Weighted U-statistics for Simple Linear Regression

*Kim Do-hyun, Kim Chul-soo**

단순한 선형회귀에 대한 무게있는 U−통계량

金道鉉， 金鐵洙*

1. Introduction

Consider an arbitrary kernel $h(x_1, \cdots, x_n)$, not necessarily symmetric, to be applied as usual to observations $X_1, \cdots, X_n$ taken m at a time. Suppose also that each term $h(X_{i_1}, \cdots, X_{i_m})$ becomes weighted by a factor $w(i_1, \cdots, i_m)$ depending only on the indices $i_1, \cdots, i_m$. In this case the U−statistics sum takes the more general form $T_n = w(i_1, \cdots, i_m)h(X_{i_1}, \cdots, X_{i_m})$. In this case that h is symmetric and the weights $w(i_1, \cdots, i_m)$ take only 0 or 1 as values. Certain "weighted U−statistics" for simple linear regression take the form T_n.

Consider the simple linear regression model

$Y_i = \alpha + \beta x_i + e_i, \quad 1 \le i \le n. \quad (1.1)$

where α and β are unknown parameters, x_i are known regression scores, and e_i are i. i. d. random variables with c. d. f. F.

In this paper we extend the procedures of Theil by using the projection of weighted U−statistic of the form weighted rank for the simple linear regression model (1.1).

2. A weighted U−statistics

For the regression model (1.1), assume that F is continuous to rule out ties among the Y 's. Also assume that $x_1 \le x_2 \le \cdots \le x_n$ with at least one strict inequality. We will consider inferences for based on weighted U−statistics defined by

$$T_\beta = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{ij}\, \phi(Y_i - \alpha - \beta x_i,\ Y_j - \alpha - \beta x_j),$$

where $\phi(u, v) = 1$ or 0 according as $u < v$ or $u > v$. The weights $a_{ij} \ge 0$ are arbitrary but assume that $a_{ij} = 0$ if $x_i = x_j$. We define the slope of the line segment from the point(x_i, Y_i) to the point (x_j, Y_j) by

$S_{ij} = (Y_j - Y_i)/(x_j - x_i), \quad i < j, \quad x_i \ne x_j.$

Note. 1. T_β is a function of the slope S_{ij} since $\phi(Y_i - \alpha - \beta x_i,\ Y_j - \alpha - \beta x_j) = 1$ when $S_{ij} > \beta$.

2. The distribution of T_β depends on the weights a_{ij}.

3. If $a_{ij} = 1$, the wilcoxon distribution can be applied.

But it is not generaly feasible to tabulate the exact distribution for smaller sample sizes. For larger sample sizes, the distribution of T_β can be approximated by a normal distribution.

2.1. An estimator associated with the Theil

師範大學 專任講師，延世大學校 講師 *

statistic to estimate β of model (1.1).

a) Let $N=\binom{n}{2}$ and form the N sample slope $S_{ij}=(Y_j-Y_i)/(x_j-x_i)$, $i<j$, $x_i\neq x_j$

b) The estimator of β is
$$\hat{\beta}=\text{median }\{S_{ij}\}.$$

Let $S^{(1)}\leq S^{(2)}\leq\cdots\cdots\leq S^{(N)}$ denote the ordered values of S_{ij}.

Then if N is odd, say $N=2k+1$, we have $\hat{\beta}=S^{(k+1)}$.

If N is even, say $N=2k$, then
$$\hat{\beta}=\frac{S^{(k)}+S^{(k+1)}}{2}$$

The estimator $\hat{\beta}$ is less sensitive to gross errors than is the classical least-square estimator.

Sen has generalized Theill's estimator to the case where the x's are not distinct.

