Application of Linear Mixed-Effects Model in Saudization Ratios Data : A Case Study

Dr. Elsherbiny Shawki Elsayed

Application of Linear Mixed-Effects Model in Saudization Ratios Data : A Case Study

Dr. Elsherbiny Shawki Elsayed

Abstract: Linear mixed-effects models are indispensable tools for analyzing balanced and unbalanced structured data in educational, medical, actuarial and social behavioral research it gives us extra flexibility in developing the appropriate model for the data. This paper introduces a mixed two- way analysis of variance with fixed university effect and random time effect. The maximum likelihood estimates of the structural parameters are obtained. The proposed model is used to analyze Saudization ratios (Saudi faculty members in total)data in three public universities in Saudi Arabia .

Key words : Linear Mixed-Effects Model; Restricted Maximum Likelihood Estimators ;Saudization Ratios ; Theil Statistic.

1-Introduction

Generalized linear models (GLMs) have attracted considerable attention over the last years. GLMs are an extension of the linear modeling process that allows models to be fit to data that follow probability distributions other than the Normal distribution, such as Poisson, Binomial, Multinomial, and etc. Generalized Linear Models also relax the requirement of equality or constancy of variances that is required for hypothesis tests in traditional linear models. The GLMs have wide area of application in actuarial studies .For example , El Bassiouni (1991) introduced a mixed model for loss ratio analysis and assumed that the loss ratio to follow lognormal distribution. This model may be treated as a mixed two – way analysis of variance with fixed insurance company effect and random time effects. His proposed model is used to analyze loss ratio data from general insurance market in Kuwait .Gedalla et al (2006) indicated how the Generalized linear models can be used to drive rating models that apply to marine liability business. Hanafy (2007) introduced a mixed model to estimating the retention rates for property and casualty insurance companies in Egypt.

The mixed linear model is a generalization of the standard linear model used in the GLM procedure , the generalization being that the data are permitted to exhibit correlation and nonconstant variability . Mixed linear models provide the flexibility of modeling variances and covariance of variables in addition to means specified in a cross sectional regression model hence can be used to model data that show correlation and non-constant variability. Random effects parameters with non constant variability such as that shown with unbalanced time series cross sectional data (i.e. spatial repeated measures time series data, nested or clustered time series data) can be modeled easily and accurately with PROC MIXED in SAS which also provides a variety of covariance structures to model random-effects parameters with non constant variability. Traditionally mixed linear models were used to model a combination of fixed and random effects that led to the name mixed model.

In the social sciences the most common mixed linear models are multilevel models , but random coefficient models are important in much wider context , including biometrics and econometrics. El-Bassiouni & Charif (2004) proposed an invariant test that combines the most powerful invariant tests against small and large alternatives for testing a null variance ratio in mixed models with zero degrees of freedom for error .The test statistic could be easily computed and the corresponding test procedure is just as easy to carry out using currently available software. The Power of the test was compared with the power of other tests advocated in the literature using two real data sets and was found to maintain high efficiency all over the parameter space. Spilke et al (2004) described the use of the mixed procedure of the SAS System for the analysis of designed experiments. Special emphasis is given to the specification of options as depending on the assumed mixed model and on the unbalancedness in the data. Liu et al (2007) considered semi parametric regression model that relates a normal outcome to covariates and a genetic pathway, where the covariate effects are modeled parametrically and the pathway effect of multiple gene expressions is modeled parametrically or non parametrically using least-squares kernel machines (LSKMs).

Kinn & Dunson (2007) discussed the problem of selecting which variables should be included in the fixed and random components of logistic mixed effects models for correlated data. A fully Bayesian variable selection was implemented using a stochastic search Gibbs sampler to estimate the exact model-averaged posterior distribution. Thaddeus & Petkova (2007) presented a method of determining maximum likelihood estimators of principal points for linear mixed models and applied their results to an anti-depressant study to identify prototypical drug and placebo response profiles.

The objective of this paper is to introduce a linear mixed-effects model designed to be used in determining the Saudization ratio in three public universities(namely: King Faisal university, King Saud university and King Abdulaziz university).This paper is organized as follows: the mixed linear model is introduced in Section 2. The maximum likelihood estimators of the parameters are presented in section 3. In section 4, the predictive performance of the model will be tested. Then , in section 5 the proposed model is used to estimate the Saudi ratios for three public universities .

