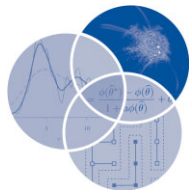
Focus Article



# Multivariate U-statistics: a tutorial with applications

Q. Yu[1], W. Tang[1], J. Kowalski[2] and X. M. Tu[1]*

U-statistics represent an important class of statistics arising from modeling quantities of interest defined by multi-subject responses such as the classic Mann–Whitney–Wilcoxon rank tests. However, classic applications of U-statistics are largely limited to univariate outcomes within a cross-sectional data setting. As longitudinal study designs become increasingly popular in today's research, it is imperative to generalize the classic theory of U-statistics to such a study setting to meet the challenges of modern clinical and translational research. In this article, we focus on applications of U-statistics to longitudinal study data. We first give a brief overview of U-statistics and then discuss how to apply this powerful class of statistics to model quantities of interest with a longitudinal data setting. In addition to generalizing U-statistics and associated inference theory to longitudinal data analysis, we also discuss a class of function response models (FRM) to bring the power of U-statistics to uncharted territory. We illustrate applications of generalized U-statistics and FRM with data from some real longitudinal studies.
 *WIREs Comp Stat* 2011 3 457–471 DOI: 10.1002/wics.178

## INTRODUCTION

Since Hoeffding's[1] foundational work, U-statistics have been widely used in both theoretical and applied statistical research. What is a U-Statistic and how is it different from the popular statistics such as the $t$ statistic? To illustrate, consider a simple example of an independently identically distributed (i.i.d.) sample $y_i$ with mean $\mu$ and variance $\sigma^2$. The sample mean $\overline{y}_n$ and variance $s_n^2$ given below are unbiased and consistent estimates of $\mu$ and variance $\sigma^2$:

$$\overline{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y}_n)^2. \qquad (1)$$

The sample mean is a sum of i.i.d. variables $y_i$. Although the sample variance is not in such a form, it can be expressed as an i.i.d. sum by applying a Taylor series expansion around $\mu^2$:

$$\begin{aligned} s_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\mu)^2 + \frac{n}{n-1}(\overline{y}_n-\mu)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\mu)^2 + o_p\left(\frac{1}{\sqrt{n}}\right), \end{aligned} \qquad (2)$$

where $o_p(\cdot)$ denotes stochastic $o(\cdot)$. For statistics and/or estimates that are in the form of an i.i.d. sum such as $\overline{y}_n$ and $s_n^2$, we can find their asymptotic distributions by applying the law of large numbers (LLN) and central limit theorem (CLT). However, the distinctive appearance of U-statistics makes it impossible to express them as an i.i.d. sum in either closed form or asymptotic approximation using a Taylor series expansion.

Many parameters and statistics of interest, such as the Mann–Whitney–Wilcoxon rank-sum and signed-rank tests[3,4] and Kendall's $\tau$,[5] cannot be defined by single-response-based statistical models. Furthermore, even for those that can be modeled by a single-subject response, the study of the asymptotic behavior of their associated statistics and estimates can be facilitated when formulated within the U-statistics setting.

*Correspondence to: xin_tu@urmc.rochester.edu

[1]Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, USA

[2]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Atlanta, GA, USA

DOI: 10.1002/wics.178

## UNIVARIATE U-STATISTICS

Consider $K$ i.i.d. samples of random vectors, $\mathbf{y}_{ki}$ $(1 \le i \le n_k, 1 \le k \le K)$. Let

$$h_{1i_1,\ldots,1i_{m_1};\ldots;Ki_1,\ldots,Ki_{m_K}} = h\left(\mathbf{y}_{1i_1},\ldots,\mathbf{y}_{1i_{m_1}};\ldots;\mathbf{y}_{Ki_1},\ldots,\mathbf{y}_{Ki_{m_K}}\right), \quad (3)$$

be some symmetric function with respect to input arguments within each sample, i.e., $h$ is constant when $\mathbf{y}_{ki_1},\ldots,\mathbf{y}_{ki_{m_k}}$ are permuted within each $k$th sample $(1 \le k \le K)$. A $K$-sample $U$-statistic with $m_k$ arguments for the $k$th sample has the general form:

$$U_n = \left[\prod_{k=1}^{K}\binom{n_k}{m_k}\right]^{-1}\sum_{k=1}^{K} \times \sum_{(i_1,\ldots,i_{m_k})\in C_{m_k}^{n_k}} h_{1i_1,\ldots,1i_{m_1};\ldots;Ki_1,\ldots,Ki_{m_K}}, \quad (4)$$

where $C_{m_k}^{n_k} = \{(ki_1,\ldots,ki_{m_k}); 1 \le ki_1 < \cdots < ki_{m_k} \le n_k\}$ denotes the set of all distinct combinations of $m_k$ indices $(i_1,\ldots,i_m)$ from the integer set $\{1,2,\ldots,n_k\}$. Note that the term *univariate* refers to the dimension of the kernel function $h$, and the arguments $\mathbf{y}$ in Eq. (3) can be a scaler or a vector.

Let $\theta = E(h_{11,\ldots,1m_1;\ldots;K1,\ldots,Km_K})$. Since

$$\begin{aligned} E(U_n) &= \left[\prod_{k=1}^{K}\binom{n_k}{m_k}\right]^{-1}\sum_{k=1}^{K} \times \sum_{(i_1,\ldots,i_{m_k})\in C_{m_k}^{n_k}} E(h_{1i_1,\ldots,1i_{m_1};\ldots;Ki_1,\ldots,Ki_{m_K}}) \\ &= \left[\prod_{k=1}^{K}\binom{n_k}{m_k}\right]^{-1}\sum_{k=1}^{K} \times \sum_{(i_1,\ldots,i_{m_k})\in C_{m_k}^{n_k}} E(h_{11,\ldots,1m_1;\ldots;K1,\ldots,Km_K}) \\ &= \theta, \end{aligned} \quad (5)$$

it follows that $U_n$ is an unbiased estimate of $\theta = E(h)$. For example, consider an i.i.d. sample $y_i$ with mean $\mu$ and variance $\sigma^2$, and let $h(y_i, y_j) = \frac{1}{2}(y_i - y_j)^2$. Then, $h(y_i, y_j)$ is invariant under permutation of $i$ and $j$. Furthermore, it is readily checked that

$$s_n^2 = \binom{n}{2}^{-1}\sum_{(i,j)\in C_2^n} h(y_i, y_j) = \frac{1}{2}(y_i - y_j)^2. \quad (6)$$

Since $E[\frac{1}{2}(y_i - y_j)^2] = \sigma^2$, $s_n^2$ is an unbiased estimate of $\sigma^2$.

To see an example where the argument is a vector, consider an i.i.d. sample of bivariate ordinal outcomes $\mathbf{z}_i = (u_i, v_i)^\top$, with $u(v)$ having $K(M)$ levels indexed by $k(m)$ $(1 \le i \le n)$. The Goodman and Kruskal $\gamma$ is a popular measure of association between two ordinal outcomes based on the notion of concordance and discordance.[6] For a pair of subjects $\mathbf{z}_i$ and $\mathbf{z}_j$, concordance and discordance are defined as:

$$(\mathbf{z}_i, \mathbf{z}_j) \equiv \begin{cases} \text{concordant} & \text{if } u_i < (>) u_j, v_i < (>) v_j \\ \text{discordant} & \text{if } u_i > (<) u_j,\ v_i < (>) v_j \\ \text{neither} & \text{if otherwise} \end{cases}.$$

Note that the 'neither' category has a positive probability because of the discrete nature of the outcomes. Goodman and Kruskal's $\gamma$ is defined as:

$$\begin{aligned} \gamma &= \frac{p_{\text{cp}} - p_{\text{dp}}}{p_{\text{cp}} + p_{\text{dp}}}, \\ p_{\text{cp}} &= \Pr((u_i - u_j)(v_i - v_j) > 0), \\ p_{\text{dp}} &= \Pr((u_i - u_j)(v_i - v_j) < 0). \end{aligned} \quad (7)$$

Let $h(\mathbf{z}_i, \mathbf{z}_j) = I_{\{(u_i-u_j)(v_i-v_j)>0\}} - I_{\{(u_i-u_j)(v_i-v_j)<0\}}$. Then, the mean of $h(\mathbf{z}_i, \mathbf{z}_j)$ is

$$\delta = E(h(\mathbf{z}_i, \mathbf{z}_j)) = p_{\text{cp}} - p_{\text{dp}}.$$

In addition, since

$$h(\mathbf{z}_i, \mathbf{z}_j) = \begin{cases} 1 & \text{if } \mathbf{z}_i \text{ and } \mathbf{z}_j \text{ are concordant} \\ -1 & \text{if } \mathbf{z}_i \text{ and } \mathbf{z}_j \text{ are discordant} \\ 0 & \text{if otherwise} \end{cases}, \quad (8)$$

it follows that the U-statistic based on $h(\mathbf{z}_i, \mathbf{z}_j)$ has the following form:

$$U_n = \binom{n}{2}^{-1}\sum_{(i,j)\in C_2^n} h(\mathbf{z}_i, \mathbf{z}_j) = \binom{n}{2}^{-1}(C - D), \quad (9)$$

where $C$ and $D$ denote the number of concordant and discordant pairs in the sample, respectively. Under the null of no association between $u$ and $v$, $\delta = p_{\text{cp}} - p_{\text{dp}} = 0$ and the U-statistic can be used to test this null.

Note that although both the sample variance $s_n^2$ in Eq. (6) and the sample version of the numerator of Goodman and Kruskal's $\gamma$ in Eq. (9) are expressed as a U-statistic, there is a fundamental difference between the two. As the concept of concordance or discordance requires two subjects' responses, it is impossible to reexpress Eq. (9) as an i.i.d. sum as in the case of

$s_n^2$ in Eq. (1). Another example of such an 'intrinsic' U-statistic is the popular Mann–Whitney–Wilcoxon rank statistics. In the Section *Functional Response Model*, we present a modern-day example of such multi-subject-response-defined models within the setting of regression analysis.

## MULTIVARIATE U-STATISTICS

Many problems of interest in modern statistical applications involve second and even higher-order moments, which are often defined by multiple correlated outcomes of different constructs and/or dimensions. Thus, even for cross-sectional data with a single snapshot, such models involve multiple parameters (representing different orders of moments of one variable or the same order moment of different variables) and correlated outcomes. On the other hand, with repeated measures in the longitudinal study, a single variable may require modeling multiple parameters to capture the changes of the variable over time. Although standard methods of moments may sometimes be used to model the quantity of interest, it is at best cumbersome.[2,7,8] Thus, it is important to extend univariate U-statistics to a multivariate setting to facilitate their applications in such a wider context involving multiparameter-defined models.

For example, we discussed how the sample variance $s_n^2$ of an i.i.d. sample $y_i$ can be expressed as a U-statistic. In many applications, the mean $\mu$ of $y_i$ is also of interest. Furthermore, we may be interested in comparing the mean $\mu$ and variance $\sigma^2$. For example, if $y_i$ is a count response, it is often of interest to see if $y_i$ is over dispersed, i.e., $\sigma^2 > \mu$.[2,9–11] If this is the case, the Poisson will be a poor model for $y_i$, and other alternatives such as the negative binomial model must be used to describe the distribution of $y_i$.[2] To formally compare $\sigma^2$ and $\mu$, we must make joint inference about $\mu$ and $\sigma^2$, yielding a bivariate U-statistic.

