Journal of Official Statistics, Vol. 22, No. 1, 2006, pp. 53–70

Estimating the Undercoverage of a Sampling Frame Due to Reporting Delays

Dan Hedlin[1], *Trevor Fenton*[2], *John W. McDonald*[3], *Mark Pont*[2], *and Suojin Wang*[4]

One of the imperfections of a sampling frame is miscoverage caused by delays in recording real-life events that change the eligibility of population units. For example, new units generally appear on the frame some time after they came into existence and units that have ceased to exist are not removed from the frame immediately. We provide methodology for predicting the undercoverage due to delays in reporting new units. The approach presented here is novel in a business survey context, and is equally applicable to overcoverage due to delays in reporting the closure of units. As a special case, we also predict the number of newborn units per month. The methodology is applied to the principal business register in the UK, maintained by the Office for National Statistics.

Key words: Frame quality; births and deaths; birth lags; right-truncated data.

1. Introduction

Most sample surveys draw their samples from a frame. More often than not, part of the target population is not accessible from the frame and in this situation the survey will suffer from undercoverage. A reporting delay or, using an equivalent term, a birth lag is defined as the time from birth (for a frame of businesses, the date when the business began to trade) to frame introduction (the date when the business came onto the sampling frame). Conversely, the death lag, causing overcoverage, is the time between cessation of activity (death) and the business being removed from the frame. Reporting delays are a considerable source of undercoverage in business surveys run by the Office for National Statistics (ONS) in the UK.

Most information on births and deaths is updated as soon as it is received in the ONS. However, some of it is held back pending further information or investigation. When the

[1] Statistics Sweden, Box 24 300, SE-104 51 Stockholm, Sweden. Email: dan.hedlin@scb.se
[2] Office for National Statistics, Cardiff Road, Newport N10 8XG, UK. Emails: trevor.fenton@ons.gov.uk and mark.pont@ons.gov.uk
[3] University of Southampton, Southampton Statistical Sciences Research Institute, Southampton SO17 1BJ, UK. Email: bigmac@soton.ac.uk
[4] Texas A&M University, Department of Statistics, College Station, Texas 77843-3143, U.S.A. Email: sjwang@stat.tamu.edu

Acknowledgments: This work was partly supported by funding from Eurostat, the Statistical Office of the European Communities. Wang's research was also supported by the U.S. National Cancer Institute (CA 57030) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106). Hedlin and Wang were at University of Southampton when the bulk of this research was undertaken. The views expressed in this article are those of the authors and do not necessarily represent those of their respective organisations.

size information indicates that the new unit has a workforce numbering twenty or more, and the unit cannot be matched against existing frame units, the recording of the unit is further delayed pending proving of the information about the unit. On average this adds about two months to the reporting delay these businesses would otherwise have. The lengths of birth lags form a highly skewed distribution. Some businesses report to the relevant authority in the UK as soon as they are set up, resulting in short lags. Others may have been operating for years below the level of annual turnover above which registration is compulsory, i.e., before their growth necessitates their registration. In these cases the lag may be very long indeed. Some businesses report to an administrative body in advance of their launch, sometimes resulting in a negative birth lag. Figure 1 shows the distribution of births by nonnegative birth lags (months). The vast majority of new businesses (85%) have been registered on the ONS frame within four months of their birth. About 10% have birth lags longer than five months.

This article presents a method for estimating the undercoverage that is caused by birth lags. A generalised linear model is fitted to historical frame records for which both birth dates and reporting delays have been recorded. The model will then be used for predicting forthcoming numbers and lags. While the method means we could accommodate economic cycles that have been observed in historical data, we have not attempted to do so as the available usable data relate only to the period January 1995–March 1998. Businesses that never come onto the frame, for example very small businesses or businesses operating entirely on the black market, are ignored, as are businesses that die before they appear on the frame.

There is a surprisingly sparse literature on reporting-delay induced undercoverage of a frame used for sample surveys, considering the importance of the problem and the fact that

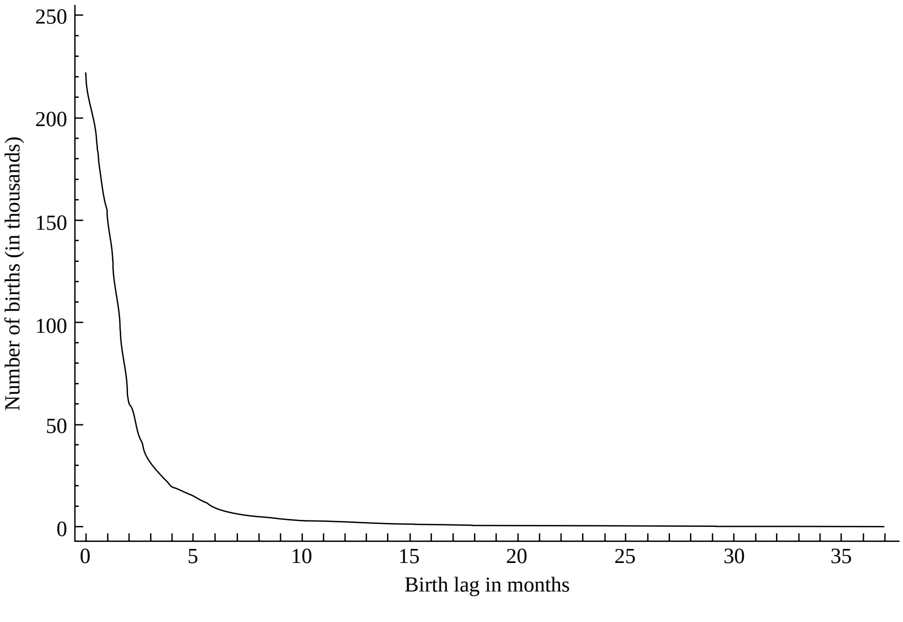

Fig. 1. Number of observed births (in thousands) against birth lag (months)

there is research on similar issues in other areas. For example, while Cox et al. (1995) discuss birth processes in detail, there is no reference to the specific problem of estimating undercoverage caused by these processes. The approach presented here to estimate the number of unobservable businesses is akin to and was inspired by estimation of the incidence of cases of AIDS in the presence of reporting delays (see Wang 1992, Sellero et al. 1996 and references therein). Our application is different: we have a very large dataset and a large contingency table. There is also a structure to our data that makes assumptions that are common in AIDS research less appealing.

