Stukel's Extended Logistic Regression Analysis with R

R ile Stukel'in Genişletilmiş Lojistik Regresyon Analizi

Vilda PURUTÇUOĞLU,[a]
Saygın KARAGÜLLE[a]

[a]Department of Statistics,
Middle East Technical University,
Ankara

Geliş Tarihi/*Received:* 27.04.2010
Kabul Tarihi/*Accepted:* 25.10.2010

Yazışma Adresi/*Correspondence:*
Vilda PURUTÇUOĞLU
Middle East Technical University,
Department of Statistics, Ankara,
TÜRKİYE/TURKEY
vpurutcu@metu.edu.tr

ABSTRACT Objective: For a logistic regression model, the degree to which predicted probabilities agree with actual outcomes can be expressed as a classification table. Being crucial in model adequacy checking, such tables may be slightly different when the same data are modeled with different statistical packages. The underlying reason is that when classifying a set of binary data, if the observations used to fit the model are also used to estimate the classification error, the resulting error-count estimate is biased. In order to cope with this problem, SAS suggests an algorithm, whereas the software is not publicly available. R is a free downloadable programme which is particularly designed for statistical computation, including the logistic regression analysis. The purpose of this study is to present a new function in R which carries out an extended logistic regression analysis of a binary data from the construction of its reduced-biased classification table, to the inference of its model parameters by calling the lrm(.) function under the Design package where necessary. **Material and Methods:** The performance of ext.logreg(.) is evaluated in terms of the accuracy of estimates and computational cost. **Results:** From the results of two binary datasets, it is observed that ext.logreg(.) via R estimates the model parameters and constructs the unbiased classification table as accurate as SAS programme under PROC logistic function without losing the computational demand. **Conclusion:** The free downloadable ext.logreg(.) function can be seen as an alternative computational tool in the analysis of logistic regression when the validation of predicted probabilities is essential.

Key Words: Logistic regression analysis; validation of predicted probabilities; unbiased classification table; R programme language

ÖZET Amaç: Lojistik regresyon modellerinde tahmin edilen olasılıkların gerçek değerlerle ne derece uyumlu oldukları sınıflandırma tablolusuyla ifade edilebilir. Model uyumlarının kontrollerinde bu tür tablolar, aynı veri farklı istatistiksel paket programlarıyla modellendikleri zaman aralarında çok küçük farklara sahip olabilirler. Bahsedilen farkın sebebi ise, iki-düzeyli tek değişkenli bir veri setinin sınıflandırmasında, eğer modele uyarlanmak için kullanılan gözlemler, sınıflandırma hatasını tahmin etmede de kullanılıyorsa, sonuçta bulunan hata-hesaplama tahmini yanlı olacaktır. Bu problemi çözmek için, SAS bir algoritma önermektedir. Buna karşın bu programlama dili herkesin kullanımına açık değildir. R, lojistik regresyon analizini de kapsayan, özellikle istatistiksel hesaplamalara yönelik olarak tasarlanmış, ücretsiz indirilebilen bir programdır. Bu çalışmanın amacı R'da kısaltılmış-yanlı sınıflandırma tablosundan, gerekli durumlarda "Design" paketi içinde bulunan lrm(.) fonksiyonunu çağırarak, model parametrelerinin tahminine kadar iki-düzeyli tek değişkenli bir verinin genişletilmiş lojistik regresyon analizini yapan yeni bir fonksiyon sunmaktır. **Gereç ve Yöntemler:** ext.logreg(.)'in performası, tahminlerin doğruluğu ve hesaplama maliyeti açısından değerlendirilmektedir. **Bulgular:** İki ayrı iki-düzeyli tek değişkenli veri setinden elde edilen cevaplara göre, R ile olan ext.logreg(.)'in model parametrelerini tahmin etmede ve yansız sınıflandırma tablosunu kurmada, hesaplama zamanında bir kayıp olmaksızın SAS programındaki PROC logistic fonksiyonu kadar doğru yaptığı gözlenmektedir. **Sonuç:** Ücretsiz indirilebilen ext.logreg(.) fonksiyonu, tahmin edilen olasılıkların doğrulanmasının gerekli olduğu lojistik regresyon analizinde, alternatif bir hesaplama aracı olarak görülebilir.

