Building a Robust Explanatory Model from a Large Data Set

Dave Kim
Department of Mathematical Sciences
U.S. Military Academy, West Point, New York

Barry Bodt
Computational and Information Sciences Directorate
US Army Research Laboratory
Aberdeen Proving Ground, MD

1 Executive Summary

Explosive growth in the size of modern data sets is found not only in the number of records but also in the number of variables each record carries. Building any kind of predictive or explanatory model from such a data set presents a novel challenge. We present a method of building an explanatory model based on David Scott's parametric L2E criterion. The empirical evidence suggests that the method exhibits a feature which enables presence of outliers to be easily detected and can also be used to select explanatory variables with (very much counter-intuitively) less ambiguity when the outliers are present than otherwise. We apply the method to develop plan evaluation metrics through analyses of OneSAF simulation runs.

2 Introduction

The enormous changes in the very notion of data that have come about lately pose many new and exciting challenges to statisticians. Primarily, most of these novel problems have to do with the sheer size of modern data sets that rapid advances in technology enable us to have. Recent works in the budding field of data mining is a good example illustrating the challenges statisticians face in dealing with large data sets. The enormity of the modern data sets forces the extension and adaptation of statistical methods starting from exploratory data analysis procedures and beyond [1].

There are two important requirements for a viable data mining method: speed and robustness. The need for speed is obvious in that any lack of speed in analysis will be greatly amplified for larger data sets, and it usually compels one to focus more on the algorithm than on the statistical model. In addition, any large data set will invariably have imperfections and its enormity will complicate any preprocessing that may be needed to clean it up. Therefore, procedures for building a statistical model that can describe a large data set with reasonable speed and sufficient robustness to resist inevitable data corruption are needed more than ever.

Rapid advance in hardware technology is an inseparable companion in this endeavor, and it can be argued that the robustness of the algorithm then takes more prominence than otherwise. A recently proposed framework for data analysis by David Scott [2] gives us a promising candidate for a robust algorithm suitable for analyzing large data sets. Scott [2] proposed minimizing the integrated squared error as an encompassing paradigm in which a large variety of

simple and complex parametric models can be built with robustness. On a space of square integrable functions, the integrated square error (ISE)

$$ISE = \int [\hat{f}(x) - f(x)]^2 dx, \tag{1}$$

where $f(x)$ is the true (and usually unknown) underlying density and $\hat{f}(x)$ is a nonparametric estimator of $f(x)$, is a time tested optimality measure in nonparametric curve estimation, for example, as a starting point in development of least squares cross validation algorithm of Bowman [3] and Rudemo [4]. An attractive property of ISE is that an equivalent functional to be minimized can be written empirically (based on iid observations $x_1,...x_n$) as--calling the equivalent functional L_2E

$$L_2E = \int \hat{f}(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}(x_i). \tag{2}$$

The minimization of this functional is done over a suitable class of functions. Scott [2] shows this criterion can be used in parametric settings if the minimization is done over a parametric family, and the resulting estimators have robustness properties that are known to be possessed by the class of minimum distance estimators.

The L_2E criterion given in Eq. (2) can be used in fitting a class of models

$$y = f(x) + \varepsilon \tag{3}$$

where x may be a vector in R^d and a parametric distribution, e.g. $N(0,\sigma^2)$, for ε may be assumed. Of course, simple and multiple regression models belong to the class. In the next section, a theoretical development of L_2E regression, both simple and multiple, is given, with an attempt at automatic variable selection method given at the end.

3 *L_2E* Regression

As was mentioned in the introduction, any model in the form of Eq. (3), whether it is linear or nonlinear in parameters, can be fitted by minimizing the L_2E functional given in Eq. (2). In this report, we will concentrate on a linear regression model which may be expressed as

$$y = b_0 + b_1x_1 + \cdots + b_kx_k + \varepsilon \tag{4}$$

and we assume ε is normally distributed with mean 0 and variance σ^2. We then have for the density of the error component in the model

$$f(\varepsilon) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\varepsilon^2}{2\sigma^2} \right). \tag{5}$$

Solving the Eq. (4) for ε and substituting into Eq. (5), we can write the L_2E functional to be minimized as

$$L_2E = \frac{1}{2\sqrt{\pi}\sigma} - \frac{2}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2} \right) \tag{6}$$

where $\varepsilon_i = y_i - b_0 - b_1 x_{1,i} - \cdots - b_k x_{k,i}$ for $i = 1, \ldots, n$.

