On Graphically Checking Goodness-of-fit of Binary Logistic Regression Models

G. Gillmann[1, 2]; **C. E. Minder**[2, 3]

[1]Swiss Federal Statistical Office, Neuchâtel, Switzerland;
[2]Department of Social and Preventive Medicine, University of Bern, Bern, Switzerland;
[3]Horten Zentrum, University of Zurich, Zurich, Switzerland

Keywords
Nonlinear models, residuals

Summary

Objectives: This paper is concerned with checking goodness-of-fit of binary logistic regression models. For the practitioners of data analysis, the broad classes of procedures for checking goodness-of-fit available in the literature are described. The challenges of model checking in the context of binary logistic regression are reviewed. As a viable solution, a simple graphical procedure for checking goodness-of-fit is proposed.
Methods: The graphical procedure proposed relies on pieces of information available from any logistic analysis; the focus is on combining and presenting these in an informative way.
Results: The information gained using this approach is presented with three examples. In the discussion, the proposed method is put into context and compared with other graphical procedures for checking goodness-of-fit of binary logistic models available in the literature.
Conclusion: A simple graphical method can significantly improve the understanding of any logistic regression analysis and help to prevent faulty conclusions.

Methods Inf Med 2009; 48: 306–310
doi: 10.3414/ME0571
received: May 15, 2008
accepted: September 20, 2008
prepublished: March 31, 2009

Correspondence to:
Gerhard Gillmann
Swiss Federal Statistical Office
Espace de l'Europe 10
2010 Neuchâtel
Switzerland
E-mail: gerhard.gillmann@bfs.admin.ch

1. Introduction

Studies of the medical literature find frequent and enduring misuse of statistical methods, also in connection with modeling (see e.g. [1]). Along with other omissions, checking goodness-of-fit is severely under-reported and probably under-used. Given the reliance on model-based inference and conclusions, this is an unsatisfactory state of affairs. In this paper we consider checking goodness-of-fit for binary logistic regression models. We chose this class of models because they are widely used. Standard statistical wisdom says that once a statistical model has been chosen and fitted, it needs to be checked against the data it is applied to [2–5]. There are two general approaches to checking the goodness-of-fit of statistical models: formal tests of goodness-of-fit and graphical analyses of residuals. Several formal tests of goodness-of-fit for binary logistic models have been proposed [6–8]. See Le Cessie and van Houwelingen [9] for a description of the associated problems. These authors also propose a new test designed to avoid the problems associated with the previously suggested tests. The power of several goodness-of-fit tests for logistic regression is examined in [10]. Although some of the tests have theoretical weaknesses [9], favorable findings for detecting moderate nonlinearity with samples exceeding n = 100 were reported. Results were less favorable for detecting interactions and incorrectly specified link functions. From the point of view of good statistical practice, the use of one of the tests of goodness-of-fit recommended in [10] should become routine. This is not onerous on the data analyst, as, for example, a version of the Hosmer-Lemeshow test is implemented in many standard packages for logistic regression.

Because, given the fitted values, binary residuals are not informative, graphical approaches based on residuals are not very promising in the case of binary logistic regression (see [11, 12] and Discussion). In this paper, we focus on an alternative graphical approach based on model-based and model-free expectations originally proposed by Copas in [13]. This approach complements any formal test and provides information regarding the nature of any deviation indicated by the test. In addition, the graphical procedure proposed presents at a glance other useful information on the quality of the model otherwise not easily available. In Section 2 we present this graphical procedure.

In Section 3, we present three examples to show the extra information that may be gained by applying the proposed method. We have used this approach routinely in the last years and have found it to be helpful. Copas and Marshall gave in [14] a thorough and insightful account of the process of logistic modeling in a complex situation where they discuss the role of similar graphical procedures for checking goodness-of-fit.

In Section 4, we review other proposals of graphically checking goodness-of-fit found in the literature. Modifications of the proposed procedure and areas for future investigation are discussed as well.

2. A Simple Graphical Method

The following procedure can be implemented easily using standard statistical software such as R, SAS or STATA. It works as follows:

1. Fit the model and obtain the estimated probabilities $p_{fit,1}, \ldots, p_{fit,n}$ corresponding to the observations 1, ..., n.
2. Let $k = \min(10, \sqrt{n})$ and $m = [n/k]$, with [.] designating the integer part.

