A default prior distribution for logistic and other regression models[*]

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

September 12, 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 algorithm. We illustrate with several examples, including a series of logistic regressions predicting voting preferences and an imputation model for a public health dataset.

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 models, least squares, linear regression, logistic regression, 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 and David Dunson for helpful comments 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, recommended by is to use Bayesian prior distributions, as recommended by . . . is to use Bayesian inference. Various prior distributions have been suggested for this purpose, most notably a Jeffreys prior distribution (Firth, 1993), but these do not supply enough information to ensure stable estimation. Here we propose a new, proper prior distribution that is still vague enough to be used as a default in routine applied work.

2 The model

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.

The model proceeds in two steps. We perform a standardization (Gelman, 2006):

- Binary input variables 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.)

- 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

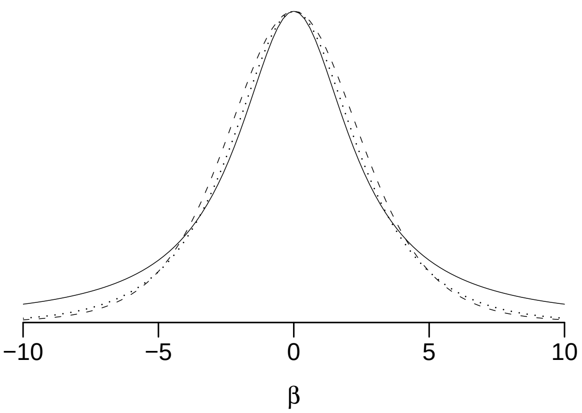

Figure 1: (solid line) Cauchy density function with scale 2.5, (dashed line) t_7 density function with scale 2.5, (dotted line) likelihood for θ corresponding to a single binomial trial of probability $\text{logit}^{-1}(\theta)$ with one-half success and one-half failure. All these curves favor values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

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 × sex), but just two inputs: age and sex. It is the input variables, not the predictors, that are standardized.

The second step of the model is to define prior distributions for the coefficients of the predictors. We use independent Student-t prior distributions 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 very close to a t density function with 7 degrees of freedom and scale 2.5. We will 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 then 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. (That is, 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 Metropolis algorithm, as is now standard with Bayesian models (see, for example, Carlin and Louis, 2001, 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:

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

2. 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").

3. 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 Oudshoom, 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.

Incorporating the prior distribution into classical logistic regression computations

In the calculation of point estimates and standard errors, the usual logistic regression algorithm 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 of β (see, for example, McCullagh and Nelder, 1989). A normal prior distribution for β can be effortlessly included in this algorithm by simply altering the least-squares step, augmenting the approximate likelihood with the information from the prior distribution.

With a t prior distribution, we can program a similar procedure, using the $\beta_k \sim \mathrm{N}(0, \sigma_k^2)$ formulation and averaging over the σ_k's at each step, treating them as missing data and performing one step of the EM algorithm. We initialize the algorithm by setting each σ_k to the value s (the scale of the prior distribution). This allows us to perform a step of the logistic regression algorithm and obtain an (approximate) estimate for each β_k; call these $\hat{\beta}_k$. Then at each step of the algorithm, the vector β is updated by maximizing the expected value of the (approximate) log-posterior density, which is simply the approximate log-likelihood (as computed using the standard generalized linear model computation), plus a sum of terms of the form, $-\frac{1}{2\sigma_k^2}\beta_k^2$:

- E-step: To average over the σ_k's in the EM algorithm, we replace $1/\sigma_k^2$ in this expression by $\mathrm{E}(1/\sigma_k^2)$, with the expectation evaluated conditional on the most recent iteration of β. From the χ^2 distribution, we get $\mathrm{E}(1/\sigma_k^2) = (\nu + 1)/(\nu s^2 + \hat{\beta}_k^2)$.

- Approximate M-step: We simply augment the data matrix by including a pseudo-data point for each coefficient k with mean 0 and variance $(\nu s^2 + \beta_k^2)/(\nu - 1)$. Performing least-squares on the augmented data matrix yields a new estimated vector of coefficients, $\hat{\beta}$, which we can then use for the next E and M step.

We have implemented these computations by altering the `glm` function in R, creating a new function, `bayesglm`, which finds the posterior mode using the above iteration. 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. Furthermore, the `standardize` function in R automatically rescales regression inputs by centering and dividing by two standard deviations (Gelman, 2006), and so using these two functions together performs our recommended procedure automatically.

