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.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
2. Optimising response burden

2.1 Introduction

In budget surveys and consumer panels the burden of the respondents can be a major problem. The work load associated with filling out diaries is considered to be an important reason for the high initial non response rate for budget surveys compared to other household surveys (Lindström, 1989; Lyberg, 1991). A second effect of response burden is the phenomenon of underreporting. It is common to find that the number of reported purchased products in the first week is higher than in the second week (Harrison, 1991; Nevraumont, 1991; Ribe, 1991). In consumer panels the response burden also influences the panel attrition (Silberstein and Jacobs, 1989). Modern techniques like bar scanning methods and electronic diaries (Saris et al., 1992) are applied to relieve the respondent’s task. But in spite of these techniques a respondent still has to do a considerable amount of work, especially in panels that aim at continuous measurement of expenditures. It may be, however, unnecessary to collect the budget data for every week. When we take only a sample of weeks this may result in a relative small loss of precision. On the other hand, it may result in a lower attrition rate as fewer respondents become fed up with their task.

In this chapter we study the relation between response burden and the precision of estimators of change in consumer panels. Response burden is defined here as the number of weeks during a certain period (e.g. three months or a year) that respondents have to fill out a diary. Although it affects initial non-response and underreporting as well, we will focus on the effect on panel attrition. Therefore we assume a simple model which describes the attrition as a function of the response burden. In exploring the effect of panel attrition to the precision of our estimators, we will concentrate on the effect to the variance, not the bias. Most literature on panel attrition deals with bias effects. Some examples are the use of Markov chain models for non-random non response to estimate gross flows in categorical data (Stasny, 1987), econometric regression analyses to correct for attrition bias (Hausman and Wise, 1979), bias reduction by sample designs (Van De Pol, 1989) and bias corrections by weighting techniques (Van De Pol, 1993). For ease of exposure we will assume that households are sampled by simple random sampling, although the theory can easily be generalised to other sampling methods. We will assume that the households that dropped out of the panel are immediately replaced using a quota sampling method (Van De Pol, 1989). Consequently the bias due to attrition is kept to a minimum, and the panel size is constant over time.

We will study two models that correspond with two panel designs. The first design deals with a panel that is dedicated to measure expenditures on consumption goods. This is the case in most consumer panels. We will call this a single purpose panel. In the second

design, however, the panel is also used for other purposes. An example of this is the Dutch Telepanel, where the burdensome questionnaires on expenditures are interchanged with questionnaires on a variety of other topics. We will call this a multi purpose panel. The theory derived here is stated in general terms as to make it applicable to a more general situation.

2.2 Preliminary notation and relations

We are interested in estimates (e.g. of consumption or purchases of fast moving consumer goods) over a certain period of \( M \) weeks. Usually \( M \) is equal to 13, a three month period, but we also may consider \( M = 4, M = 26 \) or \( M = 52 \). In every wave (or week) we have \( n \) households from a much larger population of \( N \) households. It is assumed that there is a constant attrition \( q \). For each week every panel member has a probability \( q \) to drop out of the panel independent of the other panel members and independent of what happens in other weeks. Hence, if we denote the probability to stay in the panel by \( p \) (\( p = 1 - q \)), a respondent who is in the panel at week \( j \) has probability \( p^{k-j} \) to still be a panel member at week \( k > j \). Let

\[
X_{i}^{(j)} \text{ be the amount of purchases by household } i \text{ in week } j \text{ of period } t
\]

\[
X_{j}^{(t)} = Nn^{-1} \sum_{i=1}^{n} X_{i}^{(j)} , \text{ the estimated population total for week } j \text{ of } t
\]

\[
X^{(t)} = \sum_{j=1}^{M} X_{j}^{(t)} \text{ the estimated population total for period } t.
\]

It is assumed that for given \( j \) and \( t \) the \( X_{i}^{(t)} \) are independent and identical distributed with variance \( \sigma^2 \) for the households \( i \). For different values of \( j \) and \( t \) we assume the \( X_{i}^{(t)} \) to be homoscedastic (possessing equal variances). We focus our attention to more or less daily shopping routines, and assume that this is a stable process that is in equilibrium, although the process may be different for each household. When there is a regular weekly pattern in such purchases it is reasonable to assume that for a combination \((j,u) \neq (k,v)\) we have

\[
\text{cov}(X_{j}^{(u)}, X_{k}^{(v)}) = \rho \sigma^2
\]

(2.1)

