The Stata Journal (2015)
15, Number 3, pp. 712–736

Approximate Bayesian logistic regression via penalized likelihood by data augmentation

Andrea Discacciati
Unit of Biostatistics and Unit of Nutritional Epidemiology
Institute of Environmental Medicine
Karolinska Institutet
Stockholm, Sweden
andrea.discacciati@ki.se

Nicola Orsini
Unit of Biostatistics and Unit of Nutritional Epidemiology
Institute of Environmental Medicine
Karolinska Institutet
Stockholm, Sweden
nicola.orsini@ki.se

Sander Greenland
Departments of Epidemiology and Statistics
University of California
Los Angeles, CA
lesdomes@ucla.edu

Abstract. We present a command, `penlogit`, for approximate Bayesian logistic regression using penalized likelihood estimation via data augmentation. This command automatically adds specific prior-data records to a dataset. These records are computed so that they generate a penalty function for the log likelihood of a logistic model, which equals (up to an additive constant) a set of independent log prior distributions on the model parameters. This command overcomes the necessity of relying on specialized software and statistical tools (such as Markov chain Monte Carlo) for fitting Bayesian models, and allows one to assess the information content of a prior in terms of the data that would be required to generate the prior as a likelihood function. The command produces data equivalent to normal and generalized log-F priors for the model parameters, providing flexible translation of background information into prior data, which allows calculation of approximate posterior medians and intervals from ordinary maximum likelihood programs. We illustrate the command through an example using data from an observational study of neonatal mortality.

Keywords: st0400, penlogit, penalized likelihood estimation, data augmentation, Bayesian methods, logistic models

 st0400

1 Introduction

Philosophical objections to Bayesian methods have lost much force over recent decades because examples of successful applications of these methods have grown. Nonetheless, Bayesian analyses remain uncommon in many disciplines. This slow adoption is unsurprising, given that Bayesian methods are rarely covered in basic courses and thus remain somewhat mysterious to many scientists. This coverage failure may in turn be attributed to a pervasive yet incorrect belief that Bayesian statistics requires computational formulas and software fundamentally different from familiar frequentist statistics such as p-values and confidence intervals.

Approximate Bayesian analyses, however, can be carried out easily using penalized likelihood (PL) estimation, which in turn can be implemented via data augmentation. The accuracy of these approximations are as good as or better than the accuracy of the corresponding frequency approximations that underpin maximum likelihood (ML) estimates and to date have given results similar to analyses based on posterior sampling (Greenland 2001, 2003; Cole et al. 2012; Sullivan and Greenland 2013; Cole, Chu, and Greenland 2014).

Data augmentation begins by translating prior distributions into prior-data records, an exercise that displays the information content of prior distributions in familiar terms of experimental results and sample size (Landaw, Sampson, and Toporek 1982; Bedrick, Christensen, and Johnson 1996; Higgins and Spiegelhalter 2002; Greenland 2006, 2007b, 2007a; Sullivan and Greenland 2013). Once this translation is made, Bayesian analyses can be carried out with any statistical software that implements standard likelihood methods. This approach runs faster than simulation methods like Markov chain Monte Carlo (MCMC) and does not introduce complex convergence criteria or simulation error.

In this article, we present a new command, `penlogit`, that fits penalized logistic regression via data augmentation and thus can be used to carry out approximate Bayesian logistic regression. The article is organized as follows: In section 2, we introduce PL estimation in the context of logistic regression and illustrate how it can be employed to carry out Bayesian analyses. In section 3, we describe the syntax and the options of the `penlogit` command. In section 4, we present a simulation study comparing the empirical performance of standard logistic regression and penalized logistic regression on sparse data. In section 5, we use data from an observational study on neonatal mortality to present some practical examples of Bayesian analyses using penalized logistic regression. In section 6, we conclude.

2 Methods and formulas

2.1 Penalized log likelihood

We will define a penalized log likelihood (PLL) as a log likelihood with a penalty function added to it. Suppose we have a sample of N binomial observations, each with y_i successes out of n_i trials, and a p-dimensional vector of covariates $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p})$ (including a constant, if any), $i = 1, \ldots, N$. In the case of ungrouped data, y_i is either equal to 1 or to 0, while $n_i \equiv 1$. Suppose we wish to fit a logistic regression model to these data,

$$\ln\left\{\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right\} \equiv \text{logit}\left\{\pi(\boldsymbol{x}_i)\right\} = \sum_{j=1}^{p} x_{ij}\beta_j$$

where $\pi(\boldsymbol{x}_i)$ denotes the proportion of successes in group i, given $\boldsymbol{x}_i$.

The corresponding PLL will then be

$$\text{PLL}(\boldsymbol{\beta};\boldsymbol{x}) = \underbrace{\sum_{i=1}^{N}\left[\ln\left\{\text{expit}\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right\}y_i + \ln\left\{1-\text{expit}\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right\}(n_i - y_i)\right]}_{\ln\{\mathrm{L}(\boldsymbol{\beta};\boldsymbol{x})\}} + P(\boldsymbol{\beta}) \qquad (1)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ indicates the vector of unknown regression parameters, $\ln\{\mathrm{L}(\cdot)\}$ is the log likelihood of a standard logistic regression model, $\text{expit}(x)$ is equivalent to $\exp(x)/\{1+\exp(x)\}$, and $P(\boldsymbol{\beta})$ is the penalty term (Le Cessie and van Houwelingen 1992). The purpose of the penalty is to pull or shrink the final parameter estimates away from the ML estimates, toward values $\boldsymbol{m} = (m_1, \ldots, m_p)$.

Ideally, the choice of these values should be guided by background information outside of the likelihood and should be good guesses for the parameters in $\boldsymbol{\beta}$, although a commonly used default value of 0 is often chosen for those parameters for which background information is limited or controversial. Zero is especially appropriate for coefficients of exposures in exploratory studies ("fishing expeditions"), whereas it would not be appropriate for coefficients of known outcome predictors (typically, at least age and sex) for which considerable background information is available. An advantage of penalized estimation is that the penalty can be restricted to those coefficients for which the value to shrink toward is easy to specify; then analysis becomes partial-Bayes or semi-Bayes (Cox 1975; Greenland 2000).

2.2 Bayesian perspective

From a Bayesian perspective, one can think of the penalty as arising from a prior distribution on the parameters. Specifically, a prior distribution for a model parameter is a probability distribution that incorporates a priori information—that is, information apart from the data being analyzed—that the data analyst has about a given parameter. Prior distributions that are spread out carry weak background information, whereas priors concentrated on a limited portion of the parameter space carry extensive background information. The two extreme cases are priors with $+\infty$ and 0 variance, respectively. In the former case, we have no background information at all and thus we rely only on the data for our analyses, whereas in the latter scenario, prior information is so strong that the data information about the parameter is ignored.

The latter scenario is in effect for every parameter that is omitted from the model without further checking whether it should be entered; for example, when product terms are not entered in the model, all coefficients of such terms are effectively assumed to be 0, which corresponds to using a normal prior with 0 mean and 0 variance. To incorporate less rigid background information into the parameter estimates, we instead enter the term in the model but specify a prior with a nonzero variance. We then add the logarithm of the prior density function as the penalty term in the log likelihood. The PLL is then (apart from an additive constant) equal to the logarithm of the posterior distribution of $\boldsymbol{\beta}$ given the data (Greenland 2001).

Bayesian analyses are sometimes criticized for their sensitivity to choice of prior. This sensitivity can, however, be exploited to show how sensitive or robust inferences may be to changing assumptions about background information (Bayesian sensitivity analysis) and to show weaknesses of the data in light of such information. In this view, it can be valuable to have some flexibility in the location, scale, and shape of the prior. We will thus consider two basic families of priors for logistic coefficients: the normal and the generalized log-F distributions.

Normal priors

Normal priors for β_j are symmetric and unimodal, and therefore the prior mean, mode, and median equal the same value m_j. Equivalently, they impose a log-normal distribution on $\exp(\beta_j)$, where $\exp(m_j)$ is the prior median odds ratio; however, $\exp(m_j)$ is neither the prior mode nor the prior mean odds ratio. The amount of background information carried by these priors is controlled by their variance (v_j): smaller values mean that the priors are more concentrated around m_j and therefore carry more background information.

