ON GENERALIZED SCORE TESTS

Dennis D. Boos

Institute of Statistics Mimeo Series No. 1980

September 1990

ABSTRACT

Generalizations of Rao's score test are receiving increased attention, especially in the econometrics and biostatistics literature. These generalizations are able to account for certain model inadequacies or lack of knowledge by use of empirical variance estimates. This article shows how the various forms of the generalized test statistic arise from Taylor expansion of the estimating equations. The general estimating equations structure unifies a variety of applications and helps suggest new areas of application.

Key Words: composite null hypothesis, empirical variance, estimating equations, information sandwich, Lagrange multiplier test, misspecified likelihood, observed information, robust inference.

ON GENERALIZED SCORE TESTS

1. Introduction

Rao (1948) introduced score statistics for composite null hypotheses having the form

$$S(\tilde{\theta})^T \tilde{I}_f^{-1} S(\tilde{\theta}) \ , \tag{1}$$

where $S(\theta)$ is the vector of partial derivatives of the log likelihood function, $\tilde{\theta}$ is the vector of restricted maximum likelihood estimates under H_0, and $\tilde{I}_f$ is the Fisher information of the sample evaluated at $\tilde{\theta}$. These test statistics are attractive because they only require computation of the null estimates $\tilde{\theta}$ and are asymptotically equivalent to Wald and likelihood ratio statistics under both null and Pitman alternative hypotheses. In fact many common tests statistics such as the Pearson chi-square are score statistics or are closely related. A parallel development of (1) was begun by Aitchison and Silvey (1958) under the name "Lagrange multiplier statistic," and the econometrics literature uses this latter term. Both "score statistic" and "Lagrange multiplier statistic" seem to be the names used regardless of whether expected value information $\tilde{I}_f$ or observed information is used in (1). Introductions to score and Lagrange multiplier tests may be found in Breusch and Pagan (1980), Buse (1982), Engle (1984), Hosking (1983), and Tarone (1988).

The purpose of this note is to discuss the use of score tests in the general estimating equations situation where $\hat{\theta}$ is obtained by minimizing an objective function $Q(\theta)$ or by solving the related vector equation

$$S(\theta) = -\frac{\partial Q(\theta)}{\partial \theta} = 0 \ . \tag{2}$$

Typically $-Q(\theta)$ is the log likelihood function and we are concerned about misspecification of the underlying model, or $Q(\theta)$ is chosen so that $\hat{\theta}$ is reliable over a range of possible models (robust or semiparametric methods), or $Q(\theta)$ is chosen for computational convenience (e.g., least squares). I will tend to use the language of "misspecified likelihoods" even though the results apply in general. Note that when $-Q(\theta)$ is the log likelihood function, then $S(\theta)$ is commonly called the score function.

Wald and likelihood ratio tests are fairly easy to generalize to misspecified models (Section 3) although approximations to their null distributions are not always adequate or easy to derive. The generalizations for score tests are not as simple, and I have found the literature somewhat lacking in motivation regarding their derivation. However, these generalized score tests may be the most useful of all three types of tests, and I would like to highlight a few of the details in their derivation.

Specifically, I would like to suggest the following general asymptotic principle for construction of generalized score statistics:

> "Find the asymptotic covariance matrix of $S(\tilde{\theta})$ under H_0, say Σ_S.
>
> Then define the generalized score statistic to be
>
> $$T_{GS} = S(\tilde{\theta})^T \tilde{\Sigma}_S^- S(\tilde{\theta}) ,$$
>
> where $\tilde{\Sigma}_S^-$ is a generalized inverse of a consistent estimate $\tilde{\Sigma}_S$ of Σ_S."

By consistent estimate I mean $a_n(\tilde{\Sigma}_S - \Sigma_S) \xrightarrow{p} 0$ in large samples for some scalar a_n since Σ_S may not be standardized to converge. Although a generalized inverse seems unnecessarily messy to deal with, we can usually find an explicit version which is easily computed. Moreover, the form of T_{GS} essentially guarantees that it has the correct kind of asymptotic null chi-squared distribution (e.g., Theorem 3.5, p. 128 of Serfling, 1980).

Some authors (e.g., Breusch and Pagan, p. 240) suggest that the form of Rao's statistic (1) arises from noting that with correctly specified likelihoods, the covariance matrix of $S(\theta)$ is the Fisher information, and replacing θ by the restricted estimates $\tilde{\theta}$ leads to (1). Unfortunately, this method does not generalize to using consistent estimates of the covariance matrix of $S(\theta)$ when the likelihood is misspecified. In fact, it seems almost "lucky" that the form (1) is the appropriate statistic in the correctly specified case. The equivalent Lagrange multiplier version of (1) by Aitchison and Silvey (1958) appears to be derived more in the spirit of the general asymptotic principle mentioned above.

The paper is organized as follows. Section 2 introduces the distributional structure and notation, and Section 3 briefly discusses the generalized Wald and likelihood ratio tests. Section 4 shows how the various forms of the generalized score statistic arise from asymptotic Taylor expansion. Regularity conditions will be avoided because they can hinder conceptual understanding and are best left to specific applications.

2. Notation and Structure

To give some structure to the problem, suppose that the data $Y_1,\ldots,Y_n$ are independent and $Y_i|x_i$ has density $g(y,x_i)$, where the x_i are fixed regressor vectors. For likelihood analysis the presumed model will be $f(y,x_i|\theta)$, where θ is a $b \times 1$ vector of unknown parameters. In general, $Q(\theta)$ need not arise from a parametric family, but θ will still be a meaningful $b \times 1$ vector, and we will assume that $Q(\theta)$ has the form $Q(\theta) = \sum_{i=1}^{n} q(Y_i,x_i,\theta)$. Solutions to (2) will be denoted by $\hat{\theta}$, and $\tilde{\theta}$ will be used to denote related estimators under null hypothesis constraints.

We shall assume that there is a unique θ^* in the parameter space such that $\hat{\theta} \xrightarrow{p} \theta^*$ as $n \to \infty$ and $\tilde{\theta} \xrightarrow{p} \theta^*$ under H_0 as $n \to \infty$. Hypotheses will be phrased in terms of θ^*, and one always has to decide whether such hypotheses are meaningful under misspecification. For example, in linear regression contexts where $\theta^T = (\beta_0^T, \beta_1^T)$ and $Y_i = \beta_0 + x_i^T\beta_1 + e_i$, typically $\beta_1^* = \beta_1$ regardless of the estimation method or distribution of $e_1,\ldots,e_n$, and thus tests about β_1^* are meaningful.

