# Fisher Consistency of GEE Models Under Link Misspecification

Chongsun Park and Sanford Weisberg
University of Minnesota
School of Statistics
St. Paul, Minnesota 55108-6042
Technical Report No. 602
April 7, 1995

# Fisher Consistency of GEE Models Under Link Misspecification

Chongsun Park and Sanford Weisberg*
University of Minnesota
School of Statistics
St. Paul, Minnesota 55108-6042
Technical Report No. 602

April 7, 1995

**Abstract**

Liang and Zeger (1986) introduced Generalized Estimating Equation or GEE models as a way of using the ideas of generalized linear models when repeated measures are available on subjects or clusters of subjects but normality of the response may not be appropriate. They have shown that consistent estimates of regression coefficients are obtained even if the within cluster or subject correlation is misspecified. In this article we show that under certain common circumstances Fisher consistent estimates of regression coefficients will be obtained even if the link function in the generalized linear model is misspecified.

## 1 Introduction

Generalized estimating equation, or GEE models, were introduced by Liang and Zeger (1986) and Zeger and Liang (1986) to provide a simple methodology for modeling repeated measure or clustered data using models similar to generalized linear models. In their approach, only the conditional mean structure for the response given predictors and the conditional second moment structure needs to be specified. Given these, a multivariate quasi-likelihood (Wedderburn, 1974) can be constructed and used to obtain consistent and asymptotically normal estimates of regression coefficients. Liang and Zeger (1986) show, among other things, that these results hold even if the the correlation structure between the repeated observations is misspecified.

In this article, we consider misspecification in the first moment. As with generalized linear models, the GEE models require that the conditional mean of the response given the predictors depends only on a single linear combination of the predictors and on a link function. We will show that under certain circumstances consistent estimates of

*Work supported by National Science Foundation grant DMS-9208819.

regression coefficients can be obtained even if the link function is not correctly specified. Thus, even fewer assumptions must be satisfied for GEE models to be useful.

We first review the structure of GEE models in Section 2. Section 3 gives the derivation of the result, and the necessary conditions for it to hold. Section 4 illustrates the results numerically.

# 2 Generalized Estimating Equation models

Following Zeger and Liang (1986), we assume that the $i$-th subject produces a $n_i \times 1$ vector of responses $y_i$, $i = 1, \ldots, I$ with typical elements $y_{ij}$. Let $X_i$ be a $n_i \times k$ matrix of covariates for subject $i$ with typical row $x_{ij}^T$.

We need to specify the first two conditional moments of the response given the predictors. Let $\mu_{ij} = E(y_{ij}|x_{ij})$ be the conditional expectation of $y_{ij}$ given $x_{ij}$ and suppose that

$$g(\mu_{ij}) = \beta_0 + x_{ij}^T \beta \tag{1}$$

where $\beta$ is a $k \times 1$ vector of unknown regression coefficients, $\beta_0$ is an intercept, and $g$ is called the link function. For simplicity of presentation, we will assume that the intercept $\beta_0$ is known, and it equals zero. This is not necessary for GEE models, but in the next section when the link function is assumed unknown, the intercept is not estimable because it can be absorbed into the link function.

Specifying a covariance structure requires two steps. First, as in generalized linear models, the assumption of a particular error distribution such as the Poisson suggests that $\text{var}(y_{ij}|x_{ij})$ is a known function of the mean, apart from a possibly unknown positive scale factor $\phi$. This assumption however has nothing to say about correlation structure within a subject, and for this purpose we need to specify a covariance matrix $V_i(\mu)$ for the $i$-th subject. This is given by

$$V_i = V_i(\mu_i, \rho) = A_i^{1/2} R_i(\rho) A_i^{1/2} / \phi \tag{2}$$

where $A_i = A(\mu_i)$ is an $n_i \times n_i$ diagonal matrix with $\text{var}(y_{ij}|x_{ij})$ as the $j$th diagonal element, and $R_i(\rho)$ is an $n_i \times n_i$ matrix of correlations, assumed to be a function only of an $s \times 1$ an unknown parameter vector $\rho$. Liang and Zeger referred to $R_i(\rho)$ as a "working" correlation matrix because they do not expect it to be correctly specified.

