NON-PARAMETRIC c-SAMPLE TESTS WITH REGRESSION

J. N. ADICHIE

(Received Sept. 4, 1972; revised Nov. 13, 1974)

Abstract

For testing that several regression lines are concurrent, a class of rank score tests is proposed. This class includes the usual Wilcoxon and normal scores type of tests. The performance of the proposed tests is shown to be the same as that of rank score tests in the ordinary c-sample problem.

1. Introduction and summary

For the c-sample model $Y_{ij}=\alpha_i+Z_{ij}$ $j=1,\cdots,n_i$, $i=1,\cdots,c$, where Z_{ij} are independent random variables, optimum rank score tests for the hypothesis $\alpha_i=\alpha$ for all i, have been given by Puri [7] and Kruskal and Wallis [5] among others. In this paper, we consider the c-sample problem in the regression model, and give rank score tests. The basic model considered in this paper is used quite often in practice. One may for example be interested in the effects $(\alpha_i+\beta x_{ij})$ of c treatments on experimental animals, where x_{ij} is the weight of the $(i-j)$th animal prior to the treatment, and the unknown factor β, which is assumed to be independent of the treatments is not of interest.

2. Assumptions and notations

Let Y_{ij} $(j=1,\cdots,n_i;\ i=1,\cdots,c)$ be a sequence of independent random variables with continuous distribution functions F_{ij} given by

$$\text{(2.1)}\qquad \mathrm{P}\,[Y_{ij}\leqq y]=F_{ij}(y)=F(y-\alpha_i-\beta x_{ij})\qquad i=1,\cdots,c\,.$$

Here x_{ij} are known "regression constants," α_i's are the quantities of interest while β is a nuisance parameter. Our problem is to test the hypothesis:

$$\text{(2.2)}\qquad \alpha_i=\alpha\ \text{(unknown)}\,,\qquad \text{for all } i\,,$$

against the set of alternatives that $\alpha_1,\cdots,\alpha_c$ are not all equal.

Let us write

$$C_n^2=\sum_i \sum_j x_{ij}^2\ ; \qquad N=\sum_i n_i \tag{2.3}$$

$$\lambda_{ni}=(n_i/N)\ ; \quad \gamma_{ni}=(n_i/C_n^2) \qquad i=1,\cdots,c\,. \tag{2.4}$$

Observe that the quantities above depend on an integer n which is assumed to tend to infinity. Furthermore each of C_n^2, n_i and N tends to infinity with n in such a way that

$$\begin{aligned} &\gamma_{ni}\to\gamma_i\ ; \qquad 0<\gamma_0\leqq\gamma_1,\cdots,\gamma_c(1-\gamma_0)<1 \\ &\lambda_{ni}\to\lambda_i\ ; \qquad 0<\lambda_0\leqq\lambda_1,\cdots,\lambda_c(1-\lambda_0)<1 \end{aligned} \tag{2.5}$$

where $\lambda_0, \gamma_0\leqq 1/c$. Furthermore, the regression constants are assumed to satisfy the Noether condition,

$$\lim\,[\max_{ij}|x_{ij}|/C_n]=0\ , \tag{2.6}$$

and the boundedness condition

$$\sup_n\,[N^{-1}C_n^2]<\infty\ . \tag{2.7}$$

To simplify the notation, we shall assume throughout this paper that

$$\sum_j x_{ij}=0\ , \qquad i=1,\cdots,c\,. \tag{2.8}$$

Observe that assumption (2.8) imply that the group averages of x's are all equal (to zero).

