Genomic-Enabled Prediction of Ordinal Data with Bayesian Logistic Ordinal Regression

Osval A. Montesinos-López,* Abelardo Montesinos-López,† José Crossa,‡,1 Juan Burgueño,‡ and Kent Eskridge§

*Facultad de Telemática, Universidad de Colima, C.P. 28040 Colima, Colima, México, †Departamento de Estadística, Centro de Investigación en Matemáticas (CIMAT), Guanajuato, 36240, México, ‡Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), México, Distrito Federal, México, and §University of Nebraska, Statistics Department, Lincoln, Nebraska 68583-0963

ORCID ID: 0000-0001-9429-5855 (J.C.)

ABSTRACT Most genomic-enabled prediction models developed so far assume that the response variable is continuous and normally distributed. The exception is the probit model, developed for ordered categorical phenotypes. In statistical applications, because of the easy implementation of the Bayesian probit ordinal regression (BPOR) model, Bayesian logistic ordinal regression (BLOR) is implemented rarely in the context of genomic-enabled prediction [sample size (*n*) is much smaller than the number of parameters (*p*)]. For this reason, in this paper we propose a BLOR model using the Pólya-Gamma data augmentation approach that produces a Gibbs sampler with similar full conditional distributions of the BPOR model and with the advantage that the BPOR model is a particular case of the BLOR model. We evaluated the proposed model by using simulation and two real data sets. Results indicate that our BLOR model is a good alternative for analyzing ordinal data in the context of genomic-enabled prediction with the probit or logit link.

KEYWORDS
Bayesian ordinal regression
genomic selection
probit
logit
Gibbs sampler
GenPred
shared data resource

Genomic-enabled prediction models are revolutionizing animal and plant breeding. There is some evidence that they are powerful for predicting the genomic merit of animals and plants based on high-density single-nucleotide polymorphism (SNP) marker panels and are being recommended increasingly for genomic prediction in human health (Yang and Tempelman 2012). However, most genomic-enabled prediction models assume a continuous and normally distributed phenotype. Because often this assumption is not fulfilled, researchers normally approach phenotypes in three ways: (a) they ignore the lack of normality in the phenotypes; (b) they transform the non-normal phenotype to approximate it to normality; or (c) they use generalized linear mixed models (GLMMs) to model the appropriate distribution of the phenotype (Stroup 2015).

The use of the first approach is justified for large sample sizes with the central limit theorem, which states that treatment means have an approximate normal distribution if the sample size is large enough. However, there is a lot of evidence indicating that the first approach produces highly biased results for small- and moderate-sample sizes (Stroup 2012, 2015). Transformations introduced by Bartlett (1947) for non-normal data were proposed for variance-stabilization to fulfill the assumption of homogeneous variance; they are still considered standard procedures in many agricultural disciplines. Implementation with transformations is equal to that for phenotypes normally distributed (based on the linear model). However, there is mounting evidence that transformations do more harm than good for the models required by most agricultural research, because the use of the linear model with or without transformed data produces a great loss of accuracy and power (Stroup 2015), mostly in small sample sizes.

Nelder and Wedderburn (1972) introduced generalized linear models, a major departure from the usual approach to non-normal data. GLMMs extend the linear model theory to accommodate non-normal data with heterogeneous variance and even correlated observations. Viewed through the GLMM lens, the pre-1990s understanding of non-normal data––still pervasive in the agricultural research community––is antiquated at best, obsolete at worst. Today, small sample investigations are providing an increasing body of evidence that

doi: 10.1534/g3.115.021154
Manuscript received May 13, 2015; accepted for publication August 11, 2015; published Early Online August 18, 2015.

[1] Corresponding author: Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600 México, Distrito Federal, México. E-mail: j.crossa@cgiar.org

GLMMs work as well in practice as they do in theory because they produce more accuracy and power than approaches (a) and (b). Also, now there are textbooks and software available for the implementation of GLMMs, although the implementation of approaches (a) and (b) is still dominant in agricultural research (Stroup 2015). In genomic-enabled prediction, the use of GLMMs is still in its early stages because their implementation is not straightforward given that the number of observations (n) is usually smaller than the number of covariates (p) included in the model. In addition, a complex dependence structure is observed among covariates (markers) and observations (lines) due to the joint involvement of biological processes and pathways.

To overcome this situation in the pregenomic era, Gianola (1980, 1982) and Gianola and Foulley (1983) proposed a probit (threshold) model for dealing with ordinal categorical traits in animal breeding. This probit model was extended to deal with $p >> n$ in the genomic era by González-Recio and Forni (2011) and Villanueva *et al.* (2011) for binary trials, and by Wang *et al.* (2012) and Montesinos-López *et al.* (2015) for more than two ordinal categories. Also, BGLR (Bayesian generalized linear regression), software developed for genomic-enabled prediction that is able to deal with normal, binary, ordinal, and censored data (de los Campos and Perez-Rodriguez 2013; Perez-Rodriguez and de los Campos 2014), is now available. However, no GLMMs are available for genomic-enabled prediction for counts and percentage phenotypes.

For modeling ordinal categorical phenotypes, the ordinal logistic regression model is often preferred over the ordinal probit model in statistical applications, because it provides regression coefficients that are more interpretable due to their connection to odds ratios (Zucknick and Richardson 2014). However, in genomic-enabled prediction (when $p >> n$), only the Bayesian probit model is frequently implemented, given that Bayesian methods that introduce sparseness through additional priors on the model size are very well-suited to this problem. Therefore, because of the lack of a Bayesian logistic ordinal model analogous to the Bayesian probit ordinal model that uses a data augmentation approach, the logistic model is not practical for genomic selection.

Both logistic and normal distributions are symmetric with a basic, unimodal “bell curve” shape. The only difference is that the logistic distribution has a somewhat heavier tails, which means that it is less sensitive to outlying data (and hence somewhat more robust for modeling mis-specifications or erroneous data). This is another advantage of logistic regression over probit regression. Because of its easy implementation, the use of BPOR is extremely common, even though it is less robust for modeling mis-specifications and its coefficients are less interpretable. Because of the aforementioned properties of the logistic model, some researchers have proposed approximations to logit regression. For example, Bartholomew and Knott (1999) proposed $logit(u) = k \times \Phi^{-1}(u)$, where Φ is the cumulative density function for the standard normal distribution and $k = 1.814$, whereas Camilli (1994) proposed using k = 1.702, obtained by minimizing the maximum distance between two cumulative distribution functions. Although Amemiya (1981) proposed a value of k = 1.6 and computed tables for representative values of the density function for different values of k, he did not explain why he used k = 1.6. More recently, Savalei (2006) obtained a value of k = 1.75 based on minimizing the Kullback-Leibler information. However, although some of these approximations do a reasonable job of approximating the logistic distribution, they are only approximations, and it goes without saying that an exact solution is preferred.

In this paper, we propose a Bayesian logistic ordinal regression (BLOR) model for genomic-enabled prediction by using a data augmentation approach. We illustrate our proposed method with simulation and real data. We compare the BLOR with the Bayesian probit ordinal regression (BPOR) model with and without approximation.

MATERIALS AND METHODS

Gray leaf spot (GLS) and Septoria data sets

GLS is one of the most important foliar diseases of maize worldwide. The GLS data set is composed of 278 maize lines; the ordinal trait measured in each line was GLS [1 (no disease), 2 (low infection), 3 (moderate infection), 4 (high infection), 5 (complete infection)] caused by the fungus *Cercospora zeae-maydis*, evaluated in three environments (México, Zimbabwe, and Colombia). These data are part of a data set previously analyzed by Crossa *et al.* (2011), González-Camacho *et al.* (2012), and Montesinos-López *et al.* (2015). Genotypes of all 278 lines were obtained using the 55k SNP Illumina platform. SNPs with $>$10% missing values or a minor allele frequency of $\leq$ 0.05 were excluded from the data. After line-specific quality control (applying the same quality control to each line separately), the maize data still contained 46,347 SNPs.

On the other hand, the Septoria data set contains 268 wheat lines planted in Toluca, México, in 2010, and the trait (Septoria scores) was measured using an ordinal four-point scale. Genotypes of these lines were obtained with 45,000 genotype by sequencing (GBS), following the protocol of Poland *et al.* (2012). We kept only 13,913 GBS that had $<$50% missing data; after filtering for minor allele frequency, we ended up with 6787 GBS that were used in the analysis.

For the implementation of the proposed model, we formed five data sets from these two real data sets (GLS and Septoria), four from the GLS data set and one from the Septoria data set. The first three data sets formed from GLS correspond to each environment in which they were evaluated for GLS; the last one was formed by pooling the data from the three environments (information from the three environments without taking into account the environments as covariates).

