

METHODOLOGY

Bayesian perspectives for epidemiological research. II. Regression analysis

Sander Greenland

Accepted 15 August 2006

This article describes extensions of the basic Bayesian methods using data priors to regression modelling, including hierarchical (multilevel) models. These methods provide an alternative to the parsimony-oriented approach of frequentist regression analysis. In particular, they replace arbitrary variable-selection criteria by prior distributions, and by doing so facilitate realistic use of imprecise but important prior information. They also allow Bayesian analyses to be conducted using standard regression packages; one need only be able to add variables and records to the data set. The methods thus facilitate the use of Bayesian solutions to problems of sparse data, multiple comparisons, subgroup analyses and study bias. Because these solutions have a frequentist interpretation as ‘shrinkage’ (penalized) estimators, the methods can also be viewed as a means of implementing shrinkage approaches to multiparameter problems.

Keywords Bayesian methods, biostatistics, odds ratio, relative risk, risk assessment

Introduction

A number of authors have argued that the Bayesian perspective needs to be incorporated into basic statistical training.^{1,2} This training is easily accomplished, insofar as Bayesian analysis can be carried out using conventional (frequentist) formulas for stratified analysis: one may use either information (inverse-variance) weighted averaging of current-data results with a prior distribution, or representation of the prior as a prior-data stratum (data augmentation).^{2,3} Thus, Bayesian analysis does not require posterior sampling or other special software, and does not even require explicit use of Bayes theorem. The prior-data method also has the advantage of displaying the strength of the prior in terms of an ‘informationally equivalent’ experiment, which can reveal overconfidence in prior opinions.

The current installment describes extensions of the prior-data method to regression analyses. It begins with a rationale section, focusing on Bayesian regression for sparse and highly multivariate data as an alternative to variable selection and its attendant problems. That section is followed by review of the 2×2 table case. The remaining sections cover Bayesian analyses of risk and rate regression models using only ordinary (frequentist) software. The methods derive from the theory of data augmentation priors for generalized linear models,⁴

which encompasses linear, polytomous, and ordinal regression, as well as the models discussed here.

Bayesian methods for sparse data and shrinkage

Bayesian and equivalent approaches offer significant advantages over conventional frequentist methods, especially for sparse data (data with few or no subjects at crucial combinations of variable values, e.g. few exposed cases). In epidemiology, data sparsity often results in inflated estimates from unconditional and conditional logistic regressions.^{5–7} Although the problem is most common when using many regressors, it can occur with only one regressor.⁶ Deleting variables with variable-selection algorithms (e.g. stepwise regression) is a common response to the problem. Such algorithms rarely correct the problem, because the inflated coefficients are usually selected and can remain inflated after deletion of other variables.^{5,8} Furthermore, variable selection may discard key confounders in favour of less relevant variables, because they are based on only the confounder-disease association, whereas the confounder-exposure distribution also governs confounding.⁹ Finally, as has been documented by theory and simulation, standard algorithms produce invalidly small *P*-values and invalidly narrow confidence intervals; see Rothman and Greenland¹⁰ p. 402 for citations.

Departments of Epidemiology and Statistics, University of California, Los Angeles, CA 90095-1772, USA. E-mail: lesdomes@ucla.edu

As an alternative to variable selection, *shrinkage estimation* (including empirical-Bayes, semi-Bayes, ridge regression, peanlized estimation and Stein estimation) pulls or ‘shrinks’ coefficient estimates toward prior patterns or values (the term ‘shrinkage’ is a bit misleading, because the coefficients may be pulled toward prior values larger than observed, and so may inflate rather than shrink). The relative degree of pull is roughly proportional to the variance of the original estimate, so unstable coefficients are pulled more than stable ones. This process usually improves overall accuracy of estimation and prediction (e.g. it produces sets of estimates with lower total mean-squared error).

When the prior values are all zero, the coefficients are pulled toward zero (hence the term ‘shrinkage’). The process can then be viewed as a way of allowing partial entry of regressors into the model. When the coefficients in question are of limited size, shrinkage toward zero can dramatically outperform maximum likelihood and conventional variable-selection procedures such as stepwise regression.^{5,8,11–15} Consider a regressor whose effect on the outcome and hence whose importance as a predictor or a confounder is uncertain: zero seems the most likely value for its coefficient, but one wishes to allow for the possibility that it may be non-zero.¹⁶ Conventional variable selection retains the regressor only if it passes a significance test (i.e. if its *P*-value is below some α -level such as 0.10 or 0.05) or if adjustment for it results in more than some minimum change in the study-exposure coefficient (e.g. 10%). These are both all-or-nothing strategies that can result in distorted *P*-values and confidence limits.^{17,18} Variable-selection problems also arise in constructing propensity scores.¹⁹

As an alternative, one simple but effective Bayesian shrinkage method places a prior distribution on each uncertain coefficient, with a single peak (mode) at its expected value (e.g. zero). The degree of adjustment for the regressor is then determined by the spread of the prior around this expectation: the narrower the prior, the more the prior pulls or shrinks the coefficient towards its prior expectation, thereby reducing adjustment for the regressor if its expected coefficient is zero. Thus, the analyst can control the degree of adjustment by setting the spread of the prior (e.g. by specifying the prior variance). The other factor determining adjustment is how clearly and strongly the regressor appears related to the outcome in the conventional analysis (through the variance and size of the conventional estimate, which is averaged with the prior). The final adjustment will then be an average of adjustment based on the prior expectation and conventional adjustment. Regressors already known to have important effects (typically, age and sex) can and should be excluded from this shrinkage process.^{5,8,20}

Frequentist versions of partial adjustment estimate the prior from a hierarchical data model,²¹ average the unadjusted and adjusted estimates,²² or treat the prior as a smoothing device.^{23,24} The Bayesian approach has the advantage of being able to incorporate actual prior information, and consequently can be extended to include priors for *unmeasured* regressors.^{25–29} This extension is an example of non-identified bias modelling, which will be discussed in a later installment.

