Regression Under Skew-Normal Error Model, and Predicting Arsenic from Geographic Characteristics in the Mekong Delta Region

Nabendu Pal
University of Louisiana at Lafayette, USA
(A joint work with
Dr. Uyen T. Huynh, University of Economics and Law, Vietnam National University, Ho Chi Minh City, Vietnam and
Dr. Man Nguyen, Mahidol University, Thailand)

June 16, 2022

OUTLINES

Introduction 1

Objectives 2

Methodology 3

Results 4

Scope of Further Research 5

Conclusion 6

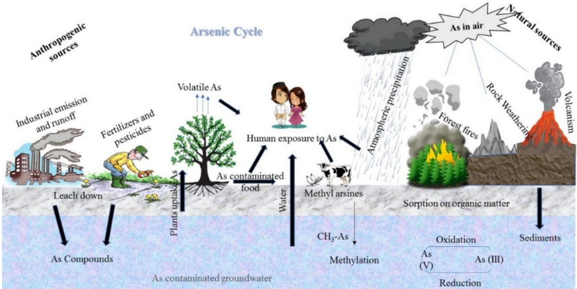
Introduction

In recent decades, environmental pollution has become a major global concern.

The South and Southeast Asian countries are witnessing an alarming rise of arsenic pollution in groundwater where a large section of the rural local population depends on agriculture, and they draw groundwater for:

- Direct irrigation
- Raising cattle
- Daily personal consumption

Arsenic (*As*), which enters the food-chain from various sources, is causing major health issues.



The current situation of arsenic pollution in Vietnam

+ Red River Delta (RRD).
+ Mekong Delta Region (MDR).

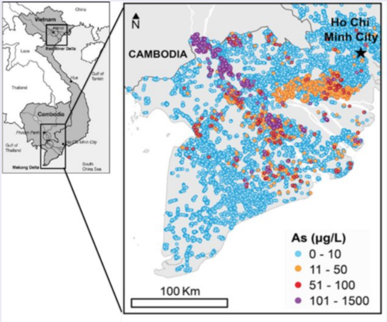
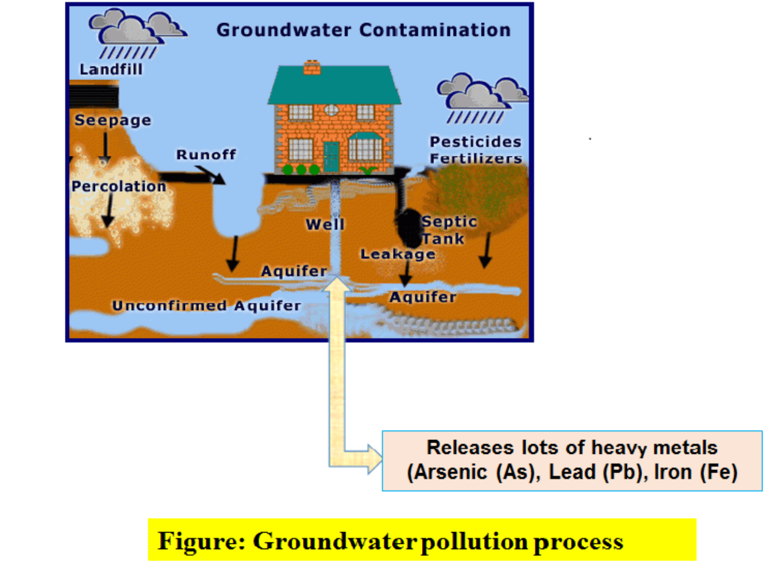


Figure 1: The arsenic concentration in MDR.

- Within the MDR, An Giang is one of the worst affected provinces that have been witnessing a very high *As* pollution in groundwater.
- Why the arsenic pollution level in Vietnam is high? It is due to several reasons:



Releases lots of heavy metals (Arsenic (As), Lead (Pb), Iron (Fe)

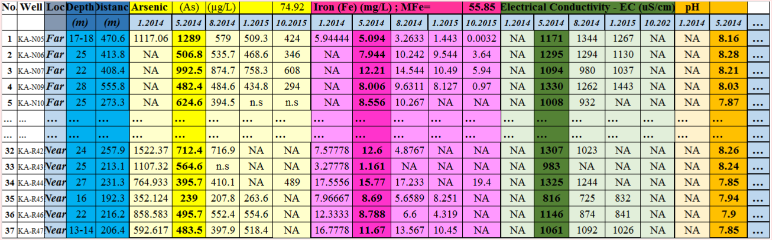
Figure: Groundwater pollution process

Figure 2: The causes of arsenic concentration in Vietnam

We have decided to focus only on *As* as it is a major pollutant.

Arsenic dataset

The research team of Faculty of Environment and Natural Resources (FENR), Ho Chi Minh City, University of Technology (HCMUT) had undertaken a massive exercise in An Giang province to collect data on various elements, including Arsenic (As).



