

Scientific Inquiry and Review (SIR)

Volume 7 Issue 3, 2023

ISSN (P): 2521-2427, ISSN (E): 2521-2435

Homepage: <https://journals.umt.edu.pk/index.php/SIR>



Article QR



Title: Bivariate and Multivariate Data Cloning through Non Linear Regression Models

Author (s): Sajid Hussain¹, Zafar Iqbal¹, Muhammad Mansoor², Rashid Ahmed¹

Affiliation (s): ¹The Islamia University of Bahawalpur, Pakistan
²Government S. E. College, Bahawalpur, Pakistan

DOI: <https://doi.org/10.32350/sir.73.01>

History: Received: December 4, 2022, Revised: March 28, 2023, Accepted: March 30, 2022,
Published: August 28, 2023

Citation: Hussain S, Iqbal Z, Mansoor M, Ahmed R. Bivariate and multivariate data cloning through nonlinear regression models. *Sci Inq Rev.* 2023;7(3):1–21.
<https://doi.org/10.32350/sir.73.01>

Copyright: © The Authors

Licensing:  This article is open access and is distributed under the terms of [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

Conflict of Interest: Author(s) declared no conflict of interest



A publication of
The School of Science
University of Management and Technology, Lahore, Pakistan

Bivariate and Multivariate Data Cloning through Non Linear Regression Models

Sajid Hussain^{1*}, Zafar Iqbal¹, Muhammad Mansoor², and Rashid Ahmed¹

¹Department of Statistics, The Islamia University of Bahawalpur, Pakistan

²Department of Statistics, Government S. E. College, Bahawalpur, Pakistan

ABSTRACT

Nonlinear regression analysis holds significant popularity in mathematical, engineering, and social science domains. Disciplines like financial matters, biology, and natural chemistry have broadly utilized nonlinear regression models (NLRMs). Cloned datasets have their own importance in such areas which provide the same fit of bivariate and multivariate nonlinear regression models for the actual datasets. This article presents a sequence of cloned datasets that give exactly the same fit of bivariate and multivariate nonlinear regression models.

Keywords: cloned data, nonlinear regression model, fictitious datasets, data visualization.

1. INTRODUCTION

If genuine information is private and cannot be shown, a matching or alternative set of data is required which provide same summary statistics as of the actual data. Cloned data refers to the alternative or matching set of data through mathematical techniques that allow rapid provisioning in testing and developments. Data cloning has its own significance as an alternative method for protecting confidential information and database. Table 1 shows four fictitious distinct cloned datasets (CDSs) created by Anscombe [1] to demonstrate the significance of graphs in statistical analysis. The summary statistics (mean, standard deviation, and correlation) as well as the parameter estimates of the fitted regression equation R^2 and estimated standard deviation of residuals are identical across these four distinct CDSs, however, they were vastly different scatter plots as shown in Figure 1. Dataset I was strongly linear with a single outlier and II appears to follow a parabolic distribution, whereas dataset III appears to adhere to a noisy linear regression model (LRM), and dataset IV appears to follow a vertical line with the regression thrown off by a single outlier. Datasets in Table 1 are significant and frequently used to show how important visible

* Corresponding Author: sajidhussain060917@gmail.com

methods are. These datasets were also known for their significant use in education. However, the method used to create the datasets was not explained in [1]. A genetic algorithm-based approach was proposed by Chatterjee and Firat [2], who generated 1,000 random datasets with comparable summary statistics and graphics for the basic datasets. Govindaraju and Haslett [3] devised a method for producing datasets by regressing the response on the covariate in the direction of their unconditional sample means, while maintaining identical LRM estimates. As a result, the variability in the response and the covariate decreased in each subsequent cloned dataset. Haslett and Govindaraju's [4] method for creating matched datasets was extended to include a multiple linear regression model, ensuring that the matched datasets have an identical fit to the original data. The idea of data-cloning emerged from both biostatistics [5, 6] and financial time series [7].

Cloning for maximum likelihood estimation using Bayesian software was achieved by the simple device of replicating the original data many times [6]. Fung et al. [8] expressed that the creation of CDSs to anonymize sensitive data was another application for datasets with the same statistical properties, as discussed in [3]. In this instance, it is critical that individual data points were altered, while the data's overall structure remained unchanged.

Haslett and Govindaraju [9] described a straightforward approach for modifying LRM data, while still obtaining the same fitted regression parameters. Ponciano et al. [10] showed how structural parameter non-identifiability can be diagnosed with Data Cloning (DC) and distinguished from other parameter estimability issues, such as when parameters are structurally identifiable but not estimable in a given data set or when they are identifiable and weakly estimable. Bayesian phylogenetics software can be used to diagnose non-identifiability with the DC approach. Additionally, it was demonstrated that DC can be used to examine and eliminate the influence of priors, particularly when prior elicitation was difficult. Finally, DC can be used to investigate at least two significant statistical issues when applied to phylogenetic inference, developing effective sampling strategies for computationally expensive posterior densities, and evaluating the identifiability of discrete parameters, such as the tree's topology.

Data confidentiality is one of the designed goals of tunable encrypted deduplication, see Amvrosiadis and Bhadkamkar [11]. Additionally, it

reduced the risk of data leakage brought by frequency analysis. Furthermore, it was identified that better ways of seeing and exploring data lead to better insights. The "Datasaurus" Cairo dataset was created by Alberto Cairo [12]. This, like Anscombe's Quartet, emphasized the significance of data visualization, despite the dataset's normal summary statistics, the plot it produced depicted a dinosaur. They started with the datasaurus and created additional datasets with the same summary statistics. Additionally, Cairo's Datasaurus data visualization prohibited to solely rely on the summary statistics of the used data.

Resultantly, according to [2], datasets should be as graphically distinct as possible. With different standard deviations but identical means and LRM estimates, [3, 4, 9] data are intended to be graphically comparable. Matejka and Fitzmaurice [13] developed a novel method for creating datasets, which are identical across a variety of statistical properties but visually distinct during the data exploration. To address the primary empirical facts of financial time series, numerous complex parametric stochastic volatility models were proposed in the subsequent literature. The models that Mao et al. [14, 15] proposed incorporated a broader asymmetric volatility function.

Hussain et al. [16] used a simple procedure to clone data for nonlinear regression models having linearizable or nonlinearize regression functions, such as aX^b , ab^X , ae^{bX} , ka^{b^X} , $ks^X b^{c^X}$, $k+ab^X$, $\frac{k}{1+bc^X}$, $A[aX_2^{-b} + (1 - a)X_1^{-b}]^{-\frac{1}{b}}$. They found that cloned data generated by linearizable or non-linearizable estimable functions of parameters have unchanged estimates. The procedure increased the sample size of cloned data without changing the parameters estimates, which was for n original sample points (x, y) . This generated n^2 observations by adding $[a_i: i = 1, 2, \dots, n]$ to the data points y over $\sum_i a_i = 0$. Due to increased sample size, cloned estimates showed smaller standard errors as compared to the original standard errors. This procedure used by [16] was sufficient for the first iteration because in the next iterations, it became tedious work. This procedure was useful for modeling but not for confidentialising or encrypting data, as in the design matrix variables remained unchanged. In this case, the term "confidentializing" referred to making the values of particular variables certain for particular people that cannot be deduced from the data. Our goal in this article is to create datasets with the same fit for nonlinear linear

regression models (NLRMs). [3, 4] methods were used to generate these cloned data sets. To get around the problem in [16], nonlinear regression models with linearizable regression functions were the prime focus of this article.

