ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2017-55668

Generalized Estimating Equations Approach to Multi-Way Contingency Tables with Longitudinal Data in Case-Control Studies

Vaka-Kontrol Çalışmalarında Uzun Süreli Verilere Sahip Çok-Yönlü Çapraz Tablolar için Genelleştirilmiş Tahmin Denklemleri Yaklaşımı

Melike KAYA BAHÇECİTAPAR,[a]
Serpil AKTAŞ ALTUNAY[a]

[a]Department of Statistics,
Hacettepe University Faculty of Science,
Ankara

Geliş Tarihi/*Received:* 13.03.2017
Kabul Tarihi/*Accepted:* 14.06.2017

Yazışma Adresi/*Correspondence:*
Melike KAYA BAHÇECİTAPAR
Hacettepe University Faculty of Science,
Department of Statistics, Ankara,
TÜRKİYE/TURKEY
mlk@hacettepe.edu.tr

ABSTRACT Objective: Log-linear analysis is a classical statistical method for the analysis of association between variables in multi-way contingency tables. Generalized Estimating Equations (GEEs) approach is popular especially for analyzing longitudinal data, since it enables to take into account the correlation of repeated measures over time within subjects by defining a so-called "working correlation structure". GEEs provide consistent regression parameter estimates even if working correlation structure is misspecified. In this paper, we suggest GEEs approach to the analysis of a multi-way contingency table with longitudinal data, which consists of more than one contingency tables obtained over time in case-control studies and examine the method of GEEs by considering four different working correlation structures for correlations between longitudinal count data in the table. **Material and Methods:** Log-linear analysis and GEEs method for longitudinal data in the multi-way contingency table with time factor are performed by SAS 9.4 statistical software program. A real genetic association case-control study with longitudinal data was illustrated to compare both methods. **Results:** Using either the classical log-linear analysis or GEEs method with an independent (IND) working correlation structure generates similar results for parameter estimates. It is found that linear model fitting longitudinal data in the multi-way contingency table is observed to be same for both log-linear analysis and GEEs approach performed under all correlation structures. **Conclusion:** This study indicates that GEEs approach provides more efficient and unbiased regression parameter estimates for the multi-way contingency table designed by responses measured repeatedly over time in the case-control study.

Keywords: Case-control studies; linear models; longitudinal studies; GEEs

ÖZET Amaç: Logaritmik doğrusal analiz, çok-yönlü tablolarda değişkenler arasındaki ilişkiyi analiz etmek için kullanılan klasik bir istatistiksel yöntemdir. Genelleştirilmiş Tahmin Denklemleri (GEEs) yaklaşımı, bir "çalışan korelasyon yapısı" tanımlayarak, zaman boyunca birimlerin tekrarlı ölçümleri arasındaki korelasyonu hesaba kattığı için, özellikle uzun süreli verilerin analizi için sık kullanılmaktadır. Çalışan korelasyon yapısı yanlış belirlenmiş olsa bile, regresyon parametre tahminleri tutarlıdır. Bu makalede, vaka-kontrol çalışmalarında zaman boyunca elde edilen birden fazla çapraz tablolardan oluşan uzun süreli verilere sahip çok-yönlü olumsallık tablosunun analizi için GEEs yaklaşımını önerdik ve tablodaki uzun süreli kesikli veriler arasındaki korelasyonlar için dört farklı çalışan korelasyon yapısı düşünülerek GEEs yöntemini inceledik. **Gereç ve Yöntemler:** Zaman faktörlü çok-yönlü çapraz tablodaki uzun süreli verilerin analizi için logaritmik doğrusal analiz ve GEEs yöntemi SAS 9.4 istatistiksel yazılım programında uygulanmıştır. Her iki yöntemi karşılaştırmak için gerçek bir genetik ilişki vaka-kontrol çalışması kullanılmıştır. **Bulgular:** Logaritmik doğrusal analiz ve bağımsız (IND) çalışan korelasyon yapısına sahip GEEs yöntemi ile benzer parametre tahminleri elde edilmiştir. Logaritmik doğrusal analiz ve tüm korelasyon yapıları düşünülerek uygulanan GEEs yönteminde de, çapraz tablodaki uzun süreli verileri inceleyen doğrusal modelin aynı olduğu bulunmuştur. **Sonuç:** Bu çalışma, vaka-kontrol çalışmasında zaman boyunca tekrarlı olarak ölçülen cevaplar tarafından tasarlanmış bir çok-yönlü çapraz tablo için GEEs yaklaşımının logaritmik doğrusal analize göre daha etkin ve yansız tahminler verdiğini göstermektedir.

