

A default prior distribution for logistic and other regression models*

Andrew Gelman[†] Aleks Jakulin[‡] Maria Grazia Pittau[§] and Yu-Sung Su[¶]

November 19, 2006

Abstract

We propose a new prior distribution for classical (non-hierarchical) logistic regression models, constructed by first scaling all nonbinary variables to have mean 0 and standard deviation 0.5, and then placing independent Student- t prior distributions on the coefficients. As a default choice, we recommend the Cauchy distribution with center 0 and scale 2.5, which in the simplest setting is a longer-tailed version of the distribution attained by assuming one-half additional success and one-half additional failure in a logistic regression. We implement a procedure to fit generalized linear models in R with this prior distribution by incorporating an approximate EM algorithm into the usual iteratively weighted least squares. We illustrate with several examples, including a series of logistic regressions predicting voting preferences, an imputation model for a public health dataset, and a hierarchical logistic regression in epidemiology.

We recommend this default prior distribution for routine applied use. It has the advantage of always giving answers, even when there is complete separation in logistic regression (a common problem, even when the sample size is large and the number of predictors is small) and also automatically applying more shrinkage to higher-order interactions. This can be useful in routine data analysis as well as in automated procedures such as chained equations for missing-data imputation.

Keywords: Bayesian inference, generalized linear model, least squares, hierarchical model, linear regression, logistic regression, multilevel model, noninformative prior distribution

1 Introduction

Nonidentifiability is a common problem in logistic regression. In addition to the problem of collinearity, familiar from linear regression, discrete-data regression can also become unstable from *separation*, which arises when a linear combination of the predictors is perfectly

*We thank Chuanhai Liu, David Dunson, and Hal Stern for helpful comments, Peter Messeri for the HIV example, Ezra Susser, Mary Beth Terry, and Rafael Guerrero Preston for the stomach cancer example, and the National Science Foundation and National Institutes of Health for financial support.

[†]Department of Statistics and Department of Political Science, Columbia University, New York, gelman@stat.columbia.edu, www.stat.columbia.edu/~gelman

[‡]Department of Statistics, Columbia University, New York

[§]Department of Statistics, Columbia University, New York

[¶]Department of Political Science, City University of New York

predictive of the outcome (Albert and Anderson, 1984, Lesaffre and Albert, 1989). Separation is surprisingly common in applied logistic regression, especially with binary predictors, and, as noted by Zorn (2005), is often handled inappropriately. For example, a common “solution” to separation is to remove predictors until the resulting model is identifiable, but, as Zorn (2005) points out, this typically results in removing the strongest predictors from the model.

An alternative approach to obtaining stable logistic regression coefficients is to use Bayesian inference. Various prior distributions have been suggested for this purpose, most notably a Jeffreys prior distribution (Firth, 1993), but these have not been set up for reliable computation and are not always clearly interpretable as prior information in a regression context. Here we propose a new, proper prior distribution that produces stable, regularized estimates while still being vague enough to be used as a default in routine applied work.

2 A default prior specification

A challenge in setting up any default prior distribution is getting the scale right: for example, suppose we are predicting vote preference given age (in years). We would not want the same prior distribution if the age scale were shifted to months. But discrete predictors have their own natural scale (most notably, a change of 1 in a binary predictor) that we would like to respect.

On one hand, scale-free prior distributions such as Jeffreys’ do not include enough prior information; on the other, what prior information can be assumed for a generic model? Our key idea is that actual effects tend to fall within a limited range. For logistic regression, a change of 5 moves a probability from 0.01 to 0.5, or from 0.5 to 0.99. We rarely encounter situations where a shift in input x corresponds to the probability of outcome y changing from 0.01 to 0.99, hence we are willing to assign a prior distribution that assigns low probabilities to changes of 10 on the logistic scale.

2.1 Standardizing input variables to a commonly-interpretable scale

The first step of the model is to standardize the input variables (Gelman, 2006b):

- Binary inputs are shifted to have a mean of 0 and to differ by 1 in their lower and upper conditions. (For example, if a population is 10% African-American and 90% other, we would define the centered “African-American” variable to take on the values 0.9 and -0.1 .)

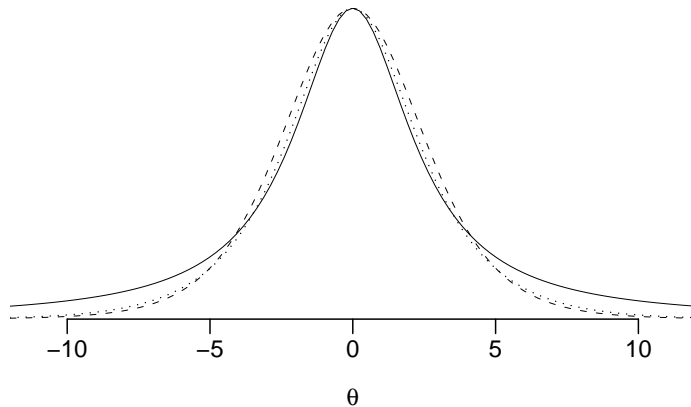


Figure 1: (solid line) Cauchy density function with scale 2.5, (dashed line) t_7 density function with scale 2.5, (dotted line) likelihood for θ corresponding to a single binomial trial of probability $\text{logit}^{-1}(\theta)$ with one-half success and one-half failure. All these curves favor values below 5 in absolute value; we choose the Cauchy as our default model because it allows the occasional probability of larger values.

- Other inputs are shifted to have a mean of 0 and scaled to have a standard deviation of 0.5. This scaling puts continuous variables on the same scale as symmetric binary inputs (which, taking on the values ± 0.5 , have standard deviation 0.5).

Following Gelman and Pardoe (2007), we distinguish between regression *inputs* and *predictors*. For example, in a regression on age, sex, and their interaction, there are four predictors (the constant term, age, sex, and age \times sex), but just two inputs: age and sex. It is the input variables, not the predictors, that are standardized.

2.2 A weakly informative t family of prior distributions

The second step of the model is to define prior distributions for the coefficients of the predictors. We use the Student- t prior distribution with mean 0, degrees-of-freedom parameter ν , and scale s , with ν and s chosen to provide minimal prior information to constrain the coefficients to lie in a reasonable range.

One way to pick a default value of ν and s is to consider the baseline case of one-half of a success and one-half of a failure for a single binomial trial with probability $p = \text{logit}^{-1}(\theta)$ —that is, a logistic regression with only a constant term. The corresponding likelihood is $e^{\theta/2}/(1 + e^{\theta})$, which is close to a t density function with 7 degrees of freedom and scale 2.5 (Liu, 2004). We shall choose a slightly more conservative choice, the Cauchy, or t_1 , distribution, again with a scale of 2.5. Figure 1 shows the three density functions: they all give preference to values less than 5, with the Cauchy allowing the occasional possibility of

very large values (a point to which we return in Section 5).