(a covariance matrix with compound symmetry). Of course, such an assumption is violated for a product that does not follow such a weekly pattern, e.g. when it is purchased on a two-weekly basis. Such types of variables are outside the scope of this chapter. This may suggest that the assumptions are rather restrictive. When, however, broad categories are used like meat, green vegetables, fruit or candies, the assumptions apply, at least in the
Dutch society, to the most important results that have to come out of a budget survey. Our main interest is to measure the changes from one period to another. The absolute consumption level of a product in itself is not a very useful figure. In terms of the variables defined above this means that we are interested in the precision of $X_{t+1} - X_t$ (and as a by-product of $X_{t}^i$). Consequently, we need to know the covariances of the terms of which these quantities consist. Let $n_{jk}$ be the number of panel members which are in both wave $j$ and in wave $k > j$ in a period $t$. Then $n_{jk}$ has a binomial distribution with parameters $n$ and $p^{k-j}$. Assume that the panel members are numbered such that the first $n_{jt}$ are in both waves. Then we can compute

$$\text{cov}(X_{t}^i, X_{t}^{j}) = E_{n_{jk}} \text{cov}(X_{t}^i, X_{t}^{j} | n_{jk}) + \text{cov}_{n_{jk}}(EX_{t}^i, EX_{t}^{j} | n_{jk})$$

$$= E \frac{N^2}{n^2} \sum \text{cov}(X_{j}^i, X_{k}^{j})$$

$$= N^2 p^{k-j} \rho \sigma^2 / n$$  \hspace{1cm} (2.2)$$

This yields the following variance of $X_{t}^i$:

$$\text{var}(X_{t}^i) = \text{var}(\sum_{j=1}^{M} X_{j}^i)$$

$$= \sum \sum \text{cov}(X_{j}^i, X_{k}^{j})$$

$$= \frac{N^2 \sigma^2}{n} \left( M + \frac{2 \rho (M-1) p}{1-p} - \frac{2 \rho p^2 (1-p^{M-1})}{(1-p)^2} \right)$$  \hspace{1cm} (2.3)$$

and for the variance of $X_{t+1} - X_{t}$:

$$\text{var}(X_{t+1} - X_{t}) = \text{var}\left( \sum_{j=1}^{M} X_{j}^{t+1} - \sum_{j=1}^{M} X_{j}^{t} \right)$$

$$= \text{var}(X_{t+1}) + \text{var}(X_{t}) - 2 \sum \sum \text{cov}(X_{j}^{t+1}, X_{k}^{t})$$

$$= \frac{2 N^2 \sigma^2}{n} \left( M + \frac{2 \rho (M-1) p}{1-p} - \frac{2 \rho p^2 (1-p^{M-1})}{(1-p)^2} - \rho p (1-p^{M-1}) \right)$$  \hspace{1cm} (2.4)$$
These (rather complex) formulas were derived under the assumption that the correlations between the purchases in the different weeks are constant. This assumption is unrealistic: one can at least expect some sampling fluctuation, seasonal fluctuation and fluctuation for special occasions like Christmas. Before we continue with the assumption of a constant correlation we report on a theoretical exercise and show some practical examples of our data where this assumption is not exactly met. In that case we have

\[ \text{cov}(X_{t}^{(u)}, X_{t}^{(v)}) = \rho_{jk}^{(u,v)} \]  

(2.5)

As a consequence the variances of our estimators can not be simplified into the formulas (3) and (4), since they will now depend on a whole matrix of correlations.

Theoretically, we studied the effect of changing single correlations from the average. When for only one combination \((j,k,u,v)\) the value \(\rho_{jk}^{(u,v)}\) differs from the other correlations, then the effect to \(\text{var}(X^{(t)})\) and \(\text{var}(X^{(t+1)} - X^{(t)})\) will depend on the position of the correlation in the matrix. If \(u = v = t\), an increase of the correlation will both increase \(\text{var}(X^{(t)})\) and \(\text{var}(X^{(t+1)} - X^{(t)})\). If \(u \neq t\) (i.e. \(u = t\) and \(v = t+1\)) then this will have no effect on \(\text{var}(X^{(t)})\), but \(\text{var}(X^{(t+1)} - X^{(t)})\) will decrease. If a correlation decreases the effects will be in the opposite direction. In a matrix where several correlations are different from the average \(\rho\) most effects will cancel out. Practically, we studied correlation matrices of expenditures on six products in a period of two quarters (26 weeks). One of these correlation matrices (for the product ‘eggs’) can be found in the appendix. We computed variances for our estimators in two ways: using the complete correlation matrix, and using the average correlation. Table 2.1 shows factors between the variances obtained from the two methods. These are the factors that should be applied to the variances we find if we use the average correlation, in order to obtain the actual variances. For the product ‘eggs’ \(\text{var}(X^{(t)})\) is overestimated by 11%, \(\text{var}(X^{(t+1)})\) is underestimated by 10%, and \(\text{var}(X^{(t+1)} - X^{(t)})\) is underestimated by 3%. Our conclusion is that on the whole the errors introduced by assuming equal correlations are small.

