

# Robust Modeling Based on $L_2E$ Applied to Combat Simulation Data

David B. Kim  
Department of Mathematical Sciences  
United States Military Academy  
West Point, NY 10996

March 15, 2003

## **Abstract**

Parametric modeling based on a minimum distance criterion gives us a robust (i.e. resistant to outliers and other types of data contamination) way of analyzing the data. The integrated squared error ( $L_2E$ ) of David Scott is one such criterion, and a system of fitting a model to the data and assessing the model can be built upon it. Combat simulation data containing numerous variables may benefit from such a robust method of analysis since it is quite likely that some of the assumptions for building, for example, a least squares linear model may be violated in the data.

We will outline the development of such a system and demonstrate the new methodologies via the analysis of simulation results from One Semi Automated Forces (OneSAF). Specifically, a heuristic variable selection method based on  $L_2E$  is used in the analysis to provide models identifying key battle parameters for a given engagement.

# 1 Introduction

Statistics is often called the science of the data. It then should come as no surprise that 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 (Hand et al. 2000).

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 (2001) gives us a promising candidate for a robust algorithm suitable for analyzing large data sets. Scott (2001) 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)

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

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 (1984) and Rudemo (1982). An attractive property of ISE is that an approximately equivalent functional to be minimized can be written empirically (based on iid observations  $x_1, \dots, x_n$ ) as—calling the equivalent functional  $L_2E$

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

The minimization of this functional may then be done over a suitable class of functions. Scott (2001) 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

$$y = f(x) + \epsilon \quad (3)$$

where  $x$  may be a vector in  $R^d$  and a parametric distribution for  $\epsilon$ , for example  $N(0, \sigma^2)$ , may be assumed. Of course, simple and multiple regression models belong to the class. In the next section, a brief summary of the theoretical development of  $L_2E$  regression, both simple and multiple, is given, which is followed by a section on the description and motivation by an example of a heuristic variable selection method. This method will be applied to OneSAF data set in the final section.

## 2 $L_2E$ 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 functional given in Eq. (2). In this report, we will concentrate on a linear regression model which may be expressed as

$$y = b_0 + b_1x_1 + \dots + b_kx_k + \epsilon \quad (4)$$

and we assume  $\epsilon$  is normally distributed with mean 0 and variance  $\sigma^2$ . We then have for the density of the error component,  $\epsilon$  in the model

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

Solving Eq. (4) for  $\epsilon$  and substituting into Eq. (5), we may write the  $L_2E$  functional to be minimized as

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

where  $\epsilon_i = y_i - b_0 - b_1x_{1,i} - \dots - b_kx_{k,i}$  for  $i = 1, \dots, n$ .

	True Values	Least Squares Estimates	$L_2E$ Estimates
$b_0$	0	0.0006	0.0104
$b_1$	-1	-1.0050	-1.0068
$b_2$	2	1.9982	2.0013
$b_3$	-1	-1.0015	-1.0100
$b_4$	2	2.0020	2.0022
$\sigma$	0.2	0.2055	0.2182

Table 1: Comparison of the parameter estimates from the least squares regression and  $L_2E$  regression (no contamination).

	True Values	Least Squares Estimates	$L_2E$ Estimates
$b_0$	0	0.0139	-0.0357
$b_1$	-1	0.1592	-1.0286
$b_2$	2	1.6500	2.0137
$b_3$	-1	0.1679	-1.0005
$b_4$	2	2.0401	2.0110
$\sigma$	0.2	8.65	0.2118

Table 2: Comparison of the parameter estimates from the least squares regression and  $L_2E$  regression (20% contamination).

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

As was noted by Scott (2001), the parameter estimates from minimizing Eq. (6) are robust in the sense that they are not as easily affected by the presence of outliers as the least squares estimates. This will be illustrated in the following simulated example. Consider the model with four covariates ( $k=4$  in Eq. (4)). The Tables 1 and 2 compare the usual least squares estimates of the parameters with  $L_2E$  estimates for two simulated data sets, one without any contamination and the other where 20% of the data were simulated from a different model ( $n=100$  for both cases.).

