Communications for Statistical Applications and Methods
2023, Vol. 30, No. 4, 355–368

https://doi.org/10.29220/CSAM.2023.30.4.355
Print ISSN 2287-7843 / Online ISSN 2383-4757

ELCIC: An R package for model selection using the empirical-likelihood based information criterion

Chixiang Chen[a], Biyi Shen[b], Ming Wang[1,c]

[a]Division of Biostatistics and Bioinformatics, University of Maryland School of Medicine, USA;
[b]Regeneron Pharmaceuticals, USA;
[c]Department of Population and Quantitative Health Sciences, Case Western Reserve University, USA

Abstract

This article introduces the R package ELCIC (`https://cran.r-project.org/web/packages/ELCIC/index.html`), which provides an empirical likelihood-based information criterion (ELCIC) for model selection that includes, but is not limited to, variable selection. The empirical likelihood is a semi-parametric approach to draw statistical inference that does not require distribution assumptions for data generation. Therefore, ELCIC is more robust and versatile in the context of model selection compared to the currently existing information criteria. This paper illustrates several applications of ELCIC, including its use in generalized linear models, generalized estimating equations (GEE) for longitudinal data, and weighted GEE (WGEE) for missing longitudinal data under the mechanisms of missing at random and dropout.

Keywords: model selection, empirical likelihood, generalized estimating equations, generalized linear models, missing data, R

1. Introduction

Model selection is an essential aspect of data analysis that is used for valid inference or improved prediction. This process encompasses many aspects, such as variable selection in the mean structure, correlation structure selection in longitudinal data analysis, and tuning parameter selection in penalized regressions. However, most existing information criteria critically depend on distribution assumptions and have limited applications in more complicated model selection problems, other than variable selection (Chen and Lazar, 2012). Information criteria like the Akaike information criterion (AIC) (Akaike, 1974), Bayesian information criterion (BIC) (Schwarz, 1978), and generalized information criteria (GIC) (Konishi and Kitagawa, 1996) are likelihood-based and rely heavily on parametric distribution assumptions. They are sensitive to distribution misspecification and data heterogeneity (Chen *et al.*, 2020). In the presence of longitudinal outcomes, some information criteria have been developed under the semi-parametric framework, such as the quasi-likelihood information criterion (QIC) (Pan, 2001), which is widely used in generalized estimation equations (GEE) (Liang and Zeger, 1986). Other examples include the missing longitudinal information criterion (MLIC) (Shen and Chen, 2012), the weighted quasi-likelihood information criterion (QICW) (Gosho, 2016), and the joint longitudinal information criterion (JLIC) (Shen and Chen, 2018). These criteria are

[1] Corresponding author: Department of Population and Quantitative Health Sciences, Case Western Reserve University, OH 44106, USA. E-mail: mxw827@case.edu

Published 31 July 2023 / journal homepage: http://csam.or.kr

applied to longitudinal data with dropout missingness; however, existing criteria suffer from a loss of variable selection power under small sample sizes (Chen *et al.*, 2019). Several R packages are available for model selection criteria, such as the AIC (R function `AIC`) and the BIC (R function `BIC`), which are provided as generic functions in R. The R package `repolr` (Parsons, 2017) provides the QIC (R function `QIC`). Additionally, the R package `wgeesel` (Xu *et al.*, 2019) offers the MLIC (R function `MLIC.gee`) and QICW (R function `QICW.gee`) for longitudinal data analysis.

To minimize the negative impact of distribution misspecification, the empirical likelihood (EL), a data-driven approach that avoids distribution specifications but still borrows likelihood properties (Owen, 1988; Qin and Lawless, 1994) has gained significant attention for data analysis and statistical inference (Owen, 2001). However, EL-based information criteria for model selection are still not well studied. Kolaczyk (1995) first proposed the empirical information criterion (EIC) based on the Kullback-Leibler divergence between discrete empirical distributions, but there was a severe convergence issue. To alleviate the computation issue, Variyath *et al.* (2010) advocated an empirical AIC and an empirical BIC based on the adjusted empirical likelihood by incorporating an additional parameter (Chen *et al.*, 2008). However, these criteria require searching for empirical likelihood estimators, which could be computationally expensive. Most recently, Chen *et al.* (2019) proposed joint empirical Bayesian information criteria (JEAIC and JEBIC) for missing longitudinal data under the mechanisms of missing at random (MAR) and dropout. Later, Chen *et al.* (2020) generalized this criterion and proposed the EL-based consistent information criterion (ELCIC), which targets model selection in a more general context; one that is not limited to variable selection. Specifically, ELCIC was initially derived from the asymptotic expansion of the marginal likelihood in a Bayesian framework, and its consistent model selection property (Shao, 1997; Variyath *et al.*, 2010) was then established in a general context by allowing the use of a regular plug-in estimator to calculate this criterion.