2-The Model

Let X_{ij} and p_{ij} denote the saudization ratio and the number of faculty members , respectively for the university i in year j($i=1,2,\ldots,a$, $j=1,2,\ldots,b$). Assume that X_{ij} has lognormal distribution. Of course it is necessary to chuck whether it is possible to describe the Saudization ratios by a lognormal distribution , but it suffices here to assert that the shape of the lognormal curve is appealing in this context and has been applied before to model ratios data like that ,see El Bassiouni (1991), Jiming & Sunil (2003) ,Katrien & Beirlan (2005) and Hanafy (2007).Also notice that , we analyze Saudization ratios at universities operating in the same field in the same country ,so it realistic to assume that the universities have fixed effects .Set $Y_{ij} = \ln X_{ij}$ and assume that ,

$$y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ij} \qquad (1)$$

Where μ is an general mean , α_i are unknown fixed effects due to university i , β_j are random effects due (time) to year j , $(\alpha\beta)_{ij}$ are the interaction between the effect of university i and the effect of (time) year j and e_{ij} are random errors . Notice that the β_j and e_{ij} are mutually independent normal random variables having zero mean and variances θ_2 and θ_1/p_{ij} respectively . Thus , the parameter space is given by :

$$\Theta = (\alpha_1 , \alpha_2, \ldots\ldots, \alpha_a, \theta_1, \theta_2 : \alpha_i \in R , i = 1 ,\ldots.a, \theta_1 \geq 0 , \theta_2 \geq 0)$$

The *i-th* university mean is given by : $\mu_i = \mu + \alpha_i + (\alpha\beta)_{ij}$ .Model (1) is called mixed two ways analysis of variance model or mixed randomize block design.

We can estimate the general mean μ as :

$$\hat{\mu} = \bar{y}_+ = \frac{\Sigma_{ij} y_{ij}}{ab} \qquad (2)$$

So, model (1) becomes :

$$y^*_{ij} = \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ij} \qquad (3)$$

Where , $y^*_{ij} = y - \bar{y}_+$

Let a=b and write model (1) in matrix form as :

$$Y = X\alpha + Z\beta + e \qquad (4)$$

where
Y is an $n \times 1$ vector of n observed records of y^*_{ij}
α is a × 1 vector of fixed effects
β is b × 1 vector of random effects
e is an $n \times 1$ vector of random, residual terms
X is a known *design matrix* of order $n \times a$, which relates the records in **y** to the fixed effects in **α**
Z is a known *design matrix* of order $n \times b$, which relates the records in **y** to the random effects in **β**. In the next section , the maximum likelihood estimates of fixed effects parameters and variance components for model (1) will be obtained.

3-The Maximum Likelihood Estimators

El Bassiouni (1991) introduced the following maximum likelihood estimation for the fixed effects parameters .Define the diagonal matrix P where :

$$p = \text{diag}\,(p_{11},\ldots\ldots\ldots, p_{ab}) \qquad (5)$$

Under the model assumption , we can prove that :

$$Y \sim N(X\alpha\,,\,\theta_1 P^{-1} + \theta_2 ZZ') \qquad (6)$$

Following Harville (1977) , the likelihood equation for α is given by :

$$XP^{1/2}\Sigma^{-1}P^{1/2}X\alpha = XP^{1/2}\Sigma^{-1}P^{1/2}Y \qquad (7)$$

From equation (7), we can get an estimate for the parameters α as follows :

$$\hat{\alpha} = \Phi^{-1}\lambda \qquad (8)$$

where Φ is *axa* matrix whose elements are given by:

$$\Phi_{rs} = p_{r+} - \textstyle\sum_{j=1}^{b}\rho_j p_{rj}^2 \qquad , r = s = 1,2,\ldots\ldots\ldots a$$

$$= -\textstyle\sum_{j=1}^{b}\rho_i p_{rj}\, p_{sj} \qquad r \neq s$$

Where $p_{r+} = \sum_{j=1}^{b} p_{ij}$ and λ is $ax1$ vector whose elements are given by:

$$\lambda_r = \textstyle\sum_{j=1}^{b} p_{rj}\,(Y_{ij}^{*} - \rho_j \sum_{i=1}^{a} p_{ij}\, Y_{ij}^{*}) \qquad , r = 1,2,\ldots\ldots\ldots\ldots,a$$

$$\text{where } \rho_j = \frac{\theta_2}{\theta_1 + \theta_2 \sum_{i=1}^{a} p_{ij}}$$