The $100(1-\alpha)\%$ equal-tailed prior limits for the odds ratio—that is, that pair of numbers such that the data analyst would give $100(1-\alpha)\%$ probability that the true odds ratio is between these two numbers, ignoring the analysis data, with equal probability of falling above or below the interval—is $\exp(m_j \pm z_{1-\alpha/2}\text{se}_{\text{prior},j})$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of a standard normal distribution.

Suppose we specify independent normal priors for the first q model parameters (with $q \leq p$). Each of these priors is characterized by its prior mean m_j and its prior variance v_j, $j = 1, \ldots, q$. Letting $\widetilde{\boldsymbol{\beta}}$ denote the vector of these coefficients, the penalty function in (1) is defined as

$$P\left(\widetilde{\boldsymbol{\beta}}\right) = -\frac{1}{2}\left\{\sum_{j=1}^{q}\frac{\left(\beta_j - m_j\right)^2}{v_j}\right\} \tag{2}$$

We note that some literature defines the penalty function as -2 times the quantity subtracted from the log likelihood; in the normal case, this makes the penalty equal the sum of squares in (2), and more generally makes the penalty a quantity added to the deviance function (which is -2 times the log likelihood).

Generalized log-F priors

Generalized log-F priors subsume normal priors as a limiting case and provide a more flexible tool to translate background information about the model parameters $\boldsymbol{\beta}$ into prior distributions (Greenland 2003, 2007a). Log-F distributions are unimodal but, unlike normal priors, can be skewed if prior information is directional, favoring protective (left-skew) or harmful (right-skew) associations (assuming $Y = 1$ indicates an adverse event). Log-F priors are the natural conjugate-prior family for logistic regression.

These priors are characterized by four parameters: the prior mode m_j, the degrees of freedom $\mathrm{df}_{1,j}$ and $\mathrm{df}_{2,j}$, and the scale parameter s_j. The $\mathrm{df}_{1,j} = \mathrm{df}_{2,j} = \mathrm{df}_j$ case produces a symmetric log-F prior that approaches normality and whose variance decreases as df_j increases. If df_j is small, the tails of the prior become heavier than those of a normal prior. To skew the prior while keeping the mode at m_j, we increase the difference between $\mathrm{df}_{1,j}$ and $\mathrm{df}_{2,j}$. If $\mathrm{df}_{1,j} < \mathrm{df}_{2,j}$, the log-$F$ distribution becomes left-skew, whereas if $\mathrm{df}_{1,j} > \mathrm{df}_{2,j}$, the distribution becomes right-skew. Unless $\mathrm{df}_{1,j} = \mathrm{df}_{2,j} = \mathrm{df}_j$, the prior mode m_j is neither the mean nor the median of the prior. The scale parameter s_j allows expansion or contraction of the prior distribution around m_j without changing its shape. If $s_j > 1$, the distribution expands, while if $0 < s_j < 1$, the distribution contracts.

To evaluate the resulting prior, we calculate the exact $100(1-\alpha)\%$ prior limits on the odds-ratio scale. Let $f_{\alpha/2}$ and $f_{1-\alpha/2}$ be the $(\alpha/2)$ and $(1-\alpha/2)$ quantiles of an F distribution with $\mathrm{df}_{1,j}$ and $\mathrm{df}_{2,j}$ degrees of freedom. The $100(1-\alpha)\%$ prior limits for the odds ratio are calculated as

$$\left[\exp(m_j) f_{\frac{\alpha}{2}}^{s_j},\, \exp\left(m_j\right) f_{1-\frac{\alpha}{2}}^{s_j}\right] \tag{3}$$

All the other percentiles can be obtained in an analogous fashion; for example, the prior median (50th percentile) is equal to $\exp(m_j) f_{0.50}^{s_j}$.

Suppose we specify independent generalized log-F priors on the coefficients in the vector $\widetilde{\boldsymbol{\beta}}$ of the first q model parameters (with $q \leq p$). Each of these priors is characterized by a parameter vector $(m_j, \mathrm{df}_{1,j}, \mathrm{df}_{2,j}, s_j)$, $j = 1, \ldots, q$. The penalty function in (1) is defined as

$$P\left(\widetilde{\boldsymbol{\beta}}\right) = \sum_{j=1}^{q} \left[\frac{\mathrm{df}_{1,j}}{2}\left(\frac{\beta_j - m_j}{s_j} + \eta_j\right) - \frac{\mathrm{df}_{1,j} + \mathrm{df}_{2,j}}{2} \ln\left\{1 + \exp\left(\frac{\beta_j - m_j}{s_j} + \eta_j\right)\right\}\right]$$

where $\eta_j = \ln\left(\mathrm{df}_{1,j}/\mathrm{df}_{2,j}\right)$ (Greenland 2001, 2003, 2009; Brown, Spears, and Levy 2002; Jones 2004).

Specifying the priors

Specification of priors for the model coefficients is the major aspect of Bayesian analysis that differentiates it from a classical frequentist analysis. One way to specify a prior for the model coefficient β_j is starting from, say, 95% prior limits and then calculating the hyperparameters (that is, the prior's parameters) from there. Suppose that reasonable 95% prior limits for the odds ratio $\exp(\beta_j)$ are $(\omega_{Lj}, \omega_{Uj}) = [\exp(\beta_{Lj}), \exp(\beta_{Uj})]$ conditional on the remaining covariates. Under normality, it is easy to calculate the corresponding mean and variance for β_j by reversing the usual steps for interval estimation:

$$m_j = \frac{(\beta_{Lj} + \beta_{Uj})}{2} = \ln\left\{(\omega_{Lj} \times \omega_{Uj})\right\}^{\frac{1}{2}} \tag{4}$$

and

$$v_j = \left\{\frac{(\beta_{Uj} - \beta_{Lj})}{2 \times 1.96}\right\}^2 = \left\{\frac{\ln\left(\frac{\omega_{Lj}}{\omega_{Uj}}\right)}{2 \times 1.96}\right\}^2 \tag{5}$$

For generalized log-F priors, calculating hyperparameters starting from prior limits is less straightforward. However, one can always start by specifying a normal prior with reasonable 95% limits and calculating its hyperparameters using (4) and (5). Then, given that log-F priors subsume normal priors as a limiting case, one can impose the same normal prior employing a rescaled symmetric log-F distribution with a large number of degrees of freedom (see section 2.3). Lastly, by increasing the difference between $\mathrm{df}_{1,j}$ and $\mathrm{df}_{2,j}$, it is possible to obtain a prior distribution with the desired skewness that correctly reflects the asymmetric prior information for β_j. Equation (3) can be used to evaluate the resulting 95% prior limits on the odds ratio scale, and this exercise can be repeated for different α to understand the implications of the prior.

Posterior distribution

Apart from a multiplicative constant k, the $\mathrm{PL}(\boldsymbol{\beta}; \boldsymbol{x})$ equals the posterior density $f(\boldsymbol{\beta}|\boldsymbol{x})$,

$$\mathrm{PL}\left(\boldsymbol{\beta}; \boldsymbol{x}\right) \propto f\left(\boldsymbol{\beta}|\boldsymbol{x}\right) = k \times \mathrm{L}\left(\boldsymbol{\beta}; \boldsymbol{x}\right) \times \prod_{j=1}^{q} f_j\left(\beta_j\right)$$

where $f_j(\beta_j)$ is the prior density for β_j (Greenland 2001). Thus, the maximum penalized-likelihood (MPL) estimate of β_j ($\beta_{\mathrm{post},j}$) is the mode of the posterior distribution, also known as the maximum a posteriori estimate (Landaw, Sampson, and Toporek 1982). Furthermore, $\beta_{\mathrm{post},j}$ is the approximate posterior mean and median, while the estimated standard error is the approximate standard deviation of the posterior distribution ($\mathrm{se}_{\mathrm{post},j}$). The odds ratio estimate $\exp(\beta_{\mathrm{post},j})$ and its $100(1-\alpha)\%$ Wald confidence limits $\exp(\beta_{\mathrm{post},j} \pm z_{1-\alpha/2}\mathrm{se}_{\mathrm{post},j})$ are the approximate posterior median and $100(1-\alpha)\%$ posterior limits, that is, the $(\alpha/2)$ and $(1-\alpha/2)$ quantiles of the posterior distribution of $\exp(\beta_j)$. Ideally, the data analyst would give $100(1-\alpha)\%$ probability that the true odds ratio is between these numbers, after analyzing the data, assuming the prior used represents what the analyst would give before seeing the data and that the regression model represents the probabilities the analyst would assign to the data when the parameters are known. These posterior limits approach the usual frequentist confidence limits when all the prior variances are allowed to grow arbitrarily large (which represents negligible prior information).