Table 1. Anscombe's CDSs with Pairs (x, y_1) , (x, y_2) , (x, y_3) and (x_4, y_4)

| x | y_1 | y_2 | y_3 | x_4 | y_4 |
|-----|-------|-------|-------|-------|-------|
| 7 | 6.42 | 7.26 | 4.82 | 8 | 7.91 |
| 4 | 5.39 | 3.10 | 4.26 | 19 | 12.5 |
| 11 | 7.81 | 9.26 | 8.33 | 8 | 8.47 |
| 13 | 12.74 | 8.74 | 7.58 | 8 | 7.71 |
| 8 | 6.77 | 8.14 | 6.95 | 8 | 5.76 |
| 5 | 5.73 | 4.74 | 5.68 | 8 | 6.89 |
| 12 | 8.15 | 9.13 | 10.84 | 8 | 5.56 |
| 6 | 6.08 | 6.13 | 7.24 | 8 | 5.25 |
| 10 | 7.46 | 9.14 | 8.04 | 8 | 6.58 |
| 9 | 7.11 | 8.77 | 8.81 | 8 | 8.84 |
| 14 | 8.84 | 8.10 | 9.96 | 8 | 7.04 |

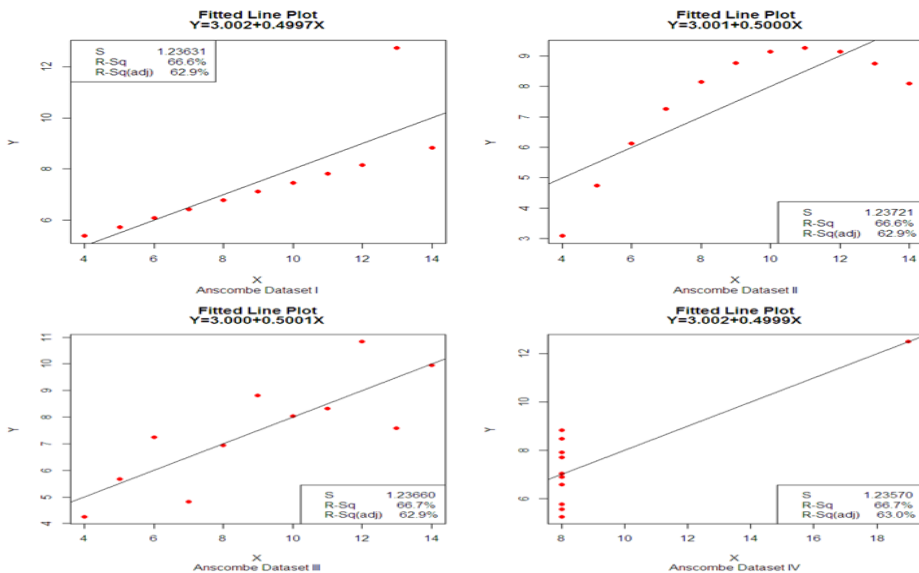


Figure 1. Scatter Plots of Anscombe's Datasets with Matched Simple Regression Models

2. THE NONLINEAR REGRESSION MODEL

2.1. The Regression Model (RM)

The Regression Model (RM) talks about the relationship between a response variable Y and one or more covariates $X^{(j)}$. The general model is:

$$Y_i = H\left(X_i^{(1)}, X_i^{(2)}, \dots, X_i^{(m)}; A_1, A_2, \dots, A_p\right) + E_i$$

Here, H is a suitable function that relies on the covariates $X_i^{(1)}, X_i^{(2)}, \dots, X_i^{(m)}$ and parameters A_1, A_2, \dots, A_p . The unstructured deviations from H are defined by means of random errors (REs) $E_i \sim N(0, \sigma^2)$.

2.2. The Linear Regression Model

In multiple LRM (MLRM), function H are characterized as linear in the parameters.

$$H\left(X_i^{(1)}, X_i^{(2)}, \dots, X_i^{(m)}; A_1, A_2, \dots, A_p\right) = A_1 \tilde{X}_i^{(1)} + A_2 \tilde{X}_i^{(2)} + \dots + A_p \tilde{X}_i^{(m)}$$

where $\tilde{X}_i^{(j)}$ can be arbitrary functions of the original covariates $X_i^{(j)}$.

2.3. The Nonlinear Regression Model

In NLRM, function H is regarded in such a way that it can't be written as linear in parameters. In case, there are infinite ways to explain the deterministic part of the model.

2.4. Linearizable Regression Functions (LRFs)

In NLRMs, functions H can be linearized by the transformation of the variable of interest and the explanatory variables. Therefore, the regression is named as function H which is linearizable if it can be converted into a function linear in the parameters.

2.5. A Few Examples of Nonlinear Regression Functions

- 1- $H(X_i; A_1, A_2) = A_1 X_i^{A_2}$
- 2- $H(X_i; A_1, A_2) = A_1 A_2^{X_i}$
- 3- $H(X_i; A_1, A_2) = A_1 e^{A_2 X_i}$
- 4- $H\left(X_i^{(1)}, X_i^{(2)}; A_1, A_2, A_3\right) = A_1 (X_i^{(1)})^{A_2} (X_i^{(2)})^{A_3}$

2.6. Linearizable Regression Function Model

A LRM with the LRF in the referred example is based on the model given below:

$$\ln(Y_i) = B_1 + B_2\tilde{X}_i^{(1)} + B_3\tilde{X}_i^{(2)} + E_i; \tilde{X}_i^{(1)} = \ln(X_i^{(1)}), \tilde{X}_i^{(2)} = \ln(X_i^{(2)})$$

Where, E_i follows the normal distribution. This model was back-converted and for this reason, the following equation was obtained:

$$Y_i = A_1(X_i^{(1)})^{A_2}(X_i^{(2)})^{A_3}\tilde{E}_i; \tilde{E}_i = \exp(E_i), i = 1, 2, \dots, n$$

The errors \tilde{E}_i follows lognormal distributed and contributed multiplicatively. The assumptions about the random deviations were accordingly now appreciably distinct for a model, which was primarily based on:

$$Y_i = A_1(X_i^{(1)})^{A_2}(X_i^{(2)})^{A_3} + E_i^*$$

with random deviations E_i^* that follows normal distribution and contributed additively.

3. DATA CLONING BY USING REGRESSING Y ON X AND X ON Y

Assuming n paired observations of X and Y say $(x_i, y_i) i = 1, 2, \dots, n$. The following procedure from [3] would generate a sequence of CDSs by obtaining the same fitted NLRM equations.

