

Pischner, Rainer; Rendtel, Ulrich

Working Paper — Digitized Version

Quer- und Längsschnittgewichtung des Sozio-oekonomischen Panels

DIW Discussion Papers, No. 69

Provided in Cooperation with:

German Institute for Economic Research (DIW Berlin)

Suggested Citation: Pischner, Rainer; Rendtel, Ulrich (1993) : Quer- und Längsschnittgewichtung des Sozio-oekonomischen Panels, DIW Discussion Papers, No. 69, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:

<https://hdl.handle.net/10419/95722>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



Diskussionspapiere
Discussion Papers

Diskussionspapier Nr. 69

**Quer- und Längsschnittgewichtung
des Sozio-oekonomischen Panels**

von

Rainer Pischner und Ulrich Rendtel

Die in diesem Papier vertretenen Auffassungen liegen ausschließlich in der Verantwortung des Verfassers und nicht in der des Instituts.

Opinions expressed in this paper are those of the author and do not necessarily reflect views of the Institute.

Deutsches Institut für Wirtschaftsforschung

Diskussionspapier Nr. 69

Quer- und Längsschnittgewichtung des Sozio-oekonomischen Panels

von

Rainer Pischner und Ulrich Rendtel

Berlin, im Juli 1993

Deutsches Institut für Wirtschaftsforschung, Berlin
Königin-Luise-Str. 5, 1000 Berlin 33
Telefon: 49-30 - 82 991-0
Telefax: 49-30 - 82 991-200

Quer- und Längsschnittgewichtung des Sozio-oekonomischen Panels (SOEP)

1

0 Einleitung

In diesem Aufsatz werden zunächst einige allgemeine Bemerkungen über das Sozio-oekonomische Panel (SOEP) gemacht, wobei ich mich mehr auf technische als auf inhaltliche Gesichtspunkte konzentrieren werde. Dann wird das methodische Konzept, das der Gewichtung des Panels zugrundeliegt, dargestellt. Generell muß zwischen Querschnitt- und Längsschnittgewichtung unterschieden werden:

Jede Panelwelle kann zunächst für sich isoliert betrachtet werden; in diesem Fall sind Querschnittsgewichtungen erforderlich. Dabei wird sich zeigen, daß die Startwelle eines Panels eine Sonderrolle spielt, so daß zwei Arten von Querschnittsgewichtungen behandelt werden müssen.

Verfolgt man Personen und Haushalte von einer Welle zur anderen, so entsteht das Problem der Längsschnittgewichtung. Insgesamt werden also im SOEP drei Gewichtungstypen gebraucht, deren theoretische Grundlagen ich zunächst darstellen werde. Anschließend werde ich zeigen, wie dieses Konzept in die Praxis umgesetzt wurde.

1 Kurze Vorstellung des Sozio-oekonomischen Panels

Seit 1984 werden mit dem Sozio-oekonomischen Panel durch wiederholte Befragungen regelmäßig Mikro-Längsschnittdaten über Personen, Familien und Haushalte in Westdeutschland bereitge-

1) Einige Teile des Referats sind sehr eng an den Aufsatz von U. Rendtel "Die Schätzung von Populationswerten in Panelerhebungen" (1991) angelehnt.

stellt (im folgenden SOEP-West genannt).² Im Juni 1990 wurde die Panelerhebung auch auf das Gebiet der damaligen DDR ausgedehnt (SOEP-Ost).³

Haushaltszusammensetzung, Beteiligung am Erwerbsleben und berufliche Mobilität, Einkommensverläufe und Wohnsituation sind die wichtigsten Themenbereiche der Erhebung. Aber auch Fragen zur Bildung, Gesundheit, Zeitverwendung, Zufriedenheit und zu Werteinstellungen werden gestellt. Es ist nicht Ziel dieses Aufsatzes, auf dieses vielfältige Spektrum der angesprochenen Themen des SOEP näher einzugehen, vielmehr sollen einige grundlegende methodische Angaben zur Stichprobenanlage und zum Umfang des Projektes gemacht werden.

Der Start des SOEP-West mit rund 6000 Privathaushalten - davon 1400 mit ausländischem Haushaltsvorstand - erfolgte im Jahr 1984. Ein hoher Ausländeranteil in der Stichprobe von beinahe 25% ergab sich aus einem disproportionalen Stichprobenansatz. Diese Überrepräsentation soll spezielle Analysen zur Situation der Ausländer in Deutschland ermöglichen.

Erhebungsinstrumente sind ein Adreßprotokoll mit den Informationen über sämtliche Haushaltsmitglieder, ein Haushaltsfragebogen - in der Regel vom Haushaltsvorstand bzw. von der Bezugsperson beantwortet, und für jede Befragungsperson ein Personenfragebogen. Die Interviews wurden zumeist mündlich durchgeführt.

Befragt wurden alle Personen eines Haushalts, die zu Beginn des Befragungsjahres mindestens 16 Jahre waren. Haushalte, in dem

2) Einen aktuellen Überblick über das Sozio-oekonomische Panel findet man bei Projektgruppe SOEP (1991)..

3) Das von der DFG geförderte Projekt wird seit 1991 vom Deutschen Institut für Wirtschaftsforschung, Berlin, in alleiniger Verantwortung durchgeführt. Vorbereitet wurde es am Sonderforschungsbereich 3 (Sfb 3) im Teilprojekt B-1 "Integrierte Mikrodatenfiles an der Universität Frankfurt" und später bis 1990 als Teilprojekt B-5 gemeinsam mit dem DIW fortgeführt. Die Befragungen selbst führt Infratest Sozialforschung, München, vor.

nicht alle Personen bereit waren, Interviews zu gewähren, galten grundsätzlich - von wenigen Ausnahmen abgesehen - als Ausfälle.

Für das SOEP-West wurden bis zum Jahr 1992 neun, für das SOEP-Ost drei Wellen erhoben.

2 Das methodische Konzept der Hochrechnung des SOEP und seine Realisierung

Es kann im Rahmen dieses Beitrags nicht über Sinn und Unsinn von Hochrechnungen und Gewichtungen⁴ im allgemeinen diskutiert werden (Vgl. Rendtel, Pötter 1992). Thema ist allein, wie die Hochrechnungen im SOEP realisiert werden.

Für die Gewichtung eines Panels ergeben sich grundsätzlich zwei Zielstellungen⁵: Zum einen müssen für jede Welle Querschnittgewichte bereitgestellt werden, zum anderen sind Längsschnittgewichte für alle möglichen Kombinationen von Befragungszeiträumen erforderlich.

2.1 Die Bestimmung der Auswahlwahrscheinlichkeit als methodischer Hintergrund

Grundgedanke aller Gewichtungen im Rahmen des SOEP ist, daß die Hochrechnungsfaktoren durch den Kehrwert der Auswahlwahrscheinlichkeit einer Stichprobeneinheit bestimmt sind. Dieser Gedanke geht auf Horvitz und Thompson (1952) zurück, die dieses Konzept indes nur für Querschnittsdaten vorstellten. Bei Panelerhebungen wie dem SOEP muß ein Schritt weitergegangen

4) Dazu eine definitorische Anmerkung: Da sich Hochrechnungsfaktoren und Gewichte grundsätzlich nur um eine konstanten Faktor unterscheiden, werde ich im folgenden die Ausdrücke Gewichtung und Hochrechnung zumeist synonym verwenden.

5) Den Hochrechnungen des SOEP liegen Arbeiten von Galler (1987) und Rendtel (1991) zugrunde.

werden, da bei Längsschnitterhebungen die einzelnen Stichprobeneinheiten in der Folgebefragung erneut aufgesucht und befragt werden. Hierdurch entstehen statistische Abhängigkeiten, die es zu berücksichtigen gilt. Der Auswahlprozeß stellt also bei Panel-Studien eine Erweiterung des bei gewöhnlichen Querschnittbefragungen erforderlichen zweistufigen Auswahlprozesses dar (vgl. dazu die Darstellungen von Rendtel 1987 und 1991).

Betrachten wir zunächst die beiden Schritte, die für eine Querschnittgewichtung notwendig sind:

Als erstes erfolgt die Auswahl der Stichprobenelemente im Rahmen des Stichprobendesigns. Das Stichprobendesign bestimmt die Wahrscheinlichkeit, daß ein Element der Grundgesamtheit in der Brutto-Stichprobe enthalten ist. Im zweiten Schritt werden die Wahrscheinlichkeiten bestimmt, daß eine Antwort im Falle der Auswahl gewährt wird. Damit ist der zweistufige Auswahlprozeß für die Querschnitterhebung abgeschlossen. In einer Panelerhebung dagegen folgen für jede weitere Welle zwei neue Stufen. Denn für die Realisation eines Interviews in einer zweiten und folgenden Welle müssen die folgenden Auswahlstufen passiert werden, die für die Hochrechnung zu modellieren sind:

Zum ersten ist die Wahrscheinlichkeit einer erneuten Kontaktaufnahme mit den realisierten Stichprobeneinheiten aus der vorhergehenden Welle zu bestimmen. Zum zweiten muß die Wahrscheinlichkeit einer erneuten Antwortgewährung in der Folgewelle geschätzt werden, sofern die Kontaktaufnahme erfolgreich war. Dieser Ansatz erfordert also neben der Bestimmung der Auswahlwahrscheinlichkeiten auf Basis des Stichprobendesigns die Analyse des Ausfallverhaltens zwischen den Panelwellen.

Eine Ausfallanalyse ist für die erste Panel-Welle schwerer zu realisieren als für die folgenden Befragungswellen, da in der

ersten Welle über die ausgefallenen Stichprobeneinheiten nur sehr wenige Informationen vorliegen (Vgl. dazu auch Rendtel 1988). Um eine möglichst theoriegeleitete Hochrechnung der ersten SOEP-Welle zu ermöglichen, wurde versucht, von "Verweigerern" noch rudimentäre Informationen zu erheben (vgl. auch Wagner et al. 1991)⁶. Für die erste Welle wurden Techniken herangezogen, die externe Informationen bei der Schätzung der Antwortwahrscheinlichkeiten berücksichtigen. Formal entspricht diese Vorgehensweise der Schätzung der Antwortwahrscheinlichkeiten der ersten Panel-Welle (vgl. Anhang Rendtel, Pötter 1992). Nach Abschluß der ersten Panelwelle stehen bei den nun ausfallenden Personen und Haushalten alle bisher erhobenen Merkmale für eine methodisch zufriedenstellende Modellierung für die Gewichtung zur Verfügung.

2.2 Der Horwitz - Thompson Schätzer

Allgemein läßt sich das Ziel einer Hochrechnung darin beschreiben, auf Basis der Stichprobenerhebung auf das Vorkommen interessierender Merkmalskombinationen in der Grundgesamtheit zu schließen. Es ist also im Sinne der Schätztheorie der unbekannte Populationsparameter

$$Y = \sum_{j=1}^N Y_j$$

zu schätzen, wobei die Indikatorvariable Y_j anzeigt, ob die j -te Stichprobeneinheit aus der Grundgesamtheit N die interessierende Merkmalskombination besitzt ($Y_j=1$) oder nicht ($Y_j=0$).

6) Tatsächlich wurden die Informationen lediglich zur Kontrolle der Stichprobenqualität genutzt.