An extension of the problem of predicting the population size is to predict the population total of some variable. Most businesses in transition between start and frame introduction are part of the target population and hence their absence from the sampling frame will result in a negative bias in estimated totals if these are based solely on samples from the frame. We propose a method of estimating this bias. A similar estimation problem is addressed in actuarial science. Insurance companies need to estimate the net sum of claims that have yet to be settled; see, e.g., Haberman and Renshaw (1996).

Section 2 describes the frame and frame-maintenance processes used at the ONS that lead to reporting delays. In Sections 3 and 4 Poisson regression models are used to predict the number of unobservable businesses. In Section 5 the precision of each model is assessed by means of a cross-validation study. Section 6 addresses the problem of bias in estimates of totals in the presence of reporting delays. The article concludes with a discussion in Section 7.

2. Birth Processes at the Office for National Statistics

The data used in our study were all births that occurred on ONS's business register, the Inter-Departmental Business Register (IDBR) (Perry 1995), between 1 January 1995 and 22 March 1998. A limitation of the data was that it was impossible to tell retrospectively whether a dead business had been closed because of a genuine death or because it had been part of a merger, takeover etc. While we focus on birth lags, the same methods could be applied to death lags, although deaths can be dealt with more easily as part of survey operations.

The administrative sources that the IDBR is built upon are two government departments: HM Customs and Excise and Inland Revenue. The former provides information relating to Value Added Tax (VAT)-registered legal units daily (weekly up to 1999). These indicate new registrations and any traders that have deregistered. The Inland Revenue provides a file of all Pay As You Earn (PAYE) employer records each quarter. In the PAYE scheme employers pay the employees' income tax and national insurance contributions. From these notifications, new registrations and deregistrations can be detected by comparison with the file from the previous quarter. Because the ONS is not notified continuously, frame introductions from the PAYE system tend to be clustered. The total number of businesses on the IDBR in 1998 was about 1.8 million (in addition to the data analysed here there were a large number of businesses that went unchanged through a period starting in 1995 and ending in February 1998).

Table 1 indicates the birth lag distribution for businesses born between 1 January 1995 and 22 March 1998. The rows of the table represent the numbers of businesses that were

Table 1. Number of observed births per lag (in months) and birth month. Partially unobservable cell counts are indicated with a ≥ symbol, totally unobserved cell counts with a dash

Birth month	Birth lag						Total number of observed births
	0	1	2	. . .	38	>38	
Jan. 95	5,444	4,982	1,910	. . .	≥ 6	–	16,054
Feb. 95	5,333	4,069	1,280	. . .	–	–	13,425
⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮
Jan. 98	7,783	4,102	≥1,346	. . .	–	–	13,231
Feb. 98	7,075	≥3,087	–	. . .	–	–	10,162
Mar. 98	≥5,888	–	–	. . .	–	–	5,888
Total	226,582	156,517	61,346	. . .	6	–	549,386

born in each month. We refer to the month a business started trading as its birth month. The columns are birth lags measured in months calculated as the number of complete months (successive periods of 30.4 days) between birth and frame introduction. With the observation window spanning the period January 1995–March 1998 the longest observable birth lag is 38 months. The count of the rightmost cell in the first row of Table 1 is unobservable (unless we gain access to data that go beyond the final date in the data currently available). Adhering to common terminology, cells with unknown counts are referred to as structural zeros (see, e.g., Agresti 1990); their unknown counts are represented in Table 1 with dashes. The term structural zero is conventional but in this case "unobservable counts" might have been more informative. With structural zeros, the table is an incomplete contingency table. The rightmost diagonal of the upper triangle containing observed counts is partially unobservable. Another way of expressing the fact that we cannot observe new businesses that have not yet been introduced to the sampling frame is to say that the series of data is right-truncated. The problem of estimating the undercoverage due to birth lags is equivalent to estimating the number of businesses that have been subjected to right-truncation.

On 31 March 1998, the undercoverage is the sum of the unknown counts in the lower triangle of Table 1. As a special case, the row totals can be predicted; they correspond to the number of births per month. Note that it is the column sums of Table 1, excluding partially truncated cells that are graphed in Figure 1.

In addition to measuring the overall length of birth lags, we have examined lags by industry classified by the Standard Industrial Classification 1992 (SIC92) and by region. There is little to choose between most of the different industries. However, it is clear that Health and Social Work has longer birth lags than any other industry. This is likely to be because registration in this sector is more dependent on the less frequent PAYE system. Most regions have very similar average lags except for Northern Ireland, which stands out as having greater than average lags. We do not take differential reporting delays in industries and regions into account in this article as they are inconsequential for national estimates. For subpopulations, however, models may need to accommodate differential lags.

As we focus on undercoverage due to birth lags, the businesses of interest are those that came onto the frame *after* they were born. In addition to this stipulation we selected for further analysis only those businesses with births between 1 January 1995 and 28 February 1998, to exclude the partly truncated diagonal in Table 1.

Table 2 and Figure 2 show some aggregates of births and the distribution of births. Except for the truncation effect clearly visible from November 1997 in Figure 2, the curve is astonishingly regular over time. Note that this curve represents the row sums of Table 1 apart from partially truncated cells. Note also that the scales of Figures 1 and 2 are very different; there is far more variability in counts between lags, especially short lags, than between birth months.

The longest birth lag we can fully observe is 37 months. Longer lags are entirely negligible as only 15 out of the 16,000 businesses that were born in January 1995 have 37 months birth lag; only 48 out of 30,000 businesses born in either January or February 1995 have 36 months birth lag or more.

Figure 3 displays number of births by day for births in 1995–1997. The two panels contrast the distribution of birthday for Aprils with that of other months. In Aprils 38 per cent of all new businesses started trading on the first of the month, in other months the proportion was even higher. The prominent peak at April 6 in Figure 3 is due to this day being the start of the taxation year in the UK. In practice, owing to differing interpretation of what constitutes the start of a business, it is frequently hard to fix on one day as the actual birthday for a business. The first of the month is often perceived as a convenient date for administrative purposes, both for the business managers and for the administrative bodies. Also, there is some clustering visible in Figure 3 in that most of the bars for dates like 10, 15 and so forth are slightly taller than most other bars. Therefore, month seems to be the smallest viable unit in the classification of number of births.