No.	**Well**	**Loc**	**Depth**	**Distanc**	**Arsenic**	(As)	(µg/L)		74.92	**Iron (Fe) (mg/L) ; MFe=**				**55.85**	**Electrical Conductivity - EC (uS/cm)**					**pH**		
			(m)	***(m)***	***1.2014***	***5.2014***	***8.2014***	***1.2015***	***10.2015***	***1.2014***	***5.2014***	***8.2014***	***1.2015***	***10.2015***	***1.2014***	***5.2014***	***8.2014***	***1.2015***	***10.202***	***1.2014***	***5.2014***	**...**
1	KA-N05	***Far***	17-18	470.6	1117.06	**1289**	579	509.3	424	5.94444	**5.094**	3.2633	1.443	0.0032	NA	**1171**	1344	1267	NA	NA	**8.16**	**...**
2	KA-N06	***Far***	25	413.8	NA	**506.8**	535.7	468.6	346	NA	**7.944**	10.242	9.544	3.64	NA	**1295**	1294	1130	NA	NA	**8.28**	**...**
3	KA-N07	***Far***	22	408.4	NA	**992.5**	874.7	758.3	608	NA	**12.21**	14.544	10.49	5.94	NA	**1094**	980	1037	NA	NA	**8.21**	**...**
4	KA-N09	***Far***	28	555.8	NA	**482.4**	484.6	434.8	294	NA	**8.006**	9.6311	8.127	0.97	NA	**1330**	1262	1443	NA	NA	**8.03**	**...**
5	KA-N10	***Far***	25	273.3	NA	**624.6**	394.5	n.s	n.s	NA	**8.556**	10.267	NA	NA	NA	**1008**	932	NA	NA	NA	**7.87**	**...**
...	...	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**
...	...	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**
32	KA-R42	***Near***	24	257.9	1522.37	**712.4**	716.9	NA	NA	7.57778	**12.6**	4.8767	NA	NA	NA	**1307**	1023	NA	NA	NA	**8.26**	**...**
33	KA-R43	***Near***	25	213.1	1107.32	**564.6**	n.s	NA	NA	3.27778	**1.161**	NA	NA	NA	NA	**983**	NA	NA	NA	NA	**8.24**	**...**
34	KA-R44	***Near***	27	231.3	764.933	**395.7**	410.1	NA	489	17.5556	**15.77**	17.233	NA	19.4	NA	**1325**	1244	NA	NA	NA	**7.85**	**...**
35	KA-R45	***Near***	16	192.3	352.124	**239**	207.8	263.6	NA	7.96667	**8.69**	5.6589	8.251	NA	NA	**816**	725	832	NA	NA	**7.94**	**...**
36	KA-R46	***Near***	22	216.2	858.583	**495.7**	552.4	554.6	NA	12.3333	**8.788**	6.6	4.319	NA	NA	**1146**	874	841	NA	NA	**7.9**	**...**
37	KA-R47	***Near***	13-14	206.4	592.617	**483.5**	397.9	518.4	NA	16.7778	**11.67**	13.567	10.45	NA	NA	**1061**	1092	1026	NA	NA	**7.85**	**...**

- It consists of measuring the following characteristics at 5 different time points
  - Geographic characteristics of the water-wells such as depth and distance from the river.
  - Heavy metals (arsenic (As), iron (Fe), lead (Pb), etc.)
  - Chemical characteristics (Salinity (Sal), pH, Electrical Conductivity (EC), etc.)
- After a careful study of the dataset, we decided to focus on the complete observations from 29 locations where arsenic concentration was measured in May and Aug 2014.

Big Question:

What can we do with this data?

To measure the *As* level at any new site ⇒ **COSTLY** and **TIME CONSUMING**

Research Question

How to build a regression model (based on the existing survey data) which can help us predict the arsenic level at a new site within the same geographic region without going through an expensive chemical analysis. It can help us save time and money. Further, at the same time, we would like to have a higher precision in our prediction.

Research Question

How to build a regression model (based on the existing survey data) which can help us predict the arsenic level at a new site within the same geographic region without going through an expensive chemical analysis. It can help us save time and money. Further, at the same time, we would like to have a higher precision in our prediction.

It is further addressed through the following five research questions.

(R1) How to build a reasonably good regression model for arsenic under the normal errors for the given MDR dataset?—**old model**

(R2) How to improve the above regression model further going beyond the normality assumption? —**new model**

(R3) How can we quantify the improvements in predicting the arsenic level using the new approach over the standard one?

The factor are influencing the arsenic contamination

First of all, we plot the scatterplots to understand the basic relationship between arsenic and depth (*Dep*), distance (*Dis*), time (May & Aug 2014). Time was found to be insignificant either as a main factor and/or having any interaction with the other independent variables. We have tried the following cases:

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

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

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

(d) *(As)* regressed on a quadratic expression involving *In(Dep)* and *In(Dis)*.

Our objective in the regression analysis is to find the “optimal" model using normal errors, and then improving it further by generalizing the normal distribution for the errors.

The “best" regression model under the normal distribution

The “best" regression model under the normal distribution is model of case (b)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon, \tag{1}$$

where $Y = ln(As)$, $X_1 = X_1^* =$ standardized (Dep), $X_2 = X_2^* =$ standardized (Dis), $X_3 = X_2^2$, $X_4 = X_1^2 X_2$ and $X_5 = X_1^2 X_2^2$ and $R^2 = 0.41$ is highest one of four cases.

The “best" regression model under the normal distribution

The “best" regression model under the normal distribution is model of case (b)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon, \tag{1}$$

where $Y = ln(As)$, $X_1 = X_1^*$ = standardized (Dep), $X_2 = X_2^*$ = standardized (Dis), $X_3 = X_2^2$, $X_4 = X_1^2 X_2$ and $X_5 = X_1^2 X_2^2$ and $R^2 = 0.41$ is highest one of four cases.

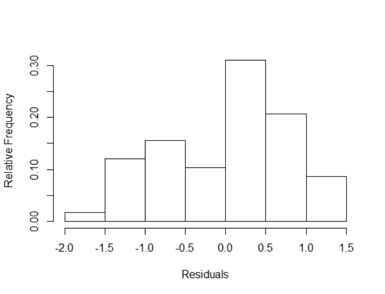


Figure 3: The histogram of the residuals of the model (1) with normal errors

It look likes somewhat skewed=> This justifies our next step has to improve the model (1) beyond the normality assumption.