Die Hochrechnung über die inversen Auswahlwahrscheinlichkeiten basiert auf der Zufallsgröße C_j , die angibt, ob die Einheit j der Stichprobe angehört ($C_j=1$) oder nicht ($C_j=0$).

Beschränkt man sich auf lineare Schätzer der Form

$$\hat{Y} = \sum_{j=1}^N \alpha_j C_j Y_j$$

mit

- j = Index für die Individuen,
- N = Gesamtpopulation,
- Y_j = Das interessierende Merkmal,
- C_j = Zufallsgröße, die angibt, ob das Individuum in der Stichprobe ($C_j=1$) ist oder nicht ($C_j=0$) und
- α_j = Hochrechnungsfaktor.

so liefert die Forderung nach Erwartungstreue, also

$$\hat{Y} = \sum_{j=1}^N \frac{1}{P(C_j=1)} C_j Y_j = \sum_{j=1}^n \frac{1}{P(C_j=1)} Y_j$$

mit n = Stichprobenumfang

den Horwitz/Thompson Schätzer

$$\alpha_j = \frac{1}{P(C_j=1)}$$

Ist die Auswahlwahrscheinlichkeit für alle Stichprobenmitglieder gleich dem Auswahlatz n/N , so ergibt sich als Spezialfall die freie Hochrechnung, d.h. die Multiplikation der Besetzungszahlen von Kontingenztabellen mit dem inversen Auswahlatz, also $\alpha_j = N/n$.

Die aus diesem Ansatz resultierende Aufgabe besteht nun darin, für alle Stichprobenelemente ihre Auswahlwahrscheinlichkeit zu bestimmen.

2.3 Die Längsschnittgewichtung

Der Auswahlvorgang durch eine Panelerhebung läßt sich als einen mehrstufigen Prozeß beschreiben, der durch die Anzahl der Panel-Wellen bestimmt ist. Die Auswahl eines Längsschnitt über T Panelwellen wird als Auswahlprozeß mit $2T$ Stufen betrachtet, der folgendermaßen charakterisiert ist. Die Indizes j für die Beobachtungseinheiten werden zur Vereinfachung der Notation fortgelassen.

Bezeichnen wir mit D die Designauswahl, mit R_t den Response und mit K_t die Kontaktaufnahme zum Zeitpunkt t , so kann der Auswahlprozeß wie folgt beschrieben werden:

1. Stufe :Design-Auswahl (Stichprobenanlage)

$$P(D=1)$$

2. Stufe : Antwort in der ersten Welle

$$P(R_1=1|D)$$

3. Stufe: Erfolgreiche Kontaktaufnahme in der 2. Welle

$$P(K_2=1|D, R_1)$$

4. Stufe: Antwort in der 2. Welle

$$P(R_2=1|D, R_1, K_2)$$

⋮
⋮
⋮

$2T$. Stufe: Antwort in der T . Welle

$$P(R_T=1|D, R, K_2, \dots, R_{T-1}, K_T).$$

Für den gesamten Auswahlprozeß ergibt sich die Wahrscheinlichkeit $P(C=1)$ als Produkt der Einzelwahrscheinlichkeiten, also

$$\begin{aligned}
P(C=1) &= P(D=1, R_1=1, K_2=1, R_2=1, \dots, K_T=1, R_T=1) \\
&= P(D=1) \\
&\quad * P(R_1=1 | D) \\
&\quad * P(K_2=1 | D, R_1) \\
&\quad * P(R_2=1 | D, R_1, K_2) \\
&\quad \dots \\
&\quad * P(R_T=1 | D, R_1, K_2, \dots, R_{T-1}, K_T)
\end{aligned}$$

Das Problem, Längsschnitte hochzurechnen, reduziert sich damit auf die Berechnung der Startwahrscheinlichkeit, also der Wahrscheinlichkeit, an der ersten Welle des Panels teilzunehmen, sowie auf die Bestimmung der Bleibewahrscheinlichkeiten. Unter Bleibewahrscheinlichkeit wird die bedingte Wahrscheinlichkeit verstanden, bei Teilnahme an der vorhergehenden Panel-Welle in die nächstfolgende Auswahlstufe zu gelangen und so im Panel zu verbleiben. Sie ist definiert als das Produkt aus der Wahrscheinlichkeit, in einer Folgewelle den Kontakt wiederherzustellen und der Wahrscheinlichkeit, anschließend ein Interview gewährt zu bekommen.

Bezeichnen wir den Hochrechnungsfaktor für eine Person bzw. für einen Haushalt j zum Zeitpunkt t mit $\alpha_j(t)$ und den Kehrwert der entsprechenden Bleibewahrscheinlichkeit mit $\delta_j(t)$, so ergibt sich der Hochrechnungsfaktor für die Periode $t+1$ als Produkt aus beiden Komponenten:

$$\alpha_j(t+1) = \alpha_j(t) \delta_j(t)$$

Auf Basis dieser Gleichung lassen sich Hochrechnungsfaktoren für beliebige Zeiträume (t_1, t_2) bestimmen, sofern Hochrechnungsfaktoren $\alpha_j(t_1)$ für die Welle t_1 und sämtliche Bleibewahrscheinlichkeiten für den Zeitraum (t_1, t_2) bekannt sind.

Da Daten aus der Vorperiode für die Personen und Haushalte vorliegen, die in der Folgewelle ausfallen, lassen sich die Wahrscheinlichkeiten zur erneuten Kontaktaufnahme und die Bleibewahrscheinlichkeiten über ein Logit-Modell schätzen, das personenbezogene Merkmale PX und haushaltsbezogene Variable HX berücksichtigt, welches die allgemeine Form

$$\ln \frac{P_j}{1-P_j} = PX_j' B_1 + HX_j' B_2$$

hat, wobei P_j für die zu schätzenden Wahrscheinlichkeiten und B_1, B_2 für die zu schätzenden Parameter stehen.

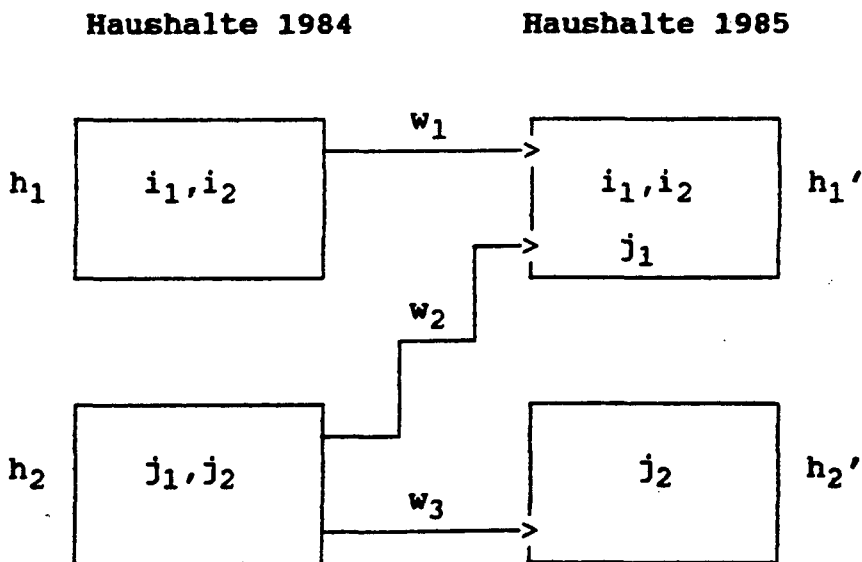
Mit diesem Ansatz ist das Problem der Längsschnittgewichtung prinzipiell gelöst.

2.4 Die Querschnittgewichtung ab der 2. Welle

Ein wenig problematischer ist die Bestimmung der Querschnittgewichte ab der zweiten Welle.

Die Situation sei beispielhaft an den Auswahlwahrscheinlichkeiten für Haushalte in der zweiten Welle des SOEP-West dargestellt. Hierzu denke man sich die Haushalte der 1984 und 1985er Grundgesamtheit nebeneinander aufgereiht.

Schaubild 1: Weiterverfolgungspfade zwischen Haushalten



Zwischen einigen Haushalten bestehen insofern Beziehungen, daß in ihnen mindestens eine gemeinsame Person wohnt. Trägt man nun die Verbindungen zwischen den Haushalten als Pfeile ein, die von den Haushalten des Jahres 1984 ausgehen, so erhält man alle möglichen Pfade, die im Rahmen des Weiterverfolgungskonzeptes des Panels realisiert werden können.

Im 1984er Haushalt h_1 leben die zwei Personen i_1 und i_2 , im Haushalt h_2 die Personen j_1 und j_2 . Zwischen 1984 und 1985 trete nun folgende Veränderung ein: Person j_1 ziehe in den ehemaligen Haushalt h_1 , der nun h_1' heißt. Der Haushalt h_1' hatte also zwei Möglichkeiten, im Jahr 1985 Mitglied des SOEP zu sein. Einmal über die Auswahl des Haushalts h_1 in der ersten Welle mit anschließender Weiterverfolgung über den Weg w_1 . Zum anderen kann er über die Auswahl des Haushalts h_2 in der 1. Welle und wegen der Weiterverfolgung der Person j_1 über den Weg w_2 im Jahr 1985 in die Panelstichprobe gelangen.

Geht man von der Unabhängigkeit der Auswahl der Haushalte bei Start des Panels aus, so ergibt sich die Wahrscheinlichkeit, den Haushalt h_1' in der '85er Panelstichprobe zu haben, als Summe der Einzelwahrscheinlichkeiten, nämlich die Haushalte h_1 oder h_2 auszuwählen und die Wege w_1 bzw. w_2 erfolgreich durchzuführen. Indes stehen Informationen über den Haushalt h_2 , den die Person j_1 verlassen hat und die dessen Auswahlwahrscheinlichkeit in der ersten Welle unter Umständen erheblich beeinflussen, nicht zur Verfügung. Die Auswahlwahrscheinlichkeit des Haushalts h_2 erklärt sich aber wieder über personenbezogene Merkmale PX_j der zugezogenen Person j und über deren haushaltsbezogene Variablen HX_j des Haushalts, in den er einzieht. Zieht man als Schätzansatz für die Auswahlwahrscheinlichkeit P_j wiederum ein Logit-Modell heran, fehlen haushaltsbezogene Merkmale HX_j' für den Haushalt h_2 , aus dem die Person j ausgezogen ist. Diese Merkmale übernehmen deshalb die Funktion eines Störterms $\epsilon_j = HX_j' \beta_2$, so daß sich der Logit-Ansatz auf

$$L_j = \ln \frac{P_j}{1-P_j} = PX_j' \beta_1 + \epsilon_j$$

reduziert. Gilt diese Beziehung sowohl für Personen aus Panelwelle t-1 als auch für die in der Folgewelle t hinzugekommenen Personen, so sind neben den personenbezogenen Merkmalen auch die Logits L_j bekannt, bzw. lassen sich diese aus den Hochrechnungsfaktoren α_j über die Beziehung $P_j = 1/\alpha_j$ bestimmen. Die Schätzung der Auswahlwahrscheinlichkeiten erfolgt dann nach folgendem Schema ab:

1. Die Querschnittsgewichte für die Welle t-1 seien bekannt. Aus ihnen lassen sich durch eine einfache Retransformation die Logits $L_j(t-1)$ berechnen.