We assign independent Cauchy prior distributions with center 0 and scale 2.5 to each of the coefficients in the logistic regression except the constant term. When combined with the standardization, this implies that the absolute difference in logit probability should be less than 5, when moving from one standard deviation below the mean, to one standard deviation above the mean, in any input variable.

If we were to apply this prior distribution to the constant term as well, we would be stating that the success probability is probably between 1% and 99% for units that are average in all the inputs. Depending on the context (for example, epidemiologic modeling of rare conditions, as in Greenland, 2001), this might not make sense, so as a default we apply a weaker prior distribution—a Cauchy with center 0 and scale 10, which implies that we expect the success probability for an average case to be between 10^{-9} and $1 - 10^{-9}$.

An appealing byproduct of applying the model to rescaled predictors is that it automatically implies more stringent restrictions on interactions. For example, consider three symmetric binary inputs, x_1, x_2, x_3 . From the rescaling, each will take on the values $\pm 1/2$. Then any two-way interaction will take on the values $\pm 1/4$, and the three-way interaction can be $\pm 1/8$. But all these coefficients have the same default prior distribution, so the total contribution of the three-way interaction (for example) is $1/4$ that of the main effect. Going from the low value to the high value in any given three-way interaction is, in the model, unlikely to change the logit probability by more than $5 \cdot (1/8 - (-1/8)) = 5/4$ on the logit scale.

3 Computation

In principle, logistic regression with our prior distribution can be computed using the Gibbs and Metropolis algorithms. We do not give details as this is now standard with Bayesian models (see, for example, Carlin and Louis, 2001, Martin and Quinn, 2002, and Gelman et al., 2003). In practice, however, it is desirable to have a quick calculation that returns a point estimate of the regression coefficients and standard errors. Such an approximate calculation fits in better with routine statistical practice and, in addition, recognizes the approximate nature of the model itself.

We consider three computational settings:

- Classical (non-hierarchical) logistic regression, using our default prior distribution in place of the usual flat prior distribution on the coefficients.

- Multilevel (hierarchical) modeling, in which some the default prior distribution is applied only to the subset of the coefficients that are not otherwise modeled (sometimes called the “fixed effects”).
- Chained imputation, in which each variable with missing data is modeled conditional on the other variables with a regression equation, and these models are fit and random imputations inserted iteratively (Van Buuren and Oudshoorn, 2000, Raghunathan, Van Hoewyk, and Solenberger, 2001).

In any of these cases, our default prior distribution has the purpose of stabilizing (regularizing) the estimates of otherwise unmodeled parameters. In the first scenario, we typically only want point estimates and standard errors (unless the sample size is so small that the normal approximation to the posterior distribution is inadequate). In the second scenario, it makes sense to embed the computation within the full Markov chain simulation. In the third scenario of missing-data imputation, we would like the flexibility of quick estimates for simple problems with the potential for Markov chain simulation as necessary. Also, because of the automatic way in which the component models are fit in a chained imputation, we would like a computationally stable algorithm that returns reasonable answers.

We have implemented these computations by altering the `glm` function in R, creating a new function, `bayesglm`, which finds an approximate posterior mode and variance using extensions of the classical generalized linear model computations, as described in the rest of this section. The `bayesglm` function allows the user to specify independent prior distributions for the coefficients in the t family, by default using the Cauchy distribution with center 0 and scale 2.5, and also can fit hierarchical models in which regression coefficients are structured in batches. Furthermore, the `standardize` function in R automatically rescales regression inputs by centering and dividing by two standard deviations (Gelman, 2006b), and so using these two functions together performs our recommended procedure automatically.

3.1 Incorporating the prior distribution into classical logistic regression computations

Working in the context of the logistic regression model,

$$\Pr(y_i = 1) = \text{logit}^{-1}(X_i\beta), \tag{1}$$

we adapt the classical maximum likelihood algorithm to obtain approximate posterior inference for the coefficients β , in the form of an estimate $\hat{\beta}$ and covariance matrix V_β .

The standard logistic regression algorithm—upon which we build—proceeds by approximately linearizing the derivative of the log-likelihood, solving using weighted least squares, and then iterating this process, each step evaluating the derivatives at the latest estimate $\hat{\beta}$ (see, for example, McCullagh and Nelder, 1989). At each iteration, the algorithm determines pseudo-data z_i and pseudo-variances $(\sigma_i^z)^2$ based on the linearization of the derivative of the log-likelihood:

$$\begin{aligned} z_i &= X_i \hat{\beta} + \frac{(1 + e^{X_i \hat{\beta}})^2}{e^{X_i \hat{\beta}}} \left(y_i - \frac{e^{X_i \hat{\beta}}}{1 + e^{X_i \hat{\beta}}} \right) \\ (\sigma_i^z)^2 &= \frac{1}{n_i} \frac{(1 + e^{X_i \hat{\beta}})^2}{e^{X_i \hat{\beta}}}. \end{aligned} \tag{2}$$

and then performs weighted least squares, regressing z on X with weight vector $(\sigma^z)^{-2}$. The resulting estimate $\hat{\beta}$ is used to update the computations in (2), and the iteration proceeds until approximate convergence.

Computation with a specified normal prior distribution. The simplest informative prior distribution assigns normal prior distributions for the components of β :

$$\beta_j \sim N(\mu_j, \sigma_j^2), \quad \text{for } j = 1, \dots, J.$$

This information can be effortlessly included in the classical algorithm by simply altering the least-squares step, augmenting the approximate likelihood with the prior distribution (see, e.g., Section 14.8 of Gelman et al., 2003). If the model has J coefficients β_j with independent $N(\mu_j, \sigma_j^2)$ prior distributions, then we add J pseudo-data points and perform weighted linear regression on “observations” z_* , “explanatory variables” X_* , and weight vector w_* , where

$$z_* = \begin{pmatrix} z \\ \mu \end{pmatrix}, \quad X_* = \begin{pmatrix} X \\ I_J \end{pmatrix}, \quad w_* = (\sigma^z, \sigma)^{-2}. \tag{3}$$

Here, z_* and w_* are vectors of length $n + J$ and X_* is a $(n + J) \times J$ matrix. With the augmented X_* , this regression is identified, and thus the resulting estimate $\hat{\beta}$ is well defined and has finite variance, even if the original data have collinearity or separation that would result in nonidentifiability of the maximum likelihood estimate.

