

# *Linear regression with Type I interval- and left-censored response data*

MARY LOU THOMPSON<sup>1,2</sup> and KERRIE P. NELSON<sup>2</sup>

<sup>1</sup>*Department of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195, USA*

<sup>2</sup>*National Research Center for Statistics and the Environment, University of Washington*


*E-mail: mlt@u.washington.edu*

Received April 2001; Revised November 2001

---

Laboratory analyses in a variety of contexts may result in left- and interval-censored measurements. We develop and evaluate a maximum likelihood approach to linear regression analysis in this setting and compare this approach to commonly used simple substitution methods. We explore via simulation the impact on bias and power of censoring fraction and sample size in a range of settings. The maximum likelihood approach represents only a moderate increase in power, but we show that the bias in substitution estimates may be substantial.

*Keywords:* environmental data, laboratory analysis, maximum likelihood

1352-8505 © 2003  Kluwer Academic Publishers

---

## 1. Introduction

The statistical practices of chemists are designed to protect against mis-identifying a sample compound and falsely reporting a detectable concentration. In environmental assessment, trace amounts of contaminants of concern are thus often reported by the laboratory as “non-detects” or “trace”, in which case the data may be left- and interval-censored respectively. Samples below the “non-detect” threshold have contaminant levels which are regarded as being below the limits of detection and “trace” samples have levels that are above the limit of detection, but below the quantifiable limit. The type of censoring encountered in this setting is called Type I censoring: the censoring thresholds (“non-detect” or “trace”) are fixed and the number of censored observations is random. We consider here the problem of linear regression modeling in the setting where the outcome of interest,  $Y$ , is subject to Type I left- and interval-censoring.

The analysis of singly censored response observations has received attention in the biostatistical (e.g., in the context of survival analysis) and in the environmental literature. Useful summaries of approaches in environmental statistics may be found in Akritas *et al.* (1994) and Gilbert (1995). The literature on linear regression with censored data has focussed primarily on the setting of right-censored (typically survival) outcomes and random censoring mechanisms, an assumption that is not met in our setting, where the censoring thresholds may, for instance, be constant. Buckley and James (1979) proposed an approach where the (right) censored values are replaced by a weighted average of the

uncensored observations, with weights being derived from a non-parametric estimate of the error distribution function. Schmee and Hahn (1979) considered an iterative least squares approach, assuming Gaussian errors. Aitkin (1981) linked this approach to maximum likelihood and use of the EM algorithm (Dempster *et al.*, 1977). Ireson and Rao (1985) considered non-parametric estimation of the slope in a simple linear regression model in the presence of random censoring. Ritov (1990) considered an estimating equation approach for right-censored outcomes that is a modification of M-estimation for regression. Wei and Tanner (1991) developed a multiple imputation approach to regression with right-censored outcomes where, again, the censoring mechanism is assumed to be random. Akritas *et al.* (1995) developed an extension of Theil-Sen estimation in the context of right censoring of both  $Y$  and a single covariate  $X$ . Akritas (1996) considered an approach where least squares is used to fit a polynomial regression model with a single covariate  $X$  using a non-parametric estimate of the conditional expectation of  $Y$ , given  $X$ , based on a window around each observed  $X$ . Zhang and Li (1996) considered the setting where  $Y$  is subject to left- and right-censoring and extended the estimating equation approach of Ritov (1990).

In this paper, we extend the work of Aitkin (1981) to develop a maximum likelihood approach for the setting which includes Type I left- and interval-censoring. We evaluate and compare this approach with once-off substitution of censored values (commonly used in practice in environmental analyzes) through a practical example and by simulation. In Section 2 we develop the methodology for estimating regression parameters in the presence of interval- and left-censored data,  $Y$ . Section 3 contains a practical example and we evaluate and compare the developed procedures in a simulation study in Section 4. Section 5 contains a discussion.

