Bayesian Statistics Homework 1

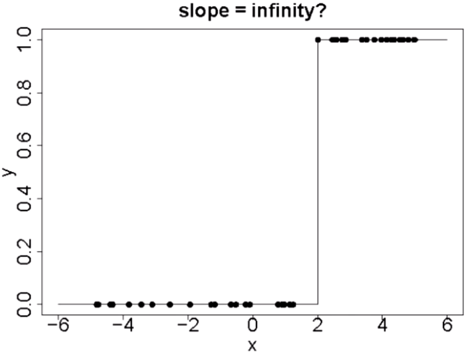
Gelman et al., *A Weakly Informative Default Prior Distribution For Logistic and Other Regression Models* (2008)

Roberto Corti

April 6, 2021

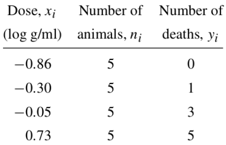
Logistic regression: we have some problems...

In some cases, Maximum Likelihood estimates of the coefficients can lead to unstable results.

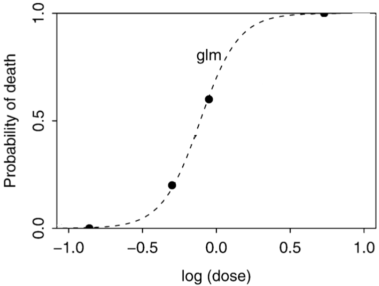


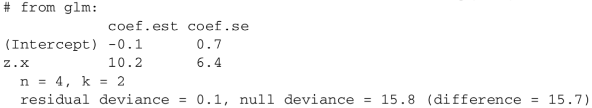
Logistic regression: we have some problems...

In some cases, Maximum Likelihood estimates of the coefficients can lead to unstable results.



