The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Likelihood Ratio, Score, and Wald Tests in a Constrained Parameter Space

Geert Molenberghs & Geert Verbeke

To cite this article: Geert Molenberghs & Geert Verbeke (2007) Likelihood Ratio, Score, and Wald Tests in a Constrained Parameter Space, The American Statistician, 61:1, 22-27, DOI: 10.1198/000313007X171322

To link to this article: https://doi.org/10.1198/000313007X171322

Likelihood Ratio, Score, and Wald Tests in a Constrained Parameter Space

Geert MOLENBERGHS and Geert VERBEKE

Likelihood ratio, score, and Wald tests statistics are asymptotically equivalent. This statement is widely known to hold true under standard conditions. But what if the parameter space is constrained and the null hypothesis lies on the boundary of the parameter space, such as, for example, in variance component testing? Quite a bit is known in such situations too, but knowledge is scattered across the literature and considerably less well known among practitioners. Motivated from simple but generic examples, we show there is quite a market for asymptotic one-sided hypothesis tests, in the scalar as well as in the vector case. Reassuringly, the three standard tests can be used here as well and are asymptotically equivalent, but a somewhat more elaborate version of the score and Wald test statistics is needed. Null distributions take the form of mixtures of χ^2 distributions. Statistical and numerical considerations lead us to formulate pragmatic guidelines as to when to prefer which of the three tests.

KEY WORDS: Boundary condition; Dose response; Generalized linear mixed model; Likelihood ratio test; Linear mixed model; One-sided test; Score test; Variance component; Wald test.

1. INTRODUCTION

Whenever one-sided hypothesis tests are of interest, two forms need to be clearly distinguished. Consider the scalar parameter case. The first, more conventional, unconstrained form compares positive values against negative values, or vice versa. The second, constrained form places a restriction on the parameter space, for example, by comparing positive alternative values versus 0 under the null hypothesis. The latter is common in variance component testing arising in a linear or generalized linear mixed model context, and in monotonic dose-response modeling, as will be illustrated in the following examples.

Geert Molenberghs is Professor of Biostatistics, Center for Statistics, Hasselt University, Diepenbeek, Belgium (E-mail: *geert.molenberghs@uhasselt.be*) and Geert Verbeke is Professor of Biostatistics, Biostatistical Centre, Katholieke Universiteit Leuven, Belgium (E-mail: *geert.verbeke@med.kuleuven.be*). The authors gratefully acknowledge support from Belgian IUAP/PAI network "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data."

1.1 The Rat Data Example

The data resulted from a randomized longitudinal experiment (Verdonck et al. 1998), in which 50 male Wistar rats were randomized to either a control group or one of the two treatment groups, where treatment consisted of a low or high dose of the testosterone inhibitor Decapeptyl. The primary goal was to investigate the drug's effect on craniofacial growth. The treatment started at the age of 45 days, and measurements were taken every 10 days, starting at the age of 50 days. Of interest was skull height, measured as the distance (in pixels) between two well-defined points on X-ray pictures taken under anesthesia. Some rats have incomplete follow-up because they did not survive anesthesia.

Let Y_{ij} denote the response taken at time t_j, for rat i. Verbeke and Molenberghs (2000, 2003) modeled the subject-specific profiles as linear functions of $t = \ln(1 + (\text{Age} - 45)/10)$:

$$Y_{ij} = \begin{cases} \beta_0 + b_{1i} + (\beta_1 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if low dose,} \\ \beta_0 + b_{1i} + (\beta_2 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if high dose,} \\ \beta_0 + b_{1i} + (\beta_3 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if control.} \end{cases} \quad (1)$$

Here, β_0 is the average response at the time of randomization, while β_1, β_2, and β_3 are the average slopes in the three different treatment groups. Further, the b_{1i} and b_{2i} are rat-specific intercepts and slopes, representing natural heterogeneity between rats relative to baseline values and evolution over time, respectively. They are assumed to be zero-mean normally distributed with covariance matrix D. The residual error terms ε_{ij} are independently normally distributed with zero mean and variance σ^2.

The so-called marginal model results form integrating over the random effects. Marginally, the elements Y_{ij} of the vector $\mathbf{Y}_i$ of all measurements on unit i are normally distributed with means $\beta_0 + \beta_k t_{ij}$, where $k = 1, 2, 3$, refers to dose group, with $\text{cov}(\mathbf{Y_i}) = V_i = Z_i D Z_i' + \sigma^2 I_{n_i}$, with Z_i the random-effects design matrix, consisting of a column of ones and a column of observation times t_{ij}, where $j = 1, \ldots, n_i$ and n_i the number of measurements obtained for rat $i = 1, \ldots, N = 50$.

