Extending the Mann–Whitney–Wilcoxon rank sum test to longitudinal regression analysis

R. Chen\textsuperscript{a*}, T. Chen\textsuperscript{a}, N. Lu\textsuperscript{a}, H. Zhang\textsuperscript{b}, P. Wu\textsuperscript{a}, C. Feng\textsuperscript{a} and X.M. Tu\textsuperscript{a}

\textsuperscript{a}Biostatistics and Computational Biology, University of Rochester Medical Center, 601 Elmwood Avenue, Box 630, Rochester, NY 14642, USA; \textsuperscript{b}Biostatistics, St. Jude Children’s Research Hospital, MS 768, 262 Danny Thomas Place, Memphis, TN 38105, USA

(Received 25 July 2013; accepted 13 May 2014)

Outliers are commonly observed in psychosocial research, generally resulting in biased estimates when comparing group differences using popular mean-based models such as the analysis of variance model. Rank-based methods such as the popular Mann–Whitney–Wilcoxon (MWW) rank sum test are more effective to address such outliers. However, available methods for inference are limited to cross-sectional data and cannot be applied to longitudinal studies under missing data. In this paper, we propose a generalized MWW test for comparing multiple groups with covariates within a longitudinal data setting, by utilizing the functional response models. Inference is based on a class of U-statistics-based weighted generalized estimating equations, providing consistent and asymptotically normal estimates not only under complete but missing data as well. The proposed approach is illustrated with both real and simulated study data.

Keywords: FRM; missing data; missing at random; outliers; sexual health; UWGEE

1. Introduction

Popular models for longitudinal data analysis with continuous outcomes such as the generalized linear mixed-effects models (GLMM) and weighted generalized estimating equations (WGEE) lack robustness in the presence of outliers. For example, in a study to evaluate the efficacy of a sexual risk-reduction intervention for sexually active teenage girls in low-income urban settings, a group at elevated risk for HIV, some adolescent girls reported very large numbers such as 450 and even 1,000,000 for their unprotected vaginal sex over a three-month period [13]. Although answers like this are clearly not legitimate values of the outcome, they do indicate the extremely high level of sexual activity among these girls, when compared with the rest of the study sample. However, the mean-based GLMM and WGEE are not capable of dealing with this type of
‘outliers’, due to the sensitivity of the sample mean to large values. Rank-based methods such as the popular Mann–Whitney–Wilcoxon (MWW) rank sum test are more effective to address such outliers. However, the MWW test is limited to two groups. The Kruskal–Wallis extension to more than two groups does not have known asymptotic distributions for its estimates [9]. In addition, none of these approaches accommodate covariates. Finally, although attempts have been made to extend the MWW test to longitudinal data, available methods often do not provide valid inference because they ignore the sampling variability in the estimated model for missing data [8,19].

In this paper, we address the aforementioned limitations by framing the MWW test within the regression framework of the functional response models (FRM). In Section 2, we provide a brief review of FRM and discuss how to recast the MWW test within the context of FRM to address its various limitations. In Section 3, we discuss inference for both cross-sectional and longitudinal data. In Section 4, we illustrate the approach with both real and simulated data.

2. Extending MWW rank sum test to multi-group and longitudinal study setting

We start with a brief review of the MWW rank sum test for between-group difference. We then generalize the classic test for more than two treatment groups within a longitudinal data setting utilizing the FRM.

2.1 MWW rank sum test

Consider two independent samples of size $n_k$ and let $y_{ki}$ be some continuous outcome from the $i$th subject within the $k$th group ($1 \leq i \leq n_k, k = 1, 2$). Let $R_{ki}$ denote the rank of $y_{ki}$ in the pooled sample. The Wilcoxon rank sum statistic has the following form [8,14]:

$$\text{Wilcoxon rank sum statistic: } W_n = \sum_{i=1}^{n_1} R_{1i}. $$

Note that the sum of the rank scores $\sum_{j=1}^{n_2} R_{2j}$ from the second group may also be used as a statistic. However, since the two sums add up to $n(n+1)/2$ ($n = n_1 + n_2$), only one of the sums can be used as in the test statistic. An alternative form of this statistic is the following Mann–Whitney statistic [8,14]:

$$\text{Mann–Whitney statistic: } U_n = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I_{\{y_{2j} - y_{1i} \leq 0\}}, \quad (1)$$

where $I_{\{u \leq 0\}}$ is a set indicator with $I_{\{u \leq 0\}} = 1$ if $u \leq 0$ and 0 otherwise. Since $W_n = U_n + n_1(n_1 + 1)/2$ (in the absence of ties), the two tests are equivalent. However, the Mann–Whitney version is readily generalized to longitudinal data in the presence of missing values, while the Wilcoxon form is not. Thus, we use the Mann–Whitney form of the test for the extension, which will be referred to as the MWW rank sum test in the discussion that follows unless otherwise stated.

Let $V_n = (\frac{n}{2})^{-1} U_n$ be the normalized MWW statistic in Equation (1) and let $\theta = E(V_n)$, where $(\frac{n}{2})$ denotes combinations of two distinct elements $(i,j)$ from the integer set $\{1, \ldots, n\}$. If $y_{ki}$ have the same distribution, then $\theta = \frac{1}{2}$. Although the reverse is generally not true, $\theta = \frac{1}{2}$ is often considered as the null hypothesis of no between-group difference in practical applications. This connotation is adopted in the following discussion unless otherwise stated.
Inference about the MWW test has been discussed in the literature (e.g. [7,8,14,18,19]). Below, we first provide a brief review of FRM and then discuss how to use the flexibility of FRM to extend this classic test to accommodate multiple groups, longitudinal data and even covariates.

2.2 Functional response models

Existing regression models are defined based on a single-subject response, severely limiting their applications in practice. For example, the classic linear regression has the form, \( E(y_i | x_i) = x_i^\top \beta \), where \( y_i \) denotes some response (a vector of predictors or covariates) and \( \beta \) is a vector of parameters. In this model, the response is a single-subject linear \( y_i \). Although modern regressions such as the generalized linear models (GLM) and even nonlinear models are more complex, the response remains the same as linear regression. For example, in the GLM \( E(y_i | x_i) = h(x_i^\top \beta) \), the linear predictor \( \eta_i = x_i^\top \beta \) is extended to a function \( h(x_i^\top \beta) \) to accommodate non-continuous outcome \( y_i \), but the response function is still the same single-subject linear \( y_i \) in the linear model.

Many quantities of interest require not only more complex response functions, but also multi-subject responses. Within the current context, the statistic in Equation (1) is defined by pairs of subjects \( y_{1i} \) and \( y_{2j} \). The FRM address the limitations of existing regression models by extending (1) the single-subject \( y_i \) to multiple subject responses and (2) the linear response to an arbitrary function

\[
E[f(y_{1i}, \ldots, y_{iq}) | x_{1i}, \ldots, x_{iq}] = h(x_{1i}, \ldots, x_{iq}; \theta),
\]

where \( f(\cdot) \) is some function, \( h(\cdot) \) some smooth function (with continuous second-order derivatives), \( y_{1i}, \ldots, y_{iq} \) are \( q \) individual responses, and \( x_{1i}, \ldots, x_{iq} \) denote vectors of predictors from \( q \) subjects. By generalizing the response in this fashion, an FRM has been successfully applied to address a range of methodological issues involving multi-subject responses and second-order moments [3,11,12,23,24].

2.3 Extending MWW statistic to multiple groups and longitudinal data with covariates

Adopting the previous notation, let

\[
f(y_{1i}, y_{2j}) = I(y_{1i} \leq y_{2j}), \quad h(\theta) = \theta, \quad i \in C_{1}^{n_1}, \quad j \in C_{1}^{n_2},
\]

where \( C_{q}^{n} \) denotes the set of \( \binom{n}{q} \) combinations of \( q \) distinct elements \( (i_1, \ldots, i_q) \) from the integer set \( \{1, \ldots, n\} \). Then, the following is an FRM:

\[
E[f(y_{1i}, y_{2j})] = h(\theta) = \theta, \quad i \in C_{1}^{n_1}, \quad j \in C_{1}^{n_2}.
\]

In the FRM above, \( \theta \) is the parameter of interest. For example, in most applications, we are interested in testing the null of no between-group difference, \( H_0 : \theta = \frac{1}{2} \).

