Functional Response Models for Addressing Outliers and Beyond

Chen, T.* Chen, R.† Wu, P.‡ Lu, N.§ He, H.¶ Zhang, H.‖
Kowalski, J.** Tu, X.M.††

Abstract

Functional response models (FRM) are widely used to address limitations of classic statistical models as well as timely issues arising in research and practice. In this paper, we discusses two most recent applications of this class of models: one in robust regression analysis against outliers and the other on modeling human interaction.

Popular semi-parametric regression models such as the generalized estimating equations (GEE) improve robustness of inference over parametric models. However, such models are not robust against outlying observations. The rank regression (RR) provides more robust estimates over the GEE. Unfortunately, the RR and its recent extensions to longitudinal data do not sufficiently address missing data. We discuss a new FRM-based approach to address outliers in longitudinal studies with missing data following the missing at random (MAR) mechanism. The second application focuses on modeling between-subject attributes, a new data type arising in modeling human interaction such as friendship between individuals. Such attributes are both conceptually and analytically differerent from whtin-subject attributes in conventional statistical models. The FRM is uniquely positioned to model this new data type.

Key Words: between-subject attribute, longitudinal data, missing at random, rank regression, semi-parametric models

1. Introduction

1.1 Limitations of Regression Models for Addressing Outliers

Regression models are widely used to model a response, or dependent variable, of interest, as a function of a set of independent variables, in almost all areas of research. The independent variables are also referred to as *explanatory* variables, *predictors*, or *covariates*, etc., depending on their roles in the statistical model. Classic regression models assume mathematical distributions for the conditional mean of the response given the independent variables, such as the normal distribution in linear regression. A major issue with the classic model is that estimates of model parameters may be biased when fitting such models to data that do not follow the assumed parametric distributions. Because of limited mathematical distributions, parametric regression models often fail to fit the data in real studies, raising concerns about conclusions drawn from such analysis.

An approach widely used to address the limitations of classic regression models is the generalized estimating equations (GEE) approach. This popular alterative only assumes the form of the conditional mean, i.e., the dependence structure of the mean of the response

*Department of Mathematics and Statistics, University of Toledo, Toledo, OH 43606
†Amazon.com, Inc., Seattle, WA 98109
‡Value Institute, Christiana Care Health System, Newark, DE 19718
§School of Medicine and Health Management, Huazhong University of Science and Technology, Wuhan, China
¶Department of Epidemiology, School of Public Health & Tropical Medicine, Tulane University, New Orleans, LA 70112
‖Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38105
**Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA. 30322
††Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642

given the set of explanatory variables, and yield valid inference, if the posited relationship between the response and explanatory variables as specified by the conditional mean is correct. Thus, the GEE provides valid inference for a broad class of data distributions.

Although significantly improving the utility of regression, the GEE still has limitations in practice. A major limitation is sensitiveness of its estimates to outlying observations. Like the sample mean, GEE estimates can become quite biased when there are extremely large or small values in the response. Various approaches have been developed to address this important issue. One is simply remove such outliers such as the popular trimmed mean. Another is replace outliers with reasonable values such as the winsorized mean. Both are subject to individual interpretations and are less objective. A more objective alternative is downweight outliers and a popular approach based on this principle is the rank regression (Naranjo et al., 1994, Jung 1996). Unfortunately, the rank regression as well as its extensions to longitudinal data do not address missing data, especially when missing data follows the missing at random (MAR) mechanism.

1.2 New Paradigm for Modeling Between-subject Attributes

The dominant regression paradigm as we know it is also limited to model within-subject attributes such as individuals' demographic information, medical illnesses, and social and health behaviors. However, the internet and recent expositions of online social media have not only created much more data, but also a new data dimension of human interaction to allow one to model its effect on individual outcomes. With data science methods, recent studies have indicated that human interaction is a key predictor of most human behaviors and social phenomena such as flu pandemic, financial crashes and political upsets (Pentland et al., 2013). Indeed, human interaction is such a strong predictor that it fundamentally changes the way we design behavioral and social intervention research studies.

For example, in a study using mobile and online social media, Pentland et al. (2013) showed that simply changing the schedules of coffee breaks from one person at a time to multiple employees simultaneously resulted in a productivity increase of $15 million a year for a Bank of America call center. Another study about helping save energy found that it is more effective to change behaviors of others connected to the person of interest than to try to change this person in the group who is consuming more energy (Pentland et al., 2013). The researchers provided small cash incentives to individuals who had the most interaction with specific high energy use consumers, rewarding them for improved behavior of offending consumers. Similar studies replicate this finding that a social influence approach is up to four times as efficient as traditional methods (Pentland et al., 2013).