As another example, consider the Goodman and Kruskal $\gamma$ again. We constructed a U-statistic to test the null of no association between two ordinal outcomes based on the numerator of this index in the Section *Univariate U-Statistics*. If inference about $\gamma$ is of interest, we may also construct a U-statistic to estimate the denominator $p_{\text{cp}} + p_{\text{dp}}$ so that we can integrate the two using the Delta method to provide inference about $\gamma$.

Applications of U-statistics to longitudinal data also require the extension of univariate U-statistics to a multivariate setting. Consider a longitudinal study with $n$ subjects and $m$ assessment times. Let $y_{it}$ denote the response from the $i$th subject at time $t$ ($1 \le i \le n$, $1 \le t \le m$). We are interested in modeling the variance of the response $y_{it}$ over time. Let

$$\sigma_t^2 = \text{Var}(y_{it}), \quad \mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{im})^\top,$$
$$\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)^\top. \tag{10}$$

A primary interest in longitudinal studies is trend over time. Within the current context, we are interested in testing the variance homogeneity over time, that is

$$H_0 : \sigma_t^2 = \sigma^2, \quad \text{for all } 1 \le t \le m. \tag{11}$$

For each $t$, the U-statistic in Eq. (6) is an unbiased estimate of $\sigma_t^2$. Inference about the above hypothesis, however, requires jointly modeling $\sigma_t^2$ across the times.

The generalization of univariate U-statistics to a multivariate setting is straightforward. Consider $K$ i.i.d. samples of random vectors, $\mathbf{y}_{ki}$ ($1 \le i \le n_k$, $1 \le k \le K$). Let

$$\mathbf{h}_{11,\ldots,1m_1;\ldots;K1,\ldots,Km_K} = \mathbf{h}(\mathbf{y}_{11}, \ldots, \mathbf{y}_{1m_1}; \ldots; \mathbf{y}_{K1}, \ldots, \mathbf{y}_{Km_K}),$$

be some symmetric vector-valued function with respect to the input arguments within each sample, i.e., $\mathbf{h}$ is constant when $\mathbf{y}_{k1}, \ldots, \mathbf{y}_{km_k}$ are permuted within each $k$th sample ($1 \le k \le K$). A $K$-sample $U$-statistic with $m_k$ arguments for the $k$th sample has the general form:

$$\mathbf{U}_n = \left[\prod_{k=1}^{K} \binom{n_k}{m_k}\right]^{-1} \sum_{k=1}^{K} \times \sum_{(i_1,\ldots,i_{m_k}) \in C_{m_k}^{n_k}} \mathbf{h}_{1i_1,\ldots,1i_{m_1};\ldots;Ki_1,\ldots,Ki_{m_K}}. \tag{12}$$

As in the one-sample case, $\mathbf{U}_n$ is an unbiased estimate of $\boldsymbol{\theta} = E(\mathbf{h})$.

To apply Eq. (12) to test the null in Eq. (11), let $\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)^\top$, and consider the multivariate U-statistic:

$$\widehat{\boldsymbol{\theta}} = \binom{n}{2}^{-1} \sum_{(i,j) \in C_2^n} \mathbf{h}(\mathbf{y}_i, \mathbf{y}_j) = \binom{n}{2}^{-1} \sum_{(i,j) \in C_2^n} \frac{1}{2} \begin{pmatrix} (y_{i1} - y_{j1})^2 \\ \vdots \\ (y_{im} - y_{jm})^2 \end{pmatrix}. \tag{13}$$

It follows that $\widehat{\boldsymbol{\theta}}$ is an unbiased estimate of $\boldsymbol{\theta}$.

As an example within a cross-sectional data setting, consider again Goodman and Kruskal's $\gamma$ in Eq. (7). Let $\mathbf{h}(\mathbf{z}_i, \mathbf{z}_j)$ in Eq. (12) be defined as, $\mathbf{h}(\mathbf{z}_i, \mathbf{z}_j) = (h_1(\mathbf{z}_i, \mathbf{z}_j), h_2(\mathbf{z}_i, \mathbf{z}_j))^\top$, where $h_1(\mathbf{z}_i, \mathbf{z}_j)$ is given in Eq. (8) and $h_2(\mathbf{z}_i, \mathbf{z}_j)$ is defined by:

$$h_2(\mathbf{z}_i, \mathbf{z}_j) = \begin{cases} 1 & \text{if } \mathbf{z}_i \text{ and } \mathbf{z}_j \text{ are concordant} \\ & \quad \text{or discordant} \\ 0 & \text{if otherwise} \end{cases}. \quad (14)$$

Then, the resulting $\widehat{\boldsymbol{\theta}} = \mathbf{U}_n$ is an unbiased estimate of $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top = (p_{\text{cp}} - p_{\text{dp}}, p_{\text{cp}} + p_{\text{dp}})^\top$. It follows that $\widehat{\gamma} = (\widehat{\theta}_1/\widehat{\theta}_2)$ is an estimate of $\gamma$. To apply the Delta method for inference about $\gamma$, we need to find the asymptotic distribution of $\widehat{\boldsymbol{\theta}}$, which we discuss next.

## INFERENCE FOR U-STATISTICS

### Complete Data

A major distinction of the U-statistic is its signature form involving a symmetric summation of dependent variables (univariate U-statistics) or vectors (multivariate U-statistics). As the dependence structure deviates from the usual independence assumption in conventional statistics in the form of sum of i.i.d. random variables or vectors, standard asymptotic methods such as LLN and CLT do not apply directly to determine the limiting distribution of U-statistics. The basic idea behind tackling the dependence of U-statistic is to create a random quantity, called *projection*, in the usual form of a sum of i.i.d. random variables or vectors that approximate the original U-statistic so well to the degree that the two become indistinguishable asymptotically, i.e., they have the same asymptotic distribution. This allows one to study the large sample behavior of U-statistic via its projection counterpart using the familiar asymptotic methods.

Consider the one-sample univariate U-statistic with $m$ arguments:

$$U_n = \binom{n}{m}^{-1} \sum_{(i_1,\ldots,i_m)\in C_m^n} h(\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_m}). \quad (15)$$

The projection $\widehat{U}_n$ of $U_n$ is defined as follows:

$$\widehat{U}_n = \sum_{i=1}^{n} E(U_n \mid \mathbf{y}_i) - (n-1)\theta,$$
$$\theta = E[h(\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_m})]. \quad (16)$$

If $\mathbf{y}_i$ $(1 \le i \le n)$ is an i.i.d. sample, so is $E(U_n \mid \mathbf{y}_i)$, as the latter is a function of $\mathbf{y}_i$. Thus, $\widehat{U}_n$ is a sum of i.i.d. random variables and standard asymptotic methods such as LLN and CLT can be applied to find its limiting distribution.

Note that the second term in Eq. (16) is a normalizing factor so that $E(\widehat{U}_n) = \theta$. Note also that unlike $U_n$, $\widehat{U}_n$ is not a statistic since $E(U_n \mid \mathbf{y}_i)$ is generally a function of $\theta$.

Consider the U-statistic for the sample variance in Eq. (6). It is readily checked that

$$E(h(y_j, y_k) \mid y_i) = h_1(y_i) = \begin{cases} \frac{1}{2}(y_i^2 - 2\mu y_i + E(y_i^2)) & \text{if } i \in \{j,k\} \\ \sigma^2 & \text{if } i \notin \{j,k\} \end{cases}.$$

Let $D_{jk} = \{(l,m) \in C_2^n; l \ne i, m \ne i\}$. It follows that

$$\begin{aligned} E(s_n^2 \mid y_i) &= \binom{n}{2}^{-1}\Bigg[\sum_{j\ne i} E(h(y_j, y_i) \mid y_i) \\ &\qquad + \sum_{(j,k)\in D_{jk}} E(h(y_j, y_k) \mid y_i)\Bigg] \\ &= \binom{n}{2}^{-1}\left[\binom{n-1}{1} h_1(y_i) + \binom{n-1}{2}\sigma^2\right] \\ &= \frac{1}{n}(y_i^2 - 2\mu y_i + E(y_i^2)) + \frac{n-2}{n}\sigma^2. \end{aligned}$$

Thus,

$$\begin{aligned} \widehat{U}_n &= \sum_{i=1}^{n}\left[\frac{1}{n}(y_i^2 - 2\mu y_i + E(y_i^2)) + \frac{n-2}{n}\sigma^2\right] \\ &\quad - (n-1)\sigma^2 \\ &= \frac{2}{n}\sum_{i=1}^{n}\frac{1}{2}(y_i^2 - 2\mu y_i + E(y_i^2)) - \sigma^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2. \end{aligned}$$

As $\widehat{U}_n$ is a function of $\mu$, it is not a statistic. Also, $\widehat{U}_n$ is exactly the first term in Eq. (2). Thus, $\widehat{U}_n$ differs from $U_n$ by a higher-order term $o_p(1/\sqrt{n})$, and the two have the same asymptotic distribution.

For a general $K$ sample multivariate U-statistic in Eq. (12), the projection of $\mathbf{U}_n$ is defined as:

$$\begin{aligned} \widehat{\mathbf{U}}_n &= \sum_{k=1}^{K}\sum_{i=1}^{n_k} E(\mathbf{U}_n \mid \mathbf{y}_{ki}) \\ &= \sum_{k=1}^{K}\sum_{i=1}^{n_k}[E(\mathbf{U}_n \mid \mathbf{y}_{ki}) - \boldsymbol{\theta}] + \boldsymbol{\theta}. \end{aligned}$$

If we center $\widehat{\mathbf{U}}_n$ at $\boldsymbol{\theta} = E(\mathbf{U}_n)$, we have:

$$\begin{aligned}\widehat{\mathbf{U}}_n - \boldsymbol{\theta} &= \sum_{k=1}^{K} \frac{m_k}{n_k} \sum_{i=1}^{n_k} (\mathbf{h}_{k1}(\mathbf{y}_{ki}) - \boldsymbol{\theta}) \\ &= \sum_{k=1}^{K} \frac{m_k}{n_k} \sum_{i=1}^{n_k} \widetilde{\mathbf{h}}_{k1}(\mathbf{y}_{ki}), \end{aligned} \tag{17}$$

where

$$\mathbf{h}_{k1}(\mathbf{y}_{k1}) = E[\mathbf{h}(\mathbf{y}_{11}, \ldots, \mathbf{y}_{1m_1}; \ldots; \mathbf{y}_{K1}, \ldots, \mathbf{y}_{Km_K}) \mid \mathbf{y}_{k1}],$$
$$\widetilde{\mathbf{h}}_{k1}(\mathbf{y}_{ki}) = \boldsymbol{\theta}, \quad 1 \le k \le K.$$

Consider testing the null of variance homogeneity between two groups within a longitudinal setting. For notational brevity, assume a pre–post setting and let

$$\begin{aligned}&\mathbf{h}(\mathbf{y}_{1i}, \mathbf{y}_{1j}; \mathbf{y}_{2l}, \mathbf{y}_{2m}) \\ &= \begin{pmatrix} \frac{1}{2}(y_{1i1} - y_{1j1})^2 \\ \frac{1}{2}(y_{1i2} - y_{1j2})^2 \end{pmatrix} - \begin{pmatrix} \frac{1}{2}(y_{2l1} - y_{2m1})^2 \\ \frac{1}{2}(y_{2l2} - y_{2m2})^2 \end{pmatrix}. \end{aligned} \tag{18}$$

Then, it follows that

$$\boldsymbol{\theta} = E[\mathbf{h}(\mathbf{y}_{1i}, \mathbf{y}_{1j}; \mathbf{y}_{2l}, \mathbf{y}_{2m})] = \begin{pmatrix} \sigma_{11}^2 - \sigma_{21}^2 \\ \sigma_{12}^2 - \sigma_{22}^2 \end{pmatrix}.$$

Under the null $H_0 : \sigma_{11}^2 - \sigma_{21}^2 = 0, \sigma_{12}^2 - \sigma_{22}^2 = 0$, $\boldsymbol{\theta} = \mathbf{0}$. Thus, if the null $H_0 : \boldsymbol{\theta} = \mathbf{0}$ holds true, the two groups have the same variance at both pre- and post-assessments.

