L2E Modeling of a Binary Response Variable

David Kim
Department of Mathematics and Computer Science
Manhattan College, Riverdale, New York

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

1 Executive Summary

We present an alternative way of fitting a binary response regression model using David Scott's L2E criterion. Application of the L2E criterion to estimation of the mean of independent and identically distributed (iid) Bernoulli random variables leads to the sensible closed form solution. Then we argue how the L2E criterion for iid case may be adapted to a non-trivial binary response regression model. The values of the parameters estimated from the L2E criterion agree with maximum likelihood estimates.

2 Introduction

Much of statistical modeling seeks the relationship between explanatory and response variables. When a single response variable is of continuous kind, a simple linear regression model is often used to examine the relationship between a single explanatory variable and a continuous response variable. When there are several explanatory variables, multiple regression models are used. There are, however, many cases where the response is not of continuous kind. Instead, the response is simply a designation of one of two possible outcomes, a binary response, e.g., success or failure, head or tail, 1 or 0. A binary response regression model is an umbrella term that includes logistic, probit, and other regression models that have been developed to explore the relationship between a binary response variable and one or more explanatory variables.

Simple linear regression and multiple regression models may be considered in terms of classical linear models:

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

where Y is a column vector of dimension n; X an n by $(k+1)$ matrix; β the vector of $k+1$ dimensions; and e is the n dimensional vector of errors, which are usually taken as independent and identically distributed Gaussian random variables. Clearly, β is the vector of the coefficients in a multiple regression model with k explanatory variables. In addition, we see that polynomial regression models may also be expressed using Eq. (1) by setting

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^k \\ 1 & x_2 & x_2^2 & \cdots & x_2^k \\ \vdots & \vdots & \ddots & & \vdots \\ 1 & \vdots & & \ddots & \vdots \\ 1 & x_n & \cdots & \cdots & x_n^k \end{bmatrix} \tag{2}$$

for a model $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots \beta_k x_i^k + e_i$, for example.

The extension of a classical linear model to more general settings that encompass the binary response variables (among other things) is called a generalized linear model. First note that we may rewrite Eq. (1) as

$$E(Y) \equiv \mu = X\beta. \tag{3}$$

If Y is a vector of binary outcomes, its expected values will range between 0 and 1. The linear part of Eq. (3) will range over the entire real line. We may maintain the linearity of the model over the parameters by one-to-one transforming the expected values onto the entire real line using a link function

$$g(\mu_i) = X[i,]\beta. \tag{4}$$

where $X[i,]$ is the ith row of the design matrix X. Such a model is called a generalized linear model. Using the identity function as the link, $g(\mu) = \mu$, we get the classical linear models. Logistic regression models are obtained when we use the logistic link, $g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$, and probit models are obtained when we use the probit link, where the link function is the inverse cumulative distribution function of a Gaussian random variable with mean 0 and standard deviation 1. The estimation of the parameters is usually done by maximum likelihood which leads to iterated numerical procedure. [1]

Logistic regression models have several useful properties which explain its wide popularity. First, there is a useful interpretation of the regression coefficients in a logistic model as the change in log odds. Second, the differences on the logistic scale may be estimated whether the data are sampled prospectively or retrospectively, that is, before the fact or after the fact, so to speak—this is property of the logistic link function not shared by other link functions [1].

3 L2E for iid Bernoulli Data

David 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)

$$ISE = \int [\hat{f}(x) - f(x)]^2 dx, \tag{5}$$

where $f(x)$ is the true (and usually unknown) underlying probability density function (pdf) 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 *L2E*

$$L2E = \int \hat{f}(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}(x_i). \tag{6}$$

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. For discrete data, $y_1, \ldots, y_n$, iid $p(y)$, Eq. (6) becomes

$$L2E = \sum_y \hat{p}^2(y) - \frac{2}{n}\sum_{i=1}^{n} \hat{p}(y_i) \tag{7}$$

where $\hat{p}(y)$ is the estimated probability mass function (pmf).

