Chapter 5

A Mortality Model for Pandemics and Other Contagion Events

Gary Venter

Abstract The crisis caused by COVID-19 has had various impacts on the mortality of different sexes, age groups, ethnic and socio-economic backgrounds and requires improved mortality models. Here a very simple model extension is proposed: add a proportional jump to mortality rates that is a constant percent increase across the ages and cohorts but which varies by year. Thus all groups are affected, but the higher-mortality groups get the biggest increases in number dying. Every year gets a jump factor, but these can be vanishingly small for the normal years. Statistical analysis reveals that even before considering pandemic effects, mortality models are often missing systemic risk elements which could capture unusual or even extreme population events. Adding a provision for annual jumps, stochastically dispersed enough to include both tiny and pandemic risks, improves the results and incorporates the systemic risk in projection distributions. Here the mortality curves across the age, cohort, and time parameters are fitted using regularised smoothing splines, and cross-validation criteria are used for fit quality. In this way, we get more parsimonious models with better predictive properties. Performance of the proposed model is compared to standard mortality models existing in the literature.

5.1 Introduction

Probabilistic mortality models usually assume that deaths are independent, identically distributed Yes/No events, which makes the number of Yes events in a period binomially distributed. For low-probability events such as deaths in a year, binomial distributions are very close to Poisson, which is more convenient for modelling. It is common, however, for heavier-tailed distributions, such as negative binomial, to give better fits, which suggests that there are unmodelled correlated effects. Population

G. Venter (✉)

Actuarial Science, School of Professional Studies, Columbia University, 203 Lewisohn Hall, 2970 Broadway, MC 4119, New York, NY 10027, USA

e-mail: gv2112@columbia.edu

© The Author(s) 2022

M. C. Boado-Penas et al. (eds.), *Pandemics: Insurance and Social Protection*, Springer Actuarial, https://doi.org/10.1007/978-3-030-78334-1_5

events, such as disease outbreaks, extreme weather and natural disasters produce year-to-year jumps and contribute to such effects on mortality rates. This is likely to be a reason the mortality models have heavier-tailed residuals. Including model provisions for such population events can improve the fit of the theoretically appropriate Poisson models, while also providing enough flexibility to account for pandemic experience.

In this chapter, a simple model with mortality jumps every year, which are negligible in ordinary years, is proposed: a single factor, drawn annually from a fixed distribution, increases all the modelled mortality means proportionally. In the example, the jump distribution fit to historical data improves on typical models, and also gives reasonable probabilities for events like the COVID-19 pandemic. The other model parameters are fit to smoothing splines, which have several advantages enumerated below.

A few related papers on pandemic mortality rates have appeared recently. For example, Özen and Şahin (2020) postulate occasional population jump events to better predict the market behaviour of mortality catastrophe bonds. Chen and Cox (2009) use a similar idea for mortality securities in general. Zhou et al. (2013) generalise this to two-population modelling, which they point out is often used for mortality securities. Alijean and Narsoo (2018) test a number of mortality models over an extended period for quality of fit, and find that a jump model like these works better than the well-known continuous models, particularly in that it is able to account for the effects of the Spanish flu epidemic around 1918. Barigou et al. (2021) fit smoothed surfaces to mortality rates, with and without pandemic effects, using methodology very similar to the smoothing splines used here. Cairns et al. (2020) model COVID-19 mortality in greater demographic detail. O'hare and Li (2017) test several standard mortality models of log rates for normality of residuals, and find they almost universally fail their tests. This is actually to be expected from Poisson-distributed mortality rates, and the models here use log links to fit the actual mortality counts. They also find evidence of correlation of residuals, which does indicate the type of systematic effects that the jump model here is aiming to avoid.

Section 5.2 provides a summary of the methodology and findings. Section 5.3 lays out methodology to do parametric regression by fitting smoothing splines across all of the age, period, and cohort variables in MCMC, using Bayesian shrinkage for the smoothing, and discusses the predictive advantages of such smoothing. This is done for both cubic and linear splines, optimising the degree of smoothing by conditional expectations instead of the typical cross-validation methods. Details of the mortality models used and the fitting process are given in Sect. 5.4. Section 5.5 contains the results, as well as a discussion of possible generalisations of the models. Section 5.6 concludes, and some computer code is in the Appendix.

5.2 Highlights of Methodology and Findings

5.2.1 *Summary of Methodology*

The approach is to apply smoothing splines to fit standard mortality models to the French data, using both Poisson and negative binomial residuals. For the best models tried, the negative binomial fit better for both male and female populations. The models were extended to add an annual mortality multiplier factor drawn from a stochastic process, while using only Poisson residuals.

Mortality data generally comes in annual blocks by calendar age at death, and year of death minus age approximates the year of birth cohort. One class of models has factors for age, period (i.e., year), and cohort. If these factors simply multiply, this is called an APC model. These are usually estimated in log form, so the model is additive, and it can be fit by regression. More complicated models multiply the period log factors by an age modifier, to reflect the fact that medical advances, etc., affect some ages more than others. The Lee and Carter (1992) model has age effects and age-modified period effects. Renshaw and Haberman (2006) add cohorts, and this is called the RH model below. The ages that trend faster can change, however, especially for models over wide age and period ranges. Developing economies, for instance, can initially have sharp improvements in pre-teen mortality, then later shift to the more usual pattern from improved treatments for diseases affecting older people. Hunt and Blake (2014) allow multiple trends over various periods, each with its own age modifiers. Venter and Şahin (2018) used this for US males, and found an additional trend for HIV-related mortality in the 1980s and 1990s for young adults, and another complex trend for ages in the 40s, perhaps related to substance abuse.