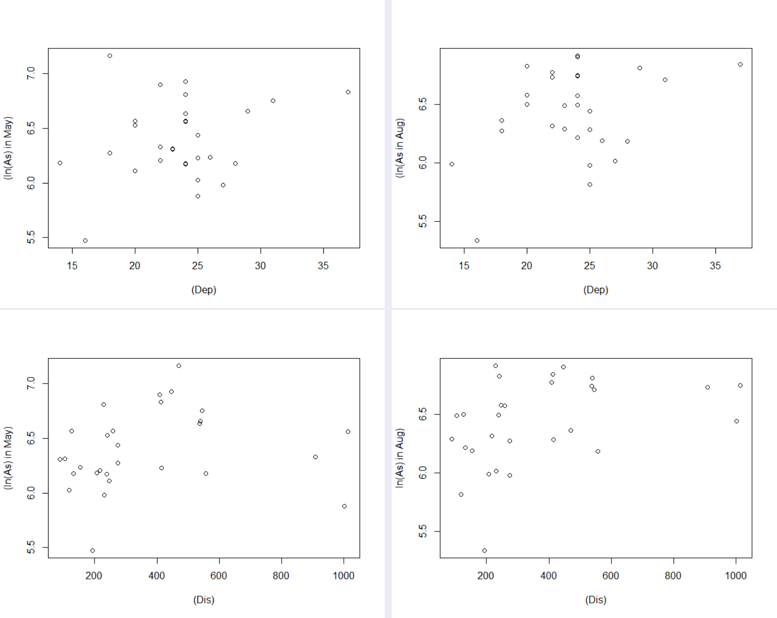


Figure 4: The scatterplots of the best model

For the best model, the normality assumption seems inconclusive as the two standard test methods (Anderson - Darling test (ADT) and Shapiro - Wilk test (SWT)) yield p-values of 0.099 and 0.091 respectively, and the equality of variances uses the standard test (Levene's test and F test) produce a large p-value of 0.8874 and 0.7783 respectively.

SND is a natural generalization of the usual normal distribution. So let us review some basic properties of SND.

Skew-Normal distribution $SND(\mu,\sigma,\lambda)$

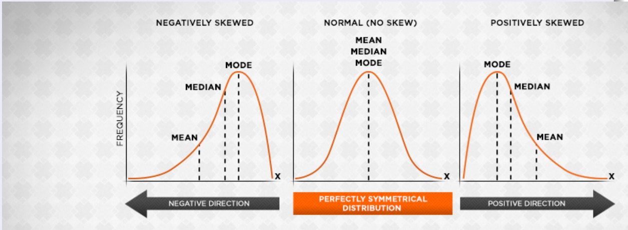
A r.v. $W \sim SND(\mu,\sigma,\lambda)$, provided its *pdf* given as

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

where

- μ is location parameter $\in \mathbb{R}$
- σ is scale parameter $\in \mathbb{R}^+$
- λ is shape (or skew) parameter $\in \mathbb{R}$
- $\phi(.)$ and $\Phi(.)$ are the standard normal *pdf* and *cdf* respectively.

Note that
+ $\lambda = 0 \Rightarrow SND(\mu,\sigma,\lambda) \equiv N(\mu,\sigma^2)$.
+$\lambda > 0 \Rightarrow SND(\mu,\sigma,\lambda)$ is positively skewed.
+ $\lambda < 0 \Rightarrow SND(\mu,\sigma,\lambda)$ is negatively skewed.



It can take a positively skewed, negatively skewed, or perfectly symmetric normal structure through its skew parameter λ. Therefore, SND is a natural generalization of ND.

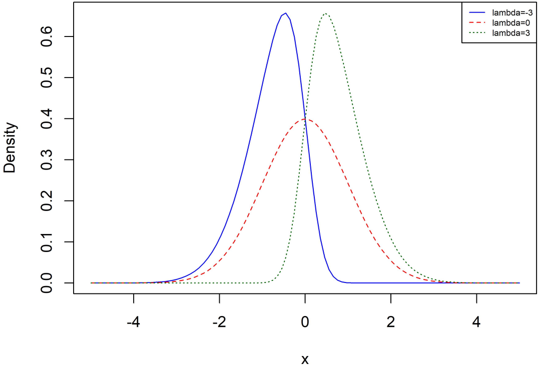


Figure 5: The three *pdf* curves with $\mu = 0, \sigma = 1$, and $\lambda = -3, 0, 3$

When lambda goes to infinity, it looks like a half-normal

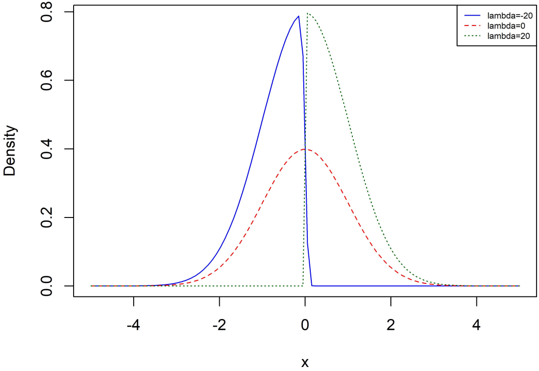


Figure 6: The three *pdf* curves with $\mu = 0, \sigma = 1$, and $\lambda = -20, 0, 20$

- The three parameter SND was first introduced by O'Hagan and Leonard (1976). In the mid-80s, Azzalini (1985, 1986) pioneered the research on SND. Based on his works, several other researchers contributed more on SND research, such as - Roberts (1988), Gupta and Brown (2001), Arellano-Valle et al. (2013), etc.
- Most of work on SND had focused primarily on its properties and characterizations but *not* much research had been done on inferences (such as confidence interval, hypothesis tests, prediction).
- In the present study, we explore how a regression model based on SND could provide a better prediction than the one based on the normal model.

First, let us see some key properties of SND some of which have been used in our research, especially the Property 6.

The useful properties of SND

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

Property 2 The r.v. $W \sim SND(\mu, \sigma, \lambda)$ if and only if $(-W) \sim SND(-\mu, \sigma, -\lambda)$.

Property 3 As $\lambda \to \pm\infty$, $W_* = \frac{(W-\mu)}{\sigma} \sim SND(0, 1, \lambda) \to \pm|Z|$.

