Psicológica (2012), 33, 345-361.

Estimation of logistic regression models in small samples. A simulation study using a weakly informative default prior distribution

Amalia Gordóvil-Merino, Joan Guàrdia-Olmos* & Maribel Peró-Cebollero

University of Barcelona, Spain

In this paper, we used simulations to compare the performance of classical and Bayesian estimations in logistic regression models using small samples. In the performed simulations, conditions were varied, including the type of relationship between independent and dependent variable values (i.e., unrelated and related values), the type of variable (i.e., binary and continuous), and different Binomial distribution values and symmetry (i.e., symmetry and positive asymmetry). Iteratively re-weighted least squares was used as the estimate method to fit the models in both the classical and Bayesian estimations. A weakly informative default distribution was chosen as the prior distribution for Bayesian estimation. The simulation results demonstrate that Bayesian estimations provide more stable distributions but are not able to solve problems generated by asymmetric distributions based on small samples. Additional research using different kinds of priors that is addressed at solving problems caused by asymmetry is needed.

It is well known that sample size is associated with the accuracy of an estimator (Claeskens, Aerts, & Molenberghs, 2003; Mila, Yang, & Carriquiry, 2003). Inference on large data sets can be highly inaccurate if applied to logistic regression (LR) with small samples (SS) (Potter, 2005). Further, estimates from the analysis of small or medium samples are biased (Maiti & Pradhan, 2009). According to Steyerberg et al. (2001), the accurate estimation of the internal validity of a predictive LR model is especially problematic with SS. The problem centers on the fact that the

* This study was supported by the Comissionat per a Universitats I Recerca from the Generalitat de Catalunya (Departament de Innovació, Universitats I Empresa), the European Social Fund and the Grup de Tècniques Estadístiques Avançades Aplicades a la Psicologia (GTEAAP). Corresponding author's address: Joan Guàrdia Olmos. Department of Methodology for the Behavioral Sciences. University of Barcelona. Passeig Vall d'Hebron, 171, 08035 Barcelona (Spain). Tel. +34933125844 Fax. +34934021359. Email: jguardia@ub.edu

regression coefficients are overestimated for predictive purposes (Van Houwelingen & Le Cessie, 1990) and that the excess error estimate is outstanding (Bautista, Arana, Martí-Bonmatí, & Paredes, 1999). Nemes, Jonasson, Genell and Steineck (2009) state that LR overestimates odds ratios in studies with small to moderate samples size. Hence, large sample theory may not be appropriate in SS (Chen, Chen, Yang, & Chen, 2008; Dasgupta & Chen, 2002; Nottingham & Birch, 1998).

Some studies have been conducted in light of these problems related to SS data sets in LR models. Most of them relate to LR-SS contexts with respect to model fit, coefficient of determination, type I error, validation and estimate methods.

Pooi (2003) investigated the performance of the likelihood ratio test when fitting LR models with SS. It is commonly assumed that the null distribution of this test is approximately Chi-square with one degree of freedom. The study concludes that the size of the test at the nominally 5% level can range from 6% to 14% in the case of SS. Lekdee and Ingsrisawang (2010) present the empirical distribution of the Wald test, Score test, likelihood ratio, Hosmer-Lemeshow test, and Deviance test as compared with the Chi-square distribution when the sample is small. The above authors find that the Hosmer-Lemeshow distribution is still close to the Chi-square distribution. Regarding the R^2 coefficient, Mittlbock and Schemper (2002) propose two adjustments for use with SS; one is a direct analogue of R^2 from the general linear model, and the other is based on shrinkage. Another related work (Liao & McGee, 2003) presents two adjusted coefficients of determination for LR that correct the overestimation problem associated with unadjusted coefficients; these adjusted coefficients of determination are especially useful when the sample size is small or the number of predictors is large. Dasgupta, Pascual and Spurrier (2001) and Dasgupta, Spurrier, Martinez and Moore (2000) compare several binary regression slopes to that of a control. Their findings indicate that effective control of type I error can be achieved through the use of an asymptotic SS test in conjunction with a pivoted version of that test. These tests are generally robust to departures from the LR model (Dasgupta & Chen, 2002). Moreover, Potter (2005) presents a permutation test for inference in LR with small data sets. As compared to the asymptotic likelihood ratio test, type I error is well controlled. With regard to validation studies, the leave-one-out method may reduce classification bias when the same cases are used to obtain a predictive model (Bautista et al., 1999). For the purpose of obtaining a valid model for another population, re-calibration may be a recommended strategy (Steyerberg, Borsboom, van Houwelingen, Eijkemans, & Habbema, 2004), even when a small validation data set is

available for updating. Schulz, Betebenner and Ahn (2004) find that Bayesian estimations based on 25% samples have predictive validity nearly equal to maximum likelihood estimates (MLE) based on full samples. Data characteristics, particularly sample size, have the strongest effect on the predictive performance of LR models (Pearce & Ferrier, 2000). According to these authors, a sample size of 50 is too small for the development of accurate models. Hence, the sample density function is skewed with SS (Nemes et al., 2009). Some alternatives have been proposed to improve point and interval estimates and to reduce the associated error. Heinze and Puhr (2010) present an SS penalty conditional likelihood bias correction method. The results reveal a reduction in bias and an improvement in the precision of point estimations. Some adjustments for skewness and kurtosis of the conditional likelihood are achieved by saddle point methods, which are used to estimate densities, likelihoods and tail probabilities (Platt, 2000). Steyerberg, Eijkemans, Harrell and Habbema (2000) compare the performance of some selection and estimate methods using small data sets. They base their conclusions on the fact that selection from a predefined set of predictors combined with external information may be adequate to ensure predictive accuracy and insight in important predictive relationships.

