2006

The Limitations due to Exposure Detection Limits for Regression Models

Enrique F. Schisterman
Albert Vexler
Brian W Whitcomb, *University of Massachusetts - Amherst*
Aiyi Liu

Published in final edited form as:
Am J Epidemiol. 2006 February 15; 163(4): 374–383.

The Limitations due to Exposure Detection Limits for Regression Models

Enrique F. Schisterman[1], **Albert Vexler**[1], **Brian W. Whitcomb**[1], and **Aiyi Liu**[1]
1*Division of Epidemiology, Statistics, and Prevention Research National Institute of Child Health and Human Development National Institutes of Health 6100 Executive Blvd. Rockville, MD 20852*

Abstract

Biomarker use in exposure assessment is increasingly common and consideration of related issues is of growing importance. Exposure quantification may be compromised when measurement is subject to a lower threshold. Statistical modeling of such data requires a decision regarding the handling of such readings. Various authors have considered this problem. In the context of linear regression analysis, Richardson and Ciampi proposed replacement of data below a threshold by a constant equal to the expectation for such data to yield unbiased estimates. Use of such an imputation has some limitations; distributional assumptions are required, and bias reduction in estimation of regression parameters is asymptotic, thereby presenting concerns to small studies. In this paper the authors propose distribution-free methods for managing values below detection limits and evaluate the biases that may result when exposure measurement is constrained by a lower threshold. The authors utilize an analytical approach as well as a simulation study to assess the effects of the proposed replacement method on estimates. These results may inform decisions regarding analytical plans for future studies as well as provide possible explanation for some amount of discordance seen in extant literature.

Keywords

Regression analysis; limit of detection; bias; biomarkers; molecular epidemiology; threshold

Abbreviations

LOD, limit of detection; LOQ, limit of quantification; ELISA, enzyme linked immunosorbent assay

The growing use of biomarkers in exposure assessment suggests the need to address issues related to their measurement. Even when levels are sufficient for measurement, some random exposure measurement error is expected, in part related to instrument precision. However, in many cases a proportion of study participants have levels at or below some experimentally determined detection limit (*dl*). Investigators are often interested in risk of negative health outcomes associated with such levels. For example, studies of serum organochlorine levels, lipophilic xenobiotics, and breast cancer have determined up to 99% of study participants to have levels below the *dl* for some toxicants under study (1).

Biomarker quantification may be compromised if instrumentation cannot detect low levels. This may occur, for example, in quantitation of immunoassays (e.g., ELISA) which require

Reprint requests to: Enrique F. Schisterman, PhD, Epidemiology Branch, Division of Epidemiology, Statistics, and Prevention Research, National Institute of Child Health and Human Development, 6100 Executive Boulevard, Room 7B03, Rockville, MD, 20852, Tel: 301-435-6893, Fax: 301-402-2084, Email:schistee@mail.nih.gov.

Grants: none to declare

antigen concentrations sufficient for binding by antibodies. Highly specific binding conditions may impair antibody sensitivity and thereby challenge quantitation of low levels (2). Alternatively, assays may detect low biomarker levels, but suffer from insufficient specificity, and measurement of exposure is hampered by background. The detection limit is often determined as a function of observed variance for a series of blanks; the terms 'limit of detection (LOD)' and 'limit of quantification (LOQ)' generally correspond to three and ten, respectively, standard deviations from serial measurement of blanks (3). As such, numerical data are observable above and below the *dl*; even among values above the threshold it may not be possible to clearly delineate between those that are "real" and those that are not. Data below the threshold are often reported by laboratories as nondetects (*ND*), and the data analyst or epidemiologist is limited to this qualitative assessment.

Statistical modeling of this data requires decisions regarding its handling (4,5). Conventional approaches include: omission, resulting in a truncated dataset; imputation of a constant like the *dl* or fraction thereof (e.g. $dl/2$, $dl/\sqrt{2}$), or; the observed values may be used directly or indirectly (4,5,6,7). Many of these imputations have their origins in well behaved distributions such as Normal (in the case of $dl/2$) and lognormal (in the case of $dl/\sqrt{2}$), and will yield correct inferences if these distributional assumptions are not grossly violated. Lubin et al. (5) propose a multiple imputation approach to handling nondetects when the exposure distribution can be assumed. Richardson and Ciampi (7) developed a coefficient of bias to linear regression coefficient estimates when exposure is measured with a detection threshold and random error, and proposed replacement of below-threshold data by the expectation for such data (i.e., $E[x|x < dl]$) to yield unbiased estimates. Application of this theory to practice also requires investigators to assume an exposure distribution function. In contrast to these approaches, there has been comparatively little attention toward implicitly and explicitly non-parametric approaches to measurement with a threshold

In this paper the authors propose distribution-free methods for managing values below the *dl* and evaluate biases that result when exposure measurement is constrained by a lower threshold. Results from an analytical approach and those of a simulation study assessing the proposed replacement method are described. The proposed method allows investigators to relax assumptions (e.g., distributional, asymptotic) necessary for use of other approaches. These results may inform decisions by investigators regarding appropriate analytical plans for future studies and provide possible explanation for discordance seen in current literature.

Statement of the problem and analytical solution

Let the observed continuous outcome, *Y*, satisfy the following linear regression model

$$Y_i = \alpha + \beta x_i + e_i, \tag{1}$$

with exposure variable x_i, random noise e_i, and regression parameters α and β. However, *x* is not observed. A lower threshold, *dl*, interferes with measurement of low exposure levels. In a simple case, we observe *z*, which equals either *x* or *ND*, according to the following

$$\text{for all } x > dl, \quad z = x$$

$$\text{for all } x \leq dl, \quad z = ND$$

Alternately, when the explanatory variable is less than dl, there is quantitative random noise, ζ rather than the qualitative response '*ND*'.

