Individual-based and Group-based Occupational Exposure Assessment: Some Equations to Evaluate Different Strategies

ERIK TIELEMANS,* LAWRENCE L. KUPPER†, HANS KROMHOUT,*‡ DICK HEEDERIK* and REMKO HOUBA*

*Environmental and Occupational Health Group, Department of Environmental Sciences, Wageningen Agricultural University, Wageningen, The Netherlands †Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, NC, U.S.A.

Basically, two strategies can be considered for the analysis of hazardous pollutants in the work environment: group-based and individual-based strategies. This paper provides existing and recently derived equations for both strategies describing the influence of several factors on attenuation and on the standard error of an estimated linear regression coefficient relating a continuous exposure variable and a continuous health outcome via a simple linear regression model. We applied these equations using exposure variability information from industry-wide surveys over the past decade in order to gain more insight into the effects of various sources of exposure variability on choices among different analysis strategies. In general, for the modeling scenario considered here, there is not a straightforward criterion for choosing an optimal analysis strategy. Researchers have to decide between individual-based strategies generating precise, though biased, estimates or group-based strategies generating less precise but essentially unbiased estimates. For most exposure variability scenarios evaluated, an individual-based strategy yielded substantial attenuation. It is the authors' contention that the choice between individual-based and group-based strategies should be based on validity, rather than on precision, of the estimated exposure-response coefficient.

© 1998 British Occupational Hygiene Society. Published by Elsevier Science Ltd.

INTRODUCTION

An important element of occupational epidemiological studies is the assessment of exposures to hazardous pollutants in the work environment. The concentration of an air pollutant can vary considerably over time and place and can be divided into exposure variability between workers and within-workers. Basically, two quantitative exposure assessment strategies can be distinguished. A pollutant can either be measured for each individual separately or on the level of so-called occupational groups. In an individual-based strategy, workers are measured repeatedly and an average exposure is calculated for each worker. The performance of individual-based strategies depends on the ratio of within-worker to between-worker variance components in exposure and the number of repeated measurements per worker. In the group-based strategy, subgroups of workers are constructed based on their job titles, tasks or other features of the work environment they share. Ideally, all workers are measured within these groups, and the group is the unit of analysis, not the individual, with group means used to quantify exposure and possibly response levels as well. For group-based analyses, statistical issues involve consideration of between-group, within-group, and within-worker exposure variability. Statistical theory on measurement error models has been described for individual-based strategies and is still a subject of debate for group-based strategies (Kromhout and Heederik, 1995). This paper provides equations describing the influence of several factors on attenuation and on the standard error of an estimated linear regression coefficient relating a continuous exposure variable (measured with error) and a continuous health outcome via a simple linear regression model.

EQUATIONS FOR ATTENUATION AND STANDARD ERROR OF ESTIMATED REGRESSION COEFFICIENT

The equations for attenuation and standard error of the estimated exposure-response coefficient hold

Downloaded from https://academic.oup.com/annweh/article-abstract/42/2/115/371032 by guest on 25 November 2018
under the assumption that any logtransformed exposure measurement \(Y_{gpi}\) satisfies the following model:

\[
Y_{gpi} = \ln(X_{gpi}) = \mu_i + \tau_g + \beta_g + c_{gpi} = \mu_{gpi} + c_{gpi},
\]

where \(\mu_{gpi} = \mu_i + \tau_g + \beta_g\), \(g = 1, 2, \ldots, G\), \(i = 1, 2, \ldots, k\), and \(j = 1, 2, \ldots, n\).

Here, \(G\) is the number of randomly chosen groups, \(k\) is the number of randomly chosen workers within a group, \(n\) is the number of repeated exposure measurements taken on each worker, and the total number \(N\) of sampled workers satisfies the relationship \(N = Gk\).

