1997

# Analysis of Environmental Data with Censored Observations

Shiping Liu
*Banking, Finance, and Securities Consulting, Global Business Intelligence Solutions, IBM, San Francisco, California*

Jye-Chyi Lu
*Department of Statistics, North Carolina State University, Raleigh, North Carolina*

Dana Kolpin
*U.S. Geological Survey*

William Meeker
*Iowa State University*, wqmeeker@iastate.edu

Environ. Sci. Technol. **1997**, *31*, 3358−3362

# Analysis of Environmental Data with Censored Observations

SHIPING LIU
*Banking, Finance, and Securities Consulting, Global Business Intelligence Solutions, IBM, San Francisco, California 94111*

JYE-CHYI LU
*Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695*

DANA W. KOLPIN*
*U.S. Geological Survey, 400 S. Clinton Street, Iowa City, Iowa 52244*

WILLIAM Q. MEEKER
*Department of Statistics, Iowa State University, Ames, Iowa 50011*

The potential threats to humans and to terrestrial and aquatic ecosystems from environmental contamination could depend on the sum of the concentrations of different chemicals. However, direct summation of environmental data is not generally feasible because it is common for some chemical concentrations to be recorded as being below the analytical reporting limit. This creates special problems in the analysis of the data. A new model selection procedure, named forward censored regression, is introduced for selecting an appropriate model for environmental data with censored observations. The procedure is demonstrated using concentrations of atrazine (2-chloro-4-ethylamino-6-isopropylamino-*s*-triazine), deethylatrazine (DEA, 2-amino-4-chloro-6-isopropylamino-*s*-triazine), and deisopropylatrazine (DIA, 2-amino-4-chloro-6-ethylamino-*s*-triazine) in groundwater in the midwestern United States by using the data derived from a previous study conducted by the U.S. Geological Survey. More than 80% of the observations for each compound for this study were left censored at 0.05 $\mu$g/L. The values for censored observations of atrazine, DEA, and DIA are imputed with the selected models. The summation of atrazine residue (atrazine + DEA + DIA) can then be calculated using the combination of observed and imputed values to generate a pseudo-complete data set. The all-subsets regression procedure is applied to the pseudo-complete data to select the final model for atrazine residue. The methodology presented can be used to analyze similar cases of environmental contamination involving censored data.

## Introduction

It is common that some observations of environmental measurements such as herbicide concentrations in soil, air, and water are recorded as below specified analytical reporting limits due to measurement capacities or economical/practical concerns. This practice, however, creates special problems in the analysis of the data. Statistically, a data set with observations recorded as being below a certain limit is called "left censored" or simply "censored". In most environmental data analyses, censored data implies that values are only reported for those observations above some predetermined value (*1*). When data are censored, censored regression, or Tobit regression (*2*), is an appropriate method for data analysis (*1*, *3*).

Although censored regression has been widely used by statisticians and economists, it has been rarely used to analyze environmental data (*4*). For environmental data, however, the total concentrations of several contaminants in the environment, each having some censored observations, may be of interest. This total concentration could be calculated using a simple summation or a weighted summation, depending of the interest of the study and compounds being considered. For example, the potential threats of atrazine contamination to humans and terrestrial and aquatic ecosystems may depend on total concentrations of several compounds such as atrazine (ATZ), deethylatrazine (DEA), and deisopropylatrazine (DIA) in the environment. DEA and DIA can both be derived from the degradation of ATZ (*5*). The importance of DEA and DIA are that they are structurally and toxicological similar to ATZ (*6*−*8*). Therefore, the risks from the same concentration level of ATZ, DEA, and DIA likely are similar. The actual risk of using groundwater as a source of drinking water may depend on the total atrazine-residue concentration (ATZ + DEA + DIA) and not that of ATZ alone. Therefore, when designing appropriate policies to improve or protect the groundwater, it is necessary to consider total atrazine residue to properly determine risk levels. Direct summation of these three atrazine compounds, however, is not feasible because censored observations of ATZ, DEA, and DIA are common in groundwater (*9*). An appropriate statistical method has been developed to deal with this issue of censored environmental data.

The purpose of this paper is to present a statistical method developed for estimating significant factors that affect the total concentrations of several contaminants from measurements that include censored observations. The procedure is demonstrated using concentrations of ATZ, DEA, and DIA in groundwater. The methodology presented in this paper, however, can be used to analyze similar cases of environmental contamination involving censored data.