Assuming that the between-subject correlation is zero, first and second moments are now completely specified apart from a few unknown parameters. Thus we can now use a multivariate quasi-likelihood approach to estimate the unknown regression coefficients $\beta$ and the nuisance parameters $\rho$ and $\phi$.

The multivariate quasi-likelihood begins with a set of score equations,

$$U = \sum_{i=1}^{I} u_i = \sum_{i=1}^{I} D_i^T V_i^{-1} (y_i - \mu_i) = 0 \tag{3}$$

where $D_i = \partial\mu_i/\partial\beta$ and $\mu_i = (\mu_{i1}, \cdots, \mu_{in_i})^T$. Consider the multiple integral

$$Q(X\beta, y) = \sum_{i=1}^{I} q(X_i\beta, y_i) = \sum_{i=1}^{I} \int_{y_i}^{\mu_i} D_i^T V_i^{-1}(y_i - t_i)dt_i \tag{4}$$

where $X\beta$ is related to the means via (1). In the univariate case, Wedderburn (1974) has shown that $Q(X\beta, y)$ has properties similar to a genuine log-likelihood function for $\beta$ under very mild assumptions; a similar proof can be extended to the multiple integral case. Consequently, estimates of $\beta$ obtained by solving the score equations should have properties similar to maximum likelihood estimates.

Slightly changing notation leads to the generalized estimating equations, given by

$$\sum_{i=1}^{I} D_i^T (A_i^{1/2} R_i(\rho) A_i^{1/2})^{-1} S_i = 0 \tag{5}$$

where $S_i = y_i - \mu_i$. The GEE estimates $\hat{\beta}_R$ are obtained by replacing $\rho$ and $\phi$ in (5) by $I^{1/2}$-consistent estimators $\hat{\rho}$ and $\hat{\phi}$, given $\beta$ known, and solving (5) for $\beta$ and $\phi$.

Under mild regularity conditions, Liang and Zeger (1986, Theorem 2) show that as the number of subjects grows large, $\hat{\beta}_R$ is a consistent estimate of $\beta$ and that given the predictors, $I^{1/2}(\beta_R - \beta)$ is asymptotically Normal with covariance matrix that depends on both the assumed covariance matrix and on the true covariance matrix. All that is required for this methodology to work well is that the matrix $I^{-1}\left(\sum D_i^T V_i^{-1} D_i\right)^{-1}$ converges to a fixed matrix, and $\rho$ and $\phi$ must be estimated consistently, given $\beta$ known. See Liang and Zeger (1986) for more details. Splus and C functions for doing these computations written by V. J. Carey and others are available from `statlib@stat.cmu.edu`.

# 3 Fisher consistency results of parameters under link violation

As noted in the last section, it is not necessary to know the exact correlation structure to get consistent and asymptotically normal estimates of regression coefficients, provided that the link function is correctly specified. In this section, we consider consistency of estimates when the link function is not correctly specified. That consistent estimates are possible with link misspecification may not be obvious, since using the wrong link suggests that the wrong functional form will be fit. This same problem has been addressed in some generality by Li and Duan (1989), but for the case of a univariate response; our result is an application of their result. The fundamental result is that Fisher consistent estimates will be obtained as long as certain conditional expectations *among the predictors* are linear. This condition is always satisfied when predictors are normally distributed, or more generally with elliptically contoured predictors, but the condition may be satisfied at least approximately for many other distributions for the predictors, or for particular values of the regression coefficients. In general, we now assume that $(y_i, X_i)$ are independent samples from some multivariate distribution. This assumption forces all of the number of observations $n_i$ per cluster to be equal, but this is done here only for convenience, as

only a more complicated argument is required with unequal $n_i$.

We now assume that the link function $g$ given by (1) is unknown. This implies that both the intercept $\beta_0$ and the length $||\beta||$ are not estimable, since these can be absorbed into $g$ by location/scale invariance. Consequently, the best we can hope to do is to estimate $c\beta$ for some nonzero constant $c$; for many purposes, such as examining a regression problem graphically (Cook and Weisberg, 1994), or for estimating a link function (Weisberg and Welsh, 1995), an estimate of $c\beta$ is all that is required.

