University of Louisville

ThinkIR: The University of Louisville's Institutional Repository

Electronic Theses and Dissertations

5-2015

Novel applications of and extensions to linear regression methods for the biomedical and materials sciences.

Joe Bible
University of Louisville

Follow this and additional works at: <https://ir.library.louisville.edu/etd>

Part of the [Biostatistics Commons](#)

Recommended Citation

Bible, Joe, "Novel applications of and extensions to linear regression methods for the biomedical and materials sciences." (2015). *Electronic Theses and Dissertations*. Paper 2030.
<https://doi.org/10.18297/etd/2030>

This Doctoral Dissertation is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. This title appears here courtesy of the author, who has retained all other copyrights. For more information, please contact thinkir@louisville.edu.

**NOVEL APPLICATIONS OF AND EXTENSIONS TO
LINEAR REGRESSION METHODS FOR THE
BIOMEDICAL AND MATERIALS SCIENCES**

By

Joe Bible
M.S., Kennesaw State University, 2011

A Dissertation
Submitted to the Faculty of the
School of Public Health and Information
Sciences of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy
in
Biostatistics: Decision Science

Department of Biostatistics and Bioinformatics
University of Louisville
Louisville, Kentucky

May 2015

**NOVEL APPLICATIONS OF AND EXTENSIONS TO
LINEAR REGRESSION METHODS FOR THE
BIOMEDICAL AND MATERIALS SCIENCES**

By

Joe Bible
M.S., Kennesaw State University, 2011

A Dissertation Approved On

April 3rd 2015

Date

by the following Dissertation Committee:

Somnath Datta, PhD
Dissertation Director

Susmita Datta, PhD

Mahendra Sunkara, PhD

KB Kulasekera, PhD

ACKNOWLEDGEMENTS

I thank Dr. Somnath Datta for his tireless efforts these past few years. I also thank my dissertation committee members Drs. Susmita Datta, Mahendra Sunkara and KB Kulasekera for their patience, continuous support, confidence and direction.

I thank the faculty and students of the Department of Biostatistics and Bioinformatics as well as the faculty and students at the Conn Center for Renewable Energy Research for their support, direction and collaboration.

Finally, I need to thank Laurel and Josh. These past few years were trying, to say the least. Thank you for being the things I could consistently look forward to.

ABSTRACT

NOVEL APPLICATIONS OF AND EXTENSIONS TO LINEAR REGRESSION METHODS FOR THE BIOMEDICAL AND MATERIALS SCIENCES

Joe Bible

April 3rd 2014

In this work we present three topics, each of which centered on either the application or modification of various linear regression methods. Our work with respect to the “Materials Genome” project while undermined by oversimplification and data integrity issues in its early stages, provides a sound platform from which the project can proceed successfully. Building upon a growing body of knowledge around the use of Weighted Generalized Estimating Equations (WGEE), our second investigation proposes an extension to that framework intended to address the inherent bias present in the analysis of clustered longitudinal data with potentially informative cluster sizes and temporal observation profiles. Having demonstrated the utility of our marginal WGEE’s with respect to mitigating induced bias our final investigation presents a comparison of our marginal WGEE’s to model estimation via Joint Likelihood maximization in certain simulation models. We find, as would be in line with expectation, comparable performance with a loss of efficiency in the marginal WGEE setting.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
LIST OF TABLES	viii
LIST OF FIGURES	x

CHAPTER

I. MATERIALS SCIENCE APPLICATION OF HIGH DIMENSIONAL REGRESSION METHODS: BAND GAP ESTIMATION	1
A. Preliminary Investigation	2
1. Methods	3
2. Results	7
3. Discussion	12
B. Solid Solutions Problem	13
1. Approach to Modeling Band Gaps of Solid Solutions . .	15
2. Data	16
3. Modeling - Binary Compounds	18
4. Defining the Mixture Model	20
C. Conclusion	26
1. Be wary of over simplification	26
2. On the acquisition of data	26
3. Brief Discussion of Applicability and Implementation of Regression Techniques in Chem-Informatic Settings . . .	27

II. CLUSTER ADJUSTED REGRESSION FOR DISPLACED SUBJECT DATA (CARDS): MARGINAL INFERENCE UNDER POTENTIALLY INFORMATIVE TEMPORAL CLUSTER SIZE PROFILES 29

A.	Introduction	29
B.	Informative Cluster Size	31
1.	Informative Cluster Size	31
2.	Temporally Varying Cluster Size	33
C.	The Case for Scientifically Motivated Marginalization and Sta- tistical Inference	34
1.	Weighted Generalized Estimating Equations	36
2.	Variance of WGEE's	40
D.	Simulation	41
1.	Simulation Results	46
E.	Periodontal Application	47
F.	Discussion	56

III. PERFORMANCE OF WEIGHTED GENERALIZED ESTI- MATING EQUATIONS WITH RESPECT TO MARGINAL MODEL ESTIMATION COMPARISON WITH JOINT MODELING 61

A.	Introduction	61
B.	Joint Modeling - Full Joint Likelihood	62
1.	Design	63
2.	Likelihood	64
3.	Likelihood Estimation and Maximization	64
C.	GEE and Marginal WGEE	65
D.	Variance Component Estimation	66
E.	Simulation	68
1.	Simulation Results	68

2.	On Numeric Estimation of Joint Model Parameters . . .	78
F.	Discussion	88
IV. Future Work		91
A.	Materials Science - Band gap estimation	91
B.	MWGEE's	92
C.	Joint Modeling	92
REFERENCES		94
A.	An R function for estimating MWGEE models	98
B.	MWGEE estimation using the <i>genmod</i> procedure	104
CURRICULUM VITAE		112

LIST OF TABLES

TABLE		Page
1.	Band gaps of the 28 chalcopyrite compounds used in our investigation.	2
2.	Valency, Melting Point, Electronegativity, Pseudopotential Radius and Atomic Number of the elements used in our primary investigation. Where the abbreviations: Grp = Group, Elm = Element, EN = Electronegativity, AN = Atomic Number, MP = Melting Point, PR = Pseudopotential Radius and VL = Valency.	3
3.	Results of OLS analysis, summary table includes the top five performing models as determined by lowest MSCVE. AN means atomic number, PR is pseudopotential radius, and VL is valency. The suffixes <i>A</i> , <i>B</i> , and <i>C</i> denote the element; for example, AN- <i>B</i> is the atomic number of the <i>B</i> -site element.	7
4.	MSCVE of SPLS analysis, evaluated for each combination of the tuning parameters	12
5.	Results of preliminary investigation, extension to Suh and Rajan's 2004 work.	12
6.	Summary of simulation results from Designs I and II with $M = 500$ and $Q = (50, 200)$, using independence working correlation. Note that in Design I we have a case ICS without TVCS making Equation 2.14 an appropriate marginal analysis. Whereas, Design II there is no temporal component of the data but, ICS is present and there is a within cluster varying binary factor making Equation 2.15 an appropriate analysis.	44

7.	Summary of simulation results from Designs III and IV (M=500, Q=(50, 200), using independence working correlation. Note that Design III incorporates ICS and TVCS, thus Equation 2.16 is an appropriate marginal analysis. In the analysis of Design IV we use the knowledge that TVCS is determined via the ordinal logit model, dependent on cluster size in order to estimate the expected values of $l_i^{k_{ij}}$ in Equation 5.	45
8.	Aggregate results of racially stratified GEE, CWGEE and marginal WGEE analyses of AL. In each analysis an AR-1 working correlation was assumed. The effects reported are those significant according to a Wald test (covariance estimates obtained via sandwich and weighted sandwich) at the $\alpha \leq .05$ level.	54
9.	Tabular results of 1,000 Monte Carlo simulations from the indicated design with $Q = 50$ clusters and 20 Quadrature Points.	69
10.	Tabular results of 1,000 Monte Carlo simulations from the indicated design with $Q = 200$ clusters and 20 Quadrature Points.	72
11.	Tabular results of Joint Model estimates from the 1,000 Monte Carlo simulations from the indicated design with $Q = 50$ clusters and 50 Quadrature Points.	75
12.	Tabular results of Joint Model estimates from the 1,000 Monte Carlo simulations from the indicated design with $Q = 200$ clusters and 50 Quadrature Points.	78

LIST OF FIGURES

FIGURE	Page
1. Graphical summary of (a) PLS, (b) SPLS and (c) Lasso and Elastic Net performance.	9
2. The Lasso solution path for our data showing the effective coefficients of various descriptors in the Lasso model as a function of the number of LARS steps. In this plot, EN1 stands for EN- <i>A</i> , AN2 stands for AN- <i>B</i> etc.	10
3. SPLS MSCVE map over the indicated grid of tuning parameters. . .	11
4. Observed band gap of various doping concentrations of $(GaAs)_x(GaSb)_{1-x}$. The x-axis indicates the mixture concentration (<i>x</i>) and the y-axis indicates the observed band gap in eV.	14
5. Diagram of the various transformations and interactions employed in building up the descriptor set for the binary modeling setting. <i>LP</i> indicates a lattice parameter and the second column represents a set of interaction terms constructed accommodating the two crystal phases observed.	17
6. MSCVE in the region surrounding the best performing model for the binary investigation.	19
7. Observed band gap of various doping concentrations of $(GaN)_y(InN)_{1-y}$. The x-axis indicates the mixture concentration (<i>y</i>) and the y-axis indicates the observed band gap in eV.	22

8.	Band gap of the $(GaN)_y(InN)_{1-y}$ solution with overlaid quadratic model fit using standard spreadsheet software. The x-axis indicates the mixture concentration (y) and the y-axis indicates the observed band gap in eV.	24
9.	Snapshots of various intervals over a twenty year period. Recall that $E[Y] = 0$ and note that as time progresses the estimated distribution (blue line), based on observed information gets further from the hypothetical complete distribution.	35
10.	Graphical representation of the available data. (a) Baseline cluster size of each of the 452 individuals retained. (b) Number of observations made on each of the 7,823 retained teeth. (c) Total number of observations contributed by each individual. . .	49
11.	Distribution of attachment loss of teeth stratified by number of teeth at baseline and total number of observations made one each tooth. .	51
12.	Distribution of various demographic information stratified by race. All values have been centered according to the respective overall sample means.	57
13.	Power plots of the rejection rates of Wald statistic (estimated from bootstrap for Joint model), $Q = 50$ and 20 Quadrature Points. WGEE - Black, Joint Model -Red	70
14.	PP plots of the empirical coverage vs. Normal (sd obtained from 1,000 Monte Carlo samples), $Q = 50$ and 20 Quadrature Points. WGEE - Black, Joint Model -Red	71
15.	Power plots of the rejection rates of Wald statistic (estimated from bootstrap for Joint model), $Q = 200$. WGEE - Black, Joint Model -Red	73
16.	PP plots of the empirical coverage vs. Normal (sd obtained from 1,000 Monte Carlo samples), $Q = 200$ and 20 Quadrature Points. WGEE - Black, Joint Model -Red	74

17.	PP plots of the empirical coverage vs. Normal and Power plots of the rejection rates of Wald statistic (estimated from bootstrap) for β_0 and β_2 , $Q = 50$ and 50 Quadrature Points. WGEE - Black, Joint Model -Red	76
18.	PP plots of the empirical coverage vs. Normal and Power plots of the rejection rates of Wald statistic (estimated from bootstrap) for β_0 and β_2 , $Q = 200$ and 50 Quadrature Points. WGEE - Black, Joint Model -Red	77
19.	Estimated likelihood for various numbers of quadrature points. The likelihood is mapped for small variations of the indicated parameter (β_0), holding the other parameters fixed at the true value. Likelihood was estimated using one random generation of the indicated design with $Q = 50$ clusters.	79
20.	Estimated likelihood for various numbers of quadrature points. The likelihood is mapped for small variations of the indicated parameter (β_1), holding the other parameters fixed at the true value. Likelihood was estimated using one random generation of the indicated design with $Q = 50$ clusters.	80
21.	Estimated likelihood for various numbers of quadrature points. The likelihood is mapped for small variations of the indicated parameter (β_2), holding the other parameters fixed at the true value. Likelihood was estimated using one random generation of the indicated design with $Q = 50$ clusters.	82
22.	Estimated likelihood for various numbers of quadrature points. The likelihood is mapped for small variations of the indicated parameter (β_3), holding the other parameters fixed at the true value. Likelihood was estimated using one random generation of the indicated design with $Q = 50$ clusters.	83

23.	Estimated likelihood for various numbers of quadrature points. The likelihood is mapped for small variations of the indicated parameter (β_0), holding the other parameters fixed at the true value. Likelihood was estimated using one random generation of the indicated design with $Q = 200$ clusters.	84
24.	Estimated likelihood for various numbers of quadrature points. The likelihood is mapped for small variations of the indicated parameter (β_1), holding the other parameters fixed at the true value. Likelihood was estimated using one random generation of the indicated design with $Q = 200$ clusters.	85
25.	Estimated likelihood for various numbers of quadrature points. The likelihood is mapped for small variations of the indicated parameter (β_2), holding the other parameters fixed at the true value. Likelihood was estimated using one random generation of the indicated design with $Q = 200$ clusters.	86
26.	Estimated likelihood for various numbers of quadrature points. The likelihood is mapped for small variations of the indicated parameter (β_3), holding the other parameters fixed at the true value. Likelihood was estimated using one random generation of the indicated design with $Q = 200$ clusters.	87

CHAPTER I

MATERIALS SCIENCE APPLICATION OF HIGH DIMENSIONAL REGRESSION METHODS: BAND GAP ESTIMATION

Modern semiconductors are singlehandedly responsible for the digital age. Their utility ranges from digital devices to energy efficient lighting and further to solar energy production. Each specific application requires the identification and production of semiconducting materials with desirable properties. One common property of interest is the electronic band gap. Heuristically, the band gap of a material can be thought of as the amount of energy needed to change the conductive properties of a semiconductive material.

Bulk electronic band structures are the result of the spacial arrangement of molecules, atoms and electrons in a crystal structure. Mapping electronic band structures from first principle (*ab initio*) methods is possible. Due to the complicated nature of the relationships between particles such methods are computationally expensive making a wholesale investigation of *interesting* materials unfeasible. Since the quantity of interest (band gap, the differential between the highest valence band and the lowest conduction band) represents only a very small amount of the information contained in the bulk electronic band structure it may be unnecessary to map the entire band structure.

TABLE 1

Band gaps of the 28 chalcopyrite compounds used in our investigation.

<i>I – III – VI₂</i> Compounds					
<i>CuAlS₂</i>	<i>CuAlSe₂</i>	<i>CuAlTe₂</i>	<i>CuGaS₂</i>	<i>CuGaSe₂</i>	<i>CuGaTe₂</i>
3.49	2.67	2.06	2.43	1.68	1.12
<i>CuInS₂</i>	<i>CuInSe₂</i>	<i>CuInTe₂</i>	<i>AgAlS₂</i>	<i>AgAlSe₂</i>	<i>AgAlTe₂</i>
1.53	1.04	1.06	3.13	2.55	2.27
<i>AgGaS₂</i>	<i>AgGaSe₂</i>	<i>AgGaTe₂</i>	<i>AgInS₂</i>	<i>AgInSe₂</i>	<i>AgInTe₂</i>
2.64	1.8	1.32	1.87	1.24	0.95
<i>II – IV – V₂</i> Compounds					
<i>ZnSiP₂</i>	<i>ZnSiAs₂</i>	<i>ZnGeP₂</i>	<i>ZnGeAs₂</i>	<i>CdSiP₂</i>	<i>CdSiAs₂</i>
2.07	1.74	2.05	1.15	2.33	1.55
	<i>CdGeP₂</i>	<i>CdGeAs₂</i>	<i>CdSnP₂</i>	<i>CdSnAs₂</i>	
	1.72	0.57	1.17	0.26	

A Preliminary Investigation

Materials scientists have been considering chemo-informatic alternatives to estimating band gaps for years. One such investigation (Suh and Rajan, 2004) laid the framework for our investigation. Suh and Rajan attempted to estimate the band gap of 28 known chalcopyrite compounds through the implementation of Partial Least Squares Regression. They used a small set of pseudo-compound properties, constructed as linear combinations of the corresponding elemental properties, as the descriptor set and used the known experimental band gaps as the response. Figure 1 presents the 28 compounds used in Suh and Rajan’s 2004 investigation as well as ours.

Each of the compounds in our investigation was a crystal structure composed of ternary compounds of the form ABC_2 . Unlike Suh and Rajan’s 2004

TABLE 2

Valency, Melting Point, Electronegativity, Pseudopotential Radius and Atomic Number of the elements used in our primary investigation. Where the abbreviations: Grp = Group, Elm = Element, EN = Electronegativity, AN = Atomic Number, MP = Melting Point, PR = Pseudopotential Radius and VL = Valency.

<i>I – III – VI₂ Compounds</i>							<i>II – IV – V₂ Compounds</i>						
Grp	Elm	EN	AN	MP	PR	VL	Grp	Elm	EN	AN	MP	PR	VL
I	Cu	1.08	29	1358.0	2.04	11	II	Zn	1.44	30	692.7	1.88	12
	Ag	1.07	47	1235.0	2.375	11		Cd	1.40	48	594.3	2.215	12
III	Al	1.64	13	933.5	1.675	3	IV	Si	1.98	14	1687.0	1.42	4
	Ga	1.70	31	302.9	1.695	3		Ge	1.99	32	1211.0	1.56	4
	In	1.63	49	429.8	2.05	3		Sn	1.88	50	505.1	1.88	4
VI	S	2.65	16	388.4	1.1	6	V	P	2.32	15	317.3	1.24	5
	Se	2.54	34	494.0	1.285	6		As	2.27	33	1089.0	1.415	5
	Te	2.38	52	722.7	1.67	6							

investigation we chose to exploit the elemental information in our modeling. As such, we obtained a set of fifteen descriptors (five for each of the three elements) the individual elemental descriptors are included in Figure 2. To put it more clearly, the compound $CuAlS_2$ has A element equal to Cu which has an electronegativity(EN)= 1.08, hereafter we will use place holder descriptor ID’s such that the electronegativity of the A element will be denoted EN- A as such EN- A of $CuAlS_2$ would be 1.08. Similarly, PR- C of $CuAlS_2$ would represent the Pseudopotential Radius of S (further abbreviations are indicated in Table 2).

1 Methods

In many regression applications, model interpretation is of primary concern. In such settings the use of black box type data mining methods is discouraged and often simpler linear models are adopted instead. Such was the case in our preliminary investigation as a number of collaborators were employing various feature selection and variable importance work independently. The intent of the

independent approaches was to further validate the candidate models through the identification of similar selected features. In the following let, $X = (x_{ij})_{1 \leq i \leq N, 1 \leq j \leq p}$ denote a $N \times p$ matrix of descriptors, $Y = y_1, \dots, y_N$ be the vector of corresponding responses and β be the vector of parameters belonging to the relationship $m(x) = f(x, \beta)$ where $m(x) = E(Y|X = x)$.

In our preliminary investigation we chose to exploit three regression techniques: Ordinary Least Squares (OLS), Sparse Partial Least Squares (SPLS) and Elastic net. It should be noted that the results will be discussed in terms of OLS, Partial Least Squares (PLS), Least Angle Shrinkage and Selection Operator (Lasso) and Elastic Net. While this would indicate that there were five methods used PLS and Lasso are specific cases of SPLS and Elastic Net respectively; as such whether three or five methods are used is a matter of semantics. Below we will briefly present each method as well as strengths and weakness of each, in the Extension Section and Discussion Section we will elaborate on the associated strengths and pitfalls associated with each method.

