ON THE DISTRIBUTION FOR STANDARDIZED RESIDUALS FROM LINEAR REGRESSION

BY

G.H. FICK and R.R. DAVIDSON

University of Calgary and University of Victoria

SUMMARY

Using a linear model that assumes only a portion of the errors to be independent, a general distribution for the standardized residuals is derived. From this distribution, the sampling distribution for the maximum likelihood estimator of a variance inflation parameter is obtained. Inferences based on this estimator are discussed.

Keywords: inference, likelihood, regression, residuals

Short Title For Running Heads: Standardized Residuals

1. INTRODUCTION

For the analysis of data under a linear regression model, the standardized residuals are used to assess many of the usual assumptions including the assumptions of independent and identically distributed errors. In particular, if it is possible that some of the error terms have variance potentially larger than the others, it would be appropriate to examine this inflation of variance.

Cook, Holschuh and Weisberg(1982) examine a model that allows for variance inflation using maximum likelihood estimation. In their article, they conclude that the small sample distribution theory appears to be intractible. In this article, we present a general distribution for standardized residuals that requires only a portion of the errors to be independent. While the distribution for standardized residuals under independent and identically distributed errors has been known for some time, this general distribution appears to be new. We include two derivations of this distribution. One method of derivation uses likelihood modulation, while the other method follows from a scaled multivariate Student distribution.

Using these results we are able to present the exact small sample distribution theory for the maximum likelihood estimator of a variance inflation parameter when we know which component of error is subject to potential variance inflation.

2. PRELIMINARIES AND NOTATION

Let us write a linear model in the form

$$\mathbf{y} = X\beta + \sigma\mathbf{z}$$

where $\mathbf{y}$ is a response vector of n observations obtained at input levels recorded in the r linearly independent columns of X. We assume the error vector

$$\mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} \sim N_n(0,\Omega)$$

where Ω = diag(T,I) is partitioned according to $\mathbf{z}$, i.e., we are allowing $\mathbf{z}_1$, the first p components of $\mathbf{z}$ to have arbitrary positive definite covariance matrix T. The analysis to be presented is independent of the choice of basis L(X), the linear space spanned by the columns of X. Let V be a n x r column orthonormal matrix with L(V) = L(X). Then V can be augmented by a matrix N so that [V N] is an n x n orthogonal matrix. Since I-VV′ = NN′ the standardized residuals $\mathbf{d}$ can be written

$$\mathbf{d} = \frac{(I-VV')\mathbf{y}}{||(I-VV')\mathbf{y}||} = \frac{(I-VV')\mathbf{z}}{||(I-VV')\mathbf{z}||} = \frac{N\mathbf{t}}{||\mathbf{t}||}$$

where $\mathbf{t} = N'\mathbf{z}$, i.e. $\mathbf{d}$ depends on $\mathbf{z}$ only through $\mathbf{t}$. It is important to note that the distribution of $\mathbf{d}$ depends only on the distribution for $\mathbf{z}$ and not on the parameters β or σ. Therefore, inference for T can be based on $\mathbf{d}$ or on equivalent statistics in $L^{\perp}(V)$,

the orthogonal complement to L(V). Next partition $V=\begin{bmatrix}V_1\\V_2\end{bmatrix}$ according to $\mathbf{z}=\begin{bmatrix}\mathbf{z}_1\\\mathbf{z}_2\end{bmatrix}$ and choose N so that

$$N=\begin{bmatrix}N_1 & 0\\N_2 & N_3\end{bmatrix}$$

where N_1 is p x q with q = rank $(I-V_1V_1')$ .Then $\mathbf{t} = N'\mathbf{z}$ can be written

$$\mathbf{t} = \begin{bmatrix}\mathbf{t}_1\\\mathbf{t}_2\end{bmatrix} = \begin{bmatrix}N_1'\mathbf{z}_1+N_2'\mathbf{z}_2\\N_3'\mathbf{z}_2\end{bmatrix} \sim N_{n-r}(0, \Omega_t) \qquad (1)$$

where Ω_t = diag{Λ, I} with $\Lambda=I+N_1'(T-I)N_1$ partitioned according to **t**. Finally, the standardized residuals $\mathbf{d} = \frac{N\mathbf{t}}{||\mathbf{t}||}$ can be partitioned according to **z**, namely

$$\mathbf{d} = \begin{bmatrix}\mathbf{d}_1\\\mathbf{d}_2\end{bmatrix} = \frac{1}{||\mathbf{t}||}\begin{bmatrix}N_1\mathbf{t}_1\\N_2\mathbf{t}_1+N_3\mathbf{t}_2\end{bmatrix}$$

