A Nonparametric Test for the Parallelism of Regression Lines Based on Kendall's Tau

Moon Sup Song*

ABSTRACT

For testing $\beta_i=\beta$, $i=1...,k$, in the regression model $Y_{ij}=\alpha_i+\beta_i x_{ij}+e_{ij}$, $j=1,...,n_i$, a simple and robust test based on Kendall's tau is proposed. Its asymptotic distribution is proved to be chi-square under the null hypothesis and noncentral chi-square under an appropriate sequence of alternatives. For the optimal designs, the asymptotic relative efficiency of the proposed procedure with respect to the least squares procedure is the same as that of the Wilcoxon test with respect to the t-test.

1. Introduction

Consider the regression model

$$Y_{ij}=\alpha_i+\beta_i x_{ij}+e_{ij},\ j=1,\ldots,n_i;i=1,\ldots,k \tag{1.1}$$

where e_{ij}'s are independent and identically distributed (iid) random variables, x_{ij}'s are known constants, α_i's are nuisance parameters, and β_i's are the regression parameters. Our problem is to test the null hypothesis

$$H_0 : \beta_1=\ldots=\beta_k=\beta \text{ (unknown)} \tag{1.2}$$

against the set of alternatives that $\beta_1,\ldots,\beta_k$ are not all equal.

In this paper, a nonparametric test for H_0 based on Kendall's rank correlation tau is proposed. For the regression model $Y_i=\alpha+\beta x_i+e_i$, $i=1,\ldots,n$, where e_i's are iid random variables, an estimator of β based on Kendall's tau was studied by Sen (1968). The procedure proposed in this paper is an extension of his idea to the problem of testing the homogeneity of the regression coefficients from k independent samples.

In the nonparametric case, comparisons of several regression coefficients are considered by Sen (1969), Hollander (1970), Potthoff (1974) and Adichie

*Kyunghee University. The author is grateful to Professor C. J. Park for his careful reading of the manuscript and valuable comments on the first draft of this paper.

(1974), among others. Hollander and Potthoff dealt with the hypothesis that two simple regression lines are parallel, when the error terms are not identically distributed. Sen(1969) proposed a class of rank order tests for the parallelism of several regression lines based on individual ranks of k different samples, while Adichie (1974) used the simultaneous ranking of all observations in the different k samples with the restriction $\alpha_1=\ldots=\alpha_k=\alpha$ (unknown). The procedures proposed by Sen and Adichie have good efficiency properties. However, the point estimators of β which are originally discussed by Adichie (1967) require trial and error solutions. Such a trial and error procedure may be quite laborious when the number of observations is not very small.

In Section 2, the test statistic is constructed. Section 3 presents the asymptotic distribution and the asymptotic efficiency of the proposed test statistic. For the optimally designed cases which are frequently encountered in practice, the asymptotic relative efficiency (A.R.E.) of the proposed test with respect to the variance-ratio test based on the least squares estimators is the same as that of the Wilcoxon test with respect to the Student's t-test.

2. Notations and Preliminaries

For each $i=1,\ldots,k$, let Y_{ij}, $j=1,\ldots,n_i$, be independent random variables with distributions

$$P(Y_{ij}\leq y)=F_{ij}(y)=F(y-\alpha_i-\beta_i x_{ij}),$$

where $F(y)$ is a continuous cumulative distribution function (cdf). Here x_{ij}'s are known constants, and α_i's and β_i's are unknown parameters. We wish to test the null hypothesis (1.2).

Without loss of generality we may assume that

$$x_{i1}\leq x_{i2}\leq\ldots\leq x_{in_i},\quad i=1,\ldots,k$$

where all the equality signs are not strict. For $i=1,\ldots,k$, let $\{x_{i1},\ldots,x_{in_i}\}$ be composed of a_i (≥ 2) distinct sets of elements, where in the t-th set there are u_{it} elements which are all equal for $t=1,\ldots,a_i$. We now define, for $i=1,\ldots,k$,

$$\bar{x}_i=(1/n_i)\sum_{j=1}^{n_i}x_{ij}\ ;\ C^2{}_{ni}=\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2\ ;$$

$$A^2{}_{ni}=(1/12)\left\{n_i(n_i^2-1)-\sum_{t=1}^{a_i}u_{it}(u_{it}^2-1)\right\}\ ; \tag{2.1}$$

$$\rho_{ni}=\sum_{j=1}^{n_i}(j-\frac{1}{2}(n_i+1))(x_{ij}-\bar{x}_i)/(A_{ni}C_{ni}).$$