However, under the current data and analytic paradigm, outcomes, or variables, are defined as measures of attributes from each individual, such as age, gender, income and hospitalization. Research studies, regardless of observational or randomized control studies, all focus on modeling relationships among such within-subject attributes, completely ignoring influences from interactions with others. Outcomes measuring human interaction are of between-subject attribute. Such variables are conceptually different from conventional variables, since they are even defined for an individual. For example, if f_{ij} is an indicator of connection between two individuals such as friendship ($f_{ij} = 1$ if connected and 0 otherwise), then f_{ij} clearly requires two individuals. As we detail below, this new data type of between-subject attribute has significant implications for statistical analysis and in particular excludes applications of conventional statistical models.

2. Functional Response Models as A Unified Paradigm

We start with a brief review of the functional response models (FRM). We then illustrate applications of the FRM to address the limitations of existing regression models for outliers and the between-subject attributes as in modeling human interaction.

2.1 Functional Response Models (FRM)

The functional response models (FRM) have the general form:

$$E\left(f_{i_1,\ldots,i_q} \mid \mathbf{x}_{i_1},\ldots,\mathbf{x}_{i_q}\right) = h\left(\mathbf{x}_{i_1},\ldots,\mathbf{x}_{i_q};\theta\right), \quad (i_1,\ldots,i_q) \in C_q^n, \tag{1}$$

where $f(\cdot)$ is some function, $h(\cdot)$ is some smooth function (e.g., continuous second-order derivatives), $\mathbf{x}_i$ denote some explanatory variables, 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 θ a vector of parameters. The response $f_{i_1,\ldots,i_q}$ in (1) can be quite a complex function of either between- or within-subject attributes. For example, if $q=1$, $f_i = y_i$ and $h(\mathbf{x}_i;\theta) = h\left(\mathbf{x}_i^\top \beta\right)$, the FRM yields the generalized linear model. For $q=2$, we may define f_{ij} as a connection indicator discussed above. The FRM has been applied to a range of applications such as extensions of the Mann-Whitney-Wilcoxon rank sum test to longitudinal and causal inference settings (Chen et al., 2014; Wu et al., 2014), reliability coefficients (Lu et al., 2013), models for population mixtures (Yu et al., 2013), and causal inference for multi-layered intervention studies (Wu et al., 2014).

2.2 Limitations of Existing Regression Models for Addressing Outliers

We start with the classic linear regression for continuous responses and develop an FRM to provide robust inference against outliers. We then extend the FRM to longitudinal data with missing values following the MAR.

Let y_i $(\mathbf{x}_i)$ denote a continuous response (a p-dimensional vector of explanatory variables). The classic linear regression model is given by:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim \text{i.i.d. } N\left(0,\sigma^2\right), \quad 1 \le i \le n, \tag{2}$$

where i.i.d. denotes "independently and identically distributed", $N\left(\mu,\sigma^2\right)$ a normal with mean μ and variance σ^2 and $\boldsymbol{\beta}$ is a p-dimensional vector of parameters. Although the least square (LS), or maximum likelihood (ML), estimate $\widehat{\boldsymbol{\beta}}^{(LS)}$ of $\boldsymbol{\beta}$ is widely used for inference about $\boldsymbol{\beta}$, this popular estimate is quite sensitive to outliers in y_i.

One way to reduce the effect of outliers on the estimate $\widehat{\boldsymbol{\beta}}^{(LS)}$ is to trim the outliers. Another is to replace these outlying observations with more reasonable values, such as 3 times the interquartile range (Schroeder et al., 2003). Both methods are quite subjective and generally yield quite different estimates depending how the outliers are trimmed or winsorized. A third alternative is to downweigh the contributions of the outliers. The rank regression (RR) is based on this principle. This approach weights the residuals by using the so-called Wilcoxon score:

$$n^{-2}\sum_{i<j}|e_i - e_j| = 2n^{-2}\sum_{i=1}^{n}\left|R(e_i) - \frac{n+1}{2}\right| e_i, \tag{3}$$

where $e_i = y_i - \mathbf{x}_i^\top \boldsymbol{\beta}$ and $R(e_i)$ denotes the rank of e_i.

