

Statistical methods for the analysis of left-censored variables

Statistische Analysemethoden für linkszensierte Variablen und Beobachtungen mit Werten unterhalb einer Bestimmungs- oder Nachweisgrenze

Abstract

In some applications statisticians are confronted with values which are reported to be below a limit of detection or quantitation. These left-censored variables are a challenge in the statistical analysis. In a simulation study, we compare different methods to deal with this type of data in statistical applications. These include measures of location, dispersion, association, and statistical modeling. Our simulation study showed that the multiple imputation approach and the Tobit regression lead to unbiased estimates, whereas the naïve methods including simple substitution of non-detects lead to unreliable estimates. We illustrate the application of the multiple imputation approach and the Tobit regression with an example from occupational epidemiology.

Keywords: left-censored variables, limit of detection, multiple imputation, Tobit regression, occupational exposure, welding

Zusammenfassung

In der statistischen Praxis treten immer wieder Variablen mit Werten unterhalb einer Bestimmungs- oder Nachweisgrenze auf. Diese sind linkszensiert und stellen daher eine Herausforderung für die statistische Analyse dar. Im Rahmen einer Simulationsstudie vergleichen wir Schätzmethode zur Berechnung von Lage- und Streuungsmaßen, Korrelationen und Regressionsparametern bei diesen Variablen. Unsere Ergebnisse zeigen, dass die multiple Imputationsmethode und die Tobit Regression zu unverzerrten Schätzungen führen. Naive Methoden, einschließlich der einfachen Substitution von zensierten Beobachtungen, ergeben hingegen unzuverlässige Schätzungen. Wir illustrieren die Anwendung der multiplen Imputationsmethode und der Tobit Regression anhand eines Beispiels aus der Epidemiologie der Arbeitswelt.

Introduction

In some applications statisticians are confronted with left-censored variables. For example in medical research, biomarkers can only be measured with a certain precision and sometimes measurements are observed below a limit of detection or quantitation. Similarly in epidemiology an exposure measurement might also be non-detectable. As noted by Helsel (2010) [1] and Lubin et al. (2004) [2], these non-detects in the data have to be analyzed with great care, because ignoring left-censored variables in the statistical analysis might lead to bias in estimates, improper models and, hence, to errors in the conclusions. We will use the abbreviation LOD (limit of determination) to refer to the cut-off value of censored variables. The aim of this paper is to compare and validate different

methods dealing with the statistical analysis of left-censored variables. The methods discussed in this paper are to restrict the datasets to measurable values, to substitute censored values by one-half, two-third, or one over the square root of two times LOD, to use multiple imputation of left-censored variables, and to apply Tobit regression. In a simulation study, we calculate the median and the interquartile range as measures of location and dispersion, the Pearson correlation coefficient as measure of association, and linear regression models as an example of statistical modeling. Finally we use data from the WELDOX study [3], [4], a cross-sectional study on welders, to show how the presented methods can be applied in a study from occupational epidemiology. Here we are interested in describing the concentrations of

Anne Lotz¹
Benjamin Kendzia¹
Katarzyna Gawrych¹
Martin Lehnert¹
Thomas Brüning¹
Beate Pesch¹

1. Institut für Prävention und Arbeitsmedizin der Deutschen Gesetzlichen Unfallversicherung, Institut der Ruhr-Universität Bochum (IPA), Bochum, Germany

respirable particles (RP) in welding fume and to quantify the predictors of this exposure.

