Thailand Statistician
April 2021; 19(2): 339-360
http://statassoc.or.th
Contributed paper

A Statistical Profile of Arsenic Prevalence in the Mekong Delta Region

Uyen Huynh [a], Nabendu Pal*[b], Buu-Chau Truong [c] and Man Nguyen [a]
[a] Department of Mathematics, Faculty of Science, Mahidol University, Bangkok, Thailand.
[b] Department of Mathematics, University of Louisiana at Lafayette, Lafayette, Louisiana, USA.
[c] Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam.
*Corresponding author; e-mail: nabendupal@gmail.com

Received: 7 December 2019
Revised: 22 March 2020
Accepted: 10 May 2020

Abstract

This work is a novel approach to model the concentration of arsenic in groundwater in An Phu district, An Giang province, in the Mekong Delta Region (MDR) of Vietnam, based on data available from a sample of water-wells. Arsenic contamination is a major problem in Vietnam, especially in the MDR where a large population depends on the groundwater pumped through tubewells for daily consumption as well as irrigation. It is a time consuming and expensive process to do a detailed chemical test to measure arsenic at every possible site of groundwater extraction. However, using a suitable statistical regression model we can construct a statistical profile of arsenic concentration over a suitable area which then can be further used to predict arsenic concentration at a new site within the same surveyed area just based on its geographic characteristics. First, we provide a brief overview of the textbook type regression model based on normally (or, Gaussian) distributed errors. Then, we provide a more general model based on the skew-normal distribution (SND) for the errors. It should be noted that the SND is a generalization of the regular normal probability distribution, and hence provides a greater flexibility in our regression model. We provide a step by step approach to estimate all parameters of the regression model which is not only new and easy, but also quite different from the approaches followed by the other researchers. The sampling distributions of the parameter estimates are then studied using the bootstrap method which enables one to construct interval estimates of the model parameters. The methodology of using SND errors to develop a suitable regression model to build a profile of arsenic prevalence in the MDR can easily be adopted by the investigators for many other similar applied research problems.

Keywords: Skew-normal distribution, nonlinear regression, parameter estimation, bootstrap method, confidence interval, Akaike information criterion (AIC).

1. Introduction

1.1. Preliminaries

Arsenic is the twentieth most abundant element in the earth's crust, fourteenth most abundant in seawater, and twelfth most abundant in the human body. It is a naturally occurring omnipresent element which is found in the atmosphere, pedosphere, hydrosphere, and biosphere. There are four oxidation states of Arsenic: gaseous arsine, in the form of AsH_3; elemental arsenic (characteristic

of O oxidation state); arsenite; and arsenate, see in Nguyen (2008). Arsenic is water soluble, and is almost never in its elemental form; rather, it forms compounds which are called arsenicals. From a geochemical point of view, arsenicals are often associated with sulphurous minerals made up of sulphur, iron, gold, silver, copper, antimony, nickel, and cobalt. Further, it is detected in more than 200 different minerals, see in Lièvremont et al. (2009).

Although arsenic is the twelfth most abundant element in the human body, it is highly toxic in excess amounts. An elemental arsenic concentration of 48 μg/L is the lethal dose for rats, which roughly translates into 125 mg lethal dosage for an average middle-aged male, see in Altug (2003); Ahuja (2008). Due to this lethal dosage, arsenic can be extremely toxic when it is found in the food chain, especially over a sustained period of time. Its toxicity is also dependent on hydrogen potential (pH), redox potential, organic matter content, and the presence of other substances like iron and magnesium.

The European Union (EU), the World Health Organization (WHO), and the United States Environmental Protection Agency (USEPA), - have all recognized arsenic contamination as one of the major human health hazards. The WHO guidelines for safe level of arsenic ingestion is a concentration of 10 μg/L in drinking water, and a limit of 100 μg/L in untreated water prior to being processed for consumption. The maximum safe limit of arsenic ingestion for an average middle-aged man is about 220 μg per day, see in Ahuja (2008). Here, the word 'ingestion' includes drinking water from groundwater sources and consuming food made from crops irrigated with arsenic-rich water and animals that were fed food/water with arsenic additives.

The USEPA defines arsenic as a persistent, bio-accumulative, and toxic chemical having the ability to accumulative in the air, soil, and water. The arsenic pollution was first detected in Taiwan in 1961 and later in Belgium, Netherlands, Germany, Italy, Hungary, Portugal, The Philippines, Ghana, USA, Chile, Mexico, Argentina, and Thailand. In 1992, the toxicity of arsenic was found as a major public health hazard in certain districts of West Bengal province in India. Later, in neighboring Bangladesh, arsenic poisoning was found to be a major public health problem. Number of people affected due to this hazard was estimated to be 23 million in 1997, and subsequently it rose to 60 million in 2005.

Apart from the natural resources, arsenic pollution in the environment is rising rapidly due to human use (or abuse) of harmful chemicals. Smelting of nonferrous ores create arsenic trioxide which escapes into the atmosphere, and eventually settles in neighboring fields and streams (with rain-water). Used electronics, batteries, filters, etc. dumped into landfills also cause arsenic leakage into the soil which eventually seeps into groundwater. Plants absorb the arsenic-contaminated groundwater, and when these plants are fed to livestock, the arsenic finds its way up the foodchain in excess amount.

1.2. Arsenic pollution in Vietnam

The soil layer in Vietnam, like those in the eastern region of India, and Bangladesh, derive its sediments from the Himalayas washed down to the Mekong and Red River deltas. In the Mekong Delta Region (MDR), which is the primary focus of this study, there is also arsenious shale that release arsenicals into the groundwater, see in Nguyen (2008).

In addition to the natural causes, the Vietnam War's 'Operation Ranch Hand'(ORH) also greatly contributed to the arsenic contamination crisis in Vietnam. The ORH was a United States military project for aerial spraying of herbicides in South Vietnam, mainly in the MDR, to clear foliage and vegetation cover to help improve enemy detection and enhanced troop deployments. Over a period of ten years, from August 1961 to October 1971, multiple herbicides were sprayed over South Vietnam (see Young and Regigani (1988), Nakamura (2007)). One of the major herbicidal chemicals was Agent Blue which damaged nearly two million hectares of land in South Vietnam, primarily near Saigon and Da Nang. Agent Blue was composed of 31.1% arsenical compounds, and the lethal concentration of Agent Blue for rats was 3.5 μg/L.

In both the MDR in the south and the Red River Delta (RRD) in the north, farmers irrigate

their land by drawing groundwater through tubewells. Widespread and intensive use of tubewells have caused the aquifers experience sustained overdraft since the mid-1990s in these two deltas. This over exploitation of the aquifers has caused saline incursion and land subsidence which in turn has exacerbated the arsenic contamination. Arsenic pollution came to be known in Vietnam in 1998 when samples of groundwater were tested from more than 150,000 tubewells supported by the UNICEF to replace unsanitary surface water. The results of the survey indicated that about 12% of the water samples (i.e 22,450 samples) had arsenic concentrations higher than 50 μg/L. Tran et al. (2011) indicated that An Giang province is one of the areas that had been recorded with the highest arsenic concentrations in groundwater in the MDR where people used not only groundwater but also surface water for drinking and irrigation. Therefore, arsenic led to serious public health hazards such as severe skin damage and/or cancer.

Urban and semi-urban water treatment facilities, which are geared toward providing safe drinking water in the MDR, exploit aquifers between thirty and seventy meters deep, while private tubewells mainly pump water from shallow aquifers which are twelve to forty-five meters deep. However, testing soil and/or water samples at every location and various depths may be very costly. Therefore, a statistical profile of a region can be built based on sampled sites characteristics and observed arsenic concentration. Once such a statistical profile (through a suitable regression model) is built, it can be exploited to predict arsenic concentration at a new location, within the same region, based on the locations geographic characteristics. The objective of our study is to demonstrate how such a statistical model can be employed effectively based on a recent survey dataset.

1.3. Background of dataset

Researchers have been carrying out surveys to study arsenic concentration in water and soil in two major areas within Vietnam, namely the MDR in southern part, and the RRD in northern part of Vietnam. One of the worst affected areas within MDR is An Giang province, bordering Cambodia. Researchers from the Faculty of Resources and Environment, University of Technology, Ho Chi Minh City (HCMC), Vietnam, studied water samples from water-wells at selected sites within An Phu district of An Giang province between January 2014 and October 2015. The primary goal of the researchers have been to study patterns, if there is any, in arsenic concentration based on the water-wells' locations (i.e., distance from the near by Bassac River), their depths, and the time of the data collection. (Apart from arsenic, the researchers also collected other information, such as - iron, lead, mercury, etc. as well as pH level.) The original dataset has many missing values, and the survey could have been much better had the statisticians been taken on board before the whole exercise. However, we have made our best effort to model the existing data with our proposed new approach.

After a careful study of the dataset, dropping some doubtful observations, we have decided to focus on a partial dataset with complete observations from 29 locations where arsenic concentration was measured in May 2014 and August 2015. (The missing observations were recorded as "n/a", and it is not clear whether they were truly "not available" or "too close to zero", i.e., too negligible to be recorded as a substantial value.) A reasonably good statistical model can be extremely useful for the farmers and landowners in particular, of the heavily agricultural Khanh An Commune that covers the sampling sites, and the applied researchers, in general, to predict arsenic level in water-wells at newer sites without undertaking an expensive process of chemical analysis. The original survey and the details of the study have been reported by Vo et al. (2015, 2016). The dataset we use in this study has been provided in Table 2 (see the Appendix), and also available in Pham (2015).

1.4. The objective of our work

The main objective of our work has been to use the skew-normal distribution (SND) errors in our regression model which provides greater flexibility, and hence improves over the usual normal error model. In Section 2 we first review the classical normal based regression model briefly, followed by an introduction of the SND, its basic properties, and how it can be used as the error distribution in a regression set-up. In Section 3 we present our detailed method of estimation of the regression

coefficient vector $\boldsymbol{\beta}$, followed by estimation of σ and λ, the SND scale and skew parameters respectively. Ours is a combination of the ordinary least squares method coupled with the method of moments estimation which has not been used before in the literature. This method is not only easy to implement, but also avoids the computational challenges one faces with maximum likelihood estimation under the SND model. Section 4 is devoted to studying the sampling distributions of the parameter estimates through bootstrap method. The sampling distributions of the key estimators help us make further inferences and check model goodness of fit through Akaike Information Criterion (AIC). Finally, in Section 5 we revisit the arsenic problem and explain the arsenic pollution trend via the explanatory variables.