Full Bayesian computation

In the full Bayesian setting, we can throw in the prior distribution and then perform the Metropolis algorithm as before, simply using the new posterior distribution. Another option is to expand the t model if a regression coefficient β_k has a t_ν distribution with mean 0 and scale s, we can break this into two distributions: $\beta_k \sim \mathrm{N}(0, \sigma_k^2)$ and $\sigma_k^2 \sim \text{Inv-}\chi^2(\nu, s^2)$. The hyperparameters ν and s here are specified in the model, and the Gibbs-Metropolis steps can be expanded to update the scale parameters σ_k as well as the coefficients β_k. Depending on how the original logistic regression has been programmed, this expansion can be computationally efficient.

4 Examples

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 the estimated coefficients for logistic regression predicting probability of Republican vote for President given sex, race, and income, as fit separately to data from the National Election Study from 1952, 1956, ..., 2000. (We have followed our general procedure of centering the binary inputs (`female` and `black`) to have means of 0, and rescaling the numerical input (`income`) to have mean 0 and standard deviation 0.5.) 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

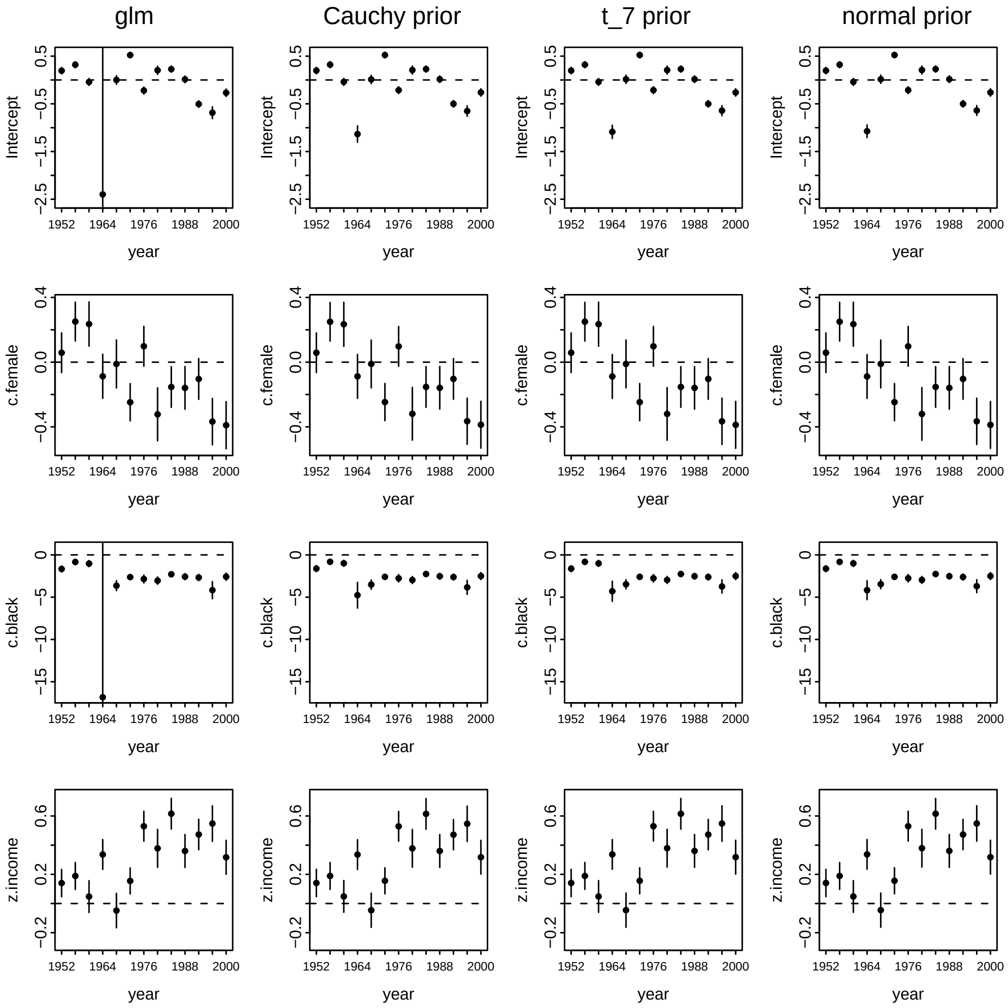

Figure 2: The left column shows the estimated coefficients (±1 standard error) for a logistic regression predicting probability of Republican vote for President given sex, race, and income, as fit separately to data from the National Election Study from 1952, 1956, . . . , 2000. (The binary inputs `female` 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, are determined by the number of iterations used by the `glm` function in R.)
(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	Number of	Number of
(log g/ml)	animals, n_i	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 reflecion we prefer the smaller estimate, which is based on downweighting the most extreme possibilities that are allowed by the likelihood.

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.