Review of tabular priors

Table 1 shows the neonatal-mortality experience during the first full year of electronic fetal monitoring at a

Table 1 Cohort data on electronic fetal monitoring and neonatal death,³⁰ and maximum-likelihood (ML) estimates, coding no monitoring as $X = 1$

	$X = 1$		$X = 0$
Deaths ($Y = 1$)	14	3	odds ratio = $\exp(b_{ML}) = 1.412$
Total	2298	694	(risk ratio = 1.409)

$b_{ML} = \log \text{odds ratio} = \ln(RR) \text{ estimate} = \ln(1.412) = 0.3449$.
 $v_{ML} = \text{estimated variance} = 0.4066$.
 95% Wald confidence limits = $\exp\{0.3449 \pm 1.96(0.4066)^{1/2}\} = 0.40, 4.9$.

Table 2 Notation for 2×2 prior table

	$X=1/S$	$X=0$	
Cases ($Y = 1$)	A_1	A_0	$RR_{table} = (A_1/N_1)/(A_0/N_0)$
Total	N_1	N_0	

teaching hospital.^{8,30} To keep the risk ratios above 1, ‘no monitoring’ is taken as the ‘exposed’ ($X = 1$) category. With this coding, a positive association was expected, but with little certainty of the size. Because the outcome (death, $Y = 1$) is extremely rare, we can use odds ratios to approximate risk ratios. Based on the standard formulas,¹⁰ the log odds ratio and estimated variance from Table 1 are 0.3449 and 0.4066, hence the odds ratio is $\exp(0.3449) = 1.4$ with 95% approximate (Wald) confidence limits of $\exp\{0.3449 \pm 1.96(0.4066)^{1/2}\} = 0.40, 4.9$.

Approximating a lognormal prior

Suppose we wish to examine the impact of averaging the results in Table 1 with a lognormal prior for the risk ratio RR , or (equivalently) using a normal prior for $\beta = \ln(RR)$ with mean = median = mode = m_{prior} and variance v_{prior} . The data-prior (data augmentation) approach translates this prior into a prior-data stratum (Table 2). One fills in the table with numbers from a hypothetical study in which the outcome is very rare (so rare that odds and risk ratios are interchangeable), and that yield m_{prior} as the maximum-likelihood estimate (MLE) of β , with the prior variance v_{prior} as its variance estimate. Specifically, we construct a prior table with

- (i) $RR_{table} = (A_1/N_1)/(A_0/N_0) \approx \exp(m_{prior})$
- (ii) $1/A_1 + 1/A_0 \approx v_{prior}$.

To best approximate normality, we set $A_1 = A_0 = A$, which makes $RR_{table} = N_0/N_1 \approx \exp(m_{prior})$ and $A \approx 2/v_{prior}$. Thus the prior-table equations are⁵ $A = 2/v_{prior}$ and $N_0 = RR_{table}N_1$, where we make N_1 and N_0 much larger than A . For example, with $m_{prior} = \ln(2)$ and $N_1 = 10^5$ we get $N_0 = 2(10^5)$, and with $v_{prior} = 0.5$ we get $A = 2/0.5 = 4$. These prior data give back $m_{prior} \approx \ln(2)$ and $v_{prior} \approx 1/4 + 1/4 = 0.5$, as desired (Table 3).

Prior specification usually begins with a prior interval for RR rather than a mean and variance for β . Suppose we give $P\%$ certainty or $P/(1-P)$ odds that RR is between RR_{low} and RR_{up} , with equal certainty (1:1 odds) of being above the interval as below. Using the lognormal approximation for the RR prior, we get $RR_{table} = N_0/N_1 = (RR_{low} \times RR_{up})^{1/2}$. Next, let Z_P be the

value such that a standard normal variate has P% chance of falling between $-Z_P$ and Z_P . Then

$$v_{\text{prior}} = \left\{ \frac{\ln(\text{RR}_{\text{up}}/\text{RR}_{\text{table}})}{Z_P} \right\}^2 \approx \frac{2}{A}$$

so

$$A \approx \frac{2}{v_{\text{prior}}} = 2 \left\{ \frac{Z_P}{\ln(\text{RR}_{\text{up}}/\text{RR}_{\text{table}})} \right\}^2$$

With 2:1 odds (2/3 certainty) of being between 1 and 4, we get $\text{RR}_{\text{table}} = (1 \times 4)^{1/2} = 2$, $P = 67\%$, $Z_P = 0.97$, and $A = 2(0.97/\ln(4/2))^2 \approx 4$, as in Table 3. With 95% certainty of being between 1/2 and 2 we get $\text{RR}_{\text{table}} = (1/2 \times 2)^{1/2} = 1$, $Z_P = 1.96$, and $A = 2\{1.96/\ln(2/1)\}^2 \approx 16$.

A data prior can be viewed as the outcome of a thought experiment. One imagines a perfectly valid study (e.g. a huge perfect randomized trial) in a very low-risk setting, and a result that would justify the RR prior. The point estimate of $\text{RR}_{\text{table}} = \exp(m_{\text{prior}})$ from this study represents the indifference point (prior median); one would bet roughly that RR is as likely above this point as below it (equal or 1:1 odds). Because N_1 and N_0 are so large, the prior information hinges on A_1 and A_0 alone. N_1 and N_0 are forced to be large only for mathematical simplicity; this artifact may be dispensed with by using offset methods (below). The methods are unchanged when one is studying rate ratios rather than odds or risk ratios; N_1 and N_0 then represent person-time rather than persons. They are also unchanged for case-control studies; N_1 and N_0 may then represent numbers of non-cases rather than totals, if so desired.