Let $\mu_{kt} = E(y_{1it})$. Since

$$\begin{aligned}\mathbf{h}_{11}(\mathbf{y}_{11}) &= E(\mathbf{h}(\mathbf{y}_{11}, \mathbf{y}_{12}; \mathbf{y}_{21}, \mathbf{y}_{22}) \mid \mathbf{y}_{11}) \\ &= \begin{pmatrix} \frac{1}{2}(y_{111} - \mu_{11})^2 + \frac{1}{2}\sigma_{11}^2 - \sigma_{21}^2 \\ \frac{1}{2}(y_{112} - \mu_{12})^2 + \frac{1}{2}\sigma_{12}^2 - \sigma_{22}^2 \end{pmatrix}, \end{aligned}$$

it follows that

$$\begin{aligned}\widetilde{\mathbf{h}}_{11}(\mathbf{y}_{11}) &= \mathbf{h}_{11}(\mathbf{y}_{11}) - \boldsymbol{\theta} \\ &= \begin{pmatrix} \frac{1}{2}\left[(y_{111} - \mu_{11})^2 - \sigma_{11}^2\right] \\ \frac{1}{2}\left[(y_{112} - \mu_{12})^2 - \sigma_{12}^2\right] \end{pmatrix}. \end{aligned}$$

Likewise,

$$\begin{aligned}\widetilde{\mathbf{h}}_{21}(\mathbf{y}_{21}) &= \mathbf{h}_{21}(\mathbf{y}_{21}) - \boldsymbol{\theta} \\ &= \begin{pmatrix} \frac{1}{2}[(y_{211} - \mu_{21})^2 - \sigma_{21}^2] \\ \frac{1}{2}[(y_{212} - \mu_{22})^2 - \sigma_{22}^2] \end{pmatrix}. \end{aligned}$$

Hence, the centered projection in Eq. (17) for this particular example is

$$\widehat{\mathbf{U}}_n - \boldsymbol{\theta} = 2\frac{1}{n_1} \sum_{i=1}^{n_k} \widetilde{\mathbf{h}}_{11}(\mathbf{y}_{11}) + \frac{1}{n_2} \sum_{i=1}^{n_k} \widetilde{\mathbf{h}}_{21}(\mathbf{y}_{21}). \tag{19}$$

Since $\widehat{\mathbf{U}}_n - \boldsymbol{\theta}$ is in the form of i.i.d. sum, we can apply CLT to find the asymptotic distribution of the projection $\widehat{\mathbf{U}}_n$. Since $\widehat{\mathbf{U}}_n$ and $\mathbf{U}_n$ have the same asymptotic distribution, this is also the asymptotic distribution of $\mathbf{U}_n$.

For the general U-statistic defined in Eq. (12), let $n = \sum_{k=1}^{K} n_k$ and assume that $\lim_{n\to\infty}(n/n_k) = \rho_k^2 < \infty$ $(1 \le k \le K)$. Let

$$\begin{aligned}\mathbf{h}_{k1}(\mathbf{y}_{k1}) &= E(\mathbf{h} \mid \mathbf{y}_{k1}), \quad \widetilde{\mathbf{h}}_{k1}(\mathbf{y}_{k1}) = \mathbf{h}_{k1}(\mathbf{y}_{k1}) - \boldsymbol{\theta} \\ \Sigma_{h_k} &= \mathrm{Var}\left[\widetilde{\mathbf{h}}_{k1}(\mathbf{y}_{k1})\right] \\ &= E\left[\widetilde{\mathbf{h}}_{k1}(\mathbf{y}_{k1})\widetilde{\mathbf{h}}_{k1}^{\top}(\mathbf{y}_{k1})\right], \quad 1 \le k \le K. \end{aligned}$$

Under mild regularity conditions, we have:

$$\begin{aligned}&\mathbf{U}_n \to_p \boldsymbol{\theta}, \quad \sqrt{n}(\mathbf{U}_n - \boldsymbol{\theta}) \to_d N \\ &\times \left(0, \Sigma_U = \sum_{k=1}^{K} \rho_k^2 m_k^2 \Sigma_{h_k}\right), \end{aligned} \tag{20}$$

where $\to_p$ $(\to_d)$ denotes convergence in probability (distribution).

For the variance homogeneity example in Eq. (18),

$$\Sigma_{h_k} = \frac{1}{4} \begin{pmatrix} \mathrm{Var}((y_{k11} - \mu_{k1})^2) & \mathrm{Cov}((y_{k11} - \mu_{k1})^2, \\ & (y_{k12} - \mu_{k2})^2) \\ & \mathrm{Var}((y_{k12} - \mu_{k2})^2) \end{pmatrix}.$$

We can readily estimate the entries using the respective sample moments. With an estimated $\widehat{\Sigma}_{h_k}$, we estimate $\Sigma_U$ by: $\Sigma_U = 4\sum_{k=1}^{2}(n_1 + n_2/n_k)\widehat{\Sigma}_{h_k}$.

We may also construct a U-statistic to estimate each $\Sigma_{h_k}$ without attempting to evaluate these variance matrices. For example, by expressing $\Sigma_{h_1}$ as:

$$\begin{aligned}\Sigma_{h_1} &= E[E(\mathbf{h}(y_{11}, y_{12}, y_{21}, y_{22}) \mid y_{11})]^2 - \boldsymbol{\theta}\boldsymbol{\theta}^{\top} \\ &= \boldsymbol{\Phi}_{h_1} - \boldsymbol{\theta}\boldsymbol{\theta}^{\top} \end{aligned}$$

we can estimate $\boldsymbol{\theta}\boldsymbol{\theta}^{\top}$ by $\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^{\top}$. To estimate $\boldsymbol{\Phi}_{h_1}$, note that

$$\begin{aligned}\boldsymbol{\Phi}_{h_1} &= E[E(\mathbf{h}(\mathbf{y}_{11}, \mathbf{y}_{12}; \mathbf{y}_{21}, \mathbf{y}_{22}) \mid \mathbf{y}_{11}) \\ &\quad \times E(\mathbf{h}(\mathbf{y}_{11}, \mathbf{y}_{13}, \mathbf{y}_{23}, \mathbf{y}_{24}) \mid \mathbf{y}_{11})] \end{aligned}$$

$$= E\{E[\mathbf{h}(\mathbf{y}_{11}, \mathbf{y}_{12}; \mathbf{y}_{21}, \mathbf{y}_{22}) \times \mathbf{h}(\mathbf{y}_{11}, \mathbf{y}_{13}, \mathbf{y}_{23}, \mathbf{y}_{24}) \mid \mathbf{y}_{11}]\}$$
$$= E[\mathbf{g}(\mathbf{y}_{11}, \mathbf{y}_{12}, \mathbf{y}_{13}, \mathbf{y}_{21}, \mathbf{y}_{22}, \mathbf{y}_{23}, \mathbf{y}_{24})].$$

Let $\mathbf{g}_{123;1234} = \mathbf{g}(\mathbf{y}_{11}, \mathbf{y}_{12}, \mathbf{y}_{13}, \mathbf{y}_{21}, \mathbf{y}_{22}, \mathbf{y}_{23}, \mathbf{y}_{24})$ denote a symmetric kernel of

$$\mathbf{h}(\mathbf{y}_{11}, \mathbf{y}_{12}; \mathbf{y}_{21}, \mathbf{y}_{22})\mathbf{h}(\mathbf{y}_{11}, \mathbf{y}_{13}, \mathbf{y}_{23}, \mathbf{y}_{24}).$$

Then, the U-statistic matrix,

$$\widehat{\mathbf{\Phi}}_{h_1} = \left[\binom{n_1}{3}\binom{n_2}{4}\right]^{-1} \sum_{(i_1,i_2,i_3)\in C_3^{n_1}} \times \sum_{(j_1,j_2,j_3,j_4)\in C_4^{n_2}} \mathbf{g}_{i_1 i_2 i_3; j_1 j_2 j_3 j_4},$$

is a consistent estimate of $\mathbf{\Phi}_{h_1}$. This alternative is especially effective when the asymptotic variance cannot be evaluated analytically.

## Incomplete Data

In longitudinal studies, missing data is a common and persistent problem. Under the missing completely at random (MCAR) assumption, the occurrence of missing data is independent of the outcome, and thus estimates based on the observed data are still consistent.[2,12] Although such a complete-data approach still provides valid inference, it is inefficient. For example, if the occurrence of missing data is random over time and across all subjects, the number of subjects with complete data may be substantially smaller than that of the original sample.

Missing data arising in many longitudinal trials is the result of treatment response and thus is typically predictable by the observed outcomes. For example, in treatment studies, missing data may occur if a patient feels that he/she has responded to the treatment and does not see any additional benefit for continuing the treatment, and thus decides to stop the treatment. Such an outcome-dependent dropout process, or missing at random (MAR) mechanism, no longer fits the MCAR assumption. Consequently, estimates based on the complete-data subsample are not only inefficient but also inconsistent (biased) as well.

To illustrate, consider again the model for variance over time in a longitudinal setting in Eq. (10). Define a set of missing (or rather observed) data indicators for each subject as follows:

$$r_{it} = \begin{cases} 1 & \text{if } y_{it} \text{ is observed} \\ 0 & \text{if } y_{it} \text{ is missing} \end{cases}, \quad \mathbf{r}_i = (r_{i1}, \ldots, r_{im})^\top. \tag{21}$$

Then, under MCAR,

$$\widehat{\sigma}_t^2 = \frac{1}{\sum_{(i,j)\in C_2^n} r_{it} r_{jt}} \sum_{(i,j)\in C_2^n} r_{it} r_{jt} h_t(\mathbf{y}_i, \mathbf{y}_j)$$
$$\to_p \frac{E[\frac{1}{2} r_{it} r_{jt} (y_{it} - y_{jt})^2]}{E^2(r_{it})}$$
$$= \frac{E^2(r_{it})\sigma_t^2}{E^2(r_{it})}$$
$$= \sigma_t^2, \tag{22}$$

where $h_t(\mathbf{y}_i, \mathbf{y}_j) = \frac{1}{2}(y_{it} - y_{jt})^2$. Thus, the U-statistic in Eq. (22) is a consistent estimate of $\sigma_t^2$.

