

Gewichtung und Hochrechnung mit dem SOEP

Martin Spieß

SOEP, DIW-Berlin
Mohrenstr. 58
10117 Berlin

10.3.2008

Warum Gewichtung?

Gewichte der Startwellen des SOEP

Gewichte der Folgewellen des SOEP

Bleibefaktoren

Querschnittsgewichte

Umsetzung und Anmerkungen

Varianzschätzung und Konfidenzintervalle

Informationen zur Hochrechnung in der SOEP-Datenbank

Literatur

Warum Gewichtung?

Zentral: Population, über die eine Aussage gemacht werden soll

Prinzipiell:

- ▶ Konkrete, endliche Grundgesamtheit. Mit dem SOEP:
 - Privathaushalte in der Bundesrepublik in e. best. Zeitraum
 - In Privathaushalten lebende Wohnbevölkerung in der Bundesrepublik am Ort der Hauptwohnung in e. best. Zeitraum
- ▶ Abstrakte Grundgesamtheit. Mit dem SOEP:
 - Privathaushalte
 - In Privathaushalten (Hauptwohnsitz) lebende Personen

Schluß von der Stichprobe auf die Grundgesamtheit (Inferenz):

- ▶ Konkrete, endliche Grundgesamtheit: „Design“-basierte Inferenz
 - Ausgangspunkt: Zufälligkeit im Ziehungsdesign begründet
 - Ziel: „Hochrechnung“ auf die endliche Grundgesamtheit (z.B. Anzahl an Haushalten i. e. best. EK-Klasse)
- ▶ Abstrakte Grundgesamtheit: „Modell“-basierte Inferenz
 - Ausgangspunkt: Zufälligkeit im Modell begründet
 - Ziel: (meist) Schluß auf Parameter eines Modells (z.B. Effekt von Ausbildung auf Einkommen).

Warum Gewichtung?

Design-basierte Inferenz im Querschnitt — formale Begründung für Gewichtung:

- ▶ Schätzung von Kennwerten in konkreter, endlicher Grundgesamtheit (N : Anzahl an Populations-Einheiten, $i = 1, \dots, N$)
- ▶ Zufallsvariable: Beobachtungsindikator c_i mit
$$c_i = \begin{cases} 1 & \text{falls Element } i \text{ in Stichprobe gelangt,} \\ 0 & \text{sonst} \end{cases}$$
- ▶ Zu schätzender Kennwert z.B. Merkmalssumme:

$$t_y = \sum_{i=1}^N y_i$$

Erwartungstreuer Schätzer

$$\hat{t}_y = \sum_{i=1}^N w_i y_i \quad \text{mit} \quad E(w_i) = 1.$$

- ▶ z.B. „inverse selection probability weighting scheme“

$$w_i = \frac{c_i}{\pi_i}$$

mit $\pi_i = E(c_i) = \Pr(c_i = 1)$ („Horvitz-Thompson“-Schätzer).

Für w_{it} gibt es auch andere Möglichkeiten (andere „weighting schemes“, z.B. Kalton und Brick, 1995).

Warum Gewichtung?

Modell-basierte Inferenz — formale Begründung für Gewichtung prinzipiell wie im design-basierten Kontext.

- ▶ Aber: Neben Beobachtungsindikator c_i werden weitere Variablen als Zufallsvariablen behandelt.

Ähnliche Begründung auch im Hinblick auf asymptotische Eigenschaften (z.B. Wooldridge, 2002).

SOEP: „inverse selection probability weighting scheme“

Warum Gewichtung?

Für valide Inferenz wird benötigt:

- ▶ W'keit die Einheit i (Haushalt, Person) zu beobachten. Dazu:
 1. Ableitung/Modellierung des Prozesses, der zur Beobachtung der Einheit führt.
 2. Bestimmung/Schätzung der Beobachtungsw'keit.
- ▶ Für spezifische (modell-basierte) Analysen: Um unnötig große Standardfehler zu vermeiden Ableitung/Modellierung dieses Prozesses nur insoweit als informativ für spezif. Analyse (z.B. Wooldridge, 2002).