As Table 1 is very large, it cannot be easily displayed in an article. To bring out the structure of the table while avoiding obscuring detail, Figure 4 gives a contour plot of a two-dimensional bar chart based on Table 1 with one bar per cell and partly truncated cells excluded. The plot can be thought of as a contour map over the landscape of bars where altitude corresponds to the height of bars. The area with the largest counts is to the far left, and then the counts fall as we proceed to the right. The contour levels are 1,096, 148, 21, and 3 (equal distances on a log scale), so Area 1 consists of cells with counts greater than or equal to 1,096. A couple of the "pot-holes" in Area 4 are counts smaller than 3. Figure 4 indicates a large degree of homogeneity within areas. The dashed horizontal lines in Figure 4 mark Aprils. Areas 2 and 3, in particular, jut out along the dotted lines indicating

Table 2. Number of observed births per year and monthly average

Born in year	Number	Average per month
1995	174,300	14,500
1996	172,600	14,400
1997	171,300	14,200
1998 (Jan. and Feb.)	19,000	9,500
Total	537,200	14,100

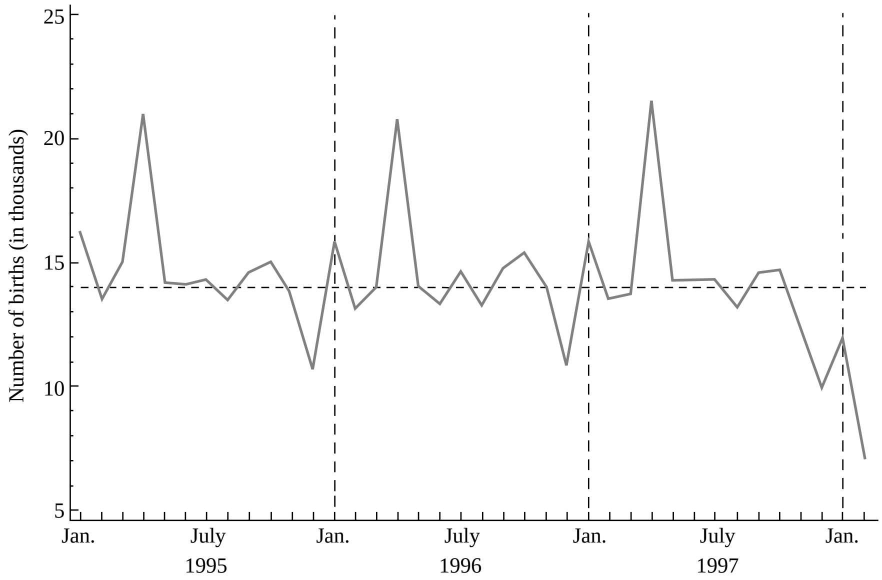

Fig. 2. Number of observed births (in thousands) per birth month

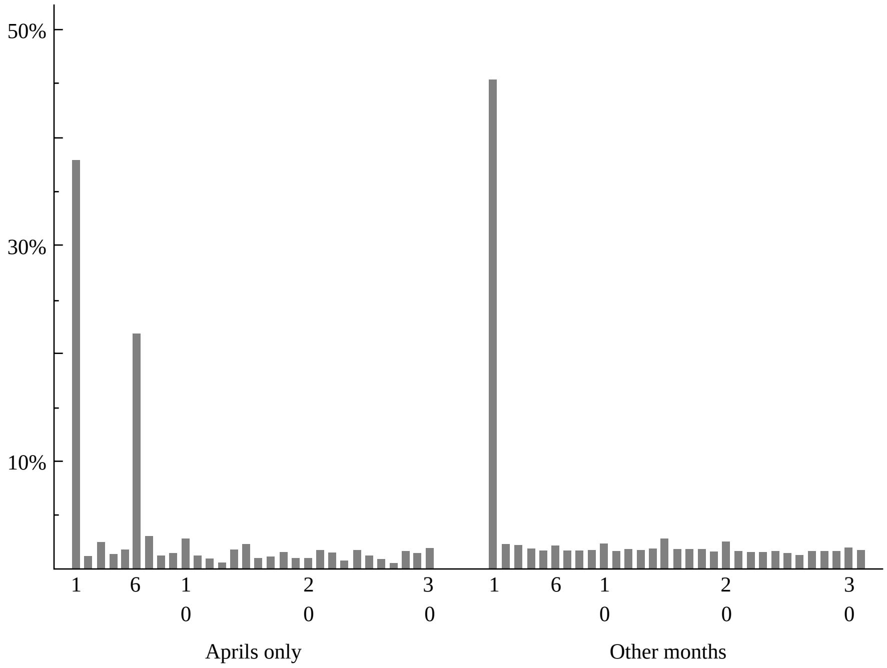

Fig. 3. Per cent of observed births per day of the month. Jan. 1995–Dec. 1997

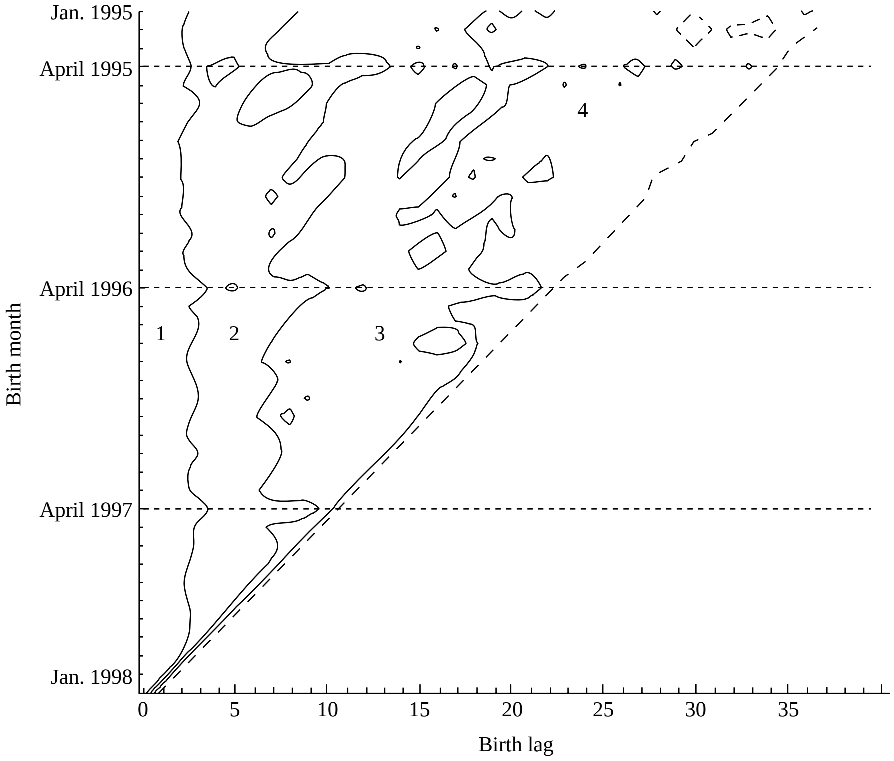

Fig. 4. A contour plot of the contingency table, Table 1. Levels for number of frame introductions: Area 1 corresponds to cells with 1,096 frame introductions or more; Area 2–4 to 148–1,095, 21–147 and 3–20 frame introductions, respectively

