

Logistic Regression with Incomplete Choice-Based Samples

MICHAEL FUGATE ACHLA MARATHE CLINT SCOVEL¹

Abstract

We have tested the Steinberg and Cardell estimator for incomplete choice-based samples where the independent covariates are generated by a non-uniform density and compared it with a benchmark procedure. We find that this estimator provides substantial improvement over the benchmark when evaluated with respect to misclassification error.

Key Words: Log likelihood; misclassification; relative improvement.

1 Introduction

In fraud detection efforts at Los Alamos National Laboratory, it is common to have data consisting of examples of fraud, but no examples of non-fraud. In our experience with the healthcare claims data and the IRS's (Internal Revenue Service) tax return data, it is often the case that small set of training data on abusive providers and fraudulent tax returns is available but no illustrations of good providers or good tax returns are available. In the health care arena where billions of dollars are being lost to fraudulent providers, a small increase in precision in modeling can save the tax payers millions of dollars.

The health care data contains providers in tens of thousands whereas only a small number (a few hundred) have labels of fraud, the rest of the providers have no labels at all. The labels on providers are binary i.e. either a provider belongs to the fraudulent set in which case he gets a value $y = 1$ or he belongs to the unknown set of providers. Each of these providers have a number of covariates associated with them. Covariates can take the form of ratios, averages, entropy, chi-square goodness of fit statistic etc.

One way to develop a predictor is to assume that the fraud sample is a random sample of the fraud population. A similar sized random sample can be taken from the complement of the fraud sample and which can be labeled as non-fraud. The resulting sample gives a training set that can be used to determine a predictor which we will call the S predictor. No attempts have been made to determine its asymptotic bias. However, empirical evaluations demonstrate it has enough predictive capability to make it useful.

Traditional sampling schemes for logistic regression are prospective sampling (also called exogenous sampling), and retrospective sampling (also called choice-based sampling). For rare

¹Los Alamos National Laboratory P.O. Box 1663, MS B265, Los Alamos NM 87545. Email: {fugate, achla, jcs}@c3.lanl.gov.

events or costly samples, retrospective sampling has many advantages and work has been done to construct estimators for such data. See Manski and Lerman [8] for example. However, in our applications, a complete choice-based sample is unavailable. Namely we do not have a sample of non-frauds. Cosslet [3, 4] has constructed an estimator when this data is supplemented by a random sample of data without dependent variable information, but it appears to be computationally intensive.

Steinberg and Cardell [10] have constructed an estimator which requires only a sample of the fraud population. We call this method the *SC* method. No labels are necessary for the rest of the population. Such an estimator can be determined from standard logistic regression packages as long as negative weights may be used in the loglikelihood function. In [2] they prove consistency of the estimator and derive the estimation of standard errors.

This paper compares the performance of estimators in terms of the misclassification rate, from the three methods i.e. *S* method, *SC* method and the simple logistic regression (*LR*) method. Note that the *LR* method has full information on the response variable y as it requires to know all the labels, hence a superior classification rate for the *LR* method is to be expected.

2 Theoretical Background

Manski and Lerman [8] give a nice discussion of a powerful technique for the determination of consistent estimators for choice-based samples. In short, it consists of writing a quasi-loglikelihood function that asymptotically converges to a function whose optimal parameter value is the correct value. One then applies Amemiya's lemma [1] to prove that the sequence of optimal parameter values converges to the correct value for almost all samples. One consequence of this technique is that if the asymptotic limit of the quasi-loglikelihood is the asymptotic limit of a loglikelihood of a consistent sample plan, the quasi-loglikelihood gives consistent estimates. That is our interpretation of the *SC* estimator which we now describe.

Let the data be (x, y) where x is a real vector and $y = 0$ or 1 is binary and the parameter of the regression $p(y = 1|x, \beta)$ is to be determined. If we take a random sample of size n the normalized conditional loglikelihood function is

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \log(p(y_i|x_i, \beta)) \\ &= \frac{1}{n} \sum_{i:y_i=1} \log(p(1|x_i, \beta)) + \frac{1}{n} \sum_{i:y_i=0} \log(1 - p(1|x_i, \beta)). \end{aligned}$$

Jennrich's [7] law of large numbers states that the limit becomes

$$p(1) \int \log(p(1|x, \beta))p(x|1)dx + p(0) \int \log(1 - p(1|x, \beta))p(x|0)dx.$$