Several authors support Bayesian estimation (BE) as an alternative to classical statistical approaches. Although there are some studies reporting satisfactory results (Marrelec, Benali, Ciuciu, Pélégrini-Issac, & Poline, 2003; Nijsse, 2003; Okada & Shigemasu, 2010; Wang & McArdle, 2005), Bayesian methodology is not yet commonly applied in psychology. In a previous study Gordóvil, Guàrdia, Peró and de la Fuente (2010), found some advantages for using BE in LR models. But what happens when working with SS and/or skewed samples? Cañadas, Lozano, de la Fuente, Vargas and Saldaña (2010) state that the Bayesian approach is specially recommended when working with SS. Although, very little is currently known about the unconditional LR estimate with SS, especially if samples are skewed. In this study, we aim to compare classical estimation (CE) and BE in SS-LR models via simulation. We present results for different types of distributions and asymmetry. We define SS as cell count with fewer effectives. We work with sample sizes equal to 100 but, the relationship between dependent variables (DV) and independent variables (IV) will provide a small number of effectives per cell (see the outline of simulated conditions). Note that DV refers to the outcome variable and IV refers to any kind of predictor variable.

METHOD

A simulation study was performed to compare CE and BE in LR models with SS. We simulated different conditions for a DV and two IV. The three variables (under all generated conditions) were set at a LR models:

$$\text{Logit}\{P(Y=1)\}=\beta_0+\beta_1 x_1+\beta_2 x_2+\varepsilon$$

The procedure was to initially manipulate certain parameters (i.e. distributional values, relationship conditions, asymmetry) to see the effect they have on the LR models.

The settings in the simulation were as follows. The DV *Y*, sets two conditions generated by two different Binomial distribution values, namely, π_1=.5 (first condition) and π_1=.2 (second condition). This is, we simulated a population in which a specific disorder was present in 50% of cases and absent in the remaining 50% (first condition for the DV: π_1=.5; π_0=.5); we also simulated another population in which a specific disorder was present in 20% of cases and absent in the remaining 80% (second condition for the DV: π_1=.2; π_0=.8). Note that the subscript 1 refers to the group with the disorder and the subscript 0 refers to the group without the disorder.

There were 2 IV, a binary variable (X_1) and a continuous variable (X_2). In regard to the binary IV X_1, two different kinds of conditions were generated by the Binomial distribution, namely, 1) unrelated condition (π_1=.5; π_0=.5, subscripts refers to the DV groups) and 2) related condition (π_1=.7; π_0=.4, subscripts refers to the DV groups). In the unrelated condition, we simulated that a specific exposure was present in the 50% of cases and absent in the remaining 50% in the group with the disorder. The same percentages affected the group without the disorder, and hence, 50% of cases without the disorder presented the exposure, and the remaining 50% did not present it. Thus, exposure was unrelated to the disorder because it was equally distributed in both groups (i.e. the group with the disorder and the group without the disorder). Under the related condition, exposure was present in the 70% of cases and absent in the remaining 30% of cases in the group with the disorder. Regarding the group without the disorder, exposure was present in the 40% of cases and absent in the remaining 60%. Hence, exposure was related to the disorder because it was differently distributed in both groups and it was more frequent in the group with the disorder compared to the group without the disorder.

Regarding the continuous IV X_2, symmetry and relation conditions were varied. The procedure was as follows. First, normal standardized variables (μ=0, σ^2=1) were defined. Afterward, we generated asymmetry using the formula below, based on Tukey's distribution (Jiménez & Martínez, 2006).

$$Y = T_{g,h}(Z) = \frac{e^{g \cdot z} - 1}{g} \cdot e^{h \cdot z^2/2}$$

Z is a random normal distributed variable, g refers to the asymmetry of the distribution (a value of 0 denotes a symmetric distribution), and h is the distribution kurtosis. Kurtosis was not studied in the performed simulations. Hence, h was set to 0, and g was set to .8 to generate positive asymmetry distributions as follows.