Anahtar Kelimeler: Lojistik regresyon analizi; tahmin edilen olasılıkların doğrulanması; yansız sınıflandırma tablosu; R programlama dili

Turkiye Klinikleri J Biostat 2011;3(1):49-56

In certain analysis where the outcomes of interest are binary such as the longitudinal study of the cardiovascular disease which categorizes women into two classes with respect to their menopause periods and their risk levels[1, 2] or the study about the relationship between coronary disease mortality and its possible risk factors,[3, 4] the researchers may encounter with computational problems in inference of the model parameters whose calculations are done via the ordinary least square method under certain assumptions or the maximum likelihood method. In order to handle such binary response variables, the logistic regression technique has been developed. Today this technique is widely applicable for advanced computational tools[5,6] and statistical package programmes such as the SAS programme language.[7] In this article, we present an R function, called ext.logreg(.), which enables us to implement a complete logistic regression analysis with binary data by constructing an unbiased classification table as the novelty, and calling the lrm(.) function within the "Design" R package (compatible any R versions ≥ 2.0.0) where necessary. The R software is a well-known free downloadable programme language which is particularly user friendly for statistical analyses. In the following section, we present the construction of the logistic regression in R. In Section 3, the details of the ext.logreg(.) function are given, and in Section 4, we implement it in the two real datasets. Finally we conclude our outputs in Section 5.

LOGISTIC REGRESSION MODEL WITH R

In a simple logistic regression where we deal with merely one predictor variable, the relationship between the binary outcome and continuous predictor variable disables us to use the ordinary least squares method in inference of the model parameters. The reason is that the sigmoidal shape of the response function causes a nonlinear relation under binary response variable (Figure 1). Indeed, even the response function is chosen as linear, the estimation becomes problematic when the error terms are non-normal and their variances are not constant.[8] So in such regression models, the underlying nonlinear relation between the predictor and the response, which is called the logistic response function, can be described as follows.

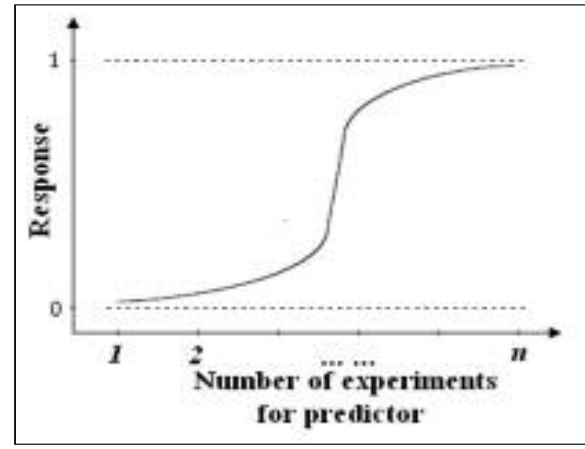

FIGURE 1: Possible relationship between a binary response variable and a continuous predictor variable whose total sample size is n.

$$E(Y_i) = \pi_i = \frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)} \quad (1)$$

where X_i and Y_i denote the th predictor and response, respectively, and β shows the regression coefficient. Equation 1 can be linearized by the logit transformation of the probability π_i, denoted by

$$\eta = \log it(\pi_i) = In\left(\frac{\pi_i}{1-\pi_i}\right) = X_i'\beta$$

Indeed, apart from the logit transformation, the mean response function π_i can be also linearized by the probit and the complementary log-log response function. Although both expressions have the same shape as the logistic response function, they are computationally problematic when we have more than one predictor variable.[8] Hereby we prefer logit, rather than its alternatives, as the linear transformation of π_i in our calculation.

In a logistic regression model, the parameters can be approximately estimated by different iterative techniques. But typically the Newton-Raphson, general optimization, and Iterated Weighted Least Squares (IWLS) method are chosen for the inference.[7] Both SAS and R choose the Newton-

Raphson approach for parameter estimation in such models.

