Bayesian Binary Regression

Paolo Onorati

Sapienza, MEMOTEF departement

Introduction

- We consider the general problem of Bayesian binary regression. In particular we focus on probit and logit model.
- We introduce a new class of distributions, the Perturbed Unified Skew Normal (pSUN), which generalizes the SUN class and show that it is conjugate to any binary regression model, provided that the link function can be expressed as a scale mixture of Gaussian densities.
- We show that, when the number of covariates p is larger than the sample size n, it is possible to produce an exact posterior simulation both in the probit and logit framework.

Modeling Bernoulli Random Variables

We refer to binary regression model when it is assumed a data generating process with Bernoulli random variables as output.

This kind of model is used in many different context, in finance a scope of application is in the granting of loans.

One typically models the $P(Y = 1)$.

Parametric and Semi-parametric Framework

In a parametric, and semi-parametric, framework normally the probability of i-th observation is equal to 1 is assumed to be a function of some covariates via a link-calibration functions structure.

Let $Y \in \{0,1\}^n$, $p \in (0,1)^n$, $X \in \mathbb{R}^{n \times p}$ and X_i be the i-th row of X; so

$$\begin{aligned} Y_i|p &\overset{ind}{\sim} Bern(p_i) \quad \forall i = 1, 2, \ldots, n\,, \\ p_i &= \Lambda(\eta(X_i))\,, \end{aligned}$$

where $\Lambda : \mathbb{R} \to [0,1]$ is a know link function and $\eta(\cdot)$ is a calibration function. Commonly $\Lambda(\cdot)$ is a CDF of some random variable symmetric around 0 and $\eta(\cdot)$ is a linear, i.e.

$$\eta(x) = x'\beta$$

Linear Symmetric Binary Regression

Let $\Lambda_n(x) = \prod_{i=1}^n \Lambda(x_i)$ for $x \in \mathbb{R}^n$, $B_x = 2\mathrm{diag}(x) - I_n$ for $x \in \{0,1\}^n$ with I_n the identity matrix of dimension n. So if $\Lambda(\cdot)$ is symmetric and $\eta(\cdot)$ is linear the likelihood of β is

$$L(y;\beta) = \Lambda_n(B_y X\beta)\,.$$

Even if theoretically all CDFs are legitimate link functions for $\Lambda(\cdot)$ mostly only the standard Gaussian, Gaddum (1933) and Bliss (1934), and standard Logistic, Berkson (1944), ones are used.

There is no closed form for MLE.

Bayesian Setting

Due to the shape of the likelihood in Probit and Logit models, computation of posterior distribution in Bayesian setting has been hard for long time.

Commonly Metropolis-Hastings algorithms were used or others approximate methods as Laplace Approximation, Rasmussen and Williams (2006).

Recently Durante (2019) find a conjugate distribution for the Bayesian probit model.

Conjugate Model for Probit Regression i

Arellano-Valle and Azzalini (2006) establish the so-called Unified Skew Normal family which includes many of the several proposals as generalization of Skew Normal appeared in the literature.

We say $Y \sim \mathrm{SUN}_{d,m}(\tau, \Delta, \Gamma, \xi, \Omega)$ if its density function is

$$f_Y(y) = \varphi_\Omega(y-\xi)\frac{\Phi_{\Gamma-\Delta'\bar{\Omega}^{-1}\Delta}(\tau+\Delta'\bar{\Omega}^{-1}\mathrm{diag}^{-\frac{1}{2}}(\Omega)(y-\xi))}{\Phi_\Gamma(\tau)},$$

where $Y = \xi + \mathrm{diag}^{1/2}(\Omega)X|\,(U+\tau>0)$, $\xi \in \mathbb{R}^d$, Γ is a m-correlation matrix, Ω is a d-covariance matrix, and $\bar{\Omega} = \mathrm{diag}^{-\frac{1}{2}}(\Omega)\,\Omega\,\mathrm{diag}^{-\frac{1}{2}}(\Omega)$, and

$$\begin{bmatrix} X \\ U \end{bmatrix} \sim N_{d+m}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \bar{\Omega} & \Delta \\ \Delta' & \Gamma \end{bmatrix}\right)$$

Notice that the Gaussian distribution is a special case of the SUN.

Conjugate Model for Probit Regression ii

Durante (2019) shows that in Bayesian probit regression if

$$\beta \sim \mathrm{SUN}_{d,m}\left(\tau, \Delta, \Gamma, \xi, \Omega\right)$$

then $\beta | Y$ is

$$\mathrm{SUN}_{d,m+n}\left(\begin{bmatrix} \tau \\ B_Y X \xi \end{bmatrix}, \begin{bmatrix} \Delta' \\ B_Y X \mathrm{diag}^{\frac{1}{2}}(\Omega)\bar{\Omega} \end{bmatrix}', \begin{bmatrix} \Gamma & 0_{m\times n} \\ 0_{n\times m} & I_n + B_Y X \Omega X' B_Y \end{bmatrix}, \xi, \Omega\right).$$

Sampling i.i.d. draws from SUN distribution relies on the acceptance-rejection algorithm of Botev (2017) that works efficiently for small value of $(m+n)$ so for large dataset one must use a Gibbs Sampler.

Extensions

Therefore our aim is to extend the results of Durante (2019) in 2 different ways:

- we introduce a wider class of conjugate family to probit model,
- we extends the result to logistic regression.

For the sake of clarity, we will denote with abuse of notation

$$\operatorname{diag}\left(\begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}\right) = \begin{bmatrix} x_1 & 0 & \dots & 0 \\ 0 & x_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & x_n \end{bmatrix}$$

The Perturbed SUN family

We introduce a new class of distributions, the Perturbed Unified Skew Normal (pSUN), which generalizes the SUN class.

