Introduction to U statistics and UGEE

Jianghao Li

Duke University

December 2, 2016
Overview

1 U Statistics
   - One sample U-statistic
   - Inference on U statistics
   - Two sample and Multivariate U-statistic

2 Functional Response Models
   - Regression model
   - UGEE
Motivation

- U-statistic is an effective way of constructing unbiased estimators.
- For non-parametric families, U statistic usually produces uniformly minimum variance unbiased estimator (UMVUE).
  - By Lehmann-Scheff theorem, if T is an sufficient and complete statistics (c.s.s) and h(T) is unbiased estimator of \( \theta \), then h(T) is unique UMVUE of \( \theta \).
  - Ordered statistic is c.s.s for non-parametric family \( \mathcal{F}_r \), the class of distribution functions having finite \( r^{th} \) moment.
  - \( \phi(X_1, X_2, ..., X_n) \) is a function of ordered statistic iff it’s symmetric.
  - Thus, to find the UMVU estimator, it suffices to find a symmetric unbiased estimator.
One sample U statistics

- For distribution family $\mathcal{F}$ with parameter $\theta$ and a sample
  $\{X_1, X_2, ..., X_r\}$, $\theta$ is called **estimable** wrt to $\mathcal{F}$, if $\exists$ a statistics
  $h(X_1, X_2, ..., X_r)$ s.t.

  $$E_F h(X_1, X_2, ..., X_r) = \theta, \forall F \in \mathcal{F}$$

- $h(x_1, x_2, ..., x_r)$ is called **kernel function** of $\theta$
- WLOG, suppose $h(x_1, x_2, ..., x_r)$ is symmetric wrt $(x_1, x_2, ..., x_r)$
  - Otherwise, let

  $$h^*(x_1, x_2, ..., x_r) = \frac{1}{r!} \sum_{\alpha_1, ..., \alpha_r} h(X_{\alpha_1}, X_{\alpha_2}, ..., X_{\alpha_r})$$
Definition

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d sample with cdf $F(x)$. Estimable parameter $\theta$ kernel $h(X_1, X_2, \ldots, X_r)$. The U Statistics constructed by $h(X_1, X_2, \ldots, X_r)$ is:

$$U(X_1, X_2, \ldots, X_n) = \frac{1}{\binom{n}{r}} \sum_{\beta_1, \ldots, \beta_r} h(X_{\beta_1}, X_{\beta_2}, \ldots, X_{\beta_r})$$

Example

- $\theta = E(X_1)$, then $h(x) = x$,
  $$U(X_1, X_2, \ldots, X_n) = \frac{1}{\binom{n}{1}} \sum_{i=1}^{n} X_i = \bar{X}$$

- $\theta = Var(X_1)$, then $h(x_1, x_2) = \frac{1}{2} (x_1 - x_2)^2$,
  $$U(X_1, X_2, \ldots, X_n) = \frac{1}{\binom{n}{2}} \sum_{i<j}^{n} \frac{1}{2} (X_i - X_j)^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$
Var(U) = E \left[ \left( \binom{n}{r} \right)^{-1} \sum_{\beta_1, \ldots, \beta_r} (h(X_{\beta_1}, X_{\beta_2}, \ldots, X_{\beta_r}) - \theta) \right]^2 \\
= \left( \binom{n}{r} \right)^{-2} E \left[ \sum_{\beta_1, \ldots, \beta_r} (h(X_{\beta_1}, X_{\beta_2}, \ldots, X_{\beta_r}) - \theta) \right] \\
= \left( \binom{n}{r} \right)^{-2} \sum_{\beta_1, \ldots, \beta_r} \sum_{\alpha_1, \ldots, \alpha_r} \text{cov}[h(X_{\beta_1}, X_{\beta_2}, \ldots, X_{\beta_r}), h(X_{\alpha_1}, X_{\alpha_2}, \ldots, X_{\alpha_r})] \\
= \left( \binom{n}{r} \right)^{-2} \sum_{c=0}^{n} \binom{n}{r} \binom{n}{c} \binom{n-r}{n-c} \zeta_c \\
\text{where } \zeta_c = \text{cov}\{ h(X_1, \ldots, X_c, X_{c+1}, \ldots, X_r), h(X_1, \ldots, X_c, X_{r+1}, \ldots, X_{2r-c}) \}
General form of the variance of U statistics:

\[
Var(U) = \frac{1}{\binom{n}{r}} \sum_{c=1}^{r} \binom{r}{c} \binom{n-r}{r-c} \zeta_c
\]

\[n \Var(U) \to r^2 \zeta_1\]

Example

\[\theta = E(X_1), \text{ then } U(X_1, X_2, ..., X_n) = \bar{X},\]

\[
\Var(\bar{X}) = \frac{1}{\binom{n}{1}} \binom{1}{1} \binom{n-1}{1} \zeta_1 = \frac{\sigma^2}{n}\]
Projection