In this setting, the sample observations are $\{Y_1, \ldots, Y_n, z_1, \ldots, z_n\}$. Without loss of generality, Y_i and z_i can be assumed to be scalars. This model can be considered in a more general context where the exposure is measured with error, η. Thus, the linear regression model is

$$Y_i = \alpha + \beta z_i + e_i, \tag{2}$$

and

$$z_i = (x_i + \eta_i) I\{x_i + \eta_i \geq dl\} + \xi_i I\{x_i + \eta_i < dl\}, \quad i = 1, \ldots, n$$

is the exposure with measurement error, $I\{\bullet\}$ is an indicator function (1 if $\{\bullet\}$ is true and 0 otherwise), and e_i, η_i, ζ_i are independent random disturbance terms related to regression error, measurement error and detection limit error with $f_\varepsilon(u)$, $f_\eta(u)$ and $f_\zeta(u)$ densities respectively and $E(e_i) = 0$, $var(e_i) = \sigma_e^2$.

The accuracy of regression parameter estimates depends upon the analytic approach to below-threshold values. One may consider substitution of observed z by z' whereby,

$$z'(a) = \begin{cases} x \text{ if } x \geq dl \\ a \text{ if } x < dl \end{cases}.$$

In this paper, least squares estimation was used to determine an a that may be used in place of censored data for unbiased estimation of regression parameters. Non-numerical as well as numerical instrument response for below-threshold measurements were addressed. Additionally, the circumstances of instrument noise bounded by the detection limit (where the probability of values above the limit being due to error alone is approximately zero), and unbounded (where values above the limit will be a mix of noise alone and signal) were considered. The authors apply the contexts of linear regression models with both known and unknown intercept and, additionally, provide an extension for application to logistic regression models.

1. Non-numerical indicator below the detection limit

In certain situations exposure variable values below *dl* are reported as '*ND*'. For example, this occurs when instrumentation is set to observe a threshold, and laboratory supplied datasets include this notation for below-threshold observations. Such responses are clearly distinguished from numerical data; a decision regarding their management is required. Substitution of unobserved z_i is performed according to the following

$$z'_i = (x_i + \eta_i) I\{x_i + \eta_i \geq dl\} + a\, I\{x_i + \eta_i < dl\}, \quad i = 1, \ldots, n, \tag{3}$$

1.1 No intercept, α, models—In cases where the intercept is known, such as reliability studies (when it is equal to zero), investigators may exclude an intercept term from models. This may also occur when modeling determinants of change from baseline, or centering data ([8,9]). Even when the intercept of a general model is unknown, transformations are available to reduce its influence on estimation of other regression parameters if the intercept is not of primary interest. In each of these circumstances the intercept may be set to a constant and the model becomes,

$$Y_i - \alpha = \beta z_i + e_i.$$

Without loss of generality, assume the intercept equals zero.

Applying least squares fitting to unknown parameter β in the model depicted by equation 2 yields an estimator in the widely known form

$$\hat{\beta}_n = \frac{\sum_{i=1}^{n} Y_i z'_i}{\sum_{i=1}^{n} (z'_i)^2} = \beta + B_n + E_n, \tag{4}$$

where B_n

$$B_n = \beta \frac{\sum_{i=1}^{n} \left(x_i - z'_i\right) z'_i}{\sum_{i=1}^{n} (z'_i)^2} = \frac{\beta}{\sum_{i=1}^{n} (z'_i)^2} \sum_{i=1}^{n} \left(- x_i \eta_i I\{x_i + \eta_i \geq dl\} + a x_i I\{x_i + \eta_i < dl\} - \eta_i^2 I\{x_i + \eta_i \geq dl\} - a^2 I\{x_i + \eta_i < dl\}\right) \tag{5}$$

signifies the bias to the regression parameter, and E_n

$$E_n = \frac{\sum_{i=1}^{n} e_i z'_i}{\sum_{i=1}^{n} (z'_i)^2},$$

signifies noise with expectation zero. Assume that exposure, x, has some distribution, F_x, and is independent of e and η. Therefore, directly from equation 5, with $\eta_i \equiv 0$, there are two solutions that yield an unbiased estimator $\hat{\beta}_n$ (i.e. $B_n = 0$)—when $a = a_1$ or $a = a_2$ (for details and solution without the restriction on η_i see Appendix 1),

$$a_1 = \frac{\sum_{i=1}^{n} x_i I\{x_i < dl\}}{\sum_{i=1}^{n} I\{x_i < dl\}} \tag{6a}$$

or

$$a_2 = 0 \tag{6b}$$

Asymptotically (as $n \to \infty$), a_1 is approximately equal to $E(x|x < dl)$ under the condition $E(x^2) < \infty$. Replacement of values below the detection limit by this value was proposed by Richardson and Ciampi (7), however as shown in equation 6b, the solution is not unique. There are important distinctions between these two methods; use of equation 6a assumes knowledge of the distribution function of x, whereas for $a = 0$ no distributional assumptions are needed.

