SAS 1018-2017

Comparing Priors in Bayesian Logistic Regression for Sensorial Classification of Rice

Geiziane Oliveira, SAS Institute, Brazil; George von Borries, Universidade de Brasília, Brazil; Priscila Zaczuk Bassinello, Embrapa Rice and Beans, Brazil.

ABSTRACT

The present study use Proc MCMC to estimate rice stickiness (sticky or loose) in binomial logistic regression. Rice quality can be primarily assessed by evaluating its texture after cooking. The classical sensory evaluation is expensive and a time-consuming method since it requires training, capability and availability of people. Therefore, the present study investigated Bayesian binomial logistic models to replace sensory evaluation of stickiness by analyzing the relationship between sensory and principal components of viscosity measurements of rice. Proc MCMC was used to produce models based on different priors, as (1) noninformative prior; (2) default prior (Gelman et al., 2008); (3) prior based on odds ratio (Sullivan and Greenland, 2012) and (4) power priors (Ibrahim and Chen, 2000). SAS MCMC showed to be easy to implement and to compare results.

INTRODUCTION

Visual characteristics of rice grain are important in determination of quality and price of cooked rice. Stickiness is a visual characteristic of texture that is typically measured by an expensive and time consuming sensorial analysis of cooked rice. Here, PROC MCMC is used to automatically estimate rice stickiness (sticky or loose) with a Bayesian binomial logistic regression model applied to linear combinations of five viscosity measurements of cooked rice. The priors used in the Bayesian modelling were based on four different suggestions of literature: (1) a noninformative prior; (2) a default prior (Gelman et al., 2008); (3) a prior based on odds ratio (Sullivan and Greenland, 2012); and (4) power priors (Ibrahim and Chen, 2000). Then, the quality of each adjustment is evaluated by comparison of agreement between sensorial and predictive (model) classification of stickiness.

MATERIAL AND METHODS

Stickiness is modeled as a response Y_i with a binomial distribution,

$$y_i | \, p_i \sim binomial(n_i, p_i),$$

where p_i is the success probability (sticky) and links to regression covariates C_1, C_2 through a logit transformation

$$logit\,(p_i) = log\left(\frac{p_i}{1-p_i}\right) = \beta_0 \; + \; \beta_1 C_1 + \beta_2 C_2 \; .$$

Once a priori distribution is defined for a parameter $\beta_i, \;\; p(\beta_i)$, the posteriori distribution is found by

$$p(\beta_i|Y) = \; \frac{L(\beta_i|Y)\, p(\beta_i)}{\int L(\beta_i|Y)\, p(\beta_i) d\beta_i} \, ,$$

with $L(\beta_i|Y)$ the likelihood of β_i. The priors on β_0, β_1 and β_2 were defined according to four different prior distributions suggested in literature:

Noninformative Prior: A noninformative prior distribution is proportional to the likelihood function, resulting in low (or none) impact in parameters of the posteriori distribution. One common prior distribution is given by the uniform distribution.

Informative Prior A: Sullivan and Greeland (2012) suggest a Bayesian analysis by data augmentation justifying that a poorly chosen prior distribution may degrade the performance of inferential procedures. The suggested process for constructing the prior data is the creation of a normal prior for each β based on its corresponding odds ratio (Kleinbaum and Klein, 2010) obtained from original data through classical logistic regression (maximum likelihood estimation) using PROC LOGISTIC (SAS/STAT®, 2013). The normal prior has centre, β_{Prior}, as

$$\beta_{prior} = \ln(OR_{prior}) = \frac{\ln(OR_{upper}) + \ln(OR_{lower})}{2}$$

where $OR = e^{\beta}$, and (OR_{lower}, OR_{upper}) are the respective 95% confidence limits. The variance v_{prior} of the normal prior is obtained by

$$v_{prior} = \left[\frac{\ln(OR_{upper}) - \ln(OR_{lower})}{2 \times 1.96}\right]^2.$$

Informative Prior B: according to Gelman et al. (2014) the model should be consistent with knowledge about the underlying scientific problem and the data collection process. Gelman et al. (2008) propose independent Student-*t* prior distribution for parameters of classical logistic regression models. The priors are applied after scaling all no binary variables to have mean 0 and standard deviation 0.5. In this study we used C_1, C_2 with mean 0 and corresponding variance. Gelman et al. use a minimal prior knowledge rather than information about a particular analysis. For this purpose, they made a conservative choice and have chosen independent Cauchy with center 0 and scale 10 to the constant term (β_0) and Cauchy with center 0 and scale 2.5 to each of the coefficients (β_1, β_2) in the logistic regression.