Ordinary Least Squares

From an estimation standpoint, OLS is the simplest form of regression. Provided that a handful of prior assumptions are tenable, OLS modeling is appropriate. Estimates of the parameters of an OLS model are obtained as the solution to the following optimization problem:

$$\hat{\beta}^{OLS} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^N \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \right\}. \quad (1.1)$$

Conveniently, this optimization problem has a closed form solution as:

$$\hat{\beta}^{OLS} = (X^T X)^{-1} X^T Y. \quad (1.2)$$

Model selection and tuning of OLS models is performed by varying the composition of X to include or exclude predictors. There are $2^p - 1$ possible sets of parameters

which contain only first order terms for any p descriptor data set.

Lasso

Lasso is a simultaneous variable selection and regression method. Unlike subset selection methods which are discrete methods in which descriptors are retained or discarded with a high level of variability, shrinkage methods provide a more continuous manner for model estimation. Variable shrinkage methods often produce models with lower variability than models estimated using subset selection techniques (Hastie *e.t. al*, 2001). The estimation of Lasso models is carried out by imposing an L_1 penalty, λ_1 , on the sum of the β estimates. Estimates of Lasso models are obtained as the solution to the following penalized least squares optimization:

$$\hat{\beta}^{Lasso} = \underset{\beta}{\operatorname{argmin}} (Y - X^T \beta)^T (Y - X^T \beta) + \lambda_1 |\beta|_1, \text{ (Tibshirani, 1996)} \quad (1.3)$$

where λ_1 is a tuning parameter that controls the amount of regularization/shrinkage (Hastie *e.t. al*, 2001).

Elastic Net

Elastic Net is an extension of Lasso which controls, via scaling of the covariance matrix, the correlation of predictors (Zou and Hastie , 2005). This decorrelation makes the elastic net a stabilization of the Lasso. Estimates of the Elastic Net model are obtained as the solution to the optimization problem:

$$\hat{\beta}^{ElasticNet} = \underset{\beta}{\operatorname{argmin}} (Y - X^T \beta)^T (Y - X^T \beta) + \lambda_1 |\beta|_1 + \lambda_2 |\beta|_2^2. \quad (1.4)$$

PLS and SPLS

Partial least squares (PLS) is a dimension reduction/latent variable regression technique which uses predictors derived as linear combinations of the original predictors in order to predict the response. The derived predictors are

estimated such that inclusion of each successive predictor in the model will lead to a lesser reduction of variance than the inclusion of the prior. This is favorable to the former method of subset selection in that there are a relatively small number of models ($2^p - 1$ in the case of LOO cross validated investigation) which need to be investigated. PLS unlike principal component analysis (PCA) takes into account the response vector as well as the set of descriptors. More explicitly PLS derives latent constructs in a manner that maximizes the covariance between the derived descriptors and responses. It is widely used in the field of Chemometrics. Sparse partial least squares (SPLS) is a generalization of PLS that adds sparsity to the latent variable regression by adding a sparsity parameter η (where $\eta = 0$ corresponds to PLS) along with the number of derived predictor terms K .

Crossvalidation

Cross validation is often used as a convenient means to judge a regression models prediction ability. For this analysis, mean squared cross validated error (MSCVE) was employed as the performance measure for comparison between different models. Leave one out (LOO) cross validation was the method of cross validation employed for this analysis, given the small number of records available this is a computationally reasonable approach. LOO is the case of cross validation whereby one object is removed at a time from the data, a model is then fitted for incomplete data and an error estimate is obtained by making a prediction for the removed object and comparing the prediction to the observed data; this process is repeated until an error estimate has been obtained for every object. MSCVE is then obtained as the mean of the squared cross validated errors. The following is a comparison of the three regression techniques employed.

2 Results

We report detailed results for the three classes of regression models. For the OLS models, we have tried all possible subsets of descriptors to find the best fitting model. Since there are only two types of chemistries (groups) in the training data, the valency parameters are collinear. However, we retained valency in our computational pipeline since future extensions of our models will be based on a larger training set for which these variables will become linearly independent. In Table 3, we only report the top 5 performing OLS models (as well as the corresponding relevant descriptors) w.r.t. MSCVE.

TABLE 3

Results of OLS analysis, summary table includes the top five performing models as determined by lowest MSCVE. AN means atomic number, PR is pseudopotential radius, and VL is valency. The suffixes *A*, *B*, and *C* denote the element; for example, AN-*B* is the atomic number of the *B*-site element.

Top Performing OLS Models (by MSCVE)	
Descriptors Selected	MSCVE
AN- <i>B</i> , AN- <i>C</i> , PR- <i>B</i> , PR- <i>C</i> , VL- <i>A</i>	0.0501
EN- <i>B</i> , AN- <i>B</i> , AN- <i>C</i> , PR- <i>C</i> , VL- <i>A</i>	0.0515
AN- <i>B</i> , AN- <i>C</i> , PR- <i>C</i> , VL- <i>A</i>	0.0525
EN- <i>B</i> , AN- <i>C</i> , MP- <i>B</i> , PR- <i>B</i> , PR- <i>C</i>	0.0527
EN- <i>C</i> , AN- <i>B</i> , AN- <i>C</i> , PR- <i>B</i> , VL- <i>A</i>	0.0529

The Lasso and Elastic net investigation was conducted by examining each of the 14 Lasso/Elastic Net step models for each of the following values of the quadratic penalty parameter $\lambda_2 = \{.01, .1, 1, \dots\}$. It was found that the 13 step Lasso yielded the best model $\text{MSCVE} = 0.0527$. The MSCVE results are graphically displayed in the rightmost panel of Figure 1. Note that the Lasso is fitted by the computationally efficient least-angle regression (LARS) algorithm. Figure 2 displays the solution path for Lasso for our data where the number of LARS steps corresponds to different degrees of regularization. This figure shows the

relative magnitudes of the coefficients of various model descriptors under different LARS steps (equivalently, under different λ_1 s). Figure 4 presents a graphical presentation of PLS/SPLS and Lasso/Elastic Net model performance.

Figure 1. Graphical summary of (a) PLS, (b) SPLS and (c) Lasso and Elastic Net performance.

Figure 2. The Lasso solution path for our data showing the effective coefficients of various descriptors in the Lasso model as a function of the number of LARS steps. In this plot, EN1 stands for EN-*A*, AN2 stands for AN-*B* etc.

The SPLS investigation fitted SPLS models across a grid of (η, K) values where $\eta = (0.1, 0.2, \dots, 0.9)$ and $K = (1, 2, \dots, 14)$. The PLS models (corresponding to $\eta = 0$) were fitted as well by varying the number of PLS components. The best performing model was found to be the one estimated using the (η, K) combination $(0.8, 8)$ with $\text{MSCVE} = 0.0619$. A table of the MSCVE values for each of the (η, K) combinations is included in Table 4. The results are graphically displayed in the left and middle panels of Figure 1 and 3.

The aggregate results of the predictions for the training data are presented in

Figure 3. SPLS MSCVE map over the indicated grid of tuning parameters.

Table 5. As can be seen from Table 5, OLS produced the lowest MSCVE (highest accuracy), followed by Lasso and SPLS. Further investigation into the selected descriptors (significantly contributing elemental properties) to each model showed a considerable amount of agreement between the elemental properties selected via our algorithmic selection and the PCA and Rough Set analysis conducted by our collaborators. Furthermore, while Suh and Rajan did not employ a cross validated analysis in their 2004 analysis, we reproduced a similar analysis based upon their design and used the reported MSCVE as a metric for comparison in early iterations of our analysis. While our approach was a dramatic improvement with regard to MSCVE over the 2004 investigation we neglected a straightforward comparison of the two. For a more detailed account of the results as well as the feature selection

TABLE 4

MSCVE of SPLS analysis, evaluated for each combination of the tuning parameters

K	η								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	0.2224	0.2251	0.2333	0.2581	0.2448	0.1940	0.1565	0.3491	0.3346
2	0.0988	0.0982	0.0930	0.0919	0.1066	0.0751	0.0707	0.0810	0.0634
3	0.0799	0.0786	0.0783	0.0743	0.0722	0.0755	0.0727	0.0711	0.0641
4	0.0830	0.0830	0.0873	0.0929	0.0923	0.0926	0.0889	0.0775	0.0833
5	0.0857	0.0857	0.0857	0.0816	0.0847	0.0796	0.0694	0.0713	0.0758
6	0.0755	0.0755	0.0755	0.0755	0.0749	0.0720	0.0766	0.0723	0.0667
7	0.0695	0.0695	0.0695	0.0695	0.0694	0.0682	0.0678	0.0671	0.0649
8	0.0649	0.0649	0.0649	0.0649	0.0652	0.0654	0.0655	0.0619*	0.0646
9	0.0656	0.0656	0.0656	0.0656	0.0656	0.0655	0.0656	0.0656	0.0648
10	0.0727	0.0727	0.0727	0.0727	0.0727	0.0727	0.0727	0.0727	0.0727
11	0.2800	0.2800	0.2800	0.2800	0.2800	0.2836	0.2989	0.2978	0.3050
12	0.6376	0.6376	0.6376	0.6376	0.6376	0.6751	0.6970	0.7640	0.6771
13	1.2626	1.2626	1.2626	1.2626	1.2626	1.3658	1.4090	1.5809	1.2872
14	2.1551	2.1551	2.1551	2.1551	2.1551	2.2571	2.3456	2.5288	1.6222

work conducted by our collaborators the reader is directed to see Dey *et. al* (2014).

TABLE 5

Results of preliminary investigation, extension to Suh and Rajan's 2004 work.

Method	Parameters/Descriptors	MSCVE
OLS	AN- <i>B</i> , AN- <i>C</i> , PR- <i>B</i> , PR- <i>C</i> , VL- <i>A</i>	0.0501
Lasso/ Elastic Net	Lasso, 13 steps	0.0527
SPLS	$\eta = 0.8$, $K = 8$	0.0619
PLS	9 components	0.0651

3 Discussion

Our preliminary efforts yielded a number of useful findings. The first being the realization that while the intent to simplify band gap estimation through the implementation of informatics based tools was sound and reasonable the approaches we had encountered and extended had dramatically over simplified the context of

the problem. For example, it is possible for a particular chalcopyrite compound to exist in several different crystal phases. That is the arrangements of unit cells within two crystals composed of the same chemical compound can be different. Band structure is the consequence of the spacial arrangement of particles within a repeating structure as such if the same compound is arranged in any of a number of different ways the band structures will differ. In the our preliminary investigation and the 2004 investigation crystal phases. This was highlighted by Dr. Jacek Jasinski in the acquisition of body of information where band gaps of a number of compounds were reported for various crystal phases.

Further investigation into the sources of retained error indicated that both investigations had ignored what type of gap was being reported (transition responsible for the reported band gap). This issue presents another dimension equally as difficult to address as the crystal phase problem. The issue being that while band gaps are usually reported as direct or indirect (direct *usually* being the $\Gamma\Gamma$ transition) our data did not include any information on the transition corresponding to the reported band gap. This presented a number of challenges not readily addressed by a straightforward literary search. Given the observed discrepancies between band gaps of the same compounds in different phases and differences in direct and indirect band gaps the observed prediction errors were in line with reasonable expectation. In light of the pitfalls encountered from the over simplification of the problem, the utility of being able to make predictions by exploiting the combinatorial nature of chemical compounds is undeniable. As such the follow up investigation (into solid solutions) was conducted with due consideration to the issues encountered in the chalcopyrite investigation.

B Solid Solutions Problem

Having learned a considerable amount from the chalcopyrite investigation, efforts were made to segue into what was referred to as the Solid Solutions Problem.

The solid solutions problem is an effort to estimate band structure properties of various mixtures of compounds. For example, we could have two crystal compounds, say, Compound 1- AB_2 and Compound 2- CB_2 with the desired element and compound level information known for each. The Solutions problem involves estimating the properties of various doping concentrations of the two *i.e.* $S_x = (AB)_x(CB)_{1-x}$. For example take, $GaAs$ and $GaSb$, $GaAs$ has a direct band gap of around 1.43eV, where $GaSb$ has a direct band gap of around .7. Figure 4 outlines the observed band gap of the solution obtained at various doping concentrations of $(GaAs)_x(GaSb)_{1-x}$. The principle objective of our solid solutions investigation was to accurately predict the desired band structures of these and new solid solutions.

Figure 4. Observed band gap of various doping concentrations of $(GaAs)_x(GaSb)_{1-x}$. The x-axis indicates the mixture concentration (x) and the y-axis indicates the observed band gap in eV.

In our investigation we limited the structure of the solid solutions to compositions of two binary compounds (like those indicated above in the

($GaAs$) _{x} ($GaSb$) _{$1-x$} example). There were two approaches that seemed applicable at the onset of our investigation, the first being a slight misspecification of the situation, which was to model the solutions properties based on the elemental concentration of the solution (ignoring the properties of the binary compounds). The second being to model the solutions properties as a function of the binary concentrations (ignoring the individual elemental properties). The utility of the first method lie in the fact that, often, there was not complete information available for all of the binary compounds whereas the elemental information was available. However, modeling the solutions properties based solely on elemental information would have been inappropriate as binary compound level information would be ignored. In practice, a hybrid approach was adopted exploiting both element and binary compound level information.

1 Approach to Modeling Band Gaps of Solid Solutions

In an effort to validate the mixture models, modeling began by specifying and fitting models to estimate the band gaps of the individual binary components of the mixture data. This choice was made in part because similar issues were encountered with the availability of compound level information (such as crystal phase) as in the chalcopyrite investigation; including elemental information in the framework would also allow for greater flexibility downstream when extending the classes of compounds under investigation. Doing so also yields a multi-pronged approach to mixture modeling.

Two tiered approach to mixture modeling:

1. Modeling of the band gaps of the individual binary compounds
 - (a) Elemental Properties (Specific Volume, Electronegativity,...)
 - (b) Binary Level Properties (Lattice Constants, Atomic Radii, Density,...)

2. Extending the binary model to accommodate mixtures

- (a) Elemental concentration (including the binary level information)
- (b) Binary concentration (with solution level information)
- (c) Hybrid

2 Data

In all, data was collected through literary searches on 32 mixture profiles. Because the data were collected from numerous different sources there were considerable discrepancies in the information available for the collected data points. In some cases sources were very specific providing the transition corresponding to the reported band gap as well as the crystal phases corresponding to the binary compounds used. In most cases one or both of these pieces of information were unavailable. Considerable effort was placed into externally validating the various sources and to fill in the missing information regarding binary level information. Ultimately, the band gaps of 31 binary compounds, 14 distinct (several were replicates obtained from multiple sources), for which a reasonable level of confidence in the validity of the reported quantities could be assumed, were retained for analysis. In the binary investigation a set of 240 predictors composed of elemental properties (atomic weight, melting point, boiling point, covalent radii, thermal conductivity etc.), binary level information (crystal phase, melting point, lattice constants, etc) and various transformations and interactions between the two were constructed. Figure 5 provides a diagram of the various interactions and transformations employed in building up the descriptor set for the binary modeling setting.

Figure 5. Diagram of the various transformations and interactions employed in building up the descriptor set for the binary modeling setting. LP indicates a lattice parameter and the second column represents a set of interaction terms constructed accommodating the two crystal phases observed.

3 Modeling - Binary Compounds

With the set of descriptors constructed (using transformations of scientific interest) attention was turned to finding an appropriate transformation of the observed band gap information in order to facilitate modeling. As a rule semiconductors have positive band gaps (negative band gaps would indicate an overlap of valence and conduction bands which would denote that a material was a conductor), as such logarithm, square root, other root and power transformations were reasonable choices for transformations. We were interested in a model of the form,

$$h(BandGap) = X\beta, \tag{1.5}$$

in order to facilitate the implementation of linear regression methods (in an effort to retain model interpretability), with X corresponding to the indicated binary descriptor set and $h(BandGap)$ being, generically, a transformation of the observed band gaps.

In the chalcopyrite investigation Elastic Net produced consistently stable models, with predictions made on non trained compounds being reasonably in line with expectation. As such we employed Elastic Net as the primary method for modeling the binary compound band gaps. A tuning scheme similar to that of the chalcopyrite investigation was employed, with the exception that tuning was performed over a grid of (α, λ_2) values where α corresponds to a parameter representing the magnitude of λ_1 . λ_1 takes values from (0,1) with 0 indicating the null model and 1 indicating a saturated fit. Tuning was performed by varying α over the set $\{.05, .06, \dots, .95\}$ and λ_2 over the set $\{0, .01, .1, 1, 10\}$. Comparisons were made for various transformations (choices of $h(.)$) and it was determined that a square root transformation would provide the best option.

In assessing the initial viability of the proposed model *i.e.* $\sqrt{BandGap + C} = X\beta$, LOO cross validation was employed and MSCVE was used as the metric for model comparison. Figure 6 shows the topography of the MSCVE

map around the region of smallest predictive error.

Figure 6. MSCVE in the region surrounding the best performing model for the binary investigation.

The best performing model had an MSCVE of .1798. This is in contrast to the chalcopyrite investigation where the best performing models had an average MSCVE around .056, neither of which was in line with an acceptable margin of error for experimental validation via synthesis. Further investigation indicated that there were discrepancies between the reported observed values and independently documented values for the same quantities. This compounded with observed discrepancies in observed 'replicate' values, that is, band gaps obtained from

multiple sources for the same compound (which should be in reasonable agreement with one another). Pointed to significant integrity issues in the collected data. At the time of the writing of this manuscript efforts were underway to validate the solutions data. With consistent mixture data, the framework is in place for a straightforward extension to the solid solutions problem upon validation of an acceptable binary model.

4 Defining the Mixture Model

At present, the inability to validate the binary models (as a result of data integrity issues) has stalled a full scale exploration of the Solid Solutions problem. As such there exists (at this time) no quantitative metric with which to compare the various approaches to defining an appropriate model structure for the mixture setting. We present here a number of approaches, for subjective consideration, to be investigated when data acquisition issues are overcome.

Generic Form

In the simple case (composition of two distinct compounds) a mixture model will need to incorporate information from both primary compounds as well as the mixture concentration. Generically, we can then define a model for the for the mixture problem as:

$$h(MixBandGap) = g(X_1, X_2, y_{con}). \quad (1.6)$$

Where for the purposes of this section X_1 and X_2 denote the various descriptor information available for the first and second compound, respectively and y_{con} denotes the concentration information. As indicated previously in the binary modeling section, the choice of $h(.)$ *i.e.* transformation on the observed band gap, can be considered a trivial matter consisting of finding a transformation which

makes a linear relation appropriate. As such, in the mixture problem the onus is placed on the form of $g(X_1, X_2, y_{con})$.