3.1. Procedure for Bivariate Nonlinear Regression Model $Y = AX^B$

The simple NLRM (a geometric or power curve) $Y = AX^B$ was linearizable due to logarithmic transformation as $\tilde{Y} = a + B\tilde{X}$ where $\tilde{Y} = \ln(Y)$, $\tilde{X} = \ln(X)$, $a = \ln(A)$, and $A = \exp(a)$. The inverse nonlinear regression model (INLRM) of $Y = AX^B$ is $(\frac{Y}{C})^{\frac{1}{D}}$, which was also linearizable as $\tilde{X} = c + d\tilde{Y}$ where $d = \frac{1}{D}$, $D = \frac{1}{a}$, $c = -\frac{\ln(C)}{D}$, $C = \exp(-\frac{c}{d})$.

1. First fit regression \tilde{Y} on \tilde{X} , namely $\tilde{Y}_1 = a + B\tilde{X}$. Also fit inverse regression (IR) \tilde{X} on \tilde{Y} , namely $\tilde{X}_1 = c + d\tilde{Y}$.
2. The regression of \tilde{Y}_1 on \tilde{X}_1 would be $\tilde{Y}_2 = a + B\tilde{X}_1$, preserving same parameter estimates. Likewise, $\tilde{X}_2 = c + d\tilde{Y}_1$. Note that $S_{\tilde{Y}_1}^2 < S_{\tilde{Y}}^2$ and $S_{\tilde{X}_1}^2 < S_{\tilde{X}}^2$.

3. The above method can be iterated with \tilde{Y}_2 and \tilde{X}_2 as done in step 1 to gain cloning sets of data having the identical linearizable regression equation (LRE). Again $S_{\tilde{Y}_2}^2 < S_{\tilde{Y}_1}^2 < S_Y^2$ and $S_{\tilde{X}_2}^2 < S_{\tilde{X}_1}^2 < S_X^2$ and so on.
4. If preferred, transform back for CDSs, having same coefficients of NLRM. It was noted that variability in Y and X of the cloned datasets fluctuated after every iteration (see Table 2).

Example 1. Consider the variables $X = (0.5, 1.5, 2.5, 5.0, 10.0)^T$ and $Y = (3.4, 7.0, 12.8, 29.8, 68.2)^T$, resulting in the nonlinear regression fit

$$\hat{Y} = 5.709057X^{1.01876} \tag{3.1}$$

Steps 1-4 given above would yield the CDSs proven in Table 2. with exactly the same equation of fitted NLRM as in (Eq. 3.1).

Table 2. Cloned Data Sets Having the same Non Linear Regression Fit $Y = AX^B$

| | Raw data | | First iteration | | Second iteration | |
|-------------|-----------------|----------------|------------------|----------------|------------------|----------------|
| | X | Y | X ₁ | Y ₁ | X ₂ | Y ₂ |
| | 0.5 | 3.4 | 0.61884 | 2.81765 | 0.51656 | 3.50135 |
| | 1.5 | 7.0 | 1.23925 | 8.62897 | 1.51542 | 7.10350 |
| | 2.5 | 12.8 | 2.21417 | 14.52011 | 2.49958 | 12.83071 |
| | 5.0 | 29.8 | 4.99031 | 29.42030 | 4.92915 | 29.36222 |
| | 10.0 | 68.2 | 11.06349 | 59.61073 | 9.72023 | 66.07545 |
| Mean | 3.9 | 24.24 | 4.02521 | 22.99955 | 3.83619 | 23.77465 |
| Variances | 14.425 | 706.448 | 18.2784 | 516.8321 | 13.50223 | 657.32050 |
| Correlation | - | 0.99684 | - | 0.99803 | - | 0.99679 |
| | Third iteration | | Fourth iteration | | Fifth iteration | |
| | X ₃ | Y ₃ | X ₄ | Y ₄ | X ₅ | Y ₅ |
| | 0.63657 | 2.91275 | 0.53332 | 3.60356 | 0.65443 | 3.00903 |
| | 1.25687 | 8.71936 | 1.53068 | 7.20638 | 1.27437 | 8.80883 |
| | 2.21927 | 14.51765 | 2.49918 | 12.86087 | 2.22429 | 14.51524 |
| | 4.91979 | 28.99566 | 4.86072 | 28.93958 | 4.85168 | 28.58561 |
| | 10.73187 | 57.91217 | 9.45374 | 64.05832 | 10.41665 | 56.29510 |
| Mean | 3.952877 | 22.61152 | 3.77553 | 23.33374 | 3.88428 | 22.24276 |
| Variances | 17.04036 | 483.35930 | 12.65003 | 612.24240 | 15.9017 | 452.4788 |
| Correlation | - | 0.99797 | - | 0.99672 | - | 0.99791 |

3.2. Procedure for Bivariate Nonlinear Regression Model $Y = AB^X$

A simple nonlinear regression model (an exponential curve) $Y = AB^X$ was linearizable due to logarithmic transformation as $\tilde{Y} = a + bX$ where

$\tilde{Y} = \ln(Y)$, $a = \ln(A)$, $A = \exp(a)$, $b = \ln(B)$, and $B = \exp(b)$. The inverse nonlinear regression model of $Y = AB^X$ was $X = \ln\left(\frac{Y}{C}\right)^{\frac{1}{\ln(D)}}$, which was also linearizable, as $X = c + d\tilde{Y}$ where $d = \frac{1}{\ln(D)}$, $D = \exp\left(\frac{1}{d}\right)$, $c = -\frac{\ln(C)}{\ln(D)}$, $C = \exp\left(-\frac{c}{d}\right)$.

1. First fit the regression of \tilde{Y} on X specifically $\tilde{Y}_1 = a + bX$. Also, fit the IR of X on \tilde{Y} particularly $X_1 = c + d\tilde{Y}$.
2. The regression of \tilde{Y}_1 on X_1 would be $\tilde{Y}_2 = a + bX_1$, maintaining the identical parameter estimates. Similarly, $X_2 = c + d\tilde{Y}_1$. Note that $S_{\tilde{Y}_1}^2 < S_{\tilde{Y}}^2$ and $S_{X_1}^2 < S_X^2$.
3. The above procedure can be iterated with \tilde{Y}_2 and X_2 as in step 1 to obtain CDSs having the same LRE. Again $S_{\tilde{Y}_2}^2 < S_{\tilde{Y}_1}^2 < S_{\tilde{Y}}^2$ and $S_{X_2}^2 < S_{X_1}^2 < S_X^2$ and so on.
4. If preferred, convert back to produce a sequence of CDSs, all with the same NLRM coefficients. Therefore, it was observed that variability in Y of the cloned datasets fluctuated after each and every generation (see Table 3).

Example 2. Consider the variables $X = (0, 1, 2, 3, 4, 5, 6, 7, 8)^T$ and $Y = (0.75, 1.20, 1.75, 2.50, 3.45, 4.70, 6.20, 8.25, 11.50)^T$ and resulting in the nonlinear regression fit

$$\hat{Y} = (0.8573324)(1.392474)^X \tag{3.2}$$

Steps 1-4 described above would generate the CDSs presented by Table 3 having exactly same NLRM fitted equation as in (Eq. 3.2).