If the posterior distribution of a given parameter is not approximately normal—or equivalently, if the penalized profile log likelihood is not very closely quadratic—Wald posterior limits may no longer be adequate. To obtain more accurate posterior limits, it is possible to use posterior sampling or penalized profile-likelihood posterior limits (Greenland 2003; Cole et al. 2012; Sullivan and Greenland 2013; Cole, Chu, and Greenland 2014).

For penalized profile-likelihood posterior limits, let $\boldsymbol{\beta}_{\text{post}}(\beta_j)$ be the vector of the restricted MPL estimates when component j of $\boldsymbol{\beta}$ is held fixed at β_j, and let $\text{PPL}(\beta_j;\boldsymbol{x}) \equiv \text{PL}(\boldsymbol{\beta}_{\text{post}}(\beta_j);\boldsymbol{x})$ be the corresponding restricted maximum. The penalized profile-likelihood posterior limits are found by solving

$$-2\ln\left\{\frac{\text{PPL}\left(\beta_j;\boldsymbol{x}\right)}{\text{PL}\left(\boldsymbol{\beta}_{\text{post}};\boldsymbol{x}\right)}\right\} = -2\left[\ln\left\{\text{PPL}\left(\beta_j;\boldsymbol{x}\right)\right\} - \ln\left\{\text{PL}\left(\boldsymbol{\beta}_{\text{post}};\boldsymbol{x}\right)\right\}\right] = S_\alpha$$

where S_α is the $(1-\alpha)$ quantile of a χ^2_1 distribution so that the probability coverage of the resulting limits approximates $100(1-\alpha)\%$. For example, if $100(1-\alpha)\% = 95\%$, then $S_\alpha = S_{0.05} = 3.84$.

Normality of either the likelihood or the prior may be sufficient to make the posterior distribution normal enough to use the Wald posterior limits (Greenland 2007b). However, when a skewed prior is used or when the data are sparse, the use of penalized profile-likelihood or posterior-sampling limits will usually be necessary (Greenland 2003, 2007a). See Greenland (2003) for an example comparing Wald, profile, and posterior-sampling limits when using highly skewed priors, and see Greenland (2007a) for a more detailed discussion on profile posterior checks.

The approximations used in PL estimation work well in the context of observational epidemiology. Approximation errors are, in fact, negligible when considering the uncertainties about the data-generation processes, and they are typically far below the magnitude of random errors and biases such as uncontrolled confounding, measurement error, and selection bias (Greenland 2001, 2007b). PL estimation is therefore a valuable alternative to posterior sampling such as MCMC, which requires specialized software and introduces complex convergence criteria. Moreover, PL estimation runs quicker than MCMC, thus simplifying Bayesian sensitivity analyses (Greenland 2006). Even if one prefers to sample from the posterior distribution, PL estimation can still provide good starting values and validity checks for the chosen sampler (Greenland 2007b; Sullivan and Greenland 2013).

2.3 Data augmentation

Instead of directly maximizing the PLL in (1), an equivalent way of carrying out approximate Bayesian logistic regression is to use data augmentation (Landaw, Sampson, and Toporek 1982; Bedrick, Christensen, and Johnson 1996; Greenland 2001, 2003; Greenland and Christensen 2001). With this procedure, one prior-data record is added to the actual dataset for each prior (plus one column of offset terms if necessary). The prior-data records will generate a penalty function that imposes the desired prior constraints on the model parameters. No specialized software is needed for Bayesian analysis using data augmentation; in fact, any statistical software that implements ML estimation will suffice. Moreover, data augmentation has the advantage of showing the strength of the priors being imposed on β_j in terms of number of cases $[(\mathrm{df}_{1,j})/(2 \times s_j^2)]$ and noncases $[(\mathrm{df}_{2,j})/(2 \times s_j^2)]$ that would supply data information about the coefficient approximately equivalent to the information supplied by the prior (Greenland 2006, 2007a).

With data augmentation, perfectly normal priors can be imposed employing symmetric log-F priors with a large number of degrees of freedom (say, $\mathrm{df}_{1,j} = \mathrm{df}_{2,j} = \mathrm{df}_j = 1800$). These priors are rescaled by the factor $s_j = \{(v_j \times \mathrm{df}_j)/4\}^{\frac{1}{2}}$, where v_j is the desired prior variance for the normal distribution. The scale factor s_j is then divided into all the regressor values in the prior data, including the offset, and the numbers of added cases and noncases are multiplied by s_j^2 to compensate for the rescaling (Greenland 2007a; Sullivan and Greenland 2013).

See Greenland (2006, 2007a) and Sullivan and Greenland (2013) for practical details on data augmentation. For more technical details, see Greenland (2001, 2003), Greenland and Christensen (2001), and references therein.

2.4 Frequentist—Bayesian parallels

Although in this paper we focus on PL from a Bayesian perspective, PL estimates can also be derived as frequentist "shrinkage" estimates when the penalty is viewed as a loss function for estimation errors. This dual interpretation illustrates how Bayesian and frequentist interpretations can be viewed as complementary, rather than conflicting (Greenland 2006, 2007b; Sullivan and Greenland 2013; Cole, Chu, and Greenland 2014). Use of normal priors, as illustrated here, corresponds to using a sum of squared error (quadratic) loss function, and is a useful tool for model expansion and for estimate stabilization when dealing with sparse data (Greenland 2001, 2007b; Sullivan and Greenland 2013).

Regression with a quadratic penalty (2), $\boldsymbol{m} = (0, \ldots, 0)$, and the v_j assumed to equal a constant v is also known as ridge regression, which can be used to allow partial entry of regressors into the model by shrinking the parameters toward 0 (Le Cessie and van Houwelingen 1992; Steyerberg 2008; Hastie, Tibshirani, and Friedman 2009); in this formulation, v is replaced by a tuning or ridge parameter λ equal to $1/v$ or $1/(2v)$. Ridge regression cannot set model parameters to 0, and thus it cannot exclude regressors from the model. However, it can be used as an alternative to conventional variable-selection methods, such as those based on significance levels (for example, p-value < 0.05 for inclusion) or changes in estimates (for example, at least 10% relative change upon inclusion), which can lead to distorted tests and estimates (Greenland 1989; Maldonado and Greenland 1993). The optimal value of the ridge parameter is usually estimated using cross-validation to minimize some measure of prediction error (for example, mean squared error or mean classification error) or using empirical Bayes (marginal ML) methods (Steyerberg 2008; Hastie, Tibshirani, and Friedman 2009; Efron 2012).

Bayesian shrinkage parallels empirical Bayes but prespecifies v based on contextual (external) information, with larger values representing greater prior uncertainty; the degree of the shrinkage is then controlled directly by the data analyst (Greenland 2007b).

Sparse data arise when there are only a few or no subjects at certain covariate patterns, or when the number of regressors approaches the number of cases or noncases. Sparse-data bias can happen not only in small samples but also in large samples, as the simulation in section 4 and the example in section 5 will show. In these situations, conventional frequentist estimates often result in inflated odds-ratio estimates and excessively wide confidence intervals, even if no other bias is present. Weakly informative priors can be used to stabilize estimates that suffer from sparse-data artifacts. Use of priors or penalties also allows the inclusion of more confounders in the model, which can potentially reduce the bias in effect estimates (Greenland 2008). A similar frequentist approach to sparse data is Firth's method, which shrinks estimates toward 0, which for logistic regression involves maximization of the PLL in (1), where $P(\boldsymbol{\beta}) = 1/2 \ln|I(\boldsymbol{\beta})|$ and $I(\boldsymbol{\beta})$ is the Fisher information matrix (Firth 1993; Heinze and Schemper 2002).

