Check for updates

BRIEF REPORT

Cloning data with unchanged estimates of estimable non-linear functions of parameters [version 1; peer review: 1 approved with reservations]

Sajid Hussain ¹, Muhammad Daniyal ¹, Roseline Oluwaseun Ogundokun ², Muhammad Yousaf Shad³, Zafar Iqbal¹, Rashid Ahmed¹

¹Department of Statistics, The Islamia University of Bahawalpur, Bahawalpur, 63100, Pakistan

²Department of Computer Science, Landmark University Omu Aran, Omu Aran, Kwara State, 251101, Nigeria

³Department of Statistics, Quaid-i-Azam University, Islamabad, 44000, Pakistan

V1 First published: 11 Feb 2021, 10:106
<https://doi.org/10.12688/f1000research.28297.1>

Latest published: 15 Mar 2022, 10:106
<https://doi.org/10.12688/f1000research.28297.2>

Abstract

Non-linear regression models occur in the fields of biology, banking, economics, and sociology, population and biological growth. The absolute growth, growth of humans, and most importantly, an economic variable is appropriately described by non-linear regression models. In this article, we present cloned datasets for bivariate and multivariate non-linear regression models with the same non-linear regression fit. The application of such cloned datasets is used for the confidentiality of sensitive data for publication purposes. In this article, we present cloned data sets which will yield the same fitted non-linear regression models.

Keywords

Cloned data, Linear regression, Non-linear regression, Residuals

Open Peer Review

Approval Status ?

1

version 2

(revision)
15 Mar 2022

version 1

11 Feb 2021

view

1. **David Medina-Ortiz**, Universidad de Chile,
Región Metropolitana, Chile

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the **Research on Research, Policy & Culture** gateway.

Corresponding author: Roseline Oluwaseun Ogundokun (ogundokun.roseline@lmu.edu.ng)

Author roles: **Hussain S:** Conceptualization, Formal Analysis, Methodology, Writing – Original Draft Preparation; **Daniyal M:** Data Curation, Investigation, Software, Writing – Original Draft Preparation; **Ogundokun RO:** Methodology, Supervision, Validation, Writing – Review & Editing; **Shad MY:** Formal Analysis, Investigation, Software; **Iqbal Z:** Data Curation, Methodology, Supervision; **Ahmed R:** Methodology, Resources, Visualization

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2021 Hussain S *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Hussain S, Daniyal M, Ogundokun RO *et al.* **Cloning data with unchanged estimates of estimable non-linear functions of parameters [version 1; peer review: 1 approved with reservations]** F1000Research 2021, 10:106 <https://doi.org/10.12688/f1000research.28297.1>

First published: 11 Feb 2021, 10:106 <https://doi.org/10.12688/f1000research.28297.1>

Introduction

In situations where the data is confidential and cannot be shown, there is a need for an alternative or matching set of data that can play the role of the actual data. The alternative or matching set of data is called cloned data. Therefore, cloned datasets can give a model-free way of representing confidential data. One possible use of these naturally cloned datasets is for the confidentiality of sensitive data for publication purposes, where having data sets with the same fit as the original data is the main advantage. [Anscombe \(1973\)](#) provided four cloned datasets to show the significance of graphs in a statistical study. All these cloned datasets have identical summary statistics (e.g., mean, variance, and correlation) but different data graphics (scatter plots). [Chatterjee and Firat \(2007\)](#) presented a technique of producing different (bivariate) datasets with the same summary statistics but dissimilar graphs by applying a genetic algorithm-based method. The idea of generating cloned data that has the same fit for simple and multiple linear regressions has been explained by Haslett and [Govindaraju \(2009, 2012\)](#) and [Govindaraju and Haslett \(2008\)](#). [Govindaraju and Haslett \(2008\)](#) gave the idea of cloning datasets by using the simple linear regression in the bivariate case. In all the cases, regression estimates are the same, and the variability decreases in the first to the next iteration. [Haslett and Govindaraju \(2009\)](#) explained the procedure for generating matching or cloning datasets for multivariate case.

The procedure by Haslett and Govindaraju gives a substitute way of presenting confidential data so that statistical analysis of multiple regression has the same fit in the original as well as in the cloned data. However, the data have been changed to be not any more confidential. The advantage is that parameter estimates from the cloned data and the original data do not consist of any model error. [Haslett and Govindaraju \(2012\)](#) regarded the issue of how to enhance the algorithm of producing several cloned datasets that will generate the same fitted regression equations. The primary method described that the fitted slope and intercept are merely estimates and that somewhat dissimilar datasets can still generate the same estimates. [Anscombe \(1973\)](#) showed using four different fictitious data sets (see [Table 1](#)) and showed that the regression estimates and their standard errors obtained are similar with their graphical significance in the literature (presented in [Figure 1](#)) but could not elaborate how such data have been obtained. [Chatterjee and Firat \(2007\)](#) used the data given in [Anscombe \(1973\)](#). They showed that all four datasets have identical summary statistics but different graphs by using the algorithm of the same datasets. [Govindaraju and Haslett \(2008\)](#) explained the procedure to generate cloned data for a simple linear regression $y_i = a + bx_i + e_i$ model as follows, assuming.

$$E(e_i) = 0, \text{var}(e_i) = \sigma^2, E(e_i e_j) = 0, E(X e_i) = 0$$

- 1) Consider n pairs of observations for X and Y, i.e., (x_i, y_i) , obtain $\hat{X}$ by simple regression of Y on X, also obtain by simple inverse regression.
- 2) The simple regression of $\hat{Y}$ on $\hat{X}$ i.e., $\hat{Y} = a + b\hat{X}$ has the same parameters.
- 3) Further, obtain $\hat{\hat{X}}$ by inverse regression and observe simple regression $\hat{\hat{Y}}$ on $\hat{\hat{X}}$ Has the same parameters.