The FRM in Equation (4) works for continuous \( y_{kj} \). For discrete outcomes, ties are present. The null in this case is expressed as [20]

\[
H_0 : E(I_{y_{1i} < y_{2j}}) + \frac{1}{2} E(I_{y_{1i} = y_{2j}}) = \frac{1}{2}.
\]

Thus, we may redefine the functional response in Equation (3) as \( f(y_{1i}, y_{2j}) = I_{y_{1i} < y_{2j}} + \frac{1}{2} I_{y_{1i} = y_{2j}} \). For notational brevity, we assume continuous \( y_{kj} \) throughout the remaining discussion unless otherwise noted.

By framing the MWW test under FRM, we can readily extend this test for comparing two groups to multi-group settings, even accommodating covariates. To this end, we adopt a popular
model for interpersonal connections in social network analysis [21]. Since ties between individual subjects in social networks involve two subjects, standard statistical models for individual outcomes such as linear and logistic regression are not appropriate for modeling ties between two subjects. Holland and Leinhardt introduced a model for such between-subject defined outcomes [5,24], which is readily adapted to our setting for the extension.

Let \( x_{ij} \) be a vector of covariates from the \( i \)th subject within the \( k \)th treatment \( (k = 1, 2) \). Consider the following FRM:

\[
h(x_{i1}, x_{j2}; \theta) = \logit^{-1} \left( \sum_{l=1}^{p} \beta_l v_l(x_{i1}, x_{j2}) + \gamma \right), \quad i \in C_1^{n_1}, \quad j \in C_1^{n_2},
\]

\[
v(x_{i1}, x_{j2}) = (v_1(x_{i1}, x_{j2}), \ldots, v_p(x_{i1}, x_{j2}))^\top, \quad \beta = (\beta_1, \ldots, \beta_p)^\top, \quad \theta = (\gamma, \beta^\top)^\top,
\]

where \( \logit^{-1}(x) = \exp(x)/(1 + \exp(x)) \) and \( v(x_{i1}, x_{j2}) \) is some function of differences between \( x_{i1} \) and \( x_{j2} \). For example, we may simply set: \( v_l(x_{i1}, x_{j2}) = x_{i1l} - x_{j2l} \), where \( x_{i1l} \) denotes the \( l \)th component of \( x_{i1} \). Since Equation (5) is a function of two subjects, each subject’s covariates no longer exert their effect individually, but rather jointly through the function \( v(x_{i1}, x_{j2}) \). In addition to the logit link, we may also use probit and other link functions. In fact, the probit is a more natural link in some applications (see Section 4 for details).

The above is essentially an extension of the popular logistic regression to our setting involving a response function defined by two subjects. If \( \beta = 0 \), it is readily seen that the null of no treatment difference under Equation (4) is equivalently expressed as \( H_0: \gamma = 0 \). If \( \beta \neq 0 \), Equation (5) accounts for the effect of covariates on the probability that \( y_{i1} \leq y_{j2} \) and \( H_0: \gamma = 0 \) again represents the null of no treatment difference.

We can also readily extend Equation (5) to \( K \) groups by defining an FRM as follows:

\[
E[f(y_{1i}, y_{kj}) \mid x_{1i}, x_{kj}] = h_k(x_{1i}, x_{kj}; \theta), \quad f(y_{1i}, y_{kj}) = I_{\{y_{i1} \leq y_{kj}\}},
\]

\[
h(x_{1i}, x_{kj}; \theta) = \logit^{-1} \left( \sum_{l=1}^{p} \beta_l v_l(x_{1i}, x_{kj}) + \sum_{q=2}^{K} I_{\{q=k\}} \gamma_q \right), \quad i \in C_1^{n_1}, \quad j \in C_1^{n_2}, \quad 2 \leq k \leq K,
\]

\[
\gamma = (\gamma_2, \ldots, \gamma_K)^\top, \quad \theta = (\gamma^\top, \beta^\top)^\top.
\]

As in the two group case, \( \beta \) indicates the effect of covariates, while \( H_0: \gamma = 0 \) is the primary null hypothesis of no treatment difference. We may also include \( I_{\{q=k\}} \) by \( v_l(x_{1i}, x_{kj}) \) interactions if \( v_l(x_{1i}, x_{kj}) \) has differential effects across \( k \).

We can further extend Equation (6) to longitudinal data analysis. For a longitudinal study with \( m \) assessments, let \( y_{klit} \) denote a response (a vector of covariates) from the \( i \)th subject within the \( k \)th group at time \( t \). The version of Equation (6) for longitudinal data is defined by

\[
E[f(y_{1it}, y_{kjt}) \mid x_{1it}, x_{kjt}] = h_k(x_{1it}, x_{kjt}; \theta), \quad f(y_{1it}, y_{kjt}) = I_{\{y_{i1} \leq y_{kjt}\}},
\]

\[
h_k(x_{1it}, x_{kjt}; \theta) = \logit^{-1} \left( \sum_{l=1}^{p} \beta_l v_l(x_{1it}, x_{kjt}) + \sum_{q=2}^{K} I_{\{q=k\}} \gamma_q \right), \quad i \in C_1^{n_1}, \quad j \in C_1^{n_2}, \quad 2 \leq k \leq K, \quad 1 \leq t \leq m,
\]

\[
\gamma_t = (\gamma_{2t}, \ldots, \gamma_{Kt})^\top, \quad \gamma = (\gamma_1^\top, \ldots, \gamma_m^\top)^\top, \quad \theta = (\gamma^\top, \beta^\top)^\top.
\]

Like the FRM in Equation (6) for cross-sectional data, \( \beta \) accounts for the effect of covariates, while \( \gamma_t \) models the between-group difference at time \( t \). For randomized trials, if \( t = 1 \)
denotes baseline, then \( \gamma_{t} = 0 \), since no difference at baseline is expected. If treatment difference is expected over the follow-up time, then \( \gamma_{t} \neq 0 \) for some or all remaining \( t \) (\( 2 \leq t \leq m \)). Thus, the null \( H_{0} : \gamma = 0 \) indicates no between-group difference at any of the follow-up times. We may also model temporal patterns for each group using linear, piece-linear or polynomial functions as in standard regression analysis. In the absence of covariates, the FRM in Equation (7) is equivalent to
\[
E[f(y_{iut}, y_{jk}t)] = h_{kt}(\theta) = \theta_{kt}, \quad \theta_{t} = (\theta_{2t}, \ldots, \theta_{Kt})^\top, \quad \theta = (\theta_{1}^\top, \ldots, \theta_{m}^\top)^\top,
\]
where \( \theta_{kt} = \text{logit}^{-1}(y_{kt}) \) (\( 1 \leq k \leq 2, 1 \leq t \leq m \)).

3. Inference for the generalized MWW model

We start with the cross-sectional data and then extend the considerations to the longitudinal setting.