2.2. Estimation for β by T_β

Consider the estimator $\hat{\beta}$ of the slope parameter defined by
$$\hat{\beta}_U=\sup\ \{\beta:\ T_\beta>\frac{a\cdots}{2}\},$$
$$\hat{\beta}_L=\inf\ \{\beta:\ T_\beta<\frac{a\cdots}{2}\},$$
and $\hat{\beta}=(\hat{\beta}_u+\hat{\beta}_L)/2$, where
$$a\cdots=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}a_{ij}.$$

Note. T_β is a nonincreasing, left-continuous function of that ranges from $a\cdots$ down to zero.It is a step function with jumps a_{ij} occurring at the point S_{ij}.

Consider the asymptotic distribution of T_β. The distribution of T_β when the slope parameter is β is the same as the distribution of T_0 when $\beta=0$. Now T_0 is a function of the ranks of the Y's and the basic approach is to consider the projection of T_0, say T_0^*, into the family of linear rank statistics and to establish that T_0 and T_0^* have the same asymptotic distributions.

Define the row and column sums of the weights by
$$a_{i\cdot}=\sum_{j=i+1}^{n}a_{ij}$$
and
$$a_{\cdot j}=\sum_{i=1}^{j-1}a_{ij}\quad\text{for } 1\leq i,\ j\leq n$$
with $a_{n\cdot}=0$ and $a_{\cdot 1}=0$. Let $A_i=a_{\cdot i}-a_{i\cdot}$ and
$$a\cdots=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}a_{ij}.$$

From the Hájek and Šidák, the projection of T_0 is
$$T_0^*=(1/n)\sum_{i=1}^{n}A_iR_i+a\cdots/2,$$
where R_i is the rank of Y_i among $\{Y_1\cdots\cdots Y_n\}$, $1\leq i\leq n$. That is, this can be verified by the following lemma and theorem.

Lemma(Hájek). Let $Z_1,\cdots,Z_n$ be independent random variables and $S=S(Z_1,\cdots,Z_n)$ any statistic satisfying $E(S^2)$. Then the random variable
$$\hat{S}=\sum_{i=1}^{n}E(S|Z_i)-(n-1)E(S)$$
satisfies
$$E(\hat{S})=E(S)$$
and
$$E(\hat{S}-S)^2=Var(S)-Var(\hat{S}).$$

The random variable $\hat{S}$ is called the projection of S.

It is also possible to apply the technique to project a statistic onto dependent random variables.

Theorem. Consider a rank statistic $T=T(R_1,\cdots,R_N)$ and put
$$\hat{a}(i,\ j)=E(T\ R_i=j),\quad 1\leq i,\ j\leq N.$$

Then the statistic
$$\hat{T}=\frac{N-1}{N}\sum_{i=1}^{N}\hat{a}(i,\ R_i)-(N-2)E(T)$$
is the projection of T into the family of linear rank statistics.

Proof. See Hâjek, p. 59.

The following two theorems are immediate from

Hâjek and Šidák(1967, p. 163).

Theorem 1: Assume $\beta=0$ and the condition

$$A: \sum_{i=1}^{m} A_i^2 / \max_{1\le i\le n} A_i^2 \to 0 \text{ as } n\to\infty.$$

Then T_0^* is asymptotically $N(a\cdots/2, \sum_{i=1}^{n} A_i^2/12)$.

Theorem 2: Assume $\beta=0$, condition A, and condition

$$B: \sum_{i<j\le n} a_{ij}^2 / \sum_{i=1}^{n} A_i^2 \to 0 \text{ as } n\to\infty.$$

Then T_0 is asymptotically $N(a\cdots/2, \sum_{i=1}^{n} A_i^2/12)$.

The exact variance of T_0 is

$$\left(\sum_{i=1}^{n} A_i^2 + \sum_{i=1}^{n-1}\sum_{j=i+1}^{m} a_{ij}^2\right)/12,$$

Now the standard test of $\beta=0$ of the model (1.1) is introduced. This is based on $\sum_{i=1}^{n}(x_i-\bar{x})Y_i$. This suggests a rank statistic

$$U=\sum_{i=1}^{n}(x_i-\bar{x})R_i$$

where $R_1,\cdots,R_n$ are the ranks of $Y_1,\cdots,Y_n$.

