Modeling Digit Preference by Penalized Composite Link Models

Extended Abstract

Carlo G. Camarda[†] Paul Eilers[⋆] Jutta Gampe[†]

[†]Max Planck Institute for Demographic Research, Rostock, Germany
[⋆]Department of Medical Statistics, Leiden University, The Netherlands

September 22, 2006

1 Introduction

Age heaping is the result of preferences for some digits in self-reported age over the adjacent digits, and measuring and correcting for this clustering at specific ages has long drawn the attention of demographers. More generally, digit preference can be found in many data sets, typical examples include self-reported age at menopause, self-reported height or weight, blood pressure measurements or number of cigarettes smoked per day (Canner et al., 1991; Crawford et al., 2002; Hessel, 1986; Klesges et al., 1995; Rowland, 1990).

Preferred digits usually are 0 and 5, but even numbers may also be preferred over odd numbers, or some numbers, like 13, may show a tendency to be avoided. Consequently digit preference leads to frequency distributions with unusual spikes at the preferred digits at the expense of the neighboring numbers.

All methods to quantify, and thereafter to correct for digit preference originate from the idea that "the figures for adjacent ages should presumably be rather similar" (Siegel and Swanson, 2004, p. 136). The most simple indexes of age preference compare the frequency of the preferred digit to what would be expected for a rectangular or linear distribution in some neighborhood of the target value. More complex approaches like Whipple's index or Myers' Blended Index (Myers, 1940) yield an index of preference for each terminal digit over a defined age range (Ewbank, 1981; Shryock et al., 1976; Siegel and Swanson, 2004).

In this paper we present a model which takes the idea that for the true distribution the counts for adjacent digits should be similar as a starting point. However, apart from this smoothness assumption no further restrictions are put on the true distribution. The observed frequency distribution results from the true, but unobserved one through misclassification of a certain proportion of the less attractive digits towards the preferred ones.

Following an idea of Eilers and Borgdorff (2004), we can put this process of age misstatement in the framework of a so called composite link model (Thompson and Baker, 1981) and introduce the idea of a smooth latent distribution via a penalization term in the likelihood. Thereby we may estimate both the latent distribution and the misclassification probabilities by a straightforward extension of a penalized iteratively reweighted least squares algorithm (McCullagh and Nelder, 1989).

In the following section we will introduce the basic methodology, referring more technical details to the appendix. Before applying our model to a demographic data set in section 4, we demonstrate its performance in a simulation study (see section 3). We conclude with an outlook to additional work in progress.

2 The Methodology

We denote by $\gamma = (\gamma_1, \ldots, \gamma_J)'$ a smooth discrete sequence of J counts, which is the expected value of the unknown latent age distribution. In a more general setting we could allow γ to depend on some covariates, however, here we define it as $\gamma = \exp(\boldsymbol{X}\beta)$, with β smooth and $\boldsymbol{X}$ simply the identity matrix.

The actually observed counts are denoted by $y = (y_1, \ldots, y_J)'$, which are realizations from a Poisson distribution with expected values $E(y_j) = \mu_j$. An additional matrix C 'composes' the vector μ from γ. In other words, the 'composition matrix' C describes how the latent distribution is mixed before generating the data, and it is characteristic for the process that generated the data. When modelling misreported age distributions, C is a matrix that 'adjusts' the expectation γ_j in order to get μ_j, which are the expected values of the actually observed counts.

2.1 Estimating the model

If we could observe the 'true' counts $z_j, j = 1, \ldots, J$, then the z_j would be distributed according to a Poisson model with smooth expectation γ_j:

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

and $\gamma = \exp\{\boldsymbol{X}\beta\}$, where the covariate matrix $\boldsymbol{X}$ simply represents the sequence of ages. Using a generalized linear model (GLM) approach (McCullagh and Nelder, 1989; Nelder and Wedderburn, 1972), we estimate the values of β with an iteratively reweighted least squares (IRWLS) algorithm. In matrix notation we solve the system of equations:

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

where $\tilde{\boldsymbol{W}} = \texttt{diag}(\tilde{\gamma})$.

