A Principled Approach to Model Validation in the Context of Estimating Age-Specific Rates

Ameer Dharamshi[1], Daniela Witten[1,2], and Monica Alexander[3,4]

[1]Department of Biostatistics, University of Washington
[2]Department of Statistics, University of Washington
[3]Department of Statistics, University of Toronto
[4]Department of Sociology, University of Toronto

October 31, 2023

Abstract

A number of statistical models are available to estimate age-specific demographic rates in a wide range of populations, partcularly in contexts where the data are sparse or unreliable. However, model development and testing has largely focused on in-sample validation, with less emphasis on model validation on new data. Validation using classical train/test sample splitting approaches is not appropriate in the context of estimating age-specific rates, as demographic outcomes across age are not independent and identically distributed. In this paper, we show that data thinning, a collection of randomization strategies used in the statistical learning literature, can be used to generate independent train and test sets by partitioning the population of each individual age group into two (or more) subpopulations subject to the same latent risks. As all age groups are represented in both folds, model fitting and testing replicate the intended use of the model, that is, whether the latent risk curve extracted from the observed data generalizes to an independent realization from the same age groups. We illustrate the advantages of data thinning over sample splitting in a simulation study based on US state-level mortality data.

1 Introduction

Obtaining reliable estimates of age-specific demographic rates is challenging in many populations, as the available data to estimate such rates may be sparse, noisy, or unreliable. In these contexts, demographers have developed a wide range of statistical models to estimate the underlying latent risk of a particular demographic event by age. For example, recent research has used techniques such as principal components (Dharamshi et al., 2023a), penalized splines (Camarda, 2019), TOPALS models (Dyrting, 2020), and other parametric approaches (Schmertmann, 2003) in applications across outcomes such as mortality, fertility, migration, and marriage age-specific rates.

While statistical models to estimate and project age-specific demographic outcomes are widely available, there has been less focus on methods to validate the performance of such models. Many authors evaluate the performance of their method compared to other extant methods by assessing in-sample measures of model fit (see for example, Bohk-Ewald et al. (2018); Camarda (2019); Denecke et al. (2023)). However, there is a lack of emphasis on out-of-sample model validation, particularly in the case of estimation (rather than projection). Classical approaches from statistical learning focus on sample splitting; that is, randomly splitting the available data into a training dataset, which is used to estimate model parameters, and a test dataset, which is used to validate how well the model performs on new data. However, these approaches are not suitable in many demographic contexts, particularly when producing estimates by age, year, geographic area (and

potentially more subgroups), *because only one observation of the outcome is available per combination of population characteristics.*

To understand the problem, note that we are interested in estimating a collection of parameters θ_i for $i = 1, \ldots, n$, where θ_i represents some demographic rate in the ith population. We could consider using some of the populations as a training set and holding the rest out as a test set. But there is problem: by doing this, we ensure that some of the θ_i's are not present in the training set at all, i.e. we cannot possibly estimate all of them well, without unreasonably strong assumptions about smoothness of the θ_i's across groups. Furthermore, sample splitting departs from the validation question of interest: sample splitting tells us how a model would perform on populations not seen in the original study, whereas our interest is in the performance of the model on an independent realization of the populations observed in the original study.

The problem of validating a model in settings where sample splitting cannot be applied is not unique to demography; it also arises in problems such as clustering (Gao et al., 2022; Chen and Witten, 2022). This has led to a recent line of work on randomization as a strategy to split a *single* observation into two independent components. In a seminal paper, Rasines and Young (2022) demonstrate that for Gaussian data, by injecting independent external Gaussian noise, one can decompose a single observation into two independent observations that sum to recover the original. Models can then be fit on one set of observations and validated on the other. This can be performed even in settings where sample splitting is not applicable. Neufeld et al. (2023) and Dharamshi et al. (2023b) show that similar randomization strategies are available for a wide class of distributions. Notably, their "data thinning" proposal includes the count-valued binomial and Poisson distributions.

