Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 5958–5976
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

On testing a subset of regression parameters under heteroskedasticity

Miin-Jye Wen[a], Shun-Yi Chen[b], Hubert J. Chen[c,*,1]

[a]*Department of Statistics, National Cheng Kung University, Tainan 701, Taiwan*
[b]*Department of Mathematics, Tamkang University, Tamsui 251, Taiwan*
[c]*Department of Statistics and Accountancy, National Cheng Kung University, Tainan 701, Taiwan*

Received 16 March 2006; received in revised form 27 September 2006; accepted 11 November 2006
Available online 6 December 2006

Abstract

Assuming a general linear model with unknown and possibly unequal normal error variances, the interest is to develop a one-sample procedure to handle the hypothesis testing on all, partial, or a subset of linear functions of regression parameters. The sampling procedure is to split up each single sample of size n_i at a controllable regressor's data point into two portions, the first consisting of the $n_i - 1$ observations for initial estimation and the second consisting of the remaining one for overall use in the final estimation in order to define a weighted sample mean based on all sample observations at each data point. Then, the weighted sample mean is used to serve as a basis for parameter estimates and test statistics for a general linear regression model. It is found that the distributions of the test statistics based on the weighted sample means are completely independent of the unknown variances. This method can be applied to analysis of variance under various designs of experiments with unequal variances.

Keywords: Regression analysis; Heteroskedasticity; One-sample procedure; Point estimation; Hypothesis testing; Partial test; Analysis of variance

1. Introduction

Assuming a general linear model with unknown and possibly unequal normal error variances, the interest is to develop a one-sample procedure to handle the statistical problems involving point estimation and hypothesis testing on all, partial, or a subset of linear functions of regression parameters. The sampling procedure is to split up each single sample of size n_i at a controllable regressor's vector data point $\boldsymbol{X}_i$ into two portions, the first consisting of the first (or randomly) $n_i - 1$ observations for initial estimation and the second consisting of the remaining one for the overall use in the final estimation. A linear combination of all sample observations at each data point is formulated, where its coefficients depend on the first portion's sample information. As a result, the proposed linear combination is a weighted unbiased sample mean for the expected regression mean response at the given vector data point. The linear combination so formulated will be used to serve as a basis for parameter estimates and test statistics for testing the null hypotheses about the parameters under a general linear regression model with unknown and unequal variances. The advantage of this proposed sampling procedure is that the distributions of the tests under the null hypotheses are exact

* Corresponding author. Tel.: +886 6 2757575x53641; fax: +886 6 2342469.
E-mail address: jpchen@stat.ncku.edu.tw (H.J. Chen).
[1] Professor Emeritus, University of Georgia, 2003.

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.11.018

and completely independent of the unknown variances, or are parameter free. Consequently, the p-values and critical values for various tests under the null hypotheses can be simulated by a computer program. To be more specific, this method can be applied to regression analysis including estimation, overall test, test for a subset of parameters, and a partial test for model selection whenever the quantitative or qualitative variables or their combination are used as predictors. In addition, the proposed inferential technique can also be applied to analysis of variance models for one-way layout, two-way layout with or without interactions, and three or higher-way layout models, under a completely randomized design, randomized block design, factorial design, Latin square design and so on. Thus, this procedure shall provide an alternative to practitioners who deal with statistical data analysis whenever they encounter heterogeneous error variances.

By looking into literature in the field of heteroskedasticity problem in the econometrics and regression models, a good number of researchers proposed several potential solutions to the estimation of heteroskedastic error variances. To name a few examples: Horn et al. (1975) proposed a set of almost unbiased estimators for the heteroskedastic error variances in a general linear model, which improved previously studied minimum norm quadratic unbiased estimators and demonstrated computational economy, positive variance estimates and decreased mean square error; no test procedure for testing the true error variances was seen. Later on, White (1980) proposed a heteroskedasticity-consistent covariance matrix estimator (HCCME) for estimating the unknown covariance matrix of the disturbance terms in a general linear model. He substituted his HCCME and the ordinary least-squares estimators of the regression parameters into a quadratic formula to construct a test statistic for testing a linear hypothesis by asymptotic theory. The disadvantage of the HCCME is that it is somewhat unreliable in finite sample and that there was no small sample distribution being given. A good review in this area can be found in Davidson and MacKinnon (1993, Chapter 16) for econometrics study. Recently, Cribari-Neto (2004) further studied the finite-sample behavior of the HCCME mainly based on White's (1980) idea. He was able to propose some improved covariance matrix estimator (called HC4 estimator), and then applied it to the quasi-t tests using Monte Carlo method. However, when performing a test on a linear hypothesis, his sample statistic was still based on limiting distribution or bootstrapping scheme; no exact small sample distribution of his estimator was given. Therefore, in this research, we proposed a one-sample sampling procedure to handle the heterogeneous error variances problem in the general linear regression model: an exact sampling procedure is defined, unbiased point estimators of regression parameters are provided, test statistics for linear hypotheses are proposed. It can be shown that these test statistics based on the proposed statistical procedure have exact small sample distributions which are parameter-free under null hypotheses.

Consider the classical general linear regression model

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_{ij} \\ &= \mathbf{P}_i'\boldsymbol{\beta} + e_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i, \end{aligned} \tag{1}$$

where Y_{ij} is a response variable, $X_{i1}, \ldots, X_{ip}$ are predictor variables and $\mathbf{P}_i' = (1, X_{i1}, \ldots, X_{ip})$ is a known row vector of predictors at ith X-data point, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is an unknown column vector of regression parameters subject to $(p+1) \leqslant k$, and e_{ij}'s are independently distributed as normal $N(0, \sigma_i^2)$ random errors, where the variances σ_i^2's are unknown and possibly unequal. It is also assumed that at each X-data point i, there are n_i responses, $Y_{i1}, \ldots, Y_{in_i}$ observed. In many experimental situations the assumption that the random errors e_{ij}'s have equal variances is not justifiable, and investigations have shown that inequality of the error variances can have a serious effect on the probabilities associated with inferences where the p-value of a test is inflated (e.g., see Bishop and Dudewicz, 1978). To eliminate such deficiencies, Bishop (1978) proposed a Stein-type (1945) two-stage sampling procedure such that the distributions of the test statistics are completely independent of the variances and that the power of the test and the length of an interval can be controlled at a desirable level. But they had no exact distribution, except for large sample approximation. Furthermore, the two-stage procedure is a design-oriented procedure which requires possibly large additional observations at the second stage. As a result, it may not be practicable for the problem of data analysis in practice due to project termination, budget limitation, unavailability of additional samples and other cost factors. There has been little work to handle this type of regression problem since Bishop (1978). Therefore, in the paper, we develop a one-sample procedure without taking additional observations in the future, which yields statistics whose distributions are also completely independent of the unknown variances. The one-sample procedure was originally studied by Chen and Lam (1989) in their interval estimation on the largest mean under heteroscedastic error variances, but was not extended to regression problem because of complexity. One quick application is that: when the two-stage sampling