areas with relatively large counts that are stretched to the right. This is partly due to the fact that there are more births in the month of April, partly due to a more skewed lag distribution for businesses born in April (the average birth lag is 2.5 months for businesses born in April and 1.6 months for businesses born in other months). There is also a diagonal pattern emerging in Areas 2 and 3 above the horizontal line that indicates April 1996. The diagonals correspond approximately to frame introduction months. Businesses that came onto the frame in the same month are located along one or two diagonal lines in the contingency table running from bottom left to top right. It is likely that what produces these diagonal ridges visible in Figure 4 is the reporting of births from the Inland Revenue's PAYE system. Since this reporting is done on a roughly quarterly basis, the notifications of new businesses come in sizeable batches. These diagonals cannot be seen in displays of marginal distributions such as Figure 1.

3. Models

The number of businesses in transition between birth and frame introduction can be viewed as a stochastic process over time. The process is not stationary since Figure 4 indicates among other things that birth lags tend to be longer for businesses born in April than for businesses born at any other time of year.

In this section we fit models to the upper triangle of the contingency Table 1, excluding partially truncated cells. It is convenient to confine the class of models to generalised linear models (e.g., McCullagh and Nelder 1989). A generalised linear model has a random component, which identifies the probability structure of a response variable Y, a link function which specifies the relationship between the expected value μ of the response, and a systematic component, which defines a linear function of the explanatory variables. The systematic component can easily accommodate the seasonality and the non-stationary structure we have observed. Another advantage is that generalised linear models are useful even if the parametric assumption underlying the model is ill-fitting, since the ML estimation of parameters uses only the link function, choice of covariates and the variance function $V(\mu)$, where $V(Y) = \phi V(\mu)$ and ϕ is known as the overdispersion parameter (Davison and Hinkley 1997, Ch. 7). Thus our approach is essentially semi-parametric.

Let r be the number of rows in the table and let m_{ij} be the expected number of businesses that were born in month i, $i = 1, 2, \ldots, r$, and that were introduced on the frame in month $d = i + j - 1$, that is with a birth lag j, $j = 1, \ldots, c$, where c is the maximum birth lag we can observe. For convenience, index j starts at 1, i.e., $j = 1$ for 0 month birth lag, etc.

We have seen that the birth rate is higher in some months, such as Aprils, than in other months. It seems plausible that a higher (or lower) birth rate for certain months should give roughly proportionally larger (or smaller) counts of new businesses for all birth lags. Hence it seems more plausible that birth months, birth lags and other effects that potentially could be part of the systematic component are multiplicative rather than additive. This leads us to the following type of log-linear model:

$$\log(m_{ij}) = u + u_{(ij)} \tag{1}$$

for $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, c - i + 1$, where u is an intercept and $u_{(ij)}$ is a parameter for cell i and j in the fully observed triangle in Table 1, with total number of rows r and columns $c = r$; here $r = 38$. Hence the link function is the logarithmic function, which conveniently converts multiplicative effects on the original scale to additive effects on the log scale. The variance function $V(\mu) = m_{ij}$ is reasonable even if the cell counts are not independent and Poisson distributed, since the overdispersion parameter can account for discrepancies between the variance of the response and the variance function.

One of the most parsimonious models (i.e., with fewest parameters) that we may be interested in is a log-linear model with just birth lag effects with

$$u_{(ij)} = u_{lag(j)} \tag{2}$$

where $u_{lag(j)}$ is a parameter associated with birth lag j only, $i = 1, 2, \ldots, 38$, $j = 1, 2, \ldots, 38 - i + 1$. Considering Figures 1 and 2, the lag effect should be far more important than a birth month effect. Although a model with a lag effect only may be oversimplified, it is of interest as a reference model. Under this model all cells in a column have the same expected value. Another log-linear model arises from the assumption that the expected cell counts are separable into quasi-independent row effects and column effects with

$$u_{(ij)} = u_{birthmonth(i)} + u_{lag(j)} \tag{3}$$

with i and j as defined in (2). See McDonald (1998) for a definition of quasi-independence and ML estimation for incomplete tables. Since the underlying stochastic process is not stationary, there is an interaction between birth months and lags, which the quasi-independence model fails to capture.

A third model is one with a month-of-the-year effect and a lag effect. The underlying assumption is that some of the rows of the contingency table show a repetitive pattern in that their effects are the same and do not depend on year. Figure 2 suggests that all Januaries are similar, and so forth. It seems reasonable to examine a model with twelve 'month' parameters, as opposed to 38 birth month parameters. The model is:

$$u_{(ij)} = u_{month(k)} + u_{lag(j)} \tag{4}$$

$i = 1, 2, \ldots, 38, j = 1, 2, \ldots, 38 - i + 1, k = i\,(\text{modulo } 12)$.

When this model is fitted to the fully observed counts in Table 1, the residuals show a clear diagonal pattern, a pattern that is visible in Table 1 itself. A "diagonal effect" can be added to the model to obtain a better fit. The diagonal effect is essentially a time effect that corresponds to frame introduction months. Further, an "April effect" can accommodate part of the observed longer lags for businesses with births in April:

$$u_{(ij)} = u_{month(k)} + u_{lag(j)} + u_{diag(d)} + \alpha j I\,(k = 4) \tag{5}$$

with i, j and k defined as in (4), $d = i + j - 1$, α is a parameter and $I(\cdot)$ is an indicator function taking value 1 if the argument is true, 0 otherwise.

The models above were fitted to the fully observed upper triangle of Table 1 using ML estimation. The usual likelihood ratio test statistic (the "G^2 statistic") and the Pearson chi-squared test statistic gave very similar results. The estimation of parameters was done with Proc Genmod in the SAS System® version 8.02 for Windows. To check the numerical stability of the ML fitting algorithm for the large table analysed, the order of columns was changed, likewise the order of the rows for (3), but the results remained the same.