Table 1. Anscombe's datasets with (X, Y1), (X, Y2), (X, Y3), and (X4, Y4), forming pairs.

X	Y ₁	Y ₂	Y ₃	X ₄	Y ₄
10	7.46	9.14	8.04	8	6.58
8	6.77	8.14	6.95	8	5.76
13	12.74	8.74	7.58	8	7.71
9	7.11	8.77	8.81	8	8.84
11	7.81	9.26	8.33	8	8.47
14	8.84	8.10	9.96	8	7.04
6	6.08	6.13	7.24	8	5.25
4	5.39	3.10	4.26	19	12.5
12	8.15	9.13	10.84	8	5.56
7	6.42	7.26	4.82	8	7.91
5	5.73	4.74	5.68	8	6.89

Figure 1. Scatter plots of Anscombe's datasets with cloned simple regression models

Iterating steps 1 & 2 can generate several fictitious or cloned datasets. They used the datasets given in [Anscombe \(1973\)](#) to generate several cloned datasets and showed that all cloned datasets giving the same regression estimates for first, second, third, fifth, and tenth iteration but different scatter plots. They also identify that $S_Y^2 > S_Y^2 > S_Y^2$ and $S_X^2 > S_X^2 > S_X^2$. The cloned datasets generated by four fictitious datasets given in [Anscombe \(1973\)](#) provided the same mean of X and Y, the correlation between X and Y, coefficient of determination R^2 , adjusted R^2 , regression fit, and standard error of the slope. But the variance of X and Y, standard error of residual and standard error of intercept decreases as the iterations increase. It shows regression towards the mean, i.e. every next cloned dataset is closer to the mean. [Haslett and Govindaraju \(2009\)](#) explained the procedure for generating cloned or matched datasets for a multivariate case that has the same fit. They consider identically independently distributed data for multiple regression models

$$Y = X\beta + \varepsilon$$

Where Y is the vector of responses, $X = (X_1, X_2, \dots, X_p)$ is the $n \times p$ design matrix, β is the unknown $p \times 1$ vector of parameters, and ε is the $n \times 1$ vector of errors. The OLS estimate $\hat{\beta}$ is

$$\hat{\beta} = (X'X)^{-1}X'Y$$

They used mean corrected form for response variable and independent variables $x_1, x_2, \dots, x_p$. Because of mean correction, the above multiple regression models can be written as

$$\hat{y} = b_1x_1 + b_2x_2 + \dots + b_px_p$$

They explained the procedure in six steps and generated $y_{new}, x_{1, new}, \dots, x_{p, new}$ which have the same fit as the original model. The cloned dataset generated by [Haslett and Govindaraju \(2009\)](#) gave the same regression fit, the sample mean of Y, X_1, X_2 but the variance of Y, X_1, X_2 , and residual standard error less than that of raw data. [Haslett and Govindaraju \(2012\)](#)

developed cloning algorithms for simple and multiple linear regression models. They fit the linear regression model of y_{new} on x (where x and y are mean-centred) on the original data and find its estimates and residuals. The residuals are added to data y one by one to create n^2 data points then fit the linear regression model of y_{new} and x to find its estimates, resulting in identical estimates for the original datasets and the cloned datasets. The above cloning algorithm can also be used in the multivariate case. In both cases, the parameter estimates of original datasets and cloned datasets are similar. They explained the following methods to generate cloned datasets

- Cloning via supplementing data by zero-mean additions-bivariate case
- Cloning via supplementing data by zero-mean additions-multivariate case
- Bivariate data cloning by regression y on x and x on y

Cloning for multiple regression via pivots

Here, we use the model presented in [Haslett and Govindaraju \(2012\)](#) to provide cloned datasets for bivariate and multivariate non-linear regression models with the same non-linear regression fit.

Methods

We consider the non-linear regression of y on X , where both X and y are non-mean-centered with data points. R software was used for all analysis.

Cloning for bivariate non-linear regression

In general, the non-linear regression model is

$$y = h(X; \beta) + \varepsilon \quad (2.1)$$

with y being the response variable, X the covariate data design matrix, which is often controlled by the researcher, β the model parameters characterizing the relationship between X and y through the regression function h , and ε the model errors that are assumed to be normally distributed with zero mean and unknown variances σ^2 .

When the regression function h is linear in the parameters β , it leads to linear regression analysis. However, linear models are not always appropriate, so one often needs to apply a non-linear regression model where h is non-linear in β .

Like in linear regression, non-linear regression provides parameter estimates based on the least square criterion. However, unlike linear regression, no explicit mathematical solution is available and specific algorithms are needed to solve the minimization problem, involving iterative numerical approximations. Here, since this is a bivariate non-linear regression on non-mean corrected data when $X = x$ a column vector. In general, provided X is of full row rank, the ordinary least square estimate of β is, of course, $\hat{\beta}$.

