Outlier Detection by Regression Diagnostics in Large Data

Preprint · October 2009

CITATIONS
0

READS
21

2 authors:

Abdul Awal Md Nurunnabi
University of Luxembourg
94 PUBLICATIONS **1,485** CITATIONS
SEE PROFILE

M. Nasser
University of Rajshahi
33 PUBLICATIONS **303** CITATIONS
SEE PROFILE

Outlier Detection by Regression Diagnostics in Large Data

A. A. M. Nurunnabi
School of Business, Uttara University
Dhaka-1230, Bangladesh
e-mail: pyal1471@yahoo.com

Mohammed Nasser
Department of statistics, University of Rajshahi
Rajshahi-6205, Bangladesh
e-mail: mnasser.ru@gmail.com

***Abstract* —— Regression analysis is a well known supervised learning techniques. To justify an effective model from regression analysis it is necessary to check and preprocess the data set. Without outliers (noise) it is impossible to get a real data. Areas in bio-informatics, astronomy, image analysis, computer vision etc, large or fat data appear with unusual observations (outliers) very naturally. Especially when the data sets are large, in these industries robust regression are commonly used in model building process. But robust regressions are not good enough in large and/or high dimensional data. Checking raw data for outliers in regression is regression diagnostics. Robust regression and regression diagnostics are complementary ideas and any one is not enough for a contaminated data. Most of the popular diagnostic methods are not sufficient for large data because of masking and swamping. In this article, both of the above ideas are shortly discussed and we show a new measure can effectively identify influential observations in linear regression for large data.**

Key words–Influential observation; learning method; outlier; regression diagnostics; robust regression.

I. INTRODUCTION

Learning methods for analyzing and modeling data are divided into two: "supervised learning" and "unsupervised learning" to the pattern recognition and machine learning community. Statistical models are extremely useful devices for extracting and understanding the essential features of a set of data. Classification, decision trees, neural networks, and regression analysis are well known supervised learning methods. In recent years regression analysis has drawn much attention for its wide range of applicability to the user of learning community. Regression can be performed using many different types of techniques, including classification, genetic algorithms, neural networks and optical flow estimation.

Supervised learning requires input data that has both predictor (independent) variables and target (dependent/response) variable, whose value is to be estimated. By various means, the process "learns" how to model (predict) the value of the target variable based on the predictor variables. Regression analysis is one of the most important supervised learning methods that have the objective to specify a relationship between the target (dependent) and explanatory (independent) variables. Data contamination is a practical problem, introduced in many ways in different research fields. Noise can greatly impair data learning. It is often more severe problem in the context of streams (large and high dimensions) because it interfaces with change adaptation. Outlier detection has been suggested for numerous applications, such as network intrusion, credit and fraud detection, clinical trials, voting irregularity analysis, severe weather prediction, geographic information system, and other data mining tasks. Outlying observations do not inevitably 'perplex' or 'mislead'; they are not necessarily 'bad' or 'erroneous', and the experimenter may be tempted in some situations not to reject an outlier but to welcome it as an indication of new and important findings.

Regression analysis is not an exception to this kind of problem and robust regression and regression diagnostics are two remedies of the problem. But most of the identification techniques are performed well only for small and few dimensional data. In this paper we have a brief discussion of robust regression in Section II, Section III focused on regression diagnostics. We propose a new diagnostic measure for identifying multiple influential observations for large data set in Section IV. Numerical demonstrations are thereafter followed by conclusion in Section VI.

II. ROBUSTNESS AND ROBUST REGRESSION

Box [2] first introduced the technical term 'robust' and the subject matter acquired recognition as a legitimate topic for investigation only in mid-sixties, mainly due to the pioneer works of Tukey [16], Huber [10], and Hampel [8]. Robustness is a property of estimation technique like unbiasdness, consistency or efficiency. It is concerned with evaluating and improving the stability of estimation techniques when data points are deviated from assumptions. The theory of robustness is not just a superfluous mathematical decoration. It plays an essential role in organizing and reducing information about the behavior of statistical procedures to a manageable form Hampel et al. [9].

There are many types of robust regression techniques in literature, they work in different ways but all give less weight to observations that would otherwise influence the

regression line and the estimation of the parameters. Some of the most popular are discussed below.

