Journal of Multivariate Analysis 150 (2016) 14–26

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Lack of fit tests for linear regression models with many predictor variables using minimal weighted maximal matchings

Forrest R. Miller[a], James W. Neill[b,*]

[a] Department of Mathematics, Cardwell Hall, Kansas State University, USA
[b] Department of Statistics, Dickens Hall, Kansas State University, USA

ARTICLE INFO

Article history:
Received 25 August 2015
Available online 19 May 2016

AMS subject classifications:
62J05
62F03

Keywords:
Linear regression
Lack of fit
Many predictors
Matchings

ABSTRACT

We develop lack of fit tests for linear regression models with many predictor variables. General alternatives for model comparison are constructed using minimal weighted maximal matchings consistent with graphs on the predictor vectors. The weighted graphs we employ have edges based on model-driven distance thresholds in predictor space, thereby making our testing procedure implementable and computationally efficient in higher dimensional settings. In addition, it is shown that the testing procedure adapts to efficacious maximal matchings. An asymptotic analysis, along with simulation results, demonstrate that our tests are effective against a broad class of lack of fit.

1. Introduction

We consider the problem of testing the adequacy of a linear regression model

$$\mathbf{y} = [\mathbf{1}_n, X]\beta + \mathbf{e}$$

where the rows of X consist of the predictor vectors $x_i = (x_{i1}, \dots, x_{ip}) \in \mathbb{R}^p$, $1 \le i \le n$, which are possibly higher dimensional. In addition, $\mathbf{e}^\top = (e_1, \dots, e_n)$ is a Gaussian distributed random vector with independent components having $\mathrm{E}(e_i) = 0$ and $\mathrm{E}(e_i^2) = \sigma^2$, and $\mathbf{1}_n \in \mathbb{R}^n$ denotes the vector consisting entirely of ones to accommodate an intercept in the model. In the following we let $W = [\mathbf{1}_n, X]$. Since parametric regression models with many predictors are frequently used in the natural and social sciences, it is important to first check the adequacy of a proposed model to avoid misleading inferences. In this paper we present regression lack of fit tests for the case of many predictors which are effective against a large class of lack of fit and are computationally efficient.

Our testing procedure involves the development of a supremum-type multiple test based on a collection of Fisher statistics, each constructed from a matching on the predictor vectors x_i. General alternatives for defining the Fisher statistics use minimal weighted maximal matchings consistent with graphs on the x_i. These graphs have edges weighted according to model-driven distance thresholds in predictor space. The matchings are constructed edge by edge, at each stage choosing the edge with smallest possible weight. It is the use of matchings which makes feasible the construction and implementation of our lack of fit tests when p is large, although as seen in Section 2 it is necessary to have $n > 2p + 2$.

[*] Corresponding author.
E-mail address: jwneill@k-state.edu (J.W. Neill).

http://dx.doi.org/10.1016/j.jmva.2016.05.005

In previous work, Miller et al. [20,21] developed a graph theoretic representation of near replicate clusterings of statistical units to obtain lack of fit tests for linear regression models. This work helps provide a framework for generalizing the classical test presented by Fisher [11], an approach for testing linear regression lack of fit investigated by several authors as referenced by Miller et al. [20,21]. Although presented in generality, implementation of our previous tests was focused on models with lower dimensional predictor vectors. In particular, a graph on the predictor vectors was used to determine a special collection of clusterings (the atoms consistent with the graph), and then an optimization procedure (a maximin method or restricted least squares approach) was applied to choose an optimal clustering to test $\mathrm{E}(\mathbf{y}) \in W$ (by which we mean $\mathrm{E}(\mathbf{y})$ is an element of the column space of W). In the current work, we use a very different special collection of clusterings consistent with the graph. These are the clusterings that group at most two vertices together, provided these two vertices form an edge of the graph, which is why we call them edge clusterings. They are called matchings in the field of combinatorial optimization. Edge clusterings possess special advantages for testing regression lack of fit, some of which were discussed in more recent work by the authors (Miller and Neill [19]). However, they also allow efficient implementation in higher dimensional models with many predictor variables, which is the emphasis of the current work.