Approximating the posterior percentiles

To get approximate posterior percentiles for β , we may summarize over the actual-data table and the prior-data table using any statistically efficient method, such

Table 3 Examples of prior-data tables for $\beta = \ln(\text{RR})$ with mode at $m_{\text{prior}} = \ln(2)$. Data for approximately normal prior for β with variance 1/2

	X=1	X=0	
Cases	4	4	$\text{RR}_{\text{table}} = \exp(m_{\text{prior}}) = 2$
Total	10^5	$2 \cdot 10^5$	

Approximate variance = $1/4 + 1/4 = 0.5$.

Approximate 95% prior limits = $\exp\{\ln(2) \pm 1.96(1/2)^{1/2}\} = 2(0.250, 4.00) = 0.50, 8.0$.

2.5%, 97.5% cutoffs for $F(8,8) = 0.226, 4.43$.

Data rescaled by $S = (0.5 \times 400/2)^{1/2} = 10$ by setting $X = 1/10$, in order to more closely approximate a normal prior for β with variance 0.5

	X=1/10	X=0	
Cases	400	400	$\text{RR}_{\text{table}} = \exp(m_{\text{prior}})^{1/10} = 2^{1/10}$
Total	10^5	$2^{1/10}10^5$	

Approximate variance = $10^2(1/400 + 1/400) = 1/2$.

Approximate 95% prior limits = $\exp\{\ln(2) \pm 1.96(1/2)^{1/2}\} = 2(.250, 4.00) = 0.50, 8.0$.

Exact 2.5%, 97.5% cutoffs for $F(800, 800) = 0.87055, 1.1487$.

Exact 95% prior limits for $\text{RR} = 2(0.87055^{10}, 1.1487^{10}) = 0.50, 8.0$.

as maximum likelihood (ML) or information (inverse-variance) weighting of the log odds ratios. If the actual data are stratified, the prior table becomes just another stratum and the summary is done over all the strata (actual and prior). The resulting point estimate and variance for a common log odds ratio approximates the posterior mode m_{post} and variance v_{post} for β . Maximum-likelihood summarization over Table 1 and Panel 1 of Table 3 gives approximate posterior mode $m_{\text{post}} \approx 0.5054$ and variance $v_{\text{post}} \approx 0.2385$ for β . These yield an approximate posterior median for RR of $\exp(m_{\text{post}}) \approx 1.66$ and approximate (Wald) 95% posterior limits of $\exp\{0.5054 \pm 1.96(0.2385)^{1/2}\} = 0.64, 4.3$.

Perfecting the normal approximation

If A is under 4, the actual prior probability of being in the interval will be somewhat less than the normal approximation indicates. For example, the exact prior probability that RR is between 1/2 and 8 implied by Table 3 is 93.3%, as opposed to the 95% implied by the normal approximation, and the exact 95% prior limits are 0.45 and 8.9. One might find this acceptable insofar as the prior is actually weaker than the approximation suggests. Nonetheless, one can instead make the prior implied by the data perfectly normal by using the rescaling factor S. This is done by

- (i) dividing X in the prior data by S, and
- (ii) setting $\text{RR}_{\text{table}} = N_0/N_1 = \exp(m_{\text{prior}}/S) = \exp(m_{\text{prior}})^{1/S}$

The prior-data table now has $X = 1/S$ in the left column and 0 in the right column, even if 1/S is a meaningless value for X (as in the monitoring example). The prior variance for β implied by this rescaled-data prior is $v_{\text{prior}} \approx S^2(2/A)$. Thus, if we want essentially perfect normality we need only make A huge and set $S = (v_{\text{prior}} \times A/2)^{1/2}$ to compensate.³¹

Making $A = 400$ forces normality to an accuracy far beyond actual epidemiological data. For example, with $m_{\text{prior}} = \ln(2)$, $v_{\text{prior}} = 0.5$, $A = 400$, and $N_1 = 10^5$, we get $S = (0.5 \times 400/2)^{1/2} = 10$ and $N_0 = \exp\{\ln(2)/10\}10^5 = 2^{1/10}10^5$. These prior data are shown in Panel 2 of Table 3. The exact and approximate 95% prior limits for RR from this table are both 0.50, 8.0, as desired.

Other tabular priors

When a prior table in which the outcome is rare is entered as a stratum in an ML summarization procedure, it imposes a prior distribution for $\beta = \ln(\text{RR})$ of the generalized-conjugate (log-F) form.³²⁻³⁴ This form is approximately normal when $A_1 = A_0 > 4$, which is the basis for the procedures here and in Greenland.² To skew this β prior, one need only make A_1 and A_0 disparate; and to make the tails of the prior heavier (thicker) than a normal, one need only make A_1 and A_0 small. Use of skewing along with a rescaling factor S allows considerable latitude in the prior shape and spread; for further details see Greenland.^{3,32}

Bayesian regression via tabular augmentation

Table 4 presents exposed-case counts, priors, maximum-likelihood results, and Bayesian results from multiple logistic regression of neonatal death simultaneously on 14 regressors in the cohort in Table 1. The 14 coefficients ($\beta_1, \dots, \beta_{14}$) were given normal priors with $v_{\text{prior}}=0.5$ and $m_{\text{prior}}=0, \ln(2)$, or $\ln(4)$. The 3 priors correspond to factors identified by clinicians as 'uncertain direction' (95% prior limits of 1/4,4), 'probably positive' (1/2,8; 83.3% probability or 5:1 odds of being above 1), and 'probably strong' (1,16; 97.5% probability of being above 1). Variables were recoded and rescaled as shown in the footnote to justify reduction to these three groups. For example, early age was recoded 0,1,2 and labour progress was recoded 0, 1/3, 2/3, 1, reflecting that one age increment and labor arrest (labor progress=1 vs 0) would be given similar (although independent) priors.