Let $I_Y = -\partial S(\theta) \big/ \partial\theta^T = \sum_{i=1}^{n} \partial^2 q(Y_i,x_i,\theta)/\partial\theta\partial\theta^T$, and let $\hat{I}_Y$, $\tilde{I}_Y$, and $\overset{*}{I}_Y$ be the values of I_Y when $\theta = \hat{\theta}$, $\tilde{\theta}$, and θ^*, respectively. Note that for maximum likelihood $\hat{I}_Y$

is the observed sample information matrix of the presumed model. Let E_g denote expectation with respect to the true densities. Then define $I_g = E_g[I_Y|x_1,\ldots,x_n] = \int I_y \prod g(y_i,x_i)dy_1 \ldots dy_n$, and let $\hat{I}_g$, $\tilde{I}_g$, and $\overset{*}{I}_g$ be the values of I_g when $\theta = \hat{\theta}$, $\tilde{\theta}$, and θ^*, respectively. When maximum likelihood is used, we may define the analogous model quantities $I_f = \int I_y \prod f(y_i,x_i,\theta)dy_1,\ldots,dy_n$ and $\hat{I}_f$, $\tilde{I}_f$, and $\overset{*}{I}_f$. I_f is called the expected value information or Fisher information of the model. These latter quantities play no role in the general asymptotics since typically $I_g \neq I_f$ and I_g and I_Y are the important quantities in the Taylor expansions where laws of large numbers give $I_g^{-1} I_Y \xrightarrow{p}$ identity matrix. Finally, with $s_i(\theta) = -\partial q(Y_i,x_i,\theta) \big/ \partial\theta$ and thus $S(\theta) = \sum_{i=1}^{n} s_i(\theta)$, define $D_Y = \sum_{i=1}^{n} s_i(\theta)s_i(\theta)^T$, $D_g = E_g D_Y$, and $D_f = E_f D_Y$. The definitions of $\hat{D}_Y$, $\tilde{D}_Y$, $\overset{*}{D}_Y$, etc., should be obvious. Under suitable regularity conditions on $f(y,x_i,\theta)$, $D_f = I_f$, but in general one would not expect $D_g = I_g$. In fact White (1982) has proposed model adequacy tests based on $\hat{I}_Y - \hat{D}_Y$ which estimates $I_g - D_g$.

Summarizing for easy reference:

$$S(\theta) = \frac{-\partial Q(\theta)}{\partial\theta} = -\sum_{i=1}^{n} \frac{\partial q(Y_i,x_i,\theta)}{\partial\theta} = \sum_{i=1}^{n} s_i(\theta)$$

$\hat{\theta}$ minimizes $Q(\theta)$ and satisfies $S(\hat{\theta}) = 0$

$\tilde{\theta}$ minimizes $Q(\theta)$ subject to null hypothesis constraints

$\overset{*}{\theta}$ probability limit of $\hat{\theta}$ and $\tilde{\theta}$.

2nd Derivative Matrices		Evaluated at		
		$\hat{\theta}$	$\tilde{\theta}$	$\overset{*}{\theta}$
$I_Y = \frac{-\partial S(\theta)}{\partial\theta^T} = \frac{\partial^2 Q(\theta)}{\partial\theta\partial\theta^T}$	→	$\hat{I}_Y$	$\tilde{I}_Y$	$\overset{*}{I}_Y$
$I_g = E_g[I_Y|x_1,\ldots,x_n]$	→	$\hat{I}_g$	$\tilde{I}_g$	$\overset{*}{I}_g$

$$g = \prod_{i=1}^{n} g(y_i,x_i) = \text{true density}$$

$$I_f = E_f[I_Y|x_1,\ldots,x_n] \qquad \rightarrow \qquad \hat{I}_f \quad \tilde{I}_f \quad \overset{*}{I}_f$$

$$f = \prod_{i=1}^{n} f(y_i,x_i|\theta) = \text{presumed density when likelihood methods used}$$

1st Derivative "Empirical Variances"

$$D_Y = \sum_{i=1}^{n} s_i(\theta)s_i(\theta)^T \qquad \rightarrow \qquad \hat{D}_Y \quad \tilde{D}_Y \quad \overset{*}{D}_Y$$

$$D_g = E_g[D_Y|x_1,\ldots,x_n] \qquad \rightarrow \qquad \hat{D}_g \quad \tilde{D}_g \quad \overset{*}{D}_g$$

$$D_f = E_f[D_Y|x_1,\ldots,x_n] \qquad \rightarrow \qquad \hat{D}_f \quad \tilde{D}_f \quad \overset{*}{D}_f$$

Composite null hypotheses are usually specified by either

I. $\underset{1\times b}{\theta^{*T}} = \left(\underset{1\times r}{\theta_1^{*T}} \; , \; \underset{1\times b-r}{\theta_2^{*T}} \right)$, H_0: $\theta_1^* = \theta_{10}$ $\qquad r < b$,

or

II. H_0: $\underset{r\times 1}{h(\theta^*)} = 0$, where $\underset{r\times b}{H(\theta)} = \dfrac{\partial h(\theta)}{\partial \theta^T}$ has full row rank r,

or

III. H_0: $\theta^* = g(\beta)$, where β is an $b - r$ dimensional parameter

and $\underset{b\times b-r}{G(\beta)} = \dfrac{\partial g(\beta)}{\partial \beta^T}$ has full column rank $b - r$.

Recall that use of θ^* admits the possibility that $\hat{\theta}$ and $\tilde{\theta}$ converge to a value different from the true θ of the generating mechanism.

3. Generalized Wald and Likelihood Ratio Tests

The asymptotic distribution of $\hat{\theta}$ is easily ascertained through Taylor expansion of (2),

$$0 = \mathrm{S}(\hat{\theta}) = \mathrm{S}(\theta^*) + \frac{\partial \mathrm{S}(\theta^*)}{\partial \theta^T}\ (\hat{\theta} - \theta^*) + \mathrm{R}_{\mathrm{n1}}\ ,$$

leading to

$$\hat{\theta} - \theta^* = \overset{*}{\mathrm{I}}_{\mathrm{g}}^{-1}\mathrm{S}(\theta^*) + \mathrm{R}_{\mathrm{n2}}\ .$$

Under suitable regularity conditions, $\hat{\theta}$ is asymptotically normal $(\theta^*, \overset{*}{\mathrm{V}}_{\mathrm{g}})$, where $\overset{*}{\mathrm{V}}_{\mathrm{g}} = \overset{*}{\mathrm{I}}_{\mathrm{g}}^{-1}\overset{*}{\mathrm{D}}_{\mathrm{g}}\overset{*}{\mathrm{I}}_{\mathrm{g}}^{-1}$. $\overset{*}{\mathrm{V}}_{\mathrm{g}}$ is of course very familiar to those working in robust statistics and with maximum likelihood under misspecification (e.g., Huber, 1967; Burguete, Gallant, and Souza, 1982; Kent, 1982; White, 1982; Royall, 1986). The term information "sandwich" is sometimes used for the estimate $\hat{\mathrm{V}}_{\mathrm{Y}} = \hat{\mathrm{I}}_{\mathrm{Y}}^{-1}\hat{\mathrm{D}}_{\mathrm{Y}}\hat{\mathrm{I}}_{\mathrm{Y}}^{-1}$ (c.f., Lin and Wei, 1989, p. 1074) since $\hat{\mathrm{I}}_{\mathrm{Y}}$ is the observed information in the sample when $-\mathrm{Q}(\theta)$ is the log likelihood function.

Turning to null hypotheses I and II given in Section 2, the appropriate generalized Wald tests are

$$\mathrm{T}_{\mathrm{GWI}} = (\hat{\theta}_1 - \theta_{10})^T[\hat{\mathrm{V}}_{\mathrm{Y}}^{-1}]_{11}(\hat{\theta}_1 - \theta_{10})\ ,$$

and

$$\mathrm{T}_{\mathrm{GWII}} = \mathrm{h}(\hat{\theta})^T(\mathrm{H}(\hat{\theta})\hat{\mathrm{V}}_{\mathrm{Y}}\mathrm{H}(\hat{\theta})^T)^{-1}\mathrm{h}(\hat{\theta})$$

(see e.g., Kent, 1982, p. 23, and White, 1982, p. 8). Gallant (1987, p. 219) and others note that the null asymptotic chi-squared approximation for Wald statistics is not always adequate and that the statistic is not invariant to reparameterization.