In this paper, we will show that the binomial and Poisson data thinning strategies proposed by Neufeld et al. (2023) and Dharamshi et al. (2023b) are particularly useful in the context of estimating age-specific demographic rates. By decomposing demographic count data into two or more independent data sets, data thinning unlocks a principled cross-validation algorithm for validating demographic estimation methods. In this setting, these tools have elegant interpretations rooted in sampling theory. Specifically, one can understand data thinning as a statistical way of turning an enumeration over a population into multiple independent enumerations over smaller populations subject to the same risks. In the context of model validation, each distinct enumeration can then be used for either model fitting or validation. This is in contrast to sample splitting, where we are forced to reserve some age groups for validation, which is inconsistent with the goals of our model: we are interested in constructing models that can recover age-specific rates from an enumeration, not from a censored dataset.

The rest of this abstract is organized as follows. In Section 2, we formally define binomial and Poisson data thinning, and interpret the associated algorithms in the context of estimating demographic rates. In Section 3, we illustrate the advantage of data thinning over classical sample splitting in a simulation study based on US state-level mortality data. We conclude with a brief discussion in Section 4.

2 Data thinning

Data thinning is a strategy for decomposing an individual random variable X into two or more independent random variables $X^{(1)}, \ldots, X^{(K)}$ that can together reconstruct the original (Neufeld et al., 2023; Dharamshi et al., 2023b). A formal definition of data thinning is given in Definition 1.

Definition 1 (Data thinning)**.** *Suppose that $X \sim P_\theta$ where P is a distribution parameterised by an unknown parameter θ. Consider the distribution G_x defined as the conditional distribution $(X^{(1)}, \ldots, X^{(K)})|X = x$. Suppose that G_x does not depend on θ and that upon sampling $(X^{(1)}, \ldots, X^{(K)})$ from G_x, the following hold:*

1. *$X^{(1)}, \ldots, X^{(K)}$ are mutually independent,*
2. *For $k = 1, \ldots, K$, $X^{(k)} \sim Q_\theta^{(k)}$ where $Q^{(k)}$ is some distribution parameterised by θ, and*
3. *X can be reconstructed from $X^{(1)}, \ldots, X^{(K)}$ by some deterministic function T (ie. $T(X^{(1)}, \ldots, X^{(K)}) = X$).*

Then, we say that P has been thinned by T.

Intuitively, data thinning starts with one data point, and breaks it into K pieces. The new data points divide the information about the unknown parameter θ contained in the original. Because the K new data points are independent, we are free to perform one task using the first new data point, another using the second, and so on, without recycling information across tasks.

Extending to a data vector of length n, suppose that for $i = 1, \ldots, n$, $X_i \sim P_{\theta_i}$ are independent, but not necessarily identically distributed data points. If we apply K-fold data thinning entry-by-entry to the entire data vector, we will produce K independent data vectors of length n where in the kth vector, $X_i^{(k)} \sim Q_{\theta_i}^{(k)}$. Once again, we are free to perform one task on the first vector, a second task on the second vector, and so on. If we choose $K = 2$, we can fit a model to the first vector and perform validation on the second. Because all of $\theta_1, \ldots, \theta_n$ are represented in each data vector, we can draw conclusions about model performance at the individual parameter level. This stands in stark contrast to classical model validation via sample splitting, which fails in this setting, because (i) it requires that $X_1, \ldots, X_n$ are both independent and identically distributed (Cox, 1975); and (ii) even if we ignore the violation of the previous assumption, the training and test set would each contain only a subset of the parameters $\theta_1, \ldots, \theta_n$, and so it would not be possible to fit and validate a model involving all of the parameters of interest.

The case of independent, but non-identically distributed data is common in demographic estimation. Frequently, we observe counts of some demographic quantity across a variety of ages, periods, and locations. These counts are often treated as independent conditioned on the unobserved risk of the corresponding population. As the risks vary across populations, the counts follow different distributions. While there are many proposed methods for estimating latent risks given observed demographic counts, there are few proposals on how to validate these models that acknowledge that the data are non-identically distributed.

