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

Chen Q[1], Elliott MR[1], Little RJA[1], Hedgeman E[2], Gillespie BW[1], Garabrant D[2]

[1]Department of Biostatistics, University of Michigan School of Public Health, 109 S Observatory, Ann Arbor, MI 48109, USA; [2]Department of Environmental Health Sciences, University of Michigan School of Public Health, 109 S Observatory, Ann Arbor, MI 48109, USA

Abstract
In recent publications, age-adjusted quantile estimates of serum dioxin concentrations are obtained by categorizing subjects into several age groups and calculating quantiles stratified by age groups.[1] In the present study, authors used quantile regression to predict quantiles of serum dioxin concentration conditional on age, which enables better adjustment of the age effect. The present paper also describes a multiple imputation (MI) approach to impute serum dioxin values below the limit of detection (LOD) of GC/MS measurement instruments, and describes a method for fitting survey-weighted quantile regression. The MI technique and quantile regression were demonstrated using the serum 2, 3, 7, 8-TCDD concentrations measured in the 2001-2002 National Health and Nutrition Examination Survey (NHANES).

Introduction
In exposure assessment, the median and upper quantiles are sometimes of more interest than the mean, from a public health perspective. In the presence of a skewed distribution, the median and upper quantiles can also catch important information that might be missed by measurements of central tendency and dispersion. Additionally, serum dioxin levels are strongly positively related to age, such that quantiles conditional on age are of great interest. Furthermore, values below LOD result in left-censoring of the data, which make estimating the age-adjusted median and lower quantiles difficult. In the cases of the NHANES 2001-2002 data, a high proportion of the data is left-censored, especially in the young age groups, which makes estimation of all quantiles problematic in these groups. The results below LOD are also called non-detects. This paper focuses on how to handle non-detects and how to estimate the quantiles of serum dioxin concentration conditional on age. We also present the age-adjusted quantile estimates for 2, 3, 7, 8-TCDD using data from the NHANES 2001-2002.

Methods
Quantile regression generalizes a univariate quantile estimate of serum dioxin concentration to a conditional quantile estimate given age, or age and other covariates.[2] It provides a complete picture of the age effect on the serum dioxin concentration when a set of quantiles are modeled. It is especially useful to examine whether the tails and the central location of the conditional quantiles vary differently with age or other covariates. With complete data, quantile regression can be performed using the QUANTREG procedure (Experimental) in SAS 9.1.[3]

If subjects in the population have various probabilities of being included in the study, then the statistical analysis should be adjusted to reflect the population from which the sample was selected. However, there is no statistical software currently available that can fit survey weighted quantile regression. In this study, we obtain quantile regression coefficients from weighted quantile regression (specifying sampling weights in the WEIGHT statement in PROC QUANTREG in SAS), and correct the standard errors of the regression coefficients using Bootstrap sampling.

Bootstrap is a method of computing standard errors from the variability of estimates based on repeated resampling of the observed data. Let $\hat{\beta}$ be the estimate of β from a weighted quantile regression based on a

sample $S = \{i : i = 1,..., n\}$ 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 and let $\hat{\beta}^{(b)}$ be the estimate of β from a weighted quantile regression based on $S^{(b)}$. With B bootstrap samples, $\left(\hat{\beta}^{(1)},..., \hat{\beta}^{(B)}\right)$ are obtained. Then the bootstrap estimate of the variance of $\hat{\beta}$ is

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

Moreover, if the survey sampling involves clusters or strata, then bootstrap sampling should be carried out within each cluster or stratum.

For data with results below the LOD, Tobit regression can be used if the goal of the study is to get parameter estimates for mean models. However, parameter estimates in quantile regression models are of interest in this study. If complete data are available, then the regular quantile regression model can be applied. There are several simple strategies that are commonly employed to impute the results below LOD, including filling in the non-detects with LOD, LOD/2, and LOD/√2. These approaches are single imputation, that is, only one value is imputed for each value below LOD. The key problem with single imputation is that inferences about parameters based on the filled-in data do not account for imputation uncertainty. An alternative approach of creating complete data set is called multiple imputation (MI).[4] "A data set with a relatively small set of MIs can allow users to derive excellent inferences for a broad range of estimands with complete-data models, provided the MIs are based on a fine model."[5] In this study, we adopted MI strategy to handle the results below the LOD. The MI procedure described by Lubin et al. (2004) was used to impute results below the LOD.[6] Using this method, we assumed some distribution for serum dioxin concentration. For example, we assume a log-normal distribution. Then for each imputation, a bootstrap sample was generated from the study sample. Left-censored regression was fitted on the bootstrap sample conditional on covariates to obtain regression coefficients and variance estimates. The natural logarithm transformed imputed values were drawn from a normal distribution with known mean and variance, and left-truncated at the corresponding natural logarithm of LODs. This procedure was repeated 5 times to generate multiple data sets.

