

Random Coefficient Model for Hazard Analysis Critical Control Points (HACCP) A Pilot Study

Xinpeng (Tom) Wang, Yating Sun, Julie Caswell

Key words: Mixed model, random coefficient model, proc mixed, proc glm, food safety, HACCP, hierarchical structure, Bayes Models.

Introduction

Our paper employs random coefficient model using proc mixed to explore the effects on the safety of seafood importation to US market. Different covariance structures are compared. The preferred model is identified. We also offer the interpretation of the results and discuss the potential policy implication.

Our analysis is based on the data collected by National Marine Fisheries Service, the U.S. Department of Commerce, the U.S. Department of Labor, the U.S. Treasury Reporting Rates of Currencies and the International Monetary Funds. However, our data is not complete. The Du Bois library at the University of Massachusetts, Amherst has maintained the quarterly U.S. Treasury Reporting Rates of Currencies since June 30, 1990, but pages of September 30, 1990 are missing. Another missing datum is September 30, 1995 rate of Denmark. It is not printed on the page. We also made use of total export values of goods and services compiled by the International Monetary Funds to calculate the sizes of countries' fishery industries. However, the latest 1999 issue only covers a period of time ended in 1998, which left us two missing data on the sizes for all countries. All of these, of course, also add fun in the modeling effort.

SAS and Mixed Model

A mixed model is one that contains both fixed and random effects. While the regular regression models have a basic assumption that the residual or error terms are independently and identically distributed, mixed model is an important new approach that allow us to relax this assumption and take more complicated data structures into account in a flexible way. For example, if a number of repeated measures are made on a subject, mixed models allow us to specify a pattern for the correlation between these measurements. There are many benefits to

be gained using mixed models. These could be an increase in the precision of the estimates or a greater insight into what underlines the data.

Since SAS Institute introduced - Proc Mixed - into their large menu of offerings in 1992, it has updated and expanded the models and options available as part of the procedure to the point that it is now a reasonable choice for researchers fitting many types of mixed models. Proc Mixed's flexibility has led many unsuspecting users to write a program, obtain results, and have no idea what model has been fit. This paper is to show economists how to fit a particular type of mixed model using SAS and draw useful inferences of the data without getting into the pitfall of pure linear matrix theories.

Although a mixed model approach has advantages over the conventional type of analysis in that it does not require complete data from all subjects in the case of repeated measures over time, we should not regard it as a panacea for all the pains of the world. For example, for assessing the effect of a food safety regulation on the export volume of countries, observations on all involved variables are supposed to be taken every year. The analysis will usually be complicated by a number of missing records for some countries, which can cause considerable difficulty in a conventional analysis. Missing data do not cause nearly so serious a problem with Proc Mixed as with Proc GLM since mixed procedure is based on the basic theory that holds even with unbalanced and missing data, so long as the missing data are random. This results in more appropriate estimates of the effects and their standard errors. The mixed model also gives great flexibility in analysis in that it can allow for a wide variety of ways in which the successive observations are correlated with one another. Beginning from version9, SAS has a new procedure to deal with missing values.

PROC MIXED handles missing level combinations of classification variables similarly to the way PROC GLM does. Both procedures delete fixed-effects parameters corresponding to missing levels in order to preserve estimability. However, PROC MIXED does not delete

missing level combinations for random-effects parameters because linear combinations of the random-effects parameters are always estimable.

Introduction to Random Coefficient Models

Data that have a nested or hierarchical structure are common in a wide variety of disciplines, especially in educational and behavioral studies. The random coefficient approach is used to analyze such data. You can find this approach in different disciplines under different guises, such as empirical Bayes models in the statistical literature (Laird and Ware, 1982), hierarchical linear models (Goldstein, 1987), and multilevel linear models (Bryk and Raudenbush, 1992) in educational and behavioral studies. Unfortunately, this approach has not been applied to economics as broadly as it should be. As a pilot study, we applied this approach in the analysis of the impact of HACCP where we deem it as a perfect fit for the data.

In general, a standard random coefficient model involves a random intercept and slope for each subject. A standard form can be found in Little's famous book on SAS and mixed model.

Letting y_{ij} denote the measurement on the j^{th} observation on the i^{th} subject, this model can be written as follows:

$$y_{ij} = a_i + x_{ij}b_i + e_{ij} \quad (1)$$

where

$$i = 1, 2, \dots, t$$

$$j = 1, 2, \dots, n$$

$$\begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim iid N \left[\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \Psi \right]$$

$$\Psi = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{bmatrix}$$

$$e_{ij} \sim iid N[0, \sigma^2]$$

One of the great advantages of the likelihood-based estimation approach to mixed models is the ability to fit a variety of covariance structure. To accommodate the complexity of the data, we consider structures modeling variability that changes across the data, that is, variability that is heterogeneous.