Table 3. Cloned Data Sets Having the same Nonlinear Regression Fit $Y = AB^X$

| Raw data | | First iteration | | Second iteration | |
|----------|------|-----------------|----------------|------------------|----------------|
| X | Y | X ₁ | Y ₁ | X ₂ | Y ₂ |
| 0 | 0.75 | -0.3781 | 0.8573 | 0.0235 | 0.7565 |
| 1 | 1.20 | 1.0332 | 1.1938 | 1.0176 | 1.2070 |
| 2 | 1.75 | 2.1660 | 1.6624 | 2.0118 | 1.7563 |
| 3 | 2.50 | 3.2370 | 2.3148 | 3.0059 | 2.5037 |
| 4 | 3.45 | 4.2041 | 3.2233 | 4.0000 | 3.4486 |
| 5 | 4.70 | 5.1325 | 4.4883 | 4.9941 | 4.6896 |
| 6 | 6.20 | 5.9642 | 6.2499 | 5.9882 | 6.1762 |

| | Raw data | | First iteration | | Second iteration | |
|-------------|-----------------|----------------|------------------|----------------|------------------|----------------|
| | 7 | 8.25 | 6.8219 | 8.7028 | 6.9824 | 8.2045 |
| | 8 | 11.50 | 7.8192 | 12.1184 | 7.9765 | 11.4143 |
| Mean | 4 | 4.47778 | 4 | 4.53456 | 4 | 4.46187 |
| Variances | 7.5 | 12.95069 | 7.45591 | 14.67662 | 7.41209 | 12.73009 |
| Correlation | - | 0.95442 | - | 0.91709 | - | 0.95498 |
| | Third iteration | | Fourth iteration | | Fifth iteration | |
| | X ₃ | Y ₃ | X ₄ | Y ₄ | X ₅ | Y ₅ |
| | -0.3524 | 0.8640 | 0.0469 | 0.7629 | -0.3268 | 0.8707 |
| | 1.0506 | 1.2008 | 1.0352 | 1.2140 | 1.0679 | 1.2078 |
| | 2.1768 | 1.6688 | 2.0234 | 1.7626 | 2.1875 | 1.6753 |
| | 3.2415 | 2.3193 | 3.0117 | 2.5075 | 3.2459 | 2.3238 |
| | 4.2029 | 3.2233 | 4.0000 | 3.4473 | 4.2017 | 3.2233 |
| | 5.1258 | 4.4796 | 4.9883 | 4.6793 | 5.1192 | 4.4709 |
| | 5.9526 | 6.2256 | 5.9766 | 6.1526 | 5.9412 | 6.2016 |
| | 6.8053 | 8.6521 | 6.9648 | 8.1596 | 6.7888 | 8.6021 |
| | 7.7968 | 12.0245 | 7.9531 | 11.3298 | 7.7744 | 11.9318 |
| Mean | 4 | 4.51756 | 4 | 4.44617 | 4 | 4.50081 |
| Variances | 7.36852 | 14.41740 | 7.32521 | 12.51401 | 7.28215 | 14.16366 |
| Correlation | - | 0.91777 | - | 0.95553 | - | 0.91844 |

3.3. Procedure for Bivariate Nonlinear Regression Model $Y = Ae^{BX}$

The simple nonlinear regression model (an exponential curve) $Y = Ae^{BX}$ was linearizable due to logarithmic transformation as $\tilde{Y} = a + BX$, where $\tilde{Y} = \ln(Y)$, $a = \ln(A)$, and $A = \exp(a)$. The inverse nonlinear regression model of $Y = Ae^{BX}$ was $X = \ln\left(\frac{Y}{C}\right)^{\frac{1}{D}}$, which was also linearizable as $X = c + d\tilde{Y}$ where $d = \frac{1}{D}$, $D = \frac{1}{d}$, $c = -\frac{\ln(C)}{D}$, $C = \exp\left(-\frac{c}{d}\right)$.

1. First fit simple LRM of \tilde{Y} on X representing $\tilde{Y}_1 = a + BX$. Also, fit the simple inverse LRM of X on \tilde{Y} describing as $X_1 = c + d\tilde{Y}$.
2. The regression of \tilde{Y}_1 on X_1 would be $\tilde{Y}_2 = a + BX_1$, saving the alike parameter estimates. In the same way, $X_2 = c + d\tilde{Y}_1$. Note that $S_{\tilde{Y}_1}^2 < S_{\tilde{Y}}^2$ and $S_{X_1}^2 < S_X^2$.
3. The approach used above can be iterated with \tilde{Y}_2 and X_2 as in step 1 to get CDSs with the same LRE. Again $S_{\tilde{Y}_2}^2 < S_{\tilde{Y}_1}^2 < S_{\tilde{Y}}^2$ and $S_{X_2}^2 < S_{X_1}^2 < S_X^2$ and so on.

- If preferred, transform back to get a sequence of CDSs, all with the same NLRM coefficients. It was observed that variability in Y of the cloned datasets fluctuate after every iteration (see Table 4).

Example 3. Consider the variables $X = (0.5, 0.8, 1.4, 2.0, 2.5)^T$ and $Y = (9.1, 8.5, 7.5, 6.7, 6.1)^T$ and resulting in the nonlinear regression fit

$$\hat{Y} = 9.989049e^{-0.1991399X} \tag{3.3}$$

Here, the CDSs would be yielded as shown in Table 4, using steps 1-4 discussed earlier, to produce same equation of fitted NLRM as given in (Eq. 3.3).

Table 4. Cloned Data Sets Having the Same Nonlinear Regression Fit $Y = Ae^{BX}$

| | Raw data | | First iteration | | Second iteration | |
|-------------|-----------------|----------------|------------------|----------------|------------------|----------------|
| | X | Y | X ₁ | Y ₁ | X ₂ | Y ₂ |
| | 0.5 | 9.1 | 0.46924 | 9.04235 | 0.50111 | 9.09792 |
| | 0.8 | 8.5 | 0.81135 | 8.51797 | 0.80076 | 8.49874 |
| | 1.4 | 7.5 | 1.43912 | 7.55866 | 1.40005 | 7.50000 |
| | 2.0 | 6.7 | 2.00487 | 6.70739 | 1.99934 | 6.70089 |
| | 2.5 | 6.1 | 2.47543 | 6.07171 | 2.49875 | 6.10149 |
| Mean | 1.44 | 7.58 | 1.44 | 7.57961 | 1.44 | 7.57980 |
| Variances | 0.6830 | 1.532 | 0.68219 | 1.51378 | 0.68138 | 1.52834 |
| Correlation | - | -0.99617 | - | -0.99976 | - | -0.99618 |
| | Third iteration | | Fourth iteration | | Fifth iteration | |
| | X ₃ | Y ₃ | X ₄ | Y ₄ | X ₅ | Y ₅ |
| | 0.47039 | 9.04035 | 0.50222 | 9.09583 | 0.47153 | 9.03835 |
| | 0.81209 | 8.51668 | 0.80151 | 8.49748 | 0.81283 | 8.51540 |
| | 1.43912 | 7.55859 | 1.40009 | 7.50000 | 1.43912 | 7.55851 |
| | 2.00420 | 6.70827 | 1.99868 | 6.70179 | 2.00353 | 6.70916 |
| | 2.47420 | 6.07322 | 2.49749 | 6.10298 | 2.47298 | 6.07474 |
| Mean | 1.44 | 7.57942 | 1.44 | 7.57962 | 1.44 | 7.57923 |
| Variances | 0.68058 | 1.51018 | 0.67977 | 1.52469 | 0.67897 | 1.50657 |
| Correlation | - | -0.99976 | - | -0.99618 | - | -0.99976 |

We have generated the cloned data sets for following nonlinear regression models $Y = \frac{1}{A+BX}$, $Y = A + \frac{B}{1+X}$, $Y = A + B\sqrt{X}$, $Y = AX^2 + BX$ and $Y = A + BX + CX^2$ by using the procedure given by [3] and presented, respectively in Table 5-9.