On the other hand if the categorical data analyzed by the logistic regression have more than two levels in any explanatory variables, the calculation can be implemented by mainly two different approaches. The first approach considers that the predictor variables do not have ordered categories, hereby, the reference class for the levels of variable is chosen by the researcher. This assumption is used in the generalized logit model. Whereas the second approach accepts that the variable possesses increasing or decreasing ordered levels such as upper, middle, and lower classes. Thereby the calculation can be done via the cumulative logit model by choosing the lowest or the highest level as the reference class. In the analysis via SAS and R programmes, the generalized logit model is implemented by default to assign this class in estimation.

EVALUATIONS OF THE LOGISTIC REGRESSION MODEL WITH R

Once the logistic model is constructed, its effectiveness can be evaluated by means of four common criteria, namely overall model evaluation, statistical test of individual predictors, goodness of fit test, and validations of predicted probabilities[7] whose details are presented below.

OVERALL MODEL EVALUATION

If we need to compare two models such as, $\log it(\pi_i) = \beta_0$, as the baseline and $\log it(\pi_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$ as the full model under the k number of predictor variables, we use the overall model evaluation approach to assess any improvement in the fitted model over the baseline. Statistically, the improvement is likely to be true provided that the calculated p-value of the test statistic under the model of interest is less than the chosen significance level. Accordingly to evaluate the alternative models, we can compare them by the likelihood ratio, score, or Wald test. Among these three criteria, the last two ones are statistically equivalent to the first criterion. Although some programme languages such as SAS are capable of performing all tests, the likelihood ratio is more popular than others due to its computational simplicity. Because the assessment of both Wald and score tests require to perform vector-matrix operations which cause complexity in the derivation.[10] Moreover, although all the three tests serve for the same purpose, they may find different models. Under such conditions, the likelihood ratio and score tests outperform Wald's results.[7] Due to its underlying advantages, R uses merely the likelihood ratio statistics for the evaluation of models.

STATISTICAL TEST OF INDIVIDUAL PREDICTORS

The significance of the likelihood ratio test implies that at least one regression coefficient is statistically important. If we observe such significance under the likelihood result, the significances of individual model coefficients can be conducted by the Wald test. In the interpretation of the Wald result, if the calculated p-value for a regression coefficient is less than the chosen significance level α, we can conclude the necessity of the corresponding regression coefficient in the final model.[7]

GOODNESS OF FIT TEST

A goodness of fit test can be carried out to assess the fit of a logistic model against actual outcomes. In such validation, one inferential test and two descriptive measures are performed by SAS. As the inferential test, the Hosmer-Lemeshow test is conducted.[7] In this method, if the calculated p-value is greater than the chosen significance level α, it is considered that the model fits to the data well. As the alternative approach for this test, the Cox and Snell R^2, and Nagelkerke R^2 can be also used for the model descriptive measures.[7] Among these techniques, the correlation of determination R^2 can be easily interpreted in an ordinary least square regression model in the sense that it shows how much of the variation in the dependent variable can be explained by the chosen independent variables. Whereas this explanation is not valid for a logistic regression model, resulting in the computation of other highlighted R^2 values in place of this classical R^2 calculation.[7] R programme computes

the Nagelkerke R^2 for the goodness of fit test in logistic models.

VALIDATIONS OF PREDICTED PROBABILITIES

The calculation of the predicted probabilities of an event from its logit provides the opportunity to validate the results. In order to check whether there exists an association of high probabilities with events and association of low probabilities with non-events, five different measures, namely Kendall's Tau, Goodman-Kruskal Gamma, Somer's D, c-statistics, and classification table, are provided by SAS and R programmes.[7]

Kendall's Tau Measure

This statistic has a range from -1 to 1 where the value 1 indicates the perfect agreement, whereas, the values closer to -1 imply the perfect disagreement. The main drawback of this measure is that it provides no adjustment when there are too many ties. In logistic regression, if the two observations coming from the same predicted value indicate different outcomes, these observed outcomes are called *tied pair*. When the data have many ties, Kendall's Tau cannot reach -1 or 1.[11]

Goodman-Kruskal Gamma Measure

Similar to the Kendall's Tau measure, the gamma statistic lies from -1 to 1 where the values closer to 1 show the perfect agreement and the values closer to -1 represent the perfect disagreement. Moreover like the Kendall's Tau, the Goodman-Kruskal Gamma measure does not penalize ties.