Now add the residuals $r = \hat{\varepsilon} = y - h(x; \hat{\beta})$ from the model fit the data, so that the original data are replicated as block n times to create an $n^2 \times 1$ vector and to each block, one of the residuals is added. The first block is $y + 1r_1$ where 1 is a vector of 1's and r_1 the first residual. The data are now $1 \otimes y + r \otimes 1$, and the design matrix becomes $1 \otimes x$. On noting that the model is still the same, i.e., a bivariate non-linear regression, if $1 \otimes y + r \otimes 1$ are now regressed on $1 \otimes x$, the OLS estimate becomes $\tilde{\beta}$ which is equal to $\hat{\beta}$. Thus, the non-linear regression estimates for cloned data are unchanged because the sum of the residuals $1^T r$ is zero. Software R has been used to obtain the numerical results.

Anything can be added, i.e., if $\{a_l : l = 1, 2, \dots, m\}$ is added to each data point in the set $\{y_i : i = 1, 2, \dots, n\}$ then the condition is that $\sum a_l = 0$. Some additions are more useful than others.

Example 1: The following cloned dataset ([Table 2](#)) is generated from the dataset $X = (1, 2, 3, 4, 5, 6)^T$ and $Y = (2.98, 4.26, 5.21, 6.10, 6.80, 7.50)^T$ for the nonlinear regression model $Y = aX^b$, a geometric or power curve. The parameter estimates for this cloned data set are summarized in [Table 2b](#), which can be suitable for the data used in different fields of life if our plotted data shows the form of model $Y = aX^b$. It can be observed that the estimates obtained by cloning procedure in [Table 2b](#) are some of the actual estimates.

Table 2. Cloned dataset having the same non-linear regression fit $Y = 2.974X^{0.5154}$.

X_{clone}	Y_{clone}	X_{clone}	Y_{clone}	X_{clone}	Y_{clone}
1	2.990626	1	3.003097	1	2.988703
2	4.270626	2	4.283097	2	4.268703
3	5.220626	3	5.233097	3	5.218703
4	6.110626	4	6.123097	4	6.108703
5	6.810626	5	6.823097	5	6.808703
6	7.510626	6	7.523097	6	7.508703
1	2.962378	1	2.950562	1	2.985865
2	4.242378	2	4.230562	2	4.265865
3	5.192378	3	5.180562	3	5.215865
4	6.082378	4	6.070562	4	6.105865
5	6.782378	5	6.770562	5	6.805865
6	7.482378	6	7.470562	6	7.505865

Table 2b. Parameter estimates of the raw and cloned dataset in Table 2.

	Estimates	Std. Error	Variables	Mean	Variance	RSE	Corr.
a	2.9740	0.015214	X	3.5	3.5	Y X	-
b	0.5154	0.003503	Y	5.475	2.80367	0.021987	0.9931
a_{clone}	2.9740	0.007380	X_{clone}	3.5	3	Y_{clone} X_{clone}	-
b_{clone}	0.5154	0.001699	Y_{clone}	5.475	2.40378	0.026125	0.9931

Example 2: The cloned dataset (Table 3) is generated from the dataset $X = (0, 1, 2, 3, 4, 5, 6, 7, 8)^T$ and $Y = (0.75, 1.20, 1.75, 2.50, 3.45, 4.70, 6.20, 8.25, 11.50)^T$ for the nonlinear regression model $= ab^X$, an exponential curve. If the sensitive observed data shows the exponential curve ($Y = ab^X$) such procedure can be useful for cloning of data. It can be observed that the estimates obtained by cloning procedure in Table 3b are similar to the actual estimates.

Example 3: The cloned dataset (Table 4) is generated from the dataset $X = (1, 2, 3, 4, 5, 6)^T$ and $Y = (1.6, 4.5, 13.8, 40.2, 125.0, 363.0)^T$ for the nonlinear regression model $Y = ae^{bX}$, an exponential curve. If the sensitive data shows the non-linear regression shape of $Y = ae^{bX}$, then such cloning procedure would be helpful. It can be observed that the estimates obtained by cloning procedure in Table 4b are equal to the actual estimates.

Example 4: The cloned dataset (Table 5) is generated from the dataset $X = (0, 1, 2, 3, 4, 5)^T$ and $Y = (58, 66, 72.5, 78, 82, 85)^T$ for the nonlinear regression model $Y = ka^{b^X}$, the Gompertz curve. Parameter estimates of the raw and cloned dataset is shown in Table 5b.

Example 5: The cloned dataset (Table 6) is generated from the dataset $X = (0.5, 0.5, 1, 1, 2, 2, 4, 4, 8, 8, 16, 16)^T$ and $Y = (0.96, 0.91, 0.86, 0.79, 0.63, 0.62, 0.48, 0.42, 0.17, 0.21, 0.03, 0.05)^T$ for the nonlinear regression model $= ks^X b^{c^X}$, the Makeham curve. The observed sensitive data shows the non-linear regression shape of makeham curve, then such cloning procedure would be beneficial as the estimates are closed. It can be observed that the estimates obtained by cloning procedure in Table 6b are some of the actual estimates.

