2001

Linear Empirical Bayes Estimation of Survival Probabilities with Partial Data

Mostafa Mashayekhi
University of Nebraska-Lincoln, mmashaye@unlnotes.unl.edu

Journal of Actuarial Practice Vol. 9, 2001

Linear Empirical Bayes Estimation of Survival Probabilities with Partial Data

Mostafa Mashayekhi*

Abstract†

In this paper we consider linear empirical Bayes estimation of survival probabilities with partial data from right-censored and possibly left-truncated observations. Such data are produced by studies in which the exact times of death are not recorded and the length of time that each subject may be under observation cannot exceed one unit of time. We obtain asymptotically optimal linear empirical Bayes estimators, with respect to the squared error loss function, under the assumption that the probability of death under observation in a unit time interval is proportional to the length of observation. This assumption is sometimes implied by Balducci's assumption and sometimes is implied by the assumption of uniform distribution of deaths.

Key words and phrases: *asymptotically optimal, credibility theory*

*Mostafa Mashayekhi, Ph.D., A.S.A., M.A.A.A., is an assistant professor of actuarial science at the University of Nebraska-Lincoln. He received his bachelor's degree in economics from the University of London's Queen Mary College and his Master's degree in econometrics and mathematical economics from the London School of Economics and Political Science. He obtained his Ph.D. in statistics from Michigan State University. His research interests include statistical decision theory (especially the compound decision theory and empirical Bayes methods), survival models, and applications of stochastic calculus in actuarial mathematics.

Dr. Mashayekhi's address is: Actuarial Science Program, Department of Finance, University of Nebraska-Lincoln, Lincoln NE 68588-0426, USA. Internet address: *mmashaye@unlnotes.unl.edu*

†The author would like to thank the anonymous referee for carefully reading the manuscript and for helpful suggestions that led to a considerable improvement in the presentation of this paper.

1 Introduction

Consider the problem of estimating the mortality rate q_x or p_x with partial data from right-censored and possibly left-truncated observations[1] from a study of n individuals. Suppose the i^{th} individual comes under observation at age $x + r_i$ and is scheduled to be under observation for u_i years until age $x + s_i$, where $u_i = s_i - r_i$ and $0 \le r_i < s_i \le 1$. The data are partial in the sense that the exact times of death are not recorded. For each i, the data only show whether the i^{th} individual did or did not die under observation. Here the observable random variables are $\delta_1, \ldots, \delta_n$ where

$$\delta_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ person dies under observation; and} \\ 0 & \text{otherwise.} \end{cases}$$

Thus a typical record of data would contain i, x, r_i, u_i, and δ_i.

Because the times of death are not known, one cannot find the product-limit estimator with these data. Even when the exact times of death are known, the product limit estimator based on left-truncated observations (Klein and Moeschberger, 1997, pp. 114–115) can produce an unreasonable estimate of p_x.

The maximum likelihood method does not provide a compelling solution in this case either. The maximum likelihood method requires a distributional assumption that makes it possible to write ${}_{u_i}q_{x+r_i}$ in terms of q_x. The three well-known assumptions that actuaries use for $0 \le t \le 1$ are: (i) the Balducci assumption, i.e., ${}_{1-t}q_{x+t} = (1-t)q_x$; (ii) the assumption of uniform distribution of deaths, i.e., ${}_tq_x = tq_x$; and (iii) the constant force of mortality, i.e., ${}_tq_x = 1 - (1 - q_x)^t$. Under each of these assumptions, except for trivial cases, the likelihood equation $\mathrm{d}\mathcal{L}/\mathrm{d}q_x = 0$, where

$$\mathcal{L} = \prod_{i=1}^{n} \left(1 - {}_{u_i}q_{x+r_i}\right)^{1-\delta_i} \left({}_{u_i}q_{x+r_i}\right)^{\delta_i}$$

does not have a closed form solution unless n is small. When there is no closed form solution, one may find a solution by numerical methods. As the likelihood equation $\mathrm{d}\mathcal{L}/\mathrm{d}q_x = 0$ may have multiple roots, it is difficult to determine, however, if the solution obtained by numerical methods is the value of the root that has optimal large sample properties.

[1] An observation is said to be right-censored if the individual being observed is alive when the study ends. An observation is said to be left-truncated if the individual entered the study after age x.

Because maximum likelihood estimators are justified mainly by their desired large sample properties, the maximum likelihood approach in this case may not be appealing.

Another method of estimation is the method of moments. This method is one of the oldest statistical estimation methods. One of its biggest advantages over other statistical estimation methods is that it produces easy-to-compute estimates. One of its disadvantages is that it may produce an estimate that is outside the possible range of the parameter. Another disadvantage of the method of moments is that it may produce multiple estimators for the same parameter.

To demonstrate this, consider, for example, estimation of q_x with partial data as described above under the assumption that

$$_{u_i}q_{x+r_i} = u_i \times q_x \tag{1}$$

for each i. The assumed equality in equation (1) is the exact form of the approximation given in equation (6.3) of London (1988). Note that equation (1) cannot be satisfied without restrictions on r_i and u_i. Specifically, equation (1) without restrictions on r_i and u_i gives $_{0.5}q_x = {_{0.5}q_{x+0.5}} = 0.5q_x$, which, for $q_x > 0$, contradicts the identity

$$q_x = {_{0.5}q_x} + (1 - {_{0.5}q_x}) \times {_{1-0.5}q_{x+0.5}}.$$

The equality in equation (1) is practically plausible in three cases only: (i) with $s_i = 1$ and $r_i = 0$ for all i in which case the equality is trivially true; (ii) under Balducci's assumption with $s_i = 1$ for all i; and (iii) under the uniform distribution of deaths assumption with $r_i = 0$ for all i. Under these three cases London (1988) (equations (6.7), (6.10), (6.13)) proposes the method of moments estimator given by

$$\hat{q}_x^{(a)} = \frac{\sum\limits_{i=1}^{n} \delta_i}{\sum\limits_{i=1}^{n} u_i} \tag{2}$$

which is obtained by setting the random variable $\sum_{i=1}^{n} \delta_i$ equal to its expected value and solving for q_x. Another observable random variable that one can equate to its expected value to yield a method of moments estimator is $\sum_{i=1}^{n} u_i^{-1}\delta_i$, which has expected value equal to $_nq_x$. This method of moments estimator is given by

$$\hat{q}_x^{(b)} = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{u_i}. \tag{3}$$