Least median of squares regression (LMS): It was developed by Rousseeuw [14]. Instead of minimizing the sum of squared residuals, $\sum e_i^2$ as in LS to estimate the parameter vector β, Rousseeuw proposed minimizing their median as follows:

$$\underset{\hat{\beta}}{\text{minimize}}\ \underset{i}{med}\ e_i^{\,2} \quad . \tag{1}$$

This estimator effectively trims the n/2 observations having the largest residuals, and uses the maximal residual value in the remaining set as the criterion to be minimized. The LMS estimators perform poorly from the point of view of asymptotic efficiency and convergence.

Least trimmed of squares regression (LTS):

Rousseeuw proposed it [14] and defined as

$$\underset{\hat{\beta}}{\text{minimize}} \sum_{i=1}^{h} \left(e^2\right)_{i:n} ; \tag{2}$$

where $\left(e_1^{\,2}\right) \le \left(e_2^{\,2}\right) \le \ldots \le \left(e_h^{\,2}\right) \le \ldots \le (e_n^2)$ are the ordered squared residuals. The formula is similar to OLS, and the only difference being that the largest $(n-h, h = \dfrac{n+p+1}{2})$ squared residuals are not used in this summation.

Reweighted Least Squares (RLS)

Reweighted least squares was developed by Rousseeuw and Leroy [15] in order to improve on the crude LMS and LTS solutions, and to obtain standard quantities like t-values, confidence intervals, and the like. We can apply a weighted least squares analysis based on the identification of the outliers. For instance, we could make use of the following weights:

$$w_i = \begin{cases} 1 & if \left| e_i / \hat{\sigma} \right| \le 2.5 \\ 0 & if \left| e_i / \hat{\sigma} \right| > 2.5 \end{cases} . \tag{3}$$

This means simply that case *i* will be retained in the weighted LS, if its LMS residual is small to moderate. The bound 2.5 is, of course, arbitrary, but quite reasonable because in a Gaussian situation there will be very few residuals larger than $2.5\hat{\sigma}$. We then apply weighted least squares defined by $minimize \sum_{i=1}^{n} w_i e_i^{\,2}$.

M-estimators

Robust M-estimators introduced by Huber [11] are based on the idea of replacing the sum of squared residuals $\sum e_i^2$ used in LS estimation by another function of the residual, $\sum_{i=1}^{n} \rho(e_i)$; where ρ is a symmetric function generally with a unique minimum at zero. Differentiating this expression with respect to the regression coefficients yields

$$\sum_{i=1}^{n} \psi(e_i) x_i = 0 , \tag{4}$$

where ψ is the derivative of ρ. It has two drawbacks; firstly, the estimators are not equivariant, and secondly the finite breakdown point is 1/n due to outlying x_i. To remove these demerits by

$$\sum_{i=1}^{n} w(x_i) \psi(e_i / \hat{S}) x_i = 0 , \tag{5}$$

where σ is estimated simultaneously by $\hat{s}$.

The lack of simple procedures for inference or reluctance to use the straightforward inference based on asymptotic justification. We have to take some help from the re-sampling techniques like '*bootstrapping*', '*jackknifing*' etc especially for small samples. These are computer intensive techniques, introduced by Efron [5].

III. Regression Diagnostics

In the customary linear regression model

$$Y = X\beta + \varepsilon , \tag{6}$$

we predict the output *Y* with least squares method via the model,

$$\hat{Y} = X\hat{\beta} \text{ (vector-matrix form);} \tag{7}$$

where $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$ and the corresponding residual vector, $r = Y - \hat{Y} = (I - H)\varepsilon$; in scalar form,

$$r_i = \varepsilon_i - \sum_{j=1}^{n} h_{ij} \varepsilon_j ; \quad i = 1,2,...,n$$

where $H = X(X^T X)^{-1} X^T$ is the leverage matrix and its *j*th diagonal element is h_{ij}. Usually r_i and h_{ij} along with their different transformed versions are used for the diagnosis of outliers and leverage points respectively.

It is known that as with outliers, high leverage points need not be influential, and influential observations are not necessarily high-leverage points. According to Belsley et al. [1], influential observation is one which either individual or together with several other observations has a demonstrably larger impact on the

calculated values of various estimates than is the case for most of the other observations. Field diagnostics is a combination of graphical and numerical tools. It is designed to detect and delete (if necessary recheck/correct) and then fit the good data by classical methods. It tries to iteratively detect possibly wrong data and reject/refit them through analysis of globally fitted model.