The full computation is then an iteratively weighted least squares, starting with a guess of β (e.g., independent draws from the unit normal distribution), then computing the derivatives of the log-likelihood to compute z and σ_z , then using weighted least squares on the pseudodata (3) to yield an updated estimate of β , then recomputing the derivatives of the

log-likelihood at this new value of β , and so forth, converging to the estimate $\hat{\beta}$. The covariance matrix V_β is simply the inverse second derivative matrix of the log-posterior density evaluated at $\hat{\beta}$ —that is, the usual normal-theory uncertainty estimate for an estimate not on the boundary of parameter space.

Approximate EM algorithm with a t prior distribution. If the coefficients β_j have t prior distributions with centers μ_j and scales s_j ,¹ we can program a similar procedure, using the formulation

$$\beta_j \sim N(\mu_j, \sigma_j^2), \quad \sigma_j^2 \sim \text{Inv-}\chi^2(\nu_j, s_j^2) \quad (4)$$

and averaging over the β_j 's at each step, treating them as missing data and performing one step of the EM algorithm to estimate the σ_j 's. Once enough iterations have been performed to reach approximate convergence, we get an estimate and covariance matrix for the vector parameter β the estimated σ_j 's.

We initialize the algorithm by setting each σ_j to the value s_j (the scale of the prior distribution) and, as before, starting with a guess of β . Then, at each step of the algorithm, we update σ by maximizing the expected value of its (approximate) log-posterior density,

$$\begin{aligned} \log p(\beta, \sigma | y) \approx & -\frac{1}{2} \sum_{i=1}^n \frac{1}{(\sigma_i^z)^2} (z_i - X_i \beta)^2 - \frac{1}{2} \sum_{j=1}^J \left(\frac{1}{\sigma_j^2} (\beta_j - \mu_j) + \log(\sigma_j^2) \right) - \\ & p(\sigma_j | \nu_j, s_j) + \text{constant}. \end{aligned} \quad (5)$$

Each iteration of the algorithm proceeds as follows:

1. Based on the current estimate of β , perform the normal approximation to the log-likelihood and determine the vectors z and σ^z using (2), as in classical logistic regression computation.
2. Approximate E-step: first run the weighted least squares regression based on the augmented data (3) to get an estimate $\hat{\beta}$ with variance matrix V_β . Then determine the expected value of the log-posterior density by replacing the terms $(\beta_j - \mu_j)^2$ in (5) by

$$E((\beta_j - \mu_j)^2 | \sigma, y) \approx (\hat{\beta}_j - \mu_j)^2 + (V_\beta)_{jj}, \quad (6)$$

which is only approximate because the normal distribution for z in (5) is only an approximation to the generalized linear model likelihood for y .

¹As discussed earlier, we set $\mu_j = 0$, $s_j = 2.5$, $\nu_j = 1$ as a default, but we describe the computation more generally in terms of arbitrary values of these parameters.

3. M-step: maximize the expected value of the log-posterior density (5) to get the estimate,

$$\hat{\sigma}_j^2 = \frac{(\hat{\beta}_j - \mu_j)^2 + (V_\beta)_{jj} + \nu_j s_j^2}{1 + \nu_j}, \quad (7)$$

which corresponds to the posterior mode of σ_j^2 given a single measurement with value (6) and an $\text{Inv-}\chi^2(\nu_j, s_j^2)$ prior distribution.

4. Recompute the derivatives of the log-posterior density given the current $\hat{\beta}$, set up the augmented data (3) using the estimated $\hat{\sigma}$ from (7), and repeat steps 1,2,3 above.

At convergence of the algorithm, we summarize the inferences using the latest estimate $\hat{\beta}$ and covariance matrix V_β .

3.2 Multilevel (hierarchical) modeling

Although not the main topic of this paper, hierarchical logistic regression models can be fit in a similar approximate manner using an extension of the above EM algorithm to average over hyperparameters. Consider the model (1) with the vector β of coefficients partitioned into *batches* $k = 0, \dots, K$. We write the individual coefficients as β_{jk} , $j = 0, \dots, J_k$, for the coefficients in batch k , label the total number of coefficients as $J = J_0 + J_1 + \dots + J_K$. We correspondingly partition the matrix X of predictors into X^0, X^1, \dots, X^K , the columns corresponding to each batch of coefficients. We define W as the $J \times K$ matrix of 0's and 1's indicating which elements of β are in which batches. (Components of β that are in batch 0 are represented by rows of 0's in W ; each of the others has a 1 in the column corresponding to its batch and 0 elsewhere.)

Batch 0 is different from all the others, representing the “fixed effects” that have independent prespecified prior distributions. As with the nonhierarchical model in Section 2, we use the t family:

$$\beta_{j0} \sim t_{\nu_{j0}}(\mu_{j0}, (s_{j0})^2), \text{ for } j = 1, \dots, J_0. \quad (8)$$

The other batches represent the “random effects,” which are modeled hierarchically. We use the notation β^0 for the subvector $(\beta_{j0}, j = 1, \dots, J_0)$ and β^+ for the subvector of the other $J_+ = J - J_0$ coefficients, with $X^+ = (X^1, \dots, X^K)$ being our shorthand for the submatrix of X excluding X^0 . We model the coefficients exchangeably within each batch:

$$\beta_{jk} \sim N(\mu_k^{\text{batch}}, (\sigma_k^{\text{batch}})^2), \text{ for } j = 1, \dots, J_k, k = 1, \dots, K. \quad (9)$$

The batch means μ_k^{batch} are themselves regression coefficients that would typically be unmodeled (or given noninformative prior distributions), so we give them t prior distributions:

$$\mu_k^{\text{batch}} \sim t_{\nu_k^{\mu.\text{batch}}}(\mu^{\mu.\text{batch}}, (s_k^{\mu.\text{batch}})^2), \text{ for } k = 1, \dots, K, \quad (10)$$

We also assign independent prior distributions to the batch variances:

$$(\sigma_k^{\text{batch}})^2 \sim \text{Inv-}\chi^2(\nu_j^{\sigma.\text{batch}}, (s_j^{\sigma.\text{batch}})^2), \text{ for } k = 1, \dots, K. \quad (11)$$

All these steps may seem overly elaborate, but it is really a simple hierarchical model, with complexities arising only because we are using proper prior distributions rather than uniform densities for the usually-unmodeled parameter vectors β^0 , μ^{batch} , and σ^{batch} .