Table 3 gives the values of test statistics for four models ordered by the number of non-redundant parameters. The p-values are not given in the table below; all are miniscule. The G^2-values in Table 3 are extremely large due to the very large cell counts and the large

Table 3. Goodness of fit for Models 1–4

Model (formula in text)	# parameters	Degrees of freedom	G^2	Decrease in G^2	Knoke-Burke-ratio
1. Lags only (2)	38	703	49,323		
2. Lags and months (4)	49	692	38,259	11,064	22%
3. Lags and birth months (3)	75	666	36,888	12,435	25%
4. Lags, months, diagonals and April effect (5)	87	654	21,829	27,494	56%

number of cells. It is not meaningful in this application to use G^2-values for significance tests since any useful model would be rejected. All null hypotheses of no difference between models are also clearly rejected in this application. We can, however, use G^2-values for the comparison of models without formal tests. Another general strategy for dealing with large counts in a contingency table is to look for nonrandom patterns among residuals for different models. In Section 5 we also study how well the models predict future observations.

The Knoke-Burke ratio (Knoke and Burke 1980) is $[1 - G^2_{alt}/G^2_{ref}] \times 100\%$, where G^2_{ref} is the value of the test statistic under a reference model (here Model 1, lag effect only) and G^2_{alt} under an alternative model that includes the reference model as a special case. Note that if the alternative model is the saturated model then the Knoke-Burke ratio attains its maximum, 100 per cent. Knoke and Burke (1980) suggest that this ratio may be used for very large datasets; a large value indicates that the alternative model is satisfactory. We refer to the models using the order number in Table 3. Clearly, Model 4 gives the best fit. It is the addition of the diagonal effect that accounts for the major part of the reduction in G^2. Adjusted residuals from Model 4 are large but show no clear pattern.

There are other approaches in the AIDS diagnosis literature. Harris (1990) and Wang (1992) discuss parametric and nonparametric methods, respectively, to estimate the size of the population. Generalised additive models include generalised linear models (Hastie and Tibshirani 1986). The link function in these models is a sum of nonparametric curve components. Davison and Hinkley (1997, Examples 7.4 and 7.12) contrast our Model 3 with a generalised additive model which gives smoother predictions of unobservable counts in a register of English and Welsh AIDS patients. In our problem we could take $\log(m_{ij}) = u + u_{month(k)} + u(j)$ with $u(j)$ being some nonparametric curve describing the marginal relationship between cell counts and birth lags. Figure 1 suggests that the flat part of the curve may not need a different parameter for each birth lag, as it does in Models 1–4. We leave these ideas for future research.

4. Predicting Undercoverage and Number of Births Per Month

The models fitted to the upper triangle of the contingency table in Table 1 are now used for predicting counts in the lower triangle. Let $T = T_o + T_s$, where T_o and T_s are the sum of observable and unobservable cell counts, respectively, and let T be predicted by $\hat{T} = T_o + \hat{T}_s$, where $\hat{T}_s$ is a predictor for T_s. Note that the models in Section 3 are not defined for cells with structural zeros. Under the natural assumption that Models 1–3 can be extended to comprise the entire table, we can take $\hat{T}_S = \sum_S \hat{m}_{ij}$. For Model 4 it is assumed that the diagonal pattern observed for the last 12 months can be extrapolated periodically; that is, to predict cells along a diagonal d' in the part of the lower-right triangle where $c + 1 \leq d' < c + 12$, the parameter associated with diagonal $d' - 12$ in the upper-left triangle is used. To predict cells along a diagonal in the next band of twelve consecutive diagonals, $c + 13 \leq d'' < c + 24$, the parameter associated with diagonal $d'' - 24$ is used, and so on. Thus, only the rightmost band of twelve diagonals in the observed triangle is used for prediction. While this may seem to underutilise the information, there does not seem to exist a periodic model for the diagonal effects that uses all observed diagonals and gives smaller prediction errors than the model just described that only uses the last twelve observed diagonals.

Table 4 gives the number of births aggregated to year levels. The observed count in 1997 is about 8–9 per cent less than the predicted count. The difference between the sum of the predicted counts under Model 4 and the observed count is $570{,}000 - 542{,}000 = 28{,}000$. Hence, in terms of number of businesses the undercoverage due to reporting delays is about 1.6 per cent (28,000 in 1.8 million).

Figure 5 shows the observed and predicted number of births per month for Models 2–4. The grey curve representing observed births in Figure 5 is the same one as in Figure 2. Judging from Figure 5 there is little to choose between the prediction methods, with only Model 3 being somewhat separated from the others. There is a 1 per cent truncation effect as early as September 1995 that each model captures.

5. Prediction Error

To assess the prediction error, we can turn the clock back, for example to the end of May 1995, and pretend that all observed businesses born afterwards are unknown. Hence there will be a 5 × 5 square subtable with observed counts in the upper-left triangle and "missing" counts in the lower-right triangle. A natural estimate of the error is obtained by estimating parameters for the upper triangular subtable and basing the prediction error on the difference between the observed and predicted counts in the lower-right triangular subtable. Using this approach, Figure 6 shows the number of births per month for data cut off at the end of April 1997. The grey curve is the number of births per month obtained from the full original table (that is, it is the same curve as in Figure 2). Models 3 and 4 are indistinguishable while Model 2 predicts the rise in births in April rather better than the other models.

Thus the end points of the solid black curves in Figure 6 show the predicted number of births for the month that corresponds to the last row of the particular triangular subtable, which has been obtained by cutting the full table off at the end of April 1997.

Figures 7 and 8 show the prediction errors for a series of subtables, from the one obtained by cutting off at the end of December 1995 to the one where data after December 1997 were discarded. In Figure 7 the final-month errors are shown, defined as the difference between the predicted number of births in the last month of the subtable and the observed number of births in the same month in the part of the original table covered by the subtable.

Table 4. Observed number of births per year and the predicted to observed ratio

Year	Observed number of births	Ratio predicted count to observed count			
		Model 1	Model 2	Model 3	Model 4
1995	175,898	1.00	1.00	1.00	1.00
1996	174,013	1.01	1.01	1.01	1.01
1997	172,570	1.09	1.08	1.09	1.08
1998	19,103	1.75	1.74	1.92	1.69