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)
```

The need of (weak) prior information

Adding a prior distribution over regression's coefficient can regularize unstable results.

Which kind of prior do we use?

- Non-informative prior
- Fully informative prior

The aim is to provide *weakly informative priors*

- able to give regularized coefficient estimates
- more appropriate for automatic use

General assumptions for weakly-informative priors

What prior information can be assumed for a generic model?

"For logistic regression, a change of 5 in the logistic scale moves a probability from 0.01 to 0.5, or from 0.5 to 0.99"

Given the logistic regression coefficients $\beta = (\beta_0, \beta_1, ..., \beta_J)$
if $\beta_i \simeq 5$, for $x = (x_1, ..., x_J) \to x' = (x_1, ..., x_i + 1, ..., x_J)$, then:

$$P(y_i = 1|x) = 0.01 \to P(y_i' = 1|x') \simeq 0.5$$

We rarely encounter this kind of situations, thus we are willing to assign for $\pi(\beta_i)$ low probabilities for values outside $[-5, 5]$

General assumptions for weakly-informative priors

In order to have a default prior that could be used in many different contexts, it has to be defined over a scale-independent range.

It is required to *standardize* the input variables:

- For binary varibales, inputs must have 0 mean and they have to differ 1 in their lower and upper values.
- Other variables must have 0 mean and standard deviation of 0.5.

General assumptions for weakly-informative priors

Which probability distribution can be used for this prior?

- prior independence: $\pi(\beta) = \pi(\beta_0) \cdot \pi(\beta_1) \cdot ... \cdot \pi(\beta_J)$
- t-Student family: flat-tailed distributions, easy and stable computations.

Typical choices can be:

- for the coefficient terms $\beta_1, ..., \beta_J$, model as a likelihood of a binomial trial with half success and half failure: $\pi(\beta_i) = t_{\nu=1}(\theta = 0, s = 2.5)$ (Cauchy distribution).

- for the costant term β_0, allow a success probability between 10^{-9} and $1 - 10^{-9}$ for units that are average in all the inputs: $\pi(\beta_0) = t_{\nu=1}(\theta = 0, s = 10)$ (Cauchy distribution).

A tool to use weakly-informative priors in R

We can now combine our weakly prior information with the likelihood through the posterior distribution:

$$\pi(\beta|X, y) \propto p(y|X, \beta)\pi(\beta)$$

We are often interested in having point estimates of the posterior distribution →**maximum a posteriori (MAP) estimate**

`bayesglm`: based on `glm`, it allows the specification prior distributions for the coefficients in the `t` family,

A tool to use weakly-informative priors in R

`bayesglm`

- **Goal**: get MAP estimates β_{MAP} and $V_{\beta_{MAP}}$
- **Computation**: Considering the prior distribution for each coefficient as a mixture of normals with unknown scale σ_i

$$\pi(\beta_i) = \pi(\beta_i|\sigma_i)\pi(\sigma_i) = \mathcal{N}(\mu_i, \sigma_i^2)\text{inv-}\chi^2(\nu_i^2, s_i^2)$$

estimates of β_{MAP} and $V_{\beta_{MAP}}$ are obtained in a iterative way:

- for a fixed σ, approximate the priors as $\pi(\beta_i) \approx \mathcal{N}(\mu_i, \sigma_i^2)$ and get $\hat{\beta}$ by using weighted least squares.
- determine the expected value of the log-posterior density $\log p(\beta, \sigma|X, y)$ and maximize it with respect to σ in order to get to get the estimate $\hat{\sigma}$

Applications of `bayesglm`

1. A model predicting voting from demographic predictors,
2. A simple bioassay model from an early article [Racine et al. (1986)] on routine applied Bayesian inference,
3. A missing-data imputation problem from our current applied work on a study of HIV virus load.
4. Evaluation of the predictive performance by using a variety of prior distributions on 45 datasets from the UCI Machine Learning data repository [Newman et al. (1998), Asuncion and Newman (2007)].

Results of `bayesglm` (1)

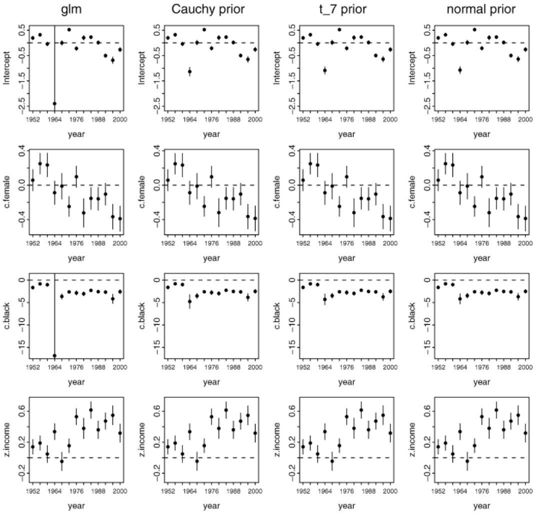
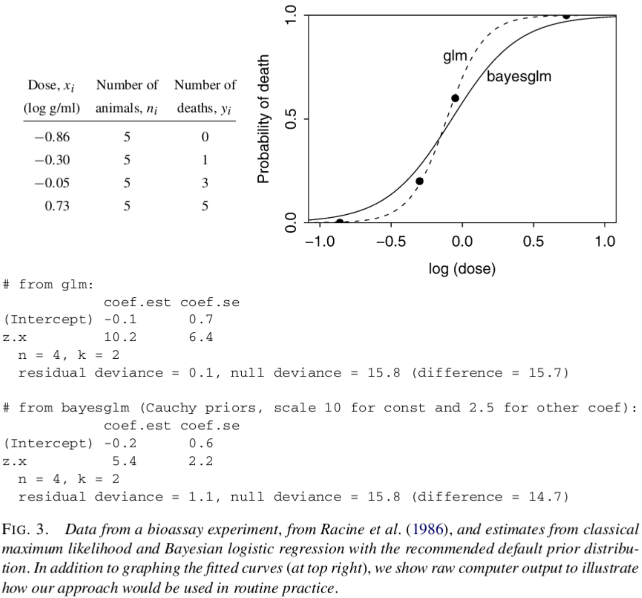


FIG. 2. *The left column shows the estimated coefficients* (±1 *standard error*) *for a logistic regression predicting the probability of a Republican vote for president given sex, race, and income, as fit separately to data from the National Election Study for each election* 1952 *through* 2000. [*The 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.*]

Results of `bayesglm` (2)



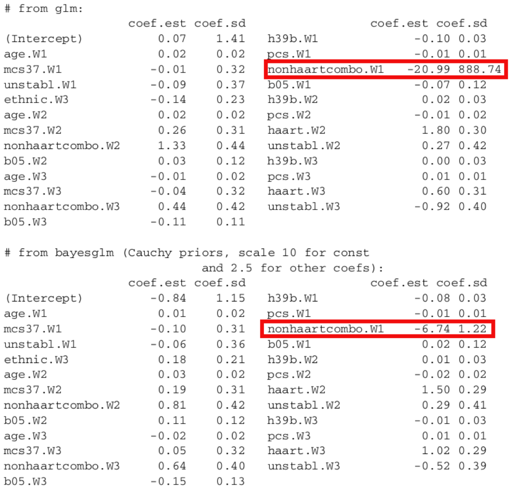
Dose, x_i (log g/ml)	Number of animals, n_i	Number of deaths, y_i
−0.86	5	0
−0.30	5	1
−0.05	5	3
0.73	5	5

