The Journal of Animal & Plant Sciences, 25(6): 2015, Page: 1543-1548
ISSN: 1018-7081

PERFORMANCE OF SEVERAL COVARIANCE STRUCTURES WITH MISSING DATA IN REPEATED MEASURES DESIGN

E. Eyduran†[1] and Y. Akbas[2]

[1]Igdir University, Faculty of Agriculture, Department of Animal Science, 76000, Igdir, Turkey
[2]Ege University, Faculty of Agriculture, Department of Animal Science, Bornova-İzmir, Turkey
†Corresponding Author: ecevit.eyduran@gmail.com
*The study is a part of PhD thesis of the first author

ABSTRACT

This investigation was carried out on annual amounts of wheat production from 65 provinces in seven geographical regions of Turkey during the years 1982 and 1999 to test performance of candidate covariance structures such as: Compound Symmetry (CS), Heterogenous Compound Symmetry (CSH), Unstructured (UN), Huynth Feldth (HF), First-Order Autoregressive (AR(1)), Heterogenous First-Order Autoregressive (ARH(1)), First-Order Ante-Dependence ANTE(1), Toeplitz (TOEP) and Heterogenous Toeplitz (TOEPH) specified for the missing repeated measures data with two fixed factors, viz., region (7 levels) and year (18 levels) using mixed model. In the generation of three missing data sets, deletion operations for the wheat production data were performed at three proportions (10%, 20%, and 30%) on the basis of Missing Completely at Random (MCAR), regardless of any factors under the assessment. The covariance structures were compared with Akaike's Information Criterion (AIC), Shwartz's Bayesian Criterion (SBC), and Corrected Akaike's Information Criterion (AICC) criteria. It was determined that CS was the covariance that produced the best fit among candidate covariance structures with -1158, -1153, -1158 at 10(%) missing data; -916, -912, and -916 at 20(%) missing data and -696, -692, and -696 at 30(%) missing data for AIC, SBC, and AICC, respectively. In conclusion, the investigation results illustrated that several covariance structures in the MIXED procedure of SAS program were easily specified for the repeated measures analysis of the missing data with more time levels.

Key words: Repeated Measures Design, Missing Data, Mixed procedure, Covariance Structure.

INTRODUCTION

The repeated measures are referred to as multiple measurements obtained from same experimental units (subjects) over time (Barcikowski and Robey, 1984; Looney and Stanley, 1989; Gürbüz *et al.,* 2003). Such repeated measures are analyzed with respect to a repeated measures design. Unlike ANOVA methods, the assumption "independency of observations" is not met in the repeated measures designs (Gürbüz *et al.,* 2003). In the non-missing repeated measures data, analysts prefer more advantageous and convenient one among univariate (Repeated ANOVA=RANOVA), multivariate (Profile Analysis (PA) and Mixed Model (MM) approaches whether or not Spherity assumption, the significant assumption of repeated measures data, is satisfied (Eyduran and Akbaş, 2010). Historically, Keselman *et al.* (1993; 1998; 1999; 2000; 2001; 2003) have documented different aspects of repeated measures designs.

Many factors are available about the selection of the suitable approach for repeated measures data with two factors, treatment and time: i) the structure of data set (balanced or unbalanced), ii) distribution of the studied trait, iii) type of variance-covariance matrices of data, iv) level of the correlation within subjects, v) the violation or satisfaction of Spherity assumption, vi) number of subjects in each level of treatment factor and vii) the non-missing or missing data etc. In the satisfaction of Spherity assumption, RANOVA is, in general, quietly sufficient to analyze non-missing repeated data (Orhan *et al.*, 2010). Conversely, in the violation of the assumption, the simplest solution for researchers is usually to apply adjusted H-F and G-G epsilon approaches for non-missing repeated measures data as a better alternative of RANOVA, a traditional method for correlated groups design. Profile analysis (PA), a multivariate extension of RANOVA, is more preferable in comparison with the adjusted epsilon approaches, and more principally, RANOVA for the non-missing data (Eyduran *et al.*, 2013). A good example regarding PA application could be found in the study of Mendes *et al.* (2005). Eyduran *et al.* (2008) suggested the specification of PA instead of RANOVA in violation of Spherity assumption for non-missing data. More sophistically, Tabachnick and Fidell (2001) addressed the ideal performance of PA when number of subjects (n) per treatment factor is higher than number of repeated measurements (p). A substantial limitation is that the previous approaches can be specified for only non-missing data. With these reasons, contrary to the previous approaches, the most important advantage of mixed model (MM) approach is that it is used more appropriately for missing and non-missing repeated measures data (Eyduran *et al.*, 2013). Compared with the