2. Es wird ein geeigneter Satz personenbezogener Daten $PX_j(t-1)$ ausgewählt, der für jede Welle erhoben wird.

3. Die Logits der Periode t-1 werden mit einer OLS-Schätzung durch $PX_j(t-1)$ erklärt. Man erhält einen Parametervektor

$$\hat{\beta}_1(t-1) = f(L, PX, t-1).$$

4. Damit können die Auswahlwahrscheinlichkeiten P_j der Periode t für neu in das Panel eintretende Personen geschätzt werden, also

$$\hat{L}_j(t) = PX_j'(t) \cdot \hat{\beta}_1(t-1)$$

$$\hat{\alpha}_j(t) = (1 + \exp(\hat{L}_j(t))) / \exp(\hat{L}_j(t)), \text{ da } \alpha_j = 1/P_j(t) \text{ gilt.}$$

5. Der Hochrechnungsfaktor für den Haushalt i, in den die Person j in der Periode t eingezogen ist, ergibt sich dann als reziproker Wert der Summe zweier Auswahlwahrscheinlichkeiten

$$\hat{\alpha}_j(t) = \frac{1}{\hat{p}_i(t) + \hat{p}_j(t)} = \frac{1}{1/\hat{\alpha}_i(t) + 1/\hat{\alpha}_j(t)}$$

3 Die Realisierung des Konzepts

3.1 Die Gewichtung der Startwellen

Wie oben ausgeführt, bedürfen die Gewichtungen der Startwellen einer gesonderten Behandlung, da die Auswahlwahrscheinlichkeit der Stichprobenelemente zu einem großen Teil über externe Informationen zu schätzen sind. Durch die spätere Einbeziehung der Haushalte der damaligen DDR mußte dieser Prozeß im SOEP zweimal durchgeführt werden. Auch wenn für beide Teilstichproben im Prinzip das gleiche Verfahren angewendet wurde, gab es doch in der praktischen Durchführung einige Unterschiede, da sowohl die internen Daten - d.h. die Informationen aus der Panelbefragung - als auch die externen Statistiken - d.h. die Rahmendaten aus der amtlichen Statistik - nicht vergleichbar sind.

3.1.1 Die Gewichtung der ersten Welle des SOEP-West

Für die Gewichtung von Startwellen ist die Berücksichtigung des Ziehungsverfahrens der Stichprobenelemente unerlässlich. Für die erste Welle des SOEP-West stimmte das Erhebungsdesign für die Stichprobe A, das sind die Haushalte, deren Haushaltsvorstand weder die türkische, italienische, spanische, jugoslawische noch griechische - also meist die deutsche - Nationalität haben, weitgehend mit dem ADM-Ziehungsverfahren überein (vgl. Kirschner 1984). Dieses basiert auf der Auswahl von Stimmbezirken einer möglichst aktuellen Bundestagswahl. Es ist gekennzeichnet durch ein zweistufiges Auswahlverfahren. Die Primäreinheiten, auch Sample-Points genannt, sind in der Stichprobe A die Wahlbezirke. Die Ziehung der Sekundäreinheiten,

also der Haushalte der ersten Welle, geschah über eine Random-Route-Begehung innerhalb der Stimmbezirke.

Für die Stichprobe B (Ausländerhaushalte) war dieses Verfahren nicht anwendbar, da Ausländer kein Wahlrecht besitzen. Deshalb mußten als Primäreinheiten Kreise bzw. kreisfreie Städte gewählt werden. Eine Personenauswahl aus dem Ausländer-Register des jeweiligen Kreises schloß sich an. Die Wohnadresse der ausgewählten Personen führte zu den Starthaushalten, den Sekundäreinheiten der Stichprobe B.

Vor der Ziehung der Primäreinheiten wurden diese geschichtet. In der Stichprobe A wurden hierfür aus den Merkmalen Bundesland, Regierungsbezirk und Boustedt-Region insgesamt 148 Regionalzellen gebildet. Die Schichtgröße wurde proportional zur Anzahl der Haushalte in den jeweiligen Regionalzellen bestimmt. Sodann wurden die Primäreinheiten innerhalb jeder Regionalzelle durch systematisches, Größenproportionales Ziehen bestimmt.

In der Stichprobe B wurden innerhalb jeder Ausländergruppe (Türken, Italiener, Spanier, Jugoslawen und Griechen) die Kreise nach Bundesland und Regierungsbezirk angeordnet. Die Primäreinheiten wurden dann wieder durch systematisches, Größenproportionales Ziehen ausgewählt.

Insgesamt wurden 584 Sample-Points für Stichprobe A und 240 für die Stichprobe B ausgewählt. Durch die Anordnung der Primäreinheiten und das systematische Ziehungsverfahren erhält man eine Art Netz, das das Erhebungsgebiet gleichmäßig überdeckte. Die Größenproportionalen Ziehungswahrscheinlichkeiten auf der Ebene der Primäreinheiten standen im Zusammenhang mit der Vorgabe, in jeder Primäreinheit eine feste Anzahl - angestrebt wurden acht Haushalte - von Sekundäreinheiten zu ziehen.

Die Art der Stichprobenziehung bestimmt den ersten Teil der Gewichtung. Er umfaßte Design- und Regionalgewichtung: Zunächst mußten Designgewichte für die insgesamt 824 Sample-Points ermittelt werden. Dies geschah durch proportionale Anpassung der einzelnen Sample-Points mit den Ausschöpfungsquoten. Anschließend erfolgte - nunmehr auf Haushaltsebene - eine Korrektur der Auswahlwahrscheinlichkeit, die das Design der Stichprobe erforderte. Denn die Auswahl der Ausländerhaushalte war zweifach überhöht: Einmal war sie disproportional angelegt und zum anderen bewirkte das personenbezogene Ziehungsverfahren eine Bevorzugung der größeren Haushalte. Ihre Auswahlwahrscheinlichkeit war damit proportional zur Anzahl der Personen über 16 Jahre. Nach diesem Schritt folgte die regionale Anpassung nach der Bevölkerung in Bundesländern und Regierungsbezirken für alle 5921 Netto-Haushalte, also der Haushalte, für die die Befragung erfolgreich durchgeführt werden konnte. Abgeschlossen wurde diese Phase mit der Korrektur der Auswahlwahrscheinlichkeit der Haushalte mit einem zweiten Wohnsitz. Diese besaßen im Vergleich zu Haushalten ohne Zweitwohnsitz *ceteris paribus* ungefähr die doppelte Wahrscheinlichkeit, in die Panelstichprobe zu gelangen. Im folgenden werden die so erstellten Gewichte vereinfacht als Designgewichte bezeichnet.

Die Designgewichte gingen als Startwerte für die Schätzung des endgültigen Hochrechnungsrahmens ein, der insgesamt 316 Restriktionen aus der amtlichen Statistik erfüllen sollte. Die verwendeten Restriktionen stammen aus den Mikrozensus von 1982 und 1985, aus denen durch Interpolation Daten für 1984 geschätzt wurden und aus der 1984er EG-Arbeitskräfte-Stichprobe.⁷

Das Anpassungsverfahren basiert auf dem Prinzip des minimalen Informationsverlustes, dessen Idee kurz skizziert werden soll (vgl. Merz 1983):

7) Der Mikrozensus 1984 wurde nicht erhoben.

Gegeben sei ein Vektor

$$q = [q_j] \quad j = 1, \dots, n$$

mit den Designgewichten für die $n=5921$ Haushalte.

Weiterhin wird ein Vektor mit den Eckdaten (Restriktionen)

$$r = [r_i] \quad i = 1, \dots, m$$

vorgegeben. Ein Element $r(i)$ des Vektors r gibt an, wieviele Haushalte es mit der Eigenschaft i in der Grundgesamtheit insgesamt gab. Der Wert für m hatte in unserem Fall also den Wert 316.

Schließlich ist eine Matrix

$$S = [s_{i,j}] ,$$

erforderlich, deren Elemente angeben, wie oft in einem Haushalt j die Eigenschaft i im Bezugsjahr vorkam.

Gesucht wird nun der Vektor

$$\alpha = [\alpha_j]$$

mit den endgültigen Hochrechnungsfaktoren, der unter Einhaltung der Nebenbedingung

$$S \alpha = r$$

die Zielfunktion

$$Z = \sum_{j=1}^N \alpha_j \log(\alpha_j/q_j) = \min!, \quad \alpha_j, q_j > \alpha_0$$

minimiert, wobei α_0 eine frei zu wählende unterere Schranke für jeden Hochrechnungsfaktor angibt. Allgemein gesprochen wird gefordert, den Vektor q möglichst wenig zu verändern, aber doch so, daß die die Restriktionen $S \alpha = r$ erfüllt werden.

Der Hochrechnungsrahmen für die Startwelle des SOEP-West berücksichtigt folgende Merkmale:

Haushaltsmerkmale:

- Haushalte mit deutschem Haushaltsvorstand (HV) nach Geschlecht, Alter des HV, Haushaltsgröße,
- Haushalte mit ausländischem HV nach Haushaltsgröße und nach Nationalität des HV.

Personenmerkmale:

- Personen nach Alter, Geschlecht, Familienstand und Nationalität,
- Ausl. Anstaltspersonen nach Geschlecht,
- (Berufs-)Schüler nach Schulart und Geschlecht,
- Erwerbstätigkeit nach Alter, Geschlecht, Nationalität und Art der Erwerbstätigkeit.

Die Hochrechnungsfaktoren α_j wurden zunächst für die Haushalte bestimmt. Dabei wurde für die Untergrenze α_0 ein Wert von 60 festgelegt.⁸ Da von Ausnahmen abgesehen in der Startwelle sämtliche Personen eines Haushaltes am Panel teilgenommen hatten, konnten die Haushaltsgewichte für sämtliche Haushaltsmitglieder - also auch auf die Kinder unter 16 Jahre - übernommen werden.

3.1.2 Die Gewichtung der ersten Welle des SOEP-Ost

Die Gewichtung der Startwelle des SOEP-Ost (Stichprobe C) erfolgte ähnlich zu derjenigen des SOEP-West erfolgen (Vgl. Pischner 1991). Allerdings waren hier die internen und externen statistischen Gegebenheiten von anderer Qualität. Zunächst einige Informationen über die Stichprobenziehung: Die Grundgesamtheit der DDR-Basisbefragung bildeten alle Privathaushalte mit deutscher Bezugsperson und alle in ihnen wohnenden Personen, die zum Befragungszeitpunkt mindestens 16 Jahre alt waren.

8) Der kleinste Hochrechnungsfaktor mit dem Wert 62 wurde für einen Ausländerhaushalt geschätzt. Die Ausländerhaushalte der ersten Welle hatten im Durchschnitt einen Hochrechnungsfaktor von ca. 700; der Durchschnitt aller Haushalte betrug ca. 2350.

