

# CE04C: Longitudinal Data Analysis: Semiparametric and Nonparametric Approaches

Annie Qu and Peter Song

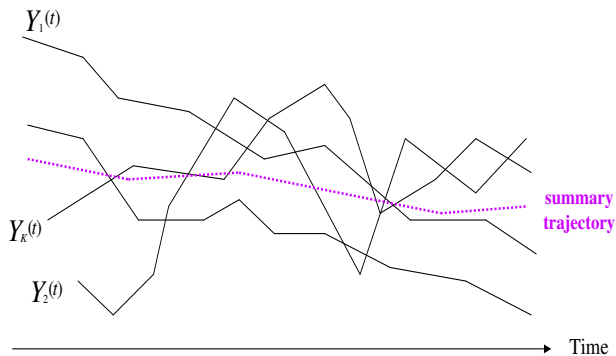
University of Illinois at Urbana-Champaign  
University of Michigan

JSM, DC  
Aug 1st, 2009

# What Is Longitudinal Data?

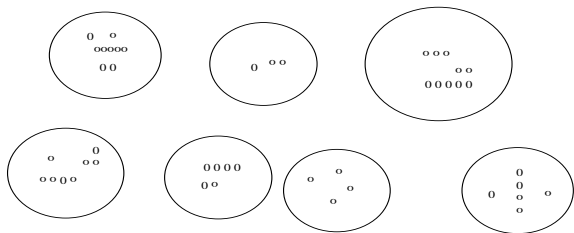
- *Longitudinal Data*: Sequentially observed over time, *longitudinal data* may be collected either from an observational study or a designed experiment, in which response variables pertain to a sequence of events or outcomes recorded at certain time points during a study period.
- Longitudinal data may be regarded as a collection of many time series, each for one subject.

# Longitudinal Data



# Clustered Data

- *Clustered data* refers to a set of measurements collected from subjects that are structured in clusters, where a group of related subjects constitutes a cluster, such as a group of genetically related members from a familial pedigree.

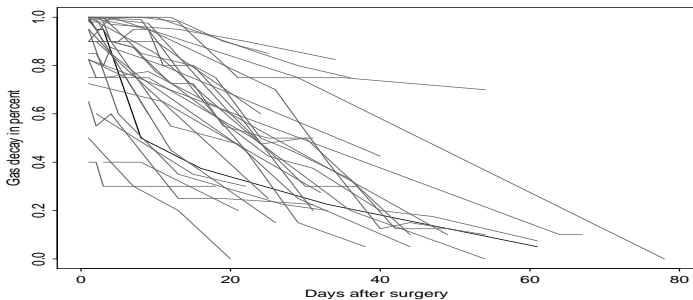




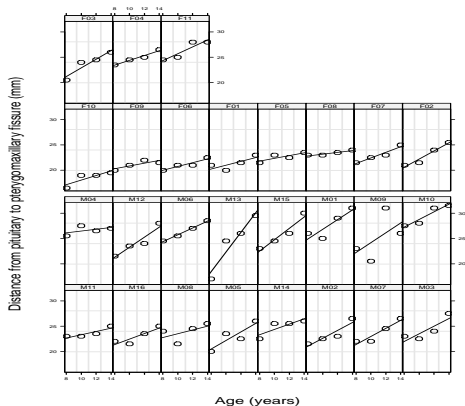
# Multilevel Data

- *Multilevel data* are collected from clusters in multi-level hierarchies, such as spatio-temporal data.
- This short course focuses on longitudinal data, and related methodology may be applied to analyze other types of correlated data such as clustered data.

# Visualizing Longitudinal Data: Spaghetti Plot



# Visualizing Longitudinal Data: Trellis Plot



Orthodontic growth patterns in 16 boys(M) and 11 girls(F) between 8 and 14 years of age. Lines represent the individual least squares fits of the simple linear regression model.



# Analysis of Longitudinal Data

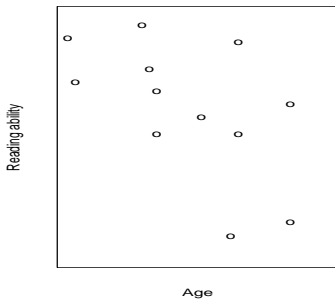
- Primary interest lies in the mechanism of change over time, including growth, time profiles or effects of covariates.
- Main advantages of a longitudinal study:
  - (1) To investigate how the variability of the response varies in time with covariates. For instance, to study time-varying drug efficacy in treating a disease, which cannot be examined by a cross-sectional study.

# Analysis of Longitudinal Data

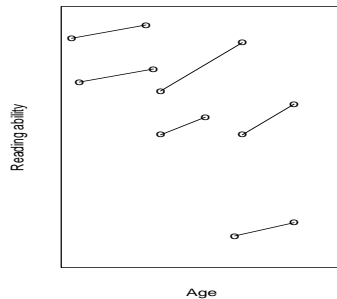
- (2) To separate the so-called *cohort* and *age* (or time) effects.  
From the figure, we learn:
- (a) Importance of monitoring individual trajectories;
  - (b) Characterize changes within each individual in the reference to his baseline status.

# Cross-sectional Analysis versus Longitudinal Analysis

(a) Cross-sectional study



(b) Longitudinal study



# Challenges in Longitudinal Data Analysis

- Complexity of the underlying probability mechanism of data generation. Likelihood inference is either unavailable or numerically too intricate to be implemented.
- Difficulty of dealing with missing data. (a) Partial information is available to hopefully “recover” the full data; (b) Constraint of preserving the same correlation structure.
- Expectation of dealing with nuisance parameters in correlation structures; when time series is long, modeling the transitional behavior (or correlation structure) become a primary task.

# Main Data Features

- The presence of repeated measurements for each subject implies that data are autocorrelated or serially correlated. Thus, statistical inference needs to take this serial correlation into account.
- The length of time series determines how much we like to learn about the correlation structure of the data.
- In many practical studies, outcomes are not normally distributed.
- Outcomes are vector-values at give a time point.
- Data contain missing values.

## Data Example 1: HIV Data

- HIV AIDS data (Huang et al., 2002; Qu and Li, 2006)
- Consists of 283 homosexual males who were HIV positive between 1984 and 1991
- Each patient had their visits after their HIV infection and had his CD4 counts repeated measured about every 6 months
- The measurements of CD4 vary from the minimum 1 to maximum 14 because of some patients missed their appointments
- HIV destroys CD4 cells, therefore it is important to monitor progression of the disease through the CD4 counts over time

## Data Example 1: HIV Data

- The response variable is the CD4 percentage over time
- Four covariates were also collected: patient's age, smoking status with 1 as smoker and 0 as nonsmoker, the CD4 cell percentage before their infection
- The linear regression model is

$$y = \beta_0 + \beta_1 \text{Smoke} + \beta_2 \text{Age} + \beta_3 \text{PreCD4} + \beta_4 \text{Time} + \varepsilon.$$

- Contains unevenly spaced measurements over time, with partial information missing for some subjects

# Data Example 1: HIV Data

ID	Time	Smoke	PreCD4	Age	CD4
1022	0.2	0	26.25	38	17
1022	0.8	0	26.25	38	30
1022	1.2	0	26.25	38	23
1022	1.6	0	26.25	38	15
1022	2.5	0	26.25	38	21
1022	3	0	26.25	38	12
1022	4.1	0	26.25	38	5
1049	0.3	0	32.375	44.5	37
1049	0.6	0	32.375	44.5	44
1049	1	0	32.375	44.5	37
1049	1.5	0	32.375	44.5	35
1049	2	0	32.375	44.5	25
1049	2.5	0	32.375	44.5	21
1049	3	0	32.375	44.5	22
1049	3.5	0	32.375	44.5	21
1049	4	0	32.375	44.5	22
1049	4.5	0	32.375	44.5	26
1049	5	0	32.375	44.5	20
1049	5.5	0	32.375	44.5	15



## Data Example 2: Binary Outcome

- Data example from Preisser & Qaqish (1999) on urinary incontinence
- The response variable is binary, indicating whether or not the subject's daily life is bothered by accidental loss of urine with 1 corresponding to bothered and 0 otherwise
- Subjects are correlated if they are from the same hospital practice
- There are 137 patients from 38 practices, and each cluster contains at least 1 patient and at most 8 patients
- There are 5 covariates, gender ('female'), age ('age'), daily leaking accidents ('dayacc'), severity of leaking ('severe') and number of times to use the toilet daily ('toilet')

## Data Example 2: Binary Outcome

- The logistic link function is assumed for the marginal model, so that

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 \text{ female} + \beta_2 \text{ age} + \beta_3 \text{ dayacc} + \beta_4 \text{ severe} + \beta_5 \text{ toilet},$$

where  $\mu_{ij}$  denotes the probability of being bothered for patient  $j$  in cluster  $i$

- Patients 8 and 44 were identified as possible outliers (Preisser & Qaqish, 1996; 1999)
- Without downweighting, GEE provides estimator which is very different from the estimator without downweighting

## Data Example 2: Urinary Incontinence Data

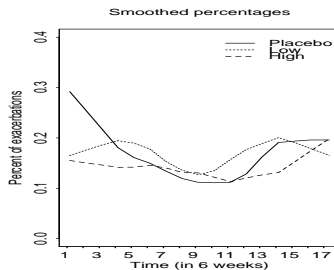
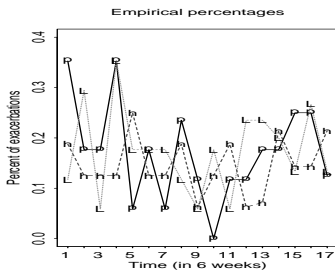
pract_id	pat_id	bothered	female	age	dayacc	severe	toilet
8	1	1	1	77	7	3	8
8	2	0	1	82	1	1	3
8	3	1	1	78	7	3	6
24	4	0	1	87	0.286	2	6
24	5	0	1	78	2	2	4
27	6	0	1	79	1	2	4
27	7	1	1	90	15	4	20
27	8	0	0	77	9.286	1	10
27	9	0	1	84	3	2	4
27	10	1	1	77	14.857	2	15
..							
..							
..							
107	44	0	1	77	3	2	20

## Data Example 3: Multiple Sclerosis Trial (MST)

- A longitudinal clinical trial to assess the effects of neutralizing antibodies on interferon beta-1 (IFNB) in relapsing-remitting multiple sclerosis (MS), a disease that destroys the myelin sheath surrounding the nerves (Petkau et al, 2004).
- Six-weekly frequent Magnetic Resonance Imaging (MRI) sub-study involving 52 patients, randomized into 3 treatment groups; 17 in placebo, 17 in low dose and 16 in high dose.
- At each of 17 scheduled visits, a binary outcome of *exacerbation* was recorded at the time of each MRI scan, according to whether an exacerbation began since the previous scan.
- Baseline covariates include age, duration of disease (in years), sex, and initial EDSS (expanded disability status scale) scores.
- Does the IFNB help to reduce the risk of exacerbation?

## Data Example 3: Multiple Sclerosis Trial (MST)

- A collection of  $N = 52$  longitudinal trajectories, which are equally spaced at 17 time points.

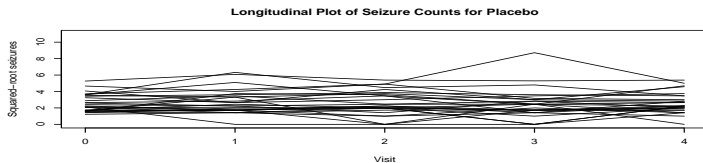
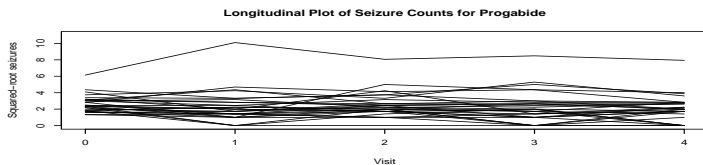


## Data Example 4: Epileptic Seizures (ES) Data

- Data were collected from a clinical trial of 59 epileptics.
- It aimed to examine the effectiveness of the drug **progabide** in treating epileptic seizures.
- For each patient, the number of epileptic seizures was recorded during a baseline period of 8 weeks.
- Patients were then randomized to two treatment arms, one with progabide, and the other with a placebo, in addition to a standard chemotherapy.
- The number of seizures was recorded in 4 consecutive two-week periods after the randomization.
- The scientific question: whether the drug progabide helped to reduce the rate of epileptic seizures.

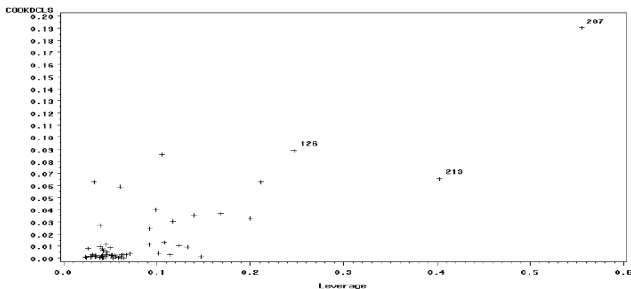
## Data Example 4: Epileptic Seizures Data

- A collection of 59 short longitudinal trajectories, which are equally spaced at 4 time points after randomization.
- Covariates: Baseline count of seizures (Disease severity), age, treatment, and interaction between age and treatment.



## Data Example 4: Epileptic Seizures Data

- ID 207 (in the treatment arm) is a possible outlier, with unduly large counts of epileptic seizures.
- Invoke SAS/IML software developed in Hammill and Preisser (2006) for GEE regression diagnostics and obtain a plot of the Cook's distance versus leverages at the cluster level.





## Data Example 4: Epileptic Seizures Data

Fit the data by GEE and QIF, respectively.

Par	Complete data				Without patient #207			
	Estimate		Std Err		Estimate		Std Err	
	GEE	QIF	GEE	QIF	GEE	QIF	GEE	QIF
intcpt	-2.522	-2.233	1.034	1.006	-2.380	-2.017	0.863	0.892
bsln	1.247	1.193	0.163	0.099	0.987	0.960	0.080	0.066
trt	-0.020	-0.046	0.190	0.141	-0.255	-0.281	0.152	0.146
logage	0.653	0.581	0.287	0.270	0.783	0.680	0.247	0.261
vst	-0.064	-0.052	0.034	0.026	-0.045	-0.047	0.035	0.031
Q-stat	-	3.7	-	-	-	5.9	-	-
TGI	-	40.26	-	-	-	21.76	-	-

## Data Example 4: Epileptic Seizures Data

- To assess the influence of the outlier ID 207 on estimation, use DFBETAS defined as follows:

$$RC(\theta_j) = \frac{|\theta_{j,gee}^{with} - \theta_{j,gee}^{without}|}{s.e.(\theta_{j,gee}^{without})} / \frac{|\theta_{j,qif}^{with} - \theta_{j,qif}^{without}|}{s.e.(\theta_{j,qif}^{without})}.$$

- If  $RC > 1$ , then the outlier would affect the GEE more severely than the QIF.

