Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. **51**(4), 2009, 499–503 doi: 10.1111/j.1467-842X.2009.00560.x

CLONING DATA: GENERATING DATASETS WITH EXACTLY THE SAME MULTIPLE LINEAR REGRESSION FIT

S. J. HASLETT AND K. GOVINDARAJU*

Massey University

Summary

This paper presents a simple computational procedure for generating 'matching' or 'cloning' datasets so that they have exactly the same fitted multiple linear regression equation. The method is simple to implement and provides an alternative to generating datasets under an assumed model. The advantage is that, unlike the case for the straight model-based alternative, parameter estimates from the original data and the generated data do not include any model error. This distinction suggests that 'same fit' procedures may provide a general and useful alternative to model-based procedures, and have a wide range of applications. For example, as well as being useful for teaching, cloned datasets can provide a model-free way of confidentializing data.

Key words: data cloning; multiple linear regression; orthogonality.

1. Introduction

Anscombe (1973) presented four diverse datasets to illustrate the importance of the graphical exploration of data. These four very well-known datasets have the same simple regression fit, R^2, and estimated residual standard deviation, s, etc. Chatterjee & Firat (2007) provided a method of generating diverse (bivariate) datasets with identical summary statistics but dissimilar graphs by using a genetic-algorithm-based approach. Govindaraju & Haslett (2007) presented a rather simpler procedure based on the orthogonality principle to obtain several naturally matching datasets giving the same simple regression equation. Regressing towards the mean is a natural concept that forms part of the material in Govindaraju & Haslett (2007), and hence the matching datasets there are not as arbitrary or as diverse as those of Anscombe (1973) and Chatterjee & Firat (2007). One possible application of such naturally matching datasets is for confidentializing sensitive real data for publication purposes, where having datasets with exactly the same fit as the original data is a major advantage, and major departures in data structure, such as are apparent in Anscombe's dataset and inherent in Chaterjeee & Firat's, are disadvantageous.

2. Methodology

We first consider the case of independent and identically distributed data, but extend the methods to an arbitrary error covariance structure. Let the multiple regression model be

*Author to whom correspondence should be addressed.
Institute of Fundamental Sciences, Massey University, Palmerston North 4442, New Zealand.
e-mail: k.govindaraju@massey.ac.nz

Acknowledgments. The authors would like to acknowledge the assistance of the Technical Editor, Associate Professor Ken Russell.

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{Y}$ is the $n \times 1$ vector of responses, $\boldsymbol{X} = (\boldsymbol{X}_1 : \boldsymbol{X}_2 : \ldots . : \boldsymbol{X}_p)$ is the $n \times p$ covariate data matrix, $\boldsymbol{\beta}$ is the unknown $p \times 1$ vector of parameters, and $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of errors. When $\boldsymbol{X}$ has full (column) rank, the ordinary least squares (OLS) estimate of $\boldsymbol{\beta}$ is

$$\boldsymbol{b} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y},$$

and the fitted multiple regression equation is

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{b}. \tag{2}$$

Let $\mathbf{1}$ denote a vector of ones and $\bar{Y}$ and $\bar{X}_i$ be the sample means of the elements in $\boldsymbol{Y}$ and $\boldsymbol{X}_i$ for $i = 1, 2, \ldots, p$, respectively. Then, (2) can be written in a mean-corrected form as

$$\hat{\boldsymbol{y}} = b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2 + \cdots + b_p\boldsymbol{x}_p, \tag{3}$$

where $\hat{\boldsymbol{y}} = \hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}$, with $\bar{\boldsymbol{Y}} = \bar{\boldsymbol{y}}\mathbf{1}$, and $\boldsymbol{x}_i = \boldsymbol{X}_i - \bar{\boldsymbol{X}}_i$, with $\bar{\boldsymbol{X}}_i = \bar{\boldsymbol{x}}_i\mathbf{1}$ for $i = 1, 2, \ldots, p$ (and where, if $\boldsymbol{X}$ contains $\mathbf{1}$ as one of its columns, then because that column becomes a column of zeros, it is dropped and p in (3) is reset to one less than the number of columns in $\boldsymbol{X}$, without any loss of generality).

Our problem is to obtain a new vector of the response variable, $\boldsymbol{Y}_{\text{new}}$, and a new data matrix, $\boldsymbol{X}_{\text{new}}$, such that

$$\boldsymbol{b} = \left(\boldsymbol{X}_{\text{new}}^\top \boldsymbol{X}_{\text{new}}\right)^{-1} \boldsymbol{X}_{\text{new}}^\top \boldsymbol{Y}_{\text{new}}.$$

In other words, what we require is a matching multivariate dataset $(\boldsymbol{Y}_{\text{new}}, \boldsymbol{X}_{\text{new},1}, \boldsymbol{X}_{\text{new},2}, \ldots, \boldsymbol{X}_{\text{new},j}, \ldots, \boldsymbol{X}_{\text{new},p})$ that will lead to exactly the same multiple linear regression equation as that for the original dataset $(\boldsymbol{Y}, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_j, \ldots, \boldsymbol{X}_p)$. Later we consider the extension to the case where the $\boldsymbol{\varepsilon}$ are correlated, with a specified (i.e. known or estimated) covariance matrix. This extension will include generating cloned data in linear mixed models.

