Verwendung von Proc Surveyselect im Rahmen des Matching-Verfahrens

Anja Marr
Institut für Medizinische Informatik,
Biometrie und Epidemiologie - IMIBE
Universitätsklinikum Essen
Hufelandstrasse 55
D-45122 Essen
anja.marr@uk-essen.de

Zusammenfassung

In Fall-Kontroll-Studien stellt sich oft die Aufgabe, dass für eine kleinere Anzahl Fälle aus einer größeren Auswahl potentieller Kontrollen Personen gefunden werden sollen, die den Fällen in Bezug auf vorgegebene Eigenschaften entsprechen. Dieses Matching-Verfahren soll mögliche statistische Störeffekte, auch Confounding genannt, auf ein Minimum reduzieren.
Nach Analyse der beiden Personengruppen (Fälle und potentielle Kontrollen) kann sich zeigen, dass bei einem gewünschten Matching-Ratio (1 Fall : n Kontrollen) mehr passende (matchende) Kontrollen zur Verfügung stehen, als benötigt werden. Hier soll dann nach einem Zufallsverfahren eine Stichprobe aus den matchenden Kontrollen gezogen werden. Oft muss aber auch erst ermittelt werden, ob bzw. in welchen Strata das gewünschte Matching-Ratio erreicht wird. (Stratum bezeichnet hier jede bei Fällen aufgetretene Wertekombination in den vorgegebenen Matching-Variablen.)
Anhand von Beispielen wird gezeigt, wie diese Aufgabe mittels grundlegender Data Steps und der Prozedur Surveyselect umgesetzt werden kann. Verschiedene Möglichkeiten in der Anwendung und wichtige Limitationen der Prozedur Surveyselect werden vorgestellt.

Schlüsselwörter: Matching, PROC SURVEYSELECT, epidemiologische Studien

1 Einleitung

In Fall-Kontroll-Studien werden Patienten mit der interessierenden Krankheit (Fälle) und nicht daran erkrankte Personen (Kontrollen) miteinander verglichen, um Risikofaktoren zu identifizieren, die die Krankheit verursachen bzw. ihre Entstehung fördern können. Bereits bei Studienkonzeption muss bedacht werden, ob Confounder in den beiden Gruppen vorhanden sein können. Dies sind Eigenschaften, die sowohl die Risikofaktoren als auch das Krankheitsauftreten beeinflussen. Kontrollen und Fälle sollen dann idealerweise in der Kombination dieser Confounder übereinstimmen, um statistische Verzerrungen zu minimieren. Das Verfahren, das diese Übereinstimmung erzielt, wird Matching genannt. Confounder, die im Matching berücksichtigt werden, sind dann Matchingvariablen.

Man unterscheidet zwischen individuellem und Häufigkeitsmatching. Bei individuellem Matching wird jedem Fall ein eigener Satz Kontrollen zugewiesen. Dadurch muss unabhängig vom gewünschten Fall-Kontroll-Ratio mindestens eine Kontrolle gefunden werden, damit der Fall in die spätere Auswertung eingehen kann. Bei Häufigkeitsmatching entfällt dieser Zwang, es wird lediglich angestrebt, für Fälle wie für Kontrollen die gleiche Verteilung bezüglich ihrer kombinierten Matchingvariablen zu erreichen.
Die Kontrollen können der gleichen Datenquelle entstammen wie die Fälle, z. B. dem Untersuchungskollektiv einer Krankenhausabteilung. Sie können aber auch unterschiedlichen Datenbeständen entstammen, Fälle beispielsweise aus Arztmeldungen und Kontrollen aus Einwohnermeldeamtsstichproben.
Idealerweise finden sich für einen Fall mehr matchende Kontrollen als benötigt. Dann muss für die Auswahl der Kontrollen ein Zufallsverfahren angewandt werden, um auswahlbedingten Verzerrungseffekten vorzubeugen. Die vorliegende Arbeit zeigt, wie die Prozedur Surveyselect genutzt werden kann, um eine Zufallsziehung unter Berücksichtigung der zu matchenden Variablen durchzuführen.

2 Aufgabenstellung

Für die vorliegende Arbeit wird festgelegt, dass Fälle und potentielle Kontrollen in einer gemeinsamen Datei vorliegen, wie z. B. in der diagnostisch untersuchten, retrospektiven Patientenreihe einer Krankenhausabteilung. Der Fall-Kontroll-Status leite sich aus dem Vorliegen einer Erkrankung ab und es werde ein Matching-Ratio von 1 Fall : 2 Kontrollen angestrebt. Die vorab festgelegten Confounder, auf die gematcht werden soll, seien Alter, Geschlecht und ein klinischer Score. Es wird auf Häufigkeit gematcht. Das Matching wird nur einmal für die Studie durchgeführt. Es gilt, zuerst zu ermitteln, ob das gewünschte Matching-Ratio erzielt werden kann, und danach die Kontrollen zufällig auszuwählen. Das Ergebnis des Zufallsverfahrens soll dokumentiert werden.

