A COMPARISON OF REGRESSION METHODS IN DATA SUBJECT TO DETECTION LIMITS: AN APPLICATION TO LUNG FIBER ANALYSIS AMONG BRAKE WORKERS

by

Yimeng Liu

B.S., Pharmaceutical Science, Wuhan University, China, 2006

MPH, The Chinese University of Hong Kong, Hong Kong, 2011

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

by

Yimeng Liu

It was defended on

July 29th, 2013

and approved by

Ada O. York, PhD, Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Gong Tang, PhD, Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Ying Ding, PhD, Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Ravi K. Sharma, PhD, Assistant Professor, Departmental of Behavior and Community Health Science, Graduate School of Public Health, University of Pittsburgh

Thesis Director: Gary M. Marsh, PhD, Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Gary M. Marsh, PhD

A COMPARISON OF REGRESSION METHODS IN DATA SUBJECT TO DETECTION LIMITS: AN APPLICATION TO LUNG FIBER ANALYSIS AMONG BRAKE WORKERS

Yimeng Liu, M.S.

University of Pittsburgh, 2013

ABSTRACT

Objective: This thesis aims to apply and compare selected regression methods with a lung fiber analysis dataset. Final results based on 19 cases will be compared to 2011 Marsh et al.'s analysis based on the first 15 cases.

Methods: Two research questions for the lung fiber dataset are: (1) is there a relationship between the lung fiber concentration of TAA and lung fiber concentration of AC? and (2) is there a relationship between the lung fiber concentration of TAA and duration of employment as a brake worker? Besides the substitution method, bivariate normal regression was used in the doubly left-censored situation in question 1, while the censored normal regression and regression modeling with count data were used in the situation with only the dependent variable subject to detection limits in question 2.

Result: (1) The estimate of the slopes between the log-scale of two lung concentrations (TAA vs AC) were 0.59, 0.57, 0.59 and 0.54 in the simple linear regression with substitution (DL, 0.5DL, DL/$\sqrt{2}$) and the bivariate normal regression, respectively. All of the slope estimates were statistically significant different from zero (p-value = 0.001, 0.003, 0.002 and 0.003). (2) The estimate of the slopes between the log-scale of the TAA lung fiber concentrations and DOE were

0.001, 0.014, 0.008, 0.020 and 0.030 in the simple linear regression with substitution (DL, 0.5DL, and DL/$\sqrt{2}$), censored normal regression and the negative binomial regression, respectively. All of the slope estimates were not statistically significant different from zero (p-value = 0.933, 0.486, 0.675, 0.390 and 0.439).

Conclusions: The consistent results from the substitution and other methods provide support for both a positive relationship between the lung concentration of TAA and AC and for no relationship between the lung concentration of TAA and DOE among 19 brake workers with mesothelioma. These findings are consistent with Marsh et al.'s findings in 2011 based on the first 15 cases. The public health significance is that the study results provide additional support for the conclusion that exposure to non-commercial amphibole asbestos, and not chrysotile, is related to the observed mesothelioma in brake workers. However, these conclusions need to be verified with a larger sample size.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

TAA	Tremolite, actinolite and anthophyllite
AC	Amosite and crocidolite
DL	Detection limit
LC	Left censoring
DOE	Duration of employment
MLE	Maximum likelihood estimation
ROS	Regression on order statistics
K-M	Kaplan-Meier
CDF	Cumulative density function
SD	Standard deviation
SE	Standard error
CI	Confidence interval
HIV	Human immunodeficiency virus
RNA	Ribonucleic acid

PREFACE

I would like to express my very great appreciation to my thesis advisor Prof. Gary M. Marsh for his patient guidance, enthusiastic encouragement and useful critiques of this research work. Besides my advisor, I would like to thank Prof. Ying Ding for her comprehensive and patient assistance in the theoretical organization of the methodology and review of the analysis results in this research work. I also would like to thank the rest of my thesis committee: Prof. Ada O. Youk, Prof. Gong Tang, and Prof. Ravi K. Sharma for their insightful comments and inspired questions.

I would like to express my great thanks to Dr. Victor Roggli for providing me the lung fiber dataset and background information checking.

My sincere thanks also go to my friends: Lu Wang, Lin-wan Chen, Fei Ding, Zhen Zeng, Andrew Potter and Chien-Wei Lin, for their constructive suggestions and encouragement on my thesis working whenever I felt frustrated.

Last, but not least, I want to thank my parents: Xiaohong Zou and Jinzhong Liu, for their infinite support and unconditional trust of my goal to be a biostatistician.

1.0 INTRODUCTION

1.1 BACKGROUND & RATIONALE

It is widely known that prior exposure to asbestos is an important risk factor of mesothelioma [1-4]. Environmental and occupational exposures are the main sources of human exposure to asbestos. Several types of occupational populations including insulation and shipyard workers exposed to high-levels of asbestos dust were reported to have elevated risk of developing mesothelioma [5-12].

A large number of workers who install and repair brakes in cars and trucks have some potential asbestos exposure (mainly chrysotile asbestos) although the nature of these exposures is thought to eliminate or greatly decrease the potential health risks involved [13-18]. The epidemiology literature provides no support for increased mesothelioma risks among brake (automotive friction products) workers [18-19] and there is ongoing debate about whether chrysotile exposures in any setting can elevate mesothelioma risks [18-23].

Among all types of asbestos fibers, commercial amphiboles (primarily crocidolite and amosite) are well known to cause the mesothelioma due to their greater bio-persistence compared to chrysotile [5,16,24-26]. In his study of 10 brake repair workers with mesothelioma, Roggli et al. found that among all workers with an elevated level of the chrysotile or non-commercial amphibole fiber concentration, there is also an elevated level of the commercial amphiboles fiber concentration [27]. Because commercial amphiboles were not used in friction

products in the US, Roggli et al. believe there must be an unrecognized source of commercial amphibole fiber exposure for these brake workers that caused their mesothelioma [27-28].As a result, the question is whether there is a true linear or positive relationship between the chrysotile and the commercial amphiboles among these brake workers.

To address this question, Finkelstein performed a linear regression analysis between the concentration of the tremolite fiber (as a reasonably good biomarker for the chrysotile) and the commercial amphiboles in these 10 workers and found a not statistically significant p-value for the beta coefficient. He concluded that the there is no relationship between the chrysotile and commercial fibers [29].

However, by applying the quantile regression analysis that accounted for the two influential points among these 10 workers, Marsh et al. found an alternative conclusion that the lung levels of commercial amphiboles was a statistically significant predictor to the lung levels of tremolite. Marsh et al. also found no evidence for the duration of employment as a brake worker as a significant predictor for the lung tremolite level. They also obtained the same result after adding to the dataset five more mesothelioma cases that were brake workers [30].

Both Finkelstein and Marsh et al. used the substitution method for non-detect fiber concentrations in their analysis (observation labeled as less than some value).While Helsel has claimed that the substitution method in dealing with the non-detect problem is inadequate and inaccurate in estimating both of the summary statistics and regression coefficients in the environment setting [31-33], Antweiler et al. found that substitution with the 0.5DL will only give slight bias on summary statistics during some reasonable conditions using a simulated dataset [34].

Because of the suspected issues with the substitution methods, the goal of this thesis is to review and evaluate other available regression methods dealing with the non-detect problem that can be applied to the lung fiber dataset. Moreover, the results will be compared among different methods using the dataset.

1.2 REVIEW OF DATA ANALYSIS WITH NON-DETECT OBSERVATIONS

The primary question of the dataset is whether there is a linear or positive relationship between the lung tissue concentration of non-commercial amphiboles (TAA) and commercial amphiboles (AC). However, both of the two fiber concentration measurements have multiple detection limits (DLs). As a result, regular regression analysis cannot be applied while methods dealing with left-censored observations with both independent and dependent variable are subject to non-detect are in need.

In environmental and occupational studies, data with lower reporting limits (non-detect) are usually reported. Some examples of the non-detect problem included the water quality studies, industrial hygiene studies, HIV RNA measurement and astronomy research [34-39]. Non-detect data usually occur when the researchers in the laboratory cannot distinguish between a true zero and the false negative. The only known information about the data point is that the true value should be somewhere between zero and a positive value, which is known or estimated as the DL.

Some non-detect problems refer to the measurement of chemicals with very low-level concentration (such as chemical contamination in the water). With the limitation of the instruments or methods, sample with a positive signal might not mean a true positive value of the

material concentration (false-positive). As a result, methods had been developed to calculate the DL according to the blank samples (samples known to be true zero) or samples with very low positive values. The ways to define the DL varies across different settings [37].

Due to the DL, observations with value less than the DL will be reported as " < DL ", leading to a left-censored dataset in the future analysis by assuming the DL is known and fixed (although sometimes we need to estimate it). Hewett et al. divided datasets subject to non-detect into two types: (1) single censored and (2) complex censored [38]. A single censored dataset is one with only one DL and all value less than this DL are reported as "< DL". A complex censored dataset is one with multiple DLs, which means there could be observations with the value between two DLs (eg. <5, 7, <10).

The non-detect problem complicates data analyses such as the estimation of the summary statistics, measurement of the association and regression coefficients. A common and rough way to treat non-detect data in many practical fields is to substitute a fraction of the DL for all left-censored observations and apply the subsequent analysis, which is also known as fabrication [31,34,35]. Examples of the substituted value include 0.5DL (mid-point), DL/$\sqrt{2}$, DL or even zero [31,35].

However, Helsel has reported that substitution can give inaccurate estimate of the statistics and considered the poor estimation of the statistics can lead to the conclusion of a significant difference, correlations or regression relationship that do not exist [31-33]. On the other hand, Antweiler et al. found that when the censored rate is less than 70%, substituting 0.5DL will only cause slight bias but substituting zero or DL will cause severe bias in estimating the summary statistics using the simulated dataset[34].

Methods to estimate the summary statistics (mean, median & SD) for variables subject to non-detect have been developed. Most of the methods can be divided into four types: maximum likelihood estimation (MLE), regression on order statistics (ROS), Kaplan-Meier (K-M) or Turnbull estimation and multiple imputation [32, 34, 38, 39].

The book "Statistics For Censored Environmental Data Using Minitab and R" report there were 15 papers published, that discussed the comparison between different methods in estimating the summary statistics [41]. However, the results of the comparison are not consistent among all the papers and the reason could be due to the difference of the real and simulated dataset each author used to test the methods [42]. Based on the 15 review papers, Helsel concluded in his book "Statistics For Censored Environmental Data Using Minitab and R" that two factors needed for choosing the appropriate method was the percent of censored observations and the total sample size [42]. He recommended using the K-M method and imputation when less than half of the observations are censored. While MLE and multiple imputations can be applied when there are more censored cases (between 50% - 80%) and a larger sample size (> 50) [42]. However, it is more appropriate to use the robust MLE and ROS when the sample size is small (< 50) after checking the distribution assumption (except for the log-normal) [42]. He also recommended to only reporting the high sample percentiles or percent above a meaningful threshold when there are more than 80% of the observations censored [42].

Regression analysis is the most commonly used method in explaining the relationship between a continuous response variable and several dependent variables. Regression methods dealing with the response Y subject to the right-censored are well developed such as the Cox proportional hazard model, estimating the coefficient of the covariates based on the semi-parametric likelihood. As for the regression analysis for data with non-detect observations, most

of the methods focused on the single DL with only the outcome Y subject to the non-detect [43-45.] For multiple DLs, MLE methods can be applied to the dataset with Y subject to left-censoring (LC) and the covariate X is fully observed. By specifying some distribution function for the outcome variable and performing the ML estimation for the coefficients in the mean model. Instead of using the probability density function, the cumulative density function is used for all LC observations in constructing the likelihood.

