Kaeding, Matthias

Conference Paper

Flexible Modeling of Binary Data Using the Log-Burr Link

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2015: Ökonomische Entwicklung - Theorie und Politik - Session: Microeconometric Modelling, No. D22-V2

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Kaeding, Matthias (2015) : Flexible Modeling of Binary Data Using the Log-Burr Link, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2015: Ökonomische Entwicklung - Theorie und Politik - Session: Microeconometric Modelling, No. D22-V2, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft

This Version is available at:
https://hdl.handle.net/10419/113043

Flexible Modeling of Binary Data Using the Log-Burr Link

February 27, 2015

Abstract

Popular link functions often fit skewed binary data poorly. We propose the log-Burr link as flexible alternative. The link nests the complementary-log-log and logit link as special cases, determined by a shape parameter which can be estimated from the data. Shrinkage priors are used for the shape parameter, furthermore the parameter is allowed to vary between subgroups for clustered data. For modeling of nonlinear effects basis function expansions are used. The associated regression coefficients are reparameterized as random effects. Inference is done in a fully Bayesian framework. Posterior simulation is done via the No-U-Turn sampler implemented in Stan, avoiding convergence problems of the Gibbs sampler and allowing for easy use of nonconjugate priors. The proposed methods and the effect of misspecification of the modeled dgp are investigated in a simulation study. The approach is applied on large scale unemployment data.
Keywords: Log-burr link; skewed binary data; shrinkage priors

1 Introduction

Most response functions for binary data are given by cdfs:

$$F(x) = \int_{-\infty}^{x} f(x)\, du.$$

Popular choices for $f(x)$ are densities from the location scale family such as the logistic and normal distribution:

$$f(x) = \sigma^{-1} g(\frac{x-\mu}{\sigma}),$$

where σ and μ are usually not identified. The shape of $f(x)$ determines the flexibility of the resulting response function. The most popular choices for $f(x)$ are symmetric and henceforth inappropriate for unbalanced binary data. Data of this kind are e.g. common in discrete duration analysis, panel data or in the context of propensity score weighting. More flexibility is gained by using a distribution indexed by one additional shape parameter which can be estimated from the data. In this paper, the choice of the log-Burr distribution for $f(x)$ is investigated. Here, the response function is given by

$$P(y = 1) = 1 - (1 + \alpha \exp(x))^{(-1/\alpha)}, \tag{1}$$

with shape parameter $\alpha \in (0, \infty)$. For $\alpha < 1$ the pdf is left-skewed, for $\alpha = 1$ symmetric, for $\alpha > 1$ right-skewed. The function nests the complementary-log-log ($\lim \alpha \to 0$) and the logit model ($\alpha = 1$) as special cases. The log-Burr distribution is a special case of the generalized logistic distribution, indexed by one additional parameter, used by Prentice (1976) for differentiating between models. We argue that the log-Burr response function is a good compromise between flexibility and parsimony and can be applied for general use.

The log-Burr link has recently been investigated by Hess (2009) who motivates the use for discrete duration data by the limiting distribution of threshold excesses of a latent continuous duration variable. Hess et al. (2014) show by simulation that misspecification of the response function leads to biased predicted probabilities. In this paper existing modeling approaches are extended, taking account the importance of the shape parameter which has not been done in this form. Following extensions are given: (1) The shape parameter is modeled using sparsity priors, allowing shrinkage in the case of high variance. (2) The shape parameter is allowed to vary between clusters, allowing flexible modeling for datasets containing left/right and non-skewed clusters, while stabilizing estimates by borrowing information between clusters. (3) Nonlinear effects can be estimated using P-splines and other basis function expansions.

