A comparative analysis of the UK and Italian small businesses using Generalized Extreme Value model

Dr Galina Andreeva, University of Edinburgh
Dr Raffaella Calabrese, University of Essex
Dr Silvia Angela Osmetti, Catholic University, Milan

Acknowledgment: Royal Society of Edinburgh for sponsoring this research

Outline

- Research questions
- Drawbacks of logistic regression model
- GEV and BGEVA models
- SME definition
- Data description
- Preliminary results
- Work in progress and further research

Research Questions

- How different/similar are significant predictors of SME insolvency between the UK and Italy?
- How different is predictive power of financial statement variables in the UK and Italy?
- How Generalized Extreme Value Models (GEV and BGEVA, Calabrese and Osmetti, 2013; Calabrese et al, 2013) compares to logistic regression?

Generalised Linear Models

Let Y be a binary r.v. of parameter $PD = P\{Y = 1\}$.
A Generalized Linear Model (GLM) considers a monotonic and twice differentiable function $g(\cdot)$, called *link function*, and an explanatory variable vector $\mathbf{x}$ such that

$$g(PD) = \mathbf{x}'\boldsymbol{\beta}.$$

assumption: the explanatory variables provide a set of *linear* predictor $\boldsymbol{\eta} = \mathbf{x}'\boldsymbol{\beta}$

drawback: the assumption of linearity is not satisfied in credit scoring (Thomas et al., 2002).

Logit Model

- The logit link is the most polular in credit scoring
- It is *symmetric* about 0.5

$$logit[PD(x)] = \ln\left[\frac{PD(x)}{1-PD(x)}\right] = -logit[1-PD(x)]$$

- The default is a **rare event** (default percentage is lower than 5%).

$\rightarrow$ The defaults are more informative than non-defaults.

$\rightarrow$ A symmetric link function is not appropriate for scoring models.

drawback: The PD is *underestimated* (Calabrese and Osmetti, 2013).

GEV distribution

To focus the attention on the tail of the response curve for the values close to 1, Calabrese and Osmetti (2013) propose to use the quantile function of the Generalized Extreme Value (**GEV**) r.v. as link function.
The cumulative distribution function is

$$F_X(x) = \exp\left\{-\left[1-\tau\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\tau}}\right\}$$

defined on $S_X = \{x : 1+\tau(x-\mu)/\sigma > 0\}$,
with $-\infty < \tau < \infty$, $-\infty < \mu < +\infty$ and $\sigma > 0$.

- τ is the shape parameter,
- μ and $\sigma(>0)$ are location and scale parameters, respectively.

GEV distribution

- for $\tau \to 0$, Type I class of extreme value distribution (Gumbel-type distribution)

$$F_X(x) = \exp(\exp(-x)) \quad -\infty < x < \infty$$

- for $\tau > 0$, Type II class of extreme value distribution (Fréchet-type distribution)

$$F_X(x) = \exp(-x^{-k}) \quad k = \frac{1}{\tau} > 0 \quad x > 0$$

- for $\tau < 0$, Type III class of extreme value distribution (Weibull-type distribution)

$$F_X(x) = \exp(-(-x)^{k}) \quad k = \frac{1}{\tau} > 0 \quad x < 0$$

GAM model

A Generalized Additive Model (GAM) (Hastie and Tibshirami, 1990) differs from a GLM in that an *additive predictor* replaces the linear predictor

$$g(PD) = \alpha + \sum_{j=1}^{p} s_j(x_j),$$

where α is a constant, s_j with $j = 1, 2, \ldots, p$ are arbitrary smooth functions.

BGEVA model

Calabrese *et al.* (2013) propose a new GAM model whose link function is the quantile function of a GEV r.v.

$$\frac{\{-ln[PD(x)]\}^{-\tau} - 1}{\tau} = \alpha + \sum_{j=1}^{p} s_j(x_j). \qquad (1)$$

called Binary Generalized Extreme Value Additive (BGEVA) model.

BGEVA R package

- GEV and BGEVA models are implemented in the **BGEVA package** (Calabrese et al., 2013) available in R-CRAN.
- The BGEVA package provides
  - parametric and smooth component summaries for logistic and log-log additive models
  - parametric and smooth component summaries for BGEVA model
  - approximate p-values for testing individual smooth terms for equality to the zero function are also provided for BGEVA model
  - plots of the smooth component estimates with confidence intervals for BGEVA model
  - variable selection via the use of shrinakge smoothers or information criteria for BGEVA model.

