Type of manuscript: Original Article

Full title: The Firth bias correction, penalization, and weakly informative priors:
A case for log-*F* priors in logistic and related regressions

Short title: Bias correction and weak priors in logistic regression

Authors' full names and affiliations:
Sander Greenland[1], Mohammad Ali Mansournia[2*]

1 Department of Epidemiology, Fielding School of Public Health, and Department of Statistics, College of Letters and Science, University of California, Los Angeles, CA, USA

2 Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran

Corresponding author's name and mailing address, telephone and fax numbers, and e-mail address:
Mohammad Ali Mansournia
Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, PO Box: 14155-6446, Tehran, Iran. Tel: +98-21-88989123; Fax: +98-21-88989127; Email: mansournia_ma@yahoo.com

Word count, abstract: 126
Word count, main text: 4638
No. Tables: 1

The Firth bias correction, penalization, and weakly informative priors: A case for log-*F* priors in logistic and related regressions

Abstract. Penalization is a very general method encompassing the Firth bias correction as a special case. This correction has been programmed in major software packages, albeit with small differences among implementations and the results they provide. We consider some questions that arise when considering alternative penalties, and note some serious interpretation problems for the Firth penalty arising from the underlying Jeffreys prior, as well as contextual objections to alternative priors based on *t* distributions. Taking simplicity of implementation and interpretation as our chief criteria, we propose that the log-*F*(1,1) prior provides a better default penalty than other proposals. Penalization based on more general log-*F* priors is trivial to implement and facilitates sensitivity analyses of penalty strength the number of added observations (prior degrees of freedom) are varied.

Keywords: Bias correction, Firth correction, Jeffreys prior, logistic regression, maximum likelihood, penalized likelihood, sparse data

1. INTRODUCTION

A very useful method of dealing with sparse data and separation in logistic, Poisson, and Cox regression is the Firth bias correction [1,2]. Thanks to the work of Heinze and colleagues, this method for has been adopted into major software packages including SAS [3-5], R [6-9], and Stata [10]. Although the correction is often described as penalized likelihood estimation, penalization is a general method encompassing the Firth correction as a special case. This view has led us to a number of conclusions regarding the appropriateness of the correction and its competitors, which we present here.

We begin the present paper by describing the Firth correction in the simplest case, where its relation to classical bias corrections and simple prior distributions is transparent. We then consider proposals for default and "weakly informative" priors based on independent normal, Cauchy, and log-*F* distributions, and illustrate them in a clinically well-understood example. That example shows how the correlations in the Jeffreys prior underlying the Firth penalty can lead to artifacts such as estimates lying outside the range of the prior median and the maximum-likelihood estimate (MLE). We argue that, for transparency, computational simplicity, and reasonableness for logistic regression, a log-*F*(1,1) prior may provide a better reference point than the Jeffreys prior or those based on *t* distributions. Regardless of the prior shape chosen, however, we advise that a properly shifted prior or (more conveniently) no prior be used for intercepts or coefficients that could reasonably have extreme values, and that stronger penalties be used to control error in multiple inference, goals that cannot be accomplished with current implementations of the Firth correction. Finally, we describe a small discrepancy among implementations and computation of standard errors for the Firth correction.

2. THE FIRTH PENALTY AND THE JEFFREYS PRIOR IN LOGISTIC REGRESSION

Consider a logistic regression model $\pi(\boldsymbol{x}) = e^{\boldsymbol{x}'\boldsymbol{\beta}}/(1+ e^{\boldsymbol{x}'\boldsymbol{\beta}})$ for the dependence of a Bernoulli outcome parameter π on a covariate vector $\boldsymbol{x}$; $\boldsymbol{x}$ may include a constant, in which case the coefficient vector $\boldsymbol{\beta}$ includes an intercept. Let $\mathbf{X}$ be a design matrix with typical row $\boldsymbol{x}$, $\boldsymbol{y}$ the corresponding vector of observed binomial counts y with denominators n in $\boldsymbol{n}$, and let $\ell(\boldsymbol{\beta})$ denote the model loglikelihood $\boldsymbol{y}'(\mathbf{X}\boldsymbol{\beta}) - \boldsymbol{n}'\ln(1+ \exp(\mathbf{X}\boldsymbol{\beta}))$. The observed (and expected) information matrix $I(\boldsymbol{\beta})$ is then $-\ell''(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{W}\mathbf{X}$, the negative Hessian of $\ell(\boldsymbol{\beta})$, where $\mathbf{W}$ is diagonal with diagonal elements $ne^{\boldsymbol{x}'\boldsymbol{\beta}}/(1+ e^{\boldsymbol{x}'\boldsymbol{\beta}})^2$. The Firth correction [1] estimates $\boldsymbol{\beta}$ as the maximum of the penalized loglikelihood

$$\ell^*(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta})+ \tfrac{1}{2}\ln|I(\boldsymbol{\beta})|$$

and the penalized information $I^*(\boldsymbol{\beta})$ is the negative Hessian $-\ell^{*\prime\prime}(\boldsymbol{\beta})$. We will omit the arguments $\boldsymbol{x}$ and $\boldsymbol{\beta}$ from subsequent notation.

