MODEL SELECTION IN GENERALIZED ESTIMATING EQUATIONS BASED ON KULLBACK'S I-DIVERGENCE

BY

ROBERT NYAMAO NYABWANGA

A THESIS SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN APPLIED STATISTICS

SCHOOL OF MATHEMATICS, STATISTICS AND ACTUARIAL SCIENCE

MASENO UNIVERSITY

Declaration

This thesis is my own work and has not been presented for a degree award in any other institution.

ROBERT NYAMAO NYABWANGA

PHD/MAT/00054/2015

Signature:.....Date:.....

This thesis has been submitted for examination with our approval as the university supervisors.

PROF. FREDRICK ONYANGO

Department of Statistics and Actuarial Science

Maseno University

Signature:.....Date:.....

DR. EDGAR OUKO OTUMBA

Department of Statistics and Actuarial Science

Maseno University

Signature:.....Date:.....

Acknowledgement

I sincerely thank my Supervisors Prof. Fredrick Onyango and Dr. Edgar Otumba for their insightful direction, constant encouragement and kind support throughout the entire duration of my thesis research. They have guided me to an interesting research area, provided me tremendous help along the way, and constantly reassured me of my progress. I am also thankful to Prof. Kepher Makambi of GeorgeTown University, USA for for providing interesting research insights into the area of longitudinal data modeling and model selection. Special thanks to Maseno University for providing partial funds in support of this study. Sincere appreciation is extended to the faculty, staff and students of the Department of Statistics and Actuarial Science of Maseno University, who have helped me in different ways.

Dedication

To my parents Fredrick Nyabwanga and Paskalia Mogoi, my wife Verah, my daughters Katrina and Mitchelle and son Ryan Francis.

Abstract

The method of Generalized Estimating Equations (GEE) is often used in analyzing correlated longitudinal data and does provide consistent estimates which are robust to misspecification of the working correlation structure. However, the estimates suffer loss of efficiency if the correlation structure is not close to the true one hence the models selected may not be generalizable, good-fit and parsimonious. The Quasi-likelihood information criterion (QIC) which results from utilizing Kullback's I-divergence as the targeted discrepancy is widely used in the GEE framework to select the best correlation structure and the best subset of predictors. However, it has been established to have success rates of less than 50% hence higher chances of selecting a misspecified structure. Use of a mis-specified structure results in efficiency loss in the GEE estimator of up to 40% compared to when the correct correlation structure is used. Also, the independence structure favored by QIC, results in efficiency loss of up to 60% in the GEE estimates. Through numerical simulations, the study sought to investigate the properties of QIC in selecting the true working correlation structure and set of covariates for the mean structure in GEEs, develop a hybrid methodology based on empirical likelihood Akaike Information Criteria (EAIC) and QIC for model selection in the GEE framework and apply the proposed hybrid methodology to the Shareholder Value Creation data. With regard to consistency in selecting the true correlation structure, we established having a selection set of only parsimonious structures and penalizing for the number of correlation and regression parameters estimate to be sufficient conditions for QIC to select the true structure with a probability approaching one as $n \to \infty$. In relation to the selection of covariates, we established that QIC had high sensitivity but low sparsity. The type I error rate converged to 0.3 as $n \to \infty$ while the type II error rates quickly diminished to zero as $n \to \infty$. The low under-fitting probabilities meant high statistical power hence rejecting any given false null hypothesis is essentially guaranteed for sufficiently large n even if the effect size is small. We further established that the hybrid methodology (EQ_{AIC}) resulted in models with lower MSE compared to models selected by QIC only. When applied to shareholder value creation data, we established an AR-1 correlation structure for the data with $\rho = 0.775$ and the key drivers to shareholder value creation ranked based on their relative importance were the growth rate of earnings, economic spread, firm size, leverage, dividend policy and level of financial distress. This justified the tendency of QIC to over-fit models since a more complex model compared to the Gordon Constant Growth model was preferred. However, the use of an AR-1 correlation structure selected by EAIC resulted to a model with lower MSE than the model selected by using QIC only. Based on the study findings we conclude that correctly specifying a working correlation structure improves efficiency of GEE estimates.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS AND ACRONYMS

ρ	Nuisance correlation parameter in GEE
AIC	Akaike Information Criteria
QIC	Quasi-Likelihood Information Criterion
EAIC	Empirical Likelihood based AIC
EQ_{AIC}	Hybrid Methodology of EAIC and QIC
QIC_m	Modified QIC
β	Regression parameters to be estimated
$\hat{\beta}$	An estimate of β
$\hat{\beta}_G$	GEE estimate of β
Corr(.)	Correlation
Cov(.)	Covariance
CV	Cross-Validation
D	Matrix of Partial derivatives
SE	Standard Error
MLE	Maximum Likelihood Estimator
MSE	Mean Squared Error
n	Number of subjects
m	Number of measurements per subject
g(.)	Link function
GLM	Generalized Linear Model
GEE	Generalized Estimating Equations
q	Number of correlation parameters
p	Number of regressors
$\dim(\theta)$	Total number of parameters estimated
$R(\rho)$	Correlation structure parameterized by ρ
Θ	Parameter space
$\Theta(k)$	k-dimensional parameter space

IN	Independence Correlation Structure
EX	Exchangeable Correlation structure
AR-1	Autoregressive structure of order 1
UN	Unstructured Correlation structure
Toep	Toeplitz correlation structure
R_0	True Working Correlation Structure
R_*	Selected Correlation structure
R(F)	Empirical Likelihood Ratio
$\Re(.)$	Empirical Likelihood Ratio for (.)
$I_{f_0 f}$	Kullback's I-divergence
log(.)	Natural logarithm (base e)
L(.)	Likelihood function
$\ell(.)$	Log-likelihood
N	Total number of observations
O_p, o_p	Notation for random sequences
O(.), o(.)	Asymptotic notation of sequences
QL(.)	Log-Quasi-likelihood
X_{itp}	p^{th} covariate measured on i^{th} subject
y_{it}	Response value for i^{th} subject at time t
$(.)^T$	Transpose of (.)
MELE	Maximum Empirical Likelihood Estimate
MVUE	Minimum Variance Unbiased Estimators
WCS	Working Correlation Structure
k_e	Cost of Equity
ROE	Return on Equity
MV	Market Value
BV	Book Value
P_r	Probability

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

1.1.1 Introduction to Longitudinal Data Modeling

Longitudinal data comprise measurements taken repeatedly over time from the same cases. This as observed by Avital et al. [4] imply that each subject is measured repeatedly either under different conditions or at different times or both with the main interest of characterizing the way the outcome changes over time, and the predictors of that change. For instance, if we consider a longitudinal study with independent subjects for which m measurements are taken for each subject, then the t^{th} measurement collects the response y_{it} and a set of covariates $X_{it} = [X_{1it}.....X_{pit}]$. If we let $\sum_{i=1}^{n} m_i = N$ be the total number of observations, then a typical longitudinal dataset will be illustrated as in Table 1.1:

Table 1.1: Structure of a Typical Longitudinal Dataset

Subject	Observation	Response	Explanatory Variables			
1	1	y_{11}	X_{111}	.	.	X_{p11}
1	2	y_{12}	X_{112}	.	.	X_{p12}
.	.	.	.	.	.	.
1	m	y_{1m}	X_{11m}	.	.	X_{p1m}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
n	1	y_{n1}	X_{1n1}	.	.	X_{pn1}
n	2	y_{n2}	X_{1n2}	.	.	X_{pn2}
.	.	.	.	.	.	.
n	m	y_{nm}	X_{1nm}	.	.	X_{pnm}

Presence of repeated measures imply intrinsic correlation for observations from the

same subject and ignoring the correlation while analyzing such data can lead to misleading, inefficient or invalid inference (Diggle et.al. [18]). The dependence might be short term in which case correlation becomes weak after a certain time-lag or long term if it lasts being strong for most of the time-lags.

Compared to time series and cross-sectional data, Ilk [40] asserts that Longitudinal data allows for the measurement of change hence makes inferences drawn from such data consistent with future studies. This implies that methodologies have to be developed to characterize the way the outcome change over time and the predictors of that change. For example in clinical trials that aim to investigate the efficacy of a new drug in treating a disease, it is often of interest to examine the pharmacokinetic behaviour of the drug when it is applied to experimental subjects. Most drugs do not have constant efficacy over time and such time-varying treatment effectiveness can only be examined through a longitudinal study. Avital et al. [4] affirmed that longitudinal data contain short series hence does not need the stationarity assumption and does not need to be collected at equispaced time points.

Diggle et al. [18] observed that analyzing of longitudinal data based individual time series trajectories helps in separating the cohort and age effects hence can characterize change over time within individuals (age effect) from differences among the subjects in reference to their baseline status (cohort effect). Further, collecting repeated measures from a single subject may help reduce the burden of recruiting a sizable number of subjects for a cross-sectional study. For example in studies of rare diseases, the number of patients available is insufficient for simple randomized trials.

Cho [13] observed that the most important feature of longitudinal data is that they are highly correlated hence makes it difficult to specify the full likelihood function when responses are non-normal. This makes the estimation of covariance structure that defines the within subject correlation the core issue in the analysis of longitudinal data since it will improve estimation efficiency hence better predictive ability of a model.

For longitudinal data with a non-normal response, Liang and Zeger [49] proposed a class of Generalized Estimating Equations (GEE) to model both univariate longitudinal

continuous and discrete outcomes by extending the quasi-likelihood method of Wedderburn [80] to correlated data. The quasi-likelihood is a methodology for regression that requires the specification of relationships between mean response and covariates and between mean response and variance. Thus it does not assume a probability distribution as in the case of a full likelihood.

According to Fitzmaurice et al. [25], GEE is a population-level approach that models the mean response across the population of subjects at each time point as a function of covariates hence provides the population-averaged estimates of the parameters. It only requires the specification of the first two moments of the response variable and a tentative "working" structure for the covariance among repeated responses. However, it creates difficulty in model selection since many traditional model selection criteria, such as AIC and BIC, need to be redefined because of the within-subject correlation of the observations and lack of an explicit likelihood function.

The within-subject correlation is accounted for by defining a working correlation structure $(R(\rho))$, where ρ is a vector of parameters that characterize the structure. Useful working correlation matrices include independence i.e $Corr(y_{ij}, y_{ik}) = 0, \forall j \neq k$; exchangeable i.e. $Corr(y_{ij}, y_{ik}) = \rho, \forall j \neq k$; Toeplitz i.e. $Corr(y_{ij}, y_{i,j+t}) = \rho_t, for j = 1, 2...n_i - t$; unstructured correlation matrix i.e. $Corr(y_{ij}, y_{ik}) = \rho_{jk}, \forall j > k$ and the first-order AR-1 working correlation structure in which $Corr(y_{ij}, y_{ik}) = \rho^{|j-k|}, \forall\ j > k$. Liang and Zeger [49] assert that GEE yields asymptotically consistent $\hat{\beta}$ even when $R(\rho)$ is misspecified. However, if it is correctly specified, the efficiency of the estimates will be greater(Fitzmaurice et al. [25], Wang and Carey [79], Sutradhar and Das [74]).

1.1.2 Model selection Framework and Principles

Suppose that the true and candidate models are respectively of the form

$$\mathbf{Y} = X_0\beta_0 + \varepsilon_0, \quad \varepsilon_0 \sim N(0, \sigma_0^2) \tag{1.1}$$

$$\mathbf{Y} = \mathbf{X}\ \beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \tag{1.2}$$

where Y_i=$[Y_{i1}, ...Y_{im_i}]^T$, i=1,2....n. Y_i are assumed to be independent, but observations within the subject are not assumed to be independent. Further, we will assume that

Y_{it} is a binary outcome for subject i at time t. X_{0it}=$[X_{0it_1}, ...X_{0it_p}]^T$ is $n \times p_0$ vector of covariates of rank p_0 which can be collected into a $m_i \times p_0$ matrix of covariates $[X_{0i1}^T, ...X_{0im}^T]^T$ while X_{it}=$[X_{it_1}, ...X_{it_p}]^T$ is $n \times p$ vector of covariates of rank p which can be collected into a $m_i \times p$ matrix of covariates $[X_{i1}^T, ...X_{im}^T]^T$. β_0=$[\beta_{0_1}.....\beta_{0_p}]^T$ and β=$[\beta_0, \beta_1.....\beta_p]^T$ are $p_0 \times 1$ and $p \times 1$ parameter vectors of respective regression coefficients; ε_0 and ε are noise vectors.

Let θ_0=(β_0^T,σ_0^2) and θ_k=(β^T,σ^2) define the parameter space for θ_0 and θ_k respectively such that their MLE are denoted by $\hat{\theta}_0$=($\hat{\beta}_0^T$,$\hat{\sigma}_0^2$) and $\hat{\theta}_k$=($\hat{\beta}^T$,$\hat{\sigma}^2$) respectively. The subsequent empirical likelihoods for the true, candidate and fitted models are $f_0(y|\theta_0)$, $f(y|\theta_k)$ and $f(y|\hat{\theta}_k)$ respectively. Further if we let $\digamma(k) = \{f(y|\hat{\theta}_k)|\theta_k \in \Theta_k\}$ be the family of k-dimensional densities corresponding to candidate models (1.2), then model selection seeks to search among a collection of classes $\digamma = [\digamma(K_1)...\digamma(K_L)]$ for the fitted model $f(y|\hat{\theta}_k)$, $k \in \{1...K_L\}$ which serves as the best approximation of $f_0(y|\theta_0)$ i.e. the fitted candidate model nearest to the true model. With p predictors in the equation (1.2), the total number of candidate models is 2^p and as p increases, identifying the optimal fitted model within the large model space can be computationally burdensome.

Akaike [2], observed that the model selected should be generalizable, a good-fit and parsimonious. A generalizable model is one that predicts future observations with a high degree of confidence and as observed by Konishi and Kitagawa [46], such a model should not differ from one supposed to depict the structure of the true model. Burham and Anderson [7] also suggested that striving for generalizability should be the key model selection objective.

The goodness-of-fit principle requires that the fitted candidate model conforms to the data used to construct it thus giving a measure of divergence between observed outcomes employed to construct the model and expected values under the fitted model. According to Koniski and Kitagwa [46], a model can fit the data very well because it is excessively complicated hence the need to consider a parsimonious model. Therefore, model selection requires that a balance between goodness-of-fit and parsimony be determined so that a model that captures the most informative features of the generating

model be preferred.

Under-fitting and over-fitting are also pertinent concepts in determining the quality of the fitted model. Burham and Anderson [7] assert that under-fit models will fail to include all the important variables leading to biased estimates and poor predictive performance. Further, such models will provide an incomplete representation of the generating model. On the other hand, over-fit models will incorporate all the important variables plus some spurious ones leading to models with high variability hence less precision of the parameter estimates. Such models are unnecessarily complex, difficult to interpret and subject to excessive sampling error. However, Burnham and Anderson [7], observes that modest over-fitting is less damaging than under-fitting since it is less detrimental to include a non-informative variable in a correctly specified model(over-fitting) compared to the failure to include an informative variable (under-fitting).

1.1.3 Model Selection in the GEE Framework

In the GEE framework, model selection focuses more on selecting the working correlation structure $R(\rho)$ and suitable covariates for the mean structure. Even though GEE approach yields consistent estimators of the model parameters even if the correlation structure $R(\rho)$ is misspecified with large n, its misspecification yields inconsistent estimates of the correlation parameters ($\hat{\rho}$) which in turn compromises the consistency of $\hat{\beta}$, leads to inflated variance estimate and eventual loss in efficiency. Therefore, as asserted by Kaurmann and Carroll [45], the asymptotic relative efficiency depends on the correct specification of $R(\rho)$. This is emphasized by Fitzmaurice et al. [25] who asserted that the robustness property of the sandwich variance estimator to misspecification of $R(\rho)$ cannot be assumed to hold in all situations. For instance, they established that when the number of subjects (n) is small and the number of repeated measures (m) for each subject is large, sandwich variance estimator is not appropriate.

Wang and Carey [79] affirmed that the asymptotic relative efficiency of the GEE parameter estimates is likely to be low when the working correlation structure is misspecified. These assertions were corroborated by Sutradhar and Das [74] who also pointed

out that, the mis-specification of a correlation structure lowers the relative efficiency of the estimate even when the sample size is finite since the Cramer-Rao lower bound variance estimate will not be achieved with an inconsistent estimate of ρ. Fitzmaurice et al. [25] recommended for efforts to be made for correct modeling of the within-subject correlation to be assured of both the consistency and efficiency of the GEE estimates.

1.1.4 Examples

We use datasets to show why using correct correlation structure in GEE analysis is important in variable selection.

Example 1.1.1. To verify the robustness of the GEE to mis-specification of the working correlation structure, an example dataset (Ohio dataset) from the geepack library was analyzed using different correlation structures. The data analyzed the health effect of air pollution on children who were followed for four years. The wheeze status, age and smoking status of mothers were recorded for 537 individuals resulting to 2148 observations. The Regression coefficients and standard errors estimated by GEE analysis with different correlation structures are presented in Table 1.2.

Table 1.2: Regression coefficients, standard errors and p-values

	CORRELATION STRUCTURE			
	Exchangeable	Unstructured	Independence	AR-1
Intercept	-1.88(0.11)(0.00)	-1.89(0.11)(0.00)	-1.88(0.14)(0.00)	-1.90(0.12)(0.00)
Age	-0.11(0.04)(0.01)	-0.12(0.04)(0.01)	-0.11(0.04)(0.01)	-0.11(0.05)(0.01)
Smoke	0.27(0.17)(0.15)	0.25(0.17)(0.16)	0.27(0.17)(0.13)	0.23(0.18)(0.20)

From Table 1.2, it can be seen that, although the conclusions based on p-values are the same, there are some differences in the magnitude of the regression coefficients. This is important, because it is far more interesting to estimate the magnitude of the association by means of the regression coefficients and the 95% confidence intervals than just estimating p-values. Further, it is observed that the estimated standard errors of the different parameters are not very similar for various correlation structures. The assumption of AR-1 correlation within the responses inflated the standard errors

than the other correlation structures hence could be inferred to be far from the correct structure.

Example 1.1.2. In the finance field, a dataset for a study by Odongo et al. [57] that examined the relationship between Shareholder value creation and the predictors: firm size, leverage, liquidity and board size is analyzed using different correlation structures. The shareholder value creation which is ratio between the market value (MV) of shares and their book value (BV) was measured as a dichotomous variable in which Shareholder value creation was 1 if $\frac{MV}{BV} > 1$ and 0 if $\frac{MV}{BV} \leq 1$. The data were recorded for 6 agricultural firms listed in the NSE for a period of 6 years (2011-2016). The Regression coefficients and standard errors estimated by GEE analysis with different correlation structures are presented in Table 1.3.

Table 1.3: GEE Regression Coefficients, SE and p-values for Shareholder Value Creation Data

	CORRELATION STRUCTURE			
	Exchangeable	Unstructured	Independence	AR-1
Intercept	-35.4(8.40)(.000)	-1.5e+16(7.8e+15)(.082)	-40.7(11.45)(.000)	-38.2(11.4)(.000)
Firmsize	1.99(0.43)(.000)	9.4e+14(4.4e+14)(.047)	2.37(0.66(.000))	2.26(0.68)(.000)
Liquidity	2.01(0.58)(.022)	3.8e+14(3.7e+13)(.000)	1.24(0.52)(.043)	1.05(0.47)(.030)
Leverage	69.2(42.01)(.079)	9e+16)(2e+15)(.000)	55.4(32.7)(.062)	63.1(34.5)(.051)
Boardsize	-1.73(0.31)(.000)	-9.3e+14(2.8e+14)(.000)	-1.86(0.58)(.007)	-1.86(0.61)(.003)

From Table 1.3 it can be seen that, the conclusions based on p-values are not the same under the different correlation structures. For example at $\alpha = 0.01$, liquidity is not significant under the exchangeable, independence and AR-1 structures but is significant when the unstructured correlation structure is assumed. Also, leverage is significant only under the unstructured correlation. Further, there are differences in the magnitude of the regression coefficients and the standard errors. Assuming the unstructured correlation matrix greatly inflates the standard errors of the predictor variables hence the GEE estimators are the least efficient. For instance, the efficiency of $\hat{\beta}_{UN}$ relative to $\hat{\beta}_I$, $\hat{\beta}_{EX}$ and $\hat{\beta}_{AR-1}$ is lower. Moreover, these GEE estimates ($\hat{\beta}_{UN}$)

deviate greatly from the estimates under the other correlation structures.

The results from Tables 1.2 and 1.3 above show that it is important to choose a suitable correlation structure before a GEE analysis is performed which as observed by Sakate and Kashi [67], will result to substantial improvement in efficiency of the GEE estimator ($\hat{\beta}_G$). Different criteria have been developed under different assumptions and for different statistical frameworks and data types.

Akaike information criterion (AIC) by Akaike [3] whose derivation utilized the Kullback's I-Divergence ($I(\theta_0, \theta_k)$) that measure the separation between the true model and a fitted model is widely used in most modeling frameworks for model selection. It's justification relies upon the conventional large-sample properties of maximum likelihood estimators. It is a measurement tool of the goodness of fit of an estimated statistical model based on information theory. The chosen model is the one that minimizes the Kullback-I divergence between the model and the truth, and the criterion is used to describe the trade-off between bias and variance in model construction. However, since GEE is not likelihood based, the use of AIC for model selection in the GEE framework is not possible.

Pan [60] by replacing the log-likelihood in AIC by the log-quasi-likelihood and re defining the penalty term, proposed the Quasi-likelihood Information Criterion (QIC) for model selection in GEE and recommended for its routine use to select the correct set of covariates for the mean structure and a working correlation structure. However, he observed that QIC was not very powerful in choosing a working correlation structure due to the fact that $Q(\beta, I)$ does not contain any information about the within-subject correlation. Similar views were held by Hin and Wang [34] who asserted that QIC was heaviliy impacted by its first term hence was not a good criteria to use in selecting a working correlation structure. Moreover, QIC's selection rates have been established to be less than 50% in most simulation studies. For example Barnett et al. [5] noted that its overall success rate was 29.4% and was biased towards selecting the unstructured correlation structure which estimates the highest number of nuisance parameters while Hyun-Joo et al. [39] established success rates of less than 25% for AR-1 and less than

45% for unstructured matrix.

The low performance of QIC has led to the conclusion that it is not powerful in choosing the correct correlation structure and as asserted by Fitzmaurice [26], the resulting estimator from a mis-specified correlation structure may be 40% less efficient compared to the estimator obtained by using the correct correlation structure.

Shinpei [71] attributed the low performance of QIC in selecting the working correlation structure to the non-consideration of the correlation parameter by Pan [60] in his derivation of QIC in which he only considered the bias that arise when estimating β. Further, Hin et al. [35] showed that the resultant criteria based on the bias correction term of QIC only such as the correlation information criterion (CIC) performed better than QIC in choosing the true correlation matrix. Since an inappropriate correlation structure may significantly impair the efficiency of $\hat{\beta}$, it is important to select the working correlation structure most appropriate for the data at hand with the ultimate aim of improving efficiency of estimates.

Jang [42] affirmed that no single model selection criteria exists that can with high success rates select the true correlation matrix, correct set of covariates and variance function in GEE modeling hence recommended that future studies should focus on combining proposed model selection criteria so as to develop model selection strategies that could improve optimality of the GEE models. Based on his recommendation, Erfanul et al. [22] applied a combination of CIC and QIC in selecting selecting the correlation structure and relevant covariates respectively in their study that sought to establish the impact of height on the occurrence of type II diabetes. However, they did not establish whether employing the combined methodology improved efficiency of estimates and this formed the basis for the hybrid methodology proposed in the study.

1.1.5 Properties of QIC in GEE model Selection

Fan and Li [23] observed that a good model selection criteria should be asymptotically consistent i.e. provided the correct model is included in the set of candidate models, it should identify that correct model asymptotically with probability one. Dziak [19]

observed that, for consistent model selection, two properties are required: sensitivity and sparsity. Sensitivity implies that the model selection criteria should retain all of the coefficients which should be retained with a probability approaching one while sparsity implies that the model selection criteria should delete all of the coefficients which should be deleted, with probability approaching one. Hirokazu et, al. [36] in their study which sought to establish whether as the sample size approaches infinity, AIC selects the true model with a probability approaching one established that AIC whose extension to the GEE framework resulted to QIC is not consistent in model selection.

In the GEE framework, consistency can be established for the selection of the true working correlation structure and selection of variables. Pan [60] developed QIC but did not establish its consistency properties in selecting the working correlation structure and variables. He however, noted that QIC was not good in selecting the correlation structure but was good in selecting the variables. Shinpei [71] examined the properties of QIC in GEE and established that the bias of QIC increased with an increase in the number of parameters. Other scholars like Carey and Wang [9], Hardin and Hilbe [32], Hin et al. [35], and Jang [42] focused primarily on establishing the success rates of QIC in selecting the correct working correlation structure compared to other selection criteria such as RJ, CIC e.t.c. Establishing theoretically or verifying numerically the consistency property or conditions for consistency of QIC has received little attention despite the importance of GEE in modeling longitudinal data. Moreover, little or no studies have established the sensitivity and sparsity of QIC in selecting the true generating model and this study sought to fill the void.

In formulating the consistency framework for QIC, indicators such as the within-subject correlation, number of measurements per subject and sample size need to be taken into account. Breslow [6] noted that increasing the within subject correlation and number of measurements will eliminate the bias in the variance component hence will have an effect on the performance of a model selection tool. However, the simulation study by Pan [60] never examined how variations in the number of measurements per subject affected the performance of QIC. In any case, he considered the independence

structure which assumes no within-subject correlation. The within subject correlations can vary from slightly correlated to heavily correlated and as observed by Shinpei [71], incorporating the correlation parameter into the derivation of QIC will improve its performance. In our simulations, we sought to establish the effect of increasing the number of measurements per subject and level of correlations on the consistency of QIC.

1.2 Statement of Problem

Model selection whose main objective is to choose the most generalizable model that balances the increase in fit against the increment in model complexity plays an important role in statistical literature. To facilitate the selection process, a variety of model selection criteria have been developed and are employed for the selection of the most appropriate models. Despite AIC being the most popular model selection criteria, it cannot be applied directly for model selection in the GEE framework since GEEs are not likelihood based. By adopting the quasi-likelihood approach, QIC was proposed for the selection of both the working correlation matrix and covariates. However, simulation studies have shown that it is much impressive in variable selection than in the selection of the true correlation matrix as its success rates in the latter have been established to be far less than 50% hence a high likelihood of selecting a mis-specified structure which results up to 40% loss of efficiency in GEE estimators compared to when the correct structure is used. Also, the independence structure favored by QIC results in efficiency loss of up to 60% in the GEE estimates

Most previous studies have primarily focused on establishing the success rates of QIC in selecting the correct working correlation structure compared to other selection criteria. Establishing theoretically or verifying numerically the consistency of QIC has received little attention despite the importance of GEE in modeling longitudinal data. Moreover, little or no studies have established the sensitivity and sparsity of QIC in selecting the true generating model hence the understanding of the bias-variance trade-off for QIC remains scanty. Further, efforts to improve performance of QIC in selecting

the true structure has resulted to numerous modifications of QIC such as CIC, mQIC and fQIC with little focus on how to use them to improve efficiency.

This study therefore sought to investigate the properties of QIC in selecting the true working correlation structure and covariates for the mean structure in GEE with focus on its consistency in selecting the true working correlation structure and correct generating model and its over-fitting and under-fitting probabilities. Further the study sought to develop a hybrid model selection procedure that deploys empirical likelihood and quasi-likelihood information criteria with the aim of improving prediction performance of models selected by QIC. We sought to demonstrate that extending of empirical likelihood method to GEE enhances the chances of selecting the correct correlation structure and eventual increase in efficiency of the GEE estimates.

1.3 Objectives of the Study

1.3.1 General Objective

The main objective of the study was to investigate model selection criteria in the GEE framework based on Kullback's I-divergence.

1.3.2 Specific Objectives

The specific objectives of the study were to:

(i) Investigate the properties of QIC in selecting the true working correlation structure in generalized estimating equations.

(ii) Investigate the Properties of QIC in selecting Covariates for the mean structure in generalized estimating equations.

(iii) Develop a hybrid methodology based on empirical likelihood Akaike Information Criteria (EAIC) and QIC for selecting models in the GEE framework

(iv) Apply the proposed hybrid methodology to select the firm specific covariates that influence the shareholder Value creation for public listed firms in the Nairobi Securities Exchange

1.4 Significance of the Study

Selecting an appropriate set of important variables helps reduce the variances of parameter estimates and by eliminating some noise variables, precision of the estimates are greatly improved. In the GEE framework, selection of the the correct correlation structure for the response variable increases the relative efficiency of the estimate even when the sample size is finite.

To statisticians, this thesis sought to provide an understanding of the bias-variance trade-off for QIC. This will enable modelers to determine the right contexts for using QIC and determining appropriate strategies for enhancing its performance. In academia, the study contributes to the ever growing knowledge in the area of variable selection in the GEE framework.

In finance literature, much of the modeling has often applied ordinary least square regression to establish relationships. Using GEE which accounts for the within-firm correlation to model the drivers of shareholder value creation will be helpful to the existing shareholders and the prospective ones in making valuable investment decisions based of the soundness of the firms

1.5 Justification of the Study

The study is justified as it provides detailed insights on the consistency, sensitivity and sparsity of QIC in model selection in the GEE framework which goes beyond the comparison of its performance with other criteria provided in most literature. Also, the study proposes a modification to Pan [60]'s QIC to incorporate the number (p) of regression parameters and the number (q) of correlation parameters as cost components into the penalty term so as to enhance its chances of selecting a parsimonious correlation structure. Further, the study proposes an hybrid methodology involving EAIC and QIC to improve efficiency of models selected using QIC. These would stimulate further studies in this area of statistics.

1.6 Mathematical Concepts

1.6.1 Basic Generalized Linear Model (GLM) Concepts

GLMs were developed as an extension to linear models, to allow for more complex relationships between the response and explanatory variables, e.g. binary or count data. They have three main components:

(i) A random component, specifying the conditional distribution of the response variable, y_i given the explanatory variables i.e. a family, or distribution e.g. the exponential family.

(ii) A linear function of the regressors, called the mean structure defined as;

$$\begin{aligned} \eta_i &= \beta_0 + \beta_i X_{i1} + ... + \beta_k X_{ik} \\ &= X_i^T \beta \end{aligned} \tag{1.3}$$

on which the expected value μ_i of y_i depends.

(iii) An invertible link function $g(\cdot)$ that is strictly monotonic and differentiable such that:

$$g(\mu_i) = \eta_i, \quad E(Y_i) = \mu_i, \quad Var(Y_i) = \phi v(\mu_i) \quad and \quad \mu_i = g^{-1}(\eta_i) \tag{1.4}$$

Examples of link functions include identity link for Normal distribution, logit or probit link for Binomial, log link for Poisson distribution e.t.c.

Exponential Family

Let $Y_i (i = 1, ..., n)$ be outcomes for n subjects. If Y_i comes from the exponential family of distributions, then according to Wedderburn and Nelder [52], its probability density function, or probability mass function takes the form:

$$f_Y(y|\theta, \phi) = exp\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\}, \quad -\infty < y < \infty \tag{1.5}$$

The function (1.5) is called the exponential dispersion family where θ is the natural or canonical parameter of the distribution, ϕ is the scale or dispersion parameter and

a, b and c are known functions. This depends on the unknown parameters μ_i and ϕ. From equation (1.5) we have the following expression

$$f_Y(y|\theta,\phi) = exp[\frac{y\theta}{a(\phi)}] \times exp[\frac{-b(\theta)}{a(\phi)}] \times exp[c(y,\phi)] \tag{1.6}$$

If we let $d(\theta) = \frac{\theta}{a(\phi)}$, $a(\theta) = exp[\frac{-b(\theta)}{a(\phi)}]$ and $b(y) = exp[c(y,\phi)]$, equation (1.6) simplifies to a form called the natural exponential family which is sufficient for basic discrete data models given as:

$$f_Y(y|\theta) = a(\theta)b(y)exp[yd(\theta)], \quad -\infty < y < \infty \tag{1.7}$$

For equation (1.5), the log-likelihood is:

$$\ell(\theta) = \{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\} \tag{1.8}$$

such that

$$\frac{d\ell(\theta)}{d\theta} = \{\frac{y - b^{'}(\theta)}{a(\phi)}\} \tag{1.9}$$

Evaluating the expected value of equation (1.9) we have.

$$E(\frac{d\ell(\theta)}{d\theta}) = \{\frac{E(y) - b^{'}(\theta)}{a(\phi)}\} \tag{1.10}$$

Since $E(\frac{d\ell(\theta)}{d\theta}) = 0$, equation (1.10) can be simplified to:

$$E(Y) = \mu = b^{'}(\theta) \tag{1.11}$$

Therefore, the mean of any exponential family random variable, is the first derivative of the cumulant function $[b(\theta)]$ whose form depends on a particular distribution and θ is a function of the mean μ (Wedderburn and Nelder [52]).

Taking the second derivative of equation (1.8) we have;

$$\frac{d^2\ell(\theta)}{d\theta^2} = \{\frac{-b^{''}(\theta)}{a(\phi)}\} \tag{1.12}$$

and since $E\{\frac{d^2\ell(\theta)}{d\theta^2}\} = -E\{\frac{d\ell(\theta)}{d\theta}\}^2$, it follows that;

$$\frac{b^{''}(\theta)}{a(\phi)} = \frac{E\{[Y - b^{'}(\theta)]^2\}}{a(\phi)^2} \tag{1.13}$$

When re-organized equation (1.13) yields the relation

$$a(\phi)b^{''}(\theta) = E\{[Y - b^{'}(\theta)]^2\} \tag{1.14}$$

This leads to the second useful general result:

$$Var(Y) = b^{''}(\theta)a(\phi) \tag{1.15}$$

$a(\cdot)$ could be any function of ϕ.

When ϕ is unknown, there is need to write $a(\phi) = \frac{\phi}{\kappa}$, such that equation (1.16) becomes;

$$Var(Y) = \frac{b^{''}(\theta)a(\phi)}{\kappa} \tag{1.16}$$

Where κ is a known constant that is 1 in most cases.

Since μ and θ are linked via equation (1.11), we can define a variance function of Y in terms of μ as

$$Var(Y) = \phi V(\mu) \tag{1.17}$$

Distributions such as the Bernoulli, Binomial, Multinomial, Poisson, Negative Binomial, Normal, Geometric, Gamma and Inverse Gaussian are members of the exponential. ϕ is fixed at 1 for the Poisson and binomial distributions.

If the term $c(y, \phi)$ in the log-likelihood is available explicitly, the full likelihood can be used to estimate β and ϕ jointly. But often $c(y, \phi)$ is not available hence estimation of ϕ needs a special consideration. One can simply estimate ϕ using the Chi-Square statistic divided by the appropriate degrees of freedom. The Chi-Square statistic is asymptotically unbiased if the model is correctly specified.

Example 1.6.1. Bernoulli Model

For dichotomous outcomes we assume that $y_i \sim Bernoulli(\pi_i)$ for i=1,2....n with $E(y_i) = \pi_i$. y_i is a realization of a random variable Y_i that can take the values one and zero with probabilities π_i and $1 - \pi_i$ respectively. The probability mass function of Y_i is written in compact form as:

$$f_Y(y, \pi_i) = \pi^y(1 - \pi)^{1-y} \mid y \in (0, 1) \tag{1.18}$$

which can also be written as:

$$f_Y(y,\pi_i) = exp\{ylog(\frac{\pi}{1-\pi}) + log(1-\pi)\} \tag{1.19}$$

Set $\theta = log(\frac{\pi}{1-\pi})$ i.e.$\pi = \frac{e^\theta}{1+e^\theta}$; $1-\pi = \frac{1}{1+e^\theta}$ and ϕ=1;

Therefore; $b(\theta) = log(1+e^\theta)$ hence;

$$b'(\theta) = \frac{e^\theta}{1+e^\theta} = \pi = \mu \tag{1.20}$$

and

$$b''(\theta) = \frac{e^\theta}{(1+e^\theta)^2} = \pi(1-\pi) \tag{1.21}$$

This implies that $V(\mu) = \mu(1-\mu)$. The mean and variance depend on the underlying probability π_i. Any factor that affects the probability will alter not just the mean but also the variance of the observations. This suggest that a linear model that allows the predictors to affect the mean but assumes that the variance is constant will not be adequate for the analysis of binary data.

Using the logit link function $g(\pi) = log(\frac{\pi}{1-\pi})$, the generalized linear model becomes

$$logit(\pi) = log(\frac{\pi}{1-\pi}) = X_i^T\beta \tag{1.22}$$

which is a logistic regression model that specifies a linear structure for the log odds or logit.

Example 1.6.2. Binomial Model

Let $Y \sim Bin(m,\pi)$ be a binomial random variable with m>0, fixed integer with density;

$$f_Y(y,\pi) = \begin{pmatrix} m \\ y \end{pmatrix} \pi^y(1-\pi)^{m-y} \mid y \in \{0,1,..,m\}, \quad 0 \le \pi \le 1 \tag{1.23}$$

Which can equally be written as:

$$f_Y(y,\pi) = exp\{ylog(\frac{\pi}{1-\pi}) + mlog(1-\pi) + log\begin{pmatrix} m \\ y \end{pmatrix}\} \tag{1.24}$$

In this form $E(Y) = m\pi$ and $Var(Y) = m\pi(1-\pi)$ indicating that E(Y) depends on m.

Consider a data transformation from $y \mapsto \frac{y}{m}$ so that $my \sim Binomial(m, \pi)$, y=0, $\frac{1}{m}$...,1 such that

$$\begin{aligned} f_Y(y,\pi) &= \begin{pmatrix} m \\ my \end{pmatrix} \pi^{my}(1-\pi)^{m-my} \\ &= exp\{mylog(\frac{\pi}{1-\pi}) + mlog(1-\pi) + log \begin{pmatrix} m \\ my \end{pmatrix}\} \\ &= \{\frac{ylog(\frac{\pi}{1-\pi}) - log(1-\pi)^{-1}}{\frac{1}{m}} + log \begin{pmatrix} m \\ my \end{pmatrix}\}_{y \in \{0, \frac{1}{m}, ..., \frac{m}{m}\}} \end{aligned} \tag{1.25}$$

If we set $\theta = log(\frac{\pi}{1-\pi})$; $b(\theta) = log(1-\pi)^{-1}) = log(1+e^{\theta})$ and $\phi = \frac{1}{m}$, then we have:

$$E(Y) = b'(\theta) = \frac{e^{\theta}}{1+e^{\theta}} = \pi = \mu \tag{1.26}$$

and

$$\begin{aligned} Var(Y) &= b''(\theta)a(\phi) \\ &= \frac{e^{\theta}}{m(1+e^{\theta})^2} \\ &= \frac{\pi(1-\pi)}{m} = \frac{\mu(1-\mu)}{m} = \phi V(\mu) \end{aligned} \tag{1.27}$$

Note that when $a(\phi) = \frac{\phi}{m}$, then $Var(y) = \phi\frac{\mu(1-\mu)}{m}$. When $\phi > 1$ we have the case of over dispersion.

If the objective is to explain a sample Y with effects represented by a linear combination of explanatory variables, then the GLM of choice will be the logistic regression whose canonical link is the logit link given as:

$$logit(\pi_i) = log(\frac{\pi_i}{1-\pi_i}) = \sum_{j=1}^{p} X_{ij}\beta_j \tag{1.28}$$

Where β_j are the parameters to be estimated.

Remark 1.6.3. If $Var(Yi) = \phi V(\mu_i)$, then, $Var(Yi)$ exceeds the variance under the fitted model hence over-dispersion arises. This however does not arise for exponential dispersion family members such as the normal and inverse Gaussian distributions where this parameter is simply the variance(σ^2). Ignoring over-dispersion results to the underestimation of standard errors of $\hat{\beta}$ leading to incorrect inferences.