$$Y = T_{g,0}(Z) = \mu + \sigma * \frac{e^{0.8 \cdot z} - 1}{0.8}$$

In the above formula, we specify notation for mean and standard deviation. This is because mean values were varied to generate unrelated and related conditions. We define unrelated conditions to those conditions in which the mean of the group with the disorder (μ_1) has the same value as the mean of the group without the disorder (μ_0). Hence, $\mu_1 = \mu_0$=100. If the means are equal among the two groups, this value is un-related to the presence or absence of the disorder. Moreover, we define related conditions to those conditions in which the mean of the group with the disorder (μ_1) has a different value compared to the mean of the group without the disorder (μ_0). This is, μ_1=100, μ_0=120. If mean value varies among the two groups, we can say that this value is related to the presence of the disorder. Taking this explanation into account, the unrelated conditions were: 1) $\mu_1=\mu_0$=100; σ=15; g_1=.0 and 2) $\mu_1=\mu_0$=100; σ=15; g_1=.8. The related conditions were: 3) μ_1=100; μ_0=120; σ=15; g_1=.0 and 4) μ_1=100; μ_0=120; σ=15; g_1=.8. As can be seen, positive asymmetry is only generated in group 1, while the value of the standard deviation remained constant throughout all simulated conditions.

Note that, unrelated and related conditions refer to the relationship between DV (the presence of the disorder) and IVs (the presence of the exposure). LR models with three variables were defined. The three variables

were: the dependent binary variable, an independent binary variable and a continuous IV. Note that, when positive asymmetry is generated, overdispersed models are defined. Each simulated condition was carried out using a sample size of n=100. Simulations were repeated 10000 times. All simulated conditions were combined and they are summarized as follows (see Table 1).

Table 1. Outline of simulated conditions using sample sizes of 100 with 10000 replications.

Type of IV	Type of Condition	Distribution Values	
		DV	IV
Binary (X_1)	Unrelated	π_1=.5; π_0=.5	π_1=.5; π_0=.5
		π_1=.2; π_0=.8	
	Related	π_1=.5; π_0=.5	π_1=.7; π_0=.4
		π_1=.2; π_0=.8	
Continuous (X_2)	Unrelated	π_1=.5; π_0=.5	μ_1=μ_0=100; σ=15; g_1=.0
		π_1=.2; π_0=.8	μ_1=μ_0=100; σ=15; g_1=.8
	Related	π_1=.5; π_0=.5	μ_1=100; μ_0=120; σ=15; g_1=.0
		π_1=.2; π_0=.8	μ_1=100; μ_0=120; σ=15; g_1=.8

Note: IV, independent variable; DV, dependent variable.

We used generalized linear models (GLM) methodology, particularly iteratively re-weighted least squares (IRLS), as the estimate method to fit the LR models. The IRLS algorithm is a simplification of MLE but is limited to exponential distributions. An advantage of using GLM methods over the individual maximum likelihood is much easier with modeling process by the GLM; the interested reader may refer to Hilbe (2009) for a useful explanation of LR estimation methods.

Regarding the BE, we used weakly informative prior knowledge to find out whether differences, compared to the CE, could have been detected. Following Gelman, Jakulin, Pittau and Su (2008), LR models were estimated using a simple adaptation of the IRLS algorithm. The choice of the parametric family for the prior distribution corresponded to the *t* family, focusing on the Cauchy distribution. This prior distribution was constructed by scaling non-binary variables to have a mean 0 and standard deviation of .5 and scaling binary variables to have a mean of 0 and to differ by 1 in the lower and upper conditions. After standardizing variables, the independent Cauchy prior distribution was assigned to coefficients in the LR models, except for the constant term. The prior distribution was centered to 0 and scaled to 2.5; see Gelman et al. (2008) for a detailed explanation of the constructed prior distribution.

According to the above authors, the inclusion of some actual prior information is enough to regularize the extreme inferences that are obtained using MLE or completely non-informative priors. Other marks in its favor are the production of stable, regularized estimates and the innovation developed by Raftery[1]; the prior scale parameter is given a direct interpretation in terms of LR parameters (Gelman et al., 2008).

Notice that when we refer to BE, we are referring to the use of the above mentioned weakly informative default prior distribution in the generated LR models. When we refer to CE, we are emphasizing a lack of any kind of prior distribution in the simulated LR models.

Data Analysis

Classical and Bayesian LR estimations were obtained for each simulated condition. We used the glm function from the stats package (R Development Core Team, 2010) to obtain the classical LR estimation, and we used the bayesglm function from the arm package (Gelman et al., 2010) to obtain the Bayesian LR estimation. The IRLS estimation algorithm was employed in both cases. Means of the coefficient estimates, means of the standard errors and correct *p*-values were computed under each simulated condition among the 10000 replications. Proportion comparison tests (related samples) and binomial tests were also computed to identify significant differences between classical and Bayesian LR estimations in the percentage of correct p-values. Simulations, calculations and figures were derived with the R environment 2.11.1 (R Development Core Team, 2010).

[1] As explained in Gelman et al. (2008), Raftery standardized the input variables and applied this procedure to Bayesian generalized linear models.

RESULTS

The percentage of correct *p*-values given by CE and BE are summarized in Table 2. A two-sample proportion test (related samples) was performed to analyze possible statistical differences over correct decision percentages beyond the CE and BE. We performed a one-tailed test by supposing that BE would outperform CE. When assumptions for the two-sample proportion test were not met, binomial tests were used.

Table 2. Percentage of correct *p*-values, the proportion comparison test (related samples) and the binomial test, for type of independent variable and type of logistic regression estimation.