Example 6: The cloned dataset (Table 7) is generated from the dataset $X = (0, 1, 2, 3, 4, 5, 6, 7, 8)^T$ and $Y = (0.75, 1.20, 1.75, 2.50, 3.45, 4.70, 6.20, 8.25, 11.50)^T$ for the nonlinear regression model $= k + ab^X$, a modified exponential curve. Sensitive data showing the pattern of modified exponential curve, procedure explained above with the help of table and their estimates would be beneficial. It can be observed that the estimates obtained by cloning procedure in Table 7b are equal to the actual estimates.

Example 7: The following cloned dataset (Table 8) is generated from the dataset $X = (0, 1, 2, 3, 4, 5, 6, 7, 8)^T$ and $Y = (1225, 2879, 4994, 11525, 16190, 22573, 30677, 38517, 39003)^T$ for the nonlinear regression model $Y = \frac{k}{1+bc^x}$, the

Table 3. Cloned dataset having the same non-linear regression fit $Y = ab^X$

X_{clone}	Y_{clone}	X_{clone}	Y_{clone}	X_{clone}	Y_{clone}
0	0.7898168	0	0.9124101	0	0.7033688
1	1.2398168	1	1.3624101	1	1.1533688
2	1.7898168	2	1.9124101	2	1.7033688
3	2.5398168	3	2.6624101	3	2.4533688
4	3.4898168	4	3.6124101	4	3.4033688
5	4.7398168	5	4.8624101	5	4.6533688
6	6.2398168	6	6.3624101	6	6.1533688
7	8.2898168	7	8.4124101	7	8.2033688
8	11.5398168	8	11.6624101	8	11.4533688
0	0.5847028	0	0.8680142	0	0.6307205
1	1.0347028	1	1.3180142	1	1.0807205
2	1.5847028	2	1.8680142	2	1.6307205
3	2.3347028	3	2.6180142	3	2.3807205
4	3.2847028	4	3.5680142	4	3.3307205
5	4.5347028	5	4.8180142	5	4.5807205
6	6.0347028	6	6.3180142	6	6.0807205
7	8.0847028	7	8.3680142	7	8.1307205
8	11.3347028	8	11.6180142	8	11.3807205
0	0.7705852	0	0.8032982	0	0.5312433
1	1.2205852	1	1.2532982	1	0.9812433
2	1.7705852	2	1.8032982	2	1.5312433
3	2.5205852	3	2.5532982	3	2.2812433
4	3.4705852	4	3.5032982	4	3.2312433
5	4.7205852	5	4.7532982	5	4.4812433
6	6.2205852	6	6.2532982	6	5.9812433
7	8.2705852	7	8.3032982	7	8.0312433
8	11.5205852	8	11.5532982	8	11.2812433

Table 3b. Parameter estimates of the raw and cloned dataset of **Table 3.**

	Estimates	Std. Error	Variables	Mean	Variance	RSE	Corr.
a	0.97	0.037728	X	4	7.50	Y X	-
b	1.36	0.007540	Y	4.5	12.95	0.139762	0.954
a_{clone}	0.96	0.015889	X_{clone}	4	6.75	Y_{clone} X_{clone}	-
b_{clone}	1.36	0.003210	Y_{clone}	4.5	11.67	0.177424	0.954

Logistic curve. If the curve of observed data is in the form of logistic, then **Table 8** procedure for cloning the data would be suitable. It can be observed that the estimates obtained by cloning procedure in **Table 8b** are identical as the actual estimates.

Cloning for multivariate non-linear regression

The algebra for the bivariate non-linear regression is unaltered for multivariate non-linear regression, except that the matrix X becomes $X_{n \times p} = [x_i^{(1)} : x_i^{(2)} : \dots : x_i^{(p)}]$, and the parameter vector and its estimates become $(p+1) \times 1$ vector, $\hat{\beta}, \tilde{\beta}$, and β .

Table 4. Cloned dataset having the same non-linear regression fit $Y = ae^{bx}$

X_{clone}	Y_{clone}	X_{clone}	Y_{clone}	X_{clone}	Y_{clone}
1	1.3412568	1	-0.219635	1	1.2394154
2	4.2412568	2	2.6803635	2	4.1394154
3	13.541257	3	11.980364	3	13.439415
4	39.941257	4	38.380364	4	39.839415
5	124.74126	5	123.18036	5	124.63942
6	362.74126	6	361.18036	6	362.63942
1	3.0524369	1	1.1087440	1	1.5468715
2	5.9524369	2	4.0087440	2	4.4468715
3	15.252437	3	13.308744	3	13.746872
4	41.652437	4	39.708744	4	40.146872
5	126.45244	5	124.50874	5	124.94687
6	364.45244	6	362.50874	6	362.94687

Table 4b. Parameter estimates of the raw and cloned dataset in Table 4.