Note that ρ_{ni} is the correlation coefficient, adjusted for ties, between $(1,\ldots,n_i)$ and $(x_{i1},\ldots,x_{in_i})$. From Theorem 6.3 of Sen (1968), we have

$$0\leq\rho_{ni}\leq 1,\ \text{for } i=1,\ldots,k.$$

We also define, for $i=1,\ldots,k$,

$$T_n{}^2=\sum_{i=1}^{k}\rho_{ni}{}^2C_{ni}{}^2\ ;\ \gamma_{ni}=\rho_{ni}{}^2C_{ni}{}^2/T_n{}^2 \tag{2.2}$$

It is assumed that, as $n\to\infty$,

$$n_i/(\sum_{i=1}^{k}n_i)\to c_i\ ;\ 0<c_0\leq c_1,\ldots,c_k\leq 1-c_0,$$

$$\gamma_{ni}\to\gamma_i\ ;\ 0<\gamma_0\leq\gamma_1,\ldots,\gamma_k\leq 1-\gamma_0,$$

where $c_0,\gamma_0\leq 1/k$. It is also assumed that

$$\rho_{ni}>0\ ;\ C_{ni}{}^2\to\infty \text{ as } n\to\infty,\ \text{for } i=1,\ldots,k. \tag{2.3}$$

Finally, we assume that $F(x)$ is absolutely continuous having a continuous density function satisfying

$$B(F)=\int_{-\infty}^{\infty}f^2(x)dx<\infty.$$

We define $c(u)$ to be 1, 0, or -1 according as u is $>$, $=$, or <0. For any real b, we also define

$$Z_{ij}(b)=Y_{ij}-bx_{ij},\ j=1,\ldots,n_i\ ;\ i=1,\ldots,k.$$

To estimate β_i in the model (1.1), we consider the statistics.

$$U_{ni}(b)=\sum_{1\leq s<t\leq ni}c(x_{it}-x_{is})c(Z_{it}(b)-Z_{is}(b)),\ \text{for } i=1,\ldots,k. \tag{2.4}$$

Thus, for each i, $U_{ni}(b)$ is equivalent to the Kendall's sample tau coefficient between the x_{ij} and the $(Y_{ij}-bx_{ij})$. Since $x_{is}\leq x_{it}$ for all $s<t$, $Z_{it}(b)-Z_{is}(b)$

is non-increasing in b for all $1\leq s<t\leq n_i$; $i,\ldots,k$. Hence from (2.4), it follows that $U_{ni}(b)$ is also non-increasing in b. Note that $Z_{i1}(\beta_i),\ldots,Z_{in_i}(\beta_i)$ are n_i iid random variables having cdf $F(y-\alpha_i)$, for $i=1,\ldots,k$, independent of $\boldsymbol{X}=(x_{i1},\ldots,x_{in_i})$. Consequently, $U_{ni}(\beta_i)$ is a strictly distribution-free statistic having a distribution symmetric about zero. Thus we may estimate β_i by choosing b which makes $U_{ni}(b)$ as close to zero as possible.

As in Sen (1968), we define the estimators $\beta_{n,i}^*$ of β_i as follows:

$$\beta_{n,i}^{(1)}=\mathrm{Sup}\{b: U_{ni}(b)>0\}\ ;\ \beta_{n,i}^{(2)}=\mathrm{Inf}\{b: U_{ni}(b)<0\}, \tag{2.5}$$

and

$$\beta_{n,i}^*=\frac{1}{2}(\beta_{n,i}^{(1)}+\beta_{n,i}^{(2)}). \tag{2.6}$$

Then $\beta_{n,i}^*$, defined in (2.5) and (2.6), is the median of the well-defined slope estimators for line i of the form

$$W_{st}^{(i)}=(Y_{it}-Y_{is})/(x_{it}-x_{is}), \tag{2.7}$$

where $x_{is}<x_{it}$, $1\leq s<t\leq n_i$; $i=1,\ldots,k$. (For the details, see Section 3 of Sen (1968)). Note also that $\beta_{n,i}^*$ is invariant in the sense that

$$\beta_{n,i}^*(\boldsymbol{Y}+a,\boldsymbol{X})=\beta_{n,i}^*(\boldsymbol{Y},\boldsymbol{X})+a, \text{ for all real } a,$$

and the distribution of $\beta_{n,i}^*$ is symmetric about the true parameter β_i.

