SAS-Makros für Epidemiologische Studien

Hans-Peter Altenburg

Deutsches Krebsforschungszentrum
Abt. Klinische Epidemiologie / C0500
D-69120 Heidelberg
E-Mail: hp.altenburg@dkfz.de

Abstract

Nested Case-Control oder Case-Cohort-Studien sind epidemiologische Studiendesigns, die häufig bei Kohortenstudien angewandt werden, um Risikokennzahlen für seltene Erkrankungen abzuschätzen. Während nested Case-Control-Studien einen retrospektiven Charakter haben, sind Case-Cohort-Studien prospektiv angelegt. Für beide Studientypen kann die SAS-Prozedur PHREG für eine Analyse verwendet werden, da als Modell für die Erkrankungsrate in beiden Fällen ein semi-parametrischer Ansatz zugrunde gelegt werden kann.

Es werden einige SAS-Makros vorgestellt, die es erlauben solche, in große Kohortenstudien eingebetteten, Studiendesigns auf einfache Art und Weise auszuwerten. Wesentlich für die Analyse ist dabei eine geeignete Auf- bzw. Vorbereitung der Datenbasis. Auch hierfür werden entsprechend geeignete Makros beschrieben.

1. Einleitung

Kohortenstudien spielen in der Epidemiologie eine zentrale Rolle. Beziehungen zwischen individuellen Charakteristiken oder Expositionen und dem Auftreten einer Erkrankung oder spezieller krankheitsbezogener Ereignisse können mit Hilfe eines Kohortenstudiendesigns entdeckt, um dann evtl. in Hypothesen und Strategien zur Krankheitsprävention umgesetzt zu werden. Kohortenstudien sind in der Regel longitudinal und prospektiv angelegt, was ihre Realisierung aufwendig und von den Kosten her gesehen teuer macht.

Kohortenstudie

Das Wort Kohorte, abgeleitet aus lat. „cohors“, Krieger(-anzahl), dem 10. Teil einer römischen Legion, beschreibt eine bestimmte Gruppe von Personen (Subpopulation) aus einer Bevölkerung, die über eine definierte zeitliche Periode beobachtet und erfaßt (identifiziert) wird. Ziel ist es, diejenigen Todesursachen oder Erkrankungen in dieser Kohorte zu identifizieren, die gehäufter oder seltener auftreten. Außerdem sollen Teilpopulationen in der Zukunft einer bestimmten Exposition eines Faktors ausgesetzt sein kann, von dem angenommen wird, daß er zu einer Erkrankung führen kann. Als Beispiel für eine große Kohortenstudie sei die EPIC-Studie (European Prospective Investigation into Cancer and Nutrition) genannt, in der europaweit die Zusammenhänge zwischen Ernährungsgewohnheiten und der Entstehung verschiedener Krebserkrankungen untersucht werden sollen.
Der Vorteil einer prospektiv angelegten Kohortenstudie liegt darin, daß die Exposition eines möglichen Faktors gemessen wurde bevor die Erkrankung diagnostiziert wird. Nachteilig wirken sich vor allem die hohen Kosten und die lange Dauer aus bis brauchbare Ergebnisse vorliegen können, da für die meisten Erkrankungen die während der Beobachtungsperiode zu erwartende Anzahl von Ereignissen klein ist gegenüber dem Umfang der Kohorte. Somit sind die meisten Studienmittel, sei es das Datenmaterial als auch das gesammelte biologische Material (wie etwa Blutproben), Probanden gewidmet, die nur wenig Einfluß auf die Studienergebnisse haben.
Als ein möglicher Ausweg aus dieser Problematik wurden von Liddell et al (1977) das alternative Studiendesign, die *nested Case-Control-Studie*, und danach von Prentice (1986)

das *Case-Cohort-Studiendesign* vorgeschlagen. Beide Studientypen verwenden als Basis eine Kohorte bzw. die Datenbasis einer Kohortenstudie.

Fall-Kontroll-Studie

Eine Fall-Kontrollstudie ist eine epidemiologische Beobachtungsstudie, welche Personen, die eine interessierende Krankheit haben (Gruppe der Fälle), mit einer Gruppe von nicht erkrankten Personen (Kontrollgruppe) vergleicht. Ziel ist es, diejenigen Faktoren zu ermitteln, unter denen die Wahrscheinlichkeit für die Erkrankung erhöht bzw. erniedrigt ist. Die Beziehung eines Expositionsattributes zur Erkrankung wird dabei durch den Vergleich der Häufigkeiten des Attributes bei Erkrankten und Nichterkrankten ermittelt. Als Standardkennzahl wird das Quotenverhältnis oder relative Chance (engl. odds ratio (OR)) verwendet. Fall-Kontroll-Studien sind die am häufigsten verwendete Form der Studiendesigns in der analytischen Epidemiologie. Ihr wichtigstes Merkmal ist die retrospektive (d.h. rückblickende) Informationsgewinnung für die Erfassung der Faktoren.