Bayesian logistic ordinal regression

Let $\boldsymbol{y} = \{y_{ij}\}$ $(i = 1, \ \ldots, I; \ j = 1, 2, \ldots, n_i)$, where i represents the genotype and j denotes the number of replicates or experimental units of each genotype. The total number of observations is $n = \sum_{i=1}^{I} n_i$. In other words, the observed vector $\boldsymbol{y_i}$ contains n_i elements, and the n-dimensional vector $\mathbf{y}$ of all responses can be written as $\boldsymbol{y}^T = (\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, \ldots, \boldsymbol{y}_I^T)$. The response variable y_{ij} represents an assignment into one of C mutually exclusive and exhaustive categories that follow an order. Therefore, the ordinal logistic regression model can be written in terms of a latent response variable l_{ij} as follows:

$$l_{ij} = \boldsymbol{x}_{\boldsymbol{ij}}^{\boldsymbol{T}} \boldsymbol{\beta} + b_i + \varepsilon_{ij} \tag{1}$$

where l_{ij} are called “liabilities”, $\varepsilon_{ij} \sim L(0,1)$, where $L(.)$ denotes the logistic distribution, and the vectors $\boldsymbol{x}_{ij}$ $(p \times 1)$ are explanatory variables associated with the fixed effects $\boldsymbol{\beta}$. The random effect $b_i \sim N(0, \sigma_b^2)$. In genomic-enabled prediction, $\boldsymbol{b} = (b_1, \ldots, b_I)^T \sim N(\mathbf{0}, \mathbf{G}\sigma_b^2)$. Since l_{ij} are unobservable and can be measured indirectly by an observable ordinal variable y_{ij}, then l_{ij} can be defined by:

$$y_{ij} = \begin{cases} 1 & if \ -\infty < l_{ij} < \gamma_1, \\ 2 & if \ \gamma_1 < l_{ij} < \gamma_2, \\ & \vdots \\ C & if \ \gamma_{C-1} < l_{ij} < \infty \end{cases}$$

This means that l_{ij} is divided by thresholds into C intervals, corresponding to C ordered categories. The first threshold, γ_1, defines the upper bound of the interval corresponding to observed outcome 1. Similarly, threshold γ_{C-1} defines the lower bound of the interval corresponding to observed outcome C. Threshold γ_c defines the boundary between the interval corresponding to observed outcomes $c - 1$ and c for (c = 1,2,…, $C - 1$). Threshold parameters are

$\boldsymbol{\gamma}^T = (\gamma_{min} < \gamma_1 < \cdots < \gamma_{C-1} < \gamma_{max})$ with $\gamma_{min} = -\infty$, and $\gamma_{max} = \infty$.

Assuming that the error term ε_{ij} of the latent response l_{ij} is distributed as $L(0,1)$, the cumulative response probability for the c category of the ordinal outcome y_{ij} is:

$$P(y_{ij} \leq c|\boldsymbol{\beta},\boldsymbol{b}) = \pi_{ij(c)} = P(l_{ij} \leq \gamma_c|\boldsymbol{\beta},\boldsymbol{b}) = \mathrm{P}\left(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + b_i + \varepsilon_{ij} \leq \gamma_c\right)$$

$$= \mathrm{P}\left(\varepsilon_{ij} \leq \gamma_c - \boldsymbol{x}_{ij}^T\boldsymbol{\beta} - b_i\right),$$

for $c = 1, 2, \ldots, C-1$.

$$= \frac{\exp(\gamma_c - \boldsymbol{x}_{ij}^T\boldsymbol{\beta} - b_i)}{1 + \exp(\gamma_c - \boldsymbol{x}_{ij}^T\boldsymbol{\beta} - b_i)} \tag{2}$$

Similarly, model (2) can be written as a cumulative logit model:

$$log\left(\frac{\pi_{ij(c)}}{1 - \pi_{ij(c)}}\right) = \gamma_c - \boldsymbol{x}_{ij}^T\boldsymbol{\beta} - b_i, \ \ for \ \ c = 1, 2, \ldots, C-1.$$

This GLMM model is described by: (1) two distributions, one for observations in the response variable $(y_{ij(1)}, y_{ij(2)}, \ldots, y_{ij(C)} |\boldsymbol{\beta}, \boldsymbol{b}) \sim \mathrm{Multinomial}\ (1, \pi_{ij(1)}, \pi_{ij(2)}, \ldots, \pi_{ij(C)})$, where $\boldsymbol{\beta}$ is the p × 1 vector of fixed effects, and another for the random effects, $b_i \sim N(0, \sigma_b^2)$ or $\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{G}\sigma_b^2)$, where b_i is the effect of line i; (2) linear predictor $\eta_{ij(c)} = \gamma_c - \boldsymbol{x}_{ij}^T\boldsymbol{\beta} - b_i$, where $\eta_{ij(c)}$ denotes the c^{th} link (c = 1,2,. . ., $C - 1$) for the fixed and random effects combination, γ_c is the intercept (threshold) for the c^{th} link, and $\boldsymbol{x}_{ij}^T$ are known row incidence vectors corresponding to fixed effects in $\boldsymbol{\beta}$. Because there are C categories, a total of $C - 1$ link functions are required to fully specify the model; and (3) link function: cumulative logit $\{\eta_{ij(c)} = log\left(\frac{\pi_{ij(c)}}{1-\pi_{ij(c)}}\right)$, ($c$ = 1,2,. . ., $C - 1$)}.

Using the inverse link for this model, we can calculate $P(y_{ij} = c\big|\boldsymbol{\beta}, \ \boldsymbol{b}) = \pi_{ij(c)}$ as follows:

$$\pi_{ij(c)} = P\left(\gamma_{c-1} < l_{ij} < \gamma_c\right)$$
$$= \frac{\exp(\boldsymbol{\gamma}_c - \boldsymbol{x}_{ij}^T\boldsymbol{\beta} - b_i)}{1 + \exp(\boldsymbol{\gamma}_c - \boldsymbol{x}_{ij}^T\boldsymbol{\beta} - b_i)} - \frac{\exp(\boldsymbol{\gamma}_{c-1} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta} - b_i)}{1 + \exp(\boldsymbol{\gamma}_{c-1} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta} - b_i)}.$$

Since we have latent variables l_{ij} distributed as $L(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + b_i, 1)$ and we observe $y_{ij} = c$ if, and only if, $\gamma_{c-1} < l_{ij} < \gamma_c$, then the joint posterior density of the parameter vector and latent variable becomes

$$P\left(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}, \sigma_\beta^2, \sigma_b^2, \boldsymbol{l}|\boldsymbol{y}\right) \propto P(\boldsymbol{y}|\boldsymbol{l}, \boldsymbol{\gamma})P(\boldsymbol{l}|\boldsymbol{\beta}, \boldsymbol{b})P(\boldsymbol{\gamma})$$
$$\times P\left(\boldsymbol{\beta}|\sigma_\beta^2\right)P\left(\boldsymbol{b}|\sigma_b^2\right)P\left(\sigma_\beta^2\right)P\left(\sigma_b^2\right).$$

Let's assume a scaled independent inverse chi-square $\chi^{-2}(\nu_b, S_b)$ prior for σ_b^2, a normal prior distribution for $\boldsymbol{\beta}|\sigma_\beta^2 \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0\sigma_\beta^2)$, a normal prior distribution for $\boldsymbol{b}|\sigma_b^2 \sim N(\boldsymbol{0}, \boldsymbol{G}\sigma_b^2)$, and also a $\chi^{-2}(\nu_\beta, S_\beta)$ prior for σ_β^2 (Gianola 2013). Following Sorensen *et al.* (1995), the prior for the $C - 1$ unknown thresholds has been given as order statistics from $\mathrm{U}(\gamma_{\min}, \ \gamma_{\max})$ distribution,

$$P(\boldsymbol{\gamma}) = (C-1)!\left(\frac{1}{\gamma_{\max} - \gamma_{\min}}\right)^{C-1} I(\boldsymbol{\gamma} \in T)$$

where $T = \{(\gamma_1, \ldots, \gamma_{\max})|\gamma_{\min} < \gamma_1 < \cdots < \gamma_{C-1} < \gamma_{\max}\}$.

The fully conditional posterior distributions are provided below and details of all derivations are given in Appendix A.

Liabilities and Pólya-Gamma values

The fully conditional posterior distribution of liability l_{ij} is a truncated normal distribution and its density is

$$P\left(l_{ij}|ELSE\right) = \frac{\phi\left(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + b_i, 1/\sqrt{\omega_{ij}}\right)}{\Phi\left(\gamma_c - \boldsymbol{x}_{ij}^T\boldsymbol{\beta} - b_i\right) - \Phi\left(\gamma_{c-1} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta} - b_i\right)} \tag{3}$$

For simplicity, *ELSE* is the data and the parameters, except the one in question. $\phi(.)$ is a normal density with parameters as indicated in the argument, Φ is the cumulative distribution function of a normal density with mean $\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + b_i$ and variance $1/\sqrt{\omega_{ij}}$, and the fully conditional posterior distribution of ω_{ij} is

$$\omega_{ij}\big|ELSE \ \sim \ PG\left(2, -l_{ij} + \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + b_i\right) \tag{4}$$

Regression coefficients ($\boldsymbol{\beta}$)

The fully conditional posterior of $\boldsymbol{\beta}$ is as follows:

$$\boldsymbol{\beta}|ELSE \ \sim \ N_p\left(\tilde{\boldsymbol{\beta}}_0, \tilde{\boldsymbol{\Sigma}}_0\right) \tag{5}$$

where $\tilde{\boldsymbol{\Sigma}}_0 = (\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2} + \boldsymbol{X}^T\boldsymbol{D}_\omega\boldsymbol{X})^{-1}$, $\tilde{\boldsymbol{\beta}}_0 = \tilde{\boldsymbol{\Sigma}}_0(\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}\boldsymbol{\beta}_0 - \boldsymbol{X}^T\boldsymbol{D}_\omega\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{X}^T\boldsymbol{D}_\omega\boldsymbol{l})$. With $\boldsymbol{l} = [\boldsymbol{l}_1^T, \ldots, \boldsymbol{l}_I^T]^T$, $\boldsymbol{l}_i = [l_{i1}^T, \ldots, l_{in_i}^T]^T$, $\boldsymbol{X} = [\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_I^T]^T$, $\boldsymbol{X}_i = [\boldsymbol{x_{i1}}, \ldots, \boldsymbol{x_{in_i}}]^T$, $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{1}_{n_1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{1}_{n_2} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{1}_{n_I} \end{bmatrix}$, $\boldsymbol{D}_\omega = diag(\boldsymbol{D}_{\omega 1}, .., \boldsymbol{D}_{\omega I})$, $\boldsymbol{D}_{\omega i} = diag(\omega_{i1}, \ldots, \omega_{in_i})$. It is important to point out that if we use a prior for $\boldsymbol{\beta} \propto$ Constant (improper uniform distribution), then in $\tilde{\boldsymbol{\Sigma}}_0$ and $\tilde{\boldsymbol{\beta}}_0$ we need to make $\boldsymbol{0}$ the term $\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}$.