As with specifying the correlation structure, we can choose a "working" link function to do estimation. A good choice is the canonical link for the corresponding generalized linear model. In this case, the quasi-likelihood $Q$ is strictly convex. By the strong law of large numbers,

$$\lim_{I\to\infty} \frac{1}{I}\sum_{i=1}^{I} q(X_i b, y_i) \to \mathrm{E}_{X,y} q(Xb, y) \tag{6}$$

if the expectation is well-defined, and hence by definition the minimizer of (4) will be a Fisher consistent estimate of the parameters that minimize the expectation on the right hand side of (6), Cox and Hinkley (1974, p. 287). minimizer of this expectation is equal to $c\beta$ for some constant $c$.

To obtain this result, we will need to impose the following linearity condition on the distribution of $X$, for all vectors $b$:

$$E(Xb|X\beta) = cX\beta \tag{7}$$

The scalar $c$ may depend on $b$. An intercept is not required because we have assumed without loss of generality that $\mathrm{E}(X_i) = 0$. This condition is satisfied by $X$ normally distributed, and more generally by elliptically contoured $X$. It is likely to hold at least approximately in many problems.

To prove Fisher consistency, we need to show that for all vectors $b$ there is a scalar $c$ such that

$$\mathrm{E}_{X,y} L(Xb, y) \geq \mathrm{E}_{X,y} L(cX\beta, y)$$

Since $L$ is strictly convex, we can use Jensen's inequality,

$$\begin{aligned}
E_{X,y}L(Xb,y) &= E_{X\beta,y}E_{X|X\beta,y}L(Xb,y) \\
&= E_{X\beta,y}E_{X|X\beta}L(Xb,y) \\
&\geq E_{X\beta,y}L(E_{X|X\beta}(Xb),y) \\
&= E_{X\beta,y}L(cX\beta,y) \\
&= E_{X\beta,y}E_{X|X\beta,y}L(cX\beta,y) \\
&= E_{X,y}L(cX\beta,y).
\end{aligned}$$

which is what we have set out to prove.

# 4 Numerical illustration

To obtain Fisher consistent estimates, we need to fit the GEE model using any working correlation structure, and using a canonical link, as long as the condition (7) is satisfied by the predictors. To illustrate this we consider a small example with sample of size 4 for each subject, 40 subjects, exchangeable correlation of 0.3 between observations on the same subject, and normal error distribution. We generated data with quadratic link and used identity link in estimating procedure since it is the canonical link for the normal model. All that will be varied in this example is the distribution of the predictors (either normal or exponential) and the choice of the coefficient vector, either $(1,1)$ or $(1,-1)$. For both the case of normal predictors or the case of exponential predictors with $\beta = (1,1)$, the linearity condition (7) is satisfied, while for exponential predictors with $\beta = (1,-1)$ this condition is not satisfied. We fit the model 20 times, and measured the angle between the true coefficients and the coefficients estimated by the GEE models. The results are summarized in Figure 1.

When the predictors are normally distributed, then the average angle between the GEE estimate and the true value is small. With exponential predictors and $\beta = (1,-1)$ then the estimates are much worse, with an average angle of 26 degrees. The case $\beta = (1,1)$ is similar to the normal as (7) is satisfied.

# 5 Summary

In this article, we have extended the results of Li and Duan (1989) to the multivariate setting of GEE modeling. We have shown that consistent estimates of regression coefficients are possible when both the link function and the correlation structure are misspecified.

A useful next step in statistical analysis might be to try to estimate the link function given the consistent estimate of the regression coefficients. This is discussed in the univariate case by Weisberg and Welsh (1995), who find that a consistent estimate of the link function is possible using an appropriate kernel smoother.

# 6 References


Cook, R.D. and Weisberg, S. (1994) *An Introduction to Regression Graphics.* J. Wiley & Sons, New York.

Li, K.C. and Duan, N. (1989) Regression analysis under link violation. *Ann. Statist.* 17, 1009-1052.

Liang, K.Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13-22.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models.* (2nd ed). Chapman and Hall, New York.

Wedderburn, R.W.M. (1974) Quasilikelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* 61, 439-47.

Weisberg, S. and Welsh, A.H. (1995) Adapting for the missing link. *Ann. Statist.* 23, in press.

Zeger, S.L. and Liang, K.Y. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42, 121-30.

