Computational Statistics and Data Analysis 54 (2010) 2284–2295

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Order selection tests with multiply imputed data

Fabrizio Consentino, Gerda Claeskens *

ORSTAT and Leuven Statistics Research Center, K.U. Leuven, 3000 Leuven, Belgium

ARTICLE INFO

Article history:
Received 2 April 2009
Received in revised form 8 April 2010
Accepted 8 April 2010
Available online 21 April 2010

Keywords:
Akaike information criterion
Hypothesis test
Multiple imputation
Lack-of-fit test
Missing data
Omnibus test
Order selection

ABSTRACT

Nonparametric tests for the null hypothesis that a function has a prescribed form are developed and applied to data sets with missing observations. Omnibus nonparametric tests such as the order selection tests, do not need to specify a particular alternative parametric form, and have power against a large range of alternatives. More specifically, likelihood-based order selection tests are defined that can be used for multiply imputed data when the data are missing-at-random. A simulation study and data analysis illustrate the performance of the tests. In addition, an Akaike information criterion for model selection is presented that can be used with multiply imputed datasets.

1. Introduction

Testing the lack-of-fit of a parametric function is well-studied. Several types of tests exist, ranging from fully parametric tests, to semiparametric and nonparametric omnibus tests. For an overview of nonparametric tests, see Hart (1997). In settings with missing data, testing for lack-of-fit is more complicated. González-Manteiga and Pérez-González (2006) developed a test based on local linear estimators for a linear regression model with missing response values but a completely observed covariate. We address in particular lack-of-fit tests for missing data situations where multiple imputation is applied. We will focus on a class of smoothing-based tests that use the idea of order selection. Our tests are applicable in parametric likelihood models and are not restricted to linear models.

Eubank and Hart (1992) introduced the order selection test in linear regression models. The idea is to test the shape of a parametric function, most often the mean of the response, by considering not a single one, but a sequence of alternative models. These alternative models are constructed by means of a series expansion of the function of interest around the hypothesized null model. A data-driven method is then applied to select the "order" of the alternative model. That is, in the sequence of alternative models, a method such as Akaike's information criterion AIC (Akaike, 1973) will select the most appropriate one. If the selected model coincides with the null model, the test does not reject the null hypothesis. However, if a model different from the null model is selected, the test will reject the null hypothesis. In those instances, the order of the chosen model, that is, the number of parameters in the model, exceeds that of the null model. Written in another way, this test statistic takes the form of a maximum of weighted likelihood ratio statistics, which clearly indicates the omnibus, or nonparametric nature of the test.

By using such a series expansion the class of alternative models is large and is *not* restricted to a single specified alternative. For example, if we would just test a linear versus a quadratic fit we would miss out on high frequency alternatives for which the quadratic term happens to be zero. Instead, we are interested in the development of tests that are sensitive to essentially any departure from the null hypothesis.

* Corresponding author. Tel.: +32 16 326993; fax: +32 16 326624.
E-mail address: Gerda.Claeskens@econ.kuleuven.be (G. Claeskens).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.04.009

The original order selection tests are extended towards testing in general likelihood models by Aerts et al. (1999) and to multiple regression models by Aerts et al. (2000). Recently, these tests have been studied for inverse regression problems by Bissantz et al. (2009). Test statistics can be based on likelihood ratio, Wald statistics, or score statistics. All of these methods are based on completely observed data.

In practice, many data sets contain one or more missing observations. We refer to Little and Rubin (2002) for an overview of methods to deal with such data. Throughout the paper we make the assumption that the data are missing-at-random; this means that the missingness depends only on the observed data. Most research focuses on the estimation under missingness. Multiple imputation methods are particularly attractive since once values are imputed, traditional, complete case methods can be applied to filled-in data sets. Single imputation, where unknown observations are each replaced by a single value, risks understating uncertainty. Inference is generally improved by imputing values several, say m times, creating m complete data sets, with $m = 5$ a typical choice based on coverage properties seen in simulation studies. Li et al. (1991a) considered hypothesis testing in this setting. In particular, for a parametric null hypothesis of the form $\theta = \theta_0$, with an alternative of the form $\theta \neq \theta_0$, they construct a Wald test by combining the results of m Wald tests, one for each of the m imputed data sets. They show that the distribution of such a test statistic can be approximated by that of an F-distribution with degrees of freedom that depend on the fraction of missing information. Meng and Rubin (1992) extend this idea to combining m likelihood ratio tests. Recently, Reiter (2007) obtained an alternative approximation to the degrees of freedom for such combined Wald test statistics that should work better for small samples.