Table 5. Cloned Data Sets Having the Same Non Linear Regression

Fit $Y = \frac{1}{A+BX}$; $A = 82.97359, B = -36.58871$

| Raw data | | First iteration | | Second iteration | |
|-----------------|----------------|------------------|----------------|------------------|----------------|
| X | Y | X ₁ | Y ₁ | X ₂ | Y ₂ |
| 2.000 | 0.0615 | 1.225281 | 0.102081 | 1.313470 | 0.026218 |
| 2.000 | 0.0527 | 1.188237 | 0.102081 | 1.313470 | 0.025318 |
| 0.667 | 0.0334 | 1.038642 | 0.017074 | 0.648053 | 0.022237 |
| 0.667 | 0.0258 | 0.918315 | 0.017074 | 0.648053 | 0.020254 |
| 0.400 | 0.0138 | 0.458482 | 0.014633 | 0.514769 | 0.015106 |
| 0.400 | 0.0258 | 0.918315 | 0.014633 | 0.514769 | 0.020254 |
| 0.286 | 0.0129 | 0.389507 | 0.013791 | 0.457862 | 0.014551 |
| 0.286 | 0.0183 | 0.701590 | 0.013791 | 0.457862 | 0.017451 |
| 0.222 | 0.0083 | -0.196641 | 0.013360 | 0.425914 | 0.011090 |
| 0.222 | 0.0169 | 0.639830 | 0.013360 | 0.425914 | 0.016789 |
| 0.200 | 0.0129 | 0.389507 | 0.013218 | 0.414932 | 0.014551 |
| 0.200 | 0.0087 | -0.121065 | 0.013218 | 0.414932 | 0.011441 |
| Third iteration | | Fourth iteration | | Fifth iteration | |
| X ₃ | Y ₃ | X ₄ | Y ₄ | X ₅ | Y ₅ |
| 0.926740 | 0.028641 | 0.970763 | 0.020381 | 0.777712 | 0.021073 |
| 0.908248 | 0.028641 | 0.970763 | 0.020104 | 0.768481 | 0.021073 |
| 0.833572 | 0.016874 | 0.638594 | 0.019057 | 0.731203 | 0.016776 |
| 0.773506 | 0.016874 | 0.638594 | 0.018291 | 0.701219 | 0.016776 |
| 0.543963 | 0.015591 | 0.572061 | 0.015855 | 0.586634 | 0.016118 |
| 0.773506 | 0.015591 | 0.572061 | 0.018291 | 0.701219 | 0.016118 |
| 0.509531 | 0.015101 | 0.543653 | 0.015545 | 0.569446 | 0.015852 |
| 0.665320 | 0.015101 | 0.543653 | 0.017056 | 0.647214 | 0.015852 |
| 0.216933 | 0.014839 | 0.527705 | 0.013327 | 0.423385 | 0.015707 |
| 0.634490 | 0.014839 | 0.527705 | 0.016734 | 0.631824 | 0.015707 |
| 0.509531 | 0.014751 | 0.522223 | 0.015545 | 0.569446 | 0.015658 |
| 0.254660 | 0.014751 | 0.522223 | 0.013577 | 0.442217 | 0.015658 |

Table 6. Cloned Data Sets Having the Same Non Linear Regression Fit $Y =$

$A + \frac{B}{1+X}$; $A = 0.086506, B = -0.091625$

| Raw data | | First iteration | | Second iteration | |
|----------|--------|-----------------|----------------|------------------|----------------|
| X | Y | X ₁ | Y ₁ | X ₂ | Y ₂ |
| 2.000 | 0.0615 | 2.332049 | 0.055964 | 1.805105 | 0.059008 |

| Raw data | | First iteration | | Second iteration | |
|-----------------|----------------|------------------|----------------|------------------|----------------|
| X | Y | X ₁ | Y ₁ | X ₂ | Y ₂ |
| 2.000 | 0.0527 | 1.565860 | 0.055964 | 1.805105 | 0.050796 |
| 0.667 | 0.0334 | 0.705670 | 0.031542 | 0.652333 | 0.032788 |
| 0.667 | 0.0258 | 0.506758 | 0.031542 | 0.652333 | 0.025696 |
| 0.400 | 0.0138 | 0.272456 | 0.021059 | 0.404582 | 0.014499 |
| 0.400 | 0.0258 | 0.506758 | 0.021059 | 0.404582 | 0.025696 |
| 0.286 | 0.0129 | 0.257787 | 0.015258 | 0.296953 | 0.013659 |
| 0.286 | 0.0183 | 0.351251 | 0.015258 | 0.296953 | 0.018698 |
| 0.222 | 0.0083 | 0.187800 | 0.011526 | 0.236035 | 0.009367 |
| 0.222 | 0.0169 | 0.325711 | 0.011526 | 0.236035 | 0.017392 |
| 0.200 | 0.0129 | 0.257787 | 0.010151 | 0.215011 | 0.013659 |
| 0.200 | 0.0087 | 0.193575 | 0.010151 | 0.215011 | 0.009740 |
| Third iteration | | Fourth iteration | | Fifth iteration | |
| X ₃ | Y ₃ | X ₄ | Y ₄ | X ₅ | Y ₅ |
| 2.072216 | 0.053842 | 1.644782 | 0.056682 | 1.863838 | 0.051862 |
| 1.444278 | 0.053842 | 1.644782 | 0.049020 | 1.340783 | 0.051862 |
| 0.687722 | 0.031054 | 0.638878 | 0.032217 | 0.671312 | 0.030599 |
| 0.504364 | 0.031054 | 0.638878 | 0.025600 | 0.502137 | 0.030599 |
| 0.284091 | 0.021273 | 0.408885 | 0.015152 | 0.295141 | 0.021472 |
| 0.504364 | 0.021273 | 0.408885 | 0.025600 | 0.502137 | 0.021472 |
| 0.270142 | 0.015859 | 0.307342 | 0.014368 | 0.281892 | 0.016421 |
| 0.358695 | 0.015859 | 0.307342 | 0.019070 | 0.365715 | 0.016421 |
| 0.203334 | 0.012377 | 0.249424 | 0.010363 | 0.218200 | 0.013172 |
| 0.334572 | 0.012377 | 0.249424 | 0.017851 | 0.342948 | 0.013172 |
| 0.270142 | 0.011095 | 0.229361 | 0.014368 | 0.281892 | 0.011975 |
| 0.208864 | 0.011095 | 0.229361 | 0.010711 | 0.223486 | 0.011975 |

