Paper 233

Fitting a Multisource Regression Model with Random Slopes, a Fisheries Application of SAS^{TM} PROC MIXED.

Robert G. Downer. Louisiana State University. Baton Rouge. LA.
Mark C. Benfield. Louisiana State University. Baton Rouge. LA

Abstract

The application of mixed effects linear models continues to grow and the available software is advancing with the methodology. When covariate measurements are made at randomly sampled units. random coefficient models are quite natural for describing the relationship between the response and the predictors. In this very general paper, fitting a multisource regression model in SAS is reviewed. The options available in PROC MIXED are presented, illustrated. and discussed through a coastal fisheries application.

1. Introduction

Mixed linear models are now extensively used in many subject areas. Their application is now quite common in repeated measures models: longitudinal data analysis and statistical genetics. The flexibility they permit will result in more frequent application in other contexts in the future.

The mixed effects linear model is given by

$$Y = \mathbf{X}\beta + \mathbf{Z}\nu + \epsilon$$

where Yis the data vector, β and ν are vectors of fixed and random effects respectively, X and **Z** are the fixed and random effect design matrices and ϵ is the unknown random error vector. The realized ϵ and ν are have assumed to be uncorrelated and have expectation 0 and variance-covariance matrices G and **R** respectively. As a result , the variance-covariance matrix for Y is $\mathbf{V} = \mathbf{ZGZ'} + \mathbf{R}$. Rather than least squares: maximum likelihood or restricted maximum likelihood is used to obtain estimates for G and **R**. Using these estimated covariance matrices: solutions to the mixed model equations are:

$$\hat{\beta} = \left(\mathbf{X}\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\hat{\mathbf{V}}^{-1}\mathbf{Y}$$

$$\hat{\nu} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\beta})$$

No random effects (i.e. elimination of **Z** and ν and $\mathbf{R} = \sigma^2\mathbf{I}$ leaves one with the linear fixed effects model for the mean. In the mixed model. independence among the elements of ϵ is not required and one can model the covariance structure of the data Y by specifying a particular form for **R** or G. Flexibility in the possible covariance structure is a strength of the mixed model (particularly for repeated measurements on subjects). but this feature is not the focus of this paper.

Inclusion of the random effect vector ν implies a sampling distribution for the fitted random effects. These effects may consist of treatment factors or continuous covariates. In analysis of covariance or multisource regression models both types of effects are present. The estimated G matrix of the mixed model will contain the variance components for specified random effects. A random coefficient model containing one random class factor and one covariate allows for the linear relationship between the covariate and the response to be a realization of possible relationships due to the (possible) presence of a random intercept and slope. In the fitting of a random effects model. changing the elements of the random effects vector results in a change in the observed log likelihood. Since one model is contained within another, this allows one to evaluate the importance of the deleted random effect. Modelling of the random effects or covariance structure is generally done while specifying a complete model for the mean (Wolfinger(1993)) and hence all fixed effects are present.

Fitting a multisource regression model in SAS will be reviewed. Using PROC MIXED to fit a random

slopes model in a coastal fisheries application is presented and discussed.

2. Model fitting options in SAS

Prior to introduction of PROC MIXED, a multisource regression (or analysis of covariance) model was generally performed in either PROC GLM or PROC REG. In PROC GLM, treatment factors (included in CLASS statements) can be fixed or random effects while covariates are assumed to be fixed. Inclusion/exclusion of an interaction between a covariate and a random treatment factor tests a homogeneous slopes hypothesis for that particular regressor. In PROC REG, one can equivalently define indicator variables for the treatment factor and all effects are considered fixed. Expected mean squares are provided for effects specified in the RANDOM statement and appropriate testing can be performed by careful examination of the expected mean squares and use of a TEST statement. However. these random effects are still considered fixed in the PROC GLM model fit. Neither GLM or REG allows for random slopes. Allowing a realized coefficient to be an observation from a distribution of potential slopes is now possible through PROC MIXED. To fit a random slopes model one may include a random treatment factor and its interaction in the RANDOM statement or define a random intercept and covariate in which the subject is specified as a random treatment factor (a formulation more commonly seen in repeated measures applications).

Automatic model selection procedures such as BACKWARD. FORWARD, STEPWISE, RSQUARE are available in PROC REG which are not available in PROC GLM. The methodology for analysis of residuals: influential points and predicted values has been thoroughly researched and included in the REG procedure (options such as CLI, INFLUENCE: P, R) and the GLM procedure (e.g. COOK D, H, PRESS: PREDICTED) For either of these procedures, R^2 or an adjusted R^2 is a classical check of the model's adequacy for explaining variation in the response.