## Atrazine Data in Groundwater

The atrazine data used for this statistical demonstration were collected in a previous study of pesticides in groundwater of the midwestern United States by the U.S. Geological Survey (*10*−*12*). A total of 303 wells were sampled for this study (Figure 1). During 1991, 589 water samples were collected from these 303 wells in March−April (preplanting) and July−August (postplanting). The number of samples containing reported concentrations (>0.05 $\mu$g/L) were ATZ (101), DEA (106), and DIA (32), respectively. The maximum ATZ concentration in groundwater was 2.10 $\mu$g/L, about 30% below the maximum contamination level (MCL) for atrazine (3 $\mu$g/L). The maximum atrazine-residue concentration, however, was 4.48 $\mu$g/L, about 50% above the atrazine MCL and more than twice the maximum concentration of ATZ alone. This, however, is not an appropriate procedure for calculating a total concentration and will be addressed later. The statistical summary for concentrations above the 0.05 $\mu$g/L analytical reporting limit is given in Table 1.

Two groups of ancillary factors also were collected for each well: hydrogeologic and land use. Because multicollinearity among the regressors causes inefficiency and inconsistency, an initial screening procedure implemented in previous research (*13*, *14*) was used to eliminate variables that have limited explanatory power. Details of specific factors

* Corresponding author e-mail address: dwkolpin@usgs.gov; fax (319) 358-3606.