Power Priors: Ibrahim et al. (2015) suggested priors to be constructed from historical data. If no reduction is attained with historical data, one could discard previous information and concentrate only on actual data and/or user experience. The formulation of the power prior is

$$\pi(\beta|D_0, a_0) \propto L(\beta|D_0)^{a_0} \ \ \pi_0(\beta|c_0)$$

with c_0 the hyperparameter for a inicial prior (π_0) and a_0 is a scalar parameter ($0 \leq a_0 \leq 1$) that controls the influence of the historical data on weights the historical data relative to the likelihood of the current study ($L(\beta|D_0)$). When $a_0 = 1$, the posterior distribution is based completely on historical data. Howerver, if $a_0 = 0$, the prior does not depend on historical data at all. Here the prior distribution is the informative prior B with different weights given to data collected in 2013 by Embrapa.

CLASSIFICATION ERROR

The adjusted models allow one to calculate the predictive probabilities of sensorial stickiness. The probability of a sample to receive a classification as loose is given by

$$P(Y = 1|C, D) = \int P(y = 1|C, \boldsymbol{\beta})P(\boldsymbol{\beta}|D)\, P(\boldsymbol{\beta})$$

with *C* indicating the covariates, *D* the available data and $\boldsymbol{\beta}$ the vector of parameters. The probability of a sample be classified as sticky is the complementary probability. The results were used to produce a confusion matrix that compares the classification a sensory researcher gave to a sample with the predictive classification given by a model (category with higher probability). Finally, the apparent error rate (APR) was obtained as (Johnson and Wichern, 2007),

$$APR = \frac{n_d}{n},$$

with n_d representing the number of discordances between researcher classification and model prediction and n the total number of samples. The result was used as the main indicator of goodness of fit for each model.

SAS IMPLEMENTATION

The working data set (RICECP) has 5 variables:

- SAMPLE: a numeric variable with identification of sampled cooked rice.
- YEAR: year of harvesting of rice, 2013 or 2014. In each year there are 72 samples.
- PEGAJB: sensorial classification of stickiness provided by researchers from EMBRAPA. This variable is the gold standard for measure of quality prediction of each adjusted model and has value 0 if if a sample was classified as sticky and 1 if a sample was classified as loose. For the year of 2013 has 30 samples classified as sticky and 42 as loose, and for 2014 there are 34 samples classified as sticky and 38 as loose.
- C1: explanatory variable resulted from linear combination of viscosity measures and responsible for 49.83% of total information available on original viscosity measures.
- C2: explanatory variable resulted from linear combination of viscosity measures, independent of C1 and responsible for 41.21% of total information available on original viscosity measures.

The variables C1 and C2 are scores obtained from the first two principal components obtained from viscosity measures (apparent amylose content, gelatinization temperature, peak, breakdown and final viscosity) provided by Embrapa Rice and Beans for the samples analyzed. Details about principal component analysis (PCA) and the calculus of C1 and C2 are available in Oliveira (2015) and Johnson and Wichern (2007).

The data set RICECP was divided in two data sets, one for each year, named as RICECP13 and RICECP14.

Display 1 shows the first 10 observations from RICECP data set.

SAMPLE	YEAR	PEGAJB	C1	C2
1	2013	1	0.27599	-0.52657
2	2013	1	0.09544	-0.54368
3	2013	1	0.16147	-0.62299
4	2013	1	0.36966	-0.28279
5	2013	1	0.32899	-0.39975
6	2013	1	0.31115	-0.50216
7	2013	0	-1.12638	1.21866
8	2013	1	-1.09412	0.80003
9	2013	0	-1.08571	1.31549
10	2013	1	-1.09369	0.90455

Display 1. First 10 observations from working data RICECP.

