Generalized Extreme Value Regression: an Application to Credit Defaults

Calabrese, Raffaella
University College Dublin
Casl, Belfield
Dublin 4, Ireland
raffaella.calabrese@ucd.ie

Osmetti, Silvia Angela
Università Cattolica, Dipartimento di Scienze statistiche
Via Necchi 9
Milano (20123), Italy
E-mail: silvia.osmetti@unicatt.it

Introduction

We aim at proposing a Generalized Linear Model (GLM) with binary dependent variable Y, whose link function defined by the Generalized Extreme Value (GEV) distribution. We define this model as GEV regression. The goal of this paper is to overcome the drawbacks shown by the logistic regression in rare events: the probability of rare events is underestimated and the logit link is a symmetric function.
Let Y denote a binary response variable and let X be an explanatory variable, the logistic response curve is

$$\pi(x) = \frac{\exp(\alpha_0 + \alpha_1 x)}{1 + \exp(\alpha_0 + \alpha_1 x)},$$

with link function is $logit(\pi(x)) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha_0 + \alpha_1 x$. The maximum likelihood method is usually used to estimate the parameters α_0 and α_1. The ordinary logistic regression shown same drawbacks when we study rare event data (binary dependent variables with a very small number of ones than zero). Firstly, when the dependent variable represents a rare event, the logistic regression could underestimate the probability of occurrence of the rare events. Secondly, commonly used data collection strategies are inefficient for rare event data (King Zeng, 2001). In order to overcome this drawback the choice-based or endogenous stratified sampling (case-control design) is used. The strategy is to select on Y by collecting observations for which $Y = 1$ and a random selection of observations for which $Y = 0$. This sampling method is usually supplemented with a prior correction of the bias of MLE estimators (McCullagh and Nelder, 1989) show a analytical approximation for the bias in the MLE estimates to account for finite sample. This bias is amplified in application with rare events.
Thirdly, the logit link is symmetric about 0.5

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

This means that the response curve for $\pi(x)$ approaches 0 at the same rate it approaches 1. If the dependent variable represents a rare event, a symmetric link function is not appropriated. Since a rare event is usually modeled by a Poisson distribution, which has positive skewness, it is coherent to choose a asymmetric link function in order to obtain a response curve that approaches 0 at a different rate it approaches 1.
In the extreme value theory the GEV distribution is used to model the tail of a distribution (Kotz Nadarajah, 2000). Since we focus our attention on the tail of the response curve for the values close to 1, we chose the GEV distribution (Kotz Nadarajah, 2000). The Gumbel and Weibull distributions represent particular cases of the GEV distribution. In GLM (Agresti, 2002) the log-log link function, related to the Gumbel distribution, is used since it is asymmetric. In this paper we generalized the previous models by the GEV distribution, so we define the GEV regression. After computing the likelihood and the score functions, we estimate the parameters by the maximum likelihood method using a iterative algorithm. In order to compute the initial point estimates

of the parameters we consider the results obtained for the log-log model.
In particular, in GLMs the log-log link function is used since it is an asymmetric function (Agresti, 2000, pp. 248-250). In a log-log model the response curve is

$$\pi(x) = \exp[-\exp(\alpha_0 + \alpha_1 x)] \tag{1}$$

the quantile function of a Gumbel random variable. It is asymmetric, in particular $\pi(x)$ approaching 0 sharply but approaching 1 slowly.
Since the dependent variable is a rare event, we focus our attention on the tail of the response curve for the values closed to one. We propose to used the Generalized Extreme Value Distribution (GEV) to define the link function in GLMs with binary dependent variable. Finally, we apply the GEV regression to empirical data on Italian firms to model the default probability.

The extreme value regression model

The generalize extreme value cumulative distribution function is given by

$$F_X(x) = \exp\left\{-\left[1-\tau\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\tau}}\right\} \quad -\infty<\tau<\infty, \quad -\infty<\mu<+\infty \quad \sigma>0 \tag{2}$$

defined on $S_X = \{x : 1+\tau(x-\mu)/\sigma > 0\}$. The Type II (Fréchet-type distribution) and the Type III (Weibull-type distribution) classes of the extreme value distribution correspond respectively to the case $\tau > 0$ and $\tau < 0$, while the Type I class (Gumbel-type distribution) arises in the limit as $\tau \to 0$. The parameter τ is referred to as the shape parameter, while μ and $\sigma(>0)$ are location and scale parameters respectively.
We propose a generalization of the log-log model called Generalized Extreme Value (GEV) regression using as link function the quantile function of the GEV distribution.
For a binary response variable Y and an explanatory variable X let $\pi(x) = P\{Y = 1|X = x\}$. We define as response curve the following function

