
Greek Statistical Institute
Proceedings of the 33rd Panhellenic Statistics Conference (2021), pp. 000−000

UNFOLDING MODELS FOR ORDINAL DATA IN CYBER RISK ASSESSMENT

S.Facchinetti[1], M Iannario[2], S.A. Osmetti[1], C. Tarantola[3]
[1]Università Cattolica del Sacro Cuore, Milano, Italy
{silvia.facchinetti, silvia.osmetti}@unicatt.it
[2]University of Naples Federico II, Napoli, Italy
maria.iannario@unina.it
[3]University of Pavia, Pavia Italy
claudia.tarantola@unipv.it

ABSTRACT

In an increasingly digitalized world, where organizations are affected by technological evolution, cyber attacks are multiplying rapidly. They have an impact on every class of business and no industry can consider itself immune to them. Quantitative loss data are rarely available while it is possible to obtain a qualitative evaluation, expressed on a rating scale, from experts of the sector. Hence, we focus on ordinal data models for cyber risk evaluation (rating) with particular emphasis on a mixture model taking into account the uncertainty in the process of scoring. We examine a set of data regarding cyber attacks that occurred worldwide before and during the pandemic due to Covid-19. The aim of our analysis is to investigate if Covid-19 has affected experts' uncertainty and assessment, and identify the relevant factors which influence the severity of an attack.

Keywords: Cyber risk; CUP models; Rating; Uncertainty

1. INTRODUCTION

Throughout the last years the use of statistical modelling in the analysis of cyber risk assessment has gained a rapidly increasing interest. While quantitative loss data are rarely available, a qualitative evaluation of the level of severity of an attack, expressed on a rating scale, from experts of the sector is able to be obtained. In this way, we can identify which types of attacks are the most

dangerous.

The rating assigned by the expert can be considered as the final outcome of a complex activity based on knowledge of the topic, collection of information but also instinct and feeling of the expert himself.

In this contribution we rely on ordinal mixture models to mimic the skilled decision-making process. In particular, we exploit the CUP mixture introduced by Tutz *et al.* (2017). It is based on a *C*ombination of an incertitude in the process of assessment (*U*ncertainty component) and a deliberate choice based on the evaluation of the respondent (*P*erception component). The latter component accounts for reasoned judgments toward the attack under evaluation as well as the set of experts' perceptions and information connected with it. The uncertainty component accounts for other unreasonable elements such as the difficulty in expressing a rating regarding a specific event about which the expert has not a clear opinion or has a limited set of information. Furthermore, it is also related to the amount of time devoted to the judgment or experts' laziness, boredom or circumstances. The mixture can be considered as a combination of the distributions of a discretised version of the underlying continuous latent variables describing these different components.

We implement the CUP mixture in rating systems for the analysis of risk assessment of worldwide cyber attacks occurred in the period 2018-2020 years (before and during the Covid-19 pandemic) generalizing the main findings in Facchinetti *et al.* (2020, 2021) referred to 2017-2019 data. The proposal discussed here extends the previous contributions from two directions: the analysis of uncertainty in the experts' assessment and the evaluation of epidemic period by considering the new tools affecting the process of scoring.

The plan of the paper is as follows. Section 2. is devoted to the description of the considered model. In Section 3. we present the data and discuss the obtained results. The paper ends with some concluding remarks and avenues for future research.

2. METHODOLOGY

In this section we briefly review standard CUP models. The perception component is described via a cumulative link model under the proportional odds assumption (McCullagh, 1980) while a discrete Uniform distribution is used for the uncertainty component.

More precisely, the observed rating s (severity level) assigned to a specific cyber attack can be considered as a realisation of a random variable S with probability distribution

$$P(S_i = s|\boldsymbol{x}_i) = \pi P_M(Y_i = s|\boldsymbol{x}_i) + (1-\pi)P(U_i = s), \quad s = 1, 2, \ldots, m. \quad (1)$$