Anahtar Kelimeler: Vaka-kontrol çalışmaları; doğrusal modeller; uzun süreli çalışmalar; GEEs

Case-control studies are often used in epidemiology and medical sciences as an easy way of comparing subjects with a disease interested (cases) with disease-free subjects (controls).[1,2] A genetic case-control study may examine whether a statistical relationship exists between a particular disease and genotypes of subjects by comparing the frequency of genotypes in groups of subjects with and without the disease.[3] Results of case-control studies are displayed in contingency tables to calculate the relationship between disease and a specific risk factor. In assessing more than one factors that may influence the disease, multi-way contingency tables might be used to analyze a set of counts or frequencies obtained by classifying observations.

Log-linear models which are a special type of generalized linear models for Poisson distributed data can be used to analyze the relationship between two factors in two-way contingency tables and three or more factors in multi-way contingency tables by taking the logarithm of the expected cell counts in the contingency table. It can also deal with multi-category or polytomous factors on the disease interested in case-control studies. Tanaka et al.[4] used log-linear models to investigate the allele frequency among disease-free subjects decreasing with age and the dependence of genotype and age in the genetic association case-control study.

Longitudinal binary responses arise, when the case-control study is designed repeatedly over time for same subjects. Several methods have been proposed for analyzing such longitudinal data in case-control studies.[5,6] Park and Kim[5] extended ordinary logistic regression models to the case-control designs with independent repeated responses from same subjects over time. Their proposed method is based on the method of Generalized Estimating Equations (GEEs)[7]. GEEs are the quasi-likelihood equations which provide quasi-likelihood estimates of regression coefficients in the model from maximization of normality-based log-likelihood function without assuming that the response is normally distributed. GEEs approach does not require distributional assumptions of responses and is widely used for longitudinal and other correlated response data, especially if responses are from binomial and Poisson distribution.[7,8] In the GEEs approach, the within-subject correlations among repeated responses from each subject are taken into account by defining a working correlation structure. GEEs provide consistent parameter and standard error estimates even if the correlation structure among responses is not specified correctly. However, Hin and Wang (2009) and Wang and Carey (2003) showed that the misspecification of working correlation structure may cause inaccurate results of the parameter estimates especially in small sample sizes.[9,10]

Besides the GEEs, generalized linear mixed model (GLMM) provides a flexible approach to handle longitudinal data from exponential family distributions (normal, binomial or Poisson) by using link functions. GLMM is an extension of generalized linear models including the random effects, which represents the influence of each subject on his/her own repeated responses. GLMM can deal with correlated binary and count responses over time. Zhang et al.(2012) examine GEE- and GLMM-based log-linear models for modelling longitudinal count responses of each subject.[11]

In this paper, unlike the usual methods, GEEs approach is used for the analysis of counts in the multi-way contingency table which summarize longitudinal data obtained by the case-control study repeatedly over time. Cell counts in the contingency table are considered as the responses in the sense that we are interested in describing the relationship between factors and disease. The use of response differs from the classical use in GEEs approach. We illustrate and compare the GEEs approach and

classical log-linear analysis provided by SAS statistical software program to analyze count data in the three-way contingency table where one of factors is time. We also examined GEEs approach for independent (IND), unstructured (UN), compound symmetry (CS) and first-order autoregressive (AR1) working-correlation structures among repeated counts in the contingency table with time factor. Additionally, it is emphasized that the model selection in GEEs is important and necessary to determine the model including significant factors for the study.