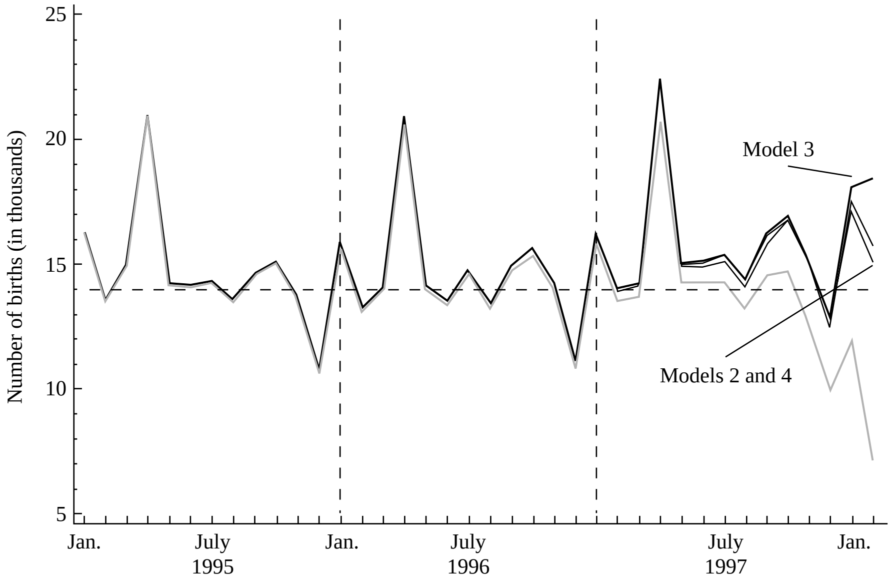

Fig. 5. Predicted number of births (in thousands) per month under Models 2–4. The observed counts are graphed with a grey line

The part of Figure 7 to the right of July 1997 is clearly influenced by the bias resulting from truncation of the original series. In the beginning of the series the error is, as expected, large due to the fact that in the beginning of the series there is less data for the estimation of parameters. It seems reasonable to forego the prediction errors before July 1996 and after July 1997. As seen in Figure 7, Model 2 gives smaller final-month errors

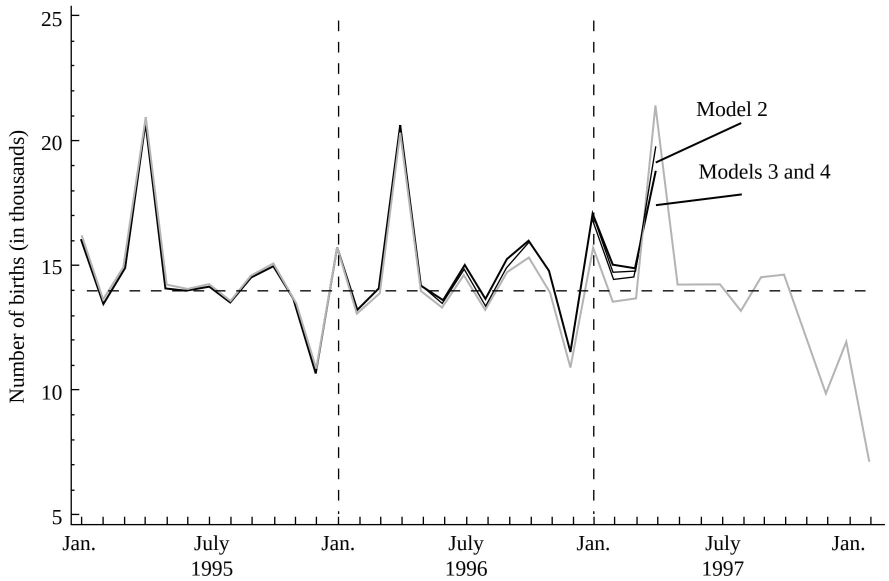

Fig. 6. Predicted number of births (in thousands) based on data up to 30 April 1997: Models 2–4 and observed counts as at 28 February 1998 (grey line)

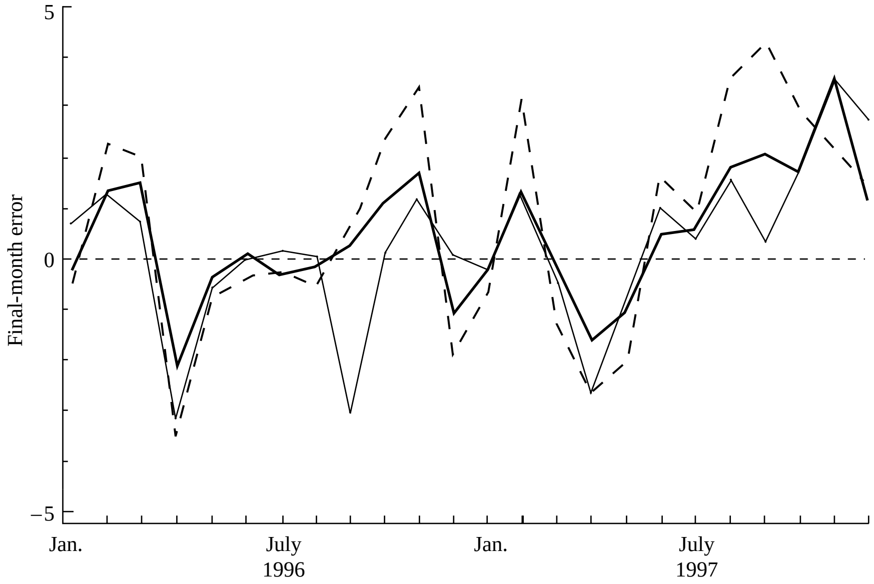

Fig. 7. Difference between predicted and observed number of births (in thousands) for the final month in successive subtables. Three models: Model 2 (thick line), Model 3 (dashed line) and Model 4 (thin line)

than Model 3 for each month in this interval. This may seem paradoxical since Model 3 has more parameters and gave a better fit to the upper triangle of the contingency table (see Table 3). However, the models play two roles here. One is to fit counts in the upper triangle of the contingency table. The other is to be a tool for prediction. Good performance in one

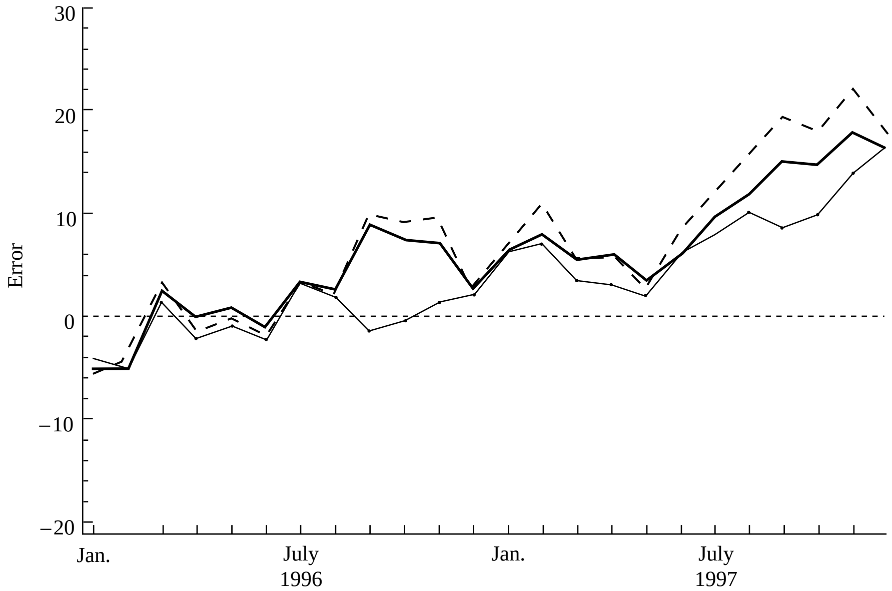

Fig. 8. Difference between the sum of predicted number of births and observed number of births (in thousands) in successive subtables. Three models: Model 2 (thick line), Model 3 (dashed line) and Model 4 (thin line)