	Covariate				
	intercept	baseline	treatment	log-age	visit
RC	0.68	0.92	0.96	1.39	3.37

# Modeling Longitudinal Data

- Express the data in matrix notation,  $(y_i, X_i, t_i)$ ,  $i = 1, \dots, N$ , where

$$y_i = (y_{i1}, \dots, y_{in_i})'$$

$$X_i = (x_{i1}, \dots, x_{in_i})$$

$$t_i = (t_{i1}, \dots, t_{in_i})'$$

- For example of Epileptic Seizures Data: For subject ID 104 (placebo, 31 yrs old, 11 seizures during the 8 weeks prior to the randomization),

$$y_1 = (5, 3, 3, 3)'$$

$$X_1 = \begin{bmatrix} 1 & 0 & 1 & 31 & 11 \\ 1 & 0 & 2 & 31 & 11 \\ 1 & 0 & 3 & 31 & 11 \\ 1 & 0 & 4 & 31 & 11 \end{bmatrix}$$

$$t_1 = (2, 4, 6, 8)'$$

## Modeling Longitudinal Data

- A parametric modeling framework assumes that  $y_i$  is a realization of  $Y_i$  drawn from a certain population of the form,

$$Y_i | (X_i, t_i) \stackrel{ind.}{\sim} p(y | X = X_i, t = t_i; \theta), \quad i = 1, \dots, N,$$

where  $\theta$  is the parameter of interest.

- What is  $\theta$ ? Typically,  $\theta = (\beta, \Gamma)$ , where
  - $\beta$  is the parameter vector involved in a regression model for the mean of the population
  - $\Gamma$  represents the other model parameters needed for the specification of a full parametric distribution  $p(\cdot | \cdot)$ , including those in the correlation structure.
- Explicitly specifying such a parametric distribution for nonnormal data is not trivial.
- Multivariate normal! Multivariate binomial? Multivariate Poisson? Multivariate Multinomial? ...

# Modeling Longitudinal Data

- We know how to handle marginals very well from the GLM theory.

$$Y_{ij}|x_{ij}, t_{ij} \sim \text{GLM}(\mu_{ij}, \sigma_{ij}^2)$$

- The mean  $\mu_{ij}$  follows a regression GLM,

$$g(\mu_{ij}) = \eta(x_{ij}, t_{ij}; \beta), \quad j = 1, \dots, n_i$$

- The dispersion (scale) parameter  $\sigma_{ij}^2$  may follow

$$\log(\sigma_{ij}^2) = \zeta(x_{ij}, t_{ij}; \varsigma).$$

- $\sigma_{ij}^2 = 1$  in Poisson and binary data, unless overdispersion (or underdispersion) occurs.

## Specification of the Mean Structure

- Several commonly used marginal models (specification of  $\eta$  function) in the literature.

- (a) Marginal GLM Model takes (the most popular one)

$$\eta(x_{ij}, t_{ij}; \beta) = x'_{ij}\beta,$$

Parameter  $\beta = (\beta_0, \dots, \beta_p)'$  is interpreted as the population-average effects of covariates. They are constant over time as well as across subjects.

- (b) Marginal Generalized Additive Model takes

$$\eta(x_{ij}, t_{ij}; \beta) = \theta_0 + \theta_1(x_{ij1}) + \dots + \theta_p(x_{ijp}),$$

$\beta$  denotes the set of nonparametric regression functions  $\theta_l, l = 0, 1, \dots, p$ . When one covariate is time  $t_{ij}$ , the resulting model characterizes a nonlinear time-varying profile of the data, particularly desirable in longitudinal data analysis.

## Specification of the Mean Structure

- (c) Semi-Parametric Marginal Model includes both parametric and nonparametric predictors, for example,

$$\eta(x_{ij}, t_{ij}; \beta) = \theta_0(t_{ij}) + x'_{ij}\Upsilon,$$

$\beta$  contains both function  $\theta_0(\cdot)$  and coefficients  $\Upsilon$ . The population-average effect of a covariate (e.g. drug treatment) is adjusted by a nonlinear time-varying baseline effect.

- (d) Time-Varying Coefficient Marginal Model follows a GLM with time-varying coefficients,

$$\eta(x_{ij}, t_{ij}; \beta) = x'_{ij}\beta(t_{ij}),$$

$\beta = \beta(t)$  represents a vector of time-dependent coefficient functions. Time-varying effects of covariates, rather than population-average constant effects, are more realistic.

## Specification of the Mean Structure

- (e) Single-Index Marginal Model is specified

$$\eta(x_{ij}, t_{ij}; \beta) = \theta_0(t_{ij}) + \theta_1(x'_{ij}\Upsilon),$$

$\beta$  includes functions  $\theta_0(\cdot)$  and  $\theta_1(\cdot)$  and the vector of coefficients  $\Upsilon$ . It is particularly useful for dimension reduction in the presence of a large number of covariates.

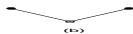
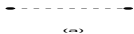
- (f) A certain combination of models (a)-(e).

- Models (a) and (d) are the focus.



# Strategies of Joining Marginal Models

- Not enough to only specify the marginal first moments of the distribution  $p(\cdot)$ .
- A much harder task is to specify higher moments of the joint distribution  $p(\cdot)$  or even the joint distribution itself.
- The marginals have to be joined by a certain suitable correlation structure.
- Three popular strategies of modeling: (a) **Quasi-likelihood Modeling**, (b) **Conditional Modeling**, and (c) **Joint Modeling**.



## Quasi-likelihood (QL) Modeling Approach

- Do not fully specify the joint distribution  $p(\cdot)$ , but only specify its first two moments, including a correlation structure.
- The minimal set of model conditions required to make a valid statistical inference.
- The QL approach explicitly specifies the covariance of the data,  $V_i = Cov(Y_i|X, t_i)$ :

$$V_i = \text{diag} \left[ \sqrt{\text{Var}(Y_{ij})} \right] R_i \text{diag} \left[ \sqrt{\text{Var}(Y_{ij})} \right]$$

where the key component is the correlation matrix  $R = [\alpha_{ts}]$  of  $Y_i$ .

- How to specify  $R$ ?
  - Pearson correlation of linear dependency
  - Odds ratio for association between categorical outcomes

## Common Types of Correlation Structures

- (1) (*Independence*) Assumes all pairwise correlation coefficients are zero:

$$\gamma(Y_{it}, Y_{is}) = 0, \quad t \neq s,$$

- (2) (*Unstructured*) Assumes all pairwise correlation coefficients are different parameters:

$$\gamma(Y_{it}, Y_{is}) = \alpha_{st} = \alpha_{ts}, \quad t \neq s,$$

- (3) (*Interchangeable, Exchangeable, Compound symmetry*) Assumes pairwise correlation coefficients are equal

$$\gamma(Y_{it}, Y_{is}) = \alpha, \quad t \neq s,$$

## Common Types of Correlation Structures

- (4) (*AR-1*) Assumes the correlation coefficients decay exponentially over time

$$\gamma(Y_{it}, Y_{is}) = \alpha^{|t-s|}, \quad t \neq s,$$

- (5) (*m-dependence*) Assumes the responses are uncorrelated if they are apart more than  $m$  units in time, or  $|t - s| > m$ ,

$$\gamma(Y_{it}, Y_{is}) = \alpha_{ts}, \text{ for } |t - s| \leq m,$$

## Which Correlation Structure Is Suitable?

Invoke a residual analysis with the following steps:

- Step I:** Fit longitudinal data by a marginal GLM under the independence correlation structure, and output fitted values  $\hat{\mu}_{it}$ .
- Step II:** Calculate the Pearson-type residuals, which presumably carry the information of correlation that was originally ignored in Step I:

$$r_{it} = \frac{y_{it} - \hat{\mu}_{it}}{\sqrt{V(\hat{\mu}_{it})}}, t = 1, \dots, n_i, i = 1, \dots, N,$$

where  $V(\cdot)$  is the variance function chosen according to the marginal model.

## Which Correlation Structure Is Suitable?

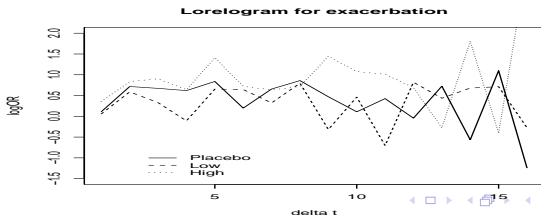
- Step III:** Compute the pairwise Pearson correlations  $\hat{\gamma}_{ts}$  of the residuals for each pair of fixed indices  $(t, s)$ , which produces a sample correlation matrix  $\hat{R} = (\hat{\gamma}_{ts})$ .
- Step IV:** Examine the pattern of matrix  $\hat{R}$ , to match with one of those listed above.

Step III may be modified as sample log-odds ratios for categorical responses, called *lorelogram* (Heagerty and Zeger, 1998):

$$\text{LOR}(t_j, t_k) = \log \text{OR}(Y_j, Y_k).$$

## Which Correlation Structure Is Suitable?

- For the example of Multiple Sclerosis Trial data, the lorelograms of the observed exacerbation incidences across the 3 treatment groups
- More rigorous decision may be made via a certain model selection criterion.



## Conditional Modeling Approach

- **Latent Variable Approach:** Conditional on a latent variable  $b$ ,  $Y = (Y_1, \dots, Y_n)'$  are independent,

$$Y = (Y_1, \dots, Y_n)' | b \sim p(y_1 | b) \cdots p(y_n | b).$$

where conditional distributions are 1-dimensional, so the GLM theory can be applied.

- The joint distribution  $p(\cdot)$  is obtained by

$$\begin{aligned} p(y | X, t) &= \int p(y, b | X, t) db \\ &= \int \prod_{i=1}^n p(y_i | b, X, t) p(b | X, t) db, \end{aligned}$$

- The correlation structure is induced from the specification of the latent variables and their distributions.
- How many latent variables?



# Transitional Model Approach

- For subject  $i$ ,

$$p(y_{i1}, \dots, y_{in_i} | X_i, t_i) = f(y_{in_i} | y_{i,n_i-1}, \dots, y_{i,1}, X_i, t_i) \times \dots \\ \dots \times f(y_{i2} | y_{i1}, X_i, t_i) f(y_{i1} | X_i, t_i)$$

- For example, the transitional logistic model

$$\text{logit}P[Y_{it} = 1 | y_{it-1}, y_{it-2}, \dots, y_{it-q}] = x'_{it}\beta + \sum_{j=1}^q \psi_j y_{it-j}$$

- Use existing software packages to fit transitional models with appropriate form of covariates.

# Joint Modeling Approach

- Directly specify the joint distribution  $p(\cdot)$  of the data.
- Mostly *ad hoc* methods, but few general frameworks available.
- Song et al. (2009) proposed so-called **vector generalized linear models** based on Gaussian copulas. Also refer to

**Song (2007, Ch. 6) “Correlated Data Analysis: Modeling, Analytics and Applications.” Springer.**

## Quasi-Likelihood Approach: GEE

- GEE (Generalized Estimating Equation) was first termed by Liang and Zeger (1986, *Biometrika*).
- The idea of estimating equations (or estimating functions) has been around in the statistical literature for more than 3 decades. For example, Fisher (1935), Kimball (1946) and Godambe (1960, *Ann. Math. Statist.*)
- Instead of the formulation of a likelihood function, **directly specify an analog to the likelihood equation** for the parameter of interest.

## Formulation of GEE

- In the longitudinal data case, the GEE takes the form

$$U(\beta) = \sum_{i=1}^N \dot{\mu}'_i V_i^{-1} (y_i - \mu_i) = 0,$$

where the **working** covariance matrix

$$V_i = \text{diag} \left[ \sqrt{\text{Var}(Y_{ij})} \right] R_i(\alpha) \text{diag} \left[ \sqrt{\text{Var}(Y_{ij})} \right]$$

- $R_i(\alpha)$  is the **working** correlation matrix.
- Where is the  $\beta$ ?
- Replace **nuisance** correlation parameters  $\alpha$  by a “good” estimate,  $\hat{\alpha}$ , in the GEE above, and then solve it for  $\hat{\beta}$ , which is the solution to the GEE.
- SAS PROC GENMOD or R gee package implemented this approach.

## Merits of the GEE Method

- It is useful to evaluate the population-average effects of covariates.
- It is simple, as it only requires to correctly specify the first two moments of the underlying distribution of the data.
- It is robust against the model misspecification on the correlation structure.
- It is easy to implement numerically using available software packages such as SAS and R. This is really under the framework of Weighted Least Squares.

## Caveats of the GEE Method

- (1) First underlying assumption is that data are relatively homogeneous, in the sense that the variation in the response is mostly due to different levels of covariates (not due to subject-specific variation).
- (2) Second underlying assumption is that the first moment mean model is correct,

$$g(\mu_{ij}) = x'_{ij}\beta$$

- (3) Third underlying assumption is that the nuisance correlation parameter  $\alpha$  is properly estimated.
- (4) Fourth underlying assumption is that missing data are missing completely at random (MCAR).
- (5) No way of performing model selection because of the lack of an objective function in the estimation procedure.
  - Quadratic Inference Function (QIF) can help to deal with (2), (3), and (5).