3.1 Cross-sectional data

Consider the \( K \)-group FRM in Equation (6) and let
\[
f_{ki} = f(y_{1i}, y_{ki}), \quad h_{ki} = h_{k}(x_{1i}, x_{kj}; \theta), \quad i = (i, j) \in C_{1}^{kk}, \quad 2 \leq k \leq K,
\]
where \( f(y_{1i}, y_{ki}) \) and \( h_{k}(x_{1i}, x_{kj}; \theta) \) are defined in Equation (6), and \( C_{1}^{kk} = C_{1}^{nk} \otimes C_{1}^{nk} \) with \( \otimes \) denoting the Cartesian product of \( C_{1}^{nk} \) and \( C_{1}^{nk} \) (\( 1 \leq k \leq K \)). Define the U-statistics-based generalized estimating equations (UGEE) as follows:
\[
U_n(\theta) = \sum_{k=2}^{K} \frac{1}{n_{1}n_{k}} \sum_{i \in C_{1}^{kk}} U_{kni}(\theta) = \sum_{k=2}^{K} \frac{1}{n_{1}n_{k}} \sum_{i \in C_{1}^{kk}} D_{ki} V_{ki}^{-1} S_{ki} = 0,
\]
where \( S_{ki} = f_{ki} - h_{ki}, D_{ki} = (\partial/\partial \theta)h_{ki} \) and \( V_{ki} \) is given by
\[
V_{ki} = \text{Var}(f(y_{1i}, y_{ki}) | x_{1i}, x_{kj}) = h_{k}(x_{1i}, x_{kj}; \theta)(1 - h_{k}(x_{1i}, x_{kj}; \theta)).
\]

For example, for the classic MWW test in Equation (3), the UGEE can be solved in closed form to obtain \( \hat{\theta} = (1/n_{1}n_{2}) \sum_{i \in C_{1}^{kk}} f(y_{1i}, y_{2i}) = (1/n_{1}n_{2}) \sum_{i \in C_{1}^{kk}} I_{[y_{1i} \leq y_{2i}]} \). In general, numerical methods such as the Newton–Raphson may be used to solve the UGEE.

As in the standard generalized estimating equations (GEE) \([10]\), the UGEE estimate \( \hat{\theta} \) obtained as solution to Equation (10) is consistent and asymptotically normal. We summarize the asymptotic properties below. As a special case of longitudinal data, Theorem 1 follows from asymptotic results that will be established for longitudinal data in the next section.

**Theorem 1**

Let
\[
v_{li} = \begin{cases} 
E\left(\sum_{k=2}^{K} U_{kni} \mid y_{1i}, x_{li}\right) & \text{if } l = 1, \\
E(U_{lni} \mid y_{li}, x_{li}) & \text{if } l \geq 2,
\end{cases} \quad B = \sum_{k=2}^{K} E(D_{ki} V_{ki}^{-1} D_{ki}^\top),
\]
\[
D_{ki} = \frac{\partial}{\partial \theta} h_{ki}, \quad \Sigma_{U} = \sum_{l=1}^{K} \rho_{l}^2 \Sigma_{l}, \quad \Sigma_{l} = \text{Var}(v_{li}), \quad \Sigma_{\theta} = B^{-1} \Sigma_{U} B^{-1},
\]
\[
n = \sum_{l=1}^{K} n_{l}, \quad \rho_{l}^2 = \lim_{n \to \infty} \frac{n}{n_{l}} < \infty.
\]
Then, under mild regularity conditions, \( \hat{\theta} \) is consistent and asymptotically normal

\[
\sqrt{n}(\hat{\theta} - \theta) \to_d N(0, \Sigma_{\theta}).
\]

The asymptotic variance \( \Sigma_{\theta} \) generally is not expressed in closed form, except under some special cases. For example, for the classic MWW test in Equation (3), \( K = 2 \) and \( B = 1 \). It follows that

\[
\Sigma_{\theta} = \Sigma_U = \rho_1^2 \Sigma_1 + \rho_2^2 \Sigma_2, \quad \Sigma_1 = E(1 - F_2(y_{1i}))^2 - \theta, \quad \Sigma_2 = E(F_2^2(y_{2i})) - \theta,
\]

where \( F_2(y) \) denotes the cumulative distribution function of \( y_{ki} \). Further, under the null \( H_0 : F_2(y) = F(y), \theta = \frac{1}{2} \). It follows from Equation (13) and the fact that \( F(y_{ki}) \) is a uniform \( U \) between 0 and 1 that

\[
\Sigma_1 = E(1 - U)^2 - \frac{1}{4} = \frac{1}{12}, \quad \Sigma_2 = \frac{1}{12}.
\]

Thus in this special case, \( \Sigma_{\theta} = \frac{1}{12}(\rho_1^2 + \rho_2^2) \) and a consistent estimate is given by \( \hat{\Sigma}_{\theta} = \frac{1}{12}(n/n_1 + n/n_2) \) \( (n = n_1 + n_2) \).

For general \( K \) and \( h(x_{1i}, x_{ki}; \theta) \), a consistent estimate of \( \Sigma_{\theta} \) is obtained by substituting respective consistent estimates in place of \( B \) and \( \Sigma_k \). A consistent estimate of \( B \) is

\[
\hat{B} = \sum_{k=2}^{K} \frac{1}{n_1 n_k} \sum_{i \in C_1} \hat{D}_{ki} \hat{V}_{ki}^{-1} \hat{D}_{ki}^T,
\]

where \( \hat{A}_i \) denotes \( A_i \) with \( \theta \) replaced by \( \hat{\theta} \). To find a consistent estimate of \( \Sigma_l \), note that

\[
\hat{\Sigma}_l = \text{Var}(\bar{v}_l) = \frac{1}{n_l - 1} \sum_{i=1}^{n_l} \hat{v}_{li} \hat{v}_{li}^T, \quad l = 1, \ldots, K,
\]

where

\[
\hat{v}_{li} = \begin{cases} \sum_{k=2}^{K} \frac{1}{n_k} \sum_{j \in C_i} \hat{D}_{kij} \hat{V}_{kij}^{-1} \hat{S}_{kij} & \text{if } l = 1, \\ \frac{1}{n_1} \sum_{j \in C_i} \hat{D}_{lij} \hat{V}_{lij}^{-1} \hat{S}_{lij} & \text{if } l \geq 2. \end{cases}
\]

Thus, a consistent estimate of \( \Sigma_{\theta} \) is given by

\[
\hat{\Sigma}_{\theta} = \hat{B}^{-1} \hat{\Sigma}_U^{-1} \hat{B}, \quad \hat{\Sigma}_U^{-1} = \sum_{l=1}^{K} \frac{n_l}{n} \hat{\Sigma}_l,
\]

where \( \hat{B} \) and \( \hat{\Sigma}_l \) are given in Equations (15) and (16).

Under the null \( H_0 : L\theta = b \), where \( L \) denotes a known matrix and \( b \) a known vector of constants, we can use the Wald statistic

\[
W = n(\hat{\theta} - \theta_0)^{-2} \sim \chi^2_{\text{df}}, \quad \hat{\theta} = \frac{1}{n_1 n_2} \sum_{i \in C_1} I_{(y_{1i} \leq y_{2i})}, \quad \hat{\Sigma}_{\theta} = \frac{1}{12} \left( \frac{n_1 + n_2}{n_1 + n_2} \right).
\]
3.2 Longitudinal data

We first consider the case of complete data and then generalize it to address missing values.

3.2.1 Complete data

Let

\[
\begin{align*}
    f_{ik} &= f(y_{1ik}, y_{2ik}), \\
    h_{ik} &= h_k(x_{1ik}, x_{2ik}; \theta), \\
    f_i &= (f_{ik1}, f_{ik2}, \ldots, f_{ikm})^T, \\
    h_i &= (h_{ik1}, h_{ik2}, \ldots, h_{ikm})^T, \quad i \in C^k, \quad 2 \leq k \leq K.
\end{align*}
\]  

(18)

Inference about \( \theta \) can be made using the UGEE with the same form as in Equation (10), but with \( f_{ik} \) and \( h_k(\theta) \) redefined in Equation (18) to account for multiple assessments. In addition, we need to redefine \( V_{ik} \) to reflect the added time dimension in \( f_{ik} \) and \( h_i \) as well as the correlations among the repeated responses \( y_{kij} \). For example, consider extending \( V_{ik} \) in Equation (11) to account for the within-subject correlation as follows:

\[
V_{ik} = A_{ik} R(\varphi) A_{ik}^T, \quad A_{ik} = \text{diag}(V_{ik}), \quad R(\varphi) = C_{m \times m}(\rho), \quad V_{ik} = h_{ik}(1 - h_{ik}),
\]

(19)

where \( \text{diag}(A_r) \) denotes a diagonal block matrix with \( A_r \) on the \( r \)th diagonal and \( C_{m \times m}(\rho) \) a compound symmetry correlation matrix with \( \rho \) being the correlation between any two components of \( f_{ik} \).