Also, \(\mu_i\) is the population mean exposure, \(\tau_g\) is the random group effect \([\tau_g \sim N(0, \sigma_{\tau g}^2)]\), \(\beta_g\) is the random effect for \(i\)-th worker in \(g\)-th group \([\beta_g \sim N(0, \sigma_{\beta g}^2)]\), and \(c_{gpi}\) is the random effect for \(j\)-th day measurement on \(i\)-th worker in \(g\)-th group \([c_{gpi} \sim N(0, \sigma_{c pi}^2)]\). The true health outcome model is assumed to be of the form:

\[
R_{gi} = \beta_i + \beta \mu_{gpi} + u_{gi},
\]

Here, \(R_{gi}\) is the continuous health response, \(u_{gi}\) is the residual error \([u_{gi} \sim N(0, \sigma_u^2)]\), \(\beta_i\) is the true exposure-response coefficient, and \(u_{gi}\), \(\tau_g\), \(\beta_g\) and \(c_{gpi}\) are mutually independent. Then, under the usual normal theory regression analysis assumptions, it follows that the unweighted least squares estimator of \(\beta_i\), when using an individual-based analysis is equal to

\[
\hat{\beta}_i = \frac{\sum_{g=1}^G \sum_{i=1}^k (Y_{gpi} - \bar{Y}_g) \bar{R}_{gpi}}{\sum_{g=1}^G \sum_{i=1}^k (Y_{gpi} - \bar{Y}_g)^2},
\]

where

\[
\bar{Y} = \frac{1}{Gk} \sum_{g=1}^G \sum_{i=1}^k Y_{gpi}, \quad \bar{R}_g = \frac{1}{n} \sum_{i=1}^n R_{gpi}.
\]

The expected value of \(\hat{\beta}_i\) can be written in the form

\[
E(\hat{\beta}_i) = \frac{\sigma_{\beta g}^2}{\sigma_{\beta g}^2 + \frac{\sigma_{c pi}^2}{n}} \beta_i + \frac{1}{1 + \frac{\sigma_u^2}{\sigma_{\beta g}^2}} \beta_i.
\]

where \(\kappa = \frac{\sigma_u^2}{\sigma_{\beta g}^2}\)

This is a well known equation for attenuation and applications can be found in both nutritional (Willett, 1990) and environmental studies (Brunekreef et al., 1987; Heederik et al., 1994). Assuming that each worker in group \(g\) is assigned that group’s mean exposure \(\bar{Y}_g\), then the least squares estimator of \(\beta_i\) using a group-based analysis is equal to

\[
\hat{\beta}^* = \frac{\sum_{g=1}^G \sum_{i=1}^k (\bar{Y}_g - \bar{Y}) \bar{R}_g}{\sum_{g=1}^G \sum_{i=1}^k (\bar{Y}_g - \bar{Y})^2} = \frac{\sum_{g=1}^G \sum_{i=1}^k (\bar{Y}_g - \bar{Y}) \bar{R}_g}{\sum_{g=1}^G \sum_{i=1}^k (\bar{Y}_g - \bar{Y})^2}.
\]

The expected value of this estimator \(\hat{\beta}^*\) can be written as (see also Appendix A)

\[
E(\hat{\beta}^*) = \frac{\sigma_{\beta g}^2}{\sigma_{\beta g}^2 + \frac{\sigma_{c pi}^2}{n}} \beta_i.
\]

When \(k = 1\), equation (6) reduces to equation (4) since \(\sigma_{\beta g}^2 + \frac{\sigma_{\beta g}^2}{n} = \sigma_{\beta g}^2\). For an individual-based analysis, the standard error of the estimator \(\hat{\beta}_i\) is equal to

\[
\text{SE}(\hat{\beta}_i) = \sqrt{\frac{1}{Gk} \left( \frac{\sigma_u^2}{\sigma_{\beta g}^2 + \frac{\sigma_{c pi}^2}{n}} + \frac{\sigma_u^2}{\sigma_u^2} \right)}.
\]

where \(G\) denotes the total number of workers. For a group-based analysis, the standard error of the estimator \(\hat{\beta}^*\) is equal to

\[
\text{SE}(\hat{\beta}^*) = \sqrt{\left( \frac{\sigma_u^2}{\sigma_{\beta g}^2 + \frac{\sigma_{c pi}^2}{n}} + \frac{\sigma_u^2}{\sigma_u^2} \right)}.
\]

where \(G\) denotes the number of groups. Equation (8) reduces to equation (7) when every worker comprises his or her own group.