SOEP Startwelle

Wie kommt die beobachtete Stichprobe in der 1. Welle zustande?

1. Ziehung der Stichprobenelemente („Brutto“-Stichprobe)
2. Kontaktieren und Beobachten der Stichprobenelemente („Netto“-Stichprobe)

Für das SOEP:

- ▶ Ziehungswahrscheinlichkeiten für die Teilstichproben des SOEP (A bis H) sind entsprechend des Designs der jeweiligen Stichprobe unterschiedlich.
- ▶ Non-Response ist von Bedeutung. D.h.:
 - ▶ Teilnahmeverweigerung in der jeweiligen Startwelle (1984 A+B, 1990 C, 1994, 1995 D, 1998 E, 2000 F, 2002 G, 2006 H).

Stichprobenziehung: SOEP besteht aus acht Teilstichproben A–H.
Grundsätzlich:

- ▶ Zweistufiges Ziehungsverfahren
 1. „Primäreinheiten“: Wahlkreise (A, D(teilweise), E, F), Kreise/kreisfreie Städte (B), Gemeinden/Landkreise (C), stat. Einheiten auf Basis kommunaler Bezirke (H)
 2. „Sekundäreinheiten“: Haushalte (A, C, D, E, F, G, H), Personen aus Ausländerregister (B)
- ▶ Systematisches, größenproportionales Ziehen. Dazu Anordnen der Einheiten,
 - ▶ Primäreinheiten nach regionalen Kriterien und Gewichtung nach Größe,
 - ▶ Sekundäreinheiten „Random Route-Verfahren“ (zufällige Startadresse und fixes Intervall).

Besonderheiten der Teilstichproben:

- ▶ A: Start 1984, BRD und West-Berlin, Haushaltsvorstand (Hv) besitzt nicht die Nationalität der B-Stichproben Hv (Netto: 4528 Hh)
- ▶ B: Start 1984, BRD und West-Berlin, Nationalität des Hv türkisch, italienisch, griechisch, jugoslawisch, spanisch (Netto: 1393 Hh)
- ▶ C: Start 1990, neue Bundesländer (Netto: 2179 Hh)
- ▶ D: Start 1994/1995, priv. Haushalte (Hh), die seit 1984 nach West-Deutschland kamen; besteht aus mehreren Stichproben mit unterschiedl. Ziehungsdesigns (Netto: 522 Hh)
- ▶ E: Start 1998, Auffrischungsstichprobe, Population wie A–D, unabh. Stichprobe (Netto: 1056 Hh)

Besonderheiten der Teilstichproben (Forts.):

- ▶ F: Start 2000, Auffrischungstichprobe, Population wie A–D bzw. E, unabh. Stichprobe (Netto: 6052 Hh)
- ▶ G: Start 2002, Hh im oberen Einkommensbereich (3835 Euro monatl. Netto-Hh-Einkommen; Netto: 1124 Hh)
 - ▶ Für SOEP Stichproben untypisches Auswahlverfahren:
 1. InfraScope Mehrthemenbefragung, telefonisch, im Jahr 2001, Einkommensangaben unvollständig.
 2. Daraus: 6330 Hh, die die Einkommensgrenze erreichen. Einverständnis für weitere Befragung: 5663 Haushalte
 3. neuerliche (geschichtete) Auswahl für mündliche Interviews: 3672 Haushalte. Einverständnis mit mündlichem Interview: 2495 Haushalte. (Auswahlkriterium war Befragten und Interviewern unbekannt).
 4. Schließlich beobachtet: 1631 Haushalte. Nicht alle erreichten die Einkommenschwelle: bleiben 1224 Haushalte.
- ▶ H: Start 2006, Auffrischungstichprobe, Population wie A–D bzw. E und F, unabh. Stichprobe (Netto: 1505 Hh)

Erster Schritt bei der Gewichtung der jeweiligen Startwelle:

- ▶ Ableitung der Ziehungsw'keiten aus dem Auswahlverfahren. Liefert die W'keit in die Brutto-Stichprobe gezogen zu werden, $\Pr(s_i = 1)$, mit

$$s_i = \begin{cases} 1 & \text{falls Element } i \text{ in Stichprobe gezogen wird,} \\ 0 & \text{sonst} \end{cases}$$

dem Selektionsindikator.