Parametric curves are often used across the age, period, and cohort log factors. For instance, Perks (1932) introduces a four-parameter mortality curve that fits pretty well across ages. However the parameters do change from time to time. Xu et al. (2019) use curves from multifactor bond-yield models to fit mortality by age for individual cohorts or cohort groups. These come with automatic projections to older ages, which are like long bond yields in these models.

Cubic splines are often fit across the factors instead of using parametric curves. They are smoother and use fewer parameters than models with parameters at every age, period, and cohort. A recent refinement is to use smoothing splines, which typically are cubic splines with a penalty included to constrain the average second derivative across the curve. Linear splines have generally been considered too jagged for such modelling, but Barnett and Zehnwirth (2000) introduce a form of linear smoothing splines that constrain second differences, which is analogous to how smoothing is done for cubic splines, and these can be fairly smooth as well. Using smoothing splines across regression variables is called semiparametric regression. A detailed source is Harezlak et al. (2018), and Venter and Şahin (2021) apply it to an actuarial model.

5.2.2 Summary of Findings

For the female population, adding the annual jump factors made this Poisson model better than the best previous (negative binomial) model. For the males, doing this was almost as good, but the contagion was not captured completely. More complex possible population event models are discussed in Sect. 5.5.

Figure 5.1 shows the fitted mortality multipliers for both populations. These are estimated in logs as additive terms. The male and female factors are generally similar. Their peaks are at 2003, a year with about 14,000 extra deaths from an extreme heat wave. 2018 also has high factors, and it also had a bad heat wave, but preparations for it were better, so it is not quite as high. The 2020 COVID-19 pandemic had about 5 times as many deaths as in 2003. Under the estimated distribution of factors, for males 2003 was about a 1 in 9 year event and 2020 about 1 in 50 years. As such population events can occur over adjacent years, these frequencies are not unreasonable.

Fig. 5.1 Fitted mortality multipliers

5.3 Semiparametric Regression in MCMC

MCMC is an effective tool for fitting shrinkage splines. In this section, first the theoretical background for this is presented, with Bayesian and frequentist interpretations. Then the design matrices for estimating splines by regression are specified, followed by the hows and whys of parameter shrinkage, including some history. Cross validation is used in semiparametric regression to compare the goodness of fit of models from a predictive viewpoint, and this is relatively simple to do in MCMC. Cross validation is also widely used to estimate the best degree of shrinkage for a model, but this is problematic. A better alternative is available when using MCMC estimation.

5.3.1 MCMC Parameter Shrinkage

Markov Chain Monte Carlo (MCMC) is a method for estimating probability distributions through stochastic sampling. “Markov chain” means that when generating a sequence of samples, each one uses the previous one, without reference to earlier history. It was developed by physicists in the late 1940s as an efficient numerical integration method to model distributions of particle interactions. After enough samples, it has distributional stability—new samples will all be from the same distribution.

In Bayesian statistics, MCMC is used to sample parameter sets from the joint distribution of the parameters and the data. This can be computed as the product of the prior distribution of the parameters and the conditional distribution of the data given the parameters. The latter is the likelihood, so the product is called the joint likelihood. By the definition of conditional distribution, it is also the product of the conditional distribution of the parameters given the data with the probability of the data. The latter is not known, but it is a constant. Thus the joint likelihood is proportional to the conditional distribution of the parameters given the data, and sampling from it with MCMC will provide an estimate of that conditional distribution.

The same thing can be done with random effects. In classical statistics, parameters are unknown constants which do not have distributions. Random effects are like parameters, but they have distributions. These are not subjective distributions but are part of the model specification, just like residual distributions are. They can be evaluated based on the fit and revised as necessary. Random-effects estimation typically just computes the mode of the joint likelihood, but it could estimate the conditional distribution of the effects given the data by MCMC like Bayesians do. Many contemporary Bayesians are abandoning subjective probabilities and consider the priors to be specified distributions that can be revised as needed, just like is done in random effects.

The terminology used here is somewhat intermediate. In the models used, there are no parameters in the frequentist sense, just random effects. But that is not a widely understood term, so it is simpler to just call them parameters. As their postulated

distributions are not subjective, the terms “prior” and “posterior” can be misleading. Modern Bayesians insist that these terms do not imply subjective probabilities, but their subjective use is so ingrained in the vocabulary that trying to define them out can lead to confusion.

Smoothing splines involve constraints that limit the second derivatives or second differences of the curves. This can be done in MCMC for each parameter individually by postulating shrinkage distributions. Here, a shrinkage distribution is any distribution with mode at or asymptotic to zero. They give more weight to parameters closer to zero, so favour such parameters. Examples include the standard normal distribution and double-exponential distributions, which mirror the exponential on the negative reals. Gamma distributions with shape parameter less than 1 have mode asymptotic to zero, and are used here for the distribution of the annual event frequency jumps.

The purpose of parameter shrinkage is not just to make smoother curves. It actually can improve the predictive accuracy of models. This will be easier to discuss once the models are formally specified.

5.3.2 *Spline Regressions*

The APC models are a good place to start, as in log form they are purely linear models. The log model value at any data point is just the sum of the age, period, and cohort parameters for that point.