Of course, in reality, we do not observe z, but the counts given by y, with $\boldsymbol{\mu} = E(y) = \boldsymbol{C\gamma}$, or $\mu_i = \sum_j c_{ij}\gamma_j$. Adapting the maximum likelihood equations leads to a modified version of the IRWLS equations, as is shown in Appendix A:

$$\breve{\boldsymbol{X}}'\tilde{\boldsymbol{W}}\breve{\boldsymbol{X}}\boldsymbol{\beta} = \breve{\boldsymbol{X}}'\tilde{\boldsymbol{W}}\{\tilde{\boldsymbol{W}}^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) + \breve{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}\}\,, \tag{2.1}$$

where $\tilde{\boldsymbol{W}} = \texttt{diag}(\tilde{\mu})$ and $\breve{\boldsymbol{X}}$ can be interpreted as a 'working X' and its elements are $\breve{x}_{ik} = \sum_j c_{ij}x_{jk}\gamma_j/\mu_i$.

2.1.1 Smoothness and Penalty

When $\boldsymbol{X}$ in $\ln(\boldsymbol{\gamma}) = \boldsymbol{\eta} = \boldsymbol{X\beta}$ is the identity matrix, it is clear that smoothness of $\boldsymbol{\beta}$ implies smoothness of $\boldsymbol{\gamma}$. In GLMs or CLMs we can force the solution vector $\boldsymbol{\beta}$ to be smooth by subtracting a roughness penalty from the log–likelihood (see Eilers and Marx, 1996):

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

where $\boldsymbol{D}_d$ is the matrix that computes the d–th differences. We similarly can proceed with the modified IRWLS, and the system of equations 2.1 simply becomes:

$$(\breve{\boldsymbol{X}}'\tilde{\boldsymbol{W}}\breve{\boldsymbol{X}} + \boldsymbol{P})\boldsymbol{\beta} = \breve{\boldsymbol{X}}'\tilde{\boldsymbol{W}}\{\tilde{\boldsymbol{W}}^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) + \breve{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}\} \tag{2.2}$$

where $\boldsymbol{P} = \lambda\boldsymbol{D}_d'\boldsymbol{D}_d$.

The additional parameter λ tunes the smoothness of the parameters $\boldsymbol{\beta}$. To choose the value of the smoothing parameter λ an appropriate information criterion, such as Akaike's Information Criterion (AIC) (see Hastie and Tibshirani (1990)) is minimized. Alternatively the Bayesian Information Criterion (BIC) can be minimized (Schwarz, 1978). See Appendix B for more details.

3 Simulation study

Before we apply the model to a demographic dataset, we want to demonstrate the performance of the method in a small simulation study. The parameters in the study are chosen to mimic a realistic scenario.

The true latent distribution is designed to follow a Gompertz distribution, that is, the expected values γ of the true counts are derived from this Gompertz density (multiplied by a fixed sample size). The misreporting mechanism of this 'true' distribution is achieved with a given 'composition' matrix C that creates the expected values μ of the actually observed counts. The misreporting proportions are given in Table 1. From this composed distribution, the actual counts y are simulated as random numbers from a Poisson distribution with means μ.

Finally, based on these data, the Penalized Composite Link Model (PCLM) is estimated, as outlined in the previous section, leading to estimates of both γ and the misreporting proportions. An example of such simulation is given in Figure 1. Here the sample size is 10000 deaths from age 28 to 102. The structure of the transfer pattern is given in Table 1.