Elemental Concentration

In Section 1, we indicated two approaches to the extension of the binary setting to the mixture problem. The first considers approaching the problem as a function of elemental concentration. The elemental concentration approach would require a re-framing of how one would consider the true state of the resultant mixture material. Say, that a mixture was the composition of two binary materials, say, AB and CB in varying concentrations, say, y_{con} such that the composite material could be defined as $(AB)_y(CB)_{(1-y)}$. In specific settings, for example, when the B element is shared it would be possible to define the material as a hybrid ternary *i.e.* $A_yC_{1-y}B$. We could then modify our generic mixture model *i.e.* Equation 1.6 as a function of the individual elemental components and their corresponding concentrations:

$$h(MixBandGap) = g(X_A, X_B, X_C, y_{con}). \quad (1.7)$$

Where X_A X_B and X_C denote the relevant elemental descriptor information corresponding to the A , B and C elements, respectively and $X_A \subseteq X_1 \cap X_2 \dots$. Such an approach is enticing because of its simplicity and would at a glance appear appropriate for mixtures like $(GaN)_y(InN)_{1-y}$, for modeling band gap as a function of concentration included as Figure 7. However, one should shy away from this sort of generalization. Implementation of the elemental concentration model ignores binary and mixture level information and is also not readily generalizable to settings where there is no shared element say for example the $(AB)_y(CD)_{1-y}$ setting; an issue that would be exacerbated when extending to ternary and larger materials.

Figure 7. Observed band gap of various doping concentrations of $(GaN)_y(InN)_{1-y}$. The x-axis indicates the mixture concentration (y) and the y-axis indicates the observed band gap in eV.

Binary Concentration

The second approach suggests using the available binary information and is a more straightforward interpretation of the general approach of Equation 1.6. In a binary concentration model it would be unnecessary to include elemental level descriptors as the necessary binary level information would convey the relevant information. Equation 1.8 demonstrates the general form of a binary concentration model:

$$h(MixBandGap) = Q(g_1(X_1), g_2(X_2), y_{con}). \quad (1.8)$$

Where X_1 and X_2 would be binary level information for the first and second compounds, respectively. This approach is more appropriate and flexible than that of the elemental concentration model. As a toy example we could define Q as:

$$Q(.) = \sqrt{\{g_1(X_1) \cdot y_{con}\}^2 + \{g_2(X_2) \cdot (1 - y_{con})\}^2} \quad (1.9)$$

If our concern was only being able to accurately estimate the properties of mixtures of binary compounds for which comprehensive binary level information was available the binary mixture model would be appropriate. In this setting attention should be turned to estimating an appropriate form for Q and incorporating mixture level information (structure of the resulting solution, corresponding mixture band gap transition etc.).

Hybrid Model

While the binary concentration model is a decided improvement with respect to the elemental concentration model, it is still limited with respect to its extensibility and ability to be internally validated. Conventional wisdom would dictate that if at one level of concentration a solution exhibits a given quantity and at a different level of concentration said solution exhibits a different quantity then there exists some relationship that can accurately map the transition from one concentration to another. Therefore, the utility of a combinatorial mixture model does not lie in modeling the relationship between the band gap of a solution composed of two compounds for which comprehensive information is available. Creating individual mixture models for specific solutions is a trivial task and can often be carried out simply as a function of the two known compound band gaps and the corresponding concentration. For example, Figure 8 provides such a mapping from one concentration to another (obtained using standard MS office software) that parametrically models the band gap of the $(GaN)_y(InN)_{1-y}$ solution as a function of concentration and the two binary compound bandgaps.

One issue with independently modeling each solution as a function of known band gaps and concentration is that it provides little information as to what form the band gap transition from one concentration to another should take when attempting to describe a solution comprised of new compounds. Another issue is that modeling the solution as a function of known band gaps makes the resulting

Figure 8. Band gap of the $(\text{GaN})_y(\text{InN})_{1-y}$ solution with overlaid quadratic model fit using standard spreadsheet software. The x-axis indicates the mixture concentration (y) and the y-axis indicates the observed band gap in eV.

model inextensible to solutions containing compounds with unknown band gaps. The implication of this second issue is that the resulting model is rather static and cannot be readily extended to a large scale search of unknown materials. The purpose of a combinatorial modeling approach is to facilitate a broad search of unknown candidate materials with desirable properties, and the binary concentration model is not flexible enough to accommodate materials with unknown properties.

The solution is a hybrid approach, whereby elemental and binary compound level information are used to model binary compound band gaps then the information for the binary compound models are passed to a mixture model where the solution band gap can be estimated as a function of the modeled binary band gaps, binary level information, concentration and solution level properties (where applicable). For example, let $g(X_1)$ and $g(X_2)$ denote functions estimating the band gaps of the first and second compounds comprising a solid solution as functions of

their corresponding elemental and binary level descriptors X_1 and X_2 , respectively.

$$h(MixBandGap) = Q(g(X_1), g(X_2), y_{con}, b_1, b_2, C). \quad (1.10)$$

Where $b_1 \subset X_1$ and $b_2 \subset X_2$ denote the binary level information contained in X_1 and X_2 and C denotes solution level properties. The flexibility of this approach come from the fact that incorporating binary and solution level information allows the form of the mixture model to change with respect to the properties of constituent compounds and resulting solution. For example, let θ denote the set of binary and solution level descriptors *i.e.* $\theta \in b_1 \cap b_2 \cap C$, assume that there are known values of θ for which the relationship between solution band gap and concentration is known to differ (for some, band gap could be linear w.r.t. concentration, others quadratic etc.). We can incorporate our knowledge of these relationships into the model by allowing the values of θ to dictate the form of Q :

$$Q(.) = \begin{cases} \sqrt{\{g(X_1) \cdot y_{con}\}^2 + \{g(X_2) \cdot (1 - y_{con})\}^2} & : \theta = \theta_0 \\ g(X_1) \cdot y_{con} + g(X_2) \cdot (1 - y_{con}) & : \theta = \theta_1 \\ \dots & \dots \end{cases}$$

The Elemental and Binary concentration models offer a good starting point but fall short of being able to accommodate the scope of the solid solutions project. The Hybrid model while flexible and readily extensible further complicates an already involved modeling problem. The good news is that through the clever construction of descriptor design matrices the problem can be readily simplified and efforts can be focused in two seemingly independent directions: binary band gap estimation and defining appropriate relations to map band gap as a function of concentration.

C Conclusion

The utility of reducing the computational burden associated with estimating electrical properties of new semiconducting materials is undeniable. Framing the problem to exploit the combinatorial nature of stoichiometric design provides such possibilities with respect to downstream extensibility and applicability would make effective implementation profitable beyond imagination. To that end, after expending such time and energy one is obliged to elucidate some consistent methodological and practical maladies which will be encountered when approaching the “Materials Genome” type problems.

1 Be wary of over simplification

The principle objective of new materials discovery via chemo-informatic modeling is simplification. However, *ab initio* methods require an extraordinary amount of information in order to accurately map bulk material band structures. While it may be possible accurately estimate electrical properties without the complex quantum calculations, it will not be possible ignoring structural information inherent to the materials and properties (for example, crystal structure and which transition is being estimated).

2 On the acquisition of data

While there exist aggregate repositories of materials data there are inconsistencies in the depth of information available for various materials even within single databases. Many materials such as *GaInN* have been the subject of numerous experiments and as such the body of available knowledge about them is vast. In other settings there may be very little, if any, information available. In such a rapidly progressing field the conventional wisdom would suggest that a literary search of recent work would provide a wealth of applicable information. While this

is the case **extreme** care must be taken when mining literary sources for information. The data collected from such sources is often not the product of independently validated research, as such it will carry with it inherent error which cannot be removed. Similarly, quantities collected via literary search often exclude relevant information with respect to the classification of the material and the resulting properties reported; this information must often be match merged against other sources in an attempt to identify specifically which material was being investigated (thus introducing another possible source of variability).

Continued success hinges on the development of a comprehensive (externally validated) repository of materials data. Such efforts should include the acquisition of *all applicable* information pertaining to each material contained therein. While consensus does not denote scientific accuracy, this repository should be the result of several independent entities working in concert to collect necessary data and independently confirm the reported quantities. With such a repository the viability of this and similar informatics based investigations is almost certain.

3 Brief Discussion of Applicability and Implementation of Regression Techniques in Chem-Informatic Settings

For the sake of anyone attempting to pursue such an investigation it would be unfair to not comment on a brief list of strengths and weaknesses inherent to the various regression methods discussed. As mentioned before Lasso and Elastic Net provide readily interpretable models in terms of the original descriptor set. This is terribly convenient as it provides insight to the relevant quantities related to the desired physical properties. These two methods have an added benefit in that the models generated using these methods exploit only a (comparably) small subset of the original descriptors effectively weeding out the irrelevant variables. This leads to *smaller* models which are typically easy to validate.

Partial Least Squares and Sparse Partial Least Squares enjoys popularity

within the chem-informatic community. However, the nature of the latent constructs underlying the PLS and SPLS modeling framework make model interpretation rather involved. Similarly, the use of latent variable techniques where the individual components are constructed as linear combination of all or many of the descriptors can make model validation extremely difficult. This issue extends to downstream difficulties with prediction accuracy as extrapolation concerns are exacerbated by the inclusion of possible noise variables. Though, the black box appeal of PLS and SPLS cannot be denied.

In any case, model cross validation is a necessity in this type of work. However, with particularly noisy data one cannot be guaranteed that model over fitting is not a concern. Detailed appraisal of the data should be conducted, especially on replicate quantities and on materials where the data were collected from multiple sources, in order to obtain a back of the envelope estimate of the baseline variability in the reported quantities. Such variability (as in discrepancies between replicate quantities obtained for the same material) cannot be overcome. Similarly, misspecification of descriptors (as in properties mined from multiple sources, but incorrectly) effectively constitute a misspecification of the model structure and will introduce yet another source of variation which will further complicate the task at hand.

CHAPTER II

CLUSTER ADJUSTED REGRESSION FOR DISPLACED SUBJECT DATA (CARDS): MARGINAL INFERENCE UNDER POTENTIALLY INFORMATIVE TEMPORAL CLUSTER SIZE PROFILES

A Introduction

It's no secret that sample size restrictions plague many otherwise well thought out experiments. In repeated measure problems issues of sample size can be mitigated through various sampling strategies; though it is seldom possible to acquire a balanced sample in spite of the best laid plans. This issue is exacerbated within the context of clustered longitudinal data as it may not only be impossible to acquire balanced information at the onset of study; it might also be the case that attrition rates vary among subjects. In the case of the data collected by Beck *et. al* (1997a) a periodontal study which collected information on teeth belonging to elderly individuals over the course of five years we see a perfect example of how obscure the distributions of available information can be. While some of the subjects retain all 32 teeth over the five year period others have only 1 tooth at baseline. Further complicating the matter some individuals lose teeth during the study and others are removed from study entirely due to various causes.

It is often the case that the number of units available at the onset of a study and a cluster members attrition are often dependent in some manner on the outcome. Such a dependence is in violation of the underlying assumptions of many

methods for analyzing clustered longitudinal data, specifically Generalized Estimating Equations (GEE). Often the motivation for analyzing clustered data is to draw conclusions about overall population effect size of various covariates. The utility of GEE methods is that they allow estimation of these (marginal) population parameters while accommodating the within cluster effects, where the within cluster effects are of either auxiliary or no interest to the study question.

GEE's have proven instrumental in the analysis of clustered data. Their flexibility affords them the luxury of wide range of applications not the least of which being within the longitudinal setting. Their behavior in the presence of dependence between the availability of data and the outcome within the context of non-longitudinal settings has received considerable attention over the past decade. In Section B we will discuss *informative cluster size* and *temporally varying cluster size* the situations where there is a dependence between the outcome and the availability of information and where cluster members are displaced such that cluster size becomes non-constant with time. In Section C we recommend various marginal analyses and motivate the implementation of each according to contextual appraisal of the data.

In Section 1 we discuss criteria for when marginal analyses of observed data can be expected to agree with the distribution of the population composed of hypothetical complete clusters. In Section D we provide the results of a number of simulations comparing marginal observed cluster inference drawn from various designs to that of the corresponding complete cluster distribution. Our work was motivated principally by the need to analyze periodontal data. In such a setting the clustering unit is the subject under study and teeth the individual cluster members. Heuristically, one can understand the relationship between the number of teeth an individual has and that individuals overall oral health. Similarly, not all teeth are retained throughout study and the number of teeth retained at a given point in study cannot, in good conscience, be considered independent of the number of teeth

present at baseline. In Section E we analyze the data from the Piedmont study using an appropriate marginalization and compare our results with previous findings.

B Informative Cluster Size

Notation

Let (Y_{ijk}, X_{ijk}) be the outcome and covariate vector of the j^{th} unit belonging to the i^{th} cluster at observation k and Q the number of clusters. In our example $i \in \{1, 2, \dots, Q\}$ would index the Q individuals under study, $j \in \{1, 2, \dots, 32\}$ would index the 32 teeth and $k \in \{1, 2, 3, 4\}$ would denote each of the four observation occasions. For the sake of simplicity throughout the rest of the paper we concern ourselves with fixed observation times such that $t_{ijk} = t_{i'j'k}$ where t_{ijk} denotes the time (in units) corresponding to the observation occasion k . Similarly, for the sake of simplicity we assume that units under study do not return once they miss an observation occasion this assumption of monotonic displacement is not necessary but does simplify notation and calculation. Let, N_i^k be the number of units belonging to the i^{th} cluster which were observed at the k^{th} observation time, N_i is the cluster size at baseline and, k_{ij} be the number of observations made on the ij^{th} unit.

1 Informative Cluster Size

We can regard ICS as dependence between outcome and cluster size. Consider a hypothetical example simpler than the periodontal data where we define Y_{ij} as a response measured on Q individuals indexed by $i \in \{1, 2, \dots, Q\}$ with N_i repeated measures indexed by $j \in \{1, 2, \dots, N_i\}$ made on each individual. Say $Y_{ij} = \alpha_i + \epsilon_{ij}$, where α_i 's are i.i.d. and $\alpha_i \sim N(\mu, \sigma^2)$, ϵ_{ij} 's are i.i.d. and $\epsilon_{ij} \sim N(0, \tau^2)$. Further assume that f is an increasing function of α_i and that $N_i = f(\alpha_i)$ *i.e.* the number of repeated measures obtained is a deterministic function of α_i taking on values in

$\{1, 2, \dots, \}$. We consider two ways of calculating the expectation of Y_{ij} :

$$\hat{\mu}_1 = \frac{\frac{1}{Q} \sum_i \sum_j Y_{ij}}{\frac{1}{Q} \sum_i N_i}, \quad (2.11)$$

and

$$\hat{\mu}_2 = \frac{1}{Q} \sum_{i=1}^Q \frac{1}{N_i} \sum_{j=1}^{N_i} Y_{ij}. \quad (2.12)$$

Because N_i is an increasing deterministic function of α_i , $E[\hat{\mu}_1] > \mu$, however, $E[\hat{\mu}_2] = E[E[Y_{ij}]] = E[\alpha_i] = \mu$. The discrepancy between Equations 2.11 and 2.12 is an example of the effects of informative cluster size and the marginalization in Equation 2.12 is analogous to that of CWGEE, Williamson, Datta and Satten (2003); also see also Benhin, Rao and Scott (2005) and Hoffman, Sen and Weinberg (2001).

A considerable amount of effort has been put into characterizing marginal analyses that will produce unbiased estimates of population effect sizes in the presence of ICS; one very popular approach is the implementation of *Weighted Generalized Estimating Equations* (WGEE's). WGEE's exploit the flexibility of the underlying standard (*i.e.* unweighted) GEE framework by reweighting individual components of the corresponding estimating equation. The intended consequence of this reweighting is the marginalization of individual contributions to the resultant estimating equation. To this point a great deal of the literature concerned with ICS considers cross sectional or repeated measure type data. Wang *et. al* (2011) extended the *Cluster Weighted Generalized Estimating Equations* (CWGEE) developed by Williamson *et al.* in order to address repeated measures data, to accommodate longitudinal data where the number of temporal measures made on each unit is constant. To that end it was shown that under certain conditions CWGEE could, through marginalization, produce unbiased estimates of the corresponding complete cluster distribution.

2 Temporally Varying Cluster Size

For the purposes of this investigation we concern ourselves with monotonic subject displacement, that is we assume that when a subject is unobservable at a given time they become unobservable at all subsequent time points. This is a natural assumption when data are collected on a mortal cohort. This is also meaningful in dental studies since a lost tooth will remain unobservable for subsequent occasions. Under this pattern of displacement the temporally varying cluster sizes will be a decreasing function of time. In special cases marginal analysis of observed temporal data can lead to unbiased estimation of the complete cluster distribution. In general, *Temporally Varying Cluster Size* (TVCS) introduces another dimension to the situation which cannot be disregarded for a proper analysis of the data. While it is often reasonable to assume that the observed information did arise from some hypothetical complete distribution we cannot simply assume that the available cluster information is a representative sample of the original complete cluster. Consider a hypothetical cluster (of reduced size) with baseline cluster size $N_1 = 100$. Assume, that the individuals ($j \in 1, 2, \dots, 100$) experience an outcome Y_{1jk} temporally over the course of 20 years. Suppose the outcome follows the model

$$Y_{1jk} = k \times \gamma_{1j} + \epsilon_{1jk},$$

with $\gamma_{1j} \sim N(0, 1)$ and $\epsilon_{1jk} \sim N(0, 1)$. This defines the distribution of a hypothetical complete cluster. Now, let us consider that each individual has the possibility of being displaced with a constant hazard proportional to $.2 \times (1 - \Phi(\gamma_{1j}))$, where $\Phi(\cdot)$ is the standard normal CDF. The implication of the dependence of the constant hazard on the cluster member specific slope (*i.e.* γ_{1j}) is that cluster members with larger positive slopes will be retained longer; on average, than those with larger negative slopes. Figure 9 plots the observed distribution given the example design and highlights how in the longitudinal setting temporally varying cluster size cannot be ignored. Looking at each of the four panels it is easy

to see how as time passes the distribution of observed information becomes more obscured from that of the underlying complete cluster. While in the first panel the observed distribution is similar to that of the complete cluster the distribution of the subjects remaining in the fourth panel cannot be considered a representative sample of the corresponding complete cluster.

Say that during a hypothetical study we observed panel 4 of Figure 9 *i.e.* 16 to 20 years. Given the mechanism responsible for subject displacement it might be possible to make inference about the complete cluster distribution. However, given that more than 60% of the original subjects are displaced by the 16th year it could be argued that the resulting distribution of retained 16 to 20 year olds would be of considerably more interest.

We introduced the previous example to underline the fact that within the longitudinal setting ICS at baseline presents a unique scenario which requires specific attention. If it is suspected that the cluster size at baseline is informative and there exists TVCS the question should be asked, is the mechanism for subject displacement partially responsible for baseline ICS and if so to what end. If the mechanisms responsible for ICS and TVCS are such that the observed subjects cease to be representative samples of the corresponding distribution from which those clusters were drawn drawing inference about the complete cluster distribution often becomes untenable.