Inference is fully Bayesian. The advantages hereof are: (1) Variance estimation accounts for the uncertainty in all parameters. (2) Estimation is simple using MCMC methods. (3) Functions (and variance estimation thereof) of parameters, e.g. marginal effects can be directly estimated. (4)

Bayesian shrinkage priors apply naturally to the (log) shape parameter. Posterior simulation is done using The No-U-Turn sampler sampler implemented in Stan, a variant of Hamiltonian Monte Carlo (HMC). Hamiltonian Monte Carlo updates all parameters in one block using gradient information, thus avoiding convergence problems of Gibbs-samplers and allowing easy use of non-conjugate priors. However, HMC depends on two tuning parameters. The No-U-Turn sampler sets these parameter adaptively and is henceforth fully automatic. Covariate effects can be measured by marginal effects or the method of Chib and Jeliazkov (2006).

2 Model Formulation

We are working with $\gamma = \log \alpha$ to bypass the positivity restriction, so that

$$P(y_i = 1) = r_i = 1 - (1 + \exp(\gamma + \eta_i))^{-\exp(-\gamma)},$$

with linear predictor $\eta_i = \boldsymbol{z}_i^\top \boldsymbol{\beta}$. The posterior distribution is given by

$$\pi(\boldsymbol{\theta}|\mathcal{D}) \propto \pi(\boldsymbol{\theta})L = \pi(\boldsymbol{\theta}) \prod_{i=1}^{n} r_i^{y_i}(1 - r_i)^{1-y_i},$$

where $\boldsymbol{\theta}$ is the vector of all parameters with prior π, $\mathcal{D}$ is the data and L is the likelihood. Following Fahrmeir and Kneib (2011), the linear predictor can be extended to include nonlinear effects, so we have

$$\eta_i = \boldsymbol{z}_i^\top \boldsymbol{\beta} + f_1(\boldsymbol{x}_{1i}) + ... + f_p(\boldsymbol{x}_{pi}). \tag{2}$$

The functions $f_j, j = 1, ..., p$ are modeled via P-splines, where

$$f_j(\boldsymbol{x}_j) = \boldsymbol{B}(\boldsymbol{x}_j)\boldsymbol{\Xi}_j,$$

$\boldsymbol{B}(\boldsymbol{x}_j)$ is a matrix of evaluations of basis functions corresponding to B-splines, given by $\mathcal{B}_j^l(\cdot)$, so that $\boldsymbol{B}_{ij} = \mathcal{B}_j^l(x_i)$, where l is the degree of the spline. The case l=0 corresponds to a step function. More on Splines can be found in Dierckx (2006). The mean level of $f_j, j = 1, ..., p$ is not identified, so that identification restrictions are necessary.

For a hierarchical data structure with J cluster, we have

$$L = \prod_{j=1}^{J} \prod_{i=1}^{n_j} r_{ij}^{y_{ij}} (1 - r_{ij})^{1-y_{ij}}.$$

Here, the shape parameter is allowed to vary between cluster, so that $\gamma = (\gamma_1, ..., \gamma_J)$.

2.1 Priors

The prior distribution has the form

$$\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma}|\gamma_0)\pi(\gamma_0)\pi(\boldsymbol{\Xi}_1|\xi_1^2)\pi(\xi_1^2)...\pi(\boldsymbol{\Xi}_p|\xi_p^2)\pi(\xi_p^2).$$

Conditionally, all priors for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\Xi}_i, i = 1, ..., p$ have the generic form

$$\pi(\boldsymbol{x}|\boldsymbol{x}_0, v_x, \boldsymbol{P}) \propto \exp(-\frac{1}{2v_x}(\boldsymbol{x} - \boldsymbol{x}_0)^\top \boldsymbol{P}(\boldsymbol{x} - \boldsymbol{x}_0)), \tag{3}$$

with varying form of (possibly rank-deficient) penalty matrix P controlling the form of penalization of some form of prior information while v_x controls the degree of penalization and can be assigned a hyperprior to allow data driven penalization. To implement possible identification constraints, the prior can be adjusted to $\pi(\boldsymbol{x}|\boldsymbol{v}_x, x_0, \boldsymbol{P})I[\boldsymbol{A}\boldsymbol{x} = 0]$ where $I[\cdot]$ is the indicator function. The different choices are given in the following.