The penalty term $\tfrac{1}{2}\ln|I|$ is the log of a Jeffreys prior density [1, sec. 3.1], and thus the maximizer $\boldsymbol{b}$ of ℓ^* is the posterior mode given this prior. There is a large theoretical literature on this prior; we note only points relevant here. In logistic regression, this prior is proper and unimodal symmetric about $\boldsymbol{\beta} = \mathbf{0}$, with heavier tails than multivariate normal and lighter tails than multivariate t-distributions [11]. The prior is extremely weak and often described as a noninformative or reference prior for "objective Bayes" analyses [12]. Its frequentist rationale however is that it removes $O(n^{-1})$ bias from $\boldsymbol{b}$, and it prevents infinite estimates arising from separation problems [1-5]. Nonetheless, as Firth notes [1, sec. 7], it does not minimize expected error or loss. It thus may be unsurprising that, as we will illustrate below, in applications to sparse data it can produce implausible estimates when compared to possibly stronger but still contextually weak penalties, such as those based on generalized-conjugate (including log-F) priors [13-15], normal priors [15,16], or information criteria [17-20].

To gauge the information in the Firth penalty, consider a matched-pair case-control study observing n pairs discordant on an indicator x in which y pairs have x=1 for the case. Following standard theory [21], the conditional logistic regression for data can be analyzed as a single binomial observation of y successes on n trials with success probability $e^{\beta}/(1+e^{\beta})$, an intercept-only logistic model; the odds ratio of interest thus equals the odds e^{β} that in a given discordant pair the case is exposed. The resulting loglikelihood is $\ell = y\beta - n\ln(1+e^{\beta})$, which has $I = ne^{\beta}/(1+e^{\beta})^2$ and is maximized at $b = \ln(y/(n-y))$, the MLE, which is $e^b = 4$ when y=8, n=10. This MLE is the posterior mode under an improper log-F(0,0) (uniform) prior for β [14].

The Firth penalty for the pair model is $\ln|ne^{\beta}/(1+e^{\beta})^2|^{\frac{1}{2}} = \ln(n)/2 + \beta/2 - \ln(1+e^{\beta})$, corresponding (up to a constant) to the log of a Jeffreys prior density $e^{\beta/2}/(1+e^{\beta})$, which is a log-$F$(1,1) (Haldane) density for β. This prior produces a 95% prior interval for the odds ratio e^{β} of (1/648,648), which extends orders of magnitude beyond effects normally seen in health and social-science studies that require logistic-regression analyses. The penalized loglikelihood is $\ell^* = (y+\tfrac{1}{2})\beta-(n+1)\ln(1+e^{\beta})$, which equals the unpenalized loglikelihood obtained by augmenting y

by ½ and n by 1. The penalized information is $I^* = (n+1)e^{\beta}/(1+e^{\beta})^2$ and the maximizer b of ℓ^* is $b = \ln((y+½)/(n-y+½))$, the same estimate obtained from the classical Haldane bias correction for the empirical logit [1, p. 31]; this produces $e^b = 3.4$ when y=8, n=10.

More generally, if $\mathbf{X}$ is a matrix of mutually exclusive (orthogonal) indicators, I is diagonal with diagonal elements $ne^{\beta}/(1+e^{\beta})^2$ where β is an element of $\boldsymbol{\beta}$. Consequently, $|I|^{½}$ and thus the Jeffreys density becomes a product of log-F(1,1) densities. As illustrated below, however, the Jeffreys prior and independence priors can give quite divergent results, since most $\mathbf{X}$ induce correlations in the Jeffreys prior.

3. HOW DISPERSED SHOULD A REFERENCE PENALTY BE?

In cases in which there may be doubt about the appropriate degree of penalization, the Firth correction can provide a reference point arguably more appropriate than the MLE itself, given its reduced bias and the fact that the resulting estimates remain finite even with complete separation [1-5]. This reference interpretation view is mirrored by its derivation from the Jeffreys prior [1,12], which is invariant under reparameterization. As discussed below, however, it turns out that this invariance comes at a high cost of interpretability, which has inspired a search for better default priors.

