Give Missings a Chance: Ein kombinierter stochastischer und regelbasierter Ansatz für bessere Regressionsergebnisse mit fehlklassifizierten monotonen Kovariaten ohne Zusatzinformationen

4. IAB-FDZ Datennutzerkonferenz

Stephan Dlugosz
ZEW — Zentrum für Europäische Wirtschaftsforschung, Mannheim

8. April 2011

Problem

Datenqualität:

- Fehlende und falsch klassifizierte Werte in der Bildungsvariablen
- Inkonsistenzen, da Bildungsvariable im Zeitverlauf monoton

Chancen für statistische Ansätze:

- Monotonie
- verschiedene Quellen für Bildungsvariable in der IEBS

Lösungsansätze I

- klassische Behandlung fehlender Werte
  - entfernen unvollständiger Beobachtungen
  - ← nur unter MCAR-Annahme unverzerrt
  - imputieren
  - ← Annahmen zum Datengenerierungsprozess notwendig
  - ← keine Nutzung des Hintergrundwissens

- Umgang mit fehlklassifizierten Werten: simulationsbasierter Ansatz (SIMEX) wie in Wichert/Wilke 2010
  - Fehlklassifikationswahrscheinlichkeiten aus Validierungsdaten
  - Simulation der Parameterverläufe bei *zusätzlichem* Fehler gleicher Art
  - Extrapolation der fehlerfreien Daten

Lösungsansätze II

- regelbasierte Bildungskorrektur nach Fitzenberger et. al. 2006
  - Festlegung der Vertrauenswürdigkeit von Einträgen
  - Regeln für die Korrektur

  $\leftarrow$ Stochastischen Zusammenhänge werden ignoriert

- neuer Ansatz:
  Robustifizierung der Schätzergebnisse gegenüber Misklassifikationen/fehlende Werte

  $\rightarrow$ “nur” MAR-Annahme und (Regressions-)Modellspezifikation

  $\rightarrow$ Hintergrundwissen (Regeln) über Nebenbedingungen berücksichtigbar

Agenda

Voraussetzungen

- Missings in kategoriellen Variablen
- Missing at Random, d.h. “missing” ist unabh. von wahrem $X_k | X_{beob}$
- ausreichend große Fallzahl
- Kovariable kategoriell

- Anteil Missings klein genug -> Konvergenz!

Idee

- “Ort” des Fehlers bekannt $\rightarrow$ Missing
- eingeschränkte Anzahl Möglichkeiten
- Imputation *aller* (denkbaren) Möglichkeiten $\rightarrow$ Muster
- Muster (optimal) gewichten

ML-Ansatz

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(o)}) = \sum_{i=1}^{n} \sum_{\mathbf{x}_{mis,i}} w_{i,(o)} l(\boldsymbol{\theta}; \mathbf{x}_i, y_i)$$

- $l(\cdot) = l_{y_i|x_i}(\boldsymbol{\beta}, \phi) + l_{x_i}(\boldsymbol{\gamma})$
- $\boldsymbol{x}_i = (\boldsymbol{x}_{mis,i}, \boldsymbol{x}_{obs,i})$
- $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi, \boldsymbol{\gamma})$
- $w_{i,(o)} = p(\mathbf{x}_{mis,i}|\mathbf{x}_{obs,i}, y_i, \boldsymbol{\theta}^{(o)}) = \frac{p(y_i|\mathbf{x}_i,\boldsymbol{\theta}^{(o)})p(\boldsymbol{x}_i|\boldsymbol{\theta}^{(o)})}{\sum_{\boldsymbol{x}_{mis,i}} p(y_i|\boldsymbol{x}_i,\boldsymbol{\theta}^{(o)})p(\boldsymbol{x}_i|\boldsymbol{\theta}^{(o)})}$

Beispiele für $l_{y_i|x_i}(\cdot)$:

- Regression: $-(y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2$
- Logit: $y_i(\boldsymbol{x}_i\boldsymbol{\beta}) - \ln[1 + \exp(\boldsymbol{x}_i\boldsymbol{\beta})]$

EM by the method of weights

Iterativer EM-Algorithmus (Ibrahim, 1990):

init Datensatz „ergänzen“ und $w_{i,1} = \left(\sum_{\mathbf{x}_{mis,i}}\right)^{-1}$ setzen

M $\max_{\boldsymbol{\beta},\phi} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \sum_{i=1}^{n} \sum_{\mathbf{x}_{mis,i}} w_{i,(s)} l(\boldsymbol{\theta}; \mathbf{x}_i, y_i)$