**ILLUSTRATION WITH EMPIRICAL COMPONENTS OF VARIANCE**

As an illustration, some calculations with the above equations were performed. Linear regression analysis of a continuous health endpoint (i.e., lung function) on exposure was considered. The true exposure response coefficient \(\beta_i\) was set at the value of \(-0.10\) while the variance of the response variable \(\sigma_u^2\) was set at 0.15. These values are in accordance with those found in studies focused on the relation between lung function and dust exposure. The study population was assumed to contain \(N = 500\) workers in each scenario.

The number of groups in each group-based strategy was assumed to be 10, so that \(k = 50\) in each group. The number of repeated measurements per worker has been set at 1, since it is often not feasible to take...
multiple measurements for each worker if the study population is large. Values of estimated components of variance \( \hat{\sigma}_{w}^2, \hat{\sigma}_{w}^2 \) and \( \hat{\sigma}_{w}^2 \) of industry-wide surveys over the past decade (Kromhout et al., 1996) were substituted for \( \hat{\sigma}_{w}^2, \hat{\sigma}_{w}^2 \) in equations 4, 6, 7 and 8 in order to provide some information about typical magnitudes of expected values of the estimators \( \hat{\beta}_g^* \) and \( \hat{\beta}_i \) and corresponding standard errors of the regression coefficients. It can be seen from Table 1 that group-based strategies will result in larger standard errors compared to individual-based strategies. Rather ineffective groupings (i.e., groupings with a large ratio of within-group to between-group variance components) yield very large standard errors. In contrast, use of grouped exposure data substantially eliminated attenuation bias. If the variance ratio is equal to 1, an individual-based analysis results in an estimated regression coefficient which is only 50% of the true coefficient. For benzene exposure in the refinery industry (Spear et al., 1987), applying an individual-based analysis with only one measurement per worker yields an underestimation of 70% (i.e., \( 1 - \hat{\beta}_i/\hat{\beta}_i = 0.7 \)). If one wants to decrease this attenuation to around 10% (i.e., \( 1 - \hat{\beta}_i/\hat{\beta}_i = 0.1 \)), at least 21 repeated measurements per worker are needed. In a large scale epidemiological study involving about 500 workers, this would imply taking 10,500 samples. Such a measurement effort is often not feasible. Group-based strategies produced almost unbiased coefficients for ineffective groupings and for large variance ratios (\( \lambda \)).

Finally, as a crude indicator of the relative behavior of confidence intervals for individual-based and group-based strategies, consider the intervals \( E(\hat{\beta}_i) \pm 1.96[SE(\hat{\beta}_i)] \) and \( E(\hat{\beta}_g^*) \pm 1.96[SE(\hat{\beta}_g^*)] \). For the individual-based strategies, only 3 of 13 such intervals (namely, those for agents quartz, inhalable dust, and solvents) contain the value \( \beta_i = -0.10 \). In contrast, all 13 of the group-based intervals contain the value \( -0.10 \), implying a decided advantage to grouping for the modelling scenario considered in this paper.

**DISCUSSION**

Accurate assessment of exposure to hazardous pollutants is an important component of occupational epidemiology (Kromhout and Heederik, 1995; Heederik et al., 1991; Armstrong et al., 1992). To give researchers more insight in deciding which strategy to use when developing exposure estimates, equations for attenuation and standard error of the estimated slope in simple linear regression have been derived in this paper. According to the derived formulas, it was clear that attenuation of an exposure-response coefficient was very small when analyzing grouped exposure data. Even for very ineffective groupings, the percentage of underestimation was negligible. This finding is in accord with the theory that grouping data should lead to almost unbiased estimated coefficients. Since this type of strategy is primarily subject to Berkson type of measurement error (Armstrong, 1990; Spear et al., 1987).