LOG-LINEAR MODELS FOR THREE-WAY CONTINGENCY TABLES

An $a \times b \times c$ three-way contingency table cross-classifies m subjects by the levels of three factors, A, B, C, which are taking possible values 1, 2, ..., a; 1, 2, …, b and 1, 2, ..., c, respectively. The cell counts m_{ijk} are the number of subjects having categories i, j and k of factors A, B and C, respectively. A multinomial distribution can be assumed with cell probabilities π_{ijk}, where $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\pi_{ijk}=1$, and expected counts $\left(\mu_{ijk}= m\,\pi_{ijk}\right)$ in ijk th cell in the table. Log-linear models can also apply to Poisson sampling for independent observed cell counts m_{ijk} with means $\left(\mu_{ijk}= E(m_{ijk})\right)$.

Under statistical independence, for multinomial sampling, A, B and C factors are mutually independent, when $\pi_{ijk} = \pi_{i++}\,\pi_{+j+}\,\pi_{++k}$ for all i, j and k, where $\pi_{i++}= P(A=i)$, $\pi_{+j+}= P(B=j)$ and $\pi_{++k}= P(C=k)$. For μ_{ijk}, the log-linear model for mutual independence has the following form

$$log\left(\mu_{ijk}\right) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C, \quad (1)$$

with the constraints such as $\sum_i \lambda_i^A = \sum_j \lambda_j^B = \sum_k \lambda_k^C = 0$. When factor C is jointly independent of factors A and B, $\pi_{ijk} = \pi_{ij+}\,\pi_{++k}$ for all i, j and k, where $\pi_{ij+} = P(A=i, B=j)$. Jointly independence has log-linear form

$$log\left(\mu_{ijk}\right) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB}. \quad (2)$$

If factors A and B are conditionally independent of each other, when factor C is fixed, $\pi_{ij\,|\,k} = P(A=i, B=j \,/\, C=k) = \pi_{i+\,/\,k}\,\pi_{+j\,/\,k}$ for all i, j and k. Conditional independence of A and B factors, given C, has the log-linear model as

$$log\left(\mu_{ijk}\right) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}. \quad (3)$$

In Equations (1),(2) and (3), λ is an overall effect or a grand mean of the logarithms of the expected counts, λ_i^A, λ_j^B and λ_k^C are the main effects of A, B and C factors. λ_{ij}^{AB}, λ_{ik}^{AC} and λ_{jk}^{BC} are the two-way interaction terms of them. The log-linear model including three two-way interactions which are conditionally independent is $log\left(\mu_{ijk}\right) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}$. The general log-linear model for a three-way contingency table is

$$log\left(\mu_{ijk}\right) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}, \quad (4)$$

where λ_{ijk}^{ABC} is the three-way interaction term representing the association between three factors. Equation (4) is also called as saturated model seven model components including three main effects, three two-way interactions and one three-way interaction of factors. Analysis of Variance (ANOVA) type parameter constraints are $\sum_i \lambda_i^A = \sum_j \lambda_j^B = \sum_i \sum_j \lambda_{ij}^{AB} = \sum_i \sum_j \sum_k \lambda_{ijk}^{ABC} = 0$.[12]

In SAS statistical software program, PROC GENMOD and PROC CATMOD can be used for log-linear modelling of count data in contingency tables. [13, 14] Both procedures fit generalized linear models to cell counts in contingency tables. Weighted least squares (WLS) with maximum likelihood (ML) estimation method can be used to fit log-linear models in PROC CATMOD procedure. PROC GENMOD procedure is especially suited for count data and also provides ML estimation of parameters for log-linear models. In order to determine model components necessary to best account for the count data in contingency table, the overall goodness of fit of each log-linear model are tested by the standard Pearson chi-square statistic (χ^2) or the likelihood ratio statistic (G^2) defined as

$$G^2 = 2\sum_i\sum_j\sum_k m_{ijk}\, log\left(\frac{m_{ijk}}{\mu_{ijk}}\right),$$

where m_{ijk} are the observed cell counts and μ_{ijk} are the estimated expected counts under the model for testing the null hypothesis of independence. However, G^2 is more commonly used, since it is minimized in maximum likelihood estimation. Smallest value of likelihood ratio chi-square G^2 indicates the well-fitting model.[12,15,16] On the other hand, Pearson residuals or standardized (adjusted) Pearson residuals of the cells can be calculated under the assumed model to assess the selection of well-fitting model among reasonable models for the contingency table based on G^2 values. Adjusted Pearson residual is distributed as $N(0,1)$ and its absolute value that exceeds 2 or 3 indicates a sign of lack of fit.[12]