The first two methods, trimming and winsorizing outliers, are readily applied when modeling longitudinal data using methods such as the generalized linear mixed-effects

model and the weighted generalized estimating equations (Tang et al., 2012). The RR has also been extended to longitudinal data (Naranjo et al. 1994; Sievers 1983; Terpstra et al. 2000; chang et al. 1999;terpstra et al., 2000.) However, none of these extensions addresses MAR. Since missing data is the norm rather the exception in longitudinal studies and the MAR is a more realistic assumption for missing data in most clinical studies, this limitation severely hampers the utility of the RR in addressing outliers in practice.

2.3 A New Regression Model for Addressing Outliers under MAR

Under (2), we have:

$$E\left(I_{\{y_i-y_j\leq 0\}} \mid \mathbf{x}_i, \mathbf{x}_j\right) = \Pr\left[\epsilon_i - \epsilon_j \leq -\boldsymbol{\beta}^\top (\mathbf{x}_i - \mathbf{x}_j)\right] = \Phi\left(-\boldsymbol{\theta}^\top (\mathbf{x}_i - \mathbf{x}_j)\right), \quad (4)$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal and $\boldsymbol{\theta} = \frac{1}{\sqrt{2}\sigma}\boldsymbol{\beta}$. Unlike the original linear model, the response $I_{\{y_i-y_j\leq 0\}}$ only depends on the rank of y_i and thus remains the same regardless of actual values of y_i. This rank-preserving feature is also unique to (4), since the Wilcoxon score, $|e_i - e_j|$, in the RR still depends on actual values of y_i and y_j. Another major distinction is that (4) is a stand-alone model for both continuous and non-continuous y_i (e.g., count outcomes), while the RR is not, since the latter is an inference technique to reduce the effect of outliers on the estimate of $\boldsymbol{\beta}$ in the linear regression model in (2).

Inference about $\boldsymbol{\theta}$ cannot be carried out using popular methods for semi-parametric models such as the GEE, since (4) is not a generalized linear model (GLM) or even a non-linear model. However, by setting $q = 2$, $f_{ij} = I_{\{y_i-y_j\leq 0\}}$ and $h(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}) = \Phi\left(-\boldsymbol{\theta}^\top (\mathbf{x}_i - \mathbf{x}_j)\right)$, it is immediately seen that (4) is an FRM. By taking advantage of the inference theories for FRM (Kowalski and Tu, 2007), we have developed methods to obtain consistent estimates of $\boldsymbol{\theta}$ and associated asymptotic normal distributions (Chen et al., 2015), not only for (4), but also for its extension to longitudinal data with missing values following the MAR mechanism by utilizing the U-statistics-based Weighted Generalized Estimating Equations (UWGEE, Kowalski and Tu, 2007).

For longitudinal data, consider a study with m assessments and let y_{it} $(\mathbf{x}_{it})$ denote a response (a vector of explanatory variables) from the ith subject at time t ($1 \leq i \leq n$, $1 \leq t \leq m$). Then, the version of the FRM for longitudinal data becomes:

$$E\left[f(y_{it}, y_{jt}) \mid \mathbf{x}_{it}, \mathbf{x}_{jt}\right] = h(\mathbf{x}_{it}, \mathbf{x}_{jt}; \boldsymbol{\theta}), \quad f(y_{it}, y_{jt}) = I_{\{y_{it}-y_{jt}\leq 0\}}, \quad (5)$$
$$(i,j) \in C_2^n, \quad 1 \leq t \leq m.$$

Like the FRM for cross-sectional data, $\boldsymbol{\theta}$ is the vector of parameters of interest. As in popular models for longitudinal data such as the generalized linear mixed-effects models (GLMM) and GEE (WGEE), we can include time in $\mathbf{x}_{it}$ to model temporal patterns and relationships concerning changes of y_{it} over time.

For addressing missing data under MAR. we assume the Monotone Missing Data Patterns (MMDP), model the missingness and integrate this missing data model with the FRM in (5) to provide joint inference about parameters for both models.