- U statistic is not in a form that is conducive to the usual treatment by direct applications of LLN and CLT.
- Basic idea to tackle the dependence of U statistics using projection
- Define \( \mathcal{V} = \{ V \mid V = \sum_{i=1}^{n} K(X_i), K(\cdot) \text{is a real function} \} \). Then for any symmetric statistics \( W(X_1, X_2, \ldots, X_n) \) with \( E(W) = 0 \), its projection onto \( \mathcal{V} \) can be proved as:

\[
W^* = \sum_{i=1}^{n} E\{ W | X_i \}
\]

- This is a projection in the sense that

\[
E(W - V^*)^2 \leq E(W - V)^2, \forall V \in \mathcal{V}
\]
Thus the projection of $U(X_1, X_2, ..., X_n) - \theta$ on $\mathcal{W}$ is:

$$U^* = \frac{r}{n} \sum_{i=1}^{n} \{ h^*(X_i) - \theta \}$$

where

$$h^*(X_i) = E(h(X_{\beta_1}, X_{\beta_2}, ..., X_{\beta_r}|X_{\beta_1} = X_i)$$

$$= E(h(X_1, X_2, ..., X_r)|X_1 = X_i)$$

$E[h^*(X_i)] = E(U) = \theta$, $\text{Var}[h^*(X_i)] = \zeta_1$ (when $n \geq 2r$)

**Theorem** (Hoeffding)

$$\sqrt{n}[U(X_1, X_2, ..., X_n) - \theta] = \sqrt{n}U^* + o_p(1) \rightarrow_d N(0, r^2\zeta_1)$$
Example: Wilcoxon signed rank test

- Wilcoxon signed rank test statistic for testing whether distribution is symmetric around 0:

\[
W_n^+ = \sum_{i=1}^{n} I\{X_i > 0\} R_i^+
\]

where \( R_i^+ \) is the rank of \(|X_i|\)

- Another way of expressing it:

\[
W_n^+ = \sum_{i=1}^{n} \sum_{j \leq i} I\{X_i + X_j \geq 0\}
\]

\[
= \sum_{i=1}^{n} I\{X_i \geq 0\} + \sum_{i=1}^{n} \sum_{j < i} I\{X_i + X_j \geq 0\}
\]
Example: Wilcoxon signed rank test

\[
\frac{1}{n(n-1)}W_n^+ = \frac{2}{n-1} \frac{1}{n} \sum_{i=1}^{n} I\{X_i \geq 0\} + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j<i} I\{X_i+X_j \geq 0\} \\
= \frac{2}{n-1} U_1 + U_2
\]

\[\therefore \sqrt{n} \left( \frac{W_n^+ - E(W_n^+)}{\binom{n}{2}} \right) = \sqrt{n}(U_2 - E(U_2)) + o_p(n^{-\frac{1}{2}})\]

\[\rightarrow \mathcal{N}(0, 4\zeta_1)\]

where \( \zeta_1 = \text{cov}(I\{X_1+X_2 \geq 0\}, I\{X_1+X_3 \geq 0\}) = \frac{1}{12} \) under the hypothesis that the distribution of \( X_i \) is symmetric around 0.
Suppose \( \{X_1, ..., X_m\} \) and \( \{Y_1, ..., Y_n\} \) are two independent samples, with cdf \( F(x) \) and \( G(y) \), respectively. For estimable \( \theta \) if we have kernel \( h(\cdot; \cdot) \) such that

\[
E_{(F,G)}(h(X_1, ..., X_r; Y_1, ..., Y_s)) = \theta
\]

\( h(\cdot; \cdot) \) is symmetric wrt \( \{X_1, ..., X_m\} \) and \( \{Y_1, ..., Y_n\} \) respectively. Then define U-statistic as:

\[
U(X_1, ..., X_m; Y_1, ..., Y_n) = \frac{1}{\binom{m}{r} \binom{n}{s}} \sum_{\beta_1, ..., \beta_r} \sum_{\alpha_1, ..., \alpha_s} h(X_{\beta_1}, ..., X_{\beta_r}; Y_{\alpha_1}, ..., Y_{\alpha_s})
\]
Inference for Two sample U statistic