This paper presents the R package `ELCIC`, which includes the proposed EL-based information criterion for model selection and provides a tutorial on its use in widely applied parametric and semiparametric models. We also extend the ELCIC to handle longitudinal data with dropout missingness. The layout of the paper is as follows. Section 2 describes the development of ELCIC. In Section 3, we illustrate the use of core functions for simulated data. Section 4 provides a brief data application of ELCIC to data extracted from the National Institute of Mental Health Schizophrenia Collaborative Study. Finally, we summarize the features of this package and discuss future work in Section 5.

2. Methodology

2.1. Empirical likelihood

Owen (1988) proposed an empirical likelihood approach for parameter estimation and inference. Let $\boldsymbol{D} = \boldsymbol{D}_{i}{}_{i=1}^{n}$ denote the full data, where $\boldsymbol{D}_i = (\boldsymbol{X}_i^T, \boldsymbol{Y}_i^T)^T$ are assumed to be independent and identically distributed (i.i.d.), with $\boldsymbol{Y}_i$ representing the outcomes of interest and $\boldsymbol{X}_i$ including the covariates under consideration, and $i = 1, \ldots, n$. Given that some pre-specified estimating equations of $\boldsymbol{g}(\boldsymbol{D}_i, \boldsymbol{\gamma})$, which have a $p \times 1$ vector of parameters $\boldsymbol{\gamma}$ satisfying $E\boldsymbol{g}(\boldsymbol{D}_i, \boldsymbol{\gamma}_0) = \boldsymbol{0}$ with the true parameters $\boldsymbol{\gamma}_0$, the empirical likelihood ratio is then defined as:

$$R^F = \sup_{\boldsymbol{\gamma}, p_1, \ldots, p_n} \left\{ \prod_{i=1}^{n} n p_i ; p_i \geqslant 0, \ \sum_{i=1}^{n} p_i = 1, \ \sum_{i=1}^{n} p_i \boldsymbol{g}\left(\boldsymbol{D}_i, \boldsymbol{\gamma}\right) = \boldsymbol{0} \right\}. \tag{2.1}$$

In contrast to traditional likelihood, empirical likelihood utilizes point mass probabilities for observations. Thus, information from the data is automatically and efficiently borrowed from the estimating equation constraints (Qin and Lawless, 1994). This desirable property demonstrates the great potential for model selection. Given the estimator denoted by $\hat{\boldsymbol{\gamma}}$, the negative logarithm of the empirical likelihood ratio can be easily calculated based on the Lagrange multiplier method (Owen, 2001) as follows:

$$l = -\log R^F\left(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\gamma}}\right) = \sum_{i=1}^{n} \log\left\{1 + \hat{\boldsymbol{\lambda}}^T \boldsymbol{g}\left(\boldsymbol{D}_i, \hat{\boldsymbol{\gamma}}\right)\right\}, \tag{2.2}$$

where the parameter estimate $\hat{\boldsymbol{\lambda}}$ can be obtained by solving the following equations using the Newton-Raphson method

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\boldsymbol{g}\left(\boldsymbol{D}_i, \hat{\boldsymbol{\gamma}}\right)}{1 + \boldsymbol{\lambda}^T \boldsymbol{g}\left(\boldsymbol{D}_i, \hat{\boldsymbol{\gamma}}\right)} = \mathbf{0}. \tag{2.3}$$

2.2. Description of ELCIC

To implement selection based on empirical likelihood, a full set of estimating functions $\boldsymbol{g}(\boldsymbol{D}_i, \boldsymbol{\gamma})$ is specified with a length (denoted by L) larger than the length of the parameter vector $\boldsymbol{\gamma}$ (defined above with dimension p) (Kolaczyk, 1995; Variyath *et al.*, 2010; Chen and Lazar, 2012). Suppose a plug-in estimator denoted as $\hat{\boldsymbol{\gamma}}EE$ has been obtained from an external estimating equation, and the corresponding Lagrange multiplier estimator $\hat{\boldsymbol{\lambda}}EE$ has been calculated under regular conditions. Then, the proposed ELCIC is defined as

$$\text{ELCIC} = -2\log R^F\left(\hat{\boldsymbol{\lambda}}_{EE}, \hat{\boldsymbol{\gamma}}_{EE}\right) + p\log n. \tag{2.4}$$

One desired property of the proposed ELCIC is that it does not require prior specifications or parametric likelihood. We will now provide three examples of estimating functions $\boldsymbol{g}(\boldsymbol{D}_i; \boldsymbol{\gamma})$ for both variable selection and more general model selection, in terms of data structures.