Returning to the independent identically distributed case for now, we show below how the generation of matching or cloned data can be achieved by manipulating any one of the covariates, say $\boldsymbol{x}_j$, using the following steps.

(1) First fit the multiple regression model as in (3), using mean corrected data.

(2) Choose any component $\boldsymbol{x}_j$, where $j = 1, 2, \ldots, p$.

(3) Let $\hat{\boldsymbol{y}} = \boldsymbol{k} + b_j\boldsymbol{x}_j$, where $\boldsymbol{k} = b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2 + \cdots + b_{j-1}\boldsymbol{x}_{j-1} + b_{j+1}\boldsymbol{x}_{j+1} + \cdots + b_p\boldsymbol{x}_p = \hat{\boldsymbol{y}} - b_j\boldsymbol{x}_j$. Perform the simple regression of $\boldsymbol{y}_k = \boldsymbol{y} - \boldsymbol{k}$ on $\boldsymbol{x}_j$ and obtain the fitted values $\hat{\boldsymbol{y}}_k$. In addition, perform the inverse simple regression of $\boldsymbol{x}_j$ on $\boldsymbol{y}_k$ and obtain the fitted values $\hat{\boldsymbol{x}}_j$.

(4) Regress each $\boldsymbol{x}_i (i \neq j)$ on $\boldsymbol{x}_j$ and obtain $\hat{\boldsymbol{x}}_i = \boldsymbol{x}_j(\boldsymbol{x}_j^\top\boldsymbol{x}_j)^{-1}\boldsymbol{x}_j^\top\boldsymbol{x}_i$. Also obtain $\boldsymbol{x}_{i,j} = \boldsymbol{x}_i - \boldsymbol{x}_j(\boldsymbol{x}_j^\top\boldsymbol{x}_j)^{-1}\boldsymbol{x}_j^\top\boldsymbol{x}_i = (\boldsymbol{I} - \boldsymbol{x}_j(\boldsymbol{x}_j^\top\boldsymbol{x}_j)^{-1}\boldsymbol{x}_j^\top)\boldsymbol{x}_i$, where $\boldsymbol{I}$ is the identity matrix.

(5) Form $\boldsymbol{y}_{k,\text{new}} = \hat{\boldsymbol{y}}_k + \sum_{i \neq j} b_i\boldsymbol{x}_{i,j}$.

(6) Carry out the multiple regression of $\boldsymbol{y}_{k,\text{new}}$ simultaneously on all the newly obtained $\{\boldsymbol{x}_{i,j}(i \neq j)\}$ and $\hat{\boldsymbol{x}}_j$, where $\hat{\boldsymbol{x}}_j = \boldsymbol{y}(\boldsymbol{y}^\top\boldsymbol{y})^{-1}\boldsymbol{y}^\top\boldsymbol{x}_j$ in which $\boldsymbol{y} = (\mathbf{1} : \hat{\boldsymbol{y}})$ is $n \times 2$.

(7) If preferred, add back $\bar{\boldsymbol{y}}$ and $\bar{\boldsymbol{x}}_i$ for $i = 1, 2, \ldots, p$ to the cloned data, and/or multiply all the cloned data by the same scale factor.

(8) Repeat steps 1 to 7 above to prepare a sequence of datasets, all with the exactly the same regression coefficients. The choice of a possibly different value of j can be made at each iteration.

Step 2 ensures that b_j is unchanged (see Govindaraju & Haslett 2007, for a proof). For the multivariate case, the regression coefficient for $\hat{\boldsymbol{x}}_j$ must be b_j because $\{\boldsymbol{x}_{i,j}(i \neq j)\}$ are orthogonal to $\hat{\boldsymbol{x}}_j$. Similarly, for each of the $\boldsymbol{x}_{i,j}$, the corresponding regression coefficient must be b_i, again because of the orthogonal construction of $\{\boldsymbol{x}_{i,j}(i \neq j)\}$.