When both the outcome Y and the covariate X are subject to multiple DLs, the situation becomes more complicated. Methods dealing with a missing or censored independent variable include conditional mean imputation, MLE, multiple imputations from other variables, and Bayesian approaches, most of which are focused on singly-censored case [46-47.] However, no standard method exists to deal with the doubly-censored case (both Y and X subject to DLs) directly. As a result, joint modeling has been used to solve the non-detect problem in both of the Y and X variables. Because we are interested in studying the potential linear or positive relationship between two concentrations, the bivariate normal distribution is the primary choice due to the property that there is a linear relationship between the conditional mean of Y given X and the X variable itself. Therefore, if we can find appropriate normal transformations for both of the concentrations, maximizing the likelihood function using the bivariate normal distribution can get the MLE estimates of the linear coefficients in the condition model.

A non-parametric method mentioned in Helsel's book that can also be applied to the doubly LC situation was the non-parametric Akritas-Theil-Sen (ATS) regression [48-49]. The ATS slope in this regression model is an extension to the Theil-Sen slope which estimates the median of the slopes between all possible pairs of data. However, this method has not been verified using the doubly censored dataset and therefore was not considered in the thesis project.

2.0 MATERIALS

2.1 DATASET

The dataset contained 19 brake repair workers who developed mesothelioma. Table 1 showed the selected demographic characteristics and fiber analysis information of the 19 subjects. Cases 1-10 were the original cases used in the analysis of Roggli and Finkelstein [27,29]. Cases 11-15 were five more cases added in Marsh el al.'s 2011 study [30]. In addition to the 15 cases, the dataset also included four more recently defined cases provided by Dr. Roggli (case 16-19).

For all workers in the dataset, the only known or suspected occupational exposure to asbestos was the brake repair work. The data included basic demographic information (age, sex, smoking status), clinical information (tumor site), employment information (duration of employment DOE, tumor type, occupation type) and the lung tissue fiber analysis of each case. The detailed information and source of the variables can be found in Roggli et al.' s paper [27].

2.2 LUNG TISSUE FIBER ANALYSIS

The lung tissue fiber analysis is the measurement of the individual lung concentration of three types of asbestos fibers -- **non-commercial amphiboles (TAA)** including tremolite, actinolite

and anthophyllite; **commercial amphiboles (AC)** including amosite and crocidolite and **chrysotile.**

The lung tissue fiber analysis was performed on formalin-fixed or paraffin embedded lung tissue of every subject. Chemical solutions like sodium hypochlorite were used to digest the lung tissue. The residue of the digestion product was then collected on a 0.4-µm Nuclepore filter. Filter with the residue was scanned with a scanning electron microscopic (SEM) at a screen size of 22.7 X 17.3 cm. The type of the fibers was determined by the elemental composition with the energy dispersive x-ray analysis (EDXA) after scanned by the SEM. The result was reported as a density measurement (no. of fibers/g).

2.3 DETECTION LIMIT

In the process of scanning, the whole filter was divided into many small fields (cells). The scanning started at a field with a specific x designation but a randomly picked y coordinate on the filter. The scanning process stopped after continuously scanning to the 100^{th} fields or 200 fibers counted. The area of the 100 cells was measured and used in the calculation of the fiber concentration. The area of the filter is $10.5 \times 10.5 \times \pi$ and the area of 100 cells was 2.3714 mm^2 for most of the observations (14600 cells in total).

Since not all of the fibers scanned will be analyzed for the fiber type, the estimated type specific count for every type of fiber in the 100 cells will be calculated as follows:

$$Estimated\ number\ of\ type\ A\ fibers = \frac{Number\ of\ type\ A\ fibers\ /\ Total\ number\ of\ fibers\ analyzed}{Total\ number\ of\ fibers\ scanned}$$

For each type of fiber, the lung tissue concentration for a subject with at least one fiber observed in the 100 fields is calculated as follows:

$$\frac{\left(Estimated\ number\ of\ type\ specific\ fibers\ /_{Area\ in\ 100\ fields}\right) * \ Area\ of\ the\ filter}{Weight\ of\ the\ wet\ tissue\ sample(gram)}$$

If there are no this type of fiber observed in the 100 fields, the lung tissue concentration is recoded as < DL. The DL for the type A fiber is calculated as follows:

$$\frac{\left(E\ /_{Area\ in\ 100\ fields}\right) * \ Area\ of\ the\ filter}{Weight\ of\ the\ wet\ tissue\ sample(gram)}$$

Where

$$E = \frac{1\ type\ A\ fiber\ /\ Total\ number\ of\ fibers\ analyzed}{Total\ number\ of\ fibers\ scanned}$$

The DL of the fiber concentration could be different with respect to different weights of the lung tissue and the area in the 100 cells in each sample. However, different type of fibers in

same subject (same tissue) will have same DL so that the TAA and AC measurement in the dataset share a same DL in every subject.

3.0 METHODS

The aim of this study is to apply available methods dealing with the non-detect problem to the lung fiber analysis dataset and compare the results and subsequent inference among different methods.

The research question was to assess if the concentration of the non-commercial amphiboles fiber (TAA) is positive related to the concentration of commercial amphiboles fiber (AC) but no relationship with the duration of employment. The concentration of TAA and AC are two continuous variables with non-detect observations. As a result, only regression analysis dealing with doubly-censored situation can be applied to study the relationship between the two fiber concentrations (TAA & AC). However, methods dealing with only a response variable subject to non-detect can also be applied to the analysis between the concentration of TAA and the duration of employment (DOE).

According to the two research questions, the methods section is organized as three parts: (1) general descriptive analysis of the dataset, (2) TAA vs AC (methods dealing with the doubly-LC situation) and (3) TAA vs DOE (methods dealing with dependent variable subject to LC).

3.1 DESCRIPTIVE ANALYSIS

The K-M method will be used to estimate the summary statistics (mean, median and SD) of two fiber concentrations subject to non-detect. In order to apply MLE methods in the subsequent analysis, it is necessary to evaluate the potential distribution (et. normal distribution) of the variables subject to non-detect. As a result, the K-M method will also be used to estimate the cumulative distribution function and quantiles of the lung fiber concentrations of TAA and AC. The estimated quantiles will be plotted versus quantiles in some specified distributions. A linear trend is expected to be observed in probability plot if the distribution selected is correct. Due to the continuous scale of the fiber concentration, normal and log-normal distributions will be used to draw the probability plot.

3.2 TAA VS AC

This part will describe the selected regression methods dealing with the doubly-LC situation applying them to answer the first research questions: Is there a positive relationship between TAA and AC?

3.2.1 Substitution

As the most common treatment of the non-detect in the practical field, regular analysis with substitution will also be included in the methods. As a result, after substituting all non-detect observations with DL, 0.5DL, DL/$\sqrt{2}$and zero in both of the two fiber concentrations (AC &

TAA), the simple linear regression will be applied to assess the relationship between the outcome variable TAA lung concentration and the independent variable AC lung concentration.

3.2.2 Bivariate normal regression

In the analysis of the relationship between the lung fiber concentration of TAA and AC, both the dependent and independent variables are subject to non-detect. In order to deal with the non-detect observations, a common parametric approach is to find some bivariate distribution to jointly model the two random variables and use ML estimation to get the estimate of the relationship coefficient as a function of the parameters in the distribution.

The bivariate normal distribution has a good property that there is a linear relationship between the conditional mean of one variable given the other variable E(Y|X) and the other variable itself X. Therefore, our goal is to find the appropriate normal transformation for the two fiber concentrations.

The MLE approach assumes the transformed TAA & AC jointly follow a bivariate normal distribution and both of the two variables are subject to LC. The censored value of the two variables follows a same unspecified distribution, which is independent with the bivariate normal distribution.

The calculation of the DL (see 2.3) shows that the value of the DL is only determined by the area of the 100 cells scanned and the weight of the tissue. According to the data provider, the range of the lung tissue used in the fiber analysis is between 0.1 and 0.33 gram and the area of the 100 cells scanned for most of the subjects are 2.3714mm^2. Therefore, it is reasonable to assume the weight of the tissue follows a certain distribution independent with the joint

distribution of the transformed TAA & AC. As a result, the censoring value of each observation also follows the same distribution $c \sim D(c)$ that is independent of the bivariate normal.

$$x_i: g(AC);\ y_i: g(TAA);\ c_i : DL$$

$$\begin{pmatrix} Y_i \\ X_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix} \begin{pmatrix} \sigma_y^2 & \sigma_{y,x} \\ \sigma_{y,x} & \sigma_x^2 \end{pmatrix}\right) \ \&\ C_i \sim D \qquad (1)$$

Due to the existence of the DLs, we only observed the maximum value between the latent variables (X_i, Y_i) and the censoring variable (C_i), representing as (T_{xi}, T_{yi}) and the indicator of whether the observation is LC ($\boldsymbol{\delta_{xi}}$, $\boldsymbol{\delta_{yi}}$). We denote the observed data as Oi = (t_{xi}, t_{yi}, δ_{xi} , δ_{yi}).

$$t_{xi} = Max\ (x_i, c_i), t_{yi} = Max\ (y_i, c_i)$$

$$\delta_{xi} = \begin{cases} 1\ if\ x_i \leq\ c_i \\ 0\ if\ x_i >\ c_i \end{cases} \ \&\ \ \delta_{yi} = \begin{cases} 1\ if\ y_i \leq\ c_i \\ 0\ if\ y_i >\ c_i \end{cases}$$

With the bivariate normal property, Y given X follows a normal distribution and the conditional expectation of Y given X has a linear relationship with the value of X. Only key equations are numbered in sequence.

$$Y|X \sim N\ (\mu_{y|x}, \sigma_{y|x})$$

$$\mu_{y|x} = E(Y|X) =\ \mu_y +\ \rho \frac{\sigma_y}{\sigma_x}(X -\ \mu_x) \qquad (2)$$

$$\sigma_{y|x} =\ \ (1 - \rho^2)\sigma_y^2$$

As a result, the slope and intercept in model $E(Y|X) = \alpha + \beta X$ can be calculated as:

$$\beta = \rho \frac{\sigma_y}{\sigma_x} ; \alpha = \mu_y - \beta \mu_x \quad (3)$$

ρ is the correlation coefficient between Y and X with a relationship of σ_Y and σ_x as follows:

$$\rho = \frac{\sigma_{x,y}}{\sigma_x \sigma_y} = \frac{\sigma_x \beta}{\sigma_y}$$

The likelihood function of the five unknown parameters in the bivariate normal will be proportional to the likelihood part with only the bivariate normal distribution because the censoring is independent with the latent bivariate normal distribution. As a result, the likelihood function of 19 observations for the five unknown parameters in the bivariate normal distribution can be written as:

$$L\left(\theta \mid O = T_y, T_x, \delta_y, \delta_x\right) \propto \prod_1^{19} f_{x,y}(t_{xi}, t_{yi})^{(1-\delta_{xi})(1-\delta_{yi})} \left[\int_{-\infty}^{t_{yi}} f_{x=t_{xi},y}(t_{xi}, y)\, dy\right]^{(1-\delta_{xi})\delta_{yi}}$$

$$\times \left[\int_{-\infty}^{t_{xi}} f_{x,y=t_{yi}}(x, t_{yi}) dx\right]^{\delta_{xi}(1-\delta_{yi})} F_{x,y}(t_{xi}, t_{yi})^{\delta_{xi}\delta_{yi}} \quad (4)$$

Within the likelihood function, θ is a vector of the five unknown parameters.

$$\theta = (\mu_y, \mu_x, \sigma_y, \sigma_x, \beta) \quad (5)$$

$f_{x,y}(x, y)$ and $F_{x,y}(x, y)$ is the probability density function and cumulative density function of the bivariate normal distribution. Because we are interested in getting the estimate of

the slope β rather than the correlation coefficientρ, we replace the ρ in the likelihood function as $\frac{\sigma_x \beta}{\sigma_y}$ to get the direct estimate of the β in the maximization.

