Small Sample Properties of Parametric and Nonparametric Estimators in Quantal Bioassay

Dongryeon Park* Sangun Park†

April 30, 2003

Abstract

In bioassay, the logit model is the most widely used parametric model. However, the exact form of the response curve is usually unknown and even very complicated, so it is likely that the true model does not follow the logit model. Therefore, according to well known asymptotic results, when the sample size is very large, we should probably use nonparametric regression rather than the logit model unless the exact form of the true response curve is known. In practice, however, we can not increase the sample size infinitely, so the asymptotic result would not be so useful. In this paper, we would like to compare the small sample properties of the logit model and the nonparametric estimator. As the nonparametric method, we choose the locally weighted quasi–likelihood estimator. A Monte Carlo study was done under various circumstances, and it turned out that the locally weighted quasi–likelihood estimator is very competitive in the small sample situation.

KEY WORDS: ED100α; Local quasi–likelihood; Logit model; Nonparametric regression; Response curve; Small sample.

*Corresponding author : Associate Professor, Department of Statistics, Hanshin University : drpark@hanshin.ac.kr

†Associate Professor, Department of Applied Statistics, Yonsei University : sangun@yonsei.ac.kr

1 Introduction

In bioassay, different concentrations of a chemical compound are applied to experimental animals, and the all–or–none reaction of the animals are then recorded. For example, in pharmacology, the effective action of a drug or vaccines is treated by an animal experiment, where the deaths or other all–or–none reactions of the animals are recorded after exposure to the drug at various levels (Müller and Schmitt, 1988). One is often interested in the dose level such that 100α % of the subjects react, which is denoted by ED100α. ED100α summarizes the potency of the chemical and may subsequently form the basis of comparisons between different compounds.

In these experiments, if the ith subject reacts at the dose level x_i, then the binary response variable Y_i is encoded by $Y_i = 1$, and if no reaction, then $Y_i = 0$. We assume that $P(Y_i = 1|X_i = x_i) = p(x_i)$ and $P(Y_i = 0|X_i = x_i) = 1 - p(x_i)$. Here the function $p(\cdot)$ denotes the response curve and we further assume that p is strictly monotone increasing, so the functional $p^{-1}(\alpha)$, which is referred to as ED100α, is well defined for $0 < \alpha < 1$. The aim is the estimation of $p^{-1}(\alpha)$.

There are two main approaches for the estimation of $p^{-1}(\alpha)$. One is the parametric approach and the other is the nonparametric approach. It is well known that the convergence rate of the mean squared error for the parametric estimator is $O(n^{-1})$ if the true model follows the assumed one, whereas the corresponding rate for the nonparametric estimators is usually $O(n^{-\delta})$ for some number $\delta \in (0, 1)$. However, if an incorrect parametric model is used, then the mean squared error for the parametric estimator does not even converge to zero, so the nonparametric regression techniques are much better than the parametric methods in the asymptotic point of view for this case.

In many literatures of bioassay, it has been pointed out that usually biological mechanisms of drug action are so complicated that the form of p is completely unknown, so the selection of the proper functional form is not an easy task. Therefore, unless we do have some prior information about the exact form of p, the nonparametric method should probably be used as long as the sample size is very large. In practice, however, it is almost impossible to increase the sample size infinitely, so we need some results for comparing the performance of two methods under the small sample size.

There is no doubt that the nonparametric and the parametric regression should not be viewed as mutually exclusive competitors, but sometimes

this is not the case. For example, when Wu (1985) proposed a sequential design for the estimation of ED100α under the small sample situation, he would like to have a good estimate of the response curve and argued that the smooth nonparametric estimate is not feasible, since he believed that the nonparametric estimate requires a large number of observations to be a good estimate, so he used the logit model. However, he did not provide any numerical evidences for his argument. There is no guarantee that the true model is logit. Then how can we be sure that the logit model is better than the nonparametric regression under the small sample situation.

