$ kategorial ist: schätze die empirische kumulative Verteilungsfunktion mit Hilfe der gemeinsamen kumulativen Verteilungsfunktion von $(\mathbf{Y}, \mathbf{Z})$ gegeben $\mathbf{X}$

$$\mathbf{F}_{\mathbf{YZ}|\mathbf{X}}(\mathbf{y}, \mathbf{z}|\mathbf{x}) = \int_{\mathbf{t} \leq \mathbf{y}} \int_{\mathbf{v} \leq \mathbf{z}} f_{\mathbf{YZ}|\mathbf{X}}(\mathbf{t}, \mathbf{v}|\mathbf{x}) d\mathbf{t} d\mathbf{v} \tag{7}$$

Nicht-parametrischer Makroansatz - Fortsetzung

- unter CIA gilt folgende, praktische Zergliederung

$$\mathbf{F_{YZ|X}(y,z|x) = F_{Y|X}(y|x)F_{Z|X}(z|x)}, \tag{8}$$

wobei die einzelnen Faktoren sich durch

$$\widehat{\mathbf{F}}_{\mathbf{Y|X}}(\mathbf{y}|\mathbf{x}) = \frac{\sum_{a=1}^{n_A} \mathbf{I(y_a \leq y)I(x_a = x)}}{\sum_{a=1}^{n_A} \mathbf{I(x_a = x)}}, \tag{9}$$

$$\widehat{\mathbf{F}}_{\mathbf{Z|X}}(\mathbf{z}|\mathbf{x}) = \frac{\sum_{b=1}^{n_B} \mathbf{I(z_b \leq z)I(x_b = x)}}{\sum_{b=1}^{n_B} \mathbf{I(x_b = x)}} \tag{10}$$

schätzen lassen

(Vgl. D'Orazio et al. (2006), S. 31-33)

Parametrische Mikroansätze - conditional mean matching

- Für $\mathbf{Y}$ und $\mathbf{Z}$ stetig kann jeder zu imputierende Werte durch

$$\tilde{\mathbf{z}}_a = E(\mathbf{Z}|\mathbf{X} = \mathbf{x}_a) = \int_{\mathcal{Z}} \mathbf{z} f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}_a; \boldsymbol{\theta}_{\mathbf{Z}|\mathbf{X}}) d\mathbf{z}, \qquad a = 1, ..., n_A \quad (11)$$

bzw.

$$\tilde{\mathbf{y}}_b = E(\mathbf{Y}|\mathbf{X} = \mathbf{x}_b) = \int_{\mathcal{Y}} \mathbf{y} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}_b; \boldsymbol{\theta}_{\mathbf{Y}|\mathbf{X}}) d\mathbf{y}, \qquad b = 1, ..., n_B \quad (12)$$

ersetzt werden $\rightarrow$ das entspricht den Werten aus der geschätzten Regressionsfunktion von $\mathbf{Z}$ bzw. $\mathbf{Y}$ auf $\mathbf{X}$ (ersetze $\boldsymbol{\theta}_{\mathbf{Z}|\mathbf{X}}$ und $\boldsymbol{\theta}_{\mathbf{Y}|\mathbf{X}}$ durch zugehörige ML-Schätzer)

- **Nachteile**: 1) prognostizierte Werte sind keine beobachteten Werte 2) Unterschätzung der Varianz $\Rightarrow$ resultierender Datensatz nicht (approximativ) repräsentativ für $f(\mathbf{x}, \mathbf{y}, \mathbf{z}; \boldsymbol{\theta})$

(Vgl. D'Orazio et al. (2006), S. 26 und S. 30-31)

Parametrische Mikroansätze - draws based on conditional predictive distributions

Durch zufällige Ziehungen aus bedingten Vorhersageverteilungen soll die datengenerierende multivariate Verteilung besser bestimmt werden.