If we add $p(1) \int \log(1 - p(1|x, \beta))p(x|1)dx$ to the second term and subtract the same from the first term and use the fact that $p(x|1)p(1) + p(x|0)p(0) = p(x)$ we obtain

$$p(1) \int \log\left(\frac{p(1|x, \beta)}{1 - p(1|x, \beta)}\right)p(x|1)dx + \int \log(1 - p(1|x, \beta))p(x)dx . \quad (1)$$

These integrals are with respect to $p(x|1)$ and $p(x)$ and can be approximated by a quasi-loglikelihood evaluated on data which consists of a choice restricted sample (a random sample from the population labeled 1) and a supplementary sample (a random sample from the general population without the 0 or 1 label). To be more precise, suppose we have $n_1 + n$ observations, where n_1 are sampled from $p(x|1)$ and so have $y = 1$, and n are sampled from $p(x)$ and have no y label.

We can then write the quasi-loglikelihood

$$\frac{p(1)}{n_1} \sum_{i=1}^{n_1} \log\left(\frac{p(1|x_i, \beta)}{1 - p(1|x_i, \beta)}\right) + \frac{1}{n} \sum_{j=1}^n \log(1 - p(1|x_j, \beta))$$

which is asymptotically the same as (1) but is instead evaluated on this complex sample. If we decompose this sum as

$$\frac{p(1)}{n_1} \sum_{i=1}^{n_1} \log(p(1|x_i, \beta)) - \frac{p(1)}{n_1} \sum_{i=1}^{n_1} \log(1 - p(1|x_i, \beta)) + \frac{1}{n} \sum_{j=1}^n \log(1 - p(1|x_j, \beta)),$$

we see that this is the likelihood function for a modified set of data consisting of a copy of the n_1 data points with their labels $y = 1$ along with another copy of those same n_1 data points with the labels $y = 0$ and all the n unlabeled supplementary samples given the labels $y = 0$ and with (not necessarily positive) weights in front of each term.

3 Implementation of the Algorithm

In the data we consider, the parent population consists of N random samples from $p(y, x) = p(y|x)p(x)$ for some conditional distribution $p(y|x)$ and some marginal $p(x)$. N_1 out of N have $y = 1$. From N we take n samples without replacement and observe only their x values. This is referred to as the supplementary sample. Out of the N_1 of the N that have the label $y = 1$, we sample n_1 without replacement. This is referred as the choice restricted sample. $r = \frac{n}{N}$ and $r_1 = \frac{n_1}{N_1}$ denote the sampling rates of the supplementary sample and the choice restricted sample respectively. Our simulations (and those of [10]) suggest that the dependence between the supplementary sample and the choice restricted sample has little effect.

Following Steinberg and Cardell [10] we summarize the data in the following way:

N = Size of the population randomly drawn from the infinite population.

n = Size of the supplementary sample.

N_1 = Size of the subpopulation with $y = 1$.

n_1 = Size of the choice restricted sample

r = sampling rate for the supplementary sample.

r_1 = sampling rate for the choice restricted sample.

We need a model for the probabilities in the quasi-likelihood function. We choose a logistic function

$$p(1|x, \beta) = \frac{e^{(\beta_0 + \beta_1 x_1)}}{1 + e^{(\beta_0 + \beta_1 x_1)}},$$

where $\beta = (\beta_0, \beta_1)$.

Since, $p(1) = N_1/N$, we can insert this into our quasi-loglikelihood and rescale to obtain the estimating equation of Steinberg and Cardell:

$$\frac{r}{r_1} \sum_{i=1}^{n_1} \log(p(1|x_i, \beta)) - \frac{r}{r_1} \sum_{i=1}^{n_1} \log(1 - p(1|x_i, \beta)) + \sum_{j=1}^n \log(1 - p(1|x_j, \beta)).$$

However, note that this is the only place where this formulation requires N and N_1 . Consequently, if instead we have $p(1)$ available, the N samples are not required: one simply has to come up with the n supplementary samples from $p(x)$ and the n_1 choice based samples from $p(x|1)$.

This quasi-loglikelihood can be regarded as taking each supplementary sample and giving it the label $y = 0$ and weight 1, and one copy of each choice restricted sample, giving it the label $y = 1$ and weight $\frac{r}{r_1}$ and another copy of each choice restricted sample and giving it the label $y = 0$ and weight $-\frac{r}{r_1}$. This pseudo-data can then be input directly into a standard logistic regression package as long as it accepts negative weights.² Table 0 puts the pseudo data succinctly.