According to Molenberghs and Verbeke [53], inclusion of beta random−effects can be used to account for over-dispersion in clustered binary and binomial data. This results to the beta−binomial model in which the Bernoulli model is mixed with a beta distribution. In this case, the prior of the conjugate of the beta distribution of a model coefficient considered a random variable is mixed with the binomial likelihood to form a beta-binomial posterior distribution. This is necessitated by the fact the two have a conjugate relationship i.e. their coefficients are similar in structure and that their kernels are equal i.e. $\pi_i^{y_i}(1-\pi_i)^{n_i-y_i} \sim \pi_i^{a-1}(1-\pi_i)^{b-1}$

Definition 1.6.4. If we let $Y \mid \pi_i \sim Bin(m, \pi_i)$, where $\pi_i \sim Beta(a, b)$, then

$$f(\pi_i|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\pi_i^{a-1}(1-\pi_i)^{b-1}, \quad 0 \leq \pi_i \leq 1, a > 0, b > 0 \tag{1.29}$$

where;

$$E(Y) = \frac{a}{a+b} \quad and \quad Var(Y) = \frac{ab}{(a+b)^2(a+b+1)} \tag{1.30}$$

Multiplying equation (1.29) with equation (1.23) we get the density function of the Beta-Binomial distribution i.e.

$$\begin{aligned} f(Y|\pi_i, a, b) &= f(Y|\pi_i, m) \times f(\pi_i|a, b) \\ &= \frac{\Gamma(a+b)\Gamma(m+1)}{\Gamma(a)\Gamma(b)\Gamma(y_i+1)\Gamma(m-y_i+1)}\pi_i^{y_i+a-1}(1-\pi_i)^{m-y_i+b-1} \\ & \quad 0 \leq \pi_i \leq 1, \quad y_i = 0, 1.., m \end{aligned} \tag{1.31}$$

The Beta-Binomial mean and variance are;

$$E(Y) = \frac{ma}{a+b} \tag{1.32a}$$

and

$$Var(Y) = \frac{mab(a+b+m)}{(a+b)^2(a+b+1)} \tag{1.32b}$$

1.6.2 Estimation

Statistical inference in GLMs is based on maximum likelihood principle

Definition 1.6.5. Let $y_1, ..., y_n$ be independent responses for n subjects. The likelihood is given by:

$$L(\theta, y) = \prod_{i=1}^{n} f_{Y_i}(y_i; \theta) \equiv L_n(\theta) \tag{1.33}$$

The log-likelihood is:

$$\ell(\theta) = logL_n(\theta) = \sum_{i=1}^{n} logf_{Y_i}(y_i; \theta) \tag{1.34}$$

The maximum likelihood estimator of $\hat{\theta}_n$,is defined by

$$\hat{\theta}_n = Sup(\ell(\theta)) \tag{1.35}$$

In this case $\hat{\theta}$ is computed for $\frac{d\ell_n(\theta)}{d\theta} = 0$

Definition 1.6.6. Suppose $Y_1, ..., Y_n$ are independent with $Y_i \sim B(m_i, \pi_i)$ and X_i is a single covariate such that;

$$logit(\pi_i) = log\{\frac{\pi_i}{1-\pi_i}\} = \beta_0 + \beta_1 x_i \tag{1.36}$$

then,

$$\pi_i = \frac{e^{\beta_0+\beta_1 X_i}}{1 + e^{\beta_0+\beta_1 x_i}} \tag{1.37}$$

where $(\beta_0, \beta_1)$are the parameters of the model to be estimated. The Likelihood function will then be given as:

$$\begin{aligned} f_Y(y, \pi) &= \prod_{i=1}^{n} \begin{pmatrix} m_i \\ y_i \end{pmatrix} \{\frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}\}^{y_i} \{\frac{1}{1 + e^{\beta_0+\beta_1 x_i}}\}^{(m_i - y_i)} \\ &= \prod_{i=1}^{n} \begin{pmatrix} m_i \\ y_i \end{pmatrix} e^{(\beta_0+\beta_1 x_i)y_i} \{\frac{1}{1 + e^{\beta_0+\beta_1 x_i}}\}^{m_i} \end{aligned} \tag{1.38}$$

The log-likelihood function is

$$\ell_n(\beta_0, \beta_1) = log \begin{pmatrix} m_i \\ y_i \end{pmatrix} + \sum_{i=1}^{n} (y_i(\beta_0 + \beta_1 x_i) - m_i log(1 + e^{\beta_0+\beta_1 x_i})) \tag{1.39}$$

By partial differentiation we have:

$$\begin{aligned}\frac{\partial \ell_n(\beta_0,\beta_1)}{\partial \beta_0} &= \sum_{i=1}^{n}(y_i - m_i.\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}) \\ &= \sum_{i=1}^{n}(y_i - m_i\pi_i) \\ &= \sum_{i=1}^{n}(y_i - \mu_i) \end{aligned} \tag{1.40a}$$

and

$$\begin{aligned}\frac{\partial \ell_n(\beta_0,\beta_1)}{\partial \beta_1} &= \sum_{i=1}^{n}(y_i x_i - m_i x_i.\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}) \\ &= \sum_{i=1}^{n} x_i(y_i - m_i\pi_i) \\ &= \sum_{i=1}^{n} x_i(y_i - \mu_i) \end{aligned} \tag{1.40b}$$

1.6.3 Quasi-Likelihood Estimation of Parameters in GLMs

Definition 1.6.7. Suppose Y_i(i=1.....n) are independent observations such that $E(Y_i) = b^{'}(\theta) = \mu_i$, where μ_i is some known function of the set of parameters $\beta_1.....\beta_p$ and $Var(Y_i) = b^{''}(\theta) = \phi\nu(\mu_i)$ where $\nu(\cdot)$ is some known variance function. Furthermore, suppose that $g(\mu_{ij}) = \eta_i = X_{ij}\beta$ such that $\frac{\partial \eta_i}{\partial \beta_j} = x_j$ and Var(y_i)= $\phi\nu(\mu_i)$, then generalizing equation (1.8) for the i clusters we have

$$\ell(\theta) = \sum_{i=1}^{n}\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\} \tag{1.41}$$

By the chain rule,

$$\begin{aligned}\frac{\partial(\ell:\theta_i,\phi))}{\partial \beta_j} &= \frac{\partial \ell}{\partial \theta_i}\cdot\frac{\partial \theta_i}{\partial \mu_i}\cdot\frac{\partial \mu_i}{\partial \eta_i}\cdot\frac{\partial \eta_i}{\partial \beta_j} \\ &= \sum_{i=1}^{n}\frac{y_i - b^{'}\theta_i}{a(\phi)}\cdot(\frac{\partial \mu_i}{\partial \theta_i})^{-1}\cdot\frac{\partial \mu_i}{\partial \eta_i}\cdot x_{ij} \\ &= \sum_{i=1}^{n}\frac{y_i - \mu_i}{a(\phi)}\cdot\frac{1}{\nu(\mu_i)}\cdot(\frac{\partial \mu_i}{\partial \beta_j})^{T} \\ &= \sum_{i=1}^{n}\frac{y_i - \mu_i}{a(\phi)\nu(\mu_i)}\cdot(\frac{\partial \mu_i}{\partial \beta_j})^{T} \end{aligned} \tag{1.42}$$

Relaxing the need for a 'distribution-based' construction of the estimating equation and

letting $Var(Yi) = \sigma^2\nu(\mu_i)$ for an arbitrary $\nu(.)$ returning a positive quantity such that:

$$\upsilon_i = \frac{y_i - \mu_i}{\sigma^2\nu(\mu_i)}, \quad E(\upsilon_i) = 0, \quad Var(\upsilon_i) = (\sigma^2\nu(\mu_i))^{-1} \quad and \quad -E(\frac{\partial \upsilon_i}{\partial \mu_i}) = \frac{1}{\sigma^2\nu(\mu_i)} \tag{1.43}$$

then υ_i has the same properties as a score random variable (Nelder and Wedderburn [55]). Since the quasi-log-likelihood is analogous to the log-likelihood such that the integrated quasi-likelihood function perform much like the likelihood function under mild conditions (McCullagh and Nelder [52]), we can replace the log-likelihood in equation (1.42) by the log-quasi-likelihood such that;

$$\frac{\partial(Q(y_i;\mu_i))}{\partial\beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\sigma^2\nu(\mu_i)} \cdot (\frac{\partial\mu_i}{\partial\beta_j})^T \tag{1.44}$$

Where $Q(y_i, \mu_i)$ is the quasi-likelihood function as defined by Wedderburn [80] obtained from equation (1.44) and takes the form

$$Q(y_i;\mu_i) = \int_{y_i}^{\mu_i} \frac{y_i - \mu_i}{\sigma^2\nu(\mu_i))}\partial\mu_i + f(y_i) \tag{1.45}$$

The quasi-likelihood estimating equations for β obtained by differentiating $Q(\mu_i; y_i)$ may be written in the form

$$\sum_{i=1}^{n} \upsilon_i(\beta) = 0 \tag{1.46}$$

where

$$\begin{aligned} \upsilon(\beta) &= \sum_{i=1}^{n} (\frac{\partial\mu_i}{\partial\beta_j})^T \cdot \frac{y_i - \mu_i}{\sigma^2\nu(\mu_i)} \\ &= \sum_{i=1}^{n} D_i^T \nu(\mu_i)^{-1}(y_i - \mu_i)/\sigma^2 \end{aligned} \tag{1.47}$$

which is the quasi-score function, $D_i = \frac{\partial\mu_i}{\partial\beta_j}$ j=i,...p and $\nu(\mu_i)$ is referred to as the 'working' variance of y_i such that $\nu(\mu_i) = diag[\nu(\mu_{i1})....\nu(\mu_{im_i})]$. The solution to equation (1.44) seeks to find the minimum with respect to β of the objective function $\sum_{i=1}^{n} \Psi(y_i, \mu_i) = 0$ where $\Psi(y_i, \mu_i) = \frac{\partial(Q(y_i;\mu_i))}{\partial\beta_j}$ hence can be viewed as an M-estimator(Sakate and Kashi [66]) with a score function:

$$\hat{\psi}(y_i, \mu_i) = \frac{y_i - \mu_i}{\nu(\mu_i)} \frac{\partial\mu_i}{\partial\beta_j^T} \tag{1.48}$$

This quasi-likelihood method is used to fit GLMs and generalized estimating equations for estimating regression coefficients. Some important results of $\upsilon(\beta)$ as established by Liang and Zeger [49] are;

(i) $\sqrt{n}\upsilon(\beta)$ is asymptotically distributed as multivariate normal with mean=0 and variance Λ where $\Lambda = \frac{1}{n}\sum_{i=1}^{n} D_i^T \nu_i^{-1}(\mu_i) Var(y_i) \nu_i^{-1}(\mu_i) D_i$

(ii) $\sqrt{n}(\hat{\beta} - \beta)$is asymptotically distributed as multivariate normal with mean 0 and variance-covariance matrix $\Sigma^{-1}\Lambda\Sigma^{-1}$ where $\Sigma^{-1} = \sum_{i=1}^{n} D^T \nu_i^{-1}(\mu_i) D_i$

1.6.4 Quasi-Likelihood Based estimation for Correlated Binary Responses

Notation

Consider independent observations from n subjects and for each subject i (i=1,2,...,n), m observations are made. Let Y_{it} denote the t^{th} observation from the i^{th} subject ($t = 1, ..., m$) and X_{it}=$\{X_{it1}, X_{it2}.....X_{itp}\}^T$ denote a $p \times 1$ vector of covariates associated with Y_{it}. Let Y_i =$[y_{i1}, ...y_{im}]^T$ denote the response vector for the i^{th} subject and $X_i = [X_{i1}^T, ...X_{im}^T]^T$ be the $m_i \times p$ corresponding covariates matrix.

Assumptions

Carey and Wang [9] identified four key assumptions that govern the use of GEE to model correlated data:

(i) $\mu_{it} = E(Y_{it} \mid X_{it}) = E(Y_{it} \mid X_i)$ i.e. the conditional mean (μ_{it}) of Y_{it} given the predictor variables X_i measured at all possible time points t is equal to a set of the same point specific explanatory variables X_{it}

(ii) Y_{it} have a mean and variance characterized by a GLM (1.5)

(iii) A true conditional $m_i \times m_i$ covariance matrix exists

(iv) Any missing data is Missing Completely at Random (MCAR) i.e. it does not depend on the values of either the observed or missing data.

GEE Modeling

According to Carey and Wang [9], GEE modeling requires the following specifications:

(i) $E(Y_{it} \mid X_{it}) = \mu_{it}$ relate to X_{it} through a known link function i.e. $g(\mu_{it}) = \eta_{it} = X_{it}^T\beta$, where $\beta = [\beta_1 ... \beta_p]^T$ is a $p \times 1$ vector of regression parameters and X_{it} is the i^{th} row of X_i.

(ii) The conditional variance of Y_{it} given X_{it}, is assumed to depend on the mean response given the effect of covariates i.e. $var(Y_{it} \mid X_{it}) = \phi V(\mu_{it})$, where V(.) is a known variance function of μ_{it} and ϕ is a scale parameter which may need to be estimated by:

$$\hat{\phi} = \frac{1}{N-p} \sum_{i=1}^{n} \sum_{t=1}^{m_i} \tilde{e}_{it}^2 \tag{1.49}$$

where $N = \sum_{i=1}^{n} m_i$, p is the covariate dimensionality and

$$\tilde{e}_{it} = \frac{y_{it} - \mu_{it}}{\sqrt{Var(\mu_{it})}} : E(\tilde{e}_{it}) = 0, E(\tilde{e}_{it}^2) = \phi, E(\tilde{e}_{it}, \tilde{e}_{ik}) = \phi corr(y_{it}, y_{ik}). \tag{1.50}$$

Mostly $V(\cdot)$ and ϕ depend on the distribution of outcomes. For instance if Y_{it} is continuous, $V(\mu_{it})$ is specified as 1 and ϕ represents the error variance. If Y_{it} is count, $V(\mu_{it}) = \mu_{it}$ and ϕ is equal to 1.

(iii) An $m \times m$ working correlation matrix $R(\rho)$ is assumed for each Y_{it} and is assumed to be a fully specified $h \times 1$ vector of unknown parameters, $\rho = [\rho_1 ... \rho_h]^T$ which is a vector of nuisance parameters. The corresponding working covariance matrix for Y_{it} is given as:

$$V_i = \phi A_i^{\frac{1}{2}} R_i(\rho) A_i^{\frac{1}{2}} \tag{1.51}$$

where A_i is an $m \times m$ diagonal matrix with $V(\mu_{it})$ as the t^{th} diagonal element i.e. $A_i = Diag\{V(\mu_{i1}).....V(\mu_{im})\}$ and $R_i(\rho)$ is the working correlation matrix structure which describes the within-subject correlation which is of size $m_i \times m_i$ and depends on a vector of association parameter denoted by ρ (Carey and Wang [9]).

The working correlation structures considered in this study were:

1. Independence Correlation structure $(R(\rho)_I)$ which assumes that there is no correlation within the clusters (Jang [42]). It is used when the multiple measurements on the same sampling unit (e.g., person) are assumed to be uncorrelated to each other and modeling assuming such structure is equivalent to a standard normal regression.In this structure

$$Corr(y_{ij}, y_{ik}) = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases}$$

Which implies

$$R(\rho)_I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

2. Exchangeable working correlation Structure (R_{EX}) that assumes equal correlation (ρ) between any pair of measurements on the same individual (Jang [42]). It is often assumed in experiments using a split-plot design, where a within-plot factor is randomly allocated to sub-plots within main plots. Also, (R_{EX}) is a choice in small samples, since it is very parsimonious.

$$Corr(y_{ij}, y_{ik}) = \begin{cases} 1 & \text{if } j = k \\ \rho & \text{if } j \neq k \end{cases}$$

Which implies

$$R(\rho)_{EX} = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

3. Toeplitz correlation structure in which the equal correlations are assumed pairs of responses equally spaced in time. It has (m-1) parameters one for each off-diagonal and is suitable when observations are approximately equispaced.

$$Corr(y_{ij}, y_{i,j+t}) = \begin{cases} 1 & \text{if } j = k \\ \rho_t & \text{for } j = 1, 2...m - t \end{cases}$$

For example if t=4, the toeplitz matrix will be;

$$R(\rho)_{toep} = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{pmatrix}$$

4. Unstructured working correlation structure which assumes different correlations between any two responses for every subject. No constraints are placed on the correlations. Every element of the correlation matrix is estimated separately.

$$Corr(y_{ij}, y_{ik}) = \begin{cases} 1 & \text{if } j = k \\ \rho_{jk} & \text{if } j \neq k \end{cases}$$

Which implies

$$R(\rho)_{UN} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots & \\ \rho_{k1} & \rho_{k2} & \rho_{k3} & \cdots & 1 \end{pmatrix}$$

For $R(\rho)_{UN}$, the number of parameters [0.5m(m-1)] grows rapidly with the number of measurements per subject (Jang [42]). However, it is the most flexible structure.

5. Order one Auto-Regressive working correlation structure in which the size of the correlations quickly decrease as the time lag between pairs of repeated measurements increase i.e.

$$Corr(y_{ij}, y_{ik}) = \rho^k$$

Which implies

$$R(\rho)_{AR-1} = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^k \\ \rho & 1 & \rho & \cdots & \rho^{k-1} \\ \vdots & \vdots & \ddots & \vdots & \\ \rho^k & \rho^{k-1} & \rho^{k-2} & \cdots & 1 \end{pmatrix}$$

$R(\rho)_{AR-1}$ is a parsimonious structure with one parameter.

Remark 1.6.8. Correlation parameters of the toeplitz, exchangeable and AR-1 structures can be estimated using elements of $R(\rho)_{UN}$ (Jang [42]).

(i) For the Toeplitz structure, $\tilde{\rho}_t = \tilde{r}_{(Toep)i,j+t}(t = 1....., m-1)$ is obtained by averaging the t^{th} diagonal components of the estimated unstructured matrix i.e.

$$\tilde{\rho}_{(Toep)t} = \frac{1}{n(m-t)\tilde{\phi}} \sum_{i=1}^{n} \sum_{j=1}^{m-t} \tilde{e}_{ij}\tilde{e}_{i(j+t)} \quad \text{where} \quad 1 \leq t \leq (m-1) \tag{1.57}$$

(ii) For the exchangeable structure, $\tilde{\rho} = \tilde{r}_{EX}$ is obtained by averaging $\tilde{r}_{UN}$ over all j and k i.e.

$$\tilde{\rho}_{EX} = \frac{1}{0.5nm(m-1)\tilde{\phi}} \sum_{i=1}^{n} \sum_{j<k} \tilde{e}_{ij}\tilde{e}_{jk}, \forall j \neq k. \tag{1.58}$$

This implies that $\tilde{\rho}$ of the exchangeable can be regarded as the average of the upper or lower off diagonal components of the estimated unstructured matrix.

(iii) For the AR-1, $\tilde{\rho} = \tilde{r}_{AR-1}$ is obtained by averaging the first off-diagonal elements of the $\tilde{r}_{UN}$ over all j and k i.e.

$$\tilde{\rho}_{AR-1} = \frac{1}{n(m-1)\tilde{\phi}} \sum_{i=1}^{n} \sum_{j=1}^{m-1} \tilde{e}_{ij}\tilde{e}_{i,(j+1)} \tag{1.59}$$

Example 1.6.9. Consider a situation where m=5 and;

$$R(\rho)_{UN} = \begin{pmatrix} 1 & 0.56 & 0.62 & 0.47 & 0.19 \\ 0.56 & 1 & 0.64 & 0.45 & 0.33 \\ 0.62 & 0.64 & 1 & 0.87 & 0.41 \\ 0.47 & 0.45 & 0.87 & 1 & 0.76 \\ 0.19 & 0.33 & 0.41 & 0.76 & 1 \end{pmatrix}$$

Then,

(i)

$$R(\rho)_{toep} = \begin{pmatrix} 1 & 0.71 & 0.49 & 0.40 & 0.19 \\ 0.71 & 1 & 0.71 & 0.49 & 0.40 \\ 0.49 & 0.71 & 1 & 0.71 & 0.49 \\ 0.40 & 0.49 & 0.71 & 1 & 0.71 \\ 0.19 & 0.40 & 0.49 & 0.71 & 1 \end{pmatrix}$$

(ii)

$$R(\rho)_{EX} = \begin{pmatrix} 1 & 0.53 & 0.53 & 0.53 & 0.53 \\ 0.53 & 1 & 0.53 & 0.53 & 0.53 \\ 0.53 & 0.53 & 1 & 0.53 & 0.53 \\ 0.53 & 0.53 & 0.53 & 1 & 0.53 \\ 0.53 & 0.53 & 0.53 & 0.53 & 1 \end{pmatrix}$$

(iii)

$$R(\rho)_{AR-1} = \begin{pmatrix} 1 & 0.71 & 0.50 & 0.36 & 0.25 \\ 0.71 & 1 & 0.71 & 0.50 & 0.36 \\ 0.50 & 0.71 & 1 & 0.71 & 0.50 \\ 0.36 & 0.50 & 0.71 & 1 & 0.71 \\ 0.25 & 0.36 & 0.50 & 0.71 & 1 \end{pmatrix}$$

Remark 1.6.10. The working correlation structures independence, exchangeable and AR-1 can likewise be embedded into the toeplitz Structure (Chen and Nicole [12]). For example if m=4, the toeplitz structure will be defined by 3 parameters say (ρ_1, ρ_2, ρ_3) such that the correlation matrix will be given as:

$$R(\rho)_{Toep} = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{pmatrix} \tag{1.64}$$

If $\rho_1 = \rho_2 = \rho_3 = 0$, R_{Toep} reduces to the independence working correlation structure whose matrix is given as:

$$R(\rho)_{IN} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{1.65}$$

If $\rho_1 = \rho_2 = \rho_3 = \rho \neq 0$, R_{Toep} reduces to the exchangeable working correlation structure whose matrix is given as:

$$R(\rho)_{EX} = \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix} \tag{1.66}$$

If $\rho_1 = \rho \neq 0$, $\rho_2 = \rho_1^2$ and $\rho_3 = \rho_1^3$, R_{Toep} reduces to the AR-1 working correlation structure whose matrix is given as:

$$R(\rho)_{AR-1} = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix} \tag{1.67}$$

Liang and Zeger [49] call equation (1.47) in conjunction with equation (1.51) generalized estimating equations. The quasi-likelihood GEE parameter estimates of β could be obtained by solving the following system utilizing iteratively re-weighted least squares method:

$$U(\hat{\beta}; R_i(\rho), \wp_i) = \sum_{i=1}^{n} D_i^T V_i^{-1}(y_i - \mu_i) = 0 \tag{1.68}$$

where n is the number of subjects and $D_i = \frac{\partial \mu_i}{\partial \beta^T}$ which is the first derivative of the response mean with respect to the regression parameters. It is a Jacobian $m \times p$ matrix given by;

$$\frac{\partial \mu_i}{\partial \beta} = \begin{pmatrix} \frac{\partial \mu_{i1}}{\partial \beta_1} & \frac{\partial \mu_{i1}}{\partial \beta_2} & \cdots & \frac{\partial \mu_{i1}}{\partial \beta_p} \\ \frac{\partial \mu_{i2}}{\partial \beta_1} & \frac{\partial \mu_{i2}}{\partial \beta_2} & \cdots & \frac{\partial \mu_{i2}}{\partial \beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mu_{i_m}}{\partial \beta_1} & \frac{\partial \mu_{i_m}}{\partial \beta_2} & \cdots & \frac{\partial \mu_{i_m}}{\partial \beta_p} \end{pmatrix}$$

$\wp_i \equiv (Y_i, X_i)$, i=1,2....n indicates the data at hand. Since the GEE depend on both β and correlation parameters ρ and have no closed-form solution, iterative two-stage estimation procedure of β and the nuisance parameters (ρ and ϕ) is required. $(y_i - \mu_i)$ is a residual vector which measures deviations of observed responses of the i^{th} subject from its mean. Solving equation (1.68) yields the quasi-likelihood-based estimator of $\hat{\beta}$.

The GEE estimation of $\hat{\beta}$ in equation (1.68), is accomplished either through the generalized weighted least squares method or the two-stage iterative process for β and the nuisance parameters. The iterative procedure for β involves solving the score equation (1.68) until the estimates obtained converge. This involve the use

of the Fisher Scoring Algorithm (Nelder and Wedderburn [55]) which involves the following steps:

(i) Compute initial estimates of for β, say $\hat{\beta}^{(0)}$, using univariate GLM i.e. assuming independence or rather using conventional logistics regression.

(ii) Given $\hat{\beta}^{(0)}$, compute method of moments estimates for ρ (if unknown). With the obtained estimates for ρ, compute $R_i(\rho)$ and consequently the estimate of covariance V_i using equation (1.51).

(iii) After t iterations we have say $\hat{\beta}^{(t)}$. Update the estimator for $\hat{\beta}$ by solving the estimating equation using the Fisher information to obtain improved estimates:

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + (\sum_{i=1}^{n} D_i^T V_i^{-1} D_i)^{-1} \times \sum_{i=1}^{n} D^T V_i^{-1}(y_i - \mu_i) \tag{1.70}$$

(iv) Evaluate for convergence using changes in $\|\hat{\beta}^{(t+1)} - \hat{\beta}^{(t)}\|$. We iterate the above procedure until convergence criterion is satisfied. Convergence occurs when there is no much improvement in the quasi likelihood estimate, or if the set threshold for the change in quasi likelihood estimate is reached.

Asymptotic Properties of $\hat{\beta}$

The following theorem provides the large sample properties for the GEE estimate $(\hat{\beta}_G)$.

Theorem 1.6.11 (Liang and Zeger [49])**.** *Under mild regularity conditions in Appendix A.1 and given that:*

(i) $\hat{\rho}$ *is* $\sqrt{n}$*-consistent given* β *and* ϕ

(ii) $\hat{\phi}$ *is* $\sqrt{n}$*-consistent given* β

(iii) $|\ \frac{d\hat{\alpha}(\beta,\phi)}{d\phi}\ | \leq T(Y,\beta)$ *which is* $O_p(1)$

(iv) the mean structure is correctly specified

Then;

$$\sqrt{n}(\hat{\beta}_G - \beta) \to N(0, V_{LZ}) \tag{1.71}$$

i.e. $\hat{\beta}_G$ *is* $\sqrt{n}$*-consistent for* $\beta : \hat{\beta}_G \to \beta$ *as* $n \to \infty$. V_{LZ} *is a covariance matrix based on the sandwich estimator given by:*

$$V_{LZ} = B^{-1}\hat{M}_{LZ}B^{-1} \tag{1.72}$$

Where;

$$B = \frac{1}{n}\sum_{i=1}^{n} D_i^T V_i^{-1} D_i \tag{1.73a}$$

$$\hat{M}_{LZ} = \frac{1}{n}\sum_{i=1}^{n} D_i^T V_i^{-1} Cov(Y_i) V_i^{-1} D_i \tag{1.73b}$$

Proof. See **Appendix A.2** □

Remark 1.6.12. As the sample size increases $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)^T \to Cov(Y_i)$. If V_i is correctly specified, $V_i = Cov(Y_i)$, hence from equations (1.72), (1.73a) and (1.73b) it follows that;

$$(\hat{M}_{LZ} - B) \xrightarrow{P} 0, \quad (\hat{B} - B) \xrightarrow{P} 0 \quad and \quad \hat{M}_{LZ}B^{-1} \to I_p$$

where I_p is a $p \times p$ identity matrix such that;

$$B^{-1}M_{LZ}B^{-1} = B^{-1} = V_{LZ} \tag{1.74a}$$

i.e. V_{LZ} reduces to $(\frac{1}{n}\sum_{i=1}^{n} D_i^T V_i^{-1} D_i)^{-1}$ which is referred to as model-based variance estimator (Kaurmann and Carroll [45]). This implies that the Cramer-Rao lower bound is attained if $R(\rho)$ in V_i (equation(1.51)) is correct. If this is true, then $\hat{\beta}_G$ will be efficient

Remark 1.6.13. Based on Liang and Zeger [49], if the mean structure, variance function $V(\mu_{it})$ and link function are correctly specified, $\hat{M}_{LZ}$ in equation (1.72) overrides a poor choice of the working correlation structure and still yields consistent estimates of the regression parameter with large n. However, a misspecified

structure yields inconsistent $\hat{\rho}$ estimates. This violates Theorem 1.6.11 hence the consistency property of $\hat{\beta}$ along with their impact on the variance estimates will no longer be assured.

1.6.5 Model Selection Concepts

In this section we establish the technical notions that facilitate the development of estimators for Kullback's I-divergence and also introduce the model selection criteria developed based on the Kullback I-divergence.

Definition 1.6.14 (**Discrepancy / Divergence**)**.** If M_T and M_C denote the true model and candidate model respectively, a discrepancy measures the lack-of-fit when data is fitted by M_C hence a model selection criteria is a statistics that estimates the discrepancy between M_T and M_C which requires a well defined distance function d(a, b) where a and b may be vectors or scalars. Essential properties that must be satisfied by d(a,b) as espoused by Linhart and Zucchini [50] are:

(a) positiveness $d(a, b) > 0$, $\forall\ a \neq b$ and d(a, b) = 0, $\forall$ a=b

(b) Symmetry d(a, b) = d(b, a)

(c) triangle inequality $d(a, c) \leq d(a, b) + d(b, c)$ $\forall$ a, b and c

Remark 1.6.15. If we let θ_0 and θ_k denote the vectors of parameters from M_T and M_C respectively and let $d(\theta_0, \theta_k)$ be the discrepancy between M_T and M_C; then

$$d(\theta_0, \theta_k) = E_{f_0}\{\delta(y, \theta_k)\} = E_{f_0}\{-2\ell(y, \theta_k)\} \tag{1.75}$$

where $\delta(y, \theta_k)$ represents a function that measures the accuracy when y is predicted by M_C, E_{f_0} is the expectation under the true model, and $\ell(y, \theta_k)$ is the log likelihood of the candidate model.

Definition 1.6.16. Kullback's I-Divergence

If we let f_0 and f denote the densities of M_T and M_C respectively and that $d(\theta_0, \theta_k) = 0 \ \forall \theta_0 = \theta_k$, then the Kullback's I-divergence between $f_0(y|\theta_0)$ and $f(y|\theta_k)$ defined with respect to $f_0(y|\theta_0)$ and denoted by $I_{f_0 f}$ is given as:

$$\begin{aligned} I_{f_0 f} &= \int f_0 log \frac{f_0(y|\theta_0)}{f(y|\theta_k)} df_0 \\ &= \int f_0 log f_0(y|\theta_0) df_0 - \int f_0 log f(y|\theta_k) df_0 \\ &= E_{f_0}\{log f_0(y|\theta_0)\} - E_{f_0}\{log f(y|\theta_k)\} \\ &= E_{f_0}\{log \frac{f_0(y|\theta_0)}{f(y|\theta_k)}\} \end{aligned} \tag{1.76}$$

where E_{f_0} is the expectation under $f_0(y|\theta_0)$. $I_{f_0 f}$ represents the information lost when model 'f' is used to approximate f_0. Clearly, the best model loses the least information relative to other models in the set (Linhart and Zucchini [50]). Key features of $I_{f_0 f}$ are:

(i) $I_{f_0 f}$ is not symmetric i.e. $I_{f_0 f}(\theta_0, \theta_k) \neq I_{f f_0}(\theta_k, \theta_0) \ \forall \theta_0 \neq \theta_k$,

(ii) $I_{f_0 f}(\theta_0, \theta_k) > 0 \ \forall \theta_0 \neq \theta_k$ and $I_{f_0 f}(\theta_0, \theta_k) = 0$ iff $\theta_0 = \theta_k$ (Kullback [47]). To see why we recall that since log is a concave function, -log is convex, thus;

$$\begin{aligned} I_{f_0 f} &= E_{f_0}[log\{\frac{f_0(y|\theta_0)}{f(y|\theta_k)}\}] = E_{f_0}[-log\{\frac{f(y|\theta_k)}{f_0(y|\theta_0)}\}] \\ &\geq -log\{E_{f_0}[\frac{f(y|\theta_k)}{f_0(y|\theta_0)}]\} \\ &= -log\{\int f_0(y|\theta_0) \frac{f(y|\theta_k)}{f_0(y|\theta_0)} dy \\ &= -log\{\int f(y|\theta_k) dy \\ &= -log(1) = 0 \qquad \text{(by Jensen inequality)} \end{aligned} \tag{1.77}$$

This is strict only for a non-degenerate random variable and a strictly convex function and since -log is strictly convex, the inequality can only be made strict if $\frac{f(y|\theta_k)}{f_0(y|\theta_0)}$ is made degenerate. This occurs when $\theta_0 = \theta_k$

(iii) Minimizing $I_{f_0 f}$ is equivalent to maximizing $E_{f_0}[log f(y|\theta_k)]$. This is because the first term in $I_{f_0 f}$ is a constant. It in no way depends on $f(y|\theta_k)$. $I_{f_0 f}$ can only be used in model selection if $E_{f_0}[log f(y|\theta_k)]$ can be estimated.

Example 1.6.17. To illustrate the use of $I_{f_0 f}$, we consider a random variable X whose sample space is $S = \{X : X \in \mathbb{Z}, 500 \leq X \leq 510\}$ and its probability distribution is as shown in Figure 1.1.

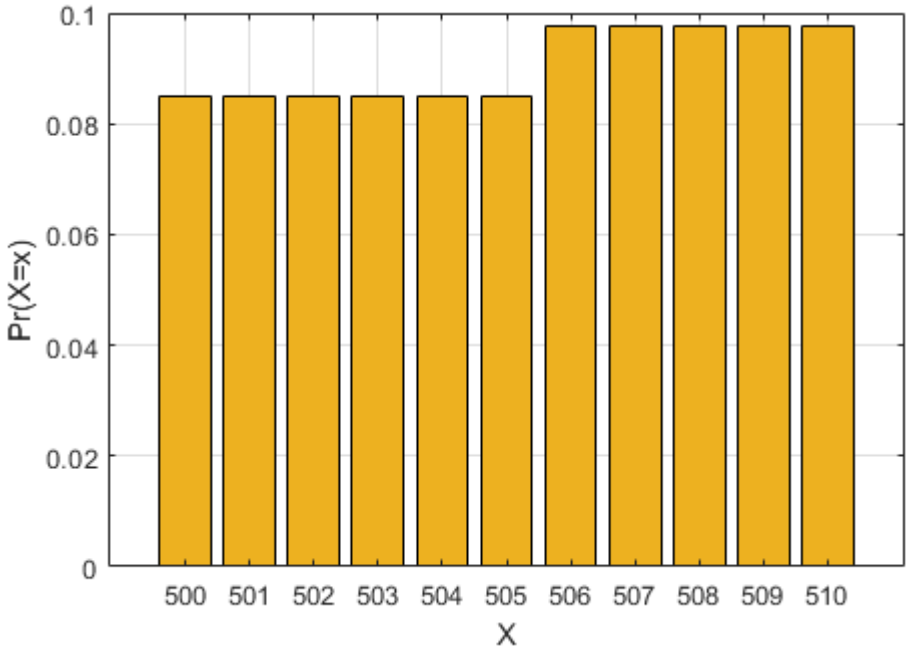

Figure 1.1: Probability Histogram Plot of the Random Variable X

We might want to reduce this data to a simple model with just one or two parameters. We would wish to determine which distribution between the uniform distribution and binomial distribution best represents the distribution of X i.e. the distribution that preserves the most information from our original data source. By calculating $I_{f_0 f}$ which is the expectation of the log difference between the probability of data in the original distribution and the approximating distributions we get the following results:

$I_{f_0 f}(Observed \parallel Uniform) = 0.015$

$I_{f_0 f}(Observed \parallel Binomial) = 0.258$

The results indicate that the information lost by using the binomial approximation is greater than that lost using the uniform approximation. This is alo shown in Figure 1.2

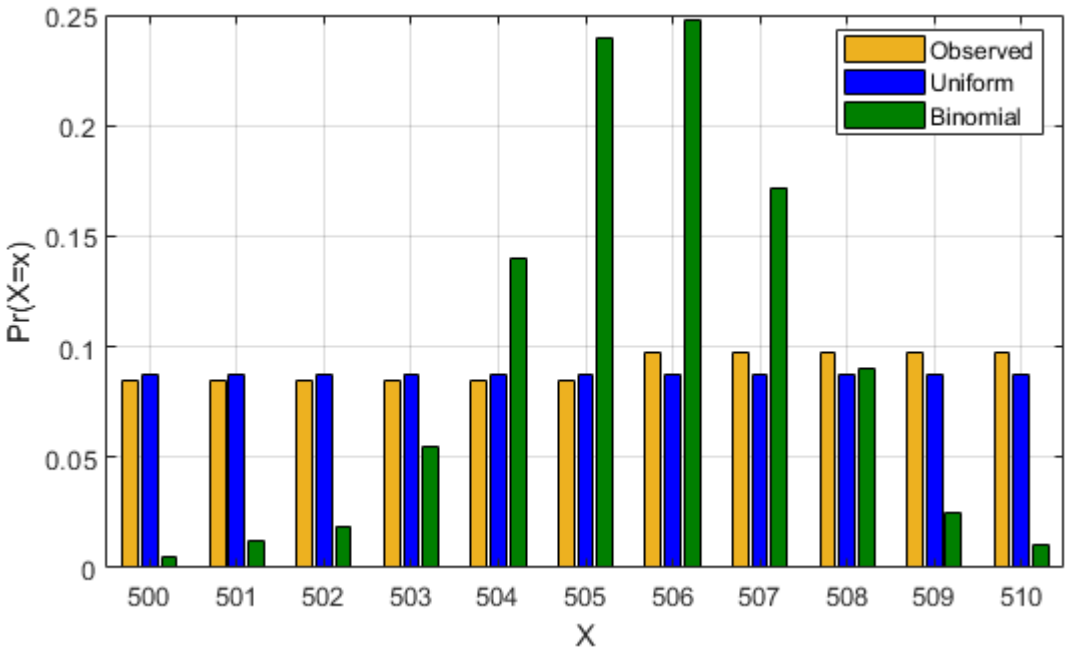

Figure 1.2: Probability Distribution of X (Observed, Uniform and Binomial)

If we have to choose one to represent the distribution of X, then it will be better off to use the uniform approximation

Definition 1.6.18. If we let $d_{f_0 f}(\theta_0, \theta_k)$ denote $d(f_0(y|\theta_0), f(y|\theta_k))$, then, according to Cavanaugh [11];

$$d_{f_0 f}(\theta_0, \theta_k) = E_{f_0}\{-2log f(y|\theta_k)\} \tag{1.78a}$$

and

$$d_{f f_0}(\theta_k, \theta_0) = E_{f}\{-2log f_0(y|\theta_0)\} \tag{1.78b}$$

Evaluating the second argument of equation (1.78a) at $f_0(y|\theta_0)$ yields

$$d_{f_0 f_0}(\theta_0, \theta_0) = E_{f_0}\{-2log f_0(y|\theta_0)\} \tag{1.79}$$

which is the minimum of $d_{f_0 f}(\theta_0, \theta_k)$

Also, evaluating the second argument of equation (1.78b) at $f(y|\theta_k)$ yields

$$d_{f f}(\theta_k, \theta_k) = E_{f}\{-2log f_0(y|\theta_k)\} \tag{1.80}$$

which is the minimum of $d_{f f_0}(\theta_k, \theta_0)$

Combining the equation (1.76) with equations (1.78a) and (1.79) we have:

$$\begin{aligned} I_{f_0 f}(\theta_0, \theta_k) &= E_{f_0} log f_0(y|\theta_0) - E_{f_0} log f(y|\theta_k) \\ 2I_{f_0 f}(\theta_0, \theta_k) &= E_{f_0} 2log f_0(y|\theta_0) - E_{f_0} 2log f(y|\theta_k) \\ &= -d_{f_0 f_0}(\theta_0, \theta_0) - (-d_{f_0 f}(\theta_0, \theta_k)) \\ &= d_{f_0 f}(\theta_0, \theta_k) - d_{f_0 f_0}(\theta_0, \theta_0) \end{aligned} \tag{1.81}$$

Equation (1.81) can be written as:

$$2I_{f_0 f}(\theta_0, \theta_k) = d_{f_0 f}(\theta_0, \theta_k) - E_{f_0}\{-2log f_0(y|\theta_0)\} \tag{1.82}$$

Since $d_{f_0 f_0}(\theta_0, \theta_0)$ is independent of θ_k, using $I_{f_0 f}(\theta_0, \theta_k)$ to rank candidate models would be similar to using $d_{f_0 f_0}(\theta_0, \theta_0)$. To discriminate among various candidate models, $d_{f_0 f}(\theta_0, \theta_k)$ is a valid substitute for $I_{f_0 f}(\theta_0, \theta_k)$ (Cavanaugh [10]). Therefore,

$$d(\theta_0, \theta_k) = E_{f_0}\{-2log f(y|\theta_k)\} \mid \theta_k = \hat{\theta}_k \tag{1.83}$$

would provide a suitable measure of separation between the generating model $f(y|\theta_0)$ and a fitted model $f(y|\hat{\theta}_k)$. The model $f(y|\hat{\theta}_k)$ that is close to $f(y|\theta_0)$ in the sense of having a small $I_{f_0 f}$ value is the best model (Cavanaugh [11]). However, $d(\theta_0, \theta_k)$ is an oracle that can be estimated but cannot be used directly for model selection since they depend on unknown generating model $f_0(y|\theta_0)$.

Likelihood-Based Model Selection Criteria: Akaike Information Criteria (AIC)

Evaluating $d(\theta_0, \hat{\theta}_k)$ is not possible since it requires the knowledge of θ_0. Akaike [3] suggested that $-2log f(y|\hat{\theta}_k)$ serves as a biased estimator of $d(\theta_0, \hat{\theta}_k)$. In order to use the bias, we investigate the bias by writing $d(\theta_0, \hat{\theta}_k)$ as follows:

$$\begin{aligned} d(\theta_0, \hat{\theta}_k) &= E_{f_0}\{E_{f_0}(-2log f(y|\theta_k))\}|_{\theta_k = \hat{\theta}_k} \\ &= E_{f_0}\{-2log f(y|\hat{\theta}_k)\} + [E_{f_0}\{E_{f_0}\{-2log f(y|\theta_k)\}|_{\theta_k = \hat{\theta}_k} \\ &\quad -E_{f_0}\{-2log f(y|\hat{\theta}_k)\}] \end{aligned} \tag{1.84}$$

The bracketed [.] quantity is referred to as the expected optimism and is useful in correcting for the negative bias incurred when $-2logf(y|\hat{\theta}_k)$ is used as an estimator of equation (1.83) (Efron [21]). The sum of $-2logf(y|\hat{\theta}_k)$ and the expected optimism say $\widetilde{O}$ provides an approximately unbiased estimator of the expected I-divergency $[d(\theta_0, \hat{\theta}_k)]$ i.e.

$$d(\theta_0, \hat{\theta}_k) = -2logf(y|\hat{\theta}_k) + \widetilde{O} \tag{1.85}$$

$-2logf(y|\hat{\theta}_k)$ is the goodness of fit term while $\widetilde{O}$ is the penalty term.

According to Cavanaugh [11], if k represents the number of functionally independent parameters in $f(y|\theta_k)$, then $\widetilde{O} \approx 2k$. Thus, under appropriate conditions, the expected value of

$$AIC = -2logf(y|\hat{\theta}_k) + 2k \tag{1.86}$$

should asymptotically approach the expected value of $d(\theta_0, \hat{\theta}_k)$ and serves as an asymptotically unbiased estimator of $E_{f_0}[d(\theta_0, \hat{\theta}_k)]$. AIC provides an approximately unbiased estimator of the expected discrepancy in settings where n is large and k is relatively small and does provide a balance between bias and variability of a candidate model hence aims at selecting a model that has few parameters but fits the data well.

Quasi-Likelihood-Based Model Selection Criterion: QIC

From equation (1.68), the solution of the score equation $U(\beta)$ is maximum if the second derivative of the log-quasi-likelihood yields a matrix that is negative definitive hence from Theorem 1.6.11 we can define the observed fisher information $I(\beta \mid y)$ as:

$$I(\beta \mid y) = -\frac{\partial^2 Q(\beta \mid y)}{\partial\beta\partial\beta^T} \tag{1.87}$$

In this regard $E_{f_0}\{-\frac{\partial^2 Q(\beta|y)}{\partial\beta\partial\beta^T}\}$ is the expected fisher information which we denote by $I(\beta)$. From equation 1.6.11, let

$$\Sigma(\beta) = I(\beta)^{-1} J(\beta) I(\beta)^{-1} \tag{1.88}$$

Where $J(\beta) = E_0\{J(\hat{\beta}|y)\}$ and $J(\beta|y) = \hat{M}_{LZ}$ in equation (1.73b)

We re-derive the Quasi-Likelihood Information Criteria(QIC) proposed by Pan [60] for the selection of a working correlation matrix and set of covariates in GEE. QIC was a modification of AIC by replacing the log-likelihood in equation (1.83) with a log-quasi-likelihood such that the overall discrepancy can be expressed as

$$d_{f_0 f} = E_{f_0}\{-2Q(y|\beta)\} \mid \beta = \hat{\beta}^I \tag{1.89}$$

where $\hat{\beta}^I$ denotes the estimator of β under the working independence model. The expected discrepancies then become

$$E_{f_0}\{d_{f_0 f}(\beta_0, \hat{\beta}_k^I)\} \tag{1.90a}$$

and

$$-2Q(\hat{\beta}_k^I|y) \tag{1.90b}$$

Where $-2Q(\hat{\beta}^I|y)$ is the quasi-likelihood under independence assumption and E_{f_0} is taken under the true model.