Polygenic effects ($\boldsymbol{b}$)

Now the fully conditional posterior of $\boldsymbol{b}$ is given as

$$\boldsymbol{b}\big|ELSE \ \sim \ N_I\left(\tilde{\boldsymbol{b}} = \boldsymbol{F}(\boldsymbol{Z}^T\boldsymbol{D}_\omega\boldsymbol{l} - \boldsymbol{Z}^T\boldsymbol{D}_\omega\boldsymbol{X}\boldsymbol{\beta}),\right.$$
$$\left.\boldsymbol{F} = \left(\sigma_b^{-2}\boldsymbol{G}^{-1} + \boldsymbol{Z}^T\boldsymbol{D}_\omega\boldsymbol{Z}\right)^{-1}\right) \tag{6}$$

Variance of polygenic effects

Next, the fully conditional posterior of σ_b^2 is

$$\sigma_b^2|ELSE \ \sim \ \chi^{-2}(\tilde{\nu}_b = \nu_b + I, \tilde{S}_b = (\boldsymbol{b}^T\boldsymbol{G}^{-1}\boldsymbol{b} + \nu_b S_b)/\nu_b + I) \tag{7}$$

Threshold effects ($\boldsymbol{\gamma}$)

The density of the fully conditional posterior distribution of the c^{th} threshold, γ_c, is

Table 1 Simulated data set 1: Average values (Mean) and SD of MLEs and the Bayesian estimators, with four sample sizes (n_i)

			BLOR		BLOR*		MLLOR		MLLOR*	
n_i	Parameter	True Value	Mean	SD	Mean	SD	Mean	SD	Mean	SD
5	β_1	−6	**-6.141**	1.935	−6.711	2.117	−6.380	2.489	−6.826	2.668
	β_2	−5	**-4.957**	2.262	−5.546	2.726	−5.596	2.731	−5.927	2.872
	β_3	7	7.550	2.746	7.815	3.112	6.659	2.698	**7.110**	2.885
	γ_1	−0.842	**-0.851**	0.190	−0.937	0.152	−0.883	0.177	−0.942	0.185
	γ_2	−0.253	**-0.254**	0.154	−0.271	0.142	−0.277	0.167	−0.298	0.179
	γ_3	0.253	0.274	0.170	0.328	0.171	**0.262**	0.203	0.281	0.218
	γ_4	0.842	0.878	0.171	0.967	0.151	**0.863**	0.211	0.920	0.220
10	β_1	−6	−6.224	1.562	−6.534	1.650	**-6.118**	1.673	−6.480	1.767
	β_2	−5	**-4.987**	1.825	−5.433	1.901	−4.717	1.500	−5.022	1.619
	β_3	7	7.306	1.971	7.762	1.825	6.606	1.836	**7.038**	1.939
	γ_1	−0.842	**-0.843**	0.100	−0.926	0.147	−0.847	0.127	−0.907	0.135
	γ_2	−0.253	**-0.239**	0.097	−0.284	0.131	−0.273	0.110	−0.296	0.119
	γ_3	0.253	0.276	0.113	0.272	0.123	0.233	0.110	**0.249**	0.120
	γ_4	0.842	0.861	0.116	0.920	0.124	**0.841**	0.115	0.897	0.123
20	β_1	−6	−6.122	1.063	−6.278	1.390	**-6.030**	0.936	−6.422	1.017
	β_2	−5	**-5.099**	1.262	−5.538	1.103	−5.181	0.962	−5.488	1.019
	β_3	7	7.271	1.114	7.479	1.394	**7.243**	1.118	7.669	1.183
	γ_1	−0.842	−0.849	0.108	−0.917	0.100	**-0.847**	0.073	−0.905	0.077
	γ_2	−0.253	**-0.252**	0.106	−0.291	0.094	−0.250	0.069	−0.270	0.074
	γ_3	0.253	0.259	0.091	0.261	0.099	**0.254**	0.068	0.272	0.073
	γ_4	0.842	0.860	0.097	0.883	0.106	**0.850**	0.079	0.907	0.084
40	β_1	−6	**-6.058**	0.804	−6.467	0.924	−6.199	0.719	−6.563	0.783
	β_2	−5	−5.175	0.817	−5.231	0.879	−4.804	0.791	**-5.101**	0.856
	β_3	7	7.163	0.815	7.674	0.958	**7.093**	0.867	7.528	0.904
	γ_1	−0.842	−0.844	0.065	−0.911	0.069	**-0.841**	0.051	−0.899	0.055
	γ_2	−0.253	**-0.258**	0.053	−0.278	0.064	−0.248	0.05	−0.267	0.056
	γ_3	0.253	0.255	0.052	0.267	0.060	**0.252**	0.044	0.271	0.048
	γ_4	0.842	0.856	0.054	0.900	0.071	**0.835**	0.047	0.893	0.050

BLOR* and MLLOR* use the parameter estimates of BPOR and MLLOR and approximate BLOR and MLLOR with $logit(u) = 1.75\times\Phi^{-1}(u)$. The best model has the value for the parameter closer to the true value; these are presented in bold. MLEs, maximum likelihood estimators; BLOR, Bayesian logistic ordinal regression; MLLOR, maximum likelihood logistic ordinal regression.

$$P(\gamma_c|ELSE) = \frac{1}{\min\left\{\min\left(l_{ij}|y_{ij}=c+1\right), \gamma_{c+1}, \gamma_{\max}\right\} - \max\left\{\max\left(l_{ij}|y_{ij}=c\right), \gamma_{c-1}, \gamma_{\min}\right\}} \tag{8}$$

Variance of regression coefficients

The fully conditional posterior of σ_β^2 is

$$\begin{aligned}\sigma_\beta^2|ELSE &\sim \chi^{-2}(\tilde{\nu}_\beta = \nu_\beta + p, \tilde{S}_\beta \\ &= \left[(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + \nu_\beta S_\beta\right]/\nu_\beta + p)\end{aligned} \tag{9}$$

The Gibbs sampler

The Gibbs sampler is implemented by sampling repeatedly from the following loop:

1. Sample liabilities from the truncated normal distribution in (3).
2. Sample ω_{ij} values from the Pólya-Gamma distribution in (4).
3. Sample the regression coefficients from the normal distribution in (5).
4. Sample the polygenic effects from the normal distribution in (6).
5. Sample the variance effect (σ_b^2) from the scaled inverted χ^2 distribution in (7).
6. Sample the thresholds from the uniform distribution in (8).
7. Sample the variance of regression coefficients (σ_β^2) from the scaled inverted χ^2 distribution in (9).
8. Return to step 1 or terminate if chain length is adequate to meet convergence diagnostics.

In the absence of polygenic effects (**b**), the aforementioned Gibbs sampler can be used only by ignoring steps 4 and 5. If all marker effects are taken into account in the design matrix, $\boldsymbol{X}$, with a prior $\boldsymbol{\beta}\sim N_p(\mathbf{0},\ \boldsymbol{I}_p\sigma_\beta^2)$ for the beta regression coefficients, we end up with a threshold Bayesian ridge regression. This is a version for ordinal categorical data of the ridge estimator of Hoerl and Kennard (1970), since the posterior expectation of $\boldsymbol{\beta}$ is equal to $E(\boldsymbol{\beta}|ELSE) = (\boldsymbol{X}^T\boldsymbol{D}_\omega\boldsymbol{X} + \boldsymbol{I}_p\sigma_\beta^{-2})^{-1}\boldsymbol{X}^T\boldsymbol{D}_\omega\boldsymbol{l}$ with pseudo-response $\boldsymbol{l}$. Another important point is that by setting each $\omega_{ij} = 1$, the aforementioned Gibbs sampler for the BLOR with the logistic link is reduced to the Gibbs sampler for the BPOR with the probit link proposed by Albert and Chib (1993). This implies that the proposed BLOR model is more general and

Table 2 Simulated data set 2: average values (Mean) and SD of MLEs and Bayesian estimators, with four POs