GENERALIZED ESTIMATING EQUATIONs (GEEs)

In a longitudinal study, suppose that there are m subjects measured at n time points. Let $\boldsymbol{Y}_i=(Y_{i1}, \ldots, Y_{in})'$ denote a $n\times1$ vector of responses of subject i ($i=1,\ldots,m$) and $\boldsymbol{X}_{ij}$ be a $(p+1)\times1$ vector of discrete or continous covariates for subject i at j^{th} time point ($j=1, \ldots, n$). A marginal model for longitudinal data is described as $g^{-1}(\mu_{ij})=\boldsymbol{X}_{ij}'\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is an unknown $(p+1)\times1$ vector of regression parameters and $\mu_{ij}=E(Y_{ij}\,|\,\boldsymbol{X}_{ij})$ depends on the covariates through a known link function $g^{-1}(.)$. The estimate of $\boldsymbol{\beta}$ is obtained by solving the following GEEs[8]:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{m} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$$

where $\boldsymbol{\mu}_i=(\mu_{i1},\ldots,\mu_{in})'$ is a mean vector for subject i, $\boldsymbol{D}_i=\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}'}$ and $\boldsymbol{V}_i$ is the variance-covariance matrix for $\boldsymbol{Y}_i$ noted by $\boldsymbol{V}_i=\phi\,\boldsymbol{A}_i^{1/2}\boldsymbol{R}_i\boldsymbol{A}_i^{1/2}$. The $\boldsymbol{R}_i$ describes $n\times n$ working correlation structure within repeated measures of subject i which account for the form of within-subject correlation of responses and $\boldsymbol{A}_i=diag\{\upsilon(\mu_{i1}), \ldots, \upsilon(\mu_{in})\}$ is a function of μ_{ij} and ϕ is a scale parameter which depend on the distribution of response. If Y_{ij} is count, $\upsilon(\mu_{ij})=\mu_{ij}$, ϕ is equal to 1 and $g^{-1}(.)$ is a log link function.[8] There are a number of choices for $\boldsymbol{R}_i$ including independent (IND), unstructured (UN), first-order autoregressive (AR1) or compound symmetry (CS) structures.[17,18] GEEs are obtained by using an iterative quasi-scoring algorithm. The inference of regression parameters in GEEs is obtained as Wald test using robust standard errors. Even when working correlation structure is misspecified, parameter estimates from GEEs are consistent.[7,18] Unlike general linear models, GEEs approach is a non-likelihood based method. The popular Akaike Information Criteria (AIC) approach as a model selection criteria cannot be directly applied, since AIC is based on maximum likelihood estimation. Model selection in GEEs framework is performed with Quasi-AIC (QIC) which is as a modification to AIC for correlated data.[19]

APPLICATION ON GENETIC ANALYSIS WORKSHOP 18 DATA

Log-linear modelling and GEEs are used for the analysis of 3×2×3 contingency table in the study of Karadağ and Aktaş (2016). The data set is obtained from Genetic Analysis Workshop 18 (GAW18) conducted between 1991 and 2011, which is interested in the association between disastolic blood pressure (DBP) as a hereditary disease and whole genome sequence genotype data. DBP values were measured at three different time points.[20] First measurements were taken between 1991 and 1996, second measurements were taken between 1997 and 2000 and third measurements were taken between 1998 and 2006. In this study, DBP is used as an identifier of high blood pressure and hypertension (DBP>90) is defined as disease (*D*). There were 445 individuals with available genotype information for *NPR3-C5orf23* gene study. The genotype distribution for subjects having hypertension (D^+) and hypertension-free (D^-) in this longitudinal study is presented by the three-way contingency table as Table 1.

TABLE 1. *NPR3-C5orf23* data set.

Visit (*V*)	Disease Status (*D*)	Genotype (*G*)			Total Visit
		CC	*CT*	*TT*	
*V*1	D^+	19	39	15	73
	D^-	117	190	65	372
*V*2	D^+	34	69	27	130
	D^-	102	160	53	315
*V*3	D^+	42	92	31	165
	D^-	94	137	49	280

The data in the 3×2×3 contingency table in Table 1 within the framework of log-linear analysis approach is fistly analyzed and log-linear analysis are performed by PROC GENMOD and PROC CATMOD procedures in SAS 9.4 statistical software program. We treat three factors as independent categorical variables. For the 3×2×3 three-way contingency table in Table 1, under all types of independence, log-linear models in Table 2 are analyzed to explore relationships between factors in Table 1.