Die mehrfach geschichtete Stichprobe umfaßte schließlich 753 Sample-Points, die wiederum über Zufallsauswahl der Startadressen und Random-Route zu 3616 Bruttoadressen führten. Nach Abschluß der Feldarbeit lagen für 2179 Haushalte und 4453 Personen Interviews vor. Im für die Hochrechnung zur Verfügung stehendem Datensatz fehlten Informationen über die Ausfallquoten in den Sample-Points, so daß eine Designgewichtung - wie für das SOEP-West vorgenommen - nicht möglich war. Allerdings hatte sich bei der Analyse der Hochrechnungsfaktoren des SOEP-West gezeigt, daß der Ausgleich zwischen Brutto- und Nettostichprobe allein schon zu einer deutlichen Erhöhung der Varianz in den Hochrechnungsfaktoren führte. So verzichteten wir auf die Designgewichtung und berücksichtigten stattdessen die in ihr enthaltene Regionalgewichtung in den Eckdaten für den iterativen Anpassungsprozeß, wodurch die Varianz der Hochrechnungsfaktoren klein gehalten wird..

Als Startgewichte wurden die für alle Einheiten gleichen, freien Hochrechnungsfaktoren $q_j = N/n$ verwendet. Aufgrund der schlechten statistischen Datenlage, die zudem noch durch die umfangreichen Wanderungsbewegungen nach Öffnung der Grenzen weiter entwertet war, definierten wir nur 115 Restriktionen für die Schätzung der Hochrechnungsfaktoren. Der Hochrechnungsrahmen wurde aus den folgenden Merkmalen gebildet:

Personenmerkmale:

- Verteilung der Wohnbevölkerung nach Geschlecht und Bezirken,
- Verteilung der Wohnbevölkerung nach Geschlecht, Alter und Familienstand.

Haushaltsmerkmale:

- Gesamtzahl der Haushalte in der DDR.

Auffallend ist, daß - mangels gesicherter Strukturdaten - nur die Gesamtzahl der Haushalte, nicht aber ihre Struktur in die Hochrechnung eingeht. Ihre Bestimmung wurde dem mit der Haushaltsstruktur korrelierendem Einfluß der demographischen Merk-

male - mit Erfolg - überlassen. Z.B. war der Anteil der Ein-Personenhaushalte in der Stichprobe mit 15,2% viel zu niedrig, nach Anpassung betrug er 25,2%, ein Wert, der sich inzwischen als recht gut erwies.⁹

2.2.2 Die Schätzung der Bleibewahrscheinlichkeiten

Die Bleibewahrscheinlichkeit ist das Produkt der Wahrscheinlichkeit der Kontaktaufnahme und der Antwortgewährung zwischen zwei Wellen. Beide Faktoren müssen getrennt modelliert werden, da sie inhaltlich stark voneinander abweichen.

Das Verfahren an sich ist jedoch in beiden Fällen gleich. Es werden die bereits betrachteten Logit-Modelle mit den Variablen geschätzt, die den vermutlich stärksten Einfluß auf Adreßermittlung bzw. Antwortgewährung haben.

Die verwandte Modellsuche eliminiert aus einem konstanten Satz von Variablen jeweils die Variable mit der geringsten Signifikanz, bis alle verbleibenden Variablen auf einem Niveau von 5% signifikant sind.

3.2.1 Die Schätzung der erneuten Kontaktaufnahme

Die folgende Übersicht zeigt die Variablen, die grundsätzlich für die Schätzung der Wahrscheinlichkeit (in %), einen Panelhaushalt wiederzufinden, verwendet werden.

9) Das statistische Bundesamt wies auf Basis von nicht dokumentierten Schätzungen in der Ausgabe 9/91 von "Zahlen-Fakten-Trends" ein Anteil der Ein-Personenhaushalte von 24,9 vH aus.

Übersicht 1

Wahrscheinlichkeiten einer erfolglosen Kontaktaufnahme in Abhängigkeit verschiedener Merkmale in der 2. Welle

Merkmal	Keine Kontaktaufnahme in	
	SOEP-West 1985	SOEP-Ost 1991
	%	%
- Insgesamt	1,9	1,5
- Nationalität		
Deutscher Haushaltsvorstand	1,5	-
Ausländischer Haushaltsvorstand	3,2	-
- Regionen nach Boustedt West/Ost		
Kerngebiet mit mehr als 500 000/100 000 Einwohnern	2,8	1,6
Nicht Kerngebiet mehr als 500 000/100 000 Einwohnern	1,4	1,4
- Haushaltsgröße		
Ein-Personenhaushalt	3,4	0,9
Zwei-Personenhaushalt	1,2	0,8
Drei-Personenhaushalt	1,2	2,4
Vier-Personenhaushalt und größer	1,9	1,5
- Typ des Wohnumfelds		
Ländlich, einsam	0,7	1,2
Hochhaus oder ähnliches	2,5	1,8
sonstiges	2,2	1,4
- Mobilität		
Haushalt ohne Adreßveränderung	0,8	0,0
Verzogener Mehr-Personen-Haushalt	7,4	11,1
Verzogener Ein-Personen-Haushalt	21,0	14,3
Neuer Haushalt	11,8	25,6
Quelle: Das Sozio-oekonomische Panel, Wellen 1 und 2, eigene Berechnungen.		

Die Merkmale, die eine erneute Kontaktaufnahme überdurchschnittlich beeinträchtigen, sind in der Übersicht hervorgehoben. Wie nicht anders zu erwarten, sind verzogene und neue Haushalte am schwersten erneut zu kontaktieren. Aber auch Haushalte mit ausländischem Haushaltsvorstand gehen überdurchschnittlich oft verloren.

Im SOEP-West hat sich die Wahrscheinlichkeit einer erfolglosen Kontaktaufnahme inzwischen deutlich verringert und bei etwa 1%

eingependelt. Ursache hierfür ist in erster Linie eine erfolgreichere Kontaktaufnahme bei verzögerten Haushalten.

3.2.2 Die Schätzung der Antwortgewährung

Übersicht 2 zeigt den Variablensatz zur Erklärung der Antwortgewährung bzw. -verweigerung

Übersicht 2

Wahrscheinlichkeiten einer Antwortverweigerung in Abhängigkeit verschiedener Merkmale in der 2. Welle

Merkmal	Antwortverweigerung in %	
	SOEP-West 1985	SOEP-Ost 1991
- Insgesamt	10,4	8,3
- Stichprobenkennzeichen		
Stichprobe A	10,2	-
Stichprobe B	10,9	-
- Alter des Haushaltsvorstands		
Älter als 74 Jahre	17,5	18,1
65 - 74 Jahre	9,9	8,0
55 - 64 Jahre	9,7	6,9
35 - 54 Jahre	9,3	6,3
25 - 34 Jahre	8,9	9,3
16 - 24 Jahre	19,9	13,7
- Geschlecht des Haushaltsvorstand		
männlich	9,8	7,6
weiblich	12,4	8,9
- Haushaltstyp		
Kein Umzug	9,6	7,8
Als ganzer Haushalt umgezogen	11,6	8,9
Haushalt abgespalten	29,2	24,6
- Interviewerstabilität		
Wechsel der Interviewernummer	15,3	8,5
Kein Wechsel der Interviewernr.	7,7	5,5
- Belastung durch Krankheiten etc.		
Keine Belastung	10,3	-
Belastung vorhanden	13,6	-

noch Übersicht 2

- Vermögensanlagen		
keine Vermögensanlagen	12,5	-
1 Vermögensanlage	13,2	-
2 Vermögensanlagen	8,3	-
3 Vermögensanlagen	8,7	-
4 Vermögensanlagen	7,6	-
5 Vermögensanlagen	9,4	-
- Sozialer Status des Haushaltsvorstands		
Sonstige	9,1	7,8
Nicht erwerbstätig	12,9	10,8
Arbeitslos	9,8	14,3
Höchste Statusgruppen	7,9	5,5
Niedrigste Statusgruppen	10,7	9,4
Keine Angabe	27,6	-
- Nettohaushaltseinkommen West/Ost		
Angabe verweigert	17,9	9,4
bis 1000/800 DM	13,8	14,3
1001 - 2000/801 - 1200 DM	11,2	8,7
2001 - 3000/1201 - 1800 DM	8,3	8,1
3001 - 4000 DM/1801 - 2500	9,6	6,3
4001/2501 und mehr DM	8,2	6,6
Quelle: Das Sozio-oekonomische Panel, Wellen 1 und 2, eigene Berechnungen.		

Hier zeigt sich ganz allgemein, daß Haushalte, die sich in schwieriger Lage oder allgemein in einer Streßsituation befinden, tendenziell häufiger ein erneutes Interview verweigerten. Alte Leute, Paare, die sich getrennt haben; Menschen, die ihre Arbeit verloren haben, haben verständlicherweise ein geringeres Interesse an der Panelbefragung erneut teilzunehmen. Weiterhin ist die Wahrscheinlichkeit einer allgemeinen Antwortverweigerung dann hoch, wenn schon in der vorherigen Befragung sensible Fragen, wie die nach der Höhe des Einkommens, verweigert wurden. Sehr bedeutend ist auf jeden Fall der Interviewereinfluß: Ein vertrautes Gesicht erhöht die Aussichten auf ein Interview beträchtlich (Vgl. dazu Rendtel 1988, 1990).

2.2.2.3 Die Schätzung der Auswahlwahrscheinlichkeit neuer Haushalte für die Querschnittsgewichtungen ab der 2. Welle

Informationen liegen natürlich nur über Personen vor, die in einen Panelhaushalt zugezogen sind und ein Interview gegeben haben. Sie fehlen selbstverständlich von all den Leuten, die nicht in einen Panelhaushalt gezogen sind. Deshalb ist eine Logit-Schätzung zur Bestimmung der Auswahlwahrscheinlichkeit - wie vorhin bereits ausgeführt - nicht möglich und es muß auf eine OLS-Schätzung zurückgegriffen werden, die aus den Hochrechnungsfaktoren der vorhergehenden Wellen abgeleitete Logits (bzw. Wahrscheinlichkeit) mit personenbezogenen Merkmale regressiert. Als Variablen wurden Nationalität, Geschlecht, Alter und Schulbildung ausgewählt. D.h. es wird über alle Personen der Stichprobe bestimmt, wie hoch die Hochrechnungsfaktoren in Abhängigkeit von persönlichen Merkmalen sind. Diese Merkmale erklärten ca. 70% der Varianz. Es wird dann angenommen, daß die - unbekannten - Hochrechnungsfaktoren der "neuen Personen" dieser Struktur folgen.

Nun konnten die Auswahlwahrscheinlichkeiten der Haushalte mit neu hinzugekommenen Personen modifiziert werden. Sie wurden im Durchschnitt verdoppelt. Die Querschnittsgewichte dieser Haushalte ergeben sich dann analog zu den Längsschnittgewichten als reziproker Wert des Produkts aus Auswahl- und Bleibewahrscheinlichkeit.

Im Gegensatz zur Längsschnittgewichtung ist bei den Querschnittsgewichtungen eine jährliche Validierung mit externen Daten z.B. des Mikrozensus möglich. Allerdings wurde nur die Gesamtzahl der geschätzten Haushalte und Personen aus dem Panel an die entsprechenden Zahlen des Mikrozensus angepaßt. Der resultierende Korrekturfaktor schwankte im Bereich von 0,98 bis 1,02. Auf diese Weise erreichten wir, daß die geschätzten Strukturen der Panelquerschnitte beibehalten werden, ohne daß

die Gesamtpopulationen (Panel versus Mikrozensus) auseinander driften können.

