Statistical Modelling 2008; **8**(4): 385–401

Modelling general patterns of digit preference

Carlo G. Camarda[1], Paul H.C. Eilers[2,3] and Jutta Gampe[1]
[1]Max Planck Institute for Demographic Research, Rostock, Germany
[2]Methodology and Statistics, Faculty of Social and Behavioural Sciences, Utrecht University, The Netherlands
[3]Data Theory Group, Leiden University, The Netherlands

Abstract: In many applications data can be interpreted as indirect observations of a latent distribution. A typical example is the phenomenon known as digit preference, i.e. the tendency to round outcomes to pleasing digits. The composite link model (CLM) is a useful framework to uncover such latent distributions. Moreover, when applied to data showing digit preferences, this approach allows estimation of the proportions of counts that were transferred to neighbouring digits. As the estimating equations generally are singular or severely ill-conditioned, we impose smoothness assumptions on the latent distribution and penalize the likelihood function. To estimate the misreported proportions, we use a weighted least-squares regression with an added L_1 penalty. The optimal smoothing parameters are found by minimizing the Akaike's information Criterion (AIC). The approach is verified by a simulation study and several applications are presented.

Key words: composite link model; digit preference; L_1 penalty; penalized likelihood; smoothing

Received May 2007; **revised** September 2007; **accepted** October 2007

1 Introduction

When people read an analog scale or report numeric results, a commonly found effect is that certain preferred end-digits are reported substantially more often than the general pattern of the distribution suggests. These digits are typically multiples of 5 and 10, possibly combined with tendencies to avoid certain unpleasant numbers like, e.g. 13. This type of misreporting leads to unusual heapings at the preferred digits and the observed data actually present a biased, though well-understood image of the true distribution. This tendency is called digit preference or age heaping, if the reported numbers refer to ages.

Different techniques have been developed in various fields to deal with this problem. Digit preference is most likely to be seen whenever laymen are involved. Hence age misreporting has long been an issue in demography (Myers, 1940; Das Gupta, 1975; Coale and Li, 1991; Siegel and Swanson, 2004). Suggested

Address for correspondence: Carlo G Camarda, Max Planck Institute for Demographic Research, Konrad-Zuse-Strasse 1, 18057 Rostock, Germany. E-mail: camarda@demogr.mpg.de

10.1177/1471082X0800800404

solutions to compensate for age heaping are the application of summary indices to quantify the extent of misreporting and ad hoc procedures to reduce digit preferences and adjust age distributions. Mari Bhat (1990) proposed a model to estimate transition probabilities of age misstatement based on iterative adjustments and generalized stable population relationships.

Other self-reported information like height and weight (Rowland, 1990), year of menopause (Crawford *et al*., 2002) or retrospective studies in fecundability (Ridout and Morgan, 1991; Pickering, 1992) and breast-feeding duration are the typical examples in epidemiology, where digit preference is obvious. But professionals are also prone to heaping of certain numbers, as is, for instance, demonstrated in blood measurement readings (Hessel, 1986; Canner *et al*., 1991; Wen *et al*., 1993; Bennett, 1994).

Besides the quantification of digit preferences and the assessment of their consequences, only a few studies aim to model, estimate and correct the process of misreporting (Heitjan and Rubin, 1990; Ridout and Morgan, 1991; Pickering, 1992; Crawford *et al*., 2002). In this paper, we expand the idea of Eilers and Borgdorff (2004), and propose a general modelling technique to estimate the unobserved latent distribution, free of the effects of misreporting. Additionally, we provide estimates for the misreported proportions. This problem can be viewed as an inverse problem, where the actually observed values are linear compositions of a latent sequence representing the true distribution. This sequence is to be estimated and the composition pattern reveals the amount of misreporting. The composite link model (CLM), proposed by Thompson and Baker (1981), provides an elegant framework for modelling indirect observations of counts. It is an extension of the generalized linear model (GLM) (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989), and can itself be easily extended to allow for smooth predictors by incorporating a penalty on the roughness of the parameter vector (Eilers, 2007). Such smoothness assumptions allow us to solve an otherwise under-determined problem.

The paper is structured as follows. A typical example of age heaping will set the stage in the following section, after which the essence of the CLM is introduced in Section 3, including the specific form of the composition matrix. Estimation of the model is covered in Section 4, including difference penalties for assuring smoothness, the estimation of the preference pattern and the choice of optimal smoothing parameters. In Section 5, we illustrate the approach via simulated data and present some applications. A critical discussion of the method concludes the paper.

2 An example of digit preference

As an example for manifest digit preference Figure 1 shows the age distribution of adult Portuguese females (ages 30–89), who died during 1940 (Instituto Nacional de

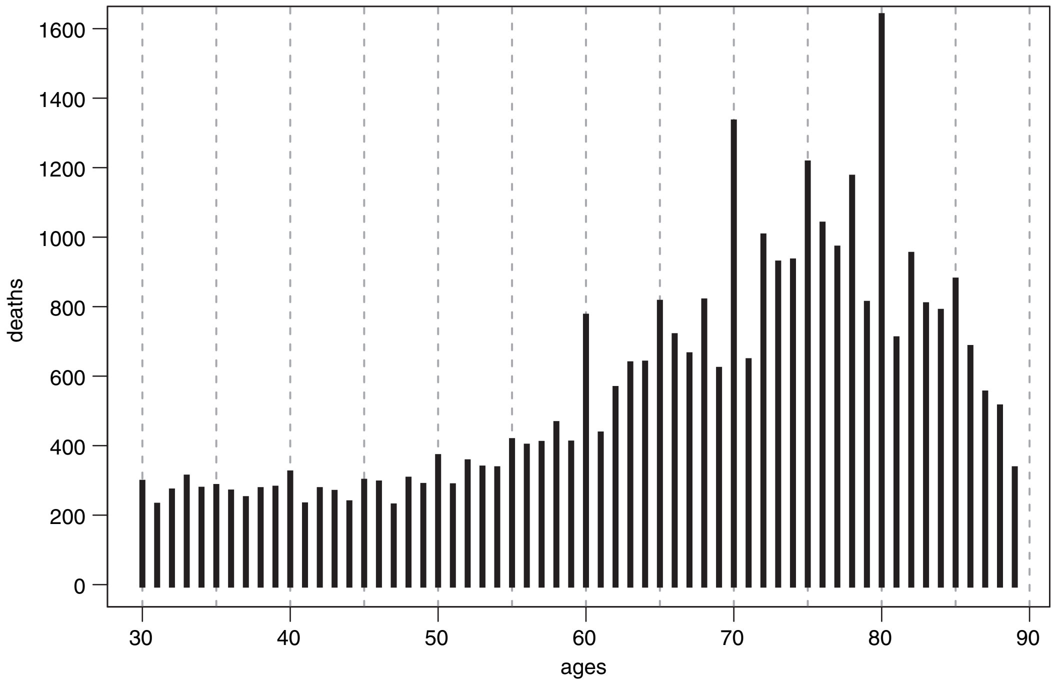

Figure 1 Age-at-death distribution for Portugal, Females, 1940.
Source: Instituto Nacional de Estatística, 1941.

