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.
7. Moment estimators for the variance component models

7.1 Introduction

We study the use of a variance component model to budget count data. The variance component model is defined in terms of interpurchase times. We have count data that consist of observed numbers of purchases. In chapters 5 and 6 we transformed the count data into data of interpurchase times by imputation. This made it possible to estimate the model parameters, although the computation time to estimate the parameters was rather high.

In this chapter we will look for an alternative estimation method that is simpler from a computational point of view and works directly on the numbers of purchases. In the variance component model (VC model) it is assumed that the interpurchase times have a lognormal distribution. This assumption identifies the distribution of \( M(i) \), the number of purchases in a time interval \([0,t]\). Unfortunately, this distribution is awkward to work with. We have to deal with convolutions that are inconvenient from a numerical point of view. As an alternative we will use asymptotic properties of the distribution of \( M(i) \) that are known from renewal theory. We will derive generalised moment estimators based on these properties. As in chapters 5 and 6, simulation studies have to prove the quality of the method.

7.2 Moment estimators

In the variance component model that we introduced in chapter 5 we assumed that the interpurchase times \( X_{i1}, X_{i2}, \ldots \) of the purchasing process of individual \( i \) satisfy the model

\[
\log(X_{ij}) = \mu + U_i + E_{ij}
\]

(7.1)

where \( \mu \) is the overall mean and the individual score \( U_i \) and the residual \( E_{ij} \) are random variables, that are stochastically independent and normally distributed with zero mean and variances \( \sigma_u^2 \) and \( \sigma_e^2 \) respectively (see section 5.2).

The number of parameters in the model is three. In order to formulate (generalised) moment estimators for the model we therefore need three empirical quantities from the data (see e.g. HOEL, PORT and STONE, 1971). The question is what three quantities will we use? The first two quantities that come to mind are the (sample) mean and variance of \( M(t) \). For the third quantity we will examine two alternatives. The first alternative is the penetration. The use
of this quantity corresponds to what we did in the case of the Poisson Gamma Spike model (see chapter 3). The second alternative is to use a quantity that we call internal correlation. We will explain in section 7.2.2 what we mean by this quantity.

### 7.2.1 Moment estimators based on mean, variance and penetration.

From renewal theory (see Cox, 1962) some general asymptotic properties of the random variable $M(t)$ are known for renewal processes. Under the assumption that the renewal process is in equilibrium we derived in chapter 6 that for the VC model $M(t)$ has asymptotic mean

$$E[M(t)] = t\exp(-\mu - \frac{1}{2}\sigma_e^2 + \frac{1}{2}\sigma_u^2), \text{ as } t \to \infty$$  \hspace{1cm} (7.2)

and asymptotic variance

$$\text{Var}(M(t)) = t^2 \exp(-2\mu - \sigma_e^2 + \sigma_u^2)(\exp(\sigma_e^2 - 1))$$

$$+ t \exp(-\mu - \frac{1}{2}\sigma_e^2 + \frac{1}{2}\sigma_u^2)(\exp(\sigma_e^2 - 1)) + O(1), \text{ as } t \to \infty$$  \hspace{1cm} (7.3)

(see equations (6.10) and (6.11)). We will ignore the term $O(1)$ in equation (7.3), and will use equations (7.2) and (7.3) as approximations. As a third quantity we will use the product period penetration. For the penetration in a period $t$ we have the following expression

$$P(M(t) > 0) = P(X_1 \leq t) = \int_0^t f_1(s)ds$$  \hspace{1cm} (7.4)

where $X_1$ is the forward recurrence time, and $f_1$ is its density function given by equation (5.18). In appendix 7.8.1 it is shown that we can write this as

$$P(M(t) > 0) = \exp(-\mu - \frac{1}{2}\sigma_e^2 + \frac{1}{2}\sigma_u^2) t - \int_0^t \int_0^y \frac{1}{\sqrt{2\pi(\sigma_u^2 + \sigma_e^2)}} \times$$

$$\times \exp\left(-\frac{(\log(y) - \mu)^2 + 2\mu\sigma_e^2 + \sigma_u^2 + 2\sigma_u^2 \log(y))}{\sigma_e^2 + \sigma_u^2}\right) dy ds$$  \hspace{1cm} (7.5)

We find (generalised) moment estimators by linking equations (7.2), (7.3) and (7.4) to the empirical data. Therefore we replace the left-hand sides of these equations by their empirical counterparts: the sample mean $m$, the sample variance $s^2$, and the penetration $p$. 

114
in the sample. The consumer panel data consist of numbers of purchases of different households in consecutive weeks. If we observe \( n \) households over a time period of \( t \) weeks, we can condense the \( n \times t \) data matrix into an \( n \)-vector \( \mathbf{M}(t) \) that consists of total numbers of household purchases in a period of \( t \) weeks. From this vector we can easily compute the sample mean \( m_t \), variance \( s_t^2 \) and penetration \( p_t \). Next substitute \( m_t \), \( s_t^2 \) and \( p_t \) into the equations (7.2), (7.3) and (7.4). From the first two equations we find expressions for \( \mu \) and \( \sigma_u^2 \) in terms of \( m_t \), \( s_t^2 \) and \( \sigma_u^2 \) (see appendix 7.8.1). We find

\[
\sigma_e^2 = \ln \left( 1 + \frac{s_t^2}{m_t^2} - \exp(\sigma_u^2) \right) \frac{1}{m_t} \tag{7.6}
\]

and

\[
\mu = \ln(t/m_t) \left( -\frac{1}{2} \sigma_e^2 + \frac{1}{2} \sigma_u^2 \right) \tag{7.7}
\]

For the penetration we get

\[
p_t = \exp(-\mu - \frac{1}{2} \sigma_e^2 + \frac{1}{2} \sigma_u^2) \int \frac{1}{\sqrt{2\pi(\sigma_e^2 + \sigma_u^2)}} \times \\
\times \exp \left( -\frac{1}{2} \left( \log(y) - \mu \right)^2 + 2 \mu \sigma_e^2 + \sigma_e^4 + 2 \sigma_u^2 \log(y) \right) \, dy \, ds \tag{7.8}
\]

Substitution of equations (7.6) and (7.7) into equation (7.8) gives us an equation with only the known sample quantities and the unknown parameter \( \sigma_e^2 \). We find an estimate \( \hat{\sigma}_e^2 \) by solving the equation numerically. Estimates for \( \sigma_u^2 \) and \( \mu \) are obtained by substitution of \( \hat{\sigma}_e^2 \) into equations (7.6) and (7.7) respectively. Although this is a clear prescription to obtain estimates, there is no guarantee that a solution will exist. We will have to study the behaviour of these estimators.