RANOVA and PA, MM with the development of the high computer technology offers more precise estimation with the specialization of discrepancy covariance structures on trend and variation of multiple sequential measurements taken repeatedly on a trait from experimental units over certain periods of time, and considers sources of variation both between and within experimental units (Littell *et al.,* 1998).

A definition on the trend of the annual wheat production across the years is quite imperative for further inspection. The time series analysis, is commonly, suggested to gain information about the trend. Whereas, mixed model specifying many covariance structures explaining variation between and within units can be a technically powerful choice for repeated measures designs at non-missing data, and in particular, missing data, to econometrically predict the trend of wheat amounts produced annually across the time periods in a consistent manner. Much more attention on the adaptation of the mixed model to the time series data should be paid on the scope of missing repeated measures data, which may provide an alternative of including extra factors and covariates into the general linear mixed model (Ser *et al.,* 2013).

In literature, more investigations on the specification of MM methodology with an assortment of covariance structures are still uncommon for missing repeated measures data. In annual wheat production of Turkey, performance of MM approach and the selection of the best covariance structure have not been yet evaluated for missing repeated measures data. Consequently, the goal of this investigation is to select the best covariance structure for very high levels (18 levels) of time factor together with a treatment factor (7 levels) for missing repeated measures (10%, 20%, and 30% missing).

MATERIALS AND METHODS

Data collection: The present data on 1170 wheat production values collected from 65 provinces with non-missing experimental units in seven geographical regions of Turkey over 18 years from 1982 until 1999 were provided (TURKSTAT, 1982-1999), and then regulated to non-missing data at repeated measures design. The considered production data from Eyduran and Akbas (2010) are depicted in Table 1.

Table1. The regions and provinces where data were obtained (Adapted from Eyduran and Akbas, 2010)

Central Anatolia	Mediterranean Sea	Aegean	Marmara	BlackSea	Southeast Anatolia	Eastern Anatolia
Ankara	Adana	Afyon	Balıkesir	Amasya	Adıyaman	Agrı
Cankırı	Antalya	Aydın	Bilecik	Artvin	Diyarbakır	Bingol
Eskisehir	Burdur	Denizli	Bursa	Bolu	Gaziantep	Bitlis
Kayseri	Hatay	İzmir	Canakkale	Corum	Mardin	Elazıg
Kırsehir	Isparta	Kutahya	Edirne	Giresun	Siirt	Erzincan
Konya	Icel	Manisa	Istanbul	Gümüshane	Sanlıurfa	Erzurum
Nevsehir	Maras	Mugla	Kırklareli	Kastamonu		Hakkari
Nigde		Usak	Kocaeli	Ordu		Kars
Sivas			Sakarya	Samsun		Malatya
Yozgat			Tekirdag	Sinop		Muş
				Tokat		Tunceli
				Zonguldak		Van

In the study, the major attention has been paid to statistically test performance of candidate covariance structures structures such as: Compound Symmetry (CS), Heterogenous Compound Symmetry (CSH), Unstructured (UN), Huynth Feldth (HF), First-Order Autoregressive (AR(1)), Heterogenous First-Order Autoregressive (ARH(1)), First-Order Ante-Dependence ANTE(1), Toeplitz (TOEP) and Heterogenous Toeplitz (TOEPH) specified for the missing repeated measures with two fixed factors, region (7 levels) and time (18 levels) using mixed model. Each of the provinces belonging to each geographical region is accepted to be an experimental unit. The following model for unbalanced data can be adopted:

$$Y_{ijk} = \text{«} + \sim_{j} + \quad_{k} + \left(\sim \quad\right)_{jk} + \text{®}_{i(j)} + e_{ijk} \qquad (1)$$