Standard methods for fitting risk and rate models, such as ML, weighted least squares (information weighting), and generalized and weighted estimating equations (GEE and WEE), are based on large-sample approximations. These methods can entail considerable bias (usually away from the null) when the number of outcomes is small.⁵⁻⁷ Here, the hydramnios result appears most affected, being an order of magnitude above clinical expectation. Stepwise regression using $\alpha=0.05$ selects only gestation, hydramnios, and multiple birth, but does not improve predictions or bring the hydramnios coefficient into a plausible range.⁸ These failings are general

defects of conventional variable-selection algorithms, reflecting that they do not exploit prior information, and that effects of omitted variables are picked up by (and hence confound) the estimates for the retained variables.⁹

Fortunately, the Bayesian approach can address the sparse-data problem through use of priors, rather than through variable selection. In this example, the Bayesian results appear more reasonable, as logically they must if the priors appear reasonable. More importantly, relative to the ML and stepwise results, the Bayesian regression produced more accurate mortality predictions for later years,⁸ and thus would have been better for planning purposes; it also suggests that the prior information, although vague, improved the validity of the fitted model. This improvement arose largely from shrinkage of the hydramnios odds ratio from 60 to 6.

Computing the Bayesian results required only entering the priors as data and re-running the logistic regression program, as described next; it is thus computationally less intensive than either typical variable-selection algorithms (which can require sequential model fitting) or typical Bayesian methods (which can require extensive Monte-Carlo simulation). More importantly, it replaces arbitrary selection criteria (e.g. $P < 0.05$ to enter) with subject-matter considerations that can improve predictive performance.

Data priors for binary indicators

Suppose we have a binary risk indicator X_j with coefficient β_j in a logistic or Poisson (exponential) regression of Y on several variables, and a prior table for β_j of the form in Table 2

Table 4 Multiple logistic regressions of neonatal-death risk in a cohort of 2992 births with 17 deaths; intercept and 14 regressors ($j=1, \dots, 14$) in each model.³¹ Shown are the prior median and 95% prior limits; estimate and 95% Wald limits for $\exp(\beta_j)$ from maximum-likelihood; and approximate posterior median and 95% Wald limits using independent normal priors for the β_j represented as prior data (see Greenland,³¹ Table 2 for complete profile-posterior and posterior-sampling results)

Regressor (X_j) ^a	Deaths with $X_j > 0$	Prior median (95% limits)	ML estimate (95% limits)	Approximate posterior median (95% limits)
Non-white	5	2 (1/2, 8)	1.9 (0.55, 6.5)	1.8 (0.73, 4.3)
Early age	3	2 (1/2, 8)	1.6 (0.39, 6.7)	1.6 (0.63, 4.1)
Nulliparity	8	2 (1/2, 8)	1.5 (0.51, 4.7)	1.5 (0.69, 3.5)
Gestation age	10	4 (1, 16)	4.9 (2.4, 10)	4.5 (2.6, 8.0)
Isoimmune	1	4 (1, 16)	3.0 (0.91, 10)	2.4 (0.94, 6.2)
Past abortion	2	1 (1/4, 4)	0.72 (0.19, 2.9)	0.83 (0.33, 2.1)
Hydramnios ^b	1	4 (1, 16)	60 (5.7, 635)	6.1 (1.6, 23)
Labour progress	2	2 (1/2, 8)	0.50 (0.06, 3.9)	1.2 (0.43, 3.5)
PCA	1	2 (1/2, 8)	3.1 (0.33, 29)	2.3 (0.67, 7.5)
No monitor	3	2 (1/2, 8)	1.2 (0.32, 4.9)	1.7 (0.70, 4.3)
Twin, triplet	3	4 (1, 16)	8.2 (1.8, 37)	5.2 (1.9, 15)
Public ward	6	2 (1/2, 8)	0.86 (0.26, 2.9)	1.3 (0.55, 3.0)
PROM	1	2 (1/2, 8)	0.54 (0.06, 4.8)	1.2 (0.43, 3.5)
Malpresented	3	4 (1, 16)	3.9 (0.88, 17)	3.9 (1.4, 10)

PCA=placental or cord abnormality, PROM=premature rupture of membranes.

Intercept (β_0) prior is normal with 95% limits of $\ln(0.0001)$, $\ln(0.005)$.

^a All are indicators except early age (0=20+, 1=15-19, 2=under 15), gestation age (0=over 38 weeks, 1=36-38, 2=33-35; <33 excluded), isoimmune (0=no, 1=Rh, 2=ABO), labour progress (0=normal, 0.33=prolonged, 0.67=protracted, 1=arrested).

^b Profile-likelihood (likelihood-ratio) confidence limits are 2.8, 478; profile-posterior (penalized likelihood-ratio) limits are 1.6, 22; Metropolis-sampling limits are 1.5, 22.

(but with entries subscripted by j). One can impose this prior on the analysis by adding a pair of prior-data records:³⁵ One record with $X_j=1/S_j$ and A_{1j} cases out of N_{1j} total, the other with $X_j=0$ and A_{0j} cases out of N_{0j} total (N_{1j} and N_{0j} are person-counts in logistic regression, person-time in Poisson regression). One also adds a new regressor to the model to indicate whether a record is from the X_j -prior or from the actual data; if we call this indicator 'Prior $_j$ ', we set $\text{Prior}_j=1$ for the two prior records, and $\text{Prior}_j=0$ for actual data.

In both the prior records, each remaining regressor is set to a single reference value, so as not to influence the estimates of the remaining coefficients. The reference value does not matter mathematically because the stratum effect is absorbed by the prior indicator Prior_j ; it could be zero or a value close to the sample mean of the regressor. For example, suppose the remaining regressors are age and systolic blood pressure (SBP). If one chose reference values of 50 years and 120 mm, both prior records would be assigned 50 years for age, 120 mm for SBP, and $\text{Prior}_j=1$ for the regression; the remaining records would be unchanged but for having the new indicator $\text{Prior}_j=0$ added. For reasons given below, however, it is best to instead recenter each regressor to make zero its reference value, by subtracting the reference value from the regressor. Then, in the actual records, age would be replaced by age-50, SBP would be replaced by SBP-120. Zero would be the reference value of these recentered regressors, and would thus be their value in the prior records.

