KM 40 (1992)
pg.69-80

MODELLING MANIPULATED DISCRETE PROCESSES: ESTIMATING PREVALENCE IN THE CASE OF NON-RESPONSE

D. Lugtenburg and P.G.H. Mulder

SUMMARY

Missing data may cause a problem in the estimation of the prevalence of a certain disease in a population by means of a survey. This chapter presents a model-based approach for dealing with missing data. The model involves several strategies, each resulting in a different estimate of this prevalence. The strategies will be compared by means of likelihood ratio testing procedures within the parametric setting. Of course, properly testing goodness of fit is not possible solely through a likelihood ratio test as long as missing data are not replaced with resampled information. The parametric model is shown to be applicable also for situations in which there is observational error.

Department of Epidemiology and Biostatistics
Erasmus University Medical School
P.O. Box 1738
3000 DR Rotterdam
The Netherlands
tel. 010-4087483

1 INTRODUCTION

In some situations observations collected are not correct or complete. One reason may be that, before outcomes of the relevant process can be observed, a second process has eliminated part of these outcomes or changed part of them. This second process "manipulates" the outcomes of the first process. Probably, the manipulating processes encountered in practice most often are "vanishing" and "recategorizing". The latter is a process where only part of the outcomes of the first process are observed correctly. If the former process is active, part of the outcomes of the first process become missing.

The nature of "vanishing" processes as they may occur in practice varies widely. For instance, in surveys where questionnaires are sent to persons of specific populations, it is only rarely that all questionnaires are returned or are completely useful for analysis. There is a certain percentage of non-response. The mechanisms that underly the existence of such missing data may affect the appropriate analysis, see e.g. Rubin (1976). An excellent introduction to this subject has been written by Little and Rubin (1987).

Three general approaches are suitable to deal with the "vanishing" problem. First, extra information about the "missing group" can be gathered by resampling. Second, the missing data can be "filled-in" or "imputed" and corresponding analyses performed. Finally, a specific structure of missingness can be assumed and conclusions can be derived from that assumption. For estimating prevalence all three general approaches have been proposed in the literature. In estimating the size of the western Arctic stock of bow head whales, Zeh et al. (1986) use the assumption that missingness is only related to visibility and obtain estimates by resampling. If several sampling procedures are available (the one that is the most expensive being the most precise) a correction formula can be derived. An example is the procedure of counting radio-labelled animals, which is not interfered with by visibility problems (Steinhorst and Samuel, 1989). The second and third of these general approaches are sometimes used simultaneously, as in the example of this paper. By supposing that all missing data show the outcome of interest (e.g. have the disease or have high blood pressure) an upper limit of the prevalence is obtained. The model postulated in this paper makes it possible to evaluate such imputation procedures. Dinse (1986) discusses similar but nonparametric estimators.

The following example will be discussed. A chemical industry needed an estimate of the percentage employees with hypertension. In a survey all employees were invited to have a medical examination. Some employees, however, decided not to attend. These refusers presumably did not form a random subgroup of the total population. A non-random process may have caused some employees to attend and others not to attend the examination. Assumptions concerning the non-randomness of this underlying missing-data-generating process are needed in order to properly estimate the prevalence of hypertension. Several simple sets of assumptions will be discussed, as well as likelihood ratio tests for discriminating between these sets.

Besides the problem of missing data, there is also the problem of misclassification. An observer may make mistakes in allocating an observed outcome to the correct category. Many

authors have written about this phenomenon. The first and the last of the three general approaches referred to earlier are proposed in the literature to deal with it. Obviously, resampling by means of repeated measurements can improve the estimates considerably, as discussed by Clayton (1985). Also, models have been proposed which *a priori* allow for misclassification. Copas (1988), for example, discusses binary regression models.