Let the Kullback I-divergence $I(\beta_0, \beta_k)$ under the independence correlation structure be indexed by the parameter vector β_k and that β_0 be the corresponding parameter of the quasi-likelihood model introduced by the true generating model. Considering a second-order Taylor expansion of $-2Q(\beta_0|y)$ about $\hat{\beta}_k$ we have:

$$-2Q(\beta_0|y) = -2Q(\hat{\beta}_k|y) + (\hat{\beta}_k - \beta_0)^T I(\hat{\beta}_k|y)(\hat{\beta}_k - \beta_0) + R_1(\beta_0, \hat{\beta}_k) \tag{1.91}$$

As $n \to \infty$, $R_1(\beta_0, \hat{\beta}_k)$ is $o_p(1)$ such that $E[R_1(\beta_0, \hat{\beta}_k)]$is o(1). Hence we have;

$$E_{f_0}\{-2Q(\beta_0|y)\} - E_{f_0}\{-2Q(\hat{\beta}|y)\}\} = E_{f_0}[(\hat{\beta}_k - \beta_0)^T I(\hat{\beta}_k|y)(\hat{\beta}_k - \beta_0)] + o(1) \tag{1.92}$$

But $E_{f_0}\{-2Q(\beta_0|y)\} = d_{f_0 f_0}(\beta_0, \beta_0)$ hence equation (1.92) can be expressed as;

$$d_{f_0 f_0}(\beta_0, \beta_0) - \{-2Q(\hat{\beta}|y)\}\} = E_{f_0}[(\hat{\beta}_k - \beta_0)^T I(\hat{\beta}_k|y)(\hat{\beta}_k - \beta_0)] + o(1) \tag{1.93}$$

Also, taking the second-order Taylor expansion of $d_{f_0 f}(\beta_0, \hat{\beta}_k^I)$ about β_0 we have;

$$d_{f_0 f}(\beta_0, \hat{\beta}_k^I) = d_{f_0 f_0}(\beta_0, \beta_0) + (\hat{\beta}_k - \beta_0)^T I(\beta_0)(\hat{\beta}_k - \beta_0) + R_2(\beta_0, \hat{\beta}_k) \tag{1.94}$$

Taking the expectation $E_{f_0}(\cdot)$ on both sides we have;

$$E_{f_0}\{d_{f_0f}(\beta_0,\beta_k)-d_{f_0f_0}(\beta_0,\beta_0)\}=E_{f_0}[(\hat{\beta}_k-\beta_0)^TI(\beta_0)(\hat{\beta}_k-\beta_0)]+o(1) \quad (1.95)$$

Let $\Upsilon(\beta_0,\hat{\beta}_k)$ define the discrepancy between $f_0(\beta_0|y)$ and $f(\hat{\beta}_k|y)$ such that

$$\begin{aligned}\Upsilon_{f_0f}(\beta_0,\hat{\beta}_k) &= E_{f_0}(-2Q(\hat{\beta}_k|y)) && (1.96)\\ &+ d_{f_0f_0}(\beta_0,\beta_0)-E_{f_0}(-2Q(\hat{\beta}_k|y)) && (1.97)\\ &+ E_{f_0}(d_{f_0f}(\beta_0,\hat{\beta}_k))-d_{f_0f_0}(\beta_0,\beta_0) && (1.98)\end{aligned}$$

Substituting equations (1.97) and (1.98) in equation (1.96) by equations (1.93)and (1.95) respectively we have;

$$\begin{aligned}\Upsilon_{f_0f}(\beta_0,\hat{\beta}_k) &= E_{f_0}(-2Q(\hat{\beta}_k|y)) && (1.99)\\ &+ E_{f_0}[(\hat{\beta}_k-\beta_0)^TI(\hat{\beta}_k|y)(\beta_k-\hat{}\,\beta_0)] && (1.100)\\ &+ E_{f_0}[(\hat{\beta}_k-\beta_0)^TI(\beta_0)(\hat{\beta}_k-\beta_0)] && (1.101)\\ &+ o(1) && (1.102)\end{aligned}$$

As $n\to\infty$, $\hat{\beta}_k\to\beta_0$ hence

$$\Upsilon_{f_0f}(\beta_0,\hat{\beta}_k)=E_{f_0}(-2Q(\hat{\beta}_k|y))+2E_{f_0}[(\hat{\beta}_k-\beta_0)^TI(\beta_0)(\hat{\beta}_k-\beta_0)]+o(1) \quad (1.103)$$

Since $E_{f_0}[(\hat{\beta}_k-\beta_0)^TI(\beta_0)(\hat{\beta}_k-\beta_0)$ is a scalar we can write equation (1.103) as;

$$\begin{aligned}\Upsilon_{f_0f}(\beta_0,\hat{\beta}_k) &= E_{f_0}(-2Q(\hat{\beta}_k|y))+2\{tr[I(\beta_0)(\hat{\beta}_k-\beta_0)^T(\hat{\beta}_k-\beta_0)]\}+o(1)\\ &= E_{f_0}(-2Q(\hat{\beta}_k|y))+2[tr\{I(\beta_0)\Sigma(\beta_0)\}]+o(1) && (1.104)\end{aligned}$$

Replacing $\Sigma(\beta)$ by equation (1.88) we have;

$$\Upsilon_{f_0f}(\beta_0,\hat{\beta}_k)=E_{f_0}(-2Q(\hat{\beta}_k|y))+2[tr\{(I(\beta_0)I(\beta_0)^{-1}J(\beta_0)I(\beta_0)^{-1}\}]+o(1) \quad (1.105)$$

By ignoring the o(1) term and letting $I(\hat{\beta}^I|y)=\hat{\Omega}_I$ and $I(\beta_0)^{-1}J(\beta_0)I(\beta_0)^{-1}=\Sigma(\hat{\beta}_k^I)=\hat{V}_r$ we define the statistic as contained in Pan [60]

$$\begin{aligned}QIC^I &= -2Q(\hat{\beta}^I|y)+2tr[\{I(\beta_0)I(\beta_0)^{-1}J(\hat{\beta}^I|y)I(\beta_0)^{-1}\}]\\ &= -2Q(\hat{\beta}^I|y)+2tr\{\hat{\Omega}_I\hat{V}_r)\} && (1.106)\end{aligned}$$

which is an asymptotically unbiased estimator of $\Upsilon_{f_0 f}(\beta_0, \hat{\beta}_k)$ if the mean structure is specified correctly.

If we assume that the GEE estimator $\hat{\beta}^R$ is obtained using a working correlation structure $R(\rho)$, equation (1.106) is modified to get of QIC^R (Pan,[60]):

$$\begin{aligned} QIC^R &= -2Q(\hat{\beta^R}|y; I, \wp) + 2tr[\{I(\beta_0)I(\beta_0)^{-1}J(\hat{\beta}^R|y)I(\beta_0)^{-1}\}] \\ &= -2Q(\hat{\beta}^R|y; I, \wp) + 2tr\{\hat{\Omega}_I \hat{V}_r\} \end{aligned} \tag{1.107}$$

$\hat{\Omega}_I$ is a $p \times p$ model-based covariance matrix for the estimated regression parameters under the independence working correlation structure and is given as:

$$\begin{aligned} \hat{\Omega}_I &= \hat{\sum}_{M(I)}^{-1} \\ &= E_{f_0}\{-\frac{d^2 Q(\beta; I, \wp}{d\beta d\beta^T}\} \mid_{\beta=\beta_0} \\ &= \frac{1}{n}\sum_{i=1}^{n} D_i^T V_i^{-1} D_i \end{aligned} \tag{1.108}$$

Where $D_i = \frac{\partial \mu_i}{\partial \beta^T}$ and μ_i is as defined in equation (1.4).

$\hat{V}_r = \hat{\sum}_{S(R)}$ is a $p \times p$ robust or sandwich variance estimator under the working correlation structure R and is given by:

$$\hat{V}_r = \hat{\Omega}_I^{-1} J(\beta|y) \hat{\Omega}_I^{-1} \tag{1.109a}$$

and

$$J(\beta|y) = \frac{1}{n}\sum_{i=1}^{n} D_i^T V_i^{-1}(Y_i - \mu_i)(Y_i - \mu_i)^T V_i^{-1} D_i \tag{1.109b}$$

$tr(\hat{\Omega}_I \hat{V}_r)$ is the sum of diagonal elements of the product matrix which measures total variability.

Remark 1.6.19. QIC^R can be decomposed into two parts:

(i) $-2Q(\hat{\beta}^R|y; I, \wp)$ is the sum of the log-quasi-likelihood function under independence correlation structure, I, evaluated at estimated regression coefficients

obtained under a 'working ' correlation structure R for the $\sum_{i=1}^{n} m_i$ observations in the data ($\wp$). It relates to the log-quasi-likelihood for independent observations hence contain no information on the anticipated within-subject correlation structure i.e. it is free from both R and R_0 , R_0 being the true within-subject correlation structure. $-2Q(\hat{\beta}^R|y; I, \wp)$ measures the goodness-of-fit of the model.

(ii) $2tr(\hat{\Omega}_I \hat{V}_r)$ contains information of the hypothesized correlation structure via $\hat{V}_r$ and acts as a penalty for over-complexity. Half of this term was used by Hin and Wang [34] to develop the Correlation Information Critieria (CIC) given by

$$CIC = tr(\hat{\Omega}_I \hat{V}_r) \tag{1.110}$$

Remark 1.6.20. If we let $\triangle(k)$ be the expected kullback I-discrepancy that reflects the separation between the generating model $f_0(\theta_0|y)$ denoted in this case as simply $f(\theta_0)$ and a fitted model $f(\hat{\theta}_k|y)$ denoted simply as $f(\hat{\theta}_k)$ such that $f(\theta_0) \in \digamma(k)$ where $\theta_0 \in \Theta(k)$, then

$$\begin{aligned} \triangle(k) &= E_{f_0}(d(\theta_0, \hat{\theta}_k)) \\ &= E_{f_0}\{-2Q(\hat{\theta}_k|y\} \\ &+ E_{f_0}\{-2Q(\theta_0|y)\} - E\{-2Q(\hat{\theta}_k|y)\} \qquad (1.111) \\ &+ E_{f_0}(d(\theta_0, \hat{\theta}_k)) - E_{f_0}\{-2Q(\theta_0|y)\} \qquad (1.112) \end{aligned}$$

In the spirit of Cavanaugh [10], the following lemma justify the asymptotic unbiasedness of QIC by asserting that equations (1.111) and (1.112) are both within o(1) of k.

Lemma 1.6.21.

$$E_{f_0}\{-2Q(\theta_0|y)\} - E_{f_0}\{-2Q(\hat{\theta}_k|y)\} = k + o(1) \tag{1.113a}$$

$$E_{f_0}\{d(\theta_0, \hat{\theta}_k)\} - E_{f_0}\{-2Q(\theta_0|y)\} = k + o(1) \tag{1.113b}$$

Proof. See **Appendix A.3** □

Remark 1.6.22. QIC is tauted as being desirable for working correlation structure and variable selection. If QIC as a model selection criterion chooses the model with minimum mean squared error in large samples under the assumption that the generating or true model is of infinite dimension, or that the set of candidate models does not contain the true model, then it is said to be asymptotically efficient (Shibata [69]).

Remark 1.6.23. Under the assumption that the true model is included in the set of candidate models, if QIC identifies the correct model asymptotically with probability one, then it will be said to be consistent.

Empirical Likelihood and Model Selection in Generalized Estimating Equations

Despite GEE enjoying the advantages of semi-parametric methods, it is limited by its lack of a likelihood since likelihood methods are quite effective in the determination of efficient estimators, construction of tests with good power properties and short confidence intervals and selection of best models from a pool of candidates (Chen and Nicole [12]). Empirical likelihood combines the reliability of non-parametric methods with the flexibility and effectiveness of likelihood approaches hence has the potential of adding value to GEE models.

For a sample $X_1, ..., X_n$ from an unknown d-variate distribution F_0 having mean $\mu_0 \in \Re^d$ (d $\geq$ 1), the empirical likelihood function for a distribution F which is the probability mass placed on X_i by F is given by

$$L(F) = \prod_{i=1}^{n} p_i \tag{1.114}$$

where $p_i = Pr(X = X_i)$ and for distributions that assign positive probability on each of the observed data points the likelihood is non zero. Without any additional constraint on p_i, L(F) is maximized by the empirical distribution function F_n which puts equal weight $\frac{1}{n}$ on each observation. Thus, according to Owen[59] the

empirical likelihood ratio for F is:

$$R(F) = \frac{L(F)}{L(F_n)} = \frac{\prod_{i=1}^{n} p_i}{\prod_{i=1}^{n} \frac{1}{n}} = \prod_{i=1}^{n} np_i \tag{1.115}$$

Suppose now one is interested in estimating a parameter μ_0, the profile empirical likelihood ratio (ELR) for a candidate μ is:

$$\Re(\mu) = Sup\{\prod_{i=1}^{n} np_i \mid p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i X_i = \mu\} \tag{1.116}$$

For the regression parameter β in GEE, the empirical likelihood ratio function is defined by

$$\Re(\beta) = Sup\{\prod_{i=1}^{n} np_i \mid p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i g(X_i^T \beta) = 0\} \tag{1.117}$$

with

$$g(X_i^T \beta) = (\frac{d\mu_i}{d\beta^T})^T V_i^{-1}(\mu)(Y_i - \mu_i) \tag{1.118}$$

where $V_i^{-1}(\mu) = A_i^{-0.5} R^{-1}(\rho)) A_i^{-0.5}$ and the maximization is taken with respect to the probabilities $p_1, ..., p_n$.

The maximum empirical likelihood estimator for β is:

$$\hat{\beta}_E = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \{\Re(\beta)\} \tag{1.119}$$

Whether or not the working correlation structure $R(\rho)$ is correctly specified, the MELE of $\hat{\beta}_E$ from equation(1.119) is consistent and asymptotically normal. However, its efficiency is compromised if $R(\rho)$ is misspecified (Owen [59]).

Definition 1.6.24. Let S be working correlation matrices such that $R_s, s = 1, ..S$, then S different empirical likelihoods $R^s(\beta), s = 1, .., S$ can be defined by equation (1.117). For each R^s, the $MELE(\hat{\beta}_E^s)$ equals the corresponding GEE estimator $\hat{\beta}_G^s$ defined by equation (1.117). By replacing the parametric likelihood in AIC with the empirical likelihood, Chen and Nicole[12] proposed the empirical likelihood versions of AIC given as:

$$EAIC_s = -2log\Re^F(\hat{\theta}_G^s) + 2dim(\theta^s) \tag{1.120}$$

where s is the index for a candidate model parameterized by $\theta^s (s = 1, ..., S)$, and $\hat{\theta}_G^s$ is the GEE estimate associated with the working correlation structure R_s. More specifically $\hat{\theta}_G^s = \begin{pmatrix} \hat{\beta}_G^s \\ \hat{\rho}_G^s \end{pmatrix}$, where $\hat{\rho}_G^s$ is the method of moment estimator of ρ given $\hat{\beta}_G^s$ and R_s.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In this chapter we review literature on model selection tools for choosing the correct set of covariates and a working correlation structures in GEE. Focus is on previously developed model selection criteria using Kullback's I-divergence as an oracle.

2.2 Generalized Estimating Equations and the Analysis of Correlated Data

To account for the within-subject dependence in longitudinal data, Liang and Zeger [83] developed the generalized estimating equation (GEE) for such data. The method only requires the specification of the first two moments and a working correlation matrix involving a small number of nuisance parameters. The GEE method is a Population Average (PA) or marginal method because it produces the average value of the individual regression lines for the regression coefficients. Neuhaus et. al. [56] compared Subject-Specific (SS) and PA approaches for analyzing correlated binary data by comparing the parameters estimated by PA and SS models algebraically and geometrically and showed that the covariate effects measured by the PA approach are closer to null values than those of the SS approach when the SS model holds and that the difference in the magnitude of the covariate effects increases with intra-subject correlation. Hanley et al.[31] observed that GEE produces reasonably accurate standard errors hence confidence intervals have the correct coverage rates compared to other methods such as the random effects models which explicitly model and estimate the between-subject variations

and incorporate them together with residual variance into standard errors. Further, they asserted that the computational complexity of GEE is a function of the number of observations per subject or cluster size rather than the number of subjects or clusters.

One of the importance of GEE is that it yields consistent estimators even if the working correlation structure is misspecified. This means that the estimator can be inefficient under a misspecified correlation structure. Wang and Hin [78], observed that correctly specifying the correct correlation structure can definitely enhance the efficiency of the parameter estimates hence selection of intra-subject correlation matrix plays a vital role in GEE as it leads to improved finite sample performance. Wang and Carey [79] assert that the asymptotic relative efficiency of the parameter estimates of the GEE method is likely to be low when the working correlation structure is misspecified. Their assertions were corroborated by Sutradhar and Das [74] who also pointed out that the mis-specification of the correlation structure lowered the relative efficiency of the estimate even when the sample size is finite. This emphasized the need for correct specification of the working correlation structure.

A working correlation structure $R(\rho)$ is a $m \times m$ correlation matrix for repeated or clustered measurements from each individual $y_i = (y_{i1}, y_{i2}...y_{im})$ fully specified by the parameter ρ. In GEE modeling, one has to specify the working correlation matrix to account for the within-subject correlation of the response variables.

2.3 Selection of Working Correlation Structure in GEE

Early approaches developed for the selection of a working correlation structure included the Rotnitzky and Jewell Criteria (RJ) by Rotnitzky and Jewell [65] to appraise the adequacy of the assumed correlation matrix using the fact that the asymptotic distribution of a modified working Wald statistic is the linear combination of independent Chi-Square random variables. However, Hin et al.[35]

showed that more often the RJ criterion preferred the exchangeable structure with one correlation parameter ρ.

Shults and Chaganty [72], while considering the minimization of the generalized error sum of squares came up with the Shults and Chaganty (SC) criterion which was later extended by Carey and Wang [9] who by adopting the Gaussian Pseudo-likelihood (GP) developed the GP(R) criteria and showed through simulation studies that it had better performance than the RJ criteria.

Akaike Information Criteria (Akaike [3]) which serves as an estimator of Kullback's I-divergence is the most popular criteria for likelihood-based GLM modeling and its development was anchored in the use of the number of parameters as a standard for comparing the candidate models hence the resultant penalty was meant to check for model complexity encouraged by the bias correction term. However, AIC could not directly be applied in the GEE framework since no distribution is assumed in GEE and also because the GEE estimator has different asymptotic properties from those of the maximum likelihood estimation. Hence a modification to the penalty term in AIC was necessary. Pan [60] proposed a modification of AIC, called 'quasi-log-likelihood under the independence model information criterion (QIC) by replacing the log-likelihood in AIC with the log-quasi-likelihood under working independence assumption.

Pan [60] examined the performance of QIC in selecting the true exchangeable correlation structure compared to AIC. He considered a candidate set that comprised the independence, exchangeable, and AR-1 structures. In his simulation design with 1000 replications, the number of measurements per subject were fixed at 3 and the number of subjects were 50 and 100. He established that for n=50, QIC's selection rate of the correct correlations structure was 67.8% while that of AIC was 83.6%. When n was 100, QIC's rate was 72.1% while that of AIC was 94.6%. The result showed that AIC was more powerful in selecting the true exchangeable structure than QIC. He attributed the low performance of QIC compared to AIC to the low efficiency of GEE estimator compared to the MLE of β which is more

efficient and the fact that information on correct correlation structure is not embedded directly in the quasi-likelihood in QIC. Further, he found that the QIC worked well in selection of predictors and not the working correlation structure. The independence structure assumed in deriving QIC is seemingly the simplest working assumption that can be adopted in all cases. However, for time-varying covariates, the resulting efficiency of the GEE estimator may be as low as 60 percent compared to the GEE estimator obtained by using the correct correlation structure (Fitzmaurice [26]). Besides, comparing the performance of QIC with AIC, Pan [60] did not examine any other properties of QIC. The current study sought to fill this void.

Barnett et al. [5] in their study on using information criteria to select the correct variance-covariance structure for longitudinal data in ecology compared the performance of AIC, QIC and the Deviance Information Criteria (DIC) for multivariate Gaussian responses. They considered a relatively small sample size (n=30) and a relatively large cluster size (m=8). Exchangeable correlation was parameterized by $\rho = \{0.2, 0.5\}$ while the AR-1 was parameterized by $\rho = \{0.3, 0.7\}$. They considered the independence, exchangeable, AR-1 and Unstructured working correlation structures. The set included the unstructured matrix that was not considered by Pan [60]. They established that for EX(0.2) and AR-1(0.3) and independence true correlation structure, QIC's success rates were between 0 to 14%. Under these settings QIC preferred the unstructured correlation matrix despite the added complexity of the 0.5m(m-1) parameters to be estimated. They further established that the performance of QIC improved when the correlation was increased to moderate level. However, in their study they only considered a fixed number of subjects (30) and measurements per subject (8) hence could not numerically demonstrate the consistency property of QIC which the study will seek to establish. In the current study within-subject correlation, number of subjects and measurements per subject were allowed to vary and inferences made on the consistency, sensitivity and sparsity of QIC.

Hin et al. [35] considered the performance of QIC in selecting the AR-1 and exchangeable structures compared to RJ criteria using simulated gaussian data. They established that for the exchangeable true structure, QIC selection rates were between 65% and 77% compared to those of the RJ criteria which were between 84% and 100%. For the AR-1 true correlation structure, the selection rates of QIC were between 64% and 81%. On the other hand RJ instead preferred the exchangeable structure with rates of between 50% and 71%. The results show that QIC outperformed RJ for AR-1 structures while RJ was superior than QIC for exchangeable correlation structures signaling that neither of the two criteria can be considered dominant in the selection of the true correlation structure in GEE. Hin et al. [35] considered only Gaussian data and did not consider discrete data with a binary outcome which is the focus of the current study.

Shinpei [71], asserted that the derivation of QIC by Pan [60] ignored the computation of the correlation parameter and recommended a formal derivation of the QIC (called formal QIC or fQIC) as an asymptotic unbiased estimate of the prediction risk based on the quasi-likelihood. In the re-derivation of formal QIC, he considered the effect of estimating the correlation matrix used in the GEE procedure and the adequacy of the risk function used. He observed that the original QIC was exactly and asymptotically equivalent to the formal QIC when the working correlation matrix was independence. He further compared the performance QIC and fQIC and established that the bias of the original QIC got larger when the number of parameters increased while fQIC kept a stable value in each model. This indicated that the performance of QIC was a function of the number of parameters to estimated which Pan [60] failed to consider in his derivation of QIC.

Unlike Shinpei [71], Deroche [16] proposed two modifications of QIC. In the first modification she multiplied the penalty term of the original QIC with 2p to penalize for the number of regression parameters. However, the resultant modified QIC selected the independence structure with a probability of one regardless of

the sample size. In the second modification she subtracted a third penalty term $(m * trace(\hat{\Omega}_I \hat{V}_I))$ with the intention of penalizing for the number of correlation parameters estimated. Unlike the first modified QIC, the second modified QIC which sought to provide a balance between the independent structure and the unstructured structure that estimates the most correlation parameters ended up favoring the unstructured structure always just like the original QIC.

Hin and Wang [34] observed that the lack of a correlation structure in the first term $[-2Q(\hat{\beta}; I; (R); D)]$ of QIC makes it an insensitive measure to use for working correlation structure selection. However, they observed that the second term $2trace(\hat{\Omega}_I V_r)$ contained information about the anticipated correlation structure through the robust variance estimator (V_r) and used half of this term to develop the correlation information criteria (CIC). They established that when the true correlation was exchangeable, the correct identification rates for QIC ranged from 62% to 72% while those o CIC were between 81% and 96%. For an AR-1 true correlation structure, the selection rates for QIC were between 57% and 73% while those of CIC ranged from 82% to 97%. They further established that the magnitude of the first term in QIC is at its minimum when β is estimated under working independence, resulting in bias towards selecting the independence correlation structure while in contrast, CIC achieves its minimum for the true working correlation structure in which $V_i^{-1} = cov(y_i)$ by the Gauss-Markov theorem. However, they concluded that CIC cannot penalize for over-parameterization.

Gosho et al. [29] in their efforts to enhance selection of a proper correlation structure for GEE modeling proposed modifications to both QIC and CIC by replacing the robust sandwich variance estimator (V_r) in the original QIC and CIC with bias corrected variance estimators V_{KC} by Kaurmann and Caroll [45], V_{MD} by Mancl and DeRouen [51] and V_{PA} by Pan [61]. Using a simulation study, they established that their proposed new criteria QIC_{MD}, QIC_{KC}, QIC_{PA},CIC_{MD},CIC_{KC} and CIC_{PA} selected the true correlation structure with higher proportions regardless of ρ compared to the original QIC and CIC. They did not however assess the

applicability of their criteria to the selection of covariates and relied on the evaluation by Pan [60] in applying their criteria to the Air Polution Data set contained in Stokes, Davis and Koch [73].

Jang [42] in her PhD dissertation on working correlation structure selection in generalized estimating equations showed that if the the correct correlation structure was AR-1, the success rates of QIC were 22.1%, 23.9% and 24.9% for n=30, 50 and 100 respectively. If the true correlation structure was exchangeable, its success rates were 10.5%, 10.9% and 10.8% for n=30, 50 and 100 respectively and if the true correlation structure was unstructured , its success rates were 34.0%, 36.8% and 40.4% for n=30, 50 and 100 respectively. Further, she established that QIC selected the independence structure the highest number of times when the true structure was AR-1 and Exchangeable. However, when the true structure was unstructured, QIC favoured AR-1 structure. She concluded that QIC's success rate in selecting the correct correlation structure was low hence it was not powerful in choosing the correct correlation structure. One possible reason is that it is not based on a likelihood that contains information about the correlation among repeated measurements.

Chen and Nicole [12] in their study on the selection of working correlation structure in GEE via empirical likelihood considered the use of Empirical Likelihood approach to select a working correlations for GEE models. They substituted the empirical likelihood for the parametric likelihood in AIC and proposed an empirical likelihood version of AIC given as $EAIC(s) = -2logR^F(\hat{\theta}^s_G) + 2dim(\theta^s)$, where s is the index of a candidate model parameterized by θ^s, (s=1,.....S), and $\hat{\theta}^s_G$ is the GEE estimate associated with the working correlation structure R_s. Through simulations they showed that EAIC was much powerful in the selection of the correct correlation structure compared to other model selection criteria including QIC, bootstrap-based criteria of minimum predictive mean squared error and CIC. They observed that, the effective way of improving the estimation efficiency within the GEE framework was to select among competing GEE models the one

that assumes the correct working correlation structure for repeated measurements. However, they did not establish whether their assertions could actually result to gain in efficiency of the GEE estimator.

2.4 Selection of covariates for the Mean Structure in GEE

One of the earliest model selection approach established for GEE was sequential testing with Wald-Z-tests on individual coefficients. It is used to test the hypothesis $H_0 : \beta_k = 0$ and is calculated using the estimate of $\hat{\beta}_k$ and dividing it by the model based standard error estimate of $\hat{\beta}_k$. However, the Wald statistic has poor properties as $\mid \hat{\beta}_k \mid$ gets large hence $var(\hat{\beta}_k)$ is replaced with the GEE robust variance estimate $\widetilde{Var}(\hat{\beta}_k)$. This gives a measure of partial association which under the null is approximately a Chi-Square with one degree of freedom. For a GEE model $E(Y_{ij} \mid X_{ij}) = g(X_{ij}^T \beta_k)$, the Wald test to test $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ can be used to form an R^2 statistic. However, testing alone was too simplistic for choosing predictive models as it provides only a single supposedly best model. This can be overcome by the use of penalized model fit statistics like AIC which can be used to compare non-nested models, and can also be used to find lists of plausible models (Cantoni et al. [8]).

Zheng [84], suggested the use of marginal R^2 (R_m^2) as measure of model fit for GEE models which was an extension of classical R^2. However, this measure ignored correlation and did not attempt to weight the residuals, even though $\hat{\beta}$ comes from a model with working covariance weights. Further, R_m^2 cannot generally be used for variable selection purposes, since like classical R^2 it would lead to choosing the largest model available.

Cantoni et al. [8], suggested a generalization of Mallows' C_p for GEE models, for estimating predictive risk under a general weighted loss function. The weights can be adjusted in order to account for correlation within subjects as well as down weighting unusual observations, potentially providing robustness against outliers

and model misspecification. They derived a GC_p statistic to estimate the resulting risk function. The resulting statistic however required Monte Carlo approximation to evaluate.

Pan [60] considered the problem and extended the classical derivation of AIC which involved estimating the relative Kullback-I divergence of each likelihood model from an unknown true model, to a GEE setting. In GEE, the likelihood is not specified, but a quasi-likelihood may be implicitly specified. He adapted the original derivation of the AIC, which involved estimating the expected model-based log-likelihood under the true model, to instead estimate an expected working-independence quasi-likelihood. This resulted in the quasi-likelihood information criteria (QIC) given in equation (1.106). Even though he established that QIC performed very well in variable selection, its dependence on working independence impedes its performance in cases of very strong correlation. In his simulation, Pan [60] only concentrated in determining the frequency of selection for the true model by QIC. He assumed the independence, compound symmetry and AR-1 correlation structures and did not determine the type I and II error rates, sensitivity and sparsity of QIC. The study sought to fill this void by studying the properties of QIC in selecting covariates for the mean structure in GEE.

Hardin and Hilbe [32] considered a modified version of QIC and established through simulations that QIC_{HH} just like QIC tends to be more sensitive to changes in the mean structure than changes in the covariance structure hence suitable in the selection of covariates.

While trying to address the limitation of GEE reliance on a working covariance matrix $R(\rho)$, Qu et al. [63] suggested the quadratic inference functions (QIF) to avoid explicit estimation of nuisance parameters. He proposed that efficiency might be improved by directly minimizing the quadratic inference function (QIF) $Q_n(\beta) = nm_n^T \hat{C}_n^- m_n$, where $\hat{C}_n(\beta) = Cov(m_n(\beta))$. However, Chen [12] compared the efficiency of the estimates resulting from GEE with the correct correlation structure and the estimates resulting from the use QIF and established that using

GEE with the true correlation structure resulted to greater efficiency. Likewise, Jamshid et al. [41] using a continuous response variable established that with a mis-specified AR-1 correlation structure, relative of efficiency of the parameter estimates of QIF over GEE was 1.23 and for a mis-specified exchangeable structure, it was 1.001. However, they established similar results when the correlation structure was correctly specified hence underscoring the importance of using the correct correlation in GEE model estimation. Similar results to Qu [63] were obtained by Adefowope et. al [1] who established a relative efficiency value of 1.1117 for AR-1 structure and 1.3082 for true exchangeable structure an indicator that QIF provided more efficient parameter estimates than GEE. This view was in contrast with the results by Chen and Nicole [12].

In trying to improve efficiency of GEE estimates, Erfanal et al. [22] examined the impact of height on the occurrence of Type II diabetes, and applied QIC to select the relevant covariates and CIC to select the appropriate correlation structure. Based on QIC values, the model with covariates height, education level and gender was selected as the best model and based on the CIC values, the unstructured correlation structure was preferred for the data. Their study showed that there existed a statistically significant inverse relationship between height of an individual and the development of Type II diabetes. However, in their study they did not assess the efficiency of estimates resulting from the model selected through the combined CIC-QIC approach compared to when QIC is used to select both the correlation structure and covariates and this forms the basis for the hybrid methodology proposed in the study.

The approach employed by Erfanul et al. [22] followed recommendations by Jang [42] who iterated that no single model selection criteria exists that can select covariates, correlation structure and variance function in GEE modeling with high rates of success and recommended that future studies should focus on combining proposed model selection criteria so as to develop model selection strategies that could improve optimality of the selected GEE models. Optimality of the models

means that the model selected by QIC should result to minimum variance unbiased estimators (MVUE). Further, Jianwen et al. [43] underscored the importance of hybrid methodologies by establishing that model selection criteria could only be effective in selecting covariates when the correlation structure is correctly specified. They further observed that GEEs with appropriate $\hat{\rho}$ have good efficiency.

Fan and Li [23] observed that a good model selection criteria should identify the correct model asymptotically with probability one provided that the correct model is included in the set of candidate models. Likewise, Dziak [19] observed that, for consistent model selection, two properties are required: sensitivity and sparsity. Sensitivity implies that the model selection criteria retains all of the coefficients which should be retained with a probability approaching one hence reduced false negative rates while sparsity implies that the model selection criteria should delete all of the coefficients which ought to be deleted with probability approaching one hence reduce the false positive rates. Little studies have focused of establishing the consistency of QIC in selecting the covariates for the mean structure despite the increased routine use of QIC in model selection as recommended by Pan [60]. For instance, Wang et al. [76] indicated that in 2014, there were 111 citations of Pan's [60] article with over 80% of them being in non-statical journals and in the first eight months of 2015, there were still over 100 citations of the article majorly in medical journals. They asserted that the increased use of QIC is without peril hence the need for studies to bring out a better understanding of QIC so that mitigation measures can be developed to overcome its shortcomings and improve on statistical performance of selected models.

2.5 Summary

From the literature on selection of the working correlation structure, it was established that the RJ criterion has high sensitivity in identifying exchangeable correlation structure but it has low specificity, which limits its usefulness, even in comparisons of structures with only one parameter (e.g.exchangeable and AR-1).

The SC criterion performs even worse than the RJ criterion in correctly identifying the underlying true covariance structure. The QIC criterion generally performs better than the SC and RJ criterion in identifying the working correlation structure. However, similar to the SC criterion, the quasi-likelihood under independence assumption which is the first term in QIC, dilutes the impact of different working correlation structures on this measure, while retaining sensitivity to differences in the mean structures of competing models. It was shown by Hin and Wang [34] that excluding the first term in QIC improved the correct identification rates of the covariance structure. Barnett et al. [5] showed that QIC criterion performed very poor when the unstructured correlation structure was included in a set of candidate models. Little attention has however been drawn to the consistency properties of QIC since most studies to date have majorly compared the performance of QIC with other model selection criteria in which QIC has been established to exhibit poor performance in the selection of the working correlation structure. The selection of misspecified working correlation structure is bound to cause loss of efficiency in the GEE estimates ($\hat{\beta}_G$). The study sought to address the issue by examining the properties of QIC and determine whether the selection of the true correlation structure improves efficiency of $\hat{\beta}_G$ and efficiency of the overall model selected by QIC.

CHAPTER 3

PROPERTIES OF QIC IN SELECTING THE TRUE CORRELATION STRUCTURE FOR GENERALIZED ESTIMATING EQUATIONS

3.1 Introduction

In this chapter, using simulations, we investigated the properties of QIC in selecting the correct correlation structure in GEE relative to changes in the number of subjects (n), measurements per subject (m) and degree of within-subject correlation (ρ). We particularly investigated its consistency in selecting the true correlation structure and established conditions under which the consistency property held. In this regard, consistency implied that with probability approaching one, QIC selected the true correlation structure as the sample size tended to infinity.

To formally state the consistency property, let ω be the set working correlation structures (R(ρ)) that involves at least one correct correlation structure. Let R_0 be the true correlation structure and R_* be the correlation structure selected by QIC. For theoretical purposes, we divide ω into over-parameterized set ω^+ and the under-specified set ω^- i.e.

$$\omega^+ = \{R \in \omega | \exists \quad \rho \in \Theta \quad s.t. \quad R(\rho) = R_0\} \tag{3.1}$$

Where Θ is the parameter space, which is a compact set and $\omega^- = \omega \setminus \omega^+$. $\forall$ $R \in \omega^+$, we assume that there exists $\rho \in \Theta^0$ such that $R(\rho) = R_0$, where Θ^0 is in the interior of Θ.

Assuming the same sufficient conditions for consistency as contained in Shinpei [70]

(C1) the GEE mean structure is correctly specified

(C2) $\forall\ R \in \omega, \sqrt{n}(\hat{\beta} - \beta) = O_p$ and $\sqrt{n}(\hat{\phi} - \phi) = O_p$

(C3) h(η_{it}) is differentiable

(C4) $\forall\ R \in \omega^+, \sqrt{n}(\hat{\rho} - \rho) = O_p$ and R(.) is a differentiable function at ρ where ρ satisfies R(ρ)=R_0

then;

$$\begin{aligned} Pr(R_*(n) &= R_0) = 1 - Pr(R_*(n) \neq R_0) \\ &= 1 - \sum_{R \in \omega \setminus \{R_0\}} Pr(R_*(n) = R) \\ &= 1 - \sum_{R \in \omega^+ \setminus \{R_0\}} Pr(R_*(n) = R) - \sum_{R \in \omega^-} Pr(R_*(n) = R) \end{aligned} \tag{3.2}$$

If:

$$\lim_{n \to \infty} Pr(R_*(n) = R) = 0, \forall R \in \omega^+ \setminus \{R_0\} \tag{3.3}$$

and

$$\lim_{n \to \infty} Pr(R_*(n) = R) = 0, \forall R \in \omega^- \tag{3.4}$$

then;

$$\lim_{n \to \infty} Pr(R_*(n) = R_0) = 1 \tag{3.5}$$

This implies that the probability of selecting the true correlation structure (R_0) converges to one as n tends to infinity. In this regard QIC will be regarded consistent in selecting the true structure.

If:

$$\lim_{n \to \infty} \sum_{R \in \omega \setminus \{R_0\}} Pr(R_*(n) = R) = 1 \tag{3.6}$$

then;

$$\lim_{n \to \infty} Pr(R_*(n) = R_0) = 0 \tag{3.7}$$

This implies that the probability of selecting the true correlation structure (R_0) converges to zero hence the selection criteria will be regarded not to be consistent.

3.2 Simulation Settings

The simulation settings were as follows:

(a) The response vector $Y_i = (y_{i1}, ..., y_{it})$ was assumed to be a Bernoulli response. i=1,2....n, where n is the total number of subjects. In the simulation studies $n = \{20, 30, 50, 100, 200\}$; t=1,2,...m; m is the number of observations per subject. In the simulation studies $m = \{3, 6, 9\}$. The values of n, m and ρ in the simulation settings were improvements to those adopted by Pan [60] and Gosho et al. [29]. This was done determine whether increasing the values of n, m and ρ improves the performance of QIC in selecting the true structure.

(b) For each subject i, its covariates were X_{it}=$[X_{1it}, X_{2it}]^T$. $X_{1it} \sim N(0, 1)$ and $X_{2it} \sim Bernoulli(0.5)$ and a within subject correlation structure dictated by R_0 true correlation structures.

(c) The true correlation structures (R_0) considered in the simulation were exchangeable (ρ), AR-1 (ρ): $\rho = \{0.2, 0.5, 0.8\}$ and the unstructured correlation matrix. The unstructured matrix was as follows:

$$R(\rho)_{UN} = \begin{pmatrix} 1.00 & 0.80 & 0.60 & 0.14 & 0.10 & 0.23 \\ 0.80 & 1.00 & 0.70 & 0.18 & 0.17 & 0.18 \\ 0.60 & 0.70 & 1.00 & 0.25 & 0.24 & 0.22 \\ 0.14 & 0.18 & 0.25 & 1.00 & 0.45 & 0.22 \\ 0.10 & 0.17 & 0.24 & 0.45 & 1.00 & 0.16 \\ 0.23 & 0.18 & 0.22 & 0.22 & 0.16 & 1.00 \end{pmatrix}$$

The matrix was the same as that used by Jang [42]. For the unstructured working correlation structure, the number of measurements per subject (m) were taken to be 3, 5 and 6.

(d) The binary response y_{it} has the conditional expectation μ_{it}:

$\mu_{it} = E(y_{it}|X_{1it}, X_{2it})$

μ_{it} can be connected with the covariates through:

$logit(\mu_{it}) = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it}$; where $i = 1...n$ and $t = 1...m$.

(e) The coefficients were set to be $\beta_0 = 0.25 = -\beta_1 = -\beta_2$ which were similar to the ones assumed in Pan [60]. These were adopted to facilitate comparison of his study results with the study results.

(f) We considered two sets of correlation structures: $\omega_1 = \{IN, EX, UN, AR{-}1\}$ and $\omega_2 = \{IN, EX, AR-1\}$. All the correlation matrices were assumed to be positive definitive.

(g) All simulations were performed using R version 3.6.0 based on the gee, geepack, MASS, MESS and MuMIn R software packages. Correlated binary data were generated using the bindata (Friedrich et al.[27]) and SimCorMultRes (Touloumis[75]) library packages.

(h) The joint distribution of the Y_i was simulated using the procedure suggested by Touloumis[75] in the SimCorMultRes package.

(i) A sample of $N = n * m_i$ independent simulated observations with response vector of size m, true correlation structure (R_0) and distribution D were generated under the marginal mean and (R_0). The model was fit under the correct marginal mean model and variance function assumptions. The number of simulation replications were 1000. This was obtained using the procedure suggested by Morris et al.[54]. Our desired coverage probability was 95% such that the standard normal variate (Z_c=1.96) and level of precision (d) was 0.01. The sample standard deviation of the response variable ($SE_{\bar{y}}$) obtained from a sample of 200 preliminary runs was 0.16134 hence $n_{sim} = (\frac{Z_c \times SE_{\bar{y}}}{d})^2 = (\frac{1.96 \times 0.16134}{0.01})^2 = 999.999 \approx 1000$, where n_{sim} was the simulation sample.

3.3 Simulation Results of the Working Correlation Structure Selection by QIC from 1000 Replications: ω_1={IN, EX, AR-1, UN}

In this section, the correct identification rates of QIC under the different true correlation structures are presented. (See Appendix B.1)

3.3.1 Selection Rates of the True AR-1 Correlation Structure by QIC

Table 3.1 present the correct identification rates of QIC for the AR-1 true correlation structure for $n = \{20, 30, 50, 100, 200\}$, $m = \{3, 6, 9\}$ and $\rho = \{0.2, 0.5, 0.8\}$.

Table 3.1: The number of times each of the working correlation structures is selected out of 1000 simulation runs by QIC: $R_0 = AR - 1$

		m=3				m=6				m=9			
R_0	n	IN	EX	AR-1	UN	IN	EX	AR-1	UN	IN	EX	AR-1	UN
AR-1	20	318	129	**207**	346	303	96	**266**	335	297	99	**273**	331
$\rho = 0.2$	30	299	108	**244**	349	301	80	**272**	347	285	84	**277**	354
	50	215	109	**294**	382	249	77	**318**	356	205	71	**323**	381
	100	186	107	**306**	401	193	68	**324**	415	173	66	**343**	418
	200	152	111	**333**	404	181	53	**347**	419	175	53	**356**	416
AR-1	20	274	169	**219**	338	267	129	**263**	341	239	123	**269**	363
$\rho = 0.5$	30	199	172	**253**	376	258	131	**273**	338	236	125	**281**	358
	50	178	146	**309**	367	192	134	**320**	354	195	103	**327**	375
	100	178	113	**321**	388	163	97	**355**	375	174	97	**344**	385
	200	182	75	**339**	404	133	89	**377**	401	158	97	**366**	378
AR-1	20	326	115	**221**	338	266	130	**259**	345	211	127	**288**	374
$\rho = 0.8$	30	307	101	**259**	333	248	114	**279**	359	200	108	**316**	376
	50	225	93	**321**	361	171	103	**348**	378	176	93	**351**	380
	100	173	76	**356**	395	125	87	**387**	404	114	93	**388**	405
	200	145	71	**387**	397	110	67	**407**	416	107	49	**419**	425

The results of the analysis show that when the the correct correlation structure was AR-1 (ρ=0.2) and m=3 , the success rates of QIC were 20.7%, 24.4%, 29.4% 30.6% and 33.3% for n=20, 30, 50, 100 and 200 respectively. When the number of measurements per subject were increased to 6, the success rates of QIC for the respective samples were 26.6%, 27.2%, 31.8% 32.4% and 34.7%. When the number of measurements per subject were further increased to 9, the success rates of QIC for the respective samples were 27.3%, 27.7%, 32.3% 34.7% and 35.6% . The results indicate that the frequency of correct identification of the true AR-

1 correlation structure increased with the number of measurements per subject although in a higher rate when $n \geq 30$. The increase was on average 10% when number of measurements per subject increased from 3 to 6 and 3% when they are increased from 6 to 9.

When the correct correlation structure was $R_0 = AR - 1(\rho = 0.5)$ and m=3, the success rates of QIC were 21.9%, 25.3%, 30.9% 32.1% and 33.9% for n=20, 30, 50, 100 and 200 respectively. When the number of measurements per subject were increased to 6, the success rates of QIC were 26.3%, 27.3%, 32.0% 35.5% and 37.7% for n=20, 30, 50, 100 and 200 respectively. When the number of measurements per subject were increased to 9, the success rates of QIC are 26.9%, 28.1%, 32.7% 34.4% and 36.6% for n=20, 30, 50, 100 and 200 respectively. The study results indicate that increasing the level of correlation to 0.5 resulted to an overall increase in the success rates of QIC in selecting the true correlation structure by approximately 5%.

When the the correct correlation structure was AR-1 (ρ=0.8) and m=3, the success rates of QIC were 22.1%, 25.9%, 32.1% 35.6% and 38.7% for n=20, 30, 50, 100 and 200 respectively. When the number of measurements per subject were increased to 6, the success rates of QIC were 25.9%, 27.9%, 34.8% 38.7% and 40.7% for n=20, 30, 50, 100 and 200 respectively. When the number of measurements per subject were increased to 9, the success rates of QIC were 28.8%, 31.6%, 35.1% 38.8% and 41.9% for n=20, 30, 50, 100 and 200 respectively which also indicated an increase in the frequency of correct identification of the true correlation structure as the level of correlation and number of measurements per subject increased.