<table>
<thead>
<tr>
<th>product</th>
<th>average correlation</th>
<th>over/underestimation factor obtained from the assumption of equal correlations</th>
<th>var(X^{(t)})</th>
<th>var(X^{(t+1)})</th>
<th>var(X^{(t+1)} - X^{(t)})</th>
</tr>
</thead>
<tbody>
<tr>
<td>eggs</td>
<td>0.24</td>
<td>1.11</td>
<td>0.90</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>poultry</td>
<td>0.16</td>
<td>0.96</td>
<td>1.05</td>
<td>1.02</td>
<td></td>
</tr>
<tr>
<td>red meat</td>
<td>0.24</td>
<td>1.03</td>
<td>0.95</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>pork</td>
<td>0.21</td>
<td>0.98</td>
<td>0.99</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>meat products</td>
<td>0.33</td>
<td>0.98</td>
<td>1.01</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>meat &amp; meat pr.</td>
<td>0.45</td>
<td>0.98</td>
<td>1.00</td>
<td>0.89</td>
<td></td>
</tr>
</tbody>
</table>
The relationships (2.3) and (2.4) hold when the respondents are required to fill in the questionnaire during all $M$ weeks of periods $t$ and $t+1$. When the respondents are required to fill in the questionnaire during only $m$ out of $M$ weeks we can distinguish two different models, corresponding with two different panel designs. For the first design we assume that immediately after the measurement a fraction $q = 1 - p$ drops out of the panel. In the weeks when no measurement with respect to the $X^{(i)}$ takes place there is no attrition. This corresponds to a single purpose panel, in which the panel is dedicated for the budget survey. In a multi purpose panel, in contrast, the panel is used for more than only the budget survey, and the attrition continues in the weeks when no measurement with respect to the $X^{(i)}$ takes place. This is the typical case of a Telepanel, which may be used for many purposes. Both models lead to slightly generalised versions of equations (2.3) and (2.4). We generalise our definition of $X^{(i)}$ to

$$X^{(i)} = Mm^{-1}Nn^{-1} \sum_{j=1}^{n} X^{(i)}_{ij}$$

so that

$$X^{(i)} = Mm^{-1}Nn^{-1} \sum_{j=1}^{n} \sum_{i} X^{(i)}_{ij}$$

where $j$ takes the values of only those weeks in which respondent $i$ fills in a questionnaire. Then for the single purpose panel we have

$$\text{var}(X^{(i)}) = \frac{M^2N^2\sigma^2}{nm^2} \left( m + \frac{2\rho(m-1)p}{1-p} - \frac{2\rho p(1-p^{m-i})}{(1-p)^2} \right)$$

(2.7)

and

$$\text{var}(X^{(i+1)} - X^{(i)})$$

$$= \frac{2M^2N^2\sigma^2}{nm^2} \left( m + \frac{2\rho(m-1)p}{1-p} - \frac{2\rho p^2(1-p^{m-i})}{(1-p)^2} \frac{\rho p(1-p^{m})}{(1-p)^2} \right)$$

(2.8)

These expressions are identical to equations (2.3) and (2.4), except that the total period $M$ is replaced by the observed period $m$, and for the factor $M^2/m^2$. For the multi purpose model we have

$$\text{var}(X^{(i)}) = \frac{MN^2\sigma^2}{nm} \left( M + \frac{2\rho(m-1)p}{1-p} - \frac{m-1}{M-1} \frac{2\rho p(1-p^{m-1})}{(1-p)^2} \right)$$

(2.9)

and
\[
\text{var}(X^{(s)} - X^{(i)}) = \frac{2MN^2 \sigma^2}{nm} \left( M \frac{2 \rho (m - 1)p}{1 - p} - \frac{m - 1}{M - 1} \frac{2 \rho p^2 (1 - p^{m-1})}{(1 - p)^2} \right) \\
- M \rho p^m \frac{2 \rho (m - 1)p^{m+1}}{1 - p} - \frac{m - 1}{M - 1} \frac{2 \rho p^{m-2} (1 - p^{m-1})}{(1 - p)^2} \right) 
\]

(2.10)

provided that the sampling design is balanced with respect to first and second order inclusions of the weeks in the sample, i.e. every week \( j \) and every combination \( (j,k) \) of weeks appears in the sample with the same frequency. This is proved in the appendix.