The rest of this section will focus on specific data thinning algorithms for common count-valued distributions that can be used to perform this validation task. We will use mortality as a running example, though other target quantities could be used instead. Notationally, we will say that D_i are the deaths corresponding to population P_i in the age-period-location indexed by i. This population is subject to the unknown risk θ_i, our target of estimation.

2.1 Binomial distribution

Suppose that we choose to model deaths with a binomial random variable $D_i|\theta_i \sim \text{Bin}(P_i, \theta_i)$. If the population size P_i is known, but the probability of death θ_i is not, then Algorithm 1 can be used to thin D_i.

Algorithm 1 (Binomial thinning)**.** *Suppose that $D_i|\theta_i \sim Bin(P_i, \theta_i)$ where P_i is known and we wish to thin D_i into K folds. Perform the following:*

1. *Choose a vector $(\epsilon_1, \ldots, \epsilon_K)$ on the unit K-simplex such that for all $k \in 1, \ldots, K$, $\epsilon_k P_i$ is an integer.*
2. *Sample $(D_i^{(1)}, \ldots, D_i^{(K)})|D_i = d \sim MultivariateHypergeometric(\epsilon_1 P_i, \ldots, \epsilon_K P_i, d)$.*

Then, $D_i^{(1)}, \ldots, D_i^{(K)}$ will be mutually independent, $D_i^{(k)}|\theta_i \sim Bin(\epsilon_k P_i, \theta)$, and $\sum_{k=1}^K D_i^{(k)} = D_i$. Furthermore, the proportion of Fisher information about θ_i contained in $D_i^{(k)}$ is ϵ_k.

Algorithm 1 starts with one binomial observation and produces K independent binomial observations, each of which represents a fraction of the original population. The intuition behind Algorithm 1 lies in step 2. The multivariate hypergeometric distribution describes the process of sampling without replacement. Applied to the scenario with binomial distributed deaths D_i in a population P_i, this distribution first divides the total population into K subpopulations of size $\epsilon_k P_i$ for $k = 1, \ldots, K$. Then, one-by-one, each of the D_i deaths are allocated to one of the K subpopulations with probability proportional to the population in each category less the number of deaths previously allocated.

It is important to note that in Algorithm 1, the risk of death in each new artificial subpopulation is θ_i. That is, the algorithm simply partitions the original population carefully such that the risk of death

is preserved (where we emphasize that the risk of death, θ_i, is unknown). It is as if before the original enumeration over the population, the data collection agency randomly assigned each person to one of the K groups. This means that any model that would have been fit to D_i can instead be fit to $D_i^{(k)}$. Similarly, any estimate of θ_i produced on one $D_i^{(k)}$ can be validated on another, as these quantities are independent.

2.2 Poisson distribution

A similar data thinning strategy for the Poisson distribution can be applied when modelling deaths as $D_i|\theta_i \sim \text{Poisson}(P_i\theta_i)$. We describe Poisson data thinning in Algorithm 2.

Algorithm 2 (Poisson thinning)**.** *Suppose that $D_i|\theta_i \sim Poisson(P_i\theta_i)$, and we wish to thin D_i into K folds. Perform the following:*

1. *Choose a vector $(\epsilon_1, \ldots, \epsilon_K)$ on the unit K-simplex.*
2. *Sample $(D_i^{(1)}, \ldots, D_i^{(K)})|D_i = d \sim Multinomial(d; \epsilon_1, \ldots, \epsilon_K)$.*

Then, $D_i^{(1)}, \ldots, D_i^{(K)}$ will be mutually independent, $D_i^{(k)}|\theta_i \sim Poisson(\epsilon_k P_i \theta_i)$, and $\sum_{k=1}^K D_i^{(k)} = X$. Furthermore, the proportion of Fisher information about θ_i contained in $D_i^{(k)}$ is ϵ_k.