Nested Case-Control-Studie

Eine nested Case-Control-Studie ist eine Fall-Kontroll-Studie bei der die Fälle und Kontrollen aus der Population (Datenbasis) einer Kohortenstudie gezogen werden. Der Vorteil liegt darin, dass die Daten für Fälle und Kontrollen bereits im Vorfeld verfügbar sind und die Effekte potentieller Confounder-Variablen reduziert werden können. Die Kontrollen werden aus der Teilmenge der Probanden unter Risiko in der Kohortenpopulation gezogen zum Zeitpunkt des Auftretens eines Falles in der Kohorte. Eine nested Case-Control-Studie besitzt die Kostenvorteile einer Fall-Kontroll-Studie ohne den Selektionsbias zu besitzen, der durch die Auswahl der Kontrollen bei einer normalen Fall-Kontroll-Studie auftreten kann. Nachteilig kann sich aber auswirken, dass Fälle und Kontrollen nicht unbedingt nach der Follow-Up-Dauer gematcht werden. Dieser Nachteil wird bei Case-Cohort-Studien ausgeglichen. Nested Case-Control-Studien sind eng verwandt mit populations-gematchten Fall-Kontrollstudien.

Beispiel für eine Nested Case-Control-Studie

Zielerkrankung: Rheumatoide Arthitis (RA)
Datenbasis bzw. Kohorte: Blutproben von n=10.000 Individuen

	nach 10 Jahren:	
n=200 mit RA		n=9800 ohne RA
		Stichprobe n=400 ohne RA
	serologischer Test	
T_+=80 T_- =120		T_+=40 T_- =360
	Odds Ratio : OR=6	

Case-Cohort-Studie

In einer Case-Cohort-Studie werden die Vorgeschichten von Fällen mit den Vorgeschichten von Nichtfällen aus der gleichen Kohorte verglichen, wobei Fälle und Nichtfälle nach der Dauer des Follow-Ups gematcht werden. Als Grundlage (Datenbasis) dient eine zufällig gezogene Stichprobe aus dem Gesamtdatenpool einer Kohortenstudie, z.B. eine 10% Subkohorte. Wichtig ist, dass die Subkohorte ausgewählt wird ohne Beachtung des Ereignisstatus! Eine Variante dieses Designs fügt (evtl. zu einem späteren Zeitpunkt) zu

dieser ausgewählten Menge noch alle Probanden hinzu, die das Zielereignis erfahren, d.h. also alle erkrankten Fälle zu einem späteren Zeitpunkt, die nicht bereits in der Subkohorte sind.

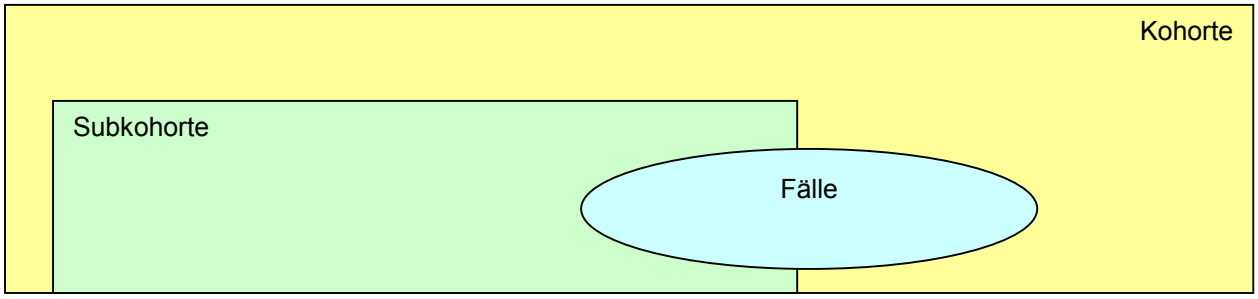

Abb. 1: Datensituation bei einer Case-Cohort-Studie

Das Ziel dieses Papers ist, einige Aspekte der Umsetzung dieser Studiendesigns in praktische Anwendungen mit dem Schwerpunkt, die für die Realisierung notwendigen Schritte mit Hilfe von SAS-Makros durchzuführen,

- SAS-Makros für die Datenaufbereitung sowie
- SAS-Makros für die Datenanalyse.

Alle im folgenden beschriebenen Makros werden im August 2001 im SAS-Ah zur Verfügung gestellt. Die URL des Inhaltsverzeichnisses lautet:

http://www.urz.uni-heidelberg.de/statistik/sas-ah/

2. Datenaufbereitung

Ein besonders wichtiger und oft auch aufwendiger Aspekt in epidemiologischen Studien ist die Aufbereitung der Daten, damit aussagekräftige Analysen durchgeführt werden können. Für die Datenaufbereitung sollen hier exemplarisch fünf Makros beschrieben werden:

- Aufteilen einer SAS-Datei mit Fällen und Kontrollen,
- Zufällige Selektion einer Teilmenge,
- Matching und
- Klassifizierung sowie Aufsplitten von quantitativen Variablen in Quantil-Gruppen.