Batch means and error terms. For later convenience in computation, we define each of the batched coefficients as a batch mean plus an error term:

$$\beta_{jk} = \mu_k^{\text{batch}} + \alpha_{jk}, \text{ for } j = 1, \dots, J^k, \text{ } k = 1, \dots, K. \quad (12)$$

Since β and μ^{batch} have already been defined, this equation serves as a definition of α . The logistic regression model (1) can then be written as,

$$\begin{aligned} \Pr(y_i = 1) &= \text{logit}^{-1} \left(X_i^0 \beta^0 + \sum_{k=1}^K X_i^k \beta^k \right) \\ &= \text{logit}^{-1} \left(X_i^0 \beta^0 + \sum_{k=1}^K X_i^k \alpha^k + \sum_{k=1}^K (XW)_{ik} \mu_k^{\text{batch}} \right) \\ &= \text{logit}^{-1} \left(X_i^0 \beta^0 + (X_i^+) \alpha + X_i W \mu^{\text{batch}} \right), \end{aligned} \quad (13)$$

separating the coefficients from each batch and then expanding into $J + K$ coefficients by pulling out the group means.

Expression of t distributions as scale mixtures. For computational reasons (as discussed in Section 3.1), it is helpful to express the t distributions (8) and (10) as mixtures of normals with unknown variances:

$$\begin{aligned} \beta_j^0 &\sim \text{N}(\mu_{j0}, (\sigma_{j0}^2)), & \sigma_{j0}^2 &\sim \text{Inv-}\chi^2(\nu_{j0}, (s_{j0})^2), & \text{ for } j = 1, \dots, J_k, \text{ } k = 1, \dots, K. \\ \mu_k^{\text{batch}} &\sim \text{N}(\mu_k^{\mu.\text{batch}}, (\sigma_k^{\mu.\text{batch}})^2), & (\sigma_k^{\mu.\text{batch}})^2 &\sim \text{Inv-}\chi^2(\nu_k^{\mu.\text{batch}}, (s_k^{\mu.\text{batch}})^2), & \text{ for } k = 1, \dots, K. \end{aligned}$$

EM algorithm for the hierarchical model. We again estimate the parameters of our model using an approximate EM algorithm, this time treating the regression coefficients (the β_{j0} 's, α_{jk} 's, and μ_k^{batch} 's) as missing data and averaging over them to estimate the variance parameters (the σ_{j0} 's, σ_k^{batch} 's, and $\sigma_k^{\mu.\text{batch}}$'s). We start the algorithm by picking arbitrary values of ... and then proceeding as follows:

1. Based on the current estimate of β , perform the normal approximation to the log-likelihood and determine the vectors z and σ^z using (2), as in classical logistic regression computation.
2. Approximate E-step: first construct augmented data included the just-computed pseudodata and the prior distributions on β^0 , α , and μ^{batch} :

$$z_* = \begin{pmatrix} z \\ \mu_0 \\ 0 \\ \mu^{\mu.\text{batch}} \end{pmatrix}, \quad X_* = \begin{pmatrix} X^0 & X^+ & XW \\ I_{J_0} & 0 & 0 \\ 0 & I_{J_+} & 0 \\ 0 & 0 & I_K \end{pmatrix}, \quad w_* = (\sigma_1^z, \sigma^0, \sigma^{\text{batch}}, \sigma^{\mu.\text{batch}})^{-2}. \quad (14)$$

Here, y_* and w_* are vectors of length $n+J+K$ and X_* is a $(n+J+K) \times (J+K)$ matrix.

Now run the weighted least squares regression based on the augmented data (14) to get an estimate $\hat{\beta}^{\text{all}}$ and variance matrix $V_{\hat{\beta}}^{\text{all}}$. Identify $\hat{\beta}^{\text{all}} = (\hat{\beta}_0, \hat{\alpha}, \hat{\mu}^{\text{batch}})$ —that is, the first J_0 elements of β^{all} correspond to the “fixed effects” β_{j0} , the next J_+ elements represent the “random” error terms α_{jk} , and the final K elements are the batch means μ_k^{batch} . We can then combine these last two parts to obtain the β_{jk} 's using (12).

We are now ready to determine the expected value of the relevant factors in the log-posterior density:

$$\begin{aligned} \text{E}((\beta_{j0} - \mu_{j0})^2 | \sigma, y) &\approx (\hat{\beta}_{j0} - \mu_{j0})^2 + (V_{\beta_0})_{jj} \\ \text{E}(\alpha_{jk}^2 | \sigma, y) &\approx \hat{\alpha}_{jk}^2 + (V_{\alpha})_{jk,jk} \\ \text{E}((\mu_k^{\text{batch}} - \mu_k^{\mu.\text{batch}})^2 | \sigma, y) &\approx (\hat{\mu}_k^{\text{batch}} - \mu_k^{\mu.\text{batch}})^2 + (V_{\mu^{\text{batch}}})_{kk}. \end{aligned} \quad (15)$$

3. M-step: maximize the expected value of the log-posterior density (5) to get the esti-

mates,

$$\begin{aligned}
\hat{\sigma}_{j0}^2 &= \frac{(\hat{\beta}_{j0} - \mu_{j0})^2 + (V_{\beta_0})_{jj} + \nu_{j0}s_{j0}^2}{1 + \nu_{j0}} \\
(\sigma^{\hat{\text{batch}}_k})^2 &= \frac{\sum_{j=1}^{J_k} (\hat{\alpha}_{jk}^2 + (V_{\alpha})_{jk,jk}) + \nu_k^{\text{batch}}(s_k^{\text{batch}})^2}{J_k + \nu_k^{\text{batch}}} \\
(\sigma^{\mu.\hat{\text{batch}}_k})^2 &= \frac{(\hat{\mu}_k^{\text{batch}} - \mu_k^{\mu.\text{batch}})^2 + (V_{\mu^{\text{batch}}})_{kk} + \nu_k^{\mu.\text{batch}}(s_k^{\mu.\text{batch}})^2}{1 + \nu_k^{\mu.\text{batch}}}. \quad (16)
\end{aligned}$$

4. Repeat the above steps and iterate until approximate convergence.

Hierarchical modeling with redundant parameterization. The above model can be slow to compute, and also the inverse-gamma prior distribution (11) can create problems for groups k where J_k , the number of coefficients in the group, is small. As discussed by Gelman (2006a), we can generalize the model by replacing the parameters α_{jk} in the above model by products of the form $\xi_k \gamma_{jk}$. The logistic regression then becomes,

$$\Pr(y_i = 1) = \text{logit}^{-1} \left(X_i^0 \beta^0 + \sum_{k=1}^K \xi_k X_i^k \alpha^k + XW \mu^{\text{batch}} \right),$$

which is the same as (13) except for the second term with the new ξ_k factors.

