Title
Hierarchical regression for epidemiologic analyses of multiple exposures.

Journal
Environmental Health Perspectives, 102(suppl 8)

ISSN
1542-4359

Author
Greenland, S

Publication Date
1994-11-01

DOI
10.1289/ehp.94102s833

Peer reviewed

Hierarchical Regression for Epidemiologic Analyses of Multiple Exposures

Sander Greenland

Department of Epidemiology, UCLA School of Public Health, Los Angeles, California

Many epidemiologic investigations are designed to study the effects of multiple exposures. Most of these studies are analyzed either by fitting a risk-regression model with all exposures forced in the model, or by using a preliminary-testing algorithm, such as stepwise regression, to produce a smaller model. Research indicates that hierarchical modeling methods can outperform these conventional approaches. These methods are reviewed and compared to two hierarchical methods, empirical-Bayes regression and a variant here called "semi-Bayes" regression, to full-model maximum likelihood and to model reduction by preliminary testing. The performance of the methods in a problem of predicting neonatal-mortality rates are compared. Based on the literature to date, it is suggested that hierarchical methods should become part of the standard approaches to multiple-exposure studies. —Environ Health Perspect 102(Suppl 8):33–39 (1994)

Key words: Bayesian statistics, hierarchical models, relative-risk gression, risk assessment

Introduction

Many epidemiologic studies, especially in occupational and environmental health, have as their objective the evaluation of multiple exposures for potentially harmful effects. While the statistics literature ostensibly deals with these problems under such topics as stepwise regression and multiple comparisons, prominent epidemiologic methodologists have been adamant in their rejection of such methods (*1–3*). These authors criticize conventional methods for (among other things) failure to take account of prior knowledge, the irrelevance of the inferential objectives on which the methods are based, and the tendency of the methods to misrepresent continuous estimation problems as discrete decision problems. I have been largely sympathetic with these criticisms; nevertheless, I have not found these alternative points of view to be entirely satisfactory with regard to the representation and solutions they offer for the multiple-inference problem (*4*).

Developments over the past few decades offer fresh approaches to the problem. Thanks to dramatic increases in computing power, hierarchical methods (such as empirical-Bayes regression) can be explored as practical alternatives or supplements to the regression analyses common in epidemiology. The hierarchical perspective is oriented specifically toward multiple-inference problems. In contrast to conventional multiple-inference methods, the Bayesian format of hierarchical methods allows straightforward accommodation of prior information and interpretation of results. At the same time, hierarchical methods can enjoy superb repeated-sampling properties (*5,6*).

Several authors have discussed the application of parametric empirical-Bayes methods to the estimation of parameters in risk-regression models (*7–10*); a related, Bayesian log-linear approach to risk modeling was given by Cornfield (*11*). Here, I provide an overview and comparison of two of the more common modeling strategies employed by epidemiologists (maximum likelihood [ML] and preliminary testing) to two hierarchical methods, parametric empirical-Bayes regression (*6*) and a variant I call "semi-Bayes" regression (*9,10*). After reviewing the basic strategies, I illustrate these methods and another hierarchical method (Bayes empirical-Bayes) in an application to a neonatal mortality study. The results illustrate how hierarchical methods can offer considerable advantages over more common approaches to epidemiologic studies of multiple exposures.

Background

Consider the following problem: An investigator gathers data on an outcome (dependent) variable y, an n-row vector of study variables x ("exposures"), and an m-row vector of nuisance variables w ("potential confounders"), with the intention of estimating the n-column vector β of exposure coefficients in a generalized linear model for the expectation of y conditional on x and w,

$$g[E(y \mid x,w)] = \alpha + x\beta + w\gamma$$

where g is a known, strictly increasing link function and y is assumed to be randomly sampled from its distribution conditional on x,w. In this multiple-estimation setting, the investigator may be concerned to maximize the "accuracy" in estimating β. In the epidemiologic literature, the concept of accuracy is rarely formalized; when it is, accuracy of point estimation is sometimes equated with mean-squared estimation error (*12*), and accuracy of interval estimation is usually equated with nominal or conservative coverage coupled with short length (*12*). The multiplicity inherent in the interval-estimation task is almost always disregarded on the ground that componentwise coverage, not overall coverage, is scientifically relevant in exploratory studies (*1–3*). The latter view finds acceptable the high probability that at least one component of β will not be covered by the componentwise 95% intervals if β has, say, 20 components.

Other multiple-inference problems arise if one is chiefly interested in estimating the expectation $E(y \mid x, w)$ or future values of y (prediction). In these problems, the coefficients are of only intermediate interest. Nonetheless, accuracy in coefficient estimation reasonably could be expected to correlate with prediction accuracy, and coefficient estimation methods can be compared for their effectiveness in producing accurate prediction.

In epidemiology, two common strategies for estimating study-variable coefficients are the following:

- Use estimates from a "full" model, one that contains all the study variables (if

This paper was presented at the 4th Japan-US Biostatistics Conference on the Study of Human Cancer held 9–11 November 1992 in Tokyo, Japan.

The author is grateful to David Draper and Dana Flanders for numerous helpful comments.

Address correspondence to Sander Greenland, 1720 Tuna Canyon Road, Topanga CA 90290.

such a model can be fit), or as many variables as can be fit. Kleinbaum et al. (*12*), Miettinen (*1*), and Rothman (*2*) make recommendations along these lines. (Note that full stratification on all variables is equivalent to such an approach, but is rarely practical because of the sparsity of the stratification.)