Similar to Algorithm 1, Algorithm 2 takes one Poisson observation D_i and produces K independent Poisson observations, $D_i^{(1)}, \ldots, D_i^{(K)}$, that each represent the deaths in a population of $\epsilon_k P_i$ individuals with risk θ_i. Once again, the intuition behind Algorithm 2 comes from sampling theory. The multinomial distribution randomly allocates each death to one of the subpopulations with probability proportional to the total subpopulation size, given by $\epsilon_k P_i$. As this process preserves the risk in the artificial subpopulations, and as the observed number of deaths in the subpopulations are independent, we can again fit a model to one $D_i^{(k)}$ and validate it on another.

3 Simulation Study

To illustrate the application of data thinning to model validation, we conducted a simulation study based on US state-level mortality data. Our goal is to demonstrate that due to the lack of uniformity in age-specific mortality rates, sample splitting leads to poor estimates on the training set, specifically in the age groups that were held out. This, in turn, leads to unreliable model validation. By contrast, if we apply data thinning, then both the training and test sets include information for all age groups.

We will use a principal component-based modelling framework for our simulations. This modelling approach constructs latent mortality risks on the log-scale as a linear combination of principal components generated from set of standard log-mortality curves. The principal components capture standard patterns in mortality, and the model seeks to determine how best to aggregate these patterns to fit new data. The use of such models in mortality estimation is discussed in Alexander et al. (2017) and Dharamshi et al. (2023a).

We used data from the United States Mortality Database (USMDB) to construct an $N \times A$ matrix, X, of state-level mortality curves spanning 60 years (United States Mortality Database, 2023). Here, $N = 3057$ is the number of state-years, and $A = 19$ is the number of age groups (<1, 1-4, 5-9, ..., 80-84, 85+). We then took the singular value decomposition, $X = U\Sigma V^\top$, and extracted the first four principal components, $V_1, \ldots, V_4$, from the columns of V. The first principal component captures the characteristic "J"-shape of a log-mortality curve, the second contains an "accident hump", and the third and fourth contain other localized mortality patterns.

To generate simulated data, we took the first four left singular values for New Hampshire in 2016 and constructed a log-mortality curve by rescaling them by the first four singular values and then multiplying by the first four principal components. This gives us a vector of length $A = 19$, comprised of elements $\theta_1, \ldots, \theta_{19}$, which will serve as our "true" log-mortality rates in this simulation. We simulated deaths for

Age	SampleSplit	DataThin	Full	Age	SampleSplit	DataThin	Full
<1	0.2650	0.1883	0.1617	45-49	0.0866	0.083	0.0714
1-4	0.4012	0.3296	0.2822	50-54	0.0757	0.0716	0.0616
5-9	0.2174	0.1999	0.1742	55-59	0.0554	0.053	0.0455
10-14	0.2709	0.2519	0.2187	60-64	0.0421	0.04	0.0341
15-19	0.2704	0.2522	0.2185	65-69	0.0369	0.0345	0.0297
20-24	0.191	0.1798	0.1551	70-74	0.0365	0.0339	0.0294
25-29	0.1504	0.1447	0.1242	75-79	0.0334	0.0312	0.0273
30-34	0.1282	0.1255	0.1077	80-84	0.0284	0.0253	0.0223
35-39	0.1127	0.1104	0.0948	85+	0.0212	0.0178	0.0152
40-44	0.0979	0.0952	0.0818				

Table 1: Root mean squared errors for estimated log-mortality rates when using sample split training sets, data thinning training sets, and the full dataset. Data thinning produces more reliable log-mortality rates than sample splitting with the same proportion of information.

each age group D_a, where a refers to the age group, using a Poisson likelihood with a population offset of 10,000.