Additionally, there are differences regarding the variance of the bias resulting from detection limit error. The variance of bias, $\mathrm{var}\left(\hat{\beta}_n \mid x\right)$, under a_1 (equation 6a) is shown as,

$$\mathrm{var}\left(\hat{\beta}_n \mid x; a_1\right) = \beta^2 o_n(1) + \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n} \left(x_i\right)^2 I\{x_i \geq dl\} + (E(x \mid x < dl))^2 \sum_{i=1}^{n} I\{x_i < dl\}}, \tag{7a}$$

$$\lim_{n \to \infty} \sup_{\beta} \mathrm{var}\left(\hat{\beta}_n \mid x; a_1\right) = \infty$$

with asymptotic properties represented by the function $o_n(1)$, which approaches 0 as n goes to infinity. Estimates will have some amount of bias for a fixed sample size. Conversely, under the second solution, (equation 6b), the variance is derived as,

$$\operatorname{var}\left(\hat{\beta}_n \mid x; a_2\right) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n} \left(x_i\right)^2 I\left\{x_i \geq dl\right\}}. \tag{7b}$$

Note that the variance solution using a_1 comprises two terms, the first being a proportion of the parameter of interest, dependent upon the sample size. The solution using a_2 is invariant of the unknown parameter β, however for "small" β, the variance may exceed that when a_1 is employed.

1.2 Models that estimate the intercept, α—Commonly, investigators have no foreknowledge of the intercept value or a need to center their data and the regression intercept must be estimated. As previously, consider the situation where non-numerical noise is returned below *dl*. As in the previous discussion, η can be assumed to be zero, and hence the imputation is $z'(a) = x\, I\{x \geq dl\} + a\, I\{x < dl\}$. Using least squares estimation, the slope parameter estimator based on z′ instead of x can be shown as,

$$\beta_n = \frac{\sum_{i=1}^{n} \left(Y_i - \frac{1}{n}\sum_{j=1}^{n} Y_j\right) z'_i}{\sum_{i=1}^{n} \left(z'_i - \frac{1}{n}\sum_{j=1}^{n} z'_j\right)^2}. \tag{8}$$

Again, bias results from using z′ for estimation instead of the true explanatory variables (see Appendix 2 for equations for bias). As previously, there are two solutions for *a* to yield an unbiased estimator shown in equation 8. The bias is zero if and only if

$$a = \frac{\sum_{i=1}^{n} x_i I(x_i < dl)}{\sum_{i=1}^{n} I(x_i < dl)} \tag{9a}$$

or

$$a = \frac{\sum_{i=1}^{n} x_i I(x_i \geq dl)}{\sum_{i=1}^{n} I(x_i \geq dl)} \tag{9b}$$

As in the previous circumstance, the Richardson and Ciampi solution (equation 9a) is valid here. Since $x\, I[x<dl]$ is not observed, application of the this solution requires some distribution, F_x, be assumed for determination of the mean of the missing values. Additionally, the asymptotic result $a \cong E(x \mid x<dl)$, as $n \to \infty$ must be assumed valid. However, the second solution (equation 9b) requires neither distributional nor asymptotic assumptions—the replacement value (*a*) may be calculated.

2. Numerical noise below the LOD

Whereas the previous sections concerned the instrumentation response of '*ND*', numerical information may be available for data below *dl*. Importantly, if detection limit error is known to be less than the *dl* itself (ζ is reasonably bounded by the *dl*), observations below *dl* may be identified as such and this becomes a special case of section 1. In this case investigators may follow the previously discussed methodologies for models with or without intercept.

When ζ is unknown or is known to have greater magnitude than *dl*, the observations due to noise can not be easily identified; the z_i are observed both above and below *dl* for all individuals and those with detection limit error are not easily discernable from those free of this error. Formally, suppose $\Pr(\zeta > dl) > 0$. Under the described scenario of detection limit error, this may be the case if the detection limit is set as two, rather than three (LOD) or ten (LOQ) standard deviations of noise. In this circumstance there is no simple approach to determining an optimal imputation, however numerical approaches for models both with and without intercept are shown in Appendix 3.

Remark: Alternatively, detection limits are occasionally determined based solely on the distribution of additive random error (or concurrent detection of some contaminant). Consider true exposure, *x*, and all concurrently detected other, ω; all samples may be reasonably thought of as the sum of these two components (i.e., $x + \omega$). When all levels of *x* are subject to the same source of error, which is independent of *x*, then the proposed imputations remain valid, however, alternately the observed values may be used for analysis in combination with commonly used methods for handling random measurement error ([10,11]).

Motivating example

We consider the association between total cholesterol and serum vitamin E in a healthy population using a population-based sample of randomly selected residents of two counties in western New York State, 35 to 79 years of age. After exclusions, a total of 857 men and women were selected for analysis. Blood specimens were analyzed for routine chemistry, hematology and a number of chronic disease and nutritional factors as well as serum vitamin E levels.

For cholesterol and vitamin E, all observations were measured above the *dl*. Regression analysis suggests a linear association between serum cholesterol and serum vitamin E($\hat{\beta} = 4.07$, $R^2 = 0.14$, $p<0.0001$). However, if one stipulates that 30% of serum vitamin E levels are below the *dl*, the question arises how to treat these "unobserved" values.

To this end estimates were compared from linear regression of serum cholesterol on serum vitamin E. For exposure data, these models used; 1. All available exposure data ("gold standard"), or; replacement of all data below the imposed threshold with 2. The mean vitamin E of all data below the threshold, or; 3. The mean of all data above the threshold, or; 4. Zero. Importantly, both the first and second circumstances rely on distributional assumptions that would be unverifiable under a true detection limit. Models were run for known intercept as well as for estimation of the intercept.