- Use estimates from a "reduced" model, including only components of β or γ that survive some preliminary-testing algorithm (such as forward selection, backward elimination, or univariate testing).

Note that in the second strategy, coefficients deleted from the model are, in effect, assigned an estimated value of zero.

Either of the above strategies may involve algorithms for reduction of the nuisance parameter γ, e.g., by backward elimination of components of w. In fact, virtually all the epidemiologic literature on variable selection has focused on methods for reduction of the nuisance parameter, rather than β(*12,13*). This focus may explain in part why published analyses often treat the components of β one at a time: for each component of β, a model is selected based on some method for reducing the number of remaining components; in each model, one component is treated as the sole study variable, while the remaining study variables are treated as nuisance variables. For example, β_1 would be estimated by forcing x_1 into the model, then applying some selection or reduction algorithm to the vector of remaining variables $(x_2,\ldots,x_n,w_1,\ldots,w_m)$. In essence, this approach treats an n-parameter inference problem as n one-parameter problems. If no reduction of the nuisance vector is applied, this approach is equivalent to simply fitting the full model and basing inference on this fit; otherwise, it may be viewed as an *ad hoc* compromise between the strategies specified above. Empirical-Bayes estimation may be viewed as a more formal compromise: as in the first strategy, coefficients will not be dropped; but, as in the second strategy, large unstable estimates from the full model may be replaced by much smaller estimates.

Hierarchical Methods

In addition to modeling the outcomes as random variables whose distributions are a function of the target parameters β, hierarchical methods model these target parameters as random variables whose joint distribution is a function of hyperparameters and prior covariates z that are thought to determine the magnitudes of the target parameters. For example, in a model in which $\beta_1,\ldots,\beta_6$ represented the carcinogenic effects of six polychlorinated biphenyls $x_1,\ldots,x_6$, the magnitude of each of these effects may be determined by the degree of chlorination of the particular compound. Thus, one would assign to each β_i ($i=1,\ldots,6$) a covariate z_i that measures the chlorination of chemical x_i. The impact of these and other such properties of exposures on carcinogenic activity could then be analyzed using a second-stage model for the first-stage (disease) coefficients β_i. Consider, for example, the regression model

$$\beta_i = z_i\pi + \delta_i,\ i=1,\ldots,n,$$
$$\text{i.e., } \beta = Z\pi + \delta = \mu + \delta, \qquad [2]$$

where z_i is a row-vector of p known prior covariates, π is a column vector of p possibly unknown prior coefficients, the δ_i are independent Gaussian (normal) random variables with mean zero and possibly unknown variance τ^2, and Z is the n-by-p matrix with rows z_i. Equation 2 represents a second stage of the sampling model (Equation 1 being the first stage); it implies that the expected value μ of β equals $Z\pi$, and that the components β_i of β have a common variance τ^2. The distribution of β is traditionally termed the prior distribution for β, because it is supposed to represent or incorporate what is known about β prior to seeing the study data. The hyperparameters μ and τ^2 in Equation 2 are hence termed the prior mean and prior variance of β. The model may be generalized by allowing the variances of the δ_i to vary, so that τ^2 becomes a vector of τ_i.

An extreme version of Equation 2 occurs when Z is the n-by-n identity matrix, in which case Equation 2 allows unrelated means for the components of β; the other extreme occurs when $z_i=1$ for all i in which case Equation 2 implies a common mean for the components. A transitional model would arise if, as above, the study variables were six types of polychlorinated biphenyls, and $z_i=(1,z_i)$ where z_i measures the degree of chlorination of chemical i. Here, the prior mean is a linear function of the chlorination covariate z_i; this allows distinct prior means for the components of β, which are nevertheless related via a covariate (chlorination) believed to be directly linked to strength of effect (as measured by β).

To model effect modification, one may add product terms among the exposures and confounders to x, and then include their coefficients in the second-stage model. Similarly, to more flexibly model dose response, one may add multiple terms (e.g., linear and quadratic) for a single exposure and then include their coefficients in the second-stage model. Building such models involve a number of complexities beyond the scope of the present discussion, however; Cornfield (*11*) presents an example involving interaction in cross-classifications.

The nuisance parameter γ can be included with β in the second-stage model, or dealt with separately. Options for γ will be further described below.

Bayesian Methods

The classical Bayesian approach to inference on β requires that one completely specify the prior distribution for β. Thus, to use Equation 2, one would have to treat the hyperparameters π and τ^2 as known quantities. Bayesian analysis would then proceed by merging the prior distribution for β with the likelihood function for β to obtain a posterior distribution for β (*14*, ch. 10). For simplicity, the present exposition will concentrate on the Gaussian approximation to this analysis.

Suppose that the prior follows Equation 2 and the likelihood for β is well approximated by a multivariate normal density with mean $\hat{\beta}$ and covariance matrix $\hat{V}$, where $\hat{\beta}$ is the maximum-likelihood estimate (MLE) and $\hat{V}$ is the inverse of the observed information matrix evaluated at β; such an approximation is adequate in large-sample logistic and Poisson regression (*15*). The posterior distribution for β will then be approximately Gaussian with mean $B\mu + (I-B)\hat{\beta}$ and covariance matrix $\hat{V}(I-B)$, where $B=(\hat{V}+\tau^2 I)^{-1}\hat{V}$ and I is the n-by-n identity matrix (*6*). Note that the posterior mean is a weighted average of the prior mean μ and the MLE $\hat{\beta}$.