Property 4 $W_*^2 = \frac{(W-\mu)^2}{\sigma^2} \sim \chi_1^2$.

Property 5 If U_1, U_2 are *i.i.d.* $\sim N(0, 1)$, Henze (1986) showed that

$$\left(\frac{\lambda}{\sqrt{1+\lambda^2}}\right)|U_1| + \left(\frac{1}{\sqrt{1+\lambda^2}}\right)U_2 \sim \ SND(0, 1, \lambda). \tag{3}$$

Property 6 If $W \sim SND(\mu, \sigma, \lambda)$, then

$$E(W) = \mu + \sigma\sqrt{\frac{2}{\pi}}\left(\frac{\lambda}{\sqrt{1+\lambda^2}}\right), \tag{4}$$

$$V(W) = E(\{W - E(W)\}^2) = \sigma^2\left\{1 - \frac{2\lambda^2}{\pi(1+\lambda^2)}\right\} \tag{5}$$

$$E(\{W - E(W)\}^3) = \sigma^3\sqrt{\frac{2}{\pi}}\left(\frac{4}{\pi} - 1\right)\left\{\frac{\lambda}{\sqrt{1+\lambda^2}}\right\}^3, \tag{6}$$

In our study we would like to show the expressions of two useful properties of SND

- the mode of $SND(\mu, \sigma, \lambda)$ and
- the median of $SND(\mu, \sigma, \lambda)$.

Property 7. The mode of $SND(\mu, \sigma, \lambda)$

The mode can be written as

$$m(\mu, \sigma, \lambda) = \mu + \sigma m_0(\lambda), \tag{7}$$

where $m_0(\lambda)$ is the mode of $SND(0, 1, \lambda)$, and

$$\boldsymbol{m_0(\lambda) = \eta_\lambda - \left(\frac{\gamma_1}{2}\right)\sqrt{1 - \eta_\lambda^2} - \frac{sign(\lambda)}{2} exp\frac{(-2\pi)}{|\lambda|}}, \tag{8}$$

$$\delta = \frac{\lambda}{\sqrt{1 + \lambda^2}}, \tag{9}$$

$$\eta_\lambda = \sqrt{\frac{2}{\pi}}\delta, \tag{10}$$

$$\gamma_1 = (2 - \frac{\pi}{2})(\sigma\sqrt{\frac{2}{\pi}})^3(1 - \frac{2\delta^2}{\pi})^{(-3/2)}. \tag{11}$$

It was mentioned and tabulated by Azzalini and Capitanio (2014).

Note that:

From (8), it is easy to see that $m_0(\lambda) = -m_0(-\lambda)$ for any λ.

Property 8. The median of $SND(\mu, \sigma, \lambda)$

The median can be expressed as

$$M(\mu, \sigma, \lambda) = \mu + \sigma M_0(\lambda), \tag{12}$$

where $M_0(\lambda)$ is the median of $SND(0, 1, \lambda)$ which can be found by solving the following equation for y:

$$\boldsymbol{T(y|\lambda) = \frac{1}{2}\left\{\Phi(y) - 0.5\right\}}, \tag{13}$$

where $T(y|\lambda)$ is the Owen's T-function given as

$$T(y|\lambda) = \frac{1}{2\pi}\int_0^{\lambda} \frac{\left[exp\left\{-y^2(1+x^2)/2\right\}\right]}{(1+x^2)}dx. \tag{14}$$

Note that
If $\lambda = -\lambda*$, where $\lambda* > 0$ then we solve

$$\boldsymbol{T(y|\lambda*) = \frac{1}{2}\left\{0.5 - \Phi(y)\right\}}. \tag{15}$$

From (15), it is easy to see that $M_0(\lambda) = -M_0(-\lambda)$, for any λ.
That's why the following table have been made for $\lambda > 0$ only for the computations for mean, mode and median of $SND(0, 1, \lambda)$.

Objectives

Motivated by the broad research question queries, the specific objectives of this research are given as follows.

(O1) Choosing the “optimal" regression model from a host of competing models under the ND errors, and then improving it further under the SND errors.

(O2) Investigating how to estimate all the model parameters under the SND errors based on a combination of the OLSE+MME.

(O3) Studying the sampling properties of the parameter estimators under the SND errors using the bootstrap method.

(O4) Comparing the two regression models under the ND and SND errors through the AIC.

(O5) Predicting the value of the response variable for a future observation under the SND errors, and comparing it with its counterpart under the ND errors in terms of prediction mean squared error ($PMSE$) and prediction mean absolute error ($PMAE$).

We will consider the first objective (O1)

Building the regression model under ND and SND errors

Under Normal Errors

We consider a multiple linear regression model as

$$Y_j = \beta_0 + \beta_1 X_{1j} + \cdots + \beta_{(p-1)} X_{(p-1)j} + \varepsilon_j, \;\; j = 1, 2, \ldots, n, \tag{16}$$

where $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) \sim^{iid} N(0, \sigma^2)$, $Y_j \sim N(\mu, \sigma^2)$, $\boldsymbol{X} = (1, X_1, \ldots, X_{(p-1)})'$ is the vector of explanatory variables. We have some well-known methods to find the estimated parameters for regression model such as

1. OLSE (Ordinary least square estimation)
2. MME (Method of moments estimation)
3. MLE (Maximum likelihood estimation)

They are all the same under ND errors.

Under SND Errors

If we replace the assumption of normality the above expression by saying that
$\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) \sim^{iid} SND(0, \sigma, \lambda), \;\; Y_j \sim SND(\mu_j, \sigma, \lambda),$

where $\mu_j = \beta_0 + \beta_1 X_{1j} + \cdots + \beta_{(p-1)} X_{(p-1)j}$.

How to estimate all parameters $\boldsymbol{\theta}$? where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma, \lambda)$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{(p-1)})'$.
What are the methods to find the estimators of the parameters?