For each dataset, we will sample split by taking 14 of the 19 observed deaths as a training set and leaving the remaining 5 as a test set. We will also data thin by applying Algorithm 2 to each age group's observed deaths with $K = 2$ and $(\epsilon_1, \epsilon_2) = (14/19, 5/19)$, and use $D_a^{(1)}$ as the training set and $D_a^{(2)}$ as the test set. We will then estimate the mortality rate from the training sets using a Poisson generalized linear model with the first four principal components (computed on the entirety of the USMDB data) as the covariates. That is, we fit the following model:

$$D_a|\theta_a \sim \text{Poisson}(10,000\theta_a), \tag{1}$$

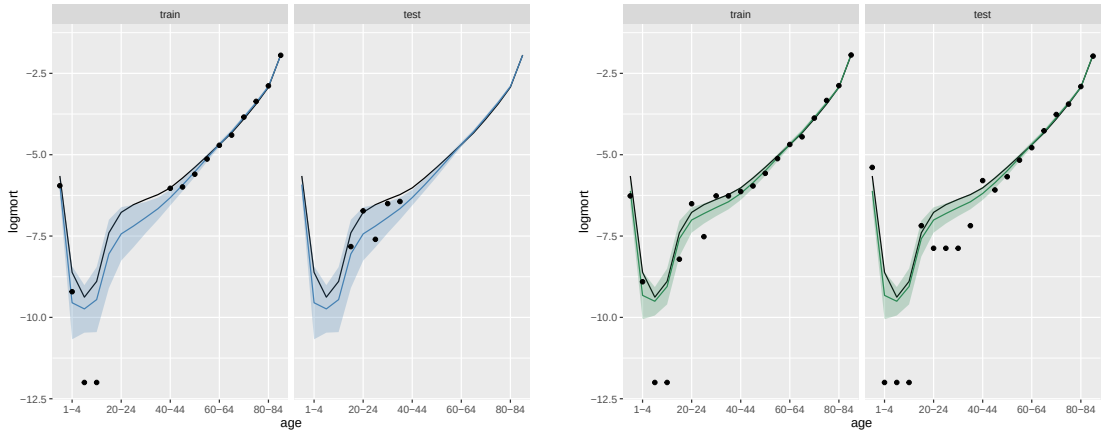
$$\log\theta_a = V_{1a}\beta_1 + V_{2a}\beta_2 + V_{3a}\beta_3 + V_{4a}\beta_4. \tag{2}$$

Note that this model exactly matches the data generating process. Our goal is to see whether sample splitting or data thinning better recover θ_a from their respective training sets. Performance will be evaluated by comparing the estimates $\hat{\theta}_a$ against the truth θ_a and computing root mean squared error (RMSE) across simulations. Nominally, the amount of information in both training sets is the same, but we expect that estimates of θ_a will be better when using the data thinning training set as all age groups are represented, thereby accounting for the non-identical mortality rates.

Table 1 presents the results of this experiment. We have also included the RMSE when fitting the model using the full dataset as a baseline. We can see that as expected, for all age groups, errors are greatest for sample splitting, lower for data thinning, and lowest for the full dataset. The difference between sample splitting and data thinning is particularly acute in the young ages. Intuitively, this is because young age mortality varies in a non-linear fashion, and can vary widely across different populations. As such, in the event sample splitting excludes a young age from the training set, estimates of the excluded age group will be unreliable.

Figure 1 provides an example of this phenomenon. In the plots, the black line represents the true log-mortality rates, the black points are the observed training and test set log-mortality rates (with zero death observations encoded as -12), and the coloured lines and intervals are the estimated log-mortality rates and 95% uncertainty intervals. In Facet 1a, when sample splitting is given a training set that excludes young adults, it greatly underestimates young adult mortality as no information on the magnitude of the accident hump is seen by the model. By contrast, in Facet 1b, the data thinning training set consists of noisy versions of the observed data for all age groups. Thus, while each data point has less information than the original, some information on the degree of the accident hump is seen by the model, leading to better estimates.

