APPENDIX AVAILABLE ON REQUEST

Research Report 152

Evaluating Heterogeneity in Indoor and Outdoor Air Pollution Using Land-Use Regression and Constrained Factor Analysis

Jonathan L. Levy et. al.

Appendix F. Results of Cholesky Residual Analysis of Normality for Constrained Factor Analyses

Note: Appendices Available on the Web appear in a different order than in the original Investigators' Report. HEI has not changed these documents. Appendices were relettered as follows:

Appendix D was originally Appendix A
Appendix E was originally Appendix B
Appendix F was originally Appendix C
Appendix G was originally Appendix D

Correspondence may be addressed to Dr Jonathan I. Levy, 715 Albany St., Talbot 4W, Boston, MA 02118.

Although this document was produced with partial funding by the United States Environmental Protection Agency under Assistance Award CR–83234701 to the Health Effects Institute, it has not been subjected to the Agency's peer and administrative review and therefore may not necessarily reflect the views of the Agency, and no official endorsement by it should be inferred. The contents of this document also have not been reviewed by private party institutions, including those that support the Health Effects Institute; therefore, it may not reflect the views or policies of these parties, and no endorsement by them should be inferred.

This document was reviewed by the HEI Health Review Committee
but did not undergo the HEI scientific editing and production process.

Appendix C: Results of Cholesky residual analysis of normality for constrained factor analyses.

As described in Section 5.2.3, we used the methodology of Houseman et al. (2006) to check on normality for our constrained factor analyses. We present figures in this appendix for indoor concentrations, noting that the qualitative conclusions are identical for outdoor concentrations.

Figures C-1 through C-6 show histograms and quantile-quantile (Q-Q) plots for three different assumed correlation structures: (1) unstructured (i.e., using the empirical variance-covariance matrix for the data); (2) constrained FA with 7 factors; and (3) constrained FA with 5 factors.

These figures demonstrate that for the weakest structural assumption, the residuals are extremely heavy-tailed but not extremely skewed. More restrictive models result in slightly more skewed residuals. While it is clear that multivariate normality is implausible, a symmetric but heavy-tailed distribution (i.e. multivariate *t*) might adequately describe the data to first approximation. Under these circumstances, the FA results can be expected to be approximately unbiased, provided that the data are missing completely at random (i.e. due to independent instrument error rather than high values of the constituent being measured, or the values of another constituent). As described in Section 6.2, missing data are largely due to analytical laboratory error or field measurement error and therefore adhere to this assumption. However, the FA results may be inefficient. Discussion of the implications of misspecified distributions in latent variable settings is available elsewhere (Little and Rubin 2002; Sanchez et al., 2009).

In spite of potential inefficiencies, it is worth mentioning that the source attribution interpretation of the FA model is destroyed by transformation; that transformations other than the logarithm are difficult to interpret in this context; and that while log-transformation improves the univariate symmetry of some of the pollutants, it worsens the univariate symmetry of others. In fact, the Cholesky residuals of log-transformed data look substantially worse (Figures C-7 and C-8). Thus, the analysis we have conducted represents the best possible of the many less-than-optimal procedures.

Figure C-1: Cholesky residual histogram for empirical variance-covariance matrix of indoor pollutant concentrations

Figure C-2: Cholesky residual Q-Q plot for empirical variance-covariance matrix of indoor pollutant concentrations