	Estimates	Std. Error	Variables	Mean	Variance	RSE	Corr.
a	0.56	0.027971	X	3.5	3.5	Y X	-
b	1.08	0.008450	Y	91	19830.87	1.210555	0.8331
a_{clone}	0.55	0.0137135	X_{clone}	3.5	3	Y_{clone} X_{clone}	-
b_{clone}	1.08	0.004206	Y_{clone}	91	16998.83	1.469659	0.8331

Table 5. Cloned dataset having the same non-linear regression fit $Y = ka^{bx}$

X_{clone}	Y_{clone}	X_{clone}	Y_{clone}	X_{clone}	Y_{clone}
0	57.93007	0	58.13133	0	57.97689
1	65.93007	1	66.13133	1	65.97689
2	72.43007	2	72.63133	2	72.47689
3	77.93007	3	78.13133	3	77.97689
4	81.93007	4	82.13133	4	81.97689
5	84.93007	5	85.13133	5	84.97689
0	58.05172	0	57.87750	0	58.03288
1	66.05172	1	65.87750	1	66.03288
2	72.55172	2	72.37750	2	72.53288
3	78.05172	3	77.87750	3	78.03288
4	82.05172	4	81.87750	4	82.03288
5	85.05172	5	84.87750	5	85.03288

Example 8: The following cloned dataset (Table 9) is generated from the dataset $X_1 = (23.81, 75.83, 9.46, 5.71, 85.78, 0.37, 8.82, 8.99, 37.65)^T$, $X_2 = (11.33, 25.92, 7.03, 29.68, 21.81, 0.57, 11.25, 19.01, 75.25)^T$ and $Y = (22.76, 76.73, 8.62, 10.98, 86.77, 0.97, 11.82, 16.63, 67.40)^T$ for the nonlinear regression model $Y = A[aX_2^{-b} + (1-a)X_1^{-b}]^{\frac{1}{1-b}}$, the constant elasticity of substitution production function. Parameter estimates of the raw and cloned dataset is shown in Table 9b.

Table 5a. Parameter estimates of the raw and cloned dataset in **Table 5**.

	Estimates	Std. Error	Variables	Mean	Variance	RSE	Corr.
a	0.615222	0.002761	X	2.5	3.5	Y X	-
b	0.7321193	0.004796	Y	73.583	104.44	0.101776	0.9859
k	0.9422	0.466140	X_{clone}	2.5	3	Y_{clone} X_{clone}	-
a_{clone}	0.615222	0.001339	Y_{clone}	73.583	89.53	0.120928	0.9859
b_{clone}	0.7321193	0.002326					
k_{clone}	0.9422	0.226110					

Table 6. Cloned dataset having the same non-linear regression fit $Y = ks^Xb^{c^X}$

X_{clone}	Y_{clone}	X_{clone}	Y_{clone}	X_{clone}	Y_{clone}	X_{clone}	Y_{clone}
0.5	0.96779	0.5	0.93143	0.5	0.93547	0.5	1.00684
0.5	0.91779	0.5	0.88143	0.5	0.88547	0.5	0.95684
1.0	0.86779	1.0	0.83143	1.0	0.83547	1.0	0.90684
1.0	0.79779	1.0	0.76143	1.0	0.76547	1.0	0.83684
2.0	0.63779	2.0	0.60143	2.0	0.60547	2.0	0.67684
2.0	0.62779	2.0	0.59143	2.0	0.59547	2.0	0.66684
4.0	0.48779	4.0	0.45143	4.0	0.45547	4.0	0.52684
4.0	0.42779	4.0	0.39143	4.0	0.39547	4.0	0.46684
8.0	0.17779	8.0	0.14143	8.0	0.14547	8.0	0.21684
8.0	0.21779	8.0	0.18143	8.0	0.18547	8.0	0.25684
16.0	0.03779	16.0	0.00143	16.0	0.00547	16.0	0.07684
16.0	0.05779	16.0	0.02143	16.0	0.02547	16.0	0.09684
0.5	0.94779	0.5	0.94873	0.5	0.94547	0.5	0.93146
0.5	0.89779	0.5	0.89873	0.5	0.89547	0.5	0.88146
1.0	0.84779	1.0	0.84873	1.0	0.84547	1.0	0.83146
1.0	0.77779	1.0	0.77873	1.0	0.77547	1.0	0.76146
2.0	0.61779	2.0	0.61873	2.0	0.61547	2.0	0.60146
2.0	0.60779	2.0	0.60873	2.0	0.60547	2.0	0.59146
4.0	0.46779	4.0	0.46873	4.0	0.46547	4.0	0.45146
4.0	0.40779	4.0	0.40873	4.0	0.40547	4.0	0.39146
8.0	0.15779	8.0	0.15873	8.0	0.15547	8.0	0.14146
8.0	0.19779	8.0	0.19873	8.0	0.19547	8.0	0.18146
16.0	0.01779	16.0	0.01873	16.0	0.01547	16.0	0.00146
16.0	0.03779	16.0	0.03873	16.0	0.03547	16.0	0.02146
0.5	0.97143	0.5	1.00873	0.5	0.93684	0.5	0.98146
0.5	0.92143	0.5	0.95873	0.5	0.88684	0.5	0.93146
1.0	0.87143	1.0	0.90873	1.0	0.83684	1.0	0.88146
1.0	0.80143	1.0	0.83873	1.0	0.76684	1.0	0.81146
2.0	0.64143	2.0	0.67873	2.0	0.60684	2.0	0.65146
2.0	0.63143	2.0	0.66873	2.0	0.59684	2.0	0.64146
4.0	0.49143	4.0	0.52873	4.0	0.45684	4.0	0.50146
4.0	0.43143	4.0	0.46873	4.0	0.39684	4.0	0.44146

Table 6. *Continued*

X_{clone}	Y_{clone}	X_{clone}	Y_{clone}	X_{clone}	Y_{clone}	X_{clone}	Y_{clone}
8.0	0.18143	8.0	0.21873	8.0	0.14684	8.0	0.19146
8.0	0.22143	8.0	0.25873	8.0	0.18684	8.0	0.23146
16.0	0.04143	16.0	0.07873	16.0	0.00684	16.0	0.05146
16.0	0.06143	16.0	0.09873	16.0	0.02684	16.0	0.07146

Table 6b. Parameter estimates of the raw and cloned dataset in Table 6.