We perform an approximate EM algorithm on this expanded parameterization, treating the β_{j0} 's, γ_{jk} 's, and μ_k^{batch} 's as missing data, and now estimating the ξ_k 's as well as all the variance parameters:

need to fill this in

...

Finally, ξ and γ do not generally have any direct interpretation, and so we combine them: $\hat{\alpha}_{jk} = \hat{\xi}_k \hat{\gamma}_{jk}$ before reporting inferences. Also, se's and variance parameters

3.3 Chained imputation

4 Examples

4.1 A series of regressions predicting vote preferences

Regular users of logistic regression know that separation can occur in routine data analyses, even when the sample size is large and the number of predictors is small. The left column of Figure 2 shows a the estimated coefficients for logistic regression predicting probability

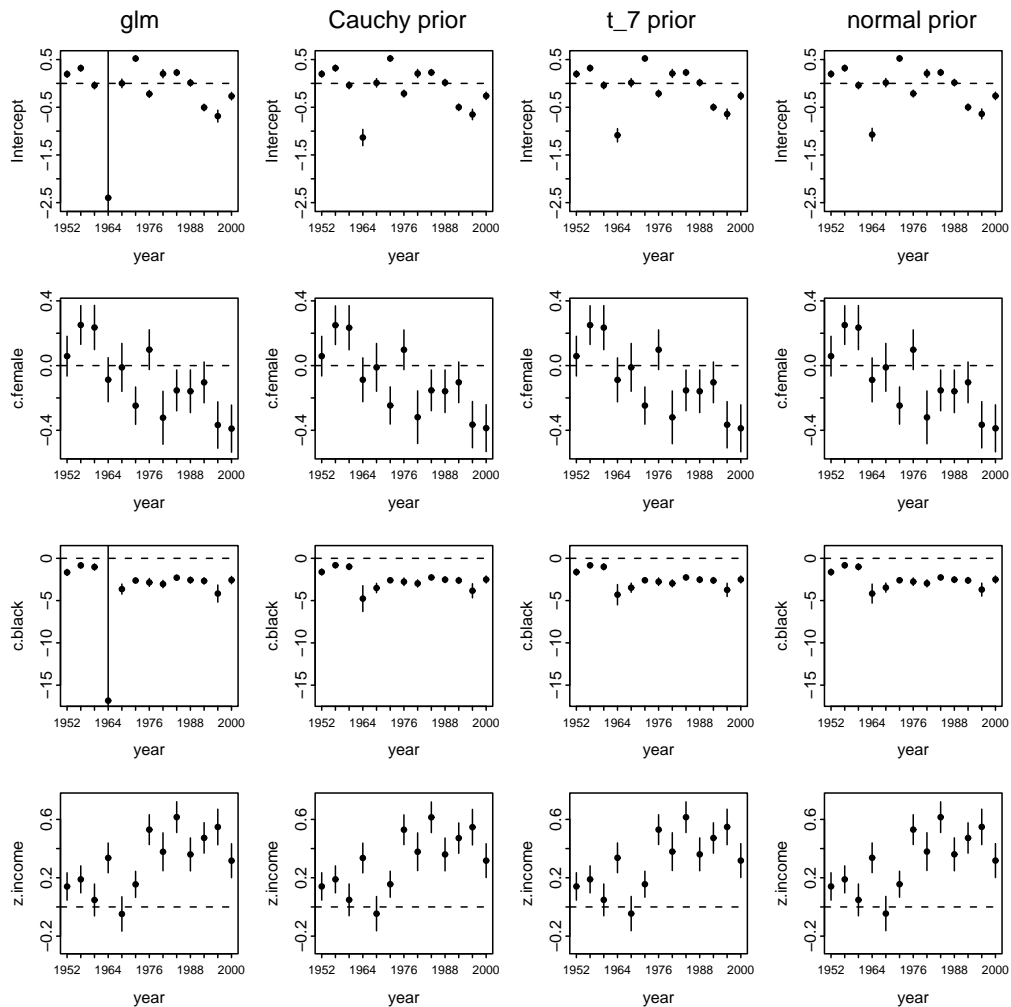


Figure 2: The left column shows the estimated coefficients (± 1 standard error) for a logistic regression predicting probability of Republican vote for President given sex, race, and income, as fit separately to data from the National Election Study for each election 1952 through 2000. (The binary inputs `female` and `black` have been centered to have means of zero, and the numerical variable `income` (originally on a 1–5 scale) has been centered and then rescaled by dividing by two standard deviations.)

There is complete separation in 1964 (with none of black respondents supporting the Republican candidate, Barry Goldwater), leading to a coefficient estimate of $-\infty$ that year. (The particular finite values of the estimate and standard error are determined by the number of iterations used by `glm` function in R before stopping.)

(other columns) Estimated coefficients (± 1 standard error) for the same model fit each year using independent Cauchy, t_7 , and normal prior distributions, each with center 0 and scale 2.5. All three prior distributions do a reasonable job at stabilizing the estimates for 1964, while leaving the estimates for other years essentially unchanged.

Dose, x_i (log g/ml)	Number of animals, n_i	Number of deaths, y_i
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

```
# from glm:
      coef.est coef.se
(Intercept) -0.1    0.7
z.x         10.2    6.4
  n = 4, k = 2
  residual deviance = 0.1, null deviance = 15.8 (difference = 15.7)

# from bayesglm (Cauchy priors, scale 10 for const and 2.5 for other coef):
      coef.est coef.se
(Intercept) -0.2    0.6
z.x         5.4    2.2
  n = 4, k = 2
  residual deviance = 1.1, null deviance = 15.8 (difference = 14.7)
```

Figure 3: Data from a bioassay experiment, from Racine et al. (1986), and estimates from classical maximum likelihood and Bayesian logistic regression with the recommended default prior distribution. The big change with the prior distribution may seem surprising at first, but upon reflection we prefer the smaller estimate, which is based on downweighting the most extreme possibilities that are allowed by the likelihood.

of Republican vote for President for a series of elections. The estimates look fine except in 1964, where there is complete separation, with all black respondents supporting the Democrats. (Fitting in R actually yields finite estimates, as displayed in the graph, but these are essentially meaningless, being a function of how long the iterative fitting procedure goes before giving up.)

The other three columns of Figure 2 show the coefficient estimates using our default Cauchy prior distribution for the coefficients, along with the t_7 and normal distributions. (In all cases, the prior distributions are centered at 0, with scale parameters set to 10 for the constant term and 2.5 for all other coefficients.) All three prior distributions do a reasonable job at stabilizing the estimated coefficient for race for 1964, while leaving the estimates for other years essentially unchanged. This example illustrates how we could use our Bayesian procedure in routine practice.