In Table 2, for i=1, 2, 3; j=1, 2 and k=1, 2 ,3; μ_{ijk} is the expected count, λ is the overall effect, λ_i^V (or V_i), λ_j^D (or D_j) and λ_k^G (or G_k) are the main effects of *V*, *D* and *G* factors and λ_{ij}^{VD} (or VD_{ij}), λ_{ik}^{VG} (or VG_{ik}) and λ_{jk}^{DG} (or DG_{jk}) are the two-way interaction terms and λ_{ijk}^{VDG} (or VDG_{ijk}) is the three-way interaction term of factors. (*VDG*) is the saturated model that includes all model components. The model (*V*, *D*, *G*) includes only the main effects of factors. All main effects and only one two-way interaction term are included in (*VD*, *G*), (*VG*, *D*) and (*DG*, *V*) models. (*VD*, *VG*), (*VD*, *DG*) and (*VG*, *DG*) are including all main effects and two different two-way interaction terms.

TABLE 2. Representations of log-linear and linear models.

Symbol	Log-Linear Model	Linear Model
(V, D, G)	$log\left(\mu_{ijk}\right) = \lambda + \lambda_i^V + \lambda_j^D + \lambda_k^G$	$log\left(\mu_{ijk}\right) = \beta_0 + \beta_1 V_i + \beta_2 D_j + \beta_3 G_k$
(*VD*, *G*)	$log\left(\mu_{ijk}\right) = \lambda + \lambda_i^V + {}^D + \lambda_k^G + \lambda_{ij}^{VD}$	$log\left(\mu_{ijk}\right) = \beta_0 + \beta_1 V_i + \beta_2 D_j + \beta_3 G_k + \beta_4 VD_{ij}$
(*VG*, *D*)	$log\left(\mu_{ijk}\right) = \lambda + \lambda_i^V + {}^D + \lambda_k^G + \lambda_{ik}^{VG}$	$log\left(\mu_{ijk}\right) = \beta_0 + \beta_1 V_i + \beta_2 D_j + \beta_3 G_k + \beta_4 VG_{ik}$
(*DG*, *V*)	$log\left(\mu_{ijk}\right) = \lambda + \lambda_i^V + \lambda_j^D + \lambda_k^G + \lambda_{jk}^{DG}$	$log\left(\mu_{ijk}\right) = \beta_0 + \beta_1 V_i + \beta_2 D_j + \beta_3 G_k + \beta_4 DG_{jk}$
(*VD*, *VG*)	$log\left(\mu_{ijk}\right) = \lambda + \lambda_i^V + \lambda_j^D + \lambda_k^G + \lambda_{ij}^{VD} + \lambda_{ik}^{VG}$	$log\left(\mu_{ijk}\right) = \beta_0 + \beta_1 V_i + \beta_2 D_j + \beta_3 G_k + \beta_4 VD_{ij} + \beta_5 VG_{ik}$
(*VD*, *DG*)	$log\left(\mu_{ijk}\right) = \lambda + \lambda_i^V + \lambda_j^D + \lambda_k^G + \lambda_{ij}^{VD} + \lambda_{jk}^{DG}$	$log\left(\mu_{ijk}\right) = \beta_0 + \beta_1 V_i + \beta_2 D_j + \beta_3 G_k + \beta_4 VD_{ij} + \beta_5 DG_{jk}$
(*VG*, *DG*)	$log\left(\mu_{ijk}\right) = \lambda + \lambda_i^V + \lambda_j^D + \lambda_k^G + \lambda_{ik}^{VG} + \lambda_{jk}^{DG}$	$log\left(\mu_{ijk}\right) = \beta_0 + \beta_1 V_i + \beta_2 D_j + \beta_3 G_k + \beta_4 VG_{ik} + \beta_5 DG_{jk}$
(*VD*, *VG*, *DG*)	$log\left(\mu_{ijk}\right) = \lambda + \lambda_i^V + \lambda_j^D + \lambda_k^G + \lambda_{ij}^{VD} + \lambda_{ik}^{VG} + \lambda_{jk}^{DG}$	$log\left(\mu_{ijk}\right) = \beta_0 + \beta_1 V_i + \beta_2 D_j + \beta_3 G_k + \beta_4 VD_{ij} + \beta_5 VG_{ik} + \beta_6 DG_{jk}$
(*VDG*)	$log\left(\mu_{ijk}\right) = \lambda + \lambda_i^V + \lambda_j^D + \lambda_k^G + \lambda_{ij}^{VD} + \lambda_{ik}^{VG} + \lambda_{jk}^{DG} + \lambda_{ijk}^{VDG}$	$log\left(\mu_{ijk}\right) = \beta_0 + \beta_1 V_i + \beta_2 D_j + \beta_3 G_k + \beta_4 VD_{ij} + \beta_5 VG_{ik} + \beta_6 DG_{jk} + \beta_7 VDG_{ijk}$