There have been many other proposals for weak priors or penalties, including several expressly derived to deal with data sparsity. A review would far surpass our present scope, since most require software beyond basic logistic regression. The strongest prior we have seen proposed as "weakly informative" is a normal(0,1.38) prior, where the 1.38 is derived from a 95% prior interval for e^{β} of (1/10,10) [22]; as illustrated below, however, this prior is too strong to be considered weakly informative in general. A common approach is thus to use a normal prior with a large variance, which must face the arbitrariness of variance choice; also, a central-limit rationale for the normal shape seems inoperative when little or no background information is used to specify the prior. Other common approaches include t-distributions with few degrees of freedom and expanded scale, which have heavier tails than Jeffreys or log-F(1,1) priors [11]; these approaches face arbitrariness in choice of degrees of freedom and scale, however.

In an article that has attracted much attention, Gelman et al. [23] proposed Cauchy (t_1 = t-distribution with one degree of freedom) priors scaled up by a factor of 2.5 as defaults for logistic coefficients (other than the intercept), and described how to add this $2.5t_1$ distribution as a penalty in standard likelihood-maximization algorithms. Unfortunately, this prior has several disadvantages relative to the Firth correction: it has no frequentist justification as a bias correction, and in fact is rather arbitrary; although available in R [23], it is not yet (to our knowledge) available as a command in major commercial software; and, unlike normal or log-*F* priors [13-15], *t* priors cannot be implemented in ordinary maximum-likelihood packages by appending simple pseudo-data.

For comparison, Gelman et al. [23, fig. 1] also consider the prior density $e^{\beta/2}/(1+e^{\beta})$ for β in an intercept-only logistic model, which again is the log-*F*(1,1) and Jeffreys density, and argue in favor of the $2.5t_1$ distribution because it provides much higher probabilities of extreme values. There are however both Bayesian and frequentist reasons that this increased dispersion argues *against* the $2.5t_1$ distribution, in favor of lighter-tailed but still highly dispersed distributions.

Consider first effects known to be huge, e.g., with odds ratios above 100 or below 1/100. Such effects would ordinarily not be good candidates for study or control by logistic regression, precisely because they would likely lead to very small or zero event counts in some categories and consequent failure of asymptotics, as well as severe misspecification bias. Instead, tight restriction or matching would be advisable to control such effects adequately; and if such effects were modeled, a zero-centered prior would be inappropriate.

Now suppose instead that huge odds ratios were so implausible that estimates in these extreme ranges would be taken as signaling large errors or biases rather than real effects, a situation which typifies most covariates when viewed in context. Then at least some shrinkage would be advisable as a precautionary measure. Degree of shrinkage is determined by the prior dispersion. In this regard, a log-*F*(1,1) prior (again, the Jeffreys prior in the intercept-only case) already allows substantial probability for huge odds ratios, producing 12.7% or about 1 chance in 8 of e^{β} being below 1/100 or above 100 ($|\beta|>4.6$), again with a 95% prior interval of (1/648,648). The $2.5t_1$ prior increases the probability of $e^{\beta}<1/100$ or $e^{\beta}>100$ to 32%, or about 1 chance in 3, with a 95% prior interval that extends into the trillions. We would expect that the consequent undershrinkage of such huge or infinite MLEs by the $2.5t_1$ penalty would leave those estimates further away from zero compared to estimates from the Firth or log- *F*(1,1) penalties, as we have observed in several examples with infinite MLEs (not shown). In the less extreme example below, however, the Firth estimate is the largest, exceeding the MLE.

Gelman et al. do point out a serious practical defect of the multivariate Jeffreys prior and hence the Firth penalty: It is not clear in general how the Jeffreys prior translates into prior probabilities for odds ratios; and, unlike with independent priors, under a Jeffreys prior the marginal prior for a given β can change in opaque ways as model covariates are added or deleted. Addition of higher-order terms (products, powers, etc.) increases this problem since those terms can be highly correlated with one another and their main effects. Finally, by extension of the hierarchy principle, it is advisable to use stronger penalties for higher-order terms than for main effects [18]; implementing such a directive is difficult starting from the Firth penalty, yet is simple for other methods.

As Gelman et al. recognize, shrinking a logistic intercept toward zero is usually inadvisable, since the intercept is rarely expected to be in a neighborhood of zero. This issue can be addressed by using an intercept prior much more dispersed than other priors. Gelman et al. propose using a $10t_1$ intercept prior, which is negligibly different from no prior at all. In data augmentation, the intercept prior can be removed entirely by omitting the prior record for the intercept [15]. A corresponding modification could be made to the Firth penalty, but (unlike the t and log-F cases) would alter the prior distribution for remaining coefficients, and such an option is not available in packaged software. Another drawback of the Firth penalty is that it has not been implemented in all major packages (e.g., it is not listed in SPSS at this time), whereas normal and log-F priors extend easily to all packages via data augmentation [14,15,24].