3 Vorgehensweise

3.1 Dateien aufbereiten

Im Idealfall finden sich für jeden Fall genügend Kontrollen, die für die Matchingvariablen die gleiche Wertekombination (Stratum) aufweisen. Dies ist in der Praxis nicht immer der Fall. Im ersten Data Step werden daher die Variablen klassiert, bei denen fürs Matchen keine völlige Übereinstimmung nötig oder möglich ist. Im vorliegenden Beispiel wird das Alter in 5-Jahres-Klassen gruppiert und der Score auf 2 Nachkommastellen gerundet. Die Klassenbreite wird in Abhängigkeit davon gewählt, wieviele Fälle bei völliger Übereinstimmung ohne Kontrollen bleiben und welche Abweichung in den Matching-Variablen akzeptabel ist. Das Ziel ist es, in möglichst allen bei Fällen auftretenden Strata die gewünschte Zahl an Kontrollen zu finden.
Im ersten Data Step werden Fälle und Kontrollen für die weitere Verarbeitung in getrennte Dateien ausgegeben. Personen, für die nicht alle Werte der Matchingvariablen bekannt sind, werden von der weiteren Verarbeitung ausgeschlossen.

```
DATA fall kontrolle;
  SET quelle.studie;

  *Altersgruppen bilden;
   IF 0<= Lebensalter <5  THEN altgruppe=1;
   IF 5<= Lebensalter <10 THEN altgruppe=2;
...
   IF 75<=Lebensalter <80 THEN altgruppe=16;
   IF 80<=Lebensalter THEN altgruppe=17;

   *Score klassieren;
   score=round(pr,0.01);

   *keine Missings bei Matching-Variablen erlauben;
   IF score=. OR altgruppe=. OR geschlecht=. THEN DELETE;

   *Gruppe 1=Fälle, Gruppe 2=Kontrollen;
   IF gruppe=1 THEN OUTPUT fall;
   ELSE IF gruppe=2 THEN OUTPUT kontrolle;
RUN;
```

3.2 Fall-Kontroll-Ratio definieren und umsetzen

Die Besetzung der Strata in der Fall-Datei und der Kontroll-Datei wird jeweils mittels Proc Freq ausgezählt und in eine Datei ausgegeben.

```
PROC FREQ DATA=fall;
  TITLE 'Verteilung der Fälle: Geschlecht, Altersklasse und Score';
  TABLES geschlecht*altgruppe*score/LIST OUT=fallout;
RUN;
```

Die beiden Ausgabe-Dateien werden dann im nächsten Data Step über die Matchingvariablen gemergt. Die Anzahl der Fälle wird mit der gewünschten Kontrollzahl multipliziert und in die Variable _nsize_ geschrieben. Die Variable _nsize_ wird von Proc Surveyselect benötigt, wenn man bei der Option Sampsize einen Dateinamen angibt. Zur Beurteilung, wie gut das Matchen funktioniert, wird die Differenz aus benötigter und vorhandener Zahl Kontrollen im Stratum gebildet.

```
DATA suchwenig;
  MERGE fallout     (DROP=percent RENAME=(count=anzfall))
        kontrollout (DROP=percent RENAME=(count=anzkont));
  BY geschlecht altgruppe score;

  *Variable, die Anzahl zu ziehender Kontrollen pro Stratum enthält;
  *pro Fall 2 Kontrollen;
  _nsize_=anzfall*2

  *Wenn anzkont missing ist, gab es keine Kontrollen im Stratum;
  IF anzkont=. THEN anzkont=0;
```

```
  *Wenn _nsize_ missing ist, gab es keine Fälle im Stratum;
  *nur Strata, aus denen gezogen werden soll;
  IF _nsize_=. THEN delete;

  stratdiff=anzkont-_nsize_;

  LABEL stratdiff='Differenz (vorhanden-nötig)'
        anzkont='Anzahl vorhandener Kontrollen'
        _nsize_='Anzahl benötigter Kontrollen';
RUN;
```

