Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 52 (2008) 2703–2713

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Increasing the power: A practical approach to goodness-of-fit test for logistic regression models with continuous predictors

Xian-Jin Xie[a,*], Jane Pendergast[b], William Clarke[b]

[a] *Division of Biostatistics, Department of Clinical Sciences and Simmons Comprehensive Cancer Center, The University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390, USA*

[b] *Department of Biostatistics, The University of Iowa College of Public Health, Iowa City, IA 52242, USA*

Received 23 April 2007; received in revised form 25 September 2007; accepted 27 September 2007
Available online 2 October 2007

Abstract

When continuous predictors are present, classical Pearson and deviance goodness-of-fit tests to assess logistic model fit break down. The Hosmer–Lemeshow test can be used in these situations. While simple to perform and widely used, it does not have desirable power in many cases and provides no further information on the source of any detectable lack of fit. Tsiatis proposed a score statistic to test for covariate regional effects. While conceptually elegant, its lack of a general rule for how to partition the covariate space has, to a certain degree, limited its popularity. We propose a new method for goodness-of-fit testing that uses a very general partitioning strategy (clustering) in the covariate space and either a Pearson statistic or a score statistic. Properties of the proposed statistics are discussed, and a simulation study demonstrates increased power to detect model misspecification in a variety of settings. An application of these different methods on data from a clinical trial illustrates their use. Discussions on further improvement of the proposed tests and extending this new method to other data situations, such as ordinal response regression models are also included.

Keywords: Continuous covariates; Cluster analysis; Generalized linear model; Goodness-of-fit test; Logistic regression

1. Introduction

Generalized linear models, and in particular, logistic regression models, are widely used in biomedical research fields. One component of model-fitting is the identification and specification of potential covariates to be included in the linear predictor. Estimation using maximum likelihood and testing the significance of the regression coefficients using either Wald or score tests are usually key goals of the analysis (Cox and Snell, 1989). Significance testing of each coefficient provides information about the relationship between the covariate and response, relative to overall variability. Goodness-of-fit tests, on the other hand, reflect whether the predicted values are an accurate representation of the observed values. Omitted predictors, a misspecified form of the predictor, or an inappropriate link function can all result in poor predictions.

[*] Corresponding author.
E-mail address: xian-jin.xie@utsouthwestern.edu (X.-J. Xie).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.09.027

If the regression of the response variable on treatment and covariates is linear or exponential, omission of important covariates only reduces the efficiency of the regression coefficient estimates, but has no effect on the consistency of the estimation. For logistic regression, this omission not only reduces the efficiency of the coefficient estimation, but also affects the consistency of the coefficient estimation. It leads to biased estimates of treatment effect, even in randomized experiments (Gail et al., 1984, 1988; Hauck et al., 1991; Robinson and Jewell, 1991). In case-control studies, the consistency of the estimators of the population odds ratio is still maintained if a correct logistic regression model is specified (Anderson, 1972; Xie and Manski, 1988; Nagelkerke et al., 1995, 2005).

The widely used chi-square statistic can be used as a measure of how far observed sample data deviate from a theoretical model providing expected counts in each of G distinct covariate patterns. Under the assumption of no lack-of-fit and suitable regularity conditions, the test statistic is asymptotically distributed as central chi-square with (G-k-1) degrees of freedom, where k is the number of regression parameters in the model (not counting the intercept). The deviance statistic can also be used for assessing the goodness-of-fit (Nelder and Wedderburn, 1972; Williams, 1987; Agresti, 1990). Under the assumption of a particular form of the underlying distribution, the deviance $(2 * [LL_s - LL_f])$ is a measure of the difference between the log likelihood of the fitted model (LL_f) and the log likelihood of the saturated model (LL_s). Under suitable regularity conditions and a properly specified model, the deviance statistic has approximately a chi-square distribution with G-k-1 degrees of freedom, where G is the number of distinct covariate patterns. These two statistics fall within a family known as power divergence statistics (Cressie and Read, 1984; Read and Cressie, 1988). Although the deviance and Pearson chi-square statistics are routinely provided in most statistical packages, their chi-square limiting null distribution is only valid when the number of observations in each covariate pattern is large. However, this condition is often unrealistic when a large number of categorical covariates or continuous covariates are present in the model.