4.2 A small bioassay experiment

We next consider a small-sample example in which the prior distribution makes a difference for a coefficient that is already identified. The example comes from Racine et al. (1986), who used a problem in bioassay to illustrate how Bayesian inference can be applied with small samples. The top part of Figure 3 presents the data, from twenty animals that were exposed to four different doses of a toxin. The bottom parts of Figure 3 show the resulting logistic regression, as fit first using maximum likelihood and then using our default Cauchy prior distributions with center 0 and scale 10 (for the constant term) and 2.5 (for the coefficient of dose). Following our general procedure, we have rescaled dose to have mean 0 and standard deviation 0.5.

With such a small sample, the prior distribution actually makes a difference, lowering the estimated coefficient of standardized dose from 10.2 ± 6.4 to 5.4 ± 2.2 . Such a large change might seem disturbing, but for the reasons discussed above, we would doubt the effect to be as large as 10.2 on the logistic scale, and the analysis shows these data to be consistent with the much smaller effect size of 5.4. The large amount of shrinkage simply confirms how weak the information is that gave the original maximum likelihood estimate.

4.3 A set of chained regressions for missing-data imputation

Multiple imputation (Rubin, 1987, 1996) is another context in which regressions with many predictors are fit in an automatic way. Van Buuren and Oudshoorn (2000) and Raghunathan, Van Hoewyk, and Solenberger (2001) discuss implementations of the chained equation approach, in which variables with missingness are imputed one at a time, each conditional on the imputed values of the other variables, in an iterative random process that is used to construct multiple imputations. In chained equations, logistic regressions or similar models can be used to impute binary variables, and when the number of variables is large, separation can arise. Our prior distribution yields stable computations in this setting, as we illustrate in with example from our current applied research.

Separation occurred in the case of imputing virus loads in a longitudinal sample of HIV-positive homeless persons (Messerli et al., 2006). The imputation analysis incorporated a large number of predictors, including demographic and health-related variables, and often with high rates of missingness. Inside the multiple imputation chained equation procedure, logistic regression was used to impute the binary variables. It is generally recommended to include a rich set of predictors when imputing missing values (Rubin, 1996). However, in this example, including all the dichotomous predictors leads to many instances of separation.

```

# from glm:
      coef.est coef.sd          coef.est coef.sd
(Intercept)      0.07  1.41   h39b.W1          -0.10 0.03
age.W1           0.02  0.02   pcs.W1          -0.01 0.01
mcs37.W1        -0.01  0.32  nonhaartcombo.W1 -20.99 888.74
unstabl.W1      -0.09  0.37   b05.W1          -0.07 0.12
ethnic.W3       -0.14  0.23   h39b.W2          0.02 0.03
age.W2           0.02  0.02   pcs.W2          -0.01 0.02
mcs37.W2         0.26  0.31   haart.W2         1.80 0.30
nonhaartcombo.W2 1.33  0.44   unstabl.W2       0.27 0.42
b05.W2           0.03  0.12   h39b.W3          0.00 0.03
age.W3          -0.01  0.02   pcs.W3           0.01 0.01
mcs37.W3        -0.04  0.32   haart.W3         0.60 0.31
nonhaartcombo.W3 0.44  0.42   unstabl.W3      -0.92 0.40
b05.W3          -0.11  0.11
n = 508, k = 25
residual deviance = 366.4, null deviance = 700.1 (difference = 333.7)

# from bayesglm (Cauchy priors, scale 10 for const and 2.5 for other coefs):
      coef.est coef.sd          coef.est coef.sd
(Intercept)     -0.84  1.15   h39b.W1          -0.08 0.03
age.W1           0.01  0.02   pcs.W1          -0.01 0.01
mcs37.W1        -0.10  0.31  nonhaartcombo.W1 -6.74 1.22
unstabl.W1      -0.06  0.36   b05.W1          0.02 0.12
ethnic.W3       0.18  0.21   h39b.W2          0.01 0.03
age.W2           0.03  0.02   pcs.W2          -0.02 0.02
mcs37.W2         0.19  0.31   haart.W2         1.50 0.29
nonhaartcombo.W2 0.81  0.42   unstabl.W2       0.29 0.41
b05.W2           0.11  0.12   h39b.W3          -0.01 0.03
age.W3          -0.02  0.02   pcs.W3           0.01 0.01
mcs37.W3         0.05  0.32   haart.W3         1.02 0.29
nonhaartcombo.W3 0.64  0.40   unstabl.W3      -0.52 0.39
b05.W3          -0.15  0.13

```

Figure 4: A logistic regression fit for missing-data imputation using maximum likelihood (top) and Bayesian inference with default prior distribution (bottom). The classical fit resulted in an error message indicating separation; in contrast, the Bayes fit (using independent Cauchy prior distributions with mean 0 and standard deviation 10 for the intercept and 2.5 for the other coefficients) produced stable estimates. We would not usually summarize results using this sort of table; however this gives a sense of how the fitted models look on the computer console.

For one example from our analysis, separation arose when estimating, for each HIV-positive persons in the sample, the probability of attendance in a group therapy called `haart`. The top part of Figure 4 shows the model as estimated using the `glm` function in R fit to the observed cases in the first year of the dataset: the coefficient for `nonhaartcombo.W1` is essentially infinity, and the regression also gives an error message indicating nonidentifiability. The bottom part of Figure 4 shows the fit using our recommended Bayesian procedure (this time, for simplicity, not recentering and rescaling the inputs, most of which are actually binary).

In the chained imputation procedure, the classical `glm` fits were nonidentifiable at many places, none of which presented any problem when we switched to `bayesglm`. We also tried the `brlr` function in R, which implements the Jeffreys prior distribution of Firth (1993). Unfortunately, we still encountered problems in achieving convergence and obtaining reasonable answers, several times obtaining an error message indicating nonconvergence of the optimization algorithm. We suspect this problem arises because `brlr` uses a general-purpose optimization algorithm that, when fitting regression models, is less stable than iteratively weighted least squares.

4.4 Complete logical separation

An example where the true β is ∞ because x and y are logically related . . .

Explore using simulation with $n = 10$ and $n = 100$

4.5 Hierarchical model of food consumption and cancer

Fang Fang example (include Rafael as coauthor). Cite Greenland papers. Will it be ok with Ezra, etc., to include this example (if we have minimal details and focus on the method, not the results)?

5 Data from a large number of logistic regressions

In the spirit of Stigler (1977), we wanted to see how large are logistic regression coefficients in some general population, to get a rough sense of what would be a reasonable default prior distribution. One way to do this is to fit many logistic regressions to available datasets and estimate the underlying distribution of coefficients.