Let

$$r_{it} = \begin{cases} 1 & \text{if } y_{it} \text{ is observed} \\ 0 & \text{if } y_{it} \text{ is unobserved} \end{cases}, \quad \mathbf{r}_i = (r_{i1}, r_{i2}, ..., r_{im})^\top, \quad (6)$$
$$\mathbf{x}_i = (\mathbf{x}_{i1}^\top, ..., \mathbf{x}_{im}^\top)^\top, \quad \mathbf{y}_i = (y_{i1}, ..., y_{im})^\top, \quad \pi_{it} = E(r_{it} \mid \mathbf{y}_i, \mathbf{x}_i),$$
$$1 \leq i \leq n, \quad 1 \leq t \leq m.$$

Also, let

$$\begin{aligned} H_{it} &= \{\mathbf{x}_{is}, \mathbf{y}_{is}; 1 \leq s \leq t-1\}, \quad \mathbf{x}_{it^-} = (\mathbf{x}_{i1}^\top, ..., \mathbf{x}_{i(t-1)}^\top)^\top, \\ \mathbf{y}_{it^-} &= (y_{i1}, ..., y_{i(t-1)})^\top, \end{aligned}$$

where $\mathbf{x}_{it^-}$ and $\mathbf{y}_{it^-}$ contain the explanatory and response variables prior to time t, respectively. Under MMDP, we model $p_{it} = \Pr\left(r_{it} = 1 \mid r_{i(t-1)} = 1, H_{it}\right)$ using logistic regression:

$$\text{logit}\left(p_{it}\left(\boldsymbol{\gamma}_t\right)\right) = \gamma_{0t} + \boldsymbol{\gamma}_{xt}^\top \mathbf{x}_{it^-} + \boldsymbol{\gamma}_{yt}^\top \mathbf{y}_{it^-}, \quad 2 \leq t \leq m. \tag{7}$$

These one-step transition probabilities are then linked to π_{it} under MAR by:

$$\pi_{it}\left(\boldsymbol{\gamma}\right) = \Pr\left(r_{it} = 1 \mid H_{it}\right) = \prod_{s=2}^{t} p_{is}\left(\boldsymbol{\gamma}_s\right), \quad 2 \leq t \leq m.$$

where $\boldsymbol{\gamma}_t = \left(\gamma_{0t}, \boldsymbol{\gamma}_{xt}^\top, \boldsymbol{\gamma}_{yt}^\top\right)^\top$ and $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_2^\top, \ldots, \boldsymbol{\gamma}_m^\top\right)^\top$.

As in the weighted generalized estimating equations (WGEE) for semi-parametric regression models, the logistic regression models in (7) allow one to test the null of MCAR. If $\boldsymbol{\gamma} \neq \mathbf{0}$, then missing data does not follow MCAR and ignoring missing data generally yields biased estimate estimates of $\boldsymbol{\theta}$.

Given an estimate $\widehat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$, we estimate $\boldsymbol{\theta}$ by solving the estimating equations of the form:

$$\mathbf{U}_n\left(\boldsymbol{\theta}\right) = \sum_{\mathbf{i} \in C_2^n} \mathbf{U}_{n,\mathbf{i}} = \sum_{\mathbf{i} \in C_2^n} D_{\mathbf{i}} V_{\mathbf{i}}^{-1} \boldsymbol{\Delta}_{\mathbf{i}} S_{\mathbf{i}} = \mathbf{0}, \tag{8}$$

where $D_{\mathbf{i}}$ and $S_{\mathbf{i}}$ are determined by (5), $\boldsymbol{\Delta}_{\mathbf{i}}$ is a function of $r_{\mathbf{i}t} = r_{it} r_{jt}$ and $\pi_{\mathbf{i}t} = \pi_{it}\pi_{jt}$, and $V_{\mathbf{i}}$ is a function of the variance of $f\left(y_{it}, y_{jt}\right)$ and a working correlation matrix (akin standard GEE). Although similar in appearance, (7) is not the standard WGEE, since $\mathbf{i}$ indexes pairs of subjects and $\mathbf{U}_n\left(\boldsymbol{\theta}\right)$ is not a sum of independent random quantities. This class of so-called UWGEE, because $\mathbf{U}_n\left(\boldsymbol{\theta}\right)$ is a U-statistics-like quantity, has nice asymptotic properties, just like the WGEE (Kowalski and Tu, 2007). Given $\widehat{\boldsymbol{\gamma}}$, estimates $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ obtained by solving (7) can be shown to be consistent and asymptotically normal under mild regularity conditions (Chen et al., 2015). Further, we have developed procedures to account for sampling variability of $\widehat{\boldsymbol{\gamma}}$ in the asymptotic variance of the UWGEE estimate $\widehat{\boldsymbol{\theta}}$ (Chen et al., 2015).

2.4 Models for Between-subject Attributes

We illustrate an application of the FRM for modeling human interaction by focusing on the network density, a popular metric for measuring connections in a social network.