In this paper, we deal with the estimation problem of ED100α under the small sample situation. We compare the small sample properties of the parametric and the nonparametric method using simulation study. For the parametric method, we choose the logit model which is the most widely used one.

As the nonparametric methods, the traditional nonparametric regression methods can be used. Müller and Schmitt (1988) defined the kernel response curve estimator in analogy to the nonparametric regression of Gasser and Müller (1984). It is known that the local polynomial regression has several more appealing features than the traditional nonparametric regression. The better performance near boundaries is one of them (Fan, 1992). Park (1999) considered the local linear regression as the response curve estimator and compared the finite sample performance with Müller and Schmitt's kernel response curve estimator.

However, these estimators ignore the binary nature of response, so they have some problems as the estimator of $P(Y = 1|X = x)$. The obvious one is that the fitted curve is not guaranteed to lie in the interval (0,1). To overcome these difficulties, a generalization of the weighting mechanism is needed. It is well known that generalized linear model (Nelder and Wedderburn, 1972) is the appropriate technique for binary response and can be applied to the nonparametric regression setting (Tibshirani and Hastie, 1987; Staniswalis, 1989). As a further extension, Wedderburn (1974) first considered a quasi–likelihood method, which requires only specification of a relationship between the mean and the variance of the response. Optimal properties of the quasi–likelihood methods have received considerable attention in the literature (Godambe and Heyde, 1987).

The kernel smoothing idea can be extended to the case where the quasi–likelihood is used. Fan, Heckman, and Wand (1995) proposed the locally weighted quasi–likelihood estimators in one–parameter exponential family,

and we choose their estimator for the nonparametric response curve estimator.

In Section 2, properties of two estimators are briefly summarized. In Section 3, the results of a Monte Carlo study are reported. We present a summary of our findings in Section 4.

2 Estimators of $ED100\alpha$

Consider binary response variables with single covariate case. Let $Y_1, \ldots, Y_n$ be the independent binary random variables with success probability $p(x_i) = P(Y = 1|X_i = x_i)$. We will assume that $p \in \mathcal{C}^2([0, 1])$, and that p is strictly monotone increasing. In the parametric generalized linear model it is usual to model a transformation of the regression function $E(Y|X = x) = p(x)$ as linear. The model is given by

$$\eta(x) = \beta_0 + \beta_1 x = g(p(x)) \tag{1}$$

where g is the link function. In the logit model, the response curve $p(x)$ is described by

$$\text{logit}(p(x)) = \log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x. \tag{2}$$

A simple algebra yields the expression of

$$p(x) = \frac{\exp(\eta(x))}{1 + \exp(\eta(x))}. \tag{3}$$

MLE $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained and whole response curve is estimated by plugging them into (3). The estimate of ED100α is then given by

$$\frac{1}{\hat{\beta}_1}\left(\log\left(\frac{\alpha}{1 - \alpha}\right) - \hat{\beta}_0\right), \tag{4}$$

and it is denoted as logit–MLE.

In many applications, the full likelihood function is unknown and one can only specify the relationship between the mean and the variance. Suppose the conditional variance is modeled as $Var(Y|X = x) = V(p(x))$ for some specific function V. In this case, the estimation of the mean can be achieved

by replacing the conditional log–likelihood by the quasi–likelihood function $Q(p(x), y)$ which satisfies

$$\frac{\partial}{\partial w}Q(w, y) = \frac{y - w}{V(w)}, \tag{5}$$

and estimating β_0 and β_1 by maximizing the quasi–likelihood

$$\sum_{i=1}^{n} Q[g^{-1}(\beta_0 + \beta_1 X_i), Y_i]. \tag{6}$$

Since we deal with binary response, $V(p) = p(1 - p)$ and in this case, the quasi–likelihood method coincides with the Bernoulli log–likelihood method.