When the errors $\boldsymbol{\varepsilon}$ in the multiple regression equation (1) are correlated, with the known or estimated full-rank covariance matrix $\boldsymbol{V}$, say, an eigen-analysis or Choleski decomposition will provide a full-rank matrix $\boldsymbol{V}^{1/2}$ with an inverse $\boldsymbol{V}^{-1/2}$ that has the property $\boldsymbol{V}^{-1/2}(\boldsymbol{V}^{-1/2})^\top = \boldsymbol{V}^{-1}$, the inverse of $\boldsymbol{V}$. Pre-multiplication of $\boldsymbol{Y}$ and $\boldsymbol{X}$ in (1) by $\boldsymbol{V}^{-1/2}$ will then yield an equation for which the ordinary least squares estimator (OLSE) used in the algorithm above is the best linear unbiased estimator (BLUE) of the model with error covariance $\boldsymbol{V}$. When $\boldsymbol{V}$ is specified in a mixed model, and the random parameters are incorporated into $\boldsymbol{V}$, the estimates of the fixed parameters in the mixed model will remain unchanged, and, in addition, the random parameter estimates will remain unchanged by construction. The algorithm above consequently provides a method of generating a sequence of multivariate datasets all with the same fixed parameter estimates (i.e. BLUEs) even when the original data are correlated.

When the design matrix $\boldsymbol{X}$ in (1) is not of full rank, as is often the case in experimental design, the following preliminary and consequent steps (which reduce $\boldsymbol{X}$ to canonical form and later back-transform it) allow the same algorithm to be used. Suppose $\boldsymbol{X}$ (which is $n \times p$) has rank r, where $r < p$. Find a basis for (the columns of) $\boldsymbol{X}$, or more simply choose r linearly independent columns of $\boldsymbol{X}$. Arrange these as an $n \times r$ matrix, denoted as $\boldsymbol{X}_r$, say. Define $\boldsymbol{\beta}_r$ via $\boldsymbol{X\beta} = \boldsymbol{X}_r\boldsymbol{\beta}_r$. Then by construction, as $\boldsymbol{X}$ and $\boldsymbol{X}_r$ have the same column space, we have $\boldsymbol{X} = \boldsymbol{X}_r\boldsymbol{L}_p$ where $\boldsymbol{L}_p = (\boldsymbol{X}_r^\top\boldsymbol{X}_r)^{-1}\boldsymbol{X}_r^\top\boldsymbol{X}$, and $\boldsymbol{X}_r = \boldsymbol{X}\boldsymbol{L}_r$ with $\boldsymbol{L}_r = (\boldsymbol{X}^\top\boldsymbol{X})^{-}\boldsymbol{X}^\top\boldsymbol{X}_r$. Here $\boldsymbol{L}_r$ and $(\boldsymbol{X}^\top\boldsymbol{X})^{-}$ are generalized inverses of $\boldsymbol{L}_p$ and $(\boldsymbol{X}^\top\boldsymbol{X})$, respectively, the choice of $(\boldsymbol{X}^\top\boldsymbol{X})^{-}$ simply reflecting how $\boldsymbol{X}_r$ has been specified. This specification allows the original transformation from $\boldsymbol{X}$ to $\boldsymbol{X}_r$ to be reversed. Hence transforming the problem from $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ to $\boldsymbol{y} = \boldsymbol{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}$ before applying the algorithm above and reversing the process afterwards gives the adjustment required if $\boldsymbol{X}$ is not full rank. Explicitly $\boldsymbol{b} = \boldsymbol{L}_r\boldsymbol{b}_r$, where $\boldsymbol{b}_r$ is the solution to the transformed problem.

3. An example

Using uncorrelated data and a full-rank design matrix, consider the fictitious variables X_1, X_2 and Y shown in Table 1 andresulting in the multiple regression fit

$$\hat{Y} = 1.48 + 0.967X_1 + 0.179X_2. \tag{4}$$

Steps 1 to 5 given above (manipulating X_2) will yield the matching or cloned data shown in Table 2, for which the fitted multiple regression equation is exactly the same as in (4).

The matching data in $X_{2,\text{new}}$ include negative values and average to zero, which may not be realistic in some contexts. This can be easily corrected by adding constant values to $X_{2,\text{new}}$ at the same time as adjusting for the cloned data average values of Y and Y_{new}. Table 3 shows this adjustment. The data shown in Table 3 again yield the same regression fit as in (4).