Table 1 displays results of the known intercept regression. Replacement of sub-threshold data by the average of sub-threshold data yielded estimates almost identical to those when all data were used. However, under usual conditions direct calculation of this quantity is not possible, and it must be estimated assuming some distribution. Conversely, replacement by zero requires no such assumptions and resulted in estimates not statistically different from those under the ideal scenario of no threshold. Table 2 displays results of linear regression models estimating both slope and intercept. Use of the proposed average of above-threshold data for replacement of missing data yielded good estimates of parameters, both intercept and slope. As with use of zero for no intercept models, use of this imputation for estimation is non-parametric and requires no distributional knowledge. Moreover, estimates of slope were hampered by slight departures from the assumptions of linear regression of the data. Use of zero for replacement of missing data resulted in estimates statistically different from those under no constraint by a detection threshold.

Logistic regression

Frequently, investigators face exposure measurement thresholds when investigating binary outcomes (e.g, presence or absence of disease) using logistic regression models for evaluation of risk. These models are often employed for analysis of case-control study data. In this context, investigators rarely have interest in intercept estimates, interpretation of which is generally meaningless (9). In such situations the discussion for condition 1A applicable, and bias is minimized when a equals zero. This is shown empirically in a simulation study described in the next section. A more detailed discussion and proof is shown in *Appendices 4 and 5*.

The solution under maximum likelihood estimation is complicated even with known intercept. When the intercept is unknown, solutions to two non-linear equations are required (Appendix 4, similar to equation 17). A solution can be determined only in rare cases and where strong assumptions regarding the observed data are employed, and the proposed methodology applied. However, this problem is beyond the scope of this paper.

Monte Carlo simulation study

We extend previous work (7) to apply a solution to the detection limit problem to binary disease variables and use of a logit-linear model. To evaluate the effects of detection limit bias on logistic regression, 10,000 datasets were simulated with $n = 300$, outcome = Y and measured exposure z to comprise the observations $\{(Y_i, z_i(a)), i=1, \ldots,n\}$, following the model,

$$P\{Y_i = 1 \mid x_i\} = (1 + exp(-c - \beta x_i))^{-1},$$

$$z_i(a) = x_i I\{x_i \geq dl\} + a\ I\{x_i < dl\}.$$

We evaluated two distributions for the exposure of interest. A bimodal distribution was chosen to exemplify a poorly behaved density, and gamma, a skewed distribution that is often assumed for biomarkers (as shown in Figure 1). After data for true exposure were generated, several detection thresholds were applied. The observed exposure after replacement of values below detection limit was determined for both $a_1\ \{a = E(x|x<dl)\}$ and $a_2\ \{a = 0\}$ according to the above, as well as for the common imputations $a = dl$ and $a = dl/2$. Bias and variance were assessed by comparing results based on use of true exposure to those subject to detection limits.

Table 3 displays the simulated Monte Carlo variance of the maximum likelihood estimator $\hat{\beta}$, $Var(a) = \overline{(\beta - \hat{\beta})^2}$, and the Monte Carlo average bias, $Bias(a) = \overline{\beta - \hat{\beta}}$, for each evaluated a under the bimodal distribution for exposure. The specified values for dl correspond to the circumstances where 25%, 50%, and 75% of data are below the threshold. In most cases imputation of zero resulted in minimally biased estimates, with Monte Carlo relative biases comparable to those observed under use of the $E(x|x<dl)$. Imputation of dl and $dl/2$ resulted in substantially greater biases. Table 4 displays the results of simulations where exposure is gamma distributed. Results for the gamma distributed exposure were similar to those for the bimodal distribution; imputation of zero and $E(x|x<dl)$ performed similarly well while use of dl or $dl/2$ resulted in substantially biased estimates.

Discussion

We have considered analysis of exposures subject to a lower threshold, a circumstance frequently confronted in epidemiological studies evaluating relations between laboratory data and health outcomes. Appropriate management of data below such a threshold is imperative for proper conclusions, and available information may not be sufficient for use of parametric approaches. Using an analytical approach to finding unbiased estimators, substitution of zero

for sub-threshold/missing data was observed to minimize bias when using no/known intercept linear regression models; when estimating the intercept, imputation of the average of data above detection limits yielded optimal estimates. Data from a population-based study was used to display the effects of the proposed solutions. Additionally, Monte Carlo simulations were used to demonstrate solutions applied to logistic regression where the intercept is known and/or excluded from models, which is appropriate for case-control study data. Imputation of zero performed optimally in these circumstances.

While this result may seem illogical, the solutions for linear regression may be understood intuitively. Imputation of zero for exposure values below detection limits essentially eliminates the leverage of those data on estimation when concerned solely with the slope of the regression line (i.e., the no intercept model). If the data perfectly meet the assumptions for this model then use of the subset of data above the threshold will yield identical estimates to those produced using all data when they are not subject to a threshold. Similarly, imputation of the expected value for all observed data (i.e. $z > dl$) eliminates the leverage of those data points on estimation of the regression line slope while also allowing for estimation of the intercept; imputation of zero clearly has implications for estimation of the linear regression intercept.

The real data example results display the effectiveness of these imputations as well as certain limitations. Estimates under the proposed imputations were not statistically different from those subjected to no detection threshold; however, nor were they identical. In combination with normal sampling variability, when data do not perfectly conform to the assumptions for linear regression this is expected to be the case. Non-linear relations can be poorly represented when a subset of data is observable. Additionally, when error is not Normally and/or identically distributed, data with disproportionate leverage on estimation may be subject to the threshold and the parameter estimates unequal to those without a detection threshold. Importantly, in these cases the estimates under a no-threshold linear model are subject to the same limitations; data transformation techniques should be considered. Under the assumptions of linear regression, imputation of zero for no intercept or the mean of observed values for intercept models are appropriate for investigators opting for non-parametric approaches.