- dh. für alle $a = 1, ..., n_A$ wird ein zufälliger Wert aus $f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}_a; \hat{\boldsymbol{\theta}}^{(ML)}_{\mathbf{Z}|\mathbf{X}})$ gezogen
  und für alle $b = 1, ..., n_B$ aus $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}_b; \hat{\boldsymbol{\theta}}^{(ML)}_{\mathbf{Y}|\mathbf{X}})$
  $\rightarrow$ dies geht nur unter Annahme von MAR
- wegen Konsistenzeigenschaft vom ML-Schätzer entspricht $\hat{\boldsymbol{\theta}}$ approximativ $\boldsymbol{\theta}$, dh. resultierender Datensatz approximativ repräsentativ für $f(\mathbf{x}, \mathbf{y}, \mathbf{z}; \boldsymbol{\theta})$

(Vgl. D'Orazio et al. (2006), S. 29-31)

Nicht-parametrische Mikroansätze - hot deck Methoden

- fehlende Werte werden durch beobachtete Werte ersetzt
- meist gibt es einen Spender- (hier B) und einen Empfängerdatensatz (hier A)
- Wahl von Spender-/Empfängerdatensatz hängt von mehreren Faktoren ab, ist ein Datensatz erheblich kleiner, dient dieser als Empfänger
- wichtige Voraussetzung: beide Datensätze stammen aus der gleichen Verteilung $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$

(Vgl. D'Orazio et al. (2006), S. 34-35)

Nicht-parametrische Mikroansätze - random hot deck

Die random hot deck Methode wählt für jeden fehlenden Wert aus dem Empfängerdatensatz zufällig einen Eintrag aus dem Spenderdatensatz aus.

- $\widehat{=}$ Schätzung der marginalen Verteilung von $\mathbf{Z}$ in B, was implizit annimmt, dass $\mathbf{Z}$ und $\mathbf{X}$ unabhängig sind
- oft gibt es eine oder mehrere kategoriale Variablen $\mathbf{X}$, anhand der, die Datensätze in Untergruppen unterteilt werden kann und die zufällige Ziehung dann innerhalb dieser Gruppen stattfindet
  $\rightarrow \widehat{=}$ Schätzung der bedingten Verteilung von $\mathbf{Z}$ gegeben $\mathbf{X}$ in B und einer zufälligen Ziehung daraus

(Vgl. D'Orazio et al. (2006), S. 38-39)

Nicht-parametrische Mikroansätze - rank hot deck

Die rank hot deck Methode kann angewendet werden, wenn eine ordinale Matchingvariable X vorhanden ist.

- ordne dazu beide Datensätze bzgl. X
- betrachte dann die empirischen kumulativen Verteilungsfunktionen von X im Empfängerdatensatz

$$\hat{F}_X^A(x) = \frac{1}{n_A} \sum_{a=1}^{n_A} I(x_A \leq x), \qquad x \in \mathcal{X}$$

und im Spenderdatensatz

$$\hat{F}_X^B(x) = \frac{1}{n_B} \sum_{b=1}^{n_B} I(x_B \leq x), \qquad x \in \mathcal{X}$$

- jedem $a = 1, .., n_A$ wird dann der Eintrag b^* aus B zugeordnet, für den gilt

$$|\hat{F}_X^A(x_a) - \hat{F}_X^B(x_{b^*})| = \min_{1 \leq b \leq n_B} |\hat{F}_X^A(x_a) - \hat{F}_X^B(x_b)|$$

(Vgl. D'Orazio et al. (2006), S. 39)

Nicht-parametrische Mikroansätze - distance hot deck

Bei der distance hot deck Methode wird jedem Eintrag aus dem Empfängerdatensatz der Eintrag aus dem Spenderdatensatz zugeordnet, der den kleinsten Abstand bezüglich der Matchingvariablen **X** hat.

- im einfachsten Fall, mit lediglich einer einzigen stetigen Matchingvariablen X, wird der Spender b^* für den a-ten Eintrag aus A so gewählt, dass für die Distanz d_{ab^*} gilt

$$d_{ab^*} = |x_a - x_{b^*}| = \min_{1 \leq b \leq n_B} |x_a - x_b| \tag{13}$$

- finden sich mehrere Spender mit dem gleichen Abstand, wird im Allgemeinen einer davon zufällig ausgewählt
- diese Definition entspricht der *unconstrained* Version