Somer's D Measure

Somer's D considers the relation between the observation and the predicted value as a regression model in such a way that one variable presents the dependent variable (y= outcome) and others are taken as the independent variables (x= estimated probability). Then it adjusts the ties in the dependent variable. The range of Somer's D lies between -1 and 1, where the value 1 and -1 refer to the perfect agreement and the perfect disagreement, respectively. On the other hand there are two asymmetric forms of the Somer's D statistic, namely D_{xy} and D_{yx}. The former implies that y stands for the dependent variable and x is the independent variable, whereas, the latter accepts the reverse relation between y and x. Thereby only D_{yx} indicates the required degree of association. Both SAS and R compute D_{xy}, but SAS, additionally, can convert D_{xy} into D_{yx}.[7]

c-Statistic

The c-statistic represents the proportion of all possible pairs of subjects (one with event and one with none-event) in such a way that the model assigns higher probability for subjects with event. Different from other methods, the c-statistic ranges from 0.5 to 1. The values close to 0.5 present that the fitted model assigns the subjects randomly into the outcome categories, whereas, the values close to 1 imply that for each possible pair with different observed outcomes, the fitted model assigns a higher probability to subject with event in that pair. The c-statistic is believed to be an effective tool for the comparison of distinct logistic models fitted to the same data. In other words, the one with the highest c-statistic value can be taken as the best fitted model among alternatives.[7]

Classification Table

In this table, the observed values for the outcome variable and the predicted values obtained from the fitted model are cross-classified. The classification of predicted probabilities into the outcome categories is based on a chosen cutoff value, which can be specified by the researcher. There are four major statistics which can be produced in a classification table. These are the *sensitivity* which denotes the proportion of correctly classified events, the *specificity* which shows the proportion of correctly classified non-events, the *false positive* which represents the proportion of observations misclassified as events over all of those classified as events, and the *false negative* which displays the proportion of observations misclassified as non-events over all of those classified as non-events.

ext.logreg(.) FUNCTION

For the construction of a logistic model, we develop an R function which combines the best alternative

choices in the analysis of logistic data by calling the lrm(.) function in "Design" package with our new computation in the validation of the predicted probabilities. Accordingly our function uses the logit transformation as the linear transformation of and implements the Newton-Raphson method for the parameter estimation. In the analysis, it performs the likelihood ratio test for overall model evaluation and carries out the Wald test to detect whether individual regression coefficients are statistically significant. Moreover it uses the generalized logit model to represent reference classes in estimation. Finally it calculates the measures of associations to validate predicted probabilities. As a novelty, we construct the classification table which has not previously calculated in the R programme. In classification table, if the same observations are taken for both constructing model and predicting probabilities, the resulting table can be biased. A suggested plan[12] to unravel the biasness can be listed as follows:

i) Remove the observation from the data during the classification.

ii) Fit a logistic model using the remaining observations.

iii) Obtain the fitted probability of the removed observation using the newly fitted logistic model.

Among all the statistical packages, SAS is the only programme which can construct such unbiased classification tables.[7] Thereby the main advantage of our function ext.logreg(.) is that it can also construct these unbiased classification tables in R. More details about inputs, outputs, and codes of the ext.logreg(.) function can be found in the Appendix.