Methods of the simulation study

To compare different methods for the statistical analysis of left-censored variables, a simulation study was applied constructing complete datasets of 250 observations without any censored value 1000 times each. The proportion of values below LOD was set to 10%, 25% and 50%. This was achieved by calculating the 10th, 25th and 50th percentile in each of the simulated datasets and using this number as LOD. We compared the estimates of the different methods based on the censored datasets with the estimates based on the complete datasets. The statistical analysis comprised estimates of quartiles, Pearson correlation coefficients and simple linear regression parameters. The quartiles were calculated according to the definition number two in the work of Hyndman and Fan (1996) [5]. The simulated variables for the estimation of quartiles were drawn from a rectangular distribution $U(0,20)$, a normal distribution $N(50,50)$ and a log-normal distribution $\ln N(3,2)$. The simulation datasets for the analysis of the Pearson correlation were taken from

bivariate normal distribution $N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$, where $\rho = 0.25, 0.50$, or 0.75 . The simulation datasets for the regression models followed the equation $y_i = 4 + 5x_i + e_i$, where $i = 1, \dots, 250$ and $e_i \sim N(0, \sigma^2)$. σ^2 was set to 100 or 1500. x was drawn from a rectangular distribution with $U(10,50)$. For the statistical analysis of left-censored variables the following methods were applied and compared: naïve methods, multiple imputation, calculation of an upper and lower bound of the quartiles, and Tobit regression. All calculations were performed with the SAS software, version 9.2 (SAS Institute Inc., Cary, NC, USA).

Naïve methods

Computationally easy to apply are the naïve methods for dealing with left-censored variables. The discard method restricts the analysis dataset to values above LOD. Simple substitution methods replace values below LOD with a constant. In this simulation study, we applied three different simple substitution methods: substitution by one-half times LOD ($\frac{1}{2} \cdot \text{LOD}$), substitution by two-thirds times LOD ($\frac{2}{3} \cdot \text{LOD}$) and substitution by one over the square root of two times LOD ($\frac{1}{\sqrt{2}} \cdot \text{LOD}$).

Multiple imputation

The multiple imputation method is described in detail in the work of Lubin et al. (2004) [2]. In brief, a lower and an upper bound for each observation are defined. It is assumed that the probability distribution of measurements within the lower and upper bound of the censoring interval depends only on observed data and follows the same distribution. In a first step, 100 bootstrap datasets

[6] are sampled with replacement from the original dataset. These bootstrap datasets are of the same size as the original dataset. Applying the Tobit regression approach, the cumulative distribution function G of the left-censored variable is estimated for each bootstrap sample. In each dataset, single imputations of each value below LOD are done as follows: Let z be a uniform random value between $G(\text{lower bound})$ and $G(\text{upper bound})$. Then, the single imputed value is equal to $G^{-1}(z)$. In the next step, for each of the 100 imputed datasets the point estimates of interest are calculated, for example the parameters of a linear regression model. Finally, these estimates are combined according to the method described by Little and Rubin (1987) [7] and the mean of the point estimates is the final point estimate.

Calculation of an upper and lower bound of the quartiles

By left-censored variables the range of the true quartiles is calculable. The upper bound of a percentile is obtained by substituting values below LOD with the LOD and then calculating the quartile according to the definition number two in the work of Hyndman and Fan (1996) [5]. Similarly, the lower bound of a percentile is obtained by substituting values below LOD with zero or the smallest possible value of the variable of interest and then calculating the quartile as usual. If the variable has no smallest possible value, the lower bound is not determinable.

Tobit regression

The basic idea of the Tobit regression is to treat the left-censored variable y as the outcome of a normally distributed latent variable y^* [8]. This leads to the following model equations

$$y = \max(\text{LOD}, y^*)$$

$$y^* = x\beta + u$$

with $u|x = N(0, \sigma^2)$. For $y = \text{LOD}$, the density of y is equal to the probability of observing $y < \text{LOD}$ and for $y > \text{LOD}$ the density of y is the same as the density of y^* . The parameters of the model are estimated using a maximum likelihood approach. Detailed information about the Tobit regression can be found in the work of Amamiya (1984) [9].

Results of the simulation study

The multiple imputation method showed the best results in the simulation study in estimating the median and the first and third quartile of a left-censored variable (Table 1). The discard method overestimated all quartiles. The three different simple substitution methods estimated the quartiles accurately in all cases where the percentage of values below LOD was smaller than the percentile of interest. Otherwise the percentiles were estimated with a systematic error. Similarly, the calculation of quartiles gave the true percentile when the percentage of values

Table 1: Results of a simulation study comparing different methods^a of quartile estimation of left-censored variables with the 'true' quartiles based on the complete datasets. The sample size is N=250 and entries are means of 1000 repetitions.