Turning to log-F priors, the choice of degrees of freedom and scale may seem arbitrary, but a log-$F(1,1)$ prior for β without rescaling has a natural interpretation as adding exactly 1 null observation regarding β; as discussed below, this interpretation generalizes easily. The interpretational and computational simplicity of the log-$F(1,1)$ prior compared to other proposals, as well as the above considerations, lead us to prefer the log-$F(1,1)$ distribution as a suitable default penalty source (omitting the penalty for intercepts). They also lead us to recommend extensions of log-F penalties for settings in which more contextual information is available, as discussed next.

4. HOW STRONG SHOULD A WEAK PENALTY BE?

Extremely weak penalties such as those described above can be said to sacrifice precision in the neighborhood of zero in exchange for limiting bias where unbiasedness is contextually unimportant, in regions implausibly far from zero. When this tradeoff seems unacceptable, more precise penalties are easily implemented in any logistic-regression package by translating each desired coefficient penalty into a prior-data record [13-16]. To illustrate, let β be an element of $\boldsymbol{\beta}$, with x the corresponding element of $\boldsymbol{x}$. Because a log-$F(m,m)$ density is proportional to $e^{\beta m/2}/(1+e^{\beta})^{m}$, penalization of β by a log-$F(m,m)$ prior can be done by adding a data record with $m/2$ successes on m trials, and zero for all covariates (including the constant) except x, which is set to 1 [14,15,24]; it thus corresponds to adding a null binomial observation of size m as the prior for β. The log-F distribution has lighter tails than a t-distribution but heavier tails than the normal [25]. The prior degrees of freedom m in a log-F prior is exactly the number of observations added by the prior, while the corresponding penalty component $m\beta/2 - m\ln(1+e^{\beta})$ adds information $m/4$; the total added observations is the total of the m across coefficients, and may be compared to the number of actual observations to gauge the relative strength of the prior.

Shrinkage increases rapidly with m, both from decreased dispersion and from lightening of tails toward the normal, with the prior becoming practically normal$(0,4/(m-1))$ for $m>10$ [15]. To choose m based on desired prior intervals for e^{β}, one can refer to percentiles from a table or function for F distributions. Asymmetric penalties can be created by assigning unequal fractions f of m to the degrees of freedom to create log-$F(fm,(1-f)m)$ priors, and location and scale parameters can be added to produce shrinkage toward nonzero values and to produce normal priors [13,14,26 appendix], but we will focus on symmetric log-F shrinkage of β toward zero.

The penalty strength m can be varied across model parameters as deemed appropriate, and the penalty (prior record) can be omitted for any coefficient (partial penalization), which corresponds to using m=0. To minimize bias from shrinkage toward zero, we omit penalties for coefficients of known strong predictors, intercepts, and other terms which are assuredly far from zero. For less clearly nonzero coefficients, however, priors from larger values of m may be considered weakly informative. For example, a log-$F(2,2)$ prior for β equals the logistic prior, and corresponds to a uniform prior on $\pi = e^{\beta}/(1+e^{\beta})$, which has a long history as a weakly informative prior for a binomial parameter. It yields a 95% prior interval of (1/39, 39) for the odds ratio e^{β}, still nearly an order of magnitude greater than typical target effects; in the matched-pair example it produces $b = \ln((y+1)/(n-y+1))$, with $e^{b} = 3$ when y=8, n=10.

Strong penalties sacrifice calibration in contextually extreme regions of the parameter space in return for greater accuracy in likely regions. Typical multiple-inference problems ("fishing expeditions") in epidemiology have good contextual reasons for using far stronger penalties than those from reference priors. For example, a log-$F(3.9,3.9)$ prior for β would give 95% probability to e^{β} being between 1/10 and 10, while a normal(0,½) or a log-$F(9,9)$ prior for β would give 95% probability to e^{β} being between ¼ and 4 [15]; yet these priors would still be considered fairly weak if most odds ratios are expected to fall between ½ and 2, as in [27], or if β was the coefficient of a higher-order term as in [26]. Strong penalties or priors may also have frequentist as well as Bayesian rationales given goals of total mean-squared error reduction and improved generalizability of risk estimates [17-20], as typified by explorations of higher-order effects such as product terms ("interactions"). In particular, the number of 2-way products among K covariates, $K(K-1)/2$, increases quadratically with K, greatly aggravating sparsity problems unless very strong penalties are applied to these terms.