APPENDIX I: R Code of the ext.logreg() Function.

```
ext.logreg = function( formula, data, cutoff = 0.5) {

if(cutoff < 0 || cutoff > 1)
{
    cat("Error: Cutoff value must be between 0 and 1\n")
    end
}
cutoff = log(cutoff/(1-cutoff))
obs0.pred0 = 0
obs1.pred0 = 0
obs0.pred1 = 0
obs1.pred1 = 0
data  = model.frame (formula , data = data)
y     = model.response (data)
n     = nrow (data)
for (i in 1:n) {
        data.wo.ith.obs. = data[-i, ]
        temp.model       = lrm (formula, data = data.wo.ith.obs.)
        fitted.value     = predict (temp.model, data[i, ])

                           if (fitted.value  < cutoff & y[i] == 0)
                            { obs0.pred0  = obs0.pred0 + 1}
         else if(fitted.value  < cutoff & y[i] == 1)
          {obs1.pred0  = obs1.pred0 + 1}
         else if(fitted.value >= cutoff & y[i] == 0)
          {obs0.pred1 = obs0.pred1 + 1}
         else
          {obs1.pred1 = obs1.pred1 + 1}
}
Sensitivity        = ( (obs1.pred1 / (obs1.pred1 + obs1.pred0) ) * 100)
specificity        = ( (obs0.pred0 / (obs0.pred0 + obs0.pred1) ) * 100)
false.positive     = ( (obs0.pred1 / (obs0.pred1 + obs1.pred1) ) * 100)
false.negative     = ( (obs1.pred0 / (obs1.pred0 + obs0.pred0) ) * 100)
overall.correct    = ( ( (obs0.pred0 + obs1.pred1) / n ) * 100)
table1 = matrix( c(obs1.pred1, obs1.pred0, obs0.pred1, obs0.pred0),
          byrow = T, ncol = 2, dimnames = list("Observed"= c("1","0"),
          Predicted = c("1","0")))
table2 = matrix( c(Sensitivity, specificity, false.positive, false.negative, overall.correct),
          byrow = T, ncol = 5, dimnames = list( c("") , c("|Sensitivity(%)|","|Specificity(%)|",
          "|False positive(%)|","|False negative(%)|", "|Overall correct(%)|") ) )
model = lrm(formula, data = data)
print(model)
cat("The Observed and the Predicted Frequencies \n")
print(table1)
print(table2)
cat("\n")
}
```

IMPLEMENTATION

In order to assess the performance of the ext.logreg(.) function, we use two datasets in R (version 2.11.1). The first data, which include 98 observations, describe an epidemic outbreak of a disease that is spread by mosquitoes.[2] Accordingly the risk factors, i.e. predictors, of this data are the "age" (denoted by AGE), the "socioeconomic status" (denoted by, D_i, i= 1, 2 as the dummy variables), which is divided by three levels (lower, middle, and upper), and the "city sector" (denoted by CS),

APPENDIX II: The ext.logreg() Function.

Description

The function, named as "ext.logreg", provides a logistic regression analysis and a reduced-bias classification table for a given dataset. The function "ext.logreg" calls the "lrm" function under the "Design" R package (compatible any R versions 2.0.0) in order to obtain a fitted logistic model.

Usage

ext.logreg (formula, data, cutoff)

Arguments

formula statistical linear model of the form y ~ x1 + x2 + ...
where y : specified name of outcome variable in data file.
x1 : specified name of the first predictor variable in data file.
x2 : specified name of the second predictor variable in data file.
...
for a model with interaction between any two predictors, say & ,
y ~ x1 + x2 + x1 * x2 + ...

data name of the dataset

cutoff required cutoff value (optional) whose default is 0.5

Examples

ext.logreg (A ~ B + C + D + E, dataset1, 0.6)

ext.logreg (Y ~ X1 + X2 + X1 * X2, dataset2, 0.4)

ext.logreg (Lung Cancer ~ Smoking Status, data=dataset3)

Outputs

* Table of frequencies for the categories of outcome variables
* Statistics: Number of observations used in the fit,
Maximum absolute value of the first derivative of the log likelihood function,
Model likelihood ratio and associated degrees of freedom, and the p-value,
c-statistic,
Somer's D_{xy}
Goodman-Kruskal gamma,
Kendall's Tau,
Nagelkerke R^2,
Brier score.
* The fitted model:
Estimated regression coefficients with corresponding standard errors, Wald test values, and p-values
* Table of observed and predicted frequencies with the specified cutoff
* Sensitivity (%)
* Specificity (%)
* False positive (%)
* False negative (%)
* Overall correct (%)

which is classified into two groups. When creating the dummy structure for the "socioeconomic status" variable, the available data choose the "upper socioeconomic class" as the reference level. Hereby the indicator of this level is taken as 0. Then in D_1 column, the middle socioeconomic class values are assigned to 1 and the remainings are equated to 0, whereas in D_2 column, the lower socioeconomic class values are set to 1 and others are presented by 0. Finally the corresponding outcomes are given as the "disease status", categorized by presence (1) and absence (0) of the illness. The analysis of the data has been already performed in SAS (version 8 and version 9.2) via PROC logistic function. Hereby we analyze the same data by the ext.logreg(.) function in R. As the inputs of ext.logreg(.), we define three arguments: *i)* The logistic model is chosen for the fitted model, *ii)* the name of dataset on which the analysis is given, and *iii)* the cutoff value of the classification table is specified.

