The Development of Web-based Graphical User Interface for Unified Modeling Data with Multi (Correlated) Responses

I Made Tirta¹, Dian Anggraeni¹

¹ Mathematics of Department, University of Jember, Jember, Indonesia
E-mail: itirta.fmipa@unej.ac.id

Abstract. Statistical models have been developed rapidly into various directions to accommodate various types of data. Data collected from longitudinal, repeated measured, clustered data (either continuous, binary, count, or ordinal), are more likely to be correlated. Therefore statistical model for independent responses, such as Generalized Linear Model (GLM), Generalized Additive Model (GAM) are not appropriate. There are several models available to apply for correlated responses including GEEs (Generalized Estimating Equations), for marginal model and various mixed effect model such as GLMM (Generalized Linear Mixed Models) and HGLM (Hierarchical Generalized Linear Models) for subject specific models. These models are available on free open source software R, but they can only be accessed through command line interface (using scrit). On the othe hand, most practical researchers very much rely on menu based or Graphical User Interface (GUI). We develop, using Shiny framework, standard pull down menu Web-GUI that unifies most models for correlated responses. The Web-GUI has accommodated almost all needed features. It enables users to do and compare various modeling for repeated measure data (GEE, GLMM, HGLM, GEE for nominal responses) much more easily through online menus. This paper discusses the features of the Web-GUI and illustrates the use of them. In General we find that GEE, GLMM, HGLM gave very closed results.

1 Introduction
Regression are among statistical methods which are frequently employed to model dependences between responses and several explanatory variables. Regression or Statistical models have been developed into various directions to handle various type and complex relationship of data. Data collected from multilevel, repeated measure or longitudinal studies are more likely to be correlated, hence various models for independent responses such as GLM, GAM, GAMLSS [1] are not appropriate. Rich variety of model for handling correlated or clustered data are available on the open source software R. They can be classified into 2 main types, namely marginal model for modeling population average, such as GEE (Generalized Estimating Equation), and the other is conditional model for modeling individual, such as GLMM (Generalized Linear Mixed Model) and HGLM (Hierarchical Generalized Linear Model). The models are much more complex than those for independent responses, since we must take into account the correlation or the clustering. Moreover, these advanced statistical models, can only be accessed using script or command line interface. Therefore it is essential to build users friendly interface that unify the various models and can be more easily accessed by nonstatistician researchers. The kind of interface aims to “closing the gap between methodologist and end user of R” [2]

GEE was first introduced in 1986 [3], and later is also called GEE order 1 (GEE1). This GEE can be regarded as a multivariate extension of GLM (Generalized Linear Model). GLM and GEE have fssmilar class of family distributios. In addition to GLM, GEE also offer several correlation
structures to account for correlation among the responses. Therefore the variance-covariance matrices are generally not diagonal matrices. The first GEE (GEE1) can be written as
\[ S(\beta) = \sum_{i=1}^{n} X_i^T \left( \frac{\partial g(\mu_i)}{\partial \eta_i} \right) \text{Var}(Y_i)^{-1}(Y_i - \mu_i) \text{ with } g(\mu) = \eta = x_i \beta \]  
for link function \( g \). Later, not only the mean parameters, are modelled, but also the scale parameter and the correlation\[4\]. This type of GEE is later popular as GEE order 2 (GEE2). With GEE 2, in addition to link function \( g_1 \) for the mean, there are also other two link functions \( g_2 \) and \( g_3 \) related to the scale and correlation parameters. The link functions are:
\[ g_2(\phi_i) = \eta_2 = x_{\phi_3} \gamma \text{ and } g_3(\rho_i) = \eta_3 = x_{\rho_3} \alpha \]  
where \( X_2 \) and \( X_3 \) are respectively vectors of predictors for the scale and correlation parameters. The GEE model has been implemented into R through geepack package \[5\]. GEE has also been extended for multinomial and ordinal responses (such as level of sickness of patient during time of study, for example weekly for 1 month) through multgee package \[6\].