Parameter estimation typically proceeds via (restricted) maximum likelihood. We consider four nested submodels. Denote the elements of D by $d_{k\ell}(k, \ell = 1, 2)$. Model 0 contains no random effects ($D = 0$) and Model 1 features a random intercept only, $D = (d_{11})$. In the remaining two models, both random effects are present, but they are assumed uncorrelated in Model 3, that is, D is 2×2 diagonal, while they are allowed to be correlated in Model 4. We will consider three hypotheses. $H_{0,1} : d_{11} = 0$ compares Models 1 and 2; $H_{0,2} : d_{22} = 0$ compares Models 2

  DOI: 10.1198/000313007X171322

and 3, and $H_{0,3} : d_{12} = d_{22} = 0$ compares Models 2 and 4. Clearly, $H_{0,1}$ and $H_{0,2}$ are scalar hypotheses, while $H_{0,3}$ is of vector type. This is one of the situations where one-sided testing next to two-sided testing can be of interest. When the linear mixed model is used merely to generate a flexible yet parsimoniously parameterized marginal model, it is sufficient that V_i be positive definite, a condition weaker than the one needed for a full random-effects interpretation, where D needs to be positive definite. In the former case, two-sided testing of D-components is sensible, while in the latter case all hypotheses lie on the boundary of the parameter space, pointing towards a one-sided framework.

1.2 The Toenail Data Example

These data were obtained from a randomized, double-blind, parallel group, multicenter study for the comparison of two oral treatments (coded A and B) for toenail dermatophyte onychomycosis (TDO, De Backer et al. 1996). TDO is a common toenail infection, requiring prolonged antifungal therapy (Roberts 1992). The development of new compounds has reduced the treatment duration to three months. In total, 2×189 patients were randomized, measured at baseline, then every month during 12 weeks of treatment and once a quarter up to 48 weeks. We consider only those patients for which the big toenail was the target nail, leading to 146 and 148 subjects, in group A and group B, respectively. Of interest is the outcome "severity of infection" (coded as 0 for not severe or 1 for severe infection), whether it decreased over time, and whether this evolution was different for the two treatment groups.

We consider a generalized linear mixed model:

$$\begin{aligned} Y_{ij}|(b_{0i}, b_{1i}) &\sim \text{Bernoulli}(\pi_{ij}), \\ \text{logit}(\pi_{ij}) &= \beta_0 + b_{0i} + (\beta_1 + b_{1i})t_{ij} + \beta_2 T_i + \beta_3 T_i t_{ij}, \end{aligned} \tag{2}$$

where T_i is the treatment indicator for subject i, t_{ij} is the timepoint at which the jth measurement is taken for the ith subject, and b_{0i} and b_{1i} are the random intercept and random slope in time, respectively, assumed to be normally distributed with mean zero and covariance matrix D. Under the null hypothesis, the random slope is removed from the model, that is, $H_0 : d_{12} = d_{22} = 0$, with notation as in the previous section. Although a generalized linear mixed model like this can be estimated using maximum likelihood, the alternative model is hard to fit in this case, rendering the score test a better option than the likelihood ratio test.

1.3 The NTP Data Example

This developmental toxicity study investigates the dose-response relationship in mice of the potentially hazardous chemical compound ethylene glycol (EG; Price et al. 1985, Aerts et al. 2002). EG has many applications, for example, as antifreeze in cooling systems. In the study, timed-pregnant CD-1 mice were dosed by gavage with EG in distilled water. Dosing occurred during the period of organogenesis and structural development of the foetuses. Doses were 0, 750, 1,500, or 3,000 mg/kg/day. A key outcome is an indicator for whether a fetus is malformed.

The dose-response relationship is modeled using generalized estimating equations (GEE, Liang and Zeger 1986). The marginal model takes the form

$$Y_{ij} \sim \text{Bernoulli}(\pi_{ij}), \qquad \text{logit}(\pi_{ij}) = \beta(d_i), \tag{3}$$

where $\beta(d_i)$ takes values $\beta_0, \beta_1, \beta_2$, or β_3, depending on whether dose d_i administered to litter i takes values 0, 750, 1,500, or 3,000 mg/kg/day, respectively. To complete GEE model specification, a working exchangeable correlation structure is assumed. Here, $Y_{ij} = 1$ if malformation occurs and 0 otherwise. The null hypothesis is $H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3$. One can also conveniently rewrite the null hypothesis as $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0$, with $\gamma_k = \beta_k - \beta_{k-1}$ $(k = 1, 2, 3)$. The unconstrained alternative would assume at least one of the γ's is different from zero. A constrained view emerges when a monotone dose-response relationship is enforced. Since GEE is not a likelihood-based method, no conventional likelihood ratio and score tests are available, making Wald tests the natural choice.