Note that $\hat{q}_x^{(a)}$ and $\hat{q}_x^{(b)}$ are linear estimators of q_x. In general let $\omega_1, \ldots, \omega_n$ be non-negative weights such that $\sum_{i=1}^{n} \omega_i = 1$. Because

$$\mathbb{E}\left[\sum_{i=1}^{n} \omega_i \frac{\delta_i}{u_i}\right] = q_x$$

the method of moments estimator is given by

$$\hat{q}_x^{(\omega)} = \sum_{i=1}^{n} \omega_i \frac{\delta_i}{u_i}. \tag{4}$$

Clearly $\hat{q}_x^{(a)}$ and $\hat{q}_x^{(b)}$ are special cases of $\hat{q}_x^{(\omega)}$.

Because $\hat{q}_x^{(\omega)}$ is linear in the δ_i/u_is, it is natural to ask if there are better linear estimators than $\hat{q}_x^{(a)}$ and $\hat{q}_x^{(b)}$. From a Bayesian perspective, one can achieve a better result using the linear Bayes estimator, which is presented in Section 2. As will be seen, the linear Bayes estimator depends on the first two moments of the prior distribution. When these moments are known the linear Bayes estimator is available. If these two moments are unknown, however, they must be estimated and one can use the linear empirical Bayes estimator described in Section 3, which also contains a discussion of the asymptotic optimality of linear empirical Bayes estimators of q_x.

2 The Linear Bayes Estimator

In a Bayes estimation problem, one is faced with a data set consisting of n observable k-dimensional random vectors (k can be 1), $\mathbf{X}_1, \ldots, \mathbf{X}_n$, and an unobservable random variable or vector θ. Given θ, $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are mutually independent.

The loss function $L(t, \theta)$ specifies the loss of estimating (predicting) θ by $t = t(\mathbf{X}_1, \ldots, \mathbf{X}_n)$. Bayesians are interested in estimators that minimize the expected loss in some sense.

Definition 1. *An estimator $\hat{\theta} = \hat{\theta}(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is called a Bayes estimator if*

$$\mathbb{E}\left[L(\hat{\theta}, \theta)\right] = \min_t \mathbb{E}\left[L(t(\mathbf{X}_1, \ldots, \mathbf{X}_n), \theta)\right]$$

where $\mathbb{E}[\,]$ denotes the expectation with respect to the joint distribution of all of the random variables involved.

In other words, a Bayes estimator for a given loss function is an estimator that minimizes the expected loss over all estimators. As the basic method of moments estimators are linear (see equation (4)), we will consider linear Bayes estimators.

Definition 2. *An estimator $\hat{\theta}^*$ is called linear Bayes if*

$$\mathbb{E}\left[L(\hat{\theta}^*, \theta)\right] = \min_{a_0, \ldots, a_n} \mathbb{E}\left[L(t(\mathbf{X}_1, \ldots, \mathbf{X}_n), \theta)\right]$$

for t a linear function of the data, i.e., $t = a_0 + \sum_{i=1}^{n} a_i \mathbf{X}_i$.

Observe that for the squared error loss function given by $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, we have $L(1 - \hat{\theta}, 1 - \theta) = (\hat{\theta} - \theta)^2 = L(\hat{\theta}, \theta)$. Hence an estimator $\hat{\theta}$ is a Bayes (linear Bayes) estimator of θ if, and only if, $(1 - \hat{\theta})$ is a Bayes (linear Bayes) estimator of $(1 - \hat{\theta})$. Therefore, the linear Bayes (linear empirical Bayes) estimator of p_x is automatically found when we find the linear Bayes (linear empirical Bayes) estimator of q_x.

The following assumption gives a formal description of the model for our estimation problem.

Assumption 1. *Let $\theta = q_x$ and $X_i = \delta_i / u_i$, then θ, $X_1, \ldots, X_n$ are random variables such that*

1.1 $\mathbb{P}[0 \le \theta \le 1] = 1, \mathbb{P}[\theta = 1] < 1,$ *and* $\mathbb{P}[\theta = 0] < 1;$

1.2 Given θ, the random variables $X_1, \ldots, X_n$ are uncorrelated; and

1.3 $u_i X_i$ is a Bernoulli random variable taking the values 0 or 1 such that

$$\mathbb{P}[u_i X_i = 1] = u_i \theta,$$

where $0 < u_i \le 1$ is a known constant for $i = 1, \ldots, n$.

Assumption 1.3 corresponds to the assumed equality given in equation (1). Under Assumption 1, $\mathbb{E}[X_i|\theta] = \theta$, and $\mathbb{V}\text{ar}[X_i|\theta] = u_i\theta(1 - u_i\theta)/u_i^2$.

Let $\mu = \mathbb{E}[\theta]$ and $\sigma^2 = \mathbb{V}\text{ar}[\theta]$. Then we have $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}\text{ar}[X_i] = \mathbb{E}[\mathbb{V}\text{ar}[X_i|\theta]] + \mathbb{V}\text{ar}[\mathbb{E}[X_i|\theta]] = u_i^{-1}\mu - \mu^2$. Therefore

$$\mathbb{E}\left[X_i^2\right] = \mathbb{V}\text{ar}[X_i] + \mu^2 = u_i^{-1}\mu \tag{5}$$

and, for $i \neq j$,

$$\mathbb{E}\left[X_iX_j\right] = \mathbb{E}\left[\mathbb{E}\left[X_iX_j|\theta\right]\right] = \mathbb{E}\left[\theta^2\right] = \mu^2 + \sigma^2. \tag{6}$$

The following theorem gives the linear Bayes estimator of θ, i.e., of q_x. Its proof is given in the appendix.