2. The Regression Set-up and the SND Errors

2.1. The classical regression model under normality

Consider a classical multiple linear regression model as

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{(p-1)} X_{(p-1)} + \varepsilon, \tag{1}$$

where Y is the dependent (or response) variable, and $\boldsymbol{X} = (1, X_1, \ldots, X_{(p-1)})'$ is the vector of independent variables. The random error ε is typically assumed to follow $N(0, \sigma^2)$ distribution, where the variance σ^2 is thought to be unknown.

Based on a random sample of size n (i.e., observations from n randomly selected subjects or units), the above generic model (1) is expanded further as

$$Y_j = \beta_0 + \beta_1 X_{1j} + \cdots + \beta_{(p-1)} X_{(p-1)j} + \varepsilon_j, \;\; j = 1, 2, \ldots, n, \tag{2}$$

where the individual errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are assumed to be $i.i.d.$ (independent and identically distributed) $N(0, \sigma^2)$. This is the typical 'textbook style' regression model which is based on the following basic assumptions:

(1a) The individual errors are all $i.i.d.$, because the sampled subjects or units are assumed to be independent, and they are subject to the same set-up.

(1b) The errors are normally distributed and have the same variance (which stems from the identical distribution assumption in (1a)).

There is another tacit assumption which says that the dependent variable Y obeys the additive law of the linearly explicable 'signal' ($\beta_0 + \beta_1 X_1 + \cdots + \beta_{(p-1)} X_{(p-1)}$) and the inexplicable 'random noise' ε. Obviously, this simple additive linear model may seem too simplistic, and may not be good enough to explain many natural phenomena. But, we do not start off with a simplistic linear model blindly, rather try to understand the basic relationship between Y and each X_i by plotting the scatterplots. If the perceived relationship between Y and X_i's show some nonlinear pattern then we often transform the variable suitably to attain some linearity before proceeding further with the estimation of the relevant parameters, and subsequent statistical inferences. So assuming that (2) holds, after making suitable adjustments to the original study variables, the regression coefficients, i.e., the vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{(p-1)})'$, is typically estimated by minimizing the sample mean squared error, say Δ_0, defined as

$$\Delta_0 = (\sum_{j=1}^{n} \varepsilon_j^2)/n = \sum_{j=1}^{n} \left(Y_j - \beta_0 - \beta_1 X_{1j} - \ldots - \beta_{(p-1)} X_{(p-1)j}\right)^2 /n, \tag{3}$$

which results into the Ordinary Least Squares Estimate (OLSE) $\widehat{\boldsymbol{\beta}}^N$ given as (where the superscript "N" in the estimator indicates the underlying normal distributed (ND) model)

$$\widehat{\boldsymbol{\beta}}^N = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{Y}, \tag{4}$$

where $\mathbb{X} = ((X_{ij}))$, $0 \leq i \leq (p-1)$, $1 \leq j \leq n$, is the design matrix of order $n \times p$, with $X_{0j} \equiv 1$, and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$, assuming that $(\mathbb{X}'\mathbb{X})$ is nonsingular. (Here j indicates the row number and i indicates the column number.) Then, the 'fitted' value of $\boldsymbol{Y}$ is defined as $\widehat{\boldsymbol{Y}}^N = \mathbb{X}\widehat{\boldsymbol{\beta}}^N = H\boldsymbol{Y}$, where the H - matrix, also known as the 'projection matrix', is defined as $H = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$. Note that the error vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$, an integral component of the model (2), is never observed. But an idea about the $i.i.d.$ errors ε_j's can be obtained from the residuals e_j^N's defined as $\boldsymbol{e}^N = (e_1^N, \ldots, e_n^N)' = (\boldsymbol{Y} - \widehat{\boldsymbol{Y}}^N)' = (I - H)\boldsymbol{Y}$. Also, the unbiased estimator of σ is found as $\widehat{\sigma}^N = \|\boldsymbol{e}^N\|/\sqrt{n-p}$.

Alternatively, one may follow nonparametric approaches like Kernel Regression or Kriging method, see in Henderson and Parmeter (2015) or a semiparametric method, see in Ichimura (1993), but these are often subjective, usually more complicated, and how effective such methods are for small to moderate sample sizes remains to be seen.

Remark 1 Though the OLSE can be seen as an estimation method of $\boldsymbol{\beta}$ free from the model (i.e., the underlying distributions of ε_j's), the sampling distribution of the OLSE is found only by using the assumed model. Further, though the OLSE is found by minimizing $\Delta_0 = \sum_{j=1}^{n} \varepsilon_j^2/n$, its precursor is $\Delta_0 = \sum_{j=1}^{n} (Y_j - E(Y_j))^2/n$, which ideally should be minimized. When ε_j's are centered at zero, there is no difference between these two expressions of Δ_0. However, they can be different when ε_j's are not centered at zero, as we will see in Subsection 2.3, and we will use the latter expression to find the OLSE under SND.

2.2. A brief overview of the SND

A random variable (r.v.) W is said to follow a skew-normal distribution with location parameter μ, scale parameter σ, and skew parameter λ (henceforth, $SND(\mu, \sigma, \lambda)$ provided its *pdf* is given as

$$f(w|\mu, \sigma, \lambda) = (2/\sigma)\, \phi\left((w-\mu)/\sigma\right) \Phi\left(\lambda(w-\mu)/\sigma\right), \tag{5}$$

where $\phi(.)$ and $\Phi(.)$ are the standard normal *pdf* and *cdf* respectively. Note that $\lambda = 0$ makes $SND(\mu, \sigma, \lambda)$ reduce to $N(0, \sigma^2)$. The above distribution (5) is positively (negatively) skewed if $\lambda > (<)\ 0$.

The above $SND(\mu, \sigma, \lambda)$ gives a nice flexibility to the usual $N(0, \sigma^2)$, and it was made popular by Azzalini (1985, 1986) by his pioneering work on various properties of this distribution. For a good review of many characterization properties of SND one may refer to Gupta et al. (2003), and Arnold and Lin (2004). For some distributional/sampling properties one can see the latest work of Thuithad and Pal (2019).

Some useful properties of $SND(\mu, \sigma, \lambda)$ are given as follows:

(i) The r.v. $W \sim SND(\mu, \sigma, \lambda)$ if and only if $W_* = (W-\mu)/\sigma \sim SND(0, 1, \lambda)$, known as the standard SND.

(ii) The r.v. $W \sim SND(\mu, \sigma, \lambda)$ if and only if $(-W) \sim SND(-\mu, \sigma, -\lambda)$.

(iii) As $\lambda \to \pm\infty$, $W_* = (W-\mu)/\sigma \sim SND(0, 1, \lambda) \to \pm|Z|$, where $Z \sim N(0, 1)$.

(iv) $W_*^2 = (W-\mu)^2/\sigma^2 \sim \chi_1^2$.

(v) If U_1, U_2 are $i.i.d.$ $\sim N(0, 1)$, then $(\lambda/\sqrt{1+\lambda^2})|U_1| + (1/\sqrt{1+\lambda^2})U_2 \sim SND(0, 1, \lambda)$.

(vi) If $W \sim SND(\mu, \sigma, \lambda)$, then $E(W) = \mu + \sigma\sqrt{2/\pi}(\lambda/\sqrt{1+\lambda^2})$;

$$E(W - E(W))^2 = \sigma^2 \left\{1 - ((2/\pi)\lambda^2/(1+\lambda^2))\right\}; \text{and}$$
$$E(W - E(W))^3 = \sigma^3\sqrt{2/\pi}((4/\pi) - 1)\left\{\lambda/\sqrt{1+\lambda^2}\right\}^3.$$

2.3. Regression model with skew-normal errors: Challenges

We consider the same expression as in (2) with the assumption that the errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are now assumed to be $i.i.d.$ $SND(0, \sigma, \lambda)$ (which gives more flexibility than the $N(0, \sigma^2)$ model). In other words, the response variable from n units are independent Y_j's where

$$Y_j = \mu_j + \varepsilon_j \sim SND(\mu_j, \sigma, \lambda), \tag{6}$$

where

$$\mu_j = \beta_0 + \beta_1 X_{1j} + \cdots + \beta_{(p-1)} X_{(p-1)j},\ 1 \leq j \leq n. \tag{7}$$

The natural question is: 'How to the estimate all parameters in the model (6)?'
In the case of normal errors, the OLSE of $\boldsymbol{\beta}$ (in (4)) is also the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$. But in the case of SND errors, as stated in (6), OLSE of $\boldsymbol{\beta}$ is different from the MLE. Worse, the MLE may not even exist, a fact that is routinely over-looked by many researchers. The likelihood function based on (6) is

$$L = L(\boldsymbol{\beta}, \sigma, \lambda|\text{data}) = \prod_{j=1}^{n} \left\{(2/\sigma)\phi((Y_j - \mu_j)/\sigma)\Phi(\lambda(Y_j - \mu_j)/\sigma)\right\}, \tag{8}$$

where μ_j is given in (7). The values of independent variables $(X_{1j}, \ldots, X_{(p-1)j}), 1 \leq j \leq n$, are fixed, $\boldsymbol{\beta} \in \mathbb{R}^p, \sigma > 0$ and $\lambda \in \mathbb{R}$. When all Y_j's are greater than $max(\mu_j)$, then the above likelihood (8) is monotonically increasing in λ, and hence the MLE of λ turns out to be $+\infty$. Similarly, when all Y_j's are smaller than $min(\mu_j)$, then the likelihood gets maximized at $\lambda = -\infty$ (irrespective of σ). Since the MLE of λ can take the values $\pm\infty$ with a positive probability, which in turn makes estimation of $\boldsymbol{\beta}$ and σ impossible, studying the sampling distribution of the MLEs becomes futile as their moments do not exist. In such a situation, typically a penalty is attached to the log-likelihood function before maximization. For example, one may consider maximizing

$$L_* = lnL - h(\lambda),$$

where $h(\lambda)$ is a suitable penalty function (such as, $h(\lambda) = (constant)\lambda^2$), with respect to the model parameters $\boldsymbol{\beta}, \sigma, \lambda$. The difficulty is that the numerical maximization of L_* is far from satisfactory as our computational regression experience has shown. All the standard software or packages suggest using a starting point (initial value of at least some of the parameters) which doesn't give any satisfactory result. (We have tried an iterative process to maximize L_* where an initial value of $\boldsymbol{\beta}$ was assigned to maximize L_* w.r.t. (σ, λ) first, and then use this suboptimal value of (σ, λ) as the initial value to maximize L_* w.r.t. $\boldsymbol{\beta}$, and then continue this iterative process with the hope that it might converge after a few cycles, but all in vain.) Therefore, in this paper our estimation of all the model parameters is based on the least squares method coupled with the method of moments estimation. To best of our knowledge, the estimation of $\boldsymbol{\beta}, \sigma$, and λ that we present here is completely new, and the estimators have fairly well structured closed forms.