# Generalized Estimating Equations: Liang & Zeger, 1986

- Attractive: only requires the first two moments of the likelihood function
- Misspecified working correlation does not affect the consistency of the regression parameter estimation ( $\hat{\beta}$ )
- Provides robust sandwich estimator for the variance of regression parameter estimator

## Drawbacks of GEE

- Misspecification of working correlation does affect the efficiency of the regression parameter estimation
- Lack of objective function, multiple roots problem (Small et al., 2000)
- Lack of inference function for testing, goodness-of-fit test for model assumptions such as LRT (Heagerty & Zeger, 2000)
- Sensitive to outliers



## Extension of the GEE

- Prentice and Zhao (1991) proposed estimating equations for jointly modelling the mean and covariance parameters
- Qu et al. (2000): introduced the QIF to improve the efficiency of GEEs
- Balan and Schiopu-Kratina (2005): derived a two-step estimation procedure for the marginal model based on the pseudolikelihood
- Chiou and Müller (2005): developed a new marginal approach based on semiparametric quasi-likelihood regression
- Hall and Severini (1998): Quasilielihood GEE approach

## Extension of the GEE

- Shults and Chaganty (1998): Quasi-least square
- Stoner and Leroux (2002): Optimal estimating equation approach
- Pourahmadi (1999, 2000): Modified Cholesky decomposition
- Wang and Carey (2003): Working correlation misspecification
- Pan and Mackenzie (2003): joint modeling of mean-covariance structures for GEE

## Quadratic Inference Function (Qu, Lindsay and Li, 2000)

- Has advantages of the estimating function approach
- Does not require the specification of the likelihood
- Provides an objective function
- Provides a valid inference function for goodness-of-fit tests, with properties similar to the LRT (Heagerty & Zeger, 2000)

# Quadratic Inference Function

- Estimation:
  - Improves efficiency of regression estimators under GEE setting
  - Robust properties for QIF estimator
- Inference function
  - Behave as minus twice log likelihood, has similar properties as likelihood ratio test
  - Testing ignorable missingness for estimating equation approaches

# Generalized Estimating Equations

- Appropriate when inference of the population-average is the focus Liang and Zeger (1986), Hardin and Hilbe (2003)
- Relate the marginal mean  $\mu_{ij}$  to the covariates:

$$g(\mu_{ij}) = x'_{ij}\beta, \quad \beta \in \mathcal{B}$$

where  $g$  is a known link function, and  $\beta = (\beta_1, \dots, \beta_q)'$  is a  $q \times 1$  vector of unknown regression parameters

- The variance of  $y_{ij}$  is a function of the mean:

$$\text{Var}(Y_{ij}) = \phi V(\mu_{ij})$$

# Generalized Estimating Equations

- GEE estimator is the solution of

$$\sum_{i=1}^N \dot{\mu}_i' V_i^{-1} (y_i - \mu_i) = 0,$$

where  $\dot{\mu}_i = \partial \mu_i / \partial \beta$  is a  $n_i \times q$  matrix, and  $V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2}$  with  $A_i$  being the diagonal matrix of marginal variances  $\text{Var}(y_{ij})$  and  $R_i(\alpha)$  being the working correlation matrix

## Quadratic Inference Function (QIF)

- Quadratic Inference Function (Qu, Lindsay and Li, Biometrika, 2000) is motivated by observing:

$$R^{-1} \approx \sum_{i=0}^k a_i M_i,$$

where  $M_0$  is the identity matrix,  $M_1, \dots, M_k$  are basis matrices and  $a_0, \dots, a_k$  are constant coefficients

# Working Correlation

- Exchangeable:  $R^{-1} = a_0 I + a_1 M_1$

$$M_1 = \begin{pmatrix} 0 & 1 & \dots & & 1 \\ & 0 & 1 & \dots & 1 \\ & & \ddots & & \\ & & & & 0 \end{pmatrix}$$

- AR-1:  $R^{-1} = a_0^* I + a_1^* M_1^* + a_2^* M_2^*$

$$M_1^* = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ & 0 & 1 & 0 & \dots & 0 \\ & & \ddots & & & \\ & & & & & 0 \end{pmatrix}, \quad M_2^* = \begin{pmatrix} 1 & 0 & \dots & 0 \\ & 0 & \dots & 0 \\ & & \ddots & \\ & & & 1 \end{pmatrix}$$



# Hybrid Working Correlation and Adaptive Estimating Equations

- Choose all available basis matrices  $I$ ,  $M_1$ ,  $M_1^*$
- Performs well if the selected basis matrices contain true correlation structure
- If there is no appropriate working correlation  

$$\sum \dot{\mu}_i' \hat{V}_i^{-1} (y_i - \mu_i) = 0$$
- But  $V_i^{-1}$  is difficult to estimate for large dimensions: the smallest eigenvalue gives highest weight
- Create adaptive estimating equations (Qu & Lindsay, 2003)

$$g_N = \sum g_i = \begin{pmatrix} \sum (\dot{\mu}_i)' A_i^{-1} (y_i - \mu_i) \\ \sum (\dot{\mu}_i)' \hat{V}_i (y_i - \mu_i) \end{pmatrix},$$

where  $\hat{V}_i = A_i^{1/2} \hat{R} A_i^{1/2}$  is the sample variance

## Quadratic Inference Function and GEE

- The GEE is a linear combination of the elements of the following extended score vector

$$\begin{aligned} \bar{g}_N(\beta) &= \frac{1}{N} \sum_{i=1}^N g_i(\beta) \\ &\approx \frac{1}{N} \begin{pmatrix} \sum_{i=1}^N \dot{\mu}_i' A_i^{-1} (y_i - \mu_i) \\ \vdots \\ \sum_{i=1}^N \dot{\mu}_i' A_i^{-1/2} M_k A_i^{-1/2} (y_i - \mu_i) \end{pmatrix}, \end{aligned}$$

where  $\mu_i = E(Y_i|X_i)$ , and the definition of  $g_i(\beta)$  should be clear from the context

## Quadratic Inference Function and GEE

- Qu et al. proposed to minimize the quadratic inference function (also used in generalized method of moments, Hansen, 1982):

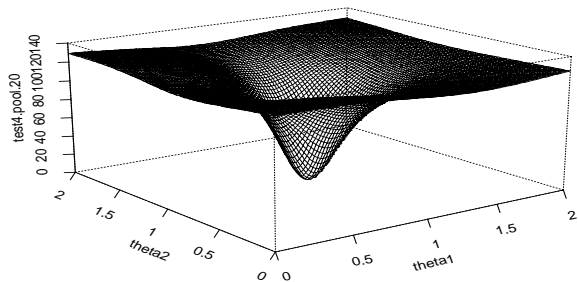
$$Q_N(\beta) = N\bar{g}_N(\beta)C_N^{-1}(\beta)\bar{g}_N(\beta),$$

where  $C_N(\beta) = N^{-1} \sum_{i=1}^N g_i(\beta)g_i'(\beta)$  is the sample covariance matrix

- The asymptotic variance of the estimator attains the minimum in the sense of Löwner ordering. QIF improves the efficiency of GEE when the working correlation is misspecified and remains as efficient as GEE when the working correlation is correct

# Properties of QIF

- Asymptotic normality  $\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} N(0, (D'\Sigma^{-1}D)^{-1})$   
where  $D = E(\partial g_i(\beta)/\partial \beta)$  and  $\Sigma = E(g_i(\beta)g_i'(\beta))$
- If  $\dim(g_N) = \dim(\beta)$ 
  - Minimizing  $Q_N$  is equivalent to solving  $g_N = 0$
  - The variance of  $\hat{\beta} = (D^{-1}\Sigma D^{-1})$  is the sandwich estimator
  - Define  $J = D'\Sigma^{-1}D$  as Godambe information
- The range of QIF:  $0 \leq Q \leq N$ , where  $N$  is the sample size of subjects

Quadratic Inference Functions for 2-dim  $\beta$ 

## Comparison of GEE and QIF Simulated Data

- Identity link:  $\mu(x_{it}, \beta) = x_{it}'\beta$
- Covariate  $x_i = (0.1, 0.2, \dots, 1.0)$
- Response  $y_i = \beta x_i + \varepsilon_i$
- Error  $\varepsilon_i \sim N_{10}(0, \Sigma)$ , where  $\Sigma$  is equi-corr or AR-1
- Number of clusters  $N = 80$ , and cluster size  $n = 10$

# Simulated relative efficiency

$$\text{SRE} = \frac{\text{m.s.e. of GEE}}{\text{m.s.e. of QIF}}$$

SRE of  $\beta$  (after 10,000 simulations) for  $E[y_{it}] = \beta x_{it}$ .

True $R$	Working $R$		
	$\rho$	exchangeable	AR-1
exchangeable	0.3	0.99	1.20
	0.7	0.99	2.04
AR-1	0.3	1.04	0.97
	0.7	1.37	0.98

# Inferential Properties of QIF

- QIF is analog to twice of negative log likelihood, so use the difference to test hypotheses, behave like likelihood ratio test
- Apply  $Q(\beta_0) - Q(\hat{\beta})$  to test  $H_0 : \beta = \beta_0$
- Apply  $Q(\psi_0, \tilde{\lambda}) - Q(\hat{\psi}, \hat{\lambda})$  to test  $H_0 : \psi = \psi_0$  if  $\beta = (\psi, \lambda)$
- Apply  $Q(\hat{\beta})$  to test goodness-of-fit for extended score moment conditions ( $E[g(\beta_0)] = 0$ ) (Hansen, 1982)



## Inferential Properties of QIF

- To test  $H_0 : \psi = \psi_0$  if  $\beta = (\psi, \lambda)$ , under the null,  $Q(\psi_0, \tilde{\lambda}) - Q(\hat{\psi}, \hat{\lambda})$  is asymptotically  $\chi_p^2$ , where  $p$  is the dimension of  $\psi$ ,  $\tilde{\lambda}$  is the minimizer of  $Q(\psi_0, \lambda)$  and  $(\hat{\psi}, \hat{\lambda})$  is the minimizer of  $Q(\psi, \lambda)$
- To test  $H_0 : \beta = \beta_0$ , suppose  $\beta$  has dimension  $p$ ,  $Q(\beta_0) - Q(\hat{\beta})$  is asymptotically  $\chi_p^2$  under the null, where  $\hat{\beta}$  is the minimizer of  $Q(\beta)$
- Goodness-of-fit test (Hansen, 1982): Suppose  $g$  has dimension  $r$  and  $\beta$  has dimension  $p$  with  $p < r$ . Then, under  $H_0 : E[g(\beta_0)] = 0$  the asymptotic distribution of  $Q(\hat{\beta})$  is  $\chi^2$  with  $r - p$  degrees of freedom

# Simulations

- Compare the relative power of the GEE's Wald test and the QIF test
- Each of  $N$  subjects hadve 4 repeated measurements of a trait, either continuous or dichotomous, and at each time two covariates, exposure ( $E$ ) and treatment type ( $T$ ) were recorded.
- Exposure  $E$  was a time-dependent continuous variable generated independently from a uniform distribution on interval  $(0, 1)$ .
- Treatment status  $T$  is a binary variable generated from a Bernoulli distribution with probability 0.5 for each subject

# Simulations

- For normal responses,

$y_{it} = \beta_0 + \beta_1 E_{it} + \beta_2 T_i + \varepsilon_{it}, i = 1, \dots, 200, t = 1, \dots, 4,$   
with  $(\varepsilon_{i1}, \dots, \varepsilon_{i4})' \sim \text{MVN}(0, \sigma^2 R(\alpha))$ , where  $\sigma^2 = 1$  and the correlation matrix  $R(\alpha)$  varies

- For binary response, an algorithm suggested by Preisser et al. was used with the same mean model subject to a logit link and sample size  $N = 500$

# Test Size and Power:

**Table:** Type I errors and Empirical relative power (ERP) between the GEE's Wald test and the QIF's score-type test ( $\text{Power}(\text{QIF})/\text{Power}(\text{GEE})$ ), for the significance of treatment based on 1000 simulation and the true correlation structure AR-1 for both continuous and binary data.

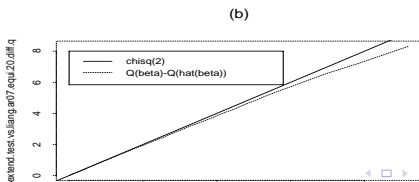
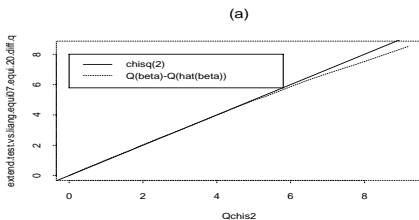
Model	Exchangeable					
	$\alpha = 0.3$			$\alpha = 0.7$		
	Type I error		Power	Type I error		Power
	QIF	GEE	ERP	QIF	GEE	ERP
Normal	0.055	0.056	1.03	0.046	0.043	1.00
Binomial	0.046	0.046	1.09	0.052	0.046	1.05
	AR-1					
Normal	0.052	0.052	1.01	0.052	0.056	1.02
Binomial	0.052	0.053	1.07	0.054	0.053	1.08

# Chi-square Tests

- All these test statistics follow  $\chi^2$  asymptotically whether or not the working correlation is true or not
- This contrasts to Rotnitzky & Jewell's (1990) score test which performs well only if the working correlation is specified correctly
- The tests are more powerful compared to other tests
- The degrees of freedom in the tests is similar to likelihood ratio test df

# Quantile-quantile Plots for Chi-square Tests

- Quantile-quantile plot of  $Q(\beta) - Q(\hat{\beta})$  relative to  $\chi_2^2$  (a)  
When working correlation is true (b) When working correlation is misspecified



# Robustness

- Efficient when the model is correct
- For the estimating equation approach, the model is correct means  $E(g) = 0$
- Robust means the estimator is consistent when the model is incorrect
- Contamination occurs often in longitudinal studies, e.g., misclassification for a binary response (coding 0 and 1 may be mistakenly switched) or covariates are contaminated
- Moment conditions  $E(g) = 0$  can be distorted by contaminated or irregular measurements
- GEE method fails to give consistent estimators if few clusters are irregular (Preisser & Qaqish, 1996, 1999; Mills et al. 2002)

## Strategies for Robustness

- Downweighting and deleting putatively contaminated clusters (Preisser & Qaqish, 1996, 1999)
- Relies on whether or how potentially problematic clusters are identified beforehand
- Adjustments to data can change the original moment conditions of the model, the moment conditions might not be valid for other non-contaminated
- QIF carries a downweighting strategy automatically in the estimation procedure through weighting matrix  $C$
- QIF does not require deletion or downweighting
- QIF behaves robustly against irregular measurements arising from either response or covariates



## Influence Function (IF)

- Define influence function

$$IF(x, T, P_\beta) = \lim_{\varepsilon \rightarrow 0} \frac{T((1 - \varepsilon)P_\beta + \varepsilon\Delta_x) - T(P_\beta)}{\varepsilon}$$

where  $T(\cdot)$  is an estimator for parameter  $\beta$ ,  $P_\beta$  is the distribution function,  $\Delta_x$  is the probability measure with mass 1 at the single-point contaminated data  $x$

- Measure how a single point changes the estimator
- A crucial property for robustness of an estimator is to have a bounded influence function (IF) (Hampel et al., 1986)