RESULTS

Log-linear model analysis is a process of assigning which model components are significant, therefore, Pearson χ^2 and G^2 statistics are assessed for each model by PROC GENMOD and PROC CATMOD procedures, respectively. Table 3 shows the results of testing fit for log-linear models in Table 2 (Table 2, Table 3).

According to the *p*-values from Table 3, the (*VD*, *G*), (*VD*, *VG*), (*VD*, *DG*) and (*VD*, *VG*, *DG*) models fit the data well. The standardized (adjusted) Pearson residuals are assessed to find the best fitting model for Table 1. The largest adjusted residuals for four models are found to be 1.2298, 1.7953, 0.5691 and 0.6965, respectively. (*VD*, *DG*) model has the smallest value among the largest adjusted residuals and therefore, it can be used in the decision of selecting the best model for Table 1.

TABLE 3. Goodness-of-fit test results for log-linear models.

Model	df	χ^2	*p*-value	G^2	*p*-value
(*V*, *D*, *G*)	12	55.1626	0.0001	56.82	<.0001
(*VD*, *G*)	10	6.3056	0.7890	6.38	0.7822
(*VG*, *D*)	8	55.1626	0.0001	56.82	<.0001
(*DG*, *V*)	10	49.3058	0.0001	51.17	<.0001
(*VD*, *VG*)	6	6.3056	0.3898	6.38	0.3818
(*VD*, *DG*)	8	0.7317	0.9994	0.73	0.9994
(*VG*, *DG*)	6	49.3058	0.0001	51.17	<.0001
(*VD*, *VG*, *DG*)	4	0.5237	0.9712	0.52	0.9712
(*VDG*)	0	0.0	-	0.0	-

df: degrees of freedom; χ^2: Pearson chi-square statistic; G^2: Likelihood ratio statistic.

Table 4 shows the parameter estimates obtained by both procedures for (*VD*, *DG*) model. It is found that both procedures do not calculate similar parameter estimates and cause different interpretations to decide whether to reject the null hypothesis of some parameter estimates, such as the category *CC* for *G* factor and the categories *V2* and D^+ for *V* and *D* factors, respectively.

TABLE 4. Parameter estimates with standard errors of the (*VD*, *DG*) model from PROC GENMOD and PROC CATMOD procedures for log-linear analysis.

Parameter		PROC GENMOD			PROC CATMOD		
		Estimate	S.E.	*p*-value	Estimate	S.E.	*p*-value
V	*V1*	0.2841	0.0791	0.0003*	-0.1570	0.0489	0.0013*
	V2	0.1178	0.0821	0.1516	0.0484	0.0439	0.2711
D	D^+	-0.3903	0.1599	0.0147*	-0.5090	0.0348	<.0001*
G	*CC*	0.6282	0.0958	<.0001*	-0.0491	0.0476	0.3020
	CT	1.0703	0.0897	<.0001*	0.5441	0.0414	<.0001*
*V*D*	*V1**D^+	-1.0996	0.1613	<.0001*	-0.3072	0.0489	<.0001*
	*V2**D^+	-0.3562	0.1432	0.0129*	0.0645	0.0439	0.1419
*D*G*	D^+**CC*	-0.3648	0.1828	0.0460*	-0.1112	0.0476	0.0195*
	D^+**CT*	-0.0624	0.1635	0.7027	0.0400	0.0414	0.3337