Aufteilen einer SAS-Datei mit Fällen und Kontrollen

Das folgende SAS-Makro dient zum Vorbereiten der Datentabelle, wie es z.B. für die Anwendung des Matching Makros benötigt wird. Die SAS-Datei „`dset`“ mit Fällen und Kontrollen wird in zwei separate Datentabellen aufgesplittet, deren Namen als Parameter „`cases`“ bzw. „`controls`“ angegeben werden können. Mit „`dset`“ wird die Eingabedatei bezeichnet. Der Parameter „`c_var`“ gibt die Variable an, welche die Information zu Fällen und Kontrollen enthält. Abhängig von der Werteliste „`c_value`“ (=Werte mit denen Kontrollen kodiert sind) erfolgt dann die Aufteilung in Fälle und Kontrollen. „`c_value`“ darf auch eine mit Leerzeichen getrennte Liste von Werten enthalten.

Makro:

```
%MACRO caco_div(dset,cases,controls,c_var,c_value);
/* ----------------------------
   Macro divides the data set "dset" in two new
   data sets "cases" and "controls" according to
   the value "c_value" of the variable "c_var".
   "c_value" denotes the keys of the controls and
   it can be a list of values. If the
   variable is not numeric the values listed at
   "c_value" must be included in ", e.g.

   possible values of c_var: A, B, E, Z
   where the controls are A, B, E
   then the values of c_value should be: "A" "B" "E".

   Macro Parameters:
   ================

   dset     data set to be divided
   cases    data set to contain the cases
   controls data set to contain the controls
   c_var    variable used to divide
   c_value  variable values for group cases

   Example to call the macro:
   =======

      %caco_div(all, cases, controls, caco, 0)

   or with character values of the variable caco:

      %caco_div(all, cases, controls, caco, "A" "B" "E")

   ------------------------------------------------- */
DATA &cases &controls ;
SET &dset ;
IF &c_var IN ( &c_value ) THEN OUTPUT &cases ;
                          ELSE OUTPUT &controls ;
RUN ;
%MEND ;
```

Aufruf:

Die Datenmenge „`all`“ enthalte die Variable „`caco`“ mit den Ausprägungen 0 und 1, wobei 1 die Fälle bezeichnet. Mit dem Makro-Aufruf

```
%caco_div(all, cases, controls, caco, 0)
```

wird dann die Datei “`all`” in zwei neue Dateien „`cases`“ und „`controls`“ aufgesplittet.

Zufällige Selektion einer Teilmenge

Für Case-Cohort-Studien muss eine Zufallsstichprobe aus der zugrundeliegenden Datenbasis gezogen werden. Das folgende Makro zieht eine Zufallsstichprobe aus einer Basisdatei. Aus der Basisdatei „`dset`“ wird eine Zufallsstichprobe gezogen und in der Datei „`subset`“ abgelegt. Der Umfang der Teilstichprobe kann über den Parameter „`proportn`“ als Zahl zwischen 0 und 1 angegeben werden. Zwei verschiedene Arten eine Teilstichprobe zu ziehen können über den Eingabeparameter „`coin`“ gesteuert werden, „exakt“ oder „Münzwurf“ (Schlüsselworte: `E` oder `exact` bzw. `C` oder `coin`). „exakt“ bedeutet, dass der Stichprobenumfang sich genau aus dem Anteil `proportn` multipliziert mit dem Stichprobenumfang der Basisdatenmenge bestimmt. Die Option „Münzwurf“ liefert eine Teilstichprobe wie sie auch das Ergebnis beim Werfen einer Münze wäre: Nur im Durchschnitt erhält man den unter `proportn` angegebenen Anteil für die Teilstichprobe.

Makro:

```
%MACRO subcoh(dset, subdset, proportn, coin, start);
/* ------------------------------------------
  Macro draws a random subsample according two
  alternative mechanisms:
  Exact (E), that means the sample size of the
             subsample is quasi "exact" (rounded)
             the proportion asked for,
  coin (C),  simulates a coin throwing process, which
             does not necessary result in an exact
             proportional sample size.

  Parameters:
  ==========

  dset       name of the basic data set,
  subdset    name of the sub data set,
  proportn   proportion of the basis data set, the
             sub sample should have,
  coin       drawing mechanism, as described above
  start      starting number for the random mechanism.

  Example:
  =======

    %subcoh(main,sub,0.10,E,1234567)

   draws a 10% sub sample from the data set main.

  ------------------------------------------ */

%IF %UPCASE(&coin)=C OR %UPCASE(&coin)=COIN  %THEN %DO ;
DATA &subdset ;
SET &dset ;
xxx_var=RANUNI(&start) ;
IF xxx_var <=&proportn ;
DROP xxx_var ;
RUN ;
                                                    %END ;
                                                %ELSE
%IF %UPCASE(&coin)=E OR %UPCASE(&coin)=EXACT %THEN %DO ;
DATA _NULL_ ;
n=%n_obs(&dset) ;
```

```
n_obs=ROUND( n * &proportn) ;
CALL SYMPUT('n_obs',n_obs) ;
RUN ;
DATA xxx_dset ;
SET &dset ;
xxx_var=RANUNI(&start) ;
RUN ;
PROC SORT DATA=xxx_dset ; BY xxx_var ;
DATA &subdset ; SET xxx_dset (OBS=&n_obs) ;
DROP xxx_var ;
RUN ;
PROC DATASETS ; DELETE xxx_dset ; RUN ;
                                                    %END ;
                                                %ELSE
%PUT ******* Error ******* Input: &coin ******* ;
DATA _NULL_ ;
             FILE PRINT ;
PUT / "INPUT  Data Set:  &dset"
    / "Sample Size:      %n_obs(&dset)"
    / "Proportion:       &proportn "
    / "Method:           &coin" +5 '(E=Exact, C=Coin)'
   // "OUTPUT Data Set:  &subdset"
    / "Sample Size:      %n_obs(&subdset)"
    ;
RUN ;
%MEND ;
```