5. A CASE STUDY

Sullivan & Greenland [15] contrasted results from logistic regressions of neonatal death on 14 risk factors, using priors ranging from highly informative to very weak. The data were very sparse, with only 17 deaths among 2,992 births. See [15] for detailed summaries, source citations, and full regression analyses including results from Markov-Chain Monte-Carlo (MCMC) as well as penalized likelihood; online supplements provide the data, along with SAS code for penalized likelihood via data augmentation in ordinary logistic, conditional logistic, Poisson, and proportional-hazards models with normal and log-F priors. Of interest here is that several coefficient MLEs were badly inflated. We focus on and present additional results for hydramnios (excess amniotic fluid), which occurred in 10 pregnancies including one death.

The unadjusted hydramnios odds ratio is (1/9)/(16/2,966) = 21 with mid-P limits of 0.88, 136 from Stata [10], quite imprecise but in accord with clinical expectations of about a 10-fold increase in death risk from the condition. This estimate is the MLE obtained from a univariate regression. Applying the Firth correction to this regression is equivalent to adding ½ to each count, producing an odds ratio of (1.5/9.5)/(16.5/2,982.5) = 28; this is an example of "Bayesian noncollapsibility" in that the posterior mode of ln(28) is outside the range of the prior mode of 0 and the MLE of ln(21) [28]. In contrast, using a ln-$F(1,1)$ prior for the log odds ratio (with or without the same prior for the intercept) produces an odds ratio of 11.

Now let β be the hydramnios-indicator coefficient and b an estimate of β from the logistic regression with all 14 covariates and the intercept; the adjusted odds ratio for hydramnios is then e^{β} with estimate e^{b}. Table 1 summarizes the e^{b} and 95% prior, Wald (log-symmetric), and profile-likelihood limits (PLLs) for e^{β} obtained from different penalties and priors, based on maximization of the indicated penalized likelihood. The MLE e^{b} of e^{β} was 60, about six times the clinical expectations and quite inflated relative to the unadjusted MLE. The 95% Wald limits were 5.7, 635, compared to PLLs of 2.8, 478; this discrepancy reflects the severe skewness of the likelihood. Results from the Jeffreys prior using the SAS FIRTH option appeared even more inflated yet more precise, giving $e^{b} = 68$ (Wald limits 9.2, 505 from Stata and 9.1, 510 from SAS; PLLs 6.1, 421); although the total distance $||\boldsymbol{b}|| = (\boldsymbol{b}\,'\boldsymbol{b})^{1/2}$ of the coefficient vector $\boldsymbol{b}$ to the origin was reduced (MLE $||\boldsymbol{b}|| = 8.82$, Firth $||\boldsymbol{b}|| = 8.63$), 7 other point estimates also moved away from zero relative to their MLEs. Direct coefficient bias correction [29] gave estimates very close to the Firth correction.

In contrast, independent log-F(0.62,0.62) priors for all coefficients including the intercept (with 0.62 chosen to match the estimated degrees of freedom of Firth penalization in this example, trace($I^{*-1}I$) = 14.04 [18,19]) produced $e^{b} = 34$ (Wald limits 2.6, 460; PLLs 1.2, 303), more consistent with clinical expectations, with all estimates moving toward zero. Similarly, independent $2.5t_1$ priors for all coefficients and $10t_1$ for the intercept using *bayesglm* in the *arm* package in R [30] produced $e^{b} = 30$ (Wald limits 2.4, 356, PLLs not given by *bayesglm*), while independent log-F(1,1) priors for all coefficients including the intercept produced $e^{b} = 23$ (Wald limits 1.4, 379; PLLs 0.76, 226), again with all estimates moving toward zero in both cases. The contrast illustrates how results from Jeffreys and independence priors can diverge; we found no clear clinical interpretation of the Jeffreys prior or the divergence in current Firth-penalization software [3-10] (which does not compute Jeffreys prior limits).

The intercept is the log odds of death among those with no risk factor (all covariates zero), and must be very negative given that the proportion dying in the entire cohort is only 17/2992 = 0.006. Thus, following our own advice to avoid shrinking the intercept toward zero, we now exclude its prior (or equivalently, assign it a log-F(0,0) prior). This has little impact on the results: keeping log-F(1,1) priors for the other coefficients produces $e^{b} = 24$ (PLLs 0.78, 243). Using instead log-F(2,2) priors produces $e^{b} = 8.1$ (PLLs 0.44, 117), again consistent with clinical expectations. However, using normal(0,1.38) priors [22] produces $e^{b} = 3.4$ (PLLs 0.37, 27), while using log-*F*(9,9) priors produces $e^b = 1.5$ (PLLs 0.41, 6.3), which appear overshrunk relative to these expectations. This overshrinkage is unsurprising given that both priors are inconsistent with the expectations (e.g., their 95th percentiles are 6.9 and 3.2, respectively), and illustrates the hazards of shrinking toward zero when a strong relation is expected. Shifting the log-*F*(9,9) priors upward to reflect expectations, with prior medians of 4 for the hydramnios odds ratio e^β and 1-4 for other covariate odds ratios, produces instead $e^b = 5.8$ (PLLs 1.6,22). Finally, using the same prior medians with variance ½ normal priors for the coefficients produced a posterior mode of $e^b = 6.1$ (PLLs 1.6, 23) from data augmentation and a posterior geometric mean of 6.0 (2.5th and 97.5th percentiles 1.6, 22) from MCMC [15].