The generalized likelihood ratio statistic may be given by

$$T_{GLR} = 2(Q(\tilde{\theta}) - Q(\hat{\theta}))$$

for any method of specifying the null hypothesis. Unfortunately, the asymptotic null distribution of T_{GLR} is generally a weighted sum of chi-squared random variables (Foutz and Srivastava, 1977; Kent, 1982; Hampel et al., 1986, p. 352) with unknown weights. In certain situations such as homoscedastic linear regression, $2(Q(\tilde{\theta}) - Q(\hat{\theta}))$ can be divided by a single weight estimate to get the usual asymptotic chi-squared distribution (e.g., Hettmansperger and Schrader, 1980).

4. Generalized Score Test

4.1 Hypotheses about Subvectors

The partitioned vector $\theta^T = (\theta_1^T, \theta_2^T)$ with hypotheses of the form I. H_0: $\theta_1^* = \theta_{10}$ is the easiest type of situation to handle. Let $S(\theta)^T = (S_1(\theta)^T, S_2(\theta)^T)$, where θ_1 and S_1 are $r \times 1$, θ_2 and S_2 are $b - r \times 1$, and $S_2(\tilde{\theta}) = 0$ from (2) with $\tilde{\theta}_1 = \theta_{10}$. The matrices defined in Section 2 are partitioned accordingly, e.g.,

$$D_Y = \begin{pmatrix} \underset{r \times r}{D_{Y11}} & \underset{r \times b-r}{D_{Y12}} \\ \underset{b-r \times r}{D_{Y21}} & \underset{b-r \times b-r}{D_{Y22}} \end{pmatrix} .$$

Rao's statistic (1) is

$$T_S = (S_1(\tilde{\theta})^T, 0) \begin{pmatrix} \tilde{I}_{f11} & \tilde{I}_{f12} \\ \tilde{I}_{f21} & \tilde{I}_{f22} \end{pmatrix}^{-1} \begin{pmatrix} S_1(\tilde{\theta}) \\ 0 \end{pmatrix} .$$

$$= S_1(\tilde{\theta})(\tilde{I}_{f11} - \tilde{I}_{f12}\tilde{I}_{f22}^{-1}\tilde{I}_{f21})^{-1}S_1(\tilde{\theta}) \;.$$

Under H_0 and a correctly specified likelihood $T_S \xrightarrow{d} \chi^2_r$. To get the generalized form we expand S_1 and S_2:

$$S_1(\tilde{\theta}) = S_1(\theta^*) - \overset{*}{I}_{Y12}(\tilde{\theta}_2 - \theta_2^*) + R_{n3}$$

$$0 = \; S_2(\tilde{\theta}) = S_2(\theta^*) - \overset{*}{I}_{Y22}(\tilde{\theta}_2 - \theta_2^*) + R_{n4} \;.$$

Then, replacing $\overset{*}{I}_{Y12}$ and $\overset{*}{I}_{Y22}$ by their asymptotically equivalent versions $\overset{*}{I}_{g12}$ and $\overset{*}{I}_{g22}$ and plugging in $\tilde{\theta}_2 - \theta_2^* \approx \overset{*}{I}_{g22}^{-1}S_2(\theta^*)$ from the second equation into the first equation, yields

$$S_1(\tilde{\theta}) = (I_r, \; - \overset{*}{I}_{g12}\overset{*}{I}_{g22}^{-1})\begin{pmatrix} S_1(\theta^*) \\ S_2(\theta^*) \end{pmatrix} + R_{n5} \;.$$

Here I_r is the r × r identity matrix. Finally, since the covariance matrix of $S(\theta^*)$ is $\overset{*}{D}_g$, the asymptotic covariance matrix of $S_1(\tilde{\theta})$ is

$$\overset{*}{V}_{gS_1} = (I_r, \; - \overset{*}{I}_{g12}\overset{*}{I}_{g22}^{-1})\overset{*}{D}_g(I_r, \; - \overset{*}{I}_{g12}\overset{*}{I}_{g22}^{-1})^T$$

$$= \overset{*}{D}_{g11} - \overset{*}{I}_{g12}\overset{*}{I}_{g22}^{-1}\overset{*}{D}_{g12}^T - \overset{*}{D}_{g12}\overset{*}{I}_{g22}^{-1}\overset{*}{I}_{g12}^T + \overset{*}{I}_{g12}\overset{*}{I}_{g22}^{-1}\overset{*}{D}_{g22}\overset{*}{I}_{g22}^{-1}\overset{*}{I}_{g12}^T \;.$$

Estimating $\overset{*}{D}_{g11}$ by $\tilde{D}_{Y11}$, etc., yields the consistent estimate $\tilde{V}_{YS_1}$ and

$$T_{GS} = S_1(\tilde{\theta})^T\tilde{V}_{YS_1}^{-1}S_1(\tilde{\theta}) \,. \tag{3}$$

This is the form for T_{GS} given by Breslow (1990, p. 567), whereas Kent (1982, p. 23) and Engle (1984, p. 822) have $\overset{*}{V}_{gS_1}^{-1}$ as

$$\overset{*}{V}_{gS_1}^{-1} = \overset{*}{I}_g^{11}\left[\left(\overset{*}{I}_g^{-1}\overset{*}{D}_g\overset{*}{I}_g^{-1}\right)_{11}\right]^{-1}\overset{*}{I}_g^{11} \;,$$

where $\overset{*}{I}{}_g^{11} = \left(\overset{*}{I}_{g11} - \overset{*}{I}_{g12}\overset{*}{I}{}_{g22}^{-1}\overset{*}{I}_{g21}\right)^{-1}$.

In some situations it may be preferable to use unrestricted estimates of θ in the estimate of $\overset{*}{V}_{gS_1}$ (see Example 1 below), and here we might use the notation $\hat{T}_{GS} = S_1(\tilde{\theta})\hat{V}_{YS_1}^{-1}S_1(\tilde{\theta})$. If it is easier to phrase the null hypothesis in terms of H_0: $\theta_2^* = \theta_{20}$, where θ_2 is $r \times 1$, then just note that $T_{GS} = S_2(\tilde{\theta})^T\tilde{V}_{YS_2}^{-1}S_2(\tilde{\theta})$, where $\tilde{V}_{YS_2}$ is notationally obtained from $\tilde{V}_{YS_1}$ by interchanging the subscripts 1 and 2. Of course under H_0 and suitable regularity conditions, $T_{GS} \rightarrow \chi_r^2$.