So far we assumed that all \( n \) households reported over a period of \( t \) weeks. The method can be generalised to the case where we observe the households for different time length \( t_i, i = 1, \ldots, n \). In that case we study the data vector that consists of elements \( t_i^{-1} \mathbf{M}(t_i) \). The technical details of this procedure are explained in appendix 7.8.1. Unfortunately, solving equation (7.8) numerically is rather time consuming. In this respect the computation time of this estimation procedure can be compared in this respect with the methods of chapters 5 and 6. In the next section we will look for an alternative method that hopefully will be faster.
7.2.2 Moment estimators based on mean, variance and internal correlation.

As an alternative approach we will use information that remains hidden when only observations \( M(t) \) are used. Intuitively, the idea is the following. Based on the frequency distribution of \( M(t) \) we can barely distinguish which part of the variance has to be ascribed to the differences between individuals and which part has to be ascribed to the variance of the interevent times. Let us split \( M(t) \) into the variables \( M(0,t_i) \) and \( M(t_i,t) \), the number of events in time interval \([0,t_i)\) and \([t_i,t)\) respectively. Then we have the equality

\[
M(t) = M(0,t_i) + M(t_i,t) .
\]

When there is no individual heterogeneity the relation between \( M(0,t_i) \) and \( M(t_i,t) \) will be weak, because there is not much dependence between what happens in \([0,t_i)\) and what happens in \([t_i,t)\). When there is individual heterogeneity, this implies a relation between \( M(0,t_i) \) and \( M(t_i,t) \), because at the individual level they share the same latent variable. We will refer to the correlation between \( M(0,t_i) \) and \( M(t_i,t) \) as the internal correlation. It can be observed in the data by extracting two data vectors \( M(0,:,t) \) and \( M(t,:,t) \) from the data matrix and computing the sample correlation. We will now relate the internal correlation to the model parameters. This equation will give us an alternative quantity that we can use in constructing moment estimators.

The internal correlation can be expressed in terms of attenuation. In chapter 4 we defined the notion of attenuation as the phenomenon that, as a result of measurement error, correlations between observed variables are lower than the correlations between the underlying latent variables. In the situation of a renewal process attenuation is a consequence of the fact that we observe the process through a finite time window. The attenuation is a function of \( t \), and is defined as

\[
c(t) = \rho(M(t),t) .
\]

where \( f \) is the latent variable that denotes the expected number of events in a time unit \([0,1)\) (see equation (4.12)). Since \( M(0,t_i) \) and \( M(t_i,t) \) share the same latent variable, the correlation between their latent scores is unity, and it follows that

\[
\rho(M(0,t_i),M(t_i,t)) = c(0,t_i)c(t_i,t) .
\]

(see equation (6.17)). If we approximate the variance by use of equation (7.3) then we find the following expression of the correlation in terms of the parameters
\[
\rho(M(0,t_1),M(t_1,t)) = \frac{\sqrt{t_1(t-t_1)} \exp(-\mu - \frac{1}{2} \sigma^2_e + \frac{1}{2} \sigma^2_u)(\exp(\sigma^2_e) - 1)}{\sqrt{t_1(t-t_1)} \exp(-\mu - \frac{1}{2} \sigma^2_e + \frac{1}{2} \sigma^2_u)(\exp(\sigma^2_u) - 1) + (\exp(\sigma^2_e) - 1)} \tag{7.9}
\]

Together with equations (7.2) and (7.3) we find a closed form solution for the model parameters. We find for the case that \( t_1 = \frac{1}{2} t \) as a moment estimator for \( \sigma^2_e \)

\[
\hat{\sigma}^2_e = \ln\left(1 + \frac{1-r_1 s^2_t}{1+r_1 m_t}\right)
\]

and from that for

\[
\hat{\sigma}^2_u = \ln\left(1 + \frac{s^2_t}{m^2_t} - \frac{\exp(\hat{\sigma}^2_e) - 1}{m_t}\right)
\]

and

\[
\hat{\mu} = \ln(t/m_t)\left(-\frac{1}{2}\hat{\sigma}^2_e + \frac{1}{2}\hat{\sigma}^2_u\right)
\]

where \( m_t, s^2_t \) and \( r_1 \) are the observed mean, variance and internal correlation in the empirical data. The derivation is shown in appendix 7.8.2. The closed form solution makes that the estimation procedure is a very fast one. Notice that if we observe that the correlation between the two half samples is equal to one, then we would find zero as an estimate for \( \sigma^2_e \) and for \( \sigma^2_u \) the logarithm of 1 plus the coefficient of variation. Notice further that if we find a large (enough) estimate for \( \sigma^2_u \), then the corresponding estimate for \( \sigma^2_e \) may become negative or even non-existing. We will study the behaviour of the estimation method in the next section.

### 7.3 Application to simulated data

We do not have a clear idea of how the two alternatives for generalised moment estimation explained in the previous section will behave in practice. In order to find out we will use the simulated data of chapter 5 that mimic the purchase data of the products pork, beef and veal. Recollect, that we distinguished two types of data. The complete data consisted of numbers of purchases of 1465 households over 13 weeks. The incomplete data contained missing observations according to the empirical response matrix of the telepanel (see Table 5.1).
For both the method that uses penetration and the method that uses internal correlation we do not have the guarantee that they will give us estimates in the domain of the parameters. Therefore the first topic of evaluation is feasibility. Table 7.1 shows us percentages of cases for which we did find estimates for the two alternatives on the 100 complete and the incomplete data sets of all three products. We see that the use of penetration does not give us estimates in all cases, i.e. there was no solution to equation (7.8). No estimates were found for most of the complete and incomplete data sets of the product veal and for some of the incomplete data sets of the product beef. On the other hand the use of internal correlation gave us a solution in all cases.

Table 7.1 Feasibility (percentage of samples where an estimator was obtained)

<table>
<thead>
<tr>
<th>data</th>
<th>method</th>
<th>pork</th>
<th>beef</th>
<th>veal</th>
</tr>
</thead>
<tbody>
<tr>
<td>complete data</td>
<td>penetration</td>
<td>100%</td>
<td>100%</td>
<td>17%</td>
</tr>
<tr>
<td></td>
<td>internal correlation</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>incomplete data</td>
<td>penetration</td>
<td>100%</td>
<td>95%</td>
<td>5%</td>
</tr>
<tr>
<td></td>
<td>internal correlation</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Why does the method that uses penetration fail in some cases? Since the data are generated according to the variance component model, this is not a consequence of a bad fit of the model. The reason must be sought in the fact that we use asymptotic approximations. We observe in Table 7.1 that the use of the penetration is rather sensitive to the fact that we use approximations. The use of internal correlations turns out to be more stable. How can we explain that the method behaves worse for the simulated data sets of veal purchases? The purchasing process of veal is a slow process compared to the other two products. The average frequency of purchasing differs dramatically. From equation (6.22) we compute that the average frequency for the product veal is 0.0183. The corresponding frequencies for pork and beef are 0.6065 and 0.4493 respectively. As a consequence the purchase process of the product veal is still far from its asymptotic behaviour. The same argument explains why the feasibility is worse for incomplete data than for complete data, because the time horizon for incomplete data is shorter.

