



STATISTICAL METHODS IN ESTIMATING QUANTILES OF SERUM DIOXIN CONCENTRATION BY AGE, WITH VALUES BELOW LIMIT OF DETECTION

Qixuan Chen, Michael Elliott, Roderick Little, Elizabeth Hedgeman, Brenda Gillespie, David Garabrant
University of Michigan, Ann Arbor, Michigan, USA

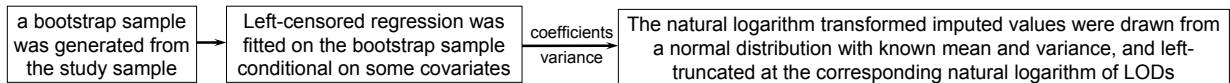
MOTIVATIONS & OBJECTIVES

- ❖ Some serum dioxin levels are below the limit of detections (LOD). → Using multiple imputation (MI) technique to obtain complete data set.
- ❖ Serum dioxin levels present a skewed distribution. → The median and other quantiles can catch important information that might be missed by measurements of central tendency and dispersion.
- ❖ Serum dioxin levels are positively associated with age. → Age-specific-estimates are of great interest.
- ❖ **Objectives:** This paper focuses on how to handle the LODs and how to estimate the age-specific-quantiles of serum dioxin concentration from a complex survey.

METHODS

Multiple imputation (MI) technique to obtain complete data sets

- Filling in the values below the LODs with LOD, LOD/2, or LOD/√2 (simple, but do not account for imputation uncertainty).
- An alternative approach of creating complete data set is called **multiple imputation** (MI).
 - The following procedure was repeated 5 times to generate the multiply imputed data sets:



Age-specific-quantile estimate

- **Quantile regression** generalizes a univariate quantile estimate of serum dioxin level to age-specific-quantile estimates.
 - There is no statistical software currently available that can fit survey weighted quantile regression.
 - With complete data, survey weighted quantile regression coefficients can be obtained by specifying sampling weights in the WEIGHT statement in the QUANTREG procedure (Experimental) in SAS 9.1.
 - The standard errors of the survey weighted quantile regression coefficients can be obtained using **Bootstrap sampling**.

Combining estimates from multiply imputed data sets

- The analysis of multiply imputed data sets can be carried out using the MIANALYZE procedure in SAS 9.1.
- Let $\hat{\beta}_m$ and \hat{V}_m , $m = 1, \dots, M$ be M complete-data estimates and their corresponding variances from the survey weighted quantile regression. The combined estimate and variance are

$$\hat{\beta}_M = \frac{1}{M} \sum_{m=1}^M \hat{\beta}_m \quad T_M = \frac{1}{M} \sum_{m=1}^M \hat{V}_m + \frac{M+1}{M(M-1)} \sum_{m=1}^M (\hat{\beta}_m - \hat{\beta}_M)^2$$

Bootstrap sampling

- Let $\hat{\beta}$ be the estimate of β from a survey weighted quantile regression based on a sample $S = \{i : i = 1, \dots, n\}$ of independent observations.
- Let $S^{(b)}$ denote a bootstrap sample, which is a sample of size n by sampling with replacement n times from S .
- Let $\hat{\beta}^{(b)}$ be the estimate of β from a survey weighted quantile regression based on $S^{(b)}$.
- With B bootstrap samples, the bootstrap estimate of the variance of $\hat{\beta}$ is

$$\hat{V}_{boot} = \frac{1}{B-1} \sum_{b=1}^B \left(\hat{\beta}^{(b)} - \frac{1}{B} \sum_{b=1}^B \hat{\beta}^{(b)} \right)^2$$

REAL EXAMPLE

2001-2002 NHANES data

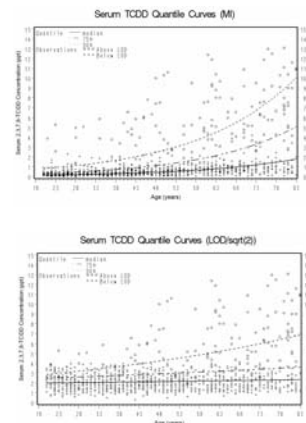
- Serum 2,3,7,8 TCDD levels was measured from a sample of 1228 subjects in the U.S. population age 20 years and over. A total of 1072 participants (87.30%) had serum TCDD concentrations below their LODs.

Statistical analysis

- Multiple imputation was performed conditional on age, BMI, gender, race, smoking, etc. (5 imputations)
- For each imputation, three quantile regression models (median, 75th percentile, and 90th percentile) were fitted over age. For comparison, the same three quantile regressions were also performed on the data by filling in the values below the LODs with LOD/√2.

Results

- The vertical axis in the figures is serum TCDD concentration in parts per trillion (ppt). Circles indicate values above their LOD, and pluses indicate values below LOD (non-detects). The top graph shows the estimated conditional quantile curves based on the model fitted on the multiply imputed data sets. For each result below the LOD, the average of the five imputed values was plotted. In contrast, the bottom graph is the estimated conditional curves based on the model fitted on the data with LOD/√2. For each result below the LOD, the value of LOD/√2 was plotted.
 - Filling in the non-detects with LOD/√2 leads to the over-estimation of age-specific-quantiles.
- The figures show that Serum TCDD levels increase with age. → It has the value to estimate the age-specific-quantiles, instead of the overall quantiles, since the overall quantiles depend on the distribution of the age in the sample.



CONCLUSIONS

- ❖ Multiple imputation can be employed to impute the values below the LOD, so that complete-data statistical methods can be implemented. For studies with large fraction of non-detects, the statistical analysis in such data are more sensitive to the imputation methods used.
- ❖ This age adjusted quantile estimates using quantile regression provide better estimates of quantiles than the traditional method of calculating the population quantiles without adjusting for age, or of adjusting for a limited number of age groups.

Financial support for this study comes from the Dow Chemical Company through an unrestricted grant to the University of Michigan.