In this paper, a framework is presented that may facilitate the modelling of both the recategorizing and the vanishing process. As has been remarked before, we will assume that the outcomes of the first process are manipulated by a second process. Therefore, only certain combinations of outcomes of both processes can be observed. First, the underlying principles will be discussed, illustrated step by step by showing the implications of modelling a vanishing process. This generally applicable model is presented in section 2 for binomial successive processes. The distribution of the actually observed outcome variate will be demonstrated to be multinomial. The model will be specified and the procedure to obtain maximum likelihood estimators (MLE) for the model parameters will be described. For the algorithm we use the composite link approach as introduced by Thompson and Baker (1981). The several different imputation procedures to estimate prevalence will be discussed within the general framework. In section 3, the above-mentioned example (prevalence of hypertension) is dealt with in more detail. The analysis is elaborated as far as necessary to serve illustrative purposes. Section 4 describes the recategorizing process within the general framework, and a discussion is presented in section 5.

2 THE MODEL

2.1 *Introduction of the model*

Consider two successive binomial processes with the second process manipulating the outcomes of the first process. The first process has two possible outcomes, z_1 and z_2. The probability that the first process actually has outcome z_1 is denoted q_1. The second process has two possible outcomes, z_{21} and z_{22}, where the probability that outcome z_{21} actually occurs may depend on the outcome of the first process. As a consequence, another two probabilities are introduced: q_2 is the probability that outcome z_{21} occurs when z_1 is the outcome of the first process, q_3 being the corresponding probability when z_2 is the outcome of the first process. The resulting process is illustrated in Figure 1.

There are four response categories r: (z_1, z_{21}), (z_1, z_{22}), (z_2, z_{21}) and (z_2, z_{22}) with r being 1 to 4, respectively. The response category r (r=1,..,4) is observed y_r times. The four response categories are complete (there are no other responses possible) and mutually exclusive. For each response category r the corresponding category probability, p_r, is assumed constant with $p_1 = q_1.q_2$, $p_2 = q_1.(1-q_2)$, $p_3 = (1-q_1).q_3$ and $p_4 = (1-q_1).(1-q_3)$. Consequently, the response variate, which describes the frequency with which this category will actually be observed, has a multinomial distribution, see e.g. Johnson and Kotz (1969).

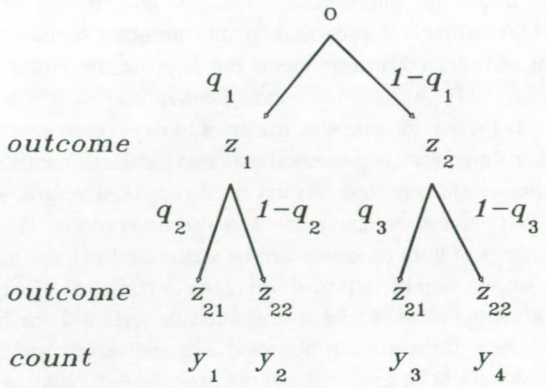

Figure 1 Schematic description of two successive processes.

If the second process is a missing-data-generating process in which z_{21} means "not missing" and z_{22} means "missing", it is easy to see that in fact one cannot observe all category counts separately. The counts that can be observed are (y_1), (y_3) and the sum of y_2 and y_4: (y_2+y_4). In matrix notation: $y_{observed} = C\,y$, with $y=(y_1, y_2, y_3, y_4)^T$ and

$$C = \begin{bmatrix} 1000 \\ 0010 \\ 0101 \end{bmatrix}.$$

In short, only combined category counts are observed. Somehow, one has to deal with this limited information when estimating the dependence of the probabilities q_j on a vector of explanatory variables. This will be discussed in subsection 2.4. In the next subsection, 2.2, the dependence model for the probabilities q_j is specified.

2.2 *Postulation of the model*

For design point i $(i=1,..., s)$ the process probabilities q_j $(j=1,2,3)$ are assumed to depend on p explanatory variables assembled in a vector x_i:

$$q_{ij}=q_j(x_i) \quad j=1,2,3; \quad i=1,...,s.$$

For every q_j, a logistic model is postulated as follows:

$$logit(q_j)=\beta_j^T x_i \quad j=1,2,3; \quad i=1,..., s$$

with β_j a vector of p coefficients specific for each probability q_j. The first element of the vector

x_i equals 1 in order to provide intercepts β_{j1}. Some of the other coefficients in β_j may be zero, meaning that the corresponding explanatory variable in x_i is not incorporated in the j^{th} process.