Using the ML estimation method, we can get the MLE estimate of the five unknown parameters by maximizing the likelihood function as follows:

$$\{\hat{\theta}_{mle} = (\hat{\mu}_Y, \hat{\mu}_x, \hat{\sigma}_Y, \hat{\sigma}_x, \hat{\beta})\} = \{arg\ max_{\theta \in \Theta}\ lnL\ (\theta \mid O)\} \quad (6)$$

According to the large sample property of the MLE estimator, the MLE estimators asymptotically jointly follow a multivariate normal distribution with the mean vector equals to the true parameter vector and the covariance matrix of the MLE estimators equals to the inverse of the Fisher information matrix$I(\theta)$ (equation 7), known as the expectation of the second derivative of the log likelihood function (equation 9). Therefore, the covariance matrix of the MLE estimators is equal to the inverse of the Fisher information matrix $[I(\theta)]^{-1}$.

The observed Fisher information matrix evaluated at when $\theta = \hat{\theta}_{mle}$ ($\bar{\mathrm{I}}(D; \hat{\theta}_{mle})$) is a natural estimator of the Fisher information matrix $I(\theta)$ (equation 11). As a result, the asymptotic covariance matrix of the MLE estimators can be estimated using the inverse matrix of $\bar{\mathrm{I}}(D; \hat{\theta}_{mle})$ (equation 12). $l(\ O; \theta)$ is the log likelihood function (equation 8).

$$\hat{\theta}_{mle} \sim AN(\theta, [I(\theta)]^{-1}) \quad (7)$$

$$l(\ O; \theta) = Log\ L\ (\theta \mid O) \quad (8)$$

$$I(\theta) = E\left(-\frac{\partial^2}{\partial\theta\partial\theta^T} l(\ O; \theta)\right) \quad (9)$$

$$\bar{\mathrm{I}}(O; \theta) = -\frac{\partial^2}{\partial\theta\partial\theta^T} l(\ O; \theta) \quad (10)$$

$$\bar{\mathrm{I}}(O;\hat{\theta}_{mle}) = -\frac{\partial^2}{\partial\theta\partial\theta^T} l(\,O;\theta)_{|\theta=\,\hat{\theta}_{mle}} \quad (11)$$

$$\widehat{Cov}\,(\hat{\theta}_{mle}) = \left[\bar{\mathrm{I}}(O;\hat{\theta}_{mle})\right]^{-1} \quad (12)$$

According to the functional invariance of the MLE estimator, we can get the MLE estimator of α as follows:

$$\hat{\alpha}_{mle} = \hat{\mu}_y - \hat{\beta}_{mle}\hat{\mu}_x \quad (13)$$

Moreover, the variance estimate of the slope and intercept can be calculated using the delta method. Using delta method, we can get the asymptotic normal distribution of a function of the MLE estimators as follows:

$$\mathrm{g}(\hat{\theta}_{mle}) \sim AN(g(\theta),\ g'(\theta)[I(\theta)]^{-1}\, g'(\theta)^T) \quad (14)$$

As a result, an estimate of the variance of the $\mathrm{g}(\hat{\theta}_{mle}) = \hat{\alpha}_{mle} = \hat{\mu}_y - \hat{\beta}_{mle}\hat{\mu}_x$ will be $g'(\hat{\theta}_{mle})\widehat{Var}\,(\hat{\theta}_{mle})\, g'(\hat{\theta}_{mle})^T$.

$$Var\left[\mathrm{g}(\hat{\theta}_{mle})\right] = g'(\theta)[I(\theta)]^{-1}\, g'(\theta)^T \quad (15)$$

$$\widehat{Var}\left[\mathrm{g}(\hat{\theta}_{mle})\right] = g'(\hat{\theta}_{mle})\left[\bar{\mathrm{I}}(O;\hat{\theta}_{mle})\right]^{-1} g'(\hat{\theta}_{mle})^T \quad (16)$$

$g'(\hat{\theta}_{mle})$ is the vector of the first derivative of the function of $g(\theta)$ with respective to the parameter vector of θ evaluated at $\theta = \hat{\theta}_{mle}$.

3.3 TAA VS DOE

This part describes the selected regression methods dealing with the situation where only the dependent variable is subject to the DL. These methods will be applied to address the second research question: Is there a positive relationship between TAA and DOE?

3.3.1 Substitution

As described previously, a simple linear regression model will be performed to assess the relationship between the outcome variable TAA lung concentration and the independent variable DOE after substituting all non-detect observations with DL, 0.5DL, DL/$\sqrt{2}$ and zero for the TAA fiber concentrations.

3.3.2 Censored normal regression

In order to solve the non-detect problem in the independent variables (TAA), assume a given distribution for the residual and use the generalized linear model form as follows:

$$g(TAA \mid DOE) = \beta_0 + \beta_1 * DOE + \varepsilon$$

Assume the residual ε follows a standard normal distribution. There are two choices for the link function $g(x)$

1. $g(x) = x$: Use the normal density to construct the likelihood function.

2. $g(x) = lnx$: Use the log-normal density to construct the likelihood function.

The link function will be chosen according to the potential distribution of the outcome TAA lung fiber concentration.

Use the ML estimation to estimate the coefficient β_1 and β_0 in the model. For all observations reported as less than some value (LC) use the cumulative density function instead of the density function in the likelihood construction as follows:

$$L = \prod f[\varepsilon_i]^{1-\delta_i} F[\varepsilon_i]^{\delta_i}$$

Where δ_i is the indicator function of left-censored observation (1: LC, 0: fully observed). $f[\varepsilon_i]$ and $F[\varepsilon_i]$ are the probability density and cumulative density function for the normal distribution, respectively.

3.3.3 Model with counts

The fiber concentration level of TAA is calculated as the total number of fibers over the weight of lung tissue. Instead of modeling the fiber concentration density, we can also use the estimated total number of TAA fiber count in 100 cells as the response outcome. As a result, we can form the mean model as follows and put the corresponding weight of the lung tissue and the area ratio (area in counted cells/ area of the filter) as an offset in the model.

$$\ln[\, E(Count_{TAA} | \, DOE)\,] = \ln offset + \beta_0 + \beta_1 * DOE$$

Where $Count_{TAA}|DOE \sim g; offset = Weight \frac{Area\ in\ 100\ cells}{Area\ of\ the\ filter}$

Within the mean model, it is assumed that the estimated conditional count in the 100 cells given the duration of employment follows some discrete distribution g. Over-dispersion and the percentage of the zero count will be checked before choosing an appropriate discrete distribution. The negative binominal distribution will be used to model the data if the over-dispersion exists (Pearson chi-square in regular Poisson model / degree of freedom > 1) and the zero-inflated Poisson model will be used if the percentage of zero count is large than 10%. If both of the situations occur, the zero-inflated negative binomial model will be used. ML estimation will be used to estimate the coefficient $\beta_0\ and\ \beta_1$ no matter which distribution is chosen.

In the regression model, the estimated TAA count in 100 cells scanned is the outcome. However, for observations with only a part of the fibers analyzed, the estimated TAA count will be calculated as a proportion of the TAA count in the fibers analyzed multiplied by the total number of fiber count in 100 cells. As a result, the estimated TAA count will be rounded to integer for the count model. Therefore, to be consistent with the calculated concentration, the offset used in the count model is calculated as the rounded estimate number of TAA fiber count in 100 cells over the TAA concentration.

$$Estimated\ number\ of\ TAA\ fibers\ in\ 100\ cells = \frac{Number\ of\ TAA\ fibers\ /\ Total\ number\ of\ fibers\ analyzed}{Total\ number\ of\ fibers\ scanned\ in\ 100\ cells}$$

4.0 RESULTS

4.1 DESCRIPTIVE ANALYSIS

Table 2 shows the censoring rate in the dataset. Among 19 observations, there are 42% (8/19) of subjects have complete data for both of the TAA and AC lung concentration. Sixteen percent (3/19) of the subjects had the TAA lung concentration observed but AC lung concentration LC while 10% (2/19) had the opposite situation. Thirty-two percent (6/19) of subjects had LC observations in both of the two lung concentrations. As a result, there are 58% of the subjects who had either one variable LC or all of the two variables LC, indicating a medium severe non-detect situation (Table 2).

Table 3 shows the summary statistics for the outcome and dependent variables. K-M method was used to estimate the mean, median and SD of variables subject to non-detect (AC&TAA). According to the K-M method, the mean lung concentrations of TAA and AC among 19 observations were 1055 and 1118 fiber/gram. The mean estimates were considered larger than the median estimates of the fiber concentrations (492, 489 fiber/gram for TAA and AC, respectively), indicating a right-skewed distribution for both of the two fiber concentrations. The log transformed mean estimate of the two fiber concentrations were 6.51 and 6.33 while the median estimate were 6.20 and 5.19, respectively (Table 3). The median estimates were very close to the mean estimates in the two log-scale concentrations. As a result, after transforming to

the log scale, the distributions of the two fiber concentrations were more symmetric than the original scale.

Figure 1 shows the estimated CDF of two fiber concentrations and Figure 2 shows the probability plot of the original and log-scale concentrations. It seems that the two log-transformed fiber concentrations are more reasonably normally distributed compared to the original scale with a clear linear trend between the estimated K-M quantile and the normal quantile.

Due to the descriptive and probability plot result, the log-transformed fiber concentrations are potentially normally distributed. As a result, the log-scale concentration of TAA and AC will be used in the following parametric methods.

4.2 TAA VS AC

4.2.1 Substitution

Figure 3 shows the histogram of the outcome variable lung fiber concentration of TAA in both the original and log scale with four different substituted values (DL, 0.5DL, DL/$\sqrt{2}$ and zero). The original scale of TAA with all four substitutions is a right-skewed distribution while with the log-transformation, the distribution becomes more symmetric (no log transformation for zero substitution).

The scatter plots between TAA and AC showed an approximately linear trend in all four substitutions (Figure 4). However, the linear trend becomes clearer and stronger when plotting the log-transformed TAA and AC concentrations with three substituted values (DL, 0.5DL and

DL/$\sqrt{2}$) (Figure 4). Because the log scale TAA concentrations after substitution are more likely to be normally distributed, we only performed the simple linear regression between the two log transformed concentrations.

The slopes between the two log-transformed fiber concentrations in a simple linear model ranged between 0.57-0.59 with three different substituted values. Moreover, all of the slopes in the three models were statistically significant different form zero (p-value = 0.001, 0.003 and 0.002) (Table 4), all indicating a significant positive linear relationship between the log-transformed TAA and the log-transformed AC.

4.2.2 Bivariate normal regression

4.2.2.1 Data analysis

The result of the ML estimation for the four parameters in the bivariate normal distribution ($\mu_x, \mu_y, \sigma_Y, \sigma_x$) and the linear coefficient (β,α) in the conditional linear model are shown in Table 5. The mean estimate of the log scale TAA and AC concentration are 6.23 and 6.41, respectively and the SD estimates are 1.24 and 1.04, respectively. The estimated slope in the conditional model (ln TAA given ln AC) was 0.54 and was significant different from zero (p-value = 0.003).

4.2.2.2 Simulation

We ran a simulation study to evaluate the performance of the MLE estimator in estimating the slope and the observation information matrix in estimating the variance of the MLE estimator when both of the sampled x & y are subject to LC. We choose the true parameters close to the

estimated value in the real dataset (19 brake workers) in order to make the simulated results meaningful to our analysis.

The samples were randomly generated from a bivariate normal distribution with $\mu_x = \mu_Y = 6,\ \sigma_x = 1.3, \sigma_Y = 1, \beta = 0.5$. Moreover, the DL for each observation was randomly generated from a uniform distribution UNIF (146/0.33, 146/0.1). If the DL was larger than the value of x(y), the observation value x(y) was left-censored at the DL. We then recorded the MLE estimate of each parameters and the standard error estimate of the MLE estimators in each random simulated sample and calculated the mean and SD.

In the simulation samples, it was possible to get a singular observed information matrix with the MLE estimate especially when the sample size was very small. As a result, we removed all samples with this situation in calculating the mean and SD for the parameter and standard error estimate. When the sample size was equal to 19, approximately 15% of the 1000 simulations did not have a non-singular observed information matrix and the percentage decreased as the sample size increased (15%, 9%, 4% and 0.1% for sample size = 19, 30, 50 and 100 , respectively).