The analysis of a multiply imputed data set is quite direct, and can be carried out using the MIANALYZE procedure in SAS.[7] First, the survey weighted quantile regression was fitted on each imputed data set. Let $\hat{\beta}_m$ and $\hat{V}_m$, $m = 1,..., M$ be M complete-data estimates and their corresponding variances from the survey weighted quantile regression. The combined estimate is

$$\hat{\beta}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m . \quad (2)$$

And the total variability associated with $\hat{\beta}_M$ is

$$T_M = \frac{1}{M}\sum_{m=1}^{M}\hat{V}_m + \frac{M+1}{M(M-1)}\sum_{m=1}^{M}\left(\hat{\beta}_m - \hat{\beta}_M\right)^2 . \quad (3)$$

The total variability is composed of two variance components, within-imputation variance and between-imputation variance.

The method in estimating conditional quantile given age with values below LOD can be summarized as follow:
Step 1: The MI procedure described by Lubin et al. (2004) was used to impute results below the LOD, and generate M complete data sets.

Step 2: For each complete data set, survey weighted quantile regressions were fitted using bootstrap sampling to obtain the consistent variance estimates for parameters using formula (1).
Step 3: The combined estimates of parameter and associated variance were obtained using formula (2) and (3) from M imputed data sets.

Real Example

The 2001-2002 NHANES data on serum 2,3,7,8 TCDD levels was drawn from a representative sample of the U.S. population age 20 years and over.[7] TCDD concentration was lipid adjusted, and measured in parts per trillion (ppt). The demographic data, health indicator, and other related information collected during household interviews, as well as the survey design information (2-year sampling weights, PSU and stratum variables) were obtained from the NHANES 2001-2002 Household Questionnaire Data Files. In total, a sample of 1228 subjects who had complete serum TCDD measures was included in this study. A total of 1072 participants (87.30%) had serum TCDD concentrations below their corresponding LODs. The median LOD levels among the 1072 non-detects was 2.69 ppt, with a range from 0.42 to 5.80 ppt. The mean population age was 46 years (age was recorded as 85 for those older than 85).

A base 10 logarithm transformation was applied for serum TCDD concentrations. Multiple imputation was performed on this data set, by assuming a log-normal distribution for the serum dioxin concentration, conditional on age, body mass index (BMI), BMI gain and loss in the past one year, race, gender, smoking, breast feeding among women, income, and education. Five imputed data sets were generated. With each imputation, three quantile regression models (median, 70^{th} percentile, and 90^{th} percentile) were fitted over age, using the method described above. The results of the combined parameter estimates, standard errors, and P-values are displayed in Table 1. For comparison, the same three quantile regressions were also performed on the data by filling in the non-detects with LOD/√2 (Table 2). The predicted conditional quantiles (90^{th} percentile, 75^{th} percentile, and median) are plotted over age in Figure 1. The vertical axis is serum TCDD concentration in parts per trillion (ppt). Circles indicate values above their LOD, and pluses indicate values below LOD (non-detects). The left graph shows the estimated conditional quantile curves based on the model in Table 1. For each result below the LOD, the average of the five imputed values was plotted. In contrast, the right graph is the estimated conditional curves based on the model in Table 2. For each result below the LOD, the value of LOD/√2 was plotted.

In Table 1, the serum TCDD concentrations are highly positively associated with age (p-value < 0.001), which shows the value of estimating serum TCDD concentrations for different ages. With the estimates of intercepts and slopes, the median, 75^{th} percentile, and 90^{th} percentile for people in all ages (age >18) could be easily estimated (for example: estimated median serum TCDD = $e^{-1.224 + 0.017 \times (age-50)}$). Comparison of the two tables and plots shows that the intercept (at age 50) of the overall median, the 75^{th}, and the 90^{th} quantiles were overestimated by using LOD/√2. In addition, because the intercept was falsely elevated, the slope over age was underestimated by using LOD/√2.