2.2.2.4 Die Behandlung temporärer Ausfälle

Personen oder Haushalte, die temporär ausfallen aber später wieder an der Befragung teilnehmen, gehen für gewichtete Längsschnittanalysen zwar verloren, werden aber bei den Querschnittsgewichtungen berücksichtigt. Die Berechnung der Hochrechnungsfaktoren erfolgt dann über eine Schätzung der Wiedereintrittswahrscheinlichkeit.

4 Hochrechnungsvariablen im Datensatz

Für die Gewichte sind im Datensatz des SOEP zwei Records bereitgestellt worden: PHRF für die Personenbezogenen HochRechnungsFaktoren, und HHRH für die haushaltsbezogenen. In den Records sind Querschnitt-, Längsschnittgewichte und die Kehrwerte der Bleibewahrscheinlichkeiten für alle Wellen enthalten.

Die Namen für die Gewichtungsvariablen setzen sich aus Wellenlabels, Haushalts-/Personenindikator und Variablenart zusammen. So bezeichnen einstellige Wellenlabels A, B, ... die Querschnittgewichte für die Wellen 1, 2, ... Sind die Wellenlabel zweistellig beginnend mit A, so ergeben sich die Bezeichnungen für die Längsschnitte AB, AC, ... für die Zeiträume erste bis zweite, dritte usw. Welle. Der Haushalts-/Personenindikator folgt unmittelbar dem Wellenbezeichner und ist entweder ein H oder ein P.

Am Ende des Variablennamens steht die Variablenart: HRF ist die Endung für Quer- und Längsschnittgewichte, die nach den be-

schriebenen Verfahren geschätzt wurden.¹⁰ Die Endung BLEIB kennzeichnet die Variablen mit den Kehrwerten der Bleibewahrscheinlichkeiten. Mithilfe dieser Variablen können - wie bereits erwähnt - beliebige Längsschnittsfaktoren, die nicht mit der ersten Welle beginnen, vom Anwender selbst berechnet werden.

Für die erste Welle SOEP-West stehen schließlich noch die Designgewichte AHDESREG und APDESREG zur Verfügung, die lediglich die Auswahlwahrscheinlichkeiten der Stichprobenanlage sowie die Regionalmerkmale von ausgefallenen und teilnahmebereiten Haushalten bei Start des Panels berücksichtigen.

5 Schlußbemerkung

Sicherlich enthält das vorgestellte Konzept einige Vereinfachungen und pragmatischen Ansätze, doch wir glauben, den wesentlichen Anforderungen einer konsistenten Quer- und Längsschnittgewichtung gerecht geworden zu sein. Es gibt zwei Punkte, die besondere Sorgfalt erfordern: Zum ersten ist die Gewichtung der Startwelle besonders kritisch, da Fehler, die hier gemacht wurden, sich in jeder weiteren Welle fortpflanzen.

Zum anderen ist die Auswahl der Variablen, die zur Schätzung der Kontaktaufnahme und zur Antwortgewährung herangezogen werden, entscheidend.

Der Anwender der im Datensatz zur Verfügung gestellten Hochrechnungsfaktoren muß in alleiniger Verantwortung prüfen, ob diese für seine Fragestellungen angemessen sind. Nicht wenige Analysen - vor allem bedingte Wahrscheinlichkeiten schätzende Regressionsverfahren - können natürlich gänzlich ohne jede Gewichtung durchgeführt werden. Da im SOEP-Datensatz alle

10) Es gibt noch Querschnittsgewichte mit der Endung IHRF, die Infratest berechnet und mit den Datensätzen geliefert hat. Auf diese Gewichte wurde in diesem Beitrag nicht eingegangen.

Variablen vorhanden und zugänglich sind, die für eigenen Ausfallanalysen und Gewichtung notwendig sind, kann jeder SOEP-Nutzer die für ihn optimale Hochrechnung auch selbst durchführen bzw. in multivariaten Analysen auch den Ausfallprozeß mit modellieren (z.B. Verbeek 1990).

Literatur:

- Galler, Heinz P. (1987): Zur Längsschnittgewichtung des Sozio-oekonomischen Panels. In: Krupp/Hanefeld, Lebenslagen im Wandel: Analysen 1987, Band 2 der Reihe: Sozio-oekonomische Daten und Analysen für die Bundesrepublik Deutschland, Frankfurt, S.295-317.
- Horvitz, D.G. und D.J. Thompson (1952): A Generalization of Sampling without Replacement from a Finite Universe. In: Journal of the American Statistical Association, Volume 47, S. 663-685.
- Infratest Sozialforschung (1985): Das Sozio-oekonomische Panel, Welle 1, Methodenbericht zur Haupterhebung, München.
- Kirschner, H. P. (1984) Allbus 1980: - Stichprobenplan und Gewichtung. In: K. U. Mayer und P. Schmidt (Hg.), Allgemeine Bevölkerungsumfrage der Sozialwissenschaften - Beiträge zu methodischen Problemen des Allbus 1980, Frankfurt, S. 114-182.
- Merz, Joachim (1983): Die konsistente Hochrechnung von Mikrodaten nach dem Prinzip des minimalen Informationsverlustes. In: Allgemeines Statistisches Archiv. Band 67, Heft 4/1983, S.342-366.
- Pischner, Rainer (1991): Eine konsistente Haushalts- und Personengewichtung für die DDR-Basisbefragung des SOEP und für die Ost-Pilotstudie des Wohlfahrtssurveys. In: Vierteljahrshefte zur Wirtschaftsforschung, 1991 Heft 1-2, Berlin, S. 50-64.
- Projektgruppe SOEP (1991): Das Sozio-oekonomische Panel (SOEP) im Jahre 1990/91. In: Vierteljahrshefte zur Wirtschaftsforschung, 1991 Heft 3-4, Berlin, S. 146-155.
- Rendtel, Ulrich (1987): Methodische Konzepte für die Hochrechnung von Panel-Daten. In: Vierteljahrshefte zur Wirtschaftsforschung, 1987 Heft 4, Berlin, S. 278-290.
- Rendtel, Ulrich (1988): Panelmortalität - Eine Analyse der Antwortausfälle in der 2. und 3. befragungswelle. In: Vierteljahrshefte zur Wirtschaftsforschung, 1988 Heft 1-2, Berlin, S. 37-59.
- Rendtel, Ulrich (1990): Teilnahmebereitschaft in Panelstudien: Zwischen Beeinflussung, Vertrauen und Sozialer Selektion. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie 42, S. 280-299.
- Rendtel, Ulrich (1991): Die Schätzung von Populationswerten in Panelerhebungen. In: Allgemeines Statistisches Archiv. Band 75, Heft 3/1991, S.225-244.
- Rendtel, Ulrich und Ulrich Pötter (1992): Über Sinn und Unsinn von Repräsentativitätsstudien. DIW-Diskussionspapier 61, Berlin.
- Verbeek, Marno (1990): On the Estimation of a Fixed Effect Model with Selectivity Bias. In: Economic Letters, 34. S. 267-270.
- Wagner, Gert, Jürgen Schupp und Ulrich Rendtel (1991): Das Sozio-ökonomische Panel - Methoden der Datenproduktion und Aufarbeitung im Längsschnitt. In: R. Hauser et al. (Hg.), Mikroanalytische Grundlagen der Gesellschaftspolitik - Erhebungsverfahren, Analysemethoden und Mikrosimulation, Weinheim (auch als DIW-Diskussionspapier Nr. 31 erschienen).

Dokumentation der für das SOEP benutzten Modelle zur Berechnung der Ausfallwahrscheinlichkeiten

Ulrich Rendtel

26. Juli 1993

Inhaltsverzeichnis

1	Ausfälle durch Kontaktverlust	2
1.1	Ausfallraten nach Mobilitätsverhalten	2
1.2	Definition der Modellvariablen	4
1.3	Geschätzte Wahrscheinlichkeiten für einen Ausfall durch Kontakt- verlust	5
2	Ausfälle durch Antwortverweigerung	7
2.1	Ausfallraten nach verschiedenen Haushaltsmerkmalen	7
2.2	Definition der Modellvariablen	15
2.3	Geschätzte Wahrscheinlichkeiten für einen Ausfall durch Antwort- verweigerung erreichter Haushalte	17

1 Ausfälle durch Kontaktverlust

In jeder Panelwelle ist es notwendig, mit den Haushalten des Vorjahres wieder Kontakt aufzunehmen. Dabei wird festgestellt, ob:

- der Haushalt noch an der alten Adresse lebt
- der Haushalt umgezogen ist
- der Haushalt sich aufgespalten hat
- der Haushalt verstorben ist
- der Haushalt das Erbungsgebiet verlassen hat
- die Haushaltsmitglieder in einen bestehenden Panelhaushalt zurückgekehrt sind

1.1 Ausfallraten nach Mobilitätsverhalten

Die folgende Tabelle dokumentiert den Erfolg der Feldarbeit bei der Kontaktaufnahme. Basis sind alle Haushalte der letzten Welle, die nicht verstorben sind oder ins Ausland gezogen sind; weiterhin alle abgespaltenen Haushalte. Ein Kontakt gilt als hergestellt, falls ein Interview erreicht wird bzw. die Verweigerung eines Interviews im Adressprotokoll vermerkt wird. Weiterhin gilt ein Kontakt als wiederhergestellt, falls der Rückzug der Haushaltsmitglieder in einen bestehenden Panelhaushalt festgestellt wird.