The results based on 1000 simulations with four different sample sizes (sz=19, 30, 50, 100) are shown in Table 6. Based on the 1000 simulations, the sample mean of the MLE estimators were close to the true value for all six parameters even with a small sample size (n=19) (6.03, 5.04, 1.19, 0.98, 0.53 and 2.73 for $\mu_x, \mu_y, \sigma_Y, \sigma_x, \beta$ and α, respectively). The distributions of the beta MLE estimates in 1000 simulations were roughly symmetric (Figure 6) even for a small sample size (n=19). The distribution of the SE estimates for beta was rightly skewed when the sample size was comparatively small (19, 30). However, the 95% CI coverage rate for the slope β ranged from 98% to 99%, indicating a slightly conservative estimate of the

confidence interval. The same situations were observed for the other parameters (data not shown).

For smaller sample size, the distributions of the SE estimates were rightly skewed so that the median of the SE estimates was more reasonable to use to compare with the SDs of the MLE estimates (a rough estimate of the true variability of the MLE estimator). As a result, the SE estimates of the beta generally under estimated the true variability of the MLE beta estimate when the sample size was 19 (0.28 and 0.44, respectively). The mean and median of the SE estimate for beta were close to the SD of the beta MLE estimate when the sample size was at least 50 (0.18, 0.17 and 0.19, respectively).

4.3 TAA VS DOE

4.3.1 Substitution

Figure 5 shows the scatter plots between the TAA and DOE in linear and log scales. Both of the scatter plots between two scales of the TAA concentration (original and log) and DOE shows no linear trend in all four substitutions. The dots in all scatter plots seem to be randomly distributed.

Table 7 shows the estimates of the linear regression coefficients. After transforming into the log scale, the slopes between the log-scaled TAA fiber concentrations and DOE in a simple linear model ranged between 0.001-0.014 with three different substituted values. However, all of the slopes in the three models are not statistically significant different form zero (p-value = 0.933, 0.486 and 0.675) (Table 7), indicating that there is no significant positive linear relationship between the log-transformed concentration of TAA and DOE.

4.3.2 Censored normal regression

The first row of Table 8 shows the estimated coefficients in the censored normal regression model with the log scale of TAA as the outcome variable and the DOE as the independent variable. The slope estimate of DOE was 0.022 which was not significant different from zero (p-value=0.39), indicating that there is no statistically significant positive linear relationship between the log-transformed TAA lung fiber concentration and the duration of employment.

4.3.3 Model with counts

Table 9 shows the distribution of the estimated TAA count in 100 cells scanned for all subjects. For every subject with no TAA fiber detected, the value of the count in 100 cells was zero. Among 19 subjects, there were 42% (8/19) with no TAA fiber detected in the 100 scanned cells. The mean count for all 19 workers was 2 while the variance of the count is 11. The histogram of the counts in 19 subjects revealed a right-skewed distribution. The variance of the count was much larger than the mean of the count and the estimated dispersion in the Poisson model was 5.43, indicating a potential over-dispersion of the data. Moreover, nearly half of the counts in all subjects were zero (42%), which is larger than 10%.

As a result, the regular Poisson model might not be appropriate to use due to the over dispersion and comparatively higher percentage of zero counts. Instead of using the Poisson distribution to model the count, we fit a regular negative binomial regression model and a zero-inflated negative binomial regression model for the count data in 100 scanned cells.

The results of the two models were very close and the likelihood test showed that there was no significant difference between the two models (p-value=0.091). As a result, only the negative binomial model results were reported.

The second row of Table 8 shows the estimate of the coefficients in the linear combinations (β, α) and the parameter alpha for the negative binomial model. The estimated alpha was 1.306 which was significantly different from zero (p-value<0.001), indicating that the negative binomial model was significantly different from the Poisson model due to over-dispersion of the data. The estimate of the slope in the negative binomial model for DOE is 0.0298 (IRR=1.03). However, this estimate was not significantly different from zero (p-value =

0.439), indicating no significant positive relationship between DOE and the lung fiber TAA concentration (Table 8).

5.0 DISCUSSION

This section includes three parts. The first and second part compares the results of questions 1 and 2 using the different methods. The last part talks about the limitations of the thesis and recommendations for data analysis with non-detects.

5.1 IS THERE A POSITIVE RELATIONSHIP BETWEEN TAA AND AC?

Figure 7 shows the scatter plot between the two lung fiber concentration TAA and AC. and the fitted lines of all four models (substitution and the bivariate normal regression). Substitution methods and the bivariate normal approach gave consistent results in the linear relationship between the log scales of the two lung fiber concentrations. The estimate of the slope in the conditional mean model of the log scale TAA concentration given the log scale AC concentration were statistically significant in both of the three substitution and the bivariate normal models. The estimate of the slopes ranged from 0.54 – 0.59 in four models as follows:

Substitution – Simple linear regression

DL: E (ln TAA | AC) = 2.79 + 0.59* ln AC

0.5DL: E (ln TAA | AC) = 2.83 + 0.57 *ln AC

DL/$\sqrt{2}$: E (ln TAA | AC) = 2.79 + 0.59 *ln AC

Bivariate normal regression

E (ln TAA | AC) = 3.04 + 0.54* ln AC

The interpretation of the double log linear model is different from the linear model. In the simple linear model with the original scale, the slope $\hat{\beta}$ can be explained as the unit increase in outcome Y with one unit increase of the X.

However, the slope $\hat{\beta}$ in the double log linear model is the unit increase in log scale of Y with one unit increase of log scale of X. After several math transformations, we can get a relationship of percentage increase of Y and X.

ln Y = $\alpha + \beta$ ln X

Y = exp ($\alpha + \beta$ ln X) = exp(α) X^{β}

If X* = $X(1 + p\%)$

Then Y* = exp(α) $[X(1 + p\%)]^{\beta}$ = exp(α) $X^{\beta}(1 + p\%)^{\beta}$ = Y $(1 + p\%)^{\beta}$

As a result, the slope estimate $\hat{\beta}$ in the double log scale model can be interpreted as a $p\%$ increase in the AC lung fiber concentration, the TAA lung fiber concentration will be multiplied with a proportion of$(1 + p\%)^{\hat{\beta}}$.

Therefore, with the estimate in the substitution and bivariate normal approach, respectively, a 10% increase of the AC lung fiber concentration, the TAA lung fiber concentration will increase by 5% (calculated as $(1 + 10\%)^{0.54} - 1$) to 6% (calculated as $(1 + 10\%)^{0.59} - 1$).

The final model between the two lung fiber concentrations in the 2011 Marsh's analysis is a median regression model with the estimated slope equal to 0.52. The interpretation of the slope estimate in the quantile regression model is different from the linear regression model since it models the conditional quantile rather than the conditional mean. Therefore, the 0.52 slope estimate in the quantile (median) regression model indicates that with every unit increase in the lung fiber concentration of AC, there is a 0.52 marginal increase in the conditional median of the lung fiber concentration of the TAA. As a result, Marsh et al.'s study found a positive linear relationship between the two lung fiber concentrations based on the median model [30].

The bivariate normal regression has a very strong assumption that both of the two variables should jointly follow a bivariate normal distribution, which requires that all of the linear combinations of these two variables also follow a normal distribution.

ML estimation and the Fisher information matrix were considered as the classical approach to get an estimate of the parameter and the variance of the estimator under the large sample assumption. As a result, it is interesting to look at the performance of the MLE methods when the sample is subject to LC.

Simulation results based on 1000 simulations with parameters close to the brake worker dataset showed that the MLE estimates are stable and accurate even when the sample size is as small as 19. However, the SE estimators for the MLE estimators (beta, alpha, mean and sigma) tend to underestimate the true variability of the estimators based on smaller sample sizes. Due to the simulation result, 50 is a recommended sample size to get a more accurate and stable variance estimate of the MLE estimator for a double LC dataset.

A rough estimate of the true variability of the MLE slope estimators is 0.43, which is the standard deviation of the MLE slope estimates based on the 1000 simulations. This value gives a

Z statistics equal to 1.16 (0.5/0.43) less than 1.98. From the simulated results with the true parameters similar to the lung fiber dataset, it will be difficult to statistically significant detect a true 0.5 slope when the sample size is as small as 19.

Although the SE estimates might be under or over-estimated in the bivariate normal regression and the analysis in this thesis did not address the two influential points in the 19 subjects but focused on the non-detect problem, the simple linear regression with substitution and the bivariate normal approach still gave very close estimates of the slope and consistently statistical significant results. Thus, it is reasonable to conclude that an important positive relationship exists between the lung fiber concentration of TAA and AC. This result is also consistent with the findings of Marsh et al based on a median regression model applied to the first 15 cases [30].

5.2 IS THERE A POSITIVE RELATIONSHIP BETWEEN TAA AND DOE?

Figure 8 shows the scatter plot between TAA and DOE and the fitted lines of three models based on the concentrations (substitution and censored normal regression). When modeling the concentrations, the substitution methods and the censored normal regression gave consistent result of the linear relationship between the log scale TAA concentration and the duration of employment. The estimate of the slopes in the condition mean model of the log scale TAA given DOE are not statistically significant in both of the two approaches.

Substitution – simple linear regression model

DL: E (ln TAA | AC) = 6.79 + 0.001*DOE

0.5DL: E (ln TAA | AC) = 6.25 + 0.014*DOE

DL/$\sqrt{2}$: E (ln TAA | AC) = 6.52 + 0.008*DOE

Censored normal regression model

E (ln TAA | AC) = 5.96 + 0.022*DOE

Because the original outcome data are the estimated counts of the TAA fibers in 100 cells, it is reasonable to model the counts directly. Due to the over-dispersion and a nearly 50% data being zero, zero-inflated Poisson and negative binomial regression model were used in the analysis of the count data.

The negative binomial mode gave a consistent result by indicating a not statistically significant positive relationship between the lung concentration of TAA and the DOE comparing to the simple linear regression with substitution and the censored normal regression.

Negative binomial model:

E [ln(TAA Count in 100 cells / offset) | DOE] = 6.16 + 0.030DOE

E [ln(TAA concentration) | DOE] = 6.16 + 0.030DOE

In summary, the inference drawn from either the model of the concentration or the model of the count in 100 cells consistently reveals that there is no statistically significant positive relationship between the TAA lung fiber concentration and the duration of employment as brake workers, which is also consistent with Marsh et al.'s findings based on the first 15 cases [30].

5.3 LIMITIONS AND RECOMMENDATIONS

There are several limitations of this study. (1) The sample size of the lung fiber analysis dataset is very small with only 19 brake workers and the censoring percentage is almost 50% for both of the dependent and independent variable. Moreover, there are not many useful covariates that can be used in the multiple imputations for the censored observations. Because of the small sample size and limited covariates, all of the methods used in the analysis require some level of distributional assumptions. Because the MLE approach is based on the principle of the large sample theory, it might not perform well when the sample size is small. (2) Due to the limited time and scale of a master's thesis, only the simulation study for the bivariate normal regression was performed to verify the performance of the bivariate normal approach in the doubly-LC situation. As a result, the comparison among different methods was only based on the lung fiber data analysis rather than the simulation study. Readers should therefore be cautious interpreting the model results. (3) The analysis in this thesis focused on the non-detect problem of the lung fiber dataset and did not consider the influential points mentioned in Marsh et al.'s paper [30]. Future studies should focus on methods addressing both of the non-detect and influential points problems.

Comparing to the MLE methods, the substitution methods with three different substituted values (DL, 0.5DL, DL/$\sqrt{2}$) seem to perform well by giving the close estimates and consistent test results of the slopes for both of the two research questions using the real dataset. As a result, the substitution methods might be still useful when there are no other methods available. However, the performance of the substitution method still needs to be verified via a simulation study.

As for the regression methods in the doubly-LC case, the bivariate normal approach is recommended when there exists appropriate normal transformation of the data and the sample size of the dataset is larger than 50. However, caution should be taken when interpreting the model coefficients after transformation.