Mortality data for the illustration is taken from Human Mortality Database (2019). Deaths are in a column listed by year then age within year. In this case this is French female and male data for years 1979–2018 and ages 50–84. These are from 74 year-of-birth cohorts, 1929 to 1968, each with 1–35 ages. The column of death counts is the dependent variable for each model. Also the population exposure for each observation is provided as another column.

A design matrix is then constructed giving dummy variable values in columns for each age, period, and cohort for each observation. For specificity, the first age, period, and cohort variables are not included. There is also a constant term in the model that the code computes outside of the design matrix. The ages, periods, and cohorts are not independent of each other and further constraints are needed, as discussed later. The design matrix thus has 34 columns for age variables, 39 for period variables, and 73 for cohort variables. To fill out the design matrix, reference columns are created for each data observation to specify the age, period, and cohort it comes from.

For a straight regression APC model, the variables are 0,1 dummies. Age i has value 1 in just those rows for which the observation is from age i , and similarly for the period and cohort variables. For linear splines, an alternate regression could be done where all the variables are for the slopes of the line segments connecting the log factors, which are the first differences of the log factors. The age i variable for this would still be 0 for observations from ages less than i and 1 for observations at age i . Later observations would add the later slopes to the existing log factors before those points, so the age i variable would also be included in them. Thus the age i variable

would be 1 for observations from ages i and greater, and the same for periods and cohorts.

For the linear splines here, the variables are the second differences of the final age, period, and cohort parameters. Then the variable for age i is still 1 for observations with age i , and 0 for earlier observations. Its value increases the final log factor at ages $k \geq i$ and so is in all those points, just like the slope parameters were. But it also increases the slope (first difference) at i , and so is in all the later first differences as well. Thus it gets added in an additional time for each subsequent point. This means that for a cell from age k , the dummy for age i is $(1 + k - i)_+$. The same is true for periods and cohorts.

Cubic splines have similar design matrices. To define these, consider one direction of variables, like age, with spline-segment variables f_1 to f_n . Then the design matrix value $f_i(z)$ for an observation at age z is a function defined for any real $0 \leq z \leq n$ so the curves can be interpolated. They are: $f_1(z) = 1$, $f_2(z) = z$, and for $i > 2$, $f_i(z) = (z - (i - 2))_+^3 - (n - (i - 2))(z - (n - 1))_+^3$. The second term is nonzero only for $z > n - 1$. The design matrix just has values for integer k . For a cell from age k , the value for the variable for age i is $f_i(k)$. These basis functions, from Hastie et al. (2017),¹ assume that the splines are linear outside of $1 < z < n$. Here they are modified by constant factors by column that change the fitted coefficients but simplify the formulas and put the fitted coefficients more on the same scale.

The average second derivative of the spline is a complicated but closed-form function of the parameters. A smoothing spline minimises the NLL plus a selected constant λ times the average second derivative. This is a penalised regression similar to ridge regression and lasso. It can be readily estimated with a nonlinear optimiser for any given λ . Some experimentation with this found that increasing λ usually shrinks all the spline parameters to a similar degree. This suggests an alternative way to define cubic smoothing splines in MCMC: use shrinkage distributions on the spline parameters to penalise large segment parameters. This also reduces the average second derivative very similarly to the usual smoothing splines, but does not map exactly to a specific shrinkage λ . It does make the splines estimable in MCMC, which has advantages discussed below.

5.3.3 Why Shrinkage?

With these design matrices, the spline models are just regression. Without parameter shrinkage they would give the same fits as the regular 0,1 dummy variables. The reason for shrinkage is based in reduced error variances. This traces back to the Stein (1956) paper, which showed that when estimating three or more means, the error variance is reduced by shrinking the estimates all towards the overall mean to some degree. Actuaries have been doing this heuristically with credibility weighting since

¹ A derivation is at <https://stats.stackexchange.com/questions/172217/why-are-the-basis-functions-for-natural-cubic-splines-expressed-as-they-are-es>.

Mowbray (1914). The same thing was extended to regression in Hoerl and Kennard (1970) with ridge regression. That minimises the negative loglikelihood NLL plus a proportion of the sum of the squared parameters. For parameters β_j this estimation minimises:

$$NLL + \lambda \sum \beta_j^2$$

They showed that there is always some positive λ that reduces the error variance from that of maximum likelihood estimation. They also called their method “regularisation” as it was derived from a slightly more general method from Tikhonov (1943) which is often translated from the Russian using that non-descriptive term.

Ridge regression is actually a form of shrinking the fitted means towards the grand mean. In practice, all the variables are standardised to have mean zero, variance one before going into the design matrix. Then each fitted value is the constant plus terms of mean zero, so shrinking the parameters shrinks the fitted values towards the overall mean. Just like in credibility, this biases the estimated means towards the overall mean while improving estimation accuracy. The problem is in selecting λ . In practice this is usually done by cross validation. The data is divided into a number of smaller subsets, and the model is estimated by fitting the data many times for different values of λ with each of the subsets excluded separately. Then the NLL can be computed for each subset using the models that excluded it, and the results compared among the different λ s.

Ridge regression eventually led to lasso—least absolute shrinkage and selection operator, following Santosa and Symes (1986) and Tibshirani (1996). This replaces the square of the parameters with their absolute values, and so minimises

$$NLL + \lambda \sum |\beta_j|$$

Lasso shrinks some parameters to exactly zero—more as λ increases—which is why it is both a shrinkage and a variable selection operator. There is good software available, e.g., in R, for this, and it quickly does extensive cross validation as well. That is why lasso has become more popular than ridge regression.