This paper is organized as follows. In Section 2 we first determine weighted graphs with vertices given by the $\{x_i\}$ and edges based on distance thresholds. Matchings on such graphs are then used to determine subspaces of the lack of fit space, and subsequently to construct lack of fit tests. The asymptotic behavior of such tests for a broad class of underlying true data generators is given, as well as the large sample behavior of matching sequences on a hypercube in $\mathbb{R}^p$. In addition, minimal weighted maximal matchings useful for detecting misspecification associated with the model $\mathrm{E}(\mathbf{y}) \in W$ are defined in Section 2. A computationally efficient algorithm to determine such matchings is presented in Section 3, along with a multiple testing procedure which follows Baraud et al. [2]. Given the unknown nature of any underlying lack of fit, this procedure is implemented with a model-driven set of matchings to enhance detection of model inadequacy associated with the specified model $\mathrm{E}(\mathbf{y}) \in W$. Section 4 provides the results of a simulation study for the cases $p = 10$ and $p = 20$. These results demonstrate efficient implementation of our testing procedure to effectively detect general lack of fit in linear regression models with many predictors. Proofs of theorems are given in the Appendix.

An extensive literature exists which addresses the problem of testing the adequacy of a specified parametric regression model. This work includes not only generalizations of Fisher's test as mentioned previously, but also the use of nonparametric regression methods to test the fit of a parametric model. Hart [14] provides a thorough review of the development and use of smoothing methodology to construct such lack of fit tests, with a focus on the $p = 1$ case. More recently, see also Eubank et al. [8]. Nonparametric regression techniques have also been employed for the multivariate predictor case (e.g. Staniswalis and Severini [23], Härdle and Mammen [13], Zheng [28], Dette [7], Fan et al. [10], Koul and Ni [17], Guerre and Lavergne [12], Song and Du [22]). However, even for smaller values of $p > 1$, the use of smoothing methods is problematic due to the curse of dimensionality. In many of these references, (y_i, x_i), $1 \le i \le n$, are considered to be independent and identically distributed $\mathbb{R}^{p+1}$ random vectors from a population, and interest involves testing $\mathrm{E}(y|X = x) = m(x)$ where $m(x)$ is a specified parametric regression function. In this setting, Lavergne and Patilea [18] presented a test for $m(x)$ with many regressors involving estimation of conditional expectations given a linear index for a class of single-index models. Combining the expectations as a single numerically estimated integral provides a test against nonparametric alternatives, which reduces the dimension of the problem yet preserves consistency.

For another approach, Khmaladze and Koul [15] presented regression model adequacy tests based on innovation martingale transforms. These tests are asymptotically distribution free for fitting a parametric model to the regression function i.e. the asymptotic null distribution is free of the specified parametric model and the error distribution but depends on the design distribution when $p > 1$. Christensen and Lin [5] also presented tests based on partial sum processes of the residuals and determined the asymptotic null distributions of the maximized partial sums in order to check for lack of fit. This work involves modifications of the test proposed by Su and Wei [27], whose test is based on a partial ordering of the residuals and the asymptotic null distribution is approximated by simulation. As the power of tests based on partial sums of residuals can be greatly influenced by the ordering chosen, Christensen and Lin [5] suggested a total ordering of the data based on a modified Mahalanobis distance and empirically demonstrated their tests were effective for certain types of model inadequacy involving multivariate predictors. Following work by Stute [24] and Stute et al. [26], Stute et al. [25] also presented regression model adequacy tests based on empirical processes of the regressors marked by the residuals, and used a wild bootstrap approximation for the process distribution.

In addition to the preceding work, Aerts et al. [1] and Fan and Huang [9] considered lack of fit tests for multiple regression where the $\{x_i\}$ are considered to be fixed. To circumvent the curse of dimensionality, these authors placed restrictions on the alternative models. For example, Aerts et al. developed tests based on functions of score statistics but require specification of a path in the additive alternative models space, which quickly becomes complex with increasing predictor dimensionality. Fan and Huang constructed tests based on the adaptive Neyman test for the multivariate case but the ability of these tests to detect various types of model inadequacy depends on the ordering of residuals, which can be challenging in higher dimensions. Christensen and Sun [6] followed Fan and Huang with Fourier transforms in the multivariate linear model context, and suggested modifications to the normalizing constants for improved small sample size while maintaining the same asymptotic distributions as Fan and Huang. As noted by Christensen and Sun, their tests also depend on the ordering of the observations to ensure that large Fourier coefficients are concentrated on the lower frequencies. As in these works, we construct lack of fit tests based on fixed predictors in $\mathbb{R}^p$. Unlike these works, we consider the case of moderate to large