Now, the proposed sample estimator of β in (1.2) is defined by

$$\beta_n^*=\sum_{i=1}^{k}\gamma_{ni}\beta_{n,i}^* \tag{2.8}$$

where γ_{ni}'s are defined in (2.2). That is, β_n^* is a weighted average of the individual slope estimators for k lines. To construct a test statistic for (1.2), we define

$$V_{ni}^2=\frac{1}{18}\left\{n_i(n_i-1)(2n_i+5)-\sum_{t=1}^{a_i}u_{it}(u_{it}-1)(2u_{it}+5)\right\} \tag{2.9}$$

for $i=1,\ldots,k$, where a_i and u_{it} are defined just before (2.1). Thus V_{ni}^2 is the variance of $U_{ni}(\beta_i)$ with the correction for tied observations in x. Now, our proposed test statistic is

$$\hat{U}_n=\sum_{i=1}^{k}[U_{ni}(\beta_n^*)/V_{ni}]^2 \tag{2.10}$$

The test is based on the following theorem, whose proof follows as a special case of Theorem 3. 1 in Section 3.

Theorem 2. 1. Under H_0 in (1. 2) and the conditions in this section, $\hat{U_n}$ in (2.10) has asymptotically a chi-square distribution with k-1 degrees of freedom (df). (2. 10)

In view of Theorem 2. 1, an asymptotic level ε test rejects H_0 in (1. 2) if $\hat{U_n}$ is greater than the upper $100\varepsilon\%$ point of the chi-square distribution with k-1 df.

3. Asymptotic distribution and asymptotic efficiency of $\hat{U}_n$

In order to study the asymptotic power properties and the asymptotic efficiency of the proposed $\hat{U_n}$-test, we consider the following sequence of alternatives

$$H_n : \beta_i=\beta+(\theta_i/T_n), i=1,\ldots,k ; \sum_{i=1}^{k}\gamma_i\theta_i=0. \tag{3. 1}$$

The proof of the limiting distribution of $\hat{U_n}$ depends on the following lemmas.

Lemma 3. 1. (Theorm 7. 1 of Sen (1968)). For $i=1,\ldots,k$, under H : $\beta_i=0$ and the conditions (2. 3),

$$U_{ni}(b/(\rho_{ni}C_{ni}))/V_{ni}$$

has asymptotically a normal distribution with mean $-4bB(F)A_{ni}/V_{ni}$ and unit variance.

If we notice that $A_{ni}/V_{ni}\rightarrow\sqrt{3}/2$ as $n\rightarrow\infty$, the following lemma can be easily proved using Lemma 3. 1.

Lemma 3. 2. For any real (a,b),

$$[U_{ni}(\beta_i-a/(\rho_{ni}C_{ni}))-U_{ni}(\beta_i-b/(\rho_{ni}C_{ni}))]/V_{ni}$$

converges (in probability) to $2\sqrt{3}(a-b)B(F)$.

Lemma 3. 3. (Theorem 6. 1 of Sen (1968)). For $i=1,\ldots,k$, under the conditions (2. 3),

$\rho_{ni}C_{ni}(\beta_{n,i}^*-\beta_i)$

has asymptotically a normal distribution with mean zero and variance $1/(12B^2(F))$.

Lemma 3.4. (i) For $i=1,\ldots,k$,

$|\rho_{ni}C_{ni}(\beta_{n,i}^*-\beta_i)|$ is bounded (in probability).

(ii) Under $\{H_n\}$ in (3.1),

$|T_n(\beta_n^*-\beta)|$ is bounded (in probability).

Proof. Using the fact that $B(F)<\infty$, Part (i) is a direct consequence of Lemma 3.3. Part (ii) is also obvious if we notice the equality

$$T_n(\beta_n^*-\beta)=\sum_{i=1}^{k}\gamma_{ni}^{\frac{1}{2}}\rho_{ni}C_{ni}(\beta_{n,i}^*-\beta_i)+\sum_{i=1}^{k}\gamma_{ni}\theta_i, \tag{3.2}$$

where the last term in (3.2) converges to zero under (3.1).

Now we have the main theorm.

