Anadolu Tarım Bilim. Derg., 2010,25(3):197-201
Anadolu J. Agric. Sci., 2010,25(3):197-201

Araştırma
Research

COMPARISON OF WORKING CORRELATION MATRICES IN GENERALIZED ESTIMATING EQUATIONS FOR ANIMAL DATA

Hasan ÖNDER[1*], Mustafa OLFAZ[1], Ercan SOYDAN[1]

[1] Ondokuz Mayıs University, Faculty of Agriculture, Department of Animal Science, Samsun-Turkey

***e-mail: Hasan Önder: hasanonder@gmail.com**

Geliş Tarihi: 19.11.2009 Kabul Tarihi: 14.10.2010

ABSTRACT: Experimental animal science is often interested in estimating the effects of some set of explanatory variables on a categorical response variable on interest. This study demonstrates the use of generalized estimating equations (GEE) with use of categorical data taken from an animal research. It is known that different correlation structures can yield different results when the data has repeated measurements. The aim of this study was to determine which correlation structure has more appropriate the animal science. Five different correlation structures were compared on data with repeated measurements for GEE. As result of this study, Independent and exchangeable correlation structures can be recommended to analyze the categorical data sets for biological sciences because of the lowest QIC values.

Keywords: Generalized Estimating Equations, Categorical data, Correlation structure, Animal science

HAYVANCILIK VERİLERİ İÇİN GENELLEŞTİRİLMİŞ TAHMİN DENKLEMLERİNDE VARSAYILAN KORELASYON MATRİSLERİNİN KARŞILAŞTIRILMASI

ÖZET: Deneysel hayvancılık bilimleri çoğu kez üzerinde çalışılan kesikli yanıt değişkeni üzerine etki eden açıklayıcı değişken kümesinin etkilerinin tahmini konu edinmektedir. Bu çalışma, hayvancılık alanından elde edilen kesikli bir veri kümesi ile genelleştirilmiş tahmin denklemlerinin kullanımını açıklamaktadır. Tekrar eden ölçüm içeren verilerde farklı korelasyon yapıları farklı sonuçlar üretebilmektedir. Bu çalışmada, hangi korelasyon yapısının hayvancılık çalışmalarına daha uygun olabileceği araştırılmıştır. Beş farklı korelasyon yapısı, tekrarlanan ölçümlü veri kümesi üzerinde genelleştirilmiş tahmin denklemleri için karşılaştırılmıştır. Çalışmada sonuç olarak, bağımsız ve değişebilir korelasyon yapılarının en düşük QIC değerine sahip olmaları nedeni ile biyolojik çalışmalarda kesikli veri kümelerinin analizi için önerilebilir olduğu anlaşılmıştır.

Anahtar Sözcükler: Genelleştirilmiş Tahmin Denklemleri, Kesikli veri, Korelasyon yapısı, Hayvancılık

1. INTRODUCTION

Regression methods have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables. It is often the case that the outcome variable is categorical, taking on two or more possible values. Over the last decade some methods like logistic regression has become the standard method of analysis for categorical response variable (Hosmer and Lemeshow, 2000). It is known that the standard logistic regression method does not easily address the exact situation where the data are clustered or have a natural hierarchy (Sturdivant and Hosmer, 2007).

The widespread availability of the Generalized Estimating Equation (GEE) method is the usability on data that consist of clustered or repeated observations (Hammill and Preisser, 2006; Reboussin, et al., 2006). GEE is useful to analyze the data that are collected in clusters where observations within a cluster may be correlated, but observations in separate clusters are independent. They can account for spatial and temporal correlations. Like generalized linear models (GLMs), GEE allows for non-linear relationships between independent variables and the dependent variable, and accommodate the dependent variable has non-normal distribution (Ward and Myers, 2007). GEE is a method to fit regression models taking the correlations among the observations into account. In this method, the correlation matrix can take different structures, if there are repeated measurements in the data set (Paradis and Claude, 2002). The use of an incorrect correlation matrix can cause an inconsistency problem in estimation (Park and Shin, 1999).

In this study, we examine the effects of different structure of correlation matrices on the data set with repeated measurement which was taken from an animal research study. It is aimed to determine which structure of correlation matrix can be suitable for animal research with repeated measurements for model fitting.