```
# 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)
```

FIG. 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. In addition to graphing the fitted curves (at top right), we show raw computer output to illustrate how our approach would be used in routine practice.*

Results of `bayesglm` (3)



```
# 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

# 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
```

FIG. 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 scale* 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 in routine data analysis.*

Results of `bayesglm` (4)

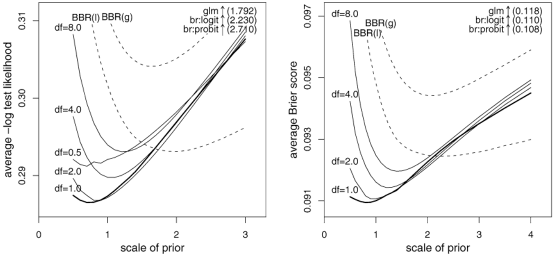


FIG. 6. *Mean logarithmic score (left plot) and Brier score (right plot), in fivefold cross-validation averaging over the data sets in the UCI corpus, for different independent prior distributions for logistic regression coefficients. Higher value on the y axis indicates a larger error. Each line represents a different degrees-of-freedom parameter for the Student-t prior family. BBR(l) indicates the Laplace prior with the BBR algorithm of Genkin, Lewis, and Madigan* (2007), *and BBR(g) represents the Gaussian prior. The Cauchy prior distribution with scale* 0.75 *performs best, while the performances of* `glm` *and* `brglm` *(shown in the upper-right corner) are so bad that we could not capture them on our scale. The scale axis corresponds to the square root of variance for the normal and the Laplace distributions.*

Results of `bayesglm` (4)

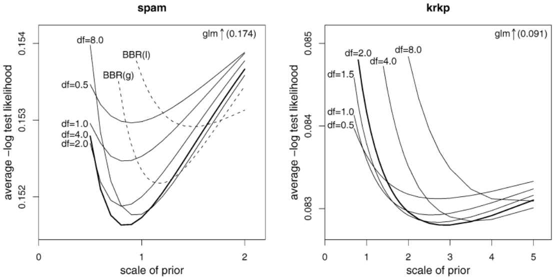


FIG. 7. *Mean logarithmic score for two datasets, "Spam" and "KRKP," from the UCI database. The curves show average cross-validated log-likelihood for estimates based on t prior distributions with different degrees of freedom and different scales. For the "spam" data, the t_4 with scale* 0.8 *is optimal, whereas for the "krkp" data, the t_2 with scale* 2.8 *performs best under cross-validation.*