While we consider our proposed regression model (6) with SND error, we would like to point out a few other recent papers along the same line, and how our work differs from those.

Cancho et al. (2010) considered a regression model with SND error to model palm oil plant weight based on the age of the plant. (To be precise, these authors considered a logistic growth model of a plant weight as a function of the plant age.) However, their model parametrization is different from ours. Their σ and λ are our $\sigma/\sqrt{1+\lambda^2}$ and $\lambda\sigma/\sqrt{1+\lambda^2}$, respectively. These authors used a characterization property of the SND with a latent variable to apply the Expectation-Maximization Algorithm (EMA) to find the MLE of the model parameters $\boldsymbol{\beta}, \sigma$ and λ. However, the EMA can be applied assuming that the MLE exists, and it converges, but there is no guarantee of that happening as mentioned earlier (see the reason after (8)).

Alhamide et al. (2016) used a multivariate skew normal distribution (MSND) for the error vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ with location parameters $-(\tau + h(\tau))\boldsymbol{\lambda}$, scale matrix Σ, skew parameter $\boldsymbol{\lambda}$, and an additional shape parameter τ. They suggested using $\boldsymbol{\lambda} = \lambda \mathbf{1}, \Sigma = \sigma^2 I_n$. But these authors claimed to have obtained the MLE of the parameters though it is not clear how (since τ and $h(\tau)$ remain unclear, and maximization of the log-likelihood function many suffer from the same difficulty as mentioned earlier).

In a similar manner, the MLE of the model parameters was suggested by Guedes et al. (2014). They also suggested Bayesian estimation using specific prior distribution and then using the MCMC algorithm. A similar approach had been taken by Pérez-Rodríguez et al. (2018).

Lachos et al. (2009) considered multivariate skew-normal independent distribution (MSNID) in a linear mixed model set-up. Sahu et al. (2003) considered a regression set up with a further generalization of SND errors. But again these authors implemented a Bayesian estimation of the model parameters assuming suitable prior distributions.

The focus of our work remains strictly frequentist as selection of priors in the aforementioned papers appear to be highly ad hoc and subjective. It is not clear how a small perturbation of these priors might alter the parameter estimates as the current literature lacks a sensitivity analysis on prior selection. (We plan to extend our current work in future in a Bayesian set-up with inclusion of a sensitivity analysis.)

3. Regression Model with SND Errors: Estimation of Parameters

3.1. Estimation of the model parameters

Under the model (6), since each $\varepsilon_j \sim SND(0, \sigma, \lambda)$,

$$\eta_j = E(Y_j) = \beta_0 + \beta_1 X_{1j} + \cdots + \beta_{(p-1)} X_{(p-1)j} + \gamma, \;\; 1 \le j \le n, \tag{9}$$

where $\gamma = \sigma\delta\sqrt{2/\pi},\ \delta = \lambda/\sqrt{1+\lambda^2}$. The sample mean squared error, as stated in Remark 1, is

$$\Delta_\gamma = \sum_{j=1}^{n} \left(Y_j - \beta_0 - \beta_1 X_{1j} - \cdots - \beta_{(p-1)} X_{(p-1)j} - \gamma\right)^2 / n, \tag{10}$$

which needs to be minimized as it was done for the normal errors case. Define $\beta_0^\gamma = \beta_0 + \gamma$, and $\boldsymbol{\beta}^\gamma = (\beta_0^\gamma, \beta_1, \ldots, \beta_{(p-1)})'$. Note that the only difference between $\boldsymbol{\beta}^\gamma$ and $\boldsymbol{\beta}$ is the presence of γ in $\boldsymbol{\beta}^\gamma$ (in the very first component of $\boldsymbol{\beta}^\gamma$). Also Δ_γ in (10) is same as Δ_0 in (3) only when $\gamma = 0$ which in turn happens only when $\lambda = 0$ (i.e., the SND takes the special case of normal distribution). Thus, minimization of $\Delta_\gamma = (\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}^\gamma)'(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}^\gamma)/n$ with respect to $\boldsymbol{\beta}^\gamma$ yields the OLSE of $\boldsymbol{\beta}^\gamma$ which the same as the one in (4), i.e.,

$$\widehat{\boldsymbol{\beta}}^{\gamma S} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{Y} = \widehat{\boldsymbol{\beta}}^N. \tag{11}$$

(The superscript "S" in the estimator of $\boldsymbol{\beta}^\gamma$ indicates the underlying SND model.) The first component of $(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{Y}$, which is a $(p \times 1)$ dimensional vector, is now estimating $(\beta_0 + \gamma)$, i.e., β_0^γ under the SND error model. Our task now is to extract the estimate of β_0 as well as that of γ, and this will be done through individual estimates of σ and λ as well. This in turn will be done through the method of moments estimation using the model residuals.

Derive the residual vector $\boldsymbol{e}^N = (e_1^N, \ldots, e_n^N)'$ as we did before under the normal errors, i.e., $\boldsymbol{e}^N = (\boldsymbol{Y} - \widehat{\boldsymbol{Y}}^N)$, i.e., $\boldsymbol{e}^N = \boldsymbol{Y} - \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{Y}$, but these residuals are now supposed to reflect unobservable errors $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ which are $i.i.d.\ SND(0, \sigma, \lambda)$. Define the first three residual raw moments as follows:

$$m_k = \left(\sum_{j=1}^{n} \left(e_j^N\right)^k / n\right), \;\; k = 1, 2, 3. \tag{12}$$

Now, let us look at the above three residual raw moments and see what they are supposed to represent. The quantity m_1 is supposed to represent $E(\varepsilon_j - \gamma)$. But since both m_1 and $E(\varepsilon_j - \gamma)$ are zero, equating m_1 with $E(\varepsilon_j - \gamma)$ doesn't provide any extra information. The quantity m_2 is supposed to reflect $E(\varepsilon_j - \gamma)^2 = \sigma^2 - \gamma^2$. Therefore, we obtain an equation by equating them as

$$m_2 = \sigma^2 - \gamma^2. \tag{13}$$

Finally, the quantity m_3 is supposed to reflect $E(\varepsilon_j - \gamma)^3 = (2 - \pi/2)\gamma^3$, and hence we obtain the equation

$$m_3 = (2 - \pi/2)\gamma^3, \tag{14}$$

where $\gamma = \sigma\delta\sqrt{2/\pi} = \sigma\lambda\sqrt{2/\pi}/\sqrt{1+\lambda^2}$.

The equation (14) gives us an estimate of γ directly as

$$\widehat{\gamma}^S = \{m_3/(2 - \pi/2)\}^{1/3}, \tag{15}$$

which helps us recover the estimate of β_0 from (11) as

$$\widehat{\beta}_0^S = \widehat{\beta}_0^{\gamma S} - \widehat{\gamma}^S = \Big\{\text{first element of } \widehat{\boldsymbol{\beta}}^{\gamma S} \text{ in (11)}\Big\} - \widehat{\gamma}^S, \tag{16}$$

So, the ordinary least squares estimate of $\boldsymbol{\beta}$ under SND is

$$\widehat{\boldsymbol{\beta}}^S = \widehat{\boldsymbol{\beta}}^{\gamma S} - (\widehat{\gamma}^S, 0, 0, \ldots, 0)'. \tag{17}$$

Remark 2 See the line after (9) where the parameter γ has been defined. While δ takes values in the interval $(-1, +1)$, the parameter σ can take any nonnegative value. Therefore, γ can take any real value. The estimate of γ obtained in (15) preserves this range as the residual third moment m_3 defined in (12) can take any real value.

Next, using (15) in (13) we can recover the estimate of σ as $\left(\widehat{\sigma}^S\right)^2 = m_2 + \left(\widehat{\gamma}^S\right)^2$, i.e.,

$$\widehat{\sigma}^S = \left\{m_2 + \left(\widehat{\gamma}^S\right)^2\right\}^{1/2}. \tag{18}$$

The last remaining piece in this model fitting is the value of λ (the skew parameter) which is embedded in γ (see the expression after (9) or (14)). Define

$$\widehat{c} = ((2/\pi)(\widehat{\sigma}^S)^2)/(\widehat{\gamma}^S)^2 - 1, \tag{19}$$

where $\widehat{\sigma}^S$ and $\widehat{\gamma}^S$ are given in (18) and (15) above. Since λ carries the sign of γ, the estimate of λ can be found as

$$\widehat{\lambda}^S = sign(\widehat{\gamma}^S)(\widehat{c})^{-1/2}, \text{ provided } \widehat{c} > 0.$$

The value of $\widehat{c} > 0$ stems from the requirement that the RHS of (19) must be > 0. If this condition does not hold, then it implies that if $\left(\widehat{\gamma}^S\right)^2$ is sufficiently large, i.e., γ should be $\pm\infty$, which in turn implies that $\lambda = \pm\infty$. Thus, if $\widehat{c} < 0$, then $\widehat{\lambda}^S$ is estimated as $\widehat{\lambda}^S = sign(\widehat{\gamma}^S)(\infty)$, which essentially says that the $i.i.d.$ distribution of ε_j's is a half-normal distribution (see the property (iii) in Subsection 2.2). However, for practical consideration, ∞ can be replaced by a large constant, say K, since $SND(0, \sigma, \lambda)$ is almost half-normal if $|\lambda| \geq 10$. Hence one can take $K = 10$. Therefore, a comprehensive estimate of $\widehat{\lambda}^S$ can be taken as

$$\widehat{\lambda}^S = \begin{cases} sign(\widehat{\gamma}^S)(\widehat{c})^{-1/2} & \text{if} \quad \widehat{c} > 0 \\ sign(\widehat{\gamma}^S)K & \text{if} \quad \widehat{c} < 0, \end{cases} \tag{20}$$

where $K = 10$.