2. MATERIAL AND METHOD

The used data to evaluate the GEE was taken from a study on hair goats in Amasya province of Turkey. In this study, 456 records of 114 animals with 4 repeated measurements were used. Milk yield was selected as response variable and categorized to three groups as 0 (good) for greater than 610 g/day, 1 (average) for between 420 and 610 g/day, and 2 (bad) for smaller than 420 g/day. This categorization made by use of economical criterion. Explanatory variables were determined as age of mother, hair color, whether horned or not, structure of ears, whether with tassel or not, and type of birth. These variables categorized as; to categorize the age of mother, we used original age form 3 to 7. We used original hair colors as black, white, grey, brown and piebald. We use 0 for horned goats and 1 for hornless goats. Structure of ears were categorized as 0 for average ear length, 1 for long ear length and 2 for short ear length. 0 for with tassel and 1 for without tassel goats. Type of birth was categorized as 0 for twins and 1 for singles.

For the GEE, the variables included in the *X* could be continuous or categorical, and the model can include additive, interactive, and nested effects among these predictors. It is possible to define estimating equations which are consistent estimator of the regression parameters β . The generalized estimating equations are;

$$\left(\frac{\partial\mu}{\partial\beta}\right)^T V^{-1}(y-\mu)=0\,. \qquad (1)$$

Where, μ is the *n* x 1 vector of the mean expected responses whose element $\mu_i(=E[y_i])$ is given by $g^{-1}(X_i^T\beta)$. Here *V* is the variance covariance matrix and can be defined as and can be used to estimate regression parameters;

$$V=\phi A^{1/2}RA^{1/2}\,, \qquad (2)$$

where, *A* is the *n x n* diagonal matrix defined by $diag\{\phi\gamma(E[y_i]\}$, that is a matrix with all its elements null except the diagonal which contains the variance of the *n* observations expected under the marginal model, and *R* is the correlation matrix of the elements of *y* (Paradis and Claude, 2002; Carl and Kühn, 2007). GEE can be solved through an iterative process which can be summarized as follows;

1. Compute an initial estimate of β, for example, with a GLM.
2. Compute an estimate of the variance covariance matrix using Equation 2.
3. Update β with;

$$\beta_{step+1}=\beta_{step}-\left[\left(\frac{\partial\mu}{\partial\beta}\right)^T V^{-1}\frac{\partial\mu}{\partial\beta}\right]^{-1}\left[\left(\frac{\partial\mu}{\partial\beta}\right)^T V^{-1}(y-\mu)\right]$$

4. Alternate between steps 2 and 3 until convergence (Paradis and Claude, 2002).

To compare models it is necessary to have a criterion, this is Quasi-likelihood under the independence model criterion (QIC) which is a modification to Akaike Information Criterion (AIC) for GEE, where the likelihood function value in AIC is replaced by the quasi likelihood function value obtained under $R_i(\theta)=I$ and the penalty term is adjusted. QIC is defined as;

$$QIC=-2Q(B)+2trace(\Gamma^{-1}\Psi) \qquad (3)$$

Where, *Q(B)* is the value of the quasi-likelihood under the independence assumption, computed by the GEE estimator of *B* (matrix of unknown coefficients) based on any working correlations. The second term in equation 3 reflects the degree of the differences between the pure and robust covariance estimates of *B*, which indicates how much the working covariance matrix is consistent with the true covariance matrix. A model that minimizes QIC is regarded as the most appropriate one among fitted models (Hwang and Takane, 2005).

In GEE, if there are repeated measurements in the data set, correlation matrix can take different structures and these structures can affect the GEE solutions. Five different structure of correlation matrices can be used in GEE method. These are; independent, first-order autoregressive (AR(1)), exchangeable, unstructured and m-dependent. These can be summarized as;

Independent: Working correlation matrix $R_i(\alpha)=I$ is a *T x T* identity matrix. This “working independence” assumption is equivalent to assuming no intra-cluster correlation and yields estimates equivalent to those from simple “pooled” models. No estimate of α is obtained, since the intra-cluster correlation is assumed to be zero (Zorn, 2001). In that case, repeated measurements are uncorrelated.

