A Joint Credit Scoring Model for Peer-to-Peer Lending and Credit Bureau

Raffaella Calabrese
Credit Research Centre and University of Edinburgh
raffaella.calabrese@ed.ac.uk

joint work with Silvia Osmetti and Luca Zanin

Credit Scoring and Credit Control conference
31 August 2017

Outline

P2P lending

- Peer-to-peer (P2P) lending allows direct lending between lenders and borrowers using a platform.
- In 2014 P2P lending generated approximately $5.5 billion loans in the US.
- To improve the predictive accuracy accuracy of scoring models for P2P, we suggest to use the information from a credit bureau if the borrower defaults on any loan.
- We propose a bivariate regression model binary unbalanced data (BivGEV model).

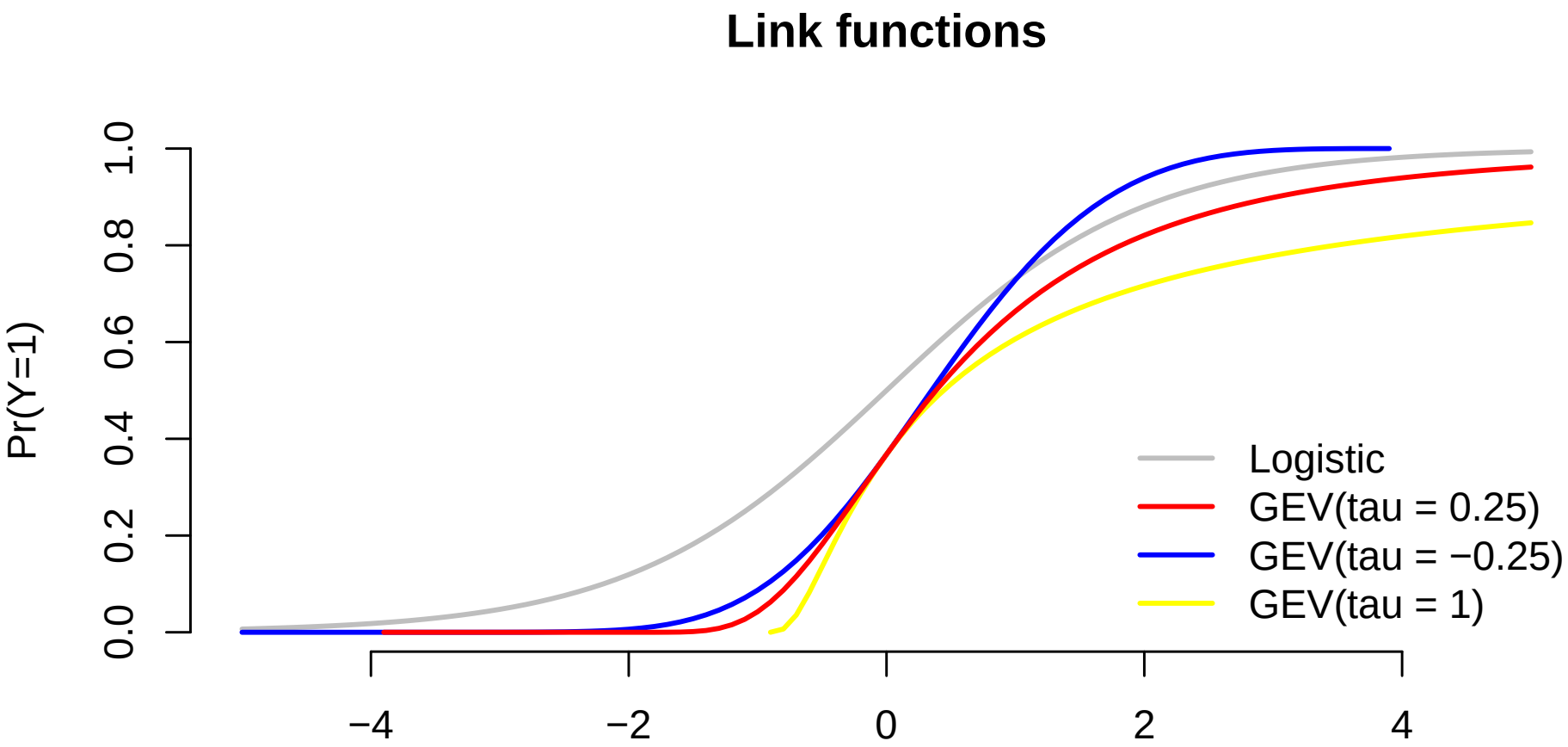
The univariate model

- Let Y be a binary response so defined

$$Y = \begin{cases} 1 & \text{if the borrower defaults} \\ 0 & \text{otherwise} \end{cases}$$

- Let $\mathbf{x} = (x_1, x_2, ..., x_p)$ be a p-covariates vector.
- We model the probability of default $P(Y = 1) = \pi(\mathbf{x}'\beta)$ where $\beta = (\beta_0, \beta_1, ..., \beta_p)$ are the regressor parameters.
- Symmetric link functions $\pi(\cdot)$, such as the logit and the probit models, are inaccurate if the binary classification is strongly unbalanced (Calabrese et al. 2015; King and Zeng, 2001; Wang and Dey, 2010).

GEV link function



The univariate model

We suggest to model the probability of default $\pi(\mathbf{x}'\beta)$ using the GEV distribution as follows

$$\pi(\mathbf{x}_{it}, s_i, t_i) = \pi(\mathbf{x}; \beta, \tau) =$$

$$= \begin{cases} \exp\left\{-\left[1+\tau(\beta_0+\sum_{j=1}^{p}\beta_j x_j)\right]_+^{-\frac{1}{\tau}}\right\} & \tau \neq 0 \\ \exp\left[-(\beta_0+\sum_{j=1}^{p}\beta_j x_j)\right] & \tau = 0 \end{cases}$$

with τ denotes the shape parameter and $x_+ = max(x,0)$.