Table 1. Expected values of the estimators \( \hat{\beta}_i \) and \( \hat{\beta}_g^* \), and corresponding standard errors for different industries and exposures; the true value of \( \beta_i \) is \(-0.10\).
Another assumption is that both within-worker and between-worker variability are constant over workers and groups. In addition, it must be emphasized that the equations are valid for the simple situation where the exposure variable is the only explanatory variable in the linear regression model. If other terms are included in the model, the derivations can become much more complicated (Armstrong, 1990; Kupper, 1984). In this paper, the health outcome is regressed on the log-transformed exposure data. In most studies, however, estimates of exposure are based on arithmetic means, which is a different exposure-response model. Although we are fully aware of the limitations of the equations used and the assumptions made, this study was presented in order to serve as a basis for development of a theoretical framework for exposure assessment in epidemiology.

REFERENCES


APPENDIX

In what follows, we briefly outline the derivations of equations 6 and 8. Equation 4 is obtained as a special case of equation 6, and equation 7 is obtained as a special case of equation 8, by setting \( k = 1 \) and noting that \( \sigma_{ww}^2 = \sigma_{wg}^2 + \sigma_{ww}^2 \) in that case.

Under models 1 and 2 in the main text, \( R_e \) and \( Y_e \) have a joint bivariate normal distribution with

\[
E(R_e) = \beta_0 + \beta_1 \mu, \\
E(Y_e) = \mu,
\]

\[
\text{Var}(R_e) = \beta_1 \left( \sigma_{wg}^2 + \frac{\sigma_{ww}^2}{k} \right) + \sigma_\epsilon^2,
\]

\[
\text{Var}(Y_e) = \sigma_{wg}^2 + \frac{\sigma_{ww}^2}{k},
\]

and

\[
\text{Cov}(R_e, Y_e) = \left( \sigma_{wg}^2 + \frac{\sigma_{ww}^2}{k} \right) \beta_1.
\]

Hence, from page 3 of Fuller (1987),

\[
E(\hat{\beta}_e^*) = \frac{\text{Cov}(R_e, Y_e)}{\text{Var}(Y_e)},
\]

which is equation 6.

Now, since the G pairs \( \{(R_e, Y_e)\} \) are mutually independent, the conditional variance

\[
\text{Var}(\hat{\beta}_e^*|Y_e) = \frac{\sum_{i=1}^{G} (Y_i - \bar{Y})^2 \text{Var}(R_i|Y_e)}{\left( \sum_{i=1}^{G} (Y_i - \bar{Y})^2 \right)^2}.
\]

Again, appealing to the bivariate normality properties of \( R_e \) and \( Y_e \) it follows that

\[
\text{Var}(\hat{\beta}_e^*|Y_e) = \frac{\beta_1^2 \left( \sigma_{wg}^2 + \frac{\sigma_{ww}^2}{k} \right) \sigma_\epsilon^2}{k} + \Theta \left( \sigma_{wg}^2 + \frac{\sigma_{ww}^2}{k} \right),
\]

so that

\[
\text{Var}(\hat{\beta}_e^*|Y_e) = \frac{\Theta}{\sum_{i=1}^{G} (Y_i - \bar{Y})^2}.
\]

Since

\[
U = \sum_{i=1}^{G} \frac{(Y_i - \bar{Y})^2}{\left( \sigma_{wg}^2 + \frac{\sigma_{ww}^2}{k} \right) k},
\]

so that \( E(U^{-1}) = (G - 3)^{-1}, \) \( G > 3, \) it follows that

\[
\text{Var}(\hat{\beta}_e^*) = \frac{\Theta}{(G - 3)\left( \sigma_{wg}^2 + \frac{\sigma_{ww}^2}{k} \right)},
\]

which is equation 8.