After estimating all the model parameters we now investigate how this helps us in further inferences. But before that we demonstrate the above estimation of parameters with a real-life dataset as shown in the next subsection.

3.2. Parameter estimates for the arsenic dataset under SND errors model

3.2.1 Model fitting under ND errors

For convenience we adopt the following notation for further consideration with regard to the example discussed in Subsection 1.3 and with summarized data given in Table 2 of Appendix.

As = Arsenic concentration (in $\mu g/L$);

Dep = Depth of the water-well (in meters);

Dis = Distance of the water-well (in meters) from the river;

$Time$ = Time of the data collection (May'14 or August'14).

The scatterplots of (As) against (Dep) and (Dis) for each value of $(Time)$ have been provided in the Appendix (Figure 8-11). Since the scatterplots do not look very linear, we tried several competing nonlinear regression models under the usual normality assumption with homoscedasticity. Our objective is to obtain the "best" possible textbook style regression model under normality first, and then improve over it further using the SND error model.

We have tried the following four possibilities:

(a) (As) regressed on a quadratic expression involving (Dep) and (Dis);

(b) $ln(As)$ regressed on a quadratic expression involving (Dep) and (Dis);

(c) $ln(As)$ regressed on a quadratic expression involving $ln(Dep)$ and $ln(Dis)$;

(d) (As) regressed on a quadratic expression involving $ln(Dep)$ and $ln(Dis)$.

In all the above four cases '$Time$' was used as a covariate, but it was found to be insignificant either as a main factor and/or having any interaction with the other independent variables.

Remark 3 In each of the above four cases, we started with the 'full model' that involved the main explanatory variables (in quadratic form), $Time$ (as a two-level categorical variable, and all the interactions. A stepwise regression approach (using package R see in RStudio Team (2019)) was then applied to arrive at a reasonable 'reduced model' that retains the most plausible significant terms as dictated by the data as well as the model assumptions.

Remark 4 For models (a) and (b), before using the quadratic expression involving (Dep) and (Dis), these two variables were standardized so that the change in units does not affect the explainability of the dependent variable artificially. (In a pure linear regression model this does not matter since the intercept adjusts itself automatically due to any change in units).

The "best" regression model, out of the four possibilities discussed above, turns out to be the reduced model due to (b) (with $R^2 = 0.41$ and $R^2_{adj} = 0.35$), followed by (a) (with $R^2 = 0.36$ and $R^2_{adj} = 0.31$), (c) (with $R^2 = 0.28$ and $R^2_{adj} = 0.24$) and (d) (with $R^2 = 0.23$ and $R^2_{adj} = 0.20$).

The "best" reduced model from (b) finally is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon, \tag{21}$$

where $Y = ln(As)$, $X_1 = standardized(Dep)$, $X_2 = standardized(Dis)$, $X_3 = X_2^2$, $X_4 = X_1^2 X_2$ and $X_5 = X_1^2 X_2^2$.

Thus, in terms of our notation in Section 2, we have $n = 58$ and $(p-1) = 5$. For the model (21), $R^2 = 0.41$ translates to a multiple correlation coefficient of $\sqrt{0.41} \approx 0.64$, which is reasonably good.

The fitted response surface of the resultant regression model under normality assumption, i.e., $\varepsilon \sim N(0, \sigma^2)$ (from (21)) is then

$$\widehat{\eta}^N = \widehat{E}(Y) = \widehat{\beta}_0^N + \widehat{\beta}_1^N X_1 + \widehat{\beta}_2^N X_2 + \widehat{\beta}_3^N X_3 + \widehat{\beta}_4^N X_4 + \widehat{\beta}_5^N X_5, \tag{22}$$

where the estimated coefficients (under the normality assumption) are given in the following unified Table 1 (see under "Normal model"). Since the right hand side of (22) is a function of (Dep) and (Dis) only, we can draw the 3D-plot of (22) in the following Figure 1.

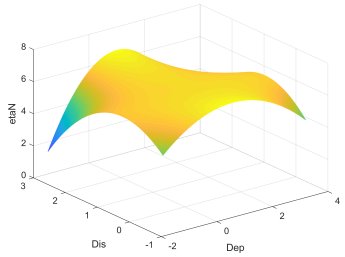

Figure 1 The fitted 3D-plot of the normal response surface

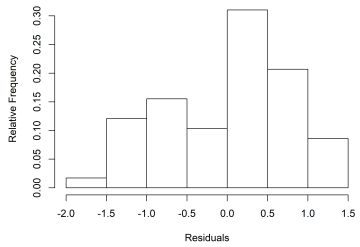

Figure 2 Histogram of residuals of model (21) with normal errors

Further, the Figure 2 shows the histogram of the residuals of the model (21) with normal errors which looks somewhat skewed. This justifies our next step of improving the model (21) by assuming $\varepsilon \sim SND(0, \sigma, \lambda)$.

3.2.2 Model fitting under SND errors

We now follow the steps presented earlier in Subsection 3.1 to implement the regression model (21) where the error ε is assumed to follow $SND(0, \sigma, \lambda)$ for the arsenic data given in Table 2.

Following the expressions in (11), (15), (17), (18) and (20), all the relevant parameters have been estimated, and the following Table 1 summarizes the estimation results for an easy comparison between the normality assumption as well as the SND assumption. To distinguish between these two models we write $\widehat{\lambda}^N$ and $\widehat{\lambda}^S$ for convenience ($\widehat{\lambda}^N$ is always zero).

Remark 5 From Table 1 it is obvious that the more flexible SND model for the regression error results into a different estimate of the intercept β_0 while the other coefficients (slopes') estimates remain the same. Thus, the more flexible SND model corrects the intercept estimate by extracting the estimate of λ, possibly nonzero, which otherwise would have been forced to take the value 0 under the usual normality assumption.

Table 1 Estimated parameters under two models

Parameters in (21)	Normal model	SND model
β_0	$\widehat{\beta}_0^N = 6.550$	$\widehat{\beta}_0^S = 6.779$
β_1	$\widehat{\beta}_1^N = -0.107$	$\widehat{\beta}_1^S = -0.107$
β_2	$\widehat{\beta}_2^N = 0.176$	$\widehat{\beta}_2^S = 0.176$
β_3	$\widehat{\beta}_3^N = -0.085$	$\widehat{\beta}_3^S = -0.085$
β_4	$\widehat{\beta}_4^N = 0.229$	$\widehat{\beta}_4^S = 0.229$
β_5	$\widehat{\beta}_5^N = -0.159$	$\widehat{\beta}_5^S = -0.159$
σ	$\widehat{\sigma}^N = 0.293$	$\widehat{\sigma}^S = 0.360$
λ	$\widehat{\lambda}^N = 0$	$\widehat{\lambda}^S = -1.320$

4. Sampling Distributions of the Parameter Estimators under SND Errors

Under the usual normal model the sampling distributions of the estimators are well known as $\widehat{\boldsymbol{\beta}}^N \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbb{X}'\mathbb{X})^{-1})$ and $\widehat{\sigma}^N = \{\|\boldsymbol{e}^N\|^2/(n-p)\}^{1/2}$ where $\widehat{\sigma}^N$ is the estimate of σ under the normal model and $\boldsymbol{e}^N = \boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^N$. Further, it is known that $\widehat{\sigma}^N$ and $\widehat{\boldsymbol{\beta}}^N$ are mutually independent.

Under SND error model, the exact sampling distributions of $\widehat{\boldsymbol{\beta}}^S, \widehat{\lambda}^S$ and $\widehat{\sigma}^S$ are impossible to get. However, an approximate sampling distribution of $\widehat{\boldsymbol{\theta}}^S = (\widehat{\boldsymbol{\beta}}^S, \widehat{\sigma}^S, \widehat{\lambda}^S)$ can be found by the bootstrap method, and this works reasonably well for $(n-p)$ not "too small". In the following we provide the estimated bias and dispersion matrix of the parameter estimates through bootstrap (using package R see in RStudio Team (2019)).

4.1. Sampling properties of the parameter estimators under SND errors

The bootstrap method is implemented through the following steps. Define $\boldsymbol{e}^S = \boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^S = (e_1^S, \ldots, e_n^S)'$, as the observed residual vector under the SND model.

Step 1: Consider a finite population of residuals as $\mathcal{P} = \{e_1^S, e_2^S, \ldots, e_n^S\}$. Draw a random bootstrap sample of size n with replacement from $\mathcal{P}$ with the observations, say, $e_1^{S^*}, e_2^{S^*}, \ldots, e_n^{S^*}$. (Note that certain residual(s) may be repeated in this bootstrap sample.)