3. Sort the observations according to the size of their estimated probabilities $p_{fit,i}$ into k bins, each bin containing m (or m + 1) observations. The counts m + 1 arise as the remaining $n - k \times m$ fitted values must also be assigned to a bin.
4. For each bin l, calculate the average $p_{f,l}$ of the estimated probabilities $p_{fit,i}$ in the bin.
5. For each bin l, determine the fraction $p_{o,l}$ of 1's among the response values. The $p_{o,l}$ are a set of model-free estimates of the average probability of success of an observation in bin l.
6. Plot $p_{o,l}$ vs. $p_{f,l}$, $l = 1, ..., k$ into a square grid with x- and y-axes each covering the interval [0, 1].

Without doubt, many applied statisticians have used methods similar to the one proposed here (see Discussion). We were, however, not able to find a corresponding reference presenting and investigating the method in some detail. We have applied the procedure presented in the sequel routinely in our data analyses of biomedical and epidemiological studies. It has proven useful, especially in the context of multiple logistic regressions with a large number of realized values of the linear predictor; it may prove uninformative in cases with only a small number of such values. As a rule of thumb, at least 100 observations should be available; the smaller the values of k and m in (2) above, the less powerful is the method.

The three examples in the next section illustrate that the resulting graph conveys valuable information on the model fit not easily accessible otherwise.

3. Three Examples

The examples in this section all are taken from a geriatric intervention study with several years of follow-up [15]. The intervention was designed to reduce dependence on help and nursing home admissions among the elderly. Its effectiveness was judged by, among other things, dependence in basic activities of daily life (BADL). A person is deemed BADL-dependent if she requires assistance in performing one or more of the following activities of daily life: bathing, dressing, eating, transferring from bed to chair, and moving around inside her own residence. Another such measure is dependence in instrumental activities of daily life (IADL). A person is deemed IADL-dependent if he requires assistance with one or more of the following activities of daily life: cooking, handling finances, handling medications, shopping, using public or private transportation, and using the telephone. The strictest, and from a health care finance point of view most interesting, such measure is NHA, the permanent admission to a nursing home. NHA has the drawback of being heavily influenced by factors independent of the functionality of the person considered, such as the presence of a partner at home, the availability of home nursing care, etc.

We proceed to illustrate the use of the graphical method proposed above by three examples, one for BADL, one for IADL and one for NHA. Each example is chosen to illustrate some of the information one can gain by applying the method proposed.

3.1 Two Examples of Logistic Regression Showing no Important Discrepancies between Model and Data

► Figure 1 shows the relation of observed and predicted probability of being BADL-independent after three years of follow-up. The graph relates to an analysis of the baseline factors associated with BADL-independence three years after beginning the intervention: intervention (yes/no), age (years), risk stratum (low/high) and BADL-independence at baseline (yes/no). Stratum is a binary indicator of risk of future dependence on help determined at baseline. The graph provides the following information not readily available from a model fit alone.

- It shows that the model achieves good predictive power: its range of predicted probabilities averaged over deciles of BADL-independence $p_{f,l}$ extends from 9.5% to 95%. The corresponding bin average fractions $p_{o,l}$ range from 9.8% to 99%, showing good agreement.
- The relation between decile-averaged observed probabilities $p_{o,l}$ and model-based decile averages $p_{f,l}$ is approximately linear.
- Together, the first two points show that the bin average of the probability of becoming BADL-independent can be predicted well at baseline by the model.
- The model is well supported by data in the domain with probability of BADL-independence above 50%. However, there is a gap between the first and second decile, from about 10% to about 50%. This is due to the large effect of baseline BADL-dependence on 3-year BADL-dependence: a person that is BADL-dependent at base-