Under MAR, however, $r_{it}$ and $y_{it}$ are dependent, and it follows that in general

$$E\left[\frac{1}{2} r_{it} r_{jt} (y_{it} - y_{jt})^2\right] \neq E^2(r_{it})\sigma_t^2.$$

In other words, $\widehat{\sigma}_t^2$ in Eq. (22) may not be consistent under MAR.

For each study dropout, if we can 'recover' all the missing responses $y_{it}$ and put them back in the sample, we can complete the missing values and obtain a consistent estimate of $\sigma_t^2$. This hypothetical scenario is, of course, unrealistic in most studies, but it underscores the basic idea of constructing consistent estimates by statistically 'recovering' such missing $y_{it}$.

Let $\pi_{it} = Pr(r_{it} = 1 \mid \mathbf{y}_i)$ denote the probability of observing the response $y_{it}$ given $\mathbf{y}_i$. For each $i$th subject $i$, $\pi_{it}$ is the probability of observing the response $y_{it}$ at time $t$. Thus, if $y_{it}$ is observed, this observation actually represents a subgroup of $1/(\pi_{i2})$ subjects whose outcomes are not observed at time $t$. Now consider the following revised estimate of $\sigma_t^2$:

$$\widehat{\sigma}_t^2 = \binom{n}{2}^{-1} \sum_{(i,j)\in C_2^n} \frac{r_{it} r_{jt}}{\pi_{it}\pi_{it}} h_t(\mathbf{y}_i, \mathbf{y}_j)$$
$$= \binom{n}{2}^{-1} \sum_{(i,j)\in C_2^n} g_t(\mathbf{y}_i, \mathbf{y}_j), \tag{23}$$

where $\mathbf{g}_{ij} = (g_t(\mathbf{y}_i, \mathbf{y}_j), \ldots, g_t(\mathbf{y}_i, \mathbf{y}_j))^\top$. In the above, each observed $y_{it}$ is weighted by $1/(\pi_{it})$ (or $1/(\pi_{it}\pi_{jt})$ for the pair $y_{it}$ and $y_{jt}$), and as a result the term $(r_{it}r_{jt})/(\pi_{it}\pi_{it})h_t(\mathbf{y}_i, \mathbf{y}_j)$ represents the contributions of all the subjects in the subgroups symbolized by $y_{it}$ and $y_{jt}$. Thus, even though none of the subjects in these subgroups is observed except for $y_{it}$ and $y_{jt}$, their presence is acknowledged in the estimate $\widehat{\sigma}_t^2$ in Eq. (23). This in effect 'recovers' the missing

data in a statistical fashion. This fact is also readily demonstrated formally since

$$\begin{aligned} E(\widehat{\sigma}_t^2) &= E\left[E\left(\frac{r_{it}r_{jt}}{\pi_{it}\pi_{jt}}h_t(y_{it},y_{jt}) \mid y_{it},y_{jt}\right)\right] \\ &= E\left[\pi_{it}^{-1}\pi_{jt}^{-1}h_t(y_{it},y_{jt})E(r_{it} \mid y_{jt})E(r_{jt} \mid y_{jt})\right] \\ &= \sigma_t^2. \end{aligned}$$

Thus, the U-statistic $\widehat{\boldsymbol{\theta}} = \binom{n}{2}^{-1}\sum_{(i,j)\in C_2^n}\mathbf{g}_{ij}$ is a consistent estimate of $\boldsymbol{\theta} = (\sigma_1^2,\ldots,\sigma_m^2)^\top$.

This inverse probability weighting (IPW) technique is the basis for the popular weighted generalized estimating equations (WGEE) to provide inference for distribution-free longitudinal regression and related models under MAR.[2,13–16] We discuss another important application of IPW within our context upon introducing the functional response models in the Section *Functional Response Model*.

While $\pi_{it}$ may be known in some study designs (e.g., deliberately discontinuing some subjects in a multiphase clinical trial), the relationship between $\mathbf{r}_i$ and $\mathbf{y}_i$ is unknown in most applications. It is generally difficult to model $\pi_{it}$ because of the large number and complexity of missing data patterns.[12,13] For longitudinal data analysis, a popular assumption is the monotone missing data pattern (MMDP). The MMDP structure eliminates a potentially large number of missing data patterns, making it feasible to model and estimate $\pi_{it}$ under MAR.

Under MMDP, if $y_{it}$ is observed at time $t$, then all $y_{is}$'s prior to time $t$ ($s < t$) are also observed. Let

$$\widetilde{\mathbf{y}}_{it} = (y_{i1},\ldots,y_{i(t-1)})^\top, \quad 2 \le t \le m.$$

Then, $\widetilde{\mathbf{y}}_{it}$ contains all the observed data prior to time $t$. Thus, under MAR,

$$\pi_{it} = \Pr(r_{it} = 1 \mid \mathbf{y}_i) = \Pr(r_{it} = 1 \mid \widetilde{\mathbf{y}}_{it}).$$

In other words, $\pi_{it}$ is a function of observed data only, making it possible to estimate these selection probabilities.

We first model the one-step transition probability to observe the response at $t$ given the observed status at the immediate prior time $t-1$, $p_{it} = \Pr(r_{it} = 1 \mid r_{i(t-1)} = 1, \widetilde{\mathbf{y}}_{it})$, using logistic regression:

$$\text{logit}(p_{it}) = \alpha_t + \boldsymbol{\beta}_t^\top\widetilde{\mathbf{y}}_{it}, \quad 2 \le t \le m. \tag{24}$$

Let $\boldsymbol{\gamma}_t = (\alpha_t, \boldsymbol{\beta}_t^\top)^\top$ denote the parameters of the logistic model above, and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_2^\top,\ldots,\boldsymbol{\gamma}_m^\top)^\top$ the collection of all these $\boldsymbol{\gamma}_t$'s. We can estimate $\boldsymbol{\gamma}$ using either maximum likelihood or estimating equations.[9]

Regardless of the approach taken, we can express the estimate $\widehat{\boldsymbol{\gamma}}$ for the set of logistic models in Eq. (24) as the solution to the following estimating equations:

$$\begin{aligned} &\sum_{i=1}^n \mathbf{w}_i(\boldsymbol{\gamma}) = \sum_{i=1}^n(\mathbf{w}_{i2}^\top,\ldots,\mathbf{w}_{im}^\top)^\top = 0, \quad 2 \le t \le m, \\ &\mathbf{w}_{it} = \frac{\partial}{\partial\boldsymbol{\gamma}_t}\{r_{i(t-1)}[r_{it}\log(p_{it}) \\ &\qquad + (1-r_{it})\log(1-p_{it})]\}. \end{aligned} \tag{25}$$

Now by invoking MMDP, we obtain:

$$\begin{aligned} \pi_{it}(\boldsymbol{\gamma}) &= \Pr(r_{i(t-1)} = 1, r_{it} = 1 \mid \widetilde{\mathbf{y}}_{i(t-1)}, \widetilde{\mathbf{y}}_{it}) \\ &= \Pr(r_{it} = 1 \mid r_{i(t-1)} = 1, \widetilde{\mathbf{y}}_{it}) \\ &\quad \times \Pr(r_{i(t-1)} = 1 \mid \widetilde{\mathbf{y}}_{i(t-1)}) \\ &= \prod_{s=2}^t p_{is}(\boldsymbol{\gamma}_s), \quad 2 \le t \le m, \quad 1 \le i \le n. \end{aligned} \tag{26}$$

The relationship above allows us to estimate $\pi_{it}(\boldsymbol{\gamma})$ from the estimates of the one-step transition probabilities $p_{it}$.

By estimating $\pi_{it}(\boldsymbol{\gamma})$ using $\pi_{it}(\widehat{\boldsymbol{\gamma}})$ and substituting the estimates, we obtain from Eq. (23) a consistent estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. However, the asymptotic variance in Eq. (20) may not correctly estimate the variation of $\widehat{\boldsymbol{\theta}}$, because it does not reflect the additional variability of estimated $\widehat{\boldsymbol{\gamma}}$. To account for this extra variability, we must revise Eq. (20) to include the variation of the estimate $\widehat{\boldsymbol{\gamma}}$. It can be shown that the asymptotic variance of $\widehat{\boldsymbol{\theta}}$ controlling for the variability of $\widehat{\boldsymbol{\gamma}}$ has the form: $\Sigma = 4(\Phi + \Psi)$, where

$$\begin{aligned} &\Phi = \text{Var}(\widetilde{\mathbf{v}}_i), \quad \Psi = -(CH^{-1}C^\top + F + F^\top), \\ &F = E(\widetilde{\mathbf{v}}_i\mathbf{w}_i^\top H^{-1}C^\top), \\ &C = E\left(\frac{\partial^\top}{\partial\boldsymbol{\gamma}}\widetilde{\mathbf{g}}_i\right), \quad H = E\left(\frac{\partial^\top}{\partial\boldsymbol{\gamma}}\mathbf{w}_i\right), \quad \widetilde{\mathbf{v}}_i = \widetilde{\mathbf{g}}_i - \boldsymbol{\theta}, \\ &\widetilde{\mathbf{g}}_i = E(\mathbf{g}_{ij} \mid \mathbf{y}_i, \mathbf{r}_i). \end{aligned} \tag{27}$$

The extra term $\Phi$ in $\Sigma$ accounts for the additional variability in the estimated $\widehat{\boldsymbol{\gamma}}$. A consistent estimate of $\Phi$ is obtained by substituting consistent estimates (e.g., moment estimates) for the respective quantities in Eq. (27).

## FUNCTIONAL RESPONSE MODEL

Existing distribution-free regression models are all defined based on a single-subject response. For example, consider a sample of size $n$, and let $y_i$ and $\mathbf{x}_i$

denote some response and a vector of predictors (or covariates) of interest ($1 \le i \le n$). The most popular regression is the linear model defined by $E(y_i \mid \mathbf{x}_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$, where $E(y_i \mid \mathbf{x}_i)$ denotes the conditional mean of $y_i$ given $\mathbf{x}_i$, and $\boldsymbol{\beta}$ a vector of parameters. In this model, the dependent variable is a single-subject response $y_i$. Although the linear regression model has been extended for modeling more complex types of response variables such as binary and count data,[9] the fact remains that the specification of the model only involves a single-subject response. For example, in the generalized linear model defined by $E[(y_i \mid \mathbf{x}_i)] = g(\mathbf{x}_i^\top \boldsymbol{\beta})$, the right side is generalized to be a function of the linear predictor, $\mathbf{x}_i^\top \boldsymbol{\beta}$, to accommodate the range restriction of nonlinear response $y_i$, but the left side remains identical to the linear model.

One inherent weakness of such single-subject-response-based regression models is their limited ability to model the moments of a response variable. As a result, many popular statistics that are complex functions of higher-order moments or defined by multiple subjects' responses cannot be studied under the classic regression paradigm.