Table 7. Cloned Data Sets Having the Same Non Linear Regression Fit $Y = A + B\sqrt{X}$; $A = -2.341445$, $B = 3.011197$

| Raw data | | First iteration | | Second iteration | |
|----------|-----|-----------------|----------------|------------------|----------------|
| X | Y | X ₁ | Y ₁ | X ₂ | Y ₂ |
| 1.0 | 1.1 | 1.389823 | 0.669752 | 1.099958 | 1.208477 |
| 1.5 | 1.3 | 1.536094 | 1.346503 | 1.571153 | 1.390611 |
| 2.0 | 1.6 | 1.769221 | 1.917031 | 2.033473 | 1.663810 |
| 2.5 | 2.0 | 2.105666 | 2.419675 | 2.490122 | 2.028076 |
| 3.0 | 2.7 | 2.764870 | 2.874101 | 2.942743 | 2.665542 |

| Raw data | | First iteration | | Second iteration | |
|-----------------|----------------|------------------|----------------|------------------|----------------|
| X | Y | X ₁ | Y ₁ | X ₂ | Y ₂ |
| 3.5 | 3.4 | 3.513706 | 3.291989 | 3.392311 | 3.303008 |
| 4.0 | 4.1 | 4.352175 | 3.680949 | 3.839463 | 3.940474 |
| Third iteration | | Fourth iteration | | Fifth iteration | |
| X ₃ | Y ₃ | X ₄ | Y ₄ | X ₅ | Y ₅ |
| 1.468250 | 0.816665 | 1.195129 | 1.307264 | 1.541545 | 0.950454 |
| 1.604771 | 1.432959 | 1.637384 | 1.473126 | 1.668620 | 1.511692 |
| 1.820930 | 1.952519 | 2.064198 | 1.721920 | 1.868668 | 1.984837 |
| 2.130380 | 2.410261 | 2.481144 | 2.053645 | 2.153013 | 2.401687 |
| 2.730323 | 2.824090 | 2.891081 | 2.634163 | 2.699051 | 2.778547 |
| 3.404599 | 3.204646 | 3.295707 | 3.214682 | 3.306735 | 3.125107 |
| 4.153208 | 3.558859 | 3.696127 | 3.795200 | 3.976065 | 3.447676 |

Table 8. Cloned Data Sets Having the Same Non Linear Regression Fit $Y = AX^2 + BX$; $A = 0.4, B = 5.0$

| Raw data | | First iteration | | Second iteration | |
|-----------------|----------------|------------------|----------------|------------------|----------------|
| X | Y | X ₁ | Y ₁ | X ₂ | Y ₂ |
| 0 | 1 | 0.1933 | 0.0000 | 0.0000 | 1.0107 |
| 1 | 5 | 0.9156 | 5.4708 | 0.9989 | 5.0388 |
| 2 | 12 | 2.0321 | 11.7428 | 1.9970 | 12.0451 |
| 3 | 20 | 3.1495 | 18.8161 | 2.9944 | 20.0064 |
| 4 | 25 | 3.7859 | 26.6907 | 3.9914 | 24.9641 |
| 5 | 36 | 5.0621 | 35.3665 | 4.9881 | 35.8353 |
| Third iteration | | Fourth iteration | | Fifth iteration | |
| X ₃ | Y ₃ | X ₄ | Y ₄ | X ₅ | Y ₅ |
| 0.1932 | 0.0000 | 0.0000 | 1.0211 | 0.1931 | 0.0000 |
| 0.9148 | 5.5123 | 0.9980 | 5.0763 | 0.9140 | 5.5523 |
| 2.0291 | 11.7893 | 1.9941 | 12.0884 | 2.0263 | 11.8338 |
| 3.1436 | 18.8311 | 2.9889 | 20.0114 | 3.1377 | 18.8445 |
| 3.7780 | 26.6377 | 3.9829 | 24.9279 | 3.7701 | 26.5848 |
| 5.0499 | 35.2093 | 4.9762 | 35.6734 | 5.0376 | 35.0546 |

Table 9. Cloned Data Sets Having the Same Non Linear Regression Fit $Y = A + BX + CX^2$; $A = 1.0, B = -0.20, C = 0.20$

| Raw data | | First iteration | | Second iteration | |
|----------|-----|-----------------|----------------|------------------|----------------|
| X | Y | X ₁ | Y ₁ | X ₂ | Y ₂ |
| 1.0 | 1.1 | 0.99760 | 1.08571 | 0.99972 | 1.09900 |

| | | | | | |
|-----------------|----------------|------------------|----------------|-----------------|----------------|
| 1.5 | 1.3 | 1.52005 | 1.29286 | 1.50226 | 1.30180 |
| 2.0 | 1.6 | 1.98929 | 1.62143 | 2.00206 | 1.60299 |
| 2.5 | 2.0 | 2.44462 | 2.07143 | 2.50109 | 2.00294 |
| 3.0 | 2.7 | 3.05159 | 2.64286 | 2.99979 | 2.70103 |
| 3.5 | 3.4 | 3.53894 | 3.33571 | 3.49833 | 3.39798 |
| 4.0 | 4.1 | 3.95791 | 4.15000 | 3.99676 | 4.09427 |
| Third iteration | | Fourth iteration | | Fifth iteration | |
| X ₃ | Y ₃ | X ₄ | Y ₄ | X ₅ | Y ₅ |
| 0.99730 | 1.08491 | 0.99945 | 1.09800 | 0.99700 | 1.08412 |
| 1.52239 | 1.29459 | 1.50450 | 1.30362 | 1.52472 | 1.29631 |
| 1.99134 | 1.62433 | 2.00410 | 1.60599 | 1.99338 | 1.62719 |
| 2.44578 | 2.07414 | 2.50217 | 2.00589 | 2.44694 | 2.07680 |
| 3.05120 | 2.64404 | 2.99959 | 2.70207 | 3.05081 | 2.64518 |
| 3.53715 | 3.33401 | 3.49667 | 3.39597 | 3.53537 | 3.33230 |
| 3.95486 | 4.14408 | 3.99355 | 4.08856 | 3.95181 | 4.13819 |

4. CLONING FOR MULTIPLE NONLINEAR REGRESSION MODEL $Y = A_0(X_i^{(1)})^{A_1}(X_i^{(2)})^{A_2}$ VIA PIVOTS

In accordance with [4], the current approach was extended to a structure of an arbitrary error covariance after discussing data that were independent and identically distributed (iid). Let us give the multiple NLRM in (Eq. 4.1).

$$Y_i = A_0(X_i^{(1)})^{A_1}(X_i^{(2)})^{A_2}E_i \quad (4.1)$$

Where Y is response vector, $X = (X^{(1)}:X^{(2)})$ is the covariate data matrix (CDM), $\alpha = (A_0, A_1, A_2)^T$ is parameters vector, and E random error vector.