- ▶ Die Kehrwerte der Ziehungsw'keiten sind in Datei DESIGN unter dem Variablennamen DESIGN zu finden.

Problem: Brutto-Stichprobe \neq Netto-Stichprobe.

W'keit eine Einheit i in der Startwelle zu beobachten, ist:

$$\Pr(c_i = 1) = \Pr(s_i = 1) \Pr(r_i = 1 | s_i = 1)$$

mit

$$r_i = \begin{cases} 1 & \text{falls Stichprobenelement } i \text{ beobachtet wird,} \\ 0 & \text{sonst} \end{cases}$$

dem Responseindikator.

Aber: $\Pr(r_i = 1 | s_i = 1)$ unbekannt.

Schätzung von $\Pr(r_i = 1|s_i = 1)$:

1. über relative Beobachtungshäufigkeiten in Regionalzellen (z.B. A, E, F)
2. anschließendes „Trimmen“ bei $10 \times$ stichprobenspez. Median
3. Randanpassung (Mikrozensus): Haushaltsgröße, Nationalität, Geschlecht, Haushaltsvorstand, Alter etc.

Gewichte der Startwelle: $1/\{\Pr(s_i = 1)\hat{\Pr}(r_i = 1|s_i = 1)\}$ abgelegt in Dateien

- ▶ HHRF auf Haushaltsebene; z.B. für Teilstichproben A und B: AHHRF
- ▶ PHRF auf Personenebene; z.B. für Teilstichproben A und B: APHRF

Folgewellen: Bleibefaktoren

Bleibefaktoren: Kehrwerte der (geschätzten) W 'keiten, Haushalte bzw. Personen in Welle t zu beobachten, gegeben sie wurden in Welle $t - 1$ beobachtet ($\Pr(c_{it}|c_{i,t-1})$).

Im SOEP: Zwei Schritte

1. Sind Hh, in Welle t wieder auffindbar?
→ Kontaktw'keiten
2. Liegen für in Welle t kontaktierte Hh Informationen vor?
→ Responsew'keiten

Beide (bedingten) W 'keiten unbekannt — daher:

1. Schätzen der Kontaktw'keiten, Logitmodell (Kovariablen z.B. Metropole – ja/nein, Einpersonen-Hh – ja/nein etc.)
2. Schätzen der Responsew'keiten, Logitmodell (Kovariablen z.B. Charakteristika Hhv, Interviewcharakteristika etc.)

jeweils spezifisch f. Teilstichproben (für Details: Spiess und Kroh, 2008).

Kehrwert des Produkts dieser geschätzten W' keiten:
Bleibefaktoren. Diese sind abgelegt in den Dateien

- ▶ HHRF, unter den Variablennamen \$HBLEIB, mit \$: B, ..., V, für Haushalte
- ▶ PHRF, unter den Variablennamen \$PBLEIB, mit \$: B, ..., V, für Personen

\$HBLEIB bzw. \$PBLEIB: Kehrwerte der geschätzten W' keiten, Hh bzw. Person in Welle \$ zu beobachten, gegeben er/sie wurde in der Vorwelle beobachtet.

Startwellengewichte und Bleibefaktoren können verwendet werden um Längsschnittgewichte zu generieren.

Annahme: Beobachtungsw'keit hängt nicht von unberücksichtigten/unbeobachteten Variablen ab, die relevant sind für die interessierende Analyse.

