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

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
1. Introduction

1.1 Motivation

Although count data of purchases are widely studied there are still many unsolved problems as how to model such data. This thesis is about purchases of meat products, or rather, counting the number of purchases, or even simpler, whether the number of purchases is zero or larger than zero. The process underlying this seemingly simple problem is complex. This is why estimation of the penetration of meat products is also complex. At first we tried to solve the problem by using simple models, but after a while these models turned out to be inadequate. So, starting with simple univariate Poisson models, we ended up with multivariate heterogeneous renewal models. This thesis describes every step in this process. From the naive to the numerical problems that arise when complex probability theory is implemented.

Information on consumer behaviour is of great interest to manufacturers and retailers of the goods that are consumed. Why would they be so interested in detailed information from their customers? Initially manufacturers and retailers think primarily in terms of sales. How many products are sold and what is the amount of money associated with it? Usually they are also very interested in changes. Are sales going up or down? At a later stage they realise that the people that buy the products vary highly. Most customers buy the products occasionally, while there is a small group of heavy buyers that is responsible for a substantive part of the sales. Data that provide this insight separate sales into two components: numbers of buyers and how much they buy. Two main variables that indicate the sales of a product are:

- **Penetration**: the proportion of households that buy an item at least once in a given period (month, quarter, year, etc.).
- **Purchase frequency**: the average number of times these buyers buy the item in a given period.

From these quantities, sales equal the number of buyers times the average number of purchased packs per buyer times the average price or size of the pack. At a later stage one may want to know how the customers vary in needs and habits. This is usually done by making profiles of the clients and by creating market segments in order to meet different customers in their needs and stay in business (see Ehrenberg, 1988). In this thesis we will not go as far as creating market segments. We will focus on simply estimating average sales and penetrations for different products and for different periods.

During the period of a year we gathered data from two thousand Dutch households every week. For this purpose we used the 'telepanel' where data are collected by using modern techniques: computers, modems and programs for Computer Assisted Self Interviewing.
Every week, the members of the households fill out a questionnaire that is sent from the central computer of the institute to the computers at their homes. The respondents have the possibility to fill out the questionnaire at a time that suits them. After completion of the interview the data are sent back to the institute automatically. Thus data are collected from a large number of households without interference of interviewers. The topics of the questionnaires depend on the projects that run at the time. They show a large variety that is reflected by the wide field of studies that were carried out on the basis of 'telepanel data'. Some examples are a budget survey (SARIS et. al., 1992), a time budget survey (KALFS, 1993), trade-union membership (VAN RIJ, 1994), evaluation of survey questions (SCHERPENZEEL, 1995) crime victimisation (DENKERS, 1996), and psychological and economical aspects of saving behaviour (NYHUS, 1996). In some cases the 'telepanel' method itself is an object of study (see e.g. GELDORP, 1993, OPPENHUISEN, 1994 and BLANKERT, 1996). The Stichting TelePanel (STP) at the Universiteit van Amsterdam initiated the 'telepanel' in 1991. Nowadays CentERdata at Tilburg University continues its exploitation and preserves a large data base that contains a part of the information collected so far. Every week the household panel delivers a new data set.

The budget survey was set up in an international co-operation to construct a Completely Automated System for Information Processing (CASIP, see SARIS, et. al., 1992). STP constructed a system for data collection based on an electronic diary of the expenditures of the households. In 1994 STP used that system to gather data on meat, poultry and eggs to report consumption for the product board for livestock and meat and for the product board for poultry and eggs. From these data we produced reports that contained figures on prices and volumes of different products.

1.2 Overview of the thesis

The layout of the thesis follows the chronological order of problems and solutions we found on our way to create the consumption reports. One of the first problems arose from the increase in the response burden when we started the project for the product boards. Before this particular project the questionnaires on fast moving consumer goods were administered once every quarter. When we started administering the questionnaires on consumption every week, there was a severe increase in panel attrition. This affected the quality of the data, both for the budget survey and for the other surveys, and made us think of changing the design in order to reduce the response burden. We studied the effects of a reduction of budget measurements on the precision of the estimators, and will discuss the results in Chapter 2.

A second problem arose from the necessity to report on product period penetrations. Incomplete data prevent the use of the sample fraction as an estimator for penetrations. For figures on volume (the total amount of products consumed or the money associated with it) a balanced survey design or a re-weighting scheme will lead to unbiased estimators that solve the problems that arise from incomplete data. For penetrations there is no simple
solution. If the number of missing observations is low one may think of standard imputation techniques. If the amount of missing information is substantial - for instance because the design is such that observations are taken every two weeks - then the use of a model seems a more appropriate approach. A suitable model may capture enough information from the incomplete data to obtain satisfactory penetration estimates. And, as a bonus, the model parameters may give extra insight in the consumption behaviour related to a certain product. The question is what model we should use. The literature discusses many models that can be applied to consumption data. For reasons that will be explained later in this thesis, we started using Poisson models. In fact, we studied a set of Poisson models that differ in complexity for dealing with heterogeneity. The use of this set of models is the topic of Chapter 3.

A third problem was caused by the fact that we report over cumulative periods. For example each quarterly report contains information about the current quarter, but also about the cumulative period of all quarters so far in the current year. We estimated penetrations for the current quarter by estimating the model parameters from the data of the current quarter, and obtained an estimate for the penetration from the model parameters. There is no problem there. For the cumulative period we aggregated the data over these periods, re-estimated the model parameters on the aggregated data set, and obtained penetration estimates from those values of the model parameters. The problem is that in some cases it occurred that the penetration estimates for the cumulative period were lower than the penetration estimate of one of the underlying periods. The product boards were very dissatisfied with penetration estimates that decreased as a function of the reference period, and demanded an improved procedure. In order to solve this problem we studied the use of multivariate heterogeneous Poisson processes. Chapter 4 shows us that, although we succeeded in describing how such a model can be estimated, we did not succeed in estimating the multivariate models for the empirical data. We attribute the failure of these models to the fact that Poisson models treat the regularity of the interpurchase times as a fixed parameter. This may not be reflected by the data. Our conclusion is that we have to abandon the Poisson assumption for the underlying process.

As a next step we studied models that are not Poisson-based. We came across variance component models for event history data and found them elegant while providing a clear interpretation of the model parameters (see Aalen and Husebye, 1992). The application of variance component models to consumer data is the topic of Chapters 5, 6 and 7. In Chapter 5 we will explain the model and will apply it to consumption data of one period (the univariate case). In doing so we came across several problems to apply the model to our situation. In Chapter 6 we will extend the variance component model to the case of more periods (the multivariate case). Our objective is to create an estimation procedure that can be applied in the setting of the project. Therefore, computational feasibility is a limiting condition to the estimation procedure. In Chapter 7 we will focus on moment-based estimators for the variance component model that need little computing time. Finally, in Chapter 8, we will evaluate the work that is laid down in this thesis and provide recommendations for future research.