Under $\beta=0$, $Y_1,\cdots,Y_n$ are i. i. d. random variables with c. d. f. $F(y-\alpha)$, we obtain the following theorem.

Theorem 3: In the simple regression model(1.1), suppose $\beta=0$ and suppose that

$$(1/n)\sum_{i=1}^{n}(x_i-\bar{x})^2 \to \delta^2 > 0$$

Then

$$U^*=\frac{1}{(n+1)\sqrt{n}}\sum_{i=1}^{n}(x_i-\bar{x})(R_i-(n+1)/2)$$ is asymptotically $N(0, \delta^2/12)$.

Proof. First, note that the rank of Y_j among $Y_1, \cdots, Y_n$ can be written as

$$R_j=1+\sum_{i=1}^{n} s(Y_j-Y_i).$$

where $s(x)=1$ if $x>0$ and 0 otherwise. Then, since

$$E[s(Y_j-Y_i)\mid Y_k=y]=\begin{cases} F(y) & k=j \\ 1-F(y) & K=i \\ 1/2 & k\ne i \text{ or } j, \end{cases}$$

we have

$$E[R_j\mid Y_k=y]=1+\sum_{i=1}^{n}E[s(Y_j-Y_i)\mid Y_k=y] =\begin{cases} 1+(n-1)F(y) & k=j \\ 1+(n-2)/2+(1-F(y)) & k\ne j \end{cases}$$

And we have

$$E[U^*\mid Y_k=y]=\frac{1}{(n+1)\sqrt{n}}\{\sum_{j\ne k}(x_j-\bar{x})[1/2-F(y)]+(x_K-\bar{x})[(n-1)F(y)-(n-1)/2]\}$$

$$=\frac{\sqrt{n}}{(n+1)}(x_K-\bar{x}[F(y)-1/2].$$

Hence the projection V_p of U^* is

$$V_p=\frac{\sqrt{n}}{(n+1)}\sum_{k=1}^{n}(x_k-\bar{x})[F(y_k)-1/2].$$

Since $F(y_k)$ has a uniform distribution on (0, 1) with mean 1/2 and variance 1/12, we have

$$\text{Var } V_p=\frac{n}{12(n+1)^2}\sum_{k=1}^{n}(x_k-\bar{x})^2 \to \frac{1}{12}\delta^2.$$

Hence Var $U^*\to\delta^2/12$, so $E(U^*-V_p)^2\to 0$.

Thus, U^* and V_p have the same distribution.

That is, V_p(and hence U^*) is asymptotically $N(0, \delta^2/12)$.

3. Confidence intervals for β

3.1. Confidence interval based on the Theil test.

For a symmetric two-sided confidence interval for β with confidence coefficient $1-\alpha$,

a) Determine the constant C_α that satisfies the equation

$$P_0\{-C_\alpha \le C \le C_\alpha\}=1-\alpha.$$

Note that $C_\alpha+2=k((\alpha/2), n)$.

This values were evaluated from the Kendall's K-statistic.

b) Obtain the ordered values $S^{(1)} < \cdots < S^{(N)}$ of the $N = \binom{n}{2}$ sample slopes $S_{ij} = (Y_j - Y_i)/(x_j - x_i)$.

c) Set $M_1 = \frac{N - C_\alpha}{2}$ and $M_2 = \frac{N + C_\alpha}{2}$.

d) The $1-\alpha$ confidence interval (β_L, β_U) is defined by $\beta_L = S^{(M_1)}$ and $\beta_U = {}^{(M_2+1)}$

Hence we have $P_\beta\{\beta_L < \beta < \beta_U\} = 1 - \alpha$.

