Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 1, 2021 (New Series), pp 115–124

Nonlinear Error-in-Variables Regression with the Error-in-Variables Distribution Estimated by Censored Data

M. S. Hamada[1] **and K. A. Kaufeld**[1]
[1]*Statistical Sciences*
Los Alamos National Laboratory, New Mexico, USA

Received: 23 October 2020; Revised: 22 November 2020; Accepted: 26 November 2020

Abstract

This article considers the analysis of data using a nonlinear regression model in which the covariate has a distribution, *i.e.*, the error-in-variables case. Moreover, some of the same data consisting of left- and right-censored data are used to estimate the covariate distribution. We show how to simultaneously fit the nonlinear error-in-variables regression model and estimate the covariate distribution using Bayesian inference. The proposed method is illustrated with a simulated data set. We also show the impact of knowing the covariate distribution and the actual covariate values. Furthermore, we show the impact of taking additional data on inference and prediction.

Key words: Bayesian inference; Left-censored; Markov chain Monte Carlo; Prediction; Right-censored.

AMS Subject Classifications: 62F15, 62J02, 62N01, 62P30

1. Introduction

It is our privilege to to contribute this article to the special issue of Statistics and Applications in honor of Professor Dey. The first author met Professor Dey when he visited the University of Waterloo in the late 1980's. At the time, research in the design of experiments for improving quality and productivity in industry had been reinvigorated by the appearance of Taguchi Methods. Professor Dey's 1985 book was timely for its mixed-level orthogonal arrays that were being promoted by the Taguchi Methods. The Wu and Hamada (2009) Experiments book refers to Professor Dey's 1985 book as well as his 1999 book with Professor Mukerjee a number of times for theoretical details and presents tables of his $\mathrm{OA}(24, 6^1, 2^{14})$, $\mathrm{OA}(54, 2^1, 3^{25})$, and $\mathrm{OA}(54, 6^1, 3^{24})$ designs for use by practitioners. The first author fondly remembers Professor Dey as a formal gentleman and seasoned scholar who kindly spent time talking to a young assistant professor about research. In this article, we present a problem that we faced on a project at work. Here we focus on data analysis although there is a design aspect that could be explored.

Suppose that we sample a population each year for I years, $i = 1, \ldots, I$. At year i, we sample a unit and record whether a feature of interest can be observed in the unit. For

Corresponding Author: Michael Hamada
Email: hamada@lanl.gov

example, cracks occur on containers based upon stresses on the container. The initiation crack area or subsequent cracks occur at some time point. If a crack is observed we know that the crack started sometime before that year. Otherwise, we know the crack will start after the recorded time. In terms of chemical reactions, we can think about a reaction occurring at a recorded time. If observed, all that we know is that the mechanism started in the unit before year i; if it is not observed, all that we know is that the mechanism will start in the unit after year i. That is, we assume that the mechanism will start at some time in all units so that there is a start time distribution. The data in which the mechanism has not started are right-censored data. The data in which the mechanism has started are left-censored data. An example of a model that displays similar characteristics is convex degradation where the degradation rate increases with the level of degradation (Meeker and Escobar, 1998). Suppose that the start time distribution is $Lognormal(\mu, \sigma^2)$, say $Lognormal(3, 0.1^2)$ with median 20.1 years and 0.95 probability interval (14.4, 27.0) years. Recall that the log start time distribution is $Normal(\mu, \sigma^2)$. The proportion of the population that the mechanism has started at time t is displayed in Figure 1.