PO	Parameter	True Value	BLOR		BLOR*		MLLOR		MLLOR*	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD
	β_1	−6	**−5.994**	0.824	−6.525	0.915	−6.230	0.723	−6.608	0.766
	β_2	−5	−5.088	0.880	−5.678	0.864	**−5.015**	0.763	−5.324	0.815
	β_3	7	7.183	0.802	7.607	0.893	**7.111**	0.651	7.535	0.689
5	γ_1	−0.842	−0.862	0.066	−0.923	0.067	**−0.853**	0.050	−0.910	0.053
	γ_2	−0.253	−0.263	0.071	−0.281	0.059	**−0.256**	0.048	−0.276	0.051
	γ_3	0.253	**0.258**	0.068	0.285	0.051	0.261	0.040	0.281	0.043
	γ_4	0.842	0.861	0.067	0.928	0.058	**0.856**	0.041	0.914	0.043
	β_1	−6	-6.156	0.854	−6.670	0.979	**-6.136**	0.607	−6.504	0.647
	β_2	−5	−5.359	0.899	−5.544	1.094	**-5.149**	0.610	−5.459	0.639
	β_3	7	7.349	0.830	7.739	0.856	**7.109**	0.645	7.536	0.682
10	γ_1	−0.842	−0.883	0.063	−0.925	0.076	**-0.866**	0.048	−0.924	0.050
	γ_2	−0.253	−0.266	0.070	−0.267	0.072	**-0.265**	0.052	−0.285	0.056
	γ_3	0.253	**0.260**	0.071	0.313	0.065	0.264	0.051	0.284	0.055
	γ_4	0.842	0.868	0.068	0.964	0.082	**0.861**	0.052	0.918	0.055
	β_1	−6	−6.529	0.730	−6.828	0.903	**-6.345**	0.645	−6.709	0.689
	β_2	−5	**-5.244**	0.759	−5.722	0.877	−5.275	0.629	−5.589	0.671
	β_3	7	7.525	0.784	7.860	0.758	**7.344**	0.735	7.780	0.772
20	γ_1	−0.842	−0.915	0.065	−0.972	0.060	**-0.883**	0.049	−0.942	0.053
	γ_2	−0.253	−0.275	0.059	−0.295	0.057	**-0.260**	0.044	−0.280	0.048
	γ_3	0.253	0.279	0.063	0.297	0.059	**0.271**	0.047	0.291	0.051
	γ_4	0.842	0.922	0.060	0.977	0.060	**0.895**	0.053	0.954	0.057
	β_1	−6	−6.794	0.803	−7.011	1.013	**-6.555**	0.654	−6.916	0.691
	β_2	−5	−5.652	0.754	−5.827	0.754	**-5.298**	0.634	−5.590	0.666
	β_3	7	8.065	0.894	8.351	0.881	**7.491**	0.752	7.925	0.798
30	γ_1	−0.842	−0.972	0.071	−1.004	0.075	**-0.898**	0.041	−0.956	0.044
	γ_2	−0.253	−0.301	0.060	−0.296	0.072	**-0.269**	0.044	−0.289	0.047
	γ_3	0.253	**0.279**	0.067	0.318	0.069	0.286	0.044	0.306	0.047
	γ_4	0.842	0.944	0.068	1.023	0.060	**0.922**	0.044	0.981	0.046

The outliers were generated with a student's *t* distribution with four degrees of freedom. BLOR* and MLLOR* use the parameter estimates of BPOR and MLLOR and approximate BLOR and MLLOR with *logit*$(u) = 1.75\times\Phi^{-1}(u)$. The best model has the value for the parameter closer to the true value; these are presented in bold. MLEs, maximum likelihood estimators; POs, percentages of outliers; BLOR, Bayesian logistic ordinal regression; MLLOR, maximum likelihood logistic ordinal regression.

includes the Gibbs sampler for the BPOR model as a particular case.

Simulation study

The purpose of this simulation study is twofold: (1) to compare the performance of the proposed BLOR with: (a) the approximation resulting from using the estimates of the BPOR with $\text{logit}(u) = 1.75\times\Phi^{-1}(u)$, denoted as BLOR*, (b) with the results of using maximum likelihood estimators (MLEs) with logit link for ordinal data (MLLOR), and (c) the approximation resulting from multiplying the MLEs with probit link for ordinal data by 1.75, denoted as MLLOR*; (2) to evaluate the performance of the BLOR in the presence of outliers. We used the value 1.75 because, according to the literature review, it is the most reasonable value (Savalei 2006).

To reach these two goals, two data sets were simulated. Both simulation studies were carried out with the following liability:

$$l_{ij} = \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_{ij},$$

where $i = 1, \ldots, 40$ and $j = 1, \ldots, n_i$, $\boldsymbol{\beta}^T = [-6, 5, 7]$ and the vectors $\boldsymbol{x}_i^T = [x_{i1}, x_{i2}, x_{i3}]$ have been drawn independently, with components following a uniform distribution within the interval [−0.1, 0.1]. The threshold parameters used were $\gamma_1 = -0.8416$, $\gamma_2 = -0.2533$, $\gamma_3 = -0.2533$, and $\gamma_4 = -0.8416$. Then the response variable was generated as follows:

$$y_{ij} = \begin{cases} 1 & \textit{if} \; -\infty < l_{ij} < \gamma_1, \\ 2 & \textit{if} \; \gamma_1 < l_{ij} < \gamma_2, \\ & \vdots \\ 5 & \textit{if} \; \gamma_4 < l_{ij} < \infty \end{cases}$$

For simulated data set 1, we used four values of sample size n_i = 5, 10, 20, and 40 and all the ε_{ij} were drawn independently from a $L(0,1)$, whereas for simulated data set 2, we used only one sample size (n_i = 40) and the error terms (ε_{ij}) were obtained from two distributions [$L(0,1)$ and a student's t distribution with four degrees of freedom, denoted as t4]. We studied four scenarios: Scenario 1, the percentage of outliers (PO) from the t4 was 5% and the remaining percentage was obtained from the $L(0,1)$ distribution; Scenario 2: the PO from the t4 was 10%, and 90% from the $L(0,1)$; Scenario 3: the PO from the t4 was 20%, and 80% from the $L(0,1)$; and Scenario 4: the PO from the t4 was 30%, and 70% from the $L(0,1)$.

The MLE estimates for the ordinal regression were obtained using the polr function of the MASS package in R (R Core Team 2015). It is important to point out that the priors used for the Bayesian methods were not informative for $\boldsymbol{\beta}|\sigma^2_\beta \sim N(\boldsymbol{\beta}_0^T = [0, 0, 0], 10000\times\boldsymbol{I}_3)$, and for the hyperparameters for thresholds, we used $\gamma_{\min} = -4$ and $\gamma_{\max} = 4$. We computed 20,000 Markov chain Monte Carlo (MCMC) samples. Bayes estimates were computed using 10,000 samples because the first 10,000 were discarded as burn-in.

Table 3 Real data sets: GLS and Septoria data sets

Model	Set (DIC)	Statistic	Probability of Each Category					BS
			1	2	3	4	5	
		Mean	0.025	0.256	0.392	0.196	0.132	0.363
	Zimbabwe	L	0.018	0.236	0.368	0.173	0.112	0.363
	(4150.29)	U	0.034	0.278	0.415	0.216	0.149	0.363
		Mean	0.054	0.442	0.267	0.169	0.069	0.350
BLOR	México	L	0.035	0.399	0.228	0.139	0.050	0.349
	(1313.82)	U	0.075	0.480	0.308	0.200	0.092	0.352
		Mean	0.204	0.248	0.259	0.212	0.077	0.390
	Colombia	L	0.179	0.219	0.232	0.187	0.058	0.389
	(2577.92)	U	0.233	0.279	0.287	0.239	0.095	0.391
		Mean	0.083	0.284	0.332	0.198	0.104	0.377
	Pooled	L	0.069	0.269	0.315	0.184	0.093	0.377
	(8327.36)	U	0.093	0.299	0.350	0.211	0.119	0.377
		Mean	0.077	0.150	0.432	0.341	–	0.335
	Septoria	L	0.049	0.112	0.376	0.294	–	0.333
	(654.33)	U	0.110	0.195	0.493	0.388	–	0.338
		Mean	0.032	0.237	0.416	0.190	0.124	0.363
	Zimbabwe	L	0.025	0.219	0.389	0.171	0.110	0.363
	(4156.86)	U	0.039	0.258	0.443	0.211	0.140	0.365
		Mean	0.057	0.436	0.280	0.156	0.070	0.350
BLOR*	México	L	0.041	0.392	0.242	0.128	0.054	0.350
	(1315.21)	U	0.075	0.479	0.323	0.187	0.089	0.353
		Mean	0.193	0.260	0.279	0.194	0.074	0.390
	Colombia	L	0.168	0.226	0.248	0.168	0.060	0.389
	(2581.66)	U	0.220	0.292	0.310	0.223	0.089	0.392
		Mean	0.082	0.277	0.358	0.184	0.100	0.377
	Pooled	L	0.074	0.259	0.341	0.168	0.090	0.377
	(8339.54)	U	0.091	0.294	0.375	0.200	0.109	0.378
		Mean	0.075	0.137	0.457	0.332	–	0.334
	Septoria	L	0.051	0.098	0.392	0.268	–	0.330
	(652.50)	U	0.104	0.176	0.527	0.393	–	0.341
		Mean	0.025	0.253	0.392	0.199	0.132	0.363
	Zimbabwe	L	0.017	0.233	0.366	0.179	0.113	0.363
	(4150.18)	U	0.033	0.274	0.416	0.219	0.148	0.363
		Mean	0.054	0.440	0.265	0.171	0.070	0.350
BPOR	México	L	0.036	0.399	0.229	0.142	0.051	0.350
	(1314.70)	U	0.076	0.479	0.304	0.203	0.092	0.352
		Mean	0.206	0.249	0.261	0.209	0.075	0.390
	Colombia	L	0.176	0.221	0.230	0.183	0.058	0.389
	(2578.42)	U	0.233	0.277	0.293	0.233	0.095	0.391
		Mean	0.071	0.256	0.331	0.218	0.123	0.377
	Pooled	L	0.042	0.183	0.313	0.191	0.104	0.374
	(8329.84)	U	0.086	0.287	0.350	0.286	0.171	0.389
		Mean	0.075	0.150	0.430	0.345	–	0.334
	Septoria	L	0.047	0.110	0.369	0.282	–	0.330
	(651.79)	U	0.109	0.191	0.500	0.402	–	0.339

BLOR* use the parameter estimates of the BPOR and approximate the BLOR with $logit(u) = 1.75\times\Phi^{-1}(u)$. Point probability estimates, credible sets for each category, DIC, and BS for the threshold Bayesian ridge regression. L and U denote lower and upper confidence sets, respectively. GLS, Gray leaf spot; DIC, deviance information criterion; BS, Brier scores; BLOR, Bayesian logistic ordinal regression; BPOR, Bayesian probit ordinal regression.