2.3. Case study I: Generalized linear models

Nelder and Wedderburn (1972) introduced the GLM to unify the theory for different models in data analysis with continuous and categorical outcomes. Under this framework, the full estimating equations $\boldsymbol{g}$ in (2.1) can be simply defined as the score functions

$$\boldsymbol{g}\left(\boldsymbol{D}_i, \boldsymbol{\beta}\right) = \boldsymbol{X}_i\left\{Y_i - \mu_i\left(\tilde{\boldsymbol{\beta}}\right)\right\}, \tag{2.5}$$

where $\mu_i(\tilde{\boldsymbol{\beta}})$ with $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\beta}^T, \mathbf{0}^T)^T$ is the conditional expectation of Y_i modeled by $f^{-1}(\boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}})$ with some pre-specified canonical link function f. As (2.5) is only valid when the mean structure is correctly specified and does not require the second moment, ELCIC under the full estimating equations in (2.5) can handle the scenario when the variance structure is mis-specified, such as with over-dispersion, which is often encountered in the analysis of count data (Variyath *et al.*, 2010; Chen *et al.*, 2020).

2.4. Case study II: Longitudinal data with GEE

For longitudinal data, Liang and Zeger (1986) introduced the marginal model to conduct statistical inference without specifying the joint distribution. A correctly specified mean structure is always the key to estimation consistency. Meanwhile, correctly identifying the “working” correlation structure can further improve the efficiency in GEE. In this case, we specify the full estimating equation so that ELCIC can simultaneously select the marginal mean and the correlation structure for repeated measurements. Existing criteria, such as the QIC (Pan, 2001), cannot handle joint selection.

To achieve joint selection and for simplicity, we assume a balanced design with J observations for each subject. For subject i, the marginal mean is denoted as $\boldsymbol{\mu}_i$ with a variance-covariance matrix $\boldsymbol{V}_i$. The over-dispersion parameter is denoted as ϕ (assumed known but can also be consistently estimated), and the correlation coefficient vector is $\boldsymbol{\rho}^c = (\rho_1^c, \ldots, \rho_{J-1}^c)^T$. Here, the superscript c indicates the type of correlation structure. For instance, under an exchangeable structure, we have $\boldsymbol{\rho}^{\text{EXC}} = (\rho, \ldots, \rho)^T$, and under an Autoregressive (AR1) structure, we have $\boldsymbol{\rho}^{AR1} = (\rho, \ldots, \rho^{(J-1)})^T$. Thus, the full estimating function in (2.1) is defined as

$$\boldsymbol{g}\left(\boldsymbol{D}_i, \boldsymbol{\beta}, \boldsymbol{\rho}^c\right) = \begin{pmatrix} \boldsymbol{H}_i^T \boldsymbol{V}_i^{-1} \left\{\boldsymbol{Y}_i - \boldsymbol{\mu}_i\left(\tilde{\boldsymbol{\beta}}\right)\right\} \\ \boldsymbol{U}_i\left(\tilde{\boldsymbol{\beta}}\right) - \boldsymbol{h}\left(\boldsymbol{\rho}^c\right)\phi \end{pmatrix}, \tag{2.6}$$

where $\tilde{\boldsymbol{\beta}}$ is defined as $(\boldsymbol{\beta}^T, \boldsymbol{0}^T)^T$, $\boldsymbol{H}_i$ denotes the first derivative of $\boldsymbol{\mu}_i$ with respect to $\tilde{\boldsymbol{\beta}}$, and $\boldsymbol{U}_i(\tilde{\boldsymbol{\beta}}) = (U_{i1}(\tilde{\boldsymbol{\beta}}), U_{i2}(\tilde{\boldsymbol{\beta}}), \ldots, U_{i(J-1)}(\tilde{\boldsymbol{\beta}}))^T$ with

$$U_{im}\left(\tilde{\boldsymbol{\beta}}\right) = \sum_{j=1}^{J-m} e_{ij}\left(\tilde{\boldsymbol{\beta}}\right) e_{i,j+m}\left(\tilde{\boldsymbol{\beta}}\right), \text{ for } m = 1, \ldots, J-1. \tag{2.7}$$

Also, e_{ij} represents the standardized residual term $(y_{ij} - \mu_{ij})/\sqrt{v_{ij}}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, J$. Finally, $\boldsymbol{h}(\boldsymbol{\rho}^c)$ is defined as $(\rho_1^c(T-1-p/n), \ldots, \rho_{J-1}^c(1-p/n))^T$. Additional details of the methods are found in Chen *et al.* (2020).