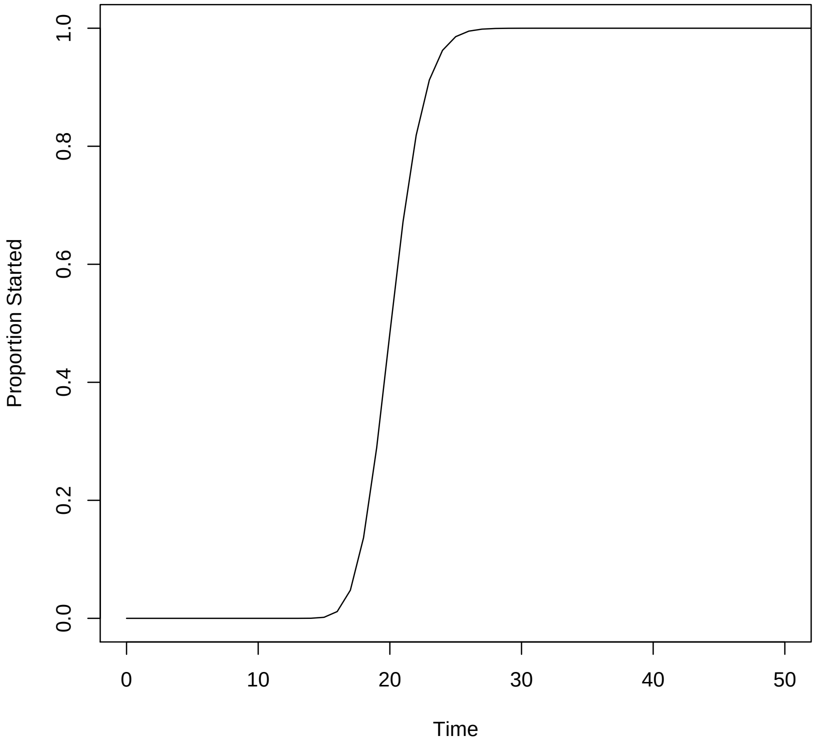

Figure 1: Proportion of population that mechanism has started versus time (years).

Suppose that for those units in which the mechanism has started we observe a quantity Y_t at time t, which is modeled as $Y_t = \beta_0(1 - \exp(-\beta_1 e_t)) + \epsilon_t$. This is a nonlinear regression model with mean $\beta_0(1 - \exp(-\beta_1 e_t))$, where e_t is the elapsed time between the time when the mechanism started s (*i.e.*, the start time) and time t (*i.e.*, $e_t = t - s$). ϵ_t is the population error assumed to be distributed as $Normal(0, \sigma_\epsilon^2)$ and is assumed independent of the start time s. Suppose that $\beta_0 = 1000$, $\beta_1 = 0.025$ and $\sigma_\epsilon = 1$. The mean $\beta_0(1 - \exp(-\beta_1 e_t))$ versus elapsed time e_t is displayed in Figure 2.

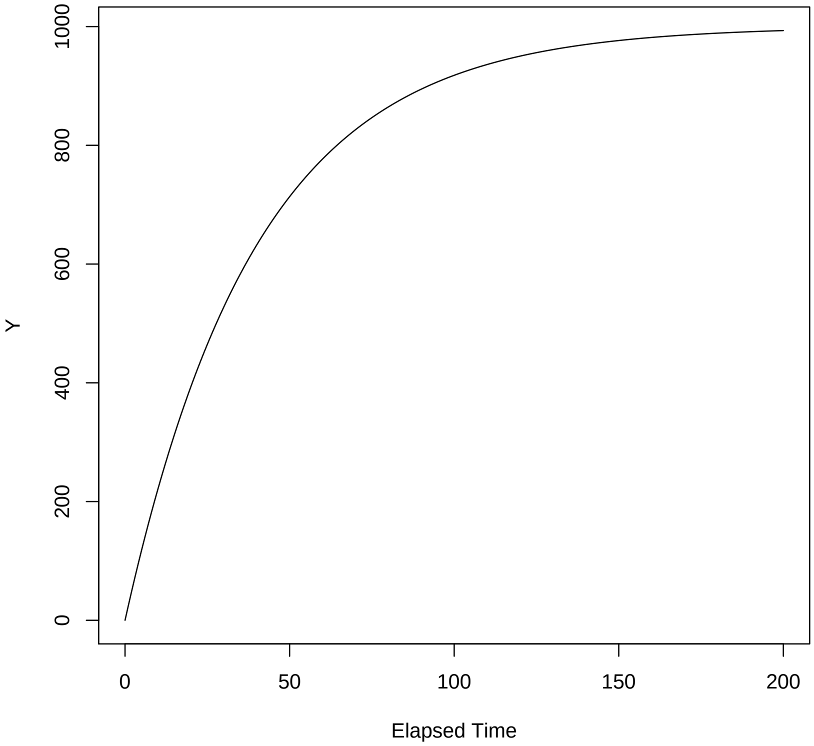

Figure 2: Quantity Y nonlinear regression mean versus elapsed time (years).