The finding that QIC identification rates of the correct correlation structure increased with increase in the level of correlation corroborates assertions by Wang and Carey [79] that when the level of correlation is high, the true correlation structure is to a greater extent differentiable from the other working correlation structures hence increased probabilities of identification. It is also notable that regardless ρ, the success rates of QIC were greater than 30% when $n \geq 30$ and the

rate of selecting the AR-1 true correlation structure generally increased with n, m and ρ. The study findings were also similar to findings by Barnett et. al[5] who established that QIC performed poorly with a weak correlation structure and did much better when the within-subject correlation increased to moderate level.

Further, the results indicate that when R_0=AR-1, QIC favoured the unstructured correlation structure which increased as the degree of correlation. This signaled QIC's preference for over-parameterized working correlation structures. This is in line with assertions by Shinpei [71] that QIC does not penalize for the number of correlation parameters estimated. The results also show that QIC rarely chose under-parameterized structures. For instance, when R_0 was AR-1, QIC selected the independence or exchangeable correlation structures $\ll 30\%$ of the time. The results are further presented in Figure 3.1.

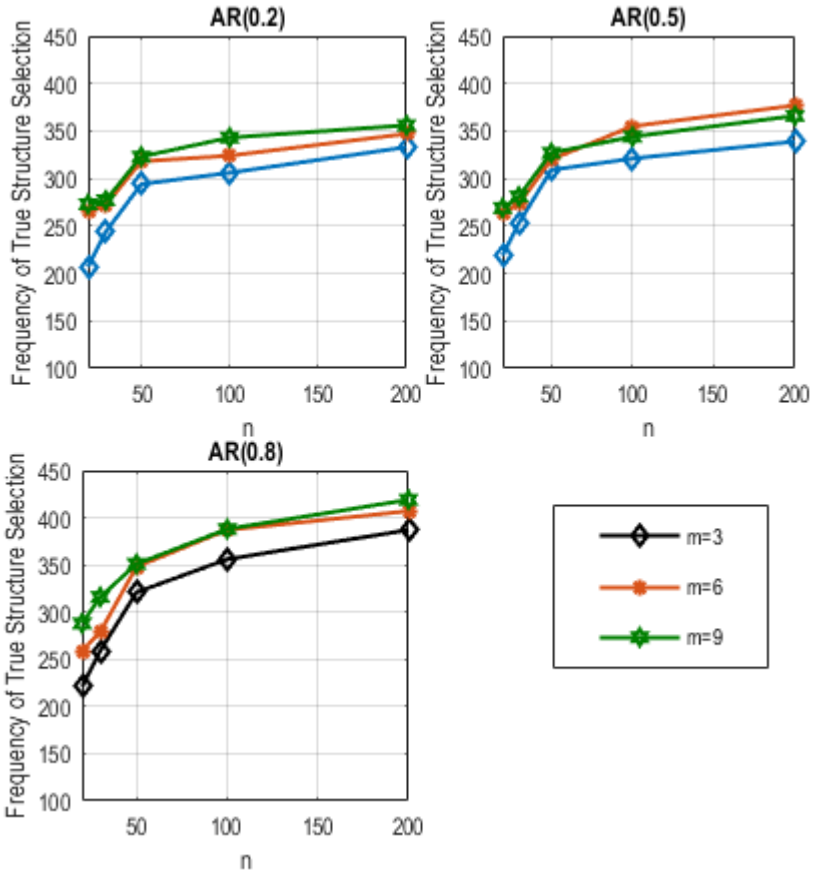

Figure 3.1: Correct Identification Rates of the True AR-1 Correlation Structure out of 1000 Replications

The results indicate that the performance of QIC in selecting the AR-1 true correlation structure improves with the degree of correlation. It is highest in all cases for a strong correlation ρ=0.8 followed by moderate correlation ρ=0.5. Its performance is lowest when their is a weak within-subject correlation. It is notable that simultaneously increasing the degree of correlation and the number of measurements per subject makes it more easier for QIC to select the correct correlation structure.

Simulation results from Gosho et al. [29] are reorganized into Table 3.2 so that we can make further comparison on the performance of QIC in selecting the AR-1 true working correlation structure. They assumed three levels of correlation 0.1, 0.3 and 0.5 and 4 and 8 measurements per subject. Also, they considered sample sizes of 10, 30 and 100 and the independence, AR-1 and exchangeable correlation structures as probable choices and never included over-parameterized structures such as the unstructured correlation structure.

Table 3.2: Selection Rates for AR-1 true correlation structure by Gosho et al. [29]

		m=4			m=8		
R_0	n	IN	EX	AR-1	IN	EX	AR-1
AR-1	10	351	328	**321**	333	326	**342**
(0.1)	30	251	352	**398**	247	358	**426**
	100	228	307	**465**	199	289	**480**
AR-1	10	320	325	**355**	247	281	**472**
(0.3)	30	211	252	**537**	193	251	**556**
	100	128	232	**640**	115	211	**674**
AR-1	10	316	342	**342**	218	291	**492**
(0.5)	30	152	242	**607**	133	227	**640**
	100	107	218	**675**	93	187	**720**

From Table 3.2, it is noted that regardless of the degree of correlation, the proportion of selecting the true correlation structure is quite low when the sample size and measurements per subject are small and it increased with an increase in n, m and ρ up to a high of 72% when n=100, m=8 and $\rho = 0.5$. This finding is similar

to the study finding which showed an increase in QIC's success rate with increase in n, m and ρ. In Gosho et al.[29], the success rates of QIC ranged from 32.1% to 72% indicating that the probability of QIC selecting the true AR-1 correlation approached one as $n \to \infty$ compared to the study success rates which ranged from 20.7% to 41.9% indicating that the the probability of QIC selecting the true AR-1 correlation approached one as $n \to \infty$ but at a slower rate hence may need unrealistically large number of subjects. This implies that the sample size of 200 is still not large enough for the asymptotic consistency to be achieved.

Stability analysis results are illustrated in Figures 3.2 and 3.3

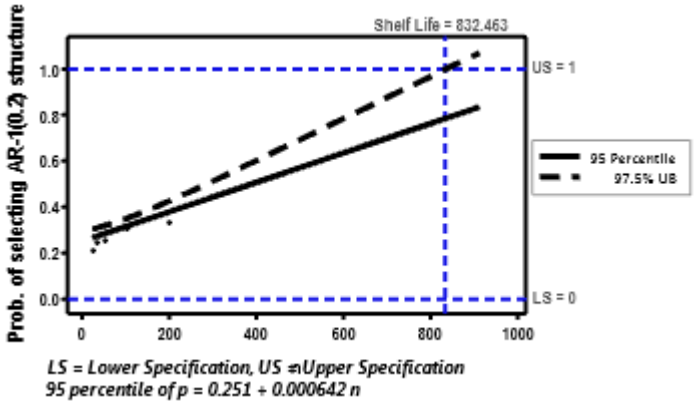

Figure 3.2: Stability Analysis: Probability of Selecting AR-1(0.2) Structure

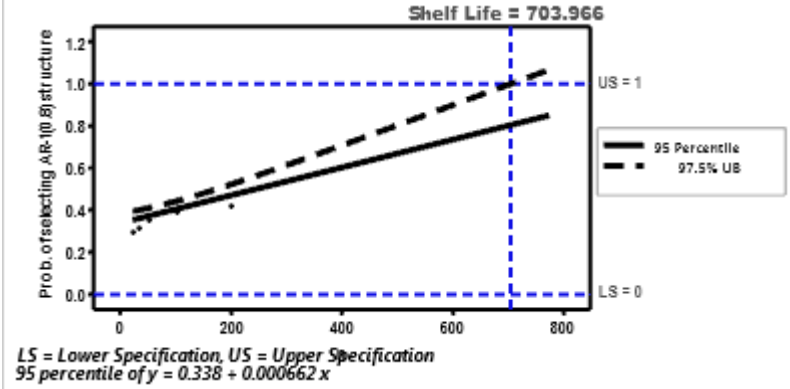

Figure 3.3: Stability Analysis: Probability of Selecting AR-1(0.8) Structure

From Figure 3.3, it is noticeable that for $\rho = 0.8$, m=9 and n=705, $Pr(R_*(n) = R_0) \simeq 0.7$ and for $\rho = 0.2$, m=3 and n=832, $Pr(R_*(n) = R_0) \simeq 0.65$(Figure 3.2). The results indicate that at $\rho = 0.8$ and m=9 allowing for a 95% upper bound, it will require at least n=705 for $Pr(R_*(n) = R_0) \to 1$ as $n \to \infty$ and that at

$\rho = 0.2$ and m=3 allowing for a 95% upper bound, it will require at least n=832 for $Pr(R_*(n) = R_0) \to 1$ as $n \to \infty$. The results show a slow rate of convergence in probability to one for QIC's selection of the true AR-1 structure.

3.3.2 Selection Rates of the True Exchangeable Structure by QIC

Results of this simulation study were used to assess the performance of QIC in selecting the true exchangeable correlation structure for different levels of ρ, n and m. The selection frequency counts of the independence, AR-1, exchangeable and unstructured correlation structures out of the 1000 independent replications are tabulated in Table 3.3

Table 3.3: Simulation Results for Selection of true exchangeable correlation structure

		m=3				m=6				m=9			
R_0	n	IN	EX	AR-1	UN	IN	EX	AR-1	UN	IN	EX	AR-1	UN
EX	20	224	**105**	205	466	189	**107**	192	509	207	**106**	109	568
$\rho = 0.2$	30	203	**94**	254	449	181	**68**	189	563	193	**71**	192	544
	50	241	**80**	282	397	170	**68**	261	501	181	**69**	247	503
	100	264	**72**	336	328	241	**65**	284	410	211	**61**	291	437
	200	270	**63**	384	283	300	**66**	320	314	233	**63**	335	371
EX	20	239	**194**	194	373	286	**211**	164	339	256	**204**	119	391
$\rho = 0.5$	30	267	**170**	218	345	299	**222**	136	343	256	**197**	178	356
	50	287	**136**	269	308	295	**157**	207	341	311	**138**	200	341
	100	250	**95**	325	330	313	**97**	273	317	313	**101**	237	349
	200	193	**54**	394	359	247	**58**	359	336	323	**74**	251	346
EX	20	250	**103**	210	426	308	**123**	156	413	311	**121**	205	363
$\rho = 0.8$	30	241	**99**	235	425	321	**107**	173	399	292	**113**	244	351
	50	269	**91**	251	389	316	**92**	199	393	287	**86**	294	333
	100	215	**83**	301	401	312	**88**	203	397	281	**73**	330	316
	200	137	**76**	356	433	278	**79**	239	404	263	**61**	377	299

The results in Table 3.3 show that when the correct correlation structure was exchangeable (ρ=0.2) and m=3 , the success rate of QIC were 10.5%, 9.4%, 8.0%

7.2% and 6.3% for n=20, 30, 50, 100 and 200 respectively. When m is increased to 6, the success rates of QIC were 10.7%, 6.8%, 6.8% 6.5% and 6.6% for n=20, 30, 50, 100 and 200 respectively. When m is further increased to 9, the success rates of QIC were 10.6%, 7.1%, 6.9% 6.1% and 6.3% for n=20, 30, 50, 100 and 200 respectively. The results indicate that the frequency of correct identification of the exchangeable true correlation structure was higher for small samples and tended to decrease with increase in sample size. Increasing the number of measurements per subject did not seem to significantly improve the performance of QIC.

When the level of correlation(ρ) was increased to 0.5, the frequency of correct identification when m=3 were 19.4%, 17.0%, 13.6% 9.5% and 5.4% for n=20, 30, 50, 100 and 200 respectively. When m was increased to 6, the success rates were 21.1%, 22.2%, 15.7% 9.7% and 5.8% for the respective sample sizes and when m was increased to 9, the success rates were 20.4%, 19.7%, 13.8% 10.1% and 7.4% for the respective sample sizes.

When the level of correlation was further increased to 0.8 which is considered strong, the frequency of correct identification when m=3 were 10.3%, 9.9%, 9.1% 8.3% and 7.6% for n=20, 30, 50, 100 and 200 respectively. When the number of measurements per subject is increased to 6, the success rates become 12.3%, 10.7%, 9.2% 8.8% and 7.9% for n=20, 30, 50, 100 and 200 respectively. When the number of measurements per subject increases to 9 the success rates are 12.1%, 11.3%, 8.6% 7.3% and 6.1% for n=20, 30, 50, 100 and 200 respectively. The results indicated a marginal increase in the frequency of correct identification of the exchangeable true correlation structure with increase in the level of correlation. However, the increase for small samples was much higher than for larger samples. Likewise, the success rates improved slightly with increase in the number of measurements per subject. Overall, poor performance of QIC in selecting the true exchangeable correlation structure was established with QIC preferring the unstructured correlation structure most of the time. QIC selection rates established in the study were similar to those of Jang [42] who established QIC's performance

to be in the range of 0% to 20% and decreased as n increased. The results were also similar those of Barnett et al. [5] who established correct identification rates of 25-30% for moderately correlated exchangeable structure. Jang [42] asserted that such performance was due to the estimation of the over-parameterized structure which becomes more precise as n increases hence increasing the likelihood of the unstructured matrix being chosen as n increases. Deroche [16] established similar success rates for the correct exchangeable correlation structure which were less than 50% of the time under all combinations of m and ρ. She established that when the number of measurements on a single person gets large, Pan's QIC had more difficulty selecting the correct correlation structure. This was also true when the degree of correlation increased. The results are further illustrated in Figure 3.4.

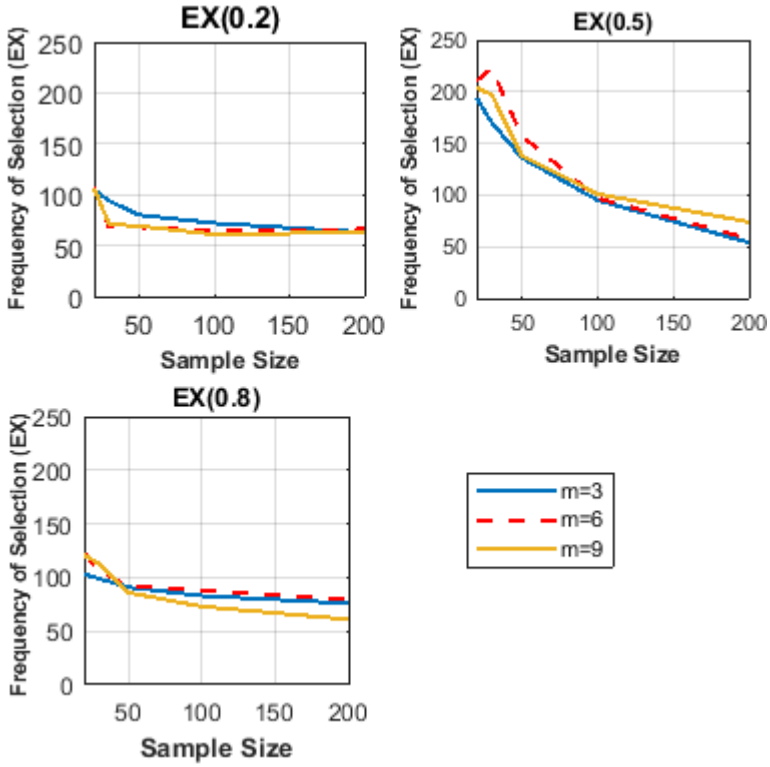

Figure 3.4: Identification rates by QIC for the true exchangeable correlation Structure

Figure 3.4 shows that the frequency of QIC selecting the exchangeable correlation

structure declined with increase in sample size such that as $n \to \infty$, the probability of QIC selecting the exchangeable correlation structure approached zero. The results indicated that QIC preferred the exchangeable structure for small samples and the decline was in spite of the increase in degree of correlation and number of measurements per subject.

The study results were compared to simulation results by Gosho et al. [29] and Pan [60] which are organized in Table 3.4 below:

Table 3.4: Results by Gosho et al. [29] and Pan[60] for Exchangeable Structure

			m=3			m=4			m=8		
AUTHOR	R_0	n	IN	EX	AR-1	IN	EX	AR-1	IN	EX	AR-1
GOSHO	EX(0.1)	10				358	**319**	323	232	**450**	318
		30				253	**408**	339	189	**503**	309
		100				208	**510**	282	84	**618**	299
	EX(0.3)	10				290	**319**	392	263	**330**	407
		30				211	**536**	253	113	**593**	294
		100				134	**676**	190	75	**650**	275
	EX(0.5)	10				324	**356**	320	317	**351**	333
		30				177	**607**	216	126	**602**	272
		100				136	**711**	153	112	**653**	235
PAN	EX(0.5)	50	138	**678**	184						
		100	140	**721**	139						

The results by Gosho et al. [29] indicated that the selection rates of the exchangeable true correlation structure increased as the sample size, number of measurements and level of correlation increased. The success rates of QIC ranged from 31.9% when n=10, m=4 and $\rho = 0.1$ to 63.5% when n=100, m=8 and $\rho = 0.5$. Their results were dissimilar to the study findings which showed a declining trend. This can be attributed to the exclusion of the unstructured correlation structure.

The results by Pan [60] who like Gosho et al. [29] only considered independence, exchangeable and AR-1 correlation structure and $R_0 = EX$ with $\rho = 0.5$ and m=3 showed the success rates of QIC to be 67.8% and 72.1% for sample sizes of

50 and 100 respectively. For the same settings, the study results were 13.6% and 9.5% respectively. These were very dissimilar results. Considering that the set of correlation structures considered in our study included the unstructured correlation structures which was not considered in the studies by Pan [60] and Gosho et al. [29], the performance of QIC in selecting the parsimonious exchangeable structure is dependent on the inclusion or exclusion of the over-parameterized unstructured correlation structure. Its inclusion seriously reduced its success rates while its exclusion greatly improved the success rates.

3.3.3 Selection Rates of the True Unstructured Structure by QIC

Frequencies of selecting the unstructured correlation structure by QIC from the 1000 independent replications are shown in Table 3.5;

Table 3.5: Selection rates of QIC for Unstructured true correlation structure

	m=3				m=5				m=6			
	IN	EX	AR-1	UN	IN	EX	AR-1	UN	IN	EX	AR-1	UN
20	341	257	177	**225**	272	180	271	**277**	292	170	201	**337**
30	320	256	177	**282**	267	188	258	**287**	259	192	196	**353**
50	322	248	122	**258**	273	159	247	**321**	233	164	186	**417**
100	303	255	128	**316**	258	148	227	**367**	237	171	135	**457**
200	307	234	113	**345**	225	137	244	**394**	198	155	144	**503**

The simulation results indicated that when the correct correlation structure was unstructured and m=3 , the success rates of QIC were 22.5%, 25.2%, 25.8% 31.6% and 34.5% for n=20, 30, 50, 100 and 200 respectively. When m was increased to 5, the selection rates were 27.7%, 28.7%, 32.1% 36.7% and 39.4% for n=20, 30, 50, 100 and 200 respectively. The study results were slightly higher than those obtained by Jang[42] who established success rates of between 17.1% and 28.0%. The difference in performance can be attributed to inclusion of the Toeplitz structure. When m was further increased to 6, QIC's selection rates were 33.7%, 35.3%, 41.7% 45.7% and 50.3% for n=20, 30, 50, 100 and 200 respectively. For the same

m=6, Jang[42] obtained success rates of 21.9%, 21.2%, 26.6% 29.0% and 35.4% which are slightly lower than our study results. The study findings indicate that the frequency of correct identification of the unstructured true correlation structure tended to increase as the sample size increased and number of measurements per subject. The results are also presented in Figure 3.5

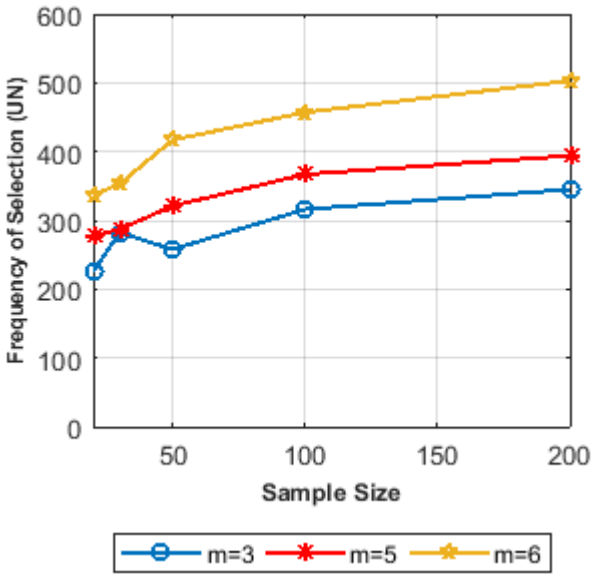

Figure 3.5: Simulation Results for WCS Selection: R_0=Unstructured

The graphical presentation show that the success rates of QIC in selecting the unstructured correlation structure were highest when the number of measurements per subject was 6 (highest) for all the sample sizes and were lowest when m was 3. This implies that QIC performance in selecting R_{UN} when it was the true correlation structure increased with increase in the sample sizes and that R_{UN} was preferred when the sample size was sufficient to estimate the 0.5m(m-1) parameters. The 34-50% success rates for m=6 were close to the 40-56% success rates established by Barnett et al. [5] for m=8. The result also indicated that QIC favours over-parameterized correlation structures with a probable reason of its inability to penalize for the number of correlation parameters.

3.4 Simulation Results of the Performance of QIC in Selecting the True Correlation Structure: ω_2={IN, EX, AR-1}

The correct identification rate for QIC of the true correlation structure out of 1000 independent simulations are determined with the candidate set being $\omega_2 = \{IN, EX, AR-1\}$ and $R_0 \in \omega_2$. In this case over-parameterized correlation structures are not considered. In the simulation, $\rho = \{0.2, 0.5\}$ and $m = \{3, 6\}$. Limiting to the two values of m and degrees of correlation was to facilitate comparison of our results with those by Gosho et al.[29] and Pan[60] who adopted similar settings.

3.4.1 Simulation Results on the Performance of QIC in Selecting the True AR-1 Correlation Structure

The number of times each correlation structure is selected from 1000 simulation runs when the true correlation structure was AR-1 are shown in Table 3.6

Table 3.6: Simulation Results when $R_0 = AR-1 | R_0 \in \omega_2$

R_0	n	m=3			m=6		
		IN	EX	AR-1	IN	EX	AR-1
AR-1 ($\rho = 0.2$)	20	283	300	**417**	223	243	**534**
	30	233	303	**464**	173	290	**537**
	50	190	343	**467**	189	198	**613**
	100	143	280	**577**	143	150	**707**
	200	144	206	**650**	90	160	**750**
AR-1($\rho = 0.5$)	20	150	327	**523**	150	196	**654**
	30	130	23	**607**	120	140	**740**
	50	130	210	**660**	110	90	**800**
	100	117	276	**707**	80	107	**813**
	200	100	177	**723**	54	126	**820**

The simulation results indicate that for $\rho = 0.2$ and m=3, the selection rates of QIC were 41.7%, 46.4%, 46.7%, 57.7% and 65.0% for sample size of 20, 30, 50,

100 and 200 respectively. When $\rho = 0.5$ for the same number of measurements per subject, QIC's selection rates for the true structure were 52.3%, 60.7%, 66.0%, 70.7% and 72.3% for the respective sample sizes of 20, 30, 50, 100 and 200. The results indicated that the selection rates increased with the increase in sample size and level of correlation. For m=6, the selection rates of the true correlation structure at $\rho = 0.2$ were 53.4%, 53.7%, 61.3%, 70.7% and 76.1% and at $\rho = 0.5$, they were 65.4%, 74.0%, 80.3%, 80.7% and 87.0% for the respective sample sizes of 20, 30, 50, 100 and 200. The results indicate that highest performance of QIC in selecting the true AR-1 correlation structure is achieved when $\rho = 0.5$ and m=6. In this quadrant the probability of QIC selecting the AR-1 correlation structure is more nearer to one hence consistency of QIC is inferred. This indicates that simultaneously increasing the number of subjects, measurements per subject and the degree of correlation increases the probability of QIC selecting the true AR-1 correlation structure. The results were similar to those of Gosho et al. [29] who established an increasing trend for the selection of AR-1 to a high of 72% for $\rho = 0.5$ and m=8 (Table 3.2). The results are further illustrated in Figure 3.6.

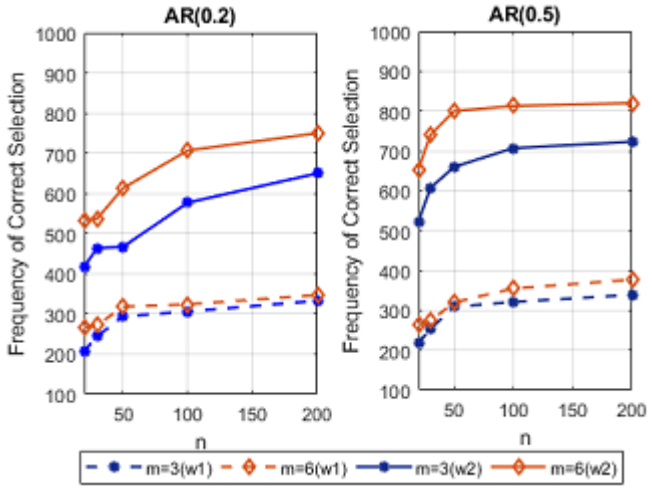

Figure 3.6: QIC's Selection Frequency of the True AR-1 Correlation Structure when only Parsimonious Structures are Considered

3.4.2 Simulation Results on the Performance of QIC in Selecting the True Exchangeable Correlation Structure

The selection frequencies of each correlation structure from 1000 simulation runs are shown in Table 3.7

Table 3.7: Simulation Results when $R_0 = EX|R_0 \in \omega_2$

R_0	n	m=3			m=6		
		IN	EX	AR-1	IN	EX	AR-1
EX($\rho = 0.2$)	20	268	**342**	390	328	**454**	217
	30	260	**410**	330	297	**509**	194
	50	300	**460**	240	266	**586**	148
	100	150	**620**	230	203	**677**	120
	200	145	**665**	190	187	**753**	60
EX($\rho = 0.5$)	20	286	**495**	219	237	**690**	73
	30	250	**527**	223	213	**733**	53
	50	208	**605**	187	157	**803**	40
	100	177	**710**	113	173	**807**	20
	200	127	**770**	103	110	870	20

When $\rho = 0.2$ and m=3, the selection rates of QIC were 34.2%, 41.0%, 46.0%, 62.0% and 66.5% for respective sample sizes of 20, 30, 50, 100 and 200. Increasing the level of correlation to 0.5 and maintaining m at 3, QIC's selection rates for the true structure increases to 49.5%, 52.7%, 60.5%, 71.0% and 77.0% for the respective sample sizes of 20, 30, 50, 100 and 200. The results showed an increase in selection rates with increase in sample size and level correlation. For m=6 and $\rho = 0.2$, the selection rates of the true correlation structure were 45.4%, 50.9%, 58.6%, 67.7% and 75.3% for the respective samples of 20,30, 50, 100 and 200 while for m=6 and $\rho = 0.5$, the selection rates were 69.0%, 73.3%, 80.3%, 80.7% and 87.0% for the respective sample sizes. This indicates that simultaneously increasing the number of subjects, measurements per subject and the degree of correlation increases the probability of QIC selecting the true exchangeable correlation structure. The

probability approaches one as $n \to \infty$ and are similar to those of Gosho et al. [29] and Pan [60] in Table 3.4. The results are further illustrated in the Figure 3.7.

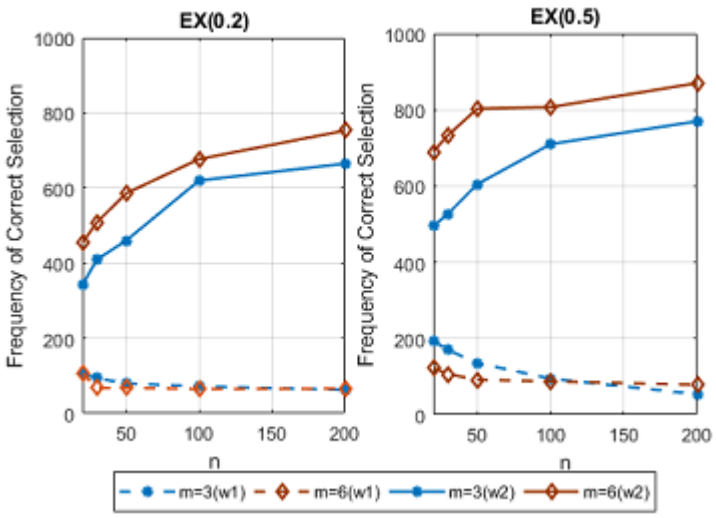

Figure 3.7: QIC's Selection Frequency of the True Exchangeable Correlation Structure when only Parsimonious Structures are Considered

The results from Figures 3.6 and 3.7 demonstrated that when the set of only parsimonious correlation structures, $\omega_2 = \{IN, EX, AR-1\}$ was considered and $R_0 \in \omega_2$, QIC asymptotically selected the correct working correlation structure with a probability of one for large n compared to when the selection set included over-parameterized correlation structures.

Conjecture 3.4.1. *Let* $\omega_1 = \{IN, EX, AR{-}1, UN\}$ *and* $\omega_2 = \{IN, EX, AR{-}1, \}$ *be sets of working correlation structures considered for selection by QIC. Let* R_0 *be the true correlation structures and* R_* *be the correlation structure selected. Then;*

(a) If ω_1 *is the set of possible correlation structures considered such that* $R_0 \in \omega_1$*, then;* $Pr(R_*(n) = R_0) \to 1$ *as* $n \to \infty$ *if and only if* $R_0 = \{AR-1, UN\}$ *but it is* $\ll 1$*. In this case we say that QIC is weakly-Consistent.*

(b) If ω_2 *is the set of possible correlation structures considered such that* $R_0 \in \omega_2$*, then* $Pr(R_*(n) = R_0) \to 1$ *as* $n \to \infty$ *hence QIC is said to be consistent.*

(c) Inclusion of over-parameterized structures such as the unstructured in the set of correlation structures considered for selection significantly reduces the

probability of selecting R_0. On the other hand, considering only the parsimonious structures significantly improves the probability of selecting R_0. i.e. if q is the number of correlation parameters and $R_0 \in (IN, EX, AR-1)$, then as $q \longrightarrow [0.5m(m-1)]$, $Pr(R_*(n) = R_0) < \sum_{R \in \omega^+ \setminus \{R_0\}} Pr(R_*(n) = R)$

3.5 Proposed Modification of the Quasi-Likelihood Information Criteria

From Conjecture 4.2.1, having $\omega_1 = \{IN, EX, AR-1, UN\}$ as the set of possible correlation matrices impedes the performance of QIC in selecting the parsimonious true correlation structures R_0; $R_0 \in \{IN, EX, AR-1\}$. This as observed by Barnet et al. [5] may be as a result of QIC not being able to penalize for the number of parameters estimated. This could also be attributed to the establishment by Yu and Shinpei [82] that $QIC^{(R)}$ as an estimator of the risk function $E_y[E_{y^*}\{-2\sum_{i=1}^{n}\sum_{t=1}^{m} Q(y^*; \hat{\beta})\}]$ where $y_i^* = (y_{i1}^*, .., y_{im}^*)$ is m-dimensional random vector independent of y_i, utilizes the independence assumption hence is not reflective of the correlation between responses. Further, the log quasi-likelihood function and the model based variance estimator $(I(\hat{\beta}^I|y))$ are estimated using the independent assumption which according to Wentao [81], makes $QIC^{(R)}$ value to be underestimated when correlation structures with higher number of correlation parameters are used in the GEE estimation. This implies that the bias correction term of QIC is not powerful enough to address the relatively larger log-quasi-likelihood values brought about by over-parameterized correlation matrices hence the need to improve the penalization weight of the bias correction term.

The goal was to develop a criteria that will select a parsimonious structure that closely approximates the true correlation matrix, without adding unnecessary parameters. In this regard, it should effectively penalize the working correlation structure with many parameters when the sample size is small or the responses are measured on many occasions and should also be able to penalize the working correlation structure which is parsimonious, but has a worse-fit.

3.5.1 Proposed Modified QIC

We proposed a modification to QIC so as to penalize for over-parameterization. The proposed modification uses the number (q) of correlation parameters which is a function of the number of repeated measurements (m) and the number (p) of regression parameters as cost components.

Definition 3.5.1. Let data from a subject i include S different working correlation structures ($R_1 ... R_S$). Each R_s has the corresponding $q_{max} = max\{q_i | i = 1...s, i \in \mathbb{Z}\}$. Also, let $M_{\bar{p}}$ be a model that is a subset of the full model with p predictor variables. Each of the subset models $M_{\bar{p}}$ has a particular number of covariates j ($0 \leq j \leq p, j \in \mathbb{Z}$). Further, let $p_{max} = max\{\bar{p}_j | j = 1...p\}$. We consider a loss function based on the weighted euclidean distance of (p,q) from the origin that takes the form;

$$d(p;q) = [wp^2 + (1-w)q^2]^{\frac{1}{2}} \qquad 0 \leq w \leq 1 \tag{3.9}$$

Since the scales of p and q are different, we consider the transformations $p* = \frac{\bar{p}_j}{p_{max}}$, where $\bar{p}_j \subset p_{max}$ and $q* = \frac{q_s}{q_{max}}$, $q_{max} = 0.5m(m-1)$ so that

$$d(p^*;q^*) = [w(p^*)^2 + (1-w)(q^*)^2]^{\frac{1}{2}} \qquad 0 \leq w \leq 1 \tag{3.10}$$

Since the dimension of covariance matrices is a function of p, we propose a penalization of the second term of QIC for the number of regression parameters by multiplying it by 2p as adopted by Deroche [16] in her proposed modification of QIC. This yielded a modified penalty term:

$$2trace(\hat{\Omega}_I \hat{V}_r) \times 2p = 4p \times trace(\hat{\Omega}_I \hat{V}_r) \tag{3.11}$$

Second, since we considered the penalty proposed by Deroche [16] in equation (3.11), we set w=0 and established a penalization factor of the form:

$$\begin{aligned} d(p^*;q^*) &= [\{\frac{q}{0.5m(m-1)}\}^2]^{\frac{1}{2}} \\ &= \frac{2q}{m(m-1)} \end{aligned} \tag{3.12}$$

Multiplying the resultant factor to $trace(\hat{\Omega}_I \hat{V}_r)$ we got the second penalty term:

$$\frac{q}{q_{max}} * trace(\hat{\Omega}_I \hat{V}_r) = \frac{2q}{m(m-2)} trace(\hat{\Omega}_I \hat{V}_r) \tag{3.13}$$

Assuming a working correlation structure $R(\rho)$ and incorporating adjustments by Shinpei [71] such that the fisher information matrix is defined using the true correlation structure i.e. $\hat{\Omega}_R$ in the penalty term in equation (3.13), we obtain a modified second penalty term:

$$\frac{q}{q_{max}} \times trace(\hat{\Omega}_I \hat{V}_r) = \frac{2q}{m(m-2)} trace(\hat{\Omega}_R \hat{V}_R) \tag{3.14}$$

The proposed modified QIC which we denote by $QIC_m(R)$ is therefore the sum of the goodness-of-fit term and the two proposed penalty terms in equations (3.11) and (3.14):

$$QIC_m(R) = -2Q(\hat{\beta}^R|y; I, \wp) + 4p \times trace(\hat{\Omega}_I \hat{V}_r) + \frac{2q}{m(m-1)} \times trace(\hat{\Omega}_R \hat{V}_R) \tag{3.15}$$

Where

$\Omega_I = \sum_{i=1}^n D_i^T A_i^{-1} D_i$, s.t $V_r = \hat{\Omega}_I^{-1} \{\frac{1}{n} \sum_{i=1}^n D_i^T V_i^{-1} (Y_i - \mu_i)(Y_i - \mu_i)^T V_i^{-1} D_i\} \hat{\Omega}_I^{-1}$

$\Omega_R = \sum_{i=1}^n D_i^T V_i^{-1} D_i$ s.t $\hat{V}_R = \hat{\Omega}_R^{-1} \{\frac{1}{n} \sum_{i=1}^n D_i^T V_i^{-1} (Y_i - \mu_i)(Y_i - \mu_i)^T V_i^{-1} D_i\} \hat{\Omega}_R^{-1}$

Remark 3.5.2. If the working correlation matrix is independence, $QIC_m(R)$'s third term becomes zero. $QIC_m(R)$ in equation (3.15), reduces to:

$$QIC_m(I) = -2Q(\hat{\beta}^I|y; I, \wp) + 4p \times trace(\hat{\Omega}_I \hat{V}_r) \tag{3.16}$$

Remark 3.5.3. If we considered $\hat{\Omega}_I$ in equation (3.14) instead of $\hat{\Omega}_R$ then, $QIC_m(R)$ will take the form

$$\begin{aligned} QIC_{m2}(R) &= -2Q(\hat{\beta}^R|y; I, \wp) + 4p \times trace(\hat{\Omega}_I \hat{V}_r) + \frac{2q}{m(m-1)} * trace(\hat{\Omega}_I \hat{V}_r) \\ &= -2Q(\hat{\beta}^R|y; I, \wp) + \{4p + \frac{2q}{m(m-1)}\} trace(\hat{\Omega}_I \hat{V}_r) \\ &= -2Q(\hat{\beta}^R|y; I, \wp) + 2\{2p + \frac{q}{m(m-1)}\} trace(\hat{\Omega}_I \hat{V}_r) \end{aligned} \tag{3.17}$$

3.5.2 Simulation Study to Compare Performance of $QIC_m(R)$ and QIC in Selecting the True Correlation Structure

We compared the performance of $QIC_m(R)$ to the original QIC using simulation. Specifically, we adopted the same model considered in Pan [60]:

$$Logit(\mu_{it}) = \beta_1 + \beta_2 X_{2it} + \beta_3(t-1) \quad t = 1, 2, 3 \quad i = 1...n \tag{3.18}$$

where $X_{2it} \sim Bernoulli(0.5)$ and $\beta_0 = 0.25 = -\beta_2 = -\beta_3$.
In the simulation, we considered the independence, exchangeable, AR-1 and Toeplitz correlation structures. Exchangeable and AR-1 matrices were parameterized with $\rho = 0.5$ while the toeplitz structure was parameterized with the parameters (0.5, 0.35). In our simulation $n = \{20, 30, 50, 100, 200\}$, m=3 and the number of simulation runs were 1000 just as in Pan [60].

3.5.3 Simulation Results Comparing Performance of $QIC_m(R)$ and QIC

Simulation results for the comparison of QIC and $QIC_m(R)$ in selecting the true correlation structure for $R_0 = \{IN, EX, AR-1, Toep\}$ are shown in Table 3.8. (See Appendix B.2)

Table 3.8: Performance of $QIC_m(R)$ compared to QIC in Selecting the true correlation Structure

		$QIC_m(R)$				QIC			
R_0	n	IN	EX	AR-1	Toep	IN	EX	AR-1	Toep
IN	20	**820**	77	73	30	**108**	197	180	423
	30	**910**	34	33	23	**160**	196	230	414
	50	**977**	7	6	0	**202**	204	170	423
	100	**1000**	0	0	0	**213**	178	170	430
	200	**1000**	0	0	0	**180**	167	187	466
EX(0.5)	20	207	**443**	263	77	211	**278**	104	337
	30	133	**584**	212	70	144	**350**	160	346
	50	87	**740**	157	17	150	**363**	120	367
	100	50	**843**	100	7	128	**312**	130	430
	200	36	**900**	64	0	143	**406**	72	379
AR-1(0.5)	20	245	199	**512**	44	156	231	**320**	293
	30	143	190	**646**	20	140	204	**309**	347
	50	123	181	**692**	4	90	222	**311**	377
	100	70	130	**800**	0	112	183	**315**	390
	200	60	113	**827**	0	90	183	**310**	417
Toep	20	46	403	551	**0**	166	290	260	**284**
	30	150	363	473	**13**	125	300	216	**359**
	50	60	370	570	**0**	130	270	247	**353**
	100	63	363	573	**0**	170	270	170	**390**
	200	40	363	597	**0**	177	160	233	**430**

The simulation results show that when R_0 is independence, $QIC_m(R)$ selects the true structure with success rates of at least 82% and reaches 100% when n=100 while the original QIC selects the independence structure with success rates of at most 22% even with increase in sample size. When R_0 is EX(0.5), $QIC_m(R)$ selects the true structure with success rates of 44.3%, 58.4%, 74%, 84.3% and 90% for respective samples of 20, 30, 50, 100 and 200 respectively. On the other hand, the original QIC selects the true structure with corresponding success rates of 27.8%, 35%, 36.3%, 31.2% and 40.6%. When R_0 is AR-1(0.5), $QIC_m(R)$ selects

the true structure with success rates of 51.2%, 64.6%, 69.2%, 80% and 82.7% for respective samples of 20, 30, 50, 100 and 200 respectively while the original QIC selects the structure with corresponding success rates of 32.0%, 30.9%, 31.1%, 31.5% and 31.0%. When R_0 is Toeplitz (0.5, 0.35), $QIC_m(R)$ does not select the true structure at all but instead prefers the parsimonious AR-1 structure. The results show that $QIC_m(R)$ does not select an over-parameterized structure at all hence meets our objective of developing a criteria that selects the best parsimonious structure. Further, if $R_0 = \{IN, EX, AR-1\}$, its probability of selecting the true structure converge to one in the limit as the number of subjects (n) becomes larger. The results are further illustrated in Figure 3.8.

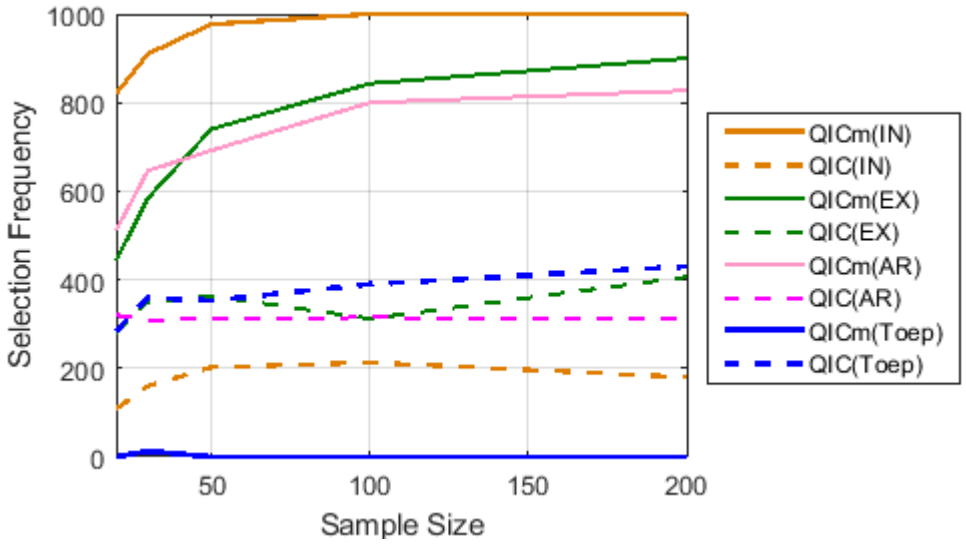

Figure 3.8: Comparison of QIC and $QIC_m(R)$ in the Selection of R_0

The results indicate that $QIC_m(R)$ select a parsimonious working correlation structure that closely approximates the true one by effectively penalizing the working correlation structure with many parameters and also penalizing the working correlation structure which is parsimonious, but with a worse-fit. This explains the low selection rates of the independence correlation structure when $R_0 = \{EX, AR-1, \}$.