Table 7.2 Average estimated parameter values and (13-week) penetrations plus standard errors for the variance component model over 100 data sets of 1465 households, using complete count data

<table>
<thead>
<tr>
<th>variable</th>
<th>true values</th>
<th>s.e.</th>
<th>$\sigma^2$</th>
<th>s.e.</th>
<th>$\sigma^2$</th>
<th>s.e.</th>
<th>penetration</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>pork</td>
<td>penetration</td>
<td>0.600</td>
<td>0.900</td>
<td>0.700</td>
<td>0.934</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>internal correlation</td>
<td>0.631</td>
<td>0.039</td>
<td>0.884</td>
<td>0.082</td>
<td>0.621</td>
<td>0.025</td>
<td>0.943</td>
</tr>
<tr>
<td></td>
<td>penetration</td>
<td>0.569</td>
<td>0.087</td>
<td>0.871</td>
<td>0.092</td>
<td>0.731</td>
<td>0.112</td>
<td>0.934</td>
</tr>
<tr>
<td>beef</td>
<td>true values</td>
<td>1.190</td>
<td>1.200</td>
<td>0.600</td>
<td>0.878</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>internal correlation</td>
<td>1.117</td>
<td>0.063</td>
<td>1.179</td>
<td>0.153</td>
<td>0.549</td>
<td>0.027</td>
<td>0.886</td>
</tr>
<tr>
<td></td>
<td>penetration</td>
<td>1.072</td>
<td>0.137</td>
<td>1.167</td>
<td>0.166</td>
<td>0.627</td>
<td>0.157</td>
<td>0.878</td>
</tr>
<tr>
<td>veal</td>
<td>true values</td>
<td>6.000</td>
<td>5.000</td>
<td>1.000</td>
<td>0.097</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>internal correlation</td>
<td>5.574</td>
<td>0.267</td>
<td>3.763</td>
<td>0.763</td>
<td>0.658</td>
<td>0.119</td>
<td>0.124</td>
</tr>
<tr>
<td></td>
<td>penetration</td>
<td>5.206</td>
<td>0.437</td>
<td>4.856</td>
<td>0.568</td>
<td>1.909</td>
<td>0.546</td>
<td>0.101</td>
</tr>
</tbody>
</table>

*For the product veal the results for moment estimates using penetration are only based on those cases where a solution was obtained (see Table 7.1)
A second topic is the accuracy of the re-estimation of the model parameters. The results for complete data can be found in Table 7.2. With respect to the estimation of $\sigma^2$, we see that moment estimation using correlation underestimates, whereas the method using penetration overestimates this parameter. If one is interested in the penetration, it is obviously better to choose the method that uses the penetration. This can easily be explained from the fact that the method fits the penetration to find the parameters, implying that the model estimate of the penetration is the same as the direct estimate. It is more interesting to see how these methods behave in the case of incomplete data, since we do not have the alternative of using direct estimators. Table 7.3 shows the results on incomplete count data.

Table 7.3 Average estimated parameter values and (13-week) penetrations plus standard errors for the variance component model over 100 data sets of 1465 households, using incomplete count data

<table>
<thead>
<tr>
<th></th>
<th>pork true values</th>
<th>s.e.</th>
<th>$\sigma^2$ s.e.</th>
<th>$\sigma^2$ s.e.</th>
<th>penetration s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mu$</td>
<td>0.600</td>
<td>0.700</td>
<td></td>
<td>0.934</td>
</tr>
<tr>
<td></td>
<td>correlation</td>
<td>0.643</td>
<td>0.090</td>
<td>0.599</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>penetration</td>
<td>0.514</td>
<td>0.139</td>
<td>0.781</td>
<td>0.149</td>
</tr>
<tr>
<td>beef</td>
<td>true values</td>
<td>1.100</td>
<td>1.200</td>
<td>0.600</td>
<td>0.878</td>
</tr>
<tr>
<td></td>
<td>correlation</td>
<td>1.110</td>
<td>0.168</td>
<td>0.560</td>
<td>0.040</td>
</tr>
<tr>
<td></td>
<td>penetration*</td>
<td>1.036</td>
<td>0.216</td>
<td>0.664</td>
<td>0.195</td>
</tr>
<tr>
<td>veal</td>
<td>true values</td>
<td>6.000</td>
<td>3.751</td>
<td>1.000</td>
<td>0.097</td>
</tr>
<tr>
<td></td>
<td>correlation</td>
<td>5.996</td>
<td>0.778</td>
<td>0.614</td>
<td>0.106</td>
</tr>
<tr>
<td></td>
<td>penetration*</td>
<td>5.531</td>
<td>0.619</td>
<td>1.609</td>
<td>0.623</td>
</tr>
</tbody>
</table>

*For the products beef and veal the results for moment estimates using penetration are only based on those cases where a solution was obtained (see Table 7.1)

For the incomplete data we find results that are similar to those for the complete data. Again we find an underestimation of $\sigma^2$ if we use the internal correlation and an overestimation if we use penetration. To estimate the penetration we find that the results are best if we use penetration, but we have to take into account that for the products beef and veal this estimate is not always available.

7.4 A comparison with Poisson based models

In this section we will compare the use of the VC model with the use of Poisson Gamma model. We studied the behaviour of Poisson based models in chapters 3 and 4, and will use the ML estimator for that model. For the VC model we will restrict ourselves to the moment estimator based on the internal correlation for reasons of feasibility and computation time. We will make the comparison on how well the penetration is estimated and will use the simulated data of section 7.3, where Table 7.3 showed us average estimated parameter values, and 13-week penetrations in the case we have incomplete data. Table 7.4 shows the comparison with the Poisson based models.
Table 7.4 A comparison between the generalised moment estimators that uses internal correlation of a variance component model and Poisson based models

<table>
<thead>
<tr>
<th></th>
<th>pork penetration</th>
<th>s.e.</th>
<th>beef penetration</th>
<th>s.e.</th>
<th>veal penetration</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>true value</td>
<td>0.934</td>
<td></td>
<td>0.878</td>
<td></td>
<td>0.097</td>
<td></td>
</tr>
<tr>
<td>Internal Correlation</td>
<td>0.945</td>
<td>0.008</td>
<td>0.885</td>
<td>0.018</td>
<td>0.125</td>
<td>0.014</td>
</tr>
<tr>
<td>Poisson</td>
<td>1.000</td>
<td>0.000</td>
<td>0.997</td>
<td>0.001</td>
<td>0.212</td>
<td>0.056</td>
</tr>
<tr>
<td>Poisson Gamma</td>
<td>0.902</td>
<td>0.007</td>
<td>0.828</td>
<td>0.010</td>
<td>0.091</td>
<td>0.009</td>
</tr>
<tr>
<td>Poisson Spike</td>
<td>0.886</td>
<td>0.009</td>
<td>0.820</td>
<td>0.012</td>
<td>0.088</td>
<td>0.009</td>
</tr>
</tbody>
</table>

We see that for the products pork and beef the use of the variance component model is a considerable improvement over the Poisson based models. For the product veal, however, this is not the case. It seems that in some cases the variance component model is better, while in other cases one should prefer the Poisson model. There are circumstances when we may expect that the Poisson model performs better. One such case is, when the time period $t$ is relatively short (or $\mu$ relatively large), because we may expect that the asymptotic approaches may be bad. Another circumstance may be if the coefficient of variation that denotes the regularity of the process is close to 1. Also in such a case using a wrong model with reasonable estimators may be better than using the right model with biased estimates.