The SAS code to fit a heterogeneous random coefficients model for the HACCP data is the following:

```
proc mixed data=wangnew covtest mmeq;
  title 'final random coefficient model';
  class IMPORT GEO;
  model KILOS=HACCP DEVELOP GEO
  EXPORTS/s noint ddfm=bw outp=predicted;
  random INT HACCP TIME/type=un
  subject=IMPORT s g gcorr v vcorr;
  repeated / type=un group=GEO r;
  contrast 'overall geographical factor' GEO
  1 -.5 -.5/chisq;
  estimate 'overall development status factor'
  DEVELOP 1/cl;
run;
```

Here, we use data for the time period between 1990 and 2000. Import volumes (KILOS, in million kilograms) to the U.S. are used as dependent variables. Time has the value between 1 and 11 for the 11 years. HACCP is a dummy variable. It equals one between 1998 and 2000 as HACCP exists, and is zero otherwise. GEO and IMPORT are classification variables. While GEO indicating geographical connection between fishery products exporting countries and the United States, import is used to indicate different importing market. GEO has three levels: Asian and Pacific countries (GEO=1), South American countries (GEO=2), and Northern countries (GEO=3). EXPORT (in million dollars) is the export value of total goods and services of each of the 33 countries. DEVELOP is a dummy variable reflecting development status. It equals one for developed countries including Australia, Canada, Denmark, Iceland, Japan, New Zealand, Norway, and United Kingdom, and is zero for the 26 developing countries.

This is a fairly complex model with four fixed effects and three random effects. The variables import and geo are classification variables, and HACCP develop geo and exports define the entire X matrix for the fixed-effects portion of the model, since the intercept is omitted with the NOINT option. Moreover, we are assuming the variation within each subject, in this case, the importing market, has an unstructured structure with group option specifying heterogeneity in the covariance structure. That is, all observations having the same level of the group effect have the same covariance parameters. Each new level of the group effect produces a new set of covariance parameters with the same structure as the original group. In this case, we are assuming that

the observations in the same geographical region will have the same covariance parameters. At the same time, we are saying that the random intercept-slopes model has different variances for intercept, slope of application of different level of HACCP regulation and slope of different time with different covariance between each pair of the three random effects. That is the variation among the subjects, which, in this case are the importing markets. Last point in errors structure, we assume importing markets are independent of each other. Hence, the subject option indicates that the variance covariance matrix for the random effect is to be block diagonal, with identical blocks.

SAS specifies the variation within each subject as R matrix and the variations among subjects as G matrix. The variance of the response variable, which, in this case, is the total weight of importation of seafood from each importing market, can be specified as $V=ZGZ'+R$ in SAS documentation. Interested reader can check out the SAS output in the Appendix for each of the three matrices. We provide the covariance parameters estimation and its significance test in APPENDIX I; hypothesis tests presented in this section indicate that all the variance components are significantly different from 0, although these tests may not be very reliable.

Model Selection

The null model likelihood ratio test suggests that it is necessary to model the covariance structure of the data. The 'Chi-Square' value is -2 times the log likelihood from the fitted model, where the null model is the one with only the fixed effects listed in the MODEL statement and $R=\sigma^2 I$. This statistic has χ^2 distribution. There is more on this in SAS documents.

It is hardly an absolute standard to say what is a good model and what is a bad model. However, that is not to say we can have no preferences over competing models. SAS provides a series of statistics and hypothesis tests to help us choose among competing options. If one covariance model is a submodel of another, you can carry out a likelihood ratio test for the significance of the more general model by computing -2 times the difference between their log likelihoods. Then compare this statistic to the χ^2 distribution with degrees of freedom equal to the difference in the number of parameters for the two models.

SAS also provides two popular model-fit criteria, Akaike's information Criterion (AIC) and Schwarz' Bayesian Creterion (SBC). These are essentially log likelihood values penalized for the number of parameters estimated. SBC imposes a heavier penalty than AIC.

Akaike's Information Criterion (AIC) (Akaike 1974) is computed as

$$AIC = l(\theta) - q \quad (2)$$

Where $l(\theta)$ is the maximized log likelihood or the residual log likelihood, and q is the effective number of covariance parameters (those not estimated to be on a boundary constraint). It can be used to compare models with the same fixed effects but different variance structures; the model having the largest AIC is deemed best.

Schwarz's Bayesian Criterion (BIC) (Schwarz 1978) is computed as

$$BIC = l(\theta) - 1/2 * q * \log N \quad (3)$$

where N is computed as described in the IC option. Again, models with larger BIC are preferred, but note that BIC penalizes models with a greater number of covariance parameters more than AIC does, and the two criteria may not agree as to which covariance model is best.

In the new SAS output of mixed procedure though, the signs of the two criterions have been reversed, so the model with smaller values of the two criterions will be preferred.

To decide if individual variable should be included in the model or not, we should also take a look at the significance test for each individual variable. However, we need to consider other variables' impact as a whole on the significance test too. This has been said, we need to keep it in mind that when the inferences drawn from individual significance test is conflicting with those drawn from the test of the overall model, we prefer the results from the test of the overall model.

In this particular case, some of the individual tests on the random effects are not significant. However, all signs from the overall model tests and AIC, SBC statistics point towards the conclusion that we do need add the extra complexity of the unstructured error structure, once the covariate has been taken into account. APPENDIX II is a table comparing AIC, SBC of the preferred model with those of some other alternative models.

Interpretation and Policy Implication