Theorem 1. *Under Assumption 1 the linear Bayes estimator $\hat{\theta}^*$ of θ under the squared error loss is given by*

$$\hat{\theta}^* = q_x^* = b_0\mu + \sum_{i=1}^{n} b_iX_i \tag{7}$$

where

$$\alpha_i = [u_i^{-1}\mu - (\mu^2 + \sigma^2)]^{-1}, \tag{8}$$

$$b_i = (1 + \sigma^2 \sum_{i=1}^{n} \alpha_i)^{-1}\sigma^2\alpha_i, \tag{9}$$

for $i = 1, \ldots, n$, and $b_0 = 1 - \sum_{i=1}^{n} b_i$.

The next question is the determination of μ and σ^2. To a purely Bayesian actuary, the prior density of θ, $\pi(\theta)$, is completely known; hence, μ and σ^2 are known so that $\hat{\theta}^*$ can be determined easily from equation (7). An actuary who is not a pure Bayesian, however, would not have an explicitly known prior distribution. In this case the actuary may use either the uniform distribution as a non-informative prior for θ or use the empirical Bayes approach to estimate μ, σ^2, α_i, and b_i in equation (7). The empirical Bayes approach is described in the next section.

Examples of priors for θ (i.e., for q_x) are:

- $\pi(\theta) = 1$ for $0 < \theta < 1$. This is a non-informative prior because it reflects the actuary's complete ignorance of any prior information on q_x. This is an extreme case.
- Suppose a mortality study is done every three years on a block of policies. In the year 2000 study the actuary feels that mortality has dropped between, say, five and 25 percent from its previous level of $q_x^{(1997)}$ in 1997. In the absence of further information the actuary's prior would be

$$\pi(\theta) = \begin{cases} \frac{1}{0.20q_x^{(1997)}} & \text{for } 0.75q_x^{(1997)} < \theta < 0.95q_x^{(1997)} \\ 0 & \text{otherwise.} \end{cases}$$

The model described in Assumption 1 is similar to the credibility theory model of Bühlmann (1967); it reduces to the Bühlmann (1967) model when $u_i = 1$ for $i = 1, 2, \ldots, n$.

3 Linear Empirical Bayes Estimators

In the empirical Bayes approach pioneered by Robbins (1955), one is faced with m independent copies of the same decision problem. In the i^{th} problem there is a random pair (X_i, θ_i) where X_i is observable and θ_i is not observable. Conditional on $\theta_i = \theta$, X_i has a specified density $f(\cdot, \theta)$ for every i. In some of the variations of the empirical Bayes estimation that were later developed (e.g., Bühlmann and Straub (1970) and its generalization in Sundt (1983), or Ghosh and Meeden (1986)) in the i^{th} problem there is an observable random vector $\mathbf{X}_i = (X_{i1}, \ldots, X_{in_i})$ where n_is are not necessarily equal. There is a non-negative loss function $L(t, \theta)$. The unobservable θ_is are assumed to be i.i.d. with unknown common distribution function $G(\cdot)$.

To put this in the context of a mortality study, suppose there are m similar portfolios of insured lives, and the i^{th} portfolio consists of n_i lives. The j^{th} individual in the i^{th} portfolio comes under observation at age $x + r_{ij}$ and is scheduled to be under observation for u_{ij} years until age $x + s_{ij}$, where $u_{ij} = s_{ij} - r_{ij}$ and $0 \le r_{ij} < s_{ij} \le 1$. For each j, the data only show whether the j^{th} individual in the i^{th} portfolio did or did not die under observation. Here the observable random variables are $\delta_{i1}, \ldots, \delta_{in_i}$ where

$$\delta_{ij} = \begin{cases} 1 & \text{if the } j^{\text{th}} \text{ person in } i^{\text{th}} \text{ portfolio dies under observation; and} \\ 0 & \text{otherwise.} \end{cases}$$

Each individual in the i^{th} portfolio is characterized by an unobservable random mortality rate $\theta_i = q_x^{(i)}$ and the θ_is are values of an unobservable random sample from the same distribution. The data consist of the available observations as shown in Table 1. The random variables X_{ij} are defined by

$$X_{ij} = \frac{\delta_{ij}}{u_{ij}}$$

for $j = 1, 2, \ldots, n_i$ and $i = 1, 2, \ldots, m$. The problem is the simultaneous estimation of the θ_is.

Table 1
Illustration of the Empirical Bayes Problem

Portfolio	Mortality Rate	Outcome Observations			Death or Survival Period		
1	θ_1	δ_{11}	$\ldots$	δ_{1n_1}	u_{11}	$\ldots$	u_{1n_1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	θ_i	δ_{i1}	$\ldots$	δ_{in_i}	u_{i1}	$\ldots$	u_{in_i}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	θ_m	δ_{m1}	$\ldots$	δ_{mn_m}	u_{m1}	$\ldots$	u_{mn_m}

To avoid needless complications, Robbins assumes the existence of a Bayes decision function t_G such that

$$\mathbb{E}\left[L(t_G(X_m), \theta_m)\right] = \min_t \mathbb{E}\left[L(t(X_m), \theta_m)\right].$$

Robbins shows that when G is not known (and, hence, t_G is not directly available) for each problem, one may use asymptotically optimal decision rules that use the data from all of the m decision problems. These decision rules asymptotically give us the same risk that we would have with the knowledge of t_G. According to Robbins' definition, a sequence of decision rules $t_m(\cdot) = t_m(X_1, \ldots, X_m; \cdot)$ is asymptotically optimal relative to G as $m \to \infty$ if

$$\mathbb{E}\left[L(t_m(X_m), \theta_m)\right] - \mathbb{E}\left[L(t_G(X_m), \theta_m)\right] \to 0 \text{ as } m \to \infty,$$

where $\mathbb{E}[]$ denotes the expectation over all random variables. Though $t_m(\cdot)$ is a decision function and not an estimator, its value $t_m(X_m) = t_m(X_1, ..., X_m; X_m)$ is an empirical Bayes estimator for the m^{th} estimation problem, and, in the context of this paper, its value $t_m(X_k) = t_m(X_1, ..., X_m; X_k)$ is the empirical Bayes estimator for the k^{th} problem, $k = 1, 2, \ldots, m$.[2]

In the linear empirical Bayes estimation problem considered by Robbins (1983), the minimizing rule is the linear Bayes rule in the sense that it minimizes the Bayes risk for the i^{th} problem within the class of all estimators of the form $aX_i + b$. Thus, t_m is asymptotically optimal if the excess of risk of using t_m over the risk of using the linear Bayes rule converges to zero as the number of problems m increases.