A large body of literature is now available [1, 3, 7, 13] for the identification of outliers and influential observations. Among them Cook's distance [4], and DFFITS [2] become very popular with the practitioners for identifying influential observations.

Cook's Distance

Cook [4] has suggested a way for reducing the influence curve, using a measure of the squared distance between the least-squares estimate based on all *n* points $\hat{\beta}$ and the estimate obtained by deleting the *i*th point, say $\hat{\beta}^{(-i)}$. This distance measure can be expressed as

$$CD_i = \frac{\left(\hat{\beta}^{(-i)} - \hat{\beta}\right)^T X^T X \left(\hat{\beta}^{(-i)} - \hat{\beta}\right)}{pMS_{\text{Res}}}. \; i=1,2,\ldots,n \qquad (8)$$

The magnitude of CD_i is usually assessed by comparing it to $F_{\alpha,p,n-p}$. If $CD_i = F_{0.5,p,n-p}$, then deleting point *i* would move $\hat{\beta}^{(-i)}$ to the boundary of an approximate 50% confidence region for β based on the complete data set. This is a large displacement and indices that the least-squares estimate is sensitive to the *i*th data point. The inventor suggested *CD* > 1 is influential case.

Welsch-Kuh Distance/DFFITS

The influence of the *i*th observation on the predicted value can be measured by the change in the prediction at x_i when the *i*th observation is omitted, relative to the standard error of, that is,

$$\frac{\left|\hat{y} - \hat{y}_i^{(-i)}\right|}{\sigma\sqrt{h_{ii}}} = \frac{\left|x_i^T(\hat{\beta} - \hat{\beta}^{(-i)})\right|}{\sigma\sqrt{h_{ii}}} \; . \qquad (9)$$

The denominator is just standardization, since $Var\left(\hat{y}_i\right) = \sigma^2 h_{ii}$. Authors [1] suggested using $\hat{\sigma}^{(-i)}$ as an estimate of σ in (9). Belsley *et al.* [1] called Welsch-Kuh's (WK) distance as DFFITS, because it is the scaled difference between $\hat{y}$ and $\hat{y}^{(-i)}$. They [1] suggested that any observation for which $\left|DFFITS_i\right| > 2/\sqrt{n}$ warrants attention.

Hadi's Influence Measure

Hadi [6] proposed a measure of the influence of the *i*th observation based on the fact that influential observations are outliers in either the response variable or in the predictors, or both. Accordingly, the influence of the *i*th observation can be measured by

$$H_i = \frac{h_{ii}}{1-h_{ii}} + \frac{p+1}{1-h_{ii}} \frac{d_i^2}{1-d_i^2}, \qquad i = 1,2,\ldots,n \qquad (10)$$

where $d_i = r_i / \sqrt{SSE}$ (SSE: sum squares residual) is the so-called normalized residual. It can be seen that observations will have large values of H_i if they are outliers in the response and/or the predictor variables, that is, if they have large values h_i, r_i, or both.

Most of the popular diagnostic methods are based on the measures of sensitivity of a single observation at a time. Single diagnostic methods are failed to detect the unusual observations in presence of masking and/or swamping phenomena.

ii) In case of group deletion diagnostics, it is a cumbersome and sometimes impossible to identify suspect group of unusual observations because of the sample size.

iii) Diagnostic techniques those do not consider robustness can make non-robust decision.

iv) It is hard to derive exact/asymptotic distribution of estimates obtained after classical diagnostic methods.

v) Existence of several regression diagnostic techniques with little guidance available as to which is appropriate.

These are based on single case deletion and unable to cope with multiple unusual observations. There are some group deletion measures (see [7]) but they are failed in case of large data sets. The problem for group deletion diagnostics is to find group of outlying observation *d*, the deletion of which produces the largest reduction to the residual sum of squares.

