Informationen zur Nutzung des SAS-Setups für das Mikrozensus Scientific Use File 1993

Der vorliegende Text soll Ihnen dabei helfen, den Rohdatensatz des Statistischen Bundesamtes, den Sie im ASCII-Format erhalten haben, korrekt in SAS einzulesen.
Dazu wird vom German Microdata Lab (GML) des Zentrums für Umfragen, Methoden und Analysen (ZUMA) ein Setup (setup93.sas) bereitgestellt.

Ausführliche Informationen über den Mikrozensus und die Mikrozensus Scientific Use Files finden Sie auf den Websites des Zentrums für Umfragen, Methoden und Analysen (ZUMA) unter:
http://www.gesis.org/Dauerbeobachtung/GML/index.htm

Weitere Informationen sind auf den Seiten des Statistischen Bundesamtes abrufbar:
http://www.destatis.de/themen/d/thm_mikrozen.php

Das vom GML bereitgestellte SAS-Setup für das Mikrozensus Scientific Use File 1993 *(setup93.sas)* dient zum Einlesen des Rohdatenmaterials und zum Erstellen eines SAS-Systemfiles. Es beinhaltet Programmanweisungen zum Ersetzen fehlender Werte, zur Umwandlung alphanumerischer Variablen in numerische sowie zum Versehen der Variablen und ihrer Ausprägungen mit entsprechenden Labels.

Das Setup gliedert sich in verschiedene Bereiche, wobei Folgendes zu beachten ist:
Zunächst werden Formate definiert, um diese später den Ausprägungen der Variablen zuzuweisen. Sollen diese Formate dauerhaft verwendet werden, müssen sie in einer separaten Datei gespeichert werden. Das entsprechende Verzeichnis ist dann mit dem speziell dafür vorgesehenen Bibliotheksnamen LIBRARY zu referenzieren. Um in späteren SAS-Sitzungen auf die Formate zugreifen zu können, muss dem Verzeichnis, in dem die Formate abgespeichert sind, wiederum der Name LIBRARY zugewiesen werden.
Beim Einlesen des Rohdatenfiles ist in der Option LRECL der INFILE-Anweisung angegeben, wie viele Stellen eine Zeile (d.h. eine Beobachtung) im Rohdatenfile umfasst. Alphanumerische Variablen sind beim Einlesen mit einem $-Zeichen versehen und werden in einem nachfolgenden Schritt in numerische Variablen umgewandelt. Fehlenden Werte, d.h. Leerstellen im Rohdatenfile, werden über IF-Anweisungen durch gültige Werte ersetzt und später über die Format-Anweisung mit Labels versehen. Weil SAS keine benutzerdefinierten Missings kennt, müssen die Werte für spätere Auswertungen gegebenenfalls manuell auf Missing zurückgesetzt werden. Optional können die IF-Anweisungen im Setup auch auskommentiert werden. Dann werden die fehlenden Werte allerdings auch nicht gelabelt.

Das Rohdatenfile des Statistischen Bundesamtes enthält alle Variablen des Mikrozensus 1993. Damit das SAS-Setup für jede mögliche Variablen-Auswahl verwendbar ist, wurden

alle Variablen erfasst. Variablen, die Sie nicht bestellt haben, sind in Ihrem Rohdatenfile auf 0 oder Leerzeichen gesetzt. Um ein Systemfile zu erzeugen, das nur die von Ihnen bestellten Variablen enthält, fügen Sie im Setup am Ende des DATA-step die Anweisung KEEP ein und nennen dort Ihre bestellten Variablen.

Um eine reibungsfreie Aufbereitung der Daten zu gewährleisten, empfiehlt es sich, die in dem Setup vorgesehenen Voreinstellungen zu berücksichtigen. Insbesondere die Anweisung COMPRESS=YES im DATA-step bewirkt, dass die erzeugte Datei deutlich reduzierten Speicherplatz benötigt.