*: *p*-value is smaller than 0.05. S.E.: Standard error

In Table 4, all parameters except two belonging to the category *V2* of *V* and categories D^+ and *CT* of *D*G* are found to be significant by PROC GENMOD. In addition to these two parameters, parameters for the main effect of *G* factor with category *CC* and the interaction term *V*D* with categories *V2* and D^+ are not found to be significant by PROC CATMOD.For GEEs analysis of GAW18 data in Table 1, it is considered that the cells in the combination of levels of *D* and *G* factors as if subjects in GEEs approach are measured repeatedly at three different time points. GEEs analysis of the data is performed by PROC GENMOD procedure with repeated statement. We consider IND, UN, AR1 and CS working correlation structures among repeated counts. Table 5 displays the QIC values for the linear models in Table 2.

TABLE 5. QIC values by GEEs under the assumptions of IND, UN, AR1 and CS working correlation structures.

	R_i =IND	R_i =UN	R_i =AR1	R_i =CS
(*V*, *D*, *G*)	-2025.9354	-1970.9739	-2022.9948	-2025.9354
(*VD*, *G*)	-14874.5887	-14759.6318	-2002.0686	-12045.9154
(*VG*, *D*)	-1347.2542	-1315.6821	-1342.9873	-1347.2543
(*DG*, *V*)	-1891.2933	-1841.0788	-1892.1134	-1894.6186
(*VD*, *VG*)	-8924.2323	-8898.9655	-8917.3526	-6605.2592
(*VD*, *DG*)	-102683.7134*	-100341.1624	-99877.4194	-102193.7447
(*VG*, *DG*)	-1131.9488	-1111.0999	-1132.0485	-1136.7712
(*VD*,*VG*, *DG*)	-71736.0073	-71069.5252	-71685.6587	-71120.1951
(*V*, *D*, *G*)	-2.950414E-32	0.000	0.000	0.000

*: The smallest QIC value. IND: Independent; UN: Unstructured; AR1: First-order autoregressive; CS: Compound symmetry

In Table 5, for all working correlation structures, (*VD*, *DG*) model has the smallest QIC value (QIC= -102683.7134) among all models. However, compared to QIC values for all working correlation structures, (*VD*, *DG*) model with IND working correlation structure is found to be best fitting model in

GEEs approach. Table 6 includes the maximum likelihood parameter estimates by GEEs approach of (*VD*, *DG*) having only the two-way terms VD_{ij} and DG_{jk} and the main effects of each factor (p-value<0.05).

TABLE 6. Parameter estimates with model-based standard errors for (*VD*, *DG*) by GEEs approach with IND working correlation structure.

Parameter		Estimate	S.E.	*p*-value
V	*V1*	0.2841	0.0239	<.0001*
	V2	0.1178	0.0248	<.0001*
D	D^+	-0.3979	0.0484	<.0001*
G	*CC*	0.6282	0.0290	<.0001*
	CT	1.0703	0.0271	<.0001*
*V*D*	$V1*D^+$	-1.0996	0.0488	<.0001*
	$V2*D^+$	-0.3562	0.0433	<.0001*
*D*G*	D^+*CC	-0.3648	0.0553	<.0001*
	D^+*CT	-0.0624	0.0495	.2069

*: *p*-value is smaller than 0.05. S.E.: Standard error

From Table 6, parameter estimates by GEEs are found to be exactly same as log-linear analysis by PROC GENMOD, but standard errors for all parameter estimates of (*VD*, *DG*) in GEEs are much smaller than those in log-linear approach (Table 6). GEEs approach provides that all parameters in the (*VD*, *DG*) model except the interaction D^+*CT (p-value=0.2069) are significant at the 5% level (p-value<0.0001). However, addition to D^+*CT, the main effect of *V* factor with the category *V2* is also found to be not significant in log-linear analysis of Table 1 (Table 1). On the other hand, the $D*V$ interaction term is significant (p-value<0.0001), indicating that the effect of visits differs for D^+ and D^-. The number of subjects per visit for each level of *D* factor illustrates this result.