Thus it seems that the relative merits of the models depend on the parameters. To obtain some idea which model is the better in which case, we generated a large number of samples using different parameter values and different time length $t$. For each data set we estimated the parameters both using the generalised moment estimators of the variance component model and the maximum likelihood estimator of the Poisson Gamma model. From the parameters we estimated the penetration. Since we know the true penetration of the generated data, we can decide which of the two methods leads to a better estimation of the penetration. Maybe we can deduce from this large number of generated data sets a rule to determine which of the two methods one should use. In total we created 1000 data sets. For each data set we draw a random value for the parameter $\mu$ (uniformly) between 0 and 5, for $\sigma_a^2$ between 0 and 3 and for $\sigma_x^2$ between 0 and 1. For the time period $t$ either a period of 2, 4, 6, 8, 10, 13, 16, 20, 26, 39 or 52 weeks was chosen (random with equal probability). Table 7.5 shows some descriptive statistics of the experiment. Row 1 to 4 show the statistics of the parameters and time period used to generate the samples. Row 5 shows the theoretical penetration that is computed by use of equation (7.5). Row 6 shows the penetration in the complete sample, i.e. the sample where all numbers of purchases are known. Next the response matrix of chapter 5 is applied to the complete data in order to obtain the incomplete data, and the two alternative models (Variance Component and Poisson Gamma) are estimated. Row 7 through 9 show statistics of the parameter estimates of the VC model. Rows 10 and 11 show the penetration estimated by use of the VC model and the PG model respectively. Finally row 11 shows statistics of the indicator that the penetration estimate of the VC model is closer to the (true) theoretical penetration. We see that the VC model is best in 71% of the cases.
Table 7.5 Descriptive statistics of 1000 data sets that are generated according to a VC model with different parameter values and different timelengths and penetrations estimated by use of moment estimators of the VC model and ML estimators of the PG model.

<table>
<thead>
<tr>
<th>data</th>
<th>minimum</th>
<th>maximum</th>
<th>mean</th>
<th>std.deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. generated ( \mu )</td>
<td>.000</td>
<td>4.996</td>
<td>2.45517</td>
<td>1.45605</td>
</tr>
<tr>
<td>2. generated ( \sigma )</td>
<td>.000</td>
<td>1.000</td>
<td>.51634</td>
<td>.29698</td>
</tr>
<tr>
<td>3. generated ( \sigma )</td>
<td>.003</td>
<td>2.999</td>
<td>1.50152</td>
<td>.85027</td>
</tr>
<tr>
<td>4. generated ( t )</td>
<td>2</td>
<td>52</td>
<td>17.26</td>
<td>14.37</td>
</tr>
<tr>
<td>5. theoretical penetration</td>
<td>.015</td>
<td>1.000</td>
<td>.57569</td>
<td>.31170</td>
</tr>
<tr>
<td>6. penetration in complete sample</td>
<td>.014</td>
<td>1.000</td>
<td>.57480</td>
<td>.31091</td>
</tr>
<tr>
<td>7. estimated ( \mu )</td>
<td>-.260</td>
<td>5.187</td>
<td>1.96615</td>
<td>1.48575</td>
</tr>
<tr>
<td>8. estimated ( \sigma )</td>
<td>.000</td>
<td>.889</td>
<td>.48412</td>
<td>.22397</td>
</tr>
<tr>
<td>9. estimated ( \sigma )</td>
<td>.000</td>
<td>4.980</td>
<td>1.33321</td>
<td>.89086</td>
</tr>
<tr>
<td>10. estimated penetration (VC)</td>
<td>.000</td>
<td>1.000</td>
<td>.55651</td>
<td>.33812</td>
</tr>
<tr>
<td>11. estimated penetration (PG)</td>
<td>.000</td>
<td>1.000</td>
<td>.50099</td>
<td>.32466</td>
</tr>
<tr>
<td>12. indicator</td>
<td>0</td>
<td>1</td>
<td>.71</td>
<td>.45</td>
</tr>
</tbody>
</table>

Now we will try to predict when the VC model is to be preferred over the PG model. We will use logistic regression (see Hosmer and Lemeshow, 1989, for an introduction of logistic regression) and perform the analyses in two stages. First we will make the prediction based on the parameters that were used to generate the data (rows 1 through 4 of Table 7.5). This analysis may confirm or reject the conjecture that we expect the variance component model to perform better in case of a small value of the parameter \( \mu \) and large \( t \)'s. Note that in practice this information is not available. A researcher has to decide on the outcome of the estimation procedure (rows 7 through 9 of Table 7.5) which of the two estimators he will take. Therefore in the second stage we will make the prediction based on the estimated parameters and the time \( t \).

Table 7.6 shows the parameter estimates of the logistic regression for the case where we use the true parameter values of the variance component model and the case where we use estimated parameter values. We will concentrate on the case of true parameter values first. The table shows us that we find a negative value for the ‘variable’ \( \mu \) of \(-0.35\). This value has the following interpretation. Suppose that we study two processes that follow the variance component model (7.1) over a period of \( t \) weeks. Let us assume that the two processes have the same values for the regularity parameter \( \alpha \) and for the heterogeneity parameter \( \sigma ^2 \), but that the first process has a value for the intensity parameter \( \mu \) that is 1 higher than the other process. A higher value of \( \mu \) indicates that the first process has a lower intensity. The interpretation of the negative sign of the value \(-0.35\) is that a higher of \( \mu \) corresponds to a lower probability that the VC model will be better model to estimate the penetration. This is exactly what we expected. The value of \(-0.35\) tells us that the probability differs by a factor \( 0.5 \exp(-0.35) / (1 + \exp(-0.35)) = 0.827 \). From the standard error we see that the effect of \( \mu \) is statistically significant with \( \alpha = 0.05 \).
Likewise we see that in the logistic regression the parameter estimates for the ‘variables’ $\sigma_e^2$ and $\sigma_u^2$ are negative. This may be explained by the fact that these parameters have a similar effect as $\mu$ on the intensity of the process (see equation (7.2)). We included a term for the interaction between $\sigma_e^2$ and $\sigma_u^2$, but its effect turned out to be insignificant. We find a parameter estimate for $t$ that is positive indicating that the VC model is more likely to be preferred as we observe the process over a longer period. The second part of Table 7.6 shows us the results of the logistic regression if we use estimated parameter values. These results are similar as in the case where we used the true parameter values. We do find a change in sign for the parameter $\sigma_v^2$, and a significant interaction between $\sigma_e^2$ and $\sigma_u^2$. From the results we can derive a rule that allows us to decide from the parameter estimates of the variance component model which method we should use. The method gives in 73.5% of the sample a good prediction.