SME definition

- EU Enterprise and Industry Directorate defines SMEs in terms of number of employees, annual turnover or annual balance sheet volume http://ec.europa.eu/enterprise/policies/sme/facts-figures-analysis/sme-definition/ index _en.htm

- The above definition is adopted, i.e. < 250 employees, < € 50 mln annual turnover /operating revenue

- Consistent with Basel II definition

- Basel II has introduced a special treatment for SMEs:
  Banks that manage small-business-related exposures in a manner similar to retail exposures may apply the less capital requiring 'retail treatment' to such exposures

Data description

- Amadeus database
- Removed those with missing values for # employees and Operating revenue
- Additional exclusions:
  - <11 Directors
  - < 5 Subsidiaries
  - Global Ultimate Owner should have <250 employees and < € 50 mln annual turnover /operating revenue
  - The largest shareholder should have <250 employees
- 2008 financial statements used to predict insolvency in 2009

Default definition

	ITALY	UNITED KINGDOM	Total	Default 1	Default 2
Active	196704	19724	216428		
Active (dormant)	0	976	976	Good	Good
Active (default of payments)	1015	5	1020		
Active (insolvency proceedings)	47	235	282	Excluded	
Bankruptcy	3112	0	3112		
Dissolved	2953	1056	4009		
Dissolved (bankruptcy)	52	0	52		
Dissolved (liquidation)	1576	0	1576		
In liquidation	2917	268	3185		
Inactive (no precision)	0	1	1	Bad	Bad
Total	208376	22265	230641		
% Bad Default1	5.09%	5.95%	5.17%		
% Bad Default2	5.12%	6.02%	5.20%		

Default definition

Italy			
	AUC	**MAE**	**MSE**
Default1 model on Default1 control	0.7290	0.0781	0.0391
Default2 model on Default2 control	0.7193	0.0850	0.0426
Default2 model on Default1 control	0.7252	0.0816	0.0392
UK			
	AUC	**MAE**	**MSE**
Default1 model on Default1 control	0.6120	0.0899	0.0452
Default2 model on Default2 control	0.5918	0.1094	0.0550
Default2 model on Default1 control	0.5974	0.0998	0.0453

- Using Logistic regression, Default 1 definition is preferred
- Additional exploration possible when more years of data are available

Explanatory variables

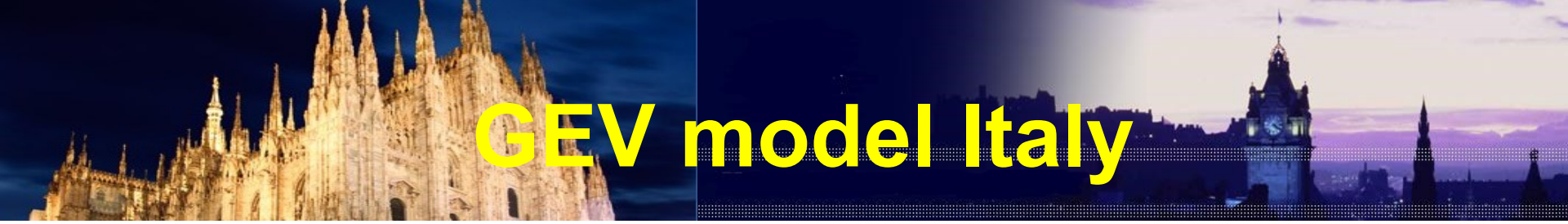
Following Altman and Sabato, 2006; Altman *et al*., 2010; Lin *et al*. 2012

1. Operating revenue
2. Total Assets
3. Profit/ Loss before Tax
4. Net Income
5. Shareholder Funds
6. Current Ratio
7. Liquidity Ratio
8. Loans
9. Capital
10. Current Liabilities
11. Noncurrent Liabilities
12. Cash Flow
13. Solvency Ratio
14. Profit per Employee
15. Profit Margin
16. EBITDA Margin
17. No Employees
18. Age of Company

Logistic regression Italy

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-2.29E+00	2.45E-02	-93.608	< 2e-16***
No Employees	2.92E-03	9.05E-04	3.232	0.001231**
Operating Revenue	-4.84E-05	6.20E-06	-7.796	6.38E-15***
Shareholder Funds	-2.29E-04	3.35E-05	-6.84	7.94E-12***
Liquidity Ratio	4.05E-02	4.17E-03	9.734	< 2e-16***
Loans	5.75E-05	1.68E-05	3.423	0.000618***
Capital	2.35E-04	3.46E-05	6.793	1.10E-11***
Current Liabilities	4.66E-05	7.12E-06	6.541	6.10E-11***
Noncurrent Liabilities	-7.57E-05	1.79E-05	-4.22	2.45E-05***
Cash Flow	-5.66E-04	6.83E-05	-8.29	< 2e-16***
Solvency Ratio	-1.57E-02	6.64E-04	-23.627	< 2e-16***
Profit Margin	-1.02E-02	1.66E-03	-6.126	9.03E-10***
EBITDA Margin	-9.89E-03	1.83E-03	-5.411	6.25E-08***
Age of Company	-3.26E-03	1.39E-04	-23.538	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