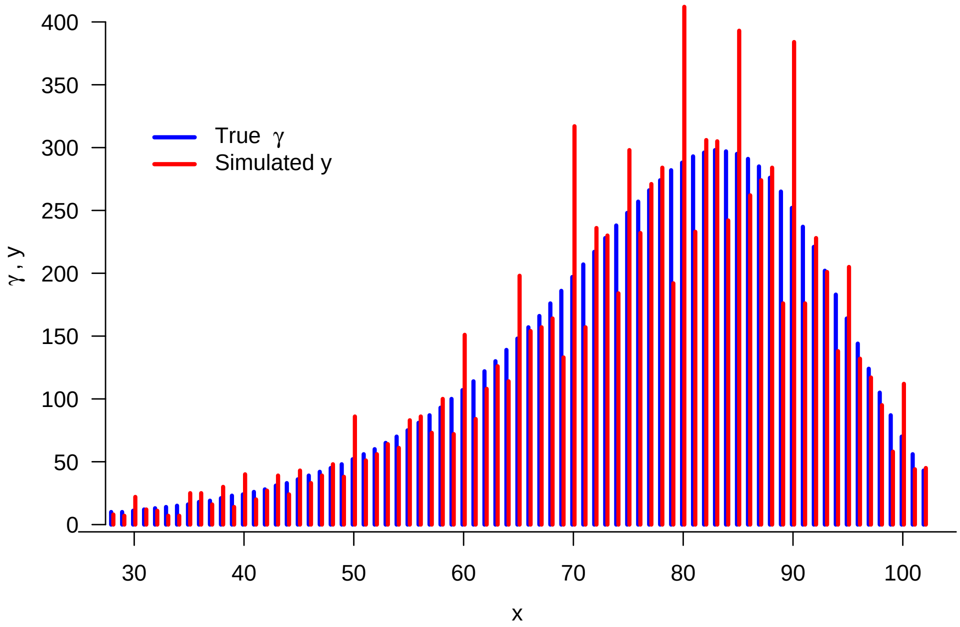

Figure 1: Example of simulated data set, assumptions are shown in Table 1

Transfer pattern	Proportions
from digit 9 to 0	0.3
from digit 0 to 1	-0.2
from digit 4 to 5	0.2
from digit 5 to 6	-0.1

Table 1: Choice of transfer patterns for the simulation study.

In the simulation study these steps were replicated 500 times, leading to 500 estimates of γ and of the proportions given in Table 1. Figure 2 shows the 'true' distribution of γ, the expected value for the misreported distribution μ, the median and the interval given by the 1%- and 99%-quantile of the 500 fitted distributions.

Figure 3 summarizes the fitted misreporting proportions for each transfer pattern[1]. Whereas

[1] In this figure a negative value has to be intended as proportion of counts at age j which has been misreported from $j+1$ to j.

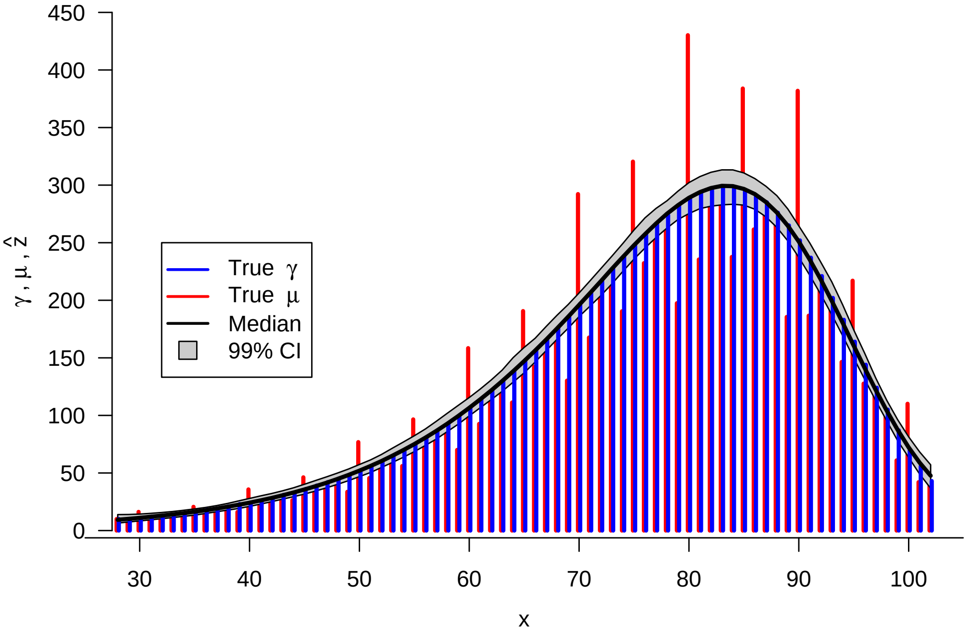

Figure 2: Summary of 500 simulations from a misreported age distribution. The 'true' distribution is depicted in blue; the red histogram shows the expected misreported distribution; the gray shadow presents the interval given by the 1%- and 99%-quantile of the fitted distributions. The black line present the median of the fitted distribution

the variability is larger where observations are fewer, the medians of the estimated proportions are always rather close to the true values.

4 Application to the population of the Philippines

Manifest digit preferences are found in the age distribution of the population of the Philippines by single years of age in 1960 (United Nations, 1962). These are census data and show systematic peaks at ages ending in 0 and, less prominently, 5. Correspondingly troughs are found at ages ending in 9, 1, 4 and 6. Moreover, particular heaping occurs at ages 12 and 18.

Shryock et al. (1976) have already used this example for measuring digit preferences with different indexes and Alho and Spencer (2005) used this example to show a possible problem with published population statistics.

Under the assumption that the actual age distribution is smooth, and that the heavy spikes are created by digit preferences we can apply the PCLM to estimate the true latent age distribution and the misreporting proportions.

