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

L2E

- L2E is an empirical measure of the L2 distance between the unknown true distribution and the estimated distribution in a parametric setting. [Scott (2001)]
- Adaptation of empirical integrated square error (ISE)

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

  for nonparametric estimation [Bowman (1984) and Rudemo (1982)] to parametric estimation.
- Letting iid observations xi's (continuous case) and yi's (discrete case), L2E functionals may be written as follows.

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

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

Generalized Linear Model

- Generalization of a classical linear model

$$E(Y) \equiv \mu = X\beta.$$

  to enable modeling of response variables that are not necessarily continuous and/or ranges over R, the entire real line.
- Using the link function g to map the mean response onto R.

$$g(\mu_i) = X[i,]\beta.$$

LINK	MODEL
g(x)=x	Classical
g(x)=log(x)	Poisson
g(x)=logit(x)	Logistic
$g(x)=\Phi^{-1}(x)$	Probit

L2E for Binary Data

- Let $Y_1,\ldots,Y_n$ be iid Bernoulli random variables with the success probability π.
- L2E for a discrete rv is given by $$L2E = \sum_y \hat{p}^2(y) - \frac{2}{n}\sum_{i=1}^{n} \hat{p}(y_i).$$
- We then have $$L2E = \sum_y \hat{\pi}^{2y}(1-\hat{\pi})^{2-2y} - \frac{2}{n}\sum_{i=1}^{n} \hat{\pi}^{y_i}(1-\hat{\pi})^{1-y_i}.$$

Letting n_1 be the number of 1's observed, we have

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

This is minimized by

$$\hat{\pi} = \frac{n_1}{n}$$

L2E for Binary Response Regression

- The empirical part of L2E utilizes iid-ness.
- 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 $\pi_1=\pi_2=\ldots=\pi_n$, the probability distributions for the response will not be identical.
- Naïve summation of pointwise L2Es is equivalent to the sum of the pointwise distances ***in the space of pmf's*** as far as minimization with respect to the parameter is concerned.

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

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

- Sum up above L2Es evaluated at each π_i.

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

- In the trivial null model: $\pi_1=\pi_2=\ldots=\pi_n$, the minimization of this L2E criterion again yields the estimate $\hat{\pi} = \frac{n_1}{n}$

- For a logistic model, substitute $\hat{\pi}_i = \frac{e^{X[i,]\beta}}{1 + e^{X[i,]\beta}}$
- For a probit model, substitute $\hat{\pi}_i = \Phi(X[i,]\beta)$