Various approaches to management of data measured with a lower limit including imputations derived from the LOD itself, such as $dl/2$ or $dl/\sqrt{2}$, have been used in laboratory and data analysis settings ([4,6,12]). A multiple imputation approach based on bootstrapping has also been proposed (5). The utility of these approaches depends upon proper determination of the exposure distribution function distribution. Recent work showed that use of $E(x|x<dl)$ for those data below detection limits allows unbiased estimation of linear and, under certain conditions, logistic regression parameters; however, this approach requires assumptions regarding the underlying exposure distribution. We have shown that unbiased estimates may also be obtained if data below the detection limit are replaced by zero for no intercept models and by $E(x|x>dl)$ for models estimating the intercept; use of these methods requires no distributional assumptions.

We performed Monte Carlo simulations of bimodal and gamma exposures with a logit-linear relation to the outcome and stipulated varying proportions of the data to be below a detection threshold. Under a no-intercept model ($\alpha = 0$) with slope parameter equal to 0.3 ($OR = 1.35$) imputation of zero resulted in similarly minimally biased results as use of $E(x|x<dl)$. This approach may be useful for logistic regression models when available information makes distributional assumptions difficult to validate, thereby extending upon previously published work (7). We have presented analytic solutions for linear regression and no-intercept logistic regression; the solutions presented here are not generalizable to other complex modeling situations.

Conclusion

Measurement of laboratory data can be limited by a detection threshold when sample exposure levels are in the range of that threshold. When a meaningful proportion of data falls below the detection threshold there is a need for simple yet valid approaches to handling the data. When investigators are confident of the validity of distributional assumptions parametric methods may be used. We have demonstrated approaches that require no distributional assumptions and are easily applied to achieve unbiased estimates. In all cases, sensitivity analyses to evaluate the chosen approach are recommended. Importantly, this paper has focused on analytic studies primarily interested in estimating linear relations between signal and response. Nevertheless, investigators should evaluate the nature of the data, detection limit, as well as the parameter to be estimated when choosing the optimal method for management of sub-threshold data.

Acknowledgements

We would like to thank Dr. Maurizio Trevisan for providing access to the MI Life study dataset. This research was supported by the Intramural Research Program of the NIH, Epidemiology Branch, DESPR, NICHD.

References

1. Cooper GS, Savitz DA, Millikan R, et al. Organochlorine exposure and age at natural menopause. Epidemiology 2002;13:729–733. [PubMed: 12410018]
2. Karasckiewicz JW. Critical factors in immunoassay optimization. Available online at http://www.kpl.com/docs/techdocs/BENCH2.PDF Accessed July 1, 2005.
3. Keith LH, Crummett W, Deegan J, et al. Principles of environmental analysis. Anal Chem 1983;55:2210–2218.
4. Helsel D. Nondetects and data analysis: Statistics for censored environmental data. Hoboken, New Jersey: John Wiley & Sons, Inc., 2005.
5. Lubin JH, Colt JS, Camann D, et al. Epidemiologic evaluation of measurement data in the presence of detection limits. Environ Health Perspect 2004;112:1691–1696. [PubMed: 15579415]
6. Hornung RW, Reed L. Estimation of average concentration in the presence of nondetectable values. Appl Occup Environ Hyg 1990;5:46–51.
7. Richardson DB, Ciampi A. Effects of exposure measurement error when an exposure variable is constrained by a lower limit. Am J Epidemiol 2003;157:355–363. [PubMed: 12578806]
8. Draper NR, Smith H. Applied regression analysis. New York, NY: John Wiley & Sons, Inc., 1998.
9. Greenland S. Introduction to regression modeling. In: Rothman KJ, Greenland S, eds. Modern epidemiology, 2nd edition. Philadelphia: Lippincott Williams & Wilkins, 1998:401–432.
10. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Stat Med 1989;8:1051–1069. [PubMed: 2799131]
11. Spiegelman D, McDermott A, Rosner B. Regression calibration method for correcting measurement-error bias in nutritional epidemiology. Am J Clin Nutr 1997;65:1179S–1186S. [PubMed: 9094918]
12. Finkelstein MM, Verma DK. Exposure estimation in the presence of nondetectable values: another look. AIHAJ 2001;62:195–198. [PubMed: 11331991]

Appendix 1

Application of the definition for z' from equation 3,

$$z'_i = (x_i + \eta_i) I\{x_i + \eta_i \geq d\} + a\, I\{x_i + \eta_i < d\}, \quad i = 1, \ldots, n$$

to equation 5 for the least squares estimator with detection limit error,

$$B_n = \beta \frac{\sum_{i=1}^{n} \left(x_i - z'_i\right) z'_i}{\sum_{i=1}^{n} (z'_i)^2}$$

yields the following,

$$= \frac{\beta}{\sum_{i=1}^{n} (z'_i)^2} \sum_{i=1}^{n} \left(- x_i \eta_i I\{x_i + \eta_i \geq d\} + a x_i I\{x_i + \eta_i < d\} - \eta_i^2 I\{x_i + \eta_i \geq d\} - a^2 I\{x_i + \eta_i < d\} \right)$$