GEV model Italy

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-1.28E+00	1.76E-02	-72.536	< 2e-16***
No Employees	2.42E-03	6.12E-04	3.945	7.97E-05***
Operating Revenue	-4.43E-05	4.11E-06	-10.781	< 2e-16***
Shareholder Funds	-1.82E-04	1.72E-05	-10.562	< 2e-16***
Liquidity Ratio	2.76E-02	2.29E-03	12.062	< 2e-16***
Loans	-2.97E-06	6.00E-06	-0.495	0.52
Capital	1.96E-04	1.65E-05	11.868	< 2e-16***
Current Liabilities	5.81E-05	2.82E-06	20.566	< 2e-16***
Noncurrent Liabilities	-5.20E-05	9.76E-06	-5.325	1.01E-07***
Cash Flow	-3.95E-04	2.23E-05	-17.732	< 2e-16***
Solvency Ratio	-1.15E-02	4.38E-04	-26.143	< 2e-16***
Profit Margin	-7.30E-03	1.18E-03	-6.187	6.12E-10***
EBITDA Margin	-6.84E-03	1.21E-03	-5.632	1.78E-08***
Age of Company	-2.44E-03	1.02E-04	-23.839	< 2e-16***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1; τ = -0.8

Logistic regression UK

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-2.39E+00	7.43E-02	-32	< 2e-16***
Operating Revenue	-1.93E-05	5.36E-06	-3.6	0.000319***
Current Ratio	1.98E-02	6.35E-03	3.118	0.001824**
Loans	3.38E-06	1.37E-06	2.458	0.013977*
Solvency Ratio	-3.81E-03	1.22E-03	-3.113	0.001849**
EBITDA Margin	-4.41E-03	1.82E-03	-2.428	0.015195*
Age of Company	-2.15E-03	3.22E-04	-6.669	2.58E-11***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

GEV model UK

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-1.40E+00	6.25E-02	-22.387	< 2e-16***
Operating Revenue	-1.61E-05	4.55E-06	-3.546	0.000392***
Current Ratio	1.65E-02	5.00E-03	3.284	0.001025**
Loans	2.45E-06	6.85E-07	3.578	0.000347***
Solvency Ratio	-3.19E-03	1.03E-03	-3.101	0.001931**
EBITDA Margin	-3.67E-03	1.53E-03	-2.395	0.016606*
Age of Company	-1.84E-03	2.75E-04	-6.696	2.14E-01***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1; τ = -0.9

Out-of-sample Performance

	Italy		UK	
	Logistic	GEV	Logistic	GEV
AUC	0.729	0.729	0.612	0.612
MAE	0.0781	0.079	0.0899	0.0901
MSE	0.0391	0.0391	0.0452	0.0452
MAE on defaults	**0.9216**	**0.9212**	**0.9463**	**0.9457**
MSE on defaults	**0.8553**	**0.8537**	**0.8959**	**0.8948**

Work in progress and further research

- Fitting BGEVA model with splines to account for non-linearity
- Investigation of different methods of handling missing values
- Considering categorical variables
- Extending the comparison to years 2001-2011
- Considering discrete time survival analysis model.

Bibliography

- Altman, E., Sabato, G. 2006. Modeling Credit Risk for SMEs: Evidence from the US Market. ABACUS, 19(6), 716-723 .
- Altman, E., Sabato, G., Wilson, N. 2010.The value of non-financial information in small and medium-sized enterprise risk management. Journal of Credit Risk. 6 (2), p.95-127.
- Calabrese R., Osmetti S., 2013. Modelling small and medium enterprise loan defaults as rare events: the generalized extreme value regression model. Journal of Applied Statistics. 40 (6), p. 1172-1188.
- Calabrese, R., Marra, G. and Osmetti, S. A. (2013). BGEVA: Binary Generalized Extreme Value Additive Models. R package version 0.1.
- Hastie, T.J., Tibshirami R.J. 1990. Generalized Additive Models. Chapman & Hall, Boca Raton.
- Lin, S.M., Ansell, J., Andreeva, G., 2012. Predicting default of a small business using different definitions of financial distress. Journal of Operational Research Society, 63, 539-548.
- Thomas, L., Edelman, D., Crook, J.C., 2002. Credit Scoring and Its Applications. SIAM, Philadelphia.