As earlier pointed out, the preference component $P_M(Y_i = s|\boldsymbol{x}_i)$ is defined via a cumulative link model on an appropriate row vector of covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ij}, \ldots x_{ip})$. Formally, we have

$$link\,[P_M(Y_i \leq s|\boldsymbol{x})] = \alpha_s - \boldsymbol{x}_i\boldsymbol{\gamma} \quad i = 1, 2 \ldots, n; \quad s = 1, 2 \ldots, m-1,$$

where $\boldsymbol{\gamma}$ is the parameter vector for the preference component, whereas $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_m = \infty$ represent the thresholds of the scale of the latent variable Y^* behind Y. Among the alternative choices for *link* functions we focus on the logit one for easiness of interpretation and robustness properties. The uncertainty component is modelled as $P(U_i = s) = 1/m$. The two components are then combined via the parameter π eventually depending on a vector of covariates $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{ij}, \ldots w_{iq})$, with a possible non empty intersection with $\boldsymbol{x}_i$. To model the effect of the covariates on the uncertainty component we use a logit link as well, $\pi = \pi(\boldsymbol{\beta}) = 1/(1 + e^{-\boldsymbol{w}_i\boldsymbol{\beta}})$, where $\boldsymbol{\beta}$ is the parameter vector for the related component. The standard cumulative link model is a special case of (1) with $\pi = 1$.

From an inferential point of view, a way to obtain stable estimates is to consider the mixture as a problem with incomplete data and use the EM algorithm (Dempster *et al.*, 1977); see the Appendix of Tutz *et al.* (2017) for further details.

3. APPLICATION TO CYBER RISK DATA

We consider a set of data collected by the experts of the "Hackmanac" society (`https://hackmanac.com/`). Hackmanac is a company based in Dubai that monitors the evolution of real global cyber threats with the aim to support companies and institutions to define their cyber defense strategy.

In particular, we investigate a sample of more than 5.000 statistical units regarding cyber attacks occurred worldwide during the years 2018, 2019 and 2020. For each attack we have information regarding the following "macro variables": `Type of attack` (main actors and motivations of the attack), `Attack`

`Technique` (adversary tactics and techniques of attack), `Target Class` (victims of cyber attack), `Continent` (where the attacks took place), and `Severity` (an ordinal classification of the gravity of an attack). Indeed, experts classify the gravity of an attack on the basis of their knowledge by the ordinal variable `Severity` assuming values 1 (low severity), 2 (medium severity), 3 (high severity) and 4 (critical severity).

The evaluation of each attack on the basis of its seriousness is the outcome of a complex activity based on various aspects such as awareness and experience about the geopolitical, social, economic, and image impact on the victims, but also sensation and feeling of the expert himself. Due to the characteristics of this decision-making process that combines knowledge of the examined event and expert awareness, in this paper we rely on CUP models.

Based on a preliminary model selection analysis, we decided to include in our model the following binary covariates for each "macro variable":

- Cybercrime as the candidate `Type of attack`
- Information and communication technologies (ICT) and Government-Military-Law-Enforcement (GOV) as `Target Class`
- Target group (TG) as an indicator of those statistical units compromised by lower than 3 attacks
- Vulnerabilities as `Attack Technique`
- `Continent` as a factor variable with Africa as reference level (AF=Africa, AM=America, AS=Asia, EU=Europa, OC=Oceania, MC=Multiple continent)
- `Covid` a dummy variable representative of the pandemic period.

In Table 1 the estimated values of the parameters and the asymptotic standard errors (in brackets) for the examined CUP model are reported. The latter were computed via (numerical) Hessian. The symbol "$\star$" indicates that the corresponding parameter is not significant at 5% level. Furthermore, the Bayesian Information Criterion (BIC) index (Schwarz, 1978) of the selected model is 12619.46 whereas the BIC index of the nested standard cumulative model is 12874.14 highlighting the added value of the proposal.

Before commenting out the obtained results, we recall that the weight of the uncertainty component in the CUP mixture is equal to $(1 - \pi)$.

`Covid` is the only covariate affecting both components of the CUP mixture; it influences negatively both the uncertainty and the perception. Thus, the fitted

Table 1: *Estimated* CUP *model for cyber risk analysis.*

Uncertainty component	
β_0	0.973 (0.142)
Covid	2.676 (0.291)
Perception component	
α_1	-2.330 (0.496)
α_2	-0.409 (0.493)
α_3	2.211 (0.478)
`Type of attack`	
Cybercrime	-1.735 (0.134)
`Target Class`	
ICT	0.866 (0.106)
GOV	0.979 (0.116)
`Attack Technique`	
Vulnerabilities	0.898 (0.119)
`Continent`	
AM	0.972 (0.458)
AS	1.809 (0.456)
EU	1.008 (0.465)
MC	0.221 (0.470) $\star$
OC	0.942 (0.508)
`Target Group`	
TG	0.411 (0.082)
`Period`	
Covid	-0.335(0.070)

model indicates a low level of uncertainty in the severity evaluation during the pandemic. This result is probably due to a more accurate evaluation expressed on cyber attacks during the Covid period than before. With regards to the perception component, we observe a lower probability to obtain an elevate level of severity during the epidemic than the previous period.