6.0 CONCLUSION

Results from this study show that simple linear regression with substituting of non-detect observations with DL, 0.5DL and DL/$\sqrt{2}$ gives consistent results (close estimate of the coefficients and same result of the test of the statistical significance) with the censored regression methods with ML estimation that account for non-detect as left-censored observations. These consistent results provide additional support for a positive relationship between the lung concentration of the TAA and AC fiber among the 19 brake workers with mesothelioma, which is consistent with Marsh et al.'s finding in 2011 based on the first 15 cases [30].

Moreover, the consistent results of the substitution approach, the censored normal regression model and the negative binomial model between TAA and DOE indicates no statistically significant positive relationship between the lung concentration of the TAA fiber and the duration of employment. The public health significance of this study is that the results provide additional support for the conclusion that exposure to non-commercial amphibole asbestos from some unrecognized source, and not chrysotile, is related to the observed mesothelioma in brake workers. However, these conclusions need to be verified with a larger sample size.

APPENDIX

TABLES AND FIGURES

Table 1: Lung fiber analyais dataset of 19 brake workers with mesothelioma

Case	Age (yr)	Tumor type/site	Occupation	Smoking	DOE[a] (yr)	Est. TAA count in 100 cells	Tissue weight (gram)	Area of 100 cells	Adjusted offset[b]	AC[c]	TAA[d]
Original 10 cases in Roggli											
1	61	E/PL	Auto machinist	XS	**37**	1	0.067	2.3714	4.59E-04	**3270**	**2180**
2	58	E/PL	Brake mechanic	C	**27**	1	0.334	2.3714	2.29E-03	**3936**	**437**
3	55	E/PL	Brake mechanic	C	**24**	11	0.363	2.5254	2.67E-03	**966**	**4115**
4	73	B/PL	Auto mechanic	C	**40**	1	0.203	2.3714	1.39E-03	**<720**	**720**
5	51	E/PE	Brake repair	XS	**11**	2	0.253	2.3714	1.73E-03	**<577**	**1155**
6	53	D/PL	Auto mechanic	ND	**7**	1	0.299	2.3714	2.05E-03	**489**	**489**
7	ND	ND/PL	Brake repair	ND	**15**	0	0.439	2.3714	3.00E-03	**333**	**<333**
8	66	B/PL	Brake repair	XS	**40**	10	4.97	2.8	4.10E-02	**122**	**244**
9	71	B/PL	Auto repair	C	**17**	5	0.194	2.3714	1.32E-03	**6148**	**3794**
10	ND	E/PL	Brakeline repair	ND	**24**	5	0.304	2.5254	2.31E-03	**1444**	**2166**
5 cases added in Marsh 2011											
11	69	B/PI	Brake & clutch repair	XS	**5**	0	0.283	2.3714	1.94E-03	**<516**	**<516**
12	70	E/PI	Auto mechanic	XS	**7**	0	0.18	2.3714	1.23E-03	**812**	**<812**
13	58	E/PI	Auto mechanic	NS	**25**	0	0.302	2.3714	2.07E-03	**<484**	**<484**
14	58	E/PI	Auto mechanic	C	**32**	0	0.069	2.3714	4.72E-04	**2117**	**<2117**
15	40	E/PE	QC parts inspector	C	**20**	1	0.297	2.3714	2.03E-03	**492**	**492**
4 cases added in 2013											
16	45	E/PI	Auto mechanic	NS	**4**	0	0.111	2.3714	7.60E-04	**<1316**	**<1316**
17	47	E/PI	Shadetree mech	NS	**17**	2	0.294	2.3714	2.01E-03	**<497**	**994**
18	53	B/PE	Auto mechanic	C	**19**	0	0.296	2.3714	1.84E-03	**<543**	**<543**
19	60	D/PI	Gen motors	C	**11**	0	0.119	2.3714	8.14E-04	**<1228**	**<1228**

ND, not determined; B, biphasic; D, desmoplastic; E, epithelial; PE, peritoneal; PL, pleural; C, current smoker; XS, ex-smoker; DOE, duration of employment; AC, commercial amphiboles (amosite + crocidolite); TAA, non-commercial amphiboles (tremolite + anthophyllite + actinolite)

[a] DOE is the years of employment as a brake worker

[b] Adjusted offset is calculated as Est. TAA count/TAA

[c] AC is the calculated concentration of the AC fiber count (fiber/gram)

[d] TAA is the calculated concentration of the TAA fiber count (fiber/gram)

Table 2: Summary of the censoring rate

Observation type	No. (%)
Both observed	8(42%)
TAA observed, AC LC	3(16%)
AC observed, TAA LC	2(10%)
Both LC	6(32%)

Table 3: Summary statistics

Variable	Mean*	Median*	SD*	% LC
TAA	1055	492	1211.8	42%
ln TAA	6.51	6.2	1.01	
AC	1118	489	1655.4	47%
ln AC	6.33	6.19	1.4	
DOE	20.11	19	11.5	-

*** Summary statistics are calculated using Kaplan-Meier method for variable with LC observations (TAA, AC)**

Table 4: Simple linear regression estimate by different substitution type - TAA vs AC

Model	Substituted value	$\widehat{\beta}$	P-value	$\widehat{\alpha}$	P-value
ln TAA | ln AC = α + β*ln AC +ε	DL	0.59	0.001	2.79	0.016
	0.5DL	0.57	0.003	2.83	0.02
	DL/√2	0.59	0.002	2.79	0.018

Table 5: Results of the bivariate normal regression

Parameter	Est.	SE	|t| Est. / SE	P-value H0: Est. = 0
Bivariate normal				
μ_{lnAC}	6.23	0.34		
μ_{lnTAA}	6.41	0.28		-*
σ_{lnAC}	1.24	0.27		
σ_{lnTAA}	1.04	0.22		
Linear model E (ln TAA| ln AC) = α + β*ln AC				
β	0.54	0.19	2.8	0.003
α	3.04	1.29	2.3	0.009

* The test of significance for the mean and SD are not interested and therefore not reported in the table.

Table 6: Simulation results for the MLE and the SE estimators*

n	$\hat{\theta}$			$SE(\hat{\theta})$			$\hat{\theta}$			$SE(\hat{\theta})$		
	Mean	**SD**	**95%CI coverage#**	**Mean**	**Median**	**SD**	**Mean**	**SD**	**95%CI coverage**	**Mean**	**Median**	**SD**
			$\hat{\mu}_x$						$\hat{\sigma}_y$			
19	6.03	0.59	0.89	0.52	0.42	0.43	0.98	0.44	1.00	0.93	0.31	13.80
30	6.02	0.42	0.91	0.41	0.36	0.19	0.99	0.32	1.00	0.30	0.27	0.15
50	6.01	0.33	0.91	0.31	0.29	0.10	0.99	0.23	1.00	0.22	0.21	0.07
100	5.99	0.22	0.95	0.22	0.21	0.05	0.99	0.16	1.00	0.15	0.15	0.03
			$\hat{\mu}_y$						$\hat{\beta}$			
19	5.94	0.79	0.89	0.51	0.38	0.51	0.53	0.43	0.98	0.87	0.28	13.89
30	5.96	0.41	0.93	0.37	0.32	0.21	0.51	0.30	0.99	0.25	0.22	0.15
50	5.99	0.28	0.93	0.27	0.25	0.10	0.51	0.19	0.98	0.18	0.17	0.06
100	6.00	0.18	0.95	0.19	0.18	0.04	0.50	0.13	0.99	0.12	0.12	0.03
			$\hat{\sigma}_x$						$\hat{\alpha}$			
19	1.19	0.46	0.99	0.42	0.36	0.29	2.73	2.95	0.91	12.60	1.95	289.20
30	1.23	0.33	1.00	0.34	0.31	0.14	2.86	2.05	0.92	1.75	1.55	0.98
50	1.26	0.25	1.00	0.26	0.25	0.08	2.90	1.34	0.93	1.25	1.17	0.43
100	1.29	0.19	1.00	0.19	0.18	0.04	2.98	0.94	0.92	0.84	0.82	0.20

* Results are based on 1000 simulation removing the cases with a singular observed information matrix. The true $\theta = \{6,6,1.3,1,0.5\}$ $\alpha = 3$

The 95% CI coverage rate is calculated as the percentage of the estimated 95% CI ($\hat{\theta} \pm 1.96 * SE(\hat{\theta})$) including the true parameter value.

Table 7: Simple linear regression estimate by different substitution type - TAA vs DOE

Model	Substituted value	$\widehat{\beta}$	P-value	$\widehat{\alpha}$	P-value
ln TAA|DOE = α + β*DOE +ε	DL	0.001	0.933	6.79	<0.001
	0.5DL	0.014	0.486	6.25	<0.001
	DL/$\sqrt{2}$	0.008	0.675	6.52	<0.001

Table 8: Results of censored normal regression and negative binomial model

Model		$\widehat{\beta}$	P-value	$\widehat{\alpha}$	P-value
Censored normal	**ln TAA | DOE = α + β*DOE +ε**	0.022	0.39	5.96	<0.001
Negative binomial	**E (ln TAA count | DOE) = ln offset + α + β*DOE**	0.03	0.439	6.16	<0.001
	alpha	1.306*	<0.001		