To find $\widehat{\beta}^S$ of β using OLS under SND errors

Now we have reconsider the above expression (16), which the assumption that the errors $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) \sim^{iid} SND(0, \sigma, \lambda)$, $\mu_j = \beta_0 + \beta_1 X_{1j} + \cdots + \beta_{(p-1)} X_{(p-1)j}$, and

$\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma, \lambda)$ is the parameters of the model where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{(p-1)})'$.

We assume that $\widehat{\boldsymbol{\beta}}^N, \widehat{\sigma}^N, \widehat{\lambda}^N = 0$ are estimated parameters of $\boldsymbol{\beta}, \sigma, \lambda$ respectively under ND model and $\widehat{\boldsymbol{\beta}}^S, \widehat{\sigma}^S, \widehat{\lambda}^S$ are estimated parameters of $\boldsymbol{\beta}, \sigma, \lambda$ respectively under SND model.

The OLSE under SND errors $\widehat{\boldsymbol{\beta}}^S$ of $\boldsymbol{\beta}$ was found by minimizing the sample mean squared error under SND errors and $\widehat{\gamma}^S$ was found via MME method under SND (see (21)), we obtain

$$\widehat{\boldsymbol{\beta}}^N = (\widehat{\beta}^N, \widehat{\beta}_1^N, \ldots, \widehat{\beta}_{(p-1)}^N) \tag{17}$$

$$\widehat{\boldsymbol{\beta}}^S = (\widehat{\beta}_0^N - \widehat{\gamma}^S, \widehat{\beta}_1^N, \ldots, \widehat{\beta}_{(p-1)}^N). \tag{18}$$

where

$$\gamma = E(\varepsilon_j) = \sqrt{\frac{2}{\pi}} \frac{\sigma\lambda}{\sqrt{1+\lambda^2}}, \tag{19}$$

Comparing these estimated parameters under ND and SND errors, we obtain

- **Only the first component of $\widehat{\boldsymbol{\beta}}^S$ is changed**.
- The remaining components of $\widehat{\boldsymbol{\beta}}^S$ are equal to the remaining components of $\widehat{\boldsymbol{\beta}}^N$

To find $\widehat{\sigma}^S, \widehat{\lambda}^S$ of σ, λ respectively using MME method under SND

Define the **first three residual raw moments** as follows:

$$m_k = \left(\sum_{j=1}^{n} \left(e_j^N\right)^k / n \right), \quad k = 1, 2, 3. \tag{20}$$

where $\boldsymbol{e}^N = (\boldsymbol{Y} - \widehat{\boldsymbol{Y}}^N)$ is residual vector and these residuals are now supposed to reflect unobservable errors $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ which are *i.i.d.* $SND(0, \sigma, \lambda)$ where $E(\varepsilon_j) = \gamma$.

- The quantity m_1 is supposed to represent $E(\varepsilon_j - \gamma)$.
- The quantity m_2 is supposed to reflect $E(\varepsilon_j - \gamma)^2 = \sigma^2 - \gamma^2$.
- The quantity m_3 is supposed to reflect $E(\varepsilon_j - \gamma)^3 = \left(2 - \frac{\pi}{2}\right)\gamma^3$.

Therefore, simple algebra leads to

$$\widehat{\sigma}^S = \left\{ m_2 + \left(\widehat{\gamma}^S\right)^2 \right\}^{1/2} \quad \text{, where } \widehat{\gamma}^S = \left(\frac{m_3}{2 - \pi/2} \right)^{1/3}, \tag{21}$$

$$\widehat{\lambda}^S = \begin{cases} sign(\widehat{\gamma}^S)(\widehat{c})^{-1/2} & \text{if} \quad \widehat{c} > 0 \\ sign(\widehat{\gamma}^S)K & \text{if} \quad \widehat{c} < 0 \end{cases} \text{, where } \widehat{c} = \frac{(2/\pi)\left(\widehat{\sigma}^S\right)^2}{\left(\widehat{\gamma}^S\right)^2} - 1 \tag{22}$$

and $K = 10$.

Parameter Estimates for the MDR Arsenic Dataset

We now follow our new method to implement the regression model (16) where $n = 58, p - 1 = 5$. (The superscript "N" and "S" in the estimators indicate the underlying ND and SND errors model, respectively.)

Table 1: Estimated parameters under two models

Parameters	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	$\widehat{\lambda}^N = 0$	$\widehat{\lambda}^S = -1.320$

To show the superiority of SND regression,
we check the goodness of prediction

PREDICTION OF ARSENIC IN MDR
USING REGRESSION MODEL WITH SND ERRORS

The predictive model for a future (a new) observation

$$Y_{(n+1)} = \boldsymbol{X}'_{(n+1)}\beta + \varepsilon_{(n+1)}, \tag{23}$$

Under normal errors $E(\varepsilon) = 0$

$$\begin{aligned} E(Y_{(n+1)}) = \boldsymbol{X}'_{(n+1)}\beta &= mean(Y_{(n+1)}), && (24)\\ &= mode(Y_{(n+1)}) && (25)\\ &= median(Y_{(n+1)}) && (26) \end{aligned}$$

The mean, mode, median are all the same under ND errors model.

Under SND errors $E(\varepsilon) = \gamma = \sqrt{\frac{2}{\pi}}\frac{\sigma\lambda}{\sqrt{1+\lambda^2}}$

$$\text{mean}(Y_{(n+1)}) = E(Y_{(n+1)}) = \eta(Y_{(n+1)}) = \boldsymbol{X}'_{(n+1)}\beta + \gamma, \tag{27}$$

$$\text{mode}(Y_{(n+1)}) = m(Y_{(n+1)}) = \boldsymbol{X}'_{(n+1)}\beta + \sigma m_0(\lambda), \tag{28}$$

$$\text{median}(Y_{(n+1)}) = M(Y_{(n+1)}) = \boldsymbol{X}'_{(n+1)}\beta + \sigma M_0(\lambda), \tag{29}$$

They are all different under the SND errors model.