The bias of estimator $\hat{\beta}_n$ is zero (i.e. $B_n = 0$) if and only if the numerator of B_n (equation 5) is equal to zero:

$$a^2 \frac{1}{n} \sum_{i=1}^{n} I\{x_i + \eta_i < d\} - a \frac{1}{n} \sum_{i=1}^{n} x_i I\{x_i + \eta_i < d\} + \frac{1}{n} \sum_{i=1}^{n} (\eta_i^2 I\{x_i + \eta_i \geq d\} + x_i \eta_i I\{x_i + \eta_i \geq d\}) = 0 \quad \text{(A1.1)}$$

Asymptotically (as $n \to \infty$), under the restrictions: 1. $E(x^2) < \infty$; 2. $E(|\zeta|)^3 < \infty$, and; 3. $E(x\zeta)^2 < \infty$; the solution of the equation 5 is almost surely equal to the solution of

$$\begin{aligned} & a^2 Pr\{x_1 + \eta_1 < d\} - a\, E(x_1,\ x_1 + \eta_1 < d) \\ & + E(\eta_1^2,\ x_1 + \eta_1 \geq d) + E(x_1 \eta_1,\ x_1 + \eta_1 \geq d) = 0. \end{aligned} \quad \text{(A1.2)}$$

The preceding determines an "a" such that bias, B_n, equals zero, asymptotically. When detection limit error can be clearly delineated, i.e., $Pr(\zeta < d) \equiv 1$, a may be either $a_1 = E(x|x < d)$ or $a_2 = 0$.

Appendix 2

For estimation of the regression parameter and unknown intercept when exposure is measured with a threshold,

$$B_n = \beta \frac{\sum_{i=1}^{n} z'_i \left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_i\right) - \sum_{i=1}^{n} \left(z'_i - \frac{1}{n}\sum_{j=1}^{n} z'_j\right)^2}{\sum_{i=1}^{n} \left(z'_i - \frac{1}{n}\sum_{j=1}^{n} z'_j\right)^2}.$$

The numerator of B_n may be demonstrated as

$$\begin{aligned} & \sum_{i=1}^{n} z'_i \left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_i\right) - \sum_{i=1}^{n} \left(z'_i - \frac{1}{n}\sum_{j=1}^{n} z'_j\right)^2 = \\ & = \sum_{i=1}^{n} x_i^2 I(x_i \geq dl) + a \sum_{i=1}^{n} x_i I(x_i < dl) - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i I(x_i \geq dl) \\ & - \frac{a}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} I(x_i < dl) - \sum_{i=1}^{n} x_i^2 I(x_i \geq dl) - a^2 \sum_{i=1}^{n} I(x_i < dl) \\ & + \frac{1}{n}\left(\sum_{i=1}^{n} x_i I(x_i \geq dl)\right)^2 + \frac{2a}{n} \sum_{i=1}^{n} x_i I(x_i \geq dl) \sum_{i=1}^{n} I(x_i < dl) + \frac{a^2}{n}\left(\sum_{i=1}^{n} I(x_i \geq dl)\right)^2 \\ & = -\left[a \sum_{i=1}^{n} I(x_i < dl) - \sum_{i=1}^{n} x_i I(x_i < dl)\right]\left[\frac{a}{n} \sum_{i=1}^{n} I(x_i \geq dl) - \frac{1}{n} \sum_{i=1}^{n} x_i I(x_i \geq dl)\right]. \end{aligned} \quad \text{(A2.1)}$$

Hence, the bias is equal to zero if the right side of equation A2.1 is equal to zero.

Appendix 3

For models not estimating the intercept with a numerical response below dl rather than 'ND', first define a substitution of observed z

$$z'_i(a) = z_i I\{z_i \geq dl'\} + a\, I\{z_i < dl'\}, \quad \text{for} \quad i = 1, \ldots, n, \tag{A3.1}$$

where *dl′* is the fixed detection limit. The bias of the least squares estimator is given by

$$B_n = \beta \frac{\sum_{i=1}^{n} (x_i - z'_i) z'_i}{\sum_{i=1}^{n} (z'_i)^2}. \tag{A3.2}$$

In this case, bias B_n is asymptotically close to 0 if *a* and *d′* are solutions of the equation

$$E(z'_1 x_1) - E(z'_1)^2 = 0. \tag{A3.3}$$

Solving *equation 12* requires the distribution functions of x_i, η_i and ζ be assumed, and may be performed numerically by grid searching *a* and *dl′* such that the left side of *equation 12* is approximately equal to zero.

When the density function of regression error is known and well behaved, then we may consider maximum likelihood estimation of β

$$\hat{\beta}_n = \arg \max_b \prod_{i=1}^{n} f_\varepsilon(Y_i - b z'_i) \tag{A3.4}$$

and the consideration of bias reduction from sections 1.1 and 1.2 is similarly relevant here.

To address a numerical response below *dl* when estimating the intercept, the replacement defined in A3.1 is employed. Using the proposed method, in the case where the distributions of all random variables in equation 1 are known, the unbiased least square estimator of parameter β based upon the sample{Y,z′} is obtained by applying *a* and *d′* such that

$$E[z'(x - E(x)] - E[z' - E(z')]^2 = 0.$$

Appendix 4

Note that, asymptotically, by the definition of least squares slope parameter estimators under the detection limit problem (e.g. equation A1.2), we attain $E\hat{\beta}_n \rightarrow \lambda\beta$, where $\lambda = \frac{\mathrm{cov}(x, z')}{\mathrm{var}(z')}$. This form is similar to an asymptotic property of the ordinary least squares estimator of the parameters from linear regression with additive measurement error (7).

