Special issue of the 2nd International Conference on Computational and Experimental Science and Engineering (ICCESEN 2015)

Generalized Linear Models for European Union Countries Energy Data

N. İyit*, H. Yonar and A. Genç
Selcuk University, Faculty of Science, Department of Statistics, 42031, Alaeddin Keykubat Campus, Konya, Turkey

The class of generalized linear models is an extension of traditional linear models that allows the mean of the response variable to be linearly dependent on the explanatory variables through a link function. Generalized linear models allow the probability distribution of the response variable to be a member of an exponential family of distributions. The exponential family of distributions include many common discrete and continuous distributions such as normal, binomial, multinomial, negative binomial, Poisson, gamma, inverse Gaussian, etc. Also link functions can be built as identity, logit, probit, power, log, and complementary log–log link functions. In this study, supply, transformation and consumption, imports and exports of solid fuels, oil, gas, electricity, and renewable energy annual data of European Union countries between 2005 and 2013 years are investigated by using generalized linear models. In this case, the response variable is taken as annual complete energy balances of European Union countries as a continuous variable having positive values, and the distribution of the response variable comes from the gamma distribution with log–link function.

DOI: 10.12693/APhysPolA.130.397
PACS/topics: 02.70.Rr, 02.50.–r, 02.50.Sk

1. Introduction

Generalized linear models (GZLMs) are first introduced by Nelder and Wedderburn [1]. The class of GZLMs allows the mean of a population to depend on a linear predictor through a nonlinear link function and allows the response probability distribution to be any member of an exponential family of distributions [2]. In recent years, the class of GZLMs has gained popularity as a statistical modeling tool.

Generalized estimating equation (GEE) approach introduced by Liang and Zeger [3] is a widely used statistical method in the analysis of longitudinal data as an extension of GZLM for correlated data [4, 5]. GEE approach specifies how the average of a response variable of a subject changes with predictors (factors and covariates) while allowing for the correlation between repeated measurements on the same subject over time. To take account of this correlation, a specification of a working correlation structure is required in GEE. GEE approach is based on the quasi-likelihood function given by Wedderburn [6] and no restrictive assumption is made about the distribution of the response variable [7].

McCullagh and Nelder [8], Firth [9], Blough et al. [10], Lindsey [11], Dobson and Barnett [12], Agresti [13], Hardin and Hilbe [14], Lipsitz et al. [15, 16], and Zeger et al. [17] are excellent references with many applications of GZLMs and GEE approach to repeated measurements in GZLMs in the literature.

In this study, it is aimed to analyze annually collected energy data, covering 28 member countries of the European Union (EU), Albania, Montenegro, Serbia, the Former Yugoslav Republic of Macedonia, and Turkey as the candidate countries to EU membership, and Norway as the European Economic Area country between 2005 and 2013 years, by using GEE approach as an extension of GZLMs to repeated measurements.

*corresponding author; e-mail: niyit@selcuk.edu.tr

2. GEE approach as an extension of GZLMs to repeated measurements data

GZLMs have three components: the random component, the systematic component, and the link function between the random and systematic components [2, 8, 11–13].

The random component of GZLMs identifies the response variable and assumes that distribution of the response variable come from the exponential family having many common discrete and continuous distributions such as normal, binomial, multinomial, negative binomial, Poisson, gamma, inverse Gaussian, etc.:

$$f_Y(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) \tag{1}$$

for some specific functions $a(\cdot), b(\cdot)$, and $c(\cdot)$, canonical parameter θ, and dispersion parameter ϕ [8]. In this study, the interest is on gamma distribution from the exponential family and some characteristics of this distribution are given in Table I.