$$\pi(x) = \exp\{-[1+\tau(\vec{\beta}'\vec{x})]^{-1/\tau}\}. \tag{3}$$

with

$$\vec{\beta}' = [\beta_0, \beta_1, ..., \beta_k] \qquad \vec{x}' = [1, x_1, ..., x_k].$$

For $\tau \to 0$ the previous model (3) becomes the response curve of the log log model and for $\tau > 0$ it becomes the Weibull response curve, a particular case of the GEV one.
The link function for this GLM is $\frac{[-ln\pi(x)]^{-\tau}-1}{\tau} = \vec{\beta}'\vec{x}$.
We propose to estimate the parameters of these models by the maximum likelihood method. The response variable has a Bernoulli distribution that belongs to the exponential family. Moreover, the log-likelihood functions showed in the following subsections satisfy the well-known condition (Barndorff-Nielsen, 1978, pp. 151). Therefore, the maximum likelihood estimators exist and are unique.
Since the score functions do not have closed-form, a iterative algorithm is used. Let $\vec{y} = (y_1, y_2, ..., y_n)$ a simple random sample of size n from Y, the log-likelihood function is

$$l(\vec{\beta}, \tau, \vec{y}) = \sum_{i=1}^{n}\left\{-y_i[1+\tau(\vec{\beta}'\vec{x}_i)]^{-1/\tau} + (1-y_i)ln[1-\exp\{-[1+\tau(\vec{\beta}'\vec{x}_i)]^{-1/\tau}\}]\right\}. \tag{4}$$

The score functions are

$$\frac{\partial l(\vec{\beta}, \tau, \vec{y})}{\partial \beta_j} = -\sum_{i=1}^{n} x_{ij}\frac{\ln[\pi(\vec{x}_i)]}{1+\tau\vec{\beta}\vec{x}_i}\frac{y_i-\pi(\vec{x}_i)}{1-\pi(\vec{x}_i)} \quad j = 0, 1, ..., k, \tag{5}$$

$$\frac{\partial l(\vec{\beta}, \tau, \vec{y})}{\partial \tau} = \sum_{i=1}^{n}\left[\frac{1}{\tau^2}ln(1+\tau\vec{\beta}'\vec{x}_i) - \frac{\vec{\beta}'\vec{x}_i}{\tau(1+\tau\vec{\beta}'\vec{x}_i)}\right]\frac{y_i-\pi(\vec{x}_i)}{1-\pi(\vec{x}_i)}ln[\pi(\vec{x}_i)]. \tag{6}$$

Since the inverse of the link function (3) is a cumulative distribution function only for the values $\{\vec{x} : 1+\tau\vec{x}' > 0\}$, in order to identify the maximum likelihood estimators, we apply a constrained optimization. Moreover since the Fisher information matrix is not diagonal, the maximum likelihood estimators of the parameters $\vec{\beta}$ and τ are dependent and they can not be computed separately. In order to estimate the parameters by a iterative algorithm we compute the initial point estimates of the parameters of the GEV models. In particular we consider the results obtained for the log-log model.
We define the initial values $\vec{\beta}^*$ and $\tau*$. We propose to set $\tau^* \simeq 0$, $\beta_j^* = 0$ for $j = 1, ..., k$ and

$$\beta_0^* = \ln\left[-ln(\bar{y})\right]$$

obtained by the log-log model. We specify that the GEV regression for $\tau^* \to 0$ becomes the log-log model. Afterwards, by substituting the values in the equation β_0^* and $\beta_j^* = 0$ in the $\dfrac{\partial l(\vec{\beta}, \tau\vec{y})}{\partial \tau} = 0$ we obtaining the τ estimate for the first step of the iterative procedure. By using such estimate of τ and β_j^* values with $j = 0, 1, ..., k$ in the equation $\dfrac{\partial l(\vec{\beta}, \tau, \vec{y})}{\partial \beta_j} = 0$, we obtain the estimates of β_j with $j = 0, 1, ..., k$ for the first step in the GEV regression.
We conclude our methodological proposal analyzing the interpretation of the GEV regression model coefficients. The interpretation involves two issues: determining the functional relationship between the dependent variable and the independent variable and appropriately defining the unit of change for the independent variable.
The estimated coefficients for the independent variables express the the rate of change of a function of the dependent variable. If we consider the GEV model with only one independent variable, the link function is

$$g(\pi(x)) = \frac{\left[-ln\pi(x)\right]^{-\tau} - 1}{\tau} = \beta_0 + \beta_1 x.$$