Clearly, it is not quite practical, to say the least, to write down the analytic solutions to this nonlinear minimization problem, but most of the nonlinear minimization routines built into many popular packages can easily be adapted to give us the numerical solution. All of the numerical work done in this report was done using the statistical program R.

3 Variable Selection Using L_2E Estimates of σ

The robustness of the L_2E estimate of σ, which is estimated by $\sqrt{MSE}$ in the least squares theory, leads us to conjecture that robust variable selection may be implemented if we use the estimate of σ as the criterion in variable selection. In the least squares theory, *MSE* is an unbiased estimator of σ^2 and is also used as a criterion in variable selection (see, for example, [5]). However, as was illustrated in the previous section, *MSE* is not robust in that a few points can easily influence the estimate. On the other hand, if *MSE* may be used in variable selection in the least squares framework, a more robust equivalent of *MSE* may lead to a more robust variable selection scheme.

Let us look at a simulated example to see how we can incorporate L_2E estimate of σ in variable selection. A data set of 100 points was simulated from the true model (with no contamination) $y = 3 - x_2 + 4x_3 - 7x_4$, and σ = 0.2. As in a forward selection procedure, a simple regression model was fit using each of $x_1, \ldots, x_4$, and the values of s_L, denoting the L_2E estimate of σ, were obtained and compared (see Table 1).

Table 1: L_2E estimates of s_L with different explanatory variables. Each variable was the only explanatory variable, yielding a simple regression model.

Explanatory variable	s_L from the model with given exp. var.
x_1	18.19
x_2	16.86
x_3	18.52
x_4	9.982

The result seen in Table 1 strongly suggests that we select x_4 as the first variable to be entered into the model. Now what happens as we add the remaining three variables one by one to the model with x_4 and compare the resulting s_L? Table 2 shows that adding x_3 results in the marked reduction in s_L.

Table 2: L_2E estimates of s_L with different explanatory variables when x_1, x_2, and x_3 were added to the model equation.

Explanatory variable	s_L from the model with given exp. var.
x_1 , x_4	9.86
x_2 , x_4	9.87
x_3 , x_4	2.69

Choosing the model with x_3 and x_4 using the "s_L criterion," we now compare what happens to "s_L" as we add x_1 and x_2. As Table 3 demonstrates, seeking the reduction in s_L helps us reach the correct model ($y = 3-x_2+4x_3-7x_4$, containing x_2, x_3, and x_4).

Table 3: L_2E estimates of s_L with different explanatory variables when x_1 and x_2 were added to the model equation which already contained x_3 and x_4.

Explanatory variable	s_L from the model with given exp. var.
x_1 , x_3 , x_4	9.80
x_2 , x_3 , x_4	0.19

Now, when the simulated data set is not contaminated (i.e.: no deliberately inserted outliers), addition of the extraneous explanatory variables results in not much change in s_L which is also the case when the least squares estimate of σ^2 is used. However, something remarkable happens when we use s_L , estimated using L2E criterion, from a contaminated data set. In such a case, addition of an extraneous variable results in the increase of s_L. Table 4 shows what happens when the same true model as above was used to simulate a data set with 20% contamination.

Table 4: L_2E estimates of s_L of a model with correct subset of explanatory variables and another with an extra variable when the simulated data set was 20% contaminated.

Explanatory variable	s_L from the model with given exp. var.
x_2 , x_3 , x_4	0.19
x_1, x_2 , x_3 , x_4	6.73

This suggests an unambiguous stopping rule: stop adding variables to the model equation when s_L increases. In summary, the following heuristic variable selection rule is proposed.

1. Choose the variable that gives the smallest s_L at each forward selection step.
2. Stop when adding any of the remaining variables gives no decrease in s_L.

The models with the appropriate selection of covariates were arrived at in many different models and at reasonable contamination levels.