2.3 *Several strategies for estimating prevalence*

Suppose that persons having the disease of interest, e.g. hypertension, fall in category z_1. Again, the second process is a missing-data-generating process where z_{21} means "not-missing" and z_{22} means "missing". Then an estimator for the prevalence (percentage observations falling in category z_1) is defined as follows. When the estimated category counts for the category counts y_{ir} are denoted $\hat{y}_{ir}$ then the prevalence estimator is:

$$prevalence = \frac{100 \sum_{i=1}^{s} (\hat{y}_{i1} + \hat{y}_{i2})}{\sum_{i=1}^{s} n_i}.$$

The conditions that define independence of both processes are straightforward. Both processes are independent if and only if $\Pr(z_{21}|z_1)=\Pr(z_{21}|z_2)$ for all design points i, which is tantamount to $q_{i2}=q_{i3}$, or $\beta_2^T x_i=\beta_3^T x_i$, for all i (this means $\beta_2=\beta_3$). This independence condition results in a "simple-guess-estimator" for the prevalence.
A generalization is the following:

$$logit(q_{i2}) - logit(q_{i3}) = \psi$$

for all i. If $\psi=0$ this simplifies to the above case $q_{i2}=q_{i3}$. If $\psi \to \infty$, then $q_{i2} \to 1$, meaning in the example introduced above that all missings are counted as normotensives. This leads to a "lower-boundary estimator" of the prevalence.

If $\psi \to -\infty$, then $q_{i3} \to 1$, meaning that all missings are counted as hypertensives. This results in an "upper-boundary estimator" of the prevalence.

It will be obvious that these last two situations are extreme situations and unlikely to be real. However, "how" unlikely are they? Interval estimation for the prevalence can be evaluated by means of a likelihood ratio testing procedure as follows. First, consider the following three situations: $\psi = -\infty$ (upper bound), $\psi = 0$ (simple guess) and $\psi = \infty$ (lower bound). For every situation the corresponding deviance is calculated. Secondly, the deviance corresponding to the MLE of ψ is calculated. This leads to an "optimal-guess estimator" for the prevalence. For each fixed ψ the likelihood ratio test statistic is calculated as the difference: deviance(ψ_{fixed}) minus deviance($\psi_{optimal}$). When this statistic is supposed to have a chi-squared distribution with 1 degree of freedom (DF for short) on the assumption that ψ_{fixed} is the true ψ, then all values of ψ with a deviance less than 3.84 (chi-squared value, 1 DF and $\alpha=0.05$) away from the deviance corresponding to $\psi_{optimal}$, correspond to values that are in the 95 per cent confidence interval of

the prevalence, because of the monotonic relationship between ψ and the prevalence.

2.4 *Fitting the model*

If the second process is a missing-data-generating process and $q_{i2}=q_{i3}$ or equivalently $\beta_2=\beta_3$, MLEs for the model parameters β_1 and β_2 can be calculated by applying standard binomial algorithms. Two separate binomial models can be fitted. The first for calculating MLEs for the model parameters of the first process: y_{i1} positive responders out of a total of $y_{i1}+y_{i3}$ observations at design point i. For the second process this is: $y_{i1}+y_{i3}$ as positive responders out of a total of n_i observations at design point i. By adding the corresponding binomial log-likelihoods it can be shown that the total log-likelihood is the appropriate multinomial log-likelihood. The more extreme models with either $q_{i2}=1$ or $q_{i3}=1$, too, can be fitted with standard binomial algorithms.

In all other situations, no standard binomial algorithms can be used. As a general method to maximize the multinomial log-likelihood, a Poisson reparametrization will be introduced.

A sufficient condition for the outcomes y_{ir} $(r=1,..,4)$ at design point i to behave like a multinomial distribution for given p_{ir} and $n_i=y_{i1}+y_{i2}+y_{i3}+y_{i4}$ is a Poisson distribution for the y_{ir}, stratified on design point i. Birch (1963) showed that both likelihoods are proportional to each other and so lead to identical MLEs in a log-linear model specification. Palmgren (1981) showed that the inverses of the Fisher information matrices are identical so that the asymptotic covariance matrices of the estimates also coincide.