2. Data Model and Analysis

In our scenario, we will use example data shown in Table 1, where one unit is sampled per year for 30 years at times 1-30. The StartStar column is 1 if the mechanism is not observed to have started; otherwise, 0 if the mechanism is observed to have started. The Y column is the quantity Y if the mechanism is observed to have started; *e.g.*, if a crack is observed, Y might be the the length of the crack. The elapsed time is not known; we only know that at time t with StartStar equal to 0, the start time s is less than t, *i.e.*, the elapsed time e_t is a random variable, $t - s$, where $s \sim Lognormal(3, 0.1^2)I(0, t)$ and $I(0, t)$ indicates that the lognormal distribution is restricted to the interval $(0, t)$. Because e_t is not known exactly, but has a distribution, the nonlinear regression model of Y is an error-in-variables model where the covariate e_t has a distribution and not an exactly known value.

Further, we use the Time-StartStar data to estimate μ and σ for the $Lognormal(\mu, \sigma^2)$ start time distribution. For a StartStar of 1, say, for Time 2, the likelihood contribution is $1 - \Phi(\frac{\ln(2)-\mu}{\sigma})$, the probability of observing a right-censored datum, where $\Phi()$ is the normal cumulative distribution function. For a StartStar of 0, say, for Time 19, the likelihood contribution is $\Phi(\frac{\ln(19)-\mu}{\sigma})$, the probability of observing a left-censored datum.

Table 1: Example Data (ordered so that right-censored data appear first; Elapsed Time is unknown to the analyst)

Time (year)	StartStar	Elapsed Time (year)	Y
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0
5	1	0	0
6	1	0	0
7	1	0	0
8	1	0	0
9	1	0	0
10	1	0	0
11	1	0	0
12	1	0	0
13	1	0	0
14	1	0	0
15	1	0	0
16	1	0	0
17	1	0	0
18	1	0	0
21	1	0	0
19	0	0.47	12.36
20	0	1.15	27.18
22	0	0.51	13.70
23	0	3.66	85.82
24	0	6.62	151.98
25	0	5.41	125.38
26	0	4.68	110.55
27	0	3.66	88.28
28	0	5.92	138.13
29	0	9.97	219.96
30	0	6.99	160.38

We use a Bayesian analysis with the following relatively diffuse prior distributions (Gelman *et al.*, 2013):

- $\beta_0 \sim Lognormal(7, 0.5^2)$ with a 0.95 probability central interval of (411.6, 2921.9)
- $\beta_1 \sim Lognormal(-4, 1^2)$ with a 0.95 probability central interval of (0.003, 0.130)
- $\sigma_t \sim HalfNormal(0, \sqrt{10}^2)$ with a 0.95 probability central interval of (0.099, 7.088)

- $\mu \sim \mathit{HalfNormal}(0, \sqrt{10}^2)$ with a 0.95 probability central interval of (0.099, 7.088)
- $\sigma \sim \mathit{HalfNormal}(0, \sqrt{10}^2)$ with a 0.95 probability central interval of (0.099, 7.088)

These prior distributions are thought to be relatively diffuse, *i.e.*, they are chosen to be quite wide so that the true values of the parameters are thought to fall within these high probability central intervals.