Since the maximum likelihood estimators of θ_1 *and* θ_2 take no account of the loss in degrees of freedom resulting from estimating α , we consider the restricted maximum likelihood method to estimate the variance components. The restricted likelihood equation for θ_1 is given by(Neumaier & Eildert ,1998) :

$$\theta_1 = \left(\textstyle\sum_{i=1}^{a}\sum_{j=1}^{b} p_{ij} Y_{ij} Z_{ij} - \sum_{i=1}^{a}(\sum_{j=1}^{b} p_{ij} Y_{ij})(\sum_{j=1}^{b} p_{ij} Z_{ij})/p_{i+}\right)/(n-a) \qquad (9)$$

Where , $Z_{ij} = Y_{ij} - \beta_j^{*}$ and $\beta_j^{*} = \rho_j \sum_{i=1}^{a} p_{ij}(Y_{ij} - \alpha_i)$,j=1,2,……,b

Also ,the restricted likelihood equation for θ_2 is given by:

$$\theta_2 = \textstyle\sum_{j=1}^{b}\beta_j^{*2} / (b - tr(Q)) \qquad (10)$$

Where $tr(Q) = \theta_1/\theta_2(\sum_{j=1}^{b}\rho_j + tr(\Phi^{-1}G))$ and G is *axa* matrix whose elements are given by : $G_{rs} = \sum_{j=1}^{b}\rho_j^2 p_{rj}\, p_{sj}$

The equations (8) ,(9) and (10) must be solved simultaneously for $\hat{\alpha}\,, \hat{\theta_1}$ *and* $\hat{\theta_2}$.

McCulloch & Searle (2000) suggested the iterative procedure to solve like these equations as follows . Set $\theta = (\theta_1, \theta_2)'$ and let $\theta^{(k)}, k = 1,2, \ldots$,denote the value produced by the the procedure on its k^{th} iteration. So we start the iteration by substituting an initial value $\theta^{(0)}$ into ,

$$\theta_1^{(k+1)} = \left(\sum_{i=1}^{a}\sum_{j=1}^{b} p_{ij} Y_{ij} Z_{ij}^{(k)} - \sum_{i=1}^{a}(\sum_{j=1}^{b} p_{ij} Y_{ij})(\sum_{j=1}^{b} p_{ij} Z_{ij}^{(k)})/p_{i+}\right)/(n-a) \quad (11)$$

And $$\theta_2^{(k+1)} = \sum_{j=1}^{b}(\beta_j^{*(k)^2})/(b - tr(Q^{(k)})) \quad (12)$$

Repeating the iteration until $\theta^{(k+1)}$ is sufficiently close to $\theta^{(k)}$ in some norm. If we have any prior information about θ, then we could use it to formulate an initial values for θ .Otherwise , we could use ANOVA estimators obtained from (9) and (10) assuming that $P = I_n$, as initial values (Breslow & Clayton ,1993). The procedure of computing the maximum likelihood estimates of parameters starts by obtaining initial estimates of variance components. These estimates of θ_1 and θ_2 are then substituted into (8) to estimate α. The estimate of α along with the initial estimates of θ_1 and θ_2 are then substituted into (11) and (12) to obtain $\theta^{(1)}$. This iterative process is to be continued until we achieve convergence after m iterations ,say , at which time we get $\hat{\theta} = \theta^{(m)}$ and $\hat{\alpha} = \alpha^{(m)}$ .After we get estimators of $\hat{\alpha}$, $\hat{\theta}_1$ and $\hat{\theta}_2$,we can predict the Saudi ratio for university i in year j using the method used in El –Bassiouni(1991) as follows:

$$\hat{X}_{ij} = \exp(\hat{\mu} + \hat{\alpha}_i^2(\hat{\theta}_2) + 0.5(\hat{\theta}_2 + \frac{\hat{\theta}_1}{p_{ij}})) \quad (13)$$

In the next section , the predictive performance of the mixed model described in section (2) will be tested using two measures, namely: Theil Statistic and Mean Square Error.

4-Testing the Performance of the Model

After estimating the parameters of model(1) ,We need to test the predictive performance of the model. This can be done by using the following measures (Zhang & Lin ,2002; Hanafy ,2007) :

- **Theil Statistic:**

$$U = \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(Y_t^s - Y_t^a)2}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(Y_t^s)2} + \sqrt{\frac{1}{T}\sum_{t=1}^{T}(Y_t^a)2}} \quad (14)$$

Where Y_t^s is the forecasted value of Y_t , Y_t^a is the actual value of Y_t , T is the number of observations and U always falls between 0 and 1.If U =0 that means the predictive performance of model is perfect and if the U = 1 that means the predictive performance of the model is bad.