Suppose that the response variable has an associated $p \times 1$ vector of covariates $\boldsymbol{x} = (x_1, \ldots, x_p)'$. The systematic component of GZLMs specifies a linear predictor η produced by the covariates in the GZLM as follows [8]:

$$\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p. \tag{2}$$

The link function of GZLMs describes the functional relationship between the systematic component given by Eq. (2) and the expected value of the random component $E(Y;\theta)$ given by Table I as a monotonic and

TABLE I

Characteristics of gamma distribution in the exponential family [8].

notation: $G(\mu, v)$	$c(y,\phi) = v\log(vy) - \log y - \log\Gamma(v)$
range of y: $(0,\infty)$	$\mu(\theta) = E(Y;\theta) = -1/\theta$
dispersion parameter: $\phi = v^{-1}$	canonical link: $\theta(\mu) = \mu^{-1}$
cumulant function: $b(\theta) = -\log(-\theta)$	variance function: $V(\mu) = \mu^2$

differentiable function $g(\mu)$ of the mean [8]. In this study, log of the mean of the random component called *log–link function* is taken as link function as follows:

$$g(\mu) = \log(\mu). \tag{3}$$

For more details, see [2, 8, 11–13, 18, 19].

GEE approach is an extension of the GZLMs for the analysis of repeated measurements data. There are three steps in the GEE approach. The first step of the GEE approach is to relate the marginal mean of the response variable $\mu_{ij} = E(y_{ij})$ to a linear combination of the covariates

$$g(\mu_{ij}) = x'_{ij}\beta \tag{4}$$

where y_{ij} is the response variable for subject i at time j, $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$ is the corresponding $p \times 1$ vector of covariates, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the $p \times 1$ vector of unknown parameters. Finally, $g(\cdot)$ is the link function [10, 17, 18].

The second step of the GEE approach is to describe the variance of the response variable as a function of the mean

$$V(y_{ij}) = V(\mu_{ij})\,\varphi, \tag{5}$$

where $V(\cdot)$ is the variance function and φ is a possibly unknown scale parameter [10, 17, 18].

The third step of the GEE approach is to choose the form of a $t_i \times t_i$ working correlation matrix $\boldsymbol{R}_i(\boldsymbol{\alpha})$ among repeated measurements over subjects for each $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{it_i})'$. The (j, j') element of $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is the known, hypothesized, or estimated correlation between y_{ij} and $y_{ij'}$. This working correlation matrix may depend on a vector of unknown parameters $\boldsymbol{\alpha}$, which is the same for all subjects. Thus, we assume that $\boldsymbol{R}_i(\boldsymbol{\alpha})$ for each subject is known except for a fixed number of parameters $\boldsymbol{\alpha}$ that we must estimate from the data. Although this correlation matrix can differ from subject to subject, we commonly use a working correlation matrix $\boldsymbol{R} = \boldsymbol{R}(\boldsymbol{\alpha})$ (Table II) that approximates the average dependence among repeated observations over subjects [10, 17, 18].

Pan [20] introduced two useful extensions of Akaike's information criterion (AIC) [21], based on the quasi-likelihood function under the independence model, as goodness-of-fit-test statistics for choosing the best working correlation structure among repeated measurements in GEE approach given in Table III.

In Table III, $\boldsymbol{I}$ represents the independent covariance structure, $\hat{\boldsymbol{V}}_R$ is robust variance estimator obtained from

TABLE II

Working correlation structures for repeated measurements in GEE approach [18].

Independent $R_{jj'} = \begin{cases} 1 \text{ if } j = j' \\ 0 \text{ otherwise} \end{cases}$	Exchangeable $R_{jj'} = \begin{cases} 1 \text{ if } j = j' \\ \alpha \text{ otherwise} \end{cases}$
First-order autoregressive $R_{jj'} = \begin{cases} 1 \text{ if } j = j' \\ \alpha^{\lvert j-j'\rvert} \text{ otherwise} \end{cases}$	M-dependent $R_{jj'} = \begin{cases} 1 \text{ if } j = j' \\ \alpha_{\lvert j-j'\rvert} \text{ if } \lvert j-j'\rvert \le m \\ 0 \text{ otherwise} \end{cases}$
Unstructured $R_{jj'} = \begin{cases} 1 \text{ if } j = j' \\ \alpha_{jj'} \text{ otherwise} \end{cases}$	

TABLE III

Goodness-of-fit-test statistics for choosing the best working correlation structure among repeated measurements in GEE approach [20].