The Bayesian results in Table 4 were derived from an ordinary logistic regression program by adding 15 prior indicators and 30 records to represent the 15 prior tables. Using $S_j=10$, we get $\text{RR}_{\text{table}}=1, 2^{1/10}$, or $4^{1/10}=1, 1.07177$, or 1.14870 ; the first record of each pair has $A_1=400, N_1=10^5$, and $X=1/10$, while the second record of each pair has $A_0=400, N_0=\text{RR}_{\text{table}}10^5=100\,000, 107\,177$, or $114\,870$, and $X=0$. Thus the prior-record pair for the hydramnios coefficient β_7 is

Cases	Total	X_7	Prior $_7$	
400	100 000	1/10	1	(all other regressors zero)
400	114 870	0	1	(all other regressors zero)

Despite the extreme data sparsity, nearly identical results follow from other posterior approximations such as nonlinear least squares, penalized likelihood, and Metropolis sampling.^{8,31}

Quantitative regressors

The method of entering the prior as tabular data may be used for the coefficient of a quantitative variable X_j . The prior data now represent a comparison of $X_j=1/S_j$ unit vs $X_j=0$ units of the variable. To aid in setting a reasonable prior, the variable should first be recentered and rescaled so that 0 is a meaningful value and a 1-unit change is contextually meaningful. For example, diastolic blood pressure could be recentered so that 80 mm is the zero point, then rescaled from millimetre to centimetre units (so 95 millimetre and 70 mm become $(95 - 80)/10=1.5$ cm and $(70 - 80)/10=-1$ cm); smoking intensity could be rescaled from cigarettes/day to packs/day; and vitamin C intake in milligrams/day could be recentered and rescaled to a 50 mg zero point and 50 mg units, so 150 mg/day becomes $(150 - 50)/50=2$ units/day. The resulting estimated

coefficients then represent coefficients for a 10 mm increase in blood pressure, a 1 pack/day increase in smoking, and a 50 mg/day increase in vitamin C. The recordings used in the present example are shown under Table 4.

Multiple priors

For independent priors, each regressor with an explicit prior receives its own prior data in which other regressors are set to their reference values, along with a distinct prior indicator to identify those data. For example, to introduce priors for the coefficients $\beta_2, \beta_3, \beta_5$ of X_2, X_3, X_5 , we would add three prior indicators $\text{Prior}_2, \text{Prior}_3, \text{Prior}_5$ to the data; these would be zero except in the prior records for the corresponding coefficients. Thus, in the two prior records for $\beta_3, X_3=1/S_3$ or 0, $\text{Prior}_3=1, \text{Prior}_2=\text{Prior}_5=0$, and all other regressors would be set to their reference values. Note that priors need not be used for every coefficient (e.g. highly stable estimates least need priors); omitting some priors corresponds to semi-Bayes estimation.^{8,20}

Product-term (interaction) priors

Typical product terms represent between-group coefficient differences. For example, the coefficient β_{jk} of a product term X_jX_k represents the change in the coefficient of X_j corresponding to a unit change in X_k . Product-term priors are thus particularly valuable for addressing issues of artefacts that arise from subgroup analyses, and the extreme instability of subgroup estimates.

Recentering and rescaling of both regressors is however crucial to sensibly interpreting β_{jk} . If (say) blood pressure X_j were left as millimetre and smoking X_k were left as cigs/day, the unit to which β_{jk} refers would be millimetre \times cigs/day, a unit so small that β_{jk} would miniscule even in the presence of a large interaction. For example, suppose the combination of a 10 mm increase in pressure and a 1-pack/day (20 cigs./day) increase in smoking multiplied the odds by 8-fold more than the product of either increase alone. Then the value of β_{jk} without rescaling would be only $\ln(8)/10(20)=0.01$, despite the large interaction.

Upon sensible recentering and rescaling, β_{jk} can be given a prior table. With prior mode m_{jk} for β_{jk} , the table has risk ratio $\text{RR}_{\text{table}}=\exp(m_{jk}/S_{jk})$; the prior records have $X_jX_k=1/S_{jk}$ or 0, and all other regressors (including X_j and X_k) are at their reference values. Care should be taken, however, to set priors for the main and product terms ($\beta_j, \beta_k, \beta_{jk}$) that are contextually coherent, which may entail prior dependencies (see subsequently).

Intercept priors

The intercept β_0 is the log odds or log rate when all regressors are zero. Hence to make sense of β_0 , zero must be a meaningful value for all the regressors. β_0 can then be given a prior, but special action is required. With prior mode m_0 for β_0 and a scaling factor S_0 , the prior is represented by a table in which $\text{RR}_{\text{table}}=\exp(m_0/S_0)$, as in the earlier format. Nonetheless, the program must be told to leave out the constant term (using the 'no-intercept' or 'no-constant' option). In place of this constant, one adds a regressor X_0 that is 1 for all actual-data records, $1/S_0$ for the first β_0 -prior record, and 0 for all other

prior records (including the second β_0 -prior record); β_0 is then the coefficient of X_0 . One also adds an indicator Prior_0 that is 1 in the two intercept-prior records, 0 for all other records. Finally, all other regressors should be zero in the intercept-prior records. In Table 1, $\exp(\beta_0)$ is the odds of death for a neonate with none of the listed risk factors; hence β_0 was given a normal prior with $v_{\text{prior}} = 1$ and $m_{\text{prior}} = \ln(7/10\,000)$, although the remaining results are unchanged upon using no prior for β_0 .

The offset method

Because tabular augmentation requires adding a tabular indicator for each prior, it can expand the data set considerably; e.g. a prior for every coefficient doubles the number of regressors (from 15 to 30 in Table 4). One can obtain identical results by replacing all the prior indicators with a single *offset* variable, which is a regressor H whose coefficient γ is forced to equal 1. H depends on the model and the prior. Forcing $\gamma = 1$ can be done using software options for constrained estimation or for offsets, or by adding a prior record for γ . H can however be omitted if it is zero for all records (actual and prior).