Type of IV	Distribution Values	Type of Estimation		Proportion Comparison Test			Binomial Test
		Classical	Bayesian	z	CI	p	p
Binary	DV=.5; IV=.5	96.25	96.85	14.34	.08-.92	<.001	
	DV=.2,IV=.5	96.48	97.35	16.98	.09-.91	<.001	
	DV=.5,IV=.7/.4	54.23	53.23	13.39	.27-.73	<.001	
	DV=.2,IV=.7/.4	41.83	39.70	21.45	.23-.77	<.001	
Continuous	DV=.5; $\mu_1=\mu_0=100$; S	95.25	95.95				<.001
	DV=.5; $\mu_1=\mu_0=100$; A	35.90	0				<.001
	DV=.2; $\mu_1=\mu_0=100$; S	95.70	96.45				<.001
	DV=.2; $\mu_1=\mu_0=100$; A	34.65	0				<.001
	DV=.5; $\mu_1=100$; $\mu_0=120$; S	100	100				<.001
	DV=.5; $\mu_1=100$; $\mu_0=120$; A	100	100				1
	DV=.2; $\mu_1=100$; $\mu_0=120$; S	99.85	99.75				1
	DV=.2; $\mu_1=100$; $\mu_0=120$; A	100	100				1

Note: IV, independent variable; DV: dependent variable; S, symmetry; A, asymmetry

This table lists results for both the classical and Bayesian LR estimations. The results are arranged according to the simulated condition. Hence, the first and second conditions of the binary variable are unrelated conditions whereas third and fourth conditions refer to related conditions. Regarding the continuous variable, first, second, third and fourth conditions correspond to unrelated conditions whereas fifth, sixth, seventh and eighth conditions correspond to related conditions.

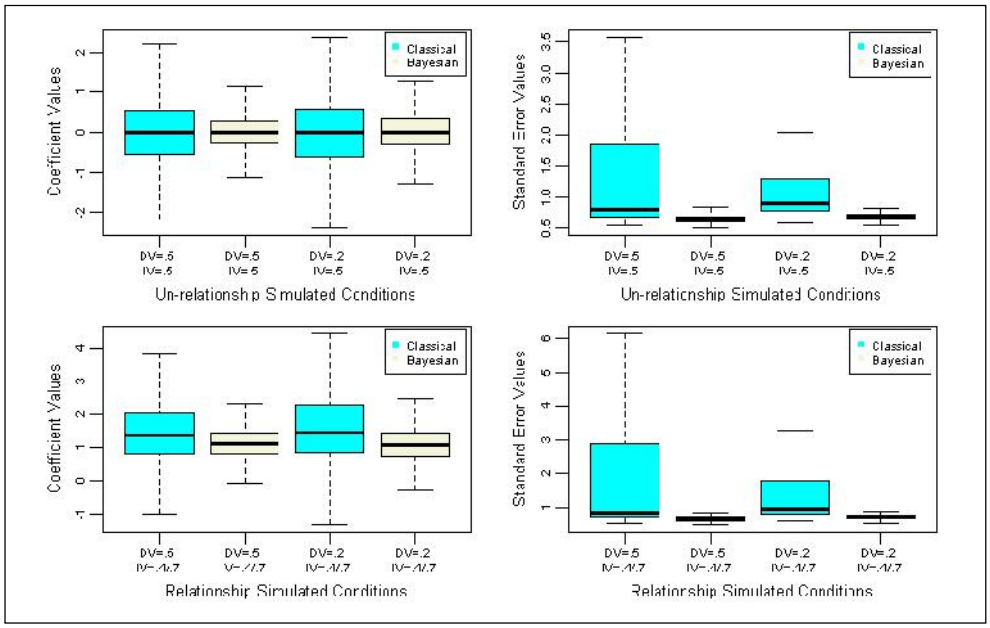
With regard to the binary IV, both the CE and BE correctly detect the unrelated conditions (percentages of correct p-values are between 96.25% and 97.35%). Percentages are significantly different between the two unrelated conditions in favor of the BE ($z=14.34$, CI=.08-.92, $p<.001$ for the .5-distributed DV and .5-distributed-IV; $z=16.98$, CI=.09-.91, $p<.001$ for the .2-distributed DV and .5-distributed IV).

Major problems arise when we study relationship conditions related to the binary IV. The percentages of correct p-values are between 39.70% and 54.23%. In these cases, the CE performs better in identifying correct p-values ($z=13.39$, CI=.27-.73, $p<.001$ for a distribution of .5 in the DV and .7/.4 in the IV and $z=21.45$, CI=.23-.77, $p<.001$ for the .2 distributed DV and .7/.4 distributed IV). Although the CE is better, this kind of estimate cannot provide percentages with respect to correct p-values higher than 54.23%.