Model implementation

The Gibbs sampler described previously for the BLOR model was implemented with the R-software (R Core Team 2015). Implementation was done with a Bayesian approach and MCMC through the Gibbs sampler algorithm, which samples sequentially from the full conditional distributions until it reaches a stationary process, converging with the joint posterior distribution (Gelfand and Smith 1990). In the real data, to reduce the potential impact of MCMC errors on prediction accuracy, we performed a total of 60,000 iterations with a burn-in of 20,000, so that 40,000 samples were used for inference. We did not apply thinning of the chains, following the suggestions of Geyer (1992), Maceachern and Berliner (1994), and Link and Eaton (2012), who provide justification for the ban on subsampling MCMC output for approximating simple features of the target distribution (*e.g.*, means, variances and percentiles), since thinning is neither necessary nor desirable, and unthinned chains are more precise. It is important to point out that implementation of the BLOR model for the real data sets was done using the hyperparameters $\nu_\beta = 3$, $S_\beta = 0.001$, $\boldsymbol{\beta}_0 = \mathbf{0}$, $\boldsymbol{\Sigma}_0 = 10000\times\boldsymbol{I}_p,)$, $\gamma_{\min} = -1000$ and $\gamma_{\max} = -1000$ for thresholds parameters, all of which were chosen to lead weakly informative priors.

Assessing prediction accuracy

We used cross-validation to estimate the prediction accuracy of the proposed models. The data set was divided into training and validation sets 10 times, with 90% of the data set used for training and 10% for testing; this was done only for the pooled data. The training set was used to fit the model and the validation set was used to evaluate the prediction accuracy of the proposed models. Since the phenotype response is ordinal categorical, we used the Brier score (Brier 1950) to measure prediction ability, which is equal to

$$BS = n^{-1}\sum_{i=1}^{n}\sum_{c=1}^{C}(\hat{\pi}_{ic}-d_{ic})^{2} \tag{10}$$

where BS denotes the Brier Score, and d_{ic} takes a value of 1 if the ordinal categorical response observed for individual i falls into category c; otherwise, d_{ic} = 0. This scoring rule uses all the information contained in the predictive distribution, not just a small portion like the hit rate or the log-likelihood score. Therefore, it is a reasonable choice for comparing categorical regression models, although there are other scoring rules that also have good properties. The range of BS in equation (10) is between 0 and 2. For this reason, we divided Brier scores (BS)/2 to get the BS bounded between 0 and 1; lower scores imply better predictions. It is important to point out that we also used the BS when analyzing the full data sets. We also used the deviance information criterion (DIC) to compare Bayesian models, as suggested by Gelman *et al.* (2003); here, the lower the DIC, the better the model.

Data availability

The two real data sets and two simulated data sets together with R codes are deposited in the link http://hdl.handle.net/11529/10254. The phenotypic data for GLS in three environments (México, Zimbabwe, and Colombia) for the 278 maize lines, the 46,347 SNPs, and the R scripts developed to fit the predictive models used in this study are given in the files PhenoGLS.RData, and MarkersGLSFinal.RDat. The repository also contains the Septoria genotypic and phenotypic data sets, in files SeptoriaGenotypic.RDat, and SeptoriaPhenotypic. RDat, respectively. The R codes to generate the simulated data sets 1 and 2, and for analyzing the real data set from Colombia are directly given in Appendices B, C, and D, respectively.

RESULTS

In the following sections, we investigate the performance of the proposed BLOR estimator through a simulation study and with real data.

Simulated data set

In Table 1, we report average estimates obtained by all methods, along with SDs; all the results are based on 50 replications. From Table 1, it is clear that as the sample size increases, the average biases and SD decrease in all cases. This confirmed the consistency properties of all the estimates. Table 1 also shows that, in general, the point estimates of the Bayesian estimates (BLOR and BLOR*) are similar to the MLEs (MLLOR, and MLLOR*, which was approximated with the MLE with probit ordinal regression); however, the approximations (BLOR* and MLLOR*) have greater bias and SD. BLOR has less bias and SD in most of the studied parameters, producing better parameter estimates than the MLLOR (which is the correct method that use maximum likelihood with the logit link function). For this reason, the proposed BLOR is an excellent alternative. However, a more in-depth simulation study is required to ensure that these findings are valid for all possible scenarios.

Table 4 GLS data set

Model	Brier Scores		
	Mean	Min	Max
BLOR	0.373	0.365	0.381
BLOR*	0.374	0.364	0.382
BPOR	0.373	0.365	0.381

BLOR* uses the parameter estimates of BPOR and approximates BLOR with $logit(u) = 1.75\times\Phi^{-1}(u)$. Brier scores (mean, minimum and maximum; lower scores indicate better prediction) evaluated for validation samples from the pooled data. GLS, Gray leaf spot; BLOR, Bayesian logistic ordinal regression; BPOR, Bayesian probit ordinal regression.

Table 2 shows that the smaller the PO, the less bias there is in the parameter estimates under the four methods (two Bayesians and two under Maximum Likelihood). However, under the two Bayesian methods, the proposed BLOR showed less bias than BLOR*, and this behavior is observed for all parameters and PO under study, except when the PO is 30% and for the parameter γ_2. Under the two maximum likelihood methods, the approximate method (MLLOR*) showed greater bias in all scenarios and parameters. Finally, although BLOR and MLLOR produced better results in term of bias, in most cases, MLLOR was better than BLOR.

Real data sets

Next we compared BLOR and BPOR with the real data set (GLS and Septoria data) described in the *Materials and Methods*. However, because with this data set the number of parameters (betas and thresholds) to be estimated is high, we compared both models with point and interval estimates of the probabilities estimated for each category on the four- and five-point ordinal scale for each data set studied. In Table 3, we compare BLOR, BPOR, and BLOR* (this is the approximate method since the estimates resulting from BPOR were used to approximate BLOR with $\text{log}it(u) = 1.75\times\Phi^{-1}(u)$) with five data sets (three locations, the resulting pooling of the three locations and the Septoria data set) made up from the data sets described in the section *Materials and Methods*. First, when comparing BLOR with BLOR*, we see that there are differences in the point probability estimates, and the lower the probabilities, the greater the differences. However, in general, the widths of the credible sets are similar. Second, when comparing BLOR with BPOR, we see that the point and credible sets for each location and Septoria produced similar results; however, for the pooled data, BLOR produced estimates with narrower confidence sets than BPOR (Table 3). We observed that the estimates produced using BPOR are less accurate because wider confidence sets are produced when the data are pooled; this could be because the assumption of errors normally distributed with mean zero and variance one when the data are pooled is not fulfilled. Also, there were no differences in the BS of the three models (BLOR, BLOR*, and BPOR).

Comparing the models with DIC, we see that in Zimbabwe, BPOR produced the lowest DIC (DIC = 4150.18), whereas BLOR* produced the highest (DIC = 4156.86). However, the DIC of BLOR (DIC = 4150.29) was very close to that of BPOR. In México, the lowest DIC was for BLOR (DIC = 1313.82), whereas the highest was for BLOR* (DIC = 1315.21). In Colombia, BLOR had the lowest DIC (2577.92) and BLOR* had the highest (2581.66). In the Septoria data set, BPOR had the lowest DIC (651.79) and BLOR had the highest (654.33). Finally, in the pooled data, BLOR (8327.36) also had the lowest DIC, whereas BLOR* had the highest (8339.539). All these results shows that even approximations that did a reasonable job (BLOR*) were sometimes very far from the exact methods (BLOR and BPOR). For this reason, whenever possible, an exact model (BLOR or BPOR) should be chosen. The fact that BPOR is sometimes the best (lower DIC) implies that for these data sets, the assumption $\varepsilon_{ij} \sim N(0,1)$ in Equation (1) is enough.

Finally, in Table 4 we present the BS for the testing sets of the pooled real GLS data with 10 cross-validations and 90% of data used for the training set and 10% for the testing set. For models BLOR and BPOR, the BS are almost identical, which means that, with regard to prediction, both models had a similar performance. However, although the approximation method (BLOR*) produced a higher BS, its prediction accuracy was very close to that of BLOR and BPOR. Although we found that the three models had a similar performance regarding prediction accuracy with this data set, more in-depth research is required to validate this observation.