CONCLUSION

In this paper, we suggest the GEEs approach for the analysis of longitudinal data in the case-control study, which are designed in the format of multi-way contingency table with time factor. We also analyzed this kind of contingency table by log-linear modelling approach and compare both approaches to find the well-fitting model for longitudinal count data in the multi-way contingency table. Results show that both log-linear modelling and GEEs give same results in model selection. Additionally, GEEs approach with IND working correlation structure and log-linear analysis provide same model with same parameter estimates. However, GEEs provide same model with significant factors under more efficient and unbiased estimates.

Conflict of Interest

Authors declared no conflict of interest or financial support.

Authorship Contributions

Idea/Concept: *Melike Kaya Bahçecitapar*
Design: *Melike Kaya Bahçecitapar*
Control/Supervision: *Melike Kaya Bahçecitapar, Serpil Aktaş Altunay*
Data Collection and/or Processing: *Serpil Aktaş Altunay*
Analysis and/or Interpretation: *Melike Kaya Bahçecitapar, Serpil Aktaş Altunay*
Literature Review: *Melike Kaya Bahçecitapar*
Writing the Article: *Melike Kaya Bahçecitapar*
Critical Review: *Serpil Aktaş Altunay*

REFERENCES

1. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. Biometrika 1979;66(3):403-11.
2. Breslow NE, Day NE. Statistical Methods in Cancer Research. Volume 1: The Analysis of Case-Control Studies. International Agency for Research on Cancer Scientific Publications; Lyon, UK: 1980. p.14-42.
3. Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zendervan KT. Basic statistical analysis in genetic case-control studies. Nat Protoc 2011;6(2):121-33.
4. Tanaka N, Kinoshita T, Asada T, Ohashi Y. Log-linear models for assessing gene-age interaction and their application to case-control studies of the apolipoprotein E (apoE) gene in Alzheimer's disease. J Hum Genet 2003;48(10):520-4.
5. Park E, Kim Y. Analysis of longitudinal data in case-control studies. Biometrika 2004;91(2):321-30.
6. Zhao Y, Jian W. Analysis of longitudinal data in the case-control studies via empirical likelihood. Commun Stat Simul Comput 2007;36(3):565-78.
7. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73(1):13-22.
8. Burton P, Gurrin L, Sly P. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. Stat Med 1998;17(11):1261-91.
9. Hin LY, Wang YG. Working-correlation-structure identification in generalized estimating equations. Stat Med 2009;28(4):642-58.
10. Wang YG, Carey V. Working correlation structure misspecification, estimation and covariate design: implications for generalized estimating equations performance. Biometrika 2003;90(1):29-41.
11. Zhang H, Yu Q, Feng C, Gunzler D, Wu P, Tu XM. A new look at the diffrence between the GEE and the GLMM when modelling longitudinal count responses. J Appl Stat 2012;39(9):2067-79.
12. Agresti A. Categorical Data Analysis. 3rd ed. New York: Wiley; 2013. p.81, 142, 143, 314-57.
13. SAS Institute Inc. SAS/STAT 9.2 User's Guide. The GENMOD Procedure (Book Excerpt). Cary, NC: SAS Institute Inc; 2009. p.2040-5.
14. SAS Institute Inc. Log-linear model analysis. SAS/STAT 9.2 User's Guide. The CATMOD Procedure (Book Excerpt). Cary, NC: SAS Institute Inc; 2009. p.1148-50.
15. Knoke D, Burke PJ. Log-Linear Models. 1st ed. Beverlys Hills, CA: Sage; 1980. p.11-24.
16. Agresti A. Analysis of Ordinal Categorical Data. 1st ed. New York: John Wiley & Sons; 1984. p.10.
17. Hedeker D, Gibbons RD. Generalized estimating equations (GEE) models. Longitudinal Data Analysis. 1st ed. New York: John Wiley & Sons; 2006. p.131-47.
18. Hardin JW, Hilbe JM. Generalized Estimating Equations. Boca Raton, FLA, USA, FL: Chapman & Hall/CRC Press; 2003. p.30, 58-76.
19. Pan W. Akaike's information criterion in generalized estimating equations. Biometrics 2001;57(1):120-5.
20. Karadağ Ö, Aktaş S. Testing the trend for genotype distribution of hypertension patients in case-control studies. Int J Hum Genet 2016;16(1-2):16-21.