AR(1): Repeated measurements have a first-order autoregressive relationship. The correlation between any two elements is equal to *r* for adjacent elements, r^2 for elements that are separated by a third, and so on. *r* is constrained so that –1< *r* <1. This autoregressive specification of working correlation matrix can be shown as $R_i(\alpha)=\rho^{|t-s|},t\ \ s$. Here, the within-observation correlation over time is an exponential function of the previous one (Zorn, 2001).

Exchangeable: This structure has homogenous correlations between elements. It is also known as a compound symmetry structure. This correlation can be shown as $R_i(\alpha)=\rho,t\ \ s$. For the exchangeable correlation structure Y_i values are assumed to covary equally across all observations within a cluster. In this specification, α is a scalar estimated by the model (Zorn, 2001).

M-dependent: Consecutive measurements have a common correlation coefficient, pairs of measurements separated by a third have a common correlation coefficient, and so on, through pairs of measurements separated by $m-1$ other measurements. Measurements with greater separation are assumed to be uncorrelated. When choosing this structure, specify a value of m less than the order of the working correlation matrix. This correlation is assumed in a random-effects model. $R_i(\alpha)$ can be chosen so that

$$[R_i]_{jk} = \begin{cases} \alpha^{\left|t_{ij}-t_{ik}\right|}, & \left|t_{ij}-t_{ik}\right| \le m, \\ 0, & \left|t_{ij}-t_{ik}\right| > m \end{cases} , \qquad (4)$$

where, t_{ij}, t_{ik} are the *j*th and *k*th observation times for the *i*th subject. This is the correlation structure for a stationary *m*-dependent process (Zeger and Liang, 1986).

Unstructured: Unstructured (or "pairwise") correlation structure is $R_i(\alpha) = \alpha_{st}, t\ \ s$. Here, no constraints are placed on the correlations across observations within a cluster; instead, they are estimated from the data without restriction. In this context, α is a T x T matrix containing the $T(T-1)/2$ unique pairwise correlations for all possible combinations of time points (Zorn, 2001).

3. RESULTS AND DISCUSSION

According to our main interest, is the differences found in the estimates across the different working correlation matrices? To compare the effects of different correlation matrices; QIC, test of model effects and parameter estimates results of GEE analysis on specified data set were given. Quasi-likelihood under the independence model criterion (QIC) values for five different working correlation matrices were given in Table 1.

Table 1. QIC values for five different working correlation matrices

	QIC
Independent	903.067
AR(1)	903.665
Exchangeable	903.067
Unstructured	949.124
m-Dependent	907.051

According to Table 1, the models which were utilized with independent and exchangeable working correlation matrices have minimum QIC values. These two models have equal QIC values; hence, these two models are the most appropriate solution for given data set. The model with AR(1) has small QIC value. The worst model has 949.124 QIC value, which was utilized with unstructured working correlation matrix.

To examine the effects of models on significance of variables, statistical significance of model effects for different working correlation matrices were given in Table 2.

Table 2 shows that intercept was found highly significant for all models. Mother age was found statistically significant for all models except the model with unstructured working correlation matrix. Hair color, horn and ear variables were found statistically insignificant for all models. Tassel was found statistically significant for all models and, models with independent and exchangeable working correlation matrices had smallest significance levels. Type of birth variable was found statistically significant only by model with unstructured working correlation matrix, which is unreliable with maximum QIC value.

These results indicated that only the model with unstructured working correlation matrix changes the decisions on significance of mother age and type of birth variables. The other models have equal significance levels on intercept and mother age variables, and have similar significance levels with small differences which is not affect the decision on other variables.

To analyze the effects of different correlation matrices on specified animal data, interpreting of parameter estimates can be helpful. Parameter estimates for different working correlation matrices were given in Table 3.

When the parameter estimates are interpreted, it can be easily understood that independent and exchangeable correlation matrices produced same results. For variable of mother age, according to independent and exchangeable models age 3, 5 and 6 are statistically significant. In the AR(1) model all ages except 7 are found statistically significant. While all ages are found statistically significant by m-dependent model, they are found insignificant by unstructured model. Mother age of 5 was found as the more effective age on milk yield by all interested models.