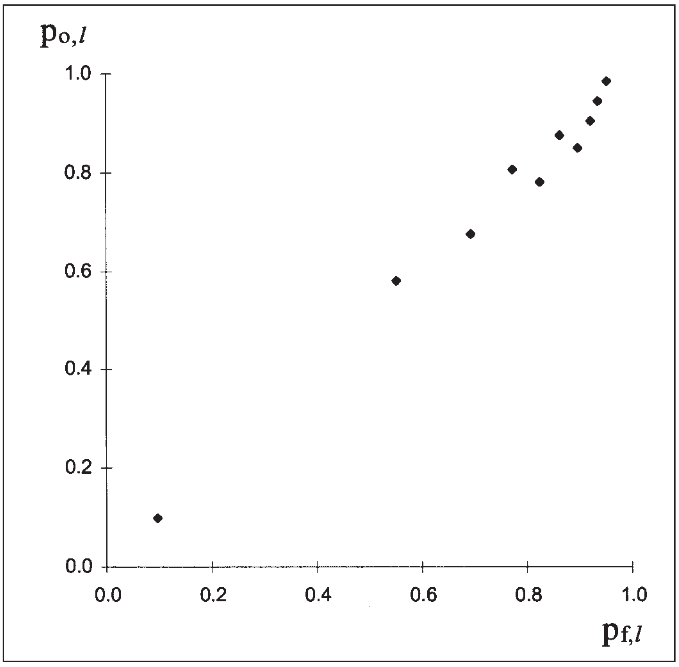

Fig. 1 A model with good predictive power, but with gaps in the support for the probability of independence in the basic activities of daily life (BADL-independence). X-axis: The values of $p_{f,l}$, l = 1, ..., k are bin averages of predicted values; y-axis: The values of $p_{o,l}$, l = 1, ..., k are bin averages of BADL. For details, see Sections 2 and 3.1.

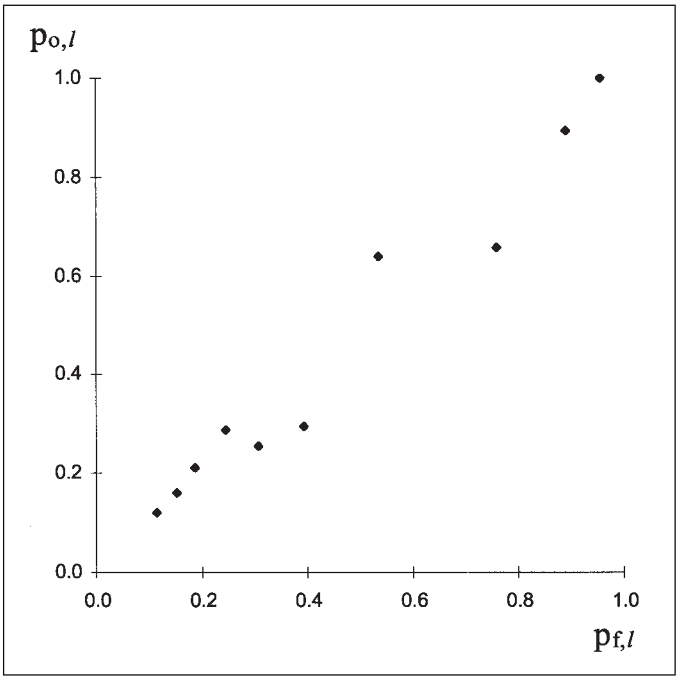

Fig. 2 A satisfactory model for the probability of dependence in the instrumental activities of daily life (IADL-dependence) with good predictive power and no gaps in its support. X-axis: The values of $p_{f,l}$, l = 1, ..., k are bin averages of predicted values; y-axis: The values of $p_{o,l}$, l = 1, ..., k are bin averages of IADL. For details, see Sections 2 and 3.1.