Tabelle 1: Ausfallraten aufgrund von Kontaktverlust im SOEP-West (Sample A und B). *N* = Anzahl der zu erreichenden Haushalte; % = Prozentsatz der Haushalte ohne Kontaktaufnahme

	Welle 2	Welle 3	Welle 4	Welle 5	Welle 6	Welle 7	Welle 8	Welle 9
	Insgesamt							
N	6051	5814	5465	5342	5156	5044	5029	5006
%	1,9	1,4	1,0	0,9	0,9	0,9	0,5	0,4
	Haushalte ohne Adressenänderung							
N	5413	5039	4808	4683	4545	4472	4448	4447
%	0,8	0,4	0,1	0,1	0,2	0,0	0,04	0,0
	verzogene Mehrpersonen-Haushalte							
N	298	307	272	274	228	186	197	195
%	7.4	3.6	4.0	5.5	0.5	1.6	0.5	0,5
	verzogene Einpersonen-Haushalte							
N	119	180	142	143	126	122	94	90
%	21,0	14,4	7,7	5,6	4,7	5,7	1,1	0.0
	abgespaltene Haushalte							
N	221	295	242	242	246	263	290	273
%	11,7	8,4	10,4	7,4	11,8	12,9	7,6	7,3

Die folgende Tabelle vergleicht das SOEP-West (Stichproben A und B) mit dem SOEP-Ost (Stichprobe C)

Tabelle 2: Vergleich der Ausfallraten durch Kontaktverlust zwischen SOEP-West und SOEP-Ost in der 2. und 3. Befragungswelle

Merkmal	Welle 2				Welle 3			
	West		Ost		West		Ost	
	N	%	N	%	N	%	N	%
insgesamt	6051	1.9	2246	1.5	5814	1.4	2304	0.5
Haushalt ohne Umzug	5413	0.8	2062	0.0	5039	0.4	2043	0.05
umgezogener Mehrpers.-H.	298	7.4	81	11.1	307	3.6	106	0.0
umgezogener Einpers.-H.	119	21.0	21	14.3	180	14.4	43	9.3
abgespaltener Haushalt	221	11.7	82	25.6	295	8.4	112	6.3

1.2 Definition der Modellvariablen

Für die Schätzung der Wahrscheinlichkeit, daß ein Haushalt durch Kontaktverlust ausfällt, wurden die folgenden Merkmale benutzt:

Merkmal	Code	Ausprägungen
Umzug	1	Haushalt nicht umgezogen
	2	Als ganzer umgezogener Mehrpersonen-Haushalt
	3	Umgezogener Einpersonen-Haushalt
	4	Abgespaltener Haushalt
Großstadt	0	sonst
	1	Kernregion (nach Boustedt) mit mehr als 500 TSD Einwohnern
Haushaltsgröße	1	1 Personen-Haushalt
	2	2 Personen-Haushalt
	3	3 Personen-Haushalt
	4	4 und mehr Personen-Haushalt
Wohnumfeld	1	Einzelhaus oder ländliches Umfeld
	2	Hochhaus oder Haus mit mehr als 10 Mietparteien
	3	sonst
Abgespaltener Haushalt	1	Verzogener Mehrpersonen-Haushalt
	2	Verzogener Einpersonen-Haushalt
	3	Abgespaltener Haushalt

1.3 Geschätzte Wahrscheinlichkeiten für einen Ausfall durch Kontaktverlust

Die im vorhergehenden Abschnitt definierten Merkmale wurden im Rahmen einer multiplen Logit-Analyse zur Schätzung von $P_K = P(\text{Kontakt}=\text{nein})$ benutzt. Für die Hochrechnung wurde ein Modell gewählt, das nur noch signifikante Variablen enthält.

Interpretation der Koeffizienten:

$$\ln \frac{P_{K,i}}{1 - P_{K,i}} = \text{const} + X'_i \beta$$

Positive Koeffizienten zeigen eine Vergrößerung des Ausfallrisikos im Vergleich zum Durchschnitt der Merkmalsausprägungen an.

In den folgenden Tabellen bedeutet „ $\cdot$ “ das Hinzufügen eines weiteren Haupteffekts und „ $\cdot$ “ die Bildung eines Interaktionstermes. Merkmal 1 (Merkmal 2 = c) bedeutet die Bildung eines konditionalen Haupteffekts für das Merkmal 1, wenn das Merkmal 2 den Wert c annimmt.

Tabelle 3: Die Schätzung der Wahrscheinlichkeit des Ausfalls eines Haushalts durch Kontaktverlust im SOEP. Koeffizienten der logistischen Regression der Kontaktermittlung von SOEP-Haushalten in der jeweiligen Befragungswelle. Darstellung: Merkmal (Code: Koeffizient)

Stichprobe A (Deutsche-West)	
Welle	Modellvariablen und Koeffizienten
2	Konstante (-2.87), Großstadt (0: -0.24/1: 0.24), Umzug (1: -2.52/2: 0.11/3: 1.53/4: 0.84)
3	Konstante (-3.62), Großstadt (0: -0.36/1: 0.36), Umzug (1: -1.79/2: -0.49/3: 1.48/4: 0.80)
4	Konstante (-3.42), Umzug (1: -3.01/2: 0.78/3: 0.98/4: 1.35)
5	Konstante (-3.76), Umzug (1: -3.09/2,3: 1.34/4: 1.75 Einpersen-HH. (Umzug =1) (0: -1.35/1: 1.35) Einpersen-HH. (Umzug =2,3) (0: -0.28/1: 0.28) Einpersen-HH. (Umzug =4) (0: -0.63/1: 0.63)
6	Konstante (-3.48), Umzug (1: -2.33/2,3: 0.64/4: 1.69) Einpersen-HH. (Umzug =1) (0: -0.75/1: 0.75) Einpersen-HH. (Umzug =2,3) (0: -0.76/1: 0.76) Einpersen-HH. (Umzug =4) (0: -0.26/1: 0.26)
7*	Konstante (-2.97), Großstadt (0: -0.39/1: 0.39), abgespaltener HH. (1: -1.10/2: -0.07/3: 1.17)
8	Konstante (-5.03) Umzug (1: -2.79/2: -0.24/3: 0.50/4: 2.53)
9	P(Kontakt=nein)= 0 falls Umzug=1,2,3 / =0.06 falls Umzug=4
Stichprobe B (Ausländer)	
2	Konstante (-2.28), Großstadt (0: -0.50/1: 0.50), Umzug (1: -1.66/2: 0.69/3: -0.07/4: 1.04) Haushaltsgröße (1: 1.23/2: 0.26/3: -0.82/4: -0.67)
3	Konstante (-2.65), Großstadt (0: -0.72/1: 0.72), Umzug (1: -3.06/2: 0.16/3: 1.64/4: 1.26)
4	Konstante (-3.34), Umzug (1: -3.60/2: -0.46/3: 2.19/4: 1.87)
5	wie Stichprobe A
6	wie Stichprobe A
7*	Konstante (-2.93), Großstadt (0: 0.064/1: -0.64), Abgespaltener HH. (1: -1.65/2: 0.58/3: 1.07), Wohnumfeld (1: -0.73/2: 1.32/3: -0.59)
8	wie Stichprobe A
9	P(Kontakt=nein)= 0 falls Umzug=1,2,3 / =0.10 falls Umzug=4
* In Welle 7 wurden alle nicht umgezogenen Haushalte wieder kontaktiert, so daß die Ausfallanalyse nur auf den umgezogenen Haushalten basiert.	

Fortsetzung von Tabelle 3

Stichprobe C (Deutsche-Ost)	
Welle	Modellvariablen und Koeffizienten
2 Ost	$P(\text{Kontakt}=\text{nein}) = \text{Umzug}(1: 0.0 / 2: 0.11 / 3: 0.14 / 4: 0.25)$
3 Ost	$P(\text{Kontakt}=\text{nein}) = \text{Umzug}(1,2: 0.0 / 3: 0.09 / 4: 0.07)$

2 Ausfälle durch Antwortverweigerung

Für die Haushalte mit Kontaktaufnahme werden zwei mögliche Ergebnisse betrachtet:

- es wird ein Haushalt-Interview realisiert
- es wird kein Haushalts-Interview realisiert.

Hierbei wird nicht nach der Art der Ausfallgründe differenziert.

2.1 Ausfallraten nach verschiedenen Haushaltsmerkmalen

Die folgenden Tabellen dokumentieren die Ausfallraten bei wieder kontaktierten Haushalten nach unterschiedlichen Haushaltsmerkmalen. Die Merkmale beziehen sie sich in der Regel auf die vorhergehende Welle. Lediglich die feldorientierten Variablen orientieren sich an der aktuellen Befragungswelle.

Die personenbezogen Merkmale beziehen sich auf den Haushaltsvorstand. Bei abgespaltenen Haushalten beziehen sich die personenbezogenen Merkmale auf die aus dem Panelhaushalt ausgezogene Person (bei mehreren Personen auf die im Adressprotokoll zuerst genannte Person).

Tabelle 4:: Teilnahmeverhalten von erreichten Haushalten nach sozio-demographischen Merkmalen des Haushaltsvorstands. N = Anzahl der zu befragenden Haushalte, % = Prozentsatz der Haushalte ohne Interview

		Welle							
		2	3	4	5	6	7	8	9
Insgesamt	N	5937	5732	5398	5285	5095	4982	4985	4977
	%	10.4	11.2	6.9	8.9	7.9	6.9	6.3	6.7
Ausfall im Vorjahr	N	-	259	197	154	169	154	183	145
	%	-	59.5	52.8	71.6	57.4	49.3	48.1	62.1
Haushalte ohne Antwortausfall im Vorjahr									
insgesamt	N	5937	5473	5201	5131	4926	4828	4802	4832
	%	10.4	8.9	5.1	7.0	6.2	5.5	4.7	5.0
Stichprobe									
A (Deutsche)	N	4611	4275	4058	3993	3834	3755	3716	3724
	%	10.2	8.7	5.2	7.1	6.2	5.3	4.6	4.9
B (Ausländer)	N	1326	1198	1143	1138	1092	1073	1086	1108
	%	10.9	9.6	5.0	6.9	6.4	6.1	5.3	5.5
Geschlecht									
männlich	N	4664	4226	3951	3840	3624	3486	3413	3372
	%	9.8	8.3	4.7	6.7	6.2	5.0	4.2	4.9
weiblich	N	1273	1247	1250	1291	1303	1342	1389	1460
	%	12.2	11.0	6.5	7.9	6.5	6.9	6.1	5.3
Alter									
75 u. älter	N	448	394	374	386	381	380	371	367
	%	18.3	13.7	6.7	6.5	6.8	5.0	3.0	5.2
65-74	N	562	513	487	480	465	462	477	503
	%	10.1	9.4	2.9	5.4	6.0	3.0	2.9	3.2
55-64	N	947	860	809	803	798	783	782	811
	%	9.8	8.7	4.2	4.7	5.9	5.5	3.8	3.4
35-54	N	2621	2401	2272	2226	2112	2017	1970	1899
	%	9.2	7.8	4.3	6.3	6.1	4.9	4.7	4.1
25-34	N	1116	1034	976	963	904	926	957	983
	%	8.9	8.0	6.1	9.9	6.1	7.7	6.1	6.2
16-24	N	243	271	283	273	266	260	245	269
	%	17.3	15.1	13.4	13.2	9.0	8.1	8.9	14.8

Fortsetzung von Tabelle 4

		Familienstand							
verheiratet, zusammen	N	3893	3600	3366	3301	3144	3029	3015	3008
	%	9.6	8.0	4.4	6.2	6.2	4.6	3.4	4.9
verheiratet, getrennt	N	104	157	119	97	120	110	96	102
	%	7.7	12.1	5.9	13.4	6.7	8.2	12.5	6.9
ledig	N	836	811	802	783	764	764	782	797
	%	12.6	9.6	9.4	11.5	6.8	8.6	9.1	9.7
geschieden	N	349	345	328	347	327	351	356	369
	%	10.3	10.4	4.6	6.3	7.0	7.1	4.8	4.9
verwitwet	N	671	560	533	542	534	532	518	515
	%	12.6	11.9	3.9	5.1	4.7	4.5	4.4	3.9
		Schulbildung							
Kein Abschluß	N	493	445	411	407	376	373	380	379
	%	10.9	11.0	6.3	8.8	6.7	5.6	5.8	5.5
Hauptschule	N	2952	2669	2493	2488	2405	2340	2314	2296
	%	11.8	9.8	4.7	6.2	6.3	5.2	4.6	4.6
Realschule	N	852	849	818	805	798	780	784	812
	%	6.8	8.8	5.6	8.3	8.0	6.3	4.9	4.7
Fachhochschule	N	223	205	205	201	183	180	184	188
	%	9.4	6.8	6.3	9.5	5.5	5.6	7.6	4.8
Abitur	N	601	588	582	569	542	552	533	557
	%	7.5	6.0	5.5	7.6	4.2	7.4	5.1	7.0
		Stellung im Beruf							
nicht erwerbs- tätig	N	1527	1325	1290	1302	1276	1329	1339	1279
	%	13.9	10.1	5.5	6.9	6.4	5.2	4.3	4.9
arbeitslos	N	206	297	260	258	265	193	199	215
	%	9.7	10.7	7.7	8.5	4.9	2.6	6.0	6.5
hoch	N	585	578	522	530	519	511	496	518
	%	7.9	6.2	4.4	7.7	7.1	8.4	5.9	5.6
mittel	N	2248	2202	2053	1982	1911	1803	1857	1932
	%	8.8	8.1	5.3	6.6	6.2	5.1	4.7	5.1
niedrig	N	1364	1071	1076	1059	954	992	911	888
	%	11.2	10.0	4.3	7.4	6.2	5.9	4.6	4.3