Example 1. Normal distribution maximum likelihood. If $Y_1,\ldots,Y_n$ are assumed to be iid normal (μ,σ^2), $\theta = (\mu,\sigma)$, then the negative log likelihood is $Q(\theta) = c + n\log\sigma + (2\sigma^2)^{-1}\Sigma(Y_i - \mu)^2$ and

$$S_\mu(\theta) = \sigma^{-2}\Sigma(Y_i - \mu)\ , \quad S_\sigma(\theta) = \Sigma[\sigma^{-3}(Y_i - \mu)^2 - \sigma^{-1}]\ ,$$

$$I_{Y\mu\mu} = n\sigma^{-2},\ \ I_{Y\mu\sigma} = 2\sigma^{-3}\Sigma(Y_i - \mu),\ \ I_{Y\sigma\sigma} = \Sigma[-\sigma^{-2}+3\sigma^{-4}(Y_i - \mu)^2]\ ,$$

$$D_{Y\mu\mu} = \sigma^{-4}\Sigma(Y_i - \mu)^2,\ \ D_{Y\mu\sigma} = \sigma^{-5}\Sigma[-\sigma^2(Y_i - \mu)+(Y_i - \mu)^3],$$

$$D_{Y\sigma\sigma}=\sigma^{-6}\Sigma[\sigma^2 - (Y_i - \mu)^2]^2\ .$$

Consider first H_0: $\mu^*=\mu_0$. The restricted mle for σ is $\tilde{\sigma} = [n^{-1}\Sigma(Y_i - \mu_0)^2]^{1/2}$ and the score statistic (1) is

$$\frac{n(\bar{Y} - \mu_0)^2}{\tilde{\sigma}^2} = \frac{n(\bar{Y} - \mu_0)^2}{S_n^2 + (\bar{Y} - \mu_0)^2}\ ,$$

where $S_n^2 = n^{-1}\Sigma(Y_i - \bar{Y})^2$. If observed information $\tilde{I}_Y$ is used in (1) place of expected information $\tilde{I}_f$, the denominator becomes $S_n^2 - (\bar{Y} - \mu_0)^2$. If the unrestricted estimator $\hat{\sigma}^2 = S_n^2$ is used in either I_f or I_Y, then (1) becomes $n(\bar{Y} - \mu_0)^2 / S_n^2$ which is the Wald statistic for this H_0.

Turning to the generalized score statistic, we find (3) yields

$$T_{GS} = \frac{n(\overline{Y} - \mu_0)^2}{\tilde{\sigma}^2}\left[1 - 2\left(\frac{\overline{Y} - \mu_0}{\tilde{\sigma}}\right)\tilde{\alpha}_3 + \left(\frac{\overline{Y} - \mu_0}{\tilde{\sigma}}\right)^2(\tilde{\alpha}_4 + 1)\right]^{-1},$$

where $\tilde{\alpha}_i = n^{-1}\Sigma(Y_i - \mu_0)^i / \tilde{\sigma}^i$, i = 3, 4. Thus T_{GS} has a curious skewness and kurtosis adjustment which should be negligible near H_0. If the unrestricted estimator $\hat{\theta} = (\overline{Y}, S_n)$ is used in I_Y and D_Y, then T_{GS} becomes the Wald statistic $n(\overline{Y} - \mu_0)^2 / S_n^2$ which seems preferable here. In general, inference via (1) about a mean vector when the data is assumed normal will be asymptotically valid even when the data is not normal. This robustness to nonnormality does not hold for inference about σ as we shall now see.

For H_0: $\sigma^* = \sigma_0$ the restricted mle for μ is $\overline{Y}$ and $S_\sigma(\tilde{\theta}) = n\sigma_0^{-3}[S_n^2 - \sigma_0^2]$. Rao's statistic (1) is

$$T_S = \frac{n}{2}\left(\frac{S_n^2}{\sigma_0^2} - 1\right)^2 .$$

Under H_0, $T_S \xrightarrow{d} [(\alpha_4 - 1)/2]\chi_1^2$, where $\alpha_4 = E(Y_1 - \mu)^4 / \sigma^4$. If the data are normal, then α_4=3 and inference based on T_S and χ_1^2 critical values is asymptotically correct. If α_4>3, then such inference can yield very liberal Type I errors.

The generalized score statistic

$$T_{GS} = \frac{n}{2}\left(\frac{S_n^2}{\sigma_0^2} - 1\right)^2\left[\tilde{\alpha}_4 - \frac{2S_n^2}{\sigma_0^2} + 1\right]^{-1}$$

has a kurtosis correction based on $\tilde{\alpha}_4 = n^{-1}\Sigma(Y_i - \overline{Y})^4 / \sigma_0^4$ and is asymptotically χ_1^2 under H_0 for any distribution with finite α_4. If the unrestricted estimator $\hat{\theta} = (\overline{Y}, S_n)$ is used to estimate $\overset{*}{V}_{gS_1}$, then

$$\hat{T}_{GS} = \frac{n}{2}\frac{S_n^2}{\sigma_0^2}\left(\frac{S_n^2}{\sigma_0^2} - 1\right)^2[\hat{\alpha}_4 - 1]^{-1} ,$$

where $\hat{\alpha}_4 = n^{-1}\Sigma(Y_i - \overline{Y})^4 / S_n^4$. As for inference about μ, I prefer $\hat{T}_{GS}$ over T_{GS} especially since there is no computational cost for using the unrestricted estimates.

Example 2. Robust linear regression with homoscedastic errors. Consider the linear model

$$Y_i = x_i^T\beta + e_i , \quad i = 1,\ldots,n ,$$

where $e_i,\ldots e_n$ are assumed iid with some unknown density g. Standard robust regression (Huber, 1981, p. 162) minimizes $\Sigma\rho(Y_i - x_i^T\beta)$ or solves $\Sigma\psi(Y_i - x_i^T\beta)x_i = 0$ where $\psi=\rho'$. Actually, the $Y_i - x_i^T\beta$ are usually rescaled by a parameter σ which is also estimated, but we shall ignore that aspect for simplicity.

For $\beta^T = (\beta_1^T,\beta_2^T)$ suppose that our null hypothesis is H_0: $\beta_2 = \beta_{20}$. We do not need the "*" notation here since $\hat{\beta} \xrightarrow{p} \beta$ except for perhaps the intercept when g is asymmetric. Let $x_i^T = (x_{i1}^T, x_{i2}^T)$ and define $X^T = [x_1|x_2|\cdots|x_n]$. Then the restricted estimates $\tilde{\beta}$ satisfy $\Sigma\psi(Y_i - x_i^T\tilde{\beta})x_{i1} = 0$ and $I_Y = \Sigma\psi'(Y_i - x_i^T\beta)x_ix_i^T$ and $D_Y = \Sigma\psi^2(Y_i - x_i^T\beta)x_ix_i^T$. Under the homoscedastic error assumption, I_Y and D_Y are asymptotically equivalent to $E_g\psi'(e_1)X^TX$ and $E_g\psi^2(e_1)X^TX$, respectively. Thus

$$V_{gS_2} = E_g\psi^2(e_1)\Big[[X^TX]_{22} - [X^TX]_{21}[X^TX]_{11}^{-1}[X^TX]_{12}\Big] .$$

An estimator $\tilde{V}_{gS_2}$ is obtained by replacing $E_g\psi^2(e_1)$ with $(n - b+r)^{-1}\Sigma\psi^2(Y_i - x_i^T\tilde{\beta})$, and the generalized score statistic is then

$$T_{GS} = S_2(\tilde{\beta})^T\tilde{V}_{gS_2}^{-1}S_2(\tilde{\beta}) ,$$

where $S_2(\tilde{\beta}) = \Sigma\psi(Y_i - x_i^T\tilde{\beta})x_{i2}$. T_{GS} was given by Sen (1982, p. 248) and extended to bounded influence regression by Markatov and Hettmansperger (1990). It is quite interesting that $E_g\psi'(e_1)$ cancels out in the definition of V_{gS_2} and therefore does not need to be estimated. Thus T_{GS} has some advantages over the generalized Wald and likelihood ratio tests which require estimates of both $E_g\psi'(e_1)$ and $E_gE\psi^2(e_1)$ and computation of the unrestricted estimates $\hat{\beta}$ (see Schrader and Hettmansperger, 1980, for details on these latter tests).