Beispiel-Gewichte für balancierte Längsschnittpopulation von 1984 bis 1986 (Haushalte):

- ▶ Balancierte Panelstichprobe: Kehrwert der (gesch.) W'keit Hh in Welle C (und B und A) zu beobachten \approx AHHRF \times BHBLEIB \times CHBLEIB

- ▶ Unbalancierte Panelstichprobe (Annahme: Ausfall ist absorbierend, keine Neuzugänge):

$$\text{AHHRF} \approx \{1/\hat{\text{Pr}}(c_{i,1} = 1)\}$$

$$\text{AHHRF} \times \text{BHBLEIB} \approx \{1/\hat{\text{Pr}}(c_{i,1} = 1, c_{i,2} = 1)\}$$

$$\text{AHHRF} \times \text{BHBLEIB} \times \text{CHBLEIB} \approx \{1/\hat{\text{Pr}}(c_{i,1} = 1, c_{i,2} = 1, c_{i,3} = 1)\}$$

Im ersten Fall: Eigentlich vorhandene Fälle werden ignoriert.

Folgewellen: Querschnittsgewichte

Ausgangspunkt:

- ▶ SOEP ist ein Längsschnittdatensatz
- ▶ Jede Welle soll auch als eigener Querschnitt auswertbar sein
- ▶ Längsschnittanalysen sollen mit beliebiger Startwelle möglich sein

Gewichte im Längsschnitt: Sich verändernde Population ist abzubilden

- ▶ Population ändert sich durch „Geburten“, „Todesfälle“, Zuwanderung aus dem bzw. Auswanderung ins Ausland
- ▶ Zusätzlich bei künstliche Einheiten → Fusionen und Aufspaltungen

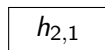
Insbesondere bei letzterem Punkt: Wie lassen sich diese zeitl. Veränderungen in den Gewichten so abbilden, dass in jeder Welle > 1 eine valide Inferenz möglich ist?

Folgewellen: Abbildung der Dynamik

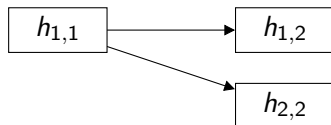
Beobachtungsw'keiten in Folgewelle (keine Ausfälle, Populationszu-/abgänge), künstliche Einheiten:



$$\pi_{12} = \pi_{11}$$



Fusion. Annahme: Ziehungen in Welle 1 unabhängig,
 $\pi_{1,2} = \pi_{1,1} + \pi_{2,1}$



Abspaltung. $\pi_{1,2} = \pi_{1,1}$,
 $\pi_{2,2} = \pi_{1,1}$

π_{it} : Beobachtungsw'keit für Hh i ($i = 1, 2$) in Welle t ($t = 1, 2$)

Folgewellen: Ergänzungen

- ▶ Ausscheiden aus der/Zuwanderungen (ganzer Hh) in die Population bildet sich u.U. nicht vollst. i.d. Stichprobe ab → „Randanpassung“, Ziehung spezifischer Stichproben oder Auffrischungstichproben (z.B. D, E, F, G, H)
- ▶ Fusion:
 - ▶ Wenn mit Haushalt 1 fusionierender Hh 2 in Welle 1 nicht Element der Population:
$$\pi_{1,2} = \pi_{1,1}$$
 - ▶ Wenn Hh 2 in Welle 1 Element der Population aber nicht Element der Stichprobe: $\pi_{1,2} = \pi_{1,1} + \tilde{\pi}_{2,1}$,
($\tilde{\pi}_{2,1}$ geschätzt unter Annahmen)
- ▶ Weitere Spezialfälle: temp. Ausfälle, Personen mit Zweitwohnsitz
- ▶ Je mehr Wellen, desto größer die W'keit des Auftretens komplexerer Verläufe. Teilweise mit obigen Bausteinen bearbeitbar.