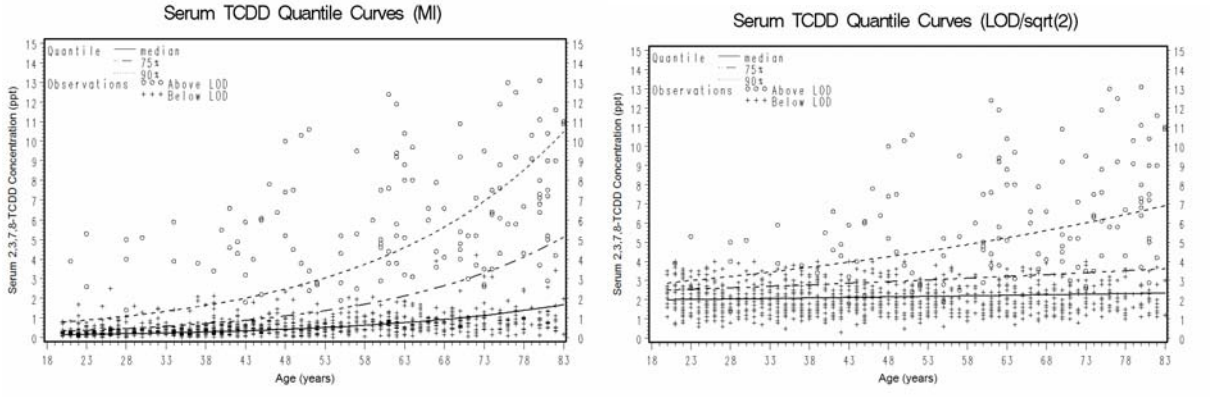
Table 1: Results of quantile regression for the data using Multiple Imputation technique

Parameter	Median			75th Percentile			90th Percentile		
	Est.	S.D.	p-value	Est.	S.D.	p-value	Est.	S.D.	p-value
Intercept	-1.224	0.128	<0.001	-0.891	0.134	<0.001	-0.472	0.098	<0.001
Age - 50	0.017	0.002	<0.001	0.019	0.002	<0.001	0.018	0.002	<0.001

Table 2: Results of quantile regression for the data filling in the non-detects with LOD/√2

Parameter	Median			75th Percentile			90th Percentile		
	Est.	S.D.	p-value	Est.	S.D.	p-value	Est.	S.D.	p-value
Intercept	0.276	0.029	<0.001	0.345	0.026	<0.001	0.336	0.045	<0.001
Age - 50	0.001	0.001	0.317	0.003	0.001	0.003	0.006	0.001	<0.001

Figure 1: Predicted conditional quantiles of serum TCDD given age



Conclusion and Discussion

Age-adjusted quantile estimation of serum dioxin concentrations is a valuable addition to the marginal quantile estimation methods and the traditional conditional mean modeling. Quantile regression allows better adjustment for age and provides more information in the presence of skewed distributions. In addition, multiple imputation can be employed to impute the values below the LOD, so that complete-data statistical methods can be implemented. For studies with relatively low LODs (few missing values and the real values that are missing close to 0), the estimation of conditional percentiles is less affected by the methods of dealing with the LOD issue, especially for upper percentiles. For dioxins, studies with small blood samples often have high LODs and, as a result, many samples have values below the LOD. Thus, the conditional quantile estimates in such data are more sensitive to the imputation methods used and to the distributional assumptions made. In the NHANES example above (in which 87% of results are below their LOD), filling in the non-detects with LOD/√2 leads to overestimation of quantiles by age.

Acknowledgements

Financial support for this study comes from the Dow Chemical Company through an unrestricted grant to the University of Michigan. The authors acknowledge Ms. Sharyn Vantine for her continuous assistance and Drs. Linda Birnbaum, Ron Hites, Paolo Boffetta and Marie Haring Sweeney for their guidance as members of our Scientific Advisory Board.

References

1. Caudil SP, Wong L-Y, Turner WE, Lee R, Henderson Alden, Patterson Jr. DG, 2007. *Chemosphere*
2. Koenker R, 2005. *Quantile Regression*, Cambridge University Press.
3. Chen C. *SUGI 30 2005*
4. Rubin, D.B., 1987. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.
5. Little, RJA and Rubin, DB (2002). *Statistical Analysis with Missing Data.* New York: John Wiley, 2nd Edition
6. Lubin JH, Colt JS, Camann D, Davis S, Cerhan JR, Severson RK, Bernstein Leslie, and Hartge P, 2004. *Environ. Health Perspect.* 112, 1691-1696.
7. SAS Institute. *SAS/STAT User's Guide Version 9*. Cary, NC: SAS Institute Inc., 2004.