Example 3. Least squares linear regression with heteroscedastic errors. Using the same notation as in Example 2, we let $\rho(x) = x^2/2$ and obtain $S(\beta) = \Sigma e_i(\beta)x_i$, $I_Y = \Sigma x_i x_i^T = X^TX$, and $D_Y = \Sigma e_i(\beta)^2 x_i x_i^T$ where $e_i(\beta) = Y_i - x_i^T\beta$. Suppose that $e_1,\dots,e_n$ are independent but no longer have the same variance. Then for testing H_0: $\beta_2 = \beta_{20}$ we might let $\tilde{D}_Y = [n/(n-b+r)]\Sigma\tilde{e}_i(\tilde{\beta})^2 x_i x_i^T$ and $T_{GS} = S_2(\tilde{\beta})^T\tilde{V}_{gS_2}^{-1}S_2(\tilde{\beta})$ with

$$\tilde{V}_{gS_2} = \tilde{D}_{Y22} - [X^TX]_{21}\,[X^TX]_{11}^{-1}\tilde{D}_{Y21}^T - \tilde{D}_{Y21}[X^TX]_{11}^{-1}[X^TX]_{21}^T$$
$$+ [X^TX]_{21}[X^TX]_{11}^{-1}\tilde{D}_{Y11}[X^TX]_{11}^{-1}[X^TX]_{21}^T\,.$$

For simple linear regression with x_i's centered so that $\Sigma x_i = 0$, $[X^TX]_{21} = 0$, and $\tilde{\beta}^T = (\bar{Y}\,, 0)$, then $\tilde{V}_{gS_2}$ simplifies to $\tilde{D}_{Y22} = [n/(n-1)]\Sigma(Y_i - \bar{Y})^2x_i^2$. $T_{GS} = [\Sigma Y_i x_i]^2 \big/ \tilde{D}_{Y22}$ may be compared to the usual $F = [\Sigma Y_i x_i]^2 \big/ [(n-2)^{-1}\Sigma\hat{e}_i(\hat{\beta})^2\Sigma x_i^2]$. MacKinnon and White (1985) have investigated related Wald t-statistics based on several versions of $\hat{D}_Y$.

Example 4. Logistic regression. Consider a dose-response situation with k dose levels $x_1,\dots,x_k$ where at the i^th^ dose we observe $\{Y_{ij}, n_{ij}, j=1,\dots,m_i\}$, and we assume that $E(Y_{ij}|n_{ij}) = p_i(\theta) = [1 + \exp(-\beta_1 - \beta_2 x_i)]^{-1}$. If $-Q(\theta)$ is the log likelihood from a presumed independent binomial model, then

$$Q(\theta) = c - \sum_{i=1}^{k}\sum_{j=1}^{m_i} [Y_{ij} \log p_i(\theta) + (n_{ij} - Y_{ij}) \log(1 - p_i(\theta))]$$

$$S(\theta) = \sum_{i=1}^{k}\sum_{j=1}^{m_i} [Y_{ij} - n_{ij}p_i(\theta)] \begin{pmatrix} 1 \\ x_i \end{pmatrix}$$

$$I_Y = \sum_{i=1}^{k}\sum_{j=1}^{m_i} n_{ij}\, p_i(\theta)(1 - p_i(\theta)) \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}$$

$$D_Y = \sum_{i=1}^{k}\sum_{j=1}^{m_i} [Y_{ij} - n_{ij}p_i(\theta)]^2 \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}$$

Under H_0: $\beta_2 = 0$ the restricted maximum likelihood estimator of $p_i(\theta)$ is $\bar{\bar{Y}} = Y_{..}/n_{..}$ where we use the notation $Y_{i.} = \sum_{j=1}^{m_i} Y_{ij}$, $Y_{..} = \sum_{i=1}^{k} \sum_{j=1}^{m_i} Y_{ij}$, etc. The usual score statistic (Cochran-Armitage) is

$$T_S = \frac{[\sum_{i=1}^{k} (Y_{i.} - n_{i.}\bar{\bar{Y}})x_i]^2}{\sum_{i=1}^{k} (x_i - \bar{x})^2 n_{i.} \bar{\bar{Y}} (1 - \bar{\bar{Y}})} ,$$

where $\bar{x} = \sum_{i=1}^{k} n_{i.} x_i / n_{..}$. If the Y_{ij} exhibit cluster effects, then the independent binomial assumption is incorrect and T_S is not asymptotically χ_1^2 under H_0. Typically $Var(Y_{ij}|n_{ij}) > n_{ij}p_i(\theta)(1 - p_i(\theta))$ and the data are said to have extra-binomial variation.

For T_{GS} we replace $p_i(\theta)$ by $\bar{\bar{Y}}$ in I_Y and D_Y and plug into (3) (with subscripts reversed since $\theta^T = (\beta_1, \beta_2)$ and the null is $\beta_2 = 0$) to get

$$T_{GS} = \frac{[\sum_{i=1}^{k} (Y_{i.} - n_{i.}\bar{\bar{Y}})x_i]^2}{\sum_{i=1}^{k} (x_i - \bar{x})^2 \sum_{j=1}^{m_i} (Y_{ij} - n_{ij}\bar{\bar{Y}})^2} .$$

Note that the only difference between T_S and T_{GS} is that the binomial variance estimate $n_{i.}(\overline{\overline{Y}}\,(1-\overline{\overline{Y}}\,)$ has been replaced by the empirical variance estimate $\Sigma_{j=1}^{m_i}(Y_{ij} - n_{ij}\overline{\overline{Y}}\,)^2$.

4.2 Hypotheses via Constraints

Here the null hypothesis is H_0: $h(\theta^*) = 0$, where the r × b matrix $H(\theta) = \partial h(\theta) \big/ \partial\theta^T$ exists and has full row rank r. Of course this case is general enough to contain the previous subsection results with $h(\theta^*) = \theta_1^* - \theta_{10}$. The estimator $\tilde{\theta}$ which minimizes $Q(\theta)$ subject to H_0 satisfies

$$\begin{aligned} S(\tilde{\theta}) - H(\tilde{\theta})^T\tilde{\lambda} &= 0 \\ h(\tilde{\theta}) &= 0\ , \end{aligned} \tag{4}$$

where λ is an r × 1 vector of Lagrange multipliers. The generalized score statistic has been given by White (1982, p. 8) and Gallant (1987, p. 219) as

$$T_{GS} = S(\tilde{\theta})^T\tilde{I}_Y^{-1}\tilde{H}^T(\tilde{H}\tilde{I}_Y^{-1}\tilde{D}_Y\tilde{I}_Y^{-1}\tilde{H}^T)^{-1}\tilde{H}\tilde{I}_Y^{-1}S(\tilde{\theta})\ , \tag{5}$$

where $\tilde{H} = H(\tilde{\theta})$. Since it is not so obvious that this statistic is the correct generalization of (1), I would like to show how it follows from the general asymptotic principle mentioned in Section 1.