6. UNPENALIZED OR PENALIZED INFORMATION?

The standard errors for Firth-corrected logistic regression produced by Stata [10] are slightly smaller than those produced by SAS [3] and R [6,8], as illustrated by the Firth intervals in Table 1. The explanation appears to be that (following Firth [1, sec. 5] and Heinze and colleagues [3,4]) SAS and R [6,8] as well as Statistica [31] use the unpenalized inverse information I^{-1} in Newton-Raphson iterations to maximize ℓ^*, taking the final I^{-1} as the estimated covariance matrix. In contrast, Stata [10] uses the penalized inverse information I^{*-1}; the smaller standard errors follow from the fact that I^* is augmented over I by the negative Hessian of $\frac{1}{2}\ln|I|$. The difference is generally minor [2,32]; for example, recall that in the matched-pair example, $I = ne^\beta/(1+e^\beta)^2$ whereas $I^* = (n+1)e^\beta/(1+e^\beta)^2$, leading to a variance ratio of $(n+1)/n$ and an estimated degrees of freedom of $n/(n+1)$.

Nonetheless, in our experience, use of I^* rather than I usually speeds convergence. These observations are unsurprising given that $-I^*$ is the gradient matrix of the score vector $\ell^{*\prime}$ and thus (unlike I) follows from the Gauss-Newton algorithm. Furthermore, in data augmentation [13], I^{*-1} arises as a first-order posterior covariance-matrix approximation. These facts do not however dictate that standard errors computed from I^* are superior, since both I^{-1} and I^{*-1} are only approximate covariance matrices. In particular, although $I^*$is the correct curvature matrix for ℓ^*, the covariance estimate I^{*-1} suffers from higher-order bias. Fortunately, the choice is rendered moot by computing intervals from the profile penalized loglikelihood, as usually recommended when n is not large or ℓ^* is not nearly quadratic [3,4,13-16], or by switching to a log-*F* prior.

7. DISCUSSION

Although we have focused on judging penalty strength from contextual consideration of the corresponding prior distribution, strength can instead (or also) be determined by cross-validation or other data-based methods [17-20], thus keeping the analysis within the frequentist empirical-Bayes sphere. Both Bayesian and empirical-Bayes analyses may be conducted and contrasted, an activity we encourage to provide a higher level of cross-validation, and to build more understanding of when these different approaches tend to agree or conflict in substance.

We have argued that, in health and social-science contexts in which these methods are reasonable approaches and penalization is desired, both the Firth correction and *t*-distribution priors are subject to serious objections. Based on simplicity of interpretation and computation, we propose that the log-$F(1,1)$ prior provides a better default penalty and reference prior, although we advise that a properly shifted prior or (more conveniently) no prior be used when for intercepts or for coefficients that could reasonably be enormous. We do not however suggest ignoring MLEs, for at the very least they serve as diagnostics: when they diverge or appear absurdly inflated, we are alerted to the fact that subsequent estimates will be profoundly sensitive to the penalty chosen to address the problem.

There are hazards associated with use of any prior or penalty, including default ones. We advise that close attention be paid to whether prior dependencies or independencies are reasonable for the model parameterization. Some independencies may turn out to be absurd when considered in context. For example, if β_1 and β_2 represent rates (or log rates or logits) of classification error or disease in two exposure categories indicated by x_1 and x_2, severe prior dependence is only to be expected since the contextual reference state is the null hypothesis of equality ($\beta_1=\beta_2$, complete dependence) [26, 28]. In such cases, contextually sensible options are to either include a correlation hyperparameter in the joint prior (which can be quite difficult to specify) [13], or else recode covariates to obtain a parameterization for which an independence prior might be reasonable, such as $c = x_1 + x_2 = 1$ and x_2, whose coefficients are β_1 and $\beta_2^* = \beta_2 - \beta_1$ [26,28].