Estatística, 1941). Systematic peaks at ages ending in 0 and, less prominently, 5 are typical features for countries with less accurate vital registration, which certainly was the case in Portugal almost seven decades ago.

Flanking the peaks, troughs are found at ages ending in 9, 1, 4 and 6. Moreover, this phenomenon seems particularly severe at older ages. Also even numbers in general seem to be preferred over odd digits.

Current age-at-death distributions are the result of the number of births and deaths, and migration flows in the past, and individual years may show particular outcomes, like epidemics, when birth cohorts are considerably smaller than the years before and after the crisis, or years of armed conflicts, when deaths are higher, especially among men. Thus there is the possibility of irregularities in an age distribution; however, the specific reasons for such irregularities are usually well understood from the historic records. In the absence of such specific past events, the assumption of a smooth age distribution is reasonable, implying that the peaks and gaps are the result of certain preferences in reported ages. If spikes or troughs in the distribution are due to events in the past, rather than digit preference, these digits will be excluded from the smoothing procedure.

The observed frequencies, therefore, can be viewed as the outcome of a misreporting process that transforms a smooth, but latent age distribution into observed data. The counts at the preferred digits are composed of the actual values at these ages plus the misclassified cases from the neighbouring categories due to the prevalent preference pattern.

3 The composite link model

We assume a smooth discrete sequence $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)'$, which is the unknown latent distribution. To ensure non-negative elements of $\boldsymbol{\gamma}$, we denote this sequence as $\boldsymbol{\gamma} = \exp(\boldsymbol{\beta})$, with $\boldsymbol{\beta}$ smooth, that is neighbouring elements of $\boldsymbol{\beta}$ being of similar size. The elements γ_j, $j = 1, \ldots, J$ are the counts that would be expected, if there were no digit preferences. However, the mechanism, which actually generates observations, operates by linearly composing the values in $\boldsymbol{\gamma}$ to a vector $\boldsymbol{\mu} = \boldsymbol{C}\boldsymbol{\gamma}$. The observed counts $\boldsymbol{y}$ are realizations from Poisson variables with $E(\boldsymbol{y}) = \boldsymbol{\mu}$, i.e.

$$P(y_j) = \frac{\mu_j^{y_j}\, \mathrm{e}^{-\mu_j}}{y_j!}\,. \tag{3.1}$$

The composition matrix $\boldsymbol{C}$ embodies the digit preference mechanism by partly redistributing certain elements of $\boldsymbol{\gamma}$ to neighbouring, preferred values in $\boldsymbol{\mu}$. In a general CLM the composition matrix $\boldsymbol{C}$ needs not be a square matrix as several categories could be lumped together. In our application, because expected counts are redistributed only partly, the matrix $\boldsymbol{C}$ is of dimension $J \times J$.

3.1 The composition matrix $\boldsymbol{C}$

The composition matrix $\boldsymbol{C}$ describes how the latent distribution $\boldsymbol{\gamma}$ was mixed before generating the data, and it is characteristic for the predominant preference pattern. Consequently, for modelling digit preferences, we have to define the matrix $\boldsymbol{C}$ according to our assumptions of the misreporting process. Eilers and Borgdorff (2004) allowed misreporting only for a few selected digits, with probabilities that did not change with the size of the underlying number, e.g. the probability for a transfer from $10x + 7$ to $10x + 8$ was assumed to be the same for all $x \in \mathbb{N}_0$.

In this paper we will assume that misreporting will only move observations to the immediate neighbouring digits, both to the left and the right. For example, observations are allowed to move from 9 to 10, but also from 9 to 8. We will not, however, consider preferences that move observations by two or more steps. For instance, we do not assume that observations get shifted from 8 to 10 nor from 12 to 10.

We denote by p_{jk} the proportion of γ_k that is moved from category k to category j. Allowing only one-step transitions implies that $p_{jk} = 0$ for $|j - k| > 1$. If we summarize these proportions in the $J \times J$ composition matrix $\boldsymbol{C}$, we obtain

$$\boldsymbol{C} = \begin{pmatrix} 1-p_{21} & p_{12} & 0 & 0 & \cdots & 0 \\ p_{21} & 1-p_{12}-p_{32} & p_{23} & \cdots & & \vdots \\ 0 & p_{32} & 1-p_{23}-p_{43} & p_{34} & \cdots & \vdots \\ 0 & 0 & p_{43} & \ddots & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & 1-p_{J-2,J-1}-p_{J,J-1} & p_{J-1,J} \\ 0 & \cdots & \cdots & 0 & p_{J,J-1} & 1-p_{J-1,J} \end{pmatrix} \tag{3.2}$$

The diagonal elements $c_{jj} = 1 - p_{j-1,j} - p_{j+1,j}$ of $\boldsymbol{C}$ specify the proportions of the γ_j that do not get redistributed. Note that all columns in $\boldsymbol{C}$ sum up to 1. In case we want to exclude a digit from this redistribution process, because we want to keep specific non-smooth features attributable to known mechanisms, we can adapt the matrix $\boldsymbol{C}$ at this very position.

It is obvious that the $2 \cdot (J-1)$ unknown elements p_{jk} cannot be estimated without imposing additional restrictions. We will estimate them via a penalized weighted least-squares approach, which will be discussed in detail in Section 4.2.

4 Estimating the CLM and the preference pattern

4.1 The CLM for a smooth latent distribution

Thompson and Baker (1981) present the CLM and the estimation algorithm very succinctly, and Eilers (2007) extended the approach to smooth latent distributions estimated by penalized likelihood. For easier reference, we describe the most crucial steps in the Poisson context here.

In case of no digit preference, we would be able to directly observe counts z_j, $j = 1, \ldots, J$, following a Poisson distribution such that

$$P(z_j) = \frac{\gamma_j^{z_j} \mathrm{e}^{-\gamma_j}}{z_j!}.$$

In our applications $\gamma_j = \exp\{\beta_j\}$, and smoothness of $\boldsymbol{\beta}$ immediately implies smoothness of $\boldsymbol{\gamma}$. In case we want to model a flexible functional dependence of the latent means $\boldsymbol{\gamma}$ on some covariate ν, we would expand this function into a B-spline basis. This leads to the more general formulation $\boldsymbol{\gamma} = \exp\{\boldsymbol{X\beta}\}$, where the design matrix $\boldsymbol{X}$ contains the basis elements covering the range of z, and the vector $\boldsymbol{\beta}$ gives the weights by which the individual B-splines in the basis get multiplied. Again, smoothness of the vector $\boldsymbol{\beta}$ implies smoothness of $\boldsymbol{\gamma}$. In our applications $\boldsymbol{X} = \boldsymbol{I}$, the identity matrix.

