Estimating Mortality in Small Areas: revising the TOPALS model

Extended abstract submitted to the
ALAP Congress, Puebla, México, 23-26 October 2018

Abstract

High stochastic variation in mortality is the rule when dealing with small populations. Therefore, estimation of mortality age-pattern is often a challenge at subnational level. However, recognising differences in age-specific mortality among regions is crucial for an adequate implementation of national health policies and an objective redistribution of resources. In this paper, we revised the TOPALS relational model which has proven its usefulness in coping with small areas estimations. Unlike the original model and its further developments, we embedded TOPALS in a completely objective estimation procedure in which arguments of the model are optimized based solely on the data in hands. Choice of the standard, which has however negligible effects on the outcomes, is motivated by demographic and historical reasons. A penalized iteratively re-weighted least-squares algorithm is proposed to estimate the model and performances are assessed on minor administrative divisions in Chile and Venezuela in 2000.

KEYWORDS: TOPALS model, Mortality estimation; Small areas; Relational model; Smoothing; Chile, Venezuela.

1 Introduction

Commonly, sub-national analysis demographic studies uses political-administrative hierarchies to divide their sub-populations. The assumption behind is that spatial inequalities are introduced by differences in local policies, or the ability of specific administrative regions to incorporates new technologies or social programs in a faster manner (Bravo and Malta, 2010; Ferguson et al., 2016). To monitor the effects of public health policies or simply to prepare long-term sub-national population projections there is a demand for mortality indicators at sub-national levels. In the process of producing them, Latin American countries are dealing with numerous challenges. On one hand, the existence of large territories with few population counts and on the other, geographically concentrated low levels of coverage in their vital statistics systems (Lima and Queiroz, 2014).

During the past decades, several studies in the region have estimated the coverage of death registration at subnational level (Hill et al., 2009; Lima et al., 2014). Their results acknowledged how spatial inequalities are no just seen in the mortality patterns but within the scope of the vital statistics systems (Freire et al., 2015). However, few studies have focused on estimating age patterns. Most of existing ones make rigid mathematical assumptions relaying on a larger surrounding administrative area pattern (Queiroz et al., 2013; Schmertmann and Gonzaga, 2016).

In the following we propose a revised version of the TOPALS model attempting to overcome some of the drawbacks and offering an objective and elegant estimation procedure. We start from the idea proposed by De Beer (2012) in which a given mortality age-pattern is the sum of a standard profile and a series of deviance which better suits the data in hands. De Beer (2012) called his model TOPALS (tool for projecting age-specific rates using linear splines). Schmertmann and Gonzaga (2016) already proven the advantages of this model in estimating mortality age-pattern in small areas and they embedded TOPALS in a Poisson settings. We move a step further by freeing TOPALS from subjective choice of the model-arguments by means of P-splines (Eilers and Marx, 1996) and estimating the model within the classic Iteratively Re-Weighted Least-Squares algorithm (McCullagh and Nelder, 1989; Nelder and Wedderburn, 1972). We thus called it TOPALS+.

In the following we will first present the data. Afterwards choice and estimation of the standard will be illustrated. The TOPALS+ model is then presented with its estimation procedure and two applications on Chile and Venezuela dataset will be finally shown.

2 Data and standard profile

One can relate living conditions gaps to diverse factors such as race, ethnicity, income, education, occupation, among others. Factors do not necessarily correspond to spatial categories. However, living conditions in Latin America highlight an unequal development along spatial lines (Curto, 1993; Prata, 1992; Schkolnik and Chackiel, 1997). The rapid urbanization process in lockstep with continuous metropolization (Rodríguez and Cunha, 2009b) has created important differences between highly concentrated and dispersed population areas (Beyer, 1967; Cohen, 2006; Montgomery et al., 2003). In this sense, the best possible age-pattern that hold similarities in terms of urbanization levels with the sub-populations under analysis is preferable, regardless of the proximity or their correspondence to the same political-administrative unit.

Working in a relational model framework, we thus decide to consider as standard age-pattern the mortality pattern for the whole group of regions which share a similar urbanization level.

For illustrative purposes, we use data from Chile and Venezuela. On one hand, number of births and deaths counts used in this study come from the national vital statistics systems. On the other, population at risk is taken from national statistics institutes estimations based on the latest census round available. Previous assessments of the vital statistic system in both countries pointed out an adequate data quality/coverage (Mikkelsen et al., 2015).