Quasi-likelihood information criterion (QIC)	$-2Q(\hat{\mu}, I) + 2\mathrm{Tr}(\hat{\Omega}_I^{-1}\hat{V}_R)$
Corrected quasi-likelihood information criterion (QICC)	$-2Q(\hat{\mu}, I) + 2p$

a general working correlation structure $\boldsymbol{R}$, $\hat{\boldsymbol{\Omega}}_I$ is another variance estimator obtained under the assumption of an independent correlation structure, p is the number of parameters in the model when $\mathrm{Tr}(\hat{\boldsymbol{\Omega}}_I^{-1}\hat{\boldsymbol{V}}_R) \approx \mathrm{Tr}(\boldsymbol{I}) = p$ [20].

3. An application on GZLMs for energy repeated measurements data using GEE approach

In this study, annually collected energy data, covering 28 member countries of the EU, 5 candidate countries to EU membership, and Norway as the European Economic Area country between year 2005 and 2013, are analysed by using GEE approach as an extension of GZLMs to repeated measurements.

For this aim, country is taken as subject variable. 28 member countries of the EU, Albania, Montenegro, Serbia, The Former Yugoslav Republic of Macedonia, and Turkey as the candidate countries to EU membership, and Norway are taken as subject variable levels. Time as year is taken as within-subjects variable. Years between 2005 and 2013 are taken as within-subject variable levels. Annual complete energy balances for all products data between 2005 and 2013 are taken as values of response variable. Supply, transformation and consumption, imports and exports of solid fuels, oil, gas, electricity, and renewable energy types of these countries between 2005 and 2013 are taken as explanatory variables to the GZLM. Supply, transformation and consumption of these energy types are taken as covariates into the model. Imports and exports of these energy types are taken as factors into the model by coding 0 and 1 depending on whether imports or exports exist or not. All data

used in this study are taken from EUROSTAT energy database [22]. All statistical computations and data analysis are performed by using IBM SPSS Statistics 21 [23] and SAS Enterprise Guide 4.3 [24] programmes.

Response variable probability distribution is taken as gamma distribution and link function is taken as log-link function. Hybrid, and maximum likelihood (ML) methods are used as parameter and scale parameter estimation methods for energy data. For modelling the within-subject variability in annual complete energy balances repeated measurements data between 2005 and 2013; independent, exchangeable, AR(1), m-dependent, and unstructured working correlation matrices are constituted. Goodness-of-fit-test statistics for choosing the best working correlation structure among energy repeated measurements data between 2005 and 2013 in GEE approach are given in Table IV. From Table IV, the most smallest information criterion (IC) values of QIC and QICC as 192.975 and 110.985 indicate that "exchangeable" and "independent" are the best working correlation structures among annual complete energy balances repeated measurements data. Cui [7] recommended using QIC when these IC select different structures. So "exchangeable" is chosen as the best working correlation structure.

In Table V, parameter estimates, standard errors of parameter estimates, lower and upper bounds of the Wald confidence intervals for parameters, the Wald chi-square test statistics values, related degrees of freedom and asymptotic significance values of statistically significant covariates and factors are given. Supply, transformation and consumption of solid fuels, oil and renewable energy covariates and also imports of solid fuels and electricity, exports of oil and electricity factors are taken as statistically significant explanatory variables into the GZLM at $\alpha = 0.05$ significance level.

TABLE IV

Goodness-of-fit-test statistics (GOF) for choosing the best working correlation structure (WCS) among annual complete energy balances repeated measurements data between 2005 and 2013 in GEE approach.

WCS	GOF	
	QIC	QICC
independent	201.544	110.985*
exchangeable	192.975*	168.874
AR(1)	309.828	216.557
M-dependent	202.295	128.052
unstructured	525.732	338.994

*The most smallest IC values for QIC and QICC indicate the best working correlation structure

TABLE V

Parameter estimates by using hybrid and ML estimation methods and also exchangeable working correlation structure belonging to complete energy balances repeated measurements data.