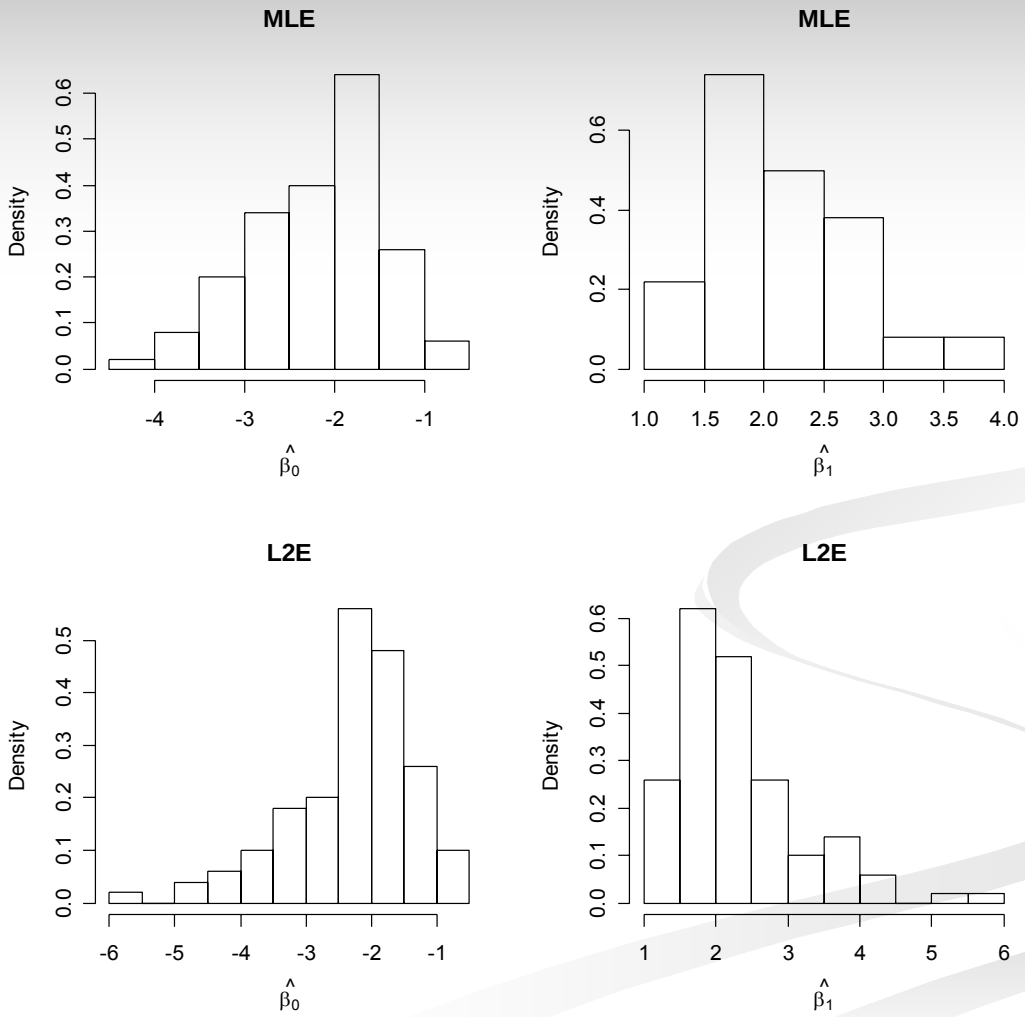
Simulation 1

- x=0.2, 0.4,…, 4.0, 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: Summary statistics of the coefficients estimated using L2E and MLE from 100 simulation runs in the Simulation 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

