Available online at www.sciencedirect.com
SCIENCE DIRECT®
Statistics & Probability Letters 73 (2005) 425–432
STATISTICS & PROBABILITY LETTERS
www.elsevier.com/locate/stapro

Comparing distribution functions of errors in linear models: A nonparametric approach ☆

Juan Mora

Departamento de Fundamentos del Análisis Económico, Universidad de Alicante, Apartado de Correos 99, E-03080 Alicante, Spain

Received 17 December 2003; received in revised form 15 February 2005
Available online 11 May 2005

Abstract

We describe how to test whether the distribution functions of errors from two linear regression models are the same, with statistics based on empirical distribution functions constructed with residuals. A smooth bootstrap method is used to approximate critical values. Simulations show that the procedure works well in practice.

Keywords: Two-sample problem; Residual-based empirical process; Smooth bootstrap

1. Introduction

Consider two linear regression models

$$Y_{ji} = X'_{ji}\beta_j + \sigma_j U_{ji}, \quad j = 1, 2, \ i = 1, \ldots, n_j, \tag{1}$$

where $\{Y_{ji}\}_{i=1}^{n_j}$ are observations, $\{X_{ji}\}_{i=1}^{n_j}$ are known design vectors in $\mathbb{R}^{k_j}$, $\beta_j \in \mathbb{R}^{k_j}$ and $\sigma_j \in (0, +\infty)$ are unknown parameters and the errors $\{U_{ji}\}_{i=1}^{n_j}$ are independent and identically distributed (i.i.d.) random variables such that $E(U_{ji}) = 0$, $E(U_{ji}^2) = 1$ (hereafter $j = 1, 2$). Assuming independence between the two samples, the objective of this article is to propose a

☆This research has been partially supported by Spanish DGI (ref. BEC2002-03097).
E-mail address: juan@merlin.fae.ua.es.

0167-7152/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2005.04.017

test procedure to face the hypotheses

$$\mathrm{H}_0: \quad F_1(\cdot) = F_2(\cdot) \quad \text{versus} \quad \mathrm{H}_1: \quad F_1(\cdot) \neq F_2(\cdot),$$

where $F_j(\cdot)$ is the distribution function of U_{ji}, assumed to be continuous. Observe that H_0 states that the distribution functions of Y_{1i} and Y_{2i} are the same except for differences in the mean, which may depend on regressors, and the variance. Testing whether Y_{1i} and Y_{2i} belong to the same location-scale family is a specific case of (1), taking $X_{ji} = 1$. However, testing whether the distribution functions of Y_{1i} and Y_{2i} are the same except for differences only in the mean is not a specific case of (1), but can be dealt with using the same method that we describe here.

The problem that we study arises in many contexts in applied work. In Economics, for example, the productivity of a firm is defined as the error from a regression model, and the researcher is often interested in comparing the distribution functions of productivity of firms from two different groups. In applied medical studies, the researcher is sometimes interested in comparing the distribution functions of certain standardized variables with data from healthy and unhealthy individuals. In these cases, the usual approach to testing for the equality of the distribution functions is to test for the equality of just some moments or, with a parametric approach, to propose parametric models for the errors and then test whether the parameters estimated are equal. Instead, we propose to compare the distribution functions without assuming any parametric form for them.

If errors were observable, we would have to test whether two observable variables come from the same distribution; thus we could use a Kolmogorov–Smirnov statistic $K_{n_1,n_2} := [n_1 n_2/(n_1+n_2)]^{1/2} \sup_{z\in\mathbb{R}} |F_{1n_1}(z) - F_{2n_2}(z)|$, where $F_{jn_j}(\cdot)$ denotes the empirical distribution function based on $\{U_{ji}\}_{i=1}^{n_j}$, or a Cramér–von Mises statistic $C_{n_1,n_2} := [n_1 n_2/(n_1+n_2)^2][\sum_{j=1}^{2}\sum_{i=1}^{n_j}\{F_{1n_1}(U_{ji}) - F_{2n_2}(U_{ji})\}^2]$. If H_0 is true and $F_1(\cdot)$ is continuous, these statistics are distribution free and their asymptotic behavior is known. So they can be used to perform a consistent test (see e.g. Shorack and Wellner (1986), Section 9.9). In our context, we do not observe errors U_{ji}, but given estimates $\widehat{\beta}_j$ and $\widehat{\sigma}_j$ we can construct residuals $\widehat{U}_{ji} = (Y_{ji} - X'_{ji}\widehat{\beta}_j)/\widehat{\sigma}_j$, for $i = 1, \ldots, n_j$, and the residual-based statistics

$$\widehat{K}_{n_1,n_2} := \left[\frac{n_1 n_2}{(n_1+n_2)}\right]^{1/2} \sup_{z\in\mathbb{R}} |\widehat{F}_{1n_1}(z) - \widehat{F}_{2n_2}(z)|,$$

$$\widehat{C}_{n_1,n_2} := \left[\frac{n_1 n_2}{(n_1+n_2)^2}\right] \sum_{j=1}^{2}\sum_{i=1}^{n_j} \{\widehat{F}_{1n_1}(\widehat{U}_{ji}) - \widehat{F}_{2n_2}(\widehat{U}_{ji})\}^2,$$

where $\widehat{F}_{jn_j}(\cdot)$ denotes the empirical distribution function based on $\{\widehat{U}_{ji}\}_{i=1}^{n_j}$. These are the statistics whose properties are studied here.

In a one-sample context, the properties of nonparametric residual-based statistics similar to these are well known. Pierce and Kopecky (1979) and Loynes (1980) prove that the replacement of errors by residuals provokes changes in the asymptotic distribution of goodness-of-fit statistics based on the empirical distribution function; their results were generalized by Bai (1994) and Koul (1996), among others. Koul (1996) also proposes the use of a two-sample statistic similar to $\widehat{K}_{n_1,n_2}$ to the test for the equality of the distribution functions of regression errors before and after a known change point, and proves that his statistic is asymptotically distribution free; but there is a crucial difference between our problem and the change point problem: in our context, it is natural