Fan, Heckman, and Wand (1995) proposed the local quasi–likelihood using kernel weights, which is given by

$$\sum_{i=1}^{n} Q[g^{-1}(\beta_0 + \beta_1(X_i - x)), Y_i]K\left(\frac{X_i - x}{h}\right) \tag{7}$$

where h is the bandwidth and K is the kernel function. Maximizing (7) with respect to β_0 and β_1 leads to the maximum local quasi–likelihood estimate

$$\hat{\eta}(x; h) = \hat{\beta}_0 \tag{8}$$

and the local quasi–likelihood response curve estimate can be computed by applying the inverse link function

$$\hat{p}(x; h) = g^{-1}(\hat{\eta}(x; h)). \tag{9}$$

To obtain the estimate of ED100α, $\hat{p}(x; h)$ should be monotone, but there is no such guarantee in practice, so it is essential to consider the monotonizing transform. Now the estimate of ED100α is given by the following intuitive estimate

$$\inf\{x : \tilde{p}(x; h) \geq \alpha\} \tag{10}$$

where $\tilde{p}(x; h)$ is the monotonized version of $\hat{p}(x; h)$.

3 Simulation Study

3.1 Models and Evaluation

A Monte Carlo study was carried out to compare the small sample properties of logit MLE and the locally weighted quasi–likelihood estimators of ED100α. As the true response curve, we used the following six models:

1. The logit model, $p_1(x) = [1 + \exp(5 - 10x)]^{-1}$
2. The complementary log–log model, $p_2(x) = 1 - \exp[-\exp(-5 + 8x)]$
3. The normal mixture model, $p_3(x) = 0.5\Phi\left(\frac{x-0.3}{0.1}\right) + 0.5\Phi\left(\frac{x-0.7}{0.1}\right)$
4. The normal mixture model, $p_4(x) = 0.3\Phi\left(\frac{x-0.3}{0.1}\right) + 0.7\Phi\left(\frac{x-0.7}{0.1}\right)$
5. The Weibull model, $p_5(x) = 1 - \exp(-2x)^2$
6. The uniform model, $p_6(x) = x$

Parameters for each models were chosen such that $0 < ED100\alpha < 1$, for $0 < \alpha < 1$. $p_1(x)$ is a symmetric sigmoid response curve, and $p_2(x)$ is a non–symmetric sigmoid curve, and $p_3(x)$ is a symmetric non–sigmoid curve with three inflection points, and $p_4(x)$ is a non–symmetric non–sigmoid curve with three inflection points, and $p_5(x)$ is a non–symmetric, strictly concave curve, and $p_6(x)$ was included since it is not an element of $\mathcal{C}^2([0,1])$ which is an interesting case because we assume $p \in \mathcal{C}^2([0,1])$.

The design points x_i's for each models were determined by Uniform(0,1) pseudo random numbers. The sample sizes under consideration were n =20, 40, 60, 80, 100. For the generation of the responses, Uniform(0,1) pseudo random numbers were constructed again and compared with $p(x_i)$ for the respective models.

In the small sample situation, the locally weighted quasi–likelihood estimate is not necessarily monotone, so the monotonizing transformation is of course required. There are several monotonizing transformation techniques discussed in the literature (Friedman and Tibshirani, 1984; Kappenman, 1987; Härdle, 1990; Hall and Huang, 2001). These transformation techniques might change the shape of the estimate dramatically in the small sample situation, but unfortunately, there is no previous study for comparing the small sample performance of the existing monotonizing transformations. Among others, we considered the method in Friedman and Tibshirani (1984) and the one in Kappenman (1987). The algorithm for computing two monotone estimates can be summarized as follows. First, construct the locally weighted quasi–likelihood estimate $\hat{p}(X_{(i)}; h_{cv})$ with the bandwidth h_{cv} chosen by the cross validation method, and then apply the following two methods, respectively. For the method in Friedman and Tibshirani (1984), find the monotone estimate $\tilde{p}(X_{(i)}; h_{cv})$ by means of the pool adjacent violators (PAV) algorithm

(Barlow *et al*, 1972). We denote this estimate as local–P. For the method in Kappenman (1987), increase the bandwidth h as small as possible to find h_o such that $\hat{p}(X_{(i)}; h)$ is monotone for all $h \geq h_o$. We denote this estimate as local–K.