1.4 One-Sidedness and Constraints

The examples underscore (i) that constrained one-sided tests are often of interest, (ii) that the preferred test (Wald/ LR / score) depends on the actual situation, and (iii) that one-sidedness can take a vector form, not only a scalar one.

The choice for one-sided tests is a common one in a number of areas, such as clinical trials (Welsh 1996; Piantadosi 1997). While these situations will not call for restrictions on the parameter space, the examples underscore such constraints are often natural. Additional situations where constraints occur are random-effects ANOVA models (Nelder 1954), overdispersion (Cox 1983; Smith and Heitjan 1993; Hines 1997; Lu 1997), clustering (Britton 1997), and homogeneity in stratified analyses (Liang 1987).

While the concepts and theory for one-sidedness are well known in the single-parameter case, the use of one-sided tests in the vector-parameter situation is less widespread. Several authors, including Gouriereoux, Holly, and Montfort (1982), Robertson, Wright, and Dykstra (1988), Silvapulle and Silvapulle (1995), and Silvapulle and Sen (2005), present results for (multivariate) one-sided tests subject to constraints on the parameter space. We should like to point out the intersection-union test (Welsh 1996), where similar issues arise. Since this a somewhat distinct strand of research, we will not consider it further in this article.

There is an important difference between constrained and unconstrained one-sided tests, in the scalar case, of the form

$$H_0 : \theta = 0 \quad \text{versus} \quad H_{Ac} : \theta > 0, \tag{4}$$

and

$$H_0 : \theta \leq 0 \quad \text{versus} \quad H_{Au} : \theta > 0, \tag{5}$$

respectively. While negative estimates for θ are allowed in (5), they are not in (4), where negative estimates are replaced by the boundary value $\widehat{\theta} = 0$. Due to this difference, different test statistics and null-distributions are needed, leading to different p values.

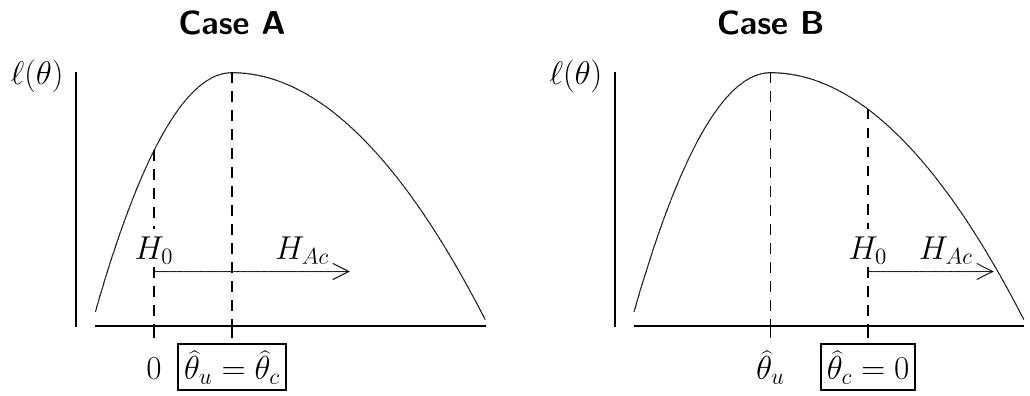

Figure 1. Graphical representation of two different situations, when developing one-sided tests. The function $\ell(\theta)$ represents the unconstrained log-likelihood function.

While the likelihood ratio test statistic case is relatively well understood (Self and Liang 1987; Stram and Lee 1994, 1995) and the score test, or Lagrange multiplier test, case has received attention (Gourieroux et al. 1982; Silvapulle and Silvapulle 1995; Hall and Præstgaard 2001; Verbeke and Molenberghs 2003), the Wald test has received little attention, Silvapulle (1992) being an exception. In this article, we will review each of the three tests in a constrained one-sided framework (Section 2). Their asymptotic equivalence will be presented, together with ways to compute p values. After applying them to the three examples in Section 3, computational issues will be discussed and guidelines for use formulated in Section 4. We will restrict to asymptotic inference, hence large sample null-distributions will be derived.

2. HYPOTHESIS TESTING IN A CONSTRAINED PARAMETER SPACE

In Section 1.4 we distinguished between the unconstrained and constrained ones-sided cases, for a scalar parameter θ. In Section 2.1 we will provide intuitive arguments regarding formulation and use of likelihood ratio, score, and Wald tests. Section 2.2 presents the general result for vector-valued parameters.