Figure 4 shows the estimated latent age distribution and Figure 5 gives the estimated

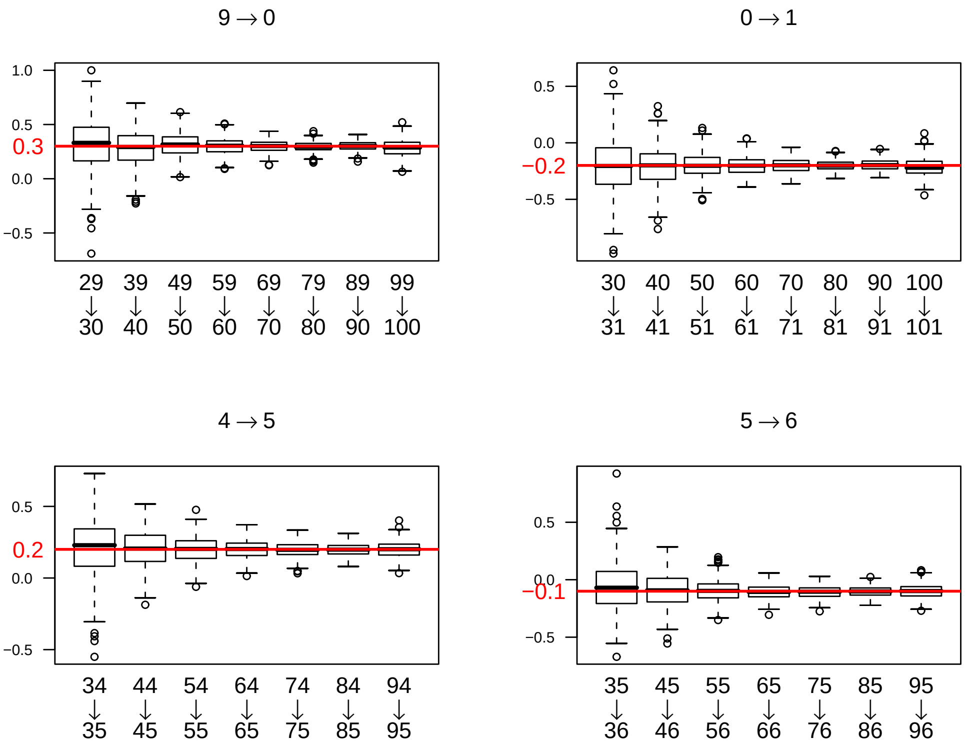

Figure 3: Summary of the misreporting proportions from 500 simulated distribution. Each panel presents a particular transfer pattern. The 'true' misreporting proportions given in the simulation are depicted in red.

misreporting pattern. Note that the misclassification increases with age, and it is relatively heavier for ages ending with 0 than with 5.

5 Outlook

The method we have presented in this paper shows how it is possible to deal with digit preferences, that result in age heaping, by combining the concept of penalized likelihood with the composite link model. The PCLM allows extraction of both the latent distribution and the pattern of misclassifications, which goes beyond the quantitative assessment of digit preferences provided by many indexes. The only assumption that is made about the underlying true distribution is smoothness.

In the current applications we specify a limited number of misclassification patterns, here mainly relating to preference for numbers ending in 0 and ending in 5, which receive contributions from their adjacent neighbors. By adding a further penalty, we are currently exploring the possibility of allowing for more general patterns of misstatement, to include both the potential for exchanges between digits which are immediate neighbors, as well as