The (density scaled) histograms of the simulated values for Simulation 1



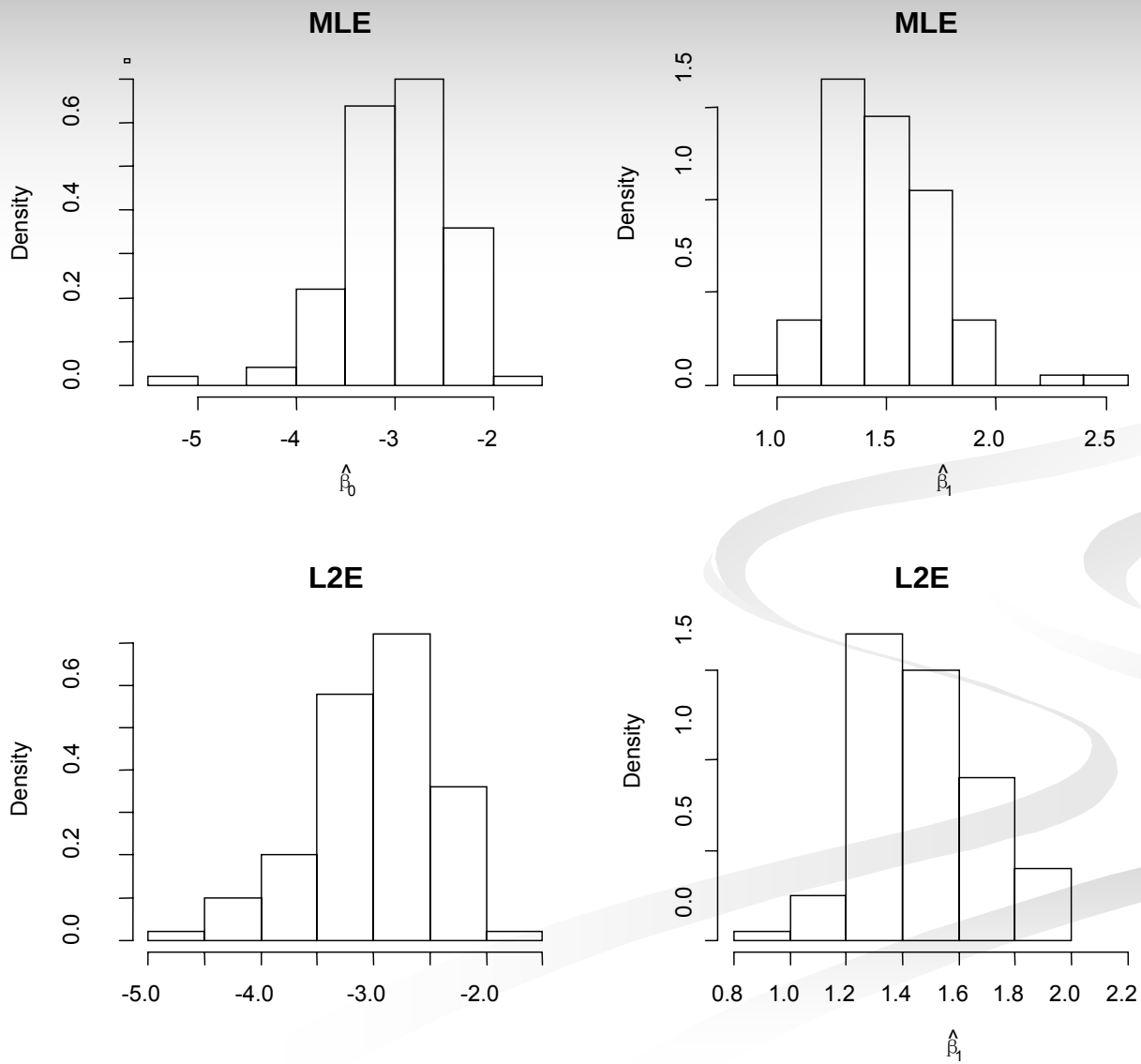
Simulation 2

- x=0.2, 0.4,…, 4.0, 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: Summary statistics of the coefficients estimated using L2E and MLE from 100 simulation runs in the Simulation 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

The (density scaled) histograms of the simulated values for Simulation 2



Simulation 3

- On a grid of 50 points on the unit square (0,1)x(0,1) ten replications on each point (*x1*,*x2*), 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: Summary statistics of the coefficients estimated using L2E and MLE from 100 simulation runs in the Simulation 3. Putative values are β_0=-3, β_1=0.5, and β_2=-1.

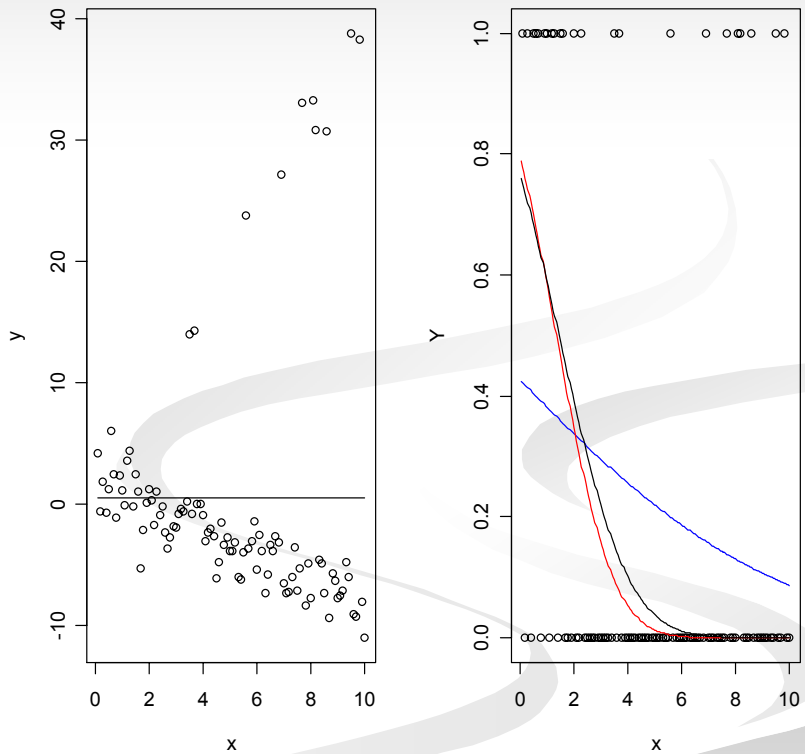
	β_0		β_1		β_2	
	MLE	L2E	MLE	L2E	MLE	L2E
Mean	-2.999	-3.120	0.628	0.753	-0.985	-1.108
SD	0.884	1.989	0.767	2.019	0.955	1.149
Median	-2.788	-2.798	0.599	0.603	-0.886	-0.903
IQR	1.059	1.139	0.937	0.954	0.952	1.015