C The Case for Scientifically Motivated Marginalization and Statistical Inference

In the marginal analysis of clustered data we assume that each cluster is drawn from some common distribution, similarly, subjects belonging to said clusters are drawn from another common distribution. The observed cluster distributions may vary but the variations are assumed to be the result of some underlying common probabilistic function. As such when we wish to estimate the distribution

Figure 9. Snapshots of various intervals over a twenty year period. Recall that $E[Y] = 0$ and note that as time progresses the estimated distribution (blue line), based on observed information gets further from the hypothetical complete distribution.

of a typical cluster or a typical subject belonging to a typical cluster the contribution of individual subjects or clusters becomes of paramount importance. If we are interested in estimating the distribution of a typical cluster it would be reasonable to conclude that each observed cluster should contribute proportional information in the estimation process thus mitigating the ability of atypically sized clusters to dominate the resulting analysis. Similarly, in the presence of TVCS if we were concerned with estimating the distribution of a typical subject belonging to a typical cluster who was observable at some point over a given interval it would be reasonable to conclude that each subject within a given cluster should be allowed to contribute proportionately to that clusters contribution to the analysis and that the sum of the marginalized subject contributions should be reweighted as to allow proportionate contribution of each cluster.

Given the often interrelated nature of ICS at baseline and TVCS, estimation of an underlying complete cluster distribution is not necessarily possible without making considerable assumptions about the form of the underlying distribution and the mechanism responsible for TVCS for examples see Seaman Pavlou and Copas (2014). We discuss in Section 1 when agreement between marginal estimation based on observed information should be in agreement with that of the hypothetical complete cluster distribution. We proceed now by justifying a number of relevant marginal analyses and motivate their implementation based on contextual relevance of the marginal distributions that they seek to describe.

1 Weighted Generalized Estimating Equations

Generalized Estimating Equations as described by Liang and Zeger (1986) provide a flexible platform for the analysis of clustered data. Through the use of a general link function GEE's are capable of modeling various non-linear outcomes as a function of linear combinations of covariates. Furthermore, GEE's accommodate the within cluster variance structure which addresses the mutually correlated

relationship of cluster specific outcomes without estimating a cluster or subject specific effect. Let, $X\beta$ define the linear component where β is a p dimensional vector of parameters, $g(\beta) = \mu = E[Y]$ define the link relating the linear component to Y such that $g^{-1} = X\beta$ and corresponding variance covariance relation as V . In the following we will consider the within subject correlation and assume that cluster members and subjects belonging to different clusters are independent, thus each subject has a specific variance covariance structure denoted by V_{ij} a $k_{ij} \times k_{ij}$ matrix and V is block diagonal with blocks defined as V_{ij} . Though it is not uncommon in practice with longitudinal data to accommodate within subject correlation structures and assume subjects are independent we choose this convention as it allows us to clearly demonstrate the application of the marginalizing weights.

Given the specified variance covariance relation we define β^{GEE} as the solution to

$$0 = \sum_{i=1}^Q \sum_{j=1}^{N_i} \frac{\partial \mu_{ij}}{\partial \beta} V_{ij}^{-1} (Y_{ij} - \mu_{ij}), \quad (2.13)$$

note that $\frac{\partial \mu_{ij}}{\partial \beta} V_{ij}^{-1} (Y_{ij} - \mu_{ij})$ is a p dimensional vector. If we wished to make marginal inference on the distribution of a typical cluster we could consider the following marginalization:

$$0 = \sum_{i=1}^Q \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{\partial \mu_{ij}}{\partial \beta} V_{ij}^{-1} (Y_{ij} - \mu_{ij}). \quad (2.14)$$

Provided that each subject experiences the same number of temporal observations or it could reasonably be assumed that the mechanism responsible for TVCS was negligible this marginalization has the interpretation of describing the distribution of a typical subject belonging to a typical cluster. This is the CWGEE marginalization proposed by Williamson *et al.* in 2003.

A shortcoming of this approach as pointed out by Huang and Leroux (2011) is that it does not properly marginalize with respect to the distribution of cluster member covariates. In other words it does not accommodate potentially informative covariate distributions (*e.g.* interdependence between outcome, cluster size and the

value of the observed covariate vector). Say, that there exists a binary covariate such that clusters could be partitioned into sub-clusters according to exposure. Further assume that the distribution of exposure between clusters varies such that there is reasonable suspicion that it is in some manner related to outcome (recall that we are interested in estimating the effect of exposure on the outcome so this supposition is not grandiose). Let, $Z_i^{r=0}$ and $Z_i^{r=1}$ denote the number of individuals belonging to the i^{th} cluster having exposure (r) equal to 0 and 1 respectively. Then, if both levels of exposure are present in each cluster we could consider the following marginalization:

$$0 = \sum_{i=1}^Q \sum_{j=1}^{N_i} \left(\frac{1}{Z_i^{r_{ij}}} \right) \frac{\partial \mu_{ij}}{\partial \beta} V_{ij}^{-1} (Y_{ij} - \mu_{ij}). \quad (2.15)$$

This marginalization allows exposed and unexposed cluster members to contribute proportionately and has the interpretation of estimating the distribution of a typical subject belonging to a typical cluster conditional on exposure. This is the DWGEE1 described by Huang and Leroux in 2011. It should be noted that this marginalization can be extended to address factor variables of several levels.

However, when the factor level distribution is such that not all clusters experience all levels of the factor the marginal interpretation of a *typical* cluster is more subjective as atypical clusters contribute more or less to the resulting estimate.

In the treatment of TVCS we propose a similar marginalization to Equation 2.14. If we wish to estimate the distribution of a typical subject belonging to a typical cluster who was observable at the beginning of a given interval as for example, say that we had a number of clusters drawn from distributions similar to that of the subjects in panel 4 of Figure 9. Marginalizing over the number of observations made on each subject has the interpretation of conditioning on the fact that a subject was observable at the onset of study such that each individual, regardless of how long they were observed, would contribute proportional information to the estimated distribution. If we then marginalized over the cluster

size each observable member would contribute proportional information to the cluster contribution and each cluster would contribute proportional information to the overall estimate. We define this marginalization as:

$$0 = \sum_{i=1}^Q \frac{1}{N_i} \sum_{j=1}^{N_i} \left(\frac{1}{k_{ij}} \right) \frac{\partial \mu_{ij}}{\partial \beta} V_{ij}^{-1} (Y_{ij} - \mu_{ij}). \quad (2.16)$$

The marginalization in 2.16 has the interpretation of estimating the distribution of the typical subject belonging to a typical cluster during a given interval conditional on observability at the onset of the interval. Note that 2.16 makes no assumptions about the mechanism responsible for TVCS. It instead seeks only to describe the observed marginal distribution and while Equations 2.15 and 2.16 appear similar they address fundamentally different situations.

Addressing TVCS a little more formally, assume that the distribution of temporal profiles of subjects belonging to a given cluster is known to be dependent on some observable component (i.e. cluster size). That is, given observed information the temporal profile of a typical subject belonging to a given cluster is a probabilistic function of observed information. It would then be reasonable to marginalize with respect to the observed temporal profile. Since it is unlikely that every possible temporal profile will be observed in each cluster we can mitigate the under contribution of clusters with fewer temporal profiles by calculating the marginal weights as the expected number of observations belonging to a given temporal profile. Let, $l_i^{r=k_{ij}}$ be the expected number of observations made on subjects belonging to the i^{th} cluster experiencing the same temporal profile as the j^{th} individual (recall that here we concern ourselves only with monotonic displacement). Then we define this marginalization as:

$$0 = \sum_{i=1}^Q \sum_{j=1}^{N_i} \left(\frac{1}{l_i^{k_{ij}}} \right) \frac{\partial \mu_{ij}}{\partial \beta} V_{ij}^{-1} (Y_{ij} - \mu_{ij}). \quad (2.17)$$

This marginalization has the interpretation of estimating the distribution of a

typical subject belonging to a typical cluster conditional on the expected TVCS distribution. It should be noted that Equation 2.17 is similar in spirit to the DWGEE2 marginal analyses proposed by Huang and Leroux (2011) extended to the longitudinal setting. A drawback to the use of this marginalization (Equation 2.17) is that assumptions must be made as to the relationship between the outcome, covariate vector and resultant cluster sizes. Similarly, sparsity in the distribution of temporal profiles can make estimation of $l_i^{k_{ij}}$ untenable. Furthermore, misspecification of the relationship between the outcome, covariate vector and observed temporal profile can obscure the validity of the intended scientific marginalization.

2 Variance of WGEE's

The sandwich estimator of the variance of β^{GEE} for clustered longitudinal data can be obtained by,

$$\Sigma^{GEE} = \Omega^{-1} M \Omega^{-1}, \quad (2.18)$$

with,

$$M = \sum_{i=1}^Q \left(\left\{ \sum_{j=1}^{N_i} \frac{\partial \mu_{ij}}{\partial \beta} V_{ij}^{-1} \epsilon_{ij} \right\} \left\{ \sum_{j=1}^{N_i} \frac{\partial \mu_{ij}}{\partial \beta} V_{ij}^{-1} \epsilon_{ij} \right\}^T \right), \quad (2.19)$$

where $\Omega = \sum_{i=1}^Q \sum_{j=1}^{N_i} \left(\frac{\partial \mu_{ij}}{\partial \beta} V_{ij}^{-1} \left(\frac{\partial \mu_{ij}}{\partial \beta} \right)^T \right)$. We propose a generalized sandwich estimator applicable to WGEE's for drawing inference about parameters obtained via marginal analysis. Let, ω_{ij} be the diagonal matrix of marginal weights applied to the j^{th} subject belonging to the i^{th} cluster. For example, in Equation 2.16, $\omega_{ij} = \text{diag} \left\{ \frac{1}{N_i k_{ij}}, \dots, \frac{1}{N_i k_{ij}} \right\}$. We can then express the general sandwich estimator for WGEE's as,

$$\Sigma^{WGEE} = \tilde{\Omega}^{-1} \tilde{M} \tilde{\Omega}^{-1}, \quad (2.20)$$

with

$$\widetilde{M} = \sum_{i=1}^Q \left(\left\{ \sum_{j=1}^{N_i} \frac{\partial \mu_{ij}}{\partial \beta} (\omega_{ij} V_{ij}^{-1}) \epsilon_{ij} \right\} \left\{ \sum_{j=1}^{N_i} \frac{\partial \mu_{ij}}{\partial \beta} (\omega_{ij} V_{ij}^{-1}) \epsilon_{ij} \right\}^T \right), \quad (2.21)$$

where $\widetilde{\Omega} = \sum_{i=1}^Q \sum_{j=1}^{N_i} \left(\frac{\partial \mu_{ij}}{\partial \beta} (\omega_{ij} V^{-1}) \left(\frac{\partial \mu_{ij}}{\partial \beta} \right)^T \right)$. The asymptotic normality of β^{WGEE} follows as a consequence of the independence of the block sums of the WGEE estimating equation. We defer the algorithmic details of parameter estimation to Section F.

D Simulation

For each of the four marginal analyses proposed in Section C we present an applicable design of which inference about the implied marginal distribution would be of interest. In each case we report the mean, bias, MSE and variance of the marginal parameter β obtained from a Monte Carlo simulation in which we calculate both the standard GEE and the relevant WGEE estimates. In each design we report the results of only one marginal WGEE analysis as designs were constructed case specific to each of the marginal WGEE analyses described in Section C.

Designs

I. Longitudinal Poisson Model : Fixed Temporal Profile

Let, $Y_{ijk} \sim \text{Pois}(\eta_{ijk})$ where $\eta_{ijk} = \exp\{X_{ijk}\beta + \alpha_i\}$ and $\beta = (-.5, 1, .25)$ corresponding to the intercept, a cluster constant binary factor, and time effect, respectively, and $\alpha_i \sim N(0, .4^2)$ a cluster specific random intercept.

Cluster size, N_i , for each of the Q clusters is determined by $N_i = \lfloor N_i^* \rfloor$ where $N_i^* = (\Phi(\alpha_i) \times 19) + 1$. Of the Q clusters half are exposed *i.e.* $X_1 = 1$ and each observed subject is observed over four time points *i.e.* ,

$X_2 \in \{1, 2, 3, 4\} \forall i, j$. Given the fact that baseline cluster size cluster size is dependent on the outcome the marginal analysis proposed by Equation 2.14 is

an appropriate candidate for estimating the marginal parameter β .

II. Binomial Model Cluster Varying Binary Exposure: Non-temporal

Let, $Y_{ij} \sim \text{ber}(\eta_{ijk})$ where $\eta_{ijk} = \exp\{X_{ijk}\beta + \alpha_i\} / (1 + \exp\{X_{ijk}\beta + \alpha_i\})$ and $\beta = (-.5, 2)$ corresponding to; the intercept and cluster varying binary factor, respectively, and $\alpha_i \sim N(0, .5^2)$ is a cluster specific random intercept.

Subcluster size (corresponding to the number of exposed and unexposed subjects in each cluster) is determined by $N_i^{X_1=0}$ and $N_i^{X_1=1}$ where $N_i^{X_1=0} \sim \text{bin}(9, p_i) + 1$ and $N_i^{X_1=1} \sim \text{bin}(9, p_i/2) + 1$ and $p_i = \Phi(\alpha_i)$. Given the fact that the number of exposed and unexposed units (and as a consequence baseline cluster size) is dependent on the outcome the marginal analysis proposed by Equation 2.15 is an appropriate candidate for estimating the marginal parameter β .

III. Longitudinal Binomial Model: TVCS Distribution Unknown

Let, $Y_{ijk} \sim \text{ber}(\eta_{ijk})$ where $\eta_{ijk} = \exp\{X_{ijk}\beta + \alpha_i\} / (1 + \exp\{X_{ijk}\beta + \alpha_i\})$ and $\beta = (-.5, 1, .25)$ corresponding to; the intercept, a cluster constant binary factor, and time effect, respectively and $\alpha_i \sim N(0, .4^2)$ is a cluster specific random intercept. Cluster size, N_i , for each of the Q clusters is determined by $N_i = \lfloor N_i^* \rfloor$ where $N_i^* = (\Phi(\alpha_i) \times 19) + 1$. Half of the Q clusters are exposed *i.e.* $X_1 = 1$, but the number of temporal observations made on a given subject is determined by k_{ij} where $k_{ij} \sim \text{bin}(3, \Phi(\alpha_i)) + 1$. Given the fact that baseline cluster size and the number of temporal observations made on a unit is dependent on the outcome (though the relationship is assumed to be unknown) the marginal analysis proposed by Equation 2.16 is an appropriate candidate for estimating the marginal parameter β .

IV. Gaussian Model - TVCS Distribution Known

Let, $Y_{ijk} \sim N(X_{ijk}\beta + \alpha_i + \gamma_{ij}, .1)$ where $\alpha_i \sim N(0, 1)$ and $\gamma_{ij} \sim N(0, .5^2)$.

The components of $\beta = (.5, 1.5, .75, .8)$; corresponding to the intercept, time

effect, cluster constant exposure, and exposure time interaction with half of the Q clusters having exposure $X_2 = 1$. The cluster size at baseline *i.e.* $N_i^1 \sim Pois(exp\{\lambda_i\})$ where $\lambda_i = 1 + \alpha_i + X_{2,i}$. The number of temporal observations made on each subject varies between 2 and 4 with the individual number of observations determined by the ordinal logit model with probabilities;

$$\begin{aligned}\pi_{k_{ij}=2} &= \frac{e^{-.01 \times N_i^1}}{1 + e^{-.01 \times N_i^1}}, \\ \pi_{k_{ij}=3} &= 1 - \pi_1 - \frac{1}{1 + e^{1.5 - .01 \times N_i^1}} \text{ and} \\ \pi_{k_{ij}=4} &= \frac{1}{1 + e^{1.5 - .01 \times N_i^1}}.\end{aligned}$$

Given the fact that baseline cluster size and the number of temporal observations made on a unit is dependent on the outcome (relationship known) the marginal analysis proposed by Equation 2.17 is an appropriate candidate for estimating the marginal parameter β .

For each of the designs (I-IV) we conduct two Monte Carlo simulations generating 500 data sets from the corresponding distributions, one with $Q = 50$ and another with $Q = 200$. For each iteration we calculate the standard GEE estimate as well as the applicable marginal WGEE estimate of the corresponding distribution. We report the summary results of the simulation in Tables 6 and 7.

TABLE 6

Summary of simulation results from Designs I and II with $M = 500$ and $Q = (50, 200)$, using independence working correlation. Note that in Design I we have a case ICS without TVCS making Equation 2.14 an appropriate marginal analysis. Whereas, Design II there is no temporal component of the data but, ICS is present and there is a within cluster varying binary factor making Equation 2.15 an appropriate analysis.

	Design I ($Q = 50$)				Design II ($Q = 50$)			
	GEE		$\sum_{i=1}^Q \frac{1}{N_i} \sum_{j=1}^{N_i} U_{ij}$		GEE		$\sum_{i=1}^Q \sum_{j=1}^{N_i} \frac{1}{Z_i} U_{ij}$	
	β_0	β_1	β_2		β_0	β_1	β_0	β_1
Mean	-0.2452	1.0035	0.2499		-0.2624	1.9022	-0.4749	1.9325
Bias	0.2548	0.0035	-0.0001		0.2376	-0.0978	0.0251	-0.0675
MSE	0.0741	0.0169	0.0001		0.0752	0.0719	0.0258	0.0909
Var	0.0091	0.0169	0.0001		0.0187	0.0623	0.0252	0.0863
Design I ($Q = 200$)								
	GEE		WGEE		GEE		WGEE	
	β_0	β_1	β_2		β_0	β_1	β_0	β_1
Mean	-0.2585	1.0227	0.2499		-0.2542	1.8879	-0.4704	1.9013
Bias	0.2415	0.0227	-0.0001		0.2458	-0.1121	0.0296	-0.0987
MSE	0.0671	0.0101	< 0.0001		0.0656	0.0281	0.0080	0.0311
Var	0.0087	0.0096	< 0.0001		0.0051	0.0155	0.0071	0.0213

TABLE 7

Summary of simulation results from Designs III and IV (M=500, Q=(50, 200), using independence working correlation. Note that Design III incorporates ICS and TVCS, thus Equation 2.16 is an appropriate marginal analysis. In the analysis of Design IV we use the knowledge that TVCS is determined via the ordinal logit model, dependent on cluster size in order to estimate the expected values of $l_{ij}^{k_{ij}}$ in Equation 5.