This paper proposes a method for BLOR using the Pólya-Gamma data augmentation approach of Scott and Pillow (2013), which produces a Gibbs sampler with full conditional distributions similar to that of the BPOR model of Albert and Chib (1993). The proposed method is reduced to the BPOR model when the sampled values, ω_{ij}, from the Pólya-Gamma distribution in Equation (4) are set to 1. This is an advantage because with the proposed model, researchers can perform an exact logistic or probit ordinal regression without having to do approximations to perform a logistic ordinal regression. The performance of the proposed method was compared with the approximation using the probit ordinal regression model in a small simulation study and real data sets (GLS and Septoria data sets) using a four- and five-point ordinal scale. On the basis of the simulation study, it is clear that the estimation of parameters using the approximation $logit(u) = (1.75)\Phi^{-1}(u)$ produces a considerable amount of bias and can give rise to wrong conclusions in association studies. However, we observed with the real data that, in terms of prediction ability, both models (BLOR and BPOR) have a similar performance even though we observed BLOR had lower DIC values in México, Colombia, and the pooled data. This means that when violation of the assumption $\varepsilon_{ij} \sim N(0,1)$ in Equation (1) is not strong, any model can be used. For this reason, we observed greater accuracy (narrow confidence sets) for the BLOR model compared with the exact BPOR model (BPOR) only with the pooled data set without a covariate for location.

Although with the real data we did not observe an advantage in prediction accuracy with the proposed BLOR model, it is very well documented in statistical literature that logistic ordinal regression is more robust for dealing with outlying data, because logistic distribution has heavier tails which was corroborated in terms of parameter estimates with the simulation done using simulated data set 2. For this reason, the proposed BLOR should be preferred because it is usually not practical to test if the error term in Equation (1) is $\varepsilon_{ij} \sim N(0,1)$ or $L(0,1)$. In addition to being more robust, the proposed method also provides regression coefficients that are more interpretable because of their connection to odds ratios (Zucknick and Richardson 2014). However, this advantage does not make sense when $p >> n$, because the main driving force in Bayesian models in the case of $p >> n$ is the prior and not the data (Gianola 2013). Even with this restriction, this paper unifies logistic and probit ordinal regression under a Bayesian framework and is a useful alternative for genomic-enabled prediction of ordinal categorical trials where available data sets have a larger number of parameters to estimate than observations. Also, the proposed method should be preferred over BPOR when outliers are present and not easily detected. This is especially true for multidimensional data, since many times a few outliers in a data set can have strong influence on parameter estimation and inference.

Finally, it is important to point out that, to devise the method proposed in this paper, we generalized the work of Scott and Pillow (2013) for ordered categorical responses. Our method is elegant, easy to implement, and produces a unified Gibbs sampler framework useful for both the logit and the probit link. For this reason, we believe it is an appealing alternative for plant and animal researchers. Also, the proposed BLOR model can be easily extended to take into account genotype × environment interactions, which play an extremely important role in plant breeding, especially when selecting candidate genotypes.

ACKNOWLEDGMENTS

We thank all International Maize and Wheat Improvement Center (CIMMYT) researchers in the Global Maize Breeding Program (GMP) and in the Global Wheat Breeding Program (GWP), as well as the researchers in the National Programs, that generate the data we use in this study.

LITERATURE CITED

Albert, J. H., and S. Chib, 1993 Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc. 88: 669–679.

Amemiya, T., 1981 Qualitative response models: a survey. J. Econ. Lit. 19: 1483–1536.

Bartholomew, D. J., and M. Knott, 1999 Latent Variable Models and Factor Analysis. Oxford University Press, New York.

Bartlett, M. S., 1947 The use of transformations. Biometrics 3: 39–52.

Brier, G. W., 1950 Verification of forecasts expressed in terms of probability. Mon. Weather Rev. 78: 1–3.

Camilli, G., 1994 Teacher's corner: origin of the scaling constant d= 1.7 in item response theory. J. Educ. Behav. Stat. 19: 293–295.

Crossa, J., P. Pérez-Rodríguez, G. de los Campos, G. Mahuku, S. Dreisigacker *et al.*, 2011 Genomic selection and prediction in plant breeding. J. Crop Improv. 25: 239–261.

de los Campos, G., and P. Perez-Rodriguez, 2013 BGLR: Bayesian generalized linear regression. R package version. Available at: http://R-Forge.R-project.org/projects/bglr/. Accessed August 31, 2015.

Gelfand, A. E., and A. F. Smith, 1990 Sampling-based approaches to calculating marginal densities. J. Am. Stat. Assoc. 85: 398–409.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 2003 Bayesian Data Analysis. Ed. 2. Chapman & Hall/CRC, Boca Raton, Florida.

Geyer, C. J., 1992 Practical Markov Chain Monte Carlo. Stat. Sci. 7: 473–483.

Gianola, D., 1980 A method of sire evaluation for dichotomies. J. Anim. Sci. 51: 1266–1271.

Gianola, D., 1982 Theory and analysis of threshold characters. J. Anim. Sci. 54: 1079–1096.

Gianola, D., 2013 Priors in whole-genome regression: the Bayesian alphabet returns. Genetics 194: 573–596.

Gianola, D., and J. L. Foulley, 1983 Sire evaluation for ordered categorical data with a threshold model. Genet. Sel. Evol. 15: 1–23.

González-Camacho, J. M., G. de los Campos, P. Pérez-Rodríguez, D. Gianola, J. E. Cairns *et al.*, 2012 Genome-enabled prediction of genetic values using radial basis function neural networks. Theor. Appl. Genet. 125: 759–771.

González-Recio, O., and S. Forni, 2011 Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genet. Sel. Evol. 43: 7.

Hoerl, A. E., and R. W. Kennard, 1970 Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12: 55–67.

Link, W. A., and M. J. Eaton, 2012 On thinning of chains in MCMC. Methods Ecol. Evol. 3: 112–115.

MacEachern, S. N., and L. M. Berliner, 1994 Subsampling the Gibbs sampler. Am. Stat. 48: 188–190.

Montesinos-López, O. A., A. Montesinos-López, P. Pérez-Rodríguez, G. de los Campos, K. M. Eskridge, and J. Crossa, 2015 Threshold models for genome-enabled prediction of ordinal categorical traits in plant breeding. G3 (Bethesda) 5: 291–300.

Nelder, J. A., and R. W. M. Wedderburn, 1972 Generalized linear models. J. R. Stat. Soc. A 135: 370–384.

Perez-Rodriguez, P., and G. de los Campos, 2014 Genome-wide regression & prediction with the BGLR statistical package. Genetics 198: 483–495.

Poland, J, J. Endelman, J. Dawson, J. Rutkoski, S. Wu *et al.*, 2012 Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome 5: 103–113.

Polson, N. G., J. G. Scott, and J. Windle, 2013 Bayesian inference for logistic models using Pólya-Gamma latent variables. J. Am. Stat. Assoc. 108 (504): 1339–1349.

R Core Team, 2015 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/. Accessed August 31, 2015.

Savalei, V., 2006 Logistic approximation to the normal: the KL rationale. Psychometrika 71: 763–767.

Scott, J., and J. W. Pillow, 2013 Fully Bayesian inference for neural models with negative-binomial spiking, pp. 1898–1906 in Advances in Neural Information Processing Systems 25, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Available at: http://papers.nips.cc/book/advances-in-neural-information-processing-systems-25-2012. Accessed August 31, 2015.

Sorensen, D. A., S. Andersen, D. Gianola, and I. Korsgaard, 1995 Bayesian inference in threshold models using Gibbs sampling. Genet. Sel. Evol. 27: 229–249.

Stroup, W. W., 2012 *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. CRC Press, Boca Raton, Florida.

Stroup, W. W., 2015 Rethinking the analysis of non-normal data in plant and soil science. Agron. J. 107: 811–827.

Villanueva, B., J. Fernandez, L. A. Garcia-Cortes, L. Varona, H. D. Daetwyler *et al.*, 2011 Accuracy of genome-wide evaluation for disease resistance in aquaculture breeding programs. J. Anim. Sci. 89: 3433–3442.

Wang, C. L., X. D. Ding, J. Y. Wang, J. F. Liu, W. X. Fu *et al.*, 2012 Bayesian methods for estimating GEBVs of threshold traits. Heredity 110: 213–219.

Yang, W., and R. J. Tempelman, 2012 A Bayesian antedependence model for whole genome prediction. Genetics 190: 1491–1501.

Zucknick, M., and S. Richardson, 2014 MCMC algorithms for Bayesian variable selection in the logistic regression model for large-scale genomic applications. Technical Report. Available at: http://arxiv.org/abs/1402.2713. Accessed August 31, 2015.