Minimizing the functional given in Eq. (7) for binary data gives us a nice closed form solution. To see this, let $Y_1, \ldots, Y_n$ be iid Bernoulli random variables and $y_1, \ldots, y_n$ be the observed values. For a Bernoulli random variable, we have

$$p(y) = \pi^y (1-\pi)^{1-y}$$

for y=0, 1. Let $\hat{\pi}$ be the estimated probability, then Eq. (7) becomes

$$L2E = \sum_y \hat{\pi}^{2y}(1-\hat{\pi})^{2-2y} - \frac{2}{n}\sum_{i=1}^{n} \hat{\pi}^{y_i}(1-\hat{\pi})^{1-y_i}. \tag{8}$$

This can be simplified by noting that the terms in the second summation over the observed values are either $\hat{\pi}$ or $1-\hat{\pi}$ for y_i=1 or 0, respectively. Letting n_1 be the number of 1's observed, we have

$$L2E = (1-\hat{\pi})^2 + \hat{\pi}^2 - \frac{2}{n}\left[n_1\hat{\pi} + (n-n_1)(1-\hat{\pi})\right], \tag{9}$$

which further simplifies to

$$L2E = 1 - 2\hat{\pi} + 2\hat{\pi}^2 - \frac{2}{n}\left[(2n_1 - n)\hat{\pi} + n - n_1\right]. \tag{10}$$

Taking the derivative with respect to $\hat{\pi}$, we get

$$\frac{\partial L2E}{\partial \hat{\pi}} = -2 + 4\hat{\pi} - \frac{4n_1}{n} + 2 = 4\hat{\pi} - 4\frac{n_1}{n}. \tag{11}$$

Thus the *L2E* functional in Eq. (10) takes its minimum at $\hat{\pi} = \frac{n_1}{n}$, which gives an eminently sensible closed form solution as the *L2E* estimate of the binomial success probability.

3 *L2E* for the Binary Response Regression

In many situations where *L2E* criteria have been successfully implemented, an iid structure in the observed data, whether directly or indirectly through the assumptions on the random components, was a necessity so as to enable us to use the empirical estimate of the expected value of the estimated pdf or pmf. In the setting of a generalized linear model for a binary response, no such iid structure seems to be present. To see this, we note that, in general, $Y_i \sim Ber(\pi_i)$, i=1, …, n, for the response variable in a binary response regression model and except for a trivial (or null) model $\pi_1 = \pi_2 = \cdots = \pi_n$, the probability distributions for the response will not be identical.

The approach we propose here may be motivated by a spirit similar to what motivates the least squares method. In a least squares curve fitting, the sum of the squared distances between the different data points and the putative curve is minimized. Since *L2E* functional itself is derived from the square of a distance measure, albeit in a more abstract space of pdf/pmf's, we may try forming a sum of the squared distances between the different data points and the curve that describes varying response probabilities using *L2E*'s. Let D^2_i be the square of the L_2 distance between the true pmf π_i and the estimated pmf $\hat{\pi}_i$. Thus we have

$$D_i^2 = \sum_y (\pi_i(y) - \hat{\pi}_i(y))^2 = l2e_i + \sum_y \pi_i^2(y) \tag{12}$$

where $l2e_i$ denotes the terms that depend on the estimated pmf $\hat{\pi}_i$ and that may be empirically estimated by an L2E functional. Minimizing the sum of D^2_i s with respect to the parameters in the estimated pmf $\hat{\pi}_i$ is precisely equivalent to minimizing the sum of $l2e_i$'s since the last terms in Eq. (12) do not depend upon $\hat{\pi}_i$ and hence may safely be dropped in search of the values of $\hat{\pi}_i$'s that minimize the sum of the squared distances. This motivates us to use the sum of *L2E*'s given in Eq. (10) with appropriate additions of indices,

$$\Lambda = \sum_{i=1}^{n}\left(1 - 2\hat{\pi}_i + 2\hat{\pi}_i^{\,2} - \frac{2}{n_i}\left[(2n_{1i} - n_i)\hat{\pi}_i + n_i - n_{1i}\right]\right) \tag{13}$$

as the criterion to be minimized in L2E modeling of a binary response variable. In the trivial null model: $\pi_1 = \pi_2 = \cdots = \pi_n$, with independent observations $Y_1, \ldots, Y_n$, the minimization of the L2E criterion in Eq. (13) again yields the estimate $\hat{\pi} = \frac{n_1}{n}$, further demonstrating this particular L2E criterion for the binary response is a sensible one.

For the logistic model,

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = X[i,]\beta \tag{14}$$

the L2E criterion to be minimized is given by substituting

$$\hat{\pi}_i = \frac{e^{X[i,]\beta}}{1+e^{X[i,]\beta}} \tag{15}$$

into Eq. (13). We will now present results of several simulations that compare L2E estimates with maximum likelihood estimates (MLE).