The GEV distribution is very flexible with the shape parameter τ controlling the tail behaviour.
The R package BGEVA is available on CRAN.

The copula function

A function $C : I^2 \to I$, with $I^2 = [0,1] \times [0,1]$ and $I = [0,1]$, is a **bivariate copula** if it is the cumulative bivariate distribution function of a rv (U,V), with uniform marginal in [0,1]

$$C_\lambda(u,v) = P(U \leq u, V \leq v), \quad 0 \leq u \leq 1 \quad 0 \leq v \leq 1$$

where the copula parameter $\lambda \in \Lambda$ describes the association between the marginals.
Copula functions capture the dependence structure between the marginals and allow the specification of multivariate distributions with arbitrary dependence structures.

Some characteristics of the main Copula functions

Copula	Dependence	Tail Dependence
Gaussian	radially symmetric	no asymptotic tail dependence
Clayton	asymmetric (exchangeable)	strong left (lower) tail dependence
Gumbel	asymmetric (exchangeable)	strong right (upper) tail dependence
Frank	radially symmetric	no asymptotic tail dependence
Joe	asymmetric (exchangeable)	strong right (upper) tail dependence

The BivGEV model

- $\mathbf{Y} = (Y_1, Y_2)$ is a binary bivariate response variable with values on (0,1);
- the marginal probabilities are
  $\pi_1(\mathbf{x}; \beta_1, \tau_1) = P(Y_1 = 1 | \mathbf{x}; \beta_1, \tau_1)$
  $\pi_2(\mathbf{x}; \beta_2, \tau_2) = P(Y_2 = 1 | \mathbf{x}; \beta_2, \tau_2)$
- The marginal probabilities are modelled using the GEV distribution.
- The BivGEV is defined using the copula function:

$$\pi_{11}(\mathbf{x}; \delta, \tau) = C_\lambda(\pi_1(\mathbf{x}; \beta_1, \tau_1), \pi_2(\mathbf{x}; \beta_2, \tau_2))$$
$$= C_\lambda\left(exp\left\{-\left[1 + \tau_1\eta_1\right]^{-1/\tau_1}\right\}, exp\left\{-\left[1 + \tau_2\eta_2\right]^{-1/\tau_2}\right\}\right)$$

- The maximum likelihood method is used to estimate the BivGEV model.