Robustness

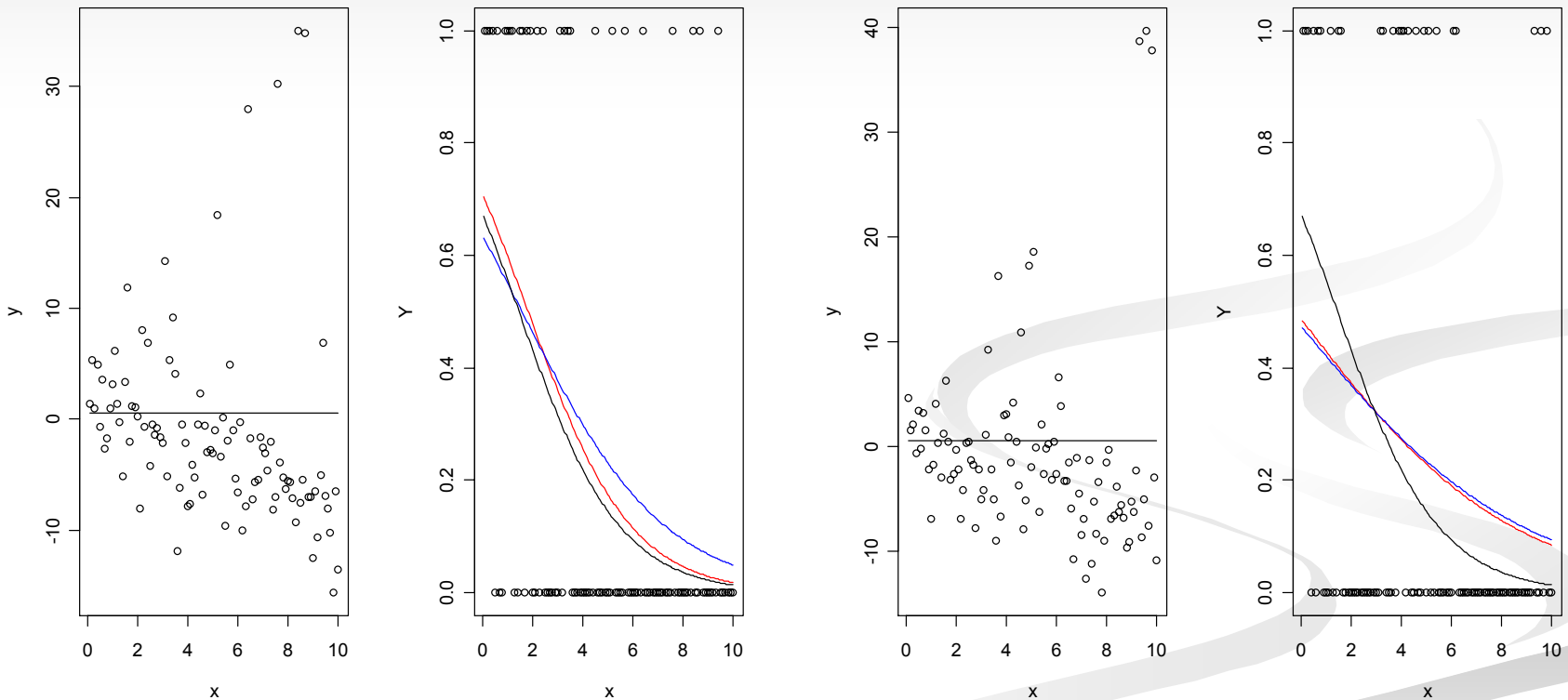
- Does the L2E derived functional we are using to fit a logistic/probit regression model give us "robust" estimates?
- L2E estimates, as minimum distance estimates, have "automatic robustness" against outliers, etc.