	Estimates	Std. Error	Variables	Mean	Variance	RSE	Corr.
b	1.206	0.128403	X	5.25	31.9773	Y X	-
c	0.29	0.350857	Y	0.510	0.1133	0.029115	-0.91
k	0.934	0.075154	X_{clone}	5.25	29.5175	$Y_{clone} X_{clone}$	-
s	0.824	0.014716	Y_{clone}	0.510	0.1053	0.037855	-0.91
b_{clone}	1.206	0.048284					
c_{clone}	0.29	0.131779					
k_{clone}	0.934	0.028093					
s_{clone}	0.824	0.005518					

Table 7. Cloned dataset having the same non-linear regression fit $Y = k + ab^X$

X_{clone}	Y_{clone}	X_{clone}	Y_{clone}	X_{clone}	Y_{clone}
0	0.86760	0	0.84130	0	0.75791
1	1.31760	1	1.29130	1	1.20791
2	1.86760	2	1.84130	2	1.75791
3	2.61760	3	2.59130	3	2.50791
4	3.56760	4	3.54130	4	3.45791
5	4.81760	5	4.79130	5	4.70791
6	6.31760	6	6.29130	6	6.20791
7	8.36760	7	8.34130	7	8.25791
8	11.61760	8	11.59130	8	11.50791
0	0.54429	0	0.82987	0	0.73207
1	0.99429	1	1.27987	1	1.18207
2	1.54429	2	1.82987	2	1.73207
3	2.29429	3	2.57987	3	2.48207
4	3.24429	4	3.52987	4	3.43207
5	4.49429	5	4.77987	5	4.68207
6	5.99429	6	6.27987	6	6.18207
7	8.04429	7	8.32987	7	8.23207
8	11.29429	8	11.57987	8	11.48207
0	0.69160	0	0.80976	0	0.67560
1	1.14160	1	1.25976	1	1.12560
2	1.69160	2	1.80976	2	1.67560
3	2.44160	3	2.55976	3	2.42560

Table 7. *Continued*

X_{clone}	Y_{clone}	X_{clone}	Y_{clone}	X_{clone}	Y_{clone}
4	3.39160	4	3.50976	4	3.37560
5	4.64160	5	4.75976	5	4.62560
6	6.14160	6	6.25976	6	6.12560
7	8.19160	7	8.30976	7	8.17560
8	11.44160	8	11.55976	8	11.42560

Table 7b. Parameter estimates of the raw and cloned dataset in Table 7.

	Estimates	Std. Error	Variables	Mean	Variance	RSE	Corr.
a	1.185529	0.127009	X	4	7.5	Y X	-
b	1.3319434	0.015977	Y	4.48	12.95	0.109390	0.954
k	-0.361129	0.194051	X_{clone}	4	6.75	$Y_{clone} X_{clone}$	-
a_{clone}	1.185529	0.053467	Y_{clone}	4.48	11.67	0.138149	0.954
b_{clone}	1.3319434	0.006726					
k_{clone}	-0.361129	0.081689					

Table 8. Cloned dataset having the same non-linear regression fit $Y = \frac{k}{1+bc^x}$

X_{clone}	Y_{clone}	X_{clone}	Y_{clone}	X_{clone}	Y_{clone}
0	1778.74	0	-632.17	0	991.62
1	3432.74	1	1021.83	1	2645.62
2	5547.74	2	3136.83	2	4760.62
3	12078.74	3	9667.83	3	11291.62
4	16743.74	4	14332.83	4	15956.62
5	23126.74	5	20715.83	5	22339.62
6	31230.74	6	28819.83	6	30443.62
7	39070.74	7	36659.83	7	38283.62
8	39556.74	8	37145.83	8	38769.62
0	3919.88	0	789.96	0	1506.03
1	5573.88	1	2443.96	1	3160.03
2	7688.88	2	4558.96	2	5275.03
3	14219.88	3	11089.96	3	11806.03
4	18884.88	4	15754.96	4	16471.03
5	25267.88	5	22137.96	5	22854.03
6	33371.88	6	30241.96	6	30958.03
7	41211.88	7	38081.96	7	38798.03
8	41697.88	8	38567.96	8	39284.03
0	686.48	0	2912.89	0	1204.81
1	2340.48	1	4566.89	1	2858.81
2	4455.48	2	6681.89	2	4973.81
3	10986.48	3	13212.89	3	11504.81
4	15651.48	4	17877.89	4	16169.81
5	22034.48	5	24260.89	5	22552.81
6	30138.48	6	32364.89	6	30656.81
7	37978.48	7	40204.89	7	38496.81
8	38464.48	8	40690.89	8	38982

