Variable selection in competing risks model

Alessandra Amendola and Marialuisa Restaino

Abstract Our aim is to investigate the performance of different variable selection methods, focusing on a statistical procedure suitable for the competing risks model. In this setting, same variables might have different degrees of influence on the risks due to multiple causes and this effect has to be taken into account in the choice of the "best" subset. The proposed procedure, based on shrinkage techniques, has been evaluated by means of empirical analysis on default risk predictions.

Key words: competing risks model, variable selection, shrinkage techniques.

1 Introduction

Since the seminal works of [1], the analysis of firms' survival are increasingly investigated in the corporate failure literature. The main interest is in building a model that is able to predict the firms' potential ending up in financial distress. However, financially distressed companies may exit the market for several reasons (bankruptcy, liquidation, merge, acquisition, etc.) and a challenging task is to identify which financial indicators may influence each reason of exit. Different variable selection techniques, such as information criterion, stepwise procedure and penalized regression, have been used for selecting predictors in different statistical frameworks (discriminant analysis, logistic regression, neural networks and survival analysis). However, only few of them have been considered in competing risks models. The aim of this paper is to investigate the determinants of the probability of different types of firms' market exit through a competing-risks hazard model, focusing in particular on the variable selection problem. We propose to fit a competing risk model

Alessandra Amendola
Department of Economics and Statistics, University of Salerno, Via Ponte don Melillo, 84084 Fisciano, Salerno (Italy) e-mail: alamendola@unisa.it

Marialuisa Restaino
Department of Economics and Statistics, University of Salerno, Via Ponte don Melillo, 84084 Fisciano, Salerno (Italy) e-mail: mlrestaino@unisa.it

by maximizing the marginal likelihood subject to a shrinkage-type penalty that encourages sparse solutions and hence facilitates the process of variable selection. The proposed approach has been compared to traditional stepwise procedure and their performance has been evaluated through an empirical analysis on a data-set of financial indicators computed from a sample of industrial firms' annual reports. The rest of the paper is structured as follows. In the next section, the statistical method is briefly reported. The empirical results are discussed in Sect. 3.

2 Statistical framework

Let $T = \min(\tilde{T}, C)$ be the observed time, which is the length of time from the predefined time of origin until failure $\tilde{T}$ or censoring C, and let D be the cause of failure. The main feature of competing risks is that from a given set of K causes, one and only one cause can be assigned to every failure. The probabilistic aspect in modeling the competing risks is the joint distribution of T and D ([4]), which is specified through the *cause-specific hazard function*, defined as the probability of failing due to a given cause k, after the time point t has been reached. It is given by:

$$\lambda_k(t) = \lim_{\Delta t \to 0} \frac{P[T \leq t + \Delta t, D = k | T \geq t]}{\Delta t}, \qquad k = 1, \ldots, K.$$

Since the cause-specific hazards are identifiable and may depend on a set of covariates, a regression on them is possible. The cause-specific hazard of cause k for a subject i is given by:

$$\lambda_{ik}(t|Z_{ik}) = \lambda_{k,0}(t) \exp\{\underline{\beta}_k^T \underline{Z}_{ik}(t)\}, \tag{1}$$

where $\lambda_{k,0}(t)$ is the *baseline cause-specific hazard* of cause k which is not necessary to be explicitly specified, $\underline{Z}_{ik}(t)$ is a vector of covariates for individual i specific to k-type risk at time t, and the vector $\underline{\beta}_k$ represents the regression coefficients of cause k to be estimated. Since the same variables could have different effects on the different risks, it is reasonable to assume that for each k $\underline{\beta}_k$ is independent of each other. In order to have an estimate of the coefficients' vector, we build the partial likelihood, using the same procedure available in univariate Cox Proportional Hazard Model ([3]). Let $0 < t_{1k} < t_{2k} < \cdots < t_{uk}$ be ordered distinct time points at which failures of any causes occur for the risk k. Assume that only one failure can happen at each failure time, i.e. there are no tied failure times in the data. The partial likelihood for specific hazard k is given by:

$$L_k(\underline{\beta}_k) = \prod_{i=1}^{n_k} \frac{\exp\{\underline{\beta}_k^T \underline{Z}_{ik}(t_{ik})\}}{\sum_{l \in R(t_{ik})} \beta_k^T \underline{Z}_{lk}(t_{lk})},$$

where n_k is the number of individuals in specific hazard k, and $R(t_{ik}) = \{l | t_{lk} \geq t_{ik}\}$ is the set of individuals at risk at time t_{ik}. The overall partial likelihood function is:

$$L(\underline{\beta}_1,\ldots,\underline{\beta}_K)=\prod_{k=1}^{K}L_k(\underline{\beta}_k). \tag{2}$$

Since not all the covariates may contribute to the prediction of survival outcomes, the problem of interest is to identify the subset of variables that are significantly associated with each failure type. Our aim is to investigate this problem by extending the lasso technique in the competing risks model. Based on [5], the lasso for the failure k is given by:

$$\hat{\underline{\beta}}_k=\underset{\underline{\beta}_k}{\operatorname{argmax}}\ \log L_k(\underline{\beta}_k)=\underset{\underline{\beta}_k}{\operatorname{argmax}}\ \sum_{i}^{n_k}\left[\exp[\underline{\beta}_k^T\underline{Z}_{ik}(t_{ik})]-\log\sum_{l\in R(t_{ik})}\exp[\underline{\beta}_k^{'}\underline{Z}_{lk}(t_{lk})]\right]$$

subject to $||\underline{\beta}_k||_1\leq\rho_k$, where $||\underline{\beta}_k||_1=|\beta_k^1|+|\beta_k^2|+\cdots+|\beta_k^p|$ is the L_1 norm of the coefficients vector $\underline{\beta}_k$ for the failure cause k, and ρ_k is the tuning parameter which quantifies the magnitude of the constraints and determines the number of coefficients estimated as zero in the model. The lasso estimation in presence of all failures is given by:

$$(\hat{\underline{\beta}}_1,\ldots,\hat{\underline{\beta}}_K)=\underset{\underline{\beta}_k}{\operatorname{argmax}}\ \log L(\underline{\beta}_1,\ldots,\underline{\beta}_K)$$

subject to $||\underline{\beta}_k||_1\leq\rho_k\quad$ for $\quad k=1,\ldots,K$. In this case, a different tuning parameter for each type of failure has been considered.

3 Empirical results

The empirical analysis is performed on a data set of Italian firms operating in the building sector in the period 2004-2009[1]. As competing risks we consider three different mutually exclusive states of exit from the market: bankruptcy, voluntary liquidation and inactivity[2]. From the overall population of active and financially distressed firms, we select a sample of $n=1462$ firms, based on the geographical distribution of the industrial firms within the region. The final sample consists of 221 companies that went bankrupt, 129 that had entered voluntary liquidation, 228 inactive and 884 active firms. The sample is divided into two parts: *in-sample set*, used for the classification ability, in order to determine how accurately a model classified businesses, and *out-of-sample set*, used for prediction accuracy. Two predictions' windows are considered: 1-year ahead and 2-years ahead.

We perform competing risks models, in which variables are selected by the lasso and the classic stepwise approach, in order to investigate and compare their performance. Then we also investigate the effect of some strategic factors on the probability of exit the market for different reasons and compare the determinants of

[1] The information on individual firms and on their financial information are obtained from the Amadeus database, provided by Bureau van Dijk.