2.5. Case study III: Longitudinal data with WGEE

WGEE is proposed for longitudinal data with dropouts under MAR by incorporating an inverse probability weight (IPW) matrix $\boldsymbol{W}_i$ to adjust for the missing data (Robins *et al.*, 1995). Note that $\boldsymbol{W}_i$ is calculated based on the inverse probability of the observed outcomes, which is defined as the weight matrix with diagonal elements $R_{ij}/\hat{\omega}_{ij}$, $j = 1, \ldots, J$, where $\omega_{ij} = Pr(R_{ij} = 1|\boldsymbol{D}_i)$, and R_{ij} is the indicator that takes a value of 1 when observing the j^{th} outcome of subject i. Note that $\omega_{ij} = \pi_{i1} \times \pi_{i2} \times \cdots \times \pi_{ij}$ where $\pi_{i1} = 1$ (outcomes at baseline are all observed) and $\pi_{ij} = Pr(R_{ij} = 1|R_{i,j-1} = 1, \boldsymbol{D}_i^{(o)})$, $j = 2, \ldots, J$ with $\boldsymbol{D}_i^{(o)}$ as observed data. Given the data $(R_{ij}, \boldsymbol{D}_i)$, π_{ij} can be estimated based on the partial likelihood from a logistic regression of $\sum_{i=1}^n \sum_{j=2}^T R_{i,j-1} \log\{\pi_{ij}(\boldsymbol{\theta})^{R_{ij}}[1 - \pi_{ij}(\boldsymbol{\theta})]^{1-R_{ij}}\}$, where $\boldsymbol{\theta}$ is a $q \times 1$ vector of regression parameters with consistent estimates obtained by solving the corresponding score function. In this dropout missingness mechanism, the full estimating function in (2.1) is defined as

$$\boldsymbol{g}\left(\boldsymbol{D}_i, \boldsymbol{\beta}, \boldsymbol{\rho}^c\right) = \begin{pmatrix} \boldsymbol{H}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{W}_i \left\{\boldsymbol{Y}_i - \boldsymbol{\mu}_i\left(\tilde{\boldsymbol{\beta}}\right)\right\} \\ \boldsymbol{U}_i\left(\tilde{\boldsymbol{\beta}}\right) - \boldsymbol{h}\left(\boldsymbol{\rho}^c\right)\phi \end{pmatrix}, \tag{2.8}$$

with notations consistent with subsection 2.4. Additional details of the methods are found in Chen *et al.* (2019, 2020).

3. Core functions

To better illustrate our package, we summarize the comparison of input and output information between the core functions in Table 1. The details are discussed in the following subsections.

Table 1: Core functions

	`ELCICglm`	`ELCICgee`	`ELCICwgee`
Criteria output	ELCIC/AIC/BIC/GIC	ELCIC	ELCIC
Repeated measurement	No	Yes	Yes
Model selection	Yes	Yes	Yes
Correlation selection	No	Yes	Yes
Missing mechanism	No missing allowed	MCAR	MAR

3.1. Cross-Sectional data

The ELCIC package provides simulated and cross-sectional data (`data(glmsimdata)`) to illustrate the example of a generalized linear model.

```
library(ELCIC)
> # load data
> data(glmsimdata)
> # extract information
> data.glm <- data.frame(y=glmsimdata$y, glmsimdata$x)
```

Note that both the covariate matrix `x` and the response `y` should be a fully observed matrix without missing data.

```
> # each participant has one record
> head(data.glm)
  y intercept        x1          x2           x3
1 2         1   -0.90899697 -0.11401724  0.151977206
2 3         1    0.63963113 -0.38777828  0.090733856
3 0         1    0.04996491 -0.58600762 -0.885259206
4 1         1   -0.29590319  0.39270464  1.320999120
5 1         1   -0.06996223 -0.02191423  1.370276318
6 5         1   -0.22548831 -0.20156386 -0.004214181
```

Suppose we are interested in the mean structure $f(\mu_1) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where $f(\cdot)$ is a known and pre-specified link function. Then, we can specify this mean structure via `models <- list(y∼x1+x2)`. In addition, the function `ELCICglm` is able to produce different criteria based on the mean structure μ_1, such as AIC, BIC, GIC, etc.

```
> models <- list(y~x1+x2)
> output<-ELCICglm(models, data.glm, family=poisson())
ELCIC for glm
************************************************
     ELCIC         AIC        BIC         GIC
  17.11135 1278.38352 1289.49487 1289.90843
************************************************
model 1: y ~ x1 + x2
************************************************
The model selected by ELCIC: y ~ x1 + x2
```

Now suppose we are interested in a sequence of mean models. The function `ELCICglm` is able to output the criteria (ELCIC, AIC, BIC, GIC) for different mean structures. For instance, there are three candidate mean models with $f(\mu_1) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, $f(\mu_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ and $f(\mu_3) = \beta_0 + \beta_1 x_1$, and we are interested in learning which one fits the data the best. We can run the codes below:

```
> models <- list(y~x1, y~x1+x2, y~x1+x2+x3)
> output<-ELCICglm(models, data=data.glm, family=poisson())
ELCIC for glm
************************************************
        model 1    model 2    model 3
ELCIC  105.0823   21.33646   22.81513
AIC   1410.5555 1278.38352 1274.30506
BIC   1417.9631 1289.49487 1289.12019
GIC   1421.3349 1289.90843 1286.97805
************************************************
model 1: y ~ x1
model 2: y ~ x1 + x2
model 3: y ~ x1 + x2 + x3
************************************************
The model selected by ELCIC: y ~ x1 + x2
```

From the output, ELCIC implies that μ_1 has the best fit in terms of the smallest criterion value. Note that AIC cannot theoretically guarantee the consistency of model selection (i.e., capturing the true model exactly) when there are multiple correct candidate models (Shao, 1997; Variyath *et al.*, 2010). In this situation, AIC tends to select a larger model that includes more nuisance variables, while ELCIC is capable of identifying the true model without involving any nuisance variables. In addition to the model demonstrated above, ELCICglm also allows the inclusion of interaction terms (e.g., $x_1 * x_2$) and other types of functions (e.g., $I(x_1^2)$).