Let $\psi(u)$ $(0<u<1)$ be a smooth non decreasing function with bounded second derivative. Also let the scores generated by ψ be defined by

$$a_n(p)=\psi_n(p)=\psi(p/(N+1)) \qquad 1\leqq p\leqq N\,, \tag{2.9}$$

and let R_{ij} be the ranks of Y_{ij} in the combined ranking of all the N observations. We shall need an estimate of β in (2.1), for that purpose, let

$$S_n(Y)=\sum_i \sum_j x_{ij}a_n(R_{ij}) \tag{2.10}$$

and define as in [1], the estimate of β based on (2.10) as follows:

$$\beta_n^*=\sup\{b:\ S_n(Y-bx)>0\}\ ; \qquad \beta_n^{**}=\inf\{b:\ S_n(Y-bx)<0\} \tag{2.11}$$

$$\hat\beta_n=\frac{1}{2}(\beta_n^*+\beta_n^{**}) \tag{2.12}$$

where $S_n(Y-bx)$ is the statistics (2.10) when the observations Y_{ij} are replaced by $(Y_{ij}-bx_{ij})$.

Now let

$$d^{(i)}_{sj}=\begin{cases}1 & s=i,\ j=1,\cdots,n_i\\ 0 & s\neq i,\ j=1,\cdots,n_s\end{cases} \tag{2.13}$$

so that

$$\bar{d}^{(i)}=N^{-1}\sum_s\sum_j d^{(i)}_{sj}=\lambda_{ni}\ .$$

For each $i=1,\cdots,c$, write

$$\hat{T}_{ni}=T_{ni}(\hat{Y})=\sum_s\sum_j(d^{(i)}_{sj}-\bar{d}^{(i)})a_n(\hat{R}_{sj}) \tag{2.14}$$

$$T^0_{ni}=T_{ni}(Y^0)=\sum_s\sum_j(d^{(i)}_{sj}-\bar{d}^{(i)})a_n(R^0_{sj}) \tag{2.15}$$

$$\hat{L}_n=\sum_i n_i^{-1}(\hat{T}_{ni}/A)^2=\sum_i W^2_{ni} \tag{2.16}$$

where $\hat{R}_{sj}$ and R^0_{sj} are the ranks of $\hat{Y}_{sj}=(Y_{sj}-\hat{\beta}x_{sj})$ and $Y^0_{sj}=(Y_{sj}-\beta_{sj})$ respectively, and

$$A^2=\int\phi^2(u)du-\left(\int\phi(u)du\right)^2\ . \tag{2.17}$$

For suitably defined ϕ, we propose $\hat{L}_n$ given in (2.16) as the test statistics for the hypothesis (2.2). Observe that $\hat{L}_n$ is easy to compute since it is based essentially on the ranks $\hat{R}_{ij}$ of the group residuals $\hat{Y}_{ij}$. For example $\hat{T}_{ni}=\sum_j a_n(\hat{R}_{ij})-\lambda_{ni}$ (constant), and if we choose the Wilcoxon score, $\phi(u)=u$, then

$$\hat{T}_{ni}=\sum_j\{\hat{R}_{ij}/(N+1)\}-n_i/2\ .$$

3. Asymptotic distribution of $\hat{L}_n$

The following theorem gives the asymptotic distribution of the $\hat{L}_n$ statistics under the hypothesis (2.2).

THEOREM 3.1. *Consider model* (2.1), *and assume that ϕ has a bounded second derivative, and that*

$$\sup_y\phi'_y(F(y))=\sup_y d/dy(\phi(F(y))) \tag{3.1}$$

is also bounded. Then, under the assumptions of Section 2, *the statistics $\hat{L}_n$ given in* (2.16) *has asymptotically under the hypothesis a chi-square distribution with* $(c-1)$ *degrees of freedom.*

PROOF. Follows as a special case of the proof of Theorem 3.2 below.

In view of Theorem 3.1, an asymptotically level ε test rejects the hypothesis (2.2), if $\hat{L}_n$ is greater than the upper $100\varepsilon\%$ point of the chi-square distribution with $(c-1)$ degrees of freedom. To study the asymptotic power of the proposed test, we shall consider in the next theorem, the limiting distribution of $\hat{L}_n$ under a sequence of alternatives given by,

$$\alpha_i=\alpha+\theta_i C_n^{-1} \qquad i=1,\cdots,c \tag{3.2}$$

where θ_i $i=1,\cdots,c$ are all real and finite.