## 2. Estimation

### 2.1 Maximum likelihood estimation

We assume that the data consist of observations  $(Y_i, X_{i1}, \dots, X_{ik}), i = 1, 2, \dots, n$  where the  $Y_i$  may be numeric or denoted as “non-detect” or “trace” in which case  $Y_i \leq c_{1i}$  (left-censored) or  $c_{1i} \leq Y_i \leq c_{2i}$  (interval-censored) respectively, where  $c_{1i}$  and  $c_{2i}$  are known. For notational convenience, we assume further that the data have been ordered so that the first  $n_1$  observations are “non-detect” (and so are left-censored), the next  $n_2$  observations are “trace” (and so are interval-censored) and the final  $n_3$  observations are numeric, where  $n = n_1 + n_2 + n_3$ .

The  $Y_i$  are assumed to follow a linear model, where:

$$\mu_i = E(Y_i | \mathbf{X}_i) = \sum_{j=0}^k \beta_j \mathbf{X}_{ij}.$$

If we assume that  $Y_i | \mathbf{X}_i \sim \mathcal{N}(\mu_i, \sigma^2), i = 1, 2, \dots, n$ , then

$$L(y_1, y_2, \dots, y_n) = \prod_{i=1}^{n_1} \Phi\left(\frac{c_{1i} - \mu_i}{\sigma}\right) \prod_{i=n_1+1}^{n_1+n_2} \left(\Phi\left(\frac{c_{2i} - \mu_i}{\sigma}\right) - \Phi\left(\frac{c_{1i} - \mu_i}{\sigma}\right)\right) \prod_{i=n_1+n_2+1}^n \phi(z_i)$$

where  $z_i = (y_i - \mu_i)/\sigma$ , and  $\phi$  and  $\Phi$  are the standard normal density and distribution functions respectively.

It can be shown that maximization of the above likelihood with respect to  $\beta_j$ ,  $j = 0, 1, 2, \dots, k$  involves solving the equations

$$\sum_{i=1}^n (y_i^* - \mu_i) x_{ij} = 0, \tag{1}$$

where

$$\begin{aligned} y_i^* &= \mu_i - \sigma \frac{\phi\left(\frac{c_{1i} - \mu_i}{\sigma}\right)}{\Phi\left(\frac{c_{1i} - \mu_i}{\sigma}\right)}, & i = 1, 2, \dots, n_1 \\ &= \mu_i - \sigma \frac{\phi\left(\frac{c_{2i} - \mu_i}{\sigma}\right) - \phi\left(\frac{c_{1i} - \mu_i}{\sigma}\right)}{\Phi\left(\frac{c_{2i} - \mu_i}{\sigma}\right) - \Phi\left(\frac{c_{1i} - \mu_i}{\sigma}\right)}, & i = n_1 + 1, \dots, n_1 + n_2 \\ &= y_i & n_1 + n_2 + 1 \leq i \leq n. \end{aligned}$$

Note that this is equivalent to the least square estimators for  $\beta_j$  based on ‘‘complete’’ data  $y_1^*, y_2^*, \dots, y_n^*$ .

The maximum likelihood estimator of the residual variance is given by:

$$\hat{\sigma}^2 = \sum_{i=n_1+n_2+1}^n \frac{(y_i - \hat{\mu}_i)^2}{D}, \tag{2}$$

where

$$D = n_3 + \sum_{i=1}^{n_1} \frac{\phi(u_i)u_i}{\Phi(u_i)} + \sum_{i=n_1+1}^{n_1+n_2} \frac{\phi(v_i)v_i - \phi(u_i)u_i}{\Phi(v_i)v_i - \Phi(u_i)u_i},$$

and where  $u_i = (c_{1i} - \hat{\mu}_i)/\hat{\sigma}$  and  $v_i = (c_{2i} - \hat{\mu}_i)/\hat{\sigma}$ .