As in the standard GEE [10], \( R(\varphi) \) is a working correlation matrix, parameterized by \( \varphi \) to account for correlations among the repeated \( y_{kij} \). If \( \varphi \) is unknown, as in most applications, it must be estimated before the UGEE in Equation (10) can be solved for \( \theta \). Although the consistency of \( \hat{\theta} \) is independent of how \( \varphi \) is estimated, the asymptotic normality of \( \hat{\theta} \) is guaranteed only when \( \sqrt{n} \)-consistent estimates of \( \varphi \) are used [8, 10]. Thus, to use the results in Theorem 1 for the longitudinal data setting, we must substitute a \( \sqrt{n} \)-consistent estimate of \( \varphi \). In particular, under the working independence model, \( \varphi \) is known and the UGEE estimate is asymptotically normal. We focus on the working independence model in what follows unless otherwise stated.

As a special case of the MWW rank sum test, \( K = 2, h_{ik}(\theta) = \theta_i \) and \( \theta = (\theta_1, \ldots, \theta_m)^T \). The UGEE in this case can be solved in closed form

\[
\hat{\theta} = \frac{1}{n_1n_2} \sum_{i \in C^{12}} f_{2i} = \frac{1}{n_1n_2} \sum_{i \in C^{12}} (I_{\{y_{1i1} \leq y_{2i1}\}}, I_{\{y_{1i2} \leq y_{2i2}\}}, \ldots, I_{\{y_{1im} \leq y_{2im}\}})^T.
\]

(20)

It follows from Theorem 1 that the UGEE estimate \( \hat{\theta} \) in Equation (20) is consistent and asymptotically normal. The asymptotic variance is readily estimated and, following Theorem 1, is given by

\[
\begin{align*}
    \hat{\Sigma}_0 &= \frac{n}{n_1} \hat{\Sigma}_1 + \frac{n}{n_2} \hat{\Sigma}_2, \\
    \hat{\Sigma}_l &= \frac{1}{n_l - 1} \sum_{i=1}^{n_l} \hat{\nu}_i\hat{\nu}_i^T, \\
    \hat{\nu}_1i &= \frac{1}{n_2} \sum_{j \in C^{12}_i} \hat{\nu}_2ij^T \hat{S}_2ij, \quad \hat{\nu}_2i = \frac{1}{n_1} \sum_{j \in C^{12}_i} \hat{\nu}_2ij^T \hat{S}_2ij.
\end{align*}
\]

(21)

Under the null of no between-group difference over all assessments, i.e. \( H_0 : \theta = \frac{1}{2} \mathbf{1}_m \), a \( n \times 1 \) column vector of 1’s, the Wald statistic for testing this null is given by

\[
W = n(\hat{\theta} - \frac{1}{2} \mathbf{1}_m)^T \hat{\Sigma}_0^{-1}(\hat{\theta} - \frac{1}{2} \mathbf{1}_m) \sim \chi^2_m.
\]

(22)
3.2.2 Monotone missing data

Define a vector of missing (or rather observed) value indicators as follows:

\[
    r_{lit} = \begin{cases} 
        1 & \text{if } y_{lit} \text{ is observed}, \\
        0 & \text{if } y_{lit} \text{ is unobserved}, 
    \end{cases} \\
    \pi_{lit} = (r_{l1}, r_{l2}, \ldots, r_{lim})^T,
\]

\[
x_{li} = (x_{l1i}, \ldots, x_{limi})^T, \quad y_{li} = (y_{l1i}, \ldots, y_{limi})^T, \\
\pi_{lit} = E(r_{lit} | y_{li}, x_{li}), \quad i \in C^{li}_1, \quad 1 \leq l \leq K, \quad 1 \leq t \leq m.
\]

We assume no missing data at baseline \( t = 1 \) such that \( r_{l1i} \equiv 1 \) for all \( i \in C^{li}_1 \) and \( 1 \leq l \leq K \). Let

\[
    \phi_{lit} = \frac{r_{lit}}{\pi_{lit}}, \quad \phi_{k\ell} = \phi_{1\ell} / \phi_{k\ell}, \quad \Delta_{k\ell} = \text{diag}(\phi_{k\ell}), \quad i = (i, j) \in C^{lk}, \\
    1 \leq l \leq K, \quad 1 \leq i \leq n, \quad 1 \leq t \leq m. \tag{24}
\]

Since \( r_{l1} \equiv 1 \), \( \pi_{l1} = 1 \) for all \( i \in C^{li}_1 \) and \( 1 \leq l \leq K \).

In the presence of missing data, the UGEE in Equation (10) generally yields biased estimates. By integrating the inverse probability weighting technique into the UGEE to account for the selection bias due to missing data, akin to weighting selected households sampled from a targeted region of interest in survey research [6], we obtain a set of U-statistics-based weighted generalized estimating equations (UWGEE) for inference about \( \theta \):

\[
    U_n(\theta) = \sum_{k=2}^{K} \sum_{i \in C^{li}} U_{kn, i} = \sum_{k=2}^{K} \sum_{i \in C^{li}} \sum_{j=1}^{K} D_{k\ell} V_{k\ell}^{-1} \Delta_{k\ell} S_{k\ell} = 0, \tag{25}
\]

where \( S_{k\ell} = f_{k\ell} - h_{k\ell} \), and \( D_{k\ell}, V_{k\ell}, S_{k\ell} \) are defined the same as in the complete data case. Unlike the UGEE, the above has an additional weight function \( \Delta_{k\ell} \) defined by \( \pi_{lit} \). Given \( \pi_{lit} \), the UWGEE in Equation (25) is again readily solved using numerical methods such as the Newton–Raphson algorithm.

In most applications, \( \pi_{k\ell i} \) is unknown and must be estimated. Under missing completely at random (MCAR), \( \pi_{lit} \) is independent of \( y_{li} \) and thus \( \pi_{lit} = E(r_{lit}) = \pi_{lit} \). In this case, \( \pi_{lit} \) is a constant independent of \( y_{li} \) and is readily estimated by the sample moment:

\[
    \hat{\pi}_{lit} = \frac{1}{n_l} \sum_{i=1}^{n_l} r_{lit}, \quad 1 \leq l \leq K, \quad 2 \leq t \leq m.
\]

Under missing at random (MAR), \( \pi_{lit} \) becomes dependent on the observed \( y_{lit} \) and \( x_{lit} \), which, within the current context, contain all \( y_{lis} \) and \( x_{lis} \) prior to time \( t (1 \leq s \leq t - 1) \). Denote such a ‘history’ by \( z_{lit} = (y_{lis}, x_{lis}; s = 1, \ldots, t - 1)^T \). Then under MAR

\[
    \pi_{lit} = \begin{cases} 
        1 & \text{if } t = 1, \\
        E(r_{lit} | z_{lit}) & \text{if } 2 \leq t \leq m. 
    \end{cases} \tag{26}
\]