Table 9: Summary of the TAA count data

Count	Freq.	Percent
0	8	42%
1	5	26%
2	2	11%
5	2	11%
10	1	5%
11	1	5%

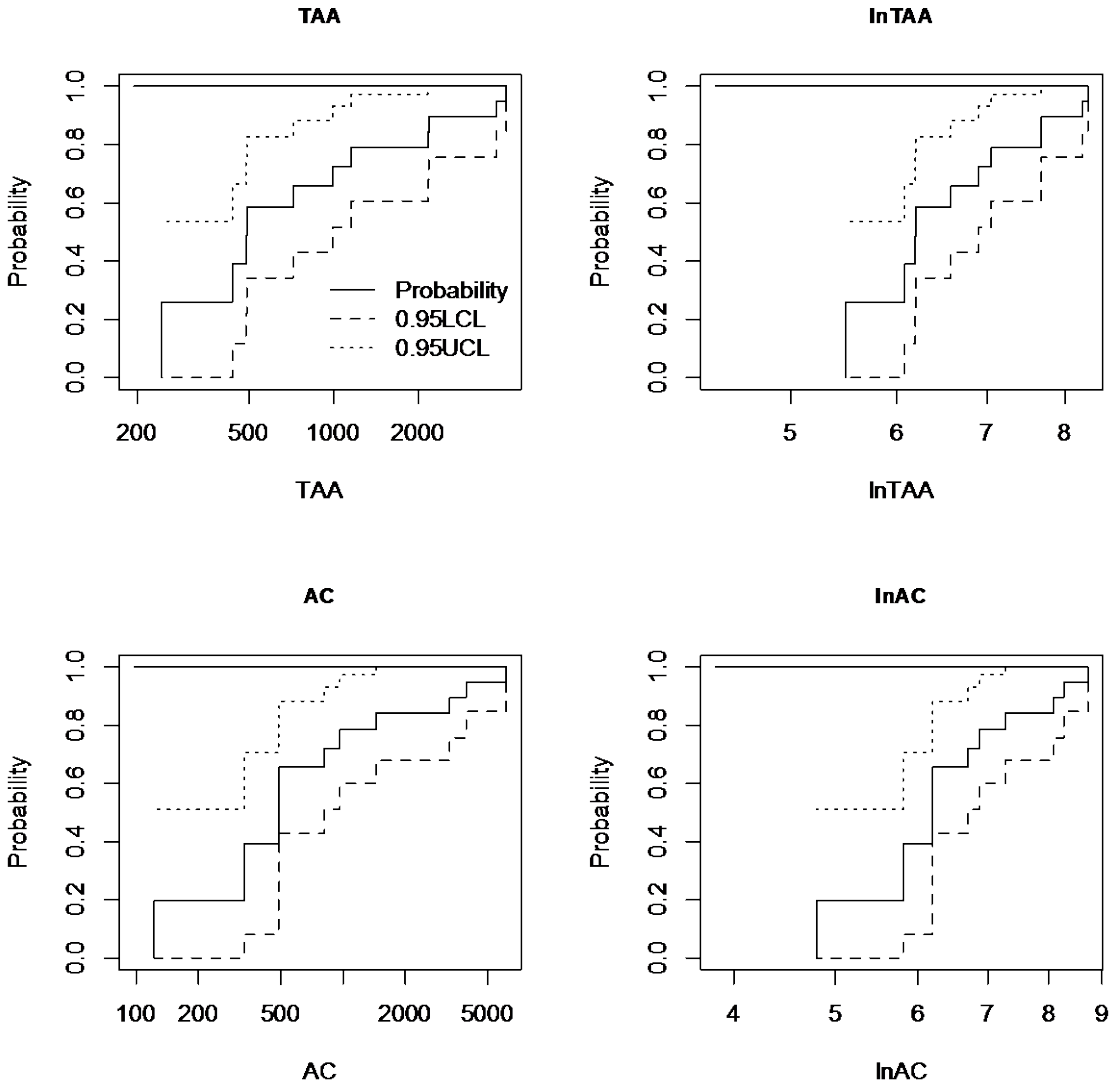

Figure 1: Estimated empirical CDF of TAA with two scales using K-M method

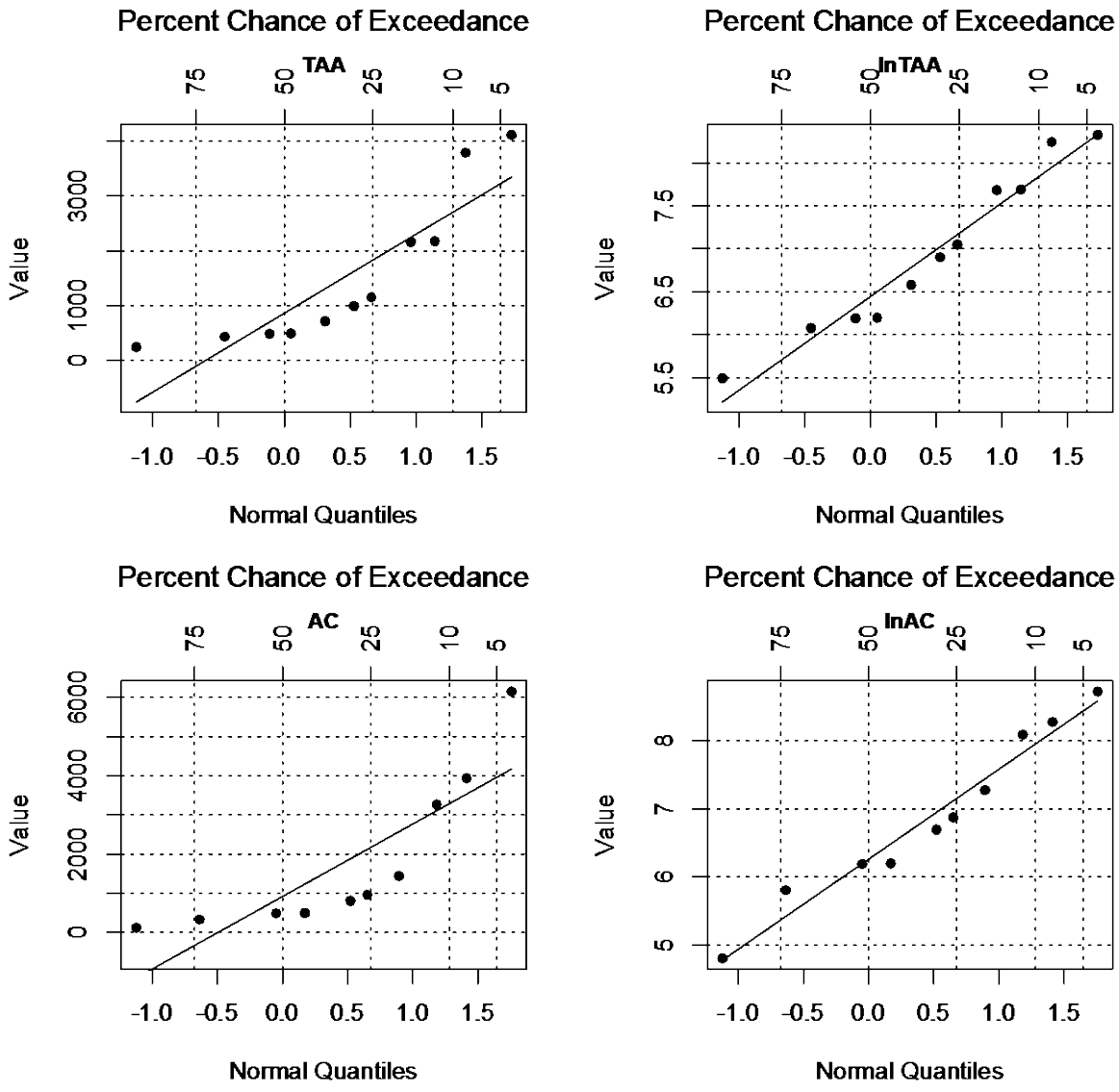

Figure 2: Normal Q-Q plots of TAA with two scales

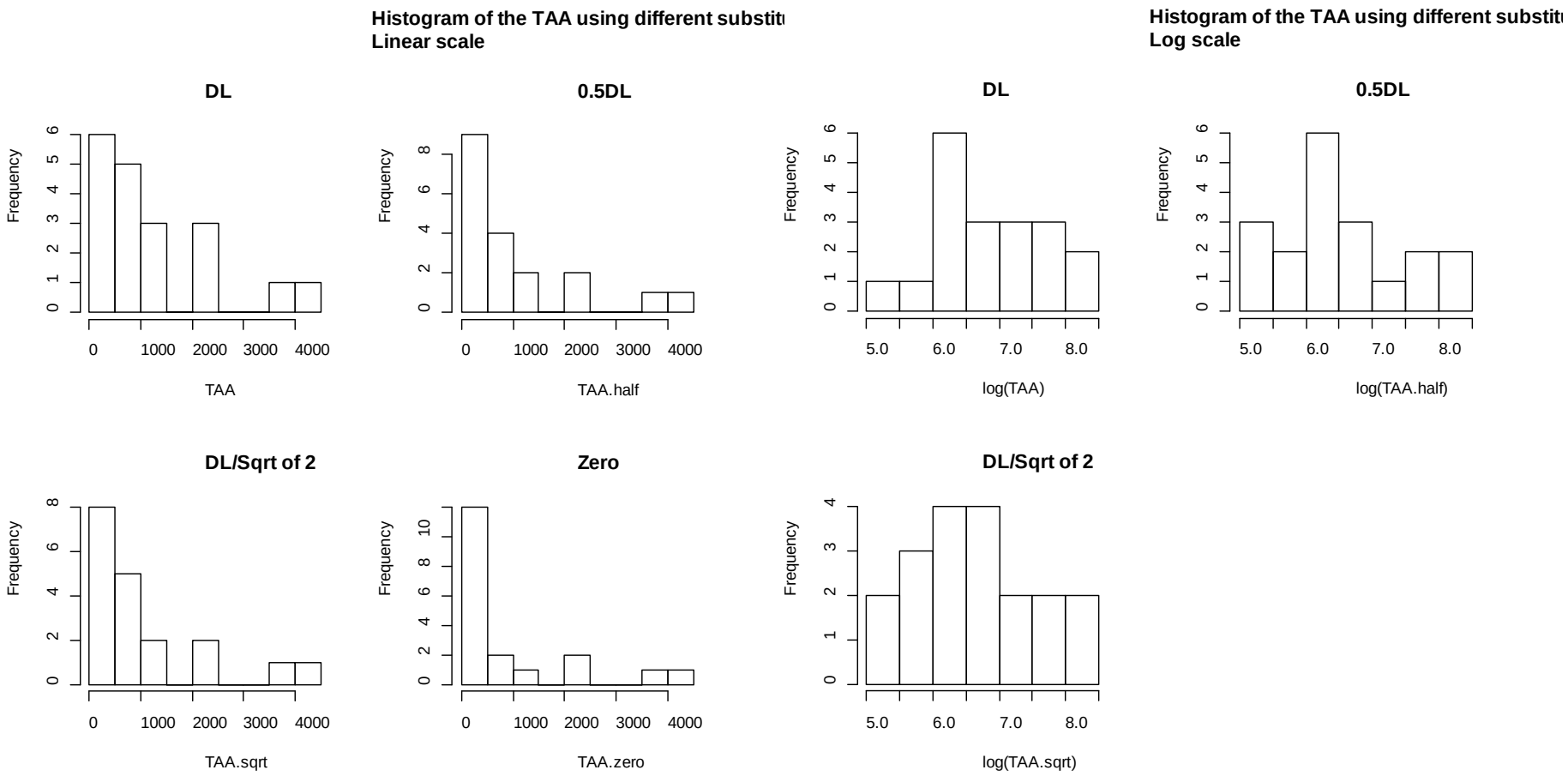

Figure 3: Histograms of TAA by different scales with substitutions

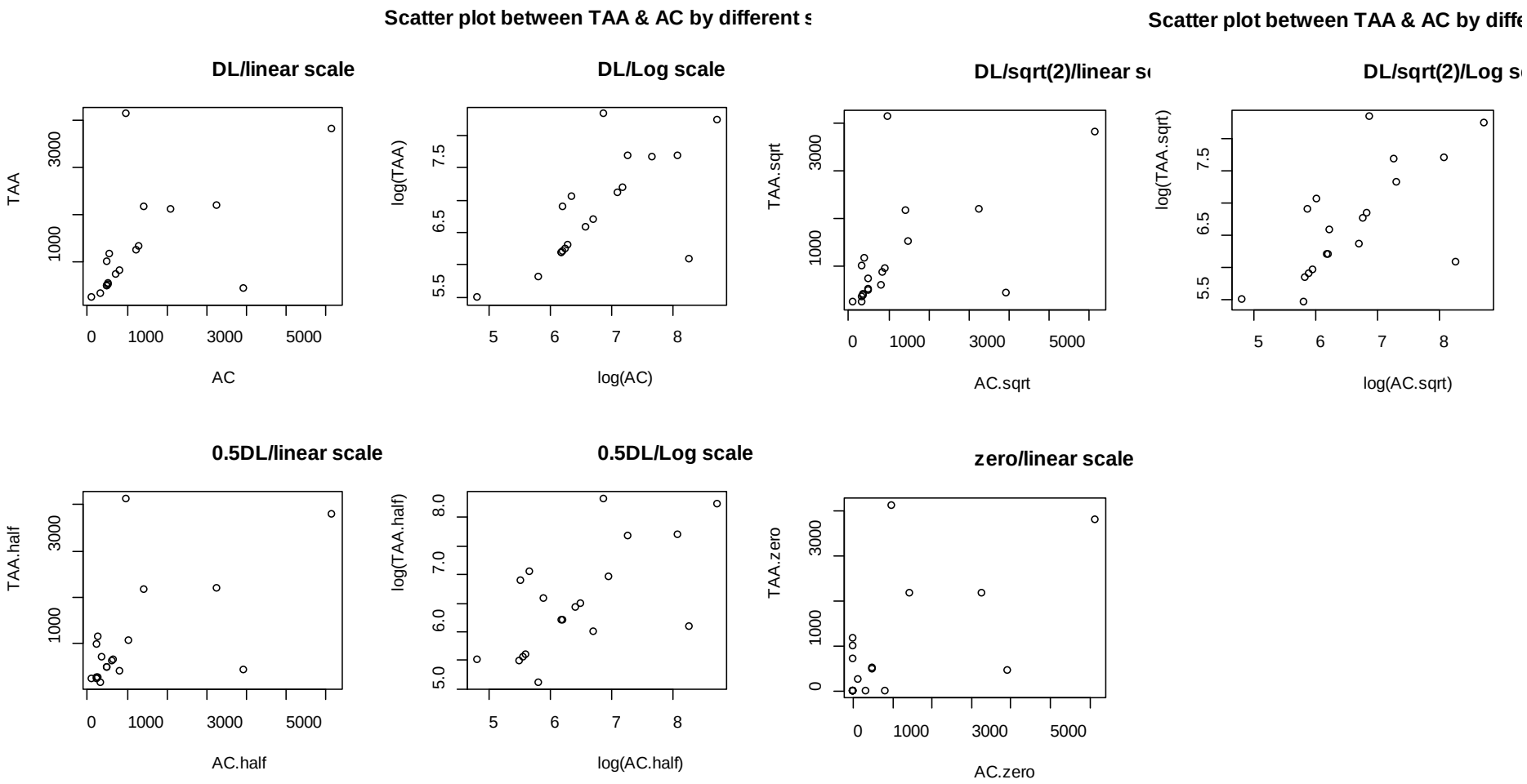

Figure 4: Scatter plots between TAA and AC by two scales with substitutions