The predictive value of Y_{n+1}

Under normal errors $E(\varepsilon) = 0$

The predictive value of Y_{n+1} under normal errors

$$\widehat{Y}^{N}_{(n+1)} = \boldsymbol{X}'_{(n+1)}\widehat{\boldsymbol{\beta}}^{N}, \tag{30}$$

Under SND errors $E(\varepsilon) = \gamma = \sqrt{\frac{2}{\pi}}\frac{\sigma\lambda}{\sqrt{1+\lambda^2}}$

The three predictors of Y_{n+1} are given as

$$\widehat{Y}^{S1}_{(n+1)} = \boldsymbol{X}'_{(n+1)}\widehat{\boldsymbol{\beta}}^{S} + \widehat{\gamma}^{S} = \boldsymbol{X}'_{(n+1)}\widehat{\boldsymbol{\beta}}^{\gamma S} = \boldsymbol{X}'_{(n+1)}\widehat{\boldsymbol{\beta}}^{N}, \tag{31}$$

$$\widehat{Y}^{S2}_{(n+1)} = \boldsymbol{X}'_{(n+1)}\widehat{\boldsymbol{\beta}}^{S} + \widehat{\sigma}^{S} m_0(\widehat{\lambda}^{S}), \tag{32}$$

$$\widehat{Y}^{S3}_{(n+1)} = \boldsymbol{X}'_{(n+1)}\widehat{\boldsymbol{\beta}}^{S} + \widehat{\sigma}^{S} M_0(\widehat{\lambda}^{S}), \tag{33}$$

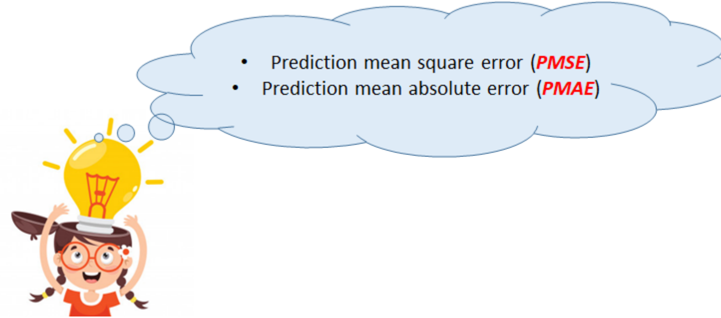
where $\widehat{\gamma}^{S}$, $\widehat{\sigma}^{S}$ are in (21), and $\widehat{\lambda}^{S}$ is in (22).

Interestingly, the mean predictor under SND model is same as that under ND model.

⇒ SND errors is generalized form of normal errors.

Among these three predictors under SND errors, which one is better?

How can we compare the performance of three predictors?



- Prediction mean square error (***PMSE***)
- Prediction mean absolute error (***PMAE***)

PMSE, PMAE

PMSE calculates the average squared differences between the predicted values of the random variable and the true value of the random variable. Similar to *PMSE*, *PMAE* measures the absolute differences between the two objects mentioned above.

Under normal model

$$PMSE(\widehat{Y}^{N}_{n+1}) = \sigma^2 \left\{1 + \boldsymbol{X}'_{(n+1)}(\mathbb{X}'\mathbb{X})^{-1}\boldsymbol{X}_{(n+1)}\right\}, \tag{34}$$

$$PMAE(\widehat{Y}^{N}_{n+1}) = \sigma\sqrt{2/\pi}\left\{1 + \boldsymbol{X}'_{(n+1)}(\mathbb{X}'\mathbb{X})^{-1}\boldsymbol{X}_{(n+1)}\right\}^{1/2}. \tag{35}$$

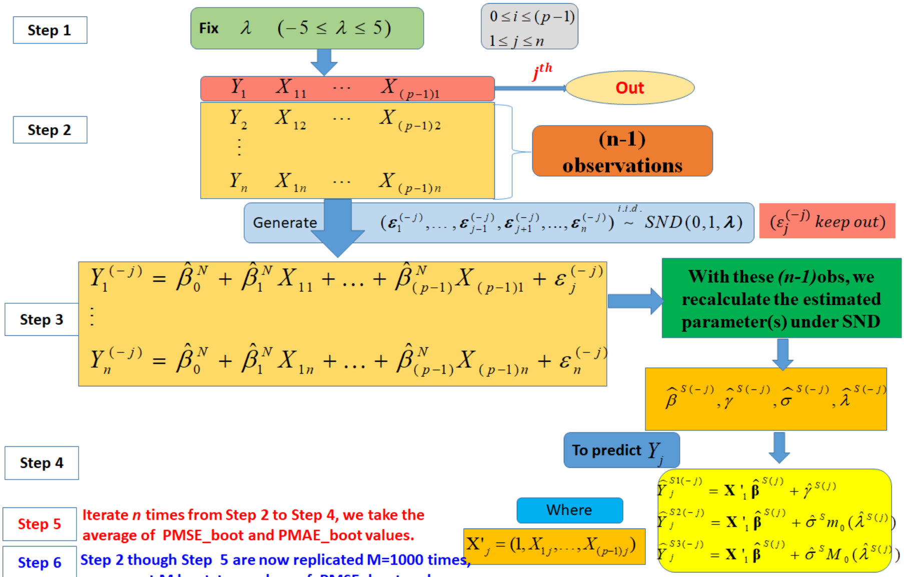
Under SND errors

PMSE and *PMAE* do not have simple expressions for the three predictors $\widehat{Y}^{S1}_{n+1}$, $\widehat{Y}^{S2}_{n+1}$, $\widehat{Y}^{S3}_{n+1}$.

⇒ we are going to aprroximate PMSE and PMAE using **bootstrap approach**.