First, expand $S(\theta^*)$ about $\tilde{\theta}$,

$$S(\theta^*) = S(\tilde{\theta}) - \tilde{I}_Y(\theta^* - \tilde{\theta}) + R_{n6}\ ,$$

and premultiply by $\tilde{H}\tilde{I}_Y^{-1}$ to get

$$\tilde{H}(\theta^* - \tilde{\theta}) = -\tilde{H}\tilde{I}_Y^{-1}(S(\theta^*) - S(\tilde{\theta})) + R_{n7}\ .$$

Next expand $h(\theta^*)$ about $\tilde{\theta}$,

$$0 = h(\theta^*) = h(\tilde{\theta}) + \tilde{H}(\theta^* - \tilde{\theta}) + R_{n8} .$$

Since $h(\tilde{\theta}) = 0$ from (4), we may put $\tilde{H}(\theta^* - \tilde{\theta}) = -R_{n8}$ into the previous equation and rearrange to get

$$\tilde{H}\tilde{I}_Y^{-1}S(\tilde{\theta}) = \tilde{H}\tilde{I}_Y^{-1}S(\theta^*) + R_{n9} ,$$

where $R_{n9} = -R_{n7} - R_{n8}$. Now premultiply by $\tilde{H}^T(\tilde{H}\tilde{I}_Y^{-1}\tilde{H}^T)^{-1}$ to get

$$(\tilde{I}_Y^{1/2})(\tilde{I}_Y^{-1/2}\tilde{H}^T(\tilde{H}\tilde{I}_Y^{-1}\tilde{H}^T)^{-1}\tilde{H}\tilde{I}_Y^{-1/2})(\tilde{I}_Y^{-1/2}S(\tilde{\theta}))$$

$$= \tilde{H}^T(\tilde{H}\tilde{I}_Y^{-1}\tilde{H}^T)^{-1}\tilde{H}\tilde{I}_Y^{-1}S(\theta^*) + R_{n10} .$$

The middle matrix on the left-hand side is a projection matrix for the column space of $\tilde{I}_Y^{-1/2}\tilde{H}^T$, and $\tilde{I}_Y^{-1/2}S(\tilde{\theta})$ is already in that space by (4). Thus the left-hand side of the last equation becomes just $S(\tilde{\theta})$. Finally, replacing $\tilde{H}$ and $\tilde{I}$ by asymptotic equivalents yields

$$S(\tilde{\theta}) = \overset{*}{H}{}^T(\overset{*}{H}\overset{*}{I}_g^{-1}\overset{*}{H}{}^T)^{-1}\overset{*}{H}\overset{*}{I}_g^{-1}S(\theta^*) + R_{n11} .$$

Since the covariance matrix of $S(\theta^*)$ is $\overset{*}{D}_g$, the asymptotic covariance matrix of $S(\tilde{\theta})$ is

$$\overset{*}{V}_{gS} = \overset{*}{H}{}^T(\overset{*}{H}\overset{*}{I}_g^{-1}\overset{*}{H}{}^T)^{-1}\overset{*}{H}\overset{*}{I}_g^{-1}\overset{*}{D}_g\overset{*}{I}_g^{-1}\overset{*}{H}{}^T(\overset{*}{H}\overset{*}{I}_g^{-1}\overset{*}{H}{}^T)^{-1}\overset{*}{H} ,$$

for which a generalized inverse may be directly verified to be

$$\overset{*}{V}{}_{gS}^{-} = \overset{*}{I}_g^{-1}\overset{*}{H}{}^T(\overset{*}{H}\overset{*}{I}_g^{-1}\overset{*}{D}_g\overset{*}{I}_g^{-1}\overset{*}{H}{}^T)\overset{*}{H}\overset{*}{I}_g^{-1} .$$

Replacing $\overset{*}{I}_g$, $\overset{*}{H}$, and $\overset{*}{D}_g$ by $\tilde{I}_Y$, $\tilde{H} = H(\tilde{\theta})$, and $\tilde{D}_Y$ yields $\tilde{V}_{gS}^{-}$ and the generalized score statistic $T_{GS} = S(\tilde{\theta})^T\tilde{V}_{gS}^{-}S(\tilde{\theta})$ given in (5).

Gallant (1987, p. 231-232) shows that if $\tilde{I}_Y^{-1}\tilde{D}_Y\tilde{I}_Y^{-1}$ is replaced in (5) by $\tilde{I}_Y^{-1}$, then (5) becomes $S(\tilde{\theta})^T\tilde{I}_Y^{-1}S(\tilde{\theta})$ which is Rao's score statistic using observed information $\tilde{I}_Y$ in place of expected information. This nontrivial reduction of (5) to (1) may be shown via (4) and a special matrix equality given by Gallant (1987, p. 241):

$$H^T(HAH^T)^{-1}H = A^{-1} - A^{-1}G(G^TA^{-1}G)^{-1}G^TA^{-1}, \qquad (6)$$

which holds for arbitrary positive definite symmetric $\underset{b\times b}{A}$, $\underset{r\times b}{H}$ of rank r, and $\underset{b\times b-r}{G}$ of rank $b-r$ such that $HG = 0$. These manipulations further illustrate that the path from (1) to (5) is not straightforward.

Example 5. Coefficient of variation. Recall the sampling situation of Example 1 where $Y_1,\ldots,Y_n$ are iid $N(\mu, \sigma^2)$, and consider H_0: $\mu = \sigma$. Three possible constraint specifications of H_0 are $h_1(\theta) = \mu/\sigma - 1$, $h_2(\theta) = \sigma/\mu - 1$, and $h_3(\theta) = \mu - \sigma$; but all three lead to restricted estimates which satisfy $\tilde{\mu}^2 + \bar{Y}\tilde{\mu} - (\bar{Y}^2 + S_n^2) = 0$ and $\tilde{\mu} = \tilde{\sigma}$. Using the fact that $H_1(\tilde{\theta})$, $H_2(\tilde{\theta})$, and $H_3(\tilde{\theta})$ are each proportional to $(1, -1)$, these specifications give the same generalized score statistic (5),

$$T_{GS} = S(\tilde{\theta})^T\tilde{I}_Y^{-1}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}\tilde{I}_Y^{-1}S(\tilde{\theta}) \Big/ [\tilde{V}_{Y11} - 2\tilde{V}_{Y12} + \tilde{V}_{Y22}],$$

where $\tilde{V}_Y = \tilde{I}_Y^{-1}\tilde{D}_Y\tilde{I}_Y^{-1}$ and $S(\theta)$, I_Y, and D_Y are given in Example 1. The invariance to h specification continues to hold if we replace $\tilde{I}_Y$ and $\tilde{D}_Y$ by the unrestricted estimates $\hat{I}_Y$ and $\hat{D}_Y$ which for this problem have a simpler form. T_{GS} is not invariant to h specification, however, if the $H_i(\tilde{\theta})$ are replaced by $H_i(\hat{\theta})$. The Wald statistics based on h_1, h_2, and h_3 are all different but simpler to compute that T_{GS}.