Many variations of the linear empirical Bayes approach have been used by statisticians; see, for example, Morris (1983) for a list of some remarkable examples. These variations usually occur in cases where there are many similar independent estimation problems and the number of observations in each problem is small. In such cases one can do significantly better by borrowing strength from data from other problems. The strength is obtained through estimation of the prior distribution (in unrestricted empirical Bayes) or estimation of the necessary moments of the prior distribution (in the case of linear empirical Bayes) by using similar data. A notable example of linear Bayes (linear empirical Bayes approach) well known to actuaries is the Bühlmann (1967) approach in credibility theory.

The variation that we are considering is slightly different from Robbins' empirical Bayes or linear empirical Bayes in the sense that our m problems are not identical when the sample sizes are different or when the durations of time that different subjects are under observation are

[2]It must be emphasized that although $t_m(X_k) = t_m(X_1, ..., X_m; X_k)$ is an estimator for the k^{th} problem, $k = 1, 2, \ldots, m$ in the context of this paper, it is not true for what Robbins does. Robbins (1955) uses so-called *delete bootstrap rules* because he has posed his problem in a non-parametric unrestricted empirical Bayes context. Non-delete bootstrap rules, although desirable, are difficult to use in the non-parametric unrestricted empirical Bayes context. In this paper, however, we consider a linear empirical Bayes estimation problem, which can be solved through the estimation of only the first two moments of the prior distribution. This has allowed us use the more desirable non-delete rules. Specifically, we have used all of the observations to find estimators for the first two moments of the prior distribution and hence the shape of the decision rule. We then have used observations from each problem to find the linear empirical Bayes estimator for that problem. This is not what Robbins (1955) has done. He considers empirical Bayes estimators for the m^{th} problem only.

not equal. Still, we may define the linear empirical Bayes estimators $\hat{\theta}_1^{\text{EB}}, \ldots, \hat{\theta}_m^{\text{EB}}$ to be asymptotically optimal if, with $\hat{\theta}_i^*$ denoting the linear Bayes estimator for the i^{th} problem, for each $i = 1, \ldots, m$ we have

$$\mathbb{E}\left[(\hat{\theta}_i^{\text{EB}} - \theta_i)^2\right] - \mathbb{E}\left[(\hat{\theta}_i^* - \theta_i)^2\right] \to 0 \text{ as } m \to \infty.$$

The model we are considering is formalized in the following assumption.

Assumption 2. $(X_{11}, \ldots, X_{1n_1}, \theta_1), \ldots, (X_{m1}, \ldots, X_{mn_m}, \theta_m)$ *are independent random vectors such that*

2.1 $\theta_1, \ldots, \theta_m$ *are identically distributed random variables with* $\mathbb{P}[0 \le \theta_i \le 1] = 1$, $\mathbb{P}[\theta_i = 1] < 1$, *and* $\mathbb{P}[\theta_i = 0] < 1$;

2.2 Conditional on θ_i, *the* $X_{i1}, \ldots, X_{in_i}$ *are uncorrelated and;*

2.3 $u_{ij}X_{ij}$ *is Bernoulli with parameter* $u_{ij}\theta_i$ *where* $0 < u_* \le u_{ij} \le 1$ *are known numbers; and*

2.4 There exists a K such that $2 \le n_i \le K < \infty$ *for all i.*

Assumption 2 is similar to Bühlmann and Straub (1970). In the Bühlmann and Straub model $(\theta_1, X_{11}, \ldots, X_{1n_1}), \ldots, (\theta_m, X_{m1}, \ldots, X_{mn_m})$ are m independent random vectors such that the θ_is are unobservable and X_{ij} is observable for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$. There are functions μ_1 and υ such that

$$\mathbb{E}[X_{it} | \theta_i] = \mu_1(\theta_i)$$

and

$$\mathbb{C}\text{ov}[X_{ir}, X_{is} | \theta_i)] = \begin{cases} \dfrac{\upsilon(\theta_i)}{p_{ir}} & \text{if } r = s \\ 0 & \text{otherwise,} \end{cases}$$

where the p_{ir}s are known constants. In Bühlmann-Straub the n_is are equal. In later variations, however, n_is are not necessarily equal. Observe that when u_{ij}s are all equal our model satisfies the above assumptions by choosing $\mu_1(\theta) = \theta$, and $\upsilon(\theta) = u^{-1}\theta(1 - u\theta)$, and $p_{ij} = 1$, with u being the common value of the u_{ij}s. Also note that in the Bühlmann (1967) model the conditional distributions are not completely specified. In our model the conditional distributions are completely specified to be Bernoulli.

Assumption 2 is used throughout the rest of this paper and therefore we will not mention it in the statement of every lemma or theorem. In the remainder of this paper all incompletely described limits are as $m \to \infty$ through positive integers.