In practice the weekly attrition \( q \) is rather small. Realistic values for \( q \) range from 0.1% to 5%, depending on the subject matter and the time horizon. Such values of \( q \) are sufficiently small to justify a linear and quadratic approximation of formulas (2.7) through (2.10) for reasonable values of \( m \). An illustration is given in Figure 2.1. Throughout this chapter we will use approximations in order to change complex relations into simpler ones. In some cases optimum values for \( m \) are intractable for the variance functions that we study, but not for their linear approximations. Figure 2.1 shows formula (2.8) and two approximations in a single purpose panel with 1% attrition a week in a period of \( M = 52 \) weeks for a product with \( \rho = 0.7 \).

Figure 2.1. Var(X^{(s)} - X^{(i)}) and its first and second order approximations for \( q \) in the single purpose panel for a product for which \( \rho = 0.7 \) and a panel with attrition rate \( q = 0.01 \).
In the region of a quarter \( \text{var}(X^{(n+1)} - X^{(n)}) \) and its approximations are almost equal. Because in the case of a single purpose model \( M \) only serves as a cut-off value, this shows that approximations are good if \( M = 13 \). We now give the formulas for the first order approximation. First and second order approximations are derived in the appendix. The linear approximation formula for \( \text{var}(X^{(n)}) \) in the case of a single purpose panel is

\[
\text{var}(X^{(n)}) = \frac{N^2 M^2 \sigma^2}{nm} \left( 1 + (m-1)\rho - (m^2 - 1)\rho q / 3 \right)
\]

(2.11)

The second term of formula (2.11) represents the well known cluster effect due to the fact that repeated measurements take place on the same respondents (see e.g. Kish, 1965). The third term shows a decrease of this cluster effect because of the attrition: an expected proportion of \( q \) panel members is replaced every week. For \( m > 0 \) formula (2.11) is a decreasing function of \( m \). So, clearly, it is optimal to have as many measurements as possible.

The first order approximation of the variance of \( \text{var}(X^{(n+1)} - X^{(n)}) \) in the case of a single purpose panel, is given by formula (2.12) and derived in the appendix.

\[
\text{var}(X^{(n+1)} - X^{(n)}) = \frac{2N^2 M^2 \sigma^2}{nm} \left( 1 - \rho + (2m^2 + 1)\rho q / 3 \right)
\]

(2.12)

The second term shows us the gain from correlations using panel data. The third term shows us that these gains are reduced by the attrition. By differentiating formula (2.12) with respect to \( m \), we can find where this approximation takes a minimum. We find such a minimum because there is an implicit relation between the number of budget measurements and the attrition in a period of \( M \) weeks. Finding such optimal values of \( m \) will be the topic of the next section in which we will study relationships between response burden and attrition. In the formulation of a multi purpose panel there is no such relation between the quarterly attrition and \( m \). This is reflected in the approximation formulas for the variances that are derived in the appendix

\[
\text{var}(X^{(n)}) = \frac{N^2 M^2 \sigma^2}{nm} \left( 1 + (m-1)\rho (1 - \frac{M+1}{3}) \right)
\]

(2.13)

and

\[
\text{var}(X^{(n+1)} - X^{(n)}) = \frac{2N^2 M^2 \sigma^2}{nm} (1 - \rho + Mm\rho q)
\]

(2.14)

Both equation (2.13) and (2.14) are decreasing functions in \( m \): we measure more accurately when more weeks are observed. This changes, however, when we assume a relationship between \( q \) and \( m \).
2.3 Models for response burden and attrition

In this section we assume a relationship between response burden and attrition. We will study the behaviour of \( \text{var}(X^{(m)} - X^{(y)}) \) as a function of the attrition. The behaviour of \( \text{var}(X^{(y)}) \) is trivial, because this variance is minimal when we have as many as possible independent observations; consequently, it decreases when attrition increases. This makes the behaviour of \( \text{var}(X^{(y)}) \) less interesting, not only from the substantive point of view, but also from the statistical point of view. When we measure differences between intervals, however, it is well known that dependent observations may give higher precision than independent observations (see KISH, 1965).