THEOREM 3.2. *Let ψ satisfy the conditions in Theorem 3.1, then under (2.5), (2.6), (2.7), (2.8) and (3.2), $\hat{L}_n$ has asymptotically a non-central chi-square distribution with $(c-1)$ degrees of freedom and non-centrality parameter,*

$$\Delta_L(F,\psi)=\sum_i \gamma_i(\theta_i-\bar{\theta})^2\left(\int \psi_y'(F(y))dF(y)\right)^2\Big/A^2 \tag{3.3}$$

where $\bar{\theta}=\sum_i \lambda_i\theta_i$, while γ_i and λ_i are as defined in (2.5).

The proof of this theorem rests on the following three lemmas.

LEMMA 3.1. *Under the sequence of alternatives in (3.2)*

$$C_n(\hat{\beta}_n-\beta) \tag{3.4}$$

is bounded in probability as $n\to\infty$.

PROOF. Similar to the proof of Lemma 3.1 of [8] and is therefore omitted.

LEMMA 3.2. *Under the conditions of Theorem 3.2,*

$$[(\hat{T}_{ni}-T_{ni}^0)/C_n]\to 0 \tag{3.5}$$

in P_n probability, $i=1,\cdots,c$.

PROOF. Without loss of generality, we may for the proof of this lemma, take $\alpha_i=\beta=0$, so that $Y_{sj}^0=Y_{sj}$. Now write

$$Y_{ij}^*=Y_{ij}-(bx_{ij}/C_n)\ , \qquad |b|<k \text{ (a generic constant)}. \tag{3.6}$$

We then have that under (3.2), $F_{ij}(y)=F(y)$, and

$$F_{ij}^*(y)=F(y-(bx_{ij}/C_n))$$

where F_{ij}^* is the distribution function of Y_{ij}^*. Because of Lemma 3.1, it is sufficient to show that

$$\lim \mathrm{E}_n[\{T_{ni}(Y^*)-T_{ni}(Y)\}/C_n]^2=0 \tag{3.7}$$

uniformly in $|b|<k$; where E_n denotes expectation under (3.2).

Observe that T_{ni} are linear rank statistics, and from Hájek's work, see Theorem 4.2 of [3], it follows that for each i

$$\lim\,[\mathrm{E}\,(T_{ni}-\mu_{ni}-Z_{ni})^2/\gamma_{ni}(1-\lambda_{ni})/C_n^2]=0 \tag{3.8}$$

where

$$\mu_{ni}=\sum_s\sum_j(d_{sj}^{(i)}-\bar{d}^{(i)})\int\phi(\bar{F}(y))dF_{sj}(y)\,, \qquad \bar{F}(y)=N^{-1}\sum_s\sum_j F_{sj}(y) \tag{3.9}$$

$$Z_{ni}(Y)=N^{-1}\sum_s\sum_j\sum_t\sum_v(d_{tv}^{(i)}-d_{sj}^{(i)})B_{tv}(Y_{sj}) \tag{3.10}$$

with

$$B_{tv}(Y_{sj})=\dot{\varphi}(F_{sj}(Y_{sj}))+Q_{sjtv}(Y_{sj})+\text{const.}\,, \tag{3.11}$$

and

$$|Q_{sjtv}(y)|\leqq k\max_{sjtv}|F_{sj}(y)-F_{tv}(y)|\,. \tag{3.12}$$

Now as in [2], we can write

$$\begin{aligned}\mathrm{E}\,(T_{ni}^*-T_{ni}^0)^2\leqq 4\,\mathrm{E}\,(T_{ni}^*-Z_{ni}^*-\mu_{ni}^*)^2+4\,\mathrm{E}\,(T_{ni}^0-Z_{ni}^0-\mu_{ni}^0)^2\\ +2\,\mathrm{E}\,(Z_{ni}^*-Z_{ni}^0)^2+2(\mu_{ni}^*-\mu_{ni}^0)^2\,.\end{aligned} \tag{3.13}$$