To overcome this fundamental limitation, consider a new class of models defined by a general nonlinear functional of several responses from multiple subjects in the form:

$$E[\mathbf{f}(y_{i_1}, \ldots, y_{i_q}) \mid \mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_q}] = \mathbf{g}(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_q}; \boldsymbol{\beta}), \quad (i_1, \ldots, i_q) \in C_q^n, \tag{28}$$

where $\mathbf{f}(\cdot)$ is some function, $\mathbf{g}(\cdot)$ some smooth function (with continuous second-order derivatives), $C_q^n$ denotes the set of $\binom{n}{q}$ combinations of $q$ distinct elements $(i_1, \ldots, i_q)$ from the integer set $\{1, \ldots, n\}$ and $\boldsymbol{\theta}$ a vector of parameters. By generalizing the response variable in such a way, this class of FRM provides a broad framework for modeling not only higher-order moments such as variance and product–moment (PM) correlations but also multi-subject-response-based models such as the classic Mann–Whitney–Wilcoxon rank tests and modern-day social network connectivity.

Consider once again the model for longitudinal variance in Eq. (10) and let

$$\mathbf{f}(\mathbf{y}_i, \mathbf{y}_j) = \frac{1}{2}\begin{pmatrix} (y_{i1} - y_{j1})^2 \\ \vdots \\ (y_{im} - y_{jm})^2 \end{pmatrix}, \quad \mathbf{g}(\boldsymbol{\theta}) = \begin{pmatrix} \sigma_1^2 \\ \vdots \\ \sigma_m^2 \end{pmatrix}, \quad (i, j) \in C_2^n. \tag{29}$$

Then, it follows from Eq. (13) that $E[\mathbf{f}(\mathbf{y}_i, \mathbf{y}_j)] = \mathbf{g}(\boldsymbol{\theta})$, an FRM for modeling the variance of a longitudinal outcome vector $\mathbf{y}_i$.

Now consider again the Goodman and Kruskal $\gamma$ in Eq. (7). Let

$$\begin{aligned} \mathbf{f}(\mathbf{z}_i, \mathbf{z}_j) &= (f_1(\mathbf{z}_i, \mathbf{z}_j), f_1(\mathbf{z}_i, \mathbf{z}_j))^\top, \quad \mathbf{g}(\boldsymbol{\theta}) = (\gamma\xi, \xi)^\top, \\ \boldsymbol{\theta} &= (\xi, \gamma)^\top, \\ f_1(\mathbf{z}_i, \mathbf{z}_j) &= I_{\{(u_i - u_j)(v_i - v_j) > 0\}} - I_{\{(u_i - u_j)(v_i - v_j) < 0\}}, \\ f_2(\mathbf{z}_i, \mathbf{z}_j) &= I_{\{(u_i - u_j)(v_i - v_j) > 0\}} + I_{\{(u_i - u_j)(v_i - v_j) < 0\}}. \end{aligned}$$

Then, the FRM $E[\mathbf{f}(\mathbf{z}_i, \mathbf{z}_j)] = \mathbf{g}(\boldsymbol{\theta})$ provides inference about $\gamma$. Note that there are two parameters in this FRM, $\xi$ and $\gamma$, with the latter being the one of primary interest.

To illustrate an FRM for a regression setting, consider modeling relationships in a social network, defined by a finite set of nodes $\{\mathbf{s}_i; 1 \le i \le n\}$, as depicted in Figure 1. The nodes represent social entities such as people, organizations and countries, whose specific ties such as friendship, competition, and collaboration constitute connectivity within the social network.

To model such network connectivity, let $y_{ij} = y(\mathbf{s}_i, \mathbf{s}_j)$ be a binary response, with the value 1 (or 0) to indicate the presence (or absence) of a tie between $\mathbf{s}_i$ and $\mathbf{s}_j$. The dyad defined by the pair of tie variables $(y_{ij}, y_{ji})$ can be (0, 0), (1, 1), (0, 1) or (1, 0), representing a null, mutual (or reciprocal), or asymmetric relationship between the nodes $\mathbf{s}_i$ and $\mathbf{s}_j$, respectively. Since self-ties are typically of no practical interest, $y_{ii}$ is defined to be 0 in most applications.

The Holland and Leinhardt model for social network connectivity is defined as[17]:

$$\begin{aligned} Pr(y_{ij} = r_{ij}, y_{ji} = r_{ji} \mid \mathbf{x}_i, \mathbf{x}_j, \boldsymbol{\theta}) &= \frac{1}{k_{ij}} \exp[r_{ij} h(\mathbf{u}_i, \mathbf{v}_j) + r_{ji} h(\mathbf{u}_j, \mathbf{v}_i) + r_{ij} r_{ji} \rho] \\ h(\mathbf{u}_i, \mathbf{v}_j) &= \mu + \mathbf{u}_i^\top \boldsymbol{\alpha} + \mathbf{v}_j^\top \boldsymbol{\beta}, \quad (i, j) \in C_2^n, \quad r_{ij}, r_{ji} = 0, 1. \end{aligned} \tag{30}$$

In this model, $\mathbf{x}_i = (\mathbf{u}_i^\top, \mathbf{v}_i^\top)^\top$ is a set of predictors, with $\mathbf{u}_i^\top \boldsymbol{\alpha}$ representing the sender (productivity), $\mathbf{v}_j^\top \boldsymbol{\beta}$ the receiver effects (attractiveness), $\rho$ the force of reciprocation, and $k_{ij}$ a normalizing factor given by:

$$k_{ij} = 1 + \exp(h(\mathbf{u}_i, \mathbf{v}_j)) + \exp(h(\mathbf{u}_j, \mathbf{v}_i)) + \exp(h(\mathbf{u}_i, \mathbf{v}_j) + h(\mathbf{u}_j, \mathbf{v}_i) + \rho).$$

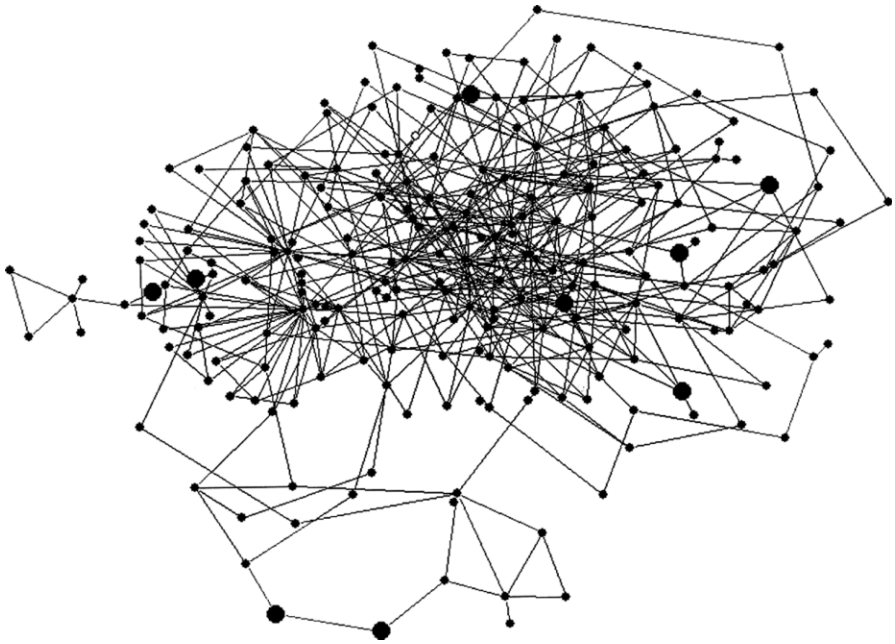


**FIGURE 1** | A hypothetical social network with dots representing nodes and edges denoting the presence of ties.

The model in Eq. (30) allows for nonsymmetric ties, i.e., $y_{ij} \neq y_{ji}$. But, if the relationship is symmetric, i.e., $y_{ij} = y_{ji}$, (30) simplifies to:

$$\Pr(y_{ij} = 1 \mid \mathbf{x}_i, \mathbf{x}_j, \boldsymbol{\theta}) = \text{logit}^{-1}[\exp(h(\mathbf{u}_i, \mathbf{v}_j) + \rho)],$$
$$h(\mathbf{u}_i, \mathbf{v}_j) = \mu + \mathbf{u}_i^\top \boldsymbol{\alpha} + \mathbf{v}_j^\top \boldsymbol{\beta}.$$

Inference about the parameters $\boldsymbol{\theta} = (\mu, \boldsymbol{\alpha}^\top, \beta^\top, \rho)^\top$ is extremely difficult under conventional regression analysis, since the dyad-based responses $\{y_{ij}; 1 \leq i, j \leq n\}$ are correlated and cannot be 'partitioned' into independent clusters. This unique characteristic of the Holland and Leinhardt model sets it apart from other popular types of clustered data. For example, in longitudinal data analysis, data from repeated assessments on the same subject are also clustered. However, by partitioning the data into independent clusters based on the observations of each subject, traditional statistical models defined based on single-subject such as the generalized linear mixed-effects models and the marginal generalized linear models can be applied to model the clustered outcomes, with inference based on maximum likelihood and generalized estimating equations. Such a partitioning is unfortunately not possible for social network data.

The dyad $(y_{ij}, y_{ji})$ yields four possible patterns indexed by a four-level nominal variable $z_{it}$:

$$\begin{aligned}
&\text{Pattern 1} : \{y_{ij} = 1, y_{ji} = 1\},\\
&\text{Pattern 2} : \{y_{ij} = 0, y_{ji} = 1\},\\
&\text{Pattern 3} : \{y_{ij} = 1, y_{ji} = 0\},\\
&\text{Pattern 4} : \{y_{ij} = 0, y_{ji} = 0\}.
\end{aligned} \tag{31}$$

As only three of the four patterns are independent, we model the probability of each of the first three using the generalized logit model.[2] Designating the first three patterns in Eq. (31) by a three-variate multinomial response $\mathbf{p}_{ij} = (p_{1ij}, p_{2ij}, p_{3ij})^\top$, where $p_{lij} = 1$ if Pattern $l$ is true and 0 otherwise, the Holland and Leinhardt model in Eq. (30) can be expressed in the form of FRM $E[\mathbf{f}(y_{ij}, y_{ji}) \mid \mathbf{x}_i, \mathbf{x}_j] = \mathbf{g}(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta})$ as follows:

$$\begin{aligned}
\mathbf{f}(y_{ij}, y_{ji}) &= (f_{1ij}, f_{2ij}, f_{3ij})^\top = (p_{1ij}, p_{2ij}, p_{3ij})^\top,\\
\mathbf{g}(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}) &= (g_{1ij}, g_{2ij}, g_{3ij})^\top,\\
g_{1ij} &= E(p_{1ij} \mid \mathbf{x}_i, \mathbf{x}_j) = \frac{1}{k_{1ij}} \exp(h(\mathbf{u}_i, \mathbf{v}_j)\\
&\quad + h(\mathbf{u}_j, \mathbf{v}_i) + \rho),\\
g_{2ij} &= E(p_{2ij} \mid \mathbf{x}_i, \mathbf{x}_j) = \frac{1}{k_{1ij}} \exp(h(\mathbf{u}_i, \mathbf{v}_j)),\\
g_{3ij} &= E(p_{3ij} \mid \mathbf{x}_i, \mathbf{x}_j) = \frac{1}{k_{1ij}} \exp(h(\mathbf{u}_j, \mathbf{v}_i)).
\end{aligned}$$