Estimates of the β_j in this GLM would be achieved by the iteratively reweighted least-squares (IWLS) algorithm, which in matrix notation is given by

$$\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X}\tilde{\boldsymbol{\beta}} = \boldsymbol{X}'\tilde{\boldsymbol{W}}\left\{\tilde{\boldsymbol{W}}^{-1}(\boldsymbol{z} - \tilde{\boldsymbol{\gamma}}) + \boldsymbol{X}\tilde{\boldsymbol{\beta}}\right\}, \tag{4.1}$$

where $\tilde{\boldsymbol{W}} = \text{diag}(\tilde{\boldsymbol{\gamma}})$. If we, however, do not observe $\boldsymbol{z}$, but realizations of the composed counts $\boldsymbol{y} \sim \text{Poisson}(\boldsymbol{\mu})$, with $\boldsymbol{\mu} = E(\boldsymbol{y}) = \boldsymbol{C}\boldsymbol{\gamma}$, or $\mu_i = \sum_j c_{ij}\gamma_j, i = 1, \ldots, J$, we can easily adapt the IWLS-scheme.

By defining $\breve{x}_{ik} = \sum_j c_{ij}x_{jk}\gamma_j/\mu_i$, the system of equations corresponding to (4.1) becomes, in matrix notation,

$$\breve{\boldsymbol{X}}'\tilde{\boldsymbol{W}}\breve{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} = \breve{\boldsymbol{X}}'\tilde{\boldsymbol{W}}\left\{\tilde{\boldsymbol{W}}^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) + \breve{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}\right\}, \tag{4.2}$$

where $\tilde{\boldsymbol{W}} = \text{diag}(\tilde{\boldsymbol{\mu}})$. A detailed derivation of (4.2) can be found in Eilers (2007).

When $\boldsymbol{X} = \boldsymbol{I}$, then $\ln(\boldsymbol{\gamma}) = \boldsymbol{\beta}$ and smoothness of $\boldsymbol{\beta}$ implies smoothness of $\boldsymbol{\gamma}$. In particular, the roughness of vector $\boldsymbol{\beta}$ can be measured with differences of order d, which can be written in matrix notation as

$$\boldsymbol{S}_d = \boldsymbol{\beta}'\boldsymbol{D}_d'\boldsymbol{D}_d\boldsymbol{\beta} = \|\boldsymbol{D}_d\boldsymbol{\beta}\|^2, \tag{4.3}$$

where $\boldsymbol{D}_d \in \mathbb{R}^{(K-d)\times K}$ is the matrix that computes d-th order differences. For $d = 1$ and $d = 2$ the $\boldsymbol{D}_d$ are

$$\boldsymbol{D}_1 = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{D}_2 = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \\ 0 & \cdots & \cdots & 1 & -2 & 1 \end{bmatrix}.$$

Both in GLMs and CLMs we can force the solution vector $\boldsymbol{\beta}$ to be smooth by subtracting a roughness penalty from the log-likelihood L (Eilers and Marx, 1996). This penalty is the roughness measure (4.3) weighted by the smoothing parameter λ:

$$L^* = L - \frac{\lambda}{2}\|\boldsymbol{D}_d\boldsymbol{\beta}\|^2 .$$

If we introduce this penalty into the likelihood for the CLM, we obtain the following system of equations:

$$(\breve{\boldsymbol{X}}'\tilde{\boldsymbol{W}}\breve{\boldsymbol{X}} + \lambda\,\boldsymbol{D}'\boldsymbol{D})\tilde{\boldsymbol{\beta}} = \breve{\boldsymbol{X}}'\tilde{\boldsymbol{W}}\left\{\tilde{\boldsymbol{W}}^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) + \breve{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}\right\}. \tag{4.4}$$

The smoothing parameter λ balances model fidelity, as expressed by the log-likelihood L, and smoothness of the parameter estimates, as expressed by the penalty term. For a given value of λ equation (4.4) can be solved iteratively. Methods for optimal choice of λ will be discussed in Section 4.3.

4.2 Finding the misreporting proportions

In order to estimate the proportions p_{jk} of misreported counts in the matrix $\boldsymbol{C}$ (cf. (3.2)), we solve a constrained weighted least-squares regression within the IWLS procedure. From the structure of the composition matrix $\boldsymbol{C}$ in equation (3.2) we see that we can write

$$\boldsymbol{\mu} = \boldsymbol{C\gamma} = \boldsymbol{\gamma} + \boldsymbol{\Gamma p}, \tag{4.5}$$

where $\boldsymbol{p} = (p_{12}, p_{23}, \ldots, p_{J-1,J}; p_{21}, \ldots, p_{J,J-1})'$, the left-to-right and the right-to-left transfer probabilities concatenated into a vector of length $2 \cdot (J - 1)$. Correspondingly, the $J \times 2 \cdot (J-1)$-matrix $\boldsymbol{\Gamma}$ is

$$\boldsymbol{\Gamma} = \begin{pmatrix} \gamma_2 & 0 & \cdots & 0 & -\gamma_1 & 0 & \cdots & 0 \\ -\gamma_2 & \gamma_3 & & \vdots & \gamma_1 & -\gamma_2 & & \vdots \\ 0 & -\gamma_3 & \ddots & 0 & 0 & \gamma_2 & \ddots & 0 \\ \vdots & & \ddots & \gamma_J & \vdots & & \ddots & -\gamma_{J-1} \\ 0 & \cdots & 0 & -\gamma_J & 0 & \cdots & \cdots & \gamma_{J-1} \end{pmatrix}.$$

Since $\boldsymbol{y} \sim \text{Poisson}(\boldsymbol{\mu})$, we can approximate the distribution of $(\boldsymbol{y} - \boldsymbol{\gamma})$ as

$$(\boldsymbol{y} - \boldsymbol{\gamma}) \text{ distributed approximately as } N(\boldsymbol{\Gamma p}, \text{diag}(\boldsymbol{\mu})). \tag{4.6}$$

As the number of unknowns in $\boldsymbol{p}$, namely $2 \cdot (J - 1)$, is considerably larger than the number J of available data points, additional restrictions have to be imposed on $\boldsymbol{p}$.