Specifically, we focused on Minor Administrative Divisions (MIAD) in Venezuela (Municipios) and Chile (Comunas) in 2000 whose urban centers concentrate less than 20,000 inhabitants. This type of classification based on number of inhabitants in cities is conventionally performed in comparative studies, and it guarantees urban condition as benchmark (Rodríguez and Cunha, 2009a; Rodríguez and Villa, 1998).

In the following, we thus aim to estimate mortality age-patterns for 182 Venezuelan (total=335) and 222 Chilean MIADs (total=342). These MIADs gathered 15.4% and 20.8% of the national population, in Chile and Venezuela, respectively. We use data from age 0 to the last open-aged group 80+.

Formally, the whole mortality data are deaths and exposures arranged in two $m \times n$ matrices, $\boldsymbol{D} = (d_{ij})$ and $\boldsymbol{E} = (e_{ij})$. Rows are classified by age at death, $\boldsymbol{a}$, $m \times 1$. Columns indexed by j identify the n Municipios (or Comunas) belonging to the same group in terms of urbanization level.

We assume that the number of deaths d_{ij} at age i in Municipio (or Comuna) j is Poisson distributed with mean $\mu_{ij}\, e_{ij}$ (Brillinger, 1986):

$$d_{ij} \sim \mathcal{P}(e_{ij}\, \mu_{ij})\,. \tag{1}$$

The value of μ_{ij} is commonly named force of mortality and its estimation is the object of all mortality models. In our framework we will treat each Municipio (or Comuna) independently, although all of them will be related to a standard mortality shape.

Specifically standard mortality data are the $m \times 1$ vectors of deaths and exposures equal to the sum over j of the previous matrices:

$$\boldsymbol{d}^s = \boldsymbol{D}\,\mathbf{1}_n = (d_i^s = \sum_j d_{ij}) \qquad \text{and} \qquad \boldsymbol{e}^s = \boldsymbol{E}\,\mathbf{1}_n = (e_i^s = \sum_j e_{ij})\,,$$

where $\mathbf{1}_n$ is a $n \times 1$ matrix of ones. Poisson assumption in (1) holds for the standard deaths and exposures, too: $\boldsymbol{d}^s \sim \mathcal{P}(\boldsymbol{\mu}^s * \boldsymbol{e}^s)$. The symbol $*$ denotes element-wise product.

Instead of relating, in a direct manner, our subnational data to the observed standard mortality pattern, we first smooth $\boldsymbol{\mu}^s$. In this way, in further analysis, we do not carry out all random fluctuations which are present in the overall standard mortality data. We achieve smoothness by a P-spline approach. For further details about this model see the seminal and the review papers by Eilers and Marx (1996) and Eilers et al. (2015), respectively. In few words, we model the logarithm of the force of mortality as follows:

$$\ln(\boldsymbol{\mu}^s) = \boldsymbol{\eta}^s = \boldsymbol{B}\,\boldsymbol{\alpha}^s\,,$$

where $\boldsymbol{B}$ is a $m \times k$ matrix of B-splines (de Boor, 1978) and $\boldsymbol{\alpha}^s$ are k associated coefficients. One of the advantage of this approach lays in its estimation procedure which simply translates in a penalized version of the Iteratively Re-Weighted Least-Squares (IRWLS) commonly used for Generalized Linear Models (McCullagh and Nelder, 1989; Nelder and Wedderburn, 1972). See next section.

3 The TOPALS+ model

Following the reasoning behind the TOPALS model (De Beer, 2012), we decide to model subnational mortality data as the sum of the standard mortality age-pattern and a set of deviance. Following Schmertmann and Gonzaga (2016), we embed TOPALS in a Poisson framework. For a given Municipio (or Comuna) j, we model the log-mortality as follows

$$\ln(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{\eta}^s + \boldsymbol{\gamma}\,, \qquad (2)$$

where the vector $\boldsymbol{\gamma} = (\gamma_i)$ denote the difference between the standard mortality pattern and the underlying force of mortality at a subnational level. Dealing with small data, we cannot obtain reasonable outcomes by simply dividing observed death rates by the smooth standard force of mortality. We thus decide to assume smoothness of the vector $\boldsymbol{\gamma}$. As in the estimation for the standard age-pattern, we describe the vector $\boldsymbol{\gamma}$ as the linear combination of equally-spaced B-splines and associated coefficients:

$$\boldsymbol{\gamma} = \boldsymbol{B}\,\boldsymbol{\alpha}\,. \tag{3}$$

Instead of subjectively selecting number and/or the location of the k B-splines with respect to age $\boldsymbol{a}$, we emulate the P-spline approach in this new setting. We take a *rich* number of B-splines over the domain (i.e. age) and simultaneously we penalize the differences of the associated coefficients in order to achieve smoothness of the estimated $\boldsymbol{\gamma}$.

As mentioned, P-splines can be estimated by penalized an IRWSL algorithm. Here we decide to briefly present this procedure to underline the simplicity of the algorithm with respect to the model (2). In formulas, we solve the following system of equations:

$$(\boldsymbol{B}'\boldsymbol{W}\boldsymbol{B} + \boldsymbol{P})\,\boldsymbol{\alpha} = \boldsymbol{B}\boldsymbol{W}\boldsymbol{z} \tag{4}$$

where, as in the GLM, $\boldsymbol{W} = \mathtt{diag}(\boldsymbol{\mu} * \boldsymbol{e})$ and $\boldsymbol{z} = \frac{\boldsymbol{d} - \boldsymbol{\mu} * \boldsymbol{e}}{\boldsymbol{\mu} * \boldsymbol{e}} + \boldsymbol{\eta}$. The additional penalty term measures the roughness of the coefficients by first-order differences tuned by a smoothing parameter λ:

$$\boldsymbol{P} = \lambda\,\boldsymbol{\Delta}'\boldsymbol{\Delta}\,. \tag{5}$$

The matrix of differences $\boldsymbol{\Delta}$ are $m \times (m-2)$ matrix which can be simply constructed in a software as `R` (R Development Core Team, 2018). See appendix.

When λ is equal to 0, the model reduces to a simple Poisson-GLM with B-splines as regressors and the vector of $\boldsymbol{\gamma}$ will theoretically be a curve with k degree-of-freedom. The larger the λ, the smoother will be the series of $\boldsymbol{\alpha}$ and, consequently, the estimated $\boldsymbol{\gamma}$. Optimal value of λ can be selected by Bayesian Information Criterion (BIC, Schwarz, 1978).

The BIC is a common tool for model selection and it corrects the log-likelihood of a fitted model for the effective dimension. The expression for BIC is given by

$$\mathrm{BIC}(\lambda) = \mathrm{DEV} + \ln(m)\,\mathrm{ED} \tag{6}$$

where DEV denotes the deviance which measures the goodness-of-fit of the model. The other term ED represents the effective dimension which is the correspondent concept of number of parameters in a smoothing context.

It is noteworthy that number of B-splines is negligible for the whole process as long as we have a sufficiently large k to describe all eventual fluctuations in the data. Smoothness will be enforced only by $\boldsymbol{P}$ by tuning λ. Degree and location of B-splines are also not important for the final outcome. Here we use equally-spaced cubic B-splines. Finally the degree of differences in $\boldsymbol{\Delta}$ indicates the “prior” function achieved when λ is extremely large. By selecting a first-order differences, we implicitly assume a constant $\boldsymbol{\gamma}$ model as "ultimate smooth" function. In other words, whenever subnational mortality data cannot provide enough information for a specific age-pattern, the model will tend to select $\boldsymbol{\eta}$ as the sum of standard mortality and an optimal constant value, leading to a proportional hazard framework. Conversely, relatively larger population may depart from $\boldsymbol{\eta}^s$ in a flexible manner, if data required so.

Being in a regression setting, we can easily compute the standard errors for $\boldsymbol{\gamma}$ and consequently for $\boldsymbol{\eta}$ from the objects in(4) after convergence. In formula the variance of $\boldsymbol{\gamma}$ is given by:

$$Var(\boldsymbol{\gamma}) = \boldsymbol{B}\,(\boldsymbol{B}'\boldsymbol{W}\boldsymbol{B} + \boldsymbol{P})^{-1}\boldsymbol{B}' \tag{7}$$

Therefore the square-root of the diagonal of (7) provides directly standard errors to build confidence intervals.