### 7.5 A method for more than one period

The model for more than one period is easily obtained by adding an index to model (7.1) that represents the period. The model becomes (see section 6.1)

$$\log(X_{ijk}) = \mu_k + U_{ik} + E_{ijk}$$

(7.10)

where $k = 1, \ldots, p$, the number of periods, $j = 1, \ldots, n_{ik}$, the number of purchases of individual $i$ in period $k$, and $i = 1, \ldots, N$, the sample size of individuals. Independence is assumed between all $E_{ijk}$ and $U_{ik}$, and within all $E_{ijk}$. For the variables $U_{ik}$ and $U_{ij}$, however, we allow that they are correlated with correlation coefficient

$$\rho_{ki} = \rho(U_{ik}, U_{ij})$$

(7.11)

In chapter 6 we derived a generalised moment estimator for $\rho_{ki}$ that we can apply in the case where the margins are fixed (see equation 6.18). Thus we can generalise the method of
moments to the case where we have more than one period. Start estimating the parameters of the marginal distributions using the moment estimators explained in section 7.2, and then estimate the correlation parameter using

\[
\hat{\rho}_{ij} = \frac{1}{\hat{\sigma}_{ik} \hat{\sigma}_{jl}} \ln \left( 1 + \frac{\hat{\rho}_{2\text{-way}} \sqrt{\left( \exp\left(\hat{\sigma}_{ik}^2\right) - 1 \right) \left( \exp\left(\hat{\sigma}_{jl}^2\right) - 1 \right)}}{\hat{c}_i(t_k) \hat{c}_j(t_l)} \right)
\]  

(7.12)

See section 6.2 for the technical details. In section 6.2.2 we explained three alternatives to find estimates for the attenuations. We will discuss the performance of this approach in section 7.6, where we will discuss the use of the moment estimators of the VC model to empirical data.

### 7.6 Empirical data

In this section we will examine the behaviour of the moment estimators of the VC model to empirical data. Therefore we will use the purchase data that we used in section 6.5 to evaluate the estimation techniques based on imputation of interpurchase times. These data sets contain information about purchases of 524 households on pork, beef and veal. The households reported numbers of purchases during 10 weeks of the third quarter and 10 weeks of the fourth quarter of 1994. We will study the behaviour of both the moment estimator that uses penetration (see section 7.2.1) and the moment estimator that uses the internal correlation (see section 7.2.2). The estimation strategy for the two alternatives is identical. We start with estimating the marginal parameters using the outlines of section 7.2. Next we estimate the dependency parameter \( \rho_{12} \) by use of equation (7.12).

It turns out that the moment estimator that uses penetration was not feasible for any of the six data sets (three products and two periods). The procedure that numerically solves equation (7.12) did not converge in any of these cases. This proves once more that the sample penetration gives us too little information in order to discriminate between the variance components. On the other hand the moment estimator that uses internal correlation was feasible for all products. The results of the estimation procedure are shown in Table 7.7.

<table>
<thead>
<tr>
<th></th>
<th>( \mu_1 )</th>
<th>( \sigma_{\mu_1}^2 )</th>
<th>( \sigma_\mu^2 )</th>
<th>( \mu_2 )</th>
<th>( \sigma_{\mu_2}^2 )</th>
<th>( \sigma_\sigma^2 )</th>
<th>( \rho_{\text{emp}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>pork</td>
<td>0.586</td>
<td>0.562</td>
<td>0.678</td>
<td>0.453</td>
<td>0.566</td>
<td>0.584</td>
<td>0.793</td>
</tr>
<tr>
<td>beef</td>
<td>1.060</td>
<td>0.759</td>
<td>0.585</td>
<td>0.939</td>
<td>0.754</td>
<td>0.548</td>
<td>0.836</td>
</tr>
<tr>
<td>veal</td>
<td>5.834</td>
<td>3.226</td>
<td>0.465</td>
<td>5.711</td>
<td>2.477</td>
<td>0.527</td>
<td>0.650</td>
</tr>
</tbody>
</table>

Given the parameter estimates of the marginal distributions and the correlation in the observed two-way table (shown in the last column of Table 7.7), we obtain estimates for the
dependency parameter $\rho_{12}$ by use of equation (7.12). As discussed earlier in section 6.2 we have three alternatives to estimate the attenuation in the two periods. The first alternative is to use the linear approximation of the asymptotic approximation. This corresponds to the use of equation (6.15). The second alternative is an improvement of the approximation by adding the intercept (see equation (6.6)). The third alternative is to use simulations (sample size 50,000) to approximate the variance. The results for the three alternatives are shown in Table 7.8. We find that the attenuation is closer to one (i.e. ‘less severe’) if we add the intercept to the asymptotic approximation than if we restrict the approximation to the linear part. If we improve the approximation of the variance by use of approximations based on a large number of simulations, then the attenuation is even closer to one. In the cases where we found estimates for the attenuation larger than one, we restricted its value to 1.

Table 7.8 Estimates for the attenuations and for the dependency parameter $\rho_{12}$ the products pork, beef and veal.

<table>
<thead>
<tr>
<th></th>
<th>asymptotic approximation</th>
<th>asymptotic approximation</th>
<th>approximation based on simulations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>linear</td>
<td>linear plus intercept</td>
<td>simulations</td>
</tr>
<tr>
<td></td>
<td>$c_i(t_1)$</td>
<td>$c_i(t_2)$</td>
<td>$c_i(t_1)$ $c_i(t_2)$ $\rho_{12}$</td>
</tr>
<tr>
<td>pork</td>
<td>0.896</td>
<td>0.926</td>
<td>0.966</td>
</tr>
<tr>
<td>beef</td>
<td>0.918</td>
<td>0.933</td>
<td>0.983</td>
</tr>
<tr>
<td>veal</td>
<td>0.909</td>
<td>0.761</td>
<td>0.986</td>
</tr>
<tr>
<td></td>
<td>1.000</td>
<td>1.000</td>
<td>0.866</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.815</td>
</tr>
</tbody>
</table>

Now we are able to make a comparison between the generalised method of moments that works directly on the count data (the method we just described) with the maximum likelihood method that uses imputation of interpurchase times (the method we described in chapter 6). Comparing Table 7.7 with Table 6.2 show that the use of moment estimators gives us lower estimates for the regularity parameter $\sigma^2$. Subsequently the parameter estimates for $\mu$ are higher. Comparing Table 7.8 with Table 6.3 we see that using generalised moment estimator we find estimates for the dependency parameter $\rho_{12}$ that are within their domain. This was not the case when we used the imputation technique, and is a result of the fact that the attenuations are not as low as in the case where we imputed.