Furthermore, we observe that also Cybercrime has a negative influence on the experts' risk perception. This result is consistent with the analysis of Facchinetti

et al. (2020, 2021) that pointed out that even if this type of offensive is quite frequent, in terms of gravity it determines attacks of minor severity. With reference to `Target Class`, `ICT` and `GOV` are associated with a higher probability of a critical severity attack with respect to the others.

The parameter associated to the examined category of `Attack technique` is positive. This indicates that the exploitation of system vulnerabilities (weakness that can be used to gain unauthorized access to a computer system) can led attacks scored with a high level of severity.

On the subject of `Continent`, the parameter related to category MC is not significant. This could be explained by the fact that attacks directed against single continents are more effective than the ones involving more of them. Moreover, we observed a lower severity level for Africa (the baseline level) with respect to the other ones.

Finally, an attack directed to a victim belonging to a target group compromised by more than 3 attacks determines a substantial higher level of severity.

4. CONCLUSION

In this paper we illustrated how CUP models can be an useful instrument for cyber risk evaluation.

The CUP mixture allows to improve results with respect to the classical assumption of the standard models used for the analysis of ordinal (rating) data by means of the added value of the uncertainty component which also represents an advantage over classical mixture models. In comparison with the latter the components are fully specified and are not from the same class of models; see among others Greene and Hensher (2003), Grün and Leisch (2008), Breen (2010). More specifically, when the uncertainty component is neglected, the strength of covariates tends to be underestimated. In addition, when uncertainty is very high, the study of the perception component without the assessment of the uncertainty one causes a loss of information and a misspecification of the model.

Further analyses will be devoted to the probability-based measures for comparing clusters (`Target group`) on ratings, while adjusting for other explanatory variables as also reported in Iannario and Tarantola (2021a, 2021b), and to a hierarchical modelling structure taking into account the homogeneity of the clusters related to several countries.

Acknowledgements: We acknowledge support from the European Cost Action "CA19130 - Fintech and Artificial Intelligence in Finance - Towards a transparent financial industry"

REFERENCES

Breen, R., Luijkx, R. (2010). Mixture models for ordinal data. *Sociological Methods and Research*, **39**, 3–24.

Dempster, A.P., Laird N.M. and Rubin, D.B.(1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Greene, W., Hensher, D. (2003). A latent class model for discrete choice analysis: Contrasts with mixed logit. *Transportation Resesearch*, Part B, **39**, 681–689.

Grün, B., Leisch, F. (2008). Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of Classification*, **25**, 225–247.

Facchinetti, S., Osmetti, S.A., Tarantola, C. (2020). How to perform cyber risk assessment via cumulative logit models, *Book of short papers - SIS 2020*, 1083-1086.

Facchinetti, S., Osmetti, S.A., Tarantola, C. (2021). A statistical approach for assessing cyber risk via ordered response models, under review for international journal.

Iannario, M. and Tarantola, C. (2021a). Effect Measures for Group Comparisons in a Two-Component Mixture Model: A Cyber Risk Analysis. In Balzano, S., Porzio, G.C., Salvatore, R., Vistocco, D., Vichi, M. (Eds.) *Statistical Learning and Modeling in Data Analysis*. Springer Nature Switzerland AG. Springer Nature Switzerland AG.

Iannario, M., Tarantola, C. (2021b). How to Interpret the Effect of Covariates on the Extreme Categories in Ordinal Data Models, *Sociological Methods & Research*, https://doi.org/10.1177/0049124120986179.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Ser.* B, 42, 109-142.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Tutz, G., Schneider, M., Iannario, M., Piccolo, D. (2017). Mixture models for ordinal responses to account for uncertainty of choice, *Advances in Data Analysis and Classification*, **11**, 281-305.