Theorem 3.1. Under $\{H_n\}$ in (3.1) and the conditions in Section 2, $\hat{U}_n$, defined in (2.10), has asymptotically a noncentral chi-square distribution with $k-1$ df and the noncentrality parameter

$$\Delta_u=12B^2(F)\sum_{i=1}^{k}\gamma_i\theta_i^2$$

Proof. Since $\beta_{n,i}^*$ is the median of the slope estimators $W_{st}^{(i)}$ in (2.7), it follows that

$$U_{ni}(\beta_{n,i}^*)=0, \quad \text{for } i=1,\ldots,k. \tag{3.3}$$

Thus, by (2.10) and (3.3), $\hat{U}_n$ can be written as

$$\hat{U}_n=\sum_{i=1}^{k}[U_{ni}(\beta_n^*)/V_{ni}-U_{ni}(\beta_{n,i}^*)/V_{ni}]^2. \tag{3.4}$$

While, by Lemma 3.2. and Lemma 3.4, the right hand side of (3.4) is equivalent (in probability) to

$$12B^2(F)\sum_{i=1}^{k}[\rho_{ni}C_{ni}(\beta_{n,i}^*-\beta_n^*)]^2 \tag{3.5}$$

Now, according to Lemma 3.3, $\rho_{ni}C_{ni}(\beta_{n,i}^*-\beta_i)$, $i=1,\ldots,k$, are independent, and each has asymptotically a normal distribution with mean zero and common variance $1/(12B^2(F))$. Here we note the following equalities:

$$\rho_{ni}C_{ni}(\beta_{n,i}^*-\beta_i)=\rho_{ni}C_{ni}(\beta_{n,i}^*-\beta)+\gamma_{ni}^{\frac{1}{2}}\vartheta_i\ ; \tag{3.6}$$

$$\sum_{i=1}^{k}[\rho_{ni}C_{ni}(\beta_{n,i}^*-\beta)]^2=\sum_{i=1}^{k}[\rho_{ni}C_{ni}(\beta_{n,i}^*-\beta_n^*)]^2$$

$$+T_n^2(\beta_n^*-\beta)^2-2T_n^2(\beta_n^*-\beta)\sum_{i=1}^{k}[\gamma_{ni}(\beta_n^*-\beta_{n,i}^*)]. \tag{3.7}$$

By (2.2) and (2.8), the last term in (3.7) is zero. Also, from the expression (3.2), $T_n(\beta_n^*-\beta)$ converges to

$$\sum_{i=1}^{k}\gamma_{ni}^{\frac{1}{2}}\rho_{ni}C_{ni}\ (\beta_{n,i}^*-\beta_i).$$

Thus, by lemma 3.3, $T_n(\beta_n^*-\beta)$ has asymptotically a normal distribution with mean zero and variance $1/(12B^2(F))$. Hence, from (3.6), (3.7) and Lemma 3.3, the theorem follows.

As mentioned before, we have $0\leq\rho_{ni}\leq 1$, for $i=1,\ldots,k$, If $\rho_{ni}=1$, we say that the independent variables are optimally designed, and if $\rho_{ni}\to 1$ as $n\to\infty$, asymptotically optimally designed (cf. Sen (1968, p. 1386)). Two particular cases where $\rho_{ni}=1$ are frequently encountered in practice. The first case is the equally spaced no replication design, i.e.,

$$x_{ij}=x_{i1}+(j-1)h_i,\ h_i>0,\ i=1\ldots,n_i.$$

The second case is the two-point design, i.e.,

$$x_{i1}=\ldots=x_{it_i}=x_{i1}^*,\ x_{i(t_{i+1})}=\ldots=x_{in^i}=x_{i2}^*,$$

where $x_{i1}^*<x_{i2}^*$ and $t_i<n_i$, for $i=1,\ldots,k$.

The most commonly used test for the hypothesis (1.2) is based on the least squares estimators $\hat{\beta}_n$ and $\hat{\beta}_{n,i}$ of the parameters β and β_i. If we work with $x'_{ij}=(x_{ij}-\bar{x}_{i\cdot})$ instead of x_{ij}(and also Y_{ij}'), the test statistic becomes

$$Z_n=\sum_{i=1}^{k}[C_{ni}^2(\hat{\beta}_{n,i}-\hat{\beta}_n)^2/(k-1)S_e^2]$$

where

$$\hat{\beta}_{n,i}=\sum_{j=1}^{n_i} Y_{ij}(x_{ij}-\bar{x}_i)/C_{ni}^2;\ \hat{\beta}_n=\sum_{i=1}^{k}\gamma_{n,i}^*\hat{\beta}_{n,i};\quad C_n^2=\sum_{i=1}^{k}C_{ni}^2;$$