Now since the expectation in (3.8) is taken under any alternative whatsoever, including the sequence in (3.2), and in view of (2.4) and (2.5) it is enough to show that both

$$\mathrm{E}\,\{(Z_{ni}^*-Z_{ni}^0)/C_n\}^2 \tag{3.14}$$

and

$$\{(\mu_{ni}^*-\mu_{ni}^0)/C_n\}^2 \qquad \text{tend to zero}\,. \tag{3.15}$$

On expanding and integrating by parts, and on making use of (3.10) through (3.13), it readily follows as in Adichie [2] that both quantities in (3.14) and (3.15) are bounded by

$$K(1-\lambda_{ni})\gamma_{ni}\max_{ij}(x_{ij}^2/C_n^2)\,\mathrm{E}_n\,[\phi_y'(F(Y_{ij}))]\,. \tag{3.16}$$

By virture of (2.6) and (3.1) the assertions in (3.14) and (3.15) follow, which due to (2.4), (2.5) and (3.13) imply (3.7). That completes the proof of the lemma.

LEMMA 3.3. *If* W_{ni} *is as defined in* (2.16) *then under the conditions of Theorem* 3.2, *the random vector* $(W_{n1},\cdots,W_{nc})$ *has a limiting normal distribution with zero means and covariance matrix whose* (i, s)*th term is*

(3.17) $$\delta_{is}-(\lambda_{ni}\lambda_{ns})^{1/2}$$

where δ_{is} is the kronecker delta.

PROOF. From Lemma 3.2, $\hat{T}_{ni}$ and $T_{ni}^{(0)}$ $i=1,\cdots,c$, have the same limiting distribution under the sequence of alternatives (3.2). Since under (3.2) with $\beta=0$ in (2.1),

$$\max_{sjtv} |F_{sj}(y)-F_{tv}(y)| \leqq |(\theta_t-\theta_s)| f(y)/C_n$$

and

$$\max (1-\lambda_{ni})/\{\lambda_{ni}(N-n_i)\}^{1/2} \leqq K/C_n$$

it follows from the results of Hájek, (see Theorem 2.2 of [3]) that for each $i=1,\cdots,c$, and as $n\rightarrow\infty$, W_{ni} is asymptotically normally distributed with zero mean and variance $(1-\lambda_{ni})$. Furthermore the joint asymptotic normality of the vector $(W_{n1},\cdots,W_{nc})$ follows readily, for example from Remark 2.4 of [3]. Also as in the proof of Theorem 2.2 of [3] $\text{Cov}(T_{ni}^{0}, T_{ns}^{0})$ can be replaced, by an asymptotically equivalent expression given by $\text{Cov}(V_{ni}, V_{ns})$ where

(3.18) $$V_{ni}=\sum_{s}\sum_{j}(d_{sj}^{(i)}-\bar{d}^{(i)})\phi(F_{sj}(Y_{sj})) \text{ .}$$

Routine computations give that $\text{Cov}(V_{ni}, V_{ns})=-\{(n_i n_s)/N\}A^2$. The asymptotic $\text{Cov}(W_{ni}, W_{ns})$ is therefore given by

(3.19) $$\{(n_i n_s)^{1/2}/N\}=-(\lambda_{ni}\lambda_{ns})^{1/2}$$

which, with the asymptotic var $(1-\lambda_{ni})$ of W_{ni} gives the covariance matrix (3.17). Observe that T_{ni}^{0} $i=1,\cdots,c$, are the same statistics whose asymptotic normality has been considered by Puri [7] under Chernoff-Savage type of assumptions. A detailed discussion of these assumptions compared with Hájek's assumptions used in this paper is given in [3].

