Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 51 (2007) 6269 – 6284

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

A study of partial F tests for multiple linear regression models

Mortaza Jamshidian[a, *, 1], Robert I. Jennrich[b], Wei Liu[c]

[a] *Department of Mathematics, California State University, Fullerton, CA 92834, USA*
[b] *Department of Mathematics, University of California, Los Angeles, CA, USA*
[c] *Statistical Sciences Research Institute and School of Maths, The University, Southampton, SO17 1BJ, UK*

Received 19 May 2005; received in revised form 12 January 2007; accepted 13 January 2007
Available online 2 February 2007

Abstract

Partial F tests play a central role in model selections in multiple linear regression models. This paper studies the partial F tests from the view point of simultaneous confidence bands. It first shows that there is a simultaneous confidence band associated naturally with a partial F test. This confidence band provides more information than the partial F test and the partial F test can be regarded as a side product of the confidence band. This view point of confidence bands also leads to insights of the major weakness of the partial F tests, that is, a partial F test requires implicitly that the linear regression model holds over the entire range of the covariates in concern. Improved tests are proposed and they are induced by simultaneous confidence bands over restricted regions of the covariates. Power comparisons between the partial F tests and the new tests have been carried out to assess when the new tests are more or less powerful than the partial F tests. Computer programmes have been developed for easy implements of these new confidence band based inferential methods. An illustrative example is provided.

Keywords: Confidence bands; Linear regression; Simultaneous inference; Statistical simulation

1. Introduction

Consider a standard multiple linear regression model given by

$$\mathbf{Y} = X\boldsymbol{\beta} + \mathbf{e}, \tag{1.1}$$

where $\mathbf{Y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ is a vector of observations, X is an $n \times (p+1)$ full column-rank design matrix with the first column given by $(1, \ldots, 1)^{\mathrm{T}}$ and the lth $(2 \leqslant l \leqslant p+1)$ column given by $(x_{1,l-1}, \ldots, x_{n,l-1})^{\mathrm{T}}$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^{\mathrm{T}}$ is a vector of unknown coefficients, and $\mathbf{e} = (e_1, \ldots, e_n)^{\mathrm{T}}$ is a vector of independent random errors with each $e_i \sim N(0, \sigma^2)$, where σ^2 is an unknown parameter.

One important problem for model (1.1) is to assess whether some of the coefficients β_i's are zero and so the corresponding covariates x_i's have no effect on the response variable Y. The model can then be simplified. To be specific, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}_1^{\mathrm{T}} = (\beta_0, \ldots, \beta_{p-k})$ and $\boldsymbol{\beta}_2^{\mathrm{T}} = (\beta_{p-k+1}, \ldots, \beta_p)$ with $1 \leqslant k \leqslant p$. If $\boldsymbol{\beta}_2$ is zero then

[*] Corresponding author.
E-mail address: mori@exchange.fullerton.edu (M. Jamshidian).

[1] The research has been supported in part by the National Science Foundation Grant DMS-0437258.

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.01.015

the covariates $x_{p-k+1}, \ldots, x_p$ have no effect on the response variable Y and model (1.1) reduces to

$$\mathbf{Y} = X_1\boldsymbol{\beta}_1 + \mathbf{e}, \tag{1.2}$$

where X_1 is formed by the first $p-k+1$ columns of the matrix X.

A commonly used statistical approach to assessing whether $\boldsymbol{\beta}_2$ is zero is to test the hypotheses

$$\mathrm{H}_0\colon \boldsymbol{\beta}_2 = \mathbf{0} \quad \text{against} \quad \mathrm{H}_a\colon \boldsymbol{\beta}_2 \neq \mathbf{0} \tag{1.3}$$

by using the partial F test, which rejects H_0 if and only if

$$\frac{[\text{Regression SS of model (1.1)} - \text{Regression SS of model (1.2)}]/k}{\text{MS residual of model (1.1)}} > f_{k,v}^{\alpha},$$

where $f_{k,v}^{\alpha}$ is the upper α point of an F distribution with k and $v = n-(p+1)$ degrees of freedom. This can be found in most text books on multiple linear regression models; see, for example, Kleinbaum et al. (1998).