Percentile	Complete datasets	Percentage below LOD	Discard	$\frac{1}{2} * \text{LOD}$	$\frac{2}{3} * \text{LOD}$	$\frac{1}{\sqrt{2}} * \text{LOD}$	Multiple imputation	Calculated range ^b
Rectangular distribution (a=0; b=20)								
Median	10.0	10%	11.0	10.0	10.0	10.0	10.0	10.0
1. Quartile	5.0		6.5	5.0	5.0	5.0	5.0	5.0
3. Quartile	15.0		15.4	15.0	15.0	15.0	15.0	15.0
Median	10.0	25%	12.5	10.0	10.0	10.0	10.0	10.0
1. Quartile	5.0		8.7	2.5	3.4	3.6	4.9	0.0–5.0
3. Quartile	15.0		16.2	15.0	15.0	15.0	15.0	15.0
Median	10.0	50%	15.0	7.5	8.3	8.5	9.9	5.0–10.0
1. Quartile	5.0		12.5	5.0	6.6	7.0	5.0	0.0–10.0
3. Quartile	15.0		17.4	15.0	15.0	15.0	15.0	15.0
Normal distribution ($\mu=50$; $\sigma^2=50$)								
Median	50.0	10%	50.9	50.0	50.0	50.0	50.0	50.0
1. Quartile	45.2		46.8	45.2	45.2	45.2	45.2	45.2
3. Quartile	54.8		55.3	54.8	54.8	54.8	54.8	54.8
Median	50.0	25%	52.3	50.0	50.0	50.0	50.0	50.0
1. Quartile	45.2		48.9	22.6	30.2	32.0	45.2	≤ 45.2
3. Quartile	54.8		56.3	54.8	54.8	54.8	54.8	54.8
Median	50.0	50%	54.8	37.5	41.7	42.7	50.0	≤ 50.0
1. Quartile	45.2		52.3	25.0	33.3	35.3	45.2	≤ 50.0
3. Quartile	54.8		58.1	54.8	54.8	54.8	54.8	54.8
Log-normal distribution ($\mu=3$; $\sigma^2=2$)								
Median	20.2	10%	24.1	20.2	20.2	20.2	20.2	20.2
1. Quartile	7.8		10.7	7.8	7.8	7.8	7.8	7.8
3. Quartile	52.4		58.3	52.4	52.4	52.4	52.4	52.4
Median	20.2	25%	31.8	20.2	20.2	20.2	20.2	20.2
1. Quartile	7.8		16.2	3.9	5.2	5.5	7.7	0.0–7.8
3. Quartile	52.4		71.4	52.4	52.4	52.4	52.4	52.4
Median	20.2	50%	52.4	15.2	16.9	17.3	20.1	10.2–20.3
1. Quartile	7.8		31.8	10.1	13.5	14.3	7.9	0–20.2
3. Quartile	52.4		102.2	52.4	52.4	52.4	52.4	52.4

^a Methods employed are discarding values below a cut-off (LOD), substituting values below LOD by $\frac{1}{2} * \text{LOD}$, $\frac{2}{3} * \text{LOD}$ or $\frac{1}{\sqrt{2}} * \text{LOD}$, multiple imputation of values below a LOD and characterizing the range of the quartile.

^b If lower and upper bound are equal, only the upper bound will be presented.

below LOD was smaller than the quartile of interest. Else, a wide range was presented which included the true quartile.

The correlation coefficient was very accurately estimated by the multiple imputation method (Table 2). The other methods underestimated the correlation coefficient in all cases. An increasing percentage of values below LOD

resulted in less accurate estimates of the correlation coefficient.