When percentages of correct p-values for the continuous IV are computed, a correct and nearly identical pattern stands out between both estimation methods under the related conditions. As shown in Table 2, 100% correct p-values are achieved in all conditions, except when 99.85% are achieved for the CE and 99.75% for the BE under a .2-distributed DV and a symmetric IV with mean values of 100 in DV group 1 and 120 in DV group 0. Regarding unrelated conditions, BE outperforms the CE when IVs are symmetrically distributed for both .5-distributed ($p<.001$) and .2-distributed ($p<.001$) DVs. We identify important problems for both estimation methods under unrelated conditions when asymmetry is generated with respect to the continuous IV. A lower performance under BE is noteworthy. While the CE detects correct p-values at a rate of between 34.65% and 35.90%, BE shows a 0% correct p-values.

Coefficient estimates and standard errors were also computed. We present results for the binary IV (see Figure 1) and the continuous IV (see Figure 2) according to the two types of simulated conditions (i.e., unrelated and related). Note that, unrelated condition boxplots are those situated on the top of the figures 1 and 2, and related condition boxplots are those situated at the bottom of the figures 1 and 2. Boxplots were created to

enable a visual comparison of CE and BE in terms of coefficients and standard errors estimates. Figures 1 and 2 were drawn without outliers so if we include outliers in these figures, the scale substantially increases, and the shape of boxes cannot be appreciated. Outliers are defined as those points more than 1.5 times the interquartile range above the third quartile and those points more than 1.5 times the interquartile range below the first quartile. Moreover, the minimum and maximum values of the coefficients and the minimum and maximum values of the standard errors were obtained (including outliers). This allowed to keep the information provided by the presence of outliers. A glance at the binary IV boxplot under unrelated conditions shows that, in the two types of estimation, coefficient values are nearly zero, as expected. Note that a value close to zero indicates unrelated values in an LR model. Moreover, the CE for both .5- and .2-distributed DVs presents greater variability than the BE. Hence, the standard errors of CE present a considerable amount of variability as compared to those derived from BE. This should be taken into account, because a high variability in standard errors threatens the stability of the model's coefficients. When outliers are included in the analysis, we find considerable differences between CE and BE in minimum and maximum values corresponding to standard errors. When DV and IV are .5-distributed, the minimum and maximum values of the classical distribution are .53 and 3,384,213, respectively, whereas the minimum and maximum values of the Bayesian distribution are .50 and .92, respectively. When the DV is .2-distributed and the IV is .5-distributed, the minimum and maximum values of the classical distribution are .58 and 3,500,966, respectively, whereas the minimum and maximum values of the Bayesian distribution are .53 and 1.14, respectively. The simulation results under the related conditions show that the values approach 1 for CE and BE. As unrelated conditions, BEs present less variability than the CEs. With regard to standard errors, a greater variability for CE is again noteworthy (see Figure 1). When the DV is .5-distributed and the IV is .7/.4-distributed, the minimum and maximum values of the classical standard errors are .54 and 3,780,183, respectively, whereas the minimum and maximum values of the Bayesian distribution are .51 and 1.01, respectively. When the DV is .2-distributed and the IV is .7/.4-distributed, the minimum and maximum values of the classical distribution are .60 and 3,653,235, respectively, whereas the minimum and maximum values of the Bayesian distribution are .55 and 1.55, respectively.

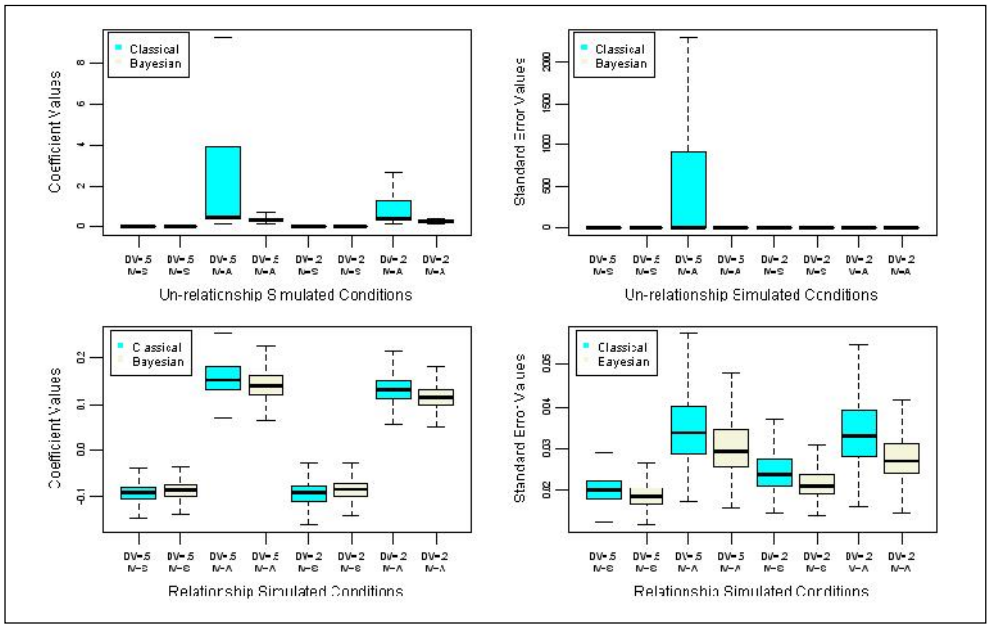


Note: Outliers are not drawn.