S0013-936X(96)00695-5 CCC: $14.00

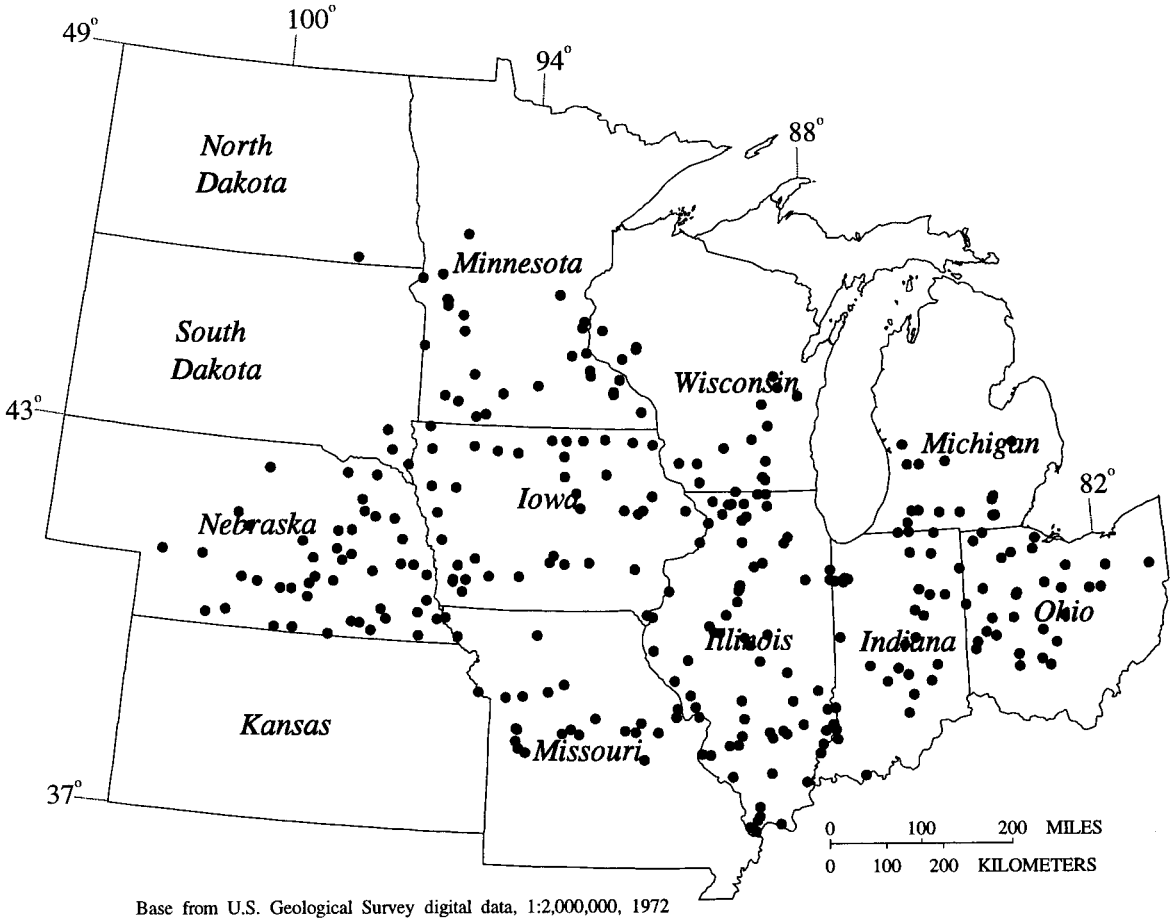


**FIGURE 1. Location of wells sampled for atrazine, deethylatrazine, and deisopropylatrazine used for this demonstration.**

**TABLE 1. Statistical Summary for Concentrations above the 0.05 μg/L Analytical Reporting Limit (Total Samples = 589)**

| | number | mean | std dev | minimum | maximum |
|---|---|---|---|---|---|
| atrazine | 101 | 0.298 | 0.399 | 0.050 | 2.100 |
| DEA | 106 | 0.211 | 0.334 | 0.050 | 2.300 |
| DIA | 32 | 0.218 | 0.230 | 0.050 | 1.170 |

collected in the survey are given elsewhere (*11, 12*). The sample statistics for selected factors are given in Table 2.

## The Statistical Model

Suppose we are interested in finding the factors that affect the total risk posed by pesticides in groundwater. The method commonly used to find the significant factors is regression analysis. Mathematically, this can be expressed as

$$G = g(\underline{x}) + \epsilon \quad (1)$$

where $G$ indicates the response (dependent) variable, $\underline{x}$ is a vector of the explanatory (independent) variables, $g(\underline{x})$ is a function of $\underline{x}$, and $\epsilon$ represents the modeling error (*15*). For this study, $\underline{x}$ is land use and hydrogeological characteristics, and the response variable, atrazine residue, $G$ is expressed as

$$G = \text{ATZ} + \text{DEA} + \text{DIA} \quad (2)$$

where ATZ, DEA, and DIA are concentration levels of atrazine and two of its degradation products in groundwater. To estimate significant factors that affect $G$ in eq 1, it is necessary to estimate values for $G$ given in eq 2 first. The problem arises because some observations of ATZ, DEA, and DIA are below the analytical reporting limit (0.05 μg/L). For levels below the analytical reporting limit, the observations were recorded as "less than 0.05 μg/L", and their precise values are unknown.

**TABLE 2. Sample Statistics for Explanatory Factors Used in the Statistical Analysis**

| variable | mean | std dev |
|---|---|---|
| percent of urban residential within 3.2 km (URBAN) | 8.661 | 12.822 |
| depth to top of aquifer (DEPTH) (m) | 5.488 | 5.089 |
| median of well open interval (OPEN) (m) | 26.956 | 25.086 |
| percent of pasture within 3.2 km (PASTURE) | 8.417 | 13.045 |
| percent of forest within 3.2 km (FOREST) | 10.265 | 10.913 |
| **Dummy Variables (yes = 1; no = 0)** | | |
| irrigation within 3.2 km (IRRID) | 0.301 | |
| chemical facility within 3.2 km (CHEM) | 0.147 | |
| golf course within 3.2 km (GOLF) | 0.130 | |
| primary water use is domestic (USEHD) | 0.501 | |
| primary water use is public supply (USEPD) | 0.301 | |
| well is unused (USEUD) | 0.112 | |
| aquifer class is unconsolidated (CLASSD) | 0.653 | |
| aquifer type is unconfined (TYPED) | 0.676 | |
| feedlot within 30 m (FEED1D) | 0.073 | |
| feedlot within 30 m-0.4 km (FEED2D) | 0.227 | |
| feedlot within 0.4–3.2 km (FEED3D) | 0.316 | |
| stream within 30 m (STR1D) | 0.077 | |
| stream within 30 m-0.4 km (STR2D) | 0.397 | |
| stream within 0.4–3.2 km (STR3D) | 0.569 | |
| sample in July or August (SUMMER) | 0.499 | |
| sample size | 589 | |