3. DISTRIBUTION THEORY

The objective of this section is to present a derivation of the distribution necessary for inference for T. When $\mathbf{z} \sim N_n(0,\Sigma)$ with Σ positive definite, the distribution for **d** is

$$h_\Sigma(\mathbf{d}) = \Gamma\left[\frac{n-r}{2}\right]/2\pi^{\frac{n-r}{2}}\,|V'\Sigma^{-1}V|^{-1/2}|\Sigma|^{-1/2}\left[\mathbf{d}'(\Sigma^{-1}-\Sigma^{-1}V(V'\Sigma^{-1}V)^{-1}V'\Sigma^{-1})\mathbf{d}\right]^{-\frac{n-r}{2}} \qquad (2)$$

for $\mathbf{d}'\mathbf{d} = 1$ and $V'\mathbf{d} = 0$.

i.e. **d** on the unit sphere in $L^{\perp}(V)$.

These distributions belong to the class of projected normal distributions (Fraser (1979)). See Fick (1984) for a derivation of the form given here. Notice that when Σ = I, $h_I(\mathbf{d})$ is a uniform distribution on the sphere in $L^{\perp}(V)$. This general model for **d** will not admit any reduction in dimensionality via sufficiency for general Σ. However, when Σ = diag{T,I} it can be shown that the likelihood for T depends on **d** only through $\mathbf{d}_1$ and hence $\mathbf{d}_1$ is sufficient for T.

When $A=I-V_1V_1'$ is nonsingular, i.e. p = q, the likelihood for T can be written

$$L(\mathbf{d}_1|T) \propto |I-K|^{1/2}\left[1-\mathbf{d}_1'K\mathbf{A}^{-1}\mathbf{d}_1\right]^{-\frac{n-r}{2}} \qquad (3)$$

where $K=I-(I+(T-I)A)^{-1}$. An illustration using this likelihood function is given in section 5.

Since $\mathbf{d}_1$ is sufficient for T interest now centres on its distribution. When p = q, $\mathbf{d}_1$ has a distribution on the interior of the p-dimensional ellipsoid $\mathbf{d}'_1A^{-1}\mathbf{d}_1 = 1$. The density for $\mathbf{d}_1$, when T = I, has been given by Ellenburg (1973)

$$g_I(\mathbf{d}_1) = \frac{\Gamma\left[\frac{n-r}{2}\right]}{\Gamma\left[\frac{n-r-p}{2}\right]\pi^{p/2}}|A|^{-1/2}(1-\mathbf{d}'_1A^{-1}\mathbf{d}_1)^{\frac{n-r-p}{2}-1} \tag{4}$$

for $\mathbf{d}'_1A^{-1}\mathbf{d}_1 < 1$.

Note that $g_I(\mathbf{d}_1)$ is constant on ellipsoids $\mathbf{d}'_1A^{-1}\mathbf{d}_1 = k$. Doornbos and Prins (1958) give this result for p=2. Butler (1984) gives a number of other references that present $g_I(\mathbf{d}_1)$. Using the representative likelihood function $L^*(\mathbf{d} \mid T)$ chosen to have $L^*(\mathbf{d} \mid I) = 1$ we can write the density $g_T(\mathbf{d}_1)$ for $\mathbf{d}_1$

$$g_T(\mathbf{d}_1) = g_I(\mathbf{d}_1)L^*(\mathbf{d}_1 \mid T) \tag{5}$$

Note that the density $g_T(\mathbf{d}_1)$ can be viewed as the "null" density $g_I(\mathbf{d}_1)$ modulated by the likelihood. This device was used by Watson (1956) and Watson and Williams (1956) in discussions of distributions on spheres. A discussion of likelihood modulation and its application to other settings can be found in Fraser (1968, 1979). The density for $\mathbf{d}_1$ can now be displayed by using (3) and (4) in (5).