In the appendix we provide a small excerpt of the `R` code used to estimate the system of equations in (4), standard errors for the fitted values and BIC.

4 Applications

Among the 222 Comunas in Chile and the 182 Municipios in Venezuela, we illustrate our approach showing outcomes from 3 different MIADs. We select these regions as representatives of different sample sizes. Specifically, we show results on a really small, a medium size and a relatively large MIAD in both countries. See Figure 1

In this long abstract, we focus on the methodological aspects of the study. However, for the final paper, we plan to deeply study the differences among different Chilean and Venezuelan MIADs providing insights for a better understanding of subnational mortality inequalities in these countries.

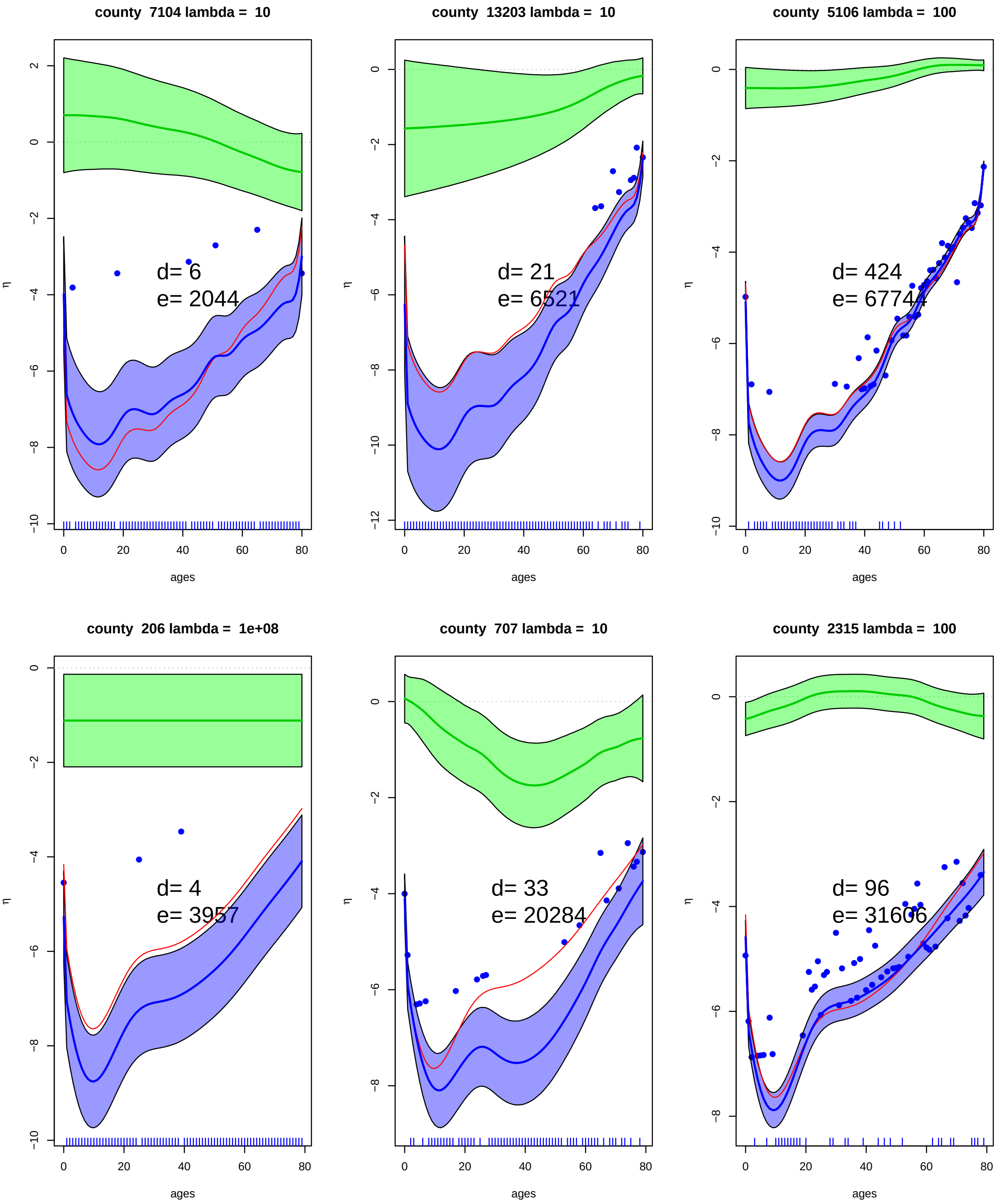

Figure 1: Actual death rates and fitted values from TOPALS+ with 95% confidence intervals (in blue). We also show estimates and confidence intervals for the deviation vector $\boldsymbol{\gamma}$. Standard age-pattern ($\boldsymbol{\eta}^s$) is depicted in red. Three Comunas in Chile (top panels) and three Municipios in Venezuela (bottom panel). The values denoted by d and e presents the total number of deaths and population at risk for each subnational dataset. Administrative number and selected λ are given in the titles.

References

Beyer, G. (1967). *The Urban Explosion in Latin America.* Cornell University Press.

Bravo, J. and J. Malta (2010). Estimating Life Expectancy in Small Population Area. In *Joint Eurostat/UNECE Work Session on Demographic Projections*, Lisbon.

Brillinger, D. R. (1986). The Natural Variability of Vital Rates and Associated Statistics. *Biometrics 42*, 693–734.

Cohen, B. (2006). Urbanization in developing countries: Current trends, future projections, and key challenges for sustainability. *Journal of Technology in Society 28*, 63–80.

Curto, S. (1993). Geographical inequalities in mortality in Latin America. *Social Science & Medicine 36*(10), 1349–1355.

De Beer, J. (2012). Smoothing and projecting age-specific probabilities of death by TOPALS. *Demographic Research 27*(20), 543–592.

de Boor, C. (1978). *A Practical Guide to Splines.* New York: Springer.

Eilers, P. H. C. and B. D. Marx (1996). Flexible Smoothing with B-splines and Penalties (with discussion). *Statistical Science 11*, 89–102.