There is a growing interest and increased use of social network analysis in a wide range of disciplines, including biomedicine, genetics, behavioral and mental health, social sciences, and health-related services. This class of models is also playing an increasingly important role in clinical trial and related intervention

research studies, especially those focusing on peer group and social support.[18,19]

## INFERENCE FOR FUNCTIONAL RESPONSE MODEL

### Complete Data

The most popular approach for inference for distribution-free models is the generalized estimating equations (GEE) defined by[2,20]:

$$\mathbf{W}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} G_i(\boldsymbol{\theta})(\mathbf{y}_i - \mathbf{g}_i) = \frac{1}{n}\sum_{i=1}^{n} G_i(\boldsymbol{\theta})S_i = 0, \tag{32}$$

where $G_i(\boldsymbol{\theta})$ is some $p \times 1$ function of $\boldsymbol{\theta}$ and $\mathbf{x}_i$ (but not $\mathbf{y}_i$) with $p$ denoting the dimension of $\boldsymbol{\theta}$. Under the some mild regularity conditions, estimates of $\boldsymbol{\theta}$ obtained by solving the above estimating equations are consistent and asymptotically normal. Although different choices of $G_i$ in Eq. (32) give rise to different estimates of $\boldsymbol{\theta}$, all such estimates are consistent. The choice of $G_i$ and associated properties of the estimates for GEE have been extensively discussed in the literature.[2,7,20–22]

For FRM, the data vector $\mathbf{f}$ is a function of responses from multiple subjects and as a result, Eq. (32) no longer defines estimating equations for such models and the asymptotic theory for estimating equations no longer applies. To illustrate the issue as well as to motivate the extension of Eq. (32) to the FRM setting, consider again the FRM in Eq. (29) for longitudinal variance. For convenience, set $m = 2$ so $\mathbf{y}_i = (y_{i1}, y_{j2})^\top$. It follows that

$$\mathbf{f}_\mathbf{i} = \mathbf{f}(\mathbf{y}_i, \mathbf{y}_j) = \left(\frac{1}{2}(y_{i1} - y_{j1})^2, \frac{1}{2}(y_{i2} - y_{j2})^2\right)^\top,$$
$$\mathbf{g}(\boldsymbol{\theta}) = \boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2)^\top, \quad \mathbf{i} = (i, j). \tag{33}$$

By setting $G = (\partial/\partial\boldsymbol{\theta})\mathbf{g}(\boldsymbol{\theta}) = \mathbf{I}_2$ (the $2 \times 2$ identity matrix) and treating $\mathbf{f}_\mathbf{i}$ and $\mathbf{g}$ above as $\mathbf{y}_i$ and $\mathbf{g}_i$ in Eq. (32), we obtain:

$$\mathbf{U}_n = \sum_{\mathbf{i}\in C_2^n} \mathbf{U}_{n\mathbf{i}} = \sum_{\mathbf{i}\in C_2^n} G(\mathbf{f}_\mathbf{i} - \mathbf{g})$$
$$= \sum_{(i,j)\in C_2^n} \left[\frac{1}{2}\begin{pmatrix}(y_{i1} - y_{j1})^2\\ (y_{i2} - y_{j2})^2\end{pmatrix} - \begin{pmatrix}\sigma_1^2\\ \sigma_2^2\end{pmatrix}\right] = 0. \tag{34}$$

The above suggests a generalization of GEE in Eq. (32) to accommodate multi-subject-response-based FRM. In addition, since $\mathbf{U}_{n\mathbf{i}} = G(\mathbf{f}_\mathbf{i} - \mathbf{g})$ is symmetric with respect to $\mathbf{y}_i$ and $\mathbf{y}_j$, $\mathbf{U}_n$ in Eq. (34) is $U$-statistic-like quantity.

For the general FRM in Eq. (28), we define a set of U-statistics-based generalized estimating equations (UGEE) as follows:

$$\mathbf{U}_n = \sum_{\mathbf{i}\in C_q^n} \mathbf{U}_{n\mathbf{i}} = \sum_{\mathbf{i}\in C_q^n} G_\mathbf{i} S_\mathbf{i} = \sum_{\mathbf{i}\in C_q^n} G_\mathbf{i}(\mathbf{f}_\mathbf{i} - \mathbf{g}_\mathbf{i}) = 0,$$
$$\mathbf{i} = (i_1, \ldots, i_q) \in C_q^n, \tag{35}$$

where $G_i(\boldsymbol{\theta})$ has the same properties as $G_i(\boldsymbol{\theta})$ in GEE, i.e., $G_i(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$ and $\mathbf{x}_{i_k}$ $(1 \le k \le q)$. Under some mild regularity conditions (akin to GEE), it follows from the asymptotic normality of multivariate U-statistics discussed in the Section *Inference for U-Statistics* that $\mathbf{U}_n$ has an asymptotic normal distribution. Furthermore, by asymptotically expanding the UGEE estimate $\widehat{\boldsymbol{\theta}}$ around $\boldsymbol{\theta}$, it can be shown that $\widehat{\boldsymbol{\theta}}$ is also asymptotically normal.[2] Note that as in the case of GEE, $G_\mathbf{i}(\boldsymbol{\theta})$ cannot depend on $\mathbf{y}_{i_k}$, since otherwise the UGEE may not even yield consistent estimate.[2]

### Incomplete Data

If $y_{i2}$ is missing for some of the subjects, the UGEE in Eq. (34) is only applicable to the subset of subjects with complete observations at posttreatment assessment time $t = 2$. Even though such a complete-data approach still yields consistent estimates of $\boldsymbol{\theta}$ under MCAR, it is not efficient. Furthermore, such estimates are generally biased under MAR.

The same approach discussed in the Section *Inference for U-Statistics* for multivariate U-statistics is readily applied to address the bias by defining a set of U-statistics-based weighted generalized estimating equations (UWGEE) to provide inference for FRM under MAR. For example, let $r_{it}$ and $\pi_{it}$ be defined the same way as in the Section *Inference for U-Statistics*. Since $y_{i1}$ is always observed at pretreatment, $r_{i1} \equiv 1$. Now consider the following revised estimating equations:

$$\mathbf{U}_n = \sum_{\mathbf{i}\in C_2^n} \Delta_\mathbf{i}(\mathbf{f}_\mathbf{i} - \mathbf{g}) = \begin{pmatrix}1 & 0\\ 0 & \frac{r_{i2}r_{j2}}{\pi_{i2}\pi_{j2}}\end{pmatrix}[\mathbf{f}(\mathbf{y}_i, \mathbf{y}_j) - \mathbf{g}]$$
$$= \sum_{(i,j)\in C_2^n} \begin{pmatrix}1 & 0\\ 0 & \frac{r_{i2}r_{j2}}{\pi_{i2}\pi_{j2}}\end{pmatrix}$$
$$\times \left[\frac{1}{2}\begin{pmatrix}(y_{i1} - y_{j1})^2\\ (y_{i2} - y_{j2})^2\end{pmatrix} - \begin{pmatrix}\sigma_1^2\\ \sigma_2^2\end{pmatrix}\right] = 0. \tag{36}$$

The above is readily solved in closed form, yielding the same estimates $\widehat{\boldsymbol{\theta}}$ as in Eq. (23).

For a general FRM, the UWGEE is a modification of Eq. (35) with the following form:

$$\mathbf{U}_n = \sum_{\mathbf{i}\in C_q^n} \mathbf{U}_{n\mathbf{i}} = \sum_{\mathbf{i}\in C_q^n} G_\mathbf{i}\Delta_\mathbf{i} S_\mathbf{i} = \sum_{\mathbf{i}\in C_q^n} G_\mathbf{i}\Delta_\mathbf{i}(\mathbf{f}_\mathbf{i} - \mathbf{g}_\mathbf{i}) = 0,$$
$$\mathbf{i} = (i_1, \ldots, i_q) \in C_q^n, \tag{37}$$

where $G_\mathbf{i}$ is defined the same as in Eq. (35) and $\Delta_\mathbf{i}$ is a function of weights such as $\pi_{it}$ in Eq. (36) determined by the probability of observing the response vector $\mathbf{y}_i$. As in the complete-data case, it can be shown that under some mild regularity conditions, both $\mathbf{U}_n$ and the UWGEE estimate $\widehat{\boldsymbol{\theta}}$ are asymptotically normal.[2] Under MMDP, we may again model $\pi_{it}$ using logistic regression and estimate the asymptotic variance of $\widehat{\boldsymbol{\theta}}$ by accounting for the variability of estimated $\Delta_\mathbf{i}$ (or rather $\pi_{it}$), as discussed in Section *Inference for U-Statistics* for multivariate U-statistics.[2]

## REAL STUDY APPLICATIONS

In this section, we present applications of U-statistics and functional response models to some real study data within a longitudinal data setting. In all the examples, we set the statistical significance at $\alpha = 0.05$.

**Example 1.** In this Penn State Young Women's Health Study,[23] the percentage of body fat was measured for a group of adolescent girls using skinfold calipers (SC) and dual-energy X-ray absorptiometry (DEXA). The goal of the study was to examine the performance of SC to see if it could be used to replace the more labor-intensive and costly DEXA as a measure of body fat for this study population.

The subjects in the study were all at the age of 12 and were assessed initially at the study entry and then every 6 months thereafter. For illustration purposes, we focus on the 89 subjects with complete baseline data at the entry into the study.

We examined the performance of SC using the concordance correlation coefficient (CCC), a popular index for assessing agreement between two continuous outcomes.[24] Let $\mathbf{y}_{it} = (y_{i1t}, y_{i2t})^\top$ denote the SC ($y_{i1t}$) and DEXA ($y_{i2t}$) measures of body fat at time $t$, with $t = 1$ denoting the baseline and $t = 2$ and 3 corresponding to the two follow-up assessments. The CCC at time $t$ is defined by:

$$\rho_{\mathrm{CCC}}(t) = \frac{2\sigma_{12t}}{\sigma_{1t}^2 + \sigma_{2t}^2 + (\mu_{1t} - \mu_{2t})^2},$$
$$\mu_{kt} = E(y_{ikt}), \quad \sigma_{kt}^2 = \mathrm{Var}(y_{ikt}),$$
$$\sigma_{12t} = \mathrm{Cov}(y_{i1t}, y_{i2t}). \tag{38}$$

The $\rho_{\mathrm{CCC}}$ ranges between $-1$ and 1; $\rho_{\mathrm{CCC}} = 1$ ($-1$) if the two outcomes completely agree (disagree), and $\rho_{\mathrm{CCC}} = 0$ if $y_{i1t}$ and $y_{i2t}$ are independent. Furthermore, $\rho_{\mathrm{CCC}} = \rho_{\mathrm{PM}} C_b$, where $\rho_{\mathrm{PM}}$ is the PM correlation measuring precision, while $C_b$ ($0 \le C_b \le 1$) is a function of scale, $\sigma_{1t}/\sigma_{2t}$, and location shift relative to the scale, $(\mu_{1t} - \mu_{2t})/\sqrt{\sigma_{1t}\sigma_{2t}}$, indicating accuracy. Thus, unlike $\rho_{\mathrm{PM}}$ and other popular association measures such as Spearman's rho, CCC captures both accuracy and precision, making it an appropriate measure for evaluating the quality of SC when used as an inexpensive alternative to DEXA for assessing percentage of body fat.

