

Deriving relative risks from aggregate data. 1. Theory

THOR NORSTRÖM

From the Swedish Institute for Social Research, University of Stockholm, S-10691 Stockholm, Sweden

SUMMARY Sociological macro analyses of the association between risk factors and mortality can be seen as a valuable supplement to epidemiological micro studies. However, sociologists and epidemiologists typically employ different measures of association and this hampers strict comparisons of findings. This study presents a synthetic approach relying on both micro and macro data. In Part 1, the mathematical relations between the relative risk and the attributable fraction on the one hand, and the regression coefficient on the other are derived in order to make cross level comparisons possible. Part 2 provides an empirical illustration of the approach.

The traditional method in epidemiology for investigating the relationship between a suspected aetiological agent and a specific cause of death is the observational study, either prospective or retrospective. A sociological approach which is becoming increasingly widespread is to estimate this relationship on the basis of aggregate ecological or time series data. With these differences in approach there come also differences in measures of association between risk factors and mortality, a state of affairs that tends to obstruct the cumulation of evidence. This is a loss of research efficiency, because the findings from the two traditions actually supplement each other. The present study outlines a synthetic approach in which both micro data (epidemiological) and macro data (sociological) are relied upon, the aim being to integrate the findings from the two. This requires an understanding of how the micro and macro measures relate to each other, which is treated in Part 1 of this study. In Part 2, the approach is empirically illustrated by micro and macro analyses of the association between unemployment and suicide.

Measures of association in epidemiology and sociology

In the epidemiological literature, the common measure of association between a risk factor and a specific cause of death is the relative risk (RR). This expresses how many times higher (or lower) the mortality rate is among the exposed, as compared with the unexposed:

$$RR = \frac{\text{Mortality rate in the exposed category}}{\text{Mortality rate in the unexposed category}} \quad (1)$$

For some purposes, eg, public health planning, it is desirable to go beyond the relative risk and to estimate the overall importance of a risk factor for a specific cause of death. In this context the attributable fraction (AF) is useful. There are several formulas for its calculation, but the following is often applied.¹

$$AF = \frac{\theta (RR - 1)}{\theta (RR - 1) + 1} \quad (2)$$

where θ is the proportion of the population exposed to the risk factor, and RR is the relative risk. The attributable fraction expresses the relative importance of the aetiological agent for the mortality at issue. Or put in another way: AF* 100 indicates by how many per cent mortality is expected to decrease if the exposure were eliminated.

For those many situations where experimental studies are not feasible, the relative risk is based on observational studies, prospective or retrospective. There are two main problems of such designs. The first one is the difficulty in obtaining adequate exposure data. Typically, the exposure measure refers to one single point of time prior to the follow up period. To keep track of changes in exposure requires designs that are expensive as well as time consuming (see Schlesselman² for a more detailed discussion). The second major problem is selection bias. This is found when correlates of the outcome variable (eg, mortality) are associated with the probability of being included in the study population.³ As an example, such bias may be present in studies of the association between unemployment and mortality, since unemployed persons as a group are probably less healthy at the outset.

To substantiate the association between a potential risk factor and a specific cause of death, a larger number of observational studies are typically relied upon. If the findings are consistent, one may have more confidence in causal inferences. There is still the disquieting possibility that all these estimates are plagued by a common source of bias, eg, selection bias. One avenue out of this dilemma is to broaden the empirical basis by including studies in which the potential bias is quite different. Such an opportunity is offered by time series studies of aggregate data.

By the same token, aggregate studies of the relationships between potential risk factors and mortality would become more powerful if the researchers more systematically related their macro findings to existing micro evidence. There is thus a need for cross level comparisons. However, as lamented by Rothman,³ an obstacle to the comparison of evidence from micro and macro inquiries is that different measures of association are used, and this permits only crude judgements of consistency. The next section describes a bridge between the micro and macro measures.

THE RELATION BETWEEN MICRO AND MACRO MEASURES

The common and preferable measure of association in time series analyses of aggregate data is the unstandardised regression coefficient (b). Given the simple model:

$$Y_t = a + bX_t + e_t \quad (3)$$

b expresses the change expected in Y when there is a one unit change in X (controlling for other determinants included in e). It is seldom recognised that there is a relation between the relative risk and the attributable fraction on the one hand, and the regression coefficient on the other.

Since the formulas for these relations do not appear either in epidemiological or sociological text books on methodology, their derivations are shown below, supplemented by a hypothetical example (to simplify, we assume no selection effects).

In this example (table), the proportion of the population which is exposed to the risk factor is 0.10 at time $t = 1$, and 0.11 at $t = 2$. The relative risk (RR) of the exposed category is 150/50, or 3.0. As noted above the regression coefficient (b) expresses the increase in the mortality rate associated with a one unit increase in the predictor. If we regard the risk factor (measured as per cent exposed) as the predictor (X) in a regression analysis, there is a one unit (ie, one per cent point) increase in X between t_1 and t_2 . Hence b is equal to the change in the mortality rate between t_1 and t_2 , ie, 61-60=1.0.