The Generalized Linear Mixed Models handle the multi-response by accommodating random effects, and the model are called mixed models since they consist of both fixed effects (related to mean) and the random effects. The family distribution for the matrix responses \( Y \) follows the family for GLM, however the distribution of random effects are limited to Gaussian distribution. The model for the random effects \( u \) can be written as
\[ g(\mu_i) = \eta_i = X_i \beta + Z_i u_i \]  
The GLMM has been implemented into R through lme4 package that can also handle multiple random effects \[7\].

The later model for random effect is HGLM where the distribution of random effect is not limited to Gaussian distribution but also other distributions, either conjugate or nonconjugate distributions (including Gaussian, Gamma, Beta, Inverse-Gamma). HGLM was introduced in 1996 \[8\] and has been implemented into R through hglm package \[9\]. The package already accommodate multiple random effects, and model for dispersion, however random effect only applied on modeling the mean. The relationship among the models are as follows: (i) GEE2 with fixed scale parameter and identity correlation link is equivalent to GEE1, while (ii) HGLM with constant dispersion parameter and Gaussian random effects is equivalent to GLMM.

2 Methods

The interface is mainly built by following shiny framework \[13\] and in this interface focus of pulldown-menu type (utilizing two files ui.r and server.r). This framework has been successfully applied for developing unified regression for independent (uni) response\[1\]. The web-GUI consists of 3 main parts: (i) The input and exploration of data which is mainly supported by graphics visualization; (ii) the building and fitting of model, (iii) checking goodness of fit of the model in order to get the most appropriate model (mainly using AIC, BIC, QAIC, QLKL). The interface is also enriched with various high quality of graphics to visualize data exploration. R has many graphics packages which can be utilized to visualize data exploration such as correlations (corrplot \[10\], pair.plot \[11\]), spline smoothers (ggplot2 \[12\]). The development involves several steps including (i) identifying variety of input-output options, (ii) identifying related (needed packages), (iii) defining (writing) user interface script, (iv) defining (writing) related server script, (v) testing the whole (integrated) interface. The detail of components and related packages are given in Table 1.

| No | Components/Model | Input Option | Output Option | Related packages |
|----|------------------|--------------|---------------|------------------|

Table 1. Summary of Features of the Web-GUI

2
| No | Components/Model | Input Option | Output Option | Relate packages |
|----|------------------|--------------|---------------|----------------|
|    | Components/Parts |              |               |                |
| 1  | Input Data       | • Internal database | • Summary of data |                |
|    |                   | • Import data (.csv, .txt) | • List of data |                |
| 2  | Exploratory Data | • General exploration | • Correlations matrix | corrplot, psych, car |
|    |                   | • Correlation among cluster within responses, | • Correlation Diagram |                |
|    |                   | • Correlation among numeric variables | • Scatter plot matrix |                |
| 3  | GEE1             | • Xs and Y | • Estimates (with robust or sandwich and naive standard error and p-values), | gee |
|    |                   | • Correlation structure | • QIC | MuMIn |
|    |                   | • id for cluster identity | | |
|    |                   | • Family distributions and link functions | | |
| 4  | GEE2             | • as GEE1 plus | • Estimates (with sandwich standard error and p values) | geepack |
|    |                   | • predictor and link for scale parameter | | |
|    |                   | • link for correlation | | |
| 5  | GEE for ordinal responses | • as GEE1 except Y on ordinal scale | • sandwich parameters estimates for each category | multgee |
|    |                   | • correlation structure | • p values | |
| 5  | GLMM             | • Y and predictors for the mean (fixed effect parameters) | • Estimates (with p-values), | lme4 |
|    |                   | • family distributions for the response | • prediction for random effects | |
|    |                   | • predictors for random effect | • AIC, BIC | |
|    |                   | • alternative formula script for complex formula (See Fig 1) | • Graphics | |
| 6  | HGLM             | • as GLMM plus | • Estimates for parameters related to the mean (with p-values), | hglm |
|    |                   | • Family (link): for random effect | • lrt (log ratio test) | |
|    |                   | • alternative formula script for complex formula | | |