For two independent samples \( \{X_1, \ldots, X_m\} \) and \( \{Y_1, \ldots, Y_n\} \), suppose \( \frac{m}{n+m} \to \lambda \) when \( N = n + m \to \infty \), then

\[
\sqrt{N}[U(X_1, \ldots, X_m; Y_1, \ldots, Y_n) - \theta] \to_d N \left( 0, \frac{r^2 \zeta_{1,0}}{\lambda} + \frac{s^2 \zeta_{0,1}}{1 - \lambda} \right)
\]

where

\[
\zeta_{c,d} = \text{cov}\left[ h(X_1, \ldots, X_c, X_{c+1}, \ldots, X_r; Y_1, \ldots, Y_d, Y_{d+1}, \ldots, Y_s), h(X_1, \ldots, X_c, X_{r+1}, \ldots, X_{2r-c}; Y_1, \ldots, Y_d, Y_{s+1}, \ldots, Y_{2s-d}) \right]
\]
Mann-Whitney-Wilcoxon rank sum test is used to test whether there is a location shift between two distributions. The statistics is written as:

\[ U = \sum_{i=1}^{m} \sum_{j=1}^{n} I\{X_i \leq Y_j\} \]

Then under null hypothesis that the two samples are identically distributed, we have asymptotic distribution:

\[ \sqrt{N}\left(\frac{1}{mn} U - \frac{1}{2}\right) \rightarrow N\left(0, \frac{1}{12\lambda} + \frac{1}{12(1 - \lambda)}\right) \]
Consider i.i.d sample of $p \times 1$ vector $X_i$, $h(X_1, \ldots, X_r)$ is a symmetric $l$ dimensional vector-valued function, then multivariate U-statistic for vector $\theta$ is:

$$U(X_1, X_2, \ldots, X_n) = \frac{1}{\binom{n}{r}} \sum_{\beta_1, \ldots, \beta_r} h(X_{\beta_1}, X_{\beta_2}, \ldots, X_{\beta_r})$$

Similarly define projection of $U - \theta$ as:

$$U^* = \sum_{i=1}^{n} E\{U|X_i\} - \theta = \frac{r}{n} \sum_{i=1}^{n} \{E(h|X_i) - \theta\}$$

Similarly we have asymptotic normality:

$$\sqrt{n}(U_n - \theta) \rightarrow_d N(0, m^2 \text{Var}(E(h|X_i)))$$
Functional responses model

- Functional response models are generalization of traditional regression models, through defining both the responses and predictors are functions.
- The class of distribution free model regression model for functional responses is defined by

$$ E[f(y_{i_1}, ..., y_{i_r})|x_{i_1}, ..., x_{i_r}] = h(x_{i_1}, ..., x_{i_r}; \theta) $$

where $f$ is a vector-valued function, $h$ is a smooth function with continuous derivatives up to second order.
- This FRM subsumes all classic regression model.
Motivating Example

Distribution-free regression model restricting for higher moments.

- When only restricting for the first moment, i.e. \( E(y_i) = \mu(x_i; \theta) \) where, GEE provides asymptotic efficient estimator when correlation is correctly specified.
- But sometimes, modeling for variance may also be of interest.
- E.g. test for overdispersion in Possion-like model.
- Solution in traditional framework: GEE II

\[
E \left( \frac{y_i}{(y_i - \mu(x_i; \theta))^2} \right) = \left( \frac{\mu(x_i; \theta)}{V(x_i; \theta)} \right)
\]
Motivating Example

- But GEE II is not a generalization of regression model, as the dependent variables involve unknown parameter.
- Alternatively, use FRM:

\[
E \left( \frac{1}{2} (y_i + y_j) \left( y_i - y_j \right)^2 \right) = \left( V(x_i; \theta) + V(x_j; \theta) + [\mu(x_i; \theta) - \mu(x_j; \theta)]^2 \right)
\]
For FRM

\[ E[f(y_{i1}, ..., y_{ir})|x_{i1}, ..., x_{ir}] = h(x_{i1}, ..., x_{ir}; \theta) \]

The U-statistic based generalized estimating equation is given by

\[ U_n(\theta) = \left( \begin{array}{c} n \\ r \end{array} \right)^{-1} \sum_{i \in C_r^n} G_i(x_i)(f_i - h_i) = 0 \]

where \( i = (i_1, ..., i_r) \), \( G_i(x_i) \) is some known \( p \times l \) matrix function, \( p \) is the dimension of \( \theta \) and \( l \) is the dimension of \( f_i \).

- Choice of \( G_i(x_i) \) is not unique and does not affect consistency of the UGEE estimate.
- Mostly set \( G_i(x_i) = D_i^T V_i^{-1} = \left( \frac{\partial h_i}{\partial \theta} \right)^T V_i^{-1} \)
Asymptotic features

Let $\Sigma = \text{Var}(E(G_i(x_i)(f_i - h_i)|y_{i1}, x_{i1}))$, $B = E(G_iD_i^T)$, then under mild regularity conditions, we have:

- $\hat{\theta}$ is consistent
- If $V$ is correctly specified, then $\theta$ is asymptotically normal:

$$\sqrt{n}(\hat{\theta} - \theta) \to_d N(0, q^2 B^{-1}\Sigma B^{-T})$$
Reference