We define response burden to be \( m \), the number of measurements in a given period of \( M \) possible measurements. For the single purpose panel, such a relationship is already implied by its definition, as each measurement causes a fraction \( q = 1 - p \) to leave the panel. By differentiating formula (2.12) with respect to \( m \), it is easily shown that in this approximation the variance takes a minimum for

\[
m_0 = \frac{3 - p}{2 \rho q} + \frac{1}{2}
\]

(2.15)

So the optimum value of \( m_0 \) decreases with both \( p \) and \( q \). This makes sense. When \( q \) goes to zero, it is desirable to have a large \( m \) because attrition is low and we can take advantage from the fact that we have correlated single source data which are well suited for measuring differences. When \( p \) goes to zero, we have no reason to be careful to keep single source data for measuring differences, so we may just as well measure as often as possible, regardless of the attrition. If, on the other hand, \( p \) goes to one, the optimum value of \( m \) becomes less than one, so \( m = 1 \) is the best possible value. From this one observation we can with certainty predict the differences between \( X^{(y)} \) at other moments.

So far it was assumed that there is no direct relation between \( q \) and \( m \). A reasonable assumption is that \( q \) may have the form of

\[
q = \lambda m^\alpha
\]

(2.16)

where both \( \lambda \) and \( \alpha \) are parameters, that can empirically be determined. Such a model comes from the theory of magnitude estimation, where one tries to relate the magnitude of a number to the magnitude of the sensation produced by a stimulus magnitude (see e.g. HAMBLIN, 1974). In our case the stimulus magnitude is \( m \), the number of budget questionnaires held in a period, and the sensation caused by it, is the response burden perceived by the panel members. If we assume that the probability of leaving the panel is proportional to the perceived response burden, we obtain formula (2.16).
For the single purpose panel the attrition model implies that the expected attrition in $M$ weeks ($q$ small) is of order $\alpha + 1$, i.e. linear when $\alpha = 0$, quadratic when $\alpha = 1$, etc. To explore the consequences of this assumption we substitute $q$ in equation (2.12). This leads to

$$\text{var}(X^{(r+1)} - X^{(r)}) = \frac{2N^2M^2\sigma^2}{nm} \left(1 - \rho - \lambda(2m^a + m^c)\rho / 3\right)$$

(2.17)

This equation has analytic minima for $\alpha = 0$ (when $\lambda = q$), $\alpha = 1$ and $\alpha = 2$. If $\alpha = 1$ the minimum is

$$m_0 = \left(\frac{3}{4} - \rho \right)^{\frac{1}{2}}$$

(2.18)

and if $\alpha = 2$ the minimum $m_0$ satisfies

$$m_0 = \sqrt{\frac{-1}{12} + \frac{1}{144} + \frac{1}{2\lambda \rho}}$$

(2.19)

The interpretation of equations (2.18) and (2.19) is similar to the interpretation of equation (2.15), where $\lambda$ has taken the role of $q$. By taking the respective roots, the values of $m_0$ decrease when $\alpha$ increases. For small values of $\rho$ the optimal value $m_0$ is high; if $\rho \to 1$, $m_0$ tends to zero. Since the effect of response burden increases with $\alpha$ the optimal value $m_0$ decreases with $\alpha$.

In the case of a multi purpose panel, the response burden comes not only from the budget survey, but also from the other questionnaires. This leads to a multivariate version of (2.16) in which the attrition $q$ has the form

$$q = \lambda m^\alpha s^\beta$$

(2.20)

Here $m$ is the number of budget surveys, $s$ is the number of questionnaires on other topics, and $\lambda$, $\alpha$ and $\beta$ are parameters that may be estimated from experiments. Suppose that $s$ is fixed, then the expected attrition in $M$ weeks ($q$ small) is of order $\alpha$, i.e. linear when $\alpha = 1$, quadratic when $\alpha = 2$, etc. Substitution of $q$ into equation (2.14) and differentiation with respect to $m$ leads to the optimum number of observed weeks

$$m_0 = \left(\frac{1 - \rho}{M^\alpha \lambda_1 \rho}\right)^{\frac{1}{\alpha + 1}}$$

(2.21)
where $\lambda_1 = \lambda s^\beta$. Note the similarity with equation (2.18) if $\alpha = 2$. Here too $m_0$ decreases with $\lambda$, because the higher the attrition associated with the budget survey, the smaller the optimum number of weeks to observe.

### 2.4 Practical considerations

Having derived these theoretical relationships, it becomes interesting to discuss their practical implications for the research design of the Telepanel budget survey. The Telepanel is a multi purpose panel. Every week two questionnaires are administered. In 1993 the budget survey was held once in the first quarter and twice in each of the last three quarters. Since respondents were supposed to report on all purchased products, no second questionnaire could be administered (so we act as if a budget measurement takes two questionnaires). In 1994, the budget survey was held every week, but respondents report only their purchases on meat, poultry and eggs. Therefore in that period the budget survey was administered next to a questionnaire on some other topic (so here we take a budget measurement as one questionnaire).