Ridge regression and lasso constraints are summed over the parameters but can be approximated in MCMC by postulating shrinkage distributions for the parameters individually. For instance, if β is normal in 0 and σ , its log density, ignoring constants, is:

$$-\log(\sigma) - 0.5(\beta/\sigma)^2$$

Then minimising the joint negative loglikelihood would mean minimising $NLL + \log(\sigma) + 0.5\beta^2/\sigma^2$. For any fixed σ , the $\log(\sigma)$ term is a constant, so this reduces to ridge regression with $\lambda = 0.5/\sigma^2$.

If β is Laplace (double exponential) distributed in 0 and s , its log density is:

$$-\log(2s) - |\beta|/s$$

For a fixed s , then, minimising the negative joint loglikelihood comes down to minimising $NLL + |\beta|/s$, so is lasso with $\lambda = 1/s$.

There is a problem with parameter shrinkage for APC models, however. Each parameter represents the effect of a given age, period, or cohort. These can rarely be taken out of the model, which would effectively happen if they were shrunk too much towards zero. That is where the splines become useful. Shrinking a slope change or cubic segment parameter to zero just extends the existing curve one step further at that point. Thus with shrinkage the smoothing splines become more compressed and do not skip any age, period, or cohort effect. The points where the curves change are called knots, and shrinkage simultaneously optimises which knots to use as well as the fitted accuracy. These are separate steps in non-MCMC spline fitting and are often not coordinated enough to optimise both simultaneously.

5.3.4 Cross Validation in MCMC

The usual way of measuring goodness-of-fit of models is penalised loglikelihood. The likelihood is penalised because of sample bias—bias in the likelihood arising from measuring it on the sample that the parameters were fit to. A bias penalty is calculated, usually as a function of number of parameters and sample size. Subtracting the penalty from the loglikelihood gives the measure. These are attempting to estimate what the loglikelihood would be on a new, independent sample. The penalty from the small sample AIC, AICc, for instance, has been shown to give an unbiased estimate of the sample bias. Of course there is still some error in such estimates.

Models with shrinkage, like ridge regression, lasso, and smoothing splines, do not have parameter counts that are appropriate for these measures. Due to shrinkage, the parameters do not use up the same degrees of freedom as they do in other models. Cross-validation is a way to estimate the sample bias when there are not good parameter counts available. The sample is divided up into subsamples and the likelihood measured on each subsample with the parameters fit without it, and the results are compared for various values of the shrinkage parameter. Thus cross validation produces a penalised likelihood. How many subsamples and how they are selected influence this process, and it also has estimation error for the sample bias.

MCMC can efficiently produce an estimate of the unbiased loglikelihood, using an approach called leave-one-out cross validation, or loo. Each sample point is taken as an omitted subsample in computing the cross-validation likelihood. Because there are many parameter sets generated in MCMC estimation, the likelihood of the point can be computed under each one of those. Gelfand (1996) showed that a weighted average of those likelihoods, with more weight given to the worse values, gives an unbiased but noisy estimate of the likelihood of a left-out point. The sum of those estimates is the loo likelihood, and as a sum it has a lower average estimation error. This method is called importance sampling, and in fact the weights are inversely proportional to the likelihood. Thus the loo estimate for a point is the harmonic mean of its likelihoods across the parameter sets.

This produces an unbiased estimate of the sample bias, but unfortunately it has turned out to have a high estimation error. Recently, Vehtari et al. (2019) improved upon it, with something akin to extreme value theory. They fit a Pareto to the tail of the distribution of likelihoods for a point, and for the 20% most extreme likelihoods use the Pareto fit instead of the actual likelihood. The sum of these over the data points gives a more accurate estimate of the likelihood for the model excluding sample bias. After the MCMC estimate is finished, this can be done quickly in an R application itself called loo.

The shrinkage parameter can then be estimated by running MCMC several times with various trial shrinkage values, and picking the one with the highest loo. This is how cross validation is used to select the shrinkage parameter in other forms of estimation as well. As with any estimation based on optimising penalised likelihood, this runs the risk of actually choosing the parameter that produces the greatest understatement of the sample bias instead of the most predictive λ .

There is an alternative approach under MCMC: also postulate a distribution for the shrinkage parameter, so it too becomes a random effect. This is called the fully Bayesian method, as now there are no parameters in the model in the frequentist sense of being an unknown constant. Usually you would want to use a distribution for this that has minimal impact on the conditional mean of this parameter. The estimates below were done with the Stan MCMC package. That assumes an initial distribution for a parameter, unless otherwise specified, as uniform on $\pm 1.7977e+308$, which is the range of real numbers in the double-precision system most computers use. That could be postulated for the log of the shrinkage parameter, as assuming a wide range for a positive parameter usually biases it upwards. I usually find that ± 8 is wide enough for that log. MCMC would of course produce a range of values for the parameter, and the whole sample would be used in risk analysis. The conditional mean is usually presented as the estimate. This is easier than cross validation in that you do not need runs for several different trial values, and it is better in that it does not have the problem of favouring shrinkage parameters with understated sample bias. MCMC sample distributions of parameters represent possible parameter sets that could have generated the data along with the distribution of those sample sets. The mean of this gives the most accurate estimate for each parameter.

5.4 Model Details

The models above were fit using semiparametric regression for Poisson and negative binomial residuals and with and without the gamma-distributed jumps. The Poisson distribution is very tight for large means, which forces those fits to emphasise the larger cells. But this is the distribution that independent deaths should follow. Comparing fits to the negative binomial thus provides an indicator of missing systematic effects in the model.

