Computational Statistics and Data Analysis 53 (2009) 2119–2128

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

New formulations for recursive residuals as a diagnostic tool in the fixed-effects linear model with design matrices of arbitrary rank

J.D. Godolphin

Department of Mathematics, University of Surrey, Guildford, Surrey, GU2 7XH, UK

ARTICLE INFO

Article history:
Available online 26 September 2008

ABSTRACT

The use of residuals for detecting departures from the assumptions of the linear model with full-rank covariance, whether the design matrix is full rank or not, has long been recognized as an important diagnostic tool. Once it became feasible to compute different kinds of residual in a straight forward way, various methods have focused on their underlying properties and their effectiveness. The recursive residuals are attractive in Econometric applications where there is a natural ordering among the observations through time. New formulations for the recursive residuals for models having uncorrelated errors with equal variances are given in terms of the observation vector or the usual least-squares residuals, which do not require the computation of least-squares parameter estimates and for which the transformation matrices are expressed wholly in terms of the rows of the Theil Z matrix. Illustrations of these new formulations are given.

1. Introduction

The use of diagnostic procedures for detecting departures from model assumptions has long been an important part of model building in Econometrics and other sciences where linear models with fixed parameters are employed. The central role played by residuals in these procedures is well known. Much attention has focused on the ordinary least squares (OLS) residuals for the routine testing of model fit and model assumptions. However, the recognition that the OLS residuals are not most suitable for many model-testing problems has led to the proposal of several other definitions of residual in the literature. Haslett and Haslett (2007) have clarified residual specification for the linear model by introducing three essentially different types of residual that categorize other definitions, and by establishing the algebraic relationships between these types. The residual specifications of Haslett and Haslett (2007) apply to the general linear model, in which the design matrix need not be full rank and the error covariance matrix need not be diagonal or even full rank so that, for example, models with fixed and random effects apply.

The OLS residuals are of marginal type, being defined as the difference between actual and fitted observations. Haslett and Haslett (2007) suggest that residuals of the conditional type, which are expressed as the difference between actual and predicted observations, are important and arguably more fundamental than marginal residuals. A useful member of the conditional type is the class of recursive residuals, see Hedayat and Robson (1970) and Brown et al. (1975). Recursive residuals differ from full-conditional residuals, where the predicted observations are based on all other data, in that they are conditional only on a subset of the data which is usually interpreted as "past history". They are therefore intended primarily for assessing linear models where there is a natural ordering among the n observations, usually measurements made over time although other sequential orderings are possible; see, in particular, Hawkins (1991) for comment on this point. It is argued that this form of conditioning is useful for detecting departures from model assumptions at various points in the ordered sequence. For instance, Brown et al. (1975) reason that recursive residuals derived over time behave exactly as on

E-mail address: j.godolphin@surrey.ac.uk.

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.09.019

the null hypothesis until a change in the model takes place when signs of this change should become apparent, unlike other methods where "one would expect the effects of the change to be spread over the full set of transformed residuals".

Recursive residuals have received much attention in the literature. Hedayat and Robson (1970) considered a test that the disturbance variance depends on the regressors, whilst Brown et al. (1975) employed cusum or cusum squared statistics to test whether the parameters of a regression relation change over a period of time, and both tests were based on recursive residuals. Belsley et al. (1980) recommended the recursive residuals in preference to OLS residuals to investigate the presence of outliers even if the model does not change over time. Galpin and Hawkins (1984) proposed on the basis of their practical experience that recursive residuals should be calculated routinely for data sets in the full-rank case as a basis for diagnostic checking. Kianifard and Swallow (1996) gave a comprehensive review of recursive residuals and discussed in detail their use for studying influence and leverage. Further properties of recursive residuals are given by Sen (1982), Haslett (1985), Hawkins (1991) and Wright (1999). Loynes (1986) discussed the case where the matrix of explanatory variables is less than full rank; he proposed that the model should be reparameterised (see, for example, Zyskind (1967)) before computing recursive residuals and then the results from the full-rank case will apply. The less than full-rank case was also discussed by Clarke and Godolphin (1992), Jammalamadaka and Sengupta (1999) and Clarke (2008, chapter 10).

