

CORRESPONDENCE



marlod: an R package to model environmental exposure and biomonitoring data with repeated measurements and values below the limit of detection

Keywords: Marginal analysis; Left censoring; Right skewness; Limit of detection; Repeated measures

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2025

Journal of Exposure Science & Environmental Epidemiology; <https://doi.org/10.1038/s41370-025-00752-8>

TO THE EDITOR:

In environmental and occupational research, when concentration measurements are taken from the same individual or study site, they are considered to be repeated measurements. Furthermore, when these samples fall below the limit of detection (LOD) of laboratory instruments, they are referred to as left-censored repeated measures data. These unquantified measurements typically represent low-level concentrations between zero and the LOD. There is an increasing need for statistical models that can effectively analyze left-censored environmental exposure and biomonitoring data with repeated measures, particularly among industrial hygienists. This is due to the growing importance of estimating the impact of exposure on disease risk and understanding the variability in occupational exposure both within and between workers. Analytical results from laboratories, including data on environmental contaminants (e.g., from hand wipes or personal breathing zone air samples) and occupational exposures (e.g., from biological media such as urine or serum) are often subject to non-detectable concentrations and tend to exhibit right-skewed distributions. The statistical modeling of exposure data that contain both repeated samples and samples below the LOD can be complex.

The author would like to introduce the R package 'marlod' [1] to analyze left-censored repeated measures exposure data when:

- data are assumed to be log-normally distributed using marginal mean regression models that employ generalized estimating equations (GEE), generalized method of moments (GMM), and quadratic inference functions (QIF) estimation methods,
- data are not assumed to follow a specific distribution using marginal quantile regression models, and
- longitudinal data contain time-varying covariates.

The 'marlod' package offers comprehensive functionality for univariable and multivariable analyses of environmental exposure and biomonitoring data. The reference manual for function documentation is accessible through <https://cran.r-project.org/web/packages/marlod/marlod.pdf>, while the vignette or user guide can be obtained either via <https://cran.r-project.org/web/packages/marlod/vignettes/marlod.html> or by using the command line '?marlod' in the R console. In existing literature, exposure data with correlated outcomes are often assumed to follow a log-normal distribution,


necessitating a natural logarithmic transformation [2]. Recently, small sample exposure data have been analyzed using a combination of marginal modeling with different substitution methods (i.e., single or multiple imputation techniques) to obtain mean regression estimates. The preferred approach uses simple working correlation structures, requiring the estimation of fewer nuisance covariance parameters [3]. Note that Bayesian methods are also suggested as viable alternatives for addressing left-censored data.

GEE is a specialized instance of marginal mean regression analysis, providing consistent regression parameter estimates even when the working correlation structure is mis-specified [4]. This property also allows marginal modeling to effectively handle data with highly right-skewed outcome distributions [3]. In addition to the GEE, the package incorporates the GMM method, a marginal analysis technique widely used in econometrics that leverages all available estimating equations [5]. Moreover, the package includes the QIF method, which demonstrates improved estimator efficiency relative to GEE, particularly when the working correlation structure is mis-specified and in large-sample settings [6].

The package also includes marginal quantile regression model for analyzing left-censored repeated measures exposure data [7]. When using traditional parametric mean regression, if the transformed exposure data do not adhere to a known distribution, then results may be suboptimal, as the estimated mean and standard deviation can be sensitive to large values. Additionally, the geometric mean, defined as the exponentiated mean of the log-transformed data, may become unstable when the distribution of the logged data is asymmetric [8]. Unlike parametric mean regression, quantile regression does not require assumptions about the underlying distribution and is robust against outliers, making it advantageous for skewed data. Moreover, quantile regression provides a comprehensive view of the entire conditional distribution of the dependent variable, eliminating the need for data transformation, regardless of the skewness of the data. The vignette includes a simulated dataset from the literature [3] to illustrate multivariable analyses in marginal mean and quantile regression models.

Lastly, this package allows users to develop either mean or quantile predictive regression models that address issues related to left censoring and time-dependent covariates, enabling a quantitative assessment of whether past and current covariates can predict current and/or future exposure levels [9]. In addition, a simulated dataset with a time-dependency mechanism is provided to facilitate selecting and applying a specific time-dependent type. Repeated measures data from clustered studies typically exhibit positive and uniform correlations among sampling outcomes (i.e., repeated measures) within the same cluster. Repeated measures data can also be longitudinal, where

outcomes are measured for each subject over time. In such studies, certain covariates may vary over time, and incorrectly treating time-dependent covariates as time-independent can lead to reduced efficiency in regression parameter estimation [10].

I-Chen Chen ¹✉

¹*Division of Field Studies and Engineering, National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention, Cincinnati, OH, USA. ✉email: okv0@cdc.gov*

REFERENCES

- Chen IC. marlod: Marginal Modeling for Exposure Data with Values Below the LOD. R package version 0.1.2. 2024. <https://CRAN.R-project.org/package=marlod>.
- Jin Y, Hein MJ, Deddens JA, Hines CJ. Analysis of lognormally distributed exposure data with repeated measures and values below the limit of detection using SAS. *Ann Occup Hyg*. 2011;55:97–112.
- Chen IC, Bertke SJ, Estill CF. Compare the marginal effects for environmental exposure and biomonitoring data with repeated measurements and values below the limit of detection. *J Expo Sci Environ Epidemiol*. 2024;34:1018–27.
- Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73:13–22.
- Hansen LP. Large sample properties of generalized method of moments estimators. *Econometrica*. 1982;50:1029–54.
- Qu A, Lindsay BG, Li B. Improving generalised estimating equations using quadratic inference functions. *Biometrika*. 2000;87:823–36.
- Chen IC, Bertke SJ, Curwin BD. Quantile regression for exposure data with repeated measures in the presence of non-detects. *J Exposure Sci Environ Epidemiol*. 2021;31:1057–66.
- Helsel DR. Less than obvious: statistical treatment of data below the detection limit. *Environ Sci Technol*. 1990;24:1766–74.
- Chen IC, Bertke SJ, Dahm MM. Quantile regression for longitudinal data with values below the limit of detection and time-dependent covariates—application to modeling carbon nanotube and nanofiber exposures. *Ann Work Expo Health*. 2024;68:846–58.
- Fitzmaurice GMA. A caveat concerning independence estimating equations with multiple multivariate binary data. *Biometrics*. 1995;51:309–17.

ACKNOWLEDGEMENTS

The author would like to thank Dr. Stephen Bertke and Dr. R. Michael Barker for their insight and feedback to the letter, and Dr. Philip Westgate and Dr. Liya Fu for their contributions to the R package.

AUTHOR CONTRIBUTIONS

ICC was responsible for developing the R package, including the reference manual for function documentation and the vignette; simulating and analyzing two datasets; drafting and revising the correspondence letter.

COMPETING INTERESTS

The author declares no competing interests.

DISCLAIMER

The findings and conclusions in this report are those of the author and do not necessarily represent the official position of the National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to I-Chen Chen.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.