# Robustness

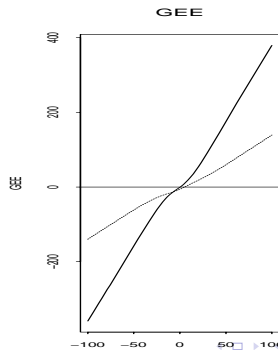
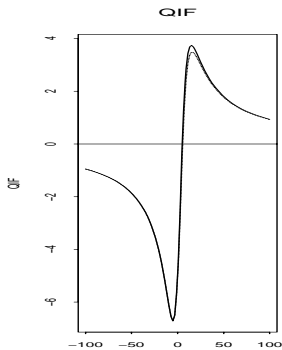
- Influence function of GEE is unbounded
- GEE:  $\sum \dot{\mu}'_i A_i^{-1/2} R^{-1} A_i^{-1/2} (y_i - \mu_i) = 0$
- Diagonal marginal variance  $A_i$  has a parametric form, a function of  $\mu$
- Working correlation  $R(\alpha)$  measures the strength of association between pairs
- Neither  $A_i$  or  $R(\alpha)$  have a downweighting effect on the irregular observations with large variation
- Minimizing the QIF is asymptotically equivalent to solving

$$\sum_i D' C^{-1} g_i = 0,$$

where  $C^{-1} = (1/N \sum g_i g_i')$  downweights clusters with large variation

# Redescending Property

- The QIF estimator has a redescending property (Qu & Song, 2004). That is, the estimating function  $D'C^{-1}g_i(z)$  is bounded and approaches to zero as  $\|z\| \rightarrow \infty$ , where  $z = y - \mu$
- The GEE estimator does not have a redescending property



# Simulation

- There are 50 clusters, cluster size of 10
- Use linear model  $y_i = \beta x_i + \varepsilon_i$ , where the true  $\beta_0 = 1$
- One outlier on one subject's response variable was introduced as  $100 * y_{it}$
- The proportions of contaminated clusters were chosen to be 0%, 10%, 20%, 50%, and 100%

True correlation is AR-1,  $\rho = 0.5$

%	AR-1		Exchangeable		Unspecified	
	GEE	QIF	GEE	QIF	GEE	QIF
0	0.001 (0.11)	0.002 (0.11)	0.004 (0.12)	0.001 (0.11)	0.004 (0.48)	0.001 (0.12)
10	2.56 (1.19)	0.02 (0.13)	2.54 (1.18)	0.02 (0.22)	3.17 (1.60)	0.02 (0.13)
20	5.15 (1.64)	0.03 (0.14)	5.07 (1.60)	0.03 (0.24)	7.30 (2.82)	0.03 (0.14)
50	<b>12.91</b> (2.68)	<b>0.09</b> (0.15)	<b>12.35</b> (2.45)	<b>0.005</b> (0.27)	<b>23.35</b> (6.12)	<b>0.09</b> (0.15)
100	<b>25.63</b> (3.58)	<b>0.27</b> (0.18)	<b>23.01</b> (2.76)	<b>0.05</b> (0.33)	<b>53.90</b> (9.78)	<b>0.27</b> (0.179)

## Example

- Data example from Preisser & Qaqish (1999) on urinary incontinence
- The response variable is binary, indicating whether the subject's daily life is bothered by accidental loss of urine (1 bothered and 0 otherwise)
- Subjects are correlated if they are from the same hospital practice
- There were a total of 137 patients from 38 hospital practices
- Each cluster contained minimum 1 patient to maximum 8 patients
- There are 5 covariates: gender, age, daily leaking accidents, severity of leaking and number of times to use toilet daily

## Example

- The logistic link function is assumed

$$\text{logit}(\mu) = \theta_0 + \theta_1 \text{female} + \theta_2 \text{age} + \theta_3 \text{dayacc} + \theta_4 \text{severe} + \theta_5 \text{toilet}$$

- The marginal variance matrix  $A_i$  is a diagonal matrix with diagonal components  $A_{ij} = \mu_{ij}(1 - \mu_{ij})$
- The exchangeable working correlation is assumed for both the GEE and the QIF
- For the QIF method, let  $M_1 = I$ ,  $M_2$  be 0 on the diagonal and 1 off the diagonal

## Example

- Patients 8 and 44 were identified as suspected outliers by Preisser & Qaqish (1999): large “dayacc” and “toilet,” but the response values are not bothered
- The ordinary GEE estimators are very sensitive to these two outliers
- Covariate “severe” is significant based on the full data, but insignificant after removing 2 patients
- Covariate “toilet” becomes significant after removing 2 patients

# Comparisons between GEE and QIF for outlying observations

	All observations		Remove 2 patients	
	GEE	QIF	GEE	QIF
Intcpt	<b>-3.18</b>	<b>-2.81</b>	<b>-3.15</b>	<b>-2.89</b>
Female	-1.24	<b>-2.02</b>	-1.94	<b>-2.80</b>
Age	-1.21	-1.16	-1.74	-1.78
Dayacc	<b>4.20</b>	<b>3.59</b>	<b>4.65</b>	<b>3.75</b>
Severe	<b>2.26</b>	1.51	<b>1.76</b>	1.63
Toilet	<b>1.09</b>	<b>2.51</b>	<b>2.64</b>	<b>2.80</b>



# Application: Testing for Missing Data Mechanism using goodness-of-fit test

- Missing indicator  $I_m = \begin{cases} 1 & \text{missing} \\ 0 & \text{o.w.} \end{cases}$
- Response variable  $Y = (Y_o, Y_m)$
- Missing completely at random (MCAR):  $I_m$  does not depend on  $Y_o, Y_m$
- Missing at random (MAR):  $I_m$  depends on  $Y_o$ , but not  $Y_m$
- Informative:  $I_m$  depends on  $Y_m$

## Motivating Example

- A real data example by Rotnitzky & Wypij (1994, Biometrics)
- Studies of pediatric asthma in Steubenville, Ohio
- Dichotomous outcomes of asthma status were recorded for children at ages 9 and 13
- The marginal probability is modeled as a logistic regression

$$\text{logit}\{\text{pr}(y_{it} = 1)\} = \beta_0 + \beta_1 I(\text{male}) + \beta_2 I(\text{age}=13)$$

where  $y_{it} = 1$  if the  $i$ th child had asthma at time  $t = 1, 2$  and  $I(E)$  is the indicator function for event  $E$

- 20% of the children had asthma status missing at age 13
- Every child had their asthma status recorded at age 9, but for some the asthma status was missing at age 13

# Motivating Problem

- There are three parameters  $\beta_0, \beta_1$  and  $\beta_2$  in the model with complete observations
- But only two identifiable parameters,  $\beta_0$  and  $\beta_1$ , for the incomplete case
- Dimensions of parameters are different for different missing patterns.

## Comparison to MCAR, MAR

- The distinction of MCAR, MAR and informative missing (Rubin, 1976) is based on the likelihood
- MLE ignoring missing data is valid for MAR (Rubin, 1976)
- The distinction between ignorable and non-ignorable missing here is based on whether mean zero assumptions of estimating equations are valid or not
- Since the estimator (GEE or QIF) is consistent if the estimating functions are unbiased  $E(g) = 0$

## Ignorability and Non-ignorability (Qu & Song, 2002)

- Suppose  $g_1$  and  $g_2$  are constructed from complete and incomplete data
- Let  $g = (g_1, g_2)'$
- If  $E_\beta(g_1) = E_\beta(g_2) = 0$ , then the missing is ignorable
- If  $E_\beta(g) \neq 0$ , then missing is nonignorable
- Example: treatment effect is positive for patients who complete trials, and 0 or negative for dropout patients, missing is not MCAR

## Goodness-of-Fit Test

- Suppose there are  $R$  missing data patterns
- Let  $g = (g_1, \dots, g_R)'$
- $\dim g_i$  could be different for different missing patterns
- Test  $H_0 : E(g) = 0$
- QIF  $Q = g' C^{-1} g = \sum_{i=1}^R g_i' C_i^{-1} g_i$ , where  $C_i = \hat{\text{var}}(g_i)$
- Based on goodness-of-fit test,  $Q(\hat{\beta}) \xrightarrow{d} \chi_{r-p}^2$  under  $H_0$ , where  $r$  is the total dimension of estimating equations

## Asthma studies

- Construct two sets of estimating equations from complete and incomplete data

$$g_1(\beta_0, \beta_1, \beta_2) = \sum (X_i^c)'(y_i^c - \mu_i^c)$$

$$g_2(\beta_0, \beta_1) = \sum (X_i^m)'(y_i^m - \mu_i^m)$$

- The QIF  $g_1' C_1^{-1} g_1 + g_2' C_2^{-1} g_2$
- The goodness-of-fit test statistic is the minimum of  $Q$ , which is 4.68
- Degrees of freedom  $3 + 2 - 3 = 2$
- The  $p$ -value from chi-squared test is 0.096 (not significant)
- Consistent with Chen and Little's (1999) Wald test
- Performs better than Wald test (Chen & Little, 1999) for small samples

## Schizophrenia data

- The outcomes are measured from the same subject over time
- Example: Adult schizophrenia trial at Harvard U. (Hogan & Laird, 1997)
- New therapy (NT) vs. standard therapy (ST)
- Longitudinal trial: week 0, 1, 2, 3, 4, 6
- Response variable: rating score 0-108, the higher the worse
- Covariates: baseline score, week, treatment
- For NT group, 46% dropout; for ST group, 34% dropout



# Schizophrenia trial

- Let  $\mu = E(y) = \beta_0 + \beta_1 * \text{trt} + \beta_2 * \text{base} + \beta_3 * \text{week}$
- $\text{trt} = 1$  (new therapy) and  $0$  (standard therapy)
- $\text{Week} = 1, 2, 3, 4, 6$
- There are 5 sets of equations for complete data and 4 missing patterns
- Estimate  $\beta$  by minimizing the QIF

$$Q = \sum_{i=1}^5 g_i' C_i^{-1} g_i, C_i = \text{var}(g_i)$$

# Schizophrenia trial

	intcpt	trt	base	week
estimates	5.49	0.89	0.62	-1.76
s.e.	3.45	1.91	0.09	0.25
t-ratio	1.59	0.47	7.17	-7.04

- New treatment is doing slightly poorly, but not statistically significant
- Consistent with results from Hogan & Laird (1997), Sun & Song (2001)
- $Q(\hat{\beta}) = 0.41 + 5.41 + 8.58 + 10.36 + 6.73 = 31.49$
- With  $df = 5 \cdot 4 - 4 = 16$ ,  $p$ -value = 0.012
- Strong indication that missing data are non-ignorable

## Sample Size and Power Determination

- Sample size determination is a paramount component at the design process of clinical trials  
Primarily compare the effect of a test treatment to that of a controlled treatment
- Goal: To transit the benefit of efficiency gain of the QIF into a design of longitudinal study, now demonstrate the sample size by the QIF compared to the GEE, and the hence the reduction of study costs
- Consider designs based on Wald test
- No need to estimate correlation parameter in the design

## A General Framework

- Consider a hypothesis testing form: A linear combination of the regression coefficients

$$H_0 : H\beta = 0 \quad \text{vs} \quad H_1 : H\beta = h_0 \neq 0$$

- Consider the Wald test in both GEE and QIF methods

$$(H\hat{\beta} - H\beta)'(HJ(\hat{\beta})^{-1}H')^{-1}(H\hat{\beta} - H\beta) \sim \chi^2(\text{rk}(H))$$

- Under  $H_0$ , we reject  $H_0$  at the  $\alpha$  level of significance if

$$(H\hat{\beta})'(HJ(\hat{\beta})^{-1}H')^{-1}(H\hat{\beta}) > \chi_{\text{rk}(H), 1-\alpha}^2.$$

- Under  $H_1$ ,

$$(H\hat{\beta})'(HJ(\hat{\beta})^{-1}H')^{-1}(H\hat{\beta}) \sim \chi_{\text{rk}(H), \lambda}^2,$$

where  $\lambda$  is the non-centrality parameter,

$$\lambda = h_0'(HJ(\hat{\beta})^{-1}H')^{-1}h_0$$

# A General Framework

- The power of the Wald test is

$$\text{power} = 1 - \beta = \int_{\chi_{rk(H), 1-\alpha}^2}^{\infty} f(x, rk(H), \lambda) dx,$$

where  $\beta$  is the type II error and  $f(x, rk(H), \lambda)$  is the probability density function of the  $\chi_{rk(H), \lambda}^2$

## Case I: Normal Longitudinal Data

The QIF and GEE have comparable efficiency, so the GEE based design is adequate

## Case II: Dichotomous Longitudinal Data

- Consider a logistic model with dichotomous outcomes:

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 d_i + \beta_2 t_{ij} + \beta_3 d_i t_{ij},$$

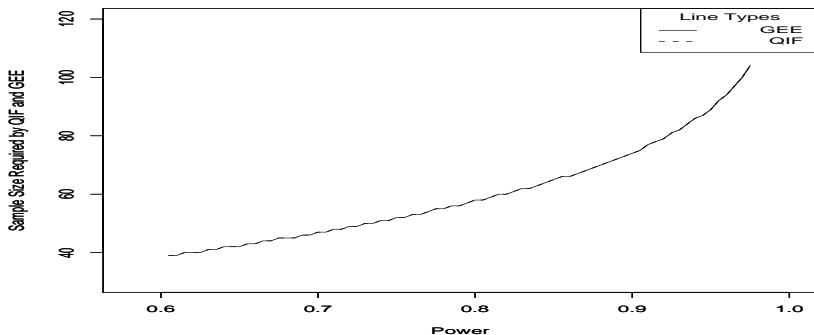
- $\mu_{ij} = P(y_{ij} = 1 | d_i, t_{ij})$ : the probability of  $y_{ij} = 1$
- $d_i$ : Indicator of treatment group:

$$d_i = \begin{cases} 1, & \text{if subject } i \text{ is in test treatment group,} \\ 0, & \text{if subject } i \text{ is in controlled treatment group,} \end{cases}$$

- $t_{ij}$ : Time of  $j$ -th visit for subject  $i$
- For convenience, consider a design with homogeneous visit times; that is,  $t_{ij} = t_j$  for all  $i$

## Simulation Results: Correctly specified correlation

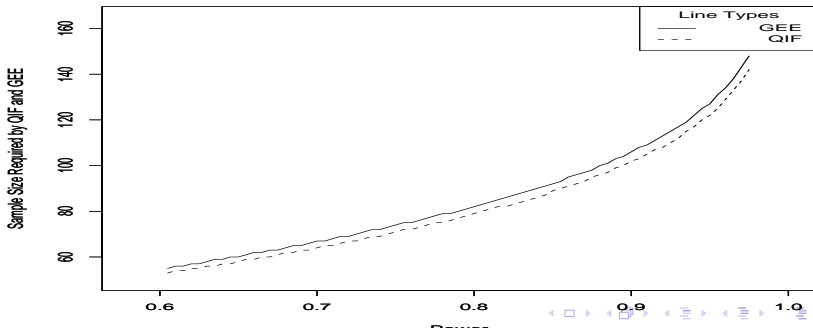
- The true correlation structure is exchangeable
- As expected, both GEE based and QIF based designs require the same sample size to reach the same statistical power





## Simulation Results: Misspecified correlation

- The true correlation structure is 1-dependence (generating the data) and the working correlation structure is exchangeable (used in the design)
- The QIF based design requires smaller sample size than the GEE based design



## Simulation Results: Misspecified correlation

$\rho$	Effect size	Power					
		0.80		0.85		0.9	
		QIF	GEE	QIF	GEE	QIF	GEE
0.2,	(6, 0, 0.4, 0.1)	88	89	99	100	113	114
	(6, 0.5, 0.4, 0.1)	90	91	102	103	116	118
	(6, 2.5, 0.4, 0.1)	96	97	108	109	123	125
0.4,	(6, 0, 0.4, 0.1)	100	107	112	120	129	138
	(6, 0.5, 0.4, 0.1)	103	110	116	124	133	142
	(6, 2.5, 0.4, 0.1)	109	116	123	131	141	150

## Remarks

- R Function: QIFSAMS (QIF Sample Size) is soon available for download
- The QIF based design provides better power to detect the treatment effect in longitudinal clinical trials
- The QIF based design requires a smaller sample size than the GEE based design in order to reach the same statistical power
- The simulation study indicates that in some occasions the QIF based design can save 25% sample size in comparison to the GEE based design

# Conclusion

- The likelihood function is often unknown or difficult to formulate
- QIF has the advantages of the estimating function approach
- It does not require the specification of the likelihood
- It provides an objective function with 0 as the lower bound, which guarantees the existence of the global minimum
- The minimum is the test statistic and the minimizer is the estimator
- The QIF approach does not require estimation of nuisance parameters involved in working correlations

# Conclusion

- The QIF estimator is highly efficient
- QIF estimator is not sensitive to outliers, downweighting outliers automatically
- The maximum likelihood and GEE based estimators are sensitive to the outliers
- Since the criteria for outliers are often arbitrary, it is not scientifically objective to remove “outliers” in order to make the model fit better or produce significant test results
- The inference function is useful for hypothesis testing for regression parameters
- Provide the goodness-of-fit test for model assumptions

# Introduction

- Standard marginal generalized linear models with parametric effects are restrictive for modeling complex covariate effects
- It is important to develop nonparametric estimation for covariate effects
- Nonparametric approaches for correlated data literature: Wang (1998, 1998), (Opsomer et al., 2001)
- Most of the nonparametric literature focuses on consistent and efficient estimation, including kernel and spline approaches by Lin and Carroll (2000, 2001), Wang (2003), Lin et al. (2004), and Wang et al. (2005)
- Inference function and model checking tools are not well developed

# Introduction

- Most of these approaches also treat response variables as normal outcomes
- Zhang (2004) proposed generalized linear mixed models for hypothesis testing for varying-coefficient models, where the response variables could be nonnormal such as binary or Poisson, however, random effects are assumed to be normal
- Lin et al. (2007) studied marginal GLMs with varying-coefficients, where the estimation procedure was based on the kernel smoothing method of local linear fitting.

## Smoothing Techniques


- Two major nonparametric methods: *kernel smoothing* and *spline smoothing*.
- Kernel smoothing is a method of local weighted average. Consider a simple nonparametric model with only one covariate  $x_i$ ,

$$y_i = \theta(x_i) + \varepsilon_i, \quad \varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2), \quad (1)$$

where  $\theta(\cdot)$  is a fully unspecified smooth function to be estimated.

- The objective is to estimate  $\theta(x)$  at an arbitrary value  $x$ . The simple kernel estimator of  $\theta(x)$ , known as the Nadaraya-Watson estimator, is given by

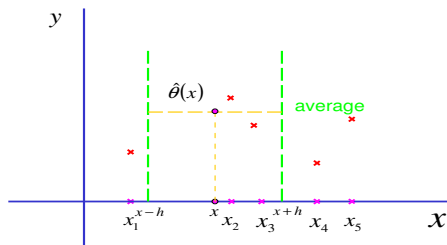
$$\hat{\theta}(x) = \frac{\sum_{i=1}^n \mathcal{K}\left(\frac{x_i - x}{h}\right) y_i}{\sum_{i=1}^n \mathcal{K}\left(\frac{x_i - x}{h}\right)} \quad (2)$$

where  $\mathcal{K}(\cdot)$  is a given kernel function and  $h$  is a bandwidth. 



## Example: Uniform Kernel Smoothing

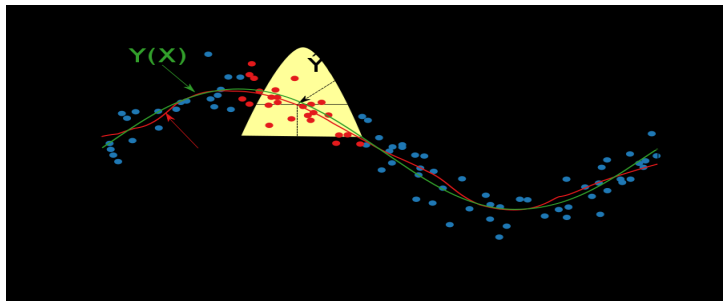
- Uniform kernel  $\mathcal{K}(u) = \frac{1}{2}I[-1 < u < 1]$ .
- The resulting local average kernel estimator is the running mean estimator.



$$\hat{\theta}(x) = \frac{\frac{y_2}{2h} + \frac{y_3}{2h}}{\frac{1}{2h} + \frac{1}{2h}} = \frac{y_2 + y_3}{2}$$

## Example: Normal Kernel Smoothing

- Normal kernel  $\mathcal{K}$  is the density of  $N(0, 1)$ , so  $\mathcal{K}_h(\cdot)$  is the density of  $N(0, h^2)$ .
- Involves all data points in the estimation, with different weights.

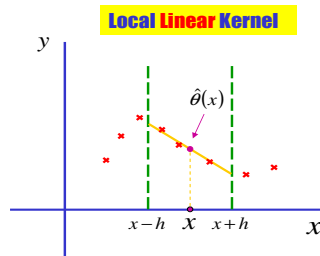
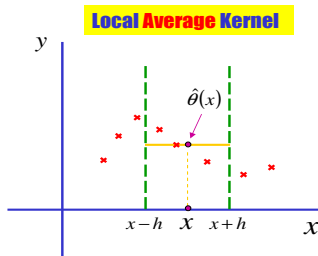


# Kernel Smoothing

- The choice of bandwidth  $h$  is important as it determines the smoothness and bias of the estimated curve.
- The choice of the kernel function  $\mathcal{K}(\cdot)$  is less important.
- R functions for kernel smoothing method
  - `ksmooth()` – it requires to pre-specify bandwidth  $h$
  - `loess.smooth()` – it requires to pre-specify bandwidth  $h$
  - `supsmu()` – it automatically specifies bandwidth  $h$  by cross-validation

# Local Polynomial Kernel Regression

Extend Nadaraya-Watson's local average estimator to a local polynomial estimator. For example, the local linear fitting.



# Local Polynomial Kernel Regression

- The N-W estimator is obtained by solving the following estimating equation:

$$\sum_{i=1}^n \mathcal{K}_h(x_i - x)(y_i - \alpha_0) = 0.$$

- Equivalently, the maximizer of the following objective function

$$U(\alpha_0) = -\frac{1}{2} \sum_{i=1}^n \mathcal{K}_h(x_i - x)(y_i - \alpha_0)^2.$$

- Equivalently,

$$\operatorname{argmax}_{\alpha_0} \sum_{i=1}^n \mathcal{K}_h(x_i - x) \ell(y_i; \alpha_0),$$

where  $\ell(y_i; \alpha_0)$  is the normal likelihood based on  $N(\alpha_0, \sigma^2)$ .

# Local Polynomial Kernel Regression

- In general, consider maximizing the following objective function with respect to a vector parameter  $\alpha = (\alpha_0, \dots, \alpha_p)'$ :

$$U(\alpha) = -\frac{1}{2} \sum_{i=1}^n \mathcal{K}_h(x_i - x) [y_i - \alpha_0 - \alpha_1(x_i - x) - \dots - \alpha_p(x_i - x)^p]^2,$$

where  $x$  is an arbitrarily fixed target value.

- The desired estimator  $\hat{\theta}(x) = \hat{\alpha}_0$ .

# Local Polynomial Kernel Regression

- It is a kind of weighted LS estimation where the normal equation is

$$X(x)'K(x)[Y - X(x)\alpha] = 0.$$

- The LS estimator is

$$\hat{\alpha} = [X(x)'K(x)X(x)]^{-1}X(x)'K(x)Y$$

where  $K(x) = \text{diag}[\mathcal{K}_h(x_1 - x), \dots, \mathcal{K}_h(x_n - x)]$ .

- Easy to implement using weighted LS numerical recipe.
- Easy to extend the procedure for nonnormal data.

## R Kernel Regression Functions

- (i) `ksmooth(y, x, kernel = "normal", bandwidth = 5)` calculates the average kernel estimator ( $p = 0$ )  $\hat{\theta}(x)$  with both pre-specified kernel function and bandwidth.
- (ii) `loess(y ~ x, span = s, degree = p)` or `loess.smooth(x, y, span = s, degree = p)` calculates loess (Local linear Regression) smoother with both pre-specified span and degree of polynomial, in which (a) span has to be in  $[0, 1]$  that represents the percent of data used in estimation at a given  $x$  and (b) degree  $p = 0$  corresponds to the local average estimation and  $p = 1$  corresponds to the local linear estimation.
- (iii) `supsmu(x, y, span = "cv")` calculates loess smoother by choosing an optimal span using CV (cross-validation).



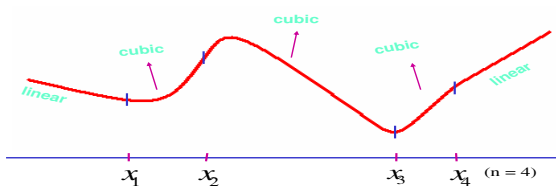
# Spline Smoothing

What is a *spline*? A spline is specified by two elements:

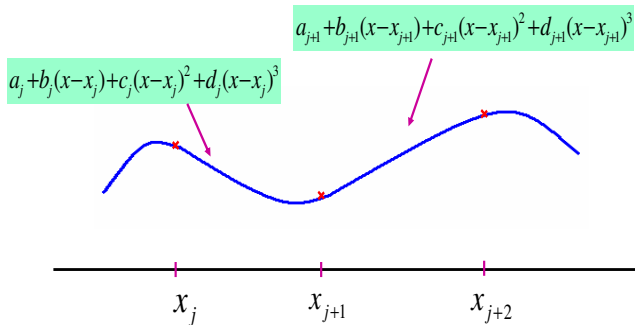
- (1)  $m$  knots, denoted by  $x_j, j = 1, \dots, m$ , and
- (2) a curve is piece-wise polynomial which is sufficiently smoothed at the given knots.

Among many types of splines, the *cubic spline*  $s(x)$  is of most interest. A cubic spline  $s(x)$  has the following properties:

- it has  $m$  knots  $x_j, j = 1, \dots, m$ ;
- a cubic polynomial on interval  $(x_j, x_{j+1})$  is imposed;
- at each knot  $x_j$ , the first derivatives  $\dot{s}(x)$  and  $\ddot{s}(x)$  are continuous;
- the third derivative  $s^{(3)}(x)$  is a step function with jumps at knots  $x_j$ 's.



# Constructing A Cubic Spline



How many free coefficients of a cubic spline are to be estimated using a given data?

$$\begin{aligned} m \text{ knots} &\leftrightarrow 4(m + 1) \text{ cubic polynomial coefficients} \\ &\quad - 3m \text{ constraints} \\ &= m + 4 \text{ free coefficients or parameters to be estimated} \end{aligned}$$

e.g., with 2 knots, there should be 6 free coefficients to be estimated.

# Cubic Spline Basis Function

**Plus Function Basis Function:** Given knots  $\{u_1, \dots, u_m\}$ ,

$$1, x, x^2, x^3, (x - u_1)_+^3, (x - u_2)_+^3, \dots, (x - u_m)_+^3$$

where

$$(x - a)_+^3 = \begin{cases} (x - a)^3, & \text{if } x > a \\ 0, & \text{if } x \leq a. \end{cases}$$

Then the cubic spline takes the form:

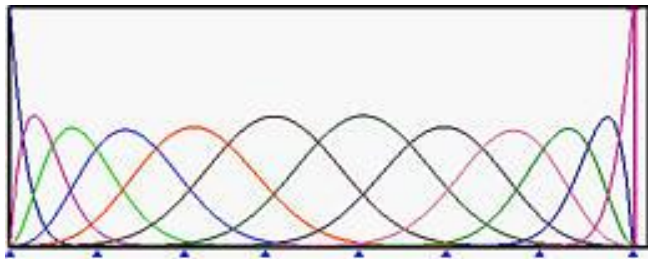
$$s(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - u_1)_+^3 + \dots + \beta_{m+4} (x - u_m)_+^3.$$

## B Spline Basis function

- Overcome the issue that the plus function basis functions are unbounded

A cubic spline formed by the B-spline basis:

$$s(x) = \beta_1 B_1(x) + \cdots + \beta_{m+4} B_{m+4}(x).$$



# Kernel Regression Analysis of Longitudinal Data

- Longitudinal GAM model (Wild and Yee, 1996; Berhane and Tibshirani, 1998):

$$g(\mu_i(t)) = \sum_j \theta_j(x_{ij}(t));$$

- Lin and Carroll (2000) attempted to incorporate serial correlation into kernel estimating equation to estimate nonparametric function  $\theta_j(\cdot)$ .
- They reported a striking (and probably counter-intuitive) discovery that the most efficient kernel based GEE estimator is obtained under the independence working correlation.
- Welsh et al. (2002) refers this to as *Locality*.

# Kernel Regression Analysis of Longitudinal Data

- Chen and Jin (2005) pointed out that a mismatch of *local observations with global variance* may cause this locality phenomenon. Further, they proposed a new kernel GEE based on *local observations with local variance*.
- Wang (2003) indicated that this locality may be caused by the use of the traditional kernel.



# Kernel-based GEE

- Illustrate this method in the context of varying-coefficient models:

$$g\{\mu_i(t)\} = \sum_{j=1}^p x_{ij}(t)\beta_j(t)$$

- Consider local linear fitting:

$$\beta_j(t) \approx a_j + b_j(t - t_0) = (1, t - t_0)(a_j, b_j)', \quad j = 1, \dots, p,$$

for  $t$  in a neighborhood of  $t_0$ .

- The local linear predictor

$$\eta_i(t) \approx \sum_j [x_{ij}(t), (t - t_0)x_{ij}(t)] \begin{bmatrix} a_j \\ b_j \end{bmatrix} = x_i(t)' \{a + (t - t_0)b\}$$

where  $a = (a_1, \dots, a_p)'$  and  $b = (b_1, \dots, b_p)'$

- The local mean is  $\tilde{\mu}_i(t_0) = g^{-1}\{\tilde{\eta}_i(t; t_0)\}$ .

# Global Variance Kernel GEE (GVKGEE)

Lin and Carroll (2000) suggested:

$$\sum_{i=1}^N \ddot{\mu}_i'(t_0) \mathcal{K}_{ih}^{1/2}(t_0) V_i(t_0)^{-1} \mathcal{K}_{ih}^{1/2}(t_0) \{y_i - \tilde{\mu}_i(t_0)\} = 0,$$

$$\begin{aligned} V_i(t_0) &= A_i^{1/2}(t_0) R_i(\delta) A_i^{1/2}(t_0) \\ \mathcal{K}_{ih}(t_0) &= \text{diag} \{ \mathcal{K}_h(t_{i1} - t_0), \dots, \mathcal{K}_h(t_{in_i} - t_0) \} \end{aligned}$$

where  $\mathcal{K}_h(\cdot) = \mathcal{K}(\cdot/h)/h$ , and  $\mathcal{K}(\cdot)$  is a traditional kernel with bandwidth  $h = h_n > 0$ .

# Local Variance Kernel GEE (LVKGEE)

Chen and Jin (2005) suggested:

$$\sum_{i=1}^N \ddot{\mu}'_i(t_0) \mathcal{K}_{ih}^{1/2}(t_0) V_i^*(t_0)^{-1} \mathcal{K}_{ih}^{1/2}(t_0) \{y_i - \tilde{\mu}_i(t_0)\} = 0,$$

$$V_i^*(t_0) = A_i^{1/2}(t_0) G_i R_i(\delta) G_i A_i^{1/2}(t_0),$$

where

$$G_i = \text{diag} [1\{\mathcal{K}_h(t_{i1} - t_0) > 0\}, \dots, 1\{\mathcal{K}_h(t_{in_i} - t_0) > 0\}],$$

and  $1\{A\}$  is an indicator of set  $A$ .

## Remarks

- Both the GVKGEE and the LVKGEE can handle time-dependent covariates.
- The LVKGEE gives better estimation efficiency under a correct correlation structure, which is unfortunately impossible in practice.
- The GVKGEE is computationally simple, with no need of estimating nuisance parameters in the working correlation  $R_j$ .
- Bandwidth selection: empirical bias bandwidth selection (EBBS) first proposed by Ruppert (1997), and later extended by Lin and Carroll (2000) to the longitudinal data setting.

# Simulation Study

- To compare GVKGEE and LVKGEE via varying-coefficient log-linear models for longitudinal count data.
- VC log-linear model:

$$\log\{\mu(t, x)\} = \beta_0(t) + \beta_1(t)x_1(t) + \beta_2(t)x_2(t),$$

with the true underlying coefficient functions

$$\beta_0(t) = 0.4 \exp(2t - 1), \beta_1(t) = 3t(1 - t), \text{ and}$$

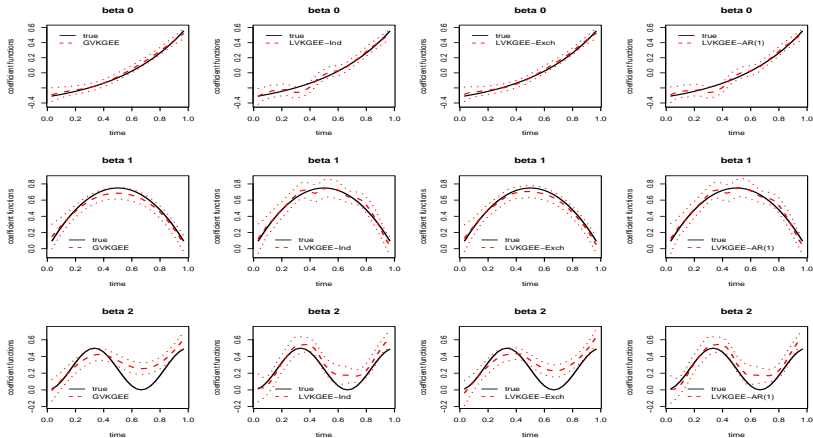
$$\beta_2(t) = 0.5 \sin^2(1.5\pi t).$$

- Curvature level increases from  $\beta_0$  to  $\beta_2$ , so it enables us to assess how the curvature affects the performance of these two methods.

## Simulation Study

- Irregular times: (1) Randomly sample  $n$  displacement points  $s_{i1} \sim U(0, 1)$ ; (2) created 30 equally spaced times by  $s_{ik} = s_{i1} + (k - 1)/30, k = 1, \dots, 30$ ; (3) Sample  $s_{ik}/30$  according to probability of selection 0.3 at each  $k$ . (4) Repeat this for each subject  $i$ .
- True correlation: AR-1.
- Used three cases of  $R_i$ : Independence, Exchangeable and AR-1.

# Simulation Study



# Simulation Study

Cumulative MSE (CMSE) under the optimal bandwidths:

Coefficient	GVKGEE	LVKGEE independent	LVKGEE exchangeable	LVKGEE AR(1)
$\beta_0$	.0340	.0424	.0423	.0419
$\beta_1$	.0847	.0988	.0988	.0985
$\beta_2$	.1142	.1114	.1092	.1092



## Remarks

- With low or medium curvature ( $\beta_0(t)$  and  $\beta_1(t)$ ), the GVKGEE outperformed all versions of the LVKGEE. In contrast, the LVKGEE estimates appear somewhat undersmoothed.
- In the case of high curvature, all versions of the LVKGEE performed clearly better than the GVKGEE. The GVKGEE seems to pay the price of bigger bias to gain higher efficiency.
- Another reason for the bias in the GVKGEE was oversmoothing, caused probably by a global bandwidth across all three coefficient functions.
- The LVKGEE performed the best under the true correlation structure, confirming the finding in Chen and Jin (2005). Correlation structure matters in the LVKGEE.

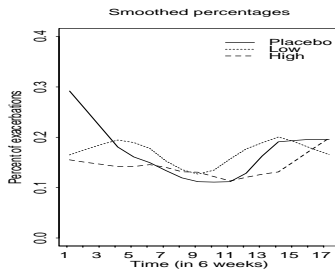
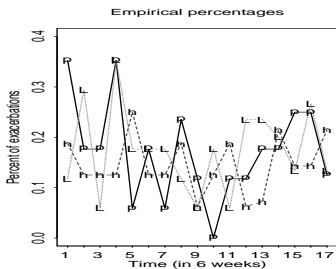
# Take-Home Message

**Curvature is a more important feature than either variance or correlation structure to determine performances of kernel-based GEE methods.**

## Application: Multiple Sclerosis Trial Data Analysis

- A longitudinal clinical trial to assess the effects of neutralizing antibodies on interferon beta-1 (IFNB) in relapsing-remitting multiple sclerosis (MS), a disease that destroys the myelin sheath surrounding the nerves (Petkau et al, 2004).
- Six-weekly frequent Magnetic Resonance Imaging (MRI) sub-study involving 52 patients, randomized into 3 treatment groups; 17 in placebo, 17 in low dose and 16 in high dose.
- At each of 17 scheduled visits, a binary outcome of *exacerbation* was recorded at the time of each MRI scan, according to whether an exacerbation began since the previous scan.
- Baseline covariates include age, duration of disease (in years), sex, and initial EDSS (expanded disability status scale) scores.
- Does the IFNB help to reduce the risk of exacerbation?

- Response  $Y_{ij}$  is binary, 1 for exacerbation and 0 otherwise.
- Covariates include treatment dosage  $\mathbf{trt}_i$ ,  $\mathbf{dur}_i$  baseline disease duration  $\mathbf{dur}_i$ , and two time variables  $\mathbf{t}_j$  and  $\mathbf{t}_j^2$ .



- The data was previously analyzed by Dyachkova et al. (1997) using GEE, where the effects of the covariates were all assumed to be constant.
- For illustration, consider the marginal logistic model

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \beta_0 + \beta_1 \text{trt}_i + \beta_2 t_j + \beta_3 t_j^2 + \beta_4 \text{dur}_i,$$

where the probability of exacerbation is

$$\pi_{ij} = \text{prob}(Y_{ij} = 1 | x_{ij}).$$

Par.	AR-1		Interchangeable	
	GEE	QIF	GEE	QIF
	Est(Std Err)	Est(Std Err)	Est(Std Err)	Est(Std Err)
intcpt	-.6793(.3490)	-.4955(.3443)	-.6847(.3502)	-.5419(.3169)
trt	-.0151(.1501)	-.0222(.1491)	-.0175(.1497)	-.0650(.1448)
time	-.0259(.0128)	-.0269(.0128)	-.0251(.0129)	-.0267(.0127)
time <sup>2</sup>	.0002(.0001)	.0002(.0001)	.0002(.0001)	.0002(.0001)
dur	-.0449(.0229)	-.0715(.0242)	-.0458(.0228)	-.0586(.0236)

- Treatment is not significant (even if more covariates are included).

- The plot of the empirical percentage of exacerbation against the time at all levels of the treatment had shown a very strong nonlinear relationship that could not simply be depicted by a polynomial function.
- Such a nonlinear pattern may be caused by the change of drug efficacy over time.
- Changing drug efficacy is a well-known phenomenon attributed to the development of drug resistance by human bodies.
- The central question was whether and how the risk of exacerbation varied in time as a function of the dose levels and the EDSS.

- Use a varying-coefficient logistic model to address the population-averaged relation between the probability of exacerbation and the time-varying effects of the covariates.
- The model takes the form

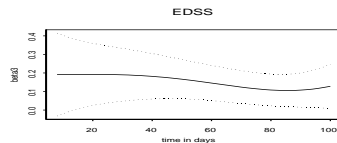
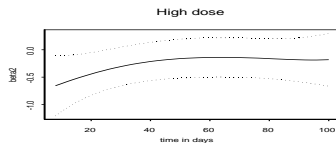
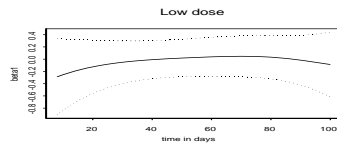
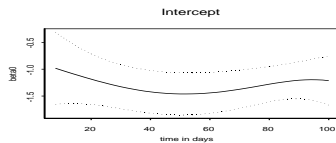
$$\log \frac{\pi_i(t)}{1 - \pi_i(t)} = \beta_0(t) + \beta_1(t) * x_{1i} + \beta_2(t) * x_{2i} + \beta_3(t) * x_{3i}(t)$$

where  $x_1 = 1$  if the treatment is low dose and otherwise 0, and  $x_2 = 1$  if the treatment is high dose and otherwise 0, and  $x_3$  is the score of EDSS.



- $\beta_0(t)$ : the adjusted time-varying effect of the treatment at placebo dose
- $\beta_1(t)$ : the adjusted difference of the treatment effect of low dose from placebo
- $\beta_2(t)$ : the adjusted difference of the treatment effect of high dose from placebo.

Fitted curves with 95% pointwise confidence bands.



- There exists a strong placebo effect.
- The low dose and placebo treatments have no statistically significant difference in reducing the risk of exacerbation.
- For the high dose, at the beginning of the trial (about first 20 days), there is some evidence that this dose lowered the odds of exacerbation than the placebo over this period.
- EDSS is an important factor relating to the risk of exacerbation. The effect of EDSS on the risk of exacerbation decreases gradually in the first 80 days and then rise up at the end.

## Varying coefficient models for longitudinal data

- The model takes both covariates and time effects into consideration (Hastie & Tibshirani, 1993; Cai, et al., 2000)
- Varying coefficient models: Hoover et al. (1998), Wu, Chiang, and Hoover (1998), Fan and Zhang (2000), Martinussen and Scheike (2001), Chiang, Rice, and Wu (2001), Huang, Wu, and Zhou (2002)
- These authors propose various estimation procedures for varying-coefficient models under longitudinal data settings, but did not discuss how to incorporate correlation information within subjects into their estimation procedures

# Varying coefficient models for longitudinal data (Qu & Li, 2006)

- Continuous outcome:

$$y_{ij} = X_i'(t_{ij})\beta(t_{ij}) + \varepsilon(t_{ij})$$

- Extend the varying coefficient model under generalized linear model settings:

$$E\{y_i(t_{ij})|X(t_{ij})\} = h\{X_i'(t_{ij})\beta(t_{ij})\} = \mu_{ij}(t_{ij})$$

where  $h(\cdot)$  is a known inverse link function

## How to model $\beta(t)$ ?

- Let  $B_{uv}(t)$  be basis functions and  $\gamma_{uv}$  be constants,  $u = 1, \dots, p$ ,  $v = 0, 1, \dots, V_u$ , where  $V_u + 1$  is the number of basis functions

$$\beta_u(t) \approx \sum_{v=0}^{V_u} \gamma_{uv} B_{uv}(t)$$

- For example, use  $q$ -degree polynomial basis and truncated power associated with knots

$$\beta_u(t) = \gamma_{u0} + \gamma_{u1}t + \dots + \gamma_{uq}t^q + \sum_{k=1}^{K_u} \gamma_{u(q+k)}(t - \kappa_k)_+^q$$

where  $K_u$  is the number of knots,  $\kappa_1 < \dots < \kappa_K$  are fixed knots and  $(z)_+^q = z^q I(z \geq 0)$ ,  $V_u = q + K_u$

# Penalized Spline

- References: Eilers & Marx (1996); Ruppert & Carroll (2000); Ruppert (2002); Yu & Ruppert (2002)
- Estimate  $\gamma = \{\gamma_{uv}, u = 1, \dots, p; v = 1, \dots, V_u\}$  by minimizing

$$N^{-1}Q + \lambda\gamma'D\gamma$$

where  $Q$  is the QIF,  $\lambda$  is a smoothing parameter, and  $D$  is a diagonal matrix with 1 if  $\gamma_{uv}$  is the coefficient of the truncated power function associated with knots and 0 otherwise

# Penalized Spline

- Penalize the model with too many knots
- Estimate smoothing parameter  $\lambda$  (Ruppert, 2002):

$$\hat{\lambda} = \arg \min_{\lambda} \frac{Q}{(N - \text{d.f.})^2},$$

where d.f. is the effective degrees of freedom of the fit

$$\text{d.f.} = \text{trace}[(\ddot{Q} + N\lambda_N D)^{-1} \ddot{Q}]$$

where  $\ddot{Q} = \partial^2 Q / \partial \beta^2$



## Asymptotic properties when $\lambda_N \rightarrow 0$

- **Result 1:** Under regularity conditions, if smoothing parameter  $\lambda_N = o(1)$ , then the spline regression parameter estimator  $\hat{\gamma}$  exists and converges to  $\gamma_0$  almost surely.
- **Result 2:** Under regularity conditions, if the smoothing parameter  $\lambda_N = o(N^{-1/2})$ , then the spline regression parameter estimator  $\hat{\gamma}$  is asymptotically normal and efficient.

That is

$$\sqrt{N}(\hat{\gamma} - \gamma_0) \xrightarrow{d} N_p(0, (J_0' C_0^{-1} J_0)^{-1}),$$

where  $J_0 = E(\partial g / \partial \beta_0)$ ,  $C_0 = E\{g(\beta_0)g'(\beta_0)\}$  and  $g$  is the estimating function based on one subject observation

# Pointwise confidence interval

- Varying coefficient  $\beta_u(t) = B_u(t)\gamma_u$  has the asymptotic distribution

$$\sqrt{N}\{\hat{\beta}_u(t) - \beta_u(t)\} \xrightarrow{d} N(0, B'_u(H^{-1}GH^{-1})_u B_u),$$

where  $(H^{-1}GH^{-1})_u$  is a sub-matrix of  $H^{-1}GH^{-1}$  associated with the variance of  $\gamma_u$

## Inferential properties of QIF

- To test whether coefficients change over time or not
- To test  $H_0$  against  $H_1$ , where  $H_0$  is nested under  $H_1$

$$H_0 : \gamma_{uv} = 0, v = 1, \dots, V_u$$

- To test whether coefficient is constant over time
- Test statistic

$$T = Q(\tilde{\gamma}) + N\lambda_N\tilde{\gamma}'D\tilde{\gamma} - Q(\hat{\gamma}) - N\lambda_N\hat{\gamma}'D\hat{\gamma},$$

where  $\tilde{\gamma}$  and  $\hat{\gamma}$  are estimators under  $H_0$  and  $H_1$  respectively

- **Result 3:** Under regularity conditions, if the smoothing parameter  $\lambda_N = o(N^{-1/2})$ , then the asymptotic distribution of  $T$  follows chi-squared with degrees of freedom equal to  $V_u$  under  $H_0$

# Asymptotic chi-squared tests

- In practice, we can set  $\lambda_N = 0$ , the test statistic  $Q(\tilde{\gamma}) - Q(\hat{\gamma})$  follows  $\chi_{V_u}^2$  asymptotically under  $H_0$ , where  $\tilde{\gamma}$  and  $\hat{\gamma}$  are estimators under  $H_0$  and  $H_1$  respectively
- **Goodness-of-fit test:** To test zero-mean assumption  
 $H_0 : E(g) = 0$   
 $Q(\tilde{\gamma}) \xrightarrow{d} \chi_{r-k}^2$  under  $H_0$  (Hansen, 1982), where  
 $k = \sum_{u=1}^p V_u + p$
- Goodness-of-fit test is useful to determine the number of knots to be selected

# Simulations

- Binary responses with the marginal distribution:

$$P(y_{ij} = 1 | t_{ij}) = \exp\{\beta(t_{ij})\} / [1 + \exp\{\beta(t_{ij})\}]$$

where  $i = 1, \dots, 200$  and  $j = 1, \dots, n_i$

- Consider 4 models:

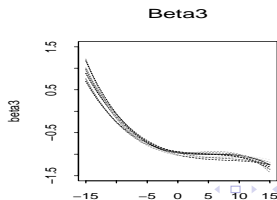
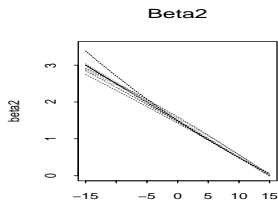
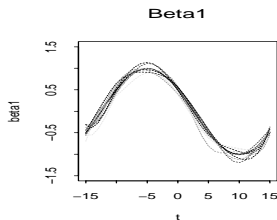
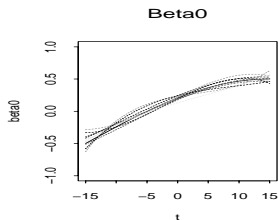
$$\begin{aligned} \beta_0(t) &= \sin\left(\frac{(t+15)\pi}{60}\right) - 0.5, & \beta_1(t) &= \cos\left\{\frac{(t-10)\pi}{15}\right\} \\ \beta_2(t) &= -0.1(t - 15), & \beta_3(t) &= \frac{(5-t)^3}{4000} - 1. \end{aligned}$$

- Fix centered time points  $\{-15, -14, \dots, 15\}$ , but except for the beginning time, the true time has 60% chance to be missing, also it varies between uniform  $(-0.5, 0.5)$  from the unskipped time
- Unbalanced data and different observed time

# Simulations

Fitted varying coefficient curves correspond to 9 deciles of mean absolute deviation of errors from 1000 simulations

$$(\text{MADE})_k = \sum_{j=0}^{30} 31^{-1} |\hat{\beta}_k(t_j) - \beta_k(t_j)| / \text{range}(\beta_k)$$



## Testing for varying coefficient

- Simulate data under  $H_0 : \beta_1(t) = 0.5$
- Basis function for  $\beta_1$  under  $H_0$  is 1
- Basis functions for  $\beta_1$  under  $H_1$  are  $1, t, t^2, t^3, (t + 10)_+^3, (t)_+^3,$  and  $(t - 10)_+^3$ .
- Under  $H_0$ ,  $Q(\tilde{\gamma}) - Q(\hat{\gamma}) \xrightarrow{d} \chi_6^2$ , where  $\tilde{\gamma}$  and  $\hat{\gamma}$  are estimators under  $H_0$  and  $H_1$  respectively

## Example from AIDs data

- 283 homosexual males who were HIV positive between 1984 and 1991 (Huang et al., 2002)
- Response variable: CD4 cell counts and percentages
- Each subject has minimum 1, maximum 14 measurements

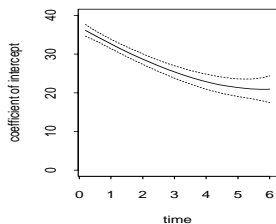
$$y_{ij} = \beta_0(t_{ij}) + \beta_1(t_{ij}) \text{ Smoke} + \beta_2(t_{ij}) \text{ Age} \\ + \beta_3(t_{ij}) \text{ Pre-CD4} + \varepsilon_{ij}$$

- Applying penalized spline and equally spaced knots
- Use goodness-of-fit test to choose 0, 5, 1, 3 knots for  $\beta_0, \beta_1, \beta_2, \beta_3$  respectively

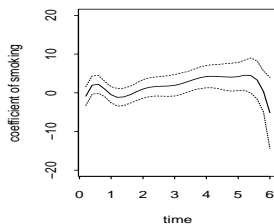


## AIDs data, varying coefficients graphs

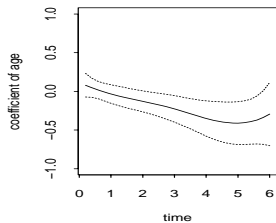
Intercept effect



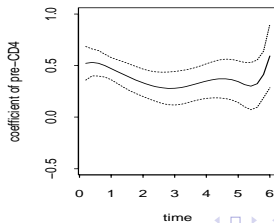
Smoking effect



Age effect



Pre-CD4 effect



## Testing varying coefficients for AIDs data

Null	Bootstrap (Huang et al., 2002)	QIF (Exchangeable)		
	$p$	T	d.f	$p$
Constant baseline	0.000	81.9	3	0.000
Smoking has no effect	0.176	13.0	9	0.163
Age has no effect	0.301	7.7	5	0.172
Constant Pre-CD4	<b>0.059</b>	12.9	6	<b>0.045*</b>

- Intercept coefficient changes over time significantly
- Overall smoking has no significant effect on CD4
- Age has no significant effect on CD4
- Pre-CD4 coefficient changes over time

## Summary for varying coefficient modeling

- The likelihood function is often unknown or difficult to formulate
- The QIF approach does not require estimation of nuisance parameters involved in working correlations
- This advantage becomes more important in nonparametric settings as there are many more parameters involved
- The inference function has an explicit asymptotic form, which allows us to test whether coefficients are time-varying or time invariant
- Provides goodness-of-fit tests for checking model assumptions
- Provides an objective criterion for choosing sufficient number of basis functions and determine how many knots are appropriate for the model

## Generalized partial linear models

- We would expect part of coefficients are fixed over time, and part of them are functions of time

$$g(\mu) = X'\beta + f(t)$$

- Lin and Carroll (2001): profile kernel method
- He et al. (2005): profile spline under GEE using robust scores, with optimal rate of convergence for estimating both  $\beta$  and  $f(\cdot)$
- Li and Nie (2007, 2008): propose partially nonlinear model via a mixed-effects approach

## Generalized partial linear models

- Bai et al. (2008): partial linear models based on the QIF, approximate  $f(t)$  by  $\pi'(t)\alpha$  given  $\beta$

$$g(\mu_{ij}(\theta)) = X'_{ij}\beta + \pi'(t_{ij})\alpha = Z'_{ij}\theta,$$

where  $Z'_{ij} = (X'_{ij}, \pi'(t_{ij}))$  and  $\theta' = (\beta', \alpha')$

- Obtain  $\hat{\theta}$  by solving the QIF using  $\mu_{ij}$  as the mean function
- Simulations in Bai et al. show that the regression spline-based QIF performs better than the profile-kernel method in Lin and Carroll (2001)

# Model selection for covariates: Introduction

- Model selection for longitudinal data is challenging, the full likelihood for longitudinal data is often difficult to specify, particularly for correlated non-Gaussian data
- Introduce BIC-type model selection criterion based on the quadratic inference function (Qu, Lindsay and Li, 2000)
- Do not require the full likelihood or quasiliikelihood
- Consistent property: selects the most parsimonious correct model with probability approaching one

# Model Selection Literature

- Quasilikelihood Information Criterion (QIC) (Pan, 2001)  
extends AIC for independent data case to GEE using quasi-likelihood
- Cantoni, Flemming and Ronchetti (2005, Biometrics)  
A generalized version of Mallows's  $C_p$
- Suitable for situations where the correlation is fairly weak

## Challenges and Motivations

- Schwarz' BIC (1978) selects the model that maximizes  $l_p - \frac{1}{2}p(\log N)$  (or equivalently minimizes  $-2l_p + p(\log N)$ ), where  $l_p$  is the log likelihood
- Can we replace the full likelihood by an alternative inference function?
- QIF plays a role similar to as that of  $-2\log(\text{likelihood})$  in the parametric setting
  - Minimizing the QIF is analogous to maximizing the likelihood.
  - $Q_N(\beta_0) - Q_N(\hat{\beta})$  is an asymptotically chi-squared test for testing  $\beta = \beta_0$ , analogous to LRT



## Some Notations

- Let  $\mathcal{M}$  be the class of candidate models
- Each member of  $\mathcal{M}$  can be identified with a unique set  $m$ , where  $m$  is a subset of  $\{1, \dots, q\}$  and contains the indices of the covariates that are included in that candidate model
- Example: consider fitting three possible covariates  $x_1, x_2, x_3$ , the candidate model that contains only  $x_1$  and  $x_3$  will be indexed by  $m = \{1, 3\}$ , and  $\beta(m) = (\beta_0, \beta_1, 0, \beta_3)^T$

# BIQIF Model Selection Criterion

## Model Selection Criterion

The QIF based BIC selects the model in  $\mathcal{M}$  which minimizes:

$$BIQIF(m) = Q_N(\hat{\beta}(m)) + |\beta(m)| \log(N),$$

where  $Q_N(\hat{\beta}(m)) = \inf_{\beta \in \mathcal{B}(m)} Q_N(\beta)$ , and  $|\beta(m)|$  denotes the number of nonzero elements in  $\beta(m)$

# Consistency Property

- Let  $\mathcal{M}^c$  be the subset of  $\mathcal{M}$  that contains the correct models, i.e.,

$$\mathcal{M}^c = \{m \in \mathcal{M} : g(E(y_{ij}|x_{ij})) = x'_{ij}\beta(m), \text{ for some } \beta(m) \in \mathcal{B}(m)\}$$

- The class of most parsimonious correct models, defined as

$$\mathcal{PM}^c = \{m \in \mathcal{M}^c : |\beta(m)| \leq |\beta(m^*)|, \forall m^* \in \mathcal{M}^c\}$$

- **Consistency Property:** Denote the model selected by *BIQIF* from  $\mathcal{M}$  by  $\tilde{m}$ . Under some regularity conditions, when  $N \rightarrow \infty$

$$P(\tilde{m} \in \mathcal{PM}^c) \rightarrow 1$$

# Simulations

- We compare with several alternative methods:
- The AIC and BIC procedure based on the full likelihood
- Z-test procedure
- The QIC procedure of Pan (2001)
- The continuous response:

$$Y_{ij} = \beta_0 + \beta_1 x_{1,ij} + \beta_2 x_{2,ij} + \beta_3 x_{3,ij} + \epsilon_{ij}, \quad i = 1, \dots, N; j = 1, \dots, 4,$$

- The true correlation is AR(1)
- The true model has  $\beta_0 = 0.2$ ,  $\beta_1 = \beta_2 = 1$  and  $\beta_3 = 0$

**Table:** The proportion of times the true model is selected out of 500 simulation runs. BIQIF1 is BIQIF using the AR(1) working correlation, BIQIF2 is BIQIF using the CS working correlation, AIC and BIC are based on the full likelihood with known covariance matrix, the Z-test uses the true covariance structure, QIC1 is the QIC procedure using the AR(1) working correlation matrix and QIC2 is the QIC procedure using the CS working correlation matrix.

$N$	$\alpha$	BIQIF <sup>1</sup>	BIQIF <sup>2</sup>	AIC	BIC	Z test	QIC <sup>1</sup>	QIC <sup>2</sup>
40	0.3	0.854	0.886	0.844	0.944	0.504	0.836	0.842
	0.5	0.874	0.884	0.852	0.948	0.350	0.886	0.856
	0.7	0.868	0.870	0.842	0.948	0.326	0.906	0.890
80	0.3	0.924	0.944	0.804	0.936	0.732	0.826	0.838
	0.5	0.944	0.936	0.856	0.962	0.644	0.870	0.846
	0.7	0.958	0.966	0.844	0.972	0.574	0.912	0.896
120	0.3	0.968	0.950	0.844	0.964	0.870	0.850	0.850
	0.5	0.976	0.958	0.870	0.978	0.816	0.908	0.894
	0.7	0.970	0.968	0.866	0.964	0.728	0.922	0.904

## Working Correlation Selection in the GEE Method

- Although both GEE and QIF produce consistent estimates of the regression coefficients under misspecified correlation, the working correlation structure affects the efficiency of the estimation – the closer the working correlation structure is to the true, the more efficient (or the more powerful) the estimation is.
- The selection of working correlation is a secondary task in comparison to the selection of mean model, but it is still practically important.

## Working Correlation Selection in the GEE Method

- In the GEE setting, QIC, as an analog to the AIC proposed by Pan (2001), is used in practice to select a working correlation structure among several candidates.
- QIC tends to favor the independence working structure, because the QL is formed under the working independence structure and hence utilizes little information about correlation.
- Hin and Wang (2009) suggested CIC (Correlation Information Criterion), that uses the penalty term in the QIC as a criterion for the selection of working correlation structure.

# CIC in the GEE Method

- The CIC is defined as

$$CIC(R_\alpha) = \text{tr} \left\{ \widehat{\Omega}_I \widehat{V}_\alpha \right\},$$

$$\widehat{\Omega}_I = -\widehat{S}_I(\widehat{\beta}_{R_\alpha}, \widehat{\sigma}_{R_\alpha}^2)$$

$$\widehat{V}_\alpha = \widehat{J}_{R_\alpha}^{-1}(\widehat{\beta}_{R_\alpha}, \widehat{\sigma}_{R_\alpha}^2)$$

- The optimal working correlation structure is the one with the minimum CIC, that is,

$$R_{opt} = \arg \min_{R_\alpha \in \mathfrak{R}} \{CIC(R_\alpha), \alpha \in \Gamma\}.$$



# Simulation Study: Normal Longitudinal Data

- Balanced longitudinal normal data of size  $N = 30$  with  $n = 5$  repeated measurements for each cluster.
- The mean model  $\mu_{ij} = 3 + 5x_{ij}$  and variance  $\sigma^2 = 1$ .
- Time-dependent covariates  $x_{ij}$  is generated from  $U(j, j + 1)$ ,
- The true correlation structure is AR-1.
- 1000 rounds of simulation.

# Simulation Study: Normal Longitudinal Data

The empirical frequencies of selecting each of the independence (IND), exchangeable (EXCH) and AR-1 correlation structures.

	$\alpha = 0.1$			$\alpha = 0.5$			$\alpha = 0.9$		
	IND	EXCH	AR-1	IND	EXCH	AR(1)	IND	EXCH	AR-1
sample size $K = 20$									
<i>QIC</i>	211	282	<b>507</b>	199	168	<b>633</b>	198	215	<b>587</b>
<i>CIC</i>	191	273	<b>536</b>	112	129	<b>759</b>	130	172	<b>698</b>
sample size $K = 100$									
<i>QIC</i>	206	243	<b>551</b>	142	103	<b>755</b>	110	191	<b>699</b>
<i>CIC</i>	170	215	<b>615</b>	35	29	<b>936</b>	17	97	<b>886</b>

The positive selection rate drops because the AR-1 structure becomes more similar to an exchangeable with  $\alpha = 0.9$  than that with  $\alpha = 0.5$ .

# Simulation Study: Binary Longitudinal Data

- Balanced longitudinal binary data of size  $N = 30$  with  $n = 5$  repeated measurements for each cluster.
- The mean logistic model  $\text{logit}(\mu_{ij}) = -1 + \frac{1}{6}x_{ij}$ .
- Time-dependent covariates  $x_{ij}$  is generated from  $U(j, j + 1)$ ,
- The true correlation structure is AR-1.
- 1000 rounds of simulation.

# Simulation Study: Binary Longitudinal Data

The empirical frequencies of selecting each of the independence (IND), exchangeable (EXCH) and AR-1 correlation structures.

	True: AR-1								
	$\alpha = 0.2$			$\alpha = 0.5$			$\alpha = 0.7$		
	IND	EXCH	AR-1	IND	EXCH	AR(1)	IND	EXCH	AR-1
<i>QIC</i>	130	207	<b>663</b>	76	108	<b>816</b>	50	98	<b>852</b>
<i>CIC</i>	97	167	<b>736</b>	30	74	<b>896</b>	14	42	<b>944</b>

## Working Correlation Selection in the QIF Method

- In the QIF setting, both AIC and BIC (Wang and Qu, 2009) can be defined by treating the QIF objective function as being similar to  $-2 \log L$ .
- Although BIC works well for the selection of regression parameters in the mean model, it performs badly to discern correlation structure.
- Song et al. (2009) suggested TGI (Trace of Godambe Information) that takes the trace of Godambe information matrix.

## Working Correlation Selection in the QIF Method

- Suppose  $J$  is the Godambe information matrix, so  $J^{-1}$  is the asymptotic covariance matrix of the QIF estimator  $\hat{\beta}$ .

$$TGI = tr(J)$$

- The "optimal" correlation matrix is the one that has the maximum TGI.

# Simulation Study: Normal Longitudinal Data

- Balanced longitudinal normal data of size  $N = 50$  with  $n = 4$  repeated measurements for each cluster.
- The mean model  $\mu_{ij} = \beta_0 + \beta_1 E_{ij} + \beta_2 d_i$ .
- Time-dependent  $E_{ij} \sim U(0, 1)$ , and  $d_i \sim \text{Bernoulli}(0.5)$ .
- 1000 rounds of simulation.

## Simulation Study

The percent of the true correlation structure being selected among 1000 simulations by the TGI and BIC criteria based on the QIF method.

Model	True Corr	$\alpha$	Selection%	
			TGI	BIC
Normal	Exch	0.3	58.9	82.8
		0.7	84.9	85.1
	AR-1	0.3	84.9	10.0
		0.7	82.5	11.2
Binomial	Exch	0.3	82.7	37.3
		0.7	92.4	39.5
	AR-1	0.3	79.7	33.9
		0.7	91.3	32.2



# SAS PROC GENMOD

- Performs the GEE for the regression coefficients  $\beta$ , in which the nuisance parameters (including the correlation and dispersion/scale parameters) are separately estimated.
- Handles data types such as normal, binomial, Poisson, and gamma.
- Use “all available data” in estimation under the assumption that missing data are MCAR.

- Take the example of Multiple Sclerosis Trial Data.
  - Response  $Y_{ij}$  is binary, 1 for exacerbation and 0 otherwise.
  - Covariates include treatment dosage  $\mathbf{trt}_i$ ,  $\mathbf{dur}_i$  baseline disease duration  $\mathbf{dur}_i$ , and two time variables  $\mathbf{t}_j$  and  $\mathbf{t}_j^2$ .
  - The marginal logistic model

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \beta_0 + \beta_1 \mathbf{trt}_i + \beta_2 \mathbf{t}_j + \beta_3 \mathbf{t}_j^2 + \beta_4 \mathbf{dur}_i,$$

where the probability of exacerbation is

$$\pi_{ij} = \text{prob}(Y_{ij} = 1 | x_{ij}).$$

# GEE in SAS

SAS code of GEE with unstructured working correlation:

```
title "UNSTRUCTURED CORRELATION";  
::::::::::(DATA IMPORT)::::::::::  
proc genmod data=exacerb;  
class id;  
model rel= trt t1 t2 dur / dist=bin link=logit;  
repeated subject=id / type=un corrw covb modelse;  
run;
```

- SAS code of GEE with interchangeable working correlation:

```
title "INTERCHANGEABLE CORRELATION (type=cs)";  
:::::(DATA IMPORT):::;  
proc genmod data=exacerb;  
class id;  
model rel= dose t1 t2 dur / dist=bin link=logit;  
repeated subject=id / type=exch corrw covb modelse;  
run;
```

# GEE in SAS

- SAS code of GEE with AR-1 working correlation:

```
title "AR-1 CORRELATION";  
::::::::::(DATA IMPORT)::::::::::  
proc genmod data=exacerb;  
class id;  
model rel= dose t1 t2 dur / dist=bin link=logit;  
repeated subject=id / type=ar corrw covb modelse;  
run;
```

# SAS MICRO QIF

- An alpha-test version of a SAS MARCO QIF (Song and Jiang, 2006), including a users' manual, is available for a secured download at the webpage:

**[www.stats.uwaterloo.ca/~song](http://www.stats.uwaterloo.ca/~song)**

- MACRO QIF works for several widely used marginal models:

Distribution	Canonical link function
Normal	Identity $g(\mu) = \mu$
Poisson	Log $g(\mu) = \log(\mu)$
Binary	Logit $g(\mu) = \log\{\mu/(1 - \mu)\}$
Gamma	Reciprocal $g(\mu) = 1/\mu$

- MACRO QIF accommodates popular working correlation structures: independence, unstructured, AR-1, and interchangeable.

# SAS MACRO QIF

- MACRO QIF outputs:
  - estimates of the model parameters
  - asymptotic covariance matrix
  - standard errors
  - $\chi^2$  statistic for goodness-of-fit test
  - model selection criteria AIC and BIC.

$$\text{AIC} = Q(\hat{\beta}) + 2(k - 1) \times \dim(\beta)$$

$$\text{BIC} = Q(\hat{\beta}) + \ln(N)(k - 1) \times \dim(\beta)$$

- MARCO QIF  $\equiv$  RPOC GENMOD for the working independence correlation.
- MARCO QIF implements listwise deletion (all available data), as in PROC GENMOD.
- MARCO QIF was coded in SAS version 9.1.3.

# SAS MACRO QIF

- Example: the marginal logistic model for Multiple Sclerosis Trial Data

```
\%qif(data=exacerb,  
      yvar=rel, xvar=dose dur t1 t2, id=id,  
      dist=bin, corr=exch, print=Y, outpar=par2,  
      outqif=qif2, outcov=cov2,outres=binres);  
run;
```



## R Package QIF

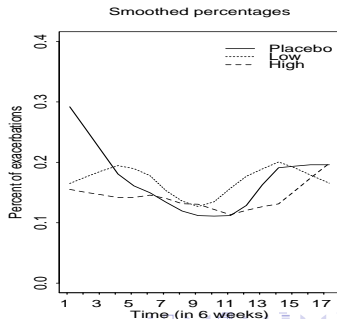
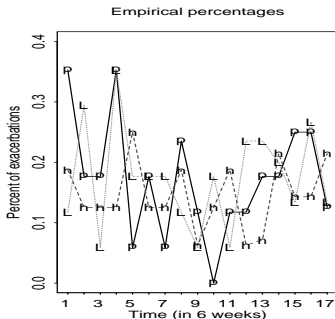
We also developed an R Package QIF using the source code of the SAS MACRO QIF.

- Download the package from the following webpage:  
`http://www-personal.umich.edu/~pxsong/qif\_package`
- Local load into R
- Need some further tests, and comments are welcome!

# Analysis of Multiple Sclerosis Trial Data

Marginal Logistic Model:

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 \text{trt}_i + \beta_2 t_j + \beta_3 t_j^2 + \beta_4 \text{dur}_i.$$



Par.	AR-1		Interchangeable	
	GEE	QIF	GEE	QIF
	Est(Std Err)	Est(Std Err)	Est(Std Err)	Est(Std Err)
intcpt	-.6793(.3490)	-.4955(.3443)	-.6847(.3502)	-.5419(.3169)
trt	-.0151(.1501)	-.0222(.1491)	-.0175(.1497)	-.0650(.1448)
time	-.0259(.0128)	-.0269(.0128)	-.0251(.0129)	-.0267(.0127)
time <sup>2</sup>	.0002(.0001)	.0002(.0001)	.0002(.0001)	.0002(.0001)
dur	-.0449(.0229)	-.0715(.0242)	-.0458(.0228)	-.0586(.0236)

- No evidence that the population-average effect of the drug treatment is significant in reducing the risk of exacerbation.
- The baseline disease severity is an important explanatory variable for the risk of exacerbation.
- Both linear and quadratic time covariates are significant, due partly to the periodic behavior of disease recurrences.

- MACRO QIF reports both goodness-of-fit statistic and AIC/BIC model selection criteria

Statistic	AR-1	Interchangeable
Q	4.3	2.5
df	5.0	5.0
AIC	14.3	12.5
BIC	23.3	21.5

- The  $p$ -value of the  $Q$  statistic, based on  $\chi_5^2$  distribution, is 0.507 under AR-1 working correlation and is 0.776 under Interchangeable working correlation.
- Thus, the marginal logistic model is appropriately specified.

# Install R Package QIF

Follow the following 4 steps:

- Step 1: Start the R software (version 2.9.0 or newer).
- Step 2: Click on the tab “packages” from the menu bar.
- Step 3: Click on “Install package(s) from local zip files ...”.
- Step 4: Find the downloaded qif package from the opened dialogue window in step 3, and open the downloaded qif package zip file. R should then automatically install it, and the qif package is ready to be loaded after the installation is finished.
- To read user’s manual, type “help(qif)” in R.

## Run R Package QIF

- To load the R package, simply type “library(qif)” into the R command window, the qif package is ready for use after this step.
- The qif package works for several types of links: identity, log and logit; and it accommodates popular covariance structures such as independence, AR-1, compound symmetry and unstructured.
- The qif function outputs: estimates of the model parameters; asymptotic covariance matrix; standard errors and p-values for coefficients; model selection criteria AIC and BIC; number of iterations it takes for algorithm to converge; fitted values as well as residuals.
- The current qif only supports equal cluster sized and equally spaced data type.

## Examples: Marginal logistic model for Multiple Sclerosis Trial data

```
Out.ind<-qif(exacerbation ~ edss + treatment + time + time2
            + duration, id=id, data=exacerb,
            family=binomial, corstr="independence")
Out.ar1<-qif(exacerbation ~ treatment + time + time2
            + duration, id=id, data=exacerb,
            family=binomial, corstr="AR-1")
Out.cs <- qif(exacerbation ~ treatment + time + time2
            + duration, id=id, data=exacerb,
            family=binomial, corstr="exchangeable")
Out.un<-qif(exacerbation ~ treatment + time + time2
            + duration, id=id, data=exacerb,
            family=binomial, corstr="unstructured")
```

To see the full list of output options, use “names(Out.un)”, for example.