(Vgl. D'Orazio et al. (2006), S. 41-42)

distance hot deck - Fortsetzung I

- bei der *constrained* Variante dient jeder Eintrag aus B nur ein Mal als Spender $\Rightarrow n_A \leq n_B$ muss gelten
- $n_A = n_B$: Spenderzuordnung so, dass

$$\sum_{a=1}^{n_A} \sum_{b=1}^{n_B} (d_{ab} w_{ab}) \tag{14}$$

unter folgenden Nebenbedingungen minimiert wird:

$$\sum_{b=1}^{n_B} w_{ab} = 1, \qquad a = 1, ..., n_A, \tag{15}$$

$$\sum_{a=1}^{n_A} w_{ab} = 1, \qquad b = 1, ..., n_B, \tag{16}$$

mit $w_{ab} \in \{0; 1\}$, wobei $w_{ab} = 1$, wenn das Paar (a, b) gematcht wurde und $w_{ab} = 0$, wenn a und b nicht einander zugeordnet wurden

(Vgl. D'Orazio et al. (2006), S. 42)

distance hot deck - Fortsetzung II

- Nebenbedingungen bei $(n_B > n_A)$:

$$\sum_{b=1}^{n_B} w_{ab} = 1, \qquad a = 1, ..., n_A, \tag{17}$$

$$\sum_{a=1}^{n_A} w_{ab} \leq 1, \qquad b = 1, ..., n_B, \tag{18}$$

  mit $w_{ab} \in \{0; 1\}$ und implizieren, dass $\sum_{a=1}^{n_A} \sum_{b=1}^{n_B} w_{ab} = n_A$ gilt

- **Hauptvorteil** der constrained Variante: die marginale Verteilung von der imputierten Variable **Z** bleibt im gematchten Datensatz erhalten
- **Nachteil**: die durchschnittliche Distanz zwischen Spender- und Empfängerwert in der Matchingvariable X ist erwartungsgemäß größer als bei der unconstrained Variante $\rightarrow$ matching noise

(Vgl. D'Orazio et al. (2006), S. 42-43)

hot deck Methoden - Anmerkungen

- wichtige Frage: stammen die Daten, die durch die hot deck Methoden erzeugt werden, wirklich aus der wahren, aber unbekannten Verteilung mit Dichtefunktion $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$?
- für endliche Stichproben kann gezeigt werden, dass alle hot deck Methoden Datensätze erzeugen, die aus einer Verteilung stammen, die sich von der wahren unterscheidet
- lässt man die Stichprobengröße gegen Unendlich laufen, nähert sich die Verteilung bei distance und rank hot deck Methoden der wahren Verteilung an

(Vgl. D'Orazio et al. (2006), S. 46)

Gemischte Methoden

Ziel: die positiven Eigenschaften der beiden Ansätze zu kombinieren → die Sparsamkeit von parametrischen Modellen & die Robustheit bzgl. Fehlspezifikationen bei Modellen von nicht-parametrischen Methoden
Die gemischten Methoden bestehen somit aus zwei Schritten.

- im ersten Schritt werden die Parameter des parametrischen Modells geschätzt
- im zweiten Schritt verwendet man eines der hot deck Verfahren bedingt auf die geschätzten Parameter aus dem ersten Schritt

(Vgl. D'Orazio et al. (2006), S. 47)

Gemischte Methoden - stetige Variablen

Bei stetigen Variablen bestehen die gemischten Modelle aus den folgenden drei Schritten:

1. Die Parameter der Regression von $\mathbf{Z}$ auf $\mathbf{X}$ werden anhand Datensatz B geschätzt.
2. Basierend auf dieser geschätzten Regressionsfunktion wird für jedes $a = 1, ..., n_A$ ein vorläufiger Wert $\tilde{\mathbf{z}}_a$ erzeugt.
3. Unter Berücksichtigung des vorläufigen Wertes $\tilde{\mathbf{z}}_a$ wird für jedes $a = 1, ..., n_A$ ein beobachteter Wert $\mathbf{z}_{b^*}$, mit b^* aus B, für den a-ten Eintrag aus A mittels einer geeigneten distance hot deck Methode imputiert.

(Vgl. D'Orazio et al. (2006), S. 47)

Gemischte Methoden - kategoriale Variablen

Liegen kategoriale Daten vor, bestehen die gemischten Modelle aus den folgenden zwei Schritten, wobei die drei vorliegenden kategorialen Variablen X, Y und Z hier univariat sein sollen, mit den Laufindizes i, j und k:

1. Die erwarteten Zellhäufigkeiten werden anhand des loglinearen Modells, das zur CIA Annahme passt, geschätzt.
2. Mit Hilfe einer hot deck Methode werden passende Z Werte aus B für jeden Eintrag aus A ausgesucht. Ein möglicher Wert, der für den a-ten Wert gemäß (i, j, k) passend wäre, wird nur genommen, falls die geschätzte Zellhäufigkeit $\tilde{n}_{ijk}$ nicht überschritten wird, andernfalls muss eine anderer Spender für den a-ten Eintrag gesucht werden.