In the newer PROC MIXED: the available options for model fitting and model checking are not as extensive and one must refit for each new combination of effects. Modelling for the mean structure is generally done after the random effects modelling has been done with the fixed effects included. Likelihood ratio tests can be performed for hierarchical models (as in the application below) while the Akaike Information Criterion (AIC), (Akaike,1974; Bozdogan 1987) and Schwartz's Bayesian criterion (SBC). (Schwartz,1978; Wolfinger 1993). 0 ne must be careful in model fitting with REML, however: as it is possible for one to add terms and get a worse fit using the restricted log likelihood.

3. Application

3.1 Background

Variability in the distribution of shrimp larvae along the coast of the Gulf of Mexico is of considerable interest to fisheries scientists. The life-history of the brown shrimp ***Farfantepenaeus*** *aztecus* includes a period of juvenile estuarine residence that begins when postlarvae gain access to estuarine systems via tidal passes. The abundance of postlarvae near shore in the northwestern Gulf is considered to be a function of a variety of factors which include the interaction of longshore current and eddies. circulation patterns: and other environmental factors. In another study, (Benfield and Aldrich, 1992). postlarvae were absent at one sampling site but found in abundance in another location separated by only a few kilometers or at the same location several hours later. This finding motivated more interest in the relationship between postlarvae abundance and relevant environmental factors. Four equidistant sampling sites off Galveston Island. Texas were randomly chosen as representative areas along the coast. Four replicate sampling tows were done on three days of four consecutive weeks at each of the four sites

3.2 Preliminary Investigation

Due to the small number of sites and their straight line configuration, a spatial covariance structure was not investigated. Correlation between counts on consecutive days was not anticipated and the relationship between the shrimp count and turbidity. water temperature: air temperature. salinity and oxygen concentration was investigated with week and day effects also included.

Exploratory analysis of the shrimp count by site revealed a large variance to mean ratio which suggested that a Poisson model would be inappropriate. Transformation of the raw counts as well as flow density (count per tow in volume per cubic meter) were then considered for modelling. Probability plots (via PROC UNIVARIATE) revealed that the distribution of the natural logarithm of the count density (shrimp

per cubic meter) was close to normal.

Site was included as a random effect while day and week were included as fixed effects in a multisource regression model in PROC MIXED. Including site as a random effect implies an overall random level (intercept) for each site. Random slopes which allow for a different relationship between the response and a given covariate were included by specifying random effects for interactions of these regressors with site.

3.3 Modelling Example

For illustration. consider the following modelling sequence. Evidence of a relationship between the mean level of the transformed count density and turbidity (e.g. via PROC REG, GLM or MIXED) has been observed and the strength of the relationship is different for each site. Site is now to be investigated as random and a turbidity effect or slope is specified as a random effect through its interaction with site. Both are included in the random statement. Suppose also that we want to control for the effect of week to week variation in the mean of the transformed response. Week is included as a fixed effect in the model statement. With **ldensity** representing the natural log of the count density, the following SAS code performs the fit:

```
proc mixed method=ml;
class  week site ;
model ldensity =    week ;     (1)
random site site*turbid;
run ;
```

The resulting SAS output is:

Model Fitting Information for **LDENSITY**

Description	Value
Observations	189 .0000
Log Likelihood	-347.520
Akaike's Information Criterion	-350.520
Schwarz's Bayesian Criterion	-355.383
-2 Log Likelihood	695.0408

Tests of Fixed Effects

Source	BDF	DDF	Type III F	Pr > F
WEEK	3	178	2.79	0.0418

To check the significance of the random interaction term site*turbid, we are testing whether $\sigma^2_{\hat{\beta}_{turbid}} = 0$. In other words, we're testing whether there is indeed variability in the turbidity slope when these sites are considered as a random sample of possible sites. Since a model without the random interaction term is contained within model (1), testing is possible through a likelihood ratio test. Deleting this random effect and refitting gives the following output:

Model Fitting Information for LDEBSITY

Description	Value
Observations	189.0000
Log Likelihood	-350.743
Akaike's Information Criterion	-352.743
Schwarz's Bayesian Criterion	-355.985
-2 Log Likelihood	701.4864

Tests of Fixed Effects

Source	BDF	DDF	Type III F	Pr > F
WEEK	3	182	3.44	0.0180

The likelihood ratio test performed by considering twice the difference of the observed log-likelihood and the observed log-likelihood for model (1) is significant $(6.45, p < .025$ as approximate $\chi^2_1)$ and hence we keep the random slope term in the model. Testing the random site effect is essentially testing variability of an intercept or the overall level of the mean transformed response by site for this random sample of possible sites. Deleting this term from the original model gives the following SAS output:

Model Fitting Information for LDEBSITY

Description	Value
Observations	189.0000
Log Likelihood	-354.800
Akaike's Information Criterion	-356.800
Schwarz's Bayesian Criterion	-360.042
-2 Log Likelihood	709.6010

Tests of Fixed Effects

Source	**NDF**	DDF	Type III F	Pr > F
UEEK	3	181	3.43	0.0184

The likelihood ratio test with comparison to model (1) is very significant $(p < .002$ approximately $\chi^2_1)$

and hence this random intercept term also remains in the model. Week was included while random effects modelling was performed so that the mean response was modelled fully while modelling the variance. Investigation of the relationship between the class factor week and the mean response is still possible although no fixed effect remains on the right hand side of the model. The resulting output gives a log-likelihood which (in the presence of these random effects) also indicates a significant effect of week on the mean transformed response.

