Measuring more or less: Estimating product period penetrations from incomplete panel data
Hoogendoorn, A.W.

Citation for published version (APA):
8. Summary, conclusions and recommendations

Consumer behaviour is a broad field that covers several scientific areas ranging from economics and marketing to psychology and anthropology. Our view in this field is a very narrow one. We focus on numbers of product purchases and try to describe the purchasing process by estimating average numbers of purchases of a household and the fractions of households that purchase at all. This task may seem an easy one, but it is not. The thesis shows an interesting path along many problems that we encountered while carrying out this task.

In consumer panels it is common practice that households report their consumption on a continuous basis. We collected data from the 'telepanel' where respondents already filled out questionnaires every week. We gave the respondents electronic diaries to keep track of their expenditures. Many respondents did not like this weekly task, and we faced an increase in panel attrition. This caused concern about the quality of the data and about the cost associated with replacing the panel members that had dropped out. In Chapter 2 we studied the effect of reducing the frequency of budget measurements on the precision of estimators. Interestingly, we found that for estimators of change the variance is a flat function of the number of measurements. There is a lot we can gain from that. It implies that almost the same accuracy can be obtained if we reduce the number of measurements for each household by 50 percent. Obviously, this will save efforts and money. In practice, the effect on the precision of estimators of change from reducing the number of measurements may even be more advantageous than we described, since we did not take a possible bias reduction into account. The results we obtained in Chapter 2 apply to a wider field than consumption reports. They apply to any survey that requires frequent measurement and endures possible panel attrition from panel burden.

The largest part of this thesis discusses the use of models for the consumption data. The need for a model arose primarily from the necessity to estimate penetrations. If one would have complete data the penetration could be estimated directly by the sample fraction of the respondents that bought the product. For incomplete data - in the sense that for some individuals we do not have information about the whole period - models are necessary in order to estimate the penetration. The most popular models for consumption data are Poisson models. The main reason for this is that Poisson models are easy to use and that they allow an analysis of product consumption. Even though the Poisson assumption means assuming a process with a constant hazard rate - which seems highly inappropriate for consumption processes - the Poisson models are very useful to obtain insight in repeat buying behaviour. In Chapter 3 we discussed an estimation scheme that uses a set of Poisson models that differ in the way they treat heterogeneity between different individuals. We successfully implemented the estimation scheme for our task to create the consumption reports for the product boards.
The fact that we had to report on the consumption of cumulative periods forced us to resolve another problem. In order to estimate product period penetrations we used the following strategy. Using the data of the period concerned, we first estimated the model parameters. Then we used the estimated model parameters to obtain an estimate for the penetration. Unfortunately this strategy does not necessarily imply that penetration estimates increase if the period grows. As a result the reports we published mentioned products that were consumed in the first quarter by more people than in the first half-year. In order to solve this problem we studied the application of Poisson processes to the situation where we have more than one period (Chapter 4). We developed new theory on multivariate heterogeneous Poisson processes and defined maximum likelihood estimators and generalised moment estimators for both the bivariate Poisson Gamma process and the bivariate Poisson Spike process. Unfortunately we did not succeed in estimating these models for the empirical data. It seems at this point that we are punished for the inappropriate Poisson assumption. Where we had no such problems for a single period, we faced a bad behaviour of the estimators in the bivariate case. The Poisson assumption determines the regularity of the process by fixing the variation coefficient of the interpurchase times to one. If this is not reflected by the empirical data, the bivariate model may interpret the heterogeneity within individuals (in time dimension) as heterogeneity between individuals. This may explain why the method does not work for consumption processes, where it might work in cases where the Poisson assumption seems more appropriate.

The bad experience with Poisson models was a reason for us to look for different models. Studying the literature on renewal processes and event history analyses, we came across a variance component model that was applied in the field of medicine (see Aalen and Husebye, 1991). We liked this model for its simplicity and because the parameters have a clear interpretation. Roughly speaking we can say that the three parameters represent the average intensity of the process, the regularity of the process at the individual level and the heterogeneity between individuals. But, again, the application of the model to empirical data turned out to be one with many obstacles. As soon as we solved one problem, some other problem emerged. Since most of these problems do not exist for the Poisson model, the popularity of this model is easy to understand. The first problem that we had to consider is the fact that the variance component model is formulated in terms of interpurchase times, while the data are in terms of numbers of purchases. Although the two stochastic quantities are related, the mathematical relationship between the two distribution functions is difficult to handle from a computational point of view. Nevertheless, we planned an easy solution. We used an imputation technique to translate the data of numbers of purchases into data of interpurchase times. Although there are several ways to impute interpurchase times from count data, we expected that the estimation results would not depend heavily on the method that was chosen. However, a simulation study showed that the estimation results do depend on the imputation method that is being used. On the other hand, the same simulation study showed that one particular imputation method - using stochastic imputation based on the uniform distribution - behaved rather well in the cases we studied. We also found that from a theoretical point of view this was the best we could get for this model. Theoretically more satisfying approaches like the EM-algorithm and MCEM methods appeared impractical for
our problem because of their computational complexity. Another complicating factor was that we had to take both left and right censoring into account. Again we had to take approximations, but we managed to find a way out that holds a reasonable balance between accuracy and computability. Thus, we ended up with a method for a single period that takes the regularity of the process into account and outperformed Poisson models in our simulation study (see Chapter 5).

Having developed a method to estimate the variance component model for a single period, the next step was to extend the method for two and more periods. In order to expand the theoretical framework we needed to model the behaviour of interpurchase times cross the boundaries between periods. We succeeded in extending the theory, and although we were able to estimate the parameters the computational work became large and the estimation process became slow. Thus we were challenged to build a faster alternative. Combining much of the experience we had gathered in this ongoing research, we came up with an idea to use generalised moment estimation for the variance component model. Using asymptotic properties of the renewal process we were able to derive generalised moment estimators. We were surprised to find that the estimator based on the mean, variance and 'internal correlation' has a closed form. This makes the estimator very attractive from a computational point of view. A simulation study showed that the use of the estimator leads to a better estimation of the penetration than estimators based on Poisson assumptions. The application of this estimator to empirical data was somewhat disappointing. The empirical data may contain a degenerated subclass of individuals that never buy meat products. The results seriously improved when we used the extra information where we asked if people ever bought meat. The final results show that using the variance component model generalised moment estimation is a computationally feasible method that is unique in the fields of models in the sense that it takes the regularity of the individual process into account.

There are many topics that we considered including in this thesis but that we had to leave out. We will mention a few. One such topic is the estimation of penetrations at the individual level. Such estimates could be very useful, since they can be combined to cross-tabulate penetrations with respect to different background variables without re-estimating the model. We made a start with such a study for Poisson models in Sikkel and Hoogendoorn, 1994. We recommend a similar study for the variance component model. A second topic that we had to leave out was to add background variables into the variance component model. Replacing the overall mean by a quantity that depends on household characteristics might improve the fit of the model. A final topic that we think is very promising to apply to this problem is to use Laplace transforms to obtain maximum likelihood estimates. Such an approach would imply the study of modern techniques for numerical inversion of Laplace transforms or, alternatively, the use of a probabilistic interpretation of the Laplace transform to generate new observations that allow immediate maximum likelihood estimation.