Let μ and σ^2 denote the mean and variance of θ_i, respectively. Observe that under Assumption 2 we have

$$\mathbb{E}\left[X_{ij}\right] = \mu \quad \text{and} \quad \mathbb{V}\text{ar}\left[X_{ij}\right] = \frac{\mu}{u_{ij}} - \mu^2.$$

Similar to equations (5) and (6), we have

$$\mathbb{E}\left[X_{ij}^2\right] = u_{ij}^{-1}\mu$$

and for $k \neq j$

$$\mathbb{E}\left[X_{ij}X_{ik}\right] = \mu^2 + \sigma^2. \tag{10}$$

Let

$$\bar{X}_{i\bullet} = \sum_{j=1}^{n_i} \omega_{ij}X_{ij}, \quad N = \sum_{i=1}^{m} n_i, \quad \bar{X}_{\bullet\bullet} = \frac{1}{N}\sum_{i=1}^{m} n_i\bar{X}_{i\bullet}, \quad \text{and}$$
$$Y_i = \frac{1}{\binom{n_i}{2}} \sum_{1\le j<k\le n_i} X_{ij}X_{ik},$$

where the ω_{ij}s are non-negative weights such that $\sum_{j=1}^{n_i} \omega_{ij} = 1$. We propose using the following estimates for μ and σ^2

$$\hat{\mu} = \bar{X}_{\bullet\bullet} \tag{11}$$

and

$$\hat{\sigma}^2 = \max(0, \bar{Y} - \hat{\mu}^2) \tag{12}$$

respectively, where

$$\bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i. \tag{13}$$

The linear empirical Bayes estimator of θ_i, based on these estimators of μ and σ^2 is given by

$$\hat{\theta}_i^{\mathrm{EB}} = \hat{q}_x^{(i)} = \hat{b}_{i0}\hat{\mu} + \sum_{j=1}^{n_i} \hat{b}_{ij} X_{ij} \tag{14}$$

where

$$\hat{\alpha}_{ij} = \begin{cases} \frac{1}{u_{ij}^{-1}\hat{\mu} - (\hat{\mu}^2 + \hat{\sigma}^2)} & \text{if } u_{ij}^{-1} - (\hat{\mu}^2 + \sigma^2) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

and

$$\hat{b}_{ij} = (1 + \hat{\sigma}^2 \sum_{j=1}^{n_i} \hat{\alpha}_{ij})^{-1} \hat{\sigma}^2 \hat{\alpha}_{ij}, \tag{16}$$

for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$, and

$$\hat{b}_{i0} = 1 - \sum_{j=1}^{n_i} \hat{b}_{ij}. \tag{17}$$

It can be proved (see Theorem 2 in the appendix) that the $\hat{\theta}_i^{\mathrm{EB}}$s are asymptotically optimal linear empirical Bayes estimators in the sense that for every $i = 1, \ldots, m$

$$\mathbb{E}\left[(\hat{\theta}_i^{\mathrm{EB}} - \theta_i)^2\right] - \mathbb{E}\left[(\hat{\theta}_i^* - \theta_i)^2\right] \to 0. \tag{18}$$

where $\hat{\theta}_i^*$ is the linear Bayes estimator of θ_i.

If we choose $\sigma = 0$, so that the class of prior distributions under consideration reduces to the class of point priors (the traditional frequentist approach) then with $m = 1$, the linear empirical Bayes estimators in equation (14) will be the same as the estimator $\hat{q}_x^{(\omega)}$ in equation (4).

A natural question to ask now is how do we choose the ω_{ij}s? Observe that according to Theorem 2 every choice of ω_{ij}s provides an asymptotically optimal estimator. However a smaller variance of $\bar{X}_{\bullet\bullet}$ means a better speed of convergence. Because the variance of $\bar{X}_{\bullet\bullet}$ is minimized when variance of each $\bar{X}_{i\bullet}$ is minimized and

$$\begin{aligned}\mathbb{V}\text{ar}\,[\bar{X}_{i\bullet}] &= \mathbb{E}\,[\mathbb{V}\text{ar}\,[\bar{X}_{i\bullet}|\theta_i]] + \mathbb{V}\text{ar}\,[\mathbb{E}\,[\bar{X}_{i\bullet}|\theta_i]] \\ &= \sum_{j=1}^{n_i} \omega_{ij}^2 \left(\frac{\mu - u_{ij}(\mu^2+\sigma^2)}{u_{ij}}\right) + \sigma^2,\end{aligned}$$

we need to minimize

$$\sum_{j=1}^{n_i} \omega_{ij}^2 \left(\frac{\mu - u_{ij}(\mu^2+\sigma^2)}{u_{ij}}\right)$$

subject to the constraint $\sum_{j=1}^{n_i} \omega_{ij} = 1$. Writing the Lagrangian

$$\sum_{j=1}^{n_i} \omega_{ij}^2 \left(\frac{\mu - u_{ij}(\mu^2+\sigma^2)}{u_{ij}}\right) - \lambda(\sum_{j=1}^{n_i} \omega_{ij} - 1)$$

and setting the partial derivatives equal to zero yields the minimizer

$$\omega_{ij}^* = \frac{c_{ij}}{\sum_{j=1}^{n_i} c_{ij}}, \tag{19}$$

where

$$c_{ij} = \frac{u_{ij}}{\mu - u_{ij}(\mu^2+\sigma^2)}. \tag{20}$$

As ω_{ij}^* depends on the unknown parameters, it is not available. Note, however, that $\mu^2 + \sigma^2 = \mathbb{E}\,[\theta^2]$. Therefore in cases when θ^2 is so small that its expectation becomes negligible we have

$$\omega_{ij}^* \cong \frac{u_{ij}}{\sum_{j=1}^{n_i} u_{ij}}.$$

The above argument also shows that the choice of weights in the moment estimator of equation (2) is a reasonable choice when q_x is small. One can also use the Chebychev's inequality, similar to the proof of consistency of $\bar{X}_{\bullet\bullet}$, to show that $\hat{q}_x^{(a)}$ of equation (2) converges in probability to q_x as $n \to \infty$. Thus when there is a large homogeneous sample available for estimation of q_x there is not much to gain by using the linear empirical Bayes method. The problem, however, is that it is not always feasible to have a large sample of homogeneous subjects. When there is a large sample of subjects that can be broken into many homogenous groups, one can show by using a variation of the weak law of large numbers (Hannan and Fabian (1985), Theorem 2.3.9) that using the estimator of equation (2) will provide a weighted average of the failure probabilities of the homogenous groups that are in the large sample. An actuary who uses such a weighted average in the determination of premiums can expect to face some anti-selection by those who feel the premium is unfair to them. Breaking the large sample into many homogeneous groups on the other hand may leave a small number of subjects in each homogeneous group. In such a case one can gain by using a linear empirical Bayes estimator instead of using the moment estimator of equation (2) for each homogenous sample separately.