Because e^β represents the odds ratio associated with a unit change in the corresponding covariate, we strongly advise that quantitative covariates be scaled in units or spans that are contextually meaningful, important, and generalizable (recognizable and transportable). This means avoiding standard deviations and other study-dependent scales (unlike Gelman et al. [23]), and using instead simple multiples of SI units within the range of the data. For example, blood pressure could be scaled in cm, adult ages in decades. Use of SI units allows checking of prior and posterior percentiles for the odds ratios e^{β} against contextual information [24], and aids interpretability and comparability of estimates e^{b} across studies [33]. We also advise centering quantitative covariates so that 0 is a meaningful reference value within the range of the data, which allows sensible interpretation of the intercept as the logit of risk when all covariates are zero; similarly, when a product term x_1x_2 is entered, it allows sensible interpretation the main-effect coefficient β_1 for x_1 as a log-odds ratio when x_2 (and hence x_1x_2) is zero. Nonetheless, we caution that if the outcome is common ($\pi > 0.10$) at some covariate levels, odds ratios become difficult to interpret correctly because they no longer approximate risk ratios; thus it will be better to compare fitted probabilities (risks) directly than to use the e^{b} as effect estimates [34].

As a practical matter we have recommended using no intercept prior because the intercept is a direct function of which covariates are included and their coding. In particular, the intercept can be extremely sensitive to how covariates are centered (is age in years since birth? or recentered to its study-specific mean? or entered instead as category indicators?) and ordered (are males=1, females=0? or females=1, males=0?). It is also extremely sensitive to the sampling design: studies may select low-risk or high-risk groups, or highly exposed groups, completely distorting the intercept relative to vital-statistics data or previous studies that might inform priors. Worse, the intercept is cut off from background information by outcome-dependent (case-control or choice-based) sampling, which makes the intercept mainly a function of the chosen outcome-sampling ratios. These dependencies render reliance on past intercept estimates dubious at best. The distortion capacity of the intercept prior will be limited if it is weak, but again (and unlike with most coefficients) the intercept is usually unlikely to be near zero. Thus, simply dropping the intercept prior seems to us the safest default.

Similar coding sensitivities will arise in main-effect coefficients when their covariates appear in product terms. Nonetheless, sensible centering may render shrinkage of such coefficients toward zero reasonable, which along with improbability of large values may justify use of weak default priors for them.

References

[1] Firth D. Bias reduction of maximum likelihood estimates. Biometrika 1993;80:27-38. Correction: Biometrika 1995;82:667.

[2] Heinze G, Schemper M. solution to the problem of monotone likelihood in Cox regression. Biometrics 2001;57:114-119.

[3] Heinze G, Ploner M. Fixing the nonconvergence bug in logistic regression with SPLUS and SAS. Computer Methods and Programs in Biomedicine 2003;71:181-187.

[4] Heinze G, Puhr R. Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets. Statistics in Medicine 2010;29:770-777.

[5] Heinze G, Ploner M. SAS and SPLUS programs to perform Cox without convergence problems. Computer Methods and Programs in Biomedicine 2002;67:217-223.

[6] Heinze G, Ploner M, Dunkler D, Southworth H. Logistf: Firth's bias reduced logistic regression. R package version 1.21. 2013.

[7] Ploner M, Heinze G. Coxphf: Cox regression with Firth's penalized likelihood. R package version 1.10. 2013.

[8] Heinze G, Ladner T. logistiX: Exact logistic regression including Firth correction. R package version 1.0-1. 2013.

[9] Kosmidis I. Brglm: Bias reduction in binomial-response generalized linear models. R package version 0.5-9. 2013.

[10] Coveney J. FIRTHLOGIT: Stata module to calculate bias reduction in logistic regression. Statistical Software Components, Department of Economics. 2008.

[11] Chen MH, Ibrahim JG, Kim S. Properties and Implementation of Jeffreys's Prior in Binomial Regression Models. Journal of the American Statistical Association 2008;103:1659-1664.

[12] Berger J, Bernardo JM, Sun D. The formal definition of reference priors. The Annals of Statistics 2009;37:905-938.

[13] Greenland S. Generalized conjugate priors for Bayesian analysis of risk and survival regressions. Biometrics 2003;59:92-99.

[14] Greenland S. Prior data for non-normal priors. Statistics in Medicine 2007;26:3578–3590.

[15] Sullivan SG, Greenland S. Bayesian regression in SAS software. International Journal of Epidemiology 2013;42:308-317.

[16] Cole SR, Chu H, Greenland S. Maximum likelihood, profile likelihood, and penalized likelihood: A primer. American Journal of Epidemiology 2014;179:252-260.

[17] Moons KG, Donders AR, Steyerberg EW, Harrell EF. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. Journal of Clinical Epidemiology 2004;57:1262-1270.

[18] Harrell F. Regression modeling strategies. New York: Springer, 2001.

[19] Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York, NY: Springer, 2008.

[20] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction. 2nd ed. New York: Springer, 2009.

[21] Breslow NE, Day NE. Statistical methods in cancer research. Vol I: the analysis of case-control data. Lyon: IARC, 1980.