Finally, having obtained estimates of the parameters, we now want to estimate the penetration. In order to do so we have to make an assumption on the behaviour of interevent times that lie in more than one period. In chapter 6 we assumed that the time between an event in one period and an event in another period is completely determined by the parameters of the period in which it started (forward inheritance). Similarly we introduced the notion of backward inheritance where interevent times are determined by the parameters of the period in which an interevent time ends. In both cases there is no change in hazard rate if we go from one period to another. Here we will assume that the hazard rate in a period is determined by the parameters of that period. There is dependency between the hazard functions from the parameter $\rho_{ii}$. Suppose that we study the renewal process over two periods of length $t_1$ and $t_2$, then the probability of no event in either of the two periods can be written as
\[ P(\text{no event in period 1 and no event in period 2}) \]
\[ = P(\text{no event in period 1}) \times P(\text{no event in period 2} \mid \text{no event in period 1}) \]
\[ = P(X_1 > t_1) \times P(X_2 > t_2 \mid X_1 > t_1) \]

In the case of an ordinary renewal process both \(X_1\) and \(X_2\) are interevent times. After logarithmic transformation the vector of interevent times is bivariate normal distributed. The probability \(P(X_1 > t_1)\) can be computed by integrating the bivariate normal density over a half-plane. In the case of a renewal process that is in equilibrium the random variable \(X_1\) is the forward recurrence time. In that case we can use lognormal approximations in order to compute these probabilities from the bivariate normal density (see appendix 7.8.3).

Table 7.9 shows (10-week) penetration estimates for the product pork obtained in two ways.

<table>
<thead>
<tr>
<th></th>
<th>pork</th>
<th>beef</th>
<th>veal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(t_1)</td>
<td>(t_2)</td>
<td>(t)</td>
</tr>
<tr>
<td>direct</td>
<td>0.809</td>
<td>0.832</td>
<td>0.875</td>
</tr>
<tr>
<td>moment</td>
<td>0.931</td>
<td>0.957</td>
<td>0.988</td>
</tr>
<tr>
<td>imputed</td>
<td>0.878</td>
<td>0.905</td>
<td>0.981</td>
</tr>
</tbody>
</table>

The first line of Table 7.9 shows direct estimates based on the complete data (sample fractions of households that bought a product). The second line shows estimates that are obtained using the variance component model and generalised moment estimates. The third line shows the estimates that are obtained if the imputation technique is used in chapter 6. The results for the different estimation methods are very large. It appears that the fit of the variance component model is rather bad. It may be that there is a group of households that never buys the product. Such a group distorts the model assumptions, and this may cause that no good estimates can be obtained.

We tried to get a better fit for the model by using extra information. At the beginning of the consumer survey we asked the persons in households whether they ever buy meat or meat products. On the basis of this information we can split the households into two groups. In the first group they (say they) will never buy meat and in a second group they (say they) will ever buy meat. The 524 households that we used in Table 7.9 was thus split into a group of 432 ever buyers (82.3%) and 92 never buyers (17.7%). In order to estimate the model parameters we used the data of the ever buyers only (the results are in the appendix). An estimate of the penetration is then a combination (in fact: the product) of the probability of being an ever buyer and the probability of buying the product in a period of time given being an ever buyer. The results in Table 7.10 show that the estimates are much better than those of Table 7.9. The results depend for a large part on the quality of the variable that
indicates whether a household is an ever buyer of a never buyer. In Table 7.10 we find direct a estimate for the penetration of pork in the second quarter ($t_2$) and the half year ($t$) that is larger than the fraction of ever buyers. This is a consequence of the fact that there are some households that bought pork, although they claimed that they would never buy a product.

<table>
<thead>
<tr>
<th></th>
<th>pork</th>
<th></th>
<th>beef</th>
<th></th>
<th>veal</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>i_1</td>
<td>0.809</td>
<td>i_2</td>
<td>0.832</td>
<td>t</td>
<td>0.875</td>
<td></td>
</tr>
<tr>
<td>i_2</td>
<td>0.702</td>
<td>i_2</td>
<td>0.769</td>
<td>t</td>
<td>0.836</td>
<td></td>
</tr>
<tr>
<td>i_3</td>
<td>0.734</td>
<td>i_2</td>
<td>0.753</td>
<td>t</td>
<td>0.794</td>
<td></td>
</tr>
<tr>
<td>i_4</td>
<td>0.679</td>
<td>i_2</td>
<td>0.694</td>
<td>t</td>
<td>0.785</td>
<td></td>
</tr>
</tbody>
</table>

### 7.7 Discussion and conclusions

In this chapter we developed generalised moment estimators for the variance component model that work directly on the count data. The method uses asymptotic properties that are known from renewal theory. We studied two versions: one that uses the penetration and one that uses the internal correlation as a sample quantity next to the mean and variance. We found that the moment estimator that uses the penetration was computational intensive and sometimes not feasible. The use of the internal correlation, however, leads to a closed form solution that is easy and fast from a computational point of view, and turned out to be feasible in most cases. Our interest is on the estimation of penetration. We showed that there are circumstances where the proposed model (using the generalised moment estimator) has much better results than Poisson Gamma models. We also saw that there are some cases where Poisson Gamma models are to be preferred. We created a decision rule for practical circumstances, where one can decide on the estimation results which model is to be preferred. We also showed that the estimation method can be generalised to the case where we have more than one period. Finally we used the developed method to estimate the penetration for empirical data. We found that the method overestimates the penetration and the observed penetration in the sample. The poor behaviour to the empirical data can be explained from the fact that in reality there will be a group of individuals that never will buy a certain meat product, whereas the model assumes that everyone buys eventually. We saw that the method can be improved by taking a group of never buyers into account. Of course, in determining the group of never buyers many new problems will arise. We keep them outside the scope of this chapter, and are of the opinion that more research on this matter is needed.
7.8 Appendix.