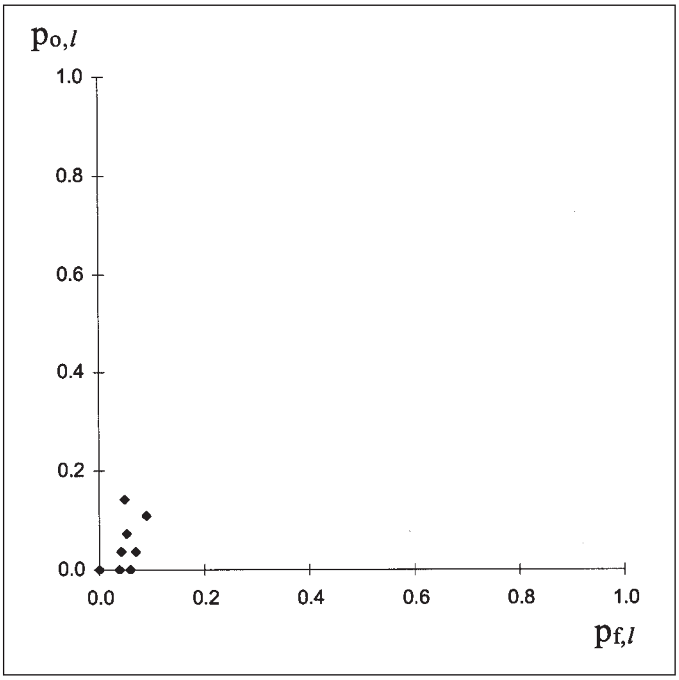

Fig. 3 A less than satisfactory model for the probability of nursing home admission (NHA) with very limited support (0% to 9.1% only). X-axis: The values of $p_{f,l}$, l = 1, ..., k are bin averages of predicted values; y-axis: The values of $p_{o,l}$, l = 1, ..., k are bin averages of NHA. Low predictive power and nonlinearities indicate a bad fit.

line is likely to be BADL-dependent three years later. This large gap also may mean a large bias in the position of the first two points. These points may be far away from the model curve, as the coordinates of each are the result of averaging over a large range.

The model appears to give a satisfactory representation of the study data and should permit reliable conclusions for the larger probabilities. The graph indicates that in future studies one might consider replacing the predictor "baseline BADL-independence" by an indicator permitting a finer gradation of the minor impairments of functionality to improve the model's predictive power in the lower probability segment of the population.

►Figure 2 shows a similar graph for a model of IADL-dependence at three years. This model does not exhibit such large gaps between the fitted probabilities and appears well supported by data in the range of between about 11% and 95% IADL-dependence. There is also less potential for serious bias in this case than in the first example.

3.2 An Example of a Less than Satisfactory Logistic Regression

►Figure 3 shows an analysis of the response "ever admitted to nursing home" (NHA) at the 3-year follow-up for the persons from the low-risk stratum at baseline. In this case, a naively interpreted logistic regression analysis indicates that the intervention results in a significant improvement. However, this conclusion may be premature due to the bad fit of the model. First, the range of the decile averages of fitted probabilities of nursing home admission varied between 0% and 9.1% only (observed probabilities in the deciles from 0% to 14.3%). Moreover, the graph shows substantial nonlinearity, suggesting that important variables were missing.

Using the information conveyed by the graph, one is lead to conclude that, despite a significant logistic regression, NHA at three years cannot be predicted well at baseline from the factors available.

4. Discussion

4.1 The Proposed Approach

In this paper, a simple but not widely used graphical method of assessing the goodness-of-fit of a logistic regression model is presented. The essence of the procedure goes back to Copas [13]. The method is related to the Hosmer-Lemeshow test [6]. In our experience, graphical methods such as the one presented are indispensable complements to routine model fits and formal tests. As is illustrated by the examples, the graphical approach presents at a glance important additional information about the properties and quality of the fitted model which otherwise is onerous to obtain. Through the proposed graphical representation, shortcomings and the predictive value of the model can be assessed immediately.

The scheme presented in Section 2 is only one of many possibilities of graphical representations of the model fit. Its advantages are that it is simple, easily implemented and easily communicated. In marked contrast to the proposals based on nonparametric smoothing discussed below, the graph proposed here illustrates well the data support of the model: the spacing of the points immediately conveys an idea in which regions there are little data to support (and test) the model. The graph also shows immediately how good the predictive quality of the model is. It is in this area where we have in fact found frequent use of similar graphs (see, for example, the excellent paper by Copas and Marshall [14]). Similar approaches are used to assess clinical predictive models (see, for example [16–18]). However, we have not found any paper giving a detailed description of the method and interpretation of the resulting graph.

Drawbacks of the proposed method include its arbitrariness as regards the number and choice of bins. It gives a discrete image of the continuous response curve and hence is coarse and biased. In theory, the method also suffers from a drawback similar to the one of the Hosmer-Lemeshow test [9]. In the context of the detailed information of the predictive power of the model obtained through our approach, we consider this to be of little practical relevance. Our extended experience with the method, as well as the usage of similar methods in the literature, lead us to conclude that the proposal made here is a reasonable compromise. As a rule of thumb about 10 bins are needed to limit bias and to permit plotting the course of the model prediction with sufficient detail. On the other hand, the number of observations per bin should not be too small in order to preserve some discriminatory power.