[22] Hamra GB, MacLehose RF, Cole SR. Sensitivity analyses for sparse-data problems—using weakly informative Bayesian priors. Epidemiology 2013;24: 233-239.

[23] Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics 2008;2:1360-1383.

[24] Greenland S. Bayesian methods for epidemiologic research. II. Regression analysis. International Journal of Epidemiology 2007;36:195-202.

[25] Jones MC. Families of distributions arising from distributions of order statistics. Test 2004;13:1–43.

[26] Greenland S. Bayesian perspectives for epidemiologic research. III. Bias analysis via missing-data methods. Int J Epidemiol 2009;38:1662-73. Corrigendum: International Journal of Epidemiology 2010;39:1116.

[27] Greenland S. When should epidemiologic regressions use random coefficients? Biometrics 2000;56:915-21.

[28] Greenland S. Simpson's paradox from adding constants in contingency tables as an example of Bayesian noncollapsibility. American Statistician 2010;64:340-344.

[29] Schaefer RL. Bias correction in maximum likelihood logistic regression. Statistics in Medicine 1983;2:71-78.

[30] Gelman A, Su YS, Yajima M, Hill J, Pittau MG, Kerman J, Zheng T, Dorie V. Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.7-03. 2014.

[31] Fijorek K, Sokolowski A. Separation-resistant and bias-reduced logistic regression: STATISTICA macro. Journal of Statistical Software 2012;47:1-12.

[32] Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression, 3rd ed. New York: Wiley, 2013, P. 392.

[33] Greenland S, Maclure M, Schlesselman JJ, Poole C, Morgenstern H. Standardized regression coefficients: a further critique and review of some alternatives. Epidemiology 1991;2:387-392.

[34] Greenland S. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. American Journal of Epidemiology 2004;160:301-5.

Table 1. Penalized-likelihood estimates e^b of the adjusted odds ratio e^β for hydramnios, using different penalties (priors) in a logistic regression of neonatal death on 14 risk factors in a cohort of 2992 births with 17 deaths (see [15] for further model details and results for other coefficients).

Priors for β	Fitting method	Model df[§]	Prior median for coefficient antilogs e^β (95% prior interval)	OR estimate e^b (95% Wald & profile-likelihood intervals)
log-F(0,0)	ML	15	undefined	60 (5.7, 635),(2.8, 478)
Jeffreys[*]	Firth penalty	14.0	1 (not given by programs)	68 (9.2, 505)[**],(6.1, 421)
log-F(0.62,0.62)[*]	DAP	14.0	$(4.46\times10^{-5}, 2.24\times10^{5})$	34 (2.6, 460),(1.2, 303)
$2.5t_1$ ($10t_1$ for intercept) [23][*†]	as per *arm* [29]	13.4	1 $(1.60\times10^{-14}, 6.25\times10^{13})$	30 (2.4, 356),(not given)
log-F(1,1)[*]	DAP	13.5	1 (1/648,648)	23 (1.4, 379),(0.76, 226)
log-F(1,1)	DAP	13.5	1 (1/648,648)	24 (1.4, 407),(0.78, 243)
log-F(2,2)	DAP	12.0	1 (1/39,39)	8.1 (0.36, 179),(0.44, 117)
Normal(0,1.38)	DAP with scale = 20	10.4	1 (1/10,10)	3.4 (0.40, 30),(0.37, 26)
log-F(9,9)	DAP	7.2	1 (1/4,4)	1.5 (0.41, 5.7),(0.41, 6.3)
log-F(9,9) centered at ln(4)[‡]	DAP	7.7	4 (1,16)	5.8 (1.6, 21),(1.6, 22)
Normal(ln(4),½)[‡]	DAP with scale = 20	8.0	4 (1,16)	6.1 (1.6, 23),(1.6, 23)
Normal(ln(4),½)[‡]	MCMC	Not given	4 (1,16)	6.0 (1.6, 22)[‡‡]

ML: maximum likelihood; DAP: data augmentation prior [13-15]; MCMC: Markov-chain Monte Carlo.

[*] Intercept included in prior; all other fits used no intercept prior.

[†] $2.5t_1$ = Cauchy distributions with center 0 and scale 2.5 for binary predictors, 2.5/(2×SD) for quantitative variables with standard deviation SD, 10 for intercept,.

[‡] Prior medians of 1-4 are used for other covariate odds ratios.

[§] No. model parameters for ML; estimated degrees of freedom = trace($I^{*-1}I$) otherwise [18,19].

[**] from Stata; (9.1, 510) from SAS.

[‡‡] Simulated posterior geometric mean and 2.5th, 97.5th percentiles from 100,000 samples using BAYES statement in SAS GENMOD [15].