Online Supplement for:

Predictors of Endotoxin Levels in U.S. Housing

Peter S. Thorne¹*, Richard Cohn², Deepak Mav²,
Samuel J. Arbes, Jr.³, Darryl C. Zeldin³

¹Environmental Health Sciences Research Center, College of Public Health,
The University of Iowa, Iowa City, IA
²Constella Group, Inc., Durham, NC
³Division of Intramural Research, National Institute of Environmental Health Sciences,
Research Triangle Park, NC

Corresponding Author and Reprint Requests:

Peter S. Thorne, Ph.D.
Department of Occupational and Environmental Health
University of Iowa, College of Public Health
100 Oakdale Campus, 176 IREH
Iowa City, IA 52242-5000
Phone: (319) 335-4216
Fax: (319) 335-4006
Email: peter-thorne@uiowa.edu
Repeated Measures ANOVA Methodology

Repeated measures analysis of variance (rANOVA) were performed to assess the relationship between each housing or occupant characteristic and the level of endotoxin concentration (EU/mg) and endotoxin load (EU/m²). Endotoxin was evaluated as a continuous variable with logarithmic transformation. For the rANOVA, we preliminarily identified 37 possible predictors of log-transformed endotoxin concentrations or loads measured at five different locations for each household, based upon knowledge gleaned through previous research and the bivariate analysis results. These are shown in Table 5 of the published manuscript. Set 1 consisted of demographic factors, set 2 was comprised of characteristics of the home, set 3 included questionnaire data on pets and vermin, set 4 included field staff-observed evidence of household characteristics, and set 5 consisted of factors specific to bedrooms. We determined the optimal subset of these predictors using a rANOVA-based model selection process, with sampling locations treated as repeated measures and each household treated as an individual observation. In effect, the rANOVA approach characterizes relationships between predictors and the distribution of multiple related endotoxin measurements in a household.

Estimation and rANOVA model optimization were based upon a maximum likelihood procedure utilizing the Akaike information criterion (AIC) statistic. We implemented a hierarchical model selection procedure in which we partitioned predictor variables of interest into five logical sets and sequentially selected the best subset of predictor variables from each set using an exhaustive search. The process was repeated using all possible orderings of the variable sets to obtain the optimal set of predictors. The best subset of bedroom-specific predictors was obtained by fitting models using only bedroom floor and bedroom bed endotoxin levels.

In the rANOVA, sampling locations were treated as repeated measures and the household was treated as “subject”. Within-household residual errors were assumed to be multivariate normal variates with mean 0 and unstructured variance-covariance matrix. The
maximum likelihood estimation procedure was employed for simultaneous estimation of regression coefficients and covariance parameters. The ratio of the Akaike's information criterion (AIC) statistic to number of observations was used as the cost of a model. The identification of the model with smallest cost, using an exhaustive search would have required $2^{33}$ (~ 8 billion) models, which was computationally infeasible. This is the reason we implemented the hierarchical approximation in which we partitioned predictor variables of interest into five logical sets. The variables in set S5 are measurable only in the bedroom, and so S5 was handled separately. The model selection process involved selecting the best subset of predictor variables from S1 using an exhaustive search. This best set of predictors was then forced to stay in the model while the best subset of variables in S2 was selected to augment the subset from S1, again using an exhaustive search. Similarly, a best subset of S3 was selected with the selected subsets from both S1 and S2 forced in to the model, and a best subset of S4 was selected with the best subsets of S1, S2, and S3 forced into the model. This process is described as follows.

The cost of a model with predictor set $x$ was denoted as $C(x)$.

- Step 1: $i = 0; x_0 = \{ \text{LOCATION} \}$
- Step 2: Let $\Omega$ be collection of all possible subsets of $S_i$ then find $s_i$ such that
  $$C(x_i \cup s_i) = \min_{s_j \in \Omega} C(x_i \cup s_j)$$
- Step 3: $x_{i+1} = x_i \cup s_i; i = i + 1$
- Step 4: if $i<4$ then go to step 2.
- Step 5: Identify $x_4$ as best set of predictors.

The above algorithm was repeated using all 24 permutations of $\{S1, S2, S3, S4\}$ to obtain the optimal set of predictors and to eliminate carry-over bias. For these data, the same final set of predictors was obtained regardless of the order in which sets S1, S2, S3, and S4 were entered.
into the model. The best subset of bedroom specific predictors (S5 in Table 5), was obtained by repeating step 2 of the above algorithm using only bedroom floor and bedroom bed endotoxin levels. That is, we fit models with all subsets of S5 to just the bedroom floor and bedroom bed data, while forcing the models to include the best subsets of S1 through S4 found by fitting all five locations in the exercise above. The analyses were conducted in SAS (SAS Institute, Cary, NC).