Eilers, P. H. C., B. D. Marx, and M. Durbán (2015). Twenty years of P-splines. *SORT. Statistics and Operations Research Transactions 39*(2), 149–186.

Ferguson, B., G. Reniers, T. Araya, J. Jones, and E. Sanders (2016). Empirical Bayes Estimation of Small Area Adult Mortality Risk in Addis Ababa, Ethiopia. Paper presented at the 2004 PAA Meeting.

Freire, F., B. Queiroz, E. Lima, M. Gonzaga, and F. Souza (2015). Mortality estimates and construction of life tables for small areas in Brazil. Paper presented at the 2015 PAA Meeting.

Hill, K., D. You, and Y. Choi (2009). Death distribution methods for estimating adult mortality: Sensitivity analysis with simulated data errors. *Demographic Research 21*, 235–254.

Lima, E. and B. Queiroz (2014). Evolution of the deaths registry system in Brazil: associations with changes in the mortality profile, under-registration of death counts, and ill-defined causes of death. *Cadernos de Saúde Pública 30*(8), 1721–1730.

Lima, E., B. Queiroz, and D. Sawyer (2014). Método de estimação de grau de cobertura em pequenas áreas: uma aplicação nas microrregiões mineiras. *Cadernos de Saúde Pública 22*(4), 409–418.

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

Mikkelsen, L., D. E. Phillips, C. AbouZahr, P. W. Setel, D. de Savigny, R. Lozano, and A. D. Lopez (2015). A global assessment of civil registration and vital statistics systems: monitoring data quality and progress. *The Lancet 386*(10001), 1395–1406.

Montgomery, M., R. Stren, B. Cohen, and H. Reed (2003). *Cities Transformed: demographic change and its implications in the developing world.* Washington: National Academies Press.

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

Prata, P. (1992). A Transição Epidemiológica no Brasil. *Cadernos de Saúde Pública 8*(2), 168–172.

Queiroz, B, ., E. Lima, F. Freire, and M. Gonzaga (2013). Adult mortality estimates for small areas in Brazil, 19802010: a methodological approach. *The Lancet 381*, S120.

R Development Core Team (2018). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rodríguez, J. and M. Cunha (2009a). Demographic Transformations, Convergences and Inequalities in Latin America: what the future holds? Presented at the XXVI IUSSP International Population Conference. Marrakech.