Let $\{\mathbf{s}_i; 1 \leq i \leq n\}$ denote the nodes of a social network sample, which may represent people or organizations. A connection outcome between two notes is given by:

$$f_{ij} = f\left(\mathbf{s}_i, \mathbf{s}_j\right) = \begin{cases} 1 & \text{if } \mathbf{s}_i \text{ and } \mathbf{s}_j \text{ are connected} \\ 0 & \text{otherwise} \end{cases}. \tag{9}$$

As noted earlier, the response f_{ij}, defined by two individuals, is a between-subject attribute. In addition to this conceptual distinction, this new data type also has significant implications for statistical analysis. For example, a popular measure of connectivity of the network is the density: $\theta = E\left(f_{ij}\right)$. We can readily estimate the density by averaging all the connection indicators, $\widehat{\theta} = \binom{n}{2}^{-1} \sum_{(i,j) \in C_2^n} f_{ij}$. However, inference about $\widehat{\theta}$ is not so

obvious, since there is no independence among the f_{ij}'s and standard asymptotic results such as the central limit theorem does not apply.

Although dependent observations also arise from other settings such as clustered repeated observations from longitudinal studies, the dependence structure among the between-subject attributes f_{ij} is quite different. For example, we can create independent clusters by grouping repeated observations from the same subject and the independence among the clusters forms the basis for all modern statistical models for longitudinal data such as the GEE. However, in general, no independent cluster exists for between-subjects. For example, even if f_{12} and f_{34} involve different pairs of individuals, they may still be correlated, since f_{12} (f_{23}) and f_{23} (f_{34}) may be correlated. Without taking the dependence into account, inference based on conventional statistical models yields biased results.

Many publications in the social network literature simply ignored this dependence issue (Feinberg et al., 2005; Valente et al., 2007; Palinkas et al., 2011; Centola, 2010; 2011). Although others realized and attempted to address this issue, all failed to recognize the distinction between the between- and within-subject attributes and continued to apply conventional statistical methods. For example, we applied two popular packages for social network data analysis, UCINET version 6.385 (Borgatti, et al., 2002; Hunter et al., 2008) and STATNET version 2014.2.0 (Goodreau et al., 2008), in a simulation study to see how they would perform. We simulated a set of connection outcomes defined by $f_{ij} = I_{\{z_i + z_j \leq 0\}}$, with z_i (z_j) generated from the standard normal $N(0,1)$ with mean 0 and variance 1 for a network of size $n = 100$. The true density in this special case is $\theta = E(f_{ij}) = 0.5$. The standard error of the estimate $\widehat{\theta} = \binom{n}{2}^{-1} \sum_{(i,j) \in C_2^n} f_{ij}$, which is readily evaluated using the theory of U-statistics, is 0.059.

Estimate $\widehat{\theta}$	Standard error				
	True	Bernoulli	STATNET	UCINET	FRM
0.51	0.059	0.0025	0.003	0.0408	0.060

Table 1: Comparison of network statistics (degree, density) between two PHQ-9 groups and estimates of parameters of FRM (estimate, standard error, p-value) for the TrevorSpace Project.

Shown in Table 1 are the estimate of θ and standard errors from the two packages as well as the naive method by treating f_{ij} as independent Binomial observations. The UCINET employed Bootstrap to estimate standard errors, while the STATNET uses the exponential random graph model with inference based on the Markov Chain Monte Carlo Maximum Likelihood estimation (Wasserman and Pattison, 1996; Snijders, 2002). All these methods yielded the same density estimate of θ, but standard errors varied tremendously across the different methods. The standard error from the FRM was nearly identical to the true value. Both the naive method and STATNET had a huge downward bias. Although the standard error from the UCINET, based on 5,000 Bootstrap samples, was much improved, the Bootstrap still failed to correct the downward bias in the standard error. Thus, inter-dependence among between-subject attributes such as the connection indicator f_{ij} in this example has a significant impact on the variability of estimates and cannot be corrected using conventional statistical models for within-subject attributes.

The FRM-based model in (9) is readily extended to more complex situations involving comparing multiple densities for different social networks and investigate effects of individual explanatory variables on densities. If $\mathbf{x}_i$ denotes a vector of explanatory variables

from the ith subject, we can study its impact on the density by the following FRM:

$$E\left(f_{ij} \mid \mathbf{x}_i, \mathbf{x}_j\right) = h\left(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}\right), \quad (i,j) \in C_2^n.$$

The model in (9) is also readily extended to network densities defined by more complicated relationships. Many people believe that connection between two individuals may not indicate a true relationship and like to look at relationships involving more individuals such as three people. The response function in (9) is readily changed to accommodate such needs. For example, by using the response:

$$f_{ijk} = f\left(\mathbf{s}_i, \mathbf{s}_j, \mathbf{s}_k\right) = \begin{cases} 1 & \text{if } \mathbf{s}_i, \mathbf{s}_j \text{ and } \mathbf{s}_k \text{ are connected} \\ 0 & \text{otherwise} \end{cases}, \tag{10}$$

we can model network densities defined by triangle relationships.