Step 2: Recreate a bootstrap version of the response variable vector $\boldsymbol{Y}$ as $\boldsymbol{Y}^* = (Y_1^*, \ldots, Y_n^*)$, where

$$Y_j^* = \widehat{\beta}_0^S + \widehat{\beta}_1^S X_{1j} + \cdots + \widehat{\beta}_{(p-1)}^S X_{(p-1)j} + e_j^{S^*},\ 1 \le j \le n.$$

Step 3: Now fit a regression model as before as in (6) - (7) with the bootstrap values of Y as

$$Y_j^* = \beta_0 + \beta_1 X_{1j} + \cdots + \beta_{p-1} X_{(p-1)j} + \varepsilon_j,\ 1 \le j \le n,$$

where ε_j's are $i.i.d.\ SND(0, \sigma, \lambda)$. Using the data $(Y_j^*, X_{1j}, \ldots, X_{(p-1)j}), 1 \le j \le n$, estimate all parameters as described in Subsection 3.1. In the other words, get a new (bootstrap) estimate of $\boldsymbol{\theta}$ under SND as $\widehat{\boldsymbol{\theta}}^{S^*} = (\widehat{\boldsymbol{\beta}}^{S^*}, \widehat{\sigma}^{S^*}, \widehat{\lambda}^{S^*})$ based on $(Y_j^*, X_{1j}, \ldots, X_{(p-1)j})$, $1 \le j \le n$.

Step 4: Repeat the above steps 1 though 3 a large number (say, Q) times. As a result, we produce Q bootstrap copies of $\widehat{\boldsymbol{\theta}}^S$ as $\widehat{\boldsymbol{\theta}}^{S^*(q)} = (\widehat{\boldsymbol{\beta}}^{S^*(q)}, \widehat{\sigma}^{S^*(q)}, \widehat{\lambda}^{S^*(q)}),\ 1 \le q \le Q$, which are used to approximate the sampling distribution of individual components of the estimated parameter vector, $\widehat{\boldsymbol{\theta}}^S$.

Remark 6 Two sampling distributional parameters of $\widehat{\boldsymbol{\theta}}^S$ are of interest, namely, - the bias vector, and the dispersion matrix of $\widehat{\boldsymbol{\theta}}^S$. To estimate these two secondary parameters, define $\overline{\widehat{\boldsymbol{\theta}}}^{S^*(\cdot)} = \sum_{q=1}^{Q} \widehat{\boldsymbol{\theta}}^{S^*(q)}/Q$. Then the bootstrap estimated bias and dispersion matrix of $\widehat{\boldsymbol{\theta}}^S$ are given as
$\widehat{B}^*(\widehat{\boldsymbol{\theta}}^S) =$ estimated bias of $\widehat{\boldsymbol{\theta}}^S = \sum_{i=1}^{Q}(\widehat{\boldsymbol{\theta}}^{S^*(q)} - \widehat{\boldsymbol{\theta}}^S)/Q$, and
$\widehat{D}^*(\widehat{\boldsymbol{\theta}}^S) =$ estimated dispersion matrix of $\widehat{\boldsymbol{\theta}}^S = \sum_{i=1}^{Q}(\widehat{\boldsymbol{\theta}}^{S^*(q)} - \overline{\widehat{\boldsymbol{\theta}}}^{S^*(\cdot)})(\widehat{\boldsymbol{\theta}}^{S^*(q)} - \overline{\widehat{\boldsymbol{\theta}}}^{S^*(\cdot)})'/Q$.

In the following we implement the above bootstrap approach for the specific dataset discussed in Subsection 3.2. Note that, for the given dataset, $(n-p) = 58-6 = 52$, which is moderately large, and hence the bootstrap method is expected to provide a reliable approximate sampling distribution of $\widehat{\boldsymbol{\theta}}^S$. The total number of parameters here is 8. Using $Q = 10^5$, we provided the bootstrap relative frequency histogram of all the eight parameters in the following Figure 3. Note that, except $\widehat{\beta}_0^S$ and $\widehat{\lambda}^S$, the sampling distributions of all other parameter estimates look fairly bell-shaped.

The diagonal elements of the $\widehat{D}^*$ matrix indicate the estimated variance (i.e., the squared standard error (SE)) of the individual components of the parameter estimate vector $\widehat{\boldsymbol{\theta}}^S$; and the off-diagonal elements of $\widehat{D}^*$ are the estimated covariances between two components of $\widehat{\boldsymbol{\theta}}^S$.

For our given arsenic data, the estimated bias and dispersion of $\widehat{\boldsymbol{\theta}}^S$ are obtained as

$$\widehat{\boldsymbol{B}}^* = \left[\ [-730.019 \quad 1.537 \quad -3.122 \quad 1.418 \quad 0.493 \quad 5.715 \quad -260.144 \quad 2472.101] \times 10^{-4}\ \right]'$$

$$\widehat{D}^* = \begin{bmatrix} 284.210 & 3.137 & 25.157 & -15.625 & -17.381 & -40.875 & 31.165 & -1509.986 \\ 3.137 & 40.281 & 7.119 & -2.032 & -40.865 & -7.217 & 0.041 & 0.414 \\ 25.157 & 7.119 & 49.337 & -20.551 & -28.218 & -21.337 & -0.181 & 8.432 \\ -15.625 & -2.032 & -20.551 & 13.229 & 10.086 & 7.201 & -0.112 & 4.857 \\ -17.381 & -40.865 & -28.218 & 10.086 & 69.803 & 26.597 & -0.164 & 6.320 \\ -40.875 & -7.217 & -21.337 & 7.201 & 26.597 & 77.606 & -1.673 & 73.793 \\ 31.165 & 0.041 & -0.181 & -0.112 & -0.164 & -0.673 & 14.443 & -206.225 \\ -1509.986 & 0.414 & 8.432 & 4.857 & 6.320 & 73.793 & -206.225 & 10355.230 \end{bmatrix} \times 10^{-4}$$

The bootstrap method can help us obtain approximate confidence intervals (CIs) of the parameters under the SND model. For example, take the first component of $\widehat{\boldsymbol{\theta}}^S$ only. The sampling distribution of $\widehat{\beta}_0^S$ doesn't look normal, hence we can apply the ordered statistics, i.e., we obtain $\widehat{\beta}_0^{S^*(q)}$, $q = 1, 2, \ldots, Q$; order these from the smallest to largest as

$$\widehat{\beta}_0^{S^*[1]} \leq \widehat{\beta}_0^{S^*[2]} \leq \cdots \leq \widehat{\beta}_0^{S^*[Q]}.$$

Given $(1-\alpha) = 0.95 =$ the confidence level, find

$$\widehat{\beta}_0^L = \widehat{\beta}_0^{S^*[Q\alpha/2]} = (\alpha/2)100^{th} \text{ percentile of } \widehat{\beta}_0^{S^*(q)}\text{'s},$$
$$\widehat{\beta}_0^U = \widehat{\beta}_0^{S^*[Q(1-\alpha/2)]} = (1-\alpha/2)100^{th} \text{ percentile of } \widehat{\beta}_0^{S^*(q)}\text{'s}.$$

The approximate $(1-\alpha) = 0.95$ level CI of β_0 is $(\widehat{\beta}_0{}^L, \widehat{\beta}_0{}^U) = (6.278, 6.925)$.

Similarly, do this for all components of $(\widehat{\boldsymbol{\theta}}^{S^*(q)})$, i.e., get the component - wise CIs as $(\widehat{\beta}_1{}^L, \widehat{\beta}_1{}^U)$ $= (-0.231, 0.0179)$, $(\widehat{\beta}_2{}^L, \widehat{\beta}_2{}^U) = (0.0378, 0.314)$, $(\widehat{\beta}_3{}^L, \widehat{\beta}_3{}^U) = (-0.157, -0.0137)$, $(\widehat{\beta}_4{}^L, \widehat{\beta}_4{}^U) =$ $(0.065, 0.392)$, $(\widehat{\beta}_5{}^L, \widehat{\beta}_5{}^U) = (-0.334, 0.012)$, $(\widehat{\sigma}^L, \widehat{\sigma}^U) = (0.261, 0.407)$, and $(\widehat{\lambda}^L, \widehat{\lambda}^U) = (-2.711, 1.159)$.

Remark 7 The above bootstrap approach of finding an approximate $(1-\alpha)$ - level CI for a parameter is quite versatile, and does not require the knowledge of the exact sampling distribution of the parameter estimate. Further, these CIs can be used to test a suitable null hypothesis against a suitable alternative hypothesis for an individual parameter. If the null hypothetical value of a parameter falls within the specified CI with confidence level $(1-\alpha)$ then the null hypothesis gets retained at level α. Otherwise, the null hypothesis gets rejected at level α.