TABLE 1

A fictitious three-variable dataset

X_1	X_2	Y
3	4.1	5
4	6.2	7
4	4.5	6
3	3.1	5
2	3.9	4
5	6.1	8
5	7.2	7
7	8.6	10
6	7.5	8
6	8.7	9

TABLE 2

Cloned data for the data in Table 1

$X_{1,\text{new}}$	$X_{2,\text{new}}$	Y_{new}
2.99351	−0.12	4.35652
4.56021	0.80	5.48777
3.88909	−0.90	5.18349
3.16617	−1.12	4.17754
2.06341	0.86	3.56536
5.54211	−0.48	6.22523
4.38755	0.62	6.42212
7.03974	−0.34	8.18343
5.30039	−0.26	7.23118
6.05784	0.94	7.44597

TABLE 3

Matching data for Table 1 data are obtained from Table 2 cloned data by adding 1.07214 to each Y_{new} element and 5.99 to the $X_{2,\text{new}}$ elements in order to preserve the same mean response for the raw and matching data

$X_{1,\text{new}}$	$X_{2,\text{new}}$	Y_{new}
2.99350	5.87	5.42866
4.56020	6.79	6.55991
3.88908	5.09	6.25563
3.16616	4.87	5.24968
2.06340	6.85	4.63750
5.54210	5.51	7.29737
4.38754	6.61	7.49426
7.03973	5.65	9.25557
5.30038	5.73	8.30332
6.05783	6.93	8.51811

Figure 1 shows the matrix plot of the raw and matching data given in Tables 1 and 3, respectively. The orthogonal manipulation done on X_2 in the algorithm above, applied once, gives $X_{2,\text{new}}$. Figure 1 shows the effect of this manipulation. The strength of the bivariate relationship between $X_{2,\text{new}}$ and Y_{new} is much weaker than that of the relationship between X_2

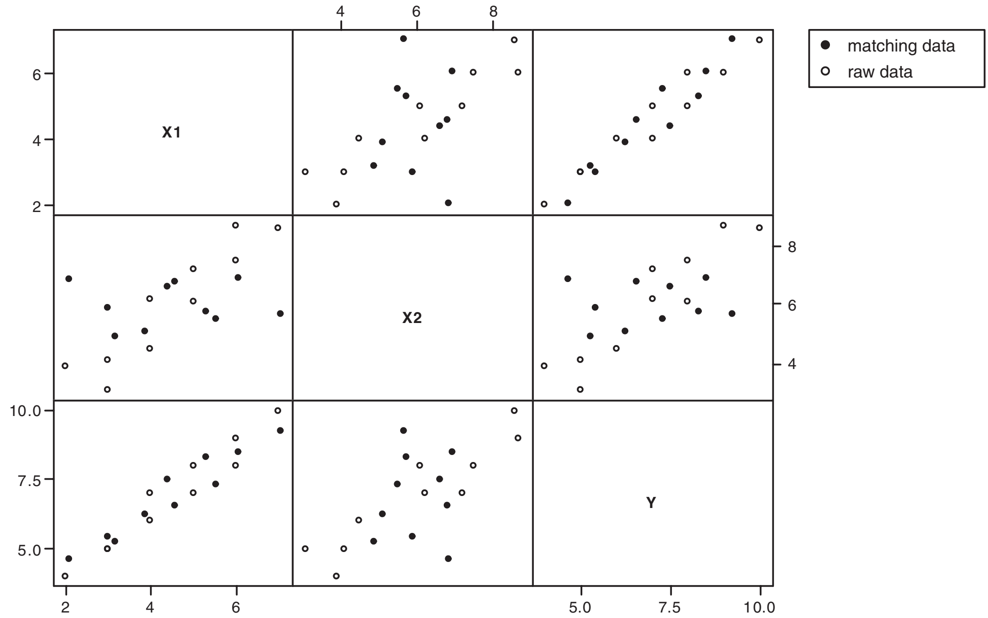

Figure 1. Matrix plot of raw and matching datasets, based on Table 1 raw data and Table 3 matching data. Data cloning was performed by manipulating X_2, and hence the matching $X_{2,\text{new}}$ data show a weaker relationship with the other variables.

and Y. However, this is not the case with $X_{1,\text{new}}$ and Y_{new}, because the orthogonal manipulation was not done with X_1.

4. Discussion

The usefulness of the algorithm in teaching to produce any number of cloned datasets with exactly the same multiple regression is clear. Perhaps less obviously, the algorithm provides an alternative way of confidentializing data so that a multiple regression analysis (which may include fitting mixed models, such as occur in small-area estimation, for example) has exactly the same fit in the cloned as in the original data, even though the data have been changed to be no longer confidential. This would seem to have important applications where model fitting to confidential unit records is required but the original data cannot be released. In this context, further research on whether methods can be developed so that cloning is possible for a collection of models would be a useful future research topic. More generally, given their parallels with multiple regression, data cloning for multivariate statistics may also warrant further research.

References

Anscombe, F.J. (1973). Graphs in statistical analysis. *Amer. Statist.* **27**, 17–21.

Chatterjee, S. & Firat, A. (2007). Generating data with identical statistics but dissimilar graphics: A follow up to the Anscombe dataset. *Amer. Statist.* **61**, 248–254.

Govindaraju, K. & Haslett, S.J. (2007). Illustration of regression towards the means. *Int. J. Math. Edu. Sci. Tech.* **39**, 544–550.