$$g_T(\mathbf{d}_1) = \frac{\Gamma\left[\frac{n-r}{2}\right]}{\Gamma\left[\frac{n-r-p}{2}\right]\pi^{p/2}}|A|^{-1/2}(1-\mathbf{d}'_1A^{-1}\mathbf{d}_1)^{\frac{n-r-p}{2}-1}(1-\mathbf{d}'_1TA^{-1}\mathbf{d}_1)^{\frac{n-r}{2}} \tag{6}$$

for $\mathbf{d}'_1A^{-1}\mathbf{d}_1 < 1$.

When A is singular so that $p > q = \text{rank}(I - V_1V'_1)$, the p-dimensional vector $\mathbf{d}_1$ is a function of a q-dimensional vector, specifically $\mathbf{d}_1 = N_1\mathbf{r}$ where $\mathbf{r} = \mathbf{t}_1/\|\mathbf{t}\|$.

For this case, we now give alternate derivation of the density g_T. If we define $\mathbf{u} = \mathbf{t}_1/\|\mathbf{t}_2\|$, it follows from the distribution of $\mathbf{t}$ given in (1) that $\mathbf{u}$ has density

$$g_T(\mathbf{u}) = \frac{\Gamma\left[\frac{n-r}{2}\right]}{\Gamma\left[\frac{n-r-q}{2}\right]\pi^{q/2}}|\Lambda|^{-1/2}(1+\mathbf{u}'\Lambda^{-1}\mathbf{u})^{-\frac{n-r}{2}} \tag{7}$$

This density is a scaled q-dimensional Student distribution with n-r-q degrees of

freedom. Since $\mathbf{r} = \mathbf{u}/(1 + \mathbf{u}'\mathbf{u})^{1/2}$ and the transformation from $\mathbf{u}$ to $\mathbf{r}$ has Jacobian $(1 - \mathbf{r}'\mathbf{r})^{-\frac{q}{2}-1}$ one obtains the density for $\mathbf{r}$

$$g_T(\mathbf{r}) = \frac{\Gamma\left[\frac{n-r}{2}\right]}{\Gamma\left[\frac{n-r-q}{2}\right]\pi^{q/2}} |\Lambda|^{-1/2}(1-\mathbf{r}'\mathbf{r})^{\frac{n-r-q}{2}-1}(1 + \mathbf{r}'(I - \Lambda^{-1})\mathbf{r})^{-\frac{n-r}{2}} \qquad (8)$$

for $\mathbf{r}'\mathbf{r} < 1$.

Note that $\mathbf{r}$ has density on the interior of the q-dimensional sphere $\mathbf{r}'\mathbf{r} = 1$. Inference for T can be based on $\mathbf{d}_1$, $\mathbf{r}$ or $\mathbf{u}$. Notice that, when specialized, the densities (6), (7) or (8) induce the same likelihood (3).

4. INFERENCE FOR VARIANCE INFLATION

We now specialize the general results from section 3 to the case of $p = 1$. In this case, the matrix T becomes a scalar τ representing the variance of z_1. This special case has been called a 'variance inflation' model by Cook, Holschuh and Weisberg (1982) when one assumes that the first (or chosen) component of error (z_1) is subject to potential variance inflation. (real variance inflation would mean that $\tau > 1$)

In the case $p = 1$, we can take the distribution theory a little further and obtain an exact test and confidence intervals.

The likelihood function given in line (3) becomes

$$L^*(d_1 \mid \tau) = (1 - \kappa)^{1/2} (1 - \kappa d_1^2/a)^{\frac{n-r}{2}}$$

where $a = 1 - V_1 V_1'$ (a scalar between 0 and 1) and $\kappa = 1-(1+a(\tau-1))^{-1}$. Notice that d_1^2 is, in fact, **minimal** sufficient.