<table>
<thead>
<tr>
<th>period</th>
<th>$m$</th>
<th>$s$</th>
<th>mean fraction of attrition per week</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st quarter 1993</td>
<td>2</td>
<td>24</td>
<td>0.0078</td>
</tr>
<tr>
<td>2nd-4th quarter 1993</td>
<td>4</td>
<td>22</td>
<td>0.0109</td>
</tr>
<tr>
<td>1994</td>
<td>13</td>
<td>13</td>
<td>0.0176</td>
</tr>
</tbody>
</table>

The data we used on attrition in the Telepanel consist of 68 weeks: all weeks of 1993 and the first 16 weeks of 1994. Table 2.2 shows the values of $m$ and $s$ in the different periods in compliance of what we explained above, together with the mean attrition per week. We see that the mean attrition per week increased from 0.78% to 1.76% due to a higher response burden caused by more budget measurements.

Fitting the multivariate attrition model (2.20) to the Telepanel data by use of non-linear regression, gave us the following values of the parameters: $\lambda = 0.00245$, $\alpha = 0.518$ and $\beta = 0.251$. The fact that $\alpha$ has a larger value than $\beta$ suggests that budget questionnaires tend to be more burdensome than questionnaires on other topics. In our data the number of questionnaires on other topics is maximal ($s = 26 - m$), and the attrition in terms of $m$ is shown as the solid line in Figure 2.2 (note that it is actually a step function).
In order to optimise the research design for the budget survey we will fix \( s \), the number of questionnaires on other topics, to the constant number of 13. The attrition as a function of \( m \) for this situation can also be found in Figure 2.2 (dashed line) and is below the line with a maximal number of questionnaires on other topics because of the lower response burden. We will use this function for the attrition in order to find an optimal value of \( m \) for the variance \( \text{var}(X_{(t+1)} - X_{(t)}) \). Having estimated the parameters of the attrition model, we find the number of budget measurements that minimises the variance for a given product by application of equation (2.21). Figure 2.3 shows the number of measurements as a function of \( \rho \). We see that if the correlation of the product is low, the optimum will be to measure on a continuous basis, i.e. take measurements in all 13 weeks of a quarter. If correlations are higher than 0.40, it is optimal to do less measurements. For most of the products we presented in Table 2.1, the optimal design is to take measurements on a continuous base. For the product ‘meat and meat products’, which has a correlation of 0.45, the optimum is to take 11 measurements in a quarter. Although this suggests that a design with continuous measurements is best, we like to point out that there is a large region in which results are obtained that are very close to the optimum. Figure 2.4 shows \( \text{var}(X_{(t+1)} - X_{(t)}) \) both for the product ‘eggs’ \( (\rho = 0.24) \) and for ‘meat and meat products’ \( (\rho = 0.45) \) as a function of \( m \). As mentioned before, for \( \rho = 0.45 \) the minimum is near \( m = 11 \). For \( \rho = 0.24 \) the variance keeps decreasing with \( m \). Yet the picture shows clearly that there is a very small increase of variance if the number of measurements would be fixed at 7. This minimum, however, is rather flat. Consequently, a sub optimal design does not give results that are much worse.
Figure 2.3 The optimal number of budget measurements at the Telepanel as a function of the correlation $\rho$. 

![Graph showing the relationship between the number of budget measurements and correlation $\rho$.]

Figure 2.4 $\text{Var}(x_i^u, x_i^v)$ in the Telepanel for 'eggs' and 'meat and meat products' as a function of $m$. 

![Graph showing the variance of budget measurements for 'eggs' and 'meat and meat products' as a function of the number of budget measurements.]
2.5 Summary and conclusions

In panel research, especially with a telepanel, research design is a difficult problem, because there are many factors to be considered. Even in a perfect world, where all respondents happily fill in their questionnaires without measurement error and without being bored by the many detailed questions about products, expenses, quantities etc., there are complex optimisation problems because products are bought in different patterns, for which different sampling schemes are optimal. In this chapter we restricted ourselves to global product categories for which it is reasonable to assume that they follow more or less the same patterns each week. We did not assume, however, a perfect world, but respondents who may become overloaded with their task. Under some very strict model assumptions we were able to calculate the variance of difference scores between two periods. Under even stricter model assumptions we could give expressions for $m_0$, the optimum number of weeks to be observed in a quarter or a year. In our practical situation it seemed reasonable have the respondent fill in a questionnaire for one of every two weeks.