Table 2. Significance of model effects for different working correlation matrices

	Independent	**AR(1)**	**Exchangeable**	**Unstructured**	**m-Dependent**
Intercept	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Mother Age	**0.000**	**0.000**	**0.000**	0.175	**0.000**
Color	0.274	0.259	0.274	0.368	0.263
Horn	0.226	0.250	0.226	0.644	0.291
Ear	0.282	0.268	0.282	0.475	0.243
Tassel	**0.013**	**0.016**	**0.013**	**0.024**	**0.024**
Type of Birth	0.599	0.456	0.599	**0.016**	0.322

Table 3. Parameter estimates for different working correlation matrices

	Independent			AR(1)			Exchangeable			Unstructured			m-Dependent		
	$\hat{\beta}$	SE	Sig.	$\hat{\beta}$	SE	Sig.	$\hat{\beta}$	SE	Sig.	$\hat{\beta}$	SE	Sig.	$\hat{\beta}$	SE	Sig.
Intercept	**1.709**	**0.345**	**0.000**	**1.779**	**0.335**	**0.000**	**1.709**	**0.345**	**0.000**	**1.351**	**0.385**	**0.000**	**1.866**	**0.327**	**0.000**
MA(3)	**-0.766**	**0.272**	**0.005**	**-0.768**	**0.254**	**0.003**	**-0.766**	**0.272**	**0.005**	-0.171	0.325	0.600	**-0.771**	**0.234**	**0.001**
MA(4)	-0.475	0.268	0.077	**-0.479**	**0.251**	**0.047**	-0.475	0.268	0.077	-0.033	0.332	0.921	**-0.525**	**0.230**	**0.023**
MA(5)	**-0.973**	**0.281**	**0.001**	**-0.993**	**0.267**	**0.000**	**-0.973**	**0.281**	**0.001**	-0.390	0.330	0.237	**-1.018**	**0.252**	**0.000**
MA(6)	**-0.580**	**0.277**	**0.036**	**-0.594**	**0.261**	**0.023**	**-0.580**	**0.277**	**0.036**	-0.187	0.345	0.588	**-0.610**	**0.242**	**0.012**
MA(7)	-0.485	0.277	0.080	-0.507	0.261	0.052	-0.485	0.277	0.080	0.132	0.343	0.701	**-0.536**	**0.243**	**0.028**
C(Black)	0.179	0.226	0.429	0.168	0.232	0.469	0.179	0.226	0.429	0.038	0.204	0.852	0.154	0.241	0.524
C(White)	-0.045	0.248	0.856	-0.062	0.253	0.806	-0.045	0.248	0.856	-0.222	0.283	0.432	-0.084	0.263	0.748
C(Piebold)	0.174	0.261	0.505	0.162	0.267	0.543	0.174	0.261	0.505	0.106	0.246	0.667	0.146	0.275	0.596
C(Grey)	0.314	0.245	0.201	0.302	0.249	0.226	0.314	0.245	0.201	0.295	0.247	0.233	0.285	0.257	0.267
Horn(Yes)	-0.115	0.095	0.226	-0.109	0.095	0.250	-0.115	0.095	0.226	-0.070	0.151	0.644	-0.101	0.096	0.291
Ear(Average)	-0.171	0.107	0.112	-0.173	0.107	0.106	-0.171	0.107	0.112	-0.066	0.138	0.632	-0.176	0.108	0.104
Ear(Long)	-0.062	0.106	0.555	-0.082	0.104	0.435	-0.062	0.106	0.555	0.180	0.186	0.333	-0.105	0.105	0.317
Tassel(Yes)	**-0.384**	**0.156**	**0.013**	**-0.373**	**0.156**	**0.016**	**-0.384**	**0.156**	**0.013**	**-0.411**	**0.182**	**0.024**	**-0.361**	**0.160**	**0.024**
TB(Twin)	-0.058	0.110	0.599	-0.081	0.109	0.456	-0.058	0.110	0.599	**0.370**	**0.154**	**0.016**	-0.109	0.110	0.322

MA: Mother Age, C: Color, TB: Type of Birth, SE: Standard Error of $\hat{\beta}$ estimates

Bold characters indicate the significant estimations.