We now specify the following dependency model for the expected responses $E(Y_{ir})$ with a model parameter μ_i representing the stratification on design point i, so that there is a one-to-one relationship between the linear logit(q_{ij}) predictors (with coefficient vectors β_j) and the linear log(p_{ir}) predictors:

$$E(Y_{i1})= n_i\, p_{i1}= \exp(\mu_i+\gamma_1^T x_i+\beta_2^T x_i)$$

$$E(Y_{i2})= n_i\, p_{i2}= \exp(\mu_i+\gamma_1^T x_i)$$

$$E(Y_{i3})= n_i\, p_{i3}= \exp(\mu_i+\beta_3^T x_i)$$

$$E(Y_{i4})= n_i\, p_{i4}= \exp(\mu_i)$$

with:

$$\beta_1^T x_i=\gamma_1^T x_i+\ln\left(\frac{\exp(\beta_2^T x_i)+1}{\exp(\beta_3^T x_i)+1}\right)$$

resulting for ψ in:

$$\psi = \beta_2^T x_i - \beta_3^T x_i \ .$$

As ψ has to be constant across i, ψ is the difference of the intercepts of process 2 and process 3: $\psi = \beta_{21} - \beta_{31}$, the other coefficients in β_2 being equal to those in β_3.

The above reparametrization to a Poisson regression model provides a way of handling the problem in question, namely that the y_{ir} at design point i are not observed directly, but that only certain combinations over r of the y_{ir} are observed: $y_i(obs) = C\, y_i$ with $y_i(obs)$ being an observed column vector of dimension $k<4$ and y_i a column vector of elements y_{ir} (r=1,...,4); the (k*4)-matrix C being the composite link matrix which is the same for all design points i. A typical C-matrix has zeroes and ones as elements. Composite Poisson models are described by Thompson and Baker (1981), who introduced the composite link model as a generalization of the generalized linear model (McCullagh and Nelder, 1983). The algorithm described by Thompson and Baker has also been applied to the example of this paper. For the vanishing process the C-matrix already was presented in subsection 2.1. The Poisson reparametrization provides a generally applicable algorithm without any further conditions.

3 AN APPLICATION: THE PREVALENCE OF HYPERTENSION

A survey was conducted on 6287 employees of Shell Pernis, ranging in age from 20 to 59 years, to find the total percentage of employees with hypertension. This was defined as a diastolic blood pressure of at least 95 mm Hg or a systolic blood pressure of at least 160 mm Hg. The target population was the total working population on January 1, 1982 at the site. The employees were invited to undergo a physical examination that year. Of about 25 per cent of the employees, however, no blood pressure values were recorded. This is partly due to the fact that a number of employees refused to attend, but also because of some computer storage problems. The latter is known because the observed blood pressure values of some employees, who were known to have attended the examination, could not be retrieved at the time of analysis in 1989.

In the general context of this paper the first process is considered to be the process that determines whether a person has high blood pressure. The second process determines whether the blood pressure values could be retrieved for the analysis in 1989. A very relevant explanatory variable for the first process is age, which was grouped into four classes: 20-29, 30-39, 40-49 and 50-59 years. One explanatory variable for the second process is the number of times a person reported sick during 1982. The underlying reasoning is simple. If an employee was sick, he did not get an invitation to attend the examination. Two classes were used, namely "reporting sick less than four times" and "reporting sick at least four times". No other explanatory variables were used. This is also an obvious determinant for process 1, especially for correctly estimating the prevalence. A summary of frequencies in the resulting 8 age-sickness classes is presented in Table I.

TABLE I

Observed percentages of definite hypertension in a survey of an industrial population for different age and sickness-absence categories

Age (years)	Reporting sick at least four times	Number of employees	Percentages data that are unknown	Within the data that are known, the percentage persons with definite hypertension
20 - 29	no	1197	24.6	8.4
20 - 29	yes	245	33.1	12.2
30 - 39	no	1318	27.5	9.6
30 - 39	yes	203	36.0	13.8
40 - 49	no	1267	21.7	14.3
40 - 49	yes	171	35.7	20.0
50 - 59	no	1642	24.5	20.7
50 - 59	yes	244	32.0	23.5

In this example the first binomial process was defined as the process that determines whether a person has high blood pressure with the probability parameter only varying across age groups. The outcomes z_1 and z_2 mean "hypertensive" and "normotensive", respectively. The second binomial process is a missing-data-generating process with z_{21} and z_{22} meaning "not missing" and "missing", respectively.