Parameter:

`dset`	Name der Basisdatenmenge,
`subdset`	Name der zufällig gezogenen Teilmenge,
`proportn`	Anteil, der gezogen werden soll,
`coin`	Ziehungsmechanismus (Parameter `E` bzw. `EXACT` oder `C` bzw. `COIN`),
`start`	Startzahl für den Zufallszahlengenerator.

Ausgegeben werden zur Kontrolle alle Eingabeparameter sowie die Stichprobenumfänge der Basisdatei und der Teilmenge. Das Makro verwendet das Makro „n_obs“ zur Bestimmung des Stichprobenumfanges einer SAS-Datentabelle (siehe C. Ortseifen et al: 4. KSFE (2000) in Gießen, Session Tipps und Tricks).

Beispielaufruf:

Aus der SAS-Datei „`main`“ soll eine 0.9%- und 10%-Stichprobe gezogen werden und in der Datei `sub` abgespeichert werden:

Programm (Anteil 0.9%**):**

```
%INCLUDE 'D:\sasdat\m\n_obs.sas' ;
%INCLUDE 'D:\sasdat\mepi\subcoh.sas' ;
OPTIONS MPRINT ;
%subcoh(main,sub,0.009,E,32894817)
%subcoh(main,sub,0.009,C,32234815)
```

Ausgabe (Anteil 0.9%):

```
INPUT  Data Set:  main
Sample Size:      2000
Proportion:       0.009
Method:           E     (E=Exact, C=Coin)
OUTPUT Data Set:  sub
Sample Size:      18
```

```
INPUT  Data Set:  main
Sample Size:      2000
Proportion:       0.009
Method:           C     (E=Exact, C=Coin)

OUTPUT Data Set:  sub
Sample Size:      21
```

Programm (Anteil 10%**):**

```
%INCLUDE 'D:\sasdat\m\n_obs.sas' ;
%INCLUDE 'D:\sasdat\mepi\subcoh.sas' ;
OPTIONS MPRINT ;
%subcoh(main,sub,0.1,E,32894817)
%subcoh(main,sub,0.1,C,32234815)
```

Ausgabe (Anteil 0.9%):

```
INPUT  Data Set:  main
Sample Size:      2000
Proportion:       0.1
Method:           E     (E=Exact, C=Coin)

OUTPUT Data Set:  sub
Sample Size:      200

INPUT  Data Set:  main
Sample Size:      2000
Proportion:       0.1
Method:           C     (E=Exact, C=Coin)

OUTPUT Data Set:  sub
Sample Size:      185
```

Beispiel Auswahl der Subkohorte in einer Case-Cohort-Studie:

Orginalkohorte: n = 458312

```
%subcoh(lung1200,lung_sub,0.1,E,3287917)
```

liefert

N_{sub}= 45831

dagegen liefert

```
%subcoh(lung1200,lung_sub,0.1,C,3287917)
```

N_{sub}= 46702

Probanden für die Subkohorte.

Matching

In der Epidemiologie können gültige Schätzungen der Stärke einer Beziehung zwischen einer Exposition und der Erkrankung nur nach Betrachtung von sog. Confoundern aufgestellt werden. Wenn sie in der Analyse ignoriert werden, so kann dies zu inkonsistenten und ineffizienten Schätzwerten führen. Zwar kann der Einfluss von Confoundern auf die Schätzwerte durch Adjustierung in einem gewissen Umfang ausgeglichen werden, jedoch nicht wenn zwischen dem Confounder und der Exposition eine Wechselwirkung besteht (d.h. der Confounder ein Effektmodifizierer ist) oder die Verteilung der Exposition in verschiednen Strata des einen (oder mehrerer Confounder) unbalanziert ist. Solche Probleme mit Confoundern können (zumindest teilweise) bereits mit Hilfe eines entsprechenden Versuchsdesigns vermieden werden. Zwei Methoden stehen hier zur Verfügung: Die totale Einschränkung (engl. total restriction) und das Matching, was eine Art von partieller Einschränkung darstellt. Hierbei werden nicht die Fälle sondern nur die Kontrollen (bzw. „Nichtfälle“) nach bestimmten Restriktionen ausgewählt.
Matching ist also ein Prozeß mit dem Ziel, die zu vergleichenden Gruppen (wie z.B. Fälle und Kontrollen) hinsichtlich ihrer Struktur der Confounder vergleichbar zu machen. Verschiedene Methoden stehen in der Epidemiologie zur Verfügung, von denen aber nicht alle im Makro umgesetzt wurden. Ein paar wichtige Matching-Verfahren seien hier genannt:

- individuelles (Paar- oder 1:1-), 1:N Matching: Hierbei werden jedem Fall eine (1:1) bzw. mehrere (1:M) Kontrollen zugeordnet, wobei die Fälle und Kontrollen nach in den Werten der Matching-Variablen übereinstimmen müssen. Paarmatching ist individuelles Matching, wobei jeweils Paare gebildet werden zwischen den entsprechenden Fällen und Kontrollen.
- Caliper-Matching (engl. caliper, Greifzirkel): Der Wert der Matching-Variablen bei den Kontrollen darf in einem gewissen Abstand zum Wert der Matching-Variablen bei dem entsprechenden Fall sein, z.B. Alter $\pm$ 2 Jahre.
- Häufigkeits-Matching (Verteilung): Hier brauchen nur die Verteilung der Matching-Variablen bei Fällen und Kontrollen übereinzustimmen.
- Kategorie-Matching (Subjekte in breiten Klassen): Die Matching-Variablen werden in breiten Klassen zusammengefasst.

In der Praxis entsteht dabei oft das Problem: Ein optimale Zuordnung ist bei einer großen Anzahl von in Frage kommenden Probanden (aus dem Reservoir der Nichterkrankten) per Hand „recht mühsam“ durchzuführen. Das folgende Makro erlaubt den Prozeß des Matching mit Hilfe eines SAS-Makros durchzuführen.

Makro -Aufruf:

```
%matching(cases,controls,idca,idco,mtchvars,
          n_ca,opt,startca,startco,outdset,print)
```

Hierbei bedeuten:

`cases`	DATA-Set, welche die Fälle enthält
`controls`	DATA-Set, welche die Kontrollen enthält,
`idca`	ID-Variable für die Identifizierung der Fälle
`idco`	ID-Variable für die Identifizierung der Kontrollen
`mtchvars`	Liste der Matching-Variablen
`n_ca`	Anzahl der auszuwählenden Kontrollen,
`opt`	Verfahren (E: Euklidischer Abstand | P: Propensity-Score)
`startca`	Startzahlen für die Erzeugung der Zufallszahl (Fälle),
`startco`	Startzahlen für die Erzeugung der Zufallszahl (Kontrollen),

`outdset`	Ausgabedatei in der die gematchten Kontrollen stehen
`print`	Druckausgabe (P | PR | PRINT) der Datei `outset`

Falls in der zugrundeliegenden Datenbasis Fälle und mögliche Kontrollen zusammen enthalten sind, muss in einem vorgeschalteten Datenschritt, diese SAS-Datentabelle in zwei getrennte Dateien aufgeteilt werden. Dies kann auch mit Hilfe des oben beschriebenen SAS Makros `caco_div` geschehen.

Da das Makro ziemlich groß ist sei für eine Kopie auf die oben bereits erwähnten Internetseiten im SAS-Ah verwiesen.

Beispielaufruf:

1:1 Matching nach Geschlecht, Alter und Geburts-Jahr.

Makro-Aufruf:

```
%match(d_case, d_cont, centerID, centerID,
       gender age b_year,1,E,6579213,6619957,M_Out,noprint)
```

Makros speziell für Ernährungsdaten:

Aufsplitten einer Nahrungsmittelgruppe in zwei neue Variablen: eine diskrete Variable mit der Ausprägung Konsument / kein Konsument und eine zweite stetige Variable, welche die Konsummenge bei Konsumenten (null oder kleiner als ausgeschlossen) enthält. Fehlende Werte (Missing) der Ursprungsvariable werden in der diskreten Variablen als 9 kodiert.

Makro:

```
%MACRO klassi(var,nvard,nvarc,vlabel,dlabel,clabel,d_form,c_form) ;
/* ------------------------------------------------
   Splits the variable VAR into two new variables
   with names NVARd (discrete) and NVARc (continuous).
   NVARd: has the values 0 if VAR = 0,
                         1 if VAR > 0,
                         9 if VAR has a missing value,
   NVARc: has the values
          . (missing)      if VAR <= 0,
          value of the var if VAR > 0.
   vlabel, clabel and dlabel: contain the label text for the
                              variable as well as the
                              discrete and continuous part
   d_form, c_form: Formats of the discrete and continuous variables.
   ------------------------------------------------ */
%IF %LENGTH(&c_form=0) %THEN %LET c_form=8.2 ;
&nvard=&var ;
&nvarc=&var ;
IF &var=. THEN &nvard=9 ;
          ELSE IF &var>0 THEN &nvard=1 ;
IF &var<=0 THEN &nvarc=. ;
LABEL &nvard="&vlabel / &dlabel"
      &nvarc="&vlabel / &clabel"
      ;
FORMAT &nvard &d_form.
       &nvarc &c_form.  ;
%MEND ;
```

Das Makro kann nur innerhalb eines DATA-Steps verwendet werden!