Simulated Data with Outliers

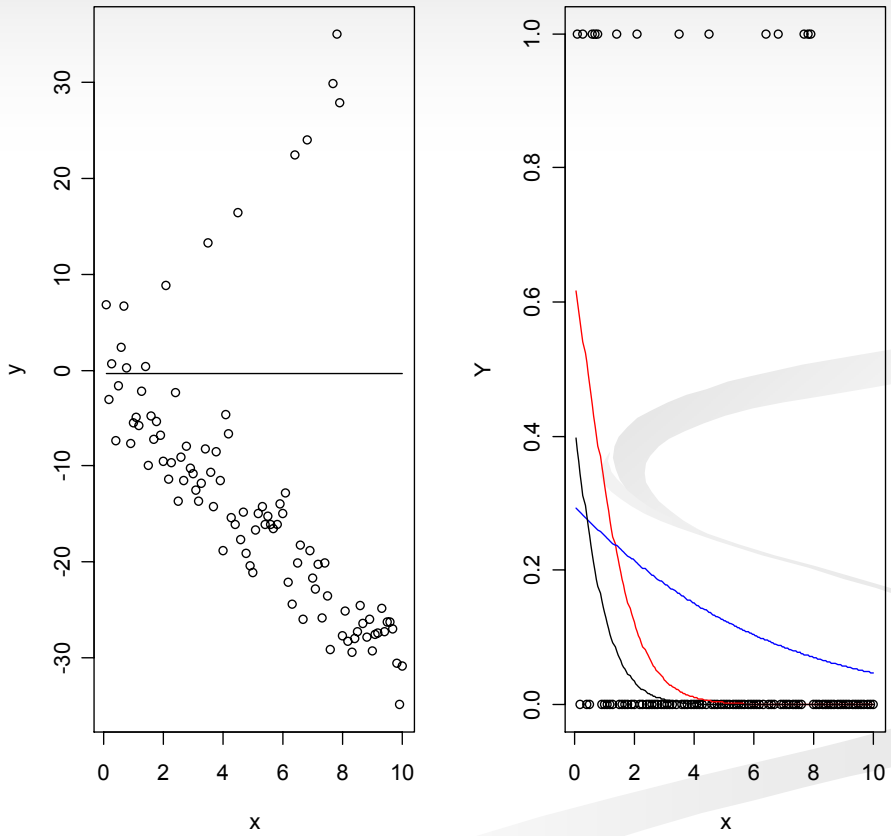
- 100 data points were generated based on a probit model: $\pi(x)=\Phi(0.73-0.50x)$, with about 10 points being clear outliers.
- Black (putative model) Blue (MLE) Red (L2E)
- With N=100 repeated runs, mean L2E estimates were 0.68 and -0.38 while mean MLE were 0.30 and -0.22



As in previous slide with logistic link (and logistic errors in the left plot).



logit link, intercept=-.34, slope=-1.5 (est: MLE -.87, -.21, L2E .54, =-1.26)



Conclusion and Ongoing Research

- Simulations demonstrate that (quasi) L2E estimates for a binary response regression model have robustness property typical of L2E estimates.
- Sampling distributions of the L2E estimators? Approximate characterization of standard errors?
- L2E estimates as diagnostic measures?

References

- Bowman, A. W., "An Alternative Method of Cross-validation for the Smoothing of Density Estimates," *Biometrika*, Vol. 71, 1984, pp. 353-360.
- Rudemo, M., "Empirical Choice of Histograms and Kernel Density Estimators," *Scandinavian Journal of Statistics*, Vol. 9, 1982, pp. 65-78.
- McCullagh, P. and Nelder, J. A., *Generalized Linear Models*, Second Ed., 1989, Chapman and Hall, London.
- Scott, D. W., "Parametric Statistical Modeling by Minimum Integrated Square Error," *Technometrics*, Vol 43, No. 3, 2001, pp. 274-285.