Parametric Empirical Bayes Methods

The "naive" parametric empirical-Bayes approach treats π and τ^2 as unknown parameters, estimates them from the data (via one of several available methods), plugs these estimates into the prior, and then uses this estimated prior in a standard Bayesian analysis to obtain an empirical-Bayes posterior distribution. This "naive" empirical-Bayes approach is unsatisfactory because it fails to account for the uncertainty in the estimated prior parameters $\tilde{\pi}$ and $\tilde{\tau}$. The resulting estimates can, however, be corrected to account for this uncertainty (*6*). With correction, the approximate mean of the empirical-Bayes posterior distribution for β under Equation 2 is

$$\beta^* = B^*\mu^* + (I-B^*)\hat{\beta}, \qquad [3]$$

where

$$B^* = (n-p-2)\,W^*\hat{V}/(n-p)$$

$$W^* = (\hat{V}+\tilde{\tau}^2 I)^{-1}, \mu^* = Z\pi^*, \tilde{\tau}^2 = nR/(n-p) - \bar{V}^*,$$

$$R = e'W^*e/\Sigma_{ij}W_{ij}^*, \ \pi^* = (Z'W^*Z)^{-1}Z'W^*\hat{\beta},$$

$$\bar{V}^* = W^*\hat{V}/\Sigma_{ij}W_{ij}^*, \ e = \hat{\beta} - \mu^*,$$

and W_{ij}^* is element ij of W^* (*6*). An approximate posterior covariance for β is

$$C^* = \hat{V}[I-(n-p)B^*/n] + A, \qquad [4]$$

where $A = 2B^*e(B^*e)'/(n-p)$ is an adjustment term that accounts for estimation of the prior variance (*6*). β^* and C^* can be used to obtain empirical-Bayes confidence regions for β; as illustrated below, these regions can be superior to the analogous conventional regions based on $\hat{\beta}$ and $\hat{V}$(*6*).

As discussed by Morris (*6*) and illustrated in simulations (*10*), the confidence intervals based on C^* can be extremely conservative in small samples when the prior variance is small relative to the first-stage variance. Improvement in small-sample behavior can be obtained by component-wise variance corrections (*6*); with the Morris (*6*, Eq. 5.6–5.9) corrections, the estimated variance of component i of the *EB* estimator is

$$v_i^* = \hat{V}_{ii} - (1-H_{ii}^*)(\hat{V}B^*)_{ii} + (\bar{V}_{ii}^* + \tilde{\tau}^2 I)W_{ii}^* A_{ii}, \qquad [5]$$

where $H^* = Z(Z'W^*Z)^{-1}Z'W^*$, $\bar{V}^* = W^*\hat{V}/\Sigma_{ij}W_{ij}$, and subscript ij on a matrix expression indicates the ijth element of the expression. Greater improvements may be obtained via further approximations (*16*) or via Monte-Carlo-based methods (*17*).

Approximate procedures become more complex if π (as well as δ) is regarded as random (*18*), but such extensions are easily handled by Monte-Carlo methods (*19*). In particular, Monte-Carlo methods render practical a fully Bayesian empirical Bayes analysis, in which π and τ^2 (the unknown hyperparameters of the prior distribution for β) are themselves assigned "hyperprior" distributions (*20*). Such distributions represent a third stage in the modeling process.

Semi-Bayes Methods

Consider Equation 2. It often is possible to specify the prior standard deviation τ on *a priori* grounds, or at least to assign τ a plausible range of values. In logistic regression with binary covariates, β represents a vector of log odds ratios, and it often is possible to set upper and lower prior bounds on these ratios. Typical environmental and occupational surveillance studies involve exposures whose effects are expected to produce small positive log odds ratios. If Z is a vector of ones, setting $\tau = 0.5$ would correspond in this case to being 95% certain that any given ratio is within an exp[2(1.96)(0.5)] = 7-fold range of about the prior geometric mean odds ratio; this would be appropriate if (for example) 19 out of 20 ratios fell between 1 and 7.

Fixing τ at some prior value in the empirical-Bayes analysis shifts both the philosophy and the mechanics of the analysis back toward the classical Bayesian approach; thus, when τ is specified in advance, I refer to the empirical-Bayes analysis as "semi-Bayes." This approach is computationally simpler than standard empirical-Bayes; for example, under Equation 2, π will require iterative estimation using standard empirical-Bayes methods, but has closed form using the semi-Bayes approach (*9*). As shown in simulations (*10*), semi-Bayes confidence regions can be superior to standard empirical-Bayes regions when τ equals or exceeds the true standard deviation of the components of β, although this superiority is purchased by a risk of subnominal coverage if τ is set lower than this.