4 Concluding Remarks

In this paper we obtain an asymptotically optimal linear empirical Bayes estimator of θ_i, with the yardstick of performance being the risk of the linear Bayes estimator. The main reason for using linear empirical Bayes estimators instead of the empirical Bayes estimators is that linear empirical Bayes estimators exist under milder conditions and are usually much easier to compute. When it is possible to reduce the risk of an asymptotically optimal linear Bayes estimator with a simple adjustment, one should not hesitate to do so.

It is easy to see that by construction we have $\hat{\theta}_i^{\text{EB}} \geq 0$. It is possible, however, that the value of $\hat{\theta}_i^{\text{EB}}$ could become more than 1. Let $\hat{\theta}_i^{**}$ be equal to $\hat{\theta}_i^{\text{EB}}$ when $\hat{\theta}_i^{\text{EB}} \leq 1$ and let $\hat{\theta}_i^{**} = 1$ otherwise. The θ_is are known to be in [0,1]; therefore, we have $\mathbb{E}\left[(\hat{\theta}_i^{**} - \theta_i)^2\right] \leq \mathbb{E}\left[(\hat{\theta}_i^{\text{EB}} - \theta_i)^2\right]$ because $|\hat{\theta}_i^{**} - \theta_i| \leq |\hat{\theta}_i^{\text{EB}} - \theta|$.

We started this paper by considering the survival probabilities as related to life insurance. The method of estimation that we present, however, may find more applications in the casualty insurance. Consider, for example, the case when an insurer who has insured a large number N of drivers is interested in assessing the risk due to severe accidents that cannot happen to a person more than once. Examples of such accidents include fatal accidents and accidents resulting in a severe disability so that the person will not be able to drive again.

Suppose that the insurer is able to classify the N policy holders according to factors such as age, area, etc. into m homogeneous groups with n_i drivers in the i^{th} group for $i = 1, \ldots, m$ such that m is large and each n_i is small. Also suppose that it is reasonable to assume the probability of an accident for the j^{th} driver in the i^{th} class during the policy period is equal to $u_{ij}\theta_i$ where u_{ij} is the duration of time the person is insured and θ_i is the probability of an accident by a typical member of the i^{th} class in a unit interval of time. Let B_{ij} denote the amount of loss the insurer will suffer if the j^{th} driver in the i^{th} class faces an accident.

In this case because each n_i is small and also because when the u_{ij}s are not equal the probabilities of accident during the policy period for different drivers are not equal, the Poisson distribution or the negative binomial distribution will not give a good approximation for the distribution of the number of accidents in each group. Therefore, using a compound Poisson model or compound negative binomial model for each class will not be accurate. In such a case, using the individual risk model (Bowers et al., 1986) for each class can produce more accurate results. In order to use the individual risk model, however, the insurer would need an estimate of θ_i for $i = 1, \ldots, m$. In such a case, the method presented in this paper can be used to obtain the desired estimates when the insurer has experience data for these m classes from a past year.

A very important question that every practitioner may ask before using any variations of the empirical Bayes approach is how large should m be? Because answering this question accurately requires knowledge of the rate of convergence of the risk of the empirical Bayes estimator, this question is often a good cause for further research when asymptotic results are obtained through application of convergence theorems such as the Lebesgue Dominated Convergence Theorem. For some results that provide a step for further research in this direction, see Hesselager (1992).

Appendix: The Proofs

In order to prove Theorem 1, we note the following: Suppose that (i) θ, $X_1, \ldots, X_n$ are random variables with finite second moments (so that they all belong to the L_2 space, and (ii) the loss function is the squared error loss function given by $L(t, \theta) = (t - \theta)^2)$. Then, from the definition of the L_2 projection (see, for example, Brockwell and Davis 1987, Chapter 2), the Bayes estimator of θ is the L_2 projection of θ on the set of all functions of $X_1, \ldots, X_n$ that belong to the L_2 space. The linear Bayes estimator of θ is the L_2 projection of θ on the closed span of $\{1, X_1, \ldots, X_n\}$.

Proof: Because $\mathbb{P}[0 \leq \theta \leq 1] = 1$, $\mathbb{P}[\theta = 1] < 1$, and $\mathbb{P}[\theta = 0] < 1$, we have $\mu = \mathbb{E}[\theta] > \mathbb{E}[\theta^2] = \mu^2 + \sigma^2$. Because $0 < u_i \leq 1$, it follows that each α_i is well defined and greater than zero. We must show that $\hat{\theta}^*$ is a version of the L_2 projection of θ on the closure of the linear span of $\{1, X_1, \ldots, X_n\}$. Thus it is enough to check that $(\hat{\theta}^* - \theta)$ is L_2 perpendicular to 1 and to X_i for $i = 1, \ldots, n$ because, if $\mathbb{E}\left[\hat{\theta}^* - \theta\right] = 0$ and $\mathbb{E}\left[(\hat{\theta}^* - \theta)X_i\right] = 0$, then for all $a_0, \ldots, a_n$

$$\mathbb{E}\left[\left(\hat{\theta}^* - \theta\right)\left(a_0 + \sum_{i=1}^{n} a_i X_i\right)\right] = 0$$

so that $\hat{\theta}^* - \theta$ is perpendicular to every element of the closed span of $\{1, X_1, \ldots, X_n\}$. We have

$$\mathbb{E}\left[\hat{\theta}^* - \theta\right] = (1 - \sum_{i=1}^{n} b_i)\mu + \sum_{i=1}^{n} b_i \mu - \mu = 0.$$

So it remains to show that $\mathbb{E}\left[(\hat{\theta}^* - \theta)X_i\right] = 0$ for each $i = 1, \ldots, n$. Because $\mathbb{E}[\theta X_i] = \mathbb{E}[\mathbb{E}[\theta X_i|\theta]] = \mathbb{E}[\theta \mathbb{E}[X_i|\theta]] = \mathbb{E}[\theta^2] = \mu^2 + \sigma^2$, it is enough to show that $\mathbb{E}\left[\hat{\theta}^* X_i\right] = \mu^2 + \sigma^2$. We have

$$\mathbb{E}\left[\hat{\theta}^* X_i\right] = b_0 \mu^2 + \sum_{j \neq i} b_j \mathbb{E}\left[X_j X_i\right] + b_i \mathbb{E}\left[X_i^2\right]. \tag{21}$$