2.1 Intuitive Arguments

The way to proceed can be derived intuitively from Figure 1. Let $\widehat{\theta}_c$ denote the constrained estimator for θ, that is, the estimator under the condition $\theta \geq 0$. In case the unconstrained estimator $\widehat{\theta}_u \geq 0$, we have $\widehat{\theta}_c = \widehat{\theta}_u$, otherwise $\widehat{\theta}_c = 0$. These two situations are the cases A and B, respectively, in Figure 1 which under H_0 occur with probability 0.5 each. In case A, we have that the amount of evidence $\widehat{\theta}_c$ provides against H_0, in favor of H_{Ac}, can be measured by either one of the classical likelihood ratio statistic, the score statistic, and the Wald statistic, given by

$$T_{\mathrm{LR}} = 2\ln\left[\ell(\widehat{\theta}_u) - \ell(0)\right],$$

$$T_{\mathrm{S}} = \left[\left.\frac{\partial \ell(\theta)}{\partial \theta}\right|_{\theta=0}\right]^2 \left[-\left.\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta}\right|_{\theta=0}\right]^{-1},$$

and

$$T_{\mathrm{W}} = \left[\widehat{\theta}_u\right]^2 \left[-\left.\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta}\right|_{\theta=\widehat{\theta}_u}\right],$$

respectively. Classical likelihood theory implies that, under H_0, T_{LR}, T_{S}, and T_{W} are asymptotically equivalent and χ^2_1 distributed (Cox and Hinkley 1990).

Under case B, we clearly see that $\widehat{\theta}_c$ does not provide any evidence against H_0 in favor of H_{Ac}. This is easily seen in case of the likelihood ratio test, where $T_{\mathrm{LR}} = 0$. This lack of evidence can be reflected in the Score and Wald tests, by setting $T_{\mathrm{S}} = T_{\mathrm{W}} = 0$. Note that the Wald statistic which is traditionally used for testing one-sided unconstrained alternatives such as (5) would be different since it would be based on $\widehat{\theta}_u$, with a one-sided rejection region.

From our intuitive arguments, we infer that under H_0, T_{LR}, T_{S}, and T_{W} are asymptotically equivalent and χ^2_1 distributed in 50% of the cases, and equal to zero in the other 50% of the cases. Hence the null distribution is a mixture of the χ^2_0 (with all probability mass at zero) and χ^2_1, with equal probability 0.5. Let us now switch to the more formal and more general results.

2.2 Likelihood Ratio, Score, and Wald Tests

Self and Liang (1987) provided a general theory showing that the null distribution of the constrained likelihood ratio test statistic is a mixture of projections of χ^2 variables onto, possibly curved, surfaces. Only in special cases the weights of these mixtures can be derived analytically (Shapiro 1988; Raubertas, Lee, and Nordheim 1986). Stram and Lee (1994, 1995) used this theory to derive the asymptotic null distributions for the special but important case of variance components in linear mixed-effects models.

Let us switch to the score test and generically parameterize the model through a vector $\boldsymbol{\theta} = (\boldsymbol{\lambda}', \boldsymbol{\psi}')'$, where testing a general hypothesis of the form $H_0 : \boldsymbol{\psi} = \mathbf{0}$ versus $H_A : \boldsymbol{\psi} \in \mathcal{C}$ is of interest. Here, $\mathcal{C}$ is a closed and convex cone in the Euclidean space, with vertex at the origin. The advantage of such a general definition is that one-sided, two-sided, and combinations of one-sided and two-sided hypotheses are included. In (4), for example, $\mathcal{C} = \mathbb{R}_+$, the positive real line.

Denote the log-likelihood function by $\ell(\boldsymbol{\theta})$, the associated score function by $\mathbf{S}_N(\boldsymbol{\theta}) = \partial\ell/\partial\boldsymbol{\theta}$ and let $H(\boldsymbol{\theta})$ be the matrix of second derivatives of the log-likelihood function. Decompose $\mathbf{S}_N$ as $\mathbf{S}_N = (\mathbf{S}'_{N\lambda}, \mathbf{S}'_{N\psi})'$, let $H_{\lambda\lambda}(\boldsymbol{\theta})$, $H_{\lambda\psi}(\boldsymbol{\theta})$ and $H_{\psi\psi}(\boldsymbol{\theta})$ be the corresponding blocks in $H(\boldsymbol{\theta})$, and define $\boldsymbol{\theta}_H = (\boldsymbol{\lambda}', \mathbf{0}')'$. $\boldsymbol{\theta}_H$ can be estimated by $\widehat{\boldsymbol{\theta}}_H = (\widehat{\boldsymbol{\lambda}}', \mathbf{0}')'$, in which $\widehat{\boldsymbol{\lambda}}$ is the maximum likelihood estimate of $\boldsymbol{\lambda}$, under H_0. Finally, let $\mathbf{Z}_N$ be equal to $\mathbf{Z}_N = N^{-1/2}\mathbf{S}_{N\psi}(\widehat{\boldsymbol{\theta}}_H)$. A one-sided score statistic is now defined as

$$T_S := \mathbf{Z}'_N H^{-1}_{\psi\psi}(\widehat{\boldsymbol{\theta}}_H)\,\mathbf{Z}_N - \inf\left\{(\mathbf{Z}_N - \mathbf{b})' H^{-1}_{\psi\psi}(\widehat{\boldsymbol{\theta}}_H)(\mathbf{Z}_N - \mathbf{b}) | \mathbf{b} \in \mathcal{C}\right\} \quad (6)$$