Figure 1. Coefficient estimates and standard errors of the binary independent variable from classical estimation and Bayesian estimation. Unrelated and related conditions in a logistic regression model.

Regarding the continuous IV studied under unrelated conditions, we can observe that the medians of the coefficients and the standard errors are close to zero. Furthermore, most distributions have almost no variability, while the minimum and maximum values between the two estimations are quite similar. Through a visual inspection of boxplots, consider the configuration that consists of a .5-distributed DV and an asymmetric IV distribution. In this case and in contrast to BE, the distribution values of the classical coefficient and standard error present variability. The minimum and maximum values of the classical distribution (including outliers) are .04 and 124,919.2, respectively, whereas the minimum and maximum values of the Bayesian estimates are .03 and .72, respectively. A somewhat similar pattern occurs for coefficients from the .2-distributed DV and asymmetric IV case, although the variability is lower (see Figure 2). In this case, the minimum and maximum values of the classical distribution are .03 and 342,992.7, respectively, whereas the minimum and maximum values of the Bayesian estimates are .03 and .24, respectively. Finally, regarding related

conditions, we observe fewer differences between classical and Bayesian coefficient estimates than in the previous analysis. The shapes of the classical distribution and the Bayesian distribution are quite similar, though slightly favoring the BE. Regarding standard errors, we also find more variability in CE. Conditions including a high median value of standard error are related to the generation of asymmetry with respect to the IV (see Figure 2). When outliers are included in the analysis, the ranges of the two types of estimates are quite similar. However, in most ranges, the CEs are remarkably wider than the BEs.



Note: Outliers are not drawn. Symmetric distributions are denoted as S, asymmetric distributions are denoted as A. Under unrelated conditions, means are equal across the two groups ($\mu_0 = \mu_1 = 100$). Under related conditions, means are different across the two groups ($\mu_1 = 100$, $\mu_0 = 120$).

Figure 2. Coefficient estimates and standard errors of the continuous independent variable based on classical estimation and Bayesian estimation. Unrelated and related conditions in a logistic regression model.

DISCUSSION

This study compares two types of estimations using simulated LR models. Bayesian methods rely not only on current knowledge (i.e., sample data) but also on prior information that may be available on the parameter of interest. In this study, BE involves introducing a weakly informative prior distribution into the LR models. This distribution was proposed by Gelman et al. (2008) as a default choice for logistic and other regression models. Prior distribution was not introduced into CE-LR models.

We generated LR models with two IVs, namely, a binary IV and a continuous IV. In the simulation procedure we specified models where values were related (related conditions) and un-related (unrelated conditions) to the DV. Regarding the binary IV, p-values were correctly detected by both the CE and BE under unrelated conditions. However, regarding related conditions, lower percentage values for both estimations were noted. The obtained coefficients and standard errors (without outliers) showed more variability in the distributions from CE. This was more notable under related conditions. When outliers were included in the distributions, CE provided extremely wide ranges in comparison to BE. This feature affected both unrelated and related conditions. Regarding the continuous IV, both estimations detected p-values correctly when related conditions were studied. However, poor performance of both estimations, especially the BE, was observed under unrelated conditions. This was particularly problematic with positive asymmetric distributions. A glance at the boxplots showed that most variability in asymmetric distributions was obtained under the unrelated conditions. We should point out that this variability is only accounted for with respect to CE. This is possibly explained by the presence of outliers, which produced distributions with considerably wide ranges.

Both boxplots and the range of distribution values suggest more variability in classical distributions compared to Bayesian distributions. This is possibly explained by the suggestion of Gelman et al. (2008) that including some actual prior information may regularize the extreme inferences that are provided by non-informative priors. According to them, the prior introduced into the model should have produced more stable and regularized estimations compared to the CE in which the prior had not been introduced.

There are, of course, several aspects that deserve further study; they are related to the present study's limitations. For instance, one important limitation in our study is that we worked with a unique sample size (n=100). Future work should take this into account by constructing LR models under

different sample sizes. Moreover, global fit of the model was not assessed. As stated earlier, extremely high standard error values from the CE have been obtained. These results indicate a problem in the convergence of the algorithm and it should be studied in future research. Other aspects that contribute to complete the work would be the study of bias and precision. Another aspect that deserves further study involves simulating other distribution values with respect to the exposure variable. In this paper, we worked with two conditions, namely, the same distribution of the exposure in the two groups (π_1=.5; π_0=.5) and different distributions of the exposure in the two groups (π_1=.7; π_0=.4). Future simulations could include more extreme differences between the distributions of the exposure variable among the groups with and without the disorder. From the variation of the distributional values, could be considered more intense relations between the DV and IVs. Complex models could have been defined (i.e. including other variables such as categorical IV with more than two categories, or studying interaction terms). Another interesting point would be to establish relationship between IVs and study the effect of collineality on CEs and BEs. Note that models based on clinical data can present collineality. It would be fruitful to design simulations similar to real data and study the obtained results.