Thus, under MAR \( r_{lit} \) only depend on the observed data \( z_{lit} \), making it possible to model \( \pi_{lit} \) in Equation (26) based on the observed data prior to time \( t \). However, it is still difficult to model and estimate \( \pi_{lit} \) without imposing the monotone missing data pattern (MMDP) assumption, because of the large number of potential missing data patterns [8,15,20]. Under MMDP, \( y_{lis} \) and \( x_{lis} \) are observed, only if all \( y_{lis} \) and \( x_{lis} \) prior to time \( t \) are all observed. The structured patterns reduce not only the number of missing data patterns, but also the complexity in modeling \( \pi_{lit} \).
Let \( p_{lt} = E(r_{lit} \mid r_{lit(-1)} = 1, z_{lt}) \) denote the one-step transition probability of observing the response at \( t - 1 \) to \( t \). We can readily model \( p_{lt} \) using a logistic regression model:

\[
p_{lt} = \logit^{-1}(\eta_{0lt} + \eta_{1lt}^Tz_{lt}), \quad i \in C_1^m, \ 1 \leq l \leq K, \ 2 \leq t \leq m, \tag{27}
\]

where \( \eta_{lt} = (\eta_{0lt}, \eta_{1lt}^T)^T \) denotes the model parameters. More complex models may be considered such as including some or all interactions of the components of \( z_{lt} \). Under MMDP, it is readily checked that

\[
\pi_{lt}(\eta_l) = E(r_{lt(-1)} \mid z_{lt}) = \prod_{s=2}^t p_{lis}(\eta_{ls}), \quad 1 \leq l \leq K, \ 2 \leq t \leq m, \tag{28}
\]

where \( \eta_l = (\eta_{l2}^T, \ldots, \eta_{lnl}^T)^T \ (1 \leq l \leq K) \).

Thus, we can estimate \( \pi_{lt} \) by estimating \( \eta_l \) of \( p_{lt} \). To estimate \( \eta_l \), we may use maximum likelihood or estimating equations. Suppose \( \eta_l \) is estimated based on the following estimating questions:

\[
Q_{nl}(\eta_l) = \sum_{i=1}^{n_l} Q_{nl,i} = 0, \quad Q_{nl,i} = (Q_{nl,2}^T, \ldots, Q_{nl,m}^T)^T, \quad 1 \leq l \leq K, \tag{29}
\]

where

\[
Q_{nl,i} = \frac{\partial}{\partial \eta_{lit}} \{r_{lit(-1)}r_{lit} \log(p_{lit}) + (1 - r_{lit}) \log(1 - p_{lit})\}, \quad t = 2, \ldots, m, \ i \in C_1^m, \ 1 \leq l \leq K. \tag{30}
\]

Let

\[
W_{nl,i}(\eta_l) = (0_{(m-1) \times (l-1)}, Q_{nl,i}^T, 0_{(m-1) \times (K-l)}^T)^T, \quad 1 \leq l \leq K,
\]

where \( 0_j \) denotes a \( J \times 1 \) column vector of \( 0 \)'s. Let \( \eta = (\eta_2^T, \ldots, \eta_m^T)^T \). We may express Equation (29) in a compact form

\[
W_n(\eta) = \sum_{l=1}^{K} (\sum_{i=1}^{n_l} W_{nl,i}(\eta_l) = \left( \sum_{i=1}^{n_1} Q_{n1,i}, \ldots, \sum_{j=1}^{n_K} Q_{nK,j} \right)^T = 0. \tag{31}
\]

Upon estimating \( \eta \), we can estimate \( \Delta_{kl} \) and then estimate \( \theta \) by substituting the estimated \( \Delta_{kl} \) into the UWGEE in Equation (25). Since the UWGEE estimate \( \hat{\theta} \) is based upon estimated \( \eta \), we must also account the variability of the estimate of \( \eta \) when estimating the variance of \( \hat{\theta} \). The theorem extends Theorem 1 to take this extra variability into consideration when constructing the variance of \( \hat{\theta} \). A proof is sketched in the appendix:

\[
U_n(\theta) = \sum_{k=2}^{K} \frac{1}{n_1 n_k} \sum_{i \in C^u} U_{kn,i} = \sum_{k=2}^{K} \frac{1}{n_1 n_k} \sum_{i \in C^l} D_{ki} \Delta_{kl} = 0. \tag{32}
\]

**Theorem 2** Let \( \hat{\theta} \) denote the UWGEE estimate from solving the estimating equations in Equation (25) upon substituting the estimate \( \hat{\eta} \) obtained from the estimating equations in
Equation (31) and some estimate \( \hat{\phi} \). Let

\[
\nu_{li} = \begin{cases} 
E \left( \sum_{k=2}^{K} U_{kn,i} | y_{li}, x_{li} \right) & \text{if } l = 1, \\
E(U_{ln,i} | y_{li}, x_{li}) & \text{if } l \geq 2,
\end{cases}
\]

\[
\Sigma_l = \text{Var}(\nu_{li}), \quad Q_l = \text{Var}(W_{nl,i}), \quad C = \sum_{k=2}^{K} E \left( \left( \frac{\partial}{\partial y} \Delta_{ki} S_{ki} \right) V_{ki}^{-1} D_{ki}^\top \right),
\]

\[
H = \sum_{l=1}^{K} \rho_l^2 E \left( \frac{\partial}{\partial \eta} W_{nl,i}(\eta_k) \right)^\top, \quad \Phi_l = CH^{-1}Q_lH^{-\top}C^\top - F_l - F_l^\top,
\]

\[
F_l = E(\nu_{li} W_{nl,i}^\top H^{-\top} C^\top), \quad \Sigma_\theta = \sum_{l=1}^{K} \rho_l^2 B^{-1}(\Sigma_l + \Phi_l)B^{-\top},
\]

(32)

where \( n \) and \( \rho_l^2 (1 \leq l \leq K) \) are defined in Equation (12). Then, under mild regularity conditions, we have the following:

(1) \( \hat{\theta} \) is consistent;
(2) if \( \hat{\phi} \) is \( \sqrt{n} \)-consistent, \( \hat{\theta} \) is asymptotically normal:

\[
\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d N(0, \Sigma_\theta).
\]

(33)

The asymptotic variance of \( \hat{\theta} \) is identical to its counterpart in Equation (13) of Theorem 1, except for an added term \( \Phi_l \), which accounts for the additional variability due to estimation of \( \eta \). A consistent estimate of \( \Sigma_\theta \) is obtained by substituting consistent estimates in place of the respective quantities in Equation (33). For example, a consistent estimate of \( B \) is \( \hat{B} = \sum_{k=2}^{K} (1/n_l n_k) \sum_{i \in X} D_{ki} \Delta_{ki} \hat{V}_{ki}^{-1} \hat{D}_{ki}^\top \), where \( A_i \) denotes \( A_i \) with \( (\theta, \eta, \phi) \) replaced by \( (\hat{\theta}, \hat{\eta}, \hat{\phi}) \). Note that Theorem 2 shows that the asymptotic variance of \( \hat{\theta} \) obtained by ignoring the sampling variability of \( \hat{\eta} \) underestimates the true variability of \( \hat{\theta} \). Also, Theorem 2 applies to the most general case where \( R(\phi) \) may be non-identity matrices. However, we again focus on the working independence model for both in this paper unless stated otherwise.