As the second process may be considered to randomly delete blood pressure records within a given age-sickness category, the percentage unknowns should be the same for employees that have outcome z_1 and for employees that have outcome z_2, implying $q_{i2}=q_{i3}$ for all classes i. This also implies that ψ, introduced in section 2.3, equals zero. This "simple-guess approach" is evaluated first. In this situation, it turns out for the first process that, besides a linear trend of the logit of the response probability on age category, the number of sickness absences (<4 or >3) in 1982, too, has explanatory power. The second process is assumed to be related only to the number of sickness absences. After fitting this model for the total process, a deviance of 16.22 with 11 DF was found, which indeed suggests a reasonable fit. Since independence applies, this deviance can easily be divided into two. For the first process this leads to a deviance of 3.52 with 5 DF and for the second process to a deviance of 12.70 with 6 DF, which latter result is not satisfactory. Various other alternatives for ψ instead of the simple- guess approach ($\psi = 0$) are to be evaluated. The results are presented in Table II.

TABLE II

Summary of several strategies towards the randomness of the underlying missing-data-generating process

Strategy name	Second process	Estimated prevalence of hypertension	Deviance	DF
simple guess	$q_{i2}=q_{i3}$	14.3	16.22	11
lower bound	$q_{i2}=1$	10.6	15.73	11
upper bound	$q_{i3}=1$	36.5	81.44	11
optimal guess	ψ estimated (0.91)	12.2	14.97	10

It seems that in this example the simple guess, the lower bound and the optimal guess are nearly equivalent strategies. However, the upper bound strategy with $\psi = -\infty$ is highly unlikely ($p<<0.001$). The optimal guess estimate for the prevalence is 12.2 per cent. In Figure 2 the deviance (as a function of the estimated prevalence) of the different strategies is interpolated; it can be seen that the 95 per cent tolerance region for the prevalence is from 10.6 to about 16 per cent, including both the lower bound estimate and the simple guess estimate.

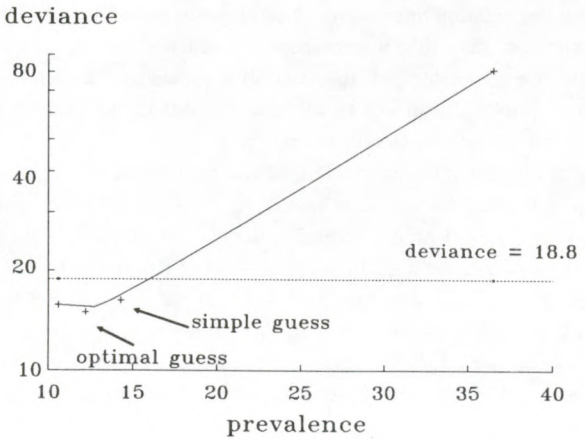

Figure 2 Deviance as a function of estimated prevalence of hypertension.

4 MODELLING A RECATEGORIZATION PROCESS

If the second process is a recategorizing process, then this process, too, can be described in terms of the general framework presented in this paper. This will be shown in this section.

If the second process is a recategorizing mechanism, in which z_{21} means "correctly reproduced" and in which z_{22} means "falsely reproduced", it is easy to see that in fact one can only observe the following combined counts: (y_1+y_4) and (y_2+y_3).

When the Poisson reparametrization is used, the composite link matrix C is:

$$C = \begin{bmatrix} 1001 \\ 0110 \end{bmatrix}$$

If the second process recategorizes the outcomes of the first process, direct estimation by binomial algorithms is possible if $q_2=1-q_3$ (or $\beta_2=-\beta_3$). In this case only the parameters β_2 can be estimated. However, for a recategorizing process this condition is very unrealistic. Moreover, one is interested in β_1 rather than β_2.

5 DISCUSSION

This paper presents a comprehensive approach to tackling the situation where two processes are observed and interest lies in only one, because the other process is a "manipulating" process. Additional information in the form of explanatory variables may be incorporated into the approach, provided that this information is available for all units. An algorithm is presented, applying the composite link approach. It is shown that sometimes it is also possible to apply standard binomial algorithms to calculate the MLEs. Of course, the EM algorithm (Dempster, Laird and Rubin, 1977) can be used as well.