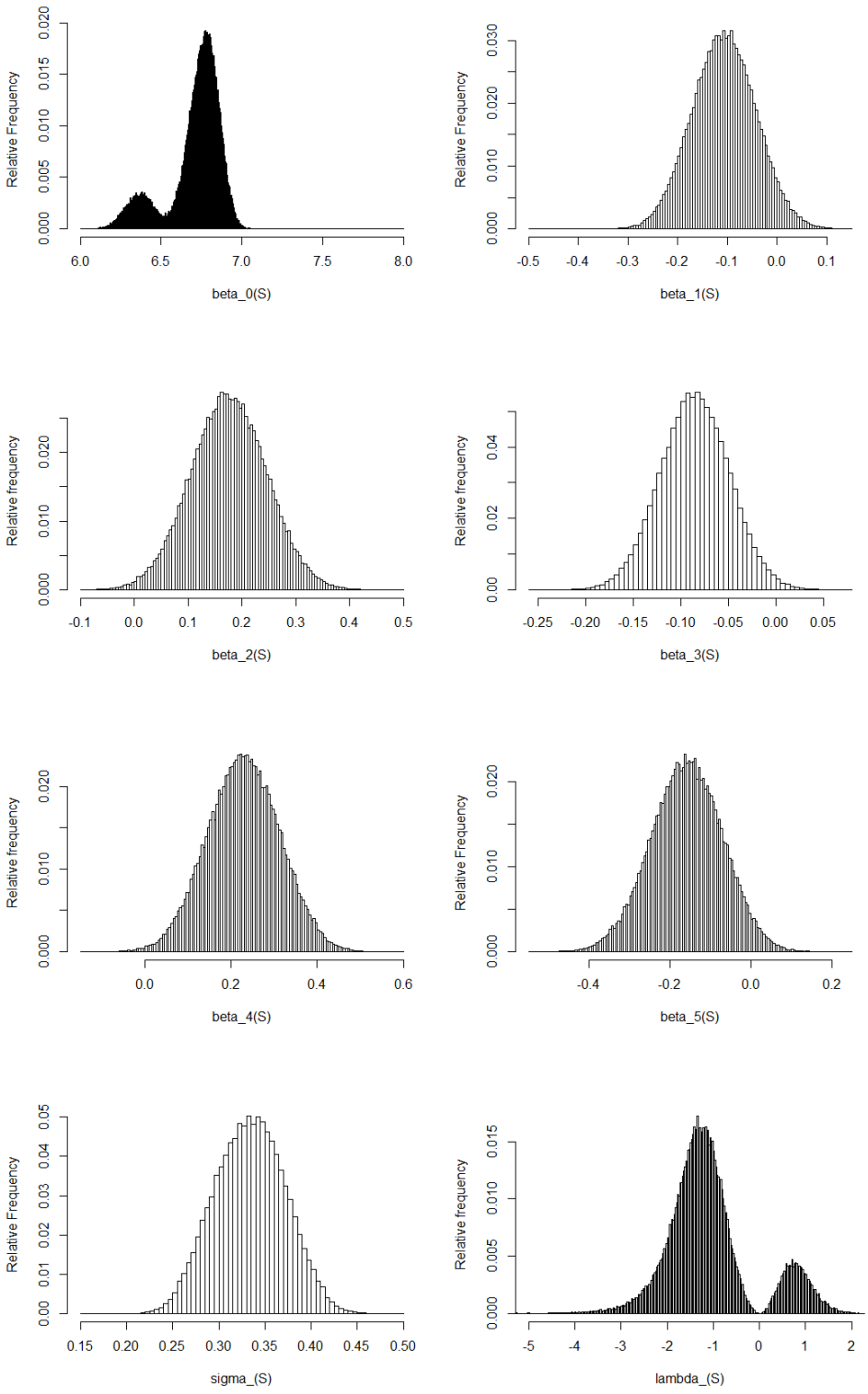

Figure 3 The bootstrap relative frequency histogram of all the eight parameters under SND errors

4.2. Model goodness of fit through AIC

A popular measure of goodness of fit (GoF) of a model to a given dataset is Akaike Information Criterion, or AIC. For a dataset of size n, with the model parameter vector $\boldsymbol{\theta}$, AIC is defined as

$$AIC = -2ln(\widehat{\boldsymbol{\theta}}) + 2k$$

where $\widehat{\boldsymbol{\theta}}$ is the estimated value of $\boldsymbol{\theta}$, and k is the total number of parameters in the model. One can also look at AIC$^* = AIC/n = AIC$ value per observation. (Though AIC* is equivalent to AIC, it gives a better idea than AIC.)

In our study we will compare AIC (or AIC*) of the ND errors model with that under the SND errors model. A smaller value of AIC (or AIC*) is desired for a better fit.

For the ND errors model we can study the AIC (or AIC*) theoretically in great details, but it is not possible for the SND errors model due to its complicated structure. Hence, for the normal model, we will study AIC* theoretically as much as possible. However, for the SND errors model we will study it through bootstrap simulation.

(i) ND errors model: $\underset{n\times 1}{\boldsymbol{Y}} = \mathbb{X}\boldsymbol{\beta} + \underset{n\times 1}{\boldsymbol{\varepsilon}} \sim N_n(\mathbb{X}\boldsymbol{\beta}, \sigma^2 I_n)$. So,

$$L(\boldsymbol{\beta}, \sigma^2) = ((2\pi)^{1/2}\sigma)^{-n} \exp\left\{-\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|^2/(2\sigma^2)\right\}.$$

Note that $\widehat{\boldsymbol{\beta}}^N = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{Y}$ and $\left(\widehat{\sigma}^N\right)^2 = \|\boldsymbol{e}^N\|^2/(n-p)$. Then,

$$L(\widehat{\boldsymbol{\beta}}^N, \left(\widehat{\sigma}^N\right)^2) = ((2\pi)^{1/2}\widehat{\sigma}^N)^{-n} \exp\left\{-\left\|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^N\right\|^2/(2\left(\widehat{\sigma}^N\right)^2)\right\}.$$

Also, $\boldsymbol{e}^N$ = residual vector = $\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^N$. So,

$$\begin{aligned} -2lnL(\widehat{\boldsymbol{\beta}}^N, \left(\widehat{\sigma}^N\right)^2) &= nln(2\pi) + nln(\left(\widehat{\sigma}^N\right)^2) + \left\|\boldsymbol{e}^N\right\|^2/\left(\widehat{\sigma}^N\right)^2 \\ &= nln(2\pi) + nln\left(\left\|\boldsymbol{e}^N\right\|^2/(n-p)\right) + (n-p), \end{aligned}$$

since $\left(\widehat{\sigma}^N\right)^2 = \|\boldsymbol{e}^N\|^2/(n-p)$.

Also, the distribution of $\|\boldsymbol{e}^N\|^2 = \sigma^2\chi^2_{(n-p)}$. So, as far as the distribution is concerned,

$$AIC|_{ND} = nln(2\pi) + n\left[ln\sigma^2 + ln(V/(n-p))\right] + (n-p) + 2(p+1),$$

(since under normal, $k = (p+1)$, where $V \sim \chi^2_{(n-p)}$).

$$\begin{aligned} \Rightarrow AIC^*|_{ND} &= (1/n)AIC|_{ND} \\ &= \left[ln(2\pi) + ln\sigma^2 + (1 + (p+2)/n) + ln(V/(n-p))\right]. \end{aligned}$$

Define $h(\sigma^2|n,p) = ln(2\pi) + ln\sigma^2 + (1 + (p+2)/n) =$ constant. Notice that V can be treated as $V = \sum_{i=1}^{(n-p)} V_i$, where each $V_i \sim \chi^2_1$, i.e., each V_i has mean 1 and variance 2. So,

$V/(n-p) = (1/(n-p)) \sum_{i=1}^{(n-p)} V_i \sim N\left(1, 2/(n-p)\right)$ approximately by the CLT, for large $(n-p)$.

Therefore by the Delta Method (Type I), for large $(n-p)$,

$$ln(V/(n-p)) \xrightarrow{d} N\left(0, 2/(n-p)\right),$$

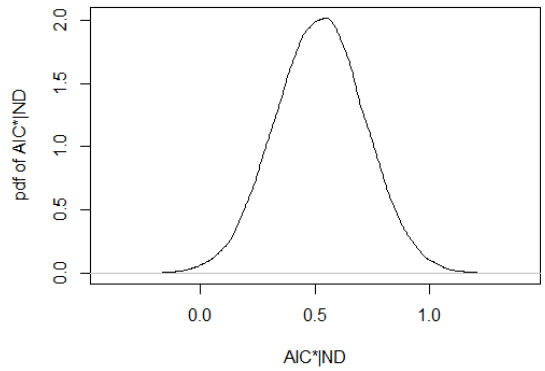

Figure 4 Approximate pdf of $\text{AIC}^*|_{ND}$

and this convergence (which is $O_p(1)$) is first order efficient in the sense that true variance of $ln(V/(n-p))$ is of order $O(1/\sqrt{n-p})$, and it matches that of $N(0,\ 2/(n-p))$.

So, $\text{AIC}^*|_{ND} \underset{approx}{\sim} N\left(h(\sigma^2|n,p), 2/(n-p)\right)$. We can plot this pdf by using $\sigma \approx \widehat{\sigma}^N = 0.293$, which is shown in Figure 4.

(ii) SND errors model: Note that, here $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma, \lambda)$ and $k = (p+2)$. So,

$$\text{AIC}|_{SND}(\widehat{\boldsymbol{\theta}}^S) = -2lnL(\widehat{\boldsymbol{\beta}}^S, \widehat{\sigma}^S, \widehat{\lambda}^S) + 2(p+2),$$

where $L(\boldsymbol{\beta}, \sigma, \lambda) = \prod_{j=1}^{n} \{(2/\sigma)\,\phi((Y_j - \mu_j)/\sigma)\Phi(\lambda(Y_j - \mu_j)/\sigma)\}$, and μ_j is given in (7).

Get $\text{AIC}^*|_{SND}(\widehat{\boldsymbol{\theta}}^S) = (1/n)\text{AIC}|_{SND}(\widehat{\boldsymbol{\theta}}^S)$ by replacing $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma, \lambda)$ by $\widehat{\boldsymbol{\theta}}^S = (\widehat{\boldsymbol{\beta}}^S, \widehat{\sigma}^S, \widehat{\lambda}^S)$. But this is just a single value for the given dataset. To study the variation of $\text{AIC}^*|_{SND}$, we can run a simulation based on the bootstrap method described earlier. [See Step - 4 of the bootstrap we discussed earlier where we generated $\widehat{\boldsymbol{\theta}}^{S^*(q)},\ 1 \le q \le Q$.] So, compute $\left[\text{AIC}^*|_{SND}(\widehat{\boldsymbol{\theta}}^{S^*(q)})\right]$, $1 \le q \le Q$; and draw the histogram of $\text{AIC}^*|_{SND}(\widehat{\boldsymbol{\theta}}^{S^*(q)})$ values, which is presented in Figure 5, and see how to it compares with that of $N(h, 2/(n-p))$ (under ND).

Remark 8 Note that $\text{AIC}^*|_{SND}$ is mostly ≤ -2.0, whereas $\text{AIC}^*|_{ND}$ is mostly ≥ 0.0. Hence, as a smaller AIC^* indicates a better GoF, our fitting of the data using a regression model based on SND errors is far better than that based on ND errors.

5. Back to Arsenic Problem

Now that we have established the superiority of the SND errors over the normal ones in fitting our regression model, it is worth going back to the actual problem of explaining the arsenic (As) level in terms of depth (Dep) and distance (Dis).