of these roles does not necessarily imply good performance in the other. Model 3 does not draw on the seasonal pattern. Stated somewhat loosely, Model 2 borrows strength from similar months in previous years. With Model 3, the predictions depend completely on single rows of the table and are much more variable. Model 2 has the additional advantage over Model 3 that it allows prediction beyond February 1998. The average final-month error in absolute terms over July 1996–June 1997 is for Models 2–4, respectively, 800, 1,680, and 910.

The largest prediction error in absolute terms for Model 2 in the interval July 1996–July 1997 is about 1,800, which occurs in November 1996. Thus the ratio of this prediction error to the average number of births per month, 14,000, is about 17 per cent. Cross-validating in the same way for the second last row gives 2,000 as the largest prediction error. The estimated error of this kind for the third last row is 1,700. The sum of all rows is about 10,000. Thus, a conservative estimate of the error of the estimated undercoverage is 10,000.

The difference between the sum of monthly predictions and observations is a measure of error more directly connected to the estimation of the undercount. These differences for a sequence of subtables are displayed in Figure 8. In the beginning of the series the difference is negative because the predictions for 1995 are too low. The difference becomes positive when the truncation effect in the original series becomes pronounced. Figure 8 makes it clear that Model 4 is better than Model 2. As seen in Figure 8, the largest prediction error in absolute terms for Model 2 in the interval July 1996–July 1997 is less than 10,000. For Model 4 the largest error is less than 6,000. The average absolute error over July 1996–June 1997 is for Models 2–4, respectively, 5,560, 6,290, and 3,220.

6. Estimating Totals in the Presence of Reporting Delays

Suppose the aim is to estimate the total $t_y = \sum_U y_k$ of a study variable $\mathbf{y}' = (y_1, y_2, \ldots, y_N)$ on a population U with unit labels $\{1, 2, \ldots, N\}$. Let U_{ij} be the set of businesses with birth month i and reporting delay j. The total of the unseen part of the population, t_{Us}, is the sum of $t_{yij} = \sum_{Uij} y_k$ over the not fully observed cells (i, j) in Table 1, each of which holds U_{ij}.

The chain ladder method is widely used in insurance practice to estimate the sum of incurred but not reported (IBNR) losses for which the clients are insured, which is a problem that is technically similar to ours. Mack (1991) and Renshaw and Verrall (1998) show that the chain ladder technique necessarily gives the same cell predictions as the quasi-independence model (i.e., our Model 3). An extension of the chain ladder technique is thus to apply Models 2 and 4 to observed totals of some frame variable to predict nonobserved cell totals of this variable.

Once the total of a frame variable has been estimated, this information can be used either to report on the possible extent of undercoverage as part of the survey's quality measures, or it can be used to update survey estimates. This can be done in a simple way by assuming that the relationship between the study variable **y** and the auxiliary variable **x** can be represented by data from the known part of the population. Such an assumption may be adequate, but would need testing. Clearly, in some situations, further research will be necessary to bridge the gap between initial estimates of undercoverage and being able to incorporate such adjustments into survey results.

The variable turnover at frame introduction was stored for the businesses whose counts are reported in Table 1. It turned out that businesses that are very large when they come onto the frame tend to have long birth lags. It is believed that few of these large businesses are genuinely new; rather they are the result of mergers and other types of restructuring, or appear large because of errors in reporting their turnover at birth. For instance, the turnover may be reported in £s rather than the required £1,000 s. To avoid duplication large businesses that are reported as new are subjected to an often lengthy proving process, which cannot usually be done without the help of the business itself.

We modelled stratum totals of turnover using the same methods we applied to the counts. Cross-validation errors that parallel those of Figure 7 are displayed in Figure 9. The estimated total undercoverage is £2.4bn. Unfortunately, the errors shown in Figure 9 are as large as the point estimate itself. The large businesses with long lags in our dataset make prediction intrinsically difficult. They enter the frame irregularly and produce large variation in total turnover per birth month. In a production system, one would wish to deal with these clerically as they occur; it was a limitation of our data that we could not distinguish with certainty true births from other restructurings.

In practice, most births are quite small and come on to the frame relatively quickly. Those businesses that have existed below the VAT registration threshold for some time, but grow to be large enough to register, also tend to be small, and it is probably unfeasible in most contexts to try to rework old data to take account of their existence for more than a couple of previous years. Genuinely large births are few and far between, and generally come about as a result of businesses restructuring. In ONS, mergers, takeovers etc. are observed through businesses' responses to surveys, so there is continuity of coverage, and

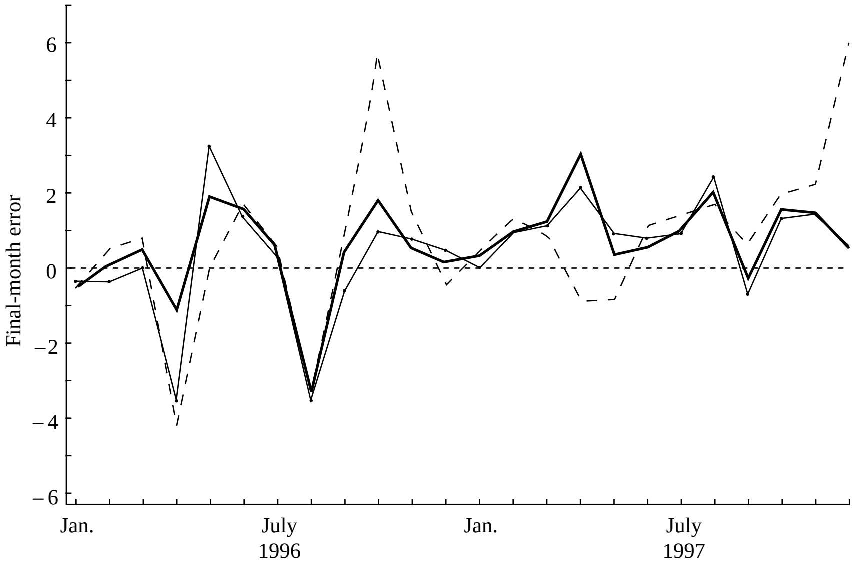

Fig. 9. Difference in £bn between predicted and observed number of births for the final month in successive subtables. Three models: Model 2 (thick line), Model 3 (dashed line) and Model 4 (thin line)

so the actual undercoverage and the estimated error are less than indicated above. This point stresses the importance of adjusting the proposed methodology to the specifics of the frame introduction process at the particular National Statistical Institute.