[2] The last state includes those firms that exit the database, but it is unknown the reason for the exit.

various exit routes. The predictive performance of the developed models are evaluated by means of some accuracy measures (i.e. C.C.R., AUC, etc.) ([2]). Results are displayed in Table 1 for in-sample and out-of-sample sets, with respect to the two time-windows. Looking at the in-sample set and considering the stepwise procedure and the lasso method separately, it can be noted that the correct classification rate is slightly higher for bankruptcy, inactive and liquidation than the single-risk and it increases at approaching the year of exit. Moreover, the AUC, which considers the impact of the I and II type errors, has a very discriminative power for bankruptcy and liquidation, compared to the the single-risk framework, and it increases its power as the failure time is approaching. It can also be observed that the lasso has a better performance than the stepwise procedure, not only for the pooled model but also for the competing risks framework. In fact, the CCR in lasso is higher for all three states and it increases at approaching the failure year. Finally, the AUC has a similar behaviour. For evaluating the predictive power of the models, we refer to the out-of-sample set. Looking at the results, we notice that even though the CCR and the AUC are higher for the single-risk model than for the other states when variables are selected by stepwise procedure, these measures improve their performance when the lasso technique is used.

Table 1 Accuracy measure.

	In sample 1-year ahead				Out-of-sample 1-year ahead			
	Stepwise method				Stepwise method			
	Bankruptcy	Inactive	Liquidation	Single-Risk	Bankruptcy	Inactive	Liquidation	Single-Risk
Correct Class Rate	0.76185	0.78819	0.78262	0.74881	0.82707	0.85263	0.84211	0.87218
Type I Error	0.40930	0.58025	0.47353	0.61441	0.66667	0.56250	0.44068	0.52564
Type II Error	0.22173	0.17134	0.19132	0.11071	0.17069	0.13713	0.13036	0.07496
AUC	0.75788	0.68797	0.73229	0.73459	0.54280	0.68957	0.77105	0.72044
	Lasso method				Lasso method			
Correct Class Rate	0.83557	0.81371	0.84399	0.75234	0.92632	0.89774	0.90526	0.88421
Type I Error	0.73798	0.66392	0.65147	0.72103	0.33333	0.62500	0.54237	0.67949
Type II Error	0.10938	0.13382	0.10561	0.06458	0.07251	0.08937	0.05116	0.04089
AUC	0.68151	0.67748	0.73091	0.72669	0.89350	0.70700	0.78829	0.72985
	In sample 2-year ahead				Out-of-sample 2-year ahead			
	Stepwise method				Stepwise method			
Correct Class Rate	0.76445	0.77875	0.77530	0.73571	0.82350	0.85118	0.83522	0.83677
Type I Error	0.42384	0.56832	0.49301	0.61593	0.11765	0.65347	0.48503	0.59615
Type II Error	0.21481	0.18020	0.19688	0.11434	0.17756	0.12113	0.13465	0.08037
AUC	0.75523	0.67646	0.72619	0.73229	0.81073	0.70975	0.75931	0.72773
	Lasso method				Lasso method			
Correct Class Rate	0.77595	0.79780	0.84116	0.73374	0.83265	0.87848	0.89392	0.84655
Type I Error	0.53974	0.63354	0.67133	0.71484	0.47727	0.68317	0.65868	0.73397
Type II Error	0.18974	0.15118	0.10569	0.07498	0.16017	0.09071	0.05408	0.04233
AUC	0.72987	0.67373	0.72515	0.72396	0.78576	0.70850	0.76421	0.72483

References

1. Altam, E.I.: Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. J. of Finance **23**, 589–609 (1968)
2. Amendola, A., Restaino, M., Sensini, L.: Variable selection in default risk model. J. of Risk Model Valid. **5**(1), 3–19 (2011)
3. Cox, D.R.: Partial likelihood. Biometrika **62**(2), 269–276 (1975)
4. Crowder, M.J.: Classical Competing Risks. Chapman and Hall/CRC Press, London (2001)
5. Tibshirani, R.: The lasso method for variable selection in the Cox model. Stat. Med. **16**, 385–395 (1997)