Erzeugen der Querschnitts-Gewichte ab 2. Welle:

1. Vorl. Gewicht =
Evtl. modifiziert(Querschnitts-Gewicht Vorwelle \times
Bleibefaktor)
2. In jeder Welle „Trimmen“ bei $10 \times$ stichprobenspez. Median
3. „Randanpassung“ um für Problem wie „Noncoverage“
(Untererfassung) etc. zu kompensieren (Haushaltsgröße,
Nationalität, Geschlecht Haushaltsvorstand, Alter)

Ausführliche Beschreibung des SOEP-Gewichtungskonzepts in
Rendtel (1995)

Die entsprechenden Querschnitts-Gewichte sind in den Dateien
HHRF (Haushalte) bzw. PHRF (Personen) zu finden (siehe auch
den nächsten Abschnitt).

Umsetzung und Anmerkungen

Für Querschnittsanalysen (Gewichte aus HHRF bzw. PHRF):

- ▶ Prinzipiell empfohlen: Standardgewichtungsfaktoren ($\$HHRF$ bzw. $\$PHRF$) und Teilstichproben A–F
 - ▶ umfassen jeweils alle Teilstichproben (außer G)
 - ▶ Überproportionalität v. Teilstichproben berücksichtigt
- ▶ Wenn Interesse nur an Zuwandererpopulation nach West-D bis 1994: z.B. D spezifische Gewichte $\$HHRFD$ bzw. $\$PHRFD$
- ▶ Es gibt Gewichte spezifisch f. Analysen der Teilstichproben:
 - ▶ A–F: $\$HHRF1$ bzw. $\$PHRF1$, erste Welle auf null gesetzt
 - ▶ A–G: $\$HHRFALL$ bzw. $\$PHRFALL$
 - ▶ G: $\$HHRFG$ bzw. $\$PHRFG$

Umsetzung und Anmerkungen

Für Längsschnittanalysen:

- ▶ Zu verwendende Gewichte hängen ab von: Definition der Population (balanciertes/unbalanciertes Panel, model-/designbasierte Inferenz)
- ▶ Gewichte können aus den mit dem SOEP ausgelieferten Bausteinen aus den Dateien HHRF, PHRF und DESIGN erzeugt werden.
- ▶ Gegebenenfalls können die Kehrwerte der Ziehungsw'keiten (DESIGN in Datei DESIGN) als Ausgangspunkt für Ableitung neuer Gewichte verwendet werden.
- ▶ Modellbasierte Analysen: Ziehungsdesign oft nicht informativ → Ziehungsw'keiten können ignoriert werden (Gewicht: $\$HHRF/DESIGN$ bzw. $\$PHRF/DESIGN$)
- ▶ Gewichte sind Funktionen von Schätzern: Nicht Berücksichtigen der Unsicherheit führt i.Allg. zu konservativer Inferenz.

Sonstige Anmerkungen:

- ▶ Bei Analysen und Vergleichen mit externen Daten beachten:
 - ▶ Abweichungen in Randverteilungen sind oft auf unterschiedl. Abgrenzungen zurückzuführen
 - ▶ SOEP: Privathaushalte bzw. Personen in Privathaushalten; Anstalts-Hh im SOEP nicht als Zufallsstichprobe aus der (endlichen Population) der Anstaltshaushalte aufzufassen
 - ▶ Teilstichprobenabgrenzung nicht über Variable \$SAMPREG („aktuelle Stichprobenregion“, z.B. in HPFAD); besser: KSAMPLE (Datei DESIGN) oder SAMPLE1 (versch. Dateien)
- ▶ Wichtig: Bei der Interpretation von Schätzwerten sollten Konfidenzintervalle bzw. Varianzen der Schätzer berücksichtigt werden.

Varianzschätzung und Konfidenzintervalle

Variablen zur Bestimmung von Konfidenzintervallen und zur Varianzschätzung in der Datei DESIGN:

- ▶ STRAT: Schichtinformationen, aus denen die Primary Sample Units (PSU; Primäreinheiten) gezogen wurden
- ▶ PSU: PSU-Nummern
- ▶ INTNR: Interviewernummer
- ▶ RGROUP: „Random Group“, s.u.