We obtain the following results using a Markov chain Monte Carlo (MCMC) algorithm (Gelman *et al.*, 2013) implemented in JAGS (Plummer, 2003) using the R (R Core Team, 2020) package rjags to call JAGS. The JAGS code for the proposed analysis is given in the Appendix that produces 400,000 draws from the posterior distribution. In the examples, we use 10,000 burnin draws (which are discarded) and 40,000,000 subsequent draws, which we thin by taking every 100th draw. Plots of the posterior draws not shown here display good mixing. Moreover, diagnostics (Gelman and Rubin's convergence diagnostic; Gelman and Rubin, 1992) also suggest convergence, *i.e.*, these draws are from the appropriate posterior distribution.

Table 2 displays the posterior summaries for the model parameters μ and σ for the start time distribution and β_0, β_1 and σ_t for the quantity Y nonlinear regression model. Note that there is substantial uncertainty associated with σ_t.

Table 2: Posterior Summaries of Model Parameters (50, 2.5, 97.5 percentiles) from Table 1 Data

Parameter	True	50%	2.5%	97.5%
μ	3.000	2.975	2.853	3.073
σ	0.100	0.127	0.077	0.269
β_0	1000.000	1106.450	502.993	2583.997
β_1	0.025	0.018	0.007	0.046
σ_t	1.000	2.134	0.100	6.999

2.1. Impact of unknown starting times

There are two impacts of not knowing the starting times. First, the start time distribution parameters are estimated from the left- and right-censored start times. Second, the elapsed times are unknown because of the unknown start times; that is, the covariate in the nonlinear regression model is not known exactly and is referred to as an error-in-variables case. Table 3 shows the impact of using the true error-in-variables (E-I-V) distribution ($\mathit{Lognormal}(3, 0.1^2)$) as well as that of using the actual elapsed times (see Table 1 for the actual elapsed times). We see that using the true EIV distribution provides no improvement, at least for this one data set, but the nonlinear regression model parameters are substantially better estimated (with reduced uncertainty) when the actual elapsed times are used as compared with Table 2.

Table 3: Posterior Summaries of Model Parameters (50, 2.5, 97.5 percentiles) for Some Hypothetical Situations

Parameter	True	50%	2.5%	97.5%
use true E-I-V distribution				
β_0	1000.000	1121.012	509.559	2652.956
β_1	0.025	0.019	0.008	0.049
σ_t	1.000	2.132	0.101	7.023
use actual elapsed times				
β_0	1000.000	1012.378	851.154	1268.457
β_1	0.025	0.025	0.019	0.030
σ_t	1.000	1.016	0.662	1.821

3. Prediction

Suppose that we want to predict a percentile of the Y distribution, say the 90th percentile, at a given time, say 30 years. Suppose that the population size is 1,000. We can draw from the start time distribution to obtain start times and predict Y using the elapsed times (30 minus start times) and the nonlinear regression model, *i.e.* draw 1,000 Y's from the Y distribution. The 90th percentile is the 900th ordered prediction. We do this 10,000 times and take the 95th percentile of the 10,000 90th percentiles to obtain 165.43; for brevity we refer to this as the 90th percentile of the population Y distribution or even shorter as the 90th percentile. For times of 45 and 60 years, the 90th percentiles of the population Y distribution are 426.42 and 605.78, respectively. Based on the proposed analysis, we can obtain a posterior predictive distribution and a 0.95 probability upper bound on the the 90th percentile of the population Y distribution. Table 4 shows the Table 1 data posterior 90th percentile at times 30, 45, and 60 years. The posterior 90th percentiles are somewhat higher that the true 90th percentiles especially at times past the data, *i.e.*, 45 and 60 years.