Table 0

Data Layout for Empirical Estimation

Weight	Pseudo y	Actual y	Covariates	Sample
1	0	Unknown	\mathbf{x}	Supplementary
r/r_1	1	1	\mathbf{x}	Choice-restricted
$-r/r_1$	0	1	\mathbf{x}	Choice-restricted

It is important to point out that the share of the subpopulation with $y = 1$, i.e. N_1 , needs to be known in order to estimate the parameters.³ The sample estimate of the covariance matrix of the estimated parameters can be obtained by using only the n observations from the supplementary sample. Let \mathbf{X} be the $n \times (p + 1)$ matrix of the covariates, where x is a p -dimensional vector, and \mathbf{V} be the $n \times n$ diagonal matrix with diagonal elements $v_i = \hat{\pi}(x_i)[1 - \hat{\pi}(x_i)]$, where $\hat{\pi}(x_i)$ is the estimated logistic probability for the i th observation in the sample based upon the model

²We use ‘‘SPLUS’’ software to find the estimates of the parameters. The covariance matrix had to be separately coded. See [11] for details on SPLUS.

³The simulation results show that the Steinberg and Cardell [10] technique is robust to different values of N_1 . Even an estimate of N_1 which is 30% off its actual value causes the misclassification error(to be explained in the next section) to go up by less than 2%.

estimated from the data set containing the $n + 2n_1$ observations. Let \mathbf{T} be the $n \times n$ diagonal matrix with elements $t_i = 2(1 - n/N)\hat{\pi}(x_i)^2 + (w - 1)\hat{\pi}(x_i)$ where $w = (n/N)/(n_1/N_1)$. Cardell and Steinberg [2] show that

$$(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{T}\mathbf{X})(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}$$

is an estimator of the covariance matrix. From this matrix we calculate the asymptotically correct standard errors of the parameter estimates.

4 Simulation Results and Examples

We look at several examples to do the comparison between the three methods i.e. the S , SC and LR method. To observe the impact of different marginal distributions $p(x)$ on the performance of the three methods, we take $p(x)$ from a symmetric distribution i.e. a standard normal; a skewed distribution i.e. a χ^{\otimes} distribution with 5 degrees of freedom and finally a uniform distribution.

For each $p(x)$ we fix the parameter $\beta = \beta^*$ of the logistic function so that the conditional distribution

$$p(1|x, \beta^*) = \frac{e^{(\beta_0^* + \beta_1^* x_1)}}{1 + e^{(\beta_0^* + \beta_1^* x_1)}}$$

is determined. The marginal is sampled from N times. After each sample is obtained, its x coordinate is inserted into $p(1|x, \beta^*)$ and a Bernoulli sample is taken with that probability to determine either $y = 0$ or 1 . All N samples then have labels. In our experience, it is common to have $n = N$ and so $r = 1$, but the sampling rate for choice restricted sample can be small. We investigate with different r_1 for different $p(x)$. The choice restricted sample is made by first counting the number N_1 of the N which have $y = 1$. From these, we sample $n_1 = r_1 N_1$ randomly without replacement.

4.1 Bench Mark Techniques

We use two bench mark methods to evaluate the Steinberg and Cardell algorithm (the SC method). The LR method is simple Logistic Regression with complete information on the dependent variable using the full parent dataset. The S method is the Logistic Regression with no information on the dependent variable except for the choice restricted sample. The dependent variable in the S method is assigned a value of 0 for all supplementary samples and the choice restricted sample retains the value $y = 1$.⁴ To compare the methods we compute misclassification rates where we determine the classifier $y = 1$ if $p(1|x, \hat{\beta}) \geq 1/2$ for the estimated value of β i.e. $\hat{\beta}$, and $y = 0$ otherwise.

Note that the S method uses a value of $y = 0$ for all supplementary samples. If $p(y = 1)$ is high, the S method should lead to higher misclassification rates. Our empirical results, which are described in the next section, show how the three methods compare as $p(1)$ (which is a function of the coefficients β_0 and β_1) and $p(x)$ change. Given that the S method assigns $y = 0$ to all the supplementary samples, and the SC method pools the supplementary and choice restricted sample data in such a way that the log likelihood function can be estimated in a consistent way, it is not

⁴This ensures that Steinberg Cardell algorithm and the S method both have identical information available.

surprising that the S method does not perform as well. In the next section we quantify how much an improvement the SC method provides over the S method.