Example 6. Common slope in parallel assay. Consider the logistic regression data of Example 4 but augmented to include a second independent set of data Y_{ij}, $i = k+1,\ldots,2k$, $j = 1,\ldots,m_{2i}$, and $x_{k+i} = x_i$ for $i = 1,\ldots,k$. For example, such data could result from a parallel assay of the relative potency of two compounds. The mean model might be $E(Y_{ij}|n_{ij}) = p_i(\theta)$, where $\theta^T = (\beta_1,\beta_2,\beta_3,\beta_4)$ and

$$p_i(\theta) = [1 + \exp(-\beta_1 - \beta_2 x_i)] \qquad i = 1,\ldots,k$$

$$p_i(\theta) = [1 + \exp(-\beta_3 - \beta_4 x_i)] \qquad i = k+1,\ldots,2k \ .$$

If we assume that the Y_{ij} are independent binomial random variables, then $Q(\theta) = Q_1(\theta) + Q_2(\theta)$, $S(\theta)^T = (S_1(\theta)^T, S_2(\theta)^T)$, and

$$I_Y = \begin{pmatrix} I_{Y1} & 0 \\ 0 & I_{Y2} \end{pmatrix} \qquad D_Y = \begin{pmatrix} D_{Y1} & 0 \\ 0 & D_{Y2} \end{pmatrix} ,$$

where $Q_1(\theta)$, $S_1(\theta)$, I_{Y1}, and D_{Y1} are Q, S, I, and D of Example 4 and $Q_2(\theta)$, $S_2(\theta)$, I_{Y2}, and D_{Y2} have exactly the same form except the summations are $\Sigma_{i=k+1}^{2k} \Sigma_{j=1}^{m_{2i}}$.

Consider the common slope null hypothesis H_0: $\beta_2=\beta_4$ which is important for defining the concept of relative potency in a parallel assay. A simple h is $h(\theta) = \beta_2 - \beta_4$ resulting in $H(\theta) = (0, 1, 0, -1)$. Using the notation $S_1(\theta)^T = (S_{11}(\theta), S_{12}(\theta))$ and $S_2(\theta)^T = (S_{21}(\theta), S_{22}(\theta))$, we find that the generalized score statistic (5) is

$$T_{GS} = \frac{[S_{12}(\hat{\theta})([\tilde{I}_{Y1}^{-1}]_{22} + [\tilde{I}_{Y2}^{-1}]_{22})]^2}{[\tilde{I}_{Y1}^{-1}\tilde{D}_{Y1}\tilde{I}_{Y1}^{-1}]_{22} + [\tilde{I}_{Y2}^{-1}\tilde{D}_{Y2}\tilde{I}_{Y2}^{-1}]_{22}} \ ,$$

where $S_{12}(\theta) = \Sigma_{i=1}^{k}[Y_{i.} - n_{i.}p_i(\tilde{\theta})]x_i$. The algebra for (5) is simplified by noting that $S_{11}(\tilde{\theta}) = S_{21}(\tilde{\theta}) = 0$ and $S_{12}(\tilde{\theta}) + S_{22}(\tilde{\theta}) = 0$.

4.3 Hypotheses via Reparameterization

Some null hypotheses are best described in terms of one set of parameters, whereas the alternative is easiest to describe in terms of another set. Lack-of-fit or goodness-of-fit tests are often of this type. For example, consider binary data in s groups where $\theta_1,...,\theta_s$ are the success probabilities but a proposed model is $\theta_i = g_i(\beta)$ where β is an $b - r$ vector of parameters.

The general description is then $\theta = g(\beta)$, g: $R^{b-r} \rightarrow R^b$, and H_0: $\theta = g(\beta)$ where $G(\beta) = \partial g(\beta) \big/ \partial \beta^T$ is assumed to have full column rank $b - r$. The generalized score statistic in this formulation is

$$T_{GS} = S(\tilde{\theta})^T[\tilde{D}_Y^{-1} - \tilde{D}_Y^{-1}\tilde{I}_Y\tilde{G}(\tilde{G}^T\tilde{I}_Y\tilde{D}_Y^{-1}\tilde{I}_Y\tilde{G})^{-1}\tilde{G}^T\tilde{I}_Y\tilde{D}_Y^{-1}]S(\tilde{\theta}) . \qquad (7)$$

We may obtain (7) from (5) by assuming H_0 has an equivalent constraint h formulation such that $HG = 0$ and use (6) with $A = \tilde{I}_Y^{-1}\tilde{D}_Y\tilde{I}_Y^{-1}$. Another route is to use the direct asymptotic expansion of S as follows.

Since $\tilde{\theta} = g(\tilde{\beta})$ minimizes $Q(g(\beta))$, $\tilde{\theta}$ satisfies $G(\tilde{\beta})^T S(g(\tilde{\beta})) \equiv \tilde{G}^T\tilde{S} = 0$. Now expand $S(\tilde{\theta})$ about $\theta^* = g(\beta^*)$ using the chain rule $\partial S(g(\beta)) \big/ \partial\beta^T = (\partial S \big/ \partial g^T)(\partial g \big/ \partial \beta^T)$,

$$S(\tilde{\theta}) = S(\theta^*) - \overset{*}{I}_Y\overset{*}{G}(\tilde{\beta} - \beta^*) + R_{n12} . \qquad (8)$$

Premultiplying by $\tilde{G}^T$ gives zero on the left-hand side and rearranging leads to

$$\tilde{\beta} - \beta^* = [\tilde{G}^T\overset{*}{I}_Y\overset{*}{G}]^{-1}\tilde{G}^T S(\theta^*) + R_{n13} .$$

Finally replace $\tilde{G}$ by $\overset{*}{G}$ and $\overset{*}{I}_Y$ by $\overset{*}{I}_g$ and substitute for $\tilde{\beta} - \beta^*$ in (8) to get

$$\begin{aligned} S(\tilde{\theta}) &= [I_s - \overset{*}{I}_g\overset{*}{G}(\overset{*}{G}{}^T\overset{*}{I}_g\overset{*}{G})^{-1}]S(\theta^*) + R_{n14} \\ &\equiv \overset{*}{C}S(\theta^*) + R_{n14} . \end{aligned}$$

Thus, the asymptotic covariance matrix of $S(\tilde{\theta})$ is $\overset{*}{\tilde{C}}\overset{*}{\tilde{D}}_g\overset{*}{\tilde{C}}^T$ and $T_{GS} = S(\tilde{\theta})^T(\tilde{C}\tilde{D}_Y\tilde{C}^T)^- S(\tilde{\theta})$. The form (7) follows by verifying that the inner matrix in (7) is a generalized inverse of $\tilde{C}\tilde{D}_Y\tilde{C}^T$.

Example 7. Lack-of-fit in binomial regression. Once again reconsider the data of Example 4 but now set $b = k$ and $\theta_i = E(Y_{ij}|n_{ij})$, $i = 1,...,b$. We would like to test the adequacy of $\theta_i = F(x_i^T\beta)$, where F is the logistic or normal or some other distribution function, and we now allow x_i and β to be $p \times 1$ vectors. If we use the binomial likelihood to estimate β, then Rao's statistic (1) is the usual Pearson chi-squared statistic

$$T_S = \sum_{i=1}^{b} n_{i.}[\hat{\theta}_i - \tilde{F}_i]^2 \Big/ \tilde{F}_i(1 - \tilde{F}_i) ,$$

where $\tilde{F}_i = F(x_i^T\tilde{\beta})$ and $\hat{\theta}_i = Y_{i.}/n_{i.}$. T_S can be significant because of mean misspecification, $\theta_i \neq F(x_i^T\beta)$, or because the binomial likelihood is wrong. In contrast, the generalized statistic T_{GS} will tend to reject only for mean misspecification and otherwise have an approximate χ^2_{b-p} distribution.