Appendix 5

The logistic regression model with exposure measurement error is given by

$$\begin{aligned} &P(Y_i = 1 \mid x_i) = \left(1 + e^{-\beta z_i}\right)^{-1}, \\ &z_i = (x_i + \eta_i) I\{x_i + \eta_i \geq d\} + \xi_i I\{x_i + \eta_i < d\}, \quad i = 1, \ldots, n, \end{aligned} \tag{A5.1}$$

where Y_i, $i=1, \ldots, n$ are Bernoulli variates whose exact distribution depends on the predictor x_i. Let $x_i + \eta_i$ be observed, if $x_i + \eta_i \geq d$, hence we assume a substitution of observed z_i by $z_i' = (x_i + \eta_i)\, I\{ x_i + \eta_i \geq d \} + a\, I\{ x_i + \eta_i < d\}$, $i=1,\ldots,n$. Applying the maximum likelihood method we obtain the estimator of β by solving the following

$$\sum_{i=1}^{n} z_i' \left(Y_i - \frac{e^{\hat{\beta}_n z_i'}}{1 + e^{\hat{\beta}_n z_i'}} \right) = 0.$$

It follows from the usual Taylor expansion that

$$\hat{\beta}_n - \beta = \left[\sum_{i=1}^{n} z_i' \left(Y_i - \frac{e^{\beta x_i}}{1 + e^{\beta x_i}} \right) + \beta \sum_{i=1}^{n} z_i' \left(x_i - z_i' \right) Q_i \right] \left[\sum_{i=1}^{n} \left(z_i' \right)^2 Q_i \right]^{-1},$$

$$Q_i \in \left(\frac{e^{\beta x_i}}{\left(1 + e^{\beta x_i}\right)^2}, \frac{e^{\hat{\beta}_n z_i'}}{\left(1 + e^{\hat{\beta}_n z_i'}\right)^2} \right). \tag{A5.2}$$

Thus, in this case, the detection limit error bias is defined by

$$B_n = \frac{\beta \sum_{i=1}^{n} z_i' \left(x_i - z_i' \right) Q_i}{\sum_{i=1}^{n} \left(z_i' \right)^2 Q_i} \tag{A5.3}$$

Therefore, even though $\eta_i = 0$, $i = 1, \ldots, n$, we obtain

$$B_n = \frac{\beta \sum_{i=1}^{n} \left(a x_i I\left(x_i < d \right) - a^2 I\left(x_i < d \right) \right) Q_i}{\sum_{i=1}^{n} \left(z_i' \right)^2 Q_i}$$

where B_n is zero only if a is zero, because generally Q_i is dependent on i, and target unknown β, and unobserved x_i.

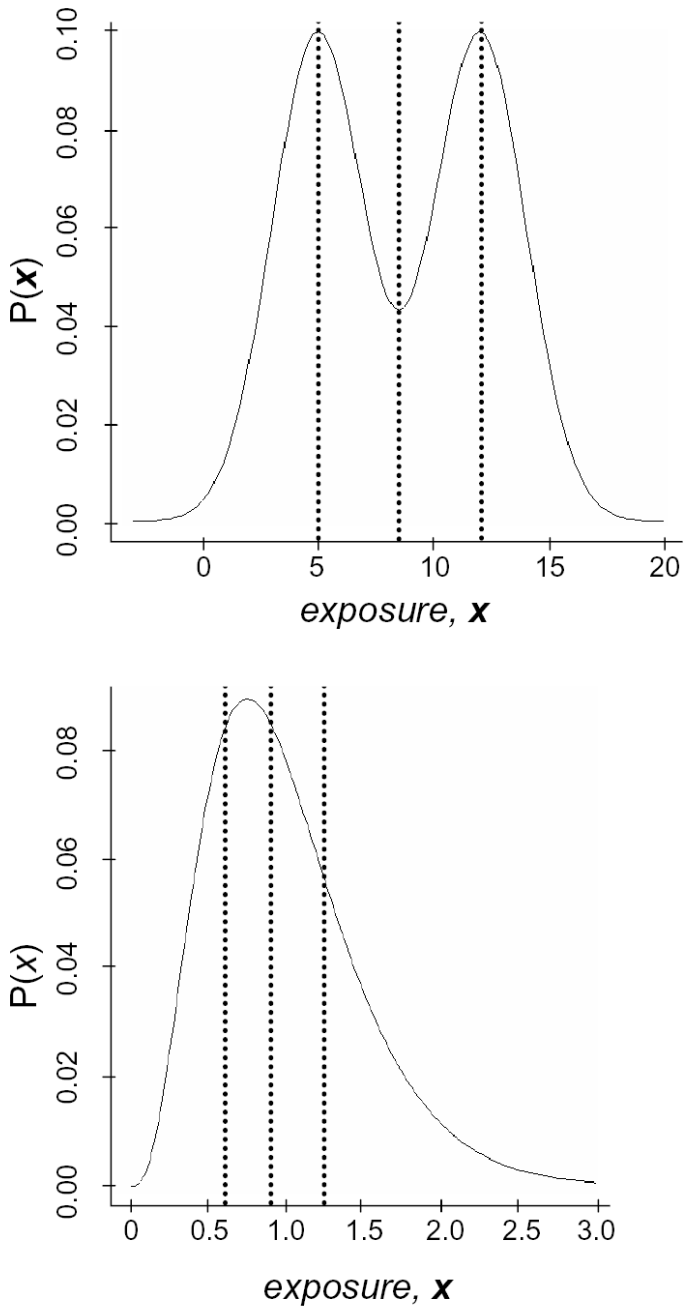

Figure 1.
Probability distribution functions for Monte Carlo simulation study; exposure distributed bimodal Normal (left panel) and gamma (right panel)
Caption: Dotted lines indicated the values utilized for the threshold corresponding to the 25^{th}, 50^{th} and 75^{th} percentiles.