The maximum likelihood estimates may be obtained via application of the EM algorithm. The sufficient statistics for the parameters in the case where there is no censoring (complete data) are  $\sum_{i=1}^n y_i^2$  and  $\sum_{i=1}^n x_{ij}y_i$ ,  $j = 0, 1, \dots, k$ . Note that

$$E[Y_i | c_{1i} \leq Y_i \leq c_{2i}] = \mu_i - \sigma \frac{\phi(v_i) - \phi(u_i)}{\Phi(v_i) - \Phi(u_i)},$$

and

$$E[Y_i^2 | c_{1i} \leq Y_i \leq c_{2i}] = \mu_i^2 + \sigma^2 - \mu_i \sigma \frac{\phi(v_i) - \phi(u_i)}{\Phi(v_i) - \Phi(u_i)} - \sigma \frac{c_{2i}\phi(v_i) - c_{1i}\phi(u_i)}{\Phi(v_i) - \Phi(u_i)}.$$

The corresponding conditional expectations for the case of left-censoring can be obtained as a special case of the above expressions.

Equations (1) and (2) for the maximum likelihood solutions based on the incomplete data are hence equivalent to those that would be obtained by replacing the censored  $y_i$  and  $y_i^2$ ,  $i = 1, 2, \dots, n_1 + n_2$ , by their conditional expectations, given the observed data and the current parameter estimates, until convergence.

The maximum likelihood approach also yields explicit (but cumbersome) expressions for the components of the information matrix associated with this model, from which the variance-covariance matrix of the parameters may be estimated.

## 2.2 Midpoint substitution

A simple approach that is commonly used in analysis of censored data in environmental settings involves replacing the censored values with a constant, such as the midpoint of the censoring interval. The estimates of model parameters are then obtained treating the data as if they were complete (Helsel and Hirsch, 1992; Davis, 1994; Gilbert, 1995; Loewenherz *et al.*, 1997). This approach has the advantage of being readily implemented with standard software. In Sections 3 and 4 below we compare the use of midpoint substitution to the maximum likelihood approach outlined above.

## 3. Example of application

In a study undertaken in Wenatchee, Washington State (Loewenherz *et al.*, 1997) children up to six years of age were monitored for evidence of pesticide exposure. Laboratory metabolite level assessment was carried out on urine samples from a sample of children in the area and the age of the children and their residential proximity to sprayed fields was determined, among other covariates. We present below the results of a linear regression analysis for the levels of logged dimethylthiophosphate (DMTP) in 67 children, as a function of their age and their residential proximity to sprayed fields (within 50 ft of sprayed fields versus 50 ft and greater). These data may be obtained from the authors on request. The DMTP measurements reflect a high degree of censoring with 49.3% of observations being non-detects(left-censored) and a further 20.9% being recorded as having trace levels of the metabolite (interval-censored).

Table 1 shows results of regression analysis with outcome log (DMTP) from both simple midpoint substitution of censored values and maximum likelihood estimation. It can be seen that the substitution approach yields smaller estimates of the coefficients for both covariates and residual variance. However, both analyses indicate possible increase in mean DMTP levels in children who live within 50 ft of sprayed fields (on average approximately 77% higher than children who live further away), but provide little evidence of a trend in DMTP levels with age. The simulation studies below confirm the bias in the simple substitution approach and illustrate the effect of increased censoring level on power to detect significant relationships.

**Table 1.** Regression analysis for Wenatchee log (DMTP).

<i>Method</i>	<i>Variable</i>	$\hat{\beta}$	$\hat{SE}(\hat{\beta})$	<i>P-value</i>	$\hat{\sigma}^2$
MLE	Intercept	- 3.66	0.452	< 0.0001	1.47
	Proximity < 50 ft	0.570	0.338	0.092	
	Age (months)	- 0.014	0.0096	0.154	
Midpoint substitution	Intercept	- 3.56	0.344	< 0.0001	1.06
	Proximity < 50 ft	0.424	0.255	0.102	
	Age (months)	- 0.0085	0.0071	0.235	

### 4. Simulation study

We carried out a range of simulations in settings that are plausible for the type of data encountered in an environmental setting such as that described above. We consider a single continuous covariate  $X$  intended to represent proximity of residence to sprayed fields, as in the above example, which has a positive skew distribution (proportional to  $\chi^2_5$  with mean 0.5 and range 0.07 to 1.75 (miles)). We assume that

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \dots, n$$

where we consider

1. Type I (fixed) censoring thresholds set so as to represent (on average) 20%, 40% and 60% censoring with (on average) equal fractions of non-detect and trace observations,
2. Sample sizes  $n = 50$  and  $100$ ,
3. Gaussian and (location shifted) chi-square error distributions.