To compute T_{GS} note that

$$S(\tilde{\theta}) = n_{i.}[\hat{\theta}_i - \tilde{F}_i]/\tilde{F}_i(1 - \tilde{F}_i) ,$$

$$\tilde{I}_Y = \text{Diag}\left[n_{i.}\left(\frac{\hat{\theta}_i}{\tilde{F}_i^2} + \frac{(1-\hat{\theta}_i)}{(1-\tilde{F}_i)^2}\right), \quad i = 1,...,b\right],$$

$$\tilde{D}_Y = \text{Diag}\left[\sum_{j=1}^{m_i}(Y_{ij} - n_{ij}\tilde{F}_i)^2 \Big/ [\tilde{F}_i(1-\tilde{F}_i)]^2, \quad i = 1,...,b\right],$$

$$\tilde{G} = \tilde{F}'X \text{ with } \tilde{F}' = \text{Diag}\left[\tilde{F}_i' = \frac{dF(z)}{dz}\bigg|_{z=x_i^T\tilde{\beta}}, \quad i = 1,...,b\right].$$

Then we have from (7)

$$T_{GS} = S(\tilde{\theta})^T\tilde{D}_Y^{-1}S(\tilde{\theta}) \;-\; S(\tilde{\theta})^T\tilde{E}_1X(X^T\tilde{E}_2X)^{-1}X^T\tilde{E}_1S(\tilde{\theta})\ ,$$

where $\tilde{E}_1 = \tilde{D}_Y^{-1}\tilde{I}_Y\tilde{F}'$ and $\tilde{E}_2 = \tilde{F}'\tilde{I}_Y\tilde{D}_Y^{-1}\tilde{I}_Y\tilde{F}'$ are diagonal and $X = [x_1|x_2|\cdots|x_b]$. If the mean specification and binomial likelihood are both correct, then the first term of T_{GS} is asymptotically equivalent to the Pearson chi-square and the second term converges in probability to zero as $\min[m_1,...,m_b] \to \infty$ and b is fixed.

4.4. Invariance under Reparametrization

For a one to one reparametrization $\theta = \tau(\beta)$ with nonsingular Jacobian $J_\beta = \partial\tau(\beta)\big/\partial\beta^T$, the invariance of Rao's score statistic (1) follows from the relations $\tilde{\theta} = \tau(\tilde{\beta})$, $S(\beta) = J_\beta^T S(\tau)$, and $I_f(\beta) = J_\beta^T I_f(\tau)J_\beta$. This invariance does not extend to (1) with I_f replaced by I_Y or to T_{GS} which must use I_Y and not I_f to estimate I_g. The lack of invariance can be seen by noting that

$$I_Y(\beta) = \frac{\partial}{\partial\beta^T}\left(J_\beta^T S(\tau(\beta))\right) = L_\beta + J_\beta^T I_Y(\tau)J_\beta\ ,$$

where L_β has the j^th^ row $[L_\beta]_j = \sum_{i=1}^{b}[S(\tau)]_i\partial[J_\beta^T]_{ij}\big/\partial\beta^T$. L_β vanishes at $\beta = \hat{\beta}$ but not at $\beta = \tilde{\beta}$. Since $D_Y(\beta) = J_\beta^T D_Y(\tau)J_\beta$ and $I_Y(\hat{\beta}) = J_{\hat{\beta}}^T I_Y(\hat{\tau})J_{\hat{\beta}}$, the generalized score statistic $\hat{T}_{GS}$ which uses unrestricted estimates in the covariance estimate is invariant to reparametrization whenever $H(\tilde{\tau}) = H(\hat{\tau})$, e.g., when $h(\theta)$ is a linear function of θ.

There are two situations where T_{GS} is invariant to reparametrization. The first situation is when $\tau(\beta)$ is a linear function of β so that $L_\beta = 0$. The second situation is when H_0 completely specifies θ, say H_0: $\theta=\theta_0$, and $T_{GS} = S(\theta_0)^T D_Y(\theta_0)^{-1}S(\theta_0)$. Here the invariance follows from $D_Y(\beta) = J_\beta^T D_Y(\tau)J_\beta$ and $S(\beta) = J_\beta^T S(\tau)$ and applies to $\hat{T}_{GS}$ as well.

5. Summary

The various versions of the generalized score statistic arise naturally from Taylor expansion of the defining equations. Although these statistics are not as simple in appeariance as Rao's original statistic (1), their wide applicability and asymptotic Type I error robustness to model misspecification make them attractive for general use.

Acknowledgement. The author wishes to thank Ron Gallant and Len Stefanski for helpful discussions during the preparation of this manuscript.

REFERENCES

Aitchison, J., and Silvey, S. D. (1958), "Maximum Likelihood Estimation of Parameters Subject to Constraints," *The Annals of Mathematical Statistics*, 19, 813-828.

Breslow, N. (1990), "Tests of Hypotheses in Overdispersed Poisson Regression and Other Quasi-Likelihood Models," *Journal of the American Statistical Association*, 85, 565-571.

Breusch, T. S., and Pagan, A. R. (1980), "The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics, " *Review of Economic Studies*, 47, 239-254.

Burguete, J. F., Gallant, A. R., and Souza, G. (1982), "On Unification of the Asymptotic Theory of Nonlinear Econometric Models," *Economic Reviews*, 1, 151-190.

Buse, A. (1982), "The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note," *The American Statistician*, 36, 153-157.

Engle, R. R. (1984), "Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics," in *Handbook of Econometrics*, Vol. II, eds. Z. Griliches and M. Intriligator, Amsterdam: North Holland.

Foutz, R. V., and Srivastava, R. C. (1977), "The Performance of the Likelihood Ratio Test when the Model is Incorrect," *Annals of Statistics*, 1183-1194.

Gallant, A. R. (1987), *Nonlinear Statistical Models*, New York: Wiley.

Hosking, J. R. M. (1983), "Lagrange Multiplier Test," *Encyclopedia of Statistical Sciences*, 4, 456-459.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, A. S. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: Wiley.

Huber, P. (1967), "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions," *Proceedings of the 5th Berkeley Symposium*, 1, 221-233.

Huber, P. J. (1981), *Robust Statistics*, New York: John Wiley.

Kent, J. T. (1982), "Robust Properties of Likelihood Ratio Tests," *Biometrika*, 69, 19-27.

Lin, D. Y., and Wei, L. J. (1989), "The Robust Inference for the Cox Proportional Hazards Model," *Journal of the American Statistical Association*, 84, 1074-1078.

MacKinnon, J. G., and White, H. (1985), "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics*, 29, 305-325.

Markatov, M., and Hettmansperger, T. P. (1990), "Robust Bounded-Influence Tests in Linear Models," *Journal of the American Statistical Association*, 85, 187-190.

Rao, C. R. (1948), "Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Applications to Problems of Estimation," *Proceedings of the Cambridge Philosphical Society*, 44, 50-57.

Schrader, R. M., and Hettmansperger, T. P. (1980), "Robust Analysis of Variance Based Upon a Likelihood Ratio Criterion," *Biometrika*, 67, 93-101.

Sen, P. K. (1982), "On M Tests in Linear Models," *Biometrika*, 69, 245-248.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: John Wiley.

Tarone, R. E. (1988), "Score Statistics," *Encyclopedia of Statistical Sciences*, 8, 304-308.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1-26.