Under Equation 2, the approximate mean and covariance of the semi-Bayes posterior distribution for β are

$$\tilde{\beta} = B\tilde{\mu} + (I-B)\hat{\beta} \qquad [6]$$

and

$$\tilde{C} = \hat{V}[I-(n-p)B/n] \qquad [7]$$

where

$$\tilde{\mu} = Z\tilde{\pi}, \ \tilde{\pi} = (Z'WZ)^{-1}Z'W\hat{\beta}, \ W = (\hat{V}+{}^2I)^{-1},$$

and $B = W\hat{V}$ (as in the classical Bayesian analysis). Note that, as τ is increased, the semi-Bayes estimate $\tilde{\beta}$ approaches the maximum-likelihood estimate $\hat{\beta}$, and $\tilde{C}$ approaches $\hat{V}$. The variance estimate for component i of each *SB* estimator is

$$\tilde{v}_i = \hat{V}_{ii} - (1-\tilde{H}_{ii})(\hat{V}\tilde{B})_{ii} \qquad [8]$$

where

$$\tilde{H} = Z(Z'\tilde{W}Z)^{-1}Z'\tilde{W}.$$

Preliminary-Test Estimates Revisited

For comparison to Bayesian methods, a preliminary-test estimator (such as a stepwise-regression result) may be viewed as an *ad hoc*, discontinuous shrinkage estimator (*21*): for any given data set, the estimate is either shrunk all the way to zero (if it gets deleted from the equation), or else it is replaced by an estimate from an equation with fewer variables than the full equation (if it is retained but other variables are deleted). This shrinkage rule is somewhat perverse relative to the usual empirical-Bayes rationale. Like empirical-Bayes, unstable estimates are most subject to change, but outliers are poorly handled. Unlike empirical-Bayes, small estimates are much more subject to change than large ones. It would seem natural, then, to expect preliminary-test estimators to have larger mean-squared error than empirical-Bayes estimators, and in fact this has been demonstrated in the multivariate normal model (*22*).

Another problem with preliminary-test estimators is the lack of a valid standard error to attach to the estimate, especially when the estimate is set to zero. A preliminary-test confidence interval may be constructed by using the information matrix for the full model evaluated at the reduced parameter-vector estimate. While this *ad hoc* device performs surprisingly well in small samples, simulation results (*10,23*) indicate that, in terms of coverage and average width, preliminary-test interval estimators still do not perform as well as other intervals.

Nuisance Parameters

A common practice in occupational and environmental epidemiology is to eliminate the nuisance parameters γ from the model by "preadjusting" the outcome variable for w (possibly after some preliminary reduction of w as well). For example, y may be the log of standardized mortality ratio (SMR), the latter being the observed number of cases of disease divided by the number expected given the age and sex (w) composition of the particular exposure group (x-level) under examination. Breslow and Day (*24*) and Checkoway et al. (*25*) present detailed discussions of this method (which implicitly assumes that joint exposure and confounder effects on rates are multiplicative). This approach allows elimination of explicit consideration of γ in the problem of inference on β; concern now focuses on inference on β in an equation

$$g[E(y_w|x)] = \alpha_w + x\beta,$$

where the subscript w indicates that some appropriate preadjustment for w has been made, as distinct from simply dropping w from Equation 1 (no preadjustment of x is needed if the exposure groups are homogeneous with respect to x).

One may instead fit the full first-stage model (Equation 1) directly, and then model both β and γ in a second-stage model analogous to Equation 2. If γ is allowed to have a different variance parameter from β, then, as this variance parameter becomes arbitrarily large, the hierarchical estimates of γ will approach the maximum-likelihood estimates. Also, the hierarchical estimates of β will approach those obtained by modeling β alone (as in Equation 2) when $\hat{\beta}$ is obtained from the full first-stage model (Equation 1). The resulting hierarchical estimates are model-based analogues of those obtained using preadjustment.

An alternative approach would be to apply some type of preliminary-test model reduction to γ when fitting Equation 1, followed by second-stage modeling of the coefficients that remain. Such an approach, however, would be subject to the same objections as other preliminary-test estimators, such as poor handling of outlying estimates.

An Example

Neonatal mortality (death among liveborn infants within the first 28 days after birth) underwent a notable decline in industrialized countries during the middle decades of this century. Neonatal mortality rates in the United States underwent a particularly important decline during the 1960s and 1970s, with the advent of a number of medical innovations, including electronic fetal monitoring and methods for management of prenatal and perinatal risk factors. An interesting question is the relative impact of interventions (such as monitoring), modifiable or potentially treatable factors (such as duration of pregnancy), and fixed patient characteristics (such as maternal age). How much did the decline in neonatal mortality arise from changes in these risk-factor distributions, especially for treatable versus fixed factors?

The Beth Israel Hospital study (*26*) provides limited data bearing on this issue. This cohort study of neonatal death included over 14,000 live births but only about 60 neonatal deaths from 1970 to 1975, after exclusions. The longitudinal nature of the study, and the documented trends over the period, provide means for evaluating the regression methods studied here (albeit indirectly) by comparing the success of the methods in predicting the observed trend.

Table 1 displays the estimates of the coefficients in a logistic regression of neonatal death rates on 14 risk factors recorded in the Beth Israel data set, using only the 2992 subjects in 1970. (Certain other variables were excluded from the original study on *a priori* grounds; for example, Apgar score was excluded because it was considered an intermediate indicator of risk.) The preliminary-test estimates were derived using a 0.10 significance level, as was done in some early (pre-1978) presentations of these data (one additional variable, malpresentation, was significant at the 0.10 level in the full model but was not significant at this level after the other nonsignificant variables were deleted). The semi-Bayes and empirical-Bayes estimates are both based on a unidimensional prior and method of moments, as used in the simulation study (*10*). Risk factor i was assigned a prior covariate value of $z_i = 1$ if the factor was expected to have a positive coefficient, $z_i = -1$ if the factor was expected to have a negative coefficient, and $z_i = 0$ if the factor was expected to have a near-zero coefficient. The prior variance τ^2 for the semi-Bayes estimates was set to 0.25, 0.5 and 1.0, round figures that also happen to be reasonable on prior grounds: For example, $\tau^2 = 0.5$ corresponds to a prior that 95% of the odds ratios for the factor effects, the $\exp(\beta_i)$, are within a $\exp(2(1.96\sqrt{0.5}) = 16$-fold span, such as 0.75 to 12. (To save space, results for $\tau^2 = 0.25$ and 1.0 are omitted from the table.)