For each of these scenarios, we consider the results of 1000 simulations using both maximum likelihood and a simple substitution approach, where censored observations are replaced by the midpoint of their censoring interval. In addition, we consider parameter estimation in the setting where information as to the non-detect threshold has not been determined, i.e., we address the question of the loss in precision by ignoring interval censoring information and considering all the censoring as being left-censored.

Tables 2–4 show summaries of the simulation results for the model:

$$Y_i = -3 - 2X_i + \varepsilon_i, \quad i = 1, 2, \dots, n.$$

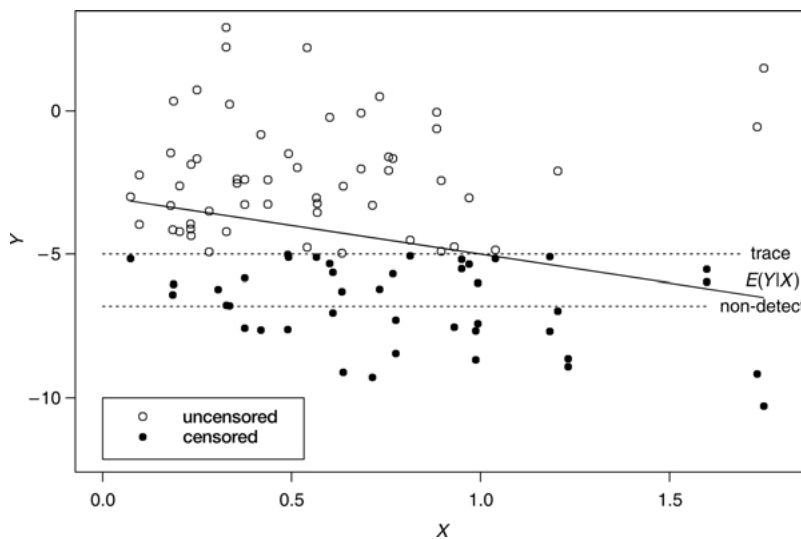


Figure 1. Simulated data with 40% censoring.

**Table 2.** Maximum likelihood estimation (midpoint substitution) with Gaussian errors.

$n$	Censoring	Power (%)	Coverage (%)	Bias( $\hat{\beta}_1$ )	Var( $\hat{\beta}_1$ )	$\hat{V}ar(\hat{\beta}_1)$	$\hat{\sigma}^2$
100	20%	63.8 (62.6)	94.5 (95.2)	0.02 (−0.15)	0.79 (0.62)	0.76 (0.67)	8.92 (8.17)
	40%	60.0 (58.6)	95.0 (93.2)	0.03 (−0.34)	0.83 (0.51)	0.83 (0.58)	9.09 (7.11)
	60%	58.2 (55.0)	95.0 (88.9)	0.06 (−0.55)	0.91 (0.41)	0.89 (0.49)	9.06 (5.97)
50	20%	47.8 (45.8)	94.2 (94.4)	0.01 (−0.16)	1.28 (1.00)	1.13 (1.01)	8.74 (8.15)
	40%	44.3 (41.8)	94.3 (95.5)	0.02 (−0.35)	1.36 (0.82)	1.22 (0.88)	8.91 (7.10)
	60%	41.9 (39.5)	94.4 (93.3)	0.05 (−0.55)	1.47 (0.66)	1.33 (0.73)	8.89 (5.95)