	Design III ($Q = 50$)										Design IV ($Q = 50$)									
	GEE					$\sum_{i=1}^Q \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{1}{k_{ij}} U_{ij}$					GEE					$\sum_{i=1}^Q \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{1}{k_{ij}} U_{ij}$				
	β_0	β_1	β_2	β_3	β_4	β_0	β_1	β_2	β_3	β_4	β_0	β_1	β_2	β_3	β_4	β_0	β_1	β_2	β_3	β_4
	Mean	Bias	MSE	Var		Mean	Bias	MSE	Var		Mean	Bias	MSE	Var		Mean	Bias	MSE	Var	
Design III ($Q = 200$)																				
GEE					$\sum_{i=1}^Q \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{1}{k_{ij}} U_{ij}$					GEE					$\sum_{i=1}^Q \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{1}{k_{ij}} U_{ij}$					
β_0	β_1	β_2	β_3	β_4	β_0	β_1	β_2	β_3	β_4	β_0	β_1	β_2	β_3	β_4	β_0	β_1	β_2	β_3	β_4	
-0.3081	0.9850	0.2503	0.0003	0.0016	-0.5007	0.9683	0.2511	0.0011	0.0044	0.8624	1.5020	0.8328	0.8067	0.0065	0.4852	1.5154	0.7565	0.7920	0.0080	
0.1919	-0.0150	0.0003	0.0007	0.0011	-0.0007	-0.0317	0.0011	0.0016	0.0044	0.3624	0.0020	0.0828	0.0067	0.0038	-0.0148	0.0154	0.0065	-0.0080	0.0065	
0.0532	0.0122	0.0016	0.0319	0.0198	0.0319	0.0198	0.0044	0.0044	0.0044	0.1971	0.0027	0.1212	0.0038	0.0037	0.0505	0.0033	0.0970	0.0042	0.0042	
0.0163	0.0120	0.0016	0.0319	0.0188	0.0319	0.0188	0.0044	0.0044	0.0044	0.0655	0.0027	0.1144	0.0037	0.0037	0.0503	0.0030	0.0970	0.0042	0.0042	

1 Simulation Results

In Designs I-IV the mechanism responsible for ICS was dependent on a cluster specific random intercept. As a consequence unweighted GEE estimates (standard GEE) of the corresponding distributions based on observed information produces biased estimates of intercept parameters (β_0 in I-III, β_0 and β_2 in IV). Appropriate marginal analyses return estimates more in line with the marginal parameters of interest. Take for example Design III, Standard GEE analysis produces notable bias in the estimate of β_0 corresponding to a relative bias of about 50% and 30% in the $Q = 50$ and $Q = 200$ settings. The proposed WGEE (Equation 2.16) in contrast reduces the relative bias to about 3% and $< 1\%$, the variance of the WGEE β_0 estimates are larger than those of the unweighted Standard GEE. This reduction in bias and increase in variance is precisely the desired behavior pointing to the compromise of the bias variance trade off.

Each of the designs was constructed such that standard GEE estimates of the intercept and or binary effect would exhibit notable bias (large relative bias). In practice we can construct situations to place intentional bias on any of a number of parameters, however, we constructed the designs to those with bias associated with one or two parameters as it kept the distributions simple in comparison. There are situations in which a small amount of bias occurred inadvertently (like that exhibited by the β_1 estimates of both the standard GEE and Marginal WGEE analyses of Design III) in such instances WGEE performs comparably or slightly better than the standard GEE and in no such case does the relative bias of the WGEE estimate exceed 3.2%.

It should be noted that Designs I-IV describe distributions where each cluster regardless of size may be regarded as a representative sample drawn from the corresponding population of clusters. Similarly, the mechanisms responsible for TVCS in designs III and IV were such that they controlled the amount of temporal information collected on each subject without obscuring the representative sample

of subjects. This is in contrast to the situation described in panels 2-4 of Figure 9.

When the mechanism responsible for ICS is such that the observable subjects can be regarded as an unrepresentative subsample of the corresponding complete cluster estimation of the corresponding complete cluster distribution will be untenable by these methods. In such cases though, the interpretation of appropriate marginal analyses still holds provided emphasis is placed on describing the resulting distribution as belonging to a typical subject belonging to a typical cluster given the observability of subjects/clusters. The latter part of the interpretation of marginal observed cluster analyses is purely contextual, say for example the data consists of periodontal outcomes of teeth belonging to elderly patients between the ages of 60-65 then the resulting marginal observed cluster estimates of the corresponding distribution would describe the expected outcome for a typical tooth belonging to a typical elderly individual.

The pitfalls of extrapolation are well known to statisticians and clinicians alike. When conducting marginal analysis based on observable information one can, unwittingly, find themselves guilty of extrapolation by virtue of simply misinterpreting the distribution which the analysis is intended to estimate. Thoughtful investigation of the mechanisms responsible for ICS and TVCS can inform an analyst as to whether conditions exist such that the results of a marginal observed cluster analysis should be in agreement with the complete cluster distribution. Even in such cases, interpreting a marginal observed cluster analysis as describing the corresponding complete cluster distribution not advisable. We recommend instead that such marginal analyses be strictly interpreted as describing the typical observed members of typical clusters.

E Periodontal Application

We illustrate our method by analyzing data collected by Beck *et. al* (1997a). Over 1000 elderly patients residing in the Piedmont Region of North Carolina

participated in the study, for each of the participants various measures of dental health were collected on available teeth including Attachment Loss (AL), Gingival Recession and Pocket Depth. Similarly, demographic and lifestyle information was collected on each individual at baseline which included sex, race (Black and White), tobacco use, level of education obtained, length of time since last visit to dentist, marital status, socioeconomic information and whether an individual resided in an urban community; the reader is directed towards Beck *et. al* (1997a) for a more detailed accounting of the study design. Beck *et. al* (1997a), Beck *et. al* (1997b) and Wang *et. al* (2011), among others, have analyzed these data in an effort to describe the conditional dependence of Attachment Loss on various demographic and lifestyle factors .

Of the more than 818 dentulous subjects followed we retain the 452 dentulous individuals which were observed at the first and second observation occasions (0 and 18 months) and which did not miss a subsequent observation occasion, *i.e.*, if an individual was observed at the fourth occasion they were also observed at the third occasion. As such, the number of temporal observations made on each individual and each tooth varies. For further clarification Figure 10 is included showing the distributions of teeth at baseline, the number of teeth summarized by total number of observations contributed and the number of records contributed by each individual. In particular, Panel (c) of Figure 10 illustrates the contrast between the number of records each individual contributes to the analysis.

We retain 452 individuals in our analysis with 7823 teeth for a total of 25,183 records. The 452 retained individuals correspond to those who were observed at both baseline, the 18 month period and who's subsequent observations followed a monotonic displacement pattern, that is to say that we retained those subjects which did not miss an observation occasion and return to study. In this particular investigation the choice to retain only the subjects with monotonic displacement was made to simplify estimation of the correlation parameters. This is in contrast to

Figure 10. Graphical representation of the available data.

- (a) Baseline cluster size of each of the 452 individuals retained.
- (b) Number of observations made on each of the 7,823 retained teeth.
- (c) Total number of observations contributed by each individual.

the investigation conducted by Wang *et. al* (2011) where only 74 individuals, which were observed at all 4 observation occasions without loss of teeth, were retained for analysis. Given the range of number of observations contributed by each individual, the range of the number of retained teeth and the distribution of the number of observations made on each tooth we consider the marginalization of Equation 2.16 instead of that of Equation 2.17 to avoid the aforementioned pitfalls associated with misspecifying the corresponding relationships and citing concerns for sparsity issues.

In our investigation we consider three analyses, standard GEE, CWGEE and the marginal WGEE described in Equation 2.16 , in each we employ a racially stratified analysis. Historically, analysis of these data has been stratified by race; Beck *et. al* (1997a) and Beck *et. al* (1997b). Rationalization for investigating the Black and White participants separately comes from disparities between the distributions of demographic, socio-economic and educational information of the two groups. For example, if we compare the mean years of education attained for the 215 Blacks which were retained in our analysis to that of the 237 retained Whites the mean of the Black group is 3.66 years higher than that of the White group, similarly the retained Black participants, on average, visited the dentist more frequently, had higher socio-economic indicators and were less likely to use tobacco than the retained White participants.

To further motivate marginal analysis Figure 11 is included to underline the relationship between number of teeth at baseline, tooth retention and AL. The discrepancies between Panels (b) and (c) of Figure 11 convey the complexity of the relationship between AL and the availability of information.

We model AL using a weighted generalized estimating equation where we assume the identity link is appropriate and accommodate within tooth temporal correlation assuming that an AR-1 correlation structure, an independence working correlation structure between teeth belonging to the same individual was assumed in the estimation of $\hat{\beta}^{WGEE}$. The weighting employed will be CWGEE (Equation

Figure 11. Distribution of attachment loss of teeth stratified by number of teeth at baseline and total number of observations made one each tooth.

(a) Distributions of AL for teeth belonging to individuals with less than 10 teeth vs. individuals with more than 23 teeth.

(b) Distributions of AL for teeth belonging to individuals with less than 10 teeth by number of observations.

(c) Distributions of AL for teeth belonging to individuals with more than 23 teeth by number of observations.

2.14) and that as described in Equation 2.16. We motivate the use of the marginalization described in Equation 2.16 upon consideration of the relationship between the number of teeth at baseline, the number of temporal observations made on a given tooth and AL described in Figure 11. Use of the CWGEE marginalization is difficult to motivate as there is not a one to one relation between the number of records each individual contributed and the number of teeth present at baseline, however it is included for comparison. With the WGEE weighting scheme each observed tooth contributes proportionate information to the subject which it belongs contribution to the overall estimating equation and the corresponding subjects contribution is weighted such that each subject contributes proportionately to the overall estimating equation.

For our racially stratified analysis we investigate the following linear model:

$$\begin{aligned}
E[AL] = & \beta_0 + \beta_1 T + \beta_2 SEX + \beta_3 EDUC + \\
& \beta_4 WHENDDS + \beta_5 MARRIED + \beta_6 SEI + \beta_7 TOBACCO + \\
& \beta_8 Urban + \beta_9 INCOME + \beta_{10}(SEX \times T) + \\
& \beta_{11}(EDUC \times T) + \beta_{12}(WHENDDS \times T) + \\
& \beta_{13}(MARRIED \times T) + \beta_{14}(SES \times T) + \beta_{15}(TOBACCO \times T) + \\
& \beta_{16}(Urban \times T) + \beta_{17}(INCOME \times T).
\end{aligned} \tag{2.22}$$

The factor effects SEX, MARRIED, TOBACCO and URBAN in Equation 2.22 are coded 0 and 1 such that , SEX= 1 implies the subject is male, and URBAN= 1 implies that a subject resided in an urban community. Here we model AL of tobacco users where $TOBACCO = 0$ corresponds to the expected AL of tobacco users. We report only the effect sizes of the significant effects, where significance is determined by a Wald test with corresponding variance estimates obtained according to Equation 2.18 in the case of the GEE analysis and Equation 2.20 in the case of CWGEE and WGEE analyses. We remit the details of parameter

estimation to the appendix. The resulting marginal WGEE model describes the expected AL of a typical tooth belonging to a typical individual sampled from the corresponding population i.e. an estimated TOBACCO use effect size of -1.4895 corresponding to the White population would indicate that in a typical examination on a typical tooth belonging to a typical individual in the White population who does not use tobacco would, on the average, have an AL value 1.4895 lower than a similar tooth belonging to tobacco users.

Piedmont Results

Table 8 presents the results of the racially stratified GEE, CWGEE and marginal WGEE analyses.

TABLE 8

Aggregate results of racially stratified GEE, CWGEE and marginal WGEE analyses of AL. In each analysis an AR-1 working correlation was assumed. The effects reported are those significant according to a Wald test (covariance estimates obtained via sandwich and weighted sandwich) at the $\alpha \leq .05$ level.

White					Black				
Standard GEE					Standard GEE				
Effect	$\hat{\beta}$	std. err	Wald	p	Effect	$\hat{\beta}$	std. err	Wald	p
Intercept	7.6862	0.5602	188.2553	< 0.0001	Intercept	6.0179	0.3930	234.4765	< 0.0001
SEX	-1.1379	0.3786	9.0317	0.0027	SEX	-0.6347	0.2892	4.8166	0.0282
SES	-0.0026	0.0011	5.3807	0.0204	TOBACCO	-1.0437	0.2971	12.3446	0.0004
TOBACCO	-1.3511	0.3798	12.6535	0.0004					
URBAN	-0.9821	0.3601	7.4393	0.0064					
CWGEE					CWGEE				
Effect	$\hat{\beta}$	std. err	Wald	p	Effect	$\hat{\beta}$	std. err	Wald	p
Intercept	8.0550	0.5924	184.8922	< 0.0001	Intercept	5.9325	0.3969	223.4071	< 0.0001
SEX	-1.3693	0.4073	11.3009	0.0008	TIME	0.1518	0.0768	3.9084	0.0480
SES	-0.0030	0.0012	5.7646	0.0164	TOBACCO	-0.9692	0.2935	10.9042	0.0010
TOBACCO	-1.5711	0.3986	15.5339	0.0001	TIME $\times$ EDUCATION	-0.0204	0.0102	4.0123	0.0452
URBAN	-1.1109	0.3797	8.5578	0.0034					
Marginal WGEE					Marginal WGEE				
Effect	$\hat{\beta}$	std. err	Wald	p	Effect	$\hat{\beta}$	std. err	Wald	p
Intercept	8.8568	0.6646	177.5851	< 0.0001	Intercept	6.1878	0.4760	169.0242	< 0.0001
SEX	-1.4392	0.5306	7.3577	0.0067	TIME	0.2110	0.0820	6.6231	0.0101
SES	-0.0032	0.0016	3.9442	0.0470	TOBACCO	-0.8630	0.3825	5.0908	0.0241
TOBACCO	-1.4895	0.5115	8.4790	0.0036	TIME $\times$ EDUCATION	-0.0233	0.0105	4.8908	0.0270
URBAN	-1.6101	0.5309	9.1971	0.0024					

The results of the standard GEE analysis, describe the expected AL of a typical tooth belonging to the pool of all teeth in the samples of White/Black individuals. Thus, individuals with more teeth and consequently on average lower values of AL are inadvertently oversampled. It can be argued that since teeth belong to a natural clustering unit *i.e.* the individual, that the expected AL of a typical retained tooth belonging to a typical White/Black individual is of considerably more relevance than that of the former. The marginal WGEE analyses results of Table 8 describe the expected AL experienced by a typical tooth belonging to a typical individual from each of the corresponding sample populations. To clarify, the GEE analyses provides the expected AL of a tooth drawn at random from the pool of teeth consisting of all teeth belonging to the corresponding sample population. Whereas, the WGEE analyses provides the expected AL of a typical retained tooth belonging to a typical individual drawn at random from the corresponding sample population.

While the effect of tobacco use is consistently significant across all six analyses the effect size varies from one analysis to another. In the White sample the same effects remain significant across all three analyses, however, there are marked discrepancies between the effect sizes in each. We see an increase in the estimated Intercept from 7.6862 to 8.8568, the estimated effect size of URBAN dwelling decreases from -0.9892 to -1.6101 , the effect size of SEX decreases from -1.1379 to -1.4392 and TOBACCO use decreases from -1.3511 to -1.4895 when comparing the GEE analysis to the WGEE analysis. Such a result is consistent with expectation given the interrelated nature of number of teeth at baseline, number of teeth present at a given observation occasion, AL and various demographic information. The Black sample tells a slightly different story, while there are still differences in the estimated effect sizes TIME and the $\text{TIME} \times \text{EDUCATION}$ interaction are significant in the Black WGEE analysis whereas they are not in the standard GEE (and neither are significant in either the standard GEE or WGEE

analysis of the White population). In a similar vein, the difference in the estimated effect size of TOBACCO use from the standard GEE to the WGEE analysis in the Black population is in the opposite direction from that of the White population. Figure 12 is included to outline a few discrepancies between the distribution of demographic information between the two samples. Note in Figure 12 we can see that the Black population on average visited the dentist more frequently (b), had more education (a), higher socioeconomic status (c) and had higher incomes (d) than their White counterparts. It is in light of these discrepancies that we defend disparities between estimated effect sizes and significant effects for each population.

F Discussion

We have provided a number of WGEE marginalizations as well a sandwich estimate of the corresponding variance of β^{WGEE} for clustered longitudinal data. The GEE framework is well established and it's use is widely supported in the analysis of clustered data. We have demonstrated through simulation and the analysis of data from an observational study how the marginal distribution of various populations can differ. We have shown that in certain circumstances it is possible to describe the hypothetical balanced data distribution from available data through the use of marginal WGEE modeling. To a more direct end we demonstrate that, contextually, the hypothetical balanced data distribution may be of little interest. We advocate the thoughtful application of marginal WGEE analyses in situations where there is reasonable concern about ICS or TVCS and present marginal WGEE's as a logical and necessary extension to the existing GEE framework.

In our present investigations we have considered only data with units experiencing monotonic temporal displacement, though, such a condition is not necessary for the application of a marginal WGEE investigation. Depending on the desired marginalization the unrestricted retention of units experiencing any of a

Figure 12. Distribution of various demographic information stratified by race. All values have been centered according to the respective overall sample means.

number of combinatorial arrangements of observations occasions can lead a number of problems in the estimation of marginal WGEE parameters. The first issue worth considering is the estimation of the correlation parameters, when all retained subjects experience continuous observation a first order approximation is appropriate. In situations where retained units are allowed to miss observations more complicated estimation methods must be employed to acquire appropriate estimates. The second being that sparsity issues can arise if there exist a number of units experiencing unique observation profiles.

We understand that coding individual analysis specific algorithms to implement marginal WGEE analyses is a major limiting factor to its wide spread acceptance. As such in various settings we have repeated a number of the analyses from this investigation in both SAS and R. There does not appear to be an existing function/package in R for appropriately addressing marginal WGEE analyses, however, use of the `< weights= >` option in Proc Genmod provides parameter estimates remarkably similar to those obtained using our formulation. It is also very important to note that often the working correlation structure is misspecified (as was the case in our applied analyses and a number of our simulations). Such a misspecification (ignoring dependencies between units) will result in smaller variance estimates for the marginal parameters if the incorrect dependencies are extended to the variance covariance calculation. One way to mitigate this issue is to write a case specific function, another is to bootstrap the variance estimator sampling the largest unit with dependencies. In our periodontal example this means sampling with replacement from each racially stratified population, obtaining estimates in this manner provides a reasonable approximation to the correct variances.

The present marginal WGEE approach is recommended as an alternative to fully parametric approaches such as joint modeling. Joint modeling while effective requires that the analyst make a number of additional assumptions regarding the distributions giving rise to the observed information. The quality of joint model

analyses is dependent on the tenability of these extra assumptions see, Chen, Zhang and Albert (2011) and Dunson, Chen and Harry (2003). In future works we intend to investigate the robustness of marginal WGEE inference as well as make side by side comparisons with joint model approaches.

Supplementary Materials

Details of WGEE Estimation

We obtain β^{WGEE} estimates through the implementation of an augmented Newton-Raphson algorithm. We begin by setting V_{ij} to the identity matrix and the initial $\hat{\beta}$ estimate is obtained as

$$\beta^0 = H_0^{-1} \left(\frac{\partial \mu}{\partial \beta} \right)_0$$

where,

$$H_0 = \sum_{i=1}^Q \sum_{j=1}^{N_i} X_{ij}^T \left(\text{diag}(\omega_{ij}) V_{ij}^{-1} \right) X_{ij},$$

and

$$\left(\frac{\partial \mu}{\partial \beta} \right)_0 = \sum_{i=1}^Q \sum_{j=1}^{N_i} X_{ij}^T \left(\text{diag}(\omega_{ij}) V_{ij}^{-1} \right) Y_{ij}.$$

After obtaining β^0 we estimate the correlation parameter (α_0) as the stage two estimator of the AR-1 correlation parameter as outlined in Wang *et. al* (2011), details provided in Appendix B. Estimation then proceeds by alternating between updating β and α as

$$\beta^{n+1} = \beta^n + H_n^{-1} \left(\frac{\partial \mu}{\partial \beta} \right)_n,$$

where,

$$H_{n+1} = \sum_{i=1}^Q \sum_{j=1}^{N_i} X_{ij}^T \left(\text{diag}(\omega_{ij}) V_{ij}(\alpha)^{-1} \right) X_{ij},$$

and

$$\left(\frac{\partial \mu}{\partial \beta}\right)_{n+1} = \sum_{i=1}^Q \sum_{j=1}^{N_i} X_{ij}^T (\text{diag}(\omega_{ij}) V_{ij}(\alpha)^{-1}) (Y_{ij} - X_{ij} \beta^n),$$

and $\alpha_n = f(\beta^n)$.