3.2. Longitudinal data

The ELCIC package also includes a simulated longitudinal dataset called (data(geesimdata)). We can use this dataset to demonstrate how to use the ELCICgee function to calculate ELCIC values for a given candidate marginal mean model and a pre-specified "working" correlation structure.

```
> # load data
> data(geesimdata)
> # extract information
> id<-geesimdata$id
> data.gee <- data.frame(id=id, y=geesimdata$y, geesimdata$x)
```

Both x and y can have missing values at random. In addition, the total number of observations for each subject and a vector indicating the observation r should be specified.

```
> # intercept should be included in the covariate matrix
> head(data.gee)
  id y intercept           x1          x2         x3
1  1 0         1 -0.948627235  0.76795271  0.2585333
2  1 0         1  0.006342953 -0.58221954  0.2585333
```

```
3  1 0         1  0.147142185  0.20105404  0.2585333
4  2 1         1  0.302887069 -0.26968169 -0.1803438
5  2 6         1  1.366432436  1.46812968 -0.1803438
6  2 2         1  0.027927011 -0.08221224 -0.1803438
> # each participant has three records
> time<-3
> # r is a vector indicating non-informative missingness: 1 for observed reco
rds,
> # and 0 for unobserved records.
> r<-rep(1,length(id))
```

Suppose we are interested in the mean structure $f(\mu_1) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ with three different correlation structures (independence, exchangeable, and ar1, which are the defaults). Then, we can specify the mean structure via `models <- list`(y ∼ x1+x2).

```
> # the outcome is categorical
> family<-poisson()
> candidate.cor.sets<-c("exchangeable","independence","ar1")
> models <- list(y ~x1+x2)
> output<-ELCICgee(models, candidate.cor.sets,data=data.gee,family,r,id,time)
ELCIC for gee
************************************************
exchangeable    independence          ar1
21.22553          42.24994           27.07361
************************************************
model 1: y ~ x1 + x2
************************************************
The mean model selected by ELCIC: y ~ x1 + x2
The correlation structure selected by ELCIC: "exchangeable"
```

Based on the output, we can see that the exchangeable correlation structure has the smallest ELCIC (21.22553), indicating that it is the most reasonable choice.

The `ELCICgee` function is powerful for running the joint selection of marginal mean structure and correlation structure. For example, if we want to compare two mean structures, $f(\mu_1) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ and $f(\mu_2) = \beta_0 + \beta_1 x_1$, across different correlation structures, we can define `models <- list(y∼x1, y∼x1+x2)`, `candidate.cor.sets<-c("exchangeable","independence","ar1")` by running the following codes:

```
> models <- list(y~x1, y~x1+x2)
> candidate.cor.sets<-c("exchangeable","independence","ar1")
> output<-ELCICgee(models, candidate.cor.sets,data=data.gee,family,r,id,time)
ELCIC for gee
************************************************
               model 1   model 2
exchangeable  145.7660  21.22553
independence  134.6632  42.24994
ar1           140.8642  27.07361
```

```
************************************************
model 1: y ~ x1
model 2: y ~ x1 + x2
************************************************
The mean model selected by ELCIC: y ~ x1 + x2
The correlation structure selected by ELCIC: "exchangeable"
```

Based on the above results, we find that the mean structure μ_1 with an exchangeable correlation structure has the smallest ELCIC and therefore might be a better choice than the other options. Note that, not limited to ELCIC, the function QICc.gee provides an output based on QIC, which is not shown here. Additionally, like ELCICglm, ELCICgee also allows for interaction terms and other types of functions.