$$\gamma_{ni}^*=C_{ni}^2/C_n^2,\quad i=1,\ldots,k,$$

and S_e^2 is the (pooled) within sample mean square due to error. It is well known that, for any $F(x)$ with finite variance $\sigma^2(\mathrm{F})$, under (3. 1) with C_n instead of T_n and γ_i^* instead of γ_i where γ_i^* is the limit of γ_{ni}^* as $n\to\infty$, $(k-1)Z_n$ has asymptotically a noncentral chi-square distribution with $k-1$ df and noncentrality parameter

$$\Delta_z=\left[\sum_{i=1}^{k}\gamma_i^*\theta_i^2\right]/\sigma^2(F).$$

Thus, according to the usual method of measuring the A.R.E. (cf. Noether (1967, p. 86)), for the optimal or asymptotically optimal designs, the A.R.E. of the $\hat{U}_n$-test relative to the Z_n-test is

$$e_{u,z}=12B^2(F)\sigma^2(F), \tag{3. 8}$$

which is the same as that of the Wilcoxon test relative to the Student's t-test (cf. Hodges and Lehmann (1956)). Thus, when $F(x)$ is normal, $e_{u,z}$ in(3. 8) is $3/\pi=0.955$. When $F(x)$ has heavy tails (such as Cauchy), $e_{u,z}$ may be infinitely large, and for any continuous $F(x)$ it cannot be less than 0. 864. But, if the independent variables are not optimally designed, the A.R.E. may not have any lower bound.

4. Conclusion

In the regression problems, long-tailed error distributions and inhomogeneity of variances have almost indistinguishable effects on the statistical inferences about regression parameters. Since the least squares estimator $\hat{\beta}$ is a weighted mean of the sample slope estimators of the form $W_{ij}=(Y_j-Y_i)/(x_j-x_i)$ with weights equal to $(x_j-x_i)^2$, a single grossly outlying observation may spoil the result, while β^* is the median of the same set of slope estimators

and therefore it will be more robust than $\hat{\beta}$. Moreover, $\hat{U}_n(\beta)$, which is the statistic in (2. 10) when β_n^* is replaced by β, is a distribution-free statistic under (1. 2) having a distribution symmetric about zero. Thus, the procedure proposed in this paper is recommended when the normality of the error terms is suspected.

Sen (1969) and Adichie (1974) proposed a class of rank order tests for the parallelism of regression lines. If we have some information about the distribution of the error terms, by a proper choice of the score generating function, we may construct an optimum test. But, trial and error solutions for the estimators of β are not easy, unless the sample size is very small. Thus, when the error terms have heavy tails without any further information, the $\hat{U}_n$-test may be simpler and more robust.

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] Hodges, J. L., and Lehmann, E. L. (1956), "The Efficiency of Some Nonparametric Competitors of the T-Test," *Ann. Math. Statist.* 27, 324-335.

[4] Hollander, M. (1970). "A Distribution-free Test for Parallelism," *J. Amer. Statist. Assoc.* 65, 387-394.

[5] Noether, G. (1967). *Elements of Non-parametric Statistics,* New York: John Wiley.

[6] Potthoff, R. F. (1974). "A Non-parametric Test of Whether Two Simple Regression Lines are Parallel," *Ann. Statist.* 2, 295-310.

[7] Sen, P. K. (1968). "Estimates of the Regression Coefficients based on Kendall's tau", *J. Amer. Statist. Assoc.* 63, 1379-1389.

[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.

<요 약>

Kendall의 Tau에 의한 회귀직선의 평행성에 관한 비모수 검정

회귀모델 $Y_{ij}=\alpha_i+\beta_i X_{ij}+e_{ij}$ $(i=1,2,\ldots,k;j=1,\ldots,n_i)$에 있어서 $\beta_i=\beta$를 검정하고자 Kendall의 tau를 사용하여 robust한 검정을 고찰하였다.

귀무가설 하에서의 이의 극한 분포는 χ^2 분포를 하며 대립가설들에 대해서는 noncentral χ^2 분포를 한다는 것을 증명하였다. optimal design에 대하여 이 논문에서 제안된 방법의 최소자승법에 대한 asymptotic relative efficiency는 t-검정에 대한 Wilcoxon 검정의 것과 똑같음을 보였다.