Jetzt können in drei Proc Print-Steps die Strata mit ungenügender, ausreichender und überreichlicher Zahl an Kontrollen dargestellt werden. Das Sum-Statement gibt dabei einen schnellen Überblick, in welcher Größenordnung sich die Differenzen zwischen benötigter und vorhandener Zahl an Kontrollen befinden.

```
PROC PRINT DATA=suchwenig LABEL N;
  TITLE 'Strata, in denen Kontrollen fehlen';
  WHERE stratdiff<0;
  VAR geschlecht altgruppe score stratdiff _nsize_ anzkont;
  SUM _nsize_ anzfall anzkont;
RUN;
```

Stellt sich jetzt heraus, dass die Menge der unterbesetzten Kontroll-Strata unakzeptabel hoch ist, muss geklärt werden, ob die Matchingvariablen etwas gröber klassiert werden dürfen oder ob das Matching-Ratio reduziert werden soll. Stellt sich andererseits heraus, dass eine sehr große Zahl von Strata überreichlich viele Kontrollen besitzt, kann überlegt werden, Matchingvariablen feiner oder gar nicht zu klassieren bzw. das Fall-Kontroll-Ratio zu erhöhen. Bei Änderungen wird das Programm bis zu diesen Proc Print-Steps so lange wiederholt, bis das resultierende Matching-Ergebnis akzeptabel ist. Kann kein zufriedenstellendes Ergebnis erzielt werden, da sich die Gruppen der Fälle und der potentielle Kontrollen zu sehr unterscheiden, sollte von der Verwendung des hier vorgestellten Programmes abgesehen und auf andere Algorithmen [1] zurückgegriffen werden.

Abschließend wird in einem weiteren Data Step die eigentliche Stichprobengröße für die Ziehung festgelegt. Bei ausreichend und überreichlich besetzten Strata bleibt die Variable _nsize_ unverändert. In unterbesetzten Strata empfiehlt es sich jedoch, _nsize_ auf die tatsächlich vorhandene Zahl Kontrollen zu reduzieren. Zwar zieht Proc Surveyselect auch in solchen Strata, wenn die Option Selectall angegeben wird. Wird _nsize_ aber nicht angepasst, ergeben die aufsummierten Kontrollzahlen in dieser Datei eine höhere Summe als die der resultierenden Stichprobenziehung. Die formale Prüfung des SAS-Logs würde dadurch aufwendiger.

```
DATA noetig;
  set suchwenig;
```

```
  *In unterbesetzten Strata dafür sorgen, dass alle vorhandenen
   Kontrollen gezogen werden;
  if stratdiff<0 and anzkont>0 then _nsize_=anzkont;

  *nicht aus Strata ziehen, die keine Kontrollen enthalten;
  if stratdiff<0 and anzkont=0 then delete;

RUN;
```

3.3 Kontrollen ziehen

Mit Version 7 wurde eine Gruppe von Prozeduren zur Analyse von Stichprobenerhebungen eingeführt [2]. Hierzu gehört Proc Surveyselect, das viele Möglichkeiten bietet, Stichproben aus einer Datei zu ziehen. Für die hier beschriebene Aufgabenstellung kommen nur Ziehungsmethoden ohne Zurücklegen in Betracht, da keine Kontrolle mehrfach gezogen werden darf. Default-Methode ist simple random sampling, wenn, wie im vorliegenden Fall, kein Size-Statement angegeben wurde.
Die zu berücksichtigenden Matchingvariablen werden im Strata-Statement aufgelistet. In der dortigen Reihenfolge muss die Quelldatei mit den verfügbaren Kontrollen sortiert sein. Die hinter Sampsize= angegebene Datei muss neben den Matchingvariablen eine Variable _nsize_ enthalten, die pro Stratum die Zahl der zu ziehenden Datensätze enthält. Hinter Out= wird die Ergebnisdatei der Ziehung benannt. Wenn, wie hier, kein Id-Statement definiert wurde, enthält diese Ergebnisdatei alle Variablen der Quelldatei. Die Option Outseed schreibt den im jeweiligen Stratum für die Zufallsauswahl verwendeten Startwert in die Ausgabedatei. Die bereits erwähnte Option Selectall bewirkt, dass alle verfügbaren Datensätze in einem Stratum gezogen werden, wenn die angeforderte Zahl die verfügbare Zahl an Datensätzen übersteigt.

```
PROC SURVEYSELECT DATA=kontrolle SAMPSIZE=noetig OUT=ergebnis1
OUTSEED SELECTALL;
  STRATA geschlecht altgruppe score;
  TITLE 'Ziehung';
RUN;
```

Proc Surveyselect gibt für bestimmte Strata Error-Meldungen im Log aus, die im Rahmen des Matchingverfahrens nicht unbedingt sinnvoll sind. So erscheint ein Error für angetroffene Strata, aus denen nicht gezogen werden soll, da in diesem Stratum keine Fälle vorliegen (_nsize_ ist missing). Ebenso erscheint ein Error, wenn _nsize_ mehr Datensätze anfordert als vorhanden und die Option Selectall nicht angegeben ist. Ist diese Option angegeben, so wird die Meldung als Note ins Log geschrieben. Unglücklicherweise gibt es aber keine Fehlermeldung, wenn aus einem Stratum gezogen werden soll, aber keine Datensätze für dieses Stratum existieren.
Aus diesen Gründen empfiehlt es sich, vor Ziehung alle Strata zu entfernen, aus denen nicht gezogen werden soll, und ebenso alle Strata aus der Sampsize-Datei zu eliminieren, für die keine Kontrollen existieren.