GEEs do not originally provide goodness of fit such as AIC or BIC. However, the quasi version, QIC can be computed through MuMIn package [14]
3 Results and Discussions

3.1 General Features of the Web GUI interface

The main aim of the Web-GUI is to enable user to GEES, and Mixed Models via menus offered via web (but as complete as if they are done via scripting). At this stage, we have developed online statistical model fitting for independent responses, covering several models described previously with general features as follows (see Figure 1).

1) Data Input: two main options available: (i) select some internally provided data (for practical purposes), (ii) import users’ own data with csv or text format (for real data analyses).
2) Data exploration: several graphical representations on relationship among variables (various correlation diagram and scatter plot diagram, some with smoother).
3) Model options: GEE1, GEE2, GEE multinom, GLMM, HGLM
4) Input-output options: as previously given in Table 1. For complex model (GLMM and HGLM), there are also option to combine menu with additional formula model through script input (Figure 1).
5) Goodness Of Fit (GOF): GEE1 and GEE2 fit are accompanied (QIC), GLMM and HGLM are accompanied by conditional-AIC, BIC.

![Figure 1. Web Appearance and The Main Menu and GLMM menu. In addition to menu, there is also space for short formula script.](image)

3.2 Numerical Illustrations

The purposes of the numerical illustration are: (i) to ensure that the main components of the interface are integrated and work well, (ii) to show, in term of the numerical calculation of similar model (between GEEs and between Mixed models), if they are comparable. The numerical illustration uses simulated data with time constant covariates (during repeated measures of Test scores Y (SHB), the observed data of Xs remain the same). The data are presummably taken from 40 subjects (id=individu) with 3 factors, 2 numerical predictor. The response is on 5 types of Test (each individu were tested/measured 5 times). The summary of the statistics are given below.

| Sekolah | Gender | Tes | id | STPA | SUN |
|---------|--------|-----|----|------|-----|
| A:50    | L:100  | T1:40 | Min. : 1.00 | Min. : 33.66 | Min. : 16.46 |
| B:50    | P:100  | T2:40 | 1st Qu.:10.75 | 1st Qu.: 63.26 | 1st Qu.: 63.74 |
| C:50    |        | T3:40 | Median :20.50 | Median : 70.13 | Median : 73.81 |
| D:50    |        | T4:40 | Mean :20.50   | Mean : 73.07   | Mean : 72.48 |
|         |        | T5:40 | 3rd Qu.:30.25 | 3rd Qu.: 82.24 | 3rd Qu.: 87.25 |

SHB          BS

Max. :40.00 Max. :104.18 Max. :116.94
3.2.1 Data Exploration

We proceed with general exploration to have ideas of the correlation and distribution of the responses, then we do alternative modellings using available models (except for multinomial) and compare the results. In addition to exploring data conventionally to get numerical information of summary statistics. In this stage users can also explore and visualize correlation, both correlation among all numeric variables (Figure 2), and correlation among repeated measures response (Figure 3). Various visualization of the correlation among repeated observations help users to choose reasonable type of candidate for correlation structures.

![Diagram Korrelasi](image)

**Figure 2.** Correlation among 3 numeric variables both graphics describe the same things, except on the right hand side presents combination of correlation, scatter plot matrix and density of the variables.

![Diagram Korrelasi](image)

**Figure 3.** Correlation among 5 repeated measures responses (a) and natural spline between one predictor (STPA) and 5 responses SHB (b)
### Alternatives Model Fittings

We apply 4 models (GEE1, GEE2, GLMM, and HGLM). The main results are given in Table 2.