Design-basierte Inferenz:

- ▶ Software: z.B. STATA, SUDAAN, IVEware
- ▶ Allgemeine Methoden: Jackknife, Bootstrap etc.
- ▶ Eine einfache Methode: „Random Groups“-Methode (z.B. Wolter, 1995), (Variable: RGROUP)

Varianzschätzung nach dem Random Group Konzept:

- ▶ „Simulation“ von (im Falle des SOEP) $R = 8$ unabhängigen, identisch gezogenen (Teil-)Stichproben.
- ▶ Berechnung des interessierenden Schätzwertes für jede Teilstichprobe
- ▶ Streuung dieser R Schätzwerte liefert Basis für Varianzschätzung.

Varianzschätzung und Konfidenzintervalle

Berechnung eines Konfidenzintervalls:

1. Berechnung der $R = 8$ Schätzwerte für die Random Groups
2. Anordnung der Einzelergebnisse nach Größe (aufsteigend)
3. Bestimmung Konfidenzintervall:
 - (a) $\Pr(\hat{y}_1 < y < \hat{y}_8) \approx 1 - 0.008$
(Konfidenzintervall zum Niveau 0.992)
 - (b) $\Pr(\hat{y}_2 < y < \hat{y}_7) \approx 1 - 0.07$
(Konfidenzintervall zum Niveau 0.93)

Schätzung der Varianz (eine mögliche Variante):

$$\hat{\sigma}_R^2 = \frac{1}{R(R-1)} \sum_{r=1}^R (\hat{y}_r - \bar{\hat{y}})^2$$

$$\text{mit } \bar{\hat{y}} = \frac{1}{R} \sum_{r=1}^R \hat{y}_r.$$

Anhang: Datei DESIGN

VAR LABELS	HHNR	'Ursprungshaushaltsnummer'
	KSAMPLE	'Stichprobenart'
	RGROUP	'Random Groups'
	DESIGN	'Inverse Ziehungswahrscheinlichkeit'
	STRAT	'Schichtung, Stratifizierungseinheiten'
	PSU	'Klumpung, Primaere Ziehungseinheiten'
	INTNR	'Interviewer-Nr'/'

Anhang: Datei HHRF

Variablen

HHNR	'Ursprungshaushaltsnummer'	SHHRFALL	'Hochrechnungsfaktor Sample A-E 2002'
HHNRAKT	'Aktuelle Haushaltsnummer'	:	:
HRGROUP	'Random Groups'	WHHRFALL	'Hochrechnungsfaktor Sample A-E 2006'
AHHRF	'Hochrechnungsfaktor 1984'	AHHRF1	'Hochrechnungsfaktor ohne Samples A,B 1984'
:	:	:	:
WHHRF	'Hochrechnungsfaktor 2006'	WHHRF1	'Hochrechnungsfaktor ohne Sample G,H 2006'
BHBLEIB	'Bleibewahrscheinlichkeit 1985'	SHHRFG	'Hochrechnungsfaktor Sample G 2002'
:	:	:	:
WHBLEIB	'Bleibewahrscheinlichkeit 2006'	WHHRFG	'Hochrechnungsfaktor Sample G 2006'
LHHRFD	'Hochrechnungsfaktor Sample D 1995'		
:	:		
WHHRFD	'Hochrechnungsfaktor Sample D 2006'		