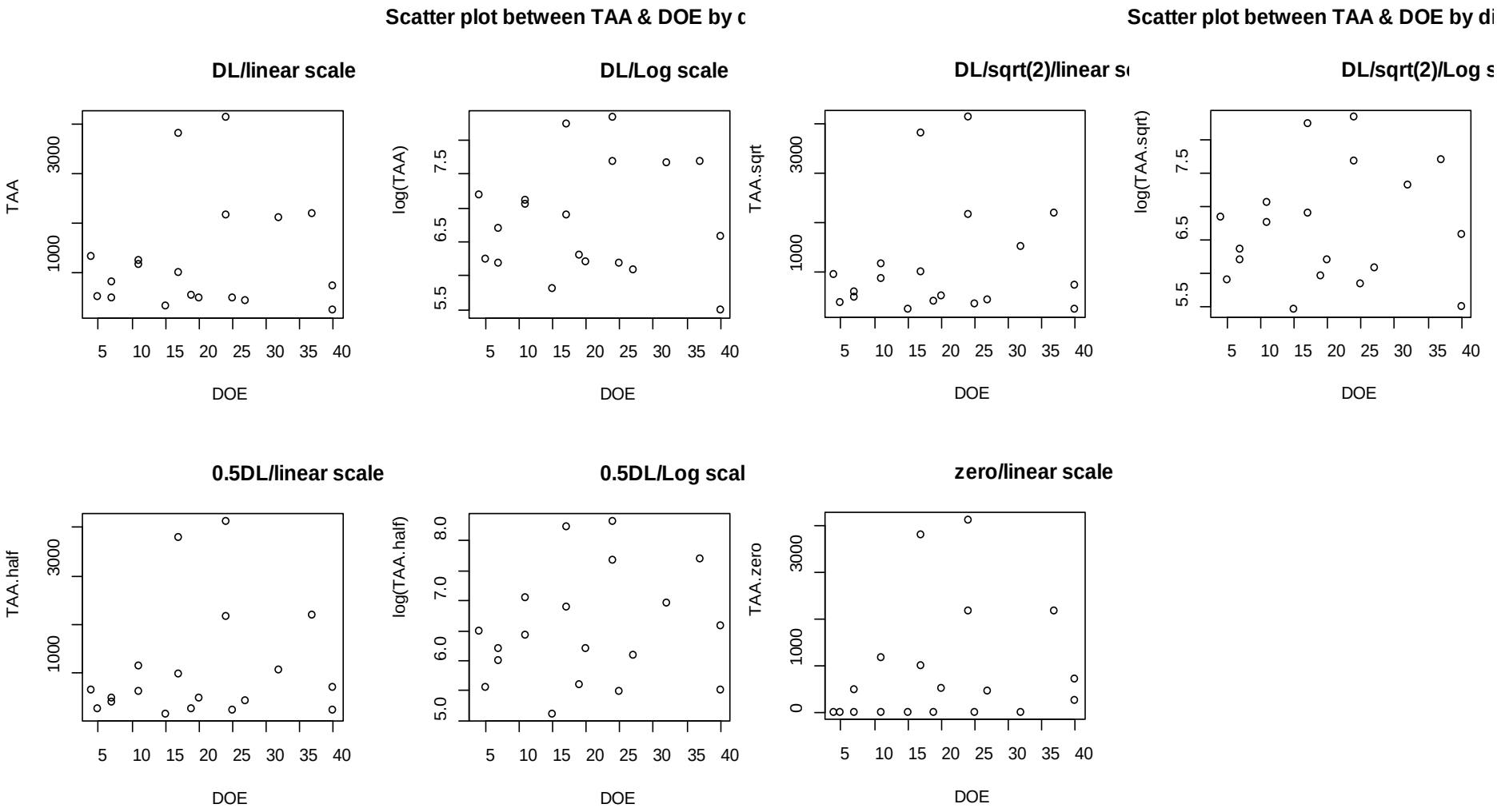

Figure 5: Scatter plots between TAA and DOE by two scales with substitutions

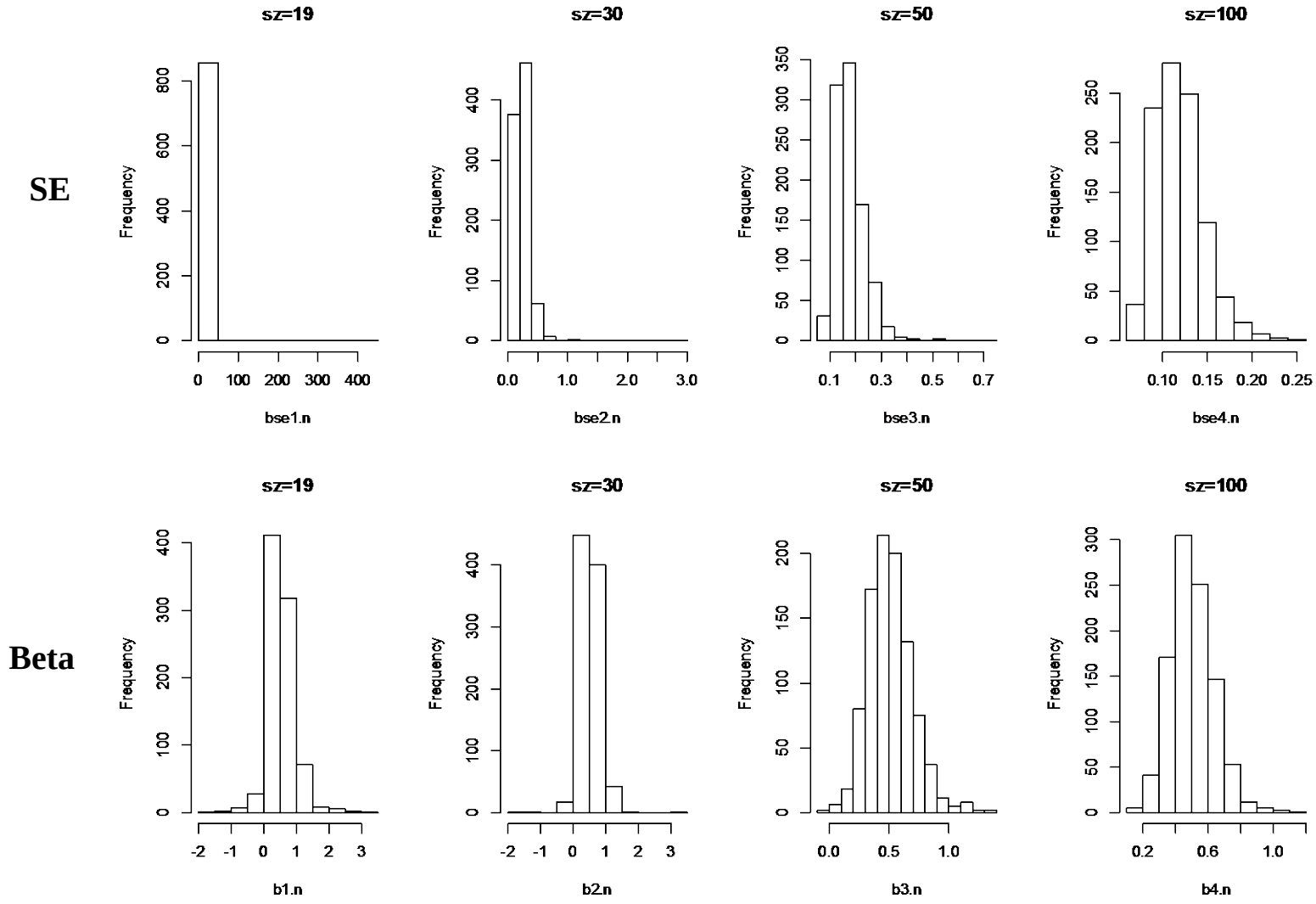

Figure 6: Histograms of the MLE of Beta estimate and its SE estimate in 1000 simulations

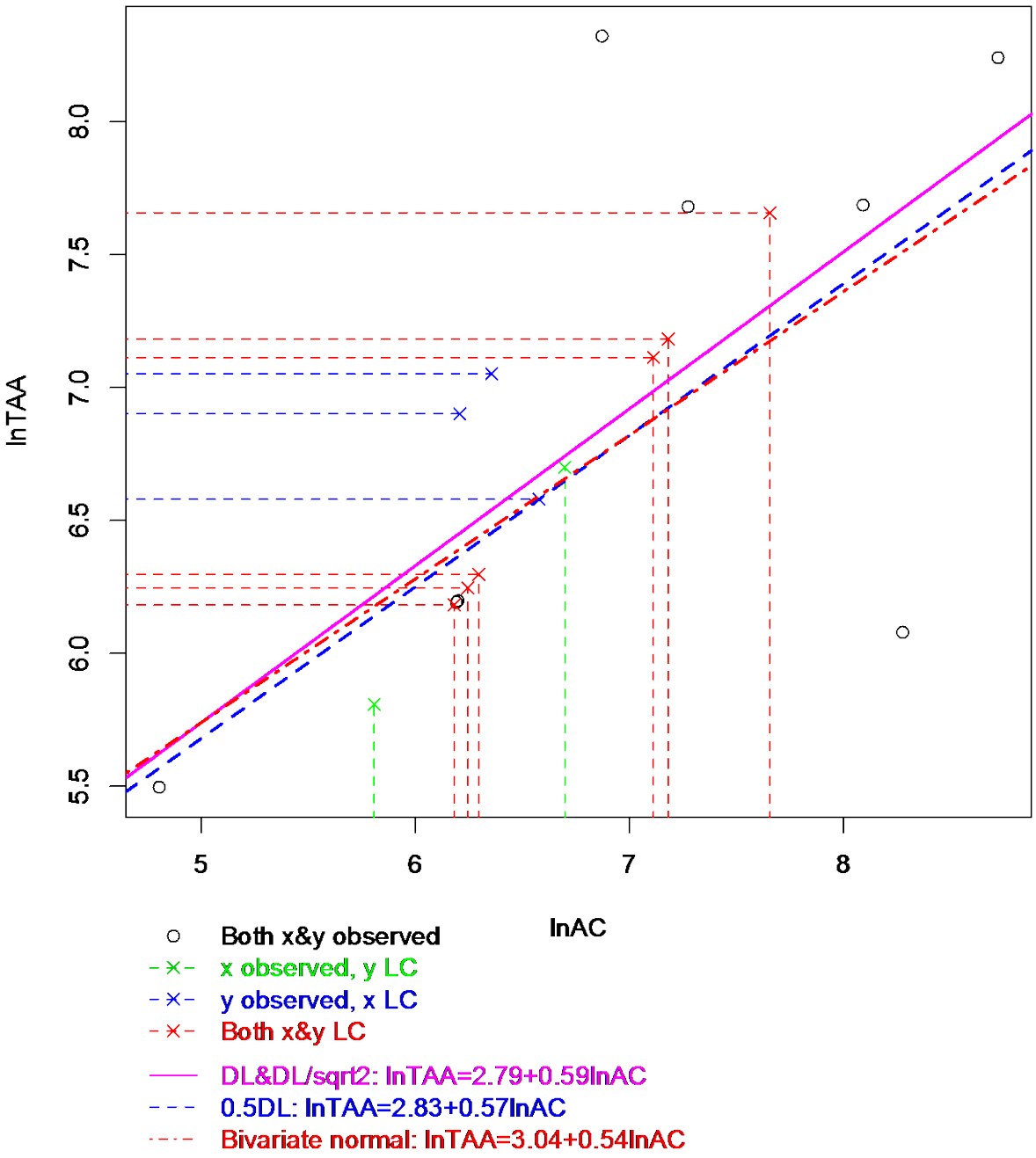

Figure 7: Scatter plot between log-scale of TAA and AC lung fiber concentration and the fitted lines of the simple linear regression model with substitution and the bivariate normal regression model