Beispiel:
Die Variable qg04 (fruits) soll in einen diskreten (Var.: QG04_d) und stetigen Teil (Var.: QG04_c) aufgesplittet werden:

Aufruf in einem DATA-Step:

```
DATA neu ; SET alt ;
. . .
%klassi(qg04,QG04_d,QG04_c,
        Fruits,discrete:.|0|>0,continuous,takeYN,10.4) ;
. . .
RUN ;
```

Über die Prozedur FREQ kann man beispielsweise kontrollieren, ob die Einteilung richtig erfolgte:

```
PROC FREQ DATA=neu ;
TABLES qg04_d ;
RUN ;
```

Ausgabe:

Häufigkeitsliste für den diskreten Teil (Variable QG04_d)

	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Fruits / discrete: . | 0 | >0 QG04_d				
No, do not take	755	0.58	755	0,58
Yes, I take	126621	97.08	127376	97.66
Missing	3053	2.34	130429	100.00

Stetige Variablen in Quantile aufteilen

In vielen epidemiologischen Studien werden oft zur besseren Erkennung eines Effektes quantitative Variablen in Quantile, wie z.B. Terzile (331/3%-Anteile), Quartile (25% Anteile) oder Quintile (20%-Anteile), aufgeteilt. Mit Hilfe des folgenden Makros kann dies für beliebige Anteile durchgeführt werden. Das Makro benutzt die Prozedur UNIVARIATE:

Makro:

```
%Macro quantils(dset,newset,q_out,outset,where,var,n_q,newvar,text1);
/* ----------------------------------------
   Quantil1.sas one variable only
   ---------------------------------------- */

TITLE1 "&text1" ;
DATA _NULL_ ;
n_q =&n_q ;                    IF n_q<2 THEN n_q=2 ;      * !!!!! ;
n_q1=n_q-1 ;                   CALL SYMPUT('n_q1',n_q1) ;
n_q2=n_q-2 ;                   CALL SYMPUT('n_q2',n_q2) ;
n_q3=n_q-3 ;                   CALL SYMPUT('n_q3',n_q3) ;
intval=FLOOR(100/n_q) ;        CALL SYMPUT('intval',intval) ;
i_start =0 ;                   CALL SYMPUT('i_start',i_start) ;
```

```
p_end1  =n_q1*intval ;        CALL SYMPUT('p_end1',p_end1) ;
p_end   =n_q *intval ;        CALL SYMPUT('p_end',p_end) ;
RUN ;

PROC UNIVARIATE DATA=&dset
%IF %LENGTH(&where) >0 %THEN (WHERE=(&where) );
   NOPRINT ;
VAR &var ;
OUTPUT  OUT=&outset
   pctlpre =q
   pctlpts =&intval TO &p_end  BY &intval ;
    ;
RUN ;

PROC PRINT DATA=&outset ; RUN ;
DATA _NULL_ ; SET &outset ;
              FILE PRINT ;
PUT / "Quantiles for the variables: &var"  //
      "Number of Quantiles: &n_q"           //
      'Variable' +15 'Quantiles (Upper Cutpoints)'
  //  "&var" @@
;
%LET nnq=0 ;
%DO j=&intval %TO &p_end  %BY &intval ;
%LET nnxx=&nnq + 1 ;
%LET  nnq=%EVAL(&nnxx) ;
         x=q&j ; PUT x 10.3 @@ ;
     CALL SYMPUT("q&nnq",x) ;
%END ;    PUT ;
RUN ;
DATA &newset ; SET &dset ;
                              IF . < &var <= &q1 THEN &newvar=1 ;
%DO i=2 %TO &n_q1 ;  %LET nnqq=&i + (-1) ;
                     %LET nnqq=%EVAL(&nnqq) ;
                      ELSE IF &&q&nnqq  < &var <= &&q&i THEN
&newvar=&i  ;
%END ;
                        %LET nnqq1=&nnqq+1 ;
                        %LET nnqq=%EVAL(&nnqq1) ;
                       ELSE IF &&q&nnqq < &var THEN &newvar=&n_q ;
RUN ;
%MEND ;
```

Beispiel:
Es soll die Variable Alter (age) in Quintile aufgeteilt werden: neue Variable agegr mit den Ausprägungen 1,2,3,4,5.

Die neue Variable kann dann z.B. weiter in (je nach Fragestellung) vier oder fünf 0-1-Dummy-Variablen aufgesplittet werden.

Programm:

```
OPTIONS MPRINT SOURCE2 ;

%INCLUDE 'D:\sasdat\mepi\quantil1.sas' ;

%quantils(lung.agetest,lung.agetest,q_out,outset,     ,age,5
,agegr,Quintiles)
PROC FREQ DATA=lung.agetest ;
```

```
TABLES agegr ;
RUN ;
```

Ausgabe:

```
Quintiles

Quantiles for the variables: age

Number of Quantiles: 5

Variable                Quantiles (Upper Cutpoints)

age     43.220     49.758     54.215     59.940     98.497
```

The FREQ Procedure

agegr	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	83580	20.00	83580	20.00
2	83514	19.99	167094	39.99
3	83619	20.01	250713	60.00
4	83560	20.00	334273	80.00
5	83565	20.00	417838	100.00

Datenanalyse - SAS-Prozeduren und statistische Grundlagen

Nested Fall-Kontroll-Studien:

- 1:1-Matching
- 1:N-Matching
- N:M-Matching

Statistische Grundlage für Nested Fall-Kontroll-Studien ist ein semi-parametrischer Ansatz für die Hazardrate:

$$\lambda(t,z(t)) = \lambda_0(t)\, r(z(t);\beta_0)$$

wobei

$r(z(t);\beta_0)$ das rel. Risiko für die Erkrankung bei einem Individuum mit Kovariablen z(t) zur Zeit t, und

$\lambda_0(t)$ das “Baseline” Risiko, wenn z=0, darstellen.

