GENERATING DIFFERENT DATA SETS FOR LINEAR REGRESSION MODELS WITH THE SAME ESTIMATES

SilvioSandoval Zocchi and Bryan F.J. Manly
University of Sao Paulo, Brazil
sszocchi@esalq.usp.br

For teaching purposes it is sometimes useful to be able to provide the students in a class with different sets of regression data which, nevertheless give exactly the same estimated regression functions. In this paper we describe a method showing how this can be done, with a simple example. We also note that the method can be generalized for situations where the regression errors are not independently distributed with a constant covariance matrix.

INTRODUCTION

For coursework and examination purposes it will sometimes be useful to be able to give students regression data sets that appear to be different, and yet give exactly the same estimates for the regression function. This can be done using a method originally proposed by Huh and Jhun (2001) for testing the significance of regression coefficients using randomization methods. In brief, Huh and Jhun proposed transforming the residuals from a multiple regression into uncorrelated variables with a constant variance, putting the residuals into a random order, and then transforming back to the scale of the original residuals. The new residuals then replace the original residuals to provide a set of data with new values for the dependent variable. As it happens, the regression estimates for the new data are exactly the same as for the original data.

THE METHOD

Consider the usual multiple regression model

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i,\ i = 1, 2, ..., n,$$

where ε_i is a random error with mean zero and constant variance σ^2, and where $x_{i1} = 1$ if there is a constant in the model. In matrix notation this becomes

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E},$$

where $\mathbf{Y}$ is an n by 1 vector of y values, $\mathbf{X}$ is an n by p matrix of x values, $\mathbf{B}$ is an n by 1 vector of regression coefficients, and $\mathbf{E}$ is an n by 1 vector of the regression errors.

With this notation it is well known that the least squares estimator of the regression coefficients is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

The estimated regression residuals are then

$$\mathbf{e} = \mathbf{Y} - \mathbf{Xb} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\ \boldsymbol{\varepsilon},$$

and it can be shown that the covariance matrix for these estimated residuals is

$$\mathbf{V}_e = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\ \sigma^2$$

where $\mathbf{I}$ is an n by n identity matrix and σ^2 is the variance of the true regression residuals.

The last equation shows that in general the variances of the estimated residuals are not equal, and the estimated residuals are correlated. This is because $\mathbf{V}_e$ has to be an identity matrix multiplied by a scalar constant for these conditions to hold, which will not usually be the case. The residuals are therefore not generally exchangeable as needed for a randomization test.

To overcome this problem, Huh and Jhun (2001) suggested that a linear transformation of the estimated residuals should be carried out to obtain uncorrelated variables with a constant variance. For example the residual variance can be estimated by the mean square error $\hat{\sigma}^2$ in the usual way matrix $\mathbf{V}_e\hat{\sigma}^2$ can be input to a principal components analysis. The principal components obtained will then be uncorrelated, and have equal variances (Manly, 2005, Chapter 6). Because of the relationship between the estimated residuals caused by fitting the regression model, p of the principal components will be identically zero, in which case they must be fixed at zero for the randomization stage of Huh and Jhun's analysis.

Once the uncorrelated variables with a constant variance are obtained these can be put in a random order. The randomized values can then be back-transformed to obtain randomized residuals for the original regression model. These are then added to the expected Y values to obtain a randomized set of data to compare with the observed data. In their paper, Huh and Jhun show that this new set of data will have exactly the same regression estimates as the original set.

EXAMPLES

Consider the data in Table 1, which is ten observations on the natural logarithms of chlorophyll-a, phosphorus and nitrogen from the lakes in a region. It represents a subset of the results for 74 lakes that were published by Smith and Shapiro (1981).

Table 1: Example regression data were Y is log(chlorophyll-a concentration), X_1 represents a constant term, X_2 is log(phosphorus concentration), and X_3 is log(nitrogen concentration).

Y	X_1	X_2	X_3	Fit	Res
4.5	1.00	5.80	2.08	4.77	-0.22
3.6	1.00	5.35	1.79	3.90	-0.24
3.3	1.00	4.68	2.40	3.75	-0.45
2.5	1.00	3.03	2.77	2.14	0.42
3.5	1.00	4.10	2.20	2.82	0.73
2.7	1.00	3.27	2.83	2.50	0.20
5.0	1.00	6.39	1.39	4.73	0.33
1.6	1.00	3.66	2.56	2.68	-1.05
2.3	1.00	3.74	2.40	2.60	-0.24
4.5	1.00	4.60	2.77	4.06	0.50
				SSE	2.57
				MSE	0.37

For these data the estimated regression coefficients for X_1 to X_3 are -4.613, 1.223 and 1.098, respectively. The error sum of squares is 2.57, the error mean square is 2.573/7 = 0.368 and the estimated covariance matrix for the residuals is

$$V_e = \begin{bmatrix} 0.233 & -0.274 & -0.312 & 0.036 & 0.019 & 0.002 & -0.006 & 0.006 & 0.024 & -0.031 \\ -0.274 & 0.274 & -0.312 & -0.036 & -0.019 & -0.011 & 0.006 & -0.006 & -0.024 & 0.031 \\ -0.312 & -0.312 & 0.312 & -0.036 & -0.019 & -0.011 & 0.006 & -0.006 & -0.024 & 0.031 \\ 0.036 & -0.036 & -0.036 & 0.261 & -0.019 & -0.011 & 0.045 & -0.006 & -0.024 & 0.031 \\ 0.019 & -0.019 & -0.019 & -0.019 & 0.285 & -0.011 & -0.006 & -0.006 & -0.024 & 0.049 \\ 0.002 & 0.011 & -0.011 & -0.036 & -0.019 & 0.279 & 0.053 & -0.006 & -0.024 & -0.031 \\ -0.006 & -0.006 & -0.006 & 0.045 & -0.019 & 0.053 & 0.168 & 0.007 & -0.024 & 0.037 \\ 0.006 & -0.006 & -0.006 & -0.036 & -0.019 & -0.011 & 0.007 & 0.308 & -0.024 & -0.031 \\ 0.024 & -0.024 & -0.024 & -0.036 & -0.019 & -0.011 & -0.006 & -0.006 & 0.293 & 0.026 \\ -0.031 & 0.031 & -0.031 & -0.036 & 0.049 & -0.031 & 0.037 & -0.006 & 0.026 & 0.160 \end{bmatrix}$$