Model 1: Noninformative prior

The following statements perform Bayesian logistic regression with noninformative uniform priors:

```
ODS GRAPHICS ON;
PROC MCMC DATA=RICECP14 NBI=4000 NMC=10000 SEED=1966 NTHIN=15
 STATISTICS=SUMMARY;
  PARMS BETA0 0 BETA1 0 BETA2 0;
  PRIOR BETA0 ~ UNIFORM(-15,15);
  PRIOR BETA1 ~ UNIFORM(-15,15);
  PRIOR BETA2 ~ UNIFORM(-15,15);
  P = LOGISTIC(BETA0 + BETA1*PRIN1 + BETA2*PRIN2);
  MODEL PEGAJB ~ BINARY(P);
ODS OUTPUT PostSummaries=PARMSM1;
RUN;
ODS GRAPHICS OFF;
```

With ODS statement, PROC MCMC produces three main graphics which aid convergence diagnostic checking: (1) trace plots allow to check whether the mean of the Markov chain has stabilized and appears constant over the graph, and whether the chain has good mixing and is "dense". The plots show if the chains appear to reached their stationary distributions; (2) autocorrelation plots indicate the degree of autocorrelation for each of the posterior samples, with high autocorrelations indicating slow mixing; (3) kernel density plots estimate the posterior marginal distributions for each parameter. Figure 1 shows the diagnostic plots for parameter β_0 using priors

$$\beta_0 \sim U(-15; 15),\ \beta_1 \sim U(-15; 15),\ \beta_2 \sim U(-15; 15)$$

The plots indicate that are evidence of convergence for the chain. The plots for parameters β_1 and β_2 are similar and allow the same conclusions.

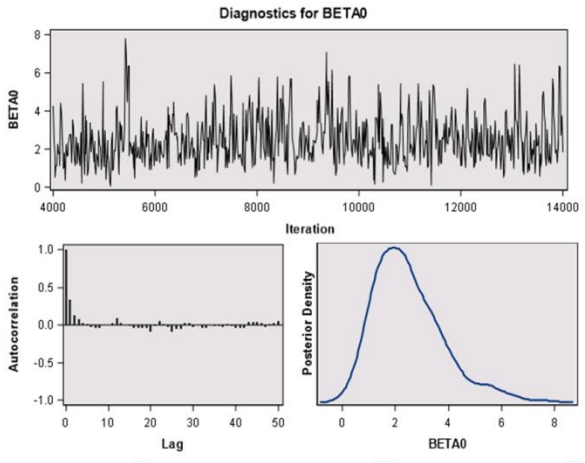

Figure 1. Diagnostic Plots for β_0.

Five statement options were used. NBI specifies the number of burn-in iterations. Note that the processing time could increase considerably if one uses a very high number for this option. NMC specifies the number of MCMC iterations after the burn-in process. SEED specifies a random seed for simulation, useful for reproducible coding. NTHIN controls the thinning rate of the simulation, keeping every *n*th simulation sample and discarding the rest. STATISTICS specifies options for posterior statistics. Summary statistics prints n, mean, standard deviation and percentiles for each parameter. PROC MCMC statement has many other options that could be checked in SAS/STAT® 14.1 (2015) user's guide.

PARMS statement declares model parameters and optional starting values. PRIOR statement specifies prior distribution for each parameter. The programming statement does the logistic transformation and assigns it to P. The MODEL statement specifies that the response variable PEGAJB has distribution of Bernoulli with probability of success P. Finally, the ODS OUTPUT option creates a data set PARMSM1 with the summary statistics from the posterior distribution.