IV. New Diagnostic Measure

Nurunnabi [12] introduced a two-step diagnostic measure originated from Pena's [13] idea attaching with formal and/or informal group deletion methods (see [7]). Group deletion helps us to reduce the maximum disturbance by deleting the suspect group of influential cases at a time and 'the deletion of which produces the largest reduction in the residual sum of squares. It helps to make the data more homogeneous. We look how every point in a sample is influenced by the specific (*i*th) observation. In our method, after deleting the suspect group at a time we again delete specific sample point from each of the observations one after another and measure the difference between the forecast with and without the specific one. Group deletion attaching with Pena's [13] idea improves the measure and show nice results in presence of masking and/or swamping phenomena. Graphical displays like index plot and character plot of explanatory and response variables could give us some ideas about the influential observations, but these plots are not useful for higher dimension of regressors. There are some suggestions in the literature for using robust regression techniques like LMS, LTS, reweighted least squares (RLS) for finding the group of suspect influential

observations. We consider one of the two proposed procedures of Hadi and Simonoff [7], where they suggested dividing the data set in two initial subsets: a 'basic' subset that contains the first k+1 clean observations and a 'nonbasic' subset that contains the remaining observations.

At the first step we find out all suspect influential (unusual) cases. We suggest using graphical display and/or regression diagnostic measures and/or robust regression methods whatever can helps to find the suspect group with maximum number of influential cases.

In the second step, we assume that d observations among a set of n observations are suspected as influential observations and to be deleted. Let us denote a set of cases 'remaining' in the analysis by R and a set of cases 'deleted' by D. Hence without loss of generality, assume that these observations are the last d rows of X and Y so that

$$X = \begin{bmatrix} X_R \\ X_D \end{bmatrix} \qquad Y = \begin{bmatrix} Y_R \\ Y_D \end{bmatrix}.$$

After formation of the deletion set indexed by D, we would compute the fitted values $\hat{Y}^{(-D)}$. Let $\hat{\beta}^{(-D)}$ be the corresponding vector of estimated coefficients when a group of observations indexed by D is omitted. We define the vector of difference between $\hat{y}_j^{(-D)}$ and $\hat{y}_{j(i)}^{(-D)}$ as

$$t_{(i)}^{(-D)} = (\hat{y}_1^{(-D)} - \hat{y}_{1(i)}^{(-D)}, ..., \hat{y}_n^{(-D)} - \hat{y}_{n(i)}^{(-D)})^T \quad (11)$$

$$= (t_{1(i)}^{(-D)}, ..., t_{n(i)}^{(-D)})^T \quad (12)$$

and

$$t_{j(i)}^{(-D)} = \hat{y}_j^{(-D)} - \hat{y}_{j(i)}^{(-D)} = \frac{h_{ji}\hat{\varepsilon}_i^{(-D)}}{1-h_{ii}}, j = 1,2,...,n \quad (13)$$

where

$h_{ji} = x_j^T (X^T X)^{-1} x_i$ and $\hat{\varepsilon}_i^{(-D)} = y_i - \hat{y}_i^{(-D)}$.

Finally, we introduce our new measure as squared standardized norm,

$$M_i = \frac{t_{(i)}^{(-D)^T} t_{(i)}^{(-D)}}{kV(\hat{y}_i^{(-D)})}, \quad (14)$$

where

$$V(\hat{y}_i^{(-D)}) = s^2 h_{ii} \qquad and \qquad s^2 = \frac{\hat{\varepsilon}^{(-D)^T}\hat{\varepsilon}^{(-D)}}{n-k}.$$

Using (11) to (14), we obtain

$$M_i = \frac{1}{ks^2 h_{ii}} \sum_{j=1}^{n} h_{ji}^2 \frac{\hat{\varepsilon}_i^{(-D)2}}{(1-h_{ii})^2}. \quad (15)$$

We use a confidence bound type cut-off value for it, and consider the ith observation to be influential if it satisfies the rule

$$|M_i| > median(M_i) + 4.5MAD(M_i), \quad (16)$$

where

$$MAD(M_i) = median\{|M_i - median(M_i)|\}/0.6745.$$

V. Numerical Examples

We consider here a large data set. We want to show the performance of our new measure and compare with the existing popular diagnostic methods.

