

Appendix to “Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors”

Christopher Jackson, Nicky Best and Sylvia Richardson

Department of Epidemiology and Public Health, Imperial College, London†

Simulation study

One aim of this study is to assess the extent of contextual, or area-level, socio-demographic effects on cardiovascular hospitalisation. It is not generally possible to distinguish between a contextual and underlying individual effect if only the area-level mean of the aggregate covariate is available. (Greenland, 2002; Sheppard, 2003). However, our data are more detailed. The aggregate covariate data essentially consist of estimates of the cross-classification of individuals between three binary covariate and 12 strata categories. These are also supplemented with individual exposure-outcome data. To assess whether it is reasonable to interpret the regression coefficients associated with ethnicity, social class and car ownership in Section 4 as individual level effects, and the coefficient of the Carstairs index as a contextual effect, a brief simulation study was performed with a design based on the hospitalisation example.

Outcomes were simulated for the entire population using the logistic individual-level model (1). The design of the dataset was based on the district-level data. Specifically, data were simulated for 12 age-sex strata within 33 areas with the same population sizes as the corresponding London districts. The probability of a simulated individual in district i having a certain combination of covariates is set to be the proportion ϕ_{ik} of individuals in that covariate category in that district. The true model depends only on individual age, sex, individual low social class (with odds ratio 1.5) and the district proportion of individuals in a low social class (odds ratio 1.1 for an increase of 0.1 in this proportion). The simulated population data were aggregated to districts, and individual-level exposures and outcomes were drawn from the full data, using the same numbers of individuals per district as the 1997–2001 HSE (see Table 2).

The true model is refitted using, in turn, individual exposure-outcome data alone, aggregate data consisting of the within-district covariate cross-classification proportions, and a combination of aggregate and individual exposure-outcome data. We assess whether the individual and contextual effects of low social class can be distinguished. The percentage bias of the estimated odds ratios, the coverage of nominal 95% confidence intervals, and the root mean square error of the estimates relative to the true values, are calculated using results from 1000 simulations. These are given in the following table.

When data are simulated from a model depending on both individual and district-level social class, the model cannot estimate both effects accurately from the individual data alone (bias -70% for the contextual effect, first block). But from the aggregate, and combined data, the bias is reduced to negligible levels. Even if the individual effect but not the contextual effect is present, or vice

†Correspondence to: Dr. Christopher Jackson, Department of Epidemiology and Public Health, Imperial College School of Medicine, Norfolk Place, London W2 1PG. Email chris.jackson@imperial.ac.uk

Table 1. Results of simulation study: percentage bias (B), coverage of 95% confidence intervals (C) and percentage root mean square error (E) of estimates of odds ratios, over 1000 simulations.

| | Individual | | | Aggregate | | | Combined | | |
|--------------------------------------|------------|------|-------|-----------|------|-------|----------|------|-------|
| | B | C | E | B | C | E | B | C | E |
| Contextual effect | | | | | | | | | |
| P(social class) (area-level) | 0.6 | 94.0 | 426.0 | -1.2 | 94.4 | 23.0 | -0.8 | 94.0 | 21.3 |
| Social class (individual) | -70.1 | 96.3 | 108.0 | -1.4 | 95.1 | 35.5 | -1.8 | 95.0 | 32.0 |
| Contextual effect (no strata) | | | | | | | | | |
| P(social class) (area-level) | 122.8 | 92.8 | 325.2 | 20.9 | 88.1 | 81.6 | 135.1 | 26.5 | 131.8 |
| Social class (individual) | -17.8 | 95.7 | 60.4 | 46.2 | 56.3 | 315.0 | -7.1 | 92.3 | 68.8 |

versa, the model can still distinguish both effects from the aggregate data (biases 6% or less, not shown). However, suppose now that the only available aggregate covariate data are the proportions in low social class within each district, rather than estimates of the proportions occupying social class *within 12 age-sex categories* within each district. There is now no within-area information, as all strata are now combined into one. When the same simulation and model fit are performed, with no strata adjustment, the model is unable to distinguish between the individual and contextual effect, even with the combined data (135% bias for the contextual effect, final block of table). We conclude that it is possible to distinguish between individual and contextual effects of the same variable using the data that we have.