This matrix has seven eigenvalues equal to the error mean square 0.368, with corresponding eigenvectors. If the 10 by 10 matrix with the i^{th} eigenvector in the i^{th} column is denoted by **H**, then the residuals transformed to uncorrelated variables with equal variances are given by

$$\mathbf{e}^* = \mathbf{H}'\mathbf{e},$$

where **e** is the vector of original residuals. For the present data this gives the uncorrelated values

$$(\mathbf{e}^*)' = (\ -0.582,\ 0.954,\ 0.477,\ -0.298,\ 0.947,\ 0.296,\ 0.000,\ 0.000,\ 0.0000,$$

where the last three values correspond to the zero eigenvalues.

To obtain a new set of data with the same regression estimates, the non-zero transformed residuals are put in a new order. For example, this might be -0.298, 0.954, 0.947, 0.152, 0.477, -0.582, 0.296, 0.000, 0.000, 0.000. These values then need to be back-transformed to find the residuals that they correspond to with the original data. As the matrix H is orthogonal the back-transformation is simply given by

$$\mathbf{e}_1 = \mathbf{H}\mathbf{e}^*_1,$$

where $\mathbf{e}^*_1$ is the reordered vector of uncorrelated variables. The new residuals are then added to the fitted values in Table 1 to get the new set of data.

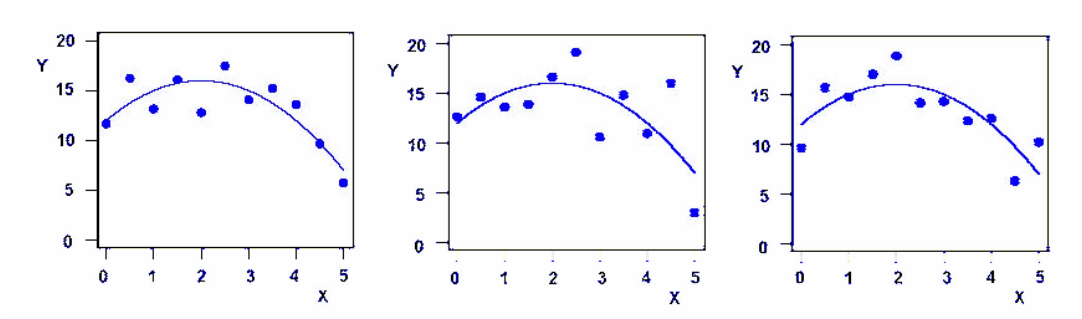
With the reordered uncorrelated variables shown above, the back-transformed residuals are -0.16, -0.64, -0.35, -0.83, 0.09, -0.07, 0.28, 0.31, 0.99, and 0.39, in order. Adding these to the fitted values shown in Table 1 gives the new Y values, which are 4.60, 3.26, 3.40, 1.30, 2.91, 2.43, 5.01, 2.99, 3.59 and 4.44, in order. It is easy to verify that the regression with these *Y* values gives the same estimated regression coefficients, standard errors, and error mean square as the *Y* values in Table 1.

We note that the matrix **H** is not unique. Different algorithms for choosing a transformation to uncorrelated residuals may therefore change the details of this example, whilst still producing different sets of data with the same regression estimates.

As another example, consider the quadratic regression model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

and the *X* values {0.0, 0.5, 1.0,.. . ,4.5, 5.0}. Figure 1 shows scatter plots and fitted models for different data sets with the estimates 12, 4, and -1 for β_0, β_1, and β_2. This illustrates again how rather different sets of data can lead to an identical fitted regression model



DISCUSSION

The emphasis in this paper has been on the use of the Huh and Jhun method to generate alternative data sets that give the same regression estimates. However, other applications may also be useful, such as generating the alternative data sets for some sort of simulation study. Also, of course, there is Huh and Jhun's original application with randomization tests for the significance of the estimated regression coefficients.

The above theory may also be of interest in an any advanced course involving regression to demonstrate that different sets of data may give the same regression estimates, or just as an interesting coursework example on an aspect of regression theory.

Finally, we note that it is not difficult to generalize the results presented here to situations where the regression errors are correlated, and may have unequal variances. These extensions will be discussed more fully elsewhere (Zocchi and Manly, in preparation).

REFERENCES

Huh, M. and Jhun, M. (2001). Random permutation testing in multiple linear regression. *Communications in Statistics - Theory and Methods*, 30, 2023-2032.

Manly, B. F. J. (2005). *Multivariate Statistical Methods: A Primer* (3rd edition). Boca Raton: Chapman and Hall/CRC.

Smith, V. H. and Shapiro, J. (1981). Chlorophyll-phosphorus relations in individual lakes: Their importance to lake restoration strategies. *Environmental Science and Technology*, 15, 444-451.