PROOF OF THEOREM 3.2. The asymptotic covariance matrix (3.17) of the vector $(W_{n1},\cdots,W_{nc})$ is singular of rank $(c-1)$. Using orthogonal (analysis of variance) transformation it follows that under (3.2), $\sum_i W_{ni}^2$ has asymptotically a noncentral chi-square distribution with $(c-1)$ degrees of freedom and noncentrality parameter given by

(3.20) $$\varDelta_L(F,\phi)=\sum_i \lim n_i^{-1}\{\mu_{ni}(\alpha_n)-\mu_{ni}(0)\}^2/A^2 \text{ .}$$

Where $\mu_{ni}(\alpha_n)$ and $\mu_{ni}(0)$ are the values of the quantity in (3.9) computed under (3.2) and (2.2) respectively.

On expanding and integrating by parts, it is easily seen that the right-hand side of (3.20) is equal to (3.3) and the proof is complete.

4. Asymptotic efficiency of $\hat{L}_n$ test

The usual method of testing the hypothesis (2.2) is based on the use of least squares estimates of the various parameters. The test statistic is (see [6], p. 286)

$$Q_n=\sum_i \sum_j [(Y_{i\cdot}-Y_{\cdot\cdot})+\bar{\beta}(x_{ij}-x_i)-\bar{\bar{\beta}}(x_{ij}-x_{\cdot\cdot})]^2/(c-1)S_e^2 \tag{4.1}$$

where

$$Y_{i\cdot}=n_i^{-1}\sum_j Y_{ij}\ , \qquad Y_{\cdot\cdot}=N^{-1}\sum\sum Y_{ij}\ ,$$

and the least squares estimators are

$$\bar{\beta}=\sum\sum (Y_{ij}-Y_i)(x_{ij}-x_i)/\sum\sum (x_{ij}-x_i)^2$$

$$\bar{\bar{\beta}}=\sum\sum (Y_{ij}-Y_{\cdot\cdot})(x_{ij}-x_{\cdot\cdot})/\sum\sum (x_{ij}-x_{\cdot\cdot})^2$$

and S_e^2 is the mean square due to error.

It is well known that for any $F(y)$ for which $\sigma^2(F)=\left[\int y^2dF(y)-\left(\int ydF(y)\right)^2\right]<\infty$, $(c-1)Q_n$ under (2.2) has asymptotically a chi-square distribution with $(c-1)$ degrees of freedom, and under (3.2) has asymptotically a noncentral chi-square distribution with $(c-1)$ degrees of freedom and noncentrality parameter,

$$\begin{aligned}\Delta_Q(F)=\lim \sum_i \sum_j [\{(\theta_i-\bar{\theta})/C_n\}-\{\sum_i \sum_j \theta_i(x_{ij}-x_{\cdot\cdot})/C_n^3\}\\ \cdot (x_{ij}-x_{\cdot\cdot})]^2/\sigma^2(F)\ .\end{aligned} \tag{4.2}$$

Under (2.8), this simplifies to

$$\Delta_Q(F)=\lim \sum_i \sum_j [(\theta_i-\bar{\theta})/C_n]^2/\sigma^2(F)=\sum_i \gamma_i(\theta_i-\bar{\theta})^2/\sigma^2(F)\ . \tag{4.3}$$

By the conventional method of measuring asymptotic efficiency, the efficiency of $\hat{L}_n$ test relative to the usual least squares test is therefore

$$\Delta_L/\Delta_Q=e(F,\phi)=\sigma^2(F)\left\{\int \phi_y'(F(y))dF(y)\right\}^2\Big/A^2\ . \tag{4.4}$$

Now (4.4) is the familiar efficiency of rank score tests relative to the classical tests in the ordinary c $(\geqq 1)$ sample problem. This result shows that even when the analysis of variance technique is compounded with regression, a suitable rank score tests can still be constructed.

One attractive feature of this efficiency expression is that it is independent of x's. As is apparent from (4.4), the efficiency expression depends only on the parent distribution function F of the observations, and on the score function ϕ.