4.2 Examples

We consider a wide variety of examples and for all of them we choose $N = n = 10000$. In example 1, x comes from a χ^2 distribution with 5 degrees of freedom. $p(y = 1)$ of 0.20 implies that the size of the subpopulation with $y = 1$ is 2000. If r_1 is 0.30, the choice-based sample is 600. Table 1 shows three different scenarios for $x \sim \chi^2(5)$. We select three different combinations of $p(y = 1)$ and r_1 which are kept the same across the three examples. In Table 1 “Class Error” refers to the total misclassification error. NA means “not applicable”. Example 1a shows that the SC method gives a misclassification error which is only 2% better than the S method. Of course, the LR method performs the best because it has complete information on y . Indeed, logistic regression is known to be a consistent estimator of the β s that generated the data. Consequently, we can compute a relative improvement score

$$RI = \frac{\text{error}(SC) - \text{error}(S)}{\text{error}(LR) - \text{error}(S)}$$

which measures how much an improvement SC is over the S method with respect to the optimal LR method. This score is 0 when SC provides no improvement over the S method and 1 when it is as good as logistic regression. In example 1a, β_0 and β_1 are chosen such that $p(1|\beta_0 = -4, \beta_1 = 0.45) = 0.40$ is small. Here SC method shows only a slight improvement over the S method in terms of classification error but the relative improvement score is 0.43. In example 1b, β_0 and β_1 are chosen such that $p(1|\beta_0 = -2, \beta_1 = 1) = 0.60$. Now the classification error given by the SC method is about 6% lower than the S method but the relative improvement decreases to 0.29. In example 3a, with high $p(y = 1)$ and lower r_1 , the misclassification error of the SC method is almost 10% lower than the S method. Here the relative improvement score is only 0.22. This implies that a large absolute improvement does not necessarily translate into a large relative improvement score and vice versa.

Table 1

Monte Carlo Simulation Results: Example 1

Example 1a:		$x \sim \chi^2(5), \beta_0 = -4, \beta_1 = 0.45$					
Method	r	r_1	$p(1)$	$\hat{\beta}_0$	$\hat{\beta}_1$	Class Error	RI
S	NA	0.30	0.20	-4.02	0.20	20.23%	} 0.43
SC	1	0.30	0.20	-2.98	0.22	18.64%	
LR	NA	NA	0.20	-3.94	0.44	16.09%	
Example 1b:		$x \sim \chi^2(5), \beta_0 = -4, \beta_1 = 7$					
S	NA	0.20	0.40	-3.49	0.17	39.39%	} 0.29
SC	1	0.20	0.40	-2.14	0.20	33.76%	
LR	NA	NA	0.40	-3.91	0.68	19.77%	
Example 1c:		$x \sim \chi^2(5), \beta_0 = -4, \beta_1 = 1.1$					
S	NA	0.10	0.60	-3.41	0.12	60.95%	} 0.22
SC	1	0.10	0.60	-1.28	0.14	51.07%	
LR	NA	NA	0.60	-4.02	1.01	17.11%	