Also shown are the posterior means from a Bayesian empirical-Bayes (BEB) analysis based on extending the hierarchical model with third-stage hyperpriors for π (the coefficient of z) and τ^2. The hyperprior for π was Gaussian with mean 0.4 and standard deviation 0.2, while the hyperprior for τ^2 was 5.5 times an inverse χ^2 distribution on 13 degrees of freedom (the residual degrees of freedom in the second-stage regression). The latter hyperprior was chosen because it has mean 0.5 (the fixed τ^2 value in the semi-Bayes analysis) and roughly 90% of the distribution falls between 0.25 and 1 (a prior standard deviation τ of 0.5 to 1.0). The hyperprior for π represents a geometric mean relative risk (per unit change in the expected harmful direction) of $\exp(0.4) = 1.5$, with a 95% hyperprior interval of $\exp[0.4 \pm 1.96(0.2)] = 1.0, 2.2$. This analysis was done via Gibbs sampling (*19,27*) averaging over the final 100 draws from 100 parallel Markov chains of 200 iterations each; using the measure of Gelman and Rubin (*28*), further iterations would probably not have changed any estimate by more than 0.5%.

For clarity and compactness, Table 1 gives standard errors only for the ML estimates. The preliminary-test (PT) estimates had 5 to 10% smaller estimated standard errors than the ML estimates; EB and BEB estimates typically had 20 to 40% smaller estimated standard errors, while the SB estimates typically had 25 to 50% smaller estimated standard errors. In the remaining discussion, I will refer to the EB, SB, and BEB estimates as the hierarchical estimates.

The results in Table 1 are typical of maximum-likelihood versus hierarchical estimates: For the hierarchical results, larger, unstable coefficients (such as for

Table 1. Maximum-likelihood, empirical-Bayes, semi-Bayes, and Bayes empirical-Bayes logistic coefficient estimates for 1970 neonatal-death data.[a]

Predictor[b]	ML	PT[c]	EB	SB	BEB	z
Nonwhite	0.634 (0.631)	0	0.573	0.561	0.483	1
Maternal age	−0.476 (0.728)	0	−0.459	−0.451	−0.345	−1
Multiparity	−0.437 (0.571)	0	−0.441	−0.437	−0.348	−1
Duration pregnancy	−1.587 (0.357)	−1.725	−1.371	−1.346	−1.385	−1
Isoimmunization	1.113 (0.614)	0	0.907	0.880	0.812	1
Previous abortion	−0.328 (0.705)	0	−0.167	−0.148	−0.196	0
Hydramnios	4.099 (1.201)	3.840	1.872	1.540	1.728	1
Dysfunctional labor	−0.231 (0.350)	0	−0.016	0.008	−0.077	1
Placental/cord abn.	1.132 (1.138)	0	0.882	0.847	0.719	1
Electronic monitor	−0.222 (0.697)	0	−0.529	-0.573	−0.398	−1
Multiple birth	2.105 (0.772)	2.143	1.447	1.366	1.408	1
Public ward	−0.146 (0.611)	0	0.295	0.357	0.219	1
PROM[d]	−0.615 (1.116)	0	0.224	0.348	0.068	1
Malpresentation	1.359 (0.756)	0	1.075	1.041	1.008	1

ML, maximum likelihood; PT, preliminary test; EB, empirical-Bayes; SB, semi-Bayes; BEB, Bayes empirical-Bayes; PROM, prolonged rupture of membrane. [a] EB, $\tilde{\tau}^2 = 0.45$; SB, $\tau^2 = 0.50$; BEB, mean $\tau^2 = 0.50$. No. subjects = 2992, no. deaths = 17, no. predictors = 14, death rate = 5.68 per 1000. ML standard errors in parentheses. z = prior covariate (expected coefficient sign). [b] All coded 1 = yes, 0 = no, except maternal age (1 = < 15 years, 2 = 15–19 years., 3 = 20+ years), duration of pregnancy (5,6,7 = 32–35, 36–38, and 39+ weeks; under 32 weeks excluded from analysis), isoimmunization (0 = none, 1 = Rh, 2 = ABO), and dysfunctional labor (0 = normal, 1 = prolonged, 2 = protracted, 3 = arrested). [c] Variables retained at 0.10 significance level. [d] PROM = 30+ hours.

hydramnios) are heavily shifted toward their estimated prior mean, while stable coefficients (such as for duration of pregnancy) are not much changed. It is interesting that three of the nonsignificant ML coefficients appear to be in the wrong direction on subject-matter grounds: dysfunctional labor, public ward, and prolonged rupture of membranes (PROM). All three are shifted in the *a priori* correct direction by the other methods, and two get the *a priori* correct sign from the hierarchical methods. There is no precise prior information to judge the remaining coefficients, however.