TABLE 1

Coefficient estimates from linear regression with known intercept of serum cholesterol on serum vitamin E with values below the LOD replaced by the average *(x|x<dl)* and zero.

	All values (reference)	***Replace by Average (x|x<dl)***	***Replace by Zero***
$\hat{\beta}$	4.07	4.06	4.19
S.E. ($\hat{\beta}$)	0.14	0.14	0.15
$\hat{\beta}$ (95% CI)	(3.80, 4.34)	(3.79, 4.33)	(3.90, 4.48)

TABLE 2

Coefficient estimates from linear regression of serum cholesterol on serum vitamin E with values below the LOD replaced by the average *(x|x<dl)*, average *(x|x>dl)* and zero.

	All values (reference)	***Replace by Average (x| x<dl)***	***Replace by Average (x| x>dl)***	***Replace by Zero***
$\hat{\alpha}$	196.88	197.39	204.53	219.76
S.E. $(\hat{\alpha})$	5.21	5.26	7.51	3.35
α (95% CI)	(186.67, 207.09)	(187.08, 207.7)	(189.81, 219.25)	(213.19, 226.33)
$\hat{\beta}$	4.07	4.03	3.02	2.89
S.E. $(\hat{\beta})$	0.36	0.36	0.46	0.24
β (95% CI)	(3.36, 4.78)	(3.32, 4.74)	(2.12, 3.92)	(2.42, 3.36)

TABLE 3

Logistic regression simulation study with bimodal Normal distributed exposure; bias and variance of the estimator under replacement value, *a*, equal to *0*, *E(x|x<dl)*, *dl* and *dl*/2

			Bias (Monte Carlo estimator of β) *a*=				Variance (Monte Carlo estimator of β) *a*=			
β	***d***	***% x < dl***	***0***	***E(x|x<d)***	***dl***	***dl/2***	***0***	***E(x|x<d)***	***dl***	***dl/2***
0.3	5	25%	0.00026	0.00019	0.00117	0.00057	0.00025	0.00025	0.00027	0.00026
	8.5	50%	0.00068	0.00059	0.01407	−0.00082	0.00026	0.00025	0.00043	0.00026
	12	75%	0.00117	−0.00796	0.05360	0.00197	0.00039	0.00038	0.00307	0.00046
0.5	5	25%	−0.00045	−0.00075	0.00276	−0.00127	0.00024	0.00024	0.00024	0.00024
	8.5	50%	−0.00062	−0.00518	0.04611	−0.00833	0.00028	0.00031	0.00225	0.00037
	12	75%	−0.00183	−0.04409	0.12608	−0.08398	0.00068	0.00239	0.01598	0.00794
0.7	5	25%	−0.00230	−0.00476	0.01814	−0.00807	0.00068	0.00068	0.00078	0.00077
	8.5	50%	0.00668	−0.03878	0.14702	−0.09131	0.00160	0.00253	0.02174	0.02174
	12	75%	−0.12120	−0.07991	0.26264	0.14123	0.10959	0.07712	0.16906	0.14076

Notes: values of $E(x|x<d)$ are; 3.405 for $d = 5$; 4.934 for $d = 8.5$; 6.700 for $d = 12$. $P\{Y_i = 1|x_i\} = (1+ exp(-c-\beta x_i))^{-1}$, $c=-5$, *c is known* $x_i = N(5,2^2)(1 - \theta_i) + N(12,2^2)\theta_i$, where, θ_i, $i \geq 1$ are independent identically Bernoulli distributed random variables with $P\{\theta_i = 1\}=1/2$

TABLE 4

Logistic regression simulation study with gamma distributed exposure; bias and variance of the estimator under replacement value, *a*, equal to *0*, *E(x|x<dl)*, *dl* and *dl*/2

			Bias (Monte Carlo estimator of β)				Variance (Monte Carlo estimator of β)			
			a =				*a*=			
β	***d***	***% x < dl***	***0***	***E*(*x*|*x*<*d*)**	***dl***	***dl/2***	***0***	***E*(*x*|*x*<*d*)**	***dl***	***dl/2***
0.3	0.634	25%	0.00153	0.00148	0.00730	−0.00165	0.01204	0.01148	0.01412	0.01373
	0.919	50%	0.00309	0.00280	0.02926	−0.00595	0.01373	0.01269	0.01580	0.01450
	1.278	75%	0.00708	0.00586	0.07383	−0.01368	0.01384	0.01289	0.01628	0.01474
1.5	0.634	25%	−0.00427	−0.00405	0.04319	−0.02375	0.01636	0.01582	0.01699	0.01732
	0.919	50%	−0.00630	−0.00599	0.18292	−0.07987	0.01933	0.01590	0.04408	0.02821
	1.278	75%	−0.01079	0.00085	0.42685	−0.14473	0.02450	0.01982	0.18934	0.04348

Notes: values of $E(x|x<d)$ are; 0.454 for $d = 0.634$; 0.615 for $d = 0.919$; 0.772 for $d = 1.278$. $P\{Y_i = 1|x_i\} = (1+ exp(-c-\beta x_i))^{-1}$, $c=-1$, *c is known*, $x_i \sim Gamma$(shape = 4,scale = 4)