In general, the web GUI has already accomodate mainly needed features, for data exploration, input and output (estimation and GOF), for the 4 modeling GEE1, GEE2, multgee, GLMM and HGLM. With the unified or integrated web-GUI, users can be more easily fitting and comparing several models and choose the most appropriate one. QAIC has taken into account the correlation structures of GEE. This will help practical researchers alot, since these types of modelings are not available on well known RGUI (RCommander [15]), neither on its original, nor on its extended (pug.in) version. More over, when it is provided on the server, the users even do not need to install anything on their computers, and they can access the modeling by using various gadgets such as tablets and smartphones. In addition, for modeling with GLMM and HGLM, complex model formula can also be inputed as a short script (feature taken from shinyAce package [16]). The menu tree is given in Appendix A.

In term of numerical illustration, after checking various combination of predictors, correlation structures and family distributions, we find that the best GEE1 model is M4, one with STPA as the only predictor and AR-1 correlation structure (with correlation parameter $\alpha=0.9369637$ and $QIC=1247.155$). QIC for GEE has not taken into account the dispersion modeling in GEE2. However, although, the QIC values does not change, the fixed parameters estimates (as well as their standard errors), actually change slightly as we change the dispersion model with GEE2.

Choosing the id (individu) as random effects, we find that the values of the fixed parameters estimates (even compare to those of GEEs) are in general comparable (not noticeably different). The estimates of GLMM are very similar (as expected) to those of HGLM with Gaussian random effects. This mixed model has $cAIC = 1500.123$, still slighty better than applying random effect with Gamma distribution ($cAIC = 1504.259$). Therefore in general, the estimates of the 4 types of modeling are comparable (similar). The comparison of the GOF are given in Table 2 and the more detail output are given in Appendix B). While GEEs can be considered as average models, the mixed model can be considered as subject spesific models. That is, in this case, every individu has each own intercept.

| No | Type of Model | Family | Predictors | Corr.Struct. | GOF Type | value |
|----|---------------|--------|------------|--------------|----------|-------|
| M1 | GEE1          | Gaussian | Sekolah,    | Exchangeable | QIC      | 1275.89 |
|    |               |         | Tes,       |              |          |       |
|    |               |         | Gender,    |              |          |       |
|    |               |         | STPA, SUN  |              |          |       |
| M2 | GEE1          | Gaussian | STPA, SUN  | Exchangeable | QIC      | 1256.35 |
| M3 | GEE1          | Gaussian | STPA, SUN  | AR1          | QIC      | 1251.723 |
| No | Type of Model | Family | Predictors | Corr.Struct. | GOF Type | GOF Value |
|----|--------------|--------|------------|--------------|----------|-----------|
| M4 | GEE1         | Gaussian | STPA       | AR1          | QIC      | 1247.155  |
|    |              |         |            |              |          | 0.9369637 |
| M5 | GEE2         | Gaussian | STPA       | AR1          | QIC      | 1247.155  |
|    | (Dispersion ~ SUN) |         |            | 0.888199     |          |           |
| M6 | (Dispersion ~ STPA) | Gaussian | STPA       | 0.8944808    | QIC      | 1247.155  |
|    |              |         |            |              |          |           |
| M7 | GLMM (id=RE) | Gaussian | STPA       | -            | cAIC     | 1585.787  |
| M8 | HGLM (RE Gauss) | Gaussian | STPA       | -            | cAIC     | 1500.123  |
| M9 | HGLM         | Gaussian | STPA       | -            | cAIC     | 1504.259  |

3.3 Future Development
The developed web-GUI has covered almost all parametric modelings for correlated responses. However, it is natural to think if these parametric models can be combined with nonparametric or additive model (GAM). The simplest way to include the spline smoother is by including either natural spine or b-spline especially in the fixed model (see [17]). A more comprehensive approach has been done by combining Aditive and Mixed model and known as GAMM (Generalized Aditive Mixed Model) and has been implemented in gamm4 package [18]. However, yet at the moment, GAMM has not been included in the current interface, but it will be considered in the next improvement, so that the users have more modeling choices, for similar data.