A Real Large data

Our example uses real data set that is taken from the engineering literature [17] and collected from the author himself. This data set was assembled for concrete containing cement plus, fly ash, blast furnace, slag, and a superplasticizer. It is an 8 predictor data set with one output variable. Total number of cases is 1030. Variables are defined as: Cement (X1), Blast Furnace Slag (X2), Fly Ash (X3), Water (X4), Superplasticizer (X5), Coarse Aggregate (X6), Fine Aggregate (X7), Age (Day 1~365; X8) Concrete compressive strength , MPa (Y; Output Variable). Results that we obtain from this real data are shown in Fig. 1. The residual versus fitted value plot, Fig. 1 (e) shows no indication of heterogeneity. We apply LMS robust method and get 96 as suspect cases, we make deleted group by putting them at the end of the design matrix as our algorithms. The last 96 observations are deleted, the deleted residuals versus deleted fitted values plot and the histogram of deleted residual (Fig. 1. (d), (f)) show clear indications of the presence of heterogeneity. We consider above mentioned three measures Cooks distance, DFFITS and Pena's S_i to identify the influential cases but their index plots (Fig. 1. (h), (j) and (l)) show they are totally failed to identify the influential cases. Index plot, Fig 1. (n), we see that proposed M_i can successfully identify 151 influential cases. Fig.1. (o), histogram of M_i also shows the heterogeneous variances in the data set that is absent in histogram of S_i.

VI. Conclusion

In this article, we propose a group deletion diagnostic measure for identifying outliers (influential observations) for large data in linear regression. The technique gives some clear indications of presence of heterogeneous variances in the data set when the existing popular diagnostic methods are not well enough. The new measure can helps us to outlier diagnostics in the fields that are involved in model building process with large data sets.