3.2. Confidence interval based on T_β

Consider the simple linear regression model (1.1) and the test of $H_0: \beta = 0$ against $H_1: \beta > 0$, where the test is based on the statistic $T_{\beta 0}$. From the theorem 2, the test which rejects H_0 if $T_{\beta 0} > (a \cdots / 2) + z_\alpha (\sum_{i=1}^{n} A_i^2 / 12)^{1/2}$ is an approximate level α test, where z_α is the quantile of order $1-\alpha$ for the standard normal distribution.

A confidence interval for β can be obtained by inverting the two-sided alternatives.

Let H(s) denote the cumulative distribution function of the discrete probability which assigns probability $a_{ij}/a\cdots$ to the points S_{ij}. Let $H^{-1}(u) = \inf\{s : H(s) \geq u\}$. Suppose that t_1 and t_2 satisfy $P_\beta(t_1 < T_\beta < t_2) = 1 - \alpha$. We note that $P(T_t < a\cdots/2) < P(\hat{\beta} < t) < P(T_t \leq a\cdots/2)$. Since the event $\{t_1 < T_\beta < t_2\}$ is equivalent to the event $\{H^{-1}(t_1) < \beta < H^{-1}(t_2)\}$, the interval $(H^{-1}(t_1), H^{-1}(t_2))$ is a $(1-\alpha)$ 100 percent confidence interval for β. By theorem 2, we determine the constants t_1 and t_2. An approximate $(1-\alpha)$ 100 percent confidence interval for β

$$(H^{-1}((1/2) - Z\alpha/2 (\sum_{i=1}^{n} A_i^2/12)^{1/2}/a\cdots),$$

$$(H^{-1}((1/2) - Z\alpha/2 (\sum_{i=1}^{n} A_i^2/12)^{1/2}/a\cdots)).$$

4. Conclusion

A natural estimate of β based on Theil statistic was the median of the pairwise slopes. Now various weights a_{ij} are introduced.

a) Let the weights be given by $a_{ij} = 1$, $i < j$, if $x_i \neq x_j$; otherwise let $a_{ij} = 0$.

If the x_i's are all distinct, then $A_i = 2i - (n+1)$, $a\cdots = n(n-1)/2$, and conditions A and B hold.

b) Let $a_{ij} = j - i$, $i < j$, if $x_i \neq x_j$; otherwise 0.

c) Let $a_{ij} = x_j - x_i$, $i < j$. Then $A_i = n(x_i - \bar{x})$ and T_β is asymptotically normal with mean $\sum_{i=1}^{n}\sum_{j=1}^{n}(x_j - x_i)/2$ and variance $n^2 \sum_{i=1}^{n}(x_i - \bar{x})^2/12$.

Another weight a_{ij} can be given by $a_{ij} = (x_j - x_i)/(j-i)$, $i < j$, but this weight a_{ij} does not have a simple representation in evaluating T.

If the above weights are considered in view of the efficiency considerations, the weight $a_{ij} = x_j - x_i$ is recommended.

That is, the weight $a_{ij} = x_j - x_i$ merits serious considerration as an alternative to the classical estimate.

Literature cited

Adichie, J. N., 1967. Estimates of regression parameters based on rank test. A. M. S. 38; 894–904.

Jaroslav Hájek and Zbynek Šidák, 1967. Theory of rank test. Academic press, New York. pp.56–165.

Lehmann, E. L., 1975. Nonparametric statistical method based on ranks. McGraw-Hill, New York., pp. 82–93.

Myles Holiander and D. A. Wolfe, 1973. Nonparametric ststistical methods, John Willey & SONS pp. 67–75.

Pranab Kumar Sen, 1981. Sequential nonparametrics, John Willey & SONS. pp. 86–90.

Randles R. H. and Wolfe D. A. 1979. Introduction to the theory of nonparametric statistics, John Willey & SONS. pp. 103–113.

국 문 초 록

순위 검정 통계량의 형태인 무게있는 U-통계량을 사용해서 단순 회귀 모형에 대한 Theil의 검정을 살펴보며 회귀계수 β에 대한 신뢰구간을 추정한다.