In the computation we set the cutoff to 0.4 and fit the following logistic model

ext.logreg (DS~AGE+D1+D2+CS, dataset, 0.4)

where (~) denotes the regression. The term on the right hand side of (~) shows the list of predictors and one on the left hand side presents the response. The fitted model and cutoff points are chosen as the same arguments of SAS in order to get comparable outputs and the results from both SAS and R programme are tabulated (Table 1). The outcome values indicate that ext.logreg(.) can construct the unbiased classification table and give the same statistics.

As the second implementation of our function to evaluate its performance in large datasets, we use the remedial reading recommendation data which have 189 observations.[7] In this dataset, the predictor variables are given as the "gender" (denoted by G), coded 1 for boys and 0 for girls, and the "reading score" (denoted by RS). On the other hand the outcome of interest is presented as the "recommendation for remedial reading class" (denoted by RRC), 1 coded for the recommendation and 0 for the non-recommendation. Similar to the analysis of the first data, initially we set the cutoff to 0.4 and then equate it to 0.1 to check its sensitivity and specificity. From the results of both programmes with distinct choices of cutoff points for the same

TABLE 1: Logistic regression analysis of the data about epidemic outbreak of a disease that is spread by mosquitoes in SAS (version 9.2) and R (version 2.11.1) via ext.logreg(.)

Analysis of Maximum Likelihood Estimates-SAS

Parameter	Estimate	Standard Error	p-value
Intercept	-2.3127	0.6426	0.0003
AGE	0.0297	0.0135	0.0276
D1	0.4088	0.5990	0.4950
D2	-0.3051	0.6041	0.6135
CS	1.5746	0.5016	0.0017

Analysis of Maximum Likelihood Estimates-R

Parameter	Estimate	Standard Error	p-value
Intercept	-2.31293	0.64259	0.0003
AGE	0.02975	0.01350	0.0276
D1	0.40879	0.59900	0.4950
D2	-0.30525	0.60413	0.6135
CS	1.57475	0.50162	0.0017

Testing the Global Null Hypothesis: β= 0

	Likelihood Ratio Statistic	Degree of Freedom	p value
SAS	21.2635	4	0.0003
R	21.26	4	0.0003

Association of Predicted Probabilities and Observed Responses

	Goodman-Kruskal			
	Kendall's Tau	Gamma	Somers' D	c-statistics
SAS	0.242	0.556	0.554	0.777
R	0.242	0.556	0.554	0.777

Classification Table with Cutoff: 0.4

		Correct		Incorrect	
	Cutoff	Event	Non- Event	Event	Non-Event
SAS	0.4	14	51	16	17
R	0.4	14	51	16	17

Percentages

	Correct	Sensitivity	Specificity	False Positive	False Negative
SAS	66.3	45.2	76.1	53.3	25.0
R	66.32653	45.16129	76.1194	53.33333	25

logistic model, it is seen that ext.logreg(.) is successful in constructing the classification table under both small and large datasets (Table 2).

TABLE 2: Logistic regression analysis of the data about the remedial reading recommendation in SAS (version 9.2) and R (version 2.11.1) via ext.logreg(.)