4. Applications

We demonstrate our considerations with both simulated and real data. We first investigate the performance of the proposed approach by simulation and then present an application to a real study on sexual health for a group of teenage girls in low-income urban settings who were at elevated risk for HIV, sexually transmitted infections (STIs) and unintended pregnancies. In all examples, we set the statistical significance at \( \alpha = 0.05 \) and Monte Carlo (MC) sample size equal to 1000.
4.1 Simulation study

4.1.1 Cross-sectional data

Consider the following analysis of covariance model:

\[
y_{li} = \beta_0 + \theta^\top x_{li} + \epsilon_{li} = \beta_0 + x_{li1} \beta + I_{|l=2}\gamma_2 + I_{|l=3}\gamma_2, \\
\epsilon_{li} \sim N \left(0, \frac{1}{\sqrt{2}}\right), \quad x_{li1} \sim N(0, \sigma_x^2), \quad \theta = (\beta, \gamma_2, \gamma_3)^\top,
\]

(34)

The above models three group means, represented by \(\beta_0, \beta_0 + \gamma_2\) and \(\beta_0 + \gamma_3\), while adjusting for one covariate \(x_{li1}\). It follows from Equation (34) that:

\[
y_{1i} - y_{kj} = \theta^\top (x_{li} - x_{kj}) + \epsilon_{li} - \epsilon_{kj},
\]

Thus, we have

\[
E(I_{\{y_{1i} - y_{kj} \leq 0\} \mid x_{1i}, x_{kj}}) = \Pr[\epsilon_{1i} - \epsilon_{kj} \leq -\theta^\top (x_{1i} - x_{kj})] = \Phi(-\theta^\top (x_{1i} - x_{kj}))
\]

\[
= \Phi(-\beta(x_{1i1} - x_{kj1}) + I_{|l=2}\gamma_2 + I_{|l=3}\gamma_3),
\]

where \(\Phi(\cdot)\) is the cumulative distribution function of the standard normal. Let

\[
f(y_{1i}, y_{kj}) = I_{\{y_{1i} - y_{kj} \leq 0\}}, \quad h_k(x_{1i}, x_{kj}; \theta) = \Phi(-\beta(x_{1i1} - x_{kj1}) + I_{|l=2}\gamma_2 + I_{|l=3}\gamma_3).
\]

Then, except for the absence of the intercept \(\beta_0\), the FRM below

\[
E[f(y_{1i}, y_{kj}) \mid x_{1i}, x_{kj}] = h_k(x_{1i}, x_{kj}; \theta), \quad 1 \leq i \leq n_l, \quad 1 \leq l \leq 3,
\]

(35)

contains the same parameters \(\theta\) as the linear model in Equation (34) and thus can be used to provide inference about \(\theta\).

If \(y_{li}\) follows the normal distribution in Equation (34), maximum likelihood provides the most efficient estimate of \(\theta\). However, if there are outlying observations of \(y_{li}\), maximum likelihood, or even the estimating equations [10], generally yields biased inference. In contrast, the FRM in Equation (35) is not affected by such outliers and continues to provide valid inference, if the ranking order of the observations is preserved.

For our simulation study, we set \(\beta_0 = \beta = \gamma_2 = \gamma_3 = 1\) and \(\sigma_x^2 = 0.2\). We considered three sample sizes 50, 100 and 300, representing small, moderate and 300 large sample sizes. Given each sample size, we simulated three group sizes \(n_l\) from a trinomial with the probability \(\frac{1}{3}\) for each of the three cells. For each of the 1000 simulated data sets, we fit the FRM in Equation (35) and computed estimates of \(\theta\) and asymptotic standard errors and tested the null that the value of parameter was equal to the true value for each of the three components of \(\theta\) based on the Wald statistic. We also computed the empirical standard errors based on the 1000 MC estimates of \(\theta\) and type I error rates based on the percent of times the null was rejected over the 1000 MC replications.

The UGEE estimates of \(\theta\) are shown in Table 1, along with standard errors (asymptotic and empirical) and type I errors based on 1000 MC replications. The UGEE estimates of \(\hat{\theta}\) were quite close to the true values, even for the small sample size \(n = 50\). The asymptotic and empirical standard errors also agreed quite well with each other, even for the small sample size. Although the type I error rates showed a small upward bias, typical for small sample sizes [1,4,16], the bias quickly decreased as the sample size increased.
Table 1. UGEE (for complete data) and UWGEE (for missing data) estimates, standard errors and type I errors (for testing the null of no time effect) for simulated longitudinal under complete and MAR missing data.

| Sample size | Estimate/ asymptotic standard error/ empirical standard error/ type I error |
|-------------|--------------------------------------------------------------------------|
| 50          | 1.06/0.56/0.60/0.070, 1.07/0.29/0.30/0.072, 1.07/0.29/0.31/0.064      |
| 100         | 1.03/0.39/0.41/0.068, 1.02/0.19/0.20/0.054, 1.02/0.19/0.20/0.051      |
| 300         | 1.01/0.22/0.22/0.046, 1.01/0.11/0.11/0.059, 1.01/0.11/0.11/0.056      |

4.1.2 Longitudinal data

We considered a longitudinal study with two groups and three assessments and simulated the longitudinal responses \( y_{ki} = (y_{k1}, y_{k2}, y_{k3}) \) for the two groups from a trivariate normal, \( N(0, C_3(0.5)) \), with \( C_3(0.5) \) denoting a \( 3 \times 3 \) compound symmetry correlation matrix. We applied the FRM in Equation (8) to compare the two groups based on the data generated. In this particular case, \( K = 2, m = 3 \) and the vector of parameters of the FRM is \( \theta = (\theta_1, \theta_2, \theta_3)^T \). We considered three sample sizes, 50, 100 and 300 per group, representing small, moderate and large sample sizes, respectively. We increased the sample size in this simulation study over the prior one to account for decreased sample sizes at the two follow-up assessments due to missing data.

To assess performance of the FRM under missing data, we assumed no missing value at baseline \( t = 1 \) and simulated the missing response \( r_{ki, t} \) at the follow-up times under MAR according to Equation (27) with the one-step transition probability \( p_{ki, t} \) modeled by the logistic regression under a one-step Markov condition below:

\[
pli = \logit^{-1}(\eta_{0lt} + \eta_{1lt}y_{li(t-1)}), \quad t = 2, 3, l = 1, 2. \tag{36}
\]

In the above, we assumed a Markov condition so that \( pli \) only depends on the most recent \( y_{li(t-1)} \) prior to time \( t \). We set \( \eta_{0lt} = 3 \) \((t = 2, 3)\) and solved the following equations for \( \eta_{0lt} \) to create about 15% and 25% missing response \( y_{li(t)} \) at time \( t = 2, 3 \), respectively:

\[
\sum_{i=1}^{n_l} p_{li2} = 0.85n_l, \quad \sum_{i=1}^{n_l} p_{li2}p_{li3} = 0.75n_l, \quad l = 1, 2. \tag{37}
\]

To ensure MMDP, we first simulated missing data indicator \( r_{li2} \) from the Bernoulli distribution, \( \text{Bern}(p_{li2}) \) \((l = 1, 2)\). If \( r_{li2} = 0 \), then set \( r_{li3} = 0 \). Otherwise, \( r_{li2} = 1 \) and \( r_{li3} \) is simulated from \( \text{Bern}(p_{li3}) \).

In the absence of missing data, the UGEE estimate of \( \theta \) is given by Equation (20), which is asymptotically normal with a consistent estimate of the asymptotic variance given by Equation (21). In the presence of missing data, the UWGEE estimate does not admit a closed-form expression, but is readily obtained by solving the UWGEE using the numerical methods. A consistent estimate of the asymptotic variance is also readily computed according to Theorem 2.

The UGEE and UWGEE estimates of \( \theta \) are shown in Table 2, along with standard errors and type I errors (for testing the null of no time effect, i.e. \( H_0: \theta_1 = \theta_2 = \theta_3 \)) under both complete and missing data cases based on 1000 MC replications. As seen, both the UGEE and UWGEE estimates of \( \hat{\theta} \) were quite accurate, even for the small sample size \( n_l = 50 \) per group. The standard errors showed a steady decrease as \( n_l \) increased. Also, the corresponding standard errors were
Table 2. UGEE (for complete data) and UWGEE (for missing data) estimates, standard errors and type I errors (for testing the null of no time effect) for simulated longitudinal under complete and MAR missing data.

| Sample size (per group) | $\theta_1$          | $\theta_2$          | $\theta_3$          | Type I error |
|-------------------------|---------------------|---------------------|---------------------|--------------|
| **Complete data (UGEE)**|                     |                     |                     |              |
| 50                      | 0.497 (0.057)       | 0.502 (0.058)       | 0.501 (0.059)       | 0.054        |
| 100                     | 0.499 (0.042)       | 0.502 (0.045)       | 0.502 (0.048)       | 0.051        |
| 300                     | 0.499 (0.023)       | 0.500 (0.024)       | 0.499 (0.025)       | 0.054        |
| **Missing data under MAR (UWGEE)**|                   |                     |                     |              |
| 50                      | 0.500 (0.061)       | 0.504 (0.071)       | 0.499 (0.069)       | 0.069        |
| 100                     | 0.502 (0.043)       | 0.501 (0.057)       | 0.501 (0.049)       | 0.051        |
| 300                     | 0.500 (0.024)       | 0.499 (0.037)       | 0.500 (0.029)       | 0.051        |

slightly larger for the UWGEE estimates because of missing data at the two follow-up times. As in the cross-sectional data case, the type I error based on the Wald statistic again showed a small upward bias for the small sample size, but the bias immediately disappeared once the sample size increased to 100 and 300. The results (not shown) from the logistic regression in Equation (36) for the missing data were also quite close to the true values.