Figure 5a shows the result of fitting separate logistic regressions to the hundreds of datasets from the `**` archive; each of these had typically dozens of binary predictors, yielding a total of `xxxx` estimated coefficients. (We excluded the intercepts from this analysis.) The

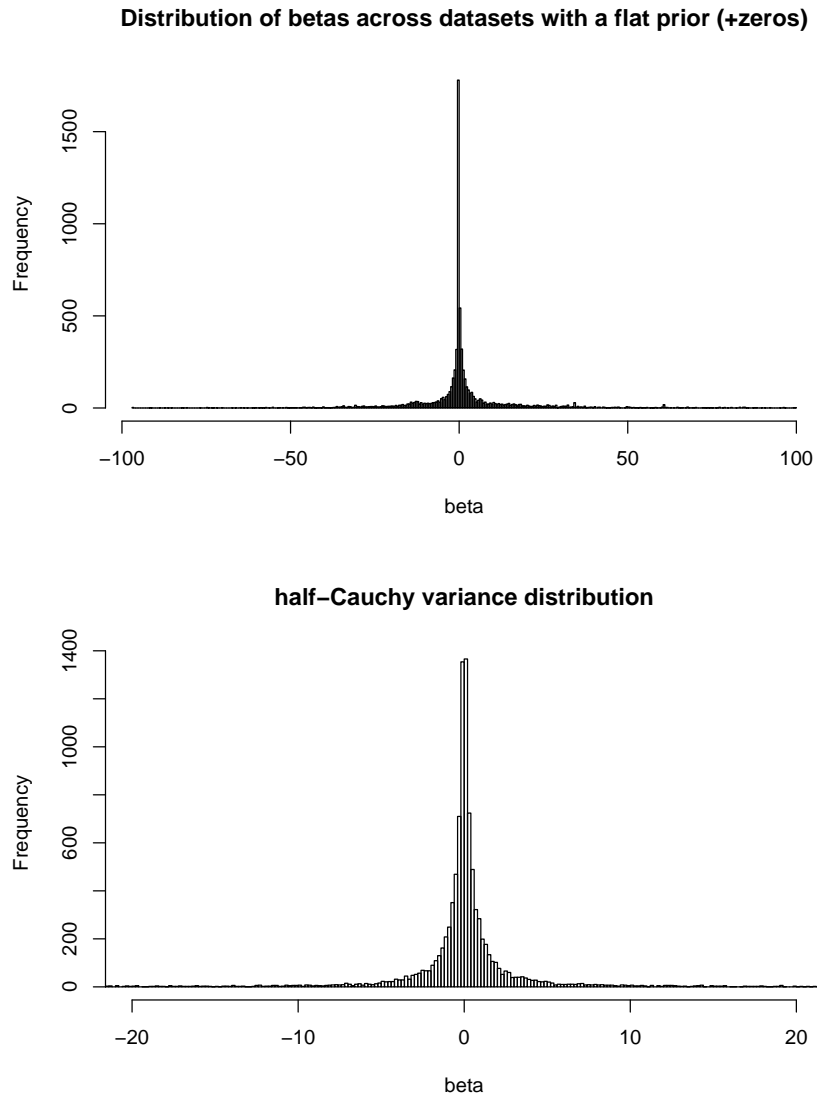


Figure 5: Distribution of thousands of estimated logistic regression coefficients as fitted to hundreds of examples, each with dozens of binary predictors. (a) Histogram of raw estimates, (b) Histogram of random posterior draws obtained by assuming an underlying Cauchy distribution of parameter values and estimating its scale using a simple hierarchical Bayes computation. The two graphs are on different scales. Data come from the ** archive.

distribution is sharply peaked about zero but with long tails. However, these are raw estimates. We can get a better sense of the distribution by shrinking these toward a model with hyperparameters estimated from the data (that is, empirical Bayes). For simplicity, we do this here by implementing one step of a Gibbs sampler ... [Aleks will supply more details here]. The result appears in Figure 5b. We are unsurprised to see that the vast majority of the coefficients are below 5 in absolute value. Comparison to Figure 1 suggests that our Cauchy distribution with scale 2.5 may in fact be overly conservative, relative to the possible logistic regression coefficients that might be encountered in real data. Given the current default (maximum likelihood) has no shrinkage at all, however, it seems to make sense to be conservative in our prior distribution.

6 Discussion

We recommend using, as a default prior model, independent Cauchy distributions on all logistic regression coefficients, each centered at 0 and with scale parameter 10 for the constant term and 2.5 for all other coefficients. Before fitting this model, we center each binary input to have mean 0 and rescale each numeric input to have mean 0 and standard deviation 0.5. When applying this procedure to classical logistic regression, we fit the model using an adaptation of the standard iteratively weighted least squares computation, using the posterior mode as a point estimate and the curvature of the log-posterior density to get standard errors. More generally, the prior distribution can be used as part of a fully Bayesian computation in more complex settings such as hierarchical models.

6.1 Other generalized linear models

6.2 Related work

Our key idea is to use minimal prior knowledge, specifically that a typical change in an input variable would be unlikely to correspond to a change as large as 10 on the logistic scale (which would move the probability from 0.01 to 0.99). This is related to the method of Bedrick, Christensen, and Johnson (1996) of setting a prior distribution by eliciting the possible distribution of outcomes given different combinations of regression inputs, and the method of Witte, Greenland, and Kim (1998) and Greenland (2001) of assigning prior distributions by characterizing expected effects in weakly informative ranges (“probably near null,” “probably moderately positive,” and so on). Our method differs from these related approaches in being more of a generic prior constraint rather than information

specific to a particular analysis. As such, we would expect our prior distribution to be more appropriate for automatic use, with these other methods suggesting ways to add more targeted prior information when necessary. One approach for going further, discussed by MacLehose et al. (2006) and Dunson, Herring, and Engel (2006), is to use mixture prior distributions for logistic regressions with large numbers of predictors. These models use batching in the parameters, or attempt to discover such batching, in order to identify more important predictors and shrink others.

Another area of related work is the choice of parametric family for the prior distribution. We have chosen the t family, focusing on the Cauchy as a conservative choice. Genkin, Lewis, and Madigan (2006) consider the Laplace (double-exponential) distribution, which has the property that its posterior mode estimates can be shrunk all the way to zero. This is an appropriate goal in projects such as text categorization (the application in that article) in which data storage is an issue, but less relevant in social science analysis of data that have already been collected.