Table 3 shows that the discard method overestimated the intercept and underestimated the slope of the regression line in all cases. The simple substitution methods also gave very imprecise estimates of the regression parameters. When a higher percentage of values below LOD was present or σ^2 increased, the error of the esti-

Table 2: Results of a simulation study comparing different methods^a to estimate the correlation coefficient in the presence of a left-censored variable y

ρ	Percentage below LOD	Discard	$\frac{1}{2} * \text{LOD}$	$\frac{2}{3} * \text{LOD}$	$1/\sqrt{2} * \text{LOD}$	Multiple imputation
0.25	10%	0.214	0.235	0.240	0.241	0.249
	25%	0.183	0.223	0.228	0.229	0.248
	50%	0.151	0.212	0.212	0.212	0.249
0.50	10%	0.439	0.471	0.481	0.483	0.501
	25%	0.387	0.449	0.459	0.461	0.499
	50%	0.329	0.428	0.428	0.428	0.500
0.75	10%	0.691	0.704	0.719	0.722	0.750
	25%	0.637	0.675	0.689	0.692	0.748
	50%	0.561	0.641	0.641	0.641	0.750

The datapairs (x_i, y_i) , $i = 1, \dots, 250$ are drawn from a bivariate normal distribution $N\left(\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$.

The entries are means of 1000 repetitions.

^a Methods employed are discarding values below a cut-off (LOD), substituting values below LOD by $\frac{1}{2} * \text{LOD}$, $\frac{2}{3} * \text{LOD}$ or $1/\sqrt{2} * \text{LOD}$ and multiple imputation of values below LOD.

Table 3: Results of a simulation study comparing different methods^a for estimating the parameters a and b of the simple regression model $y_i = a + bx_i + e_i$, where $i = 1, \dots, 250$, $a = 4$, $b = 5$, $e_i \sim N(0, \sigma^2)$, and Y is a left-censored variable. The entries are means of 1000 repetitions.

σ^2	Percentage below LOD	Parameter	Complete datasets	Discard	$\frac{1}{2} * \text{LOD}$	$\frac{2}{3} * \text{LOD}$	$1/\sqrt{2} * \text{LOD}$	Multiple imputation	Tobit regression
100	10%	a	4.0	7.3	-8.2	-2.0	-0.5	3.9	4.0
		b	5.0	4.9	5.3	5.2	5.1	5.0	5.0
100	25%	a	4.0	9.1	-21.7	-2.9	1.7	4.0	3.7
		b	5.0	4.9	5.6	5.2	5.0	5.0	5.0
100	50%	a	4.0	16.1	-20.2	21.3	31.4	3.7	4.4
		b	5.0	4.7	5.4	4.4	4.2	5.0	5.0
1500	10%	a	4.2	28.1	2.3	6.8	7.9	4.1	4.1
		b	5.0	4.4	5.0	4.9	4.9	5.0	5.0
1500	25%	a	4.2	53.2	-5.6	10.7	14.6	4.3	4.3
		b	5.0	3.8	5.2	4.8	4.7	5.0	5.0
1500	50%	a	4.2	99.0	-4.5	33.0	42.1	4.5	4.4
		b	5.0	2.9	5.0	4.1	3.9	5.0	5.0

^a Methods employed are discarding values below a cut-off (LOD), substituting values below LOD by $\frac{1}{2} * \text{LOD}$, $\frac{2}{3} * \text{LOD}$ or $1/\sqrt{2} * \text{LOD}$, multiple imputation of values below LOD and Tobit regression.

mates increased. The regression based on multiple imputation and the Tobit regression showed very accurate results for the regression parameters in all simulated settings. In an additional simulation run, we changed the simulation by using a fixed and identical set of x- and cut-off values for all simulated datasets. The results were very similar to the results presented in Table 3. Again only multiple imputation and the Tobit regression showed very accurate results for the regression parameters (data not shown).

