Evaluation of Regression Methods for Log-Normal Data

Linear Models for Environmental Exposure and Biomarker Outcomes

Akademisk avhandling

som för avläggande av medicine doktorsexamen vid Sahlgrenska akademin,
Göteborgs universitet, kommer att offentligen försvaras i sal Hamberger,
Medicinaregatan 16, torsdag den 19 mars 2015 kl. 9:00

av

Sara Gustavsson

Fakultetsopponent:
Professor Jonas Björk,
Avdelningen för Arbets- och Miljömedicin
Lunds Universitet

Avhandlingen baseras på följande arbeten:

I. Gustavsson, S. M., Johannesson, S., Sallsten, G., and Andersson, E. M. (2012). Linear Maximum Likelihood Regression Analysis for Untransformed Log-Normally Distributed Data. *Open Journal of Statistics* **2**, 389-400.

II. Gustavsson, S., Fagerberg, B., Sallsten, G., and Andersson, E. (2014). Regression Models for Log-Normal Data: Comparing Different Methods for Quantifying the Association between Abdominal Adiposity and Biomarkers of Inflammation and Insulin Resistance. *International Journal of Environmental Research and Public Health* **11**, 3521-3539.

III. Gustavsson S., and Andersson E. M., Small-Sample Inference for Linear Regression on Untransformed Log-Normal Data.
Submitted for publication.

IV. Gustavsson S., Akerstrom M., Sallsten G., and Andersson E. M., Linear Regression on Log-Normal Data with Repeated Measurements.
Submitted for publication.

UNIVERSITY OF GOTHENBURG

Göteborg 2015

Evaluation of Regression Methods for Log-Normal Data

Linear Models for Environmental Exposure and Biomarker Outcomes

Sara Gustavsson

Department of Occupational and Environmental Medicine,
Institute of Medicine
Sahlgrenska Academy at University of Gothenburg

ABSTRACT

The identification and quantification of associations between variables is often of interest in occupational and environmental research, and regression analysis is commonly used to assess these associations. While exposures and biological data often have a positive skewness and can be approximated with the log-normal distribution, much of the inference in regression analysis is based on the normal distribution. A common approach is therefore to log-transform the data before the regression analysis. However, if the regression model contains quantitative predictors, a transformation often gives a more complex interpretation of the coefficients. A linear model in original scale (non-transformed data) estimates the additive effect of the predictor, while linear regression on a log-transformed response estimates the relative effect.

The overall aim of this thesis was to develop and evaluate a maximum likelihood method (denoted ML_{LN}) for estimating the absolute effects for the predictors in a regression model where the outcome follows a log-normal distribution. The ML_{LN} estimates were compared to estimates using common regression methods, both using large-scale simulation studies, and by applying the method to a number of real-life datasets. The method was also further developed to handle repeated measurements data. Our results show that when the association is linear and the sample size is large (> 100 observations), ML_{LN} provides basically unbiased point estimates and has accurate coverage for both confidence and predictor intervals. Our results also showed that, if the relationship is linear, log-transformation, which is the most commonly used method for regression on log-normal data, leads to erroneous point estimates, liberal prediction intervals, and erroneous confidence intervals. For independent samples, we also studied small-sample properties of the ML_{LN}-estimates; we suggest the use of bootstrap methods when samples are too small for the estimates to achieve the asymptotic properties.

Keywords: log-normal distribution, linear models, absolute effects

ISBN (printed): 978-91-628-9287-6

ISBN (e-publ.): 978-91-628-9295-1