- **Mean Square Error(MSE).**

MSE is the mean of the square difference between the estimated value and its actual value .MSE for model (1) could be estimated as:

$$MSE = \frac{1}{T}\sum_{t=1}^{T}\left(Y_t^s - Y_t^a\right)^2 \qquad (15)$$

5- Case Study

The data set used in this paper consists of the Saudization ratios(X_{ij}) and the number of faculty members (P_{ij})in three public universities in Saudi Arabia (namely; King Faisal university ;King Saud university and King Abdulaziz university) for college of business (COB) as theoretical college and college of computer &information technology (COCIT) as applied college. Data are derived from annual statements from the period from 2004/2005 to 2008/2009. The Saudization ratios during this period are given in Table (1), along with the associated data on the number of faculty members .It must be more realistic to assume that the three universities have fixed effects . Thus, we will applicant the mixed linear model for the analysis of this set of data.

Table (1):Saudi ratios and the number of faculty staff for the universities mentioned .

University	Year	COB		COCIT	
		P_{ij}	X_{ij}	P_{ij}	X_{ij}
King Faisal	2004/2005	56	0.482	20	0.350
	2005/2006	56	0.482	18	0.388
	2006/2007	59	0.474	22	0.318
	2007/2008	65	0.430	25	0.320
	2008/2009	118	0.322	29	0.379
King Saud	2004/2005	262	0.702	115	0.491
	2005/2006	271	0.609	116	0.324
	2006/2007	303	0.775	121	0.521
	2007/2008	295	0.789	126	0.490
	2008/2009	326	0.721	122	0.601
King Abdulaziz	2004/2005	349	0.498	76	0.329
	2005/2006	361	0.551	108	0.354
	2006/2007	366	0.603	119	0.427
	2007/2008	355	0.706	122	0.511
	2008/2009	362	0.785	118	0.443

The initial estimates, computed from equations (9) and (10) using the usual ANOVA estimators assuming that $P = I_n$, were $\hat{\theta}_1^{(0)} = 283.023$ and $\hat{\theta}_2^{(0)} = 1.016$. For King Faisal university , COB, the procedure converged to the following estimates : $\hat{\theta}_1 = 301.417$ and $\hat{\theta}_2 = 1.059$ .These estimates were computed from equations (11) and (12).Table (2) shows the estimation of the general means and the variance components for both COB and COCIT.

Table (2): Estimation of the general means and the variance components.

university	college	$\hat{\mu}$	Exp($\hat{\mu}$)	$\hat{\theta}_1$	$\hat{\theta}_2$
King Faisal	COB	-0.5641	0.5788	301.417	1.059
	COCIT	-0.1416	0.8679	198.560	0.089
King Saud	COB	-0.3609	0.6970	245.071	0.771
	COCIT	-0.0870	0.9166	133.758	0.061
King Abdelaziz	COB	-0.3513	0.7037	230.602	0.703
	COCIT	-0.0102	0.9899	120.117	0.058

From Table (2) , we conclude that, for the universities mentioned above the general means of Saudization ratios in COB are greater than whose of COCIT. Also , there are a negative relationship among the general mean of Saudization ratios and the variance of the effect of the year $\hat{\theta}_2$ and the variance of the error θ_1/p_{ij} .Substituting the values of $\hat{\theta}_1$ and $\hat{\theta}_2$ into equation(8) we can obtain an estimate of fixed effect parameters.

Table(3): Estimation of fixed effects parameters(α_i).

university	COB	COCIT
King Faisal	24.327	23.017
King Saud	31.002	29.132
King Abdelaziz	27.654	25.308

From table (3) , we can see that the fixed effect of the university has positive effect on the Saudization ratios for all the three universities .The fixed effects of King Saud university on Saudization ratio are greater than those of King Faisal and King Abdulaziz universities. Also for our data we can make a comparison of means of the universities mentioned above by t – test as follows.

Table (4): Results of significance tests for differences of means

Difference	Estimate	S.E.	Sig.(P value>$\|t\|$)
$\hat{\mu}_1 - \hat{\mu}_2$	12.41	4.65	0.00
$\hat{\mu}_1 - \hat{\mu}_3$	10.07	4.89	0.01
$\hat{\mu}_2 - \hat{\mu}_3$	2.05	5.63	0.29

From Table (4) ,column Sig. which means probability under H_0 that a t- distributed random variable exceeds observed $|t|$,that we reject H_0 given H_0 is true we can see that the differences involving King Faisal university mean($\hat{\mu}_1$) are significant (at significance level 5%) and the difference involving King Saud university mean($\hat{\mu}_2$) and King Abdulaziz university mean($\hat{\mu}_3$) is not significant.