The main idea of this paper is to use the combined likelihood ratio tests for the m imputed data sets to perform order selection. In this way, we enlarge the testing power by not considering a single parametric test, since order selection tests are constructed to be powerful against a wide range of alternative models. This creates a straightforward to use lack-of-fit test in the setting of missing data.

Section 2 defines the order selection test first for complete data, and then proposes the new test for the case of multiply imputed data sets. Sections 3 and 4 apply the test to a data example and in a simulation study. A version of Akaike's information criterion that works with multiply imputed datasets is obtained in Section 5. Section 6 presents some extensions of the proposed method.

2. The order selection test

2.1. A model sequence for order selection

We consider a set of data $\boldsymbol{Z}_i = (Y_i, x_i)$, $i = 1, \ldots, n$ with joint density depending on a function $\gamma(\cdot)$ of interest (most often this is the mean response, conditional on covariates) and on some other nuisance parameters $\boldsymbol{\eta}$ (such as an unknown variance). We wish to test the hypothesis

$$H_0 : \gamma(\cdot) \in \mathcal{G} = \{\gamma(\cdot, \boldsymbol{\beta}_p) : \boldsymbol{\beta}_p = (\beta_0, \ldots, \beta_p) \in \Theta\}, \tag{1}$$

where the parameter space $\Theta \subset \mathbb{R}^{p+1}$. A simple example is to test for linearity of the mean response, that is, $E(Y|x) = \gamma(x, \boldsymbol{\beta}_1) = \beta_0 + \beta_1 x$. In a parametric hypothesis testing procedure, a specific parametric model would be stated for the alternative hypothesis. In nonparametric or omnibus testing, this is avoided by constructing a sequence of alternative models. These alternatives could be quite general. For regression models, following the approach of Aerts et al. (1999), we focus on additive series expansions of the true underlying function $\gamma(\cdot)$ around the null model. For convenience, we use $r = 0$ to index the null model in (1), and we define for $r = 0, 1, 2, \ldots,$

$$\gamma(x; \beta_0, \ldots, \beta_{p+r}) = \gamma(x; \beta_0, \ldots, \beta_p) + \sum_{j=1}^{r} \beta_{p+j} \psi_j(x), \tag{2}$$

where the basis functions $\psi_j(\cdot)$ are known functions. Most often these functions are taken to be (orthogonalized) polynomials, Legendre polynomials, cosine functions or wavelet functions. For all further analysis, we consider functions ψ_j that are not already used in the null model. For example, a polynomial expansion to test whether the mean $E(Y|X = x) = \beta_0 + \beta_1 x$, will take for $\psi_1(x)$ an (orthogonalized) quadratic function, for $\psi_2(x)$ a cubic function, etc. The reason for starting with a quadratic function is that the constant and linear function are already included in the null model.

In practice it is not possible to include an infinite number of terms in the series expansion in (2). The series will be truncated at a value R_n that might depend on the size of the dataset, in particular it always holds that R_n should not exceed n. The order selection test actively uses a model selection criterion to perform the test. For each $r = 0, 1, 2, \ldots, R_n$ a model with function $\gamma(\cdot; \beta_0, \ldots, \beta_{p+r})$ is fit to the data. This results in a sequence of $R_n + 1$ fitted models. A model selection criterion such as the AIC (Akaike, 1973) is applied to select one of these models. If a model different from the null model is selected, in other words, when the selected order $\hat{r} > 0$, then the null hypothesis (1) is rejected. When the selected order $\hat{r} = 0$, the null model cannot be rejected.

Asymptotic distribution theory was developed by Eubank and Hart (1992) for linear regression models. Aerts et al. (1999, 2000) extended this to likelihood-based regression models, and related the order selection test statistic to a test statistic that is the supremum of a set of weighted likelihood ratio statistics. Specifically, the null hypothesis (1) is rejected when an AIC-type criterion of the form

$$\text{aic}(r, C_n) = 2\{\log L(\hat{\boldsymbol{\eta}}, \hat{\beta}_0, \ldots, \hat{\beta}_{p+r}) - \log L(\hat{\boldsymbol{\eta}}, \hat{\beta}_0, \ldots, \hat{\beta}_p)\} - C_n r, \quad r = 0, 1, 2, \ldots, R_n,$$