Therefore, we deduce that in the GEV model it results $\beta_1 = g(\pi(x+1)) - g(\pi(x))$. For all fixed values of τ and β_0, if the parameter β_1 is positive, an increase of one unit in the independent variable is associated with a reduction of the $\pi(x)$ estimate. Otherwise, if β_1 is negative the dependent and independent variables have a direct functional relationship: an increase of one unit in the independent variable is associated with an increase of the $\pi(x)$ estimate.
Moreover, we analyze the parameter β_0: for all fixed values of τ and for a null independent variable, β_0 have a positive monotonic relationship with the dependent variable estimate. Finally, we analyze the influence of the τ parameter estimate on the $\pi(x)$ estimate and we find that for $\beta_0 = 0$ and for a null independent variable the $\pi(x)$ estimate is about equal to e^{-1} for all the values of τ. This means that the variation of $\pi(x)$ estimate depends only on the variations of the covariates and not on the variation of τ.

An Application to Credit Default

The GEV regression is proposed to model a binary dependent variable that represents a rare event. Since defaults in credit risk analysis are rare events, we apply the GEV model to empirical data on Italian Small Medium Enterprizes (SMEs) to model the default probability as a function of some covariates.
The estimate of default probability is a pivotal topic in the literature. The main cause of the importance of the prediction of the default probability is the Basel II Accord (Basel Committee on Banking Supervision (BCBS), 2004). In this framework, banks adopting the Internal-Rating-Based (IRB) approach are allowed to use their own estimates of PDs. Moreover, Basel II requires these banks to build a rating systems and provides a formula for the calculation of minimum capital requirements where the PD is the main input. For that reason, in many credit risk models such as CreditMetrics (Gupton et al., 1997), CreditRisk+ (Credit Suisse Financial Products, 1997) or CreditPortfolioView (Wilson, 1998), default probabilities are essential input parameters.
SMEs play a very important role in the economic system of many countries and particularly in Italy (about 90%

of Italian firms are SMEs (Vozzella, Gabbi 2010). Furthermore, a large part of the literature has focused on the special character of small business lending and the importance of relationships banking for solving information asymmetries. The informative asymmetries puzzle affects particulary SMEs for their difficulty to estimate and make known their fair value.

Compliant to Basel II, the default probability is one year forecasted. Therefore, let Y_t be a binary r.v. such that $y_t = 1$ if a firm is default at time t and $y_t = 0$ otherwise and let $\vec{x}$ be the covariate vector, we aim at estimating the conditional PD

$$\pi_t(\vec{x}) = P(Y_t = 1|\vec{x}_{t-1}).$$

Data used in our analysis comes from AIDA-Bureau van Dijk, a large Italian financial and balance sheets information provider. We consider defaulted and non defaulted Italian Small and Medium Enterprizes (SMEs) over the years $2005-2008$. In particular, since the default probability is one year forecasted, the covariates concern the period of time $2004-2007$. The database contains accounting data of around 210,000 Italian firms with total asset below 10 millions euro, compliant with Basel II. From these data we excluded the firms without the necessary information on the covariates.

In according with Altman and Sabato (2006) on this dataset we apply a choice-based or endogenous stratified sampling. In this sampling scheme data are stratified by the values of the response variable. We draw randomly the observations within each stratum defined by the two categories of the dependent variable (1=default, 0=non default). In particular, we consider all the units with value y=1 (firms in default). Then, we select a random sample of non-defaulted firms over the same period of defaults in order to obtain a percentage of defaults in our sample as close as possible to the default percentage for Italian SMEs (5 %). In order to analyse the properties of our model for different probabilities of the rare event $P\{Y = 1\}$, we consider also a default percentage of 1%.