To compare the performance of each estimators, the Monte Carlo MSE were computed as the average of $(\widehat{ED100\alpha} - ED100\alpha)^2$ over 1000 simulation samples.

It is known that MLE of both β_0 and β_1 exist and are unique if and only if the following condition is satisfied;

$$(x^+_{\min}, x^+_{\max}) \cap (x^-_{\min}, x^-_{\max}) \quad \text{is} \quad \text{nonempty} \tag{11}$$

where $x^+_{\max(\min)} = \max(\min)\{x_i : y_i = 1\}$ and $x^-_{\max(\min)} = \max(\min)\{x_i : y_i = 0\}$. The Monte Carlo sample which did not satisfy the condition (11) were discarded from the simulation. The evaluation of MLE of β_0 and β_1 were done by S–Plus function *glm*, and the locally weighted quasi–likelihood estimates were computed by S–Plus function *locfit* (Loader, 1999).

We considered the cases of ED100α, $\alpha = 0.5, 0.75, 0.9$. In many cases, estimating ED50 is of intrinsic interest, whereas estimating extreme percentiles is more relevant in other cases, like in quality assurance, so ED90 was included. ED75 was included because we want to check up the argument in Wu (1985). According to Wu (1985), even though the parametric quantal response model is rarely justifiable on biological or physical grounds, the successful use in practice of the parametric approach for the quantal response problem is mainly due to the key fact that the parametric curves agree very closely in a wide range of α values, especially in the range of 0.2 to 0.8.

3.2 Comparison

Simulation results are reported in Figure 1 to Figure 3. In each Figure, the height of the bar plot is the Monte Carlo MSE of each estimator.

For two monotone estimators, local–K is better than local–P for all cases, so we focus on the comparison of logit–MLE and local–K.

For the logistic model, logit–MLE should be better than local–K and this is supported by the simulation results, except ED90 with $n = 20$ case. For the complementary log–log model, local–K is better than logit–MLE for all cases, except ED50 with $n = 20$ case. For the symmetric normal mixture model, we have two different patterns. For ED50, logit–MLE is better than

local–K for all n and, on the contrary, for both ED75 and ED90, local–K is better than logit–MLE for all n. For the non–symmetric normal mixture model, local–K is better than logit–MLE for all cases, except ED50 with $n = 20$ case. For the Weibull model, we have two different patterns, like the symmetric normal mixture model, but in the reverse way. Local–K is better than logit–MLE at ED50, but logit–MLE is better than local–K at both ED75 and ED90. For the uniform model, we have the exact same pattern with the symmetric normal mixture model.

The simulation results can be summarized as follows. For the symmetric response curve model, logit–MLE should be used for the ED50 estimation, but, local–K is preferable for the estimation of both ED75 and ED90. For the concave response curve model, the exact reverse result with the symmetric model case was observed. Local–K is recommended only at ED50 estimation. For other non-symmetric response curve model, local–K should be used for all cases. Therefore, we have a criterion based on the shape of the true response curve for choosing between logit–MLE or local–K.

The simulation results tell us that the nonparametric estimator can provide a very accurate result for some cases even in small sample size.

4 Conclusion

In bioassay, the exact form of response curve is usually unknown, so it is very difficult to select the proper functional form for the parametric regression. Therefore, according to well known asymptotic results, when the sample size is very large, we should probably use the nonparametric regression rather than the parametric regression unless the true response curve is known. However, the asymptotic theory does not tell us anything about the small sample situation. If we do not have proper information about p and we can use only small n, then which method do we have to use?