Linking with the example displayed in the table, the relation between RR and b is derived as follows: denote population size N, and the death intensity of the unexposed category P_0 . The change between t_1 and t_2 yields $N/100$ additional individuals in the exposed category which produces an increment of the number of deaths shown by expression (4):

$$\frac{RRP_0N}{100} - \frac{P_0N}{100} = \frac{P_0N(RR-1)}{100} \quad (4)$$

Expression (4) thus represents the number of deaths associated with a one unit increase in the predictor (X). To obtain b, (4) is converted into the metric of the dependent variable (mortality per 100 000 of the population), ie, it is multiplied by 100 000/N:

$$b = \frac{P_0N(RR-1)100\,000}{100N} = P_0(RR-1)1000 \quad (5)$$

(The correction factor is equal to 1000 only when the dependent variable is expressed as mortality per 100 000). Expression (5) is applied when we want to obtain the b expected from given values of RR and P_0 . We can illustrate this relation by using the hypothetical data on RR and P_0 in the table:

$$b = 0.0005(3-1)1000 = 1.0$$

When working the other way round, ie, when we want the RR associated with an observed b, it is more practicable to proceed in two steps, starting from an expression which yields the attributable fraction from aggregate data. This is derived as follows. Denote the

Table. Hypothetical example of mortality in exposed and unexposed persons at $t = 1$ and $t = 2$

Time	$t = 1$				$t = 2$			
	N	%	Mortality rate per 100 000	Deaths	N	%	Mortality rate per 100 000	Deaths
Exposed	100 000	10	150	150	110 000	11	150	165
Unexposed	900 000	90	50	450	890 000	89	50	445
Total	1 000 000	100	60	600	1 000 000	100	61	610

risk factor X , and let b be the parameter expression the effect of X on Y (the mortality rate). The mortality rate produced by X will then be equal to bX , and thus:

$$AF = \frac{bX}{Y} \quad (6)$$

We may compute the AF for each unit of observation (ie, each year if annual data are used), or for the average of a time period. From the figures in the table, we get for t_1 :

$$AF = \frac{1 \cdot 10}{60} = 0.167$$

and for t_2 :

$$AF = \frac{1 \cdot 11}{61} = 0.180$$

(It is easily verified that (2) and (6) give identical results, but note that θ is expressed as a fraction, while X is expressed in the units that are used in the regression model—here, per cent).

The second step is to insert the value of AF into (2) to obtain RR. Again using the hypothetical data in the table, and rearranging (2), we have for t_1 :

$$RR = 1 + \frac{0.167}{(1 - 0.167) 0.1} = 3.0$$

(For t_2 we of course get an identical RR).

In the hypothetical example, there are only two points of time. In practice, b is of course estimated from longer time series. If X is the appropriate measure of exposure, it provides (after transformation into a fraction) the value of θ . Otherwise, θ is obtained from external sources.

Discussion

The final section in Part 2 of this research discusses some of the pros and cons of deriving relative risks and attributable fractions from aggregate data as outlined above. However, two characteristics of the measures as computed in this way should be considered here. First, they are expected to express the exogenous impact of the risk factor, net of selection effects. The rationale of this assertion is that the temporal variation of the risk factor (the effect of which is

measured by the regression coefficient) should be due to changes in exposure, and not to changes in self selection. Analytically, this can be shown as follows. Consider the following micro model:

$$Y_{it} = C_i + bX_{it} + e_{it} \quad (7)$$

where Y_{it} is the response of individual i on the outcome variable at time t ; C_i is the unmeasured time invariant attributes which are correlated with the outcome variable; and e_{it} is a random error term. The potential bias in the estimation of b is due to the correlation between Y_{it} and X_{it} that may arise because of self selection. Now consider equation (7) in aggregate form:

$$Y_t = C + bX_t + e_t \quad (8)$$

It is immediately seen that the aggregation makes C_i a constant which by definition is uncorrelated with X_t , and thus the source of bias is eliminated.

Another characteristic of the macro measures is that they include possible indirect effects of the risk factor on the unexposed category. As an example of indirect effects, we could mention the contracting of lung cancer due to other people's smoking. (In our hypothetical example, the presence of an indirect effect would have been felt in an increased mortality rate from t_1 to t_2 among the unexposed, which would have affected the regression coefficient, and hence the other measures). In some cases, it is conceivable that the total of such indirect effects is substantial, simply because of the large size of the unexposed category.

This research was funded by the Swedish Commission for Social Research. I am indebted to Anders Ahlbom, Gösta Carlsson, Walter Korpi, Joakim Palme, Ole Jørgen Skog, and an anonymous referee for their helpful comments on an earlier version of this paper. I thank Karen Leander for correcting my English.

References

- ¹ Lilienfeld AM, Lilienfeld DA. *Foundations of epidemiology*. New York: Oxford University Press, 1980.
- ² Schlesselman JJ. *Case-control studies*. New York: Oxford University Press, 1982.
- ³ Rothman K. *Modern epidemiology*. Boston/Toronto: Little, Brown and Co, 1986.