3.3. Longitudinal data with dropout missingness

This section provides a tutorial for ELCIC-based model selection in the application of longitudinal data with dropout. We downloaded the simulated data (data(wgeesimdata)) to illustrate how to obtain ELCIC from the function ELCICwgee. In contrast to the case in Section 3.2, we need to identify two models: One for the main model of the longitudinal data and another for the probability of observing the outcome, i.e., π_{ij}.

```
> data(wgeesimdata)
> id<-wgeesimdata$id
> data.wgee <- data.frame(y=wgeesimdata$y, wgeesimdata$x, x_mis1=wgeesimdata$
x_mis[,2])
> head(data.wgee, n=10)
    y intercept        x1     x2   x3     x_mis1
1   0         1   0.7701263    0    1   -0.0381310
2  NA         1   0.7701263    1    1   -0.0381310
3  NA         1   0.7701263    2    1   -0.0381310
4   1         1   0.6960051    0    0    0.4464537
5  NA         1   0.6960051    1    0    0.4464537
6  NA         1   0.6960051    2    0    0.4464537
7   0         1   0.5147030    0    1    0.1135577
8   0         1   0.5147030    1    1    0.1135577
9   1         1   0.5147030    2    1    0.1135577
10  1         1   0.8456476    0    1    0.1772856
```

Here, x is the covariate matrix for modeling longitudinal data, and x_mis is the covariate matrix for modeling the missing data mechanism. Notice that in both models, the covariate matrix is recommended to be fully observed. Again, r is a vector which indicates observations that are missing as 0 and observations that are observed as 1. In addition, we need to specify the outcome type and the number of observation times.

```
> # each participant has three records
> time<-3
> # the outcome is binary
> family<-binomial()
> # r is a vector indicating non-informative missingness: 1 for observed reco
rds,
> # and 0 for unobserved records.
> r<-wgeesimdata$obs_ind
```

Suppose we are interested in the mean structure $f(\mu_1) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Then, we are able to obtain ELCIC under different correlation structures.

```
> models <- list(y~x1+x2)
> model_mis<-r~x_mis1
> candidate.cor.sets<-c("exchangeable")
> output<-ELCICwgee(models, candidate.cor.sets,data=data.wgee,
+                          model_mis,family,r,id,time)
ELCIC for wgee
************************************************
exchangeable      independence         ar1
    18.80669          32.38873         20.17050
************************************************
model 1: y ~ x1 + x2
************************************************
The mean model selected by ELCIC: y ~ x1 + x2
The correlation structure selected by ELCIC: "exchangeable"
```

In this example, we learn that the longitudinal data model with the exchangeable correlation structure has the smallest ELCIC and is therefore a better choice. Next, we can run the joint selection for marginal mean structure and correlation structure. To be specific, if we are interested in comparing the mean structure $\mu_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ and the mean structure $\mu_2 = \beta_0 + \beta_1 x_1$, we can use the following code to specify the candidate model. Note that there exists an alternative expression `models <- list(y∼x1, y∼x1+x2)` . The results are summarized below:

```
> models <- list(y~x1, y~x1+x2)
> candidate.cor.sets<-c("exchangeable","independence","ar1")
> output<-ELCICwgee(models, candidate.cor.sets,data=data.wgee,
+                          model_mis,family,r,id,time)
ELCIC for wgee
************************************************
              model 1  model 2
exchangeable 25.00851 18.80669
independence 34.21868 32.38873
ar1          24.88281 20.17050
************************************************
model 1: y ~ x1
model 2: y ~ x1 + x2
************************************************
The mean model selected by ELCIC: y ~ x1 + x2
The correlation structure selected by ELCIC: "exchangeable"
```

In the end, the mean structure μ_1 with the exchangeable correlation structure is favored more because it has the smallest ELCIC. In addition, other functions such as `MLICwgee` (Shen and Chen,

2012) and QICwgee (Gosho, 2016) provide outputs based on MLIC and QICW, respectively. However, the results for these functions are omitted here.

4. Real data application

We provide several real data in the ELCIC package and here we use one as an illustration. The dataset impsdata originally appeared in the National Institute of Mental Health Schizophrenia Collaborative Study (Gibbons and Hedeker, 1994). A total of 386 patients were enrolled in the study, including 293 patients in the treatment group (Drug = 1) and 93 patients in the placebo group (Drug = 0). Each patient had four visits (Weeks 0, 1, 3, and 6). During each visit, the severity of the schizophrenia disorder (IMPS79) was measured, which ranged from 0 to 7. We dichotomized IMPS79 using the threshold of 4 ($Y = 1$ if IMPS79 $\geqslant 4$; otherwise, $Y = 0$) and took the square root of Time (in weeks) along similar lines to Xu *et al.* (2018). We are interested in exploring the marginal association between the risk factors (i.e., drugs, sex) and the response Y. 7.3% of the data is missing due to patient dropout. The missing mechanism should be investigated before model fitting. Xu *et al.* (2018) showed that these data sets reasonably assume the MAR mechanism, as the trajectories behave differently in the treatment and placebo groups and the mechanism of patient dropout does not depend solely on covariates.

```
> data(impsdata)
> id<-impsdata$id
> r<-impsdata$r
> data.real <- data.frame(id=id,y=impsdata$y, impsdata$x)
> head(data.real,n=10)
   id y Intercept     Time Sex Drug Time.Sex Sex.Drug Drug.Time
1   1 1         1 0.000000   1    1 0.000000        1  0.000000
2   1 0         1 1.000000   1    1 1.000000        1  1.000000
3   1 0         1 1.732051   1    1 1.732051        1  1.732051
4   1 1         1 2.449490   1    1 2.449490        1  2.449490
5   2 1         1 0.000000   1    1 0.000000        1  0.000000
6   2 0         1 1.000000   1    1 1.000000        1  1.000000
7   2 0         1 1.732051   1    1 1.732051        1  1.732051
8   2 0         1 2.449490   1    1 2.449490        1  2.449490
9   3 1         1 0.000000   1    1 0.000000        1  0.000000
10  3 0         1 1.000000   1    1 1.000000        1  1.000000
```