Conjecture 3.5.4. *Let $\omega_3 = \{IN, EX, AR-1, Toep\}$ be the set of working correlation structures considered for selection by $QIC_m(R)$. Let R_0 be the true correlation structures and R_* be the correlation structure finally selected by $QIC_m(R)$. Then*

(a)

$$\lim_{n\to\infty} Pr(R_*(n) = R_0) = \begin{cases} 1 & \text{if } R_0 = \{IN, EX, AR-1\} \\ 0 & \text{if } R_0 = Toep \end{cases}$$

i.e. the probability of $QIC_m(R)$ selecting a parsimonious true correlation structure converges to one as $n \to \infty$ hence it is consistent while its probability of selecting an over-parameterized correlation structure converges to zero as $n \to \infty$.

(b) If $R_0 \neq IN$, and $IN \in \omega_3$, then $Pr(R_*(n) = IN) \to 0$ as $n \to \infty$

CHAPTER 4

PROPERTIES OF QIC IN SELECTING COVARIATES FOR THE MEAN STRUCTURE IN GENERALIZED ESTIMATING EQUATIONS

4.1 Introduction

In this chapter, we established theoretically and verified numerically through simulations the properties of QIC in selecting covariates for the mean structure in generalized estimating equations. We examined whether QIC selected the true model asymptotically with a probability of one as $n \longrightarrow \infty$ (consistency), type I error rates (over-fitting properties), type II error rates (under-fitting properties), sensitivity and sparsity properties of QIC.

4.2 Theoretical Results

Let $Q_n(\hat{\beta}_p)$ be the log-quasi-likelihood of the full model with p parameters based on a sample size n, $Q_n(\hat{\beta}_{p_0})$ be the log-quasi-likelihood of model with p_0 correct parameters and $Q_n(\hat{\beta}_{\bar{p}})$ be the log-quasi-likelihood of sub model with $\bar{p}$ of the p parameters in the full model. If $\bar{p} > p_0$, the model with p_0 parameters is nested in the model with $\bar{p}$ parameters so that $Q_n(\hat{\beta}_{p_0})$ is obtained by setting $\bar{p} - p_0$ parameters in the larger model to constants which can be assumed to be zero without loss of generality. If $\hat{p}$ represents the fitted model, then models in which $\hat{p} < p_0$ are mis-specified and the models with $\hat{p} \geq p_0$ are correctly specified or over-specified.

In order to evaluate the probability of selecting the true model (p_0) by QIC, we made the following assumptions:

Assumption A1: All modeling specifications in GEE are correct, such that

$\hat{\Omega}_I \approx \hat{V}_r$ hence $tr(\hat{\Omega}_I \hat{V}_r) \approx p$

Assumption A2: The true model is included in the set of candidate models i.e. $p_0 \in M_c$

Assumption A3: $n \times QIC$ where n is the sample size does not change the ranking of candidate models.

Based on assumption A1 and A3, QIC can be written in the form :

$$QIC_n(p) = \frac{-2Q_n(\hat{\beta}_p)}{n} + p\frac{\psi(n)}{n} \tag{4.1}$$

where $\psi(n) = 2$. Using the general form, the model is selected that corresponds to

$$\hat{p} = \underset{p \in M_c}{\operatorname{argmin}}\{QIC_n(p)\} \tag{4.2}$$

Proposition 4.2.1. *Let $M_c = \{m_1, m_2....m_p\}$ be the set of 2^p candidate models in which the model with p parameters is the largest model. We can partition M_c into two sets: M_+ set of over-specified models i.e. candidate models that include the true model, i.e.$M_+ = \{m_i \in M_c \mid p_0 \subseteq \hat{p}\}$ and $M_- = M_c \setminus (M_+)$, the set of under-specified models i.e. $\hat{p} < p_0$. Under assumptions A1 - A3, if $\hat{p} < p_0$, $\forall \hat{p} \in M_-$, then;*

$$\lim_{n\to\infty} P_r(\hat{p}_{(n)} < p_0) = 0 \tag{4.3}$$

Proof of Proposition (4.2.1). If $\hat{p} < p_0$, then the model with $\hat{p}$ parameters is misspecified, so that

$$\operatorname*{plim}_{n\to\infty} \frac{Q_n(\hat{\beta}_{\hat{p}})}{n} < \operatorname*{plim}_{n\to\infty} \frac{Q_n(\hat{\beta}_{p_0})}{n} \tag{4.4}$$

From equation(4.1), equation (4.4) and $\lim_{n\to\infty} \frac{\psi(n)}{n} = 0$, it follows that;

$$\begin{aligned}
\lim_{n\to\infty} P_r[QIC_n(\hat{p}) \leq QIC_n(p_0)] &= \lim_{n\to\infty} P_r\{\frac{-2Q_n(\hat{\beta}_{\hat{p}})}{n} + \hat{p}\frac{\psi(n)}{n} \\
&\quad \leq \frac{-2Q_n(\hat{\beta}_{p_0})}{n} + p_0\frac{\psi(n)}{n}\} \\
&= \lim_{n\to\infty} P_r\{\frac{Q_n(\hat{\beta}_{p_0}) - Q_n(\hat{\beta}_{\hat{p}})}{n} \\
&\quad \leq 0.5(p_0 - \hat{p})\frac{\psi(n)}{n}\} \\
&= \lim_{n\to\infty} P_r\{\frac{Q_n(\hat{\beta}_{p_0}) - Q_n(\hat{\beta}_{\hat{p}})}{n}\} \leq 0 \\
&= 0 \qquad (4.5)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\lim_{n\to\infty} P_r(\hat{p}_{(n)} < p_0) &= \sum_{\hat{p}<p_0} \lim_{n\to\infty} P_r(QIC_n(\hat{p}) < QIC_n(p_0)), \quad for \quad some \quad \hat{p} < p_0 \\
&= 0 \qquad (4.6)
\end{aligned}$$

□

Proposition 4.2.2. *Under assumptions A1 - A3, if $\hat{p} > p_0$ and QIC selects model with $\hat{p}$ parameters, then*

$$\lim_{n\to\infty} P_r(\hat{p}_{(n)} > p_0) > 0 \qquad (4.7)$$

Proof of Proposition (4.2.2). Since $\hat{p} > p_0$, the model with $\hat{p}$ parameters will be selected if and only if $QIC_n(\hat{p}) < QIC_n(p_0)$ such that;

$$-2Q_n(\hat{\beta}_{\hat{p}}) + 2\hat{p} < -2Q_n(\hat{\beta}_{p_0}) + 2p_0 \qquad (4.8)$$

From equation (4.8), it follows from the likelihood ratio test that;

$$2[(Q_n(\hat{\beta}_{\hat{p}})) - (Q_n(\hat{\beta}_{p_0}))] \xrightarrow{d} T_{\hat{p}-p_0} \sim \chi^2_{\hat{p}-p_0} \qquad (4.9)$$

Therefore, from equations (4.1) and (4.9) we have:

$$n[QIC_n(p_0) - QIC_n(\hat{p})] = -2[(Q_n(\hat{\beta}_{p_0})) - (Q_n(\hat{\beta}_{\hat{p}}))] - 2(\hat{p} - p_0) \xrightarrow{d} T_{\hat{p}-p_0} - 2(\hat{p} - p_0) \qquad (4.10)$$

hence;

$$\lim_{n\to\infty} P_r(QIC_n(\hat{p}) < QIC_n(p_0)) = \lim_{n\to\infty} P_r(T_{\hat{p}-p_0} > 2(\hat{p} - p_0)) > 0 \tag{4.11}$$

From equation (4.11) it follows that:

$$\lim_{n\to\infty} P_r(\hat{p}_{(n)} > p_0) > 0 \tag{4.12}$$

This implies that the over-fitting probability of QIC converges in the limit to a value greater than zero. □

Corollary 4.2.3. *Under assumptions A1 -A3 and* $\forall \hat{p} \in M_+$

$$\lim_{n\to\infty} P_r(\hat{p}_{(n)} \geq p_0) = 1 \tag{4.13}$$

Probability of QIC selecting an over-specified model converge to one in the limit as $n \longrightarrow \infty$.

Proof of Corollary (4.2.3). The probability of selecting an over-specified model is as follows

$$P_r(\hat{p}_{(n)} \geq p_0) = 1 - P_r(\hat{p}_{(n)} < p_0) \tag{4.14}$$

But, from the results in Proposition 4.2.1 it follows that

$$\lim_{n\to\infty} Pr(\hat{p}_{(n)} \geq p_0) = 1 \tag{4.15}$$

□

Corollary 4.2.4. *From the results in Propositions 4.2.2, if we let*

$$\lim_{n\to\infty} P_r(\hat{p}_{(n)} > p_0) = \gamma, \quad \gamma > 0$$

then,

$$\lim_{n\to\infty} P_r(\hat{p}_{(n)} = p_0) = 1 - \gamma \tag{4.16}$$

QIC will not select the true model asymptotically with a probability one since $\gamma > 0$ *hence not consistent. This result combined with the results in Proposition 4.2.1 imply that QIC is conservative as a model selection criteria.*

Corollary 4.2.5. *From propositions 4.2.1 and 4.2.2, sufficient conditions for the consistency of QIC which must be satisfied simultaneously are:*

C1: $\forall \quad \hat{p} \in M_+; \lim_{n\to\infty} P_r(\hat{p}_{(n)} > p_0) = 0$

C2: $\forall \quad \hat{p} \in M_-; \lim_{n\to\infty} P_r(\hat{p}_{(n)} < p_0) = 0$

Proposition 4.2.6. *Further to the results in Proposition 4.2.1 and partitioning β_p the true value of β into truly non-zero and truly zero coefficients as follows:$A = \{j : \beta_j \neq 0, j = 1...q\}$ and $Z = \{j : \beta_j = 0, j = q+1...p\}$ where A denotes the non-zero coefficients and Z denotes the truly zero coefficients. Further, let β_A denote the vector of non-zero coefficients and β_Z denote the vector of truly zero coefficients, then;*

$$Pr\{\exists j \in A : \hat{\beta}_j = 0\} = o(1) \tag{4.17}$$

i.e. active coefficients ($\hat{\beta_A}$) are included in the model selected by QIC with probability approaching one. Conversely, the probability that any of the truly non-zero β_j will be deleted approaches zero as $n\to\infty$ (sensitivity)

Proof of Proposition 4.2.6. We assume the regularity conditions in Theorem 1.6.11, result in Proposition 4.2.1 and that true model coefficients β_{0j} are fixed. In the spirit of Dziak [20], we have;

$$\begin{aligned} Pr_{(n)}\{\exists j \in A : \hat{\beta}_j = 0\} &\leq Pr_{(n)}\{\exists j \in A : |\hat{\beta}_j - \beta_j| > \varepsilon\}, \quad \varepsilon > 0 \\ &\leq Pr\{\|\hat{\beta} - \beta_0\|^2 > \varepsilon^2\} \\ &= 0(1) \end{aligned}$$

□

Corollary 4.2.7. *From Propositions 4.2.2 and 4.2.6, $Pr_{(n)}(\beta_Z = \beta_{0Z}) \nrightarrow 1$ as $n\to\infty$ where β_{0Z} is the vector of true zero coefficients. QIC does not discard all of the true zero coefficients with a probability approaching one hence the model selected by QIC will not be sparse.*

4.3 Numerical Study

In this section, we verify the validity of our claims through simulations. The probabilities of selecting the true model are established and then applied to establish numerically the consistency, over-fitting, under-fitting, sensitivity and sparsity properties of QIC.

4.3.1 Simulation Settings for the Study of the Properties of QIC in Selecting the True Model

The simulation settings were similar to those considered in section 3.2. However, we considered an expanded model with four covariates X_1, X_2, X_3 and X_4. The binary response y_{it} has the conditional expectation μ_{it}:

$$\begin{aligned} \mu_{it} &= E\{y_{it}|X_{1,it}, X_{2,it}, X_{3,it}, X_{4,it}\} \\ &= g^{-1}(\beta_0 + \beta_1 X_{1,it} + \beta_2 X_{2,it} + \beta_3 X_{3,it} + \beta_4 X_{4,it}), \text{i=1..n, t=1..m} \end{aligned} \quad (4.18)$$

such that:

$$logit(\mu_{it}) = \beta_0 + \beta_1 X_{1,it} + \beta_2 X_{2,it} + \beta_3 X_{3,it} + \beta_4 X_{4,it} \quad (4.19)$$

$\{\beta_0, \beta_1, \beta_2\}$=$\{0.25, -0.25, -0.25\}$ and $\beta_p = 0[p \neq 1, 2]$ hence the model with $X_{1,it}$ and $X_{2,it}$ was the true model. This implied that the signals for important covariates were stronger for β_1 and β_2 thus the true model size was 2 and the true number of zero coefficients was 2. $\{X_3, X_4\} \sim U[0, 1]$. The true correlation structure R_0 was assumed to be AR-1(ρ), $\rho = \{0.2, 0.5\}$.

The simulation studies were based on 2^k factorial design. The narrow model included the intercept term and X_1 to be consistent with Pan [61]. The final model was selected from the remaining $2^3 = 8$ candidate models which were;

$$\begin{pmatrix} 1 \\ X_2 \end{pmatrix} \otimes \begin{pmatrix} 1 \\ X_3 \end{pmatrix} \otimes \begin{pmatrix} 1 \\ X_4 \end{pmatrix} = \begin{pmatrix} 1 \\ X_2 \\ X_3 \\ X_2X_3 \\ X_4 \\ X_2X_4 \\ X_3X_4 \\ X_2X_3X_4 \end{pmatrix}. \tag{4.20}$$

Combining the narrow model together with the eight models resulted to the models used in the simulation study as listed in Table 4.1. The number of models considered were more than those considered by Pan [60] who only considered five out of the possible eight models.

Table 4.1: List of Candidate Models for the Simulation Study of the Properties of QIC in Selecting the True Model

Model	Covariates
M1	Int, X_1, X_2, X_3,X_4
M2	Int, X_1, X_2, X_3
M3	Int, X_1, X_2, X_4
M4	int,X_1, X_3,X_4
M5	Int,X_1,X_2
M6	Int,X_1, X_3
M7	Int,X_1, X_4
M8	Int, X_1

Total number of simulation runs used were 1000. As observed by Li [48], at least 500 simulation runs are adequate to reduce the influence of randomness of occurrence on the estimates. Assessment of the performance of the model selection criterion was based on how many times QIC chose the true data generating model (X_1, X_2), type I and II error rates, correct deletion rates and wrong deletion rates. These were then used to establish the consistency, sensitivity and sparsity of QIC in variable selection.

4.3.2 Selection Rates of QIC for the true model: R_0=AR-1 (0.2)

QIC's selection rates of the eight possible models for different sample sizes and measurements per subject are summarized in Table 4.2. (See Appendix C)

Table 4.2: Frequencies of Candidate Models Selection by QIC: AR-1 (0.2)

m	n	M1	M2	M3	M4	M5	M6	M7	M8
	20	95	103	147	25	**308**	62	60	200
	30	60	122	143	23	**423**	32	38	158
3	50	53	133	150	12	**552**	18	25	57
	100	42	123	158	0	**663**	0	0	13
	200	25	147	163	0	**665**	0	0	0
	20	103	170	155	18	**376**	53	32	93
	30	50	170	142	12	**517**	28	21	60
6	50	47	157	142	0	**628**	0	7	20
	100	40	170	148	0	**642**	0	0	0
	200	47	136	137	0	**680**	0	0	0
	20	107	160	198	8	**412**	15	28	72
	30	62	163	193	3	**534**	4	12	27
9	50	40	161	143	0	**647**	0	2	7
	100	48	137	135	0	**680**	0	0	0
	200	32	165	126	0	**677**	0	0	0

The results in Table 4.2 show that when m=3, the rates of identification by QIC of the true model were 30.8%, 42.3%, 55.2%, 66.3% and 66.5% for sample sizes of 20, 30, 50, 100 and 200 respectively. For m=6, the true model selection rates were 37.6%, 51.7%, 62.8%, 64.2% and 68.0% for sample sizes of 20, 30, 50, 100 and 200 respectively. Further, when m increased to 9, the true model selection rates were 41.2%, 53.4%, 64.7%, 68.0% and 67.7% for the respective sample sizes. The study results showed an increasing trend for the true model selection rates which were proportionate to the increase in both the sample size and measurements per subject. The results further showed that, probabilities of selecting the models M1, M2 and M3 which were all overfit were non-diminishing while the probabilities of selecting the models M4, M6, M7 and

M8 which are all under-fit diminished to zero as n became larger hence the conclusion that QIC had higher chances of selecting over-fit models than under-fit models.

4.3.3 Selection Rates of QIC for the true model: R_0=AR-1 (0.5)

The selection rates of QIC for the different models are summarized in Table 4.3. (See Appendix C)

Table 4.3: Frequencies of Candidate Models Selection by QIC with AR-1 (0.5) True Correlation Structure

m	n	M1	M2	M3	M4	M5	M6	M7	M8
	20	86	176	107	31	**360**	47	50	143
	30	57	158	128	6	**463**	33	37	118
3	50	43	162	172	2	**555**	13	26	27
	100	38	155	137	0	**663**	5	0	2
	200	37	128	130	0	**705**	0	0	0
	20	122	168	168	18	**417**	27	22	58
	30	72	160	192	8	**535**	5	11	17
6	50	55	175	180	0	**583**	0	2	5
	100	52	138	125	0	**685**	0	0	0
	200	23	140	118	0	**719**	0	0	0
	20	120	190	176	5	**462**	17	10	22
	30	95	147	172	0	**573**	2	8	3
9	50	52	148	158	0	**640**	0	0	2
	100	33	117	147	0	**703**	0	0	0
	200	29	117	132	0	**721**	0	0	0

The results as shown in Table 4.3 indicate that under the true AR-1(0.5) correlation structure and m=3, the selection rates of the true model were 36%, 46.3%, 55.5%, 66.3% and 70.5% for respective samples of 20, 30, 50, 100 and 200. When measurements per subject were increased to 6, the selection rates of the true model increased to 41.7%, 53.5%, 58.3%, 68.5% and 71.9% for the respective samples and further increased to 46.2%, 57.3%, 64%, 70.3% and 72.1% respectively when measurements per subject were increased to 9. Further, the study results indicate that the probability of QIC selecting

under-fit models definitely converge to zero as n tends to ∞. For example, from Table 4.3, the probabilities of models M6, M7 and M8 being selected which are 0.047, 0.050 and 0.143 respectively when m=3 and n=20, diminish to zero when n is 200. These probabilities remain zero even when m is increased to 6 and 9. This implies that for moderate to large sample sizes, QIC has zero chances of selecting under-fit models. However, the probabilities of QIC selecting over-fit models M1, M2 and M3 are greater than zero in all the simulation settings.

The results show that QIC perform better in choosing more often the correct data generating model and the performance improved as the sample size, number of measurements per subject and as level of correlation increased. These results are presented in Figure 4.1.

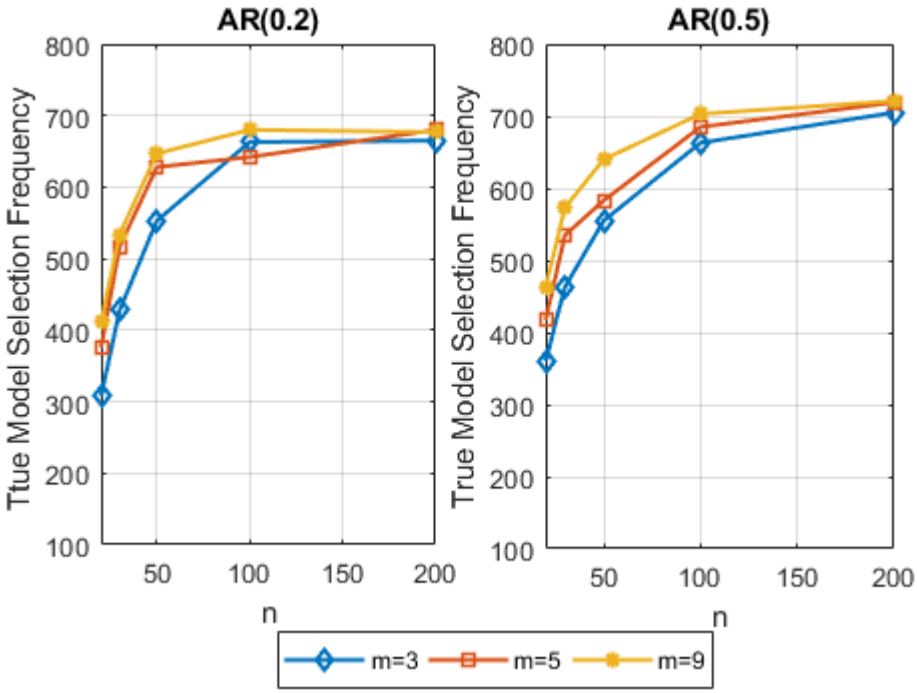

Figure 4.1: True Model Selection Frequencies by QIC ($R_0 = AR - 1$)

Figure 4.1 demonstrates the results discussed above which are in line with findings by Pan [60] who established an increase in the selection rates from 450 when n=50 to 636 when n=100 with AR-1(0.5) true correlation structure.

Moreover, the simulation results of the proportion of selection of over-specified models i.e. models that include the true model ($\hat{p} \geq p_0$) are shown in Table 4.4.

Table 4.4: Proportion of Selection of Models which Include the True Model

ρ	m	20	30	50	100	200
	3	65.3	74.8	88.8	98.7	100.0
0.2	6	80.4	87.9	97.3	100.0	100.0
	9	87.7	95.4	99.1	100.0	100.0
	3	72.9	80.6	93.2	99.3	100.0
0.5	6	87.5	95.9	99.3	100.0	100.0
	9	94.6	98.7	99.8	100.0	100.0

The results in Table 4.4 indicate that regardless of the degree of correlation and measurements per subject, the selection rates of over-specified models i.e. those that contain all the informative covariates plus some spurious ones quickly tended to 100% as n increased. However, the 100% rate was achieved more faster for larger number of measurements per subject and higher degree of correlation within those measurements. The results signify that the probability of QIC selecting an over-specified or correctly specified model converge to one in the limit for sufficiently large samples. These simulation results confirm the theoretical results in Corollary 4.2.3 that $\lim_{n\to\infty} Pr(\hat{p} \geq p_0) = 1$.

4.3.4 Type I and Type II error rates of QIC

Based on the result in Tables 4.2 and 4.3, type I and II error rates were computed. We partitioned our covariates into two sets: the set of influential covariates $X^{(1)} = \{X_{1,it}, X_{2,it}\}$ and the set of non-influential covariates $X^{(2)} = \{X_{3,it}, X_{4,it}\}$ such that $logit(\mu_{it}) = \mathbf{X}^{(1)}\beta^{(1)} + \mathbf{X}^{(2)}\beta^{(2)}$, where $\beta^{(1)} = \{\beta_1, \beta_2\}$ and $\beta^{(2)} = \{\beta_3, \beta_4\}$. The type I error was stated as:

$$Pr(Reject \quad H_0 : \beta^{(2)} = 0 \mid H_0 \quad is \quad true) \tag{4.21a}$$

while the type II error was

$$Pr(Accept \quad H_0 : \beta^{(1)} = 0 \mid H_0 \quad is \quad false) \tag{4.21b}$$

The type I and II error rates were used to determine the specificity and sensitivity of QIC respectively.

The results of the type I error rates of QIC for various combinations of n and ρ are shown in Table 4.5

Table 4.5: Model selection summary by QIC. Type I Error Rate.

		n				
ρ	m	20	30	50	100	200
0.2	3	0.345	0.325	0.336	0.328	0.335
	6	0.428	0.362	0.346	0.358	0.320
	9	0.465	0.418	0.344	0.320	0.328
0.5	3	0.369	0.343	0.377	0.330	0.295
	6	0.458	0.424	0.410	0.315	0.281
	9	0.486	0.414	0.358	0.297	0.278

The results indicate that for both $\rho = 0.2$ and $\rho = 0.5$, m=3, 6 and 9 and for n in the range, $20 \leq n \leq 200$, the range of type I error rate is $\approx 30\%$ to $\lessapprox 50\%$. However, it is noticeable that the type I error rate is higher for smaller sample sizes and decreases with increase in sample size. However, they seem not to approach zero. The Type I error rate of at least 30% indicates fairly high chances of QIC selecting over-fit models by wrongly including covariates whose coefficients are zero. The Type I error rates are also illustrated in Figure 4.2

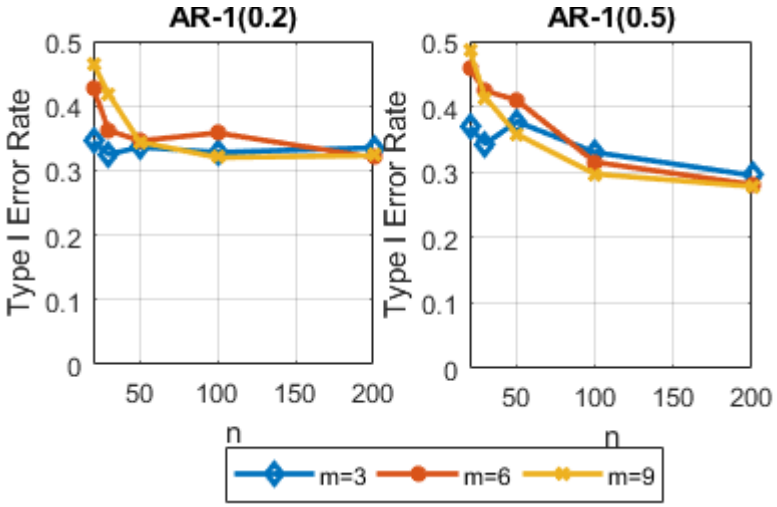

Figure 4.2: Model selection summary by QIC: Type I Error Rate

Figure 4.2 clearly shows the patterns mentioned above in which there is a decline in the

type I error rate as the sample size increases. The Type I error rates are high for small samples and decrease as the sample size increases. This implies that the number of zero coefficients included in the selected model will be significantly reduced if the sample size is made large. However, there seems to be no significant difference in QIC's type I error rate with increase in the level of within subject correlation. The results suggest that for large n, QIC has a non-vanishing chance of choosing complex models since the type I error rate decline slowly as n increases and does not seem to approach zero. The results are in line with findings by Dziak et al. [20] who established non-declining type I error rate for AIC. The high type I error rates implies reduced risk of type II error (Probability of under-fitting) hence increased statistical power of QIC i.e. increased ability to make the correct inclusion of the important variables in the selected model.

To ascertain the type I error rate convergence point, we increased the sample size to 500 and 1000 and the resulting type I error rate curve for AR-1(0.5) is shown in Figure 4.3

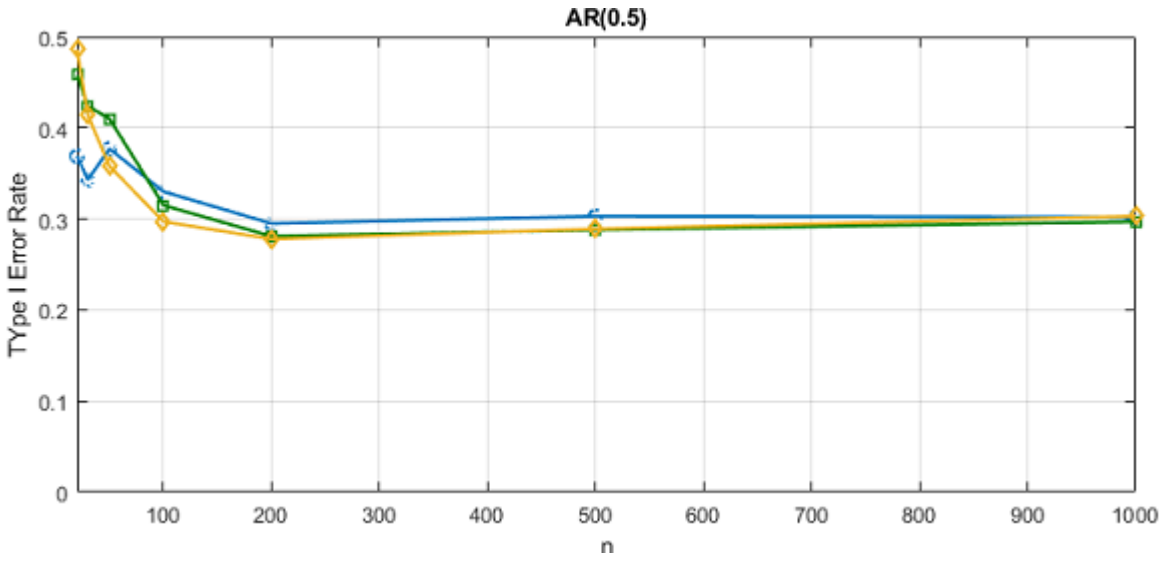

Figure 4.3: Convergence of Type I Error Rate

Figure 4.3 shows that for m=3, the type I error rate stabilizes at 30% while for m=6 and 9, the rate fluctuates around 30% and as the sample size tends to 1000, it tends towards 30% hence we conclude that QIC has a type I error rate of 30%. This confirms the results in Proposition 4.2.2 and Corollary 4.2.4. The value of γ in Corollary 4.2.4

is 0.3 hence $\lim_{n\to\infty} P_r(\hat{p}(n) > p_0) = 0.3$. This implies that QIC as a model selection criteria is approximately 70% specific.

We also established the type II error rates of QIC and its statistical power (1-type I Error). The results are illustrated in Table 4.6.

Table 4.6: Model selection summary by QIC. Type II Error and Statistical Power.

		0.2		0.5	
m	n	type II error (β)	Power of Test	type II error	Power of Test
	20	0.34	0.66	0.27	0.73
	30	0.25	0.75	0.19	0.81
3	50	0.11	0.89	0.07	0.93
	100	0.01	0.99	0.01	0.99
	200	0.00	1.00	0.00	1.00
	20	0.20	0.80	0.13	0.87
	30	0.12	0.88	0.04	0.96
6	50	0.03	0.97	0.01	0.99
	100	0.00	1.00	0.00	1.00
	200	0.00	1.00	0.00	1.00
	20	0.12	0.88	0.05	0.95
	30	0.05	0.95	0.01	0.99
9	50	0.01	0.99	0.00	1.00
	100	0.00	1.00	0.00	1.00
	200	0.00	1.00	0.00	1.00

The results show that the type II error rates were about 30% for small samples, but quickly declined to zero as the sample size increased. This implies that for small samples QIC has some chances of under-fitting at the rate of about 30%. For $n \geq 50$, QIC has little or no chances of under-fitting. This confirms the theoretical results in Proposition 4.2.1 which indicated that for large n, probability of QIC selecting an under-fit model converges to zero in the limit. The power test results show that the power of the test increases with increasing n, so that rejecting any given false null hypothesis is essentially guaranteed for sufficiently large n even if the effect size is small. This makes QIC good in predictive modeling. The result imply that for sufficiently large n, QIC is 100%

sensitive hence confirming the theoretical results in Proposition 4.2.6. The results are also illustrated in Figure 4.4

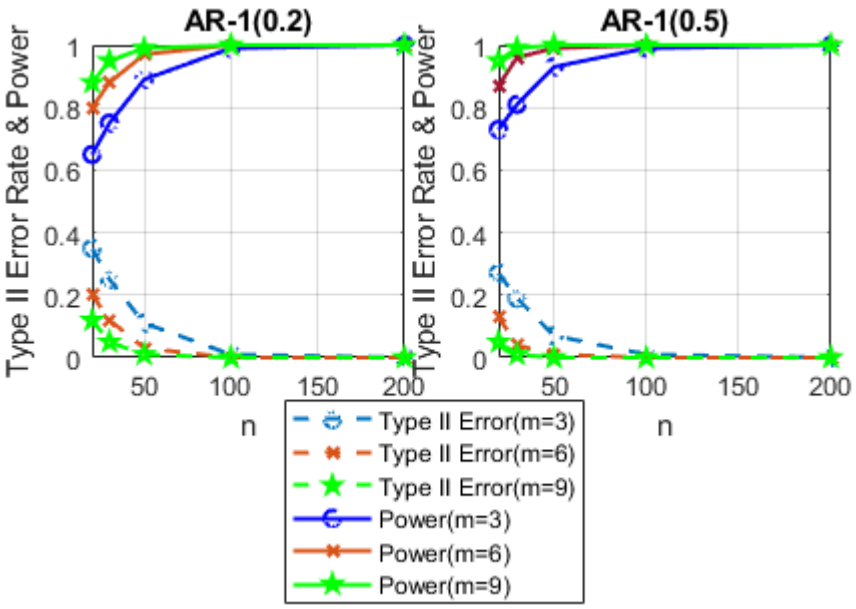

Figure 4.4: Model selection summary by QIC: Type II Error and Statistical Power.

Figure 4.4 shows declining type II error rates for QIC with increase in sample sizes and an increase in its statistical power. The statistical power of QIC ranges from 0.66 at n=20, m=3 and $\rho = 0.2$ to 1 i.e $0.66 \leq Power \leq 1$. The study results also show that high within subject correlation has an effect on the statistical power of QIC. Strong within-subject correlation will result to high statistical power and vice-versa hence, statistical power of QIC=f(n, m, ρ). Overall, the results indicate low chances of QIC making type II error hence high sensitivity. According to Dziak et al.[20], the low type II error rate implies that QIC guards against potential loss of important information.

Therefore, in concurrence with observations by Dziak et al. [20], we concluded that for a strictly predictive model, the high false positive rate of QIC may simply add noise so long as the coefficient estimates are small while a false deletion will create a confounding variable and render the model invalid. Thus, from a statistical estimation framework an over-fit model will be less damaging than an under-fit model. However, in a decision context, the relative seriousness of false exclusion or inclusion depends on the expected practical outcome, e.g., if we are modeling mortality in terms of various

possible carcinogenic pollutants, Type II error will be more hazardous to public health than Type I error.

4.3.5 Sensitivity and Sparsity of QIC in GEE Model selection

We further evaluated the performance of QIC in variable selection using the average number of correct deletion (CD)and the average number of wrong deletion(WD). Correct deletions are the average number (per simulation) of truly zero coefficients correctly estimated as zero, and wrong deletions are the average number of truly non-zero coefficients erroneously set to zero. Because $\beta = \{0.25, -0.25, 0, 0\}$, up to 2 correct deletions and 2 wrong deletions are possible. The results of the analysis are presented in Table 4.7 below:

Table 4.7: Model selection summary by QIC. Average number of correct deletions (CD) and wrong deletions (WD)

		ρ=0.2		ρ=0.5	
m	n	CD	WD	CD	WD
	20	0.62	0.35	0.72	0.27
	30	0.85	0.25	0.93	0.19
3	50	1.10	0.11	1.11	0.07
	100	1.33	0.01	1.33	0.01
	200	1.33	0.00	1.41	0.00
	20	0.75	0.20	0.83	0.13
	30	1.03	0.12	1.07	0.04
6	50	1.26	0.03	1.17	0.01
	100	1.28	0.00	1.37	0.00
	200	1.36	0.00	1.44	0.00
	20	0.82	0.12	0.92	0.05
	30	1.07	0.05	1.15	0.01
9	50	1.29	0.01	1.28	0.00
	100	1.36	0.00	1.41	0.00
	200	1.35	0.00	1.44	0.00

The results of the analysis from Table 4.7 show that when $\rho = 0.2$ and m=3, the correct deletion rates were 0.62, 0.85, 1.10, 1.33 and 1.33 respectively for sample sizes of 20, 30, 50, 100 and 200. When ρ is increased to 0.5, the correct deletion rates are 0.72, 0.93, 1.11, 1.33 and 1.41 respectively for sample sizes of 20, 30, 50, 100 and 200. The results show that increasing the level of within-subject correlation to moderate lead to marginal increase in the average number of coefficients which are set to zero correctly. Likewise, for m=3 and $\rho = 0.2$, the average number of coefficients which are set to zero by mistake remains almost similar for the two levels of correlation and diminish to zero as n increases. When the number of measurements per subject increase to 6, the correct deletion rates were 0.75, 1.03, 1.26, 1.28 and 1.36 for $\rho = 0.2$ and 0.83, 1.07, 1.17, 1.37 and 1.44 for $\rho = 0.5$ respectively for sample sizes of 20, 30, 50, 100 and 200. When m=9, the correct deletion rates are 0.82, 1.07, 1.29, 1.36 and 1.35 for $\rho = 0.2$ while for $\rho = 0.5$ and m=9, the correct deletion rates are 0.92, 1.15, 1.28, 1.41 and 1.44 respectively. The results show that as the number of subjects increase, the average number of coefficients which are set to 0 correctly increases while the average number of coefficients which are set to 0 by mistake diminishes to zero. The study results are similar to results by Shao and Rao [68] who established that the probability of under-fitting of AIC on which the derivation of QIC is based converges to zero as the sample size increases to ∞.

The results are further illustrated in Figure 4.5 which shows an increase in the average number of coefficients correctly set to zero with increase in the number of subjects and a decrease of coefficients set to zero by mistake with increase in sample size.

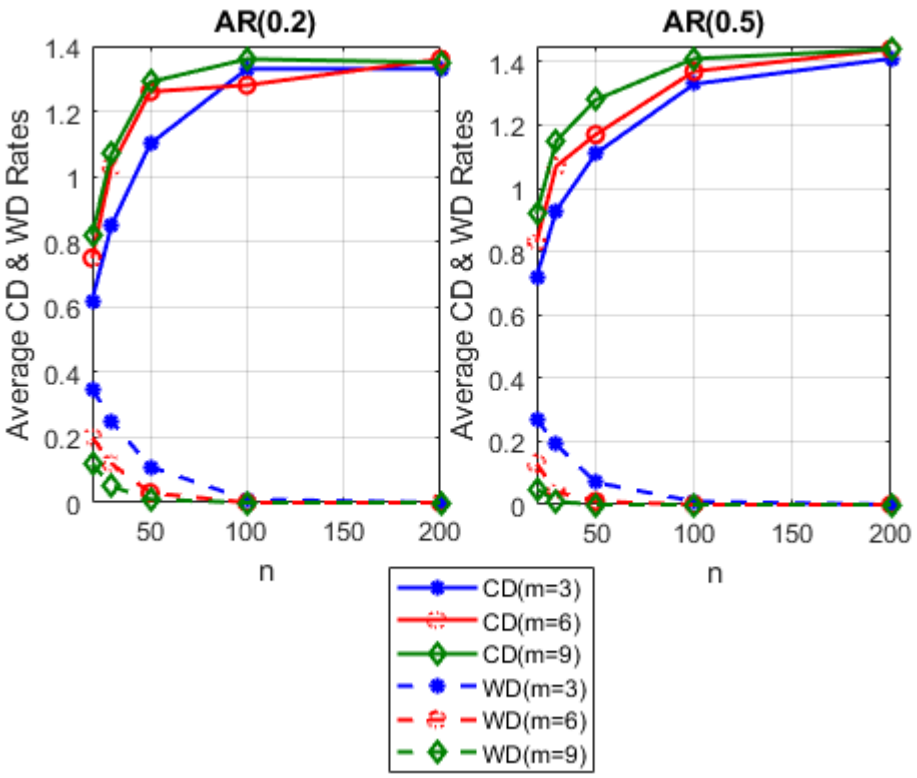

Figure 4.5: Model selection summary by QIC. Average number of coefficients which are set to 0 correctly and average number of coefficients which are set to 0 by mistake

CHAPTER 5

HYBRID METHODOLOGY (EQ_{AIC}) FOR MODEL SELECTION IN GENERALIZED ESTIMATING EQUATIONS AND EFFICIENCY GAIN

5.1 Introduction

In this chapter we considered the problem of efficiency loss in GEE estimates as a result of using a mis-specified structure. Following the results by Pan [60] that QIC is not very powerful in choosing the correct correlation structure among repeated measurements and recommendations by Jang [42] which were reinforced by Oyebayo and Mohd [58] for hybridization of model selection procedures, we sought to determine whether hybridizing QIC with empirical likelihood based AIC improves the efficiency of GEE estimator. We proposed a two-step hybrid methodology (EQ_{AIC}) that involves the use of EAIC (equation (1.120)) for selecting the correct correlation structure and then QIC (equation(1.107)) for selecting covariates with the intention of improving efficiency. We applied K-fold cross-validation to establish the mean squared errors of the model selected by EQ_{AIC} and that selected by QIC. The relative efficiency values were used to determine the efficiency gain of the proposed hybrid methodology compared to when QIC only is used to select both the working correlation structure and covariates.

5.2 Performance of EAIC in Selecting the True Working Correlation Structure

In this section we performed simulation studies to establish the performance EAIC in choosing the true working correlation structure for GEE models compared to QIC and CIC (equation(1.110)). The simulation settings were as follows:

1. The response vector $y_i = (y_{i1}, ..., y_{it})$ was assumed to be a Bernoulli response. i=1,2....n. In the simulation studies $n = \{20, 30, 50, 100, 200\}$; t=1,2,...m. In the simulation study m=3

2. The covariates were x_{1it} and x_{2it} . $x_{1it} \sim N(0,1)$ and $x_{2it} \sim Bernoulli(0.5)$ and a within subject correlation structure dictated by R_0 true correlation structures.

3. The True correlation structures R_0 considered in the simulation were Exchangeable (ρ), $\rho = \{(0.5, 0.8\}$, AR-1 (ρ), $\rho = \{(0.5, 0.8\}$, Independence and Toeplitz correlation matrices.

4. The conditional expectation ($\mu_{ij} = E(y_{it}|x_{1it}, x_{2it})$) of the binary response y_{it} was connected with the covariates through:
   $logit(\mu_{it}) = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it}$; where i =1,...,n and t=1,...,m. The coefficients were set to be ($\beta_0 = 0.25 = -\beta_1 = -\beta_2$) which are similar to the values assumed in Pan [60].

5. We considered the Toeplitz, Independence, Exchangeable and AR-1 correlation structure. This was based on the fact that Independence, Exchangeable and AR-1 working correlation structures can be embedded into the Toeplitz structure. The working correlation structures were partitioned into two sets: $\omega_1 = \{IN, EX, Toep, AR-1\}$ and $\omega_2 = \{IN, EX, AR-1\}$.

6. We parameterized the full model correlation structure by $\rho = (\rho_1, \rho_2)$ and let $\Re^F(\beta, \rho_1, \rho_2)$ define the empirical likelihood of the full model hence we had four sets of GEE estimates one for each of the four working correlation structures which were evaluated by:
   $\Re_{Toep} = R^F(\hat{\beta}^{Toep}, \hat{\rho}_1^{Toep}, \hat{\rho}_2^{Toep})$
   $\Re_{IN} = R^F(\hat{\beta}^{IN}, 0, 0)$
   $\Re_{Exch} = R^F(\hat{\beta}^{EX}, \hat{\rho}^{EX}, \hat{\rho}^{EX})$
   $\Re_{AR} = R^F(\hat{\beta}^{AR}, \hat{\rho}_1^{AR-1}, (\hat{\rho}_1^{AR-1})^2)$

7. The Simulation design was factorial with 1000 simulation replications.

8. All simulations were performed using R version 3.6.0 based on the gee, MASS, Emplik, geepack and bindata R packages. Correlated binary data for the response were generated using the bindata (Touloumis, [75]) library.