This paper has focused on logistic regression, but the same idea could be used for other generalized linear models. For Poisson regression and other models with the logarithmic link, again, we would not expect effects larger than 10 on the logarithmic scale, and so the prior distributions given here would seem like a reasonable default choice. For linear regression, the scale of the outcome is arbitrary, so we would preprocess by rescaling the outcome variable to have mean 0 and standard deviation 0.5 before applying the default prior distributions.

In the other direction, our approach (which, in the simplest logistic regression that includes only a constant term, is close to adding one-half success and one-half failure; see Figure 1) can be seen as a generalization of the work of Agresti and Coull (1988) on using Bayesian techniques to get point estimates and confidence intervals with good small-sample frequency properties. As we have noted earlier, similar penalized likelihood methods using the Jeffreys prior have been proposed by Firth (1993), Heinze and Schemper (2003), and Zorn (2005); Heinze (2006) evaluates the frequency properties of estimates and tests using method. Our approach is similar but is parameterized in terms of the coefficients and thus allows us to make use of prior knowledge on that scale. In simple cases the two methods can give similar results (for example, identical to the first decimal place in the example in Figure 3).

6.3 Concerns

A theoretical concern is that our prior distribution is defined on centered and scaled input variables; thus it implicitly depends on the data. As more data arrive, the linear transformations used in the centering and scaling will change, thus changing the implied prior distribution as defined on the original scale of the data. A natural extension here would be to formally make the procedure hierarchical, for example defining the j -th input variable X_{ij} as having a population mean μ_j and standard deviation σ_j , then defining the prior distributions for the corresponding predictors in terms of scaled inputs of the form $Z_{ij} = (X_{ij} - \mu_j)/(2\sigma_j)$. We did not go this route, however, because modeling all the input variables corresponds to a potentially immense effort which is contrary to the spirit of this method, which is to be a quick automatic solution. In practice, we do not see the dependence of our prior distribution on data as a major concern, although we imagine it could cause difficulties when sample sizes are very small.

Modeling the coefficient of a scaled variable is analogous to parameterizing a simple regression through the correlation, which depends on the distribution of x as well as the regression of y on x . Changing the values of x can change the correlation, and thus the implicit prior distribution, even though the regression is not changing at all (assuming an underlying linear relationship). That said, this is the cost of having an informative prior distribution: some scale must be used, and the scale of the data seems like a reasonable default choice.

Finally, one might argue that the Bayesian procedure, by always giving an estimate, obscures nonidentifiability and could lead the user into a false sense of security. To this objection we would reply (following Zorn, 2005): first, one is always free to also fit using maximum likelihood, and second, separation corresponds to information in the data, which is ignored if the offending predictor is removed and awkward to handle if it is included with an infinite coefficient (see, for example, the estimates for 1964 in the first column of Figure 2). Given that we do not expect to see effects as large as 10 on the logistic scale, it is appropriate to use this information.

References

- Agresti, A., and Coull, B. A. (1998). Approximate is better than exact for interval estimation of binomial proportions. *American Statistician* **52**, 119–126.
- Albert, A., and Anderson, J. A. (1984). On the existence of maximum likelihood estimates

- in logistic regression models. *Biometrika* **71**, 1–10.
- Bedrick, E. J., Christensen, R., and Johnson, W. (1996). A new perspective on priors for generalized linear models. *Journal of the American Statistical Association* **91**, 1450–1460.
- Carlin, B. P., and Louis, T. A. (2001). *Bayes and Empirical Bayes Methods for Data Analysis*, second edition. London: CRC Press.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* **39**, 1–38.
- Dunson, D. B., Herring, A. H., and Engel, S. M. (2006). Bayesian selection and clustering of polymorphisms in functionally-related genes. *Journal of the American Statistical Association*, under revision.
- Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, second edition. London: CRC Press.
- Gelman, A. (2006a). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 514–534.
- Gelman, A. (2006b). Scaling regression inputs by dividing by two standard deviations. Technical report, Department of Statistics, Columbia University.
- Gelman, A., and Pardoe, I. (2007). Average predictive comparisons for models with non-linearity, interactions, and variance components. *Sociological Methodology*.
- Genkin, A., Lewis, D. D., and Madigan, D. (2006). Large-scale Bayesian logistic regression for text categorization. Technical report, Department of Statistics, Rutgers University.
- Greenland, S. (2001). Putting background information about relative risks into conjugate prior distributions. *Biometrics* **57**, 663–670.
- Heinze, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in Medicine*, in press.
- Heinze, G., and Schemper, M. (2003). A solution to the problem of separation in logistic regression. *Statistics in Medicine* **12**, 2409–2419.
- Lesaffre, E., and Albert, A. (1989). Partial separation in logistic discrimination. *Journal of the Royal Statistical Society B* **51**, 109–116.
- Liu, C., Rubin, D. B., and Wu, Y. N. (1998). Parameter expansion to accelerate EM: the

- PX-EM algorithm. *Biometrika* **85**, 755–770.
- Liu, C. (2004). Robit regression: a simple robust alternative to logistic and probit regression. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, ed. A. Gelman and X. L. Meng, 227–238. London: Wiley.
- MacLehose, R. F., Dunson, D. B., Herring, A. H., and Hoppin, J. A. (2006). Bayesian methods for highly correlated exposure data. *Epidemiology*, under revision.
- Martin, A. D., and Quinn, K. M. (2002b). MCMCpack.
scythe.wustl.edu/mcmcpack.html
- McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*, second edition. London: Chapman and Hall.
- Racine, A., Grieve, A. P., Fluhler, H., and Smith, A. F. M. (1986). Bayesian methods in practice: experiences in the pharmaceutical industry (with discussion). *Applied Statistics* **35**, 93–150.
- Raghunathan, T. E., Van Hoewyk, J., and Solenberger, P. W. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* **27**, 85–95.
- Rubin, D. B. (1978). Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse (with discussion). *Proceedings of the American Statistical Association, Survey Research Methods Section*, 20–34.
- Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association* **91**, 473–520.
- Stigler, S. M. (1977). Do robust estimators work with real data? *Annals of Statistics* **5**, 1055–1098.
- Van Buuren, S., and Oudshoorn, C. G. M. (2000). MICE: Multivariate imputation by chained equations (S software for missing-data imputation).
web.inter.nl.net/users/S.van.Buuren/mi/
- van Dyk, D. A., and Meng, X. L. (2001). The art of data augmentation (with discussion). *Journal of Computational and Graphical Statistics* **10**, 1–111.
- Witte, J. S., Greenland, S., Kim, L. L. (1998). Software for hierarchical modeling of epidemiologic data. *Epidemiology* **9**, 563–566.
- Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis* **13**, 157–170.