For candidate covariates, we considered Time, Sex, Drug, Time*Sex, Sex*Drug, and Drug*Time effects. Additionally, we evaluated several candidate correlation structures, including independence, exchangeable, and AR1. We then used ELCIC for both model selection and correlation structure selection.

```
> # each participant has three records
> time<-4
> # the outcome is binary
> family=binomial()
> models <- list(y~Time, y~Drug,y~Time+Drug, y~Time*Drug,y~Time+Sex+Drug,
+                    y~Time+Sex+Drug+Time:Sex+Sex:Drug+Drug:Time)
> model_mis<-r~Drug+Time+Sex
> candidate.cor.sets<-c("exchangeable","independence","ar1")
> output_ELCIC<-ELCICwgee(models, candidate.cor.sets,data=data.real,model_mis,
family,}
+                         r,id,time)
ELCIC for wgee
************************************************
               model 1  model 2   model 3   model 4  model 5   model 6
exchangeable 105.79577 472.7708 103.99110 110.23972 110.3594 129.00672
independence 223.97685 477.8945 217.88094 222.16051 221.6483 236.57882
ar1           41.89357 371.4877  34.44123  40.40525  39.9565  57.40167
************************************************
model 1: y ~ Time
model 2: y ~ Drug
model 3: y ~ Time + Drug
model 4: y ~ Time * Drug
model 5: y ~ Time + Sex + Drug
model 6: y ~ Time + Sex + Drug + Time:Sex + Sex:Drug + Drug:Time
****************
The mean model selected by ELCIC: y ~ Time + Drug
The correlation structure selected by ELCIC: "ar1"
```

We now compare the result of ELCIC to the values from MLIC and QICW.

```
> output_MLIC<-MLICwgee(models, candidate.cor.sets,data=data.real,
+                         model_mis,family,r,id,time)
MLIC for wgee
************************************************
          ar1
model 1 260.8
model 2 321.3
model 3 255.0
model 4 255.2
model 5 255.7
model 6 256.8
************************************************
model 1: y ~ Time
model 2: y ~ Drug
model 3: y ~ Time + Drug
model 4: y ~ Time * Drug
model 5: y ~ Time + Sex + Drug
model 6: y ~ Time + Sex + Drug + Time:Sex + Sex:Drug + Drug:Time
************************************************
The mean model selected by MLIC: y ~ Time + Drug
The correlation structure selected by MLIC: "ar1"
> output_QICW<-QICWwgee(models,
candidate.cor.sets,data=data.real,
+                         model_mis,family,r,id,time)
```

```
QICW for wgee
************************************************
           ar1
model 1 1549.7
model 2 1871.3
model 3 1525.7
model 4 1525.8
model 5 1528.8
model 6 1533.7
************************************************
model 1: y ~ Time
model 2: y ~ Drug
model 3: y ~ Time + Drug
model 4: y ~ Time * Drug
model 5: y ~ Time + Sex + Drug
model 6: y ~ Time + Sex + Drug + Time:Sex + Sex:Drug + Drug:Time
************************************************
The mean model selected by QICW: y ~ Time + Drug
The correlation structure selected by QICW: "ar1"
```

The above results show that ELCIC selects the third model with the autocorrelation structure as the best, which is consistent with the MLIC and QICW results.

5. Discussion

This article provides details on the core functions of the R package ELCIC and illustrates how to apply these functions through three case studies. The package has a broad scope of applications in model selection across different data structures, and it includes commonly used criteria for comparison in each setup. The ELCIC framework offers a data-driven approach to model selection that overcomes limitations of classic EL-based criteria by relaxing the estimation procedure. The approach can be extended to fit model selection needs in practical settings where no existing information criterion fits well. Extensive numerical studies conducted by Chen *et al.* (2019) and Chen *et al.* (2020) have shown that ELCIC outperforms existing information criteria and is computationally efficient. ELCIC has demonstrated robustness and power in the presence of highly complex and diverse data. Our package offers a user-friendly approach to a wide range of applications. In the future, we plan to incorporate additional functions and applications in ELCIC to address more model selection problems with additional options of link functions and intermittent missingness in outcomes, missing covariates (Chen *et al.*, 2010, 2021), informative cluster size (Bible *et al.*, 2016; Shen *et al.*, 2022), time-to-event outcomes (Hickey *et al.*, 2016), multivariate outcomes with primary and secondary interests (Chen *et al.*, 2022), among others. As a starting point, ELCIC shows potential and robustness in exploring statistical issues related to model selection with highly complex and diverse data.