Methods of the application example

The measurement of welding fume in the WELDOX study served as example for the application of the methods to analysis a left-censored variable. We used data from 215 welders recruited between May 2007 and October 2009 in a cross-sectional study. The WELDOX study was approved by the Ethics Committee of the Ruhr University Bochum and was conducted in accordance with the Helsinki Declaration. Details about the WELDOX study and technical information about the exposure measurements are described by Lehnert et al. (2012) [3] and Pesch et

al. (2012) [4]. In brief, the welders were equipped with a sampling system to determine the exposure to RP during a shift and the content of manganese (Mn) in the respirable welding fume was determined. Each workplace was documented by technicians and the descriptions included the welding technique applied (gas metal arc welding, flux-cored arc welding, tungsten inert gas welding, shielded metal arc welding, and miscellaneous techniques), the materials used (mild steel, stainless steel, and miscellaneous materials) and information about the space (confined or non-confined). Confined work spaces were locations with a strongly restricted air exchange. The efficiency of the local exhaust ventilation was rated by a team of experts. Regression models were applied to determine potential predictors of the concentrations of RP. Due to the skewed distribution of the exposure variable, it was log-transformed prior to the regression analysis. Additionally to the described statistical methods in the simulation study, a third approach based on data from respirable Mn was applied. All measurement values of respirable Mn were above LOD and were closely associated with the measurement values of RP. Therefore we used the Mn data for a multiple imputation of the concentrations of RP below LOD as described in Lehnert et al. (2012) [3]. Similar to the multiple imputation approach described above, values below LOD were imputed 100 times using a Tobit regression approach with the log-transformed respirable Mn concentrations as independent variable and the log-transformed RP concentration as dependent variable.

Results of the application example

About 30% (65 out of 215) of the measurements of the concentration of RP were below LOD. The median of the concentrations of RP was 1.29 mg/m^3 . None of the 215 measurements of respirable Mn was below LOD. The log-transformed concentration of welding fume was strongly positive correlated with the log-transformed concentrations of respirable Mn (multiple imputation: $r = 0.948$). Table 4 presents the results of the regression analysis calculated by Tobit regression, multiple imputation and multiple imputation based on Mn data. The effect estimates were in good accordance. In the models, flux-cored arc welding had about two-fold higher concentrations of RP than gas metal arc welding (Tobit regression, multiple imputation, multiple imputation based on data from respirable Mn: 2.26, 2.30, 2.25). Tungsten inert gas welding had the lowest concentrations of all welding techniques. Welding of stainless steel resulted in about 0.6 fold (0.57, 0.52, 0.55) lower concentrations than welding of mild steel. The exposure in a confined work space was about two-fold (1.79, 2.01, 1.87) higher than in a non-confined space. Efficiently used local exhaust ventilation decreased the concentrations of RP about 0.4 fold (0.43, 0.40, 0.43).

Discussion

Our simulation study confirmed that the application of naïve methods can lead to a systematic error in the estimates in the statistical analysis. In presence of a left-censored variable, multiple imputation or Tobit regression should be preferred instead. However, Tobit regression can only be applied to estimate regression models whereas multiple imputation is more flexible. We demonstrated with an example from occupational epidemiology that Tobit regression and multiple imputation both worked well to determine the main predictors of the exposure to respirable welding fume and that these methods yielded similar estimates.

Median and interquartile range are robust measures of the central tendency and dispersion. The discard method always overestimated the quartiles and should not be applied. The simple substitution methods work very well as long as the percentage of left censored values is below the percentile of interest, but in all other cases multiple imputation gives the best results. Further information about methods to measure the central tendency and dispersion of a left-censored variable is summarized by Helsel (2005) [10].

Multiple imputation is also the best method in our simulation study to estimate the correlation coefficient in the presence of a left-censored variable. The other approaches lead to biased estimates. This finding is in line with a study of Lyles et al. (2001) [11]. Chu et al. (2008) [12] explored nonparametric methods to estimate the correlation coefficient and concluded that further research is needed in this field.