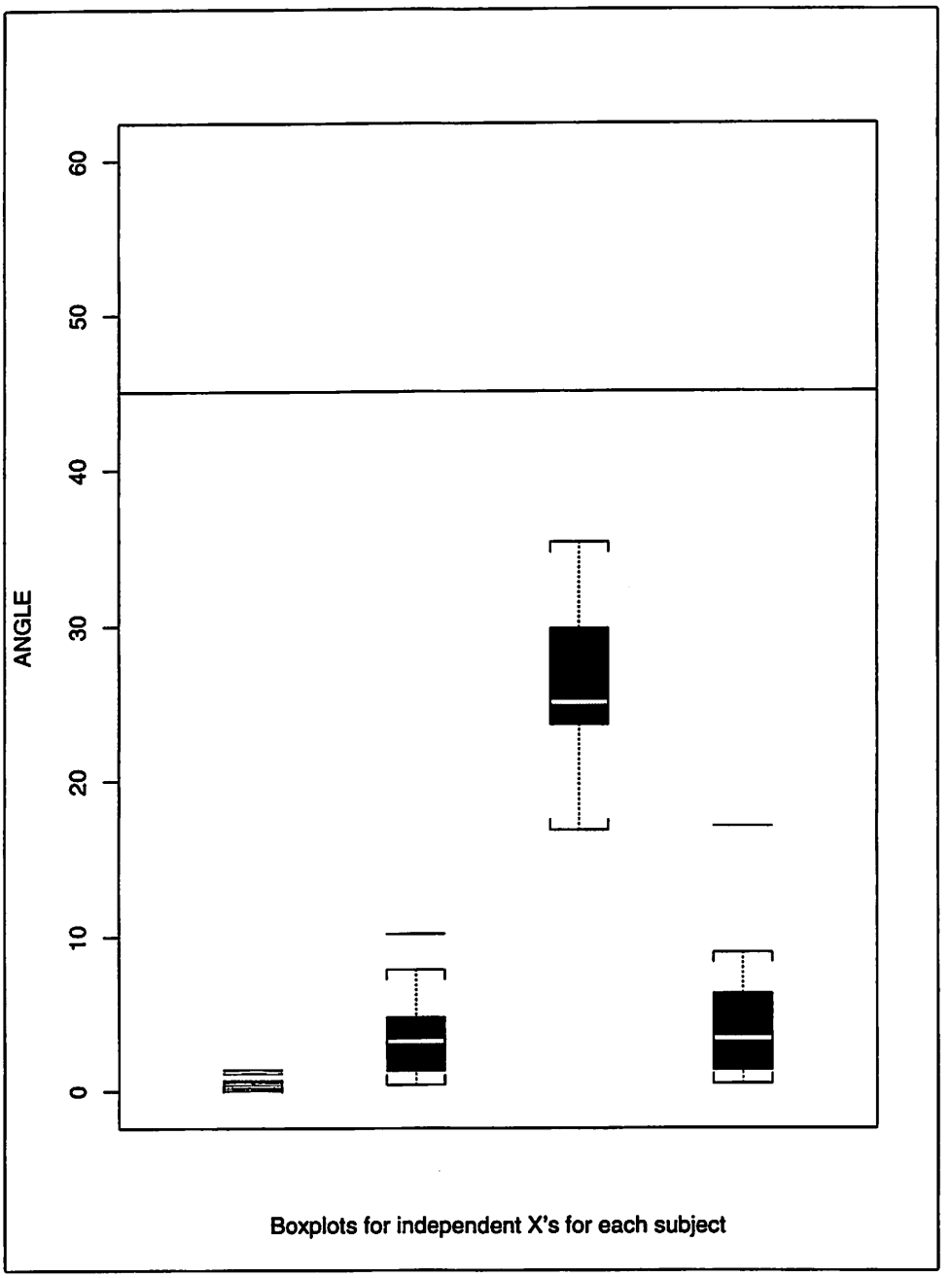


Figure 1: Boxplots for angles between true and estimated $\beta$; Solid line: Expected angle between $\beta$ and a randomly selected direction, so all methods do better than a random choice. Left-most box: Normal $x$, true link is identity, $\beta = (1, 1)$; 2nd box: normal $x$, true link is quadratic, $\beta = (1, 1)$; 3rd box: exponential $x$, true link is quadratic, $\beta = (1, -1)$; 4th box: exponential $x$, true link is quadratic, $\beta = (1, 1)$.