When a sample is censored, use of a standard estimation procedure such as simple linear regression by substituting an arbitrary value for censored data or treating censored data as missing values produces biased and inconsistent parameter estimates (*16*, *17*). Censored regression analysis provides an appropriate method to accommodate censoring in the response variable. The censored regression model is characterized by a latent regression equation

$$y_t^* = \underline{x}_t\underline{\beta} + \epsilon_t \tag{3}$$

where $y_t^*$ is the latent dependent variable, $\underline{x}_t$ is a vector of explanatory variables, $\underline{\beta}$ is a vector of parameters to be estimated, and the error term $\epsilon_t$ is assumed to be independently and normally distributed with mean 0 and variance $\sigma^2$. The observed dependent variable $y_t$ relates to the latent variable such that

$$\begin{aligned} y_t &= y_t^* \qquad \text{if } y_t^* > c \\ &= c \qquad \text{otherwise} \end{aligned} \tag{4}$$

where $c$ is the censored point. In this study, we consider the logarithmic transformation, which means that $G(y_t) = \log(y_t)$. [The standard Tobit model was considered in our preliminary analysis but was rejected based on non-nested specification test (*13*). Log here means $\log_e$.] Let $\Phi(\cdot)$ and $\phi(\cdot)$ denote the univariate standard normal distribution and density functions, respectively. Then, using eqs 3 and 4, the sample likelihood function $L$ for the lognormal Tobit model can be written as

$$L = \prod_{y_t = c} \Phi\left(\frac{\log c - \underline{x}_t\underline{\beta}}{\sigma}\right) \prod_{y_t > c} \frac{1}{\sigma y_t}\phi\left(\frac{\log y_t - \underline{x}_t\underline{\beta}}{\sigma}\right) \tag{5}$$

The above likelihood function is for only one compound. The likelihood function becomes extremely complicated for the case of three compounds, having eight possible combinations. The eight cases are (1) ATZ, DEA, and DIA are all observed; (2) ATZ and DEA are observed but DIA is censored; (3) ATZ and DIA are observed but DEA is censored; (4) DEA and DIA are observed but ATZ is censored; (5) DIA is observed but ATZ and DEA are censored; (6) DEA is observed but ATZ and DIA are censored; (7) ATZ is observed but DEA and DIA are censored; and (8) ATZ, DEA, and DIA are all censored. For each case, the likelihood function is a multiplication of either the cumulative density function or the probability density function of ATZ, DEA, and DIA, depending on what data is censored. This makes estimation extremely difficult and is one of the major reasons for estimating the parameters of the regression equation for each compound separately. Furthermore, ATZ, DEA, and DIA are log normally distributed. Therefore, the distribution of $G$ (sum of ATZ, DEA, and DIA) is unclear. The details of the likelihood function can be obtained from the authors (because of space considerations and complexity, it was not provided here). [An additional alternative approach to deal with the issue we are addressing here is to incorporate the censoring levels into the sum of the three compounds. By doing so, the sum of the three compounds is either uncensored (all three compounds are observed] interval censored (one or two compounds are censored), or left censored (all three compounds are censored). This approach, however, oversimplifies the problem for the case of at least one compound being censored. It is not clear what censoring level should be used. For example, 0.15 may be used as a censoring level for the case where all three compounds are censored. But G < 0.15 can come from an infinite number of different combinations (such as ATZ < 0.05, DIA < 0.05, and DEA < 0.05; or from ATZ < 0.10, DIA < 0.04, and DEA < 0.01). The same issue arises for the case of either one or two compounds being censored. In contrast, the methodology proposed in this study provides a precise restriction for each compound when it is censored. The procedure presented for this study is one that is both practically and theoretically justified to analyze the type of data used in this demonstration.]

By maximizing the likelihood function given in eq 5, the parameters $\underline{\beta}$ and $\sigma$ can be estimated. With the estimated parameters for each compound (ATZ, DEA, and DIA), the censored data $\log y_t$ can be imputed at its conditional mean as

$$E(\log y_t | \log y_t < \log c) = \hat{\mu} - \hat{\sigma}\frac{\phi(z)}{\Phi(z)} \tag{6}$$

where $\hat{\mu} = \underline{x}_t\underline{\beta}$ is the prediction from the estimated regression equation, $\underline{\hat{\beta}}$ is a vector of estimated coefficients and $\hat{\sigma}$ is the estimated standard deviation, and

$$z = (\log c - \hat{\mu})/\hat{\sigma} \tag{7}$$

The values of $y_t$ for the censored observations can be imputed by its conditional expectation, which can be expressed as

$$\begin{aligned} \hat{y} &= E(Y | Y \le L) \\ &= e^{\hat{\mu} + \hat{\sigma}^2/2}\,\frac{\Phi[(\log c - \hat{\mu})/\hat{\sigma} - \hat{\sigma}]}{\Phi[(\log c - \hat{\mu})/\hat{\sigma}]} \end{aligned} \tag{8}$$