Table 4: True and Table 1 Data 90th Percentiles at 30, 45, and 60 Years

Time	True	Table 1 Data
(year)	Percentile	Percentile
30	165.43	165.56
45	426.42	440.09
60	605.78	669.36

4. Impact of Taking More Samples

We can also consider the impact of taking more samples per year and taking samples for more than 30 years, *e.g.*, 60 years, using the proposed analysis. Table 5 shows the results when 60 total samples are taken. We use the notation 1@1(1)30 for the Table 1 sampling scheme, *i.e.*, 1 sample each year from years 1 to 30. Table 5 shows results for 2@1(1)30, 1@1(1)60, and 2@2(2)60; 2@2(2)60 denotes 2 samples in every even year from year 2 to year 60. Note that the first two schemes add to the Table 1 data. The 2@2(2)60 scheme uses data from the even years of the 2@1(1)30 scheme. Table 6 shows the results when 120 total samples are taken; 60 additional samples are added to the data analyzed that produced the Table 5 results. Table 6 shows results for 4@1(1)30, 2@1(1)60, and 4@2(2)60.

Increasing the sample from 1 to 2 to 4 per year (1@1(1)30, 2@1(1)30, 4@1(1)30) helps to estimate σ better; estimation for β_0 seems somewhat worse but recall these results are for one realization of the data. Spreading out the inspections across more years helps much more, *e.g.*, (1@1(1)60, 2@2(2)60) and (2@1(1)60, 4@2(2)60). Inspections on even years and more samples at each inspection helps more than inspecting every year with less samples at each inspection. It is noteworthy that none of sampling schemes had an impact on estimating σ_t so that the posterior distributions are similar to the prior distribution.

Table 5: Posterior Summaries of Model Parameters (50, 2.5, 97.5 percentiles) Using More Samples and More Years (60 total samples)

Parameter	True	50%	2.5%	97.5%
		2@1(1)30		
μ	3.000	3.013	2.938	3.077
σ	0.100	0.113	0.079	0.182
β_0	1000.000	1242.127	574.067	2762.086
β_1	0.025	0.020	0.008	0.049
σ_t	1.000	2.163	0.103	7.061
		1@1(1)60		
μ	3.000	2.998	2.907	3.079
σ	0.100	0.125	0.099	0.165
β_0	1000.000	1058.932	859.737	1523.670
β_1	0.025	0.023	0.013	0.033
σ_t	1.000	2.154	0.098	7.036
		2@2(2)60)		
μ	3.000	2.996	2.915	3.069
σ	0.100	0.108	0.086	0.141
β_0	1000.000	1117.917	916.186	1561.584
β_1	0.025	0.021	0.013	0.029
σ_t	1.000	2.168	0.096	7.199

Like Table 4 for the 1@1(1)30 sampling scheme, Tables 7 and 8 show the posterior 90th percentiles at 30, 45, and 60 years for the various sampling schemes with 60 and 120 total samples, respectively. Overall, the posterior 90th percentiles are quite close to the true 90th percentiles. The results for 2@1(1)30 and 4@1(1)30) are worse caused by the worse estimation for β_0 as noted previously. For some of the schemes, the posterior 90th

Table 6: Posterior Summaries of Model Parameters (50, 2.5, 97.5 percentiles) Using More Samples and More Years (120 total samples)

Parameter	True	50%	2.5%	97.5%
		4@1(1)30		
μ	3.000	3.026	2.983	3.065
σ	0.100	0.091	0.072	0.123
β_0	1000.000	1237.483	580.953	2966.605
β_1	0.025	0.021	0.008	0.053
σ_t	1.000	2.069	0.095	6.726
		2@1(1)60		
μ	3.000	3.004	2.943	3.061
σ	0.100	0.121	0.103	0.146
β_0	1000.000	986.211	866.839	1171.483
β_1	0.025	0.025	0.019	0.032
σ_t	1.000	2.214	0.102	7.207
		4@2(2)60		
μ	3.000	3.019	2.964	3.071
σ	0.100	0.103	0.088	0.124
β_0	1000.000	978.467	881.563	1144.787
β_1	0.025	0.026	0.020	0.032
σ_t	1.000	2.085	0.097	6.875

percentiles are slightly less the true 90th percentiles; again these results are for one realization of the data.