By means of an example the problem of testing various strategies of "correcting" for missing data is presented. The choice is based on a likelihood ratio testing procedure by incorporating in a generalized model several strategies, of which the "simple-guess" strategy relates to the model-based direct-adjustment procedure of Rosenbaum (1987). Closely related, too, is the method of Conn, Lui and McGee (1989). These authors deal with the problem of estimating the incidence of home injury deaths in a situation where the place of occurrence is often unspecified. They use a logistic regression to estimate the probability of having a home injury and use this function in a one-time imputation phase for estimating the incidence of home injury deaths.

One should be aware that the appropriateness of the model can never be proved by goodness-of-fit testing. It can only be proved by resampling within the missing-data subgroup. However, the assumption, stated in the example, that the missing-data-generating process does not depend on age seems to be supported by the data from Table II. For this specific

generalization of the model, a likelihood ratio test was performed. The observed small age effects were not statistically significant. If the model specification had given a bad fit, also a super-binomial model could have been postulated and fitted with the computer program GLIM. For the present example, this did not seem necessary. In the example, the fitted value for ψ was 0.91. This means that there is a slight tendency towards an overrepresentation of normotensives within the missing-data group, so that the optimal-guess estimate for the prevalence goes toward the lower boundary. Additional data were available for only 51 per cent of the missing cohort. In this group, attending the periodic health examination in one of the following three years, the prevalence of hypertension turned out to be 11.1 per cent. Although this subsample, too, is no random subsample of the missing data cohort, this finding supports the parametric findings.

If the first process has more than two possible realizations, e.g., in misclassification problems of ordinal responses, the algorithm may be tedious to build, but the approach is essentially not different from the approach described in this paper. For the McCullagh (1980) models the composite link matrix C may become a band matrix because misclassification for ordinal data will most probably consist of a shift of one category away from the true category.

6 REFERENCES

Birch MB (1963). Maximum likelihood in three way contingency tables. *J. R. Statist. Soc. B*, 25, 220-33.

Clayton D (1985). Using test-retest reliability data to improve estimates of relative risk: an application of latent class analysis. *Statistics in Medicine*, 4, 445-455.

Conn JM, Lui KJ and McGee DL (1989). A model-based approach to the imputation of missing data: home injury incidences. *Statistics in Medicine* 8, 263-266.

Copas JB (1988). Binary regression models for contaminated data. *J. R. Statist. Soc. B*, 50, 225-265.

Dempster AP, Laird NW and Rubin DB (1977). Maximum likelihood for incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 39, 1-38.

Dinse GE (1986). Nonparametric prevalence and mortality estimators for animal experiments with incomplete cause-of-death data. *J. Am. Statist. Assoc.* 81, 328-336.

Johnson NL and Kotz S (1969). *Distributions in statistics.* Discrete distributions. Houghton Mifflin Company, Boston.

Little RJA and Rubin DB (1987). *Statistical analysis with missing data.* John Wiley & Sons, New York.

McCullagh P (1980). Regression models for ordinal data. *J. R. Statist. Soc. B*, 42, 109-142.

McCullagh P and Nelder JA (1983). *Generalized linear models.* Chapman and Hall, London.

Palmgren J (1981). The Fisher information matrix for log-linear models arguing conditionally on observed explanatory variables. *Biometrika*, 68, 563-566.

Rosenbaum PR (1987). Model-based direct adjustment. *J. Am. Statist. Ass.* 398, 387-394.

Rubin DB (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Steinhorst RK and Samuel MD (1989). Sightability adjustment methods for aerial surveys of wildlife populations. *Biometrics* 45, 415-425.

Thompson R and Baker RJ (1981). Composite link functions in generalized linear models. *Applied Statistics*, 30, 125-131.

Zeh JE, Ko D, Krogman BD and Sonntag R (1986). A multinomial model for estimating the size of a whale population from incomplete census data. *Biometrics* 42, 1-14.

ontvangen 19- 3 -1991
geaccepteerd 23- 12-1991