3. Performance Evaluation

We evaluate the performance of the FRM models for regression analysis with outliers and social network density with simulated data.

3.1 Simulation Study

We use simulations to assess the performance of the FRM. All simulations are carried out with a Monte Carlo (MC) sample of $M = 1,000$ and a two-sided statistical significance level $= 0.05$.

3.1.1 Models for Robust Regression Against Outliers

We consider a pre-post study with two assessment times and generate data from the longitudinal model:

$$y_{it} = x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_{it}, \quad 1 \le t \le 2, \quad 1 \le i \le n,$$
$$x_{i1} \sim N\left(0, 0.2\right), \quad x_{i2} \sim \text{Bern}\left(0.5\right), \quad \boldsymbol{\varepsilon}_i = \left(\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}\right)^\top \sim N\left(\mathbf{0}, \frac{1}{2}\mathbf{I}_3\right),$$

where $\mathbf{I}_k$ denotes the $k \times k$ identity matrix, Bern(p) a Bernoulli with mean p and $N\left(\mu, \Sigma\right)$ a multivariate normal with mean μ and variance Σ. We set $\sigma_\epsilon = 1/\sqrt{2}$ such that $\theta_k = \beta_k$ $(k = 1, 2)$. We then create outliers in the sample generated by replacing the (10%) largest y_i's by those simulated from a uniform $U\left(100000, 1000000\right)$.

To simulate missing data, we assume no missing data at baseline $t = 1$ and generate missing responses under MAR with about 23% missing data at $t = 2$ from the following logistic regression:

$$\text{logit}\left(\pi_{i2}\left(\boldsymbol{\gamma}\right)\right) = \text{logit}\left(p_{i2}\left(\boldsymbol{\gamma}\right)\right) = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 \frac{R_{1i}}{n}, \quad \eta_0 = 1, \quad \eta_1 = 1.5, \quad \eta_2 = 0.5,$$

where R_{1i} denotes the rank of y_{1i}. We use the rank normalized with respect to the sample size, $\frac{R_{1i}}{n}$, rather than y_{1i} itself, in the logistic model because of extremely large outliers for some of y_{1i}.

To save space, we only report results in the case of missing data under MAR. Shown in Table 2 are the estimates of $\boldsymbol{\beta}$their corresponding standard errors as well as coverage probabilities for the sample sizes considered. As expected, the point estimates became more accurate as the sample size increased. The asymptotic standard errors were generally close to their empirical counterparts, especially for $n = 300$. Likewise, the coverage probability also improved as the sample size increased.

Estimates of $\boldsymbol{\beta}$, standard errors and coverage probabilities						
			Standard Errors		Coverage Probability	
Sample Size	Parameter	Mean	Asymptotic	Empirical	Asymptotic	Empirical
$n = 150$	β_1	0.996	0.277	0.279	0.953	0.946
	β_2	1.006	0.116	0.114	0.957	0.949
$n = 300$	β_1	1.008	0.194	0.196	0.945	0.947
	β_2	1.004	0.082	0.082	0.943	0.944

Table 2: UWGEE estimates of parameters, standard errors and coverage probabilities for simulated longitudinal data with missing values.

3.1.2 Models for Network Density

We consider two social networks. For notational brevity, let $1 \leq i \leq n_1$ denote the first and $n_1 + 1 \leq i \leq n_1 + n_2$ the second social network. We simulate the connection out $f_{\mathbf{i}}$ from a Bernoulli model as follows:

$$f_{\mathbf{i}} \mid \mathbf{x}_{\mathbf{i}} \overset{i.d.}{\sim} \text{Bernoulli}\left(h_{\mathbf{i}}\right), \quad h_{\mathbf{i}} = h\left(\mathbf{x}\left(\mathbf{s}_i, \mathbf{s}_j\right); \beta\right), \quad \mathbf{i} = (i, j) \in C_2^n,$$

where

$$h\left(\mathbf{x}\left(\mathbf{s}_i, \mathbf{s}_j\right); \beta\right) = \begin{cases} \frac{\exp(\beta_0)}{1+\exp(\beta_0)} & \text{if } i, j \text{ from 1st group} \\ \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)} & \text{if } i, j \text{ from 2nd group} \\ \frac{\exp(\beta_0+\beta_2)}{1+\exp(\beta_0+\beta_2)} & \text{if } i \text{ from 1st and } j \text{ from 2nd group} \end{cases}.$$