(k=1, 2,..., 18; j=1, 2,...,7 and i=1, 2,...,n_i)

Where;

μ : Grand mean of annual wheat production

α_j : the effect of *j* th level of “geographical region” factor on annual wheat production

β_k : *k*th year effect on annual wheat production,

$\alpha\beta_{jk}$: the effect of geographical region by year interaction on annual wheat production,

$\text{®}_{i(j)}$: the fixed effect of ith province (experimental unit) in jth geographical region

e_{ijk} : random error term (Orhan *et al*., 2010).

In the model defined above, other included effects with the exception of the term "e_{ijk}" were taken into consideration as fixed effect.

The basic assumption "multivariate normal distribution" was satisfied for the data. In the generation of three missing data sets, deletion operations for the wheat production data were made for three proportions (10%, 20%, and 30%) with Missing Completely at Random (MCAR), regardless of any factors. Performance of the covariance structures was measured with Akaike's Information Criterion (AIC), Shwartz's Bayesian Criterion (SBC), and Corrected Akaike's Information Criterion (AICC) criteria (Eyduran and Akbas, 2010). In the research, the MIXED procedure of SAS program with the specialization of different covariance structures was executed for three sets of missing repeated measures data. DDFM=Containment was also specified in the MIXED model procedure. A candidate covariance structure whose goodness of fit criteria is the nearest to zero is considered to be the best covariance structure (Eyduran and Akbas, 2010). Likelihood ratio test was employed to test validity of the least square method. When probability of the test is less than 0.05, use of the least square method is inadvisable.

RESULTS

With the deletion of 10% of data, performance of covariance structures for wheat production was evaluated and results are given in Table 2 with regard to goodness of fit criteria. Among the plausible candidate covariance structures, CS illustrated the best performance, which may be ascribed to logarithmic transformation of the production data. Consequently, the transformation noticeably narrowed very wide variation. However, TOEP was the worst covariance structure. In consideration of AIC and AICC goodness of fit criteria, a similar fit was presented by CS and HF structures of covariance, despite the fact that HF was slightly better in SBC criterion than CS. The estimate of SBC criterion of CSH covariance type was a bit superior, compared with the estimate of CS. No output was obtained for TOEPH because of infinite likelihood and UN owing to insufficient memory. For all the structures, the estimates of SBC, by depending upon sample size of the studied data, were found lower than those of the other goodness of fit criteria.

Significance results of fixed effects like region, year, and their interaction for the studied covariance structures converged in MIXED procedure of SAS software for 10% missing data are summarized in Table 3. The handled fixed effects were noted very strongly significant (P<0.001).

Table 2. Goodness of fit criteria of covariance structures for 10% missing data

	ANTE(1)	**AR(1)**	**ARH(1)**	**CS**	**CSH**	**HF**	**TOEP**
-2 Res Log Likelihood	-1464	-1357	-1387	-1162	-1214	-1198	-1514
Akaiki Information Criterion (AIC)	-1394	-1353	-1349	-1158	-1176	-1160	-1478
Burnham Handerson Criterion (AICC)	-1391	-1353	-1349	-1158	-1175	-1159	-1477
Shwartz Bayes Criterion (SBC)	-1318	-1348	-1308	-1153	-1135	-1119	-1438
Likelihood Ratio Chi-Square value	2685	2578	2609	2383	2435	2419	2735
Likelihood Ratio Test DF	34	1	18	1	18	18	17
Likelihood Ratio Test Prob. (P)	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

CS: Compound Symmetry, CSH: Heterogenous Compound Symmetry, UN: Unstructured, HF: Huynth Feldth, AR(1): First-Order Autoregressive, ARH(1): Heterogenous First-Order Autoregressive, ANTE(1): First-Order Ante-Dependence, TOEP: Toeplitz and TOEPH: Heterogenous Toeplitz.

