TESTING FOR VARYING DISPERSION IN EXPONENTIAL FAMILY NONLINEAR MODELS*

BU-CHENG WEI¹, JIAN-QING SHU¹,², WING-KAM FONG³ AND YUE-QING HU¹

¹Department of Mathematics, Southeast University, Nanjing 210096, China  
²Department of Statistics, The Chinese University of Hong Kong, Hong Kong, China  
³Department of Statistics, University of Hong Kong, Pokfulam Road, Hong Kong, China

(Received March 22, 1995; revised March 26, 1997)

Abstract. A diagnostic model and several new diagnostic statistics are proposed for testing for varying dispersion in exponential family nonlinear models. A score statistic and an adjusted score statistic based on Cox and Reid (1987, J. Roy. Statist. Soc. Ser. B, 55, 467–471) are derived in normal, inverse Gaussian, and gamma nonlinear models. An adjusted likelihood ratio statistic is also given for normal and inverse Gaussian nonlinear models. The results of simulation studies are presented, which show that the adjusted tests keep their sizes better and are more powerful than the ordinary tests.

Key words and phrases: Adjusted profile likelihood, exponential family nonlinear models, gamma model, inverse Gaussian model, score statistic, simulation study, varying dispersion.

1. Introduction

Nonconstant variance is a commonly concerned problem in regression analysis. For generalized linear models (GLM) with exponential family distributions, since the nominal variance functions are always nonconstant (except in the normal case), it is not necessary to detect nonconstant variance. However, the variance problem still exists for GLM, which becomes a test for varying dispersion (Smyth 1989), that is a test of departure from the nominal dispersion (Smith and Heitjan 1993) including overdispersion and underdispersion. This is a commonly concerned problem in recent years; see, for example, Cox (1983), Efron (1986), Cox and Snell (1989), McCullagh and Nelder (1989), Breslow (1990), Dean (1992), Gamio and Schafer (1992), Smith and Heitjan (1993), and the references therein. However, most of the above authors were interested in count data and discrete exponential families, in particular, the Poisson and binomial families in which a very common practical complication is the presence of overdispersion (Cox 1983). In this paper, our interests are placed on the following two aspects: (a) We study

* Project supported by NSFC, NSJF and CRCG 344/817/0002 of University of Hong Kong.
exponential family nonlinear models which include GLM and normal nonlinear regression as their special cases. (b) We are concerned more about the continuous exponential family distributions, in particular, the gamma, normal and inverse Gaussian distributions. A diagnostic model to detect varying dispersion is proposed for exponential family nonlinear models and several new diagnostic statistics are derived as well. Section 2 obtains a score statistic, while Section 3 derives the adjusted score statistic and the adjusted likelihood ratio statistic based on Cox and Reid (1987, 1993). In Section 4, we investigate the properties of these new tests through simulation studies. The results show that the adjusted tests keep their sizes better and are more powerful than the unadjusted tests, as expected.

Now we introduce some notations and the background material. Given a set of independent observations \((x_i, y_i), i = 1, \ldots, n\), the exponential family nonlinear models (Cordeiro and Paula (1989), Cook and Tsai (1990)) can be represented as

\[
y_i \sim ED(\mu_i, \sigma^2), \quad g(\mu_i) = f(x_i; \beta), \quad i = 1, \ldots, n,
\]

where \(\mu_i = E(y_i)\), \(g\) is a monotonic link function, \(f\) is a known differentiable function and \(\beta = (\beta_2, \ldots, \beta_p)^T\) are parameters of interest. \(ED(\mu_i, \sigma^2)\), as used by Jorgensen (1987), denotes an exponential family p.d.f. of the form

\[
p(y_i; \theta_i, \tau) = \exp \left[ \tau(y_i \theta_i - b(\theta_i) - c(y_i)) - \frac{1}{2} a(\tau, y_i) \right],
\]

where \(\tau = \sigma^{-2}\) is a constant dispersion parameter for all \(i\), \(\theta_i\) is the natural parameter and \(b\), \(c\), \(a\) are known functions. In some situations (see, for example, Jorgensen (1992)), \(ED(\mu_i, \sigma^2)\) may be replaced by \(ED(\mu_i, \sigma^2 w_i^{-1})\) with known weights \(w_i\). Since \(w_i\)'s are known, we assume \(w_i = 1\) for simplicity while the derivation is similar if \(w_i \neq 1\).

It is easily seen that if \(f(x_i) = x_i^T \beta\), then (1.1) represents a GLM; if \(g(\mu_i) = \mu_i\), then (1.1) represents a nonlinear regression. In particular, if \(g(\mu_i) = \mu_i\) with \(b(\theta_i) = \theta_i^2 / 2\), then (1.1) becomes a normal nonlinear regression. We may rewrite \(g(\mu_i) = f(x_i; \beta)\) as \(\mu_i = \mu(\beta) = g^{-1} \circ f(x_i; \beta)\) or \(\theta_i = \theta(\beta) = (g \circ b)^{-1} \circ f(x_i; \beta)\) because \(\mu_i = b(\theta_i)\), where, here and henceforth, dots over functions denote the derivatives. The deviance corresponding to each \(y_i\) is

\[
d_i(y_i, \mu_i) = d_i^o(y_i, \mu_i) - d_i^o(y_i, \mu_i),
\]

where

\[
d_i^o(y_i, \mu_i) = -2 \{y_i \theta_i - b(\theta_i) - c(y_i)\}.
\]

As discussed by Smyth (1989) and Cordeiro and McCullagh (1991), we assume that the function \(u(\tau, y_i)\) has the form

\[
u(\tau, y_i) = s(\tau) + t(y_i).
\]

It is well known that \(s(\tau) = -\log \tau\) for normal and inverse Gaussian (IG) models; \(s(\tau) = -2(\tau \log \tau - \log \Gamma(\tau))\) for gamma (GA) model, where \(\Gamma(\tau)\) is the gamma