Unconditional logistic regression

The prior-data table for X_j can be represented by a single record with A_{1j} cases and A_{0j} non-cases (for $A_{1j}+A_{0j}$ total), $X_j = 1/S_j$, and

$$H = \ln\left(\frac{A_{1j}}{A_{0j}}\right) - \frac{m_j}{S_j} = \ln\left(\frac{N_{1j}}{N_{0j}}\right),$$

where m_j is the prior mode of β_j . All other regressors in the prior record are set to zero. The resulting β_j prior is exactly of log-F (logit-beta) form with $2A_{1j}$, $2A_{0j}$ degrees of freedom (d.f.); it can be made accurately normal by setting $S_j = (v_{\text{prior}} \times A_j/2)^{1/2}$ where $A_{1j} = A_{0j} = A_j$ is very large (e.g. 400).³¹ Whether or not an intercept prior is present, the constant is replaced by a regressor X_0 that is 1 for actual-data records and 0 for prior-data records (except $X_0 = 1/S_0$ for the intercept-prior record, if present).

The Bayesian results in Table 4 are replicated by replacing the constant by X_0 and adding just one regressor H plus 15 records to represent the 15 prior tables.^{31,32} With $S_j = 10$, the prior record for X_j has $A_{1j} = 400$ cases out of $A_{1j} + A_{0j} = 800$ total, $X_j = 1/10$, $\ln(A_{1j}/A_{0j}) = 0$ and $H = 0$, $-\ln(2)/10$, or $-\ln(4)/10 = 0$, -0.0693 , or -0.1386 according to whether the prior mode for $\ln(\text{RR}_j)$ is 0, $\ln(2)$, or $\ln(4)$. For example, the prior record for the hydramnios coefficient β_7 is

Cases	Non-cases	Total	H	X_7
400	400	800	-0.1386	1/10 (all other regressors zero).

One can force $\gamma = 1$ by adding a prior-data record for γ that has $H = 1$, $10^6 e = 27\,182\,822$ cases, and 10^6 non-cases, with all other regressors (including the intercept) set to zero.³¹

Conditional-logistic and Cox regression

For these models, the prior table for β_j is translated into two records of matched case-non-case pairs.³⁶ One record represents A_{1j} pairs that have the case at $H = \ln(N_{1j}/N_{0j})$, $X_j = 1/S_j$ and the non-case at $H = 0$, $X_j = 0$. The other record represents A_{0j} pairs

that have the case at $H = 0$, $X_j = 0$ and the noncase at $H = \ln(N_{1j}/N_{0j})$, $X_j = 1/S_j$. Both prior records have all other regressors set to zero. One can force $\gamma = 1$ by adding a prior-data record that has $10^6 e = 27\,182\,822$ matched pairs with $H = 1$ for the case and 0 for the control, and another record with 10^6 matched pairs with $H = 0$ for the case and $H = 1$ for the non-case; all other regressors are set to zero.

In Cox (proportional-hazards) regression, the added matched pairs become matched risk sets with one failure (case) and one survivor (non-case) each. Each prior pair must be designated as a separate stratum (available as a ‘stratified’ or ‘matched’ survival option in most software). One must also supply values for the failure time of each matched set (failure time of the case = censoring time of the paired survivor). Because of the stratification, the time chosen will not affect the results, and so could be set to 1 for all prior pairs.³⁶

Poisson and log-linear count regression

Records for these models have only a single count. A prior for coefficient β_j can be induced by adding a record with count A_j , person-time = 1, $X_j = 1/S_j$, offset $H = \ln(A_j) - m_j/S_j$, and zero in all other entries;³² equivalently, one can dispense with H by setting the person-time to $e^H = A_j/\exp(m_j/S_j)$. The resulting β_j prior is exactly of log- χ^2 (log-gamma) form with $2A_j$ d.f., and can be made accurately normal by setting $S_j = (v_{\text{prior}} \times A_j)^{1/2}$ with A_j very large³¹ (to obtain a log-F prior in Poisson regression, one must use the prior-table method, with two records and an indicator). One can force $\gamma = 1$ by adding a prior-data record for γ that has $H = 1$, $10^6 e = 27\,182\,822$ cases, and 10^6 person-time units, with all other regressors (including the intercept) set to zero.³²

Further topics

Dependent priors and hierarchical regression

Suppose two coefficients β_3 and β_4 are dependent in our prior, in that gaining information about one would change our bets about the other. This would be true, for example, if Y were a cancer indicator and X_3 and X_4 were decades worked in two different jobs, both of which involved exposure to a chemical Z thought to increase risk of the cancer. It would also be true if X_3 and X_4 were serving-decades of consumption of two different vegetables, both of which contained a nutrient Z thought to reduce risk of the cancer.

One can build these dependencies from independent components via hierarchical (multilevel) modelling.^{5,8,20,37–39} Hierarchical regression can be done with commercial mixed-model software, or with data priors by rewriting the hierarchical model in a single-level (random-coefficient or mixed-model) form.^{5,40} Dependent priors can also be created directly by adding the prior data in the form of a multiway table in which the coefficient estimates have the desired modes, variances, and correlations, or by directly manipulating the prior design matrix.^{31,32}

Checking the model

A prior distribution can be viewed as a hierarchical extension of the initial regression model, and so itself a model.^{15,41}

Table 4 shows the importance of contrasting the prior distribution with the likelihood function in order to appreciate their relative contributions to the posterior results. In particular, it will be important to check the compatibility of the prior with the likelihood before combining. If diagnostics indicate severe incompatibility, it is advisable to expand the prior and likelihood (data) model or both to resolve the conflict.