Color, horn, and ear were not discussed here because these variables were found insignificant by all models. Tassel variable was found statistically significant by all interested models. Independent and exchangeable models yield the minimum significance levels for this variable. According to the results does with tassel produce more milk than others. Variable of birth type was determined as statistically significant by only unstructured model which is unreliable because of high QIC value. According to the more reliable independent and exchangeable models, the highest milk yield can be taken from does with mother age of five and with tassel animals. Color, horn, ear structure and birth type can not affect the milk yields.

4. CONCLUSION

In this study, we demonstrated the application of GEE method to animal science data. And also we compared the five different correlation structures on repeated data set. Our findings show that independent and exchangeable correlation structures are the appropriate choices for biological data sets. These results support the findings of Abdel-Aty and Abdalla (2004) that Autoregressive correlations between repeated choices do not accurately describe the nature of the correlation between repeated choices. And also support the study of Park and Shin (1999) stated that in many practical situations, it is commonly observed that the independent correlation yields quite consistent results with those of other working correlations. Some researches such as Hadgu and Koch (1999) claim that any working correlating structure can be specified, and yet regression coefficients estimates are still consistent even when the correlation structure is misspecified. But results of this study showed that selection of correlation structure should be essentially chosen. Otherwise, results may be misleading to the researchers and producers who put into practice these research results. Pan and Connett (2002) declared that using an appropriate working correlation structure may improve efficiency of estimation like this study. Further studies may be focused on use of generalized estimating equations with Resampling methods such as permutation tests and correction of errors due to misclassification.

5. ACKNOWLEDGEMENTS

The authors wish to thanks Prof. Dr. Yüksel Bek for helpful discussions and comments.

6. REFERENCES

Abdel-Aty, M, Abdalla, M. F., 2004, Modeling Drivers' Diversion from Normal Routes under ATIS Using Generalized Estimating Equations and Binomial Probit Link Function. Transportation 31: 327 – 348.

Carl, G., Kühn, I., 2007, Analyzing Spatial Autocorrelation in Species Distributions Using Gaussian and Logit Models. Ecological Modelling. 207: 159 – 170.

Hadgu, A, Koch, G., 1999, Application of Generalized Estimating Equations to a Dental Randomized Clinical Trial. Journal of Biopharmaceutical Statistics. 9(1): 161 – 178.

Hammill, B.G., Preisser, J.S., 2006, A SAS/IML Software Program for GEE and Regression Diagnostics. Computational Statistics & Data Analysis. 51: 1197 – 1212.

Hosmer, D. W., Lemeshow, S., 2000, Applies Logistic Regression: Second Edition. John Willey & Sons, Inc. New York. 375 page.

Hwang, H., Takane, Y., 2005, Estimation of Growth Curve Models with Structured Error Covariances by Generalized Estimating Equations. Behaviormetrica. 32(2): 155 – 163.

Pan, W., Connett, J. E., 2002, Selecting the Working Correlation Structure in Generalized Estimating Equations with Application to the Lung Health Study. Statistica Sinica 12: 475 – 490.

Paradis, E., Claude, J., 2002, Analysis of Comparative Data Using Generalized Estimating Equations. J. Ther. Biol. 218: 175 – 185.

Park, T., Shin, D., 1999, On the Use of Working Correlation Matrices in the GEE Approach for Longitudinal Data. Commun. Statist. – Simula. 28(4): 1011 – 1029.

Reboussin, B.A., Lohman, K.K., Wolfson, M., 2006, Modeling Adolescent Drug-Use Patterns in Cluster-Unit Trials with Multiple Sources of Correlation Using Robust Latent Class Regressions. AEP. 16(11): 850 – 859.

Sturdivant, R. X., Hosmer, D. W., 2007, A Smoothed Residual Based Goodness-Of-Fit Statistic for Logistic Hierarchical regression Models. Computatioan Statistic & Data Analysis. 51: 3898 – 3912.

Ward, P., Myers, R.A., 2007, Bait Loss and Its Potential Effects on Fishing Power in Pelagic Longline Fisheries. Fisheries Research. 86: 69 – 76.

Zeger, S. L., Liang, K. Y., 1986, Longitudinal Data Analysis for Discrete and Continuous Outcomes. Biometrics, 42: 121 – 130.

Zorn, C. J. W., 2001, Generalized Estimating Equation Models for Correlated Data: A Review with Applications. American Journal of Political Science, 45(2): 470 – 490.