Example 1: On each of 20 points, x=0.2, 0.4,…, 4.0, we simulated a binary response $Y \sim Ber(\pi(x))$ where $\pi(x) = \frac{\exp(-2+2x)}{1+\exp(-2+2x)}$. There were five replications on each x, giving us total of 100 data points for each simulated data sets. 100 such simulated data sets were generated and the coefficients were estimated using both L2E and MLE. Table 1 lists the summary statistics for the estimated coefficients.

Table 1: Summary statistics of the coefficients estimated using L2E and MLE from 100 simulation runs in the Example 1. Putative values are β_0=-2 and β_1=2.

	β_0		β_1	
	MLE	L2E	MLE	L2E
Mean	-2.179	-2.289	2.195	2.328
Standard deviation (SD)	0.741	0.945	0.573	0.880
Median	-2.067	-2.128	2.084	2.125
Inter Quartile Range (IQR)	1.009	1.125	0.747	0.859

Figure 1 on the next page shows the (density scaled) histograms of the simulated values. Both Table 1 and Figure 1 indicate that L2E estimates have higher variability than MLE and somewhat bigger bias.

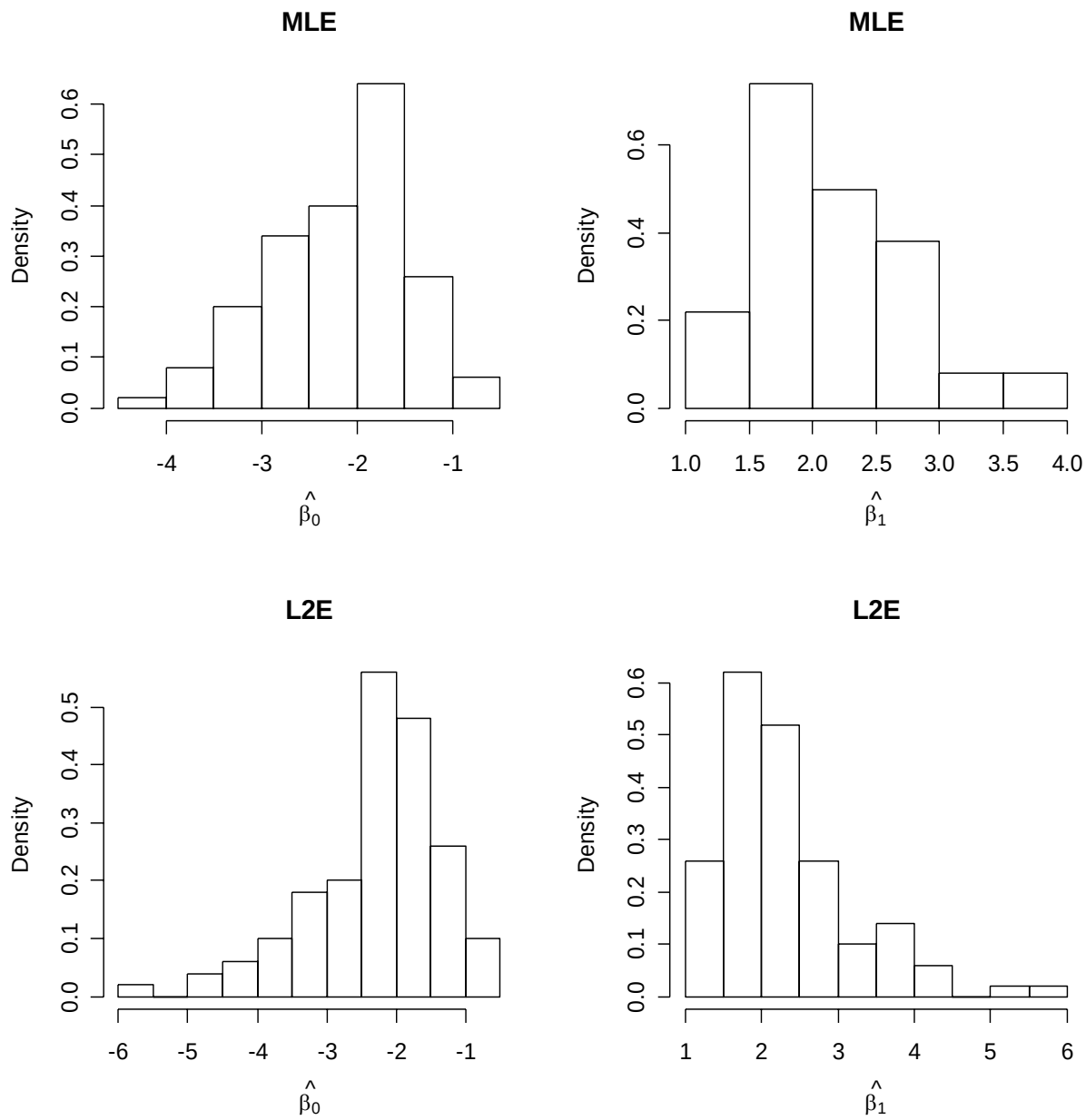

Figure 1: The (density scaled) histograms of the simulated values for Example 1.

Example 2: On each of 20 points, x=0.2, 0.4,…, 4.0, we simulated a binary response $Y \sim Ber(\pi(x))$ where $\pi(x) = \frac{\exp(-3+1.5x)}{1+\exp(-3+1.5x)}$. There were ten replications on each x, giving us total of 200 data points for each simulated data sets. 100 such simulated data sets were generated and the coefficients were estimated using both L2E and MLE. Table 2 lists the summary statistics for the estimated coefficients.

Table 2: Summary statistics of the coefficients estimated using L2E and MLE from 100 simulation runs in the Example 2. Putative values are β_0=-3 and β_1=1.5.

	β_0		β_1	
	MLE	L2E	MLE	L2E
Mean	-2.989	-3.010	1.488	1.498
SD	0.551	0.538	0.241	0.239
Median	-2.967	-2.943	1.452	1.471
IQR	0.720	0.792	0.317	0.345