In our estimation of β^{WGEE} we update β and α until

$$\sum_{q=1}^p \left| \left(H_n^{-1} \left(\frac{\partial \mu}{\partial \beta} \right)_n \right)_q \right| < .000001.$$

Estimation of α , AR-1 Correlation (Gaussian Link)

Let $z_{ijk} = Y_{ijk} - X_{ijk} \beta$ then we can define our estimator of α as

$$\hat{\alpha} = \frac{2(\zeta - \eta)}{1 + (\zeta - \eta)^2},$$

where

$$\zeta = \frac{\sum_{i=1}^Q \sum_{j=1}^{N_i} \sum_{k=2}^{k_{ij}} \omega_{ijk} (z_{ijk}^2 + z_{ijk-1}^2)}{\sum_{i=1}^Q \sum_{j=1}^{N_i} \sum_{k=2}^{k_{ij}} \omega_{ijk} (z_{ijk} z_{ijk-1})},$$

and

$$\eta = \frac{\sqrt{\sum_{i=1}^Q \sum_{j=1}^{N_i} \sum_{k=2}^{k_{ij}} \omega_{ijk} (z_{ijk} + z_{ijk-1})^2 \sum_{i=1}^Q \sum_{j=1}^{N_i} \sum_{k=2}^{k_{ij}} \omega_{ijk} (z_{ijk} - z_{ijk-1})^2}}{2 \sum_{i=1}^Q \sum_{j=1}^{N_i} \sum_{k=2}^{k_{ij}} \omega_{ijk} (z_{ijk} z_{ijk-1})}.$$

CHAPTER III

PERFORMANCE OF WEIGHTED GENERALIZED ESTIMATING EQUATIONS WITH RESPECT TO MARGINAL MODEL ESTIMATION COMPARISON WITH JOINT MODELING

A Introduction

In spite of the growing popularity of the use Weighted Generalized Estimating Equations (WGEE) to address the analysis of imbalanced clustered data, WGEE's have yet to be adopted widely as a standard method of analysis. In Chapter II we introduced a semi-parametric WGEE weighting scheme in order to address induced bias resulting from informative cluster sizes and temporal observation profiles within the clustered longitudinal data setting associated with marginal model estimation. The utility of Marginal WGEE (MWGEE) models comes from the complexity of the standard alternative, Maximum Joint Likelihood estimation. In Chapter II we demonstrated the efficacy of MWGEE's at mitigating bias. Here we present a side by side comparison of the performance of our MWGEE's with that of an alternative Joint Modeling approach. In this chapter we will demonstrate, through simulation, the small sample properties of each and discuss various practical details associated with implementation.

B Joint Modeling - Full Joint Likelihood

Parametric modeling approaches often exploit the form of some underlying probabilistic distribution which is assumed to have given rise to the observed information. This is the case with a standard GEE analysis, assumptions are made as to the relevant portions of the distribution assumed to have given rise to the observed information then an appropriate pseudo-likelihood is developed accordingly. The so called pseudo-likelihood is often considered appropriate as it is implicitly assumed that no part of the true distribution that was ignored will cause a pseudo-likelihood and a full likelihood analysis to differ substantially. The existence of Informative Cluster Size (ICS) is a violation of this implicit assumption. Much work to date has been done to highlight the deficiencies of Standard GEE within an ICS context (Williamson *et. al* 2003, Benhin *et. al* 2005, Wang *et. al* 2011 ...) and to provide applicable alternatives. One such alternative is Full Joint Likelihood Maximization (Joint Modeling).

Joint Modeling seeks to define the complete form of the distribution which gave rise to the observed data and maximize the corresponding joint likelihood. In a standard clustered data analysis it is not uncommon to assume that the size of the observed clusters is independent of the observed response, in such a case a seemingly appropriate likelihood, say, $L^S = f(\theta|X, Y)$ where X is the observed covariate matrix and Y the observed response vector can be defined and parametric estimation proceeds ignoring the cluster size information.

However, in an ICS setting the cluster size is dependent on the response. As such one would need to define the joint likelihood accommodating the cluster size distribution, say,

$$L_i^J = f(\theta|X_i, Y_i, N_i), \quad (3.23)$$

where N_i denotes the cluster size of the i^{th} cluster. However, in the longitudinal setting it may be the case that both the cluster size and the number of temporal

observations made on cluster units is dependent on the response. In such a setting one must also accommodate the distribution governing the number of temporal observations, say,

$$L_i^J = f(\theta|X_i, Y_i, N_i, K_i), \quad (3.24)$$

where K_i denotes the vector of the number of temporal observations collected on the subunits belonging to the i^{th} cluster. Our work here will focus on clustered longitudinal data with ICS and Temporally Varying Cluster Sizes (TVCS). As it would be difficult to proceed from this point defining generic conditional probabilities we introduce an example design in Section 1 for which we will explicitly define the full joint likelihood. It should be noted that in the design indicated in Section 1, the cluster size is dependent on a cluster level random intercept and the distribution of the number of temporal observations made on subunits belonging to a given cluster dependent on cluster size. In such a setting, ignoring the cluster size and temporal observation component of the full likelihood would constitute a misspecification of the likelihood and can result in invalid inference.

1 Design

Let, $Y_{ijk} \sim N(X_{ijk}\beta + \alpha_i, .1^2)$ where $\alpha_i \sim N(0, .2^2)$. The components of $\beta = (.5, 1.5, .75, .8)$ correspond to the intercept, time effect, cluster constant exposure, and exposure time interaction with half of the Q clusters having exposure $X_2 = 1$. The cluster size at baseline, *i.e.*, $N_i^1 \sim Pois(exp\{\lambda_i\}) + 1$ where $\lambda_i = 1.5 + 1.5\alpha_i$. The number of temporal observations made on each subject varies between 2 and 4 with the individual number of observations determined by the ordinal logit model with probabilities:

$$\begin{aligned}
\pi_{k_{ij}=2} &= \frac{e^{-.01 \times N_i^1}}{1 + e^{-.01 \times N_i^1}}, \\
\pi_{k_{ij}=3} &= 1 - \pi_{k_{ij}=2} - \frac{1}{1 + e^{1.5 - .01 \times N_i^1}}, \text{ and} \\
\pi_{k_{ij}=4} &= \frac{1}{1 + e^{1.5 - .01 \times N_i^1}}.
\end{aligned}$$

2 Likelihood

Under the specifications of the indicated design, we define Equation 2 as the full likelihood.

$$\begin{aligned}
L(\theta|Y, X, \alpha, N, K) &= \prod_{i=1}^Q \int_{-\infty}^{\infty} \frac{\exp\{\tau_0 + \tau_1 \alpha_i\}^{N_i-1}}{(N_i - 1)!} \exp\{-\exp[\tau_0 + \tau_1 \alpha_i]\} \cdot \\
&\quad \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} \exp\left\{\frac{-(\alpha_i)^2}{2\sigma_\alpha^2}\right\} \cdot \prod_{j=1}^{N_i} \left(I_{[k_{ij}=2]} \frac{\exp\{\gamma_1 N_i\}}{1 + \exp\{\gamma_1 N_i\}} \right. \\
&\quad \left. + I_{[k_{ij}=3]} \left[1 - \frac{\exp\{\gamma_1 N_i\}}{1 + \exp\{\gamma_1 N_i\}} - \frac{1}{1 + \exp\{\gamma_1 N_i + \gamma_2\}} \right] + \right. \\
&\quad \left. I_{[k_{ij}=4]} \frac{1}{1 + \exp\{\gamma_1 N_i + \gamma_2\}} \right) \prod_{k=1}^{k_{ij}} \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left\{\frac{-(y_{ijk} - x_{ijk}\beta - \alpha_i)^2}{2\sigma_e^2}\right\} d\alpha_i.
\end{aligned} \tag{3.25}$$

3 Likelihood Estimation and Maximization

Since our concern in this investigation is marginal parameter estimation we are uninterested in obtaining estimates of the cluster level random effects. However, in order to maximize Equation 2 we must address the matter of the unobservable random effects, we do so through the implementation of numeric approximation via Gaussian Hermite Quadrature. To that end we maximize the following:

$$\begin{aligned} \tilde{l}(\theta|Y, X, \alpha, N, K) = & \sum_{i=1}^Q \ln \left(\sum_{l=1}^{N_{GH}} \frac{1}{\sqrt{\pi}} \cdot GHw_l \cdot \frac{\exp\{\tau_0 + \tau_1 z_l\}^{N_i-1}}{(N_i-1)!} \exp\{-\exp[\tau_0 + \tau_1 z_l]\} \cdot \right. \\ & \prod_{j=1}^{N_i} \left\{ I_{[k_{ij}=2]} \frac{\exp\{\gamma_1 N_i\}}{1 + \exp\{\gamma_1 N_i\}} + I_{[k_{ij}=3]} \left[1 - \frac{\exp\{\gamma_1 N_i\}}{1 + \exp\{\gamma_1 N_i\}} - \frac{1}{1 + \exp\{\gamma_1 N_i + \gamma_2\}} \right] + \right. \\ & \left. \left. I_{[k_{ij}=4]} \frac{1}{1 + \exp\{\gamma_1 N_i + \gamma_2\}} \right\} \cdot \prod_{k=1}^{k_{ij}} \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp \left\{ -\frac{(y_{ijk} - x_{ijk}\beta - z_l)^2}{2\sigma_e^2} \right\} \right). \quad (3.26) \end{aligned}$$

Here z_l corresponds to the l^{th} Gauss-Hermite abscissa value (i.e. $GHx_l\sqrt{2\sigma_\alpha}$) and GHw_l the corresponding weight. Standard GEE, Least Squares, Poisson and Ordinal Logistic regression are used to obtain reasonable initial estimates and $\langle \text{optim}() \rangle$ is used to maximize the corresponding objective function given by Equation 3.26.

C GEE and Marginal WGEE

Within the clustered longitudinal framework, standard GEE estimation is the result of solving the following:

$$0 = \sum_{i=1}^Q \sum_{j=1}^{N_i} \frac{\partial \mu_{ij}}{\partial \beta} V_{ij}^{-1} (Y_{ij} - \mu_{ij}), \quad (3.27)$$

where V_{ij} is determined by the assumed form of the correlation structure. Marginal Weighted Generalized Estimating Equations (MWGEE) reweight individual contributions to the score. This is done in practice by imposing appropriate individual weights into the score specification (i.e. Equation 3.27). For example, Cluster Weighted GEE (CWGEE, Williamson *et. al* 2003) estimates are obtained by solving

$$0 = \sum_{i=1}^Q \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{\partial \mu_{ij}}{\partial \beta} V_{ij}^{-1} (Y_{ij} - \mu_{ij}), \quad (3.28)$$

However, as discussed in Chapter II the CWGEE framework is not appropriate when the number of temporal observations available are dependent on the response.

The design indicated in Section 1 possesses a dependence between the response and both cluster size and the observed temporal profile. As such a Marginal WGEE approach accommodating both cluster size and the number of temporal observations is needed. For the indicated design an appropriate MWGEE estimate could be obtained as the solution to:

$$0 = \sum_{i=1}^Q \frac{1}{N_i} \sum_{j=1}^{N_i} \left(\frac{1}{k_{ij}} \right) \frac{\partial \mu_{ij}}{\partial \beta} V_{ij}^{-1} (Y_{ij} - \mu_{ij}), \quad (3.29)$$

where k_{ij} denotes the number of temporal observations made on the ij^{th} subunit. For specific details pertaining to MWGEE parameter estimation the reader is referred to Section F of Chapter II.

D Variance Component Estimation

In maximum likelihood modeling settings acquiring estimates of the variance components corresponding to the parameters of interest is, at least in theory, a straightforward matter. A plug-in sandwich estimator of the variance covariance matrix can be defined in terms of the first and second derivatives (scores and Hessian) of the log(likelihood) *i.e.* $l(\theta|Y, X, \dots)$. We define this sandwich estimator as:

$$\Sigma_S = \left(\frac{\partial^2 l}{\partial \theta \partial \theta} \right)^{-1} \left(\sum_{i=1}^Q \left\{ \frac{\partial l_i}{\partial \theta} \right\} \left\{ \frac{\partial l_i}{\partial \theta} \right\}^T \right) \left(\frac{\partial^2 l}{\partial \theta \partial \theta} \right)^{-1}. \quad (3.30)$$

Σ_S is a consistent estimator of the variance component associated with θ . In our setting, however, the likelihood cannot be directly calculated and instead must be estimated. As such we must define a new plug-in variance estimator $\tilde{\Sigma}_S$ denoting the sandwich variance estimate obtained using the estimated (*via numeric approximation*) likelihood as in our example.

$$\tilde{\Sigma}_S = \left(\frac{\partial^2 \tilde{l}}{\partial \theta \partial \theta} \right)^{-1} \left(\sum_{i=1}^Q \left\{ \frac{\partial \tilde{l}_i}{\partial \theta} \right\} \left\{ \frac{\partial \tilde{l}_i}{\partial \theta} \right\}^T \right) \left(\frac{\partial^2 \tilde{l}}{\partial \theta \partial \theta} \right)^{-1}, \quad (3.31)$$

where $\tilde{l}$ denotes the log(likelihood) obtained via numeric approximation (in this case via Guass-Hermite Quadrature). Our investigation indicated considerable discrepancies between the observed variances and those estimated via $\tilde{\Sigma}_S$. In Section 2, we discuss in greater detail why the use of $\tilde{\Sigma}_S$ is inappropriate for our investigation. In light of the shortcomings of $\tilde{\Sigma}_S$, an alternative was needed in order to investigate the performance of the Joint Model with respect to power and efficiency; as such, we obtained estimates of the variance via the bootstrap.

We define our bootstrap variance estimator as:

$$\Sigma_{BS} = \frac{(\hat{\theta} - \theta_r^*)(\hat{\theta} - \theta_r^*)^T}{N_{BS} - 1}, \quad (3.32)$$

where $\hat{\theta}$ denotes the point estimate of the parameter vector obtained via maximization of the estimated log(likelihood) utilizing the complete sample and θ_r^* denotes the point estimate of the parameter vector obtained using the r^{th} bootstrap sample data set constructed by sampling with replacement Q clusters from the complete sample. While the bootstrap variance estimate was not immune to issues encountered via numeric approximation of l it was a decided improvement over estimates obtained via the sandwich estimator $\tilde{\Sigma}_S$.

Estimates of the variance for MWGEE estimates were obtained via the modified sandwich introduced in Chapter II and included for continuity below.

$$\Sigma_{WGEE} = \tilde{\Omega}^{-1} \tilde{M} \tilde{\Omega}^{-1}, \quad (3.33)$$

with

$$\tilde{M} = \sum_{i=1}^Q \left(\left\{ \sum_{j=1}^{N_i} \frac{\partial \mu_{ij}}{\partial \beta} (\omega_{ij} V_{ij}^{-1}) \epsilon_{ij} \right\} \left\{ \sum_{j=1}^{N_i} \frac{\partial \mu_{ij}}{\partial \beta} (\omega_{ij} V_{ij}^{-1}) \epsilon_{ij} \right\}^T \right), \quad (3.34)$$

and $\tilde{\Omega} = \sum_{i=1}^Q \sum_{j=1}^{N_i} \left(\frac{\partial \mu_{ij}}{\partial \beta} (\omega_{ij} V^{-1}) \left(\frac{\partial \mu_{ij}}{\partial \beta} \right)^T \right)$. We found that the modified sandwich estimator Σ_{WGEE} was an appropriate estimator of the MWGEE variance and that tests performed using the resulting estimates maintained appropriate size.

E Simulation

In order to demonstrate the small sample properties of both our MWGEE approach and the Joint Modeling approach we turn to a simulation. The design discussed in Section 1 presents an ideal setting to compare the two methods as the dependencies between the response, cluster sizes and observed temporal profiles are intentionally constructed to obscure inference obtained through standard GEE.

A Monte Carlo simulation was employed from the indicated design with 1,000 data generations and the following parameterization:

$$\beta = (.5, 1.5, .75, .8), \tau = (1.5, 1.5), \gamma = (-.5, 1.5, -.01), \sigma_e = .1 \text{ and } \sigma_\alpha = .2.$$

The results are summarized in various detail in the following section. The Wald statistics was used to obtained the power plots with the joint model standard errors obtained via bootstrap, using point estimates obtained from 50 bootstrap samples. Past MWGEE investigations indicated that MWGEE point estimates are reasonably robust to misspecification of the working correlation structure. As such in our Monte Carlo simulation we assume independence working correlation when estimating MWGEE models. In the estimation of Joint Model parameters the estimates of σ_e^2 and σ_α^2 were fixed at initialization (initial estimates were consistently in line with the true values) the corresponding likelihood was maximized with respect to the remaining parameters, we also consider two estimation schemes with 20 and 50 Gauss-Hermite Quadrature points for the Joint Model. In the Subsection 2 we will discuss specific details corresponding to the Joint Model parameter estimation procedure, specifically the choice of Quadrature points.

1 Simulation Results

$Q = 50$: 20 Quadrature Points

From Table 9 we can see that in the $Q = 50$ both MWGEE and the Joint Model perform comparably with respect to unbiased estimation of the marginal

TABLE 9

Tabular results of 1,000 Monte Carlo simulations from the indicated design with $Q = 50$ clusters and 20 Quadrature Points.

Joint				
	β_0	β_1	β_2	β_3
Mean	0.50323	1.49977	0.74973	0.80028
BIAS	0.00323	-0.00023	-0.00027	0.00028
MSE	0.00329	0.00003	0.00613	0.00006
σ_{MC}^2	0.00328	0.00003	0.00614	0.00006
σ_r^2	0.00208	0.00003	0.00334	0.00007

WGEE				
	β_0	β_1	β_2	β_3
Mean	0.49920	1.50057	0.74708	0.80084
BIAS	-0.00080	0.00057	-0.00292	0.00084
MSE	0.00206	0.00013	0.00412	0.00025
σ_{MC}^2	0.00206	0.00013	0.00411	0.00025
σ_r^2	0.00202	0.00012	0.00406	0.00023

parameter β . However, one would expect to see consistent reduction in the variance estimates for the Join Model over the MWGEE estimates. For example, the variance estimates obtained using the $\hat{\beta}$ estimates of the 1,000 Monte Carlo data generations *i.e.* σ_{MC}^2 should be smaller for the Joint Model than the MWGEE. Similarly, there should be reasonable agreement between the mean of the variance estimates obtained for the 1,000 Monte Carlo data generations *i.e.* σ_r^2 and the corresponding values of σ_{MC}^2 .

These discrepancies in the variances are telling of an underlying numeric issue involved in complicated Maximum Likelihood estimation problems. The variance issue will be exacerbated in the $Q = 200$ setting and we will discuss the root cause and various approaches to correcting them in Section 2. Figure 14 demonstrates that, in the $Q = 50$ setting, the estimator of β is approximately normal. However, Figure 13 demonstrates the effect numeric issues in variance estimation have on the power.