2.1.1 Linear effects

For linear effects we consider informative priors $N(\boldsymbol{m}_0, \boldsymbol{\Sigma}_0)$ and noninformative priors $\beta \propto 1$. The former case corresponds to $v_x = 1$, $\boldsymbol{x}_0 = \boldsymbol{m}_0, \boldsymbol{P}_0 = \boldsymbol{\Sigma}_0^{-1}$ while the latter corresponds to the limit case $\boldsymbol{\Sigma}_0^{-1} = \boldsymbol{0}$, where in an abuse of notation $\boldsymbol{0}$ denotes a matrix of zeroes.

2.1.2 Nonlinear effects

To avoid overfitting, overly rough function estimates are penalized. A good choice is the difference penalty by Eilers and Marx (1996), extended to the Bayesian case by Lang and Brezger (2004). Here $\Delta^k \beta \sim N(0, \xi^2)$, where Δ^k is the difference operator of order k. $\Xi_j, j \leq k$ are assigned a flat prior $\Xi_j \propto const$. The penalty matrix is given by $\boldsymbol{K} = \boldsymbol{D}^\top \boldsymbol{D}$ with difference

matrix $\boldsymbol{D} = \boldsymbol{D}_k = \boldsymbol{D}_1\boldsymbol{D}_{k-1}$, where

$$\boldsymbol{D}_1 = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{bmatrix}.$$

For $\pi(\xi)$, the half-cauchy distribution is used as recommended by Gelman (2006) and Nicholas G. and Scott (2011).

2.1.3 Shape parameters

- General idea: Shrinkage of γ to zero to avoid overfitting, or when estimates became unstable

3 Computation and Inference

The main building block for posterior simulation is HMC. HMC updates sample from the joint distribution

$$\pi(\boldsymbol{\theta}|\mathcal{D})N(\boldsymbol{p}|\boldsymbol{0}, \boldsymbol{M}^{-1})$$

via simulating Hamiltonian Dynamics by taking L leap-frog steps (lines 5-13 in Algorithm 1) with stepsize ϵ steps in γ and p. The matrix M is usually referred to as mass matrix. The obtained samples of u are ignored as u is just an auxillary variable introduced to simplify sampling. In Hoffman and Gelman (2014), an adaptive method for determiniation of tuning parameters L and ϵ is given, implemented in the software RStan (Stan Development Team 2014) which is used here.

Algorithm 1: Hamiltonian Monte Carlo

set θ^0;
for $s \leftarrow 2$ **to** S **do**
  sample $p^\star \sim N(0, M)$;
  $\boldsymbol{\theta}_0 \leftarrow \boldsymbol{\theta}^{(s)}$;
  $p_0 \leftarrow p^\star + (\epsilon/2)\nabla(\boldsymbol{\theta}_0)$;
  for $l \leftarrow$ **to** L **do**
    $\boldsymbol{\theta}_l = \boldsymbol{\theta}_l + (\epsilon)p_{l-1}$;
    if $(l \neq L)$ **then**
      $p_l = p_{l-1} + \epsilon\nabla(\boldsymbol{\theta}_l)$;
    else
      $p_L = p_{L-1} + (\epsilon/2)\nabla(\boldsymbol{\theta}_L)$;
    end
  end
  $\alpha(\boldsymbol{\theta}^{(s)}, \boldsymbol{\theta}^\star)) \leftarrow \min(1, \exp\{\log(\frac{(\theta^\star|\mathcal{D})}{(\theta^{(s)}|\mathcal{D})}) + \frac{1}{2}(p_L^\top M p_L - p^{\star\top} M p^\star)\}$;
  sample U$\sim Unif(0,1)$;
  if $U < \alpha(\boldsymbol{\theta}^{(s)}, \boldsymbol{\theta}^\star))$ **then**
    $\boldsymbol{\theta}^{(s+1)} \leftarrow \boldsymbol{\theta}^\star$;
  else
    $\boldsymbol{\theta}^{(s+1)} \leftarrow \boldsymbol{\theta}^{(s)}$;
  end
end