Das Partial Likelihood ist formal das gleiche wie ein bedingtes logistisches Likelihood von gematchten Fall-Kontroll-Studien.

Bedingte Logistische Regression

Gematchte Fall-Kontrollstudien in der Epidemiologie produzieren in der Regel hoch stratifizierte Datenstrukturen. Meist werden den Fällen Kontrollen auf der Basis von Variablen zugeordnet, die auch als potentielle Confounder angesehen werden können, wie etwa Alter, Geschlecht, Rauchen oder Alkoholkonsum. Die geeignete Analyseform, eine logistische Regression für diesen Datentyp durchzuführen, wird bedingte logistische Regression genannt. Sie berücksichtigt die Stratifizierung indem die Maximum-Likelihood-Schätzung auf einem bedingten Likelihood aufbaut.
Bei 1:1 gematchten Daten kann bei entsprechender Aufbereitung der Daten die Prozedur LOGISTIC verwendet werden. Der Datenvektor der Kovariablen muß dabei die Differenzen

zwischen Fällen und Kontrollen enthalten, und es muß die Option NOINT verwendet werden, welche die Schätzung des konstanten Koeffizienten (Achsenabschnitt) unterdrückt.
Einfacher ist es allerdings die Prozedur PHREG zu verwenden, die es erlaubt für beliebig gematchte Daten (1:1, 1:N oder N:M) eine bedingte logistische Regression durchzuführen.

Prozeduren:

LOGISTIC (nur 1:1 Matching)
Besonderheit: Daten müssen aufbereitet werden, damit der Datenvektor der Kovariablen die Differenzen zwischen Fällen und Kontrollen enthält.

PHREG (beliebige, N:M gematchte Daten)
Wir gehen davon aus, daß die Datenmatrix die folgende Struktur aufweist:

Datenstruktur:

Matching Variable(n) (Stratifizierung): z.B Altersgruppen ag
Fälle / Kontrollen-Indikator: z.B. 0-1-Variable `caco`, 1=Fälle
künstliche Zeitvariable: z.B. `time=2-caco` ;
Kovariablen

Falls die Zeitvariable nicht bereits schon in der Datenmatrix enthalten ist, kann sie in einem vorangehenden DATA-Step neu hinzugefügt werden. Der Wert der Dummy-Zeitvariablen muß für Kontrollen größer sein als der für Fälle.

Beispiel-SAS-Programm (N:M Matching) mit vorgeschaltetem Data-Step für die Dummy-Zeitvariable:
Es soll eine n:m, nach dem Alter (Variable `ag`) gematchte bedingte logistische Regression durchgeführt werden. Fälle (Wert: 1) und Kontrollen (Wert: 0) sind in der Variablen `caco` kodiert. Wenn der Fälle-Kontrollen-Indikator anders kodiert ist muß dieser Programmteil entsprechend modifiziert werden. Ins Programm sollen diverse Kovariablen einbezogen werden. Die Option `TIES=DISCRETE` muß nur für den Fall eines n:m Matching gewählt werden. Für ein 1:1 Matching reicht die Standardeinstellung `TIES=BRESLOW`.

```
DATA neu ;
SET daten ;
time=2 - caco ;
RUN ;
PROC PHREG DATA=neu ;
MODEL time*caco(0)= covariables / TIES=DISCRETE ;
STRATA ag ;
RUN ;
```

SAS-Makro:

```
%MACRO cond_LR(dset,newset,time,caco,cens,covar,nm,strata);
DATA &newset ;
SET &dset(KEEP=&caco &covar &strata);
&time=2 - &caco ;
RUN ;
PROC PHREG DATA=&newset ;
MODEL &time*&caco(&cens)= &covar
%IF %LENGTH(&nm)>0 %THEN / TIES=DISCRETE ; ;
%IF %LENGTH(&strata)>0 %THEN STRATA &strata ; ;
RUN ;
%MEND ;
```

Makro-Parameter:

dset	zugrundeliegende Datenmatrix
newset	Datenmatrix für die Analyse
time	Name für die neue „Dummy“-Zeitvariable
caco	Variable in der Fälle und Kontrollen kodiert sind
cens	Wert der Variable caco, der für Kontrolle steht
covar	Kovariablenliste
nm	Schlüsselwort für n:m Matching (jedes beliebige Zeichen)
strata	Stratifizierungsvariablen(-liste)

Case-Cohort-Design:

Beim Case-Cohort-Design erhält man Punktschätzer für das relative Risiko wie bei einer normalen Kohortenstudie angewandt auf die Case-Cohort-Subkohorte. Probleme gibt es allerdings beim Varianzschätzer. Dieser muß adjustiert werden, weil anstelle der Gesamtkohorte nur eine Teilstichprobe verwendet wird.