In the regression analysis both, multiple imputation and Tobit regression, lead to precise estimates of the regression parameters whereas the estimates based on the discard method and the simple substitution methods were again unreliable. Even in the case of only 10% of the measurements below LOD, the intercept estimate of the naïve methods had a strong bias. Our simulation study confirmed the results of Lubin et al. (2004) [2], who stated that "... any single value to impute missing measurement data is not advisable" and that "multiple imputation of missing data is the best approach of ensuring unbiased estimates of effects and nominal CIs". Lubin et al. (2004) [4] compared Tobit regression, multiple imputation, single imputation, inserting $\frac{1}{2} \cdot \text{LOD}$, and inserting the conditional expected value $E[Y|Y < \text{LOD}]$ in a simulation study with a regression model with zero intercept and no covariates. We could show that their result holds also in regression models including a covariate. Furthermore, our simulation study demonstrated that the discard method and other substitution methods including the insertion of $\frac{2}{3} \cdot \text{LOD}$ and $\frac{1}{\sqrt{2}} \cdot \text{LOD}$ lead to biased effect estimates.

Our example from occupational medicine showed, that multiple imputation and Tobit regression can be readily applied on real datasets. The WELDOX study shows that the major determinants of the exposure to respirable welding fume are the welding technique applied, the

Table 4: Potential determinants of exposure to respirable welding fume based on a dataset of 215 male welders from Germany of the WELDOX study

Factor	N	N _{<LOD} ^a	Tobit regression		Multiple imputation		Multiple imputation based on data from respirable manganese	
			Exp($\hat{\beta}$)	95% CI	Exp($\hat{\beta}$)	95% CI	Exp($\hat{\beta}$)	95% CI
Intercept [mg/m ³]			2.74	2.22–3.38	2.64	2.06–3.38	2.72	2.12–3.49
Gas metal arc welding	78	9	1		1		1	
Flux-cored arc welding	42	0	2.26	1.62–3.15	2.30	1.56–3.40	2.25	1.52–3.32
Tungsten inert gas welding	66	47	0.15	0.10–0.23	0.14	0.09–0.21	0.18	0.12–0.27
Shielded metal arc welding	17	8	1.05	0.61–1.79	1.04	0.55–1.96	0.68	0.37–1.26
Miscellaneous techniques	12	1	1.06	0.62–1.82	1.20	0.64–2.25	1.13	0.61–2.10
Mild steel	83	5	1		1		1	
Stainless steel	122	58	0.57	0.42–0.78	0.52	0.37–0.75	0.55	0.39–0.79
Miscellaneous materials	10	2	1.01	0.57–1.79	1.13	0.59–2.16	0.83	0.43–1.58
Non-confined space	193	65	1		1		1	
Confined space	22	0	1.79	1.20–2.68	2.01	1.26–3.22	1.87	1.17–2.99
Non-efficient local exhaust ventilation	167	47	1		1		1	
Efficient local exhaust ventilation	48	18	0.43	0.30–0.61	0.40	0.27–0.60	0.43	0.29–0.64

^a N_{<LOD}: Number of observations below the limit of determination (LOD)

material used, the possibilities of air exchange at the work space and the use of efficient local exhaust ventilation. A detailed discussion can be found in Lehnert et al. (2012) [3].

Multiple imputation and Tobit regression are recommendable for the analysis of left-censored variables. On the one hand, Tobit regression needs less computational resources as multiple imputation. But on the other hand, Tobit regression is a special method to analyze regression models and multiple imputation can be applied in a variety of statistical approaches. Tobit regression is the best approach in applications where it is solely of interest to estimate regression parameters; in all other cases the multiple imputation approach should be preferred.

Conclusion

In the presence of a left-censored variable naïve methods should not be applied. We showed that these methods are inferior to the multiple imputation method in the analysis of central tendency and dispersion, correlation and in statistical modeling. Tobit regression is equally suitable as multiple imputation to estimate regression models with a left-censored dependent variable.

Notes

Acknowledgement

The WELDOX study was financially supported by the German Social Accident Insurance (DGUV). We thank the WELDOX study group and all welders having participated. We are thankful to Ying Cheng for addressing the problems of left-censored variables in her diploma thesis which was supervised by Jörg Rahnenführer, professor for “Statistical methods in genetics and chemometrics” at the Department of Statistics of TU Dortmund University, Germany.