A Monte Carlo study was done under various circumstances and we have the criterion based on the shape of the true response curve for choosing the right method. Given data set, we can first construct the nonparametric estimate of the response curve, and from this estimate, we can determine the shape of the response curve, and then we can choose between logit–MLE and local–K.

It is expected that the result of this paper can provide useful alternative to some existing methods for estimating ED100α. For example, in Wu(1985),

what he needs at every steps of choosing next design points is the good estimate of ED100α with small n, and since we can construct the nonparametric response curve estimate based on the current design points at each step, we can choose the right method for estimating ED100α.

References

[1] Barlow, R.E., Bartholomew, D.J., Bremner, J.M., and Brunk, H.D. (1972). *Statistical Inference Under Order Restrictions*, New York, John Wiley

[2] Collett, D. (1991). *Modelling Binary Data*, London, Chapman & Hall

[3] Fan, J. (1992). Design–adaptive Nonparametric Regression. *Journal of the American Statistical Association*, **87**, 998–1004

[4] Fan, J., Heckman, N., and Wand, M. (1995). Local Polynomial Kernel Regression for Generalized Linear Models and Quasi–Likelihood Functions. *Journal of the American Statistical Association*, **90**, 141–150

[5] Friedman, J.H. and Tibshirani, R.J. (1984). The Monotone Smoothing of Scatterplots. *Technometrics*, **26**, 243–250

[6] Gasser, T. and Müller, H.G. (1984). Estimating Regression Functions and Their Derivatives by the Kernel Method. *Scandinavian Journal of Statistics*, **11**, 171–185

[7] Godambe, V.P., and Heyde, C.C. (1987). Quasi–Likelihood and Optimal Estimation. *International Statistical Review*, **55**, 231-244

[8] Hall, P. and Huang, L. (2001). Nonparametric Kernel Regression subject to Monotonicity Constraints. *Annals of Statistics*, **29**, 624–647

[9] Härdle, W. (1990). *Applied Nonparametric Regression*, Cambridge, Cambridge University Press

[10] Kappenman, R. (1987). Nonparametric Estimation of Dose-Response Curves with Application to ED50 Estimation. *Journal of Statistical Computation and Simulation*, **28**, 1–13

[11] Loader, C. (1999). *Local Regression and Likelihood*, New York, Springer-Verlag

[12] Morgan, B.J.T. (1992). *Analysis of Quantal Response Data*, London, Chapman & Hall

[13] Müller, H. and Schmitt, T. (1988). Kernel and Probit Estimates in Quantal Bioassay. *Journal of the American Statistical Association*, **83**, 750–759

[14] Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of Royal Statistical Society* Series A, **135**, 370–384

[15] Park, D. (1999). Comparison of Two Response Curve Estimators. *Journal of Statistical Computation and Simulation*, **62**, 259–269

[16] Staniswalis, J.G. (1989) The Kernel Estimates of a Regression Function in Likelihood–based Models. *Journal of the American Statistical Association*, **84**, 276–283

[17] Tibshirani, R. and Hastie, T. (1987). Local Likelihood Estimation. *Journal of the American Statistical Association*, **82**, 559–568

[18] Wedderburn, R.W.M. (1974). Quasi–Likelihood Functions, Generalized Linear Models, and the Gauss–Newton Mehod. *Biometrika*, **61**, 439–447

[19] Wu, C.F.J. (1985). Efficient Sequential Designs with Binary Data. *Journal of the American Statistical Association*, **80**, 974–984