Also to be noted is the interpretability of p-values. Future studies may include confidence interval values in order to obtain much information. Moreover and from a Bayesian viewpoint, it would be interesting to obtain probability values associated to different hypothesis. Another immediate follow-up study could include other kinds of prior distributions. This would provide a more extensive understanding about BE in SS-LR models. We acknowledge that what we denote as BE in this study is in fact a particular case of more general Bayesian inference. Note that a prior distribution can be non-informative, informative or very informative (Mila, et al.,2003). Following Gelman et al (2008), we have introduced a somewhat informative prior distribution that can be used in a wide range of applications. The above authors state that this default prior can be viewed as a baseline on top of which the user can add real prior information as necessary. Hence, two future research lines appear particularly fruitful. First, one line of research might involve studying results over SS-LR models based on real data that compares different kinds of priors. Informative priors would be constructed by incorporating relevant background knowledge. This prior knowledge may come from expert opinions, published experimental results, or a combination of both (Mila et al., 2003). Second, comparisons can be studied between non-informative, informative and very informative priors among SS-LR models via

simulation. It would be interesting to compare these priors with classical LR estimates. Finally, in both cases, research based on real data and/or simulation studies should address situations involving asymmetric samples, because this problem is not solved in the present study. Such research must take into account that distributions based on psychologically data are not often normally distributed. Hence, an approach for analyzing SS-LR models when distributions are asymmetric is much needed.

RESUMEN

Estimación en modelos de regresión logística en muestras pequeñas. Un estudio de simulación usando una distribución previa escasamente informativa. En este trabajo se utilizaron simulaciones para comparar el rendimiento de las estimaciones clásica y bayesiana en modelos de regresión logística utilizando muestras pequeñas. En las simulaciones realizadas, las condiciones fueron variadas, incluyendo el tipo de relación entre los valores de las variables dependientes e independientes (es decir, los valores vinculados y no vinculados), el tipo de variable (binario y continuo), y diferentes valores de la distribución binomial y la simetría (distribuciones simétricas y con asimetría positiva). La aplicación Iterativa de la estimación de mínimos cuadrados ponderados se utilizó como método de estimación para ajustarse a los modelos, tanto en la estimación clásica como en la bayesiana. Una distribución de tipo escasamente informativa fue elegida como la distribución a priori para la estimación bayesiana. Los resultados de la simulación muestran que las estimaciones bayesianas proporcionan una distribución más estable, pero que no son capaces de resolver los problemas generados por distribuciones asimétricas basadas en muestras pequeñas. Será preciso plantear nuevos trabajos en el ámbito del estudio del efecto de las distribuciones asimétricas utilizando diferentes tipos de distribuciones a priori.

REFERENCES

Bautista, D., Arana, E., Martí-Bonmatí, L., & Paredes, R. (1999). Validation of logistic regression models in small samples: application to calvarial lesions diagnosis. *Journal of Clinical Epidemiology, 52*, 237-241.

Cañadas, G., Lozano, L.M., de la Fuente, E.I., Vargas, C., & Saldaña, L. (2010). Análisis bayesiano de variables relacionadas con el desarrollo del síndrome de Burnout en profesionales sanitarios. *Escritos de Psicología, 3*, 33-39.

Chen, T.H., Chen, C.Y., Yang, H.C.P., & Chen, C.W. (2008). A mathematical tool for inference in logistic regression with small-sized data sets: a practical application on ISW-ridge relationships. *Mathematical Problems in Engineering, 2008*, 1-12. doi:10.1155/2008/186372

Claeskens, G., Aerts, M., & Molenberghs, G. (2003). Quadratic bootstrap method and improved estimation in logistic regression. *Statistics & Probability Letters, 61*, 383-394.

Dasgupta, N., & Chen, G. (2002). Some robustness issues for comparing multiple logistic regression slopes to a control for small samples. *Journal of Statistical Computation and Simulation, 72*, 925-935. doi:10.1080/0094965021000015486

Dasgupta, N., Pascual, F.G., & Spurrier, J.D. (2001). Small sample-techniques for comparing several logistic regression slopes to a standard. *Journal of Statistical Computation and Simulation, 71*, 141-161. doi:10.1080/00949650108812139

Dasgupta, N., Spurrier, J.D., Martinez, E., & Moore, B.C. (2000). Comparison to control in logistic regression. *Communications in Statistics-Simulation and Computation,* 29, 1039-1057. doi:10.1080/03610910008813653

Gelman, A., Jakulin, A., Pittau, M.G., & Su, Y.S. (2008). A weakly default prior distribution for logistic and other regression models. *The Annals of Applied Statistics, 2*, 1360-1383. doi:10.1214/08-AOAS191

Gelman, A., Su, Y.S., Yajima, M., Hill, J., Pittau, M.J., Kerman, J., & Theng, T. (2010). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models (Version 1.3-08) [Computer software]. Retrieved from http://CRAN.R-project.org/package=arm.