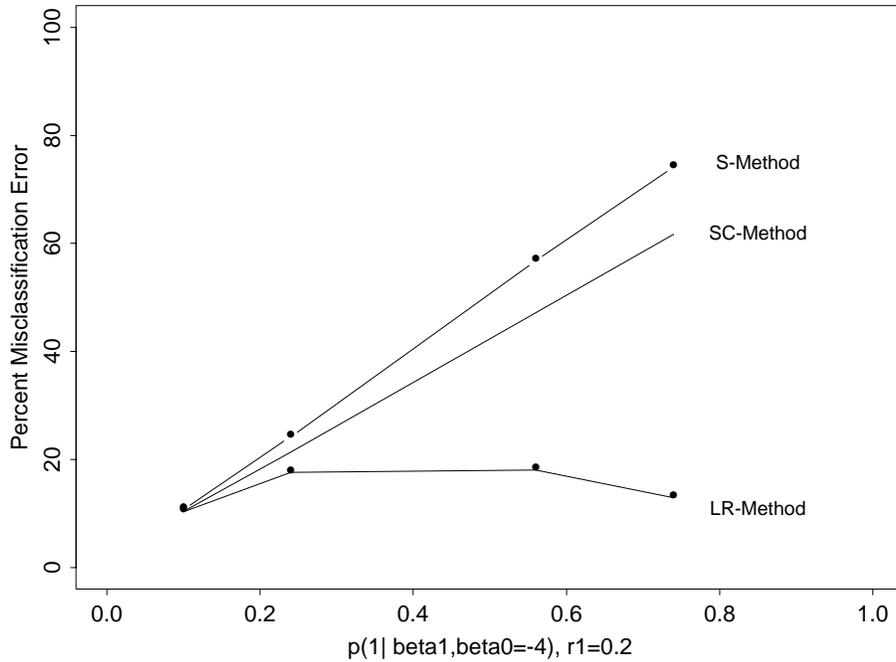


Figure 1: Misclassification Error for the Three Methods in Example 1

Figure 1 shows the performance of the three methods in terms of misclassification error as β_0 is kept constant at -4 and β_1 is varied to obtain different values of $p(y = 1|\beta) = \int p(1|x, \beta_0, \beta_1)p(x)dx$. The misclassification error rate of the *SC* method improves over the *S* method as $p(y = 1)$ increases.

Table 2 illustrates example 2 with $x \sim N(0, 1)$ and three different values of $p(y = 1)$ and r_1 . In the first case, example 2a, β_0 and β_1 are -3 and 4 respectively which result in $p(y = 1) = 0.20$. The classification error given by the *SC* method is about 4% better than the *S* method. The RI score is significant at 0.37. In example 2b where $p(y = 1)$ increases to 0.40, the *SC* method's classification error is 10% superior than the *S* method and has a relative improvement score of $RI = 0.35$. In example 2c, where $p(y = 1) = 0.60$, the *SC* method performs more than 15% better than the *S* method. In both Example 1 and 2 we have observed that as $p(y = 1)$ increases, the *SC* method shows more absolute improvement over the *S* method but the relative improvement goes down.

Table 2

Monte Carlo Simulation Results: Example 2

Example 2a:		$x \sim N(0, 1), \beta_0 = -3.8, \beta_1 = 4$					
Method	r	r_1	$p(1)$	$\hat{\beta}_0$	$\hat{\beta}_1$	Class Error	RI
<i>S</i>	NA	0.30	0.20	-3.88	1.53	18.86%	} 0.37
<i>SC</i>	1	0.30	0.20	-2.92	1.73	14.91%	
<i>LR</i>	NA	NA	0.20	-3.89	4.1	8.38%	
Example 2b:		$x \sim N(0, 1), \beta_0 = -1, \beta_1 = 4$					
<i>S</i>	NA	0.20	0.40	-2.84	0.95	41.95%	} 0.35
<i>SC</i>	1	0.20	0.40	-1.40	1.10	31.78%	
<i>LR</i>	NA	NA	0.40	-1.04	3.99	12.82%	
Example 2c:		$x \sim N(0, 1), \beta_0 = 1.2, \beta_1 = 4$					
<i>S</i>	NA	0.10	0.60	-2.91	0.57	61.56%	} 0.31
<i>SC</i>	1	0.10	0.60	-0.70	0.69	46.25%	
<i>LR</i>	NA	NA	0.60	1.26	4.09	12.51%	

Figure 2 shows the misclassification errors generated by the three methods in example 2 when β_1 is kept constant at 4 and β_0 is varied to get different values of $p(y = 1)$. The results are qualitatively the same as example 1, however, Figure 1 demonstrates that the *SC* method performs only marginally better than the *S* method at low levels of $p(y = 1|\beta)$ whereas in Figure 2, even at low levels of $p(y = 1|\beta)$, the *SC* method shows substantial improvement over the *S* method. At higher levels of $p(y = 1)$, Figure 2 shows *SC* method has better performance than the *S* method compared to Figure 1.