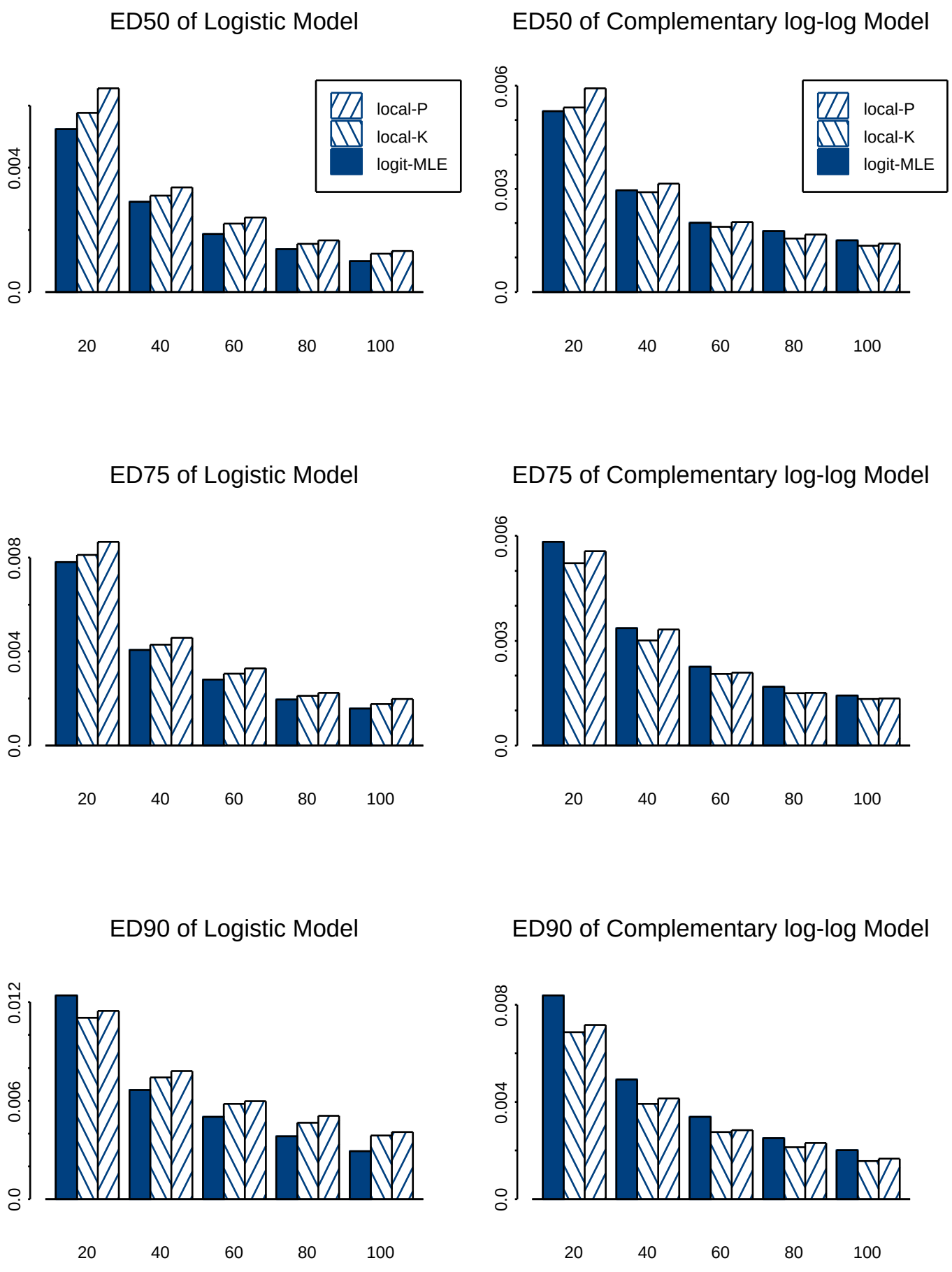

Figure 1: Monte Carlo MSE of the Logistic model and the Complementary log–log model