Analysis of Maximum Likelihood Estimates-SAS

Parameter	Estimate	Standard Error	p-value
Intercept	0.5361	0.8109	0.5085
G	0.6475	0.3248	0.0462
RS	-0.0262	0.0122	0.0324

Analysis of Maximum Likelihood Estimates-R

Parameter	Estimate	Standard Error	p-value
Intercept	0.53616	0.81088	0.5085
G	0.64749	0.32484	0.0462
RS	-0.02617	0.01223	0.0324

Testing the Global Null Hypothesis: β= 0

	Likelihood Ratio Statistic	Degree of Freedom	p-value
SAS	10.0334	2	0.0066
R	10.03	2	0.0066

Association of Predicted Probabilities and Observed Responses

	Goodman-Kruskal			
	Kendall's Tau	Gamma	Somers' D	c-statistics
SAS	0.118	0.276	0.273	0.636
R	0.118	0.276	0.273	0.636

Classification Table with Cutoff 0.4

		Correct		Incorrect	
	Cutoff	Event	Non- Event	Event	Non-Event
SAS	0.4	18	98	32	41
R	0.4	18	98	32	41

Percentages for Cutoff 0.4

	Correct	Sensitivity	Specificity	False Positive	False Negative
SAS	61.4	30.5	75.4	64.0	29.5
R	61.37566	30.50847	75.38461	64.0	29.49640

Classification Table with Cutoff 0.1

		Correct		Incorrect	
	Cutoff	Event	Non- Event	Event	Non-Event
SAS	0.1	59	2	128	0
R	0.1	59	2	128	0

Percentages for Cutoff 0.1

	Correct	Sensitivity	Specificity	False Positive	False Negative
SAS	32.3	100.0	1.5	68.4	0.0
R	32.27513	100	1.53846	68.44919	0

Finally we compare the computational demand of ext.logreg(.) and the SAS calculation in terms of the real and CPU (central processing unit) times (Table 3).

TABLE 3: The real and CPU time used in R (version 2.11.1) and SAS (version 9.2) for the analysis of two datasets.

Dataset	Program	Real time (in seconds)	CPU time
Epidemic Outbreak Data (size: 98)	R	1.89	0.00
	SAS	1.39	0.00
Remedial Reading Data (size:189)	R	2.95	0.00
	SAS	1.90	0.00

REFERENCES

1. Do K-A, Green A, Guthrie JR, Dudley EC, Burger HG, Dennerstein L. Longitudinal study of risk factors for coronary heart disease across the menopausal transition. Am J Epidemiol 2000;151(6):584-93.
2. Kutner MH, Nachtsheim CJ, Neter J, Li W. Applied Linear Statistical Models. 5thed. Boston: McGraw-Hill Irwin; 2005. p.577-611.
3. LaMonte MJ, Eisenman PA, Adams TD, Shultz BB, Ainsworth BE, Yanowitz FG. Cardiorespiratory fitness and coronary heart disease risk factors the LDS Hospital Fitness Institute cohort. Circulation 2000;102(14): 1623-8.
4. Kleinbaum DG, Klein M. Logistic Regression-A Self Learning Text. 2nd ed. New York: Springer-Verlag; 2002. p.5-32.
5. Boyacı A, Çağlı K, Ulaş MM, Avcı A, Ergün K, Emir M, et al. [The relationship between haematologic parameters and restenosis in patients with single vessel coronary artery disease following successful percutaneous invasive intervention]. Turkiye Klinikleri J Cardiol 2004;17(1):11-8.
6. Bilgili N, Akın B, Ege E, Ayaz S. [Prevalence of urinary incontinence and affecting risk factors in women]. Turkiye Klinikleri J Med Sci 2008;28(4):487-93.
7. Peng CJ, Lee KL, Ingersoll GM. An introduction to logistic regression analysis and reporting. The Journal of Educational Research 2002;96(1): 3-14.
8. Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis. 3rd ed. New York: Wiley; 2001. p.428-35.
9. Fox J. Fitting generalized linear models (Chapter 5). Writing programs (Chapter 8). An R and S-Plus Companion to Applied Logistic Regression. London: Sage Publications; 2002. p. 189-91, 273-8.
10. Hosmer DW, Lemeshow S. Multiple logistic regression (Chapter 2). Assessing the fit of the model (Chapter 5). Applied Logistic Regression. 2nd ed. New York: Wiley; 2000. p. 36-40, 147-9.
11. Yang K, Miller GJ. Handbook of Research Methods in Public Administration. 3rd ed. USA: CRC Press; 2007. p.395-8.
12. SAS Institute Inc. SAS/STAT 9.1 User's Guide (Version 8). Cary, NC: SAS Institute Inc.; 2004. p.1955-7.