4.2 Real study

Teenage girls in low-income urban settings are at elevated risk for HIV, STIs and unintended pregnancies. A randomized controlled trial was recently conducted to evaluate the efficacy of a sexual risk-reduction intervention, supplemented with post-intervention booster sessions, targeting low-income, urban, sexually active teenage girls [13]. The sexually active urban adolescent girls aged 15–19 ($n = 738$) recruited in the study in Rochester, NY, were randomized to a theory-based, sexual risk-reduction intervention or to a structurally equivalent health promotion control group. Assessments and behavioral data were collected at baseline, and then again at three and six months post-intervention. The primary interest of the study is to compare frequency of unprotected vaginal sex between the intervention and controlled condition. More details about the demographic characteristics of the study population, the treatment conditions and the assessment battery can be found in [13].

As mentioned in Section 1, a difficult problem with the data is the extremely large values some subjects reported with respect to their sexual activities. For example, 7 subjects reported over 100 episodes of unprotected vaginal sex over the past three months at the three-month follow-up, with the largest one being 1,000,000. A common approach to this issue in psychosocial research is to trim such outliers using some ad hoc rules [17]. However, these methods induce artifacts, because of their dependence on the specific rules used. Rank-based approaches such as the proposed FRM model more effectively address this issue.

Our analysis was based on 639 ($n_1 = 310, 49\%$ in the control and $n_2 = 329, 51\%$ in the intervention). We were interested in comparing the two treatment groups for the unprotected vaginal sex. Let $y_{lt}$ ($r_{lt}$) denote such an outcome (indicator for missing data) at time $t$ ($l$) for the $l$th treatment condition ($l = 1$ for the control and 2 for the intervention). As in the simulation study, we applied the FRM in (25) by jointly modeling $y_{lt}$ and $r_{lt}$. Since this is a randomized trial, we did not include any covariate...
and thus considered the following FRM for comparing the two treatment groups:

\[ E[f(y_{1it}, y_{2jt})] = h_i(\theta) = \theta, \quad i = (i,j) \in C^{12}, \quad \theta = (\theta_1, \theta_2, \theta_3)^T. \]

We modeled the missing data under MAR using logistic regression as discussed in Section 3.2. However, since some of \( y_{ilt} \) were extremely large, we used the rankings \( R_{li(t−1)} \), rather than the actual values of \( y_{li(t−1)} \), as the predictor of the logistic model:

\[ p_{li(t)} = \logit^{-1}(\eta_0 lt + \eta_1 lt R_{li(t−1)}), \quad l = 2, 3, \quad t = 2, 3. \quad (38) \]

Note that we used a Markov assumption in Equation (38) so that the missingness at time \( t \) only depends on the observed response at time \((t−1)\). Also, as ties are inevitable for the intrinsically discrete \( y_{ilt} \) in this application, we used

\[ f(y_{1lt}, y_{2lt}) = I\{y_{1lt} < y_{2lt}\} + \frac{1}{2} I\{y_{1lt} = y_{2lt}\} \quad (39) \]

as the functional response in the FRM to account for their presence. Under Equation (39)

\[ \theta_t = E(I\{y_{1lt} < y_{2lt}\}) + \frac{1}{2} E(I\{y_{1lt} = y_{2lt}\}), \]

and the null of no treatment difference at time \( t \) is \( H_0 : \theta_t = \frac{1}{2} (1 \leq t \leq 3) \).

The UWGEE estimates of the intercept (\( \eta_0 t \)) and slope (\( \eta_1 t \)) from Equation (38) for missingness at three and six months post-intervention are shown in Table 3. The occurrence of missing data did not depend on the value (ranking) of the outcome at the prior visit for either treatment condition, suggesting no evidence for rejecting MCAR. The estimated \( \theta \), standard errors and \( p \)-values for testing the null of between-group difference at each assessment time, along with the test statistics and \( p \)-values for testing the null of no temporal trend during post-treatment, are shown in Table 4. The estimated \( \theta_t \) showed a steady decrease for those in the intervention condition to engage in unprotected vaginal sex over time from baseline to three and six months post-intervention. The decline was significant at six months, but was a trend at three months (\( p \)-value is close to 0.10). The continued decline from three to six months was confirmed by the near significant \( p \)-value for testing the null \( H_0 : \theta_2 = \theta_3 \). The increased gain of the intervention effect at six months was likely due to two booster sessions the study subjects received at three months. The booster sessions, 90-min long (as opposed to four 120-min regular sessions delivered during the intervention), address behavioral patterns of girls that are expected to occur as they age and can promote maintenance of gains observed with health-behavior interventions [2].

| Table 3. UWGEE estimates of parameters of logistic regression for modeling missingness at three and six months post-intervention for Sexual Health Study. |
|---|---|---|---|
| | Month 3 | | Month 6 |
| | Intercept Control/ treat | Prior response Control/ treat | Intercept Control/ treat | Prior response Control/ treat |
| **Unprotected vaginal sex** | | | |
| Estimate | 1.386/2.036 | 0.001/−0.0006 | 2.246/2.424 | 0.002/0.003 |
| Standard error | 0.349/0.370 | 0.001/0.001 | 0.488/0.571 | 0.002/0.003 |
| \( p \)-Value | <.001/=.001 | .157/.571 | <.001/=.001 | .436/.280 |
Table 4. UWGEE estimates of parameters for comparing the number of unprotected vaginal sex from baseline to 3–6 months post-intervention for Sexual Health Study.

| Estimate | Standard error | p-Value for testing $H_0 : \theta_i = \frac{1}{2}$ |
|----------|----------------|-----------------------------------------------|
| Baseline | 0.508          | 0.566                                         | 0.733 |
| Month 3  | 0.462          | 0.600                                         | 0.112 |
| Month 6  | 0.441          | 0.616                                         | 0.017 |

Test statistic (p-value) for testing no differential treatment effect between three and six months: $H_0 : \theta_2 = \theta_3$

5. Discussion

In this paper, we extended the classic MWW test to accommodate more than two groups, covariates and longitudinal data. We achieved this generalization by utilizing the FRM, which is uniquely positioned to model relationships involving between-subject interactions, rather than attributes of a single subject, such as the MWW rank sum test within our context. For valid inference, we also considered two approaches based on a set of U-statistics-based estimating equations. The first follows the traditional route, akin to WGEE, by estimating the primary and missing data models separately, but unlike other applications the inverse probability weighting technique offers valid inference by accounting for the effect of sampling variability of the estimated weight on the estimate of the primary model. The second avoids the complex adjustment procedure by making joint inference about both the primary and missing data models through the flexibility of FRM.

We examined the performance of the proposed approach through both simulated and real study data. Results from the simulation study show that the proposed approach performed really well, with good parameter and type I estimates even for a sample as small as 50 per group. The proposed approach applies to both continuous and discrete outcomes. As demonstrated by the real study on sexual health, it handled ties well in the outcome of unprotected vaginal sex, an intrinsically count response. In addition to the MWW test, median regression may also be used to address outliers arising from the Sexual Health Study [22]. However, these methods may not work well, since they either do not address missing data in longitudinal outcomes or require a unique median. Given the discrete nature of the unprotected vaginal sex outcome in our context, ties are inevitable. Although still providing valid inference under MCAR, methods that do not address missing data may suffer power loss because of list-wise deletion.

