AUTOMATED ANALYSIS OF CASE-CONTROL DATA USING THE SAS SYSTEM

L.W. Pickle, National Cancer Institute
E. J. Martin, ORI, Inc.

The analysis of data from a large case-control epidemiologic study can be very time-consuming. After the original data have been checked for completeness and accuracy, many different cross-tabulations are required to examine the marginal effects of potential covariates on the relative risks, measures of the association between exposures and the disease. Crude odds ratios (also called cross-product ratios or approximate relative risks) computed from this series of 2x2 tables can suggest covariates to be included as confounders and effect modifiers[1] in a more comprehensive analysis. Often this will consist of a maximum likelihood or least squares method of parameter estimation using the logistic or log-linear model. For the purpose of obtaining a best-fitting model, many subsets of covariates may need to be examined and their relative fit to the data evaluated, resulting in a minimal set of covariates that adequately "explains" the data and provides estimates of the desired relative risks.

We have written sub-routines for use with the SAS[2] system of programs to eliminate the hand calculation and file manipulation required for this type of analysis. Unless a program is available that was specifically written to perform the aforementioned calculations, the odds ratios must be hand-computed from cross-tabulations produced by a general program such as SPSS[3] or SAS. Our routines automatically produce point and interval estimates of the crude odds ratios[4,5] along with the cross-tabulations, and also eliminate the necessity of preparing an input file for the programs used for subsequent analysis.

Program Description

Figure 1 outlines the system we have used. In addition to printing the stratified counts and odds ratios as requested, our current version writes out a disk file of counts and exposure, case-control, and stratifying variables suitable for input to a program that performs a logistic analysis of case-control data, including confounders and effect modifiers as described recently by Prentice.[6] The control cards for this program, FLOGREG,[7] supply the number and names of the confounders and effect modifiers to be used for the analysis. By modifying a single line, output suitable for other programs may be produced. FLOGREG, however, writes its final parameter estimates on disk, permitting us to evaluate the goodness-of-fit of our model by a new program in our system.

Figure 2 is a listing of the user-supplied sub-routines required to implement this system. The input to the program, described in line 3, consists of one record per individual, presumably validated by a previous program. Any necessary recodings or other categorizations may be inserted between lines 3 and 5, following the general SAS assignment statement rules. The only restrictions for the variables are the following:

$$\text{CASECTL} = \begin{cases} 1 & \text{for a case} \\ 0 & \text{for a control,} \end{cases}$$

$$\text{and EXP} = \begin{cases} 1 & \text{if the individual had the exposure of interest,} \\ 0 & \text{otherwise.} \end{cases}$$

Any stratifying variables (confounders and/or effect modifiers) may be included; continuous variables are allowed although in most situations categorized variables will be preferred. Once the program is modified for use with a particular data set, a change in the set of covariates included requires only the alterations of lines 6 and 31. If, for the initial models, there are so many covariates included in the model that printing the individual cross-tabulations is impractical, the DATA=_NULL_ option may be added to line 42.

Example

Figure 3 shows the program set-up and resulting output from a portion of the analysis of a case-control death certificate study of pancreatic cancer in Louisiana.[8]

Information was abstracted from death certificates of 876 persons who died of pancreatic cancer from 1960-1975 in selected Louisiana parishes. An equal number of control certificates was selected, matched to the cases by age, sex, race, year of death, and parish of residence.

In addition to the usual information on occupation, industry, and residence abstracted from the death certificates, the likelihood of Acadian ancestry was estimated based on family names and place of birth. Results of the logistic analysis shown here indicate an approximate two-fold elevation in the odds ratio among younger whites of "probable" Acadian ancestry.

Discussion

We have presented sub-routines for use with the widely available SAS system that will eliminate hand calculations required for the analysis of case-control data. This not only combines several steps of the analysis into one program, but also automatically provides the user with data required to

examine the fits of the proposed models and to support the model that is finally chosen. The flexible SAS system requires only a one-line change, for example, to modify the output file for other analytic programs or to produce different interval estimates of the odds ratios. More complete program documentation is available from the authors.

References

1. Miettinen OS: Confounding and effect modification. Am J Epidemiol 100:350-353, 1974.
2. Barr AJ, Goodnight JG, Sall JP, et al: A user's guide to SAS 76. Raleigh, NC, SAS Institute, Inc., 1976.
3. Nie NH, Hull CH, Jenkins JG, et al: SPSS: Statistical package for the social sciences (2nd ed.), New York, McGraw-Hill, 1975.
4. Woolf B: On estimating the relation between blood group and disease. Annals of Human Genetics 19:251-253, 1955.
5. Haldane JBS: The estimation and significance of the logarithm of a ratio of frequencies. Annals of Human Genetics 20:309-311, 1956.
6. Prentice R: Use of the logistic model in retrospective studies. Biometrics 32:599-606, 1976.
7. Maguire M: A computer program for the logistic analysis of frequency data (FLOGREG) Personal communication.
8. Pickle LW, Gottlieb MS: Pancreatic cancer mortality in Louisiana. Am J Public Health, In press