Thus, we not only consider interactions between subjects within each social network, but between the social networks as well. In the above, $\frac{\exp(\beta_0)}{1+\exp(\beta_0)}$ $\left(\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}\right)$ is the network density for the first (second) social network, while $\frac{\exp(\beta_0+\beta_2)}{1+\exp(\beta_0+\beta_2)}$ represents the mean connection between the two social networks. We set $n_1 = n_2 = 50$ so we can assess the performance of the FRM for relative small sample sizes.

Parameter	True value	Estimate	Standard error	
			Asymptotic	Empirical
β_0	1.38	1.33	0.071	0.069
β_1	0.81	0.84	0.12	0.11
β_2	-0.53	-0.49	0.08	0.09

Table 3: Estimates of parameters and standard errors (asymptotic and empirical) for the FRM for the simulation study.

Shown in Table 3 are estimates of parameters and (asymptotic) standard errors, along with the true values of the parameters. Under the true values of β, the density is 0.8 for the first network and 0.9 for the second, while the mean between-network connection is 0.70. The estimated parameters were quite close to the respective true values. As well, the asymptotic standard errors were also nearly identical to their empirical counterparts. Even for this small sample size, the approach worked remarkably well.

4. Discussion

In this paper, we discussed two new applications of the FRM to address outstanding statistical issues. The first is a classic problem, while the second is a timely issue. Both

problems have generated significant interest in the literature. For example, Thas et al. (2012) also independently developed the FRM for addressing outliers and termed it the probabilistic index model. However, they only considered cross-sectional data. There is also a burgeoning literature on methods for social network data analysis. In addition to network features such as the network density considered in this paper, attempts have also been made to model the impact of human interaction on behavioral and health outcomes. For example, Christakis and Fowler conducted a number of social network analyses with the Framingham Heart Study (Christakis and Fowler, 2007; 2008). This large longitudinal cohort study collects not only physical health information (e.g., BMI, happiness, depression, number of drinks), but also social networks of participates (co-workers, friends, siblings, parents and children). They attempted to estimate contagion effects using longitudinal regression models and suggested that the social influence plays a role in the spread of obesity, smoking, happiness and loneliness.

While Christakis and Fowler's work have received considerable acclaim in the popular press and in society, their analyses have come under critique. In their work, the GEE is applied to the repeated assessments of each individual, by completely ignoring interpersonal interactions among individuals within each social network. Such interdependence may have a significant impact on individual behavioral and health outcomes and ignoring this issue casts doubt on their findings. Indeed, using the same methodology, Cohen-cole and Fletcher (2008) found that other attributes, such as height and headaches, also traveled in networks, which seems quite implausible. Thus to fully account for human interaction and its impact on individual outcomes, a paradigm shift from the within- to between-subject attribute is necessary to address the underlying statistical issues.

REFERENCES

Borgatti, S.P., Everett, M.G. and Freeman, L.C. (2002). *Ucinet for Windows: Software for social network analysis*. Analytic Technologies, Harvard, MA.

Burnham, K. P., and Anderson, D. R. (1998), *Model Selection and Inference*, New York: Springer.

Centola, D. (2010), "The spread of behavior in an online social network experiment", *Science*, 329, 1194-1197.

Centola, D. (2011), "An experimental study of homophily in the adoption of health behavior", *Science*, 334, 1269-1272.

Chang, W. H., McKean, J. W., Naranjo, J. D., and Sheather, S. J. (1999), "High-breakdown rank regression",*Journal of the American Statistical Association*, 94(445), 205-219.

Chen, R., Chen, T., Lu, N., Zhang, H., Wu, P., Feng, C. and Tu, X. (2014), "Extending the Mann-Whitney-Wilcoxon rank sum test to longitudinal regression analysis", *Applied Statistics* 41(12): 2658–2675.

Chen, T., Chen, R., Wu, P., Kowalski, J. and Tu, X.M., "Rank-preserving regression for longitudinal data with missing responses",*Submitted for publication*.