Our first attempt was to add a simple ridge penalty to the least-squares problem (4.6), but this did not lead to satisfactory results. As a ridge term penalizes the squared norm $\boldsymbol{p}'\boldsymbol{p}$ of the coefficient vector $\boldsymbol{p}$, the resulting estimates tended to have elements of similar sizes, which is unlike what we would expect for digit preference patterns. We rather would suspect that particular digits attract observations while for others the respective p_{jk} should be close to zero. Therefore, instead of the L_2 norm $\boldsymbol{p}'\boldsymbol{p}$, we introduce an L_1 penalty into the weighted least-squares problem (4.6). As pointed out by Tibshirani (1996), this penalty tends to select a small number of elements p_{jk} that exhibit the strongest effects, while possibly shrinking some others to zero. Our penalty thus is $\kappa \sum |p_j|$.

We now have to face the task of optimizing a goal which contains a quadratic term and a sum of absolute values. The latter complicates numerical optimization.

We like to avoid quadratic programming or other methods that move away from the (iterative) least-squares, because we need a well-defined effective dimension to be able to compute AIC. Therefore we follow the proposal of Schlossmacher (1973) and write $\sum_j |p_j| = \sum p_j^2/|p_j|$, turning a sum of absolute values into a weighted sum of squares. Of course, to compute the weights, ones needs to know p. This problem is solved by iteration, using weights $1/|\tilde{p}_j|$, where $\tilde{p}$ is an approximation to the solution.

We iteratively solve the following system of equations:

$$(\boldsymbol{\Gamma}'\boldsymbol{V}\boldsymbol{\Gamma} + \kappa\tilde{\boldsymbol{Q}})\boldsymbol{p} = \boldsymbol{\Gamma}'\boldsymbol{V}(\boldsymbol{y} - \boldsymbol{\gamma}), \tag{4.7}$$

where $\boldsymbol{V} = \mathrm{diag}(1/\boldsymbol{\mu})$ and the matrix $\boldsymbol{Q}$ is

$$\boldsymbol{Q} = \begin{pmatrix} \frac{1}{|p_{12}|+\epsilon} & 0 & . & . & . & . & . & . \\ 0 & \frac{1}{|p_{23}|+\epsilon} & 0 & . & . & . & . & . \\ . & 0 & \ddots & 0 & . & . & . & . \\ . & . & 0 & \frac{1}{|p_{j-1,j}|+\epsilon} & 0 & . & . & . \\ . & . & . & 0 & \frac{1}{|p_{21}|+\epsilon} & 0 & . & . \\ . & . & . & . & 0 & \frac{1}{|p_{32}|+\epsilon} & 0 & . \\ . & . & . & . & . & 0 & \ddots & 0 \\ . & . & . & . & . & . & 0 & \frac{1}{|p_{J,J-1}|+\epsilon} \end{pmatrix}.$$

A small number ϵ is introduced to prevent numerical instabilities when elements of $\boldsymbol{p}$ become very small. In our experience $\epsilon = 10^{-6}$ worked well.

The additional parameter κ in (4.7) constrains the size of misreporting proportions p_{jk} and has to be estimated similarly to the smoothing parameter λ (see Section 4.3). In practice, we alternately estimate $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ for a few iterations before we start updating $\boldsymbol{p}$ from (4.7).

4.3 Optimal smoothing

The estimating equations for the penalized CLM in (4.4) and for the preference pattern (4.7) depend on the combination of the two smoothing parameters λ and κ. Once λ and κ are fixed, the estimates $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{p}}$ are determined. To choose the optimal (λ, κ)-combination we minimize AIC:

$$\mathrm{AIC}(\lambda, \kappa) = \mathrm{Dev}(\boldsymbol{y}|\boldsymbol{\mu}) + 2\,\mathrm{ED}. \tag{4.8}$$

Dev($\boldsymbol{y}|\boldsymbol{\mu}$) is the deviance of the Poisson model (3.1), and ED is the effective dimension of the model for given (λ, κ). We chose the ED as the sum of the two model components, i.e. ED $=$ ED$_1$ $+$ ED$_2$, where ED$_1$ denotes the effective dimension of the penalized CLM, and ED$_2$ refers to the penalized WLS-regression. Specifically, we have

$$\text{ED}_1 = \text{trace}\{\breve{\boldsymbol{X}}(\breve{\boldsymbol{X}}'\boldsymbol{W}\breve{\boldsymbol{X}} + \lambda\boldsymbol{P})^{-1}(\breve{\boldsymbol{X}}'\boldsymbol{W})\}$$

and

$$\text{ED}_2 = \text{trace}\{\boldsymbol{\Gamma}(\boldsymbol{\Gamma}'\boldsymbol{V}\boldsymbol{\Gamma} + \kappa\boldsymbol{Q})^{-1}\,\boldsymbol{\Gamma}'\boldsymbol{V}\}\,. \tag{4.9}$$

An efficient 2D grid-search for λ and κ is adequate to find the minimum of the AIC (see Figure 3). Both IWLS iterations and penalized WLS were implemented in R (R Development Core Team, 2007) and the code is available from the first author.

5 Simulation and applications

5.1 Simulation study

To demonstrate the performance of our approach we applied it to several simulated scenarios. Figure 2 shows one possible true distribution, i.e. the vector $\boldsymbol{\gamma}$ together

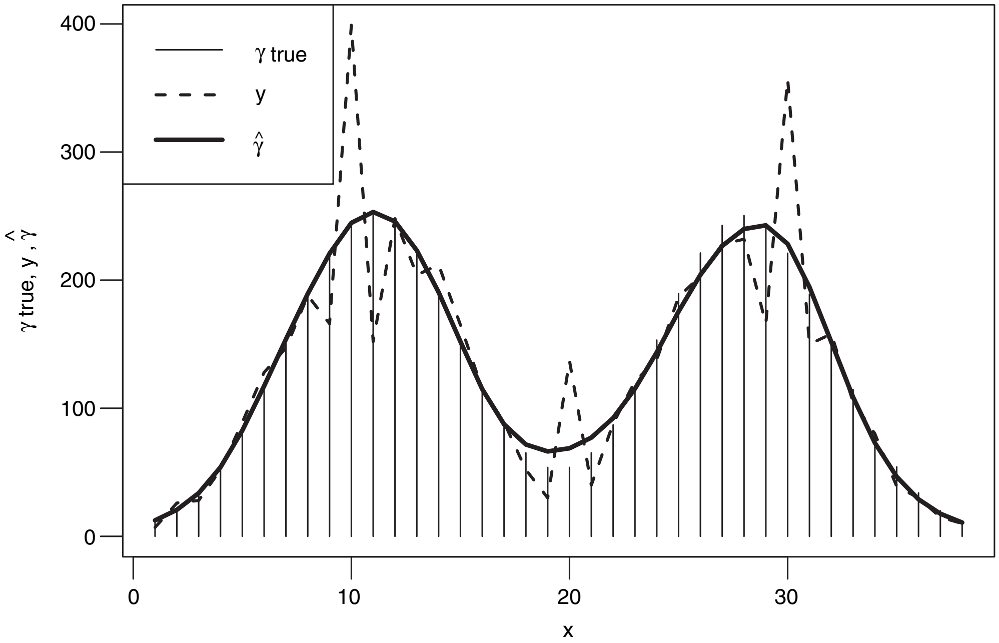

Figure 2 Raw data, true values and estimates for simulated data.