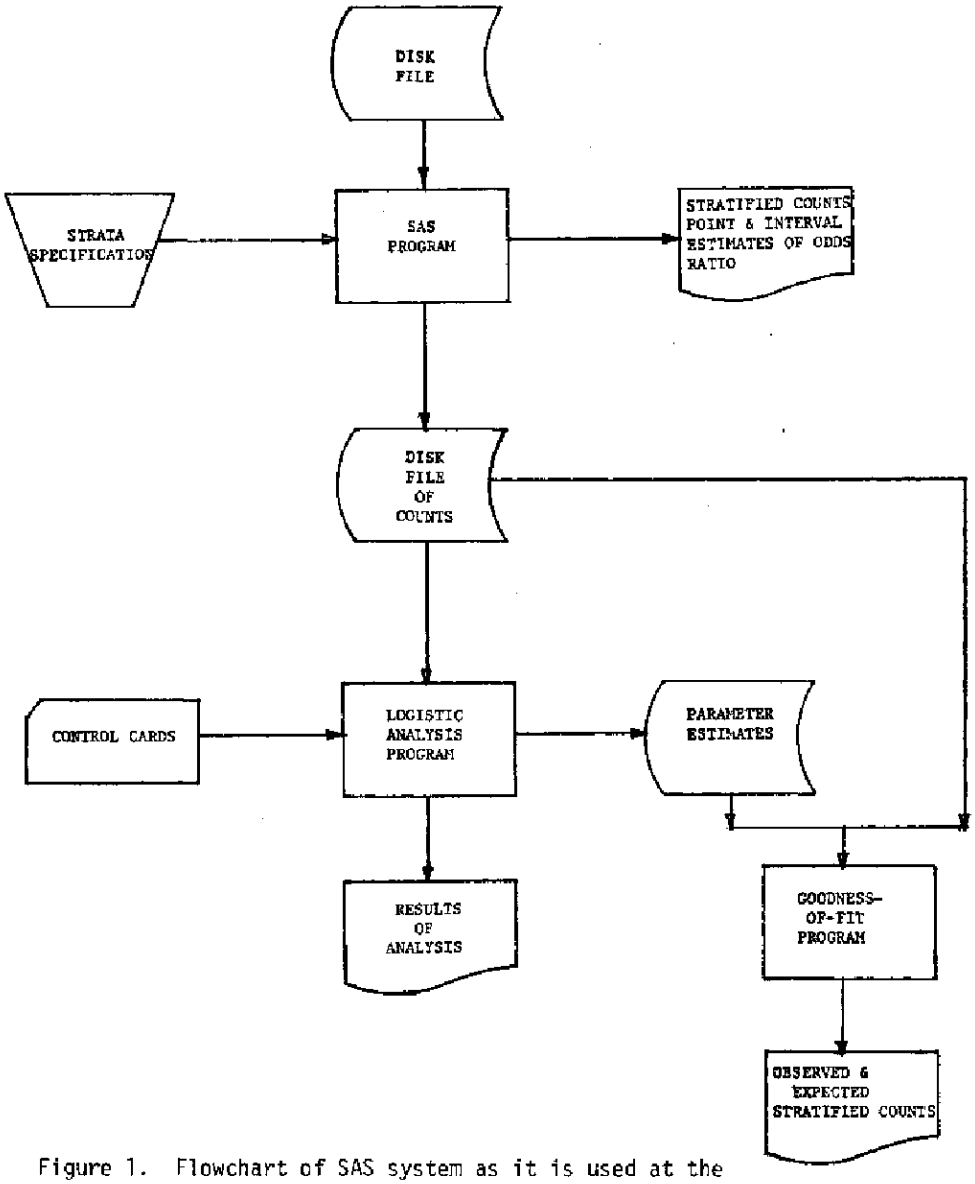

Figure 1. Flowchart of SAS system as it is used at the National Cancer Institute.

```
DATA;
INFILE INDATA ;
INPUT  CASECTL 5 DXAGE 35-36 SEXCODE 43 RACE 44 ANCES 19;
    IF RACE=1; EXP=0; IF ANCES<4 THEN EXP=1;
PROC FREQ;
    TABLES AGE*SEXCODE*EXP*CASECTL
          /NOPRINT SPARSE OUT=XTABS (DROP=PERCENT);
DATA;
    SET XTABS  END=EOJ;
  RETAIN  T 2.0     ADJ 0.5;
  RETAIN  A  B  C  D;
    IF COUNT ^= 0  THEN LINK PUT_FLOG;
    IF EOJ  THEN LINK END_RTN;
    IF CASECTL = 0 & EXP = 0  THEN D = COUNT;
    IF CASECTL = 0 & EXP = 1  THEN C = COUNT;
    IF CASECTL = 1 & EXP = 0  THEN B = COUNT;
    IF CASECTL = 1 & EXP = 1  THEN LINK ODSRATIO;
    RETURN;
  ODSRATIO:
    A = COUNT;
    REL_RISK = (A+ADJ)*(D+ADJ) / ((B+ADJ)*(C+ADJ));
    SE = SQRT(1/(A+ADJ) + 1/(B+ADJ) + 1/(C+ADJ) + 1/(D+ADJ));
    LOWER_RR = REL_RISK * EXP(-T*SE);
    UPPER_RR = REL_RISK * EXP (T*SE);
   DROP  ADJ T COUNT CASECTL EXP SE VARLINES;
    FORMAT REL_RISK LOWER_RR UPPER_RR 7.2;
    OUTPUT;
    RETURN;
  PUT_FLOG:
    FILE FLOGDATA;
    PUT COUNT EXP CASECTL AGE SEXCODE AGE;
    VARLINES + 1;
    RETURN;
  END_RTN:
    FILE PRINT;
    PUT / 'A=EXPOSED CASES, B=UNEXPOSED CASES, C=EXPOSED CONTROLS, D=UNEXPOSED
CONTROLS';
      PUT / 'CELL ADJUSTMENT = ' ADJ;
      PUT / 'CRITICAL VALUE FOR CONF. INTERVALS = ' T;
      PUT / 'NUMBER OF VARIABLE LINES FOR FLOGREG (N) = ' VARLINES;
    RETURN;
PROC PRINT  DOUBLE;  ID AGE;
TITLE1 STRATIFIED RISKS FOR PANCREATIC CANCER DUE TO ACADIAN ANCESTRY;
TITLE2 LOUISIANA DEATH CERTIFICATE STUDY - WHITES ONLY;
```

Figure 2. Listing of SAS subroutine, set up for Louisiana example.

(a)

S T A T I S T I C A L A N A L Y S I S S Y S T E M

A=EXPOSED CASES, B=UNEXPOSED CASES, C=EXPOSED CONTROLS, D=UNEXPOSED CONTROLS