Although we do not discuss them here, other penalties can be useful. For example, the lasso penalty takes the sum of absolute error as a loss function and corresponds to using Laplace (double-exponential) priors. The result can be quite different from quadratic penalization, especially in that more unstable coefficients may be shrunk all the way to 0 and thus eliminated from the model. The lasso is thus valuable when the goal is to reduce the number of variables in a predictive model rather than to simultaneously estimate all the original coefficients (Hastie, Tibshirani, and Friedman 2009).

3 The penlogit command

3.1 Description

`penlogit` provides estimates for the penalized logistic model, whose PLL is defined in (1), using data augmentation priors.

3.2 Syntax

`penlogit` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] [`,`
  `npr`*ior*`(`*varname m v* [*varname m v* ...]`)`
  `lfp`*rior*`(`*varname m* df_1 df_2 *s* [*varname m* df_1 df_2 *s* ...]`)` `ppl(`*varlist*`)`
  `nppl(`*#*`)` `bin`*omial*`(`*varname*`)` `l`*evel*`(`*#*`)` or `nol`*ist* `nocon`*stant*]

`by`, `statsby`, and `xi` are allowed; see [U] **11.1.10 Prefix commands**. `fweight`s are allowed; see [U] **11.1.6 weight**. After `penlogit` estimation, it is possible to use postestimation commands like `test`, `testparm`, `lincom`, `predict`, and `predictnl`; see [R] **test**, [R] **lincom**, [R] **predict**, and [R] **predictnl**.

By default, no priors are imposed on the model coefficients. If no priors are imposed by the user (that is, if neither option `nprior()` nor `lfprior()` is used), `penlogit` reproduces the results obtained by `logit` (see [R] **logit**).

3.3 Options

`nprior(`*varname m v* [*varname m v* ...]`)` imposes a normal prior with mean = mode = median = m and variance v on the desired model parameter (log odds-ratio).

`lfprior(`*varname m* df_1 df_2 *s* [*varname m* df_1 df_2 *s* ...]`)` imposes a generalized log-F prior with mode m, degrees of freedom df_1 and df_2, and scale factor s on the desired model parameter (log odds-ratio).

`ppl(`*varlist*`)` specifies the variables for which penalized profile-likelihood limits are required. It calls an adapted version of the `pllf` command (Royston 2007).

`nppl(`#`)` evaluates penalized profile-likelihood at # equally spaced points. The default is `nppl(100)`.

`binomial(`*varname*`)` specifies the variable containing the binomial denominator when the data are grouped (that is, when *depvar* contains the total number of successes or failures).

`level(`#`)`; see [R] **estimation options**.

`or` displays the exponentiated coefficients (odds ratios) and corresponding standard errors and confidence intervals.

`nolist` suppresses the summary of prior distributions in terms of exact prior percentiles (50th, 2.5th, and 97.5th) and data approximately equivalent to priors.

`noconstant` suppresses the constant term.

3.4 Stored results

`penlogit` stores the following in `e()`:

Scalars

`e(N)`	number of observations
`e(N_da)`	number of observations including the augmented data
`e(k)`	number of parameters
`e(ic)`	number of iterations
`e(converged)`	`1` if converged, `0` otherwise
`e(pll)`	PLL

Macros

`e(cmd)`	`penlogit`
`e(cmdline)`	command as typed
`e(depvar)`	name of dependent variable
`e(indepvars)`	names of independent variables
`e(wtype)`	weight type
`e(wexp)`	weight expression
`e(properties)`	`b V`
`e(predict)`	program used to implement `predict`

Matrices

`e(b)`	coefficient vector
`e(V)`	variance–covariance matrix of the estimators
`e(ppl)`	penalized profile-likelihood limits
`e(ilog)`	iteration log (up to 20 iterations)
`e(nprior)`	m and v of the normal priors
`e(lfprior)`	m, df_1, df_2, and s of the log-F priors

Functions

`e(sample)`	marks estimation sample

4 Simulation

In this section, we present a simulation study comparing the empirical performance of standard logistic regression and penalized logistic regression on sparse data.

We generated 1,000 samples from a standard logistic model in each of four different simulation scenarios arising from the combination of two data-generating mechanisms and two sample sizes. In all simulation scenarios, we generated 10 independent and identically distributed binary covariates x_i such that $x_i \sim \text{Bernoulli}(0.5)$, $i = 1, \ldots, 10$. The $\exp(\beta)$ associated with each of these 10 covariates was set to 4 and 10 for the first and second data-generating mechanisms, respectively. The two sample sizes were $n = 500$ and $n = 5000$. Binary outcomes y_j were sampled from a Bernoulli distribution with parameter $p_j = \text{expit}(\beta_0 + \sum_{i=1}^{10} \beta_i x_{ij})$, $j = 1, \ldots, n$. The intercept coefficient β_0 varied across the four simulation scenarios and was calculated to obtain an expected outcome $E(Y)$ (which is the marginal probability of $Y = 1$) equal to 0.05 for the simulation scenarios with $n = 500$ and equal to 0.005 for the simulation scenarios with $n = 5000$, where

$$E(Y) = \sum_{x_1=0}^{1} \dots \sum_{x_{10}=0}^{1} \left\{ E\left(Y|X_1 = x_1, \dots, X_{10} = x_{10}\right) \times \prod_{i=1}^{10} \Pr\left(X_i = x_i\right) \right\}$$
$$= \sum_{k=0}^{10} \left\{ \text{expit}\left(\beta_0 + k\beta\right) \times \binom{k}{10} 0.5^{10} \right\}$$

The four values of β_0 were -11.6 [$\beta = \ln(4), n = 500$], -14.37 [$\beta = \ln(4), n = 5000$], -18.21 [$\beta = \ln(10), n = 500$], and -21.84 [$\beta = \ln(10), n = 5000$].

The following code produces 1 of the 1,000 samples used in the first simulation scenario [$\beta = \ln(4), n = 500$]. It can be easily adapted to the other simulation scenarios and might prove useful to the reader who wants to replicate this simulation study using the Stata command `simulate`.

```
. clear
. set obs 500
number of observations (_N) was 0, now 500
. local intercept = -11.6
. local beta = ln(4)
. local xbeta ""
. foreach i of numlist 1/10 {
  2.         generate x`i´ = rbinomial(1, 0.5)
  3.         local xbeta "`xbeta´ + `beta´ * x`i´"
  4. }
. generate xb = `intercept´ `xbeta´
. generate y = rbinomial(1, invlogit(xb))
```

We analyzed the simulated data using both standard logistic regression and penalized logistic regression. First, we imposed weakly informative normal priors with mean 0 and variance 4 on each of the 10 coefficients ($\beta_1, \dots, \beta_{10}$). These priors have an exact prior median odds ratio of 1 and 95% exact prior limits of $[0.02, 50]$. Then, we imposed weakly informative normal priors with mean ln(2) and variance 1, so that the exact prior 50th, 2.5th, and 97.5th percentiles on the odds-ratio scale were 2, 0.28, and 14.21, respectively. Each of the variance-4 priors supplied data information roughly equivalent to 0.9 cases and 0.9 noncases, while each of the variance-1 priors supplied data information roughly equivalent to 2.5 cases and 2.5 noncases.

Table 1 shows the simulated 50th, 5th, and 95th percentiles of the MPL estimate of $\exp(\beta_1)$ under each scenario, for standard logistic regression (ML) and penalized logistic regression (PL). Results for the remaining nine coefficients are similar and therefore not displayed.

Table 1. Median (5th, 95th percentiles) of the simulated distribution of the MPL estimate of $\exp(\beta_1)$

Sample size	Method	Prior on β_1	Odds Ratio	
			4	10
500	ML	—	4.6 (1.9, 15)	15 (4.3, 115)[a]
	PL	Normal(0, 4)	3.7 (1.8, 9.3)	6.9 (3.1, 16)
	PL	Normal(ln(2), 1)	3.3 (1.8, 6.6)	4.8 (2.7, 8.7)
5,000	ML	—	4.2 (1.8, 12)[b]	12 (4.0, 58)[c]
	PL	Normal(0, 4)	3.8 (1.8, 9.1)	7.6 (3.2, 21)
	PL	Normal(ln(2), 1)	3.6 (1.8, 7.0)	5.7 (3.0, 11)

Each scenario was simulated 1,000 times.
[a] Excluded 6 simulations because of convergence not achieved.
[b] Excluded 8 simulations because of convergence not achieved.
[c] Excluded 45 simulations because of convergence not achieved.