7.8.1 Moment estimates from the sample mean, variance and penetration

In section 7.2.1 we derived an expression for the penetration in a period $t$ (see equation (7.5)) that we will prove here. We write $f_{N(\mu, \sigma^2)}$ for the probability density function of the normal distribution with expectation $\mu$ and variance $\sigma^2$. We write $f_{LN(\mu, \sigma^2)}$ for the probability distribution function of the lognormal distribution. We have the following equations

$$
P(M(t) > 0) = P(X_1 \leq t) = \int_0^t f_1(x)dx$$

$$= \int \int f_1(x | U = u) f_{N(0, \sigma^2)}(u) dudx$$

$$= \int \int 1 - F_{LN(\mu + u, \sigma^2)}(x) \mu + u + \frac{1}{2} \sigma^2 f_{N(0, \sigma^2)}(u) dudx$$

$$= \int \int \frac{1}{\mu + u + \frac{1}{2} \sigma^2} f_{N(0, \sigma^2)}(u) dudx - \int \int \frac{f_{N(0, \sigma^2)}(u)}{\mu + u + \frac{1}{2} \sigma^2} f_{LN(\mu + u, \sigma^2)}(y) dydudx$$

$$= \exp(-\mu - \frac{1}{2} \sigma^2 + \frac{1}{2} \sigma^2) t$$

$$- \int \int \frac{1}{\sqrt{2\pi(\sigma^2 + \sigma^2)}} \exp\left(-\frac{1}{2} \frac{(\log(y) - \mu)^2 + 2\mu \sigma^2 + \sigma^4 + 2\sigma^2 \log(y))}{\sigma^2 + \sigma^2}\right) dydudx$$

Thus we are able to integrate out over the heterogeneity variable $u$. Suppose that we observe $n$ households over a time period of length $t$. From the data vector $\mathbf{M}(t) = (M_1(t), M_2(t), \ldots, M_n(t))$ we obtain the sample mean $m$, sample variance $s^2$ and the penetration $p$ in the sample. We substitute these sample estimates on the left-hand sides of equations (7.2), (7.3) and (7.4) respectively. We find from equation (7.2)

$$m = t \exp(-\mu - \frac{1}{2} \sigma^2 + \frac{1}{2} \sigma^2)$$

(7.13)

and by substitution into (7.3) we find
$s_i^2 = m_i^2 (\exp(\sigma_u^2) - 1) + m_i (\exp(\sigma_e^2) - 1)$

Equation (7.14) allows us to express $\sigma_u^2$ in terms of $\sigma_e^2$ and the observed mean $m_i$ and variance $s_i^2$:

$$\sigma_u^2 = \ln \left( 1 + \frac{s_i^2}{m_i^2} - \frac{\exp(\sigma_e^2) - 1}{m_i} \right)$$

(7.15)

Substitution of this expression into equation (7.13) gives us the following expression for $\mu$:

$$\mu = \ln(t/m_i) \left( -\frac{1}{2} \sigma_e^2 + \frac{1}{2} \left( 1 + \frac{s_i^2}{m_i^2} - \frac{\exp(\sigma_e^2) - 1}{m_i} \right) \right)$$

(7.16)

Substituting these expressions for $\sigma_u^2$ and $\mu$ into equation (7.4) gives us the penetration as a function of the parameter $\sigma_e^2$. Solving this equation numerically gives us the estimate for $\sigma_e^2$, and substitution into (7.15) and (7.16) gives us estimates for $\sigma_u^2$ and $\mu$ respectively.

If we observe $n$ households over time periods of different length, we can create a vector

$$t^{-1} \mathbf{M}(t) = \begin{pmatrix} t_1^{-1} M_1(t_1) & t_2^{-1} M_2(t_2) & \cdots & t_n^{-1} M_n(t_n) \end{pmatrix}^T$$

of averaged numbers of purchases. From this data vector we obtain a sample mean $m$, variance $s^2$, and penetration $p$, and relate those to expressions like in equation (7.2), (7.3) and (7.4). For the expectation we derive from (7.2)

$$E[t^{-1} \mathbf{M}(t)] = \exp(-\mu - \frac{1}{2} \sigma_e^2 + \frac{1}{2} \sigma_u^2)$$

We will regard $t$ as a stochastic variable, and will assume that $t$ is independent of the rest of the process. For the asymptotic variance we have to condition on the varying time lengths $t$

$$Var(t^{-1} \mathbf{M}(t)) = Var_t(E[t^{-1} \mathbf{M}(t) \mid t]) + E_t[Var(t^{-1} \mathbf{M}(t) \mid t)]$$

$$= 0 + E_t[\exp(-2\mu - \sigma_e^2 + \sigma_u^2)(\exp(\sigma_u^2) - 1) + t^{-1}(\exp(-\mu - \sigma_e^2 + \sigma_u^2)(\exp(\sigma_u^2) - 1))]$$

$$= \exp(-2\mu - \sigma_e^2 + \sigma_u^2)(\exp(\sigma_u^2) - 1) + E_t[t^{-1}](\exp(-\mu - \sigma_e^2 + \sigma_u^2)(\exp(\sigma_u^2) - 1))$$
In order to find the moment estimators we estimate $E[r^{-1}]$ by its sample version $\bar{r}^{-1} = n^{-1} \sum i^{-1}$. We find

$$m_i = \exp(-\mu - \frac{1}{2} \sigma^2 + \frac{1}{2} \sigma^2 v),$$  \hspace{1cm} (7.17)

$$s_i^2 = m_i^2 (\exp(\sigma^2 v) - 1) + \bar{r}^{-1} m_i (\exp(\sigma^2 v) - 1)$$  \hspace{1cm} (7.18)

and

$$p_i = \frac{1}{n} \sum_{i=1}^{n} P(M(t_i) > 0)$$  \hspace{1cm} (7.19)

where

$$P(M(t_i) > 0) = \exp(-\mu - \frac{1}{2} \sigma^2 + \frac{1}{2} \sigma^2 v) \times \left( \int_0^y \frac{1}{\sqrt{2\pi(\sigma^2 v + \sigma^2)}} \times \exp \left( -\frac{1}{2} \frac{(\log(y) - \mu)^2 + 2\mu \sigma^2 v + \sigma^2 v + 2 \sigma^2 \log(y)}{\sigma^2 v + \sigma^2} \right) \right) dy ds$$

From equations (7.17) and (7.18) we can derive expressions for $\sigma^2$ and $\mu$ in terms of $\sigma^2 v$, and substitute those into equation (7.19). The sum of $n$ terms in equation (7.19) can be reduced to a sum of $T$ terms, where $T$ is the number of different time lengths that are observed. For example, in the case the period is a quarter, the households reported their purchases either 1, 2, 3, 4, ... or 13 weeks. Solving this equation for the parameter $\sigma^2$ gives an estimate, and from that estimates of the other parameters can be obtained from equations (7.17) and (7.18).