In this paper, attention is confined to the classical case of the linear model with fixed parameters where the model disturbances are mutually uncorrelated and possess a constant variance under the null hypothesis. The principal feature of the recursive residuals is that they are also uncorrelated with the same constant variance as the model disturbances. A complete set of n residuals with this property is not possible since the estimation of the model parameters generates a rank deficiency of some positive integer r so that only $n - r$ degrees of freedom are associated with the residual set. For definiteness, it is customary for recursive residuals to be described so that it is the final $n-r$ residuals which are uncorrelated with constant variance and are in a one-to-one correspondence with the matching observations. Without a convention of this sort, the recursive residuals are not uniquely defined. It is possible to formulate exact tests based on the recursive residuals for structural and related changes to the latter, i.e. more recent, part of the series, although the absence of r residuals at the beginning of the data set is seen to be a disadvantage by some authors. Other unfavourable features of recursive residuals, by comparison with the well known BLUS residuals of Theil (1965), are discussed by Magnus and Sinha (2005).

In the simple linear regression model with an unknown intercept but without a slope, the observation vector $Y = [y_1\ y_2\ \ldots\ y_n]'$ consists of n uncorrelated random variables with a common unknown mean and variance; then the matrix of explanatory variables is 1_n, a vector of n elements equal to unity, and the $n - 1$ recursive residuals $u_1, u_2, \ldots, u_{n-1}$ are specified in terms of Y by the standard Helmert transformation, i.e.

$$u_1 = \frac{y_2 - y_1}{\sqrt{2}}, \qquad u_i = \frac{\mathrm{i}y_{i+1} - (y_1 + \ldots + y_i)}{\sqrt{\mathrm{i}(\mathrm{i}+1)}} \quad (i = 2, \ldots, n-1),$$

see also Kendall and Stuart (1963, p. 250) or Cox (1975). This convenient formulation was exploited by Cox (1975) who obtained general formulae for the asymptotic relative efficiency of the cusum test advocated by Brown et al. (1975) in circumstances where there is a single possible change point, in order to confirm useful properties of the test. However, Cox's hypotheses do not seem to have been investigated in detail for cases other than this simplest of the linear models. In more general models, recursive residuals are obtained sequentially by the indirect approach of deriving the least squares estimators of the unknown parameter vector from sub-vectors of Y and then computing the ordinary residuals in each case. Some assistance in this task is provided by the Plackett formula for updating the normal equations (see Brown et al. (1975)) or Clarke (2008, Lemma 10.2); and, although Nelder (1975) and Hawkins (1991) raise doubt about the numerical stability of these computations, it seems unlikely that this is a major problem with existing software. A difficulty may be apparent if the matrix of explanatory variables is less than full rank since there seem to be no guidelines in the literature for suggesting a preliminary reparameterisation of the model before applying this procedure. The fact remains that this indirect approach of describing recursive residuals lacks the pedagogic feature of expressing the linear structure of these residuals in terms of Y for any models, whether full rank or not, which are more general than that considered by Cox (1975).

The purpose of the present paper is to derive two formulations for recursive residuals which are expressed directly in terms of the observation vector Y or, alternatively, which are expressed solely in terms of a component of the OLS residuals. Neither representation requires the computation of any least squares parameter estimates and all coefficients are given in closed form, where no matrix inversion is required in their derivation. These results apply whether the design matrix of explanatory variables has full rank or not, and then r is just the rank of this design matrix. The results therefore possess simple computational properties which parallel the approach of Foschi et al. (2003), Foschi and Kontoghiorghes (2003) and Kontoghiorghes (2004), who gave methods for deriving a set of (ordinary) residuals for seemingly unrelated regressions which do not require any matrix inversion. Furthermore, the two residual transformations presented here possess the interesting feature of being expressed wholly in terms of the rows of the Z matrix of Theil (1965), which simplifies considerably the recursive residual formulations.

Section 2 of the paper presents and derives two lemmas which provide invariant results for the linear model. Section 3 and 4 derive the two recursive residual representations and two illustrations of these formulations are given. Section 5 specifies two algorithms for computing the recursive residuals which exhibit straightforward computational properties, and a study of the computational complexity of these methods is presented.