The main objective is to predict sensory stickiness with a low apparent error rate (APR). The predicted stickiness for *i*th sample ($\hat{p}_i$) is obtained using the mean adjusted parameters in the logistic model, i.e.,

$$\hat{p}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 C_1 + \hat{\beta}_2 C_2)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 C_1 + \hat{\beta}_2 C_2)}.$$

This operation was done using macro variables to calculate the predicted values (0 = sticky; 1 = loose) and compare with original classification using PROC FREQ. The main program code is

```
* CREATIING MACRO VARIABLES FROM ESTIMATED PARAMETERS;

DATA _NULL_;
      SET PARMSM1;
      IF PARAMETER='BETA0' THEN CALL SYMPUT('B0M1',MEAN);
      IF PARAMETER='BETA1' THEN CALL SYMPUT('B1M1',MEAN);
      IF PARAMETER='BETA2' THEN CALL SYMPUT('B2M1',MEAN);
RUN;
```

```
* PREDICTION AND CLASSIFICATION;

DATA RICEPREDM1;
  SET RICECP14;
  PRED14 = EXP(&B0M1 + &B1M1 * C1 + &B2M1*C2)/(1 + EXP(&B0M1 + &B1M1 * C1 +
&B2M1*C2));
  IF PRED14 > 0.5 THEN PEGPRED = 1;
  ELSE IF PRED14 <= 0.5 THEN PEGPRED = 0;
  KEEP PRED14 PEGPRED PEGAJB;
RUN;

PROC FREQ DATA=RICEPREDM1;
  FORMAT PEGPRED PSB. PEGAJB PSB.;
  TABLES PEGPRED*PEGAJB / OUT=AERM1;
RUN;
```

With output in Display 2.

Sensorial	Predicted		
Frequency Percent	STICKY	LOOSE	Total
STICKY	33 45.83	3 4.17	36 50.00
LOOSE	1 1.39	35 48.61	36 50.00
Total	34 47.22	38 52.78	72 100.00

Display 2. Comparison of sensorial classification of stickiness with predicted using noninformative priors (Model 1).

The APR using Model 1 is about 5.56%, showing that this model could predict the sensorial classification very well. Other aspects of this model were not relevant at this moment, in this study, but could be explored.

Model 2: Informative prior of Sullivan and Greeland

The creation of a prior based on odds ratio obtained from classical logistic regression applied on original data requires the use of PROD LOGISTIC with the code

```
PROC LOGISTIC DATA= RICECP14 DESCENDING;
    MODEL PEGAJB = C1 C2;
    ODS OUTPUT OddsRatios=ODSRATIOM2(KEEP=LowerCL UpperCL);
RUN;
```

Display 3 shows partial results for the PROC LOGISTIC. The odds ratio estimates were significant only for C1 ($p < .0001$) resulting in a simpler model to adjust with only an intercept and a parameter $\boldsymbol{C_1}$ (first principal component). The prior for the intercept is normal with center 0 and large variance and the prior for $\boldsymbol{\beta_1}$has normal distribution with parameters based on the 95% confidence interval for the odds ratio, i.e., (0.007; 0.181).

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.9336	0.9454	4.1828	0.0408
C1	1	-3.3021	0.8137	16.4675	<.0001
C2	1	0.7210	0.4675	2.3789	0.1230

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
C1	0.037	0.007	0.181
C2	2.057	0.823	5.141

Display 3. Partial results from PROC LOGISTIC applied to original data.

The resulting program for the Bayesian logistic model with prior of Sullivan and Greeland has the same structure of the program for Model 1, but with a different prior.

```
PROC MCMC DATA=RICECP14 NBI=4000 NMC=10000 SEED=1966 NTHIN=15
STATISTICS=SUMMARY;
  PARMS BETA0 0  BETA1 0;
  PRIOR BETA0 ~ NORMAL(0,VAR=1000);
  PRIOR BETA1 ~ NORMAL(-3.302069, VAR=0.6621089);
  P = LOGISTIC(BETA0+BETA1*C1);
  MODEL PEGAJB ~ BINARY(P);
  ODS OUTPUT PostSummaries=PARMSM2;
RUN;
```

Even though Model 2 has fewer parameters, the APR was exactly the same as the one obtained with Model 1, i.e., same results as in Display 2.

Model 3: Informative prior of Gelman et al.