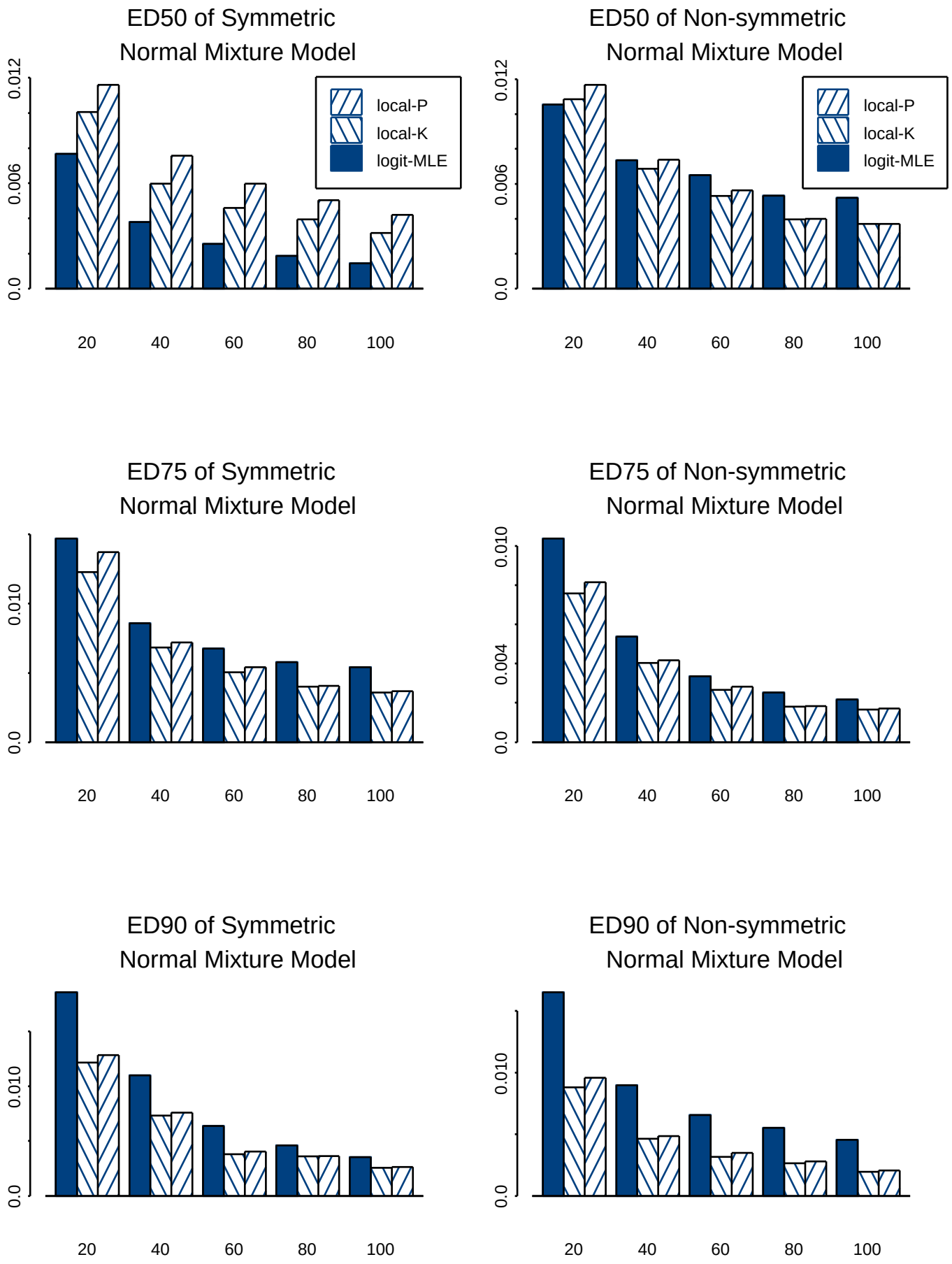

Figure 2: Monte Carlo MSE of the Symmetric normal mixture model and the Non–symmetric normal mixture model