In all four scenarios, standard logistic regression suffered from sparse-data bias, which produced a higher proportion of simulations with extreme values of $\exp(\beta_1)$ compared with penalized logistic regression. For example, in the scenario with $n = 500$ and $\exp(\beta) = 10$, standard logistic regression did not converge in 6 out of 1,000 simulations, and in the remaining 994 simulations, 5% of the estimates of $\exp(\beta_1)$ were larger than the absurd value of 115 (25% were larger than 28). On the other hand, penalized logistic regression always converged and resulted in less extreme estimates. The prior distributions used in this simulation study did not reflect any particular a priori information; still they were useful devices for providing stable inference and estimation in the presence of sparse data, by reducing sparse-data bias.

5 Examples

Greenland (2007a, 2007b) and Sullivan and Greenland (2013) used data from a study on neonatal mortality during the first full year of electronic fetal monitoring at a teaching hospital (Neutra et al. 1978) to illustrate how to conduct approximate Bayesian analysis via data augmentation. We used the same data to illustrate the `penlogit` command.

5.1 Univariate analysis

Table 2 shows the cross-tabulation of the data based on the exposure ($X = 1$, no monitoring; $X = 0$, monitoring) and the outcome ($Y = 1$, death; $Y = 0$, survival). Given that fetal monitoring was developed to rapidly detect fetal distress during labor, babies whose mothers were in the "no monitoring" group were expected to have higher odds of dying during the neonatal period (0 to 28 days) (odds ratio above 1). However, at the time of the study, the magnitude of the association was unclear.

Table 2. Cohort data on fetal monitoring and neonatal death (Neutra et al. 1978)

	Fetal monitoring[a] $X = 1$	$X = 0$	Total
Deaths ($Y = 1$)	14	3	17
Survivals ($Y = 0$)	2,284	691	2,975
Total	2,298	694	2,992

[a] "No monitoring" is coded as $X = 1$.

We fit a standard logistic regression model using the command `penlogit` by not specifying any prior on the model parameters. The response variable was the number of deaths (`deaths`), and the binary indicator for the monitoring status was the only covariate (`nomonit`). We also specified the following options: `binomial(n)` to indicate that the data are grouped, and `ppl(nomonit)` to obtain profile-likelihood confidence intervals for the covariate `nomonit`.

```
. clear
. input nomonit deaths n
       nomonit     deaths          n
  1. 0 3 694
  2. 1 14 2298
  3. end
. penlogit deaths nomonit, binomial(n) ppl(nomonit)
Logistic regression                                  No. of obs   =          2
Log likelihood = -3.7351204

------------------------------------------------------------------------------
      deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     nomonit |   .3449013   .6376887     0.54   0.589    -.9049456    1.594748
       _cons |  -5.439528   .5786022    -9.40   0.000    -6.573567   -4.305488
------------------------------------------------------------------------------

------------------------------------
      deaths | [95% PL Conf. Interval]
-------------+----------------------
     nomonit |   -.7781632    1.814143
------------------------------------
```

The ML estimate for the odds ratio was $\exp(0.345) = 1.41$, while the 95% Wald confidence limits were $\exp(0.345 \pm 1.96 \times 0.638) = [0.40, 4.92]$. Given the sparseness of the data (only three deaths among the unexposed mothers), profile-likelihood confidence limits should provide more accurate coverage, because they do not depend on the normality of the likelihood function. In this example, 95% profile-likelihood confidence intervals for the odds ratio were $[\exp(-0.778), \exp(1.814)] = [0.46, 6.13]$, indicating an asymmetric profile-likelihood.

Normal priors

Suppose that a positive but not strong association was expected. We translated this background information into a normal prior for the log odds-ratio ($\beta_{\texttt{nomonit}}$) or, equivalently, into a lognormal prior for the odds ratio [$\exp(\beta_{\texttt{nomonit}})$], such that the 95% prior limits on the odds-ratio scale were between 0.5 and 8. These limits were obtained by setting $m = \ln(2) = 0.693$ and $v = 0.50$ [see (4) and (5)]; the prior median odds ratio was thus $\exp(m) = 2$ and its 95% prior limits were $\exp\{\ln(2) \pm 1.96\sqrt{0.50}\} = [0.50, 8.00]$. With the `penlogit` command, this prior was imposed by specifying the option `nprior(nomonit ln(2) 0.5)`; we also specified the option `or` to get the results directly on the odds-ratio scale.

```
. penlogit deaths nomonit, binomial(n) ppl(nomonit) nprior(nomonit ln(2) 0.5) or
Penalized logistic regression                          No. of obs  =          2
Normal prior for nomonit: exact prior median OR (95% PL): 2.00 (0.50, 8.00)
Data approx. equivalent to prior: cases=4.54  noncases=4.54  exp(offset)=.955
Penalized log likelihood = -7.7746667
```

deaths	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
nomonit	1.657991	.8080274	1.04	0.300	.6378911	4.309412
_cons	.0037967	.0018163	-11.65	0.000	.0014866	.0096966

deaths	[95% PL Conf.	Interval]
nomonit	.6703909	4.562727

The approximate posterior median and 95% Wald posterior limits for the odds ratio were, respectively, 1.66 and $[0.64, 4.30]$, while the 95% penalized profile-likelihood posterior limits were $[0.67, 4.56]$. In this case, the profile posterior limits and the Wald were quite similar, indicating that the addition of the normal prior made the penalized profile-likelihood almost symmetric. Given the data and this specific prior information on the association between monitoring and neonatal death, we would give 95% probability that the true odds ratio is between 0.67 and 4.56.

Generalized log-F priors

Because in this example prior information was directional, pointing toward positive associations between no monitoring and neonatal death, an asymmetric prior better reflects the available background information. To illustrate, we impose an asymmetric log-F prior on the parameter $\beta_{\texttt{nomonit}}$ with a similar lower bound for the 95% prior limits as in the previous example but with no contextually meaningful upper bound. We set the prior mode as in the previous example [$m = \ln(2)$] and skewed the distribution to the right by setting $\text{df}_1 = 2000$ and $\text{df}_2 = 2$. We set the scale parameter s equal to 1. The 2.5th and 97.5th percentiles of an F distribution with 2,000 and 2 degrees of freedom are 0.271 and 39.497, respectively; thus, by using (3), an exact 95% prior interval for the odds ratio was $2(0.271, 39.497) = [0.54, 78.99]$. This prior is asymmetric and far more spread out than the normal prior of the previous example.

With the `penlogit` command, we specified the option `lfprior(nomonit ln(2) 2000 2 1)` to impose this prior.

```
. penlogit deaths nomonit, binomial(n) ppl(nomonit)
> lfprior(nomonit ln(2) 2000 2 1) or
Penalized logistic regression                        No. of obs   =          2
Log-F prior for nomonit: exact prior median OR (95% PL): 2.88 (0.54, 78.99)
Data approx. equivalent to prior: cases=1000.00 noncases=1.00  exp(offset)=500
Penalized log likelihood = -4.7798817

      deaths   Odds Ratio   Std. Err.       z     P>|z|      [95% Conf. Interval]

     nomonit     1.579411    .8384043     0.86    0.389      .5580191     4.47035
       _cons     .0039551    .0020068   -10.90    0.000       .001463    .0106921


      deaths   [95% PL Conf. Interval]

     nomonit     .6287842    5.313271
```

The approximate median of the posterior distribution for $\exp(\beta_{\texttt{nomonit}})$ was 1.58. In this example, given the sparseness of the data and the asymmetry of the prior, the resulting posterior distribution is not normal enough to trust Wald posterior limits, and penalized profile-likelihood limits are preferable. The 95% Wald and penalized profile-likelihood posterior limits were [0.56, 4.47] and [0.63, 5.31], respectively.