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 coefficient of standardized dose from 10.2 ± 6.4 to 5.4 ± 2.2. This might seem disturbing, but for the reasons discussed above, we would doubt the effect to be as large as 10 on the logistic scale, and the analysis shows these data to be consistent with the much smaller effect size of 5. The large amount of shrinkage simply confirms how weak the information is that gave the maximum likelihood estimate of 10.

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 Oudshoom (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 (Messeri 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.

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 nonidenti-

```
# 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 constrast, 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.

fiability. 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 geneneral-purpose optimization algorithm that, when fitting regression models, is less stable than iteratively weighted least squares.

5 Data from a large number of logistic regressions

In the spirit of Stigler (1977), we wanted to see how large are the 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 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. This 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.

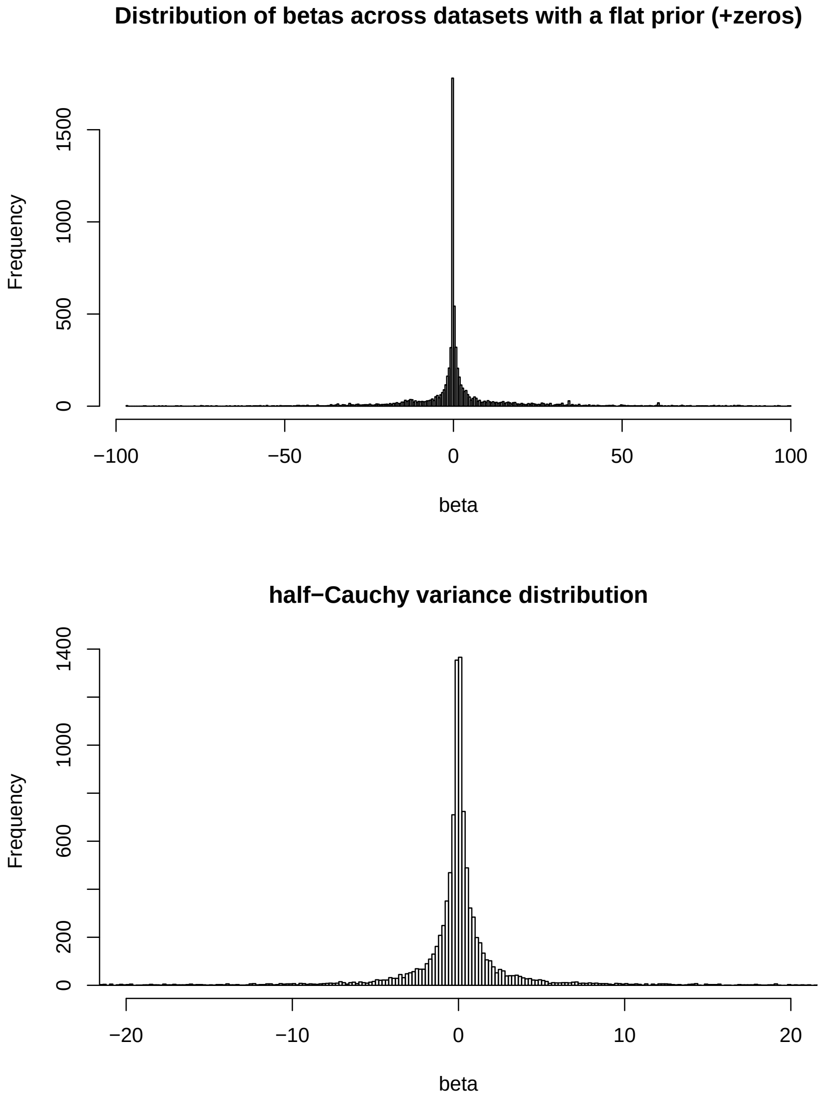

Figure 5: Distribution of thousands of estimated logistic regression coefficients as fitted to hundreds of examples, each with dozens of binary predictors. (a) Histogram of raw estimates, (b) Histogram of random posterior draws obtained by assuming an underlying Cauchy distribution of parameter values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

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.

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.

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).

Concerns

A theoretical concern is that our prior distribution is improper: being defined on centered and scaled input variables, the model 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 k-th input variable X_{ik} as having a population mean μ_k and standard deviation σ_k, then defining the prior distributions for the corresponding predictors in terms of scaled inputs of the form $Z_{ik} = (X_{ik} - \mu_k)/(2\sigma_k)$. 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 impropriety of our prior distribution as a major concern, although we imagine it could cause difficulties when sample sizes are very small.

The situation of 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.

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. (2006). 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 nonlinearity, interactions, and variance components. *Sociological Methodology.*

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. (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.

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 Oudshoom, 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/`

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.