5.2.1 Simulation Results for the Performance of EAIC Compared to QIC and CIC in Selecting the True Working Correlation Structure (R_0=AR-1, m=3)

The selection rates of EAIC compared to QIC and CIC in Selecting the true AR-1 correlation structure when m=3 are shown in Table 5.1. (See Appendix D.1)

Table 5.1: Performance of EAIC in Selecting the True working correlation structure Compared to CIC and QIC from 1000 independent replications(R_0:AR-1, m=3)

		EAIC				QIC				CIC			
R_0	n	IN	EX	AR-1	Toep	IN	EX	AR-1	Toep	IN	EX	AR-1	Toep
	20	25	245	**440**	290	160	250	**253**	337	25	260	**301**	414
	30	5	215	**560**	275	145	255	**270**	330	0	145	**455**	400
AR-1	50	0	125	**780**	95	120	210	**320**	350	0	90	**560**	350
(0.5)	100	0	35	**820**	145	75	170	**320**	435	0	10	**645**	345
	200	0	0	**871**	129	100	148	**342**	410	0	0	**687**	313
	20	0	166	**495**	339	155	165	**315**	365	0	85	**466**	449
	30	0	160	**633**	207	140	235	**320**	300	0	95	**500**	405
AR-1	50	0	57	**739**	204	114	210	**297**	379	0	65	**590**	345
(0.8)	100	0	13	**787**	200	125	230	**315**	330	0	15	**625**	360
	200	0	0	**877**	123	103	125	**380**	392	0	10	**700**	290

The results show that when $R_0 = AR - 1$ with $\rho = 0.5$, the success rates of EAIC selecting the true AR-1 structure were 44%, 56%, 78%, 82%, 87.1% for respective sample sizes of 20, 30, 50, 100 and 200. Comparatively, QIC's selection rates in selecting the true correlation structure were 25.5%, 27%, 32%, 32%, 34.2% for respective sample sizes of 20, 30, 50, 100 and 200. The selection rates of CIC in selecting the true correlation structure were 30.1%, 45.5%, 56.7%, 64.5%, 68.7% for respective sample sizes of 20, 30, 50, 100 and 200. The results indicate that the probabilities of EAIC selecting the true AR-1 correlation structure were more than twice those of QIC selecting the structure.

When the degree of correlation is increased to 0.8, the probabilities of EAIC selecting the true correlation structure ($AR-1(0.8)$) were 49.5%, 63.3%, 73.9%, 78.7% and 87.7% for sample sizes of 20, 30, 50, 100 and 200 respectively. For the same settings the

probabilities of QIC selecting the true correlation structure were 31.5%, 32.0%, 29.7%, 31.5% and 38.0% for sample sizes of 20,30, 50, 100 and 200 respectively. CIC's success rates were 46.6%, 50.0%, 59.0%, 62.5% and 70.0% for sample sizes of 20,30, 50, 100 and 200 respectively. The results show that the success rates of EAIC were higher compared to those of QIC and CIC and increased with the degree of correlation. Hence, the higher the within subject correlation, the higher the chances of EAIC selecting the true correlation structure. These are also illustrated in Figure 5.1

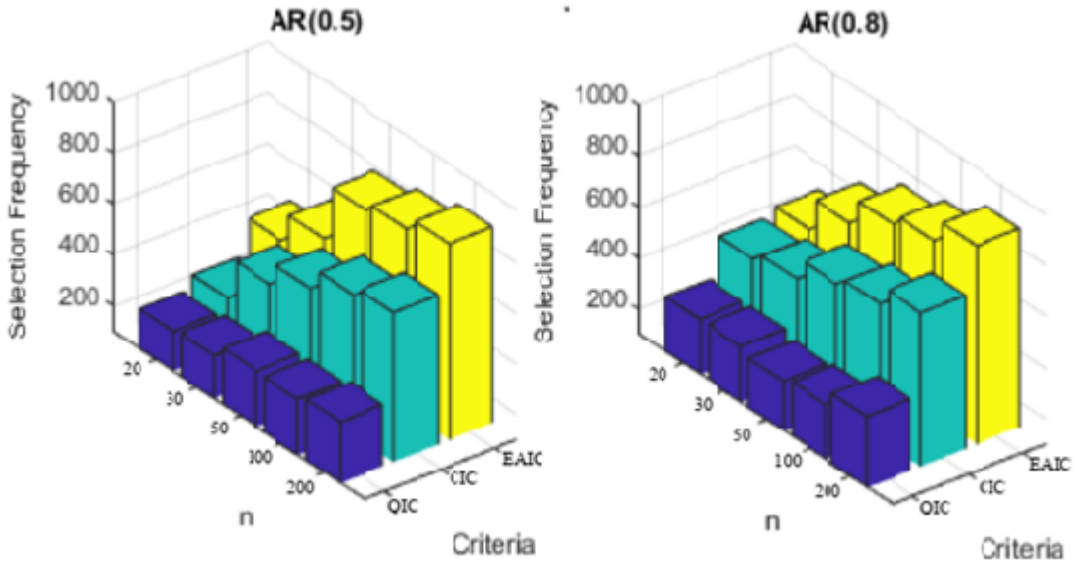

Figure 5.1: True AR-1 working correlation selection frequency by EAIC compared to QIC and CIC

Figure 5.1, show that for all ρ and n used, the performance of EAIC was superior to that of QIC and CIC in choosing the correct AR-1 structure.

5.2.2 Simulation Results for the Performance of EAIC Compared to QIC and CIC in Selecting the True Working Correlation Structure (R_0=EX, m=3)

The simulation results for the performance of EAIC compared to QIC and CIC in selecting the true exchangeable working correlation structure when m=3 are shown in Table 5.2. (See Appendix D.1)

Table 5.2: Performance of EAIC in Selecting the True working correlation structure Compared to CIC and QIC($R_0 = EX$, m=3)

		EAIC				QIC				CIC			
R	n	IN	EX	AR-1	Toep	IN	EX	AR-1	Toep	IN	EX	AR-1	Toep
	20	15	**410**	10	385	210	**300**	175	315	20	**315**	200	465
	30	0	**670**	160	170	195	**330**	125	350	0	545	130	325
EX	50	0	**770**	111	119	120	**370**	91	419	0	**585**	80	335
(0.5)	100	0	**834**	30	136	116	**361**	131	392	0	**632**	52	316
	200	0	**893**	0	107	155	**388**	85	372	0	**720**	10	270
	20	0	**583**	117	300	191	**316**	95	398	62	**340**	137	461
	30	0	**697**	111	192	177	**349**	74	400	18	**488**	102	392
EX	50	0	**789**	81	130	173	**371**	55	401	0	**598**	77	325
(0.8)	100	0	**802**	67	131	122	**365**	83	430	0	**647**	66	287
	200	0	**875**	40	85	110	**398**	41	451	0	**743**	45	212

When the degree of correlation (ρ) was 0.5, the success rates of EAIC selecting the true exchangeable structure were 41%, 67%, 77%, 83.4% and 89.3% for sample sizes of 20, 30, 50, 100 and 200 respectively . Comparatively, QIC's selection rates were 30.0%, 33.0%, 37.0%, 36.1% and 38.8% for the respective samples. Similarly, for the respective samples, the selection rates of CIC of the true correlation structure were 31.5%, 54.5%, 58.5%, 63.2% and 72.0%. The results indicate that EAIC performed better than QIC and CIC in selecting the true exchangeable structure.

When the degree of correlation (ρ) is increased to 0.8, the selection rates of EAIC were 58.3%, 69.7%, 78.9%, 80.2% and 87.5% for the respective samples of 20,30,50, 100 and 200. The selection rates of QIC for the respective samples were 31.6%, 34.9%, 37.1%, 36.5% and 39.8% while those of CIC were 34.0%, 48.8%, 59.8%, 64.7% and 74.3%. The results show that EAIC was powerful in choosing the correct structure in all settings considered and appeared to be higher if R_0=EX than when it was AR-1. On the other hand, QIC favored the correct correlation structure less than 50% of the time hence was less powerful compared to EAIC and CIC. The result further showed that increasing the degree of correlation made the within-subject correlation more easily

recognizable by a model selection criteria hence higher selection rates. These results are also illustrated in Figure 5.2

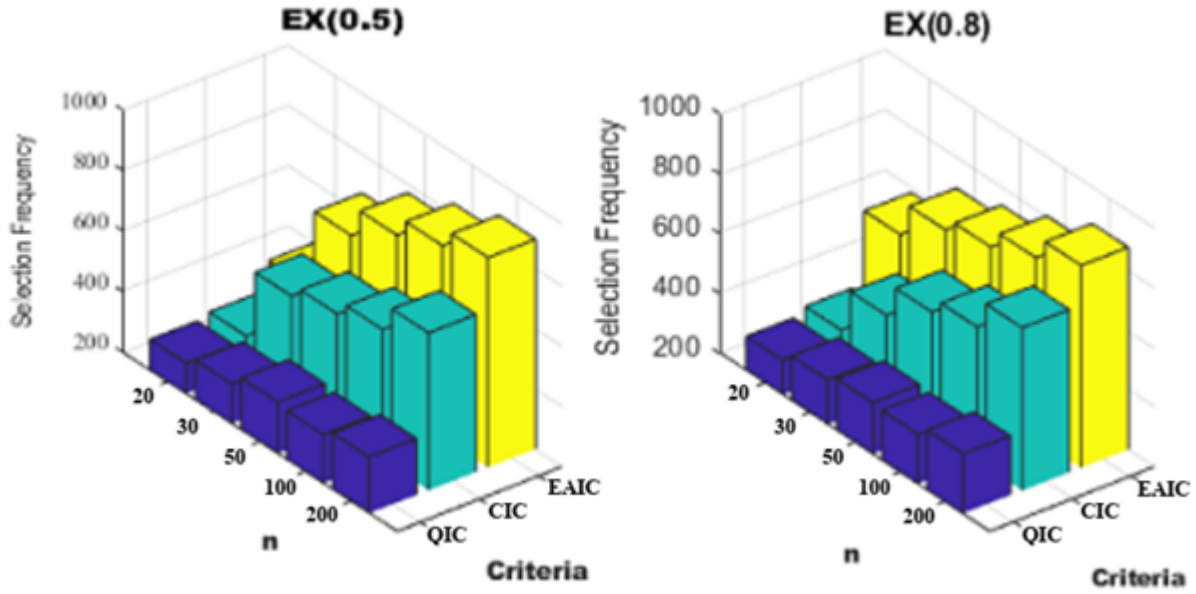

Figure 5.2: True Exchangeable working correlation selection frequency by EAIC compared to QIC and CIC

Figure 5.2 shows that for all ρ and n, EAIC selection rates of the true exchangeable structure were higher than those of QIC and CIC. Similar results were established by Chen and Nicole [12] who established an increasing trend in the selection rates of the true exchangeable structure by EAIC with increase in sample size. They also established superior performance with a stronger degree of correlation.

5.2.3 Simulation Results for the Performance of EAIC Compared to QIC and CIC in Selecting the True Working Correlation Structure (R_0={IN, Toep}, m=3)

The simulation results for the performance of EAIC, QIC and CIC in selecting the independence and Toeplitz working correlation structures are shown in Table 5.3. (See Appendix D.1)

Table 5.3: Performance of EAIC in Selecting the True working correlation structure Compared to CIC and QIC: $R_0 = \{IN, TOEP\}$, m=3

		EAIC				QIC				CIC			
R_0	n	IN	EX	AR-1	Toep	IN	EX	AR-1	Toep	IN	EX	AR-1	Toep
	20	**622**	122	136	120	**192**	218	196	394	**118**	174	166	542
	30	**638**	124	140	98	**192**	184	210	414	**116**	136	154	594
IN	50	**732**	114	106	49	**206**	208	206	380	**118**	160	196	516
	100	**718**	102	114	66	**178**	198	216	408	**114**	168	156	562
	200	**739**	113	123	25	**186**	171	201	442	**0**	277	190	533
	20	15	365	390	**230**	177	258	245	**320**	22	233	257	**488**
	30	7	420	462	**112**	165	288	238	**309**	3	257	252	**488**
Toep	50	0	455	510	**35**	123	308	237	**312**	0	188	260	**552**
	100	0	388	486	**126**	116	283	180	**421**	0	152	230	**617**
	200	0	132	357	**512**	118	248	172	**462**	0	83	152	**765**

The results in Table 5.3 show that when the true correlation structure is independence, EAIC's selection rates were 62.2%, 63.8%, 73.2%, 71.8% and 73.9% for sample sizes of 20, 30, 50, 100 and 200 respectively while those of QIC for the respective samples were 19.2%, 19.2%, 20.6%, 17.8% and 18.6%. The results show that QIC is not powerful at all in choosing the correct independence structure as its selection rates are all less than 20% for the sample sizes considered. In contrast, the probability of EAIC selecting the true independence correlation structure tends to one as n tends to ∞.

When the true working correlation structure was Toeplitz, the frequencies of EAIC choosing the correct structure were 23%, 11.2%, 3.5%, 12.6% and 51.2% for sample sizes of 20,30,50,100 and 200 respectively while those of QIC for the respective samples were 32%, 30.9%, 31.2%, 42.1% and 46.2%. The results show that the selection rates of EAIC for the correct structure drop significantly and EAIC criteria mistakenly prefers parsimonious correlation structures most of the time. Consistency of EAIC starts to set in for $R_0 = Toep$ when the sample size $\geq$200.

5.2.4 Performance of EAIC in Selecting the True correlation structure for ω_2={IN, EX, AR-1}

In this section we considered the selection set $\omega_2 = \{IN, EX, AR-1\}$ which excludes over-parameterized structures and assessed the performance of EAIC in selecting the correct working correlation structure. However, the empirical likelihood ratio was defined with the Toeplitz correlation structure even if it was not considered as a candidate for the working correlation matrix. The true exchangeable and AR-1 structures were parameterized by $\rho = 0.5$ and m=3. We considered $R_0 = \{IN, EX, AR\}$. Simulation results are presented in Table 5.4. (See Appendix D.1)

Table 5.4: Performance of EAIC in selecting the true correlation structure when IN, EX and AR-1 structures are considered

		EAIC			QIC			CIC		
R_0	n	IN	EX	AR-1	IN	EX	AR-1	IN	EX	AR-1
	20	**675**	180	145	**205**	355	440	**135**	375	490
	30	**695**	160	145	**265**	345	390	**145**	420	435
IN	50	**740**	120	140	**190**	450	360	**105**	480	415
	100	**760**	120	105	**235**	320	445	**145**	410	445
	200	**775**	103	122	**265**	350	385	**165**	410	425
	20	50	**710**	240	195	**535**	270	45	**635**	320
	30	0	**850**	150	155	**685**	160	0	**790**	210
EX	50	0	**900**	100	123	**680**	160	0	**845**	155
(0.5)	100	0	**955**	45	180	**745**	45	0	**915**	85
	200	0	**997**	3	125	**790**	85	0	**955**	45
	20	30	265	**705**	200	325	**475**	10	305	**685**
	30	15	240	**745**	140	285	**575**	10	220	**770**
AR-1	50	0	125	**875**	90	285	**625**	0	215	**785**
(0.5)	100	0	55	**945**	120	225	**655**	0	190	**810**
	200	0	0	**1000**	109	211	**680**	0	132	**868**

The results show that when R_0 is independence, the selection rates for EAIC were 67.5%, 69.5%, 74.0%, 76.0% and 77.5% respectively for sample sizes of 20, 30,50, 100 and 200 while those of QIC for the respective samples were 20.5%, 26.5%, 19.0%, 23.5%

and 26.5%. This showed that the probability of EAIC selecting the true independence structure converged to 1 as $n \to \infty$

When R_0 was exchangeable QIC was effective to some extent and performed better with success rates of 53.5%, 68.5%, 68%, 74.5% and 79% respectively for samples of 20, 30, 50,100 and 200. However, EAIC was much better than QIC as indicated by its success rates of 71%, 85%, 90%, 95.5% and 99.7% for sample sizes of 20, 30, 50, 100 and 200 respectively.

When $R_0 = AR-1$, the selection rates of QIC were 47.5%, 57.5%, 62.5%, 65.5% and 68.0% respectively for samples of 20, 30, 50,100 and 200. Comparatively, the selection rates of EAIC were 70.5%, 74.5%, 74.5%, 94.5% and 100% respectively for the respective samples of 20, 30, 50, 100 and 200. It is notable that, if the over-parameterized structures are excluded from the selection set such that $\omega_2 = \{IN, EX, AR-1\}$ and $R_0 \in \omega_2$, EAIC invariably chooses the correct structure and its consistency is achieved at n=200.

5.3 Hybrid Methodology (EQ_{AIC}) and Efficiency Improvement in Generalized Estimating Equations

In the GEE context, an inappropriate working correlation structure significantly impairs the efficiency of GEE estimator for β. A more plausible way to improve this is by using the working correlation structure most appropriate for the data at hand. In this section we sought to establish whether correct identification of the true working correlation structure improved efficiency of GEE estimator. We use an hybrid methodology that involved the use of EAIC to select the true correlation structure and QIC to select the set of informative covariates. This was implemented through an example dataset.

Example 5.3.1. We illustrated gain in efficiency by EQ_{AIC} over QIC by analyzing the Ohio Dataset contained in the geepack library. The Ohio Dataset is based on a study that analyzed the health effect of indoor and outdoor air pollution on children's wheeze status as determined by the age and smoking status of mothers. The data set was analysed by many authors (Qu et al.[63], Fitzmaurice et al.[25]) and is based on 537 children who were followed for four years and data on their age, smoking status of

mothers and wheeze status were recorded resulting to 2148 observations. The maternal smoking habit was treated as fixed at the first visit. The response is a binary outcome with 1 indicating the presence of the respiratory illness 0 and its absence . The maternal smoking habit, in the preceding year, was recorded as a binary covariate.

Assuming that the measurements for same child are serially correlated, we used the logit link function hence our logit model was given as:

$$log(\frac{\mu_{it}}{1-\mu_{it}}) = \beta_0 + \beta_1 X_{it1} + \beta_2 X_{it2} + \beta_3 X_{it1} X_{it2} \tag{5.1}$$

for 1=1....537, t=1....4. Where X_{it1}, X_{it2} and $X_{it1}X_{it2}$ are the age of child, smoking habit indicator and their interaction respectively. and the marginal variance took the form;

$$\begin{aligned} var(Y_{it}) &= \phi v(\mu_{it}) \\ &= \mu_{it}(1-\mu_{it}) \end{aligned} \tag{5.2}$$

Hence the matrix A_i is a diagonal matrix with elements $v(\mu_{it}) = \mu_{it}(1-\mu_{it})$

Empirical likelihood ratio was defined with the general correlation structure Toeplitz. EAIC and QIC were obtained with each of the four sets of GEE estimates based on the four correlation structures considered. We used the results to determine the correlation structure preferred by EAIC and that preferred by QIC for the Ohio dataset. The results are shown in Table 5.5.(See Appendix D.2)

Table 5.5: Working Correlation Structure Selection for the Wheeze Status GEE Model Using the Ohio Dataset

	Working Correlation Structures			
	IN	EX	AR-1	Toep
EAIC	656.9704	361.3486	454.8900	363.1113
QIC	3154.627	3216.383	3201.844	3215.749

The results show that EAIC preferred the exchangeable structure which has a minimum value of 361.3486, meaning that neither Toeplitz nor AR-1 were sufficient to

describe the correlation structure. In contrast, QIC chose the independence structure with a minimum QIC value of 3154.627.

Based on the set of selected working correlation structure, we applied QIC to select the best model out of the 8 possible models. The eight possible models were:

$$\begin{pmatrix} 1 \\ \mathrm{X}_{it1} \end{pmatrix} \otimes \begin{pmatrix} 1 \\ \mathrm{X}_{it2} \end{pmatrix} \otimes \begin{pmatrix} 1 \\ \mathrm{X}_{it1}X_{it2} \end{pmatrix} = \begin{pmatrix} 1 \\ \mathrm{X}_{it1} \\ \mathrm{X}_{it2} \\ \mathrm{X}_{it1}X_{it2} \\ \mathrm{X}_{it1},\ X_{it2} \\ \mathrm{X}_{it1},\ X_{it1}X_{it2} \\ \mathrm{X}_{it2},\ X_{it1}X_{it2} \\ \mathrm{X}_{it1},\ X_{it2},\ X_{it1}X_{it2} \end{pmatrix}.$$

Based on the exchangeable correlation structure selected by EAIC, the models were ranked by QIC as shown in Table 5.6.

Table 5.6: QIC Model Ranking for the Ohio Data based on the Exchangeable Correlation Structure

MODEL RANK	Intercept	X_{it1}	X_{it2}	$X_{it1}X_{it2}$	qlik	QIC
M5	-1.880	-0.1134	0.2651		-909-95	1829.5
M2	-1.783	-0.1131			-912.34	1830.1
M8	-1.900	-0.1412	0.3138	0.07083	-909.74	1830.4
M6	-1.783	-0.1214		0.02292	-912.47	1831.5
M4	-1.739			-0.09235	-913.45	1832.3
M3	-1.821		0.2761		-912.15	1832.6
M7	-1.821		0.2346	0.0704	-911.82	1833.1
M1	-1.721				-914.54	1833.2

The results as presented in Table 5.6 shows that the model with covariates X_{it1}(age) and X_{it2}(smoking status) was the choice by QIC as the best model. The model including only X_{it1} (age) was ranked second and the one including X_{it1} (age), X_{it2} (smoking status) and X_{it3} (interaction between age and smoking status) was ranked third. The intercept

only model was ranked last by QIC.

When QIC is used to select both the working correlation structure and the covariates to include in the model, the ranking of the possible models and the estimates of the corresponding coefficients are given in Table 5.7

Table 5.7: QIC Model Ranking for the Ohio Data based on the Independence Correlation Structure

MODEL RANK	Intercept	X_{it1}	X_{it2}	$X_{it1}X_{it2}$	qlik	QIC
M5	-1.884	-0.1134	0.2721		-909.95	1829.5
M2	-1.783	-0.1132			-912.34	1830.1
M8	-1.901	-0.1413	0.3140	0.07084	-909.74	1830.4
M6	-1.783	-0.0969		-0.04619	-912.27	1831.8
M4	-1.746			-0.1290	-913.35	1832.5
M3	-1.821		0.2716		-912.15	1832.6
M7	-1.821		0.2343	-0.0704	-911.82	1833.1
M1	-1.721				-914.54	1833.2

The results from Table 5.7 show a similar ranking to the rankings in Table 5.6. In both cases, Model 5 was ranked top. However, examination of the coefficients shows that there were differences in their magnitudes for the two models. This implies that using different correlation structures will result to different effect sizes for the variables. For instance a unit increase in smoking status of mothers caused an estimated increase in te odds of developing respiratory illness by a factor of $1.303(e^{0.2651})$ under EQ_{AIC} and by a factor of $1.313(e^{0.2721})$ under QIC.

To evaluate whether the use of the true correlation structure led to efficiency gain, we carried out a K-fold cross-validation in which the predictive ability of each model based on the Ohio data set is estimated using cross-validation. According to Cavanaugh [10], EQ_{AIC} procedure will be regarded asymptotically efficient if it asymptotically identifies a fitted candidate model with minimum mean squared error. The following algorithm was used to compute the MSE for the models was as given below:

(i) We randomly divided the data set into K parts say $k_1, ..., k_K$ each with approxi-

mately $\frac{n}{K}$ clusters and $\frac{mn}{K}$ data points.

(ii) With the k^{th} part held as the validation set, we fit the method on the remaining K-1 folds

(iii) We then computed the mean squared error, MSE_i using observations in the held-out fold

$$MSE_K = \frac{\sum_{i \in k}(y_i - \hat{y}_i)^2}{K} \tag{5.3}$$

Where $\hat{y}_i$ is the fit for observation i obtained from the data with part K removed.

(iv) The process was repeated K times, each time, a different group observations being treated as a validation set. The process result in K estimates of the test error, $MSE_1, ..., MSE_K$

(v) The K-fold CV error was computed by averaging MSE_i, i=1...K

$$CV_{(k)} = \frac{1}{K}\sum_{i=1}^{k} MSE_i \tag{5.4}$$

(vi) The procedure was repeated N times to reduce the influence of randomness associated with the K-fold cross validation (Li [48]) such that final CV error is:

$$CV_{(k)} = \frac{1}{N}\sum_{i=1}^{N}\{\frac{1}{K}\sum_{i=1}^{k} MSE_i\} \tag{5.5}$$

(vii) We finally computed the relative efficiency measure:

$$RE = \frac{MSE(\hat{\beta}_G^{QIC})}{MSE(\hat{\beta}_G^{EQ_{AIC}})} \tag{5.6}$$

If $RE > 1$, GEE estimates under EQ_{AIC} will be more efficient than GEE estimates under QIC. If $RE < 1$ then GEE estimates under QIC will be more efficient than GEE estimates under EQ_{AIC} and; if RE=1, then EQ_{AIC} and QIC gives the same results.

In our analysis, N=200 and $K = \{3, 5, 10\}$. The results are tabulated in Table 5.8.

Table 5.8: Efficiency Gain of EQ_{AIC} over QIC

K	MSE($\hat{\beta}_G^{EQ_{AIC}}$)	MSE($\hat{\beta}_G^{QIC}$)	RE
3	1.74628	2.32801	1.2113
5	1.73519	2.16066	1.2452
10	1.74860	2.18400	1.2490

Comparison of the MSE($\hat{\beta}_G^{EQ_{AIC}}$) and MSE($\hat{\beta}_G^{QIC}$) indicates that MSE ($\hat{\beta}_G^{EQ_{AIC}}$) < MSE($\hat{\beta}_G^{QIC}$) with MSE($\hat{\beta}_G^{QIC}$) being at least 21% more than the MSE of $\hat{\beta}_G^{EQ_{AIC}}$. The results also established relative efficiency values greater than 1 for all K hence based on Qu et al. [63], $\hat{\beta}_G^{EQ_{AIC}}$ is more efficient than $\hat{\beta}_G^{QIC}$ hence using the correct correlation structure helps achieve the goal of efficiency improvement. The gain in efficiency established in this study is 2% more than the efficiency established by Jamshid et al. [41] using QIF. The study results corroborate assertions by Chen and Nicole [12] that selecting the true correlation structure improves efficiency of GEE estimates. The results also indicate that hybridization of model selection procedures in GEE as proposed by Jang [42] and Oyebayo and Mohd [58] improves efficiency.

CHAPTER 6

APPLICATION OF THE HYBRID METHODOLOGY (EQ_{AIC}) TO MODEL THE DETERMINANTS OF SHAREHOLDER VALUE CREATION

6.1 Introduction

In this chapter, we applied the proposed hybrid methodology that involves the use of EAIC to choose an appropriate correlation structure and QIC to select the set of covariates to the shareholder value creation data. A study was designed in which a retrospective longitudinal study design was used to collect secondary quantitative data for public listed firms in the NSE. The data were obtained from the annual financial statements of the firms covering a period of 6 years (2011-2016). The sampling frame of the study was 61 firms registered in Nairobi Security Exchange(NSE) comprising 7 agricultural firms; 4 auto-mobile and accessories firms; 11 banking firms; 9 commercial services firms; 5 construction and allied firms; 4 energy and petroleum firms; 6 insurance firms; 5 investment firms; 9 manufacturing firms and 1 telecommunication firm. These were the listed companies whose shares were trading as at January 2011. The sample size of 53 firms was determined using the Cochran [14] formula as shown below:

$$\begin{aligned} n &= \frac{\sum_{i=1}^{L} \frac{N_i^2 p(1-p)}{w_i}}{\frac{N^2 d^2}{Z_{\frac{\alpha}{2}}^2} + Np(1-p)} \\ &= \frac{\frac{7^2 \times 0.5 \times 0.5}{0.11667} + ... + \frac{1^2 \times 0.5 \times 0.5}{0.016667}}{17.34277} \\ &= 53 \end{aligned} \tag{6.1}$$

Where, n= the desired sample size, Z=the standard normal deviate at the required confidence level (1.96), p=the proportion of the target population estimated to be having the characteristic being measured (0.5), d= the absolute precision defined as $d = Z_{\frac{\alpha}{2}} SE$, where SE is the standard error, α= level of statistical significance (0.05), N_i is the number of firms in each stratum i = 1.....L such that $N = N_1 + N_2 + N_3 ..... N_L$. $w_i = \frac{Ni}{N}$

is the stratum weight and L=Number of strata.

Proportionate stratified sampling method was used to select the representative sample of 53 firms based on their categorization. The sampling technique was preferred because if we let V_{Prop} represent the variance under stratified random sampling and V_{Rand} represent variance under simple random sampling and assume that variation between strata is more than variation within strata, then $V_{prop} \leq V_{rand}$ (Cochran,[14]). The sample comprised 8 agricultural firms; 3 auto-mobile and accessories firms; 10 banking firms; 8 commercial services firms; 4 construction and allied firms; 3 energy and petroleum firms; 5 insurance firms; 3 investment firms; 8 manufacturing firms and 1 telecommunication firm.

6.2 Basic Model for the Determinants of Shareholder Value Creation

The basic model that is used to make predictions on the determinants of shareholder value creation is the constant-growth model (Gordon [28]). The model predicts that changes in shareholder value creation depends on dividends and the discount rate:

$$\begin{aligned} MV &= \frac{DPS}{k_e - g} \\ &= \frac{EPS(1-b)}{k_e - g} \end{aligned} \tag{6.2}$$

The market value (MV) is the present value of the expected stream of Dividends Per Share (DPS) which is equal to the Earnings Per Share (EPS) multiplied by the pay-out ratio (1-b). This implies that DPS depends on the firm's payout ratio and the earnings growth per year (g) which depend on the retention ratio (b) and on the Return on Equity (ROE) $[ROE \times b = g]$. k_e is the cost of equity capital computed using the dividend capitalization formula i.e.

$$k_e = \frac{D_0(1+g_d)}{P_0} + g_d, \quad g_d = \frac{1}{m}\sum_{t=1}^{m}\{\frac{D_t}{D_{t-1}} - 1\} \tag{6.3}$$

where;

D_0=Current year's DPS, P_0=Current year's share price, g_d=growth rate of dividend per year, D_t= DPS at year t.

Model (6.2) assumes that dividends grow at a constant rate in perpetuity. EPS depends on the firm's ROE and the equity investment normally expressed as book value per equity share (BV) such that $[EPS = ROE \times BV]$. Equation 6.2 can therefore be re-written as:

$$\begin{aligned} MV &= \frac{BV \times ROE(1-b)}{k_e - g} \\ &= \frac{BV(ROE - b * ROE}{k_e - g} \end{aligned} \tag{6.4a}$$

Thus

$$\frac{MV}{BV} = \frac{ROE - g}{k_e - g} \tag{6.4b}$$

Equation 6.4b implies that shareholder value will be created if $\frac{MV}{BV} > 1$ and value will be destroyed if $\frac{MV}{BV} < 1$. Further, equation 6.4b indicates that economic profitability and growth are the main determinants of $\frac{MV}{BV}$ hence shareholder value creation depends on the economic spread and the volume of future investment opportunities (g) and the generic representation of the value based model is;

$$\frac{MV}{BV} = \beta_0 + \beta_1(ROE - k_e)_{it} + \beta_2 g_{it} + \beta_3(ROE - k_e)_{it} * g_{it} \tag{6.5}$$

Equation 6.5 implies that the level of economic profitability $(ROE - k_e)$ to be earned, the volume of future investment opportunities expressed as the growth of earnings per year (g) and the interaction term are the main drivers of shareholder value creation.

However, authors have tried to capture the determinants of shareholder value creation based on theoretical hypotheses on firm value. Copeland and Weston [15] tested the relationship between dividend policy and firm value. Modgiliani and Miller (MM) showed that in a world without taxes, agency costs or information asymmetry debt policy had no effect on the value of a firm. Rappaport [64] argued that accounting profitability is a very important value driver whereas Pandey [62] had the opinion that it was economic profitability that had an effect on value creation. As such the main determinants of shareholder value creation should include variables that capture the dividend policy, debt relevance policy and profitability hypothesis. Likewise, other scholars have identified board size, working capital policy, level of financial distress and the size

of the firm as determinants of shareholder value creation. The expanded model could therefore be expressed as ;

$$\begin{aligned}\frac{MV}{BV} &= \beta_0 + \beta_1(ROE - k_e)_{it} + \beta_2 g_{it} + \beta_3(ROE - k_e)_{it} * g_{it} \\ &+\beta_4 ROA_{it} + \beta_5 DPR_{it} + \beta_6 lev_{it} + \beta_7 logta_{it} + \beta_8 wcta_{it} \\ &+\beta_9 Z_{it} + \beta_{10} bsize_{it} + \epsilon_{it} \end{aligned} \tag{6.6}$$

where $\frac{MV}{BV}$ is the market value to book value ratio; $(ROE - k_e)$ is the economic profitability, ROA is the Return On Assets that represents accounting profitability, 'g' is the rate of growth of earnings per year, DPR is the dividend pay-out ratio that represents the firm's dividend policy, lev represents the the firms' financial policy, logta is the logarithm of the total assets which represents the firm size, wcta is the net working capital to to total assets which represents the working capital policy of the firms, 'Z' is the Altman's Z-score value representing the level of financial distress and bsize represents the board size.

6.3 Determinants of Shareholder Value Creation

6.3.1 The GEE model

Suppose each firm i is a cluster such that Y_{it} is the $\frac{MV}{BV}$ history $\{(0,1\}$ such that

$$Y_{it} = \begin{cases} 1 & \text{if } \frac{MV}{BV} > 1 \\ 0 & \text{if } \frac{MV}{BV} \leq 1 \end{cases}$$

The observed time t corresponds to the time values at which the $\frac{MV}{BV}$ history is measured (t=1.....6). Let $Y_i = [Y_{i1}....Y_{i6}]^T$ be the 6×1 random vector of $\frac{MV}{BV}$ history of the i^{th} firm for the 6 years studied. The covariates X_{1it}, X_{2it}, X_{3it}, X_{4it}, X_{5it}, X_{6it}, X_{7it}, X_{8it}, X_{9it} and X_{10it} are associated with the i^{th} firm and t^{th} year. Further, Suppose P_{it} is the probability that the i^{th} firm's $\frac{MV}{BV} > 1$ during the t^{th} financial year , then

$$\begin{aligned} P_{it} = P_r(Y_{it} = 1 \mid \mathbf{X}_{jit}) &= g(\beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \beta_3 X_{3it} + \beta_4 X_{4it} + \beta_5 X_{5it} \\ &+\beta_6 X_{6it} + \beta_7 X_{7it} + \beta_8 X_{8it} + \beta_9 X_{9it} + \beta_{10} X_{10it}) \end{aligned} \tag{6.7}$$

such that

$$logit(P_r(Y_{it}=1 \mid \mathbf{X}_{jit}) = \beta_0+\beta_1X_{1it}+\beta_2X_{2it}+\beta_3X_{3it}+\beta_4X_{4it}+\beta_5X_{5it} +\beta_6X_{6it}+\beta_7X_{7it}+\beta_8X_{8it}+\beta_9X_{9it}+\beta_{10}X_{10it} \quad (6.8)$$

For i=1,2.....53 and t=1,2...6

$g(u)=\frac{e^u}{1+e^u}$ is the logistic function and the logit transform yields the equation

$$log(\frac{\mu_{it}}{1-\mu_{it}}) = \beta_0+\beta_1X_{1it}+\beta_2X_{2it}+\beta_3X_{3it}+\beta_4X_{4it}+\beta_5X_{5it}+\beta_6X_{6it} +\beta_7X_{7it}+\beta_8X_{8it}+\beta_9X_{9it}+\beta_{10}X_{10it} \quad (6.9)$$

where,

- X_1=Economic profitability which is the difference between Return On Equity ($ROE=\frac{Netprofit}{BVofEquity}$) and Cost of Equity.
- X_2=growth (g) measured by the growth rate of earnings per year.
- X_3=interaction effect between growth in earnings and economic profitability,
- X_4=Firm Size measured using the logarithm of the firm's total assets. Firm size has been confirmed as a significant predictor of firm performance. According to Honjo and Harada [37], a small firm is more likely to register lesser $\frac{MV}{BV}$ because of inadequate experience in the market, limited connections and limited financial distress
- X_5= Leverage which measures the long-term solvency of a company. The analysis of financial leverage is concerned with the capital structure of the firm. Leverage will be measured by the debt ratio which compares a company's total debt to its total assets. The lower the percentage means that the company is dependent on leverage hence the stronger its equity position. $Lev=\frac{T_D}{T_A}$ where T_D=Total debt and T_A=Total assets
- X_6=Dividend Policy measured by the dividend payout Ratio which is the amount of dividends paid to stockholders relative to the amount of total net income. It shows the portion of the profits the firm decides to keep to fund operations. $DPR=\frac{D}{N_I}$, Where DPR = Dividend Payout Ratio, D = Total Dividends and N_I=Net Income

- X_7=Likelihood of financial distress of the firms measured by the Altman Z-score: $Z = 6.56T_1 + 3.26T_2 + 6.72T_3 + 1.05T_4$ Where;
  $T_1 = \frac{W_C}{T_A}$; $T_2 = \frac{R_E}{T_A}$; $T_3 = \frac{EBIT}{T_A}$ and $T_4 = \frac{MV_E}{TV_L}$
  Where, W_C=Working Capital; R_E=Retained Earnings; EBIT=earnings before interest and tax; MV_E=market value of equity; TV_L=Total value of Liabilities and TA=Total=Assets.
- X_8= Board size
- X_9= Accounting profitability measured by:$ROA = \frac{E_{AT}}{T_A}$, where ROA=Return On Assets, E_{AT}=Earnings After Tax and T_ATotal Assets. The profitability Ratio measures the ability of a company to generate earnings.
- X_{10}=working capital policy which is measured by $\frac{CA}{CL}$, where CA=current assets and CL=current liabilities

The mean vector of Y_i is

$$\mu_i = \begin{pmatrix} E(Y_{i1}) \\ E(Y_{i2}) \\ \vdots \\ E(Y_{i6}) \end{pmatrix} = \begin{pmatrix} P_{i1} \\ P_{i2} \\ \vdots \\ P_{i6} \end{pmatrix} = P_i \tag{6.10}$$

Where $P_{it} = \mu_{it} = P_r(Y_{it} = 1 \mid X_{it})$, t=1,2..6 and i=1....53. The probability of not creating value for firm i in the financial year t is $(1 - P_{it})$ and the variance of Y_{it} is

$$Var(Y_{it} = 1 \mid X_{it}) = \mu_{it}(1 - \mu_{it}) = \frac{e^{u_{it}}}{(1 + e^{u_{it}})^2}. \tag{6.11}$$

The 6×6 variance-covariance matrix of Y_i is given by

$$Var(Y_i) = \begin{pmatrix} Var(Y_{i1}) & Cov(Y_{i1}, Y_{i2}) & \dots & Cov(Y_{i1}, Y_{i6}) \\ Cov(Y_{i1}, Y_{i2}) & Var(Y_{i2}) & \dots & Cov(Y_{i2}, Y_{i6}) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(Y_{i1}, Y_{i6}) & Cov(Y_{i2}, Y_{i6}) & \dots & Var(Y_{i6}) \end{pmatrix} \tag{6.12}$$

In addition to the mean and covariance of the vector of responses Y_{it}, Liang and Zeger [83] suggested the use of an $m_i \times m_i$ working correlation matrix for Y_i, $R(\rho)$, assumed

to be fully specified by ρ such that;

$$V_i = A^{\frac{1}{2}} R_i(\rho) A^{\frac{1}{2}} \tag{6.13}$$

Where, $A_i = diag[Var(Y_{i1}), Var(Y_{i2}), ..., Var(Y_{im})]$ is a diagonal matrix which can be expressed as:

$$A_i = \begin{pmatrix} Var(Y_{i1}) & 0 & \dots & 0 \\ 0) & Var(Y_{i2}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Var(Y_{i6}) \end{pmatrix} \tag{6.14}$$

If we let $D_i = \frac{d\mu_i}{d\beta^T}$ and if $X_i's$ are observable covariates for each firm, the vector of parameters $\beta = [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}]^T$ for model 6.8 could be obtained by solving iteratively the following generalized estimating equations

$$U(\beta) = \sum_{i=1}^{53} D_i^T V_i^{-1} (Y_i - \mu_i) = 0 \tag{6.15}$$

To solve equation 6.15, we consider the mean vector $\mu_i = [\mu_{i1} ... \mu_{i6}]$ and the variance-covariance matrix V_i which varies depending on the nature of the correlation structure $R_i(\rho)$, where $(Y_i - \mu_i)$ is a residual vector which measures deviations of observed responses of the i^{th} firm from its mean. Liang and Zeger [49] established that $\hat{\beta}$ satisfies U(β)=0 hence is asymptotically unbiased in the sense that $\lim_{n\to\infty}(E[U(\beta)]) = 0$. Moreover, $\sqrt{n}(\hat{\beta}_G - \beta)$ is asymptotically multivariate normal.

6.3.2 Model Selection Procedure

EQ_{AIC} model selection procedures was applied to select the best model from the 2^{10} candidate set of models which are illustrated as:

$$\begin{pmatrix} 1 \\ X_1 \end{pmatrix} \otimes \begin{pmatrix} 1 \\ X_2 \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} 1 \\ X_{10} \end{pmatrix} = \begin{pmatrix} 1 \\ X_1 \\ X_2 \\ X_1X_2 \\ X_1X_2X_3 \\ . \\ . \\ X_1X_2X_3X_4X_5X_6X_7X_8X_9X_{10} \end{pmatrix}. \quad (6.16)$$

1. First, using the full model, we fixed the mean structure and computed the EAIC of the models under different covariance structures. The covariance structure that yielded the lowest EAIC value was considered the best.

2. Next, we fixed the selected covariance structure obtained in the first stage and computed the QIC of each sub-model selected from the preceding stage above. The model that yielded the smallest QIC was considered the best model. Model ranking was accomplished using the MuMIn R package.

3. We repeated the procedures above but we used QIC to select the working correlation structure in stage 1.

4. We applied delta ($\Delta_i = QIC_i - QIC_{min}$) which represent the information loss experienced if we are using a fitted model f_i rather than the best model f_{min} for inference to extract models with stronger evidence of being nearer the truth. According to Burnham and Anderson [7], for the true model $\Delta = 0$ and models having $\Delta_i \leq 2$ have substantial support than those in which $4 \leq \Delta_i \leq 7$ and models with $\Delta_i \geq 10$ have essentially no support. We applied $\Delta_i \leq 4$ to extract the top set of models.

5. We then computed the the MSE of $\hat{\beta}_G^{EQ_{AIC}}$ of the selected model and compared it with the MSE of $\hat{\beta}_G^{QIC}$ so as to validate our hybrid methodology. To achieve this, we made use of K-Fold cross-validation technique.

6.4 Selection of Correct Working Correlation Structure for Shareholder Value Creation Data

We first fit the the full GEE model using correlation structures independence, exchangeable, AR-1 and Toeplitz and unstructured. Empirical likelihood ratio is defined with the general correlation structure Toeplitz. EAIC and QIC values were obtained with each of the five sets of GEE estimates. We used the results to determine the correlation structure preferred by EAIC and the one preferred by QIC for the Shareholder Value Creation data (SVA). The results are shown in Table 6.1. (See Appendix E.1).