Communicating editor: E. Huang

APPENDIX A

Derivation of full conditional distributions

Liabilities and Pólya-Gamma values

The fully conditional posterior distribution of liability l_{ij} is

$$P(\boldsymbol{l}|ELSE) \propto P(\boldsymbol{l}|\boldsymbol{\beta},\boldsymbol{b})P(\boldsymbol{y}|\boldsymbol{l},\boldsymbol{\gamma})$$

$$\propto \prod_{i=1}^{I}\prod_{j=1}^{J} f(l_{ij})\sum_{c=1}^{C} I(y_{ij}=c)I\left(\gamma_{c-1}<l_{ij}<\gamma_c\right)$$

$$\propto \prod_{i=1}^{I}\prod_{j=1}^{J}\frac{\exp\left(-l_{ij}+\boldsymbol{x}_{ij}^{T}\boldsymbol{\beta}+b_i\right)}{\left[1+\exp\left(-l_{ij}+\boldsymbol{x}_{ij}^{T}\boldsymbol{\beta}+b_i\right)\right]^2}\sum_{c=1}^{C} I(y_{ij}=c)I\left(\gamma_{c-1}<l_{ij}<\gamma_c\right)$$

$$\propto \prod_{i=1}^{I}\prod_{j=1}^{J}2^{-2}\int_{0}^{\infty}\exp\left[-\frac{\omega_{ij}\left(-l_{ij}+\boldsymbol{x}_{\boldsymbol{ij}}^{\boldsymbol{T}}\boldsymbol{\beta}+b_i\right)^2}{2}\right]P(\omega_{ij};b=2,d=0)d\omega_{ij}\sum_{c=1}^{C} I(y_{ij}=c)I\left(\gamma_{c-1}<l_{ij}<\gamma_c\right)$$

The last inequality was obtained using a technique called the Pólya-Gamma method (Scott and Pillow 2013), which is useful when working with logistic likelihoods, and has the form

$$\frac{(e^{\psi})^a}{(1+e^{\psi})^b}=2^{-b}e^{\kappa\psi}\int_{0}^{\infty}e^{-\frac{\omega\psi^2}{2}}P(\omega;b,0)d\omega$$

where $\kappa=a-b/2$ and $P(\omega;b,d=0)$ denotes the density of the random variable $\omega\sim PG(b,d=0)$, where $PG(b,d)$ denotes a Pólya-Gamma distribution with parameters b and d and density

$$P(\omega;b,d)=\left\{cosh^{b}\left(\frac{d}{2}\right)\right\}\quad\frac{2^{b-1}}{\Gamma(b)}\sum_{k=0}^{\infty}(-1)^{n}\frac{\Gamma(n+b)(2n+b)}{\Gamma(n+1)\sqrt{2\pi\omega^3}}\exp\left(-\frac{(2n+b)^2}{8\omega}-\frac{d^2}{2}\omega\right),$$

where *cosh* denotes the hyperbolic cosine.

Then the joint distribution of l_{ij} and ω_{ij} is equal to

$$P(\boldsymbol{l},\boldsymbol{\omega}|ELSE)\propto\prod_{i=1}^{I}\prod_{j=1}^{J}2^{-2}\exp\left[-\frac{\omega_{ij}\left(-l_{ij}+\boldsymbol{x}_{ij}^{T}\boldsymbol{\beta}+b_i\right)^2}{2}\right]P(\omega_{ij};2,0)\sum_{c=1}^{C} I(y_{ij}=c)I\left(\gamma_{c-1}<l_{ij}<\gamma_c\right)$$

Therefore, the fully conditional posterior distribution of liability l_{ij} is a truncated normal distribution and its density is

$$P\left(l_{ij}|ELSE\right)=\frac{\phi(\boldsymbol{x}_{\boldsymbol{ij}}^{\boldsymbol{T}}\boldsymbol{\beta}+b_i,1/\sqrt{\omega_{ij}})}{\Phi\left(\gamma_c-\boldsymbol{x}_{ij}^{T}\boldsymbol{\beta}-b_i\right)-\Phi\left(\gamma_{c-1}-\boldsymbol{x}_{ij}^{T}\boldsymbol{\beta}-b_i\right)}$$

where $\phi(.)$ is a normal density with parameters as indicated in the argument, Φ is the cumulative distribution function of a normal density with mean $\boldsymbol{x}_{ij}^{T}\boldsymbol{\beta}+b_i$ and variance $1/\sqrt{\omega_{ij}}$, and the fully conditional posterior distribution of ω_{ij} is

$$P\left(\omega_{ij}|ELSE\right)\propto 2^{-2}\exp\left[-\frac{\omega_{ij}\left(-l_{ij}+\boldsymbol{x}_{ij}^{T}\boldsymbol{\beta}+b_i\right)^2}{2}\right]P(\omega_{ij};2,0)\propto\exp\left[-\frac{\omega_{ij}\left(-l_{ij}+\boldsymbol{x}_{ij}^{T}\boldsymbol{\beta}+b_i\right)^2}{2}\right]P(\omega_{ij};2,0)$$

and from here and equation (5) of Polson *et al.* (2013) we get that

$$\omega_{ij}|ELSE\sim PG(2,-l_{ij}+\boldsymbol{x}_{ij}^{T}\boldsymbol{\beta}+b_i)$$

Regression coefficients ($\boldsymbol{\beta}$)

First note that the fully conditional posterior of $\boldsymbol{l}, \boldsymbol{\beta}, \boldsymbol{\omega}$ is

$$P(\boldsymbol{l}, \boldsymbol{\beta}, \boldsymbol{\omega}|ELSE) \propto P(\boldsymbol{l}|\boldsymbol{\beta}, \boldsymbol{b})P(\boldsymbol{y}|\boldsymbol{l}, \boldsymbol{\gamma})P(\boldsymbol{\omega})P\left(\boldsymbol{\beta}\middle|\sigma_\beta^2\right)$$

$$\propto \exp\left(-\frac{1}{2}(-\boldsymbol{l} + \boldsymbol{X}\beta + \boldsymbol{Zb})^T\boldsymbol{D}_\omega(-\boldsymbol{l} + \boldsymbol{X\beta} + \boldsymbol{Zb})\right)P(\boldsymbol{\omega})P\left(\boldsymbol{\beta}\middle|\sigma_\beta^2\right)$$

where $P(\boldsymbol{\omega}) = \prod_{i=1}^{I}\prod_{j=1}^{m_i} P(\omega_{ij}; 2, 0)$. Then, the fully conditional posterior distribution of $\boldsymbol{\beta}$ is

$$P(\boldsymbol{\beta}|ELSE) \propto \exp\left(-\frac{1}{2}(-\boldsymbol{l} + \boldsymbol{X}\beta + \boldsymbol{Zb})^T\boldsymbol{D}_\omega(-\boldsymbol{l} + \boldsymbol{X\beta} + \boldsymbol{Zb}) - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\left(\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}\right)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\boldsymbol{\beta}^T\left(\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2} + \boldsymbol{X}^T\boldsymbol{D}_\omega\boldsymbol{X}\right)\boldsymbol{\beta} - 2\left(\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}\boldsymbol{\beta}_0 - \boldsymbol{X}^T\boldsymbol{D}_\omega\boldsymbol{Zb} + \boldsymbol{X}^T\boldsymbol{D}_\omega\boldsymbol{l}\right)^T\boldsymbol{\beta}\right]\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\left(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_0\right)^T\tilde{\boldsymbol{\Sigma}}_0^{-1}\left(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_0\right)\right]\right)$$

where $\tilde{\boldsymbol{\Sigma}}_0 = (\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2} + \boldsymbol{X}^T\boldsymbol{D}_\omega\boldsymbol{X})^{-1}$, $\tilde{\boldsymbol{\beta}}_0 = \tilde{\boldsymbol{\Sigma}}_0(\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}\boldsymbol{\beta}_0 - \boldsymbol{X}^T\boldsymbol{D}_\omega\boldsymbol{Zb} + \boldsymbol{X}^T\boldsymbol{D}_\omega\boldsymbol{l})$. It is important to point out that if we use a prior for $\boldsymbol{\beta} \propto Constant$ (improper uniform distribution), then in $\tilde{\boldsymbol{\Sigma}}_0$ and $\tilde{\boldsymbol{\beta}}_0$ we need to make $\mathbf{0}$ the term $\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}$. Finally, the fully conditional posterior of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}|ELSE \sim N_p\left(\tilde{\boldsymbol{\beta}}_0, \tilde{\boldsymbol{\Sigma}}_0\right)$$

Polygenic effects (b)

Now the fully conditional posterior of $\boldsymbol{b}$ is given as

$$P(\boldsymbol{b}|ELSE) \propto \exp\left(-\frac{1}{2}(-\boldsymbol{l} + \boldsymbol{X\beta} + \boldsymbol{Zb})^T\boldsymbol{D}_\omega(-\boldsymbol{l} + \boldsymbol{X\beta} + \boldsymbol{Zb})\right)\ P(\boldsymbol{b}|\sigma_b^2)$$

$$\propto \exp\left\{-\frac{1}{2}\ \left[\boldsymbol{b}^T(\sigma_b^{-2}\boldsymbol{G}^{-1} + \boldsymbol{Z}^T\boldsymbol{D}_\omega\boldsymbol{Z})\boldsymbol{b} - 2\ \left(\boldsymbol{Z}^T\boldsymbol{D}_\omega\boldsymbol{l} - \boldsymbol{Z}^T\boldsymbol{D}_\omega\boldsymbol{X\beta}\right)^T\boldsymbol{b}\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\ (\boldsymbol{b} - \tilde{\boldsymbol{b}})^T\boldsymbol{F}^{-1}\left(\boldsymbol{b} - \tilde{\boldsymbol{b}}\right)\right\}$$

This implies that the fully conditional posterior of $\boldsymbol{b}$ is

$$\boldsymbol{b}|ELSE \sim N_I\left(\tilde{\boldsymbol{b}} = \boldsymbol{F}\left(\boldsymbol{Z}^T\boldsymbol{D}_\omega\boldsymbol{l} - \boldsymbol{Z}^T\boldsymbol{D}_\omega\boldsymbol{X\beta}\right), \boldsymbol{F} = \left(\sigma_b^{-2}\boldsymbol{G}^{-1} + \boldsymbol{Z}^T\boldsymbol{D}_\omega\boldsymbol{Z}\right)^{-1}\right)$$

Variance of polygenic effects

Next, the conditional distribution of σ_b^2 is obtained. If $\sigma_b^2 \sim \chi^{-2}(\nu_b, S_b)(shape\ \ and\ \ scale)$, then

$$P\left(\sigma_b^2\middle|ELSE\right) \propto \frac{1}{(\sigma_b^2)^{\frac{\nu_b+I}{2}+1}}\exp\left(-\frac{\boldsymbol{b}^T\boldsymbol{G}^{-1}\boldsymbol{b} + \nu_b S_b}{2\sigma_b^2}\right)$$