The only difference from the noninformative model is the prior distribution to be applied. It was assigned independent Cauchy with center 0 and scale 10 to the constant term (β_0) and Cauchy with center 0 and scale 2.5 to each of the coefficients (β_1, β_2) in the logistic regression. The prior specifications in PROC MCMC are

```
PRIOR BETA0 ~ CAUCHY(0,10);
PRIOR BETA1 ~ CAUCHY(0,2.5);
PRIOR BETA2 ~ CAUCHY(0,2.5);
```

We observed the same APR as previous models, indicating a very good quality of adjustment.

Model 4: Power Priors

The use historical data involves a data set with additional previous information. Information of sensorial analysis from 2013 harvested rice was included to the original data with information from 2014 harvested rice. This data was copied many times, each copy with an indication of the weight year 2013 has in the current analysis. The program code for is

```
DATA SENSRICECP;
  SET RICECP;
  DO A = 0 TO 1 BY 0.2;
     OUTPUT;
  END;
PROC SORT DATA=SENSRICECP;
  BY A;
RUN;
```

Then, using program code in PROC MCMC, it is possible to adjust a model for each defined weight defined. The new MCMC program code is

```
PROC MCMC DATA=SENSRICECP NBI=4000 NMC=10000 SEED=1966 NTHIN=15
STATISTICS=SUMMARY;
   FORMAT PEGAJB PSB.;
   BY A;
   PARMS (BETA0 BETA1 BETA2) 0;
   PRIOR BETA0 ~ CAUCHY(0,10);
   PRIOR BETA1 ~ CAUCHY(0,2.5);
   PRIOR BETA2 ~ CAUCHY(0,2.5);
   P = LOGISTIC(BETA0 + BETA1*C1 + BETA2*C2);
   LLIKE = LOGPDF('BERNOULLI',PEGAJB,P);
   IF (YEAR = 2013) THEN LLIKE = A * LLIKE;
   MODEL GENERAL(LLIKE);
   ODS OUTPUT PostSummaries=PARMSM4;
RUN;
```

Note that the same Cauchy priors used in Model 3 were used here, but this is not a requirement. The difference is the number of models MCMC run and the weight data from 2013 has in the construction of each model. The SAS programming statements define the probability to be used in the distribution for the log-likelihood function to be constructed. LOGPDF function compute the logarithm of a Bernoulli probability density with parameter P applied at PEGAJB. Then, a different data is created considering the results for each weight of the historical data (2013). The GENERAL command in the MODEL statement defines the log-prior function to be used in the model. Note that one could use the same program code

used in Model 3 to estimate parameters for each model (different A values). The alternative used here has a smaller code and is faster to compute.

DISCUSSION AND CONCLUSION

The main difference when using the four implemented priors was the obtained estimates and standard deviations of posterior densities for each parameter. The informative prior of Sullivan and Greeland (Model 2) resulted in a model with only two parameters (β_0, β_1) with lower standard error for each parameter (left plot in Figure 2). The result observed with the power prior with $a_0 = 1$, the one the posterior distribution is based completely on historical data, was the closest to the one presented by Model 2(right plot in Figure 2). If one is interested in parameter estimates and interpretation, then we recommend the use of these models. This study focused only on quality of stickiness prediction, and for this purpose all models had the same performance.