Table 6.1: Working Correlation Structure Selection for the SVA Data

	Working Correlation Structures				
	IN	EX	AR-1	Toep	UN
EAIC	1268.282	968.789	967.706	1062.735	1151.333
QIC	941.925	1003.983	844.322	3245.05	799.670

The result indicate that EAIC chooses the AR-1 working correlation structure for the SVA data which has a minimum value of 967.706. The estimated ρ-value for the selected AR-1 correlation structure is 0.775 which implies that the estimated correlation matrix is:

$$R_{AR-1} = \begin{pmatrix} 1 & 0.775 & 0.601 & 0.466 & 0.361 & 0.280 \\ 0.775 & 1 & 0.775 & 0.601 & 0.446 & 0.361 \\ 0.601 & 0.775 & 1 & 0.775 & 0.601 & 0.466 \\ 0.466 & 0.601 & 0.775 & 1 & 0.775 & 0.601 \\ 0.361 & 0.466 & 0.601 & 0.775 & 1 & 0.775 \\ 0.280 & 0.361 & 0.466 & 0.601 & 0.775 & 1 \end{pmatrix} \tag{6.17}$$

On the other hand, QIC chooses the unstructured correlation structure with a minimum QIC value of 799.670. The estimated correlation structure in this case is;

$$R_{UN} = \begin{pmatrix} 1 & 0.566 & 0.587 & 0.465 & 0.773 & 0.189 \\ 0.556 & 1 & 0.648 & 0.547 & 0.534 & 0.330 \\ 0.587 & 0.648 & 1 & 0.871 & 0.616 & 0.316 \\ 0.465 & 0.547 & 0.871 & 1 & 0.758 & 0.534 \\ 0.773 & 0.534 & 0.616 & 0.758 & 1 & 0.795 \\ 0.189 & 0.330 & 0.316 & 0.534 & 0.795 & 1 \end{pmatrix} \tag{6.18}$$

The results indicate that whereas EAIC prefers a parsimonious AR-1 correlation structure, QIC prefers the over-parameterized unstructured correlation structure with 15 correlation parameters.

6.5 Application of QIC to Select SVA Model Based on the WCS Selected by EAIC

Based on the AR-1 working correlation structure selected by EAIC, we applied QIC to select the best model out of the 2^{10} possible models. We used the Multi-Model Inference (MuMIn) R package to generate the model selection table in which the models were ranked based on their QIC values with the model with the minimum QIC value ranked first. We extracted models whose $\Delta_i < 4$. This resulted in the top 15 models which are given Table 6.2. (See Appendix E.2)

Table 6.2: Model Selection Table by QIC When $R_0 = AR - 1$

Model No.	Int	X_6	X_2	X_5	X_4	X_9	X_1	X_7	X_3	X_8	X_{10}	QIC	Delta
479	✓	✓	✓	✓	✓		✓	✓	✓			394	0.00
351	✓	✓	✓	✓	✓		✓		✓			395	0.71
349	✓		✓	✓	✓		✓	✓				396	1.26
477	✓		✓	✓	✓		✓	✓	✓			396	1.38
223	✓	✓	✓	✓	✓		✓	✓				396	1.44
95	✓	✓	✓	✓	✓		✓					396	1.97
221	✓		✓	✓	✓		✓	✓				397	2.59
511	✓	✓	✓	✓	✓	✓	✓	✓	✓			397	2.75
471	✓	✓	✓	✓	✓		✓	✓	✓			398	2.97
93	✓		✓	✓	✓		✓					398	3.04
343	✓	✓	✓		✓		✓		✓			398	3.10
509	✓		✓	✓	✓	✓	✓	✓	✓			398	3.21
383	✓	✓	✓	✓	✓	✓	✓		✓			398	3.25
255	✓	✓	✓	✓	✓	✓	✓	✓				398	3.43
495	✓	✓	✓	✓		✓	✓	✓	✓			398	3.48

The results indicate that the model with covariates X_6 (dividend policy), X_2 (growth rate of earnings), X_5 (debt policy), X_4 (firm size), X_1 (economic spread), X_7 (Level of financial distress) and X_3 (interaction between economic spread and growth) was ranked first hence was the preferred model. The relative variable importance based on the sum of the Akaike Weights over all models including the particular explanatory variable is given in Table 6.3.

Table 6.3: Relative Variable Importance

Covariate	X_2	X_1	X_4	X_7	X_3	X_6	X_5	X_9	X_8
R. Importance	1.00	1.00	0.88	0.81	0.62	0.62	0.56	0.27	0.08
N containing Models	192	192	160	160	64	160	160	160	160

Table 6.3 shows that X_2 (growth rate of earnings) and X_1 (economic spread) are key drivers of shareholder value creation hence giving credence to the Gordon-Constant Growth model. Other important determinants are X_4(firm size), X_5 (leverage), X_3

(interaction between g and (ROE-k_e)), X_6 (dividend policy) and X_7 (firm's level of financial distress) all with relative importance values greater than 0.5.

6.6 Validation of the Model Selected by EQ_{AIC}

We first run a model selection using the unstructured working correlation structure which was selected by QIC. The top 15 models are given in the Table 6.4.

Table 6.4: Model Selection Table by QIC When R_0=Unstructured

Model	Int	X_6	X_2	X_5	X_4	X_9	X_1	X_7	X_3	X_8	X_{10}	QIC	Weight
224	✓	✓	✓	✓	✓		✓	✓		✓		401	0.00
96	✓	✓	✓	✓	✓		✓			✓		402	0.49
216	✓	✓	✓	✓	✓		✓	✓		✓		405	2.56
70	✓		✓				✓			✓		410	8.23
219	✓	✓		✓	✓		✓	✓				420	19.1
91	✓	✓		✓	✓		✓					421	19.2
217	✓			✓	✓		✓	✓				421	19.8
89	✓			✓	✓		✓					421	20.0
251	✓	✓		✓	✓	✓	✓	✓				424	22.1
123	✓	✓		✓	✓		✓					424	22.3
211	✓	✓			✓		✓	✓				424	22.6
83	✓	✓			✓		✓					424	23.0
249	✓			✓	✓	✓	✓	✓				424	23.3
121	✓			✓	✓	✓	✓					425	23.5
209	✓				✓		✓	✓				425	23.6

The results indicate that the model with explanatory variables board size (X_8), dividend policy (X_6), growth of earnings (X_2), debt policy (X_5), economic spread (X_1) and level of financial distress (X_7) was ranked as the best model. This model was not among the top 15 models selected under the EQ_{AIC} procedure.

We compared the predictive performance of the two models selected under the two settings using 10-fold cross-validation to establish efficiency gain of EQ_{AIC} over QIC. The MSE($\hat{\beta}_G^{EQ_{AIC}}$), MSE($\hat{\beta}_G^{QIC}$) and the relative error $\{RE = \frac{MSE(\hat{\beta}_G^{QIC})}{MSE(\hat{\beta}_G^{EQ_{AIC}})}\}$ for 10 iterations are given in Table 6.5. (See Appendix E.3)

Table 6.5: Efficiency of the model Selected under the EQ_{AIC} Procedure

Iteration	MSE($\hat{\beta}_G^{EQ_{AIC}}$)	MSE($\hat{\beta}_G^{QIC}$)	RE
1	7.567	9.594	1.268
2	7.944	9.220	1.161
3	1.022	10.023	9.866
4	1.016	10.302	10.140
5	1.010	10.032	9.932
6	5.215	7.129	1.275
7	8.031	9.822	1.223
8	7.790	9.731	1.249
9	1.014	10.328	10.186
10	8.596	9.830	1.144

The results show that MSE($\hat{\beta}_G^{EQ_{AIC}}$) were far much less than MSE($\hat{\beta}_G^{QIC}$). Further the relative efficiency values were all more than 1 hence it was inferred that the proposed procedure EQ_{AIC} significantly improves the efficiency of the GEE estimates ($\hat{\beta}_G$) hence the conclusion that the selection of the correct working correlation structure significantly improves the efficiency of estimates. The high $MSE(\hat{\beta}_G^{QIC})$ as asserted by Chen and Nicole [12] could be as a result of the unstructured correlation structure preferred by QIC which had more nuisance parameters to estimate and estimating them cost efficiency.

CHAPTER 7

SUMMARY OF RESULTS, CONCLUSIONS AND RECOMMENDATIONS

7.1 Introduction

This Chapter provides a summary of findings on the properties of QIC in selecting the correct working correlation structure and set of covariates for the mean structure. It also provides summary results on the hybrid methodology and its performance in terms of increasing efficiency of GEE estimates and the results of its application to modeling shareholder value creation data. Further, it provides conclusions drawn from the summary results as well as recommendations for statistical modelers and future studies

7.2 Summary of Results

The first objective of the study sought to investigate through simulations, the properties of QIC in selecting the true working correlation structure. We established that when the selection set comprised both parsimonious and over-parameterized structures, QIC selected the true AR-1 and unstructured correlation matrices with rates that increased with the sample size, measurements per subject and degree of correlation. However, the rates of selection hardly reached 50% even at n=200. The slow rate of increase was an indicator that convergence was not with a probability of one. For the true exchangeable structure, QIC's selection rates were less than 20% and declined with increase in n, m and ρ. These results showed that QIC was not consistent in selecting the true correlation structure in the presence of over-parameterized structures. In such scenarios, QIC favored the over-parameterized structures with higher probabilities even when they were not correct ones. It was established that consistency was achieved when only parsimonious structures were considered. There was a big difference in selection

rates of QIC between the two selection sets (ω_1), a set with both parsimonious and over-parameterized structures and (ω_2), a set of parsimonious structures only for AR-1 and exchangeable true structures. For instance, QIC selection rate of the true AR-1 structure was 20.7% under ω_1 (n=20, m=3, $\rho = 0.2$) while it was 41.7% under ω_2 (n=20, m=3, $\rho = 0.2$). For the same settings, the selection rates of the exchangeable structure were 10.5% and 34.2% for ω_1 and ω_2 respectively. Further, we developed a modified version of QIC in which we considered the number (q) of correlation parameters in the working correlation structure and the number (p) of regression parameters as cost components. Using simulation, the new criteria ($QIC_m(R)$) was established to select the true correlation structure with probabilities which were more than twice those of QIC. Also, whereas QIC preferred over-parameterized structures, $QIC_m(R)$ preferred a parsimonious correct structure.

The second objective of the study sought to investigate the properties of QIC in selecting the correct set of covariates for the mean structure. We sought to establish the over-fitting and under-fitting probabilities of QIC hence its consistency in selecting the true model, its sensitivity and sparsity. Theoretical results showed that the probability of QIC selecting under-fit models converged to zero in the limit as $n \longrightarrow \infty$ while its over-fitting probability converged in the limit to a value greater than zero. Using numerical simulations, we verified the result that QIC's under-fitting probabilities converged to Zero as $n \to \infty$ hence had a sensitivity rate of 100%. On the other hand the over-fitting probability of QIC converged to 0.3 as $n \longrightarrow \infty$ hence was 70% specific. This implied that QIC failed to exclude all the non-important variables from the selected model hence was inferred to have low sparsity. This was supported by the finding that QIC selected over-specified models with a probability approaching one as $n \longrightarrow \infty$. We also established that QIC had a high statistical power which ranged from 0.66 to 1 as a result of the low type II error rate. This power test results showed that rejecting any given false null hypothesis was guaranteed for sufficiently large samples even when the effect size was small making QIC good for predictive modeling.

The third objective sought to propose a model selection procedure that will improve

efficiency of the GEE estimator through the selection of the correct correlation structure. We proposed an hybrid methodology that involved the use of EAIC in selecting the working correlation structure and then QIC to select the set of covariates for the mean structure. K-fold cross-validation was used to establish the gain in efficiency of our hybrid methodology over QIC. We established EAIC's selection rates of the true correlation structure to be more than twice those of QIC. Also, the probability of EAIC selecting the true working correlation structure converged to one in the limit as $n \longrightarrow \infty$ hence was consistent. When the hybrid methodology (EQ_{AIC}) was used for model selection, the resulting model was found to be approximately 24% more efficient than the one selected by QIC only.

The fourth objective sought to apply the hybrid methodology(EQ_{AIC}) to model shareholder value creation data. Through the proposed model selection procedure, the AR-1(0.775) correlation structure was preferred for the shareholder value creation data and the model with the covariates: annual growth of earnings, economic spread, size of the firm, dividend policy, leverage and level of financial distress was selected as the best model for shareholder value creation. The relative variable importance index showed that the growth rate of earnings (g) and the economic spread ($ROE - k_e$) were the key drivers of shareholder value creation hence giving credence to the Gordon-Constant Growth model. Other important determinants were firm size, leverage, dividend policy and the firm's level of financial distress all with relative importance values greater than 0.5. When QIC was used to select both the working correlation structure and the covariates, the unstructured correlation matrix was preferred for the shareholder value creation data and the model with covariates: board size, dividend policy, growth of earnings, debt policy, economic spread and level of financial distress was ranked as the best model. This model was not among the top 15 models selected under the EQ_{AIC} procedure. Comparison of the two models selected showed that MSE($\hat{\beta}_G^{EQ_{AIC}}$) $\lll$ MSE($\hat{\beta}_G^{QIC}$). The Relative efficiency values indicated that GEE estimates under EQ_{AIC} were between 14.4% and 918.6% more efficient than those under QIC. Further, it was established that selecting the correct working correlation structure greatly improved

efficiency of the GEE estimates.

7.3 Conclusions

Based on the study findings, the following conclusions are made.

(i) QIC is not effective in selecting the correct working correlation structure for GEE modeling when the selection set includes over-parameterized correlation structures. Its effectiveness can improve if only parsimonious structures are considered.

(ii) Penalizing for the number of correlation and regression parameters estimated improves the consistency of a model selection criteria in selecting the true correlation structure

(iii) QIC does not select the true model with a probability of one hence is not consistent but rather conservative since it includes all of important variables in the model selected. However, dues to its over-fitting level, its use results to models with greater variability.

(iv) Use of the proposed Hybrid methodology (EQ_{AIC}) in GEE ensures that both the right correlation structure and covariates are selected thus leading to models with lower MSE hence higher prediction performance

(v) Gordon-Constant Growth model (Gordon, [28]) still remains important in predicting shareholder value creation.

7.4 Recommendations

7.4.1 Recommendations for Practice

Since the study findings showed that QIC was not effective in selecting the true working correlation structure in the presence of over parameterized structures, we recommend for its use when the selection set consists of parsimonious structures only. Otherwise, we recommend the use of techniques such as $QIC_m(R)$ which penalize for all parameters estimated. Secondly, since QIC was established to be conservative in model selection, we recommend for its use in predictive modeling as it will result to models with better

prediction performance. Third, we recommend the use of the proposed hybrid methodology (EQ_{AIC}) for model selection in GEE rather than the routine use of QIC. This will help in achieving the objective of efficiency along side consistency of the GEE estimates. Fourth, we recommend that researchers in the areas of financial modeling to consider the factors: firm size, debt policy, dividend policy and level of financial distress alongside the Gordon-Constant Growth model in establishing the probability of a firm creating value for its shareholders.

7.4.2 Directions for Future Studies

Our simulation experiments used complete observations. Missing values are common issues in longitudinal data analysis. The impact of missing data on the performance of QIC in selecting the correct working correlation structure and set of covariates should be examined. Future studies on this line can examine the effect of missing data on the performance of QIC and explore whether imputation methodologies remedies the effect or propose modification of QIC to take into account the effect of missing data.

In our real data analysis we assumed that the data were measured accurately. However, in real life data can contain measurement errors and these can affect the performance of a model selection criteria. Future studies can examine the influence of measurement errors on the performance of Kullback I-divergence based model selection criteria in the GEE framework.

Our proposed modified QIC ($QIC_m(R)$) was numerically established to be consistent in selecting the true correlation structure even with the inclusion of over-parameterized structures. We recommend future studies that will develop a theoretical justification for the consistency of $QIC_m(R)$ in the selection the correct working correlation structure.

In our simulation studies, the marginal mean model and variance function were correctly specified. However, the effect of a misspecified or more complicated mean structure resulting from increasing the size of p was not investigated hence we recommend future studies that could investigate the performance of QIC in a high dimensional framework in which the number of covariates (p) increase as n with a possibility of $p > n$.

In our model selection, we assumed the existence of one model that is nearest to the truth. We recommend future studies that incorporate model averaging of the possible models for the longitudinal data incorporating the GEE approach to establish whether for prediction, model averaging is better than the best-model strategy employed in this study.

REFERENCES

[1] Adefowope, O., Dillon, B., Noori, A. and Lehana T. (2008). *Comparison of generalized estimating equations and quadratic inference function using data from the National Longitudinal Survey of Children and Youth(NLSCY) database.* BMC Medical Research Methodology,8:28.

[2] Akaike, H. (1985). *Prediction and entropy, in: A. Atkinson and E. Fienberg, eds..* A Celebration of Statistics. (Springer-Verlag, New York), 1–24.

[3] Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle, in: B. N. Petrov and F. Csaki, eds..* 2^{nd} International Symposium on Information Theory (Akademia Kiado, Budapest), 267–281.

[4] Avital, C.,Nan, M.,L. and Slasor, P. (1997). *Using the General Linear Mixed Models to Analyse Unbalanced Repeated Measures and Longitudinal Data.* Statistics in Medicine. Vol. 16, 2349-2380

[5] Barnett, G., Koper, N. Annette J. D., Schmiegelow, V. and Manseau, M. (2010) *Using information criteria to select the correct variance–covariance structure for longitudinal data in ecology.* Methods in Ecology and Evolution, 1, 15–24

[6] Breslow, N. E. (2003). *Whither PQL?* UW Biostatistics Working Paper Series http://www.bepress.com/uwbiostat/paperI92

[7] Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference.*(Springer-Verlag, New York).

[8] Cantoni, E. Flemming, J. M.and Ronchetti, E.(2005). *Variable selection for marginal longitudinal generalized linear models.*Biometrics; 61:507–514.

[9] Carey, V. J. and Wang, Y., G. (2011). *Working Covariance Model Selection for GEE.*Journal of Statistics and Medicine, Vol. 70, No. 26, P 3117-3124.

[10] Cavanaugh, J., E.(2004). *Criteria for linear model selection based on Kullback's symmetric divergence*. Australian and New Zealand Journal of Statistics Vol. 46, 257-274

[11] Cavanaugh, J., E.(1999). *A large–sample model selection criterion based on Kullback's symmetric divergence*. Statistics and Probability Letters 44, 333–344.

[12] Chen, J.,Nicole, L. (2012). *Selection of Working correlation structure in Generalized Estimating Equations via Empirical Likelihood* Journal of computational and graphical Statistics. Vol.21, No.1,pp.18-41

[13] Cho, H.,Qu, A. (2013). *Model Selection for Correlated Data With Diverging Number of Parameters* Statistica Sinica.

[14] Cochran, W.(1977). *Sampling Techniques* 3^{rd} *Edition* John Wiley and sons.

[15] Copeland T.E and Weston J.F (1988) *Financial Theory and Corporate Policy*, Addison-Wesley

[16] Deroche, C., B. (2015). *Diagnostics and model selection for Generalized Linear Models and Generalized Estimating Equations (Doctoral Dissertation)*. Retieved from http://scholarcommons.sc.edu/etd/3059.

[17] DiCiccio, T., J. anf Efron, B. (1996). *Bootstrap confidence intervals (with discussion)*.Statistical Sciences, 11, 189-228.

[18] Diggle, P. J., Heagerty, P., Liang, K-Y., Zeger, S. L. (2002). *Analysis of Longitudinal Data*.Second edition. New York: Oxford University Press.

[19] Dziak, J. J. (2006) *Penalized Quadratic Inference Functions for Variable Selection in Longitudinal Research*. PhD Thesis. Pennsylvania State University

[20] Dziak, J. J. Donna L,Stephanie T., Runze L. and Lars S. J. (2019) *Sensitivity and Specifity of Information Criteria*. http://dx.doi.org/10.1101/449751

[21] Efron, B. (1986) *How biased is the apparent error rate of a prediction rule?* Journal of the American Statistical Association 81, 461-470.

[22] Erfanul H.,Mahfuzur R. and Wasimul B. (2015) *On the Selection of Relevant Covariates and Correlation Structure in Longitudinal Binary Models: Analysing the*

Impact of the Height on Type II Diabetes. Austrian Journal of Statistics, Volume 44, 3–15.

[23] Fan, J. and R. Li.(2004) *New estimation and model selection procedures for semi-parametric modeling in longitudinal data analysis*. Journal of the American Statistical Association, 99:710–723.

[24] Fan, J. and Li, R. (2001) *Variable selection via nonconcave penalized likelihood and its oracle properties.* Journal of the American Statistical Association, 96, 1348-1360.

[25] Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004). *Applied longitudinal analysis*.NJ: John Wiley and Sons.

[26] Fitzmaurice, G.,M.(1995). *A caveat concerning independence estimating equations with multiple multivariate binary data*. Biometrics 51: 309-317

[27] Friedrich L., Weingessel, A. and Kurt H. (1998). *On the generation of correlated artificial binary data*. Working Paper Series, SFB "Adaptive Information Systems and Modelling in Economics and Management Science", Vienna University of Economics.

[28] Gordon, M.(1962). *The Investment, Financing and Valuationof the Corporation*. Richard D. Irwin

[29] Gosho, M., Hamada, C. and Yoshimura, I.(2011). *Modifications of QIC and CIC for Selecting a Working Correlation Structure in Generalized Estimating Equations Method*. Japanese Journal of Biometrics, Vol. 32, NO.1, 1-12

[30] Hafidi, B. and Mkhadri, A. (2006). *A corrected Akaike criterion based on Kullback's symmetric divergence: applications in time series, multiple and multivariate regression*. Computational Statistics and Data Analysis 50, 1524–1550.

[31] Hanley, J., A. Edwardes, M., D. and Forrester, J., E. (2003). *Statistical Analysis of Correlated Data Using Generalized Estimating Equations: An Orientation*. American Journal of Epidemiology, Vol. 157, No. 4.

[32] Hardin, J., W. and Hilbe, J., M. (2003). *Generalized Estimating Equations*. Chapman and Hall/CRC.

[33] Hardin, J., W. and Hilbe, J., M. (2001). *Generalized Linear Models and Extensions*. Stata Press.

[34] Hin, L. Y. and Wang, Y. G. (2009). *Working Correlation Structure identification in generalized estimating equations*. Statistics in Medicine. 28, 642-658. 10.1002/sim.3489.

[35] Hin, L. Y., Carey, V. J., and Wang, Y. G. (2007). *Criteria for working-correlation structure selection in GEE*. The American Statistician 28, 360-364.

[36] Hirokazu, Y. Hirofumi, W. and Yasunori, F. (2015). *A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large* . Electronic Journal of Statistics 9(1), 869-897.

[37] Honjo, Y. and Harada, N. (2006). *SME Policy, financial structure and firm growth. Evidence from Japan*. Small Business Economies 27(4) 298-300

[38] Hurvich, C. and Tsai, C. L. (1989). *Regression and time series model selection in small samples*. Biometrika 76, 297-307.

[39] Hyun-Joo, K., Cavanaugh, J., E., Tad, A., D. and Fore, S. A. (2014). *Model Selection for Overdispersed Data and their application to the characterization of a host-parasite relationship*. Environ. Ecol. Stat. 21. 329-350

[40] Ilk, O. (2008). *Multivariate Longitudinal Data Analysis: Models for Binary Response and Exploratory Tools for Binary and Continuous Response*. Saarbrucken Verlag Dr. Muller (VDM)

[41] Jamshid, Y., Maryam, T. and Hamed, R.(2018). *Efficiency evaluation of QIF in the analysis of longitudinal data*. Russian Open Medical Journal.Vol. 7 Issue 3. DOI:10.15275/rusomj.2018.0310

[42] Jang, M. ,J.(2011). *Working correlation selection in generalized estimating equations*. PhD (Doctor of Philosophy) thesis, University of Iowa.

[43] Jianwen, X., Jiamao, Z. and Liya, F.(2019). *Variable selection in GEEs via Empirical Likelihood and Gaussian Pseudo-Likelihood*. Communication in Statistics.Vol. 48(4)., 1239-1250.

[44] Johnson, B., Lin, D. and Zeng, D.(2008). *Penalized Estimating Equations and Variable selection in semi-parametric regression models* (2008) Journal of America Statistics Association. 103, 672-680.

[45] Kaurmann, G.and Carroll, R., J. (2008). *A note on the efficiency of sandwich covariance matrix estimation* (2000) Journal of America Statistics Association. Vol. 96, No. 456 pp 1387-1396.

[46] Konishi, S.and Kitagwa, G.(2008). *Information Criteria and Statistical Modeling* (Springer, New York).

[47] Kullback, S. (1968) *Information theory and Statistics* New York, Dover.

[48] Li, J. (2013b). *Predictive Modelling using Random Forests and its Hybrid methods with geostatistical techniques in marine environmental*. The proceedings of the eleventh Australian Data Mining Conference November 2013.

[49] Liang, K. and Zeger, S. (1986). *Longitudinal data analysis using generalized linear models* Biometrica, 73, 12-22

[50] Linhart, H. and Zucchini, W. (1986). *Model Selection* Wiley, New York.

[51] Mancl, L. A. and DeRouen, T. A. (2001). *A covariance estimator for GEE with improved small sample properties* Biometrics, 57, 126-134

[52] McCullagh, P. and Nelder, J. (1983) *Generalized Linear Models*. London: Chapman and Hall.

[53] Molenberghs, G., and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer Science and Business Media, Inc.

[54] Morris, T. P., White, I. R. and Crowther, M. J. (2018). *Using simulation studies to evaluate statistical methods*. Statistics in Medicine, Wiley, Vol 38(11)

[55] Nelder, J. and Wedderburn, R.W.M. (1972) *Generalized Linear Models*. journal of Royal statistical association, series A. 135: p. 370-384.

[56] Neuhaus, J., Kalbfleisch, J. and Hauck, W. (1991). *A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data*. International Statistical Review 59, 25-35.

[57] Odongo,K., Thabhang, M. and Maina, L. (2014). *Determinants of Shareholder Value Creation: panel evidence of listed agricultural firms in Kenya.* MPRA Paper NO. 57116. http://mpra.ub.uni-muechen.de/57116/

[58] Oyebayo,R. and Mohdi, A. (2019). *Bayesian variable selection for multiclass classification using Bootstrap Prior Technique.* Austrian Journal of Statistics Vol.48, 63-72.

[59] Owen, A.B. (1991). *Empirical likelihood for linear models.* The Annals of Statistics, 19, 1725-1747.

[60] Pan, W. (2001a). *Akaike Information Criteria in generalized estimating equations* Biometricss 57, 120-125

[61] Pan, W. (2001b). *On the robust variance estimator in generalized estimating equations* Biometrika 88, 901-906.

[62] Pandey I.M (2005) *"What drives Shareholder Value"* Working Paper wp, No 2005 –09-04, Indian Institute of Management, Ahmedabab, India.

[63] Qu, A., Lindsay, BG., and Li, B. (2000). *Improving generalized estimating equations using QIF* Biometrika, 87(4):823-836. http://doi.org/10.1093/biomet/87.4.823

[64] Rapparport, A. (1986). *Creating Shareholder Value* The Free Press, New York

[65] Rotnitzky, A. and Jewell, N. P. (1990). *Hypothesis testing of regression parameters in semi-parametric generalized linear models for cluster correlated data* Biometrika 77(3), 485-497.

[66] Sakate, D., M. and Kashi, D., N. (2016). *A new robust model selection method in GLM with application to ecological data* Environmental Systems Research, 5:9

[67] Sakate, D., M. and Kashi, D., N. (2014). *A Deviance Based Criterion for Model selection in GLM* Statistics 48: 34-48.

[68] Shao, J. and Rao, J. S. (2000). *The GIC of model selection: A hypothesis testing approach.* J. Statistics, planning and Inference 88, 215–231.

[69] Shibata, R. (1981). *Approximate efficiency of a selection procedure for the number of regression variables.* Biometrika. 71, 43-49.

[70] Shinpei, I.(2015). *Consistent selection of working correlation structure in GEE analysis based on Stein's loss function.*Hiroshima Mathematics Journal, 45, 91–107.

[71] Shinpei I (2009). *On Properties of QIC in Generalized Estimating Equations* Graduate School of Engineering Science, Osaka University.

[72] Shults, J. and Chaganty, N. R. (1998). *Analysis of serially correlated data using quasi-least squares* Biometrics 54(4), 1622-1630.

[73] Stokes, M. E., Davis, S. C. and Koch, G. G. (1998). *Categorical Data analysis using the SAS system* 2^{nd} ed. North Carolina: SAS Institute.

[74] Sutradhar, B., C. and Das, K. (2000). *On the Accuracy of Efficiency of Estimating Equation Approach*. Biometrics, 56(2), 622-625.

[75] Touloumis, A. (2016). *Simulating Correlated Binary and Multinomial Responses under Marginal Model Specification: The SimCorMultRes Package*. The R Journal 8:2, 79-91.

[76] Wang, Y., Orla, M., Wang, Z., Bhatnagar, S.R., Schultz, J. and Moodie, E., M. (2015) . *The Perils of Quasi-Likelihood Information Criteria* Stat Journal 4, 246-254.

[77] Wang, J. Zhou and Qu, A. (2012). *Penalized Generalized Estimating for High Dimensional Longitudinal Data Analysis* Biometrics 68(2), 353-360.

[78] Wang, Y. and Hin, L. (2010). *Modeling strategies in longitudinal data analysis: Covariate, variance function and correlation structure selection* Computational Statistics and Data Analysis 54, 3359–3370.

[79] Wang, Y. G. and Carey, V. (2003). *Working correlation structure misspecification, estimation and covariate design: implications for GEE performance* Biometrika 90, 29-41.

[80] Wedderburn, R. W. M. (1974). *Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method*. Biometrika, 61, 439–447.

[81] Wentao, G.(2019). *Bootstrap-Adjusted Quasi-Likelihood Information Criteria for Mixed Model Selection*. PHD Dissertation, Bowling Green State University.

[82] Yu, I. and Shinpei, I. (2018). *Model selection criterion based on the prediction mean squared error in generalized estimating equations*. Hiroshima Math. J., **48**, 307–334.

[83] Zeger, S. L.and Liang, K., Y. (1986). *Longitudinal Data Analysis for Discrete and Continuous Outcomes*. Biometrics, 42, 121–130.

[84] Zheng, B. (2000). *Summarizing the goodness of fit on generalized linear models for longitudinal data*. Statistics in Medicine, 19, 1265-1275.

APPENDICES

Appendix A

PROOF OF THEOREM AND LEMMA

A.1 Regularity Conditions

Let Y_i, i=1,...,n be iid variable distributed with density $f(y|\theta)$ and let $\hat{\theta}$ and θ_0 be the MLE of the parameter vector θ and the true unknown parameter value respectively. Some of the regularity conditions that the density function of Y_i must satisfy are:

(i) $\theta_0 \in \Theta^0$ where Θ is the parameter space and Θ^0 is in the interior θ

(ii) the true but unknown parameter value θ_0 is identifiable i.e.

$$\theta_0 = \underset{\theta \in \Theta}{\operatorname{argmin}} E(log f(Y_i|\theta))$$

(iii) the log-likelihood function $\ell(\theta|y_1, ..., y_n) = \sum_{i=1}^{n} log f(y_i|\theta)$ is continuous in θ

(iv) $E(log f(Y_1, ..., Y_n)|\theta)$ exists

(v) the log-likelihood function is such that $\frac{1}{n}\ell(\theta|y_1, ..., y_n)$ converges almost surely to $Elog(Y_i|\theta)$ uniformly in $\theta \in \Theta$ i.e.

$$\underset{\theta \in \Theta}{\operatorname{Sup}} |\frac{1}{n}\ell(\theta|y_1, ..., y_n) - Elog(Y_i|\theta)| < \epsilon, \quad \epsilon > 0$$

(vi) the log-likelihood function is twice continuously differentiable in a neighbourhood of θ_0

(vii) the expected Fisher information matrix

$$I(\theta_0) = E\{-\frac{\partial^2 log f(Y_1, ..., Y_n)|\theta_0}{\partial\theta\partial\theta^T}\}$$

exists and is non-singular

A.2 Proof of Theorem (1.6.11)

Let $\rho^*(\beta) = \hat{\rho}[\beta, \hat{\phi}(\beta)]$. Under some regularity conditions $\sqrt{n}(\hat{\beta}_G - \beta) \to N(0, V_{LZ})$ can be approximated by;

$$\{\sum_{i=1}^{n} -\frac{\delta}{\delta\beta} U_i[\beta, \rho^*(\beta)]\}^{-1} \{U_i[\beta, \rho^*(\beta)/\sqrt{n}\}, \tag{A.1}$$

where,

$$\begin{aligned} \frac{\delta}{\delta\beta} U_i[\beta, \rho^*(\beta)] &= \frac{\delta}{\delta\beta} U_i[\beta, \rho^*(\beta)] + [\frac{\delta}{\delta\rho^*} U_i[\beta, \rho^*(\beta)][\frac{\delta}{\delta\beta}\rho^*(\beta)] \\ &= P_i + Q_i R \end{aligned} \tag{A.2}$$

Let β be fixed so that the Taylor expansion gives

$$\begin{aligned} n^{-\frac{1}{2}} \sum U_i[\beta, \rho^*(\beta)] &= n^{-\frac{1}{2}} U_i(\beta, \rho) + \frac{\sum \frac{\partial U_i(\beta,\rho)}{\partial\rho}}{n} n^{-\frac{1}{2}}(\rho^* - \rho) + o_p(1) \\ &= P^* + Q^* R^* + o_p(1) \end{aligned} \tag{A.3}$$

where the sum are over i=1...n. Now $Q^* = o_p(1)$ since $\frac{\partial U_i(\beta,\rho)}{\partial\rho}$ are linear functions of $(y_i - \mu_i)$ whose means are zero. Based on conditions (i) and (iii) in (1.6.11)

$$\begin{aligned} R^* &= \sqrt{n}\{\hat{\rho}(\beta, \phi^*) - \hat{\rho}(\beta, \phi) + \hat{\rho}(\beta, \phi) - \rho\} \\ &= \sqrt{n}\{\frac{\partial\hat{\rho}}{\partial\phi}(\beta, phi^*)(phi^* - \phi) + \hat{\rho}(\beta, \phi) - \rho\} \\ &= O_p(1) \end{aligned} \tag{A.4}$$

Consequently, $\frac{\sum_{i=1}^{n} U_i[\beta, \rho^*]}{\sqrt{n}}$ is asymptotically equivalent to P^* whose asymptotic distribution is multivariate Gaussian with zero mean and covariance matrix

$$\lim_{n\to\infty} \{\sum_{i=1}^{n} D_i' V_i^{-1} Cov(Y_i) V_i^{-1} D_i/n\} \tag{A.5}$$

Then it follows that $\sum_{i=1}^{n} Q_i = o_p(n)$, $R = O_p(1)$and that $\sum_{i=1}^{n} \frac{P_i}{n}$ converges as $n \longrightarrow \infty$ to $\frac{-\sum_{i=1}^{n} D_i' V_i^{-1} D_i}{n}$ which completes the proof.

A.3 Proof of Lemma (1.6.21)

Define $I(\theta_k) = E\langle -\frac{\delta^2 Q(\theta_k|y)}{\delta\theta_k \delta\theta_k^T}\rangle$ to denote the expected fisher information matrix and $I(\theta_k|y) = -\frac{\delta^2 Q(\theta_k|y)}{\delta\theta_k \delta\theta_k^T}$ to be the observed fisher information matrix and if we consider a

second order expansion of $-2Q(\theta_0|y)$ about $\hat{\theta}_k$ and evaluate the expectation of the result we obtain:

$$\begin{aligned} -2Q(\theta_0|y) &= E\{-2Q(\hat{\theta}_k|y)\} \\ &+ [E\{(\hat{\theta}_k - \theta_0)^T[I(\hat{\theta}_k|y)](\hat{\theta}_k - \theta_0)\} \\ &+ o(1) \end{aligned} \tag{A.6}$$

Thus we have;

$$E\{-2Q(\theta_0|y)\} - E\{-2Q(\hat{\theta}_k|y)\} = E\{(\hat{\theta}_k - \theta_0)^T[I(\hat{\theta}_k|y)](\hat{\theta}_k - \theta_0)\} + o(1) \tag{A.7}$$

Next, we consider taking a second order expansion of $d(\theta_0, \hat{\theta}_k)$ about θ_0 and we have;

$$d(\theta_0, \hat{\theta}_k) = d(\theta_0, \theta_0) + (\hat{\theta}_k - \theta_0)^T[I(\theta_0)](\hat{\theta}_k - \theta_0)\} + R(\theta_0, \hat{\theta}_k) \tag{A.8}$$

In this case $R(\theta_0, \hat{\theta}_k)$ is of $o_p(1)$ such that $E\{R(\theta_0, \hat{\theta}_k)\}$ is o(1). Using this result together with the result in (1.79) and evaluating the expectation of (A.8), we have;

$$\begin{aligned} E\{d(\theta_0, \hat{\theta}_k)\} &= E\{-2Q(y|\theta_0)\} \\ &+ E\{(\hat{\theta}_k - \theta_0)^T[I(\theta_0)](\hat{\theta}_k - \theta_0)\} \\ &+ o(1) \end{aligned} \tag{A.9}$$

Thus;

$$E\{d(\theta_0, \hat{\theta}_k)\} - E\{-2Q(y|\theta_0)\} = E\{(\hat{\theta}_k - \theta_0)^T[I(\theta_0)](\hat{\theta}_k - \theta_0)\} + o(1) \tag{A.10}$$

Accordingg to Cavanaugh [10], the quadratic forms $(\hat{\theta}_k - \theta_0)^T[I(\hat{\theta}_k|y)](\hat{\theta}_k - \theta_0)$ and $(\hat{\theta}_k - \theta_0)^T[I(\theta_0)](\hat{\theta}_k - \theta_0)$ converge to centrally distributed chi-square random variables with k degrees of freedom. Since $\theta_0 \in \Theta(k)$, the expectation of both the quadratic forms are within o(1) of k. This fact along with (A.7) and (A.10) establishes (1.113a) and (1.113b).