3.4 Final Model

More formal modelling was conducted with all variables included. Random terms for site by day and site by week were part of the original full model but three-way interactions between variables were not specified. Maximum likelihood was used to estimate the effects and variance components in all subequent modelling All fixed and random effects were initially included in a full completely saturated model giving the most complete structure possible for the mean. Sequential dropping of random effect terms (site and interactions with site) was performed first in the presence of the fixed effects. With significant random effects determined, sequential elimination of fixed effects in PROC MIXED was conducted. All submodel testing was done via likelihood ratio tests.

Even with only four sites. significance of the random effect site was not unexpected. The other significant random effects were the interactions of sites with day, site with week and site with turbidity. The random site by turbidity interaction effect (as in the illustrative example) gives evidence that the relationship between the count density and turbidity will vary depending on the selected sites.

Turbidity also played an important role in modelling of the mean with significance observed for the interactions of turbidity with temperature and salinity (main effect of turbidity also present). There was no day effect but all other main effects were significant as well as interactions between water temperature and oxygen level as well as salinity and oxygen level.

Code for the final model was as given below:

```
proc mixed method = ml;
class day week site ;
model ldensity = week turbid temp wtemp salt
                 oxygen turbid*temp turbid*salt
                 wtemp*oxygen salt*oxygen /p ;
random site site*day site*week site*turbid ;
make 'predicted' out = prd noprint;
```

The final value for -2*(log-likelihood) was 591.65. considerably better than the 695.04 of the illustrative model (1) presented above. To further evaluate the model: a pseudo R^2 was computed. The model sum of squares was taken to be $SS_{TOT} - SSE^*$ where SS_{TOT} is the total corrected sum of squares for the transformed response. $SSE^* = \sum_i (y_i - \hat{y})^2$ and y_i is the i't h element of the vector $\mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\nu}$. Hence these predicted values $\hat{y}_i$ use the estimated fixed effects β and the estimated random effects $\hat{\nu}$. The pseudo R^2 was 0.67 and the adjusted value was 0.63. Explaining this much of the variability in a response is considered quite good in fisheries research.

4. Discussion

The inclusion of random slope coefficients in modelling was very important in this coastal fisheries application. Although site could be included in GLM as a random (intercept) effect , allowing the relationships between the response and predictors to vary could only be achieved in SAS via PROC MIXED. A site-specific model has much more limited scope and hence this capability will be advantageous in many applications which have continuous covariates observed at the levels of random factors. In this application modelling with PROC MIXED was quite effective in explaining a response such as shrimp density which is affected by many other variables not considered.

The fitted model via maximum likelihood and likelihood ratio tests appears adequate. However. further model evaluation is desired. The pseudo R^2 developed was computed to have a regression-type standard available for consultation. The traditional error sum of squares is not the sole representation of randomness in the model . The SSE* used above does not give an indication of deviation from the fitted model as the realized coefficient(s) give only one possible realization of an estimated response. The interpretation of residual analysis which includes the random effects is unclear. Further research into appropriate diagnostic tools is needed.

5. References

1. Akaike, H. (1974). A new look at the Statistical Model Identification: ***IEEE Transaction*** on *Automatic* ***Control*** , **AC-19, 716-723.**

2. Benfield, M.C and Aldrich: D.V. (1992). Attraction of postlarval ***Penaeus*** *aztecus* and *P.setiferus* (L) (Crutacea:Decapoda:Penaeidae) to estuarine water in a laminar-flow choice chamber. ***Journal of Experimental Marine Biology and Ecology*** , 156 , 39-52.

3. Bozdogan, H. (1987). Model Selection and Akaike's information criteria (AIC): the general theory and its analytical extensions. ***Psychometrika,*** 52 , 345-370

4. Schwartz. G. (1978). Estimating the dimension of a model. ***Annual of Statistics*** , 6. 461-464.

5. Wolfinger. R. D. (1993). Covariance structure selection in general mixed models. ***Communications in Statistics: Simulation and Computation*** , ***22*** , 1079-1106.

6. Authors

Robert G. Downer
Department of Experimental Statistics
Louisiana State University
Baton Rouge. LA 70803-5606 USA
(225) 388-8373
rdowner@lsu.edu

Mark C. Benfield
Coastal Fisheries Institute: Dept. of Oceanography and Coast al Sciences
Louisiana State University
2 18 Wet land Resources
Baton Rouge LA 70803
(225) 388-6372
mbenfie@lsu.edu