(Vgl. D'Orazio et al. (2006), S. 50-51)

Pilotstudie Lebensqualität - Daten

	X	Y	Z
EU-SILC	Demografie... 3 Items zu materiellem Mangel	AROPE-Index At-risk-of-poverty Material-deprivation Low work intensity	
EQLS	Demografie... 3 Items zu materiellem Mangel		Lebenszufriedenheit Anerkennung Vertrauen in Presse&Regierung

EU-SILC ist hier der Empfängerdatensatz mit hauptsächlich Variablen zu ökonomischem Wohlbefinden, in den Variablen zu emotionalem Wohlbefinden aus dem Spenderdatensatz EQLS imputiert werden soll.

Pilotstudie Lebensqualität - Harmonisierung

Hier geht es darum abzuklären welche gemeinsamen Variablen der beiden Datensätze kohärent genug sind, um als Matchingvariablen in Frage zu kommen.

- Betrachtung von Frageformulierungen, Definitionen der gemessenen Konzepte, Messskalen sowie Richtlinien der Messungen
- Vergleich der marginalen Verteilungen anhand von 95% Konfidenzintervallen und geeignete statistische Tests zur Identifizierung von übereinstimmenden Variablen

$\rightarrow$ einige Variablen als Matchingvariablen ausgeschieden, andere konnten durch Transformationen harmonisiert werden

(Vgl. Leulescu und Agafitei (2013), S. 30-31)

Pilotstudie Lebensqualität - Erklärungskraft

Hier wird überprüft welche der Variablen einen Zusammenhang mit den Zielvariablen haben, die imputiert werden sollen, der so stark genug, dass ein Matchen anhand dieser Variablen sinnvoll ist.

- paarweise Korrelationen zwischen den gemeinsamen Variablen und den Zielvariablen & Tests mit der Nullhypothese "kein Zusammenhang bzw. Unabhängigkeit"
- welche der gemeinsamen Variablen hat einen starken Zusammenhang mit einem beträchtlichen Anteil der Zielvariablen

$\rightarrow$ bei den meisten Variablen trifft nicht beides zu, Kohärenz **und** hohe Erklärungskraft

(Vgl. Leulescu und Agafitei (2013), S. 33-35)

Pilotstudie Lebensqualität - CIA

Ziel: CIA bezüglich des AROPE-Index aus dem EU-SILC Datensatz und der Lebenszufriedenheit aus dem EQLS Datensatz überprüfen.

- Odds ratios mit Lebenszufriedenheit als abhängige Variable zeigen es gibt Variablen mit starkem Effekt
- ABER: selbst unter Kontrolle auf diese Variablen noch starke Korrelation zwischen Lebenszufriedenheit und den 3 gemeinsamen Items zu materiellem Mangel $\Rightarrow$ starker Zusammenhang zwischen Lebenszufriedenheit und einiger Items, die in den AROPE-Index mit eingehen spricht gegen die Annahme, dass die Lebenszufriedenheit bedingt auf die Matchingvariablen unabhängig von dem AROPE-Index ist
- Lösungsvorschlag: Näherungsvariablen (proxy variables)
- 3 Items aus dem AROPE-Index, die ebenfalls im EQLS Datensatz erhoben wurden, haben hohe Vorhersagekraft für gesamten Index $\Rightarrow$ eignen sich um Zusammenhang zwischen AROPE-Index und den zu imputierenden Variablen abzuschwächen

Pilotstudie Lebensqualität - Matchingvariablen

Zwei Gruppen Matchingvariablen.