There is still a world to be discovered in this field. In the first place the assumption that all respondents drop out of the panel with equal probability is not very realistic. It is more likely that there is a group of faithful respondents who, if it were up to them, would stay in the panel for life, and another group who drops out very rapidly. In the second place there are products that are bought with longer intervals than one week. When the purchasing process of such products is fitted to some statistical model, a pattern of correlations between different weeks may come out that would not necessarily lead to untraceable results. In the third place other classes of estimators (like composite estimators) may improve on the results we have given here. In the fourth place models to correct for measurement error (which have not been considered in this chapter) may have their impact on sampling design and estimation. Finally, time series models from which empirical Bayes estimators can be derived, can be used for optimum estimation of consumption.
2.6 Appendix

2.6.1 Design of the multi purpose panel

Let the $n$ individuals be randomly assigned to $L$ groups $G_1, G_2, ..., G_L$. The group of individual $i$ we denote by $G(i)$. The groups correspond to the sets of weeks $\Gamma_1, \Gamma_2, ..., \Gamma_L$. Each set $\Gamma_i$ consists of $m$ out of the $M$ weeks. The design is such that for every week $j$ we have that the number of individuals that have week $j$ in their set $|i: j \in \Gamma_i| = nm/M$, and for every combination $(j, k)$ of weeks we have $|i: j, k \in \Gamma_i| = nm(m-1)/(M(M-1))$. In other words, in each week the budget survey is held for a fraction $m/M$ of the respondents and in each combination of weeks for a fraction $m(m-1)/(M(M-1))$ of the respondents.

Now let $X_j^{(i)}$ be the amount of purchases by household $i$ in week $j$ of period $t$ if $j \in \Gamma_i$. This definition is a slight generalization of the definition in the main text. Now let us calculate the variances and the covariances of the variables $X_j^{(i)}$ defined according to equation (2.6).

$$\text{var}(X_j^{(i)}) = \frac{N^2 M^2}{n^2 m^2} \sum_{i \in \Gamma_i} \text{var}(X_j^{(i)}) = \frac{N^2 M}{nm} \sigma^2$$

and for $k > j$

$$\text{cov}(X_j^{(i)}, X_k^{(i)}) = \frac{N^2 M^2}{n^2 m^2} \sum_{i \in \Gamma_i} \text{cov}(X_j^{(i)}, X_k^{(i)}) = \frac{N^2 M(m-1)}{nm(M-1)} \rho \sigma^2 p^{t-k}$$

The covariances between weeks in two successive periods are similar, except for a factor $p^M$ for the attrition in $M$ weeks. So we find for $k > j$

$$\text{cov}(X_j^{(i)}, X_k^{(i+1)}) = \frac{N^2 M(m-1)}{nm(M-1)} \rho \sigma^2 p^{M-k-1}$$

from which the equations (2.3) and (2.4) can easily derived.
2.6.2 Correlation matrix of weekly purchases on eggs in first half year of 1994

2.6.3 First and second order approximations in the single purpose panel design

We start from formula (2.7). Writing $q = 1 - p$ and $c_1 = N^2 M^2 \sigma^2 / n$ we can write for the variance of $X^{(t)}$

$$\text{var}(X^{(t)}) = \frac{c_1}{m^2} \left( m + \frac{2\rho(m-1)(1-q)}{q} - \frac{2\rho(1-q)^2}{q^2} (1-(1-q)^{m-1}) \right)$$

$$= \frac{c_1}{m^2} \left( m - 2\rho(m-1) - 2\rho(1-(1-q)^{m-1}) \right)$$

$$+ \frac{2\rho(m-1) + 4\rho(1-(1-q)^{m-1})}{q} \frac{2\rho(1-(1-q)^{m-1})}{q^2}$$

$$= \frac{c_1}{m^2} \left( m - 2\rho(m-1) - 2\rho \left( \begin{array}{c} m-1 \\ 1 \end{array} \right) q - \left( \begin{array}{c} m-1 \\ 2 \end{array} \right) q^2 \right)$$

$$+ \frac{1}{q} \left[ 2\rho(m-1) + 4\rho \left( \begin{array}{c} m-1 \\ 1 \end{array} \right) q - \left( \begin{array}{c} m-1 \\ 2 \end{array} \right) q^2 + \left( \begin{array}{c} m-1 \\ 3 \end{array} \right) q^3 \right]$$