CELL ADJUSTMENT = 0.5

CRITICAL VALUE FOR CONF. INTERVALS = 2

NUMBER OF VARIABLE LINES FOR FLOGREG (N) = 16

STRATIFIED RISKS FOR PANCREATIC CANCER DUE TO ACADIAN ANCESTRY
LOUISIANA DEATH CERTIFICATE STUDY - WHITES ONLY

OBS	AGE	SEXCODE	A	B	C	D	REL_RISK	LOWER_RR	UPPER_RR
1	0	0	8	91	4	107	2.22	0.67	7.37
2	0	1	17	178	9	204	2.11	0.92	4.85
3	1	0	8	152	11	137	0.67	0.26	1.70
4	1	1	11	193	16	170	0.61	0.28	1.36

(b)

LA. PANCREAS DATA - ACADIAN ANCESTRY ANALYSIS

ITERATION NUMBER 3

NUMBER OF SUBJECTS 1316

VARIABLE	B	STD ERR	T-STAT
1 CONSTANT	-3.25503	0.32557	-9.9980
2 LN OF OR	0.79845	0.35217	2.2672
3 AGE65+	0.75585	0.34808	2.1715
4 MALE	0.11975	0.23575	0.5080
5 E-AGE65+	-1.26707	0.46926	-2.7001

THE LOG LIKELIHOOD = -308.3218
THE VARIANCE ESTIMATE IS 1.0038

THE ESTIMATE OF ODDS RATIO= 2.22210
THE MAXIMUM LIKELIHOOD 95% CONFIDENCE LIMITS=(1.11572, 4.42559)

INVERSE OF THE INFORMATION MATRIX

VARIABLE	1	2	3	4	5
1	0.10560				
2	-0.07973	0.12356			
3	-0.08364	0.07984	0.12071		
4	-0.03775	-0.00019	0.00555	0.05537	
5	0.07993	-0.12356	-0.12016	-0.00010	0.21938

CONVERGENCE REACHED

Figure 3. Results of a portion of the Acadian ancestry analysis of a series of Louisiana death certificates; (a) from SAS program, (b) from a logistic analysis program, and (c) from a goodness-of-fit program.

Figure 3 (continued) (c)

AGE65+ MALE AGE65+

LA. PANCREAS DATA - ACADIAN ANCESTRY ANALYSIS

				EXPOSED CASES				EXPOSED CONTROLS	
			P1	N1	EXPECTED	OBSERVED	P2	N2	EXPECTED	OBSERVED
0	0	0	0.07896	99	7.817	8	0.03715	111	4.123	4
0	1	0	0.08812	195	17.183	17	0.04168	213	8.877	9
1	0	1	0.04890	160	7.824	8	0.07592	148	11.236	11
1	1	1	0.05478	204	11.176	11	0.08475	186	15.764	16
TOTALS				658	44.000	44		658	40.000	40

VARIABLE LINES = 16