As with conventional modelling, there are many philosophies and methods for such diagnosis and expansion.^{1,2,21,38} Among basic diagnostics are visual comparisons of prior and likelihood summaries or graphs, e.g. comparing the prior mode and limits to the ML estimate and confidence limits, as in Table 4. Another simple check is the P -value for adding the product term $\text{Prior}_j \times X_j$ to the model. A small P -value indicates incompatibility of the prior with the likelihood, which could arise from faulty prior information, faulty actual data, a faulty likelihood model, or some combination. A large P -value, however, does *not* mean the prior and likelihood are compatible, let alone correct; at best one can only say that the diagnostic detected no problem.

In either case, one should remember that there is rarely justification for the usual likelihood models found in observational epidemiology (such as binomial and Poisson models). In fact there are often good reasons to expect them to be wrong, such as validity problems.^{29,42} Many of the latter problems will be undetectable (non-identifiable) from available data, which is why appeal to the data or to conventional results as the ultimate arbiter of truth is logically indefensible.

Accuracy of posterior approximations

When the posterior distribution is non-normal, the Wald limits $\exp\{m_{\text{post}} \pm 1.96(v_{\text{post}})^{1/2}\}$ become inaccurate. If more accurate limits are desired, one can use posterior sampling, or apply profile-likelihood methods to the augmented data to produce profile-posterior limits.^{3,31,32} Normality of either the likelihood or the prior may however suffice to make the posterior normal enough to employ the Wald limits. The hydramnios coefficient in Table 4 provides an example. There is only 1 death with hydramnios, hence the profile-likelihood limits (2.8, 478) are very different from the Wald ML limits (5.7, 635). Nonetheless, due to the normal prior, the Wald posterior limits (1.6, 23) closely approximate the profile-posterior limits (1.6, 22) and Metropolis-sampling limits (1.5, 22). See Greenland³¹ for a full comparison of profile and posterior-sampling limits for Table 4, and Greenland³² for another example comparing Wald, profile, and posterior-sampling results using highly skewed priors.

Concluding remarks: benefits and hazards of priors

Use of external exposure information can greatly improve estimation and prediction accuracy.^{5,21,38} For example, as expected with so few actual cases, the priors in Table 4 strongly affect the final estimates of several coefficients; yet, the Bayesian estimates turn out to provide more accurate mortality predictions in subsequent years of observation.⁸ Of course, most analyses do not have the benefit of new data against which to validate, but numerous real examples going back over 30 years¹¹ have found similar benefits from Bayesian and related hierarchical and shrinkage techniques, as predicted by theory and simulations.^{5,15,43}

One cannot, however, expect improvement over conventional methods using just any prior; worsening could occur from using strongly misinformed priors (priors that assign relatively low probability to values near the truth). These possibilities argue for retaining conventional frequentist results to compare to Bayesian results, and for being generous in setting the spread of the prior. When the prior is highly influential, subject-matter evaluation will also be important, and technical complications may ensue if the prior or likelihood is far from normal.^{3,32} Translation of the prior into data aids evaluation by showing how large a perfect experiment would have to be to empirically justify prior assertions.²

As illustrated above and elsewhere,^{2,5,8,20,31,43–46} vaguely informative and roughly normal priors are easy to use in epidemiological problems, yet offer notable improvements over ML and stepwise results, without incurring severe risk of distortion entailed by the latter methods. ‘Vaguely informative’ should not however be construed as non-informative. Non-informative priors correspond to using values of A_1 and A_0 of zero, which give back the ML estimate and confidence interval as the posterior mode and interval. So-called reference priors⁴⁷ correspond to using very small values and so give similar results. Such ‘objective Bayesian’ methods barely address sparse-data problems and confer none of the predictive benefits obtainable from well-informed priors; they also rarely make sense on subject-matter grounds.^{2,48,49}

Acknowledgement

The author thanks Katherine Hoggatt and the referees for helpful comments.

Conflict of interest: None declared.

KEY MESSAGES

- Bayesian and related ‘shrinkage’ methods can be used to address certain common problems in epidemiological regression, such as artefacts arising from variable selection, sparse data, subgroup analyses, and multiple comparisons.
- Bayesian regressions can be computed using ordinary software by augmenting the actual data set with prior-data records.
- The data records for a prior can represent a prior-data table generated by a thought experiment, or can be simplified into a single record.