The proposed scheme may of course be modified and improved for specific situations, e.g., by varying the number and size of bins according to the configuration of independent variables or other specific needs. What is presented in Section 2 is a simple omnibus proposal, which in our experience has worked reasonably well in many cases. Copas [13, 14] used a data-driven nonparametric smoother such as LOWESS [19] to produce the model-free estimates of the probabilities. His approach is also applied in many other papers. Use of a nonparametric smoother reduces bias and eliminates the arbitrariness of choice of bin number, bin size and bin positions and so reduces the arbitrariness of the procedure. It has, however, the disadvantage of not exhibiting the extent of data support, which in the proposed method is conveyed by the spacing of the points. Further work could aim at combining the advantages and eliminating the drawbacks of the bin based and the nonparametrical smoother-based approach.

The following Subsections 4.2 to 4.4 contain a discussion of the difficulties as well as a short outline of the development of graphical methods for assessing model quality of logistic regression models.

4.2 Difficulties with Residual-based Methods in Logistic Regression

Let p be the estimated probability for an observation Y = 0, possibly depending on some independent variables. Due to the discrete nature of the binary responses, residuals can only take the two values –p and 1 – p, corresponding to Y = 0 or Y = 1. Plotting these residuals vs. p produces a graph with two distinct bands of residuals over the span of p, i.e., graphs of residuals from logistic regression always show structure, even if the model is correct. This is in marked contrast to ordinary regression with a continuous response variable, where the presence of structure in the residual plot indicates model-inadequacy. Therefore, residual plots from logistic regression models are difficult to interpret and hence of limited use. On a more subtle level, when interpreting graphs of residuals, we intuitively rely on approximate independence between residuals and fitted values, a feature not generally available in nonlinear models such as the logistic. Furthermore, approaches based on residuals are not very promising in the case of binary logistic regression, as the residuals contain only limited information beyond the estimated expected values p_i [11, 12]. These difficulties appear to be the principal reason why checking goodness-of-fit based on residuals is hardly done with logistic regression. For the same reason, most current methods for graphically checking goodness-of-fit in logistic regression rely on some kind of smoothing. For further comments on this topic, see also the introductions to the papers by Copas [13], Fowlkes [20], Cook and Weisberg [21] and section 2.3 of Pardoe and Cook [22].

4.3 Copas Plots and Extensions

Various authors have used plots based on Copas's suggestion of 1983 [13]. Most frequently the plots were applied to compare individual predictors with smoothed values of the observed probabilities. Theoretical papers applying such methods are, among others, Landwehr, Pregibon and Shoemaker [23], Spiegelhalter [16], and Fowlkes [20]. Cook and Weisberg proposed extensions to the Copas plot, plotting not only against p, but also against other linear combinations of covariables [21]. They term the resulting graphs marginal model plots. A judiciously chosen set of such graphs may prove useful in extended examinations of a model. Pardoe and Cook extended this work using a Bayesian approach to assess model uncertainty in marginal plots [22]. This paper also provides extensive and detailed justification of why marginal plots (and hence the Copas plot) are to be preferred to residual plots in logistic regression. The tool kit assembled by Cook, Weisberg and Pardoe provide a Bayesian alternative to simulation and bootstrapping suggested by Landwehr, Pregibon and Shoemaker in [23].

The excellent paper by Copas and Marshall is a must reading in regard to applications [14]. It contains a lucid description of the process of developing a logistic model and the role of the Copas plot in this process. Papers by Harrel et al. and Steyerberg et al. use the Copas plot in the development of prediction rules [17, 18]. The paper by Harrel et al. is intended as a tutorial and contains a variety of other model assessment methods beside the Copas plot.

4.4 Other Graphical Methods for Model Assessment

In 1968, Cox and Snell proposed a general definition of residuals, extending normal linear model theory to nonlinear models [24]. They worked out the binomial case with n exceeding 5, obtaining good results. The introduction of their paper gives a lucid description of the problem area. Pregibon proposes various diagnostic plots for the assessment of logistic regression models [25]. Although the residual-based methods proposed in this paper were subsequently questioned and now are no longer much used, the paper was seminal for further methodological developments. The plots showing the effects on parameter estimates and confidence intervals of deleting each observation in turn appear to be still useful.

Apart from Copas, Landwehr, Pregibon and Shoemaker were the first to use smoothing to improve properties of residuals and derived statistics [13, 23]. They present various ways for assessing local discrepancies between data and model, such as the local mean deviance plot and partial residual plots (where they not yet apply smoothing). As a global assessment tool, they propose Copas's plot, supplemented with (computer-intensive) simulated confidence bands. Fowlkes mentions as a drawback of Landwehr-style partial residuals the pronounced structure these residual exhibit, a feature due to their binomial nature [20]. He proceeds to develop smoothed residuals designed to remove this undesirable structure, and local test statistics based on smoothed residuals. For fairly large samples of $n = 400$ observations, Fowlkes's methods appear to work well.