Auch hier dient für die Erkrankungsraten ein multiplikativer semi-parametrischer Ansatz als Grundlage

$$\lambda(t,z(t)) = \lambda_0(t)\, r(z(t);\beta_0),$$

was mit dem Pseudo-Likelihood Verfahren von Prentice (1986) (entspricht Partial-Likelihood bei Full-Cohort-Analyse) und einem speziellen Varianzschätzer (Self und Prentice 1988) analysiert werden kann. Es sind zwei Programmschritte erforderlich.

Programmschritte (Case-Cohort-Design)

1. Datenvorbereitung:

Fälle, die nicht zur Subkohorte gehören, sollten dort mit Status 0 hinzugefügt werden:

SAS-Programm:

```
DATA neu ; SET daten (KEEP ....) ;
IF status=1 THEN DO; dummy=-100; OUTPUT;
IF subcoh=1 THEN DO; dummy=0; status=0 ;
                     OUTPUT;
                END ;
```

```
                    END;
RUN ;
```

2. Programmschritt Case-Cohort-Datenanalyse mit Hilfe von PHREG:

Wegen der Varianzüberschätzung muß die Varianz extra über SAS/IML bestimmt werden (Matrixansatz von Therneau and Li (1998). Ansonsten wird wieder die Prozedur PHREG verwendet.

Makro:

```
%MACRO casecoho(dset, time, status, statval, covars, dummy,
                outset,  dfbeta_cv, idsubcoh) ;
PROC PHREG DATA=&dset ;
MODEL &time * &status(&statval) = &covars ;
OFFSET &dummy ;
OUTPUT OUT=&outset dfbeta=&dfbeta_cv ;
ID &idsubcoh ;
RUN ;
DATA xxxddd3 ;
SET &outset ; IF &idsubcoh=1 ;
RUN ;
PROC IML ; USE xxxddd3;
READ ALL VAR { &dfbeta_cv } INTO d ;
var=d' * d ;
RUN ;
PRINT , var ;
RUN ;
PROC DATASETS ; DELETE xxxddd3 ;
RUN ;
%MEND ;
```

Makro-Parameter

`dset`	aufbereitete Datenmatrix
`time`	Zeitvariable
`status`	Statusvariable (Fälle / Kontrollen)
`statval`	Wert der Statusvariablen in dem Kontrollen kodiert sind
`covars`	Kovariablenliste
`dummy`	Name der Dummyvariablen
`outset`	Name der Ausgabedatei mit den Koeffizienten dfbeta
`dfbeta_cv`	Namensliste der DFBeta-Koeffizienen
`idsubcoh`	Name der ID-Variable für die Subkohorte

Beispiel:

In einer Case-Cohort-Studie soll der Einfluß von Obst- und Gemüseverzehr sowie körperliche Aktivität und Rauchgewohnheiten (Variablen: `fruits, veget, physact, smoke`) auf die Entstehung einer bestimmten Krebsart (Variable `cancer`) überprüft werden. Die Entwicklung von Krebs wurde mit 1 kodiert. Als Zeitvariable wird das Alter (`age`) verwendet. Die Elemente der Subkohorte sind in der Variablen `subcoh` mit 1 verschlüsselt.
Für eine Anwendung des Makros müssen der oben genannte DATA-Step und das Makro mit-einander kombiniert werden.

Programm:

```
DATA neu ; SET daten (KEEP ....) ;
IF status=1 THEN DO; dummy=-100; OUTPUT;
IF subcoh=1 THEN DO; dummy=0; status=0 ;
                     OUTPUT;
                END ;
                END;
RUN ;
%casecoho(neu, age, cancer, 0, fruits veget physact smoke, dummy,
        out1, df_fruit df_veget df_physa df_smoke, subcoh)
```

Literatur:

1. Breslow, N.E. (1996): Statistics in Epidemiology: the case-control study. *J. Am. Statist. Assoc*. 91, 14-28
2. Breslow, N.E. / Day, N.E.: *Statistical Methods in Cancer Research I. The analysis of case-control studies*. Lyon, IARC 1980
3. Liddell, F.D.K., McDonald, J.C. and Thomas, D.C. (1977): Methods of cohort analysis: appraisal by application to asbestos mining /with discussion). *J. Roy. Statist. Soc. A*, 140, 469-491
4. Prentice, R.L. (1986): A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 73, 1-11
5. Self, P. and Prentice, R.L. (1988): Asymptotic distribution theory and efficiency results for case-cohort studies. *Annals of Statistics* 16, 64-81
6. Therneau, T.M. and Li, H. (1998): Computing the cox model for case cohort designs. Technical Report No. 62, Mayo Clinic Rochester, Minnesota
7. SAS/STAT User's Guide Volume 1 und 2. SAS Institute, Cary N.C.