The likelihood is maximized at

$$t = \frac{(n-r)d_1^2/a - 1}{a(1 - d_1^2/a)} + 1$$

The distribution for t can be obtained directly from (6) or (7) as

$$\frac{1/a + (t-1)}{1/a + (\tau-1)} \sim F_{1,\, n-r-1}$$

We now briefly outline some properties that follow from the distribution theory.

When considering the variance inflation model, the parameter space for τ might be restricted to $[\ 1, \infty\ )$. It should be noted that t as defined can take on values that are less than one (in fact, t can take on negative values down to $-(\ 1/a - 1\ ) < 0\ )$. Further, it can be shown that t is not unbiased. In fact it can be shown that

$$\frac{(n-r-3)t}{n-r-1} + \frac{2(1 - 1/a)}{n-r-1}$$

is an unbiased estimator for τ with variance

$$\frac{2(n-r-2)}{(n-r-5)(1/a + (\tau-1))^2}.$$

From the information inequality, a lower bound on the variance of unbiased estimators can be shown to be

$$\frac{2(n-r+2)}{(n-r-1)(1/a + (\tau-1))^2}.$$

Since the model for t is not of exponential type, we know that this bound cannot be attained uniformly in τ for any unbiased estimator (Fraser (1976 pp 344). But notice that this lower bound is the variance of the unbiased version of t based on n+4 observations.

A one sided 1 - α confidence interval for τ would be

$$\tau > -(1/a - 1) + \frac{(1/a + (t-1))}{F_{1-\alpha}}$$

where $F_{1-\alpha}$ is the 1 - α percentile from $F_{1,n-r-1}$.

The uniformly most powerful size α test for $H_0{:}\tau = 1$ versus $H_1{:}\tau > 1$ has the form 'Reject H_0' if

$$t > 1 + (F_{1-\alpha} - 1)/a$$

with power

$$1-\beta = P\left[F > \frac{F_{1-\alpha}}{1 + a(\tau - 1)}\right]$$

5. AN EXAMPLE WITH LIKELIHOOD INFERENCE

We now illustrate the results with the location - scale model in which:

$$X = 1 \text{ and } V = 1/\text{sqrt}n$$

First, we consider an analysis of the likelihood with p = 1. In Figure 1, we have graphed relative likelihood functions for n = 100 and t = 1, 3, 10, and 20. Both the horizontal and vertical scales are logarithmic. The dotted horizontal lines can be used to indicate 10% and 50% relative likelihood intervals. Typically, one would focus on the smaller likelihood limit. When this limit is greater than one, $\tau = 1$ is not plausible and variance inflation is then a possibility. Notice that the curvature of the log -

likelihood at the maximum seems to be approximately constant. In fact, it can be shown that

$$\frac{\partial^2 \log L}{\partial \log \tau^2}\Big|_{\tau = t} = -\frac{1}{2}\,\frac{n-r-1}{n-r}\,\frac{t^2}{(1/a + (t-1))^2}$$

INSERT FIGURE 1 HERE

Next, we offer an example of a likelihood analysis with p = 2. The famous data on growth rates of plants from Darwin (1878) (see also Andrews and Herzberg (1985)) contains 2 observations where variance inflation could be considered.

Table 1. Differences in eighths of an inch between cross- and self-fertilized plants of the same pair

49	23	56
-67	28	24
8	41	75
16	14	60
6	29	-48

The standardized residuals are $\mathbf{d}' =$

(0.1987 0.0146 0.2483 -0.6226 0.0500 0.0217 -0.0916 0.1421 0.3828 -0.0349 -0.0491 0.2766 -0.1057 0.0571 -0.4881)

For each pair of residuals, a likelihood for τ was determined along with 50% and 10% relative likelihood intervals. Figure 2 displays a plot of the lower 10% limit versus the lower 50% limit for each pair of standardized residuals. The 'famous' pair is flagged as are a few other pairs