Tabelle 5 : **Teilnahmeverhalten von erreichten Haushalten nach feldbedingten Merkmalen.** N = Anzahl der zu befragenden Haushalte, % = Prozentsatz der Haushalte ohne Interview

		Welle							
		2	3	4	5	6	7	8	9
		Typ des Haushalts							
nicht umgezogener Haushalt	N	5372	4810	4646	4567	4421	4353	4313	4379
	%	9.6	8.0	4.2	5.8	5.6	4.6	3.6	3.9
umgezogener Haushalt	N	370	425	373	370	322	277	274	275
	%	11.6	12.7	10.7	14.1	9.0	7.6	9.9	9.5
abgespaltener Haushalt	N	195	238	182	194	183	198	215	208
	%	29.2	21.4	17.6	23.7	16.9	23.2	21.4	22.1
		Wechsel des Interviewers							
ja	N	2041	1203	816	715	826	742	717	751
	%	14.9	17.5	12.5	19.0	12.9	14.4	10.7	12.1
nein	N	3896	4265	4385	4416	4100	4086	4085	4081
	%	7.9	6.5	3.8	5.1	4.9	3.9	3.7	3.7
		Neue Befragungsperson im Haushalt							
ja	N	257	243	218	211	209	220	198	210
	%	9.0	11.1	6.4	6.6	4.8	6.4	6.6	2.4
nein	N	5680	5230	4983	4920	4717	4608	4604	4622
	%	10.4	8.8	5.1	7.1	6.3	5.5	4.7	5.1
		Auszug einer Befragungsperson aus Haushalt							
ja	N								209
	%								7.2

Fortsetzung von Tabelle 5

		Haushalt mit Paartrennung							
insgesamt	N								94
	%								24.5
alter Haushalt	N								47
	%								14.6
neuer Haushalt	N								47
	%								34.8
		Anzahl der bereits gewährten Interviews							
vollständig	N	-	5419	5018	4826	4600	4384	4225	4060
seit Beginn	%	-	8.7	4.7	5.9	5.6	4.4	3.6	3.3
-1	N	-	-	161	246	253	294	346	389
	%	-	-	16.7	23.1	14.6	13.9	13.8	10.8
-2	N	-	-	-	46	43	73	93	127
	%	-	-	-	43.5	16.2	23.2	12.9	18.9
-3	N	-	-	-	-	24	49	63	104
	%	-	-	-	-	12.5	14.3	9.5	10.6
		Angaben zum HH-Einkommen fehlen im Vorjahr							
	N	335	310	272	237	210	203	197	226
	%	17.9	18.4	12.9	16.9	11.4	15.2	12.7	10.6

Tabelle 6: Teilnahmeverhalten von erreichten Haushalten nach Haushaltseinkommen und Anzahl der Vermögensarten im Haushalt. N = Anzahl der zu befragenden Personen, % = Prozentsatz der Haushalte ohne Interview

		Welle							
		2	3	4	5	6	7	8	9
		Haushaltseinkommen							
-1000	N	456	368	293	270	241	213	182	165
	%	13.8	10.1	5.5	9.3	7.1	6.1	6.6	8.5
1000-2000	N	1816	1521	1383	1243	1140	995	870	802
	%	11.2	9.7	5.3	6.6	6.2	4.8	5.1	4.5
2000-3000	N	1713	1572	1469	1404	1354	1329	1260	1202
	%	8.3	7.6	4.2	6.5	6.6	4.6	4.0	5.1
3000-4000	N	992	996	1008	1087	1060	1073	1069	1085
	%	9.6	6.4	4.3	6.1	5.0	4.7	3.7	4.4
4000+	N	625	706	776	890	921	1015	1224	1352
	%	8.2	8.9	5.0	6.4	5.9	6.2	4.7	4.4
		Anzahl der Vermögensarten im Haushalt							
0	N	823	769	743	735	578	661	573	604
	%	12.5	12.3	6.9	11.3	8.3	8.8	6.6	9.9
1	N	1714	1561	1468	1429	1431	1262	1256	1191
	%	13.2	10.4	5.1	6.8	6.7	5.5	4.4	4.5
2	N	1709	1549	1427	1449	1444	1310	1350	1367
	%	8.3	8.2	5.6	5.8	5.9	5.2	5.0	4.1
3	N	1224	1161	1134	1122	1107	1152	1201	1180
	%	8.7	6.7	3.9	5.9	5.7	4.1	3.7	4.2
4	N	403	388	374	343	326	377	375	447
	%	7.7	5.9	3.7	7.0	4.6	5.3	4.5	4.5
5	N	64	45	55	53	40	66	47	43
	%	9.4	6.7	9.1	11.3	2.5	7.5	14.9	4.7
		Bezug vom Sozialhilfe							
ja	N								133
	%								6.8

Tabelle 7: Vergleich der Ausfallraten zwischen SOEP-West und SOEP-Ost in der 2. und 3. Befragungswelle

Merkmal	Welle 2				Welle 3			
	West		Ost		West		Ost	
	N	%	N	%	N	%	N	%
Erreichte HH insgesamt:	5937	10.4	2213	8.3	5732	11.2	2290	11.8
davon Ausfall im Vorjahr	-	-	-	-	259	59.4	162	53.1
Kein Ausfall im Vorjahr	-	-	-	-	5473	8.9	2128	8.6
davon nach:								
Alter des HV:								
75+	503	17.5	127	18.1	385	13.7	106	11.3
65-74	576	9.9	212	8.0	513	9.5	204	7.3
55-64	975	9.7	330	7.0	862	9.0	328	7.0
35-54	2627	9.3	920	6.3	2409	7.7	897	6.8
25-34	1075	8.9	529	9.2	1035	7.9	469	9.4
-25	181	19.9	95	13.4	269	14.8	124	23.4
Geschlecht des HV:								
männlich	4639	9.8	1093	7.6	4218	8.2	1052	8.9
weiblich	1298	12.4	1120	8.9	1255	11.4	1076	8.3
Status des HV:								
Nicht erwerbstätig	1530	12.9	462	10.8	1333	10.1	484	8.2
Arbeitslos	204	9.8	14	14.3	294	10.8	202	9.4
Höchste Statusgruppe	580	7.9	383	5.5	581	6.4	276	5.8
Niedrigste Statusgruppe	1239	10.6	288	9.4	1074	9.9	335	9.5
Sonstige	2377	9.1	1066	7.8	2191	8.1	831	9.2
Höchster Schulabschluß:								
West Ost								
Abitur Abitur	823	8.0	324	7.1	793	5.9	316	7.6
mittl. Reife 10. Kl.	850	6.8	978	7.0	845	8.9	982	8.6
Hauptsch. 8. Kl.	2955	11.7	853	9.9	2673	9.7	783	8.8
Keinen Abschluß	491	11.2	-	-	436	11.2	-	-
Sonstige	818	10.8	-	-	726	7.6	-	-
Netto HH-Einkommen:								
Keine Angabe	335	17.9	32	9.4	310	18.4	40	17.5
West Ost								
-1000 -800	456	13.9	280	14.3	368	10.0	194	8.7
1000-2000 800-1200	1816	11.2	322	8.7	1521	9.6	215	8.8
2000-3000 1200-1800	1713	8.3	651	8.1	1572	7.6	511	9.8
3000-4000 1800-2500	992	9.5	669	6.2	996	6.4	623	7.5
4000+ 2500+	625	8.2	259	6.5	706	8.9	545	8.1

Fortsetzung von Tabelle 7

Merkmal	Welle 2				Welle 3			
	West		Ost		West		Ost	
	N	%	N	%	N	%	N	%
Typ des Haushalts								
Alter Haushalt, nicht umgezogen	5372	9.6	2062	7.7	4810	7.9	1915	7.3
Alter Haushalt, umgezogen	370	11.6	90	8.9	425	12.7	108	16.6
neuer Haushalt	195	29.2	61	24.6	238	21.4	105	23.8
Wechsel des Interviewers								
ja	2081	15.3	426	8.4	1208	17.5	413	11.6
nein	3856	7.6	1184	5.5	4265	6.5	1682	7.4
Interviewer Nr. fehlt	-	-	603	13.6	-	-	33	36.4
Derselbe Interviewer 1-3	-	-	-	-	2826	5.5	958	6.4
Belastung durch Pflegefall, Todesfall	191	13.6	-	-	176	9.6	109	11.0
Arbeitsplatzverlust erwartet	-	-	733	9.1	-	-	604	8.9
Haushalt lebt in alter BRD	-	-	26	15.4	-	-	63	15.8

2.2 Definition der Modellvariablen

Die in dem vorhergehenden Abschnitt aufgeführten Merkmale wurden bis auf wenige Ausnahmen in eine simultane Analyse der Ausfälle durch Antwortverweigerung einbezogen. Für die Schätzung der entsprechenden Ausfallwahrscheinlichkeit wurde jedoch nur ein Modell benutzt, in dem alle Variablen signifikant sind. Hieraus kann geschlossen werden, daß die weggelassenen Variablen bzw. Differenzierungen zwischen Merkmalsausprägungen keinen systematischen Einfluß auf die Antwortgewährung erreichter Haushalte hat. Die Definition der Modell-Variablen findet man in der folgenden Liste:

Merkmal	Code	Ausprägungen
Alter des Haushaltsvorstands	1	älter als 75 Jahre
	2	65-74 Jahre
	3	55-64 Jahre
	4	35-54 Jahre
	5	25-34 Jahre
	6	jünger als 25 Jahre
Geschlecht des Haushaltsvorstands	0	männlich
	1	weiblich
HH Typ	1	alter Haushalt, nicht umgezogen
	2	alter Haushalt, umgezogen
	3	abgespaltener Haushalt
Wechsel des Inter- viewers	0	Kein Wechsel
	1	Wechsel gegenüber letzter Welle
	2	Interviewernummer fehlt
Betreuung durch Interviewer		Anzahl der Wellen mit Betreuung durch den aktuellen Interviewer
Beginn	0	sonst
	1	HV seit Welle 1 im Panel
Auszug	0	sonst
	1	Auszug einer Befragungsperson seit letzter Welle
Paar	0	sonst
	1	Auszug des HV od. Ehepartner(Lbenspartner) des HV
Typ	0	HH Typ =1,2 und Paar=1
	1	HH Typ =1 und Paar=0
	2	HH Typ =2 und Paar=0
	3	HH Typ =3 und Paar=0
	4	HH Typ =3 und Paar=1
Ost-Berlin	0	sonst
	1	Haushalt lebt in Ostberlin