E $w_{i,o} = \frac{p(y_i|\mathbf{x}_i,\boldsymbol{\theta}^{(s)})p(\boldsymbol{x}_i|\boldsymbol{\theta}^{(s)})}{\sum_{\boldsymbol{x}_{mis,i}} p(y_i|\boldsymbol{x}_i,\boldsymbol{\theta}^{(s)})p(\boldsymbol{x}_i|\boldsymbol{\theta}^{(s)})}$

Abbruchkriterium: $\left|w_{i,s} - w_{i,s-1}\right| < c$ für ein bel. $c \in \mathbb{R}$

Bildungskorrektur

Idee:

- Entfernen der unplausiblen Werte (regelbasiert)
- Beschränkung der möglichen Muster auf monoton ansteigende (Konsistenz)
- Beschränkungen durch Zeitrestriktionen (sehr optimistisch)
  - 16 für HS, 18 für VT und HSVT, 19 für TC und UD
- Ausführen des EM-Algorithmus

Daten und Regression

- Datensatz: IABS-R04
- Analyse: Mincer-Typ Lohnregression für 1999
- männliche sozialversicherungspflichtig-vollzeitbeschäftigte Westdeutsche
- Regression: (gew.) Tobit (Einkommen beschränkt)

also:

$$\log W = (E, P, O, I, 1)\beta + \epsilon$$

Unabhängige: Bildungsgrad (E), weitere persönliche Charakteristika (P), Beruf (O), Branche (I)

Anteil Missings in der Bildungsvariablen (gewichtet): ca. 21%

Korrekturregeln nach Fitzenberger et al.

IP1:

- valide: nur AG Meldungen
- nur Vorwärtskorrektur
- ca. 2% Missings (zunächst: 32%)

IP2a:

- valide: nur AG Meldungen nach dreifacher Nennung
- nur Vorwärtskorrektur
- ca. 4% Missings (zunächst: 29%)

IP2b:

- valide: konsistente Sequenzen, sonst: nur AG Meldungen nach dreif. Nennung
- nur Vorwärtskorrektur
- ca. 3% Missings (zunächst: 30%)

IP3:

- valide: nur Meldungen von AG mit hoher “Qualität” (max. eine Änderung)
- beide Korrekturrichtungen
- ca. 3% Missings (zunächst: 31%)

Ergebnisse I

	IP1		IP2a	
	RB	EM	RB	EM
ND	-0.106***	-0.135***	-0.106***	-0.132***
HS	-0.012***	-0.585***	0.019***	-0.505***
HSVT	0.161***	0.137***	0.179***	0.154***
TC	0.317***	0.290***	0.324***	0.308***
UD	0.489***	0.515***	0.504***	0.538***

	IP2b		IP3	
	RB	EM	RB	EM
ND	-0.104***	-0.133***	-0.106***	-0.132***
HS	-0.008***	-0.540***	0.005***	-0.566***
HSVT	0.156***	0.148***	0.166***	0.146***
TC	0.319***	0.302***	0.318***	0.297***
UD	0.499***	0.531***	0.494***	0.527***

Referenz: VT, *** p<0.01, ** p<0.05, * p<0.1

Ergebnisse II - ohne Praktikanten, Werkstudenten

	IP1		IP2a	
	RB	EM	RB	EM
ND	-0.106***	-0.135***	-0.106***	-0.135***
HS	0.077***	-0.552***	0.099***	-0.009***
HSVT	0.166***	0.140***	0.181***	0.169***
TC	0.316***	0.290***	0.322***	0.321***
UD	0.486***	0.564***	0.501***	0.558***

	IP2b		IP3	
	RB	EM	RB	EM
ND	-0.104***	-0.133***	-0.106***	-0.132***
HS	0.079***	-0.483***	0.092***	-0.517***
HSVT	0.160***	0.150***	0.172***	0.149***
TC	0.317***	0.303***	0.317***	0.299***
UD	0.496***	0.531***	0.491***	0.527***

Referenz: VT, *** p<0.01, ** p<0.05, * p<0.1

Blinder-Oaxaca Zerlegung

$$\begin{aligned}\ln \bar{W}_{HS} - \ln \bar{W}_{VT} &= (\beta_{0,HS} - \beta_{0,VT}) + (\bar{\mathbf{x}}_{HS} - \bar{\mathbf{x}}_{VT})\boldsymbol{\beta}_{VT} + (\boldsymbol{\beta}_{HS} - \boldsymbol{\beta}_{VT})\bar{\mathbf{x}}_{HS} \\ &= (\beta_{0,HS} - \beta_{0,VT}) + (\bar{\mathbf{x}}_{HS} - \bar{\mathbf{x}}_{VT})\boldsymbol{\beta}_{HS} + (\boldsymbol{\beta}_{HS} - \boldsymbol{\beta}_{VT})\bar{\mathbf{x}}_{VT}\end{aligned}$$