Figure 1 shows the results of using the five regressions in Table 1 to predict the neonatal death rates in Beth Israel Hospital in 1971 to 1975, the remainder of the study period. (Not shown is the "null" regression, which would be a flat line at 5.68 per 1000.) Although only 13, 10, 7, 3, and 7 deaths occurred in 1971 to 1975, and the predicted rates from each model are not significantly different from one another, their relative relationship to the observed rate bears an interesting resemblance to small-sample simulation results (*10*): the hierarchical estimates are closer to the true parameters (the observed rates) than are the ML and PT predictions. On average, the hierarchical curves are not closer to the observed curve than the null line, but they do correctly predict the changes between all years but 1971 to 1972. The hierarchical curves are not sensitive to reasonable choices for the hyperparameters. For example, similar patterns are seen for all $0.2<\tau^2<1$ in the semi-Bayes analysis. Note that the ordinary empirical-Bayes and Bayes empirical-Bayes curves are indistinguishable; similar results were obtained using other hyperpriors for π and τ^2.

The discrepancy between the ML/PT curves and the hierarchical curves is almost entirely attributable to the failure of the former to correctly predict the trend over 1970 to 1971. Further analysis (not shown) revealed that hydramnios prevalence jumped in 1970 to 1971 from 3.3 to 9.4 per 1000, and that this jump resulted in the upward shift of the ML and PT curves, due to the large hydramnios coefficients in the ML and PT regressions. In contrast, the hierarchical curves were not so affected by the hydramnios jump because they used severely shrunken hydramnios coefficients. Most of the remaining fluctuations in the observed curve are captured by all four predicted curves, and are largely accounted for by changes in prevalence in the three strongest factors (duration of pregnancy, hydramnios, and multiple birth).

All four predicted curves move away from the observed curve to the same extent over 1971 to 1972. This suggests that most of the discrepancy between the fitted and observed curves was produced by a rapid and lasting decline in some important unmodeled risk factor (such as a medical management factor) around 1971 to 1972. Unlike the conventional analyses, the hierarchical analyses pinpoint this change as occurring in 1971 to 1972 only, as the 1970 to 1971 portion of the decline and the 1972 to 1975 changes can be explained by factors in the model.

The lesser difference among the predictions than seen in the simulated coefficient estimates could have been anticipated: Dempster et al. (*21*) noted lesser accuracy gains of ridge and Stein estimators relative to ordinary least squares when evaluated in terms of prediction error rather than coefficient-estimation error. In one sense, the PT regression does well, in that it captures the predictive power of the full ML regression while using only 3 of the original 14 covariates. On the other hand, the results conform to theoretical work (*29*) indicating that preliminary-test estimates will often do no better than least-squares estimates and tend to do worse than empirical-Bayes estimates in prediction problems.

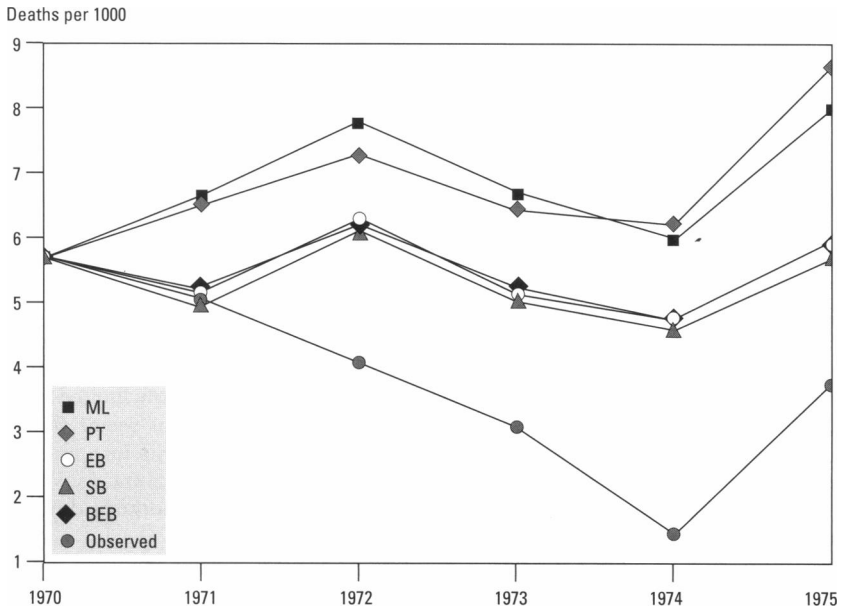

Figure 1. Observed and predicted death rates.

Discussion

In this article, I have chosen to focus on empirical-Bayes estimation rather than the more general subject of "shrinkage" estimation, which subsumes Stein estimation and ridge regression (*29–31*). While it is possible to carry out analyses similar to empirical-Bayes and semi-Bayes analyses entirely within the non-Bayesian context of logistic ridge regression (*30*), hierarchical methods have an advantage in the interpretability of the tuning parameter that controls the degree of coefficient shrinkage: in empirical-Bayes and semi-Bayes regression, the tuning parameter is the prior variance, which has a direct subject-matter interpretation, and the inferences are more easily seen to be approximately Bayesian.