Figure 13. Power plots of the rejection rates of Wald statistic (estimated from bootstrap for Joint model), $Q = 50$ and 20 Quadrature Points. WGEE - Black, Joint Model -Red

Figure 14. PP plots of the empirical coverage vs. Normal (sd obtained from 1,000 Monte Carlo samples), $Q = 50$ and 20 Quadrature Points. WGEE - Black, Joint Model -Red

$Q = 200$: 20 Quadrature Points

TABLE 10

Tabular results of 1,000 Monte Carlo simulations from the indicated design with $Q = 200$ clusters and 20 Quadrature Points.

Joint				
	β_0	β_1	β_2	β_3
Mean	0.50369	1.50001	0.75051	0.80016
Bias	0.00369	0.00001	0.00051	0.00016
MSE	0.00172	0.00001	0.00282	0.00002
σ_{MC}^2	0.00170	0.00001	0.00282	0.00002
σ_r^2	0.00172	0.00001	0.00276	0.00002

WGEE				
	β_0	β_1	β_2	β_3
Mean	0.49829	1.50065	0.75047	0.80025
Bias	-0.00171	0.00065	0.00047	0.00025
MSE	0.00050	0.00003	0.00105	0.00007
σ_{MC}^2	0.00049	0.00003	0.00105	0.00007
σ_r^2	0.00052	0.00003	0.00105	0.00006

We can see from Table 10 that the Joint Model performs well estimating the marginal parameter vector in the $Q = 200$ setting. While the variances estimates σ_{MC}^2 and σ_r^2 are in agreement, there is no improvement to efficiency w.r.t. the MWGEE estimates for β_0 and β_2 . Figure15 points to other issues encountered in the Joint Model pertaining the effect variance component estimation has on power. The MWGEE estimates are consistent with respect to both normality and variance estimation.

$Q = 50$: 50 Quadrature Points

Table 11 presents the results of the Monte Carlo Simulation $Q = 50$, when the Join Model estimates were obtained using 50 Quadrature Points. We can see that while the marginal parameter point estimates remain unbiased there are still

Figure 15. Power plots of the rejection rates of Wald statistic (estimated from bootstrap for Joint model), $Q = 200$. WGEE - Black, Joint Model -Red

Figure 16. PP plots of the empirical coverage vs. Normal (sd obtained from 1,000 Monte Carlo samples), $Q = 200$ and 20 Quadrature Points. WGEE - Black, Joint Model -Red

issues associated with estimating the variance components. Figure 17 demonstrates that while the empirical coverage with respect to the Monte Carlo samples remains approximately Normal, the variance estimation issues continue to plague the power associated with β_0 and β_2 and while the variance estimates are improved from the 20 Quadrature Point model they are not a consistent improvement w.r.t the MWGEE variances as would be expected.

TABLE 11

Tabular results of Joint Model estimates from the 1,000 Monte Carlo simulations from the indicated design with $Q = 50$ clusters and 50 Quadrature Points.

	Joint			
	β_0	β_1	β_2	β_3
Mean	0.50434	1.49976	0.74942	0.80031
BIAS	0.00434	-0.00024	-0.00058	0.00031
MSE	0.00259	0.00003	0.00504	0.00006
σ_{MC}^2	0.00257	0.00003	0.00504	0.00006
σ_r^2	0.00127	0.00003	0.00209	0.00006

$Q = 200$: 50 Quadrature Points

Increasing the number of Quadrature Points in the $Q = 200$ setting has mixed effects. On the one hand parameter estimates remain unbiased and in the estimation of β_0 the variance estimates appear to improve as can be seen from Figure 18. However, in the 20 Quadrature point setting we saw that the test associated with β_2 was around level .05 using the bootstrap variance estimate. In the 50 Quadrature Point estimation scheme the test associated with β_2 appears to be around level .1. This finding underscores the issues encountered with the choice of the number of quadrature points as one would expect that increasing the resolution would offer consistent improvement.

Figure 17. PP plots of the empirical coverage vs. Normal and Power plots of the rejection rates of Wald statistic (estimated from bootstrap) for β_0 and β_2 , $Q = 50$ and 50 Quadrature Points. WGEE - Black, Joint Model -Red

Figure 18. PP plots of the empirical coverage vs. Normal and Power plots of the rejection rates of Wald statistic (estimated from bootstrap) for β_0 and β_2 , $Q = 200$ and 50 Quadrature Points. WGEE - Black, Joint Model -Red

TABLE 12

Tabular results of Joint Model estimates from the 1,000 Monte Carlo simulations from the indicated design with $Q = 200$ clusters and 50 Quadrature Points.

	Joint			
	β_0	β_1	β_2	β_3
Mean	0.50429	1.50000	0.75187	0.80013
BIAS	0.00429	-0.00000	0.00187	0.00013
MSE	0.00103	0.00001	0.00179	0.00001
σ_{MC}^2	0.00101	0.00001	0.00178	0.00001
σ_r^2	0.00090	0.00001	0.00145	0.00002

2 On Numeric Estimation of Joint Model Parameters

The appeal of the Joint Modeling framework is undeniable. Being able to specify the complete likelihood corresponding to the distribution that gave rise to observed information mitigates a number of issues encountered using less flexible but more frequently utilized methods. Similarly, by specifying the full likelihood one expects improved efficiency and a reduction in bias over estimates obtained using misspecified likelihoods.

That said, such gains can only be achieved when accurate numeric estimation and optimization of the full likelihood are consistently achieved. When dealing with misspecified likelihood problems it is understood that loss of efficiency and bias are byproducts of over simplification. However, even when the full likelihood is correctly specified one is not guaranteed unbiasedness and improved efficiency. When estimating marginal parameters in a random effects model the onus is placed on β and the random effects are integrated out, this is the case even for standard GEE analyses. This is also the approach taken in marginal parameter estimation in the Joint Modeling setting and is commonly carried out via Gauss-Hermite Quadrature.

Gauss-Hermite Quadrature provides a very powerful tool capable of rather drastically reducing the computational burden associated with integrating out

Figure 19. Estimated likelihood for various numbers of quadrature points. The likelihood is mapped for small variations of the indicated parameter (β_0), holding the other parameters fixed at the true value. Likelihood was estimated using one random generation of the indicated design with $Q = 50$ clusters.

Figure 20. Estimated likelihood for various numbers of quadrature points. The likelihood is mapped for small variations of the indicated parameter (β_1), holding the other parameters fixed at the true value. Likelihood was estimated using one random generation of the indicated design with $Q = 50$ clusters.

random effects in likelihood maximization problems. However, the choice of the number of Quadrature Points to evaluate is not a straightforward one. In many applications it is common to find that people report the use of between 5 and 20 Quadrature Points and in many of those scenarios the consensus is that a small number (5 to 20) is sufficient. The commonplace use of Hermite Quadrature, and the frequent use of a small number of Quadrature Points, to address random effects in the estimation of complicated full likelihood problem provides an analyst with a false sense of security. One should examine the estimated likelihood very carefully prior to settling on a ‘sufficient’ number of points.

Figures 19 - 26 point to a common issue encountered when implementing Gauss-Hermite Quadrature. The figures were created by setting the marginal parameter vector to the true value and then parameter by parameter the estimated likelihood was mapped for small deviations in the parameter values. Each curve represents the estimated likelihood as determined using the indicated number of quadrature points. The blue lines in Figures 19 and 21 point to the reason why even 50 Quadrature points is insufficient for estimating the variance component for β_0 and β_2 . Due to the low number of quadrature points and more specifically the low resolution in the area of high density of the corresponding random effects distribution an artificial multi-modality is introduced to the estimated likelihood. Lesaffre and Spiessens (2001) pointed to specific issues encountered in estimating variance components and in estimating first and second order numeric derivatives as a result of this lack of resolution even when investigating a rather simple logistic model. They further cite concerns surrounding the common default settings in many canned software packages which typically employ a small number of quadrature points.

Another concern with the choice of quadrature points is how the issue seems to be exacerbated by increasing the sample size. We can see from Figures 19 - 22 that the discrepancy between the estimated $\log(\text{likelihood})$ with 10 points as

Figure 21. Estimated likelihood for various numbers of quadrature points. The likelihood is mapped for small variations of the indicated parameter (β_2), holding the other parameters fixed at the true value. Likelihood was estimated using one random generation of the indicated design with $Q = 50$ clusters.

Figure 22. Estimated likelihood for various numbers of quadrature points. The likelihood is mapped for small variations of the indicated parameter (β_3), holding the other parameters fixed at the true value. Likelihood was estimated using one random generation of the indicated design with $Q = 50$ clusters.

Figure 23. Estimated likelihood for various numbers of quadrature points. The likelihood is mapped for small variations of the indicated parameter (β_0), holding the other parameters fixed at the true value. Likelihood was estimated using one random generation of the indicated design with $Q = 200$ clusters.

Figure 24. Estimated likelihood for various numbers of quadrature points. The likelihood is mapped for small variations of the indicated parameter (β_1), holding the other parameters fixed at the true value. Likelihood was estimated using one random generation of the indicated design with $Q = 200$ clusters.

Figure 25. Estimated likelihood for various numbers of quadrature points. The likelihood is mapped for small variations of the indicated parameter (β_2), holding the other parameters fixed at the true value. Likelihood was estimated using one random generation of the indicated design with $Q = 200$ clusters.

Figure 26. Estimated likelihood for various numbers of quadrature points. The likelihood is mapped for small variations of the indicated parameter (β_3), holding the other parameters fixed at the true value. Likelihood was estimated using one random generation of the indicated design with $Q = 200$ clusters.

apposed to 100 is somewhere in the range of 35 to 45 (around the optimal region). When we increase the number of clusters to 200 the discrepancy increases to somewhere in the range of 250 to 400 as can be seen in the panels of Figure 23 - 26.

F Discussion

When applicable fully parametric maximum likelihood methods, such as Joint Modeling, offer a straightforward approach to marginal model parameter estimation. Theory dictates that when applying ML methods one can expect asymptotically unbiased parameter estimation and improved efficiency w.r.t. other methods. In situations where the true likelihood can be directly calculated ML methods should be the first choice in any setting. However, in situations like marginal parameter estimation when the data are the product of a random effects model the joint likelihood cannot be directly calculated and as a result must be estimated. When the likelihood must be approximated the validity of the analysis rests on the accuracy of the likelihood estimates obtained.

In this chapter we have presented a ML problem with a rather involved likelihood where the data are subject to cluster specific random effects. In our investigation we employed standard methods such as Gauss-Hermite Quadrature to integrate out the random effects as a means of obtaining the full likelihood and numeric optimization of the resulting approximate likelihood. Conventional wisdom would suggest that aside from logistical issues associated with processing the computational aspect of model estimation our ML problem should be little more than a matter of defining the correct likelihood and waiting for the results. Early on, however, we encountered a number of issues for which solutions were not readily available. Worse yet, in situations such as choosing the number of Quadrature Points the standard choices were clearly inadequate.

Under the specifications of the indicated design the average number of records for $Q = 50$ data sets was around 800 records; in the $Q = 200$ setting this increases

to around 3,200. Because of the issues encountered with estimating the variance components of the Joint Model parameter vector, likely due to the lack of smoothness in the estimated likelihood, variance estimates were obtained via bootstrap. The computational burden associated with acquiring point estimates of the marginal parameter vector is not of great concern, for example, optimizing the estimated likelihood for $Q = 200$ data with 50 Quadrature Points takes a little under an hour. However, bootstrapping the variance estimates brings the single node computational burden of obtaining parameter and variance estimates to around 40 hours which is still manageable.

In simulation the true state is known. Even so, choosing an appropriate number of points for accommodating the estimation of variance components is not a problem with a straightforward solution. In application, however, the true state is unknown and there is no clearly established protocol for assessing what an adequate number of Quadrature Points is and the effects a poor choice would have on the validity of the resulting analysis.

In moving forward the direct application of this Joint Model framework raises concerns as to the computational feasibility of such an investigation. In our largest simulation settings the data consisted of around 3,200 records and the computational burden of maximizing the estimated likelihood and obtaining corresponding variance estimates with respect to nine parameters (β , τ and γ) using 50 Quadrature Points was a little under two days per sample. As indicated, even 50 Quadrature Points appeared insufficient in this setting. By contrast the Periodontal data analyzed in Chapter II consisted of more than 25,000 records and the candidate model investigated contained 18 marginal model parameters (without accommodating any parameters corresponding to the cluster size or temporal profile distributions). While increasing the number of records has a seemingly linear effect on the computational burden associated with Joint Model maximization, increasing the number of parameters appears to increase the computational burden

disproportionately. As such, analyzing the marginal model investigated in Chapter II using the estimated likelihood maximization approach outlined here is computationally intractable.

We extend the WGEE framework to offer a viable alternative to direct ML estimation. While at present there does not appear to exist a canned software package that will carry out MWGEE analyses correctly (parameter and corresponding variance component estimation) the framework of MWGEE estimation is simple enough that algorithms to implement MWGEE modeling can be written rather quickly. While MWGEE's do experience a loss of efficiency compared to Joint Models we have demonstrated through simulation that the added efficiency of Joint models is dependent on the parameters of the ML estimation algorithm and is not a forgone conclusion. In a similar vein, in our simulation we assumed independence working correlation structures when estimating the MWGEE parameters. This intentional misspecification was made to illustrate the apparent robustness of MWGEE's to misspecification of the working correlation structure; a desirable property inherited from standard GEE analysis.

In conclusion we find MWGEE appropriate for addressing clustered longitudinal data with informative cluster sizes and temporal observation profiles. While it is understood that Joint Models are the preferred method for analyzing this sort of data there is considerably more to Joint Model estimation than correctly specifying the full joint likelihood. The sandwich variance estimators for MWGEE models require a very specific attention to detail but provide consistent estimators of the corresponding variance components. Whereas, estimates of the variance components of Joint Models can be severely effected by the resolution with which the likelihood can be approximated. In the future will will investigate other methods for estimating the likelihood as well as investigate the effects the choice of working correlation has on the efficiency of MWGEE estimators.

CHAPTER IV

Future Work

In this work we have discussed a number of topics pertaining to the application and extension of linear regression techniques as well as maximum likelihood estimation problems. In addressing these topics new questions and topics emerged. We enumerate here a few of those topics and questions as the focus of future investigations.

A Materials Science - Band gap estimation

1. How to address the integrity issues encountered in the collection of materials information?

Suggestions were made in regard to a framework for collecting and validating materials data to be used in an informatics based search for new materials with desirable properties. Such a plan would require the orchestration of several collaborative bodies charged with collecting/generating and validating information pertaining to materials of interest. Logistically the implementation of such an undertaking would be difficult. However, the existence of an independent body composed of a panel of researchers for the purposes of ensuring the integrity of reported materials data would not only improve the quality of the data used in our investigation but would help to raise the bar for standards as to investigative reporting in the future.

2. Determining appropriate forms of the Hybrid $Q(\cdot)$ function for modeling solid solution band gaps.

As things stand at present, unaccounted sources of variability in the body of materials data prevent a detailed treatment of the solid solutions problem. However, when the sources of variability can be identified attention can be turned to accurate estimation of constituent compound properties and mapping solutions properties as a function of concentration.

B MWGEE's

1. Investigate the properties of MWGEE's w.r.t. various misspecifications of the working correlation structure.

In this work our simulations involving MWGEE modeling intentionally assumed independence working correlation. This intentional misspecification was made in an effort to demonstrate at least to some extent the robustness MWGEE's to misspecification of the working correlation. However, we neglected a detailed investigation w.r.t. the effects of the choice of working correlation on efficiency and point estimation.

2. Broad extension of MWGEE's

We focused our attention on a specific case of clustered longitudinal data. However, the utility of re-weighted Generalized Estimating Equations extends beyond that of single case specific application. Ideally, a broad class of MWGEE's should be defined and their properties investigated to facilitate large scale acceptance and application in the future.

C Joint Modeling

1. Address likelihood estimation issues encountered when implementing Gauss-Hermite Quadrature to estimate complicated likelihoods.

We discussed in some detail issues posed by employing Gauss-Hermite

Quadrature to estimate complicated likelihoods. While often appropriate in simpler situations Gauss-Hermite Quadrature may not always be appropriate.

2. Consideration of other likelihood estimation schemes including Adaptive Gauss-Hermite Quadrature and Bayesian estimation.
3. Investigate the robustness of the Joint Model to misspecification of the cluster size and temporal profile distribution.

REFERENCES

- [1] James Beck, Lisa Cusmano, Wendy Greene-Helms, Gary Koch, and Steven Offenbacher. A 5-year study of attachment loss in community-dwelling older adults: incidence density. *Journal of Periodontal Research*, 32:506–515, 1997b.
- [2] James Beck, Tonya Sharp, Gary Koch, and Steven Offenbacher. A study of attachment loss patterns in survivor teeth at 18 months, 36 months and 5 years in community-dwelling older adults. *Journal of Periodontal Research*, 6:497–505, 1997a.
- [3] Emmanuel Benhin, Jon Rao, and Alastair Scott. Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes. *Biometrika*, 92:435–450, 2005.
- [4] Joe Bible, James Beck, and Somnath Datta. Cluster Adjusted Regression for Displaced Subject Data (CARDS): Marginal Inference under Potentially Informative Temporal Cluster Size Profiles. *Biometrics*, Submitted.
- [5] Zhen Chen, Bo Zhang, and Paul Albert. A joint Modeling approach to data with informative cluster size: Robustness to the cluster size model. *Statistics in Medicine*, 30:1825–1836, 2011.
- [6] Somnath Datta, Dipankar Bandyopadhyay, and Glen Satten. Inverse Probability of Censoring Weighted U-statistics for Right Censored Data With an Application to Testing Hypotheses. *Scandinavian Journal of Statistics*, 37:680–700, 2010.
- [7] Somnath Datta and Glen Satten. Rank-Sum Tests for Clustered Data. *Journal of the American Statistical Association*, 100:908–915, 2005.

- [8] Somnath Datta and Glen Satten. A Signed-Rank Test for Clustered Data. *Biometrics*, 64:501–507, 2008.
- [9] David Dunson, Zhen Chen, and Jean Harry. A Bayesian Approach for joint Modeling of Cluster Size and Subunit Specific Outcomes. *Biometrics*, 59:521–530, 2003.
- [10] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least Angle Regression. *Annals of Statistics*, 32:407–499, 2004.
- [11] Jie Fan and Somnath Datta. Fitting marginal accelerated failure time models to clustered survival data with potentially informative cluster size. *Computational Statistics and Data Analysis*, 55:3295–3303, 2011.
- [12] Sylvie Goetgeluk and Stijn Vansteelandt. Conditional Generalized Estimating Equations for the Analysis of Clustered and Longitudinal Data. *Biometrics*, 64:772–780, 2008.
- [13] Daniel Hall and Thomas Severini. Extended Generalized Estimating Equations for Clustered Data. *Journal of the American Statistical Association*, 93:1365–1375, 2012.
- [14] Sren Hjsgaard, Ulrich Halekoh, and Jun Yan. The R Package geepack for Generalized Estimating Equations. *Journal of Statistical Software*, 15/2:1–11, 2006.
- [15] Elaine Hoffman, Pranab Sen, and Clarice Weinberg. Within-Cluster Resampling. *Biometrika*, 88:1121–1134, 2001.
- [16] Ying Huang and Brian Leroux. Informative Cluster Sizes for Sub-Cluster Level Covariates and Weighted Generalized Estimating Equations. *Biometrics*, 67:843–851, 2011.