Equation 8 is derived from the linear model $\ln(Y) = x_t\beta + \epsilon$, where $\epsilon$ is normally distributed with mean zero and variance $\sigma$, or $\epsilon \approx N(0,\sigma)$. Therefore, a bias adjustment procedure (*18*) eliminates the main portion of the bias in the inverse transformation, $\hat{E}(Y) = e^{x_t\hat{\beta}}e^{\hat{\sigma}^2/2}$, where $\hat{\beta}$ and $\hat{\sigma}^2$ are the estimators for the linear model given above. This method, however, is for uncensored data. For the censored data, the adjustments given in eq 8 are necessary. The derivation of eq 8 is given in the Appendix. In this study, the parameters in the model for the data $Y$ in the original scale are estimated (from either a nonlinear regression or transformed from the linear regression), and we can impute the censored data at its conditional expectation given in eq 8. By combining the imputed values for the censored observations and actual data for those above the censored point, pseudo-complete data sets are obtained (*19*).

The procedure just described can be applied to ATZ, DEA, and DIA to impute the values for those observations below the analytical reporting limit. With the estimated dependent variable, atrazine-residue concentrations, the significant factors that relate to environmental contaminant concentrations can be found by standard regression procedures.

## Censored Regression Model Selection and Imputation

As noted previously, most observations of ATZ, DIA, and DEA from the data used in this demonstration were censored at 0.05 $\mu$g/L. With censored regression data, the maximum likelihood estimation method is usually used to estimate the parameters of the regression equation. The regression parameters in eq 3 can be estimated by using the LIFEREG procedure in the SAS statistical program (*20*).

To identify the significant factors that relate to atrazine-residue concentrations in groundwater, an appropriate model selection procedure has to be used. There is, however, no procedure available for selecting a term in regression with censored data. LIFEREG and other statistical programs for analyzing censored data are designed only for estimating the parameters of a given regression model. They are not programmed to perform model selection.

To select an appropriate model with censored regression analysis, the censored forward regression procedure will be used in this study (*21*). The procedure is a forward stepwise procedure used with Tobit model. In this procedure, variables are added one at a time as long as they contribute significantly to the fit. The Wald-type statistic is used in judging whether

**TABLE 3. Estimated Parameters of Censored Regression Model for Atrazine**

| variable | parameter estimate | standard error | $\chi^2$ test $\chi^2$ Va | $\chi^2$ test Pr > $\chi$ |
|---|---|---|---|---|
| INTERCPT | −5.308 | 0.520 | 104.048 | 0.000 |
| USED[a] | | | | 0.000 |
| USEHD | −0.776 | 0.438 | 3.141 | 0.076 |
| USEPD | 0.692 | 0.436 | 2.513 | 0.113 |
| USEUD | −1.079 | 0.571 | 3.578 | 0.059 |
| STR1D | 1.009 | 0.420 | 5.770 | 0.016 |
| TYPED | 0.779 | 0.286 | 7.436 | 0.006 |
| FOREST | −0.036 | 0.014 | 6.321 | 0.012 |
| STR2D | 0.540 | 0.273 | 3.914 | 0.048 |
| URBAN | −0.021 | 0.011 | 3.802 | 0.051 |
| SUMMER | 0.447 | 0.248 | 3.247 | 0.072 |
| SCALE | 1.979 | 0.163 | | |

[a] The significance level of USEHD, USPD, USUD, and USOD was determined using a Wald-type statistic. This statistic is compared to a $\chi^2$ distribution with 1 degree of freedom. The significance level used in this case is 0.10.

a new variable should be added to the model. The significance level is artificially determined as 0.10.

With the selected model for each compound, the concentration levels for those sites where observations are recorded as below the analytical reporting limit could be imputed based on eq 8. The total atrazine-residue concentration for each site can then be calculated by using observed data for those sites where observations are above the analytical reporting limit and imputed data for those below the analytical reporting limit. Finally, the all-subset model selection procedure (*22*) can be used to select the final model for atrazine residue. The adjusted $R^2$ is used for selecting the final model. Therefore, the all subsets regression procedure simply picks the model with the highest $R^2$ value.