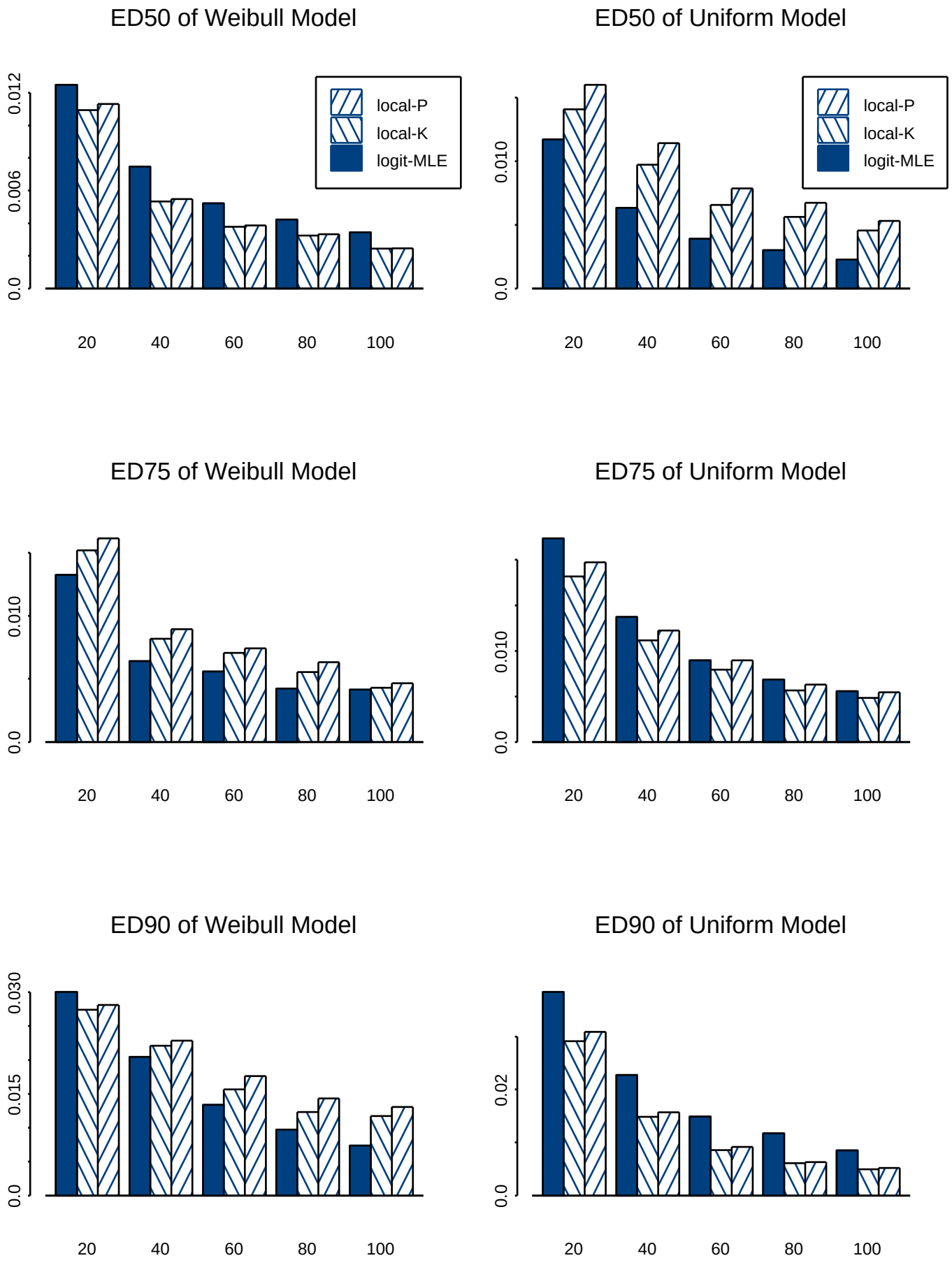

Figure 3: Monte Carlo MSE of the Weibull model and the Uniform model