5.4.1 Formulas

The form of the negative binomial used is the one in General Linear Models. The mean is μ and the variance is $\mu + \mu^2/\phi$. An indicator of how much spread a distribution can have is stdv./mean or its square, variance/mean². For the Poisson, that is $1/\mu$, and that can get small for large cells, which forces the model to fit very well at those cells. For the negative binomial, it is $1/\mu + 1/\phi$, which puts a minimum on that ratio, and increases model flexibility. Still, if there are no contagion effects, the Poisson should work well.

A gamma distribution is used for the annual mortality multipliers. The gamma in a, b has mean ab and variance ab^2 . The density at zero is asymptotic to infinity for $a < 1$. The models here assume $a = 0.1$. For females, b is estimated to be close to 0.4. That gamma then has mean 0.04, and its standard deviation is about 0.125. A large part of this distribution is close to zero, with the median at 0.00024. The largest log factor for this sample, at 2003, is about 0.1, which is at the 90th percentile of the distribution. Probably for the 2020 pandemic year the log factor will be more like 0.5, which is close to the 98.5th percentile, which seems plausible.

All the parameters except the constant and the annual log factors are assumed to be Laplace (double exponential) distributed, with mean zero and the same scale parameter s . The Laplace variance is $2s^2$, so the smaller s is, the tighter is the range around zero, so the more shrinkage it produces. Also $\log s$ is assumed to be uniform on ± 6 . For males, sampled values of it range from -3 to 1.5 , and for females they are in a tight range around -5 . The constant term, which is eventually exponentiated, is assumed uniform on ± 8 . it ends up near 0.5 for females and 2.5 for males.

For the APC model, let y be the column of death counts, x be the design matrix, v be column of parameters, c be the constant term, which here is not in the design matrix, and q the column of exposures. Using “ $\times$ ” for pointwise vector multiplication, the Poisson or negative binomial mean μ is:

$$\mu = q \times \exp(c + xv)$$

This is more complicated for models with age modifiers on the period trend, as these are not purely linear. For this, first let z be x with just the age and cohort variables, and let t be the design matrix with the period variables. Let A be a spline design matrix for all the age variables, including the first age. Now v is the vector of just the age and cohort parameters. Let u be the corresponding column of period parameters, so tu is the vector of period factors by observation, and let w be another column of age parameters for the age weights. Then set $r = Aw$, the raw age weights on trend for each observation. These need another constraint for identifiability, so let $m = \max(r)$ and set $p = \exp(r - m)$. These are the final age weights on trend for each observation, and are positive with a maximum of 1. With all of this, the mean is now:

$$\mu = q \times \exp(p \times tu + c + xv).$$

For the contagion model, with mortality multipliers by year, another design matrix B is needed with 0,1 variables for each year, including the first year, to pick out the year each observation is from. This is multiplied by the column h of log age weights selected from the gamma distribution, and then added to the log mortality. Then:

$$\mu = q \times \exp(Bh + p \times tu + c + xv)$$

5.4.2 *Fitting Process*

The overlaps among the three directions in APC models require additional constraints. In parameter shrinkage, some parameters are shrunk to zero, which takes those variables out of the model, and this is enough of a constraint to keep the design matrix from being singular. The parameters then are not readily interpretable as to what the period and cohort trends mean individually, but since they come from finding the most predictive model, it sometimes could be reasonable to take them at face value. That would be a place for actuarial judgement. Some argue that since the trends are not separately identifiable without constraints, there are no real individual effects. But there are drivers of the trends in other data. For instance, health-related behaviours, like diet, smoking, and exercise, tend to be more cohort effects than period effects. Change in demographic mix can also affect mortality, and this can have both period and cohort effects driven by immigration and demographic differences in family size.

The models in the example use linear splines, but most of the fitting procedures below work for cubic splines as well. Limited experience suggests that more parameters go to zero with cubic splines, making those models somewhat more parsimonious. That could lead to better or worse loo fit measures.

MCMC sample sets rarely have any parameters that are zero in every sample, but it can simplify the model and improve loo a bit by eliminating some of the small ones. Setting parameters to zero is like setting variables with low t-statistics to zero in regression models. That is done in regression to eliminate parameters that have a good chance of actually being zero. But in spline models, nearby parameters can be highly negatively correlated, and can have opposite signs and move simultaneously. So a parameter having zero in its sample range could still be contributing to the model. The rule followed in the fitting here was to eliminate parameters with absolute values of 0.001 or less, as long as they had ranges approximately symmetric around zero. But if that reduced the loo fit measurement, they were put back in. Sometimes offsetting adjacent parameters were both eliminated in this way.

MCMC can have difficulties in finding good parameter sets when there is a large number of highly negatively-correlated variables. This can be shortcut by first using lasso with minimal shrinkage to eliminate the least-needed variables, with the rest used in MCMC. Generally the fitting procedure above will eliminate more of these. The glmnet R package for lasso is very efficient, and includes a cross-validation routine that can indicate the minimal reasonable degree of shrinkage. Glmnet has

some difficulties with the cubic spline design matrix, and may look too parsimonious. If so, lasso can easily be fit for still lower selected values of λ by a nonlinear optimiser.

Glmnet has a Poisson option that uses a log regression and exposure multipliers. The approach taken was to first use this just on age and period variables. After eliminating some of those, the cohort variables were added in lasso to give an APC model. It is not possible to input all the APC variables initially, as the design matrix is singular until some variables are eliminated. From the initial set of 146 age, period, and cohort variables, 50 made it into the female model, with 45 for the male model. Then the APC models were fit in MCMC.