Appendix B

R-CODE FOR THE INVESTIGATION OF THE PERFORMANCE OF QIC IN SELECTING THE TRUE WORKING CORRELATION STRUCTURE

B.1 R-Code for the Performance of QIC in Selecting the True Correlation Structure

```
library(geepack);library(MESS);library(SimCorMultRes)
N=1000 # number of runs
n=20, #n=30, #n=50, #n=100, #n=200 # number of subjects
clsize=3; #clsize=6; #clsize=9 # number of measurements per subject
intercepts=0.25
betas=c(-0.25, -0.25)
#correlation Matrices#
#AR-1 (alpha=0.2)#
cor.matrix=toeplitz(c(1,0.2,0.04))..............#m=3)
cor.matrix=toeplitz(c(1,0.2,0.04,0.008,1.6e-03,3.2e-04))....#m=6)
cor.matrix=toeplitz(c(1,0.2,0.04,0.008,1.6e-03,3.2e-04,6.4e-05,1.28e-05,2.56e-06))..#m=9)
#AR-1 (alpha=0.5)#
cor.matrix=toeplitz(c(1,0.5,0.25))..............#m=3)
cor.matrix=toeplitz(c(1,0.5,0.25,0.125,0.0625,0.03125))....#m=6)
cor.matrix=toeplitz(c(1,0.5,0.25,0.125,0.0625,0.03125,1.5625e-02,7.8125e-03,3.9063e-03))..#m=9)
#AR-1 (alpha=0.8)#
cor.matrix=toeplitz(c(1,0.8,0.64))..............#m=3)
cor.matrix=toeplitz(c(1,0.8,0.64,0.512,0.4096,0.3277))....#m=6)
cor.matrix=toeplitz(c(1,0.8,0.64,0.512,0.4096,0.3277,2.6214e-01,2.0972e-01,1.6777e-01))..#m=9)
#EXCHANGEABLE (alpha=0.2)#
cor.matrix=toeplitz(c(1,0.2,0.2))..............#m=3)
```

cor.matrix=toeplitz(c(1,0.2,0.2,0.2,0.2,0.2))....#m=6)

cor.matrix=toeplitz(c(1,0.2,0.2,0.2,0.2,0.2,0.2,0.2,0.2))..#m=9)

#EXCHANGEABLE (alpha=0.5)#

cor.matrix=toeplitz(c(1,0.5,0.5))..............#m=3)

cor.matrix=toeplitz(c(1,0.5,0.5,0.5,0.5,0.5))....#m=6)

cor.matrix=toeplitz(c(1,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5))..#m=9)

#EXCHANGEABLE (alpha=0.8)#

cor.matrix=toeplitz(c(1,0.8,0.8))..............#m=3)

cor.matrix=toeplitz(c(1,0.8,0.8,0.8,0.8,0.8))....#m=6)

cor.matrix=toeplitz(c(1,0.8,0.8,0.8,0.8,0.8,0.8,0.8,0.8))..#m=9)

#UNSTRUCTURED #

$$cor.matrix = \begin{pmatrix} 1.00 & 0.80 & 0.60 & 0.14 & 0.10 & 0.23 \\ 0.80 & 1.00 & 0.70 & 0.18 & 0.17 & 0.18 \\ 0.60 & 0.70 & 1.00 & 0.25 & 0.24 & 0.22 \\ 0.14 & 0.18 & 0.25 & 1.00 & 0.45 & 0.22 \\ 0.10 & 0.17 & 0.24 & 0.45 & 1.00 & 0.16 \\ 0.23 & 0.18 & 0.22 & 0.22 & 0.16 & 1.00 \end{pmatrix}$$

min.qic =rep(0,4)

p = c(3,4,4,5)

for (j in 1:N){

x1=rep(rnorm(n),each=clsize)

x2=rep(rbinom(n,2,0.5),each=clsize)

corres=rbin(clsize=clsize,intercepts=intercepts,betas=betas,

xformula=∼x1+x2,cor.matrix=cor.matrix,link="probit")

wm1=geeglm(y∼x1+x2,family=binomial(link="logit"),id=id,corstr="independence",

data=corres$simdata)

wm2=geeglm(y∼x1+x2,family=binomial(link="logit"),id=id,corstr="exchangeable"

data=corres$simdata)

wm3=geeglm(y∼x1+x2,family=binomial(link="logit"),id=id,corstr="ar1"

```
data=corres$simdata)
wm4=geeglm(y~x1+x2,family=binomial(link="logit"),id=id,corstr="unstructured"
data=corres$simdata)
qic1=QIC(wm1)
qic2=QIC(wm2)
qic3=QIC(wm3)
qic4=QIC(wm4)
qic=c(qic1,qic2,qic3, qic4)
print( "QIC");print(qic)
id1=which.min(qic)
min.qic[id1]=min.qic[id1]+1
print(j) }
min.qic;
```

B.2 R-Code for the Comparison of $QIC_m(R)$ and QIC in Selecting the True Correlation Structure

```
library(MASS) library(bindata) library(gee) library(geepack)
# Defining the Working Correlation Matrices #
indep.corr = function(t,aplha){
diag(1,t)
}
exch.corr = function(t,alpha){
exch = matrix(1,t,t)
exch[1,] = c(1,rep(alpha,(t-1)))
for (i in 2:t){
exch[i,] = c(exch[i-1,t],exch[i-1,-t])
}
return(exch)
```

```
}
ar1.corr = function(t,alpha){
ar1 = matrix(1,t,t)
for (i in 1:(t-1)){
for (j in (i+1):t) {
ar1[i,j] = alpha^(j-i)
ar1[j,i] = ar1[i,j]
}
return(ar1)
}
toep.corr=function(t,a){
toep = diag(1,t)
m = length(a)
for (i in 1:m){
toep[abs(col(toep)-row(toep))==i] = a[i]
}
return(toep)
}
# My Gee function. Defining QIC and QIC_m(R) Functions #
mygee = function(y, x , t, fam, corr="exch", scale=1, tol=1e-7)   for binary
x1 = x[,1]
x2 = x[,2]
beta.old = glm(y ~ 0 + x1 + x2, family=fam)$coef
n = length(y)/t
x = aperm(array(t(x),c(2,t,n)),c(2,1,3))
y = matrix(y,t,n)
stop = 0
count = 0
while (stop==0) {
```

```
count = count+1
alpha.num = 0
alpha1.num = alpha2.num =0
error = p.err = matrix(0,t,n)
D = array(0,c(t,3,n))
A = array(0,c(t,t,n))
left = array(0,c(3,3,n))
right = matrix(0, 3,n)
if (corr==”ind”){
stop =1
beta.new = beta.old
for (i in 1:n) {
fitted = plogis( x[„i] %*% beta.old )
error[,i] = y[,i]-fitted
p.err[,i] = error[,i]*(fitted *(1-fitted))ˆ(-1/2(
}
phi=sum(p.errˆ2)/(n*t-3)
theta.new = c(beta.new, phi)
}
else if ( corr==”exch” ) {
for (i in 1:n) { # cal. pearson resd.
fitted = plogis( x[„i] %*% beta.old )
error[,i] = y[,i]-fitted
p.err[,i] = error[,i]*(fitted*(1-fitted))ˆ(-1/2)
D[„i] = matrix(rep(fitted*(1-fitted),3),t,3)*x[„i]
A[„i] = diag(as.vector((fitted(1-fitted))ˆ(-1)))
alpha.num = alpha.num + sum(p.err[,i]%*%p.err[,i])-sum(p.err[,i]*p.err[,i]) # exchange-
able
}
```

```
phi = sum(p.err^2)/(n*t-3)
alpha = alpha.num/(n*t*(t-1)-6)/phi # exchangeable
R = exch.corr(t,alpha)
for (i in 1:n) {
A.half = A[,,i]^(1/2)
left[,,i] = crossprod(D[,,i], A.half %*% solve(R, A.half %*% D[,,i]))
right[,i] = crossprod(D[,,i], A.half %*% solve(R, A.half %*% error[,i]))
}
beta.new = beta.old + solve(apply(left,c(1,2),sum),apply(right,1,sum))
if ( max(abs(beta.new-beta.old))=tol) {
alpha.num = 0
p.err = matrix(0,t,n)
for (i in 1:n) { # cal. pearson resd.
fitted = plogis(x[,,i]%*% beta.new )
p.err[,i] = (y[,i]-fitted)*(fitted*(1-fitted))^(-1/2)
alpha.num = alpha.num + sum(p.err[,i]%o%p.err[,i])- sum(p.err[,i]*p.err[,i])  exchange-
able
}
phi = sum(p.err^2)/(n*t-3)
alpha = alpha.num/(n*t*(t-1)-6)/phi # exchangeable
theta.new = c(beta.new, alpha, phi)
stop = 1
}
else beta.old = beta.new
}
# end if "exch"
else if (corr=="ar1") {
for (i in 1:n) { # cal. pearson resd.
fitted = plogis( x[,,i] error[,i] = y[,i]-fitted
```

p.err[,i] = error[,i]*(fitted*(1-fitted))ˆ(-1/2)

D[„i] = matrix(rep(fitted*(1-fitted),3),t,3)* x[„i]

A[„i] = diag(as.vector((fitted(1-fitted))ˆ(-1))

alpha.num = alpha.num + sum(p.err[-t,i]*p.err[-1,i]) # AR(1)

}

phi = sum(p.errˆ2)/(n*t-3)

alpha = alpha.num/(n*(t-1)-3)/phi # AR(1)

R = ar1.corr(t,alpha)

for (i in 1:n) {

A.half = A[„i]ˆ(1/2)

left[„i] = crossprod(D[„i], A.half %*% solve(R, A.half %*% D[„i]))

right[,i] = crossprod(D[„i], A.half %*% solve(R, A.half %*% error[,i]))

} beta.new = beta.old + solve(apply(left,c(1,2),sum),apply(right,1,sum))

if (max(abs(beta.new-beta.old)) ¡= tol) {

alpha.num = 0

p.err = matrix(0,t,n)

for (i in 1:n) { # cal. pearson resd.

fitted = plogis(x[„i]p.err[,i] = (y[,i]-fitted)*(fitted*(1-fitted))ˆ(-1/2)

alpha.num = alpha.num + sum(p.err[-t,i]*p.err[-1,i]) # AR(1)

}

phi = sum(p.errˆ2)/(n*t-3)

alpha = alpha.num/(n*(t-1)-3)/phi # AR(1)

theta.new = c(beta.new, alpha, phi)

stop = 1

} @ end if

else beta.old = beta.new

} #@ end else

else if (corr==”toep”){

for (i in 1:n) { # cal. pearson resd.

```
fitted = plogis( x[„i] %*% beta.old )
error[,i] = y[,i]-fitted
p.err[,i= error[,i]*(fitted*(1-fitted))^(-1/2)
D[„i] = matrix(rep(fitted*(1-fitted),3),t,3)*x[„i]
A[„i] = diag(as.vector((fitted*(1-fitted))^(-1))
alpha1.num = alpha1.num + p.err[1,i]*p.err[2,i] + p.err[2,i]*p.err[3,i] # Toeplitz
alpha2.num = alpha2.num + p.err[1,i]*p.err[3,i]
}
phi = sum(p.err^2)/(n*t-3)
alpha1 = alpha1.num/(2*n-3)/phi # Toep
alpha2 = alpha2.num/(n-3)/phi
R = toep.corr(t, c(alpha1, alpha2))
for (i in 1:n) {
A.half = A[„i]^(1/2)
left[„i] = crossprod(D[„i], A.half %*% solve(R, A.half %*% D[„i]))
right[,i] = crossprod(D[„i], A.half %*% solve(R, A.half %*% error[,i]))
}
beta.new = beta.old + solve(apply(left,c(1,2),sum),apply(right,1,sum))
if ( max(abs(beta.new-beta.old)) = tol) {
alpha1.num = alpha2.num = 0
p.err = matrix(0,t,n)
for (i in 1:n) { # cal. pearson resd.
fitted = plogis(x[„i]%*% beta.new )
p.err[,i] = (y[,i]-fitted)*(fitted*(1-fitted))^(-1/2)
alpha1.num = alpha1.num + p.err[1,i]*p.err[2,i] + p.err[2,i]*p.err[3,i] # Toeplitz
alpha2.num = alpha2.num + p.err[1,i]*p.err[3,i]
}
phi = sum(p.err^2)/(n*t-3)
alpha1 = alpha1.num/(2*n-3)/phi  Toep
```

```
alpha2 = alpha2.num/(n-3)/phi
theta.new = c(beta.new, alpha1, alpha2, phi)
stop = 1
} #@ end if
else beta.old = beta.new
}
} #@ end while
return(theta.new)
}
qicf = function(b, a, phi.i, phi, corr){
x = aperm(array(t(x),c(3,t,n)),c(2,1,3))  data transformation
y = matrix(y, t,n)
D = array(0,c(t,3,n))
A = array(0,c(t,t,n))
omega.i = I0 = matrix(0,3,3)
ql = 0
if (corr==”indep”)
R = indep.corr(t)
else if (corr==”exch”)
R = exch.corr(t,a)
else if (corr==”ar1”)
R = ar1.corr(t,a)
else if (corr==”toep”)
R = toep.corr(t,a)
for (i in 1:n){
fitted = plogis( x[,,i] error = y[,i]-fitted
D = matrix(rep(fitted*(1-fitted),3),t,3)*x[,,i]
A = diag(as.vector((fitted*(1-fitted))^(-1)))
A.half = A^(1/2)
```

```
omega.i = omega.i + crossprod(D, A%*%D)
I0 = I0 + crossprod(D, A.half %*% solve(R, A.half %*% D))
I1.left = crossprod(D, A.half %*% solve(R, A.half %*% error))
I1 = I1 + tcrossprod(I1.left, I1.left)
ql = ql + sum( y[,i]*(log(fitted)-log(1-fitted))+log(1-fitted) )
}
omega.i = omega.i/phi.i
V.r = solve(I0, I1)%*%solve(I0)
print(solve(I0)*phi)
print(V.r)
qic.v = 2*(-ql/phi.i)+ 2*sum(diag(omega.i%*%V.r))
return(qic.v)
}
# Defining the modified QIC(QIC_m(R)) Function#
Mqicf = function(b, a, phi.i, phi, corr, q){
x = aperm(array(t(x),c(3,t,n)),c(2,1,3)) # data transformation
y = matrix(y, t,n)
# D = array(0,c(t,3,n))
# A = array(0,c(t,t,n))
omega.i = I0 ¡- I1 ¡- matrix(0,3,3)
ql = 0
if (corr==”indep”)
R = indep.corr(t)
else if (corr==”exch”)
R = exch.corr(t,a)
else if (corr==”ar1”)
R = ar1.corr(t,a)
else if (corr==”toep”)
R = toep.corr(t,a)
```

```
if (corr==”indep”)
q = 0
else if (corr==”exch”)
q = 1
else if (corr==”ar1”)
q = 1
else if (corr==”toep”)
q= t-1
for (i in 1:N){
fitted = plogis( x[„i] %*% b )
error = y[,i]-fitted
D = matrix(rep(fitted*(1-fitted),3),t,3)*x[„i]
A = diag(as.vector((fitted*(1-fitted))^(-1)))
A.half = A^(1/2)
omega.R = omega.i + crossprod(D, A.half %*% solve(R, A.half %*% D))
I0 = I0 + crossprod(D, A.half %*% solve(R, A.half %*% D))
I1.left = crossprod(D, A.half %*% solve(R, A.half %*% error))
I1 = I1 + tcrossprod(I1.left, I1.left)
ql = ql + sum( y[,i]*(log(fitted)-log(1-fitted))+log(1-fitted) )
}
omega.i = omega.i/phi.i
V.r = solve(I0, I1)%*%solve(I0)
# print(solve(I0)*phi)# print(V.r)
Mqic.v = 2*(-ql/phi.i)+ 4*p*sum(diag(omega.i*V.r))
+(2*q/(t*(t-1)))*sum(diag(omega.R*V.r))
return(Mqic.v)
}
# correlated binary data generator #
bin.gen = function(x, beta, bcorr, n, t){
```

```
x = aperm(array(t(x), c(3,t,n)),c(2,1,3)) # data transformation
B = numeric(0)
for (i in 1:n) {
mu = plogis(x[,,i] y = as.vector(rmvbin(1, margprob=mu, bincorr=bcorr))
B = c(B, y)
}
return(B)
}
# Data generation #
N = 1000 # number of runs
n=20, #n=30, #n=50, #n=100, #n=200 # number of subjects
t = 3 #number of measurements per subject
beta = c(0.25, -0.25, -0.25) # true params.
alpha = 0.5 #within subject correlation
I.t = diag(rep(1,t))  t by t identity matrix
sub = rep(1:n,rep(t,n))
bcorr = indep.corr(t)
#bcorr = exch.corr(t, alpha)
#bcorr = ar1.corr(t, alpha)
#bcorr = toep.corr(t, c(0.5,0.35))
# Wc2=[IN, EXCH, AR1, TOEP] #
#SIMULATION 4              #
est.gee1 = est.gee2 = est.gee3 = est.gee4 = est.qic=est.Mqic =numeric(0)
min.qic= min.mqic = rep(0,4)
p = c(3,4,4,5)
for (j in 1:N) {
x = cbind(rep(1,t*n), rbinom(n*t,1,0.5), rep(seq(0,t-1),n)) # the long vector
p=dim(x)[2]
y = bin.gen(x, beta, bcorr,n,t)
```

```
new.gee1 = mygee(y,x,t,binomial, corr="ind")
est.gee1 = c(est.gee1, new.gee1[c(1,2,3)] )
new.gee2 = mygee(y,x,t,binomial, corr="exch")
est.gee2 = c(est.gee2, new.gee2[c(1,2,3)] )
new.gee3 = mygee(y,x,t,binomial, corr="ar1")
est.gee3 = c(est.gee3, new.gee3[c(1,2,3)])
new.gee4 = mygee(y,x,t,binomial, corr="toep")
est.gee4 = c(est.gee4, new.gee4[c(1,2,3)])
aex.hat = new.gee2[4]
aar.hat = new.gee3[4]
ast.hat = new.gee4[c(4,5)]
phi.i = 1; new.gee1[2]
qic1 = qicf(new.gee1[c(1,2,3)], 0, phi.i, phi.i, corr="indep")
qic2 = qicf(new.gee2[c(1,2,3)], aex.hat, phi.i, new.gee2[5], corr="exch")
qic3 = qicf(new.gee3[c(1,2,3)], aar.hat, phi.i, new.gee3[5], corr="ar1")
qic4 = qicf(new.gee4[c(1,2,3)], ast.hat, phi.i, new.gee4[6], corr="toep")
mqic1 = Mqicf(new.gee1[c(1,2,3)], 0, phi.i, phi.i, corr="indep")
mqic2 = Mqicf(new.gee2[c(1,2,3)], aex.hat, phi.i, new.gee2[5], corr="exch")
mqic3 = Mqicf(new.gee3[c(1,2,3)], aar.hat, phi.i, new.gee3[5], corr="ar1")
mqic4 = Mqicf(new.gee4[c(1,2,3)], ast.hat, phi.i, new.gee4[6], corr="toep")
qic= c(qic1,qic2,qic3,qic4)
mqic = c(mqic1,mqic2,mqic3,mqic4)
print("QIC"); print(qic)
print("MQIC"); print(mqic)
id5=which.min(qic);
min.qic[id3] = min.qic[id3]+1
id6=which.min(mqic); min.mqic[id4] = min.mqic[id4]+1
print(j) }
min.qic; min.mqic;
```

Appendix C

R-CODE FOR THE INVESTIGATION OF THE PERFORMANCE OF QIC IN VARIABLE SELECTION

```
    library(SimCorMultRes)
N=1000 # number of runs
n=20, #n=30, #n=50, #n=100, #n=200 # number of subjects
clsize=3; #clsize=6; #clsize=9 # number of measurements per subject
intercepts=0.25; betas=c(-0.25, -0.25, 0, 0)
#AR-1 (alpha=0.2)#
cor.matrix=toeplitz(c(1,0.2,0.04))..............#m=3)
cor.matrix=toeplitz(c(1,0.2,0.04,0.008,1.6e-03,3.2e-04))....#m=6)
cor.matrix=toeplitz(c(1,0.2,0.04,0.008,1.6e-03,3.2e-04,6.4e-05,1.28e-05,2.56e-06))..#m=9)
#AR-1 (alpha=0.5)#
cor.matrix=toeplitz(c(1,0.5,0.25))..............#m=3)
cor.matrix=toeplitz(c(1,0.5,0.25,0.125,0.0625,0.03125))....#m=6)
cor.matrix=toeplitz(c(1,0.5,0.25,0.125,0.0625,0.03125,1.5625e-02,7.8125e-03,3.9063e-03))..#m=9)
qic1=qic2=qic3=qic4=qic5=qic6=qic7=qic8=est.qic =numeric(0)
min.qic =rep(0,8)
p = c(3,4,4,5,5,6,6,7)
for (j in 1:N){
x1=rep(rnorm(n),each=clsize); x2=rep(rbinom(n,2,0.5),each=clsize)
x3=rep(runif(n,0,1),each=clsize); x4=rep(runif(n,0,1),each=clsize)
corres=rbin(clsize=clsize,intercepts=intercepts,betas=betas,
xformula=~x1+x2+x3+x4,cor.matrix=cor.matrix,link="probit")
library(geepack); library(MESS); library(MuMIn)
m1=geeglm(y~x1+x2+x3+x4,family=binomial(link="logit"),id=id,
data=corres$simdata)
```

```
m2=geeglm(y~ x1+x2+x3,family=binomial(link="logit"),id=id,
data=corres$simdata)
m3=geeglm(y~x1+x2+x4,family=binomial(link="logit"),id=id,
data=corres$simdata)
m4=geeglm(y~x1+x3+x4,family=binomial(link="logit"),id=id,
data=corres$simdata)
m5=geeglm(y~x1+x2,family=binomial(link="logit"),id=id,
data=corres$simdata)
m6=geeglm(y~x1+x3,family=binomial(link="logit"),id=id,
data=corres$simdata)
m7=geeglm(y~x1+x4,family=binomial(link="logit"),id=id,
data=corres$simdata)
m8=geeglm(y~x1,family=binomial(link="logit"),id=id,
data=corres$simdata)
qic1=QIC(m1)
qic2=QIC(m2)
qic3=QIC(m3)
qic4=QIC(m4)
qic5=QIC(m5)
qic6=QIC(m6)
qic7=QIC(m7)
qic8=QIC(m8)
qic=c(qic1,qic2,qic3, qic4, qic5,qic6,qic7, qic8)
print( "QIC");print(qic)
id2=which.min(qic)
min.qic[id2]=min.qic[id2]+1
print(j) }
min.qic;
```

Appendix D

R-CODE TO INVESTIGATE PERFORMANCE OF (EQ_{AIC}) IN IMPROVING EFFICIENCY OF $\hat{\beta}$

D.1 R-Code to Investigate Performance of EAIC in Selecting the True Working Correlation Structure

```
library(MASS)
library(emplik)
library(bindata)
library(gee)
library(geepack)
```

D.1.1 Defining the Working Correlation Structures

```
indep.corr = function(t,aplha){
diag(1,t)
}
exch.corr = function(t,alpha){
exch = matrix(1,t,t)
exch[1,] = c(1,rep(alpha,(t-1)))
for (i in 2:t){
exch[i,] = c(exch[i-1,t],exch[i-1,-t])
}
return(exch)
}
ar1.corr = function(t,alpha){
ar1 = matrix(1,t,t)
for (i in 1:(t-1)){
```

```
for (j in (i+1):t) {
ar1[i,j] = alpha^
    j-i)
ar1[j,i] = ar1[i,j]
}
return(ar1)
}
toep.corr=function(t,a){
toep = diag(1,t)
m = length(a)
for (i in 1:m){
toep[abs(col(toep)-row(toep))==i] = a[i]
}
return(toep)
```

D.1.2 Defining the Empirical Likelihood Ratio (ELR) with a Toeplitz Structure

```
elr.st = function(b,a1,a2) { # toeplitz
m = 5 # m — the dim of the est. func. g
x = aperm(array(t(x),c(3,t,n)),c(2,1,3)) # data transformation
y = matrix(y, t,n)
g = matrix(0, m,n)
R = toep.corr(t, c(a1, a2))
error = p.err = matrix(0,t,n)
D = array(0,c(t,3,n))
A = array(0,c(t,t,n))
for (i in 1:n) {
fitted = plogis( x[,,i] %*% b )
```

```
error[,i] = y[,i]-fitted
p.err[,i] = error[,i]*(fitted*(1-fitted))^(-1/2)
D[„i] = matrix(rep(fitted*(1-fitted),3),t,3)*x[„i]
A[„i] = diag(as.vector((fitted*(1-fitted))^(-1))
}
phi= sum(p.err^2)/(n*t-3)
for (i in 1:n) {
A.half = A[„i]^(1/2)
g[c(1,2,3),i] = crossprod(D[„i], A.half %*% solve(R, A.half %*% error[,i]))*phi^(-1)
g[4,i] = p.err[1,i]*p.err[2,i] + p.err[2,i]*p.err[3,i] - a1*phi*(2-3/n)
g[5,i] = p.err[1,i]*p.err[3,i] - a2*phi*(1-3/n)
}
g = t(g)
g.mu = rep(0,m) el.test(g, g.mu,gradtol=1e-9)$"-2LLR"
}
```

D.1.3 Defining QIC and CIC Functions

```
# My Gee function. Defining QIC and CIC Functions #
mygee = function(y, x , t, fam, corr="exch", scale=1, tol=1e-7)   for binary
x1 = x[,1]
x2 = x[,2]
beta.old = glm(y ~ 0 + x1 + x2, family=fam)$coef
n = length(y)/t
x = aperm(array(t(x),c(2,t,n)),c(2,1,3))
y = matrix(y,t,n)
stop = 0
count = 0
while (stop==0) {
```

```
count = count+1
alpha.num = 0
alpha1.num = alpha2.num =0
error = p.err = matrix(0,t,n)
D = array(0,c(t,3,n))
A = array(0,c(t,t,n))
left = array(0,c(3,3,n))
right = matrix(0, 3,n)
if (corr==”ind”){
stop =1
beta.new = beta.old
for (i in 1:n) {
fitted = plogis( x[„i] %*% beta.old )
error[,i] = y[,i]-fitted
p.err[,i] = error[,i]*(fitted*(1-fitted))ˆ(-1/2)
}
phi = sum(p.errˆ2)/(n*t-3)
theta.new = c(beta.new, phi)
}
else if ( corr==”exch” ) {
for (i in 1:n) { # cal. pearson resd.
fitted = plogis( x[„i] %*% beta.old )
error[,i] = y[,i]-fitted
p.err[,i] = error[,i]*(fitted*(1-fitted))ˆ(-1/2)
D[„i] = matrix(rep(fitted*(1-fitted),3),t,3)*x[„i]
A[„i] = diag(as.vector((fitted*(1-fitted))ˆ(-1)))
alpha.num = alpha.num + sum(p.err[,i]%*%p.err[,i])-sum(p.err[,i]*p.err[,i])  exchange-
able
}
```

```
phi = sum(p.err^2)/(n*t-3)
alpha = alpha.num/(n*t*(t-1)-6)/phi # exchangeable
R = exch.corr(t,alpha)
for (i in 1:n) {
A.half = A[,,i]^(1/2)
left[,,i] = crossprod(D[,,i], A.half %*% solve(R, A.half %*% D[,,i]))
right[,i] = crossprod(D[,,i], A.half %*% solve(R, A.half %*% error[,i]))
}
beta.new = beta.old + solve(apply(left,c(1,2),sum),apply(right,1,sum))
if ( max(abs(beta.new-beta.old)) <= tol) {
alpha.num = 0
p.err = matrix(0,t,n)
for (i in 1:n) { # cal. pearson resd.
fitted = plogis(x[,,i%*% beta.new )
p.err[,i] = (y[,i]-fitted)*(fitted*(1-fitted))^(-1/2)
alpha.num = alpha.num + sum(p.err[,i]%*%p.err[,i])- sum(p.err[,i]*p.err[,i])  exchange-
able
}
phi = sum(p.err^2)/(n*t-3)
alpha = alpha.num/(n*t*(t-1)-6)/phi # exchangeable
theta.new = c(beta.new, alpha, phi)
stop = 1
}
else beta.old = beta.new
}
else if (corr=="ar1") {
for (i in 1:n)  # cal. pearson resd.
fitted = plogis( x[,,i] %*% beta.old )
error[,i] = y[,i]-fitted
```

```
p.err[,i] = error[,i]*(fitted*(1-fitted))^(-1/2)
D[„i] = matrix(rep(fitted*(1-fitted),3),t,3)*x[„i]
A[„i] = diag(as.vector((fitted*(1-fitted))^(-1)))
alpha.num = alpha.num + sum(p.err[-t,i]*p.err[-1,i]) # AR(1)
}
phi = sum(p.err^2)/(n*t-3)
alpha = alpha.num/(n*(t-1)-3)/phi # AR(1)
R = ar1.corr(t,alpha)
for (i in 1:n) {
A.half = A[„i]^(1/2)
left[„i] = crossprod(D[„i], A.half %*% solve(R, A.half %*% D[„i]))
right[,i] = crossprod(D[„i], A.half %*% solve(R, A.half %*% error[,i]))
} beta.new = beta.old + solve(apply(left,c(1,2),sum),apply(right,1,sum))
if ( max(abs(beta.new-beta.old)) <= tol) {
alpha.num = 0
p.err = matrix(0,t,n)
for (i in 1:n) { # cal. pearson resd.
fitted = plogis(x[„i]p.err[,i] = (y[,i]-fitted)*(fitted*(1-fitted))^(-1/2)
alpha.num = alpha.num + sum(p.err[-t,i]*p.err[-1,i]) # AR(1)
}
phi = sum(p.err^2)/(n*t-3)
alpha = alpha.num/(n*(t-1)-3)/phi # AR(1)
theta.new = c(beta.new, alpha, phi)
stop = 1
} @ end if
else beta.old = beta.new
} #@ end else
else if (corr==”toep”){
for (i in 1:n) { # cal. pearson resd.
```

```
fitted = plogis( x[,,i] error[,i] = y[,i]-fitted
p.err[,i= error[,i]*(fitted*(1-fitted))^(-1/2)
D[,,i] = matrix(rep(fitted*(1-fitted),3),t,3)*x[,,i]
A[,,i] = diag(as.vector((fitted*(1-fitted))^(-1)))
alpha1.num = alpha1.num + p.err[1,i]*p.err[2,i] + p.err[2,i]*p.err[3,i] # Toeplitz
alpha2.num = alpha2.num + p.err[1,i]*p.err[3,i]
}
phi = sum(p.err^2)/(n*t-3)
alpha1 = alpha1.num/(2*n-3)/phi # Toep
alpha2 = alpha2.num/(n-3)/phi
R = toep.corr(t, c(alpha1, alpha2))
for (i in 1:n) {
A.half = A[,,i]^(1/2)
left[,,i] = crossprod(D[,,i], A.half %*% solve(R, A.half %*% D[,,i]))
right[,i] = crossprod(D[,,i], A.half %*% solve(R, A.half %*% error[,i]))
}
beta.new = beta.old + solve(apply(left,c(1,2),sum),apply(right,1,sum))
if ( max(abs(beta.new-beta.old)) <= tol) {
alpha1.num = alpha2.num = 0
p.err = matrix(0,t,n)
for (i in 1:n) { # cal. pearson resd.
fitted = plogis(x[,,i]%*% beta.new )
p.err[,i] = (y[,i]-fitted)*(fitted*(1-fitted))^(-1/2)
alpha1.num = alpha1.num + p.err[1,i]*p.err[2,i] + p.err[2,i]*p.err[3,i] # Toeplitz
alpha2.num = alpha2.num + p.err[1,i]*p.err[3,i]
}
phi = sum(p.err^2)/(n*t-3)
alpha1 = alpha1.num/(2*n-3)/phi  Toep
alpha2 = alpha2.num/(n-3)/phi
```

```
theta.new = c(beta.new, alpha1, alpha2, phi)
stop = 1
} #@ end if
else beta.old = beta.new
}
} #@ end while
return(theta.new)
}
qicf = function(b, a, phi.i, phi, corr){
x = aperm(array(t(x),c(3,t,n)),c(2,1,3))  data transformation
y = matrix(y, t,n)
D = array(0,c(t,3,n))
A = array(0,c(t,t,n))
omega.i = I0 = matrix(0,3,3)
ql = 0
if (corr==”indep”)
R = indep.corr(t)
else if (corr==”exch”)
R = exch.corr(t,a)
else if (corr==”ar1”)
R = ar1.corr(t,a)
else if (corr==”toep”)
R = toep.corr(t,a)
for (i in 1:n){
fitted = plogis( x[,,i] %*% b )
error = y[,i]-fitted
D = matrix(rep(fitted*(1-fitted),3),t,3)*x[,,i]
A = diag(as.vector((fitted*(1-fitted))^(-1)))
A.half = A^(1/2)
```

```
omega.i = omega.i + crossprod(D, A%*%D)
I0 = I0 + crossprod(D, A.half %*% solve(R, A.half %*% D))
I1.left = crossprod(D, A.half %*% solve(R, A.half %*% error))
I1 = I1 + tcrossprod(I1.left, I1.left)
ql = ql + sum( y[,i]*(log(fitted)-log(1-fitted))+log(1-fitted) )
}
omega.i = omega.i/phi.i
V.r = solve(I0, I1)%*%solve(I0)
print(solve(I0)*phi)
print(V.r)
qic.v = 2*(-ql/phi.i)+ 2*sum(diag(omega.i%*%V.r))
return(qic.v)
}
}
cicf = function(b, a, phi.i, phi, corr){
x = aperm(array(t(x),c(3,t,n)),c(2,1,3)) # data transformation
y = matrix(y, t,n)
# D = array(0,c(t,3,n))
# A = array(0,c(t,t,n))
omega.i = I0 = I1 = matrix(0,3,3)
if (corr==”indep”)
R = indep.corr(t)
else if (corr==”exch”)
R = exch.corr(t,a)
else if (corr==”ar1”)
R = ar1.corr(t,a)
else if (corr==”toep”)
R = toep.corr(t,a)
for (i in 1:n){
```

```
fitted = plogis( x[,,i] %*% b )
error = y[,i]-fitted
D = matrix(rep(fitted*(1-fitted),3),t,3)*x[,,i]
A = diag(as.vector((fitted*(1-fitted))^(-1))
A.half = A^(1/2)
omega.i = omega.i + crossprod(D, A%*%D)
I0 = I0 + crossprod(D, A.half %*% solve(R, A.half %*% D))
I1.left = crossprod(D, A.half %*% solve(R, A.half %*% error))
I1 = I1 + tcrossprod(I1.left, I1.left)
}
omega.i = omega.i/phi.i
V.r = solve(I0, I1)%*%solve(I0)
# print(solve(I0)*phi)
# print(V.r)
cic.v = sum(diag(omega.i%*%V.r))
return(cic.v)
}
```

D.1.4 Correlated Binary Data Generator

```
bin.gen = function(x, beta, bcorr,n,t){
x = aperm(array(t(x), c(3,t,n)),c(2,1,3)) # data transformation
B = numeric(0)
for (i in 1:n) {
mu = plogis(x[,,i] %*% beta)
y = as.vector(rmvbin(1, margprob=mu, bincorr=bcorr))
B = c(B, y)
}
return(B)
```

}

D.1.5 Data generation

N = 1000 # number of runs

n = 20 # number of subjects

n = 30;# n = 50; # n = 100; #n = 200;

t = 3; #number of measurements per subject

beta = c(0.25, -0.25, -0.25) # true params.

alpha = 0.5 #within subject correlation

alpha = 0.8;

beta = c(0.25, -0.25, -0.25) # true params.

I.t = diag(rep(1,t)) # t by t identity matrix

sub = rep(1:n,rep(t,n))

bcorr = indep.corr(t)

bcorr = exch.corr(t, alpha)

bcorr = ar1.corr(t, alpha)

bcorr = toep.corr(t, c(0.5,0.35))

Wc2=[IN, EXCH, AR1, TOEP]

SIMULATION 4#

est.gee1 = est.gee2 = est.gee3 = est.gee4 = est.aic = est.qic =numeric(0)

eaic = qic = cic= rep(0,4)

min.eaic = min.qic = min.cic =rep(0,4)

min.cic =rep(0,4)

p = c(3,4,4,5)

for (j in 1:N) {

data generation

x = cbind(rep(1,t*n), rnorm(n*t,0,1) rbinom(n*t,1,0.5),n)) the long vector

y = bin.gen(x, beta, bcorr,n,t)

```
new.gee1 = mygee(y,x,t,binomial, corr="ind")
est.gee1 = c(est.gee1, new.gee1[c(1,2,3)] )
new.gee2 = mygee(y,x,t,binomial, corr="exch")
est.gee2 = c(est.gee2, new.gee2[c(1,2,3)] )
new.gee3 = mygee(y,x,t,binomial, corr="ar1")
est.gee3 = c(est.gee3, new.gee3[c(1,2,3)])
new.gee4 = mygee(y,x,t,binomial, corr="toep")
est.gee4 = c(est.gee4, new.gee4[c(1,2,3)])
aex.hat = new.gee2[4]
aar.hat = new.gee3[4]
ast.hat = new.gee4[c(4,5)]
h1 = elr.st(new.gee1[c(1,2,3)], 0, 0)
h2 = elr.st(new.gee2[c(1,2,3)], aex.hat, aex.hat) # hi = c("-2LLR", wts)
h3 = elr.st(new.gee3[c(1,2,3)], aar.hat, aar.hat2)
h4 = elr.st(new.gee4[c(1,2,3)], ast.hat[1], ast.hat[2])
phi.i = 1; new.gee1[2]
qic1 = qicf(new.gee1[c(1,2,3)], 0, phi.i, phi.i, corr="indep")
qic2 = qicf(new.gee2[c(1,2,3)], aex.hat, phi.i, new.gee2[5], corr="exch")
qic3 = qicf(new.gee3[c(1,2,3)], aar.hat, phi.i, new.gee3[5], corr="ar1")
qic4 = qicf(new.gee4[c(1,2,3)], ast.hat, phi.i, new.gee4[6], corr="toep")
cic1 = cicf(new.gee1[c(1,2,3)], 0, phi.i, phi.i, corr="indep")
cic2 = cicf(new.gee2[c(1,2,3)], aex.hat, phi.i, new.gee2[5], corr="exch")
cic3 = cicf(new.gee3[c(1,2,3)], aar.hat, phi.i, new.gee3[5], corr="ar1")
cic4 = cicf(new.gee4[c(1,2,3)], ast.hat, phi.i, new.gee4[6], corr="toep")
elrs = c(h1, h2, h3, h4)
eaic = elrs + 2*p; # extract "-2LLR"
qic = c(qic1,qic2,qic3,qic4)
cic = c(cic1,cic2,cic3,cic4)
print("EAIC"); print(eaic)
```

```
print("QIC"); print(qic)
print("CIC"); print(cic)
id4=which.min(eaic);
min.eaic[id4] = min.eaic[id4]+1;
id5=which.min(qic);
min.qic[id5] = min.qic[id5]+1
id6=which.min(cic);
min.cic[id6] = min.cic[id6]+1
print(j)
}
min.eaic; min.qic; min.cic;
```

D.2 R-Code to Investigate Efficiency Gain in GEE When EQ_{AIC} is Used Compared to QIC Based on Ohio Data Set

```
# OHIO DATA LOADING AND SELECTION OF CORRELATION STRUCTURE
#

data(ohio)
ohio
y = ohio$resp
x = cbind(ohio$age, ohio$smoke, ohio$age:smoke) # the long vector
t =4
n= length(y)/t
p = dim(x)[2]
new.gee1 = mygee(y,x,t,binomial, corr="ind")
est.gee1 = c(est.gee1, new.gee1[c(1,2,3)] )
new.gee2 = mygee(y,x,t,binomial, corr="exch")
est.gee2 = c(est.gee2, new.gee2[c(1,2,3)] )
new.gee3 = mygee(y,x,t,binomial, corr="ar1")
```

```
est.gee3 = c(est.gee3, new.gee3[c(1,2,3)])
new.gee4 = mygee(y,x,t,binomial, corr="toep")
est.gee4 = c(est.gee4, new.gee4[c(1,2,3)])
aex.hat = new.gee2[4]
aar.hat = new.gee3[4]
ast.hat = new.gee4[c(4,5)]

    h1 = elr.st(new.gee1[c(1,2,3)], 0, 0)
h2 = elr.st(new.gee2[c(1,2,3)], aex.hat, aex.hat) # hi = c("-2LLR", wts)
h3 = elr.st(new.gee3[c(1,2,3)], aar.hat, aar.hat^2)
h4 = elr.st(new.gee4[c(1,2,3)], ast.hat[1], ast.hat[2])

    phi.i = 1; # new.gee1[2]
qic1 = qicf(new.gee1[c(1,2,3)], 0, phi.i, phi.i, corr="indep")
qic2 = qicf(new.gee2[c(1,2,3)], aex.hat, phi.i, new.gee2[5], corr="exch")
qic3 = qicf(new.gee3[c(1,2,3)], aar.hat, phi.i, new.gee3[5], corr="ar1")
qic4 = qicf(new.gee4[c(1,2,3)], ast.hat, phi.i, new.gee4[6], corr="toep")

    elrs = c(h1, h2, h3, h4)
eaic = elrs + 2*p; # extract "-2LLR"
qic = c(qic1,qic2,qic3,qic4)

    print("EAIC"); print(eaic)
print("QIC"); print(qic)

        # MODEL SELECTION BASED ON WCS SELECTED BY EQ_AIC #

library(MuMIn)
library(MESS)
data(ohio)
```

```
mm1=geeglm(resp~ 1,family=binomial(link="logit"),id=id, data=ohio, corstr = "exchangeable", std.err="san.se")
mm2=geeglm(resp~age,family=binomial(link="logit"),id=id, data=ohio, corstr = "exchangeable", std.err="san.se")
mm3=geeglm(resp~smoke,family=binomial(link="logit"),id=id, data=ohio, corstr = "exchangeable", std.err="san.se")
mm4=geeglm(resp~age:smoke,family=binomial(link="logit"),id=id, data=ohio, corstr = "exchangeable", std.err="san.se")
mm5=geeglm(resp~age+smoke,family=binomial(link="logit"),id=id, data=ohio, corstr = "exchangeable", std.err="san.se")
mm6=geeglm(resp~age+age:smoke,family=binomial(link="logit"),id=id, data=ohio, corstr = "exchangeable", std.err="san.se")
mm7=geeglm(resp~smoke+age:smoke,family=binomial(link="logit"),id=id, data=ohio, corstr = "exchangeable", std.err="san.se")
mm8=geeglm(resp~age+smoke+age:smoke,family=binomial(link="logit"),id=id, data=ohio, corstr = "exchangeable", std.err="san.se")
model.sel(mm1, mm2, mm3, mm4, mm5, mm6, mm7, mm8, rank=QIC)
```

MODEL SELECTION BASED ON WCS SELECTED BY QIC

```
m1=geeglm(resp~ 1,family=binomial(link="logit"),id=id, data=ohio, corstr = "independence", std.err="san.se")
m2=geeglm(resp~ age,family=binomial(link="logit"),id=id, data=ohio, corstr = "independence", std.err="san.se")
m3=geeglm(resp~ smoke,family=binomial(link="logit"),id=id, data=ohio, corstr = "independence", std.err="san.se")
m4=geeglm(resp~ age:smoke,family=binomial(link="logit"),id=id, data=ohio, corstr = "independence", std.err="san.se")
m5=geeglm(resp~ age+smoke,family=binomial(link="logit"),id=id, data=ohio, corstr = "independence", std.err="san.se")
```

```
m6=geeglm(resp~ age+age:smoke,family=binomial(link="logit"),id=id, data=ohio, corstr
= "independence", std.err="san.se")
m7=geeglm(resp~ smoke+age:smoke,family=binomial(link="logit"),id=id, data=ohio,
corstr = "independence", std.err="san.se")
m8=geeglm(resp~ age+smoke+age:smoke,family=binomial(link="logit"),id=id, data=ohio,
corstr = "independence", std.err="san.se")
model.sel(m1, m2, m3, m4, m5, m6, m7, m8, rank=QIC)


    # Efficiency of the models ranked 1st in each case: K.fold cross-validation #

data(ohio)
ohio
library(tidyverse)
library(modelr)
N=200
for (j in 1:N){
cv = crossv_ kfold(ohio, k = 10)
cv
models1 = map(cv$train, ~geeglm(resp~age+smoke, family=binomial(link="logit"),
id=id,corstr = "exchangeable", std.err="san.se", data = .))
models2 = map(cv$train, ~geeglm(resp~age+smoke, family=binomial(link="logit"),
id=id,corstr = "independence", std.err="san.se", data = .))
get_pred = function(model, test_data){
data = as.data.frame(test_data)
pred = add_predictions(data, model)
return(pred)
}
pred1 = map2_df(models1, cv$test, get_pred, .id = "Run")
pred2 = map2_df(models2, cv$test, get_pred, .id = "Run")
MSE1 = pred1 %>% group_by(Run) %>%
```

summarise(MSE = mean((resp - pred)^2))

$MSE1$

$MSE2 = pred2\%¿\%\text{group\_}by(\text{Run}) \% > \%$

summarise(MSE = mean($(resp - pred)^2$))

MSE2

}

mean(MSE1$MSE)

mean(MSE2$MSE)

Appendix E

R-CODE FOR THE ANALYSIS OF SHAREHOLDER VALUE CREATION DATA

E.1 Selection of Working Correlation Structure

data = read.csv("F:/SOFTWARES/data.csv")

y = data$sva

x = cbind(data$logta, data$DPR2, data$bsize, data$roe, data$wcta, data$g, data$roa, data$z1,data$lev)

t =6

n= length(y)/t

p = dim(x)[2]

gee1 = mygee(y,x,t, binomial, corr="ind")

gee2 = mygee(y,x,t, binomial, corr="exch")

gee3 = mygee(y,x,t, binomial, corr="ar1")

gee4 = mygee(y,x,t, binomial, corr="toep")

phi1 = phi.f(gee1[1:9])

phi2 = phi.f(gee2[1:9])

phi3 = phi.f(gee3[1:9])

phi4 = phi.f(gee4[1:9])

aex.hat = gee2[10]

aar.hat = gee3[10]

ast.hat = gee4[c(10,11,12)]

h1 = elr.st(as.vector(gee1[1:9]), 0,0,0)

h2 = elr.st(as.vector(gee2[1:9]), aex.hat, aex.hat, aex.hat)

h3 = elr.st(as.vector(gee3[1:9]), aar.hat, aar.hat^2, $aar.hat^3$)

$h4 = elr.st(as.vector(gee4[1:9]), ast.hat[1], ast.hat[2], ast.hat[3])$

$qic1 = qicf(as.vector(gee1[1:9]), 0, phi1, phi1, corr = "indep")$

$qic2 = qicf(as.vector(gee2[1:9]), aex.hat, phi1, phi2, corr = "exch")$

$qic3 = qicf(as.vector(gee3[1:9]), aar.hat, phi1, phi3, corr = "ar1")$

$qic4 = qicf(as.vector(gee4[1:9]), ast.hat, phi1, phi4, corr = "stat")$

$pn = c(p, p+1, p+1, p+t-1)$

$elrs = c(h1, h2, h3, h4)$

$qic = c(qic1, qic2, qic3, qic4)$

$eaic = elrs + 2 * p$

$eaic; qic$

E.2 Model Selection for SVA Data

data analysis : Model Selection

library(tidyverse)

library(MuMIn)

library(modelr)

fit.ar=geeglm(formula = sva ∼ logta + DPR2 + bsize + roe + g+roe:g +bsize+ roa +

z1 + lev, family = binomial(link = "logit"),

data = data, id = id, corstr = "ar1", std.err = "san.se")

options(na.action="na.fail")

m2=dredge(fit.ar)

model.sel(m2, rank=QIC) subset(m2,delta<4) summary(get.models(m2, 1)[[1]])

data analysis : Model Validation K-Fold Cross-validation

fit.un=geeglm(formula = sva ∼ logta + DPR2 + bsize + roe + g+roe:g +bsize+ roa

+ z1 + lev, family = binomial(link = "logit"),

data = data, id = id, corstr = "unstructured", std.err = "san.se")

options(na.action="na.fail")

m1=dredge(fit.un)

model.sel(m1, rank=QIC)

subset(m1,delta<6)

model.avg(m1.delta<6)

model.avg(m1,subset=cumsum(weight)<=.95)

summary(get.models(m1, 1)[[1]])

E.3 Establishment of Efficiency of EQ_{AIC} Over QIC

N=10

for (I in 1:10){

cv = $crossv\_kfold$(data, k = 10)

cv

models1 = map(cv$train, ~geeglm(sva~DPR2+g+lev+logta+z1+roe+g:roe,

family=binomial(link="logit"),id=id,corstr = "ar1", std.err="san.se", data = .))

models2 = map(cv$train, geeglm(sva~DPR2+g+lev+logta+z1+roe+bsize,

family=binomial(link="logit"),id=id,corstr = "unstructured", std.err="san.se", data =

.))

get_pred = function(model, $test_data$){

data = as.data.frame($test\_data$)

pred = $add\_predictions$(data, model)

return(pred)

}

pred1 = $map2\_df$(models1, cv$test, get_pred, .id = "Run")

pred2 = $map2\_df$(models2, cv$test, get_pred, .id = "Run")

MSE1 = pred1 % > % $group\_by$(Run) % > %

summarise(MSE = mean($(sva - pred)^2$))

MSE1

MSE2 = pred2 % > % $group\_by$(Run) % > %

summarise(MSE = mean((sva - pred)^2))

$MSE2$

$mean(MSE1/\text{MSE})$

mean(MSE2/MSE)

$RE = mean(MSE2/\text{MSE})/\text{mean}(\text{MSE1}/MSE)$

RE;