In order to avoid the overfitting, data are randomly divided in two groups: the sample (1485 defaulters and 29700 non defaulters) on which we apply the models and a control-sample (165 defaulters and 3300 non defaulters) on which we evaluate the predictive accuracy of the models.

For this sampling there is an obvious dependency among the observations. If the dependency is not too great or if we appeal to a superpopulation model (see Prentice, 1986), then employing a theory ignoring is should not bias the results significantly. Therefore, since we have a big sample size, we consider this sample as a simple random sample.

Moreover we choose 16 covariates, financial and economic variables, selected in according to the recent literature (see Gabbi and Vozzella, 2010; Ciampi and Gordini, 2008; Altman et al., 2006) and to cover the most relevant areas of firm activity: leverage, liquidity and profitability.

At first, we apply the multicollinearity analysis to the covariates and than we choose those (7 variables) that result significant for the PD forecast in the logistic model. The variables are: Solvency ratio, Return on Investment, Turnover per employee, Cashflow, BanksTurnover and LabourCost/Added Value.

We compare the predictive accuracy of the GEV regression model here proposed with the one of logistic regression model. The predictive accuracy of these models is assessed using two performance measures: the Mean Square Error (MSE) and the Mean Absolute Error (MAE), defined as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{7}$$

where y_i and $\hat{y}_i$ are the actual and the predicted dependent variable on loan i, respectively. Models with lower MSE and MAE forecast the dependent variable more accurately.

Since the identification of defaulters is a pivotal aim for bank internal models, we focus our attention on the tail of the response curve for the values of the dependent variable equal to one, representing the default. Moreover, we propose the GEV regression model in order to overcome the drawback of the logistic regression in the underestimation of the default probability. For all these reasons we compare the two models by computing

MAE and MSE only for the defaulters. This means that in the equations (**??**) and (7) we consider only the positive errors $y_i - \hat{y}_i > 0$ where n is equal to the number of defaulters. We indicate them by MAE^+ and MSE^+, respectively.

Since the developed models may overfit the data, resulting in over-optimistic estimates of the predictive accuracy, the MSE and the MAE must be assessed on a sample which is different from that used in estimating the model parameters. A common technique to avoid overfitting the data consists of randomly dividing the data sample in two sets: one set (sample s) is used to fit the model and the other set (control sample cs) is used to test its accuracy. In particular we choose a sample size of the control sample equal to the 10% of the total sample size.

At this point, we compare the sample and the out-of-sample predictive accuracy of our proposal with the logistic regression model on AIDA data for different sample percentages of the rare event $\{Y = 1\}$. In particular we consider the probabilities of the rare event equal to 0.05 and 0.01 in order to analyse the performance of our proposal by varying the sample frequencies of the rare event.

Table 1 reports the MAE and the MSE for each model and for each sample frequencies of rare event. In order to evaluate the weights of the errors MSE^+ and MAE^+ on the respective total errors MSE and MAE, we compute the ratios MSE^+/MSE and MAE^+/MAE weighted by sample relative frequencies of the rare events (0.05 and 0.01); their values are reported in Table 1 between round brackets. By the results reported in Table

Sample percentage of $\{Y = 1\}$	*Error*	*Models*	
		GEV regression	*Logistic regression*
5%	MAE_s^+	0.4248(4.06%)	0.8829(52.49%)
	MSE_s^+	0.2080(3.40%)	0.8171(97.50%)
	MAE_{cs}^+	0.4171 (3.99%)	0.8702 (51.86%)
	MSE_{cs}^+	0.1967 (3.22%)	0.8067 (97.43%)
1%	MAE_s^+	0.3331(0.55%)	0.9502(45.46%)
	MSE_s^+	0.1399 (0.34%)	0.9270(89.13%)
	MAE_{cs}^+	0.3234 (0.53%)	0.9320(45.46%)
	MSE_{cs}^+	0.1301 (0.32%)	0.9084(89.05%)

Table 1: Forecasting accuracy measures of different models over different sample percentage of rare event $\{Y = 1\}$ on the sample and on the out-of-sample.

1 our proposal exhibits both the MAE and the MSE lower than the respective errors of the logistic regression model for both the sample and control sample and for each sample percentages of rare events. By comparing the percentages in round brackets we deduce that for the logistic regression model the weights of positive errors are relevant. On the contrary, for our proposals these very risky errors are irrelevant.