Fortsetzung der Merkmalsliste

Merkmal	Code	Ausprägungen
Arbeitslos	0	sonst
	1	HV ist arbeitslos gemeldet
Verlust Arbeitsplatz (subjektive Einschätzung)	0	sonst
	1	sicher od. wahrscheinlich
Status des Haushaltsvorstands	0	sonst
	1	hoher Erwerbsstatus des HV
Sozialhilfe	0	sonst
	1	Haushalt bezieht Sozialhilfe
Haushalts- Einkommen West	1	K.A. beim Einkommen
	2	bis 2000 DM
	3	2000-4000 DM
	4	über 4000 DM
Haushalts- Einkommen Ost	1	K.A. beim Einkommen
	2	bis 800 DM
	3	800-1200 DM
	4	1200-1800 DM
	5	1800-2500 DM
	6	über 2500 DM
K.A.-Einkommen	0	sonst
	1	K.A. beim Einkommen
K.A.-Vermögensbilanz	0	sonst
	1	Vermögensbilanz in Welle 5 nicht beantwortet
Anzahl Anlageformen	1	Anzahl der Vermögens- anlageformen = 0
	2	Anzahl = 5
	3	sonst
Keine Anlageformen	0	sonst
	1	Anzahl der Vermögens- anlageformen = 0
Betriebsvermögen	0	sonst
	1	Besitz von Betriebsvermögen
Sparbuch als Wertanlage angegeben	0	nein
	1	ja

2.3 Geschätzte Wahrscheinlichkeiten für einen Ausfall durch Antwortverweigerung erreichter Haushalte

Die im vorhergehenden Abschnitt definierten Merkmale wurden im Rahmen einer multiplen Logit-Analyse zur Schätzung von $P_R = P(\text{Response}=\text{nein})$ benutzt.

Interpretation der Koeffizienten:

$$\ln \frac{P_{R,i}}{1 - P_{R,i}} = \text{const} + X'_i \beta$$

Positive Koeffizienten zeigen eine Vergrößerung des Ausfallrisikos im Vergleich zum Durchschnitt der Merkmalsausprägungen an.

In den folgenden Tabellen bedeutet „ $\cdot$ “ das Hinzufügen eines weiteren Haupteffekts und „ $\cdot$ “ die Bildung eines Interaktionstermes. Merkmal 1 (Merkmal 2 = c) bedeutet die Bildung eines konditionalen Haupteffekts für das Merkmal 1, wenn das Merkmal 2 den Wert c annimmt.

Tabelle 8: Die Schätzung der Wahrscheinlichkeit eines Ausfalls durch Antwortverweigerung eines wiedererreichten Haushalts im SOEP. Koeffizienten der logistischen Regression der Antwortgewährung von wieder erreichten Haushalten in der jeweiligen Befragungswelle. Darstellung: Merkmal (Code: Koeffizient)

Stichprobe A (Deutsche-West)	
Welle	Modellvariablen und Koeffizienten
2	Konstante (-1.53), Wechsel Interviewer (0: -0.25/1: 0.25), Alter HV (1: 0.66/2,3,4: 0.03/5: -0.39/6: -0.30), HH. Typ (1: -0.68/2: -0.19/3: 0.87), HH. Einkommen West (1: 0.61/2: 0.12/3: -0.35/4: -0.38)
3	Konstante (-1.22), Wechsel Interviewer (0: -0.39/1: 0.39), Alter* (Wechsel Interviewer =0) (1: -0.13/2: -0.11/3,4: -0.39/ 5: 0.26/6: 0.37), Alter* (Wechsel Interviewer =1) (1: 0.13/2: 0.11/3,4: 0.39/5: -0.26/6: -0.37), Alter (1: 0.59/2: 0.16/3,4: -0.06/5: -0.53/6: -0.16) HH. Typ (1: -0.52/2: 0.10/3: 0.42), Arbeitslos (0: -0.21/1: 0.21), K.A. Einkommen (0: -0.39/1: 0.39)
4	Konstante (-1.83), Wechsel Interviewer (Alter =1) (0: -0.44/1: 0.44), Wechsel Interviewer (Alter =2) (0: -0.74/1: 0.74), Wechsel Interviewer (Alter =3,4) (0: -0.59/1: 0.59), Wechsel Interviewer (Alter =5) (0: -0.41/1: 0.41), Wechsel Interviewer (Alter =6) (0: -0.32/1: 0.32), Alter (1: 0.21/2: -0.38/3,4: -0.24/5: 0.06/6: 0.35), HH. Typ (1: -0.45/2: 0.29/3: 0.19), K.A. Einkommen (0: -0.39/1: 0.39)

Fortsetzung von Tabelle 8

Welle	Modellvariablen und Koeffizienten
5	Konstante (-1.60), Betreuung durch Interviewer (1: 1.15/ 2: 0.41/3: 0.18/4: -0.71/5: -1.03), Alter (Wechsel Interviewer =1) (1,2: 0.52/3,4,5: -0.11/6: -0.40), HH. Typ (1: -0.49/2: 0.11/3: 0.38) K.A. Einkommen (0: -0.45/1: 0.45), Keine Anlageformen (0: -0.38/1: 0.38)
6	Konstante (-2.44), Betreuung durch Interviewer (1: 0.75/ 2: 0.58/3: 0.21/4: -0.59/5: -0.43/6: -0.52), Alter (Wechsel Interviewer =1) (1,2: 0.26/3,4,5: 0.05/6: -0.31), HH. Typ (1: -0.32/2: -0.04/3: 0.37), K.A. Einkommen (0: -0.26/1: 0.26), K.A. Vermögensbilanz (0: -0.29/1: 0.29), Betriebsvermögen (0: 0.41/1: -0.41)
7	Konstante (-1.34), Wechsel (HH. Typ =1) (0: -0.75/1: 0.75), Wechsel (HH. Typ =3) (0: -0.12/1: 0.12), HH. Typ (1: -0.66/ 2: -0.24/3: 0.90), K.A. Einkommen (0: -0.58/1: 0.58), Status HV (0: -0.30/1: 0.30)
8	Konstante (-1.15); Wechsel (0: -0.55/1: 0.55), HH. Typ (1: -0.83/2: -0.14/3: 0.97), K.A. Einkommen (0: -0.57/1: 0.57), Anzahl Anlageformen (1: -0.08/2: 0.70/3: -0.62)
9	Konstante (-1.31), Wechsel Interviewer(Beginn =0) (0: -0.17/1: 0.17), Wechsel Interviewer(Beginn =1) (0: -0.68/1: 0.68), Beginn(Alter=1) (0: -0.09/1: 0.09), Beginn(Alter=2) (0: 0.70/1: -0.70), Beginn(Alter=3) (0: 1.20/1: -1.20), Beginn(Alter=4) (0: 0.49/1: -0.49), Beginn(Alter=5) (0: 0.48/1: -0.48), Beginn(Alter=6) (0: 0.10/1: -0.10), HH. Typ (1: -0.53/2: 0.07/3: 0.46), Auszug(HH. Typ =1) (0: -0.47/1: 0.47) K.A. Einkommen (0: -0.25/1: 0.25) Keine Anlageformen (0: -0.29/1: 0.29) Geschlecht HV (0: 0.15/1: -0.15)

Fortsetzung von Tabelle 8

Stichprobe B (Ausländer)	
Welle	Modellvariablen und Koeffizienten
2	Konstante (-1.96), Wechsel Interviewer (0: -0.55/1: 0.55), HH. Typ (1: -0.03/2: -0.58/3: 0.62)
3	Konstante (-1.60), Geschlecht HV (0: -0.31/1: 0.31), HH. Typ (1,2: -0.46/3: 0.46)
4	Konstante (-1.69), Wechsel Interviewer (Alter =1,2,3) (0: -0.47/1: 0.47), Wechsel Interviewer (Alter =4) (0: -0.73/1: 0.73), Wechsel Interviewer (Alter =5) (0: -0.60/1: 0.60), Wechsel Interviewer (Alter =6) (0: -0.26/1: 0.26), HH. Typ (1: -0.34/2: 0.46/3: -0.12), HH. Einkommen (1: 0.75/2: 0.10/3: -0.85)
5	Konstante (-1.87), Betreuung durch Interviewer (1: 1.26/ 2: 0.14/3: -0.21/4: -0.70/5: -0.50, HH. Typ (1: -0.47/2: 0.89/3: -0.42), K.A. Einkommen (0: -0.43/1: 0.43)
6	Konstante (1.89), Betreuung durch Interviewer (1: 0.83/ 2: 0.37/3: -0.31/4: -0.55/5: 0.04/ 6: -0.37), HH. Typ (1: -0.41/2: 0.22/3: 0.19), K.A. Einkommen (0: -0.54/1: 0.54)
7	Konstante (-1.50), Wechsel (HH. Typ =1) (0: -0.55/1: 0.55), Wechsel (HH. Typ =2) (0: -0.98/1: 0.98), Wechsel (HH. Typ =3) (0: -1.06/1: 1.06), HH. Typ (0: -0.50/2: -0.88/3: 1.38), K.A. Einkommen (0: -0.66/1: 0.66)
8	Konstante (-2.05), Wechsel Interviewer (0: -0.48/1: 0.48), HH. Typ (1: -0.85/2: 0.22/3: 0.63)
9	Konstante (-1.79), Wechsel Interviewer (0: -0.50/1: 0.50), Beginn (0: 0.39/1: -0.39), Typ (0: 0.16/1: -0.59/2: -1.90/3: -0.03/4: 2.36), Alter (1,2,3: 0.28/4: -0.10/5: -0.65/6: 0.47), K.A. Einkommen (0: -0.66/1: 0.66), Keine Vermögensanlagen (0: -0.53/1: 0.53), Sozh. (0: 0.73/1: -0.73)

Fortsetzung von Tabelle 8

Stichprobe C (Deutsche-Ost)	
Welle	Modellvariablen und Koeffizienten
2 Ost	Konstante (-0.91), Wechsel Interviewer (0: -0.47/1: -0.04/2: 0.51), Alter HV (1: 0.41/2,3,4,5,6: -0.41), HH Typ (1,2: -0.84/3: 0.84), HH Einkommen Ost (1: 0.24/2: 0.44/3: 0.12/4: 0.00/5: -0.37/6: -0.44), Verlust Arbeitsplatz (0: -0.17/1: 0.17), Ost-Berlin (0: -0.29/1: 0.29)
3 Ost	Konstante (-1.36), HH Typ (1: -0.39/2: 0.08/3: 0.31), Wechsel Interviewer (HH Typ =1) (0: -0.28/1,2: 0.28), Wechsel Interviewer (HH Typ =2) (0: 0.42/1,2: -0.42), Wechsel Interviewer (HH Typ =3) (0: -0.36/1,2: 0.36), Alter HV (1: 0.02/2,3,4: -0.38/5: -0.20/6: 0.56), Sparbuch als Wertanlage angegeben (0: 0.35/1: -0.35)