Rodríguez, J. and M. Cunha (2009b). Urban growth and mobility in latin america. In S. Cavenaghi (Ed.), *Demographic Transformations and Inequalities in Latin America. Historical trends and recent patterns*, Investigaciones series. Rio de Janeiro: Latin American Population Association.

Rodríguez, J. and M. Villa (1998). *Distribuciòn espacial de la poblaciòn, urbanizaciòn y ciudades intermedias: hechos en su contexto.* Santiago de Chile: Economic Commission for Latin American and Caribbean, Population Division.

Schkolnik, S. and J. Chackiel (1997). América Latina: la transicin demográfica en sectores rezagados. Presented at Population International Conference of the International Union for the Scientific Study of Population. Beijing,.

Schmertmann, C. and M. Gonzaga (2016). Estimating age- and sex-specific mortality rates for small areas with TOPALS regression: an application to Brazil in 2010. *Revista Brasileira de Estudos de População 33*(3), 629–652.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics 6*, 461–464.

Appendix: Software considerations

In this appendix we outline the snippets for estimating the TOPALS+ as presented in Section 3. We work in `R` (R Development Core Team, 2018) because of its widespread use, but the following code should be easily understood by someone who is unfamiliar with this language, if the following notation is known.

The symbol `<-` is an assignment statement, an asterisk, `*`, means element-by-element multiplication, `%*%` means matrix multiplication, `t()` means transpose and `solve()` estimates a generic least square model, i.e. one would write `x <- solve(A, b)` to solve $A \cdot x = b$ for x, where b can be either a vector or a matrix.

We assume that standard log-mortality is given over m ages and assigned to an object called `etaS.hat`. Then we start to construct the B-splines basis over the vector of age `a` of length `m` by using the routine `MortSmooth_bbase()` in the library `MortalitySmooth`:

```
B <- MortSmooth_bbase(x=a, xmin=min(a), xmax=max(a), ndx=20, deg=3)
k <- ncol(B)
```

In this example the number of B-splines will be equal to $k = 23$.

Given a smoothing parameter `lambda`, we can then construct our penalty term $\boldsymbol{P}$ as in (5):

```
  Delta <- diff(diag(k), diff=1)
  tDeltaDelta <- t(Delta)%*%Delta
  P <- lambda * tDeltaDelta
```

Starting values for $\boldsymbol{\gamma}$ (`gamma` in `R`) are not crucial here. However we assume that standard mortality is equal to the subnational mortality pattern: $\boldsymbol{\eta}^s = \boldsymbol{\eta} \Rightarrow \boldsymbol{\gamma} = \mathbf{0}$

```
  gamma <- rep(0,m)
```

We now start updating `gamma` following the penalized IRWLS presented in (4):

```
  ## log-mortality
  eta <- etaS.hat + gamma
  ## force of mortality
  mu <- exp(eta)
  ## Poisson expected values
  e.mu <- e*mu
  ## diagonal Poisson weights
  W <- diag(c(e.mu))
  ## working-response
  z <- (d - e.mu)/e.mu + eta
  ## right-hand-side of IRWLS
  tBWB <- t(B) %*% W %*% B
  ## adding penalty term
  tBWBpP <- tBWB + P
  ## left-hend-side of IRWLS
```

```
tBWz <- t(B) %*% W %*% z
## estimating alpha
alpha <- solve(tBWBpP, tBWz)
## old gamma
gamma.old <- gamma
## new gamma
gamma <- B %*% alpha
```

The previous lines are repeated in a `for`-loop until two successive vectors of `gamma` do not differ much. For instance, we can break the loop when the following object

```
dgamma <- max(abs((gamma.old - gamma)/abs(gamma.old)))
```

is smaller than 10^{-6}. Commonly a handful of iterations is needed to achieve convergence.

As explained λ needs to be optimize based on a objective criterion such as BIC which could be easily computed as follows:

```
DEV <- 2 * sum(d * log(d/mu))
ED <- sum(diag(solve(tBWBpP, tBWB)))
BIC <- DEV + log(m)*ED
```

One can compute `BIC` over a series of λ and check for which value is minimized.

Standard errors for $\boldsymbol{\gamma}$ and consequently for $\boldsymbol{\eta}$ can be computed as explained in (7):

```
V  <- solve(tBWBpP)
Vs <- B%*%V%*%t(B)
se <- sqrt(diag(Vs.eta))
```