**Table 3.** Maximum likelihood estimation (midpoint substitution), non-Gaussian errors.

$n$	Censoring	Power (%)	Coverage (%)	Bias( $\hat{\beta}_1$ )	Var( $\hat{\beta}_1$ )	$\hat{V}ar(\hat{\beta}_1)$	$\hat{\sigma}^2$
100	20%	70.3 (61.8)	92.3 (95.1)	0.25 (−0.16)	0.98 (0.66)	0.78 (0.66)	8.83 (8.04)
	40%	66.6 (54.1)	92.9 (92.9)	0.31 (−0.35)	1.12 (0.61)	0.93 (0.64)	9.94 (7.85)
	60%	61.3 (47.2)	93.8 (89.1)	0.36 (−0.55)	1.32 (0.56)	1.11 (0.60)	11.16 (7.38)
50	20%	54.9 (47.6)	92.3 (95.1)	0.22 (−0.18)	1.35 (0.90)	1.13 (0.97)	8.56 (7.96)
	40%	50.5 (38.4)	93.7 (94.7)	0.28 (−0.37)	1.56 (0.84)	1.35 (0.94)	9.62 (7.75)
	60%	43.3 (31.7)	94.1 (92.2)	0.33 (−0.57)	1.82 (0.76)	1.62 (0.88)	10.81 (7.28)

**Table 4.** Maximum likelihood estimation (midpoint substitution) with Gaussian errors, left-censoring only.

$n$	Censoring	Power (%)	Coverage (%)	Bias( $\hat{\beta}_1$ )	Var( $\hat{\beta}_1$ )	$\hat{V}ar(\hat{\beta}_1)$	$\hat{\sigma}^2$
100	20%	63.1 (60.8)	94.9 (93.2)	0.06 (−0.31)	0.79 (0.49)	0.80 (0.58)	8.89 (7.09)
	40%	56.3 (51.1)	94.5 (80.2)	0.09 (−0.71)	0.94 (0.33)	0.94 (0.41)	8.99 (5.02)
	60%	45.9 (40.4)	94.7 (38.0)	0.11 (−1.12)	1.29 (0.20)	1.21 (0.25)	9.06 (3.07)
50	20%	45.6 (42.8)	94.6 (94.1)	0.13 (−0.26)	1.27 (0.74)	1.20 (0.88)	8.76 (7.10)
	40%	40.0 (34.9)	94.1 (85.3)	0.16 (−0.68)	1.51 (0.49)	1.43 (0.62)	8.89 (5.04)
	60%	30.9 (25.7)	94.7 (51.6)	0.20 (−1.10)	1.97 (0.23)	1.83 (0.37)	8.82 (3.06)

Fig. 1 shows a simulated data set for  $n = 100$  and 40% censoring under the above model with Gaussian errors. The correlation between  $Y$  and  $X$  in all cases is approximately  $-0.2$ , which is the sort of level that might be typical in the setting described in the above example. The corresponding error variances are 9 (8) with Gaussian (chi-square) errors. Simulations to evaluate size for the hypothesis test of zero slope indicated that both the maximum likelihood and midpoint substitution approaches give approximately valid type I error rates at the 5% level.

It can be seen that the estimates of  $\beta_1$  and  $\sigma^2$  using the substitution method are biased but that the power (for testing the hypothesis  $H_0 : \beta_1 = 0$ ) is comparable to that of the maximum likelihood approach. The coverage of the nominal 95% confidence interval for  $\beta_1$  based on maximum likelihood is close to 95% for all settings and is generally superior to that of the substitution approach, particularly for the case where all data are regarded as

left-censored (Table 4). Here, the midpoint substitution estimates are seriously biased and the larger variance associated with the smaller sample size ( $n = 50$ ) results in better coverage probability, although lower power.

The estimate of  $\beta_1$  based on the substitution method is, not surprisingly, less variable than that based on maximum likelihood estimation. The observed variance in the slope estimates ( $\text{var}(\hat{\beta}_1)$ ) for the maximum likelihood estimates agrees well with that based on the information matrix ( $\hat{\text{var}}(\hat{\beta}_1)$ ), particularly for larger sample size. The results in Table 3, however, indicate that, with regard to bias of the slope estimate, the maximum likelihood estimation is not robust to non-normality. Comparison of the results in Tables 2 and 4 demonstrates that there is a substantial loss in power at higher censoring levels when information regarding the non-detect threshold is ignored (i.e., when all observations are regarded as left-censored).