(Silvapulle and Silvapulle 1995; Verbeke and Molenberghs 2003). Situation (4), intuitively handled in Section 2.1, follows as a special case since, when $\widehat{\theta}_u \geq 0$, the score test at zero is nonnegative, such that the infimum in (6) becomes zero. For $\widehat{\theta}_u < 0$, the score at zero is negative and the infimum in (6) is attained for $b = 0$, resulting in $T_S = 0$.

A Wald test counterpart for (6) is given by

$$T_W := \widehat{\boldsymbol{\psi}}' V^{-1}_{\psi\psi}\widehat{\boldsymbol{\psi}} - \inf\left\{(\widehat{\boldsymbol{\psi}} - \mathbf{b})' V^{-1}_{\psi\psi}(\widehat{\boldsymbol{\psi}} - \mathbf{b}) | \mathbf{b} \in \mathcal{C}\right\}, \quad (7)$$

where V is the asymptotic variance-covariance matrix of $\boldsymbol{\theta}$ and $V_{\psi\psi}$ is the corresponding submatrix (Silvapulle 1992; Silvapulle and Silvapulle 1995, theorem 1).

2.3 Equivalence and Implications for Critical Levels

It follows from Silvapulle and Silvapulle (1995) that, provided regularity conditions hold, the likelihood ratio, score, and Wald test statistics satisfy $T_{LR} = T_S + o_p(1) = T_W + o_p(1)$, for $N \to \infty$. The regularity conditions ensure that the score $\mathbf{S}_N$ differentiably exists in a sufficiently small neighborhood around H_0, thus even outside of $\mathcal{C}$. This is true, for example, for variance components in a linear mixed model (Verbeke and Molenberghs 2003), but not for the variance in a univariate normal sample, where the value $\sigma^2 = 0$ produces a singularity. In particular, they hold in all three case studies.

Shapiro (1988, eqs. 3.1 and 3.2) has shown that the null distribution equals a weighted sum of chi-squared distributions and that determining the mixture's weights often is a complex and perhaps numerical task, with important exceptions. For example, Stram and Lee (1994, 1995) derived that, when testing for k versus $k+1$ correlated random effects in a linear mixed model, the null distribution is $0.5(\chi^2_k + \chi^2_{k+1})$. Further, when jointly testing for k parameters $\psi_m = 0$ versus $\psi_m > 0$ $(m = 1, \ldots, k)$, the null distribution is a mixture of the form (Shapiro 1988)

$$\sum_{m=0}^{k} 2^{-k}\binom{k}{m}\chi^2_m.$$

Note that such mixtures can be calculated from a weighted average of the p values corresponding to each of the constituent χ^2 distributions. This is different from the distribution of a linear combination of χ^2 variables, as discussed by Davies (1980).

Table 1. Rat Data Example. Estimated D matrices for Models 1, 2, and 3, in the constrained and unconstrained cases.

Model	Constrained (one-sided)	Unconstrained (two-sided)
1	(3.44)	(3.44)
2	$\begin{pmatrix} 3.44 & 0 \\ 0 & 0.00 \end{pmatrix}$	$\begin{pmatrix} 3.77 & 0 \\ 0 & -0.17 \end{pmatrix}$
3	$\begin{pmatrix} 3.42 & 0.0089 \\ 0.0089 & 0.00023 \end{pmatrix}$	$\begin{pmatrix} 2.83 & 0.48 \\ 0.48 & -0.33 \end{pmatrix}$

3. ANALYZING THE CASE STUDIES

Starting with the rat data, introduced in Section 1.1, the estimated D matrices for Models 1, 2, and 3, with and without constraints are presented in Table 1. The corresponding test results are displayed in Table 2. All models are estimated using the SAS procedure MIXED. For Models 2 and 3, the estimates differ between the two settings. The scalar nature of $H_{0,1}$ and $H_{0,2}$ simplifies the computations for score and Wald statistics since optimizations as in (6) and (7) are trivial. It either produces zero, as happens for hypothesis $H_{0,1}$, meaning that the one-sided and two-sided test statistics are the same, or the first and second terms in (6) are equal, resulting in a zero test statistic value. The latter is the case for $H_{0,2}$. The first situation is entirely similar to conventional one-sided testing, since also there the test statistic remains unaltered and the p value is simply halved, which is exactly the effect of using $0.5(\chi^2_0 + \chi^2_1)$ as reference distribution. The third hypothesis is of a vector type. The one-sided score and Wald tests require proper minimization and, even though the parameter estimates are on the boundary of the parameter space, the one-sided test statistics differ slightly from zero. $H_{0,2}$ and $H_{0,3}$ are accepted by all tests, illustrating the asymptotic null equivalence. Since $H_{0,1}$ is rejected, the test statistics differ considerably, in line with theoretical considerations and empirical evidence (Declerck, Aerts, and Molenberghs 1998).