4 Applications

In this section, we apply the heuristic variable selection method outlined in the previous section to "develop plan evaluation metrics" (i.e. to find explanatory variables with which we can build a multiple regression model explaining interesting features of the executed plan) through analyses of One Semi Automated Forces (OneSAF) simulation runs of an armor engagement. The scenario was based on the engagement of a "blue" armor company against "red" units situated along the way to the objective of the "blue" company. The original data set contained 143 variables at each of three time slices containing individual vehicle-level information for 228 OneSAF runs, resulting in 228×430 data matrix–where 430 is 143×3+1, the last column corresponding to the indicator variable for the "blue" mission success. For example, the data set contained the number of 105 SABOT round hits from the second "blue" platoon on "red" T80 at each of three time slices: N05ST2S1, N05ST2S2, and N05ST2S3. These three are of course heavily correlated with one another.

One of the interesting questions in analyzing such a simulation data is whether we can find a battle metric at an early stage that is very much related to the battle outcome. To that end, we examined only the variables corresponding to the first time slice, giving us 228x143 data matrix plus a column for the response variable, making it at least theoretically possible to fit a linear model including all possible covariates. Since we are going to use a variable selection method for a regression model, we need a quantitative score that has a strong bearing to the mission success as the response variable. There are two potential response variables in the data set: MBTSCORE and ERICSCOR. Figure 1 shows that indeed higher scores of these two in general correspond to the mission success (MA=1).