Factors covariates	$\hat{\beta}$	$se(\hat{\beta})$	95% Wald confidence interval		Hypothesis test		
			lower	upper	Wald χ^2	df	p-value
intercept	7.610	0.0054	7.599	7.621	1959738.307	1	0.000
supply solid fuels	4.725×10^6	1.8780×10^6	1.044×10^6	8.406×10^6	6.330	1	0.012
supply oil	1.355×10^5	3.0242×10^6	7.618×10^6	1.947×10^5	20.062	1	0.000
supply renewable energy	2.151×10^5	5.8080×10^6	1.013×10^5	3.290×10^5	13.722	11	0.000
import solid fuels	0.041	0.0013	0.039	0.044	1060.770	1	0.000
import electricity	2.600	0.1369	2.332	2.869	361.030	1	0.000
export oil	–0.742	0.0029	–0.747	–0.736	66790.365	1	0.000
export electricity	0.077	0.0025	0.073	0.082	937.657	1	0.000

By using the parameter estimates given by Table V, the fitted GZLM with gamma log–link function for complete energy balances data is given as follows: $\log(\text{complete energy balances})_i = 7.610 + 4.725 \times 10^6(\text{supply solid fuels})_i + 1.355 \times 10^5(\text{supply oil})_i + 2.151 \times 10^5(\text{supply renewable energy})_i + 0.041(\text{import solid fuels})_i + 2.600(\text{import electricity})_i - 0.742(\text{export oil})_i + 0,077(\text{export electricity})_i$ for i = Belgium, Bulgaria, Czech Republic, Denmark, Germany, Estonia, Ireland, Greece, Spain, France, Croatia, Italy, Cyprus, Latvia, Lithuania, Luxembourg, Hungary, Malta, Netherlands, Austria, Poland, Portugal, Romania, Slovenia, Slovakia, Finland, Sweden, United Kingdom, Albania, Montenegro, Serbia, The Former Yugoslav Republic of Macedonia, Turkey, and Norway.

4. Results and discussion

Top 5 EU member countries having the highest actual values of annual complete energy balances for all products and Turkey's actual values between 2005 and 2013 are demonstrated in Fig. 1. As seen from Fig. 1, Turkey (with 118532,70 thousand TOE) reached quite close to Spain's complete energy balances for all products actual value (with 119329,30 thousand TOE) in 2013.

In this study, annual complete energy balances data of EU member, candidate countries and Norway are modelled in terms of supply, transformation and consumption of solid fuels, oil and renewable energy, imports of solid fuels and electricity, exports of oil and electricity by GZLM given in Eq. (6). Actual values, estimates,

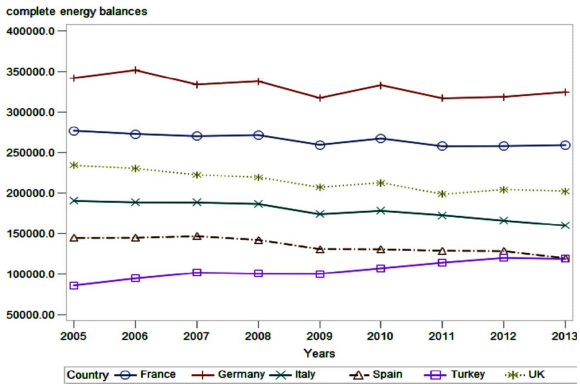

Fig. 1. Multiple line plots of annual complete energy balances for all products actual values for top 5 EU member countries and Turkey between 2005 and 2013.

and the residuals (differences between the actual values and the estimates of annual complete energy balances for all products) for the countries best fit to GZLM, and also most under and overestimated countries by GZLM in 2013 are given in Table VI. From Table VI, it is seen that Albania and Malta are the countries that best fit to the GZLM. Germany, UK, and Turkey are the most underestimated countries, France, Italy, and Greece are the most overestimated countries.

TABLE VI

Annual complete energy balances estimates for the countries best fit to the GZLM, and also the most under and overestimated countries by GZLM in 2013.