Suppose now that we wanted to contract the log-F prior without changing its shape, keeping the prior mode at ln(2). To do this, we set the scale parameter s to 0.5. The exact 95% prior limits on the odds ratio implied by a log-F distribution on $\beta_{\texttt{nomonit}}$ with mode $m = \ln(2)$, $\text{df}_1 = 2000$, $\text{df}_2 = 2$, and $s = 0.5$ are $2(\sqrt{0.271}, \sqrt{39.497}) = [1.04, 12.57]$. This prior is much less spread out than before rescaling and therefore is much more informative.

```
. penlogit deaths nomonit, binomial(n) ppl(nomonit)
> lfprior(nomonit ln(2) 2000 2 0.5) or
Penalized logistic regression                        No. of obs   =          2
Log-F prior for nomonit: exact prior median OR (95% PL): 2.40 (1.04, 12.57)
Data approx. equivalent to prior: cases=4000.00 noncases=4.00  exp(offset)=250
Penalized log likelihood = -4.8267735

      deaths   Odds Ratio   Std. Err.       z     P>|z|      [95% Conf. Interval]

     nomonit     1.779013    .6651933     1.54    0.123      .8548803    3.702143
       _cons      .003576    .0014357   -14.03    0.000       .001628    .0078549


      deaths   [95% PL Conf. Interval]

     nomonit     .9666144    4.391295
```

The stronger prior information implied by the rescaled log-F prior resulted in narrower 95% posterior limits compared with the unrescaled log-F prior. Given the sparseness of the data and the asymmetric prior distribution, we ignored the Wald posterior limits. Accepting the penalized profile-likelihood posterior limits, we would assign roughly 95% probability that the true odds ratio was between 0.97 and 4.39, given the data. The approximate posterior median was 1.78.

The results from the four univariate analyses are summarized in table 3.

Table 3. Results from approximate Bayesian analyses of the data in table 2

Prior on $\beta_{\texttt{nomonit}}$	Exact prior percentiles			Approximate posterior percentiles[a]		
	50th	2.5th	97.5th	50th	2.5th	97.5th
Normal$(0, +\infty)$[b]	—	—	—	1.41	0.40	4.92
				1.41	0.46	6.13
Normal(ln(2), 0.5)	2.00	0.50	8.00	1.66	0.64	4.30
				1.66	0.67	4.56
log-F(ln(2), 2000, 2, 1)	2.88	0.54	78.99	1.58	0.59	4.47
				1.58	0.63	5.31
log-F(ln(2), 2000, 2, 0.5)	2.40	1.04	12.57	1.78	0.85	3.70
				1.78	0.97	4.39

[a] For each prior, the 2.5th and 97.5th percentiles in the first row are Wald limits, while those in the second row are penalized profile-likelihood limits.
[b] No prior (results from table 2 alone; "percentiles" are the ML estimates and approximate confidence limits).

To illustrate the equivalence between PL estimation and data augmentation, we directly maximized the PLL of the previous example by using the `mlexp` command. The divisor 2 at the end of the penalty term is needed because Stata applies the penalty to each record in the dataset.

```
. quietly mlexp (ln(invlogit({b0}+{nomonit}*nomonit))*deaths
> + ln(1-(invlogit({b0}+{nomonit}*nomonit)))*(n-deaths) + (1000*(({nomonit}-
> ln(2))/0.5+ln(2000/2)) - 1001*ln(1+exp(({nomonit}-ln(2))/0.5+ln(2000/2))))/2)
. lincom [nomonit]_cons, or
 ( 1)  [nomonit]_cons = 0

------------------------------------------------------------------------------
             | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.779012   .6651927     1.54   0.123       .85488    3.702139
------------------------------------------------------------------------------
```

5.2 Multivariable analysis

The full dataset included 14 covariates. All were binary indicators with the exception of early age (0 = 20+ years, 1 = 15–19 years, 2 = under 15 years), gestational age (0 = 38+ weeks, 1 = 36–38 weeks, 2 = 33–35 weeks; under 33 weeks excluded), isoimmunization (0 = no, 1 = Rh, 2 = ABO), labor progress (0 = normal, 0.33 = prolonged, 0.67 = protracted, 1 = arrested), and past abortion (0 = none, 1 = 1, 2 = 2+) (see table 4).

Table 4. Priors for the 14 regressors included in the penalized logistic regression

Covariate[a]	Variable name	Prior	Exact prior percentiles		
			50th	2.5th	97.5th
No monitor	nomonit	Normal(ln(2), 0.5)	2.00	0.50	8.00
Early age	teenages	Normal(ln(2), 0.5)	2.00	0.50	8.00
Gestational age	gestage	Normal(ln(4), 0.5)	4.00	1.00	16.00
Past abortion	abort	Normal(0, 0.5)	1.00	0.25	4.00
Labor progress	dyslab	Normal(ln(2), 0.5)	2.00	0.50	8.00
Public ward	ward	Normal(ln(2), 0.5)	2.00	0.50	8.00
Malpresented	malpres	Normal(ln(4), 0.5)	4.00	1.00	16.00
Nonwhite	nonwhite	Normal(ln(2), 0.5)	2.00	0.50	8.00
Nulliparity	nullip	Normal(ln(2), 0.5)	2.00	0.50	8.00
Isoimmunization	isoimm	Normal(ln(2), 0.5)	2.00	0.50	8.00
Hydramnios	hydram	Normal(ln(4), 0.5)	4.00	1.00	16.00
PCA	placord	Normal(ln(2), 0.5)	2.00	0.50	8.00
Twin, triplet	twint	Normal(ln(4), 0.5)	4.00	1.00	16.00
PROM	prerupt	Normal(ln(2), 0.5)	2.00	0.50	8.00

PCA = placental/cord abnormality; PROM = prolonged rupture of membranes (30+ hours).

[a] Variables are binary indicators except early age (0 = 20+ years, 1 = 15–19 years, 2 = under 15 years), gestational age (0 = 38+ weeks, 1 = 36–38 weeks, 2 = 33–35 weeks; under 33 weeks excluded), isoimmunization (0 = no, 1 = Rh, 2 = ABO), labor progress (0 = normal, 0.33 = prolonged, 0.67 = protracted, 1 = arrested), and past abortion (0 = none, 1 = 1, 2 = 2+).

We fit a logistic model with all 14 variables and no priors.

```
. use http://www.imm.ki.se/biostatistics/data/neutra1978.dta, clear
(Neutra et al. (1978), Effect of fetal monitoring on neonatal death rates., NEJM)

. penlogit death nomonit teenages gestage abort dyslab ward malpres
> nonwhite nullip isoimm hydram placord twint prerupt, ppl(hydram) or

Logistic regression                               No. of obs   =       2992

Log likelihood = -81.929411
```

death	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
nomonit	1.248125	.8700669	0.32	0.751	.3183365	4.893617
teenages	1.609509	1.171998	0.65	0.513	.386254	6.706774
gestage	4.890897	1.74671	4.44	0.000	2.428816	9.848778
abort	.7202864	.5081269	-0.47	0.642	.1807274	2.87069
dyslab	.4997072	.5246818	-0.66	0.509	.0638223	3.912541
ward	.8642985	.5278453	-0.24	0.811	.2611062	2.860951
malpres	3.894239	2.944569	1.80	0.072	.8847073	17.14137
nonwhite	1.88514	1.190201	1.00	0.315	.5469274	6.49767
nullip	1.548766	.8840647	0.77	0.443	.505946	4.740974
isoimm	3.044235	1.869858	1.81	0.070	.9133672	10.14638
hydram	60.25478	72.38386	3.41	0.001	5.72066	634.6539
placord	3.101652	3.529087	0.99	0.320	.3334942	28.84681
twint	8.20637	6.338866	2.73	0.006	1.805741	37.29467
prerupt	.5407285	.6036719	-0.55	0.582	.0606309	4.822417
_cons	.000995	.0008698	-7.91	0.000	.0001793	.0055204

death	[95% PL Conf.	Interval]
hydram	2.792485	478.1916

Although the model fit successfully converged, some of the estimates were inflated because of data sparsity. For example, the binary indicator of hydramnios during pregnancy (`hydram`) had an ML estimate for the odds ratio of 60—one order of magnitude above clinical expectation—a consequence of only one death among nine hydramnios pregnancies.