The Hosmer–Lemeshow statistic (Hosmer and Lemeshow, 1980, 1989) is a practical goodness-of-fit chi-square test for general logistic regression situations, including those with continuous predictors. To implement this test, the predicted probabilities are grouped into G bins according to either data-driven percentiles of the estimated probabilities or prespecified fixed cutpoints. The test statistic is calculated by comparing the observed frequency (O_g) to the average predicted frequency (E_g) in the cell g, $g = 1, 2, \ldots, 2G$, via the familiar form of the statistic $X^2_{HL} = \sum_{g=1}^{2G} (O_g - E_g)^2/E_g$. Simulation studies indicate that, under the null hypothesis of no model lack-of-fit, the Hosmer–Lemeshow statistic can be approximated by a chi-square distribution with $G - 2$ degrees of freedom. The Hosmer–Lemeshow statistic is widely used due to its following properties: (1) it is intuitively appealing and easy to compute; (2) it has sound support from simulation studies; and (3) it is widely available in computer packages. In addition to these properties, lack of a better approach also contributes to its popularity. However, it has the following deficiencies (Hosmer et al., 1997; Pigeon and Heyse, 1999; Hosmer and Hjort, 2002; Kuss, 2002): (1) its limiting distribution has not been rigorously derived; (2) it is a conservative test and has low power to detect specific types of lack of fit (such as nonlinearity in an explanatory variable); (3) it is highly dependent on how the observations are grouped; (4) if too few groups are used to calculate the statistic (for instance, five or fewer groups), it will almost always indicate that the model fits the data; and (5) when the Hosmer–Lemeshow statistic indicates a lack of fit, it may be difficult to identify what types of subjects are not modeled well.

The Tsiatis goodness-of-fit statistic (Tsiatis, 1980) uses a different approach. Instead of grouping observations by their predicted outcomes, he partitions the multidimensional space of covariates into m distinct regions. An additive region effect for each region is added to the model to measure regional lack-of-fit. A score statistic is used to test that all of the m regional effects are zero. Tsiatis' procedure is as follows: (1), the space of covariates matrix $(x_1, x_2, \ldots, x_k)'$ is partitioned into G distinct regions in k-dimensional space denoted by $R_1, R_2, \ldots, R_G$. The indicator functions $I^{(j)}$ $(j = 1, 2, \ldots, G)$ are defined by $I^{(j)} = 1$ if $(x_1, x_2, \ldots, x_k)' \in R_j$ and $I^{(j)} = 0$ otherwise; (2), the model considered is $\ln[\pi_i/(1-\pi_i)] = \beta' X_i + \gamma' I_i$, where $\beta' = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$, $X_i' = (1, x_{1i}, x_{2i}, \ldots, x_{ki})$, $I_i' = (I_i^{(1)}, I_i^{(2)}, \ldots, I_i^{(G)})$, and $\gamma' = (\gamma_1, \gamma_2, \ldots, \gamma_G)$. Note that $\beta' X_i$ models all the original covariates and $\gamma' I_i$models the regional shifts; (3), a score statistic is then constructed to test that $\gamma_1 = \gamma_2 = \cdots = \gamma_G = 0$. More details are provided in Section 2.2 score statistic. Tsiatis's approach is conceptually elegant, but it lacks a general rule for how to partition the covariate space, especially when continuous covariates are present. How to choose the number of distinct regions m has also remained largely unstudied.

Pulkstenis and Robinson (2002) presented a goodness-of-fit method which draws on the notable strengths of both the Hosmer–Lemeshow approach and Tsiatis approach. Their method provides guidance on the choice of G and how