Table 3. Results of fixed effects under several covariance structures for 10 % missing data

Covariance Structure	**Region**		**Year**		**Region x Year Interaction**	
	F	**P**	**F**	**P**	**F**	**P**
ANTE (1)	5.65	<0.0001	10.68	<0.0001	3.02	<0.0001
AR(1)	5.29	<0.0001	13.62	<0.0001	2.93	<0.0001
ARH(1)	4.84	<0.0001	12.45	<0.0001	2.93	<0.0001
CS	5.09	<0.0001	6.87	<0.0001	2.09	<0.0001
CSH	5.10	<0.0001	7.08	<0.0001	2.10	<0.0001
HF	2.90	0.0083	6.94	<0.0001	2.12	<0.0001
TOEP	4.96	<0.0001	12.79	<0.0001	2.99	<0.0001

CS: Compound Symmetry, CSH: Heterogenous Compound Symmetry, UN: Unstructured, HF: Huynth Feldth, AR(1): First-Order Autoregressive, ARH(1): Heterogenous First-Order Autoregressive, ANTE(1): First-Order Ante-Dependence, TOEP: Toeplitz and TOEPH: Heterogenous Toeplitz.

Performance of the six covariance structures for transformed wheat production with 20% missing data regarding goodness of fit criteria is given in Table 4. Statistically ascribed to transformed data, the much more narrowed variation caused the better performance of CS structure, because of the fact that very similar correlation coefficients were estimated in the annual transformed production data between pairs of years under two-factor repeated measures design. In harmony with 20% missing data, the worst performance was again obtained by TOEP. Of the tested covariance structures, the estimates of fixed effects and goodness of fit criteria for HF, TOEPH, and UN were not estimated due to infinite likelihood and non-positive definite Hessian matrix. Similarly, SBC goodness of fit criteria estimates were closer to zero than the estimates from AIC and AICC criteria.

Significance values of the fixed effects for the fitted covariance structures to the 20% missing data are given in Table 5. For the covariance structures converged, the fixed effects were significantly found (P<0.001).

Table 4. Goodness of fit criteria of covariance structures for 20 % deletion proportion.

	ANTE(1)	AR(1)	ARH(1)	CS	CSH	TOEP
-2 Res Log Likelihood	-1170	-1104	-1126	-920	-981	-1243
Akaiki Information Criterion (AIC)	-1100	-1100	-1088	-916	-943	-1207
Burnham Handerson Criterion (AICC)	-1097	-1100	-1087	-916	-942	-1206
Shwartz Bayes Criterion (SBC)	-1024	-1096	-1046	-912	-902	-1168
Likelihood Ratio Chi-Square value	2273	2208	2229	2023	2084	2347
Likelihood Ratio Test DF	34	1	18	1	18	17
Likelihood Ratio Test Prob. (P)	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

CS: Compound Symmetry, CSH: Heterogenous Compound Symmetry, AR(1): First-Order Autoregressive, ARH(1): Heterogenous First-Order Autoregressive, ANTE(1): First-Order Ante-Dependence, TOEP: Toeplitz

Table 5. Results of fixed effects under candidate covariance structures for 20 % deletion proportion.

Covariance Structure	Region		Year		Region x Year Interaction	
	F	P	F	P	F	P
ANTE (1)	5.68	<0.0001	8.95	<0.0001	2.40	<0.0001
AR(1)	5.34	<0.0001	11.05	<0.0001	2.36	<0.0001
ARH(1)	5.13	<0.0001	10.57	0.0005	2.42	<0.0001
CS	5.11	<0.0001	5.06	<0.0001	1.92	<0.0001
CSH	5.12	<0.0001	5.12	<0.0001	2.00	<0.0001
TOEP	4.98	<0.0001	9.52	<0.0001	2.31	<0.0001

CS: Compound Symmetry, CSH: Heterogenous Compound Symmetry, AR(1): First-Order Autoregressive, ARH(1): Heterogenous First-Order Autoregressive, ANTE(1): First-Order Ante-Dependence, TOEP: Toeplitz

The results of goodness of fit criteria and fixed effects estimates for candidate covariance structures are presented in Tables 6 and 7, respectively, at 30% of missing data. CS gave the best performance in the goodness of fit criteria, but ANTE (1) has the worst performance in AIC and AICC criteria. The second best performance after CS was provided from CSH. No numerical solution was produced from the other candidate structures, i.e. HF, TOEP, TOEPH, and UN.