Country	Annual complete energy balance		
	actual value	estimate	residual
Albania	2363.00	2355.05	7.95
Malta	872.80	763.84	108.96
Germany	324488.80	158489.00	165999.80
UK	202173.80	110154.00	92019.80
Turkey	118532.70	38194.40	80338.30
France	258949.90	801678.00	–542728.10
Italy	159515.00	328095.00	-168580.00
Greece	24300.40	89536.48	–65236.08

5. Conclusion

In this study, choosing the best working correlation structure among annual complete energy balances repeated measurements data between 2005 and 2013 in GEE approach is investigated by using QIC and QICC information criterion. This is one of the most important problems while working with repeated measurements data taken from each subject in GEE approach. Wrongly specified working correlation structures given by Table II causes the misspecification of the systematic component of the GZLMs given by Eq. (2).

Acknowledgments

The authors would like to thank Professor Iskender Akkurt and the referee for their valuable contributions to this paper.

References

[1] J.A. Nelder, R.W.M. Wedderburn, *J. Roy. Statist. Soc. Ser. A* **135**, 370 (1972).

[2] G. Johnston, *SAS software to fit the generalized linear model*, Sugi Papers 93–183, SAS Institute Inc., Cary, NC 1993.

[3] K.Y. Liang, S.L. Zeger, *Biometrika* **73**, 13 (1986).

[4] P. Diggle, P. Heagerty, K.Y. Liang, S. Zeger, *Analysis of Longitudinal Data*, 2nd ed., Oxford University Press, Oxford 2002.

[5] G.M. Fitzmaurice, N.M. Laird, J.H. Ware, *Applied Longitudinal Data*, 2nd ed., Wiley, Hoboken, NJ 2004.

[6] R.W.M. Wedderburn, *Biometrika* **61**, 439 (1974).

[7] J. Cui, *Stata J.* **7**, 209 (2007).

[8] P. McCullagh, J.A. Nelder, *Generalized Linear Models*, Chapman and Hall, London 1989.

[9] D. Firth, *Generalized Linear Models*, in: *Statistical Theory and Modelling*, Eds. D.V. Hinkley, N. Reid, E.J. Snell, Chapman and Hall, London 1991, p. 55.

[10] D.K. Blough, C.W. Madden, M.C. Hornbrook, *J. Health Econ.* **18**, 153 (1999).

[11] J.K. Lindsey, *Applying Generalized Linear Models*, Springer Texts in Statistics, Springer-Verlag, New York 2000.

[12] A.J. Dobson, A.G. Barnett, *An Introduction to Generalized Linear Models*, Chapman and Hall, Boca Raton 2008.

[13] A. Agresti, *Foundations of Linear and Generalized Linear Models*, Wiley, NJ 2015.

[14] J.W. Hardin, J.M. Hilbe, *Generalized Estimating Equations*, Chapman and Hall, Boca Raton, FL 2003.

[15] S.H. Lipsitz, G.M. Fitzmaurice, E.J. Orav, N.M. Laird, *Biometrics* **50**, 270 (1994).

[16] S.H. Lipsitz, K. Kim, L. Zhao, *Statist. Med.* **13**, 1149 (1994).

[17] S.L. Zeger, K.Y. Liang, P.S. Albert, *Biometrics* **44**, 1049 (1988).

[18] C.S. Davis, *Statistical Methods for the Analysis of Repeated Measurements*, Springer Texts in Statistics, Springer-Verlag, New York 2002.

[19] G. Grover, A.S.A. Sabharwal, J. Mittal, *Int. J. Statist. Med. Res.* **2**, 209 (2013).

[20] W. Pan, *Biometrics* **57**, 120 (2001).

[21] H. Akaike, *IEEE Trans. Automat. Control* **19**, 716 (1974).

[22] Eurostat Energy Database, 2015.

[23] *IBM SPSS Statistics for Windows*, Version 21.0. Armonk, NY, IBM Corp, 2012.

[24] *SAS Enterprise Guide 7.1*, SAS, Cary, NC.