WEB APPENDIX

Lagging exposure information in cumulative exposure-response analyses

Assume that we have data set, *DS*, which is an analytical data structure for a matched case-control analysis as in our empirical example. Up to 40 controls are sampled for each case. The data set consists of one record for the case and associated controls. The variable *age_rs* specifies the age at which the risk set is enumerated (the age of case occurrence for the case or control selection). The variable *_ntot* specifies the total number of people in a risk set. The exposure histories are represented by the variables, *z1*, *z2*, … denoting the exposure accrued at each year of age through age 86.

Estimation of the 'best fitting' lag interval assuming this is a fixed constant.

The SAS code below can be used to fit a linear excess rate ratio model estimating the cumulative exposure-response trend, b_1, and the 'best' fitting lag value, *lag*. A typical lag may be implemented as a time-varying binary exposure weighting function (i.e., a step function); since this implies a discontinuous function that creates estimation problems, we facilitate estimation by means of a continuous, S-shaped exposure weighting function that closely approximates a step function. The inflection point for the function, *lag*, is directly estimated from the data, while the shape parameter for the weight function is fixed at a value that implies a steep S-shaped curve (in this example, 25). The conditional likelihood contribution for each risk set is given in (1).

```
proc nlmixed data=DS;
  parms b1=1, lag=7;
  array z{87,41} z1-z3567;
  sum=0;
   do i=_ntot to 1 by -1;
    age_rs=z{87,i};
    cumexp=0;
    do j=1 to age_rs ;
     tse=age_rs-j+0.5;
     wt=((tse/lag)**25)/(((tse/lag)**25)+(tse/lag));
     if j <=86 then cumexp=cumexp+(z{j,i}*wt/100);
    end;
    phi=1+cumexp*b1;
    sum=sum+phi;
   end;
   cc=1;
 model cc~binary(phi/sum);run;
```

The ‘parms’ statement tells SAS that the cumulative exposure effect and inflection point for the S-shaped exposure weighting function, b1 and lag, respectively, are to be estimated and sets initial values for these parameters. The array z{87,41} specifies the exposure information for each member of the risk set (in this example 87 explanatory variables for up to 41 members of each risk set. The variable ‘sum’ is used in the calculation of the likelihood contribution for each risk set. The ‘do i’ loop indexes over the _ntot members of each risk set. The ‘do j’ loop indexes over the ‘age_rs’ years of observation for members of the risk set. For each year of observation, the variable ‘tse’ represents the time-since-exposure (i.e., the time interval between the index ‘j’ and the risk set age) and the variable ‘wt’ is a time-varying exposure weighting function that varies with ‘tse’ and is bounded by 0, 1. The variable ‘cumexp’ is the cumulative exposure for person i accrued up to age ‘age_rs’ weighted by the time-varying exposure weighting function ‘wt’. The variable ‘phi’ specifies the model for the rate ratio; and, the variable ‘sum’ integrates the rate ratios over the members of a risk set. Each risk set

includes a case failure, hence the variable cc=1, and the 'model' statement specifies that we use binomial model with probability (phi/sum). This example concerns a 1:m matched case control study. We have previously described how to fit models to risk set data with multiple cases in the case-control set using SAS (1).

Estimation of mode and coefficient of variation of an assumed lognormal population distribution of induction/latency periods.

A population distribution of induction/latency periods may be estimated as a time-varying exposure weighting function which conforms to the cumulative density function for the underlying distribution of induction periods. We facilitate estimation by positing that the underlying distribution is lognormal. An exposure weighting function is defined as the cumulative density function for the lognormal distribution specified by two parameters, *lnmode* and *lncv*, which are directly estimated from the data and represent the natural log of the mode and coefficient of variation of the distribution, respectively.

```
proc nlmixed data=DS;
  array z{80,5} z1-z400;
  parms b1=1, lncv=-1.2, lnmode=1.6;
  LN_B1=log( (exp(lncv)*exp(lncv))+1 ); LN_A1=lnmode+LN_B1;
  sum=0;
    do i=_ntot to 1 by -1;
      cumexp=0;
      do j=1 to floor(_rstime);
        tse=_rstime-j + 0.5;
        wt= cdf('NORMAL',(log(tse)-LN_A1)/sqrt(LN_B1));
        if j <= 80 then cumexp=cumexp+z{j,i}*wt/100;
      end;
      phi=1+cumexp*b1; sum=sum+phi;
    end;
  L=phi/sum; cc=1;
  model cc~general(log(L)); run;
```

Note that the induction period, *L*, is said to follow a lognormal distribution if log(*L*) is normally distributed (mu, sigma^2). The mode of this lognormal variate is given by exp(mu-sigma^2) and coefficient of variation=sqrt(exp(sigma^2)-1). We derive the terms needed for the exposure weight function, wt, from the free parameters, lnmode and lncv. Estimation of these parameters may be facilitated by first estimating the mode of the distribution of induction times by fitting a model in which the induction/latency interval is assumed to be a fixed constant (as described above). The natural log of this value can then be used as a starting value for parameter *lnmode*.

Log linear models

Suppose that the observed outcomes were generated by an underlying log-linear model. We can define an expression of the expectation of the excess rate ratio at time *t*, given an exposure increment at time *j*, allowing for a population distribution of induction/latency intervals, *L*, as follows

E[$\varphi(t,d_{ij})$]

$= (\exp(\beta \times d_{ij}) - 1)E(I_L(t-j)) = (\exp(\beta \times d_{ij}) - 1)pr(L \le t-j) = (\exp(\beta \times d_{ij}) - 1)F_L(t-j)$,

where $f_L(l)$ is the population distribution of *L* and $F_L(u) = \int_0^u f_L(l)dl$.

The impact of a series of exposure increments under a log-linear model is the product of the rate ratios (i.e., exposure effects are multiplicative), implying that the expected rate ratio at time *t*, given a protracted history of exposures up time *t* allowing for induction/latency interval, *L,* is

$RR(t, D_i(t)) = \prod_{j=0}^{t} 1 + \varphi(t, d_{ij}) = \prod_{j=0}^{t} 1 + \left(\exp\left[\beta \times d_{ij}\right] - 1\right) F_L(t - j)$. The code below provides an example of SAS code to fit such a model by maximum likelihood methods. The exposure weighting function, wt, is the cumulative density function for the lognormal distribution specified by two parameters, *lnmode* and *lncv*.

```
proc nlmixed data=DS;
  array z{80,5} z1-z400;
  parms b1=1, lncv=-1.2, lnmode=1.6;
  cv=exp(lncv);
  LN_B1=log((CV*CV)+1); LN_A1=lnmode+LN_B1;
  cc=1; sum=0;
   do i=_ntot to 1 by -1;
     exp_index=floor(_rstime);
     aexp=0;phi=1;
     do j=1 to exp_index;
       tse=_rstime-j+0.5;
       wt=cdf('NORMAL',(log(tse)-LN_A1)/sqrt(LN_B1));
       if j <=80 then aexp=z{j,i}/100;
       phi=phi*(1+(exp(aexp*b{1})-1)*wt);
     end;
   sum=sum+phi;
   end;
 L=phi/sum;
 model cc~general(log(L)); run;
```

REFERENCES

1. Langholz B, Richardson DB. Fitting general relative risk models for survival time and matched case-control analysis. American journal of epidemiology 2010;171(3):377-83.