Table 6. Goodness of fit criteria of covariance structures for 30 % deletion proportion.

	ANTE(1)	AR(1)	ARH(1)	CS	CSH
-2 Res Log Likelihood	-927	-839	-877	-700	-759
Akaiki Information Criterion (AIC)	-857	-835	-839	-696	-721
Burnham Handerson Criterion (AICC)	-853	-835	-837	-696	-719
Shwartz Bayes Criterion (SBC)	-781	-831	-797	-692	-679
Likelihood Ratio Chi-Square value	1890	1802	1839	1663	1721
Likelihood Ratio Test DF	34	1	18	1	18
Likelihood Ratio Test Prob. (P)	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

CS: Compound Symmetry, CSH: Heterogenous Compound Symmetry, AR(1): First-Order Autoregressive, ARH(1): Heterogenous First-Order Autoregressive, ANTE(1): First-Order Ante-Dependence

Table 7. Results of fixed effects under candidate covariance structures for 30 % deletion proportion.

Covariance Structure	Region		Year		Region x Year Interaction	
	F	P	F	P	F	P
ANTE (1)	5.44	<0.0001	7.75	<0.0001	2.07	<0.0001
AR(1)	5.19	<0.0001	7.71	<0.0001	2.04	<0.0001
ARH(1)	5.01	<0.0001	7.33	<0.0001	2.08	<0.0001
CS	5.09	<0.0001	3.99	<0.0001	1.69	<0.0001
CSH	5.08	<0.0001	3.94	<0.0001	1.75	<0.0001

CS: Compound Symmetry, CSH: Heterogenous Compound Symmetry, AR(1): First-Order Autoregressive, ARH(1): Heterogenous First-Order Autoregressive, ANTE(1): First-Order Ante-Dependence

DISCUSSION

Considering likelihood ratio test results for missing data proportions of 10%, 20%, and 30%, the general lineral models with various covariance structures converged in SAS software program were superior to the standard least squares method. Similarly, Akbas *et al.* (2001) and Keselman *et al.* (2003) emphasized that better results of multivariate approaches are obtained in general linear model instead of least squares method. Eyduran *et al.* (2013) reported the using advantages of MIXED procedure of SAS for missing/non-missing repeated measures data, and stressed that better results are obtained in goodness of fit criteria when the first period was considered as the covariate. More virtually, in practice, considering the missing data in repeated measures design should be emphasized at various fields.

Some investigators advised for having a good performance of SBC estimates for large samples. As the missing data proportion increased from 10% to 30%, the estimates of goodness of fit criteria became closer to zero as a result of reducing observation and data dimension. For all of the missing data proportions, CS was the best one among candidate covariance structures, whereas Eyduran and Akbas (2010) reported that CS was a better covariance structure for the same production values as non-missing data.

In general linear model, the other advantages of MIXED procedure with high computer technology are to propose more precise estimation by the specialization of special covariance structures, and to emphasize variation both between and within experimental units, in opposition to traditional approaches in repeated measures data.

Extra factors and covariates into the general linear mixed models may be added with fixed effects like treatment, time, and treatment x time interaction at missing repeated measures data (Rosario *et al.,* 2007; Ser s*et al.,* 2013). Additonally, Wang and Goonewardene (2004) specified the initial measurement period as a covariance with using MIXED models in the repeated measures data.

It was found that MIXED procedure of SAS with the specification of appropriate covariance structures could be used reliably for non-missing /missing data in the general linear models (Eyduran and Akbas, 2010) and general linear mixed models (Eyduran *et. al.,* 2013). However, it is recommended for researchers to consider a covariate, the first one of the repeated measurements, and to study much larger data sets in similar investigations.

The use of the missing data at repeated measures designs with identifying probable covariance structures should be investigated in detail, chiefly for the annual production defined over time periods, which may be helpful for further modeling investigation. Unlike previous studies, it is significant to evaluate performance of missing data on candidate covariance structures at a repeated measures design having more time periods as also seen in the study, in which a pioneer point of view for econometric studies was presented.