From (21) we have

$$\begin{aligned} ln(As) \approx\ & \widehat{\beta}_0^S + \widehat{\beta}_1^S \{standardized(Dep)\} + \widehat{\beta}_2^S \{standardized(Dis)\} \\ & + \widehat{\beta}_3^S \{standardized(Dis)\}^2 + \widehat{\beta}_4^S \{standardized(Dep)\}^2 \{standardized(Dis)\} \\ & + \widehat{\beta}_5^S \{standardized(Dep)\}^2 \{standardized(Dis)\}^2, \end{aligned}$$

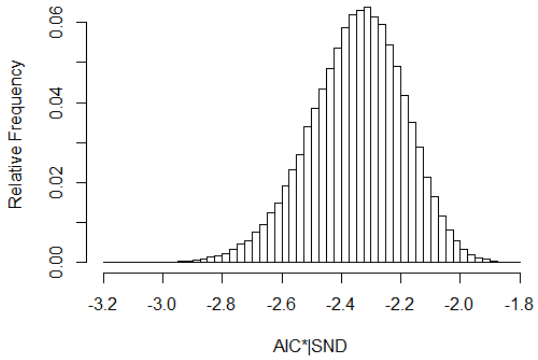

Figure 5 Bootstrap relative frequency histogram of $\text{AIC}^*|_{SND}$

where $standardized(Dep) = (Dep - \overline{Dep})/sd(Dep)$, and
$standardized(Dis) = (Dis - \overline{Dis})/sd(Dis)$, i.e.,

$$\begin{aligned}(As) \approx exp\{&\widehat{\beta}_0^S + \widehat{\beta}_1^S \{standardized(Dep)\} + \widehat{\beta}_2^S \{standardized(Dis)\} \\ &+ \widehat{\beta}_3^S \{standardized(Dis)\}^2 + \widehat{\beta}_4^S \{standardized(Dep)\}^2 \{standardized(Dis)\} \\ &+ \widehat{\beta}_5^S \{standardized(Dep)\}^2 \{standardized(Dis)\}^2\}. \end{aligned} \quad (23)$$

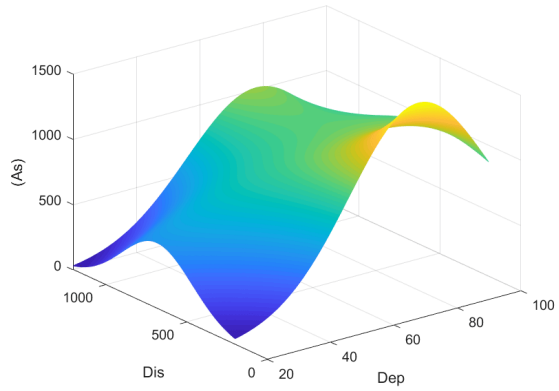

Figure 6 The fitted 3D-plot of the equation in (23)

Since (As) is a highly nonlinear function of both (Dep) and (Dis), we are going to show its behavior by plotting (i) (As) against (Dep), assuming $(Dis) = (\overline{Dis}) = average(Dis)$, i.e., $standardized(Dis) = 0$; and (ii) (As) against (Dis), assuming $(Dep) = (\overline{Dep}) = average(Dep)$, i.e., $standardized(Dep) = 0$.

For the above case - (i),

$$(As) \approx exp\left[\widehat{\beta}_0^S + \widehat{\beta}_1^S \left\{(Dep - \overline{Dep})/sd(Dep)\right\}\right]; \tag{24}$$

and for the above case - (ii),

$$(As) \approx exp\left[\widehat{\beta}_0^S + \widehat{\beta}_2^S \left\{(Dis - \overline{Dis})/sd(Dis)\right\} + \widehat{\beta}_3^S \left\{(Dis - \overline{Dis})/sd(Dis)\right\}^2\right]. \tag{25}$$

The following two figures show the plots of (24) and (25) as demonstration purposes only when (As) is plotted against each of (Dep) and (Dis) holding the other explanatory variable at its average value.

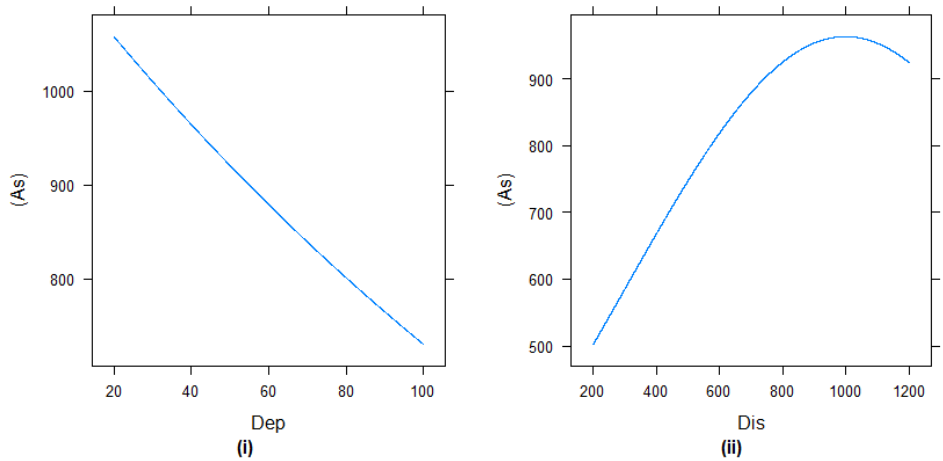

Figure 7 Plot of As against (i) Dep when $(Dis) = \overline{Dis}$; and (ii) Dis when $(Dep) = \overline{Dep}$.

Remark 9 The exact profile of arsenic prevalence in Khanh An commune of An Phu district is understood through the 3D-plot of the equation in (23) which is given in Figure 6. The landowners and/or farmers can make use of this 3D - plot to see where arsenic is high and where it is low. The 2D - plots in Figure 7 show some rough trends in arsenic w.r.t. depth (at an average distance from the river), and w.r.t. distance (at an average depth). At an average distance, arsenic is decreasing as a function of depth, which means that arsenic is more prevalent in shallow water-wells (may be due to pollution and/or presence of arsenical compounds in the upper large of soil). On the other hand, at an average depth, arsenic is increasing with distance from the river up to a certain value (approximately 1000 meters), and then it starts decreasing. Ecologists, geologists and/or environmental experts can investigate further why roughly 1000 meters from the river arsenic is highest at an average depth.

6. Concluding Remark

In this study, we have considered a regression model to fit the arsenic dataset with SND errors (which generalizes the classical regression model with ND errors) in a frequentist set-up. We have shown the details of estimating all the model parameters, and how to study their sampling properties through the bootstrap method. Our estimation method, which is a combination of the ordinary least squares method and the method of moments estimation, is quite new and easy to implement. The sampling distribution of each parameter can be used to draw further inferences on individual parameters (see Remark 7). In addition, we have presented the comparison between the SND regression and

ND regression in terms of AIC* (or AIC) for the given dataset which clearly shows the superiority of the SND regression.

With respect to the actual arsenic prevalence problem in the MDR we have analyzed the final regression model, and the emerging trends it shows for the sampled area in terms of a water-well's depth and distance. Hopefully, this new idea of SND regression can be helpful for the applied researchers in many other similar studies. Our objectives in future research will include studying the prediction of a future observation of the dependent variable based on the known value of the independent variables, and the corresponding predictive error under for the SND regression and compare it with that under the ND regression. Further, we plan to study Bayesian aspects of the SND regression model along with a sensitivity analysis with respect to suitable prior distributions.

Acknowledgements

We would like to thank the two anonymous referees who went over the first draft of this paper very meticulously, and made critical as well as constructive comments which helped us tremendously in improving the presentation of this work.

We would also like to thank Assoc. Prof. Phu Le Vo, Faculty of Environment and Natural Resources, Ho Chi Minh City University of Technology VNU-HCM, Vietnam and Prof. Rizlan Bernier-Latmani, Environmental Microbiology Laboratory (EML) at EPFL, Switzerland, for allowing us to use the arsenic dataset.

References

Ahuja S. Arsenic contamination of groundwater. New Jersey: John Wiley& Sons; 2008.

Alhamide AA, Ibrahim K, Alodat MT. Inference for multiple linear regression estimators with extended skew normal errors. Pak J Stat. 2016; 32(2): 81-96.

Altug T. Introduction to toxicology and food. New York: CRC Press; 2003.

Arelano-Valle RB, Ozan S, Bolfarine H, Lachos VH. Skew normal measurement error models. J Multivar Anal. 2005; 96(2): 265-281.

Arnold BC, Lin GD. Characterizations of the skew-normal and generalized chi distributions. Sankhya (2003-2007). 2004; 66(4): 593-606.

Azzalini A. A class of distributions which includes the normal ones. Scand J Stat. 1985; 12(2): 171-178.

Azzalini A. Further results on a class of distributions which includes the normal ones. Statistica. 1986; 46(2): 199-208.

Cancho VC, Lachos VH, Ortega EMM. A nonlinear regression model with skew-normal errors. Stat pap. 2010; 51(3): 547-558.

Henderson DJ, Parmeter CF. Applied nonparametric econometrics. New York: Cambridge University Press; 2015.

Ichimura, H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. J Econometrics. 1993; 58(1-2): 71-120.

Guedes TA, Rossi RM, Martins ABT, Janeiro V, Carneiro JWP. Applying regression models with skew-normal errors to the height of bedding plants of Stevia rebaudiana (Bert) Bertoni. Acta Sci Technol. 2014; 36(3): 463-468.

Gupta AK, Nguyen TT, Sanqui JT. Characterization of the skew-normal distribution. 2003; Ann Inst Stat Math. 56(2): 351-360.

Lachos VH, Dey DK, Cancho VG. Robust linear mixed models with skew-normal independent distributions from a Bayesian perspective. J Stat Plan Inference. 2009; 139(2): 4098-4110.

Lièvremont D, Bertin P, Lett MC. Arsenic in contaminated waters: Biogeochemical cycle, microbial metabolism and biotreatment processes. Biochimie. 2009; 91(10): 1229-1237.

Nakamura G. Defoliation during the Vietnam war. In: Extreme conflict and tropical forests. Defoliation during the Vietnam War. Springer. 2007. p. 149-158.