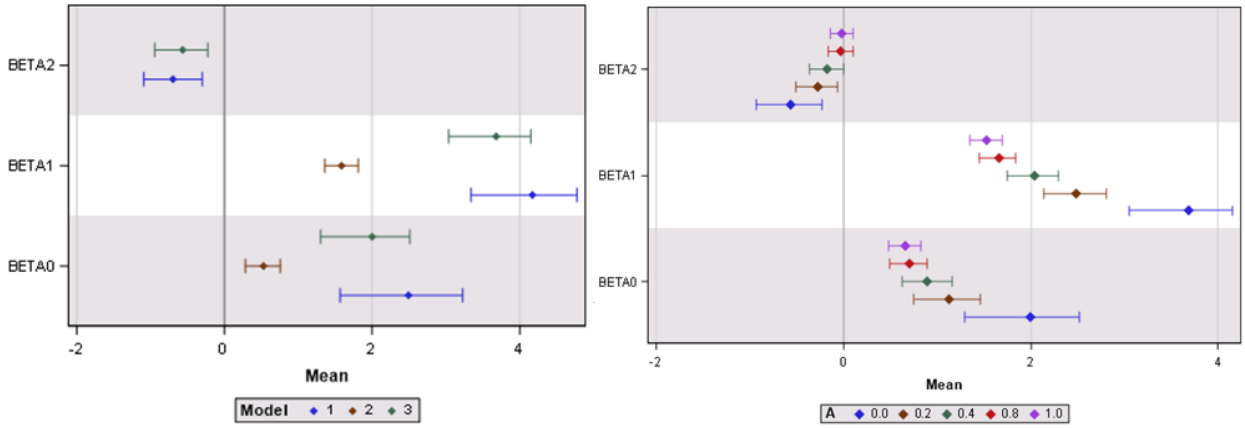

Figure 2. Left plot: Posterior Mean and Credibility Intervals (α = 5%) for parameters of models with noninformative prior (1), informative prior A (2), informative prior B (3). Right plot: : Posterior Mean and Credibility Intervals (α = 5%) for parameters of power priors models ($a_0 = A$).

Sensorial prediction/classification of Rice Stickiness is very efficient when using Bayesian Logistic Regression with PROC MCMC. The use of different priors suggested in literature produced the same quality of adjustment with difference only in the number of parameters used and size of credibility intervals. If one is interested in exploring parameters properties then we suggest the use of informative prior of Sullivan and Greeland (2012) or power prior based completely on historical data, if available. PROC MCMC is easy to use, fast to compute and very well documented.

REFERENCES

Sullivan, S.G. and Greenland, S. 2012. “Bayesian regression in SAS Software.” *Int. J. Epidemilogy*, 1-10.

Ibrahim, J.G. and Chen, M.H. 2000. “Power prior distributions for regression models.” *Statistical Science.* Vol. 15, Issue 1, 46-60.

Ibrahim, J.G.; Chen, M.H.; Gwon, Y and Chen, F. 2015. “The power prior: theory and applications.” *Statistics in Medicine*, Vol. 34, Issue 28, 3724-3749.

Gelman, A.; Jakulin, A.; Pittau, M.G. and Su, Y.S. 2008. “A weekly informative default prior distribution for logistic and other regression models.” *The Annals of Applied Statistics*, Vol. 2, Issue 4, 1360-1383.

Oliveira, G.S. (2015) *Modelos de Regressão com Resposta Ordinal para Avaliação de Textura de Arroz*. Undergraduate Monograph, Universidade de Brasília, Brazil.

SAS/STAT® 13.1 (2013). User's Guide. Cary, NC: SAS Institute Inc. Available at https://support.sas.com/documentation/onlinedoc/stat/131/logistic.pdf for PROC LOGISTIC.

SAS/STAT® 14.1 (2015). User's Guide. Cary, NC: SAS Institute Inc. Available at https://support.sas.com/documentation/onlinedoc/stat/131/mcmc.pdf for PROC MCMC or

https://support.sas.com/documentation/onlinedoc/stat/141/princomp.pdf for PROC PRINCOMP.

von Borries, G.; Bassinello, P.Z.; Rios, E. dos S., Koakuzu, S.N. and Carvalho, R.N. 2017. "Prediction Models of Rice Cooking Quality". Submitted.

Gelman, A.; Carlin, J.B.; Stern, H.S.; Dunson, D.B.; Vehtari, A. and Rubin, D.B. (2014) *Bayesian Data Analysis.* CRC Press.

ACKNOWLEDGMENTS

We are grateful to the SAS® Institute Brazil for the use of SAS through academic agreement with University of Brasília.

RECOMMENDED READING

- *SAS® / STAT 14.1 User's Guide.*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: George von Borries
Enterprise: University of Brasília
E-mail: gborries@unb.br
Web: http://georgevonborries.weebly.com/

Name: Geiziane Oliveira
Enterprise: SAS® Institute Brazil
E-mail: geiziane.oliveira@sas.com