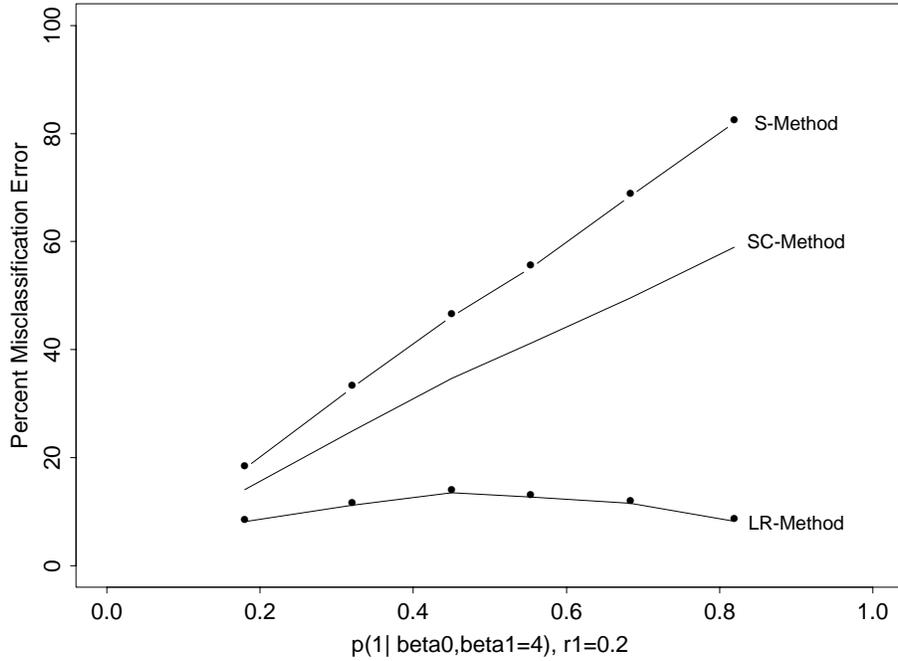


Figure 2: Misclassification Error for the Three Methods in Example 2

In example 3, we again analyze three different cases with $x \sim \text{unif}(0, 1)$. In example 3a, the values of β_0 and β_1 are kept at -4.4 and 5 respectively which result in $p(y = 1) = 0.20$. Table 3 shows the classification errors and parameter estimates obtained from the three methods. The value of r_1 is kept at 0.30. The *SC* method now performs exactly the same as the *S* method. Hence, the relative improvement score is $RI = 0.00$. Example 3b in Table 3 shows, at $p(y = 1) = 0.40$, the classification error given by the *SC* method is about only 3% lower than the *S* method though the relative improvement score is 0.22. In example 3c, with $p(y = 1) = 0.60$, *SC* method has 10% lower classification error than the *S* method with a relative improvement score of 0.28. Example 3 shows that as $p(y = 1)$ increases, the absolute and relative improvement of *SC* method over the *S* method increases.

Figure 3 shows the misclassification errors for example 3 where β_1 is kept constant at 5 and β_0 is varied in order to change $p(y = 1)$. At low levels of $p(y = 1|\beta)$, the *SC* and *S* method have very similar performance but as $p(y = 1)$ increases beyond 0.4, the *SC* method gives significantly lower classification error than the *S* method.

Table 3

Monte Carlo Simulation Results: Example 3

Example 3a:		$x \sim \text{unif}(0, 1), \beta_0 = -4.4, \beta_1 = 5$					
Method	r	r_1	$p(1)$	$\hat{\beta}_0$	$\hat{\beta}_1$	Class Error	RI
S	NA	0.30	0.20	-5.25	3.89	20.33%	} 0.00
SC	1	0.30	0.20	-4.11	3.89	20.33%	
LR	NA	NA	0.20	-4.45	5.12	18.27%	
Example 3b:		$x \sim \text{unif}(0, 1), \beta_0 = -3.1, \beta_1 = 5$					
S	NA	0.20	0.40	-4.04	2.62	39.10%	} 0.22
SC	1	0.20	0.40	-2.55	2.67	35.79%	
LR	NA	NA	0.40	-3.09	4.99	24.08%	
Example 3c:		$x \sim \text{unif}(0, 1), \beta_0 = -1.9, \beta_1 = 5$					
S	NA	0.10	0.60	-3.62	1.54	59.43%	} 0.28
SC	1	0.10	0.60	-1.41	1.59	49.43%	
LR	NA	NA	0.60	-1.84	4.97	23.61%	