The RH model was fit with the surviving variables from the APC fit, along with all the age variables for the age weights. Poisson and negative binomial versions of the RH model were fit to both populations. The negative binomial fit better for both, and had 65 variables in the female model and 48 in the male. But due to shrinkage, the female model had fewer effective parameters, according to the loo penalty. Those parameters were then used in the contagion model, with 40 additional annual jump variables, some ending up negligible.

The individual-year log jump factors in the contagion model combined can pick up some of the overall trend. Steps were taken here to inhibit that, so that they could be interpreted as jumps with no trend in them. For instance, using a low gamma a parameter pushes the jumps more towards zero, which makes it less likely that they would pick up larger trends. Also the period variable for age 1 was put into the model. It had been left out initially for identifiability reasons, but with other parameters no longer in the model, that was not necessary. Even with higher values of the gamma a , if a trend appears in the jump parameters, the jumps could still be displayed as residuals to a line fit through them. The interaction of this with the other trends could produce an overall better model in some cases, but that would have to be tested by trying different values of the gamma a . Also these parameters together could interact with the constant term. The gamma shrinkage here was enough to prevent that, but one of them could be left out as well to prevent such overlap. That could not be any of them, though, as they all are positive, with some close to zero and some large. The year with the lowest residuals to the RH model would be a logical choice to leave out in this situation.

5.5 Results

Table 5.1 summarises the fit statistics, and Fig. 5.2 graphs the resulting parameter curves for the event jump model. Loo is a penalised loglikelihood, so higher is better. It comes with a degree-of-freedom type penalty. In a good model, the penalty will be not too different from the parameter count, and ideally will be less, due to parameter shrinkage. Being much higher is an indication that the model is not working well. Probably some sample parameter sets work well for some of the observations and other sets for other observations. This would make the loo measure worse.

Table 5.1 Fit summary statistics

	Female			Male		
	Loo	Penalty	Parameters	Loo	Penalty	Parameters
APC Poisson	−9420	155	42	−9998	133	37
RH Poisson	−9053	181	78	−10029	142	73
APC NB	−8328	24	33	−8830	25	33
RH NB	−8191	34	32	−8750	29	50
Event jump Poisson	−7912	103	106	−8818	125	95

Fig. 5.2 Fitted parameter curves

The Poisson APC and RH models do not fit well for either population. The negative binomial versions have much better loo measures. The standard deviation/mean gives an indication of how much variability from the data is allowed for the model means. With this data, for the Poisson that can be as small as 1% for the largest cells. This widens to about 2.5% for the negative binomial fitted model. It is more like 3% for both distributions for the smaller cells. This forces the Poisson models to fit very tightly at the largest cells.

The models all fit better in the female population. For the males, the APC model fits better than the RH model for Poisson residuals. The event jump model is clearly the best fit for the female data, even though it uses quite a few more parameters. For the males, it is by far the best Poisson model, but it is not quite as good as the negative binomial RH model. Thus there are still some systematic elements that this model is not picking up. The trend is small for the male population, and it does not affect most ages, as seen in the top two graphs in Fig. 5.2. Leaving out the trend and trend weight variables might seem to save parameters without hurting the fit much. This turns out not to be the case. The trend and weight variables actually do not

use many parameters. The entire trend for males shows up in the cohort parameters and so does not apply to the older ages. For females, the time trend is similar to the male cohort trend, but is somewhat offset by an increase for cohorts after 1940. The journal NIUSSP has done various related studies.² They find higher smoking rates in France than in most countries, and this increased substantially for females in recent generations, particularly for those with lower education levels. It has come down a bit for males.

5.5.1 *Extensions: Generalisation, Projections and R Coding*

The event-jump model here is very simple. All the ages get the same proportional increase in mortality. A generalisation would be to have the excess deaths follow a parametric age curve, like Gompertz or others, as in Perks (1932). This would have 2–4 parameters. These parameters could even be put on semiparametric curves and allowed to evolve over time. As they are positive, their logs could be postulated to follow such curves. They might be slowly evolving, not using many change parameters, with occasional larger shifts. The population events themselves tend to extend for a few years, perhaps due to weather conditions or diseases. It might be more efficient to put the logs of the jump parameters on semiparametric curves to fit this kind of jump correlation and to save parameters. The gamma distribution for jumps has heavier-tailed options still with mode zero, such as even lower a , or heavier-tailed distributions like the log-logistic or its generalisation, the Burr.

Projected future mortality risk scenarios would include the event jump parameters. Each simulation would draw a jump for each year from the gamma distribution. MCMC program output is good for the parameter risk part of the simulation, as all the sampled parameter scenarios are available. These come with built-in dependence among the different parameters. Future projections of the trend and cohort parameters can be generated from the double-exponential distributions for the slope changes. This should use the correlation matrix of the parameters, which can be computed from the scenario output. The normal copula would be a convenient way to simulate these parameters. Also, putting the jumps on semiparametric curves would give the simulations groups of years that are fairly high simultaneously, which would be more realistic if simulated losses in nearby years is a risk of concern. Compared to other jump models, an advantage of this one is that the low-median high-mean jump distribution produced reasonable probabilities for extreme outcomes even though it did not have extreme events in the historical data used.