Acknowledgements

This research was supported in part by grant R33 DA027521 from the National Institutes of Health, and by the University of Rochester CTSA award UL1TR000042 from the National Center for Advancing Translational Sciences of the National Institutes of Health.

References

[1] DD. Boos, On generalized score tests. The American Statistician. 46 (1992), pp. 327–333.
[2] J.H.M. Greenberg, R. MacGowan, D. Celentano, V. Gonzales, N. Van Devanter, and J. Lifshay, Modeling intervention efficacy for high-risk women. The WINGS Project, Eval. Health Prof. 23(2) (2000), pp. 123–148.
Appendix

We provide a sketch of the proof for Theorem 2 from which Theorem 1 follows as a special case.

Let \( \hat{\eta} \) be the solution to Equation (31). Also, let \( W_n \) denote the normalized \((1/n)W_p\). By a stochastic Taylor series expansion of the estimating equations in Equation (25) and solving for \( \hat{\eta} - \eta \), we obtain

\[
\sqrt{n}(\hat{\eta} - \eta) = -\left( \frac{\partial}{\partial \eta} W_n \right)^{-T} \frac{1}{\sqrt{n}} \sum_{l=1}^{K} \sum_{i=1}^{n_l} W_{n,l,i} (\eta_l) + o_p(1),
\]

where \( o_p(\cdot) \) denotes the stochastic \( o(\cdot) \) [8, Chapter 1]. We have

\[
\frac{\partial}{\partial \eta} W_n = \frac{1}{n} \sum_{l=1}^{K} \sum_{i=1}^{n_l} \frac{\partial}{\partial \eta} W_{n,l,i}(\eta_l) = \frac{1}{n} \sum_{l=1}^{K} n_l \left( \frac{1}{n_l} \sum_{i=1}^{n_l} \frac{\partial}{\partial \eta} W_{n,l,i}(\eta_l) \right)
\]

\[
\rightarrow_p \sum_{l=1}^{K} \rho_l^2 E \left( \frac{\partial}{\partial \eta} W_{n,l,i}(\eta_l) \right) = H.
\]
Thus, it follows that
\[
\sqrt{n}(\hat{\eta} - \eta) = -H^{-1} \frac{\sqrt{n}}{n} \sum_{i=1}^{K} W_{n,i}(\eta_i) + o_p(1).
\]

By applying a Taylor expansion of multivariate U-statistic [8], we have
\[
\sqrt{n}U_n = - \left( \frac{\partial}{\partial \theta} U_n \right)^\top \sqrt{n}(\hat{\theta} - \theta) - \left( \frac{\partial}{\partial \eta} U_n \right)^\top \sqrt{n}(\hat{\eta} - \eta) - \left( \frac{\partial}{\partial \phi} U_n \right)^\top \sqrt{n}(\hat{\phi} - \phi) + o_p(1).
\]

Since
\[
\frac{\partial}{\partial \theta} U_n = \frac{1}{n_1n_k} \sum_{i=1}^{n_1n_k} \frac{\partial}{\partial \theta} U_{ln,i}(\theta) = \frac{1}{n_1n_k} \sum_{i=1}^{n_1n_k} \frac{\partial}{\partial \theta} (D_{ki}V_{ki}^{-1}A_{ki}S_{ki}),
\]

it follows from the properties of U-statistics that
\[
\frac{1}{n_1n_k} \sum_{i=1}^{n_1n_k} \frac{\partial}{\partial \theta} U_{ln,i}(\theta) \rightarrow_p E \left( \frac{\partial}{\partial \theta} U_{ln,i}(\theta) \right) = E \left( \frac{\partial}{\partial \theta} (D_{ki}V_{ki}^{-1}A_{ki}S_{ki}) \right)
\]
\[
= E \left( \left( \frac{\partial}{\partial \theta} \right) \Delta_{ki}V_{ki}^{-1}D_{ki}^\top \right) = -E(D_{ki}A_{ki}V_{ki}^{-1}D_{ki}^\top).
\]

Since a linear combination of consistent estimates is again consistent, it follows that
\[
\frac{\partial}{\partial \theta} U_n \rightarrow_p \sum_{k=2}^{K} -E(D_{ki}A_{ki}V_{ki}^{-1}D_{ki}^\top) = -B.
\]

Similarly
\[
\frac{\partial}{\partial \eta} U_n \rightarrow_p E \left( \frac{\partial}{\partial \eta} U_{ln,i} \right) = \sum_{k=2}^{K} E \left( \frac{\partial}{\partial \eta} \left( A_{ki}S_{ki} \right) \right) V_{ki}^{-1}D_{ki}^\top = C.
\]

Also, since \( \hat{\phi} \) is \( \sqrt{n} \)-consistent, it can be shown that
\[
\left( \frac{\partial}{\partial \phi} U_n \right)^\top \sqrt{n}(\hat{\phi} - \phi) = o_p(1).
\]

Now consider the U-statistics-like quantity \( \sqrt{n}U_n = \sqrt{n} \sum_{k=2}^{K} (1/n_1n_k) \sum_{i=1}^{n_1n_k} U_{ln,i}(\theta) \). It follows from the properties of the projection of U-statistics-like quantities that
\[
\sqrt{n} \sum_{i=1}^{n_1n_k} U_{ln,i}(\theta) = \sqrt{n} \left( \frac{1}{n_1} \sum_{i=1}^{n_1} E(U_{ln,i} \mid y_{li}, x_{li}) + \frac{1}{n_k} \sum_{i=1}^{n_k} E(U_{ln,i} \mid y_{ki}, x_{ki}) \right) + o_p(1)
\]
\[
= \sqrt{n} \left( \frac{1}{n_1} \sum_{i=1}^{n_1} E(U_{ln,i} \mid y_{li}, x_{li}) + \sqrt{n} \frac{1}{n_k} \sum_{i=1}^{n_k} E(U_{ln,i} \mid y_{ki}, x_{ki}) \right) + o_p(1).
\]

Thus, it follows that
\[
\sqrt{n}U_n = \sum_{k=2}^{K} \sqrt{n} \sum_{i=1}^{n_1n_k} U_{ln,i}(\theta)
\]
\[
= \frac{\sqrt{n}}{n_1} \sum_{k=2}^{K} E(U_{ln,i} \mid y_{li}, x_{li}) + \frac{\sqrt{n}}{n_k} \sum_{i=1}^{n_k} E(U_{ln,i} \mid y_{ki}, x_{ki}) + o_p(1)
\]
\[
= \sum_{i=1}^{K} \frac{\sqrt{n}}{n_1} \sum_{i=1}^{n_1} v_{li} + o_p(1).
\]
Thus, we have

\[
\sqrt{n}(\hat{\theta} - \theta) = B^{-1}\left[\sqrt{n}U_n + C\sqrt{n}(\hat{\eta} - \eta)\right] + o_p(1)
\]

\[
= B^{-1}\left[\sum_{l=1}^{K} \frac{\sqrt{n}}{\sqrt{n}} \sum_{i=1}^{n_l} v_{li} - CH^{-1} \sum_{l=1}^{K} \frac{\sqrt{n}}{n_l} \sum_{i=1}^{n_l} W_{n,li}(\eta_l)\right] + o_p(1)
\]

\[
= B^{-1}\left[\sum_{l=1}^{K} \frac{\sqrt{n}}{\sqrt{n_l}} \sum_{i=1}^{n_l} (v_{li} - CH^{-1}W_{n,li}(\eta_l))\right] + o_p(1).
\]

(A1)

By applying the central limit theorem to the above, we obtain the asymptotic normality of \( \hat{\theta} \) with the asymptotic variance given in Equation (32).