Anhang: Datei PHRF

Variablen

HHNR	'Ursprungshaushaltsnummer'	SPHRFALL	'Hochrechnungsfaktor alle Samples 2002'
PERSNR	'Unveraenderl. Personennummer'	:	:
HRGROUP	'Random Groups'	WPHRFALL	'Hochrechnungsfaktor alle Samples 2006'
APHRF	'Hochrechnungsfaktor 1984'	APHRF1	'Hochrechnungsfaktor ohne Samples A,B 1984'
:	:	:	:
WPHRF	'Hochrechnungsfaktor 2006'	WPHRF1	'Hochrechnungsfaktor ohne Sample G,H 2006'
BPBLEIB	'Bleibewahrscheinlichkeit 1985'	SPHRFG	'Hochrechnungsfaktor Sample G 2002'
:	:	:	:
WPBLEIB	'Bleibewahrscheinlichkeit 2006'	WPHRFG	'Hochrechnungsfaktor Sample G 2006'
LPHRFD	'Hochrechnungsfaktor Sample D 1995'		
:	:		
WPHRFD	'Hochrechnungsfaktor Sample D 2006'		

Literatur: Allgemein

Kalton, G. & Brick, J.M. (1995). Weighting Schemes for Household Panel Surveys. *Survey Methodology*, 21, 33–44.

Wolter, K.M. (1995). *Introduction to Variance Estimation*. Springer: New York.

Wooldridge, J.M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Massachusetts: The MIT Press.

Literatur: Hochrechnung SOEP

Auf <http://www.diw.de/deutsch/sop/service/doku/index.html>

Haisken-DeNew, J.P. & Frick, J.R. (2005). DTC — Desktop Companion to the German Socio-Economic Panel (SOEP).

Pischner, R. (2001). Überarbeitete Querschnittsgewichtung der Wellen G-N (1990-1997) des Sozio-ökonomischen Panels unter Einbeziehung der Ergänzungsstichprobe E (Welle O).
Webpage-Titel: Querschnittshochrechnung 90-98.

Pischner, R. (2002a). Änderungen am Konzept der Querschnittsgewichtung des Sozio-oekonomischen Panels (SOEP) 1984-2001.

Pischner, R. (2002b). Die Hochrechnung der ersten Welle der Stichprobe F des SOEP.

Pischner, R. (2003). Integrated Cross-sectional Weighting (Sub-samples A-G) for 2003.

Literatur: Hochrechnung SOEP

- Spiess, M., Kroh, M., Pischner, R. & Wagner, G.G. (forthcoming). On the Treatment of Non-Original Sample Members in the German Household Panel Study (SOEP) – Tracing, Weighting and Frequencies. (Data Documentation). Berlin: DIW.
- Spiess, M. & Kroh, M. (2008). *Documentation of Sample Sizes and Panel Attrition in the German Socio Economic Panel (SOEP) 1984 - 2006*. (Data Documentation No. 27). Berlin: DIW.
- Spiess, M. & Rendtel, U. (2000). Combining an ongoing panel with a new cross-sectional sample, DIW Discussion Paper No. 198. Webpage-Titel: Martin Spieß – Sample E Integration into SOEP.
- Spiess, M. (2001b). Combining the SOEP ongoing panel (subsamples A-E) with the new subsample F. Webpage-Titel: Sample F Integration into SOEP.
- Spiess, M. (2001a). Description of the variables: STRAT1, STRAT2 and SAMPOINT. Webpage-Titel: Documentation for the SOEP File Varianz.
- Spiess, M. (2000). Derivation of design weights: The case of the German Socio-Economic Panel (GSOEP). DIW Discussion Paper No. 197.

Sonstige SOEP-spezifische Literatur

Rendtel, U. (1995). Panelmortalität und Panelrepräsentativität. Frankfurt: Campus.

Rendtel, U., Pannenberg, M. & Daschke, S. (1997). Die Gewichtung der Zuwanderer-Stichprobe des Sozio-ökonomischen Panels (SOEP). *Vierteljahrsheft für Wirtschaftsforschung* 2/97, 271–286.

Rendtel, U., Wagner, G. & Frick, J. (1995). Eine Strategie zur Kontrolle von Längsschnittgewichten in Panelerhebungen - Das Beispiel des Sozio-ökonomischen Panels (SOEP). *Allgemeines Statistisches Archiv*, 79, 252–277.