Stepwise regression (`stepwise` command with options `pr(0.10)` and `pe(0.05)`) selected only `gestage`, `hydram`, and `twint` from the original 14 variables, but it did not bring the estimate for the hydramnios coefficient to a plausible value (odds ratio = 46.5). Moreover, stepwise regression—like other variable-selection algorithms—completely ignores background information and does not address the problem of confounding (omitted variables might confound the estimates of the selected variables). Firth's (1993) method (user-written command `firthlogit`; Coveney [2008]) did not solve the sparse-data problem either, with an estimated odds ratio for the hydramnios parameter (95% confidence limits) equal to 68.2 ([9.2, 505.3]). Thus, in this example, neither stepwise regression nor Firth's method gave satisfactory results.

We addressed the sparse-data problem by deriving penalty functions from priors. In our example (Greenland 2001), the 14 model parameters were given three possible normal priors, reflecting the background clinical information on the different risk factors.

Prior information on the risk factors was expressed in terms of 95% prior limits. In particular, 95% prior limits on the odds ratio scale were [0.25, 4], [0.5, 8], and [1, 16] for those factors identified as "uncertain", "probably positive", and "probably strong", respectively. Hyperparameters of the prior distributions were then calculated using (4) and (5), yielding the following priors: Normal(0, 0.5), Normal(ln(2), 0.5), and Normal(ln(4), 0.5) (see table 4). No prior was placed on the intercept. We reduced to 50 the points at which the penalized profile-likelihood is evaluated, using the `nppl(50)` option.

```
. penlogit death nomonit teenages gestage abort dyslab ward malpres nonwhite
> nullip isoimm hydram placord twint prerupt, nprior(nomonit ln(2) 0.5 teenages
> ln(2) 0.5 gestage ln(4) 0.5 abort 0 0.5 dyslab ln(2) 0.5 ward ln(2) 0.5
> malpres ln(4) 0.5 nonwhite ln(2) 0.5 nullip ln(2) 0.5 isoimm ln(2) 0.5
> placord ln(2) 0.5 twint ln(4) 0.5 hydram ln(4) 0.5 prerupt ln(2) 0.5)
> ppl(nomonit teenages gestage abort dyslab ward malpres nonwhite nullip
> isoimm hydram placord twint prerupt) nppl(50) or
Penalized logistic regression                        No. of obs   =       2992
```

(*output omitted*)

```
Penalized log likelihood = -141.1233
```

death	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
nomonit	1.730433	.8569543	1.11	0.268	.6555687	4.567635
teenages	1.620486	.7765477	1.01	0.314	.633496	4.145212
gestage	4.520217	1.344752	5.07	0.000	2.523069	8.098217
abort	.8317827	.3907586	-0.39	0.695	.3312292	2.088773
dyslab	1.223829	.652187	0.38	0.705	.4306356	3.478019
ward	1.272606	.5546753	0.55	0.580	.5416152	2.990177
malpres	3.853277	1.925586	2.70	0.007	1.446978	10.26121
nonwhite	1.764528	.7961023	1.26	0.208	.7287721	4.272334
nullip	1.548364	.6589724	1.03	0.304	.6723691	3.565646
isoimm	2.412273	1.159756	1.83	0.067	.9401376	6.189583
hydram	6.067147	4.13142	2.65	0.008	1.5972	23.04675
placord	2.256392	1.384533	1.33	0.185	.6778182	7.511313
twint	5.237714	2.749351	3.15	0.002	1.872121	14.65378
prerupt	1.216663	.6493625	0.37	0.713	.4274287	3.463197
_cons	.0007097	.0004794	-10.73	0.000	.0001889	.002667

death	[95% PL Conf.	Interval]
nomonit	.6827299	4.795116
teenages	.6086605	4.012307
gestage	2.516992	8.145448
abort	.3052927	1.930884
dyslab	.4125252	3.345546
ward	.5342302	2.972514
malpres	1.391036	9.91937
nonwhite	.7124589	4.212519
nullip	.6739125	3.608426
isoimm	.8498436	5.672315
hydram	1.573366	22.51026
placord	.6508624	7.170668
twint	1.797856	14.12272
prerupt	.40767	3.315775

The approximate posterior median and 95% penalized profile-likelihood limits for the hydramnios parameter on the odds ratio scale were 6.06 and $[1.57, 22.52]$, respectively. Despite the rather weak prior imposed on the hydramnios parameter, the PL estimates were far more reasonable than the ML estimates.

Figure 1 shows the normal prior for $\beta_{\texttt{hydram}}$ from table 4 (long-dashed line), the approximate profile posterior density for $\beta_{\texttt{hydram}}$ (solid line), and the profile-likelihood function for $\beta_{\texttt{hydram}}$ (rescaled to have area 1 under the curve) (short-dashed line). The dot on the short-dashed line indicates the ML estimate, while the square on the solid line indicates the maximum a posteriori. This figure exhibits the skewness of the profile likelihood due to the sparseness of the data. The reason the approximate posterior distribution is closer to the prior is that the prior contained almost three times the information in the likelihood (approximate prior information of 2 versus approximate likelihood information of 0.7 from the actual data). Moreover, the posterior distribution became almost perfectly symmetrical because of the symmetrizing effect of the normal prior.

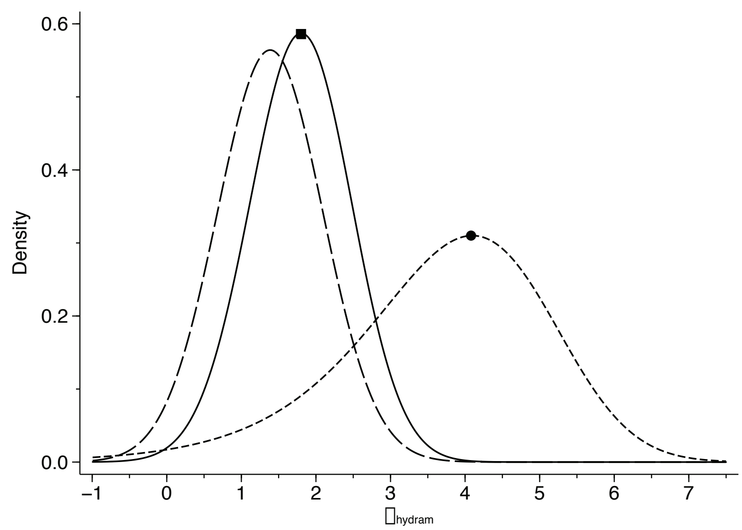

Figure 1. Normal prior for $\beta_{\texttt{hydram}}$ from table 4 (long-dashed line), approximate profile posterior density for $\beta_{\texttt{hydram}}$ (solid line), and profile-likelihood function for $\beta_{\texttt{hydram}}$ (rescaled to have area 1 under the curve) (short-dashed line)

Posterior percentiles from penalized logistic regression via data augmentation and from MCMC (Sullivan and Greenland 2013, 2014) showed exceptionally good agreement, considering the approximation error in data augmentation and the simulation error in MCMC (see table 5).

Table 5. Approximate posterior percentiles from penalized logistic regression via data augmentation and from MCMC

Covariate[a]	Variable name	Approximate posterior percentiles Data augmentation[b]			MCMC[c]		
		50th	2.5th	97.5th	50th	2.5th	97.5th
No monitor	`nomonit`	1.7	0.68	4.8	1.8	0.71	5.0
Early age	`teenages`	1.6	0.61	4.0	1.6	0.59	4.0
Gestational age	`gestage`	4.5	2.5	8.1	4.6	2.5	8.3
Past abortion	`abort`	0.83	0.31	1.9	0.79	0.29	1.9
Labor progress	`dyslab`	1.2	0.41	3.3	1.2	0.40	3.3
Public ward	`ward`	1.3	0.53	3.0	1.3	0.53	3.0
Malpresented	`malpres`	3.9	1.4	9.9	3.8	1.4	10
Nonwhite	`nonwhite`	1.8	0.71	4.2	1.8	0.70	4.2
Nulliparity	`nullip`	1.5	0.67	3.6	1.6	0.67	3.6
Isoimmunization	`isoimm`	2.4	0.85	5.7	2.3	0.81	5.6
Hydramnios	`hydram`	6.1	1.6	23	6.0	1.6	22
PCA	`placord`	2.3	0.65	7.2	2.2	0.64	7.1
Twin, triplet	`twint`	5.2	1.8	14	5.3	1.8	14
PROM	`prerupt`	1.2	0.41	3.3	1.2	0.39	3.3

PCA = placental/cord abnormality; PROM = prolonged rupture of membranes (30+ hours).