As is illustrated in Tables 1 and 2, the estimate of  $\sigma$  as well as the estimates of other model parameters are robust. In addition to the robustness,  $L_2E$  estimates follow asymptotic normal distributions under mild conditions Scott (2001).

Explanatory variable	$\hat{\sigma}$ from the model with given exp. var.
$x_1$	18.19
$x_2$	16.86
$x_3$	18.52
$x_4$	9.982

Table 3:  $L_2E$  estimates of  $\sigma$  with different explanatory variables. Each variable was the only explanatory variable, yielding a simple regression model.

Explanatory variables	$\hat{\sigma}$ 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

Table 4:  $L_2E$  estimates of  $\sigma$  with different explanatory variables when  $x_1$ ,  $x_2$ , and  $x_3$  were added to the model equation.

### 3 Toward Robust Variable Selection

The robustness of the  $L_2E$  estimate of  $\sigma$ , 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  $\sigma$  as the criterion in variable selection. In the least squares theory,  $MSE$  is an unbiased estimator of  $\sigma^2$  and is also used as a criterion in variable selection (see, for example, Draper and Smith (1998)). 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  $\sigma$  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  $\sigma = .2$ . As in a forward selection procedure, a simple regression model was fit using each of  $x_1, \dots, x_4$ , and the values of  $\hat{\sigma}$ , denoting the  $L_2E$  estimate, were obtained and compared (see Table 3).

The result seen in Table 3 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  $\hat{\sigma}$ ? Table 4 shows that adding  $x_3$  results in the marked reduction in  $\hat{\sigma}$ .

Choosing the model with  $x_3$  and  $x_4$  using “ $\hat{\sigma}$  criterion,” we now compare what happens to “ $\hat{\sigma}$ ” as we add  $x_1$  and  $x_2$ . As Table 5 demonstrates, seeking the reduction in  $\hat{\sigma}$  helps us reach the correct model ( $y = 3 - x_2 + 4x_3 - 7x_4$ , containing

Explanatory variables	$\hat{\sigma}$ from the model with given exp. var.
$x_1, x_3, x_4$	9.80
$x_2, x_3, x_4$	0.19

Table 5:  $L_2E$  estimates of  $\sigma$  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 variables	$\hat{\sigma}$ 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

Table 6:  $L_2E$  estimates of  $\sigma$  of a model with correct subset of explanatory variables and another with an extra variable when the simulated data set was contaminated.

$x_2$ ,  $x_3$ , and  $x_4$ ). 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  $\hat{\sigma}$ , which is also the case when the least squares estimate of  $\sigma^2$  is used. However, something remarkable happens when we use  $\hat{\sigma}$  estimated using  $L_2E$  criterion and use a contaminated data set. In such a case, addition of an extraneous variable results in the increase of  $\hat{\sigma}$ . Table 6 shows what happens when the same true model as above was used to simulate a data set with 20% contamination.

This suggests an unambiguous stopping rule: stop adding variables to the model equation when  $\hat{\sigma}$  increases. In summary, the following heuristic variable selection rule is proposed.

1. Choose the variable that gives the smallest  $\hat{\sigma}$  at each forward selection step.
2. Stop when adding any of the remaining variable gives no decrease in  $\hat{\sigma}$ .

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

## 4 Application to OneSAF Data

In this section, we present a result of the application of the variable selection method described above to a data set generated from One Semi Automated Forces (OneSAF). 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  $228 \times 430$  data matrix—where 430 is  $143 \times 3 + 1$ , 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  $228 \times 143$  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).

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 7 summarizes the result of the heuristic  $L_2E$  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.