### 7.8.2 Moment estimates from the sample mean, variance and internal correlation

Suppose that we observe $n$ households over a time period of length $t$. From the data vector $\mathbf{M}(t) = (M_1(t) \quad M_2(t) \quad \ldots \quad M_n(t))'$ we obtain the sample mean $m_i$, variance $s_i^2$ and an internal correlation $r_i$ by splitting the interval into $(0, \frac{1}{2} t)$ and $(\frac{1}{2} t, t)$. We substitute these sample estimates on the left-hand sides of equations (7.2), (7.3) and (7.9) respectively. As
in section 7.8.1 we find expressions for $\sigma^2_v$ and $\mu$ in terms of $\sigma^2_e$ that we can substitute into equation (7.9). This gives us

$$r_i' = \frac{\frac{1}{2} t \exp(-\frac{\mu}{2} - \frac{1}{2} \sigma^2_v + \frac{1}{2} \sigma^2_e) (\exp(\sigma^2_e) - 1)}{\frac{1}{2} t \exp(-\frac{\mu}{2} - \frac{1}{2} \sigma^2_v + \frac{1}{2} \sigma^2_e) (\exp(\sigma^2_e) - 1) + \exp(\sigma^2_e) - 1}$$

$$= \frac{m_i (\exp(\sigma^2_e) - 1)}{m_i (\exp(\sigma^2_e) - 1) + 2(\exp(\sigma^2_e) - 1)}$$

$$\frac{s^2_i}{m_i} - (\exp(\sigma^2_e) - 1)$$

$$= \frac{s^2_i}{m_i} + (\exp(\sigma^2_e) - 1)$$

by substitution of $\mu$ and $\sigma^2_v$ respectively. From this relationship we can solve $\sigma^2_v$, giving us

$$\sigma^2_v = \ln \left( 1 + \frac{1 - r_i \cdot s^2_i}{1 + r_i \cdot m_i} \right)$$

If we observe $n$ households over time periods of different length, we create a vector of averaged numbers of purchases. For the expectation and variance we have the expressions as in the previous section. For the covariance we find

$$\text{cov} \left( \frac{1}{t_a} M(t_{n1}), \frac{1}{t_{a2}} M(t_{n2}) \right)$$

$$= \text{cov} \left( E \left[ \frac{M(t_{n1})}{t_{n1}} \right], E \left[ \frac{M(t_{n2})}{t_{n2}} \right] \bigg| t_{n1}, t_{n2} \right) + E \left[ \text{cov} \left( \frac{M(t_{n1})}{t_{n1}}, \frac{M(t_{n2})}{t_{n2}} \right) \bigg| t_{n1}, t_{n2} \right]$$

$$= \sigma^2_j$$

the variance of the frequency of the heterogeneous renewal process, given by equation (6.13). We find
\[
\begin{align*}
\frac{s_i^2}{m_i} - (\exp(\sigma_e^2) - 1)t^{-1} \\
\frac{s_i^2}{m_i} + (\exp(\sigma_e^2) - 1)t^{-1}
\end{align*}
\]

and from that

\[
\sigma_e^2 = \ln \left( 1 + \frac{1 - r_t}{1 + r_t} \frac{s_i^2}{m_i} t^{-1} \right)
\]

### 7.8.3 Estimating the penetration from given parameter estimates.

Suppose that a renewal process satisfies the variance component model (7.10). In the first period \([0, t_1]\) the process is determined by the parameters \(\mu_1, \sigma_{\omega_1}^2\) and \(\sigma_e^2\). In the second period \([t_1, t_1 + t_2]\) the process is determined by parameters \(\mu_2, \sigma_{\omega_2}^2\) and \(\sigma_{\omega_2}^2\). We have to make an assumption about interevent times that cross the border between the two periods. We assume that there is a change in hazard rate. Before timepoint \(t_1\) the hazard rate is defined by the first set of parameters, and after \(t_1\) by the second set. Consequently the penetration is

\[
P(M_1(t_1) + M_2(t_2) > 0) = 1 - P(M_1(t_1) = 0 \text{ and } M_2(t_2) = 0)
\]

\[
= 1 - P(M_1(t_1) = 0)P(M_2(t_2) = 0 | M_1(t_1) = 0)
\]

\[
= 1 - P(X_1 > t_1)P(X_2 > t_2 | X_1 > t_1)
\]

We focus on the conditional probability. In the case of an ordinary renewal process, we have that

\[
P(X_2 = x_2 | X_1 = x_1) = P(\log(X_2) = \log(x_2) | \log(X_1) = \log(x_1))
\]

\[
= P(\mu_2 + U_2 + E_2 = \log(t_2) | \mu_1 + U_1 + E_1 = \log(t_1))
\]

where \(U_1 \sim N(0, \sigma_{\omega_1}^2), E_1 \sim N(0, \sigma_e^2), U_2 \sim N(0, \sigma_{\omega_2}^2), E_2 \sim N(0, \sigma_e^2)\), and where all these variables are independent, except for \(p(U_{1i}, U_{1j}) = \rho_{1i2}\). We find that the two dimensional vector \((\mu_1 + U_{1i} + E_{\omega_1}, \mu_2 + U_{1j} + E_{\omega_2})'\) is bivariate normal with parameters \(N(\mu_1, \mu_2, \sigma_{\omega_1}^2 + \sigma_{\omega_2}^2, \rho_{\omega_1\omega_2})\), where
\[ \rho_{\text{ord}} = \frac{\rho_{12}\sigma_{\epsilon_1}\sigma_{\epsilon_2}}{\sqrt{\sigma_{\epsilon_1}^2 + \sigma_{\epsilon_2}^2}} \]

We find the conditional probability \( P(X_2 > t_2 \mid X_1 > t_1) \) by integrating the bivariate normal density over the quadrant \([t_1, \infty) \times [t_2, \infty)\) and divide that integral by the integral that is obtained form integrating the bivariate normal density over the half plane \([t_1, \infty)\).

So far we assumed that the process is an ordinary renewal process. In the case that the process is in equilibrium \( X_1 \) is a forward recurrence time and does not have a lognormal density (see section 5.3.2). It is computationally convenient to use an approximating lognormal distribution. If we use approximation (A) then we find that the vector \((\mu_1 + E_{\epsilon_1}, \mu_2 + E_{\epsilon_2})\) also has a bivariate normal distribution with parameters \( N(\mu_1 + \sigma_{\epsilon_1}^2 + c_1^{(\text{A})}, \mu_2, \sigma_{\epsilon_1}^2 + 2\sigma_{\epsilon_2}^2 + c_2^{(\text{A})}, \sigma_{\epsilon_2}^2 + \rho_{\epsilon\epsilon}, \rho_{\epsilon\epsilon}^{\text{eq}}) \), where

\[ \rho_{\epsilon\epsilon}^{\text{eq}} = \frac{\rho_{12}\sigma_{\epsilon_1}\sigma_{\epsilon_2}}{\sqrt{\sigma_{\epsilon_1}^2 + 2\sigma_{\epsilon_2}^2 + c_2^{(\text{A})}\sigma_{\epsilon_2}^2}} \]

\( c_1^{(\text{A})} \) and \( c_2^{(\text{A})} \) are given constants (see equation (5.12)).