Eq. 4.1 is linearizable due to logarithmic transformation, then Eq. 4.1 becomes $\ln(Y_i) = \ln(A_0) + A_1 \ln(X_i^{(1)}) + A_2 \ln(X_i^{(2)}) + \tilde{E}$. Setting $\tilde{Y} = \ln(Y_i)$, $\tilde{X}^{(1)} = \ln(X_i^{(1)})$, $\tilde{X}^{(2)} = \ln(X_i^{(2)})$, $\tilde{E} = \ln(E_i)$, $B_0 = \ln(A_0)$, $B_1 = A_1$ and $B_2 = A_2$ we get

$$\tilde{Y} = B_0 + B_1\tilde{X}^{(1)} + B_2\tilde{X}^{(2)} + \tilde{E} \quad (4.2)$$

where \tilde{Y} is response vector, $\tilde{X} = (\tilde{X}^{(1)}:\tilde{X}^{(2)})$ is CDM, $\beta = (B_0, B_1, B_2)^T$ is unknown parameters vector and \tilde{E} errors vector. When matrix \tilde{X} of rank full as column, estimates of β by ordinary least square (OLS) is $b = (\tilde{X}^t\tilde{X})^{-1}\tilde{X}^t\tilde{Y}$, and fitted multiple LRE is

$$\tilde{Y}_1 = b_0 + b_1\tilde{X}^{(1)} + b_2\tilde{X}^{(2)} \tag{4.3}$$

These consequences follow whether or not \tilde{Y} , $\tilde{X}^{(1)}$, and $\tilde{X}^{(2)}$ are mean corrected (MCtD), as here. Due to the MCtD, (Eq. 4.3), which can be written as:

$$\hat{y} = b_1x_1 + b_2x_2 \tag{4.4}$$

where $\hat{y} = \tilde{Y} - \bar{\tilde{Y}}$, $x_1 = \tilde{X}^{(1)} - \bar{\tilde{X}}^{(1)}$ and $x_2 = \tilde{X}^{(2)} - \bar{\tilde{X}}^{(2)}$ (In order to avoid any loss of generality, the column of 1 in design matrix X is eliminated following the imply MCtD because it transforms into a column of zeros).

The identified problem here was to create a new response variable vector, Y_{clone} , and a new covariate data matrix, X_{clone} . This can be easily accomplished by transposing back to \tilde{Y}_{clone} and \tilde{X}_{clone} , such that

$$b = (\tilde{X}_{clone}^t \tilde{X}_{clone})^{-1} \tilde{X}_{clone}^t \tilde{Y}_{clone}$$

Alternatively, multivariate CDSs to be required ($Y_{clone}, X_{clone}^{(1)}, X_{clone}^{(2)}$) which produced the same multiple NLRM equation as the original dataset ($Y, X^{(1)}, X^{(2)}$).

Returning to the case of iid, how generation of CDSs can be accomplished via manipulating any one covariate, was exhibited, say x_j , where $j = 1, 2$, using the steps below.

- 1) Initially, fit multiple linearizable RM (Eq. 4.4), using MCtD data.
- 2) Select a covariate x_2 .
- 3) Let $\hat{y} = k + b_2x_2$, where $k = b_1x_1 = \hat{y} - b_2x_2$. To obtain the estimated values of \hat{y}_k and \hat{x}_2 , perform simple RM of $y_k = y - k$ on x_2 and inverse simple regression of x_2 on y_k .
- 4) Regress x_1 on x_2 and obtain \hat{x}_1 . Also, obtain $x_{1,2} = x_1 - x_2(x_2^t x_2)^{-1} x_2^t x_1 = (I - x_2(x_2^t x_2)^{-1} x_2^t) x_1$, where I is the identity matrix.
- 5) Form $y_{k,clone} = \hat{y}_k + \sum b_1 x_{1,2}$.
- 6) On all newly acquired $x_{1,2}$ and \hat{x}_2 , perform multiple LREs of $y_{k,clone}$ at the same time, where $\hat{x}_2 = y(y^t y)^{-1} y^t x_2$ in which $y = (1: \hat{y})$ is $n \times 2$.

- 7) If you prefer, you can add back \bar{Y} , $\bar{X}^{(1)}$, and $\bar{X}^{(2)}$ to the cloned data or multiply all of the cloned data by the same scale factor.
- 8) Transform back to the cloned data in step 7.
- 9) Repeat from one to eight steps to create a series of datasets with identical NLRM coefficients. At each iteration, a different possible value of j can be chosen.

Example 4. With uncorrelated data and design matrix of full-rank, consider variables X_1 , X_2 and Y in Table 9 and resulting in the multiple NLRM fit

$$\hat{Y} = 1.663079X_1^{0.6163121}X_2^{0.2931787} \quad (4.5)$$

For CDSs in Table 10, 1-9 steps specified above were used (first X_2 was used for manipulation) for which the fitted multiple NLRM equation was exactly the same as in (Eq. 4.5).

Table 10. Cloned Data Sets Having the Same Multiple Nonlinear Regression Fit $Y = A_0X_1^{A_1}X_2^{A_2}$

| X_1 | X_2 | Y | $X_{1,clone}$ | $X_{2,clone}$ | Y_{clone} |
|-------|-------|-------|---------------|---------------|-------------|
| 23.81 | 11.33 | 22.76 | 25.583 | 10.257 | 24.988 |
| 75.83 | 25.92 | 76.73 | 37.324 | 35.179 | 40.197 |
| 9.46 | 7.03 | 8.62 | 15.945 | 3.818 | 16.234 |
| 5.71 | 29.68 | 10.98 | 2.473 | 14.871 | 7.852 |
| 85.78 | 21.81 | 86.77 | 49.689 | 39.486 | 45.583 |
| 0.37 | 0.57 | 0.97 | 6.672 | 2.411 | 4.543 |
| 8.82 | 11.25 | 11.82 | 9.540 | 9.225 | 13.577 |
| 8.99 | 19.01 | 16.63 | 5.928 | 20.744 | 11.810 |
| 37.65 | 75.25 | 67.40 | 6.780 | 73.887 | 19.203 |
| 8.43 | 8.40 | 8.81 | 12.012 | 4.796 | 14.364 |
| 16.10 | 30.30 | 21.54 | 6.839 | 16.206 | 14.786 |
| 0.64 | 1.20 | 1.34 | 5.718 | 2.327 | 5.138 |
| 5.28 | 6.93 | 12.38 | 9.021 | 22.489 | 11.379 |
| 30.40 | 70.18 | 58.37 | 5.847 | 71.736 | 17.173 |
| 33.66 | 21.06 | 29.90 | 20.152 | 11.872 | 25.870 |
| 15.72 | 11.86 | 14.54 | 16.178 | 6.394 | 19.093 |
| 8.44 | 14.53 | 17.54 | 7.172 | 26.020 | 12.274 |
| 30.20 | 34.20 | 29.43 | 11.444 | 13.458 | 21.041 |
| 8.89 | 8.68 | 11.41 | 12.282 | 8.357 | 14.702 |

| X_1 | X_2 | Y | $X_{1,clone}$ | $X_{2,clone}$ | Y_{clone} |
|-------|-------|------|---------------|---------------|-------------|
| 5.14 | 2.84 | 5.45 | 20.372 | 3.117 | 14.474 |

Figure 2, represent a matrix plot of raw and cloning data in Table 10, which show the effect on X_2 done by orthogonal manipulation as described in steps 1-9 of the algorithm. Bivariate relationship strength between $X_{2,clone}$ and Y_{clone} is much weaker than X_2 and Y . However, this is not the case with $X_{1,clone}$ and Y_{clone} , because the manipulation was not done with X_1 .