REFERENCES

Akbas, Y., M. Z. Fırat and C. Yakupoglu (2001). Agricultural Information Technology Symposium, Sütçü İmam University, Agricultural Faculty, Kahramanmaraş, 20:22.

Barcikowski, R.S. and R.R. Robey (1984). Decision in single group repeated measures analysis: Statistical tests and three computer packages. The American Statistician. 38(2): 148-150.

Eyduran, E. and Y. Akbas (2010). Comparison of different covariance structure used for experimental design with repeated measurement. J. Anim. Plant Sci. 20(1): 44-51.

Eyduran, E., K. Yazgan, and T. Özdemir (2008). Utilization of profile analysis in animal science. J. Anim. Vet. Adv., 7(7):796-798.

Eyduran, E., A. Tatliyer, A. Waheed and M.M. Tariq (2013). Determination of the Most Appropriate Covariance Structure for Data with Missing Observations in Repeated Measures Design. KSU J. Nat. Sci.16(3): 32-37.

Gürbüz, F., E. Başpınar, H. Çamdeviren and S. Keskin (2003). Analysis of repeated measures designs. Van (Turkey). 130 p.

Keselman, H. J., J. Algina, R. K. Kowalchuk and R. D. Wolfinger (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. Communications in Statistics: Simulation and Computation. 27(3): 591-604.

Keselman, H. J., J. Algine, R. K. Kowalchuk and R. D. Wolfinger (1999). The analysis of repeated measurements: a comparison of mixed model satterhwaite f tests and a nonpooled adjusted degrees of freedom multivariate test. Commun. Statist.-Theory Methods. 28(12): 2967-2999.

Keselman, H. J., K.C. Carriere and L.M. Lix (1993). Testing repeated measures hypotheses when covariance matrices are heterogeneous. J. Educational Statistics. 18(4): 305-319.

Keselman, H. J., R.R. Wilcox and L.M. Lix (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs. Psychophysiology. 40: 586-596.

Keselman, H.J., J. Algina, R.R. Wilcox and R.K. Kowalcuk (2000). Testing repeated measures hypotheses when covariance matrices are heterogeneous: Revisiting the robustness of the Welch-James test again. Educational and Psychological Measurement. 60(6): 925-938.

Keselman, H.J., J. Algine and R.K. Kowalchuk (2001). The analysis of repeated measures designs: A Review. Bri. J. Math. Stat. Psychology, 54:1-20.

Littell, R.C., P.R. Henry and C.B. Ammerman (1998). Statistical analysis of repeated measures data using SAS procedures. J. Anim. Sci. 76:1216-1231.

Looney, S. W. and W.B. Stanley (1989). Exploratory repeated measures for two or more groups. The American Statistician. 43(4): 220-224.

Mendes, M., A. Karabayır, I.E. Ersoy and C. Atasoglu (2005). Effects of three different lighting programs on live weight change of bronze turkeys under semi-ıntensive conditions. Arch. Tierz. Dummerstorf. 48(1): 86-93.

Orhan, H., E. Eyduran and Y. Akbaş (2010). Defining the best covariance structure for sequential variation on live weights of anatolian merinos male lambs. J. Anim. and Plant Sci. 20(3): 158-163.

Rosario, M. F., M.A.N. Silva., A.A.D. Coelho and V.J.M. Savino (2007). Estimating and predicting feed conversion in broiler chickens by modeling covariance structure. Intl. J. Poult. Sci., 6(7): 508-514.

Ser, G., B. Kaki, A. Yesilova and A. Yilmaz (2013). Comparison of the performance of different covariance structures and estimation methods in general linear mixed model. J. Anim. Prod., 54(2): 18-23.

Tabachnick, B. G. and L.S. Fidell (2001). Using Multivariate Statistics. Allyn& Bacon, USA. 966 p.

TURKSTAT, (1982-1999). Turkish Statistical Institute. Summaries of Agricultural Statistics

Wang, Z. and L. A. Goonewardene (2004). The use of mixed models in the analysis of animal experiments with repeated measures data. Can. J. Anim. Sci. 84(1):1-11.