1. Gruppe mit strenger Überprüfung von Kohärenz und Erklärungskraft:

Spanien	**Finnland**
Geschlecht	Geschlecht
Alter	Alter
NUTS 2 Region	NUTS 2 Region
Fleisch oder Fisch jeden 2. Tag	Beschäftigungsstatus
Berufsstatus	monatl. Nettoeinkommen (Haushalt)
Wohnbesitzverhältnis des Haushalts	genereller Gesundheitszustand

2. Gruppe mit mehr Wert auf Variablen, die wichtig für die Einhaltung der CIA erschienen, dafür nachlässiger bei Kohärenz:

   Alter, Geschlecht, Bildung, Familienstand
   Beschäftigungsstatus
   Gesundheitszustand
   3 Items zum materiellen Mangel

Pilotstudie Lebensqualität - Matchingmethoden I

1. Methode anhand der Gruppe 1: *distance hot deck* (unconstrained)

- mit Distanzmaß für binäre Variablen, das auf dem Koeffizienten von Dice (similarity coefficient) basiert und folgendermaßen definiert ist

$$D_{ij} = \sqrt{1 - S_{ij}} \text{ , wobei } S_{ij} = \frac{2a}{2a + b + c} \tag{19}$$

und a für die Anzahl der Indikatoren, für die $i = 1$ und $j = 1$ gilt, b für die Anzahl der Indikatoren, für die $i = 1$ und $j = 0$ gilt und c für die Anzahl der Indikatoren, für die $i = 0$ und $j = 1$ gilt, steht

(Vgl. Leulescu und Agafitei (2013), S. 37-38)

Pilotstudie Lebensqualität - Matchingmethoden II

2. Methode anhand der Gruppe 2: *predictive mean matching*

- ein gemischtes Modell mit folgenden zwei Schritten:
- Schätzung der Parameter der Regression (hier logistische Regression) von den Zielvariablen auf die Matchingvariablen anhand EQLS $\rightarrow$ vorläufige Schätzwerte der Zielvariablen für die Objekte aus EU-SILC
- Auswahl des Spenders aus EQLS, der gemäß dem Distanzmaß der distance hot deck Methode am nächsten ist

(Vgl. Leulescu und Agafitei (2013), S. 37-38)

Pilotstudie Lebensqualität - Ergebnisse I

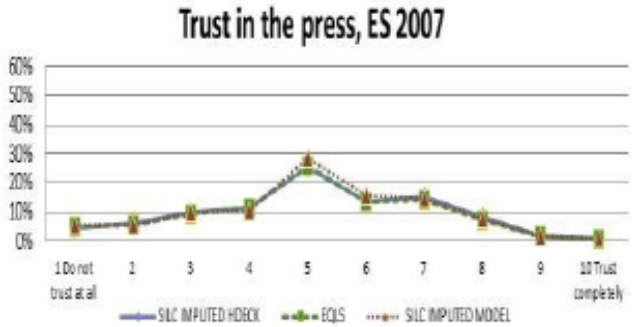


Abb. : Randverteilungen des beobachteten Datensatzes EQLS und der gematchten Datensätze mit den beiden Methoden distance hot deck und dem modellbasierten predictive mean matching (Vgl. Leulescu und Agafitei (2013),

Pilotstudie Lebensqualität - Ergebnisse II

Abb. : Vergleich der Verteilungen der Zielvariablen und der Teilpopulation mit großem materiellem Mangel und der Gesamtpopulation(Leulescu und Agafitei (2013), S. 40)

Zusammenfassung

- Erhaltung der Randverteilung hat keine gute Erfassung der gemeinsamen Verteilung zur Folge
- gute Matchingvariablen sehr wichtig bei zukünftigen Erhebungen vielleicht berücksichtigen
- CIA starke Annahme, die nicht gelten muss $\rightarrow$ siehe andere Methoden ohne CIA

Danke für Eure Aufmerksamkeit!

Literatur

D'Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. Wiley und Sons, Ltd., Sussex, England, Kap. 1 und 2.

Leulescu, A. und Agafitei, M. (2013). *Statistical matching: a model based approach for data integration*. Publications Office of the European Union, Luxemburg, Kap. 2.

Meinfelder, F. (2013). *Datenfusion: Theoretische Implikationen und praktische Umsetzung*. In: Riede, T., Bechtold, S. und Ott, N. (Hrsg.), *Weiterentwicklung der amtlichen Haushaltsstatistiken*, SCIVERO Verlag, Berlin, S. 83-98.

Noll, P. (2009). *Statistisches Matching mit Fuzzy Logic Theorie und Anwendungen in Sozial- und Wirtschaftswissenschaften*. Vieweg+Teubner |GWV Fachverlage GmbH, Wiesbaden, S. 11.