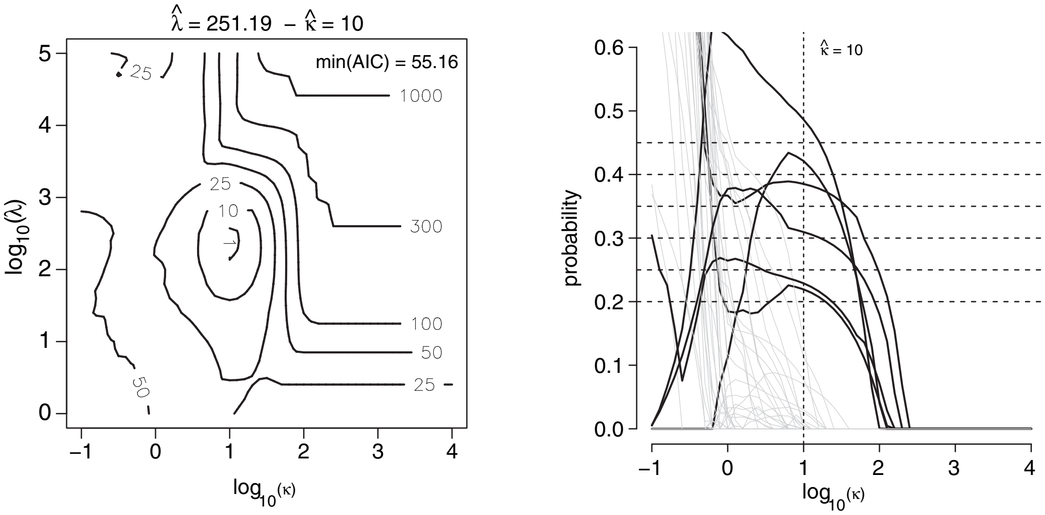

Figure 3 Left panel: AIC contour plot for the simulated data in Figure 2. Right panel: change of estimated misreporting probabilities with κ. The probabilities that are non-zero in the simulation are represented by thick black lines, the zero probabilities by thin gray lines.

with the simulated $\boldsymbol{y}$ such that $E(\boldsymbol{y}) = \boldsymbol{\mu} = \boldsymbol{C\gamma}$ and the estimated values $\hat{\boldsymbol{\gamma}}$. The assumed digit preference in this example attracted additional observations to 10, 20 and 30, from both neighbouring categories. These estimates were obtained from the optimal combination of (λ, κ) as picked from the AIC-profile shown in Figure 3, left image. The image on the the right-hand side demonstrates the effect of the L_1 penalty. On the horizontal axis the value of $\log \kappa$, i.e. the weight of the L_1 penalty, is given. For big values of κ all proportions p_{jk} are shrunk to zero. For small values of κ most proportions are far too large, but for increasing values of κ many of them are quickly damped down to zero, leaving the important ones in the model. The optimally chosen $\hat{\kappa}$ practically selects the true proportions, which are depicted by the horizontal dashed lines.

As pointed out in Section 4.2, the model actually estimates $2 \cdot (J - 1)$ misreporting proportions, which have not been restricted to be positive. A negative value of p_{jk} implies that category j receives a negative proportion of γ_k, that is, digit preference actually moves observations away from category j to k, but the amount is expressed as proportion of the receiving category k. This seemingly paradoxical behaviour is a consequence of the L_1 penalty; depending on whether $\gamma_j < \gamma_{j+1}$ or $\gamma_j > \gamma_{j+1}$, that is, whether the true distribution is increasing or decreasing at γ_j, the same preference leads to a smaller L_1 penalty when expressed via $p_{j,j+1}$ or $p_{j+1,j}$, one of them being negative.

Nevertheless, we would like to see as final results the net proportions as positive numbers. This can be easily achieved by the following transformation, which converts $2 \cdot (J-1)$ parameters to $J-1$ positive proportions:

$$\begin{aligned} &\text{if} \quad p_{j,j-1} < 0 \Rightarrow p_{j-1,j} = -\frac{\delta_j}{\gamma_j} \quad \text{and} \quad p_{j,j-1} = 0, \qquad (5.1)\\ &\text{if} \quad p_{j-1,j} < 0 \Rightarrow p_{j,j-1} = \frac{\delta_j}{\gamma_{j+1}} \quad \text{and} \quad p_{j-1,j} = 0, \end{aligned}$$

for $j = 2, \ldots, J-2$ and where $\delta_j = \mu_j - \gamma_j + \gamma_j \cdot c_{j-1,j} - \gamma_{j-1} \cdot c_{j,j-1}$. This procedure is simplified for the first step $j = 1$ and the last $j = J-1$.

The right image in Figure 3 shows these transformed and hence positive estimates. Additionally, Figure 4 summarizes true and estimated misreporting probabilities for the simulation example.

Our model allows for flexible shapes of the latent distribution, but the assumption that observations get redistributed to immediate neighbours only may be oversimplistic. To study the effect that more general patterns of digits preference may have, we modified our simulation setting as shown in Figure 5. In this scenario the digits 10 and 20 attract observations both from their next neighbours and also from digits that are two steps away (that is 8, 12, 18 and 22, respectively). Still γ and the p_{jk} were estimated based on the simpler model. As can be seen from Figure 5,

