University of Illinois	**Econ 476**	Department of Economics
Spring 2001	**Final Exam (Revised)**	Professor Roger Koenker

This is a take home final exam. It will be available on the web at 9am Monday, May 7 and it is due on Wednesday, May 11 by 5pm. You can put it in my mailbox in 484 Wohlers Hall.

The exam is motivated by my theory that a crucial objective of 476 is to prepare you to *critically* read econometric theory. The question is structured much like another lecture for 476 and questions of interpretation, or questions requiring you to fill in some missing details are interspersed. In the process of doing the exam, I hope that you will learn something useful about U-statistics. U-statistics play an increasingly important role in econometrics as recent work by Honoreánd Powell illustrate. Their early history in statistics is also closely tied to econometrics by a series of papers by Theil in which he suggested an estimator of the slope parameter in the simple bivariate regression model that was the median of all the pairwise slopes.

We will consider some second order U-statistics of the form

$$U_n = \left(\begin{array}{c} n \\ 2 \end{array} \right)^{-1} \sum\sum\nolimits_{i<j} H(X_i, X_j)$$

where $(H(\cdot,\cdot)$ is a symmetric function, i.e., $H(a,b) = H(b,a)$ for all pairs (a,b). As in many other places in mathematics H is called the kernel function. We will assume throughout that $X_1, \dots, X_n$ are iid with df F and that

$$EH^2(X_i, X_j) < \infty \qquad i \neq j$$

Let

$$\theta \equiv EH(X_i, X_j) = \int H(x_1, x_2)dF(x_1)dF(x_2)$$

and define the random variables

$$Y_i \equiv E(H(X_i, X_j)|X_i)$$

So $EY_i = \theta$, and set $V(Y_i) = \sigma^2$. It is convenient to adapt some notation from the analysis of variance, so we will write,

$$\begin{aligned} H_{ij} &= H(X_i, X_j) \\ H_{i\cdot} &= H_{\cdot i} = E(H(X_i, X_j)|X_i) \\ H_{\cdot\cdot} &= EH(X_i, X_j) = \theta \end{aligned}$$

Thus,

$$\begin{aligned} U_n - \theta &= \left(\begin{array}{c} n \\ 2 \end{array} \right)^{-1} \sum\sum\nolimits_{i<j} (H_{ij} - H_{\cdot\cdot}) \\ &= (n(n-1))^{-1} \sum\sum\nolimits_{i \neq j} [H_{ij} - H_{i\cdot} - H_{\cdot j} + H_{\cdot\cdot} + (H_{i\cdot} - H_{\cdot\cdot}) + (H_{\cdot j} - H_{\cdot\cdot})] \\ &= n^{-1} \sum (H_{i\cdot} - H_{\cdot\cdot}) + n^{-1} \sum (H_{\cdot j} - H_{\cdot\cdot}) \\ &\quad + (n(n-1))^{-1} \sum\sum\nolimits_{i \neq j} (H_{ij} - H_{i\cdot} - H_{\cdot j} + H_{\cdot\cdot}) \\ &= 2(\bar{Y}_n - \theta) + V_n, \end{aligned}$$

where $\bar{Y}_n = n^{-1}\sum Y_i$.

(a.) Show that Cov $(\bar{Y}_n, V_n) = 0$. First, provide a formal argument, then try to explain why, in light of the remark below, by making some analogy to regression.
Hint. Argue that

$$E((H_{k\cdot} - \theta)(H_{ij} - H_{i\cdot} - H_{\cdot j} + \theta)) = 0$$

for $k \notin \{i, j\}$ this is easy, for $k = i \neq j$ take expectations conditional on $X_k = X_i$ first.

(b.) Show Var $(V_n) = 2$ Var $(V_{12})/(n(n-1))$ where $V_{ij} = H_{ij} - H_{i\cdot} - H_{\cdot j} + H_{\cdot\cdot}$ Hint: Write Var $(V_n) = (n(n-1))^{-2}\sum\sum_{i\neq j}\sum\sum_{k\neq \ell}$ Cov $(V_{ij}, V_{k\ell})$ and use the approach in (a.).

(c.) Using the results of (a.) and (b.) argue that,

$$\sqrt{n}(U_n - \theta) \rightsquigarrow \mathcal{N}(0, 4\sigma^2)$$

Remark. You can view the decomposition of $U_n - \theta$ into $2(\bar{Y}_n - \theta)$ and V_n as a least squares projection. Recall that least squares projection *is* the principle underlying approximation of random variables by their conditional mean function. The theory of U-statistics developed by Hoeffding and later by Hajek and others is built on this foundation. This helps to explains the close analogy with the analysis of variance relationships.

Examples

(1.) *Sample Variance.* Consider the kernel function

$$H(x, y) = (x - y)^2$$

Let $\mu = EX, \sigma^2 = E(X_1 - \mu)^2$ and assume $\mu_4 = E(X_1 - \mu)^4 < \infty$.

(d.) Show that $Y_i = E((X_i - X_j)^2 | X_i) = (X_i - \mu)^2 + \sigma^2$, and that

$$\sigma_Y^2 = \text{ Var } (Y_i) = \mu_4 - \sigma^4$$

and conclude that

$$\sqrt{n}(U_n - \theta) \rightsquigarrow \mathcal{N}(0, 4(\mu_4 - \sigma^4))$$

where $U_n = \binom{n}{2}^{-1} \sum\sum_{i<j}(X_i - X_j)^2$.