Switching to the toenail data, all models are fitted using the SAS procedure `nlmixed`, with conventional and adaptive Gaussian quadrature to integrate over the random effects. Quadrature depends on a tuning parameter, q, the number of quadrature points. Adaptive quadrature and higher q result in higher accuracy, at computational cost. Parameter estimates for the null model, obtained with adaptive quadrature ($q = 5$), are stable. Twice the negative log-likelihood is 1257.1 and $\widehat{d_{11}} = 3.70$ (s.e. 0.34). The alternative model is less stable with inflating estimates (standard errors) and strong dependence on the quadrature method and q. For example, the difference between nonadaptive ($q = 3$) and adaptive quadrature ($q = 50$) amounts to 40% for d_{11} under the null model. In the alternative model, $\widehat{d_{11}} = 15.0$ (s.e. 2.3), $\widehat{d_{12}} = 0.9$ (s.e. 0.6), and $\widehat{d_{22}} = 4.2$ with nonadaptive quadrature ($q = 3$), as opposed to $\widehat{d_{11}} = 66.8$ (s.e. 17.3), $\widehat{d_{12}} = -4.6$ (s.e. 1.7), and $\widehat{d_{22}} = 0.9$ (s.e. 0.3) with adaptive quadrature ($q = 30$). To exhibit the relative stability of the three test statistics we consider all three, in the two-sided and one-sided cases, under adaptive ($q = 30$) and nonadaptive

Table 2. Rat, Toenail, and NTP Data Examples. Test statistics (p values) of one- as well as two-sided, likelihood ratio, score, and Wald tests for the comparison of a series of models. Rat data example: Four nested linear mixed-effects models are compared. Toenail data example: A random-intercepts model is compared with a model containing random intercepts and random slopes. Calculations are based on both adaptive Gaussian quadrature with 30 quadrature points or on nonadaptive Gaussian quadrature with three quadrature points. NTP data example: Using GEE, the hypothesis of no dose effects is tested.

Hypothesis	*sided*	*Null distribution*	*LR*	*score*	*Wald*
		Rat data example			
$H_{0,1}$	one	$\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$	174.0(<0.001)	27.2(<0.001)	19.8(<0.001)
	two	χ^2_1	174.0(<0.001)	27.2(<0.001)	19.8(<0.001)
$H_{0,2}$	one	$\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$	0.0(1.000)	0.0(1.000)	0.0(1.000)
	two	χ^2_1	1.3(0.254)	1.7(0.196)	1.8(0.182)
$H_{0,3}$	one	$\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$	0.1(0.852)	0.1(0.852)	0.04(0.905)
	two	χ^2_2	2.9(0.235)	2.0(0.362)	4.4 (0.112)
		Toenail data example			
ad, $q = 30$	one	$\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$	119.7(<0.001)	73.5(<0.001)	11.2(0.004)
	two	χ^2_2	119.7(<0.001)	73.5(<0.001)	11.2(0.002)
nonad, $q = 3$	one	$\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$	113.8(<0.001)	73.1(<0.001)	0.02(0.990)
	two	χ^2_2	113.8(<0.001)	73.1(<0.001)	0.02(0.940)
		NTP data example			
no dose	one	$\frac{1}{8}\chi^2_0 + \frac{3}{8}\chi^2_1 + \frac{3}{8}\chi^2_2 + \frac{1}{8}\chi^2_3$		43.1(<0.001)	83.6(<0.001)
	two	χ^2_3		43.1(<0.001)	83.6(<0.001)

($q = 3$) quadrature (Table 2). The likelihood ratio test statistic is somewhat unstable, the Wald statistic is extremely so, with diametrically opposed test results, in the one-sided and two-sided cases. The score test is remarkably insensitive to the choice of quadrature method.