$$- \frac{1}{q^2} \left[ 2\rho \left( \begin{array}{c} m-1 \\ 1 \end{array} \right) q - \left( \begin{array}{c} m-1 \\ 2 \end{array} \right) q^2 + \left( \begin{array}{c} m-1 \\ 3 \end{array} \right) q^3 - \left( \begin{array}{c} m-1 \\ 4 \end{array} \right) q^4 \right]$$
In a similar way we can approximate the covariance between $X^{(i)}$ and $X^{(i+1)}$

$$\text{cov}(X^{(i)}, X^{(i+1)}) = \frac{c_i}{m^2} \frac{\rho(1-q)(1-(1-q)^n)^2}{q^2}$$

$$= \frac{c_i}{m^2} \left( -\frac{\rho(1-(1-q)^n)^2}{q} + \frac{\rho(1-(1-q)^n)^2}{q^2} \right)$$

$$= \frac{c_i}{m^2} \left( -\rho \left( \frac{m}{1-q} \right)^2 - \frac{\rho(1-(1-q)^n)^2}{q^2} \right)$$

$$= \frac{c_i}{m^2} \left( m\rho - m^2 \rho q + \frac{1}{2} m(m-1)(7m+1)\rho q^2 \right)$$

This enables us to compute the second order approximation of the variance of $X^{(i+1)} - X^{(i)}$:

$$\text{var}(X^{(i+1)} - X^{(i)}) = \frac{2c_i}{m} \left( 1 - \rho + \frac{1}{2} (2m^2 + 1)\rho q - \frac{1}{2} m(m-1)(6m^2 + 2m + 2)\rho q^2 \right)$$

### 2.6.4 First and second order approximations in the multi purpose panel design

We start with formula (2.9). Writing $c_2 = N^2 M \sigma^2 / n$ we can write for the variance of $X^{(i)}$:

$$\text{var}(X^{(i)}) = \frac{c_i}{m} \left( M + \frac{2 \rho(m-1)(1-q)}{q} - \frac{2 \rho(m-1)(1-q)^2}{(M-1)q^2} \right)$$
\[ \frac{c_2}{m} \left( M - 2 \rho (m - 1) \left[ \frac{1}{q} - 1 \right] - \frac{1}{(M - 1)q^2} \left[ \binom{M - 1}{1} q - \binom{M - 1}{2} q^2 + \binom{M - 1}{3} q^3 - \binom{M - 1}{4} q^4 \right] + \frac{2}{(M - 1)q} \left[ \binom{M - 1}{1} q - \binom{M - 1}{2} q^2 + \binom{M - 1}{3} q^3 \right] - \frac{1}{(M - 1)} \left[ \binom{M - 1}{1} q - \binom{M - 1}{2} q^2 \right] \right) \]

\[ = \frac{Mc_2}{m} (1 + (m - 1) \rho - \frac{1}{3} (m - 1) (M + 1) \rho q + \frac{1}{15} (m - 1) (M - 2) (m + 1) \rho q^2) \]

In a similar way we can approximate the covariance between \( X^{(t)} \) and \( X^{(t+1)} \)

\[ \text{cov}(X^{(t)}, X^{(t+1)}) = \frac{c_2}{m} \left( M \rho (1 - q) \right)^M \]

\[ + \frac{m - 1}{M - 1} \rho (1 - q)^M \left[ \frac{2(M - 1)(1 - q)}{q} - \frac{2(1 - q)^2 (1 - (1 - q)^{M-1})}{q^2} \right] \]

\[ = \frac{Mc_2}{m} \left( m \rho - \left\{ M + \frac{1}{5} (m - 1)(4M + 1) \right\} \rho q \right) \]

\[ + \left\{ \frac{1}{2} M(M - 1) + \frac{1}{15} (m - 1)(11M^2 - 3M - 2) \right\} \rho q^2 \]

so for the variance of \( X^{(t+1)} - X^{(t)} \) we find:

\[ \text{var}(X^{(t+1)} - X^{(t)}) = \frac{2Mc_2}{m} \left( 1 - \rho + Mm \rho q - \frac{1}{6} M(5Mm - 2M - m - 2) \rho q^2 \right) \]

We believe that the process of purchasing consumer goods at the individual level, in a renewal process, is governed by a renewal process, because there are two main events (purchases) of interest to an individual: the first purchase and the renewal of the order. If we assume that the distribution of the time between two events (purchases) of a given individual is independent, we can estimate the distribution of the time between two events (purchases) of a given individual. The process can be modeled as a renewal process, starting at time 0, the first time to the first event by \( X_1 \), the time between the first and second event by \( X_2 \), and so on.
\[
\begin{align*}
\text{Expression 1:} & & \text{Expression 2:} \\
\text{Expression 3:} & & \text{Equation 4:} \\
\end{align*}
\]