Our literature review suggests that recent work elaborates the approach pioneered by Copas, a variant of which is suggested in the present paper as an omnibus graphical model assessment tool.

Acknowledgments

We thank our colleagues Ch. Schindler and A. Stuck as well as two reviewers for contributing to this paper by their support, insights, comments and suggestions as well as the permission to use data.

References

1. Strasak AM, et al. The Use of Statistics in Medical Research: A Comparison of The New England Journal of Medicine and Nature Medicine. The American Statistician 2007; 61: 47–55.
2. Davison AC. Statistical models. Cambridge: Cambridge University Press; 2003. Section 8.6.
3. Draper N, Smith H. Applied Regression Analysis. New York: Wiley & Sons; 1998.
4. Weisberg S. Applied Linear Regression. New York: Wiley & Sons; 1985.
5. Daniel C, Wood F, Gorman JW. Fitting Equations to Data: Computer Analysis of Multifactor Data. New York: Wiley & Sons; 1980.
6. Hosmer J, Lemeshow S. Applied Logistic Regression. New York: Wiley & Sons; 1989.
7. Brown CC. On a goodness-of-fit test for the logistic model based on score statistics. Communications in Statistics 1982; 11(10): 1087–1105.
8. Stukel TA. Generalized logistic models. Journal of the American Statistical Association 1988; 83: 426–431.
9. Le Cessie S, van Houwelingen JC. A Goodness-of-fit Test for Binary Regression Models Based on Smoothing Methods. Biometrics 1991; 47, 1267–1282.
10. Hosmer D, Hosmer T, le Cessie S and Lemeshow S. A Comparison of Goodness-of-Fit Tests for the Logistic Regression Model. Statistics in Medicine 1997; 16: 965–980.
11. Fienberg SE, Gong GD. Comment following the paper of Landwehr, Pregibon and Shoemaker. Journal of the American Statistical Association 1984; 79: 72–77.
12. Copas JB. Binary Regression Models for Contaminated Data (with discussion). Journal of the Royal Statistical Society B 1988; 50: 225–265.
13. Copas JB. Plotting p against x. Journal of the Royal Statistical Society C (Applied Statistics) 1983; 32: 25–31.
14. Copas JB, Marshall P. The offender group reconviction scale: A statistical reconviction score for use by probation officers. Journal of the Royal Statistical Society, Series C (Applied Statistics) 1998; 47: 159–171.
15. Stuck AE, Minder CE, Peter-Wuest I, Gillmann G, et al. A Randomized Trial of In-Home Visits for Disability Prevention in Community-Dwelling Older People at Low and High Risk for Nursing Home Admission. Archives of Internal Medicine 2000; 160: 977–986.
16. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Statistics in Medicine 1986; 5: 421–433.
17. Harrell FE, et al. Tutorial in Biostatistics. Development of a clinical prediction model for an ordinal outcome. Statistics in Medicine 1998; 17: 909–944.
18. Steyerberg EW, Vergouwe Y, Keizer HJ, Habbema JDF. Residual mass histology in testicular cancer: development and validation of a clinical prediction rule. Statistics in Medicine 2001; 20: 3847–3859.
19. Cleveland WS. Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association 1979; 74: 829–836.
20. Fowlkes E. Some Diagnostics for Binary Logistic Regression via Smothing Methods. Biometrika 1987; 74: 503–515.
21. Cook RD, Weisberg S. Graphics for Assessing the Adequacy of Regression Models. Journal of the American Statistical Association 1997; 92: 490–499.
22. Pardoe I, Cook RD. A Graphical Method for Assessing the Fit of a Logistic Regression Model. The American Statistician 2002; 56: 263–272.
23. Landwehr JM, Pregibon D, Shoemaker AC. Graphical Methods for Assessing Logistic Regression Models. Journal of the American Statistical Association 1984; 79: 61–71.
24. Cox DR, Snell EJ. A General Definition of Residuals. Journal of the Royal Statistical Society B 1968; 30: 248–275.
25. Pregibon D. Logistic Regression Diagnostics. The Annals of Statistics 1981; 9 (4): 705–724.