## 5. Discussion

We consider here linear regression analysis where the outcome variable is interval- and left-censored with censoring thresholds that are known, but may vary across observations. Data of this sort arise commonly in laboratory analyses of environmental exposure samples. The software to implement the maximum likelihood methodology with multiple covariates using *Splus* (Statistical Sciences, 1995) is available from the authors.

A common, easily implemented, approach to analysis of such data has been a simple once-off replacement of the censored observations by, say, the midpoint of their censoring interval. Our results show that, in the settings explored here, this method leads to biased parameter estimates, but has reasonable power, relative to maximum likelihood estimation. The existence of bias using substitution methods has been evaluated in other settings, see, e.g., El-Shaarawi and Esterby (1992) and Berthouex and Brown (1994). For exploratory analyses where identification of effects rather than estimation of effect size is the focus, this approach may be adequate. The maximum likelihood estimates are less subject to bias, but more variable than those resulting from simple substitution.

We further explored the impact of censoring level and sample size on parameter estimation. In the settings considered in our simulation, the reduction in power with increasing censoring fraction is surprisingly slow. As is to be expected, however, the effect on power of reduced sample size is substantial. The approach developed here will also prove valuable in study design and sample size determination. If the anticipated effect sizes and censoring fractions are specified, simulations under the assumed model will indicate the sample size needed to attain desired power or, alternatively, the censoring fraction compatible with desired power for given sample size.

Determination of the ‘‘non-detect’’ threshold represents a more time-consuming calibration problem and hence may not be undertaken in laboratory practice. Our results indicate that, particularly with high overall censoring levels, there is considerable loss in power when determination of this threshold is neglected and the outcome data are all left-censored.

The maximum likelihood approach does not appear to be robust to gross deviations from the normality assumption and it is hence particularly important that an appropriate transformation of the data be chosen. Use of a semi-parametric approach in this setting is an attractive alternative. Buckley and James (1979) suggested an iterative approach where

a non-parametric estimate of the error distribution function is obtained and the censored residuals are replaced by the weighted average of the uncensored residuals that lie in the same censoring interval, with the weights being obtained from the estimated distribution function. Turnbull (1974, 1976) and Frydman (1994) propose self-consistency algorithms for non-parametric estimation of the distribution function in settings that would accommodate left- and interval-censored data.

A glance at Fig. 1, however, will reveal a difficulty with application of this approach directly in our setting. Because of the nature of the censoring mechanism, the censored residuals are likely to be (relatively) large and negative whereas the uncensored residuals are likely to be large and positive. The uneven weight distribution will tend to bias the coefficient estimators. Some sort of symmetry assumption might be imposed to address this problem. Alternatively, the censored residuals might be replaced by estimates of their conditional expected values, based directly on the empirical distribution function and hence only indirectly on the uncensored residuals. Variance estimation in this setting would appear more challenging. In any event, estimation of a distribution function non-parametrically in the presence of substantial censoring appears a challenging and perhaps intractable problem.

In conclusion, we have developed and assessed a simple method to carry out multiple linear regression in the presence of left- and interval-censored outcome data. When the underlying Gaussian assumption is valid, the approach represents an improvement in terms of power and bias relative to commonly used once-off substitution approaches. Further work is needed to develop methods that are robust to the Gaussian assumption. In addition, we note that there are many aspects to the determination of laboratory assays that present interesting statistical questions. As noted by Akritas *et al.* (1994): "... until quantitative scientists and statisticians agree on the importance of seeing *all* the data, censoring will plague environmental data sets".

## Acknowledgment

The authors acknowledge helpful comments from the Editor. The research in this article has been funded in part by the United States Environmental Protection Agency under cooperative agreement CR825173-01-0 with the University of Washington. It has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred.