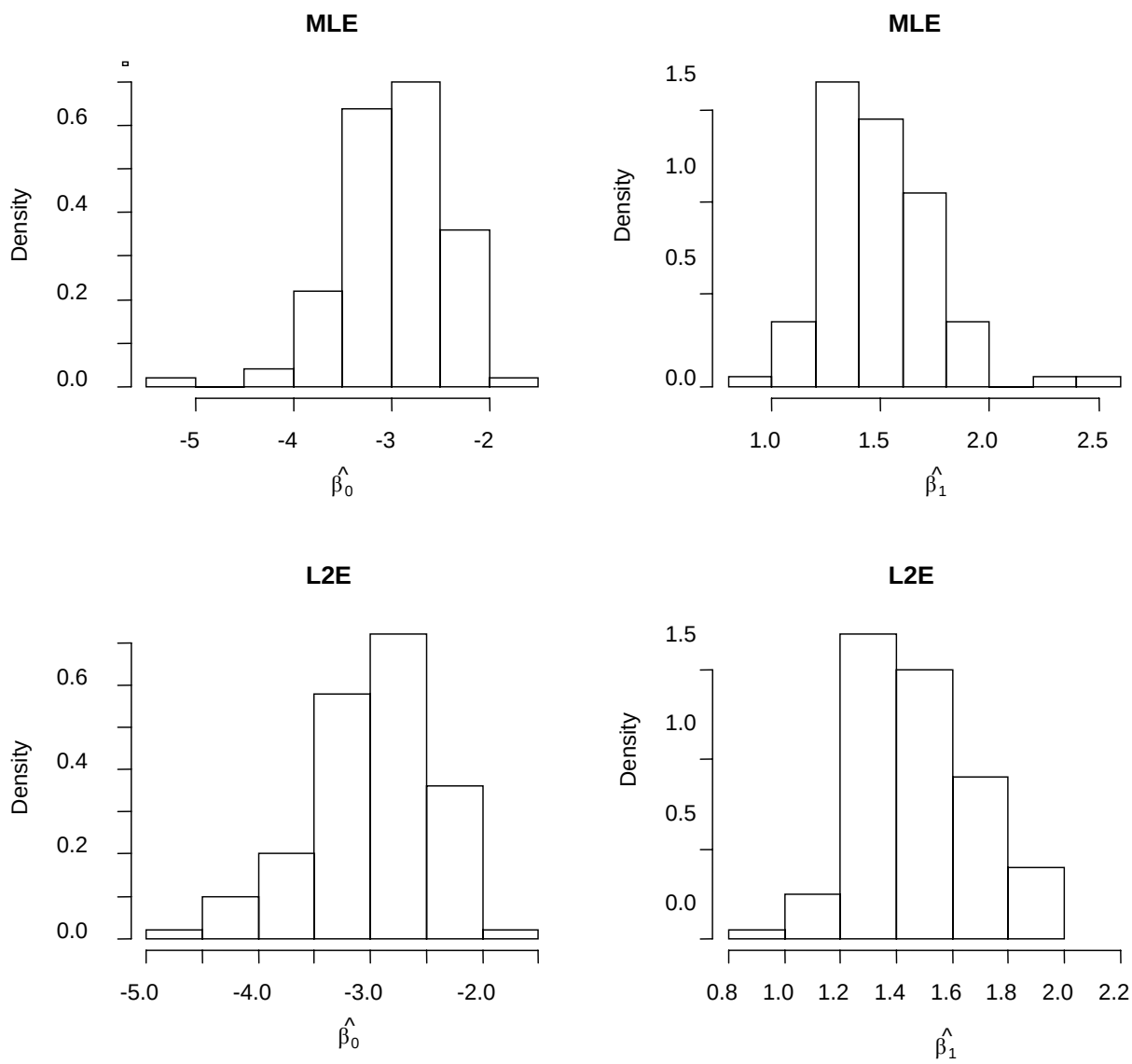

Figure 2: The (density scaled) histograms of the simulated values for Example 2.

Figure 2 on the previous page shows the (density scaled) histograms of the simulated values. What Table 2 and Figure 2 show that L2E estimates have more or less similar variability to MLE and somewhat smaller bias than MLE whereas it had higher bias than MLE in the previous example.

Example 3: On a grid of 50 points on the unit square (0,1)x(0,1), we simulated a binary response $Y \sim Ber(\pi(x_1,x_2))$ where $\pi(x) = \frac{\exp(-3+0.5x_1-1x_2)}{1+\exp(-3+0.5x_1-1x_2)}$. There were ten replications on each point (x_1,x_2), giving us total of 500 data points for each simulated data sets. 100 such simulated data sets were generated and the coefficients were estimated using both L2E and MLE. Table 3 lists the summary statistics for the estimated coefficients. Here variability and bias seem to be higher for L2E than MLE.

Table 3: Summary statistics of the coefficients estimated using L2E and MLE from 100 simulation runs in the Example 3. Putative values are β_0=-3, β_1=0.5, and β_2=-1.

	β_0		β_1		β_2	
	MLE	L2E	MLE	L2E	MLE	L2E
Mean	-2.999	-3.120	0.628	0.753	-0.985	-1.108
SD	0.884	1.989	0.767	2.019	0.955	1.149
Median	-2.788	-2.798	0.599	0.603	-0.886	-0.903
IQR	1.059	1.139	0.937	0.954	0.952	1.015

4 Discussion

The examples in the previous section showed that L2E estimates more or less agree with MLE estimates. The general L2E theory predicts that L2E estimates tend to have higher asymptotic variance than MLE. An important issue yet to be sorted out is that of starting values. Just like any other numerical minimization problem, L2E minimization depends more or less on the choice of starting values. It is possible that some of the variability seen in the simulation examples may have been due to the poor choice of starting values even as the simulation program compared the estimates from three different sets of starting values and chose the estimates corresponding to the lower local minimum of the functional.