For regression coefficients associated with basis functions, a reparameterization as mixed model is used:

$$\boldsymbol{\Xi}_j = \boldsymbol{A}_j \boldsymbol{b}_j + \boldsymbol{U}_j \boldsymbol{t}_j,$$

with priors $\boldsymbol{b}_j \propto \mathbf{1}$ and $\boldsymbol{t}_j \sim N(\mathbf{0}, \xi^2 \boldsymbol{I})$, corresponding to the unpenalized and penalized part of $\boldsymbol{\Xi}_j$, see Kneib (2006). For $\boldsymbol{U}_j$ we can set $\boldsymbol{D}_k^\top (\boldsymbol{D}_k \boldsymbol{D}_k^\top)^{-1}$. For $\boldsymbol{A}_j$ a basis of the null space of $\boldsymbol{K}_j$ given by matrices of the form

$$\begin{bmatrix} 1 & 1 & \dots & 1^{k-1} \\ 1 & 2 & \dots & 2^{k-1} \\ \vdots & \vdots & & \vdots \\ 1 & \dim(\Xi_j) & & \dim(\Xi_j)^{k-1} \end{bmatrix},$$

where any sequence of equidistant values can be used.

3.1 Parameter Interpretation

Linear and nonlinear effects are not directly interpretable. As such, the posterior mean of the marginal effects are used. Furthermore, we can marginalize over the covariates whose effect we are not interested in using the method of Chib and Jeliazkov (2006).

4 Application

4.1 Data

The Sample of Integrated Labour Market Biographies (SIAB) (Berge et al. 2013) of the Institute for Employment Research (IAB) is used to illustrate the above-presented method. The SIAB is a 2 percent random sample of all Germans that are employed (subject to social securoty contributions) between 1975 and 2010, receiving unemployment benefits or are officially registered as job seeking. The data therefore capture the majority of the German workforce (1,639,325 individuals). Due to the comprehensive nature and the longitudinal character the data is well suited to illustrate the proposed method.

4.2 Results

5 Simulation study

Here, results of a simulation study will be reported. The following questions will be answered:

- Does the proposed methodology work?
- What is the result of misspeficiation of the dgp? That is, of using a logit/probit model if the underlying response function is the log-Burr cdf and/or if the shape parameter varies by cluster.
- Does the proposed methodology finds the case where a simpler model is enough?

6 Conclusion

Bibliography

Berge, P. vom, M. König, and S. Seth (2013). *Sample of Integrated Labour Market Biographies (SIAB) 1975-2010*. Tech. rep. Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg [Institute for Employment Research, Nuremberg, Germany].

Chib, S. and I. Jeliazkov (2006). "Inference in Semiparametric Dynamic Models for Binary Longitudinal Data". In: *Journal of the American Statistical Association* 101, pp. 685–700.

Dierckx, P. (2006). *Curve and surface splitting with splines*. Oxford and New York: Clarendon Press.

Eilers, P. H. C. and B. D. Marx (1996). "Flexible smoothing with B-splines and penalties". In: *Statistical Science* 11, pp. 89–121.

Fahrmeir, L. and T. Kneib (2011). *Bayesian smoothing and regression for longitudinal, spatial and event history data*. Vol. 36. Oxford and New York: Oxford University Press.

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models". In: *Bayesian Analysis* 1, pp. 1–19.

Hess, W. (2009). *A Flexible Hazard Rate Model for Grouped Duration Data*. Working Papers. Lund University, Department of Economics.

Hess, W., G. Tutz, and J. Gertheiss (2014). *A Flexible Link Function for Discrete-Time Duration Models*. Tech. rep.

Hoffman, M. D. and A. Gelman (2014). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". In: *Journal of Machine Learning Research* 15, pp. 1351–1381.

Kneib, T. (2006). "Mixed model based inference in structured additive regression".

Lang, S. and A. Brezger (2004). "Bayesian P-Splines". In: *Journal of Computational and Graphical Statistics* 13, pp. 183–212.

Nicholas G., P. and N. Scott (2011). "On the half-Cauchy prior for a global scale parameter". In: *Pre-print*.

Stan Development Team (2014). *Stan: A C++ Library for Probability and Sampling, Version 2.5.0*.