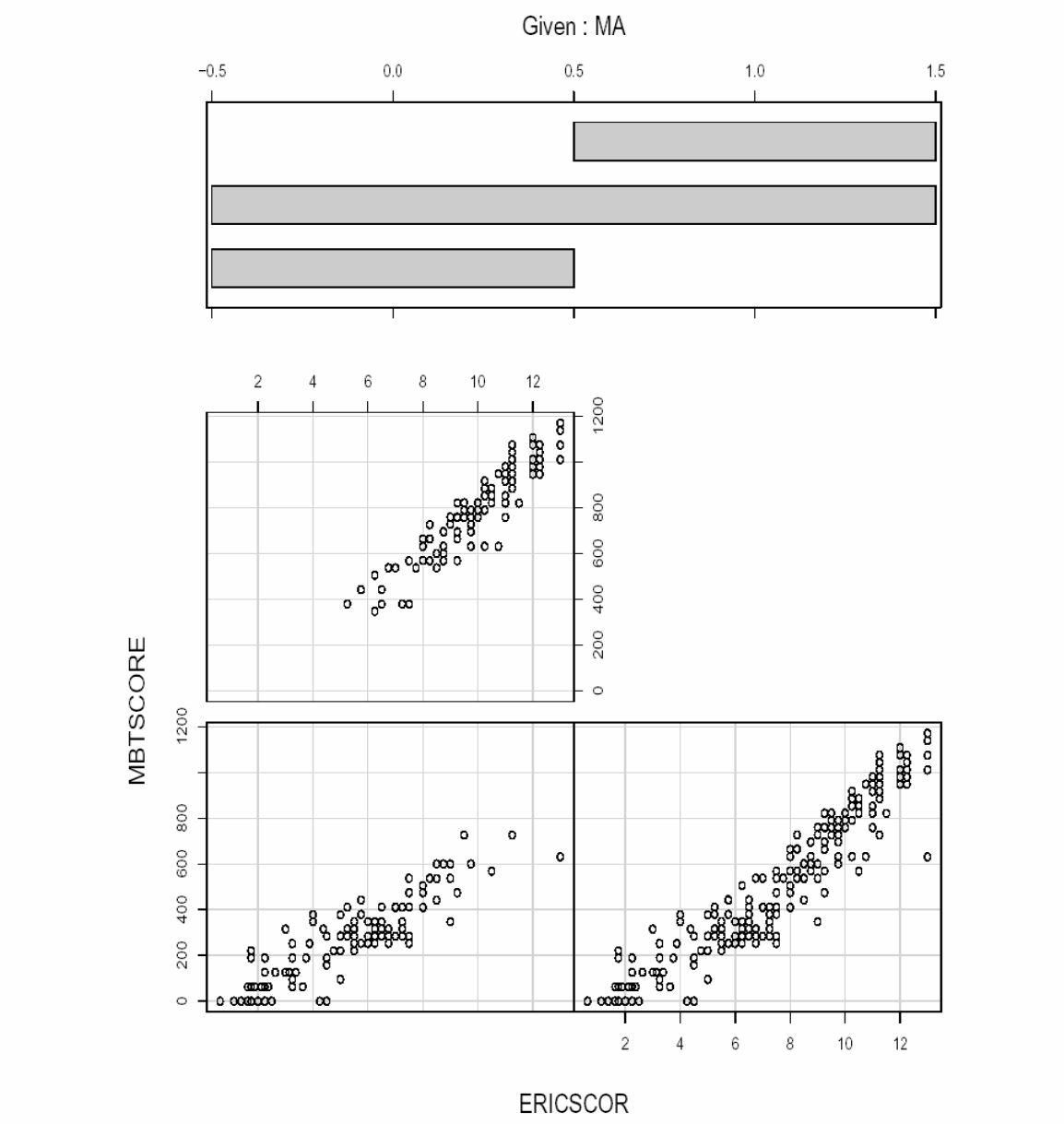

Figure 1: Conditioning plot of MBTSCORE vs. ERICSCOR with the mission success (MA) as the conditioning variable.

One thing of note, however, is that even as both quantitative measures of the mission success show strong correlation with each other and MA, the indicator variable for the mission success, there are a couple of points where ERICSCOR is high; MBTSCORE low; and it does not correspond to the mission success.

Upon the examination of the entire data set, we found that this is the reflection of the fact that ERICSCOR is the measure only of the advancement of the blue units regardless of their condition, whereas the mission success does require that the blue units reach their objective in

reasonable condition. Therefore, we chose MBTSCORE as the response variable in the regression model to be explored in the data.

Table 5 summarizes the result of the heuristic L2E variable selection method described previously, and what follows is the explanation of what each of the coded variable measures:

- MF3S1: The number of tanks in the third platoon that are mobility and firepower killed.
- F3S1: The number of tanks in the third platoon that are firepower killed.
- NKB1S1: The number of the red BMP catastrophic kills by the first platoon.
- R05HB2S1: The average range of 105 HEAT engagement on red BMPs by the second platoon.
- MF1S1: The number of tanks in the first platoon that are mobility and firepower killed.
- N05HB3S1: The number of 105 HEAT round hits on red BMPs by the third platoon.
- N05ST2S1: The number of 105 SABOT round hits on red T80s by the second platoon.

Table 5: L_2E estimates of the final multiple regression model (without any interaction) containing the variables selected using the heuristic method described in the previous section.

Selected explanatory variables	L_2E Coefficients
Intercept	584.01
MF3S1	-127.01
F3S1	-108.10
NKB1S1	111.64
R05HB2S1	0.047
MF1S1	-152.05
N05HB3S1	26.36
N05ST2S1	-51.03

A very interesting feature in the model arrived here is that the mission success and N05ST2S1 have a negative association. At first glance, this is somewhat counter-intuitive until one looks at the layout of the simulation scenario. The second platoon in the scenario actually follows another platoon, and if the number of hits on the red T80 by this following platoon is high at the time slice 1, an initial stage of the engagement, the implication is that the leading platoon has been rendered somewhat ineffective, hence potentially leading to the mission failure.

References

[1] Hand, D. J. and Blunt, G. and Kelly, M. G. and Adams, N. M., "Data Mining for Fun and Profit," *Statistical Science*, Vol. 15, No. 2, 2000, pp. 111-126.

[2] Scott, D. W., "Parametric Statistical Modeling by Minimum Integrated Square Error," *Technometrics*, Vol 43, No. 3, 2001, pp. 274-285.

[3] Bowman, A. W., "An Alternative Method of Cross-validation for the Smoothing of Density Estimates," *Biometrika*, Vol. 71, 1984, pp. 353-360.

[4] Rudemo, M., "Empirical Choice of Histograms and Kernel Density Estimators," *Scandinavian Journal of Statistics*, Vol. 9, 1982, pp. 65-78.

[5] Draper, N. R. and Smith, H., *Applied Regression Analysis*, Third Ed., 1998, John Wiley & Sons, New York.