R uses the command “fit_ss = extract(fit, permuted = FALSE)” to get the samples from a Stan run with an output called “fit,” and put them into an object called “fit_ss.” MCMC produces samples that are serially correlated, so some users prefer to permute them before using them. This does not help with most analyses, however. The non-permuted option in the extract command gives a convenient three-dimensional

² E.g., see <https://www.niussp.org/category/article/?postyear=2017>.

output: sample parameters by variable for each chain. These can be output to disk by `write.csv(fit_ss, file="samples_fit.csv")`. It is often more convenient to do this one chain at a time to preserve the variable headings. Some sample code is included in the Appendix.

5.6 Conclusions

This chapter introduced a mortality model with a jump term for contagion to fit pandemic mortality experience as well as smaller population-mortality events. It was estimated by fitting Poisson and negative binomial distributions to mortality counts with known exposures, using a log link function, and was found to be an improvement to standard continuous mortality models. The models were all fit by semiparametric regression in MCMC, with shrinkage distributions for each parameter.

The number of deaths for a homogeneous group will be binomially distributed if they are independent. This can be closely approximated by a Poisson. If model residuals are heavier-tailed, like negative binomially, some systematic drivers are missing in the model. This was found to be the case for French female and male populations. One reason could be population mortality events in some periods. These were added to the models as gamma-distributed annual mortality jumps that affect mortality for all ages proportionally. This made the female Poisson model better than the previous negative binomial model, and improved the male Poisson models but there were apparently some mortality drivers still missing for males. Possible extensions were discussed. The gamma distribution used for the jumps is wide enough to capture pandemic effects such as COVID-19 in 2020.

Semiparametric regression with linear smoothing splines is a way to improve models by reducing overfitting. This worked well here, and gave simple parameter curves across all the variable dimensions. An advantage of including the event jumps is that simulated risk scenarios can include their probabilities. It is possible that using mortality trend without shrinkage could pick up all the historical jumps, but that would lose the advantages of shrinkage and could be more difficult to project in a way that would keep this risk element.

Appendix

First is R code for running Stan for the RH model plus the event jumps, followed by the Stan code. Simpler models would use only some of these lines.

The R code starts by loading in the needed packages, and then the data. This includes columns for the deaths and exposures, and then the various design matrices. The data is combined in a list file to pass to Stan, then Stan is run. Loo is run on the output, and parameters printed and sent out to disk.


```

setwd("~/OneDrive/R/Mortality/Contagion")
library("loo")
library(rstan)
rstan_options(auto_write = TRUE)
options(mc.cores = parallel::detectCores())
library(readxl)

y = as.integer(scan('Le_mort.txt'))
expo = scan('Le_expo.txt')
#scan turns a column file into a vector
x = as.matrix(read_excel("Le_AC_nb.xlsx"))
#age and cohort only
x = x[,-c(4,5,10,14,22)]
#eliminates some variables in later runs
trendmat = as.matrix(read_excel("Le_Y_nb.xlsx"))
#year for trend
wghtm = as.matrix(read_excel("age_all.xlsx"))
#ages for weights
yearm = as.matrix(read_excel("Y01.xlsx")) #0,1 trend matrix
wghtm = wghtm[,-c(5,10,16,17,22,23,31,34)]
trendmat = as.matrix(trendmat[,-c(2,3)])
yearm = yearm[, -1]
n <- ncol(trendmat)
u <- ncol(wghtm)
t <- ncol(yearm)
N = length(y)
U = ncol(x)
c(N,U,n,u,t)
dx = list(N=N,U=U,expo=expo,y=y,x=x,n=n,u=u,t=t,
          trendmat=trendmat,wghtm=wghtm,yearm=yearm)
# this is to send input data to MCMC model
# now run model
set.seed(77)
Sys.time()
fit = stan(file = 'rh_mort_pois_double.stan',data=dx,
           verbose = FALSE, chains = 3, iter = 20000, warmup =
             15000, control = list(max_treedepth = 12))
Sys.time()
#now compute loo
log_nb1 <- extract_log_lik(fit)
loo_nb1 <- loo(log_nb1)
loo_nb1
#show parameters and write means to disk
print(fit, pars=c("cn","b", "v","w", "d","s","m2"),probs=c(.025,
                  0.25, 0.5, 0.75, 0.975), digits_summary = 3)
out <- get_posterior_mean(fit)
write.csv(out, file="out.csv")

```

Now the Stan code. This is translated into C++ then compiled for more efficient execution. Most of it is just defining the variable types and dimensions for C++. It comes in required code blocks. The means are computed in the transformed-parameters block, and the model block gives the distributions assumed for the data