Christakis, N.A. and Fowler, J.H. (2007). "The spread of obesity in a large social network over 32 years", *New England Journal of Medicine*, 357, 370-379.

Christakis, N.A. and Fowler, J.H. (2008). The collected dynamics of smoking in a large social network, *New England Journal of Medicine*, vol.358, 2249-2258.

Cohen-Cole, E. and Fletcher, J.M. (2008), "Detecting implausible social network effects in acne, height, and headaches: longitudinal analysis". *BMJ* 337: a2533.

Feinberg, M.E., Riggs, N.R., Greenberg, M.T., (2005), "Social networks and community prevention coalitions'",*J Prim Prev.* 26(4), 279-298.

Goodreau, S. M., Handcock, M. S., Hunter, D. R., Butts, C. T., Morris, M.(2008), "A STATNET Tutorial", *J. Stat. Softw.*, 24(9), 1-27.

Hunter, D.R., Handcock, M.S., Butts, C.T., Goodreau, S.M. and Morris, M. (2008), "ERGM: A package to fit, simulate and diagnose exponential-family models for networks", *J. Stat. Softw*. 24(3): 1-29.

Jung, S.-H. (1996). "Quasi-likelihood for median regression models", *Journal of the American Statistical Association*, 91(433): 251–257.

Kowalski, J. and Tu, X.M. (2007), Modern Applied U Statistics, Wiley, New York.

Lu, N., White, A., Wu, P., He, H., Hu, J., Feng, C. and Tu, X. (2013), "Social network endogeneity and its implications for statistical and causal inferences. In Social Networking: Recent Trends, Emerging Issues and Future Outlook, edited by Lu, N., White, A.M. and Tu, X.M.", *Nova Science*

Naranjo, J. and Hettmansperger, T. (1994), "Bounded influence rank regression", *Journal of the Royal Statistical Society*. Series B (Methodological) pp. 209–220.

Palinkas, L.A., Holloway, I.W., Rice, E., Fuentes, D., Wu, Q., Chamberlain,P., (2011), "Social networks and implementation of evidence based practices in public youth-serving systems: a mixed-methods study", *Implement Science* 6(113), 1-11.

Pentland, A.S. (2013), "Data driven society", *Scientific American October*, 79-83.

Schroeder, E. B., Liao, D., Chambless, L. E., Prineas, R. J., Evans, G. W., and Heiss, G. (2003), "Hypertension, blood pressure, and heart rate variability the Atherosclerosis Risk In Communities (ARIC) study", *Hypertension*, 42(6), 1106-1111.

Snijders, T.A.B., Pattison, P.E., Robins, G.L. and Handcock, M.S. (2006), "New specifications for exponential random graph models"", *Sociol. Methodol*., 36(1):99–153.

Sievers, Gerald L. (1983), "A weighted dispersion function for estimation in linear models", *Communications in Statistics-Theory and Methods*, 12, no. 10: 1161-1179.

Tang, W., He, H. and Tu, X.M. (2012), Applied Categorical and Count Data Analysis. Chapman & Hall/CRC, FL, 2012.

Terpstra, J. T., McKean, J. W., and Naranjo, J. D. (2000), "Highly efficient weighted for autoregression wilcoxon estimes for autoregression", *Statistics: A Journal of Theoretical and Applied Statistics*, 35(1), 45-80.

Thas, O., De Neve, J., Clement, L. and Ottoy, J.P. (2012), "Probabilistic index models," *J.R.Statist.Soc. B.*, 74, 623-671.

Yu, Q., Chen, R., Tang, W., He, H., Gallop, R., CritsChristoph, P., and Tu, X. M. (2013), "Distributionfree models for longitudinal count responses with overdispersion and structural zeros",*Statistics in medicine*, 32(14), 2390-2405.

Valente, T.W., Chou, C.P., Pentz, M.A., (2007)," Community coalitions as a system: Effects of network change on adoption of evidence-based substance abuse prevention",*Am J Public Health*, 97(5), 880-886.

Wasserman, S., Pattison, P. (1996), "Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p", *Psychometrika*, 61, 401–425.

Wu, P., Han, Y., Chen, T., and Tu, X. M. (2014), "Causal inference for MannWhitneyWilcoxon rank sum and other nonparametric statistics",*Statistics in medicine*, 33(8), 1261-1271.

Wu, P., Gunzler, D., Lu, N., Chen, T., Wymen, P., and Tu, X. M. (2014), "Causal inference for communitybased multilayered intervention study", *Statistics in medicine*, 33(22), 3905-3918.