Shown in Table 1 are the estimates of PM correlation coefficient $\rho_{\mathrm{PM}}$, and location and scale shift across all assessments. It is seen that SC and DEXA show a strong linear relation at baseline (as indicated by a higher $\rho_{\mathrm{PM}}$), but moderate linear associations at the follow-up assessments. Location shifts nearly doubled at times 2 and 3, when compared with that at baseline, indicating consistent differences between the two methods at the follow-up times. Scale shifts are minor, but again with the largest difference occurring at the follow-up visits. The large location shifts at the follow-up assessments would discount the use of any association measure such as the PM correlation and Spearman's $\rho$ as a measure of agreement between SC and DEXA in this study population.

In this study, missing data only occurred to DEXA, which followed the MMDP. We modeled the missingness using logistic regression, with the predictor defined as a function of the outcome observed immediately prior to $t$ under the Markov assumption, $\widetilde{\mathbf{y}}_{it} = \mathbf{y}_{i(t-1)}$, and thus

$$\mathrm{logit}(p_{it}) = \alpha_t + \beta_{1t} y_{i1(t-1)} + \beta_{2t} y_{i2(t-1)},$$
$$p_{it} = E(r_{it} = 1 \mid \widetilde{\mathbf{y}}_{it}), \quad t = 2, 3. \tag{39}$$

Shown in Table 2 are the estimates of $\boldsymbol{\beta}_t = (\beta_{1t}, \beta_{2t})^\top$, their standard errors, and corresponding $p$-values ($2 \le t \le 3$). It is seen that none of the estimated

**TABLE 1** | Estimates of Product-Moment (PM) Correlation, Location, and Scale Shift Between Skinfold Calipers (SC) and Dual-energy X-ray Absorptiometry (DEXA) at Each Assessment for the Penn State Young Women's Health Study

| | PM Correlation | Location Shift | Scale Shift |
|---|---|---|---|
| Assessment | $\rho_{\mathrm{PM}}$ | $\frac{\lvert\mu_{\mathrm{SC}} - \mu_{\mathrm{DEXA}}\rvert}{\sqrt{\sigma_{\mathrm{SC}}\sigma_{\mathrm{DEXA}}}}$ | $\sigma_{\mathrm{SC}}/\sigma_{\mathrm{DEXA}}$ |
| Age 12.5 ($t = 1$) | 0.804 | 0.562 | 0.812 |
| Age 13 ($t = 2$) | 0.622 | 1.013 | 0.859 |
| Age 13.5 ($t = 3$) | 0.623 | 1.068 | 0.867 |

**TABLE 2** | Estimates of Coefficients of Logistic Regression for Modeling Missingness under Monotone Missing Data Pattern (MMDP) for the Penn State Young Women's Health Study

| Assessment Time | Predictors | Estimates | Standard Errors | $p$ Value |
|---|---|---|---|---|
| Age 13 ($t = 2$) | SC ($\beta_{12}$) | −0.13 | 0.32 | 0.68 |
| | DEXA ($\beta_{22}$) | 0.28 | 0.25 | 0.26 |
| Age 13.5 ($t = 3$) | SC ($\beta_{13}$) | −0.08 | 0.25 | 0.74 |
| | DEXA ($y_{23}$) | −0.03 | 0.20 | 0.87 |

**TABLE 3** | Estimates of Concordance Correlation Coefficient (CCC) over Time, Standard Errors, and $p$-Value for Testing the Null of Constant CCC across Three Assessments for the Penn State Young Women's Health Study

| Estimates of CCC over Time (Asymptotic Standard Errors) | | |
|---|---|---|
| Age 12.5 ($t = 1$) | Age 13 ($t = 2$) | Age 13.5 ($t = 3$) |
| 0.683 (0.047) | 0.509 (0.055) | 0.489 (0.063) |
| Hypothesis testing $H_0 : \rho_{CCC}(t) = \rho_{CCC}$, $1 \le t \le 3$ | | |
| $p$-value $< 0.01$ | | |

coefficients was statistically different from zero. For illustration purposes, however, we proceeded under the assumption of MAR.

To model $\rho_{CCC}(t)$ in Eq. (38) using multivariate U-statistics, let

$$\begin{aligned}
\theta_{t1} &= 2\sigma_{12t},\\
\theta_{t2} &= (\mu_{1t} - \mu_{2t})^2 + (\sigma_{1t}^2 + \sigma_{2t}^2 - 2\sigma_{12t}),\\
\boldsymbol{\theta}_t &= (\theta_{t1}, \theta_{t2})^\top, \boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \boldsymbol{\theta}_3^\top)^\top,\\
\phi_t &= \frac{\theta_{t1}}{\theta_{t1} + \theta_{t2}}, \boldsymbol{\phi}(\boldsymbol{\theta}) = (\phi_1, \phi_2, \phi_3)^\top.
\end{aligned}$$

Then, we have $\boldsymbol{\rho}_{CCC} = \boldsymbol{\phi}(\boldsymbol{\theta})$. Given an estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, we can estimate $\boldsymbol{\rho}_{CCC}$ by $\widehat{\boldsymbol{\rho}}_{CCC} = \boldsymbol{\phi}(\widehat{\boldsymbol{\theta}})$ and compute the asymptotic distribution of $\widehat{\boldsymbol{\rho}}_{CCC}$ using the Delta method. Within our context, we used Fieller's method to further improve the approximation to the asymptotic distribution of the ratio statistics $\widehat{\boldsymbol{\rho}}_{CCC}$ by the Delta method.[25,26]

To estimate $\boldsymbol{\theta}$, define a multivariate U-statistic as:

$$\begin{aligned}
h_{ijt1} &= (y_{i1t} - y_{j1t})(y_{i2t} - y_{j2t}),\\
h_{ijt2} &= \frac{1}{2}[(y_{i1t} - y_{i2t})^2 + (y_{j1t} - y_{j2t})^2],\\
\mathbf{h}_{ijt} &= (h_{ijt1}, h_{ijt2})^\top, \quad \mathbf{h}_{ij} = (\mathbf{h}_{ij1}^\top, \mathbf{h}_{ij2}\mathbf{h}_{ij3}^\top)^\top,\\
\mathbf{U}_n &= \binom{n}{2}^{-1} \sum_{(i,j)\in C_2^n} \mathbf{h}_{ij}.
\end{aligned}$$

Under complete data, the estimate $\widehat{\boldsymbol{\theta}} = \mathbf{U}_n$ is unbiased and asymptotically normal. To account for MAR in this example, we revised $\mathbf{h}_{ij}$ as follows:

$$\mathbf{g}_{ijt} = \frac{r_{it}r_{jt}}{\pi_{it}\pi_{jt}}\mathbf{h}_{ijt}, \quad \pi_{it} = \prod_{s=1}^{t} p_{is}, \quad (i,j) \in C_2^n, \quad 1 \le t \le 3, \tag{40}$$

with $p_{it}$ defined in Eq. (39). We estimated the weights $\pi_{it}$ with estimates $\widehat{\boldsymbol{\beta}}_t$ from Eq. (39) based on the relationship in Eq. (26), and the asymptotic variance $\Sigma$ of $\widehat{\boldsymbol{\theta}}$ using Eq. (27).[26]

Shown in Table 3 are the estimated $\rho_{CCC}(t)$ and associated standard errors over time ($1 \le t \le 3$). The two assessment methods seemed to have the closest agreement at age 12.5, but with a steady decline over the follow-up times. To confirm this, we tested the null $H_0 : \rho_{CCC}(t) = \rho_{CCC}$ ($1 \le t \le 3$) using a linear contrast of the form:

$$H_0 : K\boldsymbol{\rho}_{CCC} = 0, \quad K = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}.$$

Under $H_0$, the Wald test statistic, $W_n = n\widehat{\boldsymbol{\rho}}^\top K^\top (K\widehat{\Sigma}K^\top)^{-1}K\widehat{\boldsymbol{\rho}}$, has an asymptotic central $\chi_2^2$ distribution with 2 degrees of freedom. For small samples, $W_n$ often yields inflated type I errors, since it completely ignores the variability in the estimated $\widehat{\Sigma}_{\boldsymbol{\rho}}$.[27] A popular alternative is Hotelling's T-square statistic, $T_n^2 = (n-p)/(p(n-1))W_n$, which follows approximately a central $F_{p,n-p}$ distribution with $p$ (numerator) and $n - p$ (denominator) degrees of freedom under $H_0$.[28] Given the relatively small sample size of this study, we reported the $p$-value in Table 3 based on the Hotelling T-square statistic.

The small $p$-value from the test of $H_0$ confirmed the discrepancies between the two methods for assessing body fat as indicated by the estimates of $\rho_{CCC}(t)$ over time. The findings substantiate the speculation in King et al.[29] that 'there may be physiological explanations for this phenomenon because of the fact that the skinfold estimation is only capable of detecting subcutaneous fat, whereas breast, lower body, and visceral fat are increasing over this age range because of the onset of menarche.

**Example 2.** The results in Example 1 show that the more convenient and less expensive SC did not provide precise measurement of percent of body fat for the population of adolescent girls in the study. To further investigate the cause of insufficient quality of SC, we would like to characterize the source of inaccuracy to see whether the problem is induced by scale shift, location shift or both.

To this end, consider the FRM for simultaneously modeling the mean and variance defined by:

$$
\begin{aligned}
f_{1iijt} &= \frac{1}{2}(y_{1it} + y_{1jt}) - \frac{1}{2}(y_{2it} + y_{2jt}),\\
f_{2ijt} &= \frac{1}{2}(y_{1it} - y_{1jt})^2 - \frac{1}{2}(y_{2it} - y_{2jt})^2,\\
\mathbf{f}_{ijt} &= (f_{1ijt}, f_{2ijt})^\top, \quad \mathbf{f}_{ij} = (\mathbf{f}_{ij1}^\top, \mathbf{f}_{ij2}^\top, \mathbf{f}_{ij3})^\top,\\
\mathbf{g}_{ijt} &= \boldsymbol{\theta}_t = (\mu_{1t} - \mu_{2t}, \sigma_1^2 - \sigma_2^2)^\top,\\
\mathbf{g}_{ij} &= (\mathbf{g}_{ij1}, \mathbf{g}_{ij2}, \mathbf{g}_{ij3})^\top, \quad \boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \boldsymbol{\theta}_3^\top)^\top. \qquad (41)
\end{aligned}
$$

The parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \boldsymbol{\theta}_3^\top)^\top$ represents the difference in $\mu_{kt}$ and $\sigma_{kt}^2$ at each of the three assessment points. Thus, we can examine location, scale shift or both at a particular assessment point as well as across all different assessment times using the FRM in Eq. (41).