- Zerlegung nicht eindeutig
- VT als Referenz

$\Rightarrow$ 2. Alternative

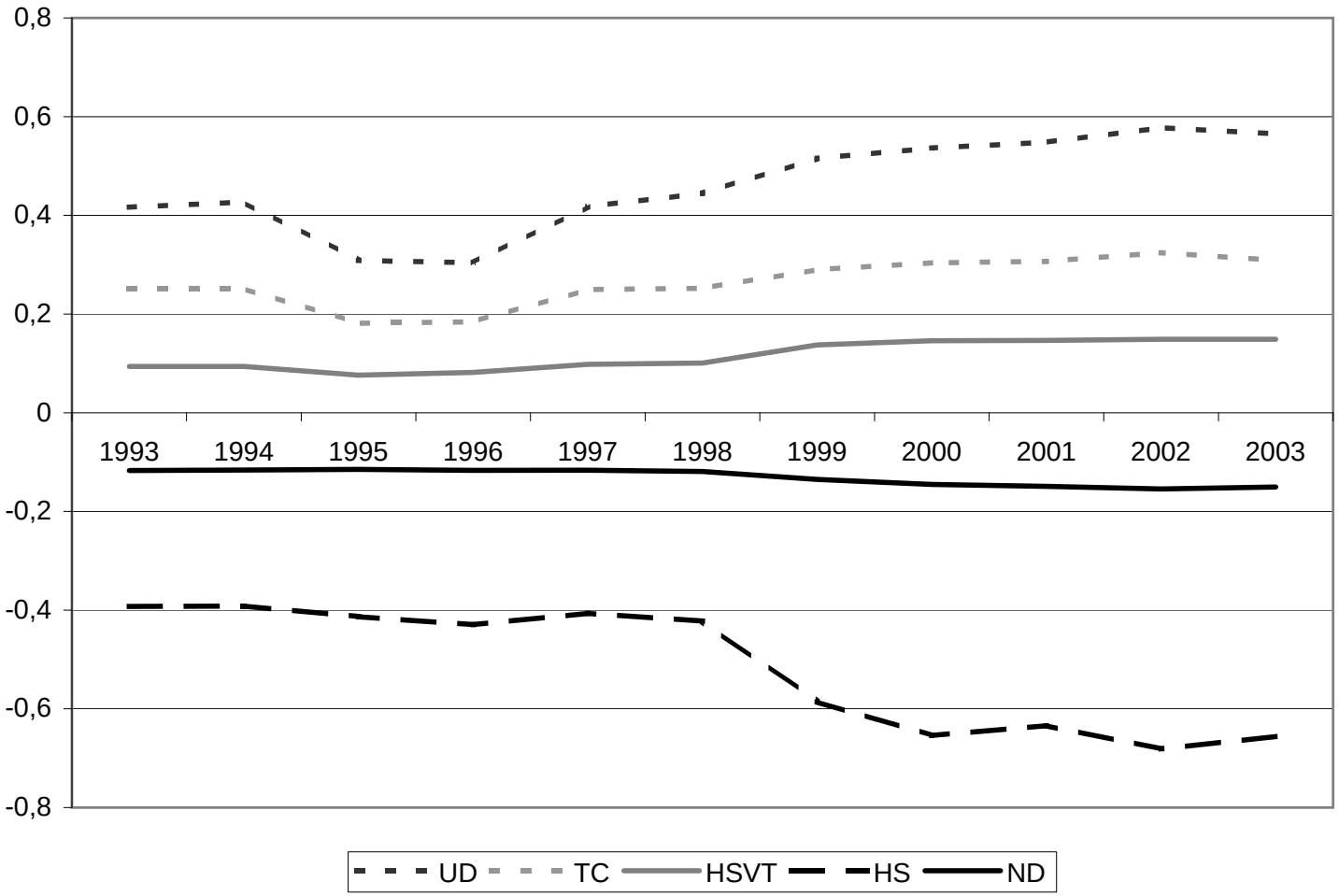
Blinder-Oaxaca Zerlegung

$$\ln \bar{W}_{HS} - \ln \bar{W}_{VT}$$
$$= (\beta_{0,HS} - \beta_{0,VT}) + (\bar{\mathbf{x}}_{HS} - \bar{\mathbf{x}}_{VT})\boldsymbol{\beta}_{HS} + (\boldsymbol{\beta}_{HS} - \boldsymbol{\beta}_{VT})\bar{\mathbf{x}}_{VT}$$

version	P/WS	wage differential	education	characteristics	coefficients
IP1	ja	-0.893	-1.118	-0.791	1.015
	nein	-0.836	-1.024	-0.776	0.964
IP2a	ja	-0.776	-1.439	-0.657	1.320
	nein	-0.205	-1.047	-0.290	1.133
IP2b	ja	-0.826	-1.281	-0.731	1.185
	nein	-0.739	-1.155	-0.688	1.104
IP3	ja	-0.859	-1.289	-0.741	1.172
	nein	-0.780	-1.160	-0.712	1.092