Since the errors for the sample and the control sample are similar, the covariates of these models are significant for default discrimination so the model is well-explained. Moreover, by comparing the errors for different sample percentages of rare events, our model improves its accuracy by reducing the occurrence probability of the rare event. On the contrary, the logistic regression model shows worst performance.

Since the main aim for banks is the forecasting of default probability, the two models are validated on a subsequent period. This means that the two models are fitted on data referring a period of time and the out-of-time predictive accuracy is measured on nondefault-defaults of 2009. In particular we evaluate the accuracy on the out of time considering the number of the firms in default for the year 2009 (64 defaulters) and we select randomly a sample of non defaulters (1354 non defaulters). Since the default probability is one year forecasted, the covariates concern the year 2008. In this case we estimate the model considering the total sample $n = 34650$.

We compare the predictive accuracy of our model with the logistic one calculating the MAE and MSE for the sample and out-of-time sample. The results in Table 2 shown both MAE and MSE lower than the respective

errors of the logistic model. The percentages in round rackets exhibits a relevant weight of the positive errors for the logistic regression and irrelevant very risky errors for our model. Moreover by comparing the errors for the different sample percentage of the rare event the logistic model shows a worst performance: the values of the MAE and MSE increase with a reduction of the percentage of the rare event. On the contrary, the GEV model improves its accuracy. For those reasons the GEV model can be consider a rare event regression model. Moreover, we compute estimate the coefficients of the models on a sample with 0.05 and 0.01 sample relative

Sample percentage of $\{Y=1\}$	*Error*	*Models*	
		GEV regression	*Logistic regression*
5%	MAE_s^+	0.4282(4.14%)	0.8815(52.53%)
	MSE_s^+	0.2103(3.50%)	0.8161(97.61)
	MAE_{cs}^+	0.4375 (3.96%)	0.9100 (48.35%)
	MSE_{cs}^+	0.2119 (3.19%)	0.8489 (94.74%)
1%	MAE_s^+	0.3509(0.58%)	0.9478(45.53%)
	MSE_s^+	0.1541 (0.39%)	0.9246(88.90%)
	MAE_{cs}^+	0.3393 (0.55%)	0.9797(47.56%)
	MSE_{cs}^+	0.1369 (0.33%)	0.9601(88.90%)

Table 2: Forecasting accuracy measures of different models over different sample percentage of rare event $\{Y=1\}$ on the sample and the out-of-time.

frequencies of defaulters and we compute the errors on a control sample with 0.01 and 0.05 sample relative frequencies of defaulters, respectively. By computing the MAE and MSE on all the loans, the errors of our model do not change in comparison with the errors calculated with a control sample with 0.05 and 0.01 sample relative frequencies of defaulters, respectively. This means that our model is robust for the sample relative frequencies of defaulters. On the contrary, for the logistic regression models the previous ratios are significantly different from one .

Therefore, The application shows that the logistic regression model underestimates the default probability. On the contrary, the GEV model overcomes this problem and accommodates skewness.

REFERENCES

Agresti A. (2002). Categorical Data Analysis. Wiley, New York.

Altman, E. and Sabato, G. (2006). Modeling Credit Risk for SMEs: Evidence from the US Market, ABACUS, 19(6), 716-723 .

Barndorff Nielsen O. (1978). Information and exponential families in statistical theory. Wiley, New York.

Ciampi F., GordiniN. (2008). Using Economic-Finantial Ratios for Small Enterprize Default Prediction Mdeling: an Empirical Analysis. Oxford Business & Economics Conference, Oxford.

Credit Suisse Financial Products (1997). CreditRisk+: A Credit Risk Management Framework, Credit Suisse First Boston.

Gupton G. M., Finger C. C., Bhatia M., 1997. CreditMetrics. Technical document, J. P. Morgan.

King G., Zeng L. (2001). Logistic Regression in Rare Events Data. Political Analysis, 9, 137-163.

Kotz S. and Nadarajah S. (2000). Extreme Value Distributions. Theory and Applications, Imperial Colleg Press, London.

McCullagh P., Nelder J.A. (1989). Generalized Linear Model, Chapman Hall, New York.

Prentice R.L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika,66, 403-411.

Vozzella P., Gabbi G. (2010). Default and Asset Correlation: An Empirical Study for Italian SMEs. Working Paper.

Wilson T. C. (1998). Portfolio credit risk. Economic Policy Review 4, 71-82.