The Saudization ratios in the universities mentioned above could be estimated from equation (13) using the estimated values of $\hat{\mu}$, $\hat{\theta}_1$ and $\hat{\theta}_2$ along with the number of faculty staff . The results appear in Table (4).

Table (5) : Estimated Saudi Ratios for the universities mentioned

University	College	Saudization ratio(%)
King Faisal	COB	48.61
	COCIT	33.09
King Saud	COB	74.26
	COCIT	47.73
King Abdulaziz	COB	68.85
	COCIT	49.17

From table (5)we can see that, the COB have Saudization ratios greater than those of COCIT in the universities mentioned above .For the COB , King Saud university has the highest Saudization ratio(74%) while King Faisal university has the lowest ratio(48.61%). Also, For the COCIT , King Abdelaziz university has the highest Saudization ratio(49.17%) while King Faisal university has the lowest ratio(33.09%). From equations (14) and (15) , the predictive performance of model (1) is tested as follows:

Table (6): shows the predictive performance tests for the three universities .

university	college	U	MSE
King Faisal	COB	0.619	0.589
	COCIT	0.106	0.097
King Saud	COB	0.531	0.511
	COCIT	0.211	0.389
King Abdelaziz	COB	0.326	0.403
	COCIT	0.089	0.130

Table (6) above shows the consistence and the good predictive of the mixed linear model used for Saudization ratios analysis in the universities mentioned above.

References

- Breslow N. and Clayton D.(1993)" Approximate inference in generalized linear mixed models" Journal of the American Statistical Association; 88: 9-25.
- Burch, B.D., Iyer, H.K(1997)" Exact confidence intervals for a variance ratio (or heritability) in a mixed linear model" Biometrics 53, 1318–1333
- Chen, Z. and Dunson, D. B. (2003)" Random effects selection in linear mixed models" Biometrics 59, 762–769
- El-Bassiouni M.Y (1991)" A mixed linear model for loss ratio analysis "ASTIN Bulletin , vol.21 , No.2 ,PP.231-238.
- El-Bassiouni M.Y.and H.A. Charif (2004) "Testing a null variance ratio in mixed models with zero degrees of freedom for error "Computational Statistics & Data Analysis 46 ,PP. 707 – 719.
- Gedalla B. ,Jackson D. and Sandars D.(2006) " A practitioners approach to marine liability pricing using generalized linear modes" Insurance: Mathematics and Economics,vol.38,PP.630-639.
- Hanafy O.M.(2007)"Estimating Retention Rates using Mixed Linear Model in the Egyptian Insurance Market " ASTIN Bulletin , vol.63 , No.14 ,PP.571-595.
- Harville, D. (1977)" Maximum likelihood approaches to variance component estimation and to related problems" Journal of the American Statistical Association 72, 320–340.
- Jiming J. and Sunil R.(2003)" Consistent Procedures for Mixed Linear Model Selection" The Indian Journal of Statistics , Vol. 65, Part 1, pp 23-42.

- Katrien A. and Beirlan J.(2005) "Applications of Generalized Linear Mixed Models in Actuarial Statistics" The Journal of Risk and Insurance, 70(4), 577-599.
- Kinny S. and Dunson D(2007) " Fixed and Random Effects Selection in Linear and Logistic Models" Biometrics 63, 690–698
- Liu D.i, X. Lin, and D. Ghosh (2007) " Semi parametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models" Biometrics 63, 1079--1088.
- McCulloch, C.E., and Searle, S.R. (2000)" Generalized, Linear, and Mixed Models" John Wiley and Sons.
- Neumaier A. and Eildert G.(1998) " Restricted Maximum Likelihood estimation of covariance in spare linear model" Journal of Genetics Selection Evolution,30 ,PP .3-26
- Spilke J., Piepho H., and Hu X.(2004)" Analysis of Unbalanced Data by Mixed Linear Models Using the mixed Procedure of the SAS System" Journal of Agronomy & Crop Science 191, 47—54.
- Stock, J.H., Watson, M.W.(1998) "Median unbiased estimation of coefficient variance in a time-varying parameter model" Journal of the American Statistical Association, 93, 349–358
- Thaddeus T. and Petkova E.(2007) " Principal Points for Linear Mixed Effects Models: Applications to Identifying Placebo Responders" The American Statistician 61:34-40.
- Wolfinger, R.D., and R.E. Kass.(2000)" Non-conjugate Bayesian analysis of variance component models" Biometrics 56:768-774.
- Zhang, D. and Lin, X. (2002)" Hypothesis testing in semi parametric additive mixed models" Biostatistics **4,** 57–74 .