Table 8b. Parameter estimates of the raw and cloned dataset in Table 8

	Estimates	Std. Error	Variables	Mean	Variance	RSE	Corr.
b	31.9624	8.38	X	4	7.5	Y X	-
c	0.46	0.032	Y	18620	220578958	1438.26	0.98
k	41044.63	1829.16	X_{clone}	4	6.75	Y_{clone} X_{clone}	-
b_{clone}	31.9624	3.72	Y_{clone}	18857	200093197	1837.82	0.98
c_{clone}	0.46	0.014					
k_{clone}	41044.63	767.30					

Table 9. Cloned dataset having the same non-linear regression fit $Y = A[aX_2^{-b} + (1-a)X_1^{-b}]^{\frac{1}{b}}$

X _{1,clone}	X _{2,clone}	Y _{clone}	X _{1,clone}	X _{2,clone}	Y _{clone}	X _{1,clone}	X _{2,clone}	Y _{clone}
23.81	11.33	18.93	23.81	11.33	18.70	23.81	11.33	21.61
75.83	25.92	72.90	75.83	25.92	72.67	75.83	25.92	75.58
9.46	7.03	4.79	9.46	7.03	4.56	9.46	7.03	7.47
5.71	29.68	7.15	5.71	29.68	6.92	5.71	29.68	9.83
85.78	21.81	82.94	85.78	21.81	82.71	85.78	21.81	85.62
0.37	0.57	-2.86	0.37	0.57	-3.09	0.37	0.57	-0.18
8.82	11.25	7.99	8.82	11.25	7.76	8.82	11.25	10.67
8.99	19.01	12.80	8.99	19.01	12.57	8.99	19.01	15.48
37.65	75.25	63.57	37.65	75.25	63.34	37.65	75.25	66.25
23.81	11.33	20.67	23.81	11.33	25.26	23.81	11.33	23.53
75.83	25.92	74.64	75.83	25.92	79.23	75.83	25.92	77.50
9.46	7.03	6.53	9.46	7.03	11.12	9.46	7.03	9.39
5.71	29.68	8.89	5.71	29.68	13.48	5.71	29.68	11.75
85.78	21.81	84.68	85.78	21.81	89.27	85.78	21.81	87.54
0.37	0.57	-1.12	0.37	0.57	3.47	0.37	0.57	1.74
8.82	11.25	9.73	8.82	11.25	14.32	8.82	11.25	12.59
8.99	19.01	14.54	8.99	19.01	19.13	8.99	19.01	17.40
37.65	75.25	65.31	37.65	75.25	69.90	37.65	75.25	68.17
23.81	11.33	19.56	23.81	11.33	23.15	23.81	11.33	25.17
75.83	25.92	73.53	75.83	25.92	77.12	75.83	25.92	79.14
9.46	7.03	5.42	9.46	7.03	9.01	9.46	7.03	11.03
5.71	29.68	7.78	5.71	29.68	11.37	5.71	29.68	13.39
85.78	21.81	83.57	85.78	21.81	87.16	85.78	21.81	89.18
0.37	0.57	-2.23	0.37	0.57	1.36	0.37	0.57	3.38
8.82	11.25	8.62	8.82	11.25	12.21	8.82	11.25	14.23
8.99	19.01	13.43	8.99	19.01	17.02	8.99	19.01	19.04
37.65	75.25	64.20	37.65	75.25	67.79	37.65	75.25	69.81

Table 9b. Parameter estimates of the raw and cloned dataset in Table 9.

	Estimates	Std. Error	Variables	Mean	Variance	RSE
A	1.36	0.064478	X₁	28.49111	1113.739	-
a	0.30	0.029150	X₂	22.42778	1008.499	-
b	-0.50	0.420536	Y	33.63111	478.7497	2.924394
A_{clone}	1.34	0.027909	X_{1,clone}	28.49111	1008.25	-
a_{clone}	0.30	0.012978	X_{2,clone}	22.42778	907.6489	-
b_{clone}	-0.50	0.185972	Y_{clone}	32.71486	430.8747	3.851240

Conclusions

In this article, we presented a cloned dataset for bivariate and multivariate non-linear regression models with the same non-linear regression fit. The application of such cloned datasets is for maintaining the confidentiality of sensitive real data for publication purposes. In this context, new methods can be developed so that cloning is possible for non-linear regression models. The question this study addresses is how cloning techniques are better than simulation and re-sampling. The simulation approach assumes that the model is known and then generates random data from the distribution of the response variable to illustrate the sampling variability in the estimates, re-sampling estimates the precision of sample statistics by using a subset of available data or drawing randomly with replacement from a set of data points. Unfortunately, these approaches do not help to explain the concept of regression or the idea of ‘moving towards’ the mean. The methods presented in this study are intended to fill this gap by yielding a sequence of matching data sets with the same fitted regression equation, for which the variability in the response variable Y and the explanatory variable X will progressively reduce. The tendency of moving towards the means rather than the conditional mean are also demonstrated.

Data Availability

All data underlying the results are available as part of the article and no additional source data are required.

Acknowledgements

This research is fully sponsored by Landmark University Centre for Research and Development, Landmark University, Omu-Aran, Nigeria.