Competing interests

The authors declare that they have no competing interests.

References

1. Helsel D. Much ado about next to nothing: incorporating nondetects in science. *Ann Occup Hyg.* 2010;54:257-62. DOI: 10.1093/annhyg/mep092
2. Lubin JH, Colt JS, Camann D, Davis S, Cerhan JR, Severson RK, Bernstein L, Hartge P. Epidemiologic evaluation of measurement data in the presence of detection limits. *Environ Health Perspect.* 2004 Dec;112(17):1691-6. DOI: 10.1289/ehp.7199

3. Lehnert M, Pesch B, Lotz A, Pelzer J, Kendzia B, Gawrych K, Heinze E, Van Gelder R, Punkenburg E, Weiss T, Mattenklott M, Hahn JU, Möhlmann C, Berges M, Hartwig A, Brüning T; Weldox Study Group. Exposure to inhalable, respirable, and ultrafine particles in welding fume. *Ann Occup Hyg.* 2012 Jul;56(5):557-67. DOI: 10.1093/annhyg/mes025
4. Pesch B, Weiss T, Kendzia B, Henry J, Lehnert M, Lotz A, Heinze E, Kafferlein HU, Van Gelder R, Berges M, Hahn JU, Mattenklott M, Punkenburg E, Hartwig A, Brüning T. Levels and predictors of airborne and internal exposure to manganese and iron among welders. *J Expo Sci Environ Epidemiol.* 2012 May-Jun;22(3):291-8. DOI: 10.1038/jes.2012.9
5. Hyndman RJ, Fan Y. Sample Quantiles in Statistical Packages. *Am Stat.* 1996;50:361-5.
6. Efron B. Bootstrap methods: Another look at the jackknife. *Ann Stat.* 1979;7:1-26. DOI: 10.1214/aos/1176344552
7. Little RJA, Rubin DB. *Statistical analysis with missing data.* New York: Wiley; 1987.
8. Tobin J. Estimation of Relationships for Limited Dependent Variables. *Econometrica.* 1958;26:24-36. DOI: 10.2307/1907382
9. Amemiya T. Tobit models: A survey. *J Econom.* 1984;24(1-2):3-61. DOI: 10.1016/0304-4076(84)90074-5
10. Helsel DR. *Nondetects and data analysis: Statistics for censored environmental data.* Hoboken, NJ: Wiley-Interscience; 2005. (Statistics in practice).
11. Lyles RH, Fan D, Chuachowong R. Correlation coefficient estimation involving a left censored laboratory assay variable. *Stat Med.* 2001 Oct;20(19):2921-33. DOI: 10.1002/sim.901
12. Chu H, Nie L, Zhu M. On estimation of bivariate biomarkers with known detection limits. *Environmetrics.* 2008;19:301-317. DOI: 10.1002/env.868

Corresponding author:

Anne Lotz

Institut für Prävention und Arbeitsmedizin der Deutschen Gesetzlichen Unfallversicherung, Institut der Ruhr-Universität Bochum (IPA), Bürkle-de-la-Camp-Platz 1, 44789 Bochum, Germany, Phone: +49-(0)234-302-4533, Fax: +49-(0)234-302-4505
lotz@ipa-dguv.de

Please cite as

Lotz A, Kendzia B, Gawrych K, Lehnert M, Brüning T, Pesch B. *Statistical methods for the analysis of left-censored variables.* *GMS Med Inform Biom Epidemiol.* 2013;9(2):Doc05. DOI: 10.3205/mibe000133, URN: urn:nbn:de:0183-mibe0001334

This article is freely available from

<http://www.egms.de/en/journals/mibe/2013-9/mibe000133.shtml>

Published: 2013-03-07

Copyright

©2013 Lotz et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by-nc-nd/3.0/deed.en>). You are free: to Share — to copy, distribute and transmit the work, provided the original author and source are credited.