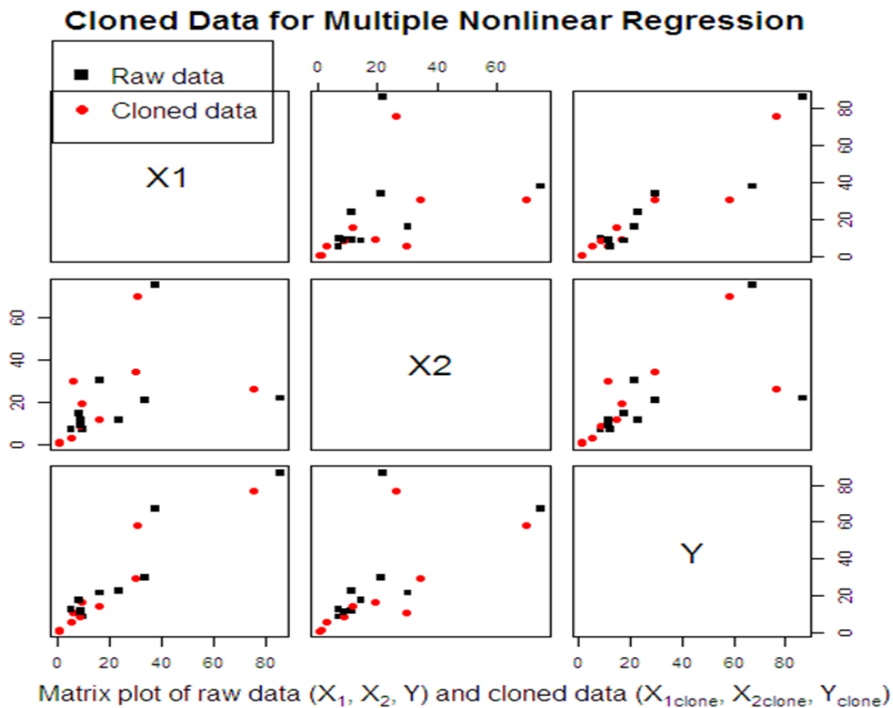


Figure 2. Matrix Plot of raw and cloned data

5. DISCUSSION

This study showed that the parameter estimates of the original datasets discussed in this article and their generated cloned datasets were identical. As a result, it was identified that data cloning had the potential to be used in a wide range of applications, including data encryption, visualization, and smoothing. The application of encryption was particularly intriguing because it can be used to generalize the databases even when regression modeling was not desired. In prior literature, cloned datasets were

generated for linear regression models. However, it had equal importance to be generated for the nonlinear regression models. In this context, new methods can be developed for nonlinear regression models to conduct cloning for the datasets or databases.

5.1. Conclusion

CDSs have been presented for bivariate and multivariate NLRMs that have linearizable regression functions including AX^B , AB^X , Ae^{BX} , $\frac{1}{A+BX}$, $A + \frac{B}{1+X}$, $A + B\sqrt{X}$, $AX^2 + BX$, $A + BX + CX^2$ and $A_0(X_i^{(1)})^{A_1}(X_i^{(2)})^{A_2}$ with exactly the same nonlinear regression coefficients. In terms of bivariate LRFs, the response and a covariate of the CDSs collapsed to their means, which had smaller variability when compared to the original dataset.

REFERENCES

1. Anscombe FJ. Graphs in statistical analysis. *Am Statistic*. 1973;27(1):17–21.
2. Chatterjee S, Firat A. Generating data with identical statistics but dissimilar graphics: A follow up to the Anscombe dataset. *Am Stat*. 2007;61(3):248–254. <https://doi.org/10.1198/000313007X220057>
3. Govindaraju K, Haslett SJ. Illustration of regression towards the mean. *Int J Matim Edu Sci Technol*. 2008;39(4):544–550. <https://doi.org/10.1080/00207390701753788>
4. Haslett SJ, Govindaraju K. Cloning data: Generating datasets with exactly the same multiple linear regression fit. *Aust New Zealand J Stat*. 2009;51(4):499–503. <https://doi.org/10.1111/j.1467-842X.2009.00560.x>
5. Lele SR, Dennis B, Lutscher F. Data Cloning: Easy maximum likelihood estimation for complex ecological models using bayesian markov chain monte carlo methods. *Ecol Lett*. 2007;10:551–563. <https://doi.org/10.1111/j.1461-0248.2007.01047.x>
6. Lele SR, Nadeem K, Schmuland B. Estimability and likelihood inference for generalized linear mixed models using data cloning. *J Am Stat Assoc*. 2010;105:1617–1625. <https://doi.org/10.1198/jasa.2010.tm09757>

7. Jacquier E, Johannes M, Polson N. MCMC Maximum likelihood for latent state models. *J Econom.* 2007;137:615–640. <https://doi.org/10.1016/j.jeconom.2005.11.017>
8. Fung BCM, Wang K, Chen R, Yu PS. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput Surv.* 2010;42(4):e14. <https://doi.org/10.1145/1749603.1749605>
9. Haslett SJ, Govindaraju K. Data cloning: Data visualization, smoothing, confidentiality, and encryption. *J Stat Plan Infer.* 2012;142:410–422. <https://doi.org/10.1016/j.jspi.2011.07.020>
10. Ponciano JM, Burleigh JG, Braun EL, Taper ML. Assessing parameter identifiability in phylogenetic models using data cloning. *Syst Biol.* 2012;61(6):955–972. <https://doi.org/10.1093/sysbio/sys055>
11. Amvrosiadis G, Bhadkamkar M. Identifying trends in enterprisedata protection systems. Paper presented at: USENIX Annual Technical Conference; July 8–10, 2015; Santa Clara, USA.
12. Download the Datasaurus: Never trust summary statistics alone; always visualize your data. Cairo website. <http://www.thefunctionalart.com/2016/08/downloaddatasaurus-never-trust-summary.html>
13. Matejka J, Fitzmaurice G. Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. Paper presented at: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems; May 6–11, 2017; Denver Colorado, USA. <https://doi.org/10.1145/3025453.3025912>
14. Mao X, Ruiz E, Veiga H. Threshold stochastic volatility: Properties and forecasting. *Int J Forecast.* 2017;33(4):1105–1123. <https://doi.org/10.1016/j.ijforecast.2017.07.001>
15. Mao X, Ruiz E, Veiga H, Czellar V. Asymmetric stochastic volatility models: Properties and particle filter-based simulated maximum likelihood estimation. *Econom Stat.* 2020;13:84–105. <https://doi.org/10.1016/j.ecosta.2019.08.002>
16. Hussain S, Daniyal M, Ogundokun RO, Muhammad YS, Iqbal Z, Ahmed, R. Cloning data with unchanged estimates of estimable non-

linear functions of parameters. *F1000Research*. 2022;10:e106.
<https://doi.org/10.12688/f1000research.28297.2>