## References

- Aitken, M. (1981) A note on the regression analysis of censored data. *Technometrics*, **23**, 161–3.
- Akritas, M.G., Murphy, S.A., and LaValley, M.P. (1995) The Theil-Sen estimator with doubly censored data and applications to astronomy. *Journal of the American Statistical Association*, **90**, 170–7.
- Akritas, M.G., Ruscitti, T.F., and Patil, G.P. (1994) Statistical analysis of censored environmental data. In *Handbook of Statistics 12*, Environmental Statistics, G.P. Patil and C.R. Rao (eds), North-Holland, New York.



- Akritis, M.G. (1996) On the use of nonparametric regression techniques for fitting parametric regression models. *Biometrics*, **52**, 1342–62.
- Berthouex, P.M. and Brown, L.C. (1994) *Statistics for Environmental Engineers*, Lewis Publishers, Boca Raton.
- Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, **66**, 429–36.
- Davis, C.B. (1994) Environmental regulatory statistics. In *Handbook of Statistics 12*, Environmental Statistics, G.P. Patil and C.R. Rao (eds), North-Holland, New York.
- Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, **39**, 1–38.
- El-Shaarawi, A.H. and Esterby, S.R. (1991) Replacement of censored observations by a constant: an evaluation. *Water Research*, **26**, 835–44.
- Frydman, H. (1994) A note on nonparametric estimation of the distribution function from interval-censored and truncated observations. *Journal of the Royal Statistical Society, B*, **56**, 71–4.
- Gilbert, R.O. (1995) A review of statistical methods for data sets with multiple censoring points. Battelle Pacific Northwest Laboratories, U.S. Environmental Protection Agency, QAD, Washington D.C.
- Helsel, D.R. and Hirsch, R.M. (1992) *Statistical Methods in Water Resources*, Elsevier, New York.
- Ireson, M.J. and Rao, P.V. (1985) Interval estimation of slope with right-censored data. *Biometrika*, **72**, 601–8.
- Loewenherz, C., Fenske, R.A., Simcox, N.J., Bellamy, G., and Kalman, D. (1997) Biological monitoring of organophosphorus pesticide exposure among children of agricultural workers in central Washington state. *Environmental Health Perspectives*, **105**, 1344–53.
- Ritov, Y. (1990) Estimation in a linear regression model with censored data. *The Annals of Statistics*, **18**, 303–28.
- Schmee, J. and Hahn, G.J. (1979) A simple method for regression analysis with censored data. *Technometrics*, **21**, 417–32.
- Statistical Sciences, Inc. (1995) *S-Plus User's Manual, Version 3.4 for Unix*. Seattle, WA.
- Turnbull, B.W. (1974) Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association*, **69**, 169–73.
- Turnbull, B.W. (1976) The empirical distribution function with arbitrarily grouped censored and truncated data. *Journal of the Royal Statistical Society, B*, **38**, 290–5.
- Wei, G.C.G. and Tanner, M.A. (1991) Applications of multiple imputation to the analysis of censored regression data. *Biometrics*, **47**, 1297–1309.
- Zhang, C-H. and Li, X. (1996) Linear regression with doubly censored data. *The Annals of Statistics*, **24**, 2720–43.

## Biographical sketches

Mary Lou Thompson, Ph.D., is Research Associate Professor in the Department of Biostatistics at the University of Washington. She received her doctorate from the Georg-August Universität in Göttingen, Germany and is an Elected Member of the International Statistical Institute. Professor Thompson also has an ongoing association with the University of Cape Town in South Africa where she has a part-time position as Associate Professor in the Department of Public Health. In addition to environmental statistics, she has research interests in the areas of reference ranges and diagnostic testing, particularly as applied to occupational health and maternal and child health.

Kerrie Nelson is a graduate student in the Department of Statistics at the University of Washington, having entered the program in 1997. She carried out her undergraduate studies at the University of Auckland, New Zealand. Aside from her work in the statistical modeling of environmental data, her main research interests lie in correlated data analysis with special focus on generalized linear mixed models.