Sometimes the fact that L2E often gives two or more different fits to a given data set may even be advantageous in that fits corresponding to different local L2E minima hold valuable diagnostic information as well as quite useful information on the local structure of the data, as was argued by Scott [5]. This suggests that a potentially useful diagnostic tool may be developed based on L2E estimates of binary response regression parameters.

Another key feature of L2E estimates in other settings has been its robustness. One problem of interest in a binary response regression is that of separation where at least one parameter estimate diverges without bound even as the likelihood converges. Separation primarily occurs in small samples with highly explanatory covariate over which there is no (or hardly any) overlap between 0's and 1's [6]. Initial exploration of L2E estimates for simulated

data sets with separation seem to have more potential fits than for the data sets without separation. In many cases, the minimization algorithm does converge at values of the estimates that are not too big. This is also an avenue of further research.

References

[1] McCullagh, P. and Nelder, J. A., *Generalized Linear Models*, Second Ed., 1989, Chapman and Hall, London.

[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] Scott, D. W., "Outlier Detection and Clustering by Partial Mixture Modeling," *Proceedings of COMPSTAT 2004,* preprint.

[6] Heinze, G. and Schemper, M., "A solution to the problem of separation in logistic regression," *Statistics in Medicine*, Vol. 21, 2002, pp. 2409-2419.