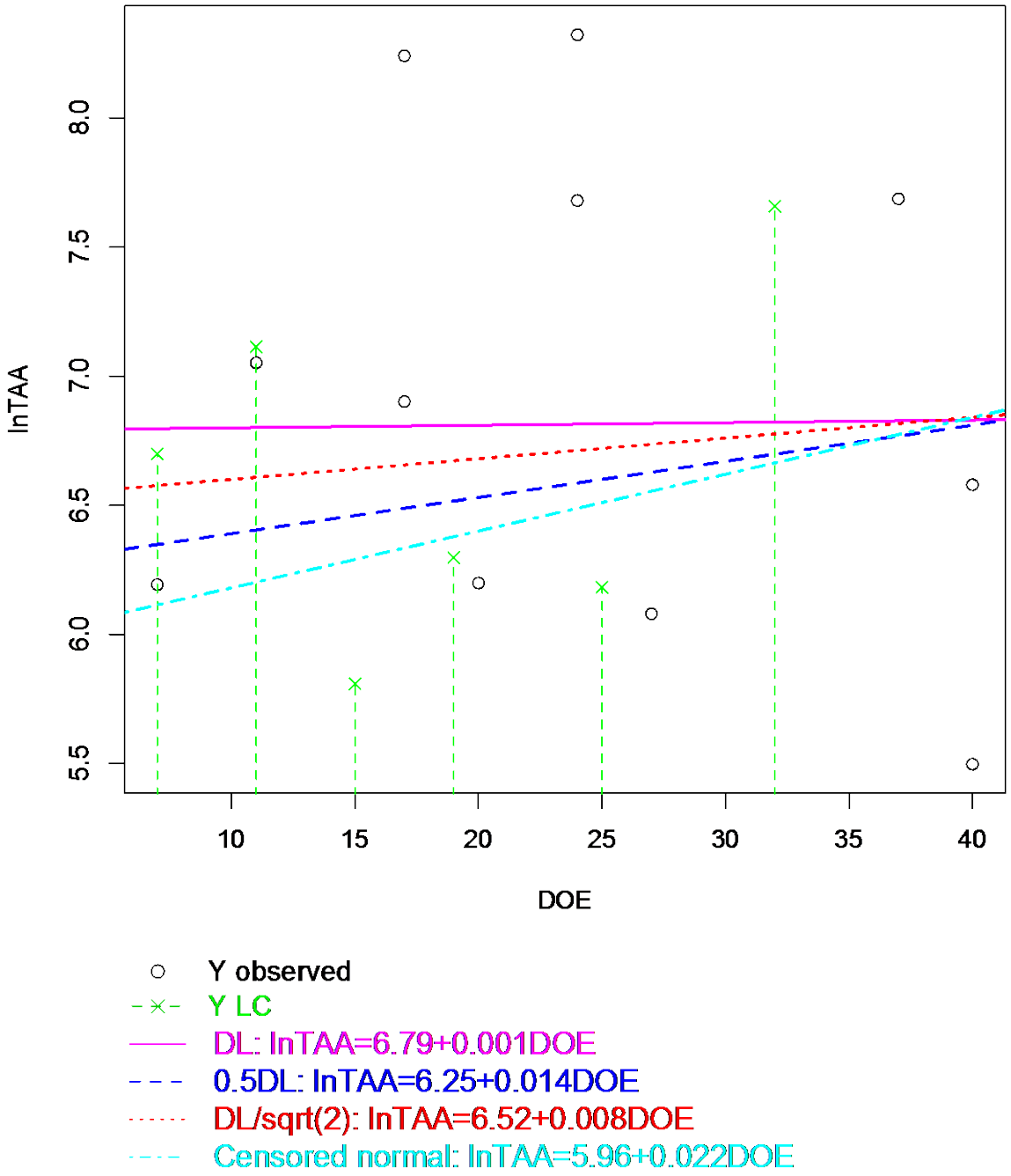

Figure 8: Scatter plot between log-scale of TAA lung fiber concentration and DOE and the fitted lines of three concentration models

BIBLIOGRAPHY

1. Selikoff IJ. *Asbestos and Diseases.* New York: Academic; 1978.
2. Selikoff IJ. Health hazards of asbestos exposure. *Ann NY Acad Sci.* 1079; 330:1-881.
3. Landrigan PJ. The third wave of asbestos disease: exposure to asbestos in place: public health control. Introduction. *Ann N Y Acad Sci.* 1991; 643:1-625.
4. Roggli VL. *Pathology of Asbestos-Associated Disease.* Boston: Little, Brown; 1992:106-107
5. Roggli VL. Malignant mesothelioma and occupational exposure to asbestos: a clinicopathological correlation of 1445 cases. *Ultrastruct Pathol.* 2002; 26(2):55-65.
6. Krstev S. Mortality among shipyard Coast Guard workers: a retrospective cohort study. *Occup Environ Med.* 2007;64(10):651-8.
7. Matanoski GM. Cancer risks and low-level radiation in U.S. shipyard workers. *J Radiat Res.* 2008; 49(1):83-91.
8. Giarelli L. Malignant mesothelioma of the pleura in the Trieste-Monfalcone area, with particular regard to shipyard workers. *Med Lav.* 1997; 88(4):316-20.
9. Edge JR. Malignant mesothelioma of the pleura in Barrow-in-Furness. *Thorax.* 1978; 33(1):26-30.
10. Connelly RR. Demographic patterns for mesothelioma in the United States. *J Natl Cancer Inst.* 1987; 78(6): 1053-60.

11. Selikoff IJ. Mortality experience of insulation workers in the United States and Canada, 1943--1976. *Ann N Y Acad Sci.* 1979; 330:91-116.
12. Ribak J. Malignant mesothelioma in a cohort of asbestos insulation workers: clinical presentation, diagnosis, and causes of death. *Br J Ind Med.* 1988; 45(3):182-187.
13. Langer AM. Mesothelioma in a brake repair worker. *Lancet.* 1982; 2(8307):1101-3.
14. Lynch JR. Brake lining decomposition products. *J Air Pollut Control Assoc*. 1968; 18(12):824-6.
15. Moore LL. Asbestos exposure associated with automotive brake repair in Pennsylvania. *Am Ind Hyg Assoc J.* 1988; 49:A12–A13.
16. Roggli VL. Amphiboles and chrysotile asbestos exposure. *Am J Ind Med.* 1988; 14:245–6.
17. Williams RL. Asbestos brake emissions. *Environ Res.* 1982; 29:70–82.
18. Wong O. Malignant mesothelioma and asbestos exposure among auto mechanics: appraisal of scientific evidence. *Regul Toxicol Pharmacol.* 2001; 34:170–7. (review paper 1)
19. Goodman M. Mesothelioma and lung cancer among motor vehicle mechanics: a meta-analysis. *Ann Occup Hyg.* 2004; 48(4):309-26. (review paper 2)
20. Anderson AE. Asbestos emissions from brake dynamometer tests. 1973;Ford Motor Co.
21. Eastern Research Group Inc. Report of the Expert Panel on health effects of asbestos and synthetic vitreous fibers: influence of fiber length, 17 March 2003. Available at www.atsdr.cdc.gov/HAC/asbestospanel/asbestostoc.html.
22. Langer AM. Reduction of the biological potential of chrysotile asbestos arising from conditions of service on brake pads. *Regul Toxicol Pharmacol.* 2003; 38:71–7.

23. Wong O. Chrysotile asbestos, mesothelioma, and garage mechanics. *Am J Ind Med.* 1992; 21:449–55

24. Dunnigan J. Linking chrysotile asbestos with mesothelioma. *Am J Ind Med.* 1988; 14(2):205-9.

25. Roggli VL. Asbestos fiber type in malignant mesothelioma: an analytical scanning electron microscopic study of 94 cases. *Am J Ind Med.* 1993; 23(4):605-14.

26. McDonald JC. Tremolite, other amphiboles, and mesothelioma. *Am J Ind Med*. 1988; 14(2):247-9.

27. Butnor KJ. Exposure to brake dust and malignant mesothelioma: a study of 10 cases with mineral fiber analyses. *Ann Occup Hyg.* 2003; 47(4): 325-30.

28. Roggli VL. Letter to the Editor, Comments on Asbestos Fibre Concentrations in the Lungs of Brake Workers: Another Look. *Ann Occup Hyg.* 2009; 53(2): 191.

29. Finkelstein MM. Asbestos Fibre Concentrations in the Lungs of Brake Workers: Reply to Roggli et al. *Ann Occup Hyg.* 2009; 53(2): 192.

30. Marsh GM. Asbestos fiber concentrations in the lungs of brake repair workers: commercial amphiboles levels are predictive of chrysotile levels. *Inhal Toxicol.* 2011; 23(12): 681-8.

31. Helsel DR. Less than obvious - statistical treatment of data below the detection limit. *Environ scitechnol.* 1990; 24:1766–74.

32. Helsel DR. More than obvious: Better methods for interpreting nondetect data. *Environ scitechnol.* 2005; 39:419A–23.

33. Helsel DR. Much ado about next to nothing: incorporating nondetects in science. *Ann Occup Hyg*. 2010 Apr; 54(3): 257-62.

34. Antweiler RC. Evaluation of statistical treatments of left-censored environmental data using coincident uncensored data sets: I. Summary statistics. *Environ Sci Technol.* 2008; 42(10): 3732-8.

35. Lubin JH. Epidemiologic evaluation of measurement data in the presence of detection limits. *Environ Health Perspect.* 2004; 112(17): 1691-6.

36. Feigelson,E.D. Statistical methods for astronomical data with upper limits. I - Univariate distributions. *Astrophysical Journal.* 1985; 293:192-206

37. Succop PA. Imputation of data values that are less than a detection limit. *J Occup Environ Hyg.* 2004; 1(7): 436-41.

38. Hewett P. A Comparison of Several Methods for Analyzing Censored Data. *Ann Occup Hyg.* 2007;51(7): 611-632.

39. Chu HT. On estimation of bivariate biomarkers with known detection limits. *Environmetrics.* 2008;19(3):301-317

40. Helsel DR. Statistical for censored environmental data using Minitab and R. Second Edition. New Jersey. John Wiley & Sons, Inc; 2012: 62-63.

41. Helsel DR. Statistical for censored environmental data using Minitab and R. Second Edition. New Jersey. John Wiley & Sons, Inc; 2012: 87-92.

42. Helsel DR. Statistical for censored environmental data using Minitab and R. Second Edition. New Jersey. John Wiley & Sons, Inc; 2012: 92-93.

43. Jacqmin-Gadda H. Analysis of left-censored longitudinal data with application to viral load in HIV infection. *Biostatistics*. 2000,1(4): 355-68.

44. Moulton LH. A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics*. 1995; 51(4): 1570-8.

45. Thompson ML. Linear regression with Type I interval- and left-censored response data. *Environmental and Ecological Statistics.* 2003; 10(2): 221-230.

46. Little R J.A. Regression With Missing X's: A Review. Journal of the American Statistical *Association.* 1992; 87(420): 1227-1237.

47. Nie L. Linear regression with an independent variable subject to a detection limit. *Epidemiology.* 2010; 21(4): 17-24.

48. Helsel DR. Statistical for censored environmental data using Minitab and R. Second Edition. New Jersey. John Wiley & Sons, Inc; 2012: 236-237.

49. Helsel DR. Statistical for censored environmental data using Minitab and R. Second Edition. New Jersey. John Wiley & Sons, Inc; 2012: 258-264.