Our bootstrap approach is called **'Leave One Out Bootstrap' (or LOOB)** where $(n-1)$ observations (out of n) of the given dataset are used to fit the regression model in order to predict the remaining observation's response variable.

Algorithmic Steps of LOOB



Step 1

Fix $\lambda \quad (-5 \le \lambda \le 5)$

$0 \le i \le (p-1)$

$1 \le j \le n$

Step 2

$Y_1 \quad X_{11} \quad \cdots \quad X_{(p-1)1}$

j^{th}

Out

$Y_2 \quad X_{12} \quad \cdots \quad X_{(p-1)2}$

$\vdots$

$Y_n \quad X_{1n} \quad \cdots \quad X_{(p-1)n}$

(n-1) observations

Generate $(\boldsymbol{\varepsilon}_1^{(-j)}, \ldots, \boldsymbol{\varepsilon}_{j-1}^{(-j)}, \boldsymbol{\varepsilon}_{j+1}^{(-j)}, \ldots, \boldsymbol{\varepsilon}_n^{(-j)}) \overset{i.i.d.}{\sim} SND(0,1,\boldsymbol{\lambda})$

$(\varepsilon_j^{(-j)}$ *keep out*$)$

Step 3

$$Y_1^{(-j)} = \hat{\beta}_0^N + \hat{\beta}_1^N X_{11} + \ldots + \hat{\beta}_{(p-1)}^N X_{(p-1)1} + \varepsilon_j^{(-j)}$$

$$\vdots$$

$$Y_n^{(-j)} = \hat{\beta}_0^N + \hat{\beta}_1^N X_{1n} + \ldots + \hat{\beta}_{(p-1)}^N X_{(p-1)n} + \varepsilon_n^{(-j)}$$

With these *(n-1)*obs, we recalculate the estimated parameter(s) under SND

$$\hat{\beta}^{S(-j)}, \hat{\gamma}^{S(-j)}, \hat{\sigma}^{S(-j)}, \hat{\lambda}^{S(-j)}$$

Step 4

To predict Y_j

$$\hat{Y}_j^{S1(-j)} = \mathbf{X}'_1 \hat{\boldsymbol{\beta}}^{S(j)} + \hat{\gamma}^{S(j)}$$

$$\hat{Y}_j^{S2(-j)} = \mathbf{X}'_1 \hat{\boldsymbol{\beta}}^{S(j)} + \hat{\sigma}^S m_0(\hat{\lambda}^{S(j)})$$

$$\hat{Y}_j^{S3(-j)} = \mathbf{X}'_1 \hat{\boldsymbol{\beta}}^{S(j)} + \hat{\sigma}^S M_0(\hat{\lambda}^{S(j)})$$

Where

$\mathbf{X}'_j = (1, X_{1j}, \ldots, X_{(p-1)j})$

Step 5 Iterate *n* times from Step 2 to Step 4, we take the average of PMSE_boot and PMAE_boot values.

Step 6 Step 2 though Step 5 are now replicated M=1000 times, we can get M bootstrap values of PMSE_boot and PMAE_boot. To get more stable values we take the average of PMSE_boot and PMAE_boot values in Step 5.

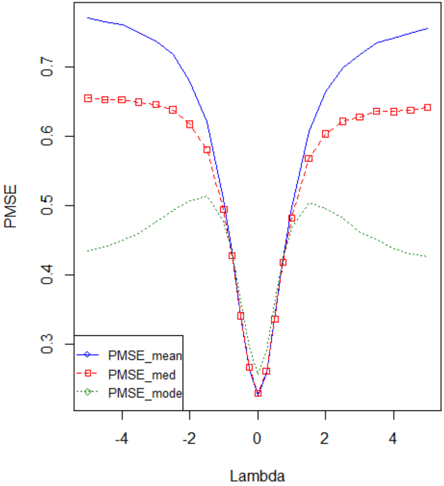


Figure 7: The plots of PMSE curves of the three predictors of $\ln(As)$ as functions of λ

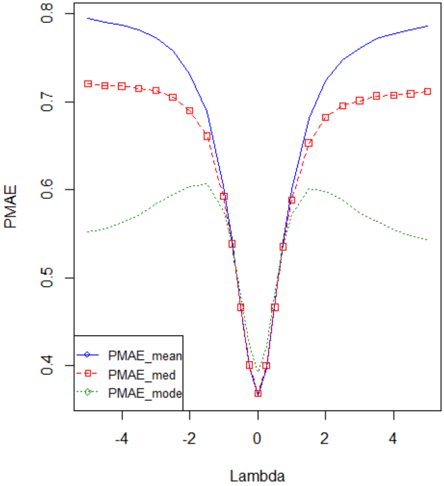


Figure 8: The plots of PMAE curves of the three predictors of $\ln(As)$ as functions of λ

Remark

- As λ moves away from 0, the mode and median predictors are showing better performance than the mean predictor (where the mode predictor is the best).
- As $\lambda = 0$ SND model boil down to ND model $\Rightarrow$ all three predictors should be coincide, and their performance should be same.

To quantify how many percent that our SND is better than ND error.

Relative Improvement

Table 2: Relative Improvement (RI) over the usual predictor of $ln(As)$ at $\lambda \approx \widehat{\lambda}^S = -1.320$

Predictor	*PMSE*	*PMAE*
$\widehat{Y}^{S1}$	0.593	0.688
$\widehat{Y}^{S2}$	0.506 (RI = 14.67%)	0.592 (RI = 11.38%)
$\widehat{Y}^{S3}$	0.555 (RI = 6.41%)	0.639 (RI = 4.34%)

- The above RIs (ranging from 4.34% to 14.67%).
- Our SND model is giving an improvement from 4.34% to 14.67% over the traditional model.