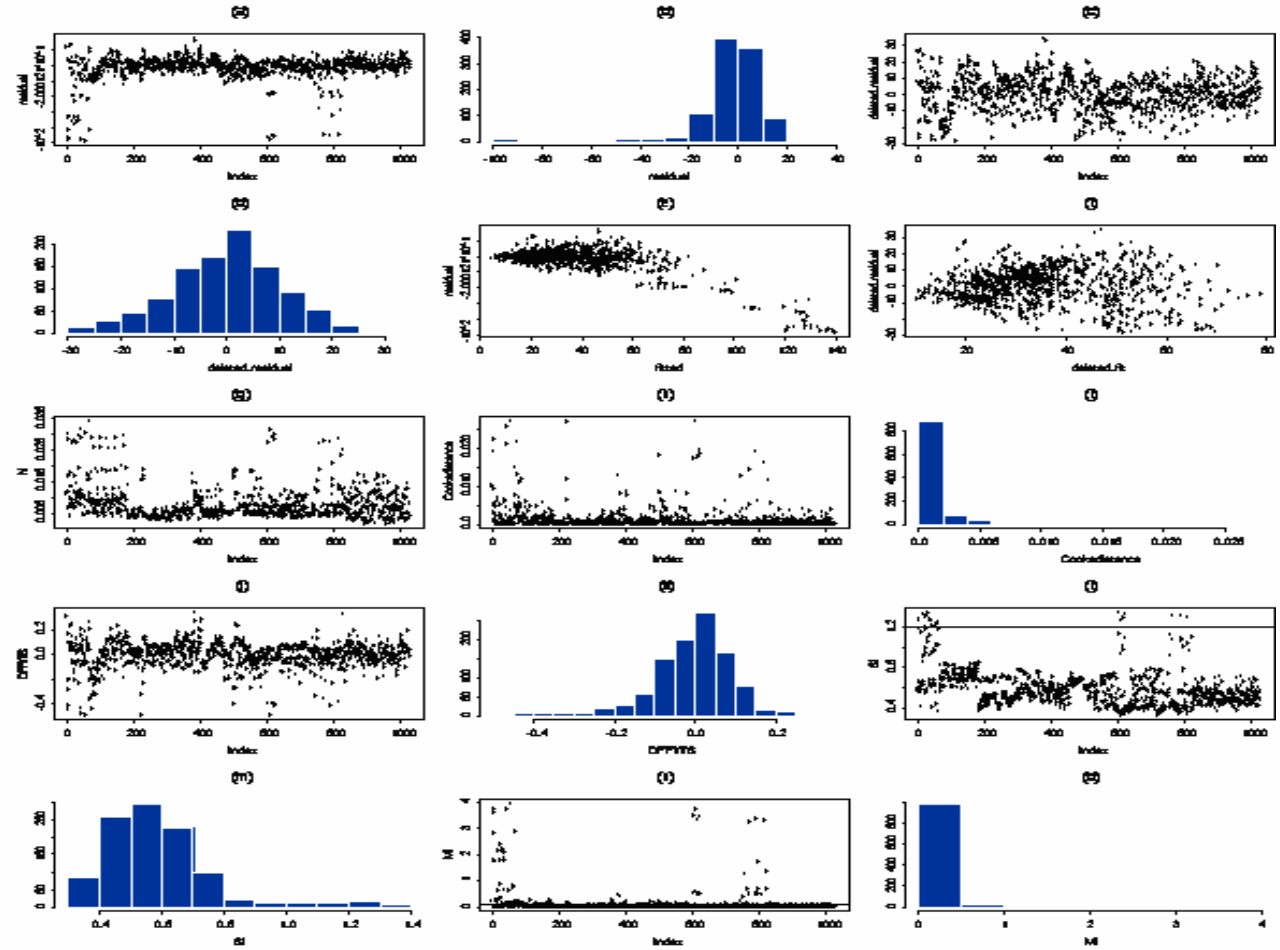

Figure 1. Influence analysis of the large real data; (a) Index plot of residuals (b) Histogram of the residuals (c) Index plot of deleted residuals (d) Histogram of the deleted residuals (e) Plot of residual versus fitted value (f) Plot of deleted residual versus deleted fitted values (g) Index plot of leverage values (h) Index plot of Cook's distance (i) Histogram of Cooks distance (j) Index plot of DFFITS (k) Histogram of DFFITS (l) Index plot of Si (m) Histogram of Si (n) Index plot of Mi, and (o) Histogram of Mi.

REFERENCES

[1] D. A. Belsley, E. Kuh, and R. E. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, New York: Wiley & Sons, 1980.

[2] G.E.P. Box, "Non-normality and Tests on Variance", Biometrika, 40, pp. 318-335, 1953.

[3] S. Chatterjee, and A. S. Hadi, Sensitivity Analysis in Linear Regression, New York: Wiley & Sons, 1988.

[4] R. D. Cook, "Detection of Influential Observations in Linear Regression", Technometrics, Vol.19, pp.15-18, 1977.

[5] B. Efron, Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics,7, pp. 1-26, 1979.

[6] A. S. Hadi, "A New Measure of Overall Potential Influence in Linear Regression", Computational Statistics and Data Analysis, Vol.14, pp.1-27, 1992.

[7] A. S. Hadi, and S. Simonoff, "Procedure for the Identification of Outliers in Linear Models," Journal of American Statistical Association, Vol. 88, pp. 1264-1272, 1993.

[8] F. R. Hampel, Contribution to the Theory of Robust Estimation, Ph. D. Thesis, Univ. of California, Berkley. 1968.

[9] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and Stahel, W. A. Robust Statistics: The Approach Based on Influence Function, New York: Wiley & Sons, 1986.

[10] P. J. Huber, Robust Estimation of a Location Parameter, Annals of Mathematical Statistics, Vol. 35, pp.73-101, 1964.

[11] P. J. Huber, Robust Regression: Asymptotics, Conjectures and Monte Carlo, Annals of Statistics, Vol.1, pp. 99-821, 1973.

[12] A. A. M. Nurunnabi, Robust Diagnostic Deletion Techniques in Linear and Logistic Regression, M. Phil. Thesis, Unpublished, Rajshahi University, Bangladesh, 2008.

[13] D. Pena, "A new statistic for influence in linear regression," Technometrics, Vol. 47, pp.1-12, 2005.

[14] P. J. Rousseeuw, Least Median of Squares Regression. Journal of American Statistical Association. Vol. 79, pp. 871-880, 1984.

[15] P. J. Rousseeuw, and A. M. Leroy, Robust Regression and Outlier Detection, New York: Wiley & Sons, 1987.

[16] J. W. Tukey, A Survey of Sampling from Contaminated Distributions: Contributions to Probability and Statistics, Olkin, I. Ed., Stanford University, California. 1960.

[17] I-C. Yeh, "Modeling of Strength of High Performance Concrete Using Artificial Neural Networks", Cement and Concrete Research, Vol. 28. (12), pp. 1797-1808, 1998.