As mentioned in the introduction, sample splitting emulates a process in which we are given a realization of new age groups not seen in the original data; as we can see from our results in Table 1, this is a very



(a) Sample Splitting

(b) Data Thinning

Figure 1: Comparison of sample splitting and data thinning. The black lines are the true log-mortality curves, the black dots are the training and test observations under each method, and the blue and green curves/regions are the estimated log-mortality curves and uncertainty intervals using the training data under sample splitting and data thinning respectively. With sample splitting, when key ages are allocated to the test set, estimates in the training set are poor as no information on local features of the mortality curve are used in training. In data thinning, some information on all parts of the mortality curve are used in both training and testing.

challenging setting! But it is not the setting of interest when estimating age-specific rates. By contrast, data thinning emulates a process in which we are given an entirely new population composed of the same set of age groups; we see from Table 1 that this setting is much easier; furthermore, this is our sampling frame of interest. Most importantly, Table 1 highlights the fact that we cannot use the age group sampling frame inherent to sample splitting to approximate the population sampling frame inherent to data thinning, as the former leads to estimates that are much less accurate than the latter.

4 Discussion & Next Steps

In this paper, we show that data thinning, a recent proposal in the statistical learning literature, can be used to validate model-based estimates of age-specific demographic rates. Unlike classical sample splitting, data thinning splits the information in each observed count, as opposed to discretely allocating individual data points to either the training set or the test set. This leads to a principled strategy for validating the ability of a model to recover important age-specific patterns. Furthermore, it allows for demographers to draw conclusions about the performance of a model across all ages, instead of just on the ages held out in the validation set. This is noteworthy as for most quantities, model performance likely varies by age. For example, young-age mortality often exhibits both small death counts and highly non-linear patterns, which make it difficult to estimate.

In the full paper, we will expand both the methodological discussion and the applications. For the former, we will discuss data thinning for more complex models than simple GLMs, as well as the consequences of model misspecification. For the latter, we will expand the simulation study to include more data types, and will add a comprehensive real data case study to showcase how data thinning should be applied by practicing demographers.

References

Alexander, M., Zagheni, E., and Barbieri, M. (2017). A flexible bayesian model for estimating subnational mortality. *Demography*, 54(6):2025–2041.

Bohk-Ewald, C., Li, P., and Myrskylä, M. (2018). Forecast accuracy hardly improves with method complexity when completing cohort fertility. *Proceedings of the National Academy of Sciences*, 115(37):9187–9192.

Camarda, C. G. (2019). Smooth constrained mortality forecasting. *Demographic Research*, 41:1091–1130.

Chen, Y. T. and Witten, D. M. (2022). Selective inference for k-means clustering. *arXiv preprint arXiv:2203.15267*.

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444.

Denecke, E., Grigoriev, P., and Rau, R. (2023). Evaluation of small-area estimation methods for mortality schedules. *arXiv preprint arXiv:2302.01693*.

Dharamshi, A., Alexander, M., Winant, C., and Barbieri, M. (2023a). Jointly estimating subnational mortality for multiple populations. *arXiv preprint arXiv:2310.03113*.

Dharamshi, A., Neufeld, A., Motwani, K., Gao, L. L., Witten, D., and Bien, J. (2023b). Generalized data thinning using sufficient statistics. *arXiv preprint arXiv:2303.12931*.

Dyrting, S. (2020). Smoothing migration intensities with p-topals. *Demographic Research*, 43:1607–1650.

Gao, L. L., Bien, J., and Witten, D. (2022). Selective inference for hierarchical clustering. *Journal of the American Statistical Association*.

Neufeld, A., Dharamshi, A., Gao, L. L., and Witten, D. (2023). Data thinning for convolution-closed distributions. *arXiv preprint arXiv:2301.07276*.

Rasines, D. G. and Young, G. A. (2022). Splitting strategies for post-selection inference. *Biometrika*.

Schmertmann, C. P. (2003). A system of model fertility schedules with graphically intuitive parameters. *Demographic research*, 9:81–110.

United States Mortality Database (2023).