Remark. If we write

$$\begin{aligned}
U_n &= (n(n-1))^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i^2 + X_j^2 - 2X_iX_j) \\
&= \frac{2}{n-1}\sum_{i=1}^{n}(X_i^2 - \bar{X}_n^2) \\
&= \frac{2}{n-1}\sum(X_i - \bar{X}_n)^2 \\
&= 2s_n^2
\end{aligned}$$

so U_n may be viewed as just a rather inefficient way to compute the ordinary sample variance.

(2.) *Gini's Mean Difference.* A classical alternative to the variance as a measure of dispersion of a distribution is Gini's mean difference

$$\hat{\theta} = \binom{n}{2}^{-1} \sum\sum_{i<j} |X_i - X_j|$$

(e.) Show that

$$Y_j = E(|X_i - X_j| \mid X_j) = X_j(2F(X_j) - 1) + \mu - 2\psi(F(X_j))$$

where $\mu = \int_0^1 F^{-1}(t)dt = \int_{-\infty}^{\infty} x dF(x)$ and $\psi(t) = \int_0^t F^{-1}(s)ds$.

(f.) Show that

$$\begin{aligned} EY_j &= 2(\mu - 2\int_0^1 \psi(t)dt) \equiv \theta \\ VY_j &= \int (\int |x-y| dF(x))^2 dF(y) - \theta^2 \end{aligned}$$

Remark. The parameter θ is related to the classical Gini coefficient used to measure the degree of inequality of income distributions. For a positive random variable X with distribution function F, let

$$\lambda(t) = \mu^{-1} \int_0^t F^{-1}(s)ds$$

denote the Lorenz curve. The proportion of aggregate income earned by the poorest proportion t of the population is given by $\lambda(t)$. The Gini coefficient is usually defined as twice the area (in the unit square) between the 45° line and the Lorenz curve, i.e.

$$\gamma = (1 - 2\int_0^1 \lambda(t)dt)$$

Thus,

$$EY_i = 2\mu\gamma$$

and we obtain a nice old way to write γ as,

$$\gamma = \frac{E|X_i - X_j|}{E(X_i + X_j)}.$$

This may be interpreted as follows: find the average disparity of income between randomly selected individuals and normalize by twice the mean income. In a recent survey Atkinson reports that for OECD countries γ varies from about .22 for Finland and Sweden to about .35 for the US. This is the empirical aspect of the exam!

(3.) (Hodges-Lehmann Estimator) As a final example I would like you to consider something a bit more sophisticated in which the U-statistic is employed as an objective function. Let

$$U_n(\alpha) = \binom{n}{2}^{-1} \sum\sum_{i<j} |(x_i + x_j)/2 - \alpha|$$

(g.) Show that for $X_1, \ldots, X_n$, iid F, with bounded continuous density, f, symmetric about α_0, the estimator

$$\hat{\alpha}_n = \text{ argmin } U_n(\alpha)$$

satisfies $\sqrt{n}(\hat{\alpha}_n - \alpha_0) \rightsquigarrow \mathcal{N}(0, \sigma^2(F))$ where

$$\hat{\sigma}^2(F) = (12(\int_{-\infty}^{\infty} f^2(x)dx)^2)^{-1}.$$

Hint. In this case we would like to approximate the objective function $U_n(\alpha)$ by a quadratic function and then argue that the minimizer of $U_n(\alpha)$ has the same limiting distribution as the minimizer of the quadratic. To achieve this rewrite the objective function as

$$D_n(\delta) = U_n(\delta/\sqrt{n}) - U_n(0)$$

Note that if $\hat{\alpha}_n = \text{ argmin } U_n(\alpha)$, then $\hat{\delta}_n = \sqrt{n}\hat{\alpha}_n = \text{ argmin } D_n(\delta)$, and assume $\alpha_0 = 0$. Now, write $D_n(\delta)$ in terms of

$$R_{ij}(\delta) = \frac{1}{2}|(X_i + X_j) - 2\delta/\sqrt{n}| + \frac{1}{2}|(X_i + X_j)|$$

Use the identity,

$$|x - y| - |x| = -y \text{ sgn } (x) + 2\int_0^y (I(x < y) - I(x < 0))ds$$

to write,

$$R_{ij}(\delta) = -\frac{\delta}{\sqrt{n}} \text{ sgn } (X_i + X_j) + \int_0^{2\delta/\sqrt{n}} (I(X_i + X_j < s) - I(X_i + X_j < 0))ds$$

Now, compute the conditional expectations of both terms and complete the argument.

Finally consider the asymptotic relative efficiency of the Hodges-Lehmann estimator versus the sample mean,

$$\text{ARE } (\hat{\alpha}_n, \bar{X}_n) = 12\sigma^2(\int f^2(x)dx)^2,$$

which measures the limiting ratio of sample sizes required to achieve the same precision with the two estimators. The Hodges-Lehmann estimator is astonishingly good compared to $\bar{X}_n$, as the following table illustrates for a few representative distributions.

	Normal	Logistic	t_5	t_3	t_1
ARE $(\hat{\alpha}_n, X_n)$	.955	1.10	1.24	1.90	∞

(h.) Explain the table briefly from a robustness viewpoint.

(i.) A question that arises naturally from examination of the above table is: How bad can $\hat{\alpha}_n$ be relative to $\bar{X}_n$, i.e. what is the least favorable F? Surprisingly, this is a question that you have already answered on a previous problem set in a completely different context. Adapt your previous answer to the new context and interpret the solution. In particular, compute the lower bound for ARE $(\hat{\alpha}_n, \bar{X}_n)$ that you obtain.