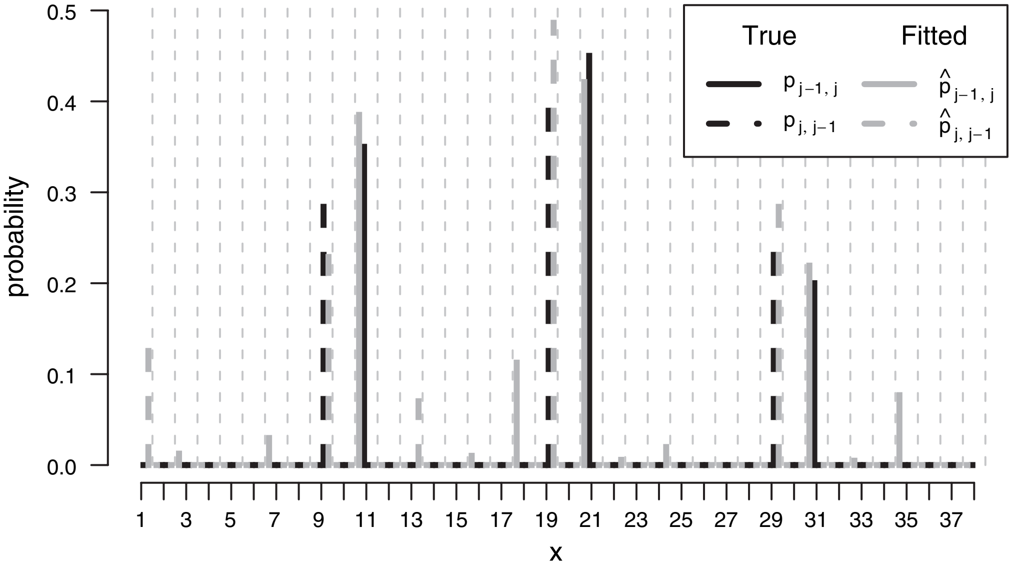

Figure 4 True misreporting probabilities and estimates for simulated data.

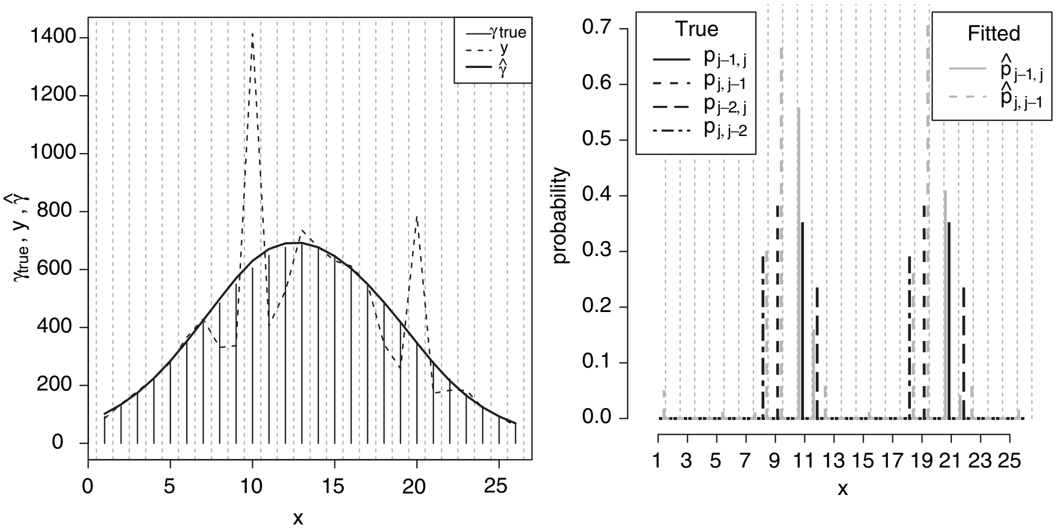

Figure 5 Results from the second simulation setting. Raw data, true values and estimates (left panel). True misreporting probabilities and estimates for simulated data (right panel).

left panel, the latent distribution $\hat{\gamma}$ is identified without problems. The two-step misreporting probabilities are reduced to two one-step components, as illustrated in the right panel of Figure 5. Instead of shifting the corresponding proportions in one sweep by two steps, which the model does not provide for, they get assigned to their next neighbours first; however, these proportions then get stacked on top of the one-step estimates to the preferred target digits (10 and 20 in this example). Hence the model 'decomposes' more complex preference patterns into subsequent simpler steps.

5.2 Portuguese ages at death

If we apply the model to the Portuguese age-at-death distribution introduced in Section 2, we obtain the results shown in Figure 6. The smooth fitted curve shows a smooth density without any age heapings. The AIC is clearly minimized for λ and κ equal to 10^4 and 15.85, respectively.

The misreporting probabilities are portrayed in Figure 7. As expected, digit preference mainly attracts observations to ages that end in 5 or 10, the latter ones showing the strongest effects. The amount of misreporting increases with age, and this fits well with the demographic experience that accurate age reporting is more problematic at the high ages. Also, for ages that are multiples of 10 there is a slightly higher tendency to receive counts from their respective right neighbours.

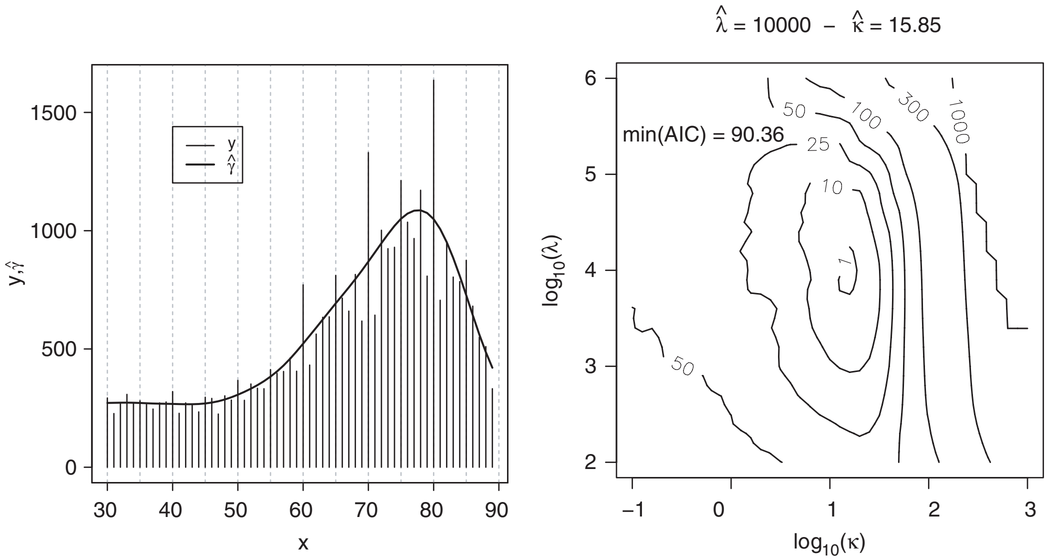

Figure 6 Results for the Portuguese data, cf. Figure 1. Observed and estimated distribution of age at death (left panel). AIC-contour plot (right panel).