Ergebnisse III

Regressionskoeffizienten über die Jahre 1993 bis 2003



Kritik

Nachteile:

- MAR, aber verwandter Algorithmus für „missing not at random“ (Ibrahim/Lipsitz 1999)
- nur diskrete Kovariaten, ggf. gruppieren
- nur parametrische Modelle
- recht langsame Konvergenz (insb. bei hohem Anteil Missings)

Vorteile:

- intuitive Vorgehensweise
- keine (zusätzliche) parametrische Annahmen
- Nutzung gegebener zusätzlicher Informationen (Restriktionen)

Literatur I

[1] X. Chen, Y. Hu and A. Lewbel (2008a).
Nonparametric identification of regression models containing a misclassified dichotomous regressor without instruments.
Economics Letters ***100***, 381-384.

[2] X. Chen, Y. Hu and A. Lewbel (2008b).
A note on the closed-form identification of regression models with a mismeasured binary regressor.
Statistics and Probability Letters ***78***, 1473-1479.

[3] X. Chen, Y. Hu and A. Lewbel (2009).
Nonparametric identification and estimation of nonclassical errors-in-variables models without additional information.
Statistica Sinica ***19***, 949-968.

Literatur II

[4] B. Fitzenberger, A. Osikominu and R. Völter (2007).
Imputation Rules to Improve the Education Variable in the IAB Employment Subsample.
Journal of the Applied Social Science Studies (Schmollers Jahrbuch) ***126***, 405-436.

[5] J.G. Ibrahim (1990).
Incomplete Data in Generalized Linear Models.
Journal of the American Statistical Association ***85***, 765-769.

[6] J.G. Ibrahim (1999), S.R. Lipsitz and M.-H. Chen (1999).
Missing Covariates in Generalized Linear Models When the Missing Data Mechanism Is Non-Ignorable.
Journal of the Royal Statistical Society B ***61***, 173-190.

Literatur III

[7] J. Ibrahim, M. Chen, S. Lipsitz and A. Herring (2005).
Missing-Data Methods for Generalized Linear Models.
Journal of the American Statistical Association ***100***, 332-346.

[8] T.A. Louis (1982).
Finding the observed information matrix when using the EM algorithm.
Journal of the Royal Statistical Society B ***44***, 226-233.

[9] L. Wichert and R.A. Wilke
Which factors safeguard employment? : An analysis with misclassified German register data.
FDZ Methodenreport, 11/2010.

Asymptotik

- Schätzer sind konsistent und normalverteilt
- asymptotische Kovarianz:

$$\Sigma = [(\Sigma_T)^{-1} - \frac{1}{\phi^2}\tilde{X}' W \tilde{M}^2 diag(\tilde{y}_i - \tilde{\mu}_i)(I - W)\tilde{X}]^{-1}$$

ϕ Skalierungsfaktor für das GLM (abhängig von Link-Fkt.)
Σ_T Kovarianz aus dem letzten M-Schritt (letzte Regression)
$\tilde{X}$ erweiterte Designmatrix
W geschätze Gewichte der „pattern“
$\tilde{M}$ letztes Gewicht des Fisher Scoring-Algorithmus
$\tilde{\mu}_i$ lineare Prognose des Modell für den erweiterten Datensatz

Allgemein: $[\Sigma_T^{-1} - sWs' + sWP(sW)']^{-1}$ mit $s = \frac{dL(\theta)}{d\theta}$ (Louis 1982)

Ausblick

Schnelle Erweiterungen (aber nicht geplant):

- beschleunigte Konvergenz durch Initialisierung mit geschätzen Gewichten
- interne Diskretisierung metrischer Kovariabler

In nächster Zeit:

- Verwendung von Log-konkaven Fehlerverteilungen (ML mit KQ-Nähe)
- Anwendung mit Validierungsdaten

Auf längere Sicht:

- Implementierung des Algorithmus für fehlklassifizierte Daten (Chen/Hu/Lewbel 2009)
- Entwicklung eines kombinierten Verfahrens