We say $Y \sim pSUN_{d,m}(Q_V, \Theta, A, b, Q_W, \Omega, \xi)$ if

$$Y = \xi + \mathrm{diag}^{\frac{1}{2}}(\Omega)X | (T \leq AX + b) \,,$$

$$X|W \sim N_d(0, \mathrm{diag}^{\frac{1}{2}}(W)\bar{\Omega}\mathrm{diag}^{\frac{1}{2}}(W)) \perp\!\!\!\perp T|V \sim N_m(0, \mathrm{diag}^{\frac{1}{2}}(V)\Theta\mathrm{diag}^{\frac{1}{2}}(V)) \,,$$

$$W \sim Q_W(\cdot) \in \mathbb{R}^d \perp\!\!\!\perp V \sim Q_V(\cdot) \in \mathbb{R}^m \,,$$

where $\bar{\Omega}, \Theta$ are correlation matrices and $Q_W(\cdot), Q_V(\cdot)$ are generic CDFs with non negative values only in the positive orthant.

Conjugate Model

Theorem
In the Bayesian binary regression with linear calibration function, if the link function is a CDF of some random variable that admits a representation as Gaussian scale mixtures, with mean 0*, then the pSUN family is a conjugate family.*

Notice that the SUN is a special case of *pSUN*, but also Generalized Hyperbolic distribution belongs to *pSUN* family, so also Student-T, skew Student-T and Bessel function distribution.

Posterior Sampling

In some situation it is possible get i.i.d. draws from a pSUN distribution but we use a more general algorithm that rely on Gibbs Sampler:

Suppose at time t we have the values Y_t, X_t, T_t, W_t, V_t, we update them in the following way:

Sample $V_{t+1} \sim V | T = T_t$
Sample $W_{t+1} \sim W | X = X_t$
Sample $X_{t+1}, T_{t+1} \sim X, T | T \leq AX + b, W = W_{t+1}, V = V_{t+1}$
  Set $\Sigma_\epsilon = \Theta_V + A\bar{\Omega}_W A'$
  Sample $\epsilon \sim TN_m(-\infty, -b, 0, \Sigma_\epsilon)$
  Set $H_\mu = \bar{\Omega}\mathrm{diag}^{1/2}(W_{t+1})A'\Sigma_\epsilon^{-1}$
  Set $H_\Sigma = \bar{\Omega} - H_\mu A\mathrm{diag}^{1/2}(W_{t+1})\bar{\Omega}$
  Sample $X_{t+1} \sim N_d(H_\mu \epsilon, H_\Sigma)$
  Set $T_{t+1} = AX_{t+1} - \epsilon$
Set $Y_{t+1} = \xi + \mathrm{diag}^{1/2}(\Omega)X_{t+1}$

Logistic Model i

We highlight that the hypothesis of previous theorem are satisfied in the case of Logistic regression, indeed Andrews and Mallows (1974) and Stefanski (1991) prove that if

$$K_0(x) = \sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2x^2)\,, \quad x > 0\,,$$
$$X|K \sim N(0, 4K^2)\,,$$
$$K \sim K_0(\cdot)$$

then X is a standard Logistic distribution.

$K_0(\cdot)$ is the CDF of Kolmogorov distribution that arise in the so-called Kolmogorov-Smirnov test, Kolmogorov (1933), Smirnov (1939).

Logistic Model ii

Notice that in order to sample from $V|T$ in the case of Logit model we must be able to sample from the posterior of Kolmogorov distribution.

We do this via an acceptance-rejection algorithm. It is used as auxiliary density an inverse Gamma distribution, the theoretically acceptance rate is always greater than 70%.

More Covariates Than Observations

A further result holds if there exists the inverse of XX', a necessary but not sufficient condition for this is $p \geq n$.

Theorem
In the Bayesian binary regression with linear calibration function, if link function is a CDF of some random variable symmetric around 0*, and prior on* β *is such that*

$$\beta \overset{d}{=} X'(XX')^{-1}U_2 \ , \ U_{2,i} \overset{i.i.d.}{\sim} G_2^*(\cdot)\,, i = 1, 2, \ldots, n\,,$$

then

$$\beta|Y \overset{d}{=} X'(XX')^{-1}U_2|U_1 \leq B_Y U_2\,,$$
$$U_{1,i} \overset{i.i.d.}{\sim} \Lambda(\cdot)\,.$$

In this case is very simple to get i.i.d. draws from the posterior distribution.

Conclusions

Our aim is improve computational performances in Bayesian binary regression in the case link function is a CDF of some random variable symmetric around 0 and calibration function is linear. Hence:

- we extend the prior conjugate class of *Durante* (2019) to a wider prior family,
- we extend the conjugate model to logistic regression case,
- we find a prior suitable when $p \geq n$ such that it is very simple to get i.i.d. draws.

References

D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1):99–102, 1974.

R. B. Arellano-Valle and A. Azzalini. On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics*, 33(3):561–574, 2006.

Joseph Berkson. Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39(227):357–365, 1944.

C. I. Bliss. The method of probits. *Science*, 79(2037):38–39, 1934.

Z. I. Botev. The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79(1): 125–148, 2017.

Daniele Durante. Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika*, 106(4):765–779, 08 2019.

J. H. Gaddum. Reports on biological standards. 3: Methods of biological assay depending on a quantal response. *Special report series (Medical Research Council (Great Britain))*, 183, 1933.

A. N. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91, 1933.

References ii

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning.* Adaptive computation and machine learning. MIT Press, 2006. ISBN 026218253X.

N. V. Smirnov. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin Mathématique de l'Université de Moscou*, 2(2):3–11, 1939.

Leonard A. Stefanski. A normal scale mixture representation of the logistic distribution. *Statistics & Probability Letters*, 11(1):69–70, 1991.