Nguyen PK. Geochemical study of arsenic behavior in aquifer of the Mekong Delta, Vietnam. PhD [dissertation]. Kyushu University; 2008.

Pérez-Rodríguez P, Acosta-Pech R, Pérez-Elizalde S, Cruz CV, Espinosa JS, Crossa J. A Bayesian genomic regression model with skew normal random errors. G3: Genes Genom Genet. 2018; 8(5): 1771-1785.

Pham CHV. Studying the mechanisms of arsenic release in groundwater in An Phu district, An Giang province. Master [dissertation]. University of Technology, Ho Chi Minh City, Vietnam; 2015.

RStudio Team. RStudio: Integrated Development for R. RStudio, Inc., Boston, MA [serial on the Internet]. 2019 [cited 2019 Apr 8]. Available from: http://www.rstudio.com/.

Sahu SK, Dey DK, Branco MD. A new class of multivariate skew distributions with applications to Bayesian regression models. The Can J Stat. 2003; 31(2):129-150.

Thiuthad P, Pal N. Point estimation of the location parameter of a skew-normal distribution: Some fixed sample and asymptotic results. J Stat Theory Pract. 2019; 13(2): 13-37.

Tran AT, Tinh TK, Vo QM. A study examining arsenic concentrations in groundwater, An Phu district, An Giang province. J Sci . 2011; 17a: 118-123.

Young AL, Regigani GM. Military use of herbicides in Vietnam: Massive quantities of herbicides were applied by the United States in a Tactical Operation Designed to reduce ambushes and disrupt enemy tactics. In: Agent Orange and its associated dioxin: assessment of a conroversy. Amsterdam: Elsevier; 1988. 10-33.

Vo LP, Bernier R, Pham CHV, Ho TNH, Nguyen TBT. Threat of arsenic occurrence in the Vietnamese Mekong Delta. J Geographical Res. 2015; 63:129-142.

Vo LP, Pham CHV, Nguyen VMM, Pham KBA, Vu VA, Nguyen TBT. Arsenic pollution in shallow groundwater in a floodplain delta: A case study in An Phu, An Giang, in Mekong Delta, Vietnam. J Sci Technol. 2016; 54.

Appendix

Table 2 Dataset from the 29 sampled locations (Here "*Std*" stands for "standardized" values.)

Location (Well)	*Std(Dep)* (X_1)	*Std(Dis)* (X_2)	*Std(ln(As)*	
			May 2014	August 2014
1	-0.803	-0.467	-0.816	0.466
2	-1.250	-0.357	-0.373	-0.379
3	0.093	-0.495	-0.658	0.226
4	-1.250	0.418	2.071	-0.130
5	0.316	0.193	-0.497	-0.344
6	-0.355	0.172	1.352	1.004
7	0.093	-0.537	1.097	1.389
8	0.988	0.755	-0.633	-0.620
9	0.316	-0.362	0.078	-1.185
10	0.093	0.318	1.431	1.358
11	3.002	0.186	1.167	1.187
12	1.659	0.715	0.951	0.828
13	-0.803	-0.941	0.431	0.249
14	-0.803	-0.493	0.317	1.144
15	0.540	-0.837	-0.485	-0.607
16	0.316	-0.976	-1.049	-1.630
17	-0.131	-1.093	-0.277	-0.327
18	-0.131	-1.029	-0.265	0.212
19	0.093	-0.924	-0.639	-0.539
20	0.093	0.684	0.615	0.916
21	-0.355	2.150	-0.216	0.890
22	0.093	2.565	0.414	0.930
23	0.316	2.517	-1.454	0.093
24	1.212	0.688	0.686	1.101
25	0.093	-0.423	0.440	0.457
26	0.764	-0.528	-1.178	-1.079
27	-1.698	-0.682	-2.565	-2.949
28	-0.355	-0.588	-0.558	-0.260
29	-2.145	-0.627	-0.626	-1.162

The standardized values (instead of the actual data points) are provided to maintain confidentiality of the original dataset. It doesn't affect the regression model fitting, and the subsequent inferences.

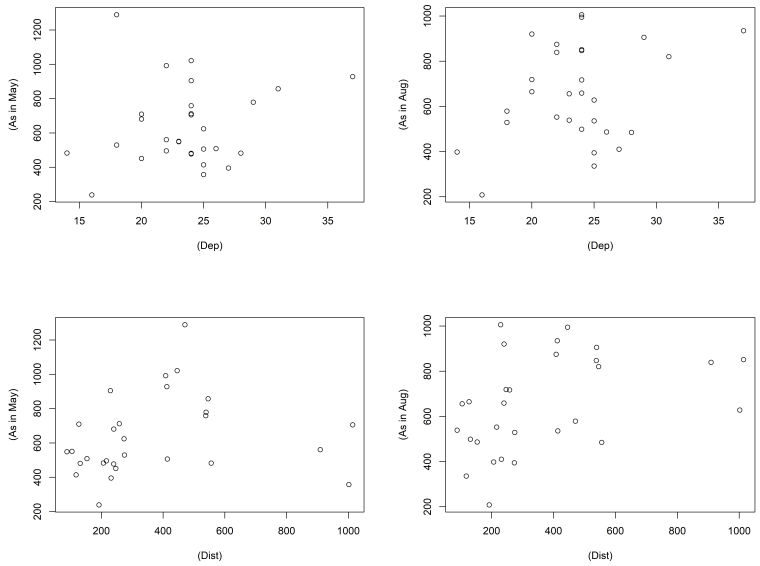

Figure 8 The scatterplots of case (a)

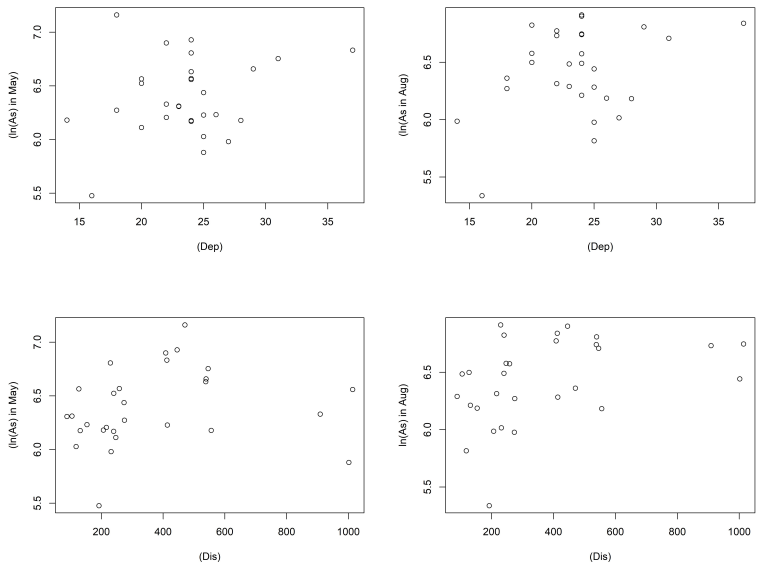

Figure 9 The scatterplots of case (b)

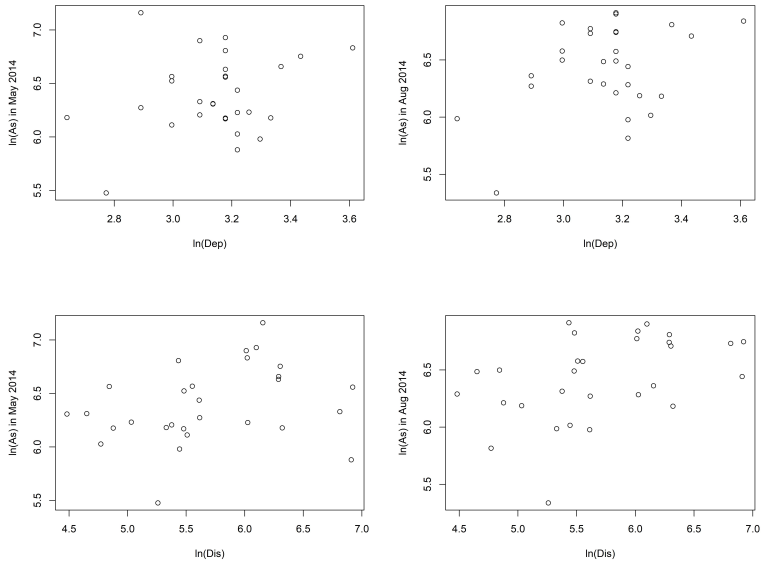

Figure 10 The scatterplots of case (c)

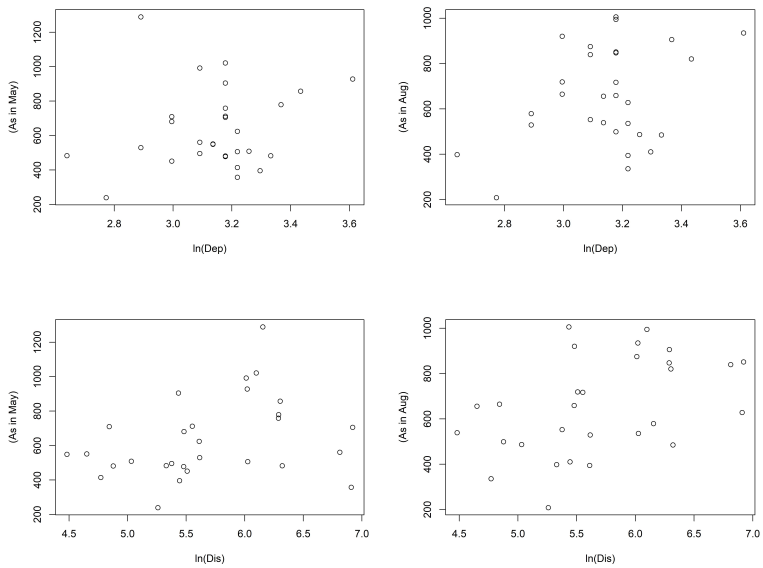

Figure 11 The scatterplots of case (d)