References

Akaike H (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**, 716–723.

Bible J, Beck JD, and Datta S (2016). Cluster adjusted regression for displaced subject data, *Biomet-*

rics, **72**, 441–451.

Chen B, Yi GY, and Cook RJ (2010). Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random, *Journal of the American Statistical Association*, **105**, 336–353.

Chen C, Han P, and He F (2022). Improving main analysis by borrowing information from auxiliary data, *Statistics in Medicine*, **41**, 567–579.

Chen C, Shen B, Liu A, Wu R, and Wang M (2021). A multiple robust propensity score method 8 for longitudinal analysis with intermittent missing data, *Biometrics*, **77**, 519–532.

Chen C, Shen B, Zhang L, Xue Y, and Wang M (2019). Empirical-likelihood-based criteria for model selection on marginal analysis of longitudinal data with dropout missingness, *Biometrics*, **75**, 950–965.

Chen C, Wang M, Wu R, and Li R (2020). A robust consistent information criterion for model selection based on empirical likelihood, *Statistica Sinica*, **32**, 1205–1223.

Chen J and Lazar NA (2012). Selection of working correlation structure in generalized estimating equations via empirical likelihood, *Journal of Computational and Graphical Statistics*, **21**, 18–41.

Chen J, Variyath AM, and Abraham B (2008). Adjusted empirical likelihood and its properties, *Journal of Computational and Graphical Statistics*, **17**, 426–443.

Gibbons RD and Hedeker D (1994). Application of random-effects probit regression models. *Journal of Consulting and Clinical Psychology*, **62**, 285–296.

Gosho M (2016). Model selection in the weighted generalized estimating equations for longitudinal data with dropout, *Biometrical Journal*, **58**, 570–587.

Hickey GL, Philipson P, Jorgensen A, and Kolamunnage-Dona R (2016). Joint modelling of time-to-event and multivariate longitudinal outcomes: Recent developments and issues, *BMC Medical Research Methodology*, **16**, 1–15.

Kolaczyk ED (1995). An information criterion for empirical likelihood with general estimating equations, *Department of Statistics, University of Chicago*.

Konishi S and Kitagawa G (1996). Generalised information criteria in model selection, *Biometrika*, **83**, 875–890.

Liang K-Y and Zeger SL (1986). Longitudinal data analysis using generalized linear models, *Biometri ka*, **73**, 13–22.

Nelder J and Wedderburn R (1972). Generalized linear models, *Journal of the Royal Statistical Society. Series A*, **135**, 370–384.

Owen AB (1988). Empirical likelihood ratio confidence intervals for a single functional, *Biometrika*, **75**, 237–249.

Owen AB (2001). *Empirical Likelihood* (2nd ed), CRC Press, London.

Pan W (2001). Akaike's information criterion in generalized estimating equations, *Biometrics*, **57**, 120–125.

Parsons N (2017). Repolr: An R package for fitting proportional-odds models to repeated ordinal scores, Avalible from: `https://CRAN.R-project.org/package=repolr`

Qin J and Lawless J (1994). Empirical likelihood and general estimating equations, *The Annals of Statistics*, **22**, 300–325.

Robins JM, Rotnitzky A, and Zhao LP (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association*, **90**, 106–121.

Schwarz G (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**, 461–464.

Shao J (1997). An asymptotic theory for linear model selection, *Statistica Sinica*, **7**, 221–264.

Shen B, Chen C, Chinchilli VM, Ghahramani N, Zhang L, and Wang M (2022). Semipara metric marginal methods for clustered data adjusting for informative cluster size with nonignorable zeros, *Biometrical Journal*, **64**, 898–911.

Shen CW and Chen YH (2012). Model selection for generalized estimating equations accommodating dropout missingness, *Biometrics*, **68**, 1046–1054.

Shen CW and Chen YH (2018). Joint model selection of marginal mean regression and correlation structure for longitudinal data with missing outcome and covariates, *Biometrical Journal*, **60**, 20–33.

Variyath AM, Chen J, and Abraham B (2010). Empirical likelihood based variable selection, *Journal of Statistical Planning and Inference*, **140**, 971–981.

Xu C, Chinchilli VM, and Wang M (2018). Joint modeling of recurrent events and a terminal event adjusted for zero inflation and a matched design, *Statistics in Medicine*, **37**, 2771–2786.

Xu C, Li Z, Xue Y, Zhang L, and Wang M (2019). An r package for model fitting, model selection and the simulation for longitudinal data with dropout missingness, *Communications in Statistics Simulation and Computation*, **48**, 2812–2829.

Received October 5, 2022; Revised May 12, 2023; Accepted May 12, 2023