This is the kernel of the scaled inverted χ^2 distribution; therefore, the fully conditional posterior is

$$\sigma_b^2|ELSE \sim \chi^{-2}(\tilde{\nu}_b = \nu_b + I, \tilde{S}_b = (\boldsymbol{b}^T\boldsymbol{G}^{-1}\boldsymbol{b} + \nu_b S_b)/\nu_b + I)$$

Threshold effects ($\boldsymbol{\gamma}$)

The density of the fully conditional posterior distribution of the c^{th} threshold, γ_c, is

$$P(\boldsymbol{\gamma}|ELSE) \propto P(y|\boldsymbol{l}, \boldsymbol{\gamma})P(\boldsymbol{\gamma})$$

$$\propto \prod_{i=1}^{I}\prod_{j=1}^{J}\sum_{c=1}^{C} I(y_{ij}=c)I\left(\gamma_{c-1}<l_{ij}<\gamma_c\right)I(\gamma\in T) \tag{A.1}$$

If Equation (A.1) is seen as a function of γ_c, it is evident that the value of γ_c must be larger than all the $l_{ij}\big|y_{ij}=c$ and smaller than all the $l_{ij}\big|y_{ij}=c+1$. Hence, as a function of γ_c, Equation (A.1) leads to the uniform density

$$P(\gamma_c|ELSE)=\frac{1}{\min\left(l_{ij}\big|y_{ij}=c+1\right)-\max\left(l_{ij}\big|y_{ij}=c\right)}I(\gamma\in T) \tag{A.2}$$

Equation (A.2) corresponds to a uniform distribution on the interval $[\min\{\min\ (l_{ij}\big|y_{ij}=c+1),\gamma_{c+1},\gamma_{\max}\},\max\{\max(l_{ij}\big|y_{ij}=c),\gamma_{c-1},\gamma_{\min}\}]$ (Albert and Chib 1993; Sorensen *et al.* 1995).

Variance of location effects

If we give $\sigma_\beta^2\sim\chi^{-2}(\nu_\beta,S_\beta)(shape\ \ and\ \ scale)$, then

$$P\left(\sigma_\beta^2|ELSE\right)\propto P\left(\sigma_\beta^2\right)P\left(\boldsymbol{\beta}|\ \sigma_\beta^2\right)=\frac{1}{(\sigma_\beta^2)^{\frac{\nu_\beta}{2}+1}}\exp\left(-\frac{\nu_\beta S_\beta}{2\sigma_\beta^2}\right)P\left(\boldsymbol{\beta}|\ \sigma_\beta^2\right)$$

$$\propto\frac{1}{(\sigma_\beta^2)^{\frac{\nu_\beta+p}{2}+1}}\exp\left(-\frac{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+\nu_\beta S_\beta}{2\sigma_\beta^2}\right)$$

This is the kernel of the scaled inverted χ^2 distribution; therefore, the fully conditional posterior is

$$\sigma_\beta^2\Big|ELSE\ \sim\chi^{-2}(\tilde{\nu}_\beta=\nu_\beta+p,\tilde{S}_\beta=[(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\Sigma_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+\nu_\beta S_\beta]\Big/\nu_\beta+p)$$

APPENDIX B

R code for simulating data set 1

We used this code for each sample size studied, and thus we only need to change ni, that denotes n_i, for the other values (10, 20, 40). For each sample size, we used this code 50 times for estimating the results in Table 1.

```
thetav = c(-5,-10,15)
Datos<-numeric(0)
nC=40;ni=5 ####Change this ni for other sample sizes
for(i in 1:nC)
{xi1<-runif(length(thetav),-0.1,0.1)
Eta=t(xi1)%*%thetav
for(j in 1:ni)
{L=Eta + rlogis(1, location = 0, scale = 1)
y=ifelse(L<=-0.8416,1,ifelse(L<=-0.2533,2,ifelse(L<0.2533,3,ifelse(L<0.8416,4,5))))
 Datos<-rbind(Datos,c(j,y,t(xi1)))
}}
colnames(Datos) = c('j','y','x1','x2','x3')
Datos
```

APPENDIX C

R codes for simulating data set 2

We used this code for each percentage of outliers (PO) studied; for this reason, we only need to change sam.out for values: 4, 8, and 12, which represent 10, 20, and 30% of outliers. For each sample size, we used this code 50 times for estimating the results in Table 2.

```
thetav = c(-5,-10,15)
Datos<-numeric(0)
nC=40; ni=40, sam.out =2 ###(this 2 represents 5% of outliers)
for(i in 1:nC)
{xi1<-runif(length(thetav),-0.1,0.1)
Eta=t(xi1)%*%thetav
for(j in 1:ni)
{ L=Eta + ifelse(j>sam.out,rlogis(1, location = 0, scale = 1), rt(1, df=4, ncp=0))
y=ifelse(L<=-0.8416,1,ifelse(L<=-0.2533,2,ifelse(L<0.2533,3,ifelse(L<0.8416,4,5))))
```

```
 Datos<-rbind(Datos,c(j,y,t(xi1)))
}}
colnames(Datos) = c(’j’,’y’,’x1’,’x2’,’x3′)
Datos
```

APPENDIX D

R code for fitting the proposed BLOR for Colombia

This code was used for all the analysis given in Table 3.

```
#####Code for the Bayesian Ordinal REgression ##################################
rm(list=ls()) # remove everything from memory in the working environment.
library(matrixcalc)
library(BayesLogit)
library(mvtnorm)
##### We load the matrix of markers ##########################################
load(’MarkersFinal.RData’)
M=as.matrix(X)
DataOrd1=data.frame(read.table(’GLSdataOsval1.csv’,sep=’,’,h=T));
DataOrd=na.omit(DataOrd1)
DataOrd=DataOrd[order(DataOrd$Stock), ]
DataOrd=subset(DataOrd, Loc==’Colombia’)
y=DataOrd[,4]
XD=DataOrd[,1:4]
##### Calculating the marker-derived genomic relationship matrix (GRM) #########
M<-scale(M,center=TRUE,scale=TRUE)
G<-tcrossprod(M)/ncol(M)
LL=t(chol(G))
##### Incidence matrix and covariance for main eff. of lines ##################
XD$Stock<-factor(x=XD$Stock, levels=rownames(M), ordered=TRUE)
ZL<-model.matrix(~XD$Stock-1)
Z=(ZL%*%LL)
X=Z
tX=t(X)
##### Starting values of Beta0, gammas(thresholds) ############################
gammas<-c(-Inf,-3.11,-2.54,-1.89,-1.27,Inf)
nt=length(gammas)-2
m = dim(X)[1]
nB=dim(X)[2]
In =diag(nB)
p=nB
Betas=rep(0,,dim(X)[2])
SigmaB=1
L<-rep(0,m)
##### Priors for the parameters ###############################################
vB=3
SB=0.001
gam.Min=-1000
gam.Max=1000
##### Number of iterations requiered for the Gibbs sampler ####################
Niter<-60000
#### Matrices for saving the output ###########################################
MBetas<-matrix(nrow=Niter,ncol=nB)
Mgammas<-matrix(nrow=Niter,ncol=nt)
Mvar_Beta<-matrix(nrow=Niter,ncol=1)
#### Function for sampling normal truncated values##############################
ntruncada<-function(lo,hi,media,std)
{
U = runif(m)
F0 = pnorm(lo,media,std)
c = pnorm(hi,media,std)-F0
muestra = qnorm(c*U+F0)*std+media
```

```
 }
##### Gibbs sampler ############################################################
for(i in 1:Niter)
{
##### Linear predictor ##########################################################
eta=as.numeric(Betas)
##### Samples from polya gamma distribution ####################################
##### Replacing w=rpg(num=m,2,-L+eta) with w=rep(1,m) we get the BPOR #######
w=rpg(num=m,2,-L+eta)
D=diag(w)
##### Sample liabilities (L) from a truncated normal distribution ##############
L = ntruncada(gammas[y],gammas[y+1],media=eta,1/sqrt(w))
##### Sample of thresholds #####################################################
newgammas<-numeric(0)
for(k in 2:(nt+1))
{
 lo<-max(c(max(L[y==k-1]),gammas[k-1],gam.Min))
 hi<-min(c(min(L[y==k]),gammas[k+1],gam.Max))
 newgammas<-append(newgammas,runif(1,lo,hi))
 }
 gammas<-c(-Inf,newgammas,Inf)
##### Sample of Betas ###########################################################
C = 1/SigmaB*In
X_tD = tX%*%D
MM=X_tD%*%X
diag(MM)=diag(MM)+diag(C)
S_a=solve(MM,tol=1e-19)
mu_a = S_a%*%(t(X)%*%D%*%L)
Betas=t(rmvnorm(1, mean =mu_a, sigma =S_a))
##### Sample of sigma_Beta#######################################################
 SigmaB = 1/rgamma(1,shape=(vB+p)/2,(SB+tcrossprod(Betas))/2)
##### Saving output #############################################################
MBetas[i,]=t(Betas)
Mgammas[i,]=t(newgammas)
Mvar_Beta[i]=SigmaB

##### Printing some posterior values estimated ################################
cat('Niter',i,'SigmaB =',SigmaB,'Beta1=',Betas[1],'\n')}
```