Sind die Daten eingelesen und ist das entsprechende Systemfile erstellt, kann die folgende Fallzahl (ohne Gewichtung, ohne Selektion) zur Kontrolle, ob der Rohdatensatz fehlerfrei eingelesen wurde, mit der des eingelesenen Datensatzes verglichen werden. Unterscheiden sich die Fallzahlen, weist dies auf einen Fehler beim Einlesen hin.

Fallzahl Mikrozensus 1993 (ohne Gewichtung, ohne Selektion) = 513.830

Des Weiteren können die Verteilungen ausgewählter Variablen des Mikrozensus SUF 1993 zur Prüfung des fehlerfreien Ablaufs des Setups herangezogen werden (vgl.: http://www.gesis.org/Dauerbeobachtung/GML/Daten/MZ/mz_1993/randverteilungen_mz93.htm).
Als Referenz zur Plausibilitätsprüfung der Mikrozensus Scientific Use Files dienen die in den Fachserien des Statistischen Bundesamtes veröffentlichten Ergebnisse des jeweiligen Mikrozensus. Für den Mikrozensus 1993 sind dies die Reihe 3 (Haushalte und Familien), die Reihe 4.1.1 (Stand und Entwicklung der Erwerbstätigkeit) und die Reihe 4.1.2 (Beruf, Ausbildung und Arbeitsbedingungen der Erwerbstätigen) der Fachserie 1.

Die Plausibilitätsprüfung des Mikrozensus SUF 1993 erfolgte durch den Vergleich des aufbereiteten Datensatzes mit den vom Statistischen Bundesamt veröffentlichten Tabellen in den entsprechenden Reihen der Fachserie 1. Dabei wurde nach den vom Statistischen Bundesamt verwendeten Bevölkerungs- und Erwerbskonzepten gewichtet und selektiert. (zur Abgrenzung und Hochrechnung der Bevölkerungsbegriffe im Mikrozensus 1993 vgl.: http://www.gesis.org/Dauerbeobachtung/GML/Daten/MZ/mz_1993/hochrechnungsfaktoren_mz93.htm).

Bedingt durch die Ziehung der 70%-Substichprobe ergeben sich Abweichungen zwischen den Häufigkeiten der Variablen des Mikrozensus Scientific Use Files und den in den Fachserien des Statistischen Bundesamtes veröffentlichten Zahlen (beruhend auf den Original-Mikrozensusdaten). Die meisten Variablen weichen nur in geringem Maße (meist 0% bis 1%, max. 5%) von den veröffentlichten Daten ab. Abweichungen über 5% wurden bei Variablen festgestellt, deren Merkmalsausprägungen mit sehr geringen Fallzahlen besetzt sind (z.B. Staatsangehörigkeit und Wirtschaftsbereiche).

Des Weiteren ist zu beachten, dass im Setup des Mikrozensus Scientific Use Files 1993 Recodierungen vorgenommen werden, die zu Abweichungen in der numerischen Bezeichnung der Value Labels im Vergleich zum Schlüsselverzeichnis führen. Im Folgenden werden die wichtigsten Regeln, die hinter diesen Recodierungen im Setup stehen, benannt:

- Die meisten Recodierungen im Setup wurden vorgenommen, um eine einheitliche Kennzeichnung der Missing-Kategorien zu erreichen. Diese sind im Falle der Ausprägung 'Entfällt' mit 9; 99 oder 999 und im Falle der Ausprägung 'Angabe fehlt/keine Angabe' mit 8; 98 und 998 gekennzeichnet. Das hat zur Folge, dass der Ausprägung 'Nein' – im Schlüsselverzeichnis in der Regel mit der Zahl 9 versehen – eine Zahl zugewiesen wurde, die sich an die anderen Ausprägungen anschließt. Die 0 für 'Angabe fehlt/keine Angabe' wurde durch die 8, 98 oder 998 ersetzt.
- Weitere Recodierungen gibt es, wenn sich innerhalb der Zahlenfolgen Lücken ergaben (z.B. 1,2,4,5,6 und 9 wurde zu 1,2,3,4,5 und 9).
- Zum Teil wurden Recodierungen vorgenommen um die Merkmalsausprägungen in eine "logische Folge" zu bringen. Das betrifft die Variablen EF108 - EF110 und EF220 und EF221. So wurde zum Beispiel bei der Variable "Stellung im Beruf" (EF110) die Ausprägung 'Direktor, Amtsleiter, Geschäftsführer oder Betriebsleiter/Werksleiter' von 0 (Schlüsselverzeichnis) auf 10 (Setup) recodiert, so dass sie sich an die Ausprägung 'Abteilungsleiter, Prokurist' (9) anschließt und nicht der Ausprägung 'Auszubildender, Praktikant, Volontär' (1) vorangestellt ist.
- In einigen Variablen wurde die Gruppe der Personen in Gemeinschaftsunterkünften, die als Missing geführt wird, von 0 auf einen anderen numerischen Wert gesetzt, damit sie sich am Ende der Häufigkeitsauszählung wiederfindet. Dies ist bei den Variablen, die die Haushalte und die Familien betreffen, der Fall.
- Zahlenfolgen in der Form von 0,1,2,3... wurden in die Form von 1,2,3,4... umgewandelt. Dies betrifft die Variable "Schulbesuch" (EF56) und Variablen, die auf der Frage nach der "Stellung im Beruf" basieren (EF115, EF227) sowie die Variablen zum "sonstigen öffentlichen und privaten Einkommen" (EF144 und EF145).

Die Regeln gelten nicht für alle im Setup aufgeführten Variablen. Als Grundlage für das vorliegende Setup diente ein Setup, das zunächst für das ZUMA-File (d.h. für eine bestimmte Merkmalsauswahl aus dem Mikrozensus Grundfile) erstellt wurde. Zu einem späteren Zeitpunkt wurde das Setup um die restlichen Variablen des Mikrozensus Grundfiles ergänzt. Dieses enthält nun alle im Mikrozensus Grundfile verfügbaren Variablen und wird von ZUMA für alle Nutzer bereitgestellt. Bei den ergänzten Variablen wurde möglichst auf Recodierungen verzichtet. Fehlende Werte wurden bei diesen Variablen wie folgt codiert: Sofern die 0 nicht besetzt war, wurden Missings mit 0 codiert. War die 0 vergeben, so wurde auf die 9; 99 oder 999 zurückgegriffen. Falls diese Zahlen ebenfalls besetzt waren, wurden die Missings mit 8 oder 10 gekennzeichnet. Betroffen sind die Variablen zur Kranken- und Rentenversicherung (EF76, EF78, EF82, EF83), zur Schichtarbeit und den geleisteten Arbeitsstunden in der Berichtswoche (EF101, EF111, EF112, EF128).

Das Setup für den Mikrozensus 1993 weist folgende Besonderheiten auf:

- Der vom Statistischen Bundesamt gelieferte Rohdatensatz enthielt 886 Sätze von Personen, die nicht tatsächlich befragt wurden und zu denen keine Angaben vorliegen, d.h. die Variablen enthalten nur fehlende Werte. Diese Ausfallsätze wurden bei der Erstellung des SAS-Systemfiles ausgeschlossen (IF EF26 NE .).
- Durch einen Programmfehler im Statistischen Bundesamt wurde die Angabe zum Eheschließungsjahr (EF39, EF214) '1990' gelöscht und statt dessen der Wert 'ohne Angabe' vergeben.

Zentrum für Umfragen, Methoden und Analysen (ZUMA)
German Microdata Lab
B2,1
68159 Mannheim
Tel: 0621-1246-265
Fax: 0621-1246-100
http://www.gesis.org/ZUMA/index.htm

Kontakt: Andrea Lengerer, Julia H. Schroedter, Hossein Shahla (GML, Mikrozensus Grundfiles)
Email: lengerer@zuma-mannheim.de; schroedter@zuma-mannheim.de; shahla@zuma-mannheim.de