## Multiple Imputation

Although imputing the censored data at its conditional mean, allowing the use of standard complete-data methods of analysis is commonly used in practice; it has the drawback in that it treats the censored data as known values. This kind of treatment ignores the actual variability in the censored data values. Research has shown that a multiple imputation with random sample of size $m = 2$ can greatly improve the confidence interval coverage probabilities, performing better than the single sample imputation method in all studied cases (*23*). When there are more data censored, the random sample size should be increased. Comparing the amount of information missing in the previous study (*23*) to the amount of censored data in this demonstration, we selected a random sample size of 5 to insure a reasonably accurate imputation result. This method was used to examine the robustness of the parameter estimation from the pseudo-complete data sets.

## Empirical Results

By using the model selection procedure discussed above, the three censored regression models for the three atrazine compounds (ATZ, DEA, and DIA) were selected. The final selected models are given Tables 3 (ATZ), 4 (DEA), and 5 (DIA). The variables listed in Tables 3−5 are in the sequence of the variables entered in the models. For instance, the variable of the best one-term model for ATZ is USEHD, the variables in the final two-term model for ATZ include USEHD and USEPD, and so on. The details about the estimated parameters and their $\chi^2$ test values are also given in Tables 3−5.

With the estimated censored regression equations, the values of censored observations can be imputed based on eq 8. The values of $\hat{\mu}$ in eq 6 were calculated by using the

**TABLE 4. Estimated Parameters of Censored Regression Model for DEA**

| variable | parameter estimate | standard error | $\chi^2$ test $\chi^2$ Va | $\chi^2$ test Pr > $\chi$ |
|---|---|---|---|---|
| INTERCPT | −4.842 | 0.530 | 83.385 | 0.000 |
| FOREST | −0.046 | 0.012 | 14.486 | 0.000 |
| TYPED | 0.831 | 0.244 | 11.598 | 0.000 |
| USED[a] | | | | 0.035 |
| USEHD | 0.138 | 0.399 | 0.119 | 0.730 |
| USEPD | 0.795 | 0.398 | 3.995 | 0.046 |
| USEUD | 0.436 | 0.462 | 0.893 | 0.345 |
| STR1D | 0.753 | 0.327 | 5.310 | 0.021 |
| SUMMER | 0.399 | 0.198 | 4.056 | 0.044 |
| STR3D | 0.435 | 0.208 | 4.366 | 0.037 |
| CLASSD | −0.726 | 0.287 | 6.403 | 0.011 |
| OPEN | −0.003 | 0.002 | 3.682 | 0.055 |
| SCALE | 1.623 | 0.131 | | |

[a] The significance level of USEHD, USPD, USUD, and USOD was determined using a Wald-type statistic. This statistic is compared to a $\chi^2$ distribution with 1 degree of freedom. The significance level used in this case is 0.10.

**TABLE 5. Estimated Parameters of Censored Regression Model for DIA**

| variable | parameter estimate | standard error | $\chi^2$ test $\chi^2$ | $\chi^2$ test Pr > $\chi$ |
|---|---|---|---|---|
| INTERCPT | −5.539 | 1.236 | 20.072 | 0.000 |
| USED[a] | | | | 0.027 |
| USEHD | −0.062 | 0.869 | 0.005 | 0.943 |
| USEPD | 1.707 | 0.848 | 4.054 | 0.044 |
| USEUD | 1.421 | 0.900 | 2.491 | 0.115 |
| STR1D | 2.094 | 0.585 | 12.831 | 0.000 |
| OPEN | −0.027 | 0.009 | 8.549 | 0.004 |
| FOREST | −0.085 | 0.032 | 7.243 | 0.007 |
| CLASSD | −2.328 | 0.807 | 8.329 | 0.004 |
| TYPED | 1.339 | 0.565 | 5.622 | 0.018 |
| STR3D | 0.989 | 0.454 | 4.741 | 0.030 |
| SUMMER | 0.798 | 0.408 | 3.823 | 0.051 |
| SCALE | 1.972 | 0.300 | | |

[a] The significance level of USEHD, USPD, USUD, and USOD was determined using a Wald-type statistic. This statistic is compared to a $\chi^2$ distribution with 1 degree of freedom. The significance level used in this case is 0.10.

estimated parameters from Tables 3−5. The conditional mean for each compound at each censored site is imputed with eq 8 also using the estimated parameters from Tables 3−5.