4 Conclusion
Web-based Graphical User Interface for modelling data with multiresponses using R based on R-shiny toolkit framework has been developed. The interface has accomodated main features of GEE1, GEE2, GEE for repeated Ordinal/multinomial responses, GLMM and HGLM. For complex models in GLMM and HGLM model formula can also be alternatively inputed via short script. This interface will help end R users alot, that they can now access the most advanced statistical modelings more easily. However, the interface still can be extended to include GAMM.
References

[1]. Tirta I M, Anggraeni D, Pandutama M 2017 Online Statistical Modeling (Regression Analysis) for Independent Responses IOP Conf. Series: Journal of Physics: Conf Series 855 012054

[2]. Wallace B C, Dahabreh I J, Trikalinos T A, Lau J, Trow P and Schmid C H 2012 Closing the Gap between Methodologists and End-Users: R as a Computational Back-End Journal of Statistics Softwares 49 (5)

[3]. Liang K Y, Zeger S L 1986 Longitudinal data analysis using generalized linear models Biometrika 73(1) 13–22

[4]. Yan J and Fine J P 2004 Estimating Equations for Association Structures Statistics in Medicine 23 859-880

[5]. Højsgaard S, Halekoh U and Yan J 2006 The R Package geepack for Generalized Estimating Equations Journal of Statistical Software 15(2) 1-11

[6]. Touloumis A 2015 R Package multgee: A Generalized Estimating Equations Solver for Multinomial Responses Journal of Statistical Software 64(8) 1-14

[7]. Bates D, Maechler M, Bolker B and Walker S 2015 Fitting Linear Mixed-Effects Models Using lme4 Journal of Statistical Software 67(1) 1-48

[8]. Lee Y and Nelder J A 1996 Hierarchical generalised linear models (with discussion). Journal of the Royal Statistical Society B 58 619-656

[9]. Ronnegard L, Shen X and Alam M 2010 hglm: A Package for Fitting Hierarchical Generalized Linear Models. The R Journal, 2(2): 20-28. URL http://journal.r-project.org/archive/2010-2/RJournal_2010-2_Roennegaard~et~al.pdf.

[10]. Taiyun Wei and Viliam Simko (2016). corrplot: Visualization of a Correlation Matrix. R package version 0.77. https://CRAN.R-project.org/package=corrplot

[11]. Revelle W 2016 psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.6.9.

[12]. Wickham H 2009 ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York)

[13]. Chang W, Cheng J, Allaire J J, Xie J and McPherson J 2017 shiny: Web Application Framework for R. R package version 1.0.5. https://CRAN.R-project.org/package=shiny

[14]. Barton K 2017 MuMIn: Multi-Model Inference. R package version 1.40.0. https://CRAN.R-project.org/package=MuMIn

[15]. Fox J 2017 Using the R Commander: A Point-and-Click Interface or R Boca Raton FL: Chapman and Hall/CRC Press.

[16]. Trestle Technology LLC 2016 shinyAce: Ace Editor Bindings for Shiny R package version 0.2.1. https://CRAN.R-project.org/package=shinyAce.

[17]. Tirta I M, Anggraini D and Octaviani L C 2016 Online and Interactive Web For Fitting GEES With Natural Splines For Longitudinal Data. IBSC:International Basic Science Conference. FMIPA Universitas Jember. 26-27 September 2016. URL: http://statslab-rshiny.fmipa.unej.ac.id/JORS/GEF/ (revised and submitted for Pertanika)

[18]. Wood S and Scheipl F 2017 gamm4: Generalized Additive Mixed Models using 'mgcv' and 'lme4'. R package version 0.2-5. https://CRAN.R-project.org/package=gamm4