[a] Variables are binary indicators except early age (`0` = 20+ years, `1` = 15–19 years, `2` = under 15 years), gestational age (`0` = 38+ weeks, `1` = 36–38 weeks, `2` = 33–35 weeks; under 33 weeks excluded), isoimmunization (`0` = no, `1` = Rh, `2` = ABO), labor progress (`0` = normal, `0.33` = prolonged, `0.67` = protracted, `1` = arrested), and past abortion (`0` = none, `1` = 1, `2` = 2+).

[b] The 2.5th and 97.5th percentiles are from penalized profile-likelihood.

[c] MCMC analysis was carried out using the **genmod** procedure in SAS 9.2. A non-informative normal prior with mean 0 and variance 1,000,000 was placed on the intercept. Number of MCMC samples was set to 100,000.

6 Conclusion

We presented a new command, **penlogit**, that fits penalized logistic regression via data augmentation. We focused on how PL can be used to carry out approximate Bayesian analyses by applying a penalty term to impose the desired prior distributions on the model parameters. Using data from an epidemiological study, we illustrated how background information on different risk factors for neonatal mortality can be translated into prior distributions and how to interpret the results. We also showed how the Bayesian approach can be useful to deal with the frequentist sparse-data problem, which neither stepwise regression nor Firth's method were able to address satisfactorily in our example.

There are several advantages of carrying out approximate Bayesian analyses using PL estimation via data augmentation with the `penlogit` command. First, data augmentation uses ML estimation and so does not require the use of specialized software or unfamiliar commands. Second, unlike MCMC, PL estimation does not introduce complex convergence criteria of the Markov chains to the posterior distribution, which is a condition difficult to verify with absolute assurance. Third, it runs much faster than MCMC and thus simplifies Bayesian sensitivity analyses. For these reasons, even if one wants to use MCMC to sample from the posterior distribution, `penlogit` can provide reasonable starting values and convergence checks for the MCMC, and can also be used for sensitivity analyses.

In epidemiologic regression examples to date, penalized profile-likelihood limits have produced posterior summaries almost indistinguishable from those derived by posterior simulation (Greenland 2001, 2003; Cole et al. 2012; Cole, Chu, and Greenland 2014); that is unsurprising, given that typical PLLs from generalized linear models are smooth, unimodal, and concave downward. PL estimation does have some limitations, however. Because it is based on relative heights of the posterior density, it is unsuitable for posterior distributions that are multimodal or have otherwise complex shapes; in those cases, posterior sampling will be necessary to visualize and summarize the distribution. More generally, and unlike MCMC, PL estimation uses the same type of asymptotic approximations as does ordinary ML, although for normal and symmetric log-F priors, it converges more rapidly to the desired behavior because of the stabilizing and symmetrizing effect of the penalty function (Sullivan and Greenland 2013).

Future developments include the creation of a set of commands to carry out PL estimation via data augmentation for conditional logistic, log-linear (Poisson), and Cox regression models as described in Greenland (2007b) and Sullivan and Greenland (2013).

7 References

Bedrick, E. J., R. Christensen, and W. Johnson. 1996. A new perspective on priors for generalized linear models. *Journal of the American Statistical Association* 91: 1450–1460.

Brown, B. W., F. M. Spears, and L. B. Levy. 2002. The log F: A distribution for all seasons. *Computational Statistics* 17: 47–58.

Cole, S. R., H. Chu, and S. Greenland. 2014. Maximum likelihood, profile likelihood, and penalized likelihood: A primer. *American Journal of Epidemiology* 179: 252–260.

Cole, S. R., H. Chu, S. Greenland, G. Hamra, and D. B. Richardson. 2012. Bayesian posterior distributions without Markov chains. *American Journal of Epidemiology* 175: 368–375.

Coveney, J. 2008. firthlogit: Stata module to calculate bias reduction in logistic regression. Statistical Software Components S456948, Department of Economics, Boston College. http://econpapers.repec.org/software/bocbocode/s456948.htm.

Cox, D. R. 1975. Partial likelihood. *Biometrika* 62: 269–276.

Efron, B. 2012. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction.* New York, NY: Cambridge University Press.

Firth, D. 1993. Bias reduction of maximum likelihood estimates. *Biometrika* 80: 27–38.

Greenland, S. 1989. Modeling and variable selection in epidemiologic analysis. *American Journal of Public Health* 79: 340–349.

———. 2000. When should epidemiologic regressions use random coefficients? *Biometrics* 56: 915–921.

———. 2001. Putting background information about relative risks into conjugate prior distributions. *Biometrics* 57: 663–670.

———. 2003. Generalized conjugate priors for Bayesian analysis of risk and survival regressions. *Biometrics* 59: 92–99.

———. 2006. Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *International Journal of Epidemiology* 35: 765–775.

———. 2007a. Prior data for non-normal priors. *Statistics in Medicine* 26: 3578–3590.

———. 2007b. Bayesian perspectives for epidemiological research: II. Regression analysis. *International Journal of Epidemiology* 36: 195–202.

———. 2008. Invited commentary: Variable selection versus shrinkage in the control of multiple confounders (with discussion). *American Journal of Epidemiology* 167: 523–531.

———. 2009. Relaxation penalties and priors for plausible modeling of nonidentified bias sources. *Statistical Science* 24: 195–210.

Greenland, S., and R. Christensen. 2001. Data augmentation priors for Bayesian and semi-Bayes analyses of conditional-logistic and proportional-hazards regression. *Statistics in Medicine* 20: 2421–2428.

Hastie, T., R. J. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. New York: Springer.

Heinze, G., and M. Schemper. 2002. A solution to the problem of separation in logistic regression. *Statistics in Medicine* 21: 2409–2419.

Higgins, J. P. T., and D. J. Spiegelhalter. 2002. Being sceptical about meta-analyses: A Bayesian perspective on magnesium trials in myocardial infarction. *International Journal of Epidemiology* 31: 96–104.

Jones, M. C. 2004. Families of distributions arising from distributions of order statistics. *TEST* 13: 1–43.

Landaw, E. W., P. F. Sampson, and J. D. Toporek. 1982. Advanced nonlinear regression in BMDP. In *Proceedings of the Statistical Computing Section*, 228–233. Washington, DC: American Statistical Association.

Le Cessie, S., and J. C. van Houwelingen. 1992. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society, Series C* 41: 191–201.

Maldonado, G., and S. Greenland. 1993. Simulation study of confounder-selection strategies. *American Journal of Epidemiology* 138: 923–936.

Neutra, R. R., S. E. Fienberg, S. Greenland, and E. A. Friedman. 1978. Effect of fetal monitoring on neonatal death rates. *New England Journal of Medicine* 299: 324–326.

Royston, P. 2007. Profile likelihood for estimation and confidence intervals. *Stata Journal* 7: 376–387.

Steyerberg, E. W. 2008. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York: Springer.

Sullivan, S. G., and S. Greenland. 2013. Bayesian regression in SAS software. *International Journal of Epidemiology* 42: 308–317. With Erratum, *International Journal of Epidemiology* 43: 1667–1668.

———. 2014. Re: Sullivan S. G., Greenland S. Bayesian regression in SAS software. Int J Epidemiol 2013; 42: 308–17. *International Journal of Epidemiology* 43: 974.

About the authors

Andrea Discacciati is a PhD student at the Institute of Environmental Medicine at Karolinska Institutet in Stockholm, Sweden.

Nicola Orsini is an associate professor of medical statistics at the Institute of Environmental Medicine at Karolinska Institutet in Stockholm, Sweden.

Sander Greenland is a professor emeritus of epidemiology at the UCLA School of Public Health and a professor of statistics at the UCLA College of Letters and Science in Los Angeles, CA.