The NTP data are modeled using GEE. The H matrix in (6) needs to be replaced by $H_{\psi\psi}V^{-1}_{\psi\psi}H_{\psi\psi}$, where H is the model-based variance, found as the limit of the first derivative of the score, and V is the variability in the score, the so-called information sandwich. The two-sided version for GEE is the score test T_S of Rotnitzky and Jewell (1990, p. 488) and the one-sided version now easily follows too. Parameter estimates are obtained with the SAS procedure GENMOD and presented in Table 3. The Wald test's null distribution is χ^2_3 in the unconstrained two-sided case, and $\frac{1}{8}\chi^2_0 + \frac{3}{8}\chi^2_1 + \frac{3}{8}\chi^2_2 + \frac{1}{8}\chi^2_3$ in the constrained one-sided case. Test results are displayed in Table 2.

4. WHAT TO DO AND HOW TO DO IT IN PRACTICE?

The asymptotic equivalence of the tests begs the question as to which one should be used in what circumstances. A combination of statistical and computational considerations will guide the choice and may differ somewhat between cases, but some patterns emerge.

When a nonlikelihood method is used, such as generalized estimating equations (NTP data), there is no genuine likelihood function and hence the Wald test is the natural option, unconstrained and constrained. Alternatively, a generalized score test can be used as presented by Rotnitzky and Jewell (1990), and implemented in the SAS procedure `genmod` for GEE. They presented two forms for both the Wald and score tests in the GEE case: the one implemented here uses the working covariance structure and follows a classical χ^2_p null distribution in the two-sided case. Alternatively, these authors propose a modified version for each of these tests, of which the null distribution is a mixture of χ^2 distributions in the two-sided case. Studying the latter test statistics in the constrained one-sided case would be of interest but falls outside of the scope of this article.

When a likelihood method is employed, such as a (generalized) linear mixed model, all three tests are possible, in principle, and some guidance is needed. Let us make a few considerations.

First, in the unconstrained case, likelihood ratio test and Wald test statistics are easy to compute and standard in most packages. The score test is slightly more difficult, since the first- and second-order derivatives of the alternative log-likelihood

Table 3. NTP Data. GEE parameter estimates (empirically corrected standard errors) for Model (3) fitted to the collapsed outcome in the EG data.

Dose group	*Parameter*	*Estimate (s.e.)*
0 mg/kg/day	β_0	−5.88 (0.98)
750 mg/kg/day	β_1	−2.22 (0.21)
1500 mg/kg/day	β_2	−0.43 (0.26)
3000 mg/kg/day	β_3	0.29 (0.27)

function, evaluated under the null hypothesis, are required. These cannot always be obtained very easily in many standard packages without additional programming. Once these derivatives are obtained, they are the straightforward building blocks for the calculation of the test statistic.

Second, the situation is subtly different in the constrained case. The constrained likelihood ratio test statistic can be obtained as in the unconstrained case, without additional computation, provided the constraints are properly imposed onto the alternative model. The constrained Wald and score test statistics (6) and (7) are composed of two parts. The first term is identical to the unconstrained counterpart, while the second term involves a constrained minimization of the quadratic forms $(\mathbf{Z}_N - \mathbf{b})' H^{-1}_{\psi\psi}(\widehat{\boldsymbol{\theta}}_H)(\mathbf{Z}_N - \mathbf{b})$ and $(\boldsymbol{\psi} - \mathbf{b})' V^{-1}_{\psi\psi}(\boldsymbol{\psi} - \mathbf{b})$, respectively, which cannot always be done analytically. In this minimization, the parameters occurring in the null hypothesis are kept fixed at their null values and minimization is over the remaining parameters, within the cone $\mathcal{C}$. In such cases, a numerical constrained optimization routine needs to be invoked. Example GAUSS code is available from the authors.

Thus, pragmatically, the likelihood ratio test is the easiest to evaluate and we suggest to consider it the default in a constrained likelihood-based approach. This is the route followed in the rats example. An exception is when the alternative model is difficult to fit, for example, because estimates are unstable, or convergence is hard to reach. Wald tests are then arguably even more vulnerable, leading to the score test as the preferred choice, as in the NTP data.

We have sketched a framework for multivariate one-sided testing in constrained parameter spaces and provided some guidelines to navigate within it. Of course, we do not claim to have provided a definitive answer, for which both additional small sample and asymptotic evaluations, accompanied with simulations, would be needed.

[Received January 2006. Revised September 2006.]

REFERENCES

Aerts, M., Geys, H., Molenberghs, G., and Ryan, L. (2002), *Topics in Modelling of Clustered Data*, London: Chapman & Hall.

Britton, T. (1997), "Tests to Detect Clustering of Infected Individuals Within Families," *Biometrics*, 53, 98–109.

Cox, D. R. (1983), "Some Remarks on Overdispersion," *Biometrika*, 70, 269–274.

Cox, D. R., and Hinkley, D. V. (1990), *Theoretical Statistics*, London: Chapman & Hall.

Davies, R. B. (1980), "The Distribution of a Linear Combination of χ^2 Random Variables," *Applied Statistics*, 29, 323–333.