Nach erfolgreicher Ziehung der Kontrollen werden wichtige formale Ziehungsdetails in den Daten dokumentiert. Dies ist mindestens das Datum der Ziehung, ggf. auch der im Stratum verwendete Seed, um die Zufallsauswahl nachvollziehen zu können.

```
DATA quelle.ziehung;
  SET ergebnis;
  KEEP id ziehdat;
  ziehdat=today();
  FORMAT ziehdat ddmmyy10.;
RUN;
```

Damit ist die gematchte Stichprobenziehung abgeschlossen. Weitere Programmschritte hängen davon ab, ob und wohin die Kontrollstichprobe exportiert wird.

3.4 Modifikationen für andere Studienvarianten

Bei individuellem Matching wird der gezogenen Kontrolle in einer zusätzlichen Variablen die ID des zugehörigen Falles zugewiesen oder es wird eine MatchID eingeführt, die für Fall und Kontrolle identische Werte besitzt (Siehe auch [3]).
Wenn im Laufe einer Studie mehrmals Stichproben aus einer Kontrolldatei gezogen werden, weil Fälle prospektiv erfasst werden, wird in der Kontrolldatei eine Statusvariable eingeführt, die anzeigt, ob die Kontrolle bereits gezogen wurde. Der neue Status gezogener Kontrollen muss dann in die Quelldatei zurückgespielt werden und neue Ziehungen dürfen nur aus der Gruppe der noch ungenutzten Kontrollen erfolgen.
Mehrmalige Stichprobenziehungen fallen auch an, wenn die selektierten Kontrollen die Studienteilnahme verweigern oder nach Aktenstudium von der Studie ausgeschlossen werden müssen. Hier muss dann das Rekrutierungsergebnis in die Berechnung von _nsize_ einfliessen.

4 Diskussion

Das vorgestellte Beispiel liess sich ohne größeren Zeitaufwand programmieren und hat sich in mehreren Projekten bereits bewährt. Das Matching-Ratio kann von 1:2 auf andere Verhältnisse geändert werden. Auch die Zahl der Matchingvariablen kann variiert werden. Bei sehr vielen Variablen kann aber der ungünstige Fall auftreten, dass nur wenige Strata von Fällen und Kontrollen gleichzeitig besetzt sind und auch Klassierung der Matchingvariablen nur unzureichend Abhilfe schafft. Wenn die Zahl der Matchingvariablen nicht reduziert werden kann, sollte andere Matching-Algorithmen oder die Verwendung von Propensity Scores geprüft werden. Propensity Scores können mittels logistischer Regression gebildet werden und fassen mehrere Variablen zu einem Einzelwert zusammen. Ein Propensity Score würde wie eine gewöhnliche Matchingvariable behandelt, die Übereinstimmung zwischen Fällen und Kontrollen sollte aber durch sehr eng gefasste Klassen erzielt werden.

Das Programm ist auch für Anwender mit geringen SAS-Kenntnissen geeignet, da bis auf Proc Surveyselect nur grundlegende Programmierschritte wie Data Steps, Proc Print und Proc Freq verwendet werden. Die Ausgaben ermöglichen einen schnellen Überblick über die Datenlage und bieten sofortige Überprüfung vorgenommener Modifikationen der Matchingvariablen oder des Matching-Ratios.
Der Vorteil in der Verwendung von Proc Surveyselect liegt in der Bandbreite der zur Verfügung stehenden Ziehungsmethoden und der Nachvollziehbarkeit des Ziehungsergebnisses.

Literatur

[1] M. Schröder, J. Hüsing, K.-H. Jöckel: An Implementation of Automated Individual Matching for Observational Studies. Methods Inf Med 2004, 43; 516-520

[2] A. An, D. Watts: New SAS® Procedures for Analysis of Sample Survey Data, SUGI Proceedings 23; Paper 247, 1998

[3] R. Diseker, K. Permanente: Simplified Matched Case-Control Sampling using PROC SURVEYSELECT. SUGI Proceedings 29, Paper 209-29, 2004