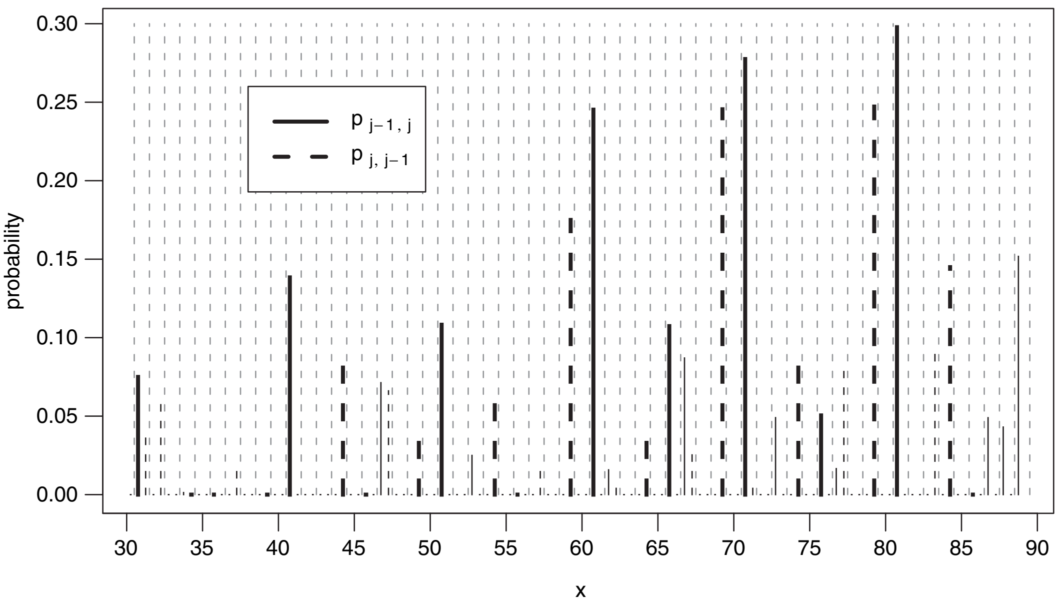

Figure 7 Misreporting probabilities for the Portuguese data, cf. Figure 1. Probabilities to digits multiples of 5 and 10 are depicted in thicker lines.

5.3 Weight data

The second example is taken from the National Health and Nutrition Examination Survey (NHANES) conducted by the US National Centre for Health Statistics (NHANES, 1980). The survey contains both the self-reported weight (in pounds) of $n = 11\,614$ women and men as well as measured weight during the examination.

Figure 8 shows the raw data of both weight variables and the fitted distribution, based on the self-reported data only. As can be seen, there is a large difference between the self-reported weights and the measured ones: people tend to round their weights to more pleasant digits, such as 0 and 5, resulting in a peculiar spiky shape to the distribution. Even though we expect that also the medical personnel doing the weight measurements will show some, though minor digit preference, we may treat the measured distribution as a proxy to the true one. Figure 8 demonstrates the close resemblance of the estimated latent distribution to the measured one, despite the severe preference pattern present in the original data.

The model was computationally quite intensive, since it fits $J = 204$ different weight-categories with $J + 2(J - 1) = 610$ parameters. Nevertheless, the penalties for the latent distribution and the misreporting probabilities worked properly in reducing the effective dimensions as well as capturing the actual weight distribution with an impressive precision. (Based on AIC, the best choice is $\log_{10} \lambda = 4.8$ and $\log_{10} \kappa = 0.2$.)

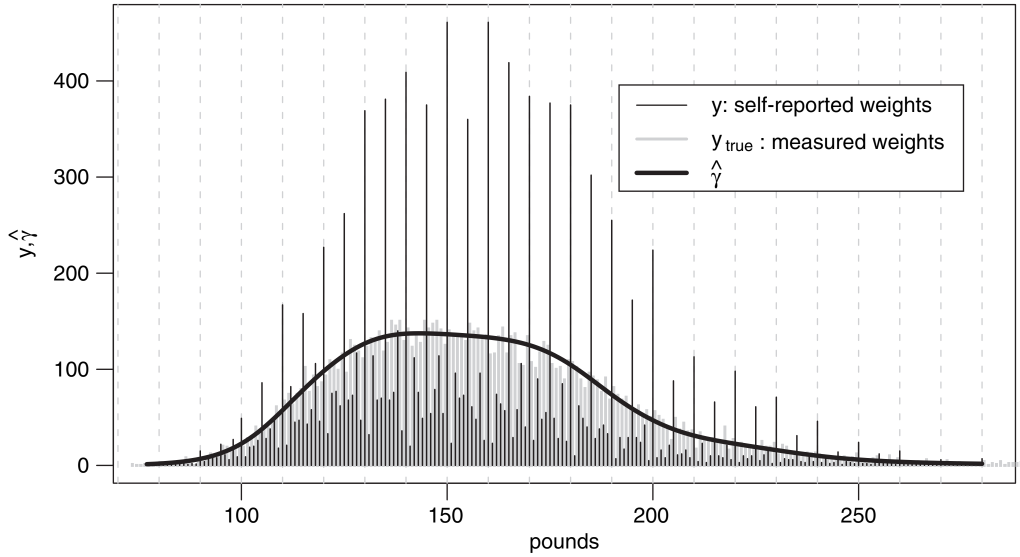

Figure 8 Self-reported and measured weight (in pounds) and fitted values for NHANES II data.

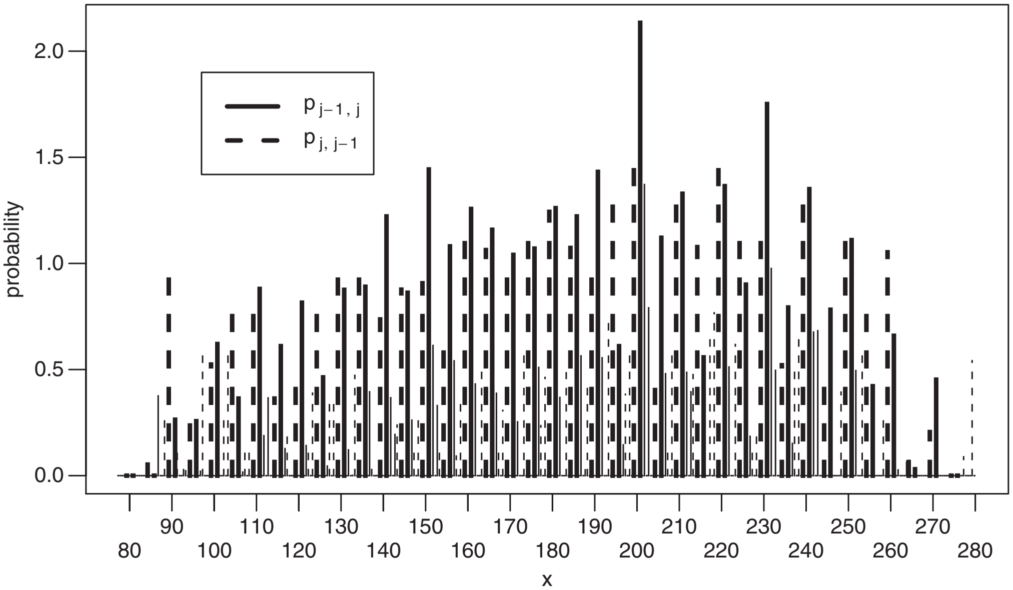

Figure 9 Misreporting probabilities for the NHANES II data. Probabilities to digits multiples of 5 and 10 are depicted in thicker lines.