De Backer, M., De Keyser, P., De Vroey, C., and Lesaffre, E. (1996), "A 12-week Treatment for Dermatophyte Toe Onychomycosis: Terbinafine 250mg/day vs. Itraconazole 200mg/day—A Double-Blind Comparative Trial," *British Journal of Dermatology*, 134, 16–17.

Declerck, L., Aerts, M., and Molenberghs, G. (1998), "Behaviour of the Likelihood Ratio Test Statistic Under a Bahadur Model for Exchangeable Binary Data," *Journal of Statistical Computation and Simulation*, 61, 15–38.

Gourieroux, C. A., Holly, A., and Montfort, A. (1982), "Likelihood Ratio Test, Wald Test, and Kuhn-Tucker Test in Linear Model with Inequality Constraints on the Regression Parameter," *Econometrica*, 50, 63–79.

Hall, D. B., and Præstgaard, J. T. (2001), "Order-Restricted Score Tests for Homogeneity in Generalised Linear and Nonlinear Mixed Models," *Biometrika*, 88, 739–751.

Hines, R. J. O. (1997), "A Comparison of Tests for Overdispersion in Generalized Linear Models," *Journal of Statistical Computation and Simulation*, 58, 323–342.

Liang, K.-Y. (1987), "A Locally Most Powerful Test for Homogeneity With Many Strata," *Biometrika*, 74, 259–264.

Liang, K.-Y., and Zeger, S.L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Lu, W. S. (1997), "Score Tests for Overdispersion in Poisson Regression Models," *Journal of Statistical Computation and Simulation*, 56, 213–228.

Nelder, J.A. (1954), "The Interpretation of Negative Components of Variance," *Biometrika*, 41, 544–548.

Piantadosi, S. (1997), *Clinical Trials: A Methodologic Perspective*, New York: Wiley.

Price, C. J., Kimmel, C. A., Tyl, R. W., and Marr, M. C. (1985), "The Developmental Toxicity of Ethylene Glycol in Mice," *Toxicology and Applied Pharmacology*, 81, 113–127.

Raubertas, R. F., Lee, C.I.C., and Nordheim, E.V. (1986), "Hypothesis Tests for Normal Means Constrained by Linear Inequalities," *Communications in Statistics—Theory and Methods*, 15, 2809–2833.

Roberts, D.T. (1992), "Prevalence of Dermatophyte Onychomycosis in the United Kingdom: Results of an Omnibus Survey," *British Journal of Dermatology*, 126 Suppl. 39, 23–27.

Robertson, T., Wright, F.T., and Dykstra, R.L. (1988), *Order Restricted Statistical Inference*, New York: Wiley.

Rotnitzky, A., and Jewell, N.P. (1990), "Hypothesis Testing of Regression Parameters in Semiparametric Generalized Linear Models for Cluster Correlated Data," *Biometrika*, 77, 485–497.

Self, S. G., and Liang, K.Y. (1987), "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions," *Journal of the American Statistical Association*, 82, 605–610.

Shapiro, A. (1988), "Towards a Unified Theory of Inequality Constrained Testing in Multivariate Analysis," *International Statistical Review*, 56, 49–62.

Silvapulle, M.J. (1992), "Robust Wald-Type Tests of One-Sided Hypotheses in the Linear Model," *Journal of the American Statistical Association*, 87, 156–161.

Silvapulle, M. J., and Sen, P. K. (2005), *Constrained Statistical Inference*, New York: Wiley.

Silvapulle, M. J., and Silvapulle, P. (1995), "A Score Test Against One-Sided Alternatives," *Journal of the American Statistical Association*, 90, 342–349.

Smith, P. J., and Heitjan, D. F. (1993), "Testing and Adjusting for Departures From Nominal Dispersion in Generalized Linear Models," *Applied Statistics*, 41, 31–41.

Stram, D. O., and Lee, J. W. (1994), "Variance Components Testing in the Longitudinal Mixed Effects Model," *Biometrics*, 50, 1171–1177.

——— (1995), Correction to "Variance Components Testing in the Longitudinal Mixed Effects Model," *Biometrics*, 51, 1196.

Verbeke, G., and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, New York: Springer.

——— (2003), "The Use of Score Tests for Inference on Variance Components," *Biometrics*, 59, 254–262.

Verdonck, A., De Ridder, L., Verbeke, G., Bourguignon, J.P., Carels, C., Kuhn, E.R., Darras, V., and de Zegher, F. (1998), "Comparative Effects of Neonatal and Prepubertal Castration on Craniofacial Growth in Rats," *Archives of Oral Biology*, 43, 861–871.

Welsh, A. H. (1996), *Aspects of Statistical Inference*, New York: Wiley.