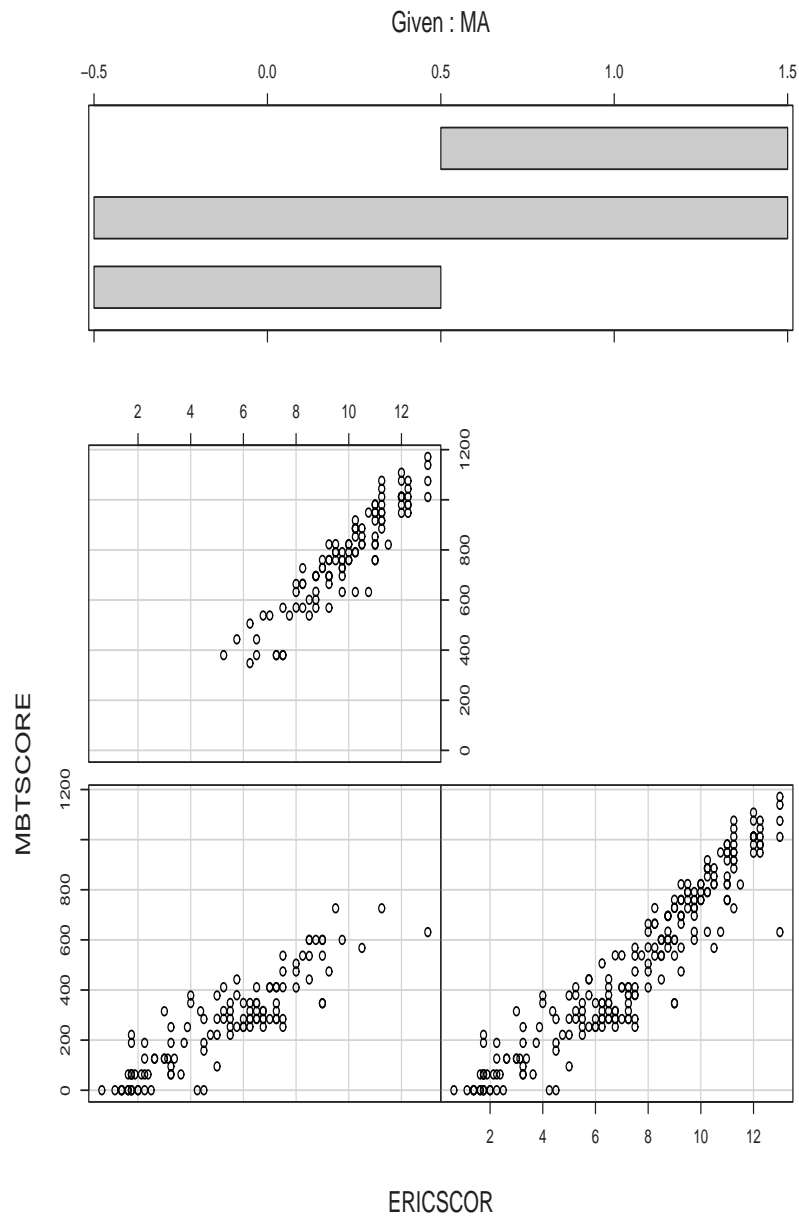


Figure 1: Conditioning plot of MBTSCORE vs. ERICSCOR with the mission success (MA) as the conditioning variable.

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

Table 7:  $L_2E$  estimates of the final multiple regression model (without any interaction) containing the variables selected using the heuristic method described in the previous section.

- 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.

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

- Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* 71(2), 353–360.
- Draper, N. R. and H. Smith (1998). *Applied regression analysis* (Third ed.). New York: John Wiley & Sons Inc.
- Hand, D. J., G. Blunt, M. G. Kelley, and N. M. Adams (2000). Data mining for fun and profit. *Statistical Science* 15(2), 111–126.
- Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* 9(2), 65–78.

Scott, D. W. (2001). Parametric statistical modeling by minimum integrated square error. *Technometrics* 43(3), 274–285.