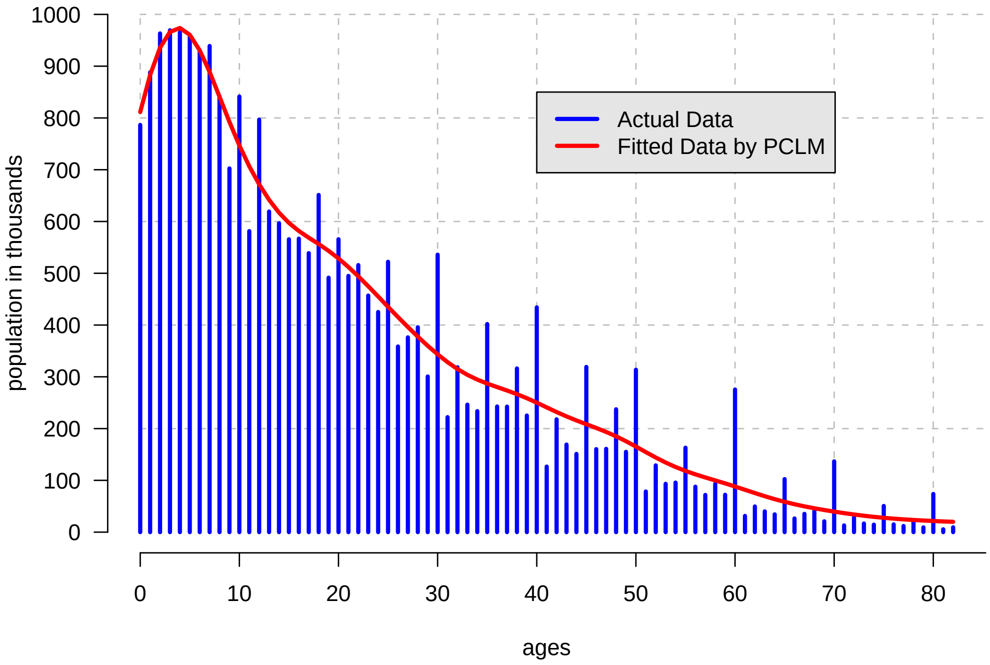

Figure 4: Distribution of the Philippines population by single age in 1960 and fitted distribution by Penalized Composite Link Model.

for exchanges between digits that are 2 steps apart (like e.g. age 78 misreported as 80 etc.).

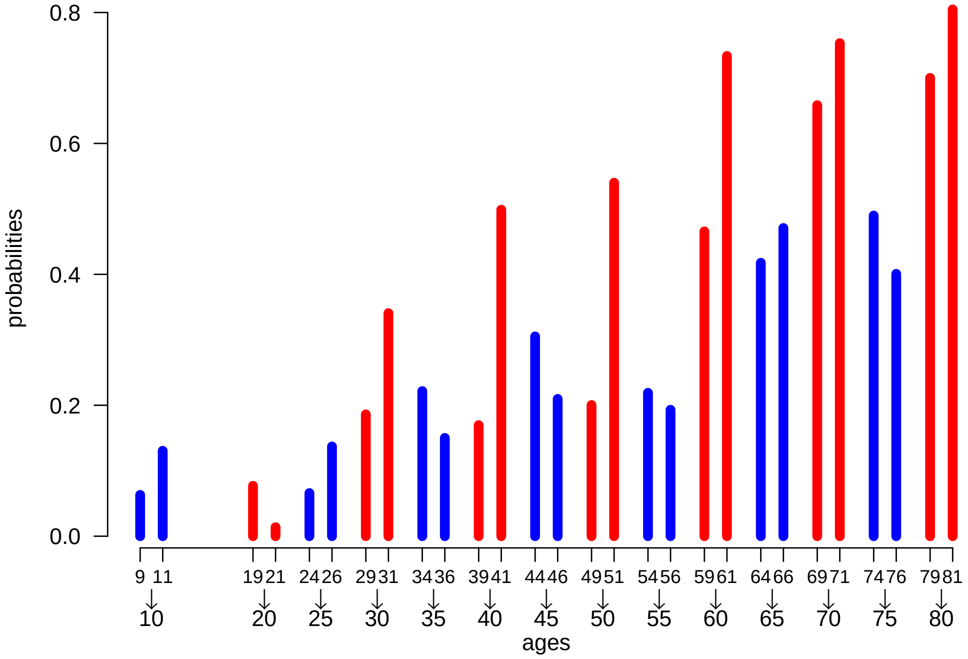

Figure 5: Misreporting probabilities from specific ages to others. Philippines, 1960.

Appendix A

Assume that z_j has a Poisson distribution with expectation γ_j and linear predictor $\eta_j = ln(\gamma_j) = \sum_k x_{jk}\beta_k$ with the logarithm as link function. The method of maximum likelihood is used to estimate the parameters $\boldsymbol{\beta}$. McCullagh and Nelder (1989, equations 2.12 and 2.13) showed that in this case the ML equations are:

$$\sum_{j=1}^{J} \frac{(z_j - \gamma_j)}{v(\gamma_j)} \frac{\partial \gamma_j}{\partial \beta_k} = \sum_{j=1}^{J} (z_j - \gamma_j) x_{jk} = 0 \tag{A-1}$$

where $v(\gamma_j)$ is the variance when $E(z) = \gamma_j$. In the case of a CLM, the new Poisson distributed variable is y_i with expectation $\mu_i = E(y_i) = \sum_j c_{ij}\gamma_j$. Hence adapting equations A-1 we find that

$$\sum_{i=1}^{I} \frac{(y_i - \mu_i)}{v(\mu_i)} \frac{\partial \mu_i}{\partial \beta_k} = 0$$