Data

We analyse 12,579 loans of 60 months provided by Lending Club from 2010 to the first quarter of 2012.

$$Y_1 = \begin{cases} 1 & \text{if the borrower is reported in default by the credit bureau} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_2 = \begin{cases} 1 & \text{if the borrower defaults on the P2P loan} \\ 0 & \text{otherwise} \end{cases}$$

The percentage of defaulted P2P loans is 24% and default credit bureau is 5%.

The determinants of the scoring models for default credit bureau and P2P lending are:

- *Loan purpose*.
- *Housing situation*: Mortgage; Rent; Own or other situation.
- *Interest rate*.
- *Annual income*.
- *Revolving utilization*.
- *Inquiries last 6 months*.
- *DTI*: Monthly debt payments to monthly income.
- *Delinquency last 2 years*.
- *Open accounts*.
- *Credit history length*.
- *Loan amount to annual income*.
- *Spatial variables defined using the first digit of the ZIP Code*.

Empirical results

Copula	*Copula parameter* λ	*Kendall-Tau*
Gaussian	0.147	0.094
Clayton	0.104	0.049
Gumbel	1.150	0.132
Frank	1.050	0.115
Joe	1.480	0.210

Copula	AIC	BIC
Gaussian	12325.60	12538.08
Clayton	12325.58	12538.06
Gumbel	12325.51	12537.99
Frank	12325.91	12538.39
Joe	12326.36	12538.84

	Default Credit Bureau	Default P2P lending
Car financing		-0.157^{**}
House		-0.094^{*}
Major purchase	-0.576^{**}	
Small business		0.379^{**}
Rent	-0.320^{**}	
Interest rate	0.123^{**}	0.055^{**}
ln(Annual income)	-0.727^{**}	-0.313^{**}
ln(Revolving utilization)	0.145^{**}	0.048^{**}
Inquiries last 6 months		0.068^{**}
Delinquency last 2 years	-0.137^{*}	
Open accounts	0.021^{**}	
DTI	-0.021^{**}	
Credit history length	0.026^{**}	0.007^{**}
Loan amount to annual income	-2.37^{**}	0.301^{**}
Intercept	4.298^{**}	1.848^{**}
τ	-0.8	-0.1

Out of sample

Model	MSE_+	MAE_+	AUC	H
Probit	0.5555	0.7392	0.6190	0.0558
		$Y_2 = 1 \mid Y_1 = 1$		
Model	MSE_+	MAE_+	AUC	H
BivGEV	0.3792	0.6109	0.5969	0.1529
BivProbit	0.3805	0.6117	0.5930	0.1520
		$Y_2 = 1 \mid Y_1 = 0$		
Model	MSE_+	MAE_+	AUC	H
BivGEV	0.5654	0.7465	0.6200	0.0783
BivProbit	0.5656	0.7463	0.6198	0.0788

Out of time

Model	MSE_+	MAE_+	AUC	H
Probit	0.5570	0.7407	0.6671	0.0790
		$Y_2 = 1 \mid Y_1 = 1$		
Model	MSE_+	MAE_+	AUC	H
BivGEV	0.3616	0.5975	0.7897	0.3907
BivProbit	0.3629	0.5984	0.7910	0.3910
		$Y_2 = 1 \mid Y_1 = 0$		
Model	MSE_+	MAE_+	AUC	H
BivGEV	0.5657	0.7472	0.6642	0.1238
BivProbit	0.5661	0.7471	0.6637	0.1197

Conclusions

- We introduced a bivariate regression model that is accurate in classifying defaults.
- We implemented the model in an R package that will be publicly available.
- We obtain that using the information from the credit bureau improves the predictive accuracy of a scoring model for P2P lending.