With the estimated mean for each compound (ATZ, DEA, DIA) at each censored site and a standard deviation, five sets of observations were generated for each site. With randomly generated values, the inverse transformation based on eq 8 was used to obtain an imputed value for each site. At the end of this process, five pseudo-complete data sets were obtained.

The statistics of the pseudo-complete data are given in Table 6. The means of the imputed minimum concentrations for censored data based on the five pseudo-complete data sets were $4.48 \times 10^{-5}$, $1.274 \times 10^{-4}$, and $9.576 \times 10^{-9}$ for ATZ, DEA, and DIA, respectively, much less than the censored limit of 0.05 $\mu$g/L. Previous research has confirmed the prevalence of ATZ and DEA concentrations in groundwater below 0.05 $\mu$g/L (*24*). The frequency of atrazine detection roughly doubles if the reporting limit is lowered from 0.05 to 0.003 $\mu$g/L.

By using the pseudo-complete data, the final model for atrazine residue was estimated by an all-subset model selection procedure (*22*), although other methods, such as the sum of squares analysis, could also have been used. The final selected models are not given here because of the issue

**TABLE 6. Statistical Summary for Concentrations (in $\mu$g/L) of Pseudo-Complete Data**

| compound | sample size | mean[a] | std dev[a] | minimum[a] | maximum[a] |
|---|---|---|---|---|---|
| atrazine residue | 589 | 0.119 | 0.358 | $1.388 \times 10^{-3}$ | 4.480 |
| atrazine | 589 | 0.057 | 0.198 | $7.048 \times 10^{-5}$ | 2.100 |
| DEA | 589 | 0.046 | 0.161 | $1.274 \times 10^{-4}$ | 2.300 |
| DIA | 589 | 0.016 | 0.072 | $9.576 \times 10^{-9}$ | 1.170 |

[a] The values are averaged across five imputations. The variations among different imputations for each statistics are limited, indicating stability in our models.

**TABLE 7. Estimated Parameters of Final Regression Model for Atrazine Residue**

| variable | parameter estimate sign | significant at 0.05 (*) or 0.10 (**)[a] |
|---|---|---|
| INTERCEP | | |
| OPEN | − | * |
| URBAN | − | |
| FOREST | − | * |
| CLASSD | − | ** |
| TYPED | + | * |
| STR1D | + | * |
| STR3D | + | * |
| SUMMER | + | ** |
| IRRID | + | * |
| USEPD | + | * |

[a] The variables selected and their associated significance levels are based on the pseudo-complete data set. Because the likelihood is very complicated, it is impractical to use the large sample normal approximation or finite sample simulation to obtain the exact significance levels. Although the significance levels reported from SAS output as given in this table are not exact due to the imputation of censored data, it is practical for most readers that typical regression procedures can be used to analyze censored data in the type of example addressed in this paper. Moreover, all five pseudo-complete data sets resulted in the same variables selected and their significance levels were all similar. Thus, the proposed procedure is reasonable to provide practical solutions to the problem addressed in this demonstration.

discussed in the multiple imputation section. The model was not unique, depending on which pseudo-complete data set is used. In Table 7, only the sign and significance level for the variables are given. Five final models selected from five pseudo-complete data sets for atrazine residue were identical in terms of variables selected, significant levels, and signs of coefficients, indicating the stability of the model. The model selected estimates based on each pseudo-complete data set and is the best model obtained from the list of 20 potential explanatory variables available (Table 1).

## Appendix: Derivation of the Mean for the Truncated Lognormal Distribution

The variable $Y$ is lognormally distributed if $T = \log Y$ is normally distributed with mean $\mu$ and variance $\sigma^2$. If $Y$ is lognormally distributed, then $Z = (\log Y - \mu)/\sigma$ is the standard normal distribution. The condition mean of $Y$, giving $Y < c$, can be expressed as

$$E(Y|Y < c) = E(\exp(\sigma Z + \mu)|\exp(\sigma Z + \mu) < c)$$
$$= E\left(\exp(\sigma Z + \mu)\middle| Z < \left(\frac{\log c - \mu}{\sigma}\right)\right)$$