Table 7: 60 Sample Data 90th Percentiles at 30, 45, and 60 Years

Time	True	1@1(1)30	2@1(1)30	1@1(1)60	2@2(2)60
(year)	Percentile	Percentile	Percentile	Percentile	Percentile
30	165.43	165.56	179.38	163.44	166.44
45	426.42	440.09	502.20	424.13	421.92
60	605.78	669.36	767.29	611.60	611.29

5. Discussion

In this article, we considered a nonlinear regression model with elapsed time as a covariate for a quantity Y. The elapsed time is the difference between the inspection time and the time when a mechanism started. At inspection, we only know that the mechanism has started or not so that the elapsed time is unknown, the error-in-variables case. Our proposed method analyzes the right- and left-censored elapsed time data to estimate the elapsed time distribution. This analysis is achieved simultaneously with analyzing the error-in-variables (E-I-V) nonlinear regression model for the Y data, where the elapsed time distribution is the E-I-V distribution. Besides the original 30 sample scheme, we showed results for various 60

Table 8: 120 Sample Data 90th Percentiles at 30, 45, and 60 Years

Time (year)	True Percentile	4@1(1)30 Percentile	2@1(1)60 Percentile	4@2(2)60 Percentile
30	165.43	187.22	161.56	167.15
45	426.42	520.54	422.68	430.83
60	605.78	803.17	603.25	610.77

sample and 120 schemes. Note that the results are based on one data set for each of these schemes where the smaller schemes data or parts of the smaller schemes data are included in the larger schemes data. Generally, the results improve for more samples per year over more years. A more extensive study using more data sets, say 500 or more, would solidify the results but would require access to a large computer cluster. Future research might consider an optimal sampling scheme that specifies how may samples and what inspection times to takes the samples. It would be natural to use a Bayesian design criterion because of the proposed Bayesian analysis method.

Acknowledgements

We thank Cee Cee Essex for her support and encouragement. We also thank an anonymous referee provided helpful comments on an earlier version of this article.

References

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2013). *Bayesian Data Analysis Third Edition*. Boca Raton: Chapman & Hall/CRC.

Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statisistical Sciences*, **7**, 457-472.

Meeker, W.Q. and Escobar, L.A. (1998). *Statistical Methods for Reliability Data*. New York: John Wiley & Sons, Inc.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, March 20-22, Vienna, Austria. R Core Team (2020). *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna. (http://www.R-project.org)

Wu, C.F.J. and Hamada, M.S. (2009). *Experiments: Planning, Analysis and Optimization, Second Edition*. New York: John Wiley and Sons, Inc.

APPENDIX

This appendix presents JAGS code for the proposed analysis. In the code:

- startStar is 1 if right-censored and 0 if left-censored, *i.e.*, the mechanism has not started or has started, respectively
- right-censored data are ordered first
- N1=19, number of right-censored data for the Table 1 data
- N2=11, number of left-censored data
- inspect is the time of the inspection (sampling)
- start is the unobserved start time
- et is the unobserved elapsed time
- resp is the response or quantity Y
- ra is β_0
- rb is β_1
- sigmaResp is σ_t
- mu is μ
- sigma is σ

```
model
{
for( i in 1 : N1 ) {
startStar[i] ~ dinterval(start[i],inspect[i])
start[i] ~ dlnorm(mu,tau) # second parameter is a precision, \textit{i.e.}, reciprocal varianc
}
for( i in (N1+1) : (N1+N2) ) {
startStar[i] ~ dinterval(start[i],inspect[i])
start[i] ~ dlnorm(mu,tau)
}
for( i in (N1+1) : (N1+N2) ) {
resp[i] ~ dnorm(muResp[i],tauResp) # second parameter is a precision
muResp[i]<- ra*(1-exp(-rb*et[i]))
et[i]<-inspect[i]-start[i]
}

#priors
ra~dlnorm(7,(1/(.5*,5)))
rb~dlnorm(-4,1)
tauResp <- 1/(sigmaResp*sigmaResp)
sigmaResp ~ dnorm(0,1.0E-1)I(0,)
mu ~ dnorm(0.0,1.0E-1)I(0,)
tau <- 1/(sigma*sigma)
sigma ~ dnorm(0,1.0E-1)I(0,)
}
```