Because

$$\frac{\partial \mu_i}{\partial \beta_k} = \sum_{j=1}^{J} c_{ij} \frac{\partial \gamma_j}{\partial \beta_k} = \sum_{j=1}^{J} c_{ij} x_{jk} \gamma_j$$

we get the likelihood equations

$$\sum_{i=1}^{I} (y_i - \mu_i) \breve{x}_{ik} = 0$$

where $\breve{x}_{ik} = \sum_j c_{ij} x_{jk} \gamma_j / \mu_i$. In this way we can proceed as for the GLM and the IRWLS equations become in matrix notation:

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

where $\tilde{\boldsymbol{W}} = \texttt{diag}(\tilde{\mu})$.

Appendix B

The AIC is equivalent to:

$$AIC = Dev(y|\mu) + 2\text{Dim} = 2\sum_{i=1}^{I} y_i \cdot \ln\left(\frac{y_i}{\mu i}\right) + 2\text{Dim}$$

where $Dev(y|\mu)$ is the deviance and Dim is the effective dimension of the model. For the latter we follow Hastie and Tibshirani (1990) to take the trace of the 'hat' matrix H that is implicit in equation 2.2:

$$\hat{z} = \breve{X}\hat{\beta} = \breve{X}(\breve{X}'W\breve{X} + P)^{-1}(\breve{X}'W)z = Hz$$

An alternative approach is given by the Bayesian Information Criterion (BIC) in which just the second part is altered:

$$BIC = Dev(y|\mu) + \ln n \mathrm{Dim} = 2 \sum_{i=1}^{I} y_i \cdot \ln \left(\frac{y_i}{\mu i} \right) + \ln n \mathrm{Dim}$$

A grid search of λ is normally sufficient to pick up the minimum of AIC and BIC.

References

Alho, J. M. and B. D. Spencer (2005). *Statistical Demography and Forecasting.* Springer Series in Statistics. Springer.

Canner, P. L., N. O. Borhani, A. Oberman, J. Cutler, R. J. Prineas, H. Langford, and F. J. Hooper (1991). The Hypertension Prevention Trial: Assessment of the Quality of Blood Pressure Measurements. *American Journal of Epidemiology 134*(4), 379–392.

Crawford, S. L., C. B. Johannes, and R. K. Stellato (2002). Assessment of Digit Preference in Self-Reported Year at Menopause: Choice of an Appropriate Reference Distribution. *American Journal of Epidemiology 156*(7), 676–683.

Eilers, Paul, H. C. and B. D. Marx (1996). Flexible Smoothing with *B*–splines and Penalties. *Statistical Science 11*(2), 89–121.

Eilers, P. H. C. and M. W. Borgdorff (2004). Modeling and correction of digit preference in tuberculin survays. *International Journal of Tuberculosis and Lung Diseases 8*, 232–239.

Ewbank, D. C. (1981). *Age Misreporting and Age-Selective Undernumeration: Sources, Pattern, and Consequences for Demographic Analysis.* Committee on Population and Demography, Report No 4. Washington, D.C.: National Academy Press.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized Additive Models.* Chapman & Hall.

Hessel, P. A. (1986). Terminal Digit Preference in Blood Pressure Measurements: Effects on Epidemiological Associations. *International Journal of Epidemiology 15*(1), 122–125.

Klesges, R., M. Debon, and J. Ray (1995). Are self-reports of smoking rate biased? evidence from the second national health and nutrition examination survey. *Journal of Clinical Epidemiology 48*, 1225–1233.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Model* (2nd ed.). Monographs on Statistics Applied Probability. Chapman & Hall.

Myers, R. J. (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, J. A. and R. W. M. Wedderburn (1972). Generalized Linear Models. *Journal of the Royal Statistical Society 135*, 370–384.

Rowland, M. L. (1990). Self–reported weight and height. *The American Journal of Clinical Nutrition 52*, 1125–1133.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

Shryock, H. S., J. S. Siegel, and Associates (1976). *Methods and Materials of Demography.* Condensed Edition by Stockwell E.G. New York: Academic Press.

Siegel, J. S. and D. A. Swanson (2004). *Methods and Materials of Demography.* Elsevier Academic Press.

Thompson, R. and R. J. Baker (1981). Composite Link Functions in Generalized Linear Models. *Applied Statistics 30*(2), 125–131.

United Nations (1962). *Demographic Yearbook, 1960.* New York: United Nations.