References

- Anscombe FJ: **Graphs in Statistical Analysis**. *Am Stat* 1973; **27**(1): 17–21.
[Publisher Full Text](#)
- 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.
- Govindaraju K, Haslett SJ: **Illustration of regression towards the mean**. *Int J Math Educ Sci Technol* 2008; **39**(4): 544–550.
- Haslett SJ, Govindaraju K: **Cloning data: Generating datasets with exactly the same multiple linear regression fit**. *Aust N Z J Stat* 2009; **51**(4): 499–503.
[Publisher Full Text](#)
- Haslett SJ, Govindaraju K: **Data cloning: Data visualization, smoothing, confidentiality, and encryption**. *J Stat Plan Inference* 2012; **142**: 410–422.
[Publisher Full Text](#)
- Hussain S: **Generation of Cloning Data with Exactly the Same Nonlinear Regression Fit**. *Unpublished M. Phil Thesis* 2017; Islamabad, Pakistan: Allama Iqbal Open University.
- Matejka J, Fitzmaurice G: **Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing**. 2017; Toronto, Ontario, Canada: Autodesk Research.
- RDevelopment Core Team: **R: A language and environment for statistical computing**. *R Foundation for Statistical Computing* 2005; Vienna, Austria. ISBN 3-900051-07-0
[Reference Source](#)

Open Peer Review

Current Peer Review Status: ?

Version 1

Reviewer Report 15 October 2021

<https://doi.org/10.5256/f1000research.31296.r95224>

© 2021 Medina-Ortiz D. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

David Medina-Ortiz

Universidad de Chile, Región Metropolitana, Chile

The authors propose a novel approach to simulate datasets in case of confidential problems using non-linear regression approaches. The authors demonstrate your proposed strategy testing with different datasets and explain how your strategy could be applied on different case of uses.

Despite of the efforts of the authors to demonstrate the usability of the proposed method, there are some points that I think is necessary to solve:

1. How does the method can be compared with classical approaches?
2. Are there some method or strategy with similar results, could be make a comparison of this strategies and demonstrate the effects of the use of the proposed method?
3. Only if it is applicable, it is possible to create a case of use selecting a reported dataset in literature used for predictive models and apply the method proposed in this work, how does the strategy change the performance of the predictive models? For example, the authors could be select some examples and demonstrate that there are not effect over the performance to change the training datasets. I thinks, this could be so interesting, because, this strategy could be used to create synthetic datasets in case when the problem of get examples is so complex.

Some minor points are listed below:

1. Please, try to share the source code of your application, submit it in a GitHub repository for example.
2. Try to show the results in a more friendly form, for example using images. Normally, if there are various tables in the text the readers of the articles ignore it, mainly because is not "attractive" despite of the high impact of the presented work.

General comment: I think the work is nice, there are a lot of applications, but I believe it is necessary to work more on the manuscript, trying to compare the results with different strategies

and demonstrate the applications in machine learning and pattern recognition problems.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

No

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Machine learning, mathematical modeling, data sciences, pattern recognition, big queries, protein engineering, cloud architectures.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Comments on this article

Version 1

Author Response 18 Nov 2021

Roseline Ogundokun

- *How does the method can be compared with classical approaches?*
- **Response:** It has been mentioned in the conclusion that classical methods like the simulation approach assume that the model is known and then generates random data from the distribution of the response variable to illustrate the sampling variability in the estimates, re-sampling estimates the precision of sample statistics by using a subset of available data or drawing randomly with replacement from a set of data points. Unfortunately, these approaches do not help to explain the concept of regression or the idea

of 'moving towards the mean. The methods presented in our study are intended to fill this gap by yielding a sequence of matching data sets with the same fitted regression equation, for which the variability in the response variable Y and the explanatory variable X will progressively reduce.

- *Are there some methods or strategies with similar results, could make a comparison of these strategies and demonstrate the effects of the use of the proposed method?*
- **Response:** Yes, there are methods available with similar results, could make a comparison of these strategies and demonstrate the effects of the use of the proposed method but they have limitations which have been discussed in the conclusion section and our study fills the gap of limitations of those methods.
- *Only if it is applicable, it is possible to create a case of use selecting a reported dataset in literature used for predictive models and applying the method proposed in this work, how does the strategy change the performance of the predictive models? For example, the authors could be select some examples and demonstrate that there are not effective over the performance to change the training datasets. I think, this could be so interesting, because this strategy could be used to create synthetic datasets in the case when the problem of getting examples are so complex*
- **Response:** The authors thank the reviewer for this valuable suggestion. We have submitted our proposal to higher education for its funding and after the acceptance of our proposal from higher authority, we will be working on the suggested work.
- *Try to show the results in a more friendly form, for example using images. Normally, if there are various tables in the text the readers of the articles ignore it, mainly because is not "attractive" despite the high impact of the presented work.*
- **Response:** The tables were necessary as part of our work to show how the data can be cloned. We will be working on improving the outlook for our article in the next project including your suggestions and comments. It should be noted that the programming language codes for example 1 (table 2 and 2b), example 1 and fig 1, and example 8 (Table 9 and 9b) are shown in the appendix session.

Competing Interests: There are no competing interests disclosed in this article.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research