and the parameters. The last block computes the log likelihoods for use in the `loo` code later.

```
data {
  int N;      // number of obs
  int U;      // number of age+cohort variables
  vector[N] expo;
  int y[N];
  matrix[N,U] x; //age and cohort matrix
  int n;      //number of trend params
  int u;      //number of wghts params
  int t;
  matrix[N,n] trendmat; //trendmat design matrix
  matrix[N,u] wghtm;    //age wght dsgn mat
  matrix[N,t] yearm;    //year 0,1 dummy dsgn mat
}
parameters { // all except v will get default uniform prior
  real<lower=-8, upper=8> cn;    //constant, in assumed range
  real<lower=-6, upper = 6> logs; //log of s, related to lambda
  real logb; //log of gamma b parameter
  vector[U] v; //age and cohort parameters
  vector[u] w; //trend-wghts parameters
  vector[n] d; //trend parameters
  vector<lower=0, upper = 1>[t] m2; //factor by age
}
transformed parameters {
  real s; // shrinkage parameter
  vector[N] mu;
  real m; // max weight
  real b; // gamma b for factor
  vector[N] colwghts;
  vector[N] trend;
  vector[N] mean2nd;
  s = exp(logs);
  b = exp(logb);
  mean2nd = yearm*m2; // 2nd mean by observation
  colwghts = wghtm*w;
  m = max(colwghts);
  colwghts = exp(colwghts-m); //makes max = 1, all >0
  trend = trendmat * d;
  for (j in 1:N) trend[j] = trend[j]*colwghts[j];
  mu = exp(trend+x*v+cn+mean2nd);
  for (j in 1:N) mu[j]=mu[j]*expo[j];
}
model { // gives priors for those not assumed uniform.
  Choose this one for lasso.
  for (i in 1:U) v[i] ~ double_exponential(0, s); //
  more weight to close to 0
  for (i in 1:u) w[i] ~ double_exponential(0, s);
  for (i in 1:t) m2[i] ~ gamma(0.1, b);
  y ~ poisson(mu);
}
```

```

generated quantities { //outputs log likelihood for
  testing purposes  vector[N] log_lik;
  for (j in 1:N) log_lik[j] = poisson_lpmf(y[j] | mu[j]);
}

```

References

- M.A.C. Alijean, J. Narsoo, Evaluation of the Kou-modified Lee-Carter model in mortality forecasting: Evidence from French male mortality data. *Risks* **6**(123) (2018)
- K. Barigou, S. Loisel, Y. Salhi, Parsimonious predictive mortality modelling by regularization and cross-validation with and without Covid-type effect. *Risks* **9**(5) (2021)
- G. Barnett, B. Zehnwirth, Best estimates for reserves. *Proc. Casualty Actuarial Soc.* **87**, 245–303 (2000)
- A.J.G. Cairns, D. Blake, A.R. Kessler, M. Kessler, The impact of Covid-19 on future higher-age mortality. Pensions Institute (2020). <http://www.pensions-institute.org/wp-content/uploads/wp2007.pdf>
- H. Chen, S.H. Cox, Modeling mortality with jumps: applications to mortality securitization. *J. Risk Insur.* **76**(3), 727–751 (2009)
- A.E. Gelfand, Model determination using sampling-based methods, in *Markov Chain Monte Carlo in Practice*, ed. by W.R. Gilks, S. Richardson, D.J. Spiegelhalter (Chapman and Hall, London, 1996), pp. 145–162
- R.J. Harezlak, D. Rupert, M.P. Wand, *Semiparametric Regression with R* (Springer, 2018)
- T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning* (Springer, Corrected 12th Printing, 2017)
- A.E. Hoerl, R. Kennard, Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970)
- Human Mortality Database, University of California, Berkeley (USA), Max Planck Institute for Demographic Research (Germany) and United Nations (2019). <https://www.mortality.org/>
- A. Hunt, D. Blake, A general procedure for constructing mortality models. *North Am. Actuarial J.* **18**(1), 116–138 (2014)
- R. Lee, L. Carter, Modelling and forecasting U.S. mortality. *J. Am. Stat. Assoc.* **87**, 659–675 (1992)
- A.H. Mowbray, How extensive a payroll exposure is necessary to give a dependable pure premium? *Proc. Casual. Actuarial Soc.* **1**, 24–30 (1914)
- C. O'hare, Y. Li, Models of mortality rates—analysing the residuals. *Appl. Econ.* **49**(52), 5309–5323 (2017)
- S. Özen, Ş Şahin, Transitory mortality jump modelling with renewal process and its impact on pricing of catastrophic bonds. *J. Comput. Appl. Math.* **376**(1) (2020)
- W. Perks, On some experiments in the graduation of mortality statistics. *J. Inst. Actuar.* **63**(1), 12–57 (1932). <https://www.cambridge.org/core/journals/journal-of-the-institute-of-actuaries/article/on-some-experiments-in-the-graduation-of-mortality-statistics/F3347914B8A72BAECF3AFD3D95E5FF50>
- A.E. Renshaw, S. Haberman, A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insur. Math. Econ.* **38**, 556–570 (2006)
- F. Santosa, W.W. Symes, Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Stat. Comput.* **7**(4), 1307–1330 (1986)
- C. Stein, Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp.* **1**, 197–206 (1956)
- R. Tibshirani, Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc. Ser. B (Methodol.)* **58**(1), 267–288 (1996)
- A.N. Tikhonov, On the stability of inverse problems. *Doklady Akademii Nauk SSSR* **39**(5), 195–198 (1943)

- A. Vehtari, D. Simpson, A. Gelman, Y. Yao, J. Gabry, Pareto smoothed importance sampling (2019). <https://arxiv.org/abs/1507.02646>
- G. Venter, Ş Şahin, Parsimonious parametrization of age-period-cohort models by Bayesian shrinkage. *Astin Bull.* **48**(1), 89–110 (2018)
- G. Venter, Ş. Şahin, Semiparametric regression for dual population mortality. *North Am. Actuar. J.* (2021). <https://doi.org/10.1080/10920277.2021.1914665>
- Y. Xu, M. Sherris, J. Ziveyi, Continuous-time multi-cohort mortality modelling with affine processes. *Scand. Actuar. J.* **2020**(6), 526–552 (2020)
- R. Zhou, J.S.-H. Li, K.S. Tan, Pricing standardized mortality securitizations: a two-population model with transitory jump effects. *J. Risk Insur.* **80**(3), 733–774 (2013)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.