While the large-sample standard errors may be questionable in the above example (with 14 predictors for 17 deaths), simulation studies (*10*) indicate that their relative magnitudes do indeed reflect the relative precision of the estimators, even in small samples. Considering mean squared error, such studies also indicate that the variance reduction of the hierarchical estimates (relative to ML) more than compensates for the fact that hierarchical estimates are biased towards their prior means. Furthermore, the ML estimates are only unbiased in the large-sample sense, and so are not guaranteed to have any small-sample bias advantage over hierarchical estimates.

Simulations provide a quantitative estimate of the gains one may expect in employing hierarchical methods in epidemiologic studies of multiple parameters. The following simulation results (*10*) seem

especially relevant to formulating recommendations for applications:

- In small samples with few exposures and subjects, ordinary empirical Bayes did not provide dramatic benefits over ordinary maximum likelihood.
- In contrast, the semi-Bayes variant of empirical Bayes appeared superior to the other simulated methods in very small studies, in that it provided narrower intervals with coverage robust to variance misspecification in such studies.
- In large samples with many covariates (i.e., in large studies), there was little difference in performance between empirical Bayes and semi-Bayes with correct specification. Semi-Bayes had the disadvantage of sensitivity to misspecification; empiricalBayes was clearly superior to ML and preliminary testing.
- Semi-Bayes estimation appeared robust to overspecification of the prior standard deviation (setting τ too large), in that overspecification did not reduce confidence-interval coverage and did not significantly decrease small-sample precision.
- Preliminary testing, perhaps the most common approach to handling many exposures, did not appear to have the precision or validity of ordinary empirical-Bayes when the number of covariates was large; it also lacked large-sample validity. Its validity could be improved by raising the significance level, but then its precision could drop below that of ordinary maximum likelihood.

Based on these observations, one might recommend semi-Bayes methods for very small studies, empirical Bayes for very large studies, and either method for the spectrum in between, with a caution to overspecify the semi-Bayes prior variance if either the sample size or number of parameters is not small. Both methods, however, can be subsumed under and replaced by Bayesian empirical Bayes (BEB) methods: ordinary empirical Bayes arises as the limiting case of BEB as the hyperpriors for π and τ^2 become diffuse; semi-Bayes arises as the limiting case of BEB as the prior for π becomes diffuse, with a one-point hyperprior for τ^2.

If used with a proper hyperprior (as in the above example), Bayesian empirical Bayes may offer the best compromise between the small-study precision of semi-Bayes and the large-sample robustness of ordinary empirical Bayes. Furthermore, due to recent developments in Monte-Carlo and approximation methods (*16,19,27,28,32–35*), Bayesian empirical-Bayes analyses are now computationally practical. Nevertheless, because of their theoretical and computational sophistication, I suspect such methods will remain far removed from routine application in the near future, despite an abundance of applications in which they could be used. In the hopes of encouraging more applications of hierarchical methods, the present article has focused on the computationally simpler ordinary empirical Bayes and semi-Bayes methods, which can be easily programmed with a matrix language (such as GAUSS, SAS IML, S-Plus, or SC), run rapidly on a personal computer, and appear to provide good approximations to fully Bayesian results in some common situations.

Given that hierarchical methods can be recommended for multiple regression analysis, there remains one major problem in implementation: design of the structure of the prior (in particular, specification of the prior design matrix Z). The problem of prior specification is a familiar one in Bayesian analysis, and has remained a major obstacle to wide use of classical Bayesian methods. Nevertheless, the problem may be less intractable in hierarchical analysis: The major specification demand is that the investigator identify subsets of parameters within which the parameters may be regarded as "exchangeable" (possibly after location and scale transforms of the parameters, which must also be specified). Here, exchangeability means that the parameters within a subset may be regarded as draws from a common prior distribution, in much the same way as effects are modeled in random-effect and mixed-model analysis of variance (*8*). This ANOVA parallel is helpful in orienting the problem to a more widely taught and used context, and in pointing out the major limitations of hierarchical methods: if subsets of exchangeable parameters cannot be identified, the methods cannot be applied. For example, in simple descriptive epidemiology, one often begins by examining the dependence of disease occurrence on basic demographic variables such as age, sex, and race; the coefficients of these variables could hardly be imagined as draws from a common distribution, even after transformations. More generally, one must be able to specify a second-stage regression model and error-covariance structure that reflects dependencies among the first-stage parameters. Such specification demands even greater subject-matter familiarity than ordinary regression analysis.

The semi-Bayes method requires additional specification of the prior standard deviation, τ. This step may be viewed as an elicitation problem similar to those encountered in classical Bayesian analysis. The objective is to specify a value that is an upper bound on the range of parameter values that would correspond to expert opinions. In practice, I have found that the potential range for this "minimal conservative" τ is very narrow, and that its elicitation is simple: even in the most vociferous epidemiologic controversies, the span of relative-risk values posited for dichotomous causal factors rarely exceeds 25-fold and is usually within 10-fold. Allowing for a 5% chance that the entire span of expert opinion is too narrow leads one to specify $\tau = 1n(25)/2(1.96) \doteq 0.8$ in the former case and $\tau = 1n(10)/2(1.96) \doteq 0.6$ in the latter. Furthermore, τ may be allowed to vary with the prior covariates.