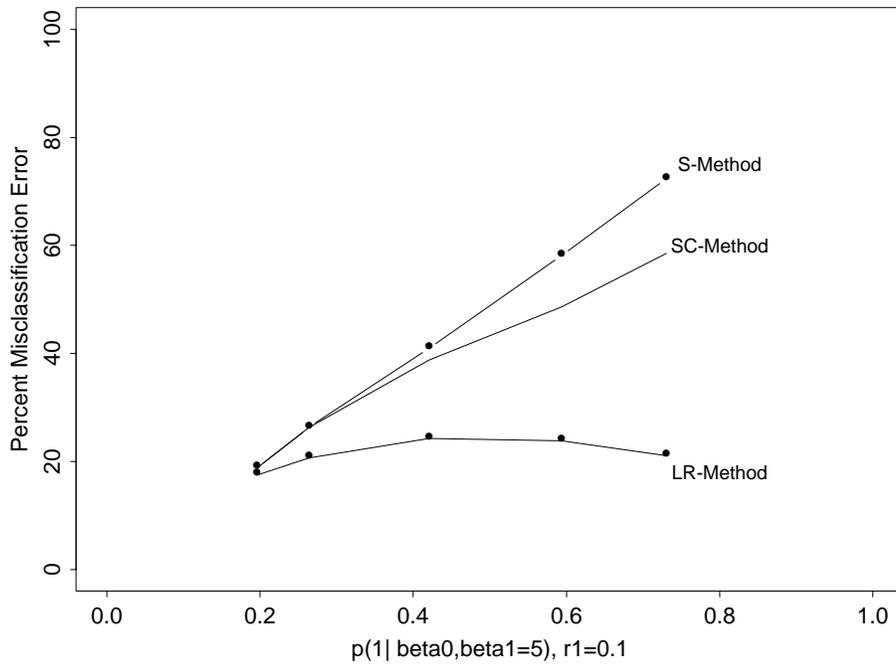


Figure 3: Misclassification Error for the Three Methods in Example 3

5 Conclusions

In this study, we compare the performance of the three methods i.e. S , SC and LR method using classification errors. S and SC methods work with incomplete information on the response variable whereas the LR method has full information on the response variable. We look at examples where the covariate comes from a $\chi^2(5)$ distribution, a standard normal distribution and a uniform distribution. Different parameters which potentially affect the performance of the methods are varied to observe their impact on the performance. We conclude that the SC method performs at least as good as the S method in all scenarios. The SC method shows substantial improvement over the S method when $x \sim N(0, 1)$ and $p(y = 1)$ is high.

References

- [1] Amemiya, T. (1973), Regression Analysis when the dependent variable is truncated normal, *Econometrica* **41**, 997-1016.
- [2] Cardell, N.S. and Steinberg D. (1991), Estimating Quantal Choice Models From Pooled Choice-based Samples and Supplementary Random Samples Without Choice Data, *preprint*.
- [3] Cosslet, S. R. (1981), Maximum likelihood estimator for choice-based samples, *Econometrica* **49**, 1289-1316.
- [4] Cosslet, S. R. (1981), Efficient estimation of discrete-choice models, *Structural analysis of discrete data with econometric applications*, 52-111, C.F. Manski and D. McFadden Eds., MIT press, Cambridge, Mass.
- [5] Hosmer, D. W. and Lemeshow S. (1989), Applied Logistic Regression, *John Wiley and Sons.*
- [6] Hsieh, D. A., Manski, C. F., and D. McFadden (1985), Estimation of response probabilities from augmented retrospective observations, *Journal of American Statistical Association* **80**, 651-662.
- [7] Jennrich, R. I. (1969), Asymptotic properties of non-linear least squares estimators, *Annals Math. Stat.* **40**, 633-643.
- [8] Manski, C. F., and Lerman, S. R. (1977), The estimation of choice probabilities from choice-based samples, *Econometrica* **45**, 1977-1988.
- [9] Manski, C. F., and D. McFadden, (1981), Alternative estimators and sample designs for discrete choice analysis, *Structural analysis of discrete data with econometric applications*, 2-50, C.F. Manski and D. McFadden Eds., MIT press, Cambridge, Mass.
- [10] Steinberg, D. and Cardell N.S. (1992), Estimating Logistic Regression Models When the Dependent Variable Has No Variance, *Communication Statistics - Theory Meth.* **21(2)**, 423-450.
- [11] Venables, W.N. and Ripley B.D. (1994), Modern Applied Statistics with S-Plus, *Springer-Verlag New York Inc.*