Thus, from equations (5) and (6) and by definition of α_i, it easily follows that the right side of equation (21) is equal to

$$\frac{\mu^2 + \sigma^2(\mu^2 + \sigma^2)\left(\sum_{i=1}^{n} \alpha_i\right) + \sigma^2 \alpha_i \times \alpha_i^{-1}}{1 + \sigma^2 \sum_{i=1}^{n} \alpha_i} = \mu^2 + \sigma^2,$$

and Theorem 1 is proved. □

Lemma 1. *Let $\hat{\mu}$ be as defined in equation (11) and $\hat{\sigma}^2$ be as defined in equation (12). Then*

$$\hat{\mu} \xrightarrow{P} \mu \tag{22}$$

and

$$\hat{\sigma}^2 \xrightarrow{P} \sigma^2. \tag{23}$$

Proof: Because $u_{ij}X_{ij}$ is Bernoulli and $u_{ij} \geq u_*$, we have $0 \leq X_{ij} < u_*^{-1}$. This gives $0 \leq \bar{X}_{i\bullet} \leq u_*^{-1}$ and hence $\mathbb{V}\text{ar}\left[\bar{X}_{i\bullet}\right] \leq \mathbb{E}\left[\bar{X}_{i\bullet}^2\right] \leq u_*^{-2}$. Therefore

$$\mathbb{V}\text{ar}\left[\bar{X}_{\bullet\bullet}\right] = \left(\sum_{i=1}^{m} n_i\right)^{-2} \sum_{i=1}^{m} n_i^2 \mathbb{V}\text{ar}\left[\bar{X}_{i\bullet}\right] \leq m^{-1} K u_*^{-2} \xrightarrow{P} 0. \tag{24}$$

Hence, equation (22) follows from equation (24), from Chebychev's inequality, and from the fact that $\mathbb{E}\left[\bar{X}_{\bullet\bullet}\right] = \mu$.

From equation (10), it follows that $\mathbb{E}\left[\bar{Y}\right] = \mu^2 + \sigma^2$. Because $0 \leq X_{ij} \leq u_*^{-1}$, we have $Y_i \leq u_*^{-2}$ and, hence, $\mathbb{V}\text{ar}\left[(Y_i\right] \leq u_*^{-4}$. Therefore $\mathbb{V}\text{ar}\left[\bar{Y}\right]) \leq m^{-1}u_*^{-4} \xrightarrow{P} 0$. By Chebychev's inequality it follows that $\bar{Y} \xrightarrow{P} \mu^2 + \sigma^2$. Because $\bar{X}_{\bullet\bullet} \xrightarrow{P} \mu$, it follows that $\bar{X}_{\bullet\bullet}^2 \xrightarrow{P} \mu^2$ and, hence, $\bar{Y} - \bar{X}_{\bullet\bullet}^2 \xrightarrow{P} \sigma^2$. Because $\sigma^2 \geq 0$, continuity of the function $g(x) = \max(0, x)$ gives equation (23). □

Lemma 2. *Suppose $\hat{\mu} \xrightarrow{P} \mu$ and $\hat{\sigma}^2 \xrightarrow{P} \sigma^2$. Let $\hat{\alpha}_{ij}$ be given by equation (15) and $\alpha_{ij} = [u_{ij}^{-1}\mu - (\mu^2 + \sigma^2)]^{-1}$. Let $\hat{b}_{ij}$ be as in equation (16) and*

$$b_{ij} = (1 + \sigma^2 \sum_{j=1}^{n_i} \alpha_{ij})^{-1} \sigma^2 \alpha_{ij}.$$

Let $\hat{b}_{i0} = 1 - \sum_{j=1}^{n_i} \hat{b}_{ij}$ and $b_{i0} = 1 - \sum_{j=1}^{n_i} b_{ij}$. Then for each $i = 1, \ldots, m$ and $k = 0, 1, \ldots, n_i$,

$$\hat{b}_{ik} - b_{ik} \xrightarrow{P} 0. \tag{25}$$

Proof: We prove the lemma by first showing that

$$\hat{\alpha}_{ij} - \alpha_{ij} \xrightarrow{P} 0. \tag{26}$$

Because $0 < u_{ij}^{-1} \leq u_*^{-1}$ we have

$$\left(\frac{\hat{\mu}}{u_{ij}} - (\hat{\mu}^2 + \hat{\sigma}^2)\right) - \frac{1}{\alpha_{ij}} = \frac{\hat{\mu} - \mu}{u_{ij}} - (\hat{\mu}^2 - \mu^2) - (\hat{\sigma}^2 - \sigma^2) \xrightarrow{P} 0. \tag{27}$$

If a_m and a'_m are two sequences such that $a_m \geq a > 0$ and $a_m - a'_m \to 0$, then eventually $a'_m \geq a/2 > 0$. Hence, eventually

$$\frac{1}{a_m} - \frac{1}{a'_m} = \frac{a'_m - a_m}{a_m a'_m} \to 0.$$

Therefore, because $\alpha_{ij}^{-1} \geq \mu - (\mu^2 + \sigma^2) > 0$, equation (26) follows from equation (27) by the fact (Bilingsley, 1986, p. 274; Royden, 1968, p. 93) that a sequence a_m converges in probability to zero if and only if every subsequence of a_m has a further subsequence that converges to zero with probability 1.