7. Discussion

As Struijs and Willeboordse (1995) point out, there is a demand for "business demographic information." However, as shown in Figure 2, naively monitoring observable births will underestimate the true trend. In addition to the estimation of births, we believe that the work initiated here provides beneficial measures of frame quality and the size of one important source of undercoverage. The number of businesses in transition between administrative sources and the frame can be estimated on a monthly basis. A time series of these estimates is a useful tool for monitoring frame quality. For example, a long-term increase could be used to investigate whether changes have occurred in the source departments, and what action needs to be taken by the statistical agency. This can be compared with more commonly used quality indicators, such as the response rate. While the response rate alone is of rather limited value to assess nonresponse bias, it is indispensable as a tool for monitoring the survey-taking climate and in-house production processes. A decrease in the final response rate of a repeated survey may indicate, for example, staffing problems. In fact, the undercoverage in terms of number of businesses is stronger than the nonresponse rate as a lack-of-quality indicator. While an increasing nonresponse rate does not necessarily correspond to increasing bias, increasing undercoverage does.

We have predicted gross totals with a log-linear model. The prediction error was estimated with a nonparametric method that has considerable natural appeal. At the end of February 1998 the undercount was 28,000 businesses, or 1.6 per cent of all registered businesses. The error of this estimate was predicted to be less than 6,000.

The sum of the turnover of the unobservable businesses could not be predicted with any accuracy because of a heavy tail in the reporting delay distribution. The heavy tail is due to the fact that many businesses that are very large when they enter the frame are not genuinely new businesses, and we were unable to distinguish these in our data. When the data are applied in an operational setting, we expect to be able to deal with such problems. Since the history of businesses is currently not stored on the business register of the ONS, it has been proposed to create a new life status variable that will store more complete information about changes to businesses. This will be a log of events that have occurred in the life of the business and will allow the separation of genuinely new businesses from businesses that are new only in a legal sense. This will permit more extensive research through making it possible to recreate the survey population at any point in time. Being able to predict accurately the bias of a frame variable enables estimation of the bias of survey variables through models of the association between the frame variable and each survey variable.

There are other approaches in actuarial science that may be of interest in this application. Overviews of the IBNR prediction problem are given by England and Verrall (2002) and De Vylder (1996, Ch. 7). It is common to assume stationarity for IBNR prediction. Klugman, Panjer, and Willmot (1998, p. 292) argue that modelling counts and the continuous variable separately has some advantages in the IBNR losses context. In the situation addressed in this article, it would be useful to compare the distribution of the study variable for different birth lags with that of the counts. Also, to investigate the effect of legal and procedural changes (for example if the VAT threshold for mandatory reporting to the relevant UK authority is changed or if new proving processes are introduced at the ONS) it is helpful to model the distribution of the counts and the study variable separately to avoid confounding.

The lags that deaths produce are similar to birth lags, although they are often longer. We have focused on births. Overcoverage due to reporting delays of deaths can be modelled with the same methods that we have used for births. An interaction term could be included in a model that encompasses both births and deaths if there is reason to believe that, for example, an increase in births often leads to a subsequent increase in deaths.

8. References

Agresti, A. (1990). Categorical Data Analysis. New York: Wiley.

Cox, B., Binder, D., Chinnappa, N., Christianson, A., Colledge, M., and Kott, P. (eds) (1995). Business Survey Methods. New York: Wiley.

Davison, A.C. and Hinkley, D.V. (1997). Bootstrap Methods and Their Application. Cambridge University Press.

De Vylder, F.E. (1996). Advanced Risk Theory. Brussels: Editions de l'Universite de Bruxelles.

England, P.D. and Verrall, R.J. (2002). Stochastic Claims Reserving in General Insurance. Paper presented to the Institute of Actuaries, London, UK, 28 Jan.

Haberman, S. and Renshaw, A.E. (1996). Generalized Linear Models and Actuarial Science. The Statistician, 45, 407–436.

Harris, J.E. (1990). Reporting Delays and the Incidence of AIDS. Journal of the American Statistical Association, 85, 915–924.

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. Statistical Science, 1, 297–310.

Klugman, S.A., Panjer, H.H., and Willmot, G.E. (1998). Loss Models: From Data to Decisions. New York: Wiley.

Knoke, D. and Burke, P. (1980). Log-Linear Models. Sage University Paper Series on Quantitative Applications in the Social Sciences (07–020). Beverly Hills and London: Sage Publications.

Mack, T. (1991). A Simple Parametric Model for Rating Automobile Insurance or Estimating IBNR Claims Reserves. ASTIN Bulletin, 21, 93–109.

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models, 2nd ed. London: Chapman and Hall.

McDonald, J.W. (1998). Quasi-Independence. In Encyclopedia of Biostatistics, P. Armitage and T. Colton (eds). New York: Wiley, 3637–3639.

Perry, J.A. (1995). The Inter-Departmental Business Register. Economic Trends, 505, 27–30. London: CSO.

Renshaw, A.E. and Verrall, R.J. (1998). A Stochastic Model Underlying the Chain-Ladder Technique. British Actuarial Journal, 4, 903–923.

Sellero, C.S., Fernández, E.V., Manteiga, W.G., Otero, X.L., Hervada, X., Fernández, E., and Taboada, X.A. (1996). Reporting Delay: A Review with a Simulation Study and Application to Spanish AIDS Data. Statistics in Medicine, 15, 305–321.

Struijs, P. and Willeboordse, A. (1995). Changes in Populations of Statistical Units. In Business Survey Methods, B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge, and P. Kott (eds). New York: Wiley, 65–84.

Wang, M.-C. (1992). The Analysis of Retrospectively Ascertained Data in the Presence of Reporting Delays. Journal of the American Statistical Association, 87, 397–406.

Received May 2003
Revised November 2005