To account for missing data, we utilized the UWGEE in Eq. (37) to estimate $\boldsymbol{\theta}$, with the weights $\pi_{it}$ estimated in Example 1 by Eq. (39). By setting $G_i = (\partial)/\partial\boldsymbol{\theta})\mathbf{g}(\boldsymbol{\theta}) = \mathbf{I}_3$, we obtained the UWGEE estimate $\widehat{\boldsymbol{\theta}}$ in closed form and estimated its asymptotic variance using Eq. (27).[30]

Shown in Table 4 are the estimated $\mu_{kt}$ and $\sigma_{kt}^2$, and their standard errors, along with the $p$-values for testing the null of mean and variance homogeneity at each time point. The results indicate a significant upward (downward) bias in the mean (variance) by SC across all assessment times. Furthermore, by testing each of the null hypotheses,

$$
\begin{aligned}
H_{01} &: \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} = \mu_{13} - \mu_{23},\\
H_{02} &: \sigma_{11}^2 - \sigma_{21}^2 = \sigma_{12}^2 - \sigma_{22}^2 = \sigma_{13}^2 - \sigma_{23}^2,
\end{aligned}
$$

we obtained a $p$-value $< 0.01$ for $H_{01}$ and 0.38 for $H_{02}$, indicating that the bias in the mean was differentially distributed over time, whereas the one in the variance was not.

## DISCUSSION

In this article, we focused on longitudinal study data and discussed extending classic multivariate U-statistics to such a study setting. In addition, we presented a class of FRM to extend the traditional regression paradigm for longitudinal data analysis. Both generalizations utilize the IPW approach to address missing data so that they provide valid inference under MAR, the most common missing data mechanism in clinical trial and cohort studies. In the latter case, the UWGEE generalizes the popular WGEE for standard distribution-free regression models to provide inference for multi-subject-response-based models such as the FRM.

The extension of U-statistics and associated applications to longitudinal data analysis is imperative, as longitudinal study designs have become the standard bench mark rather than the exception in modern clinical trials and observational studies. A major challenge in extending statistical models and methods for cross-sectional study data to the longitudinal setting is to address missing data. As missing data arising in most clinical and cohort studies hardly follows MCAR, limiting the analysis to the subset of subjects with complete data not only reduces power but also more importantly introduces bias in the estimate as well. In order for longitudinal models to be of great utility, we must overcome this major difficulty by developing procedures to provide valid inference under MAR.

For parameters and statistics defined by multi-subject responses such as the classic Goodman and Kruskal $\gamma$ and modern-day social network connectivity, we must use U-statistics and FRM to model such parameters of interest and investigate the asymptotic properties of parameter estimates. However, U-statistics and FRM can also be used to facilitate modeling of parameters defined by a single-subject response. For example, in addition to the sample variance and CCC illustrated in this article,

**TABLE 4** | Estimates of Mean and Variance of Skinfold Calipers (SC) and Dual-Energy X-ray Absorptiometry (DEXA), and Standard Errors, as well as $p$-Values for Testing the Null of Equal Mean and Variance at Each Assessment Point for the Penn State Young Women's Health Study

| Method | Mean (standard errors) | | | Variance (standard errors) | | |
|---|---|---|---|---|---|---|
| Estimates of Mean and Variance of SC and DEXA over Time (Standard Errors) | | | | | | |
| SC | 23.8 (0.37) | 25.2 (0.36) | 25.2 (0.33) | 12.0 (1.73) | 11.3 (1.48) | 9.7 (1.68) |
| DEXA | 21.6 (0.45) | 21.5 (0.44) | 21.5 (0.45) | 18.3 (2.68) | 15.4 (2.63) | 13.4 (2.41) |
| Hypothesis testing | | | | | | |
| | $\mu_{11} = \mu_{21}$ | $\mu_{12} = \mu_{22}$ | $\mu_{13} = \mu_{23}$ | $\sigma_{11}^2 = \sigma_{21}^2$ | $\sigma_{12}^2 = \sigma_{22}^2$ | $\sigma_{13}^2 = \sigma_{23}^2$ |
| $p$-value | <0.01 | <0.01 | <0.01 | <0.01 | 0.061 | 0.015 |

many statistical models of interest in modern applications also involve second- or even higher-order moments, such as $\kappa$ for agreement between categorical outcomes,[31] Cronbach Coefficient $\alpha$ for internal validity of instrument measuring latent construct,[32] and negative binomial and zero-inflated Poisson distribution for modeling over-dispersed count responses.[10,33] Although such higher-order moments may be modeled using conventional methods such as GEE II,[7,8] U-statistics-based approaches provide more elegant solutions.[2] Furthermore, it is much more convenient to generalize such approaches to longitudinal data analysis.[26,29,34–37]

FRM provides a broad framework for modeling quantities of interest defined by complex functions of multi-subject responses such as the social network example discussed in this article. Even for models defined by single-subject responses, FRM sets it apart from standard methods for modeling high-order moments. Conventional methods such as GEE II do not offer a formal regression-like setting as in standard mean-based models such as the generalized linear models with the response on the left side and the conditional mean of the response on the right side of the equation. By extending the single-subject response to allow for arbitrary functions of multi-subject responses, FRM not only addresses this fundamental limitation of the classic regression paradigm, but most importantly also provides a unified framework for concurrently modeling complex longitudinal relationships and addressing the inherent missing data. For example, modeling over-dispersed and/or zero-inflated count is beyond the realm of the traditional distribution-free regression paradigm, as the mean response, or first-order moment, does not provide sufficient information to identify all parameters when attempting to fit parametric models such as the negative and zero-inflated Poisson loglinear models within the confines of distribution-free setting.[2] Although GEE II may be used to model second-order moments necessary to identify the parameter, it is *ad hoc* and incapable of addressing MAR. In contrast, by working in tandem, FRM and UWGEE provide a natural extension of classic mean-based regression models and WGEE to simultaneously address model identifiably and MAR.[38]

## ACKNOWLEDGMENT

This research was in part supported by a Leukemia and Lymphoma Society Award and an NIH grant 4R33DA027521-02.

## REFERENCES

1. Hoeffding W. A class of statistics with asymptotically normal distribution. *Ann Math Stat* 1948, 19:293–325.
2. Kowalski J, Tu XM. *Modern Applied U Statistics*. New York: John Wiley & Sons; 2007.
3. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat* 1947, 18:50–60.
4. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics* 1945, 1:80–83.
5. Kendall MG. A new measure of rank correlation. *Biometrika* 1938, 30:81–93.
6. Agresti A. *Categorical Data Analysis*. New York: John Wiley & Sons; 1990.
7. Prentice RL, Zhao LP. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* 1991, 47:825–839.
8. Barnhart HX, Haber M, Song J. Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics* 2002, 58:1020–1027.
9. McCullagh P, Nelder JA. *Generalized Linear Models*. 2nd ed. London: Chapman and Hall; 1989.
10. Dean CB. Testing for overdispersion in Poisson and binomial regression models. *J Am Stat Assoc* 1992, 87:451–457.
11. Morrison-Beedy D, Carey MP, Feng C, Tu XM. Predicting sexual risk behaviors among adolescent and young women using a prospective diary method. *Res Nurs Health* 2008, 31:329–340.
12. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. New York: John Wiley & Sons; 1987.
13. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Stat Assoc* 1995, 90:106–121.
14. Tsiatis AA. *Semiparametric Theory and Missing Data*. New York: Springer; 2006.
15. Lu N, Tang W, He H, Yu Q, Crits-Christoph P, Zhang H, Tu XM. On the impact of parametric assumptions and robust alternatives for longitudinal data analysis. *Biom J* 2009, 51:627–643.

16. Yu Q, Tang W, Marcus S, Ma Y, Zhang H, Tu XM. Modeling sensitivity and specificity with a time-varying reference standard within a longitudinal setting. *Appl Stat* 2010, 37:1213–1230.
17. Holland PW, Leinhardt S. An exponential family of probability distributions for directed graphs. *J Am Stat Assoc* 1981, 77:33–50.
18. Helgeson VS, Cohen S. Social support and adjustment to cancer: reconciling descriptive, correlational, and intervention research. *Health Psychol* 1996, 15:135–148.
19. Berkman LF, Glass T, Brisette I, Seeman TE. From social integration to health: Durkheim in the new millennium. *Soc Sci Med* 2000, 51:843–857.
20. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986, 73:13–22.
21. Pepe MS, Anderson GL. A cautionary note on inferences for marginal regression models with longitudinal data and general correlated response. *Commun Stat Theory* 1994, 23:939–951.
22. Godambe VP. *Estimating Functions*. Oxford: Oxford University Press; 1991.
23. Lloyd T, Chinchilli VM, Eggli DF, Rollings N, Kulin HE. Body composition development of adolescent white females. *Arch Pediatr Adolesc Med* 1998, 152: 998–1002.
24. Lin, L. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989, 45:255–268.
25. Fieller EC. Some problems in interval estimation. *J R Stat Soc B* 1954, 16:175–185.
26. Ma Y, Tang W, Yu Q, Tu XM. Modeling concordance correlation coefficient for longitudinal study data. *Psychometrika* 2010, 75:99–119. doi:10.1007/S11336-009-9142-Z.
27. Guo X, Pan W, Connett JE, Hannan PT, French SA. Small-sample performance of the robust score test and its modifications in generalized estimating equation statistics. *Stat Med* 2005, 24:3479–3495.
28. Seber GAF. *Multivariate Observations*. New York: John Wiley & Sons; 1984.
29. King TS, Chinchilli VM, Carrasco JL. A repeated measures concordance correlation coefficient. *Stat Med* 2007, 26:3095–3113.
30. Zhang H, He H, Yu Q, Chen R, Lu N, Tu XM. *Generalized ANOVA for concurrently modeling mean and variance within a longitudinal data setting*. Technical Report, University of Rochester, Rochester, New York, 2011.
31. Cohen J. Weighted $\kappa$: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968, 70:213–220.
32. Cronbach LJ. Coefficient $\alpha$ and the internal structure of tests. *Psychometrika* 1951, 16:297–334.
33. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992, 34:1–14.
34. Tu XM, Feng C, Kowalski J, Tang W, Wang H, Wan C, Ma Y. Correlation analysis for longitudinal data: applications to HIV and psychosocial research. *Stat Med* 2007, 26:4116–4138.
35. Ma Y, Tang W, Feng C, Tu XM. Inference for $\kappa$ s for longitudinal study data: applications to sexual health research. *Biometrics* 2008, 64:781–789.
36. Ma Y, Gonzalez Della Valle A, Zhang H, Tu XM. A U-statistics based approach for modeling Cronbach coefficient $\alpha$ within a longitudinal data setting. *Stat Med*. In press.
37. Lu N, Gunzler D, Zhang H, Ma Y, He H, Tu XM. *On robust inference for intraclass correlation coefficients*. Technical Report, University of Rochester, Rochester, New York, 2011.
38. Yu Q, Chen R, Tang W, He H, Gallop R, Crits-Christoph P, Tu XM. *Distribution-free inference of negative and zero-inflated Poisson for longitudinal data*. Technical Report, University of Rochester, Rochester, New York, 2011.

## FURTHER READING

Chamberlain G. Asymptotic efficiency in estimation with conditional moment restrictions. *J Econ* 1987, 34:305–324.

Delong ER, Delong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a non-parametric approach. *Biometrics* 1988, 44:837–845.

Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *J Am Stat Assoc* 1952, 47:583–621.