Let $\varepsilon > 0$ and $i \in \{1, \ldots, m\}$. Observe that $\sum_{j=1}^{n_i} |\hat{\alpha}_{ij} - \alpha_{ij}| > \varepsilon$ only if for some $j \in \{1, ..., n_i\}$,

$$|\hat{\alpha}_{ij} - \alpha_{ij}| > n_i^{-1}\varepsilon > K^{-1}\varepsilon.$$

Thus we have

$$\mathbb{P}\left[\,\left|\sum_{j=1}^{n_i} \hat{\alpha}_{ij} - \sum_{j=1}^{n_i} \alpha_{ij}\right| > \varepsilon\right] \leq \mathbb{P}\left[\sum_{j=1}^{n_i} |\hat{\alpha}_{ij} - \alpha_{ij}| > \varepsilon\right] \tag{28}$$

$$\leq \sum_{j=1}^{n_i} \mathbb{P}\left[|\hat{\alpha}_{ij} - \alpha_{ij}| > K^{-1}\varepsilon\right] \to 0 \tag{29}$$

by equation (26) and the assumption that $n_i \leq K$. This means that $\sum_{j=1}^{n_i} \hat{\alpha}_{ij} - \sum_{j=1}^{n_i} \alpha_{ij} \xrightarrow{P} 0$ and, hence,

$$(1 + \hat{\sigma}^2 \sum_{j=1}^{n_i} \hat{\alpha}_{ij}) - (1 + \sigma^2 \sum_{j=1}^{n_i} \alpha_{ij}) \xrightarrow{P} 0. \tag{30}$$

Because $1 + \sigma^2 \sum_{j=1}^{n_i} \alpha_{ij} \geq 1$ it follows from equation (30) that

$$(1 + \hat{\sigma}^2 \sum_{j=1}^{n_i} \hat{\alpha}_{ij})^{-1} - (1 + \sigma^2 \sum_{j=1}^{n_i} \alpha_{ij})^{-1} \xrightarrow{P} 0. \tag{31}$$

It follows from equations (26) and (31) that for $j = 1, \ldots, n_i$, we have $\hat{b}_{ij} - b_{ij} \xrightarrow{P} 0$. Because $n_i \leq K$, it follows that

$$\sum_{j=1}^{n_i} \hat{b}_{ij} - \sum_{j=1}^{n_i} b_{ij} \xrightarrow{P} 0$$

which means we also have $\hat{b}_{i0} - b_{i0} \xrightarrow{P} 0$, and the proof is complete. □

Theorem 2. . *Let $\hat{\mu}$ be as in Lemma 1. For $j = 0, 1, \ldots, n_i$ let $\hat{b}_{ij}$ be as defined in Lemma 2. Let $\hat{\theta}_i^{EB} = \hat{b}_{i0}\hat{\mu} + \sum_{j=1}^{n_i} \hat{b}_{ij}X_{ij}$. Then $\hat{\theta}_i^{EB}$ is an asymptotically optimal linear empirical Bayes estimator in the sense that for every $i = 1, \ldots, m$ with $\hat{\theta}_i^*$ denoting the linear Bayes estimator of θ_i,*

$$\mathbb{E}\left[(\hat{\theta}_i^{EB} - \theta_i)^2\right] - \mathbb{E}\left[(\hat{\theta}_i^* - \theta_i)^2\right] \to 0. \tag{32}$$

Proof: From Lemma 2, it easily follows that $\hat{\theta}_i^{EB} - \hat{\theta}_i^* \xrightarrow{P} 0$. Because $0 \leq X_{ij} < u_*^{-1}$, we obtain that $\hat{\theta}_i^{EB}$ and $\hat{\theta}_i^*$ are both bounded. We also have $0 \leq \theta_i \leq 1$. Therefore

$$|(\hat{\theta}_i^{EB} - \theta_i)^2 - (\hat{\theta}_i^* - \theta_i)^2| = |\hat{\theta}_i^{EB} + \hat{\theta}_i^* - 2\theta_i|.|\hat{\theta}_i^{EB} - \hat{\theta}_i^*| \xrightarrow{P} 0.$$

Because $(\hat{\theta}_i^{EB} - \theta_i)^2 - (\hat{\theta}_i^* - \theta_i)^2$ is bounded, the assertion of the theorem follows by the bounded convergence theorem. □

References

Billingsley, P. *Probability and Measure*, Second Edition. New York, N.Y.: Wiley, 1986.

Bowers, N.L., Gerber, H.U., Hickman, J.C., Jones, D.A., and Nesbitt, C.J. *Actuarial Mathematics*, Schaumburg, Ill.: Society of Actuaries, 1986.

Brockwell, P.J. and Davis, R.A. *Time Series: Theory and Methods.* New York, N.Y.: Springer-Verlag, 1987.

Bühlmann, H. "Experience Rating and Credibility." *ASTIN Bulletin* 4 (1967): 199–207.

Bühlmann, H. and Straub, E. "Glaubgwürdigkeit für Schadensätze." *Bulletin of the Swiss Association of Actuaries* 70 (1970): 111–133. English translation by C.E. Brooks.

Fabian, V. and Hannan, J. *Introduction to Probability and Mathematical Statistics.* New York, N.Y.: John Willey & Sons, 1985.

Ghosh, M, and Meeden, G. "Empirical Bayes Estimation in Finite Population Sampling." *Journal of the American Statistical Association* 81 (1986): 1058–1062

Klein, J.P. and Moeschberger, M.L. *Survival Analysis.* New York, N.Y.: Springer-Verlag, 1997.

London, D. *Survival Models and Their Estimation*, Third Edition. Winsted, Conn.: Actex, 1997.

Hesselager, O. "Rates of Risk Convergence of Empirical Linear Bayes Estimators." *Scandinavian Actuarial Journal* (1992): 88–94.

Morris, C.N. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association* 78 (1983): 47–59.

Robbins, H. "An Empirical Bayes Approach to Statistics." *Proceedings of the Third Berkeley Symposium of Mathematical Statistics Probability* 1 (1955): 157–164.

Robbins, H. "The Empirical Bayes Approach to Statistical Decision Problems." *Annals of Mathematics and Statistics* 35 (1964): 1–20.

Robbins, H. "Some Thoughts on Empirical Bayes Estimation." *Annals of Statistics* 11 (1983): 713–723.

Royden, H.L. *Real Analysis*, Second Edition. New York, N.Y.: Macmillan, 1968.

Sundt, B. "Parameter Estimation in Some Credibility Models." *Scandinavian Actuarial Journal* (1983): 239–255.