Figure 9 shows the estimated misreporting pattern with expected outcomes: the probabilities detected by the model are practically only the ones to weight categories ending with 0 and 5 (depicted with thicker lines). In particular, weights of 150, 200 and 230 pounds are prone to receive counts from their neighbours, again slightly more from the right-hand category.

6 Discussion

The method we have presented in this paper demonstrates how digit preferences can be modelled by combining the CLM with the concept of penalized likelihood. The only assumption that is made about the underlying true distribution is smoothness. The approach directly addresses the process that leads to heaping of certain values. Extracting the latent distribution will be most important in many applications; however, the pattern of misclassification may also be of interest in itself. The proposed model, which goes beyond the mere quantification of digit preference provided by many indices, allows the analysis of both aspects.

The misreporting pattern was allowed to partly redistribute observations from any digit to its adjacent neighbours. Again a penalty, in this case a L_1 penalty, restrains the problem and makes estimation feasible. By allowing this rather flexible preference pattern the tendency to misreport need not be the same for identical end-digits, but may vary over the measurement range, which is often seen in real data.

As was demonstrated by the simulation study, more complicated transfer patterns still allow estimation of the latent distribution without problems. The misreporting probabilities over more than one digit are represented as contributions to several single-digit moves though, and more complex preference patterns are disguised by the model restrictions.

We currently are developing an extension of the presented model to include even more general patterns of misreporting, i.e. allow for exchanges between digits that are more than one category apart, which allows estimation of all different misclassification probabilities. Also one can envision digit preferences that improve over time, a phenomenon that is known for age reporting in demography and may also be seen in applications, where training or experience improves the quality of measurement readings over time. In this case, the transfer probabilities for different measurement occasions are expected to change smoothly. If longitudinal data are available, this trend can be handled by an additional penalty that controls the temporal pattern in the misreporting pattern.

In both extensions, additional smoothing parameters will have to be optimized so that faster algorithms to search for the optimal combination will be advisable.

Finally, we would like to point out that the way we handle a sum of absolute values in a penalty as a weighted sum of squares leads to an elegant and natural definition of the effective dimension. This seems relevant to a wide range of applications with L_1 penalties. Also this result is not dependent on the algorithm one uses to find a solution, because in (4.9) the final result can be plugged in to compute Q.

Acknowledgement

The authors wish to thank Vladimir Canudas Romo for his assistance in obtaining the Portuguese data.

References

Bennett S (1994) Blood pressure measurement error: its effect on cross-sectional and trend analyses. *Journal of Clinical Epidemiology*, **47**, 293–301.

Canner PL, Borhani NO, Oberman A, Cutler J, *et al.* (1991) The hypertension prevention trial: assessment of the quality of blood pressure measurements. *American Journal of Epidemiology*, **134**, 379–92.

Coale AJ and Li S (1991) The effect of age misreporting in China on the calculation of mortality rates at very high ages. *Demography*, **28**, 293–301.

Crawford SL, Johannes CB and Stellato RK (2002) Assessment of digit preference in self-reported year at menopause: choice of an appropriate reference distribution. *American Journal of Epidemiology*, **156**, 676–83.

Das Gupta P (1975) A general method of correction for age misreporting in census populations. *Demography*, **12**, 303–12.

Edouard L and Senthilselvan A (1997) Observer error and birthweight: digit preference in recording. *Public Health*, **111**, 77–9.

Eilers PHC (2007) Ill-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling*, **7**, 239–54.

Eilers PHC and Borgdorff MW (2004) Modeling and correction of digit preference in tuberculin surveys. *International Journal of Tuberculosis and Lung Diseases*, **8**, 232–9.

Eilers PHC and Marx BD (1996) Flexible smoothing with *B*-splines and Penalties. *Statistical Science*, **11**, 89–121.

Heitjan DF and Rubin DB (1990) Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association*, **85**, 304–14.

Hessel PA (1986) Terminal digit preference in blood pressure measurements: effects on epidemiological associations. *International Journal of Epidemiology*, **15**, 122–5.

Instituto Nacional de Estatística (1941) Òbitos por idades, meses e sexos, 1940. [Deaths by age, months and sex, 1940]. In *Anuário Demográfico Ano de 1940*. Lisboa: Impresa Nacional. Tables available at http://www.ine.pt/.

Mari Bhat PN (1990) Estimating transition probabilities of age misstatement. *Demography*, **27**, 149–63.

McCullagh P and Nelder JA (1989) *Generalized linear models*, (2nd ed.). Monographs on Statistics Applied Probability, London: Chapman & Hall.

Myers RJ (1940) Errors and bias in the reporting of ages in census data. *Transactions of the Actuarial Society of America*, **41**(Pt. 2 (104)), 395–415.

Nelder JA and Wedderburn RWM (1972) Generalized linear models. *Journal of the Royal Statistical Society*, Series A, **135**, 370–84.

NHANES (1976–1980) *National Health and Nutrition Examination Survey (NHANES)*. US National Center for Health Statistics. Data available at www.cdc.gov/nchs/nhanes.htm.

Pickering R (1992) Digit preference in estimated gestational age. *Statistics in Medicine*, **11**, 1225–38.

R Development Core Team (2007) *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Ridout MS and Morgan BJT (1991) Modelling digit preference in fecundability studies. *Biometrics*, **47**, 1423–33.

Rowland ML (1990) Self-reported weight and height. *The American Journal of Clinical Nutrition*, **52**, 1125–33.

Schlossmacher EJ (1973) An iterative technique for absolute deviations curve fitting. *Journal of the American Statistical Association*, **68**, 857–65.

Siegel JS and Swanson DA (2004) *The methods and materials of demography* (2nd ed.). Amsterdam: Elsevier Academic Press.

Thompson R and Baker RJ (1981) Composite link functions in generalized linear models, *Applied Statistics*, **30**, 125–31.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, Series B, **58**, 267–88.

Wen SW, Kramer MS, Hoey J, Hanley JA *et al.* (1993) Terminal digit preference, random error, and bias in routine clinical measurement of blood pressure, *Journal of Clinical Epidemiology*, **46**, 1187–93.