The prior specification problem is further mitigated by the fact that absolute stringency in specification of exchangeability or other aspects of the prior does not seem necessary to realize benefits from hierarchical methods. The neonatal-death example illustrates this point: the prior covariate used here is clearly naive; at the very least, an obstetrician would want to distinguish *a priori* strong risk factors (such as hydramnios) from *a priori* weak factors (such as ward). Yet, despite the naive prior, hierarchical methods produced better predictions of observable quantities than the usual methods. This result is not an isolated case: other applications of simple hierarchical methods have produced similar results in a variety of different contexts. Morris (*6*) presents an example and provides references to other examples. There is apparently much robustness in hierarchical methods, at least within the realm of applications considered to date. The chief caution seems to be that it is better to err on the vague side than the stringent side when specifying prior distributions.

Hierarchical methods have demonstrable advantages over conventional methods and are no longer seriously limited by computer hardware. It thus seems timely to introduce such methods into epidemiologic teaching and software, as was done with risk-regression methods during the 1970s and 1980s.

REFERENCES

1. Miettinen OS. Theoretical Epidemiology. New York:John Wiley and Sons, 1985.
2. Rothman KJ. Modern Epidemiology. Boston:Little Brown and Co., 1986.
3. Rothman KJ. Multiple comparisons are not a problem. Epidemiology 1:43–46 (1990).
4. Greenland S, Robins JM. Empirical-Bayes adjustments for multiple comparisons are sometimes useful. Epidemiology 2:244–251 (1991).
5. Bishop YMM, Feinberg SE, Holland PW. Discrete Multivariate Analysis: Theory and Practice. Cambridge, MA: MIT Press, 1975.
6. Morris CN. Parametric empirical Bayes: theory and applications (with discussion). J Am Stat Assoc 78:47–65 (1983).
7. Thomas DC, Semiatycki J, Dewar R, Robins J, Goldberg M, Armstrong BG. The problem of multiple inference in studies designed to generate hypotheses. Am J Epidemiol 122:1080–1095 (1985).
8. Louis TA. Using empirical Bayes methods in biopharmaceutical research. Stat in Med 10:811–829 (1991).
9. Greenland S. A semi-Bayes approach to the analysis of correlated multiple associations, with an application to an occupational cancer-mortality study. Stat in Med 11:219–230 (1992).
10. Greenland S. Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary testing, and empirical-Bayes regression. Stat in Med 12:717–736 (1993).
11. Cornfield J. Bayesian estimation for higher order cross-classifications. Milbank Mem Fund Q 48:57–70 (1970).
12. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic Research: Principles and Quantitative Methods. Belmont: Lifetime Learning Publications, 1982.
13. Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. Am J Epidemiol 129:125–137 (1989).
14. Cox DR, Hinkley DV. Theoretical Statistics. New York: Chapman and Hall, 1974.
15. McCullagh P, Nelder JA. Generalized Linear Models, 2nd ed. New York:Chapman and Hall, 1989.
16. Kass RE, Steffey O. Approximate Bayesian inference in conditionally independent hierarchical models. J Am Stat Assoc 84:717–726 (1989).
17. Laird NM, Louis TA. Bootstrapping empirical Bayes estimates to account for sampling variation (with discussion). J Am Stat Assoc 82:739–757 (1987).
18. Laird NM, Louis TA. Empirical Bayes ranking methods. J Educ Stat 14:29–46 (1989).
19. Gelfand AF, Smith AFM. Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 85:398–409 (1990).
20. Deeley JE, Lindley DV. Bayes empirical Bayes. J Am Stat Assoc 76:833–841 (1981).
21. Dempster AP, Schatzoff M, Wermuth N. A simulation study of alternatives to ordinary least squares (with discussion). J Am Stat Assoc 72:77–106 (1977).
22. Sclove SL, Morris C, Radhakrishna R. Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. Ann Math Stat 43:1481–1490 (1972).
23. Hurvich CM, Tsai C-L. The impact of model selection on inference in linear regression. American Stat 44:214–217 (1990).
24. Breslow NE, Day NE. Statistical Methods in Cancer Research, vol 2, The Design and Analysis of Cohort Studies. Lyon:International Agency for Research on Cancer, 1987.
25. Checkoway H, Pearce NE, Crawford-Brown DJ. Research Methods in Occupational Epidemiology. New York:Oxford University Press, 1989.
26. Neutra RR, Feinberg SE, Greenland S, Freidman EA. The effect of fetal monitoring on neonatal death rates. N Eng J Med 299:324–326 (1978).
27. Gilks WR, Wild P. Adaptive rejection sampling for Gibbs sampling. Appl Stat 41:337–348 (1992).
28. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences (with discussion). Stat Sci 7:457–511 (1992).
29. Copas JB. Regression, prediction, and shrinkage. J R Stat Soc Ser B 45:311–354 (1983).
30. Le Cessie S, Houwelingen JC. Ridge estimators in logistic regression. Appl Stat 41:191–201 (1992).
31. Efron B, Morris CN. Data analysis using Stein's estimator and its generalizations. J Am Stat Assoc 70:311–319 (1975).
32. Smith AFM, Skene AM, Shaw JEH, Naylor JC. Progress with numerical and graphical methods for Bayesian statistics. Statistician 36:75–82 (1987).
33. Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. J Am Stat Assoc 82:528–550 (1987).
34. Geweke J. Antihetic acceleration of Monte Carlo integration in Bayesian inference. J Economet 38:73–90 (1988).
35. Smith AFM, Roberts GO. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. J R Stat Soc Ser B 55:3-102 (1993).