To simply expressions, let's define

$$h_1(\mu,\sigma) = \frac{\log c - \mu}{\sigma}$$

$$h_2(\mu,\sigma) = \Phi(h_1(\mu,\sigma))$$

where $\Phi$ is the cumulative distribution function (cdf) of the standard normal distribution. With these notations, the above expectation becomes

$$E(Y|Y < c)$$
$$= h_2^{-1}(\mu,\sigma)\int_{-\infty}^{h_1(\mu,\sigma)} \exp(\sigma Z + \mu)(2\pi)^{-1/2}\exp(-z^2/2)\mathrm{d}z$$
$$= \exp(\mu + \sigma^2/2)h_2^{-1}(\mu,\sigma)\int_{-\infty}^{h_1(\mu,\sigma)} (2\pi)^{-1/2} \times \exp(-(z - \sigma)^2/2)\mathrm{d}z$$
$$= \exp(\mu + \sigma^2/2)\frac{\Phi[(\log c - \mu)/\sigma - \sigma]}{\Phi[(\log c - \mu)/\sigma]}$$

By replacing $\mu$ and $\sigma$ in the above equation by their estimated values of $\hat{\mu}$ and $\hat{\sigma}$, it becomes eq 8.

## Literature Cited


(1) Lawless, J. F. *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, Inc.: New York, 1982.
(2) Tobin, J. *Econometrica* **1958**, *26*, 24-36.
(3) Greene, W. H. *Econometric Analysis*, Macmillan Publishing Company: New York, 1990.
(4) Siymen, D.; Peyster, A. D. *Environ. Sci. Technol.* **1994**, *28*, 898−902.
(5) Paris, D. F; Lewis, D. L. *Residue Rev.* **1973**, *45*, 95−124.
(6) Ciba-Geigy Corporation. *Summary of Toxicological Data on Atrazine and Its Chorotrazine Metabolites*; Attachment 12, 56 FR 3526, Ciba-Geigy1993.
(7) Kaufman, D. D.; Kearney, P. C. *Residue Rev.* **1970**, *32*, 235−265.
(8) Moreau, C.; Mouvet, C. *J. Environ. Qual.* **1997**, *26*, 416−424.
(9) Kolpin, D. W.; Thurman, E. M.; Goolsby, D. A. *Environ. Sci. Technol.* **1996**, *30*, 335−340.
(10) Burkart, M. R.; Kolpin, D. W. *J. Environ. Qual.* **1993**, *22*, 646−656.
(11) Kolpin, D. W.; Burkart, M. R.; Thurman, E. M. *Open-File Rep. U.S. Geol. Surv.* **1993**, *93−114*.
(12) Kolpin, D. W.; Burkart, M. R.; Thurman, E. M. U. S. *Geol. Surv. Water-Supply Pap.* **1994**, *2413*.
(13) Liu, S.; Yen, S. T.; Kolpin, D. W. *J. Environ. Qual.* **1996**, *25*, 992−999.
(14) Liu, S.; Yen, S. T.; Kolpin, D. W. *Water Resour. Bull.* **1996**, *32*, 845−853.
(15) Johnston, J. *Econometric Methods*, 3rd ed.; McGraw-Hill Publishing Company: New York, 1984.
(16) Amemiya, T. *Advanced Econometrics*, Harvard University Press, Cambridge, MA, 1985.
(17) Liu, S.; Stedinger, J. R. *Water Resources Planning and Management and Urban Water Resources*, The American Society of Civil Engineers, 1991; pp 27−31.
(18) Miller, D. M. *Am. Statistician* **1984**, *38*, 124−126.
(19) Schmee, J.; Hahn, G. J. *Technometrics* **1979**, *21*, 417−432.
(20) SAS, Version 6, *A Statistical Software System Registered Trademark of SAS Institute Inc.*, Cary, NC, 1989.
(21) Lu, J.-C; Liu, S.; Unal, C. Written communication; Department of Statistics, North Carolina State University, Raleigh, NC, 1995.
(22) Draper, N. R.; Smith, H. *Applied Regression Analysis*, 2nd ed.; John Wiley: New York, 1981.
(23) Rubin, D. B.; Schenker, N. *J. Am. Stat. Assoc.* **1986**, *81*, 366−374.
(24) Kolpin, D. W.; Goolsby, D. A.; Thurman, E. M. *J. Envrion. Qual.* **1995**, *24*, 1125−1132.



*Received for review August 12, 1997. Revised manuscript received August 22, 1997. Accepted September 8, 1997.*[⊗]

ES960695X

[⊗] Abstract published in *Advance ACS Abstracts*, October 15, 1997.