The inferences that can be drawn from this partial F test are that if H_0 is rejected then $\boldsymbol{\beta}_2$ is deemed to be non-zero and so at least some of the covariates $x_{p-k+1}, \ldots, x_p$ affect the response variable Y, and that if H_0 is not rejected then there is not enough statistical evidence to conclude that $\boldsymbol{\beta}_2$ is not equal to zero. (Unfortunately, this latter case is often misinterpreted as $\boldsymbol{\beta}_2$ is equal to zero and so model (1.2) is accepted as more appropriate than model (1.1).) Whether H_0 is rejected or not, no information on the magnitude of $\boldsymbol{\beta}_2$ is provided directly by this approach of hypotheses testing.

The first purpose of this paper is to show that there is a simultaneous confidence band associated naturally with the partial F test and the partial F test can be interpreted more intuitively via this simultaneous confidence band. Hypotheses (1.3) can be tested by using this confidence band: the acceptance or rejection of H_0 is according to whether or not the zero hyper-plane lies completely inside the confidence band; by zero-hyperplane we mean the graph in $\mathscr{R}^{k+1}$ of the zero valued function on $\mathscr{R}^k$. The advantage of this confidence band approach over the partial F test is that it provides information on the magnitude of $\beta_{p-k+1}x_{p-k+1} + \cdots + \beta_p x_p$, whether or not H_0 is rejected. This is discussed in Section 3. However, this confidence band is over the entire range $(-\infty, \infty)$ of each of the covariates $x_{p-k+1}, \ldots, x_p$. As a linear regression model is an acceptable approximation often only over a restricted region of these covariates, the part of the confidence band outside this restricted region is useless for inference. It is therefore unnecessary to guarantee the $1-\alpha$ simultaneous coverage probability over the entire range of each of these covariates. Furthermore, inferences deduced from the part of the confidence band outside the restricted region, such as the rejection of H_0, may not be valid since the assumed model may be wrong outside the restricted region. This calls for the construction of a $1-\alpha$ simultaneous confidence band only over this restricted region of the covariates. This confidence band is narrower and so allows more precise inferences over the restricted region than the confidence band associated with the partial F test. These results are illuminated in Sections 4 and 5. Section 6 compares the powers of the partial F test and the new test induced from the confidence band over a restricted region considered in Section 4. Some concluding remarks are contained in Section 7. But we first provide some preliminaries in Section 2.

2. Some preliminaries

The main result of this section is a simple algebraic result which we will use in Sections 3 and 4; it is related to the derivation of a simultaneous confidence band over the whole range of the covariates for one linear regression model (see, Miller, 1981, Chapter 2, Section 2). Denote the estimate of $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathbf{Y} \sim \mathbf{N}(\boldsymbol{\beta}, \sigma^2(X^{\mathrm{T}}X)^{-1})$, and the estimate of σ^2 by $\hat{\sigma}^2 = \|\mathbf{Y} - X\hat{\boldsymbol{\beta}}\|^2/v \sim \sigma^2\chi_v^2/v$ where $v = n-p-1$. It is well known that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent, and $\hat{\sigma}^2$ is the *MS residual* of model (1.1). Let $\hat{\boldsymbol{\beta}}_2$ denote the estimate of $\boldsymbol{\beta}_2$. The quantity $\hat{\boldsymbol{\beta}}_2$ has the distribution $\mathbf{N}(\boldsymbol{\beta}_2, \sigma^2 A)$, where A is the $k \times k$ partition matrix from the last k rows and the last k columns of $(X^{\mathrm{T}}X)^{-1}$. Since A is symmetric and positive definite, there exists a unique symmetric positive definite matrix W such that $A = W^2$. Denote the ith column of W by $\mathbf{w}_i$ for $i = 1, \ldots, k$. Finally let $\mathbf{x}_2 = (x_{p-k+1}, \ldots, x_p)^{\mathrm{T}}$ denote the covariates corresponding to $\boldsymbol{\beta}_2$.

Lemma. *For a hyper-rectangle region C given by*

$$C = \{\mathbf{x}_2\colon a_i < x_i < b_i,\ i = p-k+1, \ldots, p\},$$