References

- ¹ Berry DA. Teaching elementary Bayesian statistics with real applications in science (with discussion). *The American Statistician* 1997;**51**:241–71.
- ² Greenland S. Bayesian perspectives for epidemiologic research. I. Foundations and basic methods (with discussion). *Int J Epidemiol* 2006;**35**:765–78.
- ³ Greenland S. Prior data for non-normal priors. *Stat Med* 2007;**26**. (in press).
- ⁴ Bedrick EJ, Christensen R, Johnson W. A new perspective on generalized linear models (1996). *J Am Stat Assoc* 1996;**91**:1450–60.
- ⁵ Greenland S. When should epidemiologic regressions use random coefficients? *Biometrics* 2000;**56**:915–21.
- ⁶ Greenland S. Small-sample bias and corrections for conditional maximum-likelihood odds-ratio estimators. *Biostatistics* 2000;**1**:113–22.
- ⁷ Greenland S, Schwartzbaum JA, Finkle WD. Problems from small samples and sparse data in conditional logistic regression analysis. *Am J Epidemiol* 2000;**151**:531–39.
- ⁸ Greenland S. Methods for epidemiologic analyses of multiple exposures: A review and a comparative study of maximum-likelihood, preliminary testing, and empirical-Bayes regression. *Stat Med* 1993;**12**:717–36.
- ⁹ Greenland S, Neutra RR. Control of confounding in the assessment of medical technology. *Int J Epidemiol* 1980;**9**:361–67.
- ¹⁰ Rothman KJ, Greenland S. *Modern Epidemiology*. Ch. 14, 2nd edn. Philadelphia: Lippincott-Raven, 1998, p. 402.
- ¹¹ Efron B, Morris CN. Data analysis using Stein's estimator and its generalizations. *J Am Stat Assoc* 1975;**70**:311–19.
- ¹² Leamer EE. *Specification Searches*. New York: Wiley, 1978.
- ¹³ Copas JB. Regression, prediction, and shrinkage. *J Royal Stat Soc Ser B* 1983;**45**:311–54.
- ¹⁴ Greenland S. Multilevel modeling and model averaging. *Scand J Work Environ Health* 1999;**25**(Suppl. 4):43–48.
- ¹⁵ Greenland S. Principles of multilevel modelling. *Int J Epidemiol* 2000;**29**:158–67.
- ¹⁶ Robins JM, Greenland S. The role of model selection in causal inference from nonexperimental data. *Am J Epidemiol* 1986;**123**:392–402.
- ¹⁷ Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. *Am J Epidemiol* 1989;**129**:125–27.
- ¹⁸ Maldonado G, Greenland S. A simulation study of confounder-selection strategies. *Am J Epidemiol* 1993;**138**:923–36.
- ¹⁹ Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol* 2006;**163**:1149–56.
- ²⁰ Greenland S. A semi-Bayes approach to the analysis of correlated associations, with an application to an occupational cancer-mortality study. *Statist Med* 1992;**11**:219–30.
- ²¹ Carlin B, Louis TA. *Bayes and Empirical-Bayes Methods of Data Analysis*. 2nd edn. New York: Chapman and Hall, 2000.
- ²² Greenland S. Reducing mean squared error in the analysis of stratified epidemiologic studies. *Biometrics* 1991;**47**:773–75.
- ²³ Titterton DM. Common structure of smoothing techniques in statistics. *Int Stat Rev* 1985;**53**:141–70.
- ²⁴ Greenland S. Smoothing observational data: a philosophy and implementation for the health sciences. *Int Stat Rev* 2006;**74**:31–46.
- ²⁵ Leamer EE. False models and post-data model construction. *J Am Stat Assoc* 1974;**69**:122–31.
- ²⁶ Graham P. Bayesian inference for a generalized population attributable fraction. *Stat Med*, 2000;**19**:937–56.
- ²⁷ Greenland S. The impact of prior distributions for uncontrolled confounding and response bias: A case study of the relation of wire codes and magnetic fields to childhood leukemia. *J Am Stat Assoc* 2003;**98**:47–54.
- ²⁸ Greenland S. Interval estimation by simulation as an alternative to and extension of confidence intervals. *Int J Epidemiol* 2004;**33**:1389–97.
- ²⁹ Greenland S. Multiple-bias modelling for analysis of observational data (with discussion). *J Roy Stat Soc A*, 2005;**168**:267–308.
- ³⁰ Neutra RR, Fienberg SE, Greenland S, Friedman EA. The effect of fetal monitoring on neonatal death rates. *New Engl J Med* 1978;**299**:324–26.
- ³¹ Greenland S. Putting background information about relative risks into conjugate priors. *Biometrics* 2001;**57**:663–70.
- ³² Greenland S. Generalized conjugate priors for Bayesian analysis of risk and survival regressions. *Biometrics* 2003;**59**:92–99.
- ³³ Lindley DV. The Bayesian analysis of contingency tables. *Ann Math Stat* 1964;**35**:1622–43.
- ³⁴ Jones MC. Families of distributions arising from distributions of order statistics. *Test* 2004;**13**:1–43.
- ³⁵ Landaw EM, Sampson PF, Toporek JD. Advanced nonlinear regression in BMDP. In *Proceedings of the Statistical Computing Section*, Washington, D.C.: American Statistical Association, 1982, pp. 228–33.
- ³⁶ Greenland S, Christensen R. Data augmentation for Bayesian and semi-Bayes analyses of conditional-logistic and proportional-hazards regression. *Stat Med* 2001;**20**:2421–28.
- ³⁷ Kass RE, Steffey D. Approximate Bayesian inference in conditionally independent hierarchical models. *J Am Stat Assoc* 1989;**84**:717–26.
- ³⁸ Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. 2nd edn. New York: Chapman and Hall/CRC, 2003.
- ³⁹ Goldstein H. *Multilevel Statistical Models*. 3rd edn. New York: Oxford, 2003.
- ⁴⁰ Witte JS, Greenland S, Kim LL, Arab LK. Multilevel modeling in epidemiology with GLIMMIX. *Epidemiology* 2000;**11**:684–88.
- ⁴¹ Good IJ. Hierarchical Bayesian and empirical Bayesian methods (letter). *The American Statistician* 1987;**41**:92.
- ⁴² Greenland S. Randomization, statistics, and causal inference. *Epidemiology* 1990;**1**:421–29.
- ⁴³ Gustafson P, Greenland S. The performance of random-coefficient regression in accounting for residual confounding. *Biometrics*, 2006;**62**:760–68.
- ⁴⁴ Witte JS, Greenland S, Haile RW, Bird CL. Hierarchical regression analysis applied to a study of multiple dietary exposures and breast cancer. *Epidemiology* 1994;**5**:612–21.
- ⁴⁵ Greenland S. Second-stage least squares versus penalized quasi-likelihood for fitting hierarchical models in epidemiologic analysis. *Stat Med* 1997;**16**:515–26.
- ⁴⁶ Aragaki CC, Greenland S, Probst-Hensch NM, Haile RW. Hierarchical modeling of gene-environment interactions: Estimating NAT2* genotype-specific dietary effects on adenomatous polyps. *Cancer Epidemiol Biomark Prev*, **6**, 307–14.
- ⁴⁷ Berger JO. The case for objective Bayesian analysis. *Bayesian Analysis*, 2006;**1**:385–472.
- ⁴⁸ Goldstein M. Subjective Bayesian analysis: principles and practice. *Bayesian Analysis* 2006;**1**:403–20.
- ⁴⁹ Greenland S. Probability logic and probabilistic induction. *Epidemiology* 1998;**9**:322–32.