Gordóvil, A., Guàrdia, J., Peró, M., & de la Fuente, E.I. (2010). Classical and Bayesian estimation in the logistic regression model applied to diagnosis of child Attention Deficit Hyperactivity Disorder. *Psychological Reports, 106*, 519-533. doi:10.2466/PRO.106.2.519-533

Heinze, G., & Puhr, R. (2010). Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets. *Statistics in Medicine, 29*, 270-277. doi:10.1002/sim.3794

Hilbe, J. M. (2009). *Logistic regression models.* Florida: Taylor & Francis.

Jiménez, J.A., & Martínez, G. (2006). Una estimación del parámetro de la distribución g de Tukey. *Revista Colombiana de Estadística*, *29*, 1-16.

Lekdee, K., & Ingsrisawang, L. (2010). The empirical distribution of Wald, Score, Likelihood Ratio, Hosmer-Lemeshow (HL), and Deviance for a small sample logistic regression model. *Proceedings of the International MultiConference of Engineering and Computer Scientists*, *3*, 2062-2065.

Liao, J.G., & McGee, D. (2003). Adjusted coefficients of determination for logistic regression. *American Statistician, 57*, 161-166. doi:10.1198/0003130031964

Maiti, T., & Pradhan, V. (2009). Bias reduction and a solution for separation of logistic regression with missing covariates. *Biometrics, 65*, 1262-1269. doi:10.1111/j.1541-0420.2008.01186.x

Marrelec, G., Benali, H., Ciuciu, P., Pélégrini-Issac, M., & Poline, J.B. (2003). Robust Bayesian estimation of the hemodynamic response function in event-related BOLD fMRI using basic physiological information. *Human Brain Mapping*, *19*, 1-17. doi:10.1002/hbm.10100

Mila, A.L., Yang, X.B., & Carriquiry, A.L. (2003). Bayesian logistic regression of Soybean Sclerotinia Stem Rot prevalence in the U.S. North-central region: accounting for uncertainty in Parameter Estimation. In S. M. Coakley & D. E. Mathre (Chairs). *New Thesis Research Contributions to Plant Disease Epidemiology* (pp.758-764). USA: Milwaukee. doi: 10.1094/PHYTO.2003.93.6.758

Mittlbock, M., & Schemper, M. (2002). Explained variation for logistic regression-small sample adjustments, Confidence intervals and predictive precision. *Biometrical Journal, 44*, 263-272.

Nemes, S., Jonasson, J.M., Genell, A., & Steineck, G. (2009). Bias in odds ratios by logistic regression modeling and sample size. *BMC Medical Research Methodology, 9*, 1-5. doi:10.1186/1471-2288-9-56

Nijsse, M. (2003) A recursion formula for bayesian probabilities. *Psychological Reports*, *93*, 1214-1216.

Nottingham, Q.J., & Birch, J.B. (1998). A note on the small sample behavior of logistic regression in a bioassay setting. *Journal of Biopharmaceutlcal Statistics, 8*, 565-576. doi:10.1080/10543409808835260

Okada, K, & Shigemasu, K. (2010). Bayesian multidimensional scaling for the estimation of Minkowski exponent. *Behavior Research Methods, 42*, 899-905. doi:10.3758/BRM.42.4.899

Pearce, J., & Ferrier, S. (2000). An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling, 128*, 127-147.

Platt, R.W. (2000). Saddlepoint approximation for small samples logistic regression problems, *Statistics in Medicine, 19*, 323-334.

Pooi, A.H., (2003). Performance of the Likelihood Ratio Test when fitting logistic regression models with small samples. *Communications in statistics-Simulation and Computation, 32,* 411-418. doi:10.1081/SAC-120017498

Potter, D.M. (2005). A permutation test for inference in logistic regression with small- and moderate-sized data sets. *Statistics in Medicine, 24*, 693-708. doi:10.1002/sim.1931

R Development Core Team. (2010). R: *A Languaje and Enviroment for Statistical Computing* (Version 2.11.1) {Computer software}. Viena: R Foundation for Statistical Computing.

Schulz, E.M., Betebenner, D., & Ahn, M. (2004). Hierarchical logistic regression in course placement. *Journal of Educational Measurement, 41*, 271-286.

Steyerberg, E.W., Borsboom, G.J.J.M., van Houwelingen, H.C., Eijkemans, M.J., & Habbema, J.D. (2004). Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in Medicine, 23*, 2567-2586. doi:10.1002/sim.1844

Steyerberg, E.W., Eijkemans, M.J., Harrell, F.E., & Habbema, J.D. (2000). Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics in Medicine, 19*, 1059-1079.

Steyerberg, E.W., Harrell, F.E., Borsboom, G.J., Eijkemans, M.J., Vergouwe, Y., & Habbema, J.D. (2001). Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology, 54*, 774-784.

Van Houwelingen, J.C., & Le Cessie, S. (1990). Predictive value of statistical models. *Statistics in Medicine, 9*, 1303-1325.

Wang, L., & McArdle, J.J. (2005) A simulation study comparison of Bayesian estimation with conventional methods for estimating unknown change points. *Structural Equation Modeling, 15*, 52-74.

(Manuscript received: 21 June 2011; accepted: 23 November 2011)