Conclusion

Conclusion

- While ND model provides a unique predictor, our proposed SND model provides 3 predictors, **one of which coincides with the one of ND model.**
- Using the twin criteria of *PMSE* and *PMAE* it has been shown through bootstrap that **the mode and median predictors** of our proposed SND model are **far superior** to the one under the **ND model** (i.e., the mean predictor of SND model).
- Hopefully the technique developed here in parameter estimation, as well as subsequent inferences can be replicated for many other similar studies.

SCOPE OF FURTHER RESEARCH

(6.2) To consider the whole dataset where many observations are missing, then employ the 'Expectation - Maximization Algorithm' (EMA) for estimation of parameters - for both two regression models based on ND and SND errors.

(6.1) To explore a natural extension for a further generalization using a multivariate SND assumption.

(6.3) To consider what effect do such Bayesian estimates have in terms of predicting the response variable.

THANK YOU
FOR YOUR KIND ATTENTION
TAIMIENPHIV

References

Akaike H. (1973). *Information theory and an extension of the maximum likelihood principle*. Second International Symposium on Information Theory (Petrov BN and Czáki E, (Eds.)), Akademiai Kiadó, Budapest. 267-281.

Alhamide AA, Ibrahim K, Alodat MT. (2015). *Multiple linear regression estimators with skew normal errors*. American Institute of Physics.

Arellano-Valle RB, Castro LM, Loschi RH. (2013). *Change Point Detection in the Skew-Normal Model Parameters*. Communications in Statistics-Theory and Methods. 42: 603-618.

Arnold BC, Lin GD. (2003). *Characterizations of the skew-normal and generalized Chi distributions*. The Indian Journal of Stat. 66: 593-606.

Azzalini A. (1985). *A class of distributions which includes the normal ones*. Scandinavian journal of statistics. 12: 171-178.

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

Azzalini A, Arellano-Valle RB. (2013). *Maximum Penalized Likelihood Estimation for Skew-normal and Skew-t Distributions*. Journal of Statistical Planning and Inference 143: 419-433.

REFERENCES

Azzalini A, Capitanio A. (2014). *The skew-normal distribution and related families*. Cambridge University Press.

Casella G, Berger RL. (2002). *Statistical inference*. Duxbury Press.

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

Figueiredo F, Gomes MI. (2013). *The skew-normal distribution in SPC*. REVSTAT – Statistical Journal. 11: 83-104.

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

Henderson DJ, Parmeter CF. (2015). *Applied nonparametric econometrics*. Cambridge University Press, New York.

Henze N. (1986). *A probabilistic representation of the 'Skew-normal' distribution*. Scand J Statist 13: 271-275.

Greene WH. (2012). *Economic Analysis*. Pearson.

REFERENCES

Gupta AK, Nguyen TT, Sanqui JT. (2004). *Characterization of the Skew-normal distribution*. Ann. Inst. Statist. Math 56: 351-360.

Gupta RC, Brown N. (2001). *Reliability studies of the skew normal distribution and its application to a Strength-Stress Model*. Communications in Statistics-Theory Methods. 11: 2427-2445.

Lachos VH, Dey DK, Cancho VG. (2009). *Robust linear mixed models with skew-normal independent distributions from a Bayesian perspective*. Journal of Statistical Planning and Inference. 139(2): 4098-4110.

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

Lim TS, Loh WY. (1995). *A comparison of tests of equality of variances*. Computational Statistics & Data Analysis . 22(3): 287-301.

Nakamura G. (2007). *Defoliation during the Vietnam war. Chapter 9: Extreme conflict and tropical forests. Defoliation during the Vietnam War*. Springer. 149-158.

Ngunkeng G, Ning W. (2014). *Information approach for the change-point detection in the skew normal distribution and its applications*. Sequential Analysis. 33: 475-490.

Nguyen PK. (2008). *Geochemical study of arsenic behavior in aquifer of the Mekong Delta, Vietnam*. PhD Dissertation. Kyushu University.

REFERENCES

O'Hagan A, Leonard T. (1976). *Bayes estimation subject to uncertainty about parameter constraints*. Biometrika. 63: 201-202.

Pérez-Rodríguez P, Acosta-Pech R, Pérez-Elizalde S, Cruz CV, Espinosa JS, Crossa J. (2018). *A bayesian genomic regression model with skew normal random errors*. G3: GENES, GENOMES, GENETICS. 8(5): 1771-1785.

Pham CHV. (2015). *Studying the mechanisms of arsenic release in groundwater in An Phu District, An Giang Province*. Master Thesis. University of Technology, Ho Chi Minh City, Vietnam.

Roberts HV. (1988). *Data analysis for managers with Minitab*. Scientific Press: Redwood City, CA.

RStudio Team. (2019). *RStudio: Integrated development for R. RStudio, Inc., Boston, MA* (version 2019 Apr 8). Available from: `http://www.rstudio.com/`.

Sahu SK, Dey DK, Branco MD. (2003). *A new class of multivariate skew distributions with applications to Bayesian regression models*. The Canadian Journal of Statistics. 31(2): 129-150.

REFERENCES

Thiuthad P, Pal N. (2019). *Point estimation of the location parameter of a skew-normal distribution: Some fixed sample and asymptotic results*. Journal of Statistical Theory and Practice. 13(2) 13-37.

Young AL, Regigani GM. (1988). *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. 10-33.

Vo LP, Bernier R, Pham CHV, Ho TNH, Nguyen TBT.(2015). *Threat of arsenic occurrence in the Vietnamese Mekong Delta*. Journal of Geographical Research. 63:129-142.

Vo LP, Pham CHV, Nguyen VMM, Pham KBA, Vu VA, Nguyen TBT. (2016). *Arsenic pollution in Shallow Groundwater in a Floodplain Delta: A Case Study in An Phu, An Giang, in Mekong Delta, Vietnam*. Journal of Science and Technology. 54.

Yap BW, Sim CH. (2011). *Comparisons of various types of normality tests* . Journal of Statistical Computation and Simulation. 81(12): 2141-2155.