The choice of ϕ would usually depend on F. If F and $f=F'$, are assumed known then the best choice of the score function is given by

$$\phi(u)=-[f'(F^{-1}(u))/f(F^{-1}(u))] \ . \tag{4.5}$$

In this situation, an asymptotically locally optimum parametric test is

$$Z_n=-2\log M_n \ , \tag{4.6}$$

where M_n is the likelihood ratio criterion. It is well known that under (3.2), Z_n has asymptotically a chi-square distribution with $(c-1)$ degrees of freedom and noncentrality parameter

$$\Delta_Z(F)=\Delta_Q\{\sigma^2(F)I(F)\}$$

where the Fisher Information $I(F)$ $\left(=\int (f'/f)^2\right)$ is assumed to be finite. It follows therefore that the efficiency of our $\hat{L}_n$ test based on (4.5), relative to Z_n is

$$e_{L,Z}(F)=\left(\int \phi'_v(F(y))dF(y)\right)^2\Big/I(F)A^2$$

which reduces to 1, for all F. In other words, if F is known, our method yields an asymptotically optimum rank test of the hypothesis.

Quite often, however, the functional form of F is really not known. In that case the commonest score functions used are $\phi(u)=u$ the Wilcoxon, and $\phi(u)=\Phi^{-1}(u)$, the normal scores. If the Wilcoxon is used, (i.e. if only the ranks are considered), then the efficiency in (4.4) reduces to

$$e_w(F)=12\sigma^2(F)\left\{\int f^2(y)dy\right\}^2 ; \tag{4.7}$$

an expression that has been studied fully by Lehmann. It is well known that $e_w(F)\geqq 0.864$ for all F. If $\phi(u)=\Phi^{-1}(u)$, the expression becomes

$$e_{\text{N.Sc.}}(F)=\sigma^2(F)\int \{f^2(y)/\phi(\Phi^{-1}(F(y)))\}^2 dy \ . \tag{4.8}$$

This expression has been studied in detail by Chernoff and Savage who found that $e_{\text{N.Sc.}}(F)\geqq 1$ for all F.

The table below gives the numerical values of $e_w(F)$ and $e_{\text{N.Sc.}}(F)$ for the common distributions. A detailed study of the efficiencies of Wilcoxon and normal scores criteria has been given by Hodges and Lehmann in [4].

Table I Numerical values of efficiency of $\hat{L}_n$ relative to the least square test Q_n

e \ F	Normal	Logistic	Uniform	Double Exponential	Exponential
$e_w(F)$	$\frac{3}{\pi}=.955$	$\frac{\pi^2}{9}=1.097$	1	1.5	3
$e_{\text{N.Sc.}}(F)$	1	$\frac{\pi}{3}=1.047$	∞	$\frac{4}{\pi}=1.273$	∞

Acknowledgement

The author is grateful to the referee for his useful comments on the paper.

UNIVERSITY OF NIGERIA

REFERENCES

[1] Adichie, J. N. (1967). Estimates of regression parameters based on rank tests, *Ann. Math. Statist.*, **38**, 894-904.

[2] Adichie, J. N. (1974). Rank score comparison of several regression parameters, *Ann. Statist.*, **2**, 396-402.

[3] Hájek, J. (1968). Asymptotic normality of simple linear rank statistics under alternatives, *Ann. Math. Statist.*, **39**, 325-346.

[4] Hodges, J. L. and Lehmann, E. L. (1960). Comparison of the normal scores and Wilcoxon tests, *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, **4**, No. 1, 307-317.

[5] Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one criterion analysis of variance, *J. Amer. Statist. Ass.*, **47**, 583-621.

[6] Lehmann, E. L. (1959). *Testing Statistical Hypothesis*, Wiley, New York.

[7] Puri, M. L. (1964). Asymptotic efficiency of a class of c-sample tests, *Ann. Math. Statist.*, **35**, 102-121.

[8] Sen, P. K. (1969). On a class of rank order tests for the parallelism of several regression lines, *Ann. Math. Statist.*, **40**, 1668-1683.