- [17] Emmanuel Lesaffre and Bart Spiessens. On the effect of the number of quadrature points in a logistic random effects model: an example. *Journal of the Royal Statistical Society: Series C*, 50:325–335, 2001.
- [18] Kung Yee Liang and Scott Zeger. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 73:13–22, 1986.
- [19] Stewart Lipsitz, Geert Molenberghs, Garrett Fitzmaurice, and Joseph Ibrahim. GEE with Gaussian Estimation of the Correlations When Data Are Incomplete. *Biometrics*, 56:528–536, 2000.
- [20] Motomi Mori, Robert Woolson, and George Woodsworth. Slope Estimation in the Presence of Informative Right Censoring: Modeling the Number of Observations as a Geometric Random Variable. *Biometrics*, 50:39–50, 1994.
- [21] Jaakko Nevalainen, Somnath Datta, and Hannu Oja. Inference on the marginal distribution of clustered data with informative cluster size. *Statistical Papers*, 55:71–92, 2014.
- [22] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.
- [23] James Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed . *Journal of the American Statistical Association*, 89:846–866, 1994.
- [24] James Robins, Andrea Rotnitzky, and Lue Ping Zhao. Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association*, 90:106–121, 1995.
- [25] Donald Rubin. Inference and Missing Data. *Biometrika*, 63:581–592, 1976.

- [26] Shaun Seaman, Menelaos Pavlou, and Andrew Copas. Methods for Observed-Cluster Inference When Cluster Size Is Informative: A Review and Clarifications. *Biometrics*, 70:449-456, 2014.
- [27] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58:267-288, 1996.
- [28] Ming Wang, Maiying Kong, and Somnath Datta. Inference for marginal linear models for clustered longitudinal data with potentially informative cluster sizes. *Statistical Methods in Medical Research*, 20:347-367, 2011.
- [29] John Williamson, Somnath Datta, and Glen Satten. Marginal Analysis of Clustered Data When Cluster Size Is Informative . *Biometrics*, 59:36-42, 2003.
- [30] Jun Yan and Jason P. Fine. Estimating Equations for Association Structures. *Statistics in Medicine*, 23:859-880, 2004.
- [31] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67:301-320, 2005.

A An R function for estimating MWGEE models

Below is R the script for the MWGEE function used in the periodontal investigation for marginal parameter and corresponding covariance estimation. If one assumes independence working correlation one can uncomment the ‘*alpha2j=0*’ line and numerically consistent results w.r.t. the marginal parameter vector β using the *weights=* option for the *geeglm()* function available as part of the *geepack* package [14] in R [22]. However, the *geeglm()* function will not handle the estimation of the of the correlation parameters correctly with the use of the *weights* option. Similarly, when the dependence structure is misspecified as is the case when accommodating the within tooth temporal correlation and ignoring the between tooth correlation no package will estimate the covariance structure correctly.

```
#####  
#R Function for by hand WGEE estimation  
#  
#This is the function used in the periodontal  
#analysis. Assuming Ar1 corr structure.  
#  
#In simulation because independence working correlation  
#was assumed it was possible to exploit the  
#geeglm() function using the <weights=> option.  
#  
#  
#geeglm() will provide identical results to the  
#gausgee() if the '#alpha2<-0' line is  
#uncommented. That is with respect to the marginal  
#parameter point estimates.  
#  
#geeglm() does not handle the estimation of the correlation  
#parameters correctly and no canned gee estimation software  
#will estimate the variance component correctly given that  
#the cluster level dependencies are ignored.
```

```

#
#SAS's gennmod procedure appears to handle estimation of the
#correlation parameters reasonably well (they appear in line
#at least to a degree with those of the specificaitons indicated
#here). However, the variance component will still need to
#be estimated independently. A sample script for genmod
#MWGEE estimation is included with with a macro for bootstrap
#variance estimation.
#####

geehand<-setClass( 'geehand' ,
                  representation( beta= 'matrix' ,Nclus='numeric' , alpha='
                                numeric' ,iter='numeric' ))

gausgee<-function(y, x, id, w) {
  X<-x
  Y<-y
  #ID is a single variable which contains Individual ID and Tooth ID.
  #In some situations (calculating the sandwich variance estimate) ID is
  #parsed to extract
  #only the Individual ID. This is to allow proper treatment of the
  #indexing in order to address
  #the dependencies properly.
  ID<-id
  Uid<-unique(id)
  j<-ID

  U<- matrix(0, nrow=ncol(X))
  b0<-bc<-b<- matrix(0, nrow=ncol(X))
  d2md2b<-matrix(0, ncol=ncol(X), nrow=ncol(X))
  dmdb<-matrix(0, nrow=ncol(X))
  omega<-matrix(0, ncol=ncol(X), nrow=ncol(X))
  M<-matrix(0, ncol=ncol(X), nrow=ncol(X))
  for( i in unique(ID)) {

```

```

n<-length(which(j==i))
vi<-diag(n)
d2md2b<-d2md2b+t(X[which(j==i),])%*%diag(w[which(j==i)])
%*%solve(vi)%*(X[which(j==i),])
dmdb<-dmdb+t(X[which(j==i),])%*%solve(vi)%*%diag(w[which
(j==i)])%*((Y[which(j==i)]-X[which(j==i),]%*%bc)
)
}

#beta differential. INITIAL

bdiff<-solve(d2md2b)%*%dmdb
b0<-bbc<-bdiff

res<-Y-(X%*%b0)
bdiff<-1
iter<-1

#Begins the NR/IRLS loop, stopping toll of .000001.

while (sum(abs(bdiff))>.000001){
#Estimate of correlation parameter alpha.
j<-ID
num<-den<-0
for(i in unique(ID)) {
for(k in which(j==i)[-1]){
num<- num+ w[k]*(res[k-1]**2 + res[k]**
2)
den<- den+(w[k]*res[k-1]*res[k])
}
}
zeta<-(num)/(2*den)

```

```

num1<-den1<-num2<-den2<-0
  for( i in unique(ID)) {
    for(k in which(j==i)[-1]){
      num1<-(num1+w[k]*(res[k-1]+res[k])**2)
      num2<-(num2+w[k]*(-res[k-1]+res[k])**2)
      den1<- (den1+(w[k]*res[k-1]*res[k]))
    }
  }
eta<-sqrt(num1*num2)/(2*den1)
alpha1<-zeta-eta

alpha1

alpha2<- (2*alpha1)/(1+alpha1**2)
#alpha2<-0
alpha2

U<- matrix(0, nrow=ncol(X))
d2md2b<-matrix(0, ncol=ncol(X), nrow=ncol(X))
dmdb<-matrix(0, nrow=ncol(X))
omega<-matrix(0, ncol=ncol(X), nrow=ncol(X))
M<-matrix(0, ncol=ncol(X), nrow=ncol(X))

for( i in unique(ID)) {
  n<-length(which(j==i))
  vi<-(matrix(rep(alpha2**abs(-n:n), n)[-1:n], ncol=n)
    [1:n,1:n])
#  vi<-diag(n)
  d2md2b<-d2md2b+t(X[which(j==i),])%*%diag(w[which(j==i)])
    %*%solve(vi)%*%(X[which(j==i),])
  dmdb<-dmdb+t(X[which(j==i),])%*%solve(vi)%*%diag(w[which
    (j==i)])%*%(((Y[which(j==i)] -X[which(j==i),])%*%bc))
  )

```

```

    }

    bdiff<-solve(d2md2b)%*%dmdb
    #updates beta.
    bc<-bc+bdiff
    iter<-iter+1
    res<-Y-X%*%bc

  }

  #Correct plugin estimator of covariance of beta.
  Mguts<- matrix(0, ncol= length(unique(ID)), nrow= ncol(X))
  t<-1
  for( i in unique(ID)) {
    n<-length(which(j==i))
    vi<-(matrix(rep(alpha2**abs(-n:n), n)[- (1:n)], ncol=n)
      [1:n,1:n])
    omega<-omega+t (X[which(j==i),])%*%diag(w[which(j==i)])%*
      %solve(vi)%*%(X[which(j==i),])
    guts<-t (X[which(j==i),])%*%diag(w[which(j==i)])%*%solve(
      vi)%*%(Y[which(j==i)]-X[which(j==i),]%*%bc)
    Mguts[, t]<- guts
  }

  t<-t+1
}

head(Mguts)

individual<-sapply( 1:length(unique(ID)), FUN = function(n)
  strsplit(unique(id)[n], split='_')[[1]][1])

M<-matrix(0, ncol=ncol(X), nrow=ncol(X))

```

```

for ( i in unique(individual)) {
  M<- M + as.matrix(rowSums(as.matrix(Mguts[,which(individual==i)
    ]))) %*% t(as.matrix(rowSums(as.matrix(Mguts[,which(
    individual==i)]))))
}

#House keeping for summary stats.
std<-as.matrix(sqrt(diag(solve(omega)%*%M%*solve(omega))), ncol=1)
colnames(std)<-"stderr"
wald<-as.matrix(bc**2/(std**2), ncol=1)
colnames(wald)<-"Wald"
p<-as.matrix(1-pchisq(wald,1), ncol=1)
colnames(p)<-"p"

cbind(bc, std, wald, p)
print(iter)
return(geehand(beta=cbind(bc, std, wald, p), alpha=alpha2, Nclus= length
(unique(ID)), iter=iter ))

}

```

B MWGEE estimation using the *genmod* procedure

It is understood that without a readily available software package to handle MWGEE model estimation very few will be willing to adopt the methodology. As such investigations were made into the use of the *genmod* procedure in SAS. The *genmod* procedure has a *weights=* option similar to the *geeglm()* function in R. Our investigation found that unlike the *geeglm()* function the *genmod* procedure with the use of the *weights=* option produced results very similar to the results obtained by the *gausgee()* we wrote for MWGEE estimation. We include here the program for obtaining the marginal parameter estimates as well as a macro for carrying out the bootstrap variance estimation. It should be noted that the *surveyselect* procedure does not return an object containing replicate sampled objects, instead it returns a column indicating the number of times a record was selected. As such when bootstrapping the variance estimate a modification to the MWGEE weights must be made in order to accommodate this such that replicate records are counted the correct number of times.

```
*****
```

```
WGEE investigation
```

```
Kennesaw Presentation Oct. 2014
```

```
Caucasian Sample UNC data
```

```
*****;
```

```
LIBNAME dent 'C:\Users\Joe\Desktop\Fall_2014\Kennesaw_Presentation';
```

```
PROC IMPORT out=dent.white
```

```
    datafile= "C:\Users\Joe\Desktop\Fall_2014\Kennesaw_Presentation\  
    forsas.csv";
```

```
    getnames=Y;
```

```
run;
```

```
PROC IMPORT out=dent.ind
```

```

datafile= "C:\Users\Joe\Desktop\Fall 2014\Kennesaw Presentation\
      indsas.csv";
getnames=Y;
run;

```

```

data dud ;
set dent.white;
set dent.ind ;
run;

```

*sets up crap for loop;

*Similarly , the first iteration of the bootstrap;

*You know how much fun creating objects in SAS is. So, I create the set
which will

contain the output from the bootstrap <agresults> outside of the loop.

At each
iteration the results <doodoo> are appended to <agresults>. ;

*Sampling with replacement by individual;

```

proc surveyselect data= dud out= dummysamp n=237 seed =1 method=URS;
      cluster indi ;
run;

```

*House cleaning (it's a Genmod thing);

```

proc sort data= dummysamp;
      by idwhite indi TIME;
run;

```

*Surveyselect does not like to return replicate records. Instead it
appends

a column which contains the number of times a record was sampled. We can

force the records to receive disproportionate weight (corresponding to the number of replicates by multiplying the marginal WGEE weights by the number of replicates.

That's what happens here.;

```
data dummysamp;
    set dummysamp;
    w = (wwhite*NumberHits);
run;
```

*It's just easier to output the results (parameter estimates) through an ODS statement

than it is to try to coerce a functional output statement. That's what happens here,

the parameter estimates are output to a set called <Parms>;

```
ods select GEEEmpPEst;
ods output GEEEmpPEst=Parms;
PROC GENMOD data=dummysamp;
    class idwhite indi;
    model ywhite = TIME SEX EDUC WHENDDS MARRIED SEIRSP TOBUSE URBAN
        ATOTINC t_SEX t_EDUC
            t_WHENDDS t_MARRIED t_SEIRSP t_TOBUSE t_URBAN t_ATOTINC
                / id=idwhite;
    repeated subject= idwhite /
        type= AR(1) CORRW;
    weight w;
run;
quit;
```

*Flips everything around to allow stacking, easier to deal with than creating a new variable at each iteration.;

```
proc transpose data=Parms out= doodoo;
    id Parm;
    var Estimate;
run;
```

```
*CREATES the set which will hold the aggregate results and subsequently
    the first record.;
data agresults;
    set doodoo;
run;
quit;
```

```
*short macro to bang out the bootstrap. Note that it starts on the
    second iteration;
```

```
*A few things to note
```

```
    The seed is specified as a macro variable (the interation number
    )
```

```
    It's not how I'd like to have done it but I needed
    reproducability and
```

```
    SAS isn't a fan of global seeds.
```

```
    .....It's just a loop that appends the results to the <agresults> set
    . ;
```

```
%macro bs(n);
```

```
    *loop magic;
```

```
    %do i=2 %to &n;
```

```
        proc surveyselect data= dud out= dummysamp n=237 seed =&
            i method=URS;
            cluster indi ;
        run;
        proc sort data= dummysamp;
```

```

        by idwhite indi TIME;
run;
data dummysamp;
    set dummysamp;
    w = (wwhite*NumberHits);
run;

ods select GEEEmpPEst;
ods output GEEEmpPEst=Parms;
PROC GENMOD data=dummysamp;
    class idwhite indi;
    model ywhite = TIME SEX EDUC WHENDDS MARRIED
        SEIRSP TOBUSE URBAN ATOTINC t_SEX t_EDUC
            t_WHENDDS t_MARRIED t_SEIRSP t_TOBUSE t_
                URBAN t_ATOTINC / id=idwhite;
    repeated subject= idwhite /
        type= AR(1) CORRW;
    weight w;
run;
quit;

proc transpose data=Parms out= doodoo;
    id Parm;
    var Estimate;
run;
proc datasets;
    append base = Agresults data= doodoo;
run;
quit;

%end;
%mend bs;

%bs(1000)

```

```

data dent.Agresults;
    set Agresults;
run;

```

```

data bsresults;
    set dent.agresults;
    drop _NAME_;
run;

```

```

ods select GEEEmpPEst;
ods output GEEEmpPEst=Parms;

```

*Produces the point estimates **NOT** resampled (FULL DATA). ;

```

PROC GENMOD data=dud exactonly;
    class idwhite indi;
    model ywhite = TIME SEX EDUC WHENDDS MARRIED SEIRSP TOBUSE URBAN
        ATOTINC t_SEX t_EDUC
            t_WHENDDS t_MARRIED t_SEIRSP t_TOBUSE t_URBAN t_ATOTINC
                / id=idwhite;
    repeated subject= idwhite(indi) /
        type= AR(1) CORRW;
    weight wwwhite;
run;
quit;

```

*Just cleans up the point estimates. ;

```

proc transpose data=Parms out= doodoo;
    id Parm;
    var Estimate;

```

```
run;
data doodoo;
    set doodoo;
    drop _NAME_;
run;
```

```
*Generates from <agresults> and the point estimate of the full data the
    Variances
    SE
    Wald stat
    P values;
```

```
proc IML;
    use bsresults;
    read all var _ALL_ into x;
    close bsresults;
    use doodoo;
    read all var _ALL_ into est;
    close doodoo;

    varnames = {'INT' 'TIME' 'SEX' 'EDUC' 'WHENDDS' 'MARRIED' '
        SEIRSP' 'TOBUSE' 'URBAN' 'ATOTINC' 't_SEX' 't_EDUC'
        't_WHENDDS' 't_MARRIED' 't_SEIRSP' 't_TOBUSE' 't_URBAN'
        't_ATOTINC'};
    statnames = {'Estimate' 'StdErr' 'Wald' 'P'};
    Ser=std(x);
    Wald = (est/Ser);
    Wald= Wald#Wald ;
    P= 1-CDF('chisquare',Wald,1);
    AnovaBS= est // Ser // Wald //P;
    AnovaBS= AnovaBS';
    print (AnovaBS);
    create BSANOVAresults from AnovaBS;
```

```
        append from AnovaBS;

quit;

*Puts everything in a pretty table;
data dent.BSANOVAResults;
        set BSANOVAResults;
run;
```

CURRICULUM VITAE

NAME: Joe Bible

ADDRESS: Department of Biostatistics and Bioinformatics
University of Louisville
Louisville, KY 40292

EDUCATION: B.S. Secondary Math Ed., Kennesaw State University, 2009
M.S. Applied Statistics, Kennesaw State University, 2011

PROFESSIONAL SOCIETIES: American Statistical Association

PUBLICATIONS: **Bible, J.**, Beck, J. D., and Datta, S. Cluster adjusted regression for displaced subject data (CARDS): marginal inference under potentially informative temporal cluster size profiles. Submitted.

Dey, P., **Bible, J.**, Dey, S. and Datta, S. Applications of feature selection and regression techniques in materials design: A Tutorial. In Computational Approaches to Materials Design: Theoretical and Practical Aspects , (Shubhabrata Datta, Ed.), IGI Global, Hershey, USA, appear (2015).

Dey, P., **Bible, J.***, Datta, S., Broderick, S., Jasinski, J., Sunkara, M., Menon, M. and Rajan, K. Informatics

aided band gap engineering for solar materials. Computational Materials Science, 83, 185195 (2014).(*equal contribution)

Bible, J., Datta, S. and Datta, S. Cluster analysis: Finding groups in data. In Informatics for Materials Science and Engineering (K. Rajan, Ed.), Elsevier, pp 5370 (2013).

PRESENTATIONS

ORAL

Cluster Adjusted Regression for Displaced Subject Data Spring 2015 Symposium Series U of L Dept. Biostatistics & Bioinformatics, January 16, 2015.

Colloquium presentation titled Marginal Model Estimation: Weighted Generalized Estimating Equations Kennesaw State Department of Statistics and Analytical Sciences, October 18, 2014.

Colloquium presentation titled Modeling Band Gap of Binary Compounds: Defining the solid solutions problem. Conn Center Renewable Energy Colloquium, August 29, 2014.

Student colloquium presentation titled Informatics Based Bandgap Estimation October, 2013 meeting of the KY chapter of the American Statistical Association.

Co-taught a short course titled Tools for Materials Genome Research at 2013 Kentucky Workshop on Renewable Energy and Energy Efficiency, KY International Convention Center

on March 24, 2013.

POSTER

Joint Statistical Meetings, Boston August 2, 2014.

SRCOS Summer Research Conference 2013, June 4, 2013.

2013 Kentucky Workshop on Renewable Energy and Energy
Efficiency, KY International Convention Center
on March 24, 2013.

FUNDED

RESEARCH:

Somnath Datta and Mahendra Sunkara (PIs). Solar: New
Materials Search for Solar Energy Conversion to Fuels,